The Simple and Infinite Joy of Mathematical Statistics

J.N. Corcoran

The Simple and Infinite Joy of Mathematical Statistics

Author: J.N. Corcoran

Institute: University of Colorado, Boulder

Date: June 29, 2022

Version: 5.0

Congratulations on deciding to learn Mathematical Statistics, or, as those in the know like to call it, “MathStat”.

I’m going to let you in on a not so little secret. MathStat isn’t really statistics– it’s probability. Yes, there is a difference.

Probability is about the future.

Suppose that you have an unfairly weighted coin that, if flipped, will result in “heads" 57% of the time. Suppose that **you are going** to flip this coin 10 times. What is the probability that you **will** see at least 7 heads? How many heads **do you expect** to see?

Did you see all of those future type words and phrases for describing probability?

Statistics, on the other hand, **is about the past**.

That is, suppose that you know next to nothing about the coin but you flipped it 10 times and observed the heads/tails *data* as

$$H, H, T, T, T, H, H, T, H, H.$$

You can use **statistics** to attempt figure out (albeit with some uncertainty) whether or not the coin **was** fair in the first place and what that single flip heads probability **might have been**.

Do you see how statistics is about looking **back** to figure out what was going on with the coin?

Much of the subject of statistics is about “reverse engineering” the data generating process in the case that the process involves some randomness or probability. Probability can stand alone as a subject, but in order to do statistics we need to understand probability.

Mathematical Statistics is the probability needed to do statistics.

So you see, you have really decided to take a probability course and not a statistics course. MathStat is a specialized area of probability, though, and before you can learn it you need to first know some basics. This is why a introductory probability course, is **an important prerequisite**.

Statistics, on the other hand, is not a prerequisite for learning Mathematical Statistics. However, if you’ve studied or used statistical analysis techniques, be prepared to see some really cool connections ahead! If you’ve ever had a course in statistical analysis where you’ve estimated things (for example averages or medians) or performed hypothesis tests that involved looking numbers up in, for example, “t-tables", this course will tell you why you did those things, when those techniques actually are valid for a problem, and what do do if they aren’t. Mathematical Statistics is an intense theoretical course in which sometimes even the link to the “data stuff" will seem completely obscured, but it is, in a word, **beautiful**.

This text evolved from notes that were originally designed for a course in Mathematical Statistics in the department of Applied Mathematics at the University of Colorado, Boulder. The course was cross-listed for

graduate students and senior undergraduates. The prerequisite is a basic undergraduate probability course. As the audience background has been quite varied, and adherence to prerequisites is often tenuous, I have included a “Preliminaries" chapter that may turn more advanced students with its simplicity. Rest assured that things will ramp up quickly after that, and that many will want to skip the first chapter altogether. (It may still be useful to skim through for the purpose of picking up notation though!) I believe that these notes are well suited for a graduate course in Mathematical Statistics even in a pure Statistics Department, where I have also had the experience of teaching MathStat. It includes many advanced concepts, examples, and challenging problems.

Enjoy, and do good things, always.

How to Read This Text

Read this text "with vim and vigor"!

Chapter 0 is a review of the probability concepts needed to understand this text. If you are taking a course in Mathematical Statistics, it is likely that this material is already known to you and you might want to skip ahead to Chapter 1. However, it still may be worth scanning Chapter 0 in order to familiarize yourself with notation we will use in this text. Section 0.8 is highly recommended reading for everyone.

Some Chapters end with a "Postscript" section. These are Sections that we included for completeness but that we would not, ourselves, cover as instructors of this course. For example, in Section 1.4, we talk about the minimum and maximum values in a random sample. We will use minimums and maximums extensively in estimation problems throughout the entire text. We will not be needing to talk about, for example, the "second smallest" value in a sample. However, Section 1.6 covers this anyway, for completeness, in a discussion about "order statistics". For the sake of "flow" in your Mathematical Statistics journey, we highly recommend skipping the "Postscript" sections. You can reference them later if needed.

Sections with an asterisk ($*$) at the end of the Section title are slightly more advanced or tedious, containing "omitted proofs", and you may decide to read them or skip them without affecting your ability to understand subsequent material. Including or not including these sections can be the difference between an advanced undergraduate course and a first year graduate course.

Contents

Chapter 0 Probability Preliminaries

This chapter is **very** elementary and is not at all representative of the overall level of this text. This chapter is also a strange sort of introduction to probability in its coverage. Some very basic central concepts and even notation for understanding probability are completely absent. Instead, it contains only the bare minimum of topics needed in order to understand the subsequent chapters. For those who insist on trying Mathematical Statistics without a basic probability background, this should help. For others already familiar with things like random variables,"pdfs", "cdfs", and expectations, it still may be helpful to only skim Sections 0.1 through 0.7 since they will serve to establish some basic terminology and notation. Sections 0.8-0.10 are recommended reading for everyone!

0.1 Between Zero and One

If you flip a "fair" coin, what is the probability that it will come up "heads"?

Most people have at least a sense of what this question is asking without any formal training in probability. Many will use the words "probability" and "chance" interchangeably and say:

"There is a 50 percent chance of it coming up heads."

This is not wrong but, **formally**, in mathematics

Probability is a number between 0 and 1.

So, the more precise answer to the original question is $1/2$ or 0.5.

If you roll a "fair" six-sided die, what is the probability that you will see a 5?

In the case of the coin and now the die, the word "fair" is there to say that there is nothing funny going on. The coin isn't warped, the die isn't shaved or repainted with extra dots, and the acts of flipping and rolling are not done in a way that favors one outcome over another. For the coin, the two outcomes "heads" and "tails" are **equally likely**, and for the die, the six outcomes are equally likely.

To answer the question, since the outcome we care about (getting a 5) is one outcome out of six equally likely outcomes, the probability of getting a 5 is $1/6$.

If you roll a fair six-sided die, what is the probability that you will see either a 5 or a 6?

Since there are now 2 outcomes out of 6 equally likely outcomes that we care about, the answer is $2/6$ or $1/3$.

Note

In the case of equally likely outcomes for an experiment, the probability that a certain event occurs is

the number of outcomes in the event of interest
divided by
the total number of possible outcomes.

Thus, in order to compute probabilities in the case of equally likely outcomes, it is important for us to be able to count things.

0.2 Counting

0.2.1 Things in a Row

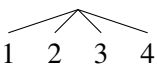
How many different ways can you list the integers 1 through 4 in various orders? Two examples are

$$3\ \ 2\ \ 1\ \ 4 \qquad \text{and} \qquad 2\ \ 4\ \ 1\ \ 3,$$

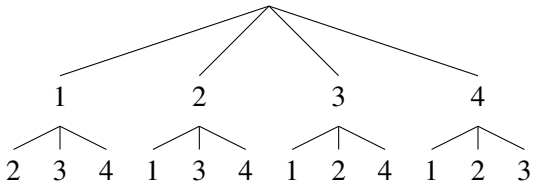
but there are obviously many more possibilities.

Listing things in various orders can help us with the problem of assigning labels to things. For example, suppose you and your three friends are going to assign yourselves administrative titles in your treehouse club. The titles are president, vice president, secretary, and treasurer. You are going to list your names on a piece of paper. The first person on the list will be president, the second will be vice president, and so on. How many different ways can you do this? This is exactly the same as the “1 2 3 4 problem" above– especially so if your names happen to be “1", “2", “3", and “4".

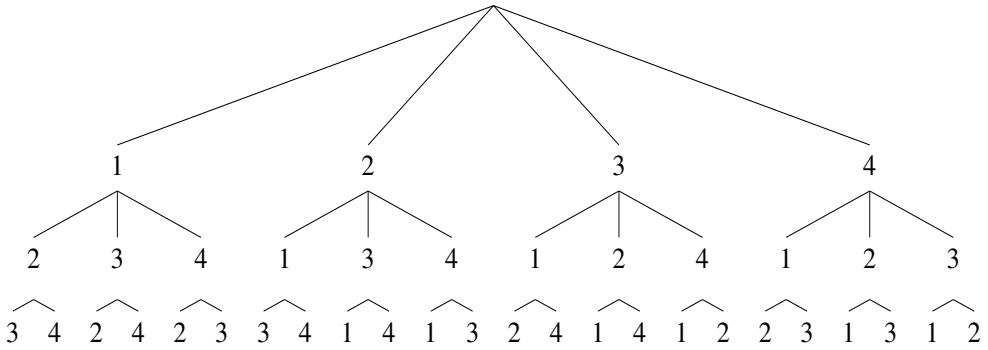
The first number in the list can be 1 or 2 or 3 or 4. Let’s visualize them as branches of a tree.



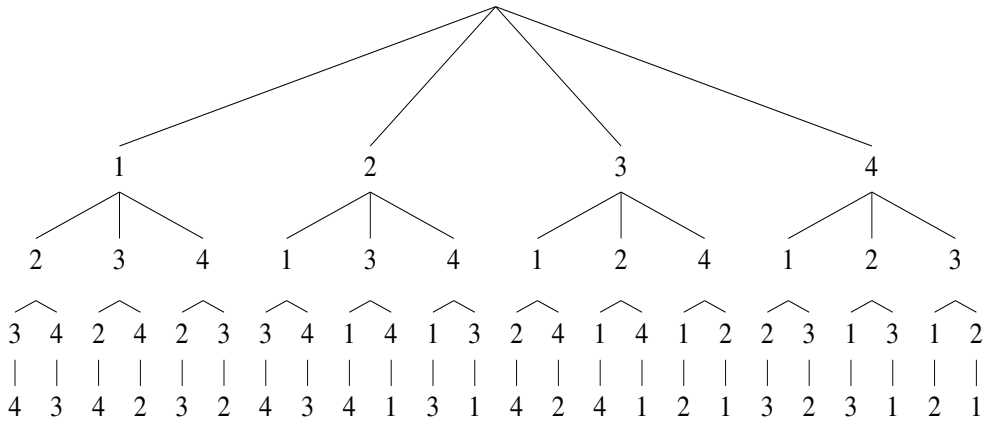
If we start our list with 1, for example, the next number can be 2 or 3 or 4. If we start our list with 2, the next number can be one of 1 or 3 or 4. In general, possibilities for “next numbers", can be envisioned like this



If our first number is 1 and our second number is 2, our third number can either be 3 or 4. So...more branches.



Finally, if we start our list with $2, 1, 4$, for example, we have no choice but to finish up with 3. Thus, we get the final “leaves" on our tree.



Following any path from top to bottom will give a different configuration of the numbers 1, 2, 3, and 4 and all possible configurations are shown. There are 4 branches at the top. Each has 3 offshoots, making $4 \times 3 = 12$ “second level branches". Each of those has 2 offshoots, making $4 \times 3 \times 2 = 24$ “third level branches" and finally, each of those has one final offshoot making $4 \times 3 \times 2 \times 1 = 24$ total branches representing configurations of 1, 2, 3, and 4. (Of course, we do not need to multiply by 1, but it gives a nice sense of completeness don’t you think?).

Definition 0.2.1

For any positive integer n, the symbol "**n**!" is read as "**n factorial**" and is defined as

$$n! := n(n-1)(n-2)\cdots(2)(1).$$

By convention (so that all will be right with the world and so that certain generalizations will go smoothly) we define $0!$ to be 1

From the trees above, we have seen that the number of ways to arrange 4 distinct numbers (or people, objects, or "things") is $4! = (4)(3)(2)(1) = 24$.

Note

In general,

$n!$ = the number of ways to arrange n distinct objects.

From our little tree drawing experiment, we have learned at least two other things. First, if the numbers can be "reused", the number of branches will not decrease. For example, if we roll a fair 6-sided die 4 times we could see outcomes like

$$3\ \ 5\ \ 3\ \ 4 \qquad \text{and} \qquad 2\ \ 1\ \ 1\ \ 1.$$

The total number of possible outcomes will be 6 (first branches) times 6 (second branches) times 6 (third branches) times 6 (fourth branches), for a total of $6^4 = 1296$ outcomes.

At this point we could make another fancy box saying something like: "If we perform an experiment with k trials and with n possible outcomes for each trial, we have a total of n^k possible outcomes for the entire experiment." However, **We won't and you shouldn't either**. If you are memorizing things for a math class (outside of definitions), the class will get more and more and more difficult as you pile up information in your head and eventually struggle to access it. If you are memorizing things for a math class, you are going to be thrown off when problems have only subtle differences from ones you've already seen or solved. On the other hand, if you **understand** and think about what you're doing at each step, the class will get easier and easier even as the material gets more complex!

0.2.2 Choosing Things: Order is Important

Here is the second thing we have learned from our "tree experiment". Suppose we have to choose 2 numbers out of our 4 numbers and that "**order is important**". This means that

$$1\ \ 4 \qquad \text{is different from} \qquad 4\ \ 1.$$

For the four people in the treehouse club, this scenario could come up if you want to choose only a president and a vice president. You could write two names down and the first in your list will be president, while the second

will be vice president. If, as before, the treehouse club's members names are 1, 2, 3, and 4, the choice

$$1 \quad 4$$

means that person 1 is president and person 4 is vice president, while the different choice

$$4 \quad 1$$

means that person 4 is president and person 1 is vice president.

Since you have 4 choices to put in the first position (envisioned as first branches of a tree) and then 3 leftover people to choose for the second position (second branches), you have a total of $4 \times 3 = 12$ possible ways to assign names to the president and vice president positions in your list.

That was easy enough, but it will be useful to be able to come up with this number using factorials. You might have noticed (or not) that

$$12 = 4 \cdot 3 = \frac{4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1} = \frac{4!}{2!},$$

but let's discuss the logic behind this. The point will be a little more clear if we increase the number of treehouse members to five. (The only reason we didn't start with five in the first place is because the tree diagram would have been really wide and too cramped for these pages!)

Suppose there are five members of the treehouse club and that their names are 1, 2, 3, 4, and 5. Further suppose that we want to know how many different ways we can choose a president and a vice president. We could list all of the members in various orders and choose the first two people in the list to be the president and vice president, respectively. For example, one possible listing of members is

$$4 \;\; 2 \;\; 5 \;\; 3 \;\; 1.$$

For this listing, 4 will be president and 2 will be vice president.

Another possible listing of members is

$$2 \;\; 4 \;\; 5 \;\; 3 \;\; 1.$$

In this case, 2 will be president and 4 will be vice president.

In all, there are $5! = 120$ possible listings of these 5 people, but for our purpose of choosing a president and vice president some are redundant. For example

$$4 \;\; 2 \;\; 5 \;\; 3 \;\; 1 \qquad \text{and} \qquad 4 \;\; 2 \;\; 1 \;\; 3 \;\; 5$$

will give the same results for our election. In fact, for this 5 person listing method, we will get the result of 4 as president and 2 as vice president $3! = 6$ different times since there are $3! = 6$ different ways to arrange the three superfluous numbers 1, 3, and 5. So, we need to divide the 120 possible ways to list 5 people by 6. In summary, the number of ways to choose 2 people (numbers/objects/things) from 5 people (numbers/objects/things) when **order is important** is

$$\frac{5!}{3!} = \frac{120}{6} = 20.$$

Note

In general, the number of ways to choose k objects out of n objects when order is important is $n!/(n-k)!$.

What we have just done is to count the number of "permutations" of k objects out of n objects. The quantity we came up with in the end is denoted in many different ways. One of the most popular is ${}_nP_k$. That is,

$$ {}_nP_k := n!/(n-k)!. $$

You will not need this notation in this course or to understand the rest of this text. If needed, we will figure out how to choose k things out of n things with logic, just as we did in the treehouse example!

0.2.3 Choosing Things: Order is Not Important

Suppose now that we want to choose two people from our five member treehouse club to fix the ladder. How many different possible "ladder fixing committees" are there? If we were listing out pairs of people, the **order is not important** because having 1 and 4 fix the ladder is the same as having 4 and 1 fix the ladder.

We will take the same approach as in the previous section, where we list out all five names and then take the first two names for our committee.

There are $5! = 120$ ways to list the 5 names. One possibility is

$$ 4\ \ 2\ \ 5\ \ 3\ \ 1. $$

This particular listing of names will give us a committee consisting of 4 and 2. As before, the order of the last 3 people is irrelevant and, using all possible listings, we will get 4 and 2 for our committee 3! different ways. So, we should divide 5! by 3! to account for this redundancy. At this point, we have $5!/3!$ lists we care about.

Note that, we still have some redundancy since the outcome

$$ 4\ \ 2\ \ 5\ \ 3\ \ 1 $$

is now the same as

$$ 2\ \ 4\ \ 5\ \ 3\ \ 1. $$

Both of these outcomes mean that our ladder committee consists of 4 and 2. In fact, every pair of people will end up being listed 2 ways. (Note that, for ease of generalization, we will use the factorial notation and say that the number of ways to list 2 people is $2! = 2$.)

So, for our $5!/3!$ surviving lists, we need to divide by 2!. In summary, the number of ways to choose 2 people (numbers/objects/things) from 5 people (numbers/objects/things) when **order is not important** is

$$ \frac{5!/3!}{2!} = \frac{5!}{2!\,3!}. $$

Note

In general, the number of ways to choose k objects out of n objects when order is not important is

$$\frac{n!}{k!\,(n-k)!}.$$

What we have just done is to count the number of "combinations" of k objects out of n objects. This quantity is almost always denoted by the symbol $\binom{n}{k}$. That is

$$\binom{n}{k} := \frac{n!}{k!\,(n-k)!}.$$

This symbol is read "**n choose k**" and we will use it a lot!

0.3 Sample Spaces, Events, and Some Simple Probabilities

In probability, we typically use capital Roman letters like A, B and C, for example, to denote "events" which are collections of outcomes from some experiment.

Suppose we perform the fabulously exciting experiment of flipping a fair coin 3 times. Using what we hope is an obvious notation for the sequences of "heads" and "tails" the eight possible outcomes are

$$HHH, HHT, HTH, THH, HTT, THT, TTH, TTT.$$

Definition 0.3.1

The set of all possible outcomes of experiment is known as the **sample space** of the experiment.

We will use the symbol Ω to denote the sample space of an experiment.

For this example we would write

$$\Omega = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}.$$

Let us define the event called "A" as the one where we see at least one "heads" in our experiment. We can write

$$A = \text{the event where we see at least one heads,}$$

or, we might use the set notation

$$A = \{HHH, HHT, HTH, THH, HTT, THT, TTH\}.$$

When we flip the coin 3 times, A "occurring" means that the result was **one** of the outcomes in this set.

What is the probability that the event A occurs? That is, if we are going to flip a fair coin 3 times, what is the probability that we will see at least 1 "heads"?

Notation

We will use $P(A)$ to denote the probability that the event A occurs.

When A is a collection of outcomes at it is here, A "occurring" means that when we performed our experiment it produced one of the outcomes in A.

Since 7 out of the 8 possible outcomes are included in the set A, this probability should be very high. This conclusion, though, is based on the idea that most outcomes are in A, but, it is also based on the idea that the coin is "fair". If the coin was severely bent or weighted in such a way that "tails" comes up 99.9 percent of the time then we should expect that it is somewhat unlikely that we will see at least 1 heads!

Because heads and tails are "equally likely" for a fair coin, the three toss outcomes, HHH, HHT, etc. are also equally likely. You might be thinking that outcomes like HHH or TTT are less likely than the others for a fair coin. To make this even more dramatic, imagine tossing the coin 20 times. Shouldn't the outcomes of 20 tails in a row be very unlikely? The fact of the matter is that it will not be any more unlikely than any other specific sequence. Just imagine flipping 20 times with the goal of getting $HTTHTHTHTHHTHTHTHTHT$. This would be hard to achieve! Similarly getting the elusive royal flush poker hand (ace, king, queen, jack and ten all in the same suit) is no more difficult than getting the hand that is exactly the 3 of clubs, 10 of diamonds, 7 of spades, 3 of hearts, and the jack of spades. (Actually, it is easier since there are 4 different possible royal flushes– one for each suit!)

Let's get back to the question though. If we flip a fair coin 3 times, what is $P(A)$, the probability that we will see at least 1 heads? Since the 8 outcomes in Ω are equally likely and since there are 7 outcomes where we see at least 1 heads, we have a "7 out of 8 chance" of seeing at least one heads. That is,

$$P(A) = 7/8.$$

It might also be convenient (though cumbersome) sometimes for us to write this out as

$$P(\{HHH, HHT, HTH, THH, HTT, THT, TTH\}) = 7/8.$$

Note

In general, if Ω is a set of equally likely outcomes of an experiment and A is a subset of Ω,

$$P(A) = \frac{\#\text{ outcomes in } A}{\#\text{ outcomes in } \Omega}.$$

This is why establishing some counting techniques in the previous sections was so important.

Example 0.3.1

If you randomly select 5 cards from a deck of 52 cards, what is the probability that you will get a royal flush? (ace, king, queen, jack, and 10 in the same suit)

Answer:

First of all, note that order is not important here. If you are playing poker, you might be excited to have 3 aces in your hand but you certainly wouldn't care that one of them was near your thumb and the other two were closer to your pinky. Therefore, as we learned in Section 0.2.3, the total number of 5 card hands is given by $\binom{52}{5}$.

Since every 5 card hand is equally likely and since there are exactly 4 hands that are a royal flush (one for each of the 4 suits), we have that

$$P(\text{ royal flush }) = 4 \Big/ \binom{52}{5}$$

Incidentally, after working out the factorials, this is $4/2,598,960 = 1.539077 \times 10^{-6}$! (← That is an exclamation point for excitement and not another factorial!) (← That is another exclamation point...)

0.4 Some Very Brief Words About Independence and "Disjointness"

Independence

In the previous section, we considered flipping a fair coin 3 times. The possible outcomes were

$$\Omega = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}.$$

Although it wasn't explicitly stated, the implication or assumption here is that the coin is flipped in a random and fair way and that the flips are independent. That is, what happens on the second flip, for example, does not depend on what happened on the first flip.

What is the probability of seeing the result HTH? Based on the previous section, and because the outcomes are equally likely, we know by counting that the answer is $1/8$. In this section, we will tackle the problem in a different way that does not require that we write out and count all the possible outcomes since this may not be a reasonable thing to do for a bigger problem.

Note that the "event" HTH means that we got heads on the first toss of the coin, tails on the second toss, and heads on the third toss. Furthermore, note that the coin doesn't care about what toss we are on at any point so

$$P(\text{ heads on any one toss }) = 1/2$$

and

$$P(\text{ tails on any one toss }) = 1/2.$$

Note that

$$P(HTH) = \frac{1}{8} = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2}.$$

Property

In general, if A and B are independent events

$$\underbrace{P(A \text{ and } B)}_{P(A \cap B)} = P(A) \cdot P(B).$$

Applying this (generalizing to 3 events) to our problem, since the flips are independent, we have that

$$\begin{aligned} P(HTH) &= P(\text{ heads on the first flip AND} \\ &\quad \text{tails on the second flip AND} \\ &\quad \text{heads on the third flip }) \\ &\overset{indep}{=} P(\text{ heads on the first flip}) \\ &\quad \cdot P(\text{tails on the second flip}) \\ &\quad \cdot P(\text{heads on the third flip }) \\ &= \tfrac{1}{2} \cdot \tfrac{1}{2} \cdot \tfrac{1}{2} = \tfrac{1}{8}. \end{aligned}$$

(The "*indep*" indicates that independence was used at that point to argue that equality.)

Disjoint Events

Disjoint events are events that can not happen together. For example, in our coin flipping experiment, the event "heads on the first flip" and the event "tails on the first flip" are disjoint. you can have one happen, or the other, but not both. This is different than the relationship between the events "heads on the first flip" and "tails on the second flip" since it is possible that we see both events occur. Some beginners to probability often confuse the ideas of independence and disjointness, but they are very different things. We can't begin to talk about events being independent if they can't even exist in the same universe! On the other hand, disjoint events are in a sense very dependent. If you know you got heads on the first flip, then you know for sure that you did not get tails on the first flip.

Suppose we are going to flip a fair coin 3 times and we want to know the probability that the first two flips are heads or the last two flips are tails. Given all equally likely outcomes

$$\Omega = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\},$$

it is easy to get the answer by counting outcomes of interest here. There are 4 outcomes (HHH, HHT, HTT, and TTT) where we have the desired result, so the desired probability is $4/8 = 1/2$. Note that the events "the first two flips are heads" and "the last two flips are tails" are disjoint. They can't happen at the same time if we are only flipping the coin three times because we would need that middle flip to be simultaneously heads and tails. In other words, the sets of outcomes $\{HHH, HHT\}$ and $\{HTT, TTT\}$ have no overlap. Counting the

total number of events in the union

$$\{HHH, HHT, HTT, TTT\} = \{HHH, HHT\} \cup \{HTT, TTT\}$$

is equivalent to counting the events in each of the two subsets and adding the two totals. Thus, from this counting perspective,

$$\begin{aligned} P(\{HHH, HHT, HTT, TTT\}) &= P(\{HHH, HHT\}) + P(\{HTT, TTT\}) \\ &= \tfrac{2}{8} + \tfrac{2}{8} = \tfrac{4}{8} = \tfrac{1}{2}. \end{aligned}$$

"Disjointness" was key. If the two subset events were not disjoint, we would have double counted some outcomes.

Property

In general, if A and B are disjoint events

$$\underbrace{P(A \text{ or } B)}_{P(A \cup B)} = P(A) + P(B).$$

This is true for disjoint events even when things are not equally likely and probabilities can not be determined by counting.

If the events are not disjoint, we must remove the doubly counted events as follows.

Property

In general, if A and B are events

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B),$$

or, in better notation

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

0.5 A Brief Review of Random Variables, PDFs, and CDFs

0.5.1 Random Variables, the Bernoulli Distribution, and a Squiggly Line

A **random variable** is a "mapping" from the set if possible outcomes of an experiment, involving some randomness or probability, to real numbers.

Example 0.5.1

Flip a, possibly unfair, coin that has some probability p of coming up "heads" and probability $1-p$ of coming up "tails", where p is some fixed number in the interval $[0,1]$.

Define

$$X = \begin{cases} 1 & , \text{ if "heads"} \\ 0 & , \text{ if "tails".} \end{cases}$$

Then X is a random variable that takes the value 1 with probability p and 0 with probability $1-p$.

We can think of it as a mapping from a set Ω into $\mathbb{R}$ (written $X : \Omega \to \mathbb{R}$) defined as

$$X(\omega) = \begin{cases} 1 & , \text{ if } \omega = H \\ 0 & , \text{ if } \omega = T \end{cases}$$

where $\Omega = \{H, T\}$ is the set of all possible outcomes for this simple experiment.

This (0/1) type of random variable comes up so often that it gets a name!

X is called a **Bernoulli random variable with parameter p**. Alternatively, we say that X **has a Bernoulli distribution with parameter p**.

Our shorthand notation for this will be to write

$$X \sim Bernoulli(p).$$

Sometimes you might instead see $X \sim Bern(p)$. The symbol "$\sim$" should be read as "has the distribution".

0.5.2 PDFs for Discrete Random Variables and the Geometric Distribution

There are two very important functions associated with random variables. One is called the

probability density function or "pdf"

and the other is called the

cumulative distribution function or "cdf".

They are usually denoted by f and F, for the pdf and cdf, respectively.

Most of the random variables we will talk about in this course will be either discrete (taking on possible values from a discrete set) or continuous (taking on possible values in a continuum). This section is about pdfs for discrete random variables. Please note that most authors use the terminology "probability density function"

and the acronym "pdf" only when talking about continuous random variables and, in the discrete case, refer to the function about to be defined here as a **probability mass function (pmf)**. In this text, we will not make this distinction. A "pdf" will simply mean different things depending on whether we are talking about discrete or continuous random variables.

First, we consider the discrete case.

Definition 0.5.1

The **probability density function (pdf)** for a discrete random variable X is the function f defined by

$$f(x) := P(X = x).$$

This is also known as a **probability mass function (pmf)**.

The right-hand side there is used to denote the probability that X takes on the specific value x. X is a random variable and x is a stand in for a specific real number that can be taken on by X.

Note

In probability and statistics, it is customary to use capital letters to denote random variables and lower case letters to denote specific fixed values.

Example 0.5.2

Suppose that $X \sim Bernoulli(p)$. What is the pdf of X?

Answer:

Since $P(X = 1) = p$ and $P(X = 0) = 1 - p$, by our definition of the pdf, we have that $f(1) = p$ and $f(0) = 1 - p$. Furthermore, $f(2.7)$, for example, is $f(2.7) = P(X = 2.7) = 0$, since X can only take on values in $\{0, 1\}$. In summary, the pdf is

$$f(x) = \begin{cases} 1-p & , \quad x = 0 \\ p & , \quad x = 1 \\ 0 & , \quad \text{otherwise.} \end{cases}$$

In mathematics, we usually call the region on which a function is defined the **domain** of the function. Noticed that the above pdf is defined everywhere (for all real numbers) but is only really "interesting" at the points $x = 0$ and $x = 1$. In general, the region where a function is non-zero is called the **support** of the function. If you use these terms interchangeably, so be it– it is not the end of the world, but really they have different meanings.

Let's now look at another discrete random variable with a distribution thatis used so frequently that it also gets a name.

Example 0.5.3

Consider an experiment consisting of a sequence of independent trials of something where each trial can have only two possible outcomes, usually labeled as "success" and "failure". We will denote success and failure by S and F, respectively.

Let p be the probability of "success" on any one trial. Here, p must live in the interval $[0, 1]$.

Let

$$X = \#\text{ trials until the first success.}$$

Then the possible values for X are $1, 2, 3, \ldots$.

For example, in an idealized world, you might imagine someone shooting baskets in basketball, and X as the number of tries they have to take until they make a basket. (In real life though, the assumptions that the success probability is constant for all trials and the trials are independent would probably be violated as the person could be learning from each failed attempt and making adjustments to get better or maybe their arms are tired and they are getting worse!)

Question:

What is the pdf for X?

Answer:

Note that

$$P(X = 1) = P(\,S \text{ on first trial}\,) = p.$$

Continuing,

$$P(X = 2) = P(\,F \text{ on 1st trial AND } S \text{ on 2nd trial}\,).$$

By independence of the trials, this is

$$P(X = 2) \overset{indep}{=} P(\,F \text{ on 1st trial}\,) \cdot P(\,S \text{ on second trial}\,)$$

$$= (1 - p) \cdot p.$$

Similarly,

$$\begin{aligned} P(X=3) &= P(\text{ F on 1st trial AND F on 2nd trial AND S on 3rd trial }) \\ &\overset{indep}{=} P(\text{ F on 1st trial }) \cdot P(\text{ F on 2nd trial }) \cdot P(\text{ S on 3rd trial }) \\ &= (1-p)\cdot(1-p)\cdot p = (1-p)^2 \cdot p. \end{aligned}$$

A pattern is forming! Indeed, we have

$$P(X=x) = (1-p)^{x-1} \cdot p$$

for $x = 1, 2, 3, \ldots$.

Since $P(X=x) = 0$ for x not in $\{1, 2, 3, \ldots\}$, we conclude that the pdf is

$$f(x) = \begin{cases} (1-p)^{x-1} \cdot p & , \quad x = 1, 2, 3 \ldots \\ 0 & , \quad \text{otherwise.} \end{cases}$$

Again we have happened across a type of random variable that is so common it gets a name. X is said to be a **geometric random variable with parameter p**. Alternatively, we say that X has a **geometric distribution**. We write

$$X \sim geom(p).$$

Note that, using the same set up with successes and failures on independent trials, sometimes the geometric random variable is defined as

$$X = \#\text{ failures before the first success.}$$

Now, if we see a success on the first trial, X will take the value 0. It is easy to adjust what we did above to show that the pdf is now

$$f(x) = \begin{cases} (1-p)^{x} \cdot p & , \quad x = 0, 1, 2, \ldots \\ 0 & , \quad \text{otherwise.} \end{cases} \tag{0.5.1}$$

Although it is not standard notation out there in the world, in this text we will write

$$X \sim geom_0(p)$$

to denote the random variable having the geometric distribution that "starts from 0" and

$$X \sim geom_1(p)$$

to denote the random variable having the geometric distribution that "starts from 1".

0.5.3 PDFs for Continuous Variables and the Exponential Distribution

Let X be a continuous random variable. Then X is assumed to take on one in a whole continuum of values. For example, suppose that we randomly select a student on campus and measure their height in inches. Let X be this height and assume we can take this measurement with infinite accuracy to as many decimal places as we want. It seems reasonable that we could find someone on campus who is between, say 64 inches and 70 inches,

and so talking about the probability $P(64 < X < 70)$ makes sense. Because the range of values in the interval $[64, 66]$ is smaller but included in $[64, 70]$, we would expect that

$$P(64 < X < 66) \leq P(64 < X < 70).$$

(Indeed, in "event notation", if A and B are events such that $A \subseteq B$, we will always have $P(A) \leq P(B)$.)

Back to heights though... As these intervals get smaller, the events become even more rare. It would in fact be very difficult to randomly select a student whose height is between 64 and 64.000001 inches. The likelihood of finding someone to be exactly $64.0\overline{0}$ is vanishingly small!

In fact, we always have the following.

Note

If X is any continuous random variable and x represents a specific number, we will always have that $P(X = x) = 0$.

In this Section, we want to define what is meant by a pdf for a continuous random variable. It can no longer be defined as $f(x) = P(X = x)$, as it was in the discrete case because this is always zero and therefore would be terribly uninteresting! Instead, we define the pdf for a continuous random variable as follows.

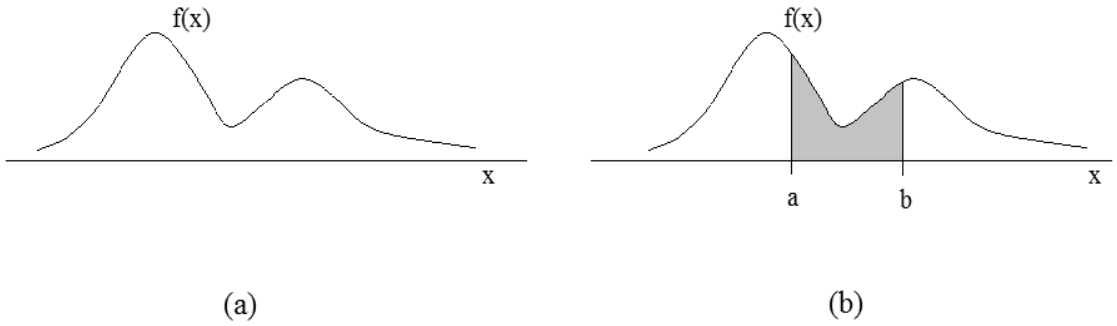
Definition 0.5.2

The **probability density function (pdf)** for a continuous random variable X is a non-negative function f under which area represents probability.

For example, if the continuous random variable X has the pdf depicted in Figure 1(a), then the probability $P(a < X \leq b)$ is the shaded region depicted in Figure 1(b). This is computed as the integral

$$P(a < X \leq b) = \int_a^b f(x)\,dx.$$

Figure 1: A pdf for a Continuous Random Variable



(a) (b)

Since the area of the vertical lines at the boundaries of the region have area zero, we can put them in or take them out without changing the area of the shaded region. Thus,

$$P(a < X \leq b) = P(a \leq X \leq b) = P(a \leq X < b) = P(a < X < b).$$

Furthermore, for any a,

$$P(X = a) = \int_a^a f(x)\,dx = 0.$$

The pdf describes how the probabilities associated with a random variable are "distributed" over the real line. Where is most of the area under the curve "piled up"? That is, where you are most likely to observe the random variable if you record an actual "realization" (numerical observation) of it in the future? Thus, we will often talk about the pdf associated with a "distribution" instead of talking about the random variable itself. For example, equation (0.5.1) is the pdf for the geometric distribution.

Property

For a function f to be a valid pdf, we must have

$$f(x) \geq 0 \text{ for all } x.$$

For a discrete distribution, we must have

$$\sum_x f(x) = 1,$$

and for a continuous distribution we must have

$$\int_{-\infty}^{\infty} f(x)\,dx = 1.$$

Example 0.5.4

Suppose that you are standing near the door of a grocery store watching customers arrive. This makes you kind of creepy but, "whatever".

Suppose further that

- the arrival rate is a constant 15.2 people per minute, and
- the number of arrivals in non-overlapping periods of time are independent.

(Yes, it probably isn't realistic to have the arrival rate be constant all day. Consider it an overly simplistic model but a model nonetheless!)

Let

$$X = \text{the time (in minutes) between any two consecutive arrivals.}$$

Then we can show (in a different course in the subject of Markov Processes) that this continuous random

variable X has the pdf

$$f(x) = \begin{cases} 15.2e^{-15.2x} & , \quad x \geq 0 \\ 0 & , \quad x < 0. \end{cases}$$

X is said to be an **exponential random variable with rate 15.2** or to have an **exponential distribution with rate 15.2**.

We will write

$$X \sim exp(rate = 15.2).$$

Note that, if people are arriving at a rate of 15.2 per minute, then, on average, the **time** between arrivals is $1/15.2$ minutes. In general, if people are arriving at a rate of λ per minute, then the mean "interarrival time" is $1/\lambda$ minutes.

We have

$$X \sim exp(rate = \lambda) \qquad \Rightarrow \qquad f(x) = \begin{cases} \lambda e^{-\lambda x} & , \quad x \geq 0 \\ 0 & , \quad x < 0. \end{cases}$$

However, many people specify this model in terms of the mean interarrival time and call that mean λ. This would correspond to a rate of $1/\lambda$ and so we write

$$X \sim exp(mean = \lambda) \qquad \Rightarrow \qquad f(x) = \begin{cases} \frac{1}{\lambda} e^{-x/\lambda} & , \quad x \geq 0 \\ 0 & , \quad x < 0. \end{cases}$$

Unfortunately, many texts use the notation $X \sim exp(\lambda)$ and it is up to you to look through earlier parts of whatever you might be reading in order to figure out which of these two pdfs they are talking about. In this text, we will almost always be thinking of it in terms of the rate parameter and we will always explicitly write "rate = λ".

0.5.4 CDFs

The acronym **cdf** stands for **cumulative distribution function**.

Definition 0.5.3

The **cumulative distribution function (cdf)** of a random variable X is usually denoted by $F(x)$ and is defined as

$$F(x) := P(X \leq x).$$

This definition holds for both discrete and continuous random variables. Note that we must always have that

$$\lim_{x \to -\infty} F(x) = 0 \qquad \text{and} \qquad \lim_{n \to \infty} F(x) = 1$$

since X is always less than or equal to ∞ and never less than or equal to $-\infty$.

Example 0.5.5 (Discrete)

Suppose that $X \sim geom_0(p)$. Note that the "event" that $\{X \leq x\}$ for a discrete random variable is made up of the disjoint events of the form $\{X = u\}$ where u is any number in the support of the random variable that is less than or equal to x. (We are saying it in this weird way because, in general, discrete random variables are not necessarily integer valued!)

Since this geometric distribution takes on values in $\{0, 1, 2, \ldots\}$, we have, for any integer x,

$$F(x) \overset{def}{=} P(X \leq x) \overset{disj}{=} \sum_{u \leq x} P(X = u) \overset{geom}{=} \sum_{u=0}^{x} P(X = u)$$

$$\overset{geom}{=} \sum_{u=0}^{x} (1-p)^u p = p \sum_{u=0}^{x} (1-p)^u$$

This is a finite geometric sum. Recall first, from Calculus, the infinite geometric sum

$$\sum_{n=0}^{\infty} r^n = \frac{1}{1-r} \qquad \text{if } |r| < 1.$$

The finite geometric sum, $\sum_{n=0}^{N} r^n$ is similar in the denominator and has the form

$$\sum_{n=0}^{N} r^n = \frac{?}{1-r}.$$

Writing the left side out as $1 + r + r^2 + \cdots + r^N$, we see that

$$\begin{aligned} ? &= (1-r)(1 + r + r^2 + \cdots + r^N) \\ &= (1 + r + r^2 + \cdots + r^N) - r(1 + r + r^2 + \cdots + r^N) \\ &= (1 + r + r^2 + \cdots + r^N) - (r + r^2 + r^3 + \cdots + r^{N+1}) \\ &= 1 - r^{N+1}. \end{aligned}$$

Thus,

$$\sum_{n=0}^{N} r^n = \frac{1 - r^{N+1}}{1-r}.$$

Going back to the cdf for the geometric distribution, we have

$$F(x) = p \sum_{u=0}^{x} (1-p)^u$$

which is a finite geometric sum with $r = 1 - p$ and $N = x$. Therefore,

$$F(x) = p \sum_{u=0}^{x} (1-p)^u = p \frac{1 - (1-p)^{x+1}}{1 - (1-p)} = 1 - (1-p)^{x+1}.$$

Here, we assumed that x was an integer. Because the distribution is integer valued, we have, for example

that

$$F(5.6) := P(X \leq 5.6) = P(X \leq 5) = F(5) = 1 - (1-p)^6.$$

Because X only takes on non-negative integer values, we have that

$$F(0) = P(X \leq 0) = 1 - (1-p) = p = P(X = 0)$$

and that, for $x < 0$,

$$F(x) = P(X \leq x) = 0.$$

In summary, if $X \sim geom_0(p)$, the cdf is

$$F(x) = \begin{cases} 1 - (1-p)^{\lfloor x \rfloor + 1} & , \quad x \geq 0 \\ 0 & , \quad x < 0 \end{cases}$$

where $\lfloor x \rfloor$ is the greatest integer function which returns the largest integer that is less than or equal to x.

Alternatively, we can of course say that

$$F(x) = \begin{cases} 1 - (1-p)^{x+1} & , \quad x = 0, 1, 2, \ldots \\ 0 & , \quad \text{otherwise} \end{cases}$$

Example 0.5.6 (Continuous)

Suppose that $X \sim exp(rate = \lambda)$. For any $x \geq 0$,

$$\begin{aligned} F(x) &= P(X \leq x) = \textstyle\int_{-\infty}^{x} f(u)\, du \\ &= \textstyle\int_{0}^{x} \lambda e^{-\lambda u}\, du = 1 - e^{-\lambda x}. \end{aligned}$$

Notice that the limits of integration changed because the pdf is 0 for $x < 0$. Technically, though it was unwritten, we have

$$\int_{-\infty}^{x} f(u)\, du = \int_{-\infty}^{0} 0\, du + \int_{0}^{x} \lambda e^{-\lambda u}\, du$$

Since the exponential distribution only takes on positive values, we have, for $x < 0$, that $F(x) = P(X \leq x) = 0$. In summary, the cdf for the exponential distribution with rate λ is

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & , \quad x \geq 0 \\ 0 & , \quad x < 0. \end{cases}$$

We have seen two examples of how to get from a distribution pdf to its cdf. Suppose we want to reverse the process. That is, given a cdf, can we compute the associated pdf?

Discrete Case:

Rather than write out a formula, it is easiest to consider the cdf for a discrete random variable by example and on a case-by-case basis. Suppose, for example, that we are dealing with a distribution that takes on only integer values. Then we can define the pdf at, say 3 by

$$f(3) = P(X = 3) = P(X \leq 3) - P(X \leq 2) = F(3) - F(2).$$

Continuing this example, for general integer values of x, we can define

$$f(x) = P(X = x) = P(X \leq x) - P(X \leq x - 1) = F(x) - F(x - 1).$$

We would also define $f(x) = 0$ whenever x is not an integer or if x is an integer that is outside the support of the distribution.

In theory, it is easy to generalize this method of recovering the pdf from a cdf for discrete random variables that do not live on integers but the notation is kind of messy. We'll just deal with it as it arises!

Continuous Case

It is much easier to write down a succinct formula for moving from a cdf to a pdf in the continuous case. Because the cdf is defined as

$$F(x) = \int_{-\infty}^{x} f(u)\, du,$$

the Fundamental Theorem of Calculus gives us the following.

Property

If X is a continuous random variable with cdf $F(x)$, the pdf is

$$f(x) = \frac{d}{dx}F(X).$$

For example, for the exponential distribution with rate λ, the cdf is $F(x) = 1 - e^{-\lambda x}$ for $x \geq 0$. The pdf is therefore

$$f(x) = \frac{d}{dx}F(X) = \frac{d}{dx}(1 - e^{-\lambda x}) = \lambda e^{-\lambda x}$$

for $x \geq 0$. It is zero elsewhere.

Rumination

We have defined a "pdf" differently for discrete and continuous random variables. This course will be "pdf-centric", but if you go on to take an advanced probability course, it will likely be "cdf-centric". That is, the cdf will be defined before the pdf, as

$$F(x) := P(X \leq x)$$

and the pdf will be defined as the derivative

$$f(x) := \frac{d}{dx}F(X),$$

if it exists. For a discrete random variable, the cdf will be a step function that is not differentiable at many points, so someone using these definitions would say that the pdf **does not exist**. When defining things in this cdf-centric way, one would never use the term "pdf" for discrete random variables.

0.6 The Expected Value, Expectation, or Mean of a Random Variable or a Distribution

We will denote the **expected value** (also called the expectation or mean) of a random variable X as $\mathsf{E}[X]$.

Definition 0.6.1

The **expected value** of X is a **probability weighted average**.

Example 0.6.1

Suppose that X is discrete, taking values in $\{0, 1, 2\}$ with the following probabilities.

x	0	1	2
$P(X = x)$	0.2	0.7	0.1

If you were going to get to observe a single realization of X, what do you expect to see?

Since the value 1 comes up with the highest probability, you might "expect" to see a 1. However, expectation

is a probability weighted average and is not limited to the actual values that X can take on. On average, you expect to see a 1 but you also might see a 0 or a 2. Since you have a higher probability of seeing a 0 versus a 2, the expected value is being pulled more towards 0 than 2.

The expected value of X, which we will define more formally in a moment, is

$$\mathsf{E}[X] = (0)(0.2) + (1)(0.7) + (2)(0.1) = 0.9.$$

Definition 0.6.2

In general, for a discrete random variable X with pdf f, the **expected value** is

$$\mathsf{E}[X] = \sum_x x \cdot P(X = x) = \sum_x x \cdot f(x).$$

The sum is taken over all real numbers though we can also just sum over the support of the distribution which you'll recall includes all x's for which $f(x) = P(X = x) \neq 0$. In other words, if you take the sum over all real numbers, you'd just get a whole lot of zeros!

Example 0.6.2

Let $X \sim geom_0(p)$. Find $\mathsf{E}[X]$.

Answer:

Since

$$f(x) = \begin{cases} (1-p)^x \cdot p & , \quad x = 0, 1, 2 \ldots \\ 0 & , \quad \text{otherwise,} \end{cases}$$

we have that

$$\mathsf{E}[X] \quad = \quad \sum_x x f(x) = \sum_{x=0}^{\infty} x \cdot (1-p)^x \cdot p.$$

Continuing,

$$\mathsf{E}[X] \quad = \quad \sum_{x=0}^{\infty} x \cdot (1-p)^x \cdot p = p \sum_{x=0}^{\infty} x(1-p)^x = p \sum_{x=1}^{\infty} x(1-p)^x \tag{0.6.2}$$

since the sum is zero when $x = 0$. Hmmmm... an infinite sum. The only one of those most of us can remember is the geometric sum

$$\sum_{n=0}^{\infty} r^n = \frac{1}{1-r} \qquad \text{if } |r| < 1$$

which can easily be adjusted to start from different places. For example

$$\sum_{n=3}^{\infty} r^n = r^3 \sum_{n=3}^{\infty} r^{n-3} = r^3 \sum_{n=0}^{\infty} r^n = r^3 \cdot \frac{1}{1-r}.$$

The sum in (0.6.2) sort of looks like a geometric sum with $r = 1 - p$. The only problem is that leading x. Check this out:

$$\mathsf{E}[X] = p \sum_{x=1}^{\infty} x(1-p)^x = p(1-p) \sum_{x=1}^{\infty} x(1-p)^{x-1}.$$

Now the summand kind of looks like a derivative. Defining $q = 1 - p$, we may rewrite it again as

$$\begin{aligned} \mathsf{E}[X] &= p(1-p) \sum_{x=1}^{\infty} x\, q^{x-1} \\ &= p(1-p) \sum_{x=1}^{\infty} \tfrac{d}{dq} q^x \\ &= p(1-p) \tfrac{d}{dq} \sum_{x=1}^{\infty} q^x \end{aligned}$$

which is a nice geometric sum. Thus

$$\begin{aligned} \mathsf{E}[X] &= p(1-p) \tfrac{d}{dq} \sum_{x=1}^{\infty} q^x \\ &= p(1-p) \tfrac{d}{dq} \tfrac{q}{1-q} \\ &= p(1-p) \tfrac{(1-q)(1)-q(-1)}{(1-q)^2} \\ &= p(1-p) \tfrac{1}{(1-q)^2} = p(1-p) \tfrac{1}{p^2} = \tfrac{1-p}{p} \end{aligned}$$

In summary, if $X \sim geom_0(p)$,

$$\mathsf{E}[X] = \frac{1-p}{p}.$$

For a continuous random variable, the expected value is defined analogously with an integral in place of the sum.

Definition 0.6.3

For a continuous random variable X with pdf f, the **expected value** is

$$\mathsf{E}[X] = \int_{-\infty}^{\infty} x \cdot f(x)\, dx.$$

Example 0.6.3

Suppose that $X \sim exp(rate = \lambda)$. Find $\mathsf{E}[X]$.

Answer:

Since the exponential distribution is a continuous distribution we must use the integral definition for this expected value.

$$\begin{aligned} \mathsf{E}[X] &= \int_{-\infty}^{\infty} x \cdot f(x)\, dx \\ &= \int_{-\infty}^{0} x \cdot 0\, dx + \int_{0}^{\infty} x \cdot \lambda e^{-\lambda x}\, dx \\ &= \int_{0}^{\infty} x \cdot \lambda e^{-\lambda x}\, dx \end{aligned}$$

Using integration by parts ($\int u\, dv = uv - \int v\, du$) with $u = x$ and $dv = \lambda e^{-\lambda x}$, we have

$$\mathsf{E}[X] = -xe^{-\lambda x}\Big|_0^{\infty} - \int_0^{\infty} (-e^{-\lambda x})\, dx.$$

That first term certainly evaluates to 0 when we plug in $x = 0$. It is also 0 at ∞ (technically a limit as $x \to \infty$). This is because $e^{-\lambda x}$ goes down to zero (it is a requirement of the exponential distribution that $\lambda > 0$) faster than x blows up. Formally, this can be shown using L'Hôpital's Rule.

We are left with

$$\mathsf{E}[X] = \int_0^{\infty} e^{\lambda x}\, dx = -\frac{1}{\lambda} e^{-\lambda x}\Big|_0^{\infty} = -\frac{1}{\lambda}(0 - 1) = \frac{1}{\lambda}.$$

Remember, the expected value of a random variable (or, equivalently, of a distribution) is also known as its expectation or mean. The $1/\lambda$ that we found here is consistent with our discussion of rates and means at the end of Section 0.5.3.

Notation

The mean or expected value of a random variable X is often denoted with the Greek letter μ. that is,

$$\mu := \mathsf{E}[X].$$

In the case of multiple random variables, say X and Y, a subscript becomes necessary to distinguish the means,

$$\mu_{\mathrm{X}} = \mathsf{E}[X] \qquad \text{and} \qquad \mu_{\mathrm{Y}} = \mathsf{E}[Y].$$

While people would surely understand what you mean if you instead use lowercase subscripts as in μ_x and μ_y, this represents a really fundamental lack of understanding about what you are doing since lowercase Roman letters in statistics almost always refer to specific numbers and not random variables which can be thought of as "potential but yet unobserved numbers". For the record, if c is a constant, then $\mathsf{E}[c] = c$. You can think of c as an uninteresting random variable X where $X = c$ with probability 1. Then $\mathsf{E}[c] = \mathsf{E}[X] = c \cdot P(X = c) = c \cdot 1 = c$. For example,

$$\mathsf{E}[3] = 3$$

and, adhering to the convention of random variables being uppercase,

$$\mathsf{E}[x] = x.$$

While we're at it here, a mean μ is also a constant. When you compute any expectation you have already sort of "summed out" the probability. Thus, while we may have defined $\mu = \mathsf{E}[X]$ for some random variable X, $\mathsf{E}[\mu]$ will just be μ again!

0.7 The Variance of a Random Variable or a Distribution

The **variance** of a random variable X is a measure of spread about its mean. We will denote the **variance** of a random variable X as $Var[X]$.

Definition 0.7.1

If we use μ to denote the mean $\mathsf{E}[X]$, then the variance of X is defined as

$$Var[X] := \mathsf{E}[(X - \mu)^2].$$

Notation

A variance is usually also denoted by the symbol σ^2 with subscripting possible if it becomes necessary to distinguish between the variances for different random variables.

If you were devise your own measure of spread for a random variable X about its mean μ, you might use the expected distance

$$\mathsf{E}[|X - \mu|].$$

This is a fine measure of spread but it is not often used. Defining the variance as we did (with the square) will be much more mathematically convenient even if it might not seem that way right now.

You may already be familiar with the expression

$$\sigma^2 := Var[X] = \mathsf{E}[X^2] - (\mathsf{E}[X])^2.$$

We will see in Section 0.10 that the expectation "operator" is in fact a linear operator so that, for random variables X and Y and constants a and b, we have

$$\mathsf{E}[aX + bY] = a\mathsf{E}[X] + b\mathsf{E}[Y].$$

In particular,

$$\mathsf{E}[X + Y] = \mathsf{E}[X] + \mathsf{E}[Y]$$

and, taking $b = -1$,

$$\mathsf{E}[X - Y] = \mathsf{E}[X] - \mathsf{E}[Y].$$

With this in mind, we have that

$$\begin{aligned} Var[X] &= \mathsf{E}[(X-\mu)^2] \\ &= \mathsf{E}[X^2 - 2\mu X + \mu^2] \\ &= \mathsf{E}[X^2] - 2\mu \underbrace{\mathsf{E}[X]}_{\mu} + \mu^2 \qquad (\mu \text{ is a constant so } \mathsf{E}[\mu^2] = \mu^2) \\ &= \mathsf{E}[X^2] - 2\mu^2 + \mu^2 \\ &= \mathsf{E}[X^2] - \mu^2 \\ &= \mathsf{E}[X^2] - (\mathsf{E}[X])^2. \end{aligned}$$

Alternate Way to Compute a Variance

$$Var[X] = \mathsf{E}[X^2] - (\mathsf{E}[X])^2$$

Sometimes it will be convenient to give a measure of spread of random variable (or equivalently, its distribution) in the same units as the original random variable. For example, if X represents the height in inches of a randomly selected student on campus, $Var[X]$ will be in squared inches. Taking the square root will bring us back to the original units and will sometimes be desirable.

Definition 0.7.2

The **standard deviation** of a random variable X with variance σ^2 is denoted by σ and is defined as

$$\sigma = \sqrt{\sigma^2} = \sqrt{Var[X]}$$

Note that

$$\sqrt{Var[X]} = \sqrt{\mathsf{E}[(X-\mu)^2]} \neq \mathsf{E}[\sqrt{(X-\mu)^2}] = \mathsf{E}[X-\mu]$$

because an expectation is a sum or integral and sums and integrals just don't work that way with square roots! For example, $\sqrt{2+3} \neq \sqrt{2} + \sqrt{3}$.

We will see what the actual relationship between $\sqrt{\mathsf{E}[(X-\mu)^2]}$ and $\mathsf{E}[\sqrt{(X-\mu)^2}]$ in the exercises for Chapter 2.

0.8 Indicator Notation and a Most Useful Appendix

We have three "named" distributions so far: the Bernoulli, the geometric, and the exponential distributions. The most common named distributions are summarized in Tables A.1 and A.2 of Appendix A. We will eventually talk about all of the columns in these tables. As of right now, the first five columns should make sense. They include names of distributions, corresponding pdfs, a description of the allowable parameter space (For example, p for the Bernoulli and geometric distributions is itself a probability and can range from 0 to 1.), the mean (expected value), and the variance for each distribution. Each pdf in the tables includes an extra part involving an "I". This is an indicator function which is defined as follows.

Definition 0.8.1

Let A be a set. The function

$$I_A(x) = \begin{cases} 1 & , \quad \text{if } x \in A \\ 0 & , \quad \text{if } x \notin A \end{cases}$$

is called an **indicator function**.

Other common notations for this indicator function are $\mathbb{1}_A(x)$, $\chi_A(x)$, and the "Iverson bracket" $[x \in A]$.

With the indicator function, we no longer need to write our pdfs piecewise or to say "and zero otherwise". For example, if $X \sim geom_0(p)$, the pdf can now be written as

$$f(x) = (1-p)^x \cdot p \cdot I_{\{0,1,2,\ldots\}}(x).$$

So, for example

$$f(2) = (1-p)^2 \cdot p \cdot I_{\{0,1,2,\ldots\}}(2) = (1-p)^2 \cdot p \cdot 1 = (1-p)^2 \cdot p$$

while

$$f(2.5) = (1-p)^{2.5} \cdot p \cdot I_{\{0,1,2,\ldots\}}(2.5) = (1-p)^{2.5} \cdot p \cdot 0 = 0.$$

In mathematical statistics, indicator functions can be very helpful! We hope to convince you of this shortly. After all, writing down pdfs in the piecewise way, where we say "and zero otherwise" is not very hard to do. Furthermore, we could easily get away with just writing down the "important piece" of the piecewise defined pdf and just leave it implied that it is "zero otherwise". Indeed, we have other reasons for using indicator functions– the first of which we will see shortly. In the meantime, let's just start getting used to them!

As another example, we can write the pdf for the exponential distribution with rate λ as

$$f(x) = \lambda e^{-\lambda x} I_{(0,\infty)}(x).$$

Rumination

You'll come to learn that we are very big on this indicator notation and then you'll notice that we do not use it when writing down cdfs. There are two reasons for this. As for the first and main reason, there will be several times in this course (the first being at the end of Section 0.9.3) where being able to factor pdfs in certain ways will tell us important things about the associated distribution only if we include the indicator in the pdf and hence in the factorization. We won't be talking about any factorization theorems for cdfs. Perhaps the more important reason though is that including indicators in cdfs can make them more, rather than less, complicated because cdfs are not necessarily "and zero otherwise".

Consider the continuous "uniform" distribution on the interval $(0, 1)$ given by the pdf

$$f(x) = 1 \cdot I_{(0,1)}(x) = I_{(0,1)}(x).$$

We write $X \sim unif(0, 1)$.

For $0 < x < 1$,

$$F(x) = P(X \leq z) = \int_0^x 1 \, du = x.$$

Since X takes on values between 0 and 1, the probability that X is less than or equal to, for example, -2.3 is

$$F(-2.3) = P(X \leq -2.3) = 0,$$

whereas the probability that X is less than or equal to say 3 is

$$F(3) = P(X \leq 3) = 1.$$

So, we don't want to "zero out" the cdf everywhere off of the interval $(0, 1)$. We could still use indicators but it is not as simplifying. For this distribution we have

$$F(x) = x \, I_{(0,1)}(x) + I_{(1,\infty)}(x).$$

0.9 Joint PDFs, Marginals, and Independence

0.9.1 The Discrete Case

Definition 0.9.1

For discrete random variables X and Y, we define the **joint pdf** as

$$f(x,y) := P(X = x \text{ and } Y = y) \overset{notation}{=} P(X = x, Y = y).$$

Note that the x and y are just place holding dummy variables. For example,

$$P(X = 1, Y = 3) = f(1,3)$$

and

$$P(X = w, Y = z) = f(w,z).$$

So, if it is necessary to be more specific about which random variables the joint pdf belongs to, we might write it as $f_{X,Y}(x,y)$. Then, it is clear that

$$f_{X,Y}(w,z) = P(X = w, Y = z).$$

Joint pdfs for discrete random variables are often given in a tabular form.

Example 0.9.1

Suppose that X takes on values in $\{-1, 0, 1\}$ and Y takes on values in $\{-5, 3\}$, the joint probabilities might be written as

		x		
		-1	0	1
y	-5	0.1	0.3	0.2
	3	0.2	0.1	0.1

Reading from the table, we have, for example, that

$$f(0,-5) = P(X = 0, Y = -5) = 0.3.$$

Question: What is $P(X = 0)$?

Answer:

$$P(X = 0) = P(X = 0, Y = -5 \text{ or } X = 0, Y = 3).$$

Since the events $\{X = 0, Y = -5\}$ and $\{X = 0, Y = 3\}$ are disjoint events, we get

$$P(X = 0) = P(X = 0, Y = -5) + P(X = 0, Y = 3) = 0.3 + 0.1 = 0.4$$

Such probabilities for X or Y alone are often tallied up in the margins of the table as

		x			
		-1	0	1	
y	-5	0.1	0.3	0.2	0.6
	3	0.2	0.1	0.1	0.4
		0.3	0.4	0.3	

For this reason, the individual X and Y pdfs are often referred to as the **marginal pdfs** for X and Y! Denote them as $f_X(x)$ and $f_Y(y)$, respectively. They may be written alone as

	x		
	-1	0	1
$\underbrace{P(X = x)}_{f_X(x)}$	0.3	0.4	0.3

and

	y	
	-5	3
$\underbrace{P(Y = y)}_{f_Y(y)}$	0.6	0.4

0.9.2 The Continuous Case

Definition 0.9.2

For continuous random variables X and Y, the **joint pdf** is a function $f(x, y)$ that defines a surface under which volume represents probability.

Analogous to the discrete case where the marginal pdfs for X and Y were gotten by summing out the values of Y and X, respectively, we have the following in the continuous case.

Definition 0.9.3

If continuous random variables X and Y have joint pdf $f(x, y)$, the **marginal pdfs** for X and Y, denoted by $f_X(x)$ and $f_Y(y)$, respectively are

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$$

and

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx.$$

Joint and marginal pdfs are easily generalizable to more than two random variables.

0.9.3 Independence

Mirroring the independence discussed in Section 0.4, random variables X and Y are independent if their joint pdf separates out nicely into an "x-part" and a "y-part". Specifically, it needs to separate out into the product of the marginal pdfs.

Definition 0.9.4

Random variables X and Y are said to be **independent** if

$$f(x,y) = f_X(x) \cdot f_Y(y).$$

Example 0.9.2

Suppose that X and Y are continuous random variables with joint pdf

$$f(x,y) = \begin{cases} xy & , \quad 0 < x < 1,\ 0 < y < 2 \\ 0 & , \quad \text{otherwise} \end{cases}$$

Alternatively, we could write this with indicators as

$$f(x,y) = xy\, I_{(0,1)}(x) I_{(0,2)}(y).$$

Since this is a pdf, the total volume under this surface should be 1:

$$\begin{aligned} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy &= \int_0^2\int_0^1 xy\,dx\,dy \\ &= \int_0^2 y \underbrace{\int_0^1 x\,dx}_{1/2}\,dy \\ &= \tfrac{1}{2}\int_0^2 y\,dy = \tfrac{1}{2}\cdot 2 = 1 \ \checkmark \end{aligned}$$

Finding probabilities here comes down to finding limits of integration. For example:

- $P(X < 3/4, Y > 1/2) = \int_0^{3/4}\int_{1/2}^2 f(x,y)\,dy\,dx = \cdots$
- $P(1/2 \le X \le 3, 1/4 \le Y < 1) = \int_{1/2}^3\int_{1/4}^1 f(x,y)\,dy\,dx$
  $= \int_{1/2}^1\int_{1/4}^1 xy\,dy\,dx = \cdots$
  (Note the limit of integration change to the "relevant part" of the x interval. The pdf is 0 for $x = 1$ to $x = 3$.)
- $P(X \le Y) = \int_0^1\int_x^2 xy\,dy\,dx = \int_0^1\int_0^y xy\,dx\,dy + \int_1^2\int_0^1 xy\,dx\,dy = \cdots$

(If you are having trouble with these limits of integration, draw the rectangle where $0 \le x \le 1$ and $0 \le y \le 2$ and then shade in the subregion where $x \le y$.

The marginal pdf for X is

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy = \int_{-\infty}^{0} 0\,dy + \int_{0}^{2} xy\,dy + \int_{2}^{\infty} 0\,dy$$

You obviously don't need to write out those integrals of 0. We are just making the point that you should, in general, integrate over all values of y from $-\infty$ to ∞. So,

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy = \int_{0}^{2} xy\,dy = 2x.$$

We are not finished until we write the support (the x values where it is non-zero) of $f_X(x)$:

$$f_X(x) = 2x, \;\; 0 < x < 1.$$

The pdf is zero otherwise. We could use an indicator to write

$$f_X(x) = 2x\, I_{(0,1)}(x),$$

or we could have taken care of this automatically using indicators from the beginning:

$$\begin{aligned} f_X(x) &= \textstyle\int_{-\infty}^{\infty} f(x,y)\,dy = \int_{-\infty}^{\infty} xy\, I_{(0,1)}(x) I_{(0,2)}(y)\,dy \\ &= x\, I_{(0,1)}(x) \textstyle\int_{-\infty}^{\infty} y I_{(0,2)}(y)\,dy \\ &= x\, I_{(0,1)}(x) \underbrace{\textstyle\int_{0}^{2} y \cdot 1\,dy}_{2} \\ &= 2x\, I_{(0,1)}(x) \end{aligned}$$

Wow! So precise and neat! Are you sold on indicators yet?

The marginal pdf for Y is

$$\begin{aligned} f_Y(y) &= \textstyle\int_{-\infty}^{\infty} f(x,y)\,dx = \int_{-\infty}^{\infty} xy\, I_{(0,1)}(x) I_{(0,2)}(y)\,dx \\ &= y\, I_{(0,2)}(y) \textstyle\int_{-\infty}^{\infty} x\, I_{(0,1)}\,dx \\ &= y\, I_{(0,2)}(y) \underbrace{\textstyle\int_{0}^{1} x \cdot 1\,dx}_{1/2} \\ &= \tfrac{1}{2} y\, I_{(0,2)}(y). \end{aligned}$$

Since $f(x,y) = f_X(x) \cdot f_Y(y)$, we have that X and Y are independent.

Note

If we just wanted to show independence, is it really necessary to go through the work of finding the marginal pdfs? Can't we just say that the pdf $f(x,y) = xy$ factors into an "x-part" and a "y-part even if we don't know quite how the constants will sort out to make marginal pdfs?

The answer is **yes, if we use indicators!!!** (See exercises.)

Example 0.9.3 (Continuous)

For the last example, we have

$$f(x,y) = xy\, I_{(0,1)}(x) I_{(0,2)}(y) = \underbrace{\Big(x I_{(0,1)}(x)\Big)}_{x-part} \cdot \underbrace{\Big(y I_{(0,2)}(y)\Big)}_{y-part}$$

$$\Rightarrow \quad X \text{ and } Y \text{ are independent}$$

Next up though is an example of a pdf that factors into an "x-part" and a "y-part but where X and Y are not independent. This is why we must include the indicators if we want to use this "lazy" (no work to find marginals) way to show independence.

Example 0.9.4

Suppose that X and Y are continuous random variables with joint pdf

$$f(x,y) = \begin{cases} 8xy & , \quad 0 < x < y < 1 \\ 0 & , \quad \text{otherwise} \end{cases}$$

For practice in finding regions of integration, let's check that this pdf integrates to 1.

$$\begin{aligned} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dy\,dx &= \int_0^1 \int_x^1 8xy\,dy\,dx \\ &= \int_0^1 8x \underbrace{\int_x^1 y\,dy}_{\frac{1}{2}(1-x^2)}\,dx \\ &= \int_0^1 4x(1-x^2)\,dx = \cdots = 1 \ \checkmark \end{aligned}$$

(To get these limits of integration, sketch the region where $0 < x < y < 1$. It is triangular. Take any x between 0 and 1. For each such fixed x, y then ranges from x to 1. If you want to do the integral in the

opposite order, take any y between 0 and 1. For each fixed y, you'll see that x ranges from 0 to y. So, the double integral can also be written as $\int_0^1 \int_y^1 8xy\,dx\,dy$.)

To find the marginal pdf, $f_X(x)$, for X, think of x as any fixed number between 0 and 1. For any fixed x, y ranges from x to 1. Thus, we get

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy = \int_x^1 8xy\,dy = 8x\int_x^1 y\,dy = 4x(1-x^2).$$

This holds for $0 < x < 1$. The pdf is 0 otherwise. We could just tack on an indicator in order to completely describe $f_X(x)$ or we could have used indicators from the beginning in the joint pdf.

Note that the joint pdf could be written with indicators in two ways:

$$f(x,y) = 8xy\,I_{(0,1)}(y)\,I_{(0,y)}(x) = 8xy\,I_{(0,1)}(x)\,I_{(x,1)}(y).$$

To find the marginal pdf for X, we should use the second representation since it has an indicator that is purely in terms of x which can be easily pulled out of an integral with respect to y:

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x,y)\,dy = \int_{-\infty}^{\infty} 8xy\,I_{(0,1)}(x)\,I_{(x,1)}(y)\,dy \\ &= 8x\,I_{(0,1)}(x)\int_{-\infty}^{\infty} y\,I_{(x,1)}(y)\,dy \\ &= 8x\,I_{(0,1)}(x)\int_x^1 y\cdot 1\,dy = 4x(1-x^2)\,I_{(0,1)}(x). \end{aligned}$$

The marginal pdf for Y is

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f(x,y)\,dx = \int_{-\infty}^{\infty} 8xy\,I_{(0,1)}(y)\,I_{(0,y)}(x)\,dx \\ &= 8y\,I_{(0,1)}(y)\int_{-\infty}^{\infty} x\,I_{(0,y)}(x)\,dx \\ &= 8y\,I_{(0,1)}(y)\int_0^y x\cdot 1\,dy = 4y^3\,I_{(0,1)}(y). \end{aligned}$$

Clearly, we do not have that $f(x,y) = f_X(x)\cdot f_Y(y)$. Thus, X and Y are not independent. We could have seen this without first finding the marginal pdfs since the indicators in the joint pdf can not be separated into an "x-part" and a "y-part".

While the definition of independence says that X and Y are independent if the joint pdf can be factored into a product of the marginal pdfs, we tried to claim here that maybe we had independence if the joint pdf just factors into an x-part and a y-part and that we can leave it up to someone else to find the marginals. This claim appears to fail for the joint pdf $f(x,y) = 8xy$, which can be factored. However, the x and y were still tied together in the condition that $0 < x < y < 1$. If we included this constraint in the joint pdf with indicators, we actually can show independence. We have the following.

Property

Random variables X and Y are **independent** if

$$f(x,y) = g(x) \cdot h(y).$$

for some functions g and h, as long as we include indicators in the joint pdf!

The proof of this property is left as an exercise.

For the previous example, the indicator part of the joint pdf was written in two ways,

$$I_{(0,1)}(y)\, I_{(0,y)}(x) \qquad \text{and} \qquad I_{(0,1)}(x)\, I_{(x,1)}(y),$$

but neither representation completely factors into an x-part and a y-part.

0.9.4 "Twisted" Indicators

Such "twisted indicators", as in the example at the end of the previous section, where the x and y parts are all "mixed together", can sometimes be sorted out and separated. To see whether or not this is the case, it is often helpful to sketch the region where the product of the indicators is equal to 1.

Consider for example,

$$I_{(0,\infty)}(xy) \cdot I_{(0,\infty)}(y - xy).$$

This product of indicators takes the value of 1 whenever both

$$0 < xy < \infty \qquad \text{and} \qquad 0 < y - xy < \infty,$$

otherwise the product will be 0.

It is helpful to rewrite that second inequality as $0 < y(1-x) < \infty$. Then we see that it will hold when $y > 0$ and $x < 1$ OR when $y < 0$ and $x > 1$. Figure 2 shows places where $I_{(0,\infty)}(xy)$ equals 1 using horizontal lines and places where $I_{(0,\infty)}(y - xy)$ is 1 using vertical lines.

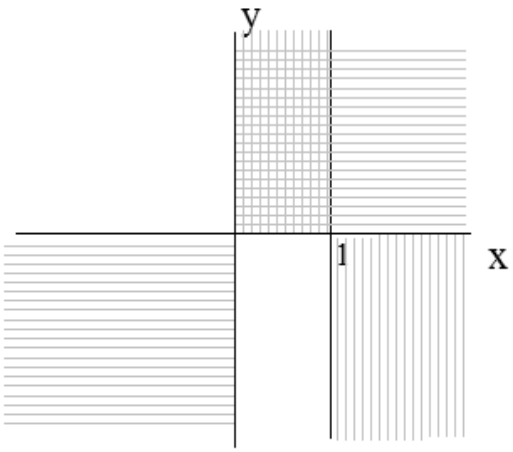
Both indicators (and hence their product) is 1 on the region where $0 < x < 1$ and $0 < y < \infty$. Thus, we have that

$$I_{(0,\infty)}(xy) \cdot I_{(0\infty)}(y - xy) = I_{(0,1)}(x) \cdot I_{(0,\infty)}(y).$$

Note that the indicators then factor into an "x-part" and a "y-part". This would help us to establish independence of random variables X and Y, assuming the rest of their joint pdf factors as well.

You will be able to split up indicators like this **whenever the region they describe is rectangular**. Take a moment to sketch the region that the random variables X and Y live on in Example 0.9.4. It is triangular, and no amount of rewriting is going to allow us to completely separate the indicators into an x part and a y part!

Figure 2: Sorting Out a Product of Indicators



0.10 Useful Properties of Expectation, Variance, and Covariance

In this section we define what is meant by "covariance" and establish several properties of the expectation and variance operators. All proofs will be given for continuous random variables. They can easily be rewritten for the discrete case as well.

1. Suppose that X is a random variable with pdf f. Let g be any function. Then
$$\mathsf{E}[g(X)] = \int_{-\infty}^{\infty} g(x)\, f(x)\, dx. \tag{0.10.3}$$

   Wait, what?!? Does that make sense to you? Shouldn't we think of $g(X)$ as a new random variable and use the pdf for this new random variable when finding the expectation? For example, maybe we should define the random variable $Y = g(X)$, figure out the pdf, f_Y, for Y and compute
$$\mathsf{E}[g(X)] = \mathsf{E}[Y] = \int_{-\infty}^{\infty} y\, f_Y(y)\, dy. \tag{0.10.4}$$

   Yes, as a matter of fact, we should. However, it turns out that the integrals in (0.10.3) and (0.10.4) will give us the same thing! So, we'll generally use (0.10.3) since it does not involve the extra work of first finding f_Y. We will prove the equivalence of (0.10.3) and (0.10.4) in Section 1.5.1 after we have some more machinery built up. Equation (0.10.3) is called the **Law of the Unconscious Statistician**. We suppose this is meant to express that it is done without thinking but we don't know anyone who can do Statistics while unconscious. Do you?

2. Expectation is a linear operator.

   Let X, Y be random variables and let a, b be constants. Then

   $$\mathsf{E}[aX + bY] = a\mathsf{E}[X] + b\mathsf{E}[Y].$$

   To prove this, note that, since the left-hand side is a function of two random variables, it needs to be computed by multiplying against the joint pdf for both random variables. This is a generalization of (0.10.3) for the case of a vector-valued random variable $Z = (X, Y)$ and the function $g(Z) = aX + bY$. We get

$$\begin{aligned}
\mathsf{E}[aX + bY] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(ax + by)f(x,y)\,dx\,dy \\
&= a\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xf(x,y)\,dx\,dy + b\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}yf(x,y)\,dx\,dy \\
&= a\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xf(x,y)\,dy\,dx + b\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}yf(x,y)\,dx\,dy \\
&= a\int_{-\infty}^{\infty}x\int_{-\infty}^{\infty}f(x,y)\,dy\,dx + b\int_{-\infty}^{\infty}y\int_{-\infty}^{\infty}f(x,y)\,dx\,dy \\
&= a\int_{-\infty}^{\infty}xf_X(x)\,dx + b\int_{-\infty}^{\infty}yf_Y(y)\,dy \\
&= a\mathsf{E}[X] + b\mathsf{E}[Y]. \quad \checkmark
\end{aligned}$$

> **Property**
>
> In general, for random variables $X_1, X_2, \ldots, X_n$ and constants $a_1, a_2, \ldots, a_n$, we have
>
> $$\mathsf{E}[\sum_{i=1}^{n} a_i X_i] = \sum_{i=1}^{n} a_i \mathsf{E}[X_i].$$

3. Expectations of products factor if things are independent.

> **Property**
>
> If X and Y are independent random variables, then
>
> $$\mathsf{E}[XY] = \mathsf{E}[X] \cdot \mathsf{E}[Y].$$

Proof :

$$\begin{aligned}
\mathsf{E}[XY] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f(x,y)\, dx\, dy \\
&\overset{indep}{=} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f_X(x) f_Y(y)\, dx\, dy \\
&= \int_{-\infty}^{\infty} y f_Y(y) \int_{-\infty}^{\infty} x f_X(x)\, dx\, dy \\
&= \int_{-\infty}^{\infty} y f_Y(y) \mathsf{E}[X]\, dy \\
&= \mathsf{E}[X] \int_{-\infty}^{\infty} y f_Y(y)\, dy \\
&= \mathsf{E}[X] \cdot \mathsf{E}[Y] \quad \checkmark
\end{aligned}$$

■

It is important to note that the reverse is not necessarily true– the expectation factoring like that does not necessarily imply that X and Y are independent. For example, let X take on values -1, 0, and 1, with respective probabilities $1/4$, $1/2$, and $1/4$, and let $Y = X^2$. Then, it is easy to see that $\mathsf{E}[X] = 0$, and, we also have

$$\mathsf{E}[XY] = \mathsf{E}[X^3] = (-1)^3(1/4) + (0)^3(1/2) + (1)^3(1/4) = 0.$$

Thus, $\mathsf{E}[XY] = \mathsf{E}[X] \cdot \mathsf{E}[Y]$, but X and Y are most certainly not independent!

Property

In general, if X and Y are independent, then for functions g and h,

$$\mathsf{E}[g(X) \cdot h(Y)] = \mathsf{E}[g(X)] \cdot \mathsf{E}[h(Y)].$$

The proof of this property is very similar to the proof of the previous property.

4. The variance of a random variable X is a measure of spread about its mean. Recall, from Section 0.7, that we use μ to denote the mean $\mathsf{E}[X]$ and σ^2 to denote $Var[X]$, and that the variance of X is defined by

$$Var[X] = \mathsf{E}[(X-\mu)^2].$$

Also, recall that we have an alternate, and usually more computationally convenient, formula

$$Var[X] = \mathsf{E}[X^2] - (\mathsf{E}[X])^2.$$

which we have already proven in Section 0.7. From either the original definition of variance or from this alternate and usually more computationally convenient formula, it is easy to verify that

Property

$$Var[aX] = a^2\, Var[X]$$

for any constant $a \in \mathbb{R}$.

Let us consider now $Var[X + Y]$ for two random variables X and Y.

In general, $Var[X + Y] \neq Var[X] + Var[Y]$.

As you might have guessed, the "square" in the definition of variance just makes things "ooky" here. In order to write down an expression for $Var[X + Y]$, we first need to discuss the concept of covariance.

5. The **covariance** between random variables X and Y is defined as follows.

Definition 0.10.1

$$Cov(X, Y) = \mathsf{E}[(X - \mu_X)(Y - \mu_Y)]$$

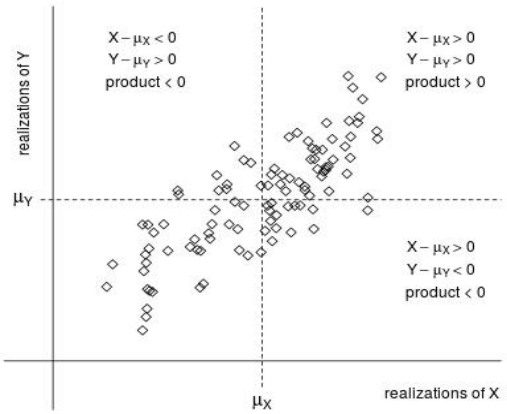
where $\mu_X = \mathsf{E}[X]$ and $\mu_Y = \mathsf{E}[Y]$.

Covariance is a measure of the strength of a linear relationship between two random variables. For example, suppose we have random variables X and Y with some realizations depicted as in Figure 3 along with dashed lines representing their means. Note that these means, μ_X and μ_Y, are true expected values based on some knowledge of the distribution of X and Y. These are different from "sample means", denoted by $\overline{X}$ and $\overline{Y}$, which are computed by averaging the finite number of data values. (Now if you didn't know μ_X and μ_Y, one's natural inclination might be to look at some data and estimate them with $\overline{X}$ and $\overline{Y}$, respectively. Assessing whether or not this is a good idea will be a central theme of this text!)

The plot in Figure 3 is separated into four quadrants based on the locations of the distribution means. In this particular case, the depicted realizations of the random X-Y pairs are such that the majority of points fall in the quadrants where both X and Y are greater than their means or both are less. In other words, most observations fall where both $X - \mu_X$ and $Y - \mu_Y$ are positive or both are negative. In even more words, most fall in places where the product $(X - \mu_X)(Y - \mu_Y)$ is positive. The covariance between X and Y is defined as the expected value of this product. Recall that this is a probability weighted average of this quantity. If we just averaged the numerical values observed for that product, we would likely produce a positive number (called the **sample covariance**). So, it is likely the true distribution covariance is also positive.

The bottom line is again that covariance is a measure of the strength of a linear relationship between two random variables. If X and Y have a positive covariance, we can roughly say that as one increases the other will tend to increase as well. (Think height and shoe size!) If X and Y have a negative covariance,

Figure 3: A Positive Covariance



we can roughly say that as one increases the other will tend to decrease. (Think oil production and gas prices!) A covariance close to zero may appear as a circular cloud of points where it doesn't appear that X and Y are even related as well as a strong distinct relationship where upward and downward tends are able to cancel each other out!

Note that

$$\begin{aligned} Cov(X,Y) &= \mathsf{E}[(X-\mu_X)(Y-\mu_Y)] \\ &= \mathsf{E}[XY-\mu_Y X-\mu_X Y+\mu_X\mu_Y] \\ &= \mathsf{E}[XY]-\mu_Y\mathsf{E}[X]-\mu_X\mathsf{E}[Y]+\mu_X\mu_Y \\ &= \mathsf{E}[XY]-\mu_Y\mu_X-\mu_X\mu_Y+\mu_X\mu_Y \\ &= \mathsf{E}[XY]-\mu_X\mu_Y \\ &= \mathsf{E}[XY]-\mathsf{E}[X]\mathsf{E}[Y] \end{aligned}$$

Here are some properties of covariance.

A) Since $Cov(X,Y) = \mathsf{E}[XY]-\mathsf{E}[X]\mathsf{E}[Y]$ we see that, if X and Y are independent, $Cov(X,Y) = 0$. Again, the reverse is not true. (See the example above in item number 2 on this list.)

B) Using properties of expectation, it is routine to show/see that

$$\begin{aligned} Cov(aX,Y) &= a\,Cov(X,Y) \\ Cov(X+Y,Z) &= Cov(X,Z)+Cov(Y,Z) \\ Cov(X,Y) &= Cov(Y,X). \end{aligned}$$

That last part allows us to easily conclude that we also have $Cov(X, bY) = b\,Cov(X, Y)$ and $Cov(X, Y + Z) = Cov(X, Y) + Cov(X, Z)$.

C) It is also easy to see, by inspection of the definitions, that

$$Var[X] = Cov(X, X)$$

Property

In general, for random variables $X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_m$ and constants $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_m$, we have

$$Cov(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{n} b_j Y_j) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j Cov(X_i, Y_j).$$

6. Now that we have defined covariance, we are ready to deal with the variance of a sum of random variables.

Property

$$Var[X + Y] = Var[X] + Var[Y] + 2Cov(X, Y)$$

To see this, write

$$\begin{aligned} Var[X + Y] &= Cov(X + Y, X + Y) \\ &= Cov(X, X) + Cov(X, Y) + Cov(Y, X) + Cov(Y, Y) \\ &= Cov(X, X) + 2Cov(X, Y) + Cov(Y, Y) \\ &= Var[X] + 2Cov(X, Y) + Var[Y] \end{aligned}$$

Plugging in $-Y$ for Y and noting that $Var[-Y] = Var[(-1)Y] = (-1)^2 Var[Y] = Var[Y]$, we see that

$$Var[X - Y] = Var[X] + Var[Y] - 2Cov(X, Y).$$

Note that, for independent X and Y, the fact that $Cov(X, Y) = 0$ implies that

$$Var[X + Y] = Var[X] + Var[Y]$$

and

$$Var[X - Y] = Var[X] + Var[Y].$$

Property

In general, for independent random variables $X_1, X_2, \ldots, X_n$ and constants $a_1, a_2, \ldots, a_n$, we have

$$Var\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i^2 Var[X_i].$$

0.11 Conditional PDFs

The information in the Section is not needed until Chapter 5. We recommend skipping this Section for now in order to get your MathStat journey underway and then coming back to it if needed.

Suppose that we randomly select a card from a standard deck of 52 playing cards. What is the probability that it is a "king"? The answer is $4/52$ or $1/13$ because all cards are equally likely to be chosen and because there are exactly 4 kings among the 52 cards.

Now, suppose that we randomly select a card from a standard deck of 52 playing cards, someone else gets a look at it, and tells us that is is a "face card". A face card is either a "king", "queen", or "jack", and there are 4 of each in the deck. With this additional information, what is the probability that it is a king?

This information will help us narrow down the possibilities. We are restricted to only 12 equally likely cards and the chance that we have selected a king is $4/12 = 1/3$.

A probability computed in light of some information is called a **conditional probability**.

The information that the card is a face card is **given information** and the probability of interest is denoted as

$$P(\text{ king } | \text{ face card }) = \frac{1}{3}$$

with the vertical line read as "given".

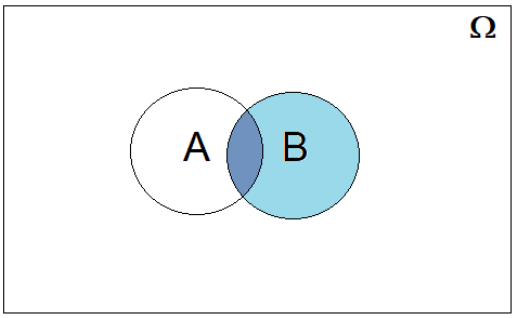
A Venn diagram, as in Figure 4, is often useful for coming up with rules of probability. The rectangle represents the entire sample space and is presumed to have area 1. Let A and B be two events in the sample space which are depicted as circles. The probability that the event A occurs, for example, is represented as the area of the A circle. If we are given that the event B has happened, we can rule out most of the rectangle and focus our attention on the B circle only. The probability of an event A happening given that we know event B has happened is a fraction of the area of B.

Definition 0.11.1

In general, if A and B are two events, the conditional probability that A occurs given that B has occurred is denoted and defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Figure 4: The Conditional Probability $P(A|B)$



Returning to our card selection example,

$$P(\text{king}|\text{face card}) = \frac{P(\text{king and a face card})}{P(\text{face card})}.$$

Since a king is always a face card, this becomes

$$P(\text{king}|\text{face card}) = \frac{P(\text{king and a face card})}{P(\text{face card})} = \frac{P(\text{king})}{P(\text{face card})} = \frac{4/52}{12/52} = \frac{1}{3},$$

which is the same thing we got when we reasoned our way through the problem.

0.11.1 Discrete Random Variables

Let X and Y be discrete random variables. The probability

$$P(X = x, Y = y)$$

is read as "the probability that X equals x and Y equals y". This is the same thing as $P(A \cap B)$ if we define A to be "the event that $X = x$" and define B to be "the event that $Y = y$". Thus, we know how to define the conditional probability that $X = x$ given that $Y = y$.

Definition 0.11.2

For discrete random variables X and Y, the conditional probability that $X = x$ given that $Y = y$ is denoted and defined as

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}.$$

In pdf/pmf notation, we can write this as

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}.$$

Note that, for any fixed y, $f_{X|Y}(x|y)$ is just a pdf in x. As such we have that

$$\sum_x f_{X|Y}(x|y) = 1$$

and, for example, that

$$P(1 \leq X \leq 3|Y = y) = \sum_{\{x:1\leq x\leq 3\}} f_{X|Y}(x|y).$$

Note that we did not write

$$P(1 \leq X \leq 3|Y = y) = \sum_{x=1}^{3} f_{X|Y}(x|y)$$

since X being discrete does not imply that X is integer-valued!

0.11.2 Continuous Random Variables

As we know, probability density functions for continuous random variables do not represent probability. They are curves under which area represents probability. If X and Y are continuous random variables all probabilities in Definition 0.11.2 are zero. However, we do still use an analogous definition for the conditional pdf.

Definition 0.11.3

For continuous random variables X and Y, the conditional pdf for X given $Y = y$ is denoted and defined as

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}.$$

As in the discrete case, for any fixed y, $f_{X|Y}(x|y)$ is just a pdf in x. As such we have that

$$\int_{-\infty}^{\infty} f_{X|Y}(x|y)\,dx = 1$$

and, for example, that

$$P(1 \leq X \leq 3|Y = y) = \int_1^3 f_{X|Y}(x|y)\,dx.$$

The fact that we have the given information that $Y = y$ does not contradict what we have said about continuous random variables and probabilities. Recall our experiment of Section 0.5.3 where we randomly selected a student on campus and measured their height in inches. We assumed that we can measure with infinite accuracy in order to make the height a truly continuous random variable. While there is zero probability that the randomly selected student has a height of 64.01782317 inches, once we have measured them we do have an exact value for their height. So, it is valid to say that a continuous random variable Y is equal to something as long as we are on the right side of the conditional line in a probability statement.

0.11.3 Conditional Distributions and Independence

In Section 0.4 we said that random variables X and Y are independent if their joint pdf factors into a product of their marginal pdfs:

$$f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y).$$

If X and Y are independent, we have

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \overset{indep}{=} \frac{f_X(x)f_Y(y)}{f_Y(y)} = f_X(x).$$

This is a highly intuitive result. If X and Y are independent, the fact that we are given a value for Y is irrelevant when looking at the distribution for X.

On the other hand, suppose we know that

$$f_{X|Y}(x|y) = f_X(x). \tag{0.11.5}$$

For this statement to make any sense, we have to assume that y is a possible value for Y, otherwise, it would never have been observed. This implies that $f_Y(y) \neq 0$ and so the left-hand side is defined. Rewriting (0.11.5), we have

$$\frac{f_{X,Y}(x,y)}{f_Y(y)} = f_X(x)$$

and therefore that

$$f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y).$$

This gives us an alternate equivalent definition of independence.

Definition 0.11.4

X and Y are said to be **independent** if

$$f_{X|Y}(x|y) = f_X(x).$$

Chapter 0 Exercises

1. Suppose that X and Y are discrete random variables with joint pdf
$$f(x,y) = c\,\frac{2^{x+y}}{x!y!}\, I_{\{0,1,2,\ldots\}}(x) I_{\{0,1,2,\ldots\}}(y).$$
   (a). Find the constant c.
   (b). Find the marginal pdfs of X and Y.
   (c). Are X and Y independent? Explain.
2. Let X and Y have the joint pdf
$$f(x,y) = ce^{-x-y} I_{(0,y)}(x) I_{(0,\infty)}(y)$$

for some constant c.

(a). Find the value of c that makes this a valid pdf.

(b). Find the marginal pdfs for X and Y.

(c). Are X and Y independent? Explain.

(d). Find the expected value of X.

3. Let $f(x) = ce^x I_{[0,1]}(x) + ce^{-x+2} I_{(1,\infty)}$ where c is some constant.

(a). Find the constant c that will make f a probability density function.

(b). Find the corresponding cumulative distribution function F.

4. A regular tetrahedron (4 sides) with sides numbered one through four is tossed twice. Let X be the larger of the two "down faces". (The tetrahedron shape does not allow a side to be face up. Just think of this as a four sided die and let X be the larger value obtained in two tosses.)

(a). Find the pdf of X.

(b). Find the expected value of X.

5. Suppose that $X \sim Bernoulli(p)$. Verify the following quantities that can be found in the table of distributions from Appendix A.

(a). Find the expected value of X.

(b). Find the variance of X.

6. Suppose that $X \sim exp(rate = \lambda)$. Verify the following quantities that can be found in the table of distributions from Appendix A.

(a). Find the expected value of X.

(b). Find the variance of X.

7. Suppose that X and Y have joint pdf given by

$$f(x,y) = \begin{cases} 1/2 & , \quad -1 < x < y < 1 \\ 0 & , \quad \text{otherwise} \end{cases}$$

(a). Rewrite this joint pdf using indicator notation. (Note: There is more than one way to do this. Choose one!)

(b). Find $\mathsf{E}[X]$ and $Var[X]$.

8. Suppose that X and Y are independent random variables with $X \sim Bernoulli(p)$ and $Y \sim Poisson(\lambda)$. (We have not talked about the Poisson distribution yet in this text but you can find the necessary pdf in the table of distributions in Appendix A!)
Find the joint probability $P(X = 1, Y = 3)$.

9. Suppose that X is a continuous random variable with pdf

$$f(x) = \frac{4}{(1+x)^5} I_{(0,\infty)}(x).$$

(a). Find, using your table of distributions, the expected value of X. That is, first identify the distribution by simply matching this pdf with one in the table and then pulling the expected value (mean) from the table without any actual computation.

(b). Find $\mathsf{E}[X^2]$. (Don't make this too much work. Can you get this from the table of distributions?)

(c). Consider a random triangle whose sides are of length X and $X + 1$. Determine the expected value of the area of the triangle. (You may also quote values from the table of distributions here.)

10. Suppose that X is a continuous random variable with pdf $f(x) = 3x^2 I_{(0,1)}(x)$.

(a). Find, using the table of distributions from Appendix A, the expected value of X. That is, first identify the distribution by simply matching this pdf with one in the table and then pulling the

expected value (mean) from the table without any actual computation.

(b). Consider a random rectangle whose sides are of length X and $(1 - X)$. Determine the expected value of the area of the rectangle. (You may also quote values from the table of distributions here.)

11. In this chapter, we had a brief discussion about the covariance between random variables X and Y and, in particular, the relationship between X and Y when they have a positive versus negative covariance. The magnitude of a covariance, however, is difficult to interpret on its own and may only be interesting when compared to another covariance. Instead, one may wish to consider a "standardized" version of covariance that is bounded between -1 and 1.

    The (linear) **correlation** between two random variables X and Y is often denoted as $\rho_{X,Y}$ and is defined as

    $$\rho_{X,Y} := \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$

    where $\sigma_{X,Y} := Cov(X, Y)$.

    (a). Show that $-1 \leq \rho_{X,Y} \leq 1$.

    (b). What can one say about the relationship between X and Y if their correlation is a perfect plus or minus 1?

12. Suppose that X and Y are continuous random variables with a joint pdf of the form

    $$f(x, y) = [g(x)\, I_G(x)] \cdot [h(y)\, I_H(y)]$$

    for some functions g and h and some sets of real numbers G and H. As the notation suggests, the sets sets are constant in that, for example, we do not have $G = G(y)$.

    Show that, even thought g and h are not necessarily pdfs, X and Y must be independent.

13. Suppose that $X_1, X_2 \overset{iid}{\sim} Poisson(\lambda)$. Find $P(X = x|X + Y = n)$. Can you name this conditional distribution?

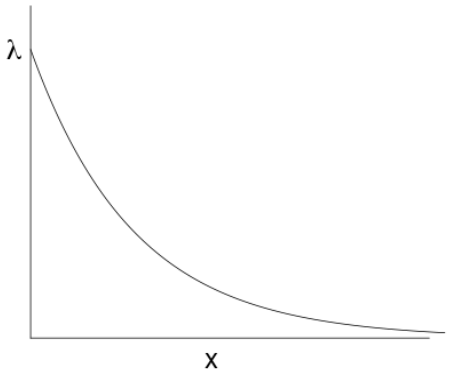
Chapter 1 MathStat Preliminaries: Four Important Tools for Mathematical Statistics

1.1 Wait. Where are we going?

Recall the exponential distribution introduced in Section 0.5.3. If $X \sim exp(rate = \lambda)$, then X has the pdf

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & , \quad x \geq 0 \\ 0 & , \quad x < 0 \end{cases}$$

which looks like this.

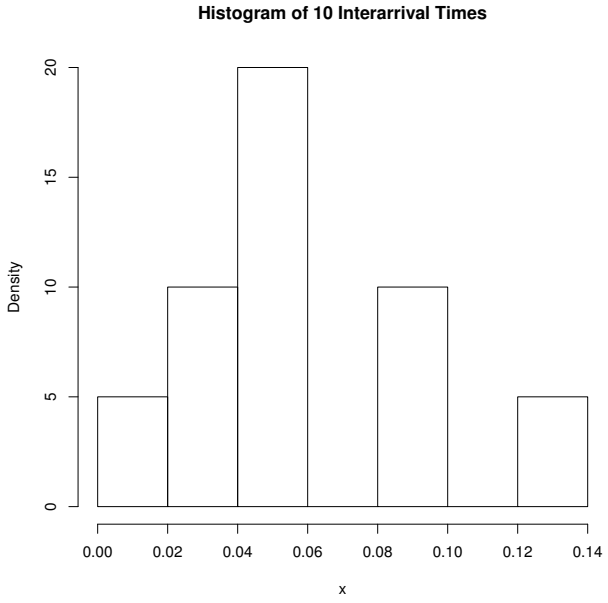


Recall that probabilities associated with X are determined by computing areas under this curve. Since the bulk of the area is on the lower end of the positive x-axis, we have higher probabilities of seeing lower values of X. (Though without any scale shown on the axes it is difficult to say what "lower" really means!)

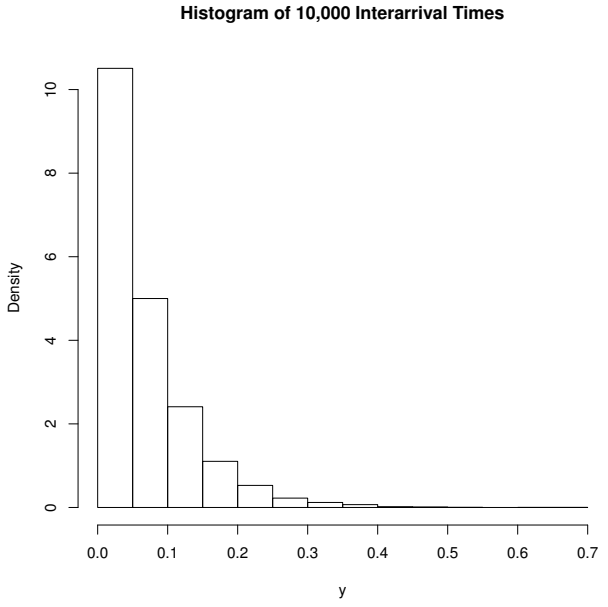
Suppose that we are going (in the future) to observe an actual numerical value for X. Actually, suppose we are going to observe n values for X. Let's call these, yet unobserved (so still random) values $X_1, X_2, \ldots, X_n$. For example, if this is the exponential distribution of Section 0.5.3 that describes a grocery store customer interarrival time, let's position ourselves at the door and get ready to record the interarrival times for the next n customers.

Once customers start arriving, we will record actual numbers for the interarrival times. These numbers are said to be "realizations" of the random variables $X_1, X_2, \ldots, X_n$. The probability that a customer interarrival time is less than 0.3 minutes is theoretically given by the area under the curve depicted above over the interval $(0, 0.3)$. If we want to try to estimate this using our "data", we should compute the proportion of values in our list of realizations that are less than 0.3. In general, if we make a histogram of these realizations, scaled in such a way that the area of a bar over an interval represents the proportion of values in that interval, then the area of that bar should be approximating the desired probability and therefore should be approximating the true area under the pdf.

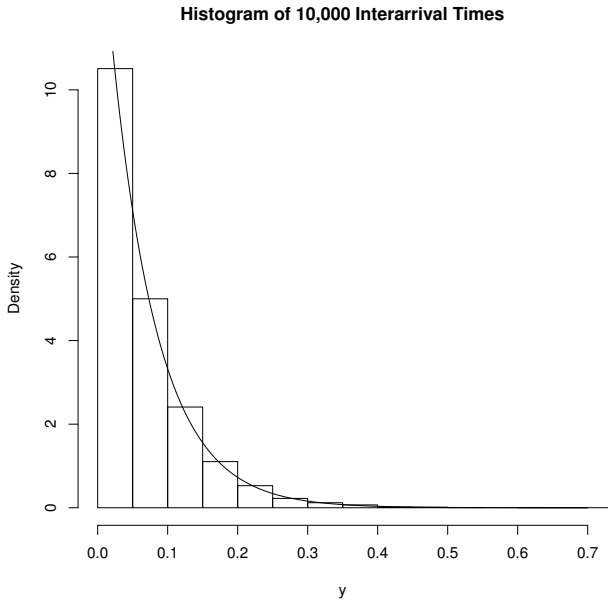
Here is such a histogram of 10 realizations of X where $X \sim exp(rate = 15.2)$.



That doesn't look much like the exponential pdf. However, with more data/realizations, we get the following.



Overlaying the pdf $f(x) = 15.2e^{-15.2x}$ for $x > 0$ gives us a pretty good fit!



Suppose now that we only have those 10,000 data points and the belief that they came from an exponential distribution, but we do not know the value of λ. How would you estimate λ?

We could use the fact that $1/\lambda$ is the expected value (mean) for the exponential distribution. Recall that this is a probability weighted average. We could just average the 10,000 values we sampled/recorded. They already have their "probability weights" built-in in the way they were generated, with, for example, lower values coming out more often. This average is known as a **sample mean** and is denoted by $\overline{x}$. If we have yet to record numerical values but are planning on recording 10,000 of them and averaging them, we would denote the sample mean with the $\overline{X}$, which is another random variable.

Using the observation above about the expected value or "distribution mean" and the sample mean, one plan for estimating λ is to think that

$$1/\lambda \approx \overline{X}$$

and to solve for λ. This gives

$$\lambda \approx 1/\overline{X}.$$

So, once we get the data and compute the sample mean, we could flip it over and use that to estimate the unknown λ. Note that those two approximations are weird things to write because, in both cases, the left-hand side are non-random constants and the right-hand sides are random variables! We will talk about what is really meant by such an approximation in Chapter 2.

An alternative way to estimate the unknown λ might be to notice that it is the y-intercept on the graph of the pdf. Could we make a more refined histogram with really thin bars and try to estimate λ using the height of the first bar? In the last histogram shown above, the first bar is quite a bit lower than 15.2. If we made each bar half as wide, the first bar would likely shoot up much higher. What if we made them one quarter as wide? Surely we couldn't keep shrinking them because we only have a finite amount of data and so we would only have a very small number of values in very small intervals. Indeed, the first bar height/area might shrink to zero!

It seems that the "y-intercept idea" is somehow not as "solid" as the sample mean idea. A large part of mathematical statistics is coming up with ways to estimate parameters after we first quantify what is meant by a "good estimator" and what is meant by a "better estimator". This is the subject of Chapter 2. Before we get started, we are going to need some more tools under our belts! Oh, and a very special trick.

1.1.1 A Very Special Trick

Suppose that we want to compute the following integral.

$$\int_0^\infty 3e^{-2x}\,dx.$$

While the computation is not too hard, we can avoid integrating completely by noting that the integrand almost looks like the pdf for an exponential distribution. In particular, ignoring the constants and just looking at "the x part" of the integrand, this appears to be an exponential pdf with rate $\lambda = 2$. Still, the constant in front of the integrand is not quite right. Let's make it right! We can write

$$\int_0^\infty 3 \underbrace{e^{-2x}}_{\substack{\text{like an}\\ \text{exp(rate=2)}\\ \text{pdf}}} dx = \frac{3}{2}\int_0^\infty \underbrace{2e^{-2x}}_{\substack{\text{is an}\\ \text{exp(rate=2)}\\ \text{pdf}}} dx = \frac{3}{2}\cdot 1 = \frac{3}{2}.$$

We managed to compute an integral by recognizing that it is basically a pdf integrated over it's entire support set after is is adjusted a bit. The more pdfs you come to know, the better your chances will be to be able to turn an integrand into a pdf that you know integrates to 1. We will call this "integrating without integrating". What a clever name!

There is, of course, an analogue to be made for "summing without summing" when working with sums and comparing them to discrete distributions.

1.2 Important Tool I: Finding Distributions of Transformations of Random Variables

1.2.1 The Discrete Case and the Binomial Distribution

Let's use this opportunity first to introduce, or possibly recall, another distribution.

Consider a sequence of n independent trials of an experiment where each trial can result in either "success" (S) or "failure" (F). Suppose that the probability of success remains the same from trial to trial. Call it p where $0 \leq p \leq 1$.

Let

$$X = \#\text{ of successes in } n \text{ trials.}$$

While it looks similar, this is different from the geometric random variable introduced in Chapter 0. There, we continued trials until the first success. Here, we will have a n (a fixed number) of trials and will count up all of the successes.

In this case, X is said to have a **binomial distribution** with parameters n and p. 

We write

$$X \sim bin(n, p).$$

As the number of successes in n trials, X can take on values in $\{0, 1, 2, \ldots, n\}$.

The pdf is

$$f(x) = P(X = x) = P(SSFSF \ldots F \text{ or } SFSFS \ldots S \text{ or} \ldots)$$

where each listed configuration of outcomes includes exactly x S's and exactly $n - x$ F's. Since the outcomes are disjoint, we get

$$f(x) = P(X = x) = P(SSFSF \ldots F) + P(SFSFS \ldots S) + \cdots \tag{1.2.1}$$

Since the trials are independent,

$$P(SSFSF \ldots F) = p \cdot p \cdot (1-p) \cdot s \cdot (1-p) \cdots (1-p) = p^x (1-p)^{n-x}.$$

In fact, every term in (1.6.3) gives that same probability since every term has the same number of S's and F's! So,

$$f(x) = P(X = x) = c \cdot p^x (1-p)^{n-x}$$

where c is the number of terms in (1.6.3).

How many ways are there to write down sequences of n S's and F's with exactly x S's? We need to choose x slots out of n in which to put the S's. There are

$$\left(\begin{array}{c} n \\ x \end{array} \right) = \frac{n!}{x!(n-x)!}$$

different ways to do this. Thus,

$$c = \left(\begin{array}{c} n \\ x \end{array} \right).$$

So, we have seen that $X \sim bin(n, p)$ means that X has pdf

$$f(x) = P(X = x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x} \, I_{\{0,1,\ldots,n\}}(x) \tag{1.2.2}$$

Check out the binomial distribution in the distribution tables in Appendix A.

We are ready for our first transformation.

Example 1.2.1

Suppose that $X \sim bin(n, p)$. Find the distribution (Name it!) for the random variable $Y := n - X$.

The answer to this one is quite "guessable", but let's go through the motions anyway since this will not always be the case.

The pdf for Y is

$$f_Y(y) \quad = \quad P(Y = y) =?$$

What choice do we have but to use the fact that $Y = n - X$? We don't know anything else about Y. Just do the problem and don't worry about a "formula" for transforming the pdf! We have,

$$f_Y(y) \quad = \quad P(Y = y) = P(n - X = y)$$

Now, we do know how to compute probabilities of the form $P(X =?)$, so, solve for X here to get

$$f_Y(y) \quad = \quad P(Y = y) = P(n - X = y) = P(X = n - y)$$

and use the pdf for X, plugging in $n - y$ where there were $x's$ in (1.2.2):

$$\begin{aligned} f_Y(y) \quad &= \quad P(Y = y) = P(n - X = y) = P(X = n - y) \\ &\overset{(1.2.2)}{=} \quad \binom{n}{n-y} p^{n-y}(1-p)^{n-(n-y)} I_{\{0,1,\ldots,n\}}(n-y) \\ &= \quad \binom{n}{n-y} p^{n-y}(1-p)^{y} I_{\{0,1,\ldots,n\}}(y) \end{aligned}$$

We simplified the exponent $n - (n - y)$ for obvious reasons, but we also simplified the indicator (see below for an explanation of how) for two reasons.

1. It's good form. After all, you are reporting a function of y. Why would you say "This holds for $n - y$ in the set $\{0, 1, 2, \ldots n\}$." as opposed to saying what y alone can be?
2. It will make it easier to recognize the distribution of Y.

To simplify the indicator $I_{\{0,1,\ldots,n\}}(n - y)$, note that is is equal to 1 whenever

$$n - y = 0, 1, 2, \ldots, n.$$

But,

$$n - y = 0 \quad \Rightarrow \quad y = n,$$

$$n - y = 1 \quad \Rightarrow \quad y = n - 1,$$

$$\vdots$$

$$n - y = n \quad \Rightarrow \quad y = 0.$$

So, y takes on values in $\{0, 1, \ldots, n\}$.

(Another note on good versus bad "form": order is unimportant for listing elements in a set, but, in my opinion, to say that "y takes on values in $\{n, n-1, \ldots, 0\}$" is kind of weird and, also in my opinion, shows that you are just "plugging and chugging" through steps of a problem without really thinking about what you're doing.)

In summary,

$$I_{\{0,1,\ldots,n\}}(n - y) = I_{\{0,1,\ldots,n\}}(y),$$

so the pdf for Y is

$$f_Y(y) = \begin{pmatrix} n \\ n-y \end{pmatrix} p^{n-y}(1-p)^{n-(n-y)}\, I_{\{0,1,\ldots,n\}}(y).$$

When trying to match this up to a distribution in our table of distributions, you will see only two discrete distributions with this type of indicator. One is the discrete uniform distribution whose pdf does not at all resemble this pdf and the other is the binomial distribution. Note that

$$\begin{pmatrix} n \\ y \end{pmatrix} = \frac{n!}{y!(n-y)!} = \frac{n!}{(n-y)!y!} = \frac{n!}{(n-y)!(n-(n-y))!} = \begin{pmatrix} n \\ n-y \end{pmatrix}.$$

Moving the p's around, we can write the pdf for Y as

$$f_Y(y) = \begin{pmatrix} n \\ y \end{pmatrix} (1-p)^y p^{n-y}\, I_{\{0,1,\ldots,n\}}(y)$$

to see that

$$\boxed{Y \sim bin(n, 1-p).}$$

Surprised? Probably not. If X is the number of successes in n trials, then $Y = n - X$ is the number of failures. Relabeling successes as failures, the "success probability" is now $1 - p$.

Probability density functions of transformations of random variables do not always turn out to be those of nice "named" distributions. When we ask you to "Find the distribution.", we will mean a named distribution, otherwise we would have just asked you to "Find the pdf." To be more clear, we will follow "Find the distribution." with "(Name it!)" Looking first at the indicators in the table of distributions is a good way to narrow down the possibilities.

1.2.2 The Continuous Case and the Gamma Distribution

In the discrete case, we didn't need some sort of general formula for making a transformation of pdfs from one random variable to another. Since pdfs were probabilities, it was easiest just to write out what we want and what we know and to go! In the continuous case, the pdf no longer represents probability (i.e. $f(x) \neq P(X = x)$), so the approach from the discrete case will not make sense. The good news though is that we do have probabilities

in the cdfs! So, the plan will be to find

$$F_Y(y) = P(Y \leq y) = \cdots$$

and then in the end to find the pdf for Y using the fact that

$$f_Y(y) = \frac{d}{dy} F_Y(y).$$

Suppose that X has pdf $f_X(x)$ and cdf $F_X(x)$.

Suppose that Y is defined as $Y = g(X)$.

We will assume that g is invertible. (If it is not, it doesn't mean we are out of luck, it just means that this approach and the formula we are about to derive won't work.)

Note that

1. g invertible $\Rightarrow$ g is either strictly increasing or strictly decreasing.
2. If g is increasing (alternatively decreasing) the g^{-1} is also increasing (alternatively decreasing).
   To see that second point in the increasing case, note that g increasing means that $x_1 \leq x_2$ implies that $g(x_1) \leq g(x_2)$ and that $x_1 \geq x_2$ implies that $g(x_1) \geq g(x_2)$. (It might help to draw a picture.)
   We want to show that g^{-1} is also increasing. Suppose that $x_1 \leq x_2$. We want to show that $g^{-1}(x_1) \leq g^{-1}(x_2)$. Suppose (incorrectly) that
   $$g^{-1}(x_1) \geq g^{-1}(x_2).$$
   We are going to take g of both sides. Since g is increasing, it preserves the order of the inequality:
   $$g(g^{-1}(x_1)) \geq g(g^{-1}(x_2)).$$
   Canceling g and g^{-1} gives
   $$x_1 \geq x_2.$$
   This contradicts the original assumption that $x_1 \leq x_2$. Thus, we must have that $g^{-1}(x_1) \leq g^{-1}(x_2)$ and thus that g^{-1} is increasing.

We are ready to find an expression for the cdf, and then the pdf, for Y.

Case One: g is increasing.

$$F_Y(y) \quad = \quad P(Y \leq y) = P(g(X) \leq y)$$

Applying g^{-1} to both sides of that inequality, and using the fact that g increasing $\Rightarrow$ g^{-1} increasing $\Rightarrow$ g^{-1}

applied to both sides of an inequality will preserve the order, we get

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(g(X) \leq y) \\ &= P(g^{-1}(g(X)) \leq g^{-1}(y)) \\ &= P(X \leq g^{-1}(y)) \\ &= F_X(g^{-1}(y)) \end{aligned}$$

Thus, the pdf for Y is

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} F_X(g^{-1}(y)) \overset{chain\ rule}{=} f_X(g^{-1}(y)) \cdot \frac{d}{dy} g^{-1}(y)$$

Note that g^{-1} increasing implies that that derivative is greater than zero.

Case Two: g is decreasing.

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y)$$

Applying g^{-1} to both sides of that inequality, and using the fact that g decreasing $\Rightarrow g^{-1}$ decreasing $\Rightarrow g^{-1}$ applied to both sides of an inequality will flip the inequality, we get

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(g(X) \leq y) \\ &= P(g^{-1}(g(X)) \geq g^{-1}(y)) \\ &= P(X \geq g^{-1}(y)) \\ &= 1 - P(X < g^{-1}(y) \\ &\overset{contin}{=} 1 - P(X \leq g^{-1}(y)) \\ &= 1 - F_X(g^{-1}(y)) \end{aligned}$$

Thus, the pdf for Y is

$$\begin{aligned} f_Y(y) &= \tfrac{d}{dy} F_Y(y) = \tfrac{d}{dy}[1 - F_X(g^{-1}(y))] \\ &\overset{chain\ rule}{=} [0 - f_X(g^{-1}(y)) \cdot \tfrac{d}{dy} g^{-1}(y)] \\ &= -f_X(g^{-1}(y)) \cdot \tfrac{d}{dy} g^{-1}(y) \end{aligned}$$

Note that, since g^{-1} is decreasing (actually strictly so), we have $\frac{d}{dy} g^{-1}(y) < 0$, so, no, we did not end up with a negative pdf! Furthermore, note that $-\frac{d}{dy} g^{-1}(y)$ is positive and equal to $|\frac{d}{dy} g^{-1}(y)|$.

In the previous increasing case, $\frac{d}{dy}g^{-1}(y)$ is positive and equal to $|\frac{d}{dy}g^{-1}(y)|$.

In summary, we can write both the increasing and decreasing cases together as follows.

Continuous Transformation PDF

Let X be a continuous random variable with pdf f_x. Let Y be a random variable defined by $Y = g(X)$ where g is invertible (and differentiable). Then the pdf for Y can be computed as

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left| \frac{d}{dy} g^{-1}(y) \right| .$$

Example 1.2.2

Let X have a "**gamma distribution** with parameters α and β" We write $X \sim \Gamma(\alpha, \beta)$. This means that X is a continuous random variable with pdf

$$f_X(x) = \frac{1}{\Gamma(\alpha)} \beta^\alpha x^{\alpha-1} e^{-\beta x} \, I_{(0,\infty)}(x)$$

for some parameters $\alpha > 0$ and $\beta > 0$.

Notes:

1. We are using the X subscript on the pdf because we will have multiple pdfs in this problem– one for X and one for a new random variable Y.
2. For some people/books, $X \sim \Gamma(\alpha, \beta)$ means that X has pdf
$$f_X(x) = \frac{1}{\Gamma(\alpha)} (1/\beta)^\alpha x^{\alpha-1} e^{-x/\beta} \, I_{(0,\infty)}(x).$$
Here, α and β are known as the "shape" and "scale" parameters, respectively.
For our form of the gamma pdf, β is known as the "inverse scale parameter".
3. The pdf involves the "gamma function" $\Gamma(\alpha)$. It is just a constant. We will define it after finishing this example. The constant $\Gamma(\alpha)$ should not be confused with $\Gamma(\alpha, \beta)$ (two arguments) which is the name of a distribution.

Let $Y = 5X$. Find the distribution of Y. (Name it!)

Answer:

Here we are looking at the transformation $Y = g(X)$ where $g(x) = 5x$.

$$y = g(x) = 5x \qquad \Rightarrow \qquad x = g^{-1}(y) = y/5$$

So,

$$\begin{aligned} f_Y(y) &= f_X(g^{-1}(y)) \cdot \left|\tfrac{d}{dy} g^{-1}(y)\right| \\ &= \tfrac{1}{\Gamma(\alpha)} \beta^\alpha \left(\tfrac{y}{5}\right)^{\alpha-1} e^{-\beta y/5} \cdot \underbrace{I_{(0,\infty)}(y/5)}_{\| \; I_{(0,\infty)}(y)} \cdot \left|\tfrac{1}{5}\right| \\ &= \tfrac{1}{\Gamma(\alpha)} \left(\tfrac{\beta}{5}\right)^\alpha y^{\alpha-1} e^{-(\beta/5)y} \cdot I_{(0,\infty)}(y) \end{aligned}$$

which is the pdf for the $\Gamma(\alpha, \beta/5)$ distribution. Thus we see that

$$\boxed{Y \sim \Gamma(\alpha, \beta/5).}$$

A pdf is not completely specified without its support. If you do not use indicator functions, you must take a moment to figure out the support of the transformed random variable. Notice that by using the indicator notation, we just made this part of the process.

An Aside: The Gamma Function

The pdf for the gamma distribution was defined using the **gamma function** which is denoted by $\Gamma(\cdot)$.

Definition 1.2.1

The gamma function, is defined, for $\alpha > 0$, as

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} \, dx.$$

Note that, for any $\beta > 0$,

$$\int_0^\infty \beta^\alpha x^{\alpha-1} e^{-\beta x} \, dx = \int_0^\infty (\beta x)^{\alpha-1} e^{-\beta x} \, \beta \, dx \overset{u=\beta x}{=} \int_0^\infty u^{\alpha-1} e^{-u} \, du = \Gamma(\alpha)$$

(Here we have used the fact that $du = \beta \, dx$ and that if x goes from 0 to ∞, then $u = \beta x$ also goes from 0 to ∞ since $\beta > 0$.)

Now,

$$\int_0^\infty \frac{1}{\Gamma(\alpha)} \beta^\alpha x^{\alpha-1} e^{-\beta x} \, dx = \frac{1}{\Gamma(\alpha)} \int_0^\infty \beta^\alpha x^{\alpha-1} e^{-\beta x} \, dx = \frac{1}{\Gamma(\alpha)} \cdot \Gamma(\alpha) = 1,$$

so basically $1/\Gamma(\alpha)$ is the constant that makes $\beta^\alpha x^{\alpha-1} e^{-\beta x}$ into a proper pdf over $x \geq 0$!

Properties of the Gamma Function

1. $\Gamma(1) = 1$
   Proof: $\Gamma(1) = \int_0^\infty x^{1-1}e^{-x}\,dx = \int_0^\infty e^{-x}\,dx = 1$.
2. For $\alpha > 1$,
$$\Gamma(\alpha) = (\alpha-1)\cdot\Gamma(\alpha-1).$$
   Proof: $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}\,dx$ Using integration by parts
$$\int u\,dv = uv - \int v\,du$$
   with $u = x^{\alpha-1}$ and $dv = e^{-x}\,dx$ (So $du = (\alpha-1)x^{\alpha-2}\,dx$ and $v = \int e^{-x}\,dx = -e^{-x}$.), we have
$$\begin{aligned}\Gamma(\alpha-1) &= -x^{\alpha-1}e^{-x}\big|_0^\infty + \int_0^\infty (\alpha-1)x^{\alpha-2}e^{-x}\,dx \\ &= 0 + (\alpha-1)\int_0^\infty x^{\alpha-2}e^{-x}\,dx = (\alpha-1)\cdot\Gamma(\alpha-1)\ \surd\end{aligned}$$
3. If $n \geq 1$ is an integer,
$$\Gamma(n) = (n-1)!.$$
   Proof: By repeated application of property 2,
$$\begin{aligned}\Gamma(n) &= (n-1)\Gamma(n-1) = (n-1)(n-2)\Gamma(n-2) \\ &= \cdots = (n-1)(n-2)\cdots(1)\underbrace{\Gamma(1)}_{1} = (n-1)!\end{aligned}$$

1.3 Important Tool II: Bivariate Transformations

Suppose that X_1 and X_2 are continuous random variables with joint pdf $f_{X_1,X_2}(x_1,x_2)$ and suppose that new random variables Y_1 and Y_2 are defined by
$$Y_1 = g_1(X_1,X_2) \qquad \text{and} \qquad Y_2 = g_2(X_1,X_2).$$
For the purpose of algebra, we can think about these functions without the random variables plugged in:
$$y_1 = g_1(x_1,x_2) \qquad \text{and} \qquad y_2 = g_2(x_1,x_2).$$
Solve for x_1 and x_2 as functions of y_1 and y_2. Call them g_1^{-1} and g_2^{-1}. That is
$$x_1 = g_1^{-1}(y_1,y_2) \qquad \text{and} \qquad x_2 = g_2^{-1}(y_1,y_2).$$
The notation g_1^{-1} and g_2^{-1} is mostly symbolic here. g_1^{-1} is not the inverse of g_1. The two equations, in general, can not be inverted individually. It is the system of two equations which is inverted. When we get x_1 written as a function of (y_1,y_2) we will simply name that function $g_1^{-1}(y_1,y_2)$. Similarly, we get x_2 written as a function of (y_1,y_2) we will name that function $g_2^{-1}(y_1,y_2)$.

Bivariate Transformation PDF

Analogous to the one-dimensional case, the joint pdf for Y_1 and Y_2 is given by

$$f_{Y_1,Y_2}(y_1,y_2) = f_{X_1,X_2}(g_1^{-1}(y_1,y_2), g_2^{-1}(y_1,y_2)) \cdot |J|$$

where $|J|$ is the absolute value of the Jacobian of the transformation which is given by the determinant

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix}.$$

We will not go through the derivation of this formula here, as it is too far a tangent into Calculus and will really distract us from Mathematical Statistics! See Appendix B if you wish to read more about Jacobians. It is completely optional (and tedious) reading! This formula can be extended in an obvious way for making a transfomation from the joint pdf of $X_1, X_2, \ldots, X_n$ to the joint pdf of $Y_1, Y_2, \ldots, Y_n$ where $Y_i = g_i(X_1, X_2, \ldots, X_n)$.

The above formula is for "bivariate-to-bivariate" transformations, but you may also find it useful for a "bivariate-to-univariate" transformation as in the following example.

Example 1.3.1

Let $X_1, X_2 \overset{iid}{\sim} \Gamma(\alpha, \beta)$.

This is read as "Let X_1 and X_2 be 'eye-eye-dee' $\Gamma(\alpha, \beta)$." the "iid" stands for **"independent and identically distributed"**. It means that X_1 and X_2 are independent and come from the exact same distribution– in this case, the $\Gamma(\alpha, \beta)$ distribution.

Alternatively, one might say that X_1, X_2 is a **random sample** from the $\Gamma(\alpha, \beta)$ distribution. The words "random sample" include the concept of independence!

Let

$$Y = \frac{X_1}{X_1 + X_2}.$$

Find the distribution (Name it!) of Y.

Why is this example in the "bivariate-to-bivariate" transformation section? It's here because perhaps the easiest way to find this distribution is to think of Y as some Y_1 and then to define a convenient Y_2 for which we can write down the joint pdf for Y_1 and Y_2 after which we can hope to be able to integrate out Y_2 to find the marginal pdf for Y_1!

Technically, Y_2 can be anything you want, but some choices are more convenient than others– especially for the integration down to the marginal pdf for Y_1 at the end.

If $\mathbf{Y_1}$ is a ratio, it is almost always a good idea to choose $\mathbf{Y_2}$ to be the denominator.

Let

$$Y_1 = \frac{X_1}{X_1 + X_2} \qquad \text{and} \qquad Y_2 = X_1 + X_2.$$

Then $y_1 = g_1(x_1, x_2) = x_1/(x_1 + x_2)$ and $y_2 = g_2(x_1, x_2) = x_1 + x_2$. Solving for x_1 ad x_2 gives

$$x_1 = y_1 y_2 \qquad \text{and} \qquad x_2 = y_2 - y_1 y_2,$$

which become the definitions of $g_1^{-1}(y_1, y_2)$ and $g_2^{-1}(y_1, y_2)$, respectively.

The Jacobian is then

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix} = \begin{vmatrix} y_2 & y_1 \\ -y_2 & 1 - y_1 \end{vmatrix} = y_2(1 - y_1) - (-y_1 y_2) = y_2.$$

Since $X_1, X_2 \overset{iid}{\sim} \Gamma(\alpha, \beta)$, their joint pdf is

$$\begin{aligned} f_{X_1,X_2}(x_1, x_2) &\overset{indep}{=} f_{X_1(x_1)} \cdot f_{X_2}(x_2) \\ &= \tfrac{1}{\Gamma(\alpha)} \beta^\alpha x_1^{\alpha-1} e^{-\beta x_1} \cdot I_{(0,\infty)}(x_1) \cdot \tfrac{1}{\Gamma(\alpha)} \beta^\alpha x_2^{\alpha-1} e^{-\beta x_2} \cdot I_{(0,\infty)}(x_2) \\ &= \tfrac{1}{[\Gamma(\alpha)]^2} \beta^{2\alpha} (x_1 x_2)^{\alpha-1} e^{-\beta(x_1+x_2)} \cdot I_{(0,\infty)}(x_1) \cdot I_{(0,\infty)}(x_2) \end{aligned}$$

Thus, the joint pdf for Y_1 and Y_2 is

$$\begin{aligned} f_{Y_1,Y_2}(y_1, y_2) &= f_{X_1,X_2}(g_1^{-1}(y_1, y_2), g^{-1}(y_1, y_2)) \cdot |J| \\ &= f_{X_1,X_2}(y_1 y_2, y_2 - y_1 y_2) \cdot |J| \\ &= \tfrac{1}{[\Gamma(\alpha)]^2} \beta^{2\alpha} (y_1 y_2^2 - y_1^2 y_2^2)^{\alpha-1} e^{-\beta y_2} \cdot I_{(0,\infty)}(y_1 y_2) \cdot I_{(0,\infty)}(y_2 - y_1 y_2) \cdot |y_2| \end{aligned}$$

We are going to need to simplify this in order to integrate out y_2 and, really, just to be "good citizens of the world". First, note that the product of indicators is the one looked at in Section 0.9.4. From there, we know that

$$I_{(0,\infty)}(y_1 y_2) \cdot I_{(0,\infty)}(y_2 - y_1 y_2) = I_{(0,1}(y_1) \cdot I_{(0,\infty)}(y_2).$$

Since this product of indicators tells us, in particular, that $y_2 > 0$, we can drop the absolute value on y_2 in the joint pdf for Y_1 and Y_2. Since we will be integrating out y_2, let's separate out the $(y_1 y_2^2 - y_1^2 y_2^2)^{\alpha-1}$

part to $(y_2^2)^{\alpha-1} \cdot (y_1 - y_1^2)^{\alpha-1}$. After multiplying in the lone $|y_2| = y_2$ from the Jacobian, we get

$$f_{Y_1,Y_2}(y_1, y_2) = \frac{1}{[\Gamma(\alpha)]^2}\, \beta^{2\alpha}\, y_2^{2\alpha-1} [y_1(1-y_1)]^{\alpha-1}\, e^{-\beta y_2}\, I_{(0,1)}(y_1) \cdot I_{(0,\infty)}(y_2).$$

Now for the marginal...

$$\begin{aligned} f_{Y_1}(y_1) &= \int_{-\infty}^{\infty} f_{Y_1,Y_2}(y_1, y_2)\, dy_2 \\ &= \frac{1}{[\Gamma(\alpha)]^2}\, \beta^{2\alpha}\, [y_1(1-y_1)]^{\alpha-1}\, I_{(0,1)}(y_1) \int_0^{\infty} \underbrace{y_2^{2\alpha-1} e^{-\beta y_2}}_{\substack{\text{like a} \\ \Gamma(2\alpha,\beta) \\ \text{pdf}}}\, dy_2 \\ &= \frac{\Gamma(2\alpha)}{[\Gamma(\alpha)]^2}\, [y_1(1-y_1)]^{\alpha-1}\, I_{(0,1)}(y_1) \int_0^{\infty} \underbrace{\frac{1}{\Gamma(2\alpha)}\, \beta^{2\alpha}\, y_2^{2\alpha-1}\, e^{-\beta y_2}}_{\substack{\text{actually a } \Gamma(2\alpha,\beta) \\ \text{pdf, so integrates} \\ \text{to 1}}}\, dy_2 \\ &= \frac{\Gamma(2\alpha)}{[\Gamma(\alpha)]^2}\, [y_1(1-y_1)]^{\alpha-1}\, I_{(0,1)}(y_1) \end{aligned}$$

If you look in your table of distributions (continuous side) you'll see that there are only two possibilities for this distribution based on the indicator. One is the uniform distribution on the interval $(0, 1)$. This is a flat line distribution with height 1 on the interval. It has pdf $f(x) = I_{(0,1)}(x)$. The marginal pdf we have computed is a little more complicated than this.

The second distribution on a finite interval, and, in particular on the interval from $(0, 1)$, is called the **Beta distribution**. This distribution has two parameters, α and β, and is given in the table as having pdf

$$f(x) = \frac{1}{\mathcal{B}(a,b)}\, x^{a-1}(1-x)^{b-1}\, dx.$$

We would write $X \sim Beta(a, b)$.

Our marginal pdf for Y_1 looks a lot like this one (written as a function of y_1). But... what about our "Γ stuff" versus this thing we have yet to define written as $\mathcal{B}(a, b)$? (This is known as the "Beta function".)

The "Γ stuff" and the "Beta stuff" are just constants. They have to match in order for these both to be pdfs! For example, we know that $f(x) = 2e^{-2x}\, I_{(0,1)}(x)$ is a pdf. It's the exponential with rate 2 pdf. From this, we know that the function $5e^{-2x}\, I_{(0,\infty)}(x)$ is **not** a pdf! It can't integrate to 1 if $2e^{-2x}\, I_{(0,\infty)}(x)$ integrates to 1. In fact,

$$\int_0^{\infty} 5e^{-2x}\, dx = \frac{5}{2} \int_0^{\infty} 2e^{-2x}\, dx = \frac{5}{2} \cdot 1 = \frac{5}{2}.$$

If you followed our transformational procedures correctly, you are guaranteed to get a pdf in the end. Thus,

our pdf

$$f_{Y_1}(y_1) = \frac{\Gamma(2\alpha)}{[\Gamma(\alpha)]^2}\,[y_1(1-y_1)]^{\alpha-1}\,I_{(0,1)}(y_1)$$

must be the pdf for the **Beta distribution with parameters a $= \alpha$ and b $= \alpha$**. We write

$$\boxed{Y \sim Beta(\alpha, \alpha).}$$

Note that we did not need to integrate anything in the previous example. When we wrote the marginal pdf for Y_1 out with an integral, you'll notices that it has separated out into a "y_1-part" and a "y_2-part". Since we included the indicators, this is enough to say that Y_1 and Y_2 are independent. That integral is just going to be a constant. The pdf for Y_1 is

$$y_1^{\alpha-1}(1-y_1)^{\alpha-1}\,I_{(0,1)}(y_1)$$

multiplied by a bunch of constants. Let someone else sort them out! If you did everything correctly, then you have a valid pdf that must integrate to 1. There is only one constant that can make this happen.

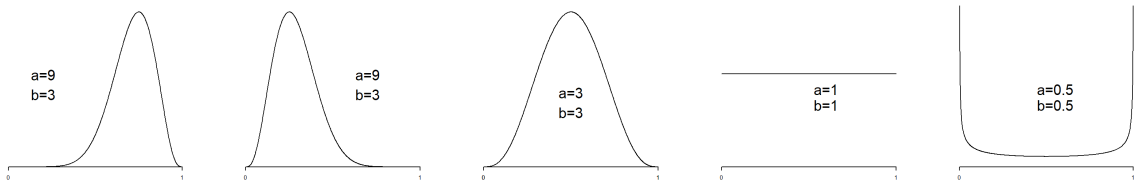
1.3.1 The Beta Distribution

Let's talk about the Beta distribution a bit more. Let X be a continuous random variable with pdf

$$f(x) = \frac{1}{\mathcal{B}(a,b)}x^{a-1}(1-x)^{b-1}\,I_{(0,1)}(x)$$

for some parameters $a, b > 0$. Here, $\mathcal{B}(a,b)$ is known as the **Beta function** which we will define in a moment. The notation (since there is no x in it) suggests (correctly!) that this is a constant.

This distribution is a nice flexible distribution that is suitable to model things that are supposed to be between 0 and 1.



Note that when $a = b = 1$, it is the uniform distribution on $(0,1)$. For $a, b < 1$, it has a "u-shape". For $a, b > 1$ it is a unimodal distribution that is tied down to 0 at both endpoints and various values of a and b in this range can move the high point around to wherever you want it!)

Just for the record though, we define the Beta function.

Definition 1.3.1

The **Beta function** is defined, for $a, b > 0$, as

$$\mathcal{B}(a,b) = \int_0^1 x^{a-1}\,(1-x)^{b-1}\,dx.$$

We have the following fun relationship.

Property

$$\mathcal{B}(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

Proof :

$$\begin{aligned}\Gamma(a)\Gamma(b) &= \left(\int_0^\infty x^{a-1}e^{-x}\,dx\right)\left(\int_0^\infty y^{b-1}e^{-y}\,dy\right)\\ &= \int_0^\infty\int_0^\infty x^{a-1}y^{b-1}e^{-x-y}\,dx\,dy\end{aligned}$$

Now make the change of variables $x = uv$, $y = u(1-v)$. The Jacobian of this transformation is

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix}$$

and u and v range over $(0,\infty)$ and $(0,1)$, respectively. We now have

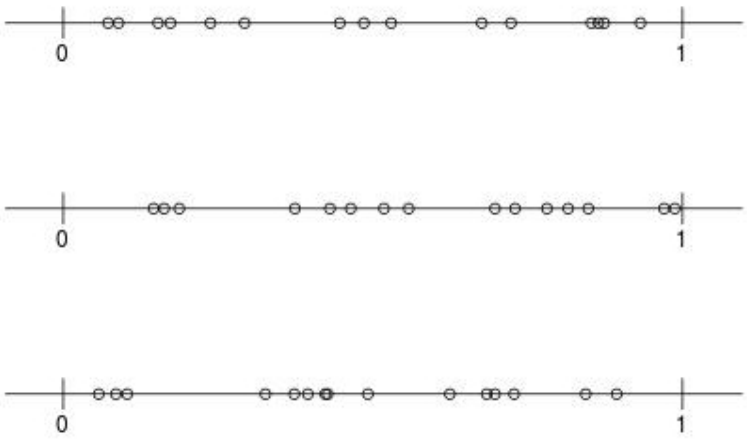
$$\begin{aligned}\Gamma(a)\Gamma(b) &= \int_0^\infty\int_0^\infty x^{a-1}y^{b-1}e^{-x-y}\,dx\,dy\\ &= \int_0^1\int_0^\infty (uv)^{a-1}[u(1-v)]^{b-1}e^{-uv-u(1-v)}|-u|\,du\,dv\\ &= \int_0^1\int_0^\infty u^{a+b-1}v^{a-1}(1-v)^{b-1}e^{-u}\,du\,dv\\ &= \left(\int_0^\infty u^{a+b-1}e^{-u}\,du\right)\left(\int_0^1 v^{a-1}(1-v)^{b-1}\,dv\right)\\ &= \Gamma(a+b)\,\mathcal{B}(a,b)\end{aligned}$$

as desired. ■

1.4 Important Tool III: Minimums and Maximums

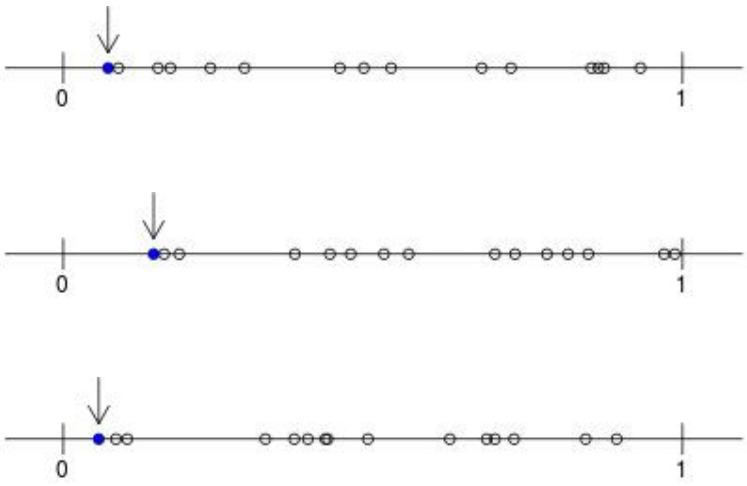
1.4.1 Introduction and Notation

Suppose that $X_1, X_2, \ldots, X_{15}$ is a random sample of size 15 from the uniform distribution over the interval $(0,1)$. Here are three different realizations of such samples.



Because these samples come from a uniform distribution, we expect them to be spread out "randomly" and "evenly" across the interval $(0, 1)$. You might think that you are seeing some sort of clustering but keep in mind that you are looking at small samples of size 15. After collecting more values your view would surely change!

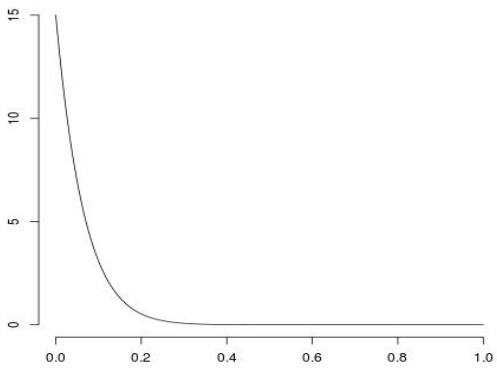
Consider the single smallest value from each of these three samples, highlighted here.



Collect the minimums onto a single graph.

Not surprisingly, they are down towards zero. It would be pretty difficult to get a sample of 15 uniforms on $(0, 1)$ that has a **minimum** up by the right endpoint of the interval. In fact, we can show, using the techniques in this Section, that if we keep collecting minimums of samples of size 15, they would have a probability

density function that looks like this.



Note that the bulk of the area under the curve is between 0 and 0.2.

Similarly, we could collect the maximum values from each sample to get a sense of how they are distributed. We would expect the bulk of this distribution to be in the upper part of the interval, say, between 0.8 and 1.

Let's introduce some notation.

Notation

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from some distribution. We denote the "order statistics" by

$$\begin{aligned} X_{(1)} &= \min(X_1, X_2, \ldots, X_n) \\ X_{(2)} &= \text{the 2nd smallest of } X_1, X_2, \ldots, X_n \\ \vdots &= \vdots \\ X_{(n)} &= \max(X_1, X_2, \ldots, X_n) \end{aligned}$$

(Another commonly used notation is $X_{1:n}, X_{2:n}, \ldots, X_{n:n}$ for the minimum through the maximum, respectively. It is maybe more clear since we know that we are talking about the minimum, next smallest, et cetera, out of n, but we will still use the first notation given here.)

In the next Section, we will derive probability density functions for the minimum and maximum of a sample.

1.4.2 The Distribution of a Minimum by Example

Let's just jump right in with an example.

Example 1.4.1

Suppose that

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} exp(rate = \lambda).$$

Let's find the distribution of

$$X_{(1)} = \min(X_1, X_2, \ldots, X_n).$$

To find the distribution of a maximum or a minimum, it is most convenient to look at cdfs.

The cdf for $X_{(1)}$, the minimum, is:

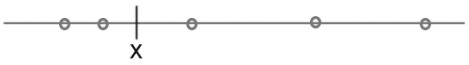
$$F_{X_{(1)}}(x) \quad = \quad P(X_{(1)} \leq x) = P(\min(X_1, X_2, \ldots, X_n) \leq x).$$

(Note that x is just an argument for the function. We can call it y, z, or anything else! There is no reason to use the cumbersome notation $x_{(1)}$.)

Since we know the distribution for the individual X_1 through X_n, it would be useful to us to figure out how the minimum value in the sample relates to these individual values. The cdf for the exponential rate λ distribution is

$$F(x) = \int_{-\infty}^{x} f(u)\, du = \int_{0}^{x} \lambda e^{-\lambda u}\, du = 1 - e^{-\lambda x} \text{ for } x > 0.$$

Note that if the minimum of a group of observed random variables is less than or equal to x, some in the group will be less than or equal to x while others may be greater than x.



It is not clear how many of them should be on each side of x. However, if we consider the equivalent expression

$$P(\min(X_1, X_2, \ldots, X_n) \leq x) = 1 - P(\min(X_1, X_2, \ldots, X_n) > x)$$

things become more clear. If the minimum of a group of random variables, $X_1, X_2, \ldots, X_n$, is greater than x, we are forced to have every $X_i > x$. On the flip side, if every X_i in $X_1, X_2, \ldots, X_n$ is greater than x, then the minimum of this group will definitely be greater than x. In other words, the "event that $\min(X_1, X_2, \ldots, X_n) > x$ is equivalent to the event that $X_1 > x$ and $X_2 > x$ and $\cdots$ and $X_n > x$. Thus,

the cdf for $X_{(1)}$, the minimum, is:

$$\begin{aligned} F_{X_{(1)}}(x) &= P(X_{(1)} \leq x) = P(\min(X_1, X_2, \ldots, X_n) \leq x) \\ &= 1 - P(\min(X_1, X_2, \ldots, X_n) > x) \\ &\overset{indep}{=} 1 - P(X_1 > x) \cdot P(X_2 > x) \cdots P(X_n > x). \end{aligned}$$

Since $X_1, X_2, \ldots, X_n$ are all identically distributed, all of these individual probabilities are the same, thus, we have that

$$\begin{aligned} F_{X_{(1)}}(x) &= P(X_{(1)} \leq x) = P(\min(X_1, X_2, \ldots, X_n) \leq x) \\ &= 1 - P(\min(X_1, X_2, \ldots, X_n) > x) \\ &\overset{indep}{=} 1 - P(X_1 > x) \cdot P(X_2 > x) \cdots P(X_n > x) \\ &\overset{ident}{=} 1 - [P(X_1 > x)]^n. \end{aligned}$$

To compute $P(X_1 > x)$, we could integrate the exponential pdf from x to ∞, or we could use our already computed cdf, $F(x) = P(X_1 \leq x) = 1 - e^{-\lambda x}$. We then have

$$P(X_1 > x) = 1 - P(X_1 \leq x) = 1 - (1 - e^{-\lambda x}) = e^{-\lambda x}.$$

So, the cdf for the minimum $X_{(1)}$ is

$$F_{X_{(1)}}(x) = 1 - [e^{-\lambda x}]^n = 1 - e^{-n\lambda x}.$$

The pdf for the minimum can be found by taking the derivative:

$$f_{X_{(1)}}(x) = \frac{d}{dx} F_Y(y) = \frac{d}{dx}[1 - e^{-n\lambda x}] = n\lambda e^{-n\lambda x}.$$

Note that $X_{(1)}$, as the minimum of a bunch of values in $(0, \infty)$, also lives in $(0, \infty)$. For completeness we write

$$f_{X_{(1)}}(x) = \frac{d}{dx} F_{X_{(1)}}(x) = \frac{d}{dx}[1 - e^{-n\lambda x}] = \lambda n e^{-n\lambda x}\, I_{(0,\infty)}(x).$$

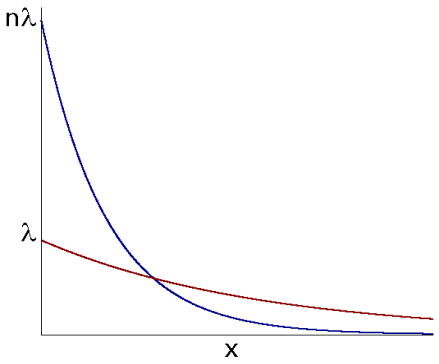
So, we have seen that

$$\boxed{X_{(1)} \sim exp(rate = n\lambda).}$$

It should surprise you to find that the minimum of exponential random variables is also exponential. There is no reason at this point in your MathStat journey to have seen that coming. Given that is is exponential

though, you should not be surprised at the larger rate.

The graph of the exponential rate λ pdf starts at λ when $x = 0$ and goes down, well, "exponentially". The pdf of the exponential rate $n\lambda$ starts higher up at $n\lambda$ and so it must come down faster if the total area under both curves is supposed to be one. That is, it has more area closer to $x = 0$, like this.



Note also that the mean of the distributions are $1/\lambda$ and $1/(n\lambda)$, respectively. So, the expected value of the minimum $X_{(1)}$ is smaller than the expected value of any of the individual X_i which is not surprising! (One might even say that it is expected!)

1.4.3 PDFs for Minimums and Maximums

In this section, we will derive the pdf and cdf for minimums and maximums for any random sample with pdf f and cdf F.

Minimums:

In the example of Section 1.4.2, we have already basically derived an expression for the cdf of the minimum in terms of F, the cdf for the original random variables.

$$\begin{aligned} F_{X_{(1)}}(x) &= P(X_{(1)} \leq x) = P(\min(X_1, X_2, \ldots, X_n) \leq x) \\ &= 1 - P(\min(X_1, X_2, \ldots, X_n) > x) \\ &\overset{indep}{=} 1 - P(X_1 > x) \cdot P(X_2 > x) \cdots P(X_n > x) \\ &\overset{ident}{=} 1 - [P(X_1 > x)]^n \\ &= 1 - [1 - F(x)]^n \end{aligned}$$

This expression for the cdf of the minimum holds for both discrete and continuous random variables. To get the pdf for the minimum in the discrete case, one would have to take differences of the cdf as described in Section 0.5.4.

For continuous random variables, We are able to get a nice formula for the pdf by taking a derivative.

$$\begin{aligned} \Rightarrow \quad f_{X_{(1)}}(x) &= \tfrac{d}{dx} F_{X_{(1)}}(x) \\ &= \tfrac{d}{dx} \{1 - [1 - F(x)]^n\} \\ &= 0 - n[1 - F(x)]^{n-1} \cdot (-f(x)) \\ &= n[1 - F(x)]^{n-1} f(x). \end{aligned}$$

PDF for a Minimum

In summary, if $X_{(1)} = \min(X_1, X_2, \ldots, X_n)$ where $X_1, X_2, \ldots, X_n$ is a random sample from a continuous distribution with pdf f and cdf F, the pdf for $X_{(1)}$ is

$$f_{X_{(1)}}(x) = n[1 - F(x)]^{n-1} f(x).$$

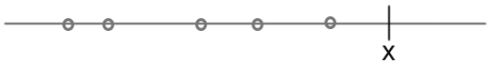
The support set for the minimum is the same as the support set for the individual X_i.

Please don't memorize this! Just work it out when you need it. It is quick and things will go a lot more smoothly for you in learning MathStat if you rely on understanding what you are doing rather than regurgitating a formula.

As per our discussion at the end of Section 0.8, if you are deriving the pdf for a minimum or maximum from scratch with cdfs, make sure to tack on an indicator at the end when reporting the final pdf!

Maximums:

Now let's derive an expression for the pdf of the maximum in terms of f and F. Consider the following 5 point "data set".



We can see that the maximum of a collection of random variables is less than or equal to x if and only if every random variable in the collection is less than or equal to x.

$$\begin{aligned} F_{X_{(n)}}(x) &= P(X_{(n)} \leq x) = P(\max(X_1, X_2, \ldots, X_n) \leq x) \\ &\overset{indep}{=} P(X_1 \leq x) \cdot P(X_2 \leq x) \cdots P(X_n \leq x) \\ &\overset{ident}{=} [P(X_1 \leq x)]^n \\ &= [F(x)]^n \end{aligned}$$

As with the minimum, this expression for the cdf of the maximum holds for both discrete and continuous random variables. In the case of discrete random variables, we can get the pdf from the cdf by taking differences as described in Section 0.5.4. In the continuous case, we can get a nice formula for the pdf by taking the derivative of the cdf.

$$\begin{aligned} f_{X_{(n)}}(x) &= \tfrac{d}{dx} F_{X_{(n)}}(x) \\ &= \tfrac{d}{dx} \{[F(x)]^n\} \\ &= n[F(x)]^{n-1} \cdot f(x) \end{aligned}$$

PDF for a Maximum

In summary, if $X_{(n)} = \max(X_1, X_2, \ldots, X_n)$ where $X_1, X_2, \ldots, X_n$ is a random sample from a continuous distribution with pdf f and cdf F, the pdf for $X_{(n)}$ is

$$f_{X_{(n)}}(x) = n[F(x)]^{n-1} \cdot f(x).$$

The support set for the maximum is the same as the support set for the individual X_i.

This is easy to derive when needed. Again, whether or not you use indicators, make sure to report the support of the pdf.

For fun, you might want to work out the pdf of the maximum of exponential random variables. The resulting pdf is not from a nice, known, "named" distribution that you will find in the table of distributions. Similarly, not all transformations of pdfs for random variables, such as in Section 1.2 will result in recognizable pdfs.

1.5 Important Tool IV: Moment Generating Functions (MGFs)

Definition 1.5.1

For a random variable X, the **moment generating function** (mgf), denoted by $M(t)$ or $M_X(t)$ if necessary, is defined as

$$M(t) = \mathsf{E}[e^{tX}].$$

If it becomes necessary to distinguish between two moment generating functions, for random variables X and Y, we will write $M_X(t)$ and $M_Y(t)$.

Moment generating functions are extremely useful tools for both computing "moments" of distributions which are the expectations $\mathsf{E}[X]$, $\mathsf{E}[X^2]$, $\mathsf{E}[X^3]$, ..., and for identifying distributions of some transformed random variables. Most notably, if you are asked to find the distribution of the "big fat sum" of iid random random variables $X_1, X_2, \ldots X_n$, you are surely going to want to try the moment generating function approach described in this section!

1.5.1 The Expectation of a Function of X

We know how to find the expectation $\mathsf{E}[X]$ but do we know how to find an expectation of a function of X such as $\mathsf{E}[g(X)]$? Before we can find any moment generating functions, we'll need to know how to do this! We begin with a simple discrete example.

Example 1.5.1

Suppose that X is a discrete random variable, taking on the values -1, 0, and 1, with probabilities

x	-1	0	1
$P(X = x)$	$4/12$	$3/12$	$5/12$

.

Now suppose that we want to find $\mathsf{E}[X^2]$. Note that this is the expected value of a new random variable

$Y := X^2$. Further, note that Y takes on possible values 0 and 1 and that Y has pdf given by

y	0	1
$P(Y = y)$	3/12	9/12

We now have

$$\mathsf{E}[X^2] = \mathsf{E}[Y] = \sum_{y=0}^{1} y \cdot P(Y = y) = 0 \cdot \frac{3}{12} + 1 \cdot \frac{9}{12} = \frac{9}{12}.$$

However, note that

$$\begin{aligned}
\mathsf{E}[X^2] &= \mathsf{E}[Y] = 0 \cdot \tfrac{3}{12} + 1 \cdot \tfrac{9}{12} \\
&= 0 \cdot \tfrac{3}{12} + 1 \cdot \left(\tfrac{4}{12} + \tfrac{5}{12}\right) \\
&= 0 \cdot \tfrac{3}{12} + 1 \cdot \tfrac{4}{12} + 1 \cdot \tfrac{5}{12} \\
&= (-1)^2 \cdot \tfrac{4}{12} + 0^2 \cdot \tfrac{3}{12} + 1^2 \cdot \tfrac{5}{12} \\
&= \sum_{x=-1}^{1} x^2 \cdot P(X = x)
\end{aligned}$$

Basically, what we are seeing here is the following.

Property

Let X be a random variable with pdf $f_x(x)$. Let $g(x)$ be some function.

- If X is discrete, we have have
$$\mathsf{E}[g(X)] = \sum_x g(x) f_X(x).$$
- If X is continuous, we have have
$$\mathsf{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx.$$

So, even though $g(X)$ is a new random variable, when finding $\mathsf{E}[g(X)]$ it is unnecessary to find a new pdf! That's kind of surprising if you think about it, but a lot of people don't think about it– they just do it sort of naturally without realizing how big of a deal it is. This is why this new formula for $\mathsf{E}[g(X)]$ (in either the discrete or continuous case) is known as the **Law of the Unconscious Statistician**!

We will now prove the Law of the Unconscious Statistician for $\mathsf{E}[g(X)]$ in the case where X is continuous and g is invertible and differentiable.

Suppose that X is continuous and g is a nice invertible and differentiable function. Define $Y = g(X)$. Then, since we know from Section 1.2 that $f_Y(y) = f_X(g^{-1}(y)) \left|\frac{d}{dy} g^{-1}(y)\right|$, we have that

$$\begin{aligned} \mathsf{E}[g(X)] &= \mathsf{E}[Y] = \int_{-\infty}^{\infty} y \cdot f_Y(y)\, dy \\ &= \int_{-\infty}^{\infty} y \cdot f_X(g^{-1}(y)) \left|\frac{d}{dy} g^{-1}(y)\right| \, dy \end{aligned}$$

Since g is invertible, we know that it is either strictly increasing or strictly decreasing. Let's consider each case separately.

Case One: g is strictly increasing

If g is strictly increasing then g^{-1} is as well. So, the derivative $\frac{d}{dy} g^{-1}(y)$ is always positive and we can drop the absolute value in the above integral. We now have

$$\mathsf{E}[g(X)] = \int_{-\infty}^{\infty} y \cdot f_X(g^{-1}(y)) \cdot \frac{d}{dy} g^{-1}(y)\, dy$$

Consider the change of variable $x = g^{-1}(y)$. Note that $dx = \frac{d}{dy} g^{-1}(y)\, dy$. Making the substitution, we have

$$\mathsf{E}[g(X)] = \int_{?}^{?'} g(x) \cdot f_X(x) dx$$

for some limits of integration $?$ and $?'$.

As for the limits, note that, even the original limits of integration went from $-\infty$ to ∞, they might actually have ended up being cut off by the support set for the pdf $f_Y(y)$. That is, the pdf $f_Y(y)$ may only be non-zero on some smaller interval. Furthermore, even if $y = g(x)$ does go from $-\infty$ to ∞, the inverse might not. For example, if $y = g(x) = \ln x$, then y goes from $-\infty$ to ∞, but $x = g^{-1}(y)$ only takes values from 0 to ∞. However, in this example,

$$\int_0^{\infty} g^{-1}(x)\, f_X(x)\, dx = \int_{-\infty}^{\infty} g^{-1}(x)\, f_X(x)\, dx$$

because the pdf for X will be zero outside of $(0, \infty)$.

So, when we make the substitution $x = g^{-1}(y)$, it will be safe to have limits all the way from $-\infty$ to ∞ since g^{-1} is increasing and since the pdf will cut off the limits in the appropriate places. That is, it is safe to use the limits $? = -\infty$ and $?' = \infty$ and we have

$$\mathsf{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx,$$

as desired.

Case Two: g is strictly decreasing

If g is strictly decreasing then g^{-1} is as well. So, the derivative $\frac{d}{dy} g^{-1}(y)$ is always negative and we have that

$|\frac{d}{dy}g^{-1}(y)| = -\frac{d}{dy}g^{-1}(y)$. Thus, we have that

$$\begin{aligned} \mathsf{E}[g(X)] &= \int_{-\infty}^{\infty} y \cdot f_X(g^{-1}(y)) \left|\frac{d}{dy}g^{-1}(y)\right| dy \\ &= -\int_{-\infty}^{\infty} y \cdot f_X(g^{-1}(y)) \cdot \frac{d}{dy}g^{-1}(y) \end{aligned}$$

Making the substitution $x = g^{-1}(y)$ and again noting that $dx = \frac{d}{dy}g^{-1}(y)$, we have that

$$\mathsf{E}[g(X)] = -\int_{?}^{?'} g(x) \cdot f_X(x)dx$$

for some limits of integration ? and ?′.

Since g^{-1} is a decreasing function, as y increases, $x = g^{-1}(y)$ decreases. As per the discussion in the increasing case, even if the random variable X takes on only a limited range of values, we can let x decrease from ∞ all the way to $-\infty$ and the pdf will cut us off appropriately. Thus,

$$\mathsf{E}[g(X)] = -\int_{\infty}^{-\infty} g(x) \cdot f_X(x)dx = \int_{-\infty}^{\infty} g(x) \cdot f_X(x),$$

as desired.

Now that we know how to compute the expectation of a function of a random variable, we can get back to moment generating functions!

1.5.2 Finding MGFs and the Poisson Distribution

Example 1.5.2 (Discrete)

A discrete random variable X is said to have the **Poisson distribution with parameter** λ if X has pdf.

$$f(x) = P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} I_{\{0,1,2,\ldots\}}(x).$$

We write $X \sim Poisson(\lambda)$.

Recall how we motivated the exponential distribution. We imagined events (arrivals at the door of a grocery store) occurring with two assumptions. They were that

1. the arrival/event rate is a constant λ per unit time, and
2. the number of arrivals/events in non-overlapping periods of time are independent.

Under these assumptions, the exponential distribution describes the amount of time between any two consecutive arrivals/events. One can show that the **number** of arrivals in one unit of time has the Poisson

distribution with parameter λ. (Again, we would direct you to a course in Markov chains or stochastic processes.)

So, suppose that $X \sim Poisson(\lambda)$. Let us find the moment generating function for X. Since this is a discrete random variable, the expectation is computed as a sum and not an integral.

$$\begin{aligned} M_X(t) &= \mathsf{E}[e^{tX}] = \sum_x e^{tx} \cdot P(X = x) \\ &= \sum_{x=0}^{\infty} e^{tx} \frac{e^{-\lambda}\lambda^x}{x!} \\ &= \sum_{x=0}^{\infty} \frac{e^{-\lambda}(\lambda e^t)^x}{x!} \end{aligned}$$

At this point, you could pull out the $e^{-\lambda}$ and recall the Taylor series expansion of e^x is

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}.$$

In our moment generating function expression, the k is actually x and the x is λe^t. So,

$$M_X(t) = e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!} = e^{-\lambda}\, e^{\lambda e^t} = e^{\lambda(e^t-1)}.$$

Alternatively, you could do what will become our "standard MathStat trick" that we usually use for integrals. The sum we need to compute looks almost like the sum over all values of the Poisson pdf except, in place of λ, we have λe^t. To make it into the Poisson pdf, we need the "$e^{-\lambda}$ part" which, in this case is $e^{-\lambda e^t}$. Putting this in (after pulling out the $e^{-\lambda}$) and adjusting for it by putting $e^{\lambda t}$ out front, we have

$$\begin{aligned} M_X(t) &= \textstyle\sum_{x=0}^{\infty} \frac{e^{-\lambda}(\lambda e^t)^x}{x!} \\ &= e^{\lambda(e^t-1)} \underbrace{\sum_{x=0}^{\infty} \frac{e^{-\lambda e^t}(\lambda e^t)^x}{x!}}_{\text{sum of Poisson pdf is 1}} \\ &= e^{\lambda(e^t-1)}. \end{aligned}$$

Take a moment to make sure you can find this in the moment generating column on the table of distributions!

Example 1.5.3 (Continuous)

Let $X \sim exp(rate = \lambda)$. Let us find the mgf of X.

$$\begin{aligned} M_X(t) &= \mathsf{E}[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx \\ &= \int_0^{\infty} e^{tx} \lambda e^{-\lambda x}\, dx \\ &= \int_0^{\infty} \lambda e^{-\lambda x + tx}\, dx \end{aligned}$$

To integrate, it is convenient to factor out the x in that exponent so we are integrating something like e^{cx}. However, by taking out the negative as well, we can see something that looks like the exponential pdf with $\lambda - t$ in place of λ.

$$M_X(t) = \int_0^{\infty} \lambda e^{-(\lambda - t)x}\, dx$$

Note that, in order for this integral to converge we must have $\lambda - t > 0$, otherwise the integrand is blowing up. (Alternatively, when comparing this to the exponential pdf, the usual parameter restriction $\lambda > 0$ has become $\lambda - t > 0$.) Either way you look at it, this mgf is only defined for $t < \lambda$.

So, for $t < \lambda$,

$$\begin{aligned} M_X(t) &= \int_0^{\infty} \lambda e^{-(\lambda - t)x}\, dx \\ &= \tfrac{\lambda}{\lambda - t} \int_0^{\infty} (\lambda - t) e^{-(\lambda - t)x}\, dx \\ &= \tfrac{\lambda}{\lambda - t} \cdot 1 = \tfrac{\lambda}{\lambda - t} \end{aligned}$$

since that integral is that of a pdf over its entire support.

Example 1.5.4 (Continuous)

Suppose that X has a normal distribution with mean μ and variance σ^2. The pdf for X is the classic "bell curve" given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}.$$

We write $X \sim N(\mu, \sigma^2)$.

There is no indicator because this pdf is defined for all $-\infty < x < \infty$. Because there is no "and zero

otherwise", there is no need to describe the support with an indicator. If we did, it would be $I_{(-\infty,\infty)}(x)$, which is always 1. For fun, why don't you multiply every function you write down today by 1!

In what follows, we will use the "exp" notation which is another way to write e to a power and is quite useful when one has superscripts on top of superscripts.

Let us find the mgf for X. For simplicity, we will start with $X \sim N(0,1)$. The moment generating function is given by

$$\begin{aligned} M_X(t) &= \mathsf{E}[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx \\ &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\, dx. \end{aligned}$$

Squaring that exponent out and bringing in the e^{tx}, we have

$$M_X(t) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\tfrac{1}{2\sigma^2}(x^2 - 2\mu x + \mu^2) + tx]\, dx.$$

We are going to try to bring the two x-terms together and to complete the square to get this back into a normal distribution form so that we can "integrate without integrating".

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\tfrac{\mu^2}{2\sigma^2}] \exp[-\tfrac{1}{2\sigma^2}(x^2 - 2\mu x) + tx]\, dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\tfrac{\mu^2}{2\sigma^2}] \exp[-\tfrac{1}{2\sigma^2}(x^2 - 2\mu x) - \tfrac{1}{2\sigma^2}(-2\sigma^2 tx)]\, dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\tfrac{\mu^2}{2\sigma^2}] \exp[-\tfrac{1}{2\sigma^2}(x^2 - 2(\mu + \sigma^2 t)x)]\, dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\tfrac{\mu^2}{2\sigma^2}] \exp[-\tfrac{1}{2\sigma^2}(x - (\mu + \sigma^2 t))^2 + \tfrac{1}{2\sigma^2}(\mu + \sigma^2 t)^2]\, dx \\ &= \exp[-\tfrac{\mu^2}{2\sigma^2}] \cdot \exp[\tfrac{1}{2\sigma^2}(\mu + \sigma^2 t)^2] \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\tfrac{1}{2\sigma^2}(x - (\mu + \sigma^2 t))^2]\, dx. \end{aligned}$$

Note that the integral is 1 because the integrand is the pdf of the normal distribution with mean $\mu + \sigma^2 t$ and variance σ^2. We are left with

$$M_X(t) = \exp[-\frac{\mu^2}{2\sigma^2}] \cdot \exp[\frac{1}{2\sigma^2}(\mu + \sigma^2 t)^2] = \exp[\mu t + \frac{1}{2}\sigma^2 t^2].$$

This holds for all t from $-\infty$ to ∞.

Why do we care about moment generating functions?

1.5.3 Finding Moments

For a random variable X, or its associated distribution, **"moments"** are expectations of the form $\mathsf{E}[X^k]$ for $k = 1, 2, 3, \ldots$. (Sometimes such moments are called "non-central moments", as opposed to the "central moments" $\mathsf{E}[(X-\mu)^k]$, where X is first "centered" around its mean $\mu = \mathsf{E}[X]$.)

Consider taking a derivative with respect to t of the moment generating function $M_X(t)$.

$$M_X(t) = \mathsf{E}[e^{tX}] \;\Rightarrow\; M'_X(t) = \frac{d}{dt}M_X(t) = \frac{d}{dt}\mathsf{E}[e^{tX}] \stackrel{?}{=} \mathsf{E}\left[\frac{d}{dt}e^{tX}\right] = \mathsf{E}[Xe^{tX}]$$

At that question mark, we pulled a derivative inside of an expectation. Since an expectation is a sum or integral, we really just exchanged a derivative and a sum or integral. In general, this may or may not be a valid thing to do. If the sum/integral converges and the sum/integral of the differentiated integrand also converges, they will converge to the same thing. In this text, things are mostly well behaved and we will make the exchange with reckless abandon!

If we plug $t = 0$ into the derivative of the moment generating function, we get

$$M'_X(0) = \mathsf{E}[Xe^0] = \mathsf{E}[X].$$

So, by differentiating the moment generating function and plugging in a 0 we get the first moment of this distribution.

With a second derivative, we have

$$M''_X(t) = \frac{d}{dt}M'_X(t) = \frac{d}{dt}\mathsf{E}[Xe^{tX}] = \mathsf{E}\left[\frac{d}{dt}Xe^{tX}\right] = \mathsf{E}[X^2e^{tX}].$$

So, plugging in $t = 0$ now gives us

$$M''_X(0) = \mathsf{E}[X^2e^0] = \mathsf{E}[X^2],$$

which is the second moment of the distribution.

Note

In general, for a random variable X with mgf $M_X(t)$, the kth moment is given by $M^{(k)}(0)$, where $M^{(k)}(t)$ is the kth derivative of $M_X(t)$ with respect to t.

Example 1.5.5

Let $X \sim Poisson(\lambda)$. One can show, using the definition of expectation and variance that $\mathsf{E}[X] = \lambda$ and $Var[X] = \lambda$. While the variance is a bit messy, you should try to show that $\mathsf{E}[X] = \lambda$ using our "summing without summing" trick of rewriting things so that you are ultimately summing a pdf over its entire support to get 1.

Let us now find the mean and variance for the Poisson distribution using moment generating functions. Recall that $M_X(t) = e^{\lambda(e^t-1)} = \exp[\lambda(e^t-1)]$.

So,

$$\begin{aligned} M'_X(t) &= \exp[\lambda(e^t-1)] \cdot \tfrac{d}{dt}[\lambda(e^t-1)] \\ &= \exp[\lambda(e^t-1)] \cdot \lambda e^t, \end{aligned}$$

which implies that

$$\mathsf{E}[X] = M'_X(0) = \lambda.$$

Furthermore,

$$\begin{aligned} M''_X(t) &= \exp[\lambda(e^t-1)] \cdot \lambda e^t + \lambda e^t \cdot \exp[\lambda(e^t-1)] \cdot \lambda e^t \\ &= \lambda e^t \cdot \exp[\lambda(e^t-1)] \cdot (1 + \lambda e^t) \end{aligned}$$

From this we get

$$M''_X(0) = \lambda(1+\lambda) = \lambda + \lambda^2.$$

So, we have that

$$Var[X] = \mathsf{E}[X^2] - (\mathsf{E}[X])^2 = \lambda + \lambda^2 - (\lambda)^2 = \lambda,$$

as expected.

1.5.4 The MGF Uniquely Identifies the Distribution

Here is an exciting claim.

A moment generating function for a random variable X uniquely determines its distribution!

For example, if you compute the mgf for some random variable X and get, for example, $\exp[\lambda(e^t - 1)]$, you know that X must be a $Poisson(\lambda)$ random variable!

This claim is arguably one of the most important reasons for learning about moment generating functions. **It is especially useful for finding the distribution of sums of independent random variables.**

We will see several examples of this in Section 1.5.6.

1.5.5 A Proof in a Simplified Setting

Suppose that we have two random variables X and Y with an infinite sequence of matching moments:

$$\mathsf{E}[X] = \mathsf{E}[Y], \qquad \mathsf{E}[X^2] = \mathsf{E}[Y^2], \qquad \mathsf{E}[X^3] = \mathsf{E}[Y^3], \qquad \text{etc ...}$$

Does this mean that X and Y must have the same distribution?

Unfortunately, the answer is no. A famous example is given by Heyde [2], where he gives the family of pdfs

$$f_a(x) = \frac{1}{x\sqrt{2\pi}}\, e^{-1/2(\ln x)^2}(1 + a\, \sin(2\pi\, \ln(x)))\, I_{(0,\infty)}(x)$$

indexed by

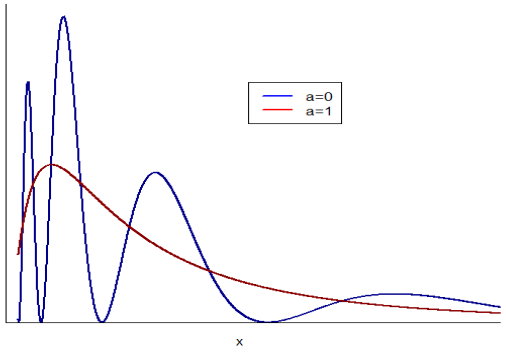
$$-1 \le a \le 1.$$

One can compute (though it is not fun), the moments

$$\mathsf{E}[X] = \sqrt{e}, \qquad \mathsf{E}[X^2] = e^2, \qquad \mathsf{E}[X^3] = e^{9/2},$$

$$\mathsf{E}[X^4] = e^8, \qquad \mathsf{E}[X^5] = e^{25/2}, \qquad \mathsf{E}[X^6] = e^{18}, \qquad \ldots$$

Note that these moments do not depend on the parameter a! (This continues for the infinite sequence.) Therefore we have (many!) different distributions, for example $f_{a_1}(x)$ and $f_{a_2}(x)$ where $a_1 \neq a_2$ with the same infinite sequence of moments. Two examples are shown here.



Now that we have this example out of the way. you might wonder whether if there is any moment criterion for identifying distributions that would ensure that two distributions are identical.

The answer is yes! If random variables X and Y both have moment generating functions $M_X(t)$ and $M_Y(t)$ that exist in some neighborhood of zero and if $M_X(t) = M_Y(t)$ for all t in this neighborhood, then X and Y have the same distributions!

We will prove the claim here in a simplified setting. The general proof of this can be found in Feller [1]. It is an inversion problem involving Laplace transform theory. (Did you notice that the mgf bears a striking resemblance to a Laplace transform?)

Proof of a special case:

Suppose that X and Y are random variables both taking only possible values in $\{0, 1, 2, \ldots, n\}$.

Further, suppose that X and Y have the same mgf for all t:

$$\sum_{x=0}^{n} e^{tx} f_X(x) = \sum_{y=0}^{n} e^{ty} f_Y(y).$$

For simplicity, will will let $s = e^t$ and we will define $c_x = f_X(x) - f_Y(x)$ for $x = 0, 1, \ldots, n$.

Now

$$\sum_{x=0}^{n} e^{tx} f_X(x) - \sum_{y=0}^{n} e^{ty} f_Y(y) = 0$$

$$\Downarrow$$

$$\sum_{x=0}^{n} s^{x} f_X(x) - \sum_{y=0}^{n} s^{y} f_Y(y) = 0$$

$$\Downarrow$$

$$\sum_{x=0}^{n} s^{x} f_X(x) - \sum_{x=0}^{n} s^{x} f_Y(x) = 0$$

$$\Downarrow$$

$$\sum_{x=0}^{n} s^{x} \left[f_X(x) - f_Y(x) \right] = 0$$

$$\Downarrow$$

$$\sum_{x=0}^{n} s^{x} c_x = 0 \qquad \forall\ s > 0$$

The above is simply a polynomial in s with coefficients $c_0, c_1, \ldots, c_n$. The only way it can be zero for all values

of s is if $c_0 = c_1 = \cdots = c_n = 0$.

So, we have that

$$0 = c_x = f_X(x) - f_Y(x) \qquad \text{for } x = 0, 1, \ldots, n.$$

Therefore

$$f_X(x) = f_Y(x) \qquad \text{for } x = 0, 1, \ldots, n.$$

In other words the density functions for X and Y are exactly the same. In other *other* words, X and Y have the same distributions!

□

1.5.6 Sums of iid Random Variables

Here is where the result of moment generating functions uniquely determining the distribution will really shine. Suppose that $X_1, X_2, \ldots, X_n$ are iid random variables from any distribution. Let $Y = \sum_{i=1}^n X_i$. Let's find the mgf for Y.

$$M_Y(t) \quad = \quad \mathsf{E}[e^{tY}] = \mathsf{E}\left[e^{t\sum X_i}\right] = \mathsf{E}\left[\prod_{i=1}^n e^{tX_i}\right]$$

$$\overset{indep}{=} \quad \prod_{i=1}^n \mathsf{E}[e^{tX_i}] \; = \prod_{i=1}^n M_{X_i}(t)$$

Note that we have not used the fact that the X_i are identically distributed. We have, however, just shown the following.

> If $X_1, X_2, \ldots, X_n$ be independent random variables. The sum
>
> $$Y = \sum_{i=1}^n X_i$$
>
> has moment generating function
>
> $$M_Y(t) = \prod_{i=1}^n M_{X_i}(t),$$
>
> where $M_{X_i}(t)$ is the moment generating function for X_i.

Now if the X_i are also identically distributed, they share the same pdf. Let's call this common pdf f. Consider the expectation

$$\mathsf{E}[e^{tX_i}] = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx.$$

(We are using continuous distribution notation, but what follows holds for expectations of discrete random variables as well.)

Note that the right-hand side, and hence the left-hand side does not depend on i. That is, $\mathsf{E}[e^{tX_i}]$ and $\mathsf{E}[e^{tX_j}]$ are going to be the same! In this case, the mgf for the sum $Y = \sum_{i=1}^{n} X_i$ is

$$M_Y(t) \overset{indep}{=} \prod_{i=1}^{n} M_{X_i}(t) \overset{ident}{=} [M_{X_1}(t)]^n .$$

This will be so useful that it deserves its own box!

> If $X_1, X_2, \ldots, X_n$ be iid random variables from a distribution with moment generating function $M_X(t)$. The sum $Y = \sum_{i=1}^{n} X_i$ has moment generating function
>
> $$M_Y(t) = [M_X(t)]^n .$$

Example 1.5.6

Suppose that $X_1, X_2, \ldots, X_n \overset{iid}{\sim} Poisson(\lambda)$. What is the distribution of $Y = \sum_{=1}^{n} X_i$?

Recall the Poisson moment generating function

$$M_{X_1}(t) = M_{X_5}(t) = \exp[\lambda(e^t - 1)].$$

For the sum we have

$$\begin{aligned} M_Y(t) &\overset{iid}{=} [M_{X_1}(t)]^n \\ &= \left(\exp[\lambda(e^t - 1)]\right)^n \\ &= \exp[n\lambda(e^t - 1)]. \end{aligned}$$

(Remember, if you raise something to a power and then to another power, you can multiply the two exponents!)

The moment generating function for Y is recognizable as that of the Poisson distribution with rate $n\lambda$. So, we can conclude that $Y \sim Poisson(n\lambda)$ since moment generating functions uniquely determine the distribution!

That was a lot easier than what would have been a generalized (to n dimensions) Jacobian approach to finding the distribution and even easier than the only slightly less cumbersome "conditioning approach" that one might see in a course on stochastic processes or Markov chains. For our next example, we consider one particular sum of random variables that we will see and use quite often.

Example 1.5.7

Let $X_1, X_2 \ldots . X_n$ be a random sample from the exponential distribution with rate λ. That is, let $X_1, X_2 \ldots . X_n \overset{iid}{\sim} exp(rate = \lambda)$. What is the distribution of $Y = \sum_{=1}^{n} X_i$?

Recall that the exponential moment generating function is

$$M_{X_1}(t) = \frac{\lambda}{\lambda - t}$$

for $t < \lambda$.

We now have

$$M_Y(t) = [M_X(t)]^n = \left(\frac{\lambda}{\lambda - t}\right)^n$$

for $t < \lambda$.

Comparing this to moment generating functions from the table of distributions in Appendix A, we see that

$$\boxed{Y \sim \Gamma(n, \lambda).}$$

The sum of iid exponential random variables has a gamma distribution. Using the notation from this text, the first parameter is the number of exponentials summed, and the second parameter matches the exponential rate parameter. This is going to be a super important result going forward.

As a final example, we will look at the sum of independent random variables that are not identically distributed.

Example 1.5.8

Let $X_1, X_2, \ldots, X_n$ be independent random variables with $X_i \sim Poisson(\lambda_i)$.

Note that these are not "iid" random variables. For this example, the random variables are not identically distributed. Although they all have a Poisson distribution, the changing λ's make them different Poisson distributions.

Let us find the distribution of $Y = \sum_{i=1}^{n} X_i$ using moment generating functions. While we have a "formula in a box", let's just derive it again as we go.

$$\begin{aligned} M_Y(t) &= \mathsf{E}[e^{tY}] = \mathsf{E}\left[e^{t\sum X_i}\right] = \mathsf{E}\left[\prod_{i=1}^{n} e^{tX_i}\right] \\ &\overset{indep}{=} \prod_{i=1}^{n} \mathsf{E}[e^{tX_i}] = \prod_{i=1}^{n} M_{X_i}(t) \end{aligned}$$

Note that the moment generating functions here are not identical, so we can not just take one and raise it to the nth power. Instead, we proceed by plugging in the individual mgfs.

$$\begin{aligned} M_Y(t) &= \prod_{i=1}^{n} M_{X_i}(t) \\ &= \prod_{i=1}^{n} \exp[\lambda_i(e^t - 1)] \\ &= \exp\left[(\sum \lambda_i)(e^t - 1)\right] \end{aligned}$$

Since this is the mgf of the Poisson distribution with rate $\sum_{i=1}^{n} \lambda_i$, we can conclude that

$$\boxed{Y \sim Poisson\left(\sum_{i=1}^{n} \lambda_i\right).}$$

This is consistent with our first example where $\lambda_i = \lambda$ for all $i = 1, 2, \ldots, n$. In that case, $\sum_{i=1}^{n} \lambda_i = \sum_{i=1}^{n} \lambda = n\lambda$.

1.6 Postscript: General Order Statistics

It feels strange to include this Section because we won't need it for anything going forward. It also feels strange not to include it along with the minimums and maximums in this Chapter. Feel free to skip this Section, as we would if we were teaching or learning Mathematical Statistics.

1.6.1 Introduction and Notation

Recall, from Section 1.4 that we have defined and denoted the order statistics for a random sample $X_1, X_2, \ldots, X_n$ as

$$\begin{aligned} X_{(1)} &= \min(X_1, X_2, \ldots, X_n) \\ X_{(2)} &= \text{the second smallest of the } X\text{'s} \\ \vdots & \quad \vdots \\ X_{(n)} &= \max(X_1, X_2, \ldots, X_{(n)}) \end{aligned}$$

We have already derived the distribution of the minimum and maximum in terms of the pdf and cdf for the individual X_i.

In this Section, we will derive the distributions for other order statistics and joint distributions for groups of order statistics. We will consider **continuous random variables only**. Imagine taking a random sample of size 15 from the geometric distribution with some fixed parameter p. The chances are very high that you will have some repeated values and not see 15 distinct values. For example, suppose we observe 7 distinct values. While it would make sense to talk about the minimum or maximum value here, it would not make sense to talk about the 12th largest value in this case. To further confuse the matter, the next sample might have a different number of distinct values! Any analysis of the order statistics for this discrete distribution would have to be well-defined in what would likely be an ad hoc way. (For example, one might define them conditional on the number of distinct values observed.)

1.6.2 The Joint Distribution of the Minimum and Maximum

Let's go for the joint cdf , which we define here, for the minimum and the maximum.

$$F_{X_{(1)},X_{(n)}}(x, y) = P(X_{(1)} \leq x, X_{(n)} \leq y)$$

It is not clear how to write this in terms of the individual X_i. Consider instead that the event $\{X_{(n)} \leq y\}$ can be written as the disjoint union

$$\{X_{(n)} \leq y\} = \{X_{(1)} \leq x, X_{(n)} \leq y\} \cup \{X_{(1)} > x, X_{(n)} \leq y\}.$$

Thus, we have

$$P(X_{(n)} \leq y) = P(X_{(1)} \leq x, X_{(n)} \leq y) + P(X_{(1)} > x, X_{(n)} \leq y). \tag{1.6.3}$$

We know how to write out the term on the left-hand side. The first term on the right-hand side is what we want to compute. As for the final term,

$$P(X_{(1)} > x, X_{(n)} \leq y),$$

note that this is zero if $x \geq y$. (In this case, $P(X_{(1)} \leq x, X_{(n)} \leq y) = P(X_{(n)} \leq y)$ and (1.6.3) gives us only $P(X_{(n)} \leq y) = P(X_{(n)} \leq y)$ which is both true and uninteresting!) So, we consider the case that $x < y$. Note

then that

$$\begin{aligned} P(X_{(1)} > x, X_{(n)} \le y) &= P(x < X_1 \le y, x < X_2 \le y, \ldots, x < X_n \le y) \\ &\overset{iid}{=} [P(x < X_1 \le y)]^n \\ &= [P(X_1 \le y) - P(X_1 \le x)]^n \\ &= [F(y) - F(x)]^n. \end{aligned}$$

Thus, from (1.6.3), we have that

$$\begin{aligned} F_{X_{(1)},X_{(n)}}(x,y) &= P(X_{(1)} \le x, X_{(n)} \le y) \\ &= P(X_{(n)} \le y) - P(X_{(1)} > x, X_{(n)} \le y) \\ &= [F(y)]^n - [F(y) - F(x)]^n \end{aligned}$$

for any $x < y$.

The joint pdf can be gotten by taking derivatives as

$$\begin{aligned} f_{X_{(1)},X_{(n)}}(x,y) &= \tfrac{d}{dx}\tfrac{d}{dy}\{[F(y)]^n - [F(y) - F(x)]^n\} \\ &= \tfrac{d}{dx}\{n[F(y)]^{n-1}f(y) - n[F(y) - F(x)]^{n-1}f(y)\} \\ &= n(n-1)[F(y) - F(x)]^{n-2}f(x)f(y), \end{aligned}$$

which holds in the case that $x < y$ and for x and y both in the support of the original distribution.

Example 1.6.1

Returning to a previous example of a sample of size 15 from the uniform distribution on $(0, 1)$, the joint pdf for the min and max is

$$f_{X_{(1)},X_{(n)}}(x,y) = 15 \cdot 14 \cdot [y - x]^{13} I_{(0,y)}(x)\, I_{(0,1)}(y).$$

The given product of indicators is one way to say that both x and y must be in $(0, 1)$ and such that $x < y$.

A Heuristic:

Since $X_1, X_2, \ldots, X_n$ are assumed to come from a continuous distribution, the min and max are also **continuous**

and **the joint pdf does not represent probability**– it is a surface under which volume represents probability. However, if we bend the rules and think of the joint pdf as probability, we can develop a heuristic method for remembering it.

Suppose (though it is not true) that

$$f_{X_{(1)},X_{(n)}}(x,y) = P(X_{(1)} = x, X_{(n)} = y).$$

This would mean that we need one value in the sample $X_1, X_2, \ldots, X_n$ to fall at x, one value to fall at y, and the remaining $n-2$ values to fall in between.

The "probability" one of the X_i is x is "like" $f(x)$. (Remember, we are bending the rules here in order to develop a heuristic. This probability is, of course, actually 0 for a continuous random variable.)

The "probability" one of the X_i is y is "like" $f(y)$.

The probability that one of the X_i is in between x and y is (actually) $F(y) - F(x)$.

The sample can fall many ways to give us a minimum at x and a maximum at y. For example, imagine that $n = 5$. We might get $X_3 = x$, $X_1 = y$ and the remaining X_2, X_4, X_5 in between x and y.

This would happen with "probability"

$$f(x)[F(y) - F(x)]^3 f(y).$$

Another possibility is that we get $X_5 = x$ and $X_2 = y$ and the remaining X_1, X_3, X_4 in between x and y.

This would also happen with "probability"

$$f(x)[F(y) - F(x)]^3 f(y).$$

We have to add this "probability" up as many times as there are scenarios. So, let's count them. There are $5!$ different ways to lay down the X_i. For each one, there are $3!$ different ways to lay down the remaining values in between that will result in the same min and max. So, we need to divide these redundancies out for a total of $5!/3! = (5)(4)$ ways to get that min at x and max at y.

In general, for a sample of size n, there are $n!$ different ways to lay down the X_i. For each one, there are $(n-2)!$ different ways that result in the same min and max. So, there are a total of $n!/(n-2)! = n(n-1)$ ways to get that

Thus, the "probability" of getting a minimum of x and a maximum of y is

$$n(n-1)f(x)[F(y) - F(x)]^{n-2} f(y),$$

which looks an awful lot like the formula we derived above!

1.6.3 The Joint Distribution for All Order Statistics

We wish now to find the pdf

$$f_{X_{(1)},X_{(2)},\ldots,X_{(n)}}(x_1, x_2, \ldots, x_n).$$

This time, we will start with the heuristic aid.

Suppose that $n = 3$ and we want to find

$$f_{X_{(1)},X_{(2)},X_{(3)}}(x_1, x_2, x_3) \quad \text{“=”} \quad P(X_{(1)} = x_1, X_{(2)} = x_2, X_{(3)} = x_3).$$

The quotes around the equals sign is there to say that the joint pdf is not really a joint probability and that we are just kind of thinking about it that way. For continuous random variables, the probability on the right-hand side will always be zero. (We feel compelled to point this out yet again as it would be quite an egregious error to think that the left-hand side is equal to the right-hand side!)

The first thing to notice is that this probability will be 0 if we don't have $x_1 < x_2 < x_3$. (Note that we use strict inequalities here. For a continuous distribution, we will never see repeated values so the minimum and second smallest, for example, could not take on the same value.)

Fix values $x_1 < x_2 < x_3$. How could a sample of size 3 fall so that the minimum is x_1, the next smallest is x_2, and the largest is x_3? We could observe

$$X_1 = x_1, X_2 = x_2, X_3 = x_3,$$

or

$$X_1 = x_2, X_2 = x_1, X_3 = x_3,$$

or

$$X_2 = x_2, X_2 = x_3, X_3 = x_1,$$

as well as other orderings. There are 3! possibilities to list and the “probability” for each of these disjoint events

is $f(x_1)f(x_2)f(x_3)$. Thus,

$$\begin{aligned} f_{X_{(1)},X_{(2)},X_{(3)}}(x_1,x_2,x_3) \quad &\text{“=”} \quad P(X_{(1)}=x_1, X_{(2)}=x_2, X_{(3)}=x_3) \\ &= \quad P(X_1=x_1, X_2=x_2, X_3=x_3) \\ &\quad\; +P(X_1=x_2, X_2=x_1, X_3=x_3) \\ &\quad\; \vdots \\ &\quad\; +P(X_1=x_3, X_2=x_2, X_3=x_1) \\ &= \quad 3!f(x_1)f(x_2)f(x_3) \end{aligned}$$

For general n, we have

$$\begin{aligned} f_{X_{(1)},X_{(2)},\ldots,X_{(n)}}(x_1,x_2,\ldots,x_n) \quad &\text{“=”} \quad P(X_{(1)}=x_1, X_{(2)}=x_2, \ldots X_{(n)}=x_n) \\ &= \quad n!f(x_1)f(x_2)\cdots f(x_n) \end{aligned}$$

which holds for $x_1 < x_2 < \cdots < x_n$ with all x_i in the support for the original distribution. The joint pdf is zero otherwise.

The Formalities:

The joint cdf,

$$P(X_{(1)} \le x_1, X_{(2)} \le x_2, \ldots, X_{(n)} \le x_n),$$

is difficult to rewrite in terms of the individual and unordered X_i.

Instead, we will consider the quantity

$$P(y_1 < X_{(1)} \le x_1, y_2 < X_{(2)} \le x_2, \ldots, y_n < X_{(n)} < x_n)$$

for values $y_1 < x_1 \le y_2 < x_2 \le y_3 < x_3 \le \cdots \le y_n < x_n$.

The event inside the probability can happen if

$$y_1 < X_1 \le x_1,\ y_2 < X_2 \le x_2,\ \ldots,\ y_n < X_n < x_n,$$

or if

$$y_1 < X_5 \le x_1,\ y_2 < X_3 \le x_2,\ \ldots,\ y_n < X_{n-2} < x_n,$$

or in many other ways. There are a total of $n!$ different ways to interchange the X_i here.

Because of the constraints on the x_i and y_i, these are disjoint events. So, we can add these $n!$ probabilities,

which will all be the same, together to get

$$P(y_1 < X_{(1)} \leq x_1, \ldots, y_n < X_{(n)} < x_n) = n!\, P(y_1 < X_1 \leq x_1, \ldots, y_n < X_n < x_n).$$

Note that

$$P(y_1 < X_1 \leq x_1, \ldots, y_n < X_n < x_n) \overset{indep}{=} \prod_{i=1}^{n} P(y_i < X_i \leq x_i) = \prod_{i=1}^{n} [F(x_i) - F(y_i)].$$

So,

$$P(y_1 < X_{(1)} \leq x_1, \ldots, y_n < X_{(n)} < x_n) = n! \prod_{i=1}^{n} [F(x_i) - F(y_i)] \tag{1.6.4}$$

The left-hand side can be written in terms of the joint pdf for all of the order statistics as

$$\int_{y_n}^{x_n} \int_{y_{n-1}}^{x_{n-1}} \cdots \int_{y_1}^{x_1} f_{X_{(1)}, X_{(2)}, \ldots, X_{(n)}}(u_1, u_2, \ldots, u_n)\, du_1 du_2 \ldots, du_n. \tag{1.6.5}$$

Taking derivatives $\frac{d}{dx_1} \frac{d}{dx_2} \cdots \frac{d}{dx_n}$ gives

$$f_{X_{(1)}, X_{(2)}, \ldots, X_{(n)}}(x_1, x_2, \ldots, x_n)$$

on the left-hand side.

Differentiating both sides of (1.6.5) with respect to $x_1, x_2, \ldots, x_n$ gives us

$$f_{X_{(1)}, X_{(2)}, \ldots, X_{(n)}}(x_1, x_2, \ldots, x_n) = n! f(x_1) f(x_2) \cdots f(x_n)$$

which holds for $x_1 < x_2 < \cdots, x_n$ and all x_i in the support of the original distribution. The pdf is zero otherwise.

1.6.4 The Distribution of $X_{(i)}$

We can get the marginal pdf for the ith order statistic $X_{(i)}$, by taking the joint pdf for all order statistics from Section 1.6.3 and integrating out the unwanted x_j.

Let's start by integrating out x_1. (This is, of course, assuming that $i \neq 1$.) Since the support of the joint pdf for

the order statistics includes the constraint $x_1 < x_2 < \cdots < x_n$, limits of integration are $-\infty$ to x_2.

$$\begin{aligned} f_{X_{(2)},\ldots,X_{(n)}}(x_2,\ldots,x_n) &= \int_{-\infty}^{x_2} f_{X_{(1)},X_{(2)},\ldots,X_{(n)}}(x_1,x_2,\ldots,x_n)\,dx_1 \\ &= \int_{-\infty}^{x_2} n!f(x_1)f(x_2)\cdots f(x_n)\,dx_1 \\ &= n!f(x_2)\cdots f(x_n)\int_{-\infty}^{x_2} f(x_1)\,dx_1 \\ &= n!f(x_2)\cdots f(x_n)F(x_2) \end{aligned}$$

for $x_2 < x_3 < \cdots < x_n$.

Now let's integrate out x_2 which goes from $-\infty$ to x_3.

$$\begin{aligned} f_{X_{(3)},\ldots,X_{(n)}}(x_3,\ldots,x_n) &= \int_{-\infty}^{x_3} f_{X_{(2)},\ldots,X_{(n)}}(x_2,\ldots,x_n)\,dx_2 \\ &= n!f(x_3)\cdots f(x_n)\int_{-\infty}^{x_3} \underbrace{F(x_2)}_{u}\underbrace{f(x_2)\,dx_2}_{du} \\ &= n!f(x_3)\cdots f(x_n)\ \tfrac{1}{2}[F(x_2)]^2\Big|_{x_2=-\infty}^{x_2=x_3} \\ &= n!f(x_3)\cdots f(x_n)\tfrac{1}{2}([F(x_3)]^2 - [\underbrace{F(-\infty)}_{0}]^2) \\ &= \tfrac{n!}{2}f(x_3)\cdots f(x_n)[F(x_3)]^2 \end{aligned}$$

which holds for $x_3 < x_4 < \cdots < x_n$.

The next time through, we will integrate out x_3 from $-\infty$ to x_4. Using $u = F(x_3)$ and $du = f(x_3)\,dx_3$, we get

$$f_{X_{(4)},\ldots,X_{(n)}}(x_4,\ldots,x_n) = \frac{n!}{(3)(2)}f(x_4)\cdots f(x_n)[F(x_4)]^3.$$

Continue until we reach $X_{(i)}$ to get

$$f_{X_{(i)},\ldots,X_{(n)}}(x_i,\ldots,x_n) = \frac{n!}{(i-1)!}f(x_i)\cdots f(x_n)[F(x_i)]^{i-1}$$

which holds for $x_i < x_{i+1} < \cdots < x_n$.

Now, we start integrating off x's from the other side.

$$\begin{aligned} f_{X_{(i)},\ldots,X_{(n-1)}}(x_i,\ldots,x_{n-1}) &= \int_{x_{n-1}}^{\infty} f_{X_{(i)},\ldots,X_{(n-1)}}(x_i,\ldots,x_n)\,dx_n \\ &= \tfrac{n!}{(i-1)!} f(x_i)\cdots f(x_{n-1})[F(x_i)]^{i-1}\int_{x_{n-1}}^{\infty} f(x_n)\,dx_n \\ &= \tfrac{n!}{(i-1)!} f(x_i)\cdots f(x_{n-1})[F(x_i)]^{i-1}[1-F(x_{n-1})] \end{aligned}$$

for $x_i < x_{i+1} < \cdots, x_{n-1}$.

$$\begin{aligned} f_{X_{(i)},\ldots,X_{(n-2)}}(x_i,\ldots,x_{n-2}) &= \int_{x_{n-2}}^{\infty} f_{X_{(i)},\ldots,X_{(n-1)}}(x_i,\ldots,x_{n-1})\,dx_{n-1} \\ &= \tfrac{n!}{(i-1)!} f(x_i)\cdots f(x_{n-2})[F(x_i)]^{i-1}\int_{x_{n-2}}^{\infty} f(x_{n-1})[1-F(x_{n-1})]\,dx_{n-1} \end{aligned}$$

Letting $u = 1 - F(x_{n-1})$ and $du = -f(x_{n-1})\,dx_{n-1}$, we get

$$\begin{aligned} f_{X_{(i)},\ldots,X_{(n-2)}}(x_i,\ldots,x_{n-2}) &= \tfrac{n!}{(i-1)!} f(x_i)\cdots f(x_{n-2})[F(x_i)]^{i-1}\left\{-\tfrac{1}{2}[1-F(x_{n-1})]^2\right\}_{x_{n-1}=x_{n-2}}^{x_{n-1}=\infty} \\ &= \tfrac{n!}{2(i-1)!} f(x_i)\cdots f(x_{n-2})[F(x_i)]^{i-1}[1-F(x_{n-2})]^2 \end{aligned}$$

for $x_i < x_{i+1}, \cdots < x_{n-2}$.

The next time through we will integrate out x_{n-2} from x_{n-3} to ∞. Note that

$$\begin{aligned} \int_{x_{n-3}}^{\infty} f(x_{n-2})\underbrace{[1-F(x_{n-2})]}_{u}{}^2\,dx_{n-2} &= -\tfrac{1}{3}[1-F(x_{n-2})]^3\Big|_{x_{n-2}=x_{n-3}}^{x_{n-2}=\infty} \\ &= \tfrac{1}{3}[1-F(x_{n-3})]^3. \end{aligned}$$

Thus,

$$f_{X_{(i)},\ldots,X_{(n-3)}}(x_i,\ldots,x_{n-3}) = \frac{n!}{(3)(2)(i-1)!} f(x_i)\cdots f(x_{n-3})[F(x_i)]^{i-1}[1-F(x_{n-3})]^3$$

for $x_i < x_{i+1} < \cdots < x_{n-3}$.

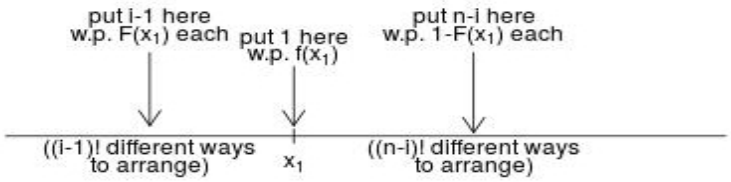
Continuing all the way down to the marginal pdf for $X_{(i)}$ alone, we get

$$f_{X_{(i)}} = \frac{n!}{(n-i)!(i-1)!}[F(x_i)]^{i-1} f(x_i)[1-F(x_{n-i})]^{n-i}$$

for $-\infty < x_i < \infty$. ($\leftarrow$ Note that this may be further restricted by indicators in $f(x_i)$.)

The Heuristic:

We once again will think of the continuous random variables $X_1, X_2, \ldots, X_n$ as discrete and $f_{X_{(i)}}(x_i)$ as the "probability" that the ith order statistic is at x_i. First note that there are $n!$ different ways to arrange the $x's$. We need to put 1 at x_i, which will happen with "probability" $f(x_i)$. We need to put $i-1$ below x_i, which will happen with probability $[F(x_i)]^{i-1}$ and we need to put $n-i$ above x_i, which will happen with probability $[1-F(x_i)]^{n-i}$. There are $(i-1)!$ different ways to arrange the x's chosen to go below x_i. These arrangements are redundant and need to be divided out. Hence, we have $(i-1)!$ in the denominator. There are $(n-i)!$ different ways to arrange the x's chosen to go above x_i. These arrangements are also redundant and need to be divided out. Thus, we also have $(n-i)!$ in the denominator.



1.6.5 The Joint Distribution of $X_{(i)}$ and $X_{(j)}$ for $i < j$

As in Section Section 1.6.4, one could start with the joint pdf for all of the order statistics and integrate out the unwanted ones. The result will be

$$f_{X_{(i)},X_{(j)}}(x_i, x_j) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!}[F(x_i)]^{i-1} f(x_i)[F(x_j)-F(x_i)]^{j-i-1} f(x_j)[1-F(x_j)]^{n-j}$$

for $-\infty < x_i < x_j < \infty$.

Can you convince yourself of this heuristically?

Chapter 1 Exercises

1. Consider a sequence of independent trials where each trial can result in one of two possible outcomes: success or failure. Suppose that the probability of success on any one trial is p for some $0 \leq p \leq 1$. Recall that, in Chapter 0, we defined the geometric random variable as counting the number of trials until the first success.

   Now, let X be the number of trials until the rth success where r is some positive integer. Then X is said to be a "negative binomial" random variable. Just as we had two different types of geometric distributions,

we may define two different types of negative binomial distributions.

(a). Find the correct pdf on your table of distributions. Then argue how to derive this pdf from the scenario described above.

(b). Find the distribution of $Y = X - r$. (Name it!) What is the interpretation of Y in terms of the success/failure experiment?

2. Let g be a continuous, invertible, and increasing function. Show that g^{-1} is also increasing.
3. Let $X \sim unif(0,1)$, find the pdf of $Y = e^X$.
4. Let $X \sim \exp(\text{rate} = \lambda)$. Find the distribution of $Y = e^{-X}$. (Name it!)
5. Let X be a random variable with the gamma distribution with parameters α and β, (ie: $X \sim \Gamma(\alpha, \beta)$).

(a). Let $c > 0$ be a constant. Find the distribution of $Y = cX$. (Name it!)

(b). Now consider a random sample, $X_1, X_2, \ldots, X_n$, gamma distribution with parameters α and β, (ie: $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \Gamma(\alpha, \beta)$). Find the distribution of $Y = \sum_{i=1}^n X_i$. (Name it!)

6. Let n be a positive integer. Suppose that $X \sim \Gamma(n/2, 1/2)$. This special gamma distribution is known as the **χ^2-squared distribution**. The parameter n is known as the **degrees of freedom** for this distribution and we write $X \sim \chi^2(n)$.

   Suppose that $X \sim N(0,1)$. (See the beginning of Chapter 3 if you do not know what this means.)

   Use moment generating functions to find the distribution of $Y := X^2$. (Name it!)
7. Let $X_1, X_2, \ldots, X_k$ be independent random variables with $X_i \sim \chi^2(n_i)$ for some positive integers $n_1, n_2, \ldots, n_k$. What is the distribution of $Y := \sum_{i=1}^n X_i$?
8. Suppose that X is a continuous random variable with pdf

$$f(x) = \frac{\gamma}{(1+x)^{\gamma+1}} I_{(0,\infty)}(x)$$

   for some parameter $\gamma > 0$.

   Then X is said to have a **Pareto distribution** with parameter γ.

   We write $X \sim Pareto(\gamma)$.

   Find the distribution of $Y = \ln(X+1)$. (Name it!)
9. Suppose that X has a Beta distribution, $X \sim Beta(a, b)$.

(a). Find the pdf for $Y = -ln\, X$.

(b). If a is a positive integer and if $b = 1$, what is the distribution of Y. (Name it!)

10. Let X have the uniform distribution over the interval $(-\pi/2, \pi/2)$. Show that $Y = tan\, X$ has a Cauchy distribution.
11. Let $X \sim unif(0,1)$.

(a). Find the distribution of $Y = -\ln X$. (Name it!)

(b). Find a transformation $Y = g(X)$ so that $Y \sim \exp(rate = \lambda)$.

12. Let X_1 and X_2 have the joint pdf

$$f_{X_1,X_2}(x_1,x_2) = 2e^{-x_1-x_2} I_{(0,x_2)}(x_1) I_{(0,\infty)}(x_2).$$

Find the joint pdf of $Y_1 = 2X_1$ and $Y_2 = X_2 - X_1$, and argue that Y_1 and Y_2 are independent.

13. Let $X_1, X_2 \sim iid\ exp(rate = 1)$. What is the pdf for X_1/X_2? (The name of the resulting distribution is the "F distibution" which is defined, more generally, in terms of chi-squared random variables.)

14. Suppose that $X_1, X_2 \overset{iid}{\sim} N(0,1)$. Find the distribution of X_1/X_2. (Name it!)

15. Let $X_1, X_2 \overset{iid}{\sim} \exp(\text{rate} = \lambda)$. Find the distribution of $Y = \frac{X_1}{X_1+X_2}$. (Name it!)

16. Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from the uniform distribution over the interval $(0,1)$.
    (a). Find the distribution of $\min(X_1, X_2, \ldots, X_n)$. (Name it!)
    (b). Find the distribution of $\max(X_1, X_2, \ldots, X_n)$. (Name it!)

17. Let X and Y be independent random variables with cdfs F_X and F_y, respectively.
    (a). Let $Z = \max(X,Y)$. Find the cdf for Z in terms of the cdfs for X and Y.
    (b). Let $Z = \min(X,Y)$. Find the cdf for Z in terms of the cdfs for X and Y.

18. Suppose that X is a continuous random variable with pdf $f(x)$. Let $Y = X^2$. (Note that this is not a one-to-one, invertible transformation.)

    Find an expression for the pdf of Y in terms of the pdf of X.

19. Suppose that X_1 and X_2 are independent random variables and that $Y_1 = g_1(X_1)$ and $Y_2 = g_2(X_2)$. Then Y_1 and Y_2 are independent. Sounds reasonable yes?

    Prove this in the case that X_1 and X_2 are continuous and g_1 and g_2 are invertible.

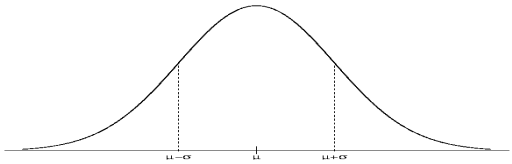
20. Let $X_1, X_2, \ldots, X_n$ be a random sample from a continuous distribution with pdf f and cdf F. Find the pdf for the *sample range* $X_{(n)} - X_{(1)}$.

Chapter 2 Qualities of Estimators: Defining Good, Better, and Best

If you knew any continuous distribution before coming into this course, it was probably the **normal distribution**, also known as the **Gaussian distribution**. It's pdf is the "bell curve" that we all know and ~~overuse~~ love. It is parameterized by the mean μ and variance σ^2 for X, and we write $X \sim N(\mu, \sigma^2)$. The pdf is

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2},$$

which looks like this.



If you do a little Calculus, you'll see that the points of inflection, where the pdf switches concavity, are precisely σ units away from the mean μ. That is, they are 1 standard deviation away from the mean.

The normal distribution with mean 0 and variance 1 is called the **standard normal distribution.** It is common for people to use a Z, instead of an X for a standard normal distribution, though an X is fine. Conversely, seeing a Z in a statistical problem does not guarantee a $N(0, 1)$ distribution.

Estimation:

Let $X_1, X_2, \ldots, X_n$ be a random sample from the $N(\mu, 1)$ distribution. If n is large and if you make a histogram of your sampled values, you should see the histogram roughly matching up with the pdf for the $N(\mu, 1)$ distribution. In particular, you will see the majority of sampled values towards the center (which is μ) and not so many out "in the tails'."

Question: How might you estimate the mean μ?

Answer: The most "natural" answer would be to use the **sample mean**:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

The true mean μ for the distribution is a probability weighted average. The sample mean is the average of the sampled values and it already has the "probability weighted" part "built in" since the values are falling mostly in the high probability regions and not so much in the low probability regions.

We will use the notation $\widehat{\mu}$ to denote an estimator for μ. So, we write

$$\widehat{\mu} = \overline{X}.$$

Note that the X in "X bar" is capitalized. At this point we are considering it as a random variable. We are **going**, in the **future**, to sample values $X_1, X_2, \ldots, X_n$ from this normal distribution and then we are **going** to average them. Once we do, we will have observations $x_1, x_2, \ldots, x_n$ which have sample mean $\overline{x}$. While μ is an unknown constant, $\widehat{\mu}$ is a random variable. It does not make sense to say that $\mu = \overline{X}$ but it does makes sense to say $\widehat{\mu} = \overline{X}$. The word "**estimator**" is used to refer to the random variable $\overline{X}$ while the word "**estimate**" is used to refer to the actual observation $\overline{x}$.

Estimating the mean μ with the sample mean $\overline{X}$ seems like the "obvious" thing to do. However, the following questions arise.

1. How "good" is this estimator? What does "good" even mean?
2. Can we do better?
3. How would you go about estimating parameters with a less obvious interpretation, such as the α from the $\Gamma(\alpha, \beta)$ distribution?

Before we can propose good estimators for parameters of various distributions, we need to be able to talk about several measures of the quality of an estimator so that we know what to aim for.

2.1 Notation, Statistics, and Unbiasedness

We will use θ to denote a generic parameter. It might represent μ in a normal distribution, λ in the exponential distribution, β in a gamma distribution, or a number of other things. It might even represent a vector of parameters such as $\theta = (\mu, \sigma^2)$ or $\theta = (\alpha, \beta)$.

From now on, we really want to emphasize that our pdfs have parameters. So, instead of writing $f(x)$, we will write $f(x; \theta)$. The semicolon will be used to separate variables and parameters. For example, when referring

to a joint pdf for $X_1, X_2, \ldots, X_n$ with parameters α and β, we might write $f(x_1, x_2, \ldots, x_n; \alpha, \beta)$.

Let $X_1, X_2, \ldots, X_2$ be a random sample from the distribution with pdf $f(x; \theta)$.

The Goal:

The goal is to estimate the parameter θ, or some function of the parameter $\tau(\theta)$[1], based on what we see in the random sample $X_1, X_2, \ldots, X_n$.

Our estimator will be a **statistic** which is defined as some function of the data. A statistic is a random variable that we will denote

$$T = t(X_1, X_2, \ldots, X_n) = t(\vec{X}).$$

The statistic may be one-dimensional such as

$$T = t(\vec{X}) = \overline{X},$$

or it may be higher dimensional such as

$$T = t(\vec{X}) = \left(\overline{X}, \sum_{i=1}^{n} X_i^2, X_{(1)}\right).$$

In general, we can have $t : \mathbb{R}^n \to \mathbb{R}^k$, but for now we will assume that $k = 1$. Similarly, we will assume for now that θ and $\tau(\theta)$ are one-dimensional.

When estimating $\tau(\theta)$ using the statistic $T = t(\vec{X})$, one thing we might want is for the estimator to be "on average" correct.

Definition 2.1.1

We say that T is an **unbiased** estimator for $\tau(\theta)$ if

$$\mathsf{E}[T] = \tau(\theta).$$

If you are not completely comfortable with properties of expectations and with computing variances, now would be a good time to go back to review Section 0.10.

[1] An example would be the case where you wanted to estimate $\tau(\theta) = \theta^2$. Should you estimate θ and square the result? Not necessarily! The answer depends on what properties you want your estimator to have.

Example 2.1.1

Let $X_1, X_2, \ldots, X_n$ be a random sample from any distribution with mean μ and variance σ^2.

(Note that μ and σ^2 denote a generic mean and generic variance. We are not necessarily talking about the normal distribution, for example, where the symbols μ and σ^2 are already used. Suppose that the random sample is from the $\Gamma(\alpha, \beta)$ distribution. Then $\mu = \alpha/\beta$ and $\sigma^2 = \alpha/\beta^2$.)

Suppose that we want to estimate the mean μ with the sample mean $\overline{X}$. This is a new random variable with its own distribution and own properties! It's mean is

$$\begin{aligned} \mu_{\overline{X}} &:= \mathsf{E}[\overline{X}] = \mathsf{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\ &= \tfrac{1}{n}\sum_{i=1}^{n} \mathsf{E}[X_i] \overset{ident}{=} \frac{1}{n}\sum_{i=1}^{n} \mu \\ &= \tfrac{1}{n} n\mu = \mu \end{aligned}$$

This is true for the sample mean for a random sample from any distribution!

Property

Let $X_1, X_2, \ldots, X_n$ be a random sample from any distribution with mean μ. Then

$$\mathsf{E}[\overline{X}] = \mu.$$

Now what can we say about the variance of $\overline{X}$ in terms of the original variance for the distribution? Variance is a measure of spread. We see that, when using $\overline{X}$ to estimate μ, on average we will be correct. However, if this estimator has a large variance, we will see wild swings between values if we take many samples. Most of the time you only get one sample to work with, so this "high variability" isn't good because it means that your one sample mean can potentially be quite far away from μ. For example, suppose the true (but unknown to you) mean is $\mu = 15.3$. Suppose you estimate it with some estimator $\widehat{\mu}$. Would you rather this estimator produce values like -117.8, 124.29, 8.55, and so on, that, when themselves averaged give you 15.3 or would you rather it produced values like 15.1, 15.8, 15.4? Remember, you only get to see one estimate so, even though the first

estimator gives you the correct value on average, the second estimator is much more desirable! In fact, you might actually prefer a little "bias" in your estimator if it is "almost unbiased" and has small variance.

Example 2.1.2

Let $X_1, X_2, \ldots, X_n$ be a random sample from any distribution with mean μ and variance σ^2.

Suppose, as in the previous example, we chose to estimate μ with $\overline{X}$. What is the variance of this estimator?

It is

$$\begin{aligned} \sigma^2_{\overline{X}} &= Var[\overline{X}] = Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n^2} Var\left[\sum_{i=1}^{n} X_i\right] \overset{indep}{=} \frac{1}{n^2}\sum_{i=1}^{n} Var[X_i] \\ &\overset{ident}{=} \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{1}{n^2}\, n\sigma^2 = \frac{\sigma^2}{n} \end{aligned}$$

So, no matter what distribution we are considering, $\overline{X}$ is always an unbiased estimator for the mean μ of the distribution. Also, the variability of $\overline{X}$ is smaller than the variability of a single X_i. For example, suppose that $2,000$ Calculus I students are divided into 50 recitation sections, each of size 40. If we look at the Midterm 1 scores for all $2,000$ students, they will likely vary quite a bit between, potentially, single digit scores and perfect scores. However, if we look at the 50 sample mean scores for each recitation section, we wouldn't expect to see such variability!

Note that we have derived an expression for the variance of the sample mean for a random sample from any distribution!

Property

Let $X_1, X_2, \ldots, X_n$ be a random sample from any distribution with variance σ^2. Then

$$Var[\overline{X}] = \frac{Var[X_1]}{n} = \frac{\sigma^2}{n}.$$

2.2 Mean Squared Error and Bias

We have produced some estimators for the mean (μ) and variance (σ^2) What about other parameters like the p in the binomial distribution of the β in the gamma distribution?

Let θ be a generic parameter. We will denote an estimator of θ by $\widehat{\theta}$. We know that if $\hat{\theta}$ is an unbiased estimator of θ then $\mathsf{E}[\hat{\theta}] = \theta$.

Definition 2.2.1

The **bias** of an estimator $\hat{\theta}$ of θ is denoted and defined as

$$B(\hat{\theta}) = \mathsf{E}[\hat{\theta}] - \theta.$$

If $\widehat{\theta}$ is an unbiased estimator of θ, its bias is zero.

The variance of a random variable X is a measure of spread about its mean $\mu = \mathsf{E}[X]$. In particular, it is

$$Var[X] = \mathsf{E}[(X - \mu)^2] = \mathsf{E}[(X - \mathsf{E}[X])^2].$$

For an unbiased estimator $\widehat{\theta}$ (which is a random variable) of θ (which is a constant), realizations of $\widehat{\theta}$ will be, on average, θ with repeated sampling. Again, in the case of one sample, this is not so useful if realizations of $\widehat{\theta}$ have really wide swings to the left and right of the true θ. We will only get to see one value and it may be way off. Thus, we would like the variance

$$Var[\widehat{\theta}] = \mathsf{E}[(\hat{\theta} - \mathsf{E}[\hat{\theta}])^2] \overset{unbiased}{=} \mathsf{E}[(\hat{\theta} - \theta)^2]$$

to be small so that this wild swinging is unlikely to happen.

In the case that $\widehat{\theta}$ is a biased estimator of θ. We still want it close to θ in the sense that we want $\mathsf{E}[(\widehat{\theta} - \theta)^2]$ to be small. It's just that this is no longer the variance of $\widehat{\theta}$ since $\mathsf{E}[\widehat{\theta}] \neq \theta$. We call $\mathsf{E}[(\widehat{\theta} - \theta)^2]$ the **mean squared error** of the estimator $\widehat{\theta}$ of θ. If $\widehat{\theta}$ is an unbiased estimator of θ, the mean squared error and variance are the same.

Definition 2.2.2

The **mean squared error** (MSE) of the estimator $\hat{\theta}$ of θ is

$$MSE(\hat{\theta}) = \mathsf{E}[(\hat{\theta} - \theta)^2].$$

By the definitions of variance and mean squared error, we can see the following.

Property

If $\hat{\theta}$ is "unbiased for θ", then

$$MSE(\hat{\theta}) = Var[\hat{\theta}].$$

In general though, we can derive a relationship between MSE and variance as follows.

$$\begin{aligned} Var[\hat{\theta}] &= \mathsf{E}[(\hat{\theta} - \mathsf{E}[\hat{\theta}])^2] = \mathsf{E}[(\hat{\theta} - \theta + \theta - \mathsf{E}[\hat{\theta}])^2] \\ &= \mathsf{E}[(\hat{\theta} - \theta - B(\hat{\theta}))^2] \end{aligned}$$

Squaring that out and running the expectation across, but keeping the $\hat{\theta} - \theta$ together gives

$$\begin{aligned} Var[\hat{\theta}] &= \mathsf{E}[(\hat{\theta} - \theta)^2] - 2B(\hat{\theta}) \underbrace{\mathsf{E}[\hat{\theta} - \theta]}_{\begin{array}{c}\mathsf{E}[\hat{\theta}] - \theta \\ || \\ B(\hat{\theta})\end{array}} + (B(\hat{\theta}))^2 \\ &= \mathsf{E}[(\hat{\theta} - \theta)^2] - (B(\hat{\theta}))^2 \\ &= MSE(\hat{\theta}) - (B(\hat{\theta}))^2 \end{aligned}$$

Property

If $\hat{\theta}$ is an estimator for θ, then

$$Var[\hat{\theta}] = MSE(\hat{\theta}) - (B(\hat{\theta}))^2.$$

From this relationship, we can again see that if $\hat{\theta}$ is unbiased for θ, $Var[\hat{\theta}] = MSE(\hat{\theta})$ since the bias of the estimator is zero.

Example 2.2.1 (Long Estimation Example!)

Let $X_1, X_2, \ldots, X_n$ be a random sample from the $exp(rate = \lambda)$ distribution. Assume that $n \geq 3$.

We will consider estimating the mean of the distribution $\tau(\lambda) = 1/\lambda$ and then λ itself.

Since $\tau(\lambda) = 1/\lambda$ is the mean for this distribution, we know that $\widehat{\tau(\lambda)} := \overline{X}$ is an unbiased estimator for $\tau(\lambda)$. It's MSE is

$$MSE(\widehat{\tau(\lambda)}) \overset{unbiased}{=} Var[\widehat{\tau(\lambda)}] = Var[\overline{X}] = \frac{Var[X_1]}{n} = \frac{(1/\lambda^2)}{n} = \frac{1}{n\lambda^2}.$$

Now consider estimating λ itself. While it is perfectly reasonable to use $1/\overline{X}$, we might not want to do this if, for example, we want an unbiased estimator for λ.

Is $\hat{\lambda} := 1/\overline{X}$ an unbiased estimator of λ?

$$\mathsf{E}[\hat{\lambda}] = \mathsf{E}\left[\frac{1}{\overline{X}}\right]$$

but THIS IS IN NO WAY SHAPE OR FORM EQUAL TO

$$\frac{1}{\mathsf{E}[\overline{X}]} = \frac{1}{1/\lambda} = \lambda.$$

Expectation just doesn't work this way.

Instead,

$$\mathsf{E}[\hat{\lambda}] = \mathsf{E}\left[\frac{1}{\overline{X}}\right] = \mathsf{E}\left[\frac{n}{\sum X_i}\right] = n\mathsf{E}\left[\frac{1}{\sum X_i}\right] = n\mathsf{E}\left[\frac{1}{Y}\right]$$

where $Y = \sum_{i=1}^{n} X_i \sim \Gamma(n, \lambda)$. (We showed that the sum of exponentials has a gamma distribution in Example 1.5.7 of Section 1.5.6.)

We now have

$$\begin{aligned}
\mathsf{E}\left[\frac{1}{Y}\right] &= \int_{-\infty}^{\infty} \frac{1}{y} f_Y(y)\, dy \\
&= \int_{0}^{\infty} \frac{1}{y} \frac{1}{\Gamma(n)} \lambda^n y^{n-1} e^{-\lambda y}\, dy \\
&= \int_{0}^{\infty} \frac{1}{\Gamma(n)} \lambda^n\, y^{n-2} e^{-\lambda y}\, dy
\end{aligned}$$

We can proceed by “integrating without integrating". Ignoring all constants, we really want to integrate

$$y^{n-2}\, e^{-\lambda y}.$$

Comparing this to the non-constant part of the gamma distribution,

$$x^{\alpha-1}e^{-\beta x},$$

we see that we are integrating something that almost looks like a $\Gamma(n-1,\lambda)$ pdf. We will put in the constants needed to make this an actual $\Gamma(n-1,\lambda)$ pdf and compensate by putting the reciprocal of the constants in front of the integral. We have

$$\begin{aligned}
\mathsf{E}\left[\tfrac{1}{Y}\right] &= \int_0^\infty \tfrac{1}{\Gamma(n)}\lambda^n\, y^{n-2}e^{-\lambda y}\, dy \\
&= \tfrac{\Gamma(n-1)}{\Gamma(n)}\lambda \underbrace{\int_0^\infty \frac{1}{\Gamma(n-1)}\lambda^{n-1}y^{n-2}e^{-\lambda y}\, dy}_{1} \\
&= \tfrac{\Gamma(n-1)}{\Gamma(n)}\lambda = \tfrac{(n-2)!}{(n-1)!}\lambda = \tfrac{1}{n-1}\lambda
\end{aligned}$$

Putting things back together we have

$$\mathsf{E}[\hat{\lambda}] = n\mathsf{E}\left[\frac{1}{Y}\right] = \frac{n}{n-1}\lambda.$$

Conclusion: This is not λ, so $\hat{\lambda} = 1/\overline{X}$ is not an unbiased estimator of λ!

Let us now try to find an unbiased estimator of λ.

$\widehat{\lambda} = 1/\overline{X}$ came close in the sense that it was just a constant times λ. If we put $(n-1)/n$ in front of $\widehat{\lambda}$ then the resulting random variable will have expectation

$$\mathsf{E}\left[\frac{n-1}{n}\frac{1}{\overline{X}}\right] = \frac{n-1}{n}\mathsf{E}\left[\frac{1}{\overline{X}}\right] = \frac{n-1}{n}\frac{n}{n-1}\lambda = \lambda.$$

Thus

$$\hat{\lambda}_2 := \frac{n-1}{n}\frac{1}{\overline{X}} = \frac{n-1}{\sum X_i}$$

is an unbiased estimator of λ!

Comparison of MSEs:

Let’s find the MSE of the original $\hat{\lambda}$.

$$MSE(\widehat{\lambda}) = \mathsf{E}[(\widehat{\lambda}-\lambda)^2] = \mathsf{E}\left[\left(\tfrac{1}{\overline{X}}-\lambda\right)^2\right] = \mathsf{E}\left[\left(\tfrac{n}{Y}-\lambda\right)^2\right]$$

where $Y \sim \Gamma(n, \lambda)$.

So,

$$\begin{aligned} MSE(\widehat{\lambda}) &= \mathsf{E}\left[\left(\tfrac{n}{Y} - \lambda\right)^2\right] \\ &= n^2 \mathsf{E}\left[\tfrac{1}{Y^2}\right] - 2n\lambda \mathsf{E}\left[\tfrac{1}{Y}\right] + \lambda^2 \end{aligned}$$

We already know that $\mathsf{E}[1/Y] = \lambda/(n-1)$. We still need to find $\mathsf{E}[1/Y^2]$.

Since $Y \sim \Gamma(n, \lambda)$, we have

$$\begin{aligned} \mathsf{E}\left[\tfrac{1}{Y^2}\right] &= \int_0^\infty \tfrac{1}{y^2} \cdot \tfrac{1}{\Gamma(n)} \lambda^n y^{n-1} e^{-\lambda y}\, dy \\ &= \int_0^\infty \cdot \tfrac{1}{\Gamma(n)} \lambda^n y^{n-3} e^{-\lambda y}\, dy. \end{aligned}$$

That $y^{n-3}e^{-\lambda y}$ part of the integrand looks like a $\Gamma(n-2, \lambda)$ pdf. In order for that pdf to be complete, we also need $\frac{1}{\Gamma(n-2)}\lambda^{n-2}$. So, we rewrite things as

$$\begin{aligned} \mathsf{E}\left[\tfrac{1}{Y^2}\right] &= \tfrac{\Gamma(n-2)}{\Gamma(n)} \lambda^2 \underbrace{\int_0^\infty \frac{1}{\Gamma(n-2)} \lambda^{n-2} y^{n-3} e^{-\lambda y}\, dy}_{\text{integral of a pdf}} \\ &= \tfrac{\Gamma(n-2)}{\Gamma(n)} \lambda^2 \cdot 1 = \tfrac{1}{(n-1)(n-2)} \lambda^2. \end{aligned}$$

Putting all of the pieces together (check this), we get

$$\boxed{MSE(\widehat{\lambda}) = \tfrac{n+2}{(n-1)(n-2)}\, \lambda^2.}$$

We will now compute the MSE for the unbiased estimator $\widehat{\lambda}_2$. Since $\widehat{\lambda}_2$ is an unbiased estimator of λ, its MSE is simply its variance:

$$MSE(\widehat{\lambda}_2) \overset{unbiased}{=} Var[\widehat{\lambda}_2] = Var\left[\tfrac{n-1}{\sum X_i}\right] = (n-1)^2 Var\left[\tfrac{1}{Y}\right]$$

where $Y \sim \Gamma(n, \lambda)$.

Note that

$$\begin{aligned} Var\left[\tfrac{1}{Y}\right] &= \mathsf{E}\left[\tfrac{1}{Y^2}\right] - \left(\mathsf{E}\left[\tfrac{1}{Y}\right]\right)^2 \\ &= \tfrac{1}{(n-1)(n-2)} \lambda^2 - \left(\tfrac{1}{n-1}\, \lambda\right)^2 \\ &= \tfrac{1}{(n-1)^2(n-2)} \lambda^2. \end{aligned}$$

So,

$$MSE(\widehat{\lambda}_2) = (n-1)^2 Var\left[\frac{1}{Y}\right] = (n-1)^2 \frac{1}{(n-1)^2(n-2)}\lambda^2 = \frac{1}{n-2}\lambda^2.$$

From two unbiased estimators of a parameter, you should choose the one with smaller variance. When comparing a biased estimator with an unbiased estimator, as in this case, it makes more sense to look at MSE. The variance of an estimator represents its spread about its mean while the MSE represents its spread about the thing you are trying to estimate!

We still don't know how to really come up with estimators of parameters. At this point we are sort of guessing and then evaluating that guess. (i.e., Is it unbiased? Does it have small variance? Small MSE?) Eventually we will learn systematic approaches for coming up with estimators. For now, we are still just going to talk about how to evaluate the quality of estimators so that we know what to aim for when coming up with methods for finding estimators.

Often we are interested in "large sample properties" of estimators. For example, when estimating the mean μ of a distribution, it seems natural to assume that the sample mean $\overline{X}$ is not only a good idea, but will become an even better estimator of the true mean for larger and larger samples. It's time to talk about limits for random variables!

In general, given a random sample $X_1, X_2, \ldots, X_n$, an estimator $\widehat{\theta}$ for a parameter θ will also be based on the sample size. For example, $\widehat{\mu} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Sometimes we will emphasize the fact that an estimator $\widehat{\theta}$ of a parameter θ depends on n by writing $\widehat{\theta} = \widehat{\theta}_n$. It would be nice to be able to say that $\widehat{\theta}_n$ converges, in some sense, to the true value θ. We know what it means for a sequence of numbers to "approach" another number. Estimators, however, are not numbers. They are random variables or "potential numbers". In order to talk about convergence, we need to use probability somehow. There are several ways to do this.

2.3 Convergence in Probability

Definition 2.3.1

A sequence of random variables $\{X_n\}$ **converges in probability** to a random variable X if, for any $\varepsilon > 0$, we have

$$\lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0.$$

(Equivalently if $\lim_{n\to\infty} P(|X_n - X| \leq \varepsilon) = 1$.)

We write $X_n \xrightarrow{P} X$.

Take a moment to think about why one of these probabilities should go to 0 and the other should go to 1.

Here are some things to note.

1. Probabilities are numbers. The limit in this definition makes sense as it is for a sequence of numbers.
2. One can talk about convergence in probability of a sequence of random variables to a constant. A constant is just a very boring random variable. The constant "3" can be thought of as a random variable X where $X = 3$ with probability 1.
3. In this text, the sequence of random variables will usually be a sequence of estimators $\widehat{\theta} = \widehat{\theta}_n$ of a parameter θ that we hope will converge in probability to the constant θ.
4. The "strictness" of the inequalities in the definition of convergence in probability is not important. It is fine to say, for example, that $\lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0$.

In order to show convergence in probability, you might find the inequalities in the next Sections useful.

2.3.1 Markov's Inequality

Markov's inequality is as follows.

Markov's Inequality

For a random variable X and any numbers $r, c > 0$,

$$P(|X| \geq c) \leq \frac{\mathsf{E}[|X|^r]}{c^r}.$$

Markov's inequality is sometimes stated in terms of a non-negative random variable in which case the absolute values can be omitted.

In order to prove Markov's inequality, we will first prove something more general. We will refer to it as the "generalized Markov inequality".

The "Generalized" Markov Inequality

If X is a random variable and $g(x)$ is a non-negative real-valued function, then for any $c > 0$,

$$P(g(X) \geq c) \leq \frac{\mathsf{E}[g(X)]}{c}.$$

Once we prove this more general inequality, we can establish Markov's inequality as a simple special case as follows.

Since $r > 0$,

$$P(|X| \geq c) = P(|X|^r \geq c^r).$$

Using $g(x) = |x|^r$ in the generalized-Markov inequality, and replacing c with c^r, we know that

$$P(|X|^r \geq c^r) \leq \frac{\mathsf{E}[|X|^r]}{c^r}.$$

So,

$$P(|X| \geq c) \leq \frac{\mathsf{E}[|X|^r]}{c^r}$$

which is Markov's inequality!

It remains to prove the generalized Markov inequality. It will be useful to use the notation

$$\int_A f(x)\,dx$$

to represent the integral over all values in the set A. For example,

$$\int_2^5 f(x)\,dx = \int_{\{x:2\leq x\leq 5\}} f(x)\,dx.$$

The purpose of this notation is to allow us to easily and compactly denote more complicated regions of integration. For example, suppose that we want to integrate a function $g(x)$ over all x in the region where $g(x) > 5$. If g goes above and below 5 many times, this region may consist of a lot of disjoint intervals which

will result in a sum of several integrals. However, with this new notation we can simply write

$$\int_{\{x:g(x)>5\}} g(x)\,dx.$$

Note that if f is the pdf for a random variable X,

$$\int_{\{x:2\leq x\leq 5\}} f(x)\,dx = P(2 \leq X \leq 5).$$

Proof : **(generalized-Markov Inequality)**

Let $f(x)$ be the pdf for X. Then,

$$\begin{aligned} \mathsf{E}[g(X)] &= \int_{-\infty}^{\infty} g(x)f(x)\,dx \\ &= \textstyle\int_{\{x:g(x)\geq c\}} g(x)f(x)\,dx + \int_{\{x:g(x)<c\}} g(x)f(x)\,dx \\ &\geq \textstyle\int_{\{x:g(x)\geq c\}} g(x)f(x)\,dx \end{aligned}$$

since both f and g are non-negative.

Since $g(x) \geq c$ on this region of integration,

$$\begin{aligned} \mathsf{E}[g(X)] &\geq \textstyle\int_{\{x:g(x)\geq c\}} g(x)f(x)\,dx \geq \int_{\{x:g(x)\geq c\}} cf(x)\,dx \\ &= c\textstyle\int_{\{x:g(x)\geq c\}} f(x)\,dx = cP(g(X) \geq c) \end{aligned}$$

as desired! ■

2.3.2 Chebyshev's Inequality

Chebyshev's inequality gives a lower bound on the probability that a random variable X is "within k standard deviations of its mean".

Chebyshev's (Tchebychev's) Inequality

Let X be a random variable with mean μ and variance $\sigma^2 < \infty$. For any $k > 0$,

$$P(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}.$$

Equivalently,

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Here, $\sigma := \sqrt{\sigma^2} > 0$ is the standard deviation of X.

Proof : (Chebyshev's Inequality)

Let $g(x) = (x - \mu)^2$. Using the generalized-Markov inequality again with $c = k^2\sigma^2$ gives us

$$P((X - \mu)^2 \geq k^2\sigma^2) \leq \frac{\mathsf{E}[(X - \mu)^2]}{k^2\sigma^2} = \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}.$$

Note that the left hand side is equivalent to $P(|X - \mu| \geq k\sigma)$. ■

2.3.3 The Sample Mean and Convergence in Probability

Let $X_1, X_2, \ldots, X_n$ be a random sample from any distribution with mean μ and finite variance σ^2. Note that the sample mean

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_n$$

depends on the sample size n. We know that $\overline{X}$ is an unbiased estimator of μ. One would think that $\overline{X}$ becomes an even better estimator of μ when n gets large. Specifically, we have the following.

The Weak Law of Large Numbers (WLLN)

Let $X_1, X_2, \ldots, X_n$ be a random sample from any distribution with mean μ and finite variance σ^2. Then,

$$\overline{X} \xrightarrow{P} \mu$$

So far we only have one type of convergence (convergence in probability) for a sequence of random variables.

Once we have others, we will see that some types of convergence are "stronger" than others and some are "weaker". This is the reason for the use of the word "weak" here.

It may look weird to you to have a limit without an apparent sequence of random variables. For this reason, we sometimes switch to the notation $\overline{X}_n$ for a sample mean in order to emphasize its dependence on the sample size. Similarly, we might rewrite an estimator $\widehat{\theta}$ of θ as $\widehat{\theta}_n$.

Proof : (WLLN)

Recall Chebyshev's Inequality. For any $k > 0$, $P(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}$.

Applying Chebyshev's Inequality to the random variable $\overline{X}$, we get

$$P(|\overline{X} - \mu_{\overline{X}}| < k\sigma_{\overline{X}}) \geq 1 - \frac{1}{k^2}.$$

But, $\mu_{\overline{X}} = \mu$ and $\sigma^2_{\overline{X}} = \sigma^2/n$, which implies that $\sigma_{\overline{X}} = \sigma/\sqrt{n}$. So, we have

$$P(|\overline{X} - \mu| < k\sigma/\sqrt{n}) \geq 1 - \frac{1}{k^2} \tag{2.3.1}$$

for any $k > 0$.

Let $\varepsilon > 0$. Choose k such that $k\sigma/\sqrt{n} = \varepsilon$. i.e., Choose $k = \varepsilon\sqrt{n}/\sigma$. Then (2.3.1) becomes

$$P(|\overline{X} - \mu| < \varepsilon) \geq 1 - \frac{\sigma^2}{\varepsilon^2 n}.$$

Therefore,

$$\lim_{n\to\infty} P(|\overline{X} - \mu| < \varepsilon) \geq \lim_{n\to\infty} \left[1 - \frac{\sigma^2}{\varepsilon^2 n}\right] = 1 - 0 = 1.$$

Since probabilities can't get greater than 1, that limit on the left-hand side must equal 1.

So, we have shown that $\overline{X} \xrightarrow{P} \mu$. ■

Example 2.3.1

Let $X_1, X_2, \ldots, X_n$ be a random sample from the exponential distribution with rate λ, then

$$\overline{X} \xrightarrow{P} \frac{1}{\lambda}$$

since $\mu = \mathsf{E}[X_i] = 1/\lambda$.

Example 2.3.2

If $X_1, X_2, \ldots, X_n$ is a random sample from the $\Gamma(\alpha, \beta)$ distribution, then

$$\overline{X} \xrightarrow{P} \frac{\alpha}{\beta}$$

since $\mu = \mathsf{E}[X_i] = \alpha/\beta$.

2.3.4 Consistent Estimators

Definition 2.3.2

Let $\widehat{\theta}_n$ be an estimator of θ. $\widehat{\theta}_n$ is a **consistent estimator** of θ if

$$\widehat{\theta}_n \xrightarrow{P} \theta.$$

So, if someone asks you to show that you have a "consistent" estimator of something, they are simply asking you to show convergence in probability.

Here is a most useful theorem!

Theorem 2.3.1

An unbiased estimator $\widehat{\theta}_n$ of θ is a consistent estimator of θ if

$$\lim_{n\to\infty} Var[\widehat{\theta}_n] = 0.$$

Proof : By Chebyshev's Inequality, we have, for any $k > 0$

$$P(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}.$$

In this theorem, our random variable is $\widehat{\theta}_n$ and, since it's unbiased, its mean is θ. Let us use σ_n^2 to denote its variance: $\sigma_n^2 = Var[\widehat{\theta}_n]$. Chebyshev's Inequality becomes

$$P(|\widehat{\theta}_n - \theta| < k\sigma_n) \geq 1 - \frac{1}{k^2}.$$

Take any $\varepsilon > 0$. Setting $k\sigma_n = \varepsilon$, we see that, by choosing $k = \varepsilon/\sigma_n$, we have

$$P(|\widehat{\theta}_n - \theta| < \varepsilon) \geq 1 - \frac{\sigma_n^2}{\varepsilon^2}.$$

Now take the limit as $n \to \infty$ of both sides. Since are given that that $\lim_{n\to\infty} \sigma_n^2 = 0$, we get

$$\lim_{n\to\infty} P(|\widehat{\theta}_n - \theta| < \varepsilon) \geq 1.$$

Since that probability can't get greater than 1, we have shown that

$$\lim_{n\to\infty} P(|\widehat{\theta}_n - \theta| < \varepsilon) = 1.$$

Thus, $\widehat{\theta}_n \xrightarrow{P} \theta$ and so $\widehat{\theta}_n$ is a consistent estimator of θ. ■

You will be happy to know that this Theorem can be generalized a little bit. Before we do this, we will need a definition.

Definition 2.3.4

$\widehat{\theta}_n$ is an **asymptotically unbiased** estimator of θ if

$$\lim_{n\to\infty} \mathsf{E}[\widehat{\theta}_n] = \theta.$$

We now have the following.

Theorem 2.3.2 'That' Theorem

An asymptotically unbiased estimator $\widehat{\theta}_n$ of θ is a consistent estimator of θ if

$$\lim_{n\to\infty} Var[\widehat{\theta}_n] = 0.$$

Proof : Recall the relationship

$$Var[\widehat{\theta}_n] = MSE[\widehat{\theta}_n] - (B[\widehat{\theta}_n])^2 \tag{2.3.2}$$

"Asymptotically unbiased" means that

$$\lim_{n\to\infty} \mathsf{E}[\widehat{\theta}_n] = \theta.$$

Since the bias is,

$$B[\widehat{\theta}_n] = \mathsf{E}[\widehat{\theta}_n] - \theta,$$

"asymptotically unbiased" gives us that the bias of our estimator goes to 0 as $n \to \infty$.

Since are given that the variance goes to zero, by (2.3.2) we now have that the MSE must go to zero.

$$\lim_{n\to\infty} MSE[\widehat{\theta}_n] = \lim_{n\to\infty} \mathsf{E}[(\widehat{\theta}_n - \theta)^2]$$

Recall the inequality

$$P(g(X) \geq c) \leq \frac{\mathsf{E}[X]}{c}.$$

Applying this here, we have, for any $\varepsilon > 0$,

$$P(|\widehat{\theta}_n - \theta| \geq \varepsilon) = P((\widehat{\theta}_n - \theta)^2 \geq \varepsilon^2) \leq \frac{\mathsf{E}[(\widehat{\theta}_n - \theta)^2]}{\varepsilon^2}.$$

Thus,

$$\lim_{n\to\infty} P(|\widehat{\theta}_n - \theta| \geq \varepsilon) \leq \lim_{n\to\infty} \frac{\mathsf{E}[(\widehat{\theta}_n - \theta)^2]}{\varepsilon^2} = 0.$$

Since a probability can't be negative, we must have

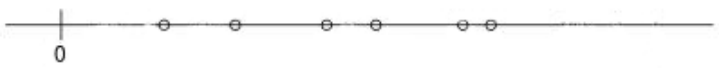
$$\lim_{n\to\infty} P(|\widehat{\theta}_n - \theta| \geq \varepsilon) = 0.$$

So, $\widehat{\theta}_n \xrightarrow{P} \theta$, as desired. ■

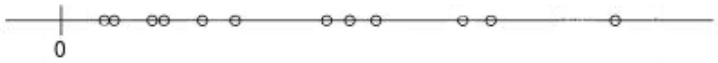
Example 2.3.3

Suppose that $X_1, X_2, \ldots, X_n \overset{iid}{\sim} unif(0, \theta)$. Show that the maximum, $X_{(n)}$ is a consistent estimator of θ.

Our sampled values are uniformly distributed over the interval $(0, \theta)$. If θ is unknown, it seems reasonable to estimate the right endpoint of the interval of possible values by using the maximum value in the sample.



It also seems like we would get a better idea about where θ is if we collect more data. When including additional data, the maximum can only go up.



In this example, we are going to show that, as the sample size increases, the maximum value in the data set is approaching the right endpoint θ in the sense that it is converging in probability to θ.

To begin, we find the distribution for $X_{(n)}$. The cdf for the $unif(0, \theta)$ distribution is

$$F(x) = \begin{cases} 0 & , \quad x < 0 \\ x/\theta & , \quad 0 \leq x < \theta \\ 1 & , \quad x \geq \theta. \end{cases}$$

Thus, the cdf for the maximum is

$$\begin{aligned} F_{X_{(n)}}(x) &= P(X_{(n)} \leq x) \\ &= P(X_1 \leq x, X_2 \leq x, \ldots, X_n \leq x) \\ &\overset{iid}{=} [P(X_1 \leq x)]^n \\ &= [x/\theta]^n \end{aligned}$$

for $0 \leq x < \theta$.

The pdf for the maximum is then

$$f_{X_{(n)}}(x) = \frac{d}{dx} F_{X_{(n)}}(x) = \frac{n}{\theta^n} x^{n-1}.$$

for x between 0 and θ. We will write this with an indicator as

$$f_{X_{(n)}}(x) = \frac{n}{\theta^n} x^{n-1} \, I_{(0,\theta)}(x).$$

The expected value of the maximum is

$$\begin{aligned} \mathsf{E}[X_{(n)}] &= \int_{-\infty}^{\infty} x \cdot f_{X_{(n)}}(x) \, dx \\ &= \int_0^{\theta} x \cdot \tfrac{n}{\theta^n} x^{n-1} \, dx \\ &= \int_0^{\theta} \tfrac{n}{\theta^n} x^n \, dx \\ &= \tfrac{n}{n+1} \, \theta. \end{aligned}$$

This is less than θ and is getting closer and closer to θ as the sample size n gets large. In particular, we have

$$\lim_{n\to\infty} \mathsf{E}[X_{(n)}] = \lim_{n\to\infty} \frac{n}{n+1}\,\theta = \theta,$$

so $X_{(n)}$ is an asymptotically unbiased estimator for θ.

If we can show that $\lim_{n\to\infty} Var[X_{(n)}] = 0$, we can use Theorem 2.3.2 to conclude that $X_{(n)} \xrightarrow{P} \theta$.

To find the variance, we first compute the second moment

$$\begin{aligned}
\mathsf{E}[X_{(n)}^2] &= \textstyle\int_{-\infty}^{\infty} x^2 \cdot f_{X_{(n)}}(x)\,dx \\
&= \textstyle\int_0^{\theta} x^2 \cdot \frac{n}{\theta^n} x^{n-1}\,dx \\
&= \textstyle\int_0^{\theta} \frac{n}{\theta^n} x^{n+1}\,dx \\
&= \textstyle\frac{n}{n+2}\,\theta^2.
\end{aligned}$$

Then

$$Var[X_{(n)}] = \mathsf{E}[X_{(n)}^2] - (\mathsf{E}[X_{(n)}])^2 = \frac{n}{n+2}\,\theta^2 - \left(\frac{n}{n+1}\theta\right)^2 = \frac{n}{(n+1)^2(n+2)}\theta^2,$$

which goes to 0 as $n \to \infty$. So, we have that

$$X_{(n)} \xrightarrow{P} \theta$$

which means that $X_{(n)}$ is a consistent estimator of θ.

2.3.5 Things About Convergence in Probability That Will Not Surprise You

Theorem 2.3.3

Suppose that $\{X_n\}$ and $\{Y_n\}$ are sequences of random variables with $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$ for some random variables X and Y.

Then

1. $X_n + Y_n \xrightarrow{P} X + Y$
2. $X_n Y_n \xrightarrow{P} XY$
3. $X_n/Y_n \xrightarrow{P} X/Y$ as long as the denominators are non-zero with probability 1
4. $g(X_n) \xrightarrow{P} g(X)$ for any continuous function g

For our estimation purposes, we will be most interested in the case where X and Y are constants which, as mentioned before, can be thought of as really boring random variables. We will prove parts 1 and 4 of this Theorem and leave the others as exercises.

Proof : **(Theorem 2.3.3, Part 1)** We will use the Triangle Inequality: $|a+b| \leq |a|+|b|$ on the random variables like this

$$|X_n + Y_n - (X+Y)| = |(X_n - X) + (Y_n - Y)| \leq |X_n - X| + |Y_n - Y|. \tag{2.3.3}$$

(Incidentally, an inequality $X \leq Y$, for example, means that $P(X \leq Y) = 1$.)

Take any $\varepsilon > 0$. Note that (2.3.3) implies that

$$P(|X_n + Y_n - (X+Y)| > \varepsilon) \leq P(|X_n - X| + |Y_n - Y| > \varepsilon). \tag{2.3.4}$$

Further, we have that

$$P(|X_n - X| + |Y_n - Y| > \varepsilon) \leq P(|X_n - X| > \varepsilon/2) + P(|Y_n - Y| > \varepsilon/2), \tag{2.3.5}$$

as described in the following aside.

Aside:

In order to have $|X_n - X| + |Y_n - Y| > \varepsilon$, we must have at least one of the events

$$\{|X_n - X| > \varepsilon/2\}, \qquad \{|Y_n - Y| > \varepsilon/2\}$$

be true. i.e.,

$$\{|X_n - X| + |Y_n - Y| > \varepsilon\} \qquad \Rightarrow \qquad \{|X_n - X| > \varepsilon/2\} \text{ or } \{|Y_n - Y| > \varepsilon/2\} \text{ or both.}$$

Equivalently

$$\{|X_n - X| + |Y_n - Y| > \varepsilon\} \subseteq \{|X_n - X| > \varepsilon/2\} \cup \{|Y_n - Y| > \varepsilon/2\}. \tag{2.3.6}$$

Recall that for events A and B,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \leq P(A) + P(B).$$

So, (2.3.6) implies that

$$\begin{aligned} P(|X_n - X| + |Y_n - Y| > \varepsilon) &\leq P(\{|X_n - X| > \varepsilon/2\} \cup \{|Y_n - Y| > \varepsilon/2\}) \\ &\leq P(|X_n - X| > \varepsilon/2) + P(|Y_n - Y| > \varepsilon/2). \end{aligned}$$

Finally,

$$\begin{aligned}
\lim_{n\to\infty} P(|X_n + Y_n - (X+Y)| > \varepsilon) &\overset{(2.3.4)}{\leq} \lim_{n\to\infty} P(|X_n - X| + |Y_n - Y| > \varepsilon) \\
&\overset{(2.3.5)}{\leq} \lim_{n\to\infty} [P(|X_n - X| > \varepsilon/2) + P(|Y_n - Y| > \varepsilon/2)] \\
&= \lim_{n\to\infty} P(|X_n - X| > \varepsilon/2) + \lim_{n\to\infty} P(|Y_n - Y| > \varepsilon/2) \\
&= 0 + 0 = 0
\end{aligned}$$

since $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$. Since $P(|X_n + Y_n - (X+Y)| > \varepsilon)$ can not be less than zero, we have

$$\lim_{n\to\infty} P(|X_n + Y_n - (X+Y)| > \varepsilon) = 0$$

and therefore $X_n + Y_n \xrightarrow{P} X + Y$, as desired. ■

We will prove Part 4 in the case that convergence is to a constant a. That is, we will assume that $X = a$ for some real number a. The more general proof requires a bit of measure theory. Furthermore, we are really only interested in convergence to constants as the entire point of our diversion into sequences of random variables is to talk about a sequence of estimators (random variables) of a constant θ getting close to θ.

Proof : (Theorem 2.3.3, Part 4, X = a)

Let $\varepsilon > 0$. g continuous implies that, for any real number a, there exists a $\delta > 0$ such that, for all x with $|x - a| < \delta$, we have $|g(x) - g(a)| < \varepsilon$.

Note that

$$|x - a| < \delta \quad \Rightarrow \quad |g(x) - g(a)| < \varepsilon$$

and yet $|g(x) - g(a)|$ may be less than ε for other x that are not within δ of a.

Plugging in the random sequence, this means that the event that $|X_n - a| < \delta$ happens implies that the event $|g(X_n) - g(a)| < \varepsilon$ but not necessarily the other way around.

Writing these events as sets, we have

$$\{|X_n - a| < \delta\} \subseteq \{|g(X_n) - g(a)| < \varepsilon\}.$$

Using the fact that $A \subseteq B$ implies that $P(A) \leq P(B)$, we have

$$P(|g(X_n) - g(a)| < \varepsilon) \geq P(|X_n - a| < \delta)$$

Since $X_n \xrightarrow{P} a$ as $n \to \infty$, the right-hand side goes to 1. Hence

$$\lim_{n\to\infty} P(|g(X_n) - g(a)| < \varepsilon) \geq 1.$$

Since the probability can not be greater than 1, we have

$$\lim_{n\to\infty} P(|g(X_n) - g(a)| < \varepsilon) = 1$$

and so $g(X_n) \xrightarrow{P} g(a)$, as desired. ■

Note

Note that we didn't need g to be continuous everywhere, only at a. We can also relax continuity of g for the more general result $g(X_n) \xrightarrow{P} g(X)$ as long as, for any point c of discontinuity of g, we have $P(X_n = c) = 0$ and $P(X = c) = 0$. In other words, if there are problem points with g, it doesn't matter because we will never be evaluating g at those points!

Example 2.3.4

If $X_n \xrightarrow{P} b$ and a is a constant, $aX_n \xrightarrow{P} ab$.

This is either a result of Part 2 of the Theorem, using the random variable sequence $\{Y_n\}$ with $Y_n = a$ with probability 1 or as a result of Part 4 of the Theorem with $g(x) = ax$.

Example 2.3.5

Let $\{a_n\}$ be a sequence of real numbers such that

$$\lim_{n\to\infty} a_n = a.$$

There are no random variables here but, as mentioned before, constants can be thought of as really boring random variables. Technically, we can talk about convergence in probability of a_n to a. The limit in this example says that, for any $\varepsilon > 0$, there exists a constant N such that $|a_n - a| < \varepsilon$ for all $n \geq N$.

So, we have, for each $n < N$, the probability $P(|a_n - a| < \varepsilon)$ is either 0 or 1. However, $P(|a_n - a| < \varepsilon) = 1$ for all $n \geq N$.

Thus,

$$\lim_{n\to\infty} P(|a_n - a| < \varepsilon) = 1,$$

and we have that $a_n \xrightarrow{P} a$.

Example 2.3.6

Suppose that $X_1, X_2, \ldots, X_n \overset{iid}{\sim} unif(0, \theta)$. In Example 2.3.3, we saw that the maximum, $X_{(n)}$ is a consistent estimator of θ. We proved this using Theorem 2.3.2.

Suppose that we did not know Theorem 2.3.2, only knew Theorem 2.3.1, and still needed to prove the result. We would need $\mathsf{E}[X_{(n)}] = \theta$ and $\lim_{n\to\infty} Var[X_{(n)}] = 0$.

We saw, however, that

$$\mathsf{E}[X_{(n)}] = \frac{n}{n+1}\theta \neq \theta.$$

All is not lost. We could define a new sequence of random variables $\{Y_n\}$ where

$$Y_n = \frac{n+1}{n} X_{(n)}.$$

By design, Y_n is an unbiased estimator for θ since

$$\mathsf{E}[Y_n] = \mathsf{E}\left[\frac{n+1}{n} X_{(n)}\right] = \frac{n+1}{n}\,\mathsf{E}[X_{(n)}] = \frac{n+1}{n} \cdot \frac{n}{n+1}\,\theta = \theta.$$

Furthermore,

$$\begin{aligned} Var[Y_n] &= Var\left[\tfrac{n+1}{n} X_{(n)}\right] = \left(\tfrac{n+1}{n}\right)^2 Var[X_{(n)}] \\ &= \left(\tfrac{n+1}{n}\right)^2 \tfrac{n}{(n+1)^2(n+2)}\,\theta^2 \\ &= \tfrac{1}{n(n+1)}\,\theta^2, \end{aligned}$$

which goes to 0 as n goes to ∞.

By Theorem 2.3.1, we now have that

$$Y_n \xrightarrow{P} \theta.$$

So, we also have

$$X_{(n)} = \frac{n}{n+1}\, Y_n \xrightarrow{P} 1 \cdot \theta = \theta$$

by Part 2 of Theorem 2.3.3. Here we have used Example 2.3.5 and the fact that $n/(n+1)$ converges to 1 as a sequence of real numbers.

2.4 Convergence In Distribution

Definition 2.4.1

Let $\{X_n\}$ be a sequence of random variables with cdfs $F_n(x) = P(X_n \leq x)$. Let X be a random variable with cdf $F(x)$.

X_n **converges in distribution** to X if

$$\lim_{n\to\infty} F_n(x) = F(x)$$

for all x where F is continuous.

We write $X_n \xrightarrow{d} X$.

Other notations are $X_n \xrightarrow{L} X$ as this is sometimes called "convergence in law", and $X_n \Rightarrow X$. When someone asks you to find the "limiting distribution" for a sequence of random variables, they are referring to convergence in distribution. Convergence in distribution is weaker in convergence in probability in the sense that $X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{d} X$ but not necessarily the other way around. We will prove this.

Example 2.4.1

Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} Pareto(1)$. Investigate the convergence in distribution of $Y_n := nX_{(1)}$ where $X_{(1)} = \min(X_1, X_2, \ldots, X_n)$.

The $Pareto(1)$ pdf is

$$f(x;\gamma) = \frac{\gamma}{(1+x)^{\gamma+1}} I_{(0,\infty)}(x) \overset{\gamma=1}{=} \frac{1}{(1+x)^2} I_{(0,\infty)}(x).$$

The cdf is

$$F(x) = \int_0^x (1+u)^{-2}\, du = \frac{-1}{1+u}\Big|_0^x = 1 - \frac{1}{1+x}$$

We need to find the cdf of Y_n and then let $n \to \infty$.

$$\begin{aligned} F_{Y_n}(y) &= P(Y_n \le y) = P(nX_{(1)} \le y) = P(X_{(1)} \le y/n) \\ &= 1 - [P(X_1 > y/n)]^n = 1 - \left(\tfrac{1}{1+y/n}\right)^n \\ &= 1 - (1+y/n)^{-n} \to 1 - e^{-y} \end{aligned}$$

as $n \to \infty$.

This is the cdf of the exponential distribution with rate 1. (If you don't recognize it, take the derivative to find the pdf.)

So,

$$Y_n \xrightarrow{d} Y \sim exp(rate = 1).$$

Example 2.4.2

Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} unif(0,1)$. Consider $Y_n = X_{(n)} = \max(X_1, X_2, \ldots, X_n)$.

The cdf for the $unif(0,1)$ distribution is $F(x) = P(X \le x) = \int_0^x 1\, du = x$, for $0 \le x < 1$. Note that the cdf is 0 for $x < 0$ and is 1 for $x \ge 1$.

The cdf of Y_n is

$$\begin{aligned} F_{Y_n}(y) &= P(Y_n \le y) = P(X_{(n)} \le y) \\ &= [P(X_1 \le y)]^n = y^n \end{aligned}$$

for $0 \le y < 1$.

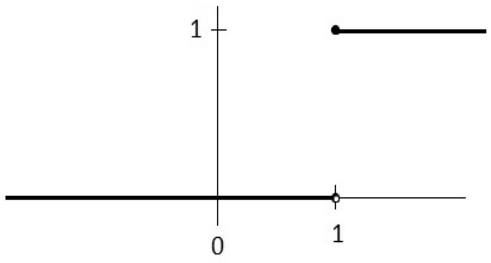
The complete cdf is

$$F_{Y_n}(y) = \begin{cases} 0 & , \quad y < 0 \\ y^n & , \quad 0 \le y < 1 \\ 1 & , \quad y \ge 1 \end{cases}$$

Letting $n \to \infty$, we get the limiting cdf

$$F(y) = \begin{cases} 0 & , \quad y < 1 \\ 1 & , \quad y \ge 1 \end{cases}$$

which looks like this



This is the cdf of a random variable that equals 1 with probability 1. We say

$$Y_n \xrightarrow{d} Y \qquad \text{where } Y = 1 \text{ w.p. } 1$$

or simply

$$Y_n \xrightarrow{d} 1.$$

Important Note: Cdfs are always right-continuous functions. This means that if and when you have a jump, you will have a "hole" on the lower piece and a "filled in point" on the higher piece. The proof of this is not that difficult but requires a lot more "machinery" in place than we have here. Sometimes, when trying to show convergence in distribution, you will end up with a left-continuous function. This is not a cdf but it is the same as the function you would get if you switched out the point and the hole which would be a cdf. Fortunately, to show convergence in distribution, we only need to show convergence of the sequence of cdfs to a cdf F "for all x where F is continuous." In other words, we don't care about these particular trouble points!

2.4.1 Convergence in Probability is Stronger

Theorem 2.4.1

$X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{d} X$

Proof : Let $F_n(x) = P(X_n \leq x)$. Let $F(x) = P(X \leq x)$. Let $\varepsilon > 0$. Let x be a point of continuity of F.

Note that

$$F_n(x) = P(X_n \leq x) = P(X_n \leq x, X \leq x + \varepsilon) + P(X_n \leq x, X > x + \varepsilon).$$

Also, note that the "event" that $\{X_n \leq x, X > x + \varepsilon\}$ is contained in the event that $\{|X_n - X| > \varepsilon\}$. That is, if the first event is true then the second is too. (The second event says that X_n and X are more than ε apart. In

the first event, X_n and X are more than ε apart but in a very specific way involving a fixed x.)

Thus,

$$F_n(x) \leq P(X_n \leq x, X \leq x + \varepsilon) + P(|X_n - X| > \varepsilon).$$

This is "trivially"

$$\leq P(X \leq x + \varepsilon) + P(|X_n - X| > \varepsilon)$$

since we have removed a restriction. So,

$$F_n(x) = F(x + \varepsilon) + P(|X_n - X| > \varepsilon).$$

On the other hand,

$$\begin{aligned} F(x - \varepsilon) &= P(X \leq x - \varepsilon) = P(X \leq x - \varepsilon, X_n \leq x) + P(X \leq x - \varepsilon, X_n > x) \\ &\leq P(X_n \leq x) + P(|X_n - X| > \varepsilon) \\ &= F_n(x) + P(|X_n - X| > \varepsilon). \end{aligned}$$

So, we have

$$F(x - \varepsilon) - P(|X_n - X| > \varepsilon) \leq F_n(x) + F(x + \varepsilon) + P(|X_n - X| > \varepsilon).$$

Letting $n \to \infty$ all the way across, and using the fact that $X_n \xrightarrow{P} X$, we get

$$F(x - \varepsilon) - 0 \leq \lim_{n \to \infty} F_n(x) \leq F(x + \varepsilon) + 0$$

or

$$F(x - \varepsilon) \leq \lim_{n \to \infty} F_n(x) \leq F(x + \varepsilon). \tag{2.4.7}$$

Notice that we did not write $F(x)$ in the middle– we are not yet sure that $F_n(x)$ has a limit! But.... letting ε become arbitrarily small in (2.4.7) makes both the right and left-hand sides go to $F(x)$, squeezing in of the limit in the center. So,

$$\lim_{n \to \infty} F_n(x) = F(x).$$

Therefore, $X_n \xrightarrow{d} X$, as desired! ■

Example 2.4.3

Counter-Example:

In general, $X_n \xrightarrow{d} X \not\Rightarrow X_n \xrightarrow{P} X$. For example, consider $X_1, X_2, \ldots$ all iid with distribution

$$P(X_i = 1) = 1/2$$

$$P(X_i = -1) = 1/2$$

Let X be a random variable with the exact same distribution. Clearly (since the cdfs are all the same and are not changing) $X_n \xrightarrow{d} X$.

However,

$$P(|X_n - X| \geq 2) = 1/2 \not\to 0$$

so $X_n \xrightarrow{P} X$.

The Reverse is Sometimes True

We have already proven that

$$X_n \xrightarrow{P} X \text{ implies } X_n \xrightarrow{d} X$$

and we have seen an example showing that

$$X_n \xrightarrow{d} X \text{ does } \textbf{not} \text{ imply } X_n \xrightarrow{P} X.$$

However, if X is a constant, we do have the reverse implication!

Theorem 2.4.2

Suppose that $X \xrightarrow{d} c$ where c is a constant. Then $X_n \xrightarrow{P} c$.

Proof : Let $\varepsilon > 0$. Note that

$$\begin{aligned} P(|X_n - c| > \varepsilon) &= P(X_n > c + \varepsilon) + P(X_n < c - \varepsilon) \\ &\leq P(X_n > c + \varepsilon) + P(X_n \leq c - \varepsilon) \\ &= 1 - F_{X_n}(c + \varepsilon) + F_{X_n}(c - \varepsilon). \end{aligned}$$

Since $X_n \xrightarrow{d} c$, $\lim_{n\to\infty} F_{X_n}(c + \varepsilon) = F(c + \varepsilon)$ and $\lim_{n\to\infty} F_{X_n}(c - \varepsilon) = F(c - \varepsilon)$, where $F(x)$ is the cdf of the "constant random variable" $X = c$. This cdf is

$$F(x) = P(X \leq x) = P(c \leq x) = \begin{cases} 0 & , \text{ if } x < c \\ 1 & , \text{ if } x \geq c \end{cases}$$

So,

$$\begin{aligned} \lim_{n\to\infty} P(|X_n - c| > \varepsilon) &\leq \lim_{n\to\infty} [1 - F_{X_n}(c + \varepsilon) + F_{X_n}(c - \varepsilon)] \\ &= 1 - 1 + 0 = 0. \end{aligned}$$

Since a probability can't get less than 0,

$$\lim_{n\to\infty} P(|X_n - c| > \varepsilon) \leq 0 \qquad \Rightarrow \qquad \lim_{n\to\infty} P(|X_n - c| > \varepsilon) = 0$$

so we have $X_n \xrightarrow{P} c$. ■

2.4.2 The Continuous Mapping Theorem

Suppose we have a sequence of random variables with $X_n \xrightarrow{d} X$. Can we say that, for example, $\overline{X}^2 \xrightarrow{d} X^2$?

The answer is yes and, in fact, convergence in distribution is preserved when put through any continuous function! When doing this, you should cite the Continuous Mapping Theorem.

The Continuous Mapping Theorem

Suppose that $X_n \xrightarrow{d} X$ and g is a continuous function. Then

$$g(X_n) \xrightarrow{d} g(X).$$

In Part 4 of Theorem 2.3.3, we proved the analogous result for convergence in probability. That result is also referred to as the Continuous Mapping Theorem. While that proof was relatively easy, the proof here is omitted

as it is lengthy, much more difficult, and would require us to take a significant detour from our discussion about estimators to build up the result.

Convergence in distribution is "weaker" than convergence in probability in the sense that

$$X_n \xrightarrow{P} X \quad \Rightarrow \quad X_n \xrightarrow{d} X$$

but

$$X_n \xrightarrow{P} X \quad \not\Rightarrow \quad X_n \xrightarrow{d} X.$$

If you are wondering why the Weak Law of Large Numbers, which is a convergence in probability result, has the word "weak" in the title, it is because convergence in probability, while "stonger" than convergence in distribution, is weaker than other types of convergence. From this point of view, convergence in distribution of a sequence of random variables seems extra weak. Indeed, it has nothing to do with the random variables getting closer together and everything to do with their distributions getting closer together. It is still quite useful as knowing the distribution in the limit tells us about the behavior the the sequence in the limit.

Here is a silly, yet illuminating, example.

Example 2.4.4

Let $X_1, X_2, \ldots$ be a sequence of iid $N(0,1)$ random variables. Let X be another independent $N(0,1)$ random variable.

The cdf for X_n is denoted by $F_n(x) = P(X_n \leq x)$, but this is just the standard normal cdf $\Phi(x)$, which has no dependence on n. (This is as opposed to an example where we have something like $X_n \sim N(\mu, 1/n)$.)

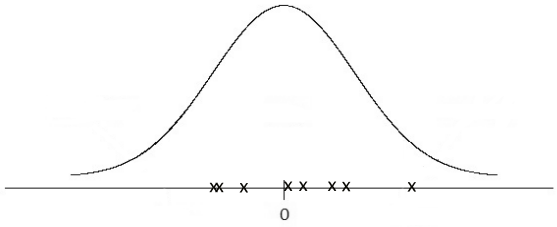
Since $X \sim N(0,1)$, we also have that its cdf is $F(x) = \Phi(x)$. Thus, we have that

$$\lim_{n\to\infty} F_n(x) = \lim_{n\to\infty} \Phi(x) = \Phi(x) = F(x),$$

so we can conclude that $X_n \xrightarrow{d} X$. Nothing is moving though– this is akin to saying that

$$\lim_{n\to\infty} 3 = 3.$$

If you were to observe these random variables, you might see values like these.



They can be anywhere between $-\infty$ and ∞ but they are mostly piled up in the high-probability center. (For the standard normal distribution, approximately 99% of the area under this curve is between -3 and 3.) By symmetry of the distribution about 0 you are equally likely to observe a value in, for example, the interval $(-2, -1)$ as you are in the interval $(1, 2)$. Again, we have convergence in distribution of the X_n to X, but if the sampled values are independent, there is no reason to expect consecutive values to be close to one another or to anything! This is why convergence in distribution is considered to be so "weak".

In Section 2.3.5 we discussed results such as

$$X_n \xrightarrow{P} X \text{ and } Y_n \xrightarrow{P} Y \quad \Rightarrow \quad X_n + Y_n \xrightarrow{P} X + Y.$$

The Section was called "Things About Convergence in Probability That Will Not Surprise You". If you agreed with that title then you will surely be surprised by this next statement.

If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} Y$, then we **do not necessarily have** that

$$X_n + Y_n \xrightarrow{d} X + Y.$$

Example 2.4.5

As in Example 2.4.4, suppose that we have a sequence $X_1, X_2, \ldots$ of iid $N(0, 1)$ random variables and X an independent $N(0, 1)$ random variable. We know that

$$X_n \xrightarrow{d} X.$$

Now define a second sequence $Y_1, Y_2, \ldots$ with

$$Y_n := -X_n.$$

By symmetry of the standard normal distribution about 0, it is easy to see that the Y_n are iid $N(0, 1)$ random

variables. As such, we know that

$$Y_n \xrightarrow{d} Y$$

where $Y \sim N(0,1)$, just as in Example 2.4.4

We can say that both sequences are converging to a standard normal random variable. For the first sequence we denote it with an X and for the second sequence we denote it with a Y. The X and Y are not really connected with the original sequences. In particular, we do not have $Y = -X$ because when we write

$$X_n \xrightarrow{d} X \sim N(0,1),$$

we are really just saying

$$X_n \xrightarrow{d} N(0,1).$$

X is used as a "placeholder" random variable from the limiting distribution.

Consider now the sequence $\{X_n + Y_n\}$. Since $Y_n = -X_n$, this is a sequence of zeros. You can think of the elements of the sequence as random variables that are equal to the value 0 with probability 1. The sequence is converging to 0, in distribution or in any other sense. On the other hand the distribution of $X + Y$ is normal with mean 0 and variance 2. Thus, we do not have $X_n + Y_n$ converging in distribution to $X + Y$. Surprise!

All that said, a random variable can be more than one-dimensional. If we look at the random vector $Z_n = (X_n, Y_n)$ and it converges in distribution to $Z = (X, Y)$, then the Continuous Mapping Theorem says that $X_n + Y_n \xrightarrow{d} X + Y$ because we can put $Z_n \xrightarrow{d} Z$ through the continuous function $g(x, y) = x + y$. We will discuss what is meant by convergence in distribution for vector-valued random variables in Section 2.4.3.

2.4.3 Slutsky's Theorem: Mixing Convergence Types

Here is a Theorem that can be used for mixing convergence types.

Slutsky's Theorem

Suppose that $X_n \xrightarrow{P} a$ and $Y_n \xrightarrow{d} Y$. Then

1. $X_n + Y_n \xrightarrow{d} a + Y$
2. $X_n Y_n \xrightarrow{d} aY$
3. $Y_n / X_n \xrightarrow{d} Y/a$ (if $a \neq 0$)

It should not be surprising that, in all cases we are only guaranteed the weaker type of convergence!

We will now prove Part 1 of Slutsky's Theorem. In order to do this, we need the notion of a **joint cumulative distribution function** (joint cdf). The joint cdf for random variables $X_1, X_2, \ldots, X_k$ is denoted and given by

$$F(x_1, x_2, \ldots, x_k) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_k \leq x_k).$$

Using this definition, we can talk about a vector-valued random variable $(X_{1,n}, X_{2,n}, \ldots, X_{k,n})$ converging in distribution to a random vector $(X_1, X_2, \ldots, X_k)$ if

$$\lim_{n\to\infty} P(X_{1,n} \leq x_1, X_{2,n} \leq x_2, \ldots, X_{k,n} \leq x_k) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n).$$

Proof : (Part 1 of Slutsky's Theorem)

We will prove that $X_n \xrightarrow{P} a$ and $Y_n \xrightarrow{d} Y$ implies that X_n and Y_n converge jointly in distribution to a and Y. In other words, we will look at the cdf of $Z_n = (X_n, Y_n)$ and show that it converges to the cdf of $Z = (a, Y)$. Once this is done, we will have the desired result by the Continuous Mapping Theorem using $Z_n \xrightarrow{d} Z$ and putting it through the continuous function $f(z_1, z_2) = z_1 + z_2$.

The joint cdf of X_n and Y_n is

$$F_n(x, y) = P(X_n \leq x, Y_n \leq y).$$

Let $F(x, y)$ be the joint cdf of (a, Y). We have that

$$F(x, y) = P(a \leq x, Y \leq y) = \begin{cases} P(Y \leq y) & , \text{ if } a \leq x \\ 0 & , \text{ if } a > x \end{cases}.$$

Let $\mathcal{C}$ denote the set of points of continuity of the cdf for Y. The points of continuity for $F(x, y)$ are then given by

$$\mathcal{D} = \{(x, y) : x \neq a, y \in \mathcal{C}\}.$$

These are the points for which we need to show that

$$\lim_{n\to\infty} F_n(x, y) = F(x, y).$$

Take any $(x, y) \in \mathcal{D}$.

Case One: $x < a$

Note that

$$\begin{aligned} F_n(x,y) &= P(X_n \leq x, Y_n \leq y) \\ &\leq P(X_n \leq x) \\ &= P(X_n - a \leq x - a) \\ &= P(a - X_n \geq \underbrace{a - x}_{>0}) \\ &\leq P(a - X_n \geq a - x) + P(a - X_n \leq -(a - x)) \\ &= P(|a - X_n| \geq a - x) = P(|X_n - a| \geq a - x). \end{aligned}$$

Since $X_n \xrightarrow{P} a$, this last expression goes to 0 as $n \to \infty$. Thus, we have

$$\lim_{n\to\infty} F_n(x,y) \leq 0$$

which implies that

$$\lim_{n\to\infty} F_n(x,y) = 0$$

since the joint cdf is a probability and must be between 0 and 1.

Recall that we are in the case where $x < a$ and so $F(x,y) = 0$. Thus, we have

$$\lim_{n\to\infty} F_n(x,y) = F(x,y)$$

when $x < a$.

Case Two: $x > a$

Note that

$$\begin{aligned} F_n(x,y) &= P(X_n \leq x, Y_n \leq y) \\ &\leq P(Y_n \leq y). \end{aligned}$$

So,

$$\lim_{n\to\infty} F_n(x,y) \leq \lim_{n\to\infty} P(Y_n \leq y) = P(Y \leq y) \tag{2.4.8}$$

since $Y_n \xrightarrow{d} Y$.

On the other hand,

$$\begin{aligned} F_n(x,y) &= P(X_n \leq x, Y_n \leq y) \\ &= P(Y_n \leq y) - P(X_n > x, Y_n \leq y) \\ &\geq P(Y_n \leq y) - P(X_n > x) \\ &= P(Y_n \leq y) - P(X_n - a > x - a) \\ &\geq P(Y_n \leq y) - P(|X_n - a| > x - a) \end{aligned}$$

Since the first term goes to $P(Y \leq y)$ and the second term goes to 0, we have

$$\lim_{n\to\infty} F_n(x,y) \geq P(Y \leq y). \tag{2.4.9}$$

By (2.4.8) and (2.4.9), we have

$$\lim_{n\to\infty} F_n(x,y) = P(Y \leq y).$$

Since we are in the case where $x > a$, $F(z_1, z_2) = P(Y \leq y)$ and we have the desired result!

■

2.4.4 Convergence in Distribution and Moment Generating Functions

We have already claimed that moment generating functions uniquely determine a distribution. A distribution, for us, is defined by a pdf which we have then used to define a cdf. So, you might find the following theorem, which we will not prove, highly believable.

Theorem 2.4.3

Let $X_1, X_2, \ldots$ be a sequence of random variables.

Let $F_n(x)$ be the cdf of X_n.

Let $M_n(t)$ be the mgf of X_n.

Let X be a random variable with cdf $F(x)$ and mgf $M(t)$.

Then,

$$\lim_{n\to\infty} M_n(t) = M(t) \qquad \Rightarrow \qquad \lim_{n\to} F_n(x) = F(x).$$

Notes:

1. Just for the record, the first limit only needs to hold only for all t in an open interval containing zero. The second limit needs to hold at all points of continuity of F.
2. This result would go in "both directions" if moment generating functions always existed, but they don't. The mgf is defined as an expectation which is defined by an integral that sometimes doesn't diverges to ∞!
3. This is how we prove the Central Limit Theorem!

2.5 The Central Limit Theorem

The Central Limit Theorem (CLT)

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean μ and variance $\sigma^2 < \infty$.

Consider the sequence of random variables $\{Z_n\}$ where

$$Z_n := \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}.$$

Then $Z_n \xrightarrow{d} Z \sim N(0,1)$ as $n \to \infty$.

Note:

- The CLT requires a finite variance. (This implies a finite mean.)
- We will prove the CLT in the case that the mgf exists. (Recall that the mgf is an expected value. We say that an expected value "exists" if it is finite.)
- The CLT can be proved more generally by using "characteristic functions" in place of mgfs. The characteristic function, $\phi(t)$, for a rv X is defined to be: $\phi(t) = \mathsf{E}[e^{itX}]$ where i is the imaginary unit. Characteristic functions always exist but since complex variables are not a prerequisite for this text we will only prove the Central Limit Theorem using moment generating functions.

2.5.1 Proof of the Central Limit Theorem

First note that

$$Z_n = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} = \frac{n(\frac{1}{n}\sum_{i=1}^n X_i - \mu)}{\sqrt{n}\sigma} = \frac{(\sum_{i=1}^n X_i - n\mu)}{\sqrt{n}\sigma} = \frac{\sum_{i=1}^n (X_i - \mu)}{\sqrt{n}\sigma}$$

For notational simplicity, let $m(t)$ denote the mgf of $X_i - \mu$ for any particular i in $\{1, \ldots, n\}$. Note that $m(0) = 1$, $m'(0) = \mathsf{E}[X_i - \mu] = 0$, and $m''(0) = \mathsf{E}[(X_i - \mu)^2] = \sigma^2$.

Expanding $m(t)$ about 0 using Taylor's formula, there exists a ξ ("ksi"), between 0 and t such that

$$m(t) = m(0) + m'(0)t + \frac{m''(\xi)t^2}{2}.$$

Note that we did not say $0 < \xi < t$ because t could be positive or negative.

Plugging in $m(0) = 1$ and $m'(0) = \mathsf{E}[X - \mu] = 0$, we have

$$m(t) = 1 + \frac{m''(\xi)t^2}{2}$$

and by adding and subtracting $\sigma^2 t^2/2$, this becomes

$$m(t) = 1 + \frac{\sigma^2 t^2}{2} + \frac{(m''(\xi) - \sigma^2)t^2}{2}.$$

Now we are ready to find the mgf of Z_n. Since the rv's $(X_i - \mu)$ are independent, it is easy to find the mgf of the sum $\sum(X_i - \mu)$.

$$\begin{aligned}
Z_n = \frac{\sum_{i=1}^n (X_i - \mu)}{\sqrt{n}\sigma} \quad \Rightarrow \quad M_{Z_n}(t) &= M_{\sum(X_i-\mu)}\left(\tfrac{t}{\sqrt{n}\sigma}\right) \\
&= \textstyle\prod_{i=1}^n M_{X_i-\mu}\left(\tfrac{t}{\sqrt{n}\sigma}\right) \\
&= \left[M_{X_1-\mu}\left(\tfrac{t}{\sqrt{n}\sigma}\right)\right]^n \\
&= \left[m\left(\tfrac{t}{\sqrt{n}\sigma}\right)\right]^n \\
&= \left[1 + \tfrac{\sigma^2 t^2}{2n\sigma^2} + \tfrac{(m''(\xi)-\sigma^2)t^2}{2n\sigma^2}\right]^n \\
&= \left[1 + \tfrac{t^2}{2n} + \tfrac{(m''(\xi)-\sigma^2)t^2}{2n\sigma^2}\right]^n
\end{aligned}$$

for ξ between 0 and $\frac{t}{\sqrt{n}\sigma}$.

Remember that our goal is to consider $\lim_{n\to\infty} M_{Z_n}(t)$.

Well, as $n \to \infty$, $t/\sqrt{n}\sigma \to 0$ so $\xi \to 0$ since $\xi < t/\sqrt{n}\sigma$.

Now, $m''(t)$ is continuous at 0, and $m''(0) = \mathsf{E}[(X - \mu)^2] = \sigma^2$, so

$$\lim_{n\to\infty} (m''(\xi) - \sigma^2) = 0.$$

Therefore,

$$M_{Z_n}(t) = \left[1 + \frac{t^2}{2n} + \frac{d(n)}{n}\right]^n,$$

where $d(n) \to 0$ as $n \to \infty$.

So we have that

$$\lim_{n\to\infty} M_{Z_n}(t) = e^{t^2/2}$$

which means that

$$Z_n \xrightarrow{d} Z \sim N(0,1).$$

□

2.5.2 Asymptotic Normality

Definition 2.5.1

We say that X_n has an **asymptotically normal** distribution with mean μ_n and variance σ_n^2 if

$$\frac{X_n - \mu_n}{\sigma_n} \xrightarrow{d} N(0,1).$$

(Note that, even without the limit, that thing on the left side has mean 0 and variance 1.)

Notation

If X_n has an asymptotically normal distribution with mean μ_n and variance σ_n^2, we will write

$$X_n \overset{asmyp}{\sim} N(\mu_n, \sigma_n^2).$$

It is not correct to write $X_n \xrightarrow{d} N(\mu_n, \sigma_n^2)$ since the first part "$X_n \xrightarrow{d}$" implies that n is going to ∞ and so there should not be any n's on the right-hand side of the arrow!

Example 2.5.1

According to the Central Limit Theorem, if $X_1, X_2, \ldots, X_n$ is a random sample from any distribution with mean μ and variance $\sigma^2 < \infty$,

$$\overline{X} = \overline{X}_n \overset{asymp}{\sim} N(\mu, \sigma^2/n).$$

Example 2.5.2

According to the Central Limit Theorem, if $X_1, X_2, \ldots, X_n$ is a random sample from any distribution with mean μ and variance $\sigma^2 < \infty$, we also have that $\sum_{i=1}^n X_i$ is asymptotically normal since

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \overset{d}{\to} Z \sim N(0,1)$$

and since

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{\frac{1}{n}\sum X_i - \mu}{\sigma/\sqrt{n}} = \frac{n}{n}\frac{\frac{1}{n}\sum X_i - \mu}{\sigma/\sqrt{n}} = \frac{\sum X_i - n\mu}{\sqrt{n}\sigma/}$$

Therefore, we see that

$$\sum_{i=1}^n X_i \overset{asymp}{\sim} N(n\mu, n\sigma^2).$$

2.5.3 A Numerical Example for the CLT

Let $\overline{X}$ denote the mean of a random sample of size 100 from the $\chi^2(50)$ distribution.

Give an approximate value for $P(49 < \overline{X} < 51)$.

Answer:

Just for the record, we can give an exact value since

$$X_1, X_2, \ldots, X_{100} \overset{iid}{\sim} \chi^2(50) = \Gamma(25, 1/2) \qquad \Rightarrow \qquad \sum_{i=1}^{100} X_i \sim \Gamma(2500, 1/2).$$

So,

$$\begin{aligned} P(49 < \overline{X} < 51) &= P(4900 < \textstyle\sum X_i < 5100) \\ &= \int_{4900}^{5100} \frac{1}{\Gamma(2500)} \left(\frac{1}{2}\right)^{2500} x^{2499} e^{-\frac{1}{2}x}\, dx \approx 0.6827. \end{aligned}$$

(This computation requires $2,500$ iterations of integration by parts and was done using Mathematica!)

With the Central Limit Theorem, we know that the distribution of $\overline{X}$ is approximately normal for large sample sizes. In statistics textbooks, a sample size of $n \geq 30$ or sometimes $n \geq 40$, is generally considered "large".

The mean for $\overline{X}$, normal or not, is

$$\mu_{\overline{X}} = \mu = 25/(1/2) = 50$$

and the variance is

$$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n} = \frac{25/(1/4)}{100} = 1.$$

We can "standardize" $\overline{X}$ into something with mean 0 and variance 1, which, for large samples, will be approximately a standard normal random variable. (See the gray box at the end of this Section.) So,

$$\begin{aligned} P(49 < \overline{X} < 51) &= P\left(\tfrac{49-50}{10/10} < \tfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} < \tfrac{51-50}{10/10}\right) \\ &= P\left(-1 < \tfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} < 1\right) \\ &\approx P(-1 < Z < 1) \end{aligned}$$

where $Z \sim N(0,1)$ is a standard normal random variable.

Now, using the z-table in Appendix C, we have

$$P(-1 < Z < 1) \overset{contin}{=} P(-1 < Z \leq 1) = P(Z \leq 1) - P(Z \leq -1)$$

$$\overset{z-table}{=} 0.8413 - (1 - 0.8413) = 0.6826.$$

This seems to be a pretty decent approximation!

Computing Probabilities for the Normal Distribution

Let $X \sim N(\mu, \sigma^2)$. Then the pdf for X is

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}.$$

While we can integrate the pdf from $-\infty$ to ∞ by converting to polar coordinates, we can not integrate over other regions very easily. In particular, there is no closed-form expression for the cdf of the normal distribution. Table C.1 in Appendix C gives numerical approximations to the cdf for the standard normal distribution. It is known as a "z-table".

Let $Z \sim N(0, 1)$. While the cdf for Z could be called F or F_Z, it is so special that it gets its own name. We define

$$\Phi(z) := P(Z \leq z).$$

To find, for example, $\Phi(2.47) = P(Z \leq 2.47)$ in Table C.1, we look down the column on the left side of the table to find the row labeled 2.4 and then move across the row until we are in the column labeled 0.07. At the intersection of this row and column we find the desired probability

$$\Phi(2.47) = P(Z \leq 2.47) = 0.9932.$$

To evaluate the cdf at negative values, use the symmetry of the $N(0, 1)$ distribution about 0.

To find probabilities for other normal distributions, we use the fact that any linear combination of normal random variables is again normal. This can easily be shown using moment generating functions and is left as an exercise at the end of this Chapter. For example, if $X_1, X_2, \ldots, X_n$ each have normal distributions (with possibly different parameters) then so do

- $a_1X_1 + a_2X_2 + \cdots + a_nX_n$ for any constants $a_1, a_2, \ldots, a_n$
- $a_1X_1 + a_2$ for any constants a_1 and a_2

  and
- $\frac{b_1X_1+b_2}{b_3}$ for constants b_1, b_2, and b_3. (This is a special case of the previous example with $a_1 = b_1/b_3$ and $a_2 = b_2/b_3$.

So, if $X \sim N(\mu, \sigma^2)$, we can easily verify that $(X - \mu)/\sigma$ has mean 0 and variance 1 and we now know that

$$Z := \frac{X - \mu}{\sigma} \sim N(0, 1).$$

On the other hand, if $Z \sim N(0, 1)$, then $\sigma Z + \mu$ has mean μ and variance σ^2 and

$$X := \sigma Z + \mu \sim N(\mu, \sigma^2).$$

2.5.4 The Delta Method

Let $X_1, X_2, \ldots, X_n$ be a random sample from any distribution with mean μ and finite variance σ^2. By the Weak Law of Large numbers, we know that

$$\overline{X} \xrightarrow{P} \mu.$$

Since this implies convergence in distribution, we can also say that

$$\overline{X} \xrightarrow{d} \mu.$$

By the Continuous Mapping Theorem we can say that

$$\overline{X}^2 \xrightarrow{d} \mu^2.$$

The question now is what can we say about the asymptotic distribution of something like $\overline{X}^2$?

Note that we are not asking about the convergence in distribution of $\overline{X}^2$. Recall that the Central Limit Theorem says that $\overline{X}$ is asymptotically normal in the sense that

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1).$$

This was denoted like this.

$$\overline{X} \overset{asymp}{\sim} N(\mu, \sigma^2/n).$$

Can we say something like

$$\overline{X}^2 \overset{asymp}{\sim} N(a_n, b_n)$$

for some sequences $\{a_n\}$ and $\{b_n\}$?

The answer to this question is given by the "**Delta Method**". As our interest with convergence of random variables is ultimately in terms of evaluating estimators (which are random variables) we state the Delta Method in terms of estimators.

The Delta Method

Suppose that we have a sequence of estimators $\widehat{\theta}_n$ for θ such that

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma^2).$$

Suppose further that $g : \mathbb{R} \to \mathbb{R}$ is continuously differentiable in a neighborhood of θ.

Then

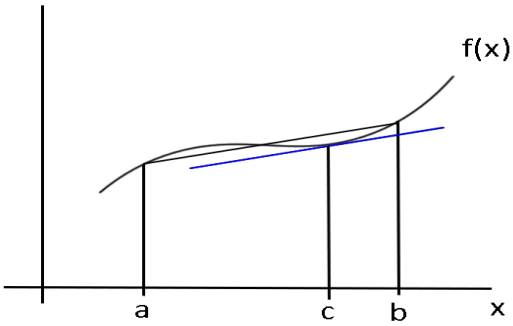
$$\sqrt{n}(g(\widehat{\theta}_n) - g(\theta)) \xrightarrow{d} N\left(0, \sigma^2[g'(\theta)]^2\right).$$

Proof :

Recall the Mean Value Theorem (MVT) from Calculus that says if f is a continuous function on a closed interval $[a, b]$ that is differentiable on (a, b), there exists a point c in (a, b) such that

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

Note that it doesn't really matter that $a < b$. These are two endpoints of an interval and there is a c somewhere between them.



Here, we will take f to be g, b to be $\widehat{\theta}_n$, and a to be θ. Note that $\widehat{\theta}_n$ is a random variable that, when observed may be less than or greater than θ. However, in one direction or the other, θ and $\widehat{\theta}_n$ form an interval with a random endpoint. By the MVT, there is a (random!) point $\widetilde{\theta}_n$ between them such that

$$g(\widehat{\theta}) - g(\theta) = g'(\widetilde{\theta}_n)(\widehat{\theta}_n - \theta).$$

Now

$$\sqrt{n}(g(\widehat{\theta}_n) - g(\theta)) = g'(\widetilde{\theta}_n)\sqrt{n}\,(\widehat{\theta}_n - \theta). \tag{2.5.10}$$

By assumption, we know that

$$\sqrt{n}\,(\widehat{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma^2).$$

What can we say about $g'(\widetilde{\theta}_n)$?

Note that

$$\widehat{\theta}_n - \theta = \frac{1}{\sqrt{n}}\sqrt{n}\,(\widehat{\theta}_n - \theta) \xrightarrow{d} 0 \cdot N(0, \sigma^2) = 0$$

by Slutsky's Theorem since $1/\sqrt{n} \xrightarrow{P} 0$.

So, we have $\widehat{\theta}_n \xrightarrow{d} \theta$ which implies that

$$\widehat{\theta}_n \xrightarrow{P} \theta$$

by Theorem 2.4.2, since θ is a constant.

Since $\widetilde{\theta}_n$ is stuck between $\widehat{\theta}_n$ and θ, and since $\widehat{\theta}_n \xrightarrow{P} \theta$, we must have $\widetilde{\theta}_n \xrightarrow{P} \theta$. (Can you prove this formally?)

So, by property 4 of Theorem 2.3.3, we have that

$$g'(\widetilde{\theta}_n) \xrightarrow{P} g'(\theta)$$

since g' is continuous.

Returning to equation 2.5.10, we have, by Slutsky's Theorem,

$$\sqrt{n}(g(\widehat{\theta}_n) - g(\theta)) = \underbrace{g'(\widetilde{\theta}_n)}_{\downarrow P,\ g'(\theta)} \underbrace{\sqrt{n}\,(\widehat{\theta}_n - \theta)}_{\downarrow d,\ N(0,\sigma^2)} \xrightarrow{d} g'(\theta) \cdot N(0,\sigma^2) = N\left(0, \sigma^2[g'(\theta)]^2\right),$$

as desired.

(Here we have used the fact that a constant c times a random variable $X \sim N(\mu, \sigma^2)$ has a $N(c\mu, c^2\sigma^2)$ distribution.)

■

Example 2.5.3

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean μ and finite variance σ^2. We know by the Central Limit Theorem that

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1).$$

This implies that

$$\sqrt{n}(\overline{X} - \mu) \xrightarrow{d} \sigma \cdot N(0,1) = N(0,\sigma^2).$$

Using the Delta Method with $g(x) = x^2$, we now know that

$$\sqrt{n}(\overline{X}^2 - \mu^2) \xrightarrow{d} N(0, \sigma^2[g'(\mu)]^2) = N(0, 4\mu^2\sigma^2).$$

Thus, we have that

$$\overline{X}^2 \overset{asymp}{\sim} N(\mu^2, 4\mu^2\sigma^2/n).$$

2.6 The Sample Variance

Let $X_1, X_2, \ldots, X_n$ be a random sample from any distribution with mean μ and variance σ^2. We have seen that $\hat{\mu} = \overline{X}$ is an unbiased estimator for μ.

Suppose that we want to estimate the variance. The interpretation of the variance ($\sigma^2 = \mathsf{E}[(X - \mu)^2]$) as the mean squared deviation from the sample mean μ leads us to the kind of natural or common sense estimator that involves averaging the squared deviations in the sample from the sample mean as follows.

$$\widehat{\sigma}_1^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n}.$$

Some books/people refer to this quantity as the **sample variance**. However, many other books/people define the sample variance to be

$$\widehat{\sigma}_2^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1}$$

because it is an unbiased estimator for σ^2 while $\widehat{\sigma}_1^2$ is a biased estimator. (See Exercise 1 in this Chapter.)

In this text, **we will always be referring to the unbiased estimator when talking about the sample variance**. The sample variance is denoted by the symbol S^2. So, in this text we have

$$S^2 := \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1}.$$

With a little manipulation, i.e. expanding out the square in the numerator and running the sum through, S^2 can also be written as

$$S^2 = \frac{\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2/n}{n-1}.$$

The first version is preferable from the point of view of interpretation. We are averaging squared distances of all data point to the same mean. The second version is preferable from the computational point of view as it can be computed knowing only two summary statistics, $\sum X_i$ and $(\sum X_i)^2$ from the data set.

S^2 is a random variable that is used as an estimator for σ^2 just as the sample mean $\overline{X}$ is used as an estimator for the mean μ. As stated, $\mathsf{E}[S^2] = \sigma^2$, and it is possible (albeit tedious!) to show that

$$Var[S^2] = \frac{\mu_4'}{n} - \frac{(\sigma^2)^2(n-3)}{n(n-1)}$$

where μ_4' is the *forth central moment* given by

$$\mu_4 = \mathsf{E}[(X-\mu)^4].$$

For distributions with a finite fourth central moment, which implies that lower order moments like $\sigma^2 = \mathsf{E}[(X-\mu)^2]$ are also finite, we can see that $Var[S^2] \to 0$ as $n \to \infty$. Thus, we have in this case that

$$S^2 \xrightarrow{P} \sigma^2$$

by Theorem 2.3.1.

We now also know that the biased version of the sample variance also converges in probability to σ^2 because

$$\widehat{\sigma}_2^2 = \frac{n-1}{n} S^2 \xrightarrow{P} 1 \cdot \sigma^2 = \sigma^2$$

since $(n-1)/n \xrightarrow{P} 1$ and $S^2 \xrightarrow{P} \sigma^2$.

2.7 Postscript: Convergence in Probability for Vector-Valued Random Variables

The absolute value in Definition 2.3.1 of convergence in probability is the Euclidean norm

$$||X_n - X|| = \sqrt{(X_n - X)^2} = |X_n - X|.$$

This observation allows us to easily generalize the definition for vector-valued random variables. If $Y_n = (X_{n1}, X_{n2}, \ldots, X_{nk})$ and $Y = (X_1, X_2, \ldots, X_k)$, then

$$||Y_n - Y|| = \sqrt{\sum_{i=1}^{k} (X_{ni} - X_i)^2}.$$

The definition for convergence in probability for a vector-valued sequence of random variables is as follows.

Definition 2.7.1

A vector-valued sequence of random variables $\{Y_n\}$ where $Y_n = (X_{n1}, X_{n2}, \ldots, X_{nk})$ **converges in probability** to a vector-valued random variable $Y = (X_1, X_2, \ldots, X_k)$ if, for any $\varepsilon > 0$, we have

$$\lim_{n\to\infty} P(||Y_n - Y|| > \varepsilon) = 0.$$

(Equivalently if $\lim_{n\to\infty} P(||Y_n - Y|| \leq \varepsilon) = 1$.)
We write $Y_n \xrightarrow{P} Y$.

Fortunately, this can be characterized in terms of one-dimensional sequence convergence.

Theorem 2.7.1

The sequence of vector-valued random variables $\{Y_n\}$ where $Y_n = (X_{n1}, X_{n2}, \ldots, X_{nk})$ converges in probability to a vector-valued random variable $Y = (X_1, X_2, \ldots, X_k)$ if and only if $X_{nj} \xrightarrow{P} X_j$ for all $j = 1, 2, \ldots, k$.

Proof :

($\Rightarrow$) Suppose that $Y_n \xrightarrow{P} Y$.

For $j = 1, 2, \ldots, k$, we have

$$P(|X_{nj} - X_j| > \varepsilon) \leq P\left(\sqrt{\sum_{i=1}^{k}(X_{ni} - X_i)^2} > \varepsilon\right) = P(||Y_n - Y|| > \varepsilon)$$

Thus we have

$$\lim_{n\to\infty} P(|X_{nj} - X_j| > \varepsilon) \leq \lim_{n\to\infty} P(||Y_n - Y|| > \varepsilon) = 0$$

since $Y_n \xrightarrow{P} Y$. This implies that

$$\lim_{n\to\infty} P(|X_{nj} - X_j| > \varepsilon) = 0,$$

as desired.

($\Leftarrow$) Now suppose that $X_{nj} \xrightarrow{P} X_j$ for all $j = 1, 2, \ldots, k$.

Write the event that

$$\sqrt{\sum_{i=1}^{k}(X_{ni} - X_i)^2} > \varepsilon \tag{2.7.11}$$

"occurs" as

$$\left\{ \sqrt{\sum_{i=1}^{k} (X_{ni} - X_i)^2} > \varepsilon \right\}.$$

One line here is simply an inequality while the other represents the "event that the inequality happens".

Note that (2.7.11) implies that at least one of the $\sqrt{(X_{ni} - X_i)^2} = |X_{ni} - X_i|$ must be greater than ε/k. (If not, there is no way that the sum can be greater than ε.)

In event notation, this is

$$\left\{ \sqrt{\sum_{i=1}^{k} (X_{ni} - X_i)^2} > \varepsilon \right\} \subseteq \bigcup_{i=1}^{k} \{|X_{ni} - X_i| > \varepsilon/k\}.$$

This implies that

$$\begin{aligned} P\left(\sqrt{\sum_{i=1}^{k} (X_{ni} - X_i)^2} > \varepsilon \right) &\leq P\left(\bigcup_{i=1}^{k} (|X_{ni} - X_i| > \varepsilon/k \right) \\ &\leq \sum_{i=1}^{k} P(|X_{ni} - X_i| > \varepsilon/k). \end{aligned}$$

(Think Venn diagrams. For example, $P(A \cup B) = P(A) + P(B) - P(A \cap B) \leq P(A) + P(B)$.)

Take the limit as $n \to \infty$ on both sides:

$$\begin{aligned} \lim_{n\to\infty} P\left(\sqrt{\sum_{i=1}^{k} (X_{ni} - X_i)^2} > \varepsilon \right) &\leq \lim_{n\to\infty} \sum_{i=1}^{k} P(|X_{ni} - X_i| > \varepsilon/k) \\ &= \sum_{i=1}^{k} \lim_{n\to\infty} P(|X_{ni} - X_i| > \varepsilon/k) \\ &= \sum_{i=1}^{k} 0 = 0 \end{aligned}$$

since $X_{ni} \xrightarrow{P} X_i$.

Thus, we have that

$$\lim_{n\to\infty} P(||Y_n - Y|| > \varepsilon) = \lim_{n\to\infty} P\left(\sqrt{\sum_{i=1}^{k} (X_{ni} - X_i)^2} > \varepsilon \right) = 0$$

and so $Y_n \xrightarrow{P} Y$.

■

Chapter 2 Exercises

1. Let $X_1, X_2, \ldots, X_n$ be a random sample from any distribution with mean μ and variance σ^2.

   Consider the sample variance

   $$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}.$$

   and a biased version of the sample variance

   $$S_2^2 := \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}.$$

   (a). Show that

   $$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2/n.$$

   (b). Use part (a) to show that S^2 is an unbiased estimator of σ^2.

   (c). Use part (b) to quickly find the expected value of S_2^2. (*Hint: Don't make this too much work!*)

2. Let $X_1, X_2, \ldots, X_n$ be independent random variables with $X_i \sim N(\mu_i, \sigma_i^2)$ for $i = 1, 2, \ldots, n$. Let $a_1, a_2, \ldots, a_n$ and a be constants.

   Use moment generating functions to show that

   $$Y := \sum_{i=1}^{n} a_i X_i + a$$

   also has a normal distribution. What are the mean and variance of Y?

3. Let $Y_1, Y_2, \ldots, Y_n$ denote a random sample from a distribution with mean μ and variance σ^2. Consider the following two estimators for μ:

   $$\widehat{\mu}_1 = \overline{Y} \qquad \text{and} \qquad \widehat{\mu}_2 = \frac{1}{4}Y_1 + \frac{Y_2 + \cdots + Y_{n-1}}{2(n-2)} + \frac{1}{4}Y_n.$$

   (a). Show that $\widehat{\mu}_2$ is unbiased for μ.

   (b). Compute and compare $Var[\widehat{\mu}_1]$ and $Var[\widehat{\mu}_2]$

4. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a population whose pdf is given by

   $$f(x) = \alpha x^{\alpha-1}/\theta^{\alpha} \cdot I_{(0,\theta)}(x)$$

   where $\alpha > 0$ is a known fixed value, but θ is unknown.

   Since θ is the right endpoint of the support for this distribution, we will consider estimating it with the maximum value in our sample. That is, we will consider the estimator

   $$\hat{\theta} = X_{(n)} = \max(X_1, X_2, \ldots, X_n).$$

   (a). Show that $\hat{\theta}$ is a biased estimator of θ.

   (b). Explain why it makes sense that $\hat{\theta}$ is biased for θ and why the "direction" of the bias makes sense.

   (c). Find a multiple of $\hat{\theta}$ that is an unbiased estimator of θ.

(d). Find the variance of your estimator from part (c).

5. Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from the $geom_0(p)$ distribution. (This is the geometric distribution that "starts from 0".)

   (a). Consider the sample mean $\overline{X}$. To what does this converge in probability? Explain.

   (b). Consider the new random variables $Y_1, Y_2, \ldots, Y_n$ where $Y_i = I_{\{X_i > 0\}}$. To what does $\overline{Y}$ converge in probability? Explain.

6. Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from any distribution with mean μ and variance σ^2. Define $\mu_4 = \mathsf{E}[(X_i - \mu)^4]$. This is called the "fourth central moment". One can show that

$$Var[S^2] = \frac{\mu_4}{n} - \frac{(n-3)(\sigma^2)^2}{n(n-1)}.$$

   (You do not have to show this.) Assuming that the fourth central moment is finite, show that S^2 is a consistent estimator of σ^2.

7. Let $X_{(n)} = \max(X_1, X_2, \ldots, X_n)$ where $X_1, X_2, \ldots, X_n$ is a random sample from any continuous distribution that has cdf $F(x)$ and pdf $f(x)$. Define $Z_n = n[1 - F(X_{(n)})]$. Find the limiting distribution (convergence in distribution) of Z_n. You may assume that F is invertible.

8. Suppose that X_n has the $\Gamma(n, \beta)$ distribution. Investigate the convergence in probability of $Y_n := X_n/n$. (That is, to what does this converge in probability?)

9. Let $X_1, X_2, \ldots, X_n$ be a random sample from the Bernoulli distribution with parameter p. Consider $Y_n = \sum_{i=1}^n X_i$. Use moment generating functions to accomplish the following.

   (a). Find the distribution of Y_n. (Name it!)

   (b). Find the limiting distribution of Y_n (Name it!) in the case that p is decreased while n is increased in such a way that $pn = \mu$ for some fixed constant $\mu > 0$.

10. Consider a random sample $X_1, X_2, \ldots X_n$ from the Pareto distribution with parameter $\gamma = 1$.

    (a). Find the limiting distribution of $Y_n = nX_{(1)}$.

    (b). Does the limiting distribution of $W_n = X(1)$ exist? If so, find it.

    (c). Does the limiting distribution of $V_n = X(n)$ exist? If so, find it.

11. Let $\{a_n\}_{n=1}^{\infty}$ be a sequence of real (and not random) numbers. Suppose that

$$\lim_{n \to \infty} a_n = a$$

    for some real number a. Show that $a_n \xrightarrow{P} a$.

12. Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from the $geom_0(p)$ distribution. (This is the geometric distribution that "starts from 0".)

    (a). Consider the sample mean $\overline{X}$. To what does this converge in probability? Explain.

    (b). Consider the new random variables $Y_1, Y_2, \ldots, Y_n$ where $Y_i = I_{\{X_i > 0\}}$. To what does $\overline{Y}$ converge in probability? Explain.

13. Consider a random sample of size 65 from the distribution with pdf $f(x) = \frac{2}{(1+x)^3} I_{(0,\infty)}(x)$. Compute the approximate probability that more than 40 of the observations are less than 3.

14. (a). Inequalities are useful for showing convergence in probability. Let X be a random variable with finite mean and let g be a convex (concave up) function. Show that

$$g(\mathsf{E}[X]) \leq \mathsf{E}[g(X)].$$

This is known as Jensen's inequality.

(b). Now suppose that g is concave (concave down). Use part (a) to show that

$$g(\mathsf{E}[X]) \geq \mathsf{E}[g(X)].$$

15. Suppose that $\{X_n\}$ is a sequence of random variables and that X is another random variable such that

$$\lim_{n\to\infty} \mathsf{E}\left[(X_n - X)^2\right] = 0.$$

Show that $X_n \xrightarrow{P} X$.

16. Suppose that $X_1, X_2, \ldots$ is a sequence of random variables such that

$$\sqrt{n}(X_n - \mu) \xrightarrow{d} N(0,1).$$

Show that this implies that $X_n \xrightarrow{P} \mu$.

17. Let $X_1, X_2, \ldots, X_n$ be a random sample from any distribution with finite fourth moments. Consider the sample variance. We already know that S^2 is an unbiased estimator of σ^2. You COULD show that

$$Var[S^2] = \frac{1}{n}\left(\mu_4' - \frac{n-3}{n-1}\sigma^4\right)$$

where μ_4' is the fourth "central" moment $\mu_4' := \mathsf{E}[(X-\mu)^4]$. Then, since this variance goes to 0 for this unbiased estimator of σ^2, we know, by Theorem 2.3.1 that $S^2 \xrightarrow{P} \sigma^2$.

(a). Show that

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 + \frac{1}{n}S^2$$

(b). Show that

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\overline{X} - \mu)^2 + \frac{1}{n}S^2$$

(c). Show that

$$\sqrt{n}(S^2 - \sigma^2) = \sqrt{n}\left[\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{n} - \sigma^2\right] - \sqrt{n}(\overline{X}-\mu)(\overline{X}-\mu) + \frac{1}{\sqrt{n}}S^2$$

(d). Use part (c) to find the asymptotic distribution of S^2.

18. Consider the sequence of random variables $\{X_n\}$ where X_n takes on the value n with probability $1/n$ and 0 with probability $1 - 1/n$.

(a). Show that $\lim_{n\to\infty} \mathsf{E}[X_n] = 1$.

(b). Show that $X_n \xrightarrow{P} 0$.

(c). Explain why this sequence of random variables does not contradict Theorem 2.3.2.

Chapter 3 A "Mostly Normal" Introduction to Confidence Intervals

Still and seemingly forever on the horizon is how to find estimators for parameters of distributions. We have found a few "common sense estimators" just by guessing and making some adjustments to our guesses. We have looked at few properties of estimators, finding their expected values, variances, and MSEs, and we have shown that some have good large sample properties in the sense that they convergence in probability and/or distribution to the parameter of interest. Soon, we will stop guessing at estimators. Soon.

In this Chapter, we give a brief introduction to confidence intervals. Most of the examples will involve the normal distribution but it is crucial that you pay attention to the process as opposed to a "resulting formula in a box" so that you will know what to do when not working with a normal distribution.

3.1 A Confidence Interval for the Mean of a Normal Distribution: σ^2 is Known

Suppose that we have $X_1, X_2, \ldots, X_n$, a random sample from the $N(\mu, \sigma^2)$ where σ^2 is known but μ is unknown.[1]

The most natural (what we might call "common sense") estimator of μ is the sample mean:

$$\widehat{\mu} = \overline{X}.$$

We are now going to move from this **point estimator** to an **interval estimator** of the form $\overline{X}$ "plus or minus something". We will make this idea more precise soon.

Since $X_1, X_2, \ldots, X_n$ are all normal random variables, $\overline{X}$ is also normal. This is because a linear combination of normal random variables, whether iid or not, is still normal[2] as seen in the Exercises in Chapter 2.

If $X_1, X_2, \ldots, X_n$, is a random sample from any distribution with mean μ and variance σ^2, we have seen that the sample mean $\overline{X}$ has mean

$$\mu_{\overline{X}} = \mathsf{E}[\overline{X}] = \mathsf{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathsf{E}[X_i] \stackrel{iid}{=} \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}(\mu + \mu + \cdots + \mu) = \mu$$

[1] It may seem kind of silly to assume that σ^2 is known when μ is unknown. Most likely, neither are known. This example is just a building block for more realistic cases!

[2] There are a few exceptions. For example, if X has a normal distribution, then $X - X = 0$ and does not have a normal distribution!

and variance

$$\sigma^2_{\overline{X}} = Var[\overline{X}] = Var\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \frac{1}{n^2}Var[\sum_{i=1}^n X_i]$$

$$\overset{indep}{=} \frac{1}{n^2}\sum_{i=1}^n Var[X_i] \overset{ident}{=} \frac{1}{n^2}\sum_{i=1}^n \sigma^2 = \frac{1}{n^2}\cdot n\sigma^2 = \sigma^2/n.$$

So, when $X_1, X_2, \ldots, X_n$ are iid $N(\mu, \sigma^2)$ distributed, we have that

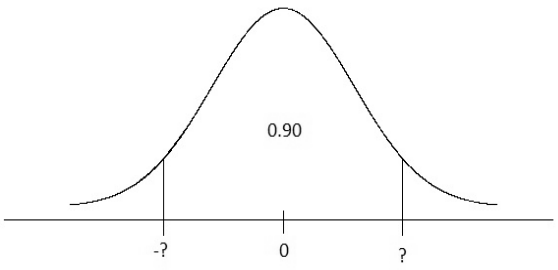
$$\overline{X} \sim N(\mu, \sigma^2/n).$$

Note that this is **not** from using the Central Limit Theorem. This is using the fact that $\overline{X}$ is a linear combination of normal random variables.

We now know that we can "standardize" $\overline{X}$ into a standard normal random variable.

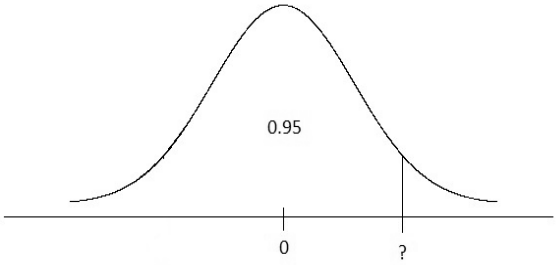
$$Z \overset{def}{=} \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

Consider trying to find two values that capture 90% of the area under a standard normal curve.



We've drawn these values as symmetric around zero for a reason we will discuss later.

Note that 10% of the area is outside those cutoff values, with 5% on each end. Therefore, the value "?" cuts off area 0.95 to its left.



"Reverse reading" our standard normal table, (looking for the probability in the body of the table in Appendix C) we see that the cutoff value "?" is about 1.645. (In this particular case, the value we were looking for fell exactly between 1.64 and 1.65 and so we took the value in the middle. In general, just go for the closest value. Interpolation is not worth much in terms of accuracy.)

So, we know that $P(-1.645 < Z < 1.645) = 0.90$. Thus,

$$\begin{aligned} 0.90 &= P(-1.645 < Z < 1.645) \\ &= P\left(-1.645 < \tfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} < 1.645\right) \end{aligned}$$

Solving for μ in the middle, we get

$$\begin{aligned} 0.90 &= P\left(-1.645\tfrac{\sigma}{\sqrt{n}} < \overline{X} - \mu < 1.645\tfrac{\sigma}{\sqrt{n}}\right) \\ &= P\left(-1.645\tfrac{\sigma}{\sqrt{n}} - \overline{X} < -\mu < 1.645\tfrac{\sigma}{\sqrt{n}} - \overline{X}\right) \\ &= P\left(\overline{X} - 1.645\tfrac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.645\tfrac{\sigma}{\sqrt{n}}\right) \end{aligned}$$

We are used to seeing random variables in the middle of inequalities. In this case however, μ is a constant, not a random variable and we have put it between two random endpoints with probability 0.90.

We say that a 90% confidence interval for μ is given by

$$\left(\overline{X} - 1.645\frac{\sigma}{\sqrt{n}}, \overline{X} + 1.645\frac{\sigma}{\sqrt{n}}\right),$$

or, in shorthand notation

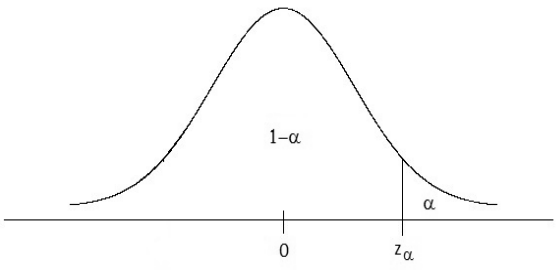
$$\overline{X} \pm 1.645\frac{\sigma}{\sqrt{n}}.$$

Interpretation:

Once a sample of some given size n is taken, we compute $\overline{x}$ and the above confidence interval endpoints turn into numbers. Once you have done this, there is no more probability involved. The true value of μ is either in that interval or it is not. The probability comes in from the sample. If you were to take another random sample of size n from this normal "population", you would get a different value for $\overline{x}$, which would give you different endpoints for the confidence interval. A third sample gives you yet another $\overline{x}$ which gives you a different confidence interval again. Repeating many times, you will have that the true value of μ is captured between the endpoints 90% of the time. Usually you only get one sample but if you are going (future) to collect one one and use the formula we derived in this Section, you have a 90% probability of that interval correctly capturing the true value of μ.

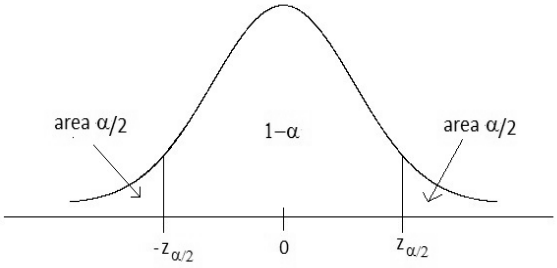
3.1.1 Standard Normal Critical Values

We will use the notation z_α to be the cutoff value for a standard normal curve that captures area α **to the right**.[3] This is also called a **critical value**.



If we want to capture a certain area, say $1-\alpha$, in the middle while distributing the remaining area α equally on both sides, the two cutoff values are $-z_{\alpha/2}$ and $z_{\alpha/2}$. (Note: In general, the values are denoted by $z_{1-\alpha/2}$ and $z_{\alpha/2}$, but, by symmetry of the $N(0,1)$ distribution about zero, we have that $z_{1-\alpha/2} = -z_{\alpha/2}$.)

[3]Be aware that some authors will use the notation z_α for a critical value that captures area α to the left!



For our 90% confidence interval, we put area 0.90 in the middle. i.e., $1 - \alpha = 0.9$. So, $\alpha/2 = 0.05$, and, in this new critical value notation, $z_{0.05} = 1.645$.

In general, if $X_1, X_2, \ldots, X_n$ is a random sample from the $N(\mu, \sigma^2)$ distribution with σ^2 known, a $100(1-\alpha)\%$ confidence interval for μ is given by

$$\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Rumination

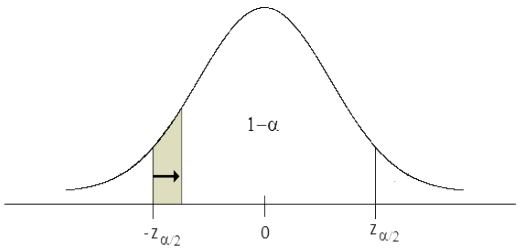
When constructing a confidence interval, we want to trap a desired area under a pdf between two values. Rather than talking about 95% area, for example, we want to be able to talk about a more general percentage like $100(1-\alpha)\%$. You might be wondering why we don't talk about an "α% confidence interval" or even a "100α% confidence interval". Confidence intervals are closely related to hypothesis testing, which we will be discussing in Chapter 4. Both can be used to convey the same information about a parameter. Historically, in the development of Statistics, hypothesis testing came first, and meaningful notation was long established by the time confidence intervals hit the scene. The notation $100(1-\alpha)\%$ for a confidence interval is used because it is aligned with hypothesis testing.

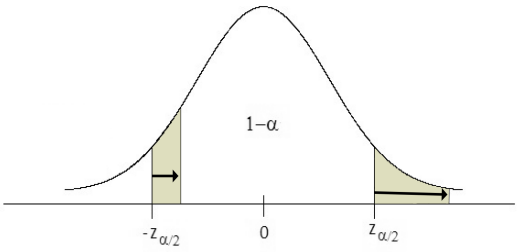
Question:

When finding a $100(1-\alpha)\%$ confidence interval, we put the desired area "in the middle" under the normal pdf. Can we put the area in other places?

Answer:

We can absolutely put the area in other places! Consider shifting the left cutoff a little bit to the right. Notice the area that is lost by this move. It is a decent amount of area because the pdf is relatively high in this place. ⇨





To compensate, the right cutoff, out in the right tail where there is less area because the pdf is lower, will have to shift further to make up the area lost by shifting ⇦ the left cutoff.

Overall, we will capture the same area but will have a longer confidence interval. Since you are giving this to someone (possibly yourself!) as your guess about where a parameter lives, it would be nice to give the shorter interval. While it is easy to see that the shortest interval when working with a normal distribution will be in the center, things are not always so clear when working with other distributions. In general, we will not be optimizing our intervals, with respect to length unless the choice of critical values that will do this is obvious.

3.2 An Approximate Large Sample Confidence Interval for the Mean of Any Distribution

Recall that the sample mean $\overline{X}$ for a random sample of size n from any distribution with mean μ and variance σ^2 has mean μ and variance σ^2/n.

Recall also that if $X_1, X_2, \ldots, X_n$ is a random sample from any distribution with finite variance, the Central Limit Theorem tells us that $\overline{X}$ has an approximately normal distribution for large n. (Greater than 30 is usually

considered "large".) So, we have the following.

> If $X_1, X_2, \ldots, X_n$ is a random sample from any distribution with mean μ and variance $\sigma^2 < \infty$, an approximate $100(1-\alpha)\%$ confidence interval for μ, when n is large, is given by
>
> $$\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

So far, we have assumed that σ^2, the variance of the distribution from which the random sample came, is known. If it is not known, we might consider estimating it with the sample variance

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}.$$

We have seen (Section 2.6) that $S^2 \xrightarrow{P} \sigma^2$. Thus, we can say that S^2 is getting "close" to σ^2, in some sense, as the sample size n gets large and we can use

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \overset{approx}{\sim} N(0,1)$$

to build a confidence interval for μ.[4]

More formally, we are using the fact that

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \xrightarrow{d} N(0,1).$$

To prove this, note that

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \cdot \frac{\sigma}{S} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \Big/ \sqrt{\frac{S^2}{\sigma^2}}\,. \tag{3.2.1}$$

We know that

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$$

and

$$S^2 \xrightarrow{P} \sigma^2 \quad \sqrt{S^2/\sigma^2} \xrightarrow{P} \sqrt{\sigma^2/\sigma^2} = 1$$

[4] Using "approx" above the "has the distribution" symbol is a very informal notation to convey that the random variable on the left side behaves approximately like the distribution on the right side. In contrast, using "asymp" means something more specific and formal.

by the Continuous Mapping Theorem for convergence in probability (see Property 4 of Theorem 2.3.3).

By Slutsky's Theorem, (3.2.1) tells us that

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} \xrightarrow{d} N(0,1)\cdot 1 = N(0,1).$$

Therefore,

> If $X_1, X_2, \ldots, X_n$ is a random sample from any distribution with mean μ and unknown variance σ^2, an approximate $100(1-\alpha)\%$ confidence interval for μ, when n is large, is given by
>
> $$\overline{X} \pm z_{\alpha/2}\frac{S}{\sqrt{n}}.$$

If the sample size is small, then we must be given a distribution since we can not appeal to the Central Limit Theorem to get an approximate distribution for $\overline{X}$. Otherwise, we are out of luck and can't find a confidence interval. We will discuss the procedure for known distributions, other than the normal distribution, later in Section 3.5. That said, we have only one case left to consider for the normal distribution.

3.3 A Confidence Interval for the Mean of a Normal Distribution: σ^2 is Unknown and Sample Size is Small

Suppose now that we have a random sample of size $n \leq 30$ from a normal distribution with unknown mean and variance.

Since the sample is from a normal distribution, we have that $\overline{X}$ has a normal distribution. In particular, $\overline{X} \sim N(\mu, \sigma^2/n)$. So, would probably start by "standardizing" $\overline{X}$ as follows.

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

We would then place this between two "z-critical values" capturing the appropriate area under the standard normal pdf and try to solve for μ "in the middle".

Unfortunately, we don't know σ^2, and if we try to "solve for μ in the middle", we will end up with a confidence interval with endpoints involving the unknown σ^2. This is not useful!

In Section 3.2, we used the estimator S^2 in place of σ^2. This was a decent idea since we saw in Section 2.6 that $S^2 \xrightarrow{P} \sigma^2$. Now, however, the sample size is small and the sample variance S^2 may not be a very good estimator of the true variance σ^2. It is not safe to say that the distribution of

$$\frac{\overline{X} - \mu}{S/\sqrt{n}}$$

is approximately $N(0,1)$.

We're going to need a new distribution.

3.3.1 The t Distribution

Suppose that $Z \sim N(0,1)$ and $W \sim \chi^2(n)$ are two independent random variables.

(Recall that we defined the $\Gamma(\alpha, \beta)$ distribution in Example 1.2.2 and the χ^2 distribution as the $\Gamma(n/2, 1/2)$ distribution.)

Define

$$T \sim \frac{Z}{\sqrt{W/n}}.$$

Then we can show (using the Jacobian method from Chapter 1) that the pdf of T is

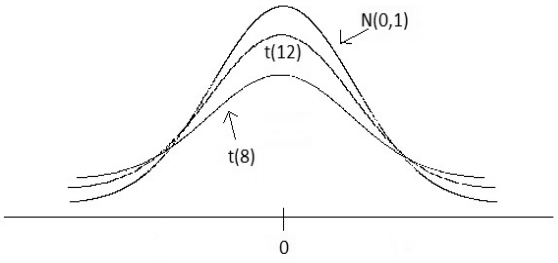
$$f_T(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \frac{1}{\sqrt{\pi n}} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}$$

for $-\infty < t < \infty$.

This distribution is known as the ***t*-distribution with n "degrees of freedom".** We write

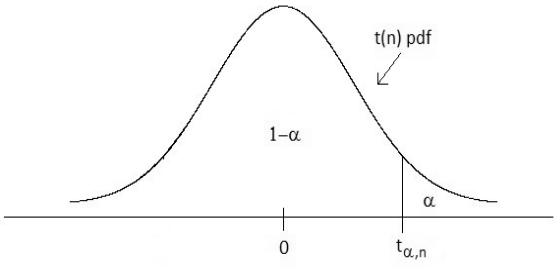
$$T \sim t(n).$$

The pdf is a symmetric bell curve centered at zero that is flatter than the $N(0,1)$ pdf. As n increases, it becomes closer and closer to the $N(0,1)$ pdf, as depicted.



As with the z curve, we would like to use the notation t_α to denote a "cutoff" or "critical value" that captures

area α to the right under the t curve. But, there are many t curves! We must include in our notation some reference to the degrees of freedom parameter. We will use $t_{\alpha,n}$.[5]



3.3.2 The Distribution of $(n-1)S^2/\sigma^2$ for a Normal Distribution

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the $N(\mu, \sigma^2)$ distribution.

Consider the quantity $\sum_{i=1}^n (X_i - \mu)^2$ expanded as

$$\begin{aligned} \sum_{i=1}^n (X_i - \mu)^2 &= \sum_{i=1}^n (X_i - \overline{X} + \overline{X} - \mu)^2 \\ &= \sum_{i=1}^n (X_i - \overline{X})^2 + 2(\overline{X} - \mu)\sum_{i=1}^n (X_i - \overline{X}) + \sum_{i=1}^n (\overline{X} - \mu)^2 \end{aligned} \tag{3.3.2}$$

Note that the middle term is zero since

$$\sum_{i=1}^n (X_i - \overline{X}) = \sum_{i=1}^n X_i - n\overline{X} = \sum_{i=1}^n X_i - n\frac{1}{n}\sum_{i=1}^n X_i = \sum_{i=1}^n X_i - \sum_{i=1}^n X_i = 0.$$

Note that the last term of (3.3.2) is $n(\overline{X} - \mu)^2$.

So, (3.3.2) becomes

$$\sum_{i=1}^n (X_i - \mu)^2 = \sum_{i=1}^n (X_i - \overline{X})^2 + n(\overline{X} - \mu)^2.$$

Dividing through by σ^2, we have

$$\frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma^2} + \frac{n(\overline{X} - \mu)^2}{\sigma^2}. \tag{3.3.3}$$

Note that

1. The term on the left-hand side has a $\chi^2(n)$ distribution since

$$X_i \sim N(\mu, \sigma^2) \Rightarrow \frac{X_i - \mu}{\sigma} \sim N(0, 1)$$

[5] We could also use the notation $t_\alpha(n)$. Be aware that other authors may use either notation as a value that captures area α to the left.

$$\Rightarrow \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(1) \text{ by Exercise 6 in Chapter 1}$$

$$\Rightarrow \frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma^2} = \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(n) \text{ by Exercise 7 in Chapter 1}$$

2. The first term on the right-hand side of (3.3.3) is
$$\frac{(n-1)S^2}{\sigma^2}.$$

3. The second term on the right-hand side has a $\chi^2(1)$ distribution since

$$\overline{X} \sim N(\mu, \sigma^2/n) \Rightarrow \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \sim N(0,1)$$

$$\Rightarrow \frac{n(\overline{X} - \mu)^2}{\sigma^2} = \left(\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma}\right)^2 \sim \chi^2(1) \text{ by Exercise 6 in Chapter 1.}$$

For simplicity, rewrite (3.3.3) as

$$W = W_1 + W_2.$$

So $W \sim \chi^2(n)$, $W_2 \sim \chi^2(n)$, and W_1 is the thing we want the distribution for. By Exercise 4 of this Chapter, we have the rather surprising result that W_1 and W_2 are independent. (At first glance, this seems crazy since they both involve $\overline{X}$. At second glance, it seems less crazy because everything here is in terms of deviations or "distances" from $\overline{X}$ which won't depend on the actual location of $\overline{X}$.)

Now

$$M_W(t) = M_{W_1+W_2}(t) \overset{indep}{=} M_{W_1}(t) \cdot W_2(t).$$

Using the appropriate χ^2 moment generating functions, we find that

$$M_{W_1}(t) = \frac{M_W(t)}{M_{W_2(t)}} = \frac{\left(\frac{1/2}{1/2-t}\right)^{n/2}}{\left(\frac{1/2}{1/2-t}\right)^{1/2}} = \left(\frac{1/2}{1/2-t}\right)^{(n-1)/2}$$

which is the moment generating function for the $\chi^2(n-1)$ random variable.

Therefore,

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

Awesome!

Rumination

At this point, we have two distributions with a "degrees of freedom" parameter. We have just seen one of these distributions connected to the sample variance S^2. Indeed, we will see shortly that they are both connected to the sample variance.

The sample variance S^2 is an estimator for σ^2. To compute

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1},$$

we need to first compute the sample mean $\overline{X}$. Once that is done, the data $X_1, X_2, \ldots, X_n$ contains only $n-1$ independent pieces of information.

For example, if your grade in a class is computed from 3 equally weighted exams and I told you that your average was 87, you could make up your own scores for the exams, but you can only make up two of them. The third score would be "locked in" by the fact that the average needs to be 87. You are "free to vary" two out of three of the scores.

Going back to the sample variance calculation, when $\overline{X}$ is known, there are only $n-1$ independent pieces of information that are "free to vary". The "degrees of freedom" associated with the calculation of S^2 is $n-1$.

3.3.3 Return to the Confidence Interval

Let $X_1, X_2, \ldots, X_n$ be a random sample from the $N(\mu, \sigma^2)$ distribution with n "small". Suppose that we do not know σ^2. Because $X_1, X_2, \ldots, X_n$ are normal, we know that $\overline{X}$ is normal, and, in particular,

$$\overline{X} \sim N(\mu, \sigma^2).$$

So,

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Thus,

$$\begin{aligned}
\frac{\overline{X}-\mu}{S/\sqrt{n}} &= \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \cdot \frac{\sigma}{S} \\
&= \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \cdot \sqrt{\frac{\sigma^2}{S^2}} \\
&= \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \Big/ \sqrt{\frac{S^2}{\sigma^2}} \\
&= \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \Big/ \sqrt{\frac{(n-1)S^2}{\sigma^2} \Big/ (n-1)}
\end{aligned}$$

So, we have a $N(0,1)$ random variable divided by the square root of an independent (See Exercise 4 of this Chapter.) $\chi^2(n-1)$ random variable, divided by it's degrees of freedom. This is how we defined a t-distribution with $n-1$ degrees of freedom. Therefore

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2) \qquad \Rightarrow \qquad \frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1).$$

Amazing how all of this stuff is coming together isn't it?

The two numbers that can capture area $1-\alpha$ in the center of a $t(n-1)$ curve are $t_{1-\alpha/2,n-1} = -t_{\alpha/2,n-1}$ and $t_{\alpha/2,n-1}$. So,

$$1-\alpha = P\left(-t_{\alpha/2,n-1} < \frac{\overline{X}-\mu}{S/\sqrt{n}} < t_{\alpha/2,n-1}\right)$$

Solving for μ "in the middle", we get that a $100(1-\alpha)\%$ confidence interval for μ is

$$\left(\overline{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}},\ \overline{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right),$$

or simply

$$\overline{X} \pm t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}.$$

> If $X_1, X_2, \ldots, X_n$ is a random sample from the normal distribution with mean μ and unknown variance σ^2, a $100(1-\alpha)\%$ confidence interval for μ, when n is small, is given by
>
> $$\overline{X} \pm t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}.$$

We actually don't need a small sample to use this result– we certainly didn't use it in the derivation above. You should technically use a t-distribution when giving a confidence interval for μ any time you have normality and σ^2 is unknown. However, the critical values for the t and standard normal distributions become virtually

indistinguishable for larger n at the level of 3 to 4 decimal places given in the tables. You can get critical values for these distributions from Tables C.1 and C.2 in Appendix C. Note that they are structured quite differently. For the z-table, critical values are in the row and column headers and probabilities are in the body of the table. For the t-table it is reversed.

Example 3.3.1

A random sample of size 10 from the normal distribution with mean μ and variance σ^2 results in a sample mean of $\overline{x} = 4.23$ and a sample variance of $s^2 = 1.76$. Give an 80% confidence interval for the mean μ.

We are in the small sample case for the mean of a normal distribution when σ^2 is unknown. This means that we need to use t-critical values. In particular, if we want to capture area 0.80 between two numbers, we want 0.10 below the lower number and 0.10 above the upper number for the $t(9)$ distribution.

The upper number is $t_{0.10,9}$. This is our notation for a number that cuts off area 0.10 above for the $t(9)$ curve. Table C.2 from Appendix C has the reverse notation for critical values in the top row. We want to go to the column associated with $t_{0.10}$ and down to the row labeled with a 9. The critical value is

$$t_{0.10,9} = 1.383.$$

(Make sure you can find this in the Table!)

The 80% confidence interval for μ is given by

$$\overline{X} \pm t_{0.10,9}\frac{S}{\sqrt{n}},$$

which is

$$4.23 \pm 1.383\frac{\sqrt{1.76}}{\sqrt{10}},$$

which gives the interval as

$$(3.65, 4.81).$$

3.3.4 Summary of Confidence Intervals for the Mean Involving Normal Distributions

Given a random sample of size n from a distribution with mean μ and variance σ^2, $100(1-\alpha)\%$ confidence interval for μ in various cases is given below.

			Distribution		
			Normal	Not Normal But Known	Unknown
Sample Size	$n \leq 30$	σ^2	$\overline{X} \pm z_{\alpha/2}\,\sigma/\sqrt{n}$	See Section 3.5	Can't do this!
		S^2	$\overline{X} \pm t_{\alpha/2,n-1}\,\frac{S}{\sqrt{n}}$		
	$n > 30$	σ^2	$\overline{X} \pm z_{\alpha/2}\,\sigma/\sqrt{n}$	$\overline{X} \pm z_{\alpha/2}\,\sigma/\sqrt{n}$	$\overline{X} \pm z_{\alpha/2}\,\sigma/\sqrt{n}$
		S^2	$\overline{X} \pm z_{\alpha/2}\,S/\sqrt{n}$	$\overline{X} \pm z_{\alpha/2}\,S/\sqrt{n}$	$\overline{X} \pm z_{\alpha/2}\,S/\sqrt{n}$

The decisions given in this Table about which confidence interval to use in a given situation reflect what people usually do in practice. Again, the critical value in the red cell in the Table should technically be $t_{\alpha/2,n-1}$ and the confidence intervals given in the green cells should technically be done using the techniques of Section 3.5.

3.4 A Difference of Means

In this Section, we will consider two samples from two different "populations" and develop a confidence interval that will be suitable for comparing the means for those populations.

3.4.1 Normal Samples, Variances Known

Suppose that we have two independent random samples: one of size n_1 from the $N(\mu_1, \sigma_1^2)$ distribution and one of size n_2 from the $N(\mu_2, \sigma_2^2)$ distribution.

We wish to derive a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$, the difference in means.

Any confidence interval starts with an estimator. Let's estimate $\mu_1 - \mu_2$ with the difference in sample means

$$\overline{X}_1 - \overline{X}_2.$$

What is the distribution of $\overline{X}_1 - \overline{X}_2$?

Aside:

While we don't think it is necessary here, we can denote the two samples with double subscripts:

$$X_{11}, X_{12}, \ldots, X_{1\,n_1}$$

and

$$X_{21}, X_{22}, \ldots, X_{2\,n_2},$$

and then we can more explicitly write out the sample means as

$$\overline{X}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1\,i} \quad \text{and} \quad \overline{X}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} X_{2\,i}$$

We know that

$$\overline{X}_1 \sim N(\mu_1, \sigma_1^2/n_1) \quad \text{and} \quad \overline{X}_2 \sim N(\mu_2, \sigma_2^2/n_2).$$

Since a linear combination of normal random variables is again normal, we know that $\overline{X}_1 - \overline{X}_2$ has a normal distribution.

The mean is

$$\mathsf{E}[\overline{X}_1 - \overline{X}_2] = \mathsf{E}[\overline{X}_1] - \mathsf{E}[\overline{X}_2] = \mu_1 - \mu_2.$$

The variance is

$$Var[\overline{X}_1 - \overline{X}_2] \overset{indep}{=} Var[\overline{X}_1] + Var[\overline{X}_2] = \sigma_1^2/n_1 + \sigma_2^2/n_2.$$

So, we have that

$$\overline{X}_1 - \overline{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

We know that any normal random variable, with its mean subtracted, and divided by its standard deviation, is a standard normal ($N(0,1)$) random variable. So, in this case, we know that

$$\frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1).$$

Let $Z \sim N(0,1)$. Suppose we want to find two numbers that capture area 0.80 in the center of the $N(0,1)$ pdf. Using our z-table, we see that these numbers are approximately ± 1.28. Thus,

$$\begin{aligned} 0.80 &= P(-1.28 < Z < 1.28) \\ &= P\left(-1.28 < \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < 1.28\right) \end{aligned}$$

Solving for $\mu_1 - \mu_2$ in the middle gives

$$0.80 = P\left(\overline{X}_1 - \overline{X}_2 - 1.28\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < 1.28\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right),$$

making an 80% confidence interval

$$\left(\overline{X}_1 - \overline{X}_2 - 1.28\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},\ \overline{X}_1 - \overline{X}_2 + 1.28\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right),$$

or more simply

$$\overline{X}_1 - \overline{X}_2 \pm 1.28\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

For the more general percentage of $100(1-\alpha)\%$, instead of 80%, we need to replace ± 1.28 with $\pm z_{\alpha/2}$.

In general, if we have a random sample of size n_1 from the $N(\mu_1, \sigma_1^2)$ distribution and an independent random sample of size n_2 from the $N(\mu_2, \sigma_2^2)$ distribution, a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is given by

$$\overline{X}_1 - \overline{X}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},$$

provided σ_1^2 and σ_2^2 are known.

3.4.2 Large Samples, Variances Unknown

For large n_1 and n_2, the two population sample variances S_1^2 and S_2^2 are good approximations of σ_1^2 and σ_2^2, respectively.

So, we have that

$$\frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \approx \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

and therefore that an approximate $100(1-\alpha)\%$ confidence interval for the difference $\mu_1 - \mu_2$ is given by

$$\overline{X}_1 - \overline{X}_2 \pm z_{\alpha/2} \sqrt{\frac{S^2}{n_1} + \frac{S_2^2}{n_2}}.$$

More formally, we are using the fact that

$$\frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \xrightarrow{d} N(0,1).$$

This can be shown by writing

$$\frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \Bigg/ \sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right) \Bigg/ \left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)} \tag{3.4.4}$$

It can be shown (See Exercise 14 of this Chapter.) that

$$\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right) \Bigg/ \left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)} \xrightarrow{P} 1.$$

For normal populations, we have that

$$\frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1).$$

If we don't have normal populations, we still have

$$\frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \xrightarrow{d} N(0,1)$$

from an application of the CLT after some manipulation. This limit is not entirely obvious– one must be careful

when dealing with double limits!

In general, if we have a random sample of size n_1 from the $N(\mu_1, \sigma_1^2)$ distribution and an independent random sample of size n_2 from the $N(\mu_2, \sigma_2^2)$ distribution, with σ_1^2 and σ_2^2 unknown, an approximate $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is given by

$$\overline{X}_1 - \overline{X}_2 \pm z_{\alpha/2}\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}v\right)}$$

provided that both n_1 and n_2 are "large".

3.4.3 At Least One Small Sample, Normality, Variances Unknown

In the case of a confidence interval for a single mean, we saw that we had to move to a t-distribution for critical values when using S^2 in place of σ^2, and that it was of particular importance to do so when working with a small sample. The same will be true when developing a confidence interval for the difference between two means.

We have two cases to consider for a confidence interval for a difference of means. The first is when $\sigma_1^2 = \sigma_2^2$ and the second is when $\sigma_1^2 \neq \sigma_2^2$. Even though we do not know the values for σ_1^2 and σ_2^2, there are situations where we might believe them to be equal. For example, suppose that we want to find a confidence interval for $\mu_1 - \mu_2$ where μ_1 is the true mean sales volume of some breakfast cereal across all grocery stores and μ_2 is the true mean sales volume of the same cereal across all grocery stores but in an exciting new package. You might expect a new flashy package design to increase sales but maybe not necessarily to change the variance of sale prices. Then again, maybe not. You probably shouldn't bring your "expectations" into this. We will eventually learn about hypothesis testing and we will develop a statistical test to determine whether or not it is reasonable to assume that $\sigma_1^2 = \sigma_2^2$.

Case: $\sigma_1^2 = \sigma_2^2$

We begin by giving the common value of σ_1^2 and σ_2^2 a more "sample neutral name" like σ^2. That is $\sigma^2 = \sigma_1^2 = \sigma_2^2$ and we will stop referring to σ_1^2 and σ_2^2 altogether.

We compute sample variances, S_1^2 and S_2^2 for each of the samples. While we are trying to estimate a single variance σ^2, we do not just combine the samples and compute a single sample variance because variance is the expected squared deviation from the mean and the samples are not assumed to have the same means.

Since S_1^2 and S_2^2 are both estimators for the common variance σ^2, we should use information from both to come up with a single estimate for the unknown σ^2. One possibility is the average value

$$\frac{S_1^2 + S_2^2}{2}.$$

This is maybe not the best estimate. It would be better to use some sort of weighted average to take into account the different sample sizes n_1 and n_2. After all, if n_1 is much larger than n_2, the sample variance from that first population (S_1^2) is probably much more accurate than the sample variance from the second population (S_2^2) and thus it should carry more weight.

Here is a weighted average:

$$\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2}.$$

This solves the issue of giving more weight to the sample variance from the larger sample but it is not mathematically convenient. Recall that, in order to develop a confidence interval, we must be able to find the distributions of certain quantities.

To this end consider a different weighted average, denoted by S_p^2 and known as the **pooled variance**.

$$S_p^2 := \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Note that

$$W := \frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2} = \frac{(n_1 - 1)S_1^2}{\sigma^2} + \frac{(n_2 - 1)S_2^2}{\sigma^2}$$

The two terms on the right-hand side are independent since the sample variances come from samples that were assumed to be independent. From Section 3.3.2, we know that the distributions of these terms are both χ^2 with $n_1 - 1$ and $n_2 - 1$ respective degrees of freedom. By Exercise 7 from Chapter 1, we then know that their sum is another χ^2 random variable with $(n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$ degrees of freedom.

We are going to base our confidence interval off of the quantity

$$\frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}. \tag{3.4.5}$$

Note that

$$\frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \Bigg/ \sqrt{\frac{S_p^2}{\sigma^2}}$$

$$= \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \Bigg/ \sqrt{\left(\frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2}\right) \Bigg/ (n_1 + n_2 - 2)} \ .$$

In conclusion, (3.4.5) can be rewritten as a standard normal random variable, divided by the square root of a χ^2 random variable divided by its degrees of freedom. This gives us that

$$\frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t(n_1 + n_2 - 2).$$

For a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$, we can place this quantity in between the two t-critical values $t_{\alpha/2,n_1+n_2-2}$ and $-t_{\alpha/2,n_1+n_2-2}$. And solve the inequalities for $\mu_1 - \mu_2$ "in the middle".

In general, if we have a random sample of size n_1 from the $N(\mu_1, \sigma_1^2)$ distribution and an independent random sample of size n_2 from the $N(\mu_2, \sigma_2^2)$ distribution, with σ_1^2 and σ_2^2 unknown but assumed to be equal, a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is given by

$$\overline{X}_1 - \overline{X}_2 \pm z_{\alpha/2}\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}v\right)}$$

Case: $\sigma_1^2 \neq \sigma_2^2$ (The Behrens-Fisher Problem)

So far, we know how to create exact or approximate confidence intervals for a difference between two means in the cases where the underlying population distributions are

- normal and σ_1^2 and σ_2^2 are known,
- normal or unknown where σ_1^2 and σ_2^2 are unknown but both samples are large, and
- normal with small samples where σ_1^2 and σ_2^2 are unknown but can be assumed to be equal.

For small samples from normal populations, the quantity

$$\frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

does not have an exact known distribution. Since we are assuming the samples are small, it also doesn't make sense to use a convergence result for a "large sample approximation". Indeed, the best way to proceed here is not clear at all. This problem is known as the **Behrens-Fisher problem**. The most common approach people take is to use something known as **Welch's Approximation**[5].

Welch's Approximation

Let S_1^2 be the sample variance for a sample of size n_1 from the $N(\mu_1, \sigma_1^2)$ distribution. Let S_2^2 be the sample variance for an independent sample of size n_1 from the $N(\mu_1, \sigma_1^2)$ distribution. Then

$$\frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \overset{approx}{\sim} t(r)$$

where r is

$$\frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}}$$

rounded down to an integer.

Recall that the t-distribution is "flatter" than the standard normal distribution. The pdfs only get close for large n. The reason the degrees of freedom approximation is rounded down here is because it will result in flatter curve for which one has to go out further in the tails of the pdf to find critical values. This results in a more conservative (longer) confidence interval in the face of the extra uncertainty in the Behrens-Fisher problem.

Example 3.4.1

A farmer has two 10 acre fields, a "northern field" and a "southern field", both on which they have planted a single variety of corn. The southern field is known to have soil that is more alkaline than that from the northern field. Just before harvesting, a random sample of 15 plants were selected from each field and their heights were measured. The plants from the northern field had an average height of 8.2 feet and the standard deviation of the heights in this group was $s_1 = 1.6$. The plants from the southern field had an

average height of 7.1 feet and the standard deviation of the heights in this group was $s_2 = 2.2$. (Note the use of lowercase letters because we have observed numbers here and are no longer working with random variables.)

Assuming that the heights of corn stalks are normally distributed, give a 90% confidence interval for $\mu_2 - \mu_1$. Here, μ_1 is the true average height for the northern field and μ_2 is the true average height of the southern field.

Note that standard deviations are given as opposed to variances and that we have changed the direction of the difference. Since we have small samples, sample standard deviations as opposed to true standard deviations, and were not given any information about the true variances possibly being equal, we will use Welch's approximation to give an approximate 90% confidence interval for $\mu_2 - \mu_1$.

The degrees of freedom calculation is symmetric in S_1^2 and S_2^2, so we do not need to adjust it for the fact that we are looking at $\mu_2 - \mu_1$ instead of $\mu_1 - \mu_2$.

$$\frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1-1} + \frac{(S_2^2/n_2)^2}{n_2-1}} = \frac{\left(\frac{1.6^2}{15} + \frac{2.2^2}{15}\right)^2}{\frac{(1.6^2/15)^2}{14} + \frac{(2.2^2/15)^2}{14}} \approx 25.5724$$

so we will use $r = 25$ degrees of freedom.

We have that

$$\frac{\overline{X}_2 - \overline{X}_1 - (\mu_2 - \mu_1)}{\sqrt{\frac{S_2^2}{n_1} + \frac{S_1^2}{n_2}}} \overset{approx}{\sim} t(25).$$

We want to capture area 0.90 between two critical values for the $t(25)$ distribution. This means that we want to split the area 0.10 into the two tails of the pdf. From Table C.2 of Appendix C, we get the upper critical value

$$t_{0.05,25} = 1.708.$$

By symmetry of the t-distribution about zero, the lower critical value is -1.708.

We have

$$0.90 \approx P\left(-t_{0.05,25} < \frac{\overline{X}_2 - \overline{X}_1 - (\mu_2 - \mu_1)}{\sqrt{\frac{S_2^2}{n_1} + \frac{S_1^2}{n_2}}} < t_{0.05,25}\right)$$

which gives the confidence interval

$$\overline{X}_2 - \overline{X}_1 \pm t_{0.05,25}\sqrt{\frac{S_2^2}{n_1} + \frac{S_1^2}{n_1}}$$

Plugging in all numerical values gives us

$$(-2.30, 0.10).$$

While the interval contains mostly negative numbers, it does contain the value 0 which tells us that it is "plausible" that the true average height for all corn stalks is the same for both fields!

The next Section is, arguably, the single most important section of this Chapter. A basic, data-oriented, "STAT 101" course will give all of the confidence interval formulas that we have come up with thus far. Students there will "plug and chug" numbers but will really be in trouble if assumptions like normality are not met. It is up to us in mathematical statistics to be able to derive confidence interval formulas in many different settings and for many different types of parameters!

3.5 General Confidence Intervals

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from some distribution that depends on some one-dimensional parameter θ. There are three steps to constructing a confidence interval for θ or a one-dimensional $\tau(\theta)$.

1. Choose a statistic $T = t(\vec{X})$ on which to base the confidence interval.
   Example: $\theta = \mu$, choose $\overline{X}$
2. Find a function of your statistic and the parameter θ whose distribution is known and is "parameter free" or at least "unknown parameter free". This will be known as a **pivotal quantity**.
   Example: $(\overline{X} - \mu)/(\sigma/\sqrt{n}) \sim N(0, 1)$
3. Put the pivotal quantity between two appropriate critical values and solve for the unknown parameter "in the middle".
   Example: $-z_{\alpha/2} < (\overline{X} - \mu)/(\sigma/\sqrt{n}) < z_{\alpha/2}$

Note that, for Step 1, we said "a statistic" and not necessarily something that makes sense as an estimator for θ, as we will see in the following example.

Example 3.5.1

Let $X_1, X_2, \ldots, X_n$ be a random sample from the exponential distribution with rate λ. Derive a $100(1-\alpha)\%$ confidence interval for λ based on the sample minimum $X_{(1)}$.

In this example, the statistic has been chosen for us. We are going to base our confidence interval on the sample minimum. In general, choosing a statistic can be difficult. You should start by thinking about estimators for the unknown parameter, but what is an estimator? There is no real definition or guidance unless you want to require certain properties like, for example, unbiasedness. The sample minimum is maybe not something that naturally comes to mind when we think about estimating the rate parameter for an exponential distribution, but we need to choose something "for which we can do Step 2". This comes with experience.

In Section 1.4.2, we saw that, if $X_{(1)}$ is the minimum of a sample of size n for the exponential distribution with rate λ, then

$$X_{(1)} \sim exp(rate = n\lambda).$$

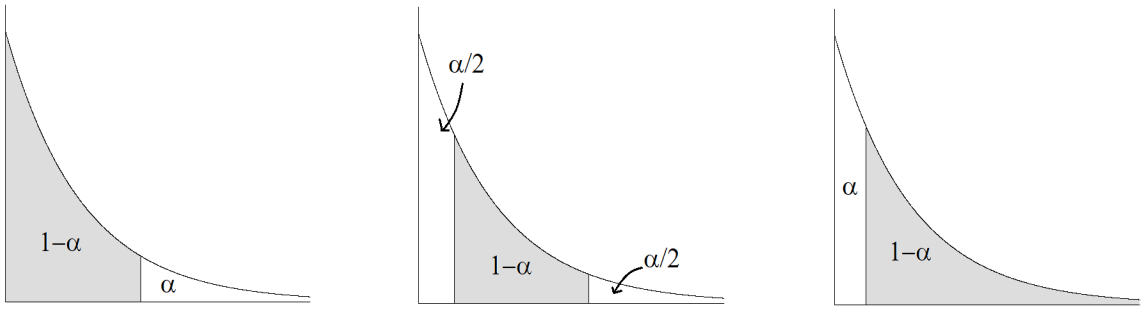
Can we come up with a function of $X_{(1)}$ and λ whose distribution is parameter free? In Exercise 5 from Chapter 1, we saw that, if $X \sim \Gamma(\alpha, \beta)$ and c is a positive constant, then $cX \sim \Gamma(\alpha, \beta/c)$. The exponential distribution with rate $n\lambda$ is the same thing as the $\Gamma(1, n\lambda)$ distribution. So, for this example, we know that

$$\lambda X_{(1)} \sim \Gamma(1, n) = exp(rate = n).$$

This distribution does not depend on the parameter λ, so $\lambda X_{(1)}$ could be the pivotal quantity we are looking for! However, to make the distribution even simpler, let's consider the quantity

$$n\lambda X_{(1)} \sim exp(rate = 1).$$

Let $Y_n := n\lambda X_{(1)}$. Can we find two numbers that capture Y_n in between with probability $1 - \alpha$? There are many ways to do this but here are three possibilities.



There is no real reason for us to put the desired area "in the middle" like we did for the normal and t distributions. For those distributions, since the bulk of the area under the pdfs happens to be in the middle, choosing critical values in this way will result in the shortest confidence interval that can be made using the given statistic. Using the shortest interval is preferable since we are giving an interval estimate of plausible values for the true parameter. We are saying that we believe that it is between two particular numbers. A shorter interval represents a more precise conclusion.

Although the shaded areas shown above are not displayed with proper scale (even relative to each other!), it makes sense that the first of the three ways to capture area $1-\alpha$ will give the smallest interval. The great height of the pdf down by zero on the x-axis implies that if we moved the left endpoint of the shaded region up a bit, the right endpoint would have to move up further in order to compensate for all the area lost on the left-hand side. It is left as an exercise (Exercise 9 in this Section) to show, using Calculus, that putting all $1-\alpha$ area to the left will indeed give the shortest interval.

Going back to the specific problem, we would like to solve

$$P(0 < Y_n < b) = 1 - \alpha$$

for b. Since Y_n is exponentially distributed with rate 1, we know that

$$P(0 < Y_n < b) = \int_0^b e^{-y}\, dy = 1 - e^{-b}.$$

Setting this equal to $1-\alpha$ and solving for b gives us that

$$b = -\ln \alpha.$$

So, we have that

$$\begin{aligned} 1-\alpha &= P(0 < Y_n < -\ln \alpha) \\ &= P(0 < n\lambda X_{(1)} < -\ln \alpha). \end{aligned}$$

Solving for λ "in the middle", we get

$$1-\alpha = P\left(0 < \lambda < -\frac{\ln \alpha}{nX_{(1)}}\right).$$

So, a $100(1-\alpha)\%$ confidence interval for the rate parameter of the exponential distribution is given by

$$\left(0, -\frac{\ln \alpha}{nX_{(1)}}\right).$$

In the next example we will derive a confidence interval for the variance of the normal distribution.

Example 3.5.2

Let $X_1, X_2, \ldots, X_n$ be a random sample from the $N(\mu, \sigma^2)$ distribution with both μ and σ^2 are unknown. Derive a $100(1-\alpha)\%$ for the variance σ^2.

For this confidence interval, let's try the "common sense estimator" and use the sample variance S^2 to make a confidence interval for σ^2.

We need to form a pivotal quantity using S^2 and σ^2. Remember that the distribution of this quantity can not depend on any unknown parameters. In this example, the unknown parameters are μ and σ^2. (If μ was assumed to be known for this problem, the distribution of the pivotal quantity could depend on μ.)

In Section 3.3.2 we learned that, for a sample from a normal distribution, we have

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

Thus, we have a function of S^2 and σ^2 whose distribution is known. Because it has a χ^2-distribution, we will put it between two χ^2-critical values. Following our notation from Section 3.3.1, we will use $\chi^2_{\alpha,n}$ to denote the value on the x-axis to the right of which we capture area α under the $\chi^2(n)$ pdf.[6]

Let $W = (n-1)S^2/\sigma^2$. Per our discussion from the previous example, we have

$$P(0 < W < \chi^2_{\alpha,n-1}) = 1-\alpha,$$

as well as

$$P(\chi^2_{1-\alpha/2,n-1} < W < \chi^2_{\alpha/2,n-1}) = 1-\alpha,$$

as well as

$$P(\chi^2_{1-\alpha,n-1} < W < \infty) = 1-\alpha.$$

(Note that the χ^2-distribution is not symmetric about zero so it doesn't make sense to use $\chi^2_{\alpha/2,n-1}$ and $-\chi^2_{\alpha/2,n-1}$.)

We can also capture area $1-\alpha$ in the middle using, for example, a lower cutoff of $\chi^2_{1-2\alpha/3,n-1}$ and an upper cutoff of $\chi^2_{\alpha/3,n-1}$. The $\chi^2(n)$ pdf can take on many different shapes depending on the value of n. (After all, it is a gamma pdf with shape parameter $n/2$.) For a fixed n, one could use Calculus to find critical values that will give the shortest confidence interval for σ^2. Here, we will just report three intervals corresponding to the three probability statements above. Plugging in $(n-1)S^2/\sigma^2$ for W and solving for σ^2 "in the middle" gives us the intervals

$$\left(\frac{(n-1)S^2}{\chi^2_{\alpha,n-1}},\infty\right), \qquad \left(\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}},\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right), \qquad \text{and} \quad \left(0,\frac{(n-1)S^2}{\chi^2_{1-\alpha,n-1}}\right).$$

In the previous example, we were able to find actual closed form expressions for critical values as functions of α and n. Here, we would have to numerically integrate regions under the $\chi^2(n-1)$ pdf or appeal to a table such as Table C.3 in Appendix C once we are given numerical values for α and n.

[6] We could also use the notation $\chi^2_\alpha(n)$. Be aware that other authors may use either notation as a value that captures area α to the left.

Super Important Note

The χ^2-distribution is central to a large number of results in Statistics. It will be important for us to be able to transform certain statistics into chi-squared random variables so that we can take advantage of these results. Of particular interest is the following transformation of the sample mean for an exponential distribution.

Suppose that

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} exp(rate = \lambda).$$

By Example 1.5.7 of Chapter 1, we know that

$$\sum_{i=1}^{n} X_i \sim \Gamma(n, \lambda).$$

By Example 1.2.2 or Exercise 5 of Chapter 1, we know that multiplying a gamma random variable by a positive constant will give us another gamma random variable with the relationship

$$X \sim \Gamma(\alpha, \beta) \qquad \Rightarrow \qquad cX \sim \Gamma(\alpha, \beta/c).$$

Thus, we know that, starting with the exponential rate λ distribution,

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \Gamma(n, n\lambda).$$

Recall that the $\chi^2(n)$ distribution is defined as the $\Gamma(n/2, 1/2)$ distribution. We could get our sample mean for the exponential distribution looking more like a chi-squared random variable my multiplying by $2n\lambda$.

$$2n\lambda\overline{X} \sim \Gamma\left(n, \frac{1}{2}\right).$$

Indeed, we already have a χ^2-distribution since

$$2n\lambda\overline{X} \sim \Gamma\left(n, \frac{1}{2}\right) = \Gamma\left(\frac{2n}{2}, \frac{1}{2}\right) = \chi^2(2n).$$

Amazing! Keep this in mind when doing Exercise 11 of this Chapter.

Chapter 3 Exercises

1. Let $X_1, X_2, \ldots, X_n$ be a random sample from the $N(\mu, \sigma^2)$ distribution. Find an unbiased estimator for σ.
2. Suppose that X_1, X_2 is a random sample of size 2 from the $N(0, \sigma^2)$ distribution. Find the distribution of
$$Y = \frac{X_1}{|X_2|}.$$
(Hint: Note that $|x| = \sqrt{x^2}$. Don't make this too hard! You **don't** have to do a "Jacobian transformation"!)
3. Let $X_1, X_2, \ldots, X_n$ be a random sample from the $N(\mu, \sigma^2)$ distribution.
   (a). Find the variance of the sample variance S^2.
   (b). Show that S^2 is a consistent estimator of σ^2.
4. Suppose that $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$.
   (a). Show that
   $$\frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma^2} + \frac{n(\overline{X} - \mu)^2}{\sigma^2}.$$
   (This has nothing to do with the particular distribution here.)
   (b). Write down the joint pdf for $X_1, X_2, \ldots, X_n$ and use the above to rewrite the "e-exponent part".
   (c). Consider the joint transformation $Y_1 = \overline{X}$, $Y_i = X_i - \overline{X}$ for $i = 2, 3, \ldots, n$.
   Use the Jacobian method to find the joint pdf for $Y_1, Y_2, \ldots, Y_n$. Show that Y_1 is independent of $Y_2, \ldots, Y_n$.
   (d). Show that $X_1 - \overline{X} = -\sum_{i=2}^n (X_i - \overline{X})$.
   Conclude that $X_1 - \overline{X}$ is independent of $\overline{X}$.
   (e). Conclude that $\overline{X}$ is independent of the sample variance S^2!
5. Let W_1 and W_2 be independent random variables with $W_1 \sim \chi^2(n_1)$ and $W_2 \sim \chi^2(n_2)$. Suppose that $n_2 > n_1$. Does $W_2 - W_1$ have a χ^2 distribution? If so, prove it and give the associated degrees of freedom. If not, explain why it is not true.
6. Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from any distribution with a finite fourth moment. In Section 2.6, we saw a not so nice expression for the variance of the sample variance S^2.
   Now suppose that the random sample is from the $N(\mu, \sigma^2)$ distribution. Find the variance of S^2 using the fact that $(n-1)S^2/\sigma^2$ and without using the forth moment formula.
7. (a). Suppose that a random sample of size 15, taken from the $N(\mu, 9)$ distribution, results in a sample mean of 5.2. Give an 85% confidence interval for the true mean μ.
   (b). Suppose that a random sample of size 150, taken from the $N(\mu, \sigma^2)$ distribution, results in a sample mean of 5.2 and a sample variance of 8.3. Give a 90% confidence interval for the true mean μ.

(c). Suppose that a random sample of size 15, taken from the $N(\mu, \sigma^2)$ distribution, results in a sample mean of 5.2 and a sample variance of 8.3. Give a 90% confidence interval for the true mean μ.

8. Let $X_1, X_2, \ldots, X_n$ be a random sample from the continuous uniform distribution on the interval from 0 to θ. Construct a 95% confidence interval for θ.

9. For Example 3.5.1, prove that using 0 as the left critical value will give the shortest possible confidence interval when using the minimum to estimate λ. (*Hint: Make the left critical value some point a. Find the corresponding right critical value needed to capture area $1-\alpha$. The length of this interval will be a function of a that can be minimized with respect to a.*)

10. Consider a random sample of size $n_1 = 9$ from the $N(\mu_1, \sigma_1^2)$ distribution and an independent random sample of size $n_2 = 12$ from the $N(\mu_2, \sigma_2^2)$ distribution. Suppose that the variances are unknown but, for some crazy reason you do know that $\sigma_1^2 = 3\sigma_2^2$. Define a random variable that has a t-distribution that can be used to find a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$.

11. Let $X_1, X_2, \ldots, X_n$ be a random sample from the exponential distribution with rate λ. In this Chapter, we constructed a $100(1-\alpha)\%$ confidence interval for λ "based on the minimum $X_{(1)}$". Now, construct $100(1-\alpha)\%$ confidence interval for λ based on the sample mean $\overline{X}$. Give your answer in terms of χ^2-critical values.

12. Let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ denote the order statistics of a random sample of size n from a distribution that has pdf

$$f(x) = \frac{3}{\theta^3} x^2 \, I_{(0,\theta)}(x).$$

(a). Show that

$$P(c < X_{(n)}/\theta < 1) = 1 - c^{3n}$$

where $0 < c < 1$.

(b). If $n = 4$ and the observed value of $X_{(4)}$ is 2.3, find a 95% confidence interval for θ based on $X_{(4)}$.

13. Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from a continuous distribution with pdf f and cdf F. You may assume that F is invertible.

(a). Find the distribution of $Y_i := F(X_i)$. Name it!

(b). Find the distribution of

$$-2\sum_{i=1}^{n} \ln F(X_i).$$

Name it!

(c). Find the distribution of

$$-2\sum_{i=1}^{n} \ln[1 - F((X_i))].$$

Name it! (*Hint: You can do this with very little extra work after part (a)!*)

(d). Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} Pareto(\gamma)$. Use part (c) to construct a $100(1-\alpha)\%$ confidence interval for γ. Leave your answer in terms of χ^2-critical value.

14. Consider a random sample of size n_1 from the $N(\mu_1, \sigma_1^2)$ distribution and an independent random sample of size n_2 from the $N(\mu_2, \sigma_2^2)$ distribution. Let S_1^2 and S_2^2 be the sample variances. Prove that

$$\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right) \Big/ \left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)} \xrightarrow{P} 1.$$

Chapter 4 A "Mostly Normal" Introduction to Hypothesis Testing

If you wish to start studying formal methods of estimation at this time,[1] you can skip to Chapter 5 and come back to this Chapter before going on to Chapter 6. There is a reason we put this "mostly normal introduction to hypothesis testing" here though. Over the years, we have found that people tend to have an easier time with the more abstract hypothesis testing presented in Chapter 6 after learning some of the basics here and then letting it all sort of "soak in" for a bit.

Imagine a population of people with some true average height μ. We'd like to know what μ is, but the population is quite large so we restrict ourselves to looking only at a relatively small sample of members of the population, measuring them, and computing $\overline{X}$ which is the average height in the sample. We might even make a confidence interval. A point or interval estimator is great but, in **hypothesis testing**, we go further and use our estimator to make decisions.

As with most concepts in this text, it is important to pay attention to the method of deriving a hypothesis test. In this "mostly normal introduction" to hypothesis testing we will spend a fair amount of time deriving tests for the mean of a normal distribution. We can put the steps we come up with into a nicely framed box for any "STAT 101 kid" to follow. Ultimately, we do not care about the particular steps! Instead, we want to understand the process for deriving them and to be able to come up with them as needed. In MathStat, it is our goal to be able to seamlessly move from deriving a test for the mean μ of a normal distribution to, for example, deriving a test for the parameter γ for a Pareto distribution without any additional instruction because really, it's all the same thing.

4.1 Getting Started, Some Intuition

Suppose that we have a random sample of size 10 from a normal distribution with mean μ and variance σ^2, and suppose that we observe a sample mean of $\overline{x} = 3.7$. Further, suppose that we wish to test the idea that $\mu \leq 3$ versus $\mu > 3$. We will do this by assuming that one of these statements is true and then looking for evidence to the contrary, given by our sample, to see if it will change our minds.

[1] We don't blame you!

We will set up two *hypotheses*, denoted by H_0 and H_1, that are, for this example, written like this.

$$H_0 : \mu \leq 3 \qquad \text{versus} \qquad H_1 : \mu > 3 \tag{4.1.1}$$

Definition 4.1.1

H_0 is called a **null hypothesis** and is, initially, assumed to be true.

H_1 is called an **alternative hypothesis** and is what we will conclude to be true if we are swayed by strong evidence from the sample.

Just because the sample mean is greater (or less) than 3 does not mean that we can conclude that μ, the overall mean for the entire population from which the sample was taken, is greater (or less) than 3. As a linear combination of normals, $\overline{X}$ has a normal distribution that can take on any value from $-\infty$ to ∞, even if the null hypothesis is true and that distribution is truly centered at a number that is less than or equal to 3. To test the hypotheses given in (4.1.1), we will assume that H_0 is true, but we will reject H_0 in favor of H_1 if $\overline{X}$ appears to be "significantly" larger than 3. To proceed, we need to make sense out of the phrase "significantly larger than 3".

4.2 Hypotheses: Simple or Composite?

A **simple hypothesis** is one that "completely specifies the distribution". For example, suppose that we have a random sample from the $N(\mu, 1)$ distribution and that the hull hypothesis is

$$H_0 : \mu = 3.$$

If H_0 is true, then you know the distribution that the random sample came from the $N(3, 1)$ distribution.

A **composite hypothesis** is one that does not completely specify the distribution. For example, suppose that we have a random sample from the $N(\mu, 1)$ distribution and that the hull hypothesis is

$$H_0 : \mu \leq 3.$$

If H_0 is true, then we know the distribution that the random sample came from some normal distribution with mean μ and variance 1 but we do not know which one. Was it the $N(2.912, 1)$ distribution? The $N(0, 1)$ distribution? The $N(-18.237, 1)$ distribution?

It is a common misconception that simple hypotheses have "equals signs" and composite hypotheses have inequalities. For example, suppose that $X_1, X_2, \ldots, X_n$ is a random sample from the $N(\mu, \sigma^2)$ distribution for

unknown μ and σ^2. Then

$$H_0 : \mu = 3$$

is not a simple hypothesis since, if it is true, we still can't say exactly which normal distribution the sample came from. Was it the $N(3, 1.7)$? The $N(3, 8.2)$?

We will begin hypothesis testing in the simplified setting where

$$H_0 : \mu = 3 \qquad \text{and} \qquad H_1 : \mu > 3. \tag{4.2.2}$$

The two hypotheses are supposed to represent all possible values for μ. For the normal distribution, the parameter μ is allowed to take on any value from $-\infty$ to ∞. In (4.2.2), we are missing the possibility that $\mu < 3$ but we start here as a building block towards the hypotheses in (4.1.1). Be aware though, even if we observe a sample mean of negative 3 million, there is still no concluding that μ is less than 3. We assume H_0 is true and reject it in favor of H_1 if $\overline{X}$ is "significantly larger than 3".

4.3 Errors in Hypothesis Testing for a Simple Null Hypothesis

In reality, the null hypothesis H_0 is either true or it is false. Going into a test, we assume that H_0 is true but we will conclude that H_1 is true instead if there is significant evidence in our sample (data) to support it. In the language of hypothesis testing, we either reject H_0 or we fail to reject H_0 but we will never "accept" H_0.

When you perform a hypothesis test, it may be the case that H_0 really is true, but that the data forces you to "reject" it, in favor of the alternative hypothesis H_1, because it was your bad luck to observe an extreme sample that is highly improbable to observe "under H_0". In this case, you have made an error through no fault of your own. Similarly, it is possible to make an error in the other direction.

In summary, if H_0 is true and we fail to reject it, we did good! Likewise, if it is false and we reject it we also did good. In the other cases, we have made one of two types of error as shown in the following table.

		Your Decision	
		Fail to Reject H_0	Reject H_0
Reality	H_0 True	✓	**Type I Error**
	H_0 False	**Type II Error**	✓

Neither type of error is inherently worse. It depends on the context of the problem and the consequences or what is at stake with each type of incorrect decision.

4.3.1 Level of Significance: α

Definition 4.3.1

For a simple null hypothesis, the **level of significance** of the test is the probability of making a Type I error. It is denoted by an α:

$$\alpha = P(\text{ Type I error }) = P(\text{ Reject } H_0 \text{ when it's true }).$$

For reasons which will become apparent later, α is also known as the **size** of the test.

To illustrate a Type I error probability, let us revisit the example from the beginning of this Chapter with a fixed variance of $\sigma^2 = 1$.

Suppose that we have a random sample, $X_1, X_2, \ldots, X_{10}$, of size 10 from the $N(\mu, 1)$ distribution and that we wish to test the hypotheses

$$H_0 : \mu = 3 \qquad \text{versus} \qquad H_1 : \mu > 3.$$

Suppose that we will perform this test by looking at the sample mean $\overline{X}$ and rejecting H_0 in favor of H_1 if $\overline{X} > 4$. (This is just an arbitrary "rejection rule" that we made up. At this time, we will not concern ourselves with where it came from or even whether or not it is a good rule!)

Question:

What is the level of significance (size) of this test?

Answer:

$$\begin{aligned}
\alpha &= P(\text{ Type I Error }) \\
&= P(\text{ Reject } H_0 \text{ when it's true }) \\
&= P(\overline{X} > 4 \text{ when } \mu = 3\,) \\
&= P\left(\tfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} > \tfrac{4-3}{1/\sqrt{10}} \text{ when } \mu = 3\right) && (\Leftarrow \text{standardize } \overline{X}) \\
&= P(Z > 3.16) && \text{where } Z \sim N(0,1) \\
&= 1 - 0.9992 = 0.0008. && (\Leftarrow \text{use } z\text{-table})
\end{aligned}$$

Typically, the level of significance is not determined from a cutoff "rejection rule" (Certainly not an arbitrary one!) but instead is specified at the beginning by the researcher who cares about the particular hypotheses. This person might say, "I am willing to let the Type I Error probability be 0.05, so what should my rejection rule be?"

4.4 Finding a Test: A Simple Null Hypothesis

Suppose that we have a random sample of size 10 from the $N(\mu, 1)$ distribution and that we wish to find a test of size or level of significance $\alpha = 0.05$ that is based on the sample mean $\overline{X}$ for the hypotheses

$$H_0 : \mu = 3 \qquad \text{versus} \qquad H_1 : \mu > 3.$$

We want to reject H_0, in favor of H_1 if the data and, specifically, the statistic $\overline{X}$ is "large" in the sense that it is above some cutoff value. The form of the rejection rule (for the rejection of H_0) is therefore:

"Reject H_0, in favor of H_1, if $\overline{X} > c$."

for some c to be determined.

Note

The form of the rejection rule is determined by the alternative hypothesis, BUT, the "greater than" in the alternative hypothesis does not automatically translate to the "greater than" in the rejection rule!

For example, suppose that you had a random sample $X_1, X_2, \ldots, X_n$ from the exponential distribution with rate λ and that you want to test

$$H_0 : \lambda = 3 \qquad \text{versus} \qquad H_1 : \lambda > 3.$$

based on looking at the sample mean $\overline{X}$.

You would want to reject H_0 in favor of H_1 if a larger value of λ is indicated by the data. Since $\overline{X}$ is an estimator of the mean of this distribution and since the mean is $1/\lambda$, a large value of λ is suggested by the data if $\overline{X}$ is small.

Therefore, the form of the rejection rule for this exponential example is to

"Reject H_0, in favor of H_1, if $\overline{X} < c$."

for some c that is chosen to give a size 0.05 test.

We have given the "form" of the test, but to give the actual test, we still need to find the value of the cutoff c. This is done using the level of significance $\alpha = 0.05$.

Note that

$$\begin{aligned}
0.05 &= P(\text{ Type I Error }) \\
&= P(\text{ Reject } H_0 \text{ when it's true }) \\
&= P(\overline{X} > c \text{ when } \mu = 3\,) \\
&= P\left(\tfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} > \tfrac{c-3}{1/\sqrt{10}} \text{ when } \mu = 3\right) \quad (\Leftarrow \text{ standardize } \overline{X}) \\
&= P\left(Z > \tfrac{c-3}{1/\sqrt{10}}\right) \text{ where } Z \sim N(0,1)
\end{aligned}$$

From the standard normal table, we know that

$$P(Z > 1.645) = 0.05.$$

So, we must have that

$$\frac{c-3}{1/\sqrt{10}} = 1.645,$$

which implies that $c = 3.52$.

In summary, we will reject H_0, in favor of H_1, at the 0.05 level of significance if, $\overline{X} > 3.52$. For the original "data" at the beginning of this Chapter, since we observed $\overline{x} = 3.7$, we would indeed reject H_0 based on this sample.

We now have a "rule" for when to reject H_0. It is known as a **rejection rule**. Since we will reject H_0 when $\overline{X}$ gets into a certain region, we will also call this a **rejection region**.

Note that our cutoff value for the rejection rule came from a transformation of the cutoff value (1.645) on the standard normal curve. The z-critical-value 1.645 cuts off area 0.05 on the right side (upper tail) of the $N(0,1)$ curve. This was translated to the value 3.52 on the $N(3,1)$ curve. If we had chosen a smaller α (say $\alpha = 0.01$), the critical value (which would need to cut off a smaller area to the right) would be higher. It is then possible that we would find that we would no longer reject H_0 with a sample mean of $\overline{x} = 3.7$. In fact, this is the case, as c would be 3.74 when $\alpha = 0.01$. (Check this!)

Note

We have now, more than once, seen a step that has been labeled with the word "standardize". This is not a part of hypothesis testing per se, but rather it comes from the fact that we are trying to compute probabilities involving the normal distribution. Don't make this part of your routine when doing a hypothesis test. It will not always be a step to take when we move on to other distributions!

4.5 Finding a Test: A Composite Null Hypothesis

Continuing with the same example, suppose now that we wish to test

$$H_0 : \mu \leq 3 \qquad \text{versus} \qquad H_1 : \mu > 3$$

at a 0.05 level of significance.

We have

$$\begin{aligned} 0.05 &= P(\text{ Type I Error }) \\ &= P(\text{ Reject } H_0 \text{ when it's true }) \\ &= P(\overline{X} > c \text{ when } \mu \leq 3\,) \\ &= P\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} > \frac{c-?}{1/\sqrt{10}} \text{ when } \mu \leq 3\right) \quad (\Leftarrow \text{standardize } \overline{X}) \end{aligned}$$

We have run into a problem. We can't subtract off μ if we don't know what it is! The Type I error probability is computed under the assumption that H_0 is true. In Section 4.4, this meant that $\mu = 3$, while here H_0 contains many possible values for μ. How can we proceed?

The answer is that our definition of α from Section 4.3.1 is just not adequate. Before we expand the definition, we will write our hypotheses a bit more generally. Recall that, throughout this text, we are using θ to denote a generic parameter. Every distribution with a parameter has a parameter "space". For example, the rate parameter λ for the exponential distribution must be positive. The mean parameter μ for the normal distribution can be anywhere from $-\infty$ to ∞. The vector-valued parameter $\theta = (\mu, \sigma^2)$ for a normal distribution lives on the space $(-\infty, \infty) \times [0, \infty)$.

Notation

For a generic parameter θ, we will denote the **parameter space** by Θ. The parameter needs to be in or "an element of" Θ. In symbols, we say that $\theta \in \Theta$.

Let Θ_0 be some subset of Θ.

We are interested in testing the hypotheses

$$H_0 : \theta \in \Theta_0 \qquad \text{versus} \qquad H_1 : \theta \in \Theta \setminus \Theta_0.$$

The backslash here is the "setminus" notation. $\Theta \setminus \Theta_0$ is everything that is left in Θ if you remove all of the elements of Θ_0. It can be written as

$$\Theta \setminus \Theta_0 = \Theta \cap \Theta_0^C$$

This notation is useful when Θ is thought of as part of a larger space such as the real number line. For example, let $\Theta = (0, \infty)$ and let $\Theta_0 = (1, \infty)$. Then $\Theta_0^C = (-\infty, 1]$ but $\Theta \setminus \Theta_0 = (0, 1]$.

We are now ready to define the size or level of significance of a hypothesis test for a general, possibly composite

null hypotheses.

Definition 4.5.1

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution that depends on a parameter θ that lives in a parameter space Θ.

Let Θ_0 be some subset of Θ.

The **size** or **level of significance** of the test of

$$H_0 : \theta \in \Theta_0 \qquad \text{versus} \qquad H_1 : \theta \in \Theta \setminus \Theta_0$$

is denoted by α and is defined as the maximum probability of making a Type I error:

$$\alpha = \max_{\theta \in \Theta_0} P(\text{ Reject } H_0 \text{ when the parameter is } \theta\,).$$

We will abbreviate this by writing

$$\alpha = \max_{\theta \in \Theta_0} P(\text{ Reject } H_0;\, \theta).$$

We may also omit the notation Θ_0 and write this as

$$\alpha = \max_{\theta \in H_0} P(\text{ Reject } H_0;\, \theta).$$

The maximum here is taken over all values of the parameter possibilities given by the null hypothesis. For example, if $H_0 : \mu \leq 3$, we may use the even more specific notation

$$\alpha = \max_{\mu \leq 3} P(\text{ Reject } H_0;\, \mu)$$

rather than saying that $\mu \in H_0$. Note that, alone, the notation

$$P(\text{ Reject } H_0;\, \mu),$$

which can be read as

"the probability we reject H_0 when the parameter is μ",

does not indicate that any error has been made. If μ is in the region indicated by the null hypothesis, then we are in Type I error territory. By setting an α using the maximum, we are controlling a "worst case scenario" or the highest probability of making a Type I error.

Fortunately, we do not find ourselves with two different definitions for the level of significance of a test. Our original definition from Section 4.3.1 for the simple null hypothesis is just a special case of this one. If we want to maximize a function $f(\mu)$ over all μ such that $\mu = 3$, for example, then we just want to plug in $\mu = 3$. In

this case, we are maximizing a function over a single point. This point is the location of the maximum as well as the location of the minimum! So, for the null hypothesis $H_0 : \mu = 3$,

$$\alpha = \max_{\mu \in H_0} P(\text{ Type I error; } \mu) = P(\text{ Type I error; } \mu = 3).$$

We are ready for our first composite null hypothesis example.

Example 4.5.1

Suppose that $X_1, X_2, \ldots, X_{10}$ is a random sample of size 10 from the $N(\mu, 1)$ distribution. Further suppose that we wish to find a test of size $\alpha = 0.05$ of

$$H_0 \quad : \quad \mu \leq 3$$

$$H_1 \quad : \quad \mu > 3$$

based on the sample mean $\overline{X}$.

Because of the form of the alternative hypothesis, we wish to reject H_0 if the mean comes across as "large" in our sample. If we decide to estimate μ by $\overline{X}$, a large value of μ will be reflected by a large value of $\overline{X}$. Therefore, the form of the test is again:

$$\text{"Reject } \quad H_0 \quad \text{ if } \quad \overline{X} > c."$$

for some c that is chosen to give a size 0.05 test.

In order to find the value of c, we write

$$\begin{aligned}
0.05 &= \max P(\text{ Type I Error }) \\
&= \max_{\mu \leq 3} P(\text{ Reject } H_0;\, \mu) \\
&= \max_{\mu \leq 3} P(\,\overline{X} > c; \mu\,) \\
&= \max_{\mu \leq 3} P\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} > \frac{c-\mu}{1/\sqrt{10};\, \mu}\right) \quad (\Leftarrow \text{standardize } \overline{X}) \\
&= \max_{\mu \leq 3} P\left(Z > \frac{c-\mu}{1/\sqrt{10}}\right) \text{ where } Z \sim N(0,1) \\
&= \max_{\mu \leq 3} \left[1 - \Phi\left(\frac{c-\mu}{1/\sqrt{10}}\right)\right]
\end{aligned}$$

where $\Phi(\cdot)$ is the cdf of the standard normal distribution.

In order to find the maximum of

$$1 - \Phi\left(\frac{c-\mu}{1/\sqrt{10}}\right),$$

we note that it is an increasing function of μ since

- $\frac{c-\mu}{1/\sqrt{10}}$ is a decreasing function of μ,
- which implies that $\Phi\left(\frac{c-\mu}{1/\sqrt{10}}\right)$ is decreasing since $\Phi(\cdot)$, as a cdf, is an increasing function, thereby increasing as it's argument increases and decreasing as it's argument decreases,
- which implies that $1 - \Phi\left(\frac{c-\mu}{1/\sqrt{10}}\right)$ is increasing in μ.

So, the maximum over all $\mu \leq 3$ occurs at the right endpoint of this interval where $\mu = 3$. Thus, we have

$$0.05 = \max_{\mu \leq 3} \left[1 - \Phi\left(\frac{c-\mu}{1/\sqrt{10}}\right)\right] = 1 - \Phi\left(\frac{c-3}{1/\sqrt{10}}\right)$$

Moving things around, we get

$$\Phi\left(\frac{c-\mu}{1/\sqrt{10}}\right) = 0.95$$

or

$$P\left(Z \leq \frac{c-\mu}{1/\sqrt{10}}\right) = 0.95$$

which again means that

$$\frac{c-\mu}{1/\sqrt{10}} = 1.645$$

or that $c = 3.52$.

In summary, we will reject H_0 in favor of H_1 at the 0.05 level of significance if $\overline{X} > 3.52$.

Let's generalize the notation of the last example a little bit. While we're at it, we'll change the direction of the inequalities in H_0 and H_1 and we'll break down what we're doing into four steps.

Example 4.5.2

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from the $N(\mu, \sigma^2)$ distribution where σ^2 is known. (Throughout this text, unless otherwise specified, the sample size n is always assumed to be known.)

Consider testing the hypotheses

$$H_0 : \mu \geq \mu_0 \qquad \text{versus} \qquad H_1 : \mu < \mu_0$$

for a fixed and known value μ_0. Use level of significance α.

Step One: Choose a statistic on which to base the test.

It makes sense to start with an estimator of μ, though any statistic can technically be an estimator for μ even if it doesn't make sense and is not considered a good estimator! As in the confidence interval Example 3.5.1 of Section 3.5 in Chapter 3, we need to choose something that we can work with when we are computing the probability to find the cutoff c in this hypothesis test.

For this test concerning the mean μ, let's try choosing the sample mean $\overline{X}$.

Step Two: Write down the form of the test.

If H_1 is true, we want to reject H_0. How would H_1 being true be be reflected in our statistic? For this example, H_1 is saying that the true mean μ is smaller than it would be under H_0. If this is true, we should see smaller values of the sample mean than we would under the assumption that H_0 is true.

The form of the test is to reject H_0, in favor of H_1 if

$$\overline{X} < c$$

for some c to be determined.

Step Three: Find the value of c.

Here is where α comes in. We have

$$\begin{aligned}
\alpha &= \max P(\text{ Type I Error }) \\
&= \max_{\mu \geq \mu_0} P(\text{ Reject } H_0 \,;\mu) \\
&= \max_{\mu \geq \mu_0} P(\,\overline{X} < c \,;\mu) \\
&= \max_{\mu \geq \mu_0} P\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < \frac{c-\mu}{\sigma/\sqrt{n}} \,;\mu\right) \quad (\Leftarrow \text{standardize } \overline{X}) \\
&= \max_{\mu \geq \mu_0} P\left(Z < \frac{c-\mu}{\sigma/\sqrt{n}}\right)
\end{aligned}$$

where $Z \sim N(0,1)$.

Notice how we have dropped the "semicolon μ". We needed the information that "the true mean is μ" in order to transform the random variable $\overline{X}$ into something whose distribution we can handle when finding probabilities and critical values. In this example, we have made a transformation to a standard normal random variable and, at this point, we no longer need to know the value of μ to compute this probability.

Because $(c-\mu)/(\sigma/\sqrt{n})$ is decreasing in μ, as μ gets larger, the quantity gets smaller and it becomes less and less likely for the random variable Z to be below it. Thus, the probability $P(Z < (c-\mu)/(\sigma/\sqrt{n}))$ is a decreasing function of μ. To maximize it, we need to take μ as small as possible. When looking at the region where $\mu \geq \mu_0$, the smallest value for μ is clearly μ_0. Thus, we have that

$$\alpha = \max_{\mu \geq \mu_0} P\left(Z < \frac{c-\mu}{\sigma/\sqrt{n}}\right) = P\left(Z < \frac{c-\mu_0}{\sigma/\sqrt{n}}\right).$$

Using our previously established critical value notation from Section 3.1.1, we must have

$$\frac{c-\mu_0}{\sigma/\sqrt{n}} = z_{1-\alpha}$$

since $z_{1-\alpha}$ is the value that captures area $1-\alpha$ under the standard normal distribution to the right and therefore area α to the left. (It might help to draw a picture!)

Solving for c, we have

$$c = \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}.$$

Step Four: Give the conclusion.

Pulling it all together, our test of size or level α for

$$H_0 : \mu \geq \mu_0 \qquad \text{versus} \qquad H_1 : \mu < \mu_0,$$

based on the statistic $\overline{X}$, is to reject H_0 if

$$\overline{X} < \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}.$$

4.5.1 Where There's an α, There's a β

We used the symbol α to denote the probability of making a Type I error for a simple null hypothesis and the maximum probability of making this error for a composite null hypothesis. We are now going to turn our attention to Type II error probabilities.

Definition 4.5.2

For a simple alternative hypothesis, we use β to denote the probability of a Type II error:

$$\beta = P(\text{ Type II error }) = P(\text{ Fail to Reject } H_0 \text{ when it's false }).$$

Recall that a Type II error exists in the "universe where H_0 is false". We have failed to reject H_0 when we should have because it is false and H_1 is true. For a simple alternative hypothesis like $H_1 : \mu = \mu_1$, H_1 being true obviously means that $\mu = \mu_1$ and we can use this fact when computing the probability of making a Type II error. If H_1 is a composite hypothesis, we'll want to control the the maximum probability of making a Type II error over every simple hypothesis contained in H_1.

Definition 4.5.3

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution that depends on a parameter θ that lives in a parameter space Θ.

Let Θ_0 be some subset of Θ.

Consider testing

$$H_0 : \theta \in \Theta_0 \qquad \text{versus} \qquad H_1 : \theta \in \Theta \setminus \Theta_0.$$

We use β to denote the maximum probability of making a Type II error:

$$\beta = \max_{\theta \in \Theta \setminus \Theta_0} P(\text{ Fail to reject } H_0 \text{ when the parameter is } \theta\,).$$

We will abbreviate this by writing

$$\beta = \max_{\theta \in \Theta \setminus \Theta_0} P(\text{ Fail to reject } H_0;\, \theta).$$

We will sometimes omit the parameter set notation and write this as

$$\beta = \max_{\theta \in H_1} P(\text{ Fail to reject } H_0;\, \theta).$$

As a probability of making an error, we obviously want β to be a small number. Equivalently, we want $1 - \beta$ to be large. Note that

$$\begin{aligned} 1 - \beta &= 1 - \max_{\theta \in H_1} P(\text{ Type II error; } \theta) \\ &= 1 - \max_{\theta \in H_1} P(\text{ Fail to reject } H_0;\, \theta) \\ &= \min_{\theta \in H_1} [1 - P(\text{ Fail to reject } H_0;\, \theta)] \\ &= \min_{\theta \in H_1} P(\text{ Reject } H_0;\, \theta). \end{aligned}$$

Again note that if θ is in H_1, we should be rejecting H_0, and so we want this quantity to be large. We call α the level of significance of a test. β doesn't really have a name like this, however, $1 - \beta$ does.

Definition 4.5.4

$1 - \beta$ is called the **power** of the test.

High power is a good thing.

In everything we've done so far, the rejection regions for our tests have been defined by

- a "test statistic" used (ex:$\overline{X}$),
- the form of the alternative hypothesis, and
- the level of significance of the test.

As it turns out, after all of this is determined, the probability of a Type II error is already "locked in".

Example 4.5.3

Suppose that $X_1, X_2, \ldots, X_{10}$ is a random sample of size 10 from the $N(\mu, 1)$ distribution. Suppose that we wish to find a test of size $\alpha = 0.05$ of

$$H_0 : \mu = 3 \qquad \text{versus} \qquad H_1 : \mu > 3$$

based on the sample mean $\overline{X}$.

We have determined, in Section 4.4, that this test is to

$$\text{"Reject } H_0 \text{ if } \overline{X} > 3.52\text{."}$$

In this case,

$$\begin{aligned}
\beta &= \max P(\text{ Type II Error }) \\
&= \max_{\mu>3} P(\text{ Fail to reject } H_0;\ \mu) \\
&= \max_{\mu>3} P(\ \overline{X} \leq 3.52 \text{ when the parameter is } \mu\ ) \\
&= \max_{\mu>3} P\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq \frac{3.52 - \mu}{1/\sqrt{10}} \text{ when the parameter is } \mu \right) \quad (\Leftarrow \text{standardize } \overline{X}) \\
&= \max_{\mu>3} P\left(Z \leq \frac{3.52 - \mu}{1/\sqrt{10}} \right) \text{ where } Z \sim N(0,1) \\
&= \max_{\mu>3} \Phi\left(\frac{3.52 - \mu}{1/\sqrt{10}} \right)
\end{aligned}$$

As $\Phi\left(\frac{3.52-\mu}{1/\sqrt{10}}\right)$ is decreasing in μ, it is maximized when μ is smallest. Technically, there is no "smallest" μ since μ is strictly greater than 3 and so it seems that there is no "largest" Type II error probability. In this case though, we define β to be the "least upper bound" on the probability of a Type II error. This least upper bound is known in mathematics as a **supremum** and it denoted with the abbreviation "sup" (pronounced

"soup"). It is the maximum of a function over a set when the maximum is, technically, not quite in the set!

We write,

$$\beta = \sup_{\mu>3} \Phi\left(\frac{3.52-\mu}{1/\sqrt{10}}\right) = \Phi\left(\frac{3.52-3}{1/\sqrt{10}}\right) = \Phi(1.645) = 0.95.$$

Ouch! We set the Type I error probability to be $\alpha = 0.05$ and we ended up with a really large Type II error probability. At first glance it might appear that we will always get $\beta = 1 - \alpha$. This is not the case.

Type I errors and Type II errors do not exist "in the same universe". H_0 is either true or or it is false. There is no "probability it is true" or "probability it is false". If it is true, our derived test will reject it with probability α and therefore will fail to reject it with probability $1 - \alpha$. If H_0 is false, we will fail to reject it with probability β and we will reject it with probability $1 - \beta$. In the example that we just completed, we saw what appears to be a "1 minus relationship" between α and β but you will not see this for all distributions/rejection rules or even for the normal distribution if we look at a simple null hypothesis versus a simple alternative hypothesis. The "1 minus relationship" in our example came from the fact that the maximizing value of μ for both calculations occurred right on the boundary between the regions described by H_0 and H_1.

While there is not necessarily a "1 minus relationship" between α and β, there is an inverse relationship in the sense that if one is forced to be very small the other one will end up larger. For example, if you want α to be very small then you will make a more extreme cutoff value for rejection in order to make it very difficult to reject H_0. By making it difficult to reject H_0 you are increasing the probability of not rejecting it when you should!

In Example 4.5.3, we set the Type I error probability and this caused us to have a large Type II error probability. As noted, once we set α, β was "locked in". In practice, if we would like to control both types of error, we would have to free up another parameter in the model and that parameter is the sample size. In the example, we would not have $n = 10$ but rather a general n in

$$\beta = \max_{\mu>3} P\left(Z \leq \frac{3.52-\mu}{1/\sqrt{n}}\right).$$

We could then choose n to get close to the desired β.

Unless otherwise specified, for every single random sample $X_1, X_2, \ldots, X_n$ in this text, n is assumed to be

known.

4.6 Non-Normal Distribution Hypothesis Tests

As stated towards the beginning of this Chapter, you should now be ready to derive hypothesis tests for parameters other than the mean and distributions other than the normal distribution without further guidance. It's a procedure after all and it is always the same. Let's try an example.

Example 4.6.1

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from the exponential distribution with rate λ. Derive a test of size (level) α for the hypotheses

$$H_0 : \lambda \leq \lambda_0 \qquad \text{versus} \qquad H_1 : \lambda > \lambda_0$$

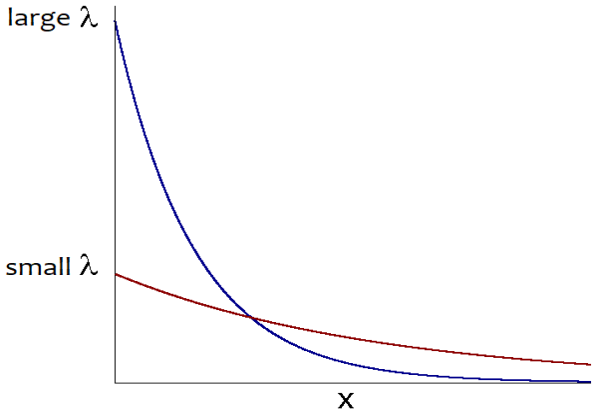
for some fixed and known $\lambda_0 > 0$. Base your test on $X_{(1)}$, the minimum value in the sample.

Step One: Choose a statistic on which to base the test.

Since the test concerns λ, it would make sense to start with an estimator of λ. As noted when we studied confidence intervals in Chapter 3, there are countless possibilities for estimators especially when we are not restricting ourselves to those with particular characteristics like unbiasedness. For now, we will just select something we can work with. Indeed, for this example, we are given that the test should be based on the sample minimum $X_{(1)}$. In Chapter 6 we will have actual methods for choosing statistics.

Step Two: Write down the form of the test.

We want to reject H_0 if, when looking at $X_{(1)}$, it seems that H_1 is true. How does H_1 being true get reflected in the statistic $X_{(1)}$? If H_1 is true, the parameter λ is larger than we thought is was under H_0. Consider a graph of the exponential pdf $f(x) = \lambda e^{-\lambda x} I_{(0,\infty)}(x)$.



It starts at a height of λ when $x = 0$ and goes down, well, exponentially. Since the total area under a pdf is 1, the higher it starts, the faster it must go down. The red curve (smaller λ) still has some significant area out in the right tail, as opposed to the blue curve which has dropped off faster and has more of it's area up against the y-axis. The "bulk" of the probability in the case of a larger λ is towards smaller values on the x-axis. In other words, the sampled values $X_1, X_2, \ldots, X_n$ will tend to be smaller when λ is larger. This makes, for example, the sample mean $\overline{X}$ tend towards smaller values, as well as the minimum, maximum, and many other quantities of interest that can be computed from the sample.

Again, we want to reject $H_0 : \lambda \leq \lambda_0$, in favor of $H_1 : \lambda > \lambda_0$ if λ seems large when we observe the sample minimum $X_{(1)}$. As a large lambda will tend to produce a sample with a small minimum, the form of our test is to

$$\text{"Reject } H_0 \text{ if } X_{(1)} < c\text{."}$$

for some c to be determined in Step Three.

Step Three: Find the value of c.

We find the value of c using α.

$$\begin{aligned} \alpha &= \max P(\text{ Type I Error }) \\ &= \max_{\lambda \leq \lambda_0} P(\text{ Reject } H_0 \,;\, \lambda) \\ &= \max_{\lambda \leq \lambda_0} P(\, X_{(1)} < c \,;\, \lambda) \end{aligned}$$

For our previous examples, the next step in this calculation was to "standardize $\overline{X}$". As mentioned, this was because $\overline{X}$ had a normal distribution and we compute probabilities for normally distributed random

variables by standardizing them and turning them into standard normal random variables. This is not what we need to do here.

We started this example with the assumption that $X_1, X_2, \ldots, X_n$ are iid from an exponential distribution. The λ in

$$P(\, X_{(1)} < c \,; \lambda)$$

says that the rate parameter for the iid exponential random variables is λ. (Recall that it can be read "when the parameter is λ".) We know, from Section 1.4.2 in Chapter 1, that if $X_1, X_2, \ldots, X_n \overset{iid}{\sim} exp(rate = \lambda)$, then $X_{(1)} \sim exp(rate = n\lambda)$. This means that

$$P(\, X_{(1)} < c \,; \lambda) \overset{continuous}{=} P(\, X_{(1)} \leq c \,; \lambda) = 1 - e^{-n\lambda c}.$$

(Here, we have used the fact that the cdf for an exponential distribution with rate λ is $F(x) = 1 - e^{-\lambda x}$.)

Going back to the α calculation, we now have

$$\begin{aligned} \alpha &= \max_{\lambda \leq \lambda_0} P(\, X_{(1)} < c \,; \lambda) \\ &= \max_{\lambda \leq \lambda_0} [1 - e^{-n\lambda c}] \\ &= 1 - e^{-n\lambda_0 c} \end{aligned}$$

since $1 - e^{-n\lambda c}$ is an increasing function of λ. (c is presumed to be positive. Otherwise, the rejection rule $X_{(1)} < c$ won't make any sense since the minimum of a random sample of exponentials can not take on negative values.)

Solving for c gives us

$$c = -\frac{1}{n\lambda_0} \ln(1 - \alpha).$$

Step Four: Give the conclusion.

Pulling it all together, our test of size or level α for

$$H_0 : \lambda \leq \lambda_0 \qquad \text{versus} \qquad H_1 : \lambda > \lambda_0,$$

based on the statistic $X_{(1)}$, is to reject H_0 if

$$X_{(1)} < -\frac{1}{n\lambda_0} \ln(1 - \alpha).$$

We now consider the same setup as the previous example and will derive another test of size α, this time based on the sample mean $\overline{X}$ instead of the sample minimum. This statistic is going to be a little harder to work with

since

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} exp(rate = \lambda) \qquad \Rightarrow \qquad \overline{X} \sim \Gamma(n, n\lambda).$$

and it is difficult to compute closed-form probabilities involving the gamma distribution.[2]

We can, of course, approximate probabilities and critical values for the gamma distribution numerically using a computer. This wasn't always the case though and, historically, people relied on tabulated values that were worked out by hand by someone else. Recall the standard normal table in Appendix C. Although it is a table of standard normal distribution probabilities, we are able to use it for computing probabilities for "non-standard" normal random variables through transformations. It would be quite unfortunate if we needed to carry around an uncountably infinite collection of tables to deal with all possible normal distributions! Similarly, we don't want to carry around an uncountably infinite collection of gamma tables, one associated with each (α, β) pair. Note that, in this exponential example, we have

$$\overline{X} \sim \Gamma(n, n\lambda) \qquad \Rightarrow \qquad n\lambda\overline{X} \sim \Gamma(n, 1).$$

That is, we can make a transformation that allows us to look up probabilities for $\overline{X}$ with only one unknown integer-valued parameter. This is a little more practical. As previously mentioned in Chapter 3, the χ^2-distribution is central to a large number of results in Statistics. If we take our transformation further to get

$$2n\lambda\overline{X} \sim \Gamma\left(n, \frac{1}{2}\right) = \Gamma\left(\frac{2n}{2}, \frac{1}{2}\right) = \chi^2(2n),$$

we could appeal to a χ^2-table for probabilities and critical values as in the following example.

Example 4.6.2

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from the exponential distribution with rate λ. Derive a test of size (level) α for the hypotheses

$$H_0 : \lambda \leq \lambda_0 \qquad \text{versus} \qquad H_1 : \lambda > \lambda_0$$

for some fixed and known $\lambda_0 > 0$, based on the sample mean $\overline{X}$. Leave your test in terms of a χ^2- critical value.

Step One: Choose a statistic on which to base the test.

This has been done for us. The statistic is $\overline{X}$.

[2]We established this claim about the distribution of the sample mean from an exponential distribution in the "super important note" at the end of Section 3.5 of Chapter 3.

Step Two: Write down the form of the test.

We want to reject H_0, in favor of H_1, if $\overline{X}$ lends support to the idea that λ is "large". As per our discussion in Example 4.6.1, a large value of λ will tend to produce smaller sample means. Thus, the form of the test will be to

$$\text{"Reject } H_0 \text{ if } \overline{X} < c\text{."}$$

for some c to be determined.

Step Three: Find the value of c.

$$\begin{aligned} \alpha &= \max P(\text{ Type I Error }) \\ &= \max_{\lambda \leq \lambda_0} P(\text{ Reject } H_0 \,;\lambda) \\ &= \max_{\lambda \leq \lambda_0} P(\,\overline{X} < c\,;\lambda) \end{aligned}$$

We were asked to make a test in terms of a χ^2 critical value. We have just seen that we can do this for this exponential sample by multiplying $\overline{X}$ by $2n\lambda$. So that we don't change anything, we will do the same thing to both sides of the inequality.

$$\begin{aligned} \alpha &= \max_{\lambda \leq \lambda_0} P(\,\overline{X} < c\,;\lambda) \\ &= \max_{\lambda \leq \lambda_0} P(\,2n\lambda\overline{X} < 2n\lambda c\,;\lambda) \end{aligned}$$

Define W to be $2n\lambda\overline{X}$. We know that $W \sim \chi^2(2n)$. Since this distribution does not depend on λ, we no longer need to carry that information through our probability statements. We have

$$\alpha = \max_{\lambda \leq \lambda_0} P(W < 2n\lambda c).$$

The probability here is an increasing function of λ since larger values of λ produce larger values of $2n\lambda c$, making it easier and easier for W to be below $2n\lambda c$. Once again, we are able to maximize a probability without Calculus by simply plugging in the largest value of λ, which is λ_0.

$$\alpha = \max_{\lambda \leq \lambda_0} P(W < 2n\lambda c) = P(W < 2n\lambda_0 c)$$

Using our critical value notation established in Chapter 3, this means that we must have

$$2n\lambda_0 c = \chi^2_{1-\alpha,2n},$$

giving us

$$c = \frac{\chi^2_{1-\alpha,2n}}{2n\lambda_0}.$$

Step Four: Give the conclusion.

Pulling it all together, our test of size or level α for

$$H_0 : \lambda \geq \lambda_0 \qquad \text{versus} \qquad H_1 : \lambda < \lambda_0,$$

based on the statistic $\overline{X}$, is to reject H_0 if

$$\overline{X} < \frac{\chi^2_{1-\alpha,2n}}{2n\lambda_0}.$$

For a random sample $X_1, X_2, \ldots, X_n$ from the exponential distribution with rate λ, we have now seen two different tests for the hypotheses

$$H_0 : \lambda \geq \lambda_0 \qquad \text{versus} \qquad H_1 : \lambda < \lambda_0.$$

Which test is better? We'll find out in the next Section!

4.7 Power Functions

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with parameter θ taking values in a parameter space Θ. Further suppose that we wish to test the hypotheses

$$H_0 : \theta \in \Theta_0 \qquad \text{versus} \qquad H_1 : \theta \in \Theta \setminus \Theta_0.$$

A *power function* is useful for comparing two different size α tests.

Definition 4.7.1

The **power function** for a hypothesis test concerning θ, will be denoted and defined as

$$\gamma(\theta) = P(\text{ Reject } H_0; \theta\,)$$

for all $\theta \in \Theta$.

It is important to note that a **power function and the power of a test are two different things**. The power of a

test is the quantity $1-\beta$, which is related to the power function as follows.

$$\begin{aligned}
1-\beta &= 1-\max P(\text{ Type II Error }) \\
&= 1-\max P(\text{ Fail to reject } H_0 \text{ when true }) \\
&= 1-\max_{\theta\in\Theta\setminus\Theta_0} P(\text{ Fail to reject } H_0;\theta\,) \\
&= 1-\max_{\theta\in\Theta\setminus\Theta_0} [1-P(\text{ Reject } H_0;\theta\,)] \\
&= 1-\max_{\theta\in\Theta\setminus\Theta_0} [1-\gamma(\theta)] \\
&= 1-\left[1-\min_{\theta\in\Theta\setminus\Theta_0}\gamma(\theta)\right] = \min_{\theta\in\Theta\setminus\Theta_0}\gamma(\theta).
\end{aligned}$$

Note that if H_1 is the simple hypothesis $H_1:\theta=\theta_1$, there is only one point to minimize over. θ being in the set $\Theta\setminus\Theta_0=\{\theta_1\}$ means that $\theta=\theta_1$. Thus

$$1-\beta=\gamma(\theta_1)$$

when H_1 has this form.

Let's find our first power function.

Example 4.7.1

In Example 4.5.2, we had a random sample $X_1, X_2, \ldots, X_n$ from the $N(\mu,\sigma^2)$. We derived a size α test for

$$H_0:\mu\geq\mu_0 \qquad \text{versus} \qquad H_1:\mu<\mu_0.$$

Our rule for the test was to

"Reject H_0 if $\overline{X}<\mu_0+z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$."

Let's find the power function for this test. We have

$$\begin{aligned}
\gamma(\mu) &= P(\text{ Reject } H_0;\mu\,) \\
&= P\left(\overline{X}<\mu_0+z_{1-\alpha}\tfrac{\sigma}{\sqrt{n}};\mu\right).
\end{aligned}$$

As a linear combination of normals, the sample mean $\overline{X}$ has a normal distribution. If we are assuming the mean is μ, we have, specifically, that $\overline{X}\sim N(\mu,\sigma^2/n)$. We will compute the above probability by

standardizing $\overline{X}$ to a $N(0,1)$ random variable.

$$\begin{aligned}\gamma(\mu) &= P\left(\overline{X} < \mu_0 + z_{1-\alpha}\tfrac{\sigma}{\sqrt{n}}; \mu\right) \\ &= P\left(\tfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} < \tfrac{\mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}} - \mu}{\sigma/\sqrt{n}}; \mu\right) \\ &= P\left(Z < \tfrac{\mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}} - \mu}{\sigma/\sqrt{n}}\right)\end{aligned}$$

where $Z \sim N(0,1)$.

We have dropped the "semicolon μ" because the distribution of Z does not depend on μ so we no longer need this information. Using $\Phi(z) = P(Z \leq z)$ to denote the cdf for the $N(0,1)$ distribution, we have that the power function is

$$\gamma(\mu) = \Phi\left(\frac{\mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}} - \mu}{\sigma/\sqrt{n}}\right).$$

For this particular example, we are unable to go further to get a closed-form expression. Here, μ_0, σ^2, n, and $z_{1-\alpha}$ are known quantities. For concreteness, suppose that the sample size is $n = 10$, that $\sigma^2 = 1$, and that α was given to be 0.10. In this case, we can use Table C.1 from Appendix C to get $z_{1-\alpha} = z_{0.90} \approx -1.28$. Further suppose that we are testing

$$H_0 : \mu \geq 2 \qquad \text{versus} \qquad H_1 : \mu > 2.$$

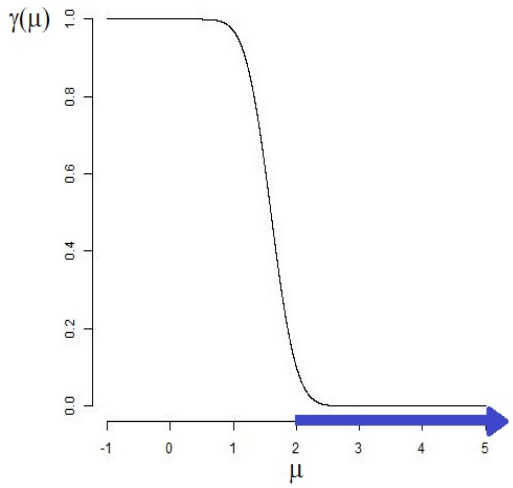
We then have that the power function is

$$\gamma(\mu) = \Phi\left(\frac{2 - 1.28\frac{1}{\sqrt{10}} - \mu}{1/\sqrt{10}}\right) = \Phi(2\sqrt{10} - 1.28 - \sqrt{10}\,\mu).$$

We can evaluate this function at various values of μ. For example,

$$\gamma(1.28) = \Phi(2\sqrt{10} - 1.28 - (\sqrt{10})(1.28)) \approx \Phi(-1.03) \overset{Table\ C.1}{=} 1 - 0.8485 = 0.1515.$$

With many more evaluations, we can see that the power function looks like this.

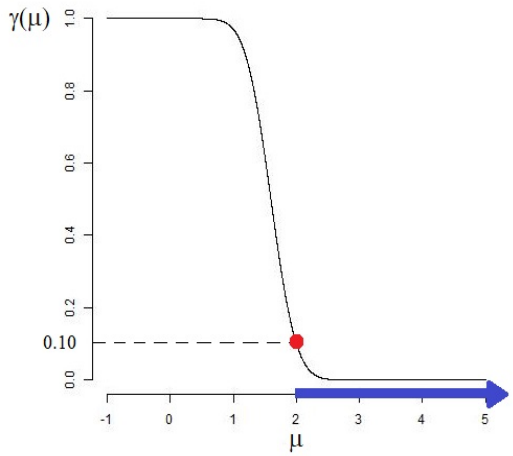


The power function gives the probability that we reject H_0 assuming that the random sample is from the $N(\mu, 1)$ distribution. For values of μ in the region indicated with a blue arrow, the null hypothesis $H_0 : \mu \geq 2$ is true. When H_0 is true, we do not want to reject H_0 and so we are glad to see that $\gamma(\mu)$, which is the probability of rejecting H_0 when the parameter is μ, is relatively small in this region. In contrast, for values of μ not in the blue arrow region, the null hypothesis $H_0 : \mu \geq 2$ is false. When H_0 is false, we want to reject it, and so we are glad to see that $\gamma(\mu)$ is larger in this region.

Note that $\alpha = 0.10$ is the maximum probability of rejecting H_0 when H_0 is true. Since the power function gives the probability of rejecting H_0, we have that

$$0.10 = \max_{\mu \geq 2} P(\text{ Reject } H_0; \mu\,) = \max_{\mu \geq 2} \gamma(\mu) = \gamma(2).$$

and we can see α as a point on our graph.



Power functions are not necessarily decreasing. For example, if we sketched a power function for a test for $H_0 : \mu \leq 2$ versus $H_1 : \mu > 2$, we would hope to see something like the plot in the above example, except that we would expect it to be mirrored horizontally over the line $\mu = 2$. Power functions are also not necessarily monotone and can look like almost anything!

In the next example, we will use power functions to compare two tests.

Example 4.7.2

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from the exponential distribution with rate λ. In Section 4.6, we developed two different hypothesis tests of size α for

$$H_0 : \lambda \leq \lambda_0 \qquad \text{versus} \qquad H_1 : \lambda > \lambda_0.$$

The first test, which we will call "Test 1" was to

$$\text{"Reject } H_0 \text{ if } X_{(1)} < -\tfrac{1}{n\lambda_0}\ln(1-\alpha)\text{."}$$

The second test, which we will call "Test 2" was to

$$\text{"Reject } H_0 \text{ if } \overline{X} < \tfrac{\chi^2_{1-\alpha,2n}}{2n\lambda_0}\text{."}$$

The power function for Test 1 is

$$\begin{aligned} \gamma_1(\lambda) &= P(\text{ Reject } H_0 \text{ using Test 1}; \lambda\,) \\ &= P\left(X_{(1)} < -\tfrac{1}{n\lambda_0}\ln(1-\alpha); \lambda\right) \end{aligned}$$

The "semicolon λ" tells us that the random sample $X_1, X_2, \ldots, X_n$ came from the exponential distribution with rate λ. This implies that the sample minimum, $X_{(1)}$, has an exponential distribution with rate $n\lambda$. Since the cdf for an exponential distribution with rate $n\lambda$ is $F(x) = 1 - e^{-n\lambda x}$, we have that

$$\begin{aligned}
\gamma_1(\lambda) &= P\left(X_{(1)} < -\tfrac{1}{n\lambda_0}\ln(1-\alpha);\lambda\right) \\
&= 1 - e^{-n\lambda\left(-\frac{1}{n\lambda_0}\ln(1-\alpha)\right)} \\
&= 1 - e^{\ln(1-\alpha)^{\lambda/\lambda_0}} \\
&= 1 - (1-\alpha)^{\lambda/\lambda_0}.
\end{aligned}$$

For concrete values for α and λ_0, we can plot this as a function of λ.

The power function for Test 2 is

$$\gamma_2(\lambda) = P\left(\overline{X} < \tfrac{\chi^2_{1-\alpha,2n}}{2n\lambda_0};\lambda\right).$$

The "semicolon λ" again tells us that the random sample $X_1, X_2, \ldots, X_n$ came from the exponential distribution with rate λ. This implies that the sample mean, $\overline{X}$, has a $\Gamma(n, n\lambda)$ distribution. We can write this power function in terms of a favored χ^2 random variable as follows.

$$\begin{aligned}
\gamma_2(\lambda) &= P\left(\overline{X} < \tfrac{\chi^2_{1-\alpha,2n}}{2n\lambda_0};\lambda\right) \\
&= P\left(2n\lambda\overline{X} < 2n\lambda\,\tfrac{\chi^2_{1-\alpha,2n}}{2n\lambda_0};\lambda\right) \\
&= P(W < \tfrac{\lambda}{\lambda_0}\chi^2_{1-\alpha,2n})
\end{aligned}$$

where $W \sim \chi^2(2n)$. Unlike Test 1, we do get a closed form expression for this power function. For concreteness, let us suppose that $n = 10$, $\alpha = 0.05$, and that we are testing

$$H_0 : \lambda \leq 1 \qquad \text{versus} \qquad H_1 : \lambda > 1.$$

From Table C.3 in Appendix C, we have that $\chi^2_{1-\alpha,20} = \chi^2_{0.95,20} = 10.851$. The power function becomes

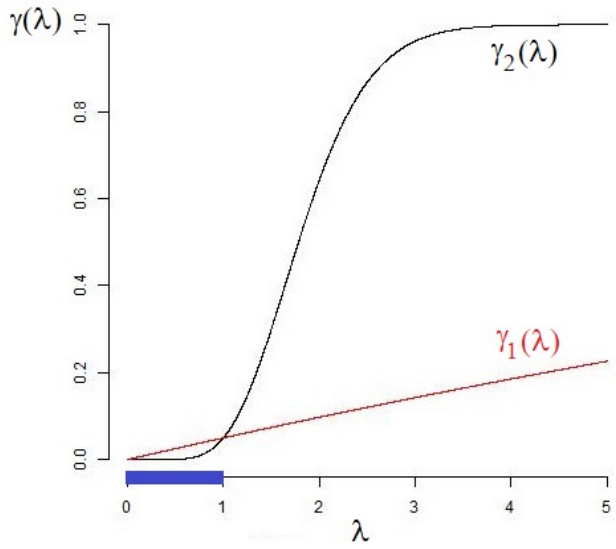
$$\gamma_2(\lambda) = P(W < 10.851\lambda).$$

Unlike the z-table, the χ^2-table is not complete enough to allow us to start evaluating $\gamma_2(\lambda)$ for many values of λ. To compute the power function at, for example, the point 1.2, note that

$$\gamma_2(1.2) = P(W < (10.851)(1.2)) = P(W < 13.0212) = \int_0^{13.0212} f_W(w)\,dw$$

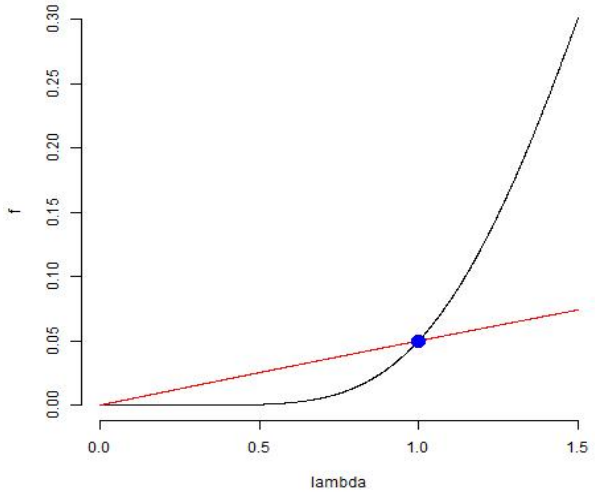
where $f_W(w)$ is the $\chi^2(20) = \Gamma(10, 1/2)$ pdf. Numerical integration gives us that $\gamma_2(1.2) \approx 0.1235$. Continuing with many evaluations of $\gamma_2(\lambda)$, we can sketch a graph of the power function. This is shown in the following Figure along with the power function for Test 1 (in red) when assuming that $\lambda_0 = 1$ and

$\alpha = 0.5$.

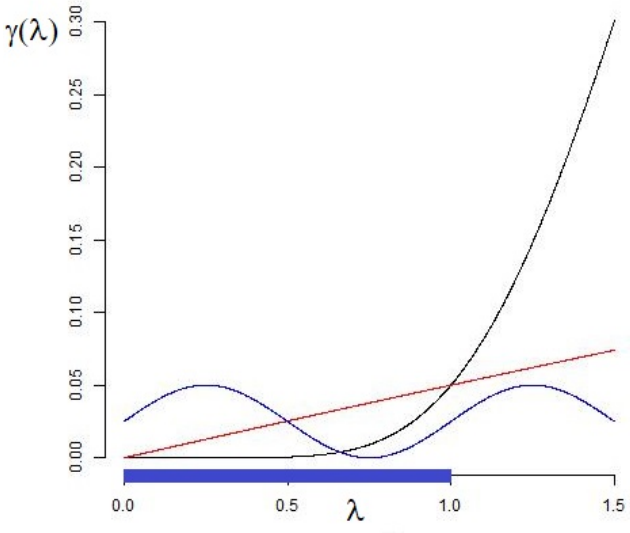


The region where H_0 is true is the region indicated in blue on the λ-axis. This is where we want the power functions, representing the probability that we reject H_0, to be low. For the rest of the λ-axis, H_0 is false and we want the probability of rejecting it to be high. In this sense, both tests are doing the right thing but Test 2 is doing much better as it's power function is lower where it should be lower and gets higher much faster than the power function for Test 1 where it should be higher. Indeed, the red curve does approach 1 as λ goes to ∞ but it takes a very long time to get there! It is not surprising that the test based on the sample mean $\overline{X}$ performs much better than the test based on the sample minimum $X_{(1)}$. For an observed sample of any decent size from an exponential distribution, the minimum is going to be hovering near 0 regardless of the value of λ that was used in producing the sample. In contrast, the sample mean will be more responsive to changes in λ.

Recall that α is the maximum probability of rejecting H_0 when it is true. For our exponential example, we can zoom in on the power function and see that the maximum value for both power functions over the region $0 \leq \lambda \leq 1$ is indeed 0.05.



As mentioned, power functions can take on many shapes. It is entirely possible to make up a third test of size $\alpha = 0.05$ for the exponential example with a power function that looks like the curve shown below in blue.



By design, the maximum value for the Test 3 power function over the region $0 \leq \lambda \leq 1$ is also at 0.05 even though it does not correspond to a point of intersection in this case. Because the function is not definitively lower where is should be lower and higher when it should be high, it is difficult to compare this test with the other two. Test 3 appears to be the worst of all three tests when λ is in (approximately) the region $(0, 0.5)$. It appears to be better than Test 1 but worse than Test 2 in (approximately) the region $(0.5, 0.65)$. It appears to be the best test out of three in the region $(0.65, 1)$ and the worst test in the region $(1, \infty)$. Surely we would prefer to use this test for

$$H_0 : \lambda \leq 1 \qquad \text{versus} \qquad H_1 : \lambda > 1$$

when $0.65 < \lambda < 1$, no? The problem is that the entire reason we are doing the test is that we don't know the location of λ! In Chapter 6, we will learn how to find a definitively "best test" whose power function is lower than that of any other possible test where it should be and higher than that of any other possible test where it should be. First, however, we will up our estimation game by moving from "common sense guesses" to formal methods of estimation.

Chapter 4 Exercises

1. Let $X_1, X_2, \ldots, X_{16}$ be a random sample from the $N(\mu, 9)$ distribution. Find a test of size 0.05 of

$$\begin{aligned} H_0 &: \mu = 3 \\ H_1 &: \mu \neq 3 \end{aligned}$$

   based on the sample mean $\overline{X}$.

2. Let $X_1, X_2, \ldots, X_n$ be a random sample from the $N(\mu, \sigma^2)$ distribution. Assume that both μ and σ^2 are unknown.

   (a). Derive a test of size (level of significance) α for

$$H_0 : \mu = \mu_0 \qquad \text{versus} \qquad H_1 : \mu_1 \neq \mu_0$$

   based on the sample mean $\overline{X}$ and sample variance S^2.

   (b). Carry out the test you derived in part (a) in the case that $\mu_0 = 1$ with the following "data set".

$$1.5,\ 3.2,\ 0.6,\ 2.4,\ -0.1.$$

   Use $\alpha = 0.05$.

3. Let $X_{(1)}, X_{(2)}, X_{(3)}, X_{(4)}$ be the order statistics of a random sample of size $n = 4$ from the $unif(0, \theta)$ distribution where $\theta > 0$. Consider the rejection rule: "The hypothesis $H_0 : \theta = 1$ is rejected in favor of $H_1 : \theta > 1$ if the observed $X_{(4)} \geq c$."

   (a). Find the constant c so that the level of significance (the "size" of the test) is $\alpha = 0.05$.

   (b). Determine the power function of the test.

4. ♣ ***** Make a pooled variance test**

5. Consider a random sample $X_1, X_2, \ldots, X_n$ from any distribution that depends on a parameter θ. Suppose we want to test the hypotheses

$$\begin{aligned} H_0 &: \theta \leq 3 \\ H_1 &: \theta > 3 \end{aligned}$$

   at level of significance α. Two of your friends, Fred and Wilma, are arguing over who has a better decision rule for performing the test. The two decision rules result in two power functions $\gamma_{Fred}(\mu)$ and

$\gamma_{Wilma}(\mu)$ shown in the graph below. ♣ *******NEED TO ADD GRAPH**

(a). Who has the better test?

(b). Where is α on this graph?

6. Consider the distribution with pdf

$$f(x;\theta) = 1 - \theta^2(x - \frac{1}{2}), \qquad 0 < x < 1, \qquad -1 < \theta < 1.$$

Find a test of size α for

$$H_0 : \theta = 0 \qquad \text{versus} \qquad H_1 : \theta \neq 0$$

based on a sample of size 1.

7. Let $X_1, X_2, \ldots, X_n$ be a random sample from the exponential distribution with rate λ. In Example 4.6.1, we derived a hypothesis test of size/level α for testing

$$H_0 : \lambda\lambda_0 \qquad \text{versus} \qquad H_1 : \lambda > \lambda_0$$

based on the minimum value of the sample.

Now, find a test of size α based on the maximum $X_{(n)}$. Compare the power functions for both tests. Can you conclude anything about which is the better test?

8. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the $N(0, \sigma^2)$ distribution.

(a). Derive a test of size α for $H_0 : \sigma^2 = \sigma_0^2$ versus $H_1 : \sigma^2 > \sigma_0^2$.

(b). Express the power function of your test from the previous problem in terms of the chi-squared distribution.

9. Consider a random sample of size n from the $unif(0, \theta)$ distribution. Find a test of size α for $H_0 : \theta \geq \theta_0$ versus $H_1 : \theta < \theta_0$ based on the sample maximum $X_{(n)}$.

Chapter 5 Estimation

So far, we have estimated means and a couple of other parameters by kind of guessing and using intuitive "common sense estimators". For example, we estimated the mean μ of several distributions with sample means. This seems like a good idea but at least two questions arise.

- Why is it a good idea?
- What are we supposed to do for other types of parameters that do not have an obvious analogue in a sample?

In this Chapter, we will replace guessing with more rigorous techniques for finding estimators.

Since we will be estimating parameters for various distributions, we will extend our notation for pdfs to emphasize these parameters. In general, instead of $f(x)$, we will write $f(x;\theta)$. Note the semicolon.[1] Commas will be reserved for separating multiple x's in joint pdfs and multiple parameters. For example, if X_1, X_2 is a random sample from the $N(\mu, \sigma^2)$ distribution, we will denote the joint pdf as $f(x_1, x_2; \mu, \sigma^2)$ or as $f(\vec{x}; \mu, \sigma^2)$.

5.1 Method of Moments Estimators (MMEs)

Suppose that we have a random sample $X_1, X_2, \ldots, X_n$ from a normal distribution with mean μ. We have decided, several times already, to estimate μ with the sample mean:

$$\widehat{\mu} = \overline{X}.$$

Now suppose that we have a random sample $X_1, X_2, \ldots, X_n$ from an exponential distribution with rate λ. The mean of this distribution is $\mu = \mathsf{E}[X_1] = 1/\lambda$. Since the mean of the distribution is one over λ, it might make sense to estimate λ with one over the sample mean:

$$\widehat{\lambda} = 1/\overline{X}.$$

We, in fact, did this in Example 2.2.1 of Chapter 2. We saw then that this estimator was a biased estimator for μ, but it still felt like a decent idea.

The estimators in both of these examples are known as *method of moments estimators* (MMEs).[2] In this Section,

[1] Some authors will use a horizontal bar and write a pdf as $f(x|\theta)$.

[2] Some people call them MOMs!

we will formalize the method which will (hopefully!) make it clear how to find MMEs for parameters that are not so "obviously" related to the mean of a distribution.

5.1.1 The Idea

In data oriented statistics, there is usually a population of individuals or objects with a "quantity of interest" that we want to know something about. We take a relatively small sample from this population, study it, and try to conclude things about the larger group. In theoretical statistics, the population is theoretical and the quantity of interest is perfectly modeled by a distribution. For this reason, some things we are about to define use the terms "population" and "distribution" interchangeably.

The idea behind "method of moments estimation" is to equate "population" (or "distribution") *moments* with sample *moments*.

We will need some definitions.

Definition 5.1.1

The k^{th} **population moment** kth moment (**distribution moment**) of X, will be denoted by μ_k, and defined as

$$\mu_k := \mathsf{E}[X^k].$$

Note that $\mu_1 = \mu = \mathsf{E}[X]$.

Some readers may recognize this as a *non-central moment* as opposed to a *central moment* where expectations for powers of X have been centered around $\mu = \mathsf{E}[X]$:

$$\mu'_k := \mathsf{E}[(X - \mu)^k].$$

In this text, when we say "moment", we are referring to a non-central moment by default.

The sample analogue of a moment of a distribution is a *sample moment*.

Definition 5.1.2

The k^{th} "sample moment" of X, denoted by M_k, is defined by

$$M_k = \frac{1}{n}\sum_{i}^{n} X_i^k.$$

Note that $M_1 = \overline{X}$.

Idea

To find a **method of moments estimator**, equate population moments to sample moments and solve for the unknown parameter(s). Use as many moments as you need in order to solve for all parameters.

5.1.2 Examples

Example 5.1.1

Let $X_1, X_2, \ldots, X_n$ be a random sample from the exponential distribution with rate λ. Find the MME of λ.

The first population moment is

$$\mu_1 = \mathsf{E}[X] = 1/\lambda.$$

The first sample moment is

$$M_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}.$$

Equate them

$$\frac{1}{\lambda} \overset{set}{=} \overline{X},$$

solve for λ,

$$\lambda = 1/\overline{X},$$

and "put a hat on it"

$$\boxed{\widehat{\lambda} = 1/\overline{X}.}$$

We saw in 2.2.1 of Chapter 2 that this is a biased estimator for λ. So, we see that method of moments estimators are not necessarily going to be unbiased. So sad.

Super Important Note

In the previous example, we wrote

$$\frac{1}{\lambda} \overset{set}{=} \overline{X} \qquad \text{and} \qquad \lambda = 1/\overline{X}.$$

These are both nonsense statements because the left-hand sides are constants and the right-hand sides are random variables!

They are both intermediate "placeholding" steps towards an end where we do have equality of random variables: $\widehat{\lambda} = 1/\overline{X}$.

Don't ever let anyone interrupt you in the middle of a method of moments estimation problem because you will look silly before you "put a hat on it"!

Of course you can avoid this issue entirely by using "hats" the whole way through, but it can make the algebra seem more cumbersome than it is.

Let's do a two-parameter example.

Example 5.1.2

For a two-parameter problem, you will generally need two sets of moments. (It is possible that you may need even more if you lose information through cancellation when solving the resulting system of equations for the unknown parameters.)

Let $X_1, X_2, \ldots, X_n$ be a random sample from the $\Gamma(\alpha, \beta)$ distribution. Find MMEs of α and β.

The first population moment is

$$\mu_1 = \mathsf{E}[X] = \alpha/\beta.$$

The first sample moment is

$$M_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}.$$

Equate them.

$$\frac{\alpha}{\beta} \overset{set}{=} \overline{X}$$

We need to solve for both α and β. We can't do this with only one equation. So, we will consider another set of moments.

The second population moment is

$$\mu_2 = \mathsf{E}[X^2] = Var[X] + (\mathsf{E}[X])^2 = \frac{\alpha}{\beta^2} + \left(\frac{\alpha}{\beta}\right)^2.$$

The second sample moment is

$$M_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

Equate them.

$$\frac{\alpha}{\beta^2} + \left(\frac{\alpha}{\beta}\right)^2 \overset{set}{=} \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

We now have a system of equations

$$\frac{\alpha}{\beta} = \overline{X} \qquad \text{and} \qquad \frac{\alpha}{\beta^2} + \left(\frac{\alpha}{\beta}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

We now solve for α and β, Let's plug $\alpha/\beta = \overline{X}$ from the first equation into the second equation:

$$\frac{1}{\beta}\overline{X} + \overline{X}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

We can then solve for β to get

$$\beta = \frac{\overline{X}}{\frac{1}{n}\sum X_i^2 - \overline{X}^2}. \qquad (\Leftarrow \text{Nonsense!})$$

From the first equation we have

$$\alpha = \beta\overline{X} = \frac{\overline{X}^2}{\frac{1}{n}\sum X_i^2 - \overline{X}^2}. \qquad (\Leftarrow \text{More nonsense!})$$

So, the method of moments estimators for α and β are

$$\widehat{\alpha} = \frac{\overline{X}^2}{\frac{1}{n}\sum X_i^2 - \overline{X}^2} \qquad \text{and} \qquad \widehat{\beta} = \frac{\overline{X}}{\frac{1}{n}\sum X_i^2 - \overline{X}^2}.$$

5.1.3 Properties of MMEs

Method of moments estimators are highly intuitive estimators that are usually relatively easy to produce. Here are some pros and cons that are generally true about MMEs under "mild conditions".[3]

[3]The population moments are functions of the parameters. The sample moments are functions of certain statistics. These functions must basically be "nice". Depending on what property you want out of an MME, the function might need to be, for example, continuous or differentiable. See examples in this Section.

Pros	Cons
intuitive	possibly biased
tend to be easy to find	not uniquely defined
usually consistent estimators	may be outside of the parameter space
usually asymptotically normal	may not exist

Most MME's are consistent as long as the functions defining them are "well-behaved". Recall that an estimator $\widehat{\theta}_n$ of θ is consistent if $\widehat{\theta}_n \xrightarrow{P} \theta$. While we will not prove this consistency claim for all MMEs, let's take a closer look at the previous two examples.

Example5.1.1 Continued

In Example 5.1.1, we had an MME $\widehat{\lambda} = 1/\overline{X}$ for the rate parameter λ for an exponential distribution. As $\overline{X}$, and hence $\widehat{\lambda}$ depend on n, we will add some notation that emphasizes this by writing the MME as

$$\widehat{\lambda}_n = 1/\overline{X}_n$$

so that "convergence arrows" might make more sense.

We have

$$\widehat{\lambda}_n = g(\overline{X}_n)$$

where $g(x) = 1/x$.

By the WLLN, we have that $\overline{X}_n \xrightarrow{P} \mu = 1/\lambda$. Although g has a discontinuity at 0, $\overline{X}_n$ will, with probability 1, not take on the value 0. So, we can use Theorem 2.3.3 along with the note about discontinuity on 122, to say that

$$\widehat{\lambda}_n = g(\overline{X}_n) \xrightarrow{P} g(1/\lambda) = \lambda.$$

Example5.1.2 Continued

In Example 5.1.2, we had the MME estimator

$$\widehat{\alpha} = \frac{\overline{X}^2}{\frac{1}{n}\sum X_i^2 - \overline{X}^2} = \frac{M_1^2}{M_2 - M_1^2}$$

for α in the $\Gamma(\alpha, \beta)$ distribution.

- From the WLLN, we know that

$$\mathsf{E}[M_1] = \mathsf{E}[\overline{X}] \xrightarrow{P} \mathsf{E}[X_1] = \mu_1 = \alpha/\beta.$$

- Since $\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \overline{Y}$ where $Y_i = X_i^2$, we know, from the WLLN, that

$$\mathsf{E}[M_2] = \mathsf{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2\right] = \mathsf{E}[\overline{Y}] \xrightarrow{P} \mathsf{E}[Y_1] = \mathsf{E}[X_1^2] = \mu_2 = \frac{\alpha}{\beta^2} + \left(\frac{\alpha}{\beta}\right)^2.$$

- From Theorem 2.7.1, we then have

$$(M_1, M_2) \xrightarrow{P} (\mu_1, \mu_2)\,.$$

  This is joint convergence in probability that we discussed in Section 2.7.

- By an analogous vector-valued result to item 4 in Theorem 2.3.3, we then have that

$$\widehat{\alpha} = g(M_1, M_2) \xrightarrow{P} g(\mu_1, \mu_2)$$

  where $g(x, y) = x^2/(y - x^2)$.

  (Although g is not continuous everywhere, the probability that $M_2 - M_1^2 = 0$ is zero so it does not cause problems here.)

  Note that

$$g(\mu_1, \mu_2) = \frac{\mu_1^2}{\mu_2 - \mu_1^2} = \frac{(\alpha/\beta)^2}{\frac{\alpha}{\beta^2} + \left(\frac{\alpha}{\beta}\right)^2 - \left(\frac{\alpha}{\beta}\right)^2} = \alpha.$$

- Thus, we have that

$$\widehat{\alpha} \xrightarrow{P} \alpha$$

  where $\widehat{\alpha}$ is the MME of α.

It is easy to adjust the definition of g to also conclude that

$$\widehat{\beta} \xrightarrow{P} \beta.$$

We have already seen, in Example 5.1.1, that MMEs are not necessarily unbiased. They are often asymptotically unbiased though and even asymptotically normal. Note that the WLLN gives us that $M_k \xrightarrow{P} \mu_k$ as long as $Var[X_i^k]$ is finite. Asymptotic normality can usually be shown representing the estimator as a function of a sample moment and combining the WLLN with the Delta Method (or a multiple-dimensional analogue of the Delta Method) from Section 2.5.4 of Chapter 2.

Method of moments estimators are not really well-defined as any order moments may be used. In our exponential example with one unknown parameter λ, we equated the first moments of the sample and population. We could equate second moments instead, solve for λ, and the result will be a different estimator for λ that is still referred

to as a method of moments estimator.

Although it does not happen that often, a major drawback of method of moments estimators is that they make take values outside of the parameter space! (See Exercise **??**) Also, they are not going to be applicable to distributions where required moments do not exist.

5.2 Maximum Likelihood Estimators (MLEs)

While method of moments is a highly intuitive approach to estimation, maximum likelihood estimators (MLEs) have many desirable properties and are considered the to be the "powerhouse" estimators of statistial inference.

Suppose that we have an unfair coin where $p = P(\text{"Heads"})$ is known to be one of 0.2, 0.3, or 0.8.

Let's suppose that we toss the coin twice and use the results to try to estimate p.

If the result of both coin flips are "Heads", we might be inclined to guess that the coin is biased towards heads and that p, the probability of getting heads is 0.8. Of course, we would be much more convinced if we flipped it 10 times and saw 9 or 10 heads, but we have to work with what we're given!

If the result of both coin flips are "Tails", we would be inclined to guess that p is most likely 0.2.

If we saw one heads and one tails in our two flips, we might guess that p is 0.3.

Idea

A maximum likelihood estimator uses, as an estimate of an unknown parameter, a value in the parameter space that makes the observed data "most likely".

Let's consider our oversimplified but "motivating" coin example with a little more rigor.

Example 5.2.1

The coin model described above can be thought of more formally as a random sample, X_1, X_2, of size 2 from the Bernoulli(p) distribution. That is,

$$X_i = \begin{cases} 1 & , \text{ if "Heads"} \\ 0 & , \text{ if "Tails"} \end{cases}$$

The pdf for each X_i is

$$f(x;p) = p^x(1-p)^{1-x}\, I_{\{0,1\}}(x).$$

The joint pdf for X_1 and X_2 is

$$f(x_1,x_2;p) \overset{indep}{=} f(x_1;p)\cdot f(x_2;p) = p^{x_1+x_2}(1-p)^{2-x_1-x_2}\, I_{\{0,1\}}(x_1) I_{\{0,1\}}(x_2)$$

For these discrete random variables, this is $P(X_1 = x_1, X_2 = x_2)$. For fixed observations x_1 and x_2, we would like to know the value of p from the simplified parameter space $\{0.2, 0.3, 0.8\}$ that maximizes the probability of observing the pair (x_1, x_2).

For this simple example we can tabulate the probabilities. The entries of this table give various values of $f(x_1, x_2; p)$.

		(x_1,x_2)			
		(0,0)	(0,1)	(1,0)	(1,1)
	0.2	0.64	0.16	0.16	0.04
p	0.3	0.49	0.21	0.21	0.09
	0.8	0.04	0.16	0.16	0.64

For each (x_1, x_2), the maximum likelihood estimator of p is the value of p that maximizes the probability of seeing that (x_1, x_2). For example, when $(x_1, x_2) = (0, 0)$ the highest probability in the corresponding column happens when $p = 0.64$. SO, this will be our guess for p when we make that observation. Continuing in this way, we have that the maximum likelihood estimator for p is

$$\widehat{p} = \begin{cases} 0.2 & , \text{ if } (x_1,x_2) = (0,0) \\ 0.3 & , \text{ if } (x_1,x_2) = (0,1) \text{ or } (1,0) \\ 0.8 & , \text{ if } (x_1,x_2) = (1,1). \end{cases}$$

In general, for fixed $(x_1, x_2, \ldots, x_n)$, we are maximizing the joint density $f(\vec{x};\theta) = f(x_1, x_2, \ldots, x_n; \theta)$ with respect to θ.

Note that, under our usual assumption that the random variables in the sample are iid,

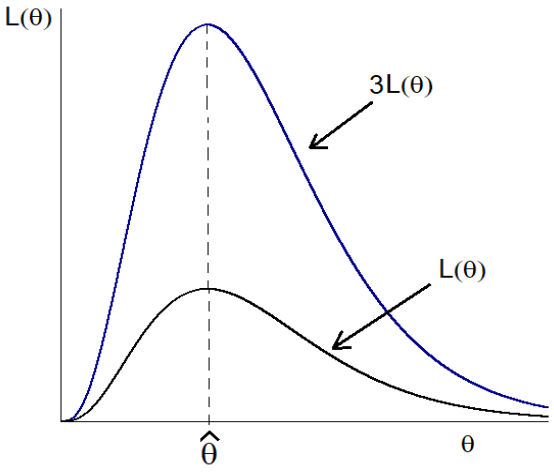
$$f(\vec{x};\theta) = f(x_1, x_2, \ldots, x_n; \theta) \overset{iid}{=} f(x_1;\theta) \cdot f(x_2;\theta) \cdots f(x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta).$$

When thought of as a function of θ (the x's are considered to be fixed constants), $f(\vec{x};\theta)$ is called a *likelihood function* and is denoted by $L(\theta)$.

$$L(\theta) = f(\vec{x};\theta).$$

This is also how we define a likelihood function even if the X_i are not iid and the joint pdf didn't come from a product of marginal pdfs.

We wish to maximize this as a function of θ. We will call the maximizer (if it exists) $\widehat{\theta}$.



Note that the value of θ that maximizes $L(\theta)$ is the same value that maximizes $3L(\theta)$, $\frac{1}{\sqrt{2\pi}}L(\theta)$, or even $(\prod_{i=1}^{n} x_i) \cdot L(\theta)$, because $\prod_{i=1}^{n} x_i$ is just another constant now.

Definition 5.2.1

Suppose that $X_1, X_2, \ldots, X_n$ have joint pdf $f(\vec{x};\theta)$.

(They may or may not be iid random variables.)

A **likelihood function** is denoted by $L(\theta)$ and is defined as any function proportional to $f(\vec{x};\theta)$ when thought of as a function of θ.

The **maximum likelihood estimator** of θ is the value of θ that maximizes $L(\theta)$.

For almost all of the "nice known named distributions" that people study, this maximum exists and is unique.

5.2.1 Examples

Example 5.2.2

Let's revisit the coin toss example with n flips and with p, the probability of getting heads on any one flip to be an unknown number in the parameter space $0 \leq p \leq 1$.

The data is $X_1, X_2, \ldots, X_n \overset{iid}{\sim} Bernoulli(p)$, so, the joint pdf is

$$\begin{aligned} f(\vec{x}; p) &\overset{iid}{=} \prod_{i=1}^{n} f(x_i; p) \\ &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} I_{\{0,1\}}(x_i) \\ &= p^{\sum x_i}(1-p)^{n-\sum x_i} \prod_{i=1}^{n} I_{\{0,1\}}(x_i). \end{aligned}$$

Since the product of indicators does not involve the unknown p, it is a constant of proportionality and hence may be dropped to form a likelihood function

$$L(p) = p^{\sum x_i}(1-p)^{n-\sum x_i}.$$

We must maximize $L(p)$ with respect to p.

In the vast majority of cases, it is more convenient to instead maximize the log of the likelihood function. Note that $y = \ln x$ is an increasing function of x, so, while it will change the shape of the likelihood, it will preserve orderings in the sense that $x_1 < x_2 \Rightarrow \ln x_1 < \ln x_2$. This means that $L(\theta)$ and $\ln L(\theta)$ are maximized at the same value of θ.

We will use $\ell(\theta)$ to denote the log-likelihood function $\ln L(\theta)$.

Continuing, we have

$$\ell(p) = \ln L(p) = \sum x_i \ln p + (n - \sum x_i) \ln(1-p).$$

Taking the derivative with respect to p gives

$$\frac{d}{dp}\ell(p) = \frac{\sum x_i}{p} - \frac{n - \sum x_i}{1-p} \overset{set}{=} 0$$

Multiply through on both sides by $p(1-p)$ to remove the denominators:

$$\sum x_i(1-p) - (n - \sum x_i)p = 0.$$

We then have

$$\sum x_i - p\sum x_i - np + p\sum x_i = 0,$$

which implies that

$$\sum x_i - np = 0 \qquad \Rightarrow p = \sum x_i/n.$$

In the end, we put capital letters back in for the x's to give an estimator[1] for any future sample drawn from the Bernoulli distribution rather than an estimate for fixed x's, and we put a hat on the parameter because you're done! The MLE for p for the Bernoulli distribution is

$$\boxed{\widehat{p} = \frac{\sum_{i=1}^{n} X_i}{n} = \overline{X}.}$$

This makes a lot of sense as an estimator. The parameter p is the probability that any one sampled value is a 1 (as opposed to a 0) and $\overline{X}$, as the sum of a lot of 0's and 1's divided by n is the proportion of 1's in the sample!

[1] See the discussion about the use of the word "estimator" versus "estimate" on page 100

Notation

Given a likelihood function $L(\theta)$, we will use $\ell(\theta)$ to denote the log-likelihood:

$$\ell(\theta) = \ln L(\theta).$$

The maximum likelihood estimator for θ is the value that maximizes $L(\theta)$ and, equivalently, $\ell(\theta)$.

We will now look at a continuous example. We still define the likelihood function to be any function proportional to the joint pdf even though the joint pdf does not represent probability in this case. We also still define the MLE to be the value that maximizes the likelihood function.

Example 5.2.3

Let $X_1, X_2, \ldots, X_n$ be a random sample with pdf $f(x;\theta) = \theta x^{\theta-1} I_{(0,\infty)}(x)$ for $\theta > 0$.

Find the maximum likelihood estimator of θ.

The joint pdf is

$$f(\vec{x};\theta) \overset{iid}{=} \prod_{i=1}^{n} f(x_i;\theta) = \prod_{i=1}^{n} \theta x_i^{\theta-1} I_{(0,1)}(x_i) = \theta^n \left(\prod_{i=1}^{n} x_i\right)^{\theta-1} \prod_{i=1}^{n} I_{(0,1)}(x_i).$$

A likelihood function is

$$L(\theta) = \theta^n \left(\prod_{i=1}^{n} x_i\right)^{\theta-1}.$$

Note that we can factor $(\prod_{i=1}^n x_i)^{\theta-1}$ into $(\prod_{i=1}^n x_i)^{\theta}$ times $(\prod_{i=1}^n x_i)^{-1}$. The later is a multiplicative constant with respect to θ and can be dropped from the likelihood. We will leave it in though and will see it disappear, showing us that it doesn't matter whether or not we include it in the first place!

The log-likelihood is

$$\ell(\theta) = n\ln\theta + (\theta-1)\ln\left(\prod_{i=1}^{n} x_i\right) = n\ln\theta - (\theta-1)\sum_{i=1}^{n}\ln x_i.$$

Note that the last term is

$$\theta\sum_{i=1}^{n}\ln x_i - \sum_{i=1}^{n}\ln x_i,$$

and the last term of this expression is going to be zero when we take the derivative with respect to θ. This term came from the inclusion of $(\prod_{i=1}^n x_i)^{-1}$ in the likelihood and it is going to disappear when we take the derivative, which is

$$\frac{\partial}{\partial\theta}\ell(\theta) = \frac{n}{\theta} + \sum_{i=1}^{n}\ln x_i \overset{set}{=} 0.$$

Solving for θ and "putting a hat on it", we get that the maximum likelihood estimator for θ is

$$\boxed{\widehat{\theta} = \frac{-n}{\sum_{i=1}^{n} X_i}.}$$

Note that $\partial^2/\partial\theta^2\ell(\theta) = -n/\theta^2 < 0$, so we did find a maximum and not a minimum or saddle point. It is unusual, when working with the standard "nice known and named" distributions to have a critical point of a likelihood actually be a minimum. In this text, we will usually omit the check that we did, in fact, find a maximum. Be careful out there in "real life estimation problems" though!

Next up is a two-parameter example.

Example 5.2.4 (Two Parameters)

Let $X_1, X_2, \ldots, X_n$ be a random sample from the $N(\mu, \sigma^2)$ distribution. Find the MLEs for μ and σ^2.

The pdf is

$$f(x; \mu, \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}.$$

The joint pdf is

$$\begin{aligned} f(\vec{x}; \mu, \theta) &\overset{iid}{=} \prod_{i=1}^n f(x_i; \mu, \sigma^2) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} \\ &= (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2}. \end{aligned}$$

A likelihood is

$$L(\mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2}.$$

(Note that we can split $(2\pi\sigma^2)^{-n/2} = (2\pi)^{-n/2}(\sigma^2)^{-n/2}$ and drop the 2π part, though we're going to leave it. Also note that we are thinking of σ^2 as a symbol and not as a "symbol squared". It is just a parameter with the name "sigma squared"!)

The log-likelihood is

$$\ell(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2$$

Since there are two variables, we take two partial derivatives and set them both equal to zero.

$$\begin{aligned} \frac{\partial}{\partial\mu}\ell(\mu, \sigma^2) &= \frac{1}{\sigma^2}\sum_{i=1}^n (x_i - \mu) \overset{set}{=} 0 \\ \frac{\partial}{\partial\sigma^2}\ell(\mu, \sigma^2) &= -\frac{n}{2}\frac{2\pi}{2\pi\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^n (x_i - \mu)^2 \overset{set}{=} 0 \end{aligned}$$

The first equation implies that

$$\sum_{i=1}^n (x_i - \mu) = 0 \quad \Rightarrow \quad \sum_{i=1}^n x_i - n\mu = 0 \quad \Rightarrow \quad \boxed{\widehat{\mu} = \overline{X}}.$$

The second equation, after cancelling the 2π and multiplying both sides by 2, becomes

$$-\frac{n}{\sigma^2} + \frac{1}{(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \mu)^2 = 0.$$

Multiplying both sides by $(\sigma^2)^2$ gives

$$-n\sigma^2 + \sum_{i=1}^{n}(x_i - \mu)^2 = 0.$$

Plugging in $\widehat{\mu} = \overline{X}$ for μ and solving for σ^2 gives the MLE

$$\boxed{\widehat{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})}{n}}.$$

Remember, no matter how you did the algebra, with upper or lower case letters, in the end, make sure everything is upper case when giving your estimator. The estimator should be a random variable unless you have actually observed numerical data!

In the next example, we will find an MLE in the case that the support of the distribution depend on the unknown parameters. (i.e. A parameter or parameters are in the indicator function.) For example, if $X_1, X_2, \ldots, X_n$ is a random sample from the uniform distribution on the interval $(0, \theta)$, we have that the pdf is $1/\theta$ for x between 0 and θ and it is 0 when x is not between 0 and θ. First though, we want to note a few things about derivatives and logs in this case.

Note

Consider the piecewise defined function

$$g(x) = \begin{cases} 2x^2 & , \quad 0 \leq x < 2 \\ 5x & , \quad x \geq 2 \end{cases}$$

The derivative is

$$g'(x) = \begin{cases} 4x & , \quad 0 < x < 2 \\ 5 & , \quad x > 2 \end{cases}$$

The log of the function is

$$\ln g(x) = \begin{cases} \ln(2x^2) & , \quad 0 \leq x < 2 \\ \ln(5x) & , \quad x \geq 2. \end{cases}$$

Did you notice how the derivative and log function did not affect the inequalities defining the two regions? In indicator notation,

$$g(x) = 2x^2\, I_{[0,2)}(x) + 5x\, I_{[2,\infty)}(x).$$

We have,

$$g'(x) = 4x\, I_{(0,2}(x) + 5\, I_{(2,\infty)}(x)$$

and

$$\ln g(x) = \ln(2x^2)\, I_{[0,2)}(x) + \ln(5x)\, I_{[2,\infty)}.$$

It is important to note that we do not take derivatives or logs of indicator functions and that, aside from a little care at the endpoints of their intervals, the indicators just "come along for the ride".

Example 5.2.5 (Parameters in the Support of the Distribution)

Let $X_1, X_2, \ldots, X_n$ be a random sample from the distribution with pdf

$$f(x; \theta, \lambda) = \theta \lambda^\theta x^{-(\theta+1)}\, I_{(\lambda,\infty)}(x)$$

with $\theta > 0$, $\lambda > 0$. This pdf is zero if x is below λ.

Find MLEs for θ and λ.

The joint pdf is

$$\begin{aligned} f(\vec{x};\theta,\lambda) &\overset{iid}{=} \prod_{i=1}^{n} f(x_i;\theta,\lambda) \\ &= \prod_{i=1}^{n} \theta\lambda^{\theta} x_i^{-(\theta+1)}\, I_{(\lambda,\infty)}(x_i) \\ &= \theta^n \lambda^{n\theta} \left(\textstyle\prod_{i=1}^{n} x_i\right)^{-(\theta+1)} \displaystyle\prod_{i=1}^{n} I_{(\lambda,\infty)}(x_i) \\ &= \theta^n \lambda^{n\theta} \left(\textstyle\prod_{i=1}^{n} x_i\right)^{-(\theta+1)}\, I_{(\lambda,\infty)}(x_{(1)}). \end{aligned}$$

We have simplified a product of indicators into a single indicator. Note that $\prod_{i=1}^{n} I_{(\lambda,\infty)}(x_i)$ is 1 if and only if every x_i in the sample is greater than λ. This happens if and only if the minimum in the sample is greater than λ and so the indicator expressions are equivalent.

As per our discussion in Example 5.2.3, we may break up

$$\left(\prod_{i=1}^{n} x_i\right)^{-(\theta+1)} \qquad \text{into} \qquad \left[\left(\prod_{i=1}^{n} x_i\right)^{-\theta}\right] \cdot \left[\left(\prod_{i=1}^{n} x_i\right)^{-1}\right]$$

and drop the second term as a constant when forming the likelihood. We will leave it in. Other than this, there are no other constants of proportionality when thinking of the joint pdf as a function of the parameters.

A likelihood function is

$$L(\theta,\lambda) = \theta^n \lambda^{n\theta} \left(\prod_{i=1}^{n} x_i\right)^{-(\theta+1)} I_{(\lambda,\infty)}(x_{(1)}).$$

Note that we can not drop the indicator when forming the likelihood because it is a function of one of the unknown parameters and not purely a function of the x_i. However, per the note preceding this example, we will not be taking the log and derivative of the indicator part when find the MLEs here. While we may bring the indicator through the calculations as in the note, we choose to leave off the indicator entirely but keep it "in mind" as it is still relevant to the problem.

$$L(\theta,\lambda) = \theta^n \lambda^{n\theta} \left(\prod_{i=1}^{n} x_i\right)^{-(\theta+1)}. \tag{5.2.1}$$

The log-likelihood is

$$\ell(\theta,\lambda) = n\ln\theta + n\theta\ln\lambda - (\theta+1)\sum_{i=1}^{n} \ln x_i.$$

The derivatives are

$$\begin{aligned} \frac{\partial}{\partial\theta}\ell(\theta,\lambda) &= \frac{n}{\theta} + n\ln\lambda - \sum_{i=1}^{n}\ln x_i \overset{set}{=} 0 \\ \\ \frac{\partial}{\partial\lambda}\ell(\theta,\lambda) &= \frac{n\theta}{\lambda} \overset{set}{=} 0. \end{aligned}$$

Note that the second equation does not give us any information. The sample size can't be zero and the parameter θ is greater than 0. This (one equation giving no information) usually happens when the parameter is in the indicator, as it was here. The goal, however, is still the same– to maximize the likelihood. So, let's look at the likelihood.

From (5.2.1), we see that the likelihood is an increasing function of λ. Thus, we want λ to be as large as possible. From our indicator, we see that the minimum value in the data set must be above λ. If you did not simplify the indicator, the product of indicators says that, equivalently, all values in the data set must be above λ. So, looking at the data as fixed, we see that the largest λ can possibly be is the minimum value of the data set. That is, the MLE for λ is

$$\boxed{\widehat{\lambda} = X_{(1)}.}$$

Plugging this in to the first of the two partial derivative equations, we solve for θ to get

$$\boxed{\widehat{\theta} = \frac{n}{\sum_{i=1}^{n}\ln X_i - n\ln X_{(1)}}.}$$

5.2.2 The Invariance Property of MLEs

Let $X_1, X_2, \ldots, X_n$ be a random sample from the exponential distribution with rate λ.

Let's find the MLE for $\tau(\lambda) = \lambda^2$.

The pdf is

$$f(x;\lambda) = \lambda e^{-\lambda x}\, I_{(0,\infty)}(x).$$

We can write it in terms of $\tau = \lambda^2$:

$$f(x;\lambda) = \sqrt{\tau} e^{-\sqrt{\tau}x}\, I_{(0,\infty)}(x).$$

Let's call this new function $\widetilde{f}(x;\tau)$. Note that

$$f(x;\lambda) = \widetilde{f}(x;\lambda^2).$$

As a function of τ, the joint pdf is now

$$\widetilde{f}(\vec{x};\tau) \stackrel{iid}{=} \prod_{i=1}^{n} \widetilde{f}(x_i;\tau) = \prod_{i=1}^{n} \sqrt{\tau} e^{-\sqrt{\tau}x_i} I_{(0,\infty)}(x_i) = \tau^{n/2} e^{-\sqrt{\tau}\sum x_i} \prod_{i=1}^{n} I_{(0,\infty)}(x_i).$$

So, a likelihood function is

$$\widetilde{L}(\tau) = \tau^{n/2} e^{-\sqrt{\tau}\sum x_i}.$$

The log-likelihood is

$$\widetilde{\ell}(\tau) = \frac{n}{2}\ln\tau - \sqrt{\tau}\sum x_i.$$

The derivative gives

$$\frac{d}{d\tau}\widetilde{\ell}(\tau) = \frac{n}{2}\frac{1}{\tau} - \frac{1}{2\sqrt{\tau}}\sum x_i \stackrel{set}{=} 0.$$

Multiplying through by 2τ, we get

$$n - \sqrt{\tau}\sum x_i = 0,$$

which implies that,

$$\tau^{1/2} = n \Big/ \sum x_i \ ,$$

or equivalently that

$$\tau = \left(n \Big/ \sum x_i\right)^2.$$

So the MLE of $\tau = \tau(\lambda) = \lambda^2$ is

$$\widehat{\tau(\lambda)} = \left(1/\overline{X}\right)^2.$$

Recall that the MLE of λ was

$$\widehat{\lambda} = 1/\overline{X}.$$

Since $\tau(\lambda) = \lambda^2$, we see that

$$\widehat{\tau(\lambda)} = \tau(\hat{\lambda}).$$

This was no accident! This will always happen for MLEs and is known as the *invariance property of MLEs*.

Definition 5.2.2

If $\hat{\theta}$ is the MLE of θ, then the MLE of a function $\tau(\theta)$ is

$$\widehat{\tau(\theta)} = \tau(\hat{\theta}).$$

This is known as the **invariance property** of MLEs.

In words, the MLE of a function of a parameter is the function with the MLE of the parameter plugged in. This is something that we are definitely going to want to take advantage of!

We will now see why MLEs are, in general, invariant.

Let $L(\theta)$ denote the likelihood function and let $\widetilde{L}$ denote the re-parameterized likelihood function. In other words, the relationship is

$$L(\theta) = \widetilde{L}(\tau(\theta)).$$

Then by the chain rule we get the second equality below:

$$\frac{\partial}{\partial\theta}L(\theta) = \frac{\partial}{\partial\theta}\widetilde{L}(\tau(\theta)) = \frac{\partial}{\partial\tau}\widetilde{L}(\tau(\theta)) \cdot \frac{\partial\tau}{\partial\theta}.$$

We can now see that if $\frac{\partial\tau}{\partial\theta} \neq 0$, then $\frac{\partial}{\partial\tau}\widetilde{L}(\tau(\theta)) = 0$ whenever $\frac{\partial}{\partial\theta}L(\theta) = 0$.

Well, $\frac{\partial}{\partial\theta}L(\theta) = 0$ at $\theta = \hat{\theta}$, because that's the definition of the MLE $\hat{\theta}$. ie:

$$\frac{\partial}{\partial\theta}L(\theta)\bigg|_{\theta=\hat{\theta}} = 0$$

Hence

$$\frac{\partial}{\partial\tau}\widetilde{L}(\tau(\theta))\bigg|_{\theta=\hat{\theta}} = \frac{\partial}{\partial\theta}L(\theta)\bigg|_{\theta=\hat{\theta}} = 0$$

Therefore $\tau(\hat{\theta})$ is a solution to $\frac{\partial}{\partial\tau}\widehat{L}(\tau(\theta)) = 0$ which is the equation that comes up in finding an MLE in the re-parameterized case. Well $\widehat{\tau(\theta)}$ is, by definition, our solution to $\frac{\partial}{\partial\tau}\widetilde{L}(\tau(\theta)) = 0$. So

$$\widehat{\tau(\theta)} = \tau(\hat{\theta}),$$

as claimed.

We are not finished talking about properties of MLEs. In order to talk about some of their asymptotic properties, we will need some machinery that is substantial enough to get it's own section.

5.3 The Cramér-Rao Lower Bound (CRLB)

The Cramér-Rao Lower Bound (CRLB) associated with a parameter θ (or for a function $\tau(\theta)$) is a lower bound on the variance of all unbiased estimators of θ (or $\tau(\theta)$).[4] Recall that, when we have two unbiased estimators for θ, say $\widehat{\theta}_1$ and $\widehat{\theta}_2$, we say (all else being equal) that $\widehat{\theta}_1$ is the better estimator if $Var[\widehat{\theta}_1] \leq Var[\widehat{\theta}_2]$. These variances are functions of θ so really we can only say that $\widehat{\theta}_1$ is the better estimator if $Var[\widehat{\theta}_1] \leq Var[\widehat{\theta}_2]$ for all θ. We might be able to find a third unbiased estimator with even lower variance. The CRLB tells us how low we could possibly go with these variances. It is not always "achievable" as a sharp lower bound. That is, the unbiased estimator with the lowest possible variance might still have a variance that is greater than the CRLB. However, if we do happen to find an unbiased estimator for θ (or $\tau(\theta)$) whose variance does achieve the CRLB, we know that we can not do any better in terms of low variance.

[4]Note the use of the phrase "associated with". People often say that the Cramér-Rao Lower Bound is "for θ" (or "for" $\tau(\theta)$). This is acceptable as long as it is not taken literally. It is not a bound on θ (or $\tau(\theta)$)!

5.3.1 Statement of the CRLB and Derivation

The Cramér-Rao Lower Bound

Let $X_1, X_2, \ldots, X_n$ be random variables with joint pdf distribution with pdf $f(\vec{x}; \theta)$.

Assume that the support of the distribution does not depend on θ.

Consider estimating some function $\tau(\theta)$.

Let $T = t(\vec{X})$ be any unbiased estimator of $\tau(\theta)$.

Then

$$Var[T] \geq \frac{[\tau'(\theta)]^2}{\mathsf{E}\left[\left(\frac{\partial}{\partial\theta} \ln f(\vec{X}; \theta)\right)^2\right]} \tag{5.3.2}$$

under certain *regularity conditions*.

Rather than listing the unmotivated regularity conditions here, we will instead figure out what they need to be as we go through the proof of the CRLB.

The expression on the right of 5.3.2 is known as the Cramér-Rao Lower Bound (CRLB) for the variance of all unbiased estimators of $\tau(\theta)$. We will denote this as $CRLB_{\tau(\theta)}$. In most of this text, the X_i are iid (a random sample) and the joint pdf has the product form

$$f(\vec{x}; \theta) \overset{iid}{=} \prod_{i=1}^{n} f(x_i; \theta),$$

but more general joint pdfs are allowed here.

Note that the X_i only appear in the denominator of the CRLB. We can think of this denominator as containing all of the "information" from the sample. In fact, the denominator is known as the *Fisher Information* and we will denote it by $I_n(\theta)$.

Definition 5.3.1

We define and denote the Fisher information for a sample as

$$I_n(\theta) := \mathsf{E}\left[\left(\frac{\partial}{\partial\theta} \ln f(\vec{X}; \theta)\right)^2\right].$$

With this notation, we can now write the CRLB for the variance of all unbiased estimators of $\tau(\theta)$ as:

$$CRLB_{\tau(\theta)} = \frac{[\tau'(\theta)]^2}{I_n(\theta)}.$$

In order to prove (5.3.2) we need to use the *Cauchy-Schwarz inequality*.

The Cauchy-Schwarz Inequality

For square-integrable functions f and g,

$$\left(\int f(x)g(x)\,dx\right)^2 \leq \left(\int f^2(x)\,dx\right)\cdot\left(\int g^2(x)\,dx\right)$$

There is an analogous Cauchy-Schwarz inequality for sums, or, more generally, inner products.

Proof :

First note that

$$\int\int [f(x)g(y) - f(y)g(x)]^2\,dy\,dx \geq 0.$$

Multiplying out the left-hand-side then gives us

$$\int\int f^2(x)g^2(y)\,dy\,dx - 2\int\int f(x)f(y)g(x)g(y)\,dy\,dx + \int\int f^2(y)g^2(x)\,dy\,dx \geq 0 \qquad (5.3.3)$$

Note that the first and third terms are exactly the same. As for the middle integral, pulling the "x stuff" out of the inner y-integral gives us

$$\begin{aligned}\int\int f(x)f(y)g(x)g(y)\,dy\,dx &= \int f(x)g(x)\int f(y)g(y)\,dy\,dx\\ &= \int f(x)g(x)\left(\int f(y)g(y)\,dy\right)\,dx\\ &= \left(\int f(y)g(y)\,dy\right)\int f(x)g(x)\,dx\\ &= \left(\int f(x)g(x)\,dx\right)^2\end{aligned}$$

So, (5.3.3) becomes

$$2\int\int f^2(x)g^2(y)\,dy\,dx - 2\left(\int f(x)g(x)\,dx\right)^2 \geq 0$$

$$\Downarrow$$

$$\int\int f^2(x)g^2(y)\,dy\,dx - \left(\int f(x)g(x)\,dx\right)^2 \geq 0$$

$$\Downarrow$$

$$\left(\int f(x)g(x)\,dx\right)^2 \leq \left(\int f^2(x)\,dx\right)\cdot\left(\int g^2(x)\,dx\right).$$

■

Note

As an expectation is a sum or an integral, one can adapt (See Exercise 18.) the above proof of the Cauchy-Schwarz inequality to show that, for functions g_1 and g_2 with $\mathsf{E}[g_1^2(X)] < \infty$ and $\mathsf{E}[g_2^2(X)] < \infty$,

$$(\mathsf{E}[g_1(X)\,g_2(X)])^2 \leq \mathsf{E}[g_1^2(X)]\cdot\mathsf{E}[g_2^2(X)].$$

This will also hold for vector-valued random variables.

We are now ready to prove (5.3.2).

Proof (CRLB):

It is sufficient to show that

$$\tau'(\theta) = \mathsf{E}\left[(T-\tau(\theta))\left(\frac{\partial}{\partial\theta}\ln f(\vec{X};\theta)\right)\right] \tag{5.3.4}$$

where $T = t(\vec{X})$. If we can show this, we will have, by the Cauchy-Schwarz inequality, that

$$\begin{aligned}
[\tau'(\theta)]^2 &= \left(\mathsf{E}\left[(T-\tau(\theta))\left(\tfrac{\partial}{\partial\theta}\ln f(\vec{X};\theta)\right)\right]\right)^2 \\
&\overset{C-S}{\leq} \mathsf{E}[(T-\tau(\theta))^2]\cdot\mathsf{E}\left[\left(\tfrac{\partial}{\partial\theta}\ln f(\vec{X};\theta)\right)^2\right] \\
&= Var[T]\cdot\mathsf{E}\left[\left(\tfrac{\partial}{\partial\theta}\ln f(\vec{X};\theta)\right)^2\right].
\end{aligned}$$

We then have the CRLB, provided that

$$0 < \mathsf{E}\left[\left(\frac{\partial}{\partial\theta}\ln f(\vec{X};\theta)\right)^2\right] < \infty$$

so that we can divide it to the other side. ■

Proof of (5.3.4):

Since $\mathsf{E}[T] = \tau(\theta)$

$$\tau'(\theta) \quad = \quad \tfrac{\partial}{\partial\theta}\tau(\theta) = \tfrac{\partial}{\partial\theta}\mathsf{E}[T] = \tfrac{\partial}{\partial\theta}\int t(\vec{x})f(\vec{x};\theta)\,d\vec{x}$$

Since we are trying to see something like $T - \tau(\theta)$, we will subtract off zero in a clever way:

$$\frac{\partial}{\partial\theta}\int t(\vec{x})f(\vec{x};\theta)\,d\vec{x} - \tau(\theta)\frac{\partial}{\partial\theta}1$$

$$= \quad \tfrac{\partial}{\partial\theta}\int t(\vec{x})f(\vec{x};\theta)\,d\vec{x} - \tau(\theta)\tfrac{\partial}{\partial\theta}\int f(\vec{x};\theta)\,d\vec{x}$$

$$= \quad \int t(\vec{x})\tfrac{\partial}{\partial\theta}f(\vec{x};\theta)\,d\vec{x} - \int \tau(\theta)\tfrac{\partial}{\partial\theta}f(\vec{x};\theta)\,d\vec{x}$$

if we assume that we are able to pull that derivative into the integral.

Continuing, we now have

$$\tau'(\theta) = \int (t(\vec{x}) - \tau(\theta))\ \frac{\partial}{\partial\theta}f(\vec{x};\theta)\,d\vec{x}$$

We want to see an expectation, which means that we want the integrand to end in $f(\vec{x};\theta)d\vec{x}$ as opposed to $\frac{\partial}{\partial\theta}f(\vec{x};\theta)\,d\vec{x}$. Note that

$$\frac{\partial}{\partial\theta}\ln f(\vec{x};\theta) = \frac{\frac{\partial}{\partial\theta}f(\vec{x};\theta)}{f(\vec{x};\theta)},$$

so

$$\frac{\partial}{\partial\theta}f(\vec{x};\theta) = \left(\frac{\partial}{\partial\theta}\ln f(\vec{x};\theta)\right)\cdot f(\vec{x};\theta).$$

We now have that

$$\tau'(\theta) = \int (t(\vec{x}) - \tau(\theta))\ \left(\frac{\partial}{\partial\theta}\ln f(\vec{x};\theta)\right)\ f(\vec{x};\theta)\,d\vec{x}$$

and this integral is the definition of

$$\mathsf{E}[(T - \tau(\theta)]\left(\frac{\partial}{\partial\theta}\ln f(\vec{X}\,;\theta)\right)].$$

■

The CRLB is said to hold under certain "regularity conditions". These are the things that we assume held in order to complete the above proof. They are

- $0 < \mathsf{E}\left[\left(\frac{\partial}{\partial\theta}\ln f(\vec{X};\theta)\right)^2\right] < \infty,$

- $\frac{\partial}{\partial\theta}\int f(\vec{x};\theta)\,d\vec{x} = \int \frac{\partial}{\partial\theta} f(\vec{x};\theta)\,d\vec{x}$, and
- $\frac{\partial}{\partial\theta}\ln f(\vec{x};\theta)$ exists.

All integrals are evaluated over the domain of $f(\vec{x};\theta)$. **Note that the second of these conditions is violated whenever the parameter θ is involved in the limits of integration.** For example, for the uniform distribution on the interval $(0,\theta)$ the integrals would go from 0 to θ and therefore we would not be able to pass the derivative with respect to θ to the inside of the integral!

The Cramér-Rao does not apply for distributions where the parameter is "in the indicator".

Example 5.3.1

Let $X_1, X_2, \ldots, X_n$ be a random sample from the Bernoulli distribution with parameter p.

a) Find the CRLB for the variance of all unbiased estimators of p. ($\tau(p) = p$)

a) Find the CRLB for the variance of all unbiased estimators of $\tau(p) = p(1-p)$.

The pdf is

$$f(x;p) = p^x(1-p)^{1-x}\, I_{\{0,1\}}(x).$$

The joint pdf is

$$f(\vec{x};p) = \prod_{i=1}^{n} f(x_i;p) = p^{\sum x_i}(1-p)^{n-\sum x_i} \prod_{i=1}^{n} I_{\{0,1\}}(x_i).$$

As in the note preceding Example 5.2.5, taking indicators through logs and derivatives is tedious and kind of defeats the entire point of their use for convenience. We will leave them off in the following calculations. In statistical inference, they should definitely not be ignored when they have information about the parameters of the problem. If this is the case, however, we would not be computing a CRLB to begin with.

To compute the Fisher information, we need the derivative of the log of the joint pdf.

$$\ln f(\vec{x};p) = \sum x_i \ln p + (n - \sum x_i)\ln(1-p)$$

$$\Downarrow$$

$$\frac{\partial}{\partial p}\ln f(\vec{x};p) = \frac{\sum x_i}{p} - \frac{n-\sum x_i}{1-p} = \frac{\sum x_i - np}{p(1-p)}$$

So,

$$\begin{aligned} I_n(p) &= \mathsf{E}\left[\left(\tfrac{\partial}{\partial p}\ln f(\vec{X};p)\right)^2\right] \\ &= \mathsf{E}\left[\left(\tfrac{\sum X_i - np}{p(1-p)}\right)^2\right] \\ &= \tfrac{1}{p^2(1-p)^2}\mathsf{E}\left[(\sum X_i - np)^2\right] \\ &= \tfrac{1}{p^2(1-p)^2}\mathsf{E}[(Y-np)^2] \end{aligned}$$

where $Y = \sum_{i=1}^n X_i$.

Since Y, as the sum of n Bernoulli(p) random variables, has the $bin(n,p)$ distribution and since the mean of the binomial is np, this expectation is simply $Var[Y]$, which is $np(1-p)$. So, the n-dimensional Fisher information is

$$I_n(p) = \frac{1}{p^2(1-p)^2}Var[Y] = \frac{1}{p^2(1-p)^2}\cdot np(1-p) = \frac{n}{p(1-p)}.$$

Finally, the CRLB for the variance of all unbiased estimators of p ($\tau(p) = p$) is

$$CRLB_p = \frac{[\tau'(p)]^2}{I_n(p)} = \frac{[1]^2}{\frac{n}{p(1-p)}} = \frac{p(1-p)}{n}.$$

For part (b), $\tau(p) = p(1-p)$ and

$$CRLB_{p(1-p)} = \frac{[\tau'(p)]^2}{I_n(p)} = \frac{[1-2p]^2}{\frac{n}{p(1-p)}} = \frac{(1-2p)^2 p(1-p)}{n}.$$

As mentioned, the CRLB is being introduced in order for us to study some theoretical properties of MLEs. Before we get back to that discussion, we will look at some useful computational simplifications.

5.3.2 Computational Simplifications for the CRLB

In this section, we will discuss some computational simplifications for the CRLB. We number them starting from 0 since the first one is only there to prove the others but not something you would actually use in when computing a CRLB.

Definition 5.3.2

In statistical inference, the **score function**[1] is the derivative of the log-likelihood

$$S(\vec{X}, \theta) = \frac{\partial}{\partial \theta} \ln f(\vec{X}; \theta).$$

[1] The name "score function" harkens back to Sir Ronald Fisher's first use of such a function in a study of family genetics. Each family in his study was given a "score" for their likelihood of carrying a certain genetic trait.

0. The expected value of the score function is zero. That is, $\mathsf{E}\left[\frac{\partial}{\partial \theta} \ln f(\vec{X}; \theta)\right] = 0$.

Proof:

$$\begin{aligned} \mathsf{E}\left[\tfrac{\partial}{\partial \theta} \ln f(\vec{X}; \theta)\right] &= \int \tfrac{\partial}{\partial \theta} \ln f(\vec{x}; \theta) \cdot f(\vec{x}; \theta)\, d\vec{x} = \int \tfrac{\partial}{\partial \theta} f(\vec{x}; \theta)\, d\vec{x} \\ &= \tfrac{\partial}{\partial \theta} \int f(\vec{x}; \theta)\, d\vec{x} = \tfrac{\partial}{\partial \theta} 1 = 0 \ \surd \end{aligned}$$

1. The Fisher information can be written in terms of a second derivative. In particular, we have that

$$\mathsf{E}\left[\left(\frac{\partial}{\partial \theta} \ln f(\vec{X}; \theta)\right)^2\right] = -\mathsf{E}\left[\frac{\partial^2}{\partial \theta^2} \ln f(\vec{X}; \theta)\right].$$

Proof: From property 0, we have that

$$\mathsf{E}\left[\frac{\partial}{\partial \theta} \ln f(\vec{X}; \theta)\right] = 0$$

which is

$$\int \frac{\partial}{\partial \theta} \ln f(\vec{x}; \theta) \cdot f(\vec{x}; \theta)\, d\vec{x} = 0.$$

Taking the derivative of both sides with respect to θ,

$$\frac{\partial}{\partial \theta} \int \frac{\partial}{\partial \theta} \ln f(\vec{x}; \theta) \cdot f(\vec{x}; \theta)\, d\vec{x} = \frac{\partial}{\partial \theta} 0.$$

The right-hand side is still zero. On the left-hand side, we bring the derivative inside the integral and use the product rule to get

$$\int \left[\frac{\partial}{\partial \theta} \ln f(\vec{x}; \theta) \cdot \frac{\partial}{\partial \theta} f(\vec{x}; \theta) + \frac{\partial^2}{\partial \theta^2} \ln f(\vec{x}; \theta) \cdot f(\vec{x}; \theta)\right] d\vec{x} = 0.$$

Running the integral through we get

$$\int \frac{\partial}{\partial\theta} \ln f(\vec{x};\theta) \cdot \underbrace{\frac{\partial}{\partial\theta} f(\vec{x};\theta)}_{\frac{\partial}{\partial\theta}\ln f(\vec{x};\theta)\cdot f(\vec{x};\theta)} d\vec{x} + \int \frac{\partial^2}{\partial\theta^2} \ln f(\vec{x};\theta) \cdot f(\vec{x};\theta)\, d\vec{x} = 0,$$

which implies that

$$\int \left(\frac{\partial}{\partial\theta} \ln f(\vec{x};\theta)\right)^2 \cdot f(\vec{x};\theta)\, d\vec{x} + \int \frac{\partial^2}{\partial\theta^2} \ln f(\vec{x};\theta) \cdot f(\vec{x};\theta)\, d\vec{x} = 0.$$

This becomes

$$\mathsf{E}\left[\left(\frac{\partial}{\partial\theta} \ln f(\vec{X};\theta)\right)^2\right] + \mathsf{E}\left[\frac{\partial}{\partial\theta^2} \ln f(\vec{X};\theta)\right] = 0,$$

or, equivalently,

$$\mathsf{E}\left[\left(\frac{\partial}{\partial\theta} \ln f(\vec{X};\theta)\right)^2\right] = -\mathsf{E}\left[\frac{\partial}{\partial\theta^2} \ln f(\vec{X};\theta)\right] \quad \surd$$

2. If $X_1, X_2, \ldots, X_n$ are independent and identically distributed, the "n-dimensional Fisher information" can be written as n times the "one-dimensional Fisher information". That is,

$$\mathsf{E}\left[\left(\frac{\partial}{\partial\theta} \ln f(\vec{X};\theta)\right)^2\right] \overset{iid}{=} n \cdot \mathsf{E}\left[\left(\frac{\partial}{\partial\theta} \ln f(X_1;\theta)\right)^2\right],$$

which can be written as

$$I_n(\theta) \overset{iid}{=} nI_1(\theta).$$

Proof:

$$\begin{aligned}
\mathsf{E}\left[\left(\tfrac{\partial}{\partial\theta}\ln f(\vec{X};\theta)\right)^2\right] &= \mathsf{E}\left[\left(\tfrac{\partial}{\partial\theta}\ln\textstyle\prod_{i=1}^n f(X_i;\theta)\right)^2\right] \\
&= \mathsf{E}\left[\left(\tfrac{\partial}{\partial\theta}\textstyle\sum_{i=1}^n \ln f(X_i;\theta)\right)^2\right] \\
&= \mathsf{E}\left[\left(\textstyle\sum_{i=1}^n \tfrac{\partial}{\partial\theta}\ln f(X_i;\theta)\right)^2\right] \\
&= \mathsf{E}\left[\left(\textstyle\sum_{i=1}^n \tfrac{\partial}{\partial\theta}\ln f(X_i;\theta)\right)\left(\textstyle\sum_{j=1}^n \tfrac{\partial}{\partial\theta}\ln f(X_j;\theta)\right)\right] \\
&= \mathsf{E}\left[\textstyle\sum_{i=1}^n\sum_{j=1}^n \tfrac{\partial}{\partial\theta}\ln f(X_i;\theta)\cdot\tfrac{\partial}{\partial\theta}\ln f(X_j;\theta)\right] \\
&= \textstyle\sum_{i=1}^n\sum_{j=1}^n \mathsf{E}\left[\tfrac{\partial}{\partial\theta}\ln f(X_i;\theta)\cdot\tfrac{\partial}{\partial\theta}\ln f(X_j;\theta)\right]
\end{aligned}$$

Now when $j \neq i$, X_i and X_j imply that $\frac{\partial}{\partial\theta}\ln f(X_i;\theta)$ is independent of $\frac{\partial}{\partial\theta}\ln f(X_j;\theta)$, and, in this case

$$\mathsf{E}\left[\frac{\partial}{\partial\theta}\ln f(X_i;\theta)\frac{\partial}{\partial\theta}\cdot\ln f(X_j;\theta)\right] = \mathsf{E}\left[\frac{\partial}{\partial\theta}\ln f(X_i;\theta)\right]\cdot\mathsf{E}\left[\frac{\partial}{\partial\theta}\ln f(X_j;\theta)\right] \overset{\boxed{0}}{=} 0\cdot 0 = 0.$$

Thus,

$$\begin{aligned}
\mathsf{E}\left[\left(\tfrac{\partial}{\partial\theta}\ln f(\vec{X};\theta)\right)^2\right] &= \textstyle\sum_{i=1}^n \mathsf{E}\left[\left(\tfrac{\partial}{\partial\theta}f(X_i;\theta)\right)^2\right] \\
&\overset{ident.}{=} n\cdot\mathsf{E}\left[\left(\tfrac{\partial}{\partial\theta}\ln f(X_1;\theta)\right)^2\right]. \quad \surd
\end{aligned}$$

Example 5.3.2

Let $X_1, X_2, \ldots, X_n$ be a random sample from the $\Gamma(2,\beta)$ distribution.

Find the CRLB for the variance of all unbiased estimators of β.

Because of computational simplification $\boxed{2}$, we will go for the one-dimensional Fisher information.

The pdf is

$$f(x;\beta) = \beta^2 x e^{-\beta x} \quad \text{for } x > 0.$$

Taking the log, we have

$$\ln f(x;\beta) = 2\ln\beta + \ln x - \beta x.$$

The derivative with respect to the parameter β is

$$\frac{\partial}{\partial\beta}\ln f(x;\beta) = \frac{2}{\beta} - x.$$

Thus, we have that the one-dimensional Fisher information is

$$I_1(\beta) \quad = \quad \mathsf{E}\left[\left(\tfrac{\partial}{\partial\beta}\ln f(X_1;\beta)\right)^2\right] = \mathsf{E}\left[\left(X_1 - \tfrac{2}{\beta}\right)^2\right] = Var[X] = \tfrac{2}{\beta^2}.$$

The one-dimensional Fisher information is almost always a variance or can be rewritten to look like one. If we did not recognize the variance here, we could have persevered with brute force by squaring $X_1 - 2/\beta$ and running the expectation through. However, taking the second derivative approach from computational simplification 1, we can, in this example, avoid the expectation entirely.

Since

$$\frac{\partial^2}{\partial\beta^2}\ln f(x;\beta) = \frac{\partial}{\partial\beta}\left(\frac{2}{\beta} - x\right) = -\frac{2}{\beta^2},$$

is a constant (meaning non-random), we have that

$$I_1(\beta) \overset{1}{=} -\mathsf{E}\left[\frac{\partial^2}{\partial\beta^2}\ln f(X_1;\beta)\right] - \mathsf{E}\left[-\frac{2}{\beta^2}\right] = -\left(-\frac{2}{\beta^2}\right) = \frac{2}{\beta^2}.$$

Either way we have

$$I_n(\beta) \overset{iid}{=} n \cdot I_1(\beta) = \frac{2n}{\beta^2}$$

and therefore that the CRLB for $\tau(\beta) = \beta$ is

$$CRLB_\beta = \frac{[1]^2}{I_n(\beta)} = \frac{\beta^2}{2n}.$$

5.4 Asymptotic Properties of MLEs

A decent estimator should probably get even better with more data. The maximum likelihood estimator is known for having good "large sample" properties.

Definition 5.4.1

An unbiased estimator whose variance attains the CRLB is said to be **efficient**.

The **efficiency** of an estimator T of $\tau(\theta)$ is the ratio

$$\frac{CRLB_{\tau(\theta)}}{Var[T]}.$$

Note that an efficient estimator has efficiency 1.

Let $\widehat{\theta}_n$ be an MLE for a parameter θ. We usually just write $\widehat{\theta}$, but we wish to bring special attention to the dependence on the sample size n.

Under certain regularity conditions, involving the existence of certain derivatives, integrals, and logs, we have several nice properties for an MLE $\widehat{\theta} = \widehat{\theta}_n$ of θ.

- $\widehat{\theta}_n$ **exists** and is **unique**.
- $\widehat{\theta}_n$ is a **consistent** estimator of θ. That is, $\widehat{\theta}_n \xrightarrow{P} \theta$.
- $\widehat{\theta}_n$ is **asymptotically unbiased** for θ. That is, $\lim_{n\to\infty} \mathsf{E}[\widehat{\theta}_n] = \theta$.
- $\widehat{\theta}_n$ is **asymptotically efficient**. That is, $\lim_{n\to\infty} CRLB_\theta / Var[\widehat{\theta}_n] = 1$.
- $\widehat{\theta}_n$ is **asymptotically normal**. In particular, $\widehat{\theta}_n \overset{asymp}{\sim} N(\theta, CRLB_\theta)$.

The last two properties obviously require the regularity conditions used in the derivation of the Cramér-Rao lower bound. We will uncover sufficient conditions for some of the other listed properties here in Section 5.4.1.

Super Important Note

We are going to start seeing random variables in indicator subscripts, such as $I_{\{1<X<5\}}$. This means the same thing as $I_{(1,5)}(X)$ and will take on the value 1 if $1 < X < 5$ and 0 if X is not between 1 and 5. Note that, regardless of whether X is discrete or continuous, $I_{\{1<X<5\}}$ is a discrete random variable. We can take its expectation as follows.

$$\begin{aligned} \mathsf{E}[I_{\{1<X<5\}}] &= 0 \cdot P(I_{\{1<X<5\}} = 0) + 1 \cdot P(I_{\{1<X<5\}} = 1) \\ &= P(I_{\{1<X<5\}} = 1) \\ &= P(1 < X < 5) \end{aligned}$$

That last equality is true since that indicator is 1 if and only if $1 < X < 5$ and so we are finding the probability of equivalent events.

The particular event that $1 < X < 5$ used in this example is not important. In general, we will always have the following.

> "The expectation of an indicator of a random event is the probability of the event being indicated."

5.4.1 A Closer Look at the MLE Asymptotic Properties*

We will now prove or partially prove and comment on some of the listed properties of MLEs. It will be convenient for us to first adjust some notation.

Whenever we have $X_1, X_2, \ldots, X_n$ iid from a distribution with pdf $f(x;\theta)$, θ is a fixed parameter. However, for a likelihood function, θ is a variable that we want to maximize over. In order to make the distinction more clear in this Section, we will leave θ as an argument of the likelihood that we want to maximize over and we will assume the random sample was generated from a pdf with a very specific value of θ that we will call θ_0:

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} f(x;\theta_0).$$

We will change the notation for the likelihood $L(\theta)$ and log-likelihood $\ell(\theta)$ to $L_n(\theta)$ and $\ell_n(\theta)$, respectively, to stress their dependence on the sample size n.

1. $\widehat{\theta}_n$ **exists** and is **unique**.

   There are many sets of sufficient conditions for the MLE to exist and be unique and none of them are really specific to statistics. When does any function have a maximum and when is it unique? For existence, it is sufficient, for example, for the likelihood to be continuous and the parameter space to be compact. (We actually don't have a compact parameter space for many of the common distributions, so, thankfully, this is not a necessary condition!) For uniqueness, it is sufficient to have the likelihood function be concave, but this is hardly necessary and there are many interesting bimodal distributions with unique maximums. For our purposes, we will not try to enumerate a priori necessary or sufficient conditions. If you need to find an MLE for a model, you will have a specific likelihood function that simply will or won't have a unique maximum.

2. $\widehat{\theta}_n$ is a **consistent** estimator of θ_0. That is, $\widehat{\theta}_n \xrightarrow{P} \theta_0$.

   Sketch of Proof:

   - For $X_1, X_2, \ldots, X_n$ iid from a distribution with pdf $f(x;\theta)$, a likelihood can be taken to be the joint pdf:

     $$L_n(\theta) = \prod_{i=1}^{n} f(X_i;\theta).$$

     The log-likelihood is then

     $$\ell_n(\theta) = \sum_{i=1}^{n} \ln f(X_i;\theta).$$

     Note that we have random variables in the pdfs on the right-hand sides. Up until now, we have worked with likelihoods with fixed (lower case) observations for the X_i, maximized likelihoods with respect to θ, and plugged in the random (upper case) X_i to ultimately report an MLE. Equivalently, we could have maximized the likelihoods, with respect to θ, with the random X_i in from the start. When written this way, $L_n(\theta)$ and $\ell_n(\theta)$ are random quantities for each fixed θ. So, it makes sense to, for example, take their expectations and talk about them converging in probability or in distribution.

   - By the WLLN, we have

     $$\frac{1}{n}\ell_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ln f(x_i;\theta) \xrightarrow{P} \mathsf{E}[\ln f(X_1;\theta)]. \tag{5.4.5}$$

     In order for the WLLN to hold, we must have that $Var[\ln f(X_1;\theta)] < \infty$. This is regularity condition we need to impose on the model.

     Because $X_1 \sim f(x;\theta_0)$, the expectation on the right-hand side of (5.4.5) is (using continuous

notation)

$$\mathsf{E}[\ln f(X_1;\theta)] = \int_{-\infty}^{\infty} \ln f(x;\theta)\, f(x;\theta_0)\, dx.$$

While this should be clear since $X_1 \sim f(x;\theta_0)$, it is often stressed using the notation $\mathsf{E}_{\theta_0}[\ln f(X_1;\theta)]$.

- In (5.4.5), we have a function of θ converging in probability to another function of θ for all θ.

  The left-hand side is, by definition, maximized at the MLE $\widehat{\theta}_n$ of θ. The right-hand side, we will soon see, is maximized at θ_0.

  Claim: With very minor conditions on the likelihood, this implies that the maximizer[5] of the left-hand side converges in probability to the maximizer of the right-hand side.

  We will not prove this claim here but instead refer the reader to [4].

  Thus, we have

  $$\widehat{\theta}_n \xrightarrow{P} \theta_0.$$

  as desired.

- To finish the proof, we will show that the right-hand side of 5.4.5 is maximized at θ_0. Note that

$$\begin{aligned}
\mathsf{E}[\ln f(X_1;\theta)] - \mathsf{E}[\ln f(X_1;\theta_0)] &= \mathsf{E}[\ln f(X_1;\theta) - \ln f(X_1;\theta_0)] \\
&= \mathsf{E}\left[\ln \tfrac{f(X_1;\theta)}{f(X_1;\theta_0)}\right]
\end{aligned}$$

- The function $\ln x$ is concave. By Jensen's inequality (See Exercise 14 in Chapter 2.) we have that

$$\mathsf{E}\left[\ln \frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right] \leq \ln\left(\mathsf{E}\left[\frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right]\right)$$

- Thus, we have

$$\begin{aligned}
\mathsf{E}[\ln f(X_1;\theta)] - \mathsf{E}[f(X_1;\theta_0)] &\leq \ln\left(\mathsf{E}\left[\tfrac{f(X_1;\theta)}{f(X_1;\theta_0)}\right]\right) \\
&= \ln\left(\textstyle\int_{-\infty}^{\infty} \tfrac{f(x;\theta)}{f(x;\theta_0)}\, f(x;\theta_0)\, dx\right) \\
&= \ln\left(\textstyle\int_{-\infty}^{\infty} f(x;\theta)\, dx\right) \\
&= \ln(1) = 0.
\end{aligned}$$

[5] More technically, this is an "argmax" which refers to the argument that maximizes a function.

- So, we have show that

$$\mathsf{E}[\ln f(X_1;\theta)] \leq \mathsf{E}[f(X_1;\theta_0)]$$

for all θ in the parameter space. In other words, $\mathsf{E}[\ln f(X_1;\theta)]$, is a function of θ that is maximized at $\theta = \theta_0$.

3. $\widehat{\theta}_n$ is **asymptotically unbiased** for θ_0. That is,

$$\lim_{n\to\infty} \mathsf{E}[\widehat{\theta}_n] = \theta_0.$$

Here are some sufficient conditions for $\widehat{\theta}_n$ to be asymptotically unbiased for θ_0.

- $\widehat{\theta}_n$ is a consistent estimator for θ_0, and
- $Var[\widehat{\theta}_n]$ is bounded uniformly in θ_0.
  (i.e. $Var[\widehat{\theta}_n] \leq M$ for some constant M that doesn't depend on θ_0)

Proof of Asymptotic Unbiasedness:

Note that

$$\left|\mathsf{E}[\widehat{\theta}_n] - \theta_0\right| \quad = \quad \left|\mathsf{E}[\widehat{\theta}_n - \theta_0]\right| \leq \mathsf{E}[|\widehat{\theta}_n - \theta_0|].$$

That last inequality is true because averaging over the absolute value of terms is always greater than or equal to averaging the terms and then taking the absolute value. In the latter case, one could have cancellation of positive and negative terms.

Take any $\varepsilon > 0$. Note that

$$\begin{aligned}
\mathsf{E}[|\widehat{\theta}_n - \theta_0|] &= \mathsf{E}[|\widehat{\theta}_n - \theta_0| \cdot I_{\{|\widehat{\theta}_n - \theta_0| \leq \varepsilon\}} + |\widehat{\theta}_n - \theta_0| \cdot I_{\{|\widehat{\theta}_n - \theta_0| > \varepsilon\}}] \\
&= \mathsf{E}[|\widehat{\theta}_n - \theta_0| \cdot I_{\{|\widehat{\theta}_n - \theta_0| \leq \varepsilon\}}] + \mathsf{E}[|\widehat{\theta}_n - \theta_0| \cdot I_{\{|\widehat{\theta}_n - \theta_0| > \varepsilon\}}] \\
&\leq \varepsilon + \mathsf{E}[|\widehat{\theta}_n - \theta_0| \cdot I_{\{|\widehat{\theta}_n - \theta_0| > \varepsilon\}}].
\end{aligned}$$

We will rewrite this expectation in a "funny way" in order to set up for the use of the Cauchy-Schwarz inequality. We have

$$\begin{aligned}
\mathsf{E}[|\widehat{\theta}_n - \theta_0|] &\leq \varepsilon + \sqrt{\left(\mathsf{E}[|\widehat{\theta}_n - \theta_0| \cdot I_{\{|\widehat{\theta}_n - \theta_0| > \varepsilon\}}]\right)^2} \\
&\leq \varepsilon + \sqrt{\mathsf{E}\left[|\widehat{\theta}_n - \theta_0|^2\right] \cdot \mathsf{E}[I^2_{\{|\widehat{\theta}_n - \theta_0| > \varepsilon\}}]} \quad \Leftarrow \text{(Cauchy-Schwarz)} \\
&\leq \varepsilon + \sqrt{\mathsf{E}\left[(\widehat{\theta}_n - \theta_0)^2\right] \cdot \mathsf{E}[I_{\{|\widehat{\theta}_n - \theta_0| > \varepsilon\}}]} \\
&= \varepsilon + \sqrt{Var[\widehat{\theta}_n] \cdot P(|\widehat{\theta}_n - \theta_0| > \varepsilon)} \\
&\leq \varepsilon + \sqrt{M} \cdot \sqrt{P(|\widehat{\theta}_n - \theta_0| > \varepsilon)}.
\end{aligned}$$

If $\widehat{\theta}_n$ is a consistent estimator for θ_0, the probability in this expression goes to 0 as $n \to \infty$. Putting this all together, we have

$$\lim_{n\to\infty} \left|\mathsf{E}[\widehat{\theta}_n] - \theta_0\right| \leq \varepsilon + \sqrt{M} \cdot 0 = \varepsilon.$$

Since ε is arbitrary, we have shown that

$$\lim_{n\to\infty} \mathsf{E}[\widehat{\theta}_n] = \theta_0$$

as desired.[6]

4. $\widehat{\theta}_n$ is **asymptotically efficient**. That is,

$$\lim_{n\to\infty} \frac{CRLB_{\theta_0}}{Var[\widehat{\theta}_n]} = 1.$$

Here, it is tempting to say that the variance of the MLE approaches the Cramér-Rao lower bound as $n \to \infty$. This would not make sense, however, since the CRLB also depends on n.

5. $\widehat{\theta}_n$ is **asymptotically normal**. In particular, $\widehat{\theta}_n \overset{asymp}{\sim} N(\theta_0, CRLB_{\theta_0})$

i.e.

$$\frac{\widehat{\theta}_n - \theta_0}{\sqrt{CRLB_{\theta_0}}} \xrightarrow{d} N(0,1)$$

Proof:

- Note that

$$CRLB_{\theta_0} = \frac{1}{I_n(\theta_0)} \overset{iid}{=} \frac{1}{n \cdot I_1(\theta_0)}.$$

[6] Be warned that some people say that estimators with this property are "unbiased in the limit". They then use the phrase "asymptotically unbiased" to refer to the situation where $\widehat{\theta}_n$ is converging in distribution to something whose expected value is θ as in Property 5 in this list. These are not the same things!

This means that we want to show that

$$\sqrt{n \cdot I_1(\theta)}\,(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N(0,1),$$

or equivalently that

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N(0, 1/I_1(\theta_0)).$$

(Recall that, for $Z \sim N(0,1)$ and $c \neq 0$, $cZ \sim N(0, c^2)$.)

- Recall that, by Taylor's Theorem, for a nice (twice differentiable) function f we can write

$$f(x) = f(c) + f'(c)(x-c) + \frac{1}{2!}f''(\xi)(x-c)^2$$

for some ξ between x and c.

Here, we will let $f = \ell_n'$, $x = \widehat{\theta}_n$, and $c = \theta_0$.

(See the proof of Property 2 in this list if this notation doesn't make sense to you.)

This gives us

$$\ell_n'(\widehat{\theta}_n) = \ell_n'(\theta_0) + \ell_n''(\theta_0)(\widehat{\theta}_n - \theta_0) + \frac{1}{2!}\ell_n'''(\widetilde{\theta}_n)(\widehat{\theta}_n - \theta_0)^2$$

for some $\widetilde{\theta}_n$ between $\widehat{\theta}_n$ and θ_0. (Note that, because $\widehat{\theta}_n$ is a random variable, $\widetilde{\theta}_n$ will end up being random as well.)

For this, we need to assume that the log-likelihood is thrice differentiable. This is one of the regularity conditions needed for this result.

- Since $\ell_n'(\widehat{\theta}_n) = 0$, we have

$$0 = \ell_n'(\theta_0) + \ell_n''(\theta_0)(\widehat{\theta}_n - \theta_0) + \frac{1}{2!}\ell_n'''(\widetilde{\theta}_n)(\widehat{\theta}_n - \theta_0)^2 \tag{5.4.6}$$

By MLE Property 2 from this list, we know that $\widehat{\theta} \xrightarrow{P} \theta_0$. This implies that $\widehat{\theta}_n - \theta_0 \xrightarrow{P} 0$, and, by the Continuous Mapping Theorem, or part 4 of Theorem 2.3.3, that $(\widehat{\theta}_n - \theta_0)^2 \xrightarrow{P} 0$.

We would like to be able to use this to say that the third term here goes to zero. However, we will have a problem if $\ell'''(\widetilde{\theta}_n)$ is not converging to a constant.

- Since $\widetilde{\theta}_n$ is between $\widehat{\theta}_n$ and θ_0, it closer to θ_0 than $\widehat{\theta}_n$. Thus, for any $\varepsilon > 0$, we have

$$P(|\widetilde{\theta}_n - \theta_0| > \varepsilon) \leq P(|\widehat{\theta}_n - \theta_0| > \varepsilon).$$

Note that

$$\lim_{n\to\infty} P(|\widetilde{\theta}_n - \theta_0| > \varepsilon) \leq \lim_{n\to} P(|\widehat{\theta}_n - \theta_0| > \varepsilon) = 0$$

since $\widehat{\theta}_n \xrightarrow{P} \theta$ by Property 2 on this list.

Thus, we have that

$$\lim_{n\to\infty} P(|\widetilde{\theta}_n - \theta_0| > \varepsilon) \leq 0$$

which implies that the limit must actually equal zero since a probability can't be negative. This shows that

$$\widetilde{\theta}_n \xrightarrow{P} \theta_0.$$

Assuming $\ell'''(\theta)$ is continuous, we have

$$\widetilde{\theta}_n \xrightarrow{P} \theta_0 \qquad \Rightarrow \qquad \ell'''_n(\widetilde{\theta}_n) \xrightarrow{P} \ell'''_n(\theta_0).$$

We will also assume that $\ell'''_n(\theta_0)$ is bounded. The continuity and boundedness assumptions should be added to our list of regularity conditions.

$\ell'''_n(\widetilde{\theta}_n) \xrightarrow{P} \ell'''_n(\theta_0)$ and $\ell'''_n(\theta_0)$ bounded imply that the final term in (5.4.6) is going, in probability, to 0. We have, in some sense,

$$0 \approx \ell'_n(\theta_0) + \ell''_n(\theta_0)(\widehat{\theta}_n - \theta_0). \tag{5.4.7}$$

While there is a formal "little oh of p argument"[7] to be made here, we will assume equality in (5.4.7) in order to not "lose the forest for the trees". That is, we want to keep our eye on the big picture here and not get bogged down in some minutiae.

- From (5.4.7), we have that

$$\widehat{\theta}_n - \theta_0 = \frac{\ell'_n(\theta_0)}{-\ell''_n(\theta_0)}. \tag{5.4.8}$$

- Note that

$$\ell_n(\theta) = \ln L_n(\theta) = \ln \prod_{i=1}^{n} f(X_i;\theta) = \sum_{i=1}^{n} \ln f(X_i;\theta),$$

and so

$$\frac{1}{n}\ell'_n(\theta_0) = \overline{Y}$$

where $Y_i = \frac{\partial}{\partial\theta} \ln f(X_i;\theta_0)$ for $i = 1, 2, \ldots, n$.

We have

$$\mathsf{E}[Y_i] = \mathsf{E}\left[\frac{\partial}{\partial\theta} \ln f(X_i;\theta_0)\right] = 0$$

by computational simplification $\boxed{0}$ from Section 5.3.2.

We also have that

$$\begin{aligned} Var[Y_i] &= \mathsf{E}[Y_i^2] - \left(\mathsf{E}[Y_i^2]\right)^2 = \mathsf{E}[Y_i^2] \\ &= \mathsf{E}\left[\left(\tfrac{\partial}{\partial\theta} \ln f(X_i;\theta_0)\right)^2\right] = I_1(\theta). \end{aligned}$$

[7] We say that a sequence of random variables $\{X_n\}$ is "little oh of p" and write $X_n = o_p(1)$ if $X_n \xrightarrow{P} 0$.

Thus, by the Central Limit Theorem, we have that

$$\frac{1}{n}\ell_n'(\theta_0) = \overline{Y} \overset{asymp}{\sim} N(0, I_1(\theta_0)/n)$$

which means that

$$\frac{\frac{1}{n}\ell_n'(\theta_0)}{\sqrt{I_1(\theta_0)/n}} = \frac{\frac{1}{n}\ell_n'(\theta_0) - 0}{\sqrt{I_1(\theta_0)/n}} \xrightarrow{d} N(0,1).$$

- From (5.4.8), we have

$$\widehat{\theta}_n - \theta_0 = \frac{\ell_n'(\theta_0)}{-\ell_n''(\theta_0)} = \frac{\left(\frac{1}{n}\ell_n'(\theta_0)\right)/\sqrt{I_1(\theta_0)/n}}{\left(-\frac{1}{n}\ell_n''(\theta_0)\right)/\sqrt{I_1(\theta_0)/n}}$$

which implies that

$$\sqrt{n}\,(\widehat{\theta}_n - \theta_0) = \frac{\left(\frac{1}{n}\ell_n'(\theta_0)\right)/\sqrt{I_1(\theta_0)/n}}{\left(-\frac{1}{n}\ell_n''(\theta_0)\right)/\sqrt{I_1(\theta_0)}}$$

where the numerator on the right-hand side is converging in distribution to a $N(0,1)$ random variable.

- As for the denominator, note that

$$-\frac{1}{n}\ell_n''(\theta_0) = \overline{W}$$

where $W_i = -\left[\frac{\partial^2}{\partial\theta^2}\ln f(X_i;\theta_0)\right]$ for $i = 1, 2, \ldots, n$. By the WLLN, this converges, in probability, to

$$\mu_W := \mathsf{E}[W_i] = -\mathsf{E}\left[\frac{\partial^2}{\partial\theta^2}\ln f(X_i;\theta_0)\right] = I_1(\theta_0).$$

Thus, we have that

$$\left(-\frac{1}{n}\ell_n''(\theta_0)\right)/\sqrt{I_1(\theta_0)/n} \xrightarrow{P} I_1(\theta_0)/\sqrt{I_1(\theta_0)} = \sqrt{I_1(\theta_0)}.$$

- By Slutsky's Theorem, we have

$$\sqrt{n}\,(\widehat{\theta}_n - \theta_0) = \frac{\left(\frac{1}{n}\ell_n'(\theta_0)\right)/\sqrt{I_1(\theta_0)/n}}{\left(-\frac{1}{n}\ell_n''(\theta_0)\right)/\sqrt{I_1(\theta_0)}} \xrightarrow{d} \frac{Z}{\sqrt{I_1(\theta_0)}} \sim N(0, 1/I_1(\theta_0)),$$

as desired!

Example 5.4.1

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from the $\Gamma(2, \beta)$ distribution.

a) Find the MLE of β.

The pdf is

$$f(x;\beta) = \beta^2 x e^{-\beta x}\, I_{(0,\infty)}(x).$$

The joint pdf is

$$\begin{aligned} f(\vec{x};\beta) &\overset{iid}{=} \prod_{i=1}^{n} f(x_i;\beta) \\ &= \prod_{i=1}^{n} \beta^2 x_i e^{-\beta x_i} I_{(0,\infty)}(x_i) \\ &= \beta^{2n} \left(\prod_{i=1}^{n} x_i\right) e^{-\beta \sum x_i} \left(\prod_{i=1}^{n} I_{(0,\infty)}(x_i)\right). \end{aligned}$$

So, a likelihood function is

$$L(\beta) = \beta^{2n} e^{-\beta \sum x_i}.$$

The log-likelihood is

$$\ell(\beta) = 2n \ln \beta - \beta \sum x_i.$$

The derivative with respect to β is

$$\frac{\partial}{\partial \beta} \ell(\beta) = \frac{2n}{\beta} - \sum x_i = 0.$$

This implies that the MLE for β is

$$\widehat{\beta} = \frac{2n}{\sum_{i=1}^{n} X_i}.$$

This can also be written in terms of $\overline{X}$, but leaving it as a sum is convenient for the next part.

b) Is this MLE an unbiased estimator of β?

$$\mathsf{E}[\widehat{\beta}] = 2n\mathsf{E}\left[\frac{1}{\sum X_i}\right] = 2n\mathsf{E}\left[\frac{1}{W}\right]$$

where $W \sim \Gamma(2n, \beta)$.

Continuing,

$$\begin{aligned}
\mathsf{E}\left[\tfrac{1}{W}\right] &= \int_{-\infty}^{\infty} \tfrac{1}{w} \cdot f_W(w)\, dw \\
&= \int_0^{\infty} \tfrac{1}{w} \cdot \tfrac{1}{\Gamma(2n)} \beta^{2n} w^{2n-1} e^{-\beta w}\, dw \\
&= \int_0^{\infty} \tfrac{1}{\Gamma(2n)} \beta^{2n} \underbrace{w^{2n-2} e^{-\beta w}}_{\substack{\text{looks like} \\ \Gamma(2n-1,\beta) \text{ pdf}}} dw \\
&= \beta \tfrac{\Gamma(2n-1)}{\Gamma(2n)} \underbrace{\int_0^{\infty} \frac{1}{\Gamma(2n-1)} \beta^{2n-1} w^{2n-2} e^{-\beta w}\, dw}_{1} \\
&= \beta \tfrac{\Gamma(2n-1)}{\Gamma(2n)} = \tfrac{1}{2n-1}\beta.
\end{aligned}$$

So,

$$\mathsf{E}[\widehat{\beta}] = 2n\mathsf{E}\left[\frac{1}{W}\right] = \frac{2n}{2n-1}\beta.$$

The estimator is not unbiased. However, we see that it is asymptotically unbiased since

$$\lim_{n\to\infty} \mathsf{E}[\widehat{\beta}] = \lim_{n\to\infty} \frac{2n}{2n-1}\beta = 1 \cdot \beta = \beta.$$

(You might prefer to stress the dependence on n by using $\widehat{\beta}_n$ to denote the estimator instead of $\widehat{\beta}$.)

c) Consider the unbiased estimator

$$\widehat{\beta}_2 := \frac{2n-1}{2n}\widehat{\beta} = \frac{2n-1}{2n}\frac{2n}{\sum X_i} = \frac{2n-1}{\sum X_i}.$$

Is this an efficient estimator?

In the previous section, we saw that the CRLB for the variance of all unbiased estimators of β is

$$CRLB_\beta = \frac{\beta^2}{2n}.$$

We need to compute the variance for our estimator $\widehat{\beta}_2$ of β and see if it is equal to the CRLB. Note the following.

- The CRLB is associated with the parameter and not the estimator.
- We are not going to compare the CRLB "for β" to the variance of $\widehat{\beta}$ because $\widehat{\beta}$ is a biased estimator for β and the CRLB for β is a lower bound on the variance of all unbiased estimators of β.

$$\begin{aligned} Var[\widehat{\beta}_2] &= Var\left[\frac{2n-1}{\sum X_i}\right] = (2n-1)^2 Var\left[\frac{1}{W}\right] \\ &= (2n-1)^2\left\{\mathsf{E}\left[\frac{1}{W^2}\right] - \left(\mathsf{E}\left[\frac{1}{W}\right]\right)^2\right\} \end{aligned}$$

$$\begin{aligned} \mathsf{E}\left[\frac{1}{W^2}\right] &= \int_{-\infty}^{\infty} \frac{1}{w^2}\cdot f_W(w)\,dw \\ &= \int_0^{\infty} \frac{1}{w^2}\cdot\frac{1}{\Gamma(2n)}\beta^{2n}w^{2n-1}e^{-\beta w}\,dw \\ &= \int_0^{\infty}\frac{1}{\Gamma(2n)}\beta^{2n}\underbrace{w^{2n-3}e^{-\beta w}}_{\substack{\text{looks like}\\ \Gamma(2n-2,\beta)\text{ pdf}}}dw \\ &= \beta^2\frac{\Gamma(2n-2)}{\Gamma(2n)}\underbrace{\int_0^{\infty}\frac{1}{\Gamma(2n-2)}\beta^{2n-2}w^{2n-2}e^{-\beta w}\,dw}_{1} \\ &= \beta^2\frac{\Gamma(2n-2)}{\Gamma(2n)} = \frac{1}{(2n-1)(2n-2)}\beta^2. \end{aligned}$$

Putting it all together, we have

$$\begin{aligned} Var[\widehat{\beta}_2] &= (2n-1)^2\left\{\mathsf{E}\left[\frac{1}{W^2}\right] - \left(\mathsf{E}\left[\frac{1}{W}\right]\right)^2\right\} \\ &= (2n-1)^2\left\{\frac{1}{(2n-1)(2n-2)}\beta^2 - \left(\frac{1}{2n-1}\beta\right)^2\right\} \\ &= \frac{1}{2n-2}\beta^2. \end{aligned}$$

Note that

$$Var[\widehat{\beta}_2] = \frac{1}{2n-2}\beta^2 > \frac{1}{2n}\beta^2 = CRLB_\beta.$$

Our estimator did not achieve the CRLB, so it is not an efficient estimator of β. It is, however, asymptotically efficient since

$$\lim_{n\to\infty}\frac{CRLB_\beta}{Var[\widehat{\beta}_2]} = \lim_{n\to\infty}\frac{[1/(2n)\beta^2]}{[1/(2n-1)]\beta^2} = \lim_{n\to\infty}\frac{2n-1}{2n} = 1.$$

d) What is the asymptotic distribution of the MLE of β?

According to the property list for MLEs, we have that

$$\widehat{\beta} \overset{asymp}{\sim} N(\beta, CRLB_\beta) = N(\beta, \beta^2/(2n)).$$

5.5 Uniformly Minimum Variance Unbiased Estimators (UMVUEs)

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with pdf $f(x;\theta)$. Consider estimating some function $\tau(\theta)$ which may be θ itself or something more general. We have been in several situations where we produced two or more unbiased estimators and chosen the one with smaller variance as the "better one". For example, suppose that the sample comes from the exponential distribution with rate θ and that we wish to estimate $\tau(\theta) = 1/\theta$. Since this is the mean of the distribution, one unbiased estimator is

$$\widehat{\tau_1(\theta)} = \overline{X}.$$

Since $X_{(1)} = \min(X_1, X_2, \ldots, X_n) \sim exp(rate = n\theta)$, we know that $\mathsf{E}[X_{(1)}] = 1/(n\theta)$ and so a second unbiased estimator of $\tau(\theta)$ is

$$\widehat{\tau_2(\theta)} = nX_{(1)}.$$

Note that

$$Var[\widehat{\tau_1(\theta)}] = Var[\overline{X}] = \frac{Var[X_1]}{n} = \frac{(1/\theta^2)}{n} = \frac{1}{n\theta^2},$$

and that

$$Var[\widehat{\tau_2(\theta)}] = Var[nX_{(1)}] = n^2\, Var[X_{(1)}] = n^2 \frac{1}{(n\theta)^2} = \frac{1}{\theta^2},$$

and so have that

$$Var[\widehat{\tau_1(\theta)}] \leq Var[\widehat{\tau_2(\theta)}]$$

for all $\theta > 0$ and for all sample sizes $n \geq 1$. (Furthermore we have that the first estimator gets better and better for larger sample sizes while the second just stays the same!)

Thus, out of these two unbiased estimators for $\tau(\theta) = 1/\theta$ we have that $\widehat{\tau_1(\theta)} = \overline{X}$ is the better estimator in terms of having smaller variance. It is important that the inequality hold for all θ. If one estimator were better for some values of θ but not others, we wouldn't be able to say that one estimator was, overall, better then the other. The value of θ is unknown– this is kind of the entire point of the estimation problem!.

For this exponential example, there are likely even more unbiased estimators to choose from. Is there a "best one" in terms of having smallest variance (for all values of θ in the parameter space) among all unbiased estimators? In mathematics, the word "uniformly" means "for all".[8] We are ready to define the "UMVUE"– the greatest estimator in all the land!

[8]For example, let $\varepsilon > 0$. A function $g(x)$ is continuous if, for every x_0 there exists a $\delta > 0$ such that for all x within a distance δ of x_0, the value $g(x)$ is within ε of $g(x_0)$. Here, δ can be dependent on x_0 but if we can find one δ that works for all x_0, we say that g is *uniformly continuous*.

Definition 5.5.1

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with pdf $f(x;\theta)$. Consider estimating a function $\tau(\theta)$. The **uniformly minimum variance unbiased estimator (UMVUE)** for $\tau(\theta)$ is the unbiased estimator that has smallest variance "uniformly in θ", meaning for all values of θ.

Note the following.

- We said "the" UMVUE and not "an" UMVUE. We will eventually prove that, if an UMVUE exists, it is unique.
- An UMVUE may not exist!
- Some people call this the "MVUE" for "minimum variance unbiased estimator".

For us, "data" is a random sample $X_1, X_2, \ldots, X_n$ from a distribution with pdf $f(x;\theta)$ where θ is typically unknown. Our goal is to estimate θ or some function $\tau(\theta)$ from the data. We often find ourselves considering some function of the data, such as, for example, the sum of the X_i, or the sample mean, or the minimum value in the data. Recall that functions of the data, such as these, are known as "statistics".

We shall often refer to the data as a vector $\vec{X} = (X_1, X_2, \ldots, X_n)$ and will usually use the letter T to denote a statistic which is some function of the data:

$$T = t(\vec{X}),$$

but we will use the letter S (where $S = s(\vec{X})$) in the special case of the statistics defined in the next section.

5.5.1 Sufficient Statistics

Let $X_1, X_2, \ldots, X_n$ be a random sample from some distribution with pdf $f(x;\theta)$, and consider, for example, the statistic $T = \sum_{i=1}^{n} X_i$.

This statistic condenses the n-dimensional random sample down to one dimension.

Question: If we know only this statistic instead of knowing the original sample, do we have less information about θ?

For example, suppose that $X_1, X_2, \ldots, X_n \overset{iid}{\sim} Bernoulli(p)$ and that we want to estimate p. Recall that each X_i is 1 with probability p, and zero otherwise. One possibility for estimating p from the data is to return the

proportion of 1's observed in the data set. This is the sum of the X_i divided by the total number n. In other words, the sample mean. If you had a house elf that carried around your data for you, do you really need her try to scoop up and juggle n clumsy pieces of data in her arms or would it be enough if she just carried around the sum $\sum_{i=1}^{n} X_i$ for you? Wouldn't that be enough? Wouldn't it be sufficient?

Definition (Informal)

A statistic S is **sufficient for** θ if it condenses the random sample $X_1, X_2, \ldots, X_n$ without sacrificing any information about θ.

"sufficient" = "enough"

If you know S, then knowing $(X_1, X_2, \ldots, X_n)$ does not give you any more information about θ.

This is the idea behind a sufficient statistic. You might want to have your house elf carry two statistics for you, such as the sum of the squares of the X_i in addition to the sum of the X_i. We can call $\sum X_i$ and $\sum X_i^2$ a set of sufficient statistics or we can have one vector-valued sufficient statistic

$$S = \left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2 \right).$$

Let's make the term "information", as it is used here, more precise.

Definition 5.5.2

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with pdf $f(x; \theta)$. S is a **sufficient statistic** for estimating θ if the conditional distribution of $(X_1, \ldots, X_n)$ given S does not depend on θ.

Example 5.5.1

An unfair coin with $p = P(\text{"Heads"})$ is tossed n times and the outcome is recorded for each toss.

i.e. We have a random sample $X_1, \ldots, X_n \overset{iid}{\sim} Bernoulli(p)$ where

$$X_i = \begin{cases} 1 & , \text{ if heads} \\ 0 & , \text{ if tails} \end{cases}$$

For guessing p, it would seem that knowing the total number of heads, $S = \sum_{i=1}^{n} X_i$, should provide as much information about p as knowing the actual 0 and 1 outcomes.

In this example, we are going to show that S is sufficient for estimating p.

Since $X_i \sim Bernoulli(p)$, the pdf is $f(x;p) = p^x(1-p)^{1-x}\, I_{\{0,1\}}(x)$.

The joint pdf for $X_1, \ldots, X_n$ is then

$$f(\vec{x};p) \overset{iid}{=} \prod_{i=1}^{n} f(x_i;p) = p^{\sum x_i}(1-p)^{n-\sum x_i} \prod_{i=1}^{n} I_{\{0,1\}}(x_i).$$

Note that $S = \sum_{i=1}^{n} X_i \sim bin(n,p)$.

The conditional pdf of $(X_1, \ldots, X_n)$ is

$$\begin{aligned} f_{X_1,\ldots,X_n|S}(x_1,\ldots,x_n|s) &\overset{discrete}{=} P(X_1 = x_1, \ldots, X_n = x_n | S = s) \\ &= \frac{P(X_1=x_1,\ldots,X_n=x_n,S=s)}{P(S=s)} \end{aligned} \tag{5.5.9}$$

The Numerator: Consider the case, just as an example, where $n = 2$ and we have some fixed numerical values in for x_1, x_2, and s. Note that $S = X_1 + X_2$. The numerator might be

$$P(X_1 = 2, X_2 = 3, X_1 + X_2 = 1)$$

This probability is clearly zero since, if $X_1 = 2$ and $X_2 = 3$, we can not have $X_1 + X_2 = 1$.

On the other hand, the probability

$$P(X_1 = 2, X_2 = 3, X_1 + X_2 = 5)$$

is not necessarily zero and, furthermore, the third argument is redundant. We can simply write

$$P(X_1 = 2, X_2 = 3, X_1 + X_2 = 1) = P(X_1 = 2, X_2 = 3)$$

So, in general, we have the numerator of (5.5.9) as

$$P(X_1 = x_1, \ldots, X_n = x_n, S = s) = \begin{cases} P(X_1 = x_1, \ldots, X_n = x_n) & , \text{ if } s = \sum x_i \\ 0 & , \text{ if } s \neq \sum x_i \end{cases}$$

$$= \begin{cases} p^{x_1}(1-p)^{1-x_1} I_{\{0,1\}}(x_1) \cdots p^{x_n}(1-p)^{n-x_n} I_{\{0,1\}}(x_n) & , \text{ if } s = \sum_{i=1}^n x_i \\ 0 & , \text{ if } s \neq \sum_{i=1}^n x_i \end{cases}$$

$$= \begin{cases} p^{\sum x_i}(1-p)^{n-\sum x_i} \prod_{i=1}^n I_{\{0,1\}}(x_i) & , \text{ if } s = \sum_{i=1}^n x_i \\ 0 & , \text{ if } s \neq \sum_{i=1}^n x_i \end{cases}$$

The Denominator: Since $S \sim bin(n,p)$,

$$P(S=s) = \begin{pmatrix} n \\ s \end{pmatrix} p^s (1-p)^{n-s} I_{\{0,1,\ldots,n\}}(s).$$

Now that we have computed the numerator and denominator, (5.5.9) becomes

$$P(X_1 = x_1, \ldots, X_n = x_n | S = s) = \begin{cases} \dfrac{p^{\sum x_i}(1-p)^{n-\sum x_x} \prod_{i=1}^n I_{\{0,1\}}(x_i)}{\begin{pmatrix} n \\ s \end{pmatrix} p^s (1-p)^{n-s} I_{\{0,1,\ldots,n\}}(s)} & , \text{ if } s = \sum_{i=1}^n x_i \\ 0 & , \text{ if } s \neq \sum_{i=1}^n x_i \end{cases}$$

Since $s = \sum x_i$ in the first expression, we can cancel many things and are left with

$$P(X_1 = x_1, \ldots, X_n = x_n | S = s) = \begin{cases} \dfrac{\prod_{i=1}^n I_{\{0,1\}}(x_i)}{\begin{pmatrix} n \\ s \end{pmatrix} I_{\{0,1,\ldots,n\}}(s)} & , \text{ if } s = \sum_{i=1}^n x_i \\ 0 & , \text{ if } s \neq \sum_{i=1}^n x_i \end{cases}$$

Since this does not depend on p, we have, by definition that S is a sufficient statistic for estimating p for this Bernoulli distribution!

Note that, for $S = s(\vec{X})$, the "$S = s$" part of the numerator of the conditional pdf will always drop out like it did above when $s = s(\vec{x})$ and that it will always be zero when $s \neq s(\vec{x})$). So, we will always have

$$f_{X_1,X_2,\ldots,X_n|S}(x_1,x_2,\ldots,x_n|s) = \begin{cases} \frac{f(\vec{x};\theta)}{f_S(s;\theta)} & , \quad s = s(\vec{x}) \\ \\ 0 & , \quad s \neq s(\vec{x}) \end{cases}$$

Since this must be "θ-free", we require (for sufficiency) that

$$\frac{f(\vec{x};\theta)}{f_S(s;\theta)} = h(\vec{x}),$$

for some function $h(\vec{x})$ where, as the notation suggests, does not depend on θ.

Equivalently, using the fact that $s = s(\vec{x})$, we have

$$f(\vec{x};\theta) = h(\vec{x}) f_S(s(\vec{x});\theta).$$

Assuming that we don't know the pdf for S, or even what S is ahead of time, we at least are looking for the joint pdf of the X's to split up into an "x part" and a part with x's and θ's mixed together and where the x's are appearing through some function that we will call $s(\vec{x})$.

(Neyman's) Factorization Criterion for Sufficiency

If the joint pdf for $X_1, X_2, \ldots, X_n$ can be written as

$$f(\vec{x};\theta) = h(\vec{x}) g(s(\vec{x});\theta)$$

for some function h that does not depend on θ and some function g that depends on the x's only through some function $s(\vec{x})$, then

$$S = s(\vec{X}) \text{ is sufficient for estimating } \theta.$$

Example 5.5.2

For the random sample from the Bernoulli distribution,

$$\begin{aligned} f(\vec{x};p) &= \textstyle\prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}\, I_{\{0,1\}}(x_i) \\ &= p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i} \textstyle\prod_{i=1}^{n} I_{\{0,1\}}(x_i) \\ &= \underbrace{\prod_{i=1}^{n} I_{\{0,1\}}(x_i)}_{h(\vec{x})} \underbrace{p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}}_{g(s(\vec{x});p)} \end{aligned}$$

In the part where the p's and x's are "mixed", the x's appear as the sum $s(\vec{x}) = \sum_{i=1}^{n} x_i$. By the Factorization Criterion, $S = \sum_{i=1}^{n} X_i$ is sufficient for estimating p.

Sufficient statistics are far from unique. If you write down the joint pdf and factor out the "pure x part", what you are left with can usually be written in many different ways. We have to ask ourselves what statistic or statistics we would need from the data in order to evaluate the $g(s(\vec{x},p))$ part. In the above example, it is sufficient that we have only the sum $S = \sum_{i=1}^{n} X_i$, but we wouldn't lose anything by throwing in extra statistics. We could say that $S = (\sum X_i, \sum X_i^2)$ is sufficient for estimating p in the above example. It is also "sufficient" if your house elf brings you all of the data. That is, we always have that the n-dimensional statistic

$$S = (X_1, X_2, \ldots, X_n)$$

is sufficient for estimating any parameter of any distribution! In Section 5.6, we discuss the concept of *minimal sufficient statistics*. Heuristically speaking, a minimal sufficient statistic is a statistic that is sufficient and has minimal dimension. (i.e. We don't throw in any extra statistics.) In actuality, this is not exactly the definition of a minimal sufficient statistic though. Once minimal sufficiency is defined, we will see an example where a one-dimensional sufficient statistic is not minimal. Still, we think this (incorrect) heuristic in terms of dimensionality gets most of the idea across.

Example 5.5.3

Let $X_1, X_2, \ldots, X_n$ be a random sample from the $unif(0,\theta)$ distribution. Find a sufficient statistic for θ.

The joint pdf is

$$f(\vec{x};\theta) = \prod_{i=1}^{n} \frac{1}{\theta} I_{(0,\theta)}(x_i) = \frac{1}{\theta^n} \prod_{i=1}^{n} I_{(0,\theta)}(x_i). \tag{5.5.10}$$

This product of indicators could be written in terms of the minimum and maximum of the sample as $I_{(0,\theta)}(x_{(1)})\, I_{(0,\theta)}(x_{(n)})$ since this is 0 or 1 whenever the full product is 0 or 1. That is, the minimum and maximum of the sample is between 0 and θ if and only if each individual x_i is between 0 and θ.

Note that the product of indicators in (5.5.10) is also equivalent to the smaller product $I_{(0,x_{(n)})}(x_{(1)})\, I_{(0,\theta)}(x_{(n)})$. This factorization is especially useful when one is trying to separate out as much "pure x stuff" as possible. In this case, we would have

$$f(\vec{x};\theta) = \underbrace{I_{(0,x_{(n)})}}_{h(\vec{x})} \cdot \underbrace{\frac{1}{\theta^n} I_{(0,\theta)}(x_{(n)})}_{g(s(\vec{x});\theta)}$$

where $s(\vec{x}) = x_{(n)}$.

So, by the Factorization Criterion, we have that $S = s(\vec{X}) = X_{(n)}$ is sufficient for estimating θ.

If we used a different representation of the indicators such as

$$f(\vec{x};\theta) = \underbrace{\frac{1}{\theta^n} I_{(0,\theta)}(x_{(1)})\, I_{(0,\theta)}(x_{(n)})}_{g(s(\vec{x});\theta)},$$

(here $h(\vec{x}) = 1$), we would say that the vector $S = (X_{(1)}, X_{(n)})$ is sufficient for θ (or that $X_{(1)}$ and $X_{(n)}$ are "jointly sufficient").

Indeed, for estimating the upper endpoint of possible values for the X's, it makes sense that looking only at the maximum in the sample is "enough". It is also "enough" to look both at the maximum and minimum though this information is "more than enough". In fact, since a sufficient statistic can be vector-valued, we can always say that $S = (X_1, X_2, \ldots, X_n)$ is sufficient for estimating θ. However, $S = X_{(n)}$ is minimal sufficient from the point of view of dimensionality and also by the formal definition of minimal sufficiency given in Section 5.6.

Super Important Note

We have been using the phrase "sufficient for θ", or, "sufficient for estimating θ". If you find yourself wanting to estimate a function like, for example, $\tau(\theta) = e^{-\theta}$, and you will be using a procedure that requires a sufficient statistic, nothing changes. If a statistic is sufficient for estimating θ, it is sufficient for estimating $e^{-\theta}$ because if the conditional distribution of $(X_1, X_2, \ldots, X_n)$ given S doesn't depend on θ it sure as heck doesn't depend on $e^{-\theta}$!

We sometimes just say that a statistic S is "sufficient for the model" where "the model" is the underlying distribution. For example, $S = \sum_{i=1}^{n} X_i$ is sufficient for the Bernoulli distribution.

5.5.2 Conditional Expectation and Variance

In this Section, we will review conditional expectation and conditional variance. (See Section 0.11 for some review of conditional probability and the definition of conditional pdfs.) We will rely heavily on these concepts in our search for the UMVUE. We will generally prove things in this Section in the case of continuous random variables even though all results also hold for discrete distributions.

Let X and Y be random variables, with pdfs $f_X(x)$ and $f_Y(y)$, respectively. Let $f_{X,Y}(x, y)$ be the joint pdf for X and Y and let $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$ be the two conditional pdfs as defined in Section 0.11.

1. $\mathsf{E}[X]$ is a constant.

   By "constant", we mean non-random. For example, we may be talking about a sequence of random variables were the nth one has an expectation that depends on n. In this case, the expectations may form a non-constant sequence of numbers. The point is that they will not be a sequence of random variables.

   The expected value of X is defined as

   $$\mathsf{E}[X] = \int_{-\infty}^{\infty} x\, f_X(x).$$

   There is nothing random on the right-hand side!

2. $\mathsf{E}[X|Y = y]$ is a function of y.

   $\mathsf{E}[X|Y = y]$ is the expected value of X taken with respect to the conditional pdf for X given that $Y = y$. We are still averaging over x values. This expectation is defined as

   $$\mathsf{E}[X|Y = y] = \int_{-\infty}^{\infty} x\, f_{X|Y}(x|y)\, dx.$$

As this is a definite integral, we will be plugging in values for x and the final answer will be a function of y.

3. $\mathsf{E}[X|Y]$ is a random variable.

   Once we have computed $\mathsf{E}[X|Y=y]$, we have a function of y. Let's call it $g(y)$. Then $\mathsf{E}[X|Y]$ is defined as $g(Y)$. This is a function with a random variable plugged in and will itself be random assuming $g(y)$ is not constant in y.

4. $\mathsf{E}[\mathsf{E}[X|Y]] = \mathsf{E}[X]$

   This is known as the law of iterated expectation. Note that since $\mathsf{E}[X|Y]$ is a function of Y, the outer expectation here is taken with respect to the pdf for Y. We have

$$\begin{aligned}
\mathsf{E}[\mathsf{E}[X|Y]] &= \int_{-\infty}^{\infty} \mathsf{E}[X|Y=y]\, f_Y(y)\, dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x\, f_{X|Y}(x|y)\, dx\, f_Y(y)\, dy \\
&= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_{X|Y}(x|y) f_Y(y)\, dy\, dx \\
&= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dy\, dx \\
&= \int_{-\infty}^{\infty} x\, f_X(x)\, dx = \mathsf{E}[X] \quad \surd
\end{aligned}$$

5. $Var[X] = Var[\mathsf{E}[X|Y]] + \mathsf{E}[Var[X|Y]]$

   Note that all rules of expectation and variance hold in "conditional land". For example, expectation is still a linear operator. In all of the proofs of expectation and variance results that we have seen (so far not involving conditional distributions), we can simply replace the pdfs with conditional pdfs and the proofs will otherwise be identical. As another example,

$$Var[X|Y] = \mathsf{E}[X^2|Y] - (\mathsf{E}[X|Y])^2$$

   for the same reason that

$$Var[X] = \mathsf{E}[X^2] - (\mathsf{E}[X])^2 .$$

   To prove the claim, note that

$$\begin{aligned}
\mathsf{E}[Var[X|Y]] &= \mathsf{E}\left[\mathsf{E}[X^2|Y] - (\mathsf{E}[X|Y])^2\right] \\
&= \mathsf{E}[\mathsf{E}[X^2|Y]] - \mathsf{E}[(\mathsf{E}[X|Y])^2] \\
&= \mathsf{E}[X^2] - \mathsf{E}[(\mathsf{E}[X|Y])^2] \quad \Leftarrow \text{(by Property 4 above)}
\end{aligned}$$

Just as we can use the fact that $Var[X] = \mathsf{E}[X^2] - (\mathsf{E}[X])^2$ to say that $\mathsf{E}[X^2] = Var[X] + (\mathsf{E}[X])^2$, we can replace the random variable X with the random variable $\mathsf{E}[X|Y]$ and write

$$\begin{aligned}
\mathsf{E}[(\mathsf{E}[X|Y])^2] &= Var[\mathsf{E}[X|Y]] + (\mathsf{E}[\mathsf{E}[X|Y]])^2 \\
&= Var[\mathsf{E}[X|Y]] + (\mathsf{E}[X])^2 \quad \Leftarrow \text{(by Property 4 again)}
\end{aligned}$$

In conclusion, we have

$$\begin{aligned}
\mathsf{E}[Var[X|Y]] &= \mathsf{E}[X^2] - \mathsf{E}[(\mathsf{E}[X|Y])^2] \\
&= \mathsf{E}[X^2] - (Var[X|Y] + (\mathsf{E}[X])^2) \\
&= Var[X] - Var[\mathsf{E}[X|Y]]
\end{aligned}$$

Thus, we have that

$$Var[X] = Var[\mathsf{E}[X|Y]] + \mathsf{E}[Var[X|Y]],$$

as desired.

5.5.3 The Rao-Blackwell Theorem

Let $\vec{X} = (X_1, X_2, \ldots, X_n)$ be a random sample from a distribution with pdf $f(x;\theta)$. Our goal continues to be to find the uniformly minimum variance unbiased estimator for a given function $\tau(\theta)$. Suppose that we have some unbiased estimator in the form of a statistic $T = t(\vec{X})$. That is, $\mathsf{E}[T] = \tau(\theta)$. If we have a sufficient statistic S, we can make your estimator even better (smaller variance) using the **Rao-Blackwell Theorem**.

The Rao-Blackwell Theorem

Suppose we want to estimate some function of θ (which may be θ itself) that we have named $\tau(\theta)$ based on some random sample $\vec{X}$.

Suppose that we have some unbiased estimator $T = t(\vec{X})$ so that

$$\mathsf{E}[T] = \tau(\theta).$$

Suppose that we have some sufficient statistic S for the distribution.

Then

$$T^* := \mathsf{E}[T|S]$$

is also unbiased for $\tau(\theta)$ and that $Var[T^*] \leq Var[T]$.

Furthermore, T^* is a function of S and the variance is strictly lower than the variance of T unless T^* and T are the same!

Proof :

- T^* is unbiased for $\tau(\theta)$ since

$$\mathsf{E}[T^*] = \mathsf{E}[\mathsf{E}[T|S]] = \mathsf{E}[T] = \tau(\theta).$$

- $Var[T^*] \leq Var[T]$ since

$$\begin{aligned} Var[T] &= Var[\mathsf{E}(T|S)] + \mathsf{E}[Var(T|S)] \\ &\geq Var[\mathsf{E}[T|S]] = Var[T^*] \end{aligned}$$

  since $\mathsf{E}[Var(T|S)] \geq 0$ since that variance inside is non-negative.
- $T^* = \mathsf{E}[T|S]$ is a function of S by the definition of that expectation.
- Equality of variances is achieved if and only if the thing we dropped to show the inequality is zero. That is, equality is achieved if and only if

$$\mathsf{E}[Var(T|S)] = 0.$$

  Since the variance inside is non-negative, this probability weighted average is zero if and only if $Var(T|S) = 0$. With no variability at all, this means that T, given S is a constant. If it is constant, it's expectation (given S) is just T itself. That is, the equality of variances is achieved if and only if

$$T^* = \mathsf{E}[T|S] = T.$$

The Rao-Blackwell Theorem tells us that, while searching for the UMVUE for $\tau(\theta)$, we can limit our attention to functions of sufficient statistics because estimators that are functions of non-sufficient statistics can be "Rao-Blackwellized" into better (lower variance) estimators. This is both exciting and not exciting at the same time. Since that the entire n-dimensional vector $S = (X_1, X_2, \ldots, X_n)$ is a sufficient statistic, have we really narrowed anything down?

Note that our proof of the Rao-Blackwell Theorem holds regardless of whether or not S is sufficient. However, we need sufficiency to ensure that $T^* = \mathsf{E}[T|S]$ does not depend on θ. Otherwise, it would not technically be a statistic which is a function of the data only. Regardless of the technical definition of a statistic, it certainly would not make a very good estimator of the unknown θ if we define it in terms of θ!

Example 5.5.4

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from the Poisson distribution with parameter λ:

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} Poisson(\lambda)$$

Our goal is to find an unbiased estimator of $\tau(\lambda) = e^{-\lambda}$.

Our first instinct here might be to take $\overline{X}$, as an unbiased estimator of λ, and consider estimating $\tau(\lambda)$ with $e^{-\overline{X}}$. This transformation likely does not preserve unbiasedness so, after computing its expectation, we might try to adjust the estimator until it is unbiased. Indeed,

$$S = \sum_{i=1}^{n} X_i \sim Poisson(n\lambda)$$

and so

$$\mathsf{E}\left[e^{-\overline{X}}\right] = \mathsf{E}\left[e^{-\frac{1}{n}S}\right] = M_S\left(-\frac{1}{n}\right)$$

where $M_S(t)$ is the moment generating function of S.

From Section 1.5.2, we know that

$$M_S(t) = e^{n\lambda(e^t - 1)}$$

for all t. Thus, we have that

$$\mathsf{E}\left[e^{-\overline{X}}\right] = M_S\left(-\frac{1}{n}\right) = e^{n\lambda(e^{-1/n}-1)}.$$

It appears that $e^{-\overline{X}}$ is not easily adjustable, certainly not by simply adding, subtracting, multiplying, or dividing by a constant, to something with expected value equal to $e^{-\lambda}$.

Let's try a different approach. The pdf for the Poisson distribution with parameter λ is

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}\, I_{\{0,1,2,\ldots\}}(x).$$

We observe that $P(X = 0) = e^{-\lambda}$. In terms of the random sample, we have $P(X_i = 0) = e^{-\lambda}$.

We know that we can always get an unbiased estimator for a probability by using an indicator. Thus, we have

$$\widehat{\tau(\lambda)} = I_{\{X_1=0\}}.$$

as an unbiased estimator. (This could just as easily have been, for example, $\widehat{\tau(\lambda)} = I_{\{X_5=0\}}$.)

Technically, we have completed the task, though it seems wasteful not to use the entire sample. Let's find a sufficient statistic for this model and "Rao-Blackwellize" this unbiased estimator to get something better. To do this, we need a sufficient statistic for this model.

The joint pdf is

$$\begin{aligned}
f(\vec{x};\lambda) &\overset{iid}{=} \textstyle\prod_{i=1}^n f(x_i;\lambda) \\
&= \textstyle\prod_{i=1}^n \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\, I_{\{0,1,2,\ldots\}}(x_i) \\
&= \frac{e^{-n\lambda}\lambda^{\sum x_i}}{\prod_{i=1}^n x_i!}\, \textstyle\prod_{i=1}^n I_{\{0,1,2,\ldots\}}(x_i) \\
&= \frac{\prod_{i=1}^n I_{\{0,1,2,\ldots\}}(x_i)}{\prod_{i=1}^n x_i!}\, e^{-n\lambda}\, \lambda^{\sum_{i=1}^n x_i}.
\end{aligned}$$

We have pulled out as much "pure x stuff" as possible in order to apply the Factorization Criterion. We have

$$h(\vec{x}) = \frac{\prod_{i=1}^{n} I_{\{0,1,2,\ldots\}}(x_i)}{\prod_{i=1}^{n} x_i!}$$

and the rest is

$$g(s(\vec{x});\lambda) = e^{-n\lambda}\,\lambda^{\sum_{i=1}^{n} x_i}.$$

We look at the g function to see how the x_i appear as a statistic or statistics. Here, they appear as a sum. That is, we have that

$$s(\vec{x}) = \sum_{i=1}^{n} x_i.$$

Thus, we have that

$$S = s(\vec{X}) = \sum_{i=1}^{n} X_i$$

is a sufficient statistic for this model by the Factorization Criterion.

Let's summarize where we are at this point. We want an unbiased estimator of $\tau(\lambda) = e^{-\lambda}$. We have one and it is $T = I_{\{X_1=0\}}$. We want to improve this estimator or at least have it use all of the information in the random sample $X_1, X_2, \ldots, X_n$. We have a sufficient statistic $S = \sum_{i=1}^{n}$. Also, we are in the Section of the text entitled "The Rao-Blackwell Theorem"! Let's try to "Rao-Blackwellize" the unbiased estimator to see if we can get something better. We consider

$$T^* = \mathsf{E}[T|S].$$

We begin by fixing S and computing $\mathsf{E}[T|S = s]$. Note that $S = \sum_{i=1}^{n} X_i \sim Poisson(n\lambda)$ and S takes on values in $\{0, 1, 2, \ldots\}$. If we have observed $S = s$, then it must be a possible value to observe– it must be in this set.

So, for some $s \in \{0, 1, 2, \ldots\}$, we have

$$\begin{aligned} \mathsf{E}[T|S=s] &= \mathsf{E}[I_{\{X_1=0\}}|S=s] \\ &= 0 \cdot P(I_{\{X_1=0\}}=0|S=s) + 1 \cdot P(I_{\{X_1=0\}}=1|S=s) \\ &= P(I_{\{X_1=0\}}=1|S=s) \\ &= P(X_1=0|S=s) \end{aligned}$$

since the indicator $I_{\{X_1=0\}}$ is equal to 1 if and only if X_1 is zero.

Now

$$\begin{aligned} \mathsf{E}[T|S=s] &= P(X_1=0|S=s) \\ &= \frac{P(X_1=0,S=s)}{P(S=s)} \\ &= \frac{P(X_1=0,\sum_{i=1}^n X_i=s)}{P(S=s)} \\ &= \frac{P(X_1=0,\sum_{i=2}^n X_i=s)}{P(S=s)}. \end{aligned}$$

Since X_1 is independent of the sum of the remaining X_i, this becomes

$$\begin{aligned} \mathsf{E}[T|S=s] &= \frac{P(X_1=0)\cdot P(\sum_{i=2}^n X_i=s)}{P(S=s)}. \\ &= \frac{\frac{e^{-\lambda}\lambda^0}{0!}\cdot\frac{e^{-(n-1)\lambda}[(n-1)\lambda]^s}{s!}}{\frac{e^{-n\lambda}[n\lambda]^s}{s!}}. \end{aligned}$$

Note that we have not included any indicators. We have already said that s is in $\{0, 1, 2, \ldots\}$ so all relevant indicators have already evaluated to 1. After simplifying this expression, we have that

$$\mathsf{E}[T|S=s] = \left(\frac{n-1}{n}\right)^s.$$

Thus, we have that

$$T^* = \mathsf{E}[T|S] = \left(\frac{n-1}{n}\right)^S = \left(\frac{n-1}{n}\right)^{\sum_{i=1}^n X_i}.$$

The Rao-Blackwell Theorem guarantees that this is an unbiased estimator of $\tau(\lambda) = e^{-\lambda}$. Furthermore, it has strictly smaller variance than the unbiased estimator $I_{\{X_1=0\}}$ as we observed that equality of variances is achieved if and only if the two estimators are the same.

5.5.4 Complete Statistics

Here is a strange definition.

Definition 5.5.3

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from the distribution with pdf $f(x;\theta)$. We say that a statistic $S = s(\vec{X})$ is **complete** for the distribution if, for any function g such that $\mathsf{E}[g(S)] = 0$ for all θ in the parameter space, we find that $g(S)$ must be 0 with probability 1.

Why might we care about this?

The main consequence of "completeness" is uniqueness of estimators with certain properties. Suppose that we have a statistic S and two functions of S, $g_1(S)$ and $g_2(S)$ that are unbiased for $\tau(\theta)$:

$$\mathsf{E}[g_1(S)] = \tau(\theta) \qquad \text{and} \qquad \mathsf{E}[g_2(S)] = \tau(\theta).$$

Note that

$$0 = \tau(\theta) - \tau(\theta) = \mathsf{E}[g_1(S)] - \mathsf{E}[g_2(S)] = \mathsf{E}[\underbrace{g_1(S) - g_2(S)}_{\substack{\text{call this} \\ g(S)}}].$$

If S is complete, since $\mathsf{E}[g(S)] = 0$ we must have that $g(S) = 0$ with probability 1. Thus, we must have that $g_1(S) = g_2(S)$ with probability 1. So, **there can only be one function of a complete statistic that is unbiased for $\tau(\theta)$!** (We will talk about the "with probability 1" statement in upcoming examples.)

Example 5.5.5

We saw that $S = \sum_{i=1}^{n} X_i$ is sufficient for the Bernoulli distribution, let us also show that is is complete.

Take any function g such that $\mathsf{E}[g(S)] = 0$ for all $0 \leq p \leq 1$.

Since $S \sim Bernoulli(p)$, we can write out this expectation as

$$0 \quad = \quad \mathsf{E}[g(S)] = \sum_{s=0}^{n} g(s) \begin{pmatrix} n \\ s \end{pmatrix} p^s (1-p)^{n-s}$$

for all $0 \leq p \leq 1$.

Note that this is a polynomial in p of degree n (or $< n$ depending on whether or not some of the coefficients involving g are zero). It will have n (or less) zeros. That is, it will have n (or less) values of p that make it

zero. However, we have assumed that it is zero for all p in the interval $[0, 1]$! Thus, it must be identically zero meaning that all of its coefficients must be zero! This forces $g(s) = 0$ for each of $s = 0, 1, \ldots, n$. While g might be non-zero for different values of s, we can conclude that $g(S) = 0$ "with probability 1" since $S = \sum_{i=1}^{n} X_i$ will only take on values in $\{0, 1, \ldots, n\}$. While it may be true that $g(4.2) = 0$, it is not a problem since $P(S = 4.2) = 0$.

Therefore, $S = \sum_{i=1}^{n} X_i$ is complete for the Bernoulli distribution.

Let's try a continuous example.

Example 5.5.6

We saw that $S = X_{(n)}$ is sufficient for the $unif(0, \theta)$ distribution, let us also show that is is complete.

Take any function g such that $\mathsf{E}[g(S)] = 0$ for all $\theta > 0$.

Since $S = X_{(n)}$ has pdf $f_S(s) = \frac{n}{\theta^n} s^{n-1} I_{(0,\theta)}(s)$, we have that

$$0 \quad = \quad \mathsf{E}[g(S)] = \int_0^\theta g(s) \tfrac{n}{\theta} s^{n-1} \, ds.$$

Pulling out n/θ, this implies that

$$\int_0^\theta g(s) s^{n-1} \, ds = 0.$$

Taking the derivative of both sides with respect to θ gives

$$g(\theta)\theta^{n-1} = 0.$$

Since this has to hold for all $\theta > 0$, we can conclude that $g(u) = 0$ for all $u > 0$. Therefore,

$$g(S) = g(X_{(n)}) = 0.$$

Thus, $S = X_{(n)}$ is complete for the $unif(0, \theta)$ distribution.

(Again, note that we have not concluded anything about the value of $g(u)$ for $u < 0$ but it is unimportant since the maximum of $unif(0, \theta)$ random variables cannot take on these values. Since $X_{(n)}$ is greater than 0 with probability 1 we at least have that $g(X_{(n)}) = 0$ with probability 1.)

5.5.5 The UMVUE

UMVUE stands for Uniformly Minimum Variance Unbiased Estimator. In mathematics, the word "uniformly" can be translated as "for all". When estimating $\tau(\theta)$ with some estimator $\widehat{\tau(\theta)}$, the variance of $\widehat{\tau(\theta)}$ will depend on θ which is unknown! If you you were to graph the variances of two unbiased estimators as a function of θ, you might see that one estimator is better (in the sense that is has smaller variance) for certain values of θ while the other is better for other values of θ. The UMVUE for $\tau(\theta)$ is the unbiased estimator that has smaller variance than all other unbiased estimators for all values of θ!

We find the UMVUE for $\tau(\theta)$ using the **Lehmann-Scheffé Theorem**.

The Lehmann-Scheffé Theorem

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with pdf $f(x;\theta)$ and suppose that $S = s(\vec{X})$ is a complete and sufficient statistic for this distribution. If $g(S)$ is some function of S that is unbiased for $\tau(\theta)$ then $g(S)$ is the UMVUE for $\tau(\theta)$.

Proof :

- Let S be complete and sufficient for this model and suppose that we have found a function g such that $\mathsf{E}[g(S)] = \tau(\theta)$. Lehmann and Scheffé claim that this is the unbiased estimator with smallest variance. Let's check this claim.
- Let T be any other unbiased estimator of $\tau(\theta)$. We want to show that $Var[g(S)]$ is smaller than $Var[T]$.
- Case One: T is not a function of S.

  In this case, we can use the Rao-Blackwell Theorem to compute $T^* = \mathsf{E}[T|S]$ to get another unbiased estimator with strictly smaller variance than that of T. Clearly then, T does not have smallest variance and therefore is not the UMVUE.

  (Note that, if Rao-Blackwellization is either going to reduce the variance or give something with the same variance as T. In the proof of the Rao-Blackwell Theorem, we saw that if $Var[T^*] = Var[T]$ then we must have that $T^* = T$ but this can't happen since T^* is a function of S and T was assumed to not be a function of S. This is how we know that T^* has a strictly smaller variance than T.)
- Case Two: T is a function of S.

  Since S is complete, we know there is only one function of S that is unbiased for $\tau(\theta)$. So, T (a function of S) must be equal to $g(S)$.
- In summary, if $g(S)$ is the unbiased estimator for $\tau(\theta)$ proposed by Lehmann and Scheffé and if T is any other unbiased estimator, we have shown that either T does not have smallest variance or $T = g(S)$. We

can't do better than $g(S)$.

■

Example 5.5.7

Let $X_1, X_2, \ldots, X_n$ be a random sample from the Poisson distribution with parameter λ. Find the UMVUE for λ.

By the Lehmann-Scheffé Theorem, we must do two things:

1. Find a complete and sufficient statistic S.
2. Find a function of S that is unbiased for $\tau(\lambda) = \lambda$.

In Example 5.5.4 we used the Factorization Criterion to find a sufficient statistic for this model. It was $S = \sum_{i=1}^{n} X_i$. Let's see if we can show that S is also complete.[1]

In order to show that $S = \sum_{i=1}^{n} X_i$ is also complete, we first note that $S \sim Poisson(n\lambda)$. So, if g is any function such that $\mathsf{E}[g(S)] = 0$ for all $\lambda > 0$, then

$$0 \quad = \quad \mathsf{E}[g(S)] = \sum_{s=0}^{\infty} g(s) \frac{e^{-n\lambda}(n\lambda)^s}{s!} = e^{-n\lambda} \sum_{s=0}^{\infty} g(s) \frac{(n\lambda)^s}{s!}.$$

Thus,

$$\sum_{s=0}^{\infty} \frac{g(s) n^s}{s!} \lambda^s = 0 \qquad \text{for all } \lambda > 0.$$

This is a power series in λ. In order to it to be zero for all $\lambda > 0$, it must be identically zero. That, we must have the coefficients $a_s = g(s)n^s/s!$ equal to zero for all non-negative integers s. This implies that $g(s) = 0$ for $s = 0, 1, 2, \ldots$. So, we have that $g(S) = 0$ with probability 1 since S is Poisson and can only take on values in $\{0, 1, 2, \ldots\}$. Therefore, S is not only sufficient but complete as well!

We have shown that $S = \sum_{i=1}^{n}$ is a complete and sufficient statistic for this model. To get the UMVUE for λ, we must find a function of $S = \sum_{i=1}^{n} X_i$ that is unbiased for λ. Trying S itself gives

$$\mathsf{E}[S] = \mathsf{E}[\sum_{i=1}^{n} X_i] = n\mathsf{E}[X_1] = n\lambda.$$

So, $\frac{1}{n}S = \frac{1}{n}\sum X_i = \overline{X}$ is a function of the complete and sufficient statistic S that is unbiased for λ. By

the Lehmann-Scheffé Theorem, we have that

$$\hat{\lambda} = \overline{X}$$

is the UMVUE for λ.

[1] We have talked about how there are many sufficient statistics for a model, including the vector-valued statistic that is the entire data set. In Section 5.6 we will see a relationship between complete statistics and minimal sufficient statistics. This suggests that we should look at a sufficient statistic with the lowest possible dimension if we are hoping to be able to show completeness!

Example 5.5.8

Let $X_1, X_2, \ldots, X_n$ be a random sample from the Poisson distribution with parameter λ. Find the UMVUE for λ^2.

We already know that $S = \sum_{i=1}^{n} X_i$ is complete and sufficient for the Poisson distribution model. By the Lehmann-Scheffé Theorem, we need to find a function of S that is unbiased for λ^2.

Let's try $\mathsf{E}[S^2]$. Recall that $S \sim Poisson(n\lambda)$ so

$$\mathsf{E}[S^2] = Var[S] + (\mathsf{E}[S])^2 = n\lambda + (n\lambda^2).$$

We know that $\overline{X} = S/n$ is unbiased for λ so we now have that

$$\mathsf{E}[S^2] = n\mathsf{E}[S/n] + n^2\lambda^2 = \mathsf{E}[S] + n^2\lambda^2.$$

Moving things around, we see that we can get λ^2 as the expected value of $(S^2 - S)/n^2$. Since this is a function of the complete and sufficient statistic S that is unbiased for λ^2, it is the UMVUE for λ^2 by the Lehmann-Scheffé. Bringing this back to the X's, we have that

$$\hat{\lambda}_{\text{UMVUE}} = \frac{(\sum X_i)^2 - \sum X_i}{n^2}.$$

Example 5.5.9

Let $X_1, X_2, \ldots, X_n$ be a random sample from the Poisson distribution with parameter λ. Find the UMVUE for $\tau(\lambda) = e^{-\lambda}$.

We know that $S = \sum_{i=1}^{n} X_i$ is complete and sufficient for the Poisson distribution model. By the Lehmann-Scheffé Theorem, we need to find a function of S that is unbiased for λ^2. We did this back in Example 5.5.4 using the Rao-Blackwell Theorem. Recall that the theorem has us starting with any unbiased estimator of $e^{-\lambda}$ and a sufficient statistic and computing an expectation. In the end, we get something that is still unbiased for $e^{-\lambda}$ and is a function of S. While the Rao-Blackwell Theorem didn't need or mention complete statistics, we have what we need for the Lehmann-Scheffé Theorem. We have a function of a complete and sufficient statistic:

$$\widehat{\tau(\lambda)} = \left(\frac{n-1}{n}\right)^{S} = \left(\frac{n-1}{n}\right)^{\sum_{i=1}^{n} X_i}.$$

that is unbiased for $e^{-\lambda}$. Therefore, we have the UMVUE for $e^{-\lambda}$.

5.5.6 Exponential Families: Finding Complete and Sufficient Statistics

We have been fortunate so far that our sufficient statistics found using the Factorization Criterion have also been complete. However, there is a different kind of factorization of the joint pdf from which we can find statistics that are sufficient and complete all in one step!

Definition 5.5.4

$X_1, X_2, \ldots, X_n$ is a random sample from a **one-parameter exponential family** if θ is one-dimensional and if $X_1, X_2, \ldots, X_n$ has a joint pdf that can be written as

$$f(\vec{x}; \theta) = a(\theta) b(\vec{x}) \exp[c(\theta) d(\vec{x})]$$

for some functions $a(\theta)$, $b(\vec{x})$, $c(\theta)$, and $d(\vec{x})$.

As the notation suggests, a and c depend only on θ and not on $\vec{x}$ while b and d depend only on $\vec{x}$ and not on θ.

There are some obvious conditions on the functions $a(\cdot)$, $b(\cdot)$, $c(\cdot)$, and $d(\cdot)$ that are needed to ensure that this is a valid pdf. For example, we can't have a of b identically 0 or we wouldn't have a pdf since it would not integrate to 1. As another example, if a is negative then b must be negative so that the pdf is non-negative.

Here is why we care about one-parameter exponential families.

Theorem 5.5.1

If $X_1, X_2, \ldots, X_n$ is a random sample from a one-parameter exponential family distribution, then $S = d(\vec{X})$ is complete and sufficient for the distribution.

The sufficiency part of this claim follows directly from the Factorization Criterion for sufficiency since

$$f(\vec{x};\theta) = \underbrace{b(\vec{x})}_{h(\vec{x})} \cdot \underbrace{a(\theta)\exp[c(\theta)d(\vec{x})]}_{g(s(\vec{x});\theta)}$$

we can see that $S = s(\vec{X}) = d(\vec{X})$ is sufficient.

The proof of the completeness part of this claim is much more difficult but we will give a very rough sketch of the idea for the discrete case.

Sketch of Proof:

Let Θ denote the parameter space for the model. We give a sketch of the proof for the case where Θ contains an interval.

- Suppose that $f(\vec{x};\theta) = a(\theta)b(\vec{x})\exp[c(\theta)d(\vec{x})]$. Without loss of generality, we assume that both $a(\theta)$ and $b(\vec{x})$ are positive.
- Consider the statistic $S := d(\vec{X})$. Let's find the pdf for S.

$$\begin{aligned} P(S=s) &= P(d(\vec{X})=s) = P(\vec{X} \text{ is such that } d(\vec{X})=s) \\ &= \sum_{\{\vec{x}:d(\vec{x})=s\}} f(\vec{x};\theta) \\ &= a(\theta)\sum_{\{\vec{x}:d(\vec{x})=s\}} b(\vec{x})\exp[c(\theta)d(\vec{x})] \\ &= a(\theta)\exp[c(\theta)\cdot s]\sum_{\{\vec{x}:d(\vec{x})=s\}} b(\vec{x}) \\ &= a(\theta)\, r(s)\exp[c(\theta)\cdot s] \end{aligned}$$

  where $r(s) := \sum_{\{\vec{x}:d(\vec{x})=s\}} b(\vec{x})$.

- Now consider any function g such that $\mathsf{E}[g(S)] = 0$ for all $\theta \in \Theta$. We then have

$$0 = \mathsf{E}[g(S)] = \sum_s g(s)\, a(\theta)\, r(s)\exp[c(\theta)\cdot s]$$

  for all $\theta \in \Theta$.

Factoring out $a(\theta)$, this implies that

$$\sum_s g(s)\, r(s) \exp[c(\theta) \cdot s] = 0$$

for all $\theta \in \Theta$.

- Note that both $r(s)$ and $\exp[c(\theta) \cdot s]$ are strictly positive. If $g(s)$ is not zero for all points in the range of $s = s(\vec{x})$, $g(s)$ will have to take on some positive and negative values over this set in a way such that this sum is zero. This may be possible for one value of θ or even multiple but countable values for θ. However, if the parameter space Θ contains an interval, it is not possible for this sum to be zero without g being zero for all $\theta \in \Theta$. Thus, $g(S) = 0$ with probability 1.

■

This "one-parameter exponential family factorization" will work for most of our "nice known and named distributions". Distributions that are ruled out from the start though are ones which have the unknown parameter in the indicator because we just can't get the full separation of data ("x-stuff") and parameters that we need.

Example 5.5.10

Let $X_1, X_2, \ldots, X_n$ be a random sample from the Bernoulli distribution with parameter p. Find a complete and sufficient statistic for this distribution.

The joint pdf is

$$f(\vec{x}; p) = p^{\sum x_i}(1-p)^{n-\sum x_i} \prod_{i=1}^{n} I_{\{0,1\}}(x_i)$$

which may be rewritten as

$$f(\vec{x}; p) = \underbrace{(1-p)^n}_{\text{"pure parameter"}} \underbrace{\prod_{i=1}^{n} I_{\{0,1\}}(x_i)}_{\text{"pure data"}} \underbrace{\left(\frac{p}{1-p}\right)^{\sum x_i}}_{\text{"mixed part"}}$$

which may again be rewritten as

$$f(\vec{x}; p) = \underbrace{(1-p)^n}_{a(p)} \underbrace{\prod_{i=1}^{n} I_{\{0,1\}}(x_i)}_{b(\vec{x})} \exp\left[\underbrace{\left(\sum x_i\right)}_{d(\vec{x})} \underbrace{\ln\left(\frac{p}{1-p}\right)}_{c(\theta)}\right].$$

So, by the "one-parameter exponential family factorization",

$$S = d(\vec{X}) = \sum_{i=1}^{n} X_i$$

is complete and sufficient for the Bernoulli distribution.

There is also an exponential family factorization for pdfs with k-dimensional parameters where $k \geq 1$.

Definition 5.5.6

$X_1, X_2, \ldots, X_n$ is a random sample from a **k-parameter exponential family** if θ is k-dimensional and if $X_1, X_2, \ldots, X_n$ has a joint pdf that can be written as

$$f(\vec{x};\theta) = a(\theta)b(\vec{x})\exp[\sum_{i=1}^{k} c_i(\theta)d_i(\vec{x})]$$

for some functions $a(\theta)$, $b(\vec{x})$, $c_1(\theta), \ldots, c_k(\theta)$, and $d_1(\vec{x}), \ldots, d_k(\vec{x})$.

Note that this definition does not require c_i to be a function of θ_i (the ith component of the vector θ). $c_i(\theta)$ is a function of the vector θ and could involve one, several, or all components of θ!

As in the one-dimensional θ case, this exponential family factorization gives complete and sufficient statistics.

Theorem 5.5.1

If $X_1, X_2, \ldots, X_n$ is a random sample from a k-parameter exponential family distribution, then $S = (d_1(\vec{X}), d_2(\vec{X}), \ldots d_k(\vec{X}))$ is a vector of complete and sufficient statistics for the distribution.

Example 5.5.11

Let $X_1, X_2, \ldots, X_n$ be a random sample from the $N(\mu, \sigma^2)$ distribution where both parameters are unknown. Find the UMVUEs for μ and σ^2

By the Lehmann-Scheffé Theorem, we must find a (set of) complete and sufficient statistic(s) for this distribution and then find functions of this (these) statistic(s) that are unbiased for μ and σ^2.

The joint pdf for the random sample is

$$f(\vec{x};\mu,\sigma^2) = (2\pi\sigma^2)^{-n/2}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]$$

which may be rewritten as

$$f(\vec{x};\mu,\sigma^2) = (2\pi\sigma^2)^{-n/2}\exp\left[-\tfrac{1}{2\sigma^2}\sum x_i^2 + \tfrac{\mu}{\sigma^2}\sum x_i - n\tfrac{\mu^2}{2\sigma^2}\right]$$

$$= \underbrace{(2\pi\sigma^2)^{-n/2}\exp\left[-n\frac{\mu^2}{2\sigma^2}\right]}_{a(\theta)}\exp\left[\underbrace{-\frac{1}{2\sigma^2}}_{c_1(\theta)}\underbrace{\sum x_i^2}_{d_1(\vec{x})} + \underbrace{\frac{mu}{\sigma^2}}_{c_2(\theta)}\underbrace{\sum x_i}_{d_2(\vec{x})}\right].$$

Here $\theta = (\mu,\sigma^2)$. By the "two-parameter exponential family factorization", $S = (\sum X_i^2, \sum X_i)$ is complete and sufficient for the $N(\mu,\sigma^2)$ distribution.

To get the UMVUEs, we must find functions of S that are unbiased for μ and σ^2. We already know these! They are

$$\hat{\mu}_{\text{UMVUE}} = \frac{\sum X_i}{n} = \overline{X} \qquad \text{and} \qquad \widehat{\sigma^2}_{\text{UMVUE}} = \frac{\sum X_i^2 - (\sum X_i)^2/n}{n-1} = S^2$$

where S^2 is the sample variance, not to be confused with the S for the sufficient statistic.

Note that, in the previous example, we used both of the complete and sufficient statistics to estimate σ^2. In general, you should NOT associate one statistic with one parameter and the other statistic with the other.

5.6 Minimal Sufficient Statistics*

Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} unif(0,\theta)$. In Example 5.5.3 we saw three different sufficient statistics for this model. They were

$$S_1 = (X_1, X_2, \ldots, X_n),$$

$$S_2 = (X_{(1)}, X_{(n)}),$$

and

$$S_3 = X_{(n)}.$$

We talked about how these statistics are sufficient or "enough" to have when estimating θ or functions of θ and, therefore, that we can include additional statistics and still have sufficiency. For example,

$$S_4 = (X_{(1)}, X_{(n)}, \sum_{i=1}^{n} X_i)$$

is sufficient for this model.

A *minimal sufficient statistic* is a sufficient statistic that does not include any extraneous information. Formally, we have the following.

Definition 5.6.1

Let $X_1, X_2, \ldots, X_n$ have joint pdf $f(\vec{x}; \theta)$. A statistic S is **minimal sufficient** for this model if

- S is sufficient, and
- S is a function of every other sufficient statistic for the model.

In the $unif(0, \theta)$ example, the statistic $S_3 = X_{(n)}$ is a function of all of the data S_1 and is a function of S_2. Specifically, in the later case,

$$S_3 = g(X_{(1)}, X_{(n)})$$

where $g(x, y) = y$ or $g(x, y) = \max(x, y)$.

Sometimes people will say that a minimal sufficient statistic is a sufficient statistic of minimal dimension. For the uniform example, we went from an n-dimensional statistic to a two-dimensional statistic, to, ultimately, a one-dimensional statistic $X_{(n)}$. If we are down to one dimension (which isn't always possible), we are down to the lowest possible dimension and therefore $X_{(n)}$ must be minimal sufficient. We will see that $X_{(n)}$ is, in fact, minimal sufficient for this model but that this dimensionality argument does not satisfy the formal definition for minimal sufficiency. As an example, consider a random sample of size 1 from the $N(0, \sigma^2)$ distribution. That is, we have an X_1 with pdf

$$f(x; \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}x^2}.$$

As always, the entire sample which is X_1 in this case is sufficient. We also know that X_1^2 is sufficient by the Factorization Criterion or by one-parameter exponential family. Note, however, that X_1^2 is a function of X_1 but the reverse is not true. We can not completely determine X_1 from X_1^2 as it could be positive or negative. Therefore, by Definition 5.6.1, X_1 is not minimal sufficient even though it is a one-dimensional statistic.

We claim, however, that X_1^2 is minimal sufficient when using a sample of size 1 from the $N(0, \sigma^2)$ distribution. This can be shown using *Bahadur's Theorem*.

Bahadur's Theorem

Is S is complete and sufficient then S is minimal sufficient.

Proof :

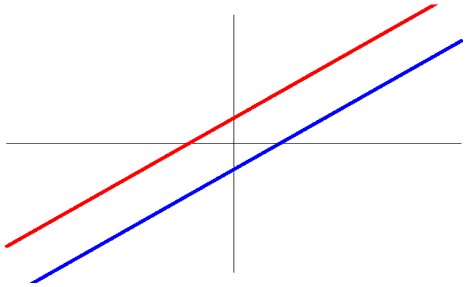
We prove Bahadur's Theorem under the assumption that a minimal sufficient statistic exists. This is almost always the case except for pathological examples [3].

- Let S be a complete and sufficient statistic for the model under consideration. We wish to show that S is minimal sufficient and to do this we need to show that it is a function of any other sufficient statistic.
- Let M be a minimal sufficient statistic for the model under consideration. By definition, M is a function of S.
- Consider the random quantity $g(S) := S - \mathsf{E}[S|M]$. Note that this is a function of S since $\mathsf{E}[S|M]$ is a function of M but M is a function of S. Thus, $g(S) := S - \mathsf{E}[S|M]$ is a function of S.
  (Recall from Section 5.5.2 that $\mathsf{E}[X|Y]$ is a function of Y.)
- Let's take the expected value of $g(S)$:
$$\mathsf{E}[g(S)] = \mathsf{E}[S - \mathsf{E}[S|M]] = \mathsf{E}[S] - \mathsf{E}[\mathsf{E}[S|M]] = \mathsf{E}[S] - \mathsf{E}[S] = 0$$
  Since S is complete, this implies that $g(S) = 0$ with probability 1, i.e., $S - \mathsf{E}[S|M] = 0$ with probability 1.
- We now have that $S = \mathsf{E}[S|M]$. The right-hand side here is a function of M so S is a function of M.
  Let U be any sufficient statistic. By definition of M as minimal sufficient, M is a function of U. Thus, S, which is a function of M is also a function of U. This shows that S is minimal sufficient!

■

5.6.1 Sufficient Statistics as Partitions*

Let X_1, X_2 be a random sample of size 2 from a distribution that has support on the entire real line. Consider the statistic $S_1 = X_1 + X_2$. When S_1 is observed as, for example, $s_1 = 1$, this can be visualized as the red line in $\mathbb{R}^2$ as depicted. When $s_1 = -1$, we get the blue line.



As we run over all $s_1 \in \mathbb{R}$, we will cover the entire plane with none of the lines intersecting. This is a way to *partition* the plane into an uncountable number of disjoint sets.

Consider now the two-dimensional statistic $S_2 = (X_1 + X_2, X_{(2)})$ where $X_{(2)} = \max(X_1, X_2)$. Suppose that the statistic is observed to be $s = (1, 0.6)$. This means that either $x_1 = 0.6$ and $x_2 = 0.4$ or that $x_1 = 0.4$ and $x_2 = 0.6$. That is, the observation of $s_2 = (1, 0.6)$ corresponds to two points in the plane. This single observed value of S_2 corresponds to two points in the plane. As we run over all $s_2 \in \mathbb{R}^2$, we will again partition the plane, this time into an uncountable collection of sets each consisting of 1 or 2 points. Note that this is a *finer* partition than the one given by S_1. The partition given by S_1 is *coarser*.

Let $X_1, X_2, \ldots, X_n$ have joint pdf $f(\vec{x}; \theta)$. A minimal sufficient statistic for this model will give the coarsest (if it exists) partition of $\mathbb{R}^n$ among all sufficient statistics.

Finding a minimal sufficient statistic with an exponential family distribution is easy because it gives us a complete and sufficient statistic. We then Bahadur's Theorem to conclude that it is minimal sufficient. (Again, this all requires the existence of a minimal sufficient statistic but they almost always exist.) If a distribution is not an exponential family distribution, we might use the Factorization Criterion to get a sufficient statistic, show it is complete by the definition of completeness, and then use Bahadur's Theorem. If we feel uncomfortable with the statement "they almost always exist" and do not want to use Bahadur's Theorem, there is another option.

Theorem 5.6.1

Suppose that $X_1, X_2, \ldots, X_n$ have joint pdf $f(\vec{x}; \theta)$.

If a statistic $T = t(\vec{X})$ has the following property:

$$\frac{f(\vec{x}; \theta)}{f(\vec{y}; \theta)} \text{ is } \theta\text{-free} \quad \Leftrightarrow \quad t(\vec{x}) = t(\vec{y})$$

then T is a minimal sufficient statistic.

Proof :

Assume that $T = t(\vec{X})$ is a statistic with the property

$$\frac{f(\vec{x}; \theta)}{f(\vec{y}; \theta)} \text{ is } \theta\text{-free} \quad \Leftrightarrow \quad t(\vec{x}) = t(\vec{y}).$$

For the rest of this proof, we will call this "the property".

Sufficiency

- Partition the support of $f(\vec{x};\theta)$ using $t(\vec{x})$. Let

$$A_t = \{\vec{x} : t(\vec{x}) = t\}.$$

- Use the axiom of choice to select and fix one representative point $\vec{x}_t$ from each A_t.

- Note that

$$\vec{x} \in A_t \quad \Leftrightarrow \quad t(\vec{x}) = t \quad \Leftrightarrow \quad \vec{x}_{t(\vec{x})} \in A_t$$

  That is, for any $\vec{x}$ in the support of $f(\vec{x};\theta)$, $\vec{x}_{t(\vec{x})}$ is in the same partition set as $\vec{x}$.

- Since $\vec{x}_{t(\vec{x})}$ is in the same partition set as $\vec{x}$, they both evaluate to t under the function $t(\cdot)$. That is,

$$t(\vec{x}) = t = t(\vec{x}_{t(\vec{x})}).$$

  Thus, by "the property" we know that $f(\vec{x};\theta) = f(\vec{x}_{t(\vec{x})};\theta)$ is θ-free. Call this ratio $h(\vec{x})$:

$$\frac{f(\vec{x};\theta)}{f(\vec{x}_{t(\vec{x})};\theta)} = h(\vec{x}).$$

- We now have

$$f(\vec{x};\theta) = h(\vec{x})\, f(\vec{x}_{t(\vec{x})};\theta).$$

  Note that a given t leads to an associated set A_t which leads to the representative point $\vec{x}_t = \vec{x}_{t(\vec{x})}$. This means that $\vec{x}_{t(\vec{x})}$ is a function of $t = t(\vec{x})$.

- So, we have the joint pdf represented as a function of $\vec{x}$ times a function of $t(\vec{x})$. By the Factorization Criterion, we have that $T = t(\vec{X})$ is sufficient for the model.

Minimality

- Let $S = s(\vec{X})$ be any other sufficient statistic. We need to show that $t(\vec{x})$ is a function of $s(\vec{x})$. To show this, we need to show that

$$s(\vec{x}) = s(\vec{y}) \qquad \Rightarrow \qquad t(\vec{x}) = t(\vec{y}).$$

- Suppose that $s(\vec{x}) = s(\vec{y})$. Since S is sufficient, we can write

$$f(\vec{x};\theta) = h(\vec{x})\, g(s(\vec{x});\theta)$$

  for some functions h and g.

- Now,

$$\frac{f(\vec{x};\theta)}{f(\vec{y};\theta)} = \frac{h(\vec{x})\, g(s(\vec{x});\theta)}{h(\vec{x})\, g(s(\vec{y});\theta)} = \frac{h(\vec{x})}{h(\vec{y})}$$

  since $s(\vec{x}) = s(\vec{y})$. Thus, we have that $f(\vec{x};\theta)/f(\vec{y};\theta)$ is θ-free.

- By "the property" we must have $t(\vec{x}) = t(\vec{y})$. In other words, starting with $s(\vec{x}) = s(\vec{y})$ produced the result $t(\vec{x}) = t(\vec{y})$ and so we have shown that $t(\vec{x})$ is a function of $s(\vec{x})$.

  Since $S = s(\vec{X})$ was an arbitrary sufficient statistic, we have that T is minimal sufficient, as desired.

■

5.7 Ancillary Statistics and Basu's Theorem

Suppose that $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ with σ^2 known. Let S^2 be the sample variance. We know that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

Since σ^2 is known, the left-hand side is just a function of the data and is therefore a statistic. Note that the distribution of this statistic does not depend on any unknown parameters.

Definition 5.7.1

A statistic whose distribution does not depend on any unknown parameters is said to be an **ancillary statistic**.

An ancillary statistic is, in some sense, the opposite of a sufficient statistic since a sufficient statistic contains all of the information about the parameter(s) of a model while an ancillary statistic contains no information about the parameter(s). We will soon seen that ancillary statistics can be useful for arguing the independence of some random quantities. Thus, it would be nice if we had a quick way to identify some ancillary statistics.

5.7.1 Location Parameters and Location-Invariant Statistics

Definition 5.7.2

θ is said to be a **location parameter** for a distribution if the pdf $f(x;\theta)$ can be written as

$$f(x;\theta) = f_0(x-\theta)$$

for some function f_0 which, as the notation suggests, does not depend on θ.

As an example, μ is a location parameter for the $N(\mu, \sigma^2)$ distribution.

Definition 5.7.3

A statistic $T = t(\vec{X})$ is a **location-invariant statistic** if

$$t(x_1 + c, x_2 + c, \ldots, x_n + c) = t(x_1, x_2, \ldots, x_n)$$

for all $c \in \mathbb{R}$.

Example 5.7.1

Let $X_1, X_2, \ldots, X_n$ be a random sample from any distribution. Let $Y_i = X_i + c$ for $i = 1, 2, \ldots, n$. Then

$$\overline{Y} = \overline{X} + c.$$

The sample variance for the Y's is equal to the sample variance of the X's since

$$S^2 = \frac{\sum_{i=1}^n (Y_i - \overline{Y})^2}{n-1} = \frac{\sum_{i=1}^n [(X_i + c) - (\overline{X} - c)]^2}{n-1} = \frac{\sum_{i=1}^n (X_i - \overline{X})}{n-1}.$$

Thus, the sample variance is a location-invariant statistic!

Theorem 5.7.1

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with pdf $f(x;\theta)$.

If θ is a location parameter and T is a location-invariant statistic, then T is ancillary.

Proof : Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with pdf $f(x;\theta)$ where θ is a location parameter:

$$f(x;\theta) = f_0(x-\theta)$$

for some f_0.

- Consider the transformation $Y_i = X_i - \theta$ for $i = 1, 2, \ldots, n$.

  We will find the joint pdf of $Y_1, Y_2, \ldots, Y_n$ using the Jacobian method.

- We have

$$\begin{aligned} y_1 &= g_1(x_1, x_2, \ldots, x_n) = x_1 - \theta \\ y_2 &= g_2(x_1, x_2, \ldots, x_n) = x_2 - \theta \\ &\vdots \\ y_n &= g_n(x_1, x_2, \ldots, x_n) = x_n - \theta, \end{aligned}$$

  and so

$$\begin{aligned} x_1 &= g_1^{-1}(y_1, y_2, \ldots, y_n) = y_1 + \theta \\ x_2 &= g_2^{-1}(y_1, y_2, \ldots, y_n) = y_2 + \theta \\ &\vdots \\ x_n &= g_n^{-1}(y_1, y_2, \ldots, y_n) = y_n + \theta. \end{aligned}$$

  (As a reminder, g_i^{-1} is not the inverse of the g_i function but when we solve the system for x_i as a function of $y_1, y_2, \ldots, y_n$, we will call that function g_i^{-1}.)

- The Jacobian of this transformation is

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_2}{\partial y_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_n} \end{vmatrix} = 1$$

because the matrix is the identity matrix.

- The joint pdf for $Y_1, Y_2, \ldots, Y_n$ is given by

$$f_{Y_1,Y_2,\ldots,Y_n}(y_1, y_2, \ldots, y_n)$$

- The joint pdf for $Y_1, Y_2, \ldots, Y_n$ is

$$\begin{aligned} f_{Y_1,Y_2,\ldots,Y_n}(y_1, y_2, \ldots, y_n) &= f_{X_1,X_2,\ldots,X_n}(g_1^{-1}(\vec{y}), g_2^1(\vec{y}), \ldots, g_n^{-1}(\vec{y})) \cdot |J| \\ &\overset{iid}{=} f(g_1^{-1}(\vec{y}); \theta)\, f(g_2^{-1}(\vec{y}); \theta) \cdots f(g_n^{-1}(\vec{y}); \theta) \cdot |J| \\ &= f(y_1 + \theta; \theta) f(y_2 + \theta; \theta) \cdots f(y_n + \theta; \theta) \cdot |1| \\ &= f_0(y_1) f_0(y_2) \cdots f_0(y_n). \end{aligned}$$

We see that the joint distribution of $Y_1, Y_2, \ldots, Y_n$ does not depend on θ.

- Let $T = t(\vec{X})$ be a location-invariant statistic.We have

$$T = t(X_1, X_2, \ldots, X_n) = t(Y_1 + \theta, Y_2 + \theta, \ldots, Y_n + \theta) = t(Y_1, Y_2, \ldots, Y_n).$$

The last equality is because T is a location-invariant statistic.

Since $T = t(Y_1, Y_2, \ldots, Y_n)$ and the distribution of the Y's does not depend on θ, we see that the distribution of T does not depend on θ. Thus, T is ancillary, as desired.

■

Example 5.7.2

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from the shifted exponential distribution with rate 1 with pdf

$$f(x; \theta) = e^{-(x-\theta)}\, I_{(\theta,\infty)}(x).$$

Clearly θ is a location parameter as we have

$$f(x; \theta) = f_0(x - \theta)$$

where $f_0(x) = e^{-x}\, I_{(0,\infty)}(x)$. (Note that $I_{(0,\infty)}(x - \theta) = I_{(\theta,\infty)}(x)$.)

The sample variance

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{(X)})^2}{n-1}$$

is location-invariant and therefore ancillary for this location parameter model. As another example, the **sample range**

$$X_{(n)} - X_{(1)}$$

is also location invariant (since the distance between the maximum and minimum is unchanged when all data is shifted by some amount) and is therefore ancillary for this example as well.

5.7.2 Scale Parameters and Scale-Invariant Statistics

Definition 5.7.5

θ is said to be a **scale parameter** for a distribution if the pdf $f(x;\theta)$ can be written as

$$f(x;\theta) = \frac{1}{\theta} f_0(x/\theta)$$

for some function f_0 which, as the notation suggests, does not depend on θ.

θ is an **inverse scale parameter** if the pdf $f(x;\theta)$ can be written as

$$f(x;\theta) = \theta f_0(\theta x).$$

As an example, λ is an inverse scale parameter for the $exp(rate = \lambda)$ distribution. As

$$f(x;\lambda) = \lambda e^{-\lambda x} I_{(0,\infty)}(x)$$

can be written as

$$f(x;\lambda) = \lambda f_0(\lambda x)$$

where f_0 is the pdf of the $exp(rate = 1)$ distribution.

Definition 5.7.6

A statistic $T = t(\vec{X})$ is a **scale-invariant statistic** if

$$t(cx_1, cx_2, \ldots, cx_n) = t(x_1, x_2, \ldots, x_n)$$

for all $c > 0$.

Theorem 5.7.1

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with pdf $f(x;\theta)$. If θ is a scale (or inverse scale) parameter and T is a scale-invariant statistic, then T is ancillary.

The proof of Theorem 5.7.7 is very similar to the proof of Theorem 5.7.4 and we leave it as an exercise for the reader.

Example 5.7.3

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from the $\Gamma(2, \beta)$ distribution. Note that β is an inverse scale parameter since

$$\begin{aligned} f(x;\beta) &= \beta^2 x e^{-\beta x} I_{(0,\infty)}(x) \\ &= \beta(\beta x) e^{-\beta x} I_{(0,\infty)}(x) \\ &= \beta(\beta x) e^{-\beta x} I_{(0,\infty)}(\beta x) \quad \text{(since } \beta > 0\text{)} \\ &= \beta f_0(\beta x) \end{aligned}$$

where f_0 is the pdf for the $\Gamma(2, 1)$ distribution.

Consider the statistic

$$T = \frac{\overline{X}}{X_{(1)}}.$$

This statistic is scale-invariant because if we multiply all of the X_i by a constant, the new sample mean is the original sample mean multiplied by the constant. Similarly, the new minimum is the old minimum multiplied by the positive constant. Specifically, if we define $Y_i = cX_i$ for $i = 1, 2, \ldots, n$ and some

positive constant c then

$$\frac{\overline{Y}}{Y_{(1)}} = \frac{c\overline{X}}{cX_{(1)}} = \frac{\overline{X}}{X_{(1)}}.$$

By Theorem 5.7.7, we have that T is ancillary. This means that the distribution of T does not depend on the unknown parameter β.

5.7.3 Basu's Theorem

So, what can we do with these ancillary statistics? As previously mentioned, an ancillary statistic is, in some sense, the opposite of a sufficient statistic since a sufficient statistic contains all of the information about the parameter(s) of a model while an ancillary statistic contains no information about the parameter(s). Indeed, we have two interesting theorems.

Theorem 5.7.1

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with pdf $f(x;\theta)$. Suppose that S is sufficient and T is any statistic that is independent of S. Then T is ancillary.

Proof : Suppose that S is sufficient and T is any statistic that is independent of S.

- Since S is sufficient, the conditional joint distribution of the X_i given $S = s$:

  $$f_{\vec{X}|S}(\vec{x}|s;\theta)$$

  is θ-free.
- In condensing the information in $(X_1, X_2, \ldots, X_n)$ into a statistic $T = t(X_1, X_2, \ldots, X_n)$ (possibly vector-valued and made up of multiple statistics), we are not going to gain any information about θ. Therefore, the conditional distribution of T given S:

  $$f_{T|S}(t|s)$$

  is θ-free.
- Since T is independent of S, we have that

  $$f_{T|S}(t|s) = f_T(t).$$

  Thus, $f_T(t)$ is θ-free and T is ancillary.

Being able to establish independence of certain statistical quantities is immensely important in statistical computations. It would be amazing if we could reverse Theorem 5.7.8 to say that sufficient statistics are independent of ancillary statistics? We can do just that if we throw in completeness.

Basu's Theorem

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with pdf $f(x;\theta)$.

If S is complete and sufficient and T is ancillary, then S and T are independent.

Proof :

- Consider the following function of the random variable S:

$$g(S) = f_T(t) - f_{T|S}(t|S).$$

Here, $f_T(t)$ is the marginal pdf of T and $f_{T|S}(t|s)$ is the conditional pdf for T given S. Note that the conditional pdf is being evaluated at a random variable and $f_{T|S}(t|S)$ is, itself, a random variable.

- We take the expectation of this function. Note that since S is the only random quantity, we take the expectation with respect to the distribution of S.

$$\begin{aligned}
\mathsf{E}[g(S)] &= \mathsf{E}[f_T(t)] - \mathsf{E}[f_{T|S}(t|S)] \\
&= f_T(t) - \mathsf{E}[f_{T|S}(t|S)] \quad (f_T(t) \text{ is not random}) \\
&= f_T(t) - \int_{-\infty}^{\infty} f_{T|S}(t|s)\, f_S(s)\, ds \\
&= f_T(t) - \int_{-\infty}^{\infty} \frac{f_{T,S}(t,s)}{f_S(s)}\, f_S(s)\, ds \\
&= f_T(t) - \int_{-\infty}^{\infty} f_{T,S}(t,s)\, ds \\
&= f_T(t) - f_T(t) = 0
\end{aligned}$$

- Since S is complete and $\mathsf{E}[g(S)] = 0$, we must have $g(S) = 0$ with probability 1. This means that

$$f_T(t) = f_{T|S}(t|S)$$

with probability 1.

In other words,

$$f_{T|S}(t|s) = f_T(t)$$

for all possible values of S. This implies that S and T are independent!

Example 5.7.4

Suppose that

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$$

where $\sigma^2 > 0$ is known.

Show that the sample $\overline{X}$ is independent of the sample range $X_{(n)} - X_{(1)}$.

Intuitively, these should be independent since the sample mean is a measure of location for the data set. You can shift the entire data set to the left or right and the sample mean will shift right along with it. The sample range, however, will remain unchanged.

The pdf is

$$f(x;\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{\sigma^2}(x-\mu)^2}.$$

The joint pdf is

$$\begin{aligned} f(\vec{x};\mu) &= (2\pi\sigma^2)^{-n/2} \exp[-\tfrac{1}{2\sigma^2}\textstyle\sum_{i=1}^n (x_i-\mu)^2 \\ &= (2\pi\sigma^2)^{-n/2} \exp\left[-\tfrac{1}{2\sigma^2}\left(\textstyle\sum_{i=1}^n X_i^2 - 2\mu\sum_{i=1}^n X_i + n\mu^2\right)\right] \\ &= (2\pi\sigma^2)^{-n/2} \exp\left[-\tfrac{n\mu^2}{\sigma^2}\right] \cdot \exp\left[-\tfrac{1}{2\sigma^2}\textstyle\sum_{i=1}^n X_i^2\right] \cdot \left[-\tfrac{\mu}{\sigma^2}\textstyle\sum_{i=1} nX_i\right]. \end{aligned}$$

Remember, σ^2 is known and just a number. We have a one-parameter exponential family factorization with

$$a(\mu) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{n\mu^2}{\sigma^2}\right],$$

$$b(\vec{x}) = \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n X_i^2\right],$$

$$c(\mu) = -\mu/\sigma^2,$$

and

$$d(\vec{x}) = \sum_{i=1}^n x_i.$$

(Note that the $(2\pi\sigma^2)^{-n/2}$ could have been made part of $b(\vec{x})$ instead of $a(\mu)$.)

By one-parameter exponential family factorization, we have that

$$S = d(\vec{X}) = \sum_{i=1}^{n} X_i$$

is complete and sufficient for this model.

μ is a location parameter since

$$f(x;\mu) = f_0(x-\mu)$$

where $f_0(x)$ is the pdf of the $N(0,\sigma^2)$ distribution.

The sample range is a location invariant statistic since adding or subtracting a constant from all data points will not change the distance between the maximum and minimum values. So, by Theorem 5.7.4, we know that $X_{(n)} - X_{(1)}$ is ancillary.

By Basu's Theorem, we know that the complete and sufficient statistic $\sum_{i=1}^{n} X_i$ is independent of an ancillary statistic $X_{(n)} - X_{(1)}$. Putting a $1/n$ in front of the sum is not going to induce dependence. If we prefer a more concise statement, we may write the exponential family factorization so that $d(\vec{X}) = \overline{X}$ is the complete and sufficient statistic.[1] So, we have shown that $\overline{X}$ is independent of the sample range $X_{(n)} - X_{(1)}$.

[1] A complete and sufficient statistic is unique up to one-to-one transformations.

5.8 Postscript: The Multi-Dimensional Cramér-Rao Lower Bound

So far we only have the Cramér-Rao lower bound for one-dimensional θ or $\tau(\theta)$. It's time to "kick it up a notch"!

5.8.1 Higher Dimensional Expectation and Variance

For a random vector $\vec{X} = (X_1, X_2, \ldots, X_n)^T$, of not necessarily iid random variables, where $\mathsf{E}[X_i] = \mu_i$, the expectation is defined componentwise:

$$\vec{\mu} = \mathsf{E}[\vec{X}] = \mathsf{E}\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} \mathsf{E}[X_1] \\ \mathsf{E}[X_2] \\ \vdots \\ \mathsf{E}[X_n] \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}.$$

The expected value of a matrix of random variables is also defined componentwise.

The **variance** of $\vec{X}$ is defined, analogous to the one-dimensional case, as

$$Var[\vec{X}] = \mathsf{E}[(\vec{X} - \vec{\mu})(\vec{X} - \vec{\mu})^T].$$

Note that this is an $n \times n$ matrix whose (i, j)-th entry is

$$\mathsf{E}[(X_i - \mu_i)(X_j - \mu_j)] = Cov(X_i, X_j).$$

The entries on the diagonal are the variances of the individual X_i.

This matrix is called a **variance-covariance matrix** of $\vec{X}$, or, sometimes more simply the **covariance matrix**, or even the variance of $\vec{X}$.

Example 5.8.1

Suppose that $X_1, X_2, \ldots, X_n$ are iid with common variance σ^2. Since they are independent, $Cov(X_i, X_j) = 0$ for $i \neq j$ and the variance-covariance matrix of $\vec{X} = (X_1, X_2, \ldots, X_n)^T$ is σ^2 times the $n \times n$ identity matrix.

It is routine to show that $Var[\vec{X}]$ can also be written as

$$Var[\vec{X}] = \mathsf{E}[\vec{X}\vec{X}^T] - \vec{\mu}\vec{\mu}^T$$

analogous to the similar formula for one-dimensional variance.

As another analogy, recall that "constants come out of variances squared". For an $n \times 1$ random vector $\vec{X}$ and an $m \times n$ matrix of constants A, we can show that

$$Var[A\vec{X}] = A\, Var[\vec{X}]\, A^T.$$

A fun (and useful!) fact about variance-covariance matrices is that they are all **non-negative definite**. This means that if $V = Var[\vec{X}]$, then $\vec{a}^T V \vec{a} \geq 0$ for any $n \times 1$ vector $\vec{a}$. The reason this must be true is because the variance of a one-dimension random variable such as $\sum_{i=1}^n a_i X_i$ must be non-negative. So, we have that

$$0 \leq Var[\sum_{i=1}^n a_i X_i] = Cov(\sum_{i=1}^n a_i X_i, \sum_{j=1}^n a_j X_j) = \sum_{i=1}^n \sum_{j=1}^n a_i a_j Cov(X_i, X_j) = \sum_{i=1}^n \sum_{j=1}^n a_i a_j v_{ij} = \vec{a}^T V \vec{a}.$$

In the one-dimensional case, if the variance σ^2 is equal to zero, we have a degenerate case where our random variable is actually a constant. Unless we mean to talk about such a case, we generally assume that $\sigma^2 > 0$. In

higher dimensions, we would like to have $\vec{a}^T V \vec{a}$ strictly greater than zero for any $n \times 1$ vector $\vec{a} \neq \vec{0}$. In this case we say that V is **positive definite**. A positive definite matrix has the benefit of being invertible and having a strictly positive determinant.

5.8.2 The Multivariate Normal Distribution

The multivariate normal distribution is extremely important in statistics so it may seem odd that we would put it in one of these "optional reading postscript sections" and also spend so little time on it. The reason for this is that we will not need the distribution for anything in this text outside of this postscript section.

A random vector $\vec{X} = (X_1, X_2, \ldots, X_n)^T$ is said to have a multivariate normal distribution if the joint pdf for the X_i is

$$f(\vec{x}; \mu, V) = (2\pi)^n |V|^{-n/2} \exp\left[-\frac{1}{2}(\vec{x} - \vec{\mu})^T V^{-1} (\vec{x} - \vec{\mu})\right] \tag{5.8.11}$$

for some $n \times 1$ vector μ and a positive definite matrix V.

We write $\vec{X} \sim MVN(\vec{\mu}, V)$, or sometimes $\vec{X} \sim MNV_n(\vec{\mu}, V)$, or sometimes even $\vec{X} \sim N(\vec{\mu}, V)$.

For $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, we have that

$$\vec{X} \sim MVN(\vec{\mu}, V)$$

where $\vec{\mu} = (\mu, \mu, \ldots, \mu)^T$ and $V = \sigma^2 I$ where I is the $n \times n$ identity matrix.

In Section 0.10 of Chapter 0 we saw that if X_i and X_j are independent then $Cov(X_i, X_j) = 0$. We also saw an example that showed that the reverse is not necessarily true. However, if $X_1, X_2, \ldots, X_n$ are normal random variables with $Cov(X_i, X_j) = 0$ for all pairs $\{(i, j) : 1 \leq i, j \leq n, i \neq j\}$, then $V = Var[\vec{X}]$ is a diagonal matrix (zeros off the diagonal) and the pdf in (5.8.11) factors, revealing that $X_1, X_2, \ldots, X_n$ are independent.

In general,

$$Cov(X_i, X_j) = 0 \qquad \Rightarrow \qquad X_i \text{ and } X_j \text{ are independent.}$$

In general,

$$X_i \text{ and } X_j \text{ are independent} \qquad \not\Rightarrow \qquad Cov(X_i, X_j) = 0.$$

If X_i and X_j are normal random variables,

$$Cov(X_i, X_j) = 0 \qquad \Leftrightarrow \qquad X_i \text{ and } X_j \text{ are independent.}$$

We are now ready to consider the multi-dimensional Cramér-Rao lower bound.

5.8.3 The Multi-Dimensional Dimensional CRLB

Suppose that $X_1, X_2, \ldots, X_n$ are one-dimensional random variables with joint pdf $f(\vec{x}; \theta)$. Let $\vec{X} = (X_1, X_2, \ldots, X_n)^T$. Then $\vec{X}$ is a **random vector**.

Suppose that θ is a $p \times 1$ vector of parameters.[9]

Suppose that we wish to estimate some

$$\tau(\theta) = (\tau_1(\theta), \tau_2(\theta), \ldots, \tau_q(\theta))^T.$$

To summarize,

- $\vec{X}$ is $n \times 1$
- θ is $p \times 1$
- $\tau(\theta) = (\tau_1(\theta), \tau_2(\theta), \ldots, \tau_q(\theta))^T$ is $q \times 1$
- $T = t(\vec{X}) = (t_1(\vec{X}), t_2(\vec{X}), \ldots, t_q(\vec{X}))^T$ is an unbiased estimator of $\tau(\theta)$.

Note that the expectation of a vector-valued random variable is taken componentwise, so this last bullet is saying that

$$\mathsf{E}[T] = \mathsf{E}\begin{bmatrix} t_1(\vec{X}) \\ t_2(\vec{X}) \\ \vdots \\ t_q(\vec{X}) \end{bmatrix} = \begin{bmatrix} \mathsf{E}[t_1(\vec{X})] \\ \mathsf{E}[t_2(\vec{X})] \\ \vdots \\ \mathsf{E}[t_q(\vec{X})] \end{bmatrix} = \begin{bmatrix} \tau_1(\theta) \\ \tau_2(\theta) \\ \vdots \\ \tau_q(\theta) \end{bmatrix} = \tau(\theta).$$

Under certain "regularity conditions" (to be explored in the upcoming proof), we have a matrix-vector version of the **Cramér-Rao lower bound**.

$$\boxed{Var[T] \geq \left(\frac{\partial \tau}{\partial \theta}\right) [I_n(\theta)]^{-1} \left(\frac{\partial \tau}{\partial \theta}\right)^T} \tag{5.8.12}$$

Here,

- $Var[T]$ is a $q \times q$ variance-covariance matrix,

[9] We apologize for the consistent inconsistency. We have always used x for a scalar variable and $\vec{x}$ for a vector. On the other hand, we have not been writing $\vec{\theta}$ even if θ is vector-valued. This was an intentional decision for "reasons".

- $\frac{\partial \tau}{\partial \theta}$ is a $q \times p$ matrix of partial derivatives

$$\frac{\partial \tau}{\partial \theta} = \begin{bmatrix} \frac{\partial \tau_1}{\partial \theta_1} & \frac{\partial \tau_1}{\partial \theta_2} & \cdots & \frac{\partial \tau_1}{\partial \theta_p} \\ \frac{\partial \tau_2}{\partial \theta_1} & \frac{\partial \tau_2}{\partial \theta_2} & \cdots & \frac{\partial \tau_2}{\partial \theta_p} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial \tau_q}{\partial \theta_1} & \frac{\partial \tau_q}{\partial \theta_2} & \cdots & \frac{\partial \tau_q}{\partial \theta_p} \end{bmatrix},$$

  and

- $I_n(\theta)$ is a $p \times p$ symmetric **Fisher information matrix** with $(i,j)th$ entry equal to

$$\mathsf{E}\left[\left(\frac{\partial}{\partial \theta_i} \ln f(\vec{X};\theta)\right)\left(\frac{\partial}{\partial \theta_j} \ln f(\vec{X};\theta)\right)\right] = -\mathsf{E}\left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln f(\vec{X};\theta)\right].$$

  Note that if $p = 1$, this matrix is a scalar and equal to the Fisher information as defined for a one-dimensional θ in Section 5.3.

The inequality in 5.8.12 is not componentwise inequality. For same-sized square matrices A and B, we are using the notation

$$A \geq B$$

to mean that the matrix $A - B$ is non-negative definite. (Recall that this means that the scalar $\vec{x}^T(A-B)\vec{x} \geq 0$ for every appropriately sized vector $\vec{x}$.) One property of non-negative definite matrices as that the diagonal entries are all positive and, indeed, one can show that the diagonal entries of the CRLB matrix are in fact lower bounds of the variances of the components of $t(\vec{X})$. The off diagonal entries, while lower bounds for covariance terms, are not particularly interesting until we use the CRLB as an asymptotic variance-covariance matrix for MLEs.

5.8.4 Proof of the CRLB

We will now prove the multi-dimensional CRLB result given in (5.8.12).

- Recall the one-dimensional CRLB proof where we showed that

$$\tau'(\theta) = \mathsf{E}\left[\Big(t(\vec{X}) - \tau(\theta)\Big)\left(\frac{\partial}{\partial \theta} \ln f(\vec{X};\theta)\right)\right]. \tag{5.8.13}$$

  and used $S(\vec{X},\theta)$ to denote the score function:

$$S(\vec{X},\theta) = \frac{\partial}{\partial \theta} \ln f(\vec{X};\theta).$$

  Since $\mathsf{E}[S(\vec{X},\theta)] = 0$, (5.8.13) can be rewritten as

$$\tau'(\theta) = Cov(t(\vec{X}) - \tau(\theta), S(\vec{X}, \theta)).$$

Since $\tau(\theta)$ is a constant, we can further simplify this to be

$$\tau'(\theta) = Cov(t(\vec{X}), S(\vec{X}, \theta)).$$

We can use the same procedure that we used for showing (5.3.1) in the one-dimensional case to show that

$$\frac{\partial}{\partial \theta_i}\tau_j(\theta) = \mathsf{E}\left[\Big(t_j(\vec{X}) - \tau(\theta)\Big)\left(\frac{\partial}{\partial \theta_i}\ln f(\vec{X};\theta)\right)\right].$$

With our new observations about writing this as a covariance, we have

$$\frac{\partial}{\partial \theta_i}\tau_j(\theta) = Cov(t_j(\vec{X}), S_i(\vec{X}, \theta)).$$

where

$$S_i(\vec{X}, \theta) = \frac{\partial}{\partial \theta_i}\ln f(\vec{X};\theta).$$

- Let $S(\vec{X}, \theta) = (S_1(\vec{X}, \theta), S_2(\vec{X}, \theta), \ldots, S_p(\vec{X}), \theta)^T$.

  Let A be a $q \times p$ matrix of constants.

  Since variance-covariance matrices are all non-negative definite, we have that

$$\underbrace{Var[\underbrace{t(\vec{X} - AS(\vec{X}, \theta))}_{q\times 1}]}_{q\times q} \geq 0.$$

  (As in the statement of the CRLB, inequality here means that the matrix is non-negative definite. It does not mean that all elements are non-negative.)

- Recall that, for random variables X and Y,

$$Var[X - Y] = Var[X] + Var[Y] - 2Cov(X, Y).$$

  The higher-dimensional analogue here is

$$\begin{aligned} 0 &\leq Var[t(\vec{X} - AS(\vec{X}, \theta))] \\ &= Var[t(\vec{X})] + Var[AS(\vec{X}, \theta)] - 2Cov(t(\vec{X}), AS(\vec{X}, \theta)) \\ &= Var[T] + AVar[S(\vec{X}, \theta)]A^T - 2Cov(t(\vec{X}), AS(\vec{X}, \theta)). \end{aligned}$$

  Note that $Var[S(\vec{X}, \theta)]$ is a $p \times p$ matrix whose $(i, j)th$ entry is

$$\begin{aligned} Cov(S_i(\vec{X},\theta), S_j(\vec{X},\theta)) &= \mathsf{E}[S_i(\vec{X},\theta)S_j(\vec{X},\theta)] - \underbrace{\mathsf{E}[S_i(\vec{X},\theta)]}_{0} \cdot \underbrace{\mathsf{E}[S_j(\vec{X},\theta)]}_{0} \\ &= \mathsf{E}\left[\left(\tfrac{\partial}{\partial\theta_i}\ln f(\vec{X};\theta)\right)\left(\tfrac{\partial}{\partial\theta_{ij}}\ln f(\vec{X};\theta)\right)\right]. \end{aligned}$$

That is, we have that

$$Var[S(\vec{X},\theta)] = I_n(\theta).$$

So, we have that

$$Var[T] + AI_n(\theta)A^T - 2Cov(t(\vec{X}), AS(\vec{X},\theta)) \geq 0.$$

- The $(i,j)th$ entry of the $q \times q$ variance-covariance matrix $Cov(t(\vec{X}), AS(\vec{X},\theta))$ is

$$\begin{aligned} \mathsf{E}\left[t_i(\vec{X})\textstyle\sum_{k=1}^{p} a_{jk}S_k(\vec{X},\theta)\right] &= \textstyle\sum_{k=1}^{p} a_{jk}\mathsf{E}[t_i(\vec{X})S_k(\vec{X},\theta)] \\ &= \textstyle\sum_{k=1}^{p} a_{jk}Cov(t_i(\vec{X}), S_k(\vec{X},\theta)) \\ &= \textstyle\sum_{k=1}^{p} a_{jk}\frac{\partial}{\partial\theta_k}\tau_i(\theta) \end{aligned}$$

from the first bullet in this proof.

Thus, we have that

$$Cov(t(\vec{X}), AS(\vec{X},\theta)) = \frac{\partial\tau(\theta)}{\partial\theta}A^T.$$

In summary, we have

$$Var[T] + AI_n(\theta)A^T - 2\frac{\partial\tau(\theta)}{\partial\theta}A^T \geq 0. \tag{5.8.14}$$

- As 5.8.14 holds for any $q \times p$ matrix A of constants, it holds if we specifically take

$$A = \frac{\partial\tau(\theta)}{\partial\theta}[I_n(\theta)]^{-1}.$$

Now (5.8.14) becomes

$$Var[T] + \frac{\partial\tau(\theta)}{\partial\theta}[I_n(\theta)]^{-1}\underbrace{I_n(\theta)([I_n(\theta)]^{-1})^T}_{identity}\left(\frac{\partial\tau(\theta)}{\partial\theta}\right)^T - 2\frac{\partial\tau(\theta)}{\partial\theta}\underbrace{([I_n(\theta)]^{-1})^T}_{=I_n(\theta)^{-1}}\left(\frac{\partial\tau(\theta)}{\partial\theta}\right)^T \geq 0.$$

This becomes

$$Var[T] + \frac{\partial\tau(\theta)}{\partial\theta}[I_n(\theta)]^{-1}\left(\frac{\partial\tau(\theta)}{\partial\theta}\right)^T - 2\frac{\partial\tau(\theta)}{\partial\theta}I_n(\theta)^{-1}\left(\frac{\partial\tau(\theta)}{\partial\theta}\right)^T \geq 0,$$

which implies that

$$Var[T] \geq \left(\frac{\partial \tau}{\partial \theta}\right) [I_n(\theta)]^{-1} \left(\frac{\partial \tau}{\partial \theta}\right)^T$$

as desired.

5.8.5 A Computational Example

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from the $N(\mu, \sigma^2)$ distribution. Let us find the CRLB for the variance of all unbiased estimators of $\theta = (\mu, \sigma^2)$.

The pdf is

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}.$$

The joint pdf is

$$f(\vec{x}; \mu, \sigma^2) = (\sqrt{2\pi\sigma^2})^{-n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}.$$

The log is

$$\ln f(\vec{x}; \mu, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

The (1,1) entry of the two-dimensional Fisher information matrix:

The derivative with respect to μ is

$$\frac{\partial}{\partial \mu} \ln f(\vec{x}; \mu, \sigma^2) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)$$

The $(1,1)$ entry of the two-dimensional Fisher information matrix will be

$$\mathsf{E}\left[\left(\frac{\partial}{\partial \mu} \ln f(\vec{X}; \mu, \sigma^2)\right)^2\right] = \mathsf{E}\left[\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)\right)^2\right] = \frac{1}{(\sigma^2)^2}\mathsf{E}\left[\left(\sum_{i=1}^{n}(X_i - \mu)\right)^2\right]$$

That sum squared can be written as a double sum and the double sum can be pulled out of the expectation. When $j \neq i$, X_i and X_j are independent so we have

$$\mathsf{E}[(X_i - \mu)(X_j - \mu)] = \mathsf{E}[X_i - \mu] \cdot \mathsf{E}[X_j - \mu] = 0 \cdot 0 = 0.$$

We have n cases where $j = i$ and we get

$$\mathsf{E}[(X_i-\mu)(X_i-\mu)] = \mathsf{E}[(X_i-\mu)^2] = Var[X_i] = \sigma^2.$$

In summary, the $(1,1)$ entry of the two-dimensional Fisher information matrix will be

$$\mathsf{E}\left[\left(\frac{\partial}{\partial\mu}\ln f(\vec{X};\mu,\sigma^2)\right)^2\right] = \frac{1}{(\sigma^2)^2}\,n\sigma^2 = \frac{n}{\sigma^2}.$$

The (1,2) and (2,1) entries of the two-dimensional Fisher information matrix:

We wish to compute

$$\mathsf{E}\left[\frac{\partial}{\partial\sigma^2}\frac{\partial}{\partial\mu}\ln f(\vec{X};\mu,\sigma^2)\right].$$

Note that

$$\frac{\partial}{\partial\sigma^2}\frac{\partial}{\partial\mu}\ln f(\vec{x};\mu,\sigma^2) = \frac{\partial}{\partial\sigma^2}\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = -\frac{1}{(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu),$$

so we have the $(1,2)$ and $(2,1)$ entries as

$$\mathsf{E}\left[-\frac{1}{(\sigma^2)^2}\sum_{i=1}^{n}(X_i-\mu)\right] = -\frac{1}{(\sigma^2)^2}\mathsf{E}[\sum_{i=1}^{n}(X_i-\mu)] = 0.$$

The (2,2) entry of the two-dimensional Fisher information matrix:

The derivative of the log-likelihood with respect to σ^2 is

$$\frac{\partial}{\partial\sigma^2}\ln f(\vec{x};\mu,\sigma^2) = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

Thus, the $(2,2)$ entry of the two-dimensional Fisher information matrix will be

$$\mathsf{E}\left[\left(\frac{\partial}{\partial\sigma^2}\ln f(\vec{X};\mu,\sigma^2)\right)^2\right] = \mathsf{E}\left[\left(\frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(X_i-\mu)^2 - \frac{n}{2\sigma^2}\right)^2\right]$$

$$= \left(\frac{1}{2(\sigma^2)^2}\right)^2 \mathsf{E}\left[\left(\sum_{i=1}^{n}(X_i-\mu)^2 - n\sigma^2\right)^2\right] = \left(\frac{1}{2(\sigma^2)^2}\right)^2 Var\left[\sum_{i=1}^{n}(X_i-\mu)^2\right]$$

$$\stackrel{iid}{=} \left(\frac{1}{2(\sigma^2)^2}\right)^2 n\, Var[(X_1-\mu)^2]$$

Note that

$$Var[(X_1-\mu)^2] = Var\left[\sigma^2\left(\frac{X_1-\mu}{\sigma}\right)^2\right] = (\sigma^2)^2 Var[W] = 2(\sigma^2)^2$$

since W, as the square of a $N(0,1)$, is $\chi^2(1)$.

Putting it all together, we have the $(2,2)$ entry of the information matrix as

$$\mathsf{E}\left[\left(\frac{\partial}{\partial\sigma^2}\ln f(\vec{X};\mu,\sigma^2)\right)^2\right] = \left(\frac{1}{2(\sigma^2)^2}\right)^2 n\, 2(\sigma^2)^2 = \frac{n}{2(\sigma^2)^2}$$

In summary, the Fisher information matrix is

$$I_n(\theta) = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2(\sigma^2)^2} \end{bmatrix}.$$

The inverse of this diagonal matrix is simply

$$[I_n(\theta)]^{-1} = \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2(\sigma^2)^2}{n} \end{bmatrix}.$$

If we are estimating $\tau(\theta) = \tau(\mu,\sigma^2) = (\mu,\sigma^2)^T$, the CRLB is

$$\left(\frac{\partial\tau}{\partial\theta}\right)[I_n(\theta)]^{-1}\left(\frac{\partial\tau}{\partial\theta}\right)^T = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2(\sigma^2)^2}{n}. \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}^T = \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2(\sigma^2)^2}{n}. \end{bmatrix}$$

5.8.6 The Asymptotic Distribution of Multi-Dimensional MLEs

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with pdf $f(x;\theta)$ where θ is a $p \times 1$ vector of parameters. Let $\widehat{\theta}_n$ be a $p \times 1$ vector of maximum likelihood estimators computed as in Example (5.2.4) where we had $p = 2$.

Under certain "regularity conditions", we still have that

$$\widehat{\theta}_n \overset{asymp}{\sim} N(\theta, CRLB_\theta)$$

only now it means that

$$\sqrt{n}\begin{bmatrix} \widehat{\theta}_1 - \theta_1 \\ \theta_2 - \theta \\ \cdots \\ \widehat{\theta}_p - \theta \end{bmatrix} \xrightarrow{d} N(\vec{0}, [I_n(\theta)]^{-1}).$$

This is a joint convergence in distribution result meaning that a multi-dimensional cdf is converging to a multi-dimensional cdf.

Chapter 5 Exercises

1. Suppose that $X_1, X_2, \ldots, X_n \overset{iid}{\sim} unif(0, \theta)$. Find an MME (method of moments estimator) of θ.
2. Suppose that $X_1, X_2, \ldots, X_n \overset{iid}{\sim} Pareto(\gamma)$. Find an MME (method of moments estimator) of γ.
3. Let $X_1, X_2, \ldots, X_n$ be a random sample from the $Beta(a, b)$ distribution. Find MMEs (method of moments estimators) for a and b.
4. Let $X_1, X_2, \ldots, X_n$ be a random sample from the distribution with pdf
$$f(x;\theta) = \frac{\Gamma(2\theta)}{[\Gamma(\theta)]^2} x^{\theta-1}(1-x)^{\theta-1} I_{(0,1)}(x).$$
Find the method of moments estimator (MME) of θ.
5. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a distribution having pdf
$$f(x;\theta) = \frac{2x}{\theta^2} I_{(0,\theta]}(x).$$
   (a). Find the MME (method of moments estimator) for θ.

(b). Find the MLE for θ.

(c). Find the MLE for the median of the distribution. (Note: The median for X is the value ξ such that $P(X \leq \xi) = 1/2$.)

(d). Compare the variances of your estimators from parts (a) and (b). Which one is better?

6. Let $X_1, X_2, \ldots, X_n$ be a random sample from the $\Gamma(\alpha, \beta)$ distribution. Suppose that α is fixed and known.

(a). Find the MME of β.

(b). Find the MLE of β.

(c). Which estimator (MME or MLE) has smaller variance.

(d). Show that your MLE is a consistent estimator of β. (Do not just quote the result we have that says the MLEs are always consistent estimators.)

7. Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from the $Pareto(\gamma)$ distribution.

(a). Find the MLE (maximum likelihood estimator) for γ.

(b). Find an unbiased estimator of γ based on the MLE from part (a).

(c). Show that your MLE is a consistent estimator of γ. (Do not just quote the result we have that says the MLEs are always consistent estimators.)

8. Let $X_1, X_2, \ldots, X_n$ be a random sample from the shifted exponential distribution with pdf

$$f(x; \lambda, \theta) = \lambda e^{-\lambda(x-\theta)} I_{(\theta,\infty)}(x)$$

for $\lambda > 0$ and $-\infty < \theta < \infty$.

(a). Find the MMEs of λ and θ.

(b). Find the MLEs of λ and θ.

(c). Compare the variances of your two estimators of θ.

9. Let $X_1, X_2, \ldots, X_n$ be a random sample from the continuous distribution with pdf

$$f(x; \theta) = \frac{2\theta(1-x)}{(2x - x^2)^{1-\theta}} I_{(0,1)}(x)$$

Here, $\theta > 0$.

(a). Find the MLE for θ.

(b). Find the distribution of $Y_i = -\ln(2X_i - X_i^2)$. Name it!

(c). Show that the MLE you found in part (a) is an asymptotically unbiased estimator of θ.

10. Let $X_1, X_2, \ldots, X_n$ be a random sample from the Pareto distribution with parameter γ.

(a). Find the Cramér-Rao lower bound (CRLB) for the variance of all unbiased estimators of γ.

(b). Does your unbiased estimator of γ from Problem 2b achieve the CRLB? (If so, we say that it is an **efficient** estimator of γ.)

11. Let $X_1, X_2, \ldots, X_n$ be a random sample from a continuous distribution with pdf $f(x; \theta_0)$. Let $\ell_n(\theta)$

be the corresponding log-likelihood function. (Here, θ_0 is used to emphasize the one true value of the parameter while we are thinking of θ as a variable in the likelihood function.)

(a). Show that

$$\frac{1}{n}\ell_n(\theta) \xrightarrow{P} \mathsf{E}[\ln f(X_1;\theta)].$$

(b). Show that $\mathsf{E}[\ln f(X_1;\theta)] \leq \mathsf{E}[\ln f(X_1;\theta_0)]$ for all θ in the parameter space.

(c). When is the inequality strict?

12. Let $X_1, X_2, \ldots, X_n$ be a random sample from the exponential distribution with rate λ. Recall that the MLE for λ is $\widehat{\lambda} = \widehat{\lambda}_n = 1/\overline{X}$.

(a). Show directly, without quoting properties of MLEs, that $\hat{\lambda}_n$ is asymptotically efficient.

(b). Find the asymptotic distribution of $\hat{\lambda}_n$. (Here you may use any known properties of MLEs.)

13. Let $X_1, X_2, \ldots, X_n$ be a random sample from the $Poisson(\lambda)$ distribution. Verify that $S = \sum_{i=1}^n X_i$ is a sufficient statistic from the definition of sufficiency. (i.e. Do not use the Factorization Criterion.)

14. Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$.

(a). If μ is unknown and σ^2 is known, show that $\sum_{i=1}^n X_i$ is sufficient for μ. Argue then that $\overline{X}$ is also sufficient for μ.

(b). If μ is known and σ^2 is unknown, show that $\sum_{i=1}^n (X_i - \mu)^2$ is sufficient for σ^2.

(c). If μ and σ^2 are both unknown, show that $\sum_{i=1}^n X_i$ and $\sum_{i=1}^n X_i^2$ are jointly sufficient for μ and σ^2. Argue then that $\overline{X}$ and S^2 are also jointly sufficient for μ and σ^2.

15. Let's see the Rao-Blackwell Theorem in action!

Let $X_1, X_2, \ldots, X_n$ be a random sample from the Poisson distribution with rate λ. We wish to find an unbiased estimator of $\tau(\lambda) = e^{-\lambda}$. In Example 5.5.4, we found two. They are

$$\widehat{\tau_1}(\lambda) = I_{\{X_1=0\}} \qquad \text{and} \qquad \widehat{\tau_2}(\lambda) = \left(\frac{n-1}{n}\right)^{\sum_{i=1}^n X_i}.$$

(a). Verify that $\widehat{\tau_2(\lambda)}$ is still unbiased for $e^{-\lambda}$.

(b). Compute and compare the variances of $\widehat{\tau_1(\lambda)}$ and $\widehat{\tau_2(\lambda)}$.

16. Let $X_1, X_2, \ldots, X_n$ be a random sample from the geometric distribution (the one starting from 0) with parameter p. The goal here is to find an unbiased estimator of p. Note that p is not the mean of the distribution and so this is not so easy!

(a). Find a one-dimensional sufficient statistic for p.

(b). Find an unbiased estimator for p based on X_1 alone.

(c). "Rao-Blackwellize" your estimator from part (b) to come up with an unbiased estimator for p based on the entire sample. (This time, you can trust the Rao-Blackwell Theorem to give you an unbiased estimator and do not need to verify unbiasedness.)

17. Let $X_1, X_2, \ldots, X_n$ be a random sample from the $unif(0, \theta)$ distribution.

(a). Show that

$$\frac{\partial}{\partial\theta}\int_0^\theta f(\vec{x};\theta)\,d\vec{x} \neq \int_0^\theta \frac{\partial}{\partial\theta} f(\vec{x};\theta)\,d\vec{x}$$

(b). Equality of the integrals in (a) was key to our proof of the Cramér-Rao lower bound. Show that the CRLB for the variance of all unbiased estimators of θ does not apply for this distribution by exhibiting a "superefficient" unbiased estimator for θ.

18. Let X be a continuous random variable with pdf f. Suppose that g_1 and g_2 are two functions such that $\mathsf{E}[g_1^2(X)] < \infty$ and $\mathsf{E}[g_2^2(X)] < \infty$. Show that

$$(\mathsf{E}[g_1(X)\,g_2(X)])^2 \leq \mathsf{E}[g_1^2(X)] \cdot \mathsf{E}[g_2^2(X)].$$

(Hint: Look at the proof of the Cauchy-Schwarz inequality given in this Chapter.)

19. Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with pdf $f(\vec{x};\theta)$ where θ is a scale parameter. Let $T = t(\vec{X})$ be a scale-invariant statistic. Show that T is ancillary.

20. Suppose that $X_1, X_2, \cdots, X_n \overset{iid}{\sim} \Gamma(\alpha, \beta)$ for $\alpha > 0$ fixed and known. Consider the statistic

$$T = \frac{X_{(1)}}{\sum_{i=1}^n X_i}.$$

(a). Show that T is independent of $\overline{X}$.

(b). Use part (a) to find $\mathsf{E}[T]$ when $\alpha = 1$.

Chapter 6 General Hypothesis Testing

In Chapter 4, we had a taste of hypothesis testing. However, we were limited in what distributions we could handle and we had to guess what statistics we should base our tests on without any real explanation other than "this seems like a good idea". We could only handle hypotheses involving the comparisons in $\{=, \neq, <, \leq, >, \geq\}$. In this Chapter, we will generalize our techniques and, to that end, we must first generalize our notation.

6.1 Language and Notation

As usual, we use θ to denote a generic parameter of interest and Θ to denote the parameter space. Let $\Theta_0 \subseteq \Theta$ be some subset of the parameter space.

A general set of hypotheses looks like this:

$$H_0 : \theta \in \Theta_0 \qquad \text{versus} \qquad H_1 : \theta \in \Theta \setminus \Theta_0.$$

The backslash here is the "setminus" notation. $\Theta \setminus \Theta_0$ is everything that is left in Θ if you remove all of the elements of Θ_0. It can be written as

$$\Theta \setminus \Theta_0 = \Theta \cap \Theta_0^C$$

This notation is useful when Θ is thought of as part of a larger space such as the real number line. For example, let $\Theta = (0, \infty)$ and let $\Theta_0 = (1, \infty)$. Then $\Theta_0^C = (-\infty, 1]$, but $\Theta \setminus \Theta_0 = (0, 1]$.

In Chapter 4, we defined tests through "rejection rules" such as

$$\text{"Reject } H_0 \text{ if } \textstyle\sum_{i=1}^n X_i^2 \leq 5\text{."}$$

This test says to reject H_0 if the statistic $\sum_{i=1}^n X_i^2$ is in the region $(-\infty, 5]$. Another way to write this would be to say that the vector $(X_1, X_2, \ldots, X_n)$ is in the region where $\sum_{i=1}^n X_i^2$. This is known as a **rejection region** or a **critical region**. If we define the critical region as

$$C = \{(x_1, x_2, \ldots, x_n) : \sum_{i=1}^n x_i^2 \leq 5\}$$

then our rejection rule is now to

$$\text{“Reject } H_0 \text{ if } \vec{X} \in C.\text{”}$$

Note that the test is completely determined by the set C and that finding a good test is equivalent to finding a good critical region C. Comparing two tests of size α, as in Example 4.7.2 in Chapter 4, is equivalent to comparing two critical regions of size α. This notation is really going to streamline the conversation!

Level of Significance

We continue to use the symbol α to denote the “level of significance" of a test. It is also known as the “size" of the test.[1]

If we are now defining a test by whether or not a sample falls into a critical region C, we may write α, for a simple null hypothesis ($H_0 : \theta = \theta_0$), as

$$\begin{aligned}
\alpha &= \max P(\text{Type I error}) \\
&= \max_{\theta \in \Theta_0} P(\text{ Reject } H_0 \text{ when the parameter is } \theta\,) \\
&= \max_{\theta \in \Theta_0} P(\text{ Reject } H_0; \theta\,) \\
&= \max_{\theta \in \Theta_0} P(\,\vec{X} \in C; \theta\,)
\end{aligned}$$

Power Function

We will use a subscript for the power function of the test to connect it to the test defined by the critical region C.

$$\begin{aligned}
\gamma_C(\theta) &= P(\text{ Reject } H_0 \text{ when the parameter is } \theta\,) \\
&= P(\text{ Reject } H_0; \theta\,) \\
&= P(\,\vec{X} \in C; \theta\,)
\end{aligned}$$

[1] This language comes from looking at probability as a “measure” and $P(A)$, for example, as measuring the size of the set A representing an event.

6.2 The "Best" Test

In this Section, we will consider the simple versus simple hypotheses

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta = \theta_1 \tag{6.2.1}$$

as a building block for more interesting sets of hypotheses.

As with any hypothesis test, the options here should be assumed to represent all possibilities for θ, i.e., $\Theta = \{\theta_0, \theta_1\}$, even if the true parameter space for the underlying distribution is actually larger.

We want a test of size α. This is equivalent to finding a critical region C of size α. That is, we will find a set C such that

$$\max_{\theta \in \Theta_0} P(\vec{X} \in C; \theta) = P(\vec{X} \in C; \theta_0) = \alpha.$$

The first equality is true since Θ_0 consists only of the single point θ_0.

Just as there may be several different rejection rules (for example, one based on $\overline{X}$, one based on $X_{(1)}$, et cetera), there may be several such sets C. Let's suppose that furry friends Kermit and Uli both have tests of size α for (6.2.1). This means, by definition, that they each have a critical region, that we will call C_K and C_U, respectively, such that

$$P(\vec{X} \in C_K; \theta_0) = \alpha \qquad \text{and} \qquad P(\vec{X} \in C_U; \theta_0) = \alpha.$$

How can we compare these two tests?

The tests/sets "act the same" when H_0 is true so we will try to differentiate them when H_0 is false and H_1 is true. In other words, we want to compare

$$P(\vec{X} \in C_K; \theta_1) \qquad \text{with} \qquad P(\vec{X} \in C_U; \theta_1).$$

Suppose that

$$P(\vec{X} \in C_K; \theta_1) \geq P(\vec{X} \in C_U; \theta_1). \tag{6.2.2}$$

Whose test is better?

We are looking at probabilities that the sample falls in critical (rejection) regions when H_1 is true. When the sample falls in a critical region, we reject H_0. So, we are looking at probabilities of rejecting H_0 when we should because H_1 is true, which is a good thing. Kermit's test is better because he has a higher probability of doing the right thing here.

What if the inequality in (6.2.2) holds using Kermit's C_K in the left hand side and **any** other critical region of size α in the entire world in the right hand side? Then Kermit would have the best test among all tests!

Definition 6.2.1

For testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, with a random sample $X_1, X_2, \ldots, X_n$, C_* is called a **best critical region of size** α (or the corresponding test is a **best test of size** α) if:

1. $P(\vec{X} \in C_*; \theta_0) = \alpha$ (This makes it “of size α".)
2. If C is any other set such that $P(\vec{X} \in C; \theta_0) = \alpha$, then
$$P(\vec{X} \in C_*; \theta_1) \geq P(\vec{X} \in C; \theta_1).$$
(This makes it “best".)

Note that part 2 of the definition of a best test can now be written in terms of power functions as

$$\gamma_{C_*}(\theta_1) \geq \gamma_C(\theta_1).$$

for all C such that $P(\vec{x} \in C; \theta_0) = \alpha$.

Recall from Section 4.7 in Chapter 4 that, in this case of a simple null hypothesis, the power of a test is related to the power function as

$$\begin{aligned} 1 - \beta &= 1 - P(\text{ Type II Error }) \\ &= 1 - P(\text{ Fail to reject } H_0 \text{ when } H_1 \text{ is true }) \\ &= P(\text{ Reject } H_0 \text{ when } H_1 \text{ is true }) \\ &= P(\text{ Reject } H_0; \theta_1) \\ &= \gamma(\theta_1) \end{aligned}$$

From Definition 6.2.1 and the comment immediately following, we say that the test based on C_* has “higher power", when $\theta = \theta_1$, than the test based on C.

A best test of size α is often refered to as a

most powerful test of size α.

How can one derive a best or most powerful test?

The Neyman-Pearson Lemma

Suppose that $X_1, X_2, \ldots, X_n$ have joint pdf $f(\vec{x}; \theta)$.

Consider the ratio

$$\lambda(\vec{x}; \theta_0, \theta_1) = \frac{f(\vec{x}; \theta_0)}{f(\vec{x}; \theta_1)}$$

and let C be the set $C = \{\vec{x} : \lambda(\vec{x}; \theta_0, \theta_1) \leq k\}$ where k is a constant such that $P(\vec{X} \in C; H_0) = \alpha$.

Then C is a **best critical region** of size α for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$.

Comments about the Neyman-Peason Lemma:

1. In order to maximize the power of the test, points that are more likely under H_1 than under H_0 should be put into the critical region.

2. If a point $(x_1, \ldots, x_n)$ is more likely under H_1 than under H_0, then the denominator of $\lambda(\vec{x}; \theta_0, \theta_1)$ is larger and so $\lambda(\vec{x}; \theta_0, \theta_1)$ is smaller. This is why the requirement for putting $(x_1, \ldots, x_n)$ into C is of the form $\lambda(\vec{x}; \theta_0, \theta_1) \leq k$, regardless of whether $\theta_0 < \theta_1$ or $\theta_1 < \theta_0$.

3. Let $L(\theta)$ be a likelihood function for the model under consideration. Recall that $L(\theta)$ is proportional to the joint pdf $f(\vec{x}; \theta)$ when it is thought of as a function of θ. This means that $L(\theta) = c \cdot f(\vec{x}; \theta)$ where c may or may not depend on $\vec{x}$. The ratio from the Neymann-Pearson Lemma is called a **likelihood ratio** since

$$\frac{f(\vec{x}; \theta_0)}{f(\vec{x}; \theta_1)} = \frac{(1/c)L(\theta_0)}{(1/c)L(\theta_1)} = \frac{L(\theta_0)}{L(\theta_1)}.$$

   For this reason, the test given by the Neyman-Pearson Lemma is one example of a **likelihood ratio test**, discussed more generally in Section 6.4.

Proof : (Neyman-Pearson Lemma)

We will prove the "N-P" Lemma for a random sample from a continuous distribution. The proof for a discrete distribution is similar.

- Let C_* be the so-called "best" critical region described by the N-P Lemma.
- Let $\vec{X} = (X_1, X_2, \ldots, X_n)$. Note that for any set A, (not necessarily a critical region), $P(\vec{X} \in A; \theta)$ may be written as
$$P(\vec{X} \in A; \theta) = \int_A f(\vec{x}; \theta)\, d\vec{x}$$
which is shorthand notation for an n-dimensional integral with respect to $dx_1\, dx_2 \ldots dx_n$ where all the limits of integration run around describing points $(x_1, x_2, \ldots, x_n) \in A$.
- Note that if $A \subseteq C_*$,
$$P(\vec{X} \in A; \theta_0) \leq k \cdot P(\vec{X} \in A; \theta_1)$$
since, in C_*, and hence in $A \subseteq C_*$, we have $f(\vec{x}; \theta_0) \leq k \cdot f(\vec{x}; \theta_1)$, so
$$P(\vec{X} \in A; \theta_0) = \int_A f(\vec{x}; \theta_0)\, d\vec{x} \leq k \int_A f(\vec{x}; \theta_1)\, d\vec{x} = k \cdot P(\vec{X} \in A; \theta_1).$$
- On the other hand, by the same reasoning, if $A \subseteq C_*^c$, (the complement of C_*), we have
$$P(\vec{X} \in A; \theta_0) > k \cdot P(\vec{X} \in A; \theta_1).$$
- Let C denote an arbitrary critical region of size α. This means that C satisfies
$$P(\vec{X} \in C; \theta_0) = \alpha.$$
- Note that we may write
$$C_* = (C_* \cap C) \cup (C_* \cap C^c)$$
and that $C_* \cap C$ and $C_* \cap C^c$ are disjoint sets.

  We then have
$$\gamma_{C_*}(\theta) = P(\vec{X} \in C_*; \theta) = P(\vec{X} \in C_* \cap C; \theta) + P(\vec{X} \in C_* \cap C^c; \theta). \tag{6.2.3}$$
- On the other hand, we may write
$$C = (C \cap C_*) \cup (C \cap C_*^c)$$
and note that $C \cap C_*$ and $C \cap C_*^c$ are disjoint sets.

So,

$$\gamma_C(\theta) = P(\vec{X} \in C; \theta) = P(\vec{X} \in C \cap C_*; \theta) + P(\vec{X} \in C \cap C_*^c; \theta). \tag{6.2.4}$$

- In order to prove the N-P Lemma, we want to show that $\gamma_{C_*}(\theta_1) \geq \gamma_C(\theta_1)$. So, we consider the difference of (6.2.3) and (6.2.4):

$$\gamma_{C_*}(\theta) - \gamma_C(\theta) = P(\vec{X} \in C_* \cap C^c; \theta) - P(\vec{X} \in C \cap C_*^c; \theta). \tag{6.2.5}$$

- Specifically, we consider the difference at θ_1 and use the fact that $C_* \cap C^c \subseteq C_*$ and $C \cap C_*^c \subseteq C_*^c$:

$$C_* \cap C^c \subseteq C_* \qquad \Rightarrow \qquad P(\vec{X} \in C_* \cap C^c; \theta_1) \geq (1/k) \cdot P((\vec{X} \in C_* \cap C^c; \theta_0)$$

$$C \cap C_*^c \subseteq C_*^c \qquad \Rightarrow \qquad P(\vec{X} \in C \cap C_*^c; \theta_1) < (1/k) \cdot P(\vec{X} \in C \cap C_*^c; \theta_0)$$

So, by (6.2.5),

$$\begin{aligned} \gamma_{C_*}(\theta_1) - \gamma_C(\theta_1) &\geq (1/k)\left[P(\vec{X} \in C_* \cap C^c; \theta_0) - P(\vec{X} \in C \cap C_*^c; \theta_0)\right] \\ &= (1/k)\left[\gamma_{C_*}(\theta_0) - \gamma_C(\theta_0)\right] \qquad \text{by (6.2.5) again} \\ &= (1/k)(\alpha - \alpha) = 0 \end{aligned}$$

- Therefore, $\gamma_{C_*}(\theta_1) \geq \gamma_C(\theta_1)$. Since this holds for any size α region C, we have shown that C_* is the best critical region of size α!

■

Example 6.2.1

Let $X_1, X_2, \ldots, X_n$ be a random sample from the $exp(rate = \theta)$ distribution. Find the best test of size α for

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta = \theta_1$$

where $\theta_0 < \theta_1$.

The N-P Theorem says to reject H_0 if

$$\lambda(\vec{x}; \theta_0, \theta_1) \leq k$$

where k is such that $P(\lambda(\vec{X};\theta_0,\theta_1) \leq k;\theta_0) = \alpha$.

We begin my computing the likelihood ratio.

$$\lambda(\vec{x};\theta_0,\theta_1) = \frac{\theta_0^n e^{-\theta_0 \sum x_i} \prod I_{(0,\infty)(x_i)}}{\theta_1^n e^{-\theta_1 \sum x_i} \prod I_{(0,\infty)}(x_i)} = \left(\frac{\theta_0}{\theta_1}\right)^n e^{-(\theta_0-\theta_1)\sum x_i}.$$

We are going to set this less than or equal to k and then we will try to isolate the X's on the left-hand side.

$$\left(\frac{\theta_0}{\theta_1}\right)^n e^{-(\theta_0-\theta_1)\sum x_i} \leq k$$

$$\Downarrow$$

$$e^{-(\theta_0-\theta_1)\sum X_i} \leq k \cdot \left(\frac{\theta_1}{\theta_0}\right)^n$$

The right-hand side is just another constant that we can call k_1:

$$e^{-(\theta_0-\theta_1)\sum X_i} \leq k_1.$$

Taking the log of both sides, we have

$$-(\theta_0-\theta_1)\sum X_i \leq k_2$$

where $k_2 = \ln(k_1)$.

Finally, we divide both sides by $-(\theta_0-\theta_1)$. Note that this quantity is positive since we were given that $\theta_0 < \theta_1$. Thus, the inequality does not flip. We have

$$\sum_{i=1}^{n} X_i \leq k_3$$

where $k_3 = -\ln(k_1)/(\theta_0-\theta_1)$.

We will see that keeping track of these constants is not important.

We have

$$\begin{aligned} \alpha &= P\left(\lambda(\vec{x};\theta_0,\theta_1) \le k;\theta_0\right) \\ &= P\left(\textstyle\sum_{i=1}^n X_i \le k_3\right) \end{aligned}$$

where k_3 will be determined to give a test of size α.

Once we find k_3 we could "backsolve" to find k but, ultimately, it depends on how we want to report the test. Do you really want to tell a client that the test should be to reject H_0, in favor of H_1, if

$$\left(\frac{\theta_0}{\theta_1}\right)^n e^{-(\theta_0-\theta_1)\sum x_i} \le k$$

where $k = -\ln(k\cdot(\theta_1/\theta_0))^n/(\theta_0-\theta_1)$?

Instead, you could just tell them to look at the sum of the data points and reject H_0, in favor of H_1, if that sum is less than or equal to some other number you found for them. There is no need to keep track of the constants!

So, for this example, a best critical region of size α has the form:

$$C = \left\{\vec{x} : \sum x_i \le k_3\right\}$$

for some k_3 which must be chosen so that

$$P(\vec{X} \in C;\theta_0) = \alpha.$$

Equivalently, we

"Reject H_0, in favor of H_1, if $\sum_{i=1}^n \le k_3$ where k_3 is determined so that the test will have size α."

Although we are not finished, this is the **form of the test**. Often the form of a test is all that we are interested in. Suppose, however, that we want to find the actual test. That is, suppose we want the value of k_3. This part is no different than it was in Chapter 4.

$$\begin{aligned} \alpha &= P(\vec{X} \in C;\theta_0) \\ &= P(\textstyle\sum_{i=1}^n X_i \le k_3;\theta_0) \end{aligned}$$

The "semicolon θ_0" is saying that we want to compute the probability when $H_0 : \theta = \theta_0$ is true. We also say that we want to compute the probability "under H_0".

If $\theta = \theta_0$, we know that

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} exp(rate = \theta_0),$$

which implies that

$$\sum_{i=1}^{n} X_i \sim \Gamma(n, \theta_0).$$

So, for a given α, n, and θ_0, we could find k_3 by solving

$$\alpha = P(G \leq k_3) \qquad \text{where} \qquad G \sim \Gamma(n, \theta_0).$$

In practice, the gamma pdf is not easy to integrate, and it is not easy to invert the resulting integral in order to solve for k_3. One option would be for us to solve for k_3 numerically, using a computer.

In Chapter 4, we noted that the χ^2 distribution plays a pretty central role in statistics. Let us try to report this test in terms of a χ^2-critical value.

Recall that $\chi^2(n) = \Gamma(n/2, 1/2)$ and that $X \sim \Gamma(\alpha, \beta) \Rightarrow cX \sim \Gamma(\alpha, \beta/c)$, provided that $c > 0$. So, we can turn G into a χ^2-random variable by multiplying by $2\theta_0$:

$$2\theta_0 G \sim \Gamma\left(n, \frac{\theta_0}{2\theta_0}\right) = \Gamma\left(n, \frac{1}{2}\right) = \Gamma\left(\frac{2n}{2}, \frac{1}{2}\right) = \chi^2(2n)$$

Now

$$\begin{aligned} \alpha &= P(G \leq k_3) \\ &= P(2\theta_0 G \leq 2\theta_0 k_3) \\ &= P(W \leq 2\theta_0 k_3) \end{aligned}$$

where $W \sim \chi^2(2n)$.

What number is W below with probability α? Using our critical value notation established in Chapter 4, it is $\chi^2_{1-\alpha,2n}$.

So, $2\theta_0 k_3 = \chi^2_{1-\alpha,2n}$ which means that

$$k_3 = \frac{\chi^2_{1-\alpha,2n}}{2\theta_0}.$$

In summary, if $X_1, X_2, \ldots, X_n$ is a random sample from the $exp(rate = \theta)$ distribution, the best test of size α of $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ is to

"Reject H_0, in favor of H_1 if $\sum X_i \leq \frac{\chi^2_{1-\alpha,2n}}{2\theta_0}$."

If you are wondering why we didn't define another constant $k_4 = 2\theta_0 k_3$, it again comes down to how we want to report the test. It is preferable to keep it in terms of the simplest statistic. We want to say that the test is to reject H_0 when the sum of the X's is below some number as opposed to when "2 times θ_0 times the sum of the X's" is below some other number. Either way is valid though.

Let's try another one.

Example 6.2.2

Consider the distribution with pdf $f(x;\theta) = \theta x^{\theta-1} I_{(0,1)}(x)$. Based on a random sample of size $n = 1$, find the best test of $H_0 : \theta = 1$ versus $H_1 : \theta = 2$ with $\alpha = 0.05$.

$$\lambda(x_1; 1, 2) = \frac{f(x_1;1)}{f(x_1;2)} = \frac{1 \cdot I_{(0,1)}(x_1)}{2x_1 \cdot I_{(0,1)(x_1)}} = \frac{1}{2x_1}$$

According to the Neyman-Pearson Lemma, we want to reject H_0 if

$$\frac{1}{2x_1} \leq k$$

where k is such that

$$P(X_1 \in C; H_0) = 0.05.$$

Solving for x_1 on one side of the inequality, it is equivalent to reject H_0 i

$$x_1 \geq k_2$$

for some k_2 determined to give a size α test.

Looking at the original pdf, we see that, When H_0 is true, $X_1 \sim unif(0, 1)$. We are ready to find k_2.

$$\begin{aligned} 0.05 &= P(\text{ Type I Error }) \\ &= P(\text{ Reject } H_0 \text{ when true }) \\ &= P(X_1 \geq k_2; 1) \\ &= 1 - k_2 \end{aligned}$$

since $X_1 \sim unif(0, 1)$.

This gives us that $k_2 = 0.95$ and the best test of size 0.05 is to

"Reject H_0, in favor of H_1 if $X_1 \geq 0.95$."

6.3 Uniformly Most Powerful Tests (UMPs)

In this section, we will consider a simple versus composite hypothesis of the form:

$$H_0 \ : \ \theta = \theta_0$$

$$H_1 \ : \ \theta \in \Theta \setminus \{\theta_0\}.$$

(H_1 will typically be one of: $\theta \neq \theta_0$, $\theta < \theta_0$, or $\theta > \theta_0$. In theory, it can even be something "weird" like H_1 is all values of θ such that $sin(3\theta) > 0.2$.)

As discussed in Section 5.5 of Chapter 5, the word "uniformly" in mathematics means "for all". In this case, we're going to be doing something "for all" θ in the alternative hypothesis set.

Definition 6.3.1

For simple H_0 and composite H_1, the critical region C is a **uniformly most powerful critical region of size** α if C is the most powerful (best) critical region for testing H_0 against every simple hypotheses in H_1.

The corresponding test is a **uniformly most powerful (UMP) test**.

Comments:

1. A UMP test may not exist. (We will see that there is usually trouble with the two-sided alternative hypothesis $H_1 : \theta \neq \theta_0$.)
2. H_1 could be simple. Then the most powerful or best test described in Section 6.2 is, by default, uniformly most powerful. It is the most powerful (best) test for all θ in the singleton set $\{\theta_1\}$!
3. A UMP test may be easily defined for the composite versus composite case:

   $$H_0 : \theta \in \Theta_0 \qquad \text{versus} \qquad H_1 : \theta \in \Theta \setminus \Theta_0$$

   C_* is a UMP critical region of size α if

   $$\alpha = \max_{\theta \in \Theta_0} \gamma_{C_*}(\theta)$$

   and

   $$\gamma_{C_*}(\theta) \geq \gamma_C(\theta) \qquad \forall\ \theta \in \Theta \setminus \Theta_0$$

   and for all critical regions C of size α. That said, we may not be able to find it using the Neyman-Pearson Lemma. If we look at the proof of the N-P Lemma and try to adapt it to a composite null hypothesis, we will see that the proof only holds when the maximum probability of making a Type I error occurs on the boundary between H_0 and H_1. This has been the case for all of our examples so far but need not be the case in general.

6.3.1 Finding a UMP Test

One possible way to find a UMP test for $H_0 : \theta = \theta_0$ versus $H_1 : \theta \in \Theta \setminus \{\theta_0\}$ is to derive the Neyman-Pearson test for a particular alternative value in the set $\Theta \setminus \theta_0$ and then to show that the test does not depend on the specific alternative value.

Example 6.3.1

Let $X_1, X_2, \ldots, X_n$ be a random sample form the exponential distribution with rate θ.

Suppose that we want to test,

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta > \theta_0.$$

We could consider the simple versus simple hypothesis:

$$H_0 : \theta = \theta_0 \qquad \text{and} \qquad H_1 : \theta = \theta_1$$

for some $\theta_1 > \theta_0$.

This is exactly what we tested in Example 6.2.1 in Section 6.2.

The best or most powerful or best test for this simple versus simple hypothesis was to reject H_0 if

$$\sum_{i=1}^{n} X_i \leq \frac{\chi^2_{1-\alpha,2n}}{2\theta_0}.$$

The intermediate steps may have depended on θ_1, but the ultimate decision rule does not. Recall though that it was necessary that θ_1 as long as it is greater than θ_0 or else the inequality in the decision rule would have flipped. (The subscripts on the χ^2-critical value would change as well.)

In summary, we get this best test for any $\theta_1 > \theta_0$. By definition of a UMP test, this decision rule gives us the UMP test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$.

Example 6.3.2

Let $X_1, X_2, \ldots, X_n$ be a random sample from the distribution with pdf

$$f(x;\theta) = \frac{3x^2}{\theta}\, e^{-x^3/\theta}\, I_{(0,\infty)}(x).$$

Let's find a UMP test for

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta > \theta_0$$

We begin by reducing the problem to the simple versus simple hypotheses:

$$H_0 : \theta = \theta_0 \qquad \text{and} \qquad H_1 : \theta = \theta_1$$

for some fixed $\theta_1 > \theta_0$.

We then use the Neyman-Pearson Lemma for the simple versus simple hypotheses:

$$\lambda(\vec{x};\theta_0,\theta_1) = \frac{\prod_{i=1}^n \left[\frac{3x_i^2}{\theta_0}\, e^{-x_i^3/\theta_0}\right]}{\prod_{i=1}^n \left[\frac{3x_i^2}{\theta_1}\, e^{-x_i^3/\theta_1}\right]} = \left(\frac{\theta_1}{\theta_0}\right)^n e^{-\left(\frac{\theta_1-\theta_0}{\theta_0\theta_1}\right)\sum x_i^3}$$

We will reject H_0 if $\lambda(\vec{x};\theta_0,\theta_1) \leq k$ where k is such that $P(\lambda(\vec{X};\theta_0,\theta_1) \leq k; H_0) = \alpha$.

$$\left(\frac{\theta_1}{\theta_0}\right)^n e^{-\left(\frac{\theta_1-\theta_0}{\theta_0\theta_1}\right)\sum x_i^3} \leq k$$

We want to isolate the x_i on one side. We can begin by dividing both sides by $(\theta_1/\theta_0)^n$

$$e^{-\left(\frac{\theta_1-\theta_0}{\theta_0\theta_1}\right)\sum x_i^3} \leq k_1$$

We then take the log of both sides.

$$-\left(\frac{\theta_1-\theta_0}{\theta_0\theta_1}\right)\sum x_i^3 \leq k_2$$

Finally, we divide by that coefficient. It is negative since $\theta_1 > \theta_0$ so the inequality flips.

$$\sum x_i^3 \geq k_3$$

for some k_3 to be determined to give a size α test.

The form of the test (for the simple versus simple hypotheses) is to

"Reject H_0, in favor of H_1, if $\sum_{i=1}^n X_i^3 \geq k_3$."

Let's find k_3. We need

$$P(\sum X_i^3 \geq k_3; H_0) = \alpha.$$

What is the distribution of $\sum X_i^3$ if H_0 is true?

Let's first consider just one of the X's:

Let $Y = X^3$. Then $y = g(x) = x^3 \Rightarrow x = g^{-1}(y) = y^{1/3}$.

Then

$$\begin{aligned} f_Y(y) &= f_X(g^{-1}(y)) \cdot \left|\tfrac{\partial}{\partial y} g^{-1}(y)\right| \\ &= \tfrac{3(y^{1/3})^2}{\theta_0}\, e^{-(y^{1/3})^3/\theta_0} \cdot I_{(0,\infty)}(y^{1/3}) \cdot \left|\tfrac{1}{3} y^{-2/3}\right| \\ &= \tfrac{1}{\theta_0} e^{-y/\theta_0} \cdot I_{(0,\infty)(y)}. \end{aligned}$$

(Note: $0 < y < \infty \Leftrightarrow 0 < y^{1/3} < \infty$. The absolute value is omitted since $y > 0$. Even if $y < 0$, the term in the absolute value is $y^{-2/3} = (y^{-1/3})^2 > 0$.)

When H_0 is true (which means that $\theta = \theta_0$) we have $Y = X^3 \sim exp(rate = 1/\theta_0)$.

Therefore,

$$\sum_{i=1}^{n} X_i^3 \sim \Gamma(n, \theta_0)$$

So,

$$\begin{aligned} \alpha &= P(\textstyle\sum X_i^3 \geq k_3; H_0) \\ &= P(G \geq k_3) \end{aligned}$$

where $G \sim \Gamma(n, \theta_0)$.

As in the previous example, this in not analytically tractable, so let's go for a chi-squared critical value. We can make the needed transformation just as before.

Recall that

$$G \sim \Gamma(n, \theta_0) \qquad \Rightarrow \qquad 2\theta_0\, G \sim \Gamma(n, 1/2) = \Gamma(2n/2, 1/2) = \chi^2(2n).$$

So,

$$\begin{aligned} \alpha &= P(G \geq k_3) \\ &= P(2\theta_0\, G \geq 2\theta_0\, k_3) \\ &= P(W \geq 2\theta_0\, k_3) \end{aligned}$$

where $W \sim \chi^2(2n)$.

Therefore

$$2\theta_0\, k_3 = \chi^2_{\alpha,2n} \qquad \Rightarrow \qquad k_3 = \frac{\chi^2_{\alpha,2n}}{2\theta_0}.$$

Let us review where we are so far in this example:

We have determined that the best or most powerful test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ where $\theta_1 > \theta_0$ is to

$$\text{Reject } H_0, \text{ in favor of } H_1, \text{ if } \sum_{i=1}^{3} X_i^3 \geq \frac{\chi^2_{\alpha,2n}}{2\theta_0}.$$

This test does not depend on the specific value of θ_1, although it did depend on the fact that $\theta_1 > \theta_0$ (otherwise an inequality would have flipped). So, it will work for any $\theta_1 > \theta_0$.

Hence, the test:

$$\text{Reject } H_0, \text{ in favor of } H_1, \text{ if } \sum_{i=1}^{3} X_i^3 \geq \frac{\chi^2_{\alpha,2n}}{2\theta_0}.$$

is a UMP test of size α for $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$.

Example 6.3.3

Let us consider Example 6.3.2 again with the hypotheses: $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.

We begin by considering testing the simple versus simple hypotheses $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$ for any fixed $\theta_1 \neq \theta_0$. The likelihood ratio is exactly the same and we set it less than or euqal to some constant k.

$$\left(\frac{\theta_1}{\theta_0}\right)^n e^{-\left(\frac{\theta_1-\theta_0}{\theta_0\theta_1}\right)\sum x_i^3} \leq k$$

We try to isolate the X_i on the left-hand side in the same way as before. After dividing both sides by $(\theta_1/\theta_0)^n$ and taking the log of both sides, we have

$$-\left(\frac{\theta_1-\theta_0}{\theta_0\theta_1}\right)\sum x_i^3 \leq k_1$$

for some constant k_1 to be determined.

The final step in isolating the X_i is to divide by $-(\theta_1 - \theta_0)/(\theta_0\theta_1)$. Here, we have a problem. We fixed a

$\theta_1 \neq \theta_0$. If it is greater than θ_0, the coefficient is negative and the inequality will flip upon division. If it is less than θ_0, the coefficient is positive and the inequality will not flip.

The best test, when $\theta_1 > \theta_0$ will have the form

"Reject H_0, in favor of H_1, if $\sum X_i^3 \leq k_2$."

The best test, when $\theta_1 < \theta_0$ will have the form

"Reject H_0, in favor of H_1, if $\sum X_i^3 \geq k_2$."

(The k_2's will be different.)

Even without finding the constant, the test that we best (most powerful) in one case will not be the best test for the other case. There is not one best test that works in both cases. Therefore, there is no UMP test for this example.

It is often, but not always, the case that a UMP does not exist for the two-sided alternative hypothesis. For example, if the coefficient had been $-(\theta_1 - \theta_0)^2/(\theta_0\theta_1)$ the sign would be the same in the case $\theta_1 > \theta_0$ and $\theta_1 < \theta_0$. We wouldn't have the problem of an inequality flipping in one case and not the other.

Rumination

Can we expand our method for finding a UMP test in the case of a composite null hypothesis?

At first glance it might seem like we can go through all of the same steps as in Examples 6.3.1 and 6.3.2 to get the form of the test and then deal with the composite null hypothesis when finding the constant k (or k_1, of k_2, or,...) with the addition of a maximization step when introducing α. However, the procedure we used in the noted examples was based on the Neyman-Pearson Lemma for finding the best test in the simple versus simple case.

While it is not obvious, one can see through careful analysis of the proof of the Neyman-Pearson Lemma, that it will still hold if the maximum probability of making a Type I error occurs on the boundary between H_0 and H_1. This has been the case for all examples in this text but does not have to be true in general.

If the maximum probability of making a Type I error does not occur on the boundary between H_0 and H_1, it does not mean that a UMP test does not exist. It just means that we can't use Neyman and Pearson's recipe for finding it. Indeed, we have no recipe to give you and anything we might try would be problem specific!

6.3.2 Exponential Families and UMP Tests

Here, for the record, is a theorem that is often given in MathStat textbooks. However, it is just consists of special cases of things we already know so we will not spend much time on it.

Theorem 6.3.1

Suppose that $X_1, X_2, \ldots, X_n$ have a joint pdf in the one-parameter exponential family:

$$f(\vec{x};\theta) = a(\theta)b(\vec{x})\exp\left[c(\theta)d(\vec{x})\right]$$

If $c(\theta)$ is an increasing function of θ. Then

- A UMP test of size α for $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$ is to reject H_0 if $d(\vec{x}) \geq k$ where k is chosen so that $P(d(\vec{X}) \geq k; \theta_0) = \alpha$.
- A UMP test of size α for $H_0 : \theta = \theta_0$ versus $H_1 : \theta < \theta_0$ is to reject H_0 if $d(\vec{x}) \leq k$ where k is chosen so that $P(d(\vec{X}) \leq k; \theta_0) = \alpha$.

If $c(\theta)$ is a decreasing function of θ. Then

- A UMP test of size α for $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$ is to reject H_0 if $d(\vec{x}) \leq k$ where k is chosen so that $P(d(\vec{X}) \leq k; \theta_0) = \alpha$.
- A UMP test of size α for $H_0 : \theta = \theta_0$ versus $H_1 : \theta < \theta_0$ is to reject H_0 if $d(\vec{x}) \geq k$ where k is chosen so that $P(d(\vec{X}) \geq k; \theta_0) = \alpha$.

Proof : [first bullet]

Without loss of generality, we will assume both functions $a(\theta)$ and $b(\vec{x})$ are positive. (They are either both positive or both negative in order for the pdf to be valid. If they are both negative, the product is still positive and we can ignore the negatives.)

- We consider the simple versus simple hypotheses

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta = \theta_1$$

for a fixed $\theta_1 > \theta_0$.

- The N-P likelihood ratio is

$$\lambda(\vec{x};\theta_0,\theta_1) = \frac{f(\vec{x};\theta_0)}{f(\vec{x},\theta_1)} = \frac{a(\theta_0)}{a(\theta_1)} \cdot \exp\left[(c(\theta_0) - c(\theta_1))\, d(\vec{x})\right].$$

- Setting $\lambda(\vec{x};\theta_0,\theta_1) \leq k$ and isolating the X_i's on the left side will give

$$\frac{a(\theta_0)}{a(\theta_1)} \cdot \exp\left[(c(\theta_0) - c(\theta_1))\, d(\vec{x})\right] \leq k$$

$$\Downarrow$$

$$\exp\left[(c(\theta_0) - c(\theta_1))\, d(\vec{x})\right] \leq k_1$$

$$\Downarrow$$

$$(c(\theta_0) - c(\theta_1))\, d(\vec{x}) \leq k_2$$

- Since $c(\theta)$ is increasing and we have assumed that $\theta_1 > \theta_0$, we know that $c(\theta_0 - c(\theta_1))$ is negative. Dividing both sides by $c(\theta_0 - c(\theta_1))$ gives us

$$d(\vec{x}) \geq k_3.$$

- We now solve for k_3 by setting $P(d(\vec{X}) \geq k_3;\theta_0) = \alpha$.

  Finding k_3 will involve the distribution of $d(\vec{X})$ under the assumption that the null hypothesis ($H_0 : \theta = \theta_0$) and will not involve θ_1. Hence, this most powerful (best) test provided by the N-P Lemma for $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ will be UMP for $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$.

■

Example 6.3.4

Let $X_1, X_2, \ldots, X_n$ be a random sample from the $N(0,\sigma^2)$ distribution.

Find a UMP test of size α for $H_0 : \sigma = \sigma_0$ versus $H_1 : \sigma > \sigma_0$.

The pdf is

$$f(x;\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}x^2}.$$

The joint pdf is

$$\Rightarrow f(\vec{x};\sigma) = (2\pi)^{-n/2}(\sigma^2)^{-n/2} \cdot \exp\left[-\frac{1}{2\sigma^2}\sum x_i^2\right].$$

This is an exponential family factorization with

$$\begin{aligned} a(\sigma) &= (2\pi)^{-n/2}(\sigma^2)^{-n/2} \\ b(\vec{x}) &= 1 && (\Leftarrow \text{the } (2\pi)^{-n/2} \text{ could go here}) \\ c(\sigma) &= -\tfrac{1}{2\sigma^2} && (\Leftarrow \text{the negative could go with } d(\vec{x})) \\ d(\vec{x}) &= \textstyle\sum_{i=1}^n x_i^2. \end{aligned}$$

Note that $c(\sigma)$ is an increasing function of σ.

So, a UMP test of size α is to reject H_0 if $\sum X_i^2 \geq k$ where k is chosen such that $P(\sum X_i^2 \geq k; \theta_0) = \alpha$.

(This is from Theorem 6.3.2 but we honestly think it is easier to figure out by just doing the same things we did in Section 6.3.1!)

Now, we find k.

When H_0 is true, $\sigma = \sigma_0$. Under this assumption, we need the distribution of $\sum X_i^2$.

A standard normal squared is a chi-squared. Specifically,

$$X_i \sim N(0,\sigma_0^2) \;\Rightarrow\; X_i/\sigma_0 \sim N(0,1) \;\Rightarrow\; (X_i/\sigma_0)^2 \sim \chi^2(1) \;\Rightarrow\; \textstyle\sum_{i=1}^n (X_i/\sigma_0^2) \sim \chi^2(n)$$

Thus, we have

$$\begin{aligned} \alpha &= P\left(\sum_{i=1}^n X_i^2 \geq k;\sigma_0\right) = P\left(\frac{\sum_{i=1}^n X_i^2}{\sigma_0^2} \geq \tfrac{k}{\sigma_0^2};\sigma_0\right) \\ &= P\left(\sum_{i=1}^n \left(\frac{X_i}{\sigma_0}\right)^2 \geq \frac{k}{\sigma_0^2};\sigma_0\right) = P(W \geq k/\sigma_0^2) \end{aligned}$$

where $W \sim \chi^2(n)$.

Should we define a new constant $k_1 = k/\sigma_0^2$? We should if we want to give the rejection rule in terms of the statistic $\sum(X_i/\sigma_0)^2$. We will keep our rejection rule in terms of $\sum X_i^2$.

We have

$$\alpha = P(W \geq k/\sigma_0^2).$$

What number is a $\chi^2(n)$ random variable larger than with probability α? In our notation, it is the critical value $\chi^2_{\alpha,n}$. So, we must have that

$$k/\sigma_0^2 = \chi^2_{\alpha,n}$$

which implies that $k = \sigma_0^2 \chi^2_{\alpha,n}$.

We know that $k_1 = \chi^2_\alpha(n)$.

So, the UMP test of size α is to

"Reject H_0, in favor of H_1, if $\sum_{i=1}^n X_i^2 \geq \sigma_0^2 \chi^2_{\alpha,n}$."

6.4 Generalized Likelihood Ratio Tests (GLRTs)

As has been mentioned, the hypothesis tests already described in this Chapter are examples of "likelihood ratio tests" since they involve ratios of joint pdfs which are equivalent to ratios of likelihood functions. As we saw in Example 6.3.3, it is not always possible to find a uniformly most powerful test. In this Section we will consider a **generalized likelihood ratio test** (GLRT) that will be applicable in a much wider range of situations. Consider it "to the UMP test as the MLE estimator is to the UMVUE". While the UMVUE is awesome in theory, in real life you are probably going to reach for an MLE.

Recall that our general hypotheses statement is

$$H_0 : \theta \in \Theta_0 \qquad \text{versus} \qquad H_1 : \theta \in \Theta \setminus \Theta_0$$

Definition 6.4.1

Let $L(\theta)$ be a likelihood for a random sample having joint pdf $f(\vec{x};\theta)$ for $\theta \in \Theta$.

The **(generalized) likelihood ratio (GLR)** is defined to be

$$\lambda(\vec{x}) = \frac{\max\limits_{\theta\in\Theta_0} L(\theta)}{\max\limits_{\theta\in\Theta} L(\theta)} = \frac{L(\widehat{\theta}_0)}{L(\widehat{\theta})}$$

where $\widehat{\theta}$ denotes the usual "unrestricted MLE" and $\widehat{\theta}_0$ denoted the MLE when restricted to the region where H_0 is true.[1]

[1] As noted in Chapter 4, these maximums should really be supremums. We write them as maximums here to make the definition more approachable to a wider audience.

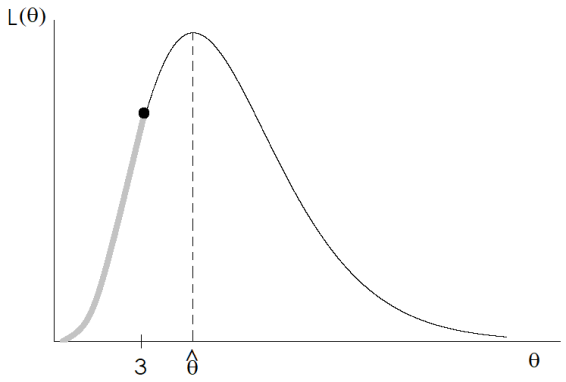
You may be wondering why the likelihood in the denominator is being maximized over all $\theta \in \Theta$ as opposed to $\theta \in \Theta \setminus \Theta_0$. The GLR certainly could have been defined in this way but the way we have defined it here will give cleaner results for which some asymptotic properties are known. $\lambda(\vec{X})$ is simply going to be a test statistic that we will use in a hypothesis test. No one is claiming that the GLR defined in this way is going to give a best or optimal test.

The Restricted MLE

Suppose that we wish to test $H_0 : \theta \leq 3$ versus $H_1 : \theta > 3$ and we have a likelihood function $L(\theta)$.

We maximize $L(\theta)$ by setting $\frac{d}{d\theta}L(\theta) = 0$.

For the purpose of illustration, let us first assume that the likelihood function is unimodal. That is, there is only one solution to $\frac{d}{d\theta}L(\theta) = 0$ and it corresponds to a maximum and not a minimum. (This is the case for the majority of the "nice known and named" distributions.) The solution is the MLE which we call $\widehat{\theta}$. In particular, suppose that the graph of $L(\theta)$ looks like this.

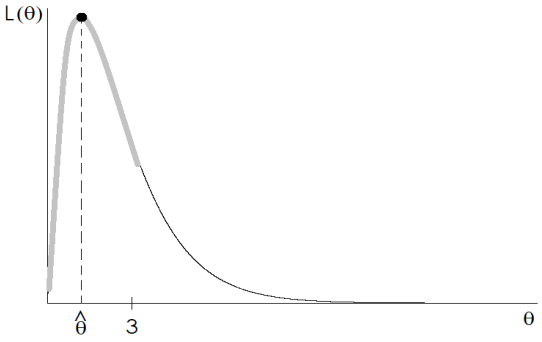


Since there are no additional turning points for $L(\theta)$, the restricted MLE will have to be at an endpoint of the region defined by H_0. In the case of the above figure, the maximum over the region where $\theta \leq 3$ (where H_0 is true) occurs at $\theta = 3$. In our new notation, we say that the restricted MLE is $\widehat{\theta}_0 = 3$.

Definition 6.4.2

The **restricted MLE** is the point that maximizes the likelihood when we restrict our attention to the region where H_0 is true.

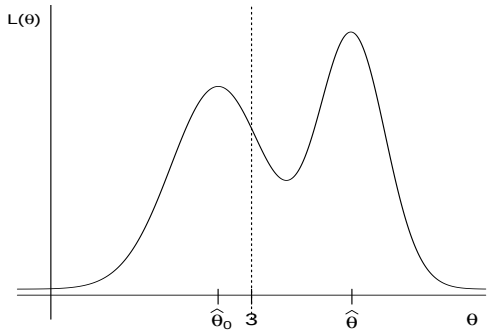
It is important to note that the MLE is a random variable and could end up on the other side of 3 like this.



In this case, the MLE when restricted to the region $\theta \leq 3$ is the same as the unrestricted MLE. Overall, the restricted MLE for this example is written as follows.

$$\widehat{\theta}_0 = \begin{cases} 3 & , \text{ if } \widehat{\theta} \geq 3 \\ \widehat{\theta} & , \text{ if } \widehat{\theta} < 3 \end{cases}$$

Of course it's possible that the likelihood function is multimodal. (This behavior would be reflected in multiple solutions to $\frac{d}{d\theta}L(\theta) = 0$.) In this case, the restricted MLE may be an actual local maximum.



We are now ready to define a generalized likelihood ratio test.

The Generalized Likelihood Ratio Test

Suppose that $X_1, X_2, \ldots, X_n$ have joint pdf $f(\vec{x}; \theta)$ with $\theta \in \Theta$. Let $\Theta_0 \subseteq \Theta$.

The **generalized likelihood ratio test** for

$$H_0 : \theta \in \Theta_0 \text{ versus } H_1 : \theta \in \Theta \setminus \Theta_0$$

is to reject H_0, in favor of H_1 if

$$\lambda(\vec{X}) \leq k$$

where $\lambda(\vec{X})$ is the GLR defined in Definition 6.4.1 and k is chosen such that

$$\max_{\theta \in \Theta_0} P(\lambda(\vec{X}) \leq k; \theta) = \alpha$$

As with the Neyman-Pearson Lemma, the direction of the inequality makes sense. The numerator is the maximized likelihood of seeing the given data under the assumption that H_0 is true. If this is small, we will tend to not believe that H_0 is true and will want to reject it.

Let's do an example.

Example 6.4.1

Let $X_1, X_2, \ldots, X_n$ be a random sample from the $N(\mu, \sigma^2)$ distribution where σ^2 is known.

Derive a GLRT of size α for

$$H_0 : \mu = \mu_0 \qquad \text{versus} \qquad H_1 : \mu \neq \mu_0.$$

(Note: There is no UMP for this problem as is often the case for a two-sided alternative hypothesis.)

The joint pdf is

$$f(\vec{x}, \mu) = (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2}.$$

Since a likelihood is any function proportional to the joint pdf, let's take

$$L(\mu) = e^{-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2}.$$

(Remember: σ^2 is assumed to be known so the $(2\pi\sigma^2)^{-n/2}$ can just be dropped as a constant of proportionality.)

We already know the usual (unrestricted) MLE for μ: $\widehat{\mu} = \overline{X}$.

Question: Now what maximizes $L(\mu)$ when H_0 is true?

Answer: This is easy since H_0 contains only one point. We are maximizing over only one value!

So,

$$\max_{\mu=\mu_0} L(\mu) = L(\mu_0).$$

The GLR is now

$$\lambda(\vec{x}) = \frac{e^{-\frac{1}{2\sigma^2}\sum(x_i-\mu_0)^2}}{e^{-\frac{1}{2\sigma^2}\sum(x_i-\overline{x})^2}} = e^{-\frac{1}{2\sigma^2}\left[\sum(x_i-\mu_0)^2-\sum(x_i-\overline{x})^2\right]}.$$

Since we're going to have to compute a probability $\max_{\mu=\mu_0} P(\lambda(\vec{X}) \leq k; \mu) = P(\lambda(\vec{X}) \leq k; \mu_0)$, let's try to simplify things a bit.

Note that

$$\begin{aligned}\sum(x_i-\mu_0)^2-\sum(x_i-\overline{x})^2 &= \sum x_i^2-2\mu_0\sum x_i+n\mu_0^2-\sum x_i^2+2\overline{x}\sum x_i-n\overline{x}^2\\ &= -2\mu_0\sum x_i+n\mu_0^2+2\overline{x}\sum x_i-n\overline{x}^2\end{aligned}$$

This kind of looks like something squared. In fact, if we pull the n out we have

$$\begin{aligned}n(-2\mu_0\tfrac{1}{n}\sum x_i+\mu_0^2+2\overline{x}\tfrac{1}{n}\sum x_i-\overline{x}^2) &= n(-2\mu_0\overline{x}+\mu_0^2+2\overline{x}^2-\overline{x}^2)\\ &= n(-2\mu_0\overline{x}+\mu_0^2+\overline{x}^2\\ &= n(\overline{x}-\mu_0)^2\end{aligned}$$

Excellent!

The GLR is

$$\lambda(\vec{x})=\exp\left[\frac{-n(\overline{x}-\mu_0)^2}{2\sigma^2}\right]$$

Usually, we would set this less than or equal to some cutoff k and then try to isolate the X's on one side. However, there is a standard normal hidden in there. (We will come to this conclusion if we just proceed as usual. We would simplify things until the Xs are isolated but then we would find ourselves with a probability involving $\overline{X}$. Under H_0, $\overline{X}\sim N(\mu_0,\sigma^2)$. To compute the probability defining α we would begin by standardizing $\overline{X}$. In doing this, we would see that we are building things back up to, essentially, where we started.)

$$\begin{gathered}\exp\left[\frac{-n(\overline{x}-\mu_0)^2}{2\sigma^2}\right]\leq k\\ \Downarrow\\ \frac{-n(\overline{x}-\mu_0)^2}{2\sigma^2}\leq k_1\\ \Downarrow\\ \frac{(\overline{x}-\mu_0)^2}{\sigma^2}\geq k_2\end{gathered}$$

This can be rewritten as

$$\left(\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}\right)^2 \geq k_2$$

We will continue in two different ways. The first will involve leaving the square in place.

$$\begin{aligned}
\alpha &= \max P(\text{ Type I Error }) \\
&= \max P(\text{ Reject } H_0 \text{ when it's true }) \\
&= \max_{\mu=\mu_0} P(\text{ Reject } H_0; \mu\,) \\
&= P(\text{ Reject } H_0; \mu_0\,) \\
&= P(\lambda(\vec{X}) \leq k; \mu_0) \\
&= P\left(\left(\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}\right)^2 \geq k_2; \mu_0\right)
\end{aligned}$$

For this probability, we are under the assumption that $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu_0, \sigma^2)$. This means that $\overline{X} \sim N(\mu_0, \sigma^2/n)$ and, therefore, that

$$\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1).$$

We saw in Exercise 6, that a $N(0,1)$ random variable, when squared, has a $\chi^2(1)$ distribution.

$$\left(\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}\right)^2 \sim \chi^2(1).$$

So, we have

$$\alpha = P(W \geq k_2)$$

where $W \sim \chi^2(1)$. This means that $k_2 = \chi^2_{\alpha,1}$.

In conclusion, the generalized likelihood ratio test for

$$H_0 : \mu = \mu_0 \qquad \text{versus} \qquad H_1 : \mu \neq \mu_0$$

is to

$$\boxed{\text{“Reject } H_0\text{, in favor of } H_1\text{, if } \left(\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}\right)^2 \geq \chi^2_{\alpha,1}.\text{”}}$$

We could also solve for $\overline{X}$ and give the equivalent rejection rule:

$$\text{“Reject } H_0\text{, in favor of } H_1\text{, if either } \quad \overline{X} \geq \mu_0 + \sqrt{\chi^2_{\alpha,1}}\,\frac{\sigma}{\sqrt{n}} \quad \text{or} \quad \overline{X} \leq \mu_0 - \sqrt{\chi^2_{\alpha,1}}\,\frac{\sigma}{\sqrt{n}}.\text{”}$$

We could even have taken the square root back when our rejection rule had the form

$$\left(\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}\right)^2 \geq k_2$$

$$\Downarrow$$

$$\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \geq k_3 \qquad \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \leq -k_3$$

where $k_3 = \sqrt{k_2}$.

We get

$$\begin{aligned}
\alpha &= \max P(\text{ Type I Error }) \\
&= \max P(\text{ Reject } H_0 \text{ when it's true }) \\
&= \max_{\mu=\mu_0} P(\text{ Reject } H_0; \mu\,) \\
&= P(\text{ Reject } H_0; \mu_0\,) \\
&= P(\lambda(\vec{X}) \leq k; \mu_0) \\
&= P\left(\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} > k_3 \text{ or } \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} < -k_3; \mu_0\right) \\
&= P(Z > k_3 \text{ or } Z < -k_3) \\
&= P(Z > k_3) + P(Z < -k_3) \qquad \Leftarrow \text{ disjoint events}
\end{aligned}$$

where $Z \sim N(0,1)$.

We can achieve this by putting area $\alpha/2$ on each side under the standard normal pdf which means we need to take $k_3 = z_{\alpha/2}$.

Thus, we have a third rejection rule which is to

"Reject H_0, in favor of H_1, if either $\overline{X} \geq \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ or $\overline{X} \leq \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$."

Have we come up with something different than our second rejection rule? That is, do we have that $z_{\alpha/2} = \sqrt{\chi^2_{\alpha,1}}$? The answer is yes and we leave this as an exercise for you!

Let's try an example that has an unknown parameter in the indicator.

Example 6.4.2

Let $X_1, X_2, \ldots, X_n$ be a random sample from the $unif(0, \theta]$ distribution.

(Note that we closed the right side of the interval. We only did this so that we won't have a problem with a "max", but this wouldn't matter at all if we were using the more general definition of the GLR that uses supremums.)

Find the GLRT of size α for $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.

The pdf is

$$f(x;\theta) = \frac{1}{\theta} I_{(0,\theta]}(x).$$

The joint pdf is

$$f(\vec{x};\theta) = \theta^{-n} \prod_{i=1}^{n} I_{(0,\theta]}(x_i) = \theta^{-n} I_{(0,\theta]}(x_{(n)})$$

A likelihood is

$$L(\theta) = \theta^{-n} \prod_{i=1}^{n} I_{(0,\theta]}(x_i) = \theta^{-n} I_{(0,\theta]}(x_{(n)}).$$

We can not drop the indicator function as because it depends on θ. However, per our discussion in Example 5.2.5 of Chapter 5, we will remove it for taking logs and derivatives but remember that it is part of the likelihood.

Taking the log of $L(\theta) = \theta^{-n}$ gives us

$$\ell(\theta) = -n \ln \theta.$$

The derivative with respect to θ is

$$\ell'(\theta) = -\frac{n}{\theta}.$$

If we set this equal to zero in an attempt to find a critical value for the likelihood, we see that there is no solution. This can not be zero as long as we have a sample of size 1 or more. The goal remains unchanged. We wish to maximize the likelihood.

Note that $L(\theta) = \theta^{-n}$ is a decreasing function of θ. So, in order to maximize it, we want to take θ as small as possible. According to the indicator, all data is between 0 and θ. Thus, the smallest possible value for θ is the maximum data point.

$$\widehat{\theta} = X_{(n)}$$

Let's move on to the restricted likelihood.

As in the previous example, since H_0 consists of only one point: $\theta = \theta_0$, the maximum of $L(\theta)$ restricted to this one point set is simply $L(\theta_0)$.

We now have the generalized likelihood ratio as

$$\lambda(\vec{x}) = \frac{\theta_0^{-n}\, I_{(0,\theta_0]}(x_{(n)})}{x_{(n)}^{-n}\, I_{(0,x_{(n)}]}(x_{(n)})} = \left(\frac{x_{(n)}}{\theta_0}\right)^n I_{(0,\theta_0]}(x_{(n)}).$$

As usual, we will reject H_0 if

$$\left(\frac{x_{(n)}}{\theta_0}\right)^n I_{(0,\theta_0]}(x_{(n)}) \leq k$$

where k is such that

$$P\left(\left(\frac{X_{(n)}}{\theta_0}\right)^n I_{(0,\theta_0]}(X_{(n)}) \leq k; \theta_0\right) = \alpha$$

Under the assumption that H_0 is true, that indicator is always 1, so we can drop it.

$$\alpha = P\left(\left(\frac{X_{(n)}}{\theta_0}\right)^n \leq k; \theta_0\right) = P\left(X_{(n)} \leq \theta_0 k^{1/n}; H_0\right) = P\left(X_{(n)} \leq k_1; \theta_0\right)$$

Finally, we solve for k_1:

$$\begin{aligned} \alpha &= P\left(X_{(n)} \leq k_1; H_0\right) \\ &= \left[P\left(X_1 \leq k_1; H_0\right)\right]^n \\ &= \tfrac{k_1}{\theta_0} \end{aligned}$$

This implies that

$$k_1 = \theta_0 \alpha^{1/n}.$$

So, the GLRT of size α, for $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ for a random sample of size n from the $unif(0, \theta]$ distribution, is to reject H_0 in favor of H_1 if

$$X_{(n)} \leq \theta_0 \alpha^{1/n}.$$

Right?

Did you fall for that? Something doesn't seem quite right. Does that rejection rule make sense? One would certainly think that is is not true that θ, the upper limit of the support set for the sample, is equal to θ_0 if we happened to observe $X_{(n)} > \theta_0$. Shouldn't we be rejecting for some values of $X_{(n)}$ that are too large? We don't just want to say: "Well of course we would automatically reject if $X_{(n)} > \theta_0$" because making up rules to suit our needs will affect the size of the test that we worked so hard to obtain.

Going back to the original rejection rule, we reject H_0 if

$$\left(\frac{X_{(n)}}{\theta_0}\right)^n I_{(0,\theta_0]} \leq k.$$

This is not equivalent to rejecting H_0 if

$$\left(\frac{X_{(n)}}{\theta_0}\right)^n \leq k_1$$

for some other constant k_1. The indicator involves θ which is not a constant. We used the fact that the indicator is 1 under the assumption that H_0 is true in order to compute a probability. However, this does not become part of the rejection rule.

Our level of significance calculation was correct because it was done under the assumption that H_0 is true. There, we determined that $k_1 = \theta_0\, \alpha^{1/n}$.

Since $k = (k_1/\theta_0)^n$, this becomes, "reject H_0 if":

$$\left(\frac{X_{(n)}}{\theta_0}\right)^n I_{(0,\theta_0]}(X_{(n)}) \leq \left(\frac{k_1}{\theta_0}\right)^n = \left(\frac{\theta_0 \alpha^{1/n}}{\theta_0}\right)^n = \alpha.$$

For a non-trivial α (ie: $\alpha > 0$), if $X_{(n)}$ is greater than θ_0, the left hand side of this inequality will be zero, hence less than α, hence we will reject, as desired.

We have selected some pretty clean examples for our GLRTs. In real life we may not be so lucky with our computations. In the next Section, we will see an asymptotic result that can sometimes allow us to get approximate large sample GLRTs.

6.5 Wilks' Theorem

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with pdf $f(x;\theta)$ for θ in a k-dimensional parameter space Θ.

Consider finding a GLRT of size α for

$$H_0 : \theta \in \Theta_0 \qquad \text{versus} \qquad H_1 : \theta \in \Theta \setminus \Theta_0$$

where Θ_0 fixes some parameters in the model and H_1 is "not H_0". For example, if $k = 5$ and $\theta = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$, H_0 might have the form

$$H_0 : \theta_2 = 3,\ \theta_5 = 1.$$

Let $L(\theta)$ be a likelihood function for the data and let $\lambda(\vec{x})$ be the generalized likelihood ratio defined in the previous section.

Wilks' Theorem

Under certain regularity conditions (given in Section 6.5.1) we have

$$-2\ln\lambda(\vec{X}) \xrightarrow{d} \chi^2(r - r_0)$$

when H_0 is true. Here, r is number of parameters in the model and r_0 is the number of free parameters under the assumption that H_0 is true.

Before actually using Wilks' Theorem, let's look at a few examples to make sure we know how to specify r and

r_0.

Example 6.5.1

Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ with σ^2 known.

Consider testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$.

Here $r = 1$ as there is one parameter in the model (σ^2 is not considered a parameter because it is just a known constant) and $r_0 = 0$ since, when H_0 is true, the parameter is known and there are no more free parameters.

Thus, by Wilks' Theorem we have that

$$-2 \ln \lambda(\vec{X}) \xrightarrow{d} \chi^2(1)$$

under H_0.

Example 6.5.2

Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$.

Consider testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$.

Here $r = 2$ as there are two parameters in the model and $r_0 = 1$ since, when H_0 is true, the parameter μ is known and there is still one free parameter σ^2.

Thus, by Wilks' Theorem we have that

$$-2 \ln \lambda(\vec{X}) \xrightarrow{d} \chi^2(1)$$

under H_0.

Example 6.5.3

Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$.

Consider testing $H_0 : \mu = \mu_0, \sigma^2 = \sigma_0^2$ versus H_1 : Not H_0.

Here $r = 2$ as there are two free parameters in the model and $r_0 = 0$ since, when H_0 is true, there are no unknown free parameters.

Thus, by Wilks' Theorem we have that

$$-2\ln\lambda(\vec{X}) \xrightarrow{d} \chi^2(2)$$

under H_0.

We will now see some examples of Wilks' Theorem in action.

Example 6.5.4

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from the $N(\mu, \sigma^2)$ distribution with σ^2 known.

Find an approximate large sample GLRT of size α for

$$H_0 : \mu = \mu_0 \qquad \text{versus} \qquad H_1 : \mu \neq \mu_0.$$

We already did this problem exactly in Example 6.4.1, but we are going to see how the approximate large sample result compares.

We begin by computing the GLR. We did this in Example 6.4.1 and got

$$\lambda(\vec{x}) = \exp\left[\frac{-n(\overline{x} - \mu_0)^2}{2\sigma^2}\right].$$

For the next part of the test, we set this less than or equal to a constant k and simplified the inequality until we were able to compute the probability of making a Type I error. This was a lot of work and can be worse or impossible for other examples.

According to Wilks' Theorem,

$$-2\ln\lambda(\vec{X}) \xrightarrow{d} \chi^2(1).$$

So, for large sample sizes, we will say that

$$-2\ln\lambda(\vec{X}) \overset{approx}{\sim} \chi^2(1).$$

We still want to reject H_0, in favor of H_1 if $\lambda(\vec{X})$, the GLR, is "small". Instead of trying to simplify things and solve for the X's, we will go the opposite direction and build this up to $-2\ln\lambda(\vec{X})$. We have

$$\lambda(\vec{X}) \leq k,$$

which implies that

$$\ln \lambda(\vec{X}) \leq k_1.$$

This implies that

$$-2 \ln \lambda(\vec{X}) \geq k_2.$$

To find k_2 we write

$$\begin{aligned}
\alpha &= \max P(\text{ Type I Error }) \\
&= \max P(\text{ Reject } H_0 \text{ when true }) \\
&= \max_{\mu=\mu_0} P(\text{ Reject } H_0; \mu\,) \\
&= P(\text{ Reject } H_0; \mu_0\,) \\
&= P(\lambda(\vec{X}) \leq k; \mu_0) \\
&= P(-2 \ln \lambda(\vec{X}) \geq k_2; \mu_0) \\
&\approx P(W \geq k_2)
\end{aligned}$$

where $W \sim \chi^2(1)$.

We need k_2 to be the critical value that captures area α in the upper tail of a $\chi^2(1)$ pdf. This means that $k_2 = \chi^2_{\alpha,1}$.

So, the approximate large sample GLRT of size α for this problem is to

"Reject H_0, in favor of H_1, if $-2 \ln \lambda(\vec{X}) \geq \chi^2_{\alpha,1}$."

This is a fine rejection rule. We have provided a test. However, note that

$$\lambda(\vec{X}) = \exp\left[\frac{-n(\overline{X}-\mu_0)^2}{2\sigma^2}\right] \quad \Rightarrow \quad -2 \ln \lambda(\vec{X}) = \frac{n(\overline{X}-\mu_0)^2}{\sigma^2} = \left(\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}\right)^2.$$

Thus, $-2\ln\lambda(\vec{X})$ has an exact $\chi^2(1)$ distribution, even for small samples!

6.5.1 Proof of Wilks' Theorem*

We will prove Wilks' Theorem in the case that θ is one-dimensional and H_0 has the form $H_0 : \theta = \theta_0$. Note that, in this case, Wilks' Theorem says that $-2\ln\lambda(\vec{X}) \xrightarrow{d} \chi^2(1)$.

Regularity Conditions: If $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$, we assume the following.

1. When indexed by θ, the densities $\{f(x;\theta) : \theta \in \Theta\}$ have common support. (i.e. No parameters in the indicators!)
2. The true parameter that generated the data under H_0, namely θ_0, is in the interior of the parameter space.
3. The derivative $\frac{\partial}{\partial\theta} f(\vec{x};\theta)$ exists. (Exception: The derivative may not exist on a set of probability measure zero.)
4. The Fisher information, $I_1(\theta_0)$ is non-zero and finite.
5. The third derivative of the log-likelihood $\ell(\theta)$ exists and satisfies

$$|\ell'''(\theta)| \leq C(x)$$

for all θ in some neighborhood of θ_0 and for some $C(x)$ satisfying $\mathsf{E}[C(X_1)] < \infty$ where $X_1 \sim f(x;\theta_0)$.

Little Oh and Big Oh:

In Section 5.4.1 of Chapter 5, we played a little fast and loose with remainder terms when proving that many maximum likelihood estimators are asymptotically normal. In this Section we will be more technical. To this end, we will need some definitions.

Definition 6.5.1

A function g is called "little oh of h" if

$$\lim_{h \searrow 0} \frac{g(h)}{h} = 0.$$

i.e. It is going to zero and going fast enough to counteract a denominator that is also going to zero.

We write $g(h) = o(h)$.

Example 6.5.5

$g(h) = h^2$ is $o(h)$ since

$$\lim_{h \searrow 0} \frac{g(h)}{h} = \lim_{h \searrow 0} \frac{h^2}{h} = \lim_{h \searrow 0} h = 0$$

Indeed, $h^k = o(h)$ for any $k > 1$.

Suppose that $g_1(h)$ and $g_2(h)$ are two functions that are $o(h)$. Then $g_1(h) + g_2(h)$ is $o(h)$ since

$$\lim_{h \searrow 0} \frac{g_1(h)g_2(h)}{h} = \lim_{h \searrow 0} g_1(h) \cdot \frac{g_2(h)}{h} = 0 \cdot 0 = 0.$$

We write this idea symbolically as

$$o(h) + o(h) = o(h)$$

since the sum of $o(h)$ functions is just another $o(h)$ function. It does not make much sense to write $2o(h)$. The functions g_1 and g_2 are not necessarily the same. The $o(h)$ terms serves to "collect" all of the negligible terms in an expression.

Example 6.5.6

$$e^{\lambda h} = 1 + \lambda h + \frac{(\lambda h)^2}{2!} + \frac{(\lambda h)^3}{3!} + \cdots = 1 + \lambda h + o(h)$$

Definition 6.5.2

A sequence of random variables $\{X_n\}$ is "little oh p of a_n" if

$$X_n/a_n \xrightarrow{P} 0.$$

We write $X_n = o_p(a_n)$.

Example 6.5.7

Suppose that $X_n \xrightarrow{P} 0$. Then $X_n = o_p(1)$.

Definition 6.5.3

A function g is called "big oh of h" if

$$\frac{g(h)}{h}$$

is bounded in the limit. That is, there exists a constant c and a constant h_0 such that

$$|g(h)| \leq c\, h$$

for all $h \geq h_0$.

We write $g(h) = O(h)$.

Example 6.5.8

Definition: A sequence of random variables $\{X_n\}$ is **bounded in probability** if, for every $\varepsilon > 0$, there exists and $M = M(\varepsilon)$ such that

$$P(|X_n| > M) < \varepsilon \;\text{ for all } n.$$

~~~~~~

Definition: A sequence of random variables $\{X_n\}$ is "big oh p of a_n" if

$$X_n/a_n$$

is bounded in probability.

We write $X_n = O_p(a_n)$.
~~~~~~

Exercise for you:

Show that

$$O_p(1) \cdot o_p(1) = o_p(1).$$

(As described before, the two different $o_p(1)$ terms are representing different functions. This is a symbolic way of saying that a $O_p(1)$ function multiplied by a $o_p(1)$ function gives another function that is still $o_p(1)$.)

Proof of Wilks' Theorem for One-Dimensional θ: $(H_0 : \theta = \theta_0)$

1. Let $\ell(\theta)$ be the usual log-likelihood. Note that, although the notation is suppressed, this is a function of $\vec{X} = (X_1, X_2, \ldots, X_n)$.

   Let $\widehat{\theta}_n$ be the MLE for θ.

   Let θ_0 be the specific θ that generated the X's.

   (This means that H_0 is true, so any convergence we are showing is going to hold "under H_0".)

2. Consider the following Taylor polynomial for $\ell(\theta_0)$, centered at $\widehat{\theta}_n$:

   $$\ell(\theta_0) = \ell(\widehat{\theta}_n) + \ell'(\widehat{\theta}_n)(\theta_0 - \widehat{\theta}_n) + \frac{1}{2}\ell''(\widehat{\theta}_n)(\theta_0 - \widehat{\theta}_n)^2 + \frac{1}{6}\ell'''(\widetilde{\theta}_n)(\theta_0 - \widehat{\theta}_n)^3$$

   for some $\widetilde{\theta}_n$ (random) between θ_0 (fixed) and $\widehat{\theta}_n$ (random).

3. Note that

   - $\ell'(\widehat{\theta}_n) = 0$ by definition of $\widehat{\theta}_n$
   - $\widehat{\theta}_n \xrightarrow{P} \theta_0$ and, in particular, $(\widehat{\theta}_n - \theta_0)^3 \xrightarrow{P} 0$.
   - Regularity condition #5 implies that $\ell'''(\widetilde{\theta}_n)$ is bounded in probability (See Exercise 3.) so we can say that $\frac{1}{6}\ell'''(\widetilde{\theta}_n) = O_p(1)$.

   Our Taylor polynomial becomes

   $$\begin{aligned} \ell(\theta_0) &= \ell(\widehat{\theta}_n) + 0 + \tfrac{1}{2}\ell''(\widehat{\theta}_n)(\theta_0 - \widehat{\theta}_n)^2 + O_p(1) \cdot o_p(1) \\ &= \ell(\widehat{\theta}_n) + 0 + \tfrac{1}{2}\ell''(\widehat{\theta}_n)(\theta_0 - \widehat{\theta}_n)^2 + o_p(1), \end{aligned}$$

   which implies that

   $$\ell(\theta_0) - \ell(\widehat{\theta}_n) = \frac{1}{2}\ell''(\widehat{\theta}_n)(\theta_0 - \widehat{\theta}_n)^2 + o_p(1).$$

4. Now

$$\begin{aligned} -2\ln\lambda(\vec{X}) &= -2\ln\tfrac{L(\theta_0)}{L(\widehat{\theta}_n)} = -2[\ell(\theta_0)-\ell(\widehat{\theta}_n)] \\ &= -\ell''(\widehat{\theta}_n)(\theta_0-\widehat{\theta}_n)^2+o_p(1) \\ &= -\tfrac{1}{n}\ell''(\widehat{\theta}_n)[\sqrt{n}(\widehat{\theta}_n-\theta_0)]^2+o_p(1) \end{aligned}$$

5. Note that

   - $-\frac{1}{n}\ell''(\theta_0) \xrightarrow{P} I_1(\theta_0)$ by the Weak Law of Large Numbers
   - $\sqrt{n}(\widehat{\theta}_n-\theta_0) \xrightarrow{d} N(0,[I_1(\theta_0)]^{-1})$ by super groovy properties of MLEs

   We can get convergence of $-\frac{1}{n}\ell''(\widehat{\theta}_n)$ using the Mean Value Theorem:

$$\frac{1}{n}\ell''(\theta_0)-\frac{1}{n}\ell''(\widehat{\theta}_n) = -\frac{1}{n}\ell'''(\widetilde{\theta}_n)(\widehat{\theta}_n-\theta_0)$$

   for some $\widetilde{\theta}_n$ between θ_0 and $\widehat{\theta}_n$.

   Using the fact that $\ell'''(\widetilde{\theta}_n)$ is bounded in probability and the fact that $\widehat{\theta}_n \xrightarrow{P} 0$, we can show that the right-hand side converges in probability to 0.

   Thus, we have that

$$-\frac{1}{n}\ell''(\widehat{\theta}_n) = -\frac{1}{n}\ell''(\theta_0)+o_p(1)$$

   which implies that

$$-\frac{1}{n}\ell''(\widehat{\theta}_n) \xrightarrow{P} I_1(\theta_0).$$

6. All together now...

$$\begin{aligned} -2\ln\lambda(\vec{X}) &= -\tfrac{1}{n}\ell''(\widehat{\theta}_n)[\sqrt{n}(\widehat{\theta}_n-\theta_0)]^2+o_p(1) \\ &= \underbrace{\frac{-\frac{1}{n}\ell''(\widehat{\theta}_n)}{I_1(\theta_0)}}_{A}\cdot\left(\underbrace{\frac{\widehat{\theta_n}-\theta_0}{\frac{1}{\sqrt{nI_1(\theta_0)}}}}_{B}\right)^2+o_p(1) \end{aligned}$$

   - $A \xrightarrow{P} 1$
   - $B \xrightarrow{d} N(0,1)$
   - $B^2 \xrightarrow{d} \chi^2(1)$ by the continuous mapping theorem
   - $o_p(1) \xrightarrow{P} 0$ by definition

   So, by Slutzky's Theorem, we have that

$$-2\ln\lambda(\vec{X}) \xrightarrow{d} \chi^2(1)$$

   under H_0.

□

Chapter 6 Exercises

1. Let $z_{\alpha/2}$ be the critical value that cuts off area $\alpha/2$ under the standard normal pdf. Let $\chi^2_{\alpha,1}$ be the critical value that cuts off area α under the $\chi^2(1)$ pdf. Show that $z_{\alpha/2} = \sqrt{\chi^2_{\alpha,1}}$.
2. Consider a random sample $X_1, X_2, \ldots, X_n$ from a discrete distribution with pdf

$$f(x;\theta) = [\theta/(\theta+1)]^x/(\theta+1)\, I_{\{0,1,2,\ldots\}}(x)$$

with $\theta > 0$.

Consider testing the hypotheses

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta = \theta_1 \end{aligned}$$

for some fixed $\theta_1 > \theta$.

   (a). Find the best test of size (level of significance) α.

   (b). Is your test from part (a) a UMP (uniformly most powerful) test of size α for H_0 against the alternate hypothesis $H_1 : \theta > \theta_0$? Explain.

3. Suppose that the third derivative of the log-likelihood $\ell(\theta)$ exists and satisfies

$$|\ell'''(\theta)| \leq C(x)$$

for all θ in some neighborhood of θ_0 and for some $C(x)$ satisfying $\mathsf{E}[C(X_1)] < \infty$ where $X_1 \sim f(x;\theta_0)$. Show that $\ell'''(\widehat{\theta}_n)$, where $\widehat{\theta}_n$ is the MLE for θ, is bounded in probability.

4. Consider a random sample $X_1, X_2, \ldots, X_n$ from a discrete distribution with pdf

$$f(x;\theta) = [\theta/(\theta+1)]^x/(\theta+1)\, I_{\{0,1,2,\ldots\}}(x)$$

with $\theta > 0$. Does a UMP test of size α of $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$ exist? If not explain. If so, find it!

5. Consider the distribution with pdf

$$f(x;\theta) = 1 - \theta^2(x - \frac{1}{2}), \qquad 0 < x < 1, \qquad -1 < \theta < 1.$$

   (a). Find the best test of size α for

$$H_0 : \theta = 0 \qquad \text{versus} \qquad H_1 : \theta \neq 0$$

based on a sample of size 1.

   (b). Find (if it exists) a UMP (uniformly most powerful) test of size α of

$$H_0 : \theta = 0 \qquad \text{versus} \qquad H_1 : \theta \neq 0$$

based on a sample of size 1.

6. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the $N(0, \sigma^2)$ distribution.
   - Derive the UMP test of size α for $H_0 : \sigma^2 = \sigma_0^2$ versus $H_1 : \sigma^2 > \sigma_0^2$.
   - Express the power function of your test from part (a) in terms of the chi-squared distribution.

7. Suppose that we have a random sample of four observations from the distribution with pdf

$$f(x;\theta) = \frac{1}{2\theta^3} x^2 e^{-x/\theta} I_{(0,\infty)}(x)$$

   (a). Find the best (most powerful) test of size α of $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, assuming that $\theta_1 > \theta_0$. Give your answer in terms of a chi-squared critical value.

   (b). Is your test uniformly most powerful for the alternative hypothesis $H_1 : \theta > \theta_0$?

8. Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from the $N(\mu, 1)$.

   (a). Find the UMP test of size α for $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$.

   (b). Explain why there is no UMP test for $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$.

9. Consider a random sample $X_1, \ldots, X_n$ from a distribution with pdf $f(x;\theta) = \theta(1-x)^{\theta-1}$, $0 < x < 1$, $\theta > 0$.

   Give the form of the GLRT for testing $H_o : \theta = 1$ against $H_1 : \theta \neq 1$?

10. Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from the distribution with pdf

$$f(x;\theta) = \theta x^{\theta-1} \cdot I_{(0,1)}(x).$$

    Derive the GLRT of size α for testing

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta \neq \theta_0,$$

    approximating the appropriate critical value based on a large sample size.

11. Let $X_1, X_2, \ldots, X_n$ be a random sample from the $N(\mu, \sigma^2)$ distribution where σ^2 is known. Find the GLR (generalized likelihood ratio) for testing the hypotheses

$$H_0 : \mu \leq \mu_0 \qquad \text{versus} \qquad H_1 : \mu > \mu_0.$$

    (Do not find the test, only the GLR!)

    (*Hint: The GLR will be defined in two pieces.* $\lambda = \lambda(\vec{x})$ = *"something when* $\overline{x}$ *is something" and "something else when* $\overline{X}$ *is something else".*)

12. Let $X_1, X_2, \ldots, X_n$ be a random sample from the uniform distribution on $(0, \theta]$. Find the exact (not asymptotic) distribution of $-2\ln\lambda(\vec{X})$ where $\lambda(\vec{X})$ is the GLR for $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.

Based on this, find another (ie: we already found one in class) GLRT of size α.

13. Let $X_1, X_2, \ldots, X_n$ be a random sample from the $N(\theta_1, \theta_2)$ distribution with $\Omega = \{(\theta_1, \theta_2) : -\infty < \theta_1 < \infty, \theta_2 > 0\}$. Find the GLRT of size α for testing

$$H_0 : \theta_1 = 0, \theta_2 > 0 \qquad \text{versus} \qquad H_1 : \theta_1 \neq 0, \theta_2 > 0$$

[OVER]

14. For exercise 13. above, show the following.

(a).
$$-2 \ln \lambda(\vec{X}) = n \ln \left\{ 1 + \frac{n\overline{X}^2}{\sum (X_i - \overline{X})^2} \right\} = n \ln \left\{ 1 + \frac{\overline{X}^2}{S^2} \right\}$$

(b). For large n, $\overline{X}^2/S^2$ is close to zero under $H_0 : \theta_1 = 0$, so, approximate the right hand side of the expression in part (a) with the first two terms of a Taylor series expanded about zero.

(c). Since n is large, replace n by $n - 1$ in your two-term Taylor series expansion, to conclude that
$$-2 \ln \lambda(\vec{X}) \approx \left(\frac{\overline{X}}{S/\sqrt{n-1}} \right)^2$$

(d). What is the distribution, under H_0, of the expression in parentheses (so, without the "squared") in part (c)?

(e). As n gets large, what is the distribution, under H_0, of the expression in parentheses in part (c) approaching?

(f). Conclude that the entire right hand side of the expression in part (c) is approaching a χ^2 distribution. What are the degrees of freedom for this distribution?

Appendix A Tables of Distributions

Table A.1: Common Discrete Distributions

Name	pdf (pmf)	Parameter Space	Mean	Variance	Moment Generating Function = $\mathsf{E}[e^{tX}]$
Uniform	$\frac{1}{n+1} I_{\{0,1,\ldots,n\}}(x)$	$n = 1, 2, \ldots$	$n/2$	$n(n+2)/12$	$\sum_{j=0}^{n} \frac{1}{n+1} e^{jt} = \frac{1-e^{(n+1)t}}{(n+1)(1-e^t)}$
Bernoulli	$p^x (1-p)^{1-x} I_{\{0,1\}}(x)$	$0 \le p \le 1$	p	$p(1-p)$	$(1-p) + pe^t$
Binomial	$\binom{n}{x} p^x (1-p)^{n-x} I_{\{0,1,\ldots,n\}}(x)$	$0 \le p \le 1$ $n = 1, 2, \ldots$	np	$np(1-p)$	$[(1-p) + pe^t]^n$
Hyper-geometric	$\frac{\binom{K}{x}\binom{M-K}{n-x}}{\binom{M}{n}} I_{\{0,1,\ldots,n\}}(x)$	$M = 1, 2, \ldots$ $K = 0, \ldots, M$ $n = 1, \ldots, M$	$n\frac{K}{M}$	$n\frac{K}{M}(1-\frac{K}{M})\frac{M-n}{M-1}$	not useful
Poisson	$\frac{e^{-\lambda}\lambda^x}{x!} I_{\{0,1,\ldots\}}(x)$	$\lambda > 0$	λ	λ	$exp[\lambda(e^t - 1)]$
Geometric	$p(1-p)^x I_{\{0,1,\ldots\}}(x)$	$0 < p < 1$	$(1-p)/p$	$(1-p)/p^2$	$\frac{p}{1-(1-p)e^t}$ for $t < -ln(1-p)$
Geometric	$p(1-p)^{x-1} I_{\{1,2,\ldots\}}(x)$	$0 < p < 1$	$1/p$	$(1-p)/p^2$	$\frac{pe^t}{1-(1-p)e^t}$ for $t < -ln(1-p)$
Negative Binomial	$\binom{r+x-1}{x} p^r (1-p)^x I_{\{0,1,\ldots\}}(x)$	$0 < p < 1$ and $r > 0$	$r(1-p)/p$	$r(1-p)/p^2$	$\left[\frac{p}{1-(1-p)e^t}\right]^r$ for $t < -ln(1-p)$
Negative Binomial	$\binom{x-1}{r-1} p^r (1-p)^{x-r} I_{\{r,r+1,\ldots\}}(x)$	$0 < p < 1$ and $r > 0$	r/p	$r(1-p)/p^2$	$\left[\frac{pe^t}{1-(1-p)e^t}\right]^r$ for $t < -ln(1-p)$
Beta-binomial	$\binom{n}{x} \frac{\mathcal{B}(x+a, n-x+b)}{\mathcal{B}(a,b)} I_{\{0,\ldots,n\}}(x)$	$a > 0, b > 0$ $n = 1, 2, \ldots$	$\frac{na}{a+b}$	$\frac{nab(n+a+b)}{(a+b)^2(a+b+1)}$	not useful
Logarithmic	$\frac{(1-p)^x}{(-x\,ln p)} I_{\{1,2,\ldots\}}(x)$	$0 < p < 1$	$\frac{(1-p)}{(-p\ ln p)}$	$\frac{(1-p)(1-p+ln p)}{-(p\ ln p)^2}$	$\frac{ln[1-(1-p)e^t]}{ln p}$ for $t < -ln(1-p)$
Discrete Pareto	$\frac{(1/x^{\gamma+1})}{\sum_{j=1}^{\infty}(1/j^{\gamma+1})} I_{\{1,2,\ldots\}}(x)$	$\gamma > 0$	$\frac{\sum_1^\infty (1/j)^\gamma}{\sum_1^\infty (1/j)^{\gamma+1}}$ for $\gamma > 1$		does not exist

Table A.2: Common Continuous Distributions

Name	pdf $= f(x)$ cdf $= F(x)$	Parameter Space	Mean	Variance	Moment Generating Function = $\mathsf{E}[e^{tX}]$
Uniform	$f(x) = \frac{1}{\beta} I_{(\alpha,\alpha+\beta)}(x)$	$-\infty < \alpha < \infty$ $\beta > 0$	$\alpha + \frac{\beta}{2}$	$\frac{\beta^2}{12}$	$\frac{e^{(\alpha+\beta)t} - e^{\alpha t}}{\beta t}$
Normal	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$	$-\infty < \mu < \infty$ $\sigma > 0$	μ	σ^2	$\exp[\mu t + \frac{1}{2}\sigma^2 t^2]$
Exponential (rate λ)	$f(x) = \lambda e^{-\lambda x} I_{(0,\infty)}(x)$	$0 < \lambda < \infty$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda - t}$ for $t < \lambda$
Bilateral exponential	$f(x) = \frac{1}{2}\beta e^{-\beta\lvert x-\alpha\rvert}$	$-\infty < \alpha < \infty$ $0 < \beta < \infty$	α	$\frac{2}{\beta^2}$	$e^{t\alpha}/(1 - t^2/\beta^2)$ for $\lvert t\rvert < \beta$
Gamma	$f(x) = \frac{1}{\Gamma(\alpha)} \beta^\alpha x^{\alpha-1} e^{-(\beta x)} I_{(0,\infty)}(x)$	$\alpha > 0$ $\beta > 0$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$	$\left(\frac{\beta}{\beta - t}\right)^\alpha$ for $t < \beta$
Weibull	$f(x) = \frac{\gamma}{\beta}\left(\frac{x-\alpha}{\beta}\right)^{\gamma-1} e^{-\left(\frac{x-\alpha}{\beta}\right)^\gamma} I_{(\alpha,\infty)}(x)$	$-\infty < \alpha < \infty$ $\beta > 0,\ \gamma > 0$	$\alpha + \beta\,\Gamma\left(1 + \frac{1}{\gamma}\right)$	$\beta^2\,[\Gamma\left(1 + \frac{2}{\gamma}\right)$ $-\Gamma^2\left(1 + \frac{1}{\gamma}\right)$	not useful $\mathsf{E}[(X-\alpha)^k] =$ $\beta^k \Gamma\left(1 + \frac{k}{\gamma}\right)$
Beta	$f(x) = \frac{1}{\mathcal{B}(a,b)} x^{a-1}(1-x)^{b-1} I_{(0,1)}(x)$	$a > 0$ $b > 0$	$\frac{a}{a+b}$	$\frac{ab}{(a+b)^2(a+b+1)}$	not useful $\mathsf{E}[X^k] =$ $\frac{\mathcal{B}(a+k,b)}{\mathcal{B}(a,b)}$
Pareto	$f(x) = \frac{\gamma}{(1+x)^{\gamma+1}} I_{(0,\infty)}(x)$	$\gamma > 0$	$1/(\gamma - 1)$ for $\gamma > 1$	$\gamma/[(\gamma-2)(\gamma-1)^2]$ for $\gamma > 2$	does not exist
Cauchy	$f(x) = \frac{1}{\beta\pi} \frac{1}{1 + \left(\frac{x-\alpha}{\beta}\right)^2}$	$-\infty < \alpha < \infty$ $\beta > 0$	does not exist	does not exist	does not exist
Logistic	$F(x) = \left[1 + e^{-\left(\frac{x-\alpha}{\beta}\right)}\right]^{-1}$	$-\infty < \alpha < \infty$ $\beta > 0$	α	$\beta^2\pi^2/3$	$e^{\alpha t}\beta\pi t \csc(\beta\pi t)$
Gumbel	$F(x) = \exp\left[-e^{-\left(\frac{x-\alpha}{\beta}\right)}\right]$	$-\infty < \alpha < \infty$ $\beta > 0$	$\alpha + \beta\gamma$ where $\gamma \approx 0.577216$	$\beta^2\pi^2/6$	$e^{\alpha t}\Gamma(1 - \beta t)$ for $t < 1/\beta$
Log normal	$F(x) = \Phi\left(\frac{ln\ x-\mu}{\sigma}\right) I_{(0,\infty)}(x)$	$-\infty < \mu < \infty$ $\sigma > 0$	$\exp[\mu + \frac{1}{2}\sigma^2]$	$\exp[2\mu + 2\sigma^2]-$ $\exp[2\mu + \sigma^2]$	does not exist $\mathsf{E}[X^k] =$ $\exp[k\mu + \frac{1}{2}k^2\sigma^2]$

Appendix B The Jacobian and a Change of Variables

B.1 The Jacobian

For a single integral, the substitution $x = x(u)$ gives us

$$\int_{x_1}^{x_2} f(x)\,dx = \int_{u_1}^{u_2} f(x(u))\,\frac{dx}{du}\,du$$

For double integrals, we consider a change of variables from the x_1x_2-plane to the y_1y_2-plane, where

$$y_1 = g_1(x_1, x_2) \qquad \text{and} \qquad y_2 = g_2(x_1, x_2).$$

We asume we can solve these equations for x_1 and x_2:

$$x_1 = g_1^{-1}(y_1, y_2) \qquad \text{and} \qquad x_2 = g_2^{-1}(y_1, y_2).$$

Now, consider with a small rectangle R in the x_1x_2-plane whose lower left corner is the point $(x_1^{(0)}, x_2^{(0)})$ and whose dimensions are Δx_1 and Δx_2.

The image of R in the y_1y_2-plane is a region S, one of whose boundary points is $(y_1^{(0)}, y_2^{(0)}) = (g_1(x_1^{(0)}, x_2^{(0)}), g_2(x_1^{(0)}, x_2^{(0)}))$.

The vector

$$\vec{v}(x_1, x_2) = g_1(x_1, x_2)\vec{i} + g_2(x_1, x_2)\vec{j}$$

is the position vector of the image of the point (x_1, x_2).

The equation of the lower side of R is $x_2 = x_2^{(0)}$, and the image curve of this side is given by the vector function $\vec{v}(x_1, x_2^{(0)})$.

The tangent vector at $(y_1^{(0)}, y_2^{(0)})$ to this image curve is

$$\vec{v}_{x_1} = \left.\frac{\partial g_1(x_1, x_2^{(0)})}{\partial x_1}\right|_{x_1 = x_1^{(0)}} \vec{i} + \left.\frac{\partial g_2(x_1, x_2^{(0)})}{\partial x_1}\right|_{x_1 = x_1^{(0)}} \vec{j}.$$

Similarly, the tangent vector at $(y_1^{(0)}, y_2^{(0)})$ to the image curve of the left side of R (given by $x_1 = X_1^{(0)}$) is given by

$$\vec{v}_{x_2} = \left.\frac{\partial g_1(x_1^{(0)}, x_2)}{\partial x_2}\right|_{x_2 = x_2^{(0)}} \vec{i} + \left.\frac{\partial g_2(x_1^{(0)}, x_2)}{\partial x_2}\right|_{x_2 = x_2^{(0)}} \vec{j}.$$

We can approximate the image region of R by a parallelogram determined by the secant vectors

$$\vec{a} = \vec{v}(x_1^{(0)} + \Delta x_1, x_2^{(0)}) - \vec{v}(x_1^{(0)}, x_2^{(0)}) \qquad \text{and} \qquad \vec{b} = v(x_1^{(0)}, x_2^{(0)} + \Delta x_2) - \vec{v}(x_1^{(0)}, x_2^{(0)}).$$

However,

$$\vec{v}_{x_1} = \lim_{\Delta x_1 \to 0} \frac{\vec{v}(x_1^{(0)} + \Delta x_1, x_2^{(0)}) - \vec{v}(x_1^{(0)}, x_2^{(0)})}{\Delta x_1}$$

and so, for Δx_1 "small", we have

$$\vec{v}(x_1^{(0)} + \Delta x_1, x_2^{(0)}) - \vec{v}(x_1^{(0)}, x_2^{(0)}) \approx \vec{v}_{x_1}\, \Delta x_1.$$

Similarly

$$\vec{v}(x_1^{(0)}, x_2^{(0)} + \Delta x_2) - \vec{v}(x_1^{(0)}, x_2^{(0)}) \approx \vec{v}_{x_2}\, \Delta x_2.$$

This means that we can approximate the image of R by a parallelogram determined by the vectors $\vec{v}_{x_1}\, \Delta x_1$ and $\vec{x_2}\, \Delta x_2$. Therefore, we can approximate the area of the image of R by the area of this parallelogram which is

$$|(\vec{v}_{x_1}\, \Delta x_1) \times (\vec{v}_{x_2}\, \Delta x_2)| = |\vec{v}_{x_1} \times \vec{v}_{x_2}| \Delta x_1 \Delta x_2.$$

Computing this cross product, we get

$$\vec{v}_{x_1} \times \vec{v}_{x_2} = \begin{vmatrix} \vec{i} & \vec{j} & \vec{k} \\ \frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial x_1} & 0 \\ \frac{\partial y_1}{\partial x_2} & \frac{\partial y_2}{\partial x_2} & 0 \end{vmatrix} = \begin{vmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial x_1} \\ \frac{\partial y_1}{\partial x_2} & \frac{\partial y_2}{\partial x_2} \end{vmatrix} \vec{k} = \begin{vmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} \end{vmatrix} \vec{k}$$

The determinant that arises in this calculation is called the Jacobian of the transformation. Call it $J(x_1, x_2)$. We can then approximate the area ΔA of the image of R by

$$\Delta A = |J(x_1, x_2)| \Delta x_1 \Delta x_2$$

So,

$$\int\int_S f(y_1, y_2)\, dA \approx \sum_{i=1}^{n} \sum_{j=1}^{m} f(y_1^{(i)}, y_2^{(i)}) \Delta A \approx \sum_{i=1}^{n} \sum_{j=1}^{m} f(g_1(x_1^{(i)}, x_2^{(j)}), g_2(x_1^{(i)}, x_2^{(j)})) |J(x_1, x_2)| \Delta x_1 \Delta x_2.$$

B.2 A Bivariate Transformation

We now proceed with the problem of finding the joint pdf of two functions of two continuous random variables.

Let the joint pdf for X_1 and X_2 be denoted by $f_{X_1,X_2}(x_1, x_2)$ and let $Y_1 = g_1(X_1, X_2)$ and $Y_2 = g_2(X_1, X_2)$

denote a one-to-one transformation from the x_1x_2-plane to the y_1y_2-plane.

Let $\mathcal{A}$ denote the two-dimensional set in the x_1x_2-plane for which $f_{X_1,X_2}(x_1, x_2)$ is non-zero. Let $\mathcal{B}$ denote the transformed region in the y_1y_2-plane.

Now, let $A \subset \mathcal{A}$ and let B denote the mapping of A under the one-t-one transformation.

$$P((Y_1, Y_2) \in B) = P((X_1, X_2) \in A) = \int \int_A f_{X_1,X_2}(x_1, x_2)\, dx_1\, dx_2.$$

By “The Jacobian" section above, this is the same as

$$\int \int_B f_{X_1,X_2}(g_1^{-1}(y_1, y_2), g_2^{-1}(y_1, y_2))|J(y_1, y_2)|\, dy_1\, dy_2$$

B.3 More Variables

All of the above can be extended to the case of many variables. That is, if $X_1, X_2, \ldots, X_n$ has joint pdf $f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n)$ and we consider the transformations

$$\begin{aligned} y_1 &= g_1(x_1, x_2, \ldots, x_n) \\ y_2 &= g_2(x_1, x_2, \ldots, x_n) \\ &\vdots \quad \vdots \quad \vdots \\ y_n &= g_n(x_1, x_2, \ldots, x_n), \end{aligned}$$

then the joint pdf of $Y_1, Y_2, \ldots, Y_n$ is given by

$$f_{Y_1,Y_2,\ldots,Y_n}(y_1, y_2, \ldots, y_n) = f_{X_1,X_2,\ldots,X_n}(g_1^{-1}(y_1, y_2, \ldots, y_n), \ldots, g_n^{-1}(y_1, y_2, \ldots, y_n)) \cdot |J(y_1, y_2, \ldots, y_n)|$$

where

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_2}{\partial y_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_n} \end{vmatrix}.$$

Appendix C Standard Normal, t, χ^2, and F Tables

Table C.1: Standard Normal Probabilities

$\Phi(z) = P(Z \leq z)$ for $Z \sim N(0,1)$

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Table C.2: t Critical Values

Critical Values for $T \sim t(\nu)$ (t_α captures area α to the right)

ν	$t_{0.40}$	$t_{0.333}$	$t_{0.25}$	$t_{0.20}$	$t_{0.125}$	$t_{0.10}$	$t_{0.05}$	$t_{0.025}$	$t_{0.01}$	$t_{0.005}$	$t_{0.001}$
1	0.325	0.577	1.000	1.376	2.414	3.078	6.314	12.706	31.821	63.657	318.31
2	0.289	0.500	0.816	1.061	1.604	1.886	2.920	4.303	6.965	9.925	22.327
3	0.277	0.476	0.765	0.978	1.423	1.638	2.353	3.182	4.541	5.841	10.215
4	0.271	0.464	0.741	0.941	1.344	1.533	2.132	2.776	3.747	4.604	7.173
5	0.267	0.457	0.727	0.920	1.301	1.476	2.015	2.571	3.365	4.032	5.893
6	0.265	0.453	0.718	0.906	1.273	1.440	1.943	2.447	3.143	3.707	5.208
7	0.263	0.449	0.711	0.896	1.254	1.415	1.895	2.365	2.998	3.499	4.785
8	0.262	0.447	0.706	0.889	1.240	1.397	1.860	2.306	2.896	3.355	4.501
9	0.261	0.445	0.703	0.883	1.230	1.383	1.833	2.262	2.821	3.250	4.297
10	0.260	0.444	0.700	0.879	1.221	1.372	1.812	2.228	2.764	3.169	4.144
11	0.260	0.443	0.697	0.876	1.214	1.363	1.796	2.201	2.718	3.106	4.025
12	0.259	0.442	0.695	0.873	1.209	1.356	1.782	2.179	2.681	3.055	3.930
13	0.259	0.441	0.694	0.870	1.204	1.350	1.771	2.160	2.650	3.012	3.852
14	0.258	0.440	0.692	0.868	1.200	1.345	1.761	2.145	2.624	2.977	3.787
15	0.258	0.439	0.691	0.866	1.197	1.341	1.753	2.131	2.602	2.947	3.733
16	0.258	0.439	0.690	0.865	1.194	1.337	1.746	2.120	2.583	2.921	3.686
17	0.257	0.438	0.689	0.863	1.191	1.333	1.740	2.110	2.567	2.898	3.646
18	0.257	0.438	0.688	0.862	1.189	1.330	1.734	2.101	2.552	2.878	3.610
19	0.257	0.438	0.688	0.861	1.187	1.328	1.729	2.093	2.539	2.861	3.579
20	0.257	0.437	0.687	0.860	1.185	1.325	1.725	2.086	2.528	2.845	3.552
21	0.257	0.437	0.686	0.859	1.183	1.323	1.721	2.080	2.518	2.831	3.527
22	0.256	0.437	0.686	0.858	1.182	1.321	1.717	2.074	2.508	2.819	3.505
23	0.256	0.436	0.685	0.858	1.180	1.319	1.714	2.069	2.500	2.807	3.485
24	0.256	0.436	0.685	0.857	1.179	1.318	1.711	2.064	2.492	2.797	3.467
25	0.256	0.436	0.684	0.856	1.178	1.316	1.708	2.060	2.485	2.787	3.450
26	0.256	0.436	0.684	0.856	1.177	1.315	1.706	2.056	2.479	2.779	3.435
27	0.256	0.435	0.684	0.855	1.176	1.314	1.703	2.052	2.473	2.771	3.421
28	0.256	0.435	0.683	0.855	1.175	1.313	1.701	2.048	2.467	2.763	3.408
29	0.256	0.435	0.683	0.854	1.174	1.311	1.699	2.045	2.462	2.756	3.396
30	0.256	0.435	0.683	0.854	1.173	1.310	1.697	2.042	2.457	2.750	3.385
35	0.255	0.434	0.682	0.852	1.170	1.306	1.690	2.030	2.438	2.724	3.340
40	0.255	0.434	0.681	0.851	1.167	1.303	1.684	2.021	2.423	2.704	3.307
45	0.255	0.434	0.680	0.850	1.165	1.301	1.679	2.014	2.412	2.690	3.281
50	0.255	0.433	0.679	0.849	1.164	1.299	1.676	2.009	2.403	2.678	3.261
55	0.255	0.433	0.679	0.848	1.163	1.297	1.673	2.004	2.396	2.668	3.245
60	0.254	0.433	0.679	0.848	1.162	1.296	1.671	2.000	2.390	2.660	3.232
∞	0.253	0.431	0.674	0.842	1.150	1.282	1.645	1.960	2.326	2.576	3.090

Table C.3: χ^2 Critical Values

Critical Values for $W \sim \chi^2(n)$ (χ^2_α captures area α to the right)

n	$\chi^2_{0.995}$	$\chi^2_{0.99}$	$\chi^2_{0.975}$	$\chi^2_{0.95}$	$\chi^2_{0.90}$	$\chi^2_{0.10}$	$\chi^2_{0.05}$	$\chi^2_{0.025}$	$\chi^2_{0.01}$	$\chi^2_{0.005}$	$\chi^2_{0.001}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879	10.828
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597	13.816
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838	16.266
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860	18.467
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750	20.515
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548	22.458
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278	24.322
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955	26.124
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589	27.877
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188	29.588
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757	31.264
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300	32.909
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819	34.528
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319	36.123
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801	37.697
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267	39.252
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718	40.790
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156	42.312
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582	43.820
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997	45.315
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401	46.797
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796	48.268
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181	49.728
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559	51.179
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928	52.620
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290	54.052
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645	55.476
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993	56.892
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336	58.301
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672	59.703
31	14.458	15.655	17.539	19.281	21.434	41.422	44.985	48.232	52.191	55.003	61.098
32	15.134	16.362	18.291	20.072	22.271	42.585	46.194	49.480	53.486	56.328	62.487
33	15.815	17.074	19.047	20.867	23.110	43.745	47.400	50.725	54.776	57.648	63.870
34	16.501	17.789	19.806	21.664	23.952	44.903	48.602	51.966	56.061	58.964	65.247
35	17.192	18.509	20.569	22.465	24.797	46.059	49.802	53.203	57.342	60.275	66.619
36	17.887	19.233	21.336	23.269	25.643	47.212	50.998	54.437	58.619	61.581	67.985
37	18.586	19.960	22.106	24.075	26.492	48.363	52.192	55.668	59.893	62.883	69.346
38	19.289	20.691	22.878	24.884	27.343	49.513	53.384	56.896	61.162	64.181	70.703
39	19.996	21.426	23.654	25.695	28.196	50.660	54.572	58.120	62.428	65.476	72.055
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766	73.402
45	24.311	25.901	28.366	30.612	33.350	57.505	61.656	65.410	69.957	73.166	80.077
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490	86.661
55	31.735	33.570	36.398	38.958	42.060	68.796	73.311	77.380	82.292	85.749	93.168
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952	99.607

Index

List of Symbols

$\chi^2_{\alpha,n}$ a critical value that captures area α to the right for the χ^2 pdf, page 180

$\ell(\theta)$ log-likelihood function, page 229

$\gamma(\cdot)$ power function, page 208

$\in$ is an element of, page 193

Ω sample space, page 7

$\Phi(\cdot)$ standard normal cdf, page 142

$\sim$ has the distribution, page 12

$\tau(\theta)$ function to be estimated, page 101

Θ a general parameter space, page 193

$E[X]$ expected value, page 22

$L(\theta)$ likelihood function, page 227

$n!$ factorial of n, page 4

$P(A)$ probability A occurs, page 8

$t_{\alpha,n}$ critical value that captures area α to the right for the $t(n)$ pdf, page 163

$X_{(1)}$ sample minimum, page 67

$X_{(n)}$ sample maximum out of n, page 67

z_α critical value that captures area α to the right for the standard normal pdf, page 157

Bibliography

[1] William Feller. *An Introduction to Probability Theory and Its Applications*. Vol. 1. Wiley, Jan. 1968. ISBN: 0471257087. URL: http://www.amazon.ca/exec/obidos/redirect?tag=citeulike04-20%7B%5C&%7Dpath=ASIN/0471257087.

[2] Chris C. Heyde. "On a Property of the Lognormal Distribution". In: *Journal of the royal statistical society series b-methodological* 25 (1963), pp. 16–18.

[3] D. Landers and L. Rogge. "Minimal sufficient σ-fields and minimal sufficient statistics. Two counterexamples". In: *The Annals of Mathematical Statistics* 43 (6) (1972), pp. 2049–2049.

[4] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998. ISBN: 0521784506.

[5] B. L. Welch. "The significance of the difference between two means when the population variances are unequal". In: *Biometrika* 29 (1938), pp. 350–362.