SM-4331 Advanced Statistics

Dr Haziq Jamil

2022-03-09

Contents

About

Updated for 2021/22 session.

These are the course notes for SM-4331 Advanced Statistics, a fourth-year module taken by students at Universiti Brunei Darussalam (UBD). The course covers the mathematical theory behind statistical inference concepts.

Part I

Introduction

What is statistics?

Statistics is a scientific subject focussed on collecting and analysing data.

- **Collecting** means designing experiments, designing questionnaires, designing sampling schemes, administration of data collection.
- **Analysing** means modelling, estimation, testing, forecasting.

Statistics is an application-oriented mathematical subject; it is particularly useful or helpful in answering questions such as:

- Does a certain new drug prolong life for AIDS sufferers?
- Is global warming really happening?
- Are O-level and A-level examinations standard declining?
- Is the house market in Brunei oversaturated?
- Is the Chinese yuan undervalued? If so, by how much?

Questions that can be answered with statistical analysis are wide-ranging, hence making it useful in a variety of fields and specialties, from the hard sciences (chemistry, geology, physics, etc.) to the social sciences (business, economics, psychology, etc.) and beyond[1] [2]. Given today's data-centric world that we live in, I posit that numerical literacy is now as important as literacy itself!

Learning statistics

There are three aspects to learning statistics:

1. **Ideas and concepts**. Understanding why statistics is needed, and what you are able to do and not do with statistics.
2. **Methods**. Knowing "how to do" (applied) statistics.
3. **Theory**. Knowing the "why" of statistics and understanding why things are the way they are. Very mathematics focused.

In this course, there is an emphasis on the **theory** aspect of statistics. It is my hope that you are already familiar with basic statistical concepts (covered in SM-2205 Intermediate Statistics!), and for those of you who will be around next semester, the SM-4337 Applied Statistics module is highly recommended to learn about applying statistics in real-life situations. Of course, for those who have taken SM-4337 will find connections between what we will be discussing in this module and what you have come across there.

This course may (at times) feel "mathematical for the sake of mathematics". In my defence, having a solid foundation in statistical theory will empower you greatly in your data analysis quest. Yes, there are software out there which seem to automagically generate the statistics of interest and even fit statistical models for the user blindly. Two things:

[1] https://www.significancemagazine.com/science/458-does-new-york-city-really-have-as-many-rats-as-people?highlight=WyJuZXciLCInbmV3IiwieW9yayIsIm5ldyB5b3JrIl0=

[2] https://thoughtcatalog.com/anonymous/2015/04/what-is-the-statistical-chance-of-finding-the-love-of-my-life/

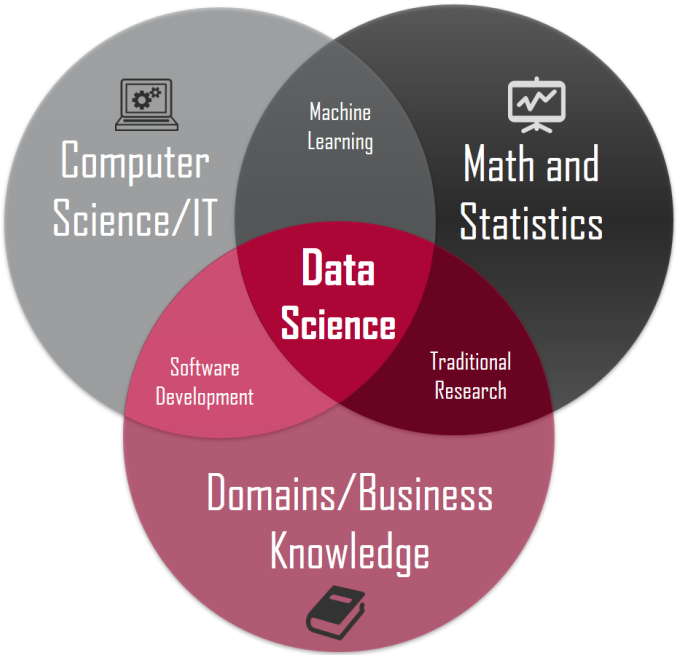
- There is still the matter of *interpretation* of these output. Will you be able to explain to your boss/stakeholder/customer/etc. the meaning of the data you helped them analyse?
- Will you be able to spot any **assumptions** that are fundamental to the model being true/useful? Remember, gargbage in garbage out.

Some words of wisdom:

> *Students who analyze data, or who aspire to develop new methods for analyzing data, should be well grounded in basic probability and mathematical statistics. Using fancy tools like neural nets, boosting, and support vector machines without understanding basic statistics is like doing brain surgery before knowing how to use a band-aid. —Larry Wasserman (in All of Statistics)*

In the grand scheme of things, mastery of statistical theory is just one part of the equation to be a practical data analyst. Given how popular the term 'data science' these days, it is worth noting the figure[3] below. Data science is seen as the intersection between Mathematics and Statistics, Computer Science, and Domain Knowledge.

Given the importance of the computer science (read: coding and programming skills) as the enabler of data science activities, I try as much as I can to encourage you to these skills up yourself. Wherever appropriate you might see `R` code embedded within the text. As a side note, `R` programming is not examinable.



Population, sample and parametric models

To motivate the use of statistics in everyday life, let us consider two practical situations where you might employ statistical methods:

1. BMW M Division has proudly unveiled the successor to their current "king of sedans", the new BMW M3 Competition (G80), sporting a 503 bhp twin-turbo 3.0 litre inline-six S58 engine with a claimed acceleration rate of 0-100 km/h in 3.9 seconds.

[3] https://towardsdatascience.com/introduction-to-statistics-e9d72d818745

2. The Authority for Info-communications Technology Industry of Brunei Darussalam (AITI) conducted the Household ICT Survey in 2018 and reported that 95% percent of individuals personally use the internet on a daily basis, a slight decrease from 97% in the year 2016. Estimates are accurate within 2% margin of error with 95% confidence.

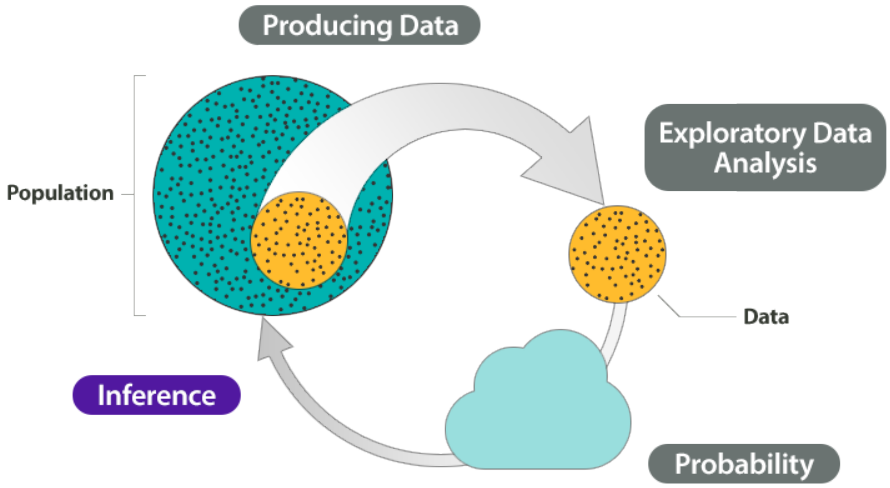
Your immediate thought should be "how can I trust these figures?". In general, it's always good to approach life with a healthy dose of skepticism. We certainly don't want to be duped by people claiming to present a version of the truth, when in reality it is a skewed version of the truth (or worse yet, false). In data we trust! But only if we are mathematically-savvy…

Population vs sample

In both cases, the conclusion is drawn on a *population* (i.e. all of the subjects concerned) based on the information from a *sample* (i.e. a subset of the population).

1. For BMW M Division, it is **impossible** to measure the entire population (obtain the acceleration rates), constituting all BMW M3 (G80) cars that have been made and are yet to be made in the future. Often this is referred to as an *infinite population* model.
2. For AITI, while possible, it is (economically) unfeasible to measure the entire population, i.e. to ask everyone in Brunei whether or not they use the internet on a daily basis. It would be very difficult to knock on everybody's door and obtain responses in a timely manner (what if they're out of the country? what if they're sick? what if they just don't want to respond?). Anyone who has done any form of survey work will understand the intricate problems that might arise.

In any case, it is important to make the distinction between population and sample. The *population* is defined to be the entire set of the objects concerned, and those objects are typically represented by some numbers. We do not know the entire population in practice. A *sample* is a (randomly selected) subset of a population, and is a set of known data in practice.



When people claim to have data (about some phenomenon), they typically imply that they have a *sample* of the population, and not the population data itself. There are exceptions of course, for instance, a country's *census* captures population data every 10 years. Another example is when the population itself is small such that all data can be collected easily.

A question that you might be thinking to yourself is the following:

> Is doing analysis on the sample good enough? Will it reveal the same insights as if we're analysing the population data?

The answer to this is that it depends on how you perform the sample! Things to look out for is definitely *bias* in data collection methods. Here are some examples:

- Asking the question "Can you live without the internet?" via **an online poll** of adults. Think about it. How exactly can you ask people *without* internet whether they can live without the internet? Clearly, the sample that you collect is biased towards those who are privileged enough to have access. And if you were wondering, yes this really happened.
- Asking people to volunteer their responses typically lead to bias. It is suggested that those who has something to complain about will voice their opinions more than those who do not.
- Other errors can crop up during data collection process which may skew the representation of the data (e.g. duplicate data, missing data, substituted data, etc.)

It is certainly important that data be collected in a methodological manner, in order for valid inferences to be drawn. Sampling methodology is beyond the scope of this course, however!

Parametric models

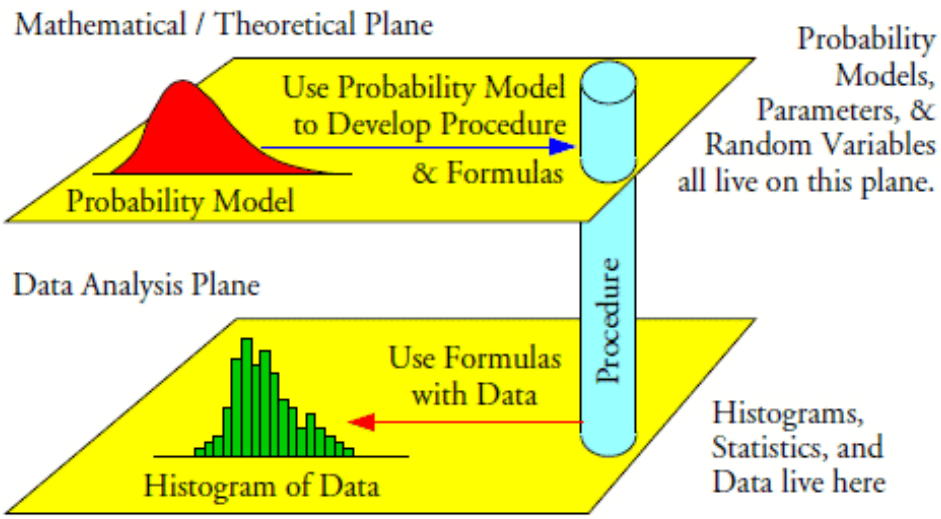
For a given problem, it is quite advantagous to **assume** that a population obeys some *probability distribution* law. To be a little bit more concrete, suppose we assign the variable x to be the quantity of interest. Then we assume that the quantity of interest has a distribution function $f(x|\theta)$ (we will recap the concept of distribution functions in the next chapter!). Furthermore,

- The form of the distribution i.e. $f(\cdot|\theta)$ is known (e.g. normal, Poisson, exponential, etc.).
- The "specifics" of the distribution is (assumed to be) **not known**, but potentially knowable if data were available.

The unknown characteristics of the distribution are traditionally represented by θ (such as the mean, variance, rate, etc.–any quantity that characterises how the distribution behaves). We call θ the *parameter(s)* of the model. Such an assumed distribution is called a **parametric model**. For the two earlier examples,

1. Let X = acceleration of BMW M3 G80 vehicles. Assume $X \sim \mathrm{N}(\mu, \sigma^2)$. Here $\theta = (\mu, \sigma^2)^\top$, where μ is the 'true' acceleration rate. In this example, the parameter is 2-dimensional instead of unidimensional.
2. Let $\{0, 1\} \ni X$ = someone in Brunei uses the internet daily. Assume $X \sim \mathrm{Bern}(p)$. Here $\theta = p$, the 'true' proportion of daily internet users in Brunei.

There is no god-given right for your quantity of interest X to follow a particular distribution! These are simply assumptions, in order to make it easy to model reality. Parametric assumptions may be correct, or they may not. In fact, strictly speaking, **all models are wrong**, but some are useful (quote attributed to the British statistician George Box).



A sample: a set of data or random variables?–A duality

A sample of size n, $\{X_1, \ldots, X_n\}$, is also called a *random* sample. It consists of n concrete numbers in a practical problem. When I say 'concrete' here, it means numbers that you can play around with–you can add them up, subtract them, plot them, take averages, and so on. These would be numbers that you collect into a spreadsheet, say.

The more contentious part of the term 'random sample' is the word 'random' itself. The word 'random' encapsulates the possibility of the concrete numbers collected in the samples being different, by virtue of:

- The sample may be taken by different people or entities.
- The sample may be obtained at a different time or location.
- The sample may be measured using different instruments (albeit measuring the same thing).
- etc.

Essentially, different samples may well be different subsets of a population.

With this, a sample may also be viewed as n (independent and identically distributed) random variables, simply because their values are conceptually not fixed (at least not until you observe them–but even then it is simply one possible realisation of potentially many others).

Now, if a sample is not random then perhaps there would be no need for statistics. But hardly ever would you find this to be the case. Rigorous mathematical methods exist to deal with this randomness, and thus viewing samples as random variables allows us to assess the performance of a statistical method.

Variability of estimates

Suppose you set out to answer questions 1 and 2 above by collecting some data. This is what you find:

BMW M example A sample of $n = 38$ gave the sample mean

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = 3.9$$

In words, the average acceleration (0-100 km h-1) of the sample of cars yielded the value 3.9 seconds. But realise that a different sample may well give a different sample mean. For instance, 29 YouTubers and other social media "influencers" were given access to the new M3 on a race track, and their sample mean yielded $\bar{X}_n = 3.4$.

What do we make of this discrepancy? Whose figure do we trust? It doesn't seem so satisfying to know that different samples will give different results. For instance, what do you tell the public or media about the acceleration figure of the new M3s? Evidently relying on a singular measure (in this case the mean) is not enough. Moreoever, it keeps on changing in value!

The key is to be able to employ *probabilistic statements* about our results. For instance, "*the acceleration figure is 3.9 plus or minus 0.01 about 95% of the time when we sample*" is a much more confident statement to make, rather than providing a figure that keeps on changing.

By treating the data $X_1, \ldots, X_n$ as random variables, it is implied that $\bar{X}_n$ is also a random variable. Everything that is random should have a distribution. If the distribution of $\bar{X}_n$ concentrates closely around the unknown μ, then it is a good estimator!

AITI example For the AITI example, there is that statement *'…accurate to within 2% margin of error with 95% confidence'*. This statement alludes to the variability of the estimate, if another random sample was obtained.

- The estimate in this case was also the sample mean (proportion of people who use the internet on a daily basis),

$$\hat{p} = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

- Mathematically, the confidence statement reads

$$\Pr(|\hat{p} - p| \leq 0.02) = \Pr\left(p \in [\hat{p} - 0.02, \hat{p} + 0.02]\right) = 0.95$$

  that is, the true value is covered 95% of the time inside an interval of width 0.02 under repeated sampling. This statement is made possible due to the *randomness* of the estimator $\hat{p}$.

Probability and statistics

Thus far, we have thrown the term 'inference' around and maybe we've all taken it for granted. Cambridge's dictionary defines *infer* as 'to reach an opinion from available information or facts'. And indeed that is what we are ultimately interested in when we analyse data. Sure, there might be that random element to the data that we have to contend with, but it is about reaching some form of conclusion one way or another using data. In the previous section, we've just implicitly described the three main activities concerning statistical inference.

1. **Point estimation**

   "What is μ?"

2. **Hypothesis testing**

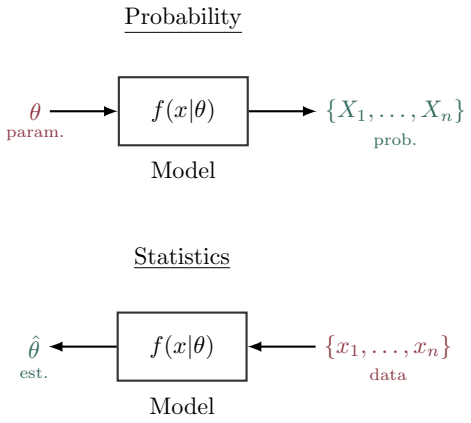
   "Is $\mu = 3.4$ and not $\mu = 3.9$?"

3. **Interval estimation**

   "What's an upper and lower bound estimate for μ?"

These three activities will be the main focus of this course, and we will formalise the notion of each one in turn. Hopefully you can now appreciate how statistics is an inherently applied subject, making use of mathematics (probability in particular) to answer problems across a variety of fields.

What is the difference between probability and statistics? The following figure might be helpful.



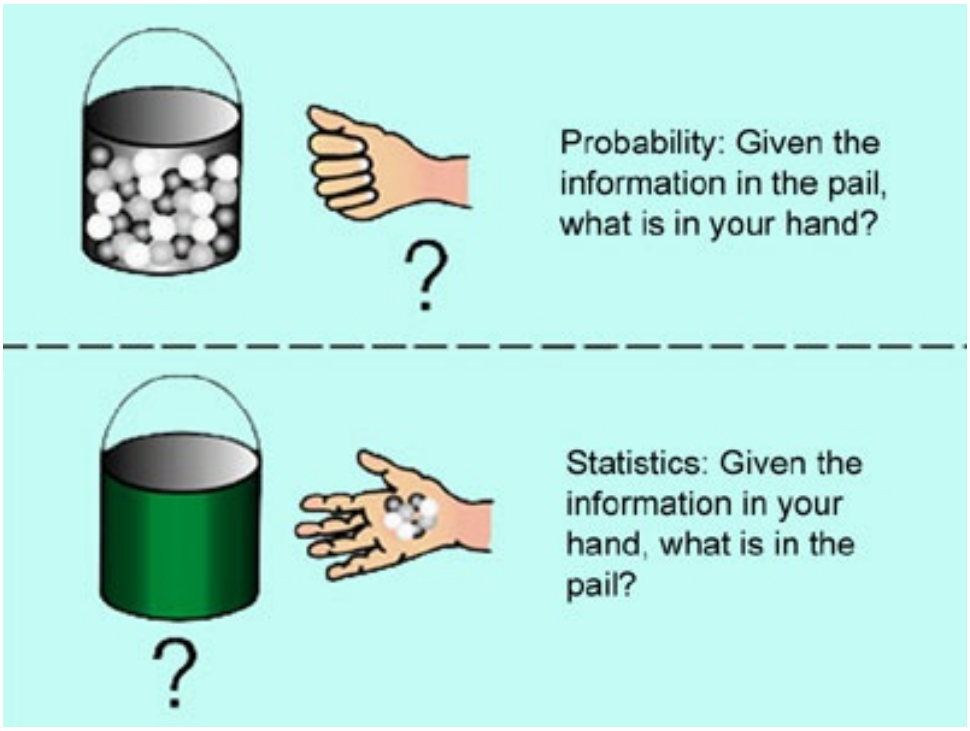
Probability is a highly mathematical subject (although maybe the name doesn't seem to suggest it, it really has its origin in abstract measure theory). In probability, we ask questions like

- What is $\mathrm{E}(X)$? (expectations)
- What is $\Pr(X > a)$? (probability calculations)

for some given value of θ in an assumed family of parametric distributions. Whereas in statistics, we are more interested in questions like

- What is θ?
- Is θ larger than θ_0?
- How confident am I that $\theta \in (\theta_l, \theta_u)$?

given some observed data. I like to think of the two as reverse processes.



Statistics definitely needs probability for sure, so we will have to master probability theory before doing any statistical inference. On to the next chapter!

Part II

Prepare

Chapter 1

Probability theory primer

As I was looking for a rationale as to why we begin statistical inference with learning about probability theory, I notice that C&B couldn't have put it better:

> The subject of probability theory is the foundation upon which all statistics is built, providing a means for modeling populations, experiments, or almost anouthing else that could be considered a random phenomenon. Through these models, statisticians are able to draw inferences about populations, inferences based on examination of only a part of the whole.
> — George Casella & Roger L. Berger in Statistical Inference

By right, probability theory and the mathematics of random events deserves one dedicate module of its own. For our purposes, it suffices to 'skim through the surface' as it were, and cover the basic and necessary ideas to move forward with statistical inference.

Learning objectives

By the end of this chapter, you will be able to:

- Compute probabilities of events by simple counting and application of various known probability results
- Understand the notion of conditional and independent events, leading up to the application of Bayes' Theorem
- Formalise mathematically the notion of random variables and make calculations using its distribution function
- Compute expectations (and variances) including via the method of moment generating functions

Readings

- Casella and Berger (2002)
  - All of Chapter 1 (skip sections 1.2.3 and 1.2.4).
  - Chapter 2, section 2.2 and 2.3 only.
  - Chapter 4, section 4.1, 4.2 and 4.5 only.
- Wasserman (2004)
  - All of Chapter 1.
  - Chapter 2, sections 2.1–2.2, 2.5–2.8.
  - Chapter 3, sections 3.1–3.5.
- Topics not covered: Counting and enumerating outcomes, moment generating functions (to be covered in the next topic), transformations of r.v., multivariate distributions (bivariate only).
- YouTube video: The medical test paradox
- YouTube video: Bayes theorem

1.1 Elementary set theory

A discussion of probability theory is almost always begun by talking about the concept of 'sets'. As the term implies, sets are a collection of things. You're sure to come across sets before, such as the set of family members in your household, or the set of all natural numbers, or something even more funky like the set of all sets (the universal set[1]).

When statisticians talk about sets it is usually in the context of conducting an "experiment"[2]. The important bits are:

- The sample space Ω is the set of possible outcomes of an experiment.
- Elements $\omega \in \Omega$ are called sample outcomes or realisations.
- Subsets of $E \subseteq \Omega$ are called events.

So whatever the context of the random process might be, we should be comfortable identifying what the sample space is, what its elements are, and what possible events might occur.

Example 1.1. In tossing a two-sided coin $n \geq 2$ times, let H denote 'heads', while T denote 'tails'. Let $\omega = (\omega_1, \dots, \omega_n)$ be the results of these coin tosses. Then the sample space is

$$\Omega = \Big\{\omega \,|\, \omega_i \in \{H, T\}\Big\}.$$

Let E be the event that the first head appears on the second toss. Then this event can be mathematically described as

$$E = \Big\{\omega \,|\, \omega_1 = T, \omega_2 = H, \omega_i \in \{H, T\} \text{ for } i > 2\Big\}.$$

1.1.1 Set operations

There are several things that we can do to sets much like we can do to numbers in arithmetic. Here is an abridged version of set operations.

- The **complement** of an event A, written A^c, is the set of all elements that are not in A: $A^c = \{\omega \mid \omega \notin A\}$.
- The complement of Ω is the empty set $\emptyset = \{\}$.
- The **union** of events A and B (thought of as "A or B or both") is defined

$$A \cup B = \{\omega \in \Omega \,|\, \omega \in A \text{ or } \omega \in B\}.$$

- The **intersection** of events A and B (thought of as "A *and* B") is defined

$$A \cap B = \{\omega \in \Omega \,|\, \omega \in A \text{ and } \omega \in B\}.$$

- Unions and intersections on sets are **commutative**, **associative**, and **distributive**[3].
  - Commutativity: $A \cup B = B \cup A$ and $A \cap B = B \cap A$.
  - Associativity: $A \cup (B \cup C) = (A \cup B) \cup C$ and $A \cap (B \cap C) = (A \cap B) \cap C$.
  - Disributive laws: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ and $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.
- DeMorgan's Laws: $(A \cup B)^c = A^c \cap B^c$ and $(A \cap B)^c = A^c \cup B^c$

[1] Which, as it turns out, is a paradox: https://en.wikipedia.org/wiki/Russell%27s_paradox

[2] Although, we don't really mean it like how scientists mean experiments to be (we're not fiddling around with buttons or chemicals). I suppose it's more about the process of the (random) data generating mechanism itself.

[3] See C&B Thm 1.14

It's pretty straightforward to prove the commutative, associative, distributive and DeMorgan's properties using the rules that precede them in the list. You may like to try this out yourself. Or you may have come across a kind of "sketch proof" involving **Venn diagrams**.

The operations of unions and intersections can be extended to infinite collections of sets as well. If $A_1, A_2, A_3, ...$ is collection of sets, all defined on a sample space Ω, then

$$\bigcup_{i=1}^{\infty} A_i = \{x \in \Omega \,|\, x \in A_i \text{ for some } i\},$$
$$\bigcap_{i=1}^{\infty} A_i = \{x \in \Omega \,|\, x \in A_i \text{ for all } i\}.$$

Example 1.2. Let $\Omega = (0, 1]$ and define $A_i = [1/i, 1]$. Then,

$$\bigcup_{i=1}^{\infty} A_i = \{1\} \cup [1/2, 1] \cup [1/3, 1] \cup \cdots = (0, 1],$$
$$\bigcap_{i=1}^{\infty} A_i = \{1\} \cap [1/2, 1] \cap [1/3, 1] \cap \cdots = \{1\}.$$

1.1.2 Partitions

We say that two events A and B are **disjoint** or **mutually exclusive** if $A \cap B = \{\}$. Disjoint sets have no points in common. Suppose that $A_1, A_2, ...$ are events defined on Ω such that they are (pairwise) disjoint, i.e.

$$A_i \cap A_j = \{\}, \text{ for } i \neq j.$$

Then the collection $A_1, A_2, ...$ forms a **partition** of Ω. Partitions divide the sample space into non-overlapping pieces.

Example 1.3. A deck of playing cards has four suits: $\clubsuit, \diamondsuit, \spadesuit, \heartsuit$. Let $A = \{\clubsuit, \diamondsuit\}$ and $B = \{\spadesuit, \heartsuit\}$. Then A and B form a partition of the sample space.

Example 1.4. The set $\{\mathbb{R}_+, \mathbb{R}_-, \{0\}\}$ (ie the set of positive reals, negative reals, and zero respectively) is a partition of the real numbers $\mathbb{R}$ since

- $\mathbb{R}_+ \cup \mathbb{R}_- \cup \{0\} = \mathbb{R}$;
- $\mathbb{R}_+ \cap \mathbb{R}_- = \mathbb{R}_+ \cap \{0\} = \mathbb{R}_- \cap \{0\} = \{\}$; and
- $\{\mathbb{R}_+$, $\mathbb{R}_-$, and $\{0\}\}$ are all non empty.

Note that $\{0\}$ is not an empty set. It contains exactly one element, the number zero.

1.2 Axiomatic probability

In principle, we can understand and easily grasp the notion of probability as the "frequency of an event occurring". But how do we operationalise this concept? That is, by what rules and mechanisms are we allowed to assign probabilities to events? If we can overcome this task and are able to assign probabilities to (random) events in an experiment, then we can start to analyse them statistically!

1.2.1 Probability as a measure

Let us take a measure-theoretic approach to defining probabilities. We will dive straight into the rigors of definitions before providing a somewhat apologetic rationale as to why such mathematical difficulties are required for probability theory.

As the name implies, measure theory is the theory about how we measure things (duh!). Measure itself is a fundamental concept in mathematics, and it would be useful to come up with a mathematical framework

for how we deal with everyday concepts like length, mass, area, volume, and so on. Importantly, such a framework allow us to reliably measure in even higher dimensions or onto more abstract constructs not yet imaginable.

Intuitively, a measure is simply a function whose input is the thing we want to measure (let's call it a set), and whose output is a non-negative number. Don't worry, a formal definition will follow, but for now, call this function μ. It would be fair to expect a measure μ to satisfy

- $A \subseteq B \Rightarrow \mu(A) \leq \mu(B)$
- $A \subseteq B \Rightarrow \mu(B - A) = \mu(B) - \mu(A)$
- If $\{A_1, A_2, ...\}$ are mutually exclusive sets (disjoint), then $\mu\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$

The first property simply says that if A is a subset of B, then the measure of A is at most the measure of B. The second property follows this up by saying that the measure of the set $B - A$, that is, the set that is obtained by starting with B and taking away the parts that is contained in A, then the measure of this created set is the difference between the measures of B and A. Finally, the third property, also known as *countable additivity*, simply states that the measure of the whole is equal to the sum of the parts. It turns out that the first and second properties follow from the third (and the fact that a measure cannot be negative)–see Definition 1.2.

So we have this intuition about what the measure should be, but what about the stuff we want to measure? For our purposes, we are interested in measuring subsets of Ω. We ask, are we able to measure all possible subsets of Ω? At a glance, perhaps if Ω is countable (e.g. $\Omega = \{1, 2, 3\}$), it is easy to describe the subsets of Ω through the power set[4] $\mathcal{P}(\Omega)$, which is the set of all possible subsets of Ω, but what about when Ω is uncountable (e.g. an interval $\Omega = [0, 1] \in \mathbb{R}$). Given a sample space Ω, we need to define the largest possible collection of subsets of Ω that can be observed and on which we can assign valid measure.

Definition 1.1 (σ-algebra)**.** A collection $\mathcal{F}$ of subsets of a set Ω is called a **σ-algebra** if it satisfies the following conditions:

i. If $A \in \mathcal{F}$, then $A^c \in cF$ *[closed under complementation].*
ii. If $A_1, A_2, \cdots \in \mathcal{F}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$ *[closed under countable unions].*
iii. $\{\} \in \mathcal{F}$ *[contains the empty set].*

As a remark, condition iii. can be replaced with $\Omega \in \mathcal{F}$ by virtue of condition i.. The σ-algebra is a collection of events or subsets of the sample space Ω, including Ω itself and the empty set $\{\}$, which is closed under countable applications of set operations. This is because DeMorgan's Law allows us to write the countable union property in iii. also as *countable intersections*: If $A_1, A_2, \cdots \in \mathcal{F}$, then by i. $A_1^c, A_2^c, \cdots \in \mathcal{F}$, and hence $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$ and also its complement. By DeMorgan's Law,

$$(\cup_{i=1}^{\infty} A_i^c)^c = \cap_{i=1}^{\infty} A_i.$$

Sets contained in $\mathcal{F}$ are called **measurable sets**.

The σ-algebra is an important condition for measure to not breakdown, because it helps draw a line as to which subsets of the sample space is measurable, and which is not. Out of interest, condition iii. in Definition 1.1 is the condition that makes $\mathcal{F}$ a σ-algebra (the σ stands for countable sum). Without this condition, one ends up with just an *algebra* of sets, one that is most likely *too small*, failing to contain sets that we would like assign a measure.

Let's take a look at some examples of σ-algebras.

Example 1.5. 1. The trivial σ-algebra:

$$\{\{\}, \Omega\}.$$

This corresponds the case of no information.

[4] For the example at hand, the power set is $\mathcal{P}(\Omega) = \{\{\}, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$

2. The power set of the sample space Ω:
$$\{A \mid A \subseteq \Omega\}.$$
This corresponds the case of full information.

3. The collection $\{\{\}, A, A^c, \Omega\}$ is a σ-algebra, for any $A \subseteq \Omega$.

4. Let $\Omega = \{a, b, c, d\}$. A possible[5] σ-algebra is
$$\{\{\}, \{a, b, c, d\}, \{a, b\}, \{c, d\}\}.$$

5. Define $B(s)$ to be a square of side length s. Let Ω be the collection of points in $(0, 1) \times (0, 1) \subset \mathbb{R}^2$ contained within the a unit square $B(1)$. Then
$$\mathcal{F} = \{\text{Collection of points contained in the square } B(s) \text{ with } s \in (0, 1)\}.$$
It should be clear there are uncountably many such squares that can be fit within the unit square.

Just as a remark, most introduction to probability measure will deal with finite or countable sets when introducing σ-algebras, giving readers an impression that it's only possible to define σ-algebras on such sets. The fifth example above gives an example of a σ-algebra which is uncountable.

The twin $(\Omega, \mathcal{F})$ is called a *measurable space*. This sort of defines the "parts" of our problem which are measurable, as per Definition 1.1. What's missing is a measure, i.e. the thing that actually tells us 'how long a piece of string is', so to speak[6]. We now define a measure as follows.

Definition 1.2 (Measure)**.** A *measure* μ is a non-negative real valued function defined on a σ-algebra, i.e. $\mu : \mathcal{F} \to \mathbb{R}_{\geq 0} \cup \{\infty\}$, where $\mathbb{R}_{\geq 0}$ are the non-negative real numbers and $\mathcal{F}$ a σ-algebra of subsets of Ω. The measure μ satisfies the following properties:

i. $\mu(\{\}) = 0$.
ii. μ is countably additive, i.e. if $A_1, A_2, ...$ are disjoint events, then
$$\mu\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i).$$

If, in addition the measure of the entire sample space is normalised (i.e. $\mu(\Omega) = 1$), then μ is called a **probability measure**. We will see this in the next section.

The triplet $(\Omega, \mathcal{F}, \mu)$ is called a *measure space* (note that without the measure it is called a measurable space). This space simply tells us the parts needed for well-defined measure to take place on the subsets of Ω.

Example 1.6. 1. The counting measure. Let Ω be a countable set [You may be creative as you like here to make this less abstract, e.g. the books on your shelf or the members of your family, although the set need not be finite]. Let $\mathcal{F} = \mathcal{P}(\Omega)$ be the power set of Ω. For all sets $A \in \mathcal{A}$, define
$$\mu(A) = \begin{cases} |A| & A \text{ has finitely many elements} \\ \infty & \text{otherwise} \end{cases}$$
where the operator $|\cdot|$ represents the *cardinality* of the set, i.e. the number of elements it contains (its size).

2. The Lebesgue measure in one dimension. Let $\Omega = \mathbb{R}$, and define $\mathcal{F}$ to contain all sets of the form

- [a,b], i.e. closed intervals,
- (a,b), i.e. open intervals,
- (a,b], i.e. open-closed intervals; and

[5] You may notice that other σ-algebras are indeed possible, e.g. the power set of Ω in this case. There is a notion of the *smallest* σ-algebra containing the collection of "basic events". Luckily for us, the event space that we will usually be working with will be the smallest σ-algebra without much technicalities, so we shall not explore this concept any further.

[6] https://idioms.thefreedictionary.com/How+long+is+a+piece+of+string%3F

- [a,b), i.e. closed-open intervals.

for all real numbers a and b. We can deduce that the σ-algebra $\mathcal{F}$ contains all possible "nice" intervals of the real line, including unbounded intervals and even singletons, which means any continuous partition of the real line can be measured (including a point, which should have measure zero). To see this, using the properties of σ-algebras,

- unbounded intervals are in $\mathcal{F}$, since, for instance

$$(x, +\infty) = \cup_{i=1}^{\infty}(x, x+i).$$

- singletons are in $\mathcal{F}$, since

$$\{x\} = \cap_{i=1}^{\infty}(x - 1/i, x + 1/i).$$

This set $\mathcal{F}$ has a special name, called the Borel σ-algebra.

All that's left is to define the measure. The Lebesgue measure μ assigns the usual concept of length to any continuous interval on $\mathbb{R}$ (to be precise, the Borel σ-algebra on $\mathbb{R}$):

$$\mu(A) = b - a$$

where A is any interval of $\mathbb{R}$ of the above forms (closed, open, open-closed, closed-open). This measure works even for singleton sets or unbounded intervals.

1.2.2 Axioms of probability

In the previous section, we defined a measure space as the triplet $(\Omega, \mathcal{F}, \mu)$. This formulation lets us work on the set of interest Ω, and defines the possible measurable subsets $\mathcal{F} \subseteq \Omega$, as well as the measuring device μ. This framework generalises the intutive notions of length, area, and volume to higher dimensions and more abstract notions.

In probability theory, we are interested in making use of measure theory to assign probabilities to events. So again in the context of conducting an "experiment",

- The sample space Ω is the set of possible outcomes $\{\omega_1, \omega_2, ...\}$ of the experiment.
- The σ-algebra $\mathcal{F} \subseteq \Omega$ would define the set of possible outcomes that are measurable, and are able to be assigned probabilities. $\mathcal{F}$ is known as the *event space*.

All that's left is to define a *probability measure* on the measurable space $(\Omega, \mathcal{F})$.

Definition 1.3 (Axioms of Probability)**.** Given a measurable space $(\Omega, \mathcal{F})$, a *probability measure* (or *probability function*) is a function $\mathbb{P} : \mathcal{F} \to [0, 1]$ that satisfies the following three conditions:

i. $\mathbb{P}(E) \geq 0, \forall E \in \mathcal{F}$.
ii. $\mathbb{P}(\Omega) = 1$.
iii. For pairwise disjoint events $A_1, A_2, ...$,

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} A_i.$$

These three conditions are commonly known as the *Axioms of Probability*, or *Kolmogorov Axioms*.

This is pretty much similar to the definition of the measure μ for a measure space, except for the unitarity requirement $\mathbb{P}(\Omega) = 1$. The first and second condition implicitly states that probabilities are always finite, by the results of Theorem 1.1 below. In contrast, measure theory allows for infinite measure.

The second condition above states that the probability of *at least* one of the elementary events in the entire space will definitely occur. One common misunderstanding here is to read the statement as "the probability of all of possible events occurring is 1", which is a rarer thing in most situations.

Figure 1.1: Andrey Nikolaevich Kolmogorov 25 April 1903–20 October 1987. Widely considered to be the father of probability theory.

As a remark, the above axioms does not tell us anything about what the functional form of $\mathbb{P}$ actually is. It is pretty abstract, but the good thing is that any such function that satisfies the above three axioms is by definition a probability function. At this point, there is still no notion of *randomness* in play. All we are doing is providing the building blocks to be able to assign a numerical representation of (un)certainty of some particular event happening.

As mentioned, any function abstract or concrete satisfying the Probability Axioms is regarded as a probability function. But what does the probability number represent, and what does it actually mean? Broadly speaking, there are two main interpretation of probabilities.

1. The **frequentist** interpretation is one that relies on "long run" frequencies. A probability of heads being 50% in a coin flip is interpreted to mean the following: If we flip the coin many times, then the proportion of heads that is observed will be 50% in the long run.

2. The **subjectivist** or Bayesian interpretation is that the probability measures an observer's strength of belief that the event is true. Put a different way, it is the measure of ignorance on the observers part on what has happened. When a coin is flipped, it has landed either heads or tails, and this much is certain. What is uncertain is my *knowledge about the coin*, rather than the outcome of the coin itself. Setting a 50% probability for heads occuring implies that I am willing to bet at a 1:1 odds that the coin landed heads.

Example 1.7 (C&B 1.2.5)**.** Consider the simple experiment of tossing a coin. Define the sample space to be $\Omega = \{H, T\}$, as representing the only two possible outcomes $H =$ heads or $T =$ tails.

What is the probability of heads occurring? The Axioms of Probability does not help us in this regard! (I mean, it does not give us a functional form for the probabilities)

Perhaps a function that assigns equal probability to either event would be a good place to start, so we require

$$\mathbb{P}(\{H\}) = \mathbb{P}(\{T\}). \tag{1.1}$$

At this point, we still don't know their values–the probabilities could be 0.1, 0.2, 0.3, or any other value. Or could they?

Since $\Omega = \{H\} \cup \{T\}$, we know that by the Probability Axioms that

$$\begin{aligned} 1 = \mathbb{P}(\Omega) &= \mathbb{P}\left(\{H\} \cup \{T\}\right) \\ &= \mathbb{P}(\{H\}) + \mathbb{P}(\{T\}) \end{aligned} \tag{1.2}$$

so the only possible value that satisfies both (1.1) and (1.2) is

$$\mathbb{P}(\{H\}) = \mathbb{P}(\{T\}) = 0.5.$$

Of course, without the restriction of equal probability in (1.1), then any two numbers satisfying (1.2) and the Probability Axioms would be valid, e.g. $\mathbb{P}(\{H\}) = 0.8$ and $\mathbb{P}(\{T\}) = 0.2$.

Example 1.8. Two six-sided dice are thrown and the outcome for both dice are recorded.

- As there are 36 possible outcomes, the sample space is

$$\Omega = \{\omega_{ij} = \{i, j\} \mid i, j = 1, \dots, 6\}$$

- Suppose we are interested in the event E defined to be *'the sum of the two scores is 6'*. These would be the events

$$E := \{\{1, 5\}, \{2, 4\}, \{3, 3\}, \{4, 2\}, \{5, 1\}\}.$$

  One may easily construct a σ-algebra $\mathcal{F}$ (for example, the power set of Ω) and verify that the event E is contained within it. So this is a measurable event.

So at this point, we might be thinking about a suitable probability function so that we may assign a probability to the event E. Especially if the two dice are fair, it seems reasonable to assume that any of the outcome in $\omega_{ij} \in \Omega$ is equally likely to occur, so we set $\mathbb{P}(\omega_{ij}) = 1/36$ for any $i, j = 1, ..., 6$. In particular, the probability of any event should be proportional to the total number of outcomes in Ω. As a quick exercise, you may check that such a probability function satisfies all the Kolmogorov Axioms.

$$\begin{aligned}\mathbb{P}(E) &= \mathbb{P}\left(\{1,5\} \cup \{2,4\} \cup \{3,3\} \cup \{4,2\} \cup \{5,1\}\right) \\ &= \mathbb{P}(\{1,5\}) + \mathbb{P}(\{2,4\}) + \mathbb{P}(\{3,3\}) \\ &\quad + \mathbb{P}(\{4,2\}) + \mathbb{P}(\{5,1\}) \\ &= \frac{5}{36}\end{aligned}$$

Alternatively, we could have also easily argued that

$$\mathbb{P}(E) = \frac{|E|}{|\Omega|} = \frac{5}{36}.$$

As a remark, it would be very cumbersome to have to check the Kolmogorov Axioms every time we encounter a probability function. For problems like the above, we won't run into any technical issues because the sample space is finite and/or countable[7]. In general, most of the problems we will come across will satisfy the axioms automatically, especially with "nice" sample space and events, so we usually don't check axioms all the time.

At this point, most textbooks go into a section about *counting*, namely using methods like combinations and permutations. I'm sure you've encountered this previously in your statistics classes, and appreciate how useful they are when trying to calculate probabilities as being "the number of outcomes in the event space" divided by "the number of outcomes in the sample space". However, our focus for this course is to get to the inference section, and the topic of counting does not contribute much to that understanding, so I shall skip it.

1.2.3 Derived probability results

Let us now look at some useful probability results that can be derived from the probability axioms.

Theorem 1.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. For any $E \in \mathcal{F}$,*

i. $\mathbb{P}(\{\}) = 0$;
ii. $\mathbb{P}(E) \leq 1$; *and*
iii. $\mathbb{P}(E^c) = 1 - \mathbb{P}(E)$.

An important thing that we learn here is that probabilities are always finite and bounded within $[0, 1]$, i.e. for any event E, $0 \leq \mathbb{P}(E) \leq 1$. So please, do not make the mistake of reporting *negative probabilities* or probabilities greater than one–they are mathematically impossible[8]!

The proof of Theorem 1.1 is left an exercise. Try this out for yourself!

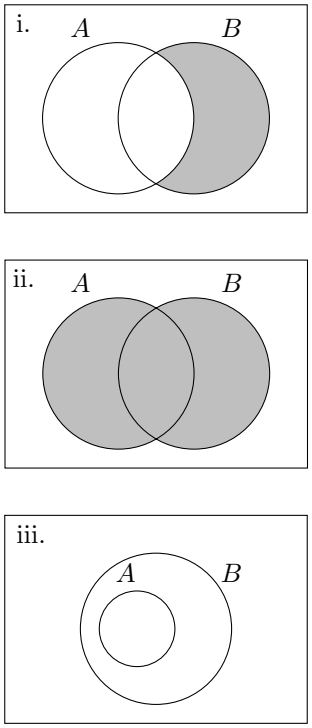
Further results regarding two events in the sample space based on the Probability Axioms:

Theorem 1.2. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. For any $A, B \in \mathcal{F}$,*

i. $\mathbb{P}(B \cap A^c) = \mathbb{P}(B) - \mathbb{P}(A \cap B)$;
ii. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$; *and*
iii. *If $A \subseteq B$, then* $\mathbb{P}(A) \leq \mathbb{P}(B)$.

[7] See Theorem 1.2.6 in C&B.
[8] At least within the framework of the Kolmogorov Axioms. See: https://en.wikipedia.org/wiki/Negative_probability

While these results are not so self-evident from the Probability Axioms, it may be useful to employ Venn diagrams to visualise the above statements.



Proof. i. Note that B is composed of the two disjoint sets $B = \{B \cap A\} \cup \{B \cap A^c\}$, so we have

$$\mathbb{P}(B) = \mathbb{P}(B \cap A) + \mathbb{P}(B \cap A^c),$$

and the desired results is obtained after rearranging.

ii. Using the identity

$$A \cup B = (A \cup B) \cap \overbrace{(A \cup A^c)}^{\Omega} = A \cup \{B \cap A^c\},$$

we have that (since the two events are disjoint)

$$\begin{aligned} \mathbb{P}(A \cup B) &= \mathbb{P}(A) + \mathbb{P}(B \cap A^c) \\ &= \mathbb{P}(A) + \mathbb{P}(B \cap A^c) \\ &= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(B \cap A). \end{aligned}$$

iii. Since $A \subseteq B$, $A \cap B = A$, using i. we get

$$0 \leq \mathbb{P}(B \cap A^c) = \mathbb{P}(B) - \overbrace{\mathbb{P}(A \cap B)}^{\mathbb{P}(A)},$$

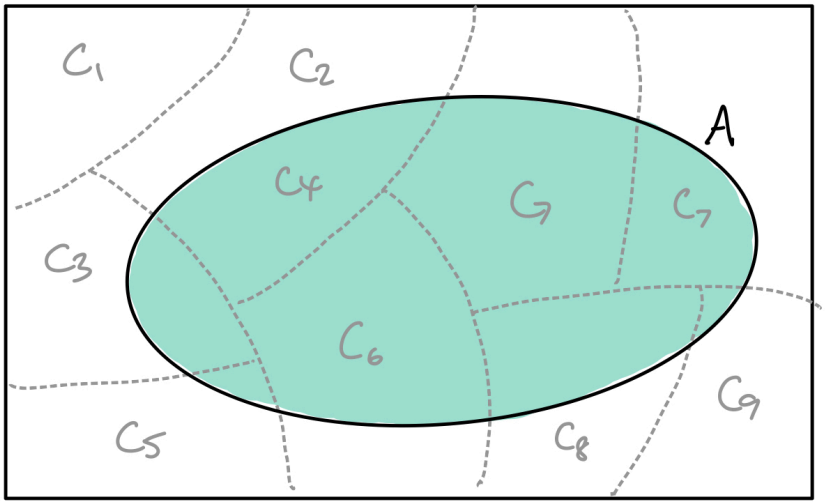
thus obtaining the desired result.

□

Theorem 1.3 (Law of Total Probability)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $A \in \mathcal{F}$ and consider a (countably infinite) partition of the sample space $C_1, C_2, \ldots$ such that $C_i \cap C_j = \{\}$ for any i, j and $\bigcup_{i=1}^{\infty} C_i = \Omega$. Then*

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A \cap C_i).$$

We may visualise the partitions of the sample space C_i as well as the event A of interest as follows:



Of course, we can only show a finite number of partitions for illustration, but this works for infinitely many countable partitions as well. We can see that the set A is simply made up of the intersections of A and the partitions. Some of these intersections will be empty, but that's OK.

Proof. Write

$$A = A \cap \Omega = A \cap \left(\bigcup_{i=1}^{\infty} C_i \right) = \bigcup_{i=1}^{\infty} (A \cap C_i)$$

Evidently the events in the union on the right hand side of the equality are disjoint, since C_i themselves are disjoint. Therefore,

$$\begin{aligned} \mathbb{P}(A) &= \mathbb{P}\left(\bigcup_{i=1}^{\infty} (A \cap C_i) \right) \\ &= \sum_{i=1}^{\infty} \mathbb{P}(A \cap C_i) \end{aligned}$$

as required. □

1.2.4 Why measure theory?

You may treat this section as optional, but it would deepen your understanding of probability theory.

Consider the uniform distribution on a random variable X on the unit interval, denoted $X \sim \text{Unif}(0, 1)$. You may have come across this before, and know that the probability that X lies in any interval contained in $[0, 1]$ is simply the length of the interval, i.e.

$$\mathbb{P}([a, b]) = \mathbb{P}([a, b)) = \mathbb{P}((a, b]) = \mathbb{P}((a, b)) = b - a, \tag{1.3}$$

for $0 \leq a \leq b \leq 1$. This definition works fine for the degenerate case $\mathbb{P}(\{a\}) = 0$ for the singleton set $\{a | a \in (0, 1)\}$. In general, if A and B are disjoint subsets of $[0, 1]$ then

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B), \tag{1.4}$$

and we can even extend this notion to that of *countable additivity*

$$\mathbb{P}\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i), \tag{1.5}$$

for disjoint sets $\{A_1, A_2, ...\}$[9].

For a uniform measure on $[0,1]$, one expects that the measure of some subset $A \subseteq [0,1]$ to be unaffected by "shifting" (with wrap-around) of that subset by some fixed amount $r \in [0,1]$. Define the *r-shift* of $A \subseteq [0,1]$ by

$$A \oplus r := \{a + r \mid a \in A, a + r \leq 1\} \cup \{a + r - 1 \mid a \in A, a + r > 1\}.$$

Then we should have

$$\mathbb{P}(A \oplus r) = \mathbb{P}(A). \tag{1.6}$$

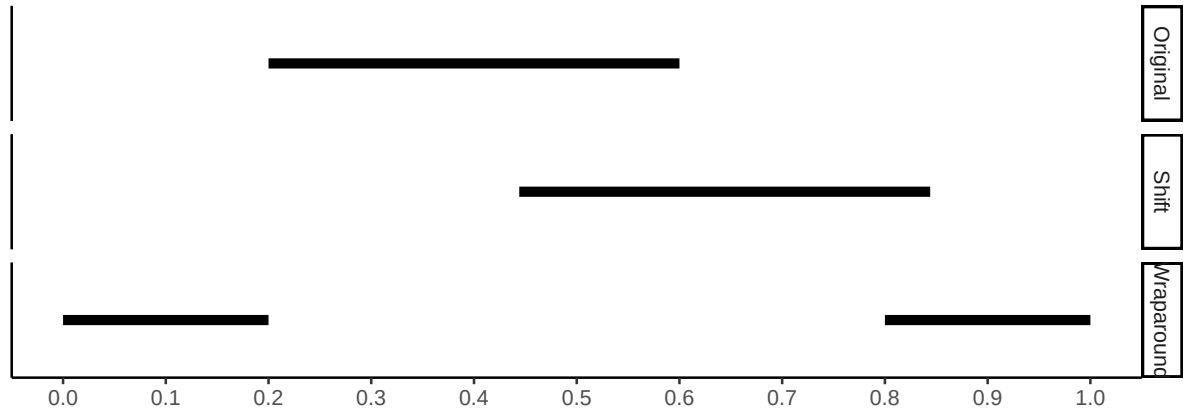


Figure 1.2: An interval in [0,1] shifted by some fixed amount, with wrap-around, should have consistent length.

At this point you might notice that all of this resonates with the previous example on the Lebesgue measure, except perhaps the shifting part, and indeed that is the case. Suppose that we dispense with measure theory and do not define things like the σ-algebra on the $[0,1]$ or the triplet $(\Omega, \mathcal{F}, \mathbb{P})$, and only use the above probability definitions given in (1.3), (1.5), and (1.6). How far can we push the boundaries of such probability definitions before things start to breakdown?

Consider these questions:

- What is the probability that X is rational?
- What is the probability that X^n is rational for some positive integer n?
- What is the probability that X is *algebraic*[10]?

All seemingly fair and interesting questions, but are they well defined? Can we actually measure them and assign probabilities to such events? Taking a step back further, we ask:

> Are all possible subsets $A \subseteq [0,1]$ measurable? Does $\mathbb{P}(A)$ even make *sense* for any event A we can think of?

It turns out the answer is no, and can be proven by contradiction with the help of equivalence relations. This shows the need for the heavy machinery that is measure theory for assigning probabilities to events[11].

Proposition 1.1. *There does not exist a definition of* $\mathbb{P}(A)$*, defined for all subsets* $A \subseteq [0,1]$*, satisfying* (1.3)*,* (1.5)*, and* (1.6)*.*

Proof. All we need to show is the existence of one such subset of $[0,1]$ whose measure is undefined. The set we are about to construct is called the Vitali set[12], after Giuseppe Vitali who described it in 1905.

[9] A concrete example of this is for the sets $A_1 = (0, 1/2)$, $A_2 = (1/2, 3/4)$, $A_3 = (3/4, 7/8)$, and so on (adding half the interval at each iteration). One finds that the measure of the countable union is $\sum_{i=1}^{\infty}(1/2)^i = 1$.

[10] An algebraic number is a number that is a root of a non-zero polynomial in one variable with integer coefficients.

[11] Or at least, for cases where "not so nice" events need to be measured.

[12] https://en.wikipedia.org/wiki/Vitali_set

Define an equivalence relation on $[0,1]$ by the following:

$$x \sim y \Rightarrow y - x \in \mathbb{Q}$$

That is, two real numbers x and y are deemed to be the same if their difference is a rational number. We would like to separate all the real numbers $x \in [0,1]$ by this equivalence relation, and collect them into groups called equivalence classes, denoted by $[x]$. Here, $[x]$ is the set $\{y \in [0,1] \mid x \sim y\}$. For instance,

- The equivalence class of 0 is the set of real numbers x such that $x \sim 0$, i.e. $[0] = \{y \in [0,1] \mid y-0 \in \mathbb{Q}\}$, which is the set of all rational numbers in $[0,1]$.
- The equivalence class of an irrational number $z_1 \in [0,1]$ is clearly not in $[0]$, thus would represent a different equivalent class $[z_1] = \{y \in [0,1] \mid y - z_1 \in \mathbb{Q}\}$.
- Yet another irrational number $z_2 \notin [z_1]$ would exist, i.e. a number $z_2 \in [0,1]$ such that $z_2 - z_1 \notin \mathbb{Q}$, and thus would represent another equivalence class $[z_2]$.
- And so on... The equivalence classes may be represented by $[0], [z_1], [z_2], ...$ where z_i are all irrational numbers that differ by an irrational number, and there are uncountably many such numbers and therefore classes.

Construct the Vitali set V as follows: Take precisely one element from each equivalent class, and put it in V. As a remark, such a V must surely exist by the Axiom of Choice[13].

Consider now the union of shifted Vitali sets by some rational value $r \in [0,1]$,

$$\bigcup_r (V \oplus r)$$

As a reminder, the set of rational numbers is countably infinite[14]. We make a few observations:

- The equivalence relation partitions the interval $[0,1]$ into a disjoint union of equivalence classes. In other words, the sets $(V \oplus r)$ and $(V \oplus s)$ are disjoint for any rationals $r \neq s$, such that $r, s \in [0,1]$. If they were not disjoint, this would mean that there exists some $x, y \in [0,1]$ with $x + r \in (V \oplus r)$ and $y + s \in (V \oplus s)$ such that $x + r = y + s$. But then this means that $x - y = s - r \in \mathbb{Q}$ so x and y are in the same equivalent class, and this is a contradiction.
- Every point in $[0,1]$ is contained in the union $\bigcup_r (V \oplus r)$. To see this, fix a point x in $[0,1]$. Note that this point belongs to some equivalent class of x, and in this equivalence class there exists some point α which belongs to V as well by construction. Hence, $\alpha \sim x$, and thus $x - \alpha = r \in \mathbb{Q}$, implying that x is a point in the Vitali set V shifted by r. Therefore,

  $$[0,1] \subseteq \bigcup_r (V \oplus r).$$

  and we may write

  $$1 = \mathbb{P}([0,1]) \leq \mathbb{P}\left(\bigcup_r (V \oplus r)\right),$$

  since the measure of any set contained in another must have smaller or equal measure. This relation is in fact implied by (1.5). Let A and B be such that $A \subseteq B$. Then we may write $B = A \cup (B - A)$ where the sets A and $B - A$ are disjoint. Hence, $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B - A)$, and since measures are non-negative and in particular $\mathbb{P}(B - A) \in [0,1]$, we have that $\mathbb{P}(B) \geq \mathbb{P}(A)$. However since the probability measure cannot be greater than 1, it must be equal to 1.
- The disjoint union $\bigcup_r (V \oplus r)$ has probability measure (according to our definitions in (1.3), (1.5), and (1.6))

  $$\begin{aligned} \mathbb{P}\left(\bigcup_r (V \oplus r)\right) &= \sum_r \mathbb{P}(V \oplus r) \\ &= \sum_r \mathbb{P}(V) \end{aligned}$$

[13] Given a collection of non-empty sets, it is always possible to construct a new set by taking one element from each set in the original collection. See https://brilliant.org/wiki/axiom-of-choice/

[14] https://www.homeschoolmath.net/teaching/rational-numbers-countable.php

Putting these three observations together gives us

$$1 = \mathbb{P}\left(\bigcup_r (V \oplus r)\right) = \sum_r \mathbb{P}(V).$$

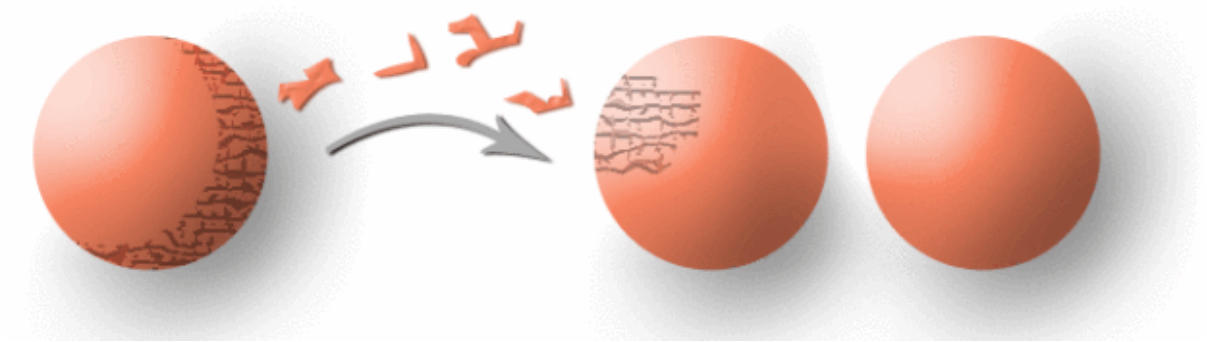
This leads to the desired contradiction: A countably infinite sum of the same quantity repeated can only equal 0, $+\infty$, or $-\infty$, but it can never equal 1. □

In summary,

- Not all subsets of uncountable sets are measurable. Admitting all subsets of uncountable sets will break mathematics.
- σ-algebras are the patch that fixes mathematics. It gatekeeps the subsets of uncountable sets and disregards those which are not measurable.
- Actually, if you have been following along, you might realise that we are at risk of breaking mathematics when dealing with uncountable sets. Strictly speaking, we only need σ-algebras when working in a set with uncountable cardinality.

Finally, what on earth is an "unmeasurable" set? Wouldn't it be (even arbitrarily) possible to just define a measure for whatever set we can think of? If the above example hasn't convinced you enough, some other mathematicians have tried to resolve this but it seems it is not possible to do so without encountering paradoxes, such as the one below.

> The Banach–Tarski paradox states that a ball in the ordinary Euclidean space can be doubled using only the operations of partitioning into subsets, replacing a set with a congruent set, and reassembly.



To be clear, no rule of mathematics are broken in the Banach-Tarski paradox, but the result defies intuition. Another statement of this paradox is that *we can chop up a pea into finitely many pieces and reassemble it into the sun* (pea-sun paradox). If we don't lay out the foundations for measuring probabilities rigorously, we can end up with nonsensical answers!

This section was highly inspired by the following references:

- Rosenthal, J. (2006). A first look at rigorous probability.
- The discussion here: https://stats.stackexchange.com/q/199280
- This YouTube video on Vitali Sets: https://youtu.be/ameugr-wjeI

1.3 Conditioning and independence

In the previous section, the probabilities we encountered are *unconditional*, in the sense that the probabilities do not depend on any other external factors or information, and only on the (fixed) information in the sample space. In contrast, we may talk about *conditional probabilities*. If the sample space gets updated based on observation of new information, then this will sure impact probability calculations.

Definition 1.4 (Conditional probabilities)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. For any $A, B \in \mathcal{F}$ such that $\mathbb{P}(B) > 0$, the *conditional probability* of A given B, written $\mathbb{P}(A|B)$, is defined to be

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

The event B is known as the *conditioning event*. For all intents and purposes, we may view $\mathbb{P}(A|B)$ as "the probability that A occurs, given that we know that B has already occurred". In this sense, the given information forms an updated sample space (as $\mathbb{P}(B|B) = 1$): All further occurrences are calibrated with respect to their relation to B. Thus, $\mathbb{P}(A|B)$ as *the fraction of times A occurs among those in which B occurs.*

Note that for mutually exclusive events A and B, $\mathbb{P}(A|B) = \mathbb{P}(B|A) = 0$ since $\mathbb{P}(A \cap B) = 0$. This makes sense because as the two events are disjoint, they have "nothing to do with each other".

In general,

$$\mathbb{P}(A|B) \neq \mathbb{P}(A).$$

This is only true when dealing with independent events. Furthermore, in general

$$\mathbb{P}(A|B) \neq \mathbb{P}(B|A).$$

Example 1.9. A medical test for a disease D has outcomes '+' and '−'. The probabilities are as follows:

	D	D^c
$+$	0.009	0.099
$-$	0.001	0.891

note: each cell represents $\mathbb{P}(A \cap B)$.

From the definition of conditional probability,

$$\mathbb{P}(+|D) = \frac{\mathbb{P}(+ \cap D)}{\mathbb{P}(D)} = \frac{0.009}{0.009 + 0.001} = 0.90$$

and

$$\mathbb{P}(-|D^c) = \frac{\mathbb{P}(- \cap D^c)}{\mathbb{P}(D^c)} = \frac{0.891}{0.099 + 0.891} \approx 0.90.$$

Suppose you go for a test and get a positive result. What is the probability you have the disease? Most will answer 0.90. Actually,

$$\mathbb{P}(D|+) = \frac{\mathbb{P}(D \cap +)}{\mathbb{P}(+)} = \frac{0.009}{0.009 + 0.099} = 0.08.$$

Notice that

- $\mathbb{P}(D \cap +) = \mathbb{P}(+|D)\mathbb{P}(D)$ after some rearranging; and
- $\mathbb{P}(+) = \mathbb{P}(+ \cap D) + \mathbb{P}(+ \cap D^c)$ since D and D^c are disjoint.

We can therefore write

$$\mathbb{P}(D|+) = \frac{\mathbb{P}(+|D)\mathbb{P}(D)}{\mathbb{P}(+|D)\mathbb{P}(D) + \mathbb{P}(+|D^c)\mathbb{P}(D^c)}.$$

For $\mathbb{P}(D|+)$ to be large, it seems $\mathbb{P}(D)$ needs to be large in addition to $\mathbb{P}(+|D)$, i.e. disease is prevalent.

1.3.1 Bayes Theorem

Following that previous example, and from the definitions of conditional probabilities, we have that, after some rearranging,

$$\mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(A \cap B),$$

and

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

So equating the two together, one can relate the two conditional probabilities $\mathbb{P}(A|B)$ and $\mathbb{P}(B|A)$ by

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

Furthermore, by using the law of total probability, we can now state Bayes' Theorem.

Theorem 1.4 (Bayes' Theorem)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $A_1, A_2, ...$ a partition of the sample space, and B be any set in $\mathcal{F}$ such that $\mathbb{P}(B) > 0$. Then, for each $i = 1, 2, ...,$*

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{j=1}^{\infty} \mathbb{P}(B|A_j)\mathbb{P}(A_j)}.$$

The above rule provides a convenient way of computing conditional probability $\mathbb{P}(A|B)$ if knowledge regarding the "reverse" conditional probability $\mathbb{P}(B|A)$ is readily available.

Figure 1.3: (Probably not) Rev. Thomas Bayes c. 1701–7 April 1761. This picture is commonly used to depict Thomas Bayes, but historians believe this not to be an accurate depiction.

Some will call $\mathbb{P}(A_i)$ the *prior probability*, and the $\mathbb{P}(A_i|B)$ *posterior probability*, especially in the context of Bayesian statistics. The terms refer to our state of knowledge before and after learning new information (respectively) that is used to update our beliefs.

Example 1.10. In a certain selection of flower seeds, 2/3 have been treated to improve germination and 1/3 have been left untreated. For the purpose of this example, we may treat these numbers as probabilities of selecting a treated or untreated flower seed.

Furthermore, the seeds which have been treated have a probability of germination of 0.8, whereas the untreated seeds have a probability of germination of 0.5.

Let's calculate the probability that a seed, selected at random:

(a) will germinate (assuming the seeds were sown and given time to germinate).
(b) a germinated seed had been treated.

First, let us define the following events:

- T = a seed has been treated
- T^c = a seed has not been treated
- G = a seed has germinated
- G^c = a seed has not germinated

We note that the events T and T^c are disjoint and partitions the sample space (a seed can either be treated or not), and so too the case with G and G^c. After some careful reading of the question, we are actually presented with the probabilities $\mathbb{P}(G|T) = 0.8$ and $\mathbb{P}(G|T^c) = 0.5$.

To answer a., we require $\mathbb{P}(G)$, which is obtained using the law of total probability:

$$\begin{aligned}\mathbb{P}(G) &= \mathbb{P}(G \cap T) + \mathbb{P}(G \cap T^c) \\ &= \mathbb{P}(G|T)\mathbb{P}(T) + \mathbb{P}(G|T^c)\mathbb{P}(T^c) \\ &= 2/3 \times 0.8 + 1/3 \times 0.5 = 0.7\end{aligned}$$

In answering b., we realise that we are after the quantity $\mathbb{P}(T|G)$. Using Bayes' Theorem,

$$\begin{aligned}\mathbb{P}(T|G) &= \frac{\mathbb{P}(G|T)\mathbb{P}(T)}{\mathbb{P}(G)} \\ &= \frac{0.8 \times 2/3}{0.7} \\ &= 0.762\end{aligned}$$

It's important to note here that $\mathbb{P}(G|T) \neq 1 - \mathbb{P}(G|T^c)$, and this is true in most cases. We cannot take complements with respect to the conditioning event!

1.3.2 Independence

In some cases, the occurrence of a particular event B has *no effect* on the probability of another event A. Mathematically, we can denote this as

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

If this were true, we can use the relationship $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$ to derive the following definition.

Definition 1.5. Two events A and B are *statistically independent* if and only if

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

What's nice about this definition is that in order to check whether to events are independent, it is sufficient to check whether their probabilities multiply out in the manner above. Note that (and it is easily checked!) that if A and B are independent then so too are

- A and B^c;
- A^c and B; and
- A^c and B^c.

Here's an experiment we can do to examine the concept of independent events. Consider tossing a fair die. Let $A = \{2, 4, 6\}$ and $B = \{1, 2, 3, 4\}$. You should be able to work out, using the above probability results and the definition of conditional probabilities, that $\mathbb{P}(A) = 1/2$, $\mathbb{P}(B) = 2/3$, and $\mathbb{P}(A \cap B) = 1/3$. Hence,

we deduce that A and B are independent, since the product of each probability event is the probability of their intersection.

If you were feeling bored and had a lot of time to spare, you could verify this empirically using an actual die. While this would be an afternoon well spent, let's use `R` to simulate some draws from the sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$, and count the number of times each events A, B and $A \cap B$ occurs.

```
# Throw a dice 10 times
sample(1:6, size = 10, replace = TRUE)
```

```
##  [1] 3 6 3 2 2 6 3 5 4 6
```

From the above, $n(A) = 6$, $n(B) = 6$, and $n(A \cap B) = 3$. Here I've used the notation $n(\cdot)$ to mean the count of the event. Do this 1,000 times, and count events automatically using the following code.

```
x <- sample(1:6, size = 1000, replace = TRUE)
head(x, 100)  # show the first 100 outcomes
```

```
##   [1] 6 1 2 3 5 3 3 1 4 1 1 5 3 2 2 1 6 3 4 6 1 3 5 4 2 5 1 1 2 3 4 5 5 3 6 1 2
##  [38] 5 5 4 5 2 1 1 3 1 6 5 1 2 4 4 6 6 3 6 6 1 6 2 1 2 4 5 5 6 3 1 4 6 1 6 1 3
##  [75] 6 4 1 6 6 3 6 5 3 6 2 5 5 3 2 2 2 4 2 2 6 4 4 6 1 6
```

```
nA <- sum(x %in% c(2, 4, 6))  # counts the frequency of 2, 4, 6
nB <- sum(x %in% c(1, 2, 3, 4))  # counts the frequency of 1, 2, 3, 4
nAB <- sum(x %in% c(2, 4))  # counts the frequency of 2, 4

# Results
c(A = nA, B = nB, AnB = nAB) / 1000
```

```
##     A     B   AnB
## 0.495 0.674 0.333
```

Empirically, we have $\hat{\mathbb{P}}(A)\hat{\mathbb{P}}(B) = 0.495 \times 0.674 = 0.33363$. This matches with the value of $\hat{\mathbb{P}}(A \cap B)$ in the table, as well as the theoretical value of $1/3$.

1.4 Random variables

Consider the following problem: Ask (randomly) 50 people whether they like (code this as "1") or dislike (code this as "0") learning statistics. What is the sample space for this experiment? This would be all 1/0 combinations such as

$$\overbrace{1000101 \cdots 10001}^{50}$$

Specifically, $\Omega = \{(X_1, X_2, \ldots, X_{50}) \,|\, X_i \in \{0, 1\}\}$. Realise that $|\Omega| = 2^{50}$. This is huge[15]!

```
2 ^ {50}
```

```
## [1] 1.1259e+15
```

[15] For context, the average American, working full-time, would have to work 25 billion years to earn 1 quadrillion dollars.

Is it practical to work with such a large sample space? Possibly not, even with fancy counting techniques.

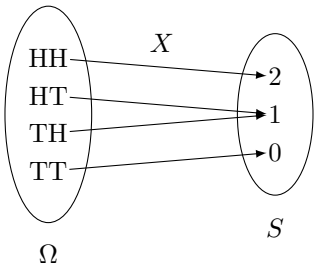
But what if we instead defined a variable $Y = \sum_{i=1}^{50} X_i$? Here, Y is the count of the number of people who like learning statistics from this sample of 50, since it only counts the values of '1's occuring. Further, the minimum value for Y is 0, and the maximum is 50. So the new sample space associated with Y is $S = \{0, 1, 2, \dots, 50\}$–much easier to deal with!

Y is defined to be a mapping from the original sample space Ω to the new space S (usually a set of real numbers). Such a mapping is called a **random variable**.

Definition 1.6 (Random variable)**.** A *random variable* X (abbreviated r.v.) on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a *measurable function*[16] from $(\Omega, \mathcal{F})$ to $\mathbb{R}$, i.e. $X : \Omega \to \mathbb{R}$.

This is much easier exaplained with an example.

Example 1.11. Flip a coin twice and let X be the number of heads. The sample space of the coin flips is $\Omega = \{\text{HH}, \text{HT}, \text{TH}, \text{TT}\}$. The sample space of X is $S = \{0, 1, 2\}$. The mapping of the random variable is illustrated as follows:



The qualifier *random* to the term 'random variable' implies that its value is not known before observing it. Random variables are conventionally denoted with uppercase letters, and the realised values of the variable will be denoted by the corresponding lowercase letters. Thus, the random variable X can take the value x.

We can see that a r.v. X assigns a real number $X(\omega)$ to each outcome ω. Can we still calculate probabilities of events? Yes.

$$\mathbb{P}(X = x) = \mathbb{P}(X^{-1}(x)) = \mathbb{P}(\{\omega \in \Omega \,|\, X(\omega) = x\})$$

More generally,

$$\mathbb{P}(X \in S) = \mathbb{P}(X^{-1}(x)) = \mathbb{P}(\{\omega \in \Omega \,|\, X(\omega) \in S\})$$

Example 1.12. For the previous example, the random variable X can be summarised as follows:

ω	$\mathbb{P}(\{\omega\})$	$X(\omega)$
TT	1/4	0
TH	1/4	1
HT	1/4	1
HH	1/4	2

x	$\mathbb{P}(X = x)$	$X^{-1}(x)$
0	1/4	TT
1	1/2	TH, HT
2	1/4	HH

In either case, the sum of the probabilities, whether in the original event space Ω or in the range of the random variable S, is equal to one.

[16] A measurable function is simply a function between the underlying sets of two measurable spaces. This will help preserve the structure of the spaces and allow things to be measured. See Wasserman, Appendix 2.13.

1.4.1 Distribution functions

With every random variable X, we associate a function called the cumulative distribution function of X.

Definition 1.7. The *cumulative distribution function (cdf)* of a r.v. X, denoted F_X, is the function $F_X : \mathbb{R} \to [0, 1]$ defined by

$$F_X(x) = \mathbb{P}(X \leq x), \text{ for all } x.$$

The cdf is sometimes just referred to as the *distribution function*. When there is no ambiguity regarding which random variable the cdf is referring to, we may drop the subcript in F_X.

Equivalently, the distribution function is written

$$F_X(x) = \mathbb{P}\left(\{\omega \in \Omega \mid X(\omega) \leq x\}\right).$$

We make some observations regarding the distribution function:

- $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to +\infty} F(x) = 1$.
- $F(x)$ is non-decreasing, i.e. $x_1 < x_2 \Rightarrow F(x_1) \leq F(x_2)$. In other words, drawing the function from left to right, it must either increase or stay the same value, but not decrease in value.
- $F(x)$ is right-continuous: for every x_0, $\lim_{x \downarrow x_0} F(x) = F(x_0)$. This means "the solid dots will be on the left of the distribution function".
- F itself *can be discontinuous* (see the next example). This is associated with whether the r.v. X is continuous or not. That is,
  - $F_X(x)$ is a continuous function $\Rightarrow$ X is continuous.
  - $F_X(x)$ is a step function $\Rightarrow$ X is discrete.

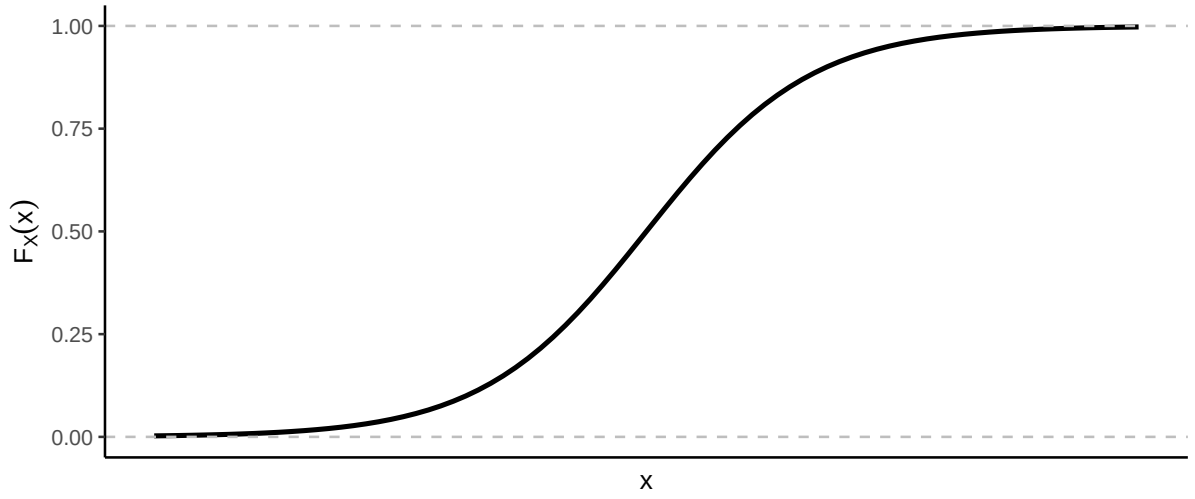


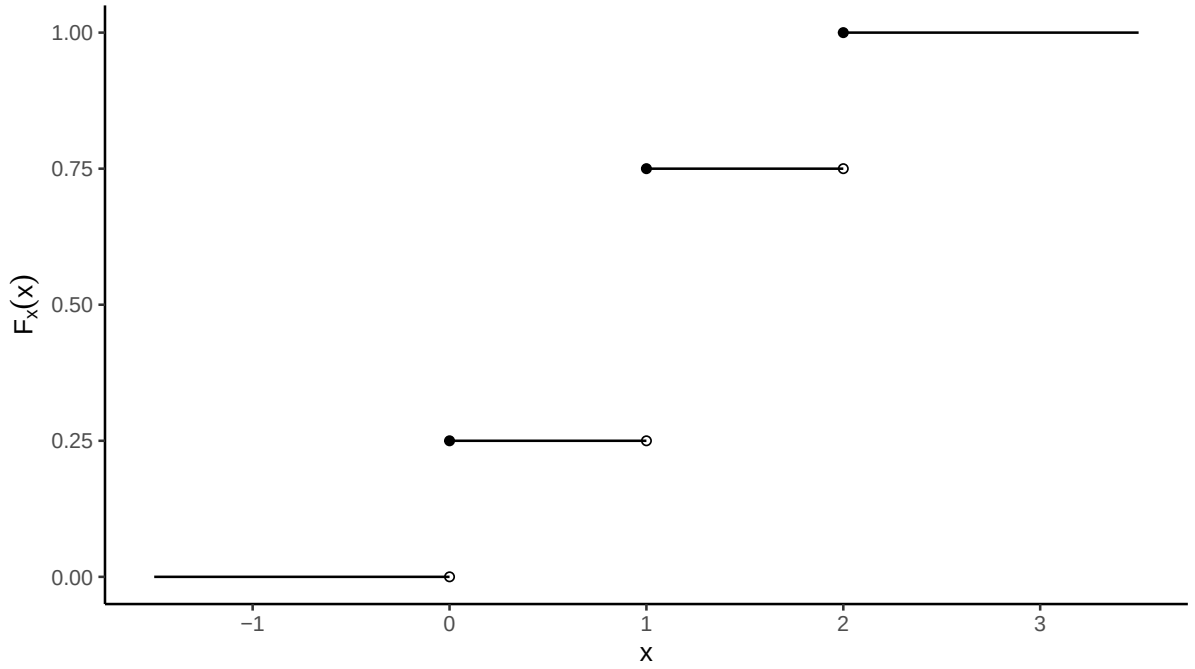
Figure 1.4: A general sketch of a (continuous) cdf.

Once again, the definition above does not give a functional form for the cdf, but the good news is that any function satisfying the above properties is a cdf. For proofs of these facts, see the reference textbooks.

Example 1.13. From Example 1.12, we have that

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 0.25 & 0 \leq x < 1 \\ 0.75 & 1 \leq x < 2 \\ 1 & x \geq 2 \end{cases}$$

This can be sketched as follows:



1.4.2 Identically distributed r.v.

Definition 1.8 (Identically distributed r.v.)**.** Let X have cdf F and let Y have cdf G. If $F(x) = G(x)$ for all x, then $\mathbb{P}(X \in A) = \mathbb{P}(Y \in A)$ for all (measurable) sets A. X and Y are said to be *identically distributed*.

Note that two identically distributed r.v. are not necessarily equal in value, only the probabilities of observing the same values are identical. Think about two independent coin flips. The probability of H/T in each flip is the same, but the outcome may not be.

Example 1.14. Consider again the experiment of tossing a coin twice. Define the random variables X and Y to be the number of heads and tails observed, respectively. The distribution of X, as we calculated previously, is

x	0	1	2
$\mathbb{P}(X = x)$	1/4	1/2	1/4

One can easily verify that the distribution of Y is

y	0	1	2
$\mathbb{P}(Y = y)$	1/4	1/2	1/4

Thus, for each $k = 0, 1, 2$, $\mathbb{P}(X = k) = \mathbb{P}(Y = k)$, so X and Y are identically distributed.

1.5 Probability functions

Going forward, we will be concentrating more on random variables and their distributions, rather than working in a probability space. While this is the case, hopefully you appreciate the probability theory that is going on behind the scenes.

Associated with a random variable X and its cdf F_X is another function, called either the probability density function (pdf) if it is continuous, or the probability mass function (pmf) if it is discrete. We shall look at both in turn.

1.5.1 Probability mass function

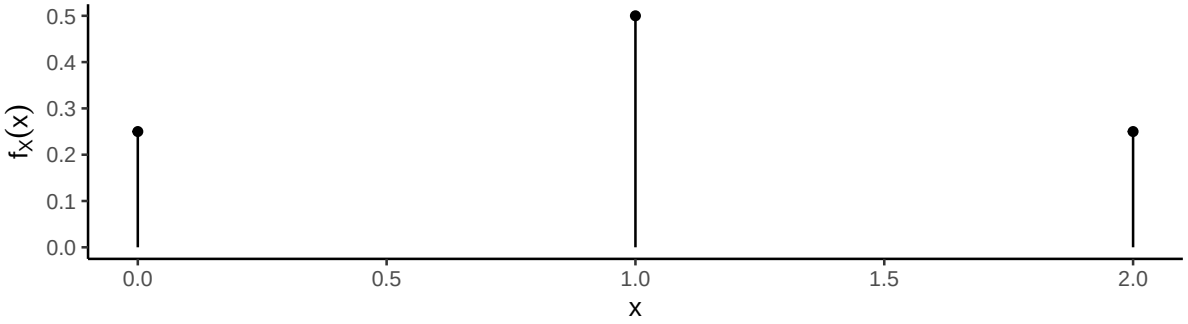
Definition 1.9 (Probability mass function)**.** A discrete random variable X may only take countably many values $\mathcal{X} = \{x_1, x_2, \dots\}$. Its *probability mass function (pmf)* is defined to be

$$f_X(x) = \mathbb{P}(X = x), \text{ for all } x \in \mathcal{X}.$$

The pmf is a function which takes input possible values that a random variable may take, and outputs the probability that the random variable takes that value. An example is given below.

Example 1.15. The pmf from Example 1.12 (the two coin flips) is given by

$$f_X(x) = \begin{cases} 1/4 & x = 0 \\ 1/2 & x = 1 \\ 1/4 & x = 2 \\ 0 & \text{otherwise} \end{cases}$$



Pmfs measure "point probabilities". Since outcomes of discrete random variables are countable, we can add up probabilities over all the points in the event. That is, for any a, b both in $\mathcal{X}$ such that $a \leq b$, we have that

$$\begin{aligned} \mathbb{P}(a \leq X \leq b) &= \mathbb{P}(X = a) + \cdots + \mathbb{P}(X = b) \\ &= \sum_{x=a}^{b} f_X(x) \end{aligned}$$

As a special case we get

$$\mathbb{P}(X \leq b) = \sum_{x \leq b} f_X(x) = F_X(b). \tag{1.7}$$

Consequently, we notice that each $f_x(x) \geq 0$ for all x (since they are probabilities), and that $\sum_x f_x(x) = 1$, as this is summing over the entire sample space of X.

1.5.2 Probability density functions

We would like to translate the very same idea of "point probabilities" over from the discrete case to the continuous case, but in doing so must exercise caution. Let X be a continuous random variable. i.e. X

is a random variable whose cdf is continuous. If such a probability function f_X exists for X, then the analogous procedure would be to consider

$$\mathbb{P}(X \leq x) = F_X(x) = \int_{-\infty}^{x} f_X(x)\,\mathrm{d}\tilde{x},$$

as this would be like summing over all possible values of $f_X(x)$ such that $X \leq x$ on a continuous scale[17], as per (1.7). In essence, the cdf F_X acts to "add up" all the "point probabilities" f_X within a required interval.

Definition 1.10 (Probability density function)**.** A continuous random variable X takes any numerical value within in an interval or collection of intervals (having an uncountable range). Its *probability density function (pdf)* is the function $f_X(x)$ that satisfies

$$F_X(x) = \int_{-\infty}^{x} f_X(\tilde{x})\,\mathrm{d}\tilde{x}, \text{ for all } x.$$

Geometrically speaking, the cdf computes the area under the pdf up to a point x, as shown in the diagram below:

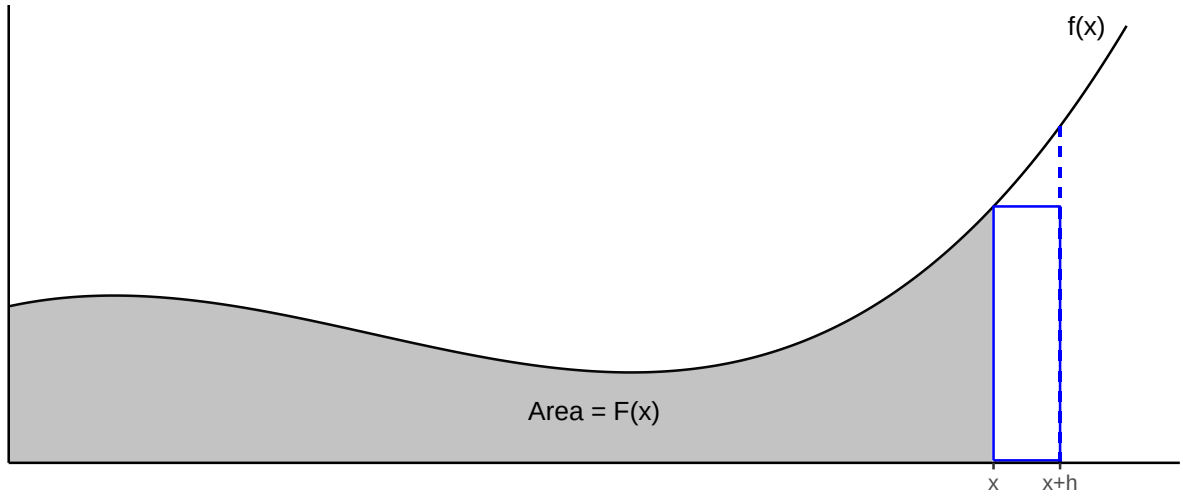


Figure 1.5: Illustration of the cdf as being the area under the pdf curve.

Consider the area under the curve up to the point $x + h$. This is given by $F(x + h)$, but may also be approximated as $F(x + h) \approx F(x) + A_{blue}$, where $A_{blue} = f(x) \cdot h$ is the area of the blue rectangle. This might be a poor approximation, and is only ever a good one when h is small. With a little rearranging, we get

$$f(x) \approx \frac{F(x + h) - F(x)}{h}$$

and argue that the RHS approaches the quantity $f(x)$ as h tends to zero, i.e.

$$f(x) = \lim_{h \to 0} \frac{F(x + h) - F(x)}{h},$$

which is the definition of the derivative. This is the idea of the Fundamental Theorem of Calculus, which tells us that

$$f_X(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_X(x).$$

Several observations regarding the probability density function:

- $f_X(x) \geq 0$ for all x. The curve of the pdf cannot dip below the x-axis.

[17] The Riemann integral is defined as the limit of the sum of the areas of bars dividing the area under the curve, as the number of bars gets larger and larger (and hence the width of the bars get smaller and smaller).

- $\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = 1$, which is essentially saying $\mathbb{P}(\Omega) = 1$.
- Point probabilities have no weight in the continuous case:

$$\mathbb{P}(X = x) = \int_x^x f(\tilde{x})\,\mathrm{d}\tilde{x} = 0.$$

  In effect, we can be less strict about the use of inequalities, since

  - $\mathbb{P}(X \leq b) = \mathbb{P}(X < b) + \cancel{\mathbb{P}(X = b)}$; and thus
  - $\mathbb{P}(a \leq X \leq b) = \mathbb{P}(a \leq X < b) = \mathbb{P}(a < X \leq b) = \mathbb{P}(a < X < b)$.

- To calculate probabilities within an interval, we can do the following:

  - $\mathbb{P}(a < X < b) = F(b) - F(a)$ (be careful, this is not true for discrete r.v.)
  - $\mathbb{P}(X > a) = 1 - F(a)$

It's wrong to think of pdfs $f(x)$ as probability functions–this only holds for discrete r.v.. Continuous pdfs do not give us probabilities unless we perform definite integrals on them. Read Wasserman (Warning after Example 2.13 on p.24) for more on this.

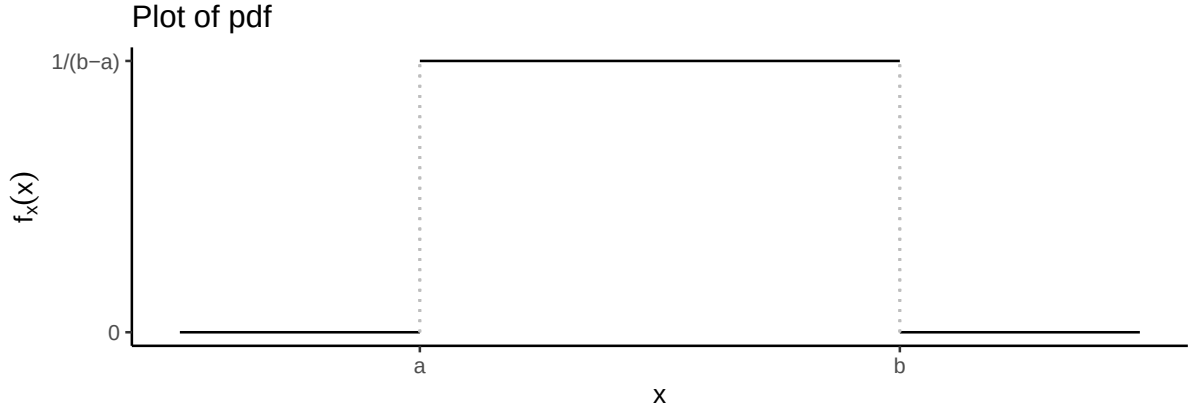
Example 1.16. Suppose that X is uniformly distributed on the interval $(a, b) \subset \mathbb{R}$. Its pdf is given by

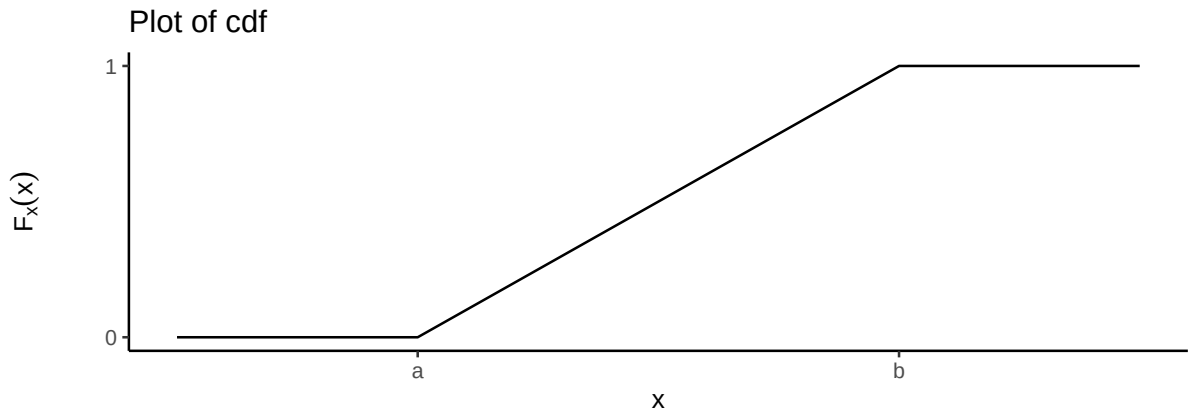
$$f_X(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{otherwise} \end{cases}$$

When $a < x < b$, the cdf is

$$\begin{aligned} F_X(x) = \int_{-\infty}^{x} f_X(\tilde{x})\,\mathrm{d}\tilde{x} &= \cancel{\int_{-\infty}^{a} f_X(\tilde{x})\,\mathrm{d}\tilde{x}}^{\,0} + \int_a^x \frac{1}{b-a}\,\mathrm{d}\tilde{x} \\ &= \left[\frac{\tilde{x}}{b-a}\right]_a^x = \frac{x-a}{b-a}, \end{aligned}$$

while $F_X(x) = 0$ for $x < a$, and $F_X(x) = 1$ for $x > b$.





Just some remarks on notation:

1. We write $X \sim F_X(x)$ to mean that "X has a distribution given by $F_X(x)$". The symbol '$\sim$' is read "is distributed as". Sometimes writing it with the pdf $X \sim f_X(x)$ is also clear in meaning. If we are dealing with a commonly used probability distribution, we would use their specially given name, e.g. $X \sim \text{Unif}(a, b)$. If X and Y are identically distributed, then we write $X \sim Y$.
2. Sometimes we just write $\int f(x)\,\mathrm{d}x$ to mean $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x$.

In case you were wondering, random variables with mixed distributions do exist, but we won't really encounter them in this course. Here's some food for thought. Let X be a discrete random variable (e.g. one that follows the two coin flip distribution), and let Y be a continuous random variable (e.g. a uniform distribution on the interval $[0, 1]$). Construct a new random variable Z by flipping a fair coin and define

$$Z = \begin{cases} X & \text{coin lands Heads} \\ Y & \text{coin lands Tails} \end{cases}$$

Since the probability of a coin toss is 50-50, in symbols we can write $\mathbb{P}(Z = X) = \mathbb{P}(Z = Y) = 0.5$. The question is, what sort of random variable is Z? If it is discrete, then how can it take on uncountably many different values when the coin lands tails? If it is continuous, then how come the point probability $\mathbb{P}(Z = X) = 1/2$ is non-zero?

The good news here is that categorising a random variable as discrete or continuous is purely arbitrary and convenient, but not required at all when measure theoretic foundations are used. There is no issue at all when dealing with such random variables, as the measure space will be well defined.

1.6 Transformations

In statistics, it is often the case that given a random variable X, we are also interested in transformations of this random variable. For example, if X is made to represent the gross domestic product (GDP) of a country, then we may be interested to study the logarithm of the GDP instead.

If $X \sim F_X(x)$ is a random variable, then a transformation of X by any function, $g(X)$ say, is also a random variable. Writing $Y = g(X)$, we may describe the probabilistic behaviour of Y in terms of that of X:

$$\mathbb{P}(Y \in A) = \mathbb{P}(g(X) \in A). \tag{1.8}$$

Sometimes we may write this expression explicitly, but is this always the case? It really depends on the transformation g.

Formally, g defines a mapping a from the original sample space of X (let's denote this $\mathcal{X}$) to a new sample space $\mathcal{Y}$ for Y. We can write $g : \mathcal{X} \to \mathcal{Y}$. For random variable X and its transformed version $Y = g(X)$, these sets are

$$\mathcal{X} = \{x \mid f_X(x) > 0\} \quad \text{and} \quad \mathcal{Y} = \{y \mid y = g(x) \text{ for } x \in \mathcal{X}\}.$$

The set $\mathcal{X}$ is called the *support* of the random variable X; the points at which the distribution is valid. We associate with g an *inverse mapping* denoted by $g^{-1} : \mathcal{Y} \to \mathcal{X}$, defined by

$$g^{-1}(A) = \{x \in \mathcal{X} \mid g(x) \in A\}.$$

Then the probability in (1.8) can be written as

$$\mathbb{P}(Y \in A) = \mathbb{P}(X \in g^{-1}(A)).$$

For discrete random variables, this is relatively straightforward, as seen in the following example.

Example 1.17. Let X be a random variable with a discrete uniform distribution on the set $\{-1, 0, 1\}$. That is,

$$f_X(x) = \begin{cases} 1/3 & x \in \{-1, 0, 1\} \\ 0 & \text{otherwise.} \end{cases}$$

Consider the transformation $Y = X^2$. Then the values of Y are

$$Y = \begin{cases} 0 & X = 0 \\ 1 & X = -1, 1 \end{cases}$$

and they take these values with probabilities

$$\mathbb{P}(Y = 0) = \mathbb{P}(X = 0) = 1/3,$$

and

$$\mathbb{P}(Y = 1) = \mathbb{P}(X = -1) + \mathbb{P}(X = 1) = 1/3 + 1/3 = 2/3.$$

As a remark, Y takes fewer values than X because the transformation is not one-to-one.

The continuous case is harder, and we need to consider the type of function and also keep track of the sample space. Suppose $x \mapsto g(x)$ represents a strictly monotonic transformation from $\mathcal{X}$ to $\mathcal{Y}$. This means that g is both *injective* (one-to-one) and *surjective* (onto); meaning that g is *bijective*. Importantly, g can be **inverted**.

We can work out the cdf of Y when g is increasing:

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) \\ &= \mathbb{P}(g(X) \leq y) \\ &= \mathbb{P}(X \leq g^{-1}(y)) \\ &= F_X(g^{-1}(y)) \end{aligned} \tag{1.9}$$

On the other hand, when g is decreasing:

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) \\ &= \mathbb{P}(g(X) \leq y) \\ &= \mathbb{P}(X > g^{-1}(y)) \quad \text{note the sign reversal} \\ &= 1 - F_X(g^{-1}(y)) \end{aligned} \tag{1.10}$$

The above is somewhat of an informal derivation of the cdf of the transformed variable Y. For more mathematical details, please see C&B Section 2.1. The main message here is that we need to keep track of the set $A_y = \{x \in \mathcal{X} \mid g(x) \leq y\}$, and it depends on whether or not g is increasing or decreasing. From here, the pdf of Y can be obtained by differentiating the cdf (in the continuous case). The following theorem formalises this approach.

Theorem 1.5 (Pdf of continuous transformations)**.** *Let X have pdf $f_X(x)$ and let $Y = g(X)$ be a strictly monotone function, i.e. g is strictly increasing or decreasing. Suppose also that $f_X(x)$ is continuous on the support of X, and that the inverse $g^{-1}(y)$ has a continuous derivative on the support of Y. Then the pdf of Y is given by*

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| & y \in \{y \mid y = g(x) \text{ s.t. } f_X(x) > 0\} \\ 0 & \text{otherwise} \end{cases}$$

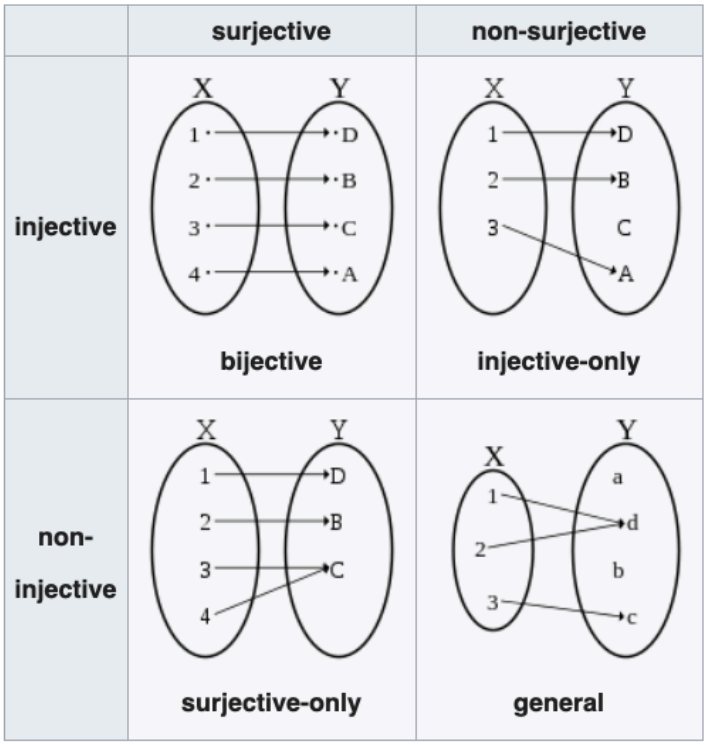


Figure 1.6: Injective vs surjective functions.

Note that the pdf of Y is valid everywhere the inverse transformation is valid on the pdf of X. For example, if X has support $\mathbb{R}$, then $Y = X^2$ has support only on $[0, \infty)$.

Proof. Differentiate and apply the chain rule to (1.9) and (1.10):

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} = \begin{cases} f_X(g^{-1}(y))\frac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y) & g \text{ increasing} \\ -f_X(g^{-1}(y))\frac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y) & g \text{ decreasing.} \end{cases}$$

□

Example 1.18. Let $X \sim \text{Unif}(-1, 1)$ (continuous). The pdf of X is given by

$$f_X(x) = \begin{cases} 1/2 & x \in (-1, 1) \\ 0 & \text{otherwise.} \end{cases}$$

Let $Y = X^2$. By symmetry, $|X| \sim \text{Unif}(0, 1)$; and $Y = |X|^2$ is a smooth, invertible function of $|X|$. Hence

$$\begin{aligned} f_Y(y) &= f_{|X|}(\sqrt{y}) \left| \frac{\mathrm{d}}{\mathrm{d}y}\sqrt{y} \right| \\ &= 1 \cdot \frac{1}{2\sqrt{y}} \end{aligned}$$

for $0 < y < 1$. Note however, $Y = X^2$ is not uniformly distributed anymore.

1.6.1 Probability integral transform

A special and very useful kind of transformation is the *probability integral transform (PIT)*. Suppose X is continuous, and let $Y = F_X(X)$. Here, Y is transformed using the cdf of X, and is considered a random variable still because it's a function of a random variable. Then, the distribution of Y is uniform on $(0, 1)$!

Theorem 1.6 (Probability integral transform (PIT))**.** *Let X have continuous cdf $F_X(x)$ and define the random variable Y as $Y = F_X(X)$. Then Y is uniformly distributed on $(0,1)$, that is, $Y \sim \text{Unif}(0,1)$ and*

$$f_Y(y) = \begin{cases} 1 & 0 < y < 1 \\ 0 & \textit{otherwise} \end{cases}$$

with $\mathbb{P}(Y \leq y) = y = F_Y(y)$ for $y \in (0,1)$.

Proof. For $Y = F_X(X)$ we have, for $0 < y < 1$,

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

At the endpoints we have $\mathbb{P}(Y \leq y) = 1$ for $y \geq 1$ and $\mathbb{P}(Y \leq y) = 0$ for $y \leq 0$, since there is zero probability outside the interval $(0,1)$. Thus Y has a uniform distribution.

Note that in the above proof, we used the fact that F_X is a monotone increasing function, and thus the equality

$$\mathbb{P}(F_X^{-1}(F_X(X)) \leq F_X^{-1}(y)) = \mathbb{P}(X \leq F_X^{-1}(y))$$

holds. There are two cases:

- Suppose F_X is strictly increasing. Then it is true that $F_X^{-1}(F_X(x)) = x$, since the inverse is uniquely defined.
- Suppose F_X is increasing but with "flat" parts. Then there are regions A of the cdf where the inverse is not uniquely defined. But for $x \in A$, the above equality still holds true because $\mathbb{P}(X \leq x) = \mathbb{P}(X \leq x^*) = y^*$, where $x^* = \inf\{x | x \in A, F_X(x) = y^*\}$. In essence, the flat cdf denotes a region of 0 probability. For instance, suppose the region $A = [x_1, x_2]$. Then $\mathbb{P}(X \leq x) = \mathbb{P}(X \leq x_1)$ for any $x \in A$.

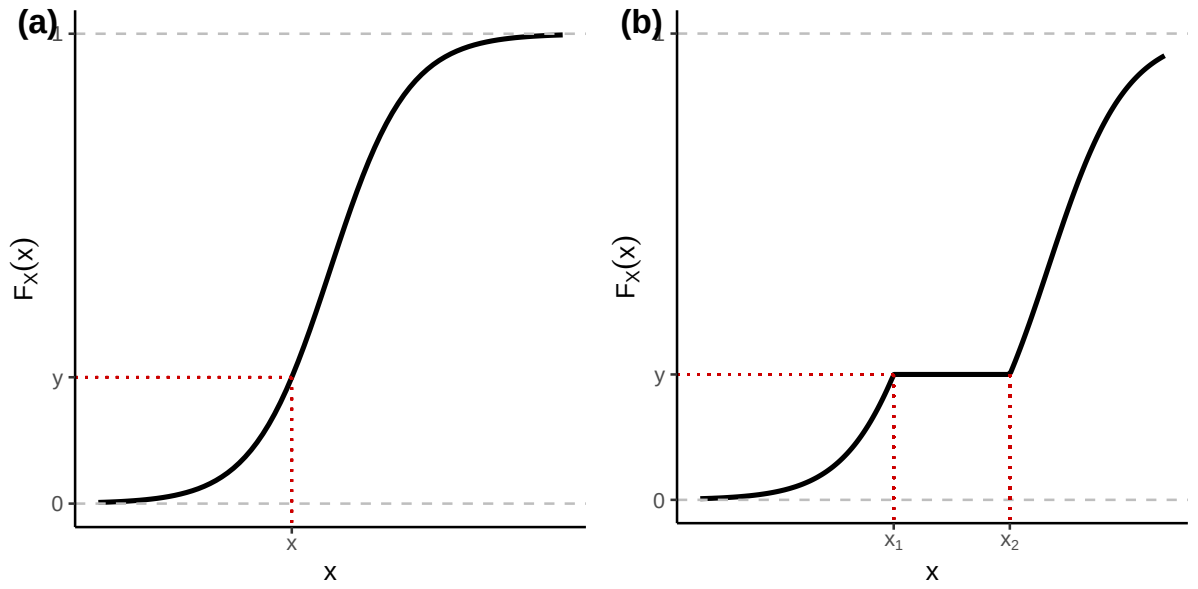


Figure 1.7: (a) a strictly increasing cdf has a unique inverse; while (b) a non-decreasing cdf has regions in which there is zero probability, so the cdf inverse is the infimum of the x in that range.

□

The PIT is useful for various statistical purposes, both theoretical and practical. A particular application of note is the *simulation* of an arbitrary random variable X on a computer[18].

Example 1.19. Let X be a random variable with an exponential distribution with unit mean. Its cdf is $F_X(x) = 1 - e^{-x}$ for $x > 0$. By the PIT, we have that $U = 1 - e^{-X}$ is uniformly distributed on $(0, 1)$.

Working a bit backwards, with $U \sim \text{Unif}(0, 1)$ suppose there is (strictly) monotone transformation $T : [0, 1] \to \mathbb{R}$ such that $T(U) := X$. We notice that

$$\begin{aligned} F_X(x) &= \mathbb{P}(X \leq x) \\ &= \mathbb{P}(T(U) \leq x) \\ &= \mathbb{P}(U \leq T^{-1}(x)) \\ &= T^{-1}(x) \end{aligned}$$

where the last step follows since U is uniformly distributed on $(0, 1)$. Therefore, F_X is the inverse function of T, or equivalently $T(u) = F_X^{-1}(u)$ for $u \in [0, 1]$. It is then possible to generate $X \sim \text{Exp}(1)$ using the algorithm below:

1. Generate $U = \{U_1, \ldots, U_n\} \sim \text{Unif}(0, 1)$.
2. Transform the samples $U \mapsto X$ using the function $T(u) = F_X^{-1}(u) = -\log(1 - u)$.
3. Then $X = \{X_1, \ldots, X_n\}$ is a sample from $\text{Exp}(1)$.

The `R` code below shows how to implement this in practice. As we can see, the PIT method and the 'direct' method using `R`'s built in function `rexp()` generates very similar results.

```
set.seed(2911)
n <- 1000

# Generate Unif(0,1) r.v.
U <- runif(n, min = 0, max = 1) %>% sort()
head(U)
```

```
## [1] 0.000795529 0.003317551 0.005630351 0.005866332 0.006031142 0.008309819
```

```
# Generate Exp(1) r.v. using PIT
X <- -log(1-U)
head(X)
```

```
## [1] 0.0007958456 0.0033230660 0.0056462615 0.0058836065 0.0060494031
## [6] 0.0083445376
```

```
# Generate Exp(1) r.v. using R built in function
Z <- rexp(n, rate = 1) %>% sort()
head(Z)
```

```
## [1] 0.0009589139 0.0020182137 0.0026925327 0.0042910245 0.0043342430
## [6] 0.0045965933
```

1.7 Multiple random variables

In the real world, data collection often involves more than one variable, so methods to analyse these kinds of data do exist. In particular, probability models may well be extended to involve more than one random variable. These are known as *multivariate models*.

[18] Also called *inverse transform sampling*

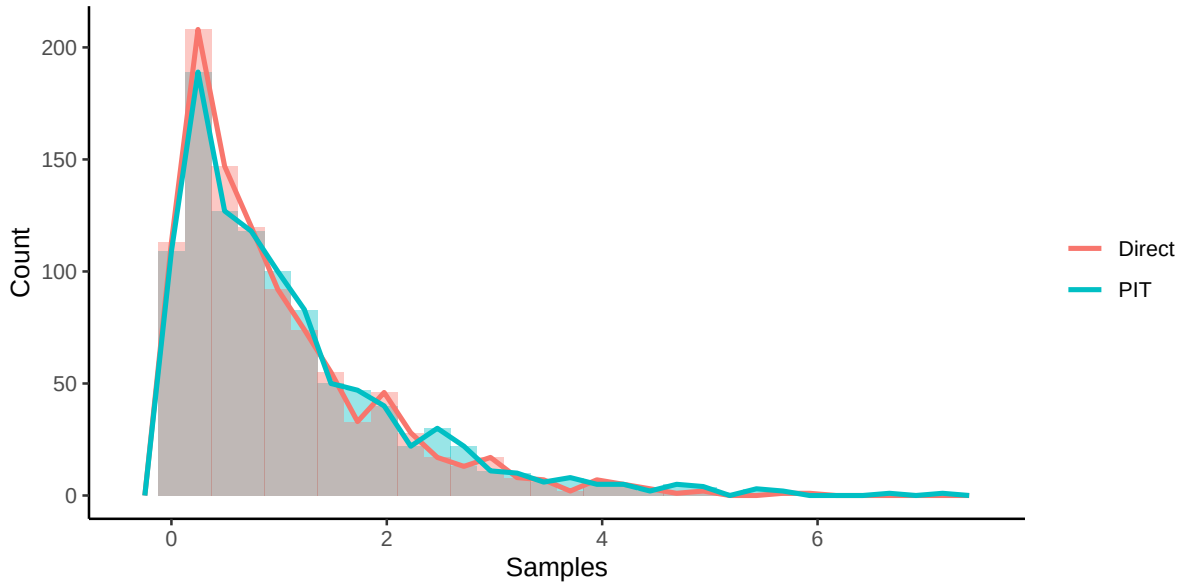


Figure 1.8: Comparison of histogram and frequency polygon of the samples generated using PIT and the 'direct' method in R.

1.7.1 Bivariate distributions

Consider the simplest kind, where we deal with only two random variables in each the discrete and continuous case.

Definition 1.11 (Joint mass function)**.** Given a pair of discrete r.v. X and Y, the joint mass function or joint pmf is defined by

$$f_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y).$$

Definition 1.12 (Joint density function)**.** A function $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$ is called a joint probability density function (pdf) of the continuous random vector (X, Y) if for any set $A \subseteq \mathbb{R}^2$,

$$\mathbb{P}((X,Y) \in A) = \iint_A f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y.$$

To be clear, bivariate random variables occur in **pairs**, so that (X, Y) is treated as one entity. Luckily, all the univariate properties carry over to the bivariate (and even multivariate) case, such as:

- $f_{X,Y}(x,y) \geq 0$ for all $(x,y) \in \mathbb{R}^2$
- $\sum_x \sum_y f(x,y) = 1$ if discrete, $\iint f(x,y)\,\mathrm{d}x\,\mathrm{d}y = 1$ if continuous
- The joint cdf is defined as

$$\begin{aligned} F_{X,Y}(x,y) &= \mathbb{P}(X \leq x, Y \leq y) \\ &= \begin{cases} \sum_{u \leq x} \sum_{v \leq x} f_{X,Y}(u,v) & \text{discrete case} \\ \int_{u \leq x} \int_{v \leq x} f_{X,Y}(u,v)\,\mathrm{d}u\,\mathrm{d}v & \text{continuous case} \end{cases} \end{aligned}$$

Example 1.20. A bivariate distribution for two discrete random variable X and Y each taking values 0 or 1 can be summarised in the 2×2 table below.

	$Y = 0$	$Y = 1$
$X = 0$	$1/9$	$2/9$
$X = 1$	$2/9$	$4/9$

For instance, $\mathbb{P}(X = 1, Y = 1) = f(1,1) = 4/9$.

A different way of expressing the above table is by explicitly listing out the probabilities, as follows:

$$f(x,y) = \begin{cases} 1/9 & x=0, y=0 \\ 2/9 & x=0, y=1 \\ 2/9 & x=1, y=0 \\ 4/9 & x=1, y=1 \end{cases}$$

Example 1.21. Consider a uniform distribution on the unit square $[0,1] \times [0,1]$. It has pdf given by

$$f(x,y) = \begin{cases} 1 & 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

This is a well-defined pdf, as $f \ge 0$ and $\int\int f(x,y)\,\mathrm{d}x\,\mathrm{d}y = 1$. Suppose we want to find $\mathbb{P}(X < 1/2, Y < 1/2)$ and $\mathbb{P}(X+Y<1)$.

For the first probability, we integrate in the set $\{(x,y) \mid 0 < x < 1/2, 0 < y < 1/2\}$:

$$\begin{aligned} \mathbb{P}(X<1/2, Y<1/2) &= \int_0^{1/2}\int_0^{1/2} \mathrm{d}x\,\mathrm{d}y \\ &= \Big[[xy]_0^{1/2}\Big]_0^{1/2} = 1/4. \end{aligned}$$

For the second probability, note that the set $\{(x,y) \mid x+y<1\}$ corresponds to $\{(x,y) \mid 0<y<1, 0<x<1-y\}$. So

$$\begin{aligned} \mathbb{P}(X+Y<1) &= \int_0^1 \mathrm{d}y \int_0^{1-y} \mathrm{d}x \\ &= \int_0^1 \mathrm{d}y [x]_0^{1-y} \\ &= \int_0^1 (1-y)\,\mathrm{d}y = [y - y^2/2]_0^1 = 1/2. \end{aligned}$$

Another way of understanding these probabilities is by thinking about them geometrically. If we plot the pdf surface, it would look something like the following sketch:

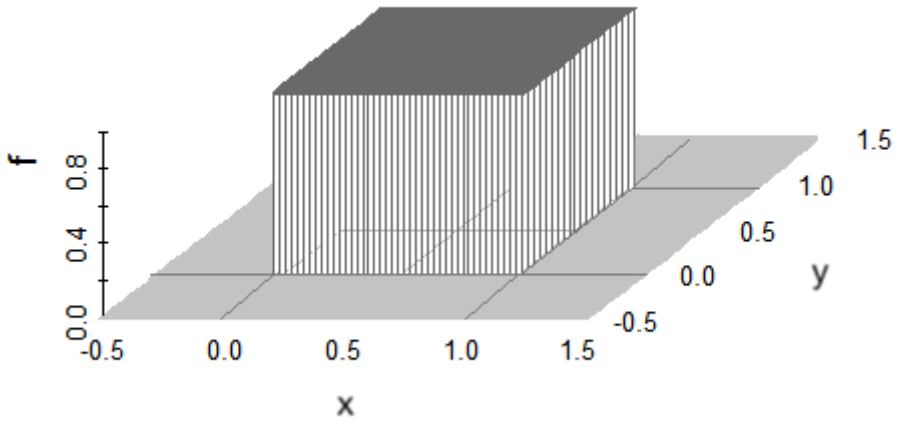


Figure 1.9: Pdf surface plot of the uniform distribution on the unit square.

Any probability of interest would be calculated by finding the volume of interest. For instance, consider again the probability $\mathbb{P}(X+Y<1)$. Viewing the surface from above effectively concentrates on the two

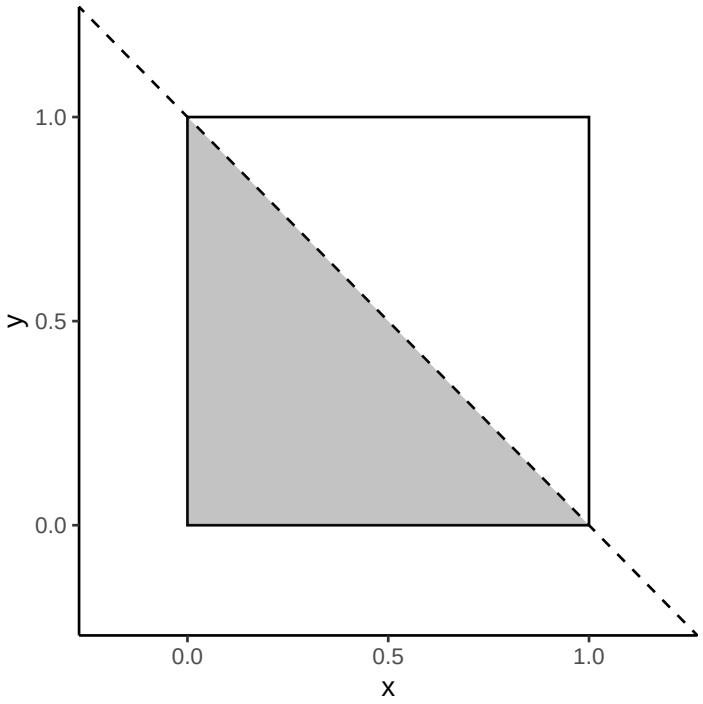


Figure 1.10: View of the pdf surface from above, focussing on the X and Y axis.

dimensions of x and y. It's straightforward to realise that the region of interest is anything occuring below the line $y = x$.

Correspondingly, we ask what is the volume of this wedge? It is the area of the shaded region (half of the unit square) multiplied by the height of the surface (1), so we get the same answer of 1/2.

1.7.2 Marginal distributions

We may think of multivariate distributions as several random variables "stitched" together, whose distribution as a whole is dependent on each of the components. Having said this, it is possible recover the distribution for one of the components in a bivariate (or multivariate) model by summing or integrating over the remaining probability distribution, depending on whether or not the other components are discrete or continuous.

Definition 1.13 (Marginal distribution)**.** For a bivariate random variable (X, Y), the marginal distributions of X and Y may be obtained respectively as

$$f_X(x) = \begin{cases} \sum_y f_{X,Y}(x,y) & \text{if } Y \text{ is discrete} \\ \int_y f_{X,Y}(x,y)\,\mathrm{d}y & \text{if } Y \text{ is continuous} \end{cases}$$

$$f_Y(y) = \begin{cases} \sum_x f_{X,Y}(x,y) & \text{if } X \text{ is discrete} \\ \int_x f_{X,Y}(x,y)\,\mathrm{d}x & \text{if } X \text{ is continuous} \end{cases}$$

A note to say that since the joint cdf is defined to be

$$F(x,y) = \mathbb{P}(X \leq x, Y \leq y),$$

the *marginal cdfs* can be obtained from the joint cdf for X by summing over all the components of Y in

the joint cdf, i.e.

$$\begin{aligned} F_X(x) &= \sum_{k \leq x} \left(\sum_y f_{X,Y}(k,y) \right) \\ &= \mathbb{P}(X \leq x, Y \leq \infty) \\ &= F_{X,Y}(x, \infty) \end{aligned}$$

Similarly, we sum up over the components of X to obtain the marginal cdf for Y, $F_Y(y) = F_{X,Y}(\infty, y)$. Note that for continuous random variables, we integrate instead: $F_X(x) = \int_{-\infty}^{x} \left(\int f_{X,Y}(\tilde{x}, y) \, \mathrm{d}y \right) \mathrm{d}\tilde{x}$.

Example 1.22. Define a joint pdf by

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

Let's compute the marginal distributions of X and Y. We'll have to use integration here since the geometry of the pdf surface in 3-dimensions is a bit complex to work with.

Firstly, we need to find the *normalising constant* c, such that the joint integral of the pdf is 1. That is, we need to find the value of c satisfying

$$\int_{x=0}^{1} \int_{y=0}^{1} cxy^2 \, \mathrm{d}x \, \mathrm{d}y = c \left[\left[\frac{x^2 y^3}{6} \right]_0^1 \right]_0^1 = 1.$$

We work out that $c = 6$.

To work out the marginal distribution of X, we integrate the pdf over all possible values of Y, i.e.

$$\begin{aligned} f_X(x) &= \int_{y=0}^{1} 6xy^2 \, \mathrm{d}y \\ &= \left[\frac{6xy^3}{3} \right]_0^1 \\ &= 2x. \end{aligned}$$

Note that $f_X(x) = 2x$ for $0 < x < 1$ is indeed a valid pdf (it integrates to 1, and also satisfies all the properties of a pdf). We can now use this to calculate probabilities involving only X, for instance

$$\mathbb{P}\left(\frac{1}{2} < X < \frac{3}{4} \right) = \int_{1/2}^{3/4} 2x \, \mathrm{d}x = \frac{5}{16}.$$

Example 1.23. Here's a trickier example. Consider the joint pdf defined by

$$f(x,y) = \begin{cases} e^{-y} & 0 < x < y < \infty \\ 0 & \text{otherwise.} \end{cases}$$

At first glance, it does not seem that the pdf depends on x at all. But actually, it does. If we look at the values at which this pdf is non-zero, it is conditional on the positive values of x such that it is lesser than y. To put it more precisely, we could write the pdf as

$$f(x,y) = \mathbb{1}_{[\{(u,v)|0<u<v<\infty\}]}(x,y) e^{-y}$$

so we can clearly see the dependence of the pdf on both x and y. Here, we have used the *indicator function* $\mathbb{1}_A(x)$ defined as

$$\mathbb{1}_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A. \end{cases}$$

To calculate the joint cdf of this bivariate distribution, we compute the integral

$$F_{X,Y}(x,y) = \int_{u=0}^{x} \int_{v=u}^{y} e^{-v} \, \mathrm{d}u \, \mathrm{d}v$$

The limits of integration are obtained as follows: For the random variable X (using the integrating variable u) we start at the smallest value it can take ($u = 0$) and proceed upwards to some arbitrary point $u = x$. For Y (using the v as the variable of integration), the smallest possible value it can take is $v = u$, since $x < y < \infty$ and the integration depends on what happens to X. From here, proceed upwards to some arbitrary point $v = y$. Working through this integral gives us

$$F_{X,Y}(x, y) = 1 - (e^{-x} + xe^{-y}), \qquad 0 < x < y < \infty.$$

This is a valid cdf, since

- $\lim_{x,y\to 0} F(x, y) = 1 - \left(\lim_{x,y\to 0} e^{-x} + \lim_{x,y\to 0} xe^{-y}\right) = 1 - (1 + 0) = 0.$
- $\lim_{x,y\to\infty} F(x, y) = 1 - \left(\lim_{x,y\to\infty} e^{-x} + \lim_{x,y\to\infty} xe^{-y}\right) = 1 - (0 + 0) = 1.$
- $F(x, y) > 0$ in that range.

From here, the marginal cdf of X is obtained as $F_X(x) = F_{X,Y}(x, +\infty) = 1 - e^{-x}$ for $0 < x < \infty$. Noting that the maximum value x can take is y, we can similarly obtain the marginal cdf of Y as $F_Y(y) = F_{X,Y}(+\infty, y) = F(+\infty, y) = F(y, y) = 1 - (e^{-y} + ye^{-y})$ for $0 < y < \infty$. It's easily checked that both of these functions are indeed cdfs.

In the continuous case, the joint cdf of (X, Y) is related to the joint pdf by the relationship

$$F_{X,Y}(x, y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(u, v)\,\mathrm{d}u\,\mathrm{d}v.$$

By the (bivariate) Fundamental Theorem of Calculus, this implies that

$$\frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y) = f_{X,Y}(x, y).$$

1.7.3 Conditional distributions

Oftentimes when two r.v. (X, Y) are observed, the values of the two variables are "related". What we mean by this is that knowledge about the value of Y gives us some information about the value of X and vice versa. Some examples:

- Height (X) and weight (Y) of a person;
- A level points score (X) and socio-economic status (Y);
- Heart rate (X) and oxygen saturation levels (Y).

To make this idea a little more concrete, think about what values the heart rate (X) of an individual can take. For a healthy individual, this might be anywhere between 40 bpm to 200 bpm (depending on their age, what activity they are doing, and so on). On average, X is 72 bpm. Oxygen saturation levels on the other hand are usually between 95 and 100 percent, but drops below this range when an intense activity is performed. Consequently, if we were to guess what the heart rate X value would be given $Y < 0.95$, it would make more sense to guess that $X = 160$ rather than $X = 72$. This concept should sound familiar!

Define the conditional distributions for discrete and continuous random variables as follows.

Definition 1.14 (Conditional distributions, discrete)**.** If X and Y are discrete, the *conditional pmf* of X given $Y = y$ is

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

Definition 1.15 (Conditional distributions, continuous)**.** If X and Y are continuous, the *conditional pdf* of X given $Y = y$ is

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

In the discrete case, the conditional distribution is derived in a similar way to how the conditional probabilities were (see Definition 1.4). Interestingly in the continuous case, the definition still looks familiar but it should be noted that plugging x and y values into the definition will not yield probabilities– one still requires integration over a set:

$$\mathbb{P}(X \in A|Y = y) = \int_A f_{X|Y}(x|y)\,\mathrm{d}x.$$

Note that as a function of x, $f_{X|Y}(x|y)$ is indeed a pdf, since in the discrete case

$$\sum_x f_{X|Y}(x|y) = \sum_x \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{\mathbb{P}(Y = y)}{\mathbb{P}(Y = y)} = 1,$$

and in the continuous case,

$$\int f_{X|Y}(x|y)\,\mathrm{d}x = \int \frac{f_{X,Y}(x, y)}{f_Y(y)}\,\mathrm{d}x = \frac{f_Y(y)}{f_Y(y)} = 1.$$

Finally, it's convenient to note that, just like as we saw for conditional probabilities, we can rearrange the equations to yield

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

Example 1.24. Let X and Y have the joint pdf $f(x, y) = x + y$ for $0 \leq x, y \leq 1$. Suppose $Y = a$ has been observed, where $a \in [0, 1]$. Firstly, the the pdf of Y is

$$f_Y(y) = \int_0^1 (x + y)\,\mathrm{d}x = \left[xy + y^2/2\right]_0^1 = y + 1/2.$$

The conditional pdf for X is

$$f_{X|Y}(x|Y = a) = \frac{f_{X,Y}(x, Y = a)}{f_Y(a)} = \frac{x + a}{a + 1/2}.$$

We can compute $\mathbb{P}(X < 1/4|Y = 1/3)$ by

$$\mathbb{P}(X < 1/4|Y = 1/3) = (1/3 + 1/2)^{-1} \int_0^{1/4} (x + 1/3)\,\mathrm{d}x = 11/80.$$

It is possible to also describe *conditional cdfs*. If we treat the conditional pmf/pdf $f(x|y)$ as a new pmf/pdf $g(x)$, then the cdf can be easily obtained in the usual way: $F(x|y) = G(x) = \mathbb{P}(X \in A|y)$.

1.7.4 Independent random variables

Previously we came across the concept of independence of probabilistic events. We can extend this notion to random variables using the conditional pmf/pdf definitions.

Definition 1.16 (Independece of r.v.)**.** Two random variables X and Y are independent if and only if for every $x \in \mathbb{R}$ and $y \in \mathbb{R}$,

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

We write $X \perp Y$.

Apparently, if there exists functions $g(x)$ and $h(y)$ (not necessarily pdfs) such that $f(x, y) = g(x)h(y)$ for all x, y, then X and Y are independent. This is proven in Lemma 4.2.7 of C&B. Hence, verifying whether two random variables are independent is made easier, since we only need to separate out the components of x and y in the joint pdf without needing to check whether or not the components themselves are pdfs.

The assumption of independence is used very often in statistical inference as it simplifies calculations quite a lot. We'll circle back to this thought when we talk about likelihood estimation.

Example 1.25. Recall the bivariate distribution on the unit square (c.f. Example 1.21). Note that the pdf of X is $f_X(x) = \int_0^1 \mathrm{d}y = 1$, and similarly $f_Y(y) = 1$. It is easy to see that X and Y are independent, since

$$f_{X,Y}(x, y) = 1 = f_X(x) f_Y(y).$$

As a consequence, to generate a random sample from (X, Y), one can randomly sample values $X \sim \text{Unif}(0, 1)$, and independently sample $Y \sim \text{Unif}(0, 1)$.

1.8 Expectations

The expected value, or expectation, of a random variable X is its average value *weighted* according to the probability distribution. Simply put, it signifies the *arithmetic mean* of a large number of independent realisations of X.

Definition 1.17 (Expectation)**.** The *expected value* or *mean* of a random variable X, denoted $\mathrm{E}(X)$, is defined to be

$$\mathrm{E}(X) = \begin{cases} \sum_x x f_X(x) = \sum_x x \mathbb{P}(X = x) & \text{if } X \text{ is discrete} \\ \int x f_X(x)\, \mathrm{d}x & \text{if } X \text{ is continuous} \end{cases}$$

provided that the integral or sum exists (is finite).

The symbol 'μ' is often used to denote the expected value. It may be represented by $\mathrm{E}\, X$, $\mathrm{E}[X]$ or even using $\mathbb{E}$ instead of E.

The expectation of a random variable is **not to be confused** with the *sample mean* of a set of observations $\{x_1, \ldots, x_n\}$, i.e. $\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$. The expectation is a purely theoretical value based on probabilities and pdfs. The sample mean incorporates *randomness* into the calculations, by virtue of the randomness of the observed set of sample values.

Example 1.26. Let $X \in \{0, 1\}$ take value 1 with probability p, and 0 with probability $1-p$. X is called a Bernoulli random variable, and we write $X \sim \text{Bern}(p)$. Then,

$$\mathrm{E}(X) = \sum_x x \mathbb{P}(X = x) = 1 \cdot p + 0 \cdot (1 - p) = p.$$

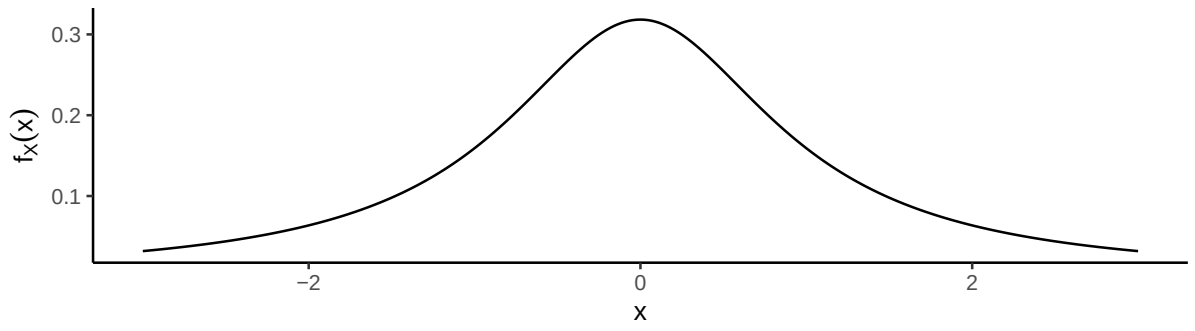
Example 1.27. Let X be a continuous random variable with pdf $f(x) = \frac{1}{b-a}$, where $a, b \in \mathbb{R}$ and $a < b$. X has what is called a uniform distribution on the interval (a, b), and we write $X \sim \text{Unif}(a, b)$. The mean of X is

$$\mathrm{E}(X) = \int_a^b \frac{x}{b-a} = \frac{a+b}{2},$$

the midpoint of the interval (a, b)! This reveals some intuition regarding uniformity of the distribution.

Do all random variables have expectations?

Example 1.28.



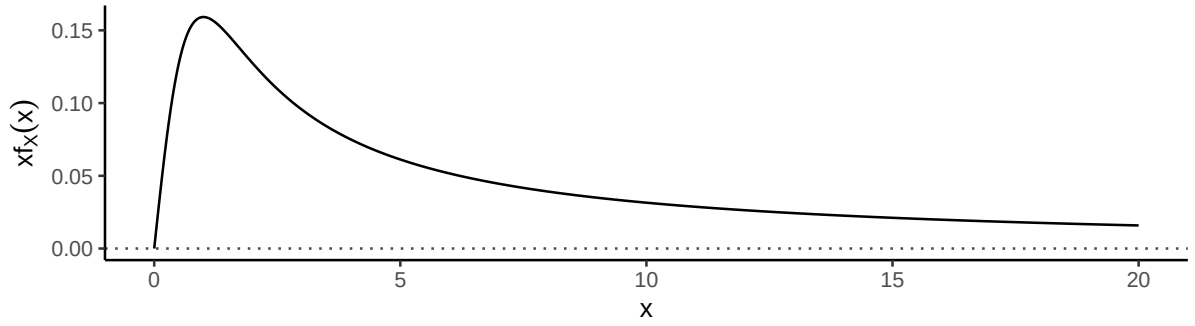
Let X be a continuous random variable with pdf $f(x) = \{\pi(1+x^2)\}^{-1}$ with support over $\mathbb{R}$. This is the Cauchy distribution[19] with location and scale parameter 0 and 1 respectively. Let's calculate the mean.

Using the substitution $u = x^2 + 1$ and $\mathrm{d}u/2 = x\,\mathrm{d}x$, we find that

$$\begin{aligned}
\mathrm{E}(X) &= \int_{-\infty}^{\infty} \frac{x\,\mathrm{d}x}{\pi(1+x^2)} \\
&= \int_{-\infty}^{0} \frac{x\,\mathrm{d}x}{\pi(1+x^2)} + \int_{0}^{\infty} \frac{x\,\mathrm{d}x}{\pi(1+x^2)} \\
&= \frac{1}{2\pi}\int_{u=\infty}^{u=1} \frac{\mathrm{d}u}{u} + \frac{1}{2\pi}\int_{u=1}^{u=\infty} \frac{\mathrm{d}u}{u} \\
&= \frac{1}{2\pi}\left[\log u\right]_{\infty}^{1} + \frac{1}{2\pi}\left[\log u\right]_{1}^{\infty} \\
&= \frac{1}{2\pi}(\infty - \infty) = \ ???
\end{aligned}$$

The mean of the Cauchy distribution is undefined. This seems a bit weird, since we can see that the pdf is somewhat bell-shaped with its peak at 0, so wouldn't we expect the mean to be zero? Not quite. The highest peak of the bell curve is known as the *mode* of the distribution, and that indeed is well defined and is zero. The *median* is also well-defined, as this is the point at which half the distribution lies below, and half lies above it–the median is zero. The median exists because the area under the pdf curve must necessarily be equal to 1, a finite value.

On the other hand, if we look at the plot of $xf(x)$ on the positive side of the real line, we see that the tail end does not drop fast enough for the area under the curve to be a finite number.



1.8.1 Expectations of functions of r.v.

Realise that if X is a r.v., then any function of X, $g(X)$, is also a random variable[20]. Often time we will want to know the mean of $g(X)$.

Theorem 1.7. *Let X be a r.v. with pdf $f_X(x)$, and let $Y = g(X)$. Then*

$$\mathrm{E}(Y) = \int g(x) f_X(x)\,\mathrm{d}x.$$

In particular, the kth **moment** of X for $k \in \mathbb{Z}$ is defined to be

$$\mathrm{E}(X^k) = \int x^k f_X(x)\,\mathrm{d}x.$$

The kth central moment is defined as $\mathrm{E}((X-\mu)^k)$, where $\mu := \mathrm{E}(X)$.

[19] Named after the French mathematician Augustin Cauchy, although in physics, it is often known by the Lorentz distribution after the Dutch Nobel Laureate Hendrik Lorentz.

[20] We can even describe the distribution for any transformation of X, see C&B Sec 2.1.

1.8.2 Properties of expectations

Let X be a r.v., and $a, b, c \in \mathbb{R}$ be constants. Here are some important properties of expectations.

- $\mathrm{E}(aX + bX + c) = a\,\mathrm{E}(X) + b\,\mathrm{E}(X) + c$ (linearity of expectations)
- If Y is a r.v. s.t. $X \perp Y$, then $\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y)$
- If $X \geq 0$ for all x, then $\mathrm{E}(X) \geq 0$
- If $a \leq X \leq b$ for all x, then $a \leq \mathrm{E}(X) \leq b$
- $\mathrm{E}(X) = \min_b \mathrm{E}((X-b)^2)$[21]

Note that the above properties work for any transformations of X too. For instance, $\mathrm{E}(ag(X) + bg(X) + c) = a\,\mathrm{E}(g(X)) + b\,\mathrm{E}(g(X)) + c$. Just think of $Y = g(X)$ as a new random variable.

If X and Y are not independent, then $\mathrm{E}(XY) \neq \mathrm{E}(X)\,\mathrm{E}(Y)$.

The last property above implies that the mean of a random variable is the minimiser of the (expected) quadratic loss function: The closer b is to X, then the smaller the value of $(X-b)^2$. We can also ask what is the expected value of b which minimises this quantity, and the answer is the mean of X. The interpretation here is that $\mathrm{E}(X)$ is a good guess of the value of X! At least, as measured by this quadratic loss function.

As a corollary to the linearity property, if $X_1, \ldots, X_n$ are r.v. and $a_1, \ldots, a_n$ are constants, then

$$\mathrm{E}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i\,\mathrm{E}(X_i).$$

Additionally, if $X_1, \ldots, X_n$ are **independent**,

$$\mathrm{E}\left(\prod_{i=1}^{n} a_i X_i\right) = \prod_{i=1}^{n} \mathrm{E}(X_i).$$

These properties are used extensively throughout statistics, so please take the time to memorise and learn these. It will make tackling questions in the later chapter much easier!

1.8.3 Variance

Aside from the mean of a random variable (its first central moment), another important concept is the second central moment, more commonly known as the variance.

Definition 1.18 (Variance)**.** Let X be a r.v. with mean μ. The *variance* of X is defined

$$\mathrm{Var}(X) = \mathrm{E}\left[(X-\mu)^2\right],$$

assuming this expectation exists. The *standard deviation* is $\mathrm{sd}(X) = \sqrt{\mathrm{Var}(X)}$.

The symbol σ^2 is often used to denote the variance, and σ the standard deviation. An alternative (and arguably easier) formula for the variance is

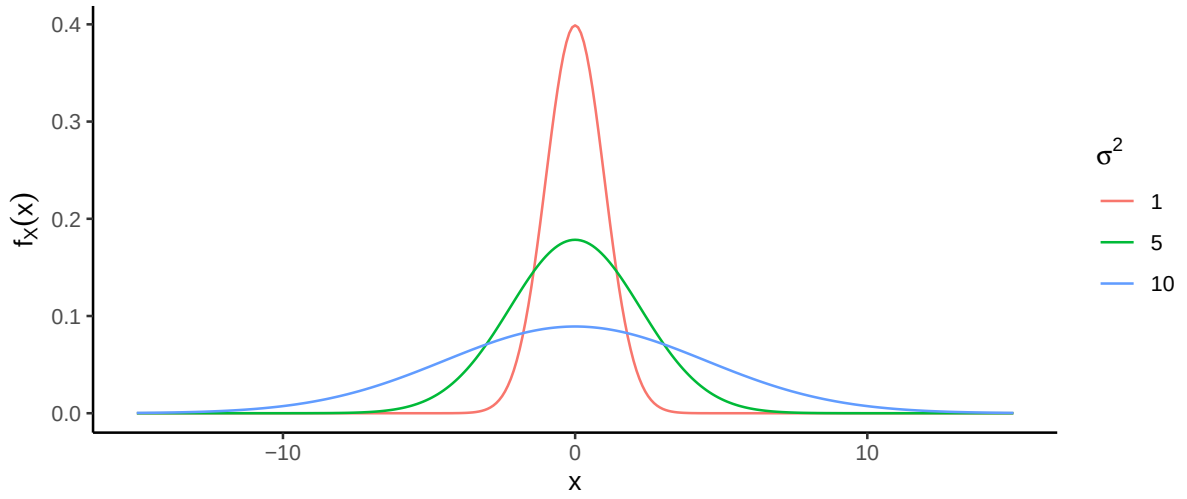
$$\sigma^2 = \mathrm{E}(X^2) - \{\mathrm{E}(X)\}^2$$

. But take caution, that this formula can be less precise than the one given in the definition above. Especially when X takes very large (or very small) values, then correspondingly $\mathrm{E}(X^2)$ and $\mathrm{E}(X)$ will be very large (very small) too. Computationally, there is a limited number of integers the computer can store (its single-precision floating point), and once this is exceeded the numbers get less precise. On the other hand, the formula in the definition avoids this issue because the difference between X and its mean is "tamed" in a way–the expected value is zero!

[21] See Example 2.2.6 C&B.

This variance is **not to be confused** with the *sample variance() of a set of observations $\{x_1, \ldots, x_n\}$, i.e. $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$. However, do inspect the two formulae for similarities!

The variance measures the spread of a distribution, that is, how far apart or close together the "mass" of a distribution are. To illustrate this, have a look at the following $\mathrm{N}(0, \sigma^2)$ pdfs for different values of σ^2.



Larger values of σ indicate a wider spread (away from the mean), and conversely, smaller values indicate a smaller spread, where values are close together near the mean. The variance will become an important criterion when determining how good or bad an parameter estimator is. More on this in Chapter 4!

1.8.4 Covariance and correlation

Suppose we had two random variables, and we wanted to see how each random variable behaves in their respective domains, but also together. That is, are larger values of X associated with larger or smaller values of X, or is there no relationship at all? We introduce the concept of covariance and correlation, both of which measure how strong the *linear relationship* is between X and Y.

Definition 1.19 (Covariance)**.** For two random variables X and Y with finite means μ_X and μ_Y respectively, the covariance between X and Y is

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

assuming this expectation exists. If the variances of the two random variables are finite, then the covariance between them exists.

As with the variance formula, there is also an alternative formula for the covariance given by

$$\mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y).$$

This formula is also susceptible to catastrophic cancellation, and should be avoided in computer programs[22]. Note that the covariance of X with itself is σ^2, which can be seen by plugging in X for Y in the formula above.

The magnitude of the covariance by itself does not reflect how strong the relationship between X and Y is, so this is where correlation comes in.

[22] See this article: https://en.wikipedia.org/wiki/Covariance#Numerical_computation

Definition 1.20 (Correlation)**.** For two r.v. X and Y with finite means μ_X and μ_Y resp., and variances σ_X^2 and σ_Y^2 resp., the correlation between X and Y is the number $\rho \in [-1, 1]$ defined by

$$\rho_{XY} := \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

The value $\rho_{XY} = 0$ implies that there is no linear relationship at all between X and Y. On the other hand, $\rho_{XY} = 1$ ($\rho_{XY} = -1$) implies a perfect positive (negative) linear relationship. In fact, $|\rho_{XY} = 1|$ if and only if $\exists a \neq 0, b \in \mathbb{R}$ such that $\mathbb{P}(Y = aX + b) = 1$. If $a > 0$ then $\rho_{XY} = 1$, and if $a < 0$ then $\rho_{XY} = -1$.

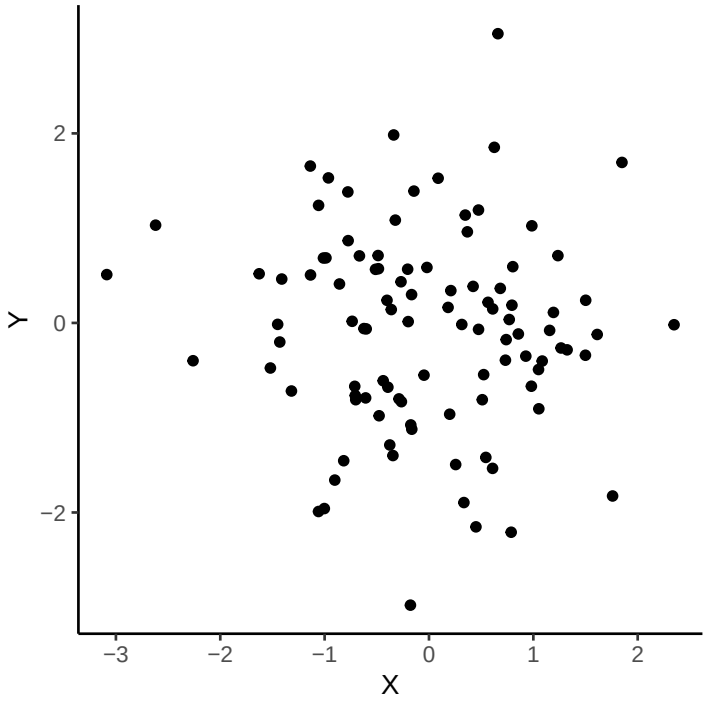
The relationship between covariance/correlation with independence is the following: If X and Y are independent, then $\text{Cov}(X, Y) = \rho_{XY} = 0$. This can easily be proven using properties of expectations.

The converse however is not true! If $\text{Cov}(X, Y) = \rho_{XY} = 0$, then X and Y are **not necessarily** independent. The reverse statement is true for normal random variables, but not in general. This is one of the properties of normal distributions which we'll cover in the next chapter.

What does *linear* relationship mean? As the name implies, a strong covariance/correlation allows us to draw a straight line through the points plotted in a 2-dimensional scatterplot. Let's take a look.

Let $X, Y \overset{\text{iid}}{\sim} \text{N}(0, 1)$. We can draw some random values in `R`, and produce a scatterplot to see the relationship between them.

```
set.seed(789)
X <- rnorm(n = 100, mean = 0, sd = 1)
Y <- rnorm(n = 100, mean = 0, sd = 1)
qplot(X, Y, geom = "point")
```



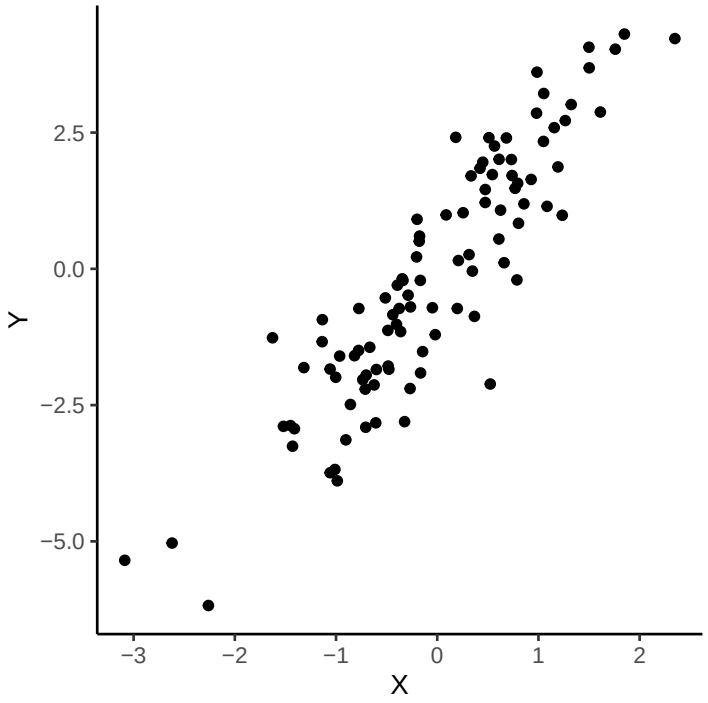
Since the objects `X` and `Y` are drawn separately, they are independent of each other. If we were to draw a straight line through the data points above, how would we draw such a line? There does not seem to be a "right" way of doing it. This illustrates the concept of zero correlation.

Now suppose $Y = 2X + Z$, where $Z \sim \text{N}(0, 1)$. Now, $\text{Cov}(X, Y) = 2$, and $\text{Var}(Y) = 2$–try and work this out yourself! Theoretically, $\rho_{XY} = 2/\sqrt{1 \cdot 2} \approx 0.71$.

```
Z <- rnorm(n = 100, mean = 0, sd = 1)
Y <- 2 * X + Z
cor(X, Y)  # calculate sample correlation
```

```
## [1] 0.9018487
```

```
qplot(X, Y, geom = "point")
```



As the two random variables are positively correlated, a pattern emerges in the scatterplot, where it seems natural to draw a straight line between the points. Positive correlation suggests that higher values of X are associated with higher values of Y, and vice versa.

1.8.5 Properties of variances and covariances

In this section, we shall state some properties of variances and covariances without proof. You may work this out for yourself, or refer to the textbooks for proofs. While the properties for variance and covariances look similar, there are subtle differences between them. Make sure you note these differences, and study them properly, and preferably commit them to memory so we can make use of these in later calculations.

Let X and Y be random variables, and $a \neq 0, b \in \mathbb{R}$ be constants.

- $\operatorname{Var}(aX + b) = a^2 \operatorname{Var}(X)$
- $\operatorname{Var}(X \pm Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) \pm 2 \operatorname{Cov}(X, Y)$
- If X and Y are independent, then $\operatorname{Var}(X \pm Y) = \operatorname{Var}(X) + \operatorname{Var}(Y)$

As a corollary, let $X_1, \ldots, X_n$ be r.v. Then,

$$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \operatorname{Var}(X_i) + \sum_{i \neq j} \operatorname{Cov}(X_i, X_j)$$

Let X, Y, W, V be r.v., and $a, b, c, d \in \mathbb{R}$. Then

- $\operatorname{Cov}(X, Y) = \operatorname{Cov}(Y, X)$
- $\operatorname{Cov}(X, b) = 0$
- $\operatorname{Cov}(aX, Y) = a \operatorname{Cov}(X, Y)$
- $\operatorname{Cov}(aX + b, cY + d) = ac \operatorname{Cov}(X, Y)$
- $\operatorname{Cov}(X + Y, W + V) = \operatorname{Cov}(X, Y) + \operatorname{Cov}(X, V) + \operatorname{Cov}(Y, W) + \operatorname{Cov}(Y, V)$

Example 1.29. Let $X \sim \mathrm{N}(0, 1)$, and $Y = 2X + 1$. Then

$$\operatorname{Var}(Y) = \operatorname{Var}(2X + 1) = 4 \operatorname{Var}(X) = 4.$$

Further,

$$\operatorname{Cov}(X, Y) = \operatorname{Cov}(X, 2X + 1) = 2 \operatorname{Cov}(X, X) = 2 \operatorname{Var}(X) = 2.$$

1.8.6 Multivariate means and covariances

In the previous section, we discussed bivariate and multivariate random variables. Let's take a look briefly at how one would define expectations and variances for these types of variables. We'll study the more general p-dimensional random variable, and hopefully you'll be able to interpolate to bivariate cases yourself.

Consider a random vector $X = (X_1, \dots, X_p)^\top$ with pdf given by $f_X(x)$ and support over $\mathbb{R}^p$. We define the mean of this random vector as the component-wise mean of the random variables.

Definition 1.21 (Mean vector)**.** Let $X \in \mathbb{R}^p$ be a random vector as described. The *mean vector* of X is

$$\mu = (\mu_1, \dots, \mu_p)^\top,$$

where each μ_k is the mean of each component of the random vector, i.e.

$$\mu_k = \mathrm{E}(X_k) = \begin{cases} \sum_x x \mathbb{P}(X_k = x) & \text{discrete case} \\ \int_x x f_{X_k}(x) \, \mathrm{d}x & \text{continuous case} \end{cases}$$

This is assuming that each of these mean components μ_k are finite.

In order to calculate the mean vector from scratch, we need the marginal distributions of each of the components of X.

In a similar manner we may define the *variance-covariance* matrix. That is, we compute each of the variance of the components of the random vector and collect them into a matrix. Additionally, we have to compute each pairwise covariance to complete populating the matrix.

Definition 1.22 (Variance-covariance matrix)**.** Let $X \in \mathbb{R}^p$ be a random vector as described. The *variance-covariance matrix* of X is the $p \times p$ square, symmetric matrix $\Sigma = (\sigma_{ij})$ where each entry is defined by

$$\sigma_{ij} = \mathrm{E}\left[(X_i - \mu_i)(X_j - \mu_j)\right]$$

for all $i, j = 1, \dots, p$.

Realise that to compute the variance-covariance matrix, one requires the pairwise joint pdfs of the components of X. For illustration, we can see that the Σ matrix looks something like this:

$$\Sigma = \begin{pmatrix} \operatorname{Var}(X_1) & \operatorname{Cov}(X_1, X_2) & \cdots & \operatorname{Cov}(X_1, X_p) \\ \operatorname{Cov}(X_2, X_1) & \operatorname{Var}(X_2) & \cdots & \operatorname{Cov}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{Cov}(X_p, X_1) & \operatorname{Cov}(X_p, X_2) & \cdots & \operatorname{Var}(X_p) \end{pmatrix}$$

The p diagonal entries of Σ are the variances of each X_k, which there are p of them. The off-diagonal entries are the covariances. It is symmetric because $\operatorname{Cov}(X_i, X_j) = \operatorname{Cov}(X_j, X_i)$.

The correlation matrix is similar in structure to the above, with the difference that each entry in the matrix is

$$\sigma_{ij} = \frac{\mathrm{E}\left[(X_i - \mu_i)(X_j - \mu_j)\right]}{\sigma_i \sigma_j} = \rho_{X_i X_j},$$

where σ_k is the standard deviation of each of the components X_k. As a result, the correlation matrix has all diagonals equal to 1, and the off-diagonals represent the correlations between the random variables.

An important type of correlation matrix is the identity matrix I_p, where all the diagonals are 1 and the off-diagonals are 0. As we discussed, this does not necessarily imply that the components of X are independent. However, in the special case of the (multivariate) normal distribution, the identity matrix correlation matrix implies p independent unit variance normal variates.

Example 1.30. Here we'll look at a bivariate random variable $X = (X_1, X_2)^\top \in [0,1]^2$, whose pdf is given by

$$f_{X_1,X_2}(x_1, x_2) = \frac{1}{2} + \frac{3}{4}(x_1^2 + x_2^2).$$

The pdf can be visualised as follows.

```
##
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':
##
##     last_plot

## The following object is masked from 'package:stats':
##
##     filter

## The following object is masked from 'package:graphics':
##
##     layout
```

Let's compute the means and variances of each components X_1 and X_2. For this, we need the marginal pdfs.

$$f_{X_1}(x) = \int_0^1 \left(\frac{1}{2} + \frac{3}{4}(x^2 + y^2)\right) \mathrm{d}y = \frac{3}{4}(x^2 + 1)$$

and by symmetry,

$$f_{X_2}(x) = \frac{3}{4}(x^2 + 1).$$

Both of these pdfs are valid, since they both integrate to 1 in the unit interval region of integration. Since they are identically distributed, $\mathrm{E}(X_1) = \mathrm{E}(X_2)$ and $\mathrm{Var}(X_1) = \mathrm{Var}(X_2)$. We compute these in turn.

$$\mathrm{E}(X_1) = \int_0^1 x \frac{3}{4}(x^2 + 1)\, \mathrm{d}x = \frac{9}{16} = \mathrm{E}(X_2),$$

and

$$\mathrm{Var}(X_1) = \int_0^1 \left(x - \frac{9}{16}\right)^2 \frac{3}{4}(x^2 + 1)\, \mathrm{d}x = 0.0836 = \mathrm{Var}(X_2).$$

As for the covariance, this is computed as follows.

$$\begin{aligned}\mathrm{Cov}(X_1, X_2) &= \int_0^1 \int_0^1 \left(x - \frac{9}{16}\right)\left(y - \frac{9}{16}\right)\left(\frac{1}{2} + \frac{3}{4}(x^2 + y^2)\right) \mathrm{d}x\, \mathrm{d}y \\ &= -0.00391\end{aligned}$$

Therefore, the mean vector for (X, Y) is

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} 9/16 \\ 9/16 \end{pmatrix}$$

while the variance-covariance matrix is

$$\Sigma = \begin{pmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) \\ \text{Cov}(X_1, X_2) & \text{Var}(X_2) \end{pmatrix} = \begin{pmatrix} 0.0836 & -0.00391 \\ -0.00391 & 0.0836 \end{pmatrix}.$$

If we were to convert the above 2×2 matrix to a correlation matrix, we would need to divide each entry by $\sqrt{\text{Var}(X_1)\,\text{Var}(X_2)} = 0.0836$, thus yielding the correlation matrix Φ

$$\Phi = \begin{pmatrix} 1 & \rho(X_1, X_2) \\ \rho(X_1, X_2) & 1 \end{pmatrix} = \begin{pmatrix} 1 & -0.05 \\ -0.05 & 1 \end{pmatrix}.$$

1.8.7 Conditional expectations and variance

In the previous section, we looked at conditional probability functions, both in the discrete and continuous case. These conditional pmfs/pdfs are also useful for calculating *conditional expectations*, i.e. the "average" value of a random variable X given some information about another random variable Y which might affect it. We define it as follows.

Definition 1.23 (Conditional expectation)**.** The *conditional expectation* of a function of a random variable X, $g(X)$ say, given a value of another random variable $Y = y$, is

$$\text{E}\left[g(X)|Y = y\right] = \begin{cases} \sum_x g(x) \overbrace{\mathbb{P}(X = x|Y = y)}^{f_{X|Y}(x|y)} & \text{if } X \text{ is discrete} \\ \int g(x) f_{X|Y}(x|y)\,\mathrm{d}x & \text{if } X \text{ is continuous} \end{cases}$$

In particular, we might be interested in $\text{E}(X|Y)$, which is obtained using $g(X) = X$ in the above definition. Notice that to calculate the conditional expectations, we require the conditional distributions.

Conditional expectations behave quite like regular expectations, so that all of the properties of the usual expectations are applicable. However, whereas $\text{E}(X)$ is a number (non-random), $\text{E}(X|Y = y)$ is a function of y. More importantly, if we have not observed Y, then $\text{E}(X|Y)$ is a **random variable**.

There are some additional properties related to conditional expectations that are somewhat intuitive. We list them here for reference.

- $\text{E}(g(X)|X) = g(X)$–"the given variable is a constant"
- $\text{E}(g(X)Y|X) = g(X)\,\text{E}(Y|X)$–since $g(X)$ given X is a constant, we can pull it out of the expectation.
- If $X \perp Y$, then $\text{E}(Y|X) = \text{E}(Y)$–since the distribution of Y does not depend on X.

Let's take a look at an example.

Example 1.31. Suppose we draw $Y \sim \text{Unif}(0, 1)$. After we observe $Y = y \in [0, 1]$, we draw $X|(Y = y) \sim \text{Unif}(y, 1)$. That is, conditional on the observed value of Y, we draw another uniform distribution whose value is at least Y and at most 1. Intuitively, since this is a uniform distribution, we expect that $\text{E}(X|Y = y)$ to be half-way between y and 1, i.e. $(1 + y)/2$.

In fact this is indeed the case, since $f_{X|Y}(x|y) = (1-y)^{-1}$, so

$$\begin{aligned}
\mathrm{E}(X|Y=y) &= \int_y^1 x f_{X|Y}(x|y)\,\mathrm{d}x \\
&= \frac{1}{1-y}\int_y^1 x\,\mathrm{d}x \\
&= \frac{1-y^2}{2(1-y)} \\
&= \frac{(1-y)(1+y)}{2(1-y)} \\
&= \frac{1+y}{2}.
\end{aligned}$$

However, if Y has not been observed yet, then $\mathrm{E}(X|Y) = (1+Y)/2$ is a random variable whose value is $\mathrm{E}(X|Y=y) = (1+y)/2$ once observed.

If $\mathrm{E}(X|Y)$ is a random variable, then it must have a distribution. Sometimes this is easy to figure out, but other times it is not so straightforward. However, we have a result to easily obtain the mean of this random variable $\mathrm{E}(X|Y)$.

Theorem 1.8 (Rule of iterated expectations/Law of total expectations)**.** *If X and Y are two random variables, then*

$$\mathrm{E}_Y\left[\mathrm{E}(X|Y)\right] = \mathrm{E}(X),$$

provided the expectation exists. More generally, $\mathrm{E}(g(X)) = \mathrm{E}\left[\mathrm{E}(g(X)|Y)\right]$ *for any function g.*

The interpretation of the above theorem is the following: The total average $\mathrm{E}(X)$ is the average of the case-by-case averages $\mathrm{E}(X|Y)$ over Y.

Proof. We can show this for the continuous case, but the proof is easily adapted for the discrete case.

$$\begin{aligned}
\mathrm{E}_Y\left[\mathrm{E}(X|Y)\right] &= \int \left(\int x f_{X|Y}(x|y)\,\mathrm{d}x\right) f_Y(y)\,\mathrm{d}y \\
&= \int\int x \cdot \overbrace{f_{X|Y}(x|y) f_Y(y)}^{f_{X,Y}(x,y)}\,\mathrm{d}y\,\mathrm{d}x \\
&= \int x \cdot \overbrace{\int f_{X,Y}(x,y)\,\mathrm{d}y}^{f_X(x)}\,\mathrm{d}x \\
&= \mathrm{E}(X).
\end{aligned}$$

□

This is a nice example taken from Wikipedia which intuitively explains the Rule of Iterated Expectations:

Example 1.32. Suppose that only two factories supply light bulbs to the market. Factory X's bulbs work for an average of 5000 hours, whereas factory Y's bulbs work for an average of 4000 hours. It is known that factory X supplies 60% of the total bulbs available. What is the expected length of time L that a purchased bulb will work for?

Without knowing anything about Theorem 1.8, we would intuitively work out that

$$\mathrm{E}(L) = 5000 \times 0.6 + 4000 \times 0.4 = 4600.$$

In fact, this is an application of the law of total expectations. Let

- $\mathbb{P}(X) = 0.6$ be the probability that the purchased bulb was manufactured by factory X (information given);

- $\mathbb{P}(X) = 1 - 0.6 = 0.4$ be the probability that the purchased bulb was manufactured by factory Y (deduced by law of total probability);
- $\mathrm{E}(L|X) = 5000$ is the expected lifetime of a bulb manufactured by X (information given);
- $\mathrm{E}(L|Y) = 4000$ is the expected lifetime of a bulb manufactured by Y (information given);

Then $\mathrm{E}(L)$, the expected length of time of a purchased bulb at random will be

$$\mathrm{E}(L) = \mathrm{E}(L|X)\mathbb{P}(X) + \mathrm{E}(L|Y)\mathbb{P}(Y) = 4600.$$

A related concept is the "conditional variance" of a random variable, defined below.

Definition 1.24 (Conditional variance)**.** The *conditional variance* of a function of a random variable X, $g(X)$ say, given a value of another random variable $Y = y$, is

$$\mathrm{Var}(g(X)|Y = y) = \mathrm{E}\left[(g(X) - \mathrm{E}(g(X)|Y = y))^2 \,\middle|\, Y = y\right],$$

provided this expectation exists.

In particular,

$$\mathrm{Var}(X|Y = y) = \mathrm{E}\left[(X - \mathrm{E}(X|Y = y))^2 \,\middle|\, Y = y\right].$$

There is an alternative formula for the conditional variance, similar to what we have seen in the regular variance case:

$$\mathrm{Var}(X|Y = y) = \mathrm{E}\left(X^2|Y = y\right) - \left\{\mathrm{E}(X|Y = y)\right\}^2.$$

The warnings about catastrophic cancellation applies here as well!

Similar to the law of total expectations, we have a result involving conditional variances.

Theorem 1.9 (Law of total variance)**.** *If X and Y are two random variables, then*

$$\mathrm{Var}_Y(X) = \mathrm{E}_Y\left[\mathrm{Var}(X|Y)\right] + \mathrm{Var}_Y\left[\mathrm{E}(X|Y)\right].$$

The above theorem also works for more general functions of X. Note that, in this context, both $\mathrm{Var}(X|Y)$ and $\mathrm{E}(X|Y)$ are random variables, as they depend on the (unknown) value of Y. The law of total variance states that the total variability of X is the sum of two parts:

1. The average of the variance of X over all possible values of the r.v. Y. This is called the average *within-sample variance*.
2. The variance of the conditional expectation of X given Y. This is called the *between-sample variance* (of the conditional averages).

The next section provides some additional explainers involving conditional expectations and variances, and also gives a proof for Theorem 1.9.

It is always advisable to keep track of which distribution the expectation is taken under. For instance,

$$\mathrm{E}_X(X) = \int x f_X(x)\,\mathrm{d}x$$

is the expectation of X using the pdf of X. There is little ambiguity here so we needn't put a subscript for the expectation and it's fine to write it simply as $\mathrm{E}(X)$. On the other hand,

$$\mathrm{E}_Y(X) = \int X f_Y(y)\,\mathrm{d}y = X\int f_Y(y)\,\mathrm{d}y = X$$

which makes sense because X is constant in Y so gets pulled out of the expectation. Keeping track of expectations is even more important when using iterated expectations under the law of total expectations and total variance!

Example 1.33. Here's a problem faced by actuaries (in the field of insurance and actuarial science). Insurance companies are interested in the *loss amount* of insuring their customers. Those that are classified as "high risk" of course tend to have a higher loss amount compared to those classified as "low risk", and each category has its own distribution.

Let U be the uniform distribution on the unit interval $(0, 1)$. Suppose that a large population of insureds is composed of "high risk" and "low risk" individuals. The proportion of insured classified as "low risk" is $p \in (0, 1)$. Let the random loss amount X of a "low risk" insured be U; and let the random loss amount X of a "high risk" insured be U shifted by a positive constant $w > 0$, i.e. $U + w$.

What is the variance of the loss amount of an insured randomly selected from this population?

For convenience, let $Y = 1$ if a person is categorised as "high risk" and $Y = 0$ if they are categorised as "low risk". Then Y is a Bernoulli random variable

$$Y = \begin{cases} 1 & \text{w.p. } 1-p \\ 0 & \text{w.p. } p \end{cases}$$

The information that we have is the following: $X|(Y = 0) \sim \text{Unif}(0, 1)$, while $X|(Y = 1) \sim \text{Unif}(w, w+1)$. This is because a random variable shifted by a constant is still uniform, and the endpoints are shifted by that constant. Thus, we know (or at least we can check from Wikipedia!–don't worry though, we'll cover properties of commonly used distributions in the next chapter)

$$\begin{aligned} \mathrm{E}(X|Y=0) &= \frac{1}{2} & \mathrm{E}(X|Y=1) &= w + \frac{1}{2} \\ \mathrm{Var}(X|Y=0) &= \frac{1}{12} & \mathrm{Var}(X|Y=1) &= \frac{1}{12} \end{aligned}$$

The law of total variance gives us a way to compute the total variance $\mathrm{Var}(X)$:

$$\begin{aligned} \mathrm{Var}(X) &= \mathrm{E}_Y\left[\mathrm{Var}(X|Y)\right] + \mathrm{Var}_Y\left[\mathrm{E}(X|Y)\right] \\ &= \mathrm{E}_Y\left[\frac{1}{12}\right] + \mathrm{Var}_Y\left[wY + \frac{1}{2}\right] \\ &= \frac{1}{12} + w^2\,\mathrm{Var}(Y) \\ &= \frac{1}{12} + w^2 p(1-p) \end{aligned}$$

In the above we made use of properties of expectations and variances to make computing the total variance easier. For the first part of the sum, notice that the conditional variance of Y is $1/12$ no matter the value of Y, hence it is a constant, and the expectation of a constant is that constant itself. As for the second part, the conditional expectation $\mathrm{E}(X|Y)$ can be written as a function of Y as $wY + 1/2$–if $Y = 0$ then the expectation is $1/2$, but if it is $Y = 1$ then the expectation is $w + 1/2$. Since w is a constant, it is pulled out and squared, and the $+\ 1/2$ bit has zero variance.

Out of interest, the total expectation can be computed as

$$\mathrm{E}(X) = p \cdot \mathrm{E}(X|Y=0) + (1-p)\,\mathrm{E}(X|Y=1) = \frac{1}{2} + (1-p)w,$$

but it would be wrong to do something similar to the variance. That is,

$$\mathrm{Var}(X) \neq p \cdot \mathrm{Var}(X|Y=0) + (1-p)\,\mathrm{Var}(X|Y=1) = \frac{1}{12}.$$

It is not possible for the total variance to be the same as the conditional variances, because the mean loss in the two cases are different! The uncertainty in the risk classes introduces additional varibility in the loss for a randomly selected insured individual, so should be higher than $1/12$.

1.8.8 Additional explainers

Let (X, Y) be a bivariate random variable with some pdf $f(x, y)$. To illustrate the concepts we've been talking about so far, consider drawing a random sample of size n from this distribution. The scatterplot is shown below, with values of X on the vertical axis, and Y on the horizontal axis.

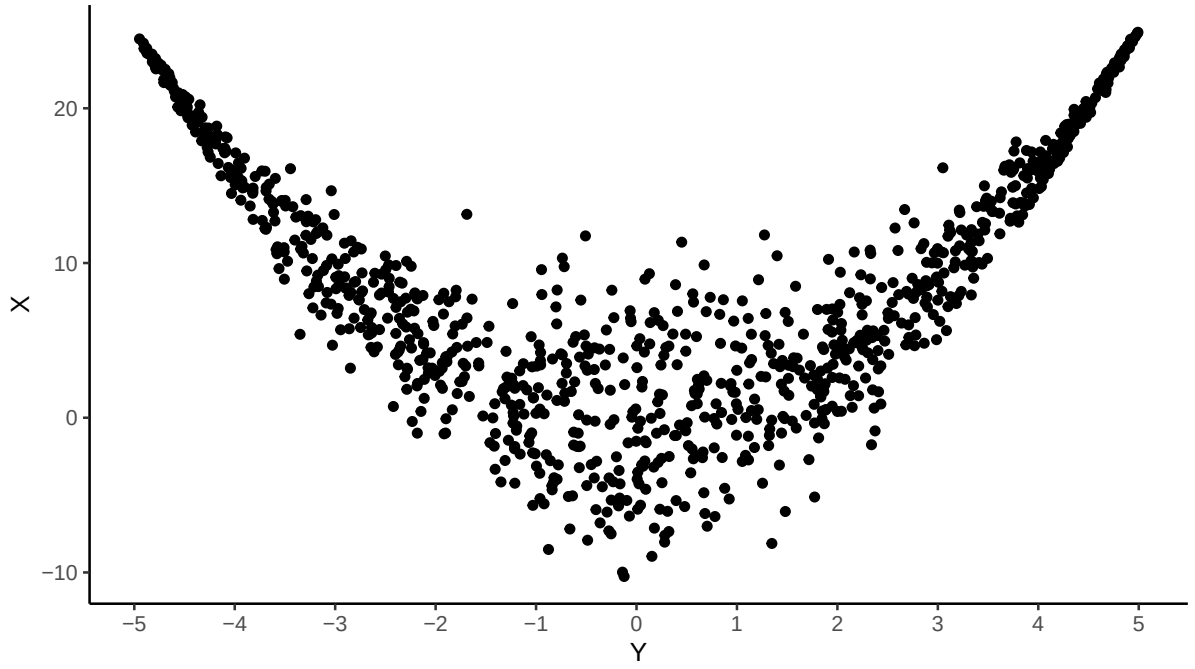


Figure 1.11: Scatter plot of randomly sampled values from a bivariate distribution. We can see a U-shape relationship between the two random variables.

The first thing to note is that there is clearly a dependence between X and Y. In the region of $X > 0$, values of Y increase as X increases, and we see a similar pattern in the opposite direction, i.e. in the region of $X < 0$, values of Y increase as X decreases. In other words, values of Y increases as $|X|$ increases. We say there is a *non-linear* dependence of X on Y (a straight line plotted through the points will not sufficiently capture the relationship between them).

Next, suppose we colour the the points by the values of X. In the plot below, large values of X are coloured on the yellow end of the spectrum, while small values of X are coloured dark blue. The marginal distribution of X is akin to "flattening" the horizontal axis until every point lies in a straight vertical line. The vertical histogram on the right side gives a representation of how distributed the points are along this flat vertical line. The *mass* of the distribution seems to lie in the middle, perhaps somewhere in between 0 and 5. If we calculated the sample mean of all the observed X values, this would given (an estimate) for the *marginal* expectation of X. This is marked in the plot by the horizontal dashed line.

Similarly, we can get the marginal distribution for Y as well as the marginal expectation of Y by flattening the vertical axis.

When we speak about "flattening" the horizontal axis, this is essentially what we do when we compute the marginal distribution,

$$f_X(x) = \int_y f_{X,Y}(x, y)\,\mathrm{d}y$$

as we are *marginalising* over the values of Y. What results is the distribution of X!

From here, we can compute the *marginal* expectation of X, i.e.

$$\mathrm{E}(X) = \int_x x \cdot \overbrace{\left(\int_y f_{X,Y}(x, y)\,\mathrm{d}y\right)}^{f_X(x)} \mathrm{d}x$$

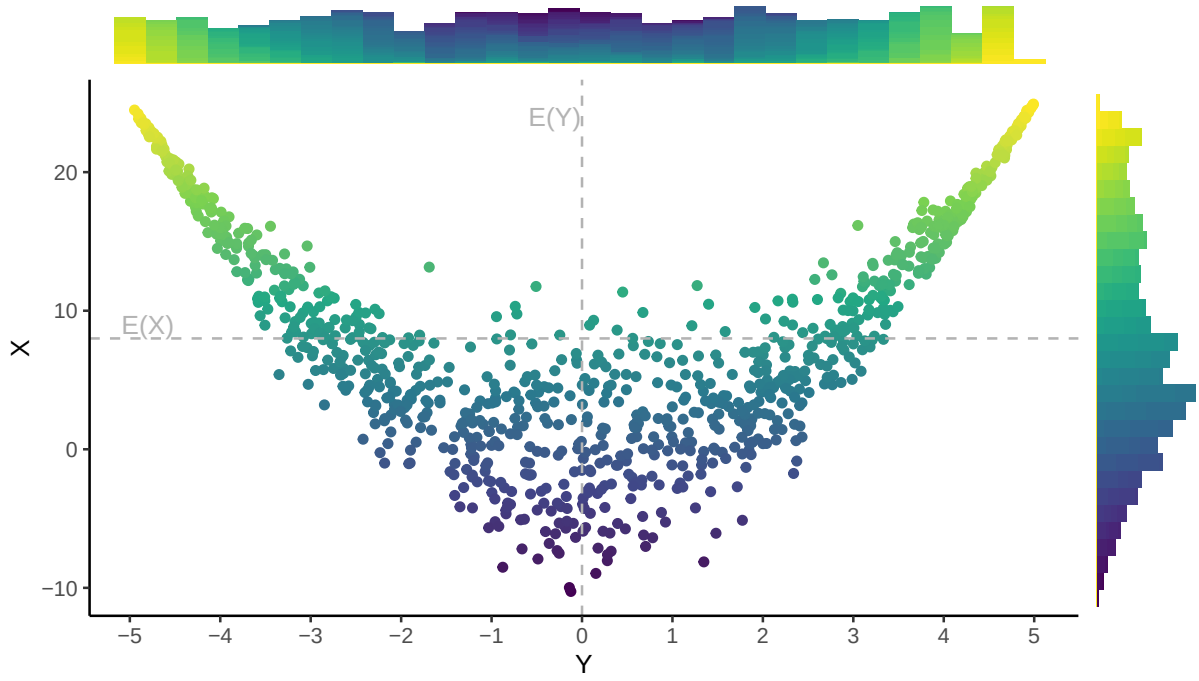


Figure 1.12: Illustration of the marginal distributions of X and Y together with their respective means.

Now, imagine that we concentrate on vertical strips of the scatterplot centred on various values of Y. When we concentrate on just one strip, the distribution of X then changes, and thus the expected value *conditional on this value of* Y also changes. We clearly see that the distribution changes depending on where we draw the vertical strip, i.e. what the value of Y is. This is the concept for *conditional distributions*.

From these conditional distributions, we can clearly visualise the and *conditional expectations* and *conditional variances* through the boxplots at different locations of Y (see Figure **??**). For instance, the boxplots at $Y = 0$ are definitely wider than that at $Y = 5$–there is more variability at $Y = 0$ and less at $Y = 5$. Also, at $Y = 0$, the mean of the points within the green strip is about 0–we can think of this as roughly representing $\mathrm{E}(X|Y = 0) = 0$. The conditional mean changes for different locations of Y. For $Y = 5$ for instance, the conditional mean of X is about 20.

In Figure **??** as well we have drawn a quadratic line through the data points. This line actually represents the curve $g(y) = \mathrm{E}(X|Y = y)$. That is to say, the mean of X conditional on the value of $Y = y$ is represented along this curve. The relationship between the curve and the total expectation $\mathrm{E}(X)$ is that $\mathrm{E}(X)$ is the average of all values of the curve $\mathrm{E}(X|Y = y)$ weighted by the probability distribution of Y. Since the distribution of Y is pretty uniform (see Figure **??**(fig:condexp2)), this explains why the dotted horizontal line is somewhere in the middle of the black curve. This explains the law of total expectations.

We can also illustrate the law of total variance using this graphic. Consider any point in this scatterplot, and its distance to the overall mean $\mathrm{E}(X)$. It can be split into two parts, as follows:

$$\overbrace{X - \mathrm{E}(X)}^{\text{blue part}} = \overbrace{X - \mathrm{E}(X|Y)}^{\text{red part}} + \overbrace{\mathrm{E}(X|Y) - \mathrm{E}(X)}^{\text{green part}} \tag{1.11}$$

This is illustrated in the following graphic:

If this is the case, using the formula for the variance as well as properties of expectations on (1.11), we

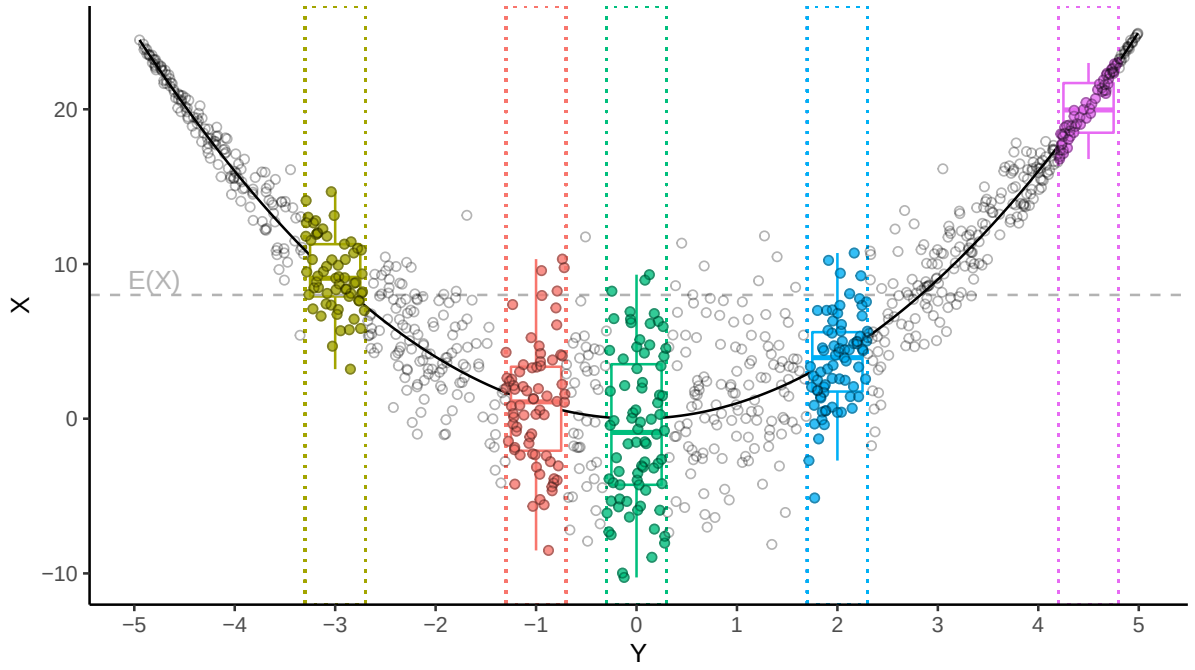


Figure 1.13: The distribution of X conditional on different values of Y.

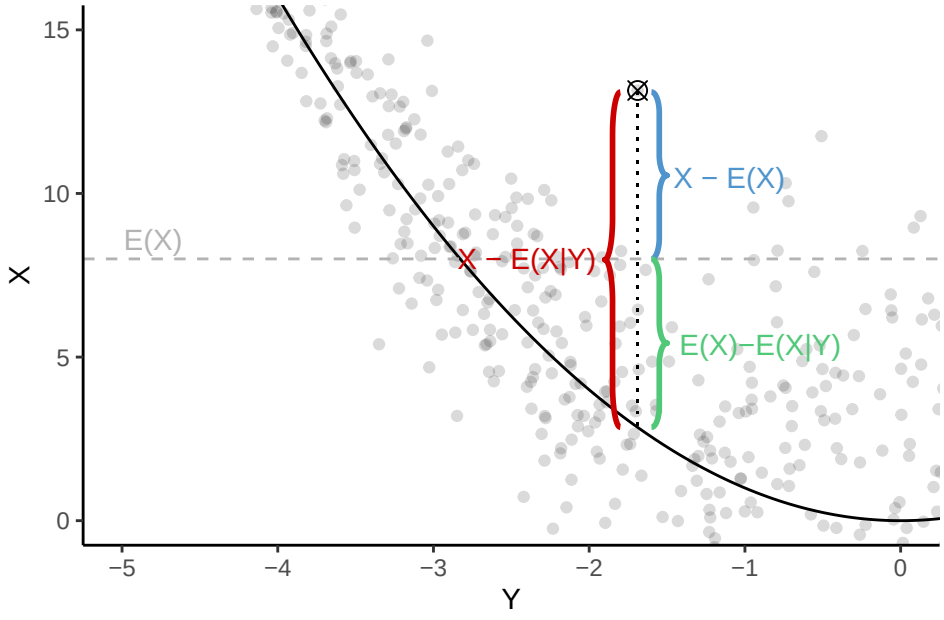


Figure 1.14: The distribution between any point and the grand mean is the sum of two parts: a) the distance between the point and the conditional mean; and b) the distance between the conditional mean and the grand mean.

get the relationship

$$\begin{aligned}
\{X - \mathrm{E}(X)\}^2 &= \{X - \mathrm{E}(X|Y) + \mathrm{E}(X|Y) - \mathrm{E}(X)\}^2 \\
&= \{X - \mathrm{E}(X|Y)\}^2 + \{\mathrm{E}(X|Y) - \mathrm{E}(X)\}^2 \\
&\quad + 2\{X - \mathrm{E}(X|Y)\}\{\mathrm{E}(X|Y) - \mathrm{E}(X)\} \\
\Rightarrow \mathrm{E}\{X - \mathrm{E}(X)\}^2 &= E\{X - \mathrm{E}(X|Y)\}^2 + \mathrm{E}\{\mathrm{E}(X|Y) - \mathrm{E}(X)\}^2 \\
&\quad + 2\,\mathrm{E}\{X - \mathrm{E}(X|Y)\}\{\mathrm{E}(X|Y) - \mathrm{E}(X)\} \quad (\to 0) \\
\Rightarrow \mathrm{Var}(X) &= \mathrm{E}\{\mathrm{Var}(X|Y)\} + \mathrm{Var}\{\mathrm{E}(X|Y)\}
\end{aligned}$$

as required. Some additional algebra to support the above:

$$\begin{aligned}
\mathrm{E}\{X - \mathrm{E}(X|Y)\}^2 &= \mathrm{E}\left\{\mathrm{E}\left[\{X - \mathrm{E}(X|Y)\}^2 \mid Y\right]\right\} \\
&\quad \text{by the Law of Total Expectations} \\
&= \mathrm{E}\{\mathrm{Var}(X|Y)\}
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}\{\mathrm{E}(X|Y) - \mathrm{E}(X)\}^2 &= \mathrm{E}\{\mathrm{E}(X|Y) - \mathrm{E}(\mathrm{E}(X|Y))\}^2 \\
&\quad - \mathrm{Var}\{\mathrm{E}(X|Y)\}
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}\left\{(X - \mathrm{E}(X|Y))\overbrace{(\mathrm{E}(X|Y) - \mathrm{E}(X))}^{g(Y)}\right\} &= \mathrm{E}\{g(Y)(X - \mathrm{E}(X|Y))\} \\
&= \mathrm{E}_Y[\mathrm{E}\{g(Y)(X - \mathrm{E}(X|Y))\} \mid Y] \\
&= \mathrm{E}_Y[g(Y)\{(\mathrm{E}(X|Y) - \mathrm{E}(X|Y))\} \mid Y] \\
&= 0
\end{aligned}$$

1.9 Moment generating functions

Interestingly, the concept of *moments* in statistics is borrowed from physics, in the sense that the moment of a function are measures relating to the shape of a function's graph. Wikipedia states that if a function represents mass, the first moment is the center of the mass, and the second moment is the rotational inertia. In statistics, we concern ourselves with moments of probability functions.

Definition 1.25 (Moments)**.** The kth *moment* of a random variable X is

$$m_k = \mathrm{E}(X^k).$$

The kth *central moment* of X is

$$\mu_k = \mathrm{E}\left[(X - \mu)^k\right],$$

where $\mu = \mu_1' = \mathrm{E}(X)$. The kth *standardised moment* of X is

$$\tilde{\mu}_k = \mathrm{E}\left[\left(\frac{X - \mu}{\sigma}\right)^k\right],$$

where $\sigma^2 = \mu_2$.

Some common moments of a random variable X are as follows.

- The first moment is the *mean* or *expectation*.
- The second central moment is the *variance*.
- The third standardised moment is the *skewness*, a measure of asymmetry of the pdf about its mean.
- The fourth standardised moment is the *kurtosis*, a measure of "tailed-ness" of aa pdf.

1.9.1 Moment generating functions

There is a function in statistics which is useful for finding moments of a random variable, called the *moment generating function*.

Definition 1.26 (Moment generating function)**.** Let $X \sim f_X(x)$. For $t \in \mathbb{R}$, the *moment generating function (mgf)* of X is defined by

$$M_X(t) = \mathrm{E}(e^{tX}) = \begin{cases} \sum_x e^{tx} f_X(x) & \text{if } X \text{ discrete} \\ \int e^{tx} f_X(x)\,\mathrm{d}x & \text{if } X \text{ continuous} \end{cases}$$

provided this expectation exists in "some neighbourhood of 0".

While the primary use of the mgf is for finding moments of a random variable, it has also these other uses (read on further down for properties of mgfs):

- Characterising a distribution.
- Finding distributions of sums of random variables.
- As a tool in statistical proofs

How exactly does it "generate" moments? Consider finding the first derivative of $M_X(t)$ evaluated at $t = 0$:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} M_X(t)\bigg|_{t=0} &= \frac{\mathrm{d}}{\mathrm{d}t}\,\mathrm{E}(e^{tX})\bigg|_{t=0} \\ &= \mathrm{E}\left[\frac{\mathrm{d}}{\mathrm{d}t} e^{tX}\right]\bigg|_{t=0} \\ &= \mathrm{E}[Xe^{tX}]\Big|_{t=0} \\ &= \mathrm{E}(X). \end{aligned} \tag{1.12}$$

We get exactly the first moment. These steps can be iterated to generate the k-th moment of X, i.e. by taking k derivatives and setting $t = 0$.

Theorem 1.10. *If X has mgf $M_X(t)$, then*

$$\mathrm{E}(X^k) = M_X^{(k)}(0) = \frac{d^k}{dt^k} M_X(t)\bigg|_{t=0}.$$

That is, the k-th moment is equal to the k-th derivative of $M_X(t)$ evaluated at $t = 0$.

There are two subtle things going on here when we use the mgf. The first one is that obtaining moments relies on being able to *interchange the order of differentiation and integration*, because for continuous distributions,

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathrm{E}(e^{tX}) = \frac{\mathrm{d}}{\mathrm{d}t}\left(\int e^{tx} f_X(x)\,\mathrm{d}x\right)$$

and in (1.12) above, it is assumed we can interchange the order of differentiation and integration so that

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathrm{E}(e^{tX}) = \int \left(\frac{\mathrm{d}}{\mathrm{d}t} e^{tx}\right) f_X(x)\,\mathrm{d}x = \mathrm{E}\left(Xe^{tX}\right).$$

For "nice" and well-behaved random variables, there are generally no problems. See Section 2.4 of C&B.

The second thing is that statement about the expectation "existing in some neighbourhood of 0". This is so that when we plug in $t = 0$ in the expression for the mgf, we get a valid expectation. If this is not the case, then the moment does not exist.

Example 1.34. Let $X \sim \text{Exp}(1/r)$ with $f_X(x) = re^{-rx}$ for $x \in [0, \infty)$. This is an exponential distribution with rate parameter r. Then for $t < r$,

$$\begin{aligned} M_X(t) &= \int_0^\infty e^{tx} \cdot re^{-rx} \, \mathrm{d}x \\ &= r \int_0^\infty e^{(t-r)x} \, \mathrm{d}x \\ &= \frac{r}{r-t}. \end{aligned}$$

Thus, $\text{E}(X) = M_X'(t)\big|_{t=0} = \frac{r}{(r-t)^2}\Big|_{t=0} = 1/r$. Indeed, the expectation of the exponential distribution is given as the inverse rate. We'll meet the exponential distribution in more detail in the next chapter.

Here are some properties of mgfs that are worth noting.

- If $Y = aX + b$, then $M_Y(t) = e^{bt} M_X(at)$.
- If $X_1, \ldots, X_n$ are independent and $Y = \sum_{i=1}^n X_i$, then $M_Y(t) = \prod_{i=1}^n M_{X_i}(t)$.
- If X and Y are random variables s.t. $M_X(t) = M_Y(t)$ for all t in an open interval around 0, then $F_X(x) = F_Y(x)$ for all x.

The mgf has the property that it uniquely defined a distribution. That is, if two distributions have identical mgfs then they have the same distribution. This way, we can usually identify distributions from a recognisable mgf. We'll see some examples of this in action in the later chapters.

1.10 Exercises

1. Using only the three axioms of probability, prove the following statements:

   (a) $\Pr(\{\}) = 0$
   (b) If $A \subseteq B$ then $\Pr(A) \leq \Pr(B)$ *Hint: Write* $B = A \cup (B \cap A^c)$
   (c) $0 \leq \Pr(A) \leq 1$
   (d) $\Pr(A^c) = 1 - \Pr(A)$
   (e) If $A \cap B = \{\}$, then $\Pr(A \cup B) = \Pr(A) + \Pr(B)$

2. This is called the "Monty Hall Problem". A prize is placed at random behind one of three doors. You pick a door. To be concrete, let's suppose you always pick door 1. Now Monty Hall chooses one of the other two doors, opens it and shows you that it is empty. He then gives you the opportunity to keep your door or switch to the other unopened door. Should you stay or switch? Intuition suggests it doesn't matter. The correct answer is that you should switch. Prove it.

   It will help to specify the sample space and the relevant events carefully. Thus write $\Omega = \{(\omega_1, \omega_2) \,|\, \omega_i \in \{1, 2, 3\}\}$ where ω_1 is where the prize is and ω_2 is the door Monty opens.

3. There are three cards. The first is green on both sides, the second is red on both sides and the third is green on one side and red on the other. We choose a card at random and we see one side (also chosen at random). If the side we see is green, what is the probability that the other side is also green? Many people intuitively answer 1/2. Show that the correct answer is 2/3.

4. (a) For independent events $A_1, \ldots, A_n$, show that

   $$\Pr(A_1 \cup \cdots \cup A_n) = 1 - \prod_{i=1}^n (1 - \Pr(A_i)).$$

   (b) A pair of dice is rolled n times. How large must n be so that the probability of rolling at least one double six is more than 1/2?

5. The probability that a child has blue eyes is 1/4. Assume independence between children. Consider a family with 3 children.

(a) If it is known that at least one child has blue eyes, what is the probability that at least two children have blue eyes?
(b) If it is known that the youngest child has blue eyes, what is the probability that at least two children have blue eyes?

6. Prove the following statements.

(a) If $A \perp B$, then $A^c \perp B^c$.
(b) $\Pr(A \cap B \cap C) = \Pr(A|B \cap C)\Pr(B|C)\Pr(C)$.

7. Let X be distributed according to

$$f_X(x) = \begin{cases} 1/4 & 0 < x < 1 \\ 3/8 & 3 < x < 5 \\ 0 & \text{otherwise} \end{cases}$$

(a) Show that f_X is indeed a probability density function.
(b) Find the cumulative distribution function of X.

8. Suppose we toss a coin once and let p be the probability of heads. Let X denote the number of heads and let Y denote the number of tails. Prove that X and Y are independent.

9. Let

$$f_{X,Y}(x,y) = \begin{cases} c(x+y^2) & 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

(a) Find the value of c.
(b) Find $\Pr(X < 1/2|Y = 1/2)$.

10. Consider a sequence of independent coin flips, each of which has probability p of being heads. Define a random variable X as the length of the run (of either heads or tails) started by the first trial. For example, $X = 3$ if either $TTTH$ or $HHHT$ is observed. Find the distribution of X and find $\mathrm{E}(X)$.

11. For a random variable X with mean μ and variance $\mathrm{Var}(X)$ and any given constant $c \in \mathbb{R}$, prove that

(a) $\mathrm{Var}(X) = \mathrm{E}(X^2) - \mu^2$.
(b) $\mathrm{Var}(X) = \mathrm{E}\left(X(X-1)\right) + \mu - \mu^2$.
(c) $\mathrm{E}\left((X-c)^2\right) = \mathrm{Var}(X) + (\mu - c)^2$ so that the minimum mean squared deviation occurs when $c = \mu$.

12. Suppose we play a game where we start with c dollars. On each play of the game you either double or halve your money, with equal probability. What is your expected fortune after n trials?

13. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{Unif}(0,1)$ and let $Y_n = \max\{X_1, \ldots, X_n\}$. Find $\mathrm{E}(Y_n)$. *Hint: Find out the distribution of Y_n by looking at the cdf of Y_n.*

14. Let $X \sim \mathrm{Unif}(0,1)$. Let $0 < a < b < 1$. Let

$$Y = \begin{cases} 1 & 0 < x < b \\ 0 & \text{otherwise} \end{cases}$$

and let

$$Z = \begin{cases} 1 & a < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

(a) Are Y and Z independent? Why/why not?

(b) Find $\mathrm{E}(Y|Z)$. *Hint: What values z can Z take? Find first* $\mathrm{E}(Y|Z = z)$.

Hand-in questions

1. A certain river floods every year. Suppose that the low-water mark is set at 1 and the high-water mark Y has distribution function

$$F_Y(y) = \Pr(Y \leq y) = 1 - \frac{1}{y^2}, \qquad 1 \leq y < \infty$$

   (a) Verify that $F_Y(y)$ is a cdf. **[1 mark]**
   (b) Find $f_Y(y)$, the pdf of Y. **[2 marks]**
   (c) If the low-water mark is reset at 0 and we use a unit of measurement that is 1/10 of that given previously, the height water mark becomes $Z = 10(Y - 1)$. What is the expected value of Z? **[2 marks]**

2. A pdf is defined by

$$f_{X,Y}(x, y) = \begin{cases} c(x + 2y) & 0 \leq x \leq 2, 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

   (a) Find the value of c. **[1 mark]**
   (b) Find the marginal distribution of X. **[2 marks]**
   (c) Find the joint cdf of X and Y. **[2 marks]**

3. Suppose we generate a random variable X in the following way. First we flip a fair coin. If the coin is heads, take X to have a Unif$(0, 1)$ distribution. If the coin is tails, take X to have a Unif$(3, 4)$ distribution.

   (a) Find the mean of X. **[2 marks]**
   (b) Find the standard deviation of X. **[3 marks]**

Chapter 2

Commonly-used probability models

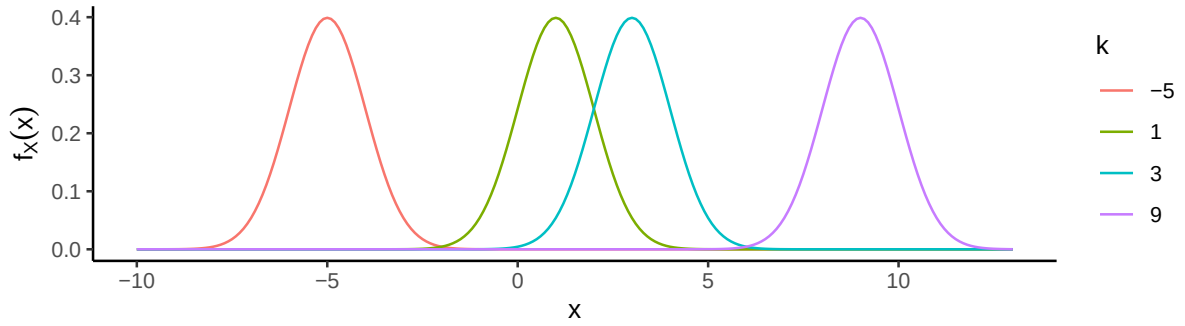
Distributions in statistics serve two main purposes:

1. To describe the assumed behaviour of the observations made in an experiment, survey or other study;
2. To calibrate the values of derived statistics used in constructing confidence regions, hypothesis tests, etc.

Some distributions are much used for both purposes (the normal distribution being the prime example).

In this chapter, we will focus on some distributions used for the first purpose. Distributions used mainly for the second purpose (these include the χ^2, t and F distributions) will be described later, in Chapter 3.

We will meet a *family* of distributions–a family which is indexed by one or more *parameters* (c.f. parametric family). This allow us to vary certain characteristics of the distribution while staying with one functional form. For example, consider several random variables $X_k \sim \mathrm{N}(k, 1)$, for a couple of values of k. That is, X_k is a normal distribution with mean k and variance 1. These are distinct distributions (e.g., $\Pr(X_{10} < 0) \ll \Pr(X_0 < 0)$) yet have similar characteristics (e.g. they are all bell-shaped).



Starting from this chapter onwards, we will denote probability using the notation 'Pr' rather than the blackboard bold symbol '$\mathbb{P}$' as we did previously. Either symbol is fine, but I feel '$\mathbb{P}$' is used to denote the probability measure of an event, and is traditionally used in probability theory. Of course when we write $\Pr(X < a)$ we really mean the probability measure of the event $\{X < a\}$ so really the measure theoretic stuff is there regardless (or maybe we should write it $\mathbb{P}(\{X < a\})$ to be proper?).

The point is, write it any way you want. It should be fine. Just so you know, I'll be using 'Pr' from here on out.

Learning objectives

By the end of this chapter, you will be:

- Familiar with commonly used discrete and continuous probability models, including their general properties and characteristics, and situations in which that distribution is used to model data.
- Able to compute probabilities involving normal distributions using the standard normal tables.
- Discovering some relationships between certain distributions, and how these relationships can be exploited in approximate calculations.

Readings

- Casella and Berger (2002)
  - Chapter 3, sections 3.1 3.2 3.3
- Wasserman (2004)
  - Chapter 2, sections 2.3 and 2.4.
  - Chapter 3, section 3.6.
- Topics not covered: Cauchy, lognormal and double exponential (Laplace) distributions, exponential families, location and scale families

2.1 Discrete models

In this section, we'll be describing commonly used **discrete** probability distributions.

2.1.1 Point mass distribution

The random variable X has a point mass distribution at a, written $X \sim \delta_a$, if $\Pr(X = a) = 1$, in which case

$$F(x) = \begin{cases} 0 & x < a \\ 1 & x \geq a. \end{cases}$$

The probability mass function is $f(x) = 1$ for $x = a$, and 0 otherwise.

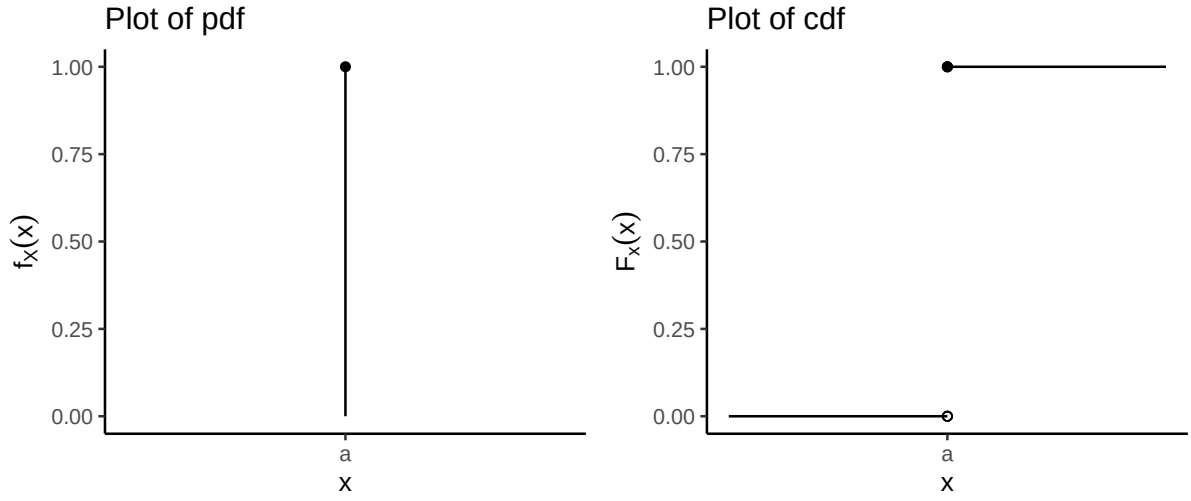


Figure 2.1: Pdf and cdf of the point mass distribution.

The mean and variance are trivial: $\mathrm{E}(X) = a$ and $\mathrm{Var}(X) = 0$, because the "random" variable X takes on the value a with probability 1 (certainty).

There isn't much practical use for point mass distributions to be honest, but sometimes they are used to describe *mixture* distributions. For example, a random variable might be equal to 0 half of the time, but may be normally distributed the other half of the time.

2.1.2 Uniform distribution

Let $k > 1$ be a given integer. The discrete uniform distribution on $\{1, \dots, k\}$ has pmf

$$f(x) = \frac{1}{k}, \qquad x = 1, \dots, k.$$

We write $X \sim \mathrm{Unif}\{1, \dots, k\}$. Its mean and variance are

- $\mathrm{E}(X) = \frac{k+1}{2}$; and
- $\mathrm{Var}(X) = \frac{k^2-1}{12}$.

The mean of the discrete uniform is intuitive; it is the half-way point between 1 and k. The discrete uniform (and the point mass) is appeallingly simple but has relatively few "real" statistical applications.

Proof. Using the arithmetic series formulae $\sum_{i=1}^{n} x = n(n+1)/2$ and $\sum_{i=1}^{n} x^2 = n(n+1)(2n+1)/6$, we have

$$\begin{aligned}
\mathrm{E}(X) &= \sum_{x=1}^{k} \frac{x}{k} \\
&= \frac{k(k+1)}{2k} \\
&= \frac{k+1}{2},
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}(X^2) &= \sum_{x=1}^{k} \frac{x^2}{k} \\
&= \frac{k(k+1)(2k+1)}{6k} \\
&= \frac{(k+1)(2k+1)}{6},
\end{aligned}$$

hence

$$\begin{aligned}
\mathrm{Var}(X) &= \mathrm{E}(X^2) - \mathrm{E}^2(X) \\
&= \frac{(k+1)(2k+1)}{6} - \frac{(k+1)^2}{4} \\
&= \frac{2(k+1)(2k+1) - 3(k+1)^2}{12} \\
&= \frac{k^2-1}{12}.
\end{aligned}$$

□

If $k = 1$, then it is the point mass distribution. But actually, we can use whatever labels we want for the values $1, 2, \dots, k$. For example, suppose we are picking between 3 colours uniformly, then we can describe a uniform distribution on $\{1 = \text{red}, 2 = \text{green}, 3 = \text{blue}\}$.

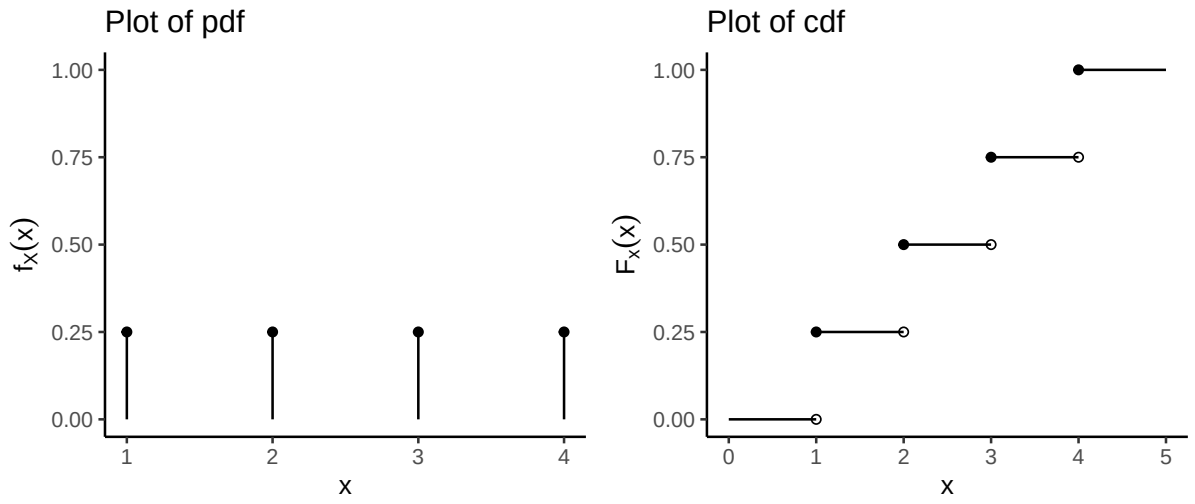


Figure 2.2: Pdf and cdf of the discrete uniform distribution for $k = 4$.

2.1.3 Bernoulli distribution

Suppose we are interested in the outcome of a (single) random trial, which can either be "success" or "failure" only. Examples include

- A coin flip can land either Heads or Tails.
- The colour of the suit of a randomly drawn card from a pack of playing cards can be either Red or Black
- A dice roll outcome can either be an Even or an Odd number.
- Babies being born being Girl or Boy.

Typically we assign the value '1' to denote success, and '0' to denote failure. This has no qualitative meaning whatsoever, the important thing is that there are only two distinct possible outcomes. As a consequence, any discrete random variable that can take on only two possible outcomes is a Bernoulli random variable.

Let X be the r.v. denoting the outcome of success ($X = 1$) or failure ($X = 0$) of a binary trial. Further let the pmf for X be

$$f(x|p) = \begin{cases} p & x = 1 \text{ (success)} \\ 1 - p & x = 0 \text{ (failure)} \end{cases}$$

We say that X has a Bernoulli distribution written $X \sim \text{Bern}(p)$.

- The expectation is
$$\text{E}(X) = \sum_x x f(x) = 1 \cdot p + 0 \cdot (1 - p) = p.$$
- The variance is
$$\text{Var}(X) = \sum_x (x - \mu)^2 f(x) = (1 - p)^2 \cdot p + (0 - p)^2 \cdot (1 - p) = p(1 - p).$$

Again the expectation is intuitive here. If a proportion p of the time we get success, then surely the expectation must be this proportion p.

The pmf for the Bernoulli distribution can also be written $f(x) = p^x(1 - p)^{1-x}$. Try plugging in $x = 1$ and $x = 0$ into this function. Do you get the appropriate probabilities?

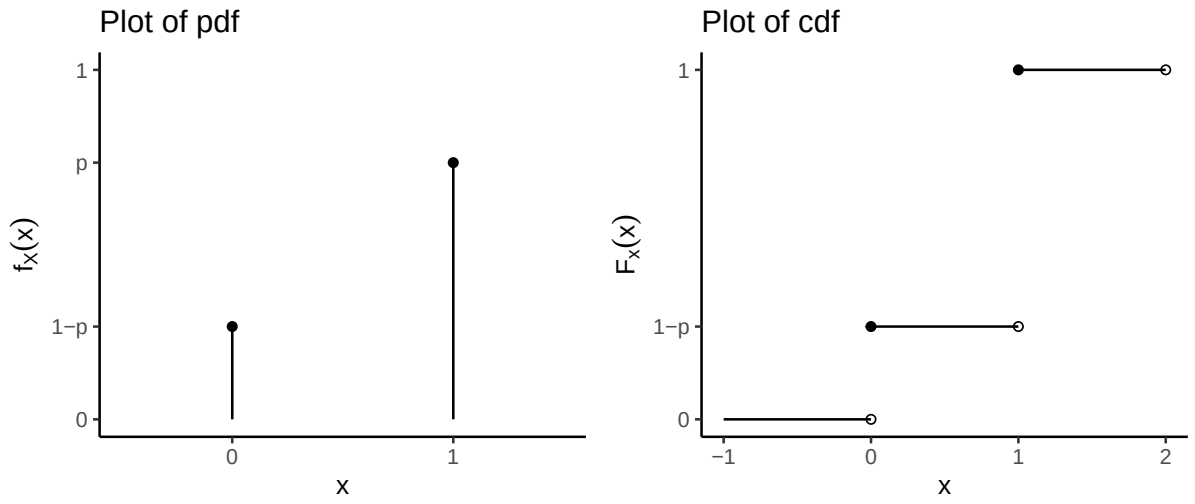


Figure 2.3: Pdf and cdf of the Bernoulli distribution.

2.1.4 Binomial distribution

A related distribution to the Bernoulli is the binomial distribution. It describes the distribution of the number of "successes" in n independent and identical binary "trials". That is, suppose we have a situation such that

- A finite number n trials are carried out.
- Each trial is independent of each other.
- The outcome of each trial is either success or failure (binary trials).
- The probability $0 \leq p \leq 1$ of a successful outcome is the same for each trial.

Let X be the number of success outcomes in n trials. Then X has a binomial distribution, written $X \sim \text{Bin}(n, p)$. The pmf of X is

$$f(x|n,p) = \binom{n}{x} p^x (1-p)^{n-x}.$$

X has support (possible values it can take) over $\{0, 1, 2, \ldots, n\}$. The mean and variance are $\text{E}(X) = np$ and $\text{Var}(X) = np(1-p)$.

Proof. Here's the proof for the mean.

$$\begin{aligned}
\text{E}(X) &= \sum_{x=0}^{n} x \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \\
&= \sum_{x=1}^{n} x \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \\
&= \sum_{x=1}^{n} n \cdot \overbrace{\frac{(n-1)!}{(x-1)!(n-x)!}}^{\binom{n-1}{x-1}} \cdot p^{x-1+1} (1-p)^{(n-1)-(x-1)} \\
&= np \overbrace{\sum_{x-1=0}^{n} \binom{n-1}{x-1} p^{x-1} (1-p)^{n-x}}^{=1} \\
&= np.
\end{aligned}$$

Obtaining the variance follows similar steps. □

Try to replicate the proof above and obtain $\text{E}(X^2)$ for the binomial distribution. After that, you may obtain $\text{Var}(X)$ using the usual formula. An alternative way is to use mgfs–keep on reading.

For the more astute of you, you might have realised that the binomial distribution is simply counting the number of successes in many independent Bernoulli trials. Indeed, that is the case.

Lemma 2.1. *Let* $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bern}(p)$*, then*

$$X = \sum_{i=1}^{n} X_i \sim \text{Bin}(n, p)$$

The lemma is proven using mgfs, and is left to you as an exercise. First, obtain the mgf for Bernoulli and binomial distributions. Then use the sum property of the mgfs characterise the distribution of the sum.

Using the lemma above, we can more easily derive that

$$\text{E}(X) = \text{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{E}(X_i) = np$$

and

$$\text{Var}(X) = \text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i) = np(1-p)$$

simply using properties of expectations and variances.

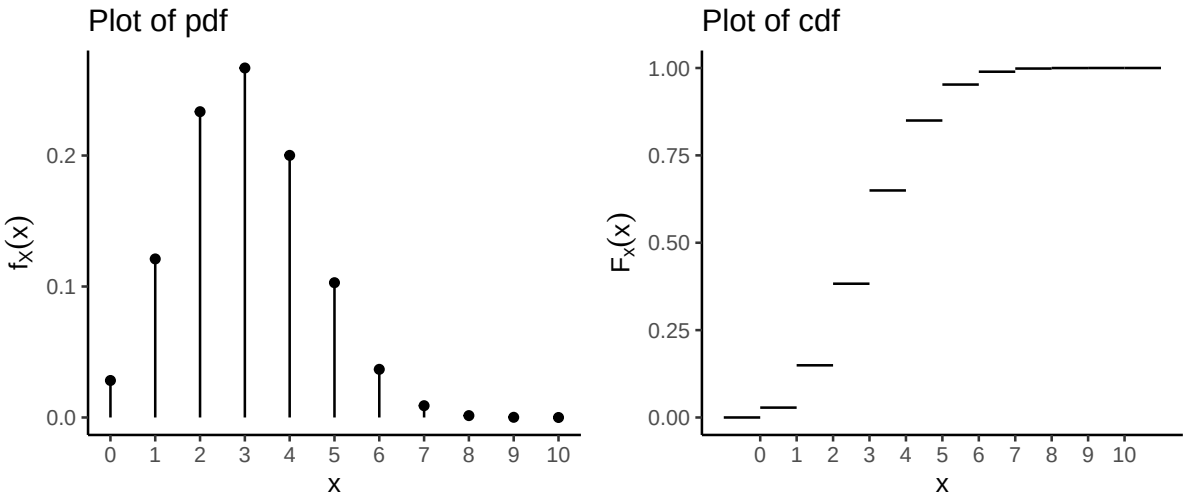


Figure 2.4: Pdf and cdf of the binomial distribution with $p = 0.3$.

2.1.5 Geometric distribution

The geometric distribution is a type of "waiting time" distribution. It involves counting the number of Bernoulli trials to get the first success. Let X be distributed geometrically, $X \sim \text{Geom}(p)$, where p is the probability of success. Clearly,

$$f(x|p) = (1-p)^{x-1}p.$$

The support of X is $\{1, 2, 3, \ldots\}$; it is countably infinite.

This is a valid pmf since

$$\begin{aligned}\sum_{x=1}^{\infty} f(x|p) &= \sum_{x=1}^{\infty} (1-p)^{x-1}p \\ &= \frac{p}{1-(1-p)} = 1.\end{aligned}$$

This required knowledge of the infinite geometric series $\sum_{k=0}^{\infty} ar^k = a/(1-r)$ for $|r| < 1$. The mean and variance of the geometric distribution are:

- $\mathrm{E}(X) = \frac{1}{p}$. The smaller the p, the longer we have to wait for a success.
- $\mathrm{Var}(X) = \frac{1-p}{p^2}$.

There is another formulation for the geometric distribution: Let Y be the number of failures before the first success occurs. Then

$$f(y|p) = (1-p)^y p.$$

Y has support $\{0, 1, 2, \dots\}$. X and Y are related through $Y = X - 1$. Thus it is easy to check that

$$\mathrm{E}(Y) = \frac{1-p}{p} \text{ and } \mathrm{Var}(Y) = \frac{1-p}{p^2}.$$

We shall mainly use the first version of the geometric distribution in this course, but be aware of the alternative version as well.

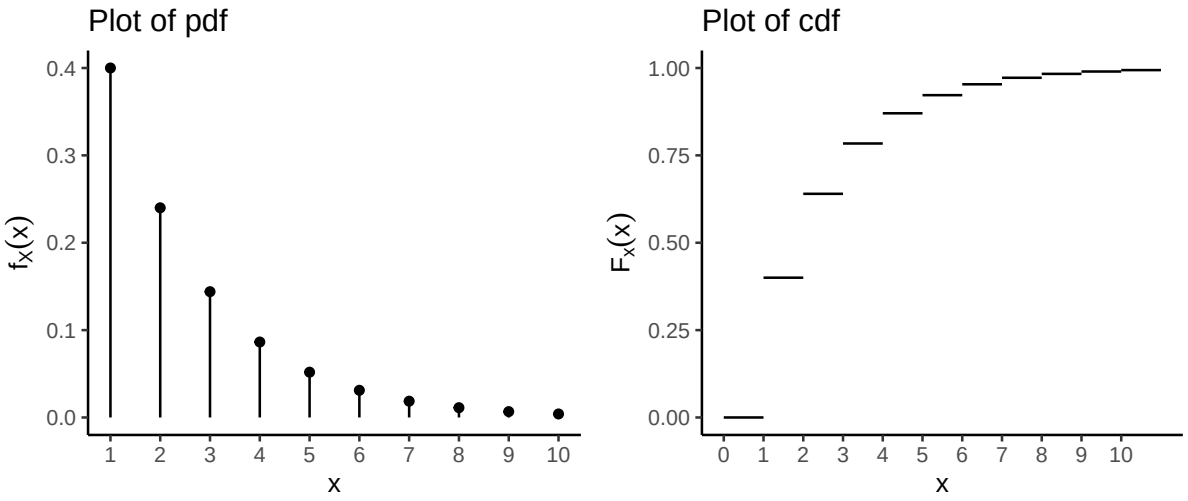


Figure 2.5: Pdf and cdf of the geometric distribution with $p = 0.4$.

2.1.6 Negative binomial

Suppose we count the number of Bernoulli trials required to get a fixed number of successes, r, each with probability of success p. This leads to the negative binomial distribution. Denote this by $X \sim \mathrm{NBin}(r, p)$. The pmf is

$$f(x|r,p) = \binom{x-1}{r-1} p^r (1-p)^{x-r}.$$

The pmf is easy to justify: In order to get $X = x$, a total of $r - 1$ successes must have occurred in the previous $x - 1$ number of trials. Then, the pmf follows directly from the binomial pmf.

Clearly, the support of X is $\{r, r+1, r+2, \dots\}$. The expectation and variance of X are

- $\mathrm{E}(X) = \frac{r}{p}$.
- $\mathrm{Var}(X) = \frac{r(1-p)}{p^2}$.

Note that if $r = 1$, then X is the geometric distribution.

The name 'negative binomial' comes from noting that $Y = X - r$, the number of failures seen before the rth success, has pmf

$$f(y|r,p) = (-1)^y \binom{-r}{y} p^r (1-p)^{r-y},$$

which looks suspiciously close to the binomial pmf[1].

[1] Details in C&B, p.95

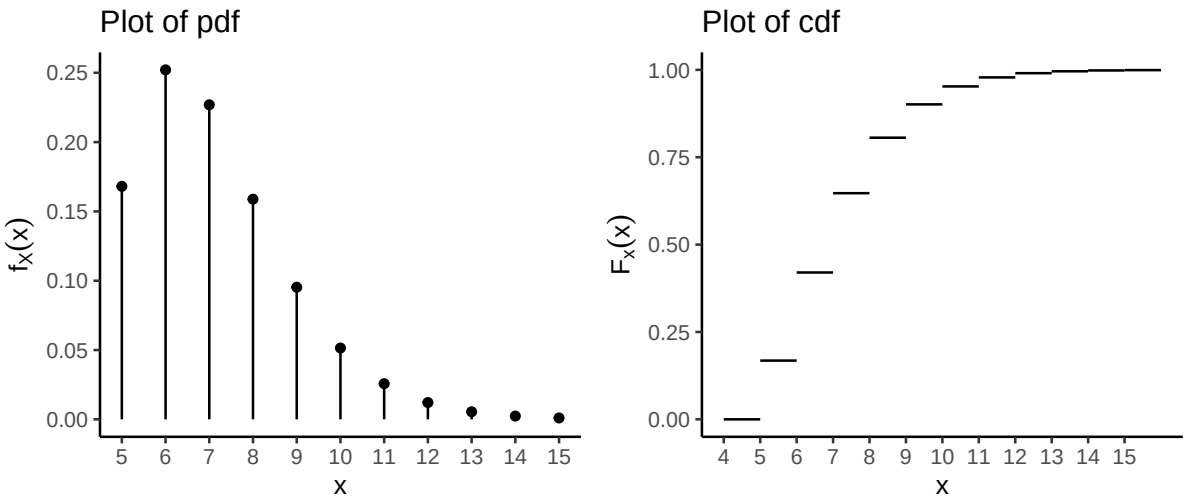


Figure 2.6: Pdf and cdf of the negative binomial distribution with $p = 0.3$ and $r = 5$.

Just a caution to say that the negative binomial has many different parameterisations. It all depends whether

- The random variable X is counting the number of trials until a fixed success, or counting the number of successes before a fixed number of failures occur;
- p denotes success or failure;
- r denotes success or failure.

See here for more details: https://en.wikipedia.org/wiki/Negative_binomial_distribution#Alternative_formulations

2.1.7 Poisson distribution

The Poisson is the most standard assumption for the distribution of a count of events that occur (separately and independently, by assumption) in time or space. Some examples:

- Amount of e-mails received in 24-hour period.
- Number of calls received by a call centre per hour.
- The number of photons hitting a detector in a particular time interval.
- The number of patients arriving in an emergency room between 10pm and 11pm.

Let X be the number of occurrences in this interval, such that the mean number of occurrences λ in the given interval (sometimes called the rate or intensity) is known and is finite. Then $X \sim \text{Poi}(\lambda)$, and

$$f(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!},$$

for $x = 0, 1, 2, ...$

To work out the mean, we make use of the Taylor series expansion. Recall that $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$. Using this fact we can derive the moments through the mgf.

$$\begin{aligned} M_X(t) &= \sum_{x=0}^{\infty} \frac{e^{tx}e^{-\lambda}\lambda^x}{x!} \\ &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!} \\ &= e^{-\lambda}e^{\lambda e^t} = \exp\{\lambda(e^t - 1)\}. \end{aligned}$$

Hence $\mathrm{E}(X) = M_X'(0) = \lambda$ and $\mathrm{E}(X^2) = M_X''(0) = \lambda^2 + \lambda$, so $\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X) = \lambda$.

The Poisson family is closed under addition. If X and Y are independent Poisson r.v. with means λ and μ, then

$$X + Y \sim \mathrm{Poi}(\lambda + \mu)$$

The proof uses mgf and the characterizing property of the mgf.

Have a go at the proof using properties of the mgf.

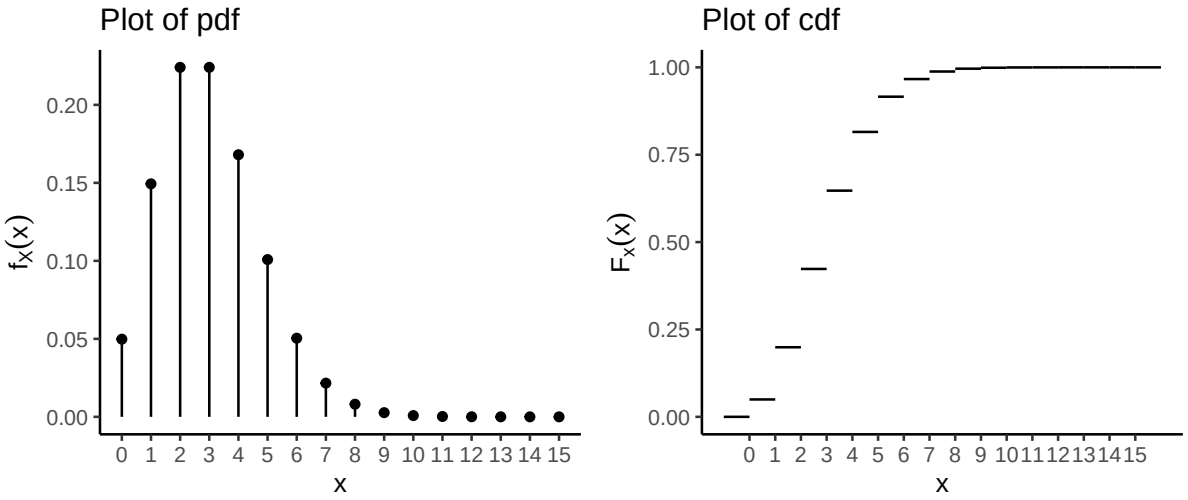


Figure 2.7: Pdf and cdf of the negative binomial distribution with $\lambda = 3$.

2.2 Continuous models

In this section, we'll study commonly used continuous distributions.

2.2.1 Continuous uniform distribution

The continuous uniform distribution is usually taken to have support on an interval, say $a \leq x \leq b$. Let $X \sim \mathrm{Unif}(a, b)$. The pdf is

$$f_X(x) = \frac{1}{b-a}$$

for $x \in [a, b]$ and 0 otherwise. The mean and variance are

- $\mathrm{E}(X) = \frac{a+b}{2}$.
- $\mathrm{Var}(X) = \frac{(a-b)^2}{12}$.

Note the similarities between the means and variances above and in the discrete case!

The plot of the pdf gives a "rectangular" shape, so probabilities can also be found geometrically, as we previously saw in Chapter 1.

2.2.2 Exponential distribution

The exponential distribution is often used to describe the distribution of measured time intervals 'duration data' or 'waiting-time data'. E.g.

- the amount of time until an earthquake occurs.

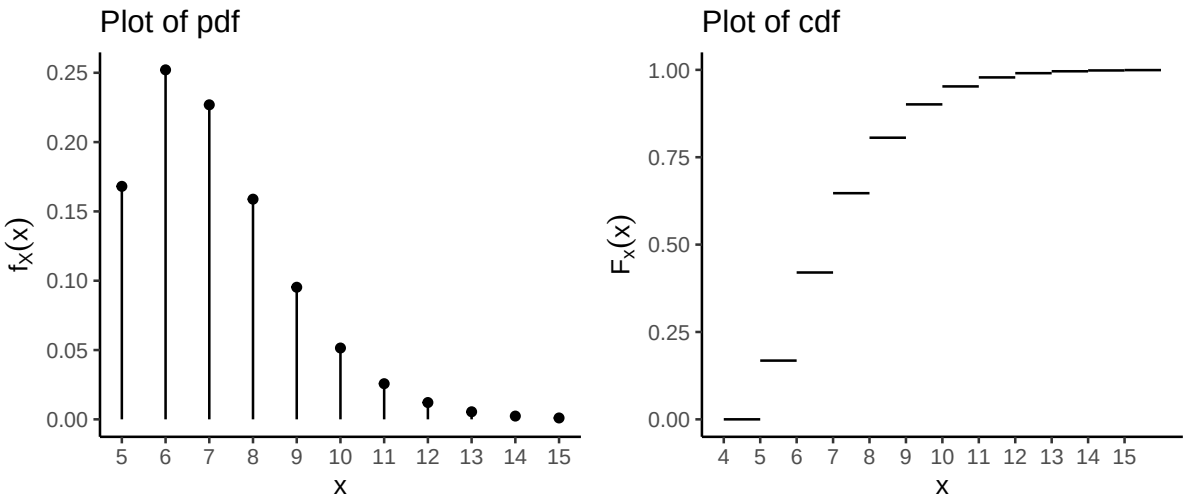


Figure 2.8: Pdf and cdf of the negative binomial distribution with $p = 0.3$ and $r = 5$.

- the time between two lightbulbs failing.
- the length (in minutes) of faculty staff meetings at UBD.
- the average waiting time at a hospital's A&E.

Let $X \sim \text{Exp}(\lambda)$. The pdf is

$$f_X(x) = \frac{1}{\lambda} e^{-x/\lambda}.$$

X has support over $[0, \infty)$, and $\lambda > 0$ is known as the "scale" parameter. The value $r = 1/\lambda$ is known as the "rate". The mean and variance are:

- $\text{E}(X) = \lambda$.
- $\text{Var}(X) = \lambda^2$.

Since λ is a scale parameter, the following property holds: $aX \sim \text{Exp}(a\lambda)$ for $a > 0$. What this is saying is that if we stretch (or contract) the scale of X by some amount $a > 0$, then the underlying distribution will still be the same.

The pdf experiences "exponential decay"–long wait times between two events occurring becomes more and more unlikely.

The exponential distribution has a very special property: it is memoryless, in the sense that for all $t > s > 0$,

$$\Pr(X > t + s | X > s) = \Pr(X > t)$$

> Given that we have been waiting for an event to occur for s units of time, the probability that we wait a further t units of time is independent to the first fact!

For example[2], assume that bus waiting times are exponentially distributed. A memoryless wait for a bus would mean that the probability that a bus arrived in the next minute is the same whether you just got to the station or if you've been sitting there for twenty minutes already.

We can show that X is a positive r.v. and memoryless if and only if it is exponentially distributed.

You will prove the memoryless fact in one of the exercises for this chapter.

[2] https://perplex.city/memorylessness-at-the-bus-stop-f2c97c59e420?gi=3602158da66b

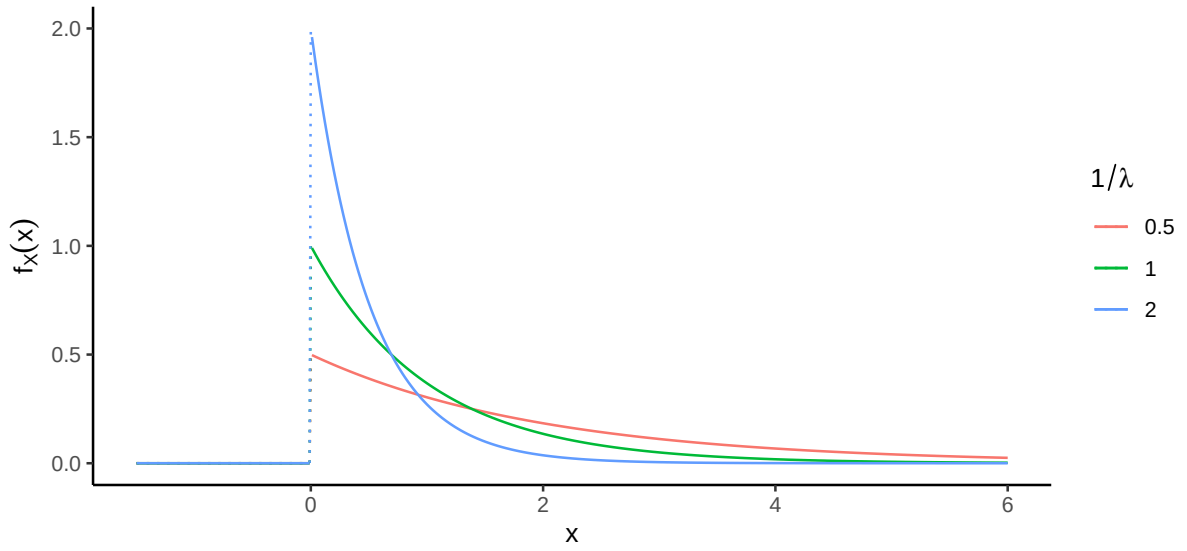


Figure 2.9: Pdf of the exponential distribution.

2.2.3 Gamma distribution

The gamma distribution generalises the exponential distribution. It is also used for modelling durations (lengths of time intervals). Let $X \sim \Gamma(\alpha, \beta)$. The pdf is

$$f_X(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$$

where $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} \, \mathrm{d}y$ is called the Gamma function. Again here, X has support over $[0, \infty]$. $\alpha > 0$ is the "shape" parameter, and $\beta > 0$ is the "scale" parameter. The mean and variance are as follows:

- $\mathrm{E}(X) = \alpha\beta$.
- $\mathrm{Var}(X) = \alpha\beta^2$.

Some other properties of the gamma distribution are:

- $\Gamma(1, \lambda) \equiv \mathrm{Exp}(\lambda)$.
- $aX \sim \Gamma(\alpha, a\beta)$ for $a > 0$.
- If $X_i \sim \Gamma(\alpha_i, \beta)$, then $\sum_{i=1}^n X_i \sim \Gamma(\sum_{i=1}^n \alpha_i, \beta)$.

Be aware that there is an alternative parameterisation of the exponential and gamma distribution using "scale" parameters:

- $Y \sim \mathrm{Exp}(\lambda)$, where $f_Y(y) = \frac{1}{\lambda} e^{-y/\lambda}$. Here λ is the …
- $Y \sim \Gamma(\alpha, s)$, where $f_Y(y) = \frac{1}{\Gamma(\alpha)s^\alpha} y^{\alpha-1} e^{-y/s}$. Here s is the **scale** parameter. The shape parameter is obtained via $\beta = 1/s$.

2.2.4 Beta distribution

The beta distributions are distributions on the unit interval $[0, 1]$, or on any other interval $[a, b]$ by transformation $X \mapsto X(b-a) + a$. It is used to model the behaviour of random variables limited to intervals of finite length in a wide variety of disciplines. Let $X \sim \mathrm{Beta}(\alpha, \beta)$. The pdf is

$$f_X(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}.$$

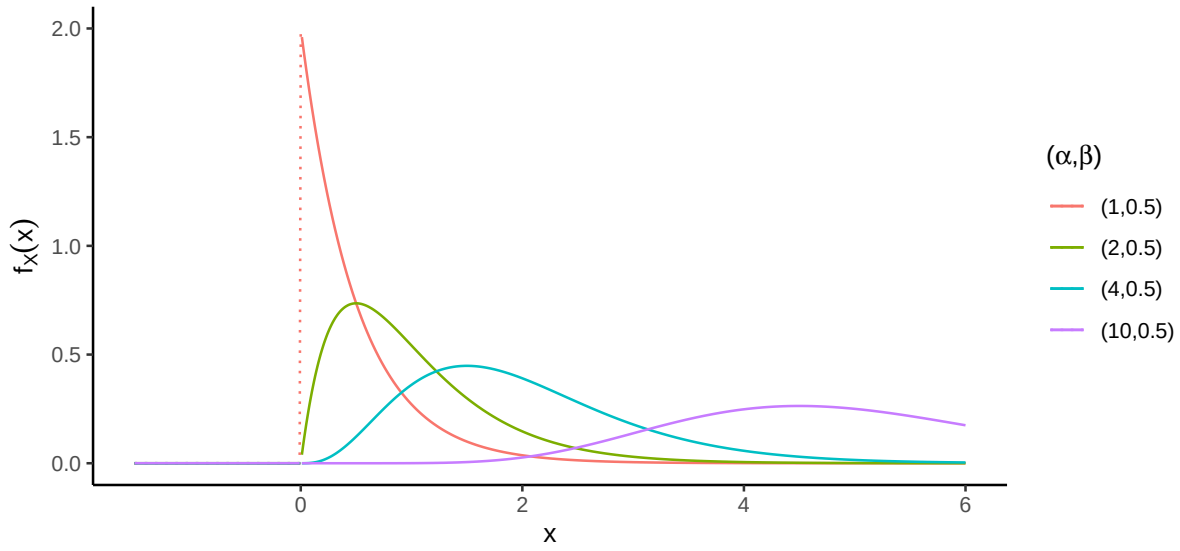


Figure 2.10: Pdf of the gamma distribution.

where $B(\alpha, \beta)$ is the so-called beta function

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}.$$

The parameters $\alpha > 0$ and $\beta > 0$ are known as the "shape" parameters. The mean and variance are

- $\mathrm{E}(X) = \frac{\alpha}{\alpha+\beta}$.
- $\mathrm{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

On interesting fact is that when $\alpha = \beta = 1$, we have the uniform distribution. I.e., $\mathrm{Beta}(1, 1) \equiv \mathrm{Unif}(0, 1)$.

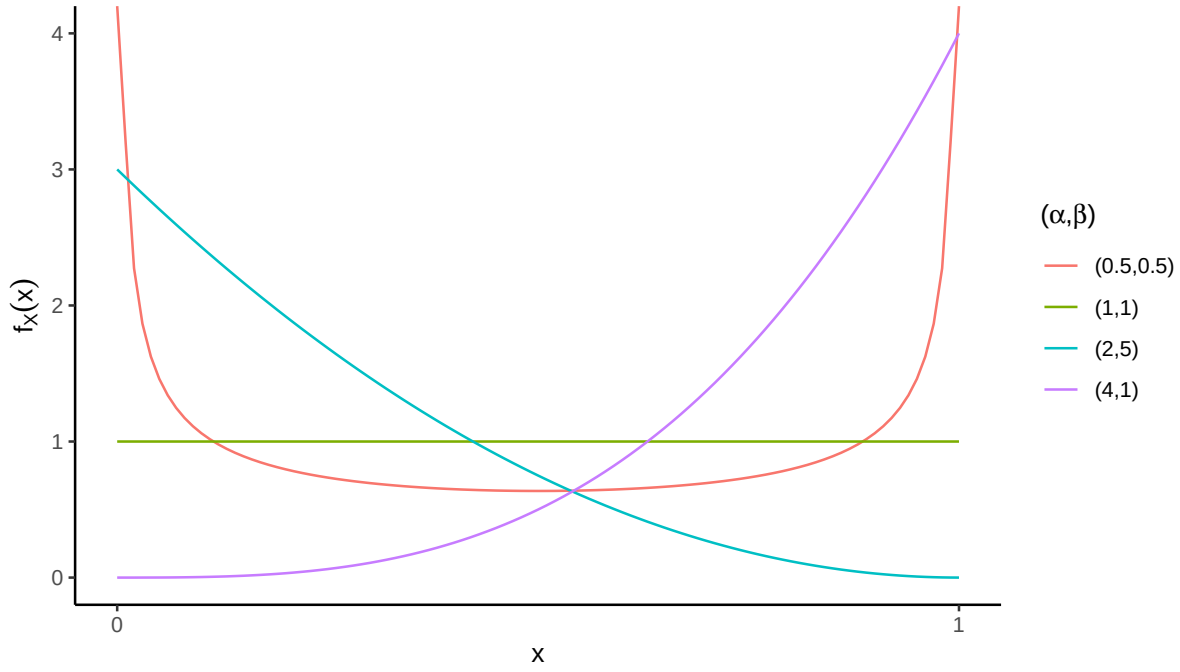


Figure 2.11: Pdf of the beta distribution.

2.3 Normal distribution

The normal distribution[3] is arguably the most important distribution in statistics. There are several reasons for this, including:

- Many naturally occurring phenomena can be modelled as following a normal distribution.
- The central limit theorem (CLT): The distribution of the mean of a sample tends to converge to a normal distribution, as more and more samples are collected.
- Often, the normal distribution is used for the error term in standard statistical models (e.g. linear regression).

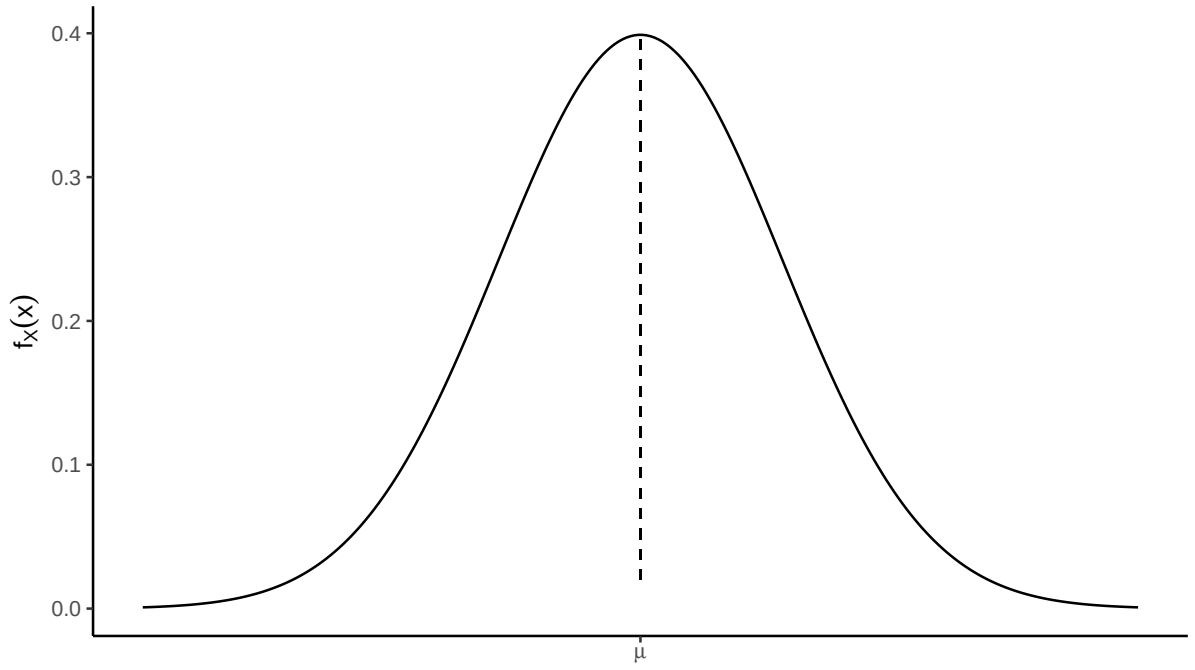
Let us introduce the normal distribution. Let X be distributed according to a normal distribution with mean μ and variance σ^2. We write $X \sim \mathrm{N}(\mu, \sigma^2)$. The pdf of X is

$$f_X(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

X has support over $\mathbb{R}$.

The normal distribution is so important that all my previous teachers have made their students (including me) memorise the pdf of the normal distribution. In continuing tradition, you should too!

The normal distribution is **symmetric** about μ. The mode and median of X is also μ.



The celebrated mathematician Carl Friedrich Gauss has the honour of having the normal distribution named after him. The normal distribution is often known as the Gaussian distribution. Actually, it is named as such because the *kernel* (i.e. the insides) of the pdf contains a Gaussian function, i.e. functions of the form $f(x) = a\exp\left(-\frac{(x-b)^2}{2c^2}\right)$, which Gauss studied quite a lot.

[3]Here's a nice short exploration of the normal distribution: https://bookdown.org/cquirk/LetsExploreStatistics/lets-explore-the-normal-distribution.html

Figure 2.12: Carl Friedrich Gauß. 30 April 1777 – 23 February 1855.

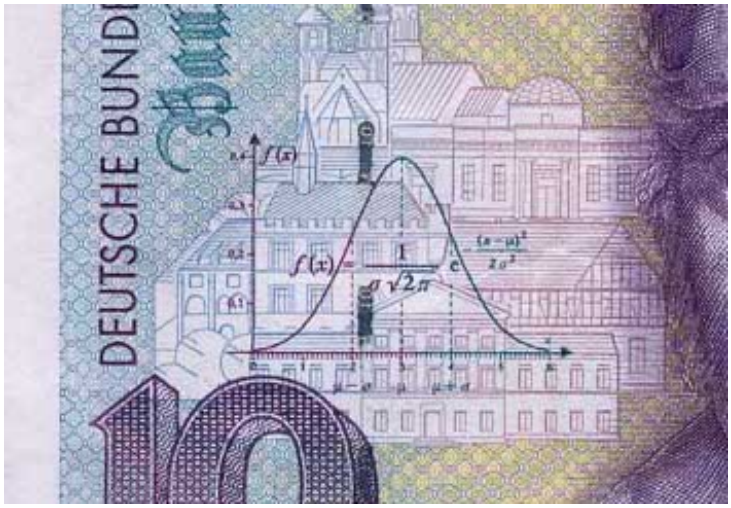
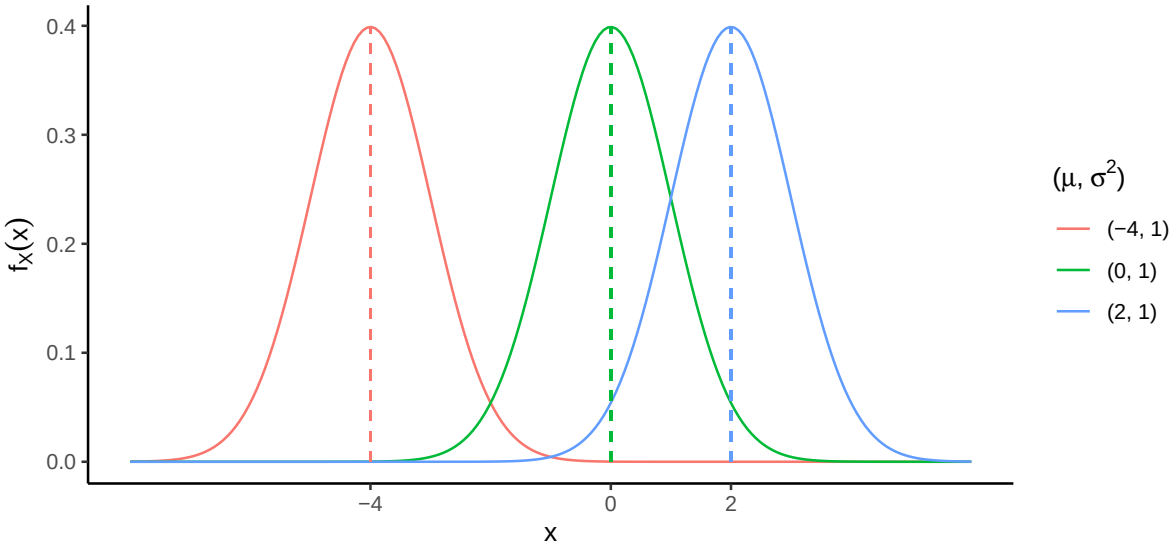


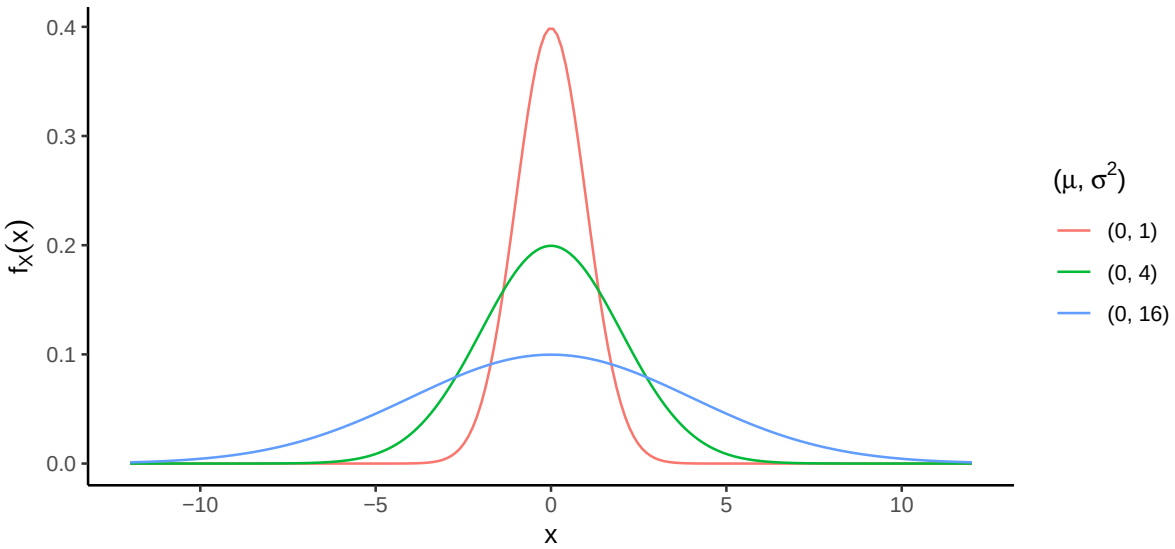
Figure 2.13: 10 Deustche Mark banknote. You know you've made it far when your pdf (used to) appears on a country's banknote.

2.3.1 Location and scale parameter

The μ parameter is also called the "location" parameter, since it determines where the bell curve is placed. Changing the location parameter changes the location of the bell curve.



The σ^2 parameter is also called the "scale" parameter, since it determines how spread out the curve is. Here we see the effect of changing the scale parameter.



2.3.2 Linear transformations of normal random variables

For any constants $c, d \in \mathbb{R}$, the r.v. $Y = cX + d$ also has a normal distribution. This fact makes the normal distribution such a convenient distribution to work with. Using the properties of expectations and variances,

- $\mathrm{E}(Y) = \mathrm{E}(cX + d) = c\mu + d$.
- $\mathrm{Var}(Y) = \mathrm{Var}(cX + d) = c^2\sigma^2$.

The facts above are proven using mgf. See the exercises at the end of this chapter.

In particular, a very important transformation is the *standardisation*

$$Z = \frac{X - \mu}{\sigma}$$

resulting in the *standard normal distribution* $Z \sim \mathrm{N}(0, 1)$. The standard normal has mean 0 and variance 1. It has pdf

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}},$$

specially denoted by the greek letter 'ϕ'.

2.3.3 The normal cdf

The cdf of the normal distribution $X \sim \mathrm{N}(\mu, \sigma^2)$ is

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\tilde{x}-\mu)^2}{2\sigma^2}} \,\mathrm{d}\tilde{x} =: \Phi\left(\frac{x-\mu}{\sigma}\right),$$

where $\Phi(z) = \int_{-\infty}^{z} \phi(\tilde{z}) \,\mathrm{d}\tilde{z}$ is the cdf of $Z = \frac{X-\mu}{\sigma}$.

Values of $\Phi(\cdot)$ must be read from a table[4], as the integrals above are *intractable* (no closed form solution). Some results worth noting:

- $\Pr(Z \leq -a) = \Phi(-a) = 1 - \Phi(a)$
- $\Pr(a \leq Z \leq b) = \Phi(b) - \Phi(a)$
- $P(-a \leq Z \leq b) = \Phi(a) + \Phi(b) - 1$
- $P(-a \leq Z \leq -b) = \Phi(b) - \Phi(a)$
- $P(|Z| \leq a) = P(-a \leq Z \leq a) = 2\Phi(a) - 1$
- $P(|Z| \geq a) = P(\{Z < -a\} \cup \{Z > a\}) = 2(1 - \Phi(a))$

Of course, we can use computers to calculate these probabilities as well. In `R`, we do:

```
pnorm(1.96, mean = 0, sd = 1)
```

```
## [1] 0.9750021
```

Some values of $\Phi(\cdot)$ worth remembering:

1. $\Phi(0) = 0.5$
2. $\Phi(1.64) \approx 0.95$
3. $\Phi(1.96) \approx 0.975$
4. $\Phi(2.33) \approx 0.99$
5. $\Phi(2.58) \approx 0.995$

These values often come up when doing confidence intervals or hypothesis-test type questions. The third one, for example, says that

$$\begin{aligned}\Pr(\mu - 1.96\sigma \leq X \leq \mu + 1.96\sigma) &= \Pr\left(\left|\frac{X-\mu}{\sigma}\right| \leq 1.96\right) \\ &= 2\Phi(1.96) - 1 \approx 0.95,\end{aligned}$$

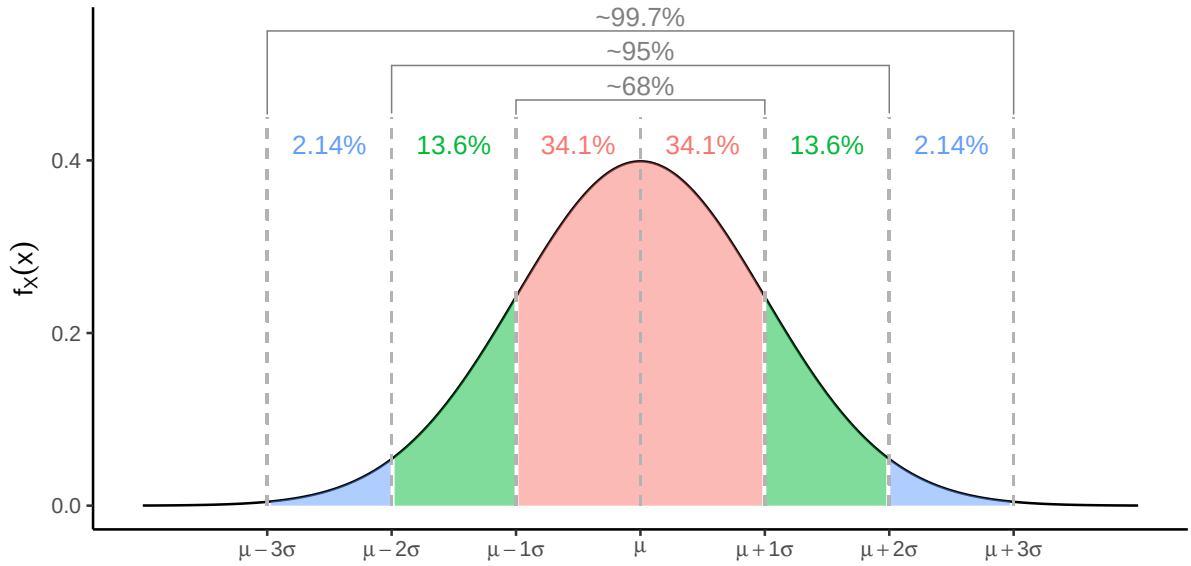
giving us an approximate probability statement as follows:

> The probability that a standardised normal random variable is within ± 1.96 in value is approximately 0.95.

[4] Download the statistical tables from here: https://haziqj.ml/stat-tables/

2.3.4 68–95–99.7 Rule

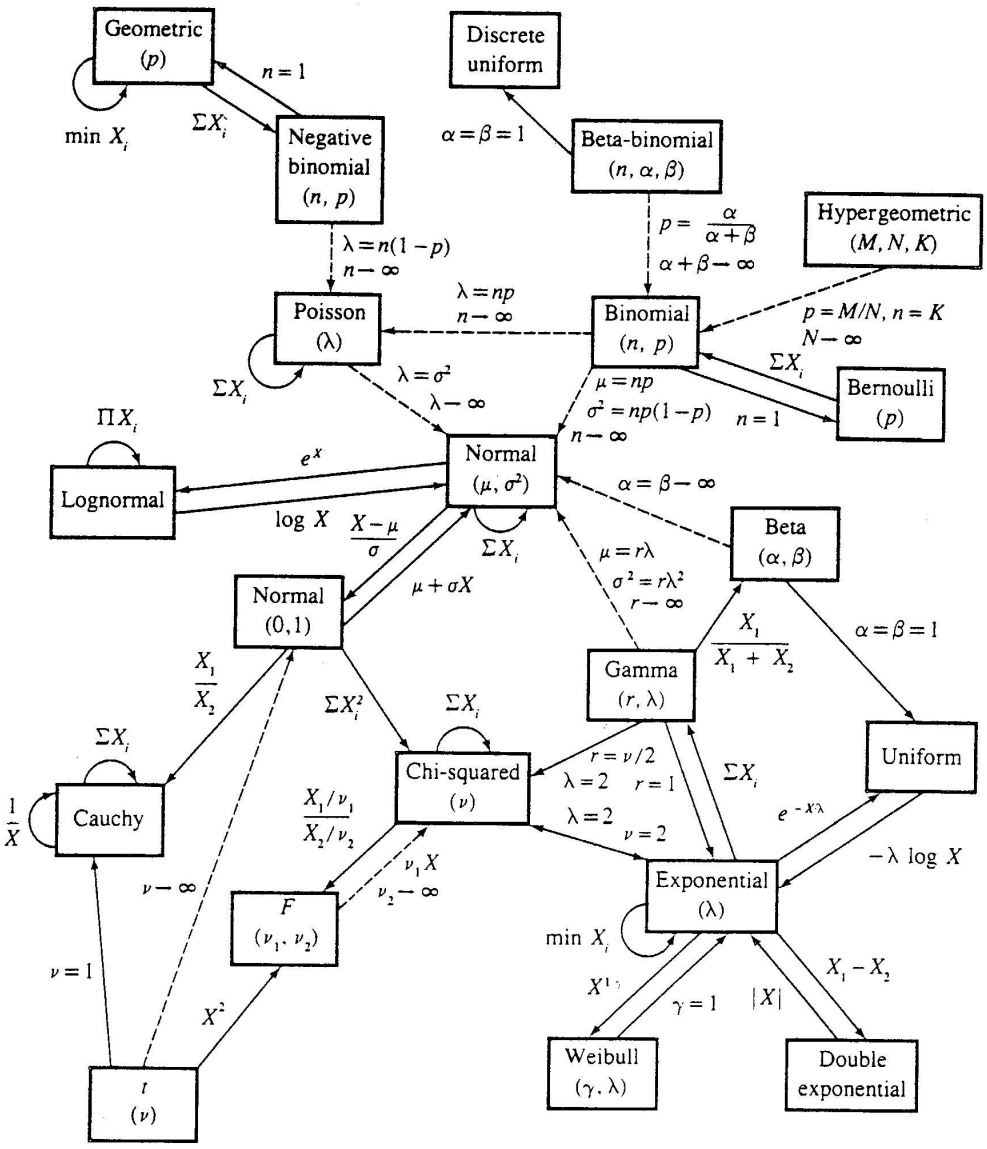
Incidentally, there is a shorthand to remember the percentage of values that lie within a band around the mean in a normal distribution.



2.4 Some relationships

There are several interesting relationships between the various distributions we have encountered. Making use of these relationships allow us to find an alternative and possibly easier way of calculating required probabilities or expectations, although some relationships might only be approximate. We discuss them below.

630 Table of Common Distributions



Relationships among common distributions. Solid lines represent transformations and special cases, dashed lines represent limits. Adapted from Leemis (1986).

2.4.1 Poisson-Binomial relationship

The Poisson distribution plays a useful approximation role for some of the other main discrete distributions. Let $X \sim \text{Bin}(n, p)$. Then

$$X \approx \text{Poi}(np)$$

when n is large and p is small. Typically the rule of thumb is $n > 20$ and $np < 5$ or $n(1-p) < 5$.

We'll attempt a kind of a sketch proof for this approximate relationship. Let $\lambda = np$. Consider the limit of as $n \to \infty$ of the binomial pmf:

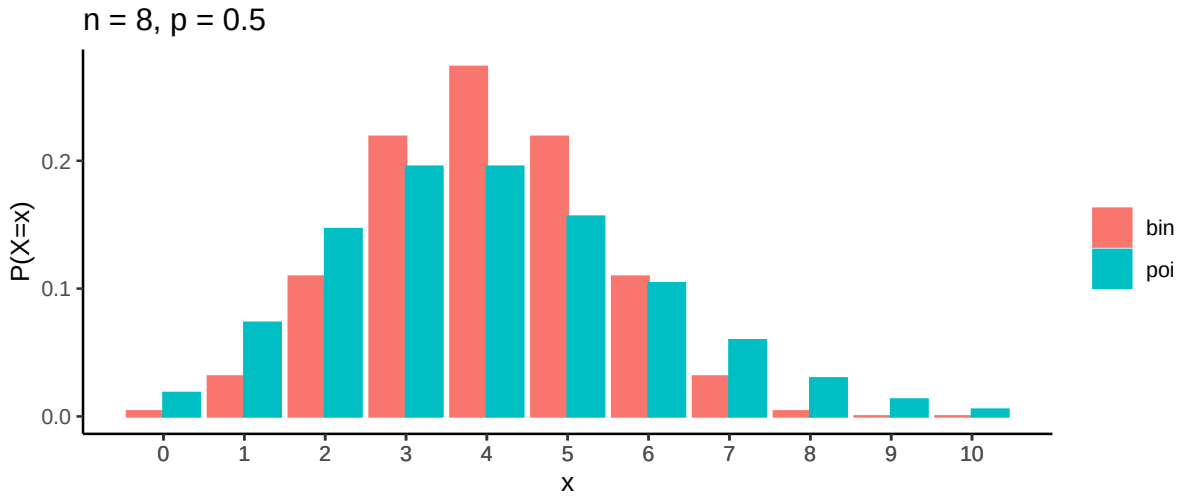


Figure 2.14: When n is small, then the two distributions are not quite identical.

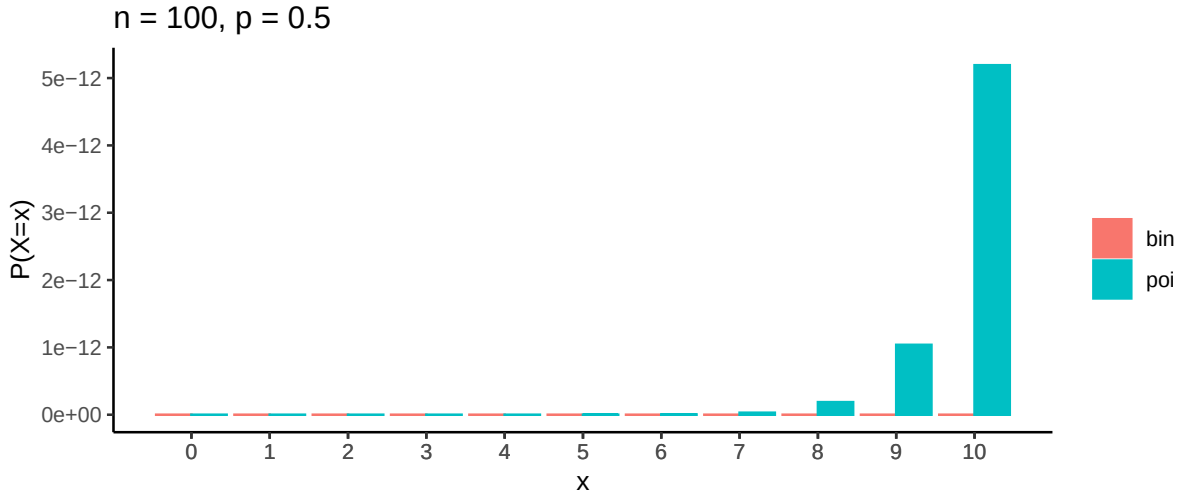


Figure 2.15: When n is large but p is not small, then the approximation does not work.

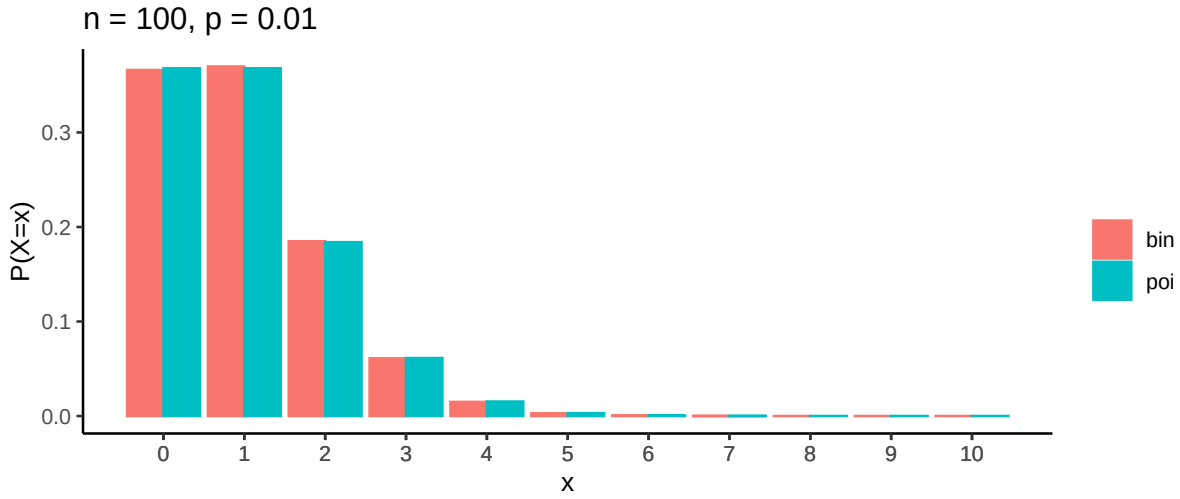


Figure 2.16: The two conditions (n large and p small) is required for the approximation to work.

$$\begin{aligned}
\lim_{n\to\infty} \Pr(X = x) &= \lim_{n\to\infty} \frac{n!}{x!(n-x)!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \\
&= \frac{\lambda^x}{x!} \lim_{n\to\infty} \underbrace{\frac{n!}{n^x(n-x)!}}_{\to 1} \underbrace{\left(1 - \frac{\lambda}{n}\right)^n}_{\to e^{-\lambda}} \underbrace{\left(1 - \frac{\lambda}{n}\right)^{-x}}_{\to 1} \\
&= \frac{e^{-\lambda}\lambda^x}{x!} \\
&= \Pr(Y = x), Y \sim \text{Poi}(\lambda).
\end{aligned}$$

In the above, the limit of the first time is 1 because:

$$\begin{aligned}
\frac{n!}{n^x(n-x)!} &= \frac{n(n-1)(n-2)\cdots 3 \cdot 2 \cdot 1}{n \cdot n \cdots n \cdot (n-x)(n-x-1)\cdots 3 \cdot 2 \cdot 1} \\
&= \frac{n}{n}\frac{n-1}{n} \cdots \frac{n-x+1}{n}
\end{aligned}$$

and each term converges to 1 as $n \to \infty$. Furthermore, the appearance of the exponential in the second term is through the very definition of exponentials, i.e.

$$e^a = \lim_{n\to\infty} \left(1 + \frac{a}{n}\right)^n.$$

2.4.2 Poisson-Exponential

We discussed previously how

- the Poisson distribution is used to model the number of occurences in a given unit of time (or space); and
- the exponential distribution is used to model the waiting time until occurence of an event.

The two are actually connected: The exponential distribution is the probability distribution of the time between events in a Poisson point process, i.e., a process in which events occur continuously and independently at a constant average rate.

To see how they are related, consider the following example. Let

- N_t be the number of phone calls during time period t; and
- X_t be the waiting time (minutes) until the next phone call from one at t.

N_t can be modelled using a Poisson distribution, and X_t by an exponential distribution. By definition, the two events are equivalent: $\{X_t > x\} \equiv \{N_t = N_{t+x}\}$. What this is saying is that the event that I have to wait more than x minutes for a phone call is the same as the event that there are no phone calls between time t and $t + x$, for any given t. If so, then

$$\Pr(X_t \leq x) = 1 - \Pr(N_t - N_{t+x} = 0).$$

Note that $\Pr(N_t - N_{t+x} = 0)$ is also the same as saying that there are no calls in x amount of time, $\Pr(N_x = 0)$. Assume that N_t is a Poisson process with rate λ per unit time t. So $N_x \sim \text{Poi}(\lambda x)$ and $\Pr(N_x = 0) = e^{-\lambda x}$. Substituting this into the above, we get

$$\Pr(X_t \leq x) = 1 - e^{-\lambda x}$$

which is the cdf of an $\text{Exp}(\lambda)$ distribution.

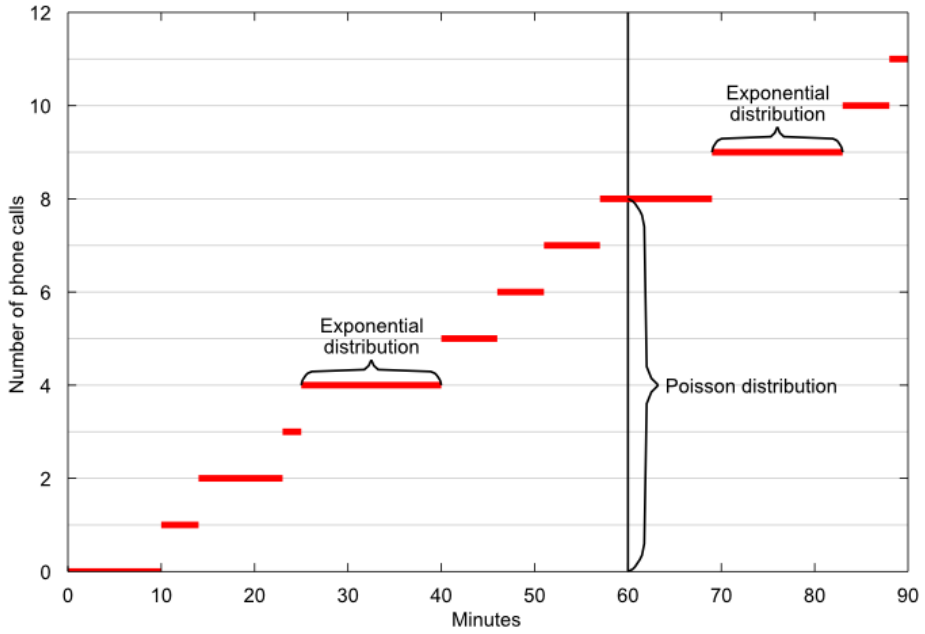


Figure 2.17: Poisson-exponential process.

2.4.3 Poisson-Gamma

Since the gamma distribution generalises the exponential, we also have a more general relationship between the Poisson and the gamma distribution. Both are closely related when the gamma shape parameter is an integer. Specifically, if $X \sim \Gamma(\alpha, \beta)$, then for any $x > 0$,

$$\Pr(X > x) = \Pr(Y < \alpha),$$

where $Y \sim \text{Poi}(x/\beta)$.

The special case for the exponential distribution is easily seen: Set $\alpha = 1$, then

$$\Pr(X > x) = \Pr(Y < 1) = \Pr(Y = 0) = e^{-x/\beta}.$$

So X has an $\text{Exp}(\beta)$ distribution (scale parameter β).

2.4.4 Normal approximations

The normal family can be used–largely on account of the Central Limit Theorem–to approximate various other distributions.

- $\text{Poi}(\lambda) \approx \text{N}(\lambda, \lambda)$, for large values of λ.
- $\text{Bin}(np) \approx \text{N}(np, np(1-p))$, for large n (and p not too close to 0 or 1).
- $\Gamma(\alpha, \beta) \approx \text{N}(\alpha\beta, \alpha\beta^2)$ for large values of α.

We will officially cover the central limit theorem in detail in the next chapter. For now, you may think of it as follows. Suppose that we're interested in the distribution of the sample mean (which, by now, you will agree is a random variable and hence has a distribution). The central limit theorem tells us precisely what the distribution of the sample mean will be when the number of samples we collect increases. It turns out to be the normal distribution!

When approximating a discrete distribution, the normal approximation is *much improved* by use of a 'continuity correction'.

Example 2.1. Let $X \sim \text{Bin}(25, 0.6)$. So $\text{E}(X) = 25 \times 0.6 = 15$ and $\text{Var}(X) = 25 \times 0.6 \times 0.4 = 6$. The normal approximation is $X \approx \text{N}(15, 6)$. A binomial probability such as

$$\Pr(X \leq 13) = \sum_{x=0}^{13} \binom{25}{x} 0.6^x 0.4^{25-x} = 0.267$$

can be approximated as

$$\Pr(X \leq 13) \approx \Pr\left(Z \leq \frac{13 - 15}{\sqrt{6}}\right) = 0.207, \quad Z \sim \text{N}(0, 1)$$

Evidently this is not a very good approximation. However, for discrete X, $\Pr(X \leq 13)$ and $\Pr(X \leq 13.5)$ are identical, and approximating the latter gives a better result:

$$\Pr(X \leq 13.5) \approx \Pr\left(Z \leq \frac{13.5 - 15}{\sqrt{6}}\right) = 0.270, \quad Z \sim \text{N}(0, 1).$$

```
pbinom(13, size = 25, prob = 0.6)
```

```
## [1] 0.2677178
```

```
pnorm(13.5, mean = 25 * 0.6, sd = sqrt(25 * 0.6 * 0.4))
```

```
## [1] 0.2701457
```

Apply these continuity corrections in your calculations!

Discrete	Continous
$X = c$	$c - 0.5 < X < c + 0.5$
$X < c$	$X < c + 0.5$
$X \leq c$	$X < c + 0.5$
$X > c$	$X > c - 0.5$
$X \geq c$	$X > c - 0.5$

2.5 Exercises

1. The flow of traffic at certain street corners can sometimes be modelled as a sequence of Bernoulli trials by assuming that the probability of a car passing during any given second is a constant p and that there is no interaction between the passing of cars at different seconds. If we treat seconds as indivisible time units (trials), the Bernoulli model applies. Suppose a pedestrian can cross the street only if no car is to pass during the next 3 seconds. Find the probability that the pedestrian has to wait for exactly 4 seconds before starting to cross.

2. A standard drug is known to be effective in 80% of cases. A new drug is tested on 100 patients and found to be effective in 85 cases. Evaluate this apparent evidence that the new drug is superior. *Hint: calculate, using a normal approximation, the probability of getting 85 or more successes if in fact the new and old drugs are equally effective.*

3. Suppose that the number of chocolate chips in a certain type of cookie follows a Poisson distribution, and that we want the proportion of cookies containing at least two chocolate chips to be greater than 0.99. Find the smallest value of the mean of the distribution that ensures this probability.

4. A truncated discrete distribution is one in which a particular outcome or set of outcomes cannot be observed and is eliminated from the sample space. In particular, if X has sample space $\{0, 1, 2, ...\}$ but 0 cannot be observed (e.g, X might be the size of a group making a booking for the theatre, or the number of vehicles involved in a road accident) the zero-truncated random variable X_T has pmf

$$\Pr(X_T = x) = \frac{\Pr(X = x)}{\Pr(X > 0)}$$

for $x = 1, 2, ...$ If $X \sim \text{Poi}(\lambda)$, find the pmf, mean and variance of X_T.

5. Show that

$$\int_x^\infty \frac{1}{\Gamma(\alpha)} z^{\alpha-1} e^{-z} \, \mathrm{d}z = \sum_{y=0}^{\alpha-1} \frac{x^y e^{-x}}{y!}, \qquad \alpha = 1, 2, 3, ...$$

Hint: Use integration by parts, and the fact that $\Gamma(n) = (n-1)!$ *for positive integers* n. Express this formula as a probabilistic relationship between Poisson and gamma random variables.

6. The *Pareto distribution*, with parameters α and β, has pdf

$$f(x) = \frac{\beta \alpha^\beta}{x^{\beta+1}}, \qquad 0 < \alpha < x < \infty, \quad \beta > 0.$$

   (a) Verify that $f(x)$ is a pdf.

   (b) Derive the mean and variance of this distribution.

   (c) Prove that the variance does not exist if $\beta \leq 2$.

7. Show, using the mgf $M_X(t) = \exp[\lambda(e^t - 1)]$, that if $X \sim \text{Poi}(\lambda)$ then $\text{E}(X) = \text{Var}(X) = \lambda$.

8. Let $N \sim \text{Poi}(\lambda)$ and suppose we toss a coin N times, where p is the probability that the coin lands heads up. Let X and Y be the number of heads and tails respectively. Show that X and Y are independent.

9. Let $X \sim \text{Poi}(\lambda)$ and $Y \sim \text{Poi}(\mu)$ and assume that X and Y are independent. Show that the distribution of X given that $X + Y = n$ is $\text{Bin}(n, \pi)$, where $\pi = \frac{\lambda}{\lambda+\mu}$. *Use the following hints:*

   - *If* $X \sim \text{Poi}(\lambda)$ *and* $Y \sim \text{Poi}(\mu)$*, and* X *and* Y *are independent, then* $X + Y \sim \text{Poi}(\lambda + \mu)$.
   - $\{X = x, X + Y = n\} = \{X = x, Y = n - x\}$.

10. Let $X_1, ..., X_n \overset{\text{iid}}{\sim} \text{Exp}(\beta)$. Let $Y = \max\{X_1, ..., X_n\}$. Find the pdf of Y. *Hint:* $Y \leq y$ *iff* $X_i \leq y$ *for* $i = 1, ..., n$.

11. In each of the following cases verify the expression given for the mgf, and in each case, use the mgf to calculate $\text{E}(X)$ and $\text{Var}(X)$.

   (a) $\Pr(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$, $M_X(t) = e^{\lambda(e^t-1)}$, $x = 0, 1, 2, ...$, $\lambda > 0$.

   (b) $\Pr(X = x) = p(1-p)^x$, $M_X(t) = \frac{p}{1-e^t(1-p)}$, $x = 0, 1, 2, ...$, $0 < p < 1$.

   (c) $f_X(x) = \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sigma\sqrt{2\pi}}$, $M_X(t) = e^{\mu t + \sigma^2 t^2/2}$, $x \in \mathbb{R}$, $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}_{>0}$.

12. Suppose the random variable T is the length of life of an object (possibly the lifetime of an electrical component or of a subject given a particular treatment). The *hazard function* $h_T(t)$ associated with the random variable T is defined by

$$h_T(t) = \lim_{\delta \to 0} \frac{\Pr(t \leq T < t + \delta | T \geq t)}{\delta}.$$

It is meant to denote the "event rate at time t, conditional on survival until time t or later". Thus, we can interpret $h_T(t)$ as the rate of chance of the probability that the object survives a little past time t, given that the object survives to time t. Show that if T is a continuous random variable, then

$$h_T(t) = \frac{f_T(t)}{1 - F_T(t)} = -\frac{\mathrm{d}}{\mathrm{d}t} \log(1 - F_T(t)).$$

Hints:

- *Use the definition of conditional probability.*
- *The derivative of a function g at x is defined as*

$$g'(x) = \lim_{\delta \to 0} \frac{g(x+\delta) - g(x)}{\delta}.$$

- *The derivative of the cdf is the pdf.*

13. Evidence concerning the guilt or innocence of a defendant in a criminal investigation can be summarized by the value of an exponential random variable X whose mean μ depends on whether the defendant is guilty. If innocent, $\mu = 1$; if guilty, $\mu = 2$. The deciding judge will rule the defendant guilty if $X > c$ for some suitably chosen value of c.

    (a) If the judge wants to be 95 percent certain that an innocent man will not be convicted, what should be the value of c?

    (b) Using the value of c found in part a., what is the probability that a guilty defendant will be convicted?

14. An image is partitioned into two regions, one white and the other black. A reading taken from a randomly chosen point in the white section will give a reading that is normally distributed with $\mu = 4$ and $\sigma^2 = 4$, whereas one taken from a randomly chosen point in the black region will have a normally distributed reading with parameters $(6, 9)$. A point is randomly chosen on the image and has a reading of 5. If the fraction of the image that is black is α, for what value of α would the probability of making an error be the same, regardless of whether one concluded that the point was in the black region or in the white region? *Hint: For $X \sim \mathrm{N}(\mu, \sigma^2)$, express $\Pr(X = c)$ as $\lim_{\epsilon \to 0} \Pr(c < X < X + \epsilon)$. Then use the definition of derivatives and the fact that $F'(x) = f(x)$.*

Hand-in questions

1. Let $M_X(t)$ be the mgf of X, and define $S(t) = \log M_X(t)$. Show that $S'(0) = \mathrm{E}(X)$ and $S''(0) = \mathrm{Var}(X)$. *Remark: $S(t)$ is called the cumulant-generating function of X.* **[3 marks]**

2. A model for the movement of a stock supposes that if the present price of the stock is s, then, after one period, it will be either us with probability p or ds with probability $1 - p$. Assuming that successive movements are independent, approximate the probability that the stock's price will be up at least 30 percent after the next 1000 periods if $u = 1.012$, $d = 0.990$, and $p = 0.52$. **[5 marks]**

3. Abu goes fishing every Sunday. The number of fish he catches follows a Poisson distribution. On a proportion π of the days he goes fishing, he does not catch anything. He makes it a rule to take home the first and then every other fish that he catches (i.e. the first, third, fifth, and so on).

    (a) Using a Poisson distribution, find the mean number of fish he catches. **[3 marks]**

    (b) Show that the probability that he takes home the last fish he catches is $(1 - \pi^2)/2$. *Hint: Use the fact that* $\sum_{k=0}^{\infty} \frac{e^{-\lambda} \lambda^{2k+1}}{(2k+1)!} = (1 - e^{-2\lambda})/2$. **[3 marks]**

4. There are two types of batteries in a bin. When in use, type i batteries last (in hours) an exponentially distributed time with rate λ_i, $i = 1, 2$. A battery that is randomly chosen from the bin will be a type i battery with probability p_i, $p_1 + p_2 = 1$. If a randomly chosen battery is still operating after t hours of use, what is the probability it will still be operating after an additional s hours? **[3 marks]**

Chapter 3

Inequalities, convergences, and normal random samples

In this chapter, we lay more advanced groundwork for the upcoming inference chapters. Here, we will take a look at the concept of *convergence* as applied to random variables. You would have come across convergence of sequences before, but when that sequence is random, does "convergence" even have a sensible meaning? It turns out we can define such concepts even for random variables. Establishing convergence concepts allows us to also establish very important limit theorems used frequently in statistics.

Also included in this chapter is a study of the distributions derived from normal random samples. As the normal distributions is used all the time in inferential statistics, the properties of normal random samples has been studied extensively. In particular, normal random samples give rise to the χ^2, t and F distributions–all of which I'm sure you have heard of in the context of statistical testing. Here, we will take a look in detail as to how they are derived, and how we can use them when calculating probabilities of interest.

Learning objectives

By the end of this chapter, you will be able to:

- Apply probability inequalities and inequalities based on expectations to compute approximate bounds and in proving other mathematical theorems.
- Understand the notion of three types of convergences as applied to random variables, namely convergence in probability, convergence in distribution, and mean-square convergence.
- Be familiar with two important limit theorems in statistics: The (weak) Law of Large Numbers; and the Central Limit Theorem.
- Use Slutzky's theorem and the delta method in probability-based calculations and inference problems.
- Define and use the distributions derived from normal random samples: χ^2, Student's t, and F distributions.

Readings

- Casella and Berger (2002)
  - Chapter 5, sections 5.1–5.3 and 5.5.
- Wasserman (2004)
  - All of chapter 4.
  - All of chapter 5.
- Topics not covered here: Order statistics, almost-sure convergence, consistency (will be covered in Part 4), strong LLN, multivariate delta method, Hoeffding's inequality, Mill's inequality,

3.1 Introduction

Statistical inference involves data points, which as we discussed previously, displays a duality between being information and also random variables. Typically, collecting data $X_1, \dots, X_n$ in an experiment means recording several observations on a a particular variable of interest X. For example.

- Time to failure for n identical circuit boards.
- Yield (in tonnes) of n seasonal harvest for *Laila* variety paddy.
- Voter preferences for n individuals in the US.

We can **model** this mathematically by declaring $X = (X_1, \dots, X_n)^\top$ to be random variables sampled from a population whose pdf or pmf is $f_X(x)$. We typically write

$$(X_1, \dots, X_n)^\top \overset{\text{iid}}{\sim} f_X.$$

There are two points to note here:

1. We have indeed treated all observations X as being sampled from a multivariate distribution. In general this is the proper way of doing it, unless we impose a further assumption of independence and identical distributions (see next section).
2. In most practical situations, we usually collect more than one *kind* of data, e.g. demographic data for study participants age, weight, height, etc. This is called *multivariable data*. We will focus on *univariate* statistics in this course for the most part.

3.1.1 Independent and identical random variable

In an experiment, the samples are usually taken in such a way

- that the value of one observation has no effect on or relationship with any of the other observations (i.e. $X_1, \dots, X_n$ are independent); and
- the pdf/pmf of each observation is $f(x)$ (i.e. identical).

In this case (remember definition of independence for pdfs),

$$f_X(x_1, \dots, x_n) = \prod_{i=1}^{n} f(x_i)$$

In particular, if the population pdf/pmf is a member of a *parametric family*, say one of those introduced in Chapter 2, then we can write

$$f_X(x_1, \dots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

This is because identical distribution assumption would mean each of the random variables is assumed to come from the same family of distribution with the same parameter. We could then use the random samples to *infer* about the (unknown) parameter θ. More on this in the next chapter.

Side note: Finite population sampling

We have just defined sampling from an *infinite* population. Sometimes, sampling is done from a *finite* population, that is, the population consists only of possible observations $\{x_1, \dots, x_N\}$.

There are several approaches to this which may or may not yield independent samples:

- sampling with vs without replacement
- simple random sampling vs complex random sampling
- single-stage sampling vs multi-stage sampling
- etc.

Very important topic in **survey methodology**. For more details see C&B §5.1, as well as 2019 lecture slides (Chapter 2). In this course, we deal only with the infinite population model.

3.1.2 Statistic

Throughout the course we will refer to an object called a "statistic". Note the singular use of the word.

Definition 3.1 (Statistic)**.** A statistic is any function $T_n = T(X_1, \dots, X_n)$. It cannot depend on unknown parameters, only on observables.

In essence, a statistic is a manipulation of information, usually in a way to condence the information from samples. Suppose $X_1, \dots, X_n$ is a random sample with mean μ and variance σ^2. Here are some examples of statistics (the first two are really common statistics)

- The sample mean
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$
- The (unbiased) sample variance
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$
- The maximum of the sample
$$\max(X_1, \dots, X_n)$$
- Any value of the sample itself
$$X_i$$
These however, are **not** statistics, because they involve parameter values of the underlying probability distribution.
- The quantity
$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$
- Any expectation
$$\mathrm{E}(g(X)) = \int g(x) f(x|\theta)\,\mathrm{d}x$$

Practically speaking, we can calculate statistics if we have the data points. But we cannot calculate non-statistics because we have to ask ourselves "what is the value of θ" in order to proceed!

Since the sample mean and sample variance are commonly used statistics, it's important to know the following properties:

Lemma 3.1. *Let $X_1, \dots, X_n$ be a random variable with mean μ and variance $\sigma^2 < \infty$. Then*

- $\mathrm{E}(\bar{X}) = \mu$.
- $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$.
- $\mathrm{E}(S^2) = \sigma^2$.

The proof of this can be found in C&B, Theorem 5.2.6.

Do try out the proof of this lemma by yourself. The first two should be doable by now, but perhaps the last one is a little bit challenging.

3.1.3 Sampling distribution

Realise that

- A statistic T_n is itself a random variable.
- If it is random, it has a distribution, even if we might not know what it is.
- Along with having a distribution, all of concepts and properties we discussed in Chapters 1 & 2 apply.

Think about the statement above, "a statistic T_n is itself a random variable"–can you rationalise why this is? Suppose you collect some data and plug these values into a statistics function (e.g. the sample mean). Will the value of the sample mean be the same each time, or will it depend on the (random) values of the data?

A very common theme in inferential statistics is to figure out what the distribution of T_n is in repeated sampling. The reason for this is in parameteric statistics, the statistic T_n may serve as an *estimator* for the true unknown value θ. But since we've established that T_n itself is a random variable, and different samples will give different values, how do we convince ourselves that T_n is actually a *good* estimator?

One way perhaps is to consider the values of T_n under *repeated sampling*, and whether its value is "typically close" to the true value θ? In the next chapter especially, we will take a look at this problem in detail.

3.1.4 Large-sample approximation

Some statistics have easily-derived sampling distribution; others do not. For instance, suppose each $X_i \sim \mathrm{N}(\mu, \sigma^2)$. Then it is well known that

$$\bar{X} \sim \mathrm{N}(\mu, \sigma^2/n). \tag{3.1}$$

The above fact (3.1) is very important and pops up all the time in statistics. Have a go at proving the distribution of the sample mean.

Generally speaking statistics derived from normal random samples have 'easy' distributions (we'll see this later). But what is the distribution of

$$n^{-1} \sum_{i=1}^{n} \tan^{-1}(X_i)?$$

If being exact is difficult, maybe we can compromise by using approximate distributions, which can be found by using *asymptotic* arguments. That is, we consider the behaviour of the distribution of the complicated statistics T_n as $n \to \infty$. This is what we often refer to as its large-sample approximation.

But of course, we have to take care of the mathematics because when dealing with infinite quantities, a lot of things can go wrong. What we need to establish is that the limiting behaviour of our statistics of interest actually do what we expect them to do. For this, we need to study inequalities and convergences.

3.2 Inequalities

Inequalities are useful tools in establishing various properties of statistical inference methods. They may also provide estimates for probabilities with little assumption on probability distributions.

There are four main inequalities that we will learn:

- Markov's inequality
- Chebyshev's inequality
- Cauchy-Schwarz inequality
- Jensen's inequality

3.2.1 Markov's inequality

In probability theory, Markov's inequality gives an upper bound for the probability that a *non-negative* random variable exceeds some positive constant.

Lemma 3.2 (Markov's inequality)**.** *Let $X \geq 0$ be a non-negative random variable and $\mathrm{E}(X) < \infty$. Then, for any $t > 0$,*

$$\Pr(X \geq t) \leq \frac{\mathrm{E}(X)}{t}.$$

Markov's inequality relate probabilities to expectations, and provides bounds for the cumulative distribution function of a random variable.

Proof. Let $f(x)$ be the pdf of X. Since $X \geq 0$,

$$\begin{aligned}
\mathrm{E}(X) = \int_{-\infty}^{\infty} x\, f(x)\, \mathrm{d}x &= \int_{0}^{\infty} x\, f(x)\, \mathrm{d}x \\
&= \int_{0}^{t} x\, f(x)\, \mathrm{d}x + \int_{t}^{\infty} x\, f(x)\, \mathrm{d}x \\
&\geq \int_{t}^{\infty} x\, f(x)\, \mathrm{d}x \\
&\geq t \int_{t}^{\infty} f(x)\, \mathrm{d}x \\
&= t \Pr(X \geq t)
\end{aligned}$$

□

Corollary 3.1. *For any random variable X and any constant $t > 0$,*

$$\Pr(|X| \geq t) \leq \frac{E|X|}{t} \quad \textit{provided } E|X| < \infty$$
$$\Pr(|X| \geq t) \leq \frac{E(|X|^k)}{t^k} \quad \textit{provided } E(|X|^k) < \infty$$

Note that there is no mistake in the corollary statement above (if you were wondering about whether there should be a power of k in the probability statement).

Can you prove the corollary to Markov's inequality? The steps are similar to the proof of the original Markov's inequality.

The tail probability $\Pr(|X| \geq t)$ is a useful measure in insurance and risk management in finance. The more moments X has, the smaller the tail probabilities are.

3.2.2 Chebyshev's inequality

In probability theory, Chebyshev's inequality guarantees that no more than a certain fraction of values can be more than a certain distance from the mean.

Lemma 3.3 (Chebyshev's inequality)**.** *Suppose a random variable X has mean μ and variance σ^2. Then, for any $t > 0$,*

$$\Pr(|X - \mu| \geq t\sigma) \leq \frac{1}{t^2}.$$

The proof of Chebyshev's inequality follows directly from Markov's inequality. You will prove this in the exercises.

Because it can be applied to completely arbitrary distributions (provided they have a known finite mean and variance), the inequality generally gives a poor bound, compared to what might be deduced if more aspects are known about the distribution involved.

Note that

$$\begin{aligned}\Pr(|X-\mu| \geq t\sigma) &= \Pr\left(\{X \leq \mu - t\sigma\} \cup \{X \geq \mu + t\sigma\}\right) \\ &= 1 - \Pr(\mu - t\sigma \leq X \leq \mu + t\sigma) \\ &= 1 - \Pr\left(|X-\mu| \leq t\sigma\right)\end{aligned}$$

Example 3.1. Suppose X has mean 0 and variance 1. By Chebyshev's inequality,

$$\begin{aligned}\Pr(|X| \geq 1) &\leq 1.00 \\ \Pr(|X| \geq 2) &\leq 0.25 \\ \Pr(|X| \geq 3) &\leq 0.11\end{aligned}$$

In contrast, suppose that we know that X is normally distributed. Then

$$\begin{aligned}\Pr(|X| \geq 1) &\leq 0.318 \\ \Pr(|X| \geq 2) &\leq 0.046 \\ \Pr(|X| \geq 3) &\leq 0.003\end{aligned}$$

Recall the 68-95-99.7 rule when we discussed the normal distribution in Chapter2.

Calculate the above probabilities in `R`:

```
2 * (pnorm(-c(1, 2, 3)))
```

```
## [1] 0.317310508 0.045500264 0.002699796
```

3.2.3 Cauchy-Schwartz inequality

This is a very useful inequality that crops up in many different areas of mathematics, such as linear algebra, analysis, probability theory, vector algebra, etc.

Lemma 3.4 (Cauchy-Schwartz inequality)**.** *Let $E(X^2) < \infty$ and $E(Y^2) < \infty$. Then*

$$|E(XY)|^2 \leq E(X^2)E(Y^2).$$

Subtle point: $|E(XY)|^2 = \mathrm{E}^2(XY)$.

Proof. Consider the expectation $\mathrm{E}((tX+Y)^2) \geq 0$ for some constant $t \in \mathbb{R}$. Expanding out, we have

$$\mathrm{E}((tX+Y)^2) = \overbrace{\mathrm{E}(X^2)}^{a} t^2 + \overbrace{2\,\mathrm{E}(XY)}^{b} t + \overbrace{\mathrm{E}(Y^2)}^{c}$$

For some constants $a, b, c \in \mathbb{R}$, the polynomial $at^2 + bt + c$ remains non-negative iff $a \geq 0$ and the discriminant $b^2 - 4ac \leq 0$. Thus,

$$4\,\mathrm{E}^2(XY) - 4\,\mathrm{E}(X^2)\,\mathrm{E}(Y^2) \leq 0,$$

and dividing by 4 throughout, we have the desired result. □

As a consequence of the Cauchy-Schwartz inequality, we have the covariance inequality.

Corollary 3.2 (The covariance inequality)**.** *Let X and Y be random variables. Then*

$$\mathrm{Var}(Y) \geq \frac{\mathrm{Cov}(Y,X)\,\mathrm{Cov}(Y,X)}{\mathrm{Var}(X)}$$

You will prove the covariance inequality in one of the exercises at the end of this chapter.

3.2.4 Jensen's inequality

Before discussing the next kind of inequality, we shall first discuss convex functions.

Definition 3.2.

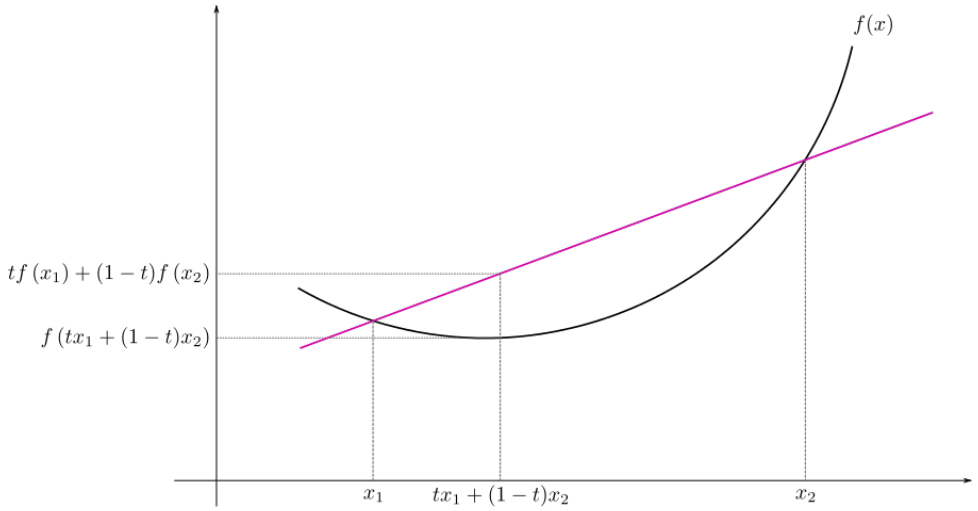
- A function g is **convex** if for any x, y and any $\alpha \in [0,1]$,

$$g(\alpha x + (1-\alpha)y) \leq \alpha g(x) + (1-\alpha)g(y).$$

- If $g''(x) > 0$ for all x, then g is convex.
- A function g is **concave** if $-g$ is convex.

Example 3.2. Examples of convex functions: $g_1(x) = x^2$ and $g_2(x) = e^x$, since $g_1''(x) = 2 > 0$ and $g_2(x) = e^x > 0$ for all x.

Examples of concave functions: $g_3(x) = -x^2$ and $g_4(x) = \log(x)$.



In the context of probability theory, we consider expectations of *convex* functions of random variables

Lemma 3.5 (Jensen's inequality)**.** *Let X be a random variable and g a convex function. Then,*

$$\mathrm{E}\left[g(X)\right] \geq g\left(\mathrm{E}\, X\right)$$

It follows directly from Jensen's inequality, the following:

- $\mathrm{E}(X^2) \geq \{\mathrm{E}(X)\}^2$
- $\mathrm{E}(1/X) \geq 1/\mathrm{E}\, X$
- $\mathrm{E}(\log X) \geq \log(\mathrm{E}\, X)$

3.3 Convergence of random variables

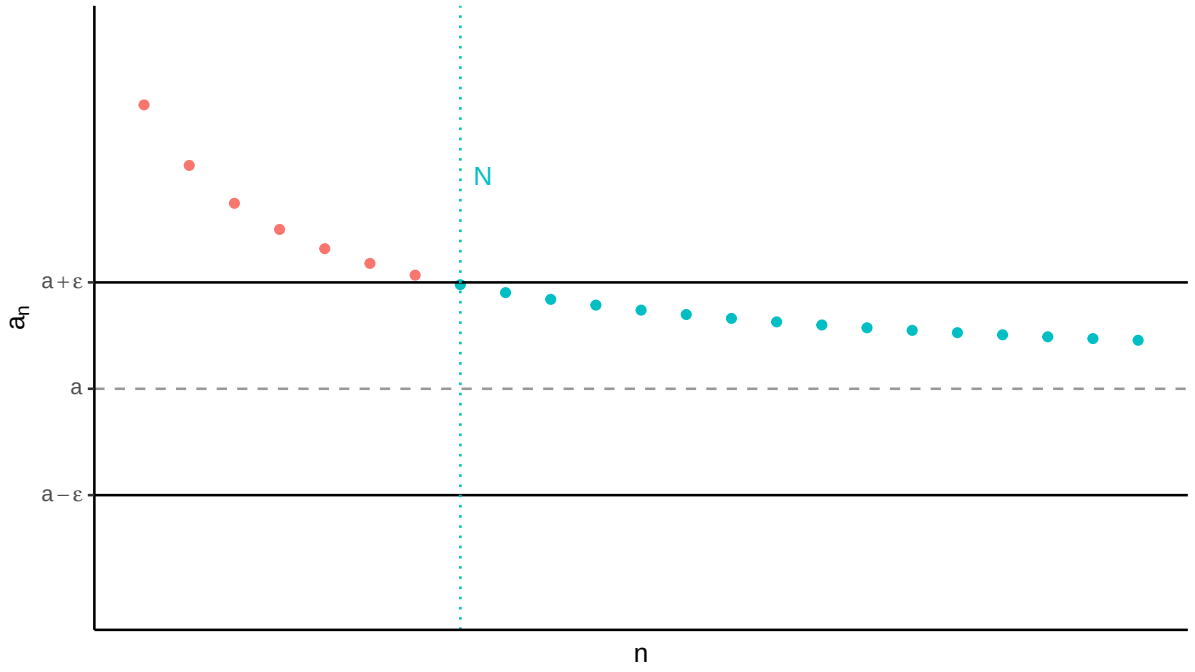
Recall the limits of sequences of real numbers (a_n), $n \in \mathbb{N}$.

Definition 3.3 (Limit of a real sequence)**.** We call a the limit of the real sequence (a_n) if for each real number $\epsilon > 0$, $\exists$ a natural number $N(\epsilon) \in \mathbb{N}$ such that, for every natural number $n \geq N$, we have $|a_n - a| < \epsilon$.

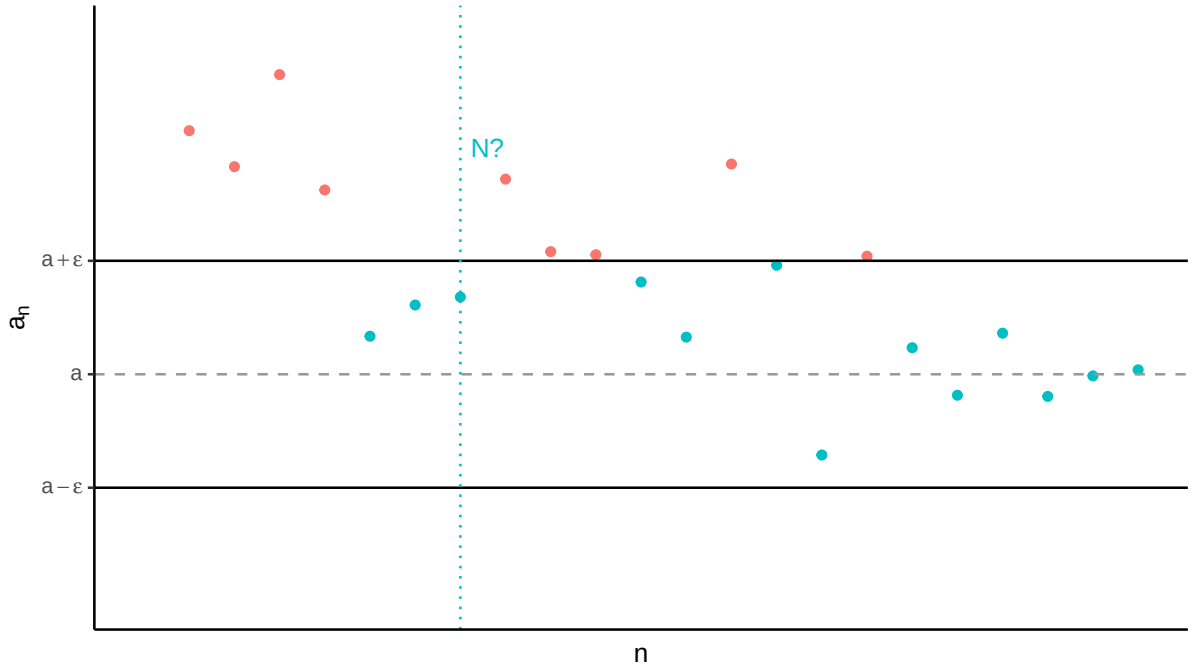
We write $\lim_{n\to\infty} a_n = a$, or simply $a_n \to a$. This also means that $|a_n - a| \to 0$ as $n \to \infty$. For every measure of closeness ϵ, the sequence's terms are eventually that close to the limit.

Some examples:

- If $a_n = c$ for some constant $c \in \mathbb{R}$, then $a_n \to c$.
- If $a_n = 1/n$, then $a_n \to 0$.
- $\lim_{n\to\infty}(1 + 1/n)^n = e$.



What if (a_n) is a random sequence (i.e. their values are not deterministic)? Does the concept of limits even make sense? Is it possible to "trap" the sequence between an upper and lower bound as the sequence progresses? This is what we will be exploring in this section.



We can in fact say similar things about sequences of **random variables**, e.g. X is the limit of a sequence (X_n) if $|X_n - X| \to 0$ as $n \to \infty$. There are some subtle issues here:

1. $|X_n - X|$ itself is a random variable, i.e. it takes difference values in the sample space Ω. Therefore, $|X_n - X| \to 0$ should hold (almost) entirely on the sample space. This calls for a probability statement.
2. Since random variable have distributions, we may also consider convergence of their distributions $F_{X_n}(x) \to F_X(x)$ for all x.

We need better tools to rigorously discuss the concept of convergence of random variables. Let $X_1, X_2, ...$ be a sequence of random variable, and X be another random variable. The main types of convergence for random variable that we will study are as follows:

1. Convergence in probability
2. Convergence in distribution
3. Convergence in mean-square

3.3.1 Convergence in probability

This is probably the most intuitive concept of what we would expect for a random variable to converge.

Definition 3.4 (Convergence in probability)**.** X_n converges to X in probability if for any constant $\epsilon > 0$,

$$\lim_{n\to\infty} \Pr(|X_n - X| \geq \epsilon) = 0.$$

We write $X_n \xrightarrow{\text{P}} X$, or $\text{plim}_{n\to\infty} X_n = X$.

An equivalent definition is

$$\lim_{n\to\infty} \Pr(|X_n - X| < \epsilon) = 1.$$

In words: "the probability of an 'unusual' outcome becomes smaller and smaller as the sequence progresses". Here, X may be a random variable or a constant.

Example 3.3. Let $X_i \overset{\text{iid}}{\sim} \text{Unif}(0, 1)$. Define another random variable $Y_n = \min\{X_1, ..., X_n\}$. Does Y_n converge to something? Draw some samples:

```
set.seed(123)
X <- runif(20); Y <- rep(NA, 20)
for (i in 1:20) Y[i] <- min(X[1:i])
round(X, 2)
```

```
##  [1] 0.29 0.79 0.41 0.88 0.94 0.05 0.53 0.89 0.55 0.46 0.96 0.45 0.68 0.57 0.10
## [16] 0.90 0.25 0.04 0.33 0.95
```

```
round(Y, 2)
```

```
##  [1] 0.29 0.29 0.29 0.29 0.29 0.05 0.05 0.05 0.05 0.05 0.05 0.05 0.05 0.05 0.05
## [16] 0.05 0.05 0.04 0.04 0.04
```

A good guess with be $Y_n \to 0$, so let's prove this. We want to show that

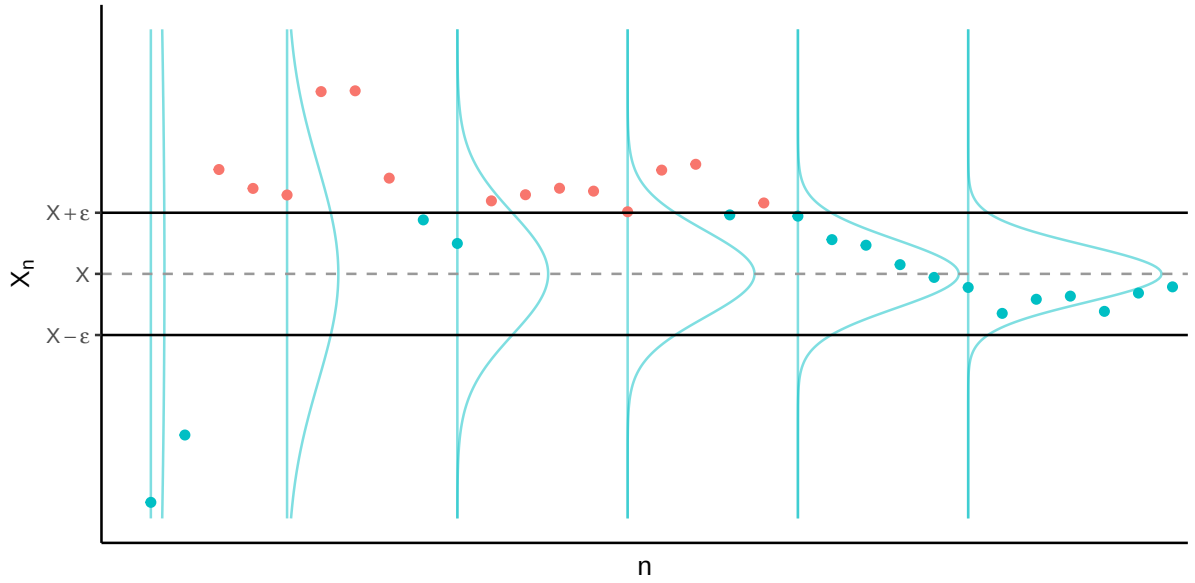
$$\Pr(|Y_n - 0| \geq \epsilon) = \Pr(Y_n \geq \epsilon) \to 0$$

as $n \to \infty$. There are two cases, i) $\epsilon > 1$ or ii) $\epsilon \leq 1$. If i), then $\Pr(Y_n \geq \epsilon) = 0$ and we are done. However, if $\epsilon \leq 1$, then

$$\begin{aligned}
\Pr(Y_n \geq \epsilon) &= \Pr(\min\{X_1, ..., X_n\} \geq \epsilon) \\
&= \Pr(X_1 \geq \epsilon, ..., X_n \geq \epsilon) \\
&= \Pr(X_1 \geq \epsilon) \cdots \Pr(X_n \geq \epsilon) \text{ by independence} \\
&= (1 - \epsilon)^n \to 0
\end{aligned}$$

as $n \to \infty$. Hence $Y_n \xrightarrow{\text{P}} 0$.

Here's another way of thinking about convergence in probability. Consider a random variable X_n converging to X as $n \to \infty$. Suppose at every point n we can draw a probability distribution curve of the random variable X_n. For small values of n, the curve might look flat, since very early on it's not converging yet. But as the values of n increases, we can steadily see that the curve becomes more concentrated around the value of X. Eventually, it will be so concentrated around X that the probability of X_n taking values outside a certain band ($\pm\epsilon$) will be virtually zero.



3.3.2 Convergence in distribution

Instead of considering the convergence of the random variable itself, we consider the convergence of the *distribution* of the sequence of random variables.

Definition 3.5 (Convergence in distribution)**.** X_n converges to X in distribution if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x).$$

We write $X_n \xrightarrow{\text{D}} X$.

- Again here X may be a constant, since a constant is a random variable with probability mass concentrated on a single point.
- We can also write $X_n \xrightarrow{\text{D}} F_X$, where F_X is the cdf of X. However, the notation $X_n \xrightarrow{\text{P}} F_X$ does not make sense!

Convergence in probability implies convergence in distribution, but not the other way around (**unless** the limiting random variable is a point mass).

Example 3.4. Let $X \sim \text{N}(0,1)$ and $X_n = -X$ for all $n \geq 1$. Then, clearly $F_{X_n} \equiv F_X$ (by linearity of normal distributions). Hence, $X_n \xrightarrow{\text{D}} X$.

However, X_n does not converge in probability to X, as for any $\epsilon > 0$,

$$\begin{aligned}\Pr(|X_n - X| \geq \epsilon) &= \Pr(2|X| \geq \epsilon) \\ &= \Pr(|X| \geq \epsilon/2) > 0.\end{aligned}$$

So we cannot have that $\Pr(|X_n - X| \geq \epsilon) \to 0$ as $n \to \infty$.

3.3.3 Mean-square convergence

In most practical situations, proving convergence in probability or distribution can be quite tough. It is therefore more convenient to consider the mean-square convergence.

Definition 3.6 (Mean-square convergence)**.** X_n converges in mean-square to X if

$$\lim_{n\to\infty} \mathrm{E}\left[(X_n - X)^2\right] = 0.$$

We write $X_n \xrightarrow{\text{m.s.}} X$.

It follows that from Markov's inequality,

$$\begin{aligned}\Pr(|X_n - X| \geq \epsilon) &= \Pr(|X_n - X|^2 \geq \epsilon^2) \\ &\leq \frac{\mathrm{E}\left[(X_n - X)^2\right]}{\epsilon^2}\end{aligned}$$

Therefore, if $X_n \xrightarrow{\text{m.s.}} X$, it also holds that $X_n \xrightarrow{\text{P}} X$.

Convergence in mean-square implies convergence in probability, but not the other way around.

Example 3.5. Let

$$X_n = \begin{cases} n^2 & \text{w.p. } 1/n \\ 0 & \text{w.p. } 1 - 1/n \end{cases}$$

Then, for any $\epsilon > 0$, $\Pr(|X_n| \geq \epsilon) = \Pr(X_n = n^2) = 1/n \to 0$ as $n \to \infty$. Hence, $X_n \xrightarrow{\text{P}} 0$.

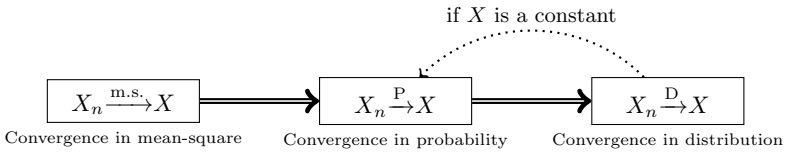
However,

$$\mathrm{E}(X_n^2) = n^2 \cdot \Pr(X_n = n^2) + 0 \cdot \Pr(X_n = 0) = n \to \infty$$

hence X_n does not converge in mean square to 0.

3.3.4 Relationship between convergences

The following diagram ties the three notions of convergences for random variables together.



You can find the proof of the above statements in Wasserman (Theorem 5.4). The proof is fairly easy to follow but for brevity will not be repeated here.

As we saw previously,

- Convergence in distribution does not imply convergence in probability.
- Convergence in probability does not imply convergence in mean-square.
- If $X_n \xrightarrow{\text{D}} c \in \mathbb{R}$ then $X_n \xrightarrow{\text{P}} c$.

It is typically easier to prove convergence in mean-square, which thus also implies convergence in probability and in distribution.

3.3.5 Slutzky's Theorem

On another practical note, once we figure out that $X_n \to X$ in probability say, we are usually also interested in whether functions of X_n also converge. For instance, suppose we calculate two sample means $\bar{X}_n$ and $\bar{Y}_n$ from the same experiment. Can we combine the two? I.e., does $\frac{1}{2}(\bar{X}_n + \bar{Y}_n)$ converge meaningfully?

Instead of trying to prove "manually" convergence for new functions of random variables, we have Slutzky's theorem to the rescue!

Theorem 3.1 (Slutzky's Theorem)**.** *Let X_n, Y_n, X, and Y be random variable, g a continuous function, and c a real constant. Then,*

- *If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then*
  - $X_n + Y_n \xrightarrow{P} X + Y$;
  - $X_n Y_n \xrightarrow{P} XY$; *and*
  - $g(X_n) \xrightarrow{P} g(X)$.
- *If $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{D} c$, then*
  - $X_n + Y_n \xrightarrow{D} X + c$;
  - $X_n Y_n \xrightarrow{D} cX$; *and*
  - $g(X_n) \xrightarrow{D} g(X)$.

Caution! If $X_n \xrightarrow{\text{D}} X$ and $Y_n \xrightarrow{\text{D}} Y$, it does **not** in general imply that $X_n + Y_n \xrightarrow{\text{D}} X + Y$.

Unfortunately the proof is not straightforward, so we shall skip it for now. Interested readers can definitely look up the proof in the suggested textbooks.

Example 3.6. Consider two sequences:

- $X_n = X$ where $X \sim \text{N}(0, 1$; and
- $Y_n = 2 + e^{-n}$.

Evidently $X_n \to \text{N}(0, 1)$ in distribution as $n \to \infty$, since all the values of X_n are drawn from a standard normal. Further, Y_n itself is not a random variable–it is a regular sequence, and we can clearly see that $Y_n \to 2$ as $n \to \infty$. Technically, we can still say $Y_n \xrightarrow{\text{D}} 2$, so Slutzky's theorem applies:

$$Z_n = X_n Y_n \xrightarrow{\text{D}} 2 \times \text{N}(0, 1) \equiv \text{N}(0, 4).$$

Similarly,

$$W_n = X_n + Y_n \xrightarrow{\text{D}} 2 + \text{N}(0, 1) \equiv \text{N}(2, 1).$$

3.4 Limit theorems

In this section, we'll cover two very important theorems in statistics: The Law of Large Numbers and The Central Limit Theorem.

3.4.1 The (weak) Law of Large Numbers

Perhaps the best application of convergence in probability.

Theorem 3.2 (The weak law of large numbers; WLLN)**.** *Let $X_1, X_2, ...$ be iid random variables with mean μ and variance σ^2. Let $\bar{X}_n$ denote the sample mean, i.e.*

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then, as $n \to \infty$,

$$\bar{X}_n \xrightarrow{P} \mu.$$

The LLN is very natural: When the sample size increases, the sample mean becomes more and more close to the population mean. Furthermore, the distribution of $\bar{X}_n$ degenerates to a single point distribution at μ, the true mean.

Proof. Recall that $\mathrm{E}(\bar{X}_n) = \mu$ and $\mathrm{Var}(\bar{X}_n) = \sigma^2/n$. Choose an $\epsilon > 0$ such that $\epsilon = t\sigma/\sqrt{n}$. By Chebyshev's inequality,

$$\begin{aligned}\Pr(|\bar{X}_n - \mu| \geq \overbrace{t\sigma/\sqrt{n}}^{\epsilon}) &\leq \frac{1}{t^2} \\ &= \frac{\sigma^2}{n\epsilon^2} \to 0\end{aligned}$$

as $n \to \infty$. Hence, $\bar{X}_n \xrightarrow{\mathrm{P}} \mu$. □

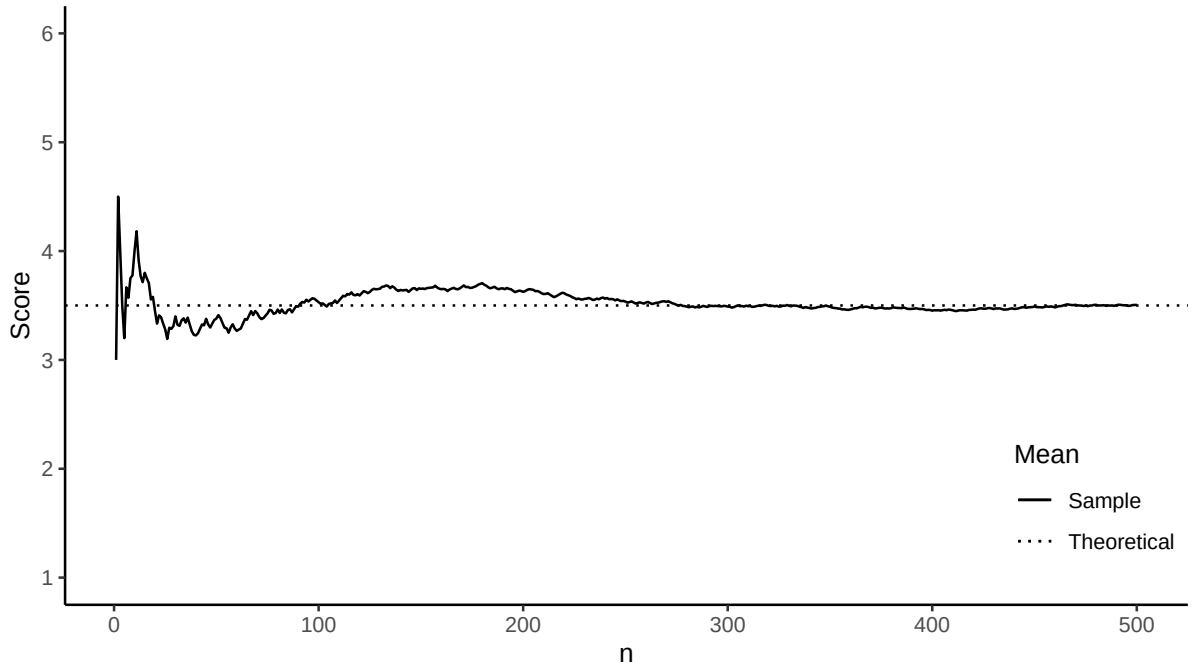
As an illustration of the WLLN, consider an experiment where we throw a six-sided die repeatedly and independently. Let $X_1, X_2, ...$ be the scores of the dice throws. We know that the true mean is $\mu = 3.5$. Let's simulate some dice throws:

```
set.seed(123)
(X <- sample(6, size = 20, replace = TRUE))
```

```
##  [1] 3 6 3 2 2 6 3 5 4 6 6 1 2 3 5 3 3 1 4 1
```

```
Xbar <- cumsum(X) / seq_along(X)
round(Xbar, 2)
```

```
##  [1] 3.00 4.50 4.00 3.50 3.20 3.67 3.57 3.75 3.78 4.00 4.18 3.92 3.77 3.71 3.80
## [16] 3.75 3.71 3.56 3.58 3.45
```



It would be very good practice to work out the true mean ($\mu = 3.5$) of the scores of the dice throws, using the formula for the expectations of discrete probability models.

3.4.2 The Central Limit Theorem

The LLN assures us that $\bar{X}_n$ eventually will be indistinguishable from μ w.p. 1. However, we would still be interested in the distribution of $\bar{X}_n$ in order to make *probabilistic statements* about $\bar{X}_n$.

Theorem 3.3 (Central Limit Theorem; CLT)**.** *Let $X_1, \ldots, X_n$ be iid random variable with mean μ and variance σ^2, and let $\bar{X}_n$ denote the sample mean. Then*

$$\bar{Z}_n := \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} \mathrm{N}(0,1)$$

as $n \to \infty$.

In words: "the standardised sample mean $\bar{Z}_n$ is approximately standard normal when the sample size is large". This is remarkable because we assume nothing about the distribution of the individual X_is! The CLT is one of the reasons why the normal distribution is the most useful and important distribution in statistics.

Alternative statements for the CLT include

- $\sqrt{n}(\bar{X}_n - \mu)/\sigma \approx \mathrm{N}(0,1)$
- $\sqrt{n}(\bar{X}_n - \mu) \approx \mathrm{N}(0, \sigma^2)$
- $\bar{X}_n - \mu \approx \mathrm{N}(0, \sigma^2/n)$
- $\bar{X}_n \approx \mathrm{N}(\mu, \sigma^2/n)$

Some other remarks:

- The CLT gives us information about the *variability* of the sample mean statistic in repeated sampling, see the slides after the next example.
- The CLT tells us nothing about the *accuracy* of any implied approximation for finite n.

- However, it still yields remarkably accurate approximations in many situations, even with modest n.
- A version of the proof involves mgfs, as you will see in the exercises.
- The CLT is responsible for the normal approximations to the binomial, Poisson, gamma, etc.!

Example 3.7. Recall the dice example above. The CLT implies that

$$\bar{X}_n \approx \mathrm{N}\left(3.5, \frac{105}{36n}\right), \tag{3.2}$$

since $\mathrm{Var}(X_i) = 105/36$.

See that variance of 105/36 in the example above? Try and obtain this figure yourself using the usual definitions of variances, or better yet, employ the results from the binomial distribution.

To illustrate this, we can take many samples of size n and compute the sample mean for each set, we then obtain many sample means. The standardised histogram of those samples resembles the normal pdf in (3.2).

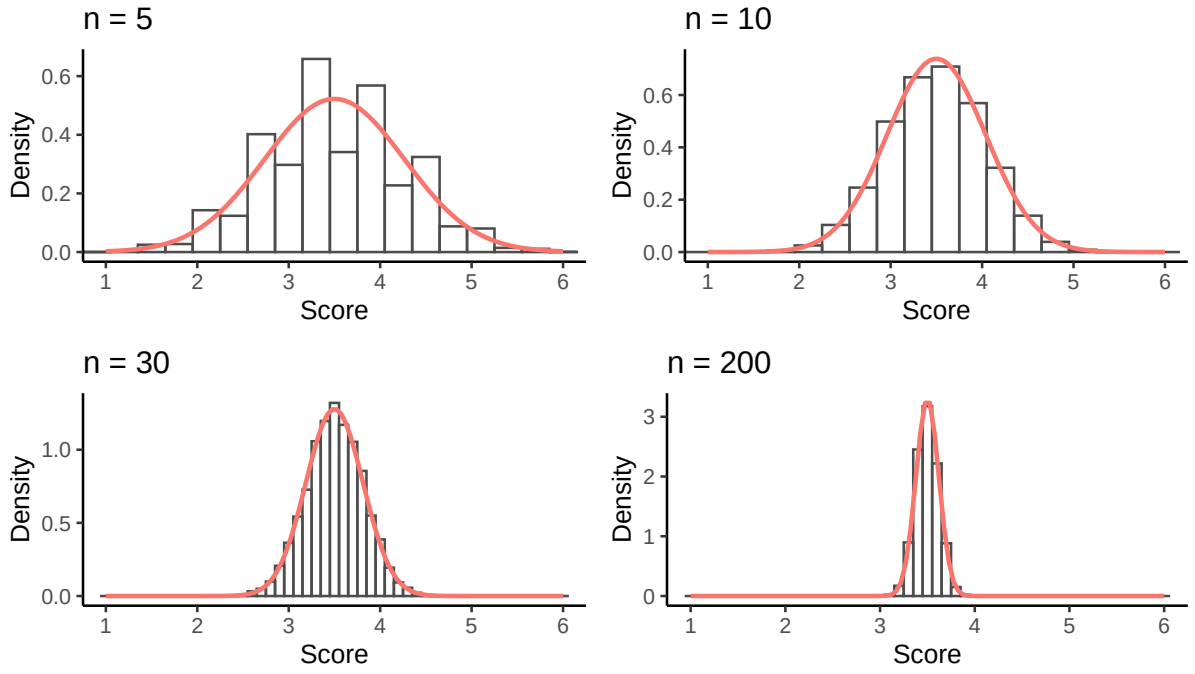
Here's the `R` code to replicate dice rolls and the sample means. The idea is to generate a sample of size n of dice roll scores repeatedly `B` times.

```
my_clt_fn <- function(n = 5, B = 10000) {
  res <- rep(NA, B)
  for (i in 1:B) {
    X <- sample(1:6, size = n, replace = TRUE)
    res[i] <- mean(X)
  }
  res
}
```

We can also use this to retrieve $\bar{X}_{20} = 3.45$ using the same random seed.

```
set.seed(123); my_clt_fn(n = 20, B = 1)
```

```
## [1] 3.45
```



3.4.3 Gauging the error of sample mean estimator

The CLT can be used as a quick way to obtain confidence statements for the sample mean. A natural estimator for the population mean $\mu = \mathrm{E}(X_i)$ is the sample mean $\bar{X}_n$. By the CLT, we can easily gauge the error of this estimation as follows:

$$\begin{aligned}\Pr(|\bar{X}_n - \mu| > \epsilon) &= \Pr(\big|\overbrace{\sqrt{n}(\bar{X}_n - \mu)/\sigma}^{\approx \mathrm{N}(0,1)}\big| > \sqrt{n}\epsilon/\sigma) \\ &\approx 2(1 - \Phi(\sqrt{n}\epsilon/\sigma))\end{aligned}$$

So with ϵ, n, and σ given, we can find the value $\Phi(\sqrt{n}\epsilon/\sigma)$ from the standard normal table. For instance, let $\epsilon := 2\sigma/\sqrt{n} = 2\sqrt{\mathrm{Var}(\bar{X}_n)}$[1]. Then

$$\begin{aligned}\Pr(|\bar{X}_n - \mu| \le \epsilon) &= 1 - \Pr(|\bar{X}_n - \mu| > \epsilon) \\ &\approx 2\Phi(2) - 1 \\ &= 0.954\end{aligned}$$

Hence, if one estimates μ by $\bar{X}_n$, and repeats it a large number of times, about 95% of times, μ is within $2 \times \mathrm{s.d.}(\bar{X}_n)$ distance away from $\bar{X}_n$

Does this look familiar to you? Recall the "68-95-99.7" rule!

3.4.4 CLT with σ^2 unknown

Typically, $\sigma^2 = \mathrm{Var}(X_i)$ is unknown in practice. We estimate it using the (unbiased) sample variance estimator

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

Note that the estimate of $\sqrt{\mathrm{Var}(\bar{X}_n)} = \sigma/\sqrt{n}$, given by $S_n/\sqrt{n}$, is called the **standard error** of the sample mean. In full,

$$\mathrm{SE}(\bar{X}_n) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

In fact, it still holds that as $n \to \infty$,

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \xrightarrow{\mathrm{D}} \mathrm{N}(0,1)$$

which implies that replacing σ with S_n in CLT applications yields the same results. Phew!

3.5 Delta method

We may be interested in the distribution of a transformation of a random variable instead of the actual random variable itself. For this, we use the delta method.

Theorem 3.4 (The delta method)**.** *Suppose that X_n is a sequence of random variable satisfying $\sqrt{n}(X_n - \mu)/\sigma \xrightarrow{D} \mathrm{N}(0,1)$. Let g be a differentiable function s.t. $g'(\mu) \neq 0$. Then*

$$\frac{\sqrt{n}(g(X_n) - g(\mu))}{|g'(\mu)|\sigma} \xrightarrow{D} \mathrm{N}(0,1).$$

[1] You might be wondering why I chose this value. It's totally abitrary. It allows us to rearrange the probability statement in order to make use of the CLT to bound the probability statement within 2 standard deviations from the mean.

In other words,

$$X_n \approx \mathrm{N}\left(\mu, \sigma^2/n\right) \quad \Rightarrow g(X_n) \approx \mathrm{N}\left(g(\mu), (g'(\mu))^2\sigma^2/n\right).$$

Example 3.8. Suppose we observe $X_1, \ldots, X_n \sim \mathrm{Bern}(p)$. A reasonable estimator for p is the sample mean $\hat{p} := \bar{X}_n = n^{-1}\sum_{i=1}^n X_i$. According to the CLT, $\hat{p} \approx (p, p(1-p)/n)$ for large n, since $\mathrm{Var}(X_i) = p(1-p)$.

Another popular parameter is $\frac{p}{1-p}$, the *odds*. This is a transformation of p using $g : p \mapsto \frac{p}{1-p}$, for which $g'(p) = \frac{1}{(1-p)^2}$. Using the delta method, we deduce that

$$\frac{\hat{p}}{1-\hat{p}} \approx \mathrm{N}\left(\frac{p}{1-p}, \frac{p}{n(1-p)^3}\right).$$

We'll definitely make use of the delta method in the next chapter, when we consider point estimation.

3.6 Normal random samples

Given that the normal distribution is very often used, the properties of normal random samples have been studied extensively.

Theorem 3.5. *Let* $\{X_1, \ldots, X_n\}$ *be a sample from* $\mathrm{N}(\mu, \sigma^2)$*, and let*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i, \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2, \qquad \textit{and} \qquad SE(\bar{X}) = S/\sqrt{n}.$$

Then,

- $\bar{X}$ *and* S^2 *are independent random variables*
- $\bar{X} \sim \mathrm{N}(\mu, \sigma^2/n)$
- $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$
- $\frac{\sqrt{n}(\bar{X}-\mu)}{S} = \frac{\bar{X}-\mu}{SE(\bar{X})} \sim t_{n-1}$

The above theorem mentions two kinds of distribution (that you may have heard of) but we are yet to discuss. We'll circle back to the proof othis theorem after covering the χ^2 and t distributions.

3.6.1 χ^2-distribution

The χ^2-distribution is an important distribution in statistics. It is closely linked with the normal, Student's t and F distributions. Inference for the variance parameter σ^2 relies on χ^2-distributions. More importantly, most goodness-of-fit tests are based on χ^2-distributions.

Definition 3.7 (χ^2-distribution)**.** Let $Z_1, \ldots, Z_k \overset{\text{iid}}{\sim} \mathrm{N}(0,1)$, i.e. each Z_i has pdf $f(z_i) = (2\pi)^{-1/2}e^{-z_i^2/2}$ for $i = 1, \ldots, k$. Then,

$$X = Z_1^2 + \cdots + Z_k^2 = \sum_{i=1}^k Z_i^2$$

follows a χ^2-distribution with $k \in \mathbb{N}$ degrees of freedom. We write $X \sim \chi^2_k$.

Out of curiosity, the pdf of a χ^2_k distribution is $f(x) = Cx^{k/2-1}e^{-x/2}$, where the normalising constant C is equal to $2^{-k/2}\Gamma^{-1}(k/2)$ ($\Gamma(\cdot)$ is the gamma function). The form of the pdf is less important to know than the definition of χ^2_k distribution given in Definition 3.7.

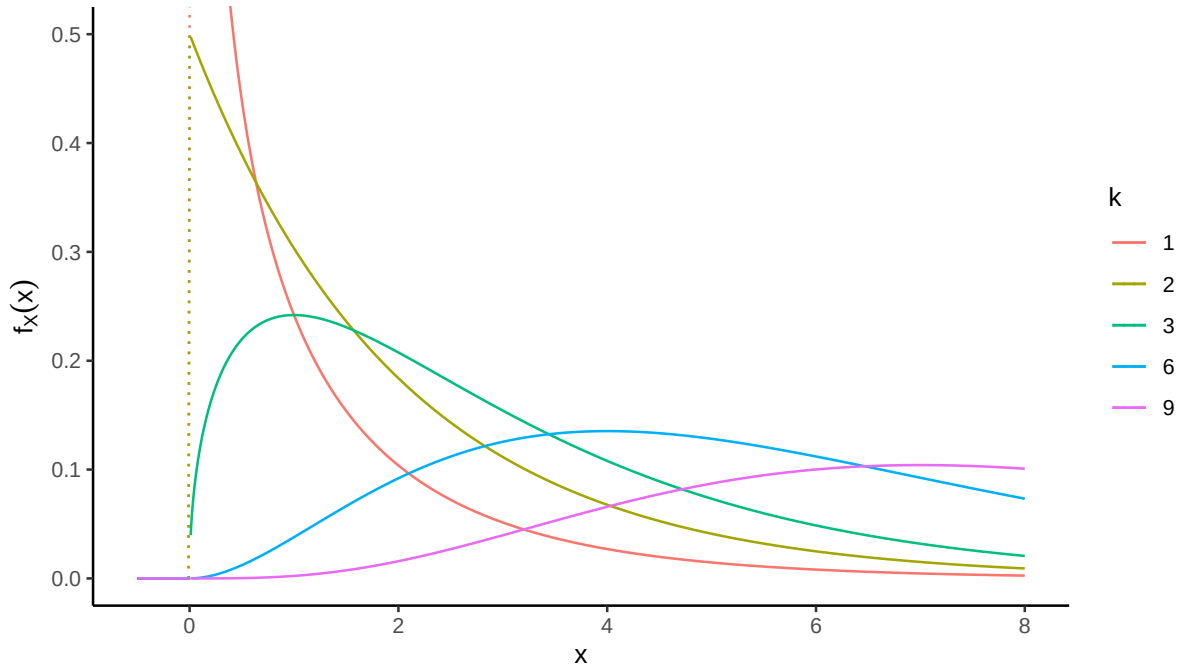
Here are some important properties of the χ^2_k distribution.

- X has support over $[0, \infty)$.
- $\mathrm{E}(X) = k$.

- $\text{Var}(X) = 2k$.
- If $X_1 \sim \chi^2_{k_1}$ and $X_2 \sim \chi^2_{k_2}$, and $X_1 \perp X_2$, then $X_1 + X_2 \sim \chi^2_{k_1+k_2}$.

There is a question at the end of this chapter where you will prove the above statements.

Pdf of χ^2_k



Probabilities such as

$$\Pr(\chi^2_k \leq x) = \int_0^x f_X(\tilde{x})\,\mathrm{d}\tilde{x}$$

where f_X is the pdf of χ^2_k cannot be found in closed form. Instead, the integral is calculated using computer approximations for the integral above. In `R`,

```
pchisq(2, df = 3)
```

```
## [1] 0.4275933
```

Alternatively, statistical tables are used. You will find tables for percentiles of the χ^2-distribution. That is, you are able to find the value of $x := \chi^2_k(\alpha)$ such that

$$\Pr(\chi^2_k \leq x) = \int_0^x f_X(\tilde{x})\,\mathrm{d}\tilde{x} = A = 1 - \alpha$$

for various values of A and k.

Example 3.9. Let $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{N}(\mu, \sigma^2)$. Then, $Z_i = \frac{Y_i - \mu}{\sigma} \sim \text{N}(0, 1)$, and hence

$$\frac{1}{\sigma^2}\sum_{i=1}^n (Y_i - \mu)^2 = \sum_{i=1}^n Z_i^2 \sim \chi^2_n.$$

Note that

$$\frac{1}{\sigma^2}\sum_{i=1}^n (Y_i - \mu)^2 = \frac{1}{\sigma^2}\sum_{i=1}^n (Y_i - \bar{Y}_n)^2 + \frac{n}{\sigma^2}(\bar{Y}_n - \mu)^2. \tag{3.3}$$

Since $\bar{Y}_n \sim \mathrm{N}(\mu, \sigma^2/n)$, it must be that $\frac{n}{\sigma^2}(\bar{Y}_n - \mu)^2 \sim \chi^2_1$. Thus, by the properties of the χ^2-distribution, the decomposition in (3.3) may be written as $\chi^2_n = \chi^2_{n-1} + \chi^2_1$. In particular, we now know

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2 \sim \chi^2_{n-1}.$$

3.6.2 Student's t-distribution

This is another important distribution in statistics, because:

- The t-test is a widely used distribution for statistical tests in many application.
- Confidence intervals for normal mean with unknown variance may be constructed based on the t-distribution.

Definition 3.8 (t-distribution)**.** Suppose we have two random variable $Z \sim \mathrm{N}(0,1)$ and $X \sim \chi^2_k$ such that X and Z are independent. Then, the distribution of the random variable

$$T = \frac{Z}{\sqrt{X/k}}$$

is called the t-distribution with $k \in \mathbb{N}$ degrees of freedom. We write $T \sim t_k$.

The pdf for $T \sim t_k$ is given by

$$f(t) \propto \left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}},$$

but once again the actual form of the pdf is not as important as the definition of the t-distribution.

Some important properties of the t-distribution:

- T is continuous and symmetric over $(-\infty, \infty)$.
- $\mathrm{E}(T) = 0$, provided $\mathrm{E}(|T|) < \infty$ $(k > 1)$.
- $\mathrm{Var}(T) = \frac{k}{k-2}$.
- Technically, $k \in \mathbb{R}$, but we will usually deal with $k \in \mathbb{N}$.

Pdf of t_k

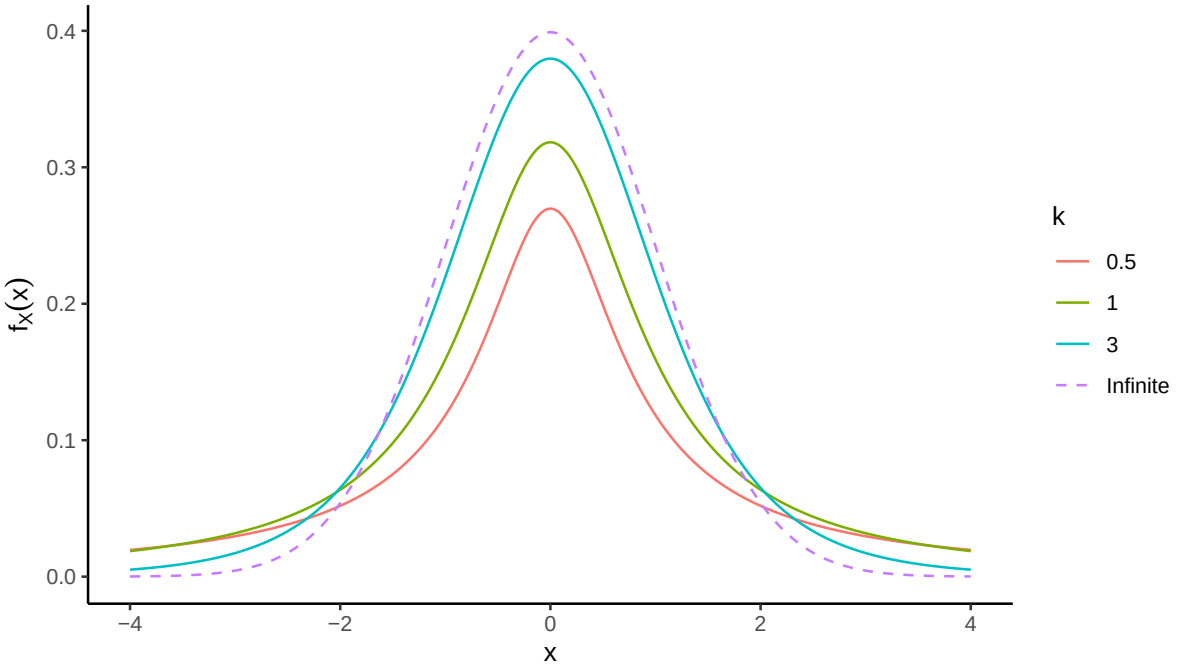


Figure 3.1: William Sealy Gosset. 13 June 1876 – 16 October 1937.

The t-distribution[2] has what is known as **heavy tails**. That is, if $T \sim t_k$, its mgf is undefined and hence $\mathrm{E}(|T|^k) = \infty$. Comparing this to the normal distribution: $X \sim \mathrm{N}(\mu, \sigma^2)$, $\mathrm{E}(|X|^k) < \infty$ for any $k > 0$. This 'heavy-tails' property is a useful property in modelling abnormal phenomena or outliers (e.g. in financial or insurance data). c.f. "robust statistics"

The connection between the t_k distribution and the normal distribution, is that the t_k actually approaches the standard normal as the degrees of freedom increases.

Lemma 3.6. $t_k \xrightarrow{D} \mathrm{N}(0,1)$ *as* $k \to \infty$.

Proof. If $X \sim \chi^2_k$, then by definition $X = Z_1^2 + \cdots + Z_k^2$, where $Z_i \overset{\text{iid}}{\sim} \mathrm{N}(0,1)$. By the LLN,

$$\frac{X}{k} = \frac{Z_1^2 + \cdots + Z_k^2}{k} \xrightarrow{\mathrm{P}} \mathrm{E}(Z_1^2) = 1.$$

as $k \to \infty$. Therefore, $\sqrt{X/k} \xrightarrow{\mathrm{P}} 1$, and in particular,

$$T = \frac{Z}{\sqrt{X/k}} \xrightarrow{\mathrm{D}} \mathrm{N}(0,1)$$

following Slutzky's theorem. □

3.6.3 Proof of Theorem 3.5

Back to this theorem. Let's prove it.

[2]Explore the t-distribution vs normal distribution here: https://eripoll12.shinyapps.io/t_Student/

Proof. ii. follows directly from properties of normal distributions, and earlier we showed that $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\bar{X})^2 \sim \chi^2_{n-1}$ which settles iii.

To prove i., consider any X_j, $j \in \{1,\dots,n\}$ and $\mathrm{Cov}(X_j-\bar{X},\bar{X})$:

$$\begin{aligned}\mathrm{Cov}(X_j-\bar{X},\bar{X}) &= \mathrm{Cov}(X_j,\bar{X}) - \mathrm{Cov}(\bar{X},\bar{X}) \\ &= \mathrm{Cov}\left(X_j, \frac{1}{n}\sum_{i=1}^{n} X_i\right) - \mathrm{Var}(\bar{X}) \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{Cov}(X_j,X_i) - \sigma^2/n = \sigma^2/n - \sigma^2/n = 0\end{aligned}$$

Since the covariance is zero and they are normal, they are independent.

Following this, if $\bar{X}$ is independent of $X_j-\bar{X}$ for any j, it stands to reason that $\bar{X}$ is also independent of $\tilde{X} = (X_1-\bar{X},\dots,X_n-\bar{X})^\top$, and also of

$$\tilde{X}^\top\tilde{X} = \begin{pmatrix}X_1-\bar{X} & \cdots & X_n-\bar{X}\end{pmatrix}\begin{pmatrix}X_1-\bar{X}\\ \vdots \\ X_n-\bar{X}\end{pmatrix} = \sum_{i=1}^{n}(X_i-\bar{X})^2 = (n-1)S^2,$$

and thus also of S^2. Here we used the fact that if $X \perp Y_i$, then $g(X) \perp g(Y_i)$, and also $g(X) \perp \{g(Y_1)+\cdots+g(Y_n)\}$.

Finally, putting everything together,

$$\frac{\overbrace{\sqrt{n}(\bar{X}-\mu)/\sigma}^{\mathrm{N}(0,1)}}{\sqrt{\frac{\overbrace{(n-1)S^2/\sigma^2}^{\chi^2_{n-1}}}{n-1}}} = \frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{\bar{X}-\mu}{\mathrm{SE}(\bar{X})} \sim t_{n-1}.$$

□

This is why for normal distributions where σ^2 is unknown, and is estimated by the unbiased sample variance s^2, the standardised sample mean follows a t-distribution! This gives rise to the t-test.

3.6.4 F-distribution

The F-distribution is another notable distribution in statistics. It commonly arises as the null distribution of a test statistic, particularly in the analysis of variance (ANOVA).

Definition 3.9 (F-distribution)**.** Let $X_1 \sim \chi^2_{k_1}$ and $X_2 \sim \chi^2_{k_2}$. Then, the distribution of

$$Y = \frac{X_1/k_1}{X_2/k_2}$$

is called the F-distribution with (k_1,k_2) degrees of freedom. We write $Y \sim F_{k_1,k_2}$.

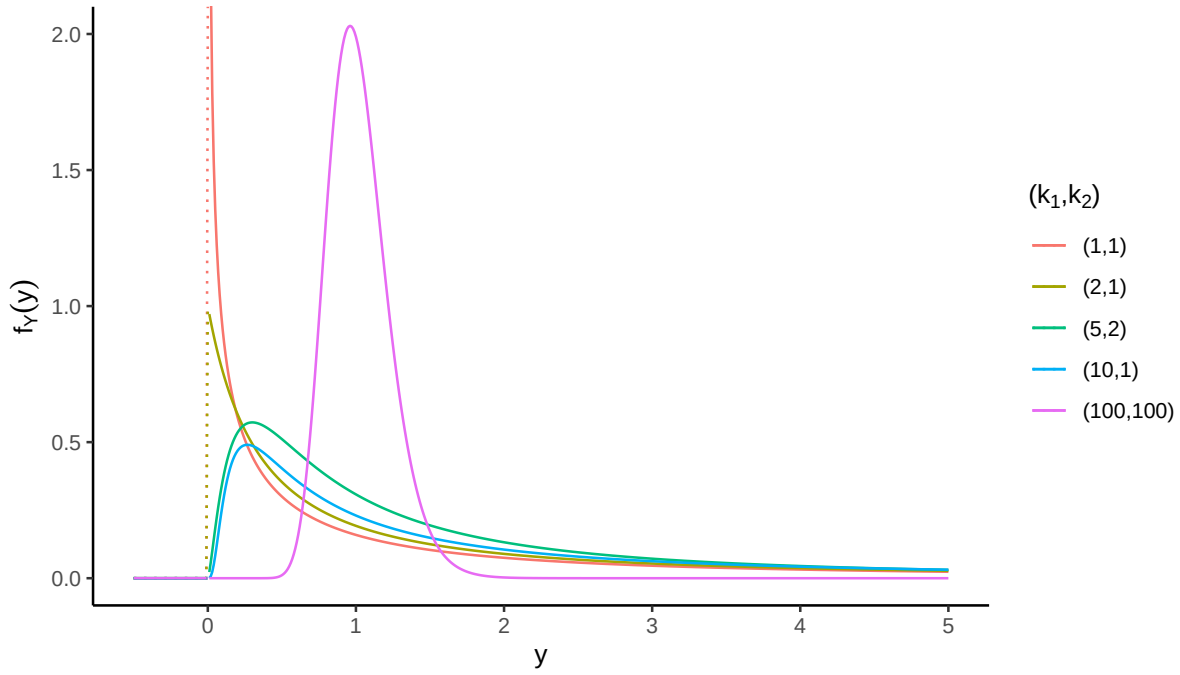
Not even going to bother writing down the pdf! See for yourself: https://en.wikipedia.org/wiki/F-distribution. Remember the definition, though.

Some important properties of the F-distribution:

- Y is continuous and has support over $[0,\infty)$, provided $k_1 > 1$.
- $\mathrm{E}(Y) = \frac{k_2}{k_2-2}$, provided $k_2 > 2$.
- $\mathrm{Var}(Y) = \frac{2k_2^2(k_1+k_2-2)}{k_1(k_2-2)^2(k_2-4)}$, provided $k_2 > 4$.
- Technically, $k_1,k_2 \in \mathbb{R}_{>0}$, but we will usually deal with $k_1,k_2 \in \mathbb{N}$.
- If $Y \sim F_{k_1,k_2}$, then $Y^{-1} \sim F_{k_2,k_1}$.

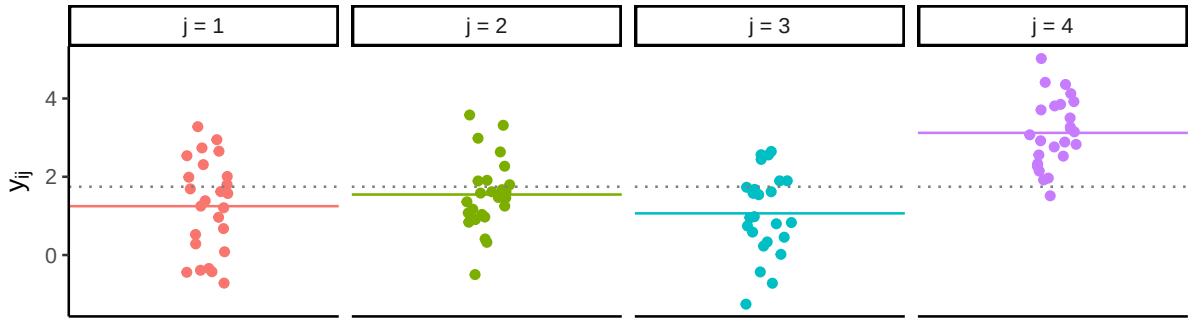
- If $T \sim t_k$, then $T^2 \sim F_{1,k}$.

Attempt to prove some of these in the exercises!



3.6.5 The analysis of variance

The ANOVA, despite its name, is a (collection of) methods used to analyse differences among group means in a sample.



The setup is as follows: Let $Y_{ij} \sim \mathrm{N}(\mu_j, \sigma^2)$, $i = 1, \dots, n_j$ and $j = 1, \dots, m$ with both μ_j and σ^2 unknown. Let $n = \sum_{j=1}^{m} n_j$ be the total sample size. Define

- the grand mean $\bar{Y} = n^{-1} \sum_{i,j} Y_{ij}$; and
- the group means $\bar{Y}_j = n_j^{-1} \sum_{i=1}^{n_j} Y_{ij}$, $j = 1, \dots, m$.

Consider the "total sum of squares" $TSS = \sum_{i,j} (Y_{ij} - \bar{Y})^2$, which can be decomposed into

$$TSS = \overbrace{\sum_{i,j} (Y_{ij} - \bar{Y}_j)^2}^{WSS} + \overbrace{\sum_{j} n_j (\bar{Y}_j - \bar{Y})^2}^{BSS}$$

where

- WSS is the "within sum of squares" (how much variation among individuals in each group); and
- BSS is the "between sum of squares" (how much variation in the mean among groups).

There is a concept of *degrees of freedom*: $n-1$ in the TSS, $m-1$ in the BSS, and therefore $n-m$ in the WSS.

This gives rise to the ANOVA table:

Source	SS	d.f.	MSS	F-statistic
Between	$\sum_j n_j(\bar{Y}_j - \bar{Y})^2$	$m-1$	$\frac{\sum_j n_j(\bar{Y}_j-\bar{Y})^2}{m-1}$	$\frac{\sum_j n_j(\bar{Y}_j-\bar{Y})^2/(m-1)}{\sum_{i,j}(Y_{ij}-\bar{Y}_j)^2/(n-m)}$
Within	$\sum_{i,j}(Y_{ij} - \bar{Y}_j)^2$	$n-m$	$\frac{\sum_{i,j}(Y_{ij}-\bar{Y}_j)^2}{n-m}$	
Total	$\sum_{i,j}(Y_{ij} - \bar{Y})^2$	$n-1$		

Suppose we want to test the hypothesis that all group means are identical (i.e. $\mu_j = \mu, \forall j$), what is the distribution of F?

We have seen that

$$TSS/\sigma^2 = \frac{1}{\sigma^2}\sum_{i,j}(Y_{ij} - \bar{Y})^2 \sim \chi^2_{n-1}.$$

In fact, we can also show similarly that

$$WSS/\sigma^2 = \frac{1}{\sigma^2}\sum_{i,j}(Y_{ij} - \bar{Y}_j)^2 \sim \chi^2_{n-m}.$$

Using these two facts, we deduce that

$$BSS/\sigma^2 = \frac{1}{\sigma^2}\sum_{j} n_j(\bar{Y}_j - \bar{Y})^2 \sim \chi^2_{m-1}$$

from the property of χ^2-distributions.

So now,

$$F = \frac{\text{mean } BSS}{\text{mean } WSS} = \frac{\overbrace{1/\sigma^2 \textstyle\sum_j n_j(\bar{Y}_j - \bar{Y})^2}^{\chi^2_{m-1}}/(m-1)}{\underbrace{1/\sigma^2 \textstyle\sum_{i,j}(Y_{ij} - \bar{Y}_j)^2}_{\chi^2_{n-m}}/(n-m)}$$

is a ratio of two χ^2-distributions, which means that F follows an F-distribution with $(m-1, n-m)$ degrees of freedom.

3.7 Exercises

1. In this exercise, you will prove the central limit theorem using moment generating functions. Let $X_1, X_2, ...$ be a sequence of iid random variables with mean $\mathrm{E}(X_i) = \mu$ and variance $\mathrm{Var}(X_i) = \sigma^2 > 0$, and whose mgfs exist in a neighbourhood of 0 (i.e., $M_{X_i}(t)$ exists for $|t| < h$ for some positive h).

   (a) Define $Y_i = (X_i - \mu)/\sigma$, and let $M_Y(t)$ denote the common mgf of the Y_is. Show that $M_Y^{(0)} = 1$, $M_Y^{(1)} = 0$, and $M_Y^{(2)} = 1$.

   (b) Using the properties of mgfs, show that the mgf of $Z := \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$ is given by

   $$M_Z(t) = \left(M_Y(t/\sqrt{n})\right)^n. \tag{3.4}$$

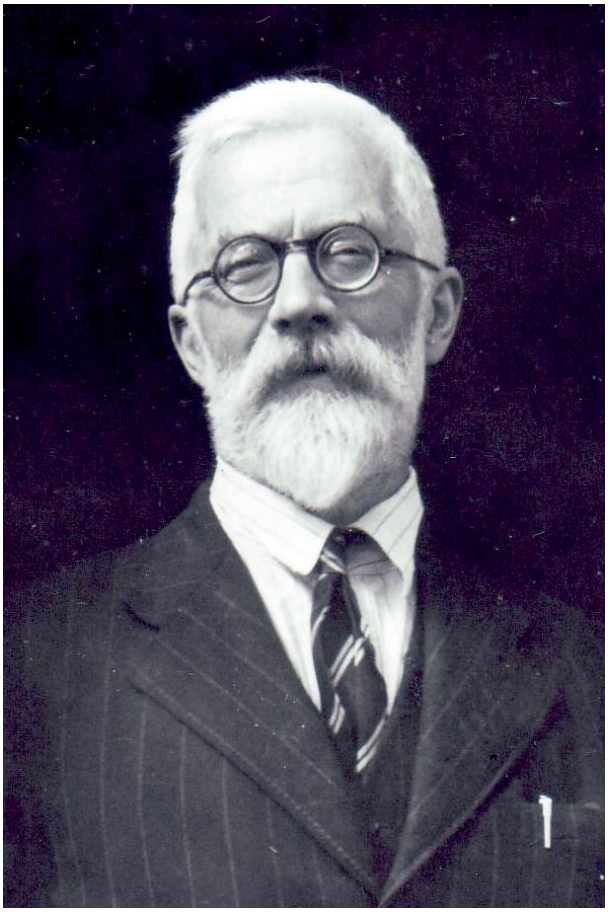


Figure 3.2: Sir Ronald Aylmer Fisher. 17 February 1890 – 29 July 1962.

(c) Recall that the Taylor series expansion of a real function $g(x)$ around $a \in \mathbb{R}$ is given by the power series

$$g(x) = \sum_{k=0}^{\infty} \frac{g^{(k)}}{k!}(x-a)^k = g(a) + \frac{g'(a)}{1!}(x-a) + \frac{g''(a)}{2!}(x-a)^2 + \cdots.$$

By Taylor expanding $M_Y(t/\sqrt{n})$ about 0, show that

$$M_Y(t/\sqrt{n}) = 1 + \frac{(t/\sqrt{n})^2}{2!} + R_n$$

where R_n is a remainder term from the Taylor series.

(d) Hence, using the fact that $nR_n \to 0$ as $n \to \infty$, show that the limiting distribution of Z is $\mathrm{N}(0,1)$.

2. (a) Suppose we have a sequence $X_1, X_2, \ldots$ of iid random variables with $\mathrm{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$. Show that the statistic $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$, where $\bar{X}_n$ is the sample mean, is an **unbiased** and **consistent** estimator for σ^2 (i.e. it converges in probability to σ^2). *Hint: Show that $S_n^2 = c_n n^{-1}\sum_{i=1}^n X_i^2 - d_n \bar{X}_n^2$ where $c_n, d_n \to 1$ as $n \to \infty$. Apply the LLN to $n^{-1}\sum_{i=1}^n X_i^2$ and to $\bar{X}_n^2$. Then use Slutzky's theorem.*

(b) Using Slutzky's theorem, show that

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \xrightarrow{\mathrm{D}} \mathrm{N}(0,1).$$

Hint: First argue that $\sigma/S_n \xrightarrow{P} 1$ using Slutzky's theorem. You may use the fact that $g(x) = c/\sqrt{x}$ is continuous for $x > 0$, where $c \in \mathbb{R}_{>0}$ is a constant. Then use the CLT.

3. Suppose that $\bar{X}$ and S^2 are calculate from an iid random sample $X_1, \ldots, X_n$ with $\mathrm{Var}(X_i) = \sigma^2$. We know that $\mathrm{E}(S^2) = \sigma^2$ from Q2(a). Prove that $\mathrm{E}(S) \leq \sigma$. *Hint: Use Jensen's inequality.*

4. Fill in the details of the proof of the delta method: Suppose that X_n is a sequence of random variables satisfying $\sqrt{n}(X_n - \mu)/\sigma \xrightarrow{\mathrm{D}} \mathrm{N}(0,1)$, and let g be a differentiable function s.t. $g'(\mu) \neq 0$.

(a) Show that $X_n \xrightarrow{\mathrm{P}} \mu$ in probability.

(b) Show, using Slutzky's theorem, that

$$\frac{\sqrt{n}(g(X_n) - g(\mu))}{|g'(\mu)|\sigma} \xrightarrow{\mathrm{D}} \mathrm{N}(0,1).$$

Hint: Use the Taylor approximation $g(X_n) \approx g(\mu) + g'(\mu)(X_n - \mu)$.

5. Let $\{X_1, \ldots, X_n\}$ be a random sample from a $\mathrm{N}(\mu, \sigma^2)$ population.

(a) Let $M = \sum_{i=1}^n (X_i - \bar{X})^2$, where $\bar{X}$ is the sample mean. Work out the distribution of M/σ^2.

(b) Let $\alpha = 0.05$. Using the χ^2 probability tables, determine the values of $\chi^2_{14}(\alpha/2)$ and $\chi^2_{14}(1-\alpha/2)$, i.e. the top and bottom $\alpha/2$ point of the χ^2_{14} distribution where $\Pr(Y > \chi^2_k(a)) = a$ when $Y \sim \chi^2_k$.

6. Suppose that we plan to take a random sample of size n from a normal distribution with mean μ and standard deviation $\sigma = 2$.

(a) Suppose $\mu = 4$ and $n = 20$.

i. What is the probability that the mean $\bar{X}$ of the sample is greater than 5?

ii. What is the probability that $\bar{X}$ is smaller than 3?

iii. What is $\Pr(|\bar{X} - \mu| \leq 1)$ in this case?

(b) How large should n be in order that $\Pr(|\bar{X} - \mu| \leq 0.5) \geq 0.95$ for every possibly value of μ?

(c) It is claimed that the true value of μ is 5 in a population. A random sample of size $n = 100$ is collected from this population, and the mean for this sample is $\bar{X} = 5.8$. Based on the result in (b), what would you conclude from this value of $\bar{X}$?

7. In all of the following sub-questions, use only probability tables and distributional properties of the relevant random variables to calculate the required probabilities.

   (a) If Z is a random variable with a standard normal distribution, what is $\Pr(Z^2 < 3.841)$?

   (b) Suppose that X_1 and X_2 are independent $\mathrm{N}(0, 4)$ random variables. Compute $\Pr(X_1^2 < 36.84 - X_2^2)$.

   (c) Suppose that $X_1, X_2, X_3 \overset{\text{iid}}{\sim} \mathrm{N}(0, 1)$, while Y independently follows a χ^2_5 distribution. Compute $\mathrm{P}(X_1^2 + X_2^2 < 7.236Y - X_3^2)$.

8. Let X_i, $i = 1, 2, 3$ be independent with $\mathrm{N}(i, i^2)$ distributions. For each of the following situations, use the X_is to construct a statistic with the indicated distribution:

   (a) χ^2-distribution with 3 degrees of freedom;

   (b) t-distribution with 2 degrees of freedom; and

   (c) F-distribution with 1 and 2 degrees of freedom.

9. Let $\{Y_{ij}\}$ be sample from $\mathrm{N}(\mu_j, \sigma^2)$, $i = 1, \ldots, n_j$ and $j = 1, \ldots, m$. In total there are $n = \sum_{j=1}^{m} n_j$ samples. Further, let $S = \sum_{i=1}^{n_j} \sum_{j=1}^{m} (Y_{ij} - \bar{Y})^2$, where $\bar{Y} = \frac{1}{n} \sum_{i=1} \sum_{j=1}^{m} Y_{ij}$.

   (a) Define the sample group means to be $\bar{Y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{ij}$. Add and subtract the sample group mean $\bar{Y}_j$ into the squared sum in S to show that

$$\sum_{i=1}^{n_j} \sum_{j=1}^{m} (Y_{ij} - \bar{Y})^2 = \sum_{i=1}^{n_j} \sum_{j=1}^{m} (Y_{ij} - \bar{Y}_j)^2 + \sum_{j=1}^{m} n_j (\bar{Y}_j - \bar{Y})^2$$

   (b) What is the distribution of $\bar{Y}$ and $\bar{Y}_j$?

   (c) Assuming that $\mu_j = \mu$, for all $j = 1, \ldots, m$ and using your answer to (b), determine then the following distributions

      i. $\frac{1}{\sigma^2} \sum_{i=1}^{n_j} \sum_{j=1}^{m} (Y_{ij} - \mu)^2$
      ii. $\frac{n}{\sigma^2} (\bar{Y} - \mu)^2$
      iii. $\frac{1}{\sigma^2} \sum_{i=1}^{n_j} \sum_{j=1}^{m} (Y_{ij} - \bar{Y})^2$
      iv. $\frac{1}{\sigma^2} \sum_{j=1}^{m} n_j (\bar{Y}_j - \mu)^2$
      v. $\frac{1}{\sigma^2} \sum_{i=1}^{n_j} \sum_{j=1}^{m} (Y_{ij} - \bar{Y}_j)^2$

   Hint: Use the sum of squares decomposition with $\bar{Y}$ and $\bar{Y}_j$, and then use the properties of χ^2-distributions.

10. This question relates to the normal approximations of several commonly used distributions. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x)$. Answer the following questions in cases where $f(x)$ is the pmf/pdf of $\mathrm{Bern}(p)$, $\mathrm{Poi}(\lambda/n)$, and $\Gamma(\alpha/n, \beta)$:

    (a) What is the distribution of $Y = \sum_{i=1}^{n} X_i$?

    (b) Using the CLT, what is the (approximate) distribution of Y/n?

    (c) What can you then say about the distribution of Y as $n \to \infty$?

Hand-in questions

1. (a) State and prove Chebyshev's inequality for a random variable X. **[3 marks]**
   (b) Let Y be a random variable with mean $\mu = 50$ and variance $\sigma^2 = 25$. Find the bounds on the probability of the random variable having a value between 40 and 60. **[2 marks]**

2. (a) Let $X \sim \chi^2_k$. Prove that $\mathrm{E}(X) = k$ and $\mathrm{Var}(X) = 2k$. **[2 marks]**
   (b) If $X_1 \sim \chi^2_{k_1}$, $X_2 \sim \chi^2_{k_2}$, and $X_1 \perp X_2$, show that $X_1 + X_2 \sim \chi^2_{k_1+k_2}$. **[2 marks]**

3. Let $\lambda_n = 1/n$ for $n = 1, 2, ...$ Let $X_n \sim \mathrm{Poi}(\lambda_n)$.

   (a) Show that $X_n \xrightarrow{\mathrm{P}} 0$ as $n \to \infty$. **[2 marks]**

   (b) Let $Y_n = nX_n$. Show that $Y_n \xrightarrow{\mathrm{P}} 0$ as $n \to \infty$. **[2 marks]**

4. Let $\bar{X}_n$ and S^2_n be the usual mean and variance statistics for the sample $X_1, \ldots, X_n$. Suppose that a new observation X_{n+1} becomes available, show that

   (a) $\bar{X}_{n+1} = (X_{n+1} + n\bar{X}_n)/(n+1)$. **[1 mark]**
   (b) $nS^2_{n+1} = (n-1)S^2_n + n(X_{n+1} - \bar{X}_n)^2/(n+1)$ **[3 marks]**

Part III

Inference

Chapter 4

Point estimation

In this chapter, we will be studying the first of three statistical inference activities, namely *point estimation*. When we studied probability and distributions, the parameters of those distributions were treated as known information, and from there we calculate probabilities or expectations related to those distributions. In statistical inference however, the intent is the opposite. That is, the parameters of the distributions are unknown and we are supposed to *infer* what they are from the clues that are available to us. The clues here refer to the data that is collected, assumed to be distributed according to some probability distribution whose parameters are to be found out. Figuring out what the values of the parameters should be is called *point estimation*.

Actually, there are many ways and methods of performing point estimation. By far the most common way (at least that is a starting point in these kind of discussions) is the method of maximum likelihood. It is an intuitive measure of the plausibility a certain parameter of interest given the observed data. We shall study this extensively, and in fact, the next two chapters do indeed revolve around the concept of likelihood.

Learning objectives

By the end of this chapter, you will be able to:

- Construct the likelihood function based on any given assumed probability distribution, and hence compute the maximum likelihood estimator.
- Obtain a point estimator by way of the method of moments.
- Evaluate the quality of any point estimator by means of its bias, variance, and mean-squared error.
- Asymptotically evaluate the qualities of point estimators in suitable circumstances, and in particular be familiar with the large sample properties of the MLE.

Readings

- Casella and Berger (2002)
  - Chapter 6, sections 6.1, 6.2 (excluding 6.2.3 and 6.2.4), and 6.3 (excluding 6.3.2).
  - Chapter 7, sections 7.1, 7.2 (excluding 7.2.3 and 7.2.4), and 7.3 (excluding 7.3.4).
  - Chapter 10, sections 10.1 (excluding 10.1.4).
- Wasserman (2004)
  - Chapter 6, sections 6.1, 6.2, 6.3.1
  - Chapter 9, sections 9.1–9.5, 9.7–9.9
- Topics not covered here: Ancillary statistics, complete statistics, Basu's theorem, the formal likelihood principle, Bayes estimators, the EM algorithm, loss function optimality, equivariance of MLE, (asymptotic) relative efficiency, bootstrap se, robustness, M-estimators.

4.1 The likelihood

Consider a statistical model for a random vector $X = (X_1, \dots, X_n)^\top$ whose distribution depends on (an unknown) parameter θ.

- Write $f(x|\theta)$ for the joint pdf/pmf of X when θ is **known**.
- Then, given $X = x$ is observed, the function of θ defined by

$$L(\theta|x) = f(x|\theta)$$

is called the *likelihood function* for θ based on data x.

Note the key distinction between

- f, which is considered *a function of* x (and, for example, must sum or integrate to 1)
- L, which is considered *a function of* θ.

For any fixed value of θ, say $\theta = \theta_1$, $L(\theta_1|x)$ is a *statistic*–a scalar-valued transformation of the observed values of $X = x$.

The purpose of $L(\theta|x)$ is to compare the *plausibility* of different candidate values of θ, given the observed data x.

If $L(\theta_1|x) > L(\theta_2|x)$, then the data x were more likely to occur under the hypothesis that $\theta = \theta_1$ than under the hypothesis that $\theta = \theta_2$. In that sense, θ_1 is a more plausible value than θ_2 for the unknown parameter θ.

Example 4.1. Consider a sequence of n coin tosses, and let X_i denote the outcome of the ith coin toss. Assume that $X_i \sim \text{Bern}(p)$, where p is the probability of heads. We know that the total number of heads $\sum_{i=1}^n X_i$ is distributed $\text{Bin}(n, p)$.

Suppose the outcome of $n = 10$ coin tosses happens to be

$$\{H, T, H, T, T, H, H, T, H, H\},$$

i.e. $X = \sum_{i=1}^n X_i = 6$ (the total number of heads).

The likelihood function is just the pmf of the binomial with p as its input, and $X = 6$ as the given information:

$$L(p|X) = \binom{10}{6} p^6 (1-p)^4$$

Clearly, and even intuitively, $L(0.6|X) > L(0.2|X)$, say. That is, it is more plausible that $p = 0.6$ rather than $p = 0.2$, given that 6 out of 10 heads turned up.

4.1.1 Calculating the likelihood

In `R`, the function `dbinom()` computes the pmf for the binomial distribution. That is, suppose that we have $X \sim \text{Bin}(10, 0.6)$ and we wanted to calculate $\Pr(X = x)$ we type

```
dbinom(x = 0:10, size = 10, prob = 0.6) %>%
    round(digits = 4)
```

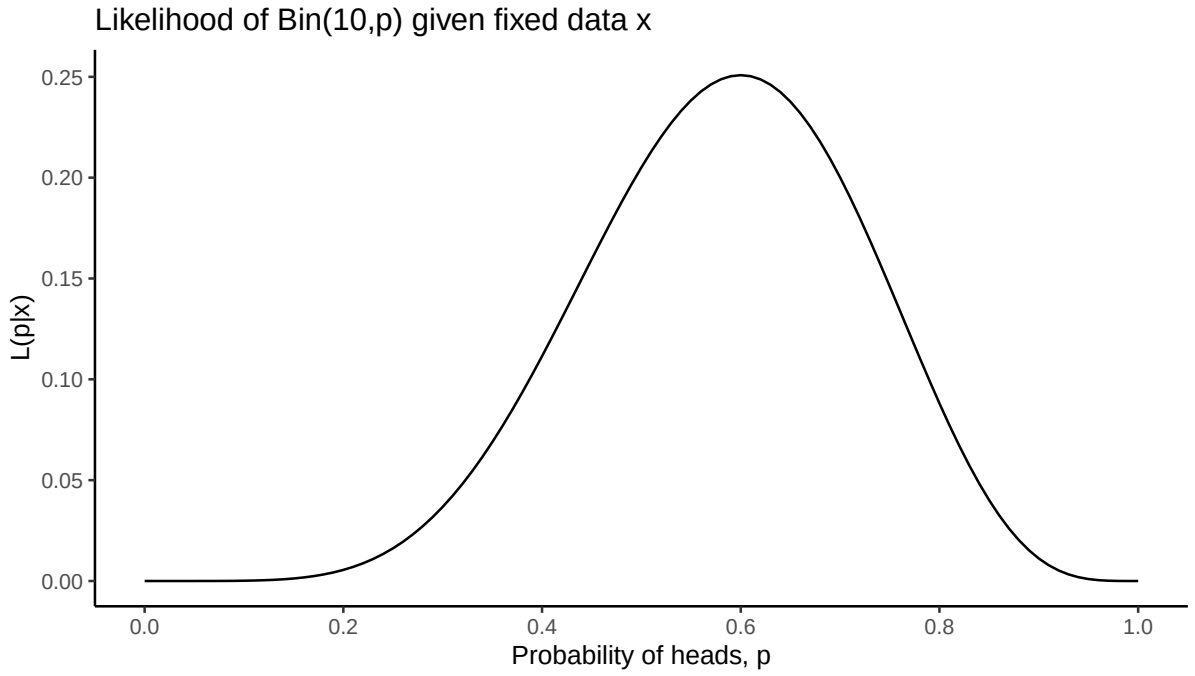
```
##  [1] 0.0001 0.0016 0.0106 0.0425 0.1115 0.2007 0.2508 0.2150 0.1209 0.0403
## [11] 0.0060
```

Since $L(\theta|x) = f(x|\theta)$, we use the same `dbinom()` to calculate the likelihood, except now we are interested in the value of the likelihood of a range of parameter values $p \in [0, 1]$ given a particular occurrence (e.g. getting 6 heads):

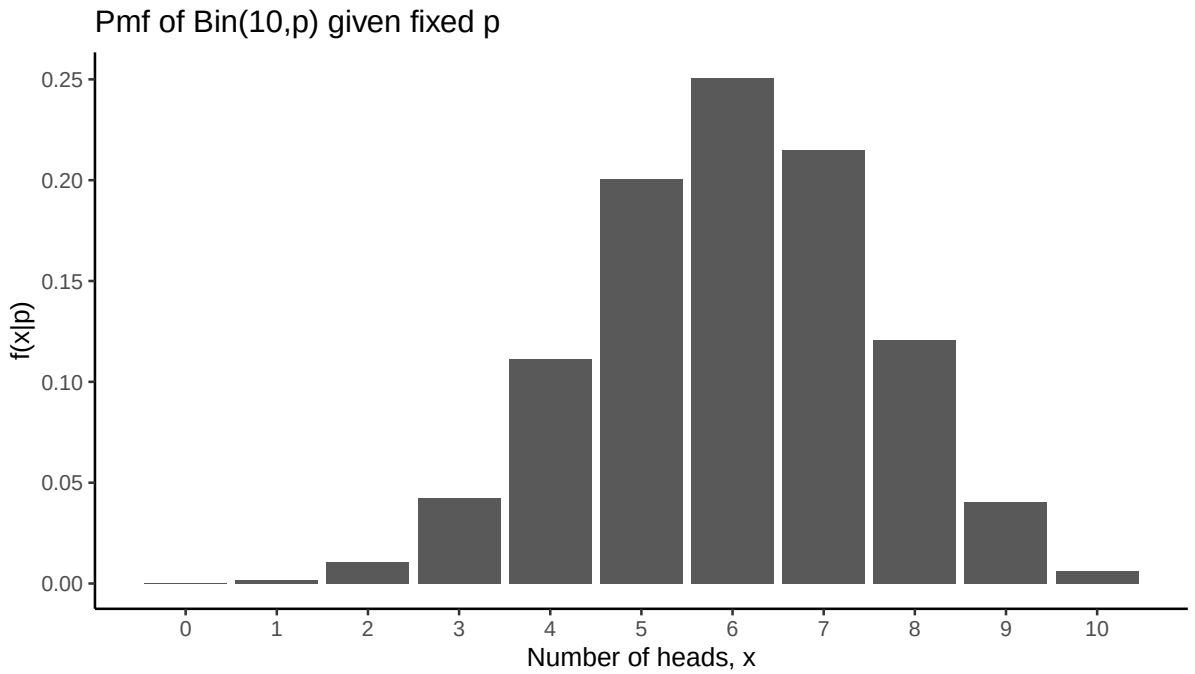
```
dbinom(x = 6, size = 10, prob = seq(0, 1, by = 0.1)) %>%
    round(digits = 4)
```

```
##  [1] 0.0000 0.0001 0.0055 0.0368 0.1115 0.2051 0.2508 0.2001 0.0881 0.0112
## [11] 0.0000
```

Plotting the likelihood, we clearly see how the likelihood is a function of p the parameter (on the horizontal axis), and the function outputs the likelihood on the vertical axis. Moreover, the likelihood function is continuous, even though the underlying data distribution is discrete (binomial). That's because we are evaluating the likelihood of the probability of success p in the continuous interval $[0, 1]$.



And just to reiterate, the probability mass function, when plotted, will show vertical bars or lines because here the input is of the pmf is x, the number of heads, which is clearly discrete.



4.1.2 Likelihood ratio

The term likelihood ratio gives the *relative* plausibility of two candidate parameter values.

Definition 4.1 (Likelihood ratio)**.** The relative plausibility of candidate parameter values, θ_1 and θ_2 say, is measured by the likelihood ratio

$$\frac{L(\theta_1|x)}{L(\theta_2|x)}$$

Interpretation: for example, if $\frac{L(\theta_1|x)}{L(\theta_2|x)} = 10$, then the observed data x were 10 times more likely under truth θ_1 than under truth θ_2.

The use of *likelihood ratios* to compare the plausibility of different θ values means that any constant factor in the likelihood–that is, any factor not depending on θ–can be neglected.

Example 4.2. Suppose $X_i \sim \text{Poi}(\lambda)$ independently $(i = 1, \dots, n)$ and we have observed $X = x$. Here,

$$\begin{aligned} L(\lambda|x) = f(x_1, \dots, x_n|\lambda) &= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \\ &= \text{const.} \times e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i} \end{aligned}$$

- The product $\frac{1}{x_1!} \cdots \frac{1}{x_n!}$ are not needed, since they do not involve λ.
- The non-constant part of the likelihood depends on x only through $T(x) = \sum_{i=1}^{n} x_i$.

As a remark, the function $T(x) = \sum_{i=1}^{n} x_i$ is called a *sufficient statistic* for θ: the value of $T(x)$ is all that is needed in order to compute the likelihood (ignoring constants)!

4.1.3 Log likelihood

In practice, especially when observations are independent, it is usually most convenient to work with the (natural) logarithm of the likelihood,

$$l(\theta) = \log L(\theta|x),$$

since this converts products into sums, which are easier to handle.

Example 4.3. n independent Poisson continued.

$$\begin{aligned} l(\lambda|x) &= \log \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \\ &= \text{const.} - n\lambda + \left(\sum_{i=1}^{n} x_i\right) \log \lambda \end{aligned}$$

In terms of the log likelihood, then, any two candidate values of θ are compared via the log-likelihood-ratio,

$$\log \frac{L(\theta_1|x)}{L(\theta_2|x)} = l(\theta_1) - l(\theta_2)$$

On the log scale, it is additive constants that can be ignored.

4.2 Sufficiency

We have introduced the notion of *sufficient statistic* already, informally, as a data summary that provides all that is needed in order to compute the likelihood. Here we will give a formal definition, and then prove the factorization theorem, which

- provides a straightforward way of checking whether a particular statistic is sufficient
- allows a sufficient statistic, to be identified by simple inspection of the likelihood function (as we did in the example of n Poissons)

Definition 4.2. A statistic $T(X)$ is said to be a sufficient statistic for θ if the conditional distribution of X, given the value of $T(X)$, does not depend on θ.

In this precise sense, a sufficient statistic $T(X)$ carries all of the information about θ that is contained in X. The notion is that, given the observed value $T(x)$ of $T(X)$, all further knowledge about x is uninformative about θ.

In particular, this is useful for data reduction: if $T(X) \in \mathbb{R}$ is a scalar sufficient statistic, then all of the information in $\{X_1, \ldots, X_n\}$ relating to θ is contained in the single-number summary $T(X)$.

4.2.1 The factorisation theorem

It is difficult to use the definition to check if a statistic is sufficient or to find a sufficient statistic. Luckily, there is a theorem that makes it easy to find sufficient statistics.

Theorem 4.1. *A statistic $T(X)$ is sufficient for θ if and only if, for all x and θ,*

$$f(x|\theta) = h(x)g(T(x)|\theta)$$

That is to say, the density f can be factored into a product such that one factor h does not depend on θ, and the other factor, which *does* depend on θ, depends on x only through the sufficient statistic $T(x)$.

Example 4.4. Let $X_1, \ldots, X_n$ be an independent random sample from $\mathrm{N}(\mu, 1)$. The pdf of X can be written

$$\begin{aligned} f(x|\mu) &= \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right) \\ &= \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - \mu)^2\right) \\ &= \underbrace{\frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)}_{h(x)} \quad \underbrace{\exp\left(-\frac{n}{2}(\bar{x} - \mu)^2\right)}_{g(\bar{x}|\mu)} \end{aligned}$$

Therefore, $\bar{X}$ is a sufficient statistic.

Example 4.5. A town has bus routes numbered $1, 2, \ldots, \theta$, with θ being unknown. Naqiyyah spends a day observing bus numbers and collects data X_i, $i = 1, \ldots, n$, representing them.

Each X_i has pmf $f(x|\theta) = \Pr(X = x) = 1/\theta$, so the joint pmf (assuming independence of the observations) is

$$f(x|\theta) = \begin{cases} \frac{1}{\theta^n} & \max(x_1, \ldots, x_n) \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

Hence, if we let $T(x) = \max(x_1, \ldots, x_n)$ then

$$f(x|\theta) = \overbrace{1}^{h(x)} \cdot \frac{\overbrace{\mathbb{1}_{t \leq \theta}(t)}^{g(t|\theta))}}{\theta^n},$$

which implies that $T(X) = \max(X_1, \ldots, X_n)$ is a sufficient statistic for θ.

4.2.2 Minimal sufficient statistic

There clearly is no unique sufficient statistic in any problem. For if $T(X)$ is a scalar sufficient statistic, then, for example

i. $s(T(X))$ is sufficient, for every1-1 function $s(\cdot)$.
ii. The pair $\{T(X), X_1\}$ is sufficient too.
iii. The full data set $\{X_1, \ldots, X_n\}$ is *always* (trivially) sufficient.

Use the factorisation theorem to check these assertions, or convince yourself with suitable examples!

The idea of a *minimal* sufficient statistic is to eliminate redundancy of the kind evident in ii. or iii. (but not i.) above, in order to achieve *maximal* reduction of the data from X to $T(X)$.

Definition 4.3 (Minimal sufficient statistic)**.** A sufficient statistic $S(x)$ is said to be minimal sufficient if, for any other sufficient statistic $T(x)$, $S(X)$ is a function of $T(X)$. I.e., there exists a function k such that $S(x) = k(T(x))$.

Intuitively, a minimal sufficient statistic most efficiently captures all possible information about the parameter θ.

The definition is clear enough in its meaning, but is not constructive: it does not help us to *find* a minimal sufficient statistic in any given situation. For this, we have the following theorem.

Theorem 4.2 (Lehmann-Scheffé)**.** *$T(x)$ is minimal sufficient if for every sample points x and y,*

$$\frac{f(x|\theta)}{f(y|\theta)} \text{ is constant in } \theta \Leftrightarrow T(x) = T(y)$$

Example 4.6. Consider the r.v.s $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Unif}(\theta, \theta + 1)$. The joint pdf of X is

$$f(x|\theta) = \begin{cases} 1 & \theta < x_1, \ldots, x_n < \theta + 1 \\ 0 & \text{otherwise} \end{cases}$$

This can be usefully re-expressed as

$$\begin{aligned} f(x|\theta) &= 1 \cdot \mathbb{1}_{\{x_1, \ldots, x_n > \theta\}}(x)\, \mathbb{1}_{\{x_1, \ldots, x_n < \theta+1\}}(x) \\ &= 1 \cdot \mathbb{1}_{t_1 = \min(x_i) > \theta}(t_1)\, \mathbb{1}_{t_2 = \max(x_i) < \theta+1}(t_2) \\ &= \underbrace{1}_{h(x)} \cdot \underbrace{\mathbb{1}_{\{t_1 > \theta\} \cap \{t_2 < \theta+1\}}}_{g(t_1, t_2|\theta)} \end{aligned}$$

We clearly see that the two-component statistic

$$T(X) = (\min(X_1, \dots, X_n), \max(X_1, \dots, X_n))$$

is sufficient. Furthermore, for any two sample points x and y, $f(x|\theta)/f(y|\theta)$ takes the constant value 1 (for all θ for which the ratio is defined) iff both $\min(x_i) = \min(y_i)$ and $\max(x_i) = \max(y_i)$.

This suggests that $T(X)$ is a minimal sufficient statistic for this problem. Note than then the minimal sufficient statistic in a one-parameter problem is not necessarily a scalar!

Obviously, if a sufficient statistic is scalar, then it must be minimal!

4.3 Point estimators

Consider the following setup. A random sample $X_1, \dots, X_n$ is obtained, such that each value is assumed to be distributed according to the pdf $f(x|\theta)$. Here the function form f is known but the parameter θ of the pdf is unknown. Often, we may specify $\theta \in \Theta$, where Θ is the *parameter space*. Note that θ may be a vector $\theta = (\theta_1, \dots, \theta_p)^\top$.

Example 4.7. Just a simple example explaining parameter spaces–it may be unidimensional or multidimensional.

For $\mathrm{N}(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)^\top$, so $p = 2$ and $\Theta = \mathbb{R} \times \mathbb{R}_{\geq 0}$.

For $\mathrm{Poi}(\lambda)$, $\theta = \lambda$ and $\Theta = \mathbb{R}_{\geq 0}$.

The goal of **point estimation** is the following:

> Provide a single "best guess" of θ, based on observations $X_1, \dots, X_n$.

Formally, we may write

$$\hat{\theta} = T(X_1, \dots, X_n) = T(X)$$

as a point estimator for θ, where $T(X)$ is a statistic. That is to say, we can give a best guess of what θ might be by manipulating the only observable stuff that we know, which is the data $X_1, \dots, X_n$.

We use the term "estimator" to denote the function that gives the estimate. On the other hand, an "estimate" is the realised value of the estimator function. In other words, the estimator $T(X)$ is a *random variable*, whereas the estimate $T(x)$ is a realised value for the observed data $X = x$.

The standard convention is to denote estimators/estimates of parameters with hats on the respective symbols (e.g. $\hat{\theta}$), whereas true values do not have hats (c.f. θ or θ_0).

So how exactly might we come up with an estimator? Surely there are many different ways, but what is clear is that a good estimator should definitely be very close to the true value! That is, a good estimator should make the quantity $|\hat{\theta} - \theta|$ as small as possible, despite

i. the true value θ being unknown; and

ii. the value of $\hat{\theta}$ changes with the sample observed (it is random!).

Luckily, we did study some useful tools with regards to convergences of random variables in the last chapter. In particular, we will make use of the sampling properties of the random variable $\hat{\theta}$ to quantify (and qualify) its worth as an estimator for θ.

In the next parts, we will consider three main aspects of point estimation

1. General methods for *finding* a point estimator

   a. Method of moments (MOM)
   b. Method of maximum likelihood (ML)

2. Methods for *assessing the performance* of point estimators

   a. Bias
   b. Variance
   c. Mean squared error

3. Large sample properties of estimators

4.4 Method of moments

One such method of estimation of population parameters is called the *method of moments*.

Definition 4.4 (Method of moments estimator)**.** Suppose that $U(X)$ is any statistic such that

$$\mathrm{E}\left(U(X)\right) = m(\theta)$$

where $m(\cdot)$ is invertible. Then

$$\hat{\theta} = m^{-1}(U(X))$$

is called the method of moments (MOM) estimator of θ based on U.

The moment here is the mean, i.e. the first moment, of $U(X)$. A more precise name for this estimator would be 'the MOM estimator based on the first moment of U'. As we can see, the key idea is to express population moments (i.e. expectations) as functions of the parameters of interest, which we can then substitute for the sample moments.

While it is common to only use the first moments, there are two main situations where moments other than the first moment are needed:

1. **When $m(\theta)$ either does not involve θ, or is otherwise not invertible.** We might then consider using instead the second moment, $\mathrm{E}(U^2) = m_2(\theta)$, say. If $m_2(\theta)$ is invertible, then MOM based on U_2 can be used in order to define an estimator.

2. **When $\theta = (\theta_1, \ldots, \theta_p)^\top$ is a vector.** I.e., there is more than one unknown parameter. The number of moments used (the number of equations to solve) must be equal to the dimensionality of θ (the number of unknowns).

Let's now look at two examples illustrating the method of moments.

Example 4.8. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{Unif}(0, \theta)$. Consider $U(X) = \bar{X}_n = n^{-1}\sum_{i=1}^n X_i$, the sample mean. Then, since $\mathrm{E}(X_i) = \theta/2$, we have

$$\begin{aligned}\mathrm{E}\left(U(X)\right) &= \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) \\ &= \frac{1}{n}\sum_{i=1}^n \mathrm{E}\left(X_i\right) = \theta/2.\end{aligned}$$

So the MOM estimator of θ based on U is $\hat{\theta} = 2\bar{X}_n$.

```
theta <- 3
(X <- runif(50, min = 0, max = theta)) %>%
    round(3)
```

```
##  [1] 2.871 1.360 2.033 1.718 0.309 2.699 0.738 0.126 0.984 2.864 2.669 2.078
## [13] 1.922 2.983 1.967 2.126 1.632 1.782 0.867 0.441 2.889 2.707 2.072 2.386
## [25] 0.074 1.433 2.275 0.649 0.955 0.695 0.428 1.244 1.241 1.107 0.457 0.416
## [37] 0.699 1.398 0.798 2.573 0.137 1.327 2.397 0.366 1.683 0.620 0.383 2.260
## [49] 2.685 1.123
```

```
2 * mean(X)  # MOM estimator
```

```
## [1] 2.945851
```

Example 4.9. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Poi}(\lambda)$. Consider, for example, $U(X) = \sum_{i=1}^{n} [X_i = 0]$, the number of zeroes found in the sample. Then, since

$$\begin{aligned} \text{E}([X_i = 0]) &= \sum_{k=0}^{\infty} [k = 0] \Pr(X_i = k) \\ &= \Pr(X_i = 0) \\ &= e^{-\lambda}, \end{aligned}$$

we have that $\text{E}\left(U(X)\right) = \sum_{i=1}^{n} \text{E}([X_i = 0]) = ne^{-\lambda}$. Hence, the MOM estimator for λ based on U is

$$\hat{\lambda} = -\log(U/n).$$

4.5 Method of maximum likelihood

The other point estimation method we will be studying is the maximum likelihood estimator. As the name implies, this uses the likelihood function.

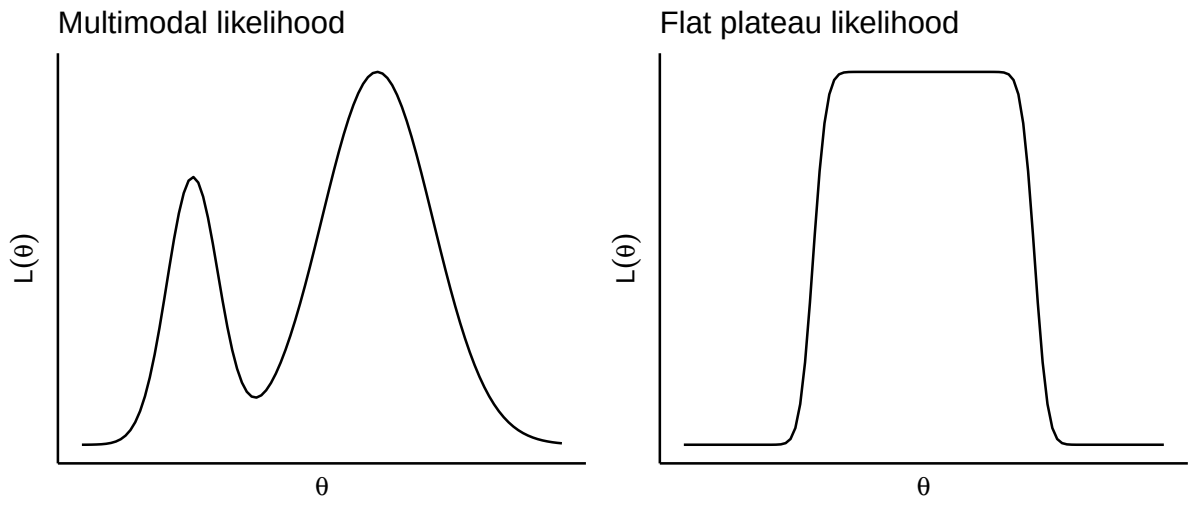
Definition 4.5 (Maximum likelihood (ML) estimator)**.** The ML estimator of θ is $\hat{\theta}$ which is such that

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta|X)$$

The ML estimator is the value of θ which is the most likeliest value as judged by the likelihood function, given the data that was observed. Geometrically, we are interested in the peak of the graph of $L(\theta|X)$ against θ.

In practice, $\hat{\theta}$ is most often found by locating the maximum of the *log-likelihood* $l(\theta|X) = \log L(\theta|X)$, which is computationally and algebraically simpler. The main reason for this is products of likelihoods become sums of log-likelihoods, which is arguably easier to deal with.

Unfortunately, uniqueness is not guaranteed. But in many 'standard' statistical models, the MLE *is* uniquely defined by the likelihood function.



4.5.1 Finding the MLE

To find the maxima of any graph, you might recall that such a point is called a stationary point and is obtained by differentiating. In other words, we locate $\hat{\theta}$ by solving $l'(\hat{\theta}) = 0$, and then checking that the stationary point is a maximum. However, several points of note must be said on this topic:

- This still leaves open the possibility that the likelihood has multiple local maxima, at each of which the derivative is zero. It is wise to check (θ) for multimodal behaviour, e.g. by drawing a sketch of the function.
- This strategy works for 'simple' enough problems, e.g. unidimensional parameters, or multidimensional parameter situations which reduce to complete information system (sets of simultaneous equations).
- Numerical methods can be employed if explicit analytical forms for the MLE cannot be found. These estimators are found more often by iterative procedures built into computer software (e.g. Newton-Raphson, Fisher scoring, quasi-Newton, gradient descent, conjugate gradients, etc.).
- Even then we might run into numerical issues (e.g. flat likelihood, multimodality, precision issues, etc.).

Let us now consider a simple example of finding the MLE of the mean of a normal random sample.

Example 4.10. Suppose that $Y_1, \ldots, Y_n$ is an iid random sample from $\mathrm{N}(\mu, 1)$, with μ unknown. Then, the log-likelihood function is

$$\begin{aligned} l(\mu) &= \log\left\{(\sqrt{2\pi})^{-n} e^{-\sum_{i=1}^{n}(Y_i-\mu)^2/2}\right\} \\ &= \text{const.} - \frac{1}{2}\sum_{i=1}^{n}(Y_i - \mu)^2 \end{aligned}$$

The derivative with respect to μ gives us

$$l'(\mu) = \frac{1}{2}\sum_{i=1}^{n}(Y_i - \mu)$$

Equating this to zero gives the MLE for μ

$$\begin{aligned} \frac{1}{2}\sum_{i=1}^{n}(Y_i - \mu) &= 0 \\ \sum_{i=1}^{n} Y_i - n\mu &= 0 \\ \Rightarrow \mu = \frac{1}{n}\sum_{i=1}^{n} Y_i &=: \bar{Y}_n \end{aligned}$$

Thus, $\hat{\mu} = \bar{Y}_n$.

Finding the MLE numerically using R involves some kind of optimising function. In R, we can use the function called `optim()`. We first construct a function called `lik()` which is essentially a function whose input is the parameter value `theta` (μ in this case). The `optim()` function then searches for the optima by way of quasi-Newton methods (using pseudo/approximate derivatives).

```
X <- rnorm(n = 100, mean = 8, sd = 1)
mean(X)
```

```
## [1] 8.021617
```

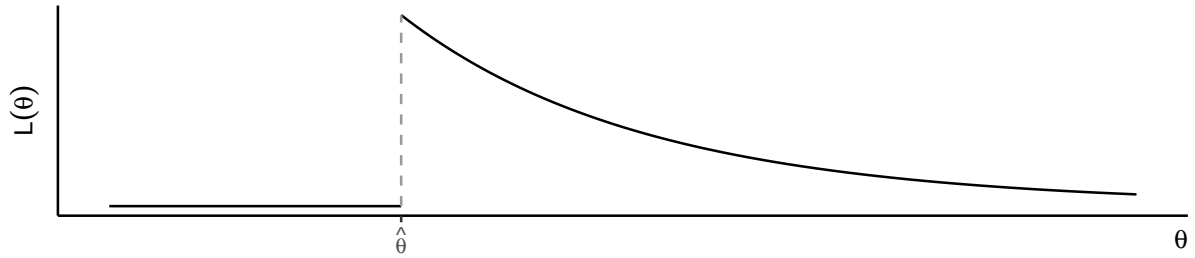
```r
# Optimising the likelihood function
lik <- function(theta) -sum(dnorm(x = X, mean = theta, sd = 1, log = TRUE))
theta0 <- 1  # starting value
res <- optim(par = theta0, fn = lik, method = "BFGS", lower = -Inf,
             upper = Inf)
res$par
```

```
## [1] 8.021617
```

Sometimes, a sketch of $l(\theta)$ reveals that the MLE does not satisfy $l'(\hat{\theta}) = 0$. Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Unif}(0, \theta)$. Then the pdf of X is $f(x|\theta) = 1/\theta^n$ for $X_1, \dots, X_n < \theta$. The likelihood is therefore

$$L(\theta|X) = \begin{cases} \frac{1}{\theta^n} & \theta > \max(X_1, \dots, X_n) \\ 0 & \text{otherwise} \end{cases}$$

which is maximised at $\hat{\theta} = \max(X_1, \dots, X_n)$.



4.5.2 Invariance of MLE

The MLE is invariant under parameter transformation:

Lemma 4.1 (Invariance of MLE). *Suppose $X \sim f(x|\theta)$, and $\psi = \psi(\theta)$ is a one-to-one transformation. Let $\hat{\theta}$ be the MLE for θ, i.e.*

$$\hat{\theta} = \arg\max_{\theta} L(\theta|X).$$

Then, the MLE for ψ is

$$\hat{\psi} = \psi(\hat{\theta}).$$

Example 4.11. Let $\hat{\pi}$ be the MLE for π after observing data $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Bern}(\pi)$. The log-odds of an event happening is given by $\nu = \log(\pi/\log(1-\pi))$, which is a one-to-one transformation of π. Therefore, the MLE for ν is given by

$$\hat{\nu} = \log \frac{\hat{\pi}}{1 - \hat{\pi}}.$$

Note that $\hat{\psi}$ can be infinite-valued, if $\hat{\theta} = 0$ or $\hat{\theta} = 1$.

This is extremely useful! That means, if we already found the MLE $\hat{\theta}$ for θ, then the MLE of any function of the parameter $g(\theta)$ is simply $g(\hat{\theta})$. We do not need to find the MLE again by reparameterising the likelihood function.

4.6 Evaluating estimators

An estimator is assessed through its distribution in repeated sampling from the assumed model. A 'good' estimator of an unknown parameter θ is a function $T(X)$ which typically, in repeated sampling, takes values that are close to the true value of θ, whatever the true value of θ may be. We discuss three such properties:

1. Bias
2. Variance
3. Mean squared error

4.6.1 Bias

Bias is a measure of the typical deviation of the estimator away from its true value.

Definition 4.6 (Bias)**.** The bias of an estimator $\hat{\theta}$ is defined to be

$$\text{Bias}_\theta(\hat{\theta}) = \text{E}_\theta(\hat{\theta} - \theta) - \theta = \text{E}_\theta(\hat{\theta}) - \theta.$$

Note that the random variable in the expectation is actually $\hat{\theta}$. Thus, the distribution of $\hat{\theta}$ must be known (the pdf $f(\hat{\theta})$ must be known) in order to calculate the expectation. The subscript θ makes clear the fact that the expectation is taken under the distribution using θ as the true value of the parameter. That is,

$$\text{Bias}_\theta(\hat{\theta}) = \int (\hat{\theta} - \theta) f(\hat{\theta} \mid \theta) \, \text{d}\hat{\theta}.$$

The specific case that occurs when $\text{E}_\theta(\hat{\theta}) = \theta$ results in $\text{Bias}_\theta(\hat{\theta}) = 0$ for all possible values of θ. We call $\hat{\theta}$ an *unbiased estimator* for θ in such cases.

Definitely small bias for estimators is desirable, especially more so unbiased estimators.

4.6.2 Variance and standard error

As $\hat{\theta}$ is a random variable, we could calculate its variance.

Definition 4.7 (Variance)**.** The variance of an estimator $\hat{\theta}$ is defined to be

$$\text{Var}_\theta(\hat{\theta}) = \text{E}_\theta\left[(\hat{\theta} - \text{E}(\hat{\theta}))^2\right]$$

This just uses the regular definition of the variance for random variables,

$$\text{Var}_\theta(\hat{\theta}) = \int (\hat{\theta} - \text{E}\,\hat{\theta})^2 f(\hat{\theta} \mid \theta) \, \text{d}\hat{\theta}.$$

It also has the same interpretation as the variance of random variables, in that it gives a sense of the spread/scale of this estimator. Furthermore, we could look at the 'standard deviation', which has a special name when it comes to estimators.

Definition 4.8 (Standard error)**.** The standard error of the estimator $\hat{\theta}$ is defined as the standard deviation of the variance of the estimator, i.e.

$$\text{se}(\hat{\theta}) = \sqrt{\text{Var}_\theta(\hat{\theta})}$$

Obviously, we desire an estimator whose variability (in repeated sampling) is low. That is, an estimator whose calculated value is not too variable across multiple sampling scenarios would be really great to have, compared to an estimator which keeps on changing in value drastically.

These two properties measure different things about estimators:

- Bias is a measure of *accuracy*.
- Variance is a measure of *precision*.

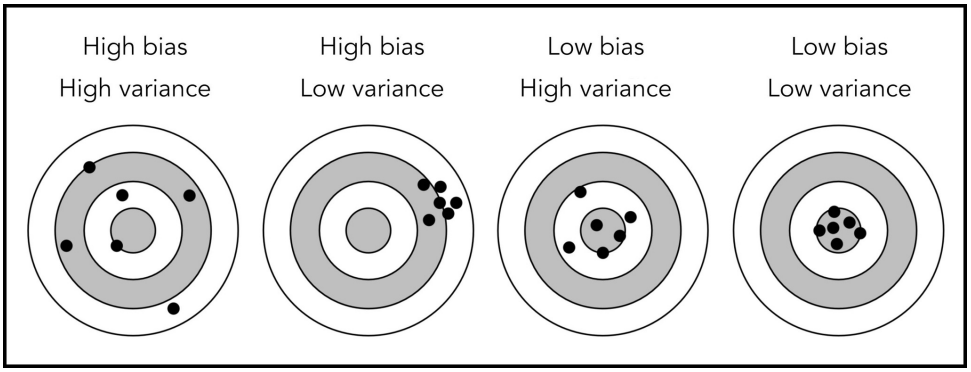


Figure 4.1: The difference between bias and variance.

4.6.3 Mean squared error

The bias and variance are related through the following concept of mean squared error.

Definition 4.9 (Mean squared error of estimator)**.** The MSE of the estimator $\hat{\theta}$ is defined as

$$\text{MSE}_\theta(\hat{\theta}) = \text{E}_\theta\left[(\hat{\theta}-\theta)^2\right] = \{\text{Bias}_\theta(\hat{\theta})\}^2 + \text{Var}_\theta(\hat{\theta}).$$

As an exercise, prove the bias-variance decomposition above. Here some hints on how to get started:

- Add and subtract the expectation of $\hat{\theta}$ in the expression:

$$\left(\hat{\theta} - \text{E}\,\hat{\theta} + \text{E}\,\hat{\theta} - \theta\right)^2$$

  and expand the square.
- Try to show that $\text{E}\left[(\hat{\theta} - \text{E}\,\hat{\theta})(\text{E}\,\hat{\theta} - \theta)\right] = 0$ using properties of expectations.

There is a clear and direct relationship between the MSE of an estimator, and its bias and variance. Note that for a fixed MSE value,

- Reducing the bias of an estimator necessarily increases its variance.
- Conversely, reducing the variance of an estimator implies that bias will increase.

This is known as the **bias-variance** trade-off. It is typically impossible to do both simultaneously.

The bias-variance trade-off does not mean an estimator with low bias **and** low variance is impossible to achieve.It simply means *improving* one aspect of an estimator will worsen it in the other aspect.

Example 4.12. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Unif}(0, \theta)$. We previously found two different estimators for θ:

- MLE: $\hat{\theta}_{ML} = \max_i(X_i)$
- MOM: $\hat{\theta}_{MOM} = 2\bar{X}$

Let us examine these in terms of bias, variance and mse.

Clearly $\hat{\theta}_{MOM}$ is unbiased: $\text{E}(\hat{\theta}_{MOM}) = 2\,\text{E}(\bar{X}) = 2\,\text{E}(X_i) = 2 \times \theta/2 = \theta$.

For the bias of $\hat{\theta}_{ML}$, let's first get the pdf (of $\hat{\theta}_{ML}$). Proceed via the cdf:

$$F_{\hat{\theta}_{ML}}(x) = \Pr(\hat{\theta}_{ML} < x) = \Pr(\max(X_1, \ldots, X_n) < x) = \prod_{i=1}^{n} \Pr(X_i < x) = \left(\frac{x}{\theta}\right)^n$$

Then, differentiating this gives us the pdf $f_{\hat{\theta}_{ML}}(x) = nx^{n-1}/\theta^n$. So now we find the mean:

$$\mathrm{E}\,\hat{\theta}_{ML} = \int_0^\theta x\frac{nx^{n-1}}{\theta^n}\,\mathrm{d}x = \left[\frac{nx^{n+1}}{\theta^n(n+1)}\right]_0^\theta = \frac{n\theta}{n+1}$$

Therefore, the bias is $\mathrm{Bias}(\hat{\theta}_{ML}) = -\theta/(n+1) \neq 0$. Note that this tends to 0 as $n \to \infty$, but can be substantial when n is small.

For $\hat{\theta}_{MOM}$, we have

$$\mathrm{Var}(\hat{\theta}_{MOM}) = 4\,\mathrm{Var}(\bar{X}) = \frac{4\,\mathrm{Var}(X_i)}{n} = \frac{\theta^2}{3n}.$$

Note that $\mathrm{Var}(\hat{\theta}_{MOM}) \to 0$ as $n \to \infty$ at the rate of $1/n$. This is typical behaviour of 'good' estimators.

For $\hat{\theta}_{ML}$:

$$\begin{aligned}\mathrm{Var}(\hat{\theta}_{ML}) = \mathrm{E}(\hat{\theta}_{ML}^2) - \mathrm{E}^2(\hat{\theta}_{ML}) &= \int_0^\theta x^2\frac{nx^{n-1}}{\theta^n}\,\mathrm{d}x - \frac{n^2\theta^2}{(n+1)^2} \\ &= \theta^2\left(\frac{n}{(n+1)^2(n+2)}\right)\end{aligned}$$

so $\mathrm{Var}(\hat{\theta}_{ML}) \to 0$ as $n \to \infty$ but at a faster rate of $1/n^2$.

For $\hat{\theta}_{MOM}$, we have

$$\mathrm{MSE}(\hat{\theta}_{MOM}) = \mathrm{Var}(\hat{\theta}_{MOM}) = \frac{\theta^2}{3n}.$$

For $\hat{\theta}_{ML}$:

$$\mathrm{MSE}(\hat{\theta}_{ML}) = \mathrm{Bias}(\hat{\theta}_{ML}^2) + \mathrm{Var}(\hat{\theta}_{ML}) = \theta^2\left(\frac{n}{(n+1)^2(n+2)} + \frac{1}{(n+1)^2}\right).$$

Notice that since $\mathrm{MSE}(\hat{\theta}_{MOM})$ is $o(1/n)$ and $\mathrm{MSE}(\hat{\theta}_{ML})$ is $o(1/n^2)$,

$$\mathrm{MSE}(\hat{\theta}_{ML}) \leq \mathrm{MSE}(\hat{\theta}_{MOM})$$

for all n (and tends to 0 as $n \to \infty$). So $\hat{\theta}_{ML}$, even though it is biased, is clearly to be preferred on the basis of MSE.

4.7 Cramér-Rao lower bound (CRLB)

It is difficult to find an estimator which simultaneously is low in bias and variance. If we instead focus on a class of *unbiased* estimators, then we have a theorem to benchmark their performance.

Theorem 4.3 (Cramér-Rao inequality for unbiased estimators)**.** *Let $X \sim f(x|\theta)$ satisfying some regularity conditions*[1] *Let $\hat{\theta} = \hat{\theta}(X)$ be an **unbiased** estimator, i.e. $\mathrm{E}_\theta(\hat{\theta}) = \theta$. Then, for any $\theta \in \Theta$,*

$$\mathrm{Var}_\theta(\hat{\theta}) \geq \frac{1}{\mathrm{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right]}.$$

[1] These regularity conditions are essentially that we are able to switch the order of integration and differentiation, and that the $\mathrm{Var}_\theta(\hat{\theta}) < \infty$. See Thm 7.3.9 of C&B.

If an estimator's variance is close to the CRLB, it can be regarded as *efficient*. A class of estimators achieving the CRLB are said to be *optimal*, known as the *minimum variance unbiased estimator (MVUE)*. Although, the CRLB is not necessarily achieved by any estimator.

Proof. The proof is an application of the Cauchy-Schwarz inequality via the covariance inequality

$$\mathrm{Var}(Y) \geq \frac{\{\mathrm{Cov}(Y,U)\}^2}{\mathrm{Var}(U)}$$

for r.v.s U and Y. We consider the more general case for **biased** estimators $\hat{\theta}(X)$. Let

$$\begin{aligned} U &= l'(\theta) = \frac{\partial}{\partial\theta}\log f(X|\theta) \\ Y &= \hat{\theta}(X) \end{aligned}$$

Firstly, we note that $\mathrm{Var}(U) = \mathrm{E}(U^2)$ since $\mathrm{E}(U) = 0$:

$$\mathrm{E}(U) = \int \frac{\frac{\partial}{\partial\theta}f(x|\theta)}{\cancel{f(x|\theta)}}\cancel{f(x|\theta)}\,\mathrm{d}x = \frac{\partial}{\partial\theta}\int f(x|\theta)\,\mathrm{d}x = 0.$$

Further, because $\mathrm{E}(U) = 0$, we have $\mathrm{Cov}(Y,U) = \mathrm{E}(UY) - \mathrm{E}(U)\,\mathrm{E}(Y) = \mathrm{E}(UY)$, and so

$$\begin{aligned} \mathrm{Cov}(Y,U) = \mathrm{E}\left(Y\cdot\frac{\partial}{\partial\theta}\log f(X|\theta)\right) &= \mathrm{E}\left(Y\cdot\frac{\frac{\partial}{\partial\theta}f(X|\theta)}{f(X|\theta)}\right) \\ &= \int \hat{\theta}(x)\frac{\frac{\partial}{\partial\theta}f(x|\theta)}{\cancel{f(x|\theta)}}\cancel{f(x|\theta)}\,\mathrm{d}x \\ &= \frac{\partial}{\partial\theta}\left[\int \hat{\theta}(x)f(x|\theta)\,\mathrm{d}x\right] \\ &= \frac{\partial}{\partial\theta}\,\mathrm{E}[\hat{\theta}(X)] = \psi'(\theta). \end{aligned}$$

Here, we have assumed that the expectation of $\hat{\theta}(X)$ is not θ but some function of θ, $\psi(\theta)$ say, since the estimator is biased.

We have now proved the general case of the CRLB which states

$$\mathrm{Var}(\hat{\theta}) \geq \frac{[\psi'(\theta)]^2}{\mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right]}.$$

For unbiased estimators, $\psi(\theta) = \theta$, and hence

$$\psi'(\theta) = \frac{\partial}{\partial\theta}(\theta) = 1,$$

which completes the proof. □

Remark: The derivative of the log-likelihood, $S(\theta) = \frac{\partial}{\partial\theta}\log f(X|\theta)$, is known as the *score function*. The property that $\mathrm{E}(S(\theta)) = 0$ is fundamental to the theory of maximum likelihood.

4.7.1 Fisher information

The quantity in the RHS denominator of Theorem 4.3 is known as the *information number* or *Fisher information*.

Definition 4.10 (Fisher information (unidimensional))**.** Let $X \sim f(x|\theta)$, where $\theta \in \mathbb{R}$. The Fisher information is defined to be the expectation of the second moment of the score function, i.e.

$$\mathcal{I}(\theta) = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right] \in \mathbb{R}$$

In simple terms, the Fisher information measures the amount of information that an observable random variable X carries about an unknown parameter θ of the statistical model that models X.

The Fisher information for multidimensional parameters can be defined similarly (c.f. Fisher information matrix).

Lemma 4.2. *Let $X \sim f(x|\theta)$, where $\theta \in \mathbb{R}$, and $S(\theta) = \frac{\partial}{\partial\theta}\log f(X|\theta)$. Under certain regularity conditions,*

- $\mathrm{E}[S(\theta)] = 0$.
- $\mathcal{I}(\theta) = \mathrm{Var}[S(\theta)]$.
- $\mathcal{I}(\theta) = -\mathrm{E}[S'(\theta)]$.

To be proven in Ex. sheet 4!

Lemma 4.3 (Fisher information is additive)**.** *Let $X_1, \ldots, X_n \overset{iid}{\sim} f(x|\theta)$. Suppose $\mathcal{I}_1(\theta)$ is the Fisher information from a single observation X_i, i.e. $\mathcal{I}_1(\theta) = -\mathrm{E}[l''(\theta|X_i)]$. Then the full Fisher information is $\mathcal{I}(\theta) = n\mathcal{I}_1(\theta)$.*

Proof.

$$\mathcal{I}(\theta) = -\mathrm{E}[l''(\theta|X)] = -\mathrm{E}\left[\sum_{i=1}^{n} l''(\theta|X_i)\right]$$

□

Example 4.13. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$. Let $\hat{\mu} = \bar{X}_n$; then $\mathrm{Var}(\hat{\mu}) = \sigma^2/n$. The score function is given as

$$l'(\mu|X) = \frac{\partial}{\partial\mu}\left(\text{const.} - \sum_{i=1}^{n}\frac{(X_i-\mu)^2}{2\sigma^2}\right) = \sum_{i=1}^{n}\frac{X_i-\mu}{\sigma^2},$$

while the Fisher information is obtained as

$$\mathcal{I}(\mu) = \mathrm{Var}\left(l'(\mu|X)\right) = \sum_{i=1}^{n}\mathrm{Var}\left(\frac{X_i-\mu}{\sigma^2}\right) = \sum_{i=1}^{n}\frac{\mathrm{Var}(X_i)}{\sigma^4} = \frac{n}{\sigma^2}.$$

Hence, the CRLB is σ^2/n, and the estimator $\hat{\mu} = \bar{X}_n$ achieves it. Therefore, $\bar{X}_n$ is the MVUE of μ.

4.7.2 Variance reduction: Rao-Blackwellisation

We can reduce the variance of an unbiased estimator by conditioning on a sufficient statistic.

Theorem 4.4 (Rao-Blackwell)**.** *Suppose that $U(X)$ is unbiased for θ, and $S(X)$ is sufficient for θ. Then the function of S defined by*

$$\phi(S) = \mathrm{E}_\theta(U|S)$$

- *is a statistic, i.e. $\phi(S)$ does not involve θ;*
- *is an unbiased statistic, i.e. $\mathrm{E}(\phi(S)) = \theta$; and*
- *has $\mathrm{Var}_\theta(\phi(S)) \leq \mathrm{Var}_\theta(U)$, with equality iff U is itself a function of S.*

In other words, $\phi(S)$ is a uniformly better unbiased estimator for θ. Thus the Rao-Blackwell theorem provides a systematic method of variance reduction for an estimator that is not a function of the sufficient statistic.

Proof. Since S is sufficient, the distribution of X given S does not involve θ, and hence $\mathrm{E}_\theta(U(X)|S)$ does not involve θ. Further, $\mathrm{E}(\phi(S)) = \mathrm{E}\left[\mathrm{E}(U|S)\right] = \mathrm{E}(U) = \theta$.

To prove the last part, note that

$$\begin{aligned}\mathrm{Var}(U) &= \mathrm{E}\left[\mathrm{Var}(U|S)\right] + \mathrm{Var}\left[\mathrm{E}(U|S)\right] \\ &= \mathrm{E}\left[\mathrm{Var}(U|S)\right] + \mathrm{Var}(\phi(S)) \\ &\geq \mathrm{Var}(\phi(S))\end{aligned}$$

with equality iff $\mathrm{Var}(U|S) = 0$, i.e. iff U is a function of S. □

Example 4.14. Suppose we have data $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{Poi}(\lambda)$ pertaining to the number of road accidents per day, and we want to estimate the probability of having no accidents $\theta = e^{-\lambda} = \Pr(X_i = 0)$.

An unbiased estimator of θ is

$$U(X) = \begin{cases} 1 & X_1 = 0 \\ 0 & \text{otherwise} \end{cases}$$

But this is likely to be a poor estimator, since it ignores $X_2, X_3, \ldots, X_n$.

We can see that $S(X) = \sum_{i=1}^n X_i$ is sufficient since the joint pdf can be expressed as

$$f(x|\lambda) = \frac{1}{x_1! \cdots x_n!} \cdot e^{-n\lambda} \lambda^{\sum_{i=1}^n x_i}.$$

Now apply the Rao-Blackwell theorem:

$$\begin{aligned}\phi(S) = \mathrm{E}(U|S) = \mathrm{E}\left(U \,\middle|\, \sum_{i=1}^n X_i = S\right) &= \Pr\left(X_1 = 0 \,\middle|\, \sum_{i=1}^n X_i = S\right) \\ &= \left(1 - \frac{1}{n}\right)^S,\end{aligned}$$

where the conditional probability in the last step comes from the Poisson-binomial relationship (see Ex sheet 2, Q11: Suppose $X_i \overset{\text{iid}}{\sim} \mathrm{Poi}(\lambda_i)$, then $X_1 \big| (\sum_{i=1}^n X_i = N) \sim \mathrm{Bin}(N, \pi)$, where $\pi = \lambda_1 / \sum_{i=1}^n \lambda_i$).

By the Rao-Blackwell theorem, $\mathrm{Var}(\phi) < \mathrm{Var}(U)$ (strict inequality since U is not a function of S), so prefer $\phi(S)$ over U as an estimator.

But is $\phi(S) = (1 - 1/n)^S$ unbiased? This is guaranteed by the RB theorem. Check: Since $S \sim \mathrm{Poi}(n\lambda)$ (sum of Poisson r.v.s is Poisson), we get

$$\begin{aligned}\mathrm{E}(\phi(S)) &= \sum_{s=0}^{\infty} \left(1 - \frac{1}{n}\right)^s \frac{e^{-n\lambda}(n\lambda)^s}{s!} \times e^{-\lambda} e^{\lambda} \\ &= e^{-\lambda} \sum_{s=0}^{\infty} \underbrace{\frac{e^{-\lambda(n-1)}[\lambda(n-1)]^s}{s!}}_{\text{pmf of } \mathrm{Poi}(\lambda(n-1))} = e^{-\lambda}.\end{aligned}$$

A similar calculation can give us the variance of this estimator.

4.8 Large sample properties of estimators

All of the criteria we have considered thus far have been finite-sample criteria. In contrast, we might consider asymptotic properties which describe the behaviour as sample size becomes infinite.

We shall discuss three properties:

1. Consistency
2. Efficiency
3. Asymptotic normality

In particular, we shall see that ML estimators are (generally) consistent, efficient (achieves CRLB), and has an asymptotic normal distribution.

4.8.1 Consistency

Definition 4.11 (Consistent estimator)**.** An estimator $\hat{\theta}_n := \hat{\theta}(X_1, \dots, X_n)$ is a consistent estimator for θ if $\hat{\theta}_n \to \theta$ in probability as $n \to \infty$.

Consistency is a natural condition for a reasonable estimator as $\hat{\theta}_n$ should converge to θ if we have a (theoretically) infinite amount of information. Therefore, a **non-consistent estimator should not be used in practice!**

A practical way of checking consistency is to check mean square convergence: If $\hat{\theta}_n \xrightarrow{m.s.} \theta$ then $\hat{\theta}_n$ is consistent (since convergence in mean square implies convergence in probability). Further, since

$$\mathrm{MSE}(\hat{\theta}_n) = \mathrm{E}\left[(\hat{\theta}_n - \theta)^2\right] = \left\{\mathrm{Bias}(\hat{\theta}_n)\right\}^2 + \mathrm{Var}(\hat{\theta}_n),$$

we can also check that both the bias and variance converges to 0.

4.8.2 Consistency vs unbiasedness

Consistency and bias are two distinct concepts:

- Unbiasedness ($\mathrm{E}(\hat{\theta}_n) = \theta$) is a statement about the expected value of the *sampling distribution* of the estimator.
- Consistency ($\mathrm{plim}_{n\to\infty} \hat{\theta}_n = \theta$) is a statement relating to the sequence of estimators $\hat{\theta}_1, \hat{\theta}_2, \dots$. It tells us where the estimator is tending to as the sample size increases.

Both are desirable properties of estimators, though it might be possible for one to be satisfied but not the other (see next example). As mentioned, and as we shall see, we are probably better off using a consistent but biased estimator rather than an inconsistent but unbiased estimator.

Example 4.15. Let $X_1, \dots, X_n$ be a sample from $\mathrm{N}(\mu, \sigma^2)$. Consider the following estimators for μ and σ^2:

- $\hat{\mu} = X_1$; and
- $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$

The estimator $\hat{\mu}$ is unbiased since $\mathrm{E}(X_1) = \mu$, but it is not consistent since the distribution of $\hat{\mu}$ is always $\mathrm{N}(\mu, \sigma^2)$ and will never concentrate around μ even with infinite sample size.

It is a fact that $\mathrm{E}(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2$, which shows that $\hat{\sigma}^2$ is biased in finite samples, but this bias vanishes as $n \to \infty$. We can also show

$$\mathrm{Var}(\hat{\sigma}^2) = \frac{2\sigma^4(n-1)}{n^2} \to 0$$

as $n \to \infty$. Therefore, $\mathrm{MSE}(\hat{\sigma}^2) \to 0$, and $\hat{\sigma}^2$ is therefore consistent.

4.8.3 Consistency of MLEs

Theorem 4.5 (Consistency of MLE)**.** *Let $X_1, \dots, X_n \overset{iid}{\sim} f(x|\theta)$, and let $\hat{\theta}_n := \arg\max_\theta L(\theta|X)$ denote the MLE of θ. Let $\psi(\theta)$ be a continuous function of θ. Under certain regularity conditions, we have that for every $\epsilon > 0$ and every $\theta \in \Theta$,*

$$\lim_{n\to\infty} \Pr(|\psi(\hat{\theta}_n) - \psi(\theta)| \geq \epsilon) = 0.$$

That is, $\psi(\hat{\theta}_n)$ is a consistent estimator of $\psi(\theta)$.

In particular, consider the identity function $\psi(\theta) = \theta$. Then the theorem states that the MLE $\hat{\theta}_n$ is consistent. Some notes:

- The regularity conditions mentioned can be found in Miscellanea 10.6.2 of C&B.
- The above theorem is stating the result for unidimensional θ, but there are similar multidimensional statements too.
- We shall defer the proof until we discuss asymptotic normality.

4.8.4 Efficiency

Efficiency of an estimator concerns the (asymptotic) variance of an estimator. The CRLB gives the benchmark for efficiency.

Definition 4.12 (Asymptotic efficiency)**.** A sequence of estimators $\hat{\theta}_n := \hat{\theta}(X_1, \dots, X_n)$ is said to be asymptotically efficient for a parameter θ if

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathrm{D}} \mathrm{N}\left(0, v(\theta)\right),$$

as $n \to \infty$, where $v(\theta)$ is the Cramér-Rao lower bound

$$v(\theta) = \frac{1}{\mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X_1|\theta)\right)^2\right]} = \mathcal{I}_1(\theta)^{-1}.$$

Some remarks:

- The property that $a_n(\hat{\theta}_n - \theta)$ converges in distribution to $\mathrm{N}(0, \sigma^2)$ is called *asymptotic normality*, and σ^2 is called the *asymptotic variance*.
- An asymptotically efficient estimator has its asymptotic variance achieving the CRLB.

4.8.5 Asymptotic normality and consistency

The phrase 'efficient and consistent' is somewhat redundant, because efficiency is defined only when the estimator is asymptotically normal, and as we shall show, asymptotic normality implies consistency.

Lemma 4.4. *Suppose that $\hat{\theta}_n$ is an estimator for θ such that*

$$\frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sigma} \xrightarrow{D} \mathrm{N}(0, 1)$$

then $\hat{\theta}_n$ is consistent for θ.

Proof. Notice that

$$\hat{\theta}_n - \theta = \frac{\sigma}{\sqrt{n}} \frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sigma} \xrightarrow{\mathrm{D}} 0$$

by Slutzky's theorem. Thus, $\hat{\theta}_n - \theta \xrightarrow{\mathrm{P}} 0$ which implies $\hat{\theta}_n \xrightarrow{\mathrm{P}} \theta$, and hence $\hat{\theta}_n$ is consistent. □

4.8.6 Efficiency of MLE

We've seen that MLEs are consistent. Under even stronger regularity conditions, we find that they are also efficient.

Theorem 4.6 (Asymptotic efficiency of MLE)**.** *Let $X_1, \dots, X_n \overset{iid}{\sim} f(x|\theta)$, and let $\hat{\theta}_n := \arg\max_\theta L(\theta|X)$ denote the MLE of θ. Under certain regularity conditions, we have that*

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{D} \mathrm{N}\left(0, \mathcal{I}_1(\theta)^{-1}\right),$$

where $\mathcal{I}_1(\theta)$ is the (unit) Fisher information for θ. That is, $\hat{\theta}_n$ is a consistent and asymptotically efficient estimator for θ.

In fact, this theorem also holds more widely–the restriction to iid cases presents a simple proof, but is not essential.

Sketch. Taylor expand the score $l'(t|X)$ about the parameter value θ:

$$l'(t|X) = l'(\theta|X) + (t-\theta)l''(\theta|X)$$

(ignoring the higher order terms). Evaluate this at the maxima $t = \hat{\theta}_n$, we get

$$\begin{aligned} \cancelto{0}{l'(\hat{\theta}_n|X)} &= l'(\theta|X) + (\hat{\theta}_n - \theta)l''(\theta|X) \\ \Rightarrow \sqrt{n}(\hat{\theta}_n - \theta) &= -\frac{\frac{1}{\sqrt{n}}l'(\theta|X)}{\frac{1}{n}l''(\theta|X)} \end{aligned}$$

As one of the exercises at the end of this chapter, you will show that

$$-\frac{1}{\sqrt{n}}l'(\theta|X) \xrightarrow{\mathrm{D}} \mathrm{N}\left(0, \mathcal{I}_1(\theta)\right)$$

and

$$\frac{1}{n}l''(\theta|X) \xrightarrow{\mathrm{P}} \mathcal{I}_1(\theta),$$

Using Slutzky's theorem, we get

$$\sqrt{n}(\hat{\theta}_n - \theta) = -\frac{\frac{1}{\sqrt{n}}l'(\theta|X)}{\frac{1}{n}l''(\theta|X)} \xrightarrow{\mathrm{D}} \mathrm{N}\left(0, \mathcal{I}_1(\theta)^{-1}\right)$$

□

4.8.7 Efficiency of transformations of MLE

Let $\psi(\theta)$ be a continuous function of θ. Using the delta method, the following result can be obtained:

$$\sqrt{n}(\psi(\hat{\theta}_n) - \psi(\theta)) \xrightarrow{\mathrm{D}} \mathrm{N}\left(0, |\psi'(\theta)|^2 v(\theta)\right).$$

This is assuming that $\psi(\cdot)$ is differentiable at the value θ.

Therefore, the transformed MLE $\psi(\hat{\theta})$ is a consistent and asymptotically efficient estimator of $\psi(\theta)$. Look back to the proof of the CRLB above and notice that the asymptotic variance of $\psi(\hat{\theta}_n)$ is exactly the general version of the CRLB (using the unit Fisher information):

$$v(\theta) = \frac{\left[\psi'(\theta)\right]^2}{\mathcal{I}_1(\theta)}.$$

4.8.8 Application of asymptotic normality

The practical implication of the theorem is that the repeated-sampling distribution of $\hat{\theta}_n$, in large samples, is approximately

$$\hat{\theta} \approx \mathrm{N}\left(\theta, \mathcal{I}(\theta)^{-1}\right).$$

In particular, we can calculate an *approximate standard error* for $\hat{\theta}$ by estimating the quantity $\mathcal{I}(\theta)$. We have two choices:

1. The obvious 'plug-in' estimator using the *expected* Fisher information

$$\mathrm{se}(\hat{\theta}_n) \approx 1\Big/\sqrt{\mathcal{I}(\hat{\theta}_n)}.$$

   This is not usually the best choice, however.

2. It is better (and generally more accurate) to use instead the *observed* Fisher information

$$\mathrm{se}(\hat{\theta}_n) \approx 1\Big/\sqrt{-l''(\hat{\theta}_n|X)},$$

   which is based directly on the curvature of the log-likelihood of $\hat{\theta}$.

Example 4.16. Suppose that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{Poi}(\lambda)$. Then

$$\begin{aligned} l(\lambda|X) &= \text{const.} - n\lambda + \sum_{i=1}^{n} X_i \log\lambda \\ l'(\lambda|X) &= -n + \sum_{i=1}^{n} X_i/\lambda \\ -l''(\lambda|X) &= \sum_{i=1}^{n} X_i/\lambda^2 \end{aligned}$$

Hence $l'(\lambda) = 0$ is solved at $\hat{\lambda}_n = \sum_{i=1}^{n} X_i/n =: \bar{X}_n$.

The large-sample variance of $\hat{\lambda}_n$ is

$$\mathcal{I}(\theta)^{-1} = \mathrm{E}\left[-l''(\lambda|X)\right]^{-1} = \lambda^2 \Big/ \mathrm{E}\left(\sum_{i=1}^{n} X_i\right) = \lambda^2/n\lambda = \lambda/n.$$

As a note, this variance is actually exact, since $\mathrm{Var}(\hat{\lambda}) = \mathrm{Var}(\bar{X}_n) = \mathrm{Var}(X_i)/n = \lambda/n$.

The estimated standard error for $\hat{\lambda}_n$ is

$$\mathrm{se}(\hat{\lambda}_n) \approx 1\Big/\sqrt{-l''(\hat{\lambda}_n|X)} = 1\Big/\sqrt{n\hat{\lambda}_n/\hat{\lambda}^2} = \sqrt{\hat{\lambda}_n/n}.$$

In this example, the plug-in estimator for $\mathcal{I}(\theta)$ happens to be the same as the observed information $-l''(\hat{\theta})$. Sometimes this happens, sometimes they are different.

4.9 Exercises

1. Let X be a sample of size 1 from a $\mathrm{N}(0, \sigma^2)$ population. Is $|X|$ a sufficient statistic for σ?

2. Let $X_1, \ldots, X_n$ be independent random variables with pdf

$$f_{X_i}(x|\theta) = \begin{cases} \exp(i\theta - x) & x \geq i\theta \\ 0 & \text{otherwise} \end{cases}$$

   for $i = 1, \ldots, n$. Prove that $T = \min(X_i/i)$ is a sufficient statistic for θ.

3. Let $X_1, \ldots, X_n$ be a random sample from a distribution whose pdf is $f(x|\theta) = (2\pi)^{-1/2} \exp(-(x-\theta)^2/2)$ for $x, \theta \in \mathbb{R}$. Find a minimal sufficient statistic for θ. *Hint: Use Theorem 10 in the lecture slides.*

4. Show that

   (a) the statistic $(\sum_i X_i, \sum_i X_i^2)$ is sufficient, but not minimal sufficient, in the $\mathrm{N}(\mu, \mu)$ family;

   (b) the statistic $\sum_i X_i^2$ is minimal sufficient in the $\mathrm{N}(\mu, \mu)$ family; and

   (c) the statistic $(\sum_i X_i, \sum_i X_i^2)$ is minimal sufficient in the $\mathrm{N}(\mu, \mu^2)$ family.

5. One observation is taken on a discrete random variable X with pmf $f(x|\theta)$ where $\theta \in \{1, 2, 3\}$. Find the MLE of θ.

x	$f(x\|\theta=1)$	$f(x\|\theta=2)$	$f(x\|\theta=3)$
0	1/3	1/4	0
1	1/3	1/4	0
2	0	1/4	1/4
3	1/6	1/4	1/2
4	1/6	0	1/4

6. Let $X_1, \ldots, X_n$ be iid with one of two pdfs. If $\theta = 0$, then

$$f(x|\theta) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

   while if $\theta = 1$,

$$f(x|\theta) = \begin{cases} 1/(2\sqrt{x}) & 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

   Find the MLE of θ.

7. Let $X_1, \ldots, X_n$ be iid with pmf

$$f(x|\theta) = \theta^x (1-\theta)^{1-x}$$

   for $x \in \{0, 1\}$ and $0 < \theta < 1/2$.

   (a) Find the MLE of θ, and the MOM estimator based on $\bar{X}$.

   (b) Find the mean squared error of each of the estimators.

   (c) Which estimator is preferred? Justify your choice.

8. If $X_1, \ldots, X_n$ are iid $\mathrm{N}(\mu, \sigma^2)$, show that the MLEs are

$$\hat{\mu} = \bar{X} \qquad \text{and} \qquad \hat{\sigma}^2 = \frac{n-1}{n} S^2,$$

   and obtain the bias, variance and mean squared error of each estimator.

9. Let $Y_1, \ldots, Y_n$ be a sample from a Poisson distribution with mean $\theta > 0$ unknown.

   (a) Let $Y = Y_1 + \cdots + Y_n$. Find the mean and variance of the distribution of Y. *Hint: Find out the mgf of Y.*

   (b) Obtain the MLE for θ and its standard error.

   (c) Suppose now that only the first m ($m < n$) observations of the sample are known explicitly, while for the other $n - m$ only their sum, Z say, is known. Determine the MLE of θ.

10. Let $X_1, \ldots, X_n$ be a sample from $\mathrm{Unif}(0, \theta)$ where $\theta > 0$ is an unknown parameter. Find the MLE $\hat{\theta}$ for θ. Derive the distribution for $\hat{\theta}$ and therefore show that $\hat{\theta}$ is a consistent estimator in the sense that $\hat{\theta} \xrightarrow{\mathrm{P}} \theta$ when $n \to \infty$. *Hint:* $\Pr(\max_i X_i \le y) = \prod_i \Pr(X_i \le y)$.

11. Let $X_1, \ldots, X_n$ be a random sample from a Bernoulli distribution, i.e. $\Pr(X_i = 1) = p = 1 - \Pr(X_i = 0)$ for $i = 1, \ldots, n$ where $p \in (0, 1)$ is unknown. Let $\theta = p^2$.

    (a) Find the Cramér-Rao lower bound for the variance of unbiased estimators for θ.

    (b) Find the MLE $\hat{\theta}$ for the parameter θ.

    (c) Show that $\mathrm{E}(\hat{\theta}) \neq \theta$.

12. Prove Lemma 28 in the lecture slides: Let $X \sim f(x|\theta)$, where $\theta \in \mathbb{R}$, and $S(\theta) = \frac{\partial}{\partial\theta} \log f(X|\theta)$. Prove that under certain regularity conditions (which you may assume to hold),

    (a) $\mathrm{E}[S(\theta)] = 0$

    (b) $\mathcal{I}(\theta) = \mathrm{Var}[S(\theta)]$

    (c) $\mathcal{I}(\theta) = -\mathrm{E}[S'(\theta)]$

13. Suppose that $X_1, \ldots, X_n \sim \mathrm{Poi}(\lambda)$. Let $\theta = e^{-\lambda} = \Pr(X_i = 0)$. Consider two estimators for θ:

    a.
    $$U(X) = \begin{cases} 1 & X_1 = 0 \\ 0 & \text{otherwise} \end{cases}$$

    b.
    $$\phi(X) = \left(1 - \frac{1}{n}\right)^{\sum_{i=1}^n X_i}$$

    Using the fact that $S = \sum_{i=1}^n X_i \sim \mathrm{Poi}(n\lambda)$, show that $\mathrm{Var}(\phi) < \mathrm{Var}(U)$.

14. (a) Show that
    $$\frac{1}{\sqrt{n}} l'(\theta|X) = \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n W_i \right)$$
    where $W = \frac{\partial f(X_i|\theta)/\partial\theta}{f(X_i|\theta)}$ has mean 0 and variance $\mathcal{I}_1(\theta)$. Here $\mathcal{I}_1(\theta)$ is the unit Fisher information, i.e. the Fisher information obtained from a single observation X_1. Now use the central limit theorem to establish the convergence to $\mathrm{N}(0, \mathcal{I}_1(\theta))$.

    (b) Show that
    $$-\frac{1}{n} l''(\theta|X) = \frac{1}{n} \sum_{i=1}^n W_i^2 - \frac{1}{n} \sum_{i=1}^n \frac{\partial^2 f(X_i|\theta)/\partial\theta^2}{f(X_i|\theta)}$$
    and that the mean of the first piece is $\mathcal{I}_1(\theta)$ and the mean of second piece is 0. Apply the WLLN.

15. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{Unif}(0, \theta)$. In the lectures, we found the MLE $\hat{\theta} = \max\{X_1, \ldots, X_n\}$ to be a biased estimator, with $\mathrm{Bias}(\hat{\theta}) = -\theta/(n+1) \neq 0$.

    (a) Find a **bias-corrected** version of $\hat{\theta}$ (call it $\hat{\phi}$), of the form $\hat{\phi} = c\hat{\theta}$ for a suitably chosen constant c.

    (b) What is the variance of the estimator $\hat{\phi}$?

    (c) In the lectures we found that
    $$\mathrm{MSE}(\hat{\theta}) = \theta^2 \left(\frac{n}{(n+1)^2(n+2)} + \frac{1}{(n+1)^2} \right).$$
    Find the mse of $\hat{\phi}$ and compare it with the mse of $\hat{\theta}$.

Hand-in questions

1. (a) If $X_1, \ldots, X_n$ is a random sample from the $\Gamma(\alpha, \beta)$ distribution, with α known, find the MLE of β. **[3 marks]**

   (b) Is the MLE for β unbiased? What is its variance? **[3 marks]**

2. Let $X_1, \ldots, X_n$ be a random sample from the pdf $f(x|\theta) = \theta x^{-2}$ for $0 < \theta \leq x < \infty$.

   (a) What is a sufficient statistic for θ? **[3 marks]**

   (b) Find the MLE of θ. **[4 marks]**

3. Let $X_1, \ldots, X_n$ be a random sample from $\mathrm{N}(\mu, \sigma^2)$. Find the Fisher information for μ and σ^2. *Hint: It may be easier to consider $\theta = \sigma^2$ in your calculations.* **[4 marks]**

Chapter 5

Hypothesis testing

In point estimation, the goal was to obtain the best estimate for an unknown parameter value θ. Another kind of inference activity that one might want to do is to test specific hypotheses about θ. For instance, we might want to know what the data says about θ taking on a specific value $\theta = \theta_0$. Traditionally, there are two viewpoints to hypothesis testing. One is the Fisherian view, and the other is the Neyman-Pearson method.

The Fisherian view of hypothesis testing is to construct a suitable test statistic and derive a probability value for the event that the observed test statistic falls in an extreme (or even more so) value based on the data, under the assumption that a particular null hypothesis is true. The idea here is that the so-called p-values gives an indication of the likelihood of the null hypothesis being plausible. Low p-values give support for the rejection of the null in favour of an alternative hypothesis.

In the Neyman-Pearson approach, a specific cut-off region must be specific beforehand, and the test statistic or equivalently the p-value is compared to this cut-off region. If it falls within a 'rejection region', then the null hypothesis is rejected. The p-values still retain their meaning, but the framework of the test is now reduced to a statistical decision (reject/not reject). The advantage of this framework is that it allows us to gauge the performances of the test (i.e. evaluating the false positives and false negatives).

The structure of this chapter is to discuss the Fisher and Neyman-Pearson approaches to hypothesis testing. In each approach, we will discuss the general method of finding a test and executing it accordingly. This usually involves deriving a suitable test statistic, finding out what its distribution is, and use that to calculate p-values.

We will find that the maximum likelihood method gives a nice optimality criterion for statistical testing. In addition, when the test statistic does not have a known distribution, we can make use of asymptotic distributions just like we did in Chapter 4.

Learning objectives

By the end of this chapter, you will be able to:

- Construct appropriate test statistics based on the problem at hand, and compute p-values accordingly.
- Use the likelihood ratio test approach to construct test statistics.
- Evaluate statistical tests based on size and power.
- Use asymptotic evaluations in the case where distributions are not easily derived.

Readings

- Casella and Berger (2002)
  - Chapter 8, sections 8.1, 8.2 (8.2.1 only), and 8.3 (8.3.1, 8.3.2 and 8.3.4 only).

  - Chapter 10, section 10.3.
- Wasserman (2004)
  - All of Chapter 10
- Topics not covered here: Bayesian tests, union-intersection and intersection-union tests, score test

5.1 Introduction

The task: to assess what the data say about the plausibility of a specific hypothesis about θ, e.g. a simple hypothesis of the form

$$H_0 : \theta = \theta_0$$

where θ_0 is a specified candidate value for θ, typically corresponding to an underlying subject-matter theory. Some examples:

- In tossing a coin, $\theta = 1/2$ means that the coin is 'fair'
- Is the true average height of males in Brunei truly $\mu = 1.65$?
- In linear regression, test the significance of the slope parameter $\beta_1 = 0$

A hypothesis under test is often called the *null hypothesis*, because it often relates to the absence (or nullity) of some conceivable **effect**. In the coin hypothesis example, $\theta = 1/2$ corresponds to absence of bias towards heads or tails.

The null hypothesis is often more complex than this, specifying a *set* of θ values, say $\theta \in \Theta_0$, rather than a single value. This is known as a composite hypothesis.

From Chapter 4, we already have a notion of *relative* plausibility for two candidate parameter values θ_1 and θ_2, namely the likelihood ratio

$$\frac{L(\theta_1|x)}{L(\theta_2|x)}.$$

Plainly, the use of the LR boils down to either "accepting" the θ_1 value, or rejecting it in favour of θ_2. For instance, if this ratio is found to be much larger than 1, then θ_1 is much more plausible than θ_2 on the basis of the data x.

We will see how likelihood ratios are the key to an *optimal* assessment of the plausibility of a hypothesis.

5.1.1 A general paradigm

A general paradigm:

- Identify, somehow, a *test statistic* $W(X)$, which is such that larger values of W represent stronger evidence against H_0;
- Measure the *strength* of the evidence against H_0 in any realised value $W(x)$ by calculating the p-value (see next slide).

If the p-value is very small, then evidence as strong as $W(x)$ (or stronger) is found only rarely under H_0, and so $W(x)$ represents strong evidence against H_0.

5.1.2 p-values

Definition 5.1 (p-value)**.** Let $W(X)$ be a test statistic such that large values of W give evidence that H_1 is true. For each sample point x, define the p-value to be

$$p_\theta(x) = \sup_{\theta \in \Theta_0} \Pr_\theta \left(W(X) \geq W(x) \right).$$

In statistical hypothesis testing, the p-value (or probability value) is the probability of obtaining test results at least as extreme as the results actually observed during the test, assuming that the null hypothesis is correct. Some general remarks:

- The p-value is a statistic.
- The p-value is indeed a **p**robability value which lies between 0 and 1.
- The p-value reports the result of a test on a more continuous scale, rather than just the dichotomous decision "Reject/Do not reject H_0".

Example 5.1. Let $X_1, \ldots, X_{20} \in \{T, H\}$ be the outcomes of an experiment of tossing a coin 20 times, i.e.

$$\Pr(X = H) = \pi = 1 - \Pr(X = T), \quad \pi \in (0, 1).$$

Let $W = [X_1 = H] + \cdots + [X_{20} = H]$. Then $W \sim \text{Bin}(20, \pi)$, and an estimate of π is $\hat{\pi} = \bar{X}$. We would like to assess whether or not the hypothesis that "the coin is fair" is true. That is,

$$H_0 : \ \pi = 0.5 \qquad \text{v.s.} \qquad H_1 : \ \pi \neq 0.5$$

Let W be the test statistic, and suppose we observe $W = 17$. Intuitively, large values of W indicate evidence against the coin being fair, and would favour the `Heads' outcome more than`Tails'.

Under the assumption $H_0 : \pi = 0.5$ is true, then

$$p(X) = \Pr_{\pi=0.5}(W \geq 17) = \sum_{w=17}^{20} \frac{20!}{w!(20-w)!} 0.5^w (1 - 0.5)^{20-w} = 0.0013$$

This is the (one-sided) p-value favouring the 'Heads' outcome.

On the other hand, small values of W *also* indicate evidence against the coin being fair; evidence in favour of a 'Tails' outcome being more likely than a 'Heads'. Let Y be the number of Tails observed, so $Y = 20 - W$, which has a $\text{Bin}(20, 1 - \pi)$ distribution.

The p-value for the observed $Y = 20 - 17 = 3$ observation would be

$$p(X) = \Pr_{\pi=0.5}(Y \leq 3) = \sum_{y=0}^{3} \frac{20!}{y!(20-y)!} 0.5^y (1 - 0.5)^{20-y} = 0.0013$$

This is the (one-sided) p-value favouring the 'Tails' outcome, which is the same as above due to symmetry. Combining the two p-values together gives the two-sided p-value, $p(X) = 0.0026$. This gives a measure of how unlikely $H_0 : \pi = 0.5$ holds given the observed 17 out of 20 'Heads' outcome.

As a remark, the answer cannot possibly be resulted from the estimator $\hat{\pi}$, for

- if $\hat{\pi} = 0.9$, then H_0 is unlikely to be true.
- if $\hat{\pi} = 0.45$, then H_0 is may be true (but also may be untrue).
- if $\hat{\pi} = 0.7$, then what?

Furthermore, $\hat{\pi} = \bar{X}$ is a random variable, so will vary from sample to sample!

5.1.3 Accept H_0?

It is not possible to "prove" a negative. When the p-value is large, it means that there is a lack of evidence to prove something exists–it does not prove something does not exist!

Not reject $\neq$ Accept

A statistical test is incapable to accept a hypothesis. A large p-value is indeed indicative of the null hypothesis being likely, but the philosophically correct attitude would be to conclude that **there is insufficient evidence to reject the null** (as opposed to accepting the null).

With this in mind, note that for the most part we will be viewing the statistical testing problem as a problem in which one of two actions is going to be taken: the assertion of H_0 or H_1.

At the end of the day, we can never know for certain what the truth is; we can only act on probability and likelihood based on the observed data.

5.1.4 Uniformity of p-values

Here's an interesting fact:

Theorem 5.1 (Uniformity of p-values)**.** *If θ_0 is a point null hypothesis for the parameter of continuous X, then a correctly calculated p-value $p_W(X)$ based on any test statistic W, is such that*

$$p_w(X) \sim \text{Unif}(0,1)$$

in repeated sampling under H_0.

This result is useful especially for *checking the validity* of a complicated p-value calculation:

1. Simulate (on a computer) several new data sets from the null distribution.
2. For each simulated data set, apply the p-value calculation and save the result.
3. Assess the collection of resulting p-values–do they seem to be uniformly distributed?

Proof. This is a consequence of the *probability integral transform*: Suppose that a continuous r.v. T has cdf $F_T(t), \forall t$. Then the r.v. $Y = F_T(T) \sim \text{Unif}(0,1)$ because:

$$F_Y(y) = \Pr(Y \leq y) = \Pr(F_T(T) \leq y) = \Pr\left(T \leq F_T^{-1}(y)\right) = F_T\left(F_T^{-1}(y)\right) = y,$$

which is the cdf of a Unif$(0,1)$ distribution.

For any data x,

$$p_W(x) = \Pr_{\theta_0}\left(W(X) \geq W(x)\right) = 1 - F(W(x)),$$

where F is the cdf (under H_0) of $W(X)$. Hence, $p_W(x) = 1 - Y$ where $Y \sim \text{Unif}(0,1)$ by the probability integral transform. But clearly if $Y \sim \text{Unif}(0,1)$, then so is $1 - Y$. □

Probability Integral Transform

5.2 Likelihood ratio test

The likelihood ratio test (LRT) is a general approach to finding a test statistic.

Definition 5.2 (Likelihood ratio test)**.** For a model with parameter space Θ, the likelihood ratio test statistic for testing a specified null hypothesis

$$H_0 : \theta \in \Theta_0$$

where $\Theta_0 \subset \Theta$, is

$$W_{LR}(X) = \frac{\sup_{\theta \in \Theta} L(\theta|X)}{\sup_{\theta \in \Theta_0} L(\theta|X)}.$$

The statistic $W_{LR}(X)$ measures the *implausibility* of the most plausibile θ value in Θ_0, relative to the most plausible value in the whole of Θ. Thus, **larger values** of $W_{LR}(X)$ represent **stronger evidence** against H_0, i.e. large values ⇒ reject H_0.

Note that

$$\hat{\theta} = \sup_{\theta \in \Theta} L(\theta|X)$$

is the (unconstrainted) ML estimator for θ. Further, define

$$\tilde{\theta} = \sup_{\theta \in \Theta_0} L(\theta|X)$$

as the constrained ML estimator under H_0. Then the LRT statistic can be written

$$W_{LR}(X) = \frac{f(\hat{\theta}|X)}{f(\tilde{\theta}|X)},$$

where $X = (X_1, \ldots, X_n)^\top \sim f(x|\theta)$.

Remark: It is easy to see that $W_{LR}(X) \geq 1$.

5.2.1 Log likelihood ratio test statistic

As with the likelihood, it is often more convenient to consider the logarithm of the likelihood ratio test statistic:

$$\begin{aligned}\log W_{LR}(X) = \log \frac{L(\hat{\theta}|X)}{L(\tilde{\theta}|X)} &= l(\hat{\theta}|X) - l(\tilde{\theta}|X) \\ &= \log f(\hat{\theta}|X) - \log f(\tilde{\theta}|X)\end{aligned}$$

The sampling distribution is of interest, but usually unknown, except in a few special cases. Two strategies:

- Identify a different statistic with an "easy'' distribution in the (log) LR statistic, which is an increasing function of the actual (log) LR statistic, and use this to instead.
- Use asymptotic results to find an approximate distribution. We'll cover this in later sections.

5.2.2 Example: Normal with known variance

Example 5.2. Suppose that n patients use a new drug for hypertention, and we wish to assess the drug's effectiveness. Measurements of blood pressure are taken before and after treatment, resulting in the measured different X_i for patient i.

Let's assume that

- The BP measurements are all iid: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$ with known variance.
- The effect of the drug is the same improvement μ for all patients.

We wish to test the null hypothesis $H_0 : \mu = 0$.

The log-likelihood is

$$l(\mu|X) = \text{const.} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$$

and so, recalling that $\hat{\mu} = \bar{X}$, the log of the LR statistic is

$$\begin{aligned}\log W_{LR} &= l(\hat{\mu}|X) - l(\tilde{\mu}|X) \\ &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{n}X_i^2 \\ &= \frac{n\bar{X}^2}{2\sigma^2}.\end{aligned}$$

Now notice that this statistic is an increasing function of $|\bar{X}|$.

We use the distribution of the sample mean statistic, $\bar{X} \sim \mathrm{N}(\mu, \sigma^2/n)$. So for a given data vector $X = x$, the p-value is

$$\begin{aligned}\Pr_{\mu=0}\left(|\bar{X}| \geq |\bar{x}|\right) &= 2\Pr_{\mu=0}\left(\bar{X} \geq |\bar{x}|\right) \\ &= 2\Pr\left(\frac{\bar{X}-0}{\sigma/\sqrt{n}} \geq \frac{|\bar{x}|-0}{\sigma/\sqrt{n}}\right) \\ &= 2(1-\Phi(\sqrt{n}|\bar{x}|/\sigma))\end{aligned}$$

Let's put in some numbers:

- $n = 10$ patients
- $\sigma = 4.3$ mmHg
- $\bar{x} = -12.8$ mmHg

–an apparent reduction in average blood pressure after treatment. Now compute the p-value:

$$p(\bar{x}) = 2\left(1-\Phi\left(\frac{\sqrt{n}|\bar{x}|}{\sigma}\right)\right) = 2\left(1-\Phi\left(\frac{\sqrt{10}\times 12.8}{4.3}\right)\right) \approx 10^{-11}$$

A very small value indeed, indicating very strong evidence against the null hypothesis (i.e. clear evidence the drug has an effect).

However, note the assumptions above. Are they realistic?

5.2.3 Example: Normal with unknown variance (t-test)

Example 5.3. Suppose that $X = (X_1, \ldots, X_n)^\top$ is a random sample from $\mathrm{N}(\mu, \sigma^2)$. We are interested in testing hypotheses

$$H_0 : \ \mu = \mu_0 \quad \text{v.s.} \quad H_1 : \ \mu \neq \mu_0,$$

where μ_0 is given, and σ^2 is unknown and is a nuisance parameter. Recall the log-likelihood function as being

$$l(\mu, \sigma^2|X) = \text{const.} - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2,$$

and maximising this without restriction yields

$$\hat{\mu} = \bar{X} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

On the other hand, under H_0, μ is fixed at μ_0, while the constrained MLE for σ^2 is

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_0)^2.$$

The LR statistic (after simplification) is then

$$W_{LR} = \frac{L(\hat{\mu}, \hat{\sigma}^2)}{L(\mu_0, \tilde{\sigma}^2)} = \left(\frac{\tilde{\sigma}^2}{\hat{\sigma}^2}\right)^{n/2}.$$

Since $\tilde{\sigma}^2 = \hat{\sigma}^2 + (\bar{X}-\mu_0)^2$, it holds that $\tilde{\sigma}^2/\hat{\sigma}^2 = 1 + T^2/(n-1)$, where

$$T = \frac{\bar{X}-\mu_0}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\Big/n}} = \frac{\bar{X}-\mu_0}{S/\sqrt{n}} \sim t_{n-1}.$$

We thus see that W_{LR} is an increasing function of $|T|$, and hence the p-value in this case is obtained from a table of the t_{n-1} distribution rather than the standard normal.

This, the so-called t-test, is probably the most commonly used of all procedures in statistical practice! Now you know how it is derived...

For the t-test, under H_0, $X_i \overset{\text{iid}}{\sim} \text{N}(\mu_0, \sigma^2)$. So we simulate a data set $\{X_1, \ldots, X_n\}$ using these parameters: $n = 15, \sigma = 2, \mu_0 = 2$.

```
X <- rnorm(n = 15, mean = 2, sd = 2)
round(X, 3)
```

```
##  [1]  4.448  2.720  2.802  2.221  0.888  5.574  2.996 -1.933  3.403  1.054
## [11] -0.136  1.564 -0.052  0.542  0.750
```

The p-value for the t-test is $\Pr\left(|Y| > |\sqrt{n}(\bar{x} - \mu_0)/s|\right)$, where $Y \sim t_{n-1}$ (the two-tail probability of "extreme events"). For instance,

```
test.stat.obs <- abs(sqrt(15) * (mean(X) - 2) / sd(X))
pval <- 2 * pt(test.stat.obs, df = 15 - 1, lower.tail = FALSE)
pval
```

```
## [1] 0.6800548
```

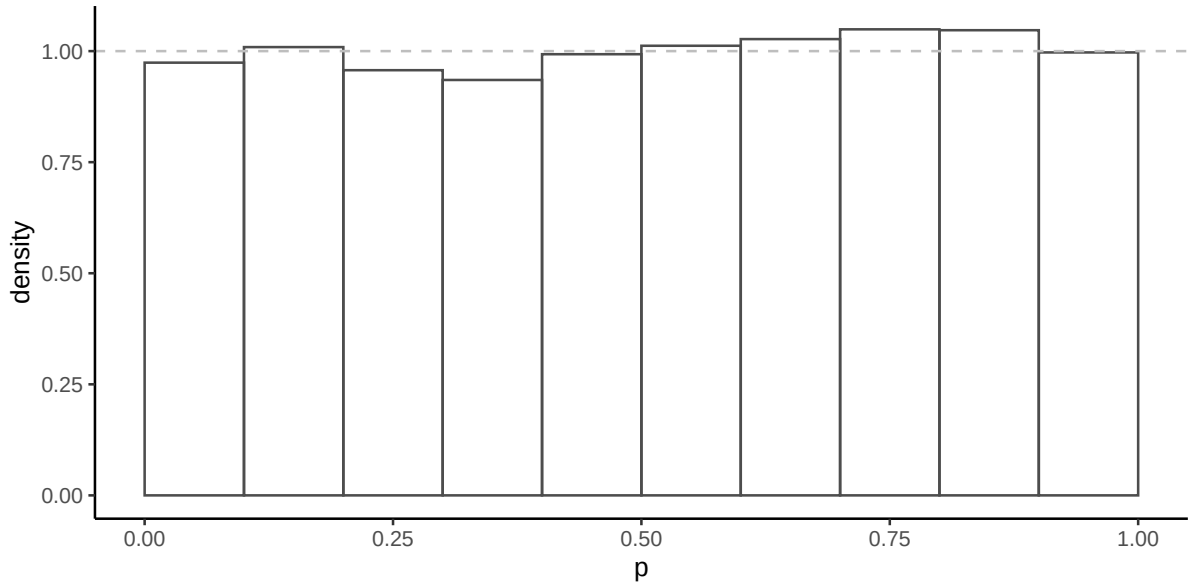
Simulate this `B=10000` times in a `for` loop:

```
B <- 10000
res <- rep(NA, B)  # create a vector to collect the p-values
for (i in 1:B) {
  X <- rnorm(n = 15, mean = 2, sd = 2)
  test.stat.obs <- abs(sqrt(15) * (mean(X) - 2) / sd(X))
  pval <- 2 * pt(test.stat.obs, df = 15 - 1, lower.tail = FALSE)
  res[i] <- pval
}

head(res)
```

```
## [1] 0.42203902 0.73961301 0.57621365 0.35418373 0.06711971 0.10304763
```

Plot a histogram of the simulated p-values. We should observe uniformity:



5.3 The Neyman-Pearson approach

The 'Neyman-Pearson' approach to testing hypotheses is to reject H_0 if $W(X) \in R$, where R is a suitably defined *critical region*. If W is designed to measure the evidence against H_0, then most often R takes the form

$$R = \{x \mid W(x) \geq c\}$$

for some constant c.

Example 5.4. From Example 5.2, we saw that $W = \exp(n\bar{X}^2/2\sigma^2)$ for testing $H_0 : \mu = 0$ from a normal sample with known variance. The rejection region is therefore

$$\begin{aligned} R &= \{x \mid \exp(n\bar{X}^2/2\sigma^2) \geq c\} \\ &= \left\{x \,\middle|\, |\bar{X}| \geq \sqrt{2\sigma^2 \log c/n}\right\} \end{aligned}$$

So the LR test rejects $H_0 : \mu = 0$ if the sample mean exceeds a specified amount.

5.3.1 Performance of a test

In deciding to "accept" or reject the null hypothesis H_0, an experimenter might be making a mistake. The performance of a test is measured by two criteria: the size and power of a test.

Definition 5.3 (Size of a test)**.** For $0 \leq \alpha \leq 1$, the size α of a test is defined as

$$\alpha := \sup_{\theta \in \Theta_0} \Pr_\theta \left(W(X) \in R\right)$$

The size of a test measures the probability of rejecting the null hypothesis under the assumption that the null hypothesis is true.

Definition 5.4 (Power of a test)**.** For $0 \leq B(\theta) \leq 1$, the power $B(\theta)$ of a test is defined as

$$B(\theta) := \Pr_\theta \left(W(X) \in R\right), \qquad \theta \notin \Theta_0$$

The power function of a test is defined as the probability of rejecting the null hypothesis *correctly* (i.e. $\theta \notin \Theta_0$) in favour of the alternative.

A *good* test (W, R) has small size α and large power $B(\theta)$ at all values of θ outside of the null hypothesis.

Example 5.5. Continuation of normal example with known variance: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$ with σ^2 known and null hypothesis $H_0 : \mu = 0$.

The rejection region from Example 5.4 is alternatively written as

$$R = \{x \mid \exp(n\bar{X}^2/2\sigma^2) \geq c\} = \left\{x \,\middle|\, \left|\frac{\bar{X}}{\sigma/\sqrt{n}}\right| \geq \sqrt{2\log c}\right\}.$$

So for instance, $R = \{x \mid |\sqrt{n}\bar{X}/\sigma| \geq 1.96\}$ is a critical region of size 0.05.

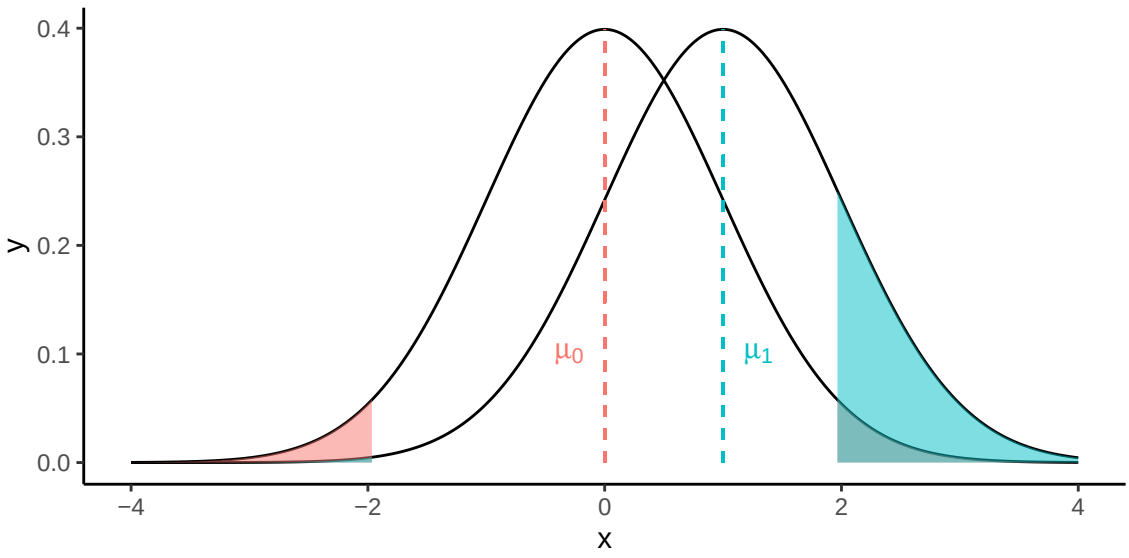
For our illustrative data, with $\sigma = 4.3$ and $\bar{x} = -12.8$,

$$\left|\frac{\bar{X}}{\sigma/\sqrt{n}}\right| = \left|\frac{-12.8}{4.3/\sqrt{10}}\right| = 9.413 > 1.96.$$

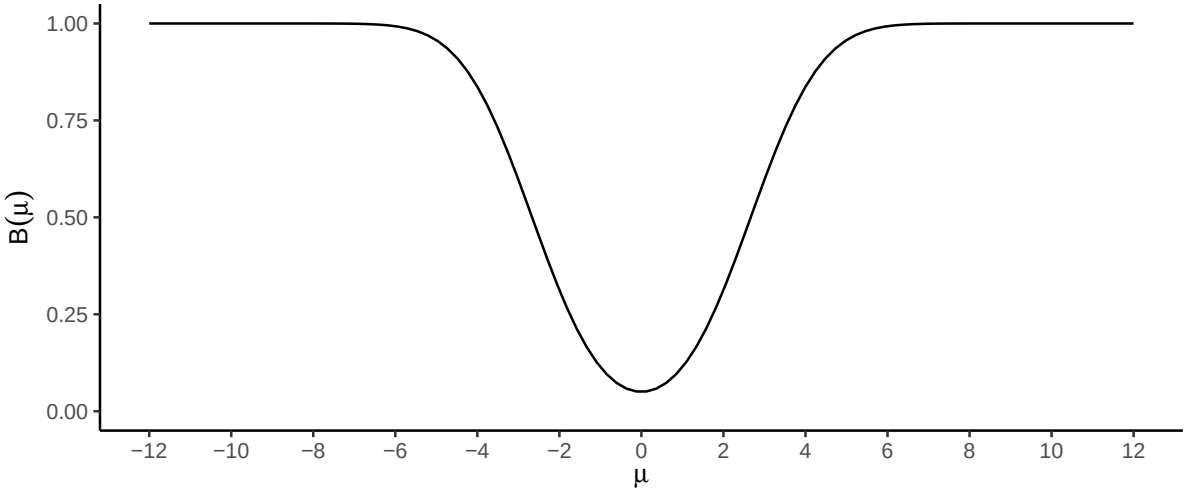
For a test of size $\alpha = 0.05$, the power of the test is

$$\begin{aligned} B(\mu) &= \Pr\left\{\left|\frac{\bar{X} - \mu + \mu}{\sigma/\sqrt{n}}\right| \geq 1.96\right\} \\ &= \Pr\left\{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq -1.96 - \frac{\mu}{\sigma/\sqrt{n}}\right\} + \Pr\left\{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \geq 1.96 - \frac{\mu}{\sigma/\sqrt{n}}\right\} \\ &= \Phi\left(-1.96 - \sqrt{n}\mu/\sigma\right) + \left[1 - \Phi\left(1.96 - \sqrt{n}\mu/\sigma\right)\right] \end{aligned}$$

This represents the two tail probabilities based on the rejection region.



For our illustrated example ($\sigma = 4.3, n = 10$), the power function is plotted below. This plots the power of the test assuming some value of μ is true. If $\mu = -12.8$ (as observed in the data) then the power is almost 1!



5.3.2 Relation to p-values

The conclusion of the test (with the illustrated data) is that "H_0 is rejected at the 5% level (of significance)". This is interpeted to mean

> If H_0 were true, we would reject H_0 using this test only 5% of the time in repeated sampling. So this is fairly strong evidence against H_0.

But that is not a very informative summary of the evidence! In fact, with these data, we would *also* reject H_0 at the 1% level, and at the 0.1% level, etc.

It would be much more informative to ask: "What is the *smallest* size of test based on W that would reject H_0 based on the data x?" The answer is precisely the p-value, $p_W(x)$.

So the two approaches are closely linked, with the p-value giving the most informative assessment of the strength of evidence against H_0.

5.4 Type I and II errors

The quantities α and $\beta(\theta) := 1 - B(\theta)$ are called the probability of a 'Type I error' and a 'Type II error' respectively.

Definition 5.5 (Type I and II error)**.** The Type I error (false positive) is defined to be

$$\alpha = \Pr(\text{Reject } H_0 \mid H_0 \text{ is true}).$$

The Type II error (false negative) is defined to be

$$\beta(\theta) = \Pr(\text{Fail to reject } H_0 \mid H_0 \text{ is false}) = 1 - B(\theta).$$

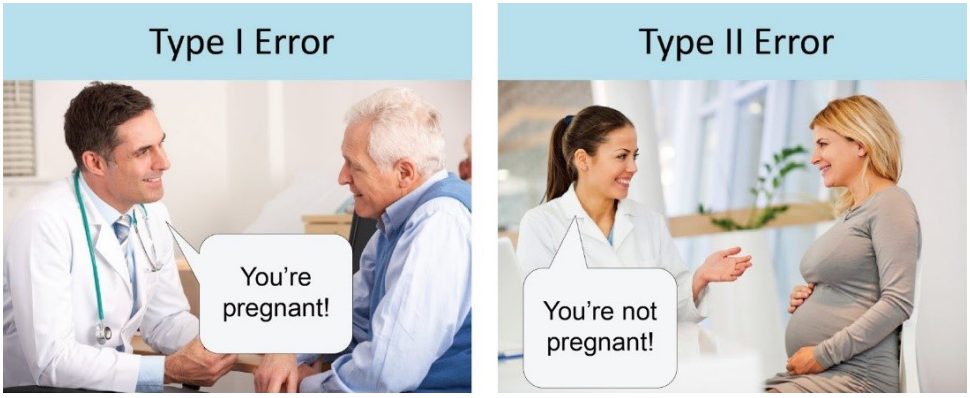
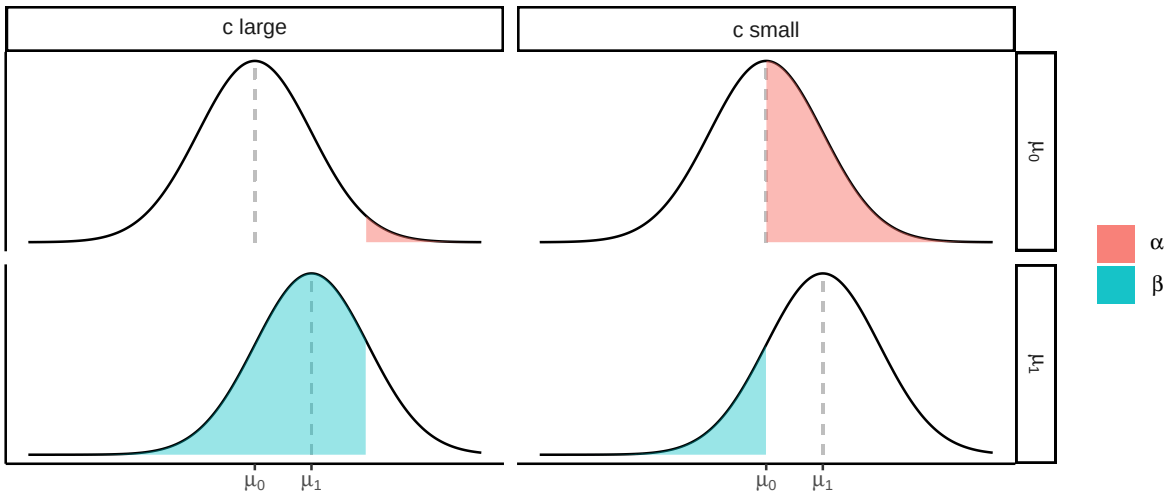


Figure 5.1: Summary of Type I and II errors.

	H_0 is true	H_1 is false
Do not reject H_0	Correct inference (true negative) prob. $= 1 - \alpha$	Type II error (false positive) prob. $= \alpha$
Reject H_0	Type I error (false positive) prob. $= \alpha$	Correct inference (true negative) prob. $= 1 - \beta(\theta)$

5.4.1 Minimising errors

The aim is to make both Type I and II errors as small as possible, simultaneously. However, for a large value of c in the rejection region will give small α and large β, and vice versa for a small value of c.



This conflict is usually resolved by *fixing* α, say at 0.05 or 0.01, and then using a test (W, R) that makes $\beta(\theta)$ as small as possible for all $\theta \notin \Theta_0$. Some remarks:

1. Suppose that H_0 is true, rejection of the null hypothesis occurs if p-value is small. But the probability of this error (Type I) is not greater than the size of the test α. Hence, it is under control.
2. Unfortunately, we do not have explicit control on the probability β of making a Type II error. But we can certainly gauge the conditions resulting in large β and try to avoid them.
3. It is more conclusive to end a test with H_0 rejected, as the decision "Not reject" does not imply that H_0 is accepted.

5.4.2 Optimality of the LR test

If we can't control the Type II error of a test, are we out of luck? The Neyman-Pearson approach provides some neat theory!

Lemma 5.1 (Neyman-Pearson)**.** *Consider testing the simple hypothesis $H_0 : \theta = \theta_0$, suppose that*

- *θ_1 is any other candidate value of θ;*
- *$W_{LR}(X) = \frac{L(\theta_1|X)}{L(\theta_0|X)}$;*
- *$R_{LR} = \{x \mid W_{LR}(X) \geq c\}$ s.t. $\Pr_{\theta_0}(W_{LR} \in R_{LR}) = \alpha$.*

*Then **no** other size α test pair (W, R) has $\Pr_{\theta_1}(W \in R)$ <u>greater than</u> $\Pr_{\theta_1}(W_{LR} \in R_{LR})$.*

The proof is omitted (see for e.g. C&B Thm 8.3.12 or on Wikipedia). The implication is that since this result applies for every possible value of θ_1, the LR test (W_{LR}, C_{LR}) is said to be the *uniformly most powerful* (UMP) test of size α. This makes the use of W_{LR} very compelling for hypothesis testing, whether via the p-value approach or the critical-region approach.

5.5 One-sided tests

Sometimes we wish to measure the evidence (against H_0) in one direction only.

Example 5.6. Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$ with σ^2 known. Consider testing

$$H_0 : \mu \leq 0 \quad \text{v.s.} \quad H_1 : \mu > 0$$

The unrestricted MLE is $\hat{\mu} = \bar{X}$, while the restricted MLE is $\tilde{\mu} = 0$ if $\bar{X} > 0$. So for $\bar{X} > 0$, we have (as before)

$$W_{LR}(X) = \frac{L(\hat{\mu}|X)}{L(0|X)} = \exp\left(n\bar{X}^2/2\sigma^2\right).$$

But if $\bar{X} \leq 0$, $W_{LR}(X) = 1$, because $\hat{\mu} = 0$ in such a case.

The p-value from data x is (using the monotonicity of $\bar{X}$ in the LRT statistic)

$$p(x) = \begin{cases} \Pr(\bar{X} > \bar{x}) = 1 - \Phi(\sqrt{n}\bar{x}/\sigma) & \bar{x} > 0 \\ 1 & \bar{x} \leq 0 \end{cases}$$

Hence, relative to the 'two-sided' test that we saw previously, the p-value is *halved* if $\bar{x} > 0$, and ignores the precise value of $\bar{x}$ if $\bar{x} \leq 0$.

It's a good idea to sketch the likelihood function above.

Further remarks:

1. Performing a one-sided test instead of a two-sided test thus makes any apparent evidence against H_0 seem stronger (since the p-value is halved).

2. In practice there are rather few situations where performing a one-sided test, which assumes that we know in advance that departures from H_0 are in one direction only, can be justified. When assessing the effect of a new drug, for example, the convention is to assess evidence for an effect in either direction, positive or negative.

3. The two-sided test is said to be more *conservative* than the one-sided test: The one-sided test risks over-stating the strength of evidence against H_0 if the underlying assumption–that evidence against H_0 counts in one direction only–is actually false.

5.6 Approximate tests

5.6.1 Asymptotic distribution of LRTs

We cannot always derive easily the distribution of W_{LR} under H_0. But a general *large-sample approximation* to the null distribution of W_{LR} comes from the following result

Theorem 5.2. *For testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, suppose $X_1, \ldots, X_n \overset{iid}{\sim} f(x|\theta)$, and $f(x|\theta)$ satisfies the usual regularity conditions. Let $\hat{\theta}_n$ be the MLE for θ. Then under H_0, as $n \to \infty$,*

$$-2\log\left[\frac{L(\theta_0|X)}{L(\hat{\theta}_n|X)}\right] = 2\log W_{LR}(X) \xrightarrow{D} \chi^2_1.$$

- For the two-sided testing situation, we can always get an approximate p-value for the observed data as $p(x) = \Pr\left(Y \geq 2\log W_{LR}(x)\right)$, where $Y \sim \chi^2_1$.
- Remarkably, this result applies *whatever* the distribution of the X_is are. It is partly a result of the asymptotic normality of $\hat{\theta}$ (see proof).

Proof. Taylor expanding $l(\theta|bX)$ around $\hat{\theta}$ gives

$$l(\theta|X) = l(\hat{\theta}|X) + \cancel{(\theta - \hat{\theta})l'(\hat{\theta}|X)} + \frac{(\theta-\hat{\theta})^2}{2!}l''(\hat{\theta}|X) + \cdots$$

Consider then quantity $2\log W_{LR}$ under the assumption that $H_0 : \theta = \theta_0$ is true:

$$\begin{aligned} 2\log W_{LR} &= 2l(\hat{\theta}|X) - 2l(\theta_0|X) \\ &\approx \cancel{2l(\hat{\theta}|X)} - \cancel{2l(\hat{\theta}|X)} - (\theta_0 - \hat{\theta})^2 l''(\hat{\theta}|X) \end{aligned}$$

Recall that $-l''(\hat{\theta}|X)$ is the so-called *observed Fisher information* (Part 4 slides, p.71), and that $-\frac{1}{n}l''(\hat{\theta}|X) \xrightarrow{\mathrm{P}} \mathcal{I}_1(\theta_0)$ (Ex. sheet 4, Q14b).

Since MLEs are, under certain regularity conditions, asymptotically efficient, we have that as $n \to \infty$ under H_0,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathrm{D}} \mathrm{N}\left(0, \mathcal{I}_1(\theta_0)^{-1}\right).$$

It follows that $\sqrt{\mathcal{I}_1(\theta_0)} \cdot \sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathrm{D}} \mathrm{N}(0,1)$ and that

$$\mathcal{I}_1(\theta_0) \cdot n(\hat{\theta} - \theta_0)^2 \xrightarrow{\mathrm{D}} \chi^2_1,$$

and hence

$$2\log W_{LR} = -\frac{l''(\hat{\theta}|X)}{n} \cdot n(\hat{\theta} - \theta_0)^2 \xrightarrow{\mathrm{D}} \chi^2_1$$

by application of Slutzky's theorem. □

5.6.2 Wilk's theorem

The above theorem can be extended to cases where the null hypothesis concerns vectors of parameters, i.e. $\Theta \subseteq \mathbb{R}^p$. We state it here without proof.

Theorem 5.3 (Wilk's theorem)**.** *Let $X_1, \dots, X_n \overset{iid}{\sim} f(x|\theta)$ with $f(x|\theta)$ satisfying the usual regularity conditions. Consider testing the composite hypothesis for $\theta \in \mathbb{R}^p$*

$$H_0 : \theta \in \Theta_0 \quad v.s. \quad H_1 : \theta \in \Theta \ \ \Theta_0.$$

Then

$$-2 \log \left[\frac{\sup_{\theta \in \Theta_0} L(\theta|X)}{\sup_{\theta \in \Theta} L(\theta|X)} \right] = 2 \log W_{LR}(X) \xrightarrow{D} \chi^2_k,$$

as $n \to \infty$, where $k = \dim(\Theta) - \dim(\Theta_0)$. The degrees of freedom k of this limiting distribution is the difference between the number of free parameters specified by $\theta \in \Theta_0$ and the number of free parameters specified by $\theta \in \Theta$.

Example 5.7. Let $X_1, \dots, X_n$ be independent, and $X_i \sim \mathrm{N}(\mu_i, 1)$. Consider the null hypothesis

$$H_0 : \mu_1 = \cdots = \mu_n.$$

The likelihood function (up to a constant of proportionality) is

$$L(\mu_1, \dots, \mu_n) \propto \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} (X_i - \mu_i)^2 \right\},$$

Then, the unconstrained MLE are $\hat{\mu}_i = X_i$, while the constrained MLE is $\tilde{\mu} = \bar{X}$. Hence,

$$W_{LR} = \frac{L(\hat{\mu}_1, \dots, \hat{\mu}_n)}{L(\tilde{\mu}, \dots, \tilde{\mu})} = \exp\left\{ \frac{1}{2} \sum_{i=1}^{n} (X_i - \bar{X})^2 \right\}.$$

The asymptotic distribution of $2 \log W_{LR}$ is

$$2 \log W_{LR} = \sum_{i=1}^{n} (X_i - \bar{X})^2 \xrightarrow{\mathrm{D}} \chi^2_{n-1}$$

as $n \to \infty$ by Wilk's theorem. Thus, the null hypothesis is rejected for large values of $2 \log W_{LR}$ as compared to the χ^2_{n-1} distribution. The (approximate) p-value is

$$p(x) = \Pr\left(Y > \sum_{i=1}^{n} (x_i - \bar{x})^2 \right), \qquad Y \sim \chi^2_{n-1}$$

It turns out that $2 \log W_{LR}$ has an **exact** χ^2_{n-1} distribution since $(n-1)^{-1} 2 \log W_{LR} = S^2$ (the unbiased sample variance), and we saw previously that $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.

5.6.3 The Wald test

Another common method of constructing a large-sample test statistic is based on an estimator that has an asymptotic normal distribution (e.g. the MLE).

Definition 5.6 (Wald test)**.** Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} f(x|\theta)$ and suppose we would like to test $H_0 : \theta = \theta_0$. Let $\hat{\theta}_n$ be an estimator for θ which is asymptotically normal, i.e. as $n \to \infty$,

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathrm{D}} \mathrm{N}\left(0, \mathcal{I}_1(\theta)^{-1}\right),$$

where $\mathcal{I}_1(\theta)$ is the (unit) Fisher information about θ. Write $\mathrm{se}(\hat{\theta}_n)$ as the estimate of the s.d. of $\hat{\theta}_n$, $n/\sqrt{\mathcal{I}_1(\theta)^{-1}}$. A Wald test is a test based on a statistic of the form

$$Z_n := \frac{\hat{\theta}_n - \theta_0}{\mathrm{se}(\hat{\theta}_n)} \approx \mathrm{N}(0, 1)$$

where θ_0 is the hypothesised value of θ (under H_0).

Some remarks regarding the Wald test:

- The asymptotic efficiency property actually affords us

$$Z_n := \frac{\hat{\theta}_n - \theta_0}{\mathrm{sd}(\hat{\theta}_n)} \approx \mathrm{N}(0,1)$$

  but $\mathrm{sd}(\hat{\theta}_n)$ may depend on some unknown parameters. If $\mathrm{sd}(\hat{\theta}_n)/\mathrm{se}(\hat{\theta}_n) \xrightarrow{\mathrm{P}} 1$ then we may use the standard error instead.

- As discussed in Chapter 4, there are two versions of obtaining the standard error:
  - Using the plug-in estimator: $\mathrm{se}(\hat{\theta}_n) = 1\Big/\sqrt{\mathcal{I}(\hat{\theta}_n)}$
  - Using the observed Fisher information: $\mathrm{se}(\hat{\theta}_n) = 1\Big/\sqrt{-l''(\hat{\theta}_n|X)}$

- The Wald test is very practical since there are no assumptions made on the distribution of the data X_i. Of course, it is an approximate test and the "reliability" of the test depends on the sample size. In fact, the Wald test can be shown to have an asymptotic size α and power 1 [1].

There are some disadvantages to the Wald test:

- The Wald test is **not** invariant to a non-linear transformation/reparameterisation of the hypothesis. One might get different answers to the test of $H_0 : \theta = 1$ and $H_0 : \log\theta = 0$ (although they ask the same thing). The reason for this is there is no relationship (in general) between the two standard errors (e.g. $\mathrm{se}(\hat{\theta}_n)$ and $\mathrm{se}(\log\hat{\theta}_n)$) so they need to be approximated somewhat independently.
- The Wald test actually uses two approximations: 1) the normality from the asymptotic efficiency property; and 2) the use of (approximate) standard errors. In contrast, the LRT only uses "one" approximation, and that is the large-sample χ^2 distribution of $2\log W_{LR}$.

Example 5.8. To deal with a coffee shop's customer complaint that the amount of chilled coffee in their bottled drinks is less than the advertised 300ml, 20 bottles were decanted and the coffee measured, yielding data X_i as follows:

282	301	311	271	293	268	302	301	293	256
278	301	309	294	282	281	305	301	285	279

The sample mean and the (unbiased) sample standard deviation are

$$\bar{x} = 289.7 \text{ ml} \qquad s = 14.8.$$

which are taken as estimates of the population mean and standard deviation μ and σ respectively.

By the CLT, the sample mean estimator is asymptotically efficient: $\sqrt{n}\,(\bar{X} - \mu) \xrightarrow{\mathrm{D}} \mathrm{N}\,(0, \sigma^2)$. From this, an (approximate) standard error of the estimator $\bar{X}$ is $\mathrm{se}(\bar{X}) = s/\sqrt{n}$. To test

$$H_0 :\ \mu = 300 \quad \text{v.s.} \quad H_1 :\ \mu < 300,$$

we apply the Wald test with an observed test statistic value of

$$z = \frac{\bar{X} - 300}{14.8/\sqrt{20}} = -3.121.$$

The critical region for a test of size $\alpha = 0.01$ is $\{x \mid Z \leq -2.326\}$. Thus the test rejects $H_0 : \mu = 300$ at the 1% significance level.

Alternatively, the p-value can be calculated:

$$p(x) = \Pr_{\mu=300}\left(\frac{\bar{X} - \mu}{\mathrm{se}(\bar{X})} \leq \frac{\bar{x} - \mu}{\mathrm{se}(\bar{X})}\right) = \Phi\,(-3.121) = 0.0009$$

Either way, the conclusion is that there is significant evidence which supports the claim that the bottled coffee is less than the advertised value of 300 ml.

[1] Check out §10.3.2 in C&B

5.7 Exercises

1. Suppose we observe n iid Bern(θ) random variables, denoted by $Y_1, \ldots, Y_n$. Show that the LRT of $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$ will reject H_0 if $\sum_{i=1}^n Y_i > c$.

2. A sample X of size 1 was obtained where X has one of the following distributions:

X	H_0	H_1
x_1	0.2	0.1
x_2	0.3	0.4
x_3	0.3	0.1
x_4	0.2	0.4

   (a) Compare the likelihood ratio W for each possible value X, and order the x_i according to W.

   (b) What is the likelihood ratio test of H_0 versus H_1 at level $\alpha = 0.2$? What is the test at level $\alpha = 0.5$?

3. A random sample, $X_1, \ldots, X_n$ is drawn from a Pareto population with pdf

$$f(x|\theta, \nu) = \frac{\theta \nu^\theta}{x^{\theta+1}} \mathbb{1}_{[\nu,\infty)}(x), \theta, \nu > 0.$$

   (a) Find the MLEs of θ and ν.

   (b) Show that the LRT of

$$H_0 : \theta = 1, \nu \text{ unknown} \qquad \text{v.s.} \qquad H_1 : \theta \neq 1, \nu \text{ unknown}$$

   has critical region of the form $\{x \mid T(x) \leq c_1 \text{ or } T(x) \geq c_2\}$, where $0 < c_1 < c_2$ and

$$T = \log \left[\frac{\prod_{i=1}^n X_i}{\{\min_i X_i\}^n} \right].$$

4. We will derive the distribution of the test statistic for comparing two normal means from two independent samples with unknown variances. Let

   - $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{N}(\mu_x, \sigma^2)$; and
   - $Y_1, \ldots, Y_m \overset{\text{iid}}{\sim} \text{N}(\mu_y, \sigma^2)$ independently of the Xs.

   Denote by $\bar{X}$ and $\bar{Y}$ the sample mean of the Xs and Ys respectively. Further let S_x^2 and S_y^2 denote the unbiased sample variance for X and Y respectively, and let

$$S^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}$$

   be the **pooled sample variance** of the data.

   (a) Write down the distribution of

      i. $\bar{X} - \bar{Y}$
      ii. $(n-1)S_x^2/\sigma^2 + (m-1)S_y^2/\sigma^2$

   (b) Based on your answers to part (a), derive the distribution of the statistic

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{S\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

   (c) Explain how you would use the above statistic to conduct a test of the hypothesis $H_0 : \mu_x = \mu_y$.

5. Let $X_1, \ldots, X_n$ be an independent random sample from N(μ, σ^2), where both the mean and variance parameters are unknown. We shall derive a statistical test for σ^2.

(a) Let $X_1, \ldots, X_n$ be an independent random sample from $\mathrm{N}(\mu, \sigma^2)$, where both the mean and variance parameters are unknown. We shall derive a statistical test for σ^2.

(b) For testing $H_0 : \sigma^2 = \sigma_0^2$ against $H_1 : \sigma^2 \neq \sigma_0^2$, write down the likelihood ratio test statistic and show that the rejection region of this test simplifies to, for some constant k,

$$R = \left\{ x \;\middle|\; \left(\frac{\hat{\sigma}^2}{\sigma_0^2}\right)^{n/2} \exp\left(-\frac{n}{2} \cdot \frac{\hat{\sigma}^2}{\sigma_0^2}\right) \leq k \right\}.$$

(c) By considering the function $f(x) = x^a e^{-ax}$, argue that the corresponding rejection regions are

$$\frac{\hat{\sigma}^2}{\sigma_0^2} \leq k_1 \qquad \text{or} \qquad \frac{\hat{\sigma}^2}{\sigma_0^2} \geq k_2$$

for some constants k_1 and k_2.

(d) Find k_1 and k_2 that makes the size of test 0.05.

Remark: this test is **exact** for normal samples and we do not require the use of asymptotics.

6. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(\theta, 1)$. Consider testing

$$H_0 : \theta = 0 \qquad \text{v.s.} \qquad H_1 : \theta = 1$$

Let the rejection region be $R = \{x \mid T(x) > c\}$ where $T(x) = n^{-1} \sum_{i=1}^n X_i$.

(a) Find c so that the test has size α.
(b) Find the power under H_1, i.e. find $B(1)$.
(c) Show that $B(1) \to 1$ as $n \to \infty$.

7. Let $\hat{\theta}_n$ be the MLE of a parameter θ and let $\mathrm{se}(\hat{\theta}_n)$ be the standard error for the parameter θ. Consider testing

$$H_0 : \theta = \theta_0 \qquad \text{v.s.} \qquad H_1 : \theta \neq \theta_0.$$

Using the Wald test with rejection region $R = \{x \mid |Z_n| > z(\alpha/2)\}$, where

$$Z_n = \frac{\hat{\theta}_n - \theta_0}{\mathrm{se}(\hat{\theta}_n)},$$

and $z(a)$ is the top a-th point, $0 \leq a \leq 1$, of the standard normal distribution, show that the power of the test $B(\theta) \to 1$ as $n \to \infty$ for any value of $\theta > \theta_0$.

8. A survey of the use a particular product was conducted in four areas, and a random sample of 200 potential users was interviewed in each area. In area i, for $i = 1, 2, 3, 4$, X_i of the 200 said that they used the product. Construct a likelihood ratio test to test whether the proportion of the population using the product is the same in each of the four areas. Carry out the test at 5% level for the case $X_1 = 76$, $X_2 = 53$, $X_3 = 59$ and $X_4 = 48$.

9. In a given city it is assumed that the number of automobile accidents in a given year follows a Poisson distribution. In past years, the average number of accidents per year was 15, and this year it was 10. Is it justified to claim that the accident rate has dropped? *Hint: Cast this into a statistical testing problem: Define the null hypothesis, calculate the p-value of observing the data under the null, and give your conclusion.*

10. The number of chocolate chips in a packet of Chipsmore cookies is well described by a Poisson distribution with mean 130 (chocolate chips per packet). Following the Kraft takeover of Cadbury (who produces Chipsmore cookies), the mean number of chocolate chips per packet reduced to 75. Fans of the beloved cookie bellowed in anger at the apparent evidence that Kraft has reduced the number of chocolate chips in Chipsmore cookies. Assess the apparent evidence and come to your own conclusion. *Hint: Set up a hypothesis, calculate the likelihood ratio, and obtain an approximate p-value.*

11. In 1861, 10 essays appeared in the New Orleans Daily Crescent. They were signed "Quintus Curtius Snodgrass" and some people suspected they were actually written by Mark Twain. To investigate this, we will consider the proportion of three letter words found in an author's work. From 8 of Twain's essays, the proportions are:

0.225	0.262	0.217	0.240	0.230	0.229	0.235	0.217

From 10 Snodgrass essays, the proportions are:

0.209	0.205	0.196	0.210	0.202	0.207	0.224	0.223	0.220	0.201

Perform a Wald test for equality of the means. Report the p-value and a 95% confidence interval for the difference of means. What do you conclude?

Hand-in questions

1. Suppose that $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are two independent random samples from two exponential distributions with mean θ and μ respectively.

   (a) Find the LRT of $H_0 : \theta = \mu$ versus $H_1 : \theta \neq \mu$. **[4 marks]**

   (b) Show that the test in part (a) can be based on the statistic

   $$T = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} X_i + \sum_{j=1}^{m} Y_j}.$$

   [2 marks]

   (c) Specify the asymptotic distribution of the LRT statistic under H_0 found in a. **[1 mark]**

2. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Unif}(0, \theta)$ and let $Y = \max\{X_1, \ldots, X_n\}$. We want to test $H_0 : \theta = 1/2$ versus $H_1 : \theta > 1/2$.

   (a) Explain briefly why the Wald test would not be appropriate in this case. **[1 mark]**

   Suppose we decide to test this hypothesis by rejecting H_0 when $Y > c$.

   (b) Derive an expression for the power function $B(\theta)$. **[2 marks]**

   (c) What choice of c will make the size of the test 0.05? **[2 marks]**

   (d) In a sample of size $n = 20$ with $Y = 0.48$, what conclusion about H_0 would you make? **[2 marks]**

Chapter 6

Interval estimation

Learning objectives

By the end of this chapter, you will be able to:

- do this
- do that
- and this

Readings

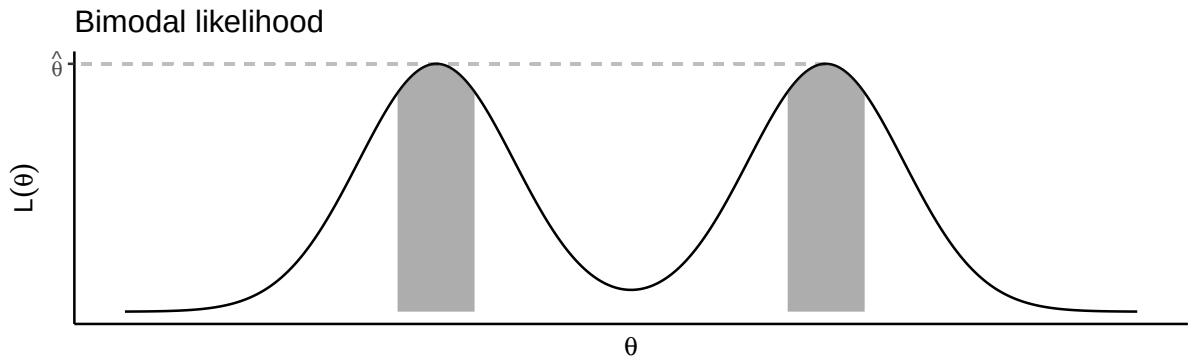
- Casella and Berger (2002)
  - Chapter 9, sections 9.1, 9.2 (9.2.1 and 9.2.2 only), 9.3 (9.3.1 only)
  - Chapter 10, section 10.4.
- Wasserman (2004)
  - Chapter 6, section 6.3.2
  - All of Chapter 8 (Bootstrap)
- Topics not covered here: Bayesian intervals, pivots based on cdfs, test-related optimality, Bayesian optimality, loss function optimality, sinterval using core statistic

6.1 Introduction

The task: to report a *set* $C \subset \Theta$ of plausible values for the unknown parameter θ, rather than a single point estimate. The set $C = C(x)$ is

- a set determined by the value of the observed data $X = x$ (thus, the set is a random variable!); and
- will often be an *interval* in $\mathbb{R}$ (if $\theta \in \mathbb{R} =: \Theta$)–hence 'interval estimation'.

Sometimes, the set of most plausible values may not be an interval.



6.1.1 Coverage probability

We'll start with some formal definitions. Let $C(X)$ be a region of the parameter space Θ, determined by the sample X.

Definition 6.1 (Coverage probability)**.** For any given value of θ, the coverage probability of $C(X)$ is

$$\Pr_{\theta}(\theta \in C(X)) =: c(\theta)$$

In words: coverage is the proportion of times that the (random) interval $C(X)$ contain the parameter value of interest θ. Of course, we are interested how well the interval covers the **true value** of the parameter.

6.1.2 Confidence regions

Since we do not know the true value of θ, we can only guarantee a coverage probability equal to the infimum of $c(\theta)$ (called the *confidence coefficient*). We call such a set a *confidence region*.

Definition 6.2 (Confidence region)**.** The set $C(X)$ is said to be a confidence region with confidence coefficient c if

$$c = \inf_{\theta} c(\theta).$$

In applications, c is typically *fixed* at some suitably large value such as 95% or 99%. That is, we want to "build" a confidence region that has a high chance of capturing the true value of θ.

Some remarks:

1. The random variable here is the set $C(X)$. The confidence coefficient is simply a statement about the repeated sampling properties of such a set.

   $C(X)$ includes the true θ in at least $100c\%$ of samples.

In a frequentist setting, $\Pr_{\theta}(\theta \in C(X))$ does not refer to "the probability of θ being in C'' (however, in the Bayesian setting it does). Rather, these probability statements refer to X and its randomness, and not θ.

2. We have so far more generally described a <u>set</u> $C(X)$, but if it is a random *interval*, $C(X) = [L(X), U(X)]$ say, then $C(X)$ is said to be a *confidence interval* with confidence coefficient c.

3. Estimating an unknown parameter θ with a set, rather than a point, seems imprecise. However, we gain some assurance that we capture the true value θ within the set.

Example 6.1. For a sample $X_1, X_2, X_3, X_4 \overset{\text{iid}}{\sim} \text{N}(\mu, 1)$, an interval estimator of μ could be

$$C(X) = [\bar{X} - 1, \bar{X} + 1].$$

That is, we assert that μ is within this interval. Realise that the probability that we are exactly correct when we estimate μ by $\bar{X}$ is $\Pr(\mu = \bar{X}) = 0$. On the other hand,

$$\begin{aligned}\Pr\left(\mu \in [\bar{X} - 1, \bar{X} + 1]\right) &= \Pr(-1 \leq \bar{X} - \mu \leq 1) \\ &= \Pr\left(-2 \leq \sqrt{4}(\bar{X} - \mu) \leq 2\right) \approx 0.95\end{aligned}$$

Thus, we have a 95% chance of covering the unknown parameter with our interval estimator. Sacrificing precision in our estimate results in an increased confidence of a true assertion.

Expanding on the previous example: Let $\mu = 0$ be the true value. A random sample of size 4 is obtained as follows

```
set.seed(123)
(X <- rnorm(4, mean = 0, sd = 1))
```

```
## [1] -0.56047565 -0.23017749  1.55870831  0.07050839
```

```
mean(X)
```

```
## [1] 0.2096409
```

A 95% confidence interval based on the sample mean is

```
c(mean(X) - 1, mean(X) + 1)
```

```
## [1] -0.7903591  1.2096409
```

In this case, the true value $\mu = 0$ is indeed contained within the interval. If we repeated this experiment many times, what proportion of the intervals would contain $\mu = 0$?

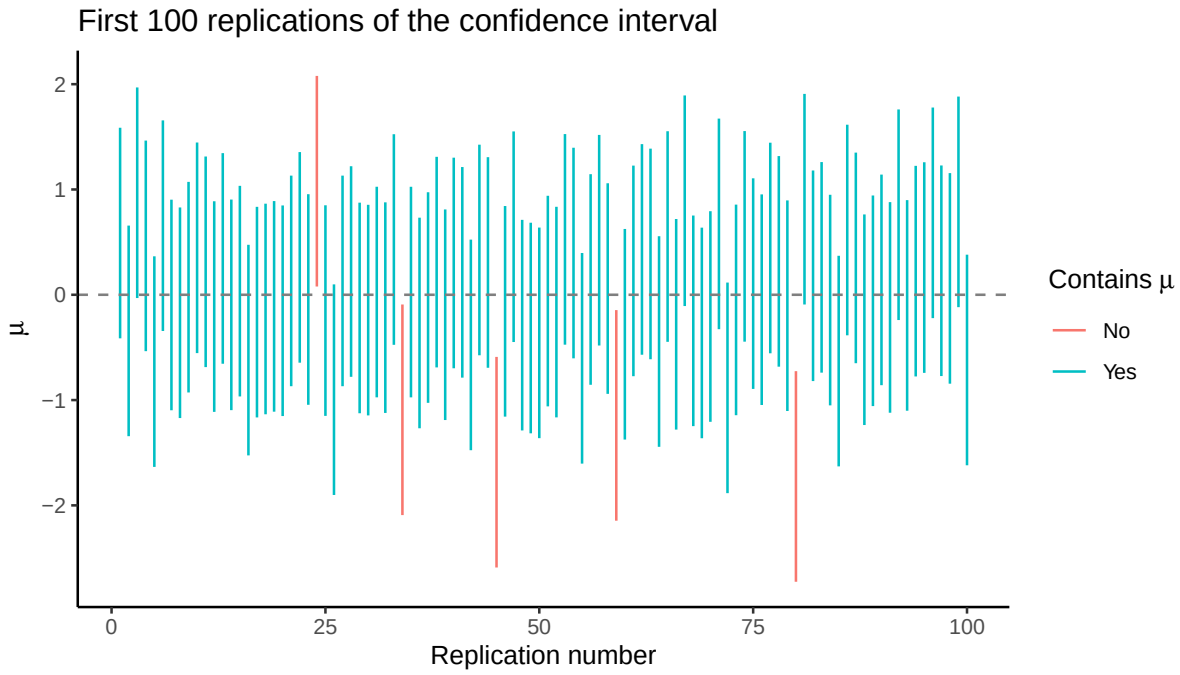
Here's the `R` code:

```
B <- 10000  # number of replications
res <- data.frame(L = rep(NA, B), U = NA, contain = NA)  # prepare the results data frame

for (i in 1:B) {
  X <- rnorm(4, mean = 0, sd = 1)
  L <- mean(X) - 1
  U <- mean(X) + 1
  contain <- (L <= 0) & (0 <= U)  # is 0 contained?
  res[i, ] <- c(L, U, contain)
}
mean(res$contain)  # coverage rate
```

```
## [1] 0.9537
```

As expected, we get a ~95% coverage with the interval $[\bar{X} - 1, \bar{X} + 1]$. Graphically, we can see this below. Of the first 100 random replications and construction of confidence intervals, here exactly 5 do not contain the true value, whereas 95 confidence intervals contain the true value (95%).



6.1.3 Methods for obtaining confidence regions

We will consider two general approaches:

1. Use of a *pivot*
2. Inversion of a hypothesis test

As we shall see, the second is really just a special case of the first.

In the same spirit, large-sample theory of maximum likelihood and of likelihood ratio tests will be found to deliver approximate confidence regions in situations where it is hard to evaluate coverage probabilities exactly.

6.2 Pivots

Definition 6.3 (Pivot)**.** Suppose that the distribution of X is determined by an unknown parameter θ. A *pivotal quantity*, or just pivot for short, is any function $Q(X, \theta)$ whose distribution is the same for all values of θ.

That is, the random variable $Q(X, \theta)$ is independent of all parameters θ: The function $Q(X, \theta)$ will usually explicitly contain both parameters and statistics, but for any set $\mathcal{A}$, $\Pr_\theta(Q(X, \theta) \in \mathcal{A})$ cannot depend on θ. From this, we can construct a confidence set for θ by

$$\{\theta \mid Q(x, \theta) \in \mathcal{A}\}$$

- Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$. Here, the three functions

$$Q_1 = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \qquad Q_2 = \frac{\bar{X} - \mu}{S/\sqrt{n}} \qquad Q_3 = \frac{(n-1)S^2}{\sigma^2}$$

  are all pivots. Q_2 and Q_3 may be used respectively for interval estimation of μ and σ.

- Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$. Here, λ is a scale parameter, so each X_i/λ is a pivot, and so is

$$Q(X) = \bar{X}/\lambda.$$

What are their distributions? Check that the distribution of $\bar{X}/\lambda$ is $\Gamma(n, 1/n)$. Hint: First check that $X_i/\lambda \sim \text{Exp}(1)$!

6.2.1 From pivot to confidence interval

Let $Q(X, \theta)$ be a pivot, and c a specified confidence coefficient (such as $c = 0.95$).

Proposition 6.1 (Pivotal confidence interval)**.** *Suppose we can find constants a and b such that*

$$\Pr_\theta(a \le Q(X, \theta) \le b) = c.$$

Then, $C(X) = \{\theta \,|\, a \le Q(X, \theta) \le b\}$ is a 100c% confidence interval for θ

Proof. This immediately follows from the definition. □

Example 6.2. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{N}(\mu, \sigma^2)$. To construct a confidence interval for μ, let's use the pivot

$$Q = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

So if a and b are such that

$$\Pr(Q \le a) = \Pr(Q \ge b) = (1 - c)/2$$

then $a = -b$ and

$$\Pr_\mu(-b \le Q \le b) = c \Leftrightarrow \Pr_\mu\left(\bar{X} - bS/\sqrt{n} \le \mu \le \bar{X} + bS/\sqrt{n}\right) = c.$$

Thus, the interval

$$C(X) = \left[\bar{X} - b\frac{S}{\sqrt{n}}, \bar{X} + b\frac{S}{\sqrt{n}}\right]$$

is a $100c\%$ confidence interval for μ.

As an illustration, suppose that $n = 20$, $\bar{X} = 8.31$ and $S = 1.97$. Then,

$$C(X) = 8.31 \pm 2.093 \cdot \frac{1.97}{\sqrt{20}} = [7.38, 9.23].$$

is a 95% confidence interval for μ.

Check, from the statistical tables, that $b = 2.093$ for $c = 0.95$ and $n = 20$.

Example 6.3. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{N}(\mu, \sigma^2)$. To construct a confidence interval for σ^2, let's use the pivot

$$Q = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

So if a and b are such that

$$\Pr(Q \le a) = \Pr(Q \ge b) = (1 - c)/2$$

then

$$\Pr_{\sigma^2}(a \le Q \le b) = c \Leftrightarrow \Pr_{\sigma^2}\left((n-1)S^2/b \le \sigma^2 \le (n-1)S^2/a\right) = c.$$

Thus, the interval

$$C(X) = \left[\frac{(n-1)S^2}{b}, \frac{(n-1)S^2}{a}\right]$$

is a $100c\%$ confidence interval for σ^2.

As an illustration, suppose that $n = 20$, $S^2 = 4.8$. Then,

$$C(X) = \left[\frac{19 \times 4.8}{32.85}, \frac{19 \times 4.8}{8.907}\right] = [2.78, 10.24].$$

is a 95% confidence interval for σ^2.

Check, from the statistical tables, that $a = 8.907$ and $b = 32.85$ for $c = 0.95$ and $n = 20$

Notice how wide this interval is! I.e., the point estimate of σ^2 was $S^2 = 4.8$, while the 95% confidence interval includes values more than twice that.

Accurate estimation of a variance, in general, requires n to be fairly large. $n = 20$ is clearly not large enough to allow σ^2 to be determined very accurately.

Consider $n = 250$. Then the lower and upper limits of the χ^2_{249} are 207.2 and 294.6 respectively. The confidence interval is then

$$C(X) = \left[\frac{249 \times 4.8}{294.6}, \frac{249 \times 4.8}{207.2}\right] = [4.06, 5.77].$$

Example 6.4. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$. To construct a confidence interval for λ, we could use $\bar{X}/\lambda \sim \Gamma(n, 1/n)$, but it's much more convenient to use

$$Q = \frac{2n\bar{X}}{\lambda} \sim \Gamma(2n/2, 2) \equiv \chi^2_{2n}.$$

Here, we have used the fact that if $Y \sim \Gamma(\alpha, \beta)$ with $\alpha = k/2$ and $\beta = 2$, then $Y \sim \chi^2_k$.

So if a and b are such that $\Pr(Q \leq a) = \Pr(Q \geq b) = (1 - c)/2$, then

$$\Pr_\lambda(a \leq Q \leq b) = c \Leftrightarrow \Pr_\lambda\left(2n\bar{X}/b \leq \lambda \leq 2n\bar{X}/a\right) = c.$$

Thus, the interval

$$C(X) = \left[\frac{2n\bar{X}}{b}, \frac{2n\bar{X}}{a}\right]$$

is a $100c\%$ confidence interval for λ.

As an illustration, suppose that $n = 20$ and $\bar{X} = 8.3$. Then,

$$C(X) = \left[\frac{2 \times 20 \times 8.3}{59.34}, \frac{2 \times 20 \times 8.3}{24.43}\right] = [5.59, 13.59].$$

is a 95% confidence interval for μ.

::: {.mycheck} Check, from the statistical tables, that $a = 24.43$ and $b = 59.34$ for $c = 0.95$:::

6.3 Inverting a test statistic

Suppose that W_{θ_0} is a test statistic measuring the evidence against $H_0 : \theta = \theta_0$. When X is continuous, we saw that the p-value $p_{W_{\theta_0}}(X)$ is distributed as Unif$(0, 1)$ under H_0. Hence, the p-value itself is a pivot since it is free of θ!

Proposition 6.2 (Confidence interval from pivoting p-values)**.**

$$C(X) = \left\{\theta_0 \;\middle|\; p_{W_{\theta_0}}(X) \geq 1 - c\right\}$$

is a $100c\%$ confidence region for θ.

Proof.

$$\Pr_\theta(\theta \in C(X)) = \Pr_\theta\left(p_{W_{\theta_0}}(X) \geq 1-c\right) = \int_{1-c}^{1} \mathrm{d}k = c.$$

□

Let $A(\theta) = \{x \mid W_\theta(x) < w\} = R^c$ for some constant w be the acceptance region of a hypothesis test, i.e. the set in the sample space such that H_0 is "accepted". For a confidence region $C(X)$ with confidence coefficient c, include all those θ values which, when tested, would result in a p-value of at least $1-c$. That is, we want the set of X such that

$$p_W(X) = \Pr(W(X) \in R) = 1 - \Pr(W(X) \in A) \geq 1-c$$

For example,

- For a 95% confidence region, include in $C(X)$ all those values of θ_0 which are such that the p-value of evidence against H_0 is at least 0.05.
- Or, in terms of the Neyman-Pearson approach: include in $C(X)$ all those values of θ that would not be rejected by a test of size 0.05.

We'll look at a more concrete example next.

Example 6.5. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$ with σ^2 known, and consider testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$. Previously, we saw that for a fixed size α, the rejection region is given by

$$R = \left\{x \;\middle|\; \left|\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right| \geq z(\alpha/2)\right\}$$

where $z(\alpha)$ is the top-α point of the standard normal distribution. The test does not reject H_0 should the observed sample $X = x$ fall in the region $\{x \mid |\bar{x}-\mu_0| \leq z(\alpha/2)\sigma/\sqrt{n}\}$. Those values of μ that would not be rejected fall in the region

$$C(X) = \left[\bar{x} - z(\alpha/2)\frac{\sigma}{\sqrt{n}}\ ,\ \bar{x} + z(\alpha/2)\frac{\sigma}{\sqrt{n}}\right],$$

which makes up a $100(1-\alpha)\%$ confidence interval for μ.

Why? First note that H_0 is "accepted'' for sample points in the acceptance region

$$A = \left\{x \;\middle|\; \bar{x} - z(\alpha/2)\frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{x} + z(\alpha/2)\frac{\sigma}{\sqrt{n}}\right\} = R^c.$$

Since the test has size α,

$$\Pr_{\mu_0}(W(X) \in R) = \alpha \Leftrightarrow \Pr_{\mu_0}(W(X) \in A) = 1-\alpha.$$

But this probability statement is true for every μ_0. Thus,

$$\Pr(\mu \in C(X)) = \Pr(W(X) \in A) = 1-\alpha =: c.$$

Some remarks.

There is a correspondence between confidence sets and acceptance regions for a hypothesis test:

- A hypothesis test fixes the parameter value (under H_0, $\mu = \mu_0$ say) and asks *what sample values* are consistent with that fixed value, i.e. the test is accepted if it falls in

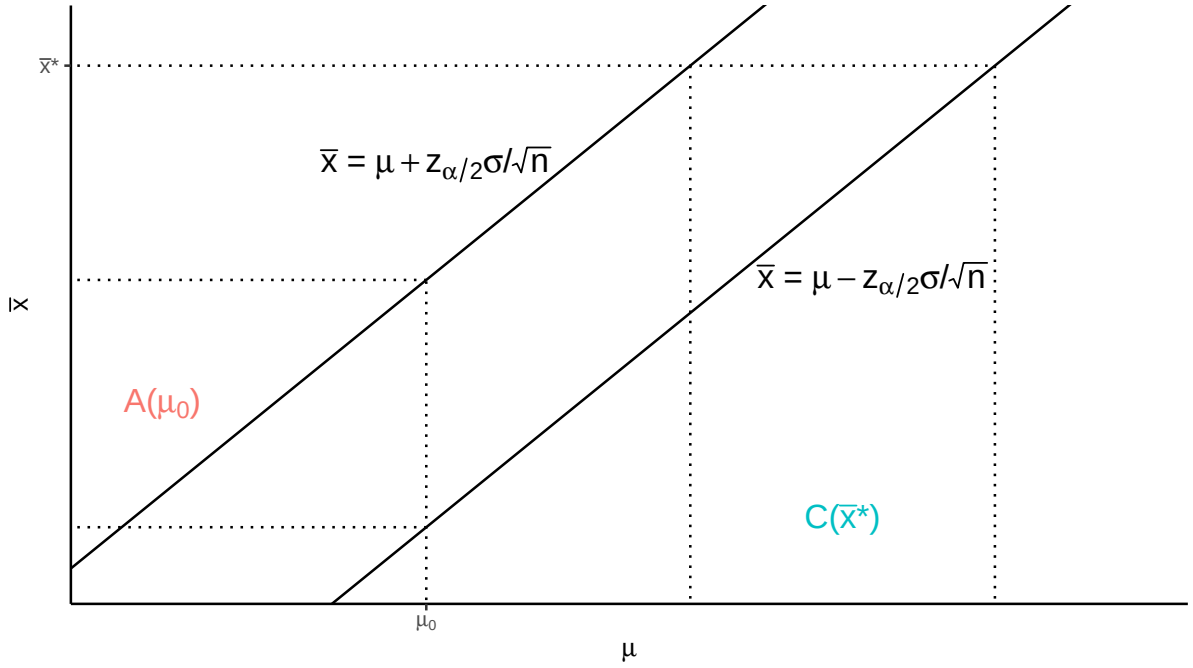
$$A(\mu_0) = \left\{x \;\middle|\; \mu_0 - z(\alpha/2)\frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu_0 + z(\alpha/2)\frac{\sigma}{\sqrt{n}}\right\}$$

- A confidence set fixes the sample value (say we observe $X = x^*$) and asks *what parameter values* make this sample value most plausible, i.e. the confidence set are the values of μ which fall within

$$C(x^*) = \left\{ \mu \;\middle|\; \bar{x} - z(\alpha/2)\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z(\alpha/2)\frac{\sigma}{\sqrt{n}} \right\}$$

The two are connected by the tautology

$$x \in A(\mu_0) \Leftrightarrow \mu \in C(x).$$



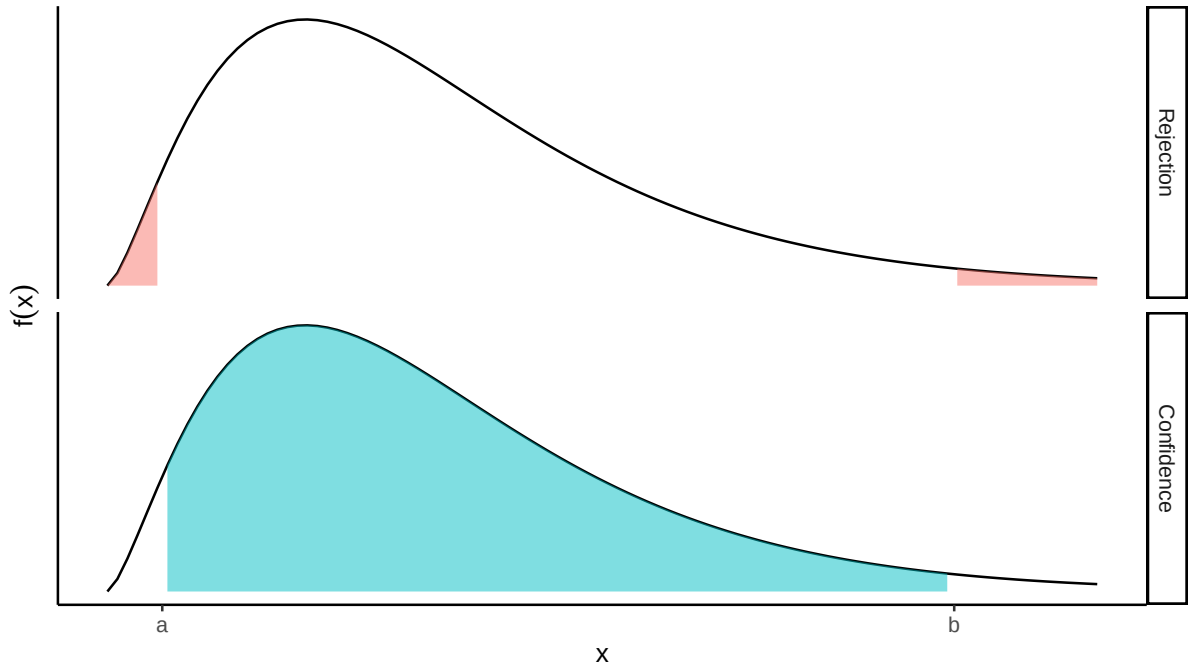
Example 6.6. In Ex. sheet 5, Q5 we looked at a hypothesis test for the variance of a normal distribution. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$ where both parameters are unknown. The LRT test rejects $H_0 : \sigma^2 = \sigma_0^2$ for samples in the rejection region

$$R = \left\{ x \;\middle|\; \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n\sigma_0^2} \leq k_1 \;\text{ or }\; \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n\sigma_0^2} \geq k_2 \right\}$$

The acceptance region can be alternatively written as

$$A = \left\{ x \;\middle|\; a \leq \frac{(n-1)s^2}{\sigma_0^2} \leq b \right\}$$

for some constants a and b based on the χ^2_{n-1} distribution. From this, we can see that we get the same confidence interval for σ^2 based on a pivotal quantity as in Example 6.3.



6.3.1 Discrete distributions

When X is discrete, the p-value is no longer uniform under H_0. The p-value in that case is *stochastically greater* than a Unif$(0,1)$ r.v. in the sense that its cdf F satisfies $F(x) \leq x$, $\forall x$.

Proof. For a continuous r.v. T, we saw that $Y = F_T(T) \sim \text{Unif}(0,1)$. However, if T is discrete[1], the inverse F_T^{-1} is not defined, and

$$F_Y(y) = \Pr(Y \leq y) = \Pr(F_T(T) \leq y) \leq y.$$

Now for any data x,

$$p_W(x) = \Pr_{\theta_0}(W(X) \geq W(x)) = \Pr_{\theta_0}(-W(X) \leq -W(x)) = F_{-W}(-W).$$

Let $Y = -W$ which is discrete, so by the above, the p-values are *stochastically greater* than a Unif$(0,1)$ r.v.. □

As a result, if X is discrete, the construction of $C(X)$ as in Proposition 6.2 above results in a *conservative* confidence region. A conservative confidence region allows for a large range with greater probability that the parameter falls in that range.

Proposition 6.3 (p-value inversion gives a conservative confidence region)**.** *When X is discrete,*

$$C(X) = \left\{\theta_0 \;\middle|\; p_{W_{\theta_0}}(X) \geq 1 - c\right\}$$

is a 100c% confidence region for θ, but this confidence region is said to be conservative.

Proof.

$$\Pr_\theta\left(\theta \in C(X)\right) = \Pr_\theta\left(p_{W_{\theta_0}}(X) \geq 1 - c\right) \geq 1 - (1 - c) = c$$

□

[1] See here: https://stats.stackexchange.com/q/73778

Example 6.7. Suppose that X is a single binary random variable with

$$\Pr_\theta(X = 1) = 1 - \Pr_\theta(X = 0) = \theta, \qquad 0 < \theta < 1.$$

Consider the LR test of $H_0 : \theta = \theta_0$. The MLE is $\hat{\theta} = X$, so the LR statistic is

$$W_{LR}(X) = \frac{L(\hat{\theta}|X)}{L(\theta_0|X)} = \frac{\hat{\theta}^X(1-\hat{\theta})^{1-X}}{\theta_0^X(1-\theta_0)^{1-X}} = \begin{cases} \frac{1}{\theta_0} & X = 1 \\ \frac{1}{1-\theta_0} & X = 0. \end{cases}$$

Thus, the p-value based on the observed data $X = x$ is

$$p_{W_{LR}}(x) = \Pr_{\theta_0}\left(W_{LR}(X) \geq W_{LR}(x)\right) = \begin{cases} \theta_0 & |x - \theta_0| > 1/2 \\ 1 & |x - \theta_0| \leq 1/2 \end{cases}$$

Suppose we set the confidence coefficient to be $c = 0.95$. Then, included in $C(X)$ are all values of θ_0 such that $p_{W_{LR}}(x) \geq 0.05$. If $x = 1$, this is the interval $[0.05, 1)$; and by symmetry if $x = 0$ it is $(0, 0.95]$.

Coverage of such a confidence interval?

$$c(\theta) = \Pr_\theta(\theta \in C(X)) = \begin{cases} 1 - \theta & \theta < 0.05 \\ 1 & 0.05 \leq \theta \leq 0.95 \\ \theta & \theta > 0.95 \end{cases}$$

so we see $c(\theta) > 0.95$ for all θ, so the confidence interval is indeed conservative.

6.4 Desirable confidence sets

We have seen two methods for deriving confidence sets (and there are others), and in fact different methods yield different confidence sets. Is there a best one?

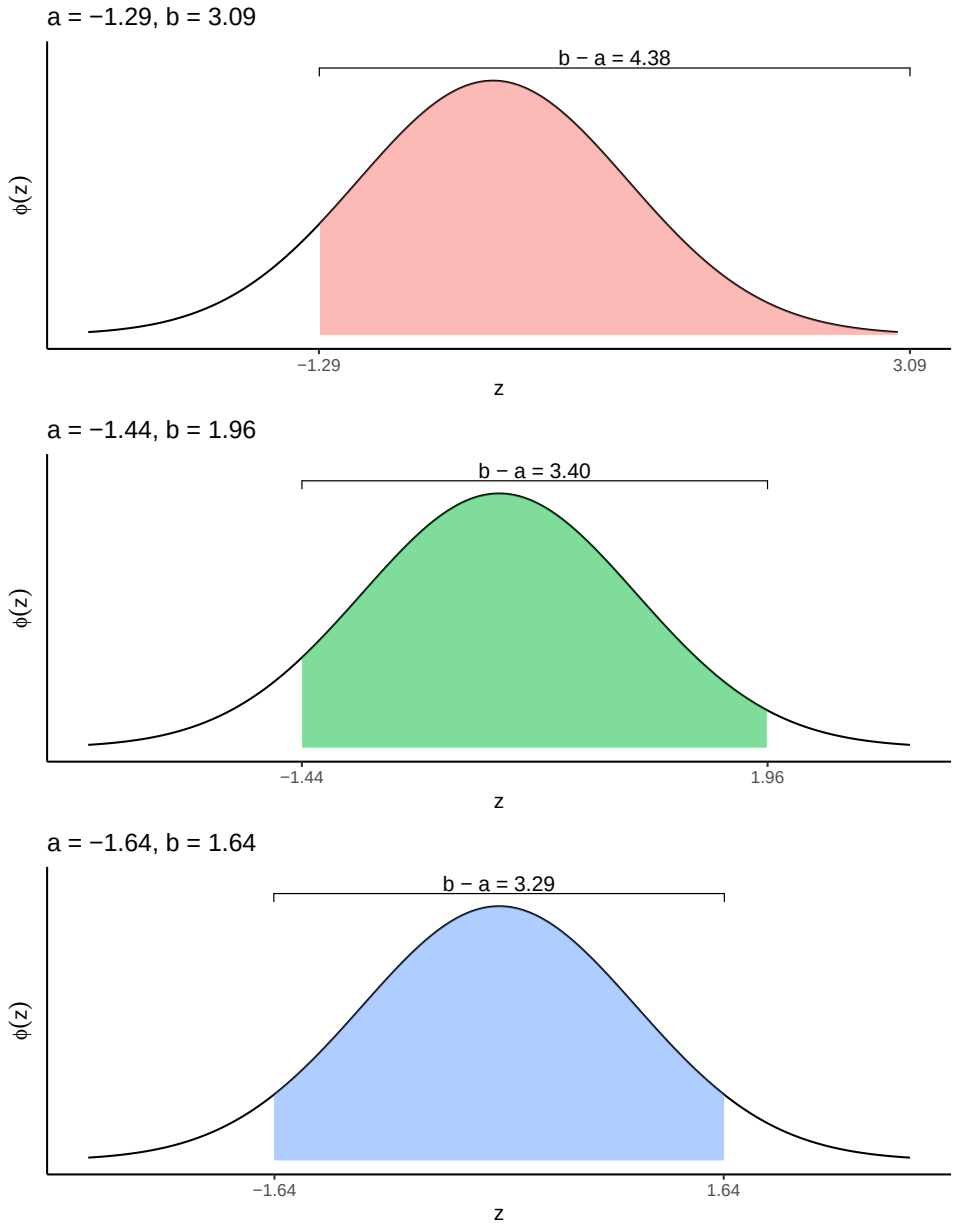
We desire a confidence set $C(X)$ which has

- small size (for a confidence interval $C(X) = [L(X), U(X)]$, this means its length $U(X) - L(X)$); and
- large coverage probability $\Pr(\theta \in C(X))$.

Often hard to construct–clearly, to increase coverage we need only increase its size.

In Example 6.5, we saw the use of the top and bottom $\alpha/2$ points of the standard normal being used. I.e., the size α was split equally among the two tails of the distribution. Is this necessary?

Suppose $1 - \alpha = 0.9$. Then any of the following pairs give 90% intervals:



It turns out the strategy of splitting α **equally** is optimal if the distribution is unimodal (note: it does not have to be symmetric!).

Theorem 6.1. *Let $f(x)$ be a unimodal pdf. If the interval $[a, b]$ satisfies*

- $\int_a^b f(x)\,\mathrm{d}x = c$*;*
- $f(a) = f(b) > 0$*; and*
- $a \leq x^* \leq b$ *where x^* is the mode of $f(x)$,*

then $[a, b]$ is the shortest among all intervals that satisfy 1.

The proof of this is omitted. See instead C& B Thm 9.3.2.

Example 6.8. Suppose $X \sim \Gamma(\alpha, \beta)$. The quantity $Y = X/\beta$ is a pivot for β, with $Y \sim \Gamma(\alpha, 1)$. We can get a confidence interval by finding a and b to satisfy

$$\Pr(a \leq Y \leq b) = c.$$

However, choosing a and b to satisfy $f_Y(a) = f_Y(b)$ is not optimal, because the interval on β is of the form

$$C(X) = \left\{ x \;\middle|\; \frac{x}{b} \le \beta \le \frac{x}{a} \right\},$$

so the length of the interval is $(1/a - 1/b)x$. That is, it is proportional to $1/a - 1/b$ and not $b - a$.

6.5 Intervals based on ML methods

Recall the following two asymptotics:

1. Asymptotic efficiency of the MLE,

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\text{D}} \text{N}(0, \mathcal{I}_1(\theta)^{-1}).$$

2. Large sample distribution of W_{LR} for testing $H_0 : \theta = \theta_0$,

$$-2 \log \left[\frac{L(\theta_0|X)}{L(\hat{\theta}_n|X)} \right] \xrightarrow{\text{D}} \chi^2_1$$

   (or the one based on Wilk's theorem).

These are two 'automatic' pivots based on the large sample distributions. So quite generally, an approximate confidence region can be based off maximum likelihood methods.

Under certain regularity conditions the MLE is asymptotically normal, and we can make use of the fact that

$$\Pr\left(-z(\alpha/2) \le \sqrt{\mathcal{I}(\theta)}(\hat{\theta}_n - \theta) \le z(\alpha/2) \right)$$

to build a $100(1-\alpha)\%$ confidence set for θ. But this is hard to invert into an interval for θ. So we simplify things by using the observed Fisher information instead.

Definition 6.4. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x|\theta)$, and $\hat{\theta}_n$ be the MLE of θ. The interval

$$\left[\hat{\theta}_n - z(\alpha/2) \cdot \text{se}(\hat{\theta}_n) \;,\; \hat{\theta}_n + z(\alpha/2) \cdot \text{se}(\hat{\theta}_n) \right],$$

with $\text{se}(\hat{\theta}_n) = 1/\sqrt{-l''(\hat{\theta}_n)}$, is an approximate $100(1-\alpha)\%$ confidence interval for θ.

This is otherwise known as the Wald interval.

Definition 6.5. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x|\theta)$, and $\hat{\theta}_n$ be the MLE of θ. The set

$$C(X) = \left\{ \theta \;\middle|\; -2 \log \left[\frac{L(\theta|X)}{L(\hat{\theta}_n|X)} \right] \le \chi^2_1(\alpha) \right\}$$

is an approximate $100(1-\alpha)\%$ confidence interval.

This is simply an inversion of the rejection region for the large-sample LRT test. The confidence region include values of θ such that $2 \log W_{LR}(X) \sim \chi^2_1$ is small. For example, $\Pr(2 \log W_{LR} \le 3.84) \approx 0.95$, so

$$C(X) = \left\{ \theta \;\middle|\; 2l(\theta|X) \ge 2l(\hat{\theta}|X) - 3.84 \right\}$$

is an approximate 95% confidence region for θ.

Which to use?

- Undoubtedly, the Wald interval is simpler computationally. In comparison, the LRT interval demands the solution of a non-linear equation in order to find the end points.

- However, the Wald interval is **not** invariant to a change in parameter, e.g. $\tau = g(\theta)$ (this is also a disadvantage of the Wald test), whereas the LRT interval is.
- The LRT interval approach works much better than the Wald interval when the likelihood is asymmetric or multi-modal.

Example 6.9. Consider a single Poisson count, $Y \sim \text{Poi}(\mu)$. There is no exact pivot in this case, so we'll build some approximate confidence sets.

The log-likelihood gives

$$\begin{aligned} l(\mu|y) &= \text{const.} - \mu + y \log \mu \\ l'(\mu|y) &= -1 + y/\mu \\ l''(\mu|y) &= -y/\mu^2 \end{aligned}$$

From this, we have $\hat{\mu} = y$, while $-l''(\hat{\mu}) = 1/y$ (provided that $y > 0$–the method has problems if not!).

Consider also the parameter transformation $\tau = \log \mu$, which is a fairly standard one to use in Poisson models[2]. Then

$$\begin{aligned} l(\tau|y) &= \text{const.} - e^{\tau} + y\tau \\ l'(\tau|y) &= -e^{\tau} + y \\ l''(\tau|y) &= -e^{\tau} \end{aligned}$$

so (not surprising, since the MLE is invariant to continuous transformations)

$$\hat{\tau} = \begin{cases} \log y & y > 0 \\ -\infty & y = 0 \end{cases}$$

and $-l''(\hat{\tau}) = y$ (notice that the standard error is not invariant).

Approximate 95% confidence intervals for μ…

(a) based on the MLE $\hat{\mu}$:

$$[y - 1.96\sqrt{y}, y + 1.96\sqrt{y}]$$

based on the MLE $\hat{\tau}$ (and converting it back via $\mu = e^{\tau}$):

$$[e^{\log y - 1.96/\sqrt{y}}, e^{\log y + 1.96/\sqrt{y}}]$$

(b) based on the LRT:

$$\{2(-\mu + y \log \mu) \geq 2(-y + y \log y) - 3.84\}$$

Here are some results for $y = 10$ and $y = 50$ comparing the three kinds of confidence intervals.

	(a)	(b)	(c)
$y = 10$	[3.8, 16.2]	[5.4, 18.6]	[5.0, 17.5]
$y = 50$	[36.1, 63.9]	[37.9, 66.0]	[37.4, 65.2]

6.6 The bootstrap method

Bootstrap is a computational method for estimating standard errors and confidence intervals, especially when inference involves a statistic whose distribution is unknown.

Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x|\theta)$, where both f and θ are unknown. We are interested to conduct inference about a statistic $T = T(X_1, \ldots, X_n)$.

[2] If interested, check out Poisson regression (log-linear models)

- If $T(X) = \bar{X}_n$, then the CLT applies as $n \to \infty$ so we can know approximately its distribution and standard error.
- What about other statistics? E.g.
  - Skewness $\gamma = \mathrm{E}\left[(X-\mu)^3\right]/\sigma^3$
  - Kurtosis $\kappa = \mathrm{E}\left[(X-\mu)^4\right]/\sigma^4$

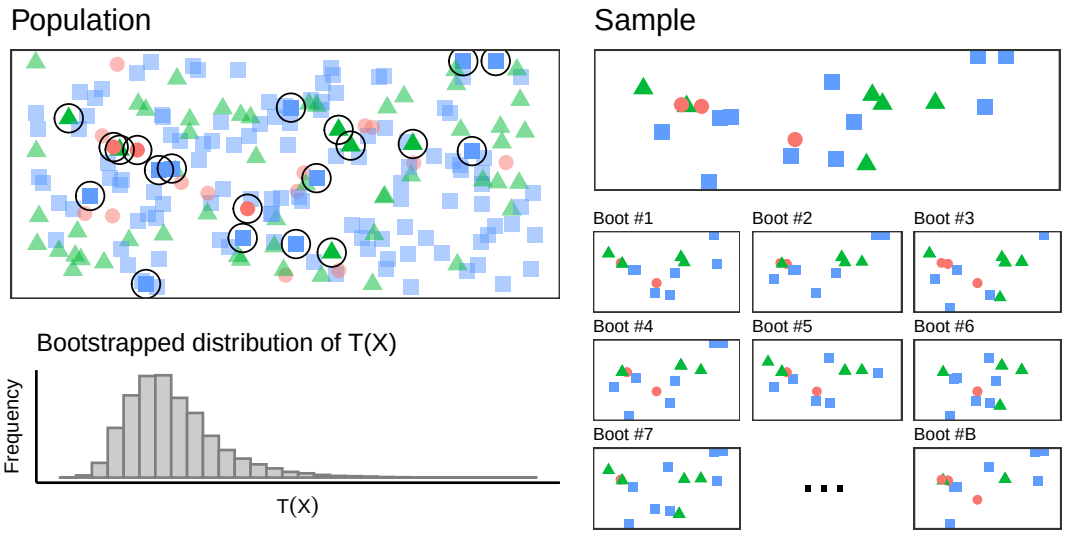


Figure 6.1: Bootstrap.

Main idea

A point estimate $T(X)$ is obtained using a sample from the population. This is all the data we have. We then draw a *bootstrap sample* $\{X_1^*, \ldots, X_n^*\}$ from the sample and calculate the statistic $T(X^*)$. Repeat this many times to get an idea of the *variability* of the statistic.

6.6.1 Empirical distribution

To see why this works, consider the *empirical distribution function* of a data set.

Definition 6.6 (Empirical distribution)**.** Let $X_1, \ldots, X_n$ be iid with common cdf $F(x)$. The empirical distribution function is defined as

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[X_i \leq x]$$

The empirical cdf just counts the number of elements in the sample less than a given value–it is literally doing what a cumulative frequency plot would do. Notice that

- For a fixed x, the r.v. $\mathbb{1}[X_i \leq x]$ is Bernoulli with param. $p = \Pr(X_i \leq x) = F(x)$.
- Hence, $n\hat{F}_n(x) \sim \mathrm{Bin}\left(n, F(x)\right)$, and so we know the mean and variance.
- Importantly, $\hat{F}_n(x)$ is an *unbiased* estimator of $F(x)$.
- It is also consistent: $\hat{F}_n(x) \xrightarrow{\mathrm{P}} F(x)$ by the law of large numbers.

6.6.2 Bootstrap variance estimation

GOAL: To estimate the variance of a statistic $T(X)$

$$\mathrm{Var}_F(T) = \int \left\{T(x) - \mathrm{E}(T(x))\right\}^2 \mathrm{d}F(x)$$

The bootstrap method has two steps:

1. Estimate $\mathrm{Var}_F(T)$ with $\mathrm{Var}_{\hat{F}_n}(T)$, i.e. using the empirical distribution.
2. Approximate $\mathrm{Var}_{\hat{F}_n}(T)$ with $\widehat{\mathrm{Var}}_{\hat{F}_n}(T)$ using simulation, i.e. bootstrap resampling.

So actually there are two sources of error:

$$\mathrm{Var}_F(T) \overset{\text{estimation error}}{\approx} \mathrm{Var}_{\hat{F}_n}(T) \overset{\text{simulation error}}{\approx} \widehat{\mathrm{Var}}_{\hat{F}_n}(T)$$

As a remark, the estimation in Step 1 is typically consistent due to the LLN. Thus, the size of the error depends on the sample size.

6.6.2.1 Step 1: Bootstrap variance estimation

Actually, Step 1 is what we have been doing so far. It simply uses the data to compute the variance of our statistic, assuming that the functional form of $F(x)$ is known.

Example 6.10. Suppose $T(X) = \bar{X}_n$. Then we know that $\mathrm{Var}_F(T) = \sigma^2/n$, where $\sigma^2 = \int (x-\mu)^2 \,\mathrm{d}F(x)$ and $\mu = \int x \,\mathrm{d}F(x)$. Still, this involves and unknown quantity σ^2, so we use an estimate instead:

$$\begin{aligned}\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \hat{\mu})^2 &= \frac{n}{n-1}\cdot\frac{1}{n}\sum_{i=1}^n (x_i-\hat{\mu})^2 \,\mathbb{1}[X_i \le x_i] \\ &= \frac{n}{n-1}\int (x-\hat{\mu})^2 \,\mathrm{d}\hat{F}_n(x),\end{aligned}$$

where $\hat{\mu} = \int x \,\mathrm{d}\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n x_i \,\mathbb{1}[X_i \le x_i] = \bar{x}$.

In the above example, Step 1 is sufficient. But when we cannot write down a simple formula for $\mathrm{Var}_{\hat{F}_n}(T)$ we need to do bootstrap.

6.6.2.2 Step 2: Bootstrap variance estimation

Recap the problem again: From our sample $\{X_1, \dots, X_n\}$ we compute $T = T(X)$, and we want to estimate $\mathrm{Var}_{\hat{F}_n}(T)$ (variance of T using the empirical cdf of X, e.g. think $\hat{\sigma}^2/n$) but we are unable to, for whatever reason.

Hypothetically if we had a "random sample" of our test statistic $\{T_1^*, \dots, T_B^*\}$, where each T_k^* is computed from a new sample $\{X_1^*, \dots, X_n^*\}$ obtained from the empirical cdf, then

$$\widehat{\mathrm{Var}}_{\hat{F}_n}(T) = \frac{1}{B}\sum_{k=1}^B (T_k^* - \bar{T}_B)^2 \xrightarrow{\mathrm{P}} \mathrm{E}_{\hat{F}_n}((T - \mathrm{E}\,T)^2) = \mathrm{Var}_{\hat{F}_n}(T)$$

as $B \to \infty$, so we have found a consistent estimator for $\mathrm{Var}_{\hat{F}_n}(T)$.

The question is, how do we sample from the empirical cdf?

Suppose we observe $X = \{x_1, \dots, x_n\}$. Order these to create

$$x_{(1)}, x_{(2)}, \dots, x_{(n)}$$

where $x_{(1)} = \min_i x_i$ and $x_{(n)} = \max_i x_i$ and $x_{(k)} \le x_{(k+1)}$. By definition of the empirical cdf,

$$\hat{F}_n(x_{(k)}) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}[X \le x_{(k)}] = \frac{k}{n}.$$

Evidently, the empirical cdf assigns mass $1/n$ on each data point x_i. Therefore, to simulate $\{X_1^*, \dots, X_n^*\} \sim \hat{F}_n(x)$, it suffices to draw n observations *with replacement* from $\{X_1, \dots, X_n\}$.

6.6.2.3 Summary of bootstrap procedure

Using the bootstrap procedure below, we may obtain an estimator v_{boot} for $\text{Var}(T)$, the variance of a statistic of interest.

Definition 6.7 (Bootstrap variance estimation).

- Draw $\{X_1^*, \dots, X_n^*\} \sim \hat{F}_n(x)$ by sampling with replacement from the set $\{X_1, \dots, X_n\}$.
- Compute $T^* = T(X_1^*, \dots, X_n^*)$.
- Repeat steps 1 and 2 B number of times to obtain $\{T_1^*, \dots, T_B^*\}$.
- Compute

$$v_{boot} := \widehat{\text{Var}}_{\hat{F}_n}(T) = \frac{1}{B}\sum_{k=1}^{B}(T_k^* - \bar{T}_B)^2$$

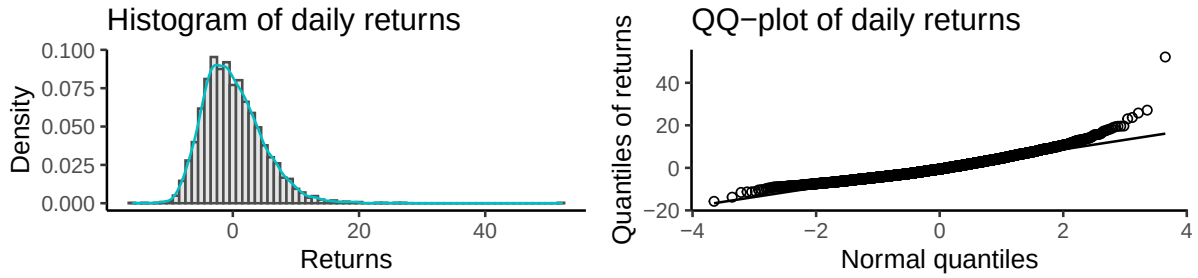
where $\bar{T}_B = B^{-1}\sum_{k=1}^{B} T_k^*$.

The above steps are what is used to calculate the variance of the estimator in practice in a variety of problems where the variance of the estimator would be unobtainable otherwise. Depending on the actual function of the statistic T, the above bootstrap procedure is quite simple to implement, and does not require too much computational power.

Example 6.11. We'll inspect the daily returns of the Shanghai Stock Exchange Composite Index in December 1994. An inspection of plots below all indicate non-normality (positive skew).

The "tailed-ness'' of a distribution is measured by the kurtosis $\kappa = \text{E}\left[(X-\mu)^4\right]/\sigma^4$ and we may use the plug-in estimator below to estimate κ:

$$\hat{\kappa} = \frac{1}{nS^4}\sum_{i=1}^{n}(X_i - \bar{X})^4$$



The estimate kurtosis is $\hat{\kappa} = 7.84$, indicating daily returns are heavy-tailed. In comparison, the kurtosis of any univariate normal distribution is 3. How accurate is this estimate? Use bootstrap to compute the standard errors.

```
mean((x - mean(x)) ^ 4) / sd(x) ^ 4  # estimate of kurtosis
```

```
## [1] 7.840612
```

```
n <- length(x)
B <- 1000
res <- rep(NA, B)  # vector to hold results
for (k in 1:B) {
  xstar <- sample(x = x, size = n, replace = TRUE)
  res[k] <- mean((xstar - mean(xstar)) ^ 4) / sd(xstar) ^ 4
}
head(res)  # this is T*
```

```
## [1] 10.661504  7.766468  4.223420  7.587617  3.980878  7.679705
```

```
sd(res)  # bootstrap standard error
```

```
## [1] 3.136696
```

6.7 Bootstrap confidence intervals

Now that we've seen how to compute the bootstrap standard error, we can build confidence intervals using it There are three kinds of bootstrap cis:

1. Normal bootstrap interval
2. Pivotal bootstrap interval
3. Percentile bootstrap interval

Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} f(x|\theta)$ whose distribution is unknown, and we are interested in constructing a ci for the parameter θ. For each of the cis, we need to obtain bootstrap samples $\{\hat{\theta}_1^*, \dots, \hat{\theta}_B^*\}$ of the estimator $\hat{\theta} = \theta(X_1, \dots, X_n)$ using the procedure in Definition 6.7.

6.7.1 Normal bootstrap interval

From the bootstrap samples obtain

$$\text{se}_{boot}(\hat{\theta}) = \sqrt{\frac{1}{B}\sum_{i=1}^{B}\left(\hat{\theta}_i^* - \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_i^*\right)^2}.$$

Definition 6.8 (Normal bootstrap interval)**.** Suppose the estimator $\hat{\theta}$ for θ is asymptotically normal. The interval

$$\left[\hat{\theta} - z(\alpha/2)\cdot \text{se}_{boot}(\hat{\theta}) \ , \ \hat{\theta} + z(\alpha/2)\cdot \text{se}_{boot}(\hat{\theta})\right],$$

is an approximate $100(1-\alpha)\%$ confidence interval for θ.

The idea is to replace $\text{se}(\hat{\theta}_n)$ in the Wald interval with the bootstrap se. Note that this interval is not very accurate unless the distribution of $\hat{\theta}$ is close to normal.

6.7.2 Bootstrap percentile interval

Arrange the bootstrapped quantities $\hat{\theta}_i^*$ in ascending order to obtain the ordered quantities

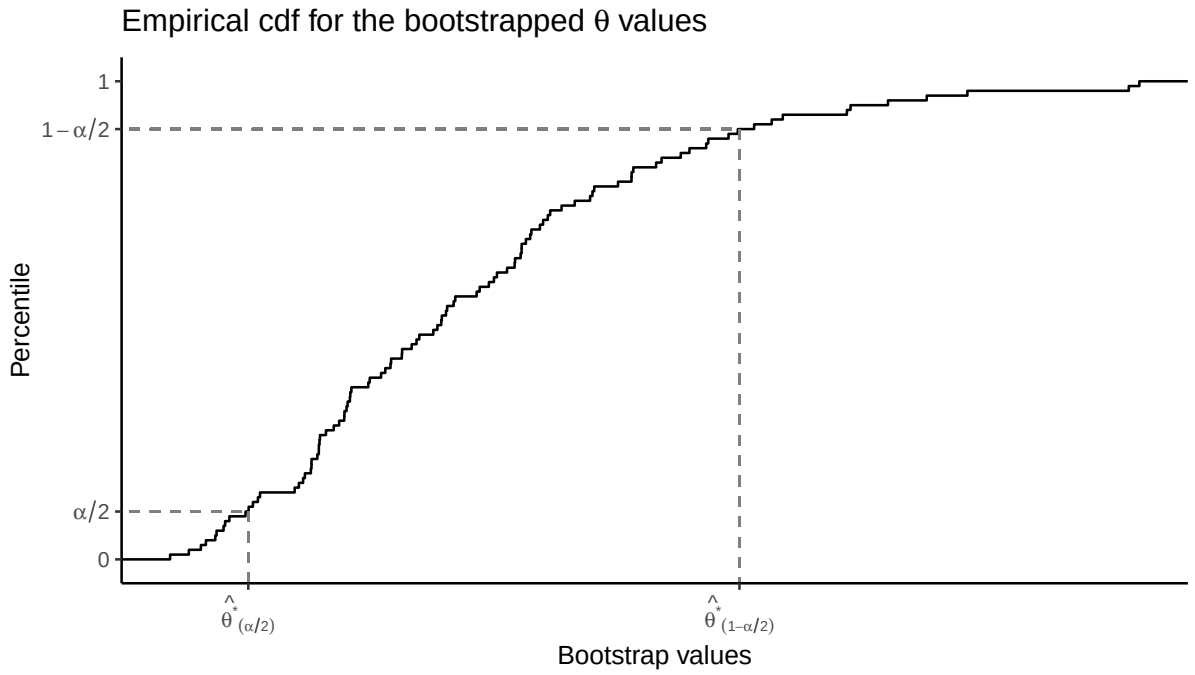
$$\hat{\theta}_{(1)}^*, \hat{\theta}_{(2)}^*, \dots, \hat{\theta}_{(B)}^*.$$

Let $\hat{\theta}_{(\alpha)}^*$ be the $\lfloor B\alpha \rfloor$-th smallest value among the $\hat{\theta}_i^*$. In other words, $100\alpha\%$ of the ordered $\hat{\theta}_{(i)}^*$ are smaller than $\hat{\theta}_{(\alpha)}^*$.

Definition 6.9 (Bootstrap percentile interval)**.** An approximate $100(1-\alpha)\%$ confidence interval based on the bootstrap percentiles is given by

$$\left[\hat{\theta}_{(\alpha/2)}^* \ , \ \hat{\theta}_{(1-\alpha/2)}^*\right]$$

The logic here is that the bootstrap method suggests that the true parameter value for $\hat{F}_n(x)$ will lie in this interval about $100(1-\alpha)\%$ of the time. Hopefully, the ci for θ based on $\hat{F}_n(x)$ will converge to the ci for θ based on $F(x)$.



6.7.3 Bootstrap pivotal interval

Define the pivotal quantity $Q = \hat{\theta} - \theta$, and denote the cdf of Q by $G(r) = \Pr(\hat{\theta} - \theta \leq r)$. Define further the top α point of the distribution of this pivot by $r(\alpha)$ s.t. $G(r(\alpha)) = 1 - \alpha$. The fact that

$$\begin{aligned} 1 - \alpha &= \Pr\left(r(1-\alpha/2) \leq \hat{\theta} - \theta \leq r(\alpha/2)\right) \\ &= \Pr\left(\hat{\theta} - r(\alpha/2) \leq \theta \leq \hat{\theta} - r(1-\alpha/2)\right), \end{aligned}$$

this gives an exact $100(1-\alpha)\%$ confidence interval for θ of the form

$$\left[\hat{\theta} - r(\alpha/2), \hat{\theta} - r(1-\alpha/2)\right].$$

Of course, this is a valid interval if the pivot Q is free of θ, which unfortunately it is not (since its distribution G depends on θ). However, in the bootstrap approach we need not care about this!

The argument is that the behaviour of $Q = \hat{\theta} - \theta$ is not far off from $\hat{Q} = \hat{\theta}^* - \hat{\theta}$, in which case we make use of the estimate of $G(r)$ given by

$$\hat{G}(r) = \frac{1}{B} \sum_{k=1}^{B} \mathbb{1}[\hat{\theta}_k^* - \hat{\theta} \leq r],$$

the empirical distribution using the bootstrap samples $\hat{\theta}_k^*$. We replace $r(\alpha/2)$ and $r(1-\alpha/2)$ by their bootstrap counterparts $r^*(\alpha/2)$ and $r^*(1-\alpha/2)$ s.t. $\hat{G}(r^*(\alpha)) = 1 - \alpha$. Then,

$$\begin{aligned} 1 - \alpha &= \Pr\left(r^*(1-\alpha/2) \leq \hat{\theta}^* - \hat{\theta} \leq r^*(\alpha/2)\right) \\ &\approx \Pr\left(r^*(1-\alpha/2) \leq \hat{\theta} - \theta \leq r^*(\alpha/2)\right) \\ &= \Pr\left(\hat{\theta} - r^*(\alpha/2) \leq \theta \leq \hat{\theta} - r^*(1-\alpha/2)\right), \end{aligned}$$

so we can build a ci based off of this fact.

In practice however, it's easier to use the bootstrap percentiles, since

$$r^*(\alpha) = \hat{\theta}^*_{(1-\alpha)} - \hat{\theta}$$

by definition. It follows that

$$\Pr\left(\hat{\theta} - r^*(\alpha/2) \leq \theta \leq \hat{\theta} - r^*(1-\alpha/2)\right) = \Pr\left(2\hat{\theta} - \hat{\theta}^*_{(1-\alpha/2)} \leq \theta \leq 2\hat{\theta} - \hat{\theta}^*_{(\alpha/2)}\right) \approx 1 - \alpha.$$

Definition 6.10 (Bootstrap pivotal interval)**.** An approximate $100(1-\alpha)\%$ confidence interval based on the bootstrap pivotal quantity is

$$\left[2\hat{\theta} - \hat{\theta}^*_{(1-\alpha/2)} \ , \ 2\hat{\theta} - \hat{\theta}^*_{(\alpha/2)}\right],$$

where $\hat{\theta}^*_{(\alpha)}$ denotes the 100α-th perncetile of the ordered bootstrap estimates $\hat{\theta}^*_i$s.

6.7.4 Which one to use?

In general, all three methods give similar performance, provided that

- the (empirical) distribution of $\hat{\theta}$ is roughly "nice", i.e. unimodal, symmetric, not skewed, unbiased.
- the empirical distribution $F_n(x)$ of the data represents the population distribution $F(x)$ well. If it doesn't, then no bootstrapping method will be reliable[3].

In all cases, these confidence intervals are approximate, i.e. the coverage probability $\Pr(\theta \in C(X))$ is not exactly $1-\alpha$. More accurate methods exist but are not discussed here.

Example 6.12. This example was used by Bradley Efron, the inventor of the bootstrap. The data are LSAT scores (for entrance to law school) and GPA.

i	LSAT	GPA
1	576	3.39
2	635	3.30
3	558	2.81
4	578	3.03
5	666	3.44
6	580	3.07
7	555	3.00
8	661	3.43
9	651	3.36
10	605	3.13
11	653	3.12
12	575	2.74
13	545	2.76
14	572	2.88
15	594	2.96

Each data point is of the form $X_i = (Y_i, Z_i)$, where $Y_i = \text{LSAT}_i$ and $Z_i = \text{GPA}_i$.

The law school is interested in the correlation coefficient

$$\rho = \frac{\iint (y-\mu_y)(z-\mu_z)\,\mathrm{d}F(y,z)}{\sqrt{\int (y-\mu_y)^2\,\mathrm{d}F(y)\int (z-\mu_z)^2\,\mathrm{d}F(z)}}.$$

The plug-in estimate is the sample correlation

$$\hat{\rho} = \frac{\sum_{i=1}^n (Y_i-\bar{Y})(Z_i-\bar{Z})}{\sqrt{\sum_{i=1}^n (Y-\bar{Y})^2 \sum_{i=1}^n (Z_i-\bar{Z})^2}}.$$

The estimated correlation is $\hat{\rho} = 0.776$. Note that $\hat{\rho} \in [0,1]$ and it is not entirely obvious what its distribution might be. Several choices do exist for distributions within the unit interval of course, for instance $\text{Unif}(0,1)$ or the Beta distribution–but are these good distributions to impose on our statistic? Let's use bootstrap to estimate the 95% ci for ρ.

[3] https://stats.stackexchange.com/a/357498

```
(rho <- cor(y, z))  # 'law' data frame in R package 'bootstrap'
```

```
## [1] 0.7763745
```

```
B <- 1000
rhostar <- rep(NA, B)
for (i in 1:B) {
  samp <- sample(1:15, size = 15, replace = TRUE)
  rhostar[i] <- cor(y[samp], z[samp])
}
round(head(rhostar), 3)
```

```
## [1] 0.684 0.898 0.955 0.675 0.910 0.864
```

```
(bootse <- sd(rhostar))  # bootstrap se
```

```
## [1] 0.1269466
```

Now, compute the three kinds of intervals.

```
# normal interval
c(rho - qnorm(0.975) * bootse, min(rho + qnorm(0.975) * bootse, 1))
```

```
## [1] 0.5275637 1.0000000
```

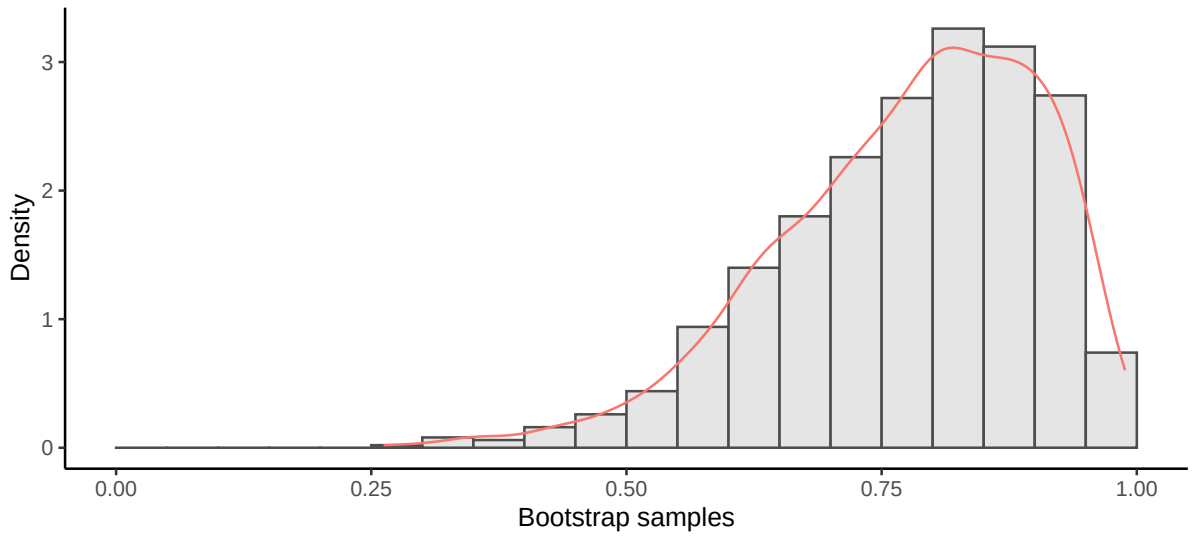
```
# percentile interval
a <- as.numeric(quantile(rhostar, probs = 0.025))
b <- as.numeric(quantile(rhostar, probs = 0.975))
c(a, b)
```

```
## [1] 0.4904234 0.9568777
```

```
# pivotal interval
c(2 * rho - b, min(2 * rho - a, 1))
```

```
## [1] 0.5958713 1.0000000
```

The three methods are not too far off each other, but with a larger sample size they may show closer agreement. The plot below shows the distribution of $\widehat{\rho}^*$ (a bit skewed).



6.8 Exercises

1. Let $X_1, \ldots, X_n$ be iid $N(\theta, 1)$. A 95% confidence interval for θ is $\bar{x} \pm 1.96/\sqrt{n}$. Let p denote the probability that an additional independent observation, X_{n+1}, will fall in this interval. Is p greater than, less than, or equal to 0.95? Prove your answer. *Hint: Consider the distribution of* $X_{n-1} - \bar{X}$.

2. The length (in millimetres) X_i of cuckoos' eggs found in hedge sparrow nests can be modelled with the distribution

$$\Pr(X_i \leq x \mid \alpha, \beta) = \begin{cases} 0 & x < 0 \\ (x/\beta)^\alpha & 0 \leq x \leq \beta \\ 1 & x > \beta \end{cases}$$

   (a) Find a two-dimensional sufficient statistic for (α, β).

   (b) The following data for the length of cuckoo's eggs were collected:

22.0	23.9	20.9	23.8	25.0	24.0	21.7
23.8	22.8	23.1	23.1	23.5	23.0	23.0

   Find the MLEs of α and β.

   (c) Construct 95% confidence interval estimate for β based on this data set, assuming that α is known and equal to its MLE. *Hint: The parameter* β *is a scale parameter, so each* X_i/β *is a pivot.*

3. Derive a confidence interval for a binomial p by inverting the LRT of $H_0 : p = p_0$ versus $H_1 : p \neq p_0$.

4. Let $X_1, \ldots, X_n$ be iid $\mathrm{N}(\theta, \sigma^2)$ where σ^2 is known. For each of the following hypotheses, write out the acceptance region of a level α test, and the $1-\alpha$ confidence interval that results from inverting the test.

   (a) $H_0 : \theta = \theta_0$ v.s. $H_1 : \theta \neq \theta_0$.

   (b) $H_0 : \theta \geq \theta_0$ v.s. $H_1 : \theta < \theta_0$.

   (c) $H_0 : \theta \leq \theta_0$ v.s. $H_1 : \theta > \theta_0$.

5. Suppose that $X_1, \ldots, X_{30}$ is a random sample from $\mathrm{N}(\mu, \sigma^2)$ where both parameters are unknown. If the observed values of $\bar{X}$ and S^2 are respectively $\bar{x} = 12.9$ and $s^2 = 4.6$, calculate a 99% confidence interval for μ and σ.

6. We saw previously (Example 6.3) that a 95% confidence interval for the normal variance is rather wide when $n = 20$–the ratio of the upper to lower limits was $10.24/2.78 = 3.7$. How large a value of n would be needed in order for the interval $[L(X), U(X)]$ to be short enough that $(U-L)/L \leq 0.2$? I.e., roughly, short enough that the interval puts bounds $\pm 10\%$ on the point estimate S^2.

7. Suppose that $Y_1, \ldots, Y_6$ are iid from the gamma distribution whose pdf is

$$f(y|\alpha, \beta) \propto \frac{1}{\beta^\alpha} y^{\alpha-1} e^{-y/\beta} \qquad y, \alpha, \beta > 0,$$

   with known shape $\alpha = 6$ and unknown scale parameter β. Find a pivot based on $Y_1, \ldots, Y_6$. If the observed value of $\sum_{i=1}^6 Y_i$ is $\sum_{i=1}^6 y_i = 6.6$, calculate a 99% confidence interval for $\mathrm{E}(Y_i)$. *Hint: Using the properties of the gamma distribution would really help out here.*

8. (a) If $X_1, \ldots, X_n$ are iid exponential random variable with mean μ, show that $Y = \min(X_1, \ldots, X_n)$ is also exponentially distributed.

   (b) Compare the lengths of 95% confidence intervals for μ based on the two different pivots $\bar{X}/\mu$ and Y/μ when $n = 3$ and $n = 10$.

9. Suppose that $Y_1, \ldots, Y_{12}$ are monthly counts of insurance claims, assumed independently Poisson distributed with (monthly) mean μ.

   (a) Show that the annual count $T = Y_1 + \cdots + Y_{12}$ is sufficient for μ.

   (b) If the observed value of T is $t = 96$, calculate two approximate 99% confidence intervals for μ based on

   i. the MLE.
   ii. the likelihood ratio.

10. Let $X_1, \ldots, X_n$ be distinct observations. Let $X_1^*, \ldots, X_n^*$ denote a bootstrap sample, and let $\bar{X}_n^* = \frac{1}{n}\sum_{i=1}^n X_i^*$. Find

    (a) $\mathrm{E}(\bar{X}_n^* \mid X_1, \ldots, X_n)$.
    (b) $\mathrm{Var}(\bar{X}_n^* \mid X_1, \ldots, X_n)$.
    (c) $\mathrm{Var}(\bar{X}_n^*)$.

11. Let $X_1, \ldots, X_n$ be a random sample from Bern(p) where $p \in (0, 1)$ is unknown. Let $\theta = p^2$.

(a) Show that the MLE $\hat{\theta}$ for the parameter θ is biased.

(b) Outline a bootstrap procedure for estimating the bias of $\hat{\theta}$.

Hand-in questions

1. Find a $1-\alpha$ confidence interval for θ, given $X_1, \ldots, X_n$ iid with pdf $f(x|\theta) = 1$ for $x \in (\theta - 1/2, \theta + 1/2)$. **[3 marks]**

2. (a) Let $X_1, \ldots, X_n$ be a sample from Bern(p), and $Y_1, \ldots, Y_m$ a sample from Bern(q). The two samples are independent of each other. **[3 marks]**

   (b) Find an approximate 95% confidence interval for $p - q$ when both n and m are large. **[2 marks]**

   (c) 100 people are given a standard antibiotic to treat an infection and another 100 are given a new antibiotic. In the first group, 90 people recover; in the second group, 85 people recover. Let p be the probability of recovery with the standard antibiotic and q the probability of recovery with the new antibiotic. Provide an 80% confidence interval for the difference $\theta = p - q$. **[3 marks]**

3. Let $X_1, \ldots, X_n$ be a sample from N(μ, 1). Let $\theta = e^{\mu}$, and we estimate θ by $\hat{\theta} = e^{\bar{X}}$, where $\bar{X} = n^{-1}\sum_{i=1}^n X_i$. Describe how to obtain a bootstrap estimator $v_{boot} = \widehat{\mathrm{Var}}(\hat{\theta})$ for $\mathrm{Var}(\hat{\theta})$. **[3 marks]**

Appendix A

Exam tips