SM-4331 Advanced Statistics

Dr Haziq Jamil

2021-11-18

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.

Contents

Welcome to SM-4331! This is a Level 4 Major Option module weighing 4 MCs. Students typically take this module in their fourth year (final or penultimate semester). It is highly recommended for students having a strong interest in statistics and probability, especially students whose final year project involves a statistical component. SM-2205 Intermediate Statistics is a pre-requisite for this module, for which all students should have taken. SM-4331 Advanced Statistics complements SM-4337 Applied Statistics and SM-4339 Survival Analysis very nicely, so it will be beneficial to take these modules together.

This class is all about deepening your understanding about statistical inference. There will be an emphasis on the mathematical aspects and theorem proving of important statistical concepts, and less on practical applications (this is left for SM-4337). Thus, it is a class about asking the question 'why?', rather than 'how?'. However, most concepts introduced will be accompanied with `R` code for students to explore in their own time. By the end of this module, hopefully, students will appreciate and be able to understand why things work the way they do in the statistical world.

Past student feedback on this module is that it is on the difficult side. Well, it is *advanced* statistics, after all. Taking past feedback into account, I have redesigned this module to make it so that students get to follow the content better. For more details, see the Class Philosophy section below.

Goals

We will first revisit in further detail the fundamental building blocks of mathematical statistics, beginning with set theory, probability theory and probability distributions. We will also learn about convergences of random variables in order to understand several very important results in statistics (e.g. the law of large numbers and the central limit theorem). The syllabus then focusses on the three important statistical inference activities: point estimation, interval estimation, and hypothesis testing. To cap things off, we will tackle linear regression, arguably the most important statistical tool at a practitioner's disposal, from a mathematical aspect.

Incidental learning outcomes

Besides the core content of mathematical statistics that we will cover in this module, I really hope my students will be able to realise the following incidental learning outcomes:

1. To be able to present your arguments in a logical and cohesive manner, at the same time gain confidence in public speaking.

2. To train you to **read more**. I simply cannot emphasise enough how important reading is to your intellectual development. It will undoubtedly also improve your grammar and expand your vocabulary. Need some inspirational quotes? Here are several:

    - *A reader lives a thousand lives before he dies. The man who never reads lives only one.*
    - *Reading is the gateway skill that makes all other learning possible.*
    - *She read books as one would breathe air, to fill up and live.*
    - *A book is a device to light the imagination.*

3. To connect statistical theory with applied computation via exploration of `R` code. If you've never learnt a programming language before, now is a good time to start. Sometimes the code is given so you can just copy and paste into an `R` terminal and see what happens for yourself!

4. To be self-independent in your studies. I understand that attending weekly, physical classes gives you a form of structure, makes you comfortable. Having video lectures that you can view in your own time is the opposite of this. Moreover, there are no notes to be had. This is an opportunity to instil self-discipline in yourself if you haven't had any to begin with, or to strengthen it if you've had some! If you stick to the schedule, complete all the tasks, you will do well.

Module information

Class format

Blended learning (mixture of face-to-face and online teaching). As per current university guidelines, I am instructed to conduct at least 40% of the course through online learning. Therefore, the format is as follows:

- **Video lectures**. Each topic will be presented via videos released incrementally topic by topic. You may view these at your leisure, but you should aim to complete the viewings before the next set of videos are released. You are encouraged to take notes. There will not be any face-to-face or Zoom lectures.

- **Tutorials**. We will have a 2-hour tutorial class (face-to-face) every other week. You will volunteer or be called upon to present the answers to the exercises. Another format we might do is "breakout sessions" where you split into small groups and tutor each other.

- **Breather/Recap sessions**. We will dedicate a 2-hour face-to-face class every other week to recap the completed part. During these sessions, you will have the opportunity to clarify any concept or idea that you are still unsure of. No new materials will be taught, it is only for answering questions. You may not ask about the "starred" questions in the tutorials directly.

Two weekly sessions are timetabled: *Tuesdays 1410-1600* and *Wednesdays 1150-1340*. Tuesday sessions will be for tutorials and Wednesdays for breather/recap sessions. Note that since we are doing video lectures, we will only see each other every other week. Classes will be in FOS 2.18, unless otherwise told.

Assessment

Formative assessment

- 1 × mock exam in Week 14
- Exercise sheets

Summative assessment

- **60% examination**: Closed-book, real-life and invigilated with "typical" worked out solutions type questions. The scheduled date for the exam is *Tuesday, 4 May 2021, 2.00-4.00pm* at Chancellor Hall (more details closer to the date). Answer 4 out of 5 questions. Calculators are allowed. A minimal formulae sheet will be provided, as well as statistical tables.

- **20% topical tests**: Open-book, multiple choice format, online submission through Canvas. There will be seven tests in total, corresponding to each topic/part of the module. In total, there are 100 equally weighted questions (0.2 points each) so each test will have 10-15 questions depending on the topic. Each test will be available as soon as a new topic is taught, and will be available until the next topic is taught. You may take the test at any time in between. You will have 12 hours to complete the test once you start it. **NO RETAKES**.

- **10% tutorials**: There are 7 planned tutorial sessions, corresponding to 7 exercise sheets. The majority of the questions are for practice (formative assessment), but there will 1-3 "starred" questions which you will have to hand in for grading. The total marks for all of these star questions will be 100, which then counts for 10% of the overall summative mark.

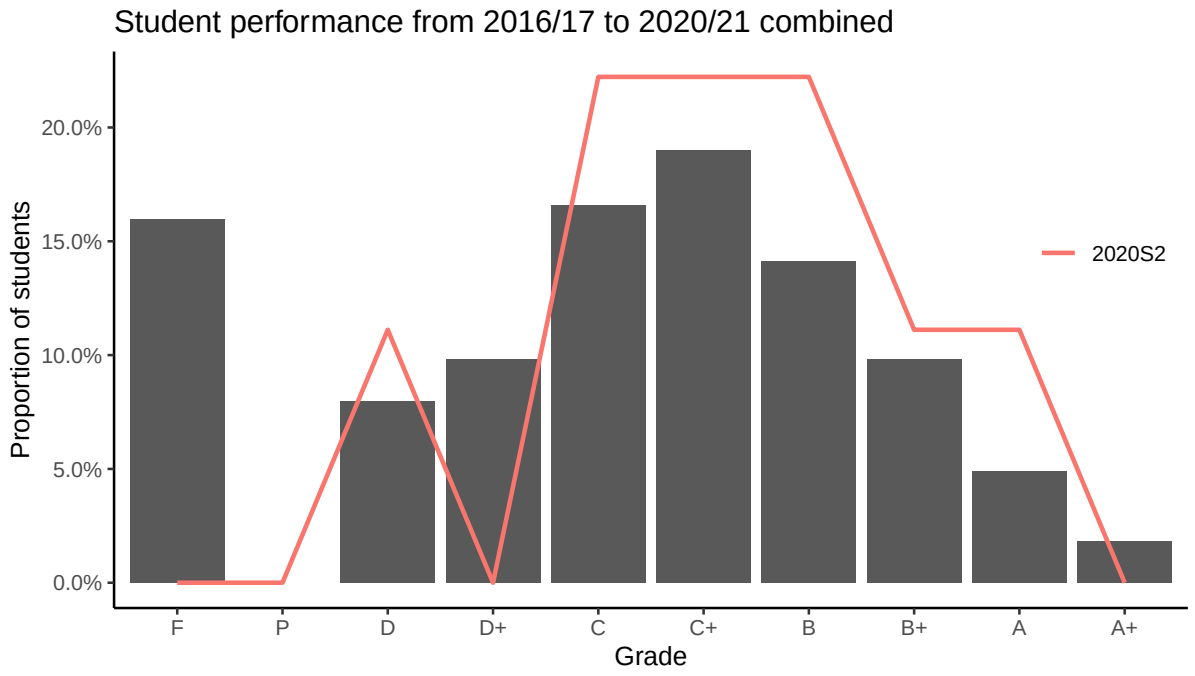
- **10% participation**: I will give you marks based on your participation during tutorial sessions, breather/recap sessions, as well as online Canvas discussion groups. The marks will reflect your participation level for the entire semester. The rubric is found in the table below.

- **(BONUS) 5% notes**: If you take notes for this course, you may submit them to me for grading. I will grade based on aesthetics (tidiness, organisation, readability) as well as content (did you grasp the key concepts? did you do all the little "green checks" from the lectures?). Check the schedule for submission dates.

Marks	1	2	3	4	5
Attendance and promptness	Late to classes every time, or poor attendance (<20%).	Is late half of the time. Attendance is irregular (20-60%)	Has been late not more than twice. Attendance is regular (60-80%)	Frequently attends on time and attendance is regular (80-90%)	Always on time and consistently attends classes (90-100%).
Contribution and engagement	Impedes the learning of others. Questions or comments are often distracting from learning. Group work often disrupted.	Rarely asks questions or offers ideas in class. Seldom contributes to group work.	Offers ideas and asks questions on occasion which help to clarify discussion for self. Good group work skills.	Offers ideas and asks questions in class which help to clarify discussion for all. Very good group work skills.	Consistently offers ideas and asks questions that clarify and extend discussions for all. Shows superior leadership
Preparedness	Poor & unfinished questions. Not prepared at all for class.	Partially or barely adequate completion of questions.	Generally completes questions with care.	Completes questions with thoroughness.	Produces a high quality level of answers.
Participation effort	Very little effort. Unwilling to be called upon to present.	Inconsistent effort. Rarely volunteers to present.	Good effort. Volunteers to present.	Makes a very good, consistent effort. Almost always is the one to present answers.	Works to the best of their ability. Extremely dependable when called upon to present.

Key data

- Past class sizes: 2016S2 = 57, 2017S2 = 70, 2018S2 = 18, 2019S2 = 9, 2020S2 = 9 (avg: 32.6)
- SFE grade average: 3.7 / 5.0 (74.0 %)



Course policy

Here I detail several policies pertaining to my course that I have in place.

Communication policy

I am usually quick to respond to student e-mails, despite receiving about a gazillion of them each day. Perhaps it is partly due to my dislike of leaving things unattended and obsession to get things moving (this is my problem, not yours). In an effort to preserve my mental health, perhaps have a think first whether that e-mail is necessary. Here are some guidelines.

- If you are unable to come to tutorials or recap/breather class, and you're afraid of losing participation marks, then you are responsible for giving me a **note in hard copy** that documents the reason for the missed class. An e-mail is unnecessary unless the impromptu absence involved missing an exam.
- If you are emailing for an extension (star questions) or a remake (topical tests), the answer is always "no". The deadlines are flexible already as it is, so schedule your time accordingly.
- If you missed a class and wanted to know what was discussed, I sincerely think you're missing the point. The classes are meant for you to ask the questions and for me to see your level of understanding. By not being there you miss out on both counts.
- If you have a question about the syllabus, perhaps read this syllabus.
- If you have a genuine question about the course contents, sure you could e-mail me. But why not post it up as a Canvas discussion and earn points?

I do not give out my WhatsApp number to my students. I think there should be clear boundaries between students and instructors, and this, in my opinion crosses that boundary. I have e-mail on my phone (for better or worse) so it really makes no difference if you WhatsApp or if you e-mail. E-mail communication is official, which means that you are accountable for what you say to me, and more importantly for what I say to you. Note that Canvas message is equal to e-mail.

Attendance policy

Students

I understand that sometimes unexpected things come about that delays us or prevents us altogether from class. Really, I do–I have spent 9 years in formal tertiary training learning about random events. However, I also believe that *those who want to will make time, and those who don't want to will make excuses.* So unless there is an excusable reason for your absence (or tardiness) then your participation marks will be affected.

Teacher

Ah, the life of an early-career academic. I often get asked what exactly does a professor do? Well, we split our time between research, teaching and admin work, often not in equal parts unfortunately. What this means for you is that sometimes there are "urgent" non-teaching matters that apparently require my attention, which may happen to clash with our schedule. I don't expect this to happen very often (in the last 4 semesters I have only had to cancel/reschedule classes due to these reasons: marshall duties during convocation, UBD open day, Wawasan 2035 meeting, research travel). But if they do, I promise that

- I will firstly do my best to keep our scheduled time together. Quite frankly, I'd rather talk about statistics with you than anything else.
- Failing that, I will keep you informed well ahead of time.
- I will reschedule the missed hours at a suitable time for everyone's benefit.

Conduct

- The medium of instruction is english. This means that I will officially lecture in english, converse in english, and write in english. I expect you to do the same. This is often misconstrued as shying away from our mother tongue (malay), but I assure you my intentions are noble. The majority of students will converse in anything other than proper english outside my class, so for the little time we spend together, at the very least you can get to practice your language skills.

- I intend to provide a safe and conducive environment for learning for my students, especially in tutorial classes. This means that you will not be ridiculed for making mistakes, or not knowing things, or forgetting things, etc. The purpose is to genuinely gauge your developmental level, and the only way I can do that is if you try. We will also not rudely interject or diminish each other's ideas, or worse yet, each other's character. Please be respectful.

- I expect each one of you to do the tests by yourself. This means no collaborative work, although it's fine to "Google" stuff if you need to or consult a dictionary for some tricky words. You may think, well it is open-book after all, so why can't we discuss among ourselve the solutions? Here's a secret: The tests are not meant to test your knowledge or ability, they are there to *force* you to refer back to your notes and reading materials :) Together with the note-writing process, these tests will reinforce the mind-map you have created in your brain, allowing you better recall, so you'll do better in the exam, which you are sitting for by yourself. Trust the process, trust me, and don't cheat.

Learning Management System

I'm a believer in LM systems like Canvas. Therefore this is the main avenue for me to distribute learning materials (slides, exercise sheets, solutions, etc.) as well as post announcements. Canvas has an app for students that you may wish to download onto your smartphones (you can get notifications too). Alternatively, keep checking Canvas on your computers on a regular basis.

Canvas will also be the avenue for submissions. COVID-19 has made me realise that I much prefer marking on my iPad. If you can upload a PDF copy of your work (I'm referring to the "starred" tutorial questions) that would be really great. Otherwise, you may drop a hard copy of your work into my pigeonhole.

Resources

This course

- My lecture slides
- This book

Textbooks

I estimate that you will probably only absorb 40% of the material through my lectures alone. Please supplement your understanding by reading the texts. These books are written by professors and course instructors who have uncountably more experience than I have, and are more able to explain the statistical concepts much better than I ever could. I used these books myself during my undergraduate years so I trust they will beneficial for you as they were for me!

-
-

Miscellaneous

- A Twitter thread on education resources
- Taking Good Notes
- Fun with Attendance and Grades (i.e. Students Should Attend Class)

Really recommended:

- All About that Bayes: Probability, Statistics, and the Quest to Quantify Uncertainty. Talk by Dr. Kristin Lennox on YouTube.
- https://www.qualitydigest.com/inside/standards-column/secret-foundation-statistical-inference-120115.html

Part I

Introduction

What is statistics?

Statistics is a scientific subject 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?

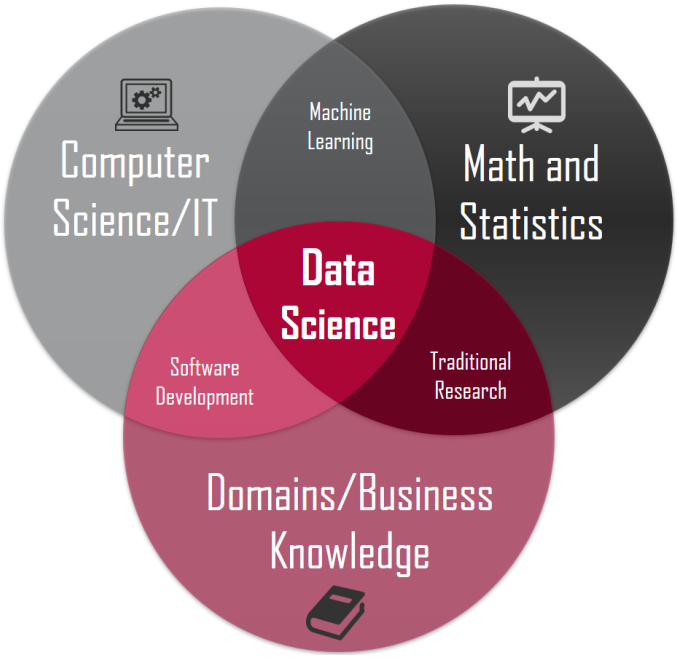


Figure 1: Data science.

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.

> *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)*

Population, sample and parametric models

Two practical situations

Figure 2: Data science.

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.

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?"

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.

Figure 3: Data science.

2. For AITI, while possible, it is (economically) infeasible to measure the entire population, i.e. to ask everyone in Brunei whether or not they use the internet on a daily basis.

The *population* is an 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.

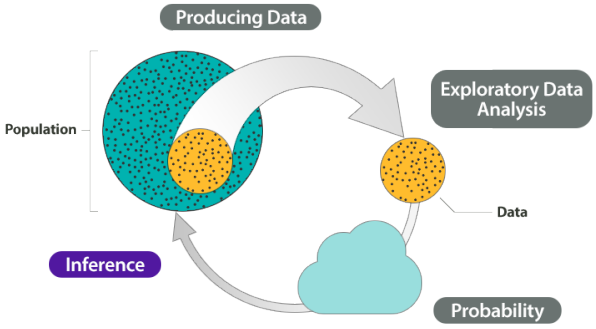


Figure 4: Data science.

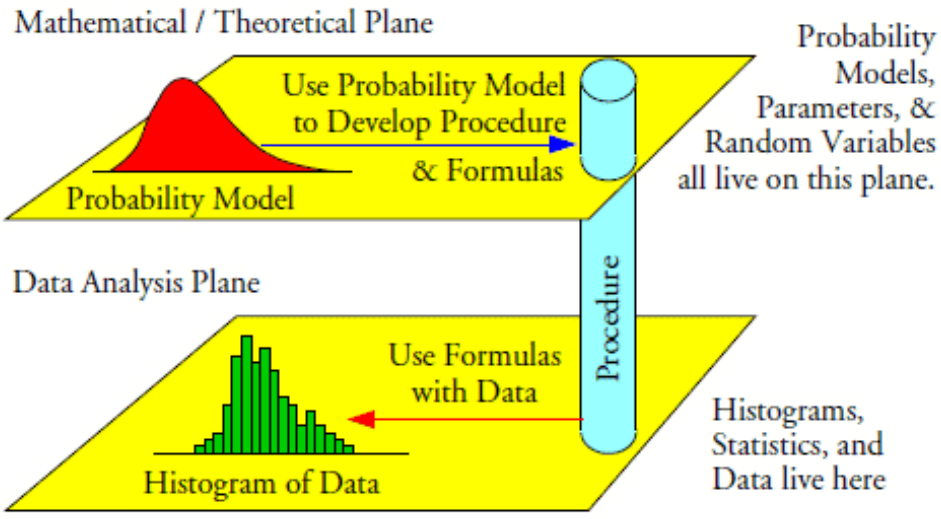
Parametric models

For a given problem, we typically assume a population to follow a *probability distribution* with pdf/pmf $f(x|\theta)$.

- 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 represented by θ (such as the mean, variance, rate, etc.). 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.
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.



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

A sample of size n, $\{X_1, \dots, X_n\}$, is also called a *random* sample. It consists of n concrete numbers in a practical problem. The word 'random' captures the characteristic of the sample (of the same size) being different.

- The sample may be taken by different people or entities.
- The sample may be obtained under different circumstances.
- etc. Essentially, they would be different subsets of a population.

Furthermore, a sample is also viewed as n independent and identically distributed (iid) random variables, when we assess the performance of a statistical method.

Variability of estimates

- For the BMW M example, suppose a sample of $n = 38$ used gave the sample mean

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

- A different sample may well give a different sample mean. Example: 29 YouTubers and other social media "influencers" were given access to the new M3 on a race track, and their sample mean yielded $\bar{X}_n = 3.4$.
- Is the sample mean $\bar{X}_n$ a good estimator for the unknown 'true' acceleration μ? Obviously, we cannot use the concrete number 3.9 to assess how good this estimator is, as a different sample may give a different average value.

By treating $X_1, \dots, X_n$ as random variables, $\bar{X}_n$ is also a random variable, so has a distribution. If the distribution of $\bar{X}_n$ concentrates closely around the unknown μ, then it is a good estimator!

- 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,

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

- 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

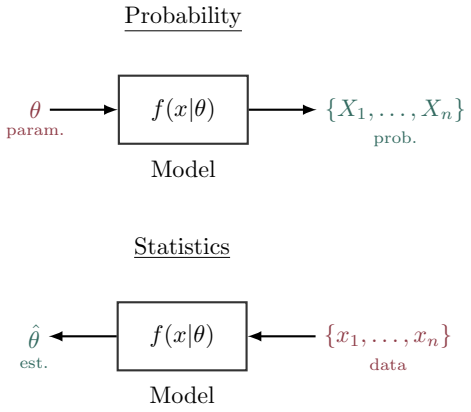
We've just implicitly described the three main activities concerning statistical inference.

1. **Point estimation**
   "What is μ?"
2. **Hypothesis testing**
   "Is $p = 0.95$ and not $p = 0.97$?"

3. **Interval estimation**

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

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.

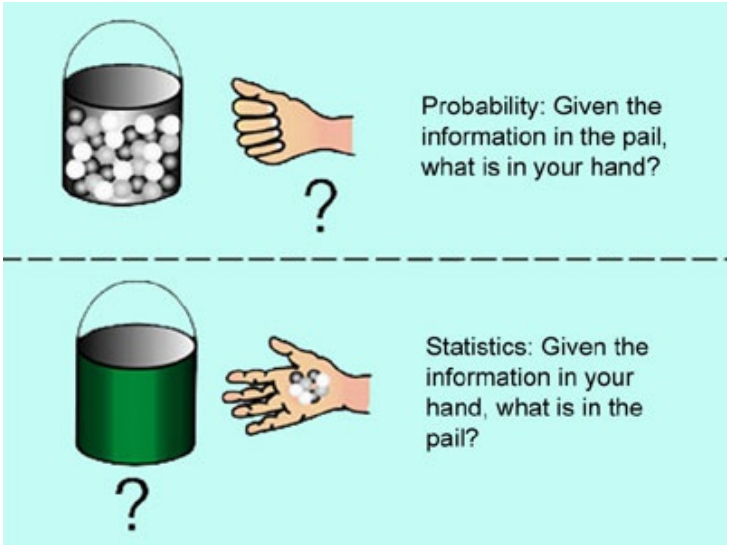


In probability, we ask questions like

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

Whereas in statistics, we are interested in questions like

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



Part II

Prepare

Chapter 1

Probability theory primer

Learning objectives

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

- do this
- do that
- and this

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

1.1.1 Sample space

In conducting an "experiment"…

- 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.

Example 1.1. In tossing a coin $n \geq 2$ times, let H denote 'heads', while T denote 'tails'. Let $\omega = (\omega_1, \ldots, \omega_n)$. Then

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

Let E be the event that the first head appears on the second toss. Then

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

1.1.2 Set operations

- The **complement** of an event A, written A^c, is the set of all elements that are not in A: $A^c = \{\omega | \omega \notin A\}$.
- 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** (see C&B Thm 1.14).

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,

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

1.1.3 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.

1.2 Axiomatic probability

1.2.1 Probability as a measure

We understand probability to mean the "frequency of an event occurring". If we can assign probabilities to (random) events in an experiment, then we can start to analyse them statistically.

We take an *axiomatic approach* to defining probabilities, rooted in *measure theory* (due to Kolmogorov, 1933). Let $(\Omega, \mathcal{B}, \mathbb{P})$ be a measure space.

- $\mathcal{B}$ is a σ-algebra on Ω, the subsets of Ω that are feasible for measuring.
- $\mathbb{P}$ is a measure on $(\Omega, \mathcal{B})$, i.e. the method that is used for measuring.

Figure 1.1: Andrey Nikolaevich Kolmogorov. 25 April 1903 – 20 October 1987.

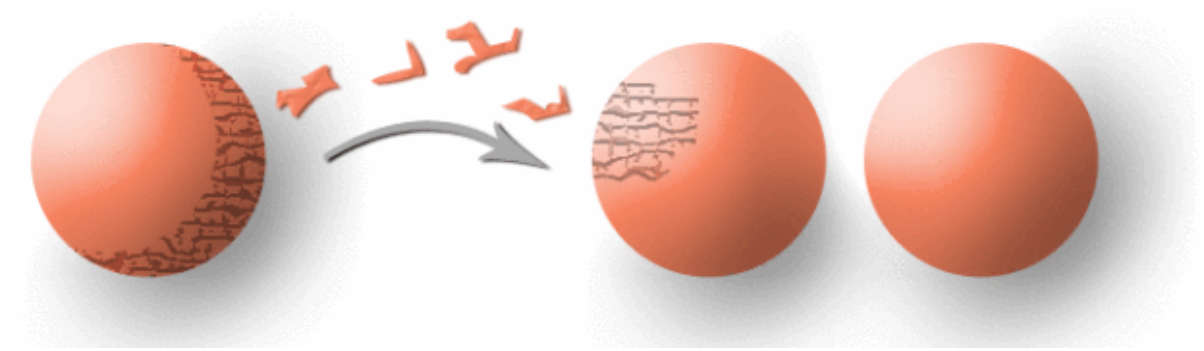
Remark. See Defn 1.2.1 in C&B and the following examples, as well as §1.9 in Wasserman.

Remark. There are alternative formulations/approaches to defining probabilities, e.g. Cox's Theorem (logical probabilities).

[Further explanation of measure theory]

Why do we bother with measure theory?

> 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.



If we don't lay out the foundations for measuring probabilities rigorously, we can end up with nonsensical answers!

1.2.2 Axioms of probability

$\mathbb{P} : \mathcal{B} \to [0, 1]$ is a **probability measure** on $(\Omega, \mathcal{B})$ if it satisfies the the following three axioms:

- Axiom 1: $\mathbb{P}(A) \geq 0, \forall A \in \Omega$.
- Axiom 2: $\mathbb{P}(\Omega) = 1$.
- Axiom 3: For pairwise disjoint events $A_1, A_2, ...,$

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

Remark. There are two main interpretation of probabilities.

1. The **frequentist** interpretation is that 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** interpretation is that the probability measures an observer's strength of belief that the event is true.

In either interpretation, the three axioms must be satisfied.

1.2.3 Derived probability results

Let A and B be measurable events from the sample space Ω. The following results can be derived using only the three axioms:

- $\mathbb{P}(\{\}) = 0$
- $0 \leq \mathbb{P}(A) \leq 1$
- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
- If $A \subseteq B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$
- $\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A \cap C_i)$ for any partition $C_1, C_2, ...$ of Ω (*Law of Total Probability*)

1.3 Conditional probabilities

1.3.1 Conditional probability

Update the sample space based on new information, and thus update probability calculations.

Definition 1.1 (Conditional probabilities)**.** If A and B are events in Ω, and $\mathbb{P}(B) > 0$, then the *conditional probability* of A given B is

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

- The given information is now the "new" sample space: $\mathbb{P}(B|B) = 1$. *All further occurrences are calibrated with respect to their relation to* B. Think of $\mathbb{P}(A|B)$ as *the fraction of times* A *occurs among those in which* B *occurs.*
- For mutually exclusive events A and B, $\mathbb{P}(A|B) = \mathbb{P}(B|A) = 0$.
- In general, $\mathbb{P}(A|B) \neq \mathbb{P}(A)$ and $\mathbb{P}(A|B) \neq \mathbb{P}(B|A)$.

Example 1.4. A medical test for a disease D has outcomes '+' and '−'. The probabilities are:

	D	D^c
+	0.009	0.099
−	0.001	0.891

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 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.

1.3.2 Bayes Theorem

Theorem 1.1 (Bayes' Rule)**.** *Let* $A_1, A_2, ...$ *be a partition of the sample space, and let* B *be any set. Then, for each* $i = 1, 2, ...,$

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

This is easily proven using definitions of conditional probabilities, as well as the law of total probability.

Remark. Some will call $\mathbb{P}(A_i)$ the **prior probability**, and the $\mathbb{P}(A_i|B)$ **posterior probability**.

1.4 Independent events

1.4.1 Independence

In some cases, the occurrence of a particular event B has *no effect* on the probability of another event A:

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

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

Definition 1.2. Two events A and B are *statistically independent* if and only if

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

If A and B are independent then so too are

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

Figure 1.2: (Probably not) Rev. Thomas Bayes c. 1701 – 7 April 1761

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 $n(A \cap B) = 3$. Do this 1,000 times, and count events automatically.

```
x <- sample(1:6, size = 1000, replace = TRUE)
head(x, 100)
```

```
##   [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 $1/3$.

1.5 Random variables

Ask (randomly) 50 people whether they like ("1") or dislike ("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!

```
2 ^ {50}
```

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

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

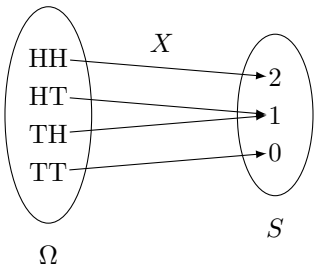
What if we defined $Y = \sum_{i=1}^{50} X_i$?

- Y is the count of the number of people who like learning statistics from this sample of 50.
- The minimum value for Y is 0, and the maximum is 50. So the new sample space for Y is $S = \{0, 1, 2, \ldots, 50\}$. Much easier to deal with!

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

Remark. Random variables (r.v.) are conventionally denoted with uppercase letters, and the realised values of the variable will be denoted by the corresponding lowercase letters. Thus, the r.v. X can take the value x.

Example 1.5. 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 r.v. is illustrated as follows:



We can see that a r.v. X is a *mapping*[1] $X : \Omega \to \mathbb{R}$ that assigns a real number $X(\omega)$ to each outcome ω. Then,

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

Example 1.6. The r.v. 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

1.6 Distribution functions

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

Definition 1.3. The cumulative distribution function (cdf) of a r.v. X, is the function $F_X : \mathbb{R} \to [0, 1]$ defined by

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

Example 1.7. From Example 1.6, 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] Technically, a measurable function. See Wasserman (2004, Appendix 2.13).

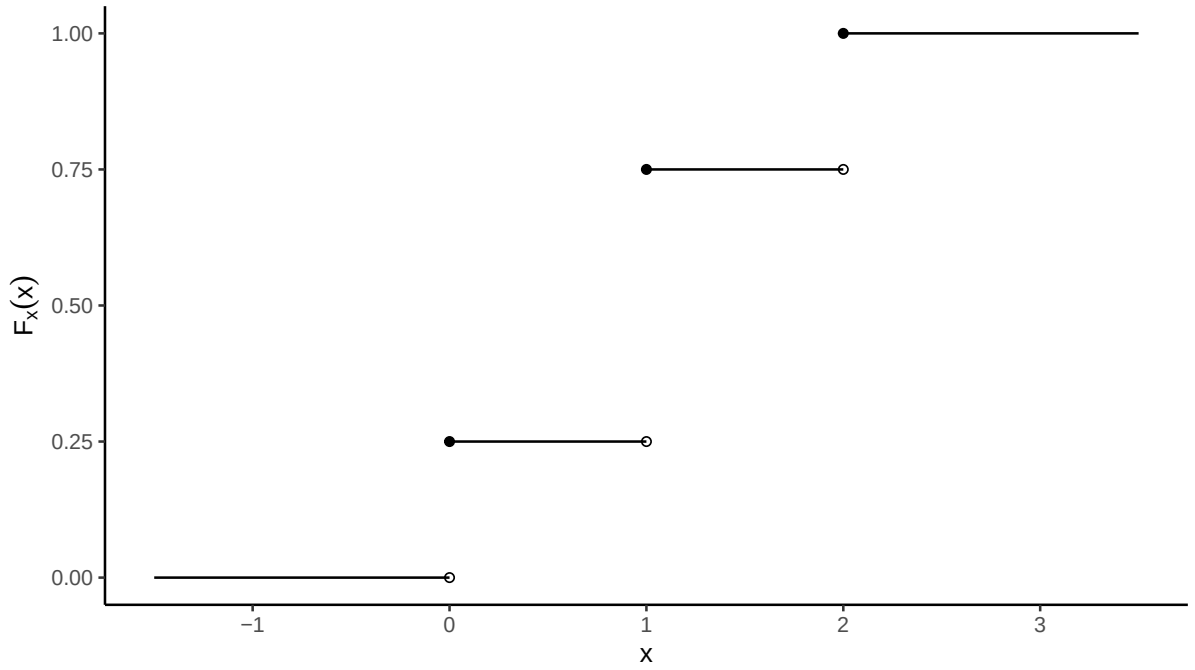


Figure 1.3: Distribution function of the random variable X

1.6.1 Properties of cdfs

- $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to+\infty} F(x) = 1$.
- $F(x)$ is non-decreasing: $x_1 < x_2 \Rightarrow F(x_1) \leq F(x_2)$.
  - 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 step function".

In fact, any function satisfying the above properties is a cdf. For proofs of these facts, see the reference textbooks.

Clearly, F itself *can be discontinuous* (we saw this in the previous 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.

1.6.2 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. X and Y are said to be **identically distributed**.

Remark. 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.

1.7 Probability functions

1.7.1 Probability mass function

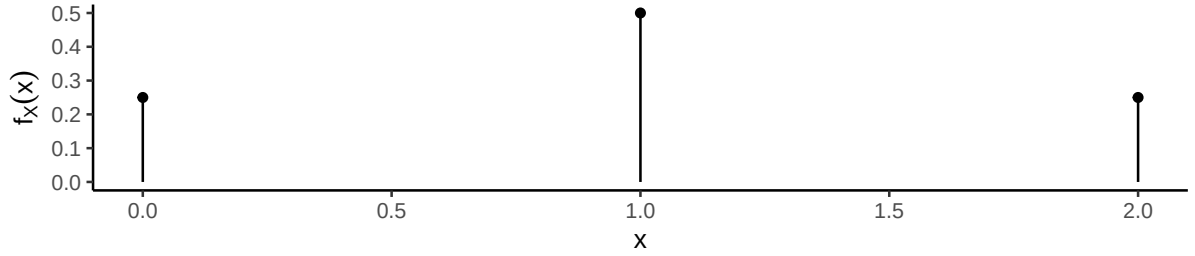
Associated with a r.v. X and its cdf F_X is another function, called either the probability density function (pdf) if it is continuous, or the probability mass function (pmf) if it is discrete.

Definition 1.4 (Probability mass function)**.** A discrete r.v. X only take countably many values $\mathcal{X} = \{x_1, x_2, \dots\}$. Its probability mass function (pmf) is

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

Example 1.8. The pmf from Example 1.6 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 r.v.s are countable, we can add up probabilities over all the points in the event. For any a, b both in $\mathcal{X}$ such that $a \leq b$, we have that

$$\mathbb{P}(a \leq X \leq b) = \sum_{x=a}^{b} f_X(x).$$

As a special case we get

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

Consequently,

- Each $f_x(x) \geq 0$ for all x; and
- $\sum_x f_x(x) = 1$.

1.7.2 Probability density functions

We want to translate the same idea of "point probabilities" over to the continuous case, but must be more careful here. Let X be a continuous r.v. (i.e., its cdf is continuous). The analogous procedure would be to consider

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

and using the Fundamental Theorem of Calculus, we have that

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

We can see that the cdf is like "adding up" the "point probabilities" $f_X(x)$ to obtain interval probabilities.

Definition 1.5 (Probability density function)**.** A continuous r.v. X takes any numerical value in an interval or collection of intervals (having an uncountable range). Its probability density function (pdf) is the function $f_X(x)$ that satisfies

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

Note that

- $f_X(x) \geq 0$ for all x;
- $\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = 1$; and
- $\mathbb{P}(X = x) = \int_{x}^{x} f(\tilde{x})\,\mathrm{d}\tilde{x} = 0$.

Don't think of $f(x)$ as probability functions–this only holds for discrete r.v.. Read Wasserman (Warning after Example 2.13 on p.24).

Example 1.9. 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

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

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

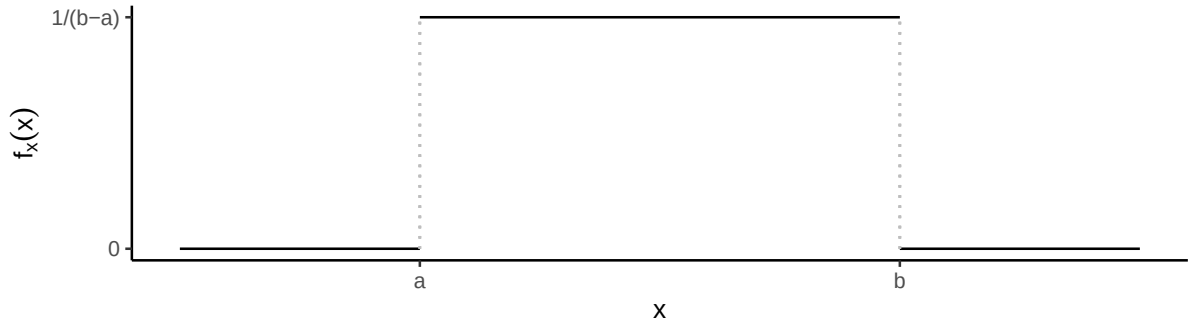


Figure 1.4: Plot of pdf

Continuous r.v. miscellanea

- On notation: 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 we write $X \sim f_X(x)$.
  - Or by their specially given name, e.g. $X \sim \text{Unif}(a, b)$.
  - If X and Y are identically distributed, then $X \sim Y$.
- Note that since $\mathbb{P}(X = 0)$ if X is continuous, then
  - $\mathbb{P}(X \leq b) = \mathbb{P}(X < b) + \cancelto{0}{\mathbb{P}(X = b)}$; and so

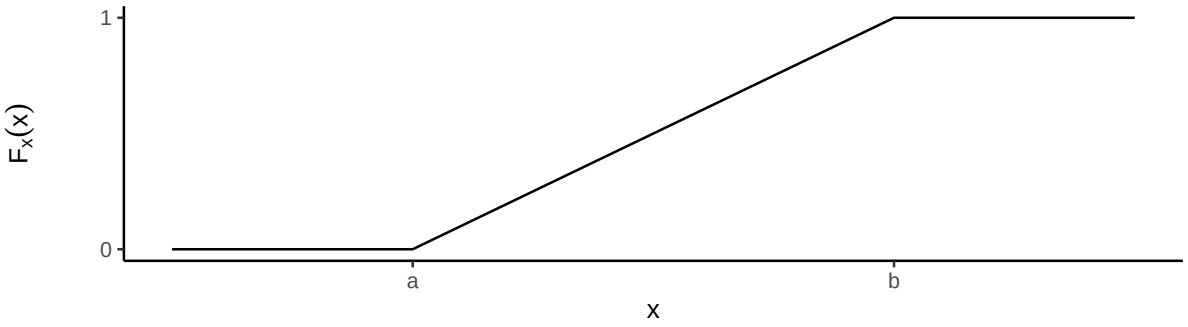
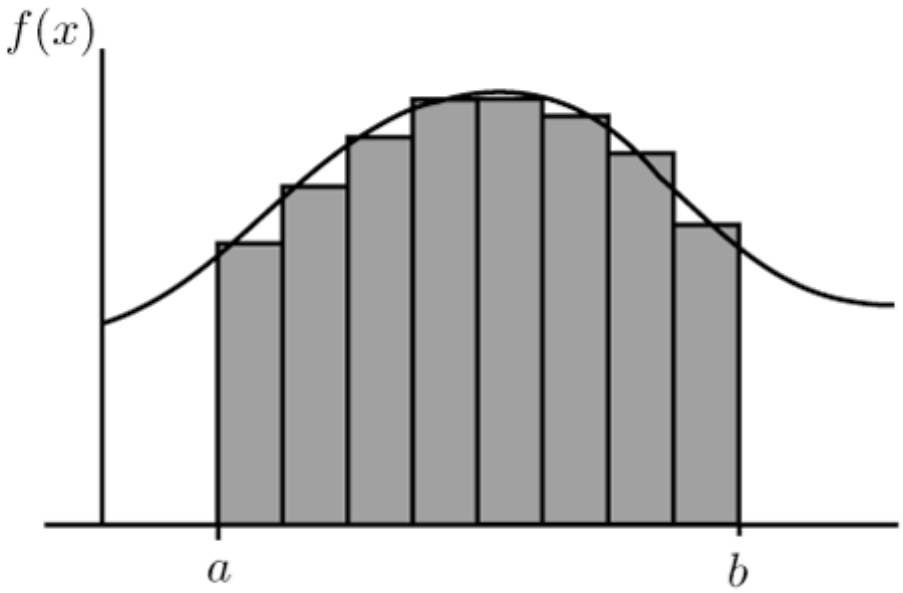


Figure 1.5: Plot of cdf

- $\mathbb{P}(a \leq X \leq b) = \mathbb{P}(a \leq X < b) = \mathbb{P}(a < X \leq b) = \mathbb{P}(a < X < b)$.

- Sometimes we just write $\int f(x)\,\mathrm{d}x$ to mean $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x$.
- The following are true:
  - $\mathbb{P}(a < X < b) = F(b) - F(a)$ (be careful, this is not true for discrete r.v.)
  - $\mathbb{P}(X > a) = 1 - F(a)$
- As a side note, mixed discrete and continuous distributed r.v. do exist, but we won't be covering them in this course.
- The Riemann integral is defined as the limit of the sum of the areas of these bars, as the number of bars gets larger and larger (and hence the width of the bars get smaller and smaller).



1.8 Multiple random variables

1.8.1 Bivariate distributions

Probability models may involve more than one random variable, known as *multivariate models*. Consider the simplest kind, dealing with only two r.v.s in each discrete and continuous case.

Definition 1.6 (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 1.7 (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 $A \subseteq \mathbb{R}^2$,

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

All the univariate properties carry over to the bivariate (and multivariate) case:

- $f_{X,Y}(x,y) \geq 0$ 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 $F_{X,Y}(x,y) = \mathbb{P}(X \leq x, Y \leq y)$

Example 1.10. A bivariate distribution for two discrete r.v. X and Y each taking values 0 or 1:

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

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

Example 1.11. Consider a uniform distribution on the unit square $[0,1] \times [0,1]$. It has pdf given by

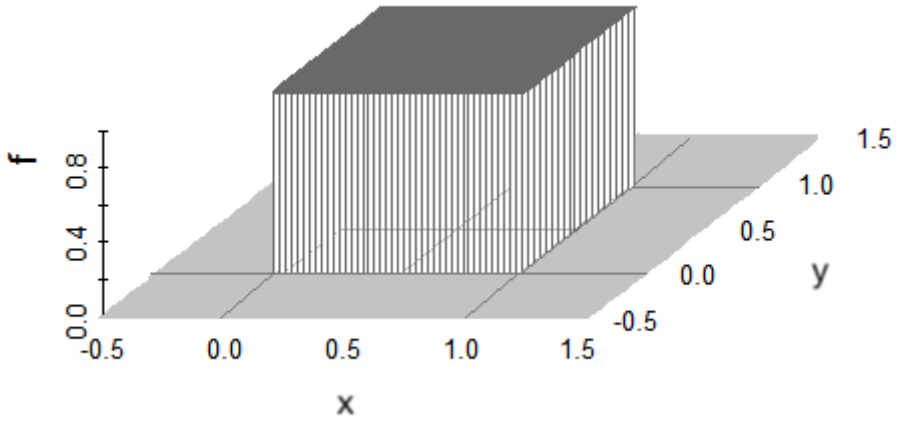
$$f(x,y) = \begin{cases} 1 & 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

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

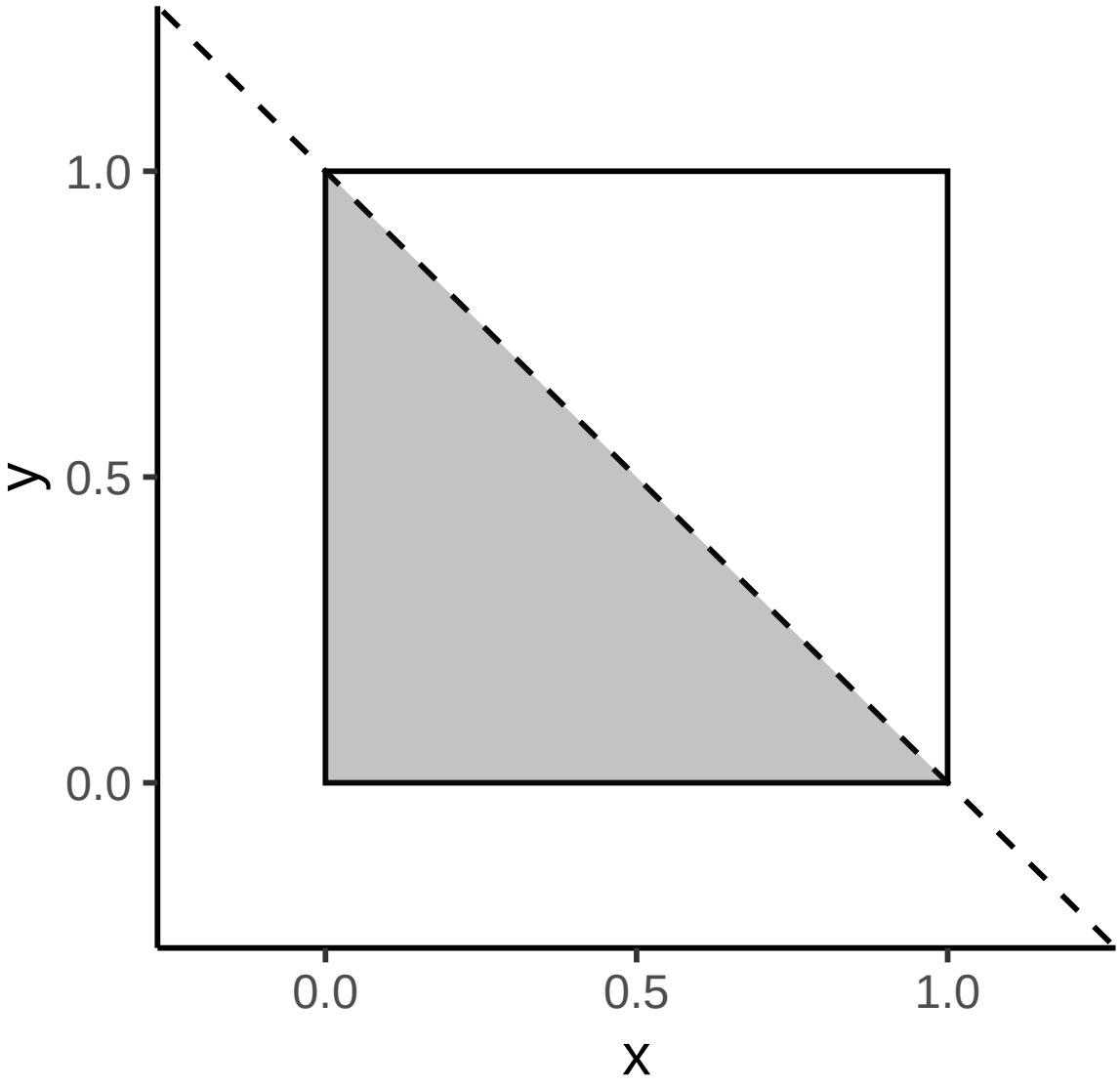
$$\begin{aligned} \mathbb{P}(X < 1/2, Y < 1/2) &= \int_0^{1/2} \int_0^{1/2} \mathrm{d}x\,\mathrm{d}y \\ &= \left[[xy]_0^{1/2} \right]_0^{1/2} = 1/4. \end{aligned}$$

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

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



This is the view from above. What is the volume of this wedge? It is the area of the shaded region multiplied by height 1.



1.8.2 Marginal distributions

We can recover the distribution for one of the r.v. in a bivariate (or multivariate) model by summing/integrating over the remaining probibility distribution.

Definition 1.8 (Marginal distribution)**.** For a bivariate r.v. (X, Y), the marginal distributions may be obtained as

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

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

See the textbooks for some examples.

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

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

the marginal cdfs can be obtained from the joint cdf for X as

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

and similarly $F_Y(y) = F_{X,Y}(\infty, y)$ for Y. Note that for continuous random variables, $\int_{-\infty}^{x} \left(\int f_{X,Y}(\tilde{x}, y)\,\mathrm{d}y \right) \mathrm{d}\tilde{x}$.

1.8.3 Conditional distributions

Oftentimes when two r.v. (X, Y) are observed, the values of the two variables are related. 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).

Knowledge about the value of Y gives us some information about the value of Y. This should sound familiar.

Definition 1.9 (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 1.10 (Conditional distributions, continuous)**.** If X and Y are discrete, the conditional of of X given $Y = y$ is

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

- As a function of x, $f_{X|Y}(x|y)$ is indeed a pdf, since

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

- The probability of X given $Y = y$ is computed as

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

- We can rearrange the equations to yield

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

Example 1.12. 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

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

1.8.4 Independent random variables

Previously we came across the concept of independence of probabilistic events. We can extend this notion to random variables.

Definition 1.11 (Independece of r.v.)**.** Two r.v. 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.
- The assumption of independence is used very often in statistical inference as it simplifies calculations quite a lot.

Example 1.13. Recall the bivariate distribution on the unit square (c.f. Example 1.11). 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 $X \sim \mathrm{Unif}(0,1)$, and independently sample $Y \sim \mathrm{Unif}(0,1)$.

1.9 Expectations

1.9.1 Expected values

The expected value, or expectation, of a random variable is merely its *average value weighted* according to the probability distribution. It signifies the **typical value** of an observation of a random variable.

Definition 1.12 (Expectation)**.** The expected value or mean of a r.v. of 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. Other symbols used are $\mathrm{E}\,X$ and $\mathrm{E}[X]$.
- The expectation is **not to be confused** with the sample mean of a set of observations $\{x_1, \dots, x_n\}$, i.e. $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$.

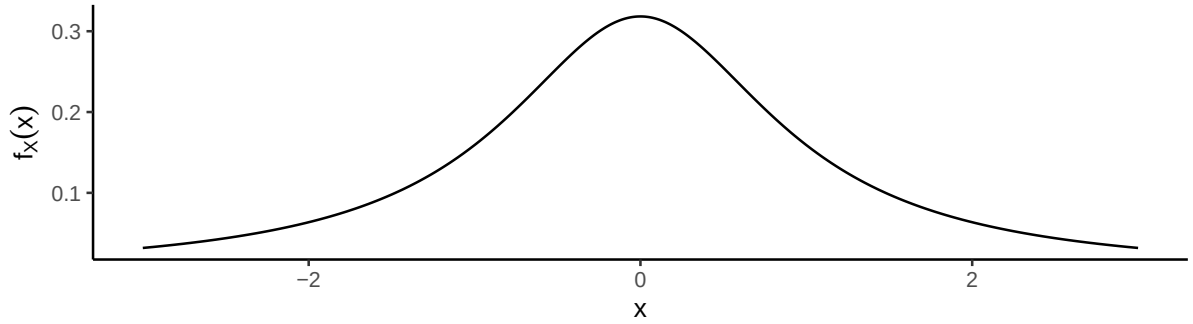
Example 1.14. Let $X \in \{0, 1\}$ take value 1 with probability p, and 0 with probability $1 - p$. X is a Bernoulli r.v., and we write $X \sim \mathrm{Bern}(p)$. Then,

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

Example 1.15. Let X be a cts. r.v. such that $X \sim \mathrm{Unif}(a, b)$. Then

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

Example 1.16.



Let X be a cts. r.v. with pdf $f(x) = \{\pi(1+x^2)\}^{-1}$ with support over $\mathbb{R}$. This is the Cauchy distribution[2] with location and scale parameter 0 and 1 respectively.

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

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

so the mean is undefined.

1.9.2 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[3]. Often time we will want to know the mean of $g(X)$.

Theorem 1.2. *Let X be a r.v. with pdf $f_X(x)$, and let $Y = g(X)$. Then*

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

[2]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.

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

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 $\mu := \mathrm{E}(X)$.

1.9.3 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 [you should really know these!]. They also work for $g(X)$ too.

- $\mathrm{E}(aX + bX + c) = a\,\mathrm{E}(X) + b\,\mathrm{E}(X) + c$ (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 $a \leq \mathrm{E}(X) \leq b$
- $\mathrm{E}(X) = \min_b \mathrm{E}((X-b)^2)$ (see Example 2.2.6 C&B)

As a corollary, 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,

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

1.9.4 Variance

Aside from the mean of a r.v., perhaps the most important moment is the second central moment, more commonly known as the variance.

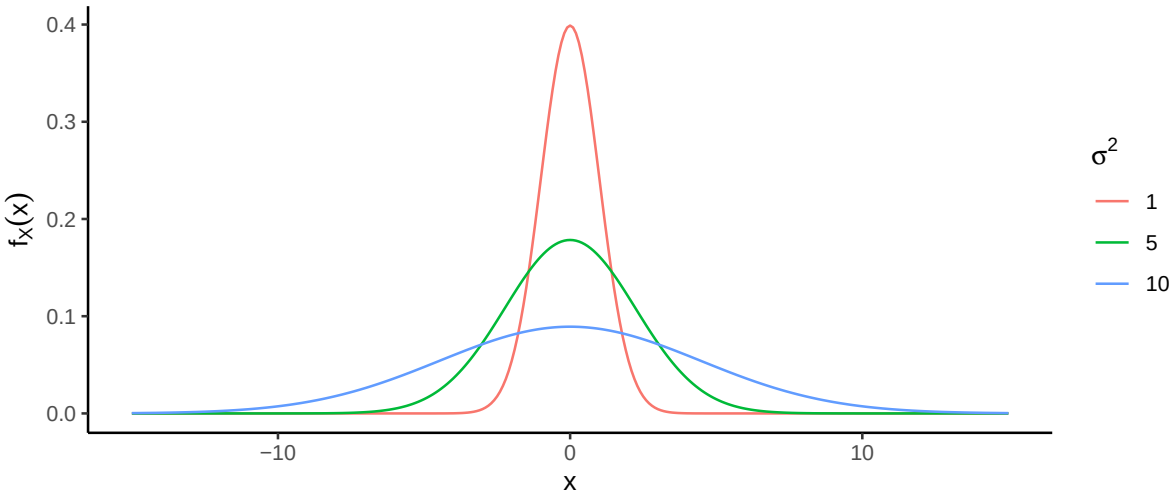
Definition 1.13 (Variance)**.** Let X be a r.v. with mean μ. The variance of X is defined

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

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 formula is $\sigma^2 = \mathrm{E}(X^2) - \{\mathrm{E}(X)\}^2$.
- This variance is **not to be confused** with the sample variance of a set of observations $\{x_1, \ldots, x_n\}$, i.e. $s^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}_n)^2$ (although, 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.



1.9.5 Covariance and correlation

The covariance and correlation between X and Y measure how strong the linear relationship is between X and Y.

Definition 1.14 (Covariance and 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 covariance between X and Y is

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

Their correlation is the number defined by

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

- An alternative formula is $\mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y)$.
- The covariance of X with itself is σ^2, while the correlation of X with itself is 1. Try and work this out yourself!

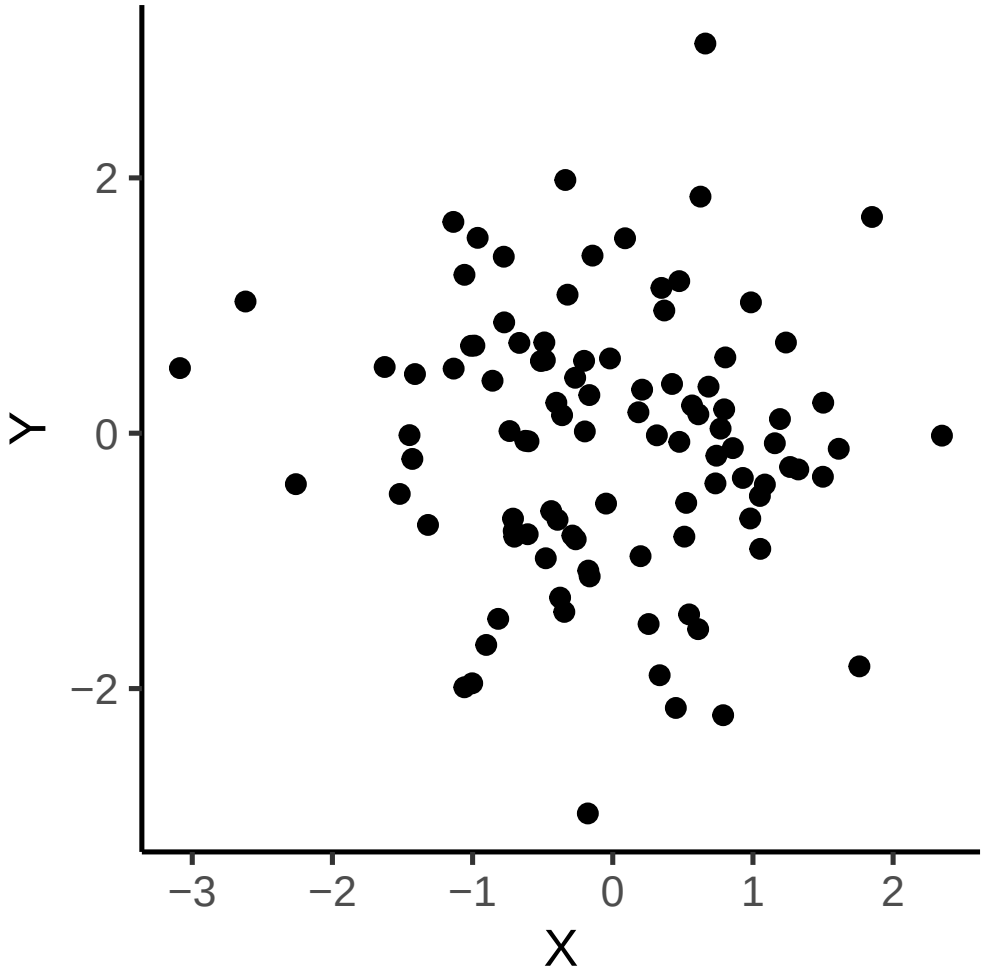
The magnitude of the covariance by itself does not reflect how strong the relationship between X and Y is, so this is where the correlation comes in.

- ρ_{XY} takes values between -1 and 1.
- $\rho_{XY} = 0$ 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, $|\rho_{XY} = 1|$ iff $\exists a \neq 0, b \in \mathbb{R}$ s.t. $\mathbb{P}(Y = aX + b) = 1$. If $a > 0$ then $\rho_{XY} = 1$, and if $a < 0$ then $\rho_{XY} = -1$.
- If X and Y are independent, then $\mathrm{Cov}(X,Y) = \rho_{XY} = 0$. Ttry and prove this!

Remark. If $\mathrm{Cov}(X,Y) = \rho_{XY} = 0$, then X and Y are **not necessarily** independent.

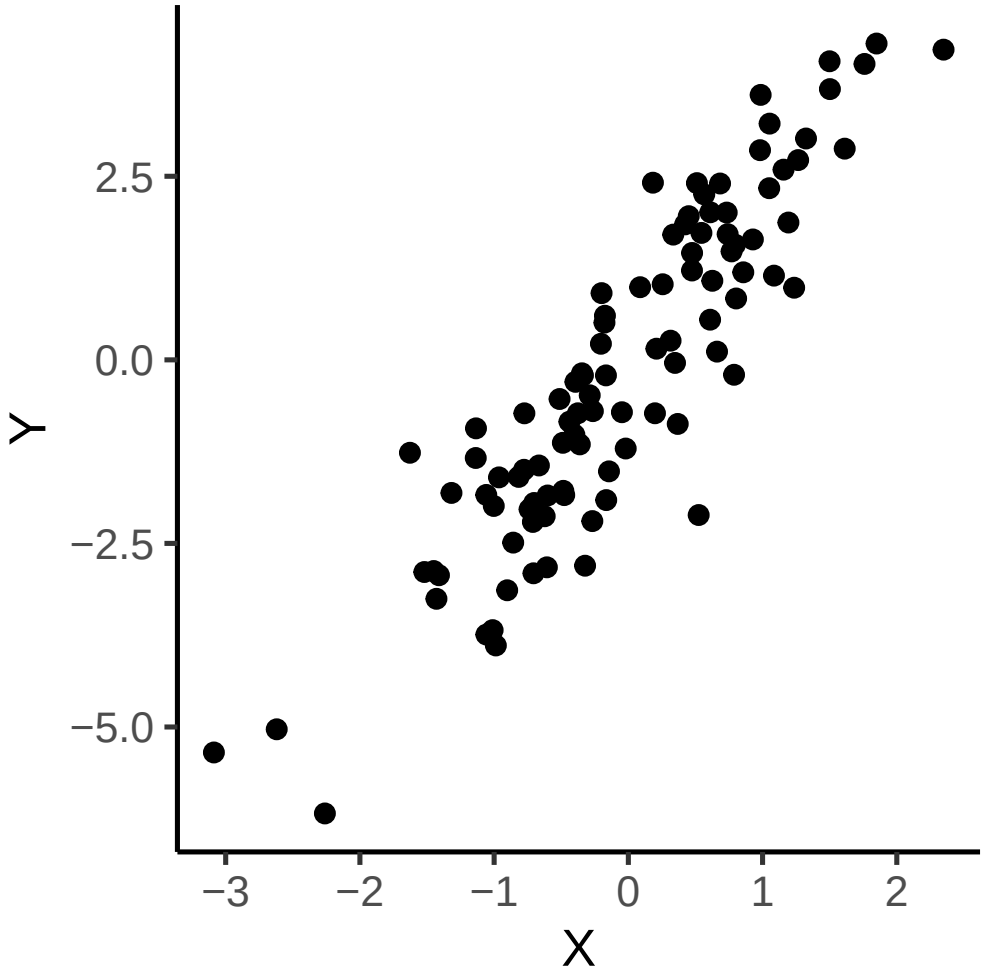
Let $X, Y \overset{\text{iid}}{\sim} \mathrm{N}(0,1)$. We can draw some random values in `R`, and produce a scatterplot to see the relationship between them.

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



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

```
Z <- rnorm(n = 100, mean = 0, sd = 1)
Y <- 2 * X + Z
qplot(X, Y, geom = "point")
```



1.9.6 Properties of variances and covariances

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 1.17. 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,

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

.

1.9.7 Variance-covariance matrix

Consider a random vector $(X_1, \dots, X_n)^\top$ whose mean is $(\mu_1, \dots, \mu_n)^\top$. The variance-covariance matrix, usually denoted $\Sigma \in \mathbb{R}^{n\times n}$, is defined to be

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

The correlation matrix is similar in structure to the above, except the off-diagonals are filled with $\rho_{X_i X_j}$ and the diagonals are all 1. Can you figure out why this is?

1.9.8 Conditional expectations

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 r.v. Y which might affect it.

Definition 1.15 (Conditional expectation)**.** The conditional expectation of a function of a r.v. X, $g(X)$ say, given a value of a nother r.v. $Y = y$, is

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

- All of the properties of the usual expectations are applicable.
- However, whereas $\mathrm{E}(X)$ is a number (non-random), $\mathrm{E}(X|Y = y)$ is a function of y. If we have not observed Y, then $\mathrm{E}(X|Y)$ is a random variable.

Example 1.18. Suppose we draw $Y \sim \mathrm{Unif}(0, 1)$. After we observe $Y = y \in [0, 1]$, we draw $X|(Y = y) \sim \mathrm{Unif}(y, 1)$. Intuitively, we expect that $\mathrm{E}(X|Y = y)$ to be half-way between y and 1, i.e. $(1 + y)/2$.

In fact, $f_{X|Y}(x|y) = (1 - y)^{-1}$, 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 r.v. whose value is $\mathrm{E}(X|Y = y) = (1 + y)/2$ once observed.

If $\mathrm{E}(X|Y)$ is a r.v., what is its mean?

Theorem 1.3 (Rule of iterated expectations/Law of total expectations)**.** *If X and Y are two r.v., 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 total average $\mathrm{E}(X)$ is the average $\mathrm{E}_Y(\cdot)$ of the case-by-case averages $\mathrm{E}(X|Y)$ over Y.

Proof.

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

□

1.9.9 Conditional variance

Definition 1.16 (Conditional variance)**.** The conditional variance of a r.v. X given $Y = y$ is

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

- An alternative formula:

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

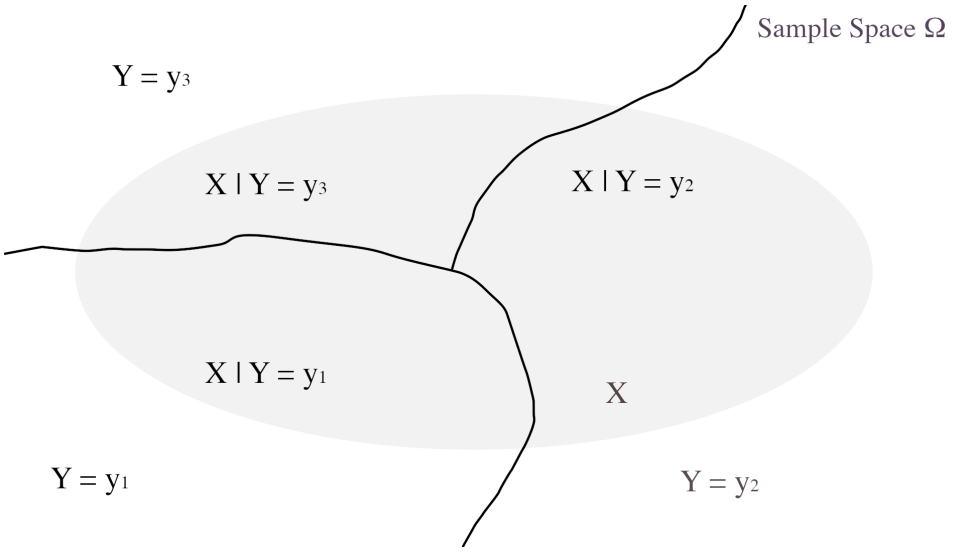
The law of total variance states that

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

Note that, in this context, both $\mathrm{Var}(X|Y)$ and $\mathrm{E}(X|Y)$ are random variables. The variance 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).

See also: https://math.stackexchange.com/a/3377007



1.10 Moment generating functions

1.10.1 Moment generating functions

As the name implies, the moment generating function (mgf) is used for finding *moments* of a r.v.. Other uses:

- Characterising a distribution
- Finding distributions of *sums of r.v.*
- As a tool in statistical proofs

Definition 1.17 (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".

1.10.2 Generating moments

Consider the following:

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

We can iterate the steps again to generate the k-th moment of X by taking k derivatives and setting $t = 0$. Note that this relies on being able to *interchange the order of differentiation and integration.* See §2.4 C&B. For "nice distributions" generally there are no problems.

Theorem 1.4. *If X has mgf $M_X(t)$, then*

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

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

Example 1.19. Let $X \sim \mathrm{Exp}(1/r)$ with $f_X(x) = re^{-rx}$ for $x \in [0, \infty)$. Then for $t < r$,

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

$\mathrm{E}(X) = M_X'(t)\big|_{t=0} = \frac{r}{(r-t)^2}\Big|_{t=0} = 1/r.$

1.10.3 Properties of mgf

- 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 r.v. 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.

1.11 Exercises

Chapter 2

Commonly-used probability models

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 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 Introduction

Distributions in statistics serve two main purposes:

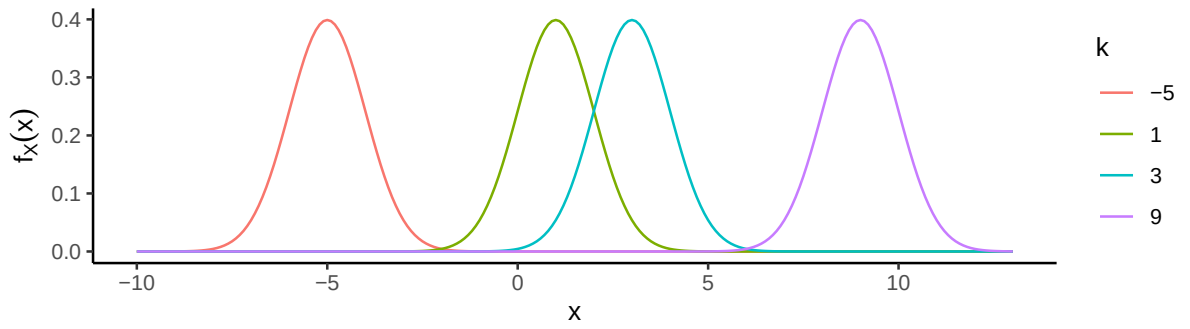
1. To describe the assumed behaviour of the observations made in an experiment, survey or other study;
2. To calibrate the values of derived statistics used in constructing confidence regions, hypothesis tests, etc.

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

In this Part 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 Part 3.

We deal with a *family* of distributions. This family is indexed by one or more *parameters* (c.f. parametric family), which allow us to vary certain characteristics of the distribution while staying with one functional form.

For example, consider r.v.s $X_k \sim \mathrm{N}(k, 1)$. These are distinct distributions yet have similar characteristics.



2.2 Discrete models

2.2.1 Point mass distribution

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

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

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

2.2.2 Uniform distribution

Let $k > 1$ be a given integer. The discrete uniform distribution on $\{1, \ldots, k\}$ has pmf

$$f(x) = \frac{1}{k}, \qquad x = 1, \ldots, k.$$

We write $X \sim \text{Unif}\{1, \ldots, k\}$.

- $\text{E}(X) = \frac{k+1}{2}$.
- $\text{Var}(X) = \frac{k^2-1}{12}$.
- If $k = 1$, then it is the point mass distribution.

The discrete uniform (and the point mass) is appeallingly simple but has relatively few "real" statistical applications.

2.2.3 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.

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 pmf can also be written $f(x) = p^x(1-p)^{1-x}$.
- The expectation is
$$\text{E}(X) = \sum_x xf(x) = 1 \cdot p + 0 \cdot (1-p) = p.$$
- The variance is
$$\text{Var}(X) = \sum_x (x-\mu)^2 f(x) = (1-p)^2 \cdot p + (0-p)^2 \cdot (1-p) = p(1-p).$$

Consider the pmf for the Bernoulli distribution above. What do you get when you plug in $x = 1$ and $x = 0$?

2.2.4 Binomial distribution

The binomial 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.

Other properties and results

- If $n = 1$ then X is a Bernoulli r.v.
- $\Pr(X = 0) = (1-p)^n$; $\Pr(X = 1) = p^n$
- Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bern}(p)$, then

$$X = \sum_{i=1}^{n} X_i \sim \text{Bin}(n, p)$$

One way to prove the above statement is by using mgfs.

From this we can more easily derive

$$\text{E}(X) = \text{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{E}(X_i) = np$$

and

$$\text{Var}(X) = \text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i) = np(1-p)$$

2.2.5 Geometric distribution

The geometric distribution is a type of 'waiting time' distribution. We count 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

$$\sum_{x=1}^{\infty} f(x|p) = \sum_{x=1}^{\infty} (1-p)^{x-1}p = \frac{p}{1-(1-p)} = 1.$$

- $\text{E}(X) = \frac{1}{p}$. The smaller the p, the longer we have to wait for a success.
- $\text{Var}(X) = \frac{1-p}{p^2}$.

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 $\{0, 1, 2, \ldots\}$. X and Y are related through $Y = X - 1$. Thus it is easy to check that

$$\text{E}(Y) = \frac{1-p}{p} \text{ and } \text{Var}(Y) = \frac{1-p}{p^2}.$$

We shall mainly use the first version of the geometric distribution in this course, but be aware of the alternative version as well.

2.2.6 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 $X \sim \text{NBin}(r, p)$. 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 $\{r, r+1, r+2, \ldots\}$.

- $\mathrm{E}(X) = \frac{r}{p}$.
- $\mathrm{Var}(X) = \frac{r(1-p)}{p^2}$.
- 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].

2.2.7 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 $X \sim \mathrm{Poi}(\lambda)$, 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.

$$\begin{aligned} M_X(t) &= \sum_{x=0}^{\infty} \frac{e^{tx} e^{-\lambda} \lambda^x}{x!} \\ &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!} \\ &= e^{-\lambda} e^{\lambda e^t} = \exp\{\lambda(e^t - 1)\}. \end{aligned}$$

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.

2.3 Continuous models

2.3.1 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.

[1] Details in C&B, p.95

- $\mathrm{E}(X) = \frac{a+b}{2}$.
- $\mathrm{Var}(X) = \frac{(a-b)^2}{12}$.

The plot of the pdf gives a "rectangular" shape, so probabilities can also be found geometrically, as we previously saw in Chapter 1.

2.3.2 Exponential distribution

The exponential distribution is often used to describe the distribution of measured time intervals 'duration data' or 'waiting-time data'. E.g.

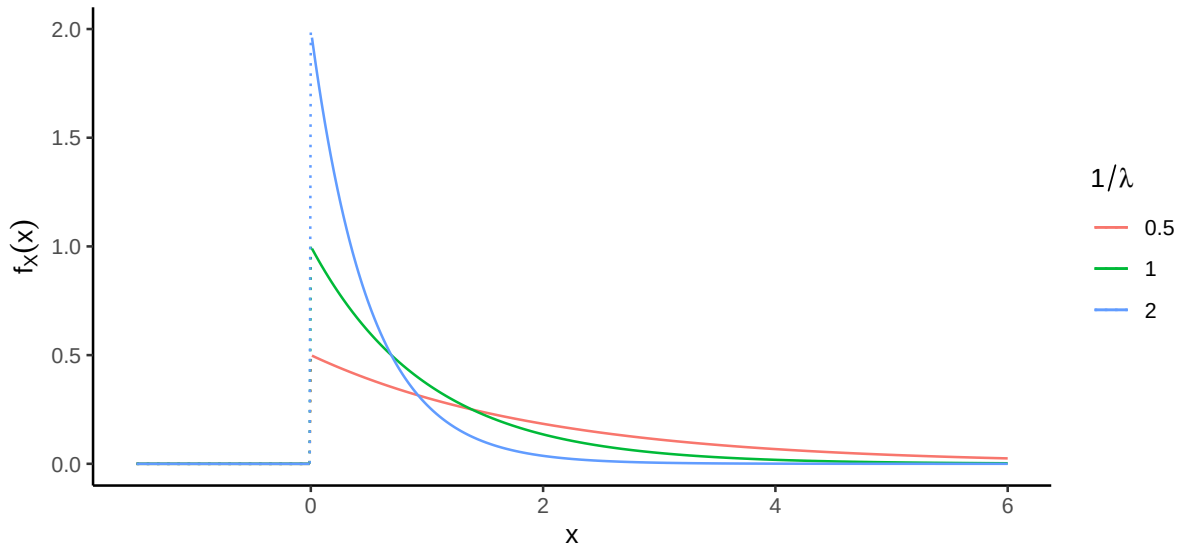
- the amount of time until an earthquake occurs.
- the time between two lightbulbs failing.
- the length (in minutes) of faculty staff meetings at UBD.
- the average waiting time at a hospital's A&E.

Let $X \sim \mathrm{Exp}(\lambda)$. The pdf is

$$f_X(x) = \frac{1}{\lambda} e^{-x/\lambda}.$$

- X has support over $[0, \infty]$.
- $\lambda > 0$ is known as the "scale" parameter. The value $1/\lambda$ is known as the "rate".
- $\mathrm{E}(X) = \lambda$.
- $\mathrm{Var}(X) = \lambda^2$.
- $aX \sim \mathrm{Exp}(a\lambda)$ for $a > 0$.

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.

2.3.3 Gamma distribution

The gamma distribution generalises the exponential. It is also used for modelling durations (lengths of time intervals). Let $X \sim \Gamma(\alpha, \beta)$. The pdf is

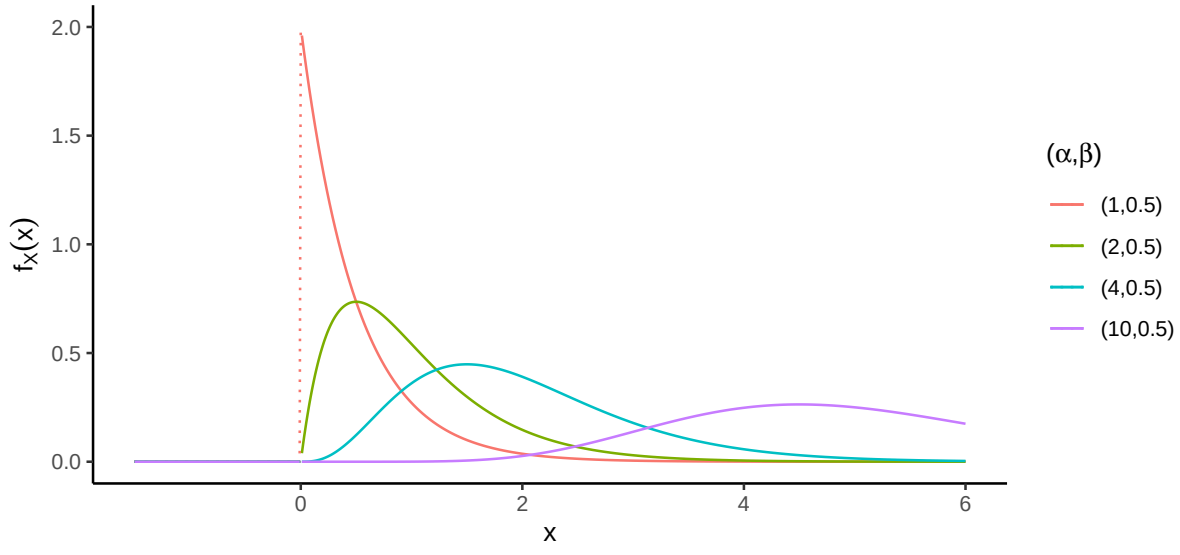
$$f_X(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$$

- X has support over $[0, \infty]$.
- $\alpha > 0$ is the "shape" parameter, and $\beta > 0$ is the "scale" parameter.
- $\mathrm{E}(X) = \alpha\beta$.
- $\mathrm{Var}(X) = \alpha\beta^2$.
- $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y}\, \mathrm{d}y$ is called the Gamma function.
- $\Gamma(1, \lambda) \equiv \mathrm{Exp}(\lambda)$.
- $aX \sim \Gamma(\alpha, a\beta)$ 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 $\beta = 1/s$.

Looks similar to the exponential pdf, but more generic. Effect of changing the shape parameter:



[2] https://perplex.city/memorylessness-at-the-bus-stop-f2c97c59e420?gi=3602158da66b

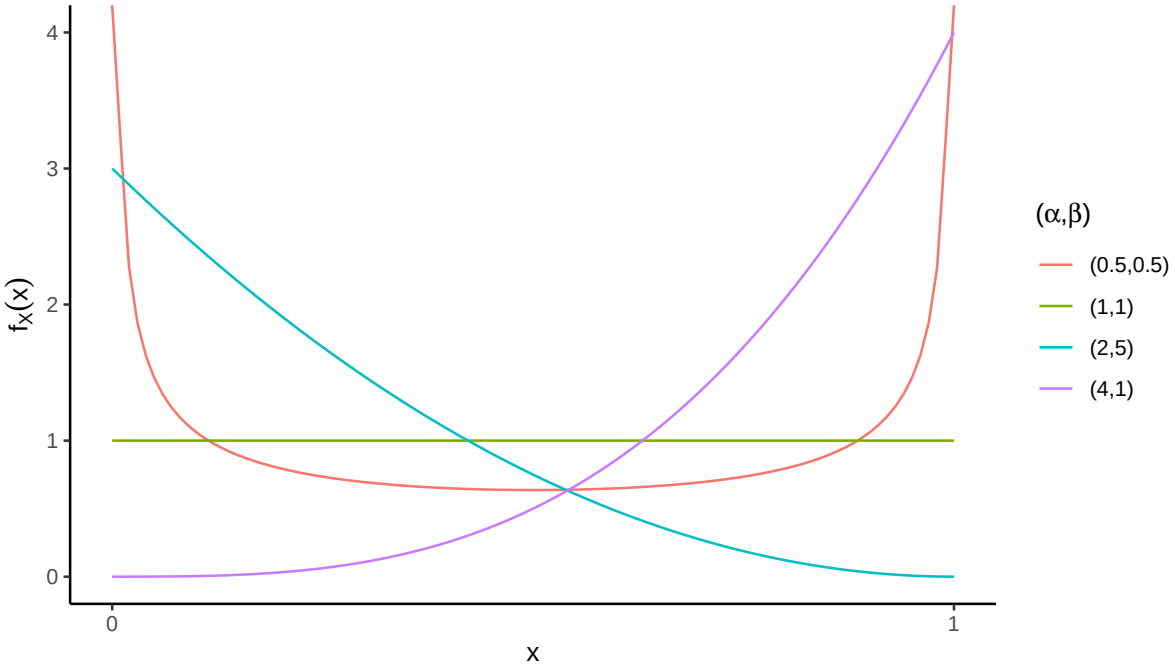
2.3.4 Beta distribution

The beta distributions are distributions on the unit interval $[0, 1]$, or on any other interval $[a, b]$ by transformation $X \mapsto aX + b$. It is used to model the behaviour of random variables limited to intervals of finite length in a wide variety of disciplines. Let $X \sim \text{Beta}(\alpha, \beta)$. The pdf is

$$f_X(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}.$$

- X has support over $[0, 1]$.
- $\alpha > 0$ and $\beta > 0$ are known as the "shape" parameters.
- $\text{E}(X) = \frac{\alpha}{\alpha+\beta}$.
- $\text{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.
- $B(\alpha, \beta)$ is the beta function $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.
- $\text{Beta}(1, 1) \equiv \text{Unif}(0, 1)$.

Pdf of beta distribution



2.4 Normal distribution

The normal distribution[3] is the most important distribution in statistics.

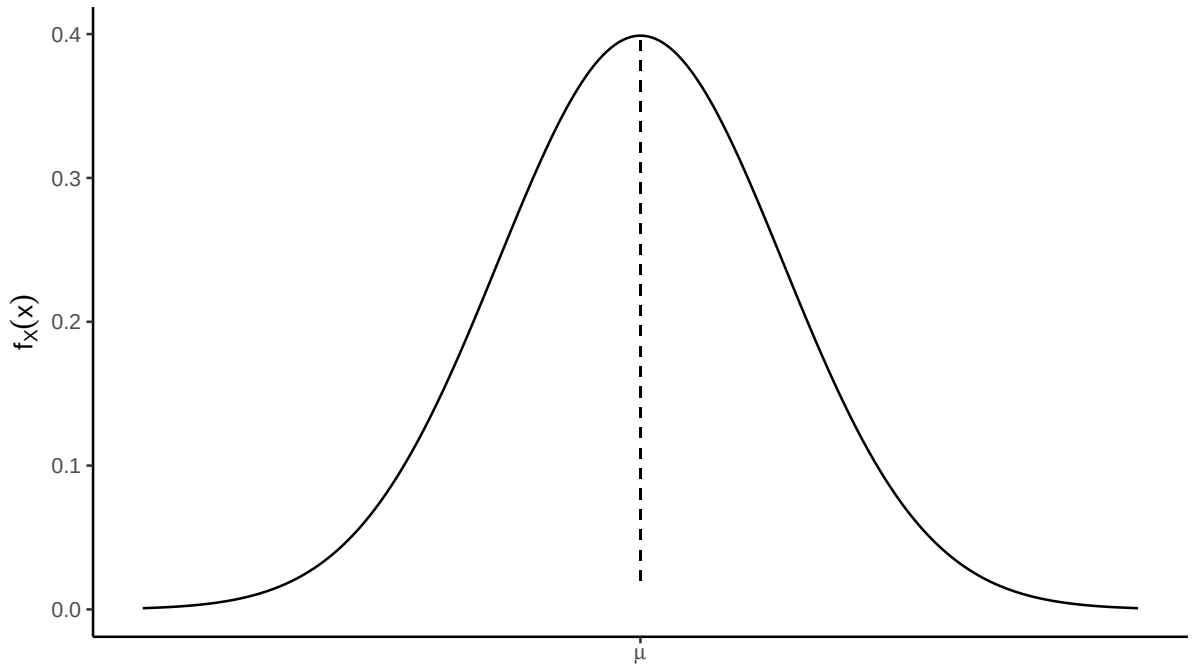
- Many naturally occurring phenomena can be modelled as following a normal distribution.
- The central limit theorem (CLT): The distribution of the mean of a sample tends to converge to a normal distribution, as more and more samples are collected.
- Often, the normal distribution is used for the error term in standard statistical models (e.g. linear regression).

[3]Here's a nice short exploration of the normal distribution: https://bookdown.org/cquirk/LetsExploreStatistics/lets-explore-the-normal-distribution.html

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

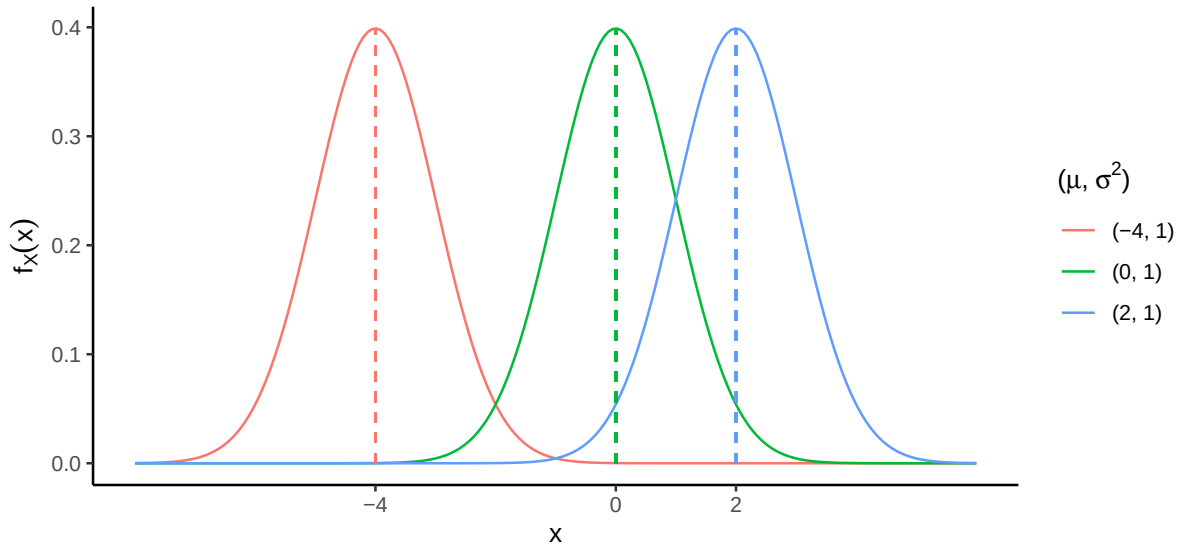
$$f_X(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

- X has support over $\mathbb{R}$.
- $\mathrm{E}(X) = \mu$.
- $\mathrm{Var}(X) = \sigma^2$.
- The normal distribution is **symmetric** about μ.
- The mode and median of X is also μ.



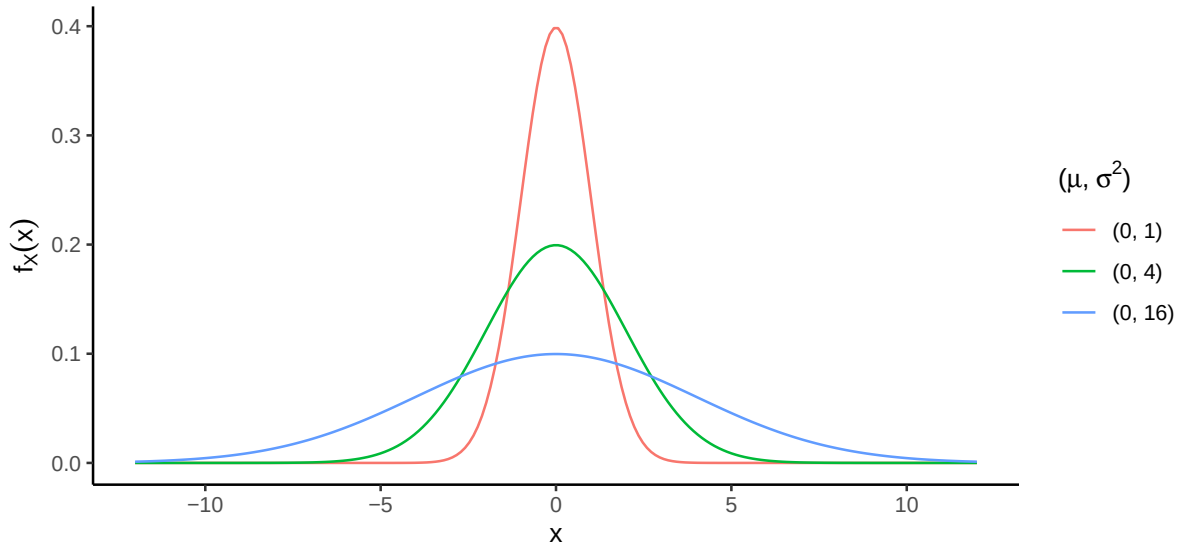
2.4.1 Location parameter

The μ parameter is also called the "location" parameter, since it determines where the bell curve is placed.



2.4.2 Scale parameter

The σ^2 parameter is also called the "scale" parameter, since it determines how spread out the curve is.



2.4.3 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.

- $\mathrm{E}(Y) = \mathrm{E}(cX + d) = c\mu + d$.
- $\mathrm{Var}(Y) = \mathrm{Var}(cX + d) = c^2\sigma^2$.

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** $Z \sim \mathrm{N}(0,1)$. It has pdf

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}},$$

specially denoted by the greek letter 'ϕ'.

Figure 2.1: Carl Friedrich Gauß. 30 April 1777 – 23 February 1855.

2.4.4 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, as the integrals above are *intractable* (no closed form solution). Download the statistical tables. Some results worth noting:

- $\Pr(Z \leq -a) = \Phi(-a) = 1 - \Phi(a)$
- $\Pr(a \leq Z \leq b) = \Phi(b) - \Phi(a)$
- $P(-a \leq Z \leq b) = \Phi(a) + \Phi(b) - 1$
- $P(-a \leq Z \leq -b) = \Phi(b) - \Phi(a)$
- $P(|Z| \leq a) = P(-a \leq Z \leq a) = 2\Phi(a) - 1$
- $P(|Z| \geq a) = P(\{Z < -a\} \cup \{Z > a\}) = 2(1 - \Phi(a))$

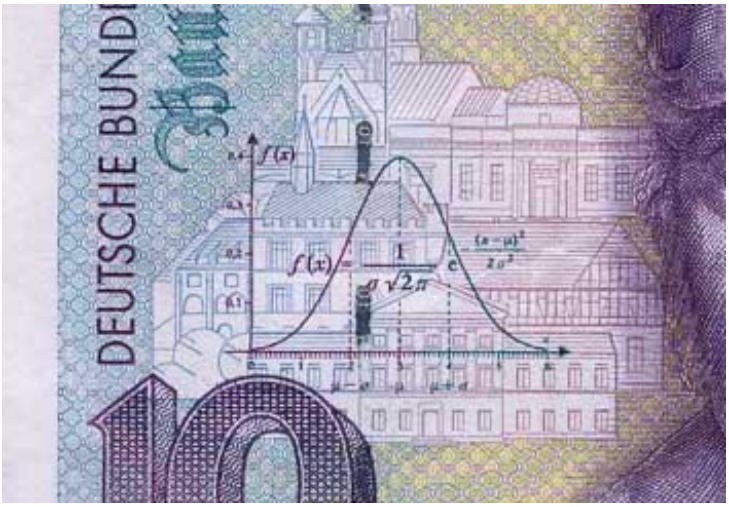


Figure 2.2: 10 Deustche Mark banknote.

We can use `R` to calculate probabilities:

```
pnorm(1.96, mean = 0, sd = 1)
```

```
## [1] 0.9750021
```

Some values of $\Phi(\cdot)$ worth remembering:

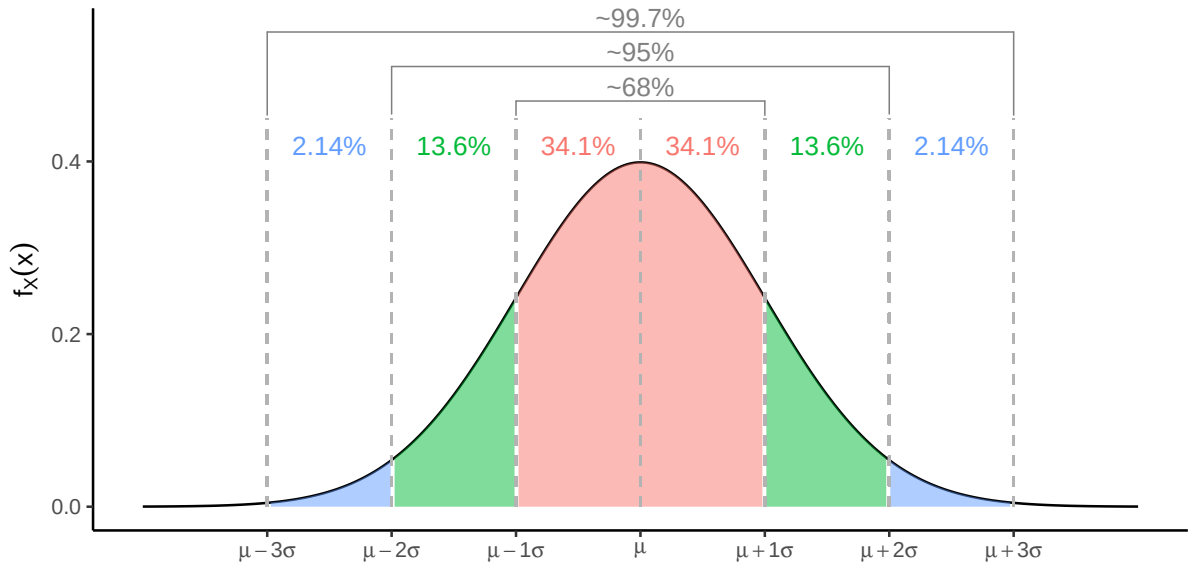
- $\Phi(0) = 0.5$
- $\Phi(1.64) \approx 0.95$
- $\Phi(1.96) \approx 0.975$

The last one, for example, says that

$$\begin{aligned}\Pr(\mu - 1.96\sigma \leq X \leq \mu + 1.96\sigma) &= \Pr\left(\left|\frac{X-\mu}{\sigma}\right| \leq 1.96\right) \\ &= 2\Phi(1.96) - 1 \approx 0.95\end{aligned}$$

2.4.5 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.



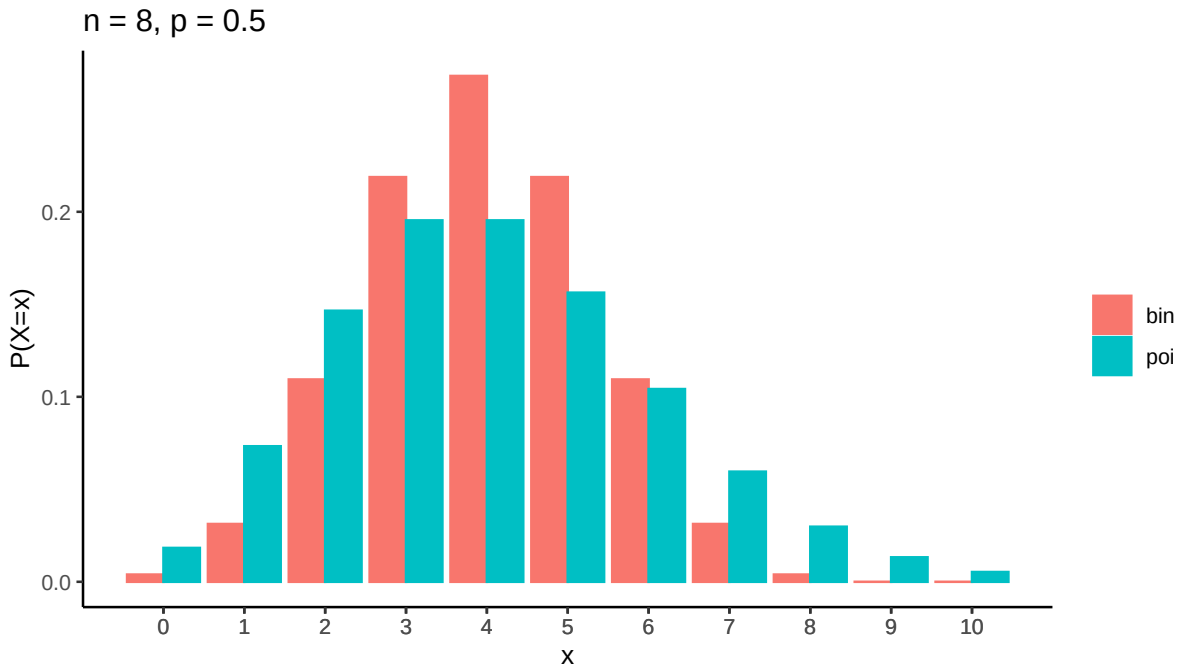
2.5 Some relationships

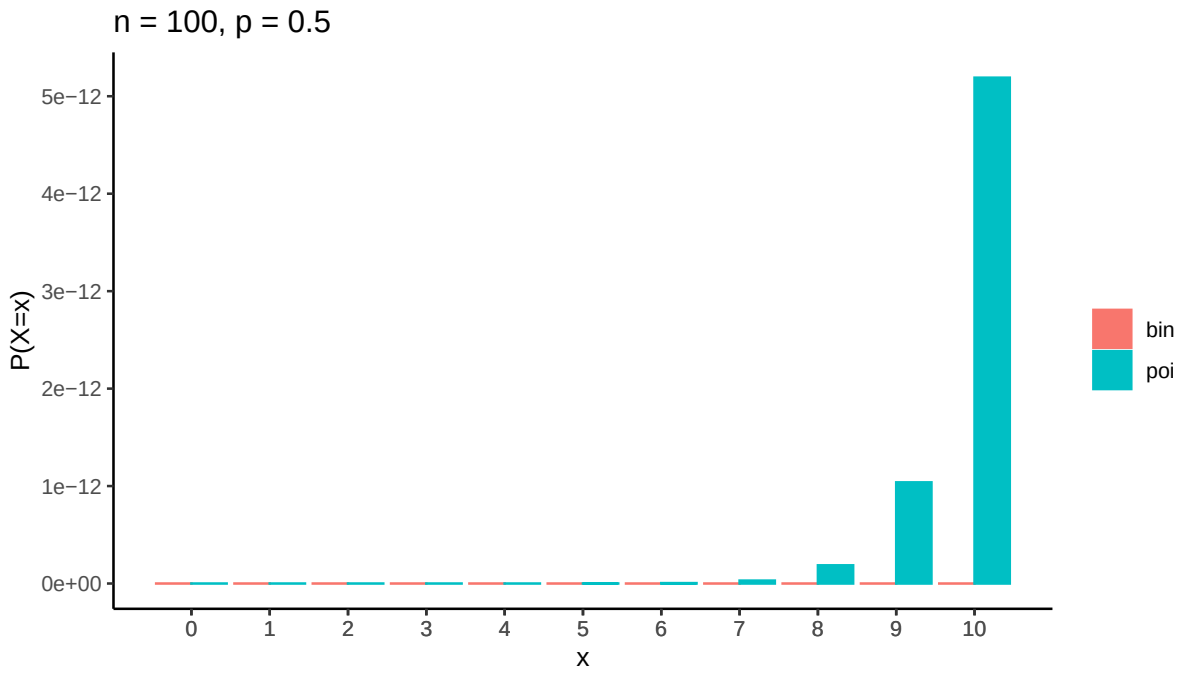
2.5.1 Poisson-Binomial relationship

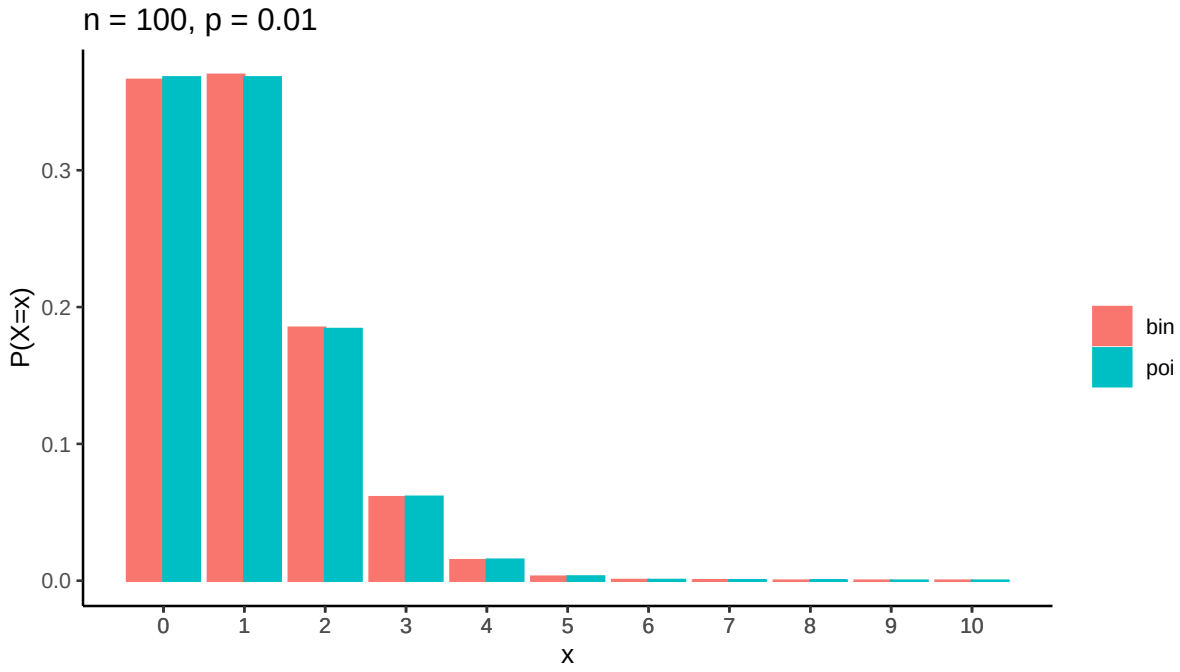
The Poisson distribution plays a useful approximation role for some of the other main discrete distributions. Let $X \sim \text{Bin}(n, p)$. Then

$$X \approx \text{Poi}(np)$$

when n is large and p is small. Typically the rule of thumb is $n > 20$ and $np < 5$ or $n(1-p) < 5$.







Let $\lambda = np$. Consider the limit of as $n \to \infty$ of the binomial pmf:

$$\begin{aligned}
\lim_{n\to\infty} \Pr(X = x) &= \lim_{n\to\infty} \frac{n!}{x!(n-x)!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \\
&= \frac{\lambda^x}{x!} \lim_{n\to\infty} \underbrace{\frac{n!}{n^x(n-x)!}}_{\to 1} \underbrace{\left(1 - \frac{\lambda}{n}\right)^n}_{\to e^{-\lambda}} \underbrace{\left(1 - \frac{\lambda}{n}\right)^{-x}}_{\to 1} \\
&= \frac{e^{-\lambda}\lambda^x}{x!} \\
&= \Pr(Y = x), Y \sim \text{Poi}(\lambda).
\end{aligned}$$

Some details...

$$\begin{aligned}\frac{n!}{n^x(n-x)!} &= \frac{n(n-1)(n-2)\cdots 3\cdot 2\cdot 1}{n\cdot n\cdots n\cdot (n-x)(n-x-1)\cdots 3\cdot 2\cdot 1}\\ &= \frac{n}{n}\frac{n-1}{n}\cdots\frac{n-x+1}{n}\end{aligned}$$

and each term converges to 1 as $n\to\infty$. Also by definition,

$$e^a = \lim_{n\to\infty}\left(1+\frac{a}{n}\right)^n.$$

2.5.2 Poisson-Exponential

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.

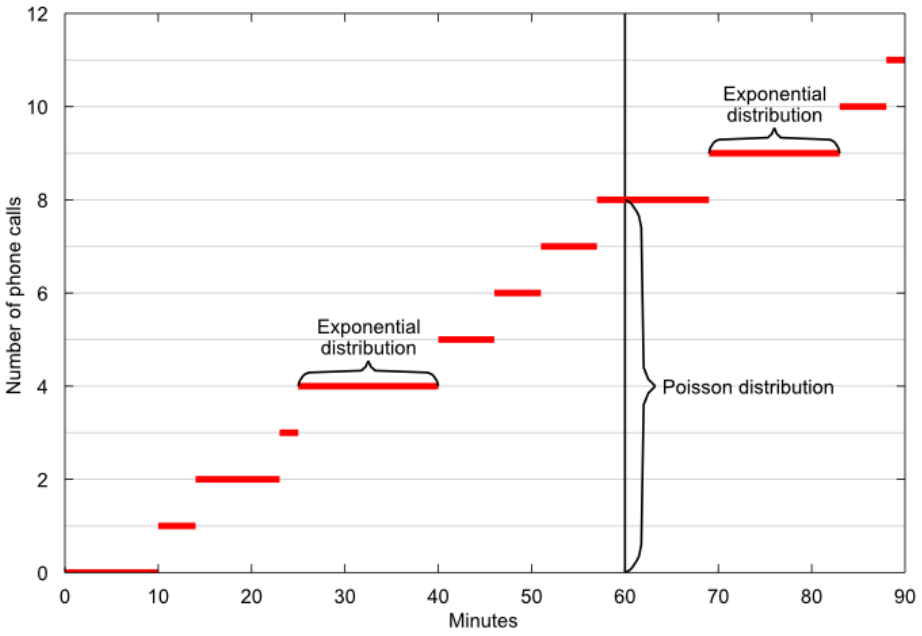


Figure 2.3: Poisson-exponential process.

Let

- N_t be the number of phone calls during time period t; and
- X_t be the waiting time until the next phone call from one at t.

By definition, the two events are equivalent: $\{X_t > x\} \equiv \{N_t = N_{t+x}\}$. Then $\Pr(X_t \le x) = 1 - \Pr(N_t - N_{t+x} = 0)$.

- $\Pr(N_t - N_{t+x} = 0)$ is the probability of no calls between time period $t+x$ and t, which 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

  $$\Pr(X_t \le x) = 1 - e^{-\lambda x}$$

  which is the cdf of an $\text{Exp}(\lambda)$ distribution.

2.5.3 Poisson-Gamma

More generally, the Poisson and gamma (which includes exponential) 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

$$\Pr(X > x) = \Pr(Y < 1) = \Pr(Y = 0) = e^{-x/\beta}.$$

2.5.4 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
$X = c$	$c - 0.5 < X < c + 0.5$
$X < c$	$X < c + 0.5$
$X \leq c$	$X < c + 0.5$
$X > c$	$X > c - 0.5$
$X \geq c$	$X > c - 0.5$

Chapter 3

Inequalities, convergences, and normal random samples

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 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

3.1.1 Random sampling

Collection of data $X_1, \ldots, X_n$ in an experiment consists of several observations on a variable of interest X.

- 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 $X_1, \ldots, X_n$ to be random variables sampled from a population whose pdf or pmf is $f_X(x)$. We typically write

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} f_X.$$

We impose the distribution f_X onto the data as an assumption! As the British statistician George Box once famously said, "all models are wrong, but some are useful".

3.1.2 Independent and identical r.v.

Usually, the samples are 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,

$$f_X(x_1, \dots, x_n) = \prod_{i=1}^{n} f(x_i)$$

In particular, if the population pdf/pmf is a member of a *parametric family*, say one of those introduced in Part 2, then we can write

$$f_X(x_1, \dots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

We could then use the random samples to *infer* about the (unknown) parameter θ. More on this in the coming parts.

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.

3.1.3 Statistic

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.

Here are two very commonly used statistics:

- The sample mean
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$
- The (unbiased) sample variance
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Lemma 3.1. *Let $X_1, \dots, X_n$ be a r.v. 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.

3.1.4 Sampling distribution

Realise that

- A statistic T_n is itself a r.v..
- If it is random, it has a distribution.
- Along with having a distribution, all of concepts and properties we discussed in Parts 1 & 2 apply.

Think about the statement above, "a statistic T_n is itself a r.v."–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.

- In parametric statistics for example, the statistic T_n may serve as an *estimator* for the true unknown value θ.
- How do we know that T_n is a *good* estimator? It takes different values with repeated sampling, is it typically close to θ? How close?

3.1.5 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

$$\bar{X} \sim \mathrm{N}(\mu, \sigma^2/n). \tag{3.1}$$

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

$$n^{-1} \sum_{i=1}^{n} \tan^{-1}(X_i)?$$

We use 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 $n \to \infty$. For this, we first 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* r.v. exceeds some positive constant.

Lemma 3.2 (Markov's inequality)**.** *Let $X \geq 0$ be a non-negative r.v. 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 r.v..

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 r.v. X and any constant $t > 0$,*

$$\begin{aligned}
\Pr(|X| \geq t) &\leq \frac{E|X|}{t} \quad \textit{provided } E|X| < \infty \\
\Pr(|X| \geq t) &\leq \frac{E(|X|^k)}{t^k} \quad \textit{provided } E(|X|^k) < \infty
\end{aligned}$$

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 r.v. 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,

$$\Pr(|X| \geq 1) \leq 1.00$$
$$\Pr(|X| \geq 2) \leq 0.25$$
$$\Pr(|X| \geq 3) \leq 0.11$$

In contrast, suppose that we know that X is normally distributed. Then

$$\Pr(|X| \geq 1) \leq 0.318$$
$$\Pr(|X| \geq 2) \leq 0.046$$
$$\Pr(|X| \geq 3) \leq 0.003$$

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.

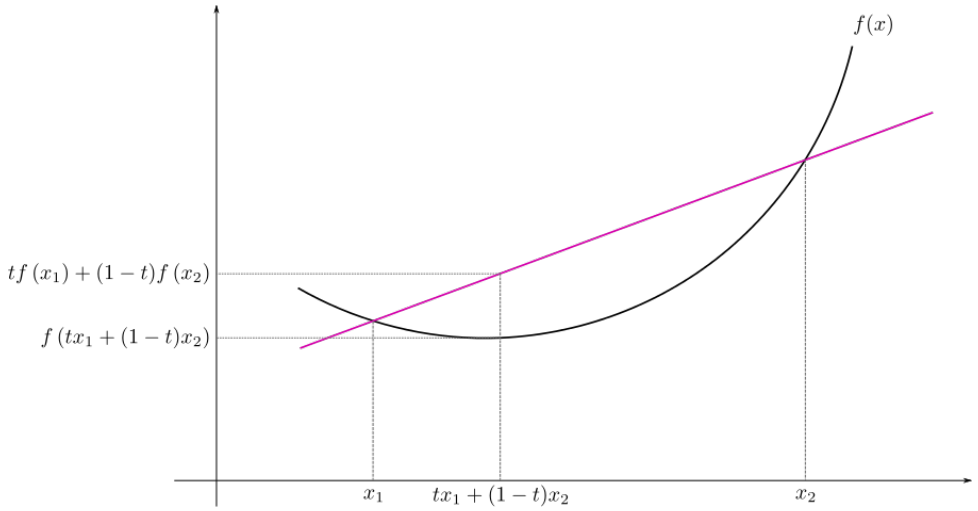
- A function g is **convex** if for any x, y and any $\alpha \in [0,1]$,

$$g(\alpha x + (1-\alpha)y) \leq \alpha g(x) + (1-\alpha)g(y).$$

- 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 r.v.s

Lemma 3.5 (Jensen's inequality)**.** *Let X be a r.v. 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

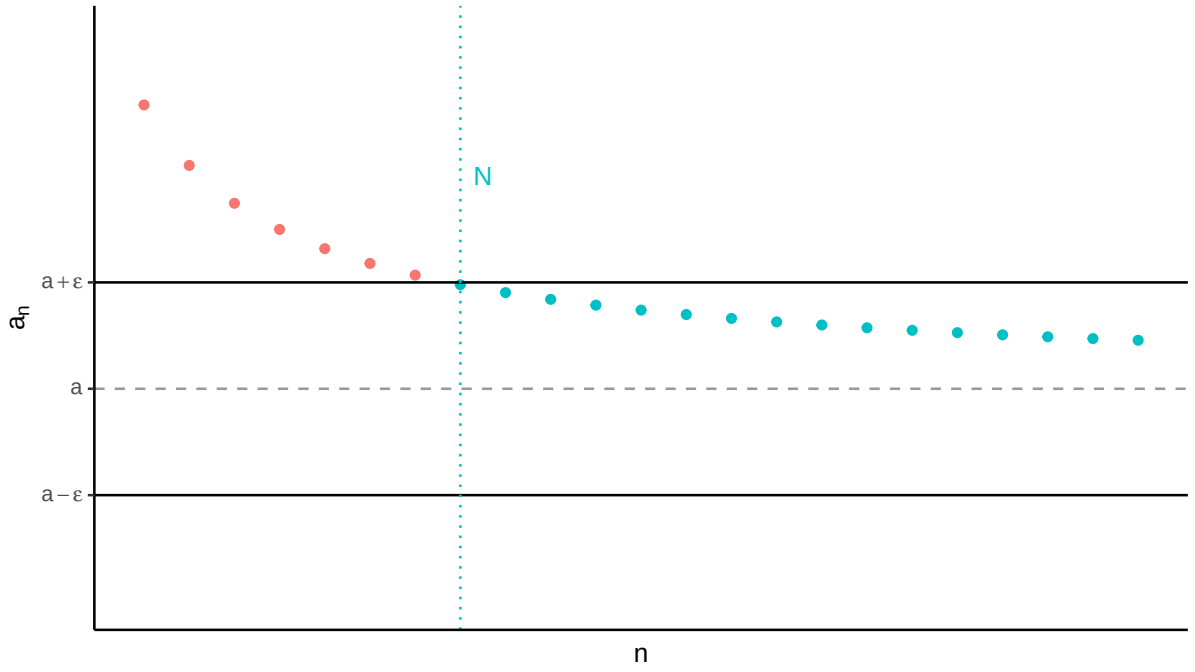
Recall the limits of sequences of real numbers (a_n), $n \in \mathbb{N}$.

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 $|a_n - a| < \epsilon$.

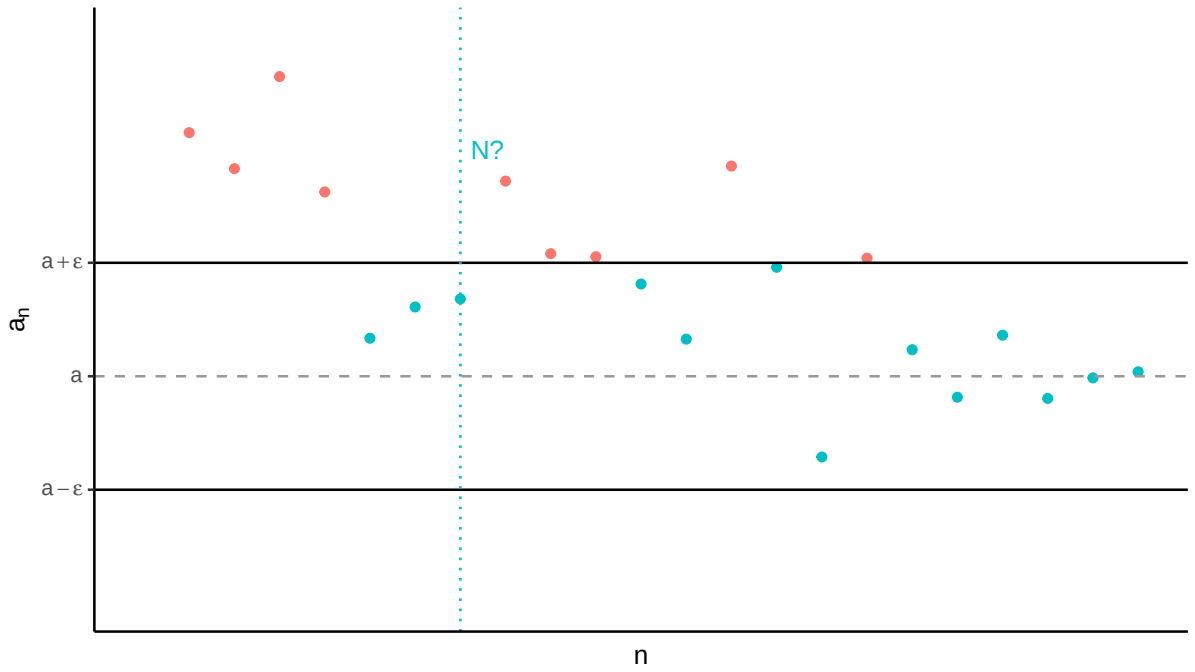
We write $\lim_{n\to\infty} a_n = a$, or simply $a_n \to a$. This also means that $|a_n - a| \to 0$ as $n \to \infty$. For every measure of closeness ϵ, the sequence's terms are eventually that close to the limit.

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 r.v., 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 r.v. 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 r.v.s. Let $X_1, X_2, ...$ be a sequence of r.v., and X be another r.v.. The main types of convergence for r.v. 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

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 r.v. or a constant.

Example 3.3. Let $X_i \overset{\text{iid}}{\sim} \text{Unif}(0, 1)$. Define another r.v. $Y_n = \min\{X_1, ..., X_n\}$. 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
```

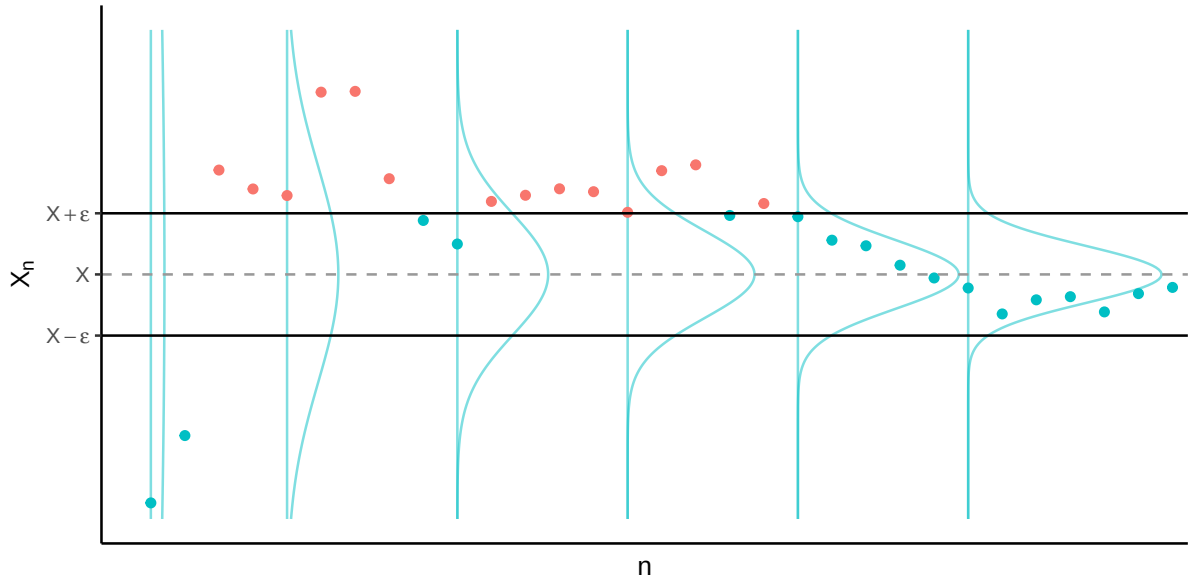
A good guess with be $Y_n \to 0$, so let's prove this. We want to show that

$$\Pr(|Y_n - 0| \geq \epsilon) = \Pr(Y_n \geq \epsilon) \to 0$$

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 $Y_n \xrightarrow{\text{P}} 0$.



3.3.2 Convergence in distribution

Instead of considering the convergence of the r.v. itself, we consider the convergence of the *distribution* of the sequence of r.v.s.

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 r.v. 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 r.v. 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

It is sometimes more convenient to consider the mean-square convergence:

Definition 3.6 (Mean-square convergence)**.** X_n converges in mean-square to X if

$$\lim_{n\to\infty} \text{E}\left[(X_n - X)^2\right] = 0.$$

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

$$X_n = \begin{cases} n^2 & \text{w.p. } 1/n \\ 0 & \text{w.p. } 1 - 1/n \end{cases}$$

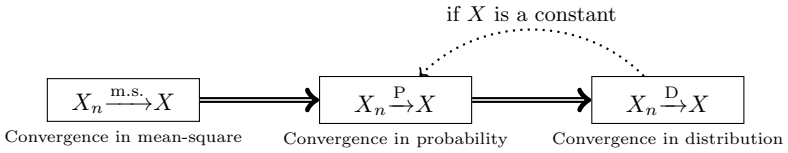
Then, for any $\epsilon > 0$, $\Pr(|X_n| \geq \epsilon) = \Pr(X_n = n^2) = 1/n \to 0$ as $n \to \infty$. Hence, $X_n \xrightarrow{\text{P}} 0$.

However,

$$\mathrm{E}(X_n^2) = n^2 \cdot \Pr(X_n = n^2) + 0 \cdot \Pr(X_n = 0) = n \to \infty$$

hence $X_n \not\xrightarrow{\text{m.s.}} 0$.

3.3.4 Relationship between convergences



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

Theorem 3.1 (Slutzky's Theorem)**.** *Let X_n, Y_n, X, and Y be r.v., 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 $X_n + Y_n \xrightarrow{\text{D}} X + Y$.

3.4 Limit theorems

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 r.v.s 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 $\text{E}(\bar{X}_n) = \mu$ and $\text{Var}(\bar{X}_n) = \sigma^2/n$. 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, $\bar{X}_n \xrightarrow{\text{P}} \mu$. □

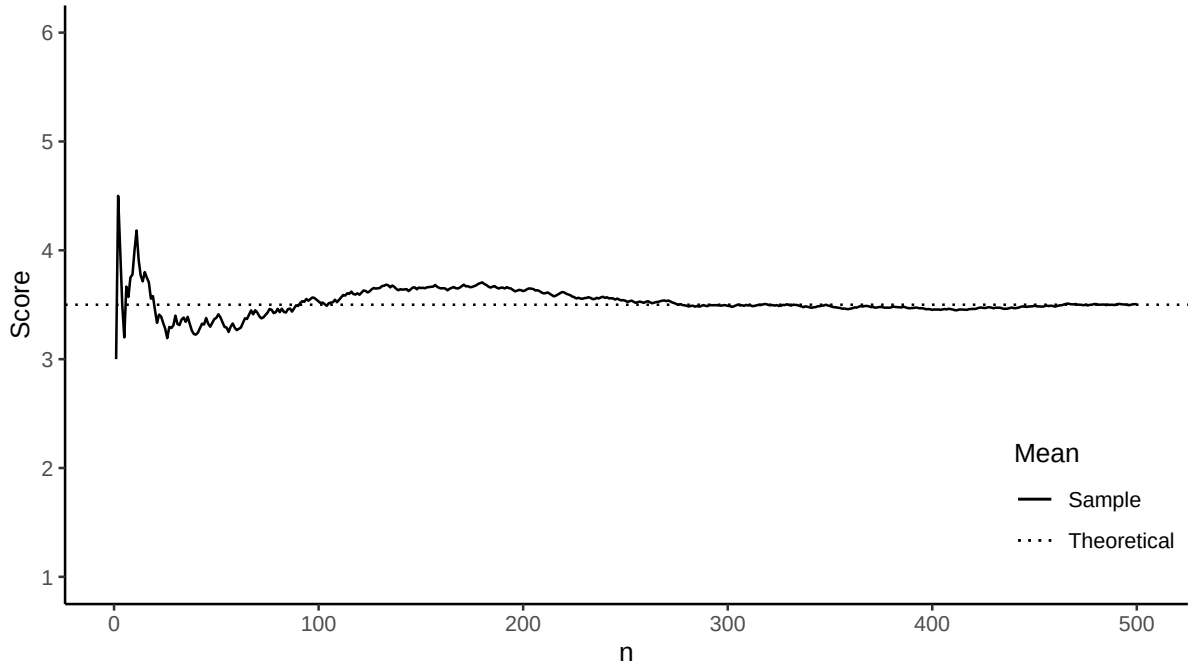
As an illustration of the WLLN, consider an experiment where we throw a six-sided die repeatedly and independently. Let $X_1, X_2, ...$ be the scores of the dice throws. We know that the true mean is $\mu = 3.5$. Let's simulate some dice throws:

```
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 r.v. 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 3.6. 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).

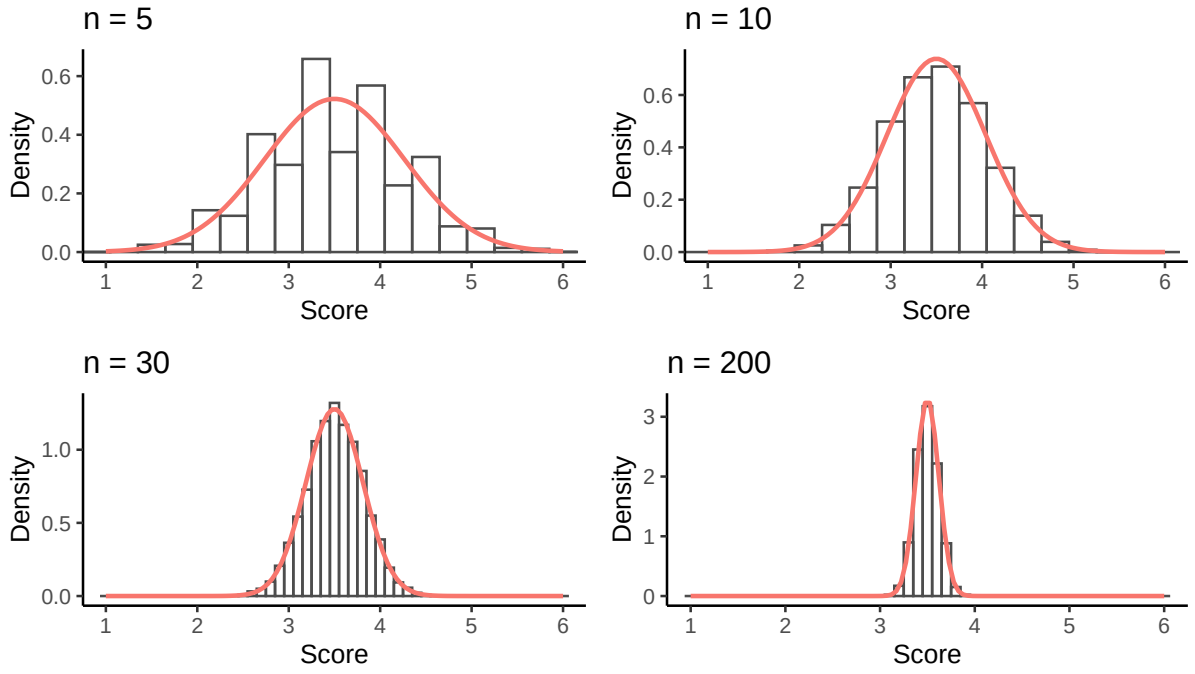
Here's the `R` code to replicate dice rolls and the sample means. The idea is to generate a sample of size n of dice roll scores repeatedly `B` times.

```
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

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 $\epsilon := 2\sigma/\sqrt{n} = 2\sqrt{\mathrm{Var}(\bar{X}_n)}$. 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 r.v. instead of the actual r.v. itself. For this, we use the delta method.

Theorem 3.4 (The delta method)**.** *Suppose that X_n is a sequence of r.v. 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*

$$\frac{\sqrt{n}(g(X_n) - g(\mu))}{|g'(\mu)|\sigma} \xrightarrow{D} \mathrm{N}(0,1).$$

In other words,

$$X_n \approx \mathrm{N}(\mu, \sigma^2/n) \quad \Rightarrow g(X_n) \approx \mathrm{N}\left(g(\mu), (g'(\mu))^2\sigma^2/n\right).$$

Example 3.7. Suppose we observe $X_1, \dots, X_n \sim \text{Bern}(p)$. 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 $\text{Var}(X_i) = p(1-p)$.

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

$$\frac{\hat{p}}{1-\hat{p}} \approx \text{N}\left(\frac{p}{1-p}, \frac{p}{n(1-p)^3}\right).$$

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 $\{X_1, \dots, X_n\}$ be a sample from $\text{N}(\mu, \sigma^2)$, and let*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i, \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2, \qquad \text{and} \qquad SE(\bar{X}) = S/\sqrt{n}.$$

Then,

- *$\bar{X}$ and S^2 are independent r.v.s*
- *$\bar{X} \sim \text{N}(\mu, \sigma^2/n)$*
- *$(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$*
- *$\frac{\sqrt{n}(\bar{X}-\mu)}{S} = \frac{\bar{X}-\mu}{SE(\bar{X})} \sim t_{n-1}$*

Given that the above theorem mentions two kinds of distribution (that you may have heard of) but we have 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 $Z_1, \dots, Z_k \overset{\text{iid}}{\sim} \text{N}(0,1)$, i.e. each Z_i has pdf $f(z_i) = (2\pi)^{-1/2}e^{-z_i^2/2}$ for $i = 1, \dots, k$. 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 $X \sim \chi^2_k$.

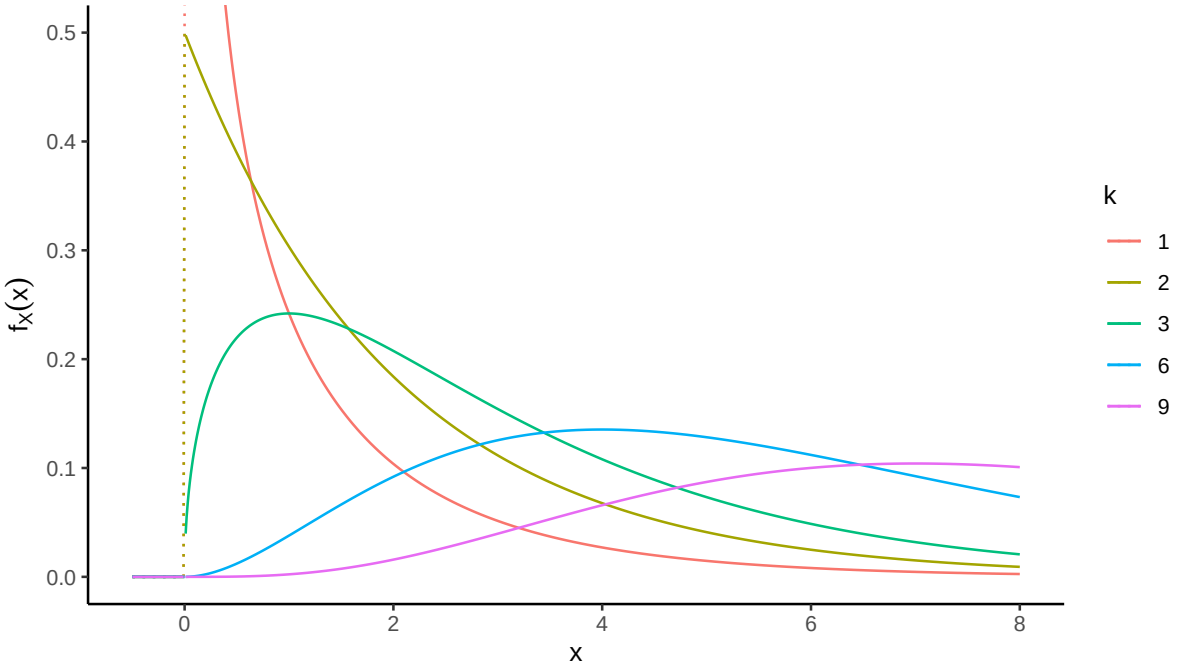
Out of curiosity, the pdf of a χ^2_k distribution is $f(x) = Cx^{k/2-1}e^{-x/2}$, where the normalising constant C is equal to $2^{-k/2}\Gamma^{-1}(k/2)$ ($\Gamma(\cdot)$ is the gamma function). The form of the pdf is less important to know than the definition of χ^2_k distribution given in Definition 3.7.

Here are some important properties of the χ^2_k distribution.

- X has support over $[0, \infty)$.
- $\text{E}(X) = k$.
- $\text{Var}(X) = 2k$.
- 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

$$\Pr(\chi_k^2 \leq x) = \int_0^x f_X(\tilde{x})\,\mathrm{d}\tilde{x}$$

where f_X is the pdf of χ_k^2 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 $x := \chi_k^2(\alpha)$ such that

$$\Pr(\chi_k^2 \leq x) = \int_0^x f_X(\tilde{x})\,\mathrm{d}\tilde{x} = A = 1 - \alpha$$

for various values of A and k.

Example 3.8. Let $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$. Then, $Z_i = \frac{Y_i - \mu}{\sigma} \sim \mathrm{N}(0, 1)$, and hence

$$\frac{1}{\sigma^2}\sum_{i=1}^n (Y_i - \mu)^2 = \sum_{i=1}^n Z_i^2 \sim \chi_n^2.$$

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 $\frac{n}{\sigma^2}(\bar{Y}_n - \mu)^2 \sim \chi_1^2$. Thus, by the properties of the χ^2-distribution, the decomposition in (3.3) may be written as $\chi_n^2 = \chi_{n-1}^2 + \chi_1^2$. In particular, we now know

$$\frac{1}{\sigma^2}\sum_{i=1}^n (Y_i - \bar{Y}_n)^2 \sim \chi_{n-1}^2.$$

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 r.v. $Z \sim \mathrm{N}(0,1)$ 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

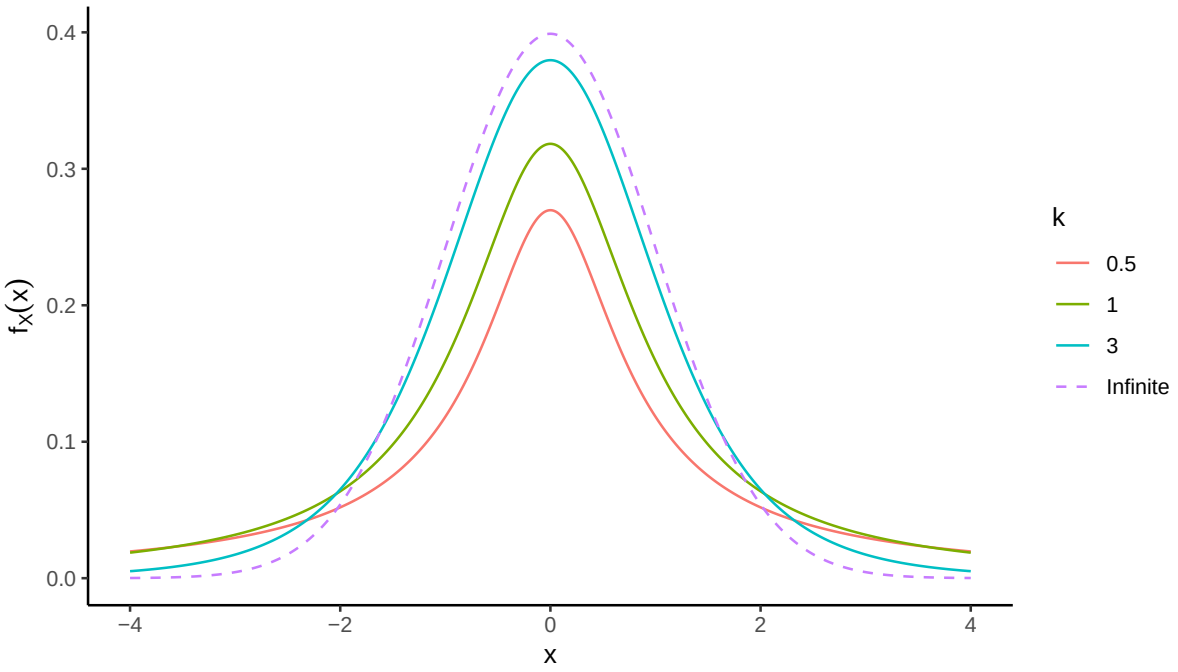
$$f(t) \propto \left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}},$$

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



The t-distribution[1] 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 $k > 0$.

[1] Explore the t-distribution vs normal distribution here: https://eripoll12.shinyapps.io/t_Student/

Figure 3.1: William Sealy Gosset. 13 June 1876 – 16 October 1937.

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.

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,

$$\frac{\overbrace{\sqrt{n}(\bar{X} - \mu)/\sigma}^{\mathrm{N}(0,1)}}{\sqrt{\frac{\chi^2_{n-1}}{\frac{(n-1)S^2/\sigma^2}{n-1}}}} = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\bar{X} - \mu}{\mathrm{SE}(\bar{X})} \sim t_{n-1}.$$

□

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

$$Y = \frac{X_1/k_1}{X_2/k_2}$$

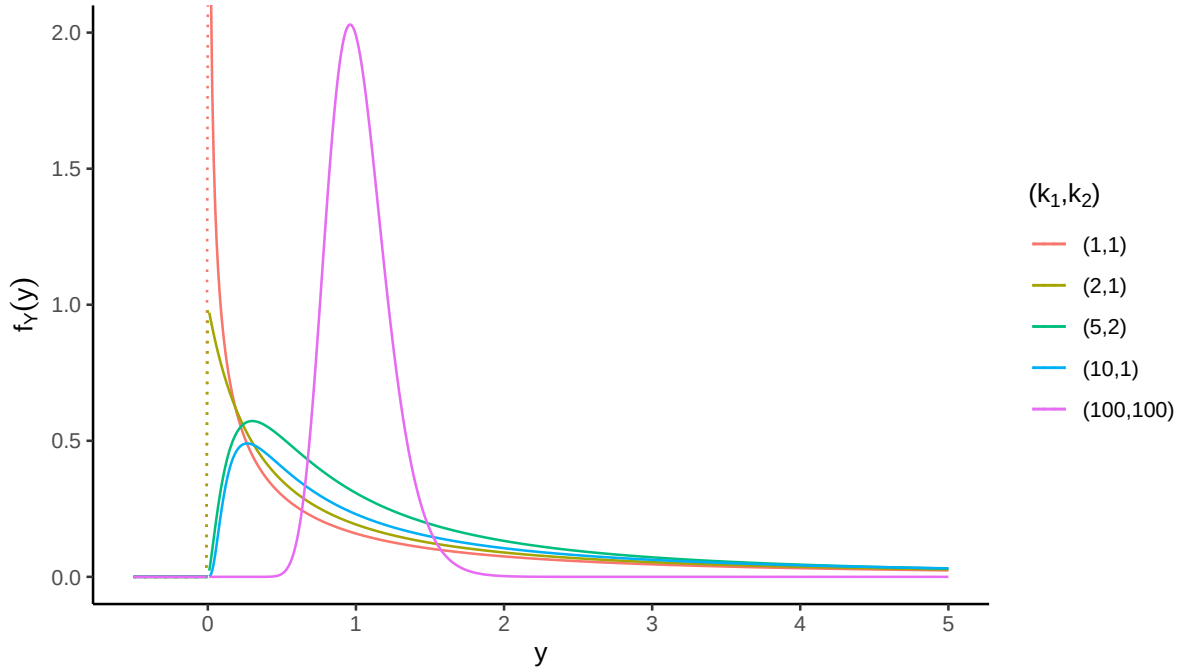
is called the F-distribution with (k_1, k_2) degrees of freedom. We write $Y \sim F_{k_1,k_2}$.

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:

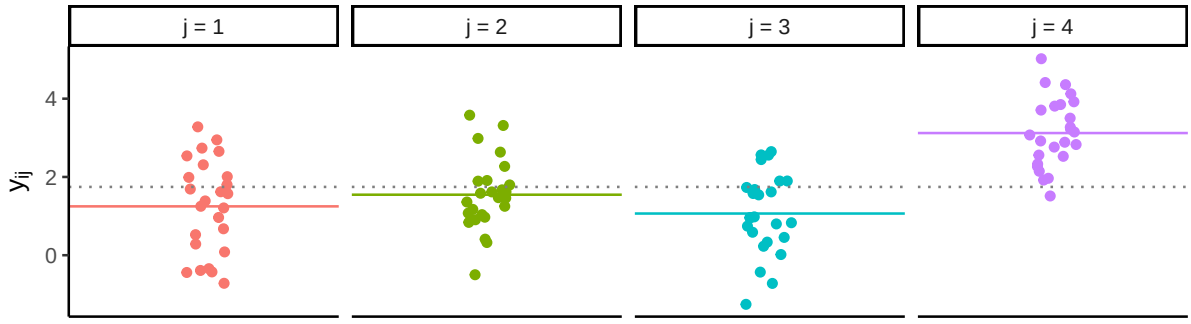
- Y is continuous and has support over $[0, \infty)$, provided $k_1 > 1$.
- $\mathrm{E}(Y) = \frac{k_2}{k_2 - 2}$, provided $k_2 > 2$.
- $\mathrm{Var}(Y) = \frac{2k_2^2(k_1 + k_2 - 2)}{k_1(k_2-2)^2(k_2-4)}$, provided $k_2 > 4$.
- Technically, $k_1, k_2 \in \mathbb{R}_{>0}$, but we will usually deal with $k_1, k_2 \in \mathbb{N}$.
- 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,

$$F = \frac{\text{mean } BSS}{\text{mean } WSS} = \frac{\overbrace{1/\sigma^2 \textstyle\sum_j n_j(\bar{Y}_j - \bar{Y})^2}^{\chi^2_{m-1}}/(m-1)}{\overbrace{1/\sigma^2 \textstyle\sum_{i,j}(Y_{ij} - \bar{Y}_j)^2}^{\chi^2_{n-m}}/(n-m)}$$

is a ratio of two χ^2-distributions, which means that F follows an F-distribution with $(m-1, n-m)$ degrees of freedom.

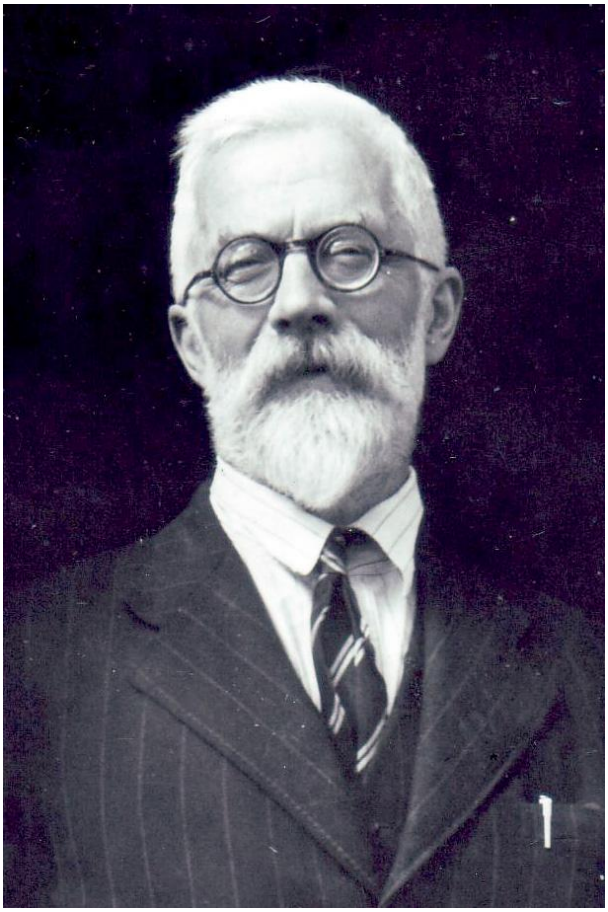


Figure 3.2: Sir Ronald Aylmer Fisher. 17 February 1890 – 29 July 1962.

Part III

Inference

Chapter 4

Point 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 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 $X = (X_1, \ldots, X_n)^\top$ 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

$$x = \{H, T, H, T, T, H, H, T, H, H\}.$$

Then $L(0.6|x) > L(0.5|x)$.

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)
```

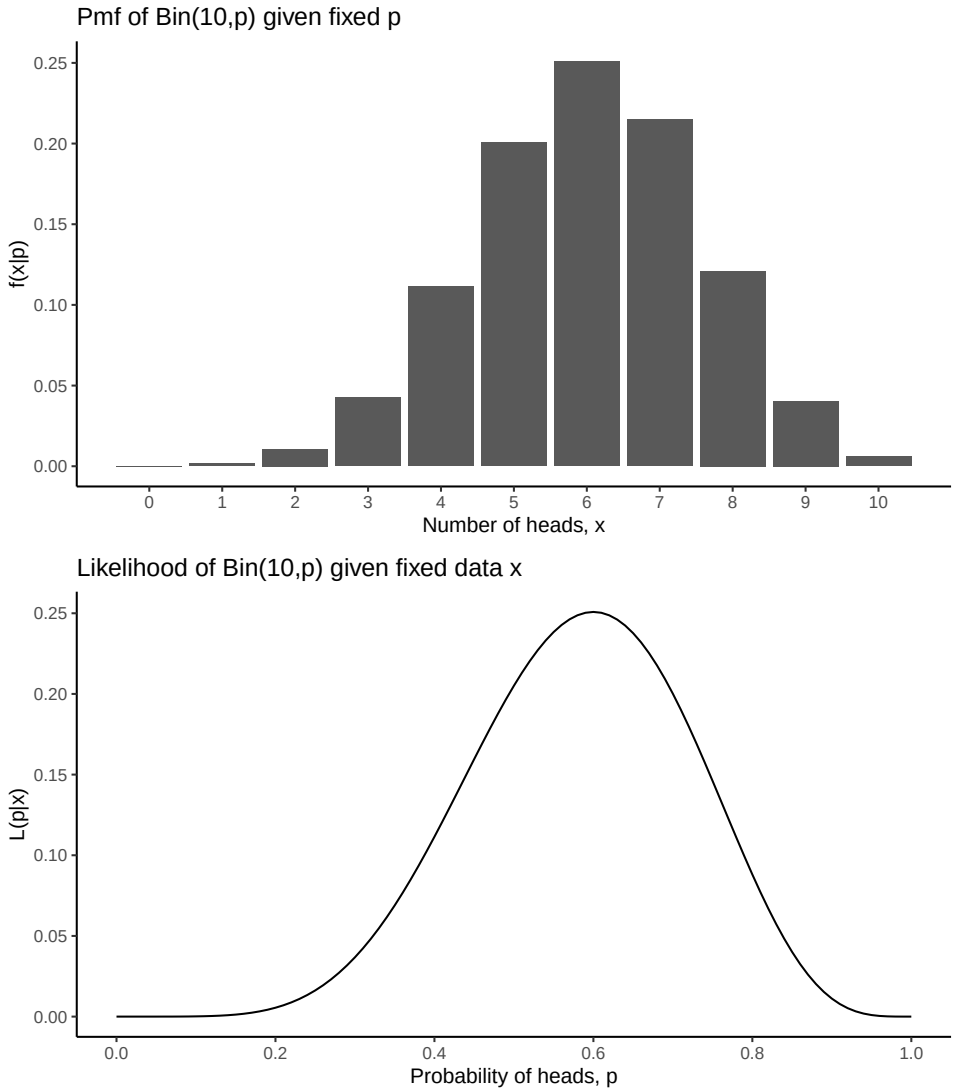
```
##  [1] 0.0001 0.0016 0.0106 0.0425 0.1115 0.2007 0.2508 0.2150 0.1209 0.0403
## [11] 0.0060
```

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
```

Plot of the likelihood



4.1.2 Likelihood ratio

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 $X_1, \dots, X_n$ 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 $1, 2, \dots, \theta$, with θ being unknown. Naqiyyah spends a day observing bus numbers and collects data X_i, $i = 1, \dots, n$, 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

$$f(x|\theta) = \begin{cases} \frac{1}{\theta^n} & \max(x_1, \dots, x_n) \le \theta \\ 0 & \text{otherwise} \end{cases}$$

Hence, if we let $T(x) = \max(x_1, \dots, x_n)$ then

$$f(x|\theta) = \overbrace{1}^{h(x)} \cdot \frac{\overbrace{\mathbb{1}_{t\le\theta}(t)}^{g(t|\theta))}}{\theta^n},$$

which implies that $T(X) = \max(X_1, \dots, X_n)$ 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 $\{X_1, \dots, X_n\}$ 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 $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Unif}(\theta, \theta+1)$. The joint pdf of X is

$$f(x|\theta) = \begin{cases} 1 & \theta < x_1, \dots, x_n < \theta + 1 \\ 0 & \text{otherwise} \end{cases}$$

This can be usefully re-expressed as

$$\begin{aligned} f(x|\theta) &= 1 \cdot \mathbb{1}_{\{x_1,\dots,x_n>\theta\}}(x)\, \mathbb{1}_{\{x_1,\dots,x_n<\theta+1\}}(x) \\ &= 1 \cdot \mathbb{1}_{t_1=\min(x_i)>\theta}(t_1)\, \mathbb{1}_{t_2=\max(x_i)<\theta+1}(t_2) \\ &= \underbrace{1}_{h(x)} \cdot \underbrace{\mathbb{1}_{\{t_1>\theta\}\cap\{t_2<\theta+1\}}}_{g(t_1,t_2|\theta)} \end{aligned}$$

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

Recall: $X_1, \dots, X_n \sim f(x|\theta)$ is a random sample, where 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. For $\text{N}(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)^\top$, so $p = 2$ and $\Theta = \mathbb{R} \times \mathbb{R}_{\geq 0}$.

For $\text{Poi}(\lambda)$, $\theta = \lambda$ and $\Theta = \mathbb{R}_{\geq 0}$.

The goal of **point estimation**:

> 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.

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).

A good estimator should make $|\hat{\theta} - \theta|$ as small as possible, despite

i. θ being unknown; and

ii. the value of $\hat{\theta}$ changes with the sample observed.

We will make use of the sampling properties of the r.v. $\hat{\theta}$ to quantify (and qualify) its worth as an estimator for θ.

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

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'.

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).

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 $\text{E}(U(X)) = \sum_{i=1}^{n} \text{E}([X_i = 0]) = ne^{-\lambda}$. Hence, the MOM estimator for λ based on U is

$$\hat{\lambda} = -\log(U/n).$$

4.5 Method of maximum likelihood

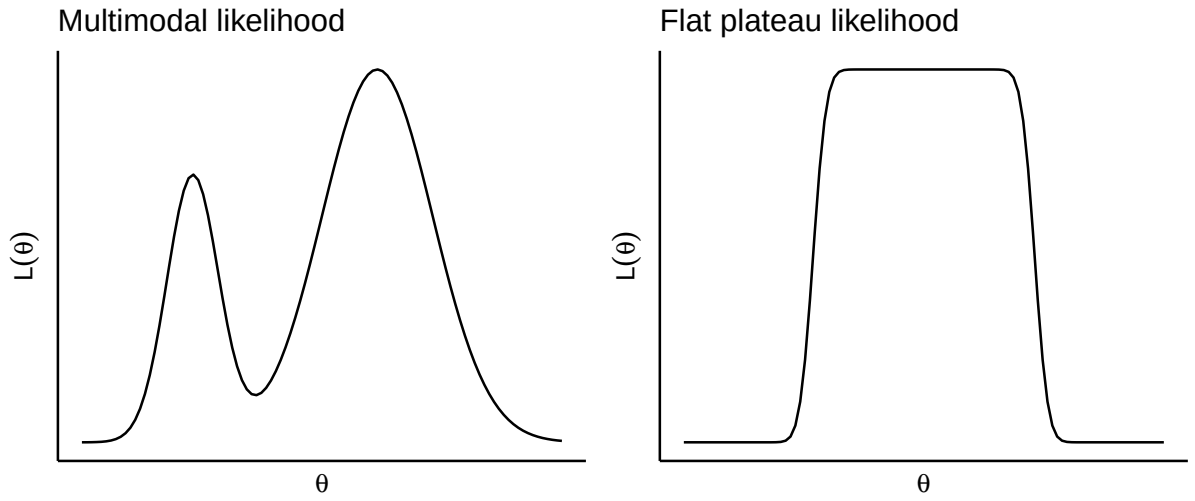
Definition 4.5 (Maximum likelihood (ML) estimator)**.** The ML estimator of θ is $\hat{\theta}$ which is such that

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta|X)$$

That is, 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. 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.

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

Typically we locate $\hat{\theta}$ by solving $l'(\hat{\theta}) = 0$, and then checking that the stationary point is a maximum. Several points on this:

- 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.).

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

```
X <- rnorm(n = 100, mean = 8, sd = 1)
mean(X)
```

```
## [1] 8.021617
```

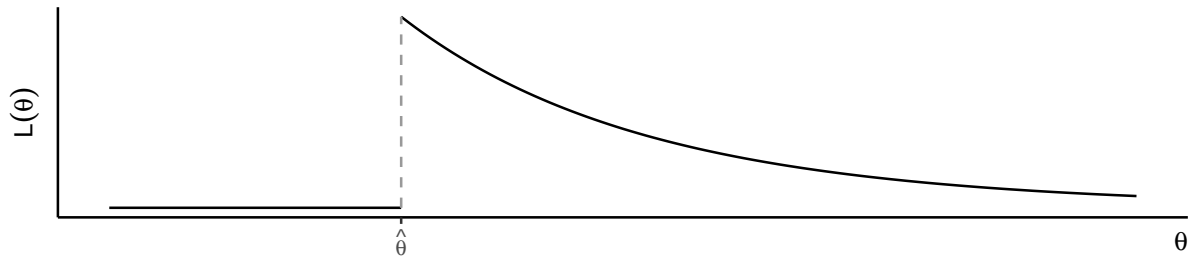
```
# Optimising the likelihood function
lik <- function(theta) -sum(dnorm(x = X, mean = theta, sd = 1, log = TRUE))
theta0 <- 1  # starting value
res <- optim(par = theta0, fn = lik, method = "BFGS", lower = -Inf,
             upper = Inf)
res$par
```

```
## [1] 8.021617
```

Sometimes, a sketch of $l(\theta)$ reveals that the MLE does not satisfy $l'(\hat{\theta}) = 0$. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Unif}(0, \theta)$. Then the pdf of X is $f(x|\theta) = 1/\theta^n$ for $X_1, \ldots, X_n < \theta$. The likelihood is therefore

$$L(\theta|X) = \begin{cases} \frac{1}{\theta^n} & \theta > \max(X_1, \ldots, X_n) \\ 0 & \text{otherwise} \end{cases}$$

which is maximised at $\hat{\theta} = \max(X_1, \ldots, X_n)$.



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 $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bern}(\pi)$. 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 $\hat{\theta} = 1$.

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

Definition 4.6 (Bias)**.** The bias of an estimator $\hat{\theta}$ is defined to be

$$\text{Bias}_\theta(\hat{\theta}) = \text{E}_\theta(\hat{\theta}) - \theta.$$

- The subscript θ makes clear the fact that the expectation is taken under the distribution using θ as the true value of the parameter.
- When $\mathrm{E}_\theta(\hat{\theta}) = \theta$, $\mathrm{Bias}_\theta(\hat{\theta}) = 0$ for all possible values of θ, and in this case $\hat{\theta}$ is called an **unbiased estimator** for θ.
- Small bias, and even unbiasedness, is desirable.

4.6.2 Variance and standard error

Definition 4.7 (Variance)**.** The variance of an estimator $\hat{\theta}$ is defined to be

$$\mathrm{Var}_\theta(\hat{\theta}) = \mathrm{E}_\theta\left[(\hat{\theta} - \mathrm{E}(\hat{\theta}))^2\right]$$

This just uses the regular definition of the variance for random variables.

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.

$$\mathrm{se}(\hat{\theta}) = \sqrt{\mathrm{Var}_\theta(\hat{\theta})}$$

Obviously, we desire an estimator whose variability (in repeated sampling) is low.

These two properties measure different things about estimators:

- Bias is a measure of *accuracy*.
- Variance is a measure of *precision*.

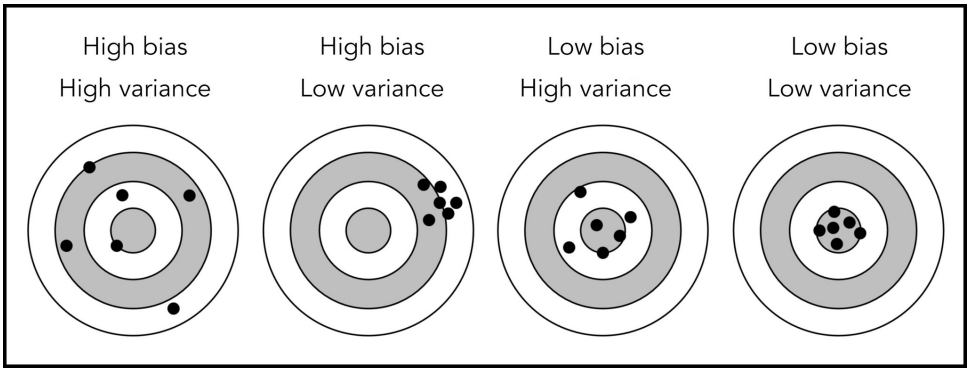


Figure 4.1: The difference between bias and variance.

4.6.3 Mean squared error

Definition 4.9 (Mean squared error of estimator)**.** The MSE of the estimator $\hat{\theta}$ is defined as

$$\mathrm{MSE}_\theta(\hat{\theta}) = \mathrm{E}_\theta\left[(\hat{\theta} - \theta)^2\right] = \{\mathrm{Bias}_\theta(\hat{\theta})\}^2 + \mathrm{Var}_\theta(\hat{\theta}).$$

As an exercise, prove the bias-variance decomposition above. Hare some hints on how to get started:

- hint 1
- hint 2

There is a clear and direct relationship between the MSE of an estimator, and its bias and variance. For a given MSE,

- Reducing the bias of an estimator implies that its variance will increase.
- 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:

$$\text{E}\,\hat{\theta}_{ML} = \int_0^\theta x \frac{nx^{n-1}}{\theta^n}\,\text{d}x = \left[\frac{nx^{n+1}}{\theta^n(n+1)}\right]_0^\theta = \frac{n\theta}{n+1}$$

Therefore, the bias is $\text{Bias}(\hat{\theta}_{ML}) = -\theta/(n+1) \neq 0$. Note that this tends to 0 as $n \to \infty$, but can be substantial when n is small.

For $\hat{\theta}_{MOM}$, we have

$$\text{Var}(\hat{\theta}_{MOM}) = 4\,\text{Var}(\bar{X}) = \frac{4\,\text{Var}(X_i)}{n} = \frac{\theta^2}{3n}.$$

Note that $\text{Var}(\hat{\theta}_{MOM}) \to 0$ as $n \to \infty$ at the rate of $1/n$. This is typical behaviour of 'good' estimators.

For $\hat{\theta}_{ML}$:

$$\begin{aligned}\text{Var}(\hat{\theta}_{ML}) = \text{E}(\hat{\theta}_{ML}^2) - \text{E}^2(\hat{\theta}_{ML}) &= \int_0^\theta x^2 \frac{nx^{n-1}}{\theta^n}\,\text{d}x - \frac{n^2\theta^2}{(n+1)^2} \\ &= \theta^2\left(\frac{n}{(n+1)^2(n+2)}\right)\end{aligned}$$

so $\text{Var}(\hat{\theta}_{ML}) \to 0$ as $n \to \infty$ but at a faster rate of $1/n^2$.

For $\hat{\theta}_{MOM}$, we have

$$\text{MSE}(\hat{\theta}_{MOM}) = \text{Var}(\hat{\theta}_{MOM}) = \frac{\theta^2}{3n}.$$

For $\hat{\theta}_{ML}$:

$$\text{MSE}(\hat{\theta}_{ML}) = \text{Bias}(\hat{\theta}_{ML}^2) + \text{Var}(\hat{\theta}_{ML}) = \theta^2\left(\frac{n}{(n+1)^2(n+2)} + \frac{1}{(n+1)^2}\right).$$

Notice that since $\text{MSE}(\hat{\theta}_{MOM})$ is $o(1/n)$ and $\text{MSE}(\hat{\theta}_{ML})$ is $o(1/n^2)$,

$$\text{MSE}(\hat{\theta}_{ML}) \leq \text{MSE}(\hat{\theta}_{MOM})$$

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 $\theta \in \Theta$,*

$$\mathrm{Var}_\theta(\hat{\theta}) \geq \frac{1}{\mathrm{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right]}.$$

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. □

[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.

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 $\mathcal{I}(\theta)$. 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

$$\begin{aligned} l(\lambda|X) &= \text{const.} - n\lambda + \sum_{i=1}^{n} X_i \log\lambda \\ l'(\lambda|X) &= -n + \sum_{i=1}^{n} X_i/\lambda \\ -l''(\lambda|X) &= \sum_{i=1}^{n} X_i/\lambda^2 \quad \text{(the observed Fisher information)} \end{aligned}$$

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.

Chapter 5

Hypothesis testing

Hello

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 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 $X = (X_1, \dots, X_n)^\top$ 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, $X_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu_0, \sigma^2)$. So we simulate a data set $\{X_1, \dots, X_n\}$ 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 $\Pr(|Y| > |\sqrt{n}(\bar{x} - \mu_0)/s|)$, 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
```

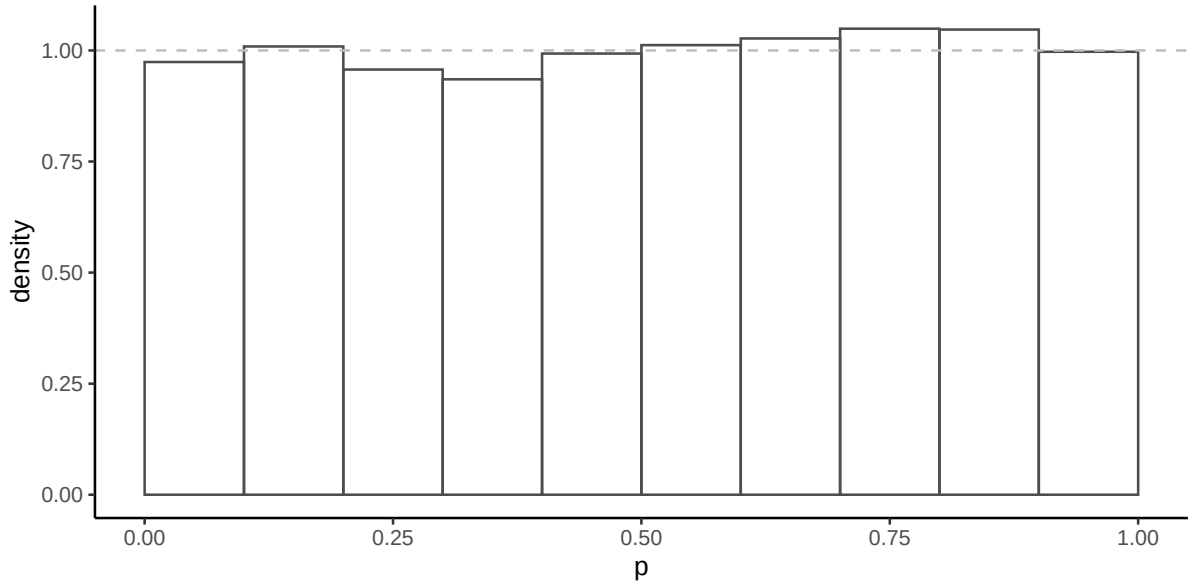
Simulate this `B=10000` times in a `for` loop:

```
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.

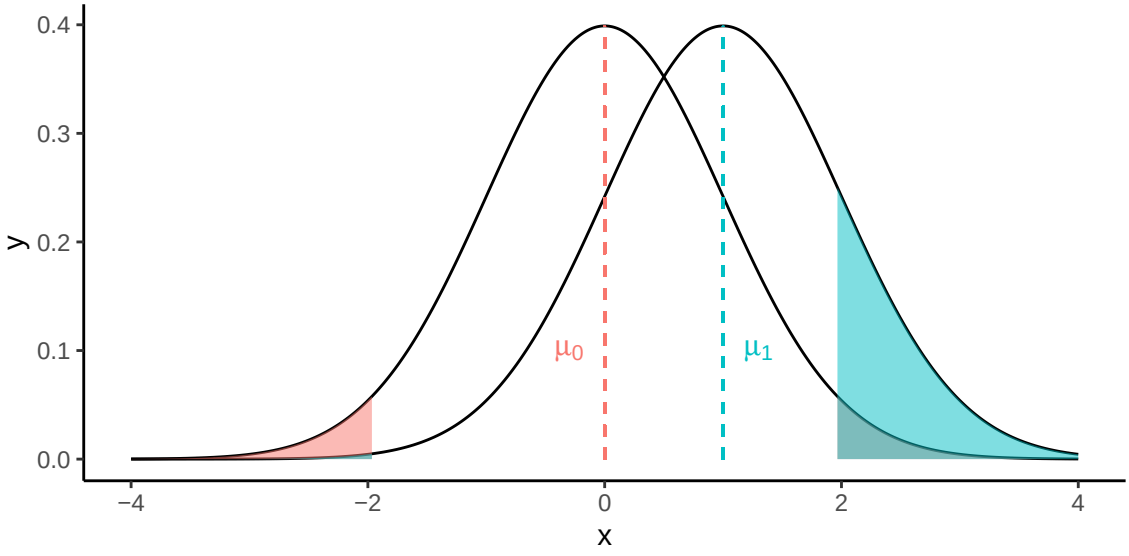
For our illustrative data, with $\sigma = 4.3$ and $\bar{x} = -12.8$,

$$\left| \frac{\bar{X}}{\sigma/\sqrt{n}} \right| = \left| \frac{-12.8}{4.3/\sqrt{10}} \right| = 9.413 > 1.96.$$

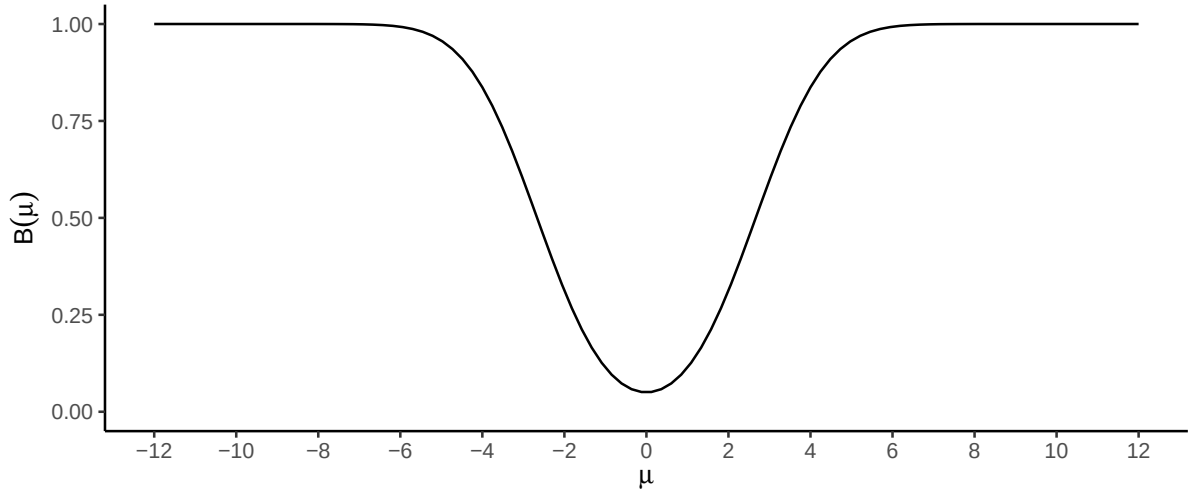
For a test of size $\alpha = 0.05$, the power of the test is

$$\begin{aligned} B(\mu) &= \Pr \left\{ \left| \frac{\bar{X} - \mu + \mu}{\sigma/\sqrt{n}} \right| \geq 1.96 \right\} \\ &= \Pr \left\{ \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq -1.96 - \frac{\mu}{\sigma/\sqrt{n}} \right\} + \Pr \left\{ \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \geq 1.96 - \frac{\mu}{\sigma/\sqrt{n}} \right\} \\ &= \Phi \left(-1.96 - \sqrt{n}\mu/\sigma \right) + \left[1 - \Phi \left(1.96 - \sqrt{n}\mu/\sigma \right) \right] \end{aligned}$$

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

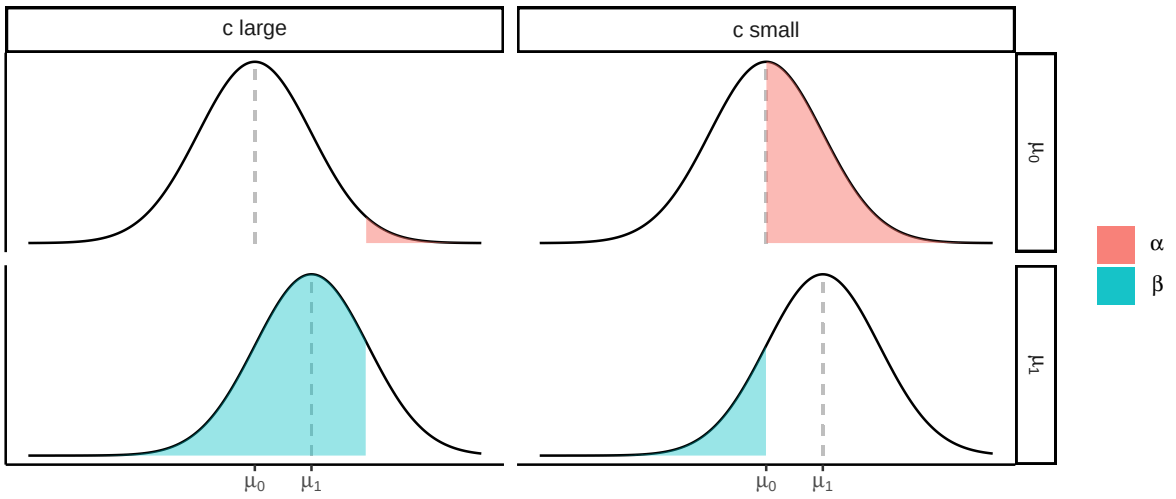
$$\beta(\theta) = \Pr(\text{Fail to reject } H_0 \mid H_0 \text{ is false}) = 1 - B(\theta).$$

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 $\Pr_{\theta_1}(W \in R)$ greater than $\Pr_{\theta_1}(W_{LR} \in R_{LR})$.*

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 $X_1, \ldots, X_n \overset{iid}{\sim} f(x|\theta)$ with $f(x|\theta)$ satisfying the usual regularity conditions. Consider testing the composite hypothesis for $\theta \in \mathbb{R}^p$*

$$H_0 : \theta \in \Theta_0 \quad \textit{v.s.} \quad H_1 : \theta \in \Theta \ \ \Theta_0.$$

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 $X_1, \ldots, X_n$ 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

$$L(\mu_1, \ldots, \mu_n) \propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(X_i - \mu_i)^2\right\},$$

Then, the unconstrained MLE are $\hat{\mu}_i = X_i$, while the constrained MLE is $\tilde{\mu} = \bar{X}$. Hence,

$$W_{LR} = \frac{L(\hat{\mu}_1, \ldots, \hat{\mu}_n)}{L(\tilde{\mu}, \ldots, \tilde{\mu})} = \exp\left\{\frac{1}{2}\sum_{i=1}^{n}(X_i - \bar{X})^2\right\}.$$

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 $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x|\theta)$ 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: $\text{se}(\hat{\theta}_n) = 1\Big/\sqrt{\mathcal{I}(\hat{\theta}_n)}$
  - Using the observed Fisher information: $\text{se}(\hat{\theta}_n) = 1\Big/\sqrt{-l''(\hat{\theta}_n|X)}$
- 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. $\text{se}(\hat{\theta}_n)$ and $\text{se}(\log\hat{\theta}_n)$) 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: $\sqrt{n}\,(\bar{X} - \mu) \xrightarrow{\text{D}} \text{N}\,(0, \sigma^2)$. From this, an (approximate) standard error of the estimator $\bar{X}$ is $\text{se}(\bar{X}) = s/\sqrt{n}$. 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:

$$p(x) = \Pr_{\mu=300}\left(\frac{\bar{X} - \mu}{\text{se}(\bar{X})} \leq \frac{\bar{x} - \mu}{\text{se}(\bar{X})}\right) = \Phi\,(-3.121) = 0.0009$$

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

Chapter 6

Interval estimation

Learning objectives

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

- do this
- do that
- and this

Readings

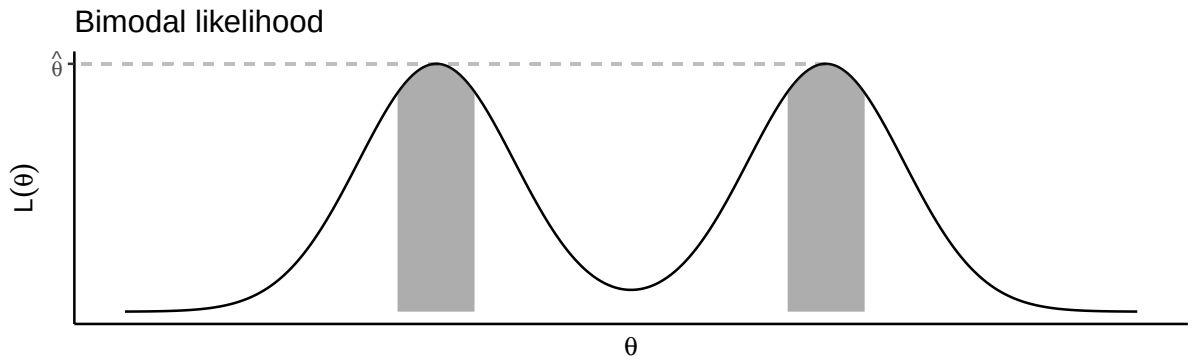
- Casella and Berger (2002)
  - Chapter 9, sections 9.1, 9.2 (9.2.1 and 9.2.2 only), 9.3 (9.3.1 only)
  - Chapter 10, section 10.4.
- Wasserman (2004)
  - Chapter 6, section 6.3.2
  - All of Chapter 8 (Bootstrap)
- Topics not covered here: Bayesian intervals, pivots based on cdfs, test-related optimality, 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 $\mu = 0$?

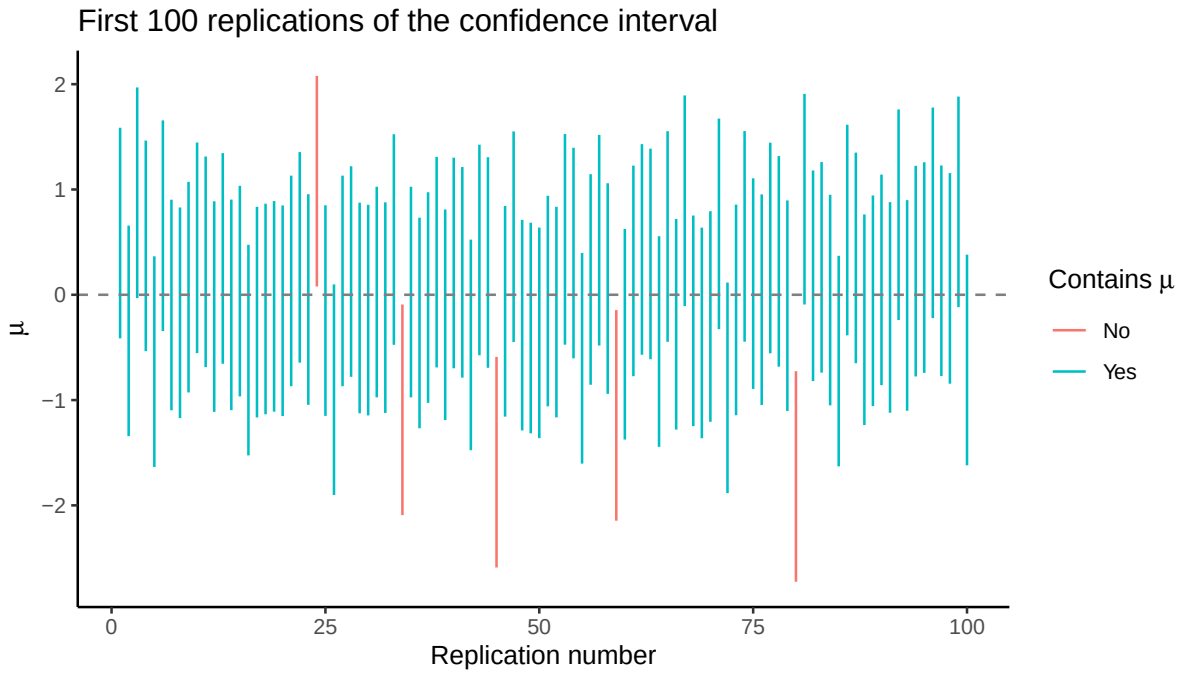
Here's the `R` code:

```
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

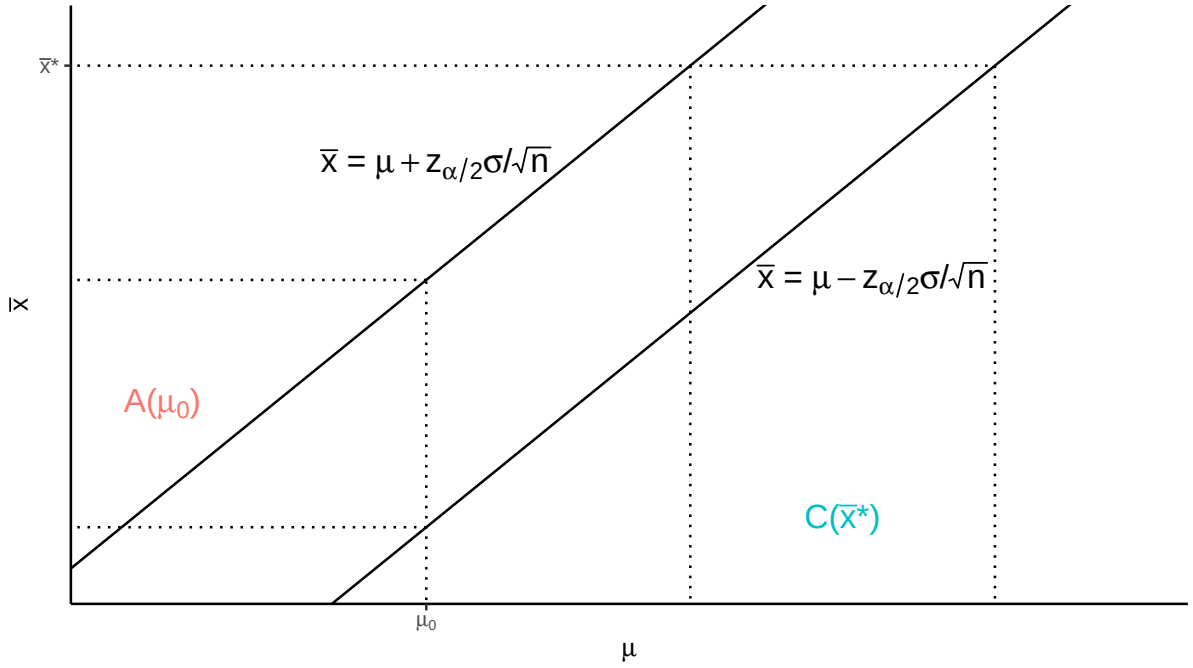
$$A(\mu_0) = \left\{x \;\middle|\; \mu_0 - z(\alpha/2)\frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu_0 + z(\alpha/2)\frac{\sigma}{\sqrt{n}}\right\}$$

- 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

$$x \in A(\mu_0) \Leftrightarrow \mu \in C(x).$$



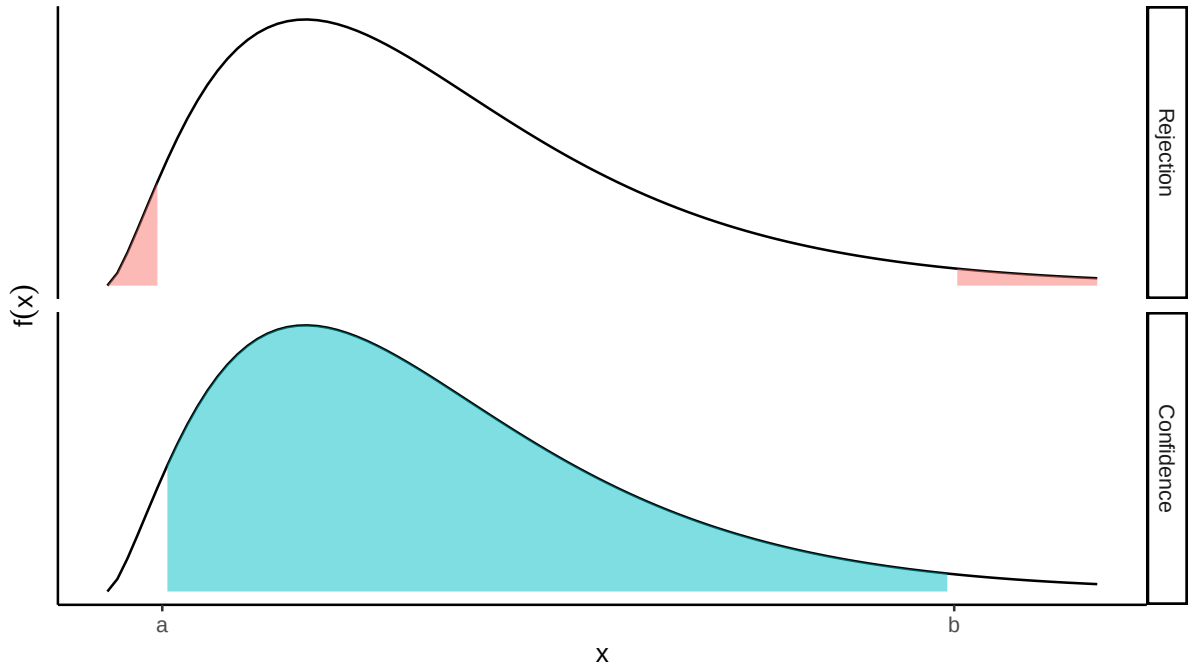
Example 6.6. In Ex. sheet 5, Q5 we looked at a hypothesis test for the variance of a normal distribution. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$ where both parameters are unknown. The LRT test rejects $H_0 : \sigma^2 = \sigma_0^2$ for samples in the rejection region

$$R = \left\{ x \;\middle|\; \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n\sigma_0^2} \leq k_1 \;\text{ or }\; \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n\sigma_0^2} \geq k_2 \right\}$$

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?

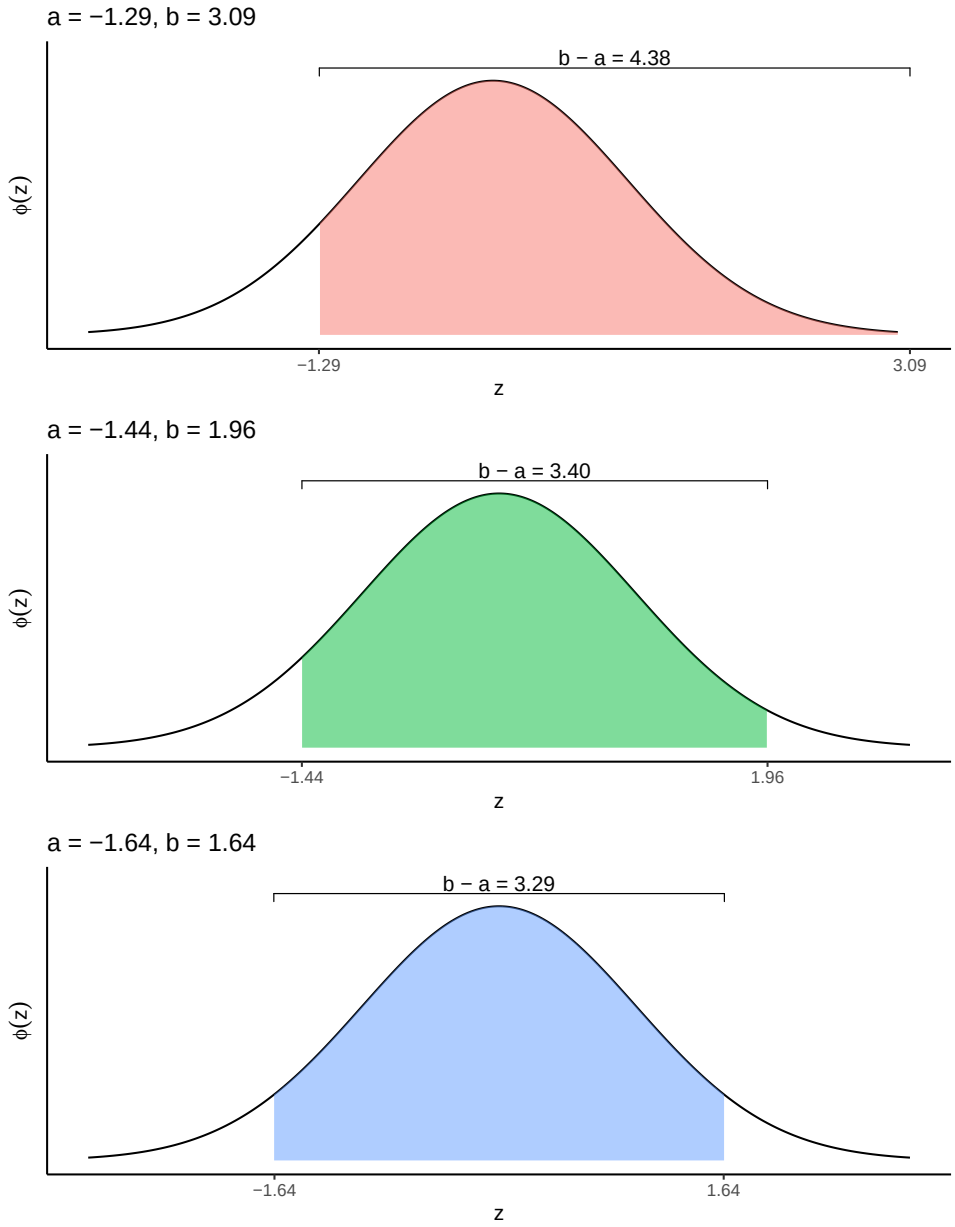
We desire a confidence set $C(X)$ which has

- 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 $\kappa = \mathrm{E}\left[(X-\mu)^4\right]/\sigma^4$

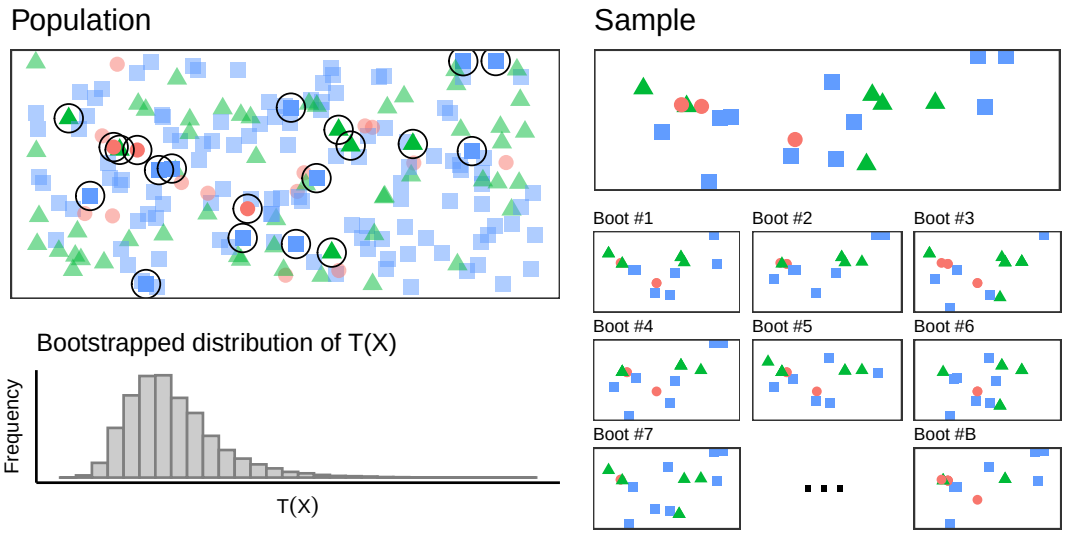


Figure 6.1: Bootstrap.

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

$$v_{boot} := \widehat{\text{Var}}_{\hat{F}_n}(T) = \frac{1}{B}\sum_{k=1}^{B}(T_k^* - \bar{T}_B)^2$$

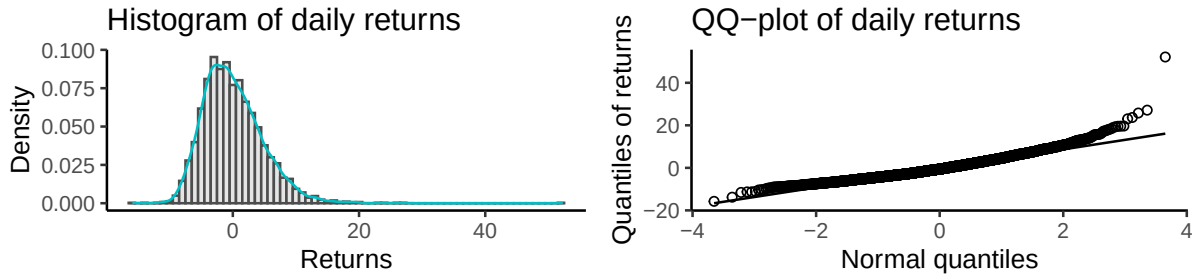
where $\bar{T}_B = B^{-1}\sum_{k=1}^{B} T_k^*$.

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

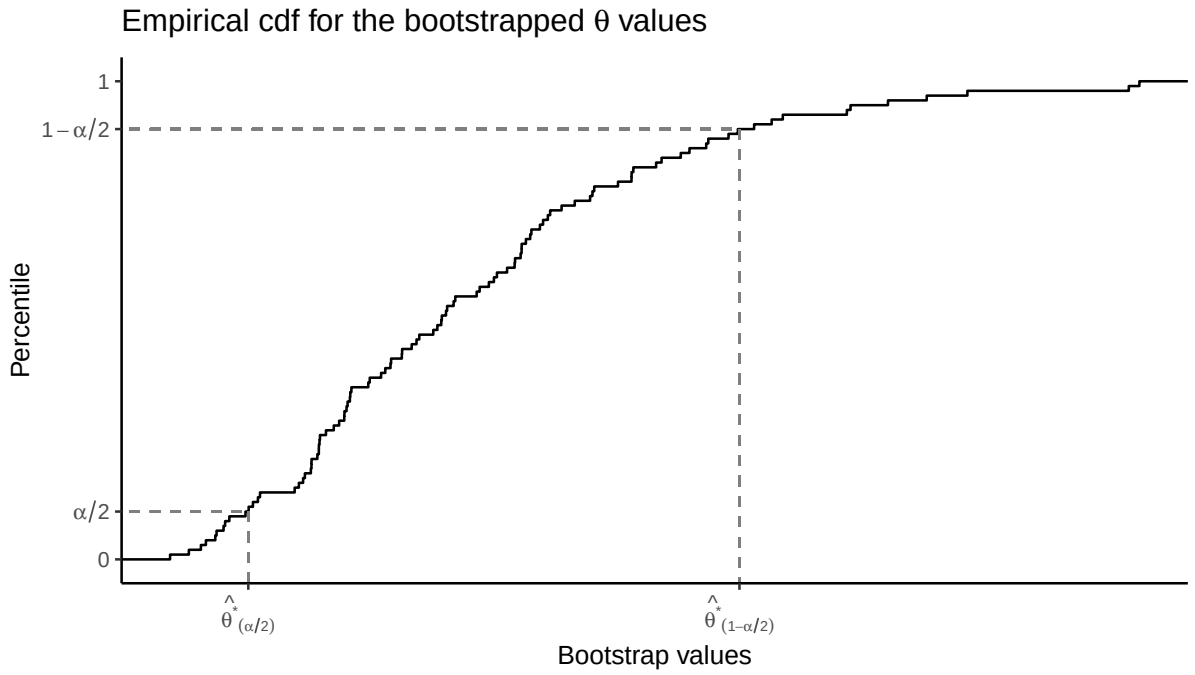
$$\hat{\theta}_{(1)}^*, \hat{\theta}_{(2)}^*, \dots, \hat{\theta}_{(B)}^*.$$

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
```

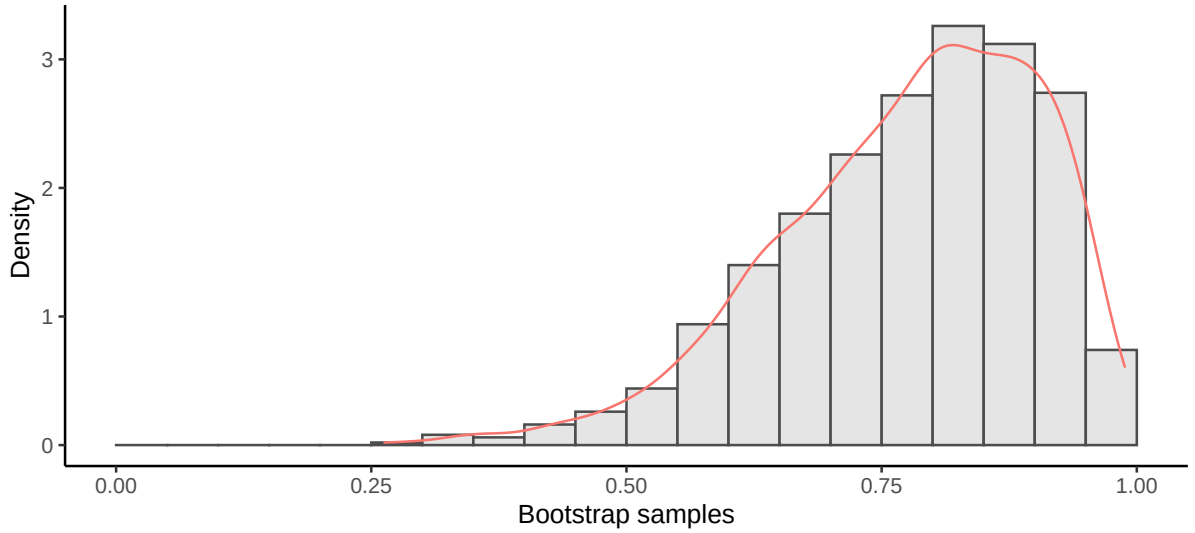
```
# percentile interval
a <- as.numeric(quantile(rhostar, probs = 0.025))
b <- as.numeric(quantile(rhostar, probs = 0.975))
c(a, b)
```

```
## [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 $\hat{\rho}^*$ (a bit skewed).



Appendix A

Exam tips