

MAT220: Probability and Statistics

Lakeland University Japan

D. O'Connell

Contents

I	Descriptive Statistics	5
1	Concepts in Statistics	6
1.1	What is Statistics?	7
1.2	Types of Data	8
1.3	Sampling as a Concept	12
1.4	Sampling Methods	14
1.5	When Data Misleads Us	18
2	Visualizing Data	22
2.1	Data Visualization as Presentation	23
2.2	Representing Quantitative Data	24
2.3	Representing Categorical Data	29
2.4	Representing Time-Dependent Data	32
2.5	Representing Paired Data	34
2.6	Case Study: Konbinis in Japan	37
3	Central Tendency and Variation	43
3.1	Central Tendency	44
3.2	Variation	52
3.3	Box and Whisker Plots	59
II	Probability Theory	61
4	Mathematical Foundations of Probability	62
4.1	Set Theory	63
4.2	Operations on Sets	68

4.3	Counting Configurations	72
4.4	Counting Permutations	74
4.5	Counting Combinations	78
5	Introduction to Probability Theory	82
5.1	What is Probability?	83
5.2	The Rules of Probability	88
5.3	Compound Probabilities	90
6	More on Probability	102
6.1	Problems Involving Counting	103
6.2	Beyond Classical Probability	111
6.3	Discrete Probability Distributions	115
7	Conditional Probability	125
7.1	Conditional Probability in General	126
7.2	Decision Trees	131
7.3	Famous Examples	138
8	The Binomial Distribution	143
8.1	A brief review: counting and distributions	144
8.2	Example: binomial experiments	149
8.3	The general theory	153
8.4	Exercises	157
III	The Normal Distribution	162
10	The Normal Distribution	163
10.1	What is a Normal Distribution?	164
10.2	The Standard Normal Distribution	167
10.3	Areas and Probability	169
10.4	Appendix: z -scores table	176
11	The Central Limit Theorem	178
11.1	A Brief Review of the Normal Distribution	178
11.2	Sampling Distributions	182

11.3	The Central Limit Theorem	183
11.4	Normal Approximation to the Binomial Distribution	187
11.5	Worked Example: Vintage Clothes	192
12	An Introduction to Estimation	194
12.1	A Review of the Central Limit Theorem	195
12.2	The Idea of Estimation	196
12.3	Constructing Confidence Intervals	200
12.4	The General Formula for Confidence Intervals	203
13	More on Estimation	206
13.1	A Review of Estimation	207
13.2	Estimation with Unknown σ	208
13.3	Estimating a Binomial Proportion	213
13.4	Worked Example: Basketball Scouting	216
IV	Hypothesis Testing	219
14	Introduction to Hypothesis Testing	220
14.1	Introduction	221
14.2	Hypothesis Testing	223
14.3	A Sample Test: The Two Coins	228
15	Hypothesis Testing the Mean	231
15.1	A Review of Hypothesis Testing	232
15.2	Testing μ with Known σ	234
15.3	Testing μ with Unknown σ	237
15.4	Testing the Binomial Success Probability p	242
15.5	Summary	245
15.6	Appendix: Tables for Student's t -Distributions	246
16	Hypothesis Tests of Paired Data	247
16.1	A Review of Hypothesis Tests for the Population Mean	248
16.2	What are Paired Populations?	249
16.3	The Sampling Distribution for Differences of Pairs	250

16.4 Hypothesis Testing Paired Data	254
16.5 Example: Movie Ratings	256
17 Hypothesis Tests of Independent Data	258
17.1 A Review of Paired Data	259
17.2 Independent Populations	260
17.3 Hypothesis Tests for Independent Means	263
17.4 Testing Two Binomial Probabilities	264
17.5 Example: Orange Skittles	266
17.6 A Summary of Hypothesis Tests	268
V Correlation and Regression	270
18 Linear Correlation	271
18.1 What is Correlation?	272
18.2 The Correlation Coefficient	276
18.3 Limitations of Correlation	284
19 Linear Regression	287
19.1 Linear Regression	288
19.2 The Least-Squares Method	289
19.3 Using the Regression Line for Predictions	295

Part I

Descriptive Statistics

Lecture 1

Concepts in Statistics

This is the first lecture of MAT220, so our goal is to establish the basic language of statistics. We begin by asking what statistics *is*, and why it is not merely a collection of formulas. We then discuss types of data, including the four levels of measurement: nominal, ordinal, interval and ratio. After that, we introduce sampling, compare several sampling methods, and explain why different methods can lead to different conclusions. Finally, we discuss the main ways in which data can mislead us: sampling problems, non-sampling errors and mistakes in inference.

1.1 What is Statistics?

Statistics is the mathematical study of *data*. We can think of data as information collected from something in the real world. The purpose of statistics is to collect, organise, analyse, interpret and present this information so that we can make better decisions.

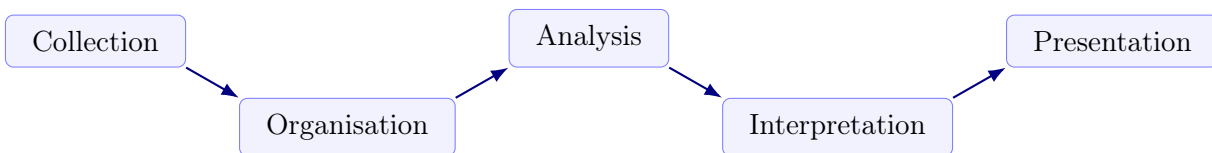
Statistics is used throughout science, business, government, medicine, sport, education and technology. Any time somebody tries to understand a complicated real-world system using data, they are using statistical thinking.

Definition

Statistics is the study of data, especially the collection, organisation, analysis, interpretation and presentation of data.

1.1.1 The workflow of a statistical study

A good statistical study mixes real-world context with mathematical techniques in order to discover new information about a system. One basic format for a statistical study is the following workflow:



Here:

- **Collection** involves measuring some system in the real world and recording data from it.
- **Organisation** involves rearranging the data into a useful format.
- **Analysis** involves using informed mathematics to describe useful features of the data set.
- **Interpretation** involves explaining what the results mean within a real-world context.
- **Presentation** involves communicating the results clearly, concisely and honestly.

Notice that the above workflow is not simply a collection of mathematical tools. Instead, a statistical study is an informed blend of mathematical techniques and real-world, practical considerations.

1.1.2 Why statistics can feel boring

Statistics can often feel boring and dry because it is formulaic. In fact, lots of the formulas used in statistics have a deep mathematical theory behind them, but practical statistics rarely uses that theory directly. Evaluated from the perspective of mathematical sophistication, statistics can often feel boring because it is so simple, and many of the most common formulas use only basic arithmetic.

In fact, statistics is also very interesting for exactly the same reason: even from the most basic mathematical operations, statistics creates a bunch of clever formulas that can be used to describe overall structures in data. Statistics is interesting in that it is such a powerful and useful tool that

comes from such humble beginnings. The practical use of statistics cannot be overstated: the world is made up of information, and many properties about you and society can be converted into data and analysed.

1.1.3 Descriptive and inferential statistics

There are two main pillars of statistics: *descriptive* statistics and *inferential* statistics.

- **Descriptive statistics** is the art of describing the data that we already have. This includes tables, charts, averages, percentages, variation and other summaries.
- **Inferential statistics** is the science of making conclusions about a larger group based on a small amount of data. This includes confidence intervals, hypothesis testing and statistical modelling.

Example. Suppose we measure the commute times of 50 students.

- If we calculate the average commute time of those 50 students, we are doing **descriptive statistics**.
- If we use those 50 students to estimate the average commute time of all students at the university, we are doing **inferential statistics**.

1.2 Types of Data

1.2.1 Data acquisition

In reality, we do not always have access to a complete data set, and instead, usually, we only have access to the data that we are able to collect. Inferential statistics builds tools that help us figure out things about a larger group without perfectly knowing everything about it. These are extremely powerful and insightful methods; however, *our data is only as good as our method of acquiring it*. If we happen to have collected bad data in the first place, then no amount of fancy mathematics will be helpful in the analysis. Unwanted effects can accidentally be built into a statistical analysis, and this can cause incorrect conclusions to be made. It is our job to correctly identify these potential effects and to minimise them.

Important point

Bad data cannot usually be saved by good mathematics. If the data was collected in a confused, biased or careless way, then the final statistical conclusion may be unreliable even if the mathematics is technically correct.

1.2.2 Quantitative and categorical data

Data can roughly be divided into two broad types:

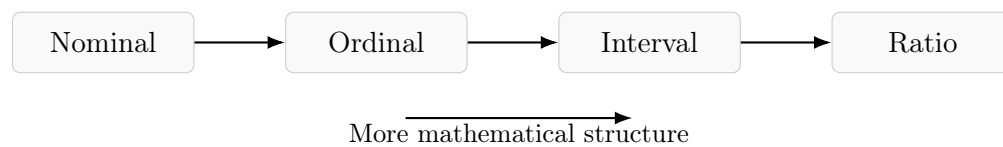
- **Quantitative data** is numerical data.

- **Categorical data** is non-numerical data.

Quantitative data has numerical structure, so it can often be analysed using arithmetic operations. Categorical data is more like a list of labels, names or groups, and therefore has much less mathematical structure to play around with.

1.2.3 Levels of measurement

There are many different types of data that we may collect in the real world. So, depending on the type of data we have access to, we may or may not be able to perform certain mathematical operations on it in a meaningful way. One way to navigate this potentially messy subject is to use *levels of measurement*, introduced by Stevens in 1946. There are four basic levels of measurement:



Roughly speaking, the more mathematical structure we have, the more operations that we can perform on our data. We will now review these levels of measurement one-by-one.

What is the point?

Levels of measurement help us be direct and explicit about our data. By classifying the type of data we have, we tell the audience what kinds of mathematical operations are meaningful and what kinds of claims would be inappropriate.

Nominal scales

Nominal data consists of labels or categories with no intrinsic order; it is *categorical*. There is no numerical structure to this level of measurement, so all we can do is test whether two observations are the same or different, and count their frequency of occurrence. But, there is no prescribed order or arithmetic that we have access to.

Nominal scale

Nominal data consists of qualitative categories with no natural order. Examples include:

- favourite colours,
- blood type,
- nationality,
- movie genre,
- train line used for commuting.

For nominal data, it makes sense to test equality and count frequencies.

Ordinal scales

The word “ordinal” means something similar to the word “order”. As such, ordinal data is data that has a meaningful order or ranking. However, the gaps between ranks are not assumed to be equal or measurable in any sense.

Ordinal scale

Ordinal data consists of categories or values with a meaningful order, but unclear distances between neighbouring values.

Examples:

- ranking of competitors,
- pain scale,
- satisfaction responses such as “very dissatisfied”, “dissatisfied”, “neutral”, “satisfied”, “very satisfied”,
- class standing such as first year, second year, third year and fourth year.

For ordinal data, medians and percentiles are often meaningful. Means can sometimes be used in practice, but one should be careful because the distance between neighbouring categories may not be equal.

Interval scales

Interval data has a meaningful order and equal differences between values. However, zero does not represent absolute absence of the quantity. Because of this, ratio statements such as “twice as much” are not generally meaningful.

Interval scale

Interval data has meaningful order and meaningful differences, but no genuine zero.

Examples:

- temperature measured in Celsius,
- year of birth.

For example, 20°C is not “twice as hot” as 10°C in a physically meaningful sense, because 0°C does not mean the absence of temperature. However, the difference between 20°C and 30°C is the same size as the difference between 30°C and 40°C.

Ratio scales

Ratio data has order, equal intervals and a genuine zero representing absence of the quantity. Because of that, both differences and ratios are meaningful.

Ratio scale

Ratio data has meaningful order, meaningful differences and a genuine zero.

Examples:

- weight,
- height,
- length,
- time duration,
- number of students in a classroom,
- salary.

For example, a commute time of 60 minutes is genuinely twice as long as a commute time of 30 minutes, because there is an absolute zero of 0 minutes that can be used as a baseline. This kind of comparison makes sense because 0 minutes means no time has passed.

Levels of measurement: summary

Level	Main structure	Allowed comparisons	Example
Nominal	Categories only	same or different	blood type
Ordinal	Categories with order	greater or less than	satisfaction rating
Interval	Equal differences, no true zero	differences	Celsius temperature
Ratio	Equal differences and true zero	differences and ratios	height

Exercise

Classify each variable as nominal, ordinal, interval or ratio.

- (a) A student's favourite convenience store.
- (b) A satisfaction rating from 1 to 5.
- (c) A temperature measured in Celsius.
- (d) The time taken to commute to campus.

Solution

- (a) Nominal, because the responses are category labels.
- (b) Ordinal, because the responses have an order, but the gaps between ratings may not be equal.
- (c) Interval, because differences are meaningful but zero Celsius is not a genuine absence of temperature.
- (d) Ratio, because time duration has a genuine zero and ratios are meaningful.

1.3 Sampling as a Concept

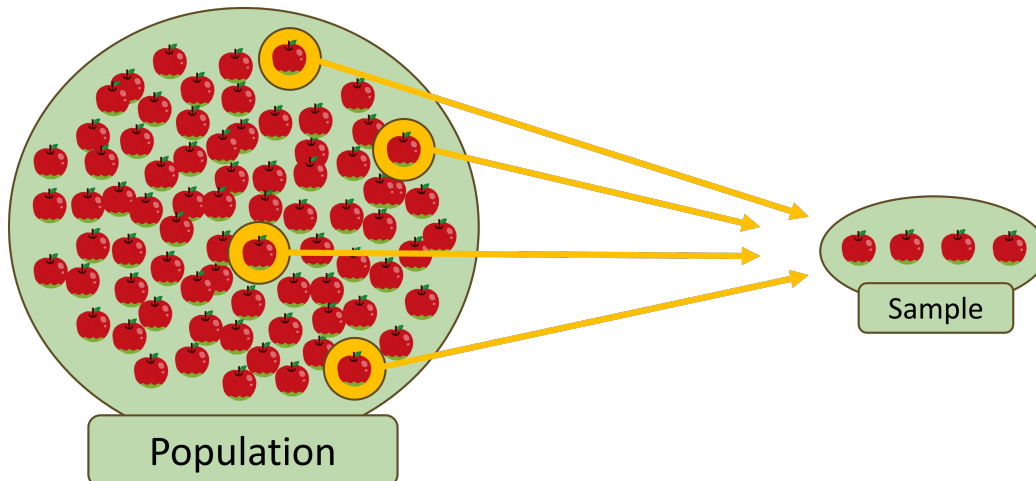
1.3.1 Core definitions

As previously mentioned, when studying the real world it is very often the case that the collection of things we want to study is simply too large to work with. So, in practice we will often just take a smaller portion and study that instead. This realistic, practical consideration motivates the following terminology.

Definitions

- **Population:** the complete group of people or things you want to learn about.
- **Sample:** the smaller group you actually collect data from, chosen from the population.
- **Population parameter:** a true number that describes the whole population, even if you do not know it exactly.
- **Sample statistic:** a number calculated from the sample, often used to estimate a population parameter.

In other words, a population is the collection we actually want to study. The facts that we want to discover are called population parameters, but often we can't discover these directly because the population is too big. So, instead we take a subset, which is just a smaller group selected from the overall population. We can then perform easier calculations on this smaller group, and this gives us some partial information known as a sample statistic. Here is a visual example of a sample:



1.3.2 Worked example: Average Salaries in Tokyo

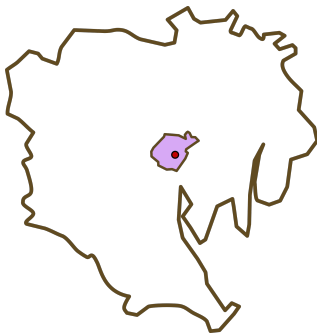
The process of obtaining a sample is known as *sampling* from the population. Sampling is an extremely important part of statistics, and large parts of our course will be dedicated to the relationship between samples and the overall population. However, we have to be very careful when sampling, because we may accidentally introduce some effects that are not representative of the true population. We will now illustrate this idea with an example.

Suppose that Jenny is an intern at a financial firm in Chiyoda, central Tokyo. She has been tasked by her boss to find out what the average salary of residents of Tokyo is. So, Jenny needs to perform a statistical study in order to find this out. Since Jenny does not work for the municipal government, she does not have access to complete salary data about the population of Tokyo. Instead, she decides to take a sample of 30 residents of Tokyo, and then take the average salary of these people.

When performing her statistical study, Jenny realises that her results will depend heavily on the people that she asks. For instance, the sample will be different based on:

- the time of day that she takes the sample,
- the location where she takes her survey,
- the age of the people involved,

and so on. Below are four different possible ways that Jenny might take a sample of 30 people, together with the problems inherent to each.

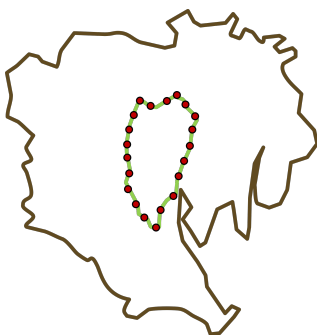
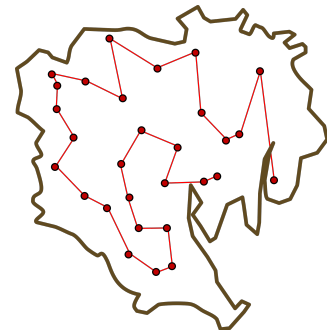


Method 1: Selecting 30 participants from her office in Chiyoda.

Lack of diversity: it is certainly convenient to take a sample of people from her local neighbourhood. However, her numbers may not represent Tokyo overall. Some neighbourhoods are richer or poorer, older or younger, or have different types of jobs. So even if she asks 30 people, she may accidentally measure the character of that neighbourhood rather than Tokyo as a whole. In the case of Chiyoda, her sample is likely to overestimate average salary.

Method 2: Randomly selecting 30 participants from anywhere in Tokyo.

Logistical difficulty: although it may feel unbiased to sample people randomly from across the city, actually doing so would be expensive and time-consuming. Not only would it be difficult to travel to the various neighbourhoods she has randomly selected, she would also run into other potential biases depending on how she selects a representative to ask. Nevertheless, a plan that looks unbiased on paper may fail in practice because of resource constraints.



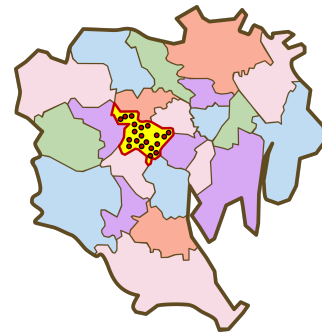
Method 3: Travelling the Yamanote line, stopping at each of the 30 stations and asking one person at each station.

Lack of accessibility: sampling “one person per station” may include tourists and outsiders who do not live in Tokyo. It can also accidentally introduce a bias towards tourist stations, commuter hubs or business districts, which can skew the sample toward certain kinds of workers. Put differently, this sample may

have a bias towards a particular type of Tokyo resident: those who work or travel around the Yamanote line.

Method 4: Randomly selecting one of the 23 wards of Tokyo, and then randomly selecting 30 participants from within that ward.

Lack of diversity: although it is more convenient than traveling throughout all of Tokyo, and less biased than Jenny’s local neighbourhood of Chiyoda, there will *still* be a lack of diversity in this type of sample. Randomness does not guarantee that the diversity of a population is fully represented in small samples. Instead, it only makes the first stage of selection random.



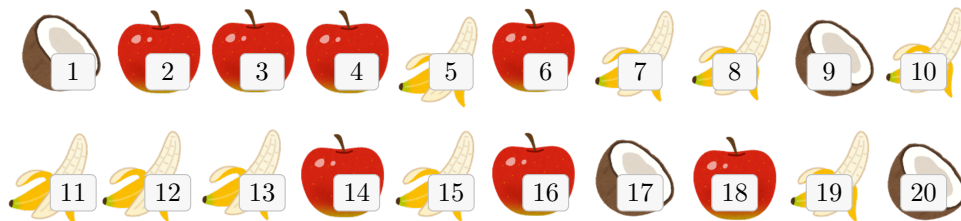
So, what is the best sampling method here? In fact, it depends on what Jenny is trying to do. In practice, there is often no single “best method”. Instead, sampling is often a balancing act between resources and error, and sometimes we must compromise. Ideally, these biases ought to be understood by the researcher running the study, and this should be made as explicit as possible when communicating results.

Sampling in practice

Sampling is often a balancing act between resources and error. The most accurate method of data collection is usually more expensive, more difficult or more time-consuming. Sometimes we need to make a compromise, but we should make the compromise explicit when presenting the results.

1.4 Sampling Methods

Motivated by Jenny in the previous example, we will now review some common sampling methods. Throughout this section, we will suppose that we have a population consisting of 20 pieces of fruit: 7 apples, 9 bananas and 4 coconuts.



Our goal is to choose a sample of size 5 from this population using different sampling methods.

1.4.1 Simple random sample

Definition: Simple Random Sample

A **simple random sample** is obtained so that every possible sample of the given size is equally likely.

For example, from our fruit population, one possible simple random sample of size 5 is:

Sample 5 9 10 12 18

Another possible simple random sample of size 5 is:

Sample 2 4 8 14 17

In a simple random sample, we do not deliberately control how many apples, bananas and coconuts appear in the sample. The sample is fair in its selection rule, but it may still be unbalanced by chance.

1.4.2 Systematic sample

Definition: Systematic Sample

A **systematic sample** is obtained by choosing a random starting point and then selecting every k th member of the population after that.

For example, if we start at 1 and choose every 4th member, we obtain:

Sample 1 5 9 13 17

If we start at 3 and choose every 4th member, we obtain:

Sample 3 7 11 15 19

Systematic sampling can be convenient, especially when the population is already arranged in a list. However, it can be dangerous if the list itself has a repeating pattern. For example, if every 4th item in a factory list comes from the same machine, then choosing every 4th item may accidentally sample only one machine.

1.4.3 Stratified sample

Definition: Stratified Sample

A **stratified sample** is obtained by first dividing the population into groups, called **strata**, based on a relevant characteristic, and then randomly sampling from each stratum.

In our fruit example, the natural strata are the 3 types of fruit:

apples, bananas, coconuts.

A stratified sample might deliberately choose some apples, some bananas and some coconuts, so that the sample represents all three fruit types.

Example stratified sample of size 5:

Sample 2 6 9 14 19

Here we have sampled from each fruit type. The goal is not just randomness, but representation across important groups.

Stratified sampling is useful when the population has important subgroups and we want to make sure that each subgroup is represented.

1.4.4 Cluster sample

Definition: Cluster Sample

A **cluster sample** is obtained by dividing the population into clusters, usually based on natural or convenient groupings, and then randomly selecting one or more whole clusters.

For example, suppose our list of fruit is divided into clusters of size 5:

$\{1, 2, 3, 4, 5\}, \quad \{6, 7, 8, 9, 10\}, \quad \{11, 12, 13, 14, 15\}, \quad \{16, 17, 18, 19, 20\}.$

If we then randomly choose the second cluster, our sample is:

Sample 6 7 8 9 10

If we randomly choose the third cluster, then our sample is:

Sample 11 12 13 14 15

Cluster sampling is useful when it is expensive or difficult to sample individuals scattered across the whole population. For example, a researcher might randomly choose several classrooms and survey every student in those classrooms.

1.4.5 A common confusion: stratified versus cluster sampling

Stratified sampling and cluster sampling can feel similar because they both divide the population into groups. The difference is the role that the groups play.

Method	How groups are formed	How the sample is chosen
Stratified sampling	Groups are formed using important characteristics, such as fruit type, gender, year level or ward.	Randomly sample from every stratum.
Cluster sampling	Groups are often natural or convenient clusters, such as classrooms, neighbourhoods or list blocks.	Randomly choose whole clusters.

In short: in **stratified sampling**, we sample *from every group*, but in **cluster sampling**, we sample *whole groups themselves*.

1.4.6 Other sampling methods

There are many other sampling methods. Some are useful in practice, but they often come with serious limitations.

Method	Basic idea
Convenience sampling	Sample whoever is easiest to reach.
Voluntary response sampling	People choose whether to participate.
Panel sampling	Follow the same group of people over time.
Haphazard sampling	Sample in an informal or careless way without a clear random rule.
Snowball sampling	Existing participants help recruit future participants.
Extreme-case sampling	Deliberately sample extreme cases, such as the smallest and largest values.

These methods are not automatically wrong, but they must be interpreted carefully. A convenience sample, for example, may be cheap and fast, but it may tell us more about who was easy to reach than about the full population.

1.4.7 Returning to Jenny's Tokyo study

We can now describe Jenny's methods using sampling language:

- Method 1: **Asking 30 people near her office in Chiyoda.** This is closest to a convenience sample. The main concern is that Chiyoda may not represent Tokyo.
- Method 2: **Randomly sampling 30 people from Tokyo.** This is closest to a simple random sample. The main concern is that it is accurate in principle, but difficult in practice.
- Method 3: **Asking one person at each Yamanote station.** This is closest to a route-based convenience sample. The main concern is that this may sample commuters and tourists rather than residents.

- Method 4: **Randomly picking one ward, then sampling inside it.** This is closest to a cluster-style sample. The main concern is that one ward may not represent all wards.

The vocabulary does not solve the practical problem by itself, but it helps us describe the strengths and weaknesses of the study clearly.

1.5 When Data Misleads Us

There are three main reasons why a statistical conclusion might fail.

Three ways a statistical conclusion can fail

1. **Sampling problems:** the sample does not honestly represent the population.
2. **Non-sampling errors:** the data itself is distorted, even if the sample was chosen well.
3. **Inferential errors:** the conclusion goes beyond what the data actually supports.

A good statistical study must deal with all three. It is not enough to have a large sample, and it is not enough to use the correct formula. The whole chain of reasoning must be trustworthy.

1.5.1 Sampling error and sampling bias

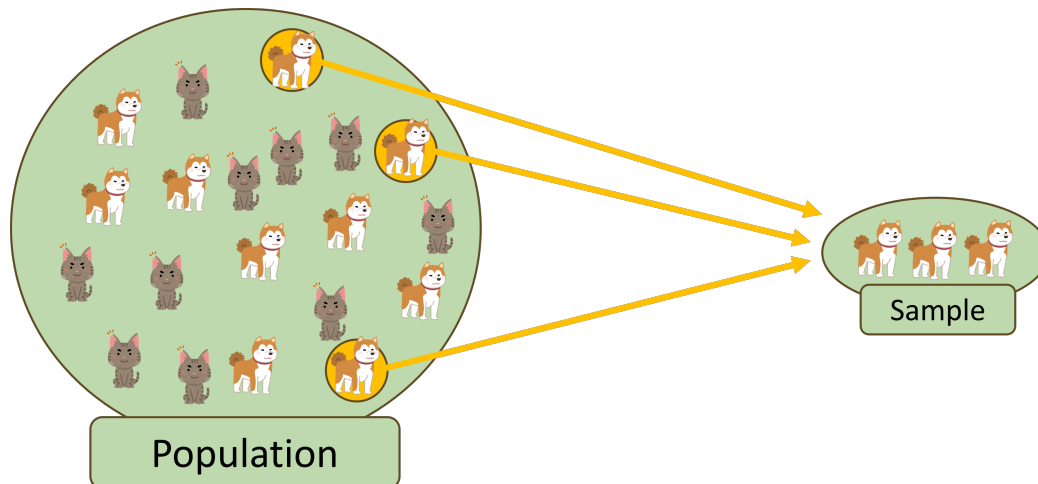
There are two important types of sampling problems.

Sampling error versus sampling bias

- **Sampling error** is a random mismatch between the sample and the population. This can happen even when the sampling method is fair.
- **Sampling bias** is systematic unfairness in the sampling method. The method itself makes some parts of the population more or less likely to be included.

Example: sampling error

Suppose the population contains cats and dogs in equal numbers. We take a simple random sample of size 3, and by chance all 3 selected animals are dogs.



This is a sampling error, because the method was fair but the particular sample was unlucky.

Example: sampling bias

Suppose Jenny only asks people in a wealthy neighbourhood such as Chiyoda about salary. This is not merely bad luck. The sampling method itself is systematically unfair because it overrepresents wealthy people.

This is a sampling bias. If Jenny repeats the same method many times, she will keep getting samples that tend to overestimate the average salary of Tokyo.

1.5.2 Non-sampling errors

Even if the sample is chosen well, the data itself can still be distorted. These distortions are called **non-sampling errors**.

Definition: Non-sampling Error

A **non-sampling error** is an error in a statistical study that is not caused by the random act of choosing a sample.

Common non-sampling errors include:

- unclear or leading questions,
- inaccurate measurements,
- dishonest answers,
- missing responses,
- recording mistakes,
- data entry mistakes,
- faulty equipment,

and so on. Importantly, these errors may occur even if the sampling method itself is reasonable.

1.5.3 Example: a biased survey question

Suppose you survey commuters in Tokyo and ask:

“How awful and terrible was your train ride this morning?”

This is a bad question because it assumes the train ride was awful and/or terrible, and therefore pushes the respondent toward a negative answer. A better version would be:

“How was your train ride this morning?”

Very bad Bad Fine Good Very good

Here the wording is much more neutral, and does not encourage a particular value judgement from the respondent.

1.5.4 Common ways conclusions become biased

Some biases are not just technical errors. Instead, they are sometimes failures in how we select, remember or interpret evidence. Here are three common examples of this phenomenon.

Survivorship bias

Survivorship bias happens when we only look at the examples that survived a selection process.

Example: “Architecture in the past was so beautiful, and modern buildings are so ugly.”

This may be misleading because many ugly old buildings were destroyed, while beautiful old buildings were preserved. The surviving buildings are not a random sample of past buildings.

Selection bias

Selection bias happens when the data is selected in a way that favours a certain conclusion.

Example: “I asked people in Kabukicho whether Tokyo is noisy and chaotic. They said yes. Therefore Tokyo is noisy and chaotic.”

Kabukicho is not a representative sample of Tokyo. The conclusion may tell us more about Kabukicho than about Tokyo as a whole.

Funding bias or conflict of interest

Funding bias can occur when a study is funded by a group that benefits from a particular conclusion.

For example, if an industry funds research about whether its own product is harmful, we should pay careful attention to the study design, the data, the analysis and the interpretation.

1.5.5 Inferential errors

Even when the sample is reasonable and the data is recorded accurately, the final conclusion can still be wrong. This happens when the inference goes beyond what the evidence supports.

Common inferential errors include:

- generalising from a small sample to a huge population,
- treating correlation as causation,
- ignoring lurking variables,
- ignoring confounding variables,
- using a statistic correctly but interpreting it incorrectly,
- making a conclusion stronger than the evidence justifies.

1.5.6 Correlation and causation

Two variables may be related without one causing the other. If two things happen together, this does not automatically mean that one is responsible for the other.

Example. Suppose that on very hot days in Tokyo, convenience stores sell more bottled water and more people report heat exhaustion.

It would be silly to conclude that buying bottled water causes heat exhaustion. A more reasonable explanation is that hot weather affects both variables.

In this example, bottled water sales and heat exhaustion are correlated, but the causal explanation involves another variable: temperature.

1.5.7 A checklist for trusting a statistical claim

Before trusting a statistical conclusion, ask the following questions.

1. **Population:** What population is the claim really about?
2. **Sample:** Who or what was actually sampled?
3. **Sampling method:** How were the sample members selected?
4. **Data quality:** Were the measurements, questions and records reliable?
5. **Missing data:** Who did not respond, and could that matter?
6. **Variables:** Are there lurking or confounding variables?
7. **Conclusion:** Does the conclusion follow from the data, or does it overreach?

Summary

A statistical claim is only as strong as the full chain behind it:

collection $\longrightarrow$ organisation $\longrightarrow$ analysis $\longrightarrow$ interpretation $\longrightarrow$ presentation.

If one part of the chain is weak, the final conclusion may be weak as well.

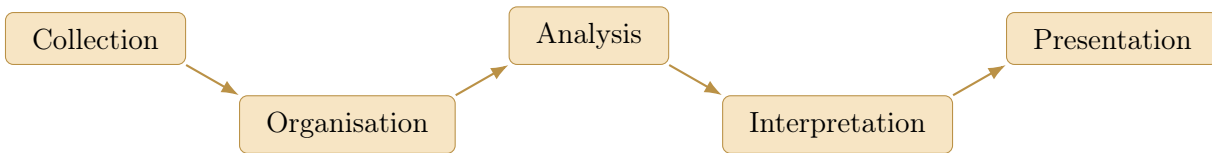
Lecture 2

Visualizing Data

In Lecture 1, we introduced the basic language of statistics: data, variables, populations, samples and the workflow of a statistical study. In this lecture, we will start our discussion of descriptive statistics by focusing on the visual presentation of data. At this stage in the course, we have not yet introduced any formulas. Instead, today we will study the idea of *communication*: how to turn a data set into a visual summary that others can quickly scan and understand. In keeping with the main point of the previous lecture, different data sets may have different structures, and therefore the “correct” presentation will depend on this structure.

2.1 Data Visualization as Presentation

Last lecture we saw that, on a basic level, a statistical study can be broken down into five key components:



We emphasized that this workflow is more like a *chain* of dependencies: if any of the boxes is flawed in its own way, then that is like “breaking the chain of validity”, and the results of the study will become unreliable. For instance, if we collect inaccurate or biased data in the first place, then it does not matter how good our statistical analysis is, the result will be as unreliable as our initial data. Or, if you collect accurate data in an unbiased way, but you apply the wrong formulas to it, then the overall study will be dubious.

Later on in the course, we will spend a lot of time discussing the analysis and interpretation stages of a statistical study. This is an extremely important element of statistical practice: not only do we need to know which formulas to correctly apply, but we also need to understand *why* these formulas are correct and *what they imply*. However, in today’s lecture we will assume that all of these deeper considerations are already established, and we will jump straight to the final step of the process: presentation.

Definition

Data visualization is the presentation of data using visual objects such as tables, graphs, charts, diagrams and maps. The goal is to make important structure in the data easier to see.

Of course, statistical studies have the potential to be extremely insightful. However, our insight is only ever as good as our *representation* of it: if data is presented poorly, then any underlying lessons or narratives may be impossible to identify. In other words, accurate data is useless unless we can clearly see the patterns inside of it.

2.1.1 Presentation is more art than science

Generally speaking, there is no perfect way to visualize a data set. As we will see, different data sets admit many different representations. Therefore, we as the presenters of the data have to make informed choices about which elements of our data we want to emphasize. Different presentations will have different effects on the audience, and therefore they will communicate slightly different messages. For this reason, data presentation is more of an art form than a strict science: there is no “formula” that we can apply to create the best presentation of data, and instead we need to use a mix of practical judgement and aesthetic taste. As we will see, often the right mix will depend on our data, our context, and our audience.

Important point

Data presentation is more art than science. A graph is not just a neutral container for information, instead it carries the choices made by the presenter along the way, and these choices will affect what the audience notices.

Of course, this does not mean that *anything* goes. As we will see, graphs can be genuinely bad, ugly, misleading or outright dishonest. Throughout today's lecture, we will slowly identify some key guidelines for good data presentation. Along the way, we will survey some basic representations of data. We will discuss:

1. quantitative data,
2. categorical data,
3. time-dependent data, and
4. paired data.

Finally, we will finish with a case study designed for this lecture, so that you can see several ways in which the same data can be presented.

2.2 Representing Quantitative Data

Recall that a data set is called “quantitative” whenever it consists of numerical measurements or counts for which arithmetic is meaningful. Because numerical data has a mathematical structure (order, arithmetic, and so on), it is often useful to keep this extra structure in mind when representing the data visually.

Generally, it is not particularly useful to present large amounts of quantitative data directly. To illustrate this, consider the following table of data, which lists the shoe sizes of 42 people, measured in centimetres:

Shoe sizes of 42 people

25	27	28	24	26	26
26	24	30	25	25	24
26	24	27	24	24	23
27	24	28	26	25	23
24	24	25	24	24	24
25	24	25	23	26	25
27	25	24	27	27	25

The table above simply displays the raw data. Although this presentation is very open and honest, it is not particularly insightful because we cannot easily see any useful patterns in the data. For instance, just from the numbers alone, we cannot quickly answer potentially useful questions such as:

- Which shoe size is most common?
- How are the shoe sizes roughly distributed in the group?
- Are there any outliers?

The answers to these questions are contained within the data *somewhere*, but they are hidden. It would be much better if these hidden structures were uncovered through a different representation.

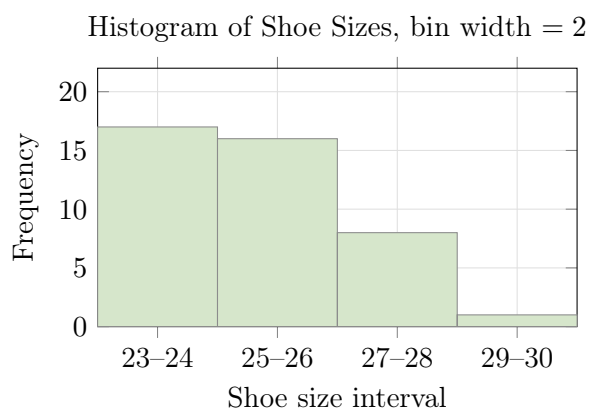
2.2.1 Frequency tables and histograms

When working with quantitative data, it is often useful to group the data into intervals of the same size. These intervals are often called *classes*, or more commonly *bins*. From this grouping, we can then create a frequency table, pictured below. For clarity, we redraw the same raw shoe-size data below next to its grouped frequency table.

Shoe sizes of 42 people							
25	27	28	24	26	26	→	
26	24	30	25	25	24		
26	24	27	24	24	23		
27	24	28	26	25	23		
24	24	25	24	24	24		
25	24	25	23	26	25		
27	25	24	27	27	25		
						Class Interval	Frequency
						23–24	17
						25–26	16
						27–28	8
						29–30	1

Here we have chosen to collect the 42 data points into bins of size 2, which are simply intervals of length 2 that run across the data. As you can see, the frequency table is already an improvement, since we can immediately see that lots of the data is clustered around the 23 to 26 range.

We will now use the information in the frequency table to plot a graph: the x -axis displays the intervals, and the y -axis displays the counts. Diagrams of this type are known as *histograms*.



Definition: Histogram

A **histogram** is a graph used to represent quantitative data by grouping values into intervals called bins. The height of each bar represents how many data values fall into that bin.

Histograms display the same basic information that can be found in the frequency table. However, the information is much more *accessible*, because we can quickly see how the frequencies relate to each other.

2.2.2 Relative frequency distributions

A frequency tells us how many observations fall into a given bin. Sometimes, however, we care more about proportions than the raw frequencies themselves. Fortunately, these proportions can be calculated with a simple procedure.

Definition: Relative Frequency

The **relative frequency** of a class is

$$\text{Relative Frequency} = \frac{\text{Bin Frequency}}{\text{Total Count}}.$$

It tells us the proportion of the data set that lies inside that bin.

For example, using our shoe-size data, the first class has frequency 17 and there are 42 people in total. So, its relative frequency is:

$$\frac{17}{42} \approx 0.405 = 40.5\%.$$

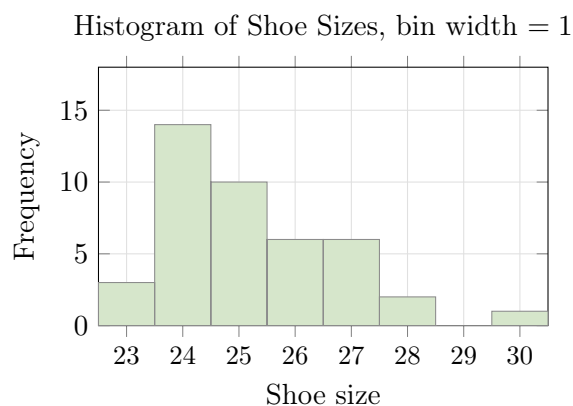
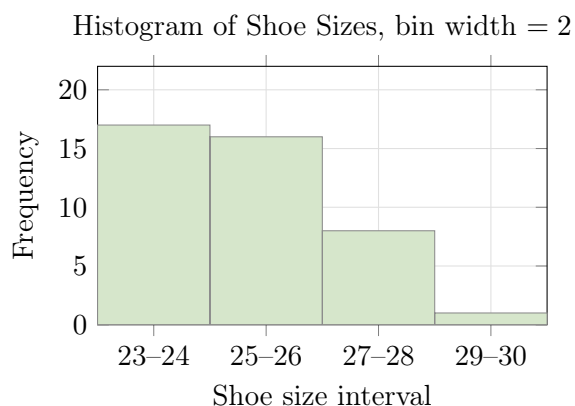
The table below is what is known as a *relative frequency table*.

Class Interval	Frequency	Relative Frequency
23–24	17	$17/42 \approx 40.5\%$
25–26	16	$16/42 \approx 38.1\%$
27–28	8	$8/42 \approx 19.0\%$
29–30	1	$1/42 \approx 2.4\%$

Relative frequency is useful when comparing data sets of different sizes. For example, if one class has 42 students and another class has 120 students, raw counts may not be directly comparable. Percentages often make the comparison clearer.

2.2.3 Abuse of histograms: changing bin width

Histograms are useful, but they are also very easy to manipulate. The main way to do this is by altering the *bin width*, since this may cause different shapes to occur as data shuffles around into different bins. For example, our shoe-size data can also be counted into bins of width 1, and this results in a very different-looking histogram:



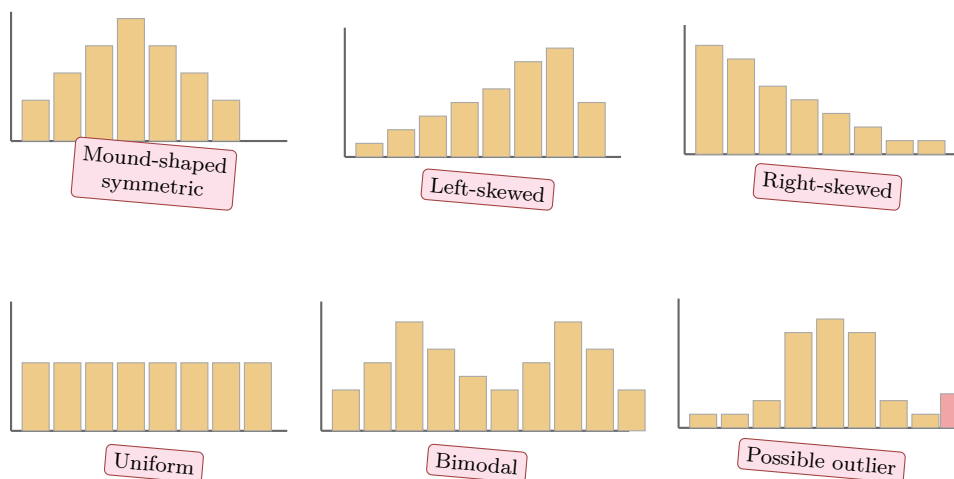
As you can see, the histogram on the right-hand side provides a higher resolution description of the data. Here we can see the distribution of each individual value. Notice how some of this data was obscured using a bin width of 2: in the case of the shoe sizes in the 23–24 bin, the first histogram merely tells us that there are 17 entries. However, the histogram on the right communicates that there are *many* more people with a shoe size 24 than there are of size 23. This extra information is entirely lost when we expand the bin size to width 2.

Important warning

Changing the bin width can change the visual story of a histogram. A responsible presenter should choose bins that reveal the structure of the data, rather than bins that exaggerate or hide a preferred conclusion.

2.2.4 Distribution shapes

Usefully, the overall shape of a histogram tells us something about the general way that the data is distributed. Some common histogram shapes are drawn below.



We will now describe these important shapes one-by-one.

Mound-shaped symmetric distributions

A mound-shaped symmetric distribution is one in which the histogram rises toward the middle and then falls away in a roughly balanced way. The two sides do not have to be perfectly identical, although in lots of cases they may well be. The main idea is that the centre is the most populated part of the distribution, and the data is spread out in both directions from that centre.

Mound-shaped symmetric distribution

A **mound-shaped symmetric distribution** is a distribution whose histogram has a central mound and whose left and right sides are approximately balanced.

Many naturally varying measurements, such as adult height, are often roughly mound-shaped, but not all randomly sampled data has this shape.

Skewed distributions

A skewed distribution is one where one tail is longer than the other, i.e. there is an inherent asymmetry. Confusingly, we name the skewness after the direction in which there is a *longer* tail, i.e. if the longer tail is on the left of the distribution, then we call that graph “left-skewed”.

Skewed distributions

A distribution is **skewed** when one side has a longer tail than the other.

- **Right-skewed:** the longer tail is on the right.
- **Left-skewed:** the longer tail is on the left.

The long tail often contains unusually large or unusually small observations. For example, income data is often right-skewed because most people earn ordinary amounts, while a small number of people earn extremely large amounts.

Uniform distributions

A uniform distribution is one in which every class has the same frequency, causing the histogram to look flat or rectangular.

Uniform distribution

A **uniform distribution** is a distribution where the classes have approximately equal frequencies.

Exact uniform distributions are uncommon in observational real-life data, but they are theoretically important, especially later during our discussions of probability theory. Uniform distributions represent a situation in which no interval is particularly special compared to any other.

Bimodal distributions

A bimodal distribution has two noticeable peaks. This can often indicate that the data set is mixing two different populations, or that there are two distinct “explanatory forces” acting on the population.

The idea is simple: the graph looks like two copies of the symmetric mound-shaped distribution. An individual peak may suggest a subgroup, mechanism, or repeated pattern in the data. In the case of a bimodal distribution there are *two* mounds, which can mean that there are *two* different properties or forces acting on the data.

Bimodal distribution

A **bimodal distribution** is a distribution with two main peaks, usually separated by at least one lower-frequency class.

For example, we could measure pedestrian traffic passing through a popular urban train station like Shinjuku. The (relative) frequency distribution will display two clear mounds: one centred around 8 a.m. and another centred around 6 p.m. This is because there are two peaks in traffic as all of the commuters go to and from work, respectively.

2.2.5 What histograms tell us

Histograms and relative-frequency histograms can show us how quantitative data is roughly distributed. By looking at histograms, we can often conclude:

- whether the data distribution is symmetric, skewed, uniform or bimodal,
- whether there are possible outliers,
- which data intervals contain the most data, and
- how spread out the data is.

2.3 Representing Categorical Data

Recall that categorical data consists of labels, names, groups or categories. Categorical data does not necessarily have any inherent mathematical structure that we can use to our advantage. For example, if our categories were blood types, then there is no sense in which we can “add” type A-negative and type O-positive together in some meaningful way. Since there is so little mathematical structure attached to categorical data, we are quite restricted in what we can do: the main operations we can perform are to count frequencies and to compare proportions. These two operations lead to two common representations of categorical data:

- **bar charts**, which represent frequencies or percentages, and
- **pie charts**, which represent proportions of a whole.

We will now discuss these separately.

2.3.1 Bar charts

A bar chart represents individual categories using separate bars in a graph, similar to a histogram. The length or height of each bar represents a value, usually a frequency or percentage.

Definition: Bar Chart

A **bar chart** is a graph for categorical data in which each category is represented by a separate bar. The height or length of the bar represents the frequency, percentage or value of that category.

Good bar charts follow three basic rules.

1. Bars can be vertical or horizontal.
2. Bars should have uniform width and uniform spacing.
3. The same measurement scale should be used for every bar.

The gaps between bars are important. A histogram represents intervals on a numerical scale. A bar chart represents distinct category labels, so the bars are usually visually separated.

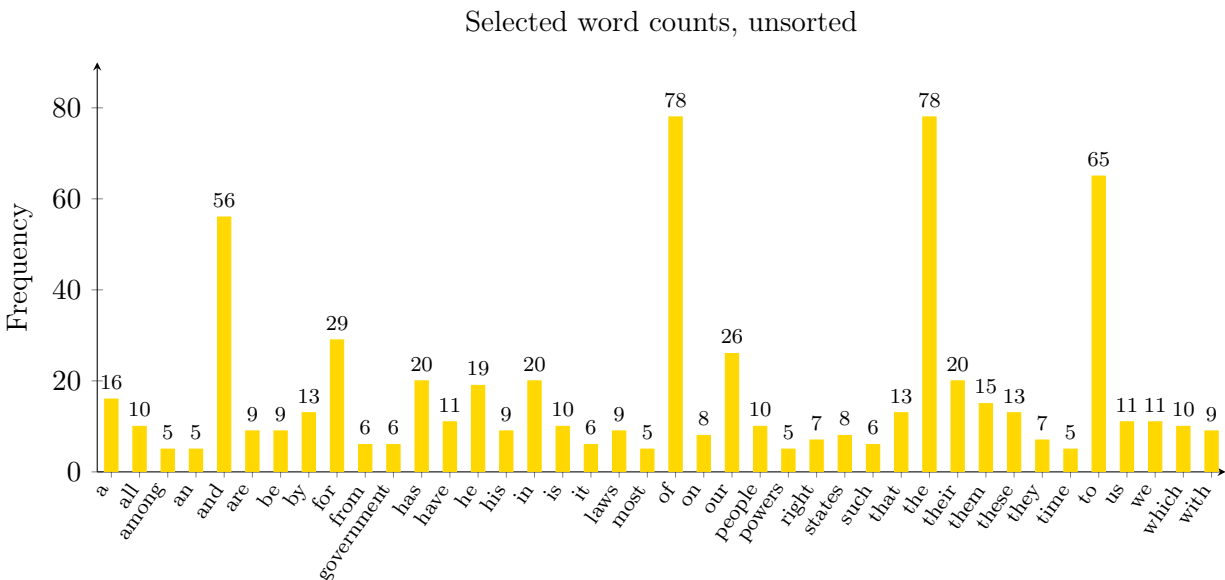
A Pareto chart is a bar chart whose bars are arranged in decreasing order. The most common categories are placed on the left, and the least common categories are placed on the right.

Definition: Pareto Chart

A **Pareto chart** is a bar chart in which the bars are arranged from largest to smallest. It is useful when we want to quickly identify the most frequent categories.

2.3.2 Example: The Declaration of Independence

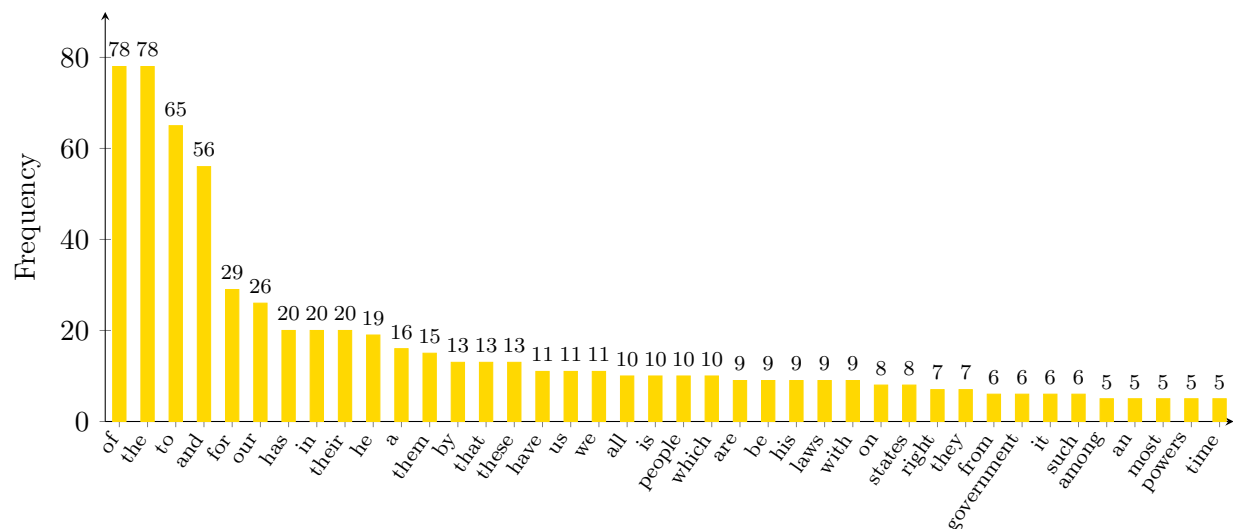
Suppose we count frequently occurring words in the Declaration of Independence. Using this particular word count, the main body contains 1333 words, so a bar chart that contains every word frequency would be quite cluttered. Instead, we will plot all words occurring at least 5 times as a bar chart:



As you can see, the graph is readable, but still quite crowded. Because there are so many individual

bars, it becomes hard to quickly answer questions like “which words are among the most frequent?” This can be improved by using a Pareto chart instead:

Selected word counts, Pareto order



The Pareto chart contains the same counts, but the ordering makes the dominant categories much easier to identify. In other words, the ordering and harmony of in the Pareto chart is easier on the eye.

2.3.3 Pie charts

A pie chart represents proportions as slices of a circle. To make a pie chart, we first calculate *proportions* by counting how often each category occurs, then dividing by the total count.

Definition: Proportion

A **proportion** is a part-to-whole comparison:

$$\text{Proportion} = \frac{\text{Frequency}}{\text{Total Count}}.$$

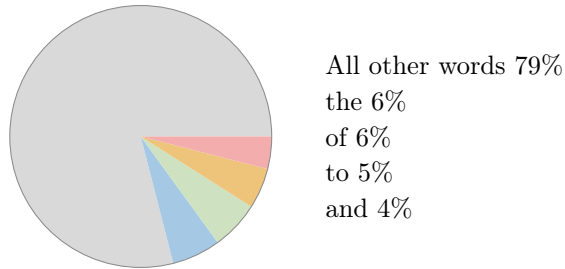
As an example, we use the data from the previous section. Observe that the word “the” occurs 78 times in the Declaration of Independence, out of a total of 1333 words. Therefore, the proportion of the word “the” is:

$$\frac{78}{1333} \approx 0.0585 = 5.85\% \approx 6\%,$$

that is, “the” appears approximately 6% of the time.

Definition: Pie Chart

A **pie chart** is a graph in which categories are represented as wedges of a circle. The size of each wedge represents the proportion of the total belonging to that category.



A readable pie chart has only a few substantial slices.

Important warning

Pie charts work best when there are only a few wedges. If a pie chart contains many tiny slices, the audience may not be able to view or compare the categories accurately.

Exercise

Suppose a class of 40 students is asked for their favourite convenience store. The results are:

7-Eleven : 18, FamilyMart : 12, Lawson : 8, Other : 2.

- (a) What proportion of students chose 7-Eleven?
- (b) Would a pie chart be reasonable for this data set?
- (c) Would a Pareto chart be reasonable for this data set?

Solution

- (a) The proportion is $18/40 = 0.45 = 45\%$.
- (b) Yes. There are only four categories, so a pie chart would still be readable.
- (c) Yes. A Pareto chart would also be reasonable, especially if we want to show the most popular store quickly.

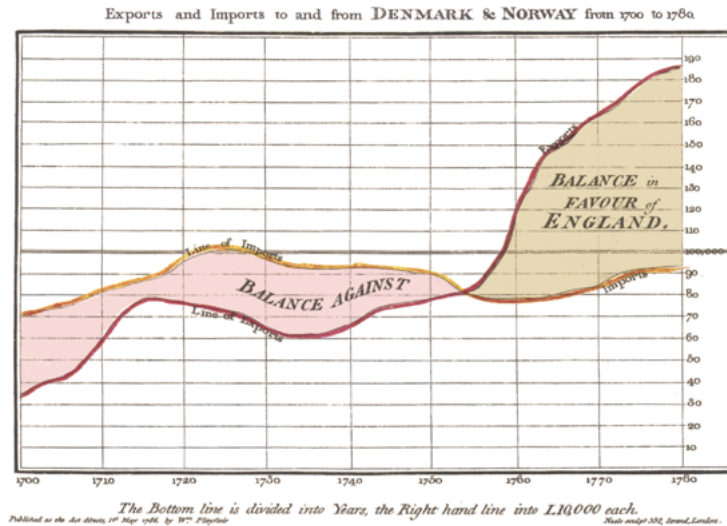
2.4 Representing Time-Dependent Data

Sometimes we collect data by repeatedly observing the same object. This allows us to track changes in a property over time. For example, we might measure our body weight every week, track the growing height of a sunflower every day, record the number of hours studied each evening, or follow the popularity of a word across many decades. This kind of data is called *time-dependent data* or *time-series data*.

Definition: Time-Series Graph

A **time-series graph** is a graph in which time-dependent data is plotted in chronological order. The horizontal axis usually represents time, and the vertical axis represents the quantity being tracked.

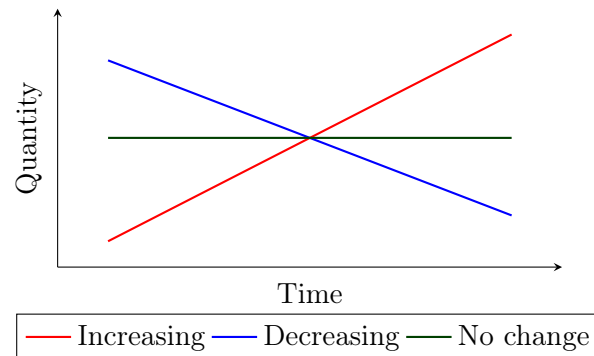
One of the earliest known examples of a statistical time-series graph was produced by William Playfair in the late eighteenth century:



As you can see, a time-series graph is extremely intuitive: we can quite clearly see the change in a quantity over time. We will now briefly review several useful types of change.

2.4.1 Linear change

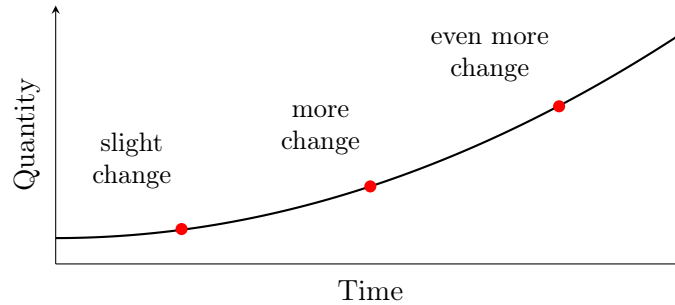
The simplest kind of change over time is *linear*. This means that the data set traces out a straight line when plotted on a time-series graph. Geometrically, the rate of change is represented by the *gradient* of the line:



The increasing line shows a quantity that grows over time. The decreasing line shows a quantity that falls over time. The flat line shows no change. In each case, the main visual idea is the same: the shape of the line tells us the direction and speed of change.

2.4.2 Non-linear change

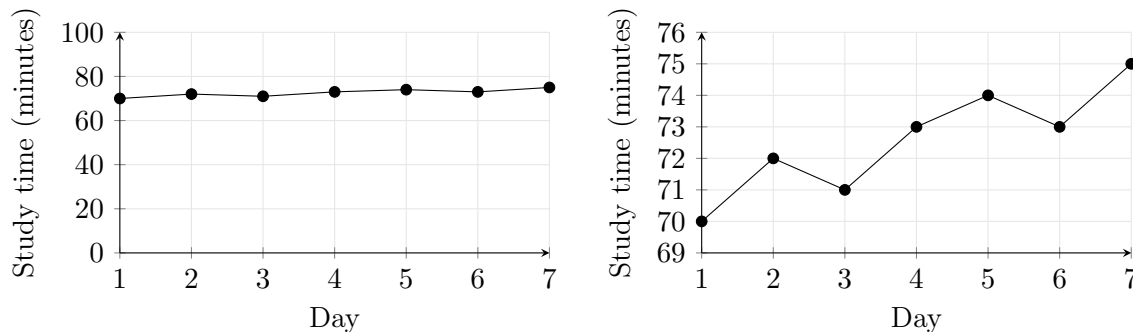
In most cases, change does not happen at a constant rate. Instead, the change itself can *speed up* or *slow down* depending on the process being measured. Consider the following:



In the graph above, we have an example of *acceleration*, since the rate of change itself is increasing with time.

2.4.3 Scale and narrative

Time-series graphs are especially sensitive to scale: changing the vertical axis can make a small change look dramatic, or make a large change look normal. To illustrate this, consider the two graphs below:



The two graphs above are both built from the same underlying data set. However, by adjusting the scale of the vertical axis we can make the data appear to be very different. The choice of scale subtly encodes *different narratives*: the first graph suggests that there is little change in the data, whereas the second graph suggests that the data is quite volatile. This observation motivates the following.

Important warning

Different scales can create different narratives. When reading a time-series graph, always check the axes before interpreting the story.

This is one of the most common ways that data visualizations mislead people. A graph may be technically correct, but still visually misleading if the scale is chosen to exaggerate the effect.

2.5 Representing Paired Data

Sometimes the data we collect naturally comes in pairs. This usually happens when we study the same object and collect different properties at the same time, or when we want to compare two properties of each observation. Examples would include:

- Left and right foot length for the same person.
- Travel time and distance travelled for students on their way to campus.
- The height and weight of people living in Tokyo.

Definition: Paired Data

Paired data occurs when each observation consists of two related measurements. We often study paired data to investigate whether there is a relationship between the two variables.

2.5.1 Example: commute time and distance

Consider the following data set of 12 students.

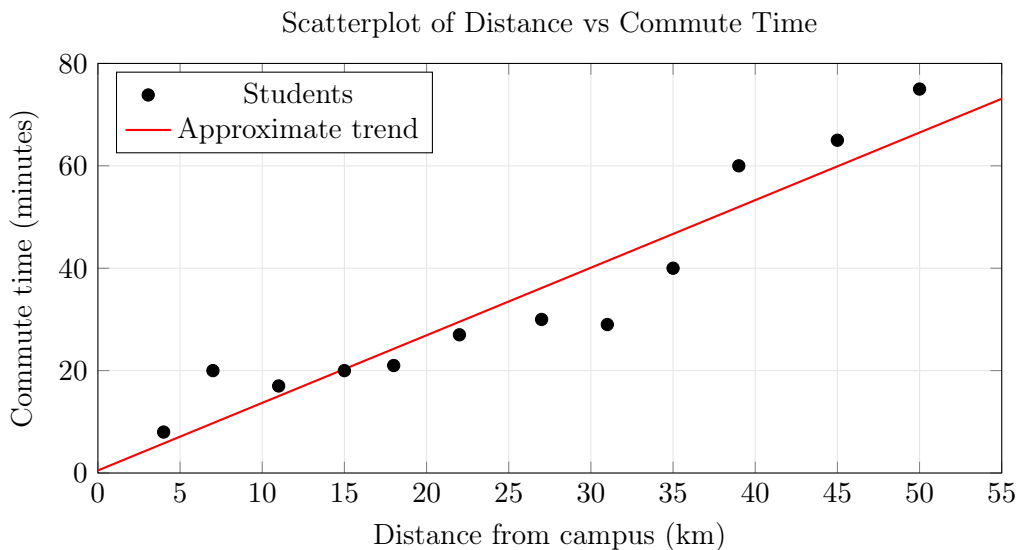
Student	Distance from campus (km)	Commute time (minutes)
1	4	8
2	7	20
3	11	17
4	15	20
5	18	21
6	22	27
7	27	30
8	31	29
9	35	40
10	39	60
11	45	65
12	50	75

This data set is naturally paired because each row contains two linked pieces of information about the same individual student. We can split this pair into two variables: the first variable is distance from campus, and the second variable is commute time. A *scatterplot* is the natural graph for this situation.

Definition: Scatterplot

A **scatterplot** is a graph for paired quantitative data. Each observation is represented by a point, where the horizontal coordinate gives one variable and the vertical coordinate gives the other variable.

To create the scatterplot, we plot one variable on the x -axis and the other on the y -axis:

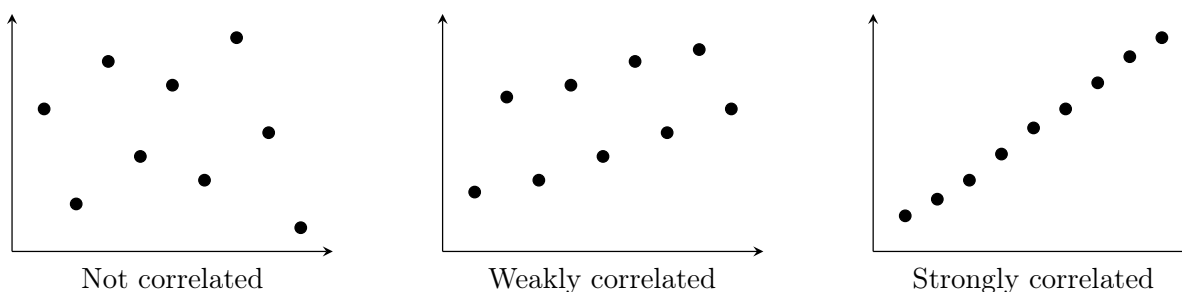


2.5.2 Correlation

The scatterplot drawn above shows a clear positive relationship between the two variables, meaning that students who live farther from campus tend to have longer commute times. Of course this is not a perfect relationship, since commute times generally depend on more than just distance alone. However, there is still a clear, general pattern between the two variables. When paired data shows a general pattern, we say that the variables are *correlated*.

Definition: Correlation

Correlation means that two variables tend to vary together. A scatterplot can show whether paired data is not correlated, weakly correlated or strongly correlated.



Later in the course, we will study some methods for how to describe correlation numerically. But, for now, it is enough simply to recognise the basic idea:

- If the points form no clear pattern, the variables may be uncorrelated.
- If the points roughly move in one direction but with a lot of scatter, the variables may be weakly correlated.
- If the points lie close to a line, the variables may be strongly linearly correlated. If they lie close to a curve, the variables may be strongly associated.

A negative correlation would show a downward trend instead.

In the commute-time example, we drew an approximate line through the data. The act of finding a best-fitting line is called **linear regression**. We will study this more formally later in the course.

Linear regression

Linear regression is a method for finding a line that summarises the relationship between two quantitative variables. In this lecture, we only use it visually. The formal method will come later.

2.5.3 Correlation is not causation

As we discussed in Lecture 1, correlation does not automatically mean causation. A scatterplot can suggest that two variables are related, but it does not by itself prove that one variable causes the other.

Example. Suppose distance from campus and commute time are strongly correlated. It is reasonable to think that living farther away often contributes to a longer commute. However, the exact commute time may also depend on train lines, transfer stations, bicycle access, traffic, walking speed and whether the student lives near a convenient station.

A good scatterplot may *suggest* a causal relationship, but the interpretation still requires context.

2.6 Case Study: Konbinis in Japan

Suppose that we have a company that produces anpan. They want to sell their products in convenience stores (konbinis) somewhere in Japan, and they are trying to do some market research to properly inform their decision. They task a poor intern, Jenny, to do some research about the distribution of konbinis nationwide, and then to present the results. Simply put, Jenny's task is to:

- figure out how konbinis are distributed within Japan, and then
- present the data to her bosses.

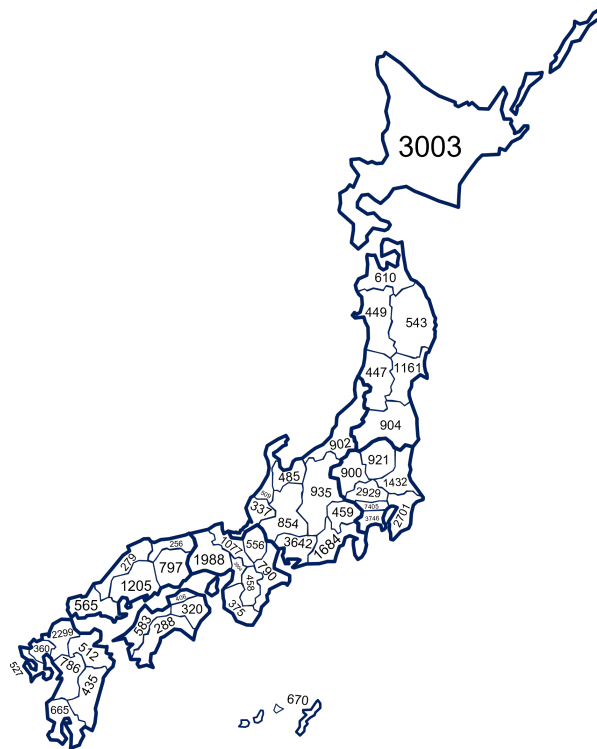
So, Jenny uses a data set of convenience stores listed by prefecture:

Prefecture	Stores	Prefecture	Stores	Prefecture	Stores	Prefecture	Stores
Hokkaido	3003	Aomori	610	Shizuoka	1684	Ishikawa	509
Saga	360	Kagawa	406	Kochi	288	Nagasaki	527
Yamanashi	459	Aichi	3642	Tottori	256	Iwate	543
Fukui	337	Okayama	797	Shimane	279	Saitama	2929
Tokyo	7405	Tochigi	921	Okinawa	670	Fukuoka	2299
Ehime	583	Yamaguchi	565	Niigata	902	Shiga	556
Miyagi	1161	Toyama	485	Nagano	935	Osaka	3954
Gifu	854	Kagoshima	665	Miyazaki	435	Hyogo	1988
Ibaraki	1432	Akita	449	Oita	512	Mie	790
Hiroshima	1205	Yamagata	447	Wakayama	375	Nara	458
Fukushima	904	Gunma	900	Kumamoto	786	Tokushima	320
Chiba	2701	Kyoto	1077	Kanagawa	3746		

We will now explore some ways in which to present this data visually.

2.6.1 Map representations

The data written above is precise, but it is not very easy to interpret. As you can see, it contains all of the relevant information, however, it is very difficult to make sense of any underlying patterns in the data. Even if we were to plot this honestly on a map of Japan, the results would be uninformative:

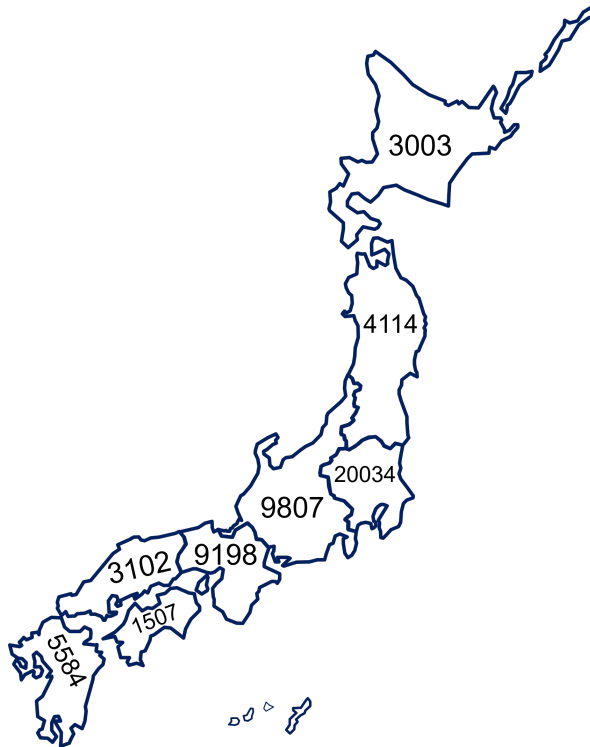


As you can see, this map is not a good way to represent this data. Although it has hints of a good idea, there are three main issues:

1. the numbers are too small to read,
2. it is still difficult to see any sort of pattern in the data,
3. even if we *could* see all of the numbers, their exact values may not actually be relevant.

One approach to solving this might be to reduce the resolution of the map by aggregating the data into regions instead of prefectures. Firstly, we may rewrite the counts into the main regions of Japan:

Region	Convenience Stores
Hokkaido	3003
Tohoku	4114
Kanto	20034
Chubu	9807
Kansai	9198
Chugoku	3102
Shikoku	1597
Kyushu	5584
Okinawa	670



As you can see, this data tells a clearer story: Kanto has the largest number of convenience stores by far, followed by Chubu and Kansai. However, we have lost detail in the process of making the data easier to read. We will now resolve this.

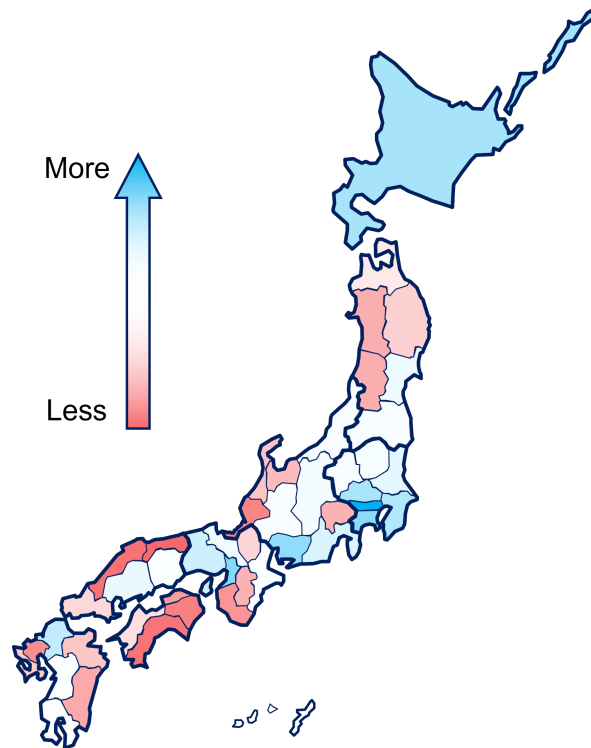
2.6.2 Coloured representations

Given the context of this presentation, it is most likely that Jenny doesn't actually need to present the exact numerical data of her study. In other words, the raw data may display *too much* information to be genuinely useful. We can suppress this information whilst maintaining its patterns by encoding quantity via a colour gradient. In the table below, the higher counts are coloured further towards blue, and the lower counts are coloured further towards red, with white being in the middle.

Prefecture	Stores	Prefecture	Stores	Prefecture	Stores	Prefecture	Stores
Hokkaido	3003	Aomori	610	Shizuoka	1684	Ishikawa	509
Saga	360	Kagawa	406	Kochi	288	Nagasaki	527
Yamanashi	459	Aichi	3642	Tottori	256	Iwate	543
Fukui	337	Okayama	797	Shimane	279	Saitama	2929
Tokyo	7405	Tochigi	921	Okinawa	670	Fukuoka	2299
Ehime	583	Yamaguchi	565	Niigata	902	Shiga	556
Miyagi	1161	Toyama	485	Nagano	935	Osaka	3954
Gifu	854	Kagoshima	665	Miyazaki	435	Hyogo	1988
Ibaraki	1432	Akita	449	Oita	512	Mie	790
Hiroshima	1205	Yamagata	447	Wakayama	375	Nara	458
Fukushima	904	Gunma	900	Kumamoto	786	Tokushima	320
Chiba	2701	Kyoto	1077	Kanagawa	3746		

We can take this data and simply shade each prefecture with its corresponding colour. Of course,

this is a loss of information compared to simply presenting the raw counts. However, the payoff is greater: we can immediately see the geographic pattern in the data.



As you can see, the patterns in the data are more detailed than the regional map plotted previously. Here, we see that there are clear hotspots of concentration around urban areas: Tokyo, Nagoya, Osaka, and to some extent Sendai, Hiroshima and Fukuoka. Conversely, outside of populated areas we see a large drop-off in konbini frequency: Shimane, Tottori, Shikoku and northern Tohoku all present a lower-than-average count. Aside from these two general observations, we also notice that Hokkaido is a clear outlier.

Although the above is a better representation than the previous graphs, there is still a sense in which it can be improved: within the context of Jenny and her company, it is likely that the CEOs already know that urban areas have more convenience stores than the national average. Therefore, the study doesn't quite display anything new. However, this does not mean that there is nothing else hiding away inside.

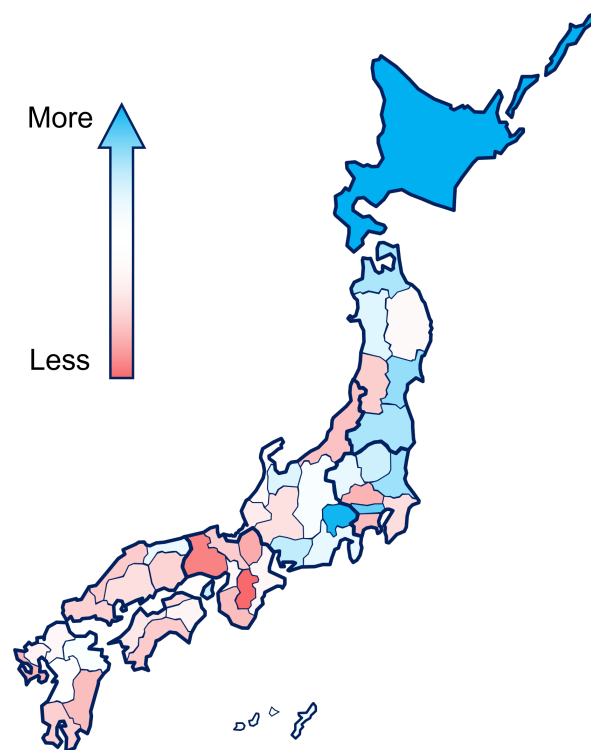
To illustrate a different pattern in this data, we will now divide the store count by the population, then multiply by 100,000, effectively creating a "konbini per 100,000 people" measure. In the table below (again labelled with our colour gradient) we display this extra information.

Prefecture	Stores	Per 100k
Hokkaido	3003	57.43
Yamanashi	459	56.64
Tokyo	7405	52.65
Miyagi	1161	50.40
Ibaraki	1432	49.92
Fukushima	904	49.29
Aomori	610	49.24
Aichi	3642	48.26
Tochigi	921	47.62
Toyama	485	46.83
Akita	449	46.77
Gunma	900	46.38
Shizuoka	1684	46.33
Tottori	256	46.22
Okinawa	670	45.63
Nagano	935	45.62

Prefecture	Stores	Per 100k
Oita	512	45.53
Kumamoto	786	45.19
Ishikawa	509	44.91
Iwate	543	44.83
Fukuoka	2299	44.74
Osaka	3954	44.72
Mie	790	44.60
Tokushima	320	44.46
Saga	360	44.33
Fukui	337	43.91
Ehime	583	43.65
Gifu	854	43.14
Hiroshima	1205	43.01
Chiba	2701	42.96
Kagawa	406	42.69
Okayama	797	42.18

Prefecture	Stores	Per 100k
Yamaguchi	565	42.07
Kagoshima	665	41.85
Yamagata	447	41.83
Kyoto	1077	41.75
Kochi	288	41.62
Shimane	279	41.54
Niigata	902	40.96
Miyazaki	435	40.65
Wakayama	375	40.63
Kanagawa	3746	40.54
Nagasaki	527	40.13
Saitama	2929	39.87
Shiga	556	39.31
Hyogo	1988	36.35
Nara	458	34.56

This table tells a very different story from the raw total counts: some prefectures (especially Yamanashi and those in northern Tohoku) suddenly switch from red to blue, meaning that they are suddenly over-represented. Conversely, some of the urban areas from before now pass into the red as they are scaled by population. Tokyo, however, stays highly represented in the data, and Hokkaido is still an outlier. What is perhaps more interesting is the pattern displayed after plotting onto our map of Japan. Compare the graph below to that of the raw count.



We see that there is a clear East versus West split in terms of konbini per capita: the east of Japan typically scores higher, and the west of Japan typically scores lower. In the context of Jenny's

presentation, she may therefore suggest investigating the east of Japan further, depending on whether the company values store density, total store count, or something else.

2.6.3 Some observations on data visualization

Finally, we finish with some important observations regarding data visualization.

Important point

Data visualization is usually an intentional *loss* of information in the pursuit of insight. Aside from honesty and accuracy, you should be guided by the principle that your presentation should be useful and interesting, ideally showing the audience something they might not otherwise see.

Presentation of data is always context dependent. As we have seen throughout this lecture, there can always be ways to warp narratives based on choices made by the presenter. Whenever you observe data in the future, it is a good idea to ask yourself the following questions.

1. **Data type:** What structure does the data have: quantitative, categorical, time-dependent, paired, or perhaps some combination?
2. **Purpose:** What question is the graph trying to answer?
3. **Scale:** Are the axes and scales honest, and easy to read?
4. **Grouping:** If the data is grouped, are the bins or categories fair and reasonable?
5. **Loss of information:** What details could have been hidden or simplified?
6. **Conflicts of interest:** Who created this data, and why might they be presenting it in this way?
7. **Narrative:** What conclusion does the visualization encourage the audience to make?
8. **Alternative views:** Would another representation tell a different (but equally valid) story?

Lecture 3

Central Tendency and Variation

In Lecture 1, we introduced the basic language of statistics, including data, variables, populations, samples and sampling methods. In Lecture 2, we studied how to represent data visually using histograms, bar charts, pie charts, time-series graphs and scatterplots. In this lecture, we begin the numerical side of descriptive statistics. Our main goal is to learn how to summarise a data set using a small number of carefully chosen statistics. We will focus on two major ideas: *central tendency*, which describes where the middle of the data is, and *variation*, which describes how spread out the data is.

3.1 Central Tendency

Throughout this lecture we will be working with quantitative data. We are going to need to talk about this data in a structured way, so we will first introduce some important notation. Throughout all of the formulae used in this lecture, we are going to write the data values using the letter x , labelled with subscripts:

$$x_1, x_2, x_3, x_4, \dots, x_n.$$

The subscript simply gives us an easy label to refer to later on. Notice that the labels on the data start at 1 and count up until we reach n . This final number n (whatever it may be) tells us the total size of the data set. When order matters, we will write our data sets from smallest to largest, that is:

$$x_1 \leq x_2 \leq x_3 \leq x_4 \leq \dots \leq x_n.$$

3.1.1 The meaning of central tendency

When we collect a lot of data, we often want to summarise it with a single representative number. One way to do this is to create a description of where the “middle” of the data is. This way, we can make claims about where the data points *tend to be*. This is exactly the idea of central tendency.

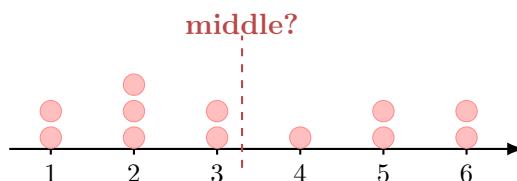
Definition: Central Tendency

Central tendency is the attempt to describe the middle of a data set.

For example, consider the data set

$$(1, 1, 2, 2, 2, 3, 3, 4, 5, 5, 6, 6).$$

We can plot this data on a number line by representing each observation as a small ball:



In the diagram, should the middle of the data be this dashed line, or perhaps something else? The task of central tendency is to figure out a meaningful way to find the location of the middle. The intuition is clear, however the mathematical question is less simple, because there are many ways to interpret the word “middle”. This is why there are several different measures of central tendency.

3.1.2 Summation notation

In the formulas we will later introduce, we are going to need to perform some arithmetic operations on data, which means that we also need to figure out how to write things down in an efficient way. To illustrate this, suppose that we were working with a data set of size $n = 1000$ and we wanted to add up all of the values of the data. In this situation, we don’t want to keep writing out:

$$x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10} + x_{11} + x_{12} + \dots$$

by hand over and over again. To solve this issue, we instead use *summation notation*.

Summation notation

$$\sum_{i=1}^n x_i \quad \text{means} \quad x_1 + x_2 + x_3 + \cdots + x_n.$$

The letter i is the **index**. It tells us which term we are currently adding. The lower number tells us where to start, and the upper number tells us where to stop.

As you can see from the above, summation notation is simply an efficient way to write out a really big sum. It should be noted that we can use the notation to write out more than just basic addition of x_i 's. For example, we may write:

$$\sum_{i=1}^n 2x_i = 2x_1 + 2x_2 + 2x_3 + \cdots + 2x_n.$$

In the above, we see that instead of summing up all the individual x_i values, we sum up all the terms $2x_i$.

Example. Suppose

$$(x_1, x_2, x_3) = (4, 10, 12).$$

Then

$$\sum_{i=1}^3 x_i = x_1 + x_2 + x_3 = 4 + 10 + 12 = 26,$$

$$\sum_{i=1}^2 x_i = x_1 + x_2 = 4 + 10 = 14,$$

and

$$\sum_{i=2}^3 x_i^2 = x_2^2 + x_3^2 = 10^2 + 12^2 = 100 + 144 = 244.$$

Exercise

Suppose

$$(x_1, x_2, x_3, x_4, x_5) = (2, 4, 6, 8, 10).$$

Compute:

$$(a) \sum_{i=1}^4 x_i$$

$$(b) \sum_{i=3}^5 x_i$$

$$(c) \sum_{i=1}^5 x_i$$

$$(d) \sum_{i=2}^3 x_i$$

$$(e) \sum_{i=1}^3 \frac{x_i}{2}$$

Solution

$$(a) \sum_{i=1}^4 x_i = x_1 + x_2 + x_3 + x_4 = 2 + 4 + 6 + 8 = 20.$$

$$(b) \sum_{i=3}^5 x_i = x_3 + x_4 + x_5 = 6 + 8 + 10 = 24.$$

$$(c) \sum_{i=1}^5 x_i = x_1 + x_2 + x_3 + x_4 + x_5 = 2 + 4 + 6 + 8 + 10 = 30.$$

$$(d) \sum_{i=2}^3 x_i = x_2 + x_3 = 4 + 6 = 10.$$

$$(e) \sum_{i=1}^3 \frac{x_i}{2} = \frac{x_1}{2} + \frac{x_2}{2} + \frac{x_3}{2} = \frac{2}{2} + \frac{4}{2} + \frac{6}{2} = 1 + 2 + 3 = 6.$$

3.1.3 Three common averages: mean, median and mode

In everyday language, the word “average” is ambiguous. It can refer to several different statistics. We will now review the three most common averages: the **mean**, the **median** and the **mode**. They are all measures of central tendency, but they answer slightly different questions.

Statistic	Basic idea	Main weakness
Mean	The balanced middle of the data.	Sensitive to outliers.
Median	The middle value after sorting.	Does not use the numerical size of every value.
Mode	The most frequent value.	May not be unique, and may not say much for numerical data.

We will now review these one-by-one, starting with the median.

Median

The median is the value that splits the whole data set in half. Once the median is found, at least half of the data lies at or below the median, and at least half lies at or above it.

Definition: Median

The **median** is the middle value of an ordered data set.

In order to find the median, we have to first pay attention to the value of n : if it is odd or even then we will have to perform slightly different operations. The step-by-step process is:

Step 1: Order all values from smallest to biggest.

Step 2: If n is odd, pick the middle entry. This is the $\frac{n+1}{2}$ -th entry.

Step 3: If n is even, look at the two middle entries and calculate the number halfway between them.

For example, consider the data set:

$$(1, 2, 3, 4, 5).$$

Here there are 5 data points, i.e. $n = 5$. So, to find the median we add +1 to n and then halve it: $\frac{5+1}{2} = 3$. So, we simply go to the 3rd entry in the data set and read off its value. In this case, the median is 3.

Consider now the data set:

$$(2, 3, 4, 5, 6, 7).$$

Here there are now 6 numbers, meaning that $n = 6$ which is even. The median position is $\frac{6+1}{2} = 3.5$. This means that the median will be a number that is precisely halfway between the 3rd and 4th entries in the data set. By inspection, we see that these two entries correspond to the data points 4 and 5. Therefore the median is halfway between these two numbers, i.e. the median is $\frac{4+5}{2} = 4.5$.

Mean

The mean is what we would ordinarily call the “average” of a data set. For a sample

$$x_1, x_2, \dots, x_n,$$

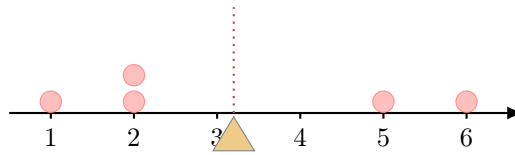
the mean is written $\bar{x}$ and is defined as follows.

Definition: Sample Mean

For a sample $x_1, x_2, \dots, x_n$, the **sample mean** is

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}.$$

In other words, the mean $\bar{x}$ is calculated by summing up all the values in the data set, and then dividing by how many there are. Geometrically, we can imagine the mean as a balancing point of the data. If you imagine each data point as a weight sitting on a number line, the mean is the point where the whole system balances. For example, for the data set $(1, 2, 2, 5, 6)$ we can imagine the mean as the balancing point in the middle:



Numerically, in this example we calculate:

$$\bar{x} = \frac{1 + 2 + 2 + 5 + 6}{5} = \frac{16}{5} = 3.2.$$

One important property of the mean is that the positive and negative deviations from the mean always cancel.

A useful fact about the mean

For any data set,

$$\sum_{i=1}^n (x_i - \bar{x}) = 0.$$

In other words, the total signed distance from the data to the mean is always zero.

Keeping with the data set $(1, 2, 2, 5, 6)$ from above, we can verify this fact by hand. Observe that there are three values to the left of the mean $\bar{x} = 3.2$, namely 1, 2 and 2, and there are two values greater than the mean, namely 5 and 6. Since the values to the left of $\bar{x}$ are smaller than 3.2, the values of $x_i - \bar{x}$ will be negative. In absolute value, the actual distances are

$$(3.2 - 1) + (3.2 - 2) + (3.2 - 2) = 2.2 + 1.2 + 1.2 = 4.6,$$

while for those data points to the right of $\bar{x}$ we have:

$$(5 - 3.2) + (6 - 3.2) = 1.8 + 2.8 = 4.6.$$

These values demonstrate that the sum of the distances to the left of the mean equals the sum of the distances to the right of the mean. This is why it makes sense to call the mean the balanced middle.

Alternatively, this fact about the mean may also be verified algebraically:

$$\begin{aligned}
 \sum_{i=1}^n (x_i - \bar{x}) &= (x_1 - \bar{x}) + (x_2 - \bar{x}) + \cdots + (x_n - \bar{x}) \\
 &= x_1 + x_2 + \cdots + x_n - \bar{x} - \bar{x} - \cdots - \bar{x} \\
 &= x_1 + x_2 + \cdots + x_n - n \cdot \bar{x} \\
 &= \sum_{i=1}^n x_i - n \cdot \frac{\sum_{i=1}^n x_i}{n} \\
 &= \sum_{i=1}^n x_i - \sum_{i=1}^n x_i \\
 &= 0.
 \end{aligned}$$

Mode

Simply put, the mode is the most frequently occurring value in the data set.

Definition: Mode

The **mode** is the value that occurs most often in a data set.

In the case that there are competing data points that occur with the same highest frequency, we will say that the data set has no unique mode.

Example. The data set

$$(1, 1, 1, 3, 4)$$

has mode 1, because 1 occurs three times.

The data set

$$(1, 1, 2, 2, 3)$$

has no unique mode, because 1 and 2 both occur twice.

3.1.4 A worked example

Consider the data set

$$(1, 1, 2, 2, 3, 4, 5).$$

The mean is

$$\bar{x} = \frac{1 + 1 + 2 + 2 + 3 + 4 + 5}{7} = \frac{18}{7} \approx 2.57.$$

There are 7 data points, so the middle entry is the

$$\frac{7+1}{2} = 4,$$

so the median is the 4th entry. The 4th entry is 2, so the median is 2.

The values 1 and 2 both occur twice, and no value occurs more often. Therefore, there is no unique mode.

Exercise

Consider the data set

$$(1, 1, 2, 2).$$

Find:

- (a) the mean,
- (b) the median,
- (c) the mode.

Solution

- (a) The mean is

$$\bar{x} = \frac{1 + 1 + 2 + 2}{4} = \frac{6}{4} = 1.5.$$

- (b) There are 4 data points, so the two middle values are the 2nd and 3rd entries. These are 1 and 2, so the median is

$$\frac{1 + 2}{2} = 1.5.$$

- (c) The values 1 and 2 both occur twice. Therefore, there is no unique mode.

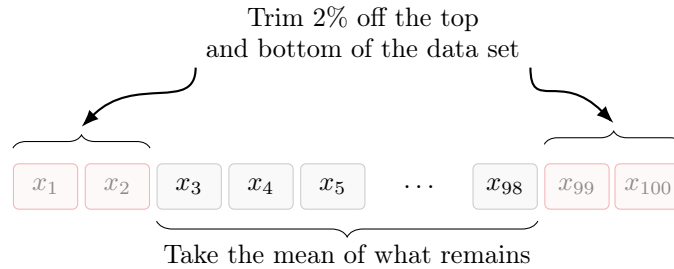
3.1.5 Trimmed means

The mean uses every value in the data set. This is often useful, but it also allows the mean to be affected by extreme outliers: if we had a single piece of data far off to one side, then the mean would have to compensate a lot in order to maintain its balance. In some situations this feature of the mean might be considered permissible. However, in other situations we might want to intentionally remove the effects of extreme outliers. This motivates the notion of a *trimmed mean*.

Definition: Trimmed Mean

An $x\%$ **trimmed mean** is obtained by removing the lowest $x\%$ and the highest $x\%$ of an ordered data set, and then calculating the mean of the remaining values.

Put differently, we create an $x\%$ trimmed mean by cutting off the bottom $x\%$ and the top $x\%$ of an ordered data set, and then taking the mean of what is left over. For example, if we have an ordered data set of size 100, a 2% trimmed mean would remove the two smallest values and the two largest values. It then takes the mean of the 96 values that are left over:



Trimmed means are especially helpful when trying to average out a collection of human judgements. For example, in figure skating, a panel of judges give scores to an individual performance. If one judge gives an unusually high score and another gives an unusually low score due to inherent bias or human error, then those extreme scores might affect the mean in an unfair way. To help regulate this effect, a performance score can be calculated with a trimmed mean to help reduce the influence of these extreme judgements.

Exercise: Trimmed means

Consider the ordered data set

1, 1, 2, 3, 4, 5, 5, 6, 6, 6, 7, 7, 8, 9, 10, 11, 16, 20, 23, 24.

Here $n = 20$.

- Compute the 5% trimmed mean.
- Compute the 20% trimmed mean.
- Explain why a 50% trimmed mean is not meaningful here.

Solution

- (a) Since 5% of 20 is 1, we remove the smallest value and the largest value. The remaining data are

1, 2, 3, 4, 5, 5, 6, 6, 6, 7, 7, 8, 9, 10, 11, 16, 20, 23.

The sum of these values is 149, so the 5% trimmed mean is

$$\frac{149}{18} \approx 8.28.$$

- (b) Since 20% of 20 is 4, we remove the lowest four values and the highest four values. The remaining data are

4, 5, 5, 6, 6, 6, 7, 7, 8, 9, 10, 11.

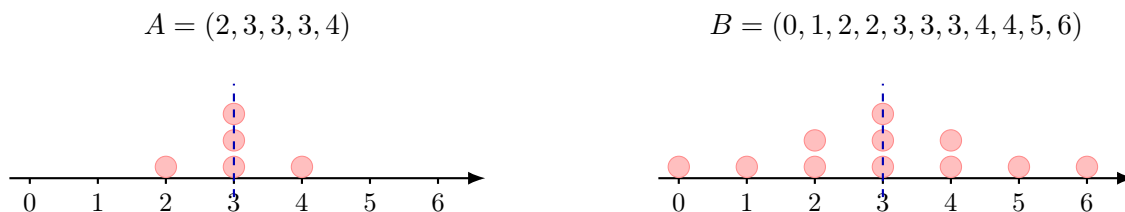
The sum of these values is 84, so the 20% trimmed mean is

$$\frac{84}{12} = 7.$$

- (c) Trimming 50% from the bottom and 50% from the top removes the entire data set. There are no values left to average, so a 50% trimmed mean is not meaningful.

3.2 Variation

Measures of central tendency allow us to identify where our data tends to be, in relation to a middle. But, central tendency cannot tell us how spread out data is. For example, consider the two data sets drawn below:



These two data sets have the same mean (both are 3), but data set B is clearly spread out much more than data set A . There is no sense in which an individual description of the centre of data will ever be able to capture the feature of spread. Instead, we need to introduce a new concept: variation.

Definition: Variation

Variation describes how spread out the data is.

3.2.1 Range

Consider the two data sets, $A = (1, 1, 1, 1, 2)$ and $B = (1, 2, 3, 3, 5)$, which are drawn below.



Visually, it is very clear that B is more spread out, and A is more tightly packed. One way to quantify this is by looking at the distance between the largest and smallest value in the data sets: for A , the whole data set is contained between 1 and 2, whereas for B the whole data set spreads out between 1 and 5.

This simple observation motivates our first definition of spread: the *range* of a data set.

Definition: Range

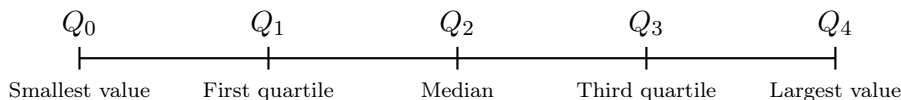
The **range** of a data set is

$$\text{range} = \max(\text{data}) - \min(\text{data}).$$

For example, our data set $A = (1, 1, 1, 1, 2)$ has range $2 - 1 = 1$, and our data set $B = (1, 2, 3, 3, 5)$ has range $5 - 1 = 4$. So, the range confirms what we saw visually: B is more spread out than A .

3.2.2 Quartiles

The word “quarter” means one fourth of something. In distinction to this, the word *quartile* means a value that helps divide an ordered data set into four quarters. Since there are four quarters in total, we need five lines to divide them up: three in the middle and two at the endpoints:



The five values

$$Q_0, \quad Q_1, \quad Q_2, \quad Q_3, \quad Q_4$$

mark the minimum, first quartile, median, third quartile and maximum.

Quartiles

For an ordered data set:

- Q_0 is the smallest value.
- Q_1 is the median of the bottom half of the data.
- Q_2 is the median of the whole data set.
- Q_3 is the median of the top half of the data.
- Q_4 is the largest value.

Different textbooks sometimes use slightly different conventions for quartiles, especially when the data set has an odd number of values. In this course, we will use the following practical rule: first find the median, then find the median of the lower half and the median of the upper half. If n is odd, we will not include the median in either half when finding Q_1 and Q_3 .

3.2.3 Interquartile range

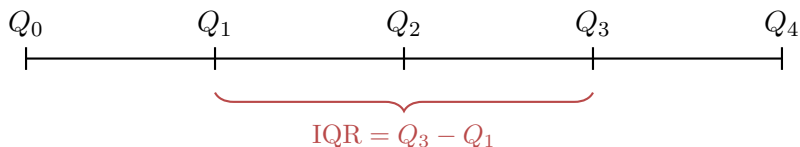
Definition: Interquartile Range

The **interquartile range**, abbreviated **IQR**, is

$$\text{IQR} = Q_3 - Q_1.$$

It measures the spread of the middle half of the data.

Visually, the interquartile range measures this distance:



The IQR is often more resistant to outliers than the range, because it focuses on the middle half of the data and ignores the lower and upper quarters.

Worked example: quartiles and IQR

Consider the ordered data set:

$$(1, 1, 2, 3, 4, 5, 5, 6, 6, 6, 7, 7, 8, 9, 10, 11, 16, 20, 23, 24).$$

To calculate the quartiles, we will first find the median. Since there are 20 data points in total, the median will lie halfway between the 10th and 11th entries. These entries are 6 and 7, therefore, the median is $Q_2 = \frac{6+7}{2} = 6.5$. To find the value of Q_1 , we need to calculate the median of the bottom half of the data. Writing this out, we are looking for the middle of the data:

$$(1, 1, 2, 3, 4, 5, 5, 6, 6, 6).$$

The median of this bottom half lies halfway between 4 and 5, so

$$Q_1 = \frac{4+5}{2} = 4.5.$$

Similarly, the top half of the data is:

$$(7, 7, 8, 9, 10, 11, 16, 20, 23, 24).$$

The median of this top half lies halfway between 10 and 11, so $Q_3 = \frac{10+11}{2} = 10.5$.

Putting this all together, the five-number summary is:

$$Q_0 = 1, \quad Q_1 = 4.5, \quad Q_2 = 6.5, \quad Q_3 = 10.5, \quad Q_4 = 24,$$

and the interquartile range can be calculated as:

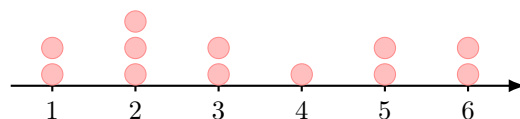
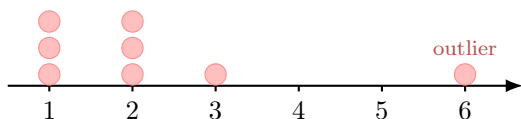
$$\text{IQR} = Q_3 - Q_1 = 10.5 - 4.5 = 6.$$

3.2.4 Variance

The range only uses two numbers: the minimum and the maximum. This makes it quick to compute, but it also makes it blind to most of the data set. The range is also extremely sensitive to outliers. For instance, consider the following two data sets:

$$A = (1, 1, 1, 2, 2, 2, 3, 6)$$

$$B = (1, 1, 2, 2, 2, 3, 3, 4, 5, 5, 6, 6)$$



Both data sets have the same range: $6 - 1 = 5$. However, visually there is still a sense in which data set A is less spread out than data set B . In data set A most of the data is packed around 1, 2 and 3, and there is a single outlier that causes the range to equal 5. In contrast, in data set B the data is much more uniformly spread, and there are no outliers.

We have just identified a fundamental flaw with the range of a data set: although it is useful as an initial measure of variation, the range is extremely sensitive to outliers and can therefore miss the

internal structure of a data set. To fix this, we will now introduce a more careful measure of spread known as *variance*.

Definition: Sample Variance

For a sample $x_1, x_2, \dots, x_n$ with sample mean $\bar{x}$, the **sample variance** is

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}.$$

As you can see, the formula is a lot more complicated than those we have seen before. So, we will break this down step-by-step to explain what s^2 is actually trying to say. Below is an annotation of the formula:

First, the $x_i - \bar{x}$ term measures how far the data point x_i is from the mean $\bar{x}$

We then square this distance

Finally, we take a sample-adjusted average of these squared distances

In a nutshell, the variance s^2 measures a sample-adjusted average of the squared distances between each data value and the mean:

1. The expression $x_i - \bar{x}$ measures how far away the point x_i is from the centre $\bar{x}$. It is not quite a distance, because it can take negative values, but is instead called a *deviation*.
2. The expression $(x_i - \bar{x})^2$ then squares all of these deviations.
3. Next, the summation notation $\sum_{i=1}^n (x_i - \bar{x})^2$ says to add up all of these squared deviations, i.e.

$$\sum_{i=1}^n (x_i - \bar{x})^2 = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2.$$

4. Finally, we divide by $n - 1$ to get the sample variance.

We will now elaborate on why *this* is the formula, in particular.

Why do we square the deviations?

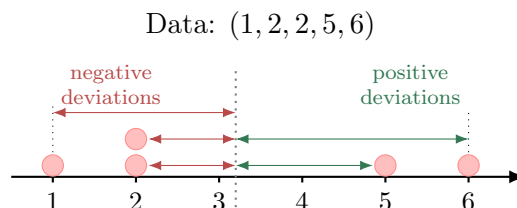
It is reasonable to ask why the variance formula squares the deviations. Instead, why not just add up the raw deviations from the mean? In fact, the issue is hidden away in the important property of the mean that we saw earlier:

$$\sum_{i=1}^n (x_i - \bar{x}) = 0.$$

This says that the sum of all the raw deviations will always equal zero. If we were to try to use the raw deviations without squaring them, then we would always get:

$$\frac{\sum_{i=1}^n (x_i - \bar{x})}{n - 1} = \frac{0}{n - 1} = 0,$$

no matter what data we started with. For example, consider the data set:



whose mean is $\bar{x} = 3.2$. If we tried to add up the raw deviations, we would get:

$$(1 - 3.2) + (2 - 3.2) + (2 - 3.2) + (5 - 3.2) + (6 - 3.2) = 0.$$

Squaring solves this problem in general. Once we square a deviation, both negative and positive deviations have non-negative squares:

$$(-2)^2 = 4, \quad 2^2 = 4.$$

This means that the sum is forced to be non-negative, and therefore avoids the problem mentioned above.¹

Why do we divide by $n - 1$?

Normally, when we take an average, we add values and divide by the number of values. So it is natural to ask why the sample variance divides by $n - 1$ instead of n .

The short answer is that s^2 is actually a **sample statistic** and not a **population parameter**. Usually, we use a sample to estimate the spread of a larger population. When we use the sample mean $\bar{x}$ inside the formula, we have already estimated one quantity from the sample, and this makes the squared deviations slightly too small on average. Dividing by $n - 1$ corrects for this effect.

Bessel's correction

The use of $n - 1$ instead of n in the sample variance formula is called **Bessel's correction**. In this course, the main point is practical:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

is the standard formula for sample variance.

We will not prove this correction in this course. For now, you should know which formula to use and what it is measuring.

¹Also, squaring gives especially large weight to deviations with large absolute value.

A worked example: variance

Example. Consider the data set

$$(1, 1, 2, 3).$$

First compute the mean:

$$\bar{x} = \frac{1 + 1 + 2 + 3}{4} = \frac{7}{4} = 1.75.$$

Now compute the deviations and squared deviations:

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	-0.75	0.5625
1	-0.75	0.5625
2	0.25	0.0625
3	1.25	1.5625

The sum of the squared deviations is

$$0.5625 + 0.5625 + 0.0625 + 1.5625 = 2.75.$$

Since $n = 4$, we divide by $n - 1 = 3$:

$$s^2 = \frac{2.75}{3} \approx 0.917.$$

3.2.5 Standard deviation

Variance is useful, but it has one awkward feature: it is measured in squared units. For example, if the data is measured in minutes, then the variance is measured in minutes squared. This quantity can be hard to directly relate to our initial data. We can solve this issue by simply taking the square root of the variance, so the result will be in the same units as we started with. This is precisely the definition of the standard deviation.

Definition: Sample Standard Deviation

The **sample standard deviation** is

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}.$$

The standard deviation is in the same units as the data. Because of this, it is easier to interpret as a distance from the mean.

In the previous example, we found that

$$s^2 \approx 0.917.$$

Therefore,

$$s = \sqrt{0.917} \approx 0.958.$$

So the standard deviation is approximately 0.958.

3.2.6 Coefficient of variation

Sometimes we want to compare the spread of data sets with very different scales. For example, a standard deviation of 2 is large if the mean is 4, but very small if the mean is 2000. To make the spread *relative* to the size of the data set, we can use the coefficient of variation.

Definition: Coefficient of Variation

For positive ratio-scale data, the coefficient of variation is a unitless measure of relative spread. It is defined as the ratio:

$$CV = \frac{s}{\bar{x}}.$$

A larger coefficient of variation means more variation compared with the typical size of the values.

This statistic is most useful when the mean is positive and not close to zero. If the mean is very close to zero, dividing by the mean can make the coefficient of variation misleading or unstable.

3.2.7 Sample statistics and population parameters

As discussed in Lecture 1, inferential statistics often uses sample statistics to estimate population parameters. Because of this, we use different notation for samples and populations.

Quantity	Sample statistic	Population parameter
Mean	$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$	$\mu = \frac{\sum_{i=1}^N x_i}{N}$
Variance	$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$	$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$
Standard deviation	$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$	$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}}$

Here, n is the sample size, while N is the population size. Observe that for the sample variance and sample standard deviation we have used Bessel's correction. But we do not need to do that for the population variance or population standard deviation, because we are no longer trying to make an estimate.

3.3 Box and Whisker Plots

3.3.1 The five-number summary

A box and whisker plot is a visual representation of the five-number summary.

Five-number summary

The **five-number summary** of a data set is:

lowest value, Q_1 , median, Q_3 , highest value.

Equivalently,

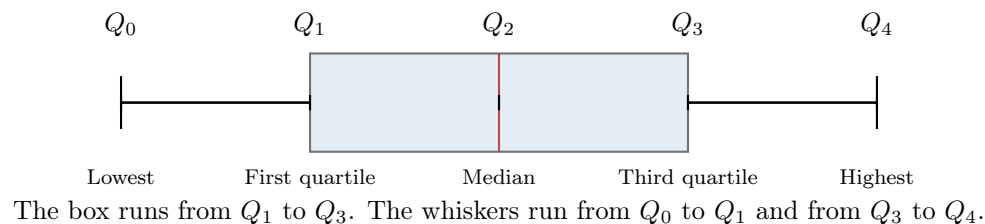
Q_0 , Q_1 , Q_2 , Q_3 , Q_4 .

Box and whisker plots are useful because they show both central tendency and variation at the same time. The median shows the middle of the data, the box shows the middle half of the data, and the whiskers show the outer parts of the data.

3.3.2 Anatomy of a box and whisker plot

In the simple version of a box and whisker plot used in this lecture:

- the left whisker begins at the lowest value Q_0 ,
- the left side of the box is Q_1 ,
- the line inside the box is the median Q_2 ,
- the right side of the box is Q_3 ,
- the right whisker ends at the highest value Q_4 .



Definition: Box and Whisker Plot

In the simple version used in this lecture, a **box and whisker plot** is a graph that displays a data set using its five-number summary: lowest value, Q_1 , median, Q_3 and highest value.

3.3.3 Constructing a box and whisker plot

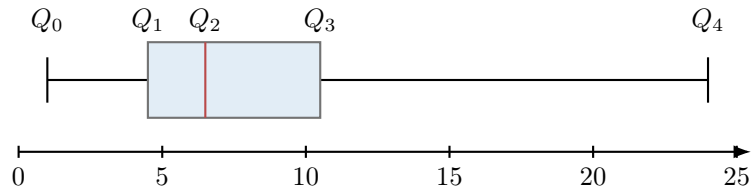
Consider the data set:

1, 1, 2, 3, 4, 5, 5, 6, 6, 6, 7, 7, 8, 9, 10, 11, 16, 20, 23, 24.

We have previously determined the five-number summary:

$$Q_0 = 1, \quad Q_1 = 4.5, \quad Q_2 = 6.5, \quad Q_3 = 10.5, \quad Q_4 = 24.$$

The resulting box and whisker plot is:



Notice that the right whisker is much longer than the left whisker. This tells us that the upper end of the data is more spread out than the lower end.

Part II

Probability Theory

Lecture 4

Mathematical Foundations of Probability

In Lectures 1–3, we studied the basic language of statistics, visual descriptions of data, and numerical summaries such as the mean, median, variance and standard deviation. We now move from descriptive statistics towards probability theory. However, before getting there, we will first need to introduce some mathematical concepts that are used heavily throughout Probability Theory. In this lecture, we will do exactly this, by introducing the set theory and counting methods that will be used throughout later portions of our course.

4.1 Set Theory

4.1.1 What is a set?

In order to talk about a group of things mathematically, we need a way to write them down all together. That is what a set is: it is simply an unordered list of items.

Definition: Set

A **set** is an unstructured collection of things.

We can imagine a set as a mathematical box in which to put things. There is no inherent structure to the box, except that it exists and contains the things that it does. In fact, *that's it*: a set is simply the mathematical way to collect things together in one place. Given how open-ended this notion is, we can form sets from basically any objects. For example:

- the set of all numbers greater than 2,
- the set of all people in the classroom who have brown hair,
- the set of all students who arrived by train today,
- the set of all penguins named Carl.

We call the objects inside a set the *elements* of a set, they are like the members of the set. In mathematical notation, we write sets using the curly brackets $\{$ and $\}$. For example the notation:

$$A = \{1, 2, 3, 4, 5, 6, 7\}$$

is how we write the set A whose elements are 1, 2, 3, 4, 5, 6 and 7. To emphasize: it is best to think of sets as having no inherent order. Of course, we have to list the elements of a set in *some* sort of order, however this does not matter at all: a set is only specified by its elements. For instance, the set A can be equally written as:

$$A = \{3, 4, 1, 7, 2, 5, 6\}, \quad \text{or} \quad A = \{3, 1, 4, 5, 7, 6, 2\}.$$

If an object x happens to be an element of the set A , then we write

$$x \in A,$$

which is simply read “ x is in A ”. If x were *not* an element of A , then we would write a slash through the above symbol, similar to the way we write $\neq$. The notation:

$$x \notin A$$

reads “ x is not in A ”. For example, if $A = \{1, 3, 5, 7, 9\}$, then $3 \in A$ but $4 \notin A$.

4.1.2 Rules for sets

Sets are the fundamental building block of almost all other mathematical structures, since before ever talking about the structure that a collection of things may or may not have, we first need to collect them all into the same place. In that sense, the theory of sets can be seen as a foundation of

modern mathematics. However, although we may think of sets as a “proto-structure”, there are still some basic ground rules.

Basic rules for sets

Rule 1: You cannot write the same element twice.

Rule 2: If two sets have all of the same elements, then they are equal.

Rule 3: Sets can have other sets as elements, but a set cannot contain itself.

Rule 4: There is a special set with absolutely nothing in it, called the empty set, written $\emptyset$.

Rule 5: There is no universal set containing absolutely everything.

Rule 1 states that we only write elements of a set *once*. This means, for example, the two sets below are equal:

$$\{1, 1, 2\} = \{1, 2\}$$

that is, when writing out a set we do not write any duplicates. In other words, a set does not count frequency of its members, only their presence. The second rule means that

$$\{1, 2, 3\} = \{3, 2, 1\} = \{1, 3, 2\},$$

which is another way of saying that the ordering of the elements within a set does not matter. Rule 3 allows us to write things like

$$A = \{1, 2, \{3, 4\}\},$$

which is a set containing *three* elements: the numbers 1 and 2, and the set $\{3, 4\}$. But, it does not allow us to write something like $A = \{A\}$, because this would lead to self-referential paradoxes. In symbols, Rule 3 forces the relation $A \notin A$ for all sets A . The fourth rule introduces the empty set, written $\emptyset$, which is the unique set containing nothing at all. This special set can be seen as a sort of zero element within set theory; a formal way to represent nothing. It should be noted that according to Rule 3, the empty set is not the same as the set containing the empty set:

$$\emptyset \neq \{\emptyset\},$$

since the first has zero elements and the second set has one element: $\emptyset$ itself. Finally, Rule 5 says that there is no set containing absolutely every mathematical object. This rule is imposed for technical reasons: if we had a set U that contained absolutely everything, then this set *is itself a thing*, meaning that by definition it ought to contain itself. But, this would contradict Rule 3.

Exercise: Equality of sets

Determine whether the following sets are equal.

- (a) $A = \{1, 3, 5, 7, 9\}$ and $B = \{9, 7, 5, 3, 1\}$.
- (b) $A = \{13, 6, 7, 9, 2\}$ and $B = \{6, 2, 13, 9\}$.
- (c) $A = \{2, 4, 6, 8\}$ and $B = \{2, 2, 2, 2, 4, 6, 8, 8\}$.
- (d) $A = \{14\}$ and $B = \{7 + 7\}$.

Solution

- (a) Yes. The two sets have exactly the same elements, so $A = B$.
- (b) No. The set A contains 7, but the set B does not. Therefore $A \neq B$.
- (c) Yes. We ignore repeated copies of elements, so $A = B$.
- (d) Yes. Since $7 + 7 = 14$, the two sets contain exactly the same element. Therefore $A = B$.

4.1.3 Cardinality

Since a set has no inherent structure, there is not much to talk about. However, what we can do is count up the elements contained within a set to get a notion of size. This “size of a set” has a special name: it is called the *cardinality* of a set.

Definition: Cardinality

For a finite set A , the **cardinality** of A is the number of elements in A . It is written

$$|A|.$$

For example,

$$|\{a, b, c, d, e\}| = 5, \quad |\emptyset| = 0.$$

4.1.4 Subsets

As the name may suggest, a subset of a set A is some kind of smaller collection of things taken from A . Formally, we can define a subset as follows.

Definition: Subset

For any two sets A and B , we say that A is a **subset** of B , written

$$A \subseteq B,$$

if every element of A is also an element of B .

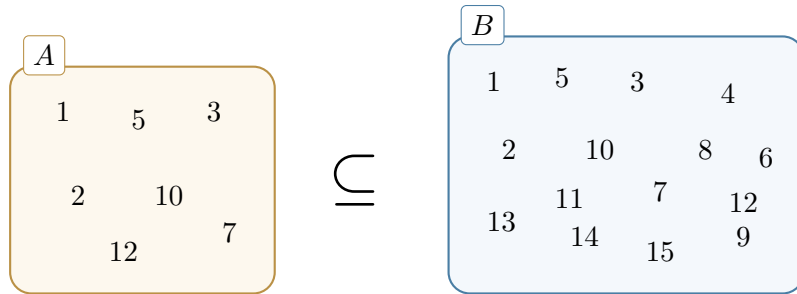
For example, the set

$$A = \{1, 2, 3, 5, 7, 10, 12\}$$

is a subset of the much larger set

$$B = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15\},$$

because every element contained in A also happens to be in B . In terms of structureless boxes, we can imagine a subset as follows.



The set A can be verified to be a subset of B by simply checking that each element of A is also present in B . Given that A has size 7, we can simply do this one-by-one.

There are two subset facts that are particularly important. For every set A , it is always the case that both

$$A \subseteq A \quad \text{and} \quad \emptyset \subseteq A.$$

The first statement says that every set is a subset of itself. This is always true, in some trivial sense: every element x in A is by default in A . The second statement above says that the empty set is a subset of every set, because there is no element of $\emptyset$ that could fail to be in A . This statement is often called *vacuously true*.

Exercise: Subsets

Suppose that

$$A = \{1, 2, 3, 4\}, \quad B = \{2, 3, 5, 7, 9\},$$

and let S be the set of all prime numbers. Determine whether the following statements are true or false.

- (a) $A \subseteq B$.
- (b) $A \subseteq S$.
- (c) $B \subseteq S$.
- (d) $\{2, 3\} \subseteq B$.
- (e) $A \subseteq A$.
- (f) $\emptyset \subseteq A$.

Solution

- (a) False. The elements 1 and 4 are in A , but they are not in B .
- (b) False. The numbers 1 and 4 are not prime.
- (c) False. The number 9 is in B , but 9 is not prime.
- (d) True. Both 2 and 3 are elements of B , so $\{2, 3\} \subseteq B$.
- (e) True. Every set is a subset of itself.
- (f) True. The empty set is a subset of every set.

4.1.5 The powerset

Suppose that we have a set A , with n -many elements:

$$A = \{x_1, x_2, \dots, x_n\}$$

We know from the previous section that we can form a subset of A by selecting some elements of A and putting them into a set, and then neglecting others if needs be. Starting from the bottom up, we know that for any element x in A , we can form subsets consisting of *only* this element: $\{x\} \subseteq A$. If we did this for every element of A , then we would create a lot of subsets of size 1:

$$\text{Subsets of size 1 : } \{x_1\}, \{x_2\}, \{x_3\}, \dots, \{x_n\}.$$

We can also try to list out all of the subsets of size 2. This would consist of all pairs of elements:

$$\text{Subsets of size 2 : } \{x_1, x_2\}, \{x_2, x_3\}, \{x_1, x_3\}, \dots$$

There appears to be many more of these subsets, but in principle they can all be written down. In fact, we can repeat this process to create subsets of size 3, 4, 5, and so on. Eventually this process will terminate, because A only has n elements, meaning that there is a unique subset of size n : A itself.

Hypothetically, after we were to complete this process, we would have written down *all* the subsets of A . To nobody's surprise, this collection can also be made into another set. We call this "set of all subsets of A " the *powerset* of A .

Definition: Powerset

The **powerset** of a set A , written $\mathcal{P}(A)$, is the set of all subsets of A .

For example, if

$$A = \{1, 2, 3\},$$

then

$$\mathcal{P}(A) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}.$$

As you can see, the powerset seems to be much larger than the original. In the case above, the set $A = \{1, 2, 3\}$ has 3 elements, whereas its powerset has 8 elements. This is not a coincidence: it happens to be the case that $2^3 = 8$. Generally, we have the following rule.

Size of a powerset

If a finite set A has n elements, then

$$|\mathcal{P}(A)| = 2^n.$$

We will justify this rule later in the lecture.

4.2 Operations on Sets

4.2.1 Building new sets

There are a few ways to build new sets from old ones. Within the context of probability theory, we will focus on three main operations:

1. the complement of a set,
2. the intersection of sets, and
3. the union of sets.

As we will see in next lecture, these three operations will allow us to properly phrase (and therefore calculate) probabilities of events containing the words “not”, “and” and “or”. We will now review each of these operations briefly.

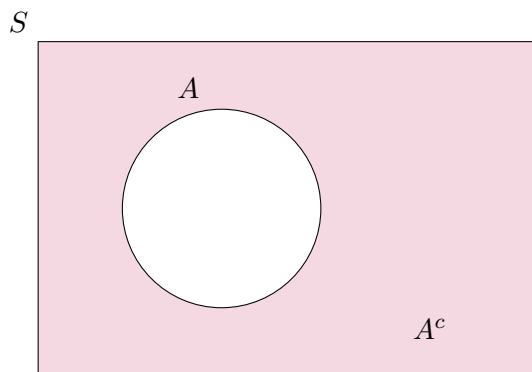
4.2.2 The Complement

The complement of A , sometimes written A^c , is simply the set of all things not in A . According to Rule 5, we cannot simply look at *all* the things not in A . Instead, we need to take these complements relative to some larger set S .

Definition: Complement

Let $A \subseteq S$. The **complement** of A in S is the set of all elements x in S that are not in A .

Visually, we can imagine the complement like the shaded area below:



For example, consider the set $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. If we let A be the set of even numbers in S , i.e. $A = \{2, 4, 6, 8, 10\}$, then the complement of A in S will be all those numbers in S that are not in A . We can find these by writing out S and removing all the elements contained in A . Whatever is left will be the complement. Writing out S and removing the elements of A , we have

$$\{1, \cancel{2}, 3, \cancel{4}, 5, \cancel{6}, 7, \cancel{8}, 9, \cancel{10}\} \longrightarrow \{1, 3, 5, 7, 9\}.$$

Therefore the complement A^c equals the set $\{1, 3, 5, 7, 9\}$, i.e. the odd numbers in S .

4.2.3 The Intersection

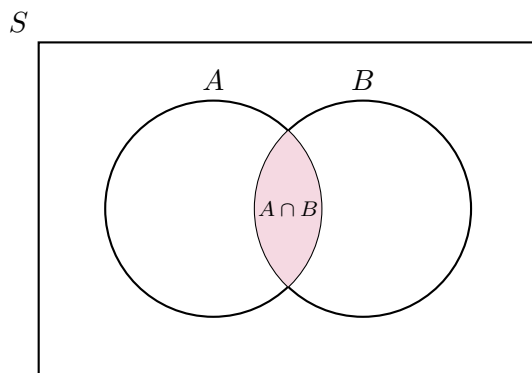
The intersection of A and B , sometimes written $A \cap B$, is the set of all things that are in A and B at the same time.

Definition: Intersection

The **intersection** of A and B is

$$A \cap B = \{x : x \in A \text{ and } x \in B\}.$$

Visually, the intersection of two sets might look like the shaded region below:



However, it should be noted that it's entirely possible for two sets to have an empty intersection. This would happen if the two circles in the diagram above did not overlap.

As an example, let $A = \{1, 2, 3, 4\}$ and $B = \{2, 3, 5, 7, 9\}$. For the intersection, we are looking for any elements that are in A and B at the same time. By inspection, we see that

$$A \cap B = \{2, 3\},$$

because 2 and 3 are the only elements that are in both sets.

4.2.4 The Union

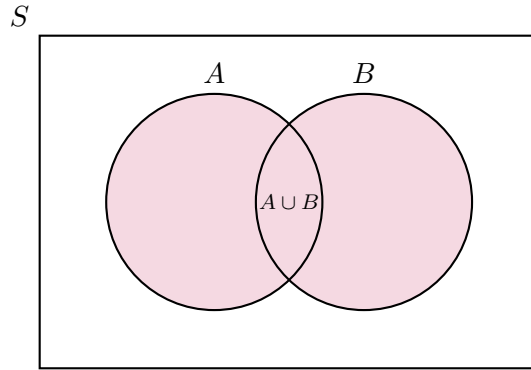
The union of A and B , sometimes written $A \cup B$, is the set of all things that are either in A or in B , or both. We can imagine this set as a set formed by joining together both A and B into one set.

Definition: Union

The **union** of A and B is

$$A \cup B = \{x : x \in A \text{ or } x \in B\}.$$

Visually, the union of two sets may look like the diagram below:



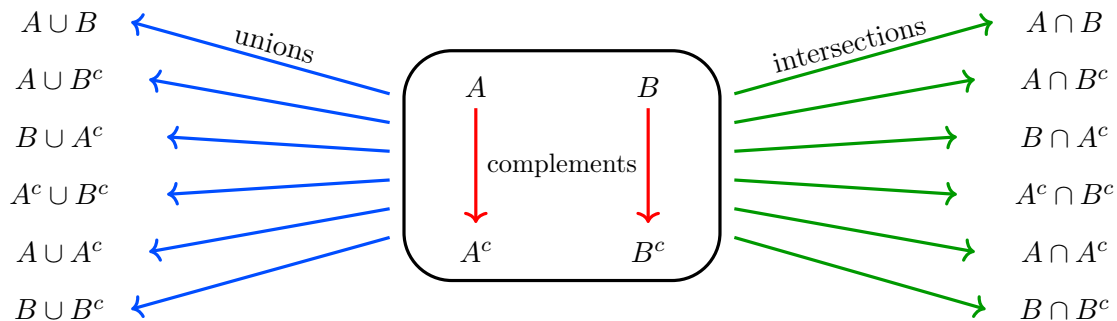
For example, suppose that $A = \{1, 2, 3\}$ and $B = \{2, 3, 4, 5\}$. Then the union will be the collection of all of these things together:

$$A \cup B = \{1, 2, 3, 4, 5\}.$$

Note that according to Rule 1 of Section 1.2, we **don't** write the elements 2 and 3 twice.

4.2.5 Combining operations

Using the three operations of complements, intersections and unions, we can build many other sets from a small amount of starting information. From just two sets A and B , both viewed as subsets of some larger set S , we can apply these three operations to create many new sets:



Exercise: Operations on sets

Let

$$S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\},$$

and consider the three subsets

$$A = \{2, 4, 6, 8, 10\}, \quad B = \{2, 3, 5, 7, 9\}, \quad C = \{1, 3, 5, 7, 9\}.$$

Write out the following sets.

- (a) A^c
- (b) $A \cap B$
- (c) $(A \cap B) \cup C$
- (d) $A \cap C$
- (e) $A \cap B \cap C$

Solution

- (a) Since A is the set of even numbers in S , its complement is

$$A^c = \{1, 3, 5, 7, 9\}.$$

- (b) The only element in both A and B is 2, so

$$A \cap B = \{2\}.$$

- (c) Since $A \cap B = \{2\}$, we have

$$(A \cap B) \cup C = \{2\} \cup \{1, 3, 5, 7, 9\} = \{1, 2, 3, 5, 7, 9\}.$$

- (d) There are no elements in both A and C , so

$$A \cap C = \emptyset.$$

- (e) Since $A \cap C = \emptyset$, there cannot be anything in all three sets. Therefore

$$A \cap B \cap C = \emptyset.$$

4.2.6 Cardinality and operations

Set operations interact with cardinality in useful ways. Some basic facts are:

$$|A \cap B| \leq |A| \quad \text{and} \quad |A \cap B| \leq |B|.$$

This is because the intersection only contains elements that were already in both sets.

Similarly,

$$|A| \leq |A \cup B| \quad \text{and} \quad |B| \leq |A \cup B|.$$

This is because the union contains everything from A and everything from B .

However, when counting the size of $A \cup B$, we must avoid double-counting the overlap. This gives the formula

$$|A \cup B| = |A| + |B| - |A \cap B|.$$

Example. Let

$$A = \{1, 2, 3, 4\}, \quad B = \{3, 4, 5, 6, 7\}.$$

Then

$$|A| = 4, \quad |B| = 5, \quad A \cap B = \{3, 4\}.$$

So

$$|A \cap B| = 2.$$

Therefore

$$|A \cup B| = |A| + |B| - |A \cap B| = 4 + 5 - 2 = 7.$$

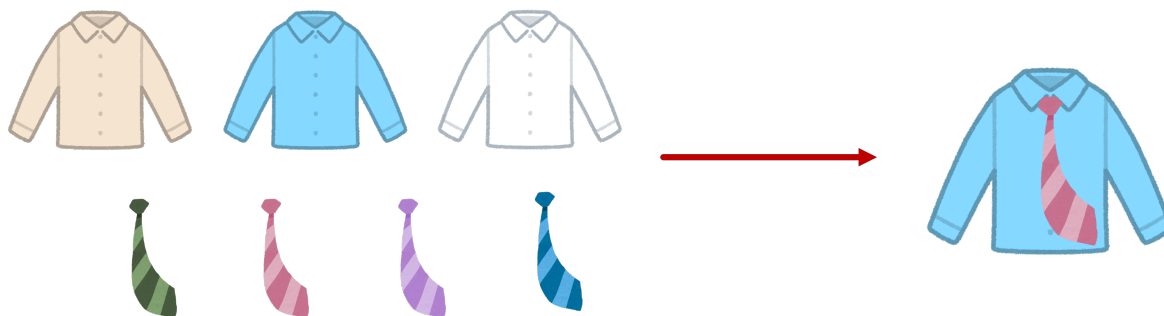
Indeed,

$$A \cup B = \{1, 2, 3, 4, 5, 6, 7\}.$$

4.3 Counting Configurations

4.3.1 The multiplication rule

Suppose that you have 3 shirts and 4 ties, and you want to wear one of each to go to work this morning. For example, one possible choice could be the following, the blue shirt and the red tie:



How many possible outfits can we make? Our problem of selecting an outfit can be broken down into two separate choices: selecting a shirt, and then selecting a tie independently. Picking a shirt first then offers us four possible outfits: the shirt plus any one of the four possible ties. Running through these choices for all the possible shirts, we end up with the following structured grid, which displays all possible configurations of shirt and tie.



As you can see, there are 12 total possible outfits, which is precisely the number of shirts (3) multiplied by the number of ties (4).

Instead of drawing pretty pictures of clothes, we can also convert the problem into mathematical language. If we label the three shirts by s_1, s_2, s_3 , and we write the four ties as t_1, t_2, t_3, t_4 , then an outfit would consist of a *pair* (s_i, t_j) . The diagram above can be rewritten:

	t_1	t_2	t_3	t_4
s_1	(s_1, t_1)	(s_1, t_2)	(s_1, t_3)	(s_1, t_4)
s_2	(s_2, t_1)	(s_2, t_2)	(s_2, t_3)	(s_2, t_4)
s_3	(s_3, t_1)	(s_3, t_2)	(s_3, t_3)	(s_3, t_4)

3 shirt choices $\times$ 4 tie choices = 12 total configurations.

This grid-like structure of pairs is known as a *Cartesian product*. If we have two sets: $S = \{s_1, s_2, s_3\}$ and $T = \{t_1, t_2, t_3, t_4\}$ consisting of the shirts and ties, respectively, then we can “multiply” these two sets together to create the set of all pairs of the form (s_i, t_j) . This new set, the Cartesian product of S and T , is written $S \times T$. What we have just discovered with our example of shirts and ties is that $|S \times T| = |S| \times |T|$. Within the context of counting configurations, we have the following general rule.

Multiplication Rule for Configurations

Suppose that a choice is made in stages. If the first stage has m options and the second stage has n options, then the total number of configurations is

$$m \times n.$$

For more than two stages, multiply the number of options at each stage.

Exercise: Counting configurations

You go to your favourite curry restaurant. They have a set deal where you choose:

- one of 5 starters,
- one of 15 main dishes,
- one of 5 desserts,
- one of 10 drinks.

How many unique meals can you create in total?

Solution

There are 5 choices for the starter, 15 choices for the main dish, 5 choices for the dessert, and 10 choices for the drink. By the multiplication rule, the total number of unique meals is

$$5 \times 15 \times 5 \times 10 = 3750.$$

So, there are 3750 unique meals in total. This means that you could visit every day for more than 10 years and never have the same meal twice.

4.4 Counting Permutations

In a moment we will introduce permutations and start counting them. However, in order to get there, we first need to introduce a new piece of notation.

Definition: Factorial

For a positive whole number n , the **factorial** of n is

$$n! = n(n-1)(n-2) \cdots 3 \cdot 2 \cdot 1.$$

We also define

$$0! = 1.$$

For example,

$$4! = 4 \times 3 \times 2 \times 1 = 24,$$

and

$$10! = 10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 3628800.$$

We stop at 1 because if we multiplied by 0, then every factorial would equal 0, which would be useless.

4.4.1 What is a permutation?

As we have seen in Section 1, a set of objects $A = \{a_1, a_2, \dots, a_n\}$ does not care about the order that we write its elements in. However, we *can* decide to care about this, and write the elements of the set A out in different orders. To distinguish from the curly brackets of sets (which don't care about the order), we are going to write *ordered* lists using round brackets (and), so that, for example:

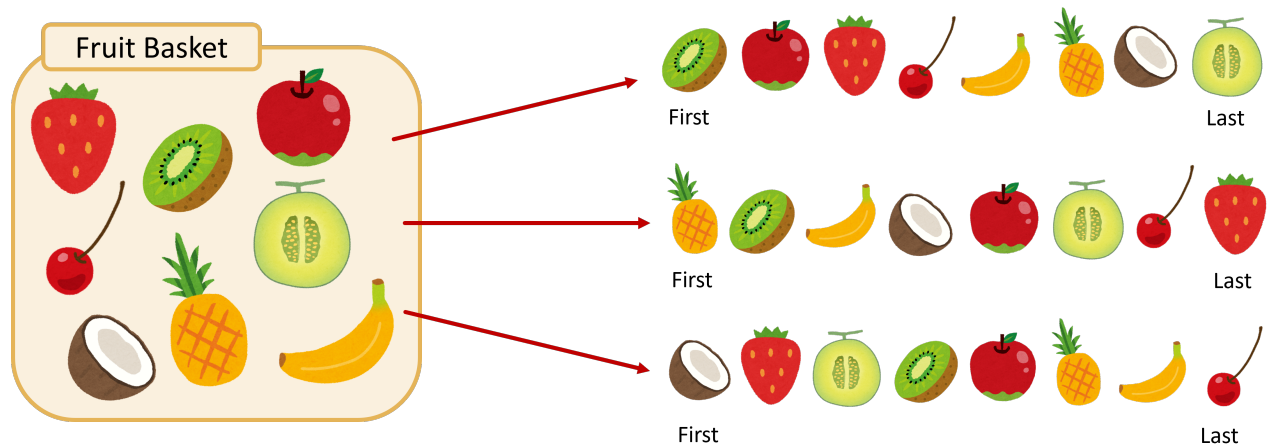
$$(a_1, a_2, \dots, a_n) \quad \text{and} \quad (a_2, a_1, \dots, a_n)$$

are considered different, even though they consist of the same n elements. Very simply, a permutation may be defined as follows.

Definition: Permutation

A **permutation** is an arrangement of objects where order matters.

As an example, suppose that we had a fruit basket consisting of 8 distinct pieces of fruit. If we were feeling particularly hungry, then we might want to eat every single piece of fruit in the basket. Since we have to eat them one at a time, there are many different possible orders. Three possible orders are listed below.



Three possible permutations of the 8 pieces of fruit.

4.4.2 Counting permutations

In order to try and understand how many permutations a given collection has, we can start with small numbers and see if we can notice a pattern. So, we start with $n = 1$ and keep increasing it:

$$n = 1 : \quad 1 \text{ way,}$$

$$n = 2 : \quad 2 \times 1 = 2 \text{ ways,}$$

$$n = 3 : \quad 3 \times 2 \times 1 = 6 \text{ ways,}$$

$$n = 4 : \quad 4 \times 3 \times 2 \times 1 = 24 \text{ ways.}$$

Based on this, we see an obvious pattern emerging: we ought to use $n!$ to count the number of permutations of n objects. In fact, throughout this course we will use a slightly more general formula, which is given below.

General Rule for Permutations

The number of ways to arrange r objects in order from a collection of n distinct objects is

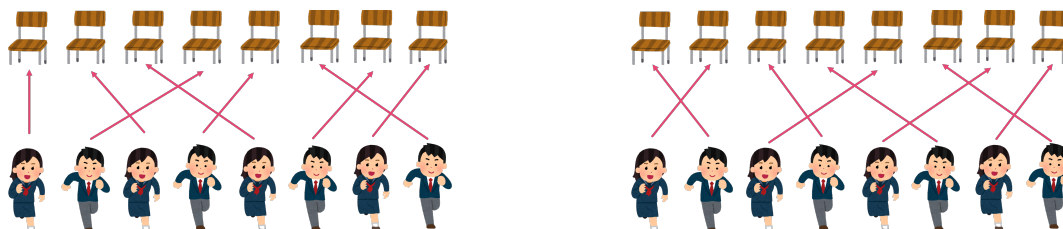
$$P_{n,r} = \frac{n!}{(n-r)!},$$

where n and r are whole numbers and $n \geq r$.

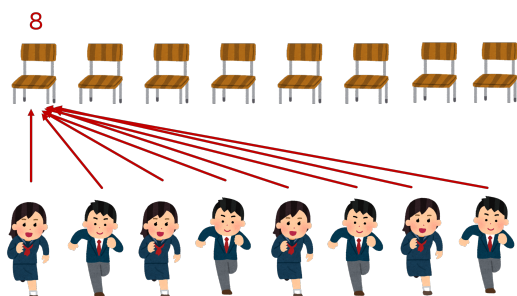
Upon first glance, this formula may appear to be quite random. So, we will now carefully justify it with an in-depth example.

4.4.3 Explaining the formulas using 8 students and 8 chairs

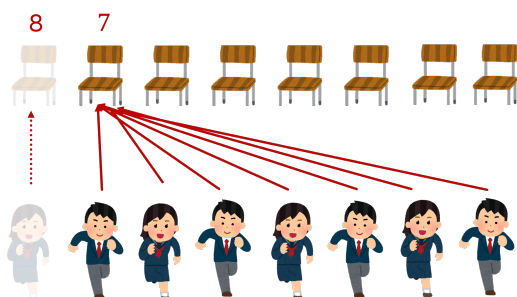
Suppose that we have a class with 8 students, and there are 8 chairs. We would like to create a “seating arrangement” by assigning the 8 students to the 8 chairs in the room. For example, the image below contains two possible seating arrangements.



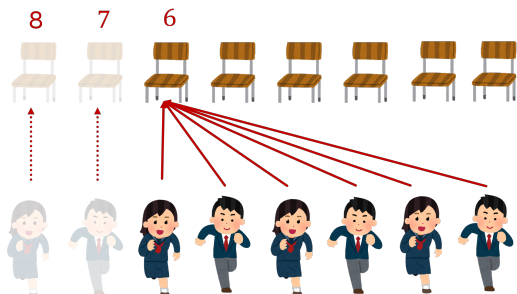
In effect, each assignment is the same as a permutation: we are taking the set of students and saying “you sit down first, you sit down second” and so on. So, the total number of seating arrangements ought to be $8! = 40320$. We will now explain this step-by-step.



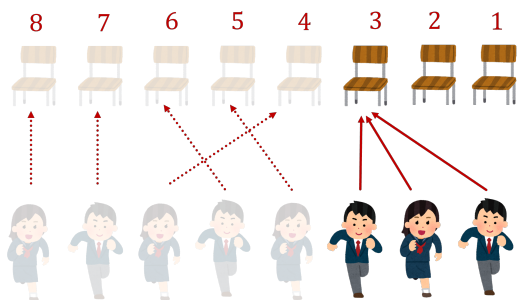
Step 1: The first chair. Since we start with 8 students, there are 8 possible choices for who can sit in the first chair.



Step 2: The second chair. When we try to repeat the process again for the second chair, we now only have 7 students and 7 empty chairs left. So, instead of there being 8 possible choices, there are only 7.



Step 3: The third chair. After we choose one of the 7 students to sit down in the second chair, we can then move on to the third. In this case, we have now already assigned *two* students, meaning that there are only 6 students who can possibly sit in the third chair. In other words, there are now 6 possible choices for the third chair.



Step 4: Continue the pattern. We may keep repeating this reasoning over and over again to get 5, then 4, then 3, then 2, and finally 1 possible choice for the remaining chairs.

In order to get the total number of possible configurations, we may use the multiplication rule for this 8-step process to conclude that the total number of seating arrangements is

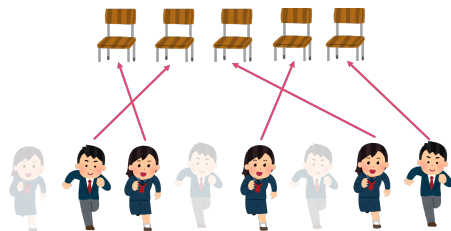
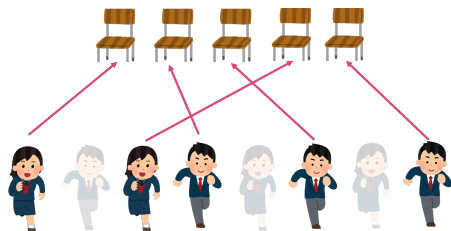
$$8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 8!.$$

This recovers the factorial formula mentioned previously. But, what about the formula

$$P_{n,r} = \frac{n!}{(n-r)!}$$

mentioned previously? We will now explain this by modifying our setup.

Suppose now that we remove 3 chairs from the classroom, meaning that we only have 5 chairs remaining. So, if we try to sit 8 students down, then only 5 of them will be able to sit down and the other 3 will have to stand. Our next question is: what is the number of possible ways that we can sit 5 of the 8 students down? Two such arrangements are depicted below.



We can repeat the same reasoning as before, so we see that the first chair has 8 choices, the second chair has 7 choices and so on. However, since there are now only 5 possible chairs to fill, we must stop our procedure *early*, after 5 chairs. This means that our answer will not be $8!$, but instead it will be given by the first 5 terms in the derivation from before:

$$P_{8,5} = 8 \times 7 \times 6 \times 5 \times 4.$$

Here $r = 5$, and we observe that we can write the same formula as:

$$P_{8,5} = \frac{8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{3 \times 2 \times 1}.$$

This is the same formula as that mentioned previously. The idea here is that the numerator of $P_{n,r}$ tells us how to order n total objects, and the denominator $(n - r)!$ tells us how to order the $(n - r)$ -many objects that we did not select. We divide by this number in order to remove them from the count. In the case of the chairs above, the $(n - r)!$ term corresponds to the 6 ways to order the 3 students who remain standing.

4.5 Counting Combinations

When discussing permutations, we cared about the order in which we arranged the students. That is, we lined up all the chairs in order and assigned different students by saying “you go first, you go second, ...”. We can imagine that there might be another way to do things: more generally, perhaps we could simply select r -many students and not care in which order they are placed in chairs. This would amount to simply choosing 5 students from the collection of 8. Two such choices are pictured below.



In fact, we can also count up “all the possible ways to choose r -many objects from n -many”. These choices aren’t permutations, because in this situation we don’t care about their order. Instead, the choices are just picking some members of the group.

As a matter of fact, there is a formula for all the possible choices of r -many objects amongst a collection of n -many. In mathematics, we call all of these unordered choices “combinations”, and we use the notation $C_{n,r}$ to denote the number of ways to choose r -many things from a collection of size n . The formula for $C_{n,r}$ is very similar to that of $P_{n,r}$, and it is given below.

General Rule for Combinations

The number of combinations of n objects taken r at a time is

$$C_{n,r} = \frac{n!}{r!(n - r)!},$$

where n and r are whole numbers and $n \geq r$.

Notice that here we have the formula for $P_{n,r}$ but we divide by an extra $r!$ term. Why do we do this? Well, in the case of $P_{n,r}$ we are asking for ordered selections, and the same choice of r -many objects can be ordered in $r!$ many different ways. However, for unordered combinations, we don’t actually care about those $r!$ many things, so we divide the overall formula by $r!$ to remove all of those choices from our count.

To illustrate the formula, we may calculate the number of ways to select 5 students from the group of 8:

$$C_{8,5} = \frac{8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{(5 \cdot 4 \cdot 3 \cdot 2 \cdot 1)(3 \cdot 2 \cdot 1)}.$$

Of course, this can be put directly into a calculator. However, we can also observe that a factor of $5!$ exists on the top and bottom of this particular fraction. Therefore, we can cancel these to get:

$$C_{8,5} = \frac{8 \cdot 7 \cdot 6}{3 \cdot 2 \cdot 1} = \frac{336}{6} = 56.$$

So, there happen to be 56 ways to choose 5 students from 8 students.

4.5.1 Pascal's Triangle

As a matter of fact, the formulas $C_{n,r}$ happen to satisfy some interesting mathematical properties. We will start with two very simple observations that are worth remembering:

$$C_{n,0} = 1 \quad \text{and} \quad C_{n,n} = 1.$$

These equalities can be easily verified by inspecting the formulas algebraically. We have

$$C_{n,0} = \frac{n!}{0!(n-0)!} = \frac{n!}{0!n!} = 1,$$

because $0! = 1$, and similarly,

$$C_{n,n} = \frac{n!}{n!(n-n)!} = \frac{n!}{n!0!} = 1.$$

Heuristically speaking, there is a clear interpretation: there is exactly one way to select nothing from n objects, and there is also exactly one way to select everything.

Perhaps more interestingly, for $1 \leq r \leq n$, these $C_{n,r}$ formulas also satisfy the identity

$$C_{n,r-1} + C_{n,r} = C_{n+1,r},$$

which is commonly called *Pascal's Identity*. For the unconvinced, we may verify this algebraically.

Algebraic verification. We calculate:

$$C_{n,r-1} + C_{n,r} = \frac{n!}{(r-1)!(n-r+1)!} + \frac{n!}{r!(n-r)!}.$$

To add these fractions, use the common denominator

$$r!(n-r+1)!.$$

Then

$$\begin{aligned} C_{n,r-1} + C_{n,r} &= \frac{rn!}{r!(n-r+1)!} + \frac{(n-r+1)n!}{r!(n-r+1)!} \\ &= \frac{rn! + (n-r+1)n!}{r!(n-r+1)!} \\ &= \frac{(n+1)n!}{r!(n-r+1)!} \\ &= \frac{(n+1)!}{r!((n+1)-r)!} \\ &= C_{n+1,r}. \end{aligned}$$

There is also a simple counting explanation. To choose r objects from $n + 1$ objects, focus on one special object. Either the special object is chosen, or it is not chosen.

- If the special object is chosen, then we still need to choose $r - 1$ objects from the other n objects. This gives $C_{n,r-1}$ possibilities.
- If the special object is not chosen, then we need to choose all r objects from the other n objects. This gives $C_{n,r}$ possibilities.

Together, these two cases give

$$C_{n,r-1} + C_{n,r} = C_{n+1,r}.$$

For low values of n and r , we can arrange the values of $C_{n,r}$ into a triangle. At the top of the triangle is $C_{0,0} = 1$, the next row is $C_{1,0} = 1$ and $C_{1,1} = 1$. But, things start to get interesting once we add more rows. In the figure below, the values of $C_{n,r}$ are written in the form of a triangle.

Row 0:			1									$C_{0,0}$							
Row 1:			1		1							$C_{1,0}$		$C_{1,1}$					
Row 2:			1		2		1					$C_{2,0}$		$C_{2,1}$		$C_{2,2}$			
Row 3:			1		3		3		1			$C_{3,0}$		$C_{3,1}$		$C_{3,2}$		$C_{3,3}$	
Row 4:		1		4		6		4		1		$C_{4,0}$		$C_{4,1}$		$C_{4,2}$		$C_{4,3}$	$C_{4,4}$
Row 5:	1		5		10		10		5		1	$C_{5,0}$	$C_{5,1}$	$C_{5,2}$	$C_{5,3}$	$C_{5,4}$	$C_{5,5}$		
$\vdots$					$\vdots$											$\vdots$			

Observe that each interior number in the triangle is given by the sum of the two numbers above it, this is Pascal's Identity at work. For example, the “6” in Row 4 (coloured red) is given by the sum of the “3” and the “3” that are above it in Row 3. By applying this addition rule, we can form as many new rows of the triangle as we like.

Observe that each row is numbered by the “ n ” in $C_{n,r}$, and the entries for that row are given by running across the different values of r between 0 and n . This means that we don't need to calculate the value of $C_{n,r}$ directly, instead we can simply read off its value using the triangle.

4.5.2 Counting subsets with combinations

There is an important connection between combinations and powersets. Suppose A is a set with n elements. If we want to count the number of subsets of A with exactly r elements, then we are choosing r elements from n elements, and order does not matter. Therefore, the number of subsets of size r is

$$C_{n,r}.$$

Example. Let

$$A = \{a, b, c, d\}.$$

This set has size 4. We can count its subsets according to their size:

- $C_{4,0} = 1$ counts the single subset of size 0, namely $\emptyset$.
- $C_{4,1} = 4$ counts the subsets $\{a\}, \{b\}, \{c\}, \{d\}$.
- $C_{4,2} = 6$ counts the subsets $\{a, b\}, \{a, c\}, \{a, d\}, \{b, c\}, \{b, d\}, \{c, d\}$.
- $C_{4,3} = 4$ counts the subsets $\{a, b, c\}, \{a, b, d\}, \{a, c, d\}, \{b, c, d\}$.
- $C_{4,4} = 1$ counts the single subset of size 4, namely A itself.

Adding these together gives

$$1 + 4 + 6 + 4 + 1 = 16 = 2^4.$$

This agrees with the powerset rule:

$$|\mathcal{P}(A)| = 2^4 = 16.$$

This connection also explains why the rows of Pascal's triangle add to powers of 2.

$$\begin{array}{rcl}
 \text{Row 0:} & & 1 = 1 \\
 \text{Row 1:} & & 1 + 1 = 2 \\
 \text{Row 2:} & & 1 + 2 + 1 = 4 \\
 \text{Row 3:} & & 1 + 3 + 3 + 1 = 8 \\
 \text{Row 4:} & & 1 + 4 + 6 + 4 + 1 = 16 \\
 \text{Row 5:} & & 1 + 5 + 10 + 10 + 5 + 1 = 32 \\
 & \vdots & \vdots \qquad \qquad \vdots
 \end{array}$$

Lecture 5

Introduction to Probability Theory

In Lecture 4, we introduced the set theory and counting methods that are needed for probability theory. Today, we will use those ideas to describe probability mathematically. The central idea is that an event is represented by a subset of possible outcomes, and the probability of that event is a number between 0 and 1. When working with classical probability theory, questions about uncertainty can be turned into questions about counting. Throughout this lecture we will explore this idea in depth.

5.1 What is Probability?

5.1.1 Uncertainty

In the first few lectures of this course, we introduced the basic language of data and statistics. We saw that data is usually collected from messy real-world situations, and then converted into something that can be analysed mathematically. Today we will deal with another kind of mess: uncertainty.

Consider the following two statements:

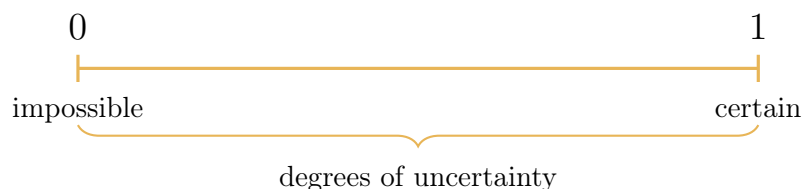
$$2 + 2 = 4 \quad \text{and} \quad 2 + 2 = 5.$$

As you probably know, the first statement is true and the second statement is false. Not only is this obvious, it is *unequivocal*: there is no sense in which the statements could be otherwise. In contrast, consider the two statements:

“Dr. O’Connell’s middle name is John”, and “Dr. O’Connell’s middle name is Jennifer”.

As you can see, these non-mathematical statements aren’t so clear anymore: if you happened to not know Dr. O’Connell, then you would be uncertain of his middle name. However, even though you don’t know my middle name, you can still imagine that having a middle name of Jennifer would be unusual for an Englishman, and having a middle name of John is more plausible. So, although you cannot definitely point to the exact truth of these statements, you can still compare their plausibility.

Probability theory tries to capture this uncertainty using numbers. In reality, the world is full of uncertain things that are not purely logical, and probability theory gives us a way to quantify these beliefs so that they can be compared. If we interpret the number 0 to mean “definitely impossible” and the number 1 to mean “absolutely certain”, then all of our uncertain statements would exist somewhere in between:



Basic idea

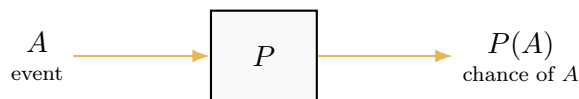
Probability is the mathematics of uncertainty. It gives us a way to assign numbers to events, so that we can compare how likely those events are.

It should be noted that probability is **not** a spectrum of truth.¹ Instead, it is a spectrum of certainty. So, a probability value of 0.5 does not mean that something is “half true”, instead it means that we assign a 50% chance to the statement being true.

¹For the curious: there is actually a logical theory that does this, called “fuzzy logic”. But this is **not** what probability is.

5.1.2 Defining probability

Mathematically, probability is a function that assigns a number to an event. If the letter A represents something that may or may not happen, then we write $P(A)$ for the probability that A happens. In this sense, we can view the letter P as a little machine that eats events in the world, and spits out a numerical measure of their certainty:



Generally, we refer to these “things that can possibly happen” as *events*, and we denote them with capital letters such as A, B, C and so on. Formally, the probability assigned to an event can be summarised as follows.

Probability values

For any event A , its probability satisfies

$$0 \leq P(A) \leq 1.$$

Here $P(A) = 0$ means that A is impossible, while $P(A) = 1$ means that A is certain.

Since probabilities are mapped somewhere onto the line between 0 and 1, we can directly compare degrees of uncertainty of different events. Suppose that we have two events A and B , and we happened to know their probabilities $P(A)$ and $P(B)$. If $P(A) < P(B)$, then event B is more likely than event A . For example, if

$$P(A) = 0.2 \quad \text{and} \quad P(B) = 0.7,$$

then B is more likely to happen than A , because $P(B)$ is closer to the absolute certainty of 1 than $P(A)$ is.

Expressing probabilities

Since any probability is a number between 0 and 1, we can write it down in several equivalent ways. For example, if the probability of an event A is 50%, we could write it in any of the following ways:

$$P(A) = \frac{1}{2} = 0.5 = 50\% = \frac{250}{500}.$$

Here $\frac{1}{2}$ is a reduced fraction, 0.5 is a decimal, 50% is a percentage, and $\frac{250}{500}$ is an unreduced fraction. In the last case, perhaps we are working with a system that has 500 possibilities, and 250 of the possible outcomes are favourable. The unreduced fraction $\frac{250}{500}$ tells us this information, but the reduced fraction $\frac{1}{2}$ does not.

5.1.3 Classical probability

In this lecture, we mostly focus on *classical probability*. This is the setting in which there are finitely many possible outcomes and each one is equally likely.

Definition: Classical probability

A probability problem is a problem in **classical probability** when there are finitely many possible outcomes and each outcome is equally likely.

Examples include:

- rolling a fair dice,
- flipping a fair coin,
- dealing cards from a well-shuffled deck,
- selecting answers to true/false questions.

In this setting, probability can usually be calculated by counting.

A note on randomness

The question of true randomness is deeply philosophical. Some people believe the world is totally deterministic, meaning everything is determined from past events and the laws of physics. Other people believe there are genuinely random things, for example observations in quantum mechanics, but even there, things are debated. In any case, it does not matter for our purposes whether true randomness exists or not. In this course we use probabilities to quantify uncertainty. So, even if things are not truly random in reality, *they may still appear random to us*.

Randomness in this course

For our purposes, an outcome is random if we cannot predict it with certainty before the experiment is performed.

This practical interpretation is enough for the kind of probability calculations we will do in this course.

5.1.4 Assigning probabilities by counting

When working with classical probability, we can calculate the probabilities of individual events by taking a ratio of the “positive cases” compared to the collection of all possibilities. For any event, we can assign a probability using the formula:

$$\text{Probability of an event} = \frac{\text{number of outcomes favourable to the event}}{\text{total number of outcomes}}.$$

Put differently, we count the number of outcomes that would cause the event to be true, and then divide by the total number of possible outcomes. In this fraction, the numerator will always be less than or equal to the denominator, so when we take the ratio we will always obtain a value between 0 and 1.

The Golden Formula for Classical Probabilities

If all outcomes are equally likely, then

$$P(A) = \frac{\text{number of outcomes favourable to } A}{\text{total number of outcomes}}.$$

We will now demonstrate how to use this formula to calculate some basic probabilities.

The Probability of an Individual Outcome

The probability of any individual outcome can be determined with this formula. For example:

- for a fair coin, the probability of the coin landing on heads is $\frac{1}{2}$, because there are only two possibilities, one of which is heads.
- When rolling a fair dice, the probability of landing on a 5 will be $\frac{1}{6}$, because there is only one 5 written on the dice, and there are 6 numbers in total.
- When randomly drawing a card from a shuffled deck, the probability of picking the ace of spades is $\frac{1}{52}$, because there are 52 cards in a deck, and only one of them is the ace of spades.

Generally, for any system in classical probability, the formula above tells us:

$$\text{Probability of any individual outcome} = \frac{1}{\text{total number of outcomes}}.$$

The probability of a more-complicated event

We can also use our Golden Formula to calculate more complicated probabilities. As previously mentioned, this can be achieved by looking through the space of possibilities and counting up all of the outcomes that make the event true. As an example, suppose that we are rolling a fair dice. Let the event A be:

$$A = \text{“we land on an even number”},$$

and let the event B be:

$$B = \text{“we land on a number less than 5”}.$$

To determine these probabilities, we need to calculate the number of “positive outcomes” for each, and then divide by 6, since there are 6 sides of the dice. For the event A we have:

$$P(A) = \frac{\text{how many numbers are even}}{\text{how many numbers there are in total}} = \frac{3}{6} = \frac{1}{2},$$

and similarly for B , the probability is:

$$P(B) = \frac{\text{how many numbers are less than 5}}{\text{how many numbers there are in total}} = \frac{4}{6} = \frac{2}{3}.$$

Example: the Toy Cabinet

Suppose that your overly enthusiastic cousin has a cabinet with 58 different toys inside, and he picks one at random to show you. Before he does this, you do not know which one he will pick, so from your perspective his choice is essentially random. Observe the cabinet below:



Let event A be: “the toy chosen is a blue car”. By examining the above image, we see that there are two blue cars in the cabinet. Therefore, according to our Golden Formula

$$P(A) = \frac{2}{58}.$$

Let event B be: “the toy chosen comes from the fourth row”. In the image above, we count and see that the fourth row contains 8 toys, so event B contains 8 outcomes. Therefore:

$$P(B) = \frac{8}{58}.$$

The calculation is the same in both cases: count the favourable toys, then divide by the total number of toys.

Exercise: True/false answer sequences

Suppose that three homework questions are true/false questions. To choose the correct answers randomly, Dr. O’Connell writes down all possible true/false sequences and chooses one from a hat. There are 8 possible sequences of 3 T/Fs in a row:

TTT, TTF, TFT, FTT, TFF, FTF, FFT, FFF

- (a) What is the probability that all 3 answers are False?
- (b) What is the probability that exactly two answers are False?
- (c) If a student blindly guesses all three answers, what is the probability that they get all 3 correct?

Solution

These three questions can all be answered using the Golden Formula mentioned previously. Note first that there are 8 possible true/false sequences.

(a) Only one outcome is favourable, namely FFF. Therefore

$$P(\text{all 3 False}) = \frac{1}{8}.$$

(b) The favourable outcomes are TFF, FTF and FFT. Therefore

$$P(\text{exactly two False}) = \frac{3}{8}.$$

(c) There is exactly one correct sequence, so

$$P(\text{all 3 correct}) = \frac{1}{8}.$$

5.2 The Rules of Probability

5.2.1 Simple events, sample spaces and events

We now organise the vocabulary of probability more carefully. The most important objects are simple events, the sample space and events.

Important definitions

- (1) A **simple event** consists of exactly one outcome of the experiment.
- (2) The **sample space**, usually written S , is the set of all possible outcomes.
- (3) An **event** is any subset of the sample space, meaning it can contain one outcome or many outcomes (or none at all).
- (4) The probability of an event A is written $P(A)$. This is a number between 0 and 1.
- (5) The sum of the probabilities of all simple events is always 1.

To see these definitions in action, let us return to the toy cabinet example. In this case, the sample space will be the set of all 58 toys in the cabinet:

$$S = \{\text{all toys in the cabinet}\}.$$

In this example, a simple event would be the selection of an individual toy, something like “picking this toy in particular”. Mathematically, this would be any subset of S that is of size one. A general event would be any subset of S at all, and these subsets would represent more general statements such as “picking a toy from the fourth row” or “picking a blue car”.

Just to emphasise this: the event *is* the subset of the sample space. For example, in the toy cabinet example above, we considered the event A meaning “selecting a blue car”, and the probability was determined to be $\frac{2}{58}$. The mathematical definition of the event A would be the set containing two objects: the two blue cars taken from the cabinet.

5.2.2 Probability is secretly set theory

Let's consider another example. Suppose that Dr. O'Connell has 11 students in his class. He lines them up in two rows: 6 students in the front row, and 5 students in the back row:



For argument's sake, what would be the probability that Dr. O'Connell randomly selects somebody from the first row? Well, assuming that each choice were equally likely, according to our golden formula:

$$P(A) = \frac{\text{number of students in the front row}}{\text{number of students in total}} = \frac{6}{11}.$$

In terms of the terminology from before, the sample space S corresponds to the set of all students, and the event A would be the subset consisting of those students in the first row:

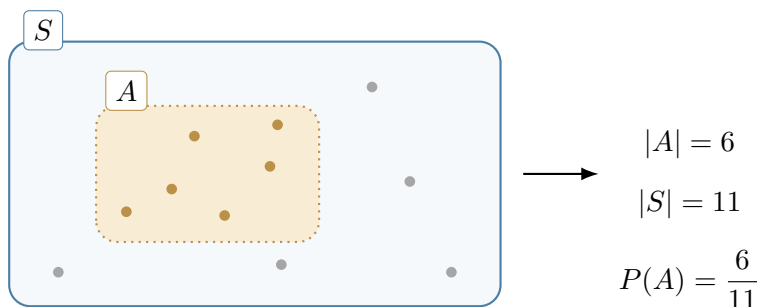
$$A = \{ \text{student 1}, \text{student 2}, \text{student 3}, \text{student 4}, \text{student 5}, \text{student 6} \}$$

$$S = \{ \text{student 1}, \text{student 2}, \text{student 3}, \text{student 4}, \text{student 5}, \text{student 6}, \text{student 7}, \text{student 8}, \text{student 9}, \text{student 10}, \text{student 11} \}$$

By a quick count, we see that $|A| = 6$ and $|S| = 11$. Taking the ratio of these returns the probability $P(A)$. So, in set-theoretic language, our golden formula now reads:

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

This says that to calculate the probability of an event A , we count the size of the subset A and divide by the size of the sample space S . More abstractly still, we are performing a count like the diagram below.



For future reference, we will restate our formula.

The Golden Formula for Classical Probabilities Restated

If all outcomes are equally likely, then the probability of an event A can be determined by

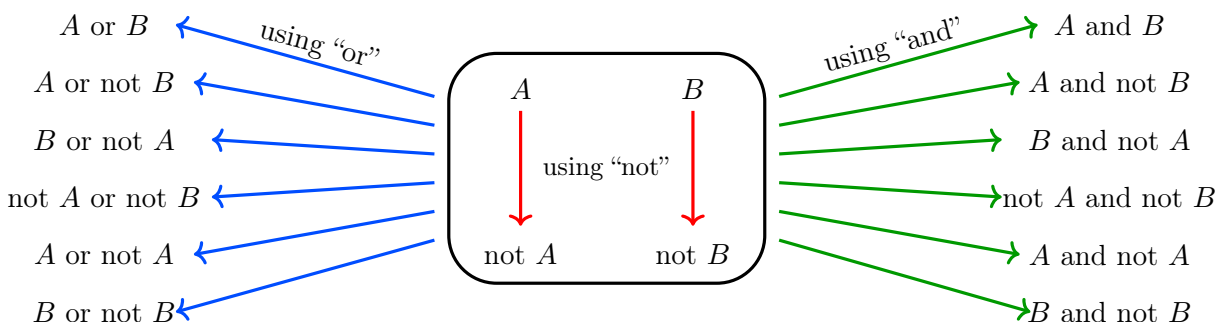
$$P(A) = \frac{|A|}{|S|},$$

that is, the cardinality of A divided by the cardinality of the sample space S .

5.3 Compound Probabilities

5.3.1 Building compound events

Suppose we have two events A and B . Based on our discussion so far, we know how to determine the probabilities $P(A)$ and $P(B)$: we simply count up the number of outcomes that correspond to A or B , and then divide that number by the size of the sample space. As a matter of fact, we can build many other events from A and B by applying three logical building blocks, namely: *not*, *and*, and *or*. For instance, applications of these three words to A and B allow us to create many other statements, which are all events in their own right:



This diagram looks suspiciously similar to a diagram we saw in Lecture 4 when discussing set theory. In fact, it is the *same diagram*. The three operations of *not*, *and*, and *or* can be directly translated into set theory as follows:

Language	Set operation	Meaning
not A	A^c	the complement of A
A and B	$A \cap B$	the intersection of A and B
A or B	$A \cup B$	the union of A and B

According to the previous section, we can calculate the probability of any event by counting up the size of the subset, and then dividing by the size of the sample space. In the case of the three statements above, this creates three new formulas:

$$P(\text{not } A) = \frac{|A^c|}{|S|},$$

$$P(A \text{ and } B) = \frac{|A \cap B|}{|S|},$$

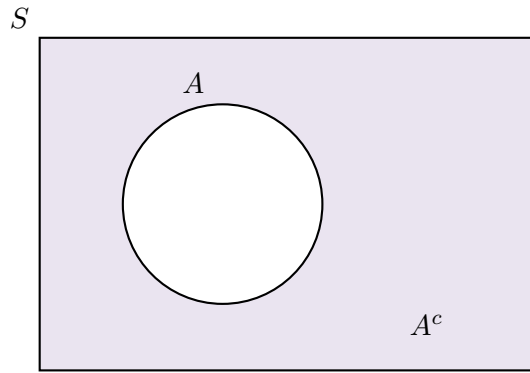
and

$$P(A \text{ or } B) = \frac{|A \cup B|}{|S|}.$$

Therefore, if we can figure out the sizes of these sets, then we can create formulas to calculate the corresponding compound probabilities. We will now explain these one-by-one.

5.3.2 The complement: not A

The complement A^c means everything in the sample space that is not in the subset A . In the picture below, the complement is the shaded area.



According to the classical probability formula, to calculate the probability of not- A , we should count the size of the complement A^c :

$$P(\text{not } A) = \frac{|A^c|}{|S|}.$$

Since A and A^c together cover the whole sample space, we have

$$|A| + |A^c| = |S|.$$

Therefore

$$|A^c| = |S| - |A|.$$

Substituting this into the probability formula gives

$$P(A^c) = \frac{|A^c|}{|S|} = \frac{|S| - |A|}{|S|} = \frac{|S|}{|S|} - \frac{|A|}{|S|} = 1 - P(A).$$

Complement rule

For any event A ,

$$P(\text{not } A) = 1 - P(A).$$

Example. In the true/false sequence example, let E be the event “FFF is chosen”. Since there are 8 equally likely sequences,

$$P(E) = \frac{1}{8}.$$

The complement E^c is the event “FFF is not chosen”. There are 7 favourable outcomes, namely everything except FFF. Therefore,

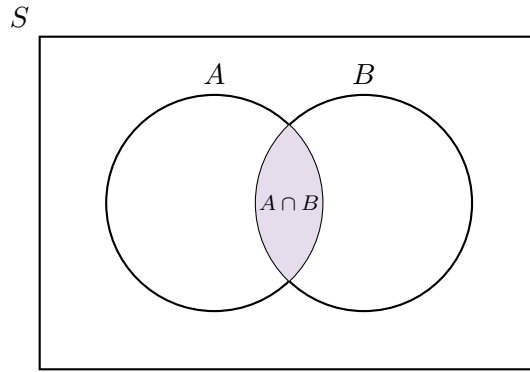
$$P(E^c) = \frac{7}{8}.$$

Equivalently, using the complement rule,

$$P(E^c) = 1 - P(E) = 1 - \frac{1}{8} = \frac{7}{8}.$$

5.3.3 The intersection: A and B

The intersection of A and B , written $A \cap B$, is the set of outcomes that are in both A and B at the same time. Visually, it is the overlap of the two sets:



According to our Golden Formula, the probability of the compound event A and B should involve the size of their intersection:

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

Unfortunately, there is no general formula for the size of $A \cap B$ in terms of only $|A|$ and $|B|$. Two large sets might overlap a lot, overlap a little, or not overlap at all. So, to calculate $P(A \text{ and } B)$, we need information about how the two events are related to each other.

5.3.4 Conditional probability

Suppose that we try to take our formula for $P(A \text{ and } B)$ and perform some algebra. We will be clever and artificially insert a copy of 1, disguised in the form $\frac{|B|}{|B|}$. Performing some basic algebra, we obtain:

$$P(A \text{ and } B) = \frac{|A \cap B|}{|S|} = \frac{|A \cap B|}{|S|} \frac{|B|}{|B|} = \frac{|A \cap B|}{|B|} \cdot \frac{|B|}{|S|}.$$

Thus we have managed to express our probability as a product of two other ratios. The second of these fractions is familiar: it is the definition of $P(B)$. In contrast, the first ratio written above is new: it is a special probability known as a *conditional probability*.

Definition: Conditional probability

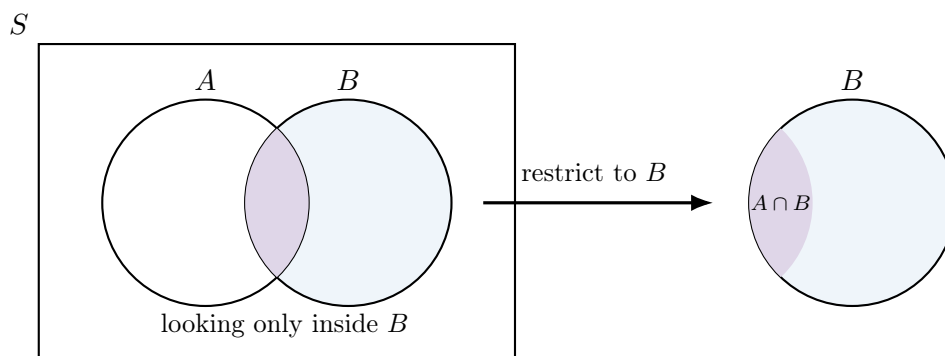
The conditional probability $P(A \mid B)$ is the probability that A happens, assuming that B has happened. In classical probability,

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

provided that $|B| \neq 0$.

This notion of conditional probability seems confusing at first glance. So, we will now explain what is happening using set theory.

As previously mentioned, there is no perfect formula that we can use to calculate the size of the intersection $A \cap B$. One way to figure out this size is to look inside the set B and then to count up all of the instances of A that we find. This is like taking our entire sample space and selectively ignoring everything that is not already in B :



If we were to try and compute the probabilities based on this new set up, we notice that the sample space is no longer S , instead it is the subset B . In that sense, since the sample space is now B , if we were to try to count the probability of A happening in this restricted context, we would need to use B in the denominator in our golden formula:

$$P(A, \text{ within } B) = \frac{\text{number of elements of } A \text{ that are also in } B}{\text{the size of } B}.$$

Any element of A that is also in B falls precisely in the intersection $A \cap B$. Therefore, the probability associated to the above allows us to recover the previous formula for conditional probability:

$$P(A \mid B) = P(A, \text{ within } B) = \frac{\text{number of elements of } A \text{ that are also in } B}{\text{the size of } B} = \frac{|A \cap B|}{|B|}.$$

As an example, suppose that we roll a fair dice. Let

$$A = \{\text{we roll an even number}\} \quad \text{and} \quad B = \{\text{we roll a number less than 4}\}.$$

Written out as subsets, the events A and B read:

$$A = \{2, 4, 6\}, \quad \text{and} \quad B = \{1, 2, 3\}.$$

So, you throw the dice up into the air. Imagine that before the dice lands, God came down from the heavens and whispered into your ear: “the outcome will be less than 4”. This means that, in that moment, you know for *certain* that it is not possible to roll a 4, 5 or 6. In this situation, if you were to evaluate your chances of event A happening and try to calculate a new probability, you would cut your space of possibilities from

$$\{1, 2, 3, 4, 5, 6\}$$

down to the smaller set

$$\{1, 2, 3\}.$$

Inside this smaller set, only one outcome is even, namely 2. Therefore

$$P(A \mid B) = \frac{1}{3}.$$

Using set notation,

$$A \cap B = \{2\},$$

so

$$P(A \mid B) = \frac{|A \cap B|}{|B|} = \frac{1}{3}.$$

Putting this all together, we see that

$$P(B) = \frac{3}{6} = \frac{1}{2}, \quad P(A \text{ and } B) = \frac{1}{6}, \quad P(A \mid B) = \frac{1}{3}.$$

Funnily enough, it just so happens that:

$$\frac{1}{6} = \frac{1}{2} \cdot \frac{1}{3}.$$

But, this is no coincidence: it is a manifestation of the general formula we mentioned before.

$$P(A \text{ and } B) = \frac{|A \cap B|}{|S|} = \frac{|A \cap B|}{|B|} \cdot \frac{|B|}{|S|} = P(A \mid B)P(B).$$

General multiplication rule

When the relevant conditional probability is defined, for any events A and B ,

$$P(A \text{ and } B) = P(B)P(A \mid B).$$

Equivalently,

$$P(A \text{ and } B) = P(A)P(B \mid A).$$

5.3.5 Independent and dependent events

In some cases, the formula for $P(A \text{ and } B)$ can be heavily simplified. In order to properly explain how, we need to first make the distinction between independent and dependent events.

Independent and dependent events

Let A and B be events.

- A and B are **independent** if learning that one happened does not change the probability of the other.
- A and B are **dependent** if learning that one happened *does* change the probability of the other.

As a reminder, probability is the science of uncertainty. So, the idea behind dependent events is that knowing the event B happened might affect the probability of A . This can potentially happen whenever the information about B is somehow relevant for the event A .

In the previous example of rolling the dice, the events A being “rolling an even number” and B being “rolling a number less than 4” are *dependent* events, because knowing that B would happen actually changed our probability from $\frac{1}{2}$ to $\frac{1}{3}$. In other words, information about one event affected our knowledge about the possibility of the other.

However, we can also imagine situations where knowing one piece of information gives us no reason to change our assessment of probability. For example, let the event C be “we roll a number greater than 2”. Then, as a subset,

$$C = \{3, 4, 5, 6\}.$$

If, for some reason, we absolutely *knew* that the outcome of our dice roll would be larger than 2, then we could restrict our sample space from

$$\{1, 2, 3, 4, 5, 6\}$$

down to the subset

$$\{3, 4, 5, 6\}.$$

Given that this restricted set consists of two odd numbers and two even numbers, the probability of rolling an even number is still $\frac{1}{2}$. This is the same as $P(A)$, so knowing that C will happen does not actually change the amount of certainty that we have about the system. Therefore, the events A and C are independent.

In symbols, A and B are independent when

$$P(A | B) = P(A),$$

and equivalently,

$$P(B | A) = P(B).$$

5.3.6 Multiplication rule for independent events

If A and B are independent, then $P(A | B) = P(A)$ can be used to drastically simplify the formula for $P(A \text{ and } B)$. We may perform a substitution to give the following multiplication rule.

Multiplication rule for independent events

If A and B are independent events, then

$$P(A \text{ and } B) = P(A)P(B).$$

Example: Two coin flips. Suppose that we flip a fair coin twice.

Let A be the event “the first flip is tails”, and let B be the event “the second flip is tails”. These events are independent, because the result of the first flip does not affect the result of the second flip.

Therefore,

$$P(A \text{ and } B) = P(A)P(B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}.$$

Equivalently, the sample space is

$$S = \{HH, HT, TH, TT\},$$

and only TT is favourable.

Example: Two separate decks. Suppose that we shuffle two separate decks of cards.

Let A be the event “the top card of deck 1 is an ace”, and let B be the event “the top card of deck 2 is an ace”. These events are independent, and each has probability

$$\frac{4}{52} = \frac{1}{13}.$$

So

$$P(A \text{ and } B) = \left(\frac{1}{13}\right) \left(\frac{1}{13}\right) = \frac{1}{169}.$$

Equivalently, there are $52 \times 52 = 2704$ possible ordered pairs of top cards, and $4 \times 4 = 16$ of these pairs consist of two aces. Hence

$$\frac{16}{2704} = \frac{1}{169}.$$

The following is an example of dependent events.

Example: Two aces from one deck. Suppose that we shuffle one deck of 52 cards. Let

$$A = \{\text{the first card is an ace}\},$$

and

$$B = \{\text{the second card is an ace}\}.$$

These events are dependent, because if the first card is an ace, then there are only 3 aces left among the remaining 51 cards when we try to take a second. We calculate:

$$P(A) = \frac{4}{52} = \frac{1}{13},$$

and

$$P(B | A) = \frac{3}{51}.$$

Therefore,

$$P(A \text{ and } B) = P(A)P(B | A) = \frac{1}{13} \cdot \frac{3}{51} = \frac{1}{221}.$$

How to calculate $P(A \text{ and } B)$

Step 1: Decide whether the events are independent. If $P(A) = P(A | B)$, then A and B are independent.

Step 2: If A and B are independent, use

$$P(A \text{ and } B) = P(A)P(B).$$

Step 3: For any events, independent or dependent, we may use

$$P(A \text{ and } B) = P(A)P(B | A),$$

or equivalently,

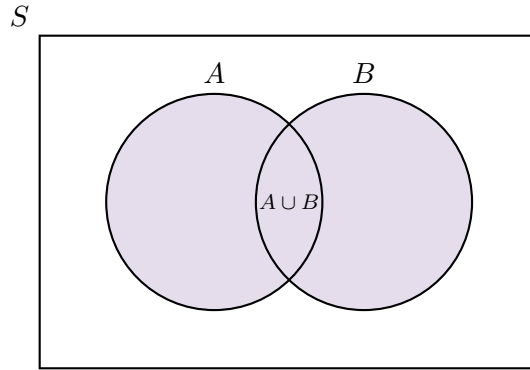
$$P(A \text{ and } B) = P(B)P(A | B).$$

5.3.7 The union: A or B

According to Section 3.1, the compound event “ A or B ” corresponds to the set-theoretic union $A \cup B$. Thus, in order to calculate the probability $P(A \text{ or } B)$, we can count up the size of this subset and divide it by the size of the sample space:

$$P(A \text{ or } B) = \frac{|A \cup B|}{|S|}.$$

The union of A and B , written $A \cup B$, is the set of outcomes that are in A , in B , or in both.



As with the intersection discussed earlier, generally we cannot deduce the size of the union $A \cup B$ based only on the sizes of A and B . In fact, we noticed last lecture that it is very tempting to naively write:

$$|A \cup B| = |A| + |B|.$$

However, this is wrong in general, because the overlap $A \cap B$ accidentally gets counted twice. For example, if

$$A = \{1, 2, 3\}, \quad \text{and} \quad B = \{2, 3, 4\},$$

then their union is the set

$$A \cup B = \{1, 2, 3, 4\},$$

which has cardinality equal to 4. However,

$$|A| + |B| = 3 + 3 = 6 \neq 4.$$

The issue is that there are two copies of the intersection $\{2, 3\}$ hiding away inside the sum $3 + 3 = 6$.

Generally speaking, to fix this issue we can subtract one copy of the intersection:

$$|A \cup B| = |A| + |B| - |A \cap B|.$$

We can use this observation to derive a general formula for $P(A \text{ or } B)$ in terms of $P(A)$, $P(B)$ and $P(A \text{ and } B)$:

$$\begin{aligned} P(A \text{ or } B) &= \frac{|A \cup B|}{|S|} \\ &= \frac{|A| + |B| - |A \cap B|}{|S|} \\ &= \frac{|A|}{|S|} + \frac{|B|}{|S|} - \frac{|A \cap B|}{|S|} \\ &= P(A) + P(B) - P(A \text{ and } B). \end{aligned}$$

This is known as the *addition rule*, and it deserves its own box.

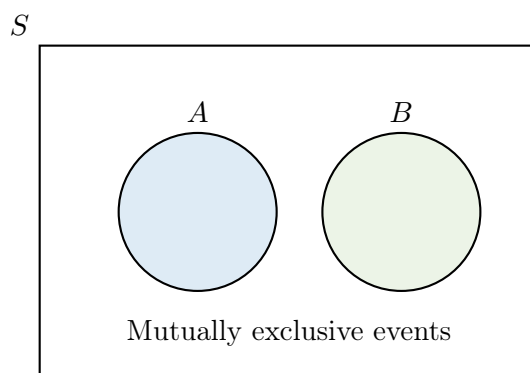
Addition rule

For any events A and B ,

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

5.3.8 Mutually exclusive events

Sometimes two events cannot happen at the same time. For example, a single coin flip cannot land on heads and tails at the same time. Such events are called *mutually exclusive*. Visually, mutually exclusive events have no overlap:



Definition: Mutually exclusive events

Two events A and B are **mutually exclusive** if they cannot happen at the same time. Equivalently, $A \cap B = \emptyset$, meaning that

$$P(A \text{ and } B) = \frac{|A \cap B|}{|S|} = \frac{0}{|S|} = 0.$$

Using this observation, we now notice that in our formula for the probability of A or B , the third term vanishes for mutually exclusive events:

$$\begin{aligned} P(A \text{ or } B) &= \frac{|A \cup B|}{|S|} \\ &= \frac{|A| + |B| - \cancel{|A \cap B|}}{|S|} \\ &= \frac{|A|}{|S|} + \frac{|B|}{|S|} - \cancel{\frac{|A \cap B|}{|S|}} \\ &= P(A) + P(B) - \cancel{P(A \text{ and } B)} \\ &= P(A) + P(B). \end{aligned}$$

This means that our formula will reduce to the simple addition of the probabilities of A and B . This is known as the addition rule for mutually exclusive events.

Addition rule for mutually exclusive events

If A and B are mutually exclusive, then

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

As an example, suppose that we shuffle a standard deck of 52 cards. Define the events:

$$A = \{\text{the top card is an ace}\},$$

$$B = \{\text{the top card is a spade}\},$$

and

$$C = \{\text{the top card is a king}\}.$$

The events A and B are not mutually exclusive, because the ace of spades is both an ace and a spade. However, A and C are mutually exclusive, because the top card cannot be both an ace and a king at the same time.

Exercise: Ace, spade and king

Compute the following.

(a) $P(A \text{ or } B)$.

(b) $P(A \text{ or } C)$.

Solution

(a) We have

$$P(A) = \frac{4}{52}, \quad P(B) = \frac{13}{52}, \quad P(A \text{ and } B) = \frac{1}{52}.$$

Therefore,

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) = \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52} = \frac{4}{13}.$$

(b) Since A and C are mutually exclusive,

$$P(A \text{ or } C) = P(A) + P(C).$$

Therefore,

$$P(A \text{ or } C) = \frac{4}{52} + \frac{4}{52} = \frac{8}{52} = \frac{2}{13}.$$

Summary: Compound probabilities

The main compound events in this lecture are built from the words *not*, *and*, and *or*. In set notation, these correspond to

$$\text{not } A \longleftrightarrow A^c, \quad A \text{ and } B \longleftrightarrow A \cap B, \quad A \text{ or } B \longleftrightarrow A \cup B.$$

For complements, use

$$P(\text{not } A) = 1 - P(A).$$

For intersections, use the multiplication rule

$$P(A \text{ and } B) = P(A)P(B \mid A),$$

or equivalently,

$$P(A \text{ and } B) = P(B)P(A \mid B).$$

If the events are independent, this simplifies to

$$P(A \text{ and } B) = P(A)P(B).$$

For unions, use the addition rule

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

If the events are mutually exclusive, then $P(A \text{ and } B) = 0$, so this simplifies to

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

Lecture 6

More on Probability

In Lecture 5, we introduced the basic vocabulary of probability: simple events, sample spaces, events, complements, intersections, unions, conditional probability and independence. We also saw that classical probability is especially nice because all outcomes are equally likely, so probabilities can be calculated by counting. Today we will practise that method in more complicated examples, especially examples where the sample space is built using multiplication, permutations or combinations. After that, we will begin moving beyond classical probability, where the outcomes are not necessarily equally likely. This leads naturally to probability distributions.

6.1 Problems Involving Counting

Today we will spend some time reviewing some exercises in classical probability, and then we will move away from this setting by studying situations in which outcomes are not equally likely. Before doing this, we will first briefly review the theory from last lecture.

6.1.1 A brief review of the formalism

Throughout Lecture 5, we introduced the basic formalism of probability theory. There are several key notions from our discussion that are worth recalling:

1. An *outcome* is something that can happen as a result of an experiment or operation.
2. The collection of all outcomes is known as the *sample space*. Formally, this is a set S containing all the outcomes ω .
3. An event is loosely defined to be a statement that we can make about our system. Formally, events are defined to be subsets A of the sample space S .
4. Simple events are special events which only contain one outcome, i.e. $A = \{\omega\}$ for some outcome ω in S .
5. The probability of an event, written $P(A)$, is a number between 0 and 1, where 0 represents impossibility, and 1 represents certainty.

Throughout Lecture 5, we worked with *classical probability*, which makes the extra assumption that all outcomes are equally likely, and the sample space S consists of finitely many outcomes in total. If the sample space has size $|S| = n$, then the probability of each outcome is $\frac{1}{n}$. As a matter of fact, we can write this in several equivalent ways. Technically, P is a function that needs events as inputs, *not* outcomes. So, the probability of an outcome happening can be written in terms of simple events as:

$$P(\{\omega\}) = \frac{1}{n}.$$

The central formula from the previous lecture was the Golden Formula for classical probability. We first introduced this naively, and then carefully studied it based on the properties of sets. Two equivalent versions of the Golden Formula are stated below.

The Golden Formula

If all outcomes are equally likely, then

$$P(A) = \frac{\text{number of favourable outcomes}}{\text{total number of outcomes}} = \frac{|A|}{|S|}.$$

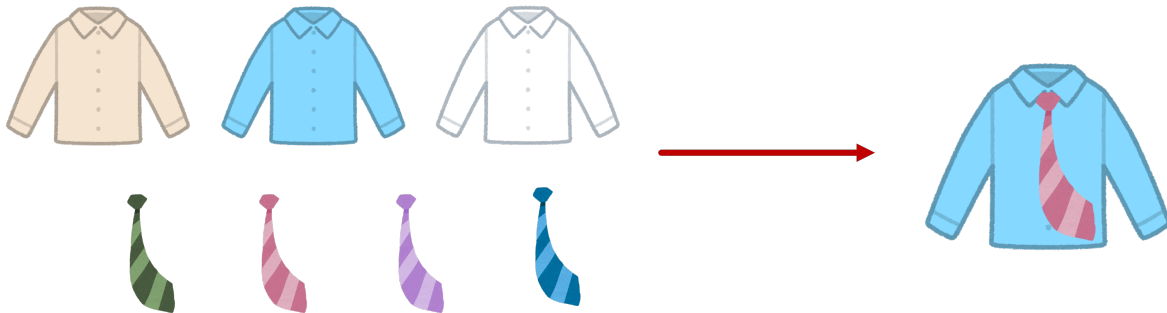
In particular, notice that if we take a simple event $A = \{\omega\}$, then we can apply the Golden Formula to recover the probability $\frac{1}{n}$ for each outcome.

The Golden Formula only works for classical probability, that is, it works when the sample space S is finite and all outcomes in S are equally likely. Often, the main difficulty is not the computation itself, but the act of *counting everything up*. We will now carefully walk through three examples of ways in which we may apply the Golden Formula in different contexts. These are based on the discussion of Lecture 4:

- choices made from several categories,
- arrangements of objects into different orders,
- random selections from a large collection of objects.

6.1.2 Example 1: picking shirts and ties

Recall our example from Lecture 4: we have a wardrobe with 3 shirts and 4 ties, and we want to pick one of each to make an outfit. For example, one such configuration is pictured below.



Using this set up, we can start picking objects at random, and therefore make our outfit a question of probability. Consider the following three questions.

Exercise 1

Suppose that we pick a shirt and then a tie at random, so that each object has an equal chance of being selected.

- What is the probability of choosing the brown shirt?
- What is the probability of choosing a white shirt with a red tie?
- What is the probability of choosing an outfit that has blue in it?

In order to answer this question, we will first create a sample space. According to the set up, the outcomes are the full choices of outfits. As we saw in Lecture 4, these can be arranged into a grid:



As you can see, there are $12 = 3 \times 4$ possible outfits. This means that the sample space S of this system has size 12, i.e. $|S| = 12$.

- (a) We can answer the first question in a few separate ways. Firstly, we may apply the Golden Formula on an informal level to perform the count. Let A be the event “picking a brown shirt and any tie”. Then

$$P(A) = \frac{\text{number of outfits with a brown shirt}}{\text{total number of outfits}} = \frac{4}{12} = \frac{1}{3}.$$

Alternatively, on a practical level we can simply observe that the choice of shirts is entirely separate from the choice of ties, and therefore we can just view the choice of shirts as a probabilistic system in its own right. We can write a new sample space

$$S' = \{\text{white shirt, blue shirt, brown shirt}\}.$$

There are three shirts and each is chosen with the same probability, meaning that the chance of picking the brown shirt is simply $\frac{1}{3}$. Selecting a tie afterwards does nothing to the chance of selecting a brown shirt.

- (b) For the next question, we are being asked to find the probability for an individual outcome: the white shirt and red tie. We may use the Golden Formula directly to see that

$$P(\text{white shirt with red tie}) = \frac{\text{number of white shirt with red tie outfits}}{\text{total number of outfits}} = \frac{1}{12}.$$

Alternatively, we can define the event A to mean “selecting a white shirt” and event B to mean “selecting a red tie”. Then the statement “selecting a white shirt with a red tie” can be reinterpreted as

A and B : selecting a white shirt and a red tie.

Since these two events are independent, we can use the multiplication formula from before, to conclude that

$$P(A \text{ and } B) = P(A) \cdot P(B) = \frac{1}{3} \cdot \frac{1}{4} = \frac{1}{12}.$$

- (c) Finally, we calculate the probability of selecting something blue. Again, there are several ways to approach the solution. Perhaps the simplest method is to examine the sample space and simply highlight all those outfits that happen to have blue in them:

As you can see, there are 6 outfits. Thus, we can naively apply the Golden Formula to conclude that

$$P(\text{something blue}) = \frac{\text{number of outfits with blue in it}}{\text{total number of outfits}} = \frac{6}{12} = \frac{1}{2}.$$

Alternatively, we may define two events

$$B = \{\text{outfits with a blue shirt}\} \quad \text{and} \quad C = \{\text{outfits with a blue tie}\}.$$

The subset B is pictured above in red, and the subset C is pictured above in blue. We note that these two events are *not* mutually exclusive, since there is an outfit consisting of a blue shirt with blue tie. This means that we must subtract this when computing the or statement:

$$P(B \text{ or } C) = P(B) + P(C) - P(B \text{ and } C) = \frac{4}{12} + \frac{3}{12} - \frac{1}{12} = \frac{6}{12} = \frac{1}{2}.$$

Of course, all of these questions can be answered more formally by actually writing out the event in symbols. We can write our system as a collection of ordered pairs corresponding to the choices of shirt/tie and the resulting outfits:



According to the diagram above, the brown shirt is third and corresponds to s_3 . We can write the sample space as the collection of all of the ordered pairs in the grid above:

$$S = \{(s_1, t_1), (s_1, t_2), \dots, (s_3, t_4)\},$$

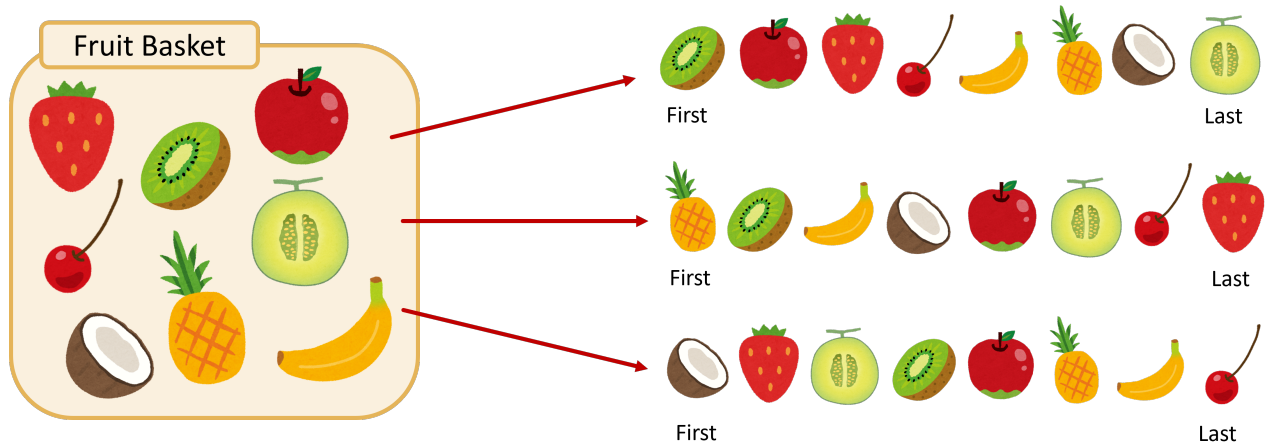
and the event containing all of the outfits with a brown shirt is

$$A = \{(s_3, t_1), (s_3, t_2), (s_3, t_3), (s_3, t_4)\}.$$

Using the formula $P(A) = \frac{|A|}{|S|}$, we recover $P(A) = \frac{1}{3}$ from before.

6.1.3 Example 2: ordering fruit

Recall from Lecture 4 the example of the fruit basket. There, we had 8 fruit arranged in a basket and we selected them one-by-one. As we saw, there were $8! = 40320$ different ways to do this, three of which are pictured below:



Exercise 2

Suppose that we randomly pick out the fruit in order. In other words, suppose that we randomly select a permutation of the 8 fruit in the basket.

- What is the probability of choosing a kiwi first?
- What is the probability of selecting an ordering which has the strawberry third?
- What is the probability of choosing an ordering which has a melon in the second slot and a cherry in the third slot?

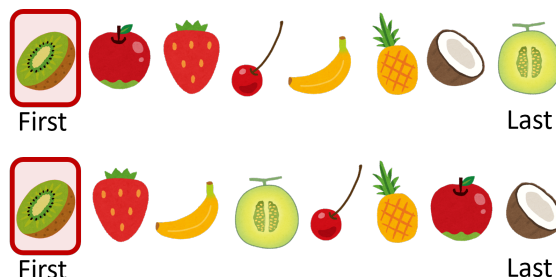
- (a) We will argue this in two separate ways. Firstly, perhaps the easiest method is to simply observe: for the first position in the ordering, each of the 8 fruit in the basket has an equal probability of being selected. There are 8 fruit, and only one is a kiwi, therefore the probability of selecting the kiwi first ought to be

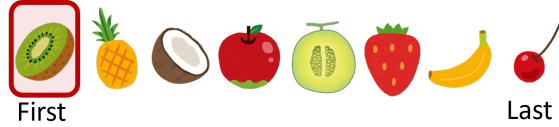
$$P(\text{kiwi first}) = \frac{\text{number of kiwis}}{\text{total number of fruit}} = \frac{1}{8}.$$

This can be confirmed if we instead answer the question on the level of full permutations. As the question states, each ordering has an equally likely chance, so we may equally calculate the probability of selecting a kiwi first by taking the ratio

$$P(\text{kiwi first}) = \frac{\text{number of permutations that have a kiwi first}}{\text{total number of permutations}}.$$

For the denominator, we saw in Lecture 4 that there are $8! = 40320$ unique ways to order the 8 pieces of fruit. Regarding the numerator above, we may calculate this by simply fixing the kiwi in place. For example, below are three of the possible orderings with the kiwi fixed:



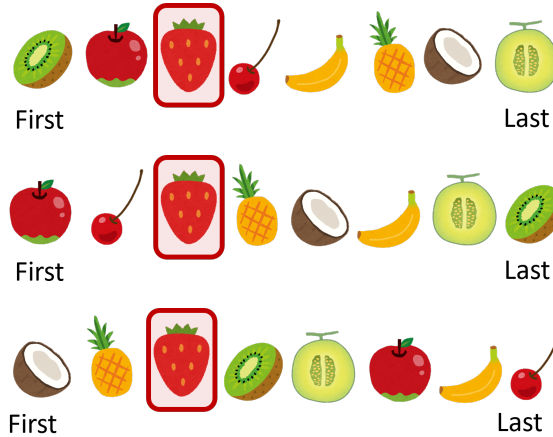


To calculate the total number of ways to make these orderings, we notice that there are 7 pieces of fruit that need to be ordered and there are 7 slots in which to order them. So, the number of total possible orderings is $7!$, which cancels down to return the same probability we calculated earlier:

$$P(\text{kiwi first}) = \frac{\text{number of permutations that have a kiwi first}}{\text{total number of permutations}}$$

$$= \frac{7!}{8!} = \frac{\cancel{7} \times \cancel{6} \times \cancel{5} \times \cancel{4} \times \cancel{3} \times \cancel{2} \times \cancel{1}}{8 \times \cancel{7} \times \cancel{6} \times \cancel{5} \times \cancel{4} \times \cancel{3} \times \cancel{2} \times \cancel{1}} = \frac{1}{8}.$$

- (b) By the same logic, we can fix the strawberry in place and permute everything around it. For example, three such orderings could be:

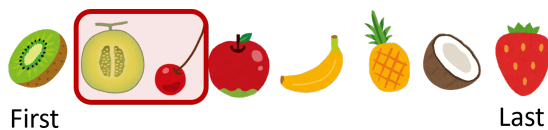


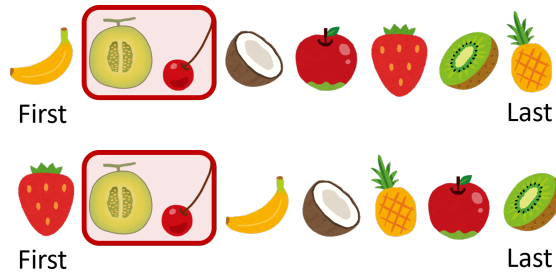
As with the previous question, there are 7 fruits that need to be ordered around the strawberry, and there are $7!$ ways to do this. The probability again reduces to

$$P(\text{strawberry third}) = \frac{\text{number of permutations that have a strawberry third}}{\text{total number of permutations}}$$

$$= \frac{7!}{8!} = \frac{\cancel{7} \times \cancel{6} \times \cancel{5} \times \cancel{4} \times \cancel{3} \times \cancel{2} \times \cancel{1}}{8 \times \cancel{7} \times \cancel{6} \times \cancel{5} \times \cancel{4} \times \cancel{3} \times \cancel{2} \times \cancel{1}} = \frac{1}{8}.$$

- (c) Finally, we are being asked to find the probability of picking a permutation with the melon in the second position and the cherry in the third. We will make two different arguments. Firstly, we can operate on a pure count as before. However, in contrast to our first two questions, here we have *two* slots already occupied. For example, some permutations may look like:





In this case, we are effectively being asked to count the number of ways to reorder the 6 fruits around the melon and cherry. Since there are 6 fruits, the total number of orderings will be $6!$, and therefore our probability is

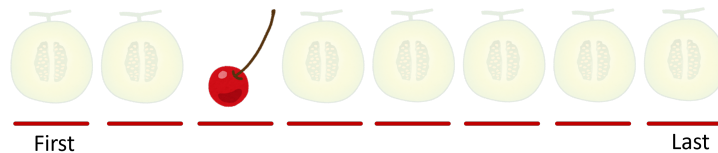
$$P(\text{melon second, cherry third}) = \frac{\text{number of permutations with melon second, cherry third}}{\text{total number of permutations}}$$

$$= \frac{6!}{8!} = \frac{\cancel{6} \times \cancel{5} \times \cancel{4} \times \cancel{3} \times \cancel{2} \times \cancel{1}}{8 \times 7 \times \cancel{6} \times \cancel{5} \times \cancel{4} \times \cancel{3} \times \cancel{2} \times \cancel{1}} = \frac{1}{8 \times 7} = \frac{1}{56}.$$

Alternatively, we can interpret our question as a compound event. If we let

$$A = \text{picking a melon second}, \quad B = \text{picking a cherry third},$$

then our desired probability is $P(A \text{ and } B)$. According to our formula for “and” statements, we need to calculate the conditional probability $P(A|B)$. Our discussion in parts (a) and (b) tells us that the probability of finding a cherry in the third slot is $\frac{1}{8}$, since there are 8 potential slots and each is equally likely to contain the cherry. In contrast, to calculate the conditional probability $P(A|B)$, we need to calculate the chance of the melon being picked second, given that we *know* that the cherry will land in the third slot. Since we know the cherry will be third, this reduces the number of available slots from 8 to 7:



There is an equal probability for the melon to land in each of these positions. In particular, we conclude that the chance of finding the melon in the second position is $\frac{1}{7}$. Putting this all together, we can calculate the compound probability by multiplying this conditional probability with $P(B)$:

$$P(A \text{ and } B) = P(A|B) \cdot P(B) = \frac{1}{7} \cdot \frac{1}{8} = \frac{1}{56}.$$

6.1.4 Example 3: a royal flush

There are several different ways to play poker. One simple version is called *Five-Card Draw*, where, as the name may suggest, each player gets given 5 cards in their hand. The players are free to rearrange these cards in their hand as they like, so the ordering does not matter at all. According to the rules, the best possible hand is the “Royal Flush”, which consists of the top 5 cards: 10, Jack, Queen, King and Ace, all in the same suit.

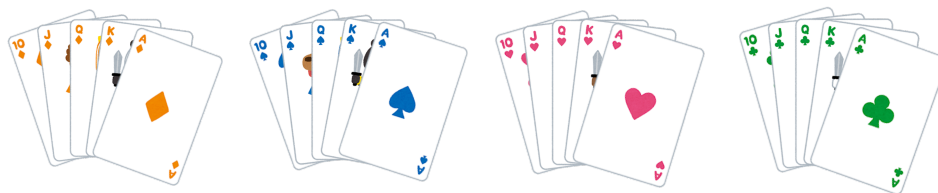
Exercise 3

What is the probability that a player receives a royal flush?

To answer this question, we will use the Golden Formula together with some combinatorics. The probability ought to be

$$P(\text{royal flush}) = \frac{\text{number of possible royal flushes}}{\text{number of possible hands}}.$$

Fortunately, the numerator is very easy to calculate. There are 4 suits, therefore there are 4 possible royal flushes:



To calculate the denominator, we recall that a standard deck of cards has 52 cards. According to the rules of Five-Card Draw, an individual poker hand consists of 5 cards and the order of the cards in the player's hand does not matter. This means that we are looking to count the number of ways that we can select 5 cards from 52, without caring about their order. To do so, we use the $C_{n,r}$ formula from Lecture 4, where here $n = 52$ and $r = 5$:

$$C_{52,5} = \frac{52!}{5!47!} = 2,598,960.$$

With this in hand (pun intentional), we can calculate the probability to be

$$P(\text{royal flush}) = \frac{\text{number of possible royal flushes}}{\text{number of possible hands}} = \frac{4}{2,598,960} = \frac{1}{649,740}.$$

So, the chance of getting a royal flush is 1 in 649,740. This example is a good reminder that the Golden Formula itself is very simple. The difficult part is often the denominator: before we can calculate the probability, we need to understand the structure of the sample space well enough to count it.

6.2 Beyond Classical Probability

6.2.1 The hidden assumption in classical probability

So far, almost every probability calculation we have done has used the Golden Formula:

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

When applicable, this formula is extremely useful. But, its use is limited, because it assumes that the n outcomes all carry the same probability of occurring (namely $\frac{1}{n}$). This assumption is often justified because of some symmetry in the system we are working with: a fair coin has two opposite sides, a 6-sided dice is a cube, and so on. However, in reality many systems aren't perfectly symmetric like this. In these cases, we can still discuss questions of probability, but we have to relax our assumption that every outcome has the same chance of happening.

For example, suppose that a coin has an uneven probability of landing on either heads or tails, say:

$$P(H) = 0.4, \quad P(T) = 0.6.$$

Here the sample space is still $S = \{H, T\}$, however the probabilities are no longer equal at 50%. Put differently, the two outcomes are no longer equally likely. Notice here that we cannot use our simple Golden Formula to determine the probability of events:

$$P(\text{landing on heads}) \neq \frac{\text{number of heads in } S}{\text{size of } S} = \frac{1}{2}.$$

This is just a simple example, but it motivates the need for a more general treatment of probabilities. We will now discuss this in detail. Throughout this section we will assume that all of the rules of probability theory still apply, however, we will *not* suppose that every outcome is equally likely.

6.2.2 Assigning probabilities to simple events

Let S be a finite sample space, say:

$$S = \{\omega_1, \omega_2, \dots, \omega_n\}.$$

Here the ω_i are the n unique outcomes of the system. Instead of assuming that each ω_i has a fixed probability of $\frac{1}{n}$, we will relax this assumption and instead simply write the probability of each individual outcome as:

$$p(\omega_i),$$

which is just a shorthand notation for $P(\{\omega_i\})$.

In order to comply with the other rules of probability theory, we still need to require that:

1. each probability $p(\omega_i)$ is a number between 0 and 1, and

2. the sum of all their probabilities equals 1:

$$\sum_{i=1}^n p(\omega_i) = p(\omega_1) + p(\omega_2) + \cdots + p(\omega_n) = 1.$$

As you can verify, our example of the unfair coin satisfies these properties. Using this setup, we can create a new “Golden Formula” for this more general setting.

Finite probability model

Let $S = \{\omega_1, \omega_2, \dots, \omega_n\}$ be a finite sample space of size n . A probability model assigns a number $p(\omega_i)$ to each outcome ω_i , such that

$$0 \leq p(\omega_i) \leq 1$$

and

$$\sum_{\omega \in S} p(\omega) = 1.$$

Then, for any event $A \subseteq S$, we define

$$P(A) = \sum_{\omega \in A} p(\omega).$$

This last formula is the general replacement for the Golden Formula. It says that in order to calculate the probability of an event A , we have to look at all of the outcomes ω contained in A and simply add up their probabilities.

As a matter of fact, our Golden Formula for classical probability is simply an *instance* of this more general idea. To see this, suppose that we have an event A containing say m outcomes, i.e.

$$A = \{\omega_1, \dots, \omega_m\}.$$

If we assume for a moment that each ω_i carries the same probability of $\frac{1}{n}$ where n is the size of the sample space S , then some simple algebra reveals:

$$P(A) = \sum_{\omega \in A} p(\omega) = p(\omega_1) + \dots + p(\omega_m) = \underbrace{\frac{1}{n} + \dots + \frac{1}{n}}_{m \text{ many}} = \frac{1 + \dots + 1}{n} = \frac{m}{n} = \frac{|A|}{|S|}.$$

So, we see that our Golden Formula for classical probability has been swallowed into a larger scheme.

6.2.3 The familiar formulas still work

Even though we have moved beyond classical probability, the main probability formulas from Lecture 5 still work. For completeness, they are stated again below.

Probability formulas

For events A and B :

$$P(\text{not } A) = 1 - P(A).$$

For and-statements:

$$P(A \text{ and } B) = \begin{cases} P(A) \cdot P(B), & \text{if } A \text{ and } B \text{ are independent,} \\ P(A) \cdot P(B|A), & \text{in general.} \end{cases}$$

For or-statements:

$$P(A \text{ or } B) = \begin{cases} P(A) + P(B), & \text{if } A \text{ and } B \text{ are mutually exclusive,} \\ P(A) + P(B) - P(A \text{ and } B), & \text{in general.} \end{cases}$$

In Lecture 5, we derived these formulas from properties of sets. However, they can still be derived using the general sum formula mentioned previously. This can all be justified using the algebra of the big-sum notation $\sum$. However, for the sake of brevity we will not do this, and instead we will justify these formulas heuristically.

For complements, observe that the sample space always splits into two non-overlapping pieces:

$$S = A \cup A^c,$$

that is, everything either belongs to A , or does not belong to A . Therefore

$$P(A) + P(A^c) = 1,$$

which rearranged gives:

$$P(A^c) = 1 - P(A).$$

For “or” statements, we are met with the usual issue of the overlap: if we were to try and simply add $P(A)$ and $P(B)$, then everything in $A \cap B$ would accidentally be added up twice. Therefore we subtract one copy:

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

For “and” statements, the general rule comes from conditional probability (which we will discuss at length in Lecture 7). If $P(A) \neq 0$, then

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

Rearranging gives

$$P(A \text{ and } B) = P(A) \cdot P(B|A).$$

If A and B are independent, then knowing that A happened does not change the probability of B . In that case,

$$P(B|A) = P(B),$$

so

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

6.2.4 Example: an unfair dice

Suppose that we have an unfair dice with the following probabilities:

$$p(1) = 0.1, \quad p(2) = 0.1, \quad p(3) = 0.1, \quad p(4) = 0.1, \quad p(5) = 0.5, \quad p(6) = 0.1.$$

This is not a fair dice, since the probability of landing on a 5 is much higher than all of the other outcomes. However, it is still a perfectly valid probability model, because each individual probability is between 0 and 1, and their overall sum equals 1 exactly:

$$0.1 + 0.1 + 0.1 + 0.1 + 0.5 + 0.1 = 1.$$

We will now use our new formula to calculate some probabilities associated to this system.

Exercise 4

Using the unfair dice above, calculate the following:

- (a) What is the probability of landing on a number different from 5?
- (b) What is the probability of landing on an even number?
- (c) What is the probability of landing on a number less than 3?

Solution

- (a) Let A be the event “land on 5”. Then the event “land on a number different from 5” is the complement of A . Therefore

$$P(\text{not } 5) = 1 - P(5) = 1 - 0.5 = 0.5.$$

- (b) The even outcomes are 2, 4 and 6. Therefore

$$P(\text{even}) = p(2) + p(4) + p(6) = 0.1 + 0.1 + 0.1 = 0.3.$$

- (c) The outcomes less than 3 are 1 and 2. Therefore

$$P(\text{less than } 3) = p(1) + p(2) = 0.1 + 0.1 = 0.2.$$

6.2.5 Example: an unfair coin

Suppose that we return to our unfair coin, where:

$$P(H) = 0.4, \quad \text{and} \quad P(T) = 0.6.$$

Exercise 5

Suppose that we flip this unfair coin independently several times in a row.

- (a) What is the probability of getting two heads in a row?
- (b) What is the probability of getting a heads and a tails, in any order?
- (c) What is the probability of getting three tails in a row?

Solution

- (a) Let A be the event “heads on the first flip” and B be the event “heads on the second flip”. The flips are independent, so we may simply multiply to yield:

$$P(HH) = P(H) \cdot P(H) = 0.4 \times 0.4 = 0.16.$$

- (b) To get one heads and one tails in any order, there are two possible sequences:

$$HT \quad \text{or} \quad TH.$$

These are mutually exclusive, so we add their probabilities:

$$P(HT \text{ or } TH) = P(HT) + P(TH).$$

Since the coin flips are independent,

$$P(HT) = 0.4 \times 0.6 = 0.24,$$

and

$$P(TH) = 0.6 \times 0.4 = 0.24.$$

Therefore

$$P(HT \text{ or } TH) = 0.24 + 0.24 = 0.48.$$

- (c) The probability of getting three tails in a row is

$$P(TTT) = P(T) \cdot P(T) \cdot P(T) = 0.6 \times 0.6 \times 0.6 = 0.216.$$

As an extension of part (b), we can write down the full sample space for two flips:

$$S = \{HH, HT, TH, TT\}.$$

The four probabilities are obtained by multiplication:

Outcome	HH	HT	TH	TT
Probability	0.16	0.24	0.24	0.36

These are unequal, and generally the probability has shifted towards TT and away from HH . To confirm, these probabilities still add up to 1 as they should:

$$0.16 + 0.24 + 0.24 + 0.36 = 1.$$

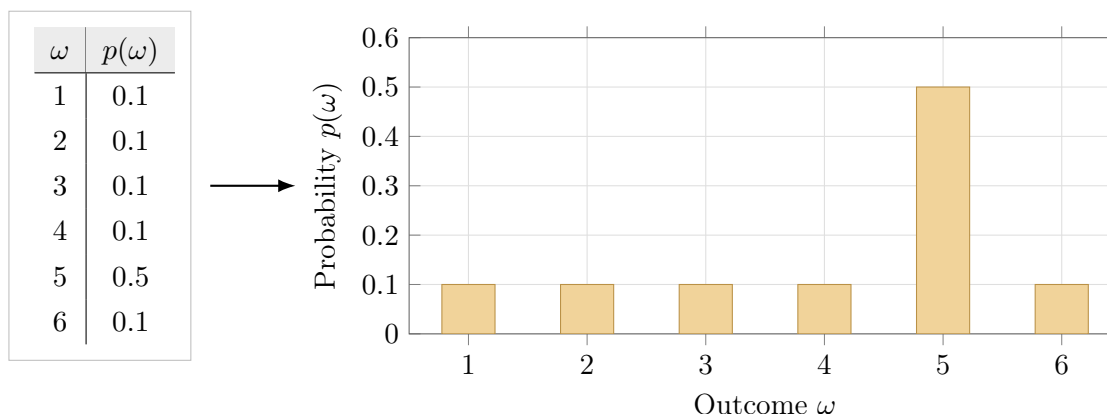
6.3 Discrete Probability Distributions

6.3.1 From probability weights to distributions

In the unfair dice example of Section 2.3, we assigned probabilities to each outcome:

$$p(1) = 0.1, \quad p(2) = 0.1, \quad p(3) = 0.1, \quad p(4) = 0.1, \quad p(5) = 0.5, \quad p(6) = 0.1.$$

We are now going to make a conceptual leap: this collection of information can be interpreted as a population of *data*, meaning that we can write it out in a table and plot it as a histogram:



This is very reminiscent of our discussion in Lecture 2. However, now we are carrying some probabilistic context: the heights of the bars are not pure counts, but *probabilities*. The histogram, therefore, tells us how the probability of our entire system is distributed across the individual outcomes. Thus we have discovered *probability distributions*.

Probability distribution

A **probability distribution** tells us how probability is spread (or, distributed) across the possible outcomes or values of a random experiment.

We will now briefly touch on the nature of the horizontal axis of a probability distribution.

6.3.2 Random variables

In the above example, the x axis of the histogram is labelled by the sample space – the numbers from 1 to 6. In usual mathematics, we might label this axis with a *variable*, which is a mystery letter x that is allowed to take on any values on the real number line. Ordinary graphs are then described by sliding the x variable along the number line and computing its various outputs. In the case of our probability distributions, we are working with a situation in which we do not simply “slide” our variable around, but instead, we allow our variable to take some value in the sample space *randomly*. For that reason, we interpret the variable on the x -axis as a *random variable*. We often write these random variables with capital letters, like X .

We will discuss random variables at length in later lectures – especially in our treatment of the Binomial Distribution in Lecture 8. For now, it is enough to understand the following definition.

Definition: random variable

A **random variable** is a rule that assigns a numerical value to each outcome of a random experiment.

The idea here is that a given sample space might not consist of quantitative data that we can plot on a histogram. So, random variables give us a way to turn the outcomes into numbers, so that we

can use descriptive statistics on the resulting probability distributions.
 As an example, suppose that we have a fair coin, and we flip it twice. The sample space is qualitative:

$$S = \{HH, HT, TH, TT\}$$

and the resulting probabilities are

$$P(HH) = P(HT) = P(TH) = P(TT) = \frac{1}{4}.$$

From this information we can define a random variable X by taking a count, for example:

$$X = \text{the number of heads.}$$

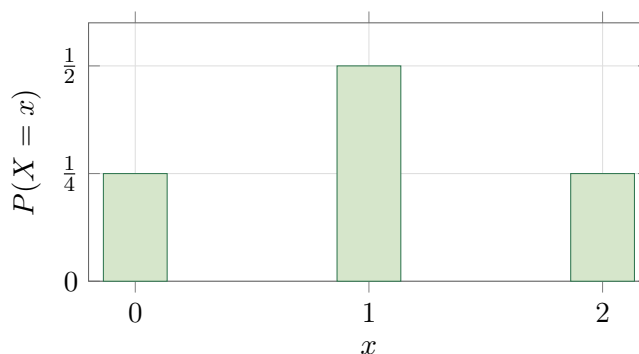
This allows us to describe the same system using numbers:

$$X(HH) = 2, \quad X(HT) = 1, \quad X(TH) = 1, \quad X(TT) = 0.$$

and from here we can re-evaluate the probabilities to:

$$P(X = 0) = P(TT) = \frac{1}{4}, \quad P(X = 1) = P(HT) + P(TH) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}, \quad P(X = 2) = P(HH) = \frac{1}{4}.$$

From this data, we obtain a genuine probability distribution, albeit quite a simple one:



6.3.3 Discrete probability distributions

Recall that real numbers, like 3, may have either a discrete or a continuous interpretation: we may think of “3” as counting up some quantity of things, or we may think of “3” as describing the length of a continuous interval. Throughout this lecture we will only deal with *discrete* random variables, since our sample spaces are finite sized.

Definition: discrete probability distribution

If X is a discrete random variable with possible values

$$x_1, x_2, \dots, x_n,$$

then a **discrete probability distribution** assigns a probability to each value:

$$P(X = x_1), \quad P(X = x_2), \quad \dots, \quad P(X = x_n).$$

These probabilities must satisfy

$$0 \leq P(X = x_i) \leq 1$$

and

$$P(X = x_1) + P(X = x_2) + \dots + P(X = x_n) = 1.$$

Once we have the distribution, probability questions reduce to simple addition. For example, in our coin flip example from before we can evaluate $P(X \geq 1)$, the probability of getting one or more heads, by simply adding up the probabilities associated to the outcomes that would make that event true:

$$P(X \geq 1) = P(X = 1) + P(X = 2) = \frac{2}{4} + \frac{1}{4} = \frac{3}{4}.$$

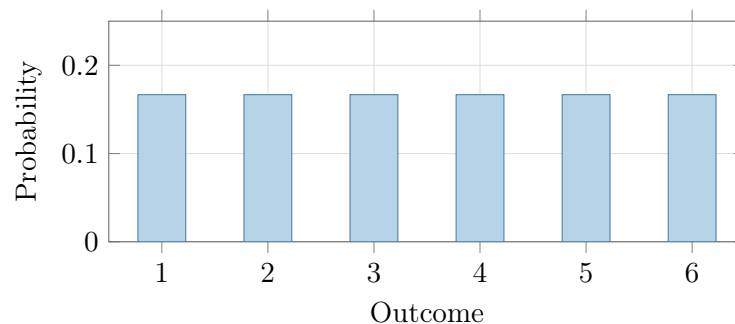
In words, this is the probability of getting at least one head. Similarly,

$$P(X < 2) = P(X = 0) + P(X = 1) = \frac{1}{4} + \frac{2}{4} = \frac{3}{4}$$

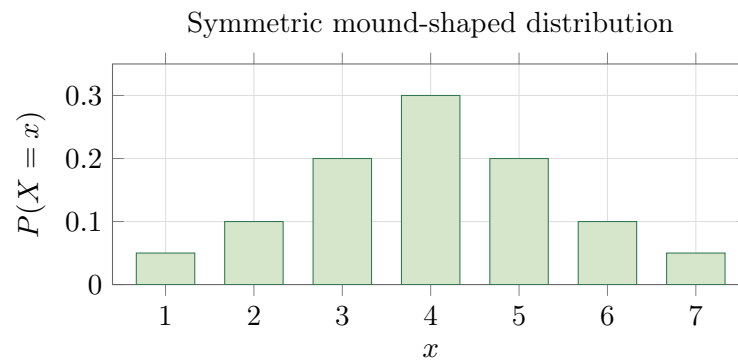
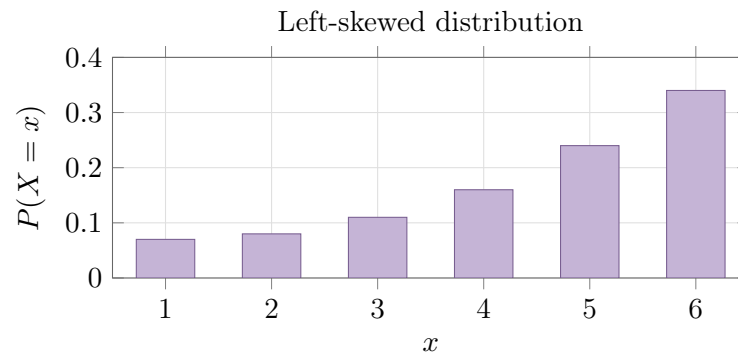
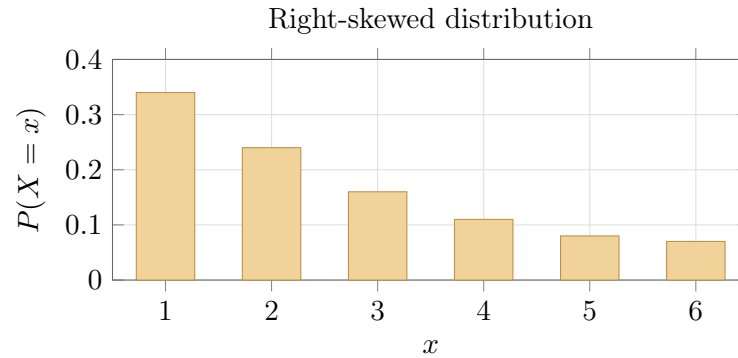
calculates the probability of getting strictly less than 2 heads.

6.3.4 The shape of a distribution

As we saw in Lecture 3, the shape of a histogram can quickly inform us of the overall structure of a distribution. The same idea applies here with probability distributions. Perhaps the simplest case is that of classical probability. Here each outcome has the same exact probability, and therefore any such distribution will be uniform:



However, other systems may admit different shapes. For example, the probability may be skewed to the left or to the right, or even symmetric mound-shaped:



Throughout Lectures 8 and 9 we will talk extensively about discrete probability distributions. As we will see, binomial distributions are particularly interesting setting for probability.

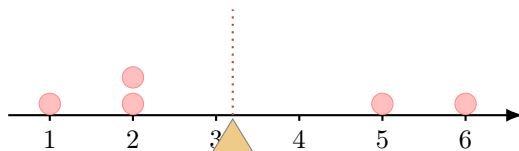
6.3.5 Expected value: the balancing point

Probability distributions are formed by taking a list of probabilities associated to some system, interpreting them as *data*, and then applying descriptive statistics to them. We will now take this one step further by asking: where is the centre?

For an ordinary data set, we used the mean as one notion of centre. Recall that the mean was defined as a geometric center of mass, or “balancing point” for the data: the defining property

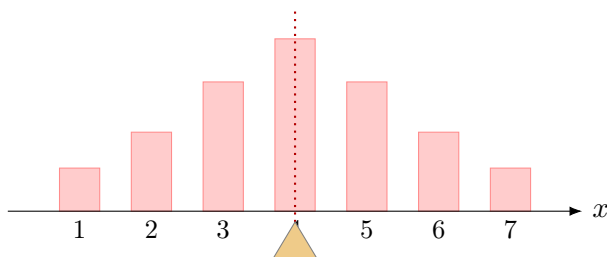
$$\sum_{i=1}^n (x_i - \bar{x}) = 0$$

said that the sum of all the signed distances between data points and the mean $\bar{x}$ must equal zero. From here we recovered the usual formula for $\bar{x}$ as the sum of all data points divided by the number of them. Visually, we pictured the data as a collection of weighted balls on the number line:



Assuming that the balls all have the same weight of 1, the terms $(x_i - \bar{x})$ then give us the *moment*, i.e. how much the weighted ball at x_i tries to rotate the entire system about the pivot point $\bar{x}$.

Trying to calculate the “balancing point” of a probability distribution works in a similar way. We can imagine the distribution as a collection of weighted blocks on the number line. The geometric center of mass of this system will then be the “balancing point” of the system of blocks:



The only important difference here is that the bars will now all have different weights. At the very least, we know that the *sum* of the bars will equal 1. However, their effect on the system will depend on their distance *and* weight. For now, let's write our mystery balancing point with the letter b . Since the different weights of the bars will affect the system, the defining property of the balancing point will now look like:

$$\sum \text{weights} \times \text{distances} = \sum_{i=1}^n p_i(x_i - b) = 0.$$

We may now use some algebra to rearrange this to solve for b . First, we will simply write out the sum in full:

$$p_1(x_1 - b) + p_2(x_2 - b) + \cdots + p_n(x_n - b) = 0.$$

We can now expand out all of the brackets individually, to get:

$$p_1x_1 - p_1b + p_2x_2 - p_2b + \cdots + p_nx_n - p_nb = 0,$$

and then we can collect all of the terms with b 's together:

$$p_1x_1 + p_2x_2 + \cdots + p_nx_n - (p_1 + p_2 + \cdots + p_n)b = 0.$$

From here, we can remember that the sum of all the probabilities has to equal 1, so in fact the b term simplifies heavily:

$$p_1x_1 + p_2x_2 + \cdots + p_nx_n - b = 0.$$

Finally, we can rearrange to get our desired description of the balancing point itself:

$$b = p_1x_1 + p_2x_2 + \cdots + p_nx_n.$$

We have just derived a formula for the “mean” of a probability distribution. Since we are now working with a system of probabilities instead of a data set, this number has a different interpretation: it is the average expected behaviour of the system. For this reason, we don’t call b the “mean” anymore; instead it is known as the *expected value*.

Expected value

For a discrete random variable X with values $x_1, x_2, \dots, x_n$, and probabilities

$$p_i = P(X = x_i),$$

the **expected value** of X is

$$E(X) = \sum_{i=1}^n p_i x_i = p_1x_1 + p_2x_2 + \cdots + p_nx_n.$$

As an example, suppose that we have our random variable X that counted the number of heads in two coin flips:

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

Then

$$E(X) = 0 \cdot \frac{1}{4} + 1 \cdot \frac{2}{4} + 2 \cdot \frac{1}{4} = 1.$$

So if X counts the number of heads in two fair coin flips, then the long-run average number of heads is 1.

6.3.6 Variance: average squared distance from the centre

Recall that for an ordinary data set, measures of central tendency cannot tell us how spread out the data is. Instead, we need to use some kinds of measure of *variation*. In Lecture 3 we introduced several of these, however, the most important for now is the *variance*. For population data, the variance was defined as:

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N},$$

and we had a similar formula for sample variation. In simple terms, the numerator of σ^2 adds up all of the squared distances between our data points x_i and the center $\bar{x}$. Dividing by N , the number of data points, creates a type of “average squared distances from the center”. In a way, we can see this as taking the mean of the new data set $(x_i - \bar{x})^2$. This is an appropriate measure of spread: if the number σ^2 is very high, then on average the data is quite far away from the middle, and if the number σ^2 is very low, then on average the data is very close to the middle.

For probability distributions, the expected value $E(X)$ plays the role of the mean. This means to recreate the formula above, we would need to:

- take the distances between points and the centre $E(X)$, then
- square all these distances, then
- take the “mean” of all of these squared distances.

In symbols, this step-by-step process would be:

$$X - E(X) \longrightarrow (X - E(X))^2 \longrightarrow E((X - E(X))^2).$$

This formula gives us the average expected squared distance from the centre of the distribution, which is the probabilistic version of variance.

Variance

The **variance** of a discrete random variable X is

$$\text{Var}(X) = E((X - E(X))^2).$$

Equivalently, if X has values $x_1, x_2, \dots, x_n$ with probabilities $p_1, p_2, \dots, p_n$, then

$$\text{Var}(X) = \sum_{i=1}^n p_i (x_i - E(X))^2.$$

Geometrically, variance is the average squared distance from the balancing point. A small variance means the distribution is clustered tightly around its centre. A large variance means the distribution is spread far away from its centre.

Careful

The formula is

$$\text{Var}(X) = E((X - E(X))^2),$$

not

$$(E(X - E(X)))^2.$$

The difference matters. Since $E(X - E(X)) = 0$, the second expression would always be 0, which would be useless as a measure of spread.

6.3.7 Example: rolling two dice

Suppose that we roll two fair dice, and then we add the two results together to get a single number. For example, rolling a 2 and a 5 would give the answer “7”, or rolling a 3 and a 6 would give the answer “9”. We define a random variable X by:

$$X = \text{the sum of the two dice.}$$

Since we are rolling two dice, there are 36 different possible outcomes, which we could describe with the set of ordered pairs:

$$S = \{(1, 1), (1, 2), \dots, (6, 6)\}.$$

Since the dice are independent and fair, these 36 different outcomes are all equally likely. However, since X is defined to be the sum of the two dice, the possible values of X are *not* equally likely. To

see this, notice that there is only one way to get a sum of 2:

$$(1, 1),$$

but there are six ways to get a sum of 7:

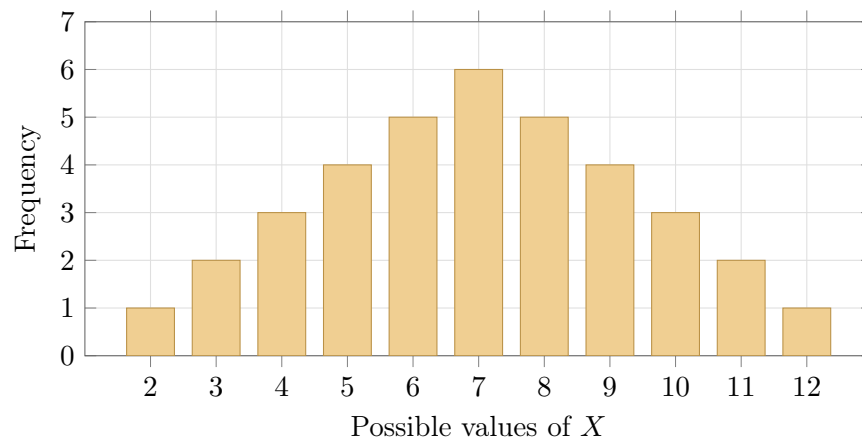
$$(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1).$$

This would mean that X taking the value of 2 should be less likely than taking the value of 7. Looking at all the combinations of the sum of the two dice, we have the following:

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

We may now count up how often each sum appears and divide by 36 to get the following probability distribution:

Sum x	2	3	4	5	6	7	8	9	10	11	12
Frequency	1	2	3	4	5	6	5	4	3	2	1
Probability $P(X = x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$



Exercise 6

Let X be the sum obtained by rolling two fair dice. Use the distribution table above to calculate the following.

- (a) $P(X = 7)$.
- (b) $P(X \geq 10)$.
- (c) $P(X \text{ is even})$.
- (d) $P(X < 5)$.

Solution

(a) There are 6 ways to get a sum of 7, so

$$P(X = 7) = \frac{6}{36} = \frac{1}{6}.$$

(b) The values at least 10 are 10, 11, 12. Therefore

$$P(X \geq 10) = \frac{3}{36} + \frac{2}{36} + \frac{1}{36} = \frac{6}{36} = \frac{1}{6}.$$

(c) The even sums are 2, 4, 6, 8, 10, 12. Therefore

$$P(X \text{ is even}) = \frac{1}{36} + \frac{3}{36} + \frac{5}{36} + \frac{5}{36} + \frac{3}{36} + \frac{1}{36} = \frac{18}{36} = \frac{1}{2}.$$

(d) The values less than 5 are 2, 3, 4. Therefore

$$P(X < 5) = \frac{1}{36} + \frac{2}{36} + \frac{3}{36} = \frac{6}{36} = \frac{1}{6}.$$

Finally, let us compute the expected value of the two-dice sum:

$$E(X) = \sum_{x=2}^{12} xP(X = x).$$

Using the table, this becomes

$$E(X) = 2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + 4 \cdot \frac{3}{36} + 5 \cdot \frac{4}{36} + 6 \cdot \frac{5}{36} + 7 \cdot \frac{6}{36} + 8 \cdot \frac{5}{36} + 9 \cdot \frac{4}{36} + 10 \cdot \frac{3}{36} + 11 \cdot \frac{2}{36} + 12 \cdot \frac{1}{36}.$$

So

$$E(X) = 7.$$

This does not mean that we will always roll a 7. It means that if we roll two dice many times and average the sums, the long-run average will approach 7, i.e. we would *expect* $X = 7$ to be the most likely outcome, on average.

Lecture 7

Conditional Probability

In this lecture, we develop *conditional probability*. As we briefly saw in Lecture 5, conditional probabilities are those probabilities that are calculated after we have learned some kind of extra information about the system that we are working with. We will now study this concept more carefully by looking at some general calculations outside of classical probability. We will also introduce *decision trees*, which give us a practical way to organise multi-step probability problems. Finally, we discuss three famous examples involving conditional probabilities: the Monty Hall problem, the Birthday Paradox, and medical testing.

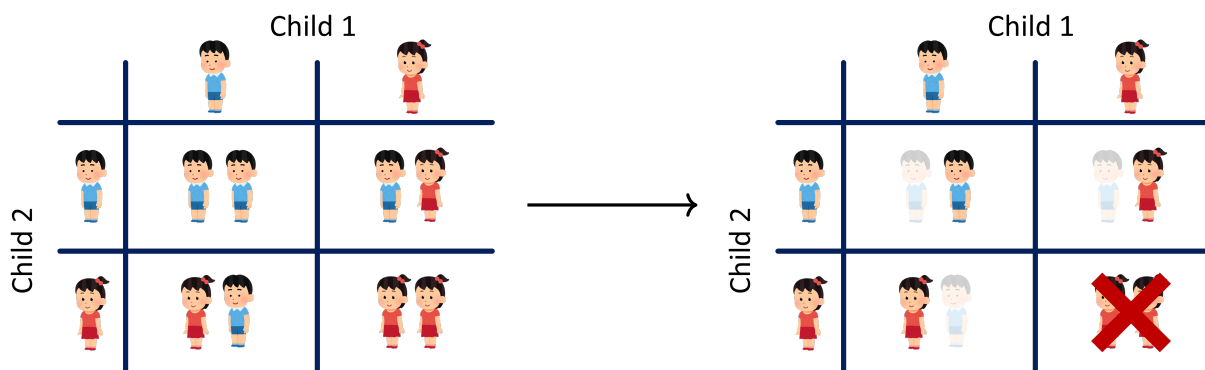
7.1 Conditional Probability in General

Suppose that you know your neighbour has two children. You go over to their house and knock on the door, and a boy answers the door. What is the probability that the other child is a boy?

Amazingly, the answer is actually $\frac{1}{3}$. This is quite surprising when heard for the first time, however, it is indeed true. Let's break down the problem briefly. Note first that there are four possible ordered outcomes:

$$S = \{BB, BG, GB, GG\}.$$

Naively, we would expect there to be a $\frac{1}{4}$ chance of any of these outcomes. However, by seeing that at least one child is a boy, we receive an information update: we now *know* that the “GG” option is no longer available. This reduces the space of possibilities to:



In other words, the sample space has effectively been reduced to:

$$\{(\text{Boy}, \text{Boy}), (\text{Girl}, \text{Boy}), (\text{Boy}, \text{Girl})\}.$$

Only one of these three outcomes has two boys. Therefore:

$$P(\text{two boys, given at least one boy}) = \frac{1}{3}.$$

We have just demonstrated a calculation of *conditional probability*, which calculates uncertainty based on information updates or extra assumptions. As you can see, the results of conditional probabilities are often surprising and counterintuitive. Today we will introduce a few methods to navigate this potentially confusing mess. We will start with a brief review of some theory that we have seen in the past few lectures, and then we will apply this to some basic computations. After that, we will work through *decision trees*, which allow us to organise probabilities into a very understandable system of outcomes. Finally, we will end today with some famous examples of counterintuitive results involving conditional probability: The Monty Hall problem, the Birthday paradox, and the validity of medical testing.

7.1.1 A brief review

Recall that in classical probability, all individual outcomes are equally likely. As we discussed in Lecture 6, this is a special assumption that offers us a unique way to determine probabilities:

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

The formula above says that since every outcome is assumed to be equally likely, we can determine $P(A)$ by simply counting up the size of any event A and calculating the proportion of the sample space S that A occupies.

In contrast, when working with general probability, we relax this assumption that every outcome has an equal chance of happening. Writing ω for an individual outcome, in general probability we do not assume that all probabilities $p(\omega)$ are the same. As a consequence, we cannot simply count up the size of an event A in order to determine its probability. Instead, we must add up the probabilities of the outcomes contained within A . The result of this sum will then be the probability $P(A)$:

$$P(A) = \sum_{\omega \text{ in } A} p(\omega).$$

As we mentioned in the previous lecture, the formulas for compound events are still valid in the general setting. Their justification is different, however, they can still be used as expected.

1. To determine the probability of an event A *not* happening, we may use the formula:

$$P(\text{not } A) = 1 - P(A).$$

2. To determine the probability of two events A and B happening at the same time, we may use either of the formulas:

$$P(A \text{ and } B) = \begin{cases} P(A) \cdot P(B), & \text{if } A \text{ and } B \text{ are independent,} \\ P(A | B) \cdot P(B), & \text{are dependent, provided } P(B) \neq 0. \end{cases}$$

3. To determine the probability of either of two events A or B happening (or both), we may use either of the formulas:

$$P(A \text{ or } B) = \begin{cases} P(A) + P(B), & \text{if } A \text{ and } B \text{ are mutually exclusive,} \\ P(A) + P(B) - P(A \text{ and } B), & \text{if } A \text{ and } B \text{ are not mutually exclusive.} \end{cases}$$

We mentioned in Lecture 5 that the symbol $P(A | B)$ represents the probability of A occurring given that we *know* B will definitely occur. In that discussion, we mentioned that for classical probabilities, we can use the formula for $P(A \text{ and } B)$ to discover the formula for $P(A | B)$. Recall that if we intentionally insert a copy of $1 = \frac{|B|}{|B|}$ into the classical version of our Golden Formula, we get:

$$P(A \text{ and } B) = \frac{|A \cap B|}{|S|} = \frac{|A \cap B|}{|S|} \cdot \frac{|B|}{|B|} = \frac{|A \cap B|}{|B|} \cdot \frac{|B|}{|S|}.$$

The second factor is just $P(B)$, and the first factor gives us a new formula:

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

In probabilistic terms, this formula determines the probability of A given that we start with B as our sample space instead of S . Put differently, it is the probability of A given that B is guaranteed.

In the case of general probability theory, there is a similar formula:

$$P(A | B) = \frac{\sum_{\omega \in A \cap B} p(\omega)}{\sum_{\omega \in B} p(\omega)} = \sum_{\omega \in A \cap B} \frac{p(\omega)}{P(B)}, \quad \text{provided } P(B) \neq 0.$$

This formula appears to be quite unattractive. For now, it is good enough for you to recognise that it exists, and in a moment we will demystify this with a clear example.

In either case (classical or otherwise), we may always rearrange the formula for $P(A \text{ and } B)$ to obtain another expression of conditional probability:

$$P(A \text{ and } B) = P(A | B) \cdot P(B) \quad \longrightarrow \quad P(A | B) = \frac{P(A \text{ and } B)}{P(B)}, \quad P(B) \neq 0.$$

The same process applies if we decide to switch the roles of A and B :

$$P(B \text{ and } A) = P(B | A) \cdot P(A) \quad \longrightarrow \quad P(B | A) = \frac{P(A \text{ and } B)}{P(A)}, \quad P(A) \neq 0.$$

These formulas give us a useful method for finding conditional probabilities: if we happened to know both $P(A \text{ and } B)$ and the probability of the event that we are assuming to know, then we can use either of the two formulas above to calculate conditional probabilities. We will now demonstrate this via two examples.

7.1.2 Example 1: Conditional Probability in a classical setting

In Lecture 5 we saw conditional probability briefly. At that time, we touched on the following example. Suppose that you roll a fair dice, and consider the two events:

$$A = \{\text{the dice lands on an even number}\},$$

and

$$B = \{\text{the dice lands on a number less than 4}\}.$$

Since we are working with an ordinary 6-sided dice, the sample space and events A and B can be written in set notation as:

$$S = \{1, 2, 3, 4, 5, 6\}, \quad A = \{2, 4, 6\}, \quad B = \{1, 2, 3\}.$$

Now, imagine that as you throw the dice and it is spinning up in the air, time slows down for a moment and God descends from the heavens. He whispers in your ear “the dice will land on a number less than 4”. Now, in that moment, you would absolutely know that the event B will happen. In turn, this might adjust your uncertainty for event A happening.

To figure out the new probability of A happening given that B is guaranteed, we can effectively reduce the entire sample space to the smaller subset of numbers less than 4:

$$S = \{1, 2, 3, \cancel{4}, \cancel{5}, 6\} \longrightarrow \{1, 2, 3\}$$

In this restricted system, we notice that there are now three possible outcomes, and only one of them happens to be an even number. Therefore, the probability of landing an even number will now *change* from the initial $\frac{1}{2}$ to $\frac{1}{3}$. What we have just done is calculate the conditional probability of event A *given* B . In symbols, the conditional probability $P(A \mid B)$ is calculated by rearranging the formula for the probability of A and B :

$$P(A \text{ and } B) = P(A \mid B) \cdot P(B) \longrightarrow P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}.$$

By inspection, we can see that the intersection of the sets A and B is the set $A \cap B = \{2\}$. Thus, in symbols, the argument we made above can be written:

$$P(A \mid B) = \frac{|A \cap B|}{|B|} = \frac{|\{2\}|}{|\{1, 2, 3\}|} = \frac{1}{3}.$$

Perhaps it is simpler to directly calculate the probabilities $P(A \text{ and } B)$ and $P(B)$ first, and then just substitute them into the formula:

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{1}{6} \div \frac{3}{6} = \frac{1}{3}.$$

7.1.3 Example 2: Conditional Probability in a general case

Now suppose that we replace our dice with an unfair one, with new probabilities:

$$p(1) = 0.1, \quad p(2) = 0.2, \quad p(3) = 0.1, \quad p(4) = 0.1, \quad p(5) = 0.1, \quad p(6) = 0.4.$$

Using the same events and sample space as before:

$$S = \{1, 2, 3, 4, 5, 6\}, \quad A = \{2, 4, 6\}, \quad B = \{1, 2, 3\}.$$

Observe now that we are *not* working with classical probability, meaning that we cannot simply count up the sizes of subsets in order to determine probabilities. Instead, we have to add up the probabilities of the individual outcomes contained within each event. So, the probabilities of A and B are now:

$$\begin{aligned} B = \{1, 2, 3\} &\implies P(B) = p(1) + p(2) + p(3) = 0.1 + 0.2 + 0.1 = 0.4 \\ A = \{2, 4, 6\} &\implies P(A) = p(2) + p(4) + p(6) = 0.2 + 0.1 + 0.4 = 0.7 \end{aligned}$$

So, according to this unfair system, there is a 70% chance of landing on an even number whenever we throw the dice. Suppose now that, again, when we throw the dice up into the air, God comes down and tells us that the outcome will definitely be a number less than 4. In this situation, our uncertainty regarding event A will again change, because we have removed the possibilities of landing a 4, a 5, or a 6. Again, based on our given information, the new system has a reduced sample space:

$$S = \{1, 2, 3, \cancel{4}, \cancel{5}, 6\} \longrightarrow \{1, 2, 3\}$$

Roughly speaking, the probabilities of the outcomes 1, 2 and 3 now need to be adjusted so that they all add up to 1 in this new system. Mathematically, the way to do that is to divide each of them by the probability $P(B) = 0.4$. Observe, when we do that, the sum of all the possible outcomes indeed adds to 1:

$$\frac{p(1)}{P(B)} + \frac{p(2)}{P(B)} + \frac{p(3)}{P(B)} = \frac{0.1}{0.4} + \frac{0.2}{0.4} + \frac{0.1}{0.4} = \frac{0.4}{0.4} = 1.$$

So, the conditional probabilities of individual outcomes inside B can be found by dividing their original probabilities by the total probability of the reduced sample space. In particular, if we look at the new probability for landing on a 2, we have:

$$p(2 | B) = \frac{p(2)}{P(B)} = \frac{0.2}{0.4} = \frac{1}{2}.$$

Here, $p(2 | B)$ is shorthand for $P(\{2\} | B)$. On some level, this makes sense: notice that in the original system, the three outcomes 1, 2 and 3 have probabilities 0.1, 0.2 and 0.1, respectively. The probabilities of landing a 1 or a 3 are equal at 0.1, and the probability of landing a 2 is twice as much at 0.2. So, if we were to suddenly restrict our attention to only these three outcomes, we would expect that this ratio is still maintained: $P(2 | B)$ should be twice as big as $P(1 | B)$ and $P(3 | B)$, and these latter two should be equal. The only way to achieve this is with the probabilities

$$P(1 | B) = \frac{1}{4}, \quad P(2 | B) = \frac{1}{2}, \quad \text{and} \quad P(3 | B) = \frac{1}{4}.$$

In order to calculate the conditional probability $P(A | B)$ in this system, we can first calculate $P(A \text{ and } B)$ and then use the same formula as before. To emphasize, we are *not* working in classical probability, so to determine the probability of the compound event A and B , we need to look inside the intersection $A \cap B$ and add up each of the probabilities of the outcomes in that set. Since $A \cap B = \{2\}$, there is only one outcome (namely, landing on a 2), and therefore there is only one term in the sum. Simply, we conclude that $P(A \text{ and } B) = 0.2$.

Now, we can use the formula for conditional probability to confirm our previous argument:

$$P(A | B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{0.2}{0.4} = 0.5,$$

which is the same answer as before.

Exercise

Calculate the conditional probability $P(B | A)$ for the unfair dice.

Solution

We want to calculate $P(B | A)$, meaning the probability that the dice lands on a number less than 4, given that we already know it landed on an even number. The intersection is

$$A \cap B = \{2\},$$

which gives us $P(A \text{ and } B) = 0.2$, as before. However, now we need to calculate $P(A)$ before applying the formula. We have:

$$P(A) = p(2) + p(4) + p(6) = 0.2 + 0.1 + 0.4 = 0.7,$$

and therefore

$$P(B | A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{0.2}{0.7} = \frac{2}{7}.$$

7.1.4 Bayes' Theorem

As the previous exercise demonstrates, the two conditional probabilities

$$P(A | B) \quad \text{and} \quad P(B | A)$$

are not the same in general. However, they are still related to each other. According to our discussion in Section 1.1, we have two different ways to determine the probability $P(A \text{ and } B)$:

$$P(A \text{ and } B) = P(A | B) \cdot P(B), \quad \text{and} \quad P(A \text{ and } B) = P(B | A) \cdot P(A).$$

We may set these two equal to each other and rearrange to discover that:

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

This relationship is known as *Bayes' Theorem*.

Bayes' Theorem

For a pair of events A and B , if $P(A) \neq 0$ and $P(B) \neq 0$ then

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

Bayes' Theorem is particularly useful when we want to calculate one conditional probability when we know the other. For example, based on our example of the unfair dice in the previous section, we have:

$$P(B | A) = \frac{P(A | B) \cdot P(B)}{P(A)} = \frac{0.5 \times 0.4}{0.7} = \frac{2}{7}.$$

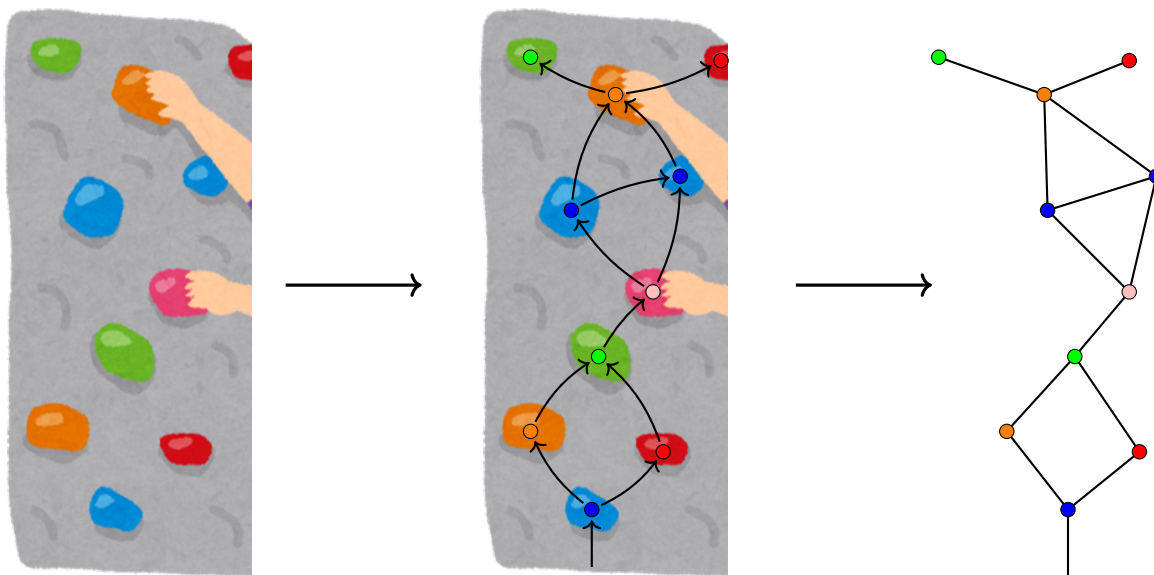
7.2 Decision Trees

Conditional probabilities naturally arise when studying situations in which successive choices might affect each other. In small systems, it can be extremely helpful to simply try to map out the space

of all possible things that can happen. One method for doing this is to use a *decision tree*, which is a diagram that branches out depending on each step of a multi-stage process. To motivate this, we will start with a rather complicated example.

7.2.1 A motivating example: the rock climber

Suppose that we follow the hand of somebody climbing up a rock face. We can idealise this situation by creating a sort of “graph” which displays how the network of coloured rocks is connected to each other:



In this picture, as the climber moves up the rock, their hand will trace a path from the bottom blue rock upwards to either the green or red rock at the top. Notice that at some stages of this process, the rock climber is presented with a choice of how to proceed. For example, at the very start of the rockface, the climber can either choose to go from the blue rock to the orange, or perhaps to the red. At other points, the climber cannot choose what to do, it is simply determined. For example, from the first orange rock they must follow the path to the next green rock.

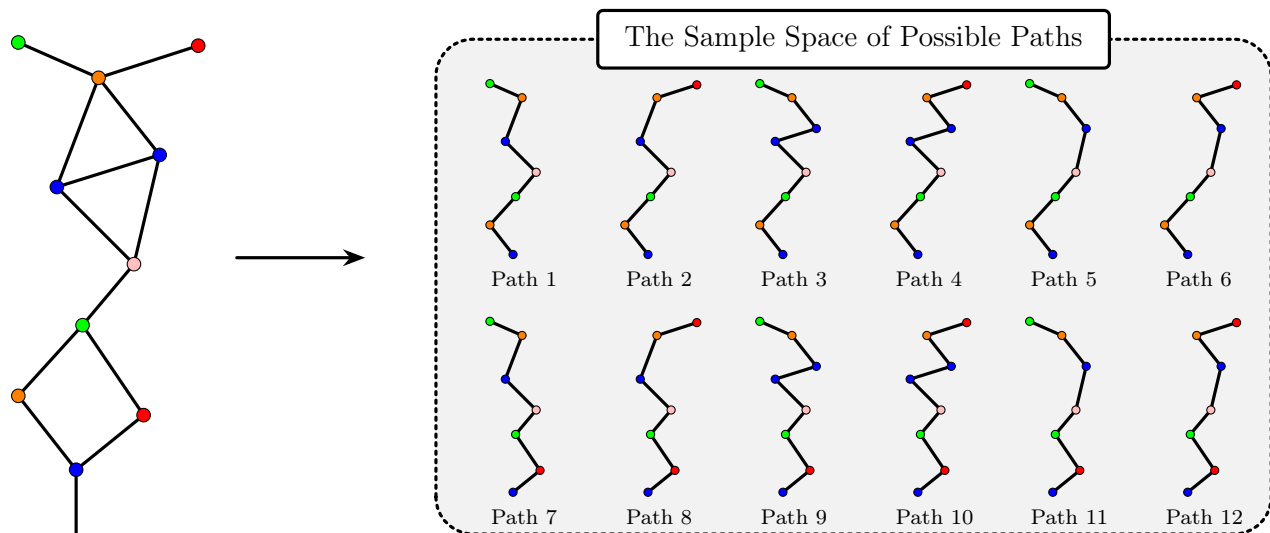
Suppose that every time the climber makes a choice, there is a 50% chance that they pick the left rock, and a 50% chance that they pick the right rock. Our central question is:

What is the probability that the climber takes any given path?

Assuming that there are 12 paths, a first natural guess might be to apply the Golden Formula of classical probability:

$$P(\text{a given path}) = \frac{1}{\text{total number of paths}}.$$

Firstly, we should describe the sample space, which in this case is the space of all paths from the bottom to the top. With a bit of time, you can convince yourself that there are actually 12 unique paths through the rock face. They are pictured below:



Thus the sample space in this setting is the set

$$S = \{\text{Path 1}, \text{Path 2}, \text{Path 3}, \dots, \text{Path 12}\}.$$

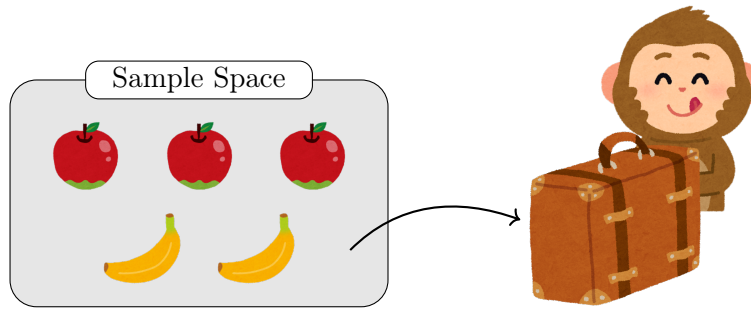
It seems reasonable, therefore, to try and use the above formula to calculate the probability of a given path. For example, for Path 1 we can guess:

$$P(\text{Path 1}) \stackrel{?}{=} \frac{1}{\text{total number of paths}} = \frac{1}{12}.$$

As a matter of fact, this is not correct. Later on we will calculate this properly, but for now we have to simply observe: in the space of paths, notice that some paths are more complicated than others. These different paths **will not** be equally likely outcomes, so the above formula is not the correct approach. We will return to this in a moment. But first, we will illustrate the concept again in a simpler setting.

7.2.2 Example: the jungle and the thief

Suppose that you are in the jungle, and you leave your bag unattended. A local monkey climbs down from his tree and rummages through your bag randomly. As it so happens, there are 5 pieces of fruit in your bag that you packed for lunch: 3 apples and 2 bananas. The monkey starts stealing fruit randomly and eats whatever he finds before looking in the bag again. In probability theory, this process is called *drawing without replacement*.



Exercise 1

- (a) What is the probability that the monkey steals an apple first?
- (b) What is the probability that the monkey steals a banana first?

Solution

There are 5 fruits total, so we may perform a simple count:

- (a) $P(\text{apple first}) = \frac{3}{5}$.
- (b) $P(\text{banana first}) = \frac{2}{5}$.

Now, we will focus on a successive question:

What is the probability he steals a banana *and then* an apple?

It is clear that we should be using a conditional probability here, since the second outcome is clearly affected by the first.

To be precise about the ordering, let B_1 mean “banana on the first draw” and let A_2 mean “apple on the second draw”. Then “banana then apple” is the event “ B_1 and A_2 ”.

We can calculate this by using the conditional probability formula. We may calculate the probability using the formula:

$$P(B_1 \text{ and } A_2) = P(A_2 | B_1) \cdot P(B_1).$$

Let’s evaluate this formula one step at a time:

Step 1: We saw previously that $P(B_1) = \frac{2}{5}$.

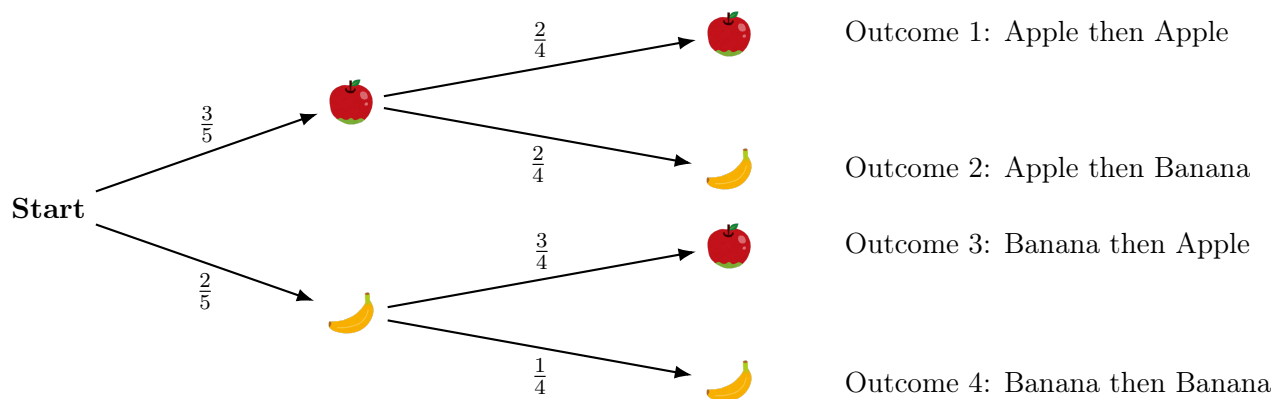
Step 2: We now compute $P(A_2 | B_1)$. Observe that if a banana is removed (and eaten by the monkey), then on the second attempt at taking fruit from the bag, the monkey now has a *different* sample space. In the first attempt at taking fruit, there were 5 pieces of fruit. However, since a banana has already been removed, now there are only 4 pieces of fruit left: 3 apples and 1 banana. This new sample space has size 4, and there are 3 apples that the monkey could select. So, the conditional probability will be:

$$P(A_2 | B_1) = \frac{3}{4}.$$

Step 3: We now take these two numbers and multiply them to get the probability of taking a banana and then an apple:

$$P(B_1 \text{ and } A_2) = P(B_1) P(A_2 | B_1) = \frac{2}{5} \cdot \frac{3}{4} = \frac{3}{10}.$$

Similar reasoning can be applied to any of the other possibilities for our cheeky monkey. We can arrange this entire process into a tree, drawn below:



As you can see, the calculation we did earlier, namely:

$$P(\text{Banana then Apple}) = P(B_1) \cdot P(A_2 | B_1) = \frac{2}{5} \cdot \frac{3}{4} = \frac{3}{10},$$

can also be visually confirmed by checking the branch that corresponds to Outcome 3 in the tree above, because here we are simply multiplying the numbers written along the branch corresponding to Outcome 3. In fact, we have just seen our first instance of a general trick: if we can arrange the outcomes of a multi-step experiment as a tree, then the probability of any outcome can be obtained by multiplying every probability that occurs along that branch.

We can use this observation to quickly calculate the other three outcomes in the tree above:

- Outcome 1: $P(A \text{ then } A) = \frac{3}{5} \cdot \frac{2}{4} = \frac{3}{10}$,
- Outcome 2: $P(A \text{ then } B) = \frac{3}{5} \cdot \frac{2}{4} = \frac{3}{10}$,
- Outcome 4: $P(B \text{ then } B) = \frac{2}{5} \cdot \frac{1}{4} = \frac{1}{10}$.

Observe that the sum of these four probabilities is

$$\frac{3}{10} + \frac{3}{10} + \frac{3}{10} + \frac{1}{10} = 1.$$

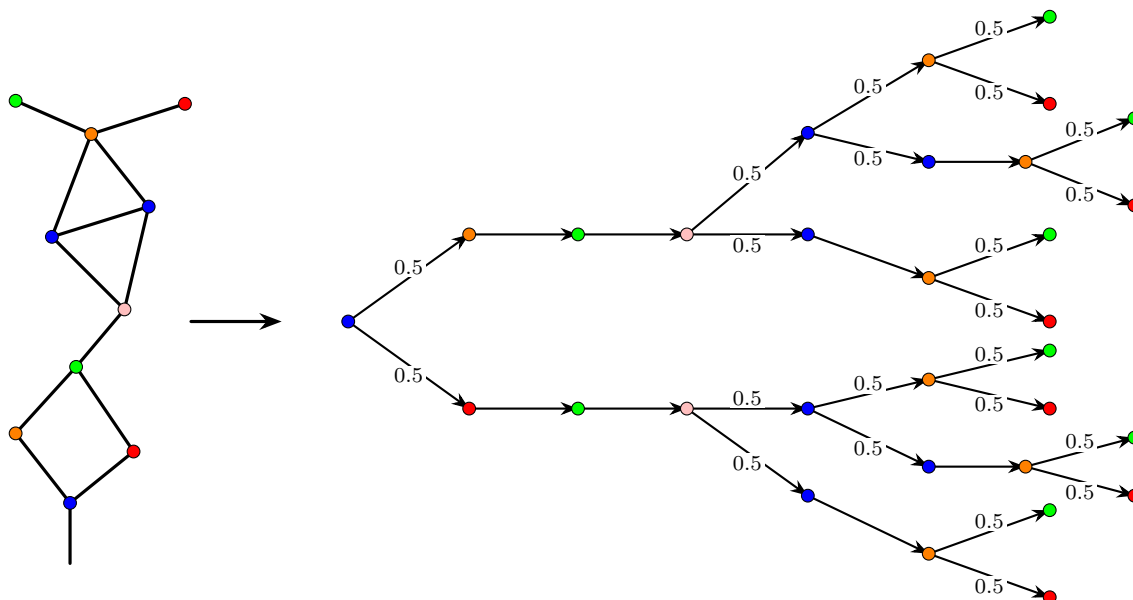
This is the same as saying that it is guaranteed that one of the outcomes must occur.

7.2.3 Returning to the Rock Climber

We've just seen that a decision tree can be used to organize all possible outcomes of a multi-step experiment. To find the probability of a particular path, we simply need to multiply the probabilities

that are written along that particular path. We can use this technique to figure out the probabilities associated with each of our rock climber's routes through the wall.

To begin with, we write the tree for this system and then label each “choice point” with the assumed 50-50 probabilities. The full tree is drawn below.



The above tree is organised so that each path is in order: the first path corresponds to Path 1, the second path corresponds to Path 2, and so on. You should observe here that some paths are more complicated than others, for instance, Path 12 looks simpler than Path 9 or 10. This is because there are some routes up the wall that only require making three choices, and others require making four choices. As a matter of fact, any path that involves making the choice to move to the left from the pink rock will be more complicated than those paths that involve the choice to move to the right instead. In any case, we can use our theory of decision trees to calculate the probability of any given path up the rock face. As with the example of the klepto-monkey, in this case we simply multiply the values written along the branch to obtain the overall probability. These are summarised below.

Probabilities of Each Path

We make a distinction between the number of choices that each path requires:

- Paths 1, 2, 3, 4, 7, 8, 9 and 10 all require making four choices. On the decision tree, there are thus 4 copies of 0.5 written on the branches corresponding to these paths, so their overall probability will be:

$$\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{16}.$$

- Paths 5, 6, 11 and 12 require making only *three* total choices. Therefore, the probability of taking each one of these paths is greater than the others:

$$\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}.$$

We may always double-check that our calculations are correct by confirming that all of our answers add up to 1. We have:

$$8 \left(\frac{1}{16} \right) + 4 \left(\frac{1}{8} \right) = \frac{1}{2} + \frac{1}{2} = 1,$$

which is correct. Notice here that we have also discovered something else: there is a 50% probability that the climber will take one of the simpler-looking paths, and there is a 50% probability that they will not.

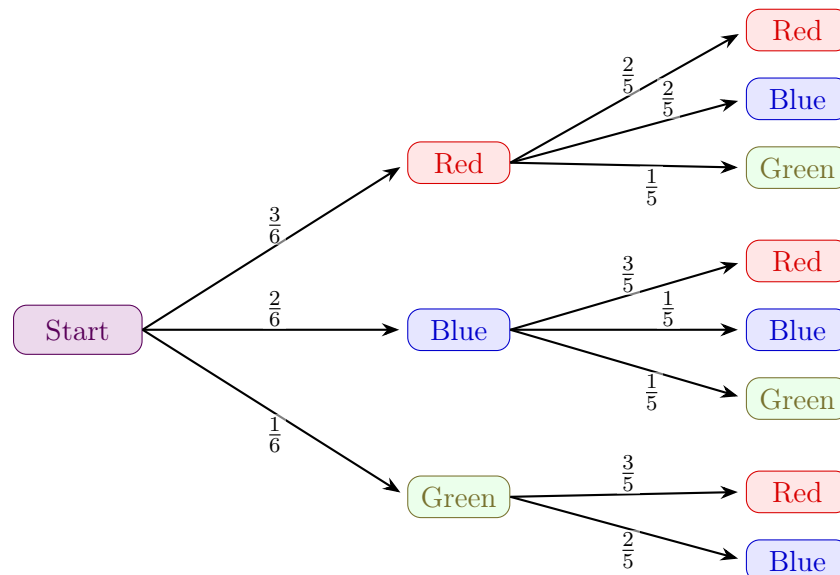
7.2.4 Exercise: Drawing Without Replacement

A box has 6 balls inside: 3 red, 2 blue and 1 green. You take 2 balls at random, without replacing the first ball back into the box.

Exercise 2

- (a) What is the probability of getting 2 reds in a row?
- (b) What is the probability of getting a green then a blue?
- (c) What is the probability of getting a blue and a red ball, in either order?

Before getting to the full solution, let's observe what the tree looks like. For completeness, we have labelled all of the probabilities along each arrow:



As you can see, on the bottom branches there are no longer three possible outcomes. Why? Because there is only 1 green ball in the box. If we select it first, then the box would only have red and blue balls left over. This tree is fully populated with all of the probabilities and conditional probabilities. Instead of calculating these directly using formulas, we can simply argue heuristically, in a similar manner to that of our monkey thief of Section 2.2. For example: to calculate the probability of picking another red ball after just having drawn a red, we are being asked to look at the situation in which one red has already been removed, meaning that the sample space has now been reduced to only 2 reds, plus the 2 blues and 1 green ball that remain in the box. Since there are 2 possible red

balls to select out of 5, the conditional probability $P(\text{red} \mid \text{red})$ will be $\frac{2}{5}$. With this in mind, we will now answer the questions above.

Solution

(a) Red then red:

$$P(\text{red then red}) = \frac{3}{6} \cdot \frac{2}{5} = \frac{6}{30} = \frac{1}{5}.$$

(b) Green then blue:

$$P(\text{green then blue}) = \frac{1}{6} \cdot \frac{2}{5} = \frac{2}{30} = \frac{1}{15}.$$

(c) Blue and red means either red then blue or blue then red:

$$P(\text{red then blue}) = \frac{3}{6} \cdot \frac{2}{5} = \frac{1}{5},$$

and

$$P(\text{blue then red}) = \frac{2}{6} \cdot \frac{3}{5} = \frac{1}{5}.$$

These are mutually exclusive outcomes, so:

$$P(\text{blue and red}) = \frac{1}{5} + \frac{1}{5} = \frac{2}{5}.$$

7.3 Famous Examples

7.3.1 The Monty Hall Problem

Suppose that you are on a game show. There are three doors. Behind one door is a car, and behind the other two doors are goats. The doors are randomized, so you do not know what is behind which door. So, you pick door 1, hoping that there is a car behind. The host, who knows what is behind the doors, then opens a different door and reveals a goat. Assume that the host always opens a different door containing a goat and always offers you the chance to switch. You now have a choice:

stay with your original door or switch to the remaining unopened door.

Should you switch?

Amazingly, under these standard assumptions, the answer is yes: it makes more sense to switch. The probability of winning the car by staying is $\frac{1}{3}$, while the probability of winning the car by switching doors is $\frac{2}{3}$. This is rather confusing, so we will argue it in three different ways.

Argument 1: transfer the probability

Let

$$A = \{\text{your first choice is the door with the car behind it}\}.$$

Then:

$$P(A) = \frac{1}{3},$$

and therefore:

$$P(\text{not } A) = \frac{2}{3}.$$

If your first choice was correct, then switching loses. If your first choice was wrong, then switching wins, because the host removes the other goat. Therefore:

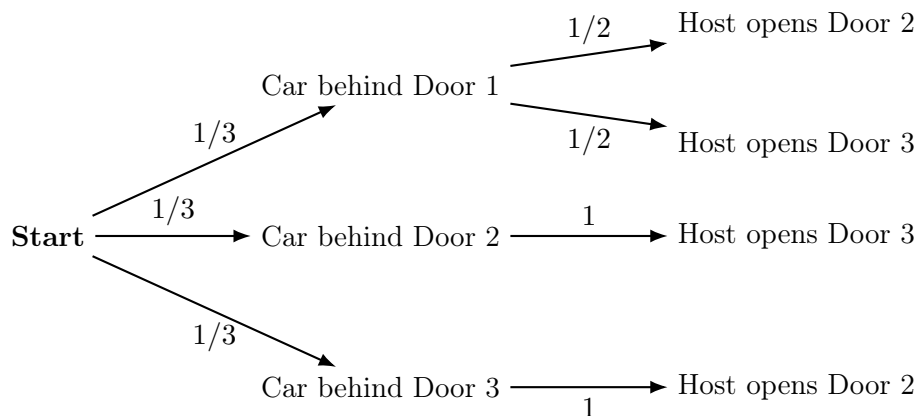
$$P(\text{switching wins}) = P(\text{your first choice was wrong}) = \frac{2}{3}.$$

Argument 2: make the game bigger

Suppose instead that there were 1000 doors, where one door has a car behind it and the other 999 doors have goats. Suppose that you picked door number 1 as before. The host, knowing where the car actually is, then opens 998 of the doors and reveals goats behind them. When you are left with only two doors left: your door and the other, you may start to doubt your original choice. If there were 1000 doors, your original choice would have a probability of $\frac{1}{1000}$ of being correct. That probability does not suddenly become larger just because the host opened doors that he already knew were goats. Instead, your door carries the same $\frac{1}{1000}$ probability of containing the car, and the other remaining door is overwhelmingly more likely to have the car. Clearly, switching doors would be a better idea.

Argument 3: a decision tree

Assuming that you initially picked Door 1, the space of possible outcomes is below. It should be emphasized here that the host *knows* which door the car is actually behind. If he wants to intentionally pick a door that reveals a goat, then sometimes he has a choice and sometimes he does not. When the host has two goat doors available, assume he chooses between them with equal probability.



If the car happened to be behind Door 1, then switching loses. According to the decision tree above, the probability of this happening will be the sum of the probabilities for the top two branches, giving a result of $\frac{1}{3}$. If the car is *not* behind Door 1, then it is either behind Door 2 or Door 3. The probability of this happening is $\frac{2}{3}$. So, we see that switching will cause you to win the car with a probability of $\frac{2}{3}$, and switching will cause you to *lose* the car with a probability of $\frac{1}{3}$.

7.3.2 The Birthday Paradox

Let A be the event “at least two people in the room share a birthday”. In this case, the probability $P(A)$ will change radically depending on n , the number of people in the room. For example, if $n = 2$, then we would expect that the chance that these two people share a birthday is extremely low. However, for $n > 365$, it is guaranteed that two people in the room must share a birthday, since there are only 365 possible birthdays if we ignore leap years. As we can see, when n is very low, so is $P(A)$, and when n is high, $P(A)$ eventually becomes equal to 1. Here is a question:

At what value of n does $P(A)$ become more than 50%?

Put differently: how many people do there need to be in the room so that it is *more likely than not* that two people share a birthday? In fact, the answer is quite surprising: we only require 23 people!

This surprising fact is known as the “Birthday paradox”. It is an example of a “veridical paradox”, which is simply a mathematical result that is so surprising that it *feels wrong*. However, as we will now demonstrate, the Birthday Paradox follows from some basic applications of probability.

Resolving the Birthday Paradox

In order to resolve the birthday paradox, we are going to flip it on its head and rephrase the problem. Instead of computing $P(A)$, we are going to compute $P(\text{not } A)$, the probability of the complement of A . Recall that the complement event is the probability-theoretic analogue of the word “not”. So here, the complement event A^c means “everybody in the room has a distinct birthday”.

Recall that the probability of the complement is simply $1 - P(A)$. So, if $P(\text{not } A) < 50\%$, then $P(A) > 50\%$, because

$$P(A) + P(\text{not } A) = 1.$$

Let’s assume that the choice of each birthday is independent and equally likely, so that any birthday has probability $1/365$. We will now calculate $P(\text{not } A)$ for low n to see how the process works.

For $n = 2$, we are trying to calculate the probability that two people have distinct birthdays. Clearly we are allowed to pick Person 1’s birthday freely. But, in order for the two birthdays to be distinct, we must pick a birthday for Person 2 that avoids our first choice. There are 364 days that are distinct from our first choice, so (if treated as a random event), the probability that Person 2 has a distinct birthday will be

$$P(\text{not } A) = \frac{\text{number of days distinct from Person 1's birthday}}{\text{total number of days in a year}} = \frac{364}{365} \approx 0.997 = 99.7\%.$$

Suppose now that we add a third person to the room, i.e., $n = 3$. We have just established that the probability of two distinct birthdays is $\frac{364}{365}$. But, we also need the third person’s choice of birthday to be distinct from Person 1’s *and* Person 2’s. The probability for the next person’s birthday is conditional on the previous birthdays. To calculate the probability of an “and” statement like this, we need to multiply the associated probabilities:

$$\begin{aligned}
P(\text{not } A) &= \frac{364}{365} \times \frac{\text{number of days distinct from Person 1 and Person 2's birthday}}{\text{total number of days in a year}} \\
&= \frac{364}{365} \cdot \frac{363}{365} \approx 0.992 = 99.2\%.
\end{aligned}$$

We can keep repeating this pattern for larger and larger n . Every time that we increase the value of n by 1, we see that we keep “using up” available birthdays for the next person to be able to pick from. This means that we keep needing to multiply by smaller and smaller numbers after every step. For example:

$$\text{For } n = 4 : P(\text{not } A) = \left(\frac{364}{365}\right) \left(\frac{363}{365}\right) \left(\frac{362}{365}\right) \approx 0.983 = 98.3\%$$

$$\text{For } n = 5 : P(\text{not } A) = \left(\frac{364}{365}\right) \left(\frac{363}{365}\right) \left(\frac{362}{365}\right) \left(\frac{361}{365}\right) \approx 0.973 = 97.3\%$$

$$\text{For } n = 6 : P(\text{not } A) = \left(\frac{364}{365}\right) \left(\frac{363}{365}\right) \left(\frac{362}{365}\right) \left(\frac{361}{365}\right) \left(\frac{360}{365}\right) \approx 0.960 = 96.0\%$$

For general n , we obtain the formula:

$$P(\text{not } A) = \left(\frac{365}{365}\right) \left(\frac{364}{365}\right) \left(\frac{363}{365}\right) \cdots \left(\frac{365 - (n - 1)}{365}\right).$$

The paradox will be resolved by finding the smallest value of n so that the probability $P(\text{not } A)$ drops below 50%, because that would mean that $P(A)$ goes *above* 50%. If we simply use the formula for bigger and bigger values of n , we will see that for $n = 23$ we have:

$$P(\text{not } A) = \left(\frac{365}{365}\right) \left(\frac{364}{365}\right) \left(\frac{363}{365}\right) \cdots \left(\frac{343}{365}\right) \approx 0.493 < 0.5.$$

For comparison, when $n = 22$, we get $P(\text{not } A) \approx 0.524 > 0.5$. So $P(A) > 50\%$ when $n = 23$.

7.3.3 Medical Testing

Suppose that 1% of a population has a disease. There happens to be a medical test for this particular disease, in which:

- if a person has the disease, then the test detects this 99% of the time, but
- if a person does *not* have the disease, then the test will give a false positive 5% of the time.

Suppose that you take the test and you test positive. What is the probability that you have the disease? Amazingly, the answer is about 16.7%. We will now demonstrate how.

To begin with, let's rewrite the above conditions using conditional probabilities.

Let:

$$D = \{\text{the person has the disease}\},$$

and

$$T = \{\text{the person tests positive}\}.$$

The question is asking for the conditional probability:

$$P(D \mid T),$$

that is, the probability of actually having the disease after testing positive. Note that this is not the same as the other way around: $P(T \mid D)$ calculates the probability of a person testing positive, given that they actually have the disease. The setup of the question tells us that

$$P(T \mid D) = 0.99, \quad P(D) = 0.01, \quad P(D^c) = 0.99, \quad \text{and} \quad P(T \mid D^c) = 0.05.$$

We can use this information in combination with Bayes' Theorem to calculate the missing conditional probability:

$$P(D \mid T) = \frac{P(T \mid D)P(D)}{P(T)}.$$

In order to do this, we need to first calculate $P(T)$, the overall probability of testing positive.

Observe that there are two possible ways to test positive:

- either the person has the disease and correctly tests positive, or
- the person does not have the disease but falsely tests positive.

These are mutually exclusive, so we can reinterpret the event T as an “or” statement and calculate its probability as:

$$P(T) = P(T \mid D)P(D) + P(T \mid D^c)P(D^c) = (0.99)(0.01) + (0.05)(0.99) = 0.0594.$$

Using this in Bayes' theorem, we recover the infamous 16.7%:

$$P(D \mid T) = \frac{0.0099}{0.0594} = \frac{1}{6} \approx 0.167.$$

Why this feels surprising

The test is very good at detecting the disease when somebody actually has it. However, the disease is rare and there are a relatively large number of false positives. These false positives drag down the reliability of a single positive result, causing such a surprising result.

Lecture 8

The Binomial Distribution

In this lecture we cover one of the final topics in our discussion of probability theory: binomial experiments. As we will see, these are particular setups in which there are two possible outcomes, namely “success” or “failure”. Given that there are only two possibilities, these systems are quite easy to describe and to calculate, meaning that we can relate them to our previous material in an easy way. After our presentation of the general theory, we will discuss various examples of binomial experiments.

8.1 A brief review: counting and distributions

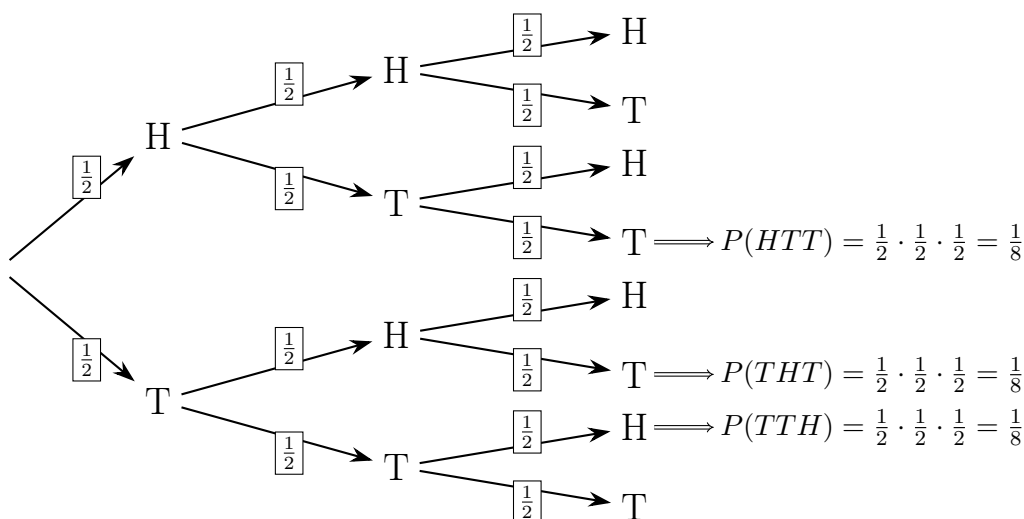
8.1.1 Calculating the probability of 2 Tails in 3 coin flips

Suppose that we have a fair coin, and we flip it 3 times. What is the probability of getting exactly 2 Tails? Fortunately, with a small number like 3 we can simply write out all of the possible outcomes:

- A single coin flip has 2 outcomes: H or T .
- Two coin flips have 4 outcomes: HH, HT, TH, TT .
- Three coin flips have 8 outcomes (the full sample space):

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

There are several ways we can try to calculate a probability like $P(\text{exactly 2 Ts})$. One method could be to use a decision tree (like in Lecture 7) to calculate the probability of each favourable outcome:



The three events HTT , TTH and THT are mutually exclusive, so we can simply add up these individual outcomes and this will give the probability of the compound event:

$$P(\text{exactly 2 Ts}) = P(HTT \text{ or } THT \text{ or } TTH) = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8}.$$

Alternatively, we could arrive at the same answer by working with a simpler calculation of compound events (like in Lecture 5). We can observe that in this situation all of the outcomes are equally likely, so we could also just use the Golden Formula directly:

$$P(\text{event}) = \frac{\text{number of outcomes favourable to the event}}{\text{total number of outcomes}}.$$

In this case there are 3 outcomes that contain precisely two tails, namely HTT , THT and TTH , so:

$$P(\text{exactly 2 Ts}) = \frac{\text{number of outcomes with 2 Ts}}{\text{total number of outcomes}} = \frac{3}{8}.$$

Solution

Total outcomes: $2^5 = 32$. The number of outcomes with exactly 2 tails is $C_{5,2} = 10$ (row 5 of Pascal's triangle). Therefore,

$$P(\text{exactly 2 tails}) = \frac{C_{5,2}}{2^5} = \frac{10}{32} = \frac{5}{16}.$$

8.1.3 Probability distributions

A variable X is called a *random variable* if the value that X takes in a given experiment or observation is determined by a random outcome. There are two main types:

- A *discrete random variable* can take on only a finite number of values, or a countable number of values.
- A *continuous random variable* can take on any of the countless number of values in an interval.

Exercise: discrete or continuous?

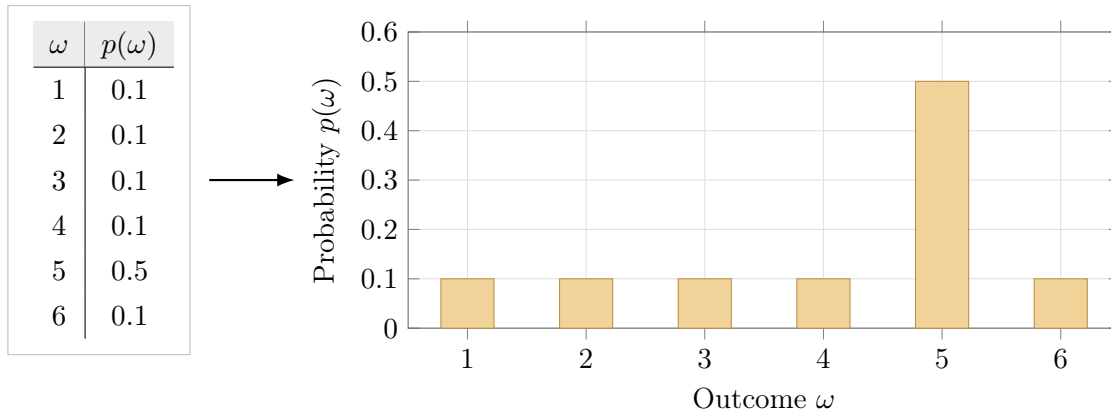
Classify each random variable as discrete or continuous.

- (a) The time it takes a selected student to register for the Fall term.
- (b) The number of text messages received by a selected student on a randomly chosen day.
- (c) The number of miles an electric vehicle can drive on a full charge.
- (d) Pick a random sample of 50 registered voters in a district and find the number who voted in the last county election.

Solution

- (a) Continuous.
- (b) Discrete.
- (c) Continuous.
- (d) Discrete.

A *discrete probability distribution* is an assignment of probabilities to each distinct value of a discrete random variable. In Lecture 6, we saw that probability distributions can be visualised as bar charts. For example, the graph:



visually explains how probability is spread out amongst the possible values 1, 2, 3, 4, 5 and 6.

Normally we write random variables with capital letters, such as X or Y . If X is allowed to take the values 1, 2, 3, 4, 5, 6, then statements like

$$X = 4 \quad \text{or} \quad X = 5$$

represent possible values of the random variable. The probability that X takes the value 4 can then be written:

$$P(X = 4).$$

On a probability distribution, this is represented by the height of the bar at 4. In the example above, we have

$$P(X = 4) = 0.1.$$

We can also ask for the probability that X lies in a range of values. For example:

$$P(X < 3)$$

means the probability that X takes a value less than 3. Since the values less than 3 are 1 and 2, we can calculate the total probability by taking the sum:

$$P(X < 3) = P(X = 1) + P(X = 2) = 0.1 + 0.1 = 0.2.$$

In Lecture 6, we also saw a formula for the *expected value* of a random variable, which acted like the mean of a probability distribution. The expected value of a probabilistic system is a weighted average. For a finite discrete random variable, the expected value is

$$E(X) = \sum_{i=1}^n p_i x_i = p_1 x_1 + \cdots + p_n x_n.$$

This formula can be carefully derived by trying to calculate the geometric center of mass of the probability distribution – for the full explanation you should refer back to Lecture 6. Informally, $E(X)$ is the long-term expected behaviour of X .

The variance of a probability distribution measures how spread out the possible values of X are around their expected value. It is defined by

$$\text{Var}(X) = E((X - E(X))^2).$$

Observe the similarities this formula has with the usual definition of population variance for a data set. In summary:

Expected value and variance

The **expected value** of a discrete random variable is

$$E(X) = \sum_{i=1}^n p_i x_i.$$

It is the probabilistic version of the mean, and therefore gives a measure of the center of the distribution.

The **variance** of a random variable is

$$\text{Var}(X) = E((X - E(X))^2).$$

It measures how much the distribution spreads out around its expected value.

8.1.4 The law of large numbers

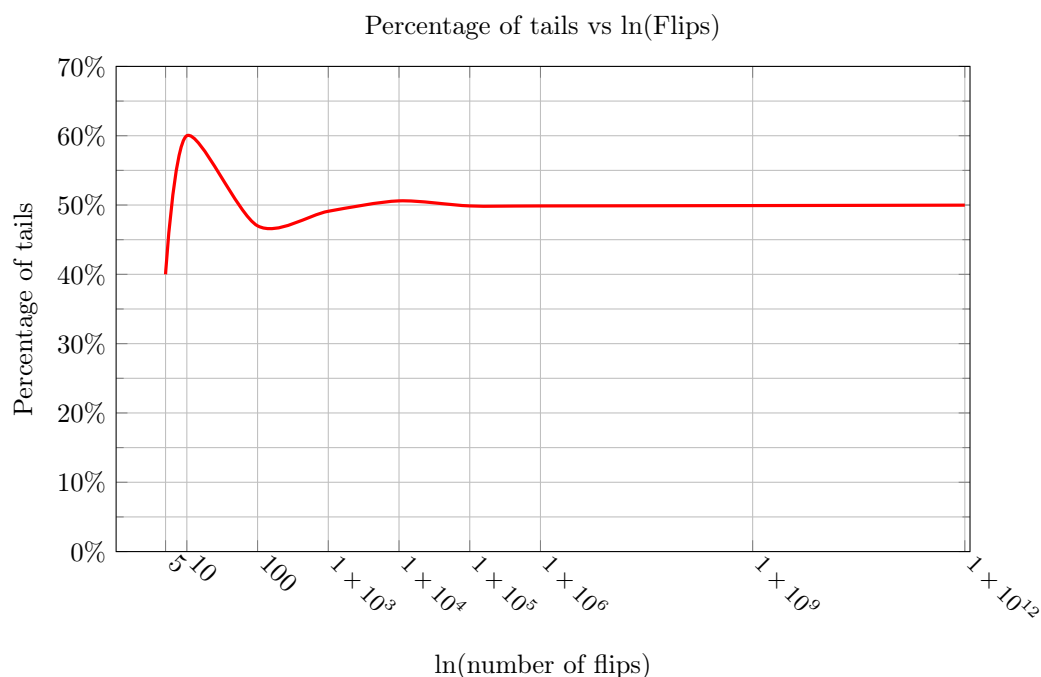
The expected value of a random variable is a way to calculate the average, long-term expected result of an experiment. We will now elaborate on this idea by introducing another interpretation of probability.

The *law of large numbers* says that if you repeat an experiment many times, then the average result tends to get closer to the theoretical average. As an example of this, suppose that we have a fair coin. We expect that the probability of flipping the coin and landing on tails is 50%, i.e. $P(T) = \frac{1}{2}$. Our interpretation of probability says that we expect there to be about 50% tails in the long run.

To see this in action, suppose that we start flipping a coin over and over again. Using a random number generator, below is an example of the proportion of tails in the sample of coin flips:

- 5 flips: HHTTH. $T = 2$, $H = 3$, so tails = 40%.
- 10 flips: HHTTHHTTTT. $T = 6$, $H = 4$, so tails = 60%.
- 100 flips: $T = 47$, $H = 53$, so tails = 47%.
- 1000 flips: $T = 491$, $H = 509$, so tails = 49.1%.
- 10000 flips: $T = 5062$, $H = 4938$, so tails = 50.6%.
- 100000 flips: $T = 49872$, $H = 50128$, so tails = 49.87%.
- 1 million flips: tails = 49.87%.
- 1 billion flips: tails = 49.92%.
- 1 trillion flips: tails = 49.982%.

We can visualize this by plotting the graph (where here we take a logarithmic scale $\ln(\cdot)$ so that the coin flips fit on the same axis):



As you can see, in the beginning there is quite a lot of random fluctuation in the results. However, over time these random changes mostly cancel out, and the proportion of Tails converges to the theoretical average of 50%. This offers us one possible interpretation of probability:

The Frequentist Interpretation of probability

When we say $P(A) = x\%$, we mean that if we repeated the experiment a huge number of times, we expect the long-run proportion of outcomes corresponding to event A would get closer and closer to $x\%$.

In this sense, the expected value $E(X)$ can be better understood as the average result that we would get if we repeated the experiment over and over again under independent conditions. Eventually, the running average of this repeated activity will have results that will converge to something, and this something is the expected value. As we will see later, the expected value is especially useful when discussing Binomial experiments.

8.2 Example: binomial experiments

8.2.1 Example 1: 10 flips of a fair coin

Suppose we have a fair coin, and we flip it 10 times in a row. What is the probability of finding exactly 5 tails?

Solution

Based on the previous section, we can calculate $P(\text{exactly 5 tails})$ using the formula:

$$P(\text{exactly 5 tails}) = \frac{C_{10,5}}{2^{10}}.$$

Of course, we may use the usual formula $C_{n,r} = \frac{n!}{r!(n-r)!}$, but for narrative consistency let's instead use Pascal's triangle. The 10th row of the triangle is:

1 10 45 120 210 252 210 120 45 10 1

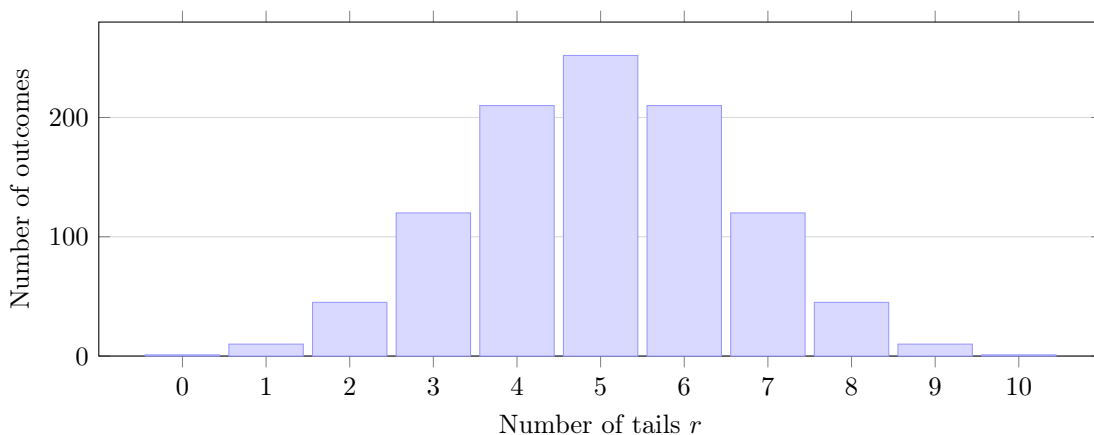
So, we move across to the entry with $r = 5$ (remembering that the triangle starts counting at 0) to see that $C_{10,5} = 252$. The total number of outcomes is $2^{10} = 1024$. Therefore:

$$P(5 \text{ tails}) = \frac{C_{10,5}}{2^{10}} = \frac{252}{1024} \approx 0.246 \approx 24.6\%.$$

We will now try to explore the structure of the sample space of 10 coin flips. Based on the above, we know that there are 1024 members of this space. These will range from *HHHHHHHHHH* to *TTTTTTTTTT* and everything in between. A natural way to navigate this large space is to group outcomes together based on how many tails they contain.¹ A helpful way to understand the sample space is to group outcomes by how many tails they contain. Based on the 10th row of Pascal's triangle, the sample space of 1024 outcomes can be organised as follows.

# of tails r	0	1	2	3	4	5	6	7	8	9	10
# of outcomes	1	10	45	120	210	252	210	120	45	10	1

The table above can be visualised as a bar chart:



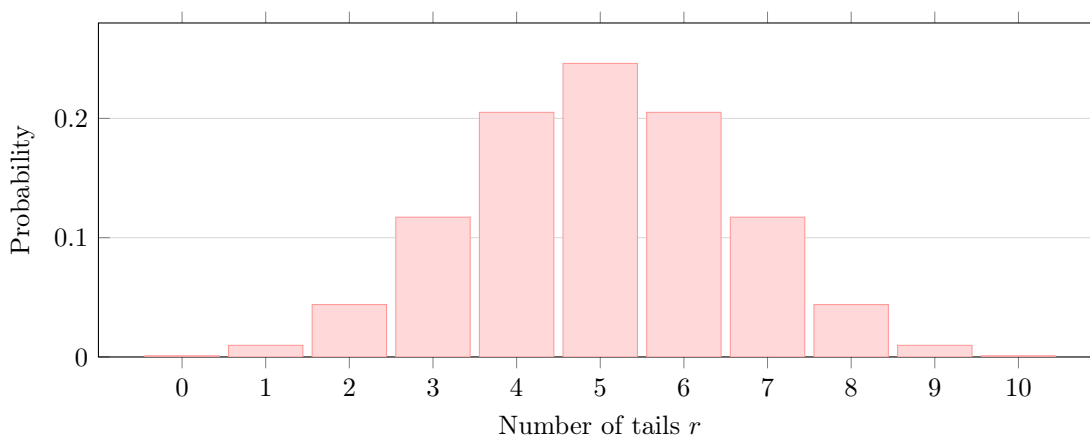
Here the x -axis consists of a discrete count: simply, it is the number of possible tails. The y -axis is the number of corresponding outcomes. So, in a sense, this graph is simply a visualisation of the 10th row of Pascal's triangle.

¹Of course, this could equally be done by starting with heads. But, since we have been using tails this whole time, we will commit to it.

Based on our discussion above, we know that the probabilities in this system can be calculated by dividing $C_{10,r}$ by $2^{10} = 1024$. Therefore, we can perform the division and then plot the results to see how *probability itself* is distributed amongst the outcomes in this system.

r tails	0	1	2	3	4	5	6	7	8	9	10
Probability (fraction)	$\frac{1}{1024}$	$\frac{10}{1024}$	$\frac{45}{1024}$	$\frac{120}{1024}$	$\frac{210}{1024}$	$\frac{252}{1024}$	$\frac{210}{1024}$	$\frac{120}{1024}$	$\frac{45}{1024}$	$\frac{10}{1024}$	$\frac{1}{1024}$
Approx. percentage	0.1%	1.0%	4.4%	11.7%	20.5%	24.6%	20.5%	11.7%	4.4%	1.0%	0.1%

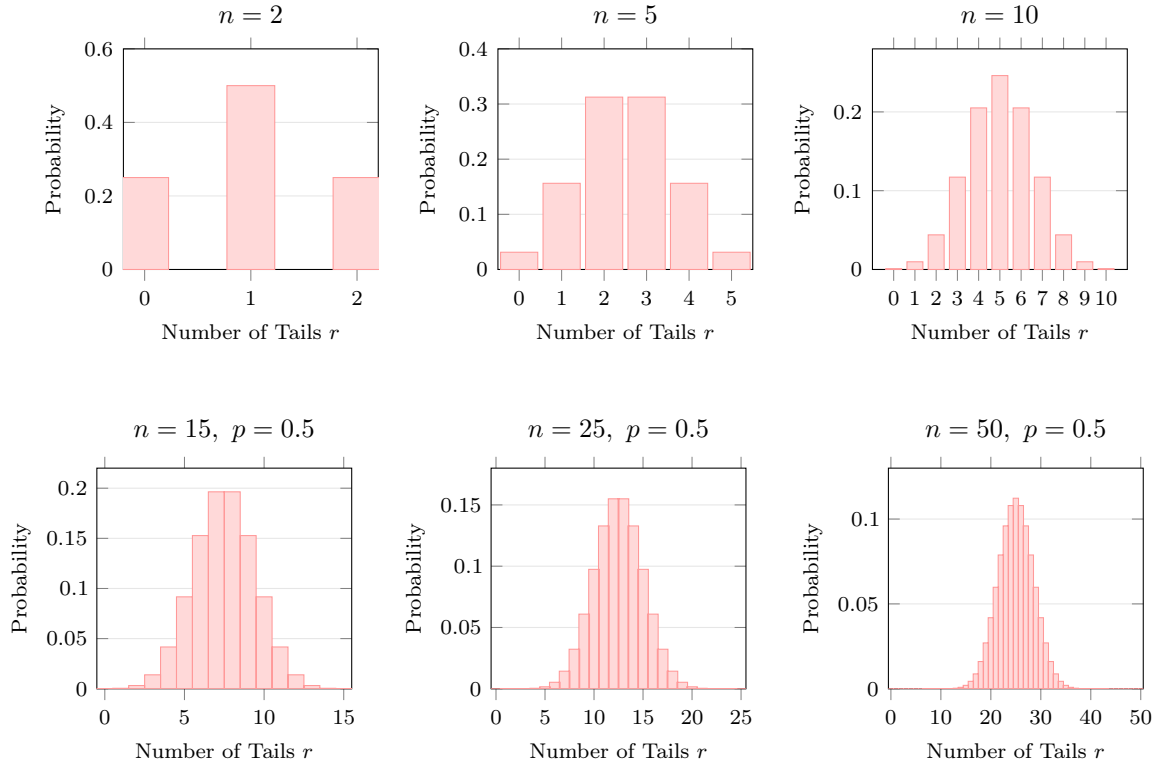
Observe the value of 24.6% that we obtain earlier. Using the above table, we can plot out this distribution of probabilities:



Notice that the shape is *precisely* the same as that of the bar chart from before. The only thing that has changed is the scale of the y -axis. This particular graph has a name: it is called a **binomial distribution**.

8.2.2 The binomial distribution for other values of n

Generally, we can create binomial distributions for other values of n , the total number of coin flips. The shape of these graphs will map out the way in which probability is distributed across the many, many possible outcomes of coin flips. The shape comes from the n^{th} row of Pascal's triangle, and after dividing by 2^n those numbers become probabilities. Generally speaking, if we were to keep making coin flips (i.e. so that n grows), then there will be more possible values of r , and the distribution becomes smoother and more concentrated around the middle (near $r = n/2$). We will speak about this "smoothness" in great detail in Lecture 10.



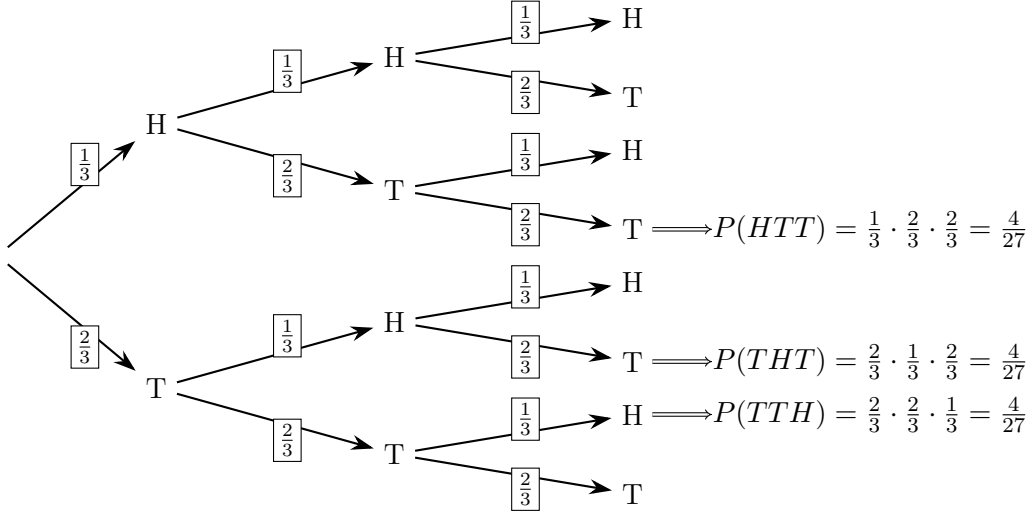
8.2.3 Example 2: flipping an unfair coin

Suppose now that we have a coin that has been tampered with in some way, so that the probabilities of landing on heads or tails are *no longer equally likely*. Let's assume that our "unfair" coin has probabilities:

$$P(H) = \frac{1}{3}, \quad P(T) = \frac{2}{3}.$$

Question: what is the probability that we get exactly 2 tails in 3 flips?

In the fair coin case discussed above, all of the 8 outcomes are equally likely, so we could simply count up all of the favourable outcomes (there were 3) then divide by 8. However, for our unfair coin, it is *not* the case that all outcomes are equally likely. So, we cannot just count favourable outcomes and divide by 8. Instead, we can use the decision tree approach. The space of possible outcomes is the same as before, however, now the probabilities on each branch of the tree have changed:



Exactly 2 tails corresponds to the outcomes HTT , THT , TTH . These are mutually exclusive events, so our desired probability can be obtained by summing:

$$P(\text{exactly 2 tails}) = P(HTT \text{ or } THT \text{ or } TTH) = \frac{4}{27} + \frac{4}{27} + \frac{4}{27} = \frac{12}{27}.$$

Notice that the probability calculation of each of these outcomes is effectively the same: to calculate $P(HTT)$ or $P(THT)$ or $P(TTH)$, we simply multiply two copies of $P(T) = \frac{2}{3}$ together with one copy of $P(H) = \frac{1}{3}$. In fact, this style of calculation will work to calculate any further probabilities. For n -many flips, we have:

$$P(r \text{ many } T\text{'s}) = C_{n,r} \left(\frac{2}{3}\right)^r \left(\frac{1}{3}\right)^{n-r}.$$

Even more generally, if we were to adjust the probabilities of landing on tails or heads to other values, we could calculate the probability of r -many tails in n -many flips using the formula:

$$P(r \text{ many } T\text{'s in } n \text{ flips}) = C_{n,r} (P(T))^r (P(H))^{n-r}.$$

Here $(P(T))^r (P(H))^{n-r}$ is the probability of one specific ordering of r -many T 's and $(n-r)$ -many H 's, and $C_{n,r}$ counts the number of different orderings of r -many T 's and $(n-r)$ -many H 's.

8.3 The general theory

8.3.1 Definition of a binomial experiment

In Section 2 we saw two different versions of coin flips: one fair, and one “unfair” in the sense that the two possible outcomes had different probabilities. As a matter of fact, these are two examples of a more general process called a *binomial experiment*. There are four important defining features of a binomial experiment.

Definition: Binomial Experiment

A binomial experiment has the following properties:

- (B1) There is a fixed number of trials. Call this number n .
- (B2) The n trials are independent and repeated under identical conditions.
- (B3) Each trial has only two outcomes: success (S) and failure (F).
- (B4) On each trial the probability of success is the same. Call it p , and call the probability of failure q . Since each trial is either success or failure, $p + q = 1$, so $q = 1 - p$.

Put differently, a binomial experiment is a fixed number of independent trials, each with two possible outcomes and the same success probability. In a binomial experiment, the central problem is to determine the probability of r successes out of n trials. The assumption (B2) is used to simplify the calculation of these successive probabilities.

Example: fair coin flips

In our first example from the previous section:

- (B1) n was the number of coin flips.
- (B2) The coin flips are independent. In fact, they are not related to the outcomes of previous experiments at all (even physically).
- (B3) Success was defined as “we got a tail” and failure was defined as “we did not get a tail”.
- (B4) For a fair coin, $p = P(T) = 0.5$ and $q = 0.5$.

For the central question of determining probabilities associated with this experiment, this is precisely what our previous discussion was doing. We already saw that

$$P(r \text{ tails out of } n \text{ flips}) = \frac{C_{n,r}}{2^n}$$

for a fair coin.

For the unfair coin, the binomial setup is essentially the same, except that we adjusted the probabilities to $p = \frac{2}{3}$ and $q = \frac{1}{3}$. We saw previously that the associated probabilities for different values of r can be calculated with the formula:

$$P(r \text{ many } T\text{'s}) = C_{n,r} \left(\frac{2}{3}\right)^r \left(\frac{1}{3}\right)^{n-r} = C_{n,r} \cdot P(T)^r \cdot P(H)^{n-r}.$$

General Formula for Binomial Probability

For a binomial experiment with parameters n , p , $q = 1 - p$, the probability of getting exactly r successes is

$$P(r \text{ successes out of } n \text{ trials}) = C_{n,r} p^r q^{n-r}.$$

Here:

- p^r is the success part of the probability for one specific order.
- q^{n-r} is the failure part of the probability for that same specific order.
- $p^r q^{n-r}$ is the probability of one particular sequence that has r successes and $(n - r)$ failures.
- $C_{n,r}$ counts how many different sequences have r successes and $(n - r)$ failures.

8.3.2 What does p do to the distribution?

Generally speaking, we can use the formula

$$P(r \text{ successes out of } n \text{ trials}) = C_{n,r} p^r q^{n-r}.$$

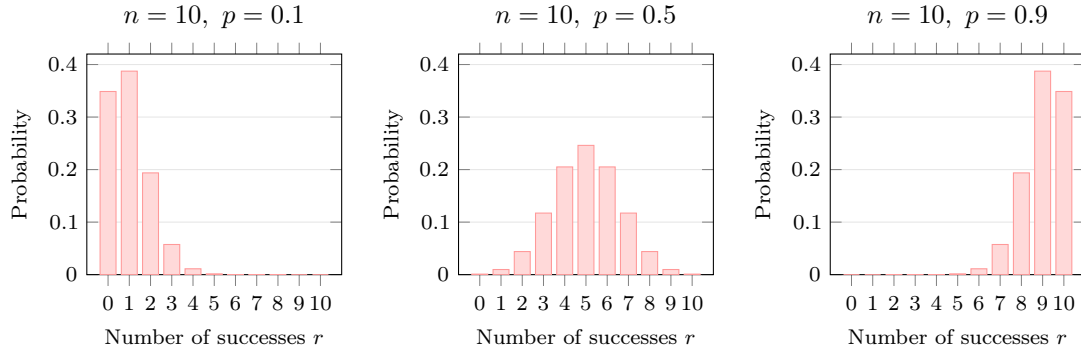
to describe a binomial distribution associated to every binomial experiment. Simply: we take the x -axis to be the possible values of the random variable (from 0 to n), and we take the y -axis to be the output of the formula above.

In the case of a fair coin, the success probability $p = 0.5$, which is exactly equal to q . This causes the resulting binomial distribution to be perfectly symmetric. Notice this in the graphs detailed in Section 2. If we start to adjust the success probability away from $p = 0.5$, then this will cause the distribution to lose its symmetry.

For illustration purposes, let's suppose that we are working with a binomial experiment of $n = 10$ trials. In the case that the probability $p = 0.5$, the resulting distribution is the figure in Section 2.1. We will now briefly discuss what happens if we make p bigger or smaller.

- If p is near 0, say $p = 0.1$, the chance of getting a successful outcome to any trial is much lower. Therefore, the chance of getting a large number of successes, such as 7, 8, 9 or 10 successful outcomes in 10 trials is *extremely low*. Most likely, you will only get 0, 1 or 2 successful outcomes in 10 attempts. As such, the binomial distribution associated to a system like this will actually have the majority of its probability distributed towards the left side of the x -axis, closer to the values of 0, 1 and 2.
- If p is near 1, say $p = 0.9$, then it would now be *extremely likely* to receive a successful trial. Therefore, it would be very unlucky to receive only one or two successful outcomes in 10 attempts, and instead you would expect to get somewhere around 7, 8, 9 or 10 successes. On the level of the binomial distribution, having such a high chance of success will push the distribution to the *right* and cause most of the probability to be centred around the 8, 9, 10 area of the x -axis.

The graphs below depict the actual binomial distributions for $n = 10$ trials with success probabilities of 0.1, 0.5 and 0.9, respectively.



8.3.3 Expected value and variance for the binomial distribution

For a binomial distribution, there are two especially important formulas:

Binomial expected value and variance

If X counts the number of successes in a binomial experiment with parameters n , p and $q = 1 - p$, then

$$E(X) = np, \quad \text{Var}(X) = npq.$$

These formulas are useful because they let us quickly understand the centre and spread of a binomial distribution without calculating the entire distribution by hand.

Technically, we could verify the above two formulas using some algebra. However, we will not do this and will instead argue loosely: imagine the n trials as n slots:



Each slot contributes either 1 success or 0 successes. Since the probability of success in each slot is p , according to our Frequentist interpretation of probability, each slot contributes a long-term expected average of p successes. Since there are n slots, the total expected number of successes is

$$p + p + \cdots + p = np,$$

which recovers the formula $E(X) = np$. As for the variance formula, we will not derive this here. However, we will note that the formula

$$\text{Var}(X) = npq$$

tells us how volatile the number of successes is:

- If p is close to 0 or close to 1, then the experiment is less volatile and more tightly packed around $E(X)$, because the outcome is more predictable.
- If p is near 0.5, then the experiment is more volatile, because success and failure are both much more plausible on each trial.

8.4 Exercises

8.4.1 The roulette wheel

Suppose that there are 20 students in our class. You study hard and after a few years you successfully graduate from LUJ. Dr. O’Connell is so proud of you that he decides to take you all to the casino in order to play on the roulette wheel. Each student spins the wheel once.



There are 36 numbered slots on the roulette wheel, and the green slot makes 37 total outcomes. Dr. O’Connell gives you the following offer:

You get \$50,000 if the ball lands on the green slot, and \$0 if any other outcome happens.

The major question we will answer is:

Exercise

What is the probability that exactly 3 students will win?

To solve this question, we will use the general binomial formula mentioned in Section 3.2. However, before we start to calculate anything, we must first confirm that this situation is actually a binomial experiment to begin with.

- B1: There are 20 students and they each try the roulette wheel once, so $n = 20$ trials.
- B2: Assuming that each spin is independent and the roulette wheel is fair, the trials are indeed independent from each other.
- B3: There are only two possible outcomes to each trial:

Success = lands on green, Failure = lands elsewhere.

- B4: each slot of the roulette wheel has an equally likely chance of being the outcome. Since there are 37 slots on the wheel and only one green slot to land on, the probabilities of success and failure are:

$$p = P(\text{success}) = \frac{1}{37}, \quad q = P(\text{failure}) = 1 - p = \frac{36}{37}.$$

Now that we have confirmed our system can be seen as a binomial experiment, we can solve the main question.

Solution

We use the binomial formula:

$$P(r \text{ successes out of } n \text{ trials}) = C_{n,r} p^r q^{n-r}.$$

In this case, $n = 20$ since there are 20 students, $r = 3$, $p = \frac{1}{37}$ and $q = \frac{36}{37}$. We substitute all of these values into the formula and see:

$$P(\text{exactly 3 wins}) = C_{20,3} \left(\frac{1}{37}\right)^3 \left(\frac{36}{37}\right)^{17} = \frac{20!}{3! \cdot 17!} \left(\frac{1}{37}\right)^3 \left(\frac{36}{37}\right)^{17}.$$

Numerically, this is approximately

$$P(\text{exactly 3 wins}) \approx 0.014 \approx 1.4\%.$$

As an alternate explanation of this probability: we are performing 20 independent trials and we want 3 successes and 17 failures. One particular outcome that represents this could be:

$$S, S, S, F, F, \dots, F,$$

i.e. the first 3 students are successful and then the 17 students after that all fail. According to the assumption (B2), all of these trials are independent from each other. So, the probability associated with this outcome can be calculated by simply multiplying the probability of each outcome. We have:

$$P(S \text{ then } S \text{ then } S \text{ then } F \text{ then } \dots \text{ then } F) = \frac{1}{37} \cdot \frac{1}{37} \cdot \frac{1}{37} \cdot \frac{36}{37} \cdot \dots \cdot \frac{36}{37},$$

which is the same as writing $p^3 q^{17}$. This is the formula to calculate a single outcome that corresponds to 3 successes and 17 failures. But, there are $C_{20,3}$ -many different ways to place the 3 successes among 20 trials. These sequences correspond to mutually exclusive outcomes, so we add their probabilities: producing $C_{20,3} p^3 q^{17}$.

$$P(\text{exactly 3 wins}) = \underbrace{p^3 q^{17} + p^3 q^{17} + \dots + p^3 q^{17}}_{C_{20,3}\text{-many times}} = C_{20,3} \cdot p^3 q^{17},$$

which is the same formula as before.

8.4.2 Loot boxes

A game has a rare loot item with a 0.26% chance of appearing on each play. Each play costs \$2.49. We will treat “success” as getting the loot.

Questions

- (a) What is the probability of getting exactly 2 wins in 100 attempts?
- (b) What is the probability of getting exactly 3 wins in 1000 attempts?
- (c) How much money should you expect to pay before you find the loot?

Solution

The probability of success is

$$p = 0.0026, \quad q = 1 - p = 0.9974.$$

(a) Here $n = 100$ and $r = 2$. So:

$$P(X = 2) = C_{100,2}(0.0026)^2(0.9974)^{98}.$$

Numerically, this is approximately

$$P(X = 2) \approx 0.0259 = 2.59\%.$$

(b) Here $n = 1000$ and $r = 3$. So:

$$P(X = 3) = C_{1000,3}(0.0026)^3(0.9974)^{997}.$$

Numerically, this is approximately

$$P(X = 3) \approx 0.218 = 21.8\%.$$

(c) We are looking for the number of attempts n so that the expected number of wins reaches 1. Since $E(X) = np$, we solve:

$$1 = n(0.0026) \quad \implies \quad n = \frac{1}{0.0026} \approx 384.6.$$

So, on average, we would expect to need about 385 plays for one loot item. Since each play costs \$2.49, this would cost

$$385 \times 2.49 = \$958.65.$$

8.4.3 Kaitenzushi

A kaitenzushi restaurant allows you to play a game and win a prize after you insert 5 plates into the machine. The probability of winning the prize is 25%.

Questions

- (a) What is the probability of playing 5 times and winning twice?
- (b) What is the probability of winning 2 times in 2 attempts?
- (c) On average, how many plates should you expect to eat before winning a prize?

Solution

Success is defined as winning the prize. Then

$$p = 0.25, \quad q = 0.75.$$

(a) Here $n = 5$ and $r = 2$, so:

$$P(X = 2) = C_{5,2}(0.25)^2(0.75)^3.$$

Numerically,

$$P(X = 2) = 10(0.0625)(0.421875) = 0.263671875 \approx 26.4\%.$$

(b) Winning 2 times in 2 attempts means winning both attempts. Since the attempts are independent:

$$P(\text{win then win}) = 0.25 \times 0.25 = 0.0625 = 6.25\%.$$

(c)

(d) Since $E(X) = np$, we want $E(X) = 1$, so:

$$1 = n(0.25) \quad \implies \quad n = 4.$$

So we expect about 4 attempts before winning one prize. Since each attempt requires 5 plates, this corresponds to

$$4 \times 5 = 20$$

plates on average.

8.4.4 Rolling a dice

Imagine you roll a standard 6-sided die three times.

We will treat “success” as rolling a 6. Then each roll has only two outcomes (6 or not 6), the rolls are independent, and the success probability is constant:

$$n = 3, \quad p = P(\text{roll a 6}) = \frac{1}{6}, \quad q = \frac{5}{6}.$$

Question

What is the probability of getting exactly two 6's?

Solution

To get exactly two 6's, we need $r = 2$ successes out of $n = 3$ trials. The binomial formula gives

$$P(r \text{ successes}) = C_{n,r} p^r q^{n-r}.$$

So

$$P(\text{exactly two 6's}) = C_{3,2} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right) = 3 \cdot \frac{1}{36} \cdot \frac{5}{6} = \frac{15}{216} = \frac{5}{72} \approx 6.9\%.$$

Narrative interpretation: there are $C_{3,2} = 3$ ways to choose which two of the three rolls are the 6's, and each such pattern has probability $\left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)$.

8.4.5 The basketball cat

A cat is shooting basketball free throws. Cats are not very good at basketball, so suppose the probability a shot goes in is $p = 0.5$. The cat takes 6 shots.

We treat “success” as “the shot goes in”. Then:

$$n = 6, \quad p = 0.5, \quad q = 0.5.$$

Questions

- (a) What is the probability that the cat makes exactly 4 shots?
- (b) What is the probability that the cat makes at least 4 shots?

Solution

Let X be the number of made shots.

- (a) Exactly 4 makes means $r = 4$ successes out of $n = 6$ trials:

$$P(X = 4) = C_{6,4}(0.5)^4(0.5)^2 = C_{6,4}(0.5)^6 = \frac{C_{6,4}}{64} = \frac{15}{64} \approx 23.4\%.$$

- (b) “At least 4” means 4 or 5 or 6 makes, so we add those mutually exclusive cases:

$$\begin{aligned} P(X \geq 4) &= P(X = 4) + P(X = 5) + P(X = 6) \\ &= \frac{C_{6,4}}{64} + \frac{C_{6,5}}{64} + \frac{C_{6,6}}{64} = \frac{15 + 6 + 1}{64} = \frac{22}{64} = \frac{11}{32} \approx 34.4\%. \end{aligned}$$

Part III

The Normal Distribution

Lecture 10

The Normal Distribution

In earlier lectures, we studied descriptive statistics, probability theory, conditional probability and binomial distributions. In this lecture we study the most famous continuous probability distribution: the normal distribution. We will learn what the normal curve is, what the standard normal distribution is, and how z -scores let us move between different normal distributions. We will also learn why probability for continuous variables is measured using *area under the curve* rather than bar heights. Next lecture we will continue our study of the normal distribution.

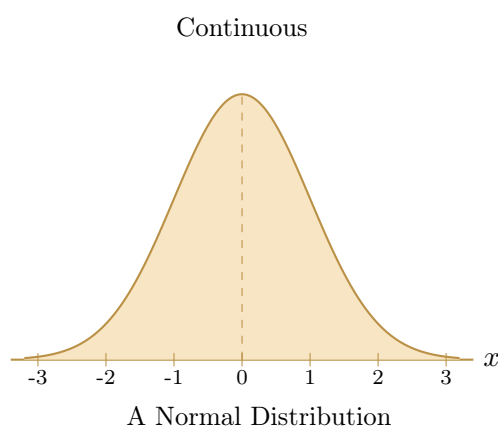
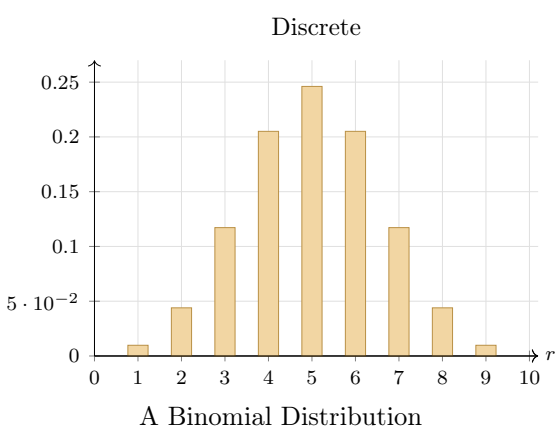
10.1 What is a Normal Distribution?

10.1.1 Discrete and continuous random variables

Recall that a *random variable* is a variable whose value is determined by chance. As we saw in Lecture 6, there are two main types of random variable:

- **Discrete random variables**, which can take separate values such as $0, 1, 2, 3, \dots$, and
- **Continuous random variables**, which can take any real value in an interval.

In Lecture 8 we studied a probability distribution built from a discrete random variable, namely the binomial distribution. Today we move to a probability distribution built from a continuous random variable instead.



Continuous probability distributions have more structure than the discrete distributions we have seen so far. That means there is more to talk about, and therefore they are potentially more confusing. So, our strategy will be to learn the continuous case by using *limits* to relate them to what we already know.

10.1.2 A metaphor for limits

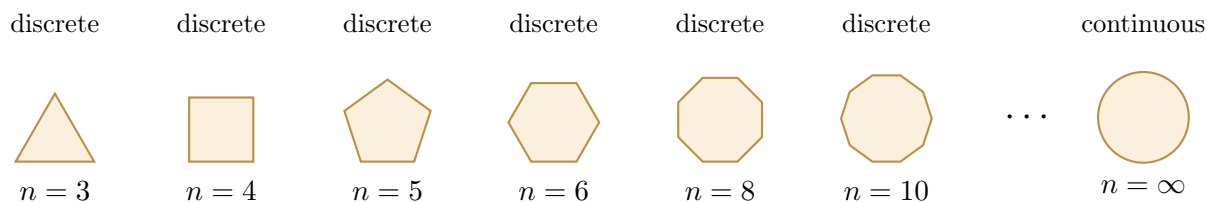
Suppose that we have a bunch of logs whose cross-sections have different shapes:



In the case of the first log, having a triangular cross-section will make it very awkward to roll. If we were to move on to the next log, which has a square cross-section, then we would notice that it is

slightly easier. After this, we would notice that the log with a pentagonal cross-section would be even easier, and again the log with a hexagonal cross-section would be easier still. We notice that as the number of sides of the cross-section increase, the log gets easier and easier to roll. Of course, a perfectly circular log would be the easiest.

If we were to have many, many logs and kept increasing the number of sides, we would get a sequence of logs that may look something like this:

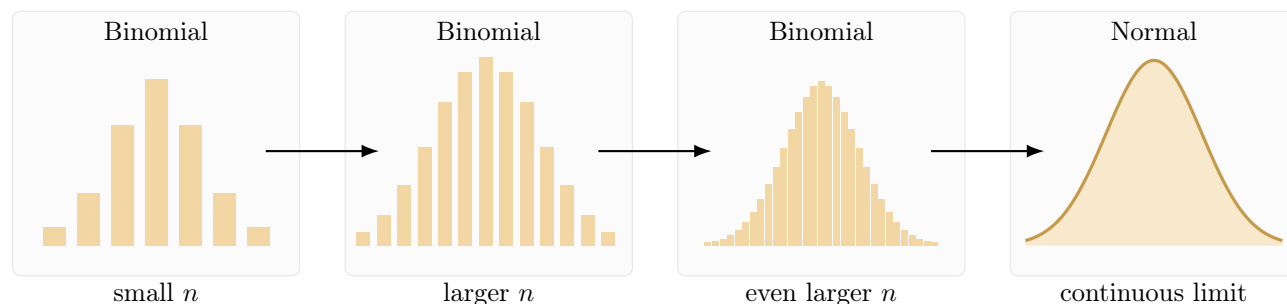


The limit of this procedure would be a circle, which we can think of as a shape with infinitely-many sides. If we interpret the number of sides as a discrete quantity, we see that in the process of adding more sides to our shapes, the polygons approach a continuous shape in the limit.

10.1.3 The limiting process

Our analogy of rolling logs has a deeper philosophy that may be applied to random variables: we can “approach” a normal distribution by considering a sequence of binomial distributions in which we increase the number of trials, and hence the number of bars. Put differently: a continuous curve can be described as a limit of many discrete bars.

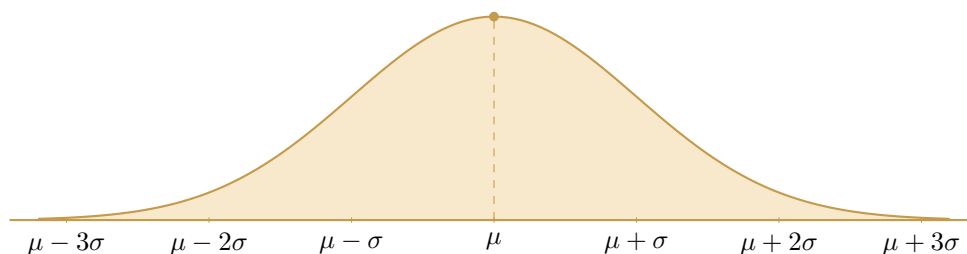
Recall that in a binomial distribution, the total number of bars in the graph is given by $n + 1$, since the possible values run from 0 to n . As n increases, the bars in the binomial distribution become thinner and more numerous. The overall distribution appears smoother and smoother, and eventually begins to resemble a curve:



This limiting bell-shaped curve is called a *normal distribution*. Now, it’s worth mentioning that this limiting procedure is a well-defined technical result in mathematics known as the de Moivre-Laplace Theorem. There are important technical features, but we will explore these in next lecture.

10.1.4 The normal distribution and the normal curve

The graph of the normal distribution is sometimes called a *normal curve*, or simply a *bell curve*, because of its shape.



The normal distribution is one of the most important objects in statistics. In practice, we see the normal distribution in human characteristics, physical measurements, industrial data, measurement error, and many other contexts.

A normal distribution has several important features.

Important properties of a normal distribution

1. The curve is bell-shaped, with its highest point above the mean μ .
2. The curve is symmetric about the vertical line through μ .
3. The tails approach the horizontal axis but never touch or cross it.
4. The total area under the entire curve is 1.

10.1.5 Mathematical description

Since a normal distribution is represented by a graph, we can ask what the associated function is. This function has a special name: it is called the *normal density function*. Although we will never need to work with this function directly, it is written below:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Despite the obvious complexity of this function, we should make a few important observations. First, notice that the function has a single variable x . This is the continuous random variable that becomes the horizontal axis of the graph. Next, we see that apart from x , the function $f(x)$ needs two important numbers as parameters:

- μ is the mean of the distribution.
- σ is the standard deviation of the distribution.

As you can see from the function above, the numbers μ and σ are the only pieces of extra information in the function (apart from constants such as π , of course). Therefore every normal distribution is

completely *characterized* by the numbers μ and σ . This means that different choices of μ and σ give different normal curves. For this reason, we often write a normal distribution as:

$$X \sim \mathcal{N}(\mu, \sigma^2).$$

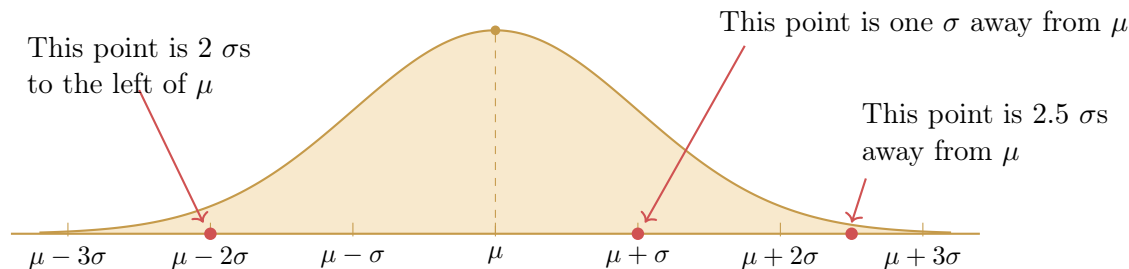
In words, this should be read “the continuous random variable X follows a normal distribution with mean μ and variance σ^2 .”

This is quite a special feature of the normal distribution. The population mean μ tells us where the centre of the distribution is, and the population standard deviation σ is a measure of how spread out the data is. Usually, these are just two numbers that help describe a dataset. However, in the case of the normal distribution, the numbers μ and σ *completely describe* the distribution. So, the symbols μ and σ are all the information we need in order to navigate the curve.

10.2 The Standard Normal Distribution

10.2.1 The z -score

According to our discussion so far, we see that any normal distribution $\mathcal{N}(\mu, \sigma^2)$ has an x -axis that can be described in terms of μ and σ . If we randomly select a value of the random variable X , then this will land somewhere on the x -axis. Since the x -axis is centred at μ and measured in units of σ , we can express where our variable is in terms of distance from the centre. For example:



The z -score, also called the *standard score*, turns this information into a pure number, by telling us how many standard deviations a measurement lies from the mean.

Definition of the z -score

If x is a value from a normal distribution with mean μ and standard deviation σ , then its z -score is

$$z = \frac{x - \mu}{\sigma}.$$

This formula can be interpreted one piece at a time:

- $x - \mu$ measures the distance x is from the mean;
- dividing by σ converts that distance into a count of standard deviations;
- the result may be positive or negative, and this corresponds to the right or left of the mean, respectively.

For example, in the diagram above:

- the point that is at $\mu - 2\sigma$ would have a z -score of -2 ,
- the point that is at $\mu + \sigma$ would have a z -score of 1 , and
- the point that is at $\mu + 2.5\sigma$ would have a z -score of 2.5 .

Sometimes we know the z -score and want to recover the original value x . Starting with the formula

$$z = \frac{x - \mu}{\sigma},$$

we can rearrange to get:

$$x = \mu + z\sigma.$$

The z -score and the raw score contain the same information, just written in two different ways.

Exercise: z -scores and raw scores

Let

$$X \sim \mathcal{N}(10, 2^2),$$

so that $\mu = 10$ and $\sigma = 2$.

- (a) How far away is the value $x = 12$ in terms of σ ?
- (b) How far away is the value $x = 4$ in terms of σ ?
- (c) What value of x is $+2\sigma$ away from the mean?

Solution

(a)

$$z = \frac{x - \mu}{\sigma} = \frac{12 - 10}{2} = \frac{2}{2} = 1.$$

So $x = 12$ is 1 standard deviation above the mean.

(b)

$$z = \frac{x - \mu}{\sigma} = \frac{4 - 10}{2} = \frac{-6}{2} = -3.$$

So $x = 4$ is 3 standard deviations below the mean.

(c) If x is $+2\sigma$ away from the mean, then $z = 2$. Hence

$$2 = \frac{x - 10}{2}.$$

Rearranging gives

$$x = 2 \cdot 2 + 10 = 14.$$

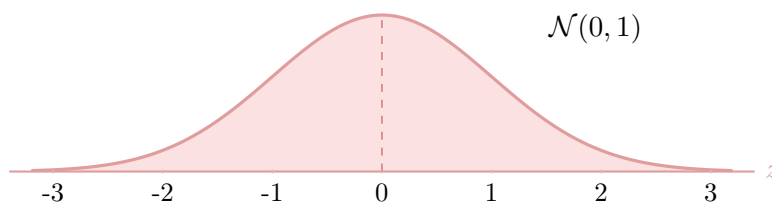
10.2.2 The standard normal distribution

The *standard normal distribution* is the normal distribution with mean 0 and standard deviation 1. We usually write this as

$$Z \sim \mathcal{N}(0, 1).$$

Note that we use the special letter Z for a standard normal random variable, instead of the letter X that we wrote before. As a convention, we will draw the standard normal distribution $\mathcal{N}(0, 1)$ in

red, so that we can visually distinguish it from general normal distributions. The standard normal distribution is depicted below.



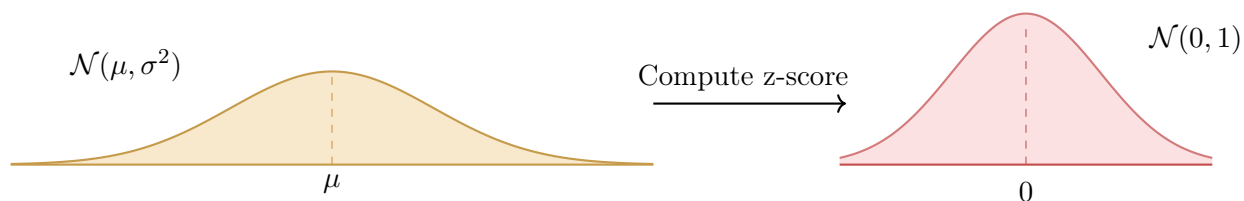
Observe the horizontal axis here: we are centred at 0 and we count up or down in increments of 1.

10.2.3 Why the standard normal is useful

Any normal distribution can be converted into the standard normal distribution by converting its measurements to z -scores. That is why the standard normal distribution is so important: instead of learning a different area table for every possible normal curve, we use one universal reference curve.

The big idea

Every normal distribution can be translated into the standard normal world. Once that has been done, probabilities can be read from a single table of standard normal areas.



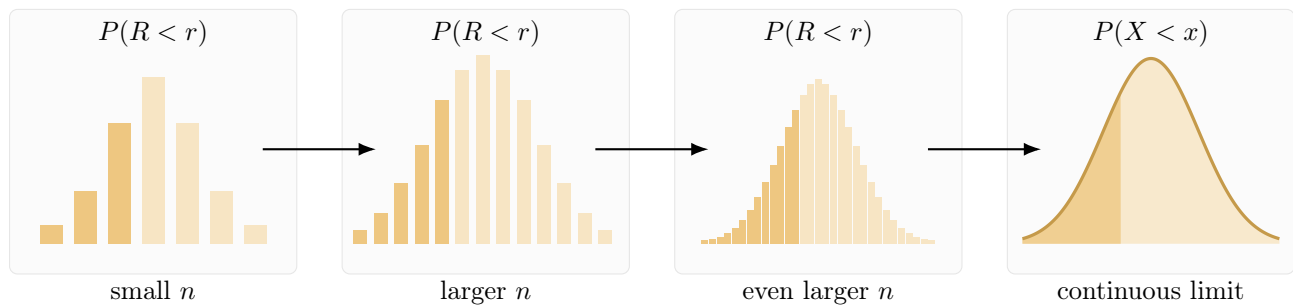
10.3 Areas and Probability

10.3.1 From bar heights to area

When working with a binomial distribution, the height of each bar gives the exact probability that the discrete random variable will take on that value. For example, in the binomial distribution for $n = 10$ and $p = 0.5$:

- the probability of exactly $r = 5$ is the height of the bar at $r = 5$, which is about 0.24;
- the probability of $r < 5$ is obtained by adding the heights of the bars at $r = 0, 1, 2, 3, 4$, which gives about 0.38.

For larger and larger values of n , this process of adding many thin bars begins to look more and more like a continuous area:



As the above picture demonstrates, for continuous random variables, we think of probabilities in terms of *areas under the curve*, rather than (a sum of) the heights of bars. This is the main conceptual shift in today's lecture.

Discrete versus continuous probability

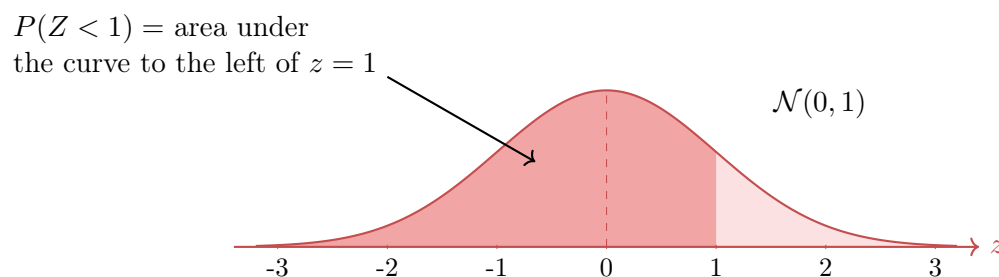
- For a **discrete** random variable, the probability $P(X < x)$ is found by adding the probabilities of separate outcomes.
- For a **continuous** random variable, the probability $P(X < x)$ is found by measuring area under a curve.

10.3.2 How areas are found

Consider again the standard normal distribution $\mathcal{N}(0, 1)$. According to the discussion above, when we write

$$P(Z < z),$$

we mean the area under the standard normal curve to the left of the point z . For example, the probability $P(Z < 1)$ would be given by the shaded area:



If you know some calculus, then these areas can be found using integration. Technically, we need to integrate the function defining $\mathcal{N}(0, 1)$ between two endpoints:

$$P(Z < z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt.$$

However, this is often difficult to compute by hand, so in practice we often use a *z-table* instead. We can think of a *z-table* as a list of all of the answers for common regions of the standard normal distribution. A simplified version of a *z-table* is written below.

z	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0
$P(Z < z)$	0.02	0.07	0.16	0.31	0.50	0.69	0.84	0.93	0.98

Here, the second row gives the area to the left of z for particular values of z . For example, in the diagram above, the table tells us that the shaded area is 0.84, meaning that 84% of the total area under the curve lies to the left of $z = 1$. In probabilistic terms, this says that our random variable Z has about an 84% chance of taking a value less than 1.

Of course, the above table only works for the 9 particular values of z ranging from -2.0 up to 2.0 . Throughout our course we will also use the left-tailed z -scores table included in the Appendix below. This is a far more complicated table, and it provides the area corresponding to $Z < z$ for values of z up to 2 decimal places. In the absence of this rather large table, some calculators can also obtain accurate values of $P(Z < z)$ using specific functions.

10.3.3 Understanding areas

A z -table lists areas under the standard normal curve for the standard normal random variable Z . We interpret the table as giving an approximate answer for the probability

$$P(Z < z),$$

that is, the area to the *left* of z . Using the basic properties of the normal distribution, we can use this table to describe other areas as well.

How to find other areas

- **Areas to the left of z :** read off $P(Z < z)$ directly.
- **Areas to the right of z :** use $P(Z > z) = 1 - P(Z < z)$.
- **Areas between two values:** use $P(a < Z < b) = P(Z < b) - P(Z < a)$.
- **Negative values:** by symmetry, for $a > 0$,

$$P(Z < -a) = 1 - P(Z < a).$$

For a continuous random variable, the probability of any *single exact point* is zero. So, for example,

$$P(Z = 0.5) = 0.$$

We stress that this does not mean that the value of 0.5 is impossible. Instead, it simply means that it has *probability* equal to zero, since the one dimensional line under the graph corresponding to $X = 0.5$ has zero area. Moreover, non-zero probabilities will only occur when we consider *intervals* such as $0.45 < X < 0.55$, because these occupy some non-zero area under the curve.

10.3.4 Computing probabilities for other normal distributions

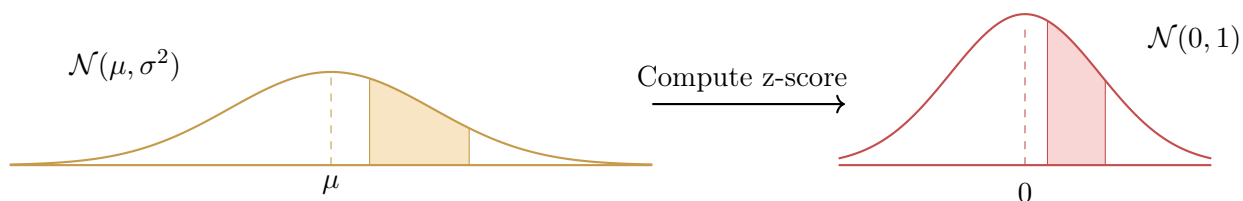
In Section 2 we mentioned that we can use the z -score formula

$$z = \frac{x - \mu}{\sigma}$$

to convert any normal distribution $\mathcal{N}(\mu, \sigma^2)$ into the standard normal distribution $\mathcal{N}(0, 1)$. What is useful here is that this transformation process also preserves the proportions of areas, meaning that we can determine probabilities in $\mathcal{N}(\mu, \sigma^2)$ by using z -scores and the z -table above. For a random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, we have:

$$P(X < x) = P\left(Z < \frac{x - \mu}{\sigma}\right).$$

In other words, to determine the value of $P(X < x)$, we simply compute the z -score of the point x , and then look at the corresponding area in the z -table above. This notion also works for all other areas. For example, a probability like $P(x_1 < X < x_2)$ can be equally determined using z -scores for x_1 and x_2 :



This is the true strength of z -scores: they allow us to handle *any* normal distribution using the standard normal table.

Worked example

Let $X \sim \mathcal{N}(10, 2^2)$. Find the probability that a randomly selected value lies between 11 and 14. In symbols, find $P(11 \leq X \leq 14)$.

Solution

Step 1: Convert the x -values to z -values.

$$z_1 = \frac{11 - 10}{2} = 0.5, \quad z_2 = \frac{14 - 10}{2} = 2.0.$$

Step 2: Read the left-tail areas from the table.

$$P(Z < 0.5) = 0.69, \quad P(Z < 2.0) = 0.98.$$

Step 3: Subtract to get the interval probability.

$$P(11 \leq X \leq 14) = P(Z < 2.0) - P(Z < 0.5) = 0.98 - 0.69 = 0.29.$$

So the probability is

$$0.29.$$

In words, there is about a 29% chance that a randomly selected value from this normal model lies between 11 and 14.

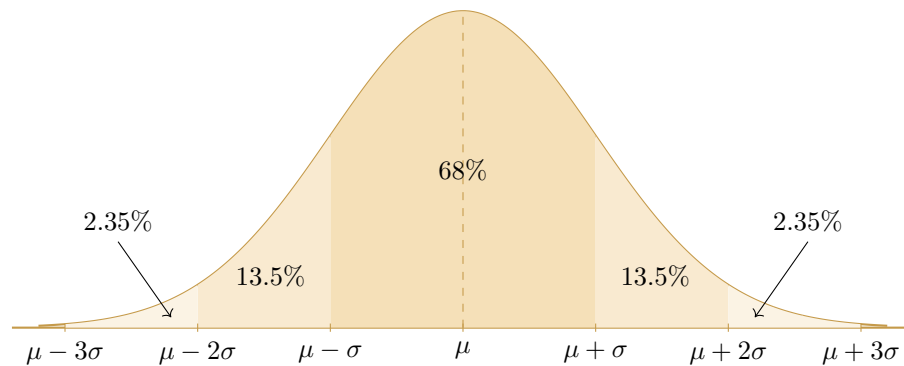
10.3.5 The 68-95-99.7 rule

For a normal distribution, the standard deviation gives a very useful geometric guide to the data.

Empirical rule

For a normal distribution:

- about 68% of the data lie within 1 standard deviation of the mean;
- about 95% of the data lie within 2 standard deviations of the mean;
- about 99.7% of the data lie within 3 standard deviations of the mean.



The remaining 0.3% lies in the two extreme tails, about 0.15% in each tail.

10.3.6 Worked exercise: areas under the standard normal curve

Use the mini z -table above, together with symmetry, to find the following areas.

Exercise

- Find the area to the left of $z = -1.00$.
- Find the area to the left of $z = 1.5$.
- Without using the z -table, find the area to the left of $z = 0$.
- Without using the z -table, find the area underneath the entire curve.
- Find the area to the right of $z = 1$.
- Find the area of the interval $-1 < Z < 1.5$.
- Find $P(Z = 0.5)$.

Solution

(a) From the table,

$$P(Z < -1.00) = 0.16.$$

(b) From the table,

$$P(Z < 1.5) = 0.93.$$

(c) By symmetry, half of the total area lies to the left of 0, so

$$P(Z < 0) = 0.50.$$

(d) The total area under the curve is always 1.

(e)

$$P(Z > 1) = 1 - P(Z < 1) = 1 - 0.84 = 0.16.$$

(f)

$$P(-1 < Z < 1.5) = P(Z < 1.5) - P(Z < -1).$$

Using the table,

$$P(-1 < Z < 1.5) = 0.93 - 0.16 = 0.77.$$

(g) Since a single exact value has zero area,

$$P(Z = 0.5) = 0.$$

10.3.7 Exercise: Japanese male heights

The height of Japanese males can be modelled using a normal distribution with mean $\mu = 171$ cm and standard deviation $\sigma = 5.6$ cm. Suppose that a Japanese male is selected at random.

Exercise

Let

$$X \sim \mathcal{N}(171, 5.6^2).$$

(a) What is the probability that the selected Japanese male is taller than 176.6 cm?

(b) What is the probability that the selected Japanese male is exactly 176.6 cm tall?

Solution

(a) First we compute the z -score:

$$z = \frac{x - \mu}{\sigma} = \frac{176.6 - 171}{5.6} = \frac{5.6}{5.6} = 1.$$

Therefore,

$$P(X > 176.6) = P(Z > 1).$$

Using the z -scores table we see that $P(Z < 1) = 0.84$, so:

$$P(Z > 1) = 1 - P(Z < 1) = 1 - 0.84 = 0.16.$$

Thus,

$$P(X > 176.6) = 0.16.$$

So there is about a 16% chance that the selected Japanese male is taller than 176.6 cm.

(b) Since X is a continuous random variable, the probability of one exact value is zero:

$$P(X = 176.6) = 0.$$

10.4 Appendix: z -scores table

STANDARD NORMAL DISTRIBUTION: $P(Z < z)$ for Negative z from -3.99 to 0.0

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.9	0.00005	0.00005	0.00004	0.00004	0.00004	0.00004	0.00004	0.00004	0.00003	0.00003
-3.8	0.00007	0.00007	0.00007	0.00006	0.00006	0.00006	0.00006	0.00005	0.00005	0.00005
-3.7	0.00011	0.00010	0.00010	0.00010	0.00009	0.00009	0.00008	0.00008	0.00008	0.00008
-3.6	0.00016	0.00015	0.00015	0.00014	0.00014	0.00013	0.00013	0.00012	0.00012	0.00011
-3.5	0.00023	0.00022	0.00022	0.00021	0.00020	0.00019	0.00019	0.00018	0.00017	0.00017
-3.4	0.00034	0.00032	0.00031	0.00030	0.00029	0.00028	0.00027	0.00026	0.00025	0.00024
-3.3	0.00048	0.00047	0.00045	0.00043	0.00042	0.00040	0.00039	0.00038	0.00036	0.00035
-3.2	0.00069	0.00066	0.00064	0.00062	0.00060	0.00058	0.00056	0.00054	0.00052	0.00050
-3.1	0.00097	0.00094	0.00090	0.00087	0.00084	0.00082	0.00079	0.00076	0.00074	0.00071
-3.0	0.00135	0.00131	0.00126	0.00122	0.00118	0.00114	0.00111	0.00107	0.00104	0.00100
-2.9	0.00187	0.00181	0.00175	0.00169	0.00164	0.00159	0.00154	0.00149	0.00144	0.00139
-2.8	0.00256	0.00248	0.00240	0.00233	0.00226	0.00219	0.00212	0.00205	0.00199	0.00193
-2.7	0.00347	0.00336	0.00326	0.00317	0.00307	0.00298	0.00289	0.00280	0.00272	0.00264
-2.6	0.00466	0.00453	0.00440	0.00427	0.00415	0.00402	0.00391	0.00379	0.00368	0.00357
-2.5	0.00621	0.00604	0.00587	0.00570	0.00554	0.00539	0.00523	0.00508	0.00494	0.00480
-2.4	0.00820	0.00798	0.00776	0.00755	0.00734	0.00714	0.00695	0.00676	0.00657	0.00639
-2.3	0.01072	0.01044	0.01017	0.00990	0.00964	0.00939	0.00914	0.00889	0.00866	0.00842
-2.2	0.01390	0.01355	0.01321	0.01287	0.01255	0.01222	0.01191	0.01160	0.01130	0.01101
-2.1	0.01786	0.01743	0.01700	0.01659	0.01618	0.01578	0.01539	0.01500	0.01463	0.01426
-2.0	0.02275	0.02222	0.02169	0.02118	0.02068	0.02018	0.01970	0.01923	0.01876	0.01831
-1.9	0.02872	0.02807	0.02743	0.02680	0.02619	0.02559	0.02500	0.02442	0.02385	0.02330
-1.8	0.03593	0.03515	0.03438	0.03362	0.03288	0.03216	0.03144	0.03074	0.03005	0.02938
-1.7	0.04457	0.04363	0.04272	0.04182	0.04093	0.04006	0.03920	0.03836	0.03754	0.03673
-1.6	0.05480	0.05370	0.05262	0.05155	0.05050	0.04947	0.04846	0.04746	0.04648	0.04551
-1.5	0.06681	0.06552	0.06426	0.06301	0.06178	0.06057	0.05938	0.05821	0.05705	0.05592
-1.4	0.08076	0.07927	0.07780	0.07636	0.07493	0.07353	0.07215	0.07078	0.06944	0.06811
-1.3	0.09680	0.09510	0.09342	0.09176	0.09012	0.08851	0.08691	0.08534	0.08379	0.08226
-1.2	0.11507	0.11314	0.11123	0.10935	0.10749	0.10565	0.10383	0.10204	0.10027	0.09853
-1.1	0.13567	0.13350	0.13136	0.12924	0.12714	0.12507	0.12302	0.12100	0.11900	0.11702
-1.0	0.15866	0.15625	0.15386	0.15151	0.14917	0.14686	0.14457	0.14231	0.14007	0.13786
-0.9	0.18406	0.18141	0.17879	0.17619	0.17361	0.17106	0.16853	0.16602	0.16354	0.16109
-0.8	0.21186	0.20897	0.20611	0.20327	0.20045	0.19766	0.19489	0.19215	0.18943	0.18673
-0.7	0.24196	0.23885	0.23576	0.23270	0.22965	0.22663	0.22363	0.22065	0.21770	0.21476
-0.6	0.27425	0.27093	0.26763	0.26435	0.26109	0.25785	0.25463	0.25143	0.24825	0.24510
-0.5	0.30854	0.30503	0.30153	0.29806	0.29460	0.29116	0.28774	0.28434	0.28096	0.27760
-0.4	0.34458	0.34090	0.33724	0.33360	0.32997	0.32636	0.32276	0.31918	0.31561	0.31207
-0.3	0.38209	0.37828	0.37448	0.37070	0.36693	0.36317	0.35942	0.35569	0.35197	0.34827
-0.2	0.42074	0.41683	0.41294	0.40905	0.40517	0.40129	0.39743	0.39358	0.38974	0.38591
-0.1	0.46017	0.45620	0.45224	0.44828	0.44433	0.44038	0.43644	0.43251	0.42858	0.42465
0.0	0.50000	0.49601	0.49202	0.48803	0.48405	0.48006	0.47608	0.47210	0.46812	0.46414

STANDARD NORMAL DISTRIBUTION: $P(Z < z)$ for Positive z from 0.0 to 3.99

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
3.1	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
3.5	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983
3.6	0.99984	0.99985	0.99985	0.99986	0.99986	0.99987	0.99987	0.99988	0.99988	0.99989
3.7	0.99989	0.99990	0.99990	0.99990	0.99991	0.99991	0.99992	0.99992	0.99992	0.99992
3.8	0.99993	0.99993	0.99993	0.99994	0.99994	0.99994	0.99994	0.99995	0.99995	0.99995
3.9	0.99995	0.99995	0.99996	0.99996	0.99996	0.99996	0.99996	0.99996	0.99997	0.99997

Lecture 11

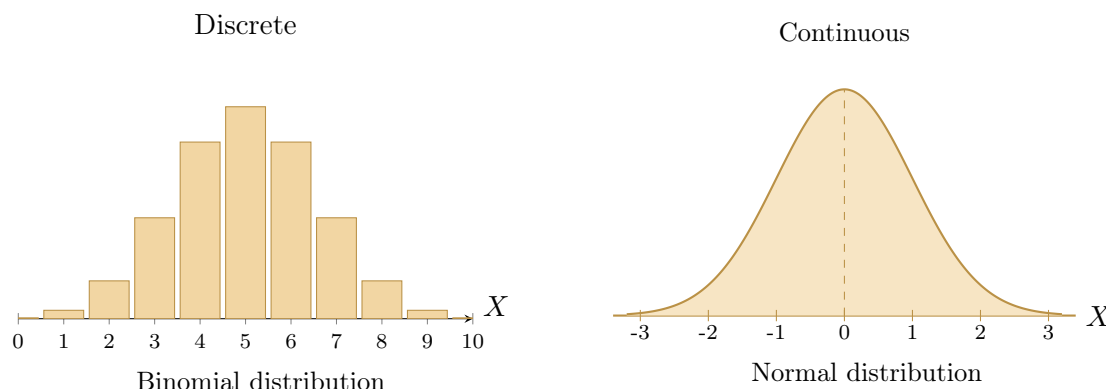
The Central Limit Theorem

In Lecture 10, we studied the normal distribution and the use of z -scores. In this lecture, we explain why the normal distribution appears so often in statistics. The key idea is the so-called *Central Limit Theorem*, which gives a method for associating a normal distribution to sample means from many systems, even if the original system is not normally distributed to begin with. In fact, this result is so important that we will dedicate this entire lecture to it. First we will take some care to properly state the result. Then, we will use it to demonstrate how binomial distributions may be approximated with normal curves.

11.1 A Brief Review of the Normal Distribution

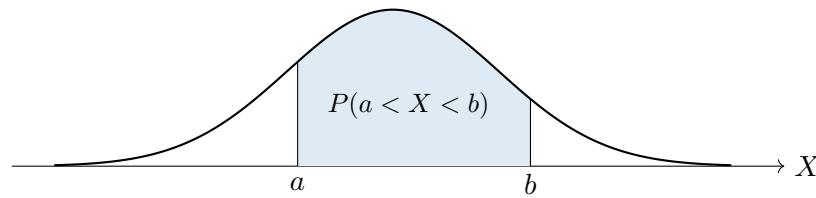
11.1.1 Discrete and continuous random variables

Last lecture we introduced *normal distributions*, which were probability distributions built from continuous random variables, instead of discrete ones. This looks very different from the binomial distributions that we have seen previously:



Here, the fundamental difference is in the possible values on the x -axis. In the case of a binomial distribution, the random variable counts the number of successes. This gives discrete values to the random variable, and then there are probabilities associated to each individual possible value that X can take. In contrast, for a normal distribution the random variable X can take *any real value*.

Instead of taking heights of bars, questions about probabilities are then phrased in terms of *areas*: the probability that X takes a value in some region is given by the size of the area under the curve, as pictured below.



Here, the probability of finding our random variable X within the region between a and b is given by calculating the size of the blue shaded region.

11.1.2 Describing a normal distribution

Generally speaking, the mean and the standard deviation represent two different ways to describe certain features of a data set. In the case of the normal distribution, these two numbers are the only pieces of information that we need to define the entire distribution. This means that we can write down a normal distribution in mathematical notation as follows.

Normal distribution

A normal distribution is a continuous probability distribution with notation

$$X \sim \mathcal{N}(\mu, \sigma^2).$$

Here μ is the mean, σ is the standard deviation, and σ^2 is the variance.

The mean and standard deviation are enough to fully describe a normal distribution. As with data sets, the mean μ tells us where the centre of the distribution is, and the standard deviation σ tells us how spread out the distribution is. Normal distributions always satisfy the following useful properties.

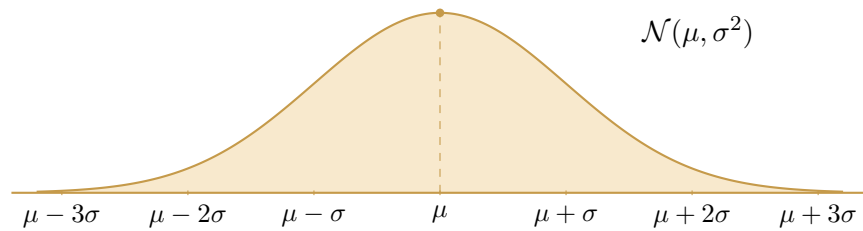
Four important properties

A normal distribution has the following properties.

1. The curve is bell-shaped, with its highest point over the mean μ .
2. The curve is symmetric about the vertical line through μ .
3. The tails approach the horizontal axis but never touch or cross it.
4. The total area under the entire curve is 1.

11.1.3 z-scores

The four properties written above mean that we can navigate a normal distribution as follows:



Here, the graph is symmetric and centered directly at μ . Since σ is defined in the same physical units as μ , we can start from the centre of the graph and “walk outwards” in increasing values of σ . This allows us to describe potential values of X in terms of their distance from μ . For instance, if X takes a value x that is far away from μ , then $x - \mu$ might be 3σ or -3σ , and if X takes a value relatively close to μ , then it could be 0.5σ away.

If we wanted to get rid of the σ in our measure of “distance from μ ”, then we can simply divide by σ to be left with a pure number. This number is known as the z -score:

$$z = \frac{x - \mu}{\sigma}.$$

In a nutshell, the z -score of a value x counts how far x is from μ *in terms of* σ . The larger z is, the further away x is from μ . If the overall z -score is negative, then that means the value x is to the left of μ , and if the z -score is positive, then that means x is to the right of μ .

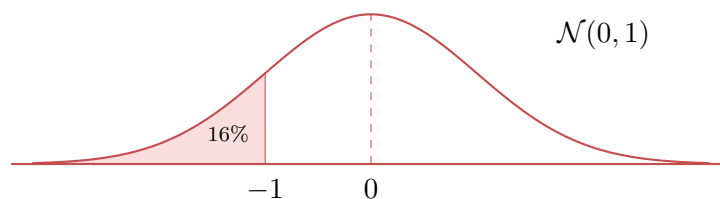
11.1.4 The standard normal distribution

Throughout Lecture 10 we also spent some time discussing the “standard” normal distribution. The idea here was that this is just about the simplest possible version of a normal distribution: the mean equalled 0 and the standard deviation was 1. Since this particular distribution is so well-behaved, we used the special letter Z to write its random variable. In mathematical notation, the standard normal distribution can be written $Z \sim \mathcal{N}(0, 1)$.

Conveniently, we can approximate areas (and therefore probabilities) of the standard normal distribution by using z -tables. These tables list out the areas to the left of common values of Z . For instance, a small version of this table might look like:

z	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0
$P(Z < z)$	0.02	0.07	0.16	0.31	0.50	0.69	0.84	0.93	0.98

More detailed tables can be found in the appendix of Lecture 10. Visually, entries in this table tell us the area to the *left* of a given value. For example, the table says that $P(Z < -1) \approx 0.16$, meaning that the area to the left of -1 in $\mathcal{N}(0, 1)$ is approximately 16% of the total area:



Tables such as that above only provide probabilities to the *left* of given values. Other probabilities can be calculated using the properties of the normal distribution. Some useful facts are listed below.

General rules for probabilities

For $Z \sim \mathcal{N}(0, 1)$:

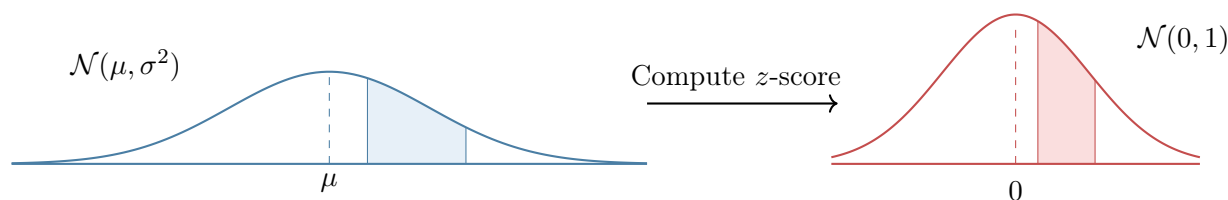
- $P(Z > z) = 1 - P(Z < z)$
- $P(z_1 < Z < z_2) = P(Z < z_2) - P(Z < z_1)$
- $P(Z > z) = P(Z < -z)$
- $P(Z = z) = 0$

11.1.5 Calculating other probabilities

If $X \sim \mathcal{N}(\mu, \sigma^2)$ is a general normal distribution, then we can convert a value x into the standard normal distribution by computing its z -score:

$$z = \frac{x - \mu}{\sigma}.$$

This transformation preserves area, and therefore the probability of finding X within a particular region can be calculated by first converting to the standard normal distribution, and then performing the calculation there:



Example: heights

To illustrate this point, suppose that male heights in Japan are normally distributed with mean 171 cm and standard deviation 5.6 cm. We can calculate, for example, the probability of meeting somebody taller than 180 cm by working with z -scores. Here we are looking for $P(X > 180)$, so we can compute the relevant z -score:

$$z = \frac{180 - 171}{5.6} \approx 1.61.$$

This transformation preserves the probability, so the answer to our question can be found by computing:

$$P(X > 180) \approx P(Z > 1.61),$$

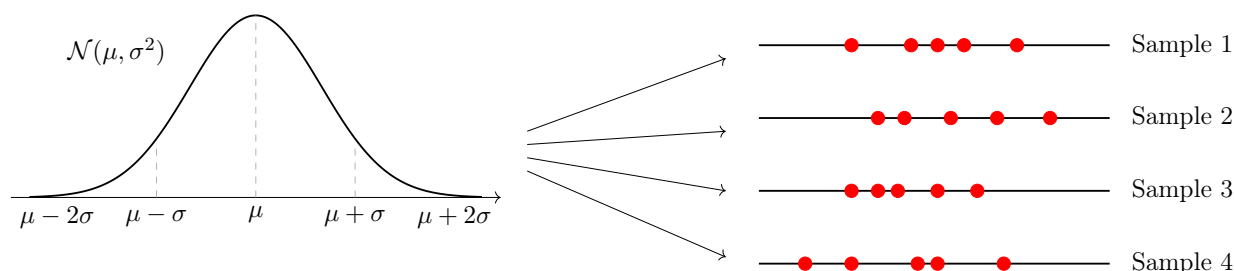
which happens to be about 5.4%.

11.2 Sampling Distributions

11.2.1 Sampling a random variable

Consider now a general normal distribution $X \sim \mathcal{N}(\mu, \sigma^2)$. As we previously mentioned, the probability of finding X within any given region can be determined by the area under the curve at that region. Therefore, we would expect that X is more likely to land in regions that correspond to larger areas.

Suppose now that we take a *sample* of random values from this distribution. For the sake of simplicity, let's pick 5 values randomly. Using the terminology of Lecture 1, this is known as a *simple random sample of size 5*. Generally, these 5 values could land anywhere on the x -axis. So, if we were to keep taking simple random samples of size 5 over and over again, they may look quite different:



For each individual sample, we can always compute a sample mean:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5}.$$

Now, here is the crucial observation: since our sample consists of random values, the exact value of the sample mean $\bar{x}$ will also be random. Therefore, we can interpret the sample mean as a random variable in its own right, and we write this random variable as $\bar{X}$. We will now explore what the distribution of the random variable $\bar{X}$ looks like.

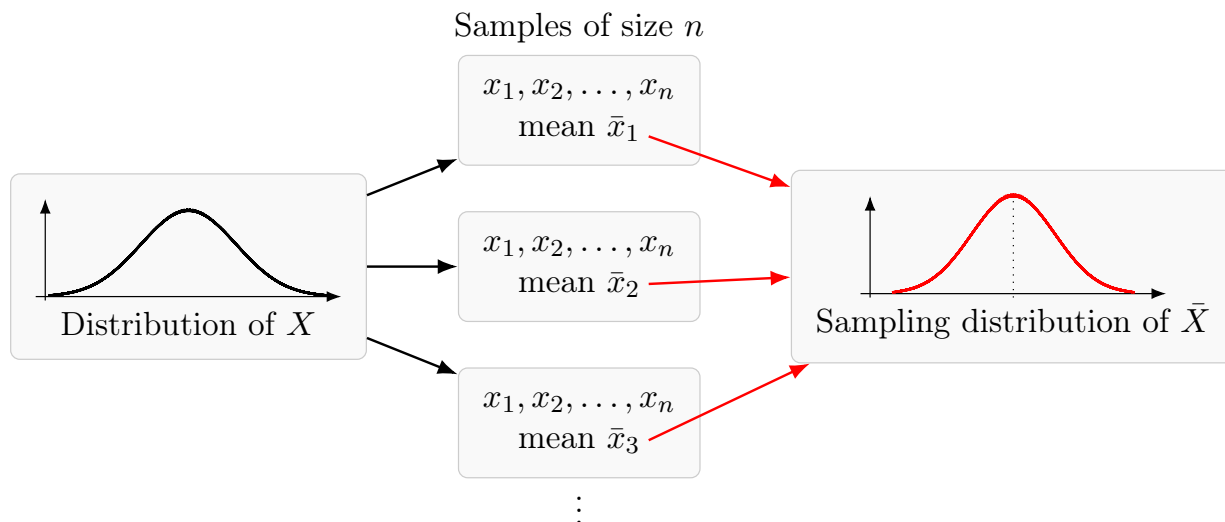
11.2.2 The sampling distribution of the sample mean

Imagine taking all possible simple random samples of size n from a population. We can compute the sample mean associated to each sample and collect all of these results together into a new probability distribution. This new object is known as a *sampling distribution*.

Sampling distribution of the sample mean

Given a probability distribution, the **sampling distribution of $\bar{X}$ for samples of size n** is the probability distribution formed from all sample means taken from all simple random samples of size n .

We can imagine the construction of a sampling distribution as something like this:



In other words, we start with a probability distribution, take all the simple random samples of size n , compute all of the associated sample means $\bar{x}$, and then collect them all together. Since the samples were randomly drawn to begin with, the sample means $\bar{x}$ can be treated as values of a random variable $\bar{X}$, and consequently, the collection of all possible $\bar{x}$'s has the structure of a probability distribution.

A sample of size n should be thought of as random variables $X_1, \dots, X_n$, each with the same distribution as X , and

$$\bar{X} = \frac{X_1 + \dots + X_n}{n}.$$

11.3 The Central Limit Theorem

As we might expect, different probability distributions will generally have different sampling distributions. The Central Limit Theorem is a result that tells us how these sampling distributions will look, depending on the shape of the original distribution. We will now present two related results: one for normal distributions, and another for more general probability distributions.

Throughout this course we will not prove these theorems. Instead, our main focus will be on the *application* of the Central Limit Theorem in techniques like Estimation and Hypothesis Testing. So, our goal is simply to state and understand the result.

11.3.1 Version 1: the sampling distribution when the original population is normal

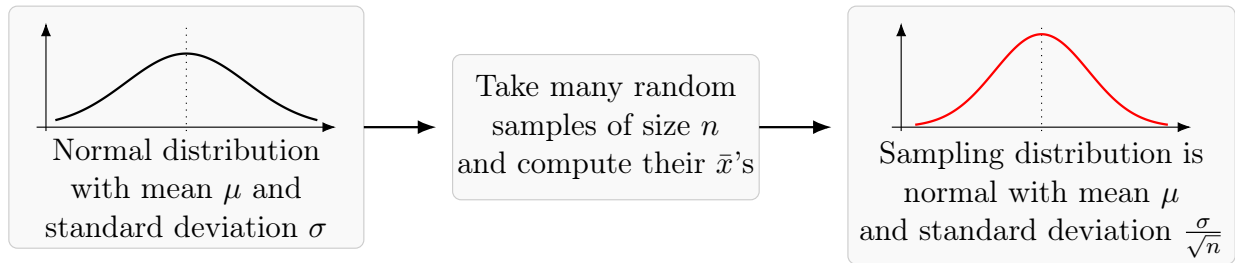
Let's first suppose that our original probability distribution is normal, i.e. $X \sim \mathcal{N}(\mu, \sigma^2)$. Our first result relates the sampling distribution of $\bar{X}$ to the parameters μ and σ .

The Central Limit Theorem: Version 1

Let X be a random variable such that $X \sim \mathcal{N}(\mu, \sigma^2)$. Then:

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

This result says that when we start with a normal distribution with mean μ and variance σ^2 , the collection of all sample means will arrange themselves as another normal distribution, with the same center μ but now with variance $\frac{\sigma^2}{n}$:



Roughly speaking, if we take a sample of values from $\mathcal{N}(\mu, \sigma^2)$, then the sample mean $\bar{x}$ will average out the random noise in these individual values and return something more central. Therefore, on average, we would expect these sample means to be closer to the true center of the distribution. This idea partially explains why the variance $\frac{\sigma^2}{n}$ of the sampling distribution is smaller than the original distribution. As we can see from the formula, the larger the value of n , the more tightly packed the sampling distribution will be – this is because samples of larger size are less sensitive to individual outliers which may skew the average.

11.3.2 Version 2: when the original population is not necessarily normal

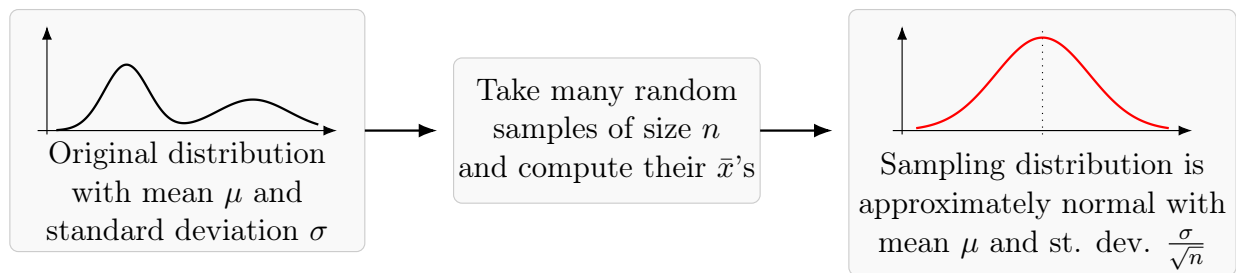
The first result says that if we build a sampling distribution of $\bar{X}$ from a normal distribution, then the result is another normal distribution. Amazingly, there is a much stronger version of the Central Limit Theorem:

Central Limit Theorem

Suppose X has a probability distribution with finite mean μ and finite variance σ^2 . For a large enough sample size n (often $n \geq 30$), the sample mean $\bar{X}$ is approximately normally distributed, with parameters:

$$\bar{X} \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

In other words, the Central Limit Theorem says that the sampling distribution of $\bar{X}$ will look approximately normal, with mean μ and variance $\frac{\sigma^2}{n}$.



In both of our results, the standard deviation of the sampling distributions had the same general form. As a matter of fact, this form has a special name: it is called the *standard error*.

Standard error

The standard deviation of the sampling distribution of $\bar{X}$ is

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

This is called the **standard error**.

The standard error measures how much sample means vary from sample to sample. As the sample size n increases, the standard error decreases, meaning that the resulting sampling distribution will be more tightly packed around the center μ . As mentioned previously, this matches our intuition that larger samples are less sensitive to outliers, i.e. they give more stable averages.

11.3.3 Why the Central Limit Theorem matters

In the real world we don't usually have access to population parameters. For instance, if we were performing a study on the average height of people in Tokyo, then we cannot go out and measure the heights of the millions of people in the city. Instead, in reality, we will only ever be able to take *samples* of the population and compute some sample statistics.

The Central Limit Theorem provides a mathematical way to connect sample statistics to the true population mean. Although we may never understand the true structure of the population's distribution, we can at least expect that if we take our samples to be large enough (usually $n \geq 30$), then the distribution of all their sample means is approximately normal. In particular, the center of the sampling distribution is still the *true* population mean. Thus, the Central Limit Theorem offers us some extra expressivity – we can now talk about the theoretical relationship between a sample mean $\bar{x}$ and the population mean μ . This idea of inferring population parameters from sample statistics lies at the beating heart of Inferential Statistics. In the next few lectures we will explore this idea in great detail: we will see some techniques for actually *estimating* μ based on some sample data.

11.3.4 A Worked Example: Guinea Pig Weights

Suppose that a large population of guinea pigs has mean 950g and standard deviation 150g. Assume that the population is not normally distributed. Now suppose we take a random sample of 30 guinea pigs and compute the sample mean $\bar{x}$. We will answer the question:

What is the probability that the average weight of 30 guinea pigs is between 900g and 1000g?

Since $n = 30$, we can use the Central Limit Theorem to approximate the sampling distribution of $\bar{X}$:

$$\bar{X} \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) = \mathcal{N}\left(950, \frac{150^2}{30}\right).$$

Equivalently, the standard error is

$$\sigma_{\bar{X}} = \frac{150}{\sqrt{30}} \approx 27.39.$$

So we may also write

$$\bar{X} \approx \mathcal{N}(950, 27.39^2).$$

To answer the question above, we are looking to find the probability $P(900 \leq \bar{X} \leq 1000)$. To do so, we need to first convert both endpoints to z -scores:

$$z_1 = \frac{900 - 950}{150/\sqrt{30}} \approx -1.83,$$

and

$$z_2 = \frac{1000 - 950}{150/\sqrt{30}} \approx 1.83.$$

Therefore,

$$P(900 \leq \bar{X} \leq 1000) \approx P(-1.83 \leq Z \leq 1.83).$$

Using a standard normal table, we have:

$$P(Z < 1.83) = 0.96638 \quad \text{and} \quad P(Z < -1.83) = 0.03362,$$

from which we can conclude that:

$$P(900 \leq \bar{X} \leq 1000) \approx P(-1.83 \leq Z \leq 1.83) = 0.96638 - 0.03362 = 0.93276.$$

So, the probability of taking a random sample of 30 guinea pigs with an average weight between 900g and 1000g is about 93%.

Why this is useful

If we were forced to calculate this probability directly, then we would be left with a very awkward computation. Firstly, if we tried to list all possible samples of 30 guinea pigs from 1000 guinea pigs, the number of possible samples would be

$$\binom{1000}{30} \approx 2.43 \times 10^{57}.$$

To calculate the exact probability directly using our Golden Formula for classical probability, we would need to:

1. list all possible samples of size 30;
2. compute the sample mean for every sample;
3. count how many of those sample means lie between 900 and 1000, and then
4. divide by $C_{1000,30}$, the total number of possible samples.

Of course, this would be much more difficult than simply approximating the solution via the Central Limit Theorem.

11.4 Normal Approximation to the Binomial Distribution

11.4.1 The Problem with Large Binomial Experiments

Recall from Lecture 8 that a binomial experiment is a probabilistic system with the following properties:

1. the same trial is repeated several times,
2. there are only ever two possible outcomes to each trial (which we called *success* and *failure*),
3. each trial has the same probabilities, and
4. the probabilities are independent.

Binomial distributions were specified by two main numbers: n , which denoted the number of trials, and p , which denoted the probability of success. The central question in a binomial experiment is to figure out the exact probability of getting a particular number, or range of numbers, of successes in n attempts. To properly write this mathematically, we defined a discrete random variable X whose job was to count the number of successes in n trials. We then saw that the probability of getting r successes in n trials was given by the formula:

$$P(X = r) = C_{n,r} p^r q^{n-r},$$

where here $q = 1 - p$.

Imagine now that we have a rather large binomial experiment, say $n = 50$ with $p = 0.25$. If we were asked to compute a complicated probability such as $P(X \geq 15)$, then this would require us to compute many individual probabilities and add them up:

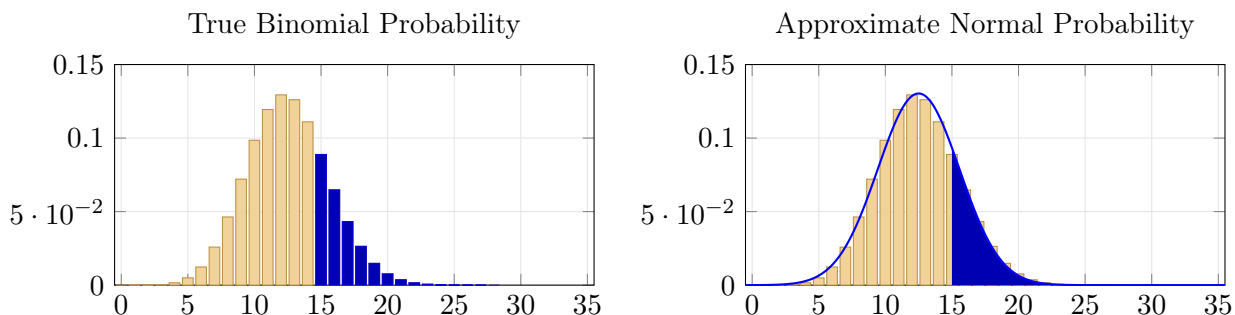
$$\begin{aligned} P(X \geq 15) &= P(X = 15) + P(X = 16) + \cdots + P(X = 49) + P(X = 50) \\ &= C_{50,15} p^{15} q^{35} + C_{50,16} p^{16} q^{34} + \cdots + C_{50,49} p^{49} q^1 + C_{50,50} p^{50} q^0, \end{aligned}$$

or perhaps via the complement:

$$\begin{aligned} P(X \geq 15) &= 1 - P(X < 15) \\ &= 1 - (P(X = 0) + P(X = 1) + \cdots + P(X = 13) + P(X = 14)) \\ &= 1 - P(X = 0) - P(X = 1) - \cdots - P(X = 13) - P(X = 14) \\ &= 1 - (C_{50,0} p^0 q^{50}) - (C_{50,1} p^1 q^{49}) - \cdots - (C_{50,13} p^{13} q^{37}) - (C_{50,14} p^{14} q^{36}) \end{aligned}$$

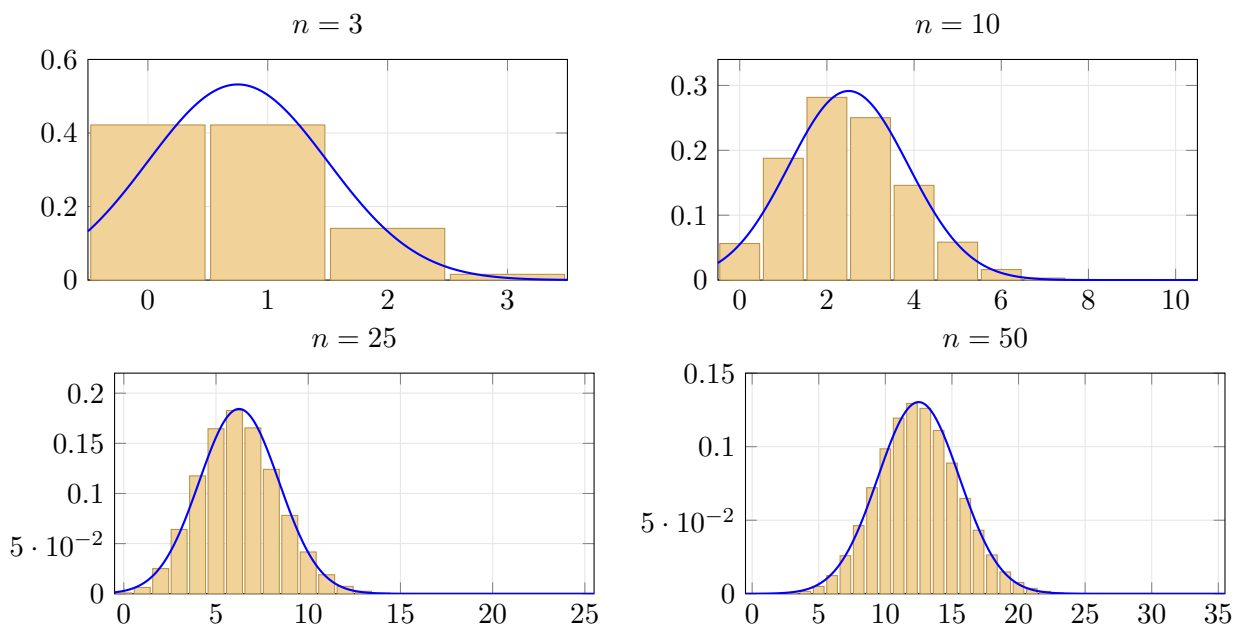
Either way we try to do this, we are met with a large and difficult computation. Instead, in a situation like this, it may be easier to simply approximate the distribution with a normal curve and

use z -scores to get a rough answer. Graphically, the true probability corresponds to the sum of the heights of all of the bars from 15 to 50, versus computing the area to the right of 15 underneath a normal curve that approximately matches the distribution:



Of course, this point becomes clearer if we consider a much larger binomial experiment, say $n = 1000$. Suddenly, calculating a probability like $P(X \geq 400)$ would involve 601 different computations of the formula $C_{n,r}p^r q^{n-r}$, followed by a rather large sum.

In Lecture 10 we motivated normal distributions by seeing how binomial distributions begin to resemble normal curves after appropriate scaling. Visually, taking n to be larger and larger caused the number of bars to increase until the image began to resemble a continuous curve. This means that as n increases, the binomial distribution will start to look more and more like a normal curve, and therefore the approximation of probabilities using z -scores will become more and more appropriate. In the image below we fit normal curves to smaller binomial distributions. Notice the normal curves start to fit the binomial distributions better and better as n gets larger:



We will now see how these normal curves can be created via an application of the Central Limit Theorem.

11.4.2 Applying the Central Limit Theorem to a Binomial Distribution

Suppose that we have a binomial experiment with fixed n , p and q . Let X be the random variable that counts the number of successes. In a way, we can understand each individual trial as a separate probabilistic system, with its own random variable attached. If we define:

$$Y_i = \begin{cases} 1 & \text{if trial } i \text{ is successful} \\ 0 & \text{if trial } i \text{ fails,} \end{cases}$$

then X can be rewritten as the sum:

$$X = Y_1 + Y_2 + \cdots + Y_{n-1} + Y_n.$$

Each term in this sum contributes either a 1 for success, or a 0 for failure. Therefore the sum X will count the total number of successes after n trials. We now notice: if we divide X by n , then we get something that looks conveniently like a sample mean:

$$\frac{X}{n} = \frac{Y_1 + Y_2 + \cdots + Y_{n-1} + Y_n}{n}.$$

The Central Limit Theorem says that the sample mean of the random variables $Y_1, \dots, Y_n$ will arrange itself approximately into a normal distribution $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$, where μ and σ^2 are the mean and variance of a single trial variable Y_i . In this case, our original variable is one of the Y_i – the variable associated to a single trial of the experiment. When working with Binomial experiments, we know that the analogue of the mean is the expected value. For the single variable Y_i , we have:

$$Y_i = \begin{cases} 1 & \text{if the trial is successful} \\ 0 & \text{if the trial fails.} \end{cases}$$

The expected value¹ is:

$$E(Y_i) = P(Y_i = 0) \cdot 0 + P(Y_i = 1) \cdot 1 = q \cdot 0 + p \cdot 1 = p.$$

Moreover, the variance is:

$$\begin{aligned} \text{Var}(Y_i) &= E((Y_i - E(Y_i))^2) \\ &= E((Y_i - p)^2) \\ &= P(Y_i = 0)(0 - p)^2 + P(Y_i = 1)(1 - p)^2 \\ &= q(-p)^2 + p(q)^2 \\ &= qp^2 + pq^2 \\ &= pq(p + q) \\ &= pq. \end{aligned}$$

¹Recall that, for a discrete random variable, the expected value of a random variable is roughly

$$E(X) = \sum \text{probability} \times \text{outcome}$$

To summarise: the variable $\frac{X}{n}$ acts as a sort-of sample mean of the outcomes of each individual trial in the binomial experiment. Each of these trials has the same expected value equal to p , and the same variance equal to pq . Applying the Central Limit Theorem tells us that $\frac{X}{n}$ is approximately normal, specifically:

$$\frac{X}{n} \approx \mathcal{N}\left(p, \frac{pq}{n}\right).$$

We can use this information to create a new approximation for our original variable X . Some complicated algebra involving normal distributions² gives us the following fact:

$$\frac{X}{n} \approx \mathcal{N}\left(p, \frac{pq}{n}\right) \quad \implies \quad X \approx \mathcal{N}(np, npq).$$

Thus we have arrived at a normal approximation of X .

Normal approximation to a binomial distribution

If X is binomially distributed according to n trials with success probability p , and using the rough rule $np > 5$ and $nq > 5$, then X can be approximated by the normal distribution:

$$X \approx \mathcal{N}(np, npq).$$

Written differently, we say that $X \approx Y$, where $Y \sim \mathcal{N}(np, npq)$.

Note that the conditions $np > 5$ and $nq > 5$ are imposed in order to make sure that the original binomial distribution is large enough and not too skewed to have a meaningful approximation. Notice, for instance, the diagrams in Section 4.1: the approximation of the $n = 3, p = 0.25$ case doesn't really fit the original distribution.

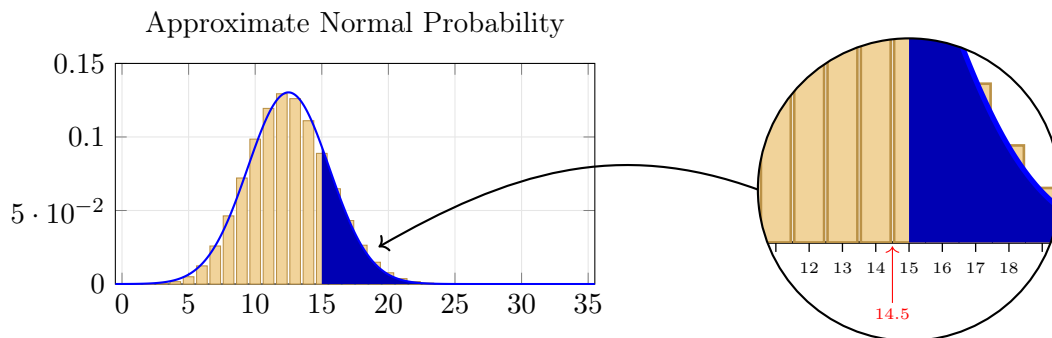
11.4.3 The Continuity Correction

According to the previous section, when a binomial distribution is large enough and not too skewed, we can approximate it with a normal distribution $\mathcal{N}(np, npq)$. The latter then allows us to compute z -scores and derive probabilities of areas. Although potentially useful, there is a genuine geometrical issue when approximating bars with a smoother area. Let's consider again the example of $n = 50$ and $p = 0.25$ from Section 4.1. If we zoom into the diagram, we notice that the approximate continuous area happens to cut the bar corresponding to $X = 15$ in half:

²If I must: we can substitute $\frac{x}{n}$ into the probability density function that defines the Normal distribution to see that

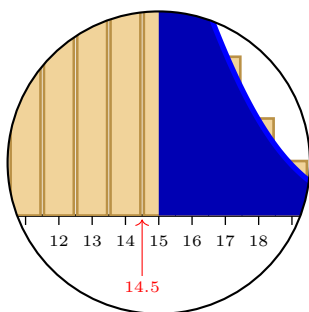
$$f\left(\frac{x}{n}\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\left(\frac{x}{n} - \mu\right)^2}{2\sigma^2}\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - n\mu)^2}{2n^2\sigma^2}\right).$$

This *almost* looks like the definition of $\mathcal{N}(n\mu, n^2\sigma^2)$, except for the first term. If we add an extra $\frac{1}{n}$ to the front of the expression, then we notice that $\frac{1}{n}f\left(\frac{x}{n}\right)$ now matches perfectly with the definition of $\mathcal{N}(n\mu, n^2\sigma^2)$.

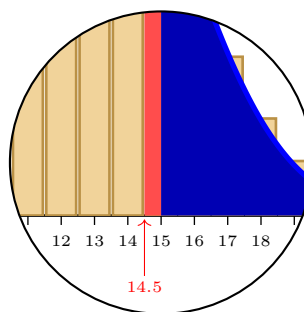


Therefore, the area corresponding to the blue region is going to be smaller than the true probability. We can correct this by making our area approximation start counting from the *start* of the bar, instead of in the middle:

Uncorrected Approximation



Corrected Approximation



Generally, this will give a closer approximation than the uncorrected version. This adjustment is called the *continuity correction*.

Rules for continuity corrections

When approximating a binomial random variable X by a normal random variable Y , move each cutoff by 0.5 so that the normal area fully covers the bars of the binomial distribution.

Binomial probability		Normal approximation
$P(X \geq r)$	becomes	$P(Y \geq r - 0.5)$
$P(X > r)$	becomes	$P(Y > r + 0.5)$
$P(X \leq r)$	becomes	$P(Y \leq r + 0.5)$
$P(X < r)$	becomes	$P(Y < r - 0.5)$
$P(a \leq X \leq b)$	becomes	$P(a - 0.5 \leq Y \leq b + 0.5)$

For example, if we wanted to approximate $P(15 \leq X)$ then we would compute $P(14.5 \leq Y)$, and if we wanted to approximate $P(X \leq 12)$ then we would compute $P(Y \leq 12.5)$.

11.5 Worked Example: Vintage Clothes

We will now walk through an example of a binomial approximation. Suppose that I head down to Shimokitazawa to try and buy some cool vintage clothes. From past experience, I know that there is about a 25% chance that I will find something interesting to buy in any given store. We assume the stores are independent and that this probability stays the same from store to store. I walk into 25 different stores. We will try to find an approximate answer to the question:

What is the probability that 8 or more of the 25 stores will have good clothes?

Assuming that we can model this as a binomial experiment, let's define the random variable

X = the number of stores with good clothes.

By assumption, the variable X follows a binomial distribution with $n = 25$, $p = 0.25$ and $q = 1 - 0.25 = 0.75$.

11.5.1 Method 1: an exact binomial calculation

For future reference, we will first compute this result exactly. We are being asked to find $P(X \geq 8)$. This can be calculated directly as:

$$\begin{aligned} P(X \geq 8) &= P(X = 8) + P(X = 9) + \cdots + P(X = 25) \\ &= C_{25,8}(0.25)^8(0.75)^{17} + C_{25,9}(0.25)^9(0.75)^{16} + \cdots + C_{25,25}(0.25)^{25}(0.75)^0. \end{aligned}$$

Using a calculator, we find that:

$$P(X \geq 8) \approx 0.27349.$$

11.5.2 Method 2: normal approximations

In order to use a normal approximation, first we have to check the approximation conditions:

$$np = 25(0.25) = 6.25 \quad \text{and} \quad nq = 25(0.75) = 18.75.$$

These are both greater than 5, so a normal approximation is reasonable. Since we have $n = 25$, $p = 0.25$ and $q = 0.75$, the approximating normal distribution is:

$$X \approx \mathcal{N}(np, npq) = \mathcal{N}(6.25, 4.6875).$$

Without applying a continuity correction, we calculate an approximate probability by computing the z -score:

$$z = \frac{8 - 6.25}{\sqrt{4.6875}} \approx 0.81.$$

Therefore:

$$P(X \geq 8) \approx P(Z > 0.81) = 1 - P(Z < 0.81) \approx 1 - 0.79103 = 0.20897.$$

As we can see, this is noticeably smaller than the exact answer of 0.27349. However, if we make the continuity correction before computing the probability, we would change $P(X \geq 8)$ into $P(Y \geq 7.5)$. Plugging this new value into the z -score formula, we get:

$$z = \frac{7.5 - 6.25}{2.17} \approx 0.58 \quad \implies \quad P(Y \geq 7.5) \approx P(Z \geq 0.58) \approx 0.28096.$$

Notice that this is significantly closer to the true value of 0.27349.

Lecture 12

An Introduction to Estimation

In this lecture, we will use our theory of the normal distributions for *estimation*. The basic question is simple: if we cannot measure an entire population, can we still make a reasonable guess about an unknown population parameter, such as μ ? As we will see, the answer is yes. We will assume, for now, that the population standard deviation σ is already known. That way, we can use the Central Limit Theorem to connect observed sample data to the true population mean. Using the machinery of normal distributions, we will then create *estimates* for the value of μ by constructing tolerable error bounds around sample means $\bar{x}$ – a construction known as a *confidence interval*.

12.1 A Review of the Central Limit Theorem

12.1.1 The sampling distribution

Suppose that we have some population distribution with mean μ and standard deviation σ . If we take a simple random sample of size n , then we can compute its sample mean:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

If we repeat this process again and again, we get many possible sample means. The collection of all of these sample means forms a new distribution, called a *sampling distribution*.

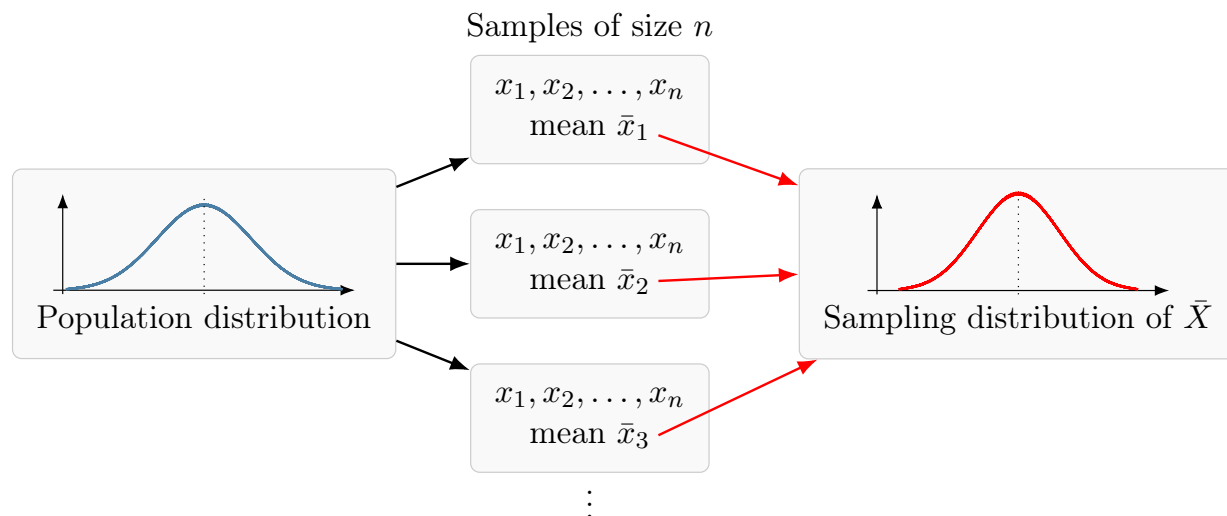
Sampling distribution of the sample mean

Given a probability distribution, the **sampling distribution of $\bar{X}$ for samples of size n** is the probability distribution formed from all sample means taken from all simple random samples of size n .

The important point here is that the samples are taken randomly. This means that the exact value of $\bar{x}$ is also going to be random. Before we take a particular sample we don't know which actual values we will get, and therefore we also don't know what the exact value of $\bar{x}$ will be. So, we can upgrade the sample mean into a random variable, written $\bar{X}$. This is used as a placeholder for all the possible values that the sample mean can take, depending on all of the possible random samples we could take from the population.

12.1.2 The Central Limit Theorem

Generally speaking, the Central Limit Theorem is a mathematical result that tells us what the sampling distribution of $\bar{X}$ looks like. If the original population is normal, then the sampling distribution of $\bar{X}$ is exactly normal:



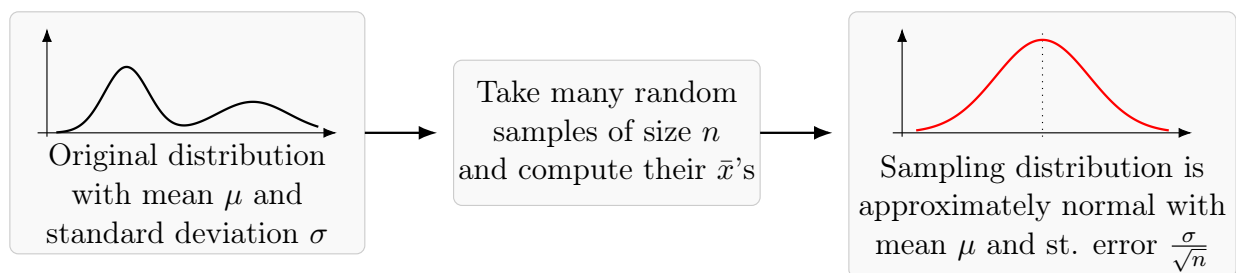
Moreover, if the original population is *not* normal, then the Central Limit Theorem says that the resulting sampling distribution of $\bar{X}$ is approximately normal, provided that n is large enough. Formally, one statement of the Central Limit Theorem is as follows.

Central Limit Theorem

If a population has mean μ and standard deviation σ , then for large enough n , the sampling distribution of $\bar{X}$ satisfies:

$$\bar{X} \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

This means that the sampling distribution is centred at the true population mean μ , but its spread is smaller than the spread of the original population. The larger the sample size n , the smaller the spread of the sample means.



In the statement of the Central Limit Theorem, the variance of the sampling distribution is equal to the particular quantity $\frac{\sigma^2}{n}$. The square root of this, $\frac{\sigma}{\sqrt{n}}$, is the standard deviation of the sampling distribution. To avoid confusing this with the standard deviation σ of the original population, we use different terminology.

Standard error

The standard deviation of the sampling distribution of $\bar{X}$ is

$$\frac{\sigma}{\sqrt{n}}.$$

This is called the **standard error**.

12.2 The Idea of Estimation

12.2.1 Estimation with Known σ

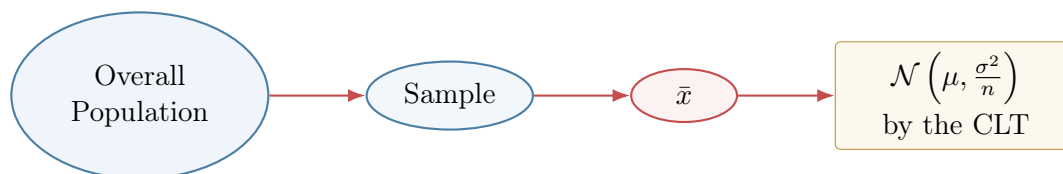
In real statistical problems, we usually do not know the true population mean μ . Knowing something like the true average of a population is extremely useful, and it is certainly a common parameter that we would like to know. However, in practice, we cannot go out into the world and measure every single object in a population in order to calculate the true mean.

In this lecture, we are going to assume that:

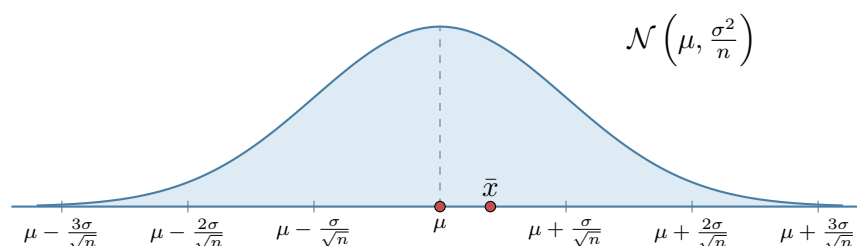
- the population mean μ is unknown, but
- the population standard deviation σ happens to be *known*.

This is obviously an unrealistic scenario. However, theoretically it is the simplest setup to estimate μ . Next lecture we will relax this assumption and estimate μ in the realistic scenario in which σ is also unknown.

The benefit of assuming that σ is known is that we may use the Central Limit Theorem to connect sample means to the true population mean:



Graphically, if we take a simple random sample, then the Central Limit Theorem tells us that the observed value of $\bar{x}$ will sit somewhere in this normal distribution:

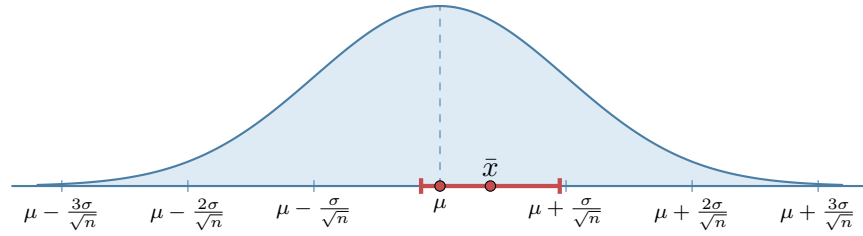


Assuming that σ is known allows us to exploit the structure of the sampling distribution $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$. For instance, we don't know where μ is, but we know that $\bar{X}$ has about a 68% chance of landing within one standard error of μ . Why? Because the randomness of $\bar{X}$ is encoded in the shape of the distribution – z -score calculations from Lecture 10 tell us that 68% of the area of the distribution is within one standard error of the center.

Importantly, these sorts of ideas are still true *even if we don't know where μ actually is*. According to the Central Limit Theorem, the randomness of $\bar{X}$ follows a normal distribution with the true population mean at the center. As we will see in a moment, even if we don't know what μ is, we can at least *relate* it to an observed value $\bar{x}$.

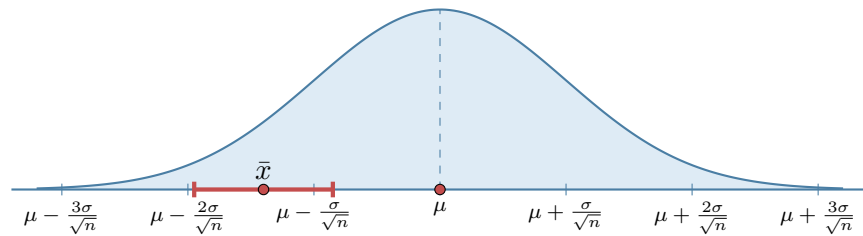
12.2.2 Intervals around sample means

If we pick a sample at random, then there is no guarantee that the resulting mean $\bar{x}$ will be exactly the same as μ . As a matter of fact, we would generally expect that every sample mean is slightly different from the true mean, ultimately due to random noise. However, at the very least, we can always define a sense in which $\bar{x}$ is *close enough* to μ . Given that we don't know what μ is, instead of saying “this is where μ is exactly”, it seems the best we can say is “ μ is probably here, within some acceptable margin of error”. Mathematically, this “acceptable margin of error” will look like an interval:



In the above picture, our randomly chosen value of $\bar{x}$ has an interval around it of a fixed length. Although μ is different from $\bar{x}$, it still lies in this neighbourhood of $\bar{x}$.

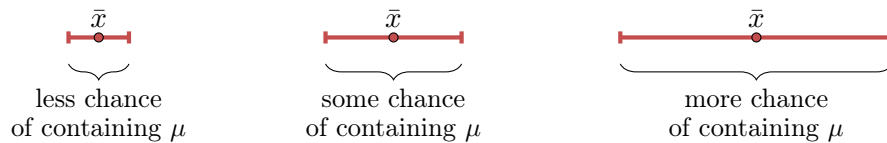
Of course, sometimes the sample will be statistically unlikely, which will cause the resulting value of $\bar{x}$ to be far away from the center:



This is bound to happen due to the nature of randomness. In fact, given some random value of $\bar{X}$, the best we can say is something like:

We think with $c\%$ confidence that this interval contains μ .

This is ultimately a statement about uncertainty: we don't *know* where μ is, but instead can say that intervals of this particular length would capture it a certain proportion of the time. Of course, if we made the interval longer, then it would be more likely to hit μ , and if we made the interval shorter then it would be less likely to contain μ :



So, it seems that our degree of confidence is directly related to the length of the interval: if we wanted to be very certain that the interval around our randomly-obtained $\bar{x}$ contained μ , then we should make the interval very large. Conversely, if we make the interval very small, then we are less certain that it will contain μ . This gives us the idea of a *confidence interval*.

Confidence interval

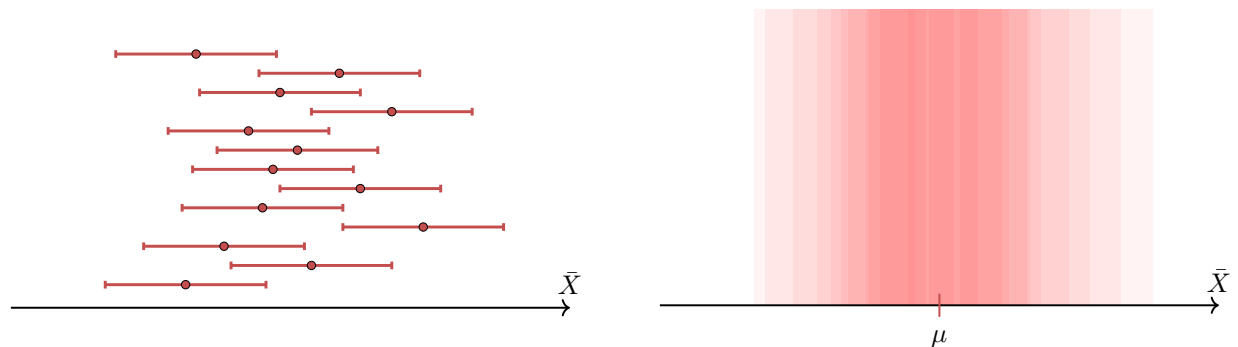
A **confidence interval** is an interval estimate for an unknown population parameter. The confidence level describes how often the method captures the true parameter in repeated sampling.

We often write the confidence level with the letter c . The idea here is that c can be written as a percentage, and there are different confidence intervals corresponding to different levels of confidence. If we want to be more confident, then we increase c and this will increase the length of the resulting interval.

12.2.3 Interpretation of the Confidence Level

For the sake of illustration, let's assume that we have constructed a confidence interval with confidence level $c = 95\%$. By construction, this means that the random interval produced by the method will contain the true population mean μ 95% of the time. In other words, if we were to take 100 simple random samples and compute 95% confidence intervals around their sample means, then on average we would expect that 95 of these intervals would contain μ and 5 would not.

The picture below helps illustrate the long-run meaning of the confidence level: if we kept taking samples over and over again, then most of the resulting confidence intervals would contain μ :



In the image on the right, we have taken the various confidence intervals from the left-hand side and overlaid them. As you can see, many of the intervals overlap near the true population mean μ .

Before discussing the construction of confidence intervals, we must first dispel a common misconception.

Common misconception

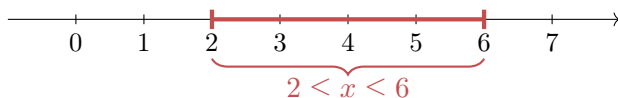
Suppose that we construct a 95% confidence interval around some observed $\bar{x}$. Here, the “95%” does **not** mean that there is a 95% chance that the interval contains μ . Instead, it refers to the long-run success rate of the overall method.

Saying something like “there is a 95% chance that this interval contains μ ” implies that μ is somehow changing its value randomly. This is not true: μ is unknown, but its value is absolutely fixed. For any interval around a chosen $\bar{x}$, the mean μ is either in the interval or it is not. Instead, it is the value of $\bar{x}$ that is randomly moving around in the distribution (as represented with the notation $\bar{X}$). The confidence interval moves around with the value of $\bar{x}$, and the confidence level refers to *this* movement.

12.3 Constructing Confidence Intervals

12.3.1 A reminder about intervals

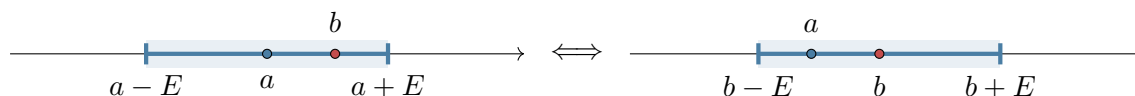
Generally speaking, a connected interval is fully described by its endpoints: we simply write where the interval starts and finishes, and from there it is understood that the interval itself is everything in between. For example, the interval $[2, 6]$ means all numbers x satisfying $2 \leq x \leq 6$:



Suppose now that a and b are two points on the real line. Saying that “ b is within distance E of a ” means the same thing as saying that “ a is within distance E of b ”. In symbols:

$$a - E \leq b \leq a + E \quad \Longleftrightarrow \quad b - E \leq a \leq b + E.$$

Geometrically, we see that if b is in an interval of length $2E$ centered at a , then a will also be in the interval of length $2E$ centered at b :



When constructing our confidence intervals, our goal will be to specify the two endpoints of each of the intervals. We are going to be looking for two endpoints $\bar{x}_1$ and $\bar{x}_2$ on either side of an observed $\bar{x}$, such that $\bar{x}$ is exactly in the center. That way, in symbols we have:

$$\mu \text{ is in the interval centered at } \bar{x} \quad \Longleftrightarrow \quad \bar{x} - E \leq \mu \leq \bar{x} + E.$$

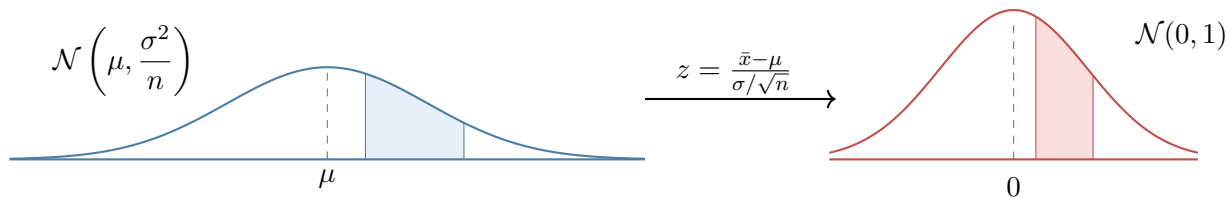
This observation is the key idea behind confidence intervals: if $\bar{X}$ is close to μ , then μ is also close to $\bar{X}$.

12.3.2 The z -score for the sampling distribution

According to the Central Limit Theorem, we know that the sampling distribution of $\bar{X}$ can for large enough n be approximated by: $\bar{X} \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$. Since this is a normal distribution, we can relate it to the Standard Normal Distribution $\mathcal{N}(0, 1)$ by creating a z -score. The formula for the z -score measures the distance between a value and the center, and then divides by the standard deviation of the distribution. In the case of the sampling distribution, the center is μ and the distribution is measured in terms of the standard error $\frac{\sigma}{\sqrt{n}}$. Therefore, the z -score formula that we have to use is:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}.$$

This converts the sampling distribution of $\bar{X}$ into the standard normal distribution, and allows us to compute areas in the sampling distribution using a z -table:

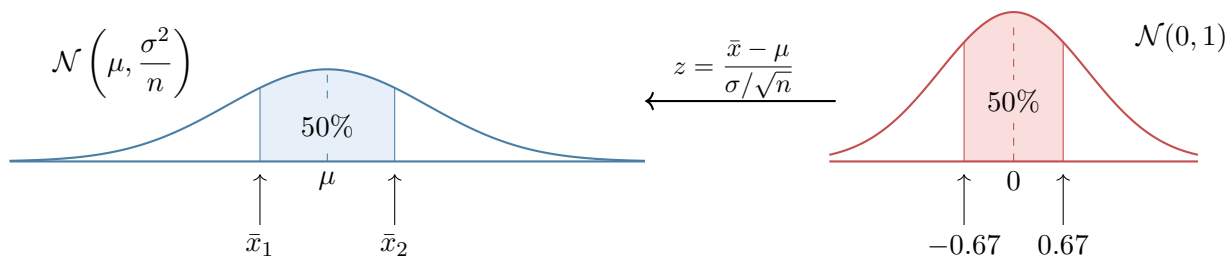


12.3.3 Finding a 50% confidence interval

Suppose we want $\bar{X}$ to be close to μ with probability 50%. We can describe confidence intervals around the values of $\bar{X}$ by first looking for 50% of the area centered around 0 in the standard normal distribution. From there, we can use the z -score formula and *work backwards* to figure out the bounds of the same area, this time in the sampling distribution $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$. According to the z -table, we can see that the values of z_1 and z_2 that give equal tail areas and satisfy

$$P(z_1 < Z < z_2) = 0.5$$

are precisely $z_1 = -0.67$ and $z_2 = 0.67$. We can now use these two values to work backwards through the z -score formula and find the corresponding values in the sampling distribution:



These values will bound 50% of the area around μ , meaning that $\bar{X}$ will have a 50% chance of landing in this region.

We can find the values of $\bar{x}_1$ and $\bar{x}_2$ by substituting in the values of ± 0.67 for z and solving algebraically. Using $z = 0.67$, we have:

$$0.67 = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \implies \bar{x} = \mu + \frac{0.67\sigma}{\sqrt{n}}.$$

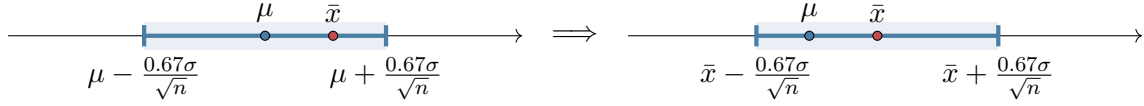
Similarly, using $z = -0.67$ gives:

$$-0.67 = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \implies \bar{x} = \mu - \frac{0.67\sigma}{\sqrt{n}}.$$

These two endpoints can now be used to describe an interval around μ for which $\bar{X}$ has a 50% chance of being inside. In symbols, by construction our interval satisfies:

$$P\left(\mu - \frac{0.67\sigma}{\sqrt{n}} < \bar{X} < \mu + \frac{0.67\sigma}{\sqrt{n}}\right) = 0.5.$$

Here is where we use our observation from Section 4.1: if $\bar{X}$ is nearby μ 50% of the time, then that means that μ is also nearby $\bar{X}$ with probability 0.5:



Therefore, if we keep the same length of the interval, we can center it around $\bar{X}$ and by construction this interval will contain μ 50% of the time. In other words, we have just created a 50% confidence interval around $\bar{X}$. We write this as:

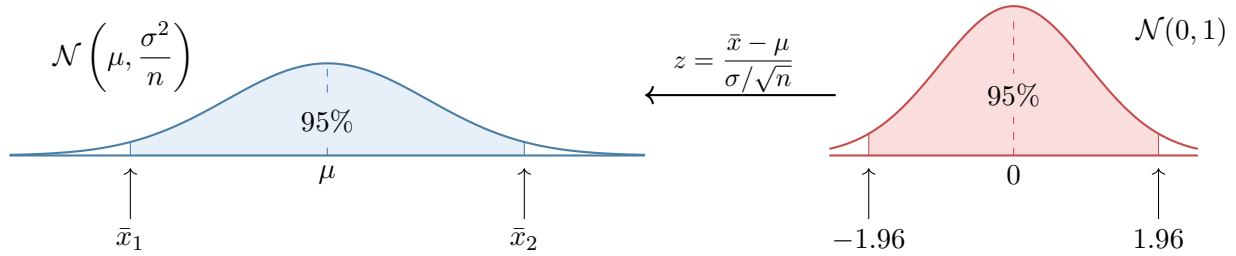
$$\left[\bar{X} - \frac{0.67\sigma}{\sqrt{n}}, \bar{X} + \frac{0.67\sigma}{\sqrt{n}} \right].$$

After observing a particular sample mean $\bar{x}$, we write the confidence interval around it as:

$$\left[\bar{x} - \frac{0.67\sigma}{\sqrt{n}}, \bar{x} + \frac{0.67\sigma}{\sqrt{n}} \right].$$

12.3.4 Finding a 95% confidence interval

Suppose now that we would like to construct a 95% confidence interval instead. The same logic still holds, except that this time, we simply seek out the values of z that will give a 95% area centered around 0 in $\mathcal{N}(0, 1)$. Again, by using a z -table, these happen to be ± 1.96 . We now use the z -score formula to find the corresponding values of the endpoints in the sampling distribution:



Now using $z = \pm 1.96$ instead of ± 0.67 , we find:

$$1.96 = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \implies \bar{x} = \mu + \frac{1.96\sigma}{\sqrt{n}}$$

$$-1.96 = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \implies \bar{x} = \mu - \frac{1.96\sigma}{\sqrt{n}}$$

By the same logic as before, the region between $\bar{x}_1$ and $\bar{x}_2$ will contain $\bar{X}$ 95% of the time:

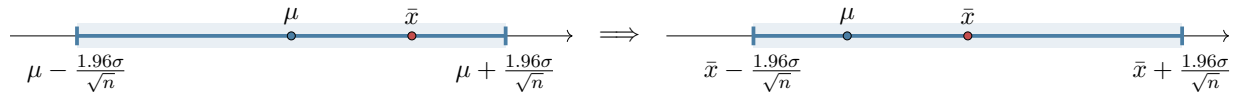
$$P\left(\mu - \frac{1.96\sigma}{\sqrt{n}} < \bar{X} < \mu + \frac{1.96\sigma}{\sqrt{n}}\right) = 0.95.$$

Therefore, if we take the same interval and instead center it at $\bar{X}$, then μ will be contained in this new interval 95% of the time as well:

$$P\left(\bar{X} - \frac{1.96\sigma}{\sqrt{n}} < \mu < \bar{X} + \frac{1.96\sigma}{\sqrt{n}}\right) = 0.95.$$

This time, geometrically we have created a wider interval around $\bar{x}$:¹

¹This image is not to scale. In fact, the real interval should be about 3 times larger than the interval drawn in the previous section.



In symbols, a 95% confidence interval around $\bar{X}$ can be written:

$$\left[\bar{X} - \frac{1.96\sigma}{\sqrt{n}}, \bar{X} + \frac{1.96\sigma}{\sqrt{n}} \right],$$

or alternatively:

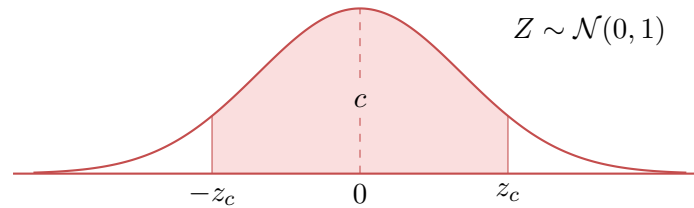
$$\left[\bar{x} - \frac{1.96\sigma}{\sqrt{n}}, \bar{x} + \frac{1.96\sigma}{\sqrt{n}} \right],$$

if a value of $\bar{x}$ has already been observed.

12.4 The General Formula for Confidence Intervals

12.4.1 Critical values

Suppose now that we want to construct a confidence interval with some confidence level c . Keeping with the same ideas as the previous section, we need to first look for the area of the standard normal distribution that corresponds to the confidence level c :



In general, c can take different values. Therefore, the exact values of $\pm z_c$ will change depending on how large or small you would like to make the area. This number z_c has a special name: it is called the *critical value* for confidence level c .

Critical value

For a confidence level c , the **critical value** z_c is the positive number satisfying

$$P(-z_c < Z < z_c) = c,$$

where

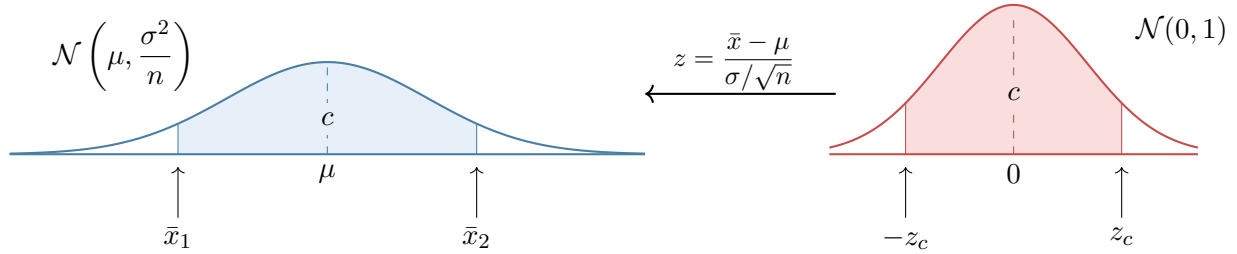
$$Z \sim \mathcal{N}(0, 1).$$

The precise value of z_c can presumably be found from a z -table or a calculator. However, in practice, we often compute confidence intervals for common levels of confidence such as 90% or 99%. For convenience, some common critical values are listed in the table below.

Level of confidence c	Critical value z_c
0.70, or 70%	1.04
0.75, or 75%	1.15
0.80, or 80%	1.28
0.85, or 85%	1.44
0.90, or 90%	1.645
0.95, or 95%	1.96
0.98, or 98%	2.33
0.99, or 99%	2.58

12.4.2 The general confidence interval

Suppose now that we want to construct a confidence interval with some confidence level c . Keeping with the same ideas as the previous section, we need to first look for the central area c of the standard normal distribution. Once we have found the critical value that corresponds to c , we can solve for the corresponding endpoints of $\bar{X}$:



Solving

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

for $\bar{x}_1$ and $\bar{x}_2$, we find that the endpoints in the sampling distribution are:

$$\mu - \frac{z_c \sigma}{\sqrt{n}} \quad \text{and} \quad \mu + \frac{z_c \sigma}{\sqrt{n}}.$$

Therefore, by construction we have found an interval centered at μ that contains $\bar{X}$ with probability c . In symbols:

$$P\left(\mu - \frac{z_c \sigma}{\sqrt{n}} < \bar{X} < \mu + \frac{z_c \sigma}{\sqrt{n}}\right) = c.$$

By the symmetry of being “nearby”, we can take this interval and center a random interval around the random variable $\bar{X}$, and this interval will therefore contain μ with the same probability:

$$P\left(\bar{X} - \frac{z_c \sigma}{\sqrt{n}} < \mu < \bar{X} + \frac{z_c \sigma}{\sqrt{n}}\right) = c.$$

This gives the general confidence interval formula.

Confidence interval for μ with known σ

Given a sample mean $\bar{x}$, known population standard deviation σ , sample size n , and confidence level c , the confidence interval for μ is:

$$\left[\bar{x} - z_c \frac{\sigma}{\sqrt{n}}, \bar{x} + z_c \frac{\sigma}{\sqrt{n}} \right].$$

12.4.3 Margin of error

Notice that the quantity

$$E = z_c \frac{\sigma}{\sqrt{n}}$$

appears in the general formula for confidence intervals. This distance E tells us the acceptable error bound of our estimate. For that reason, E is called the **margin of error**. Using this terminology, a confidence interval can be written simply as follows.

Margin of error

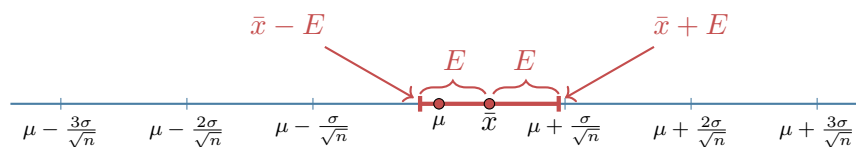
For a confidence interval for μ with known σ , the margin of error is

$$E = z_c \frac{\sigma}{\sqrt{n}}.$$

The confidence interval is

$$[\bar{x} - E, \bar{x} + E].$$

Geometrically, the length E is one-half of the total length of the confidence interval:



Therefore, to find a confidence interval for μ with known σ , we follow these steps:

Step 1: Use the confidence level c to find the critical value z_c .

Step 2: Calculate the margin of error:

$$E = z_c \frac{\sigma}{\sqrt{n}}.$$

Step 3: Write the interval:

$$[\bar{x} - E, \bar{x} + E].$$

Lecture 13

More on Estimation

In Lecture 12 we introduced the concept of *estimation*, and used it to guess the value of the population mean μ given some sample mean $\bar{x}$. In doing so, we deliberately made the assumption that σ , the true population standard deviation, was somehow already known. This helped simplify the concepts involved, which was very helpful at the time. However, we can imagine that in the real world, it is probably the case that if we don't know μ , then we don't know σ either. In this lecture we will explore the idea of estimation without assuming that σ is known. We will do this in two settings: estimating μ as in Lecture 12, and estimating the success probability p in a binomial experiment.

13.1 A Review of Estimation

Last lecture we introduced the concept of *confidence intervals* to estimate the true population mean μ . The main idea was to take a sample mean $\bar{x}$ and to use the Central Limit Theorem to construct the sampling distribution of $\bar{X}$, $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$. This distribution has the true population mean μ at the very center, and therefore we can use it to relate $\bar{x}$ to μ . From here, we created small intervals around $\bar{x}$ which, by construction, would contain the true mean a certain amount of the time.

A common real-world use of this style of estimation appears in *quality control*. For the purposes of illustration, suppose that there is a factory that produces a large number of screws. In order to fit into whatever they are designed to fit into, these screws need to be about 20 mm long, with a 0.1 mm tolerance either way. This means that, for instance, a screw of 19.9 mm is acceptable, whereas a screw of length 19.89 mm is not.

In principle, we could guarantee a high quality of the screws by inspecting every single screw that the factory produces and measuring its length. That way, we could add up all of these numbers and divide by the total number of screws to get their true average length. However, in practice this would be far too time-consuming, and practically impossible. So, in a situation like this we can instead take some sample data and merely *estimate* the true average length of the screws. If this number were within the range between 19.9 and 20.1, then the screws are acceptable. This is the idea of *quality control*.

Quality control

Quality control uses sample data to decide whether a production process is behaving acceptably. Often it is unrealistic to test every single product, so it is easier and cheaper to use a statistical estimate instead.

13.1.1 Review: estimating μ with known σ

Keeping with the example of the factory, suppose that long-term studies tell us that the population standard deviation is known, with the exact value:

$$\sigma = 1.2 \text{ mm.}$$

Let's also suppose that quality control takes a simple random sample of $n = 50$ screws and finds a sample mean of $\bar{x} = 19.98$ mm.

We will now construct a 95% confidence interval for the true mean length μ . From Lecture 12, the confidence interval for μ with known σ can be calculated from the margin of error E . If we center the interval at an observed value of $\bar{x}$, the confidence interval will be:

$$\left[\bar{x} - z_c \frac{\sigma}{\sqrt{n}}, \bar{x} + z_c \frac{\sigma}{\sqrt{n}} \right].$$

For a 95% confidence interval, we know from the table in Lecture 12 that $z_c = 1.96$. So, the margin of error can be calculated:

$$E = z_c \frac{\sigma}{\sqrt{n}} = 1.96 \cdot \frac{1.2}{\sqrt{50}} \approx 0.333.$$

Therefore, the 95% confidence interval centered around our observed sample mean $\bar{x} = 19.98$ will be:

$$\text{CI} = [19.98 - 0.333, 19.98 + 0.333] = [19.647, 20.313].$$

Interpretation

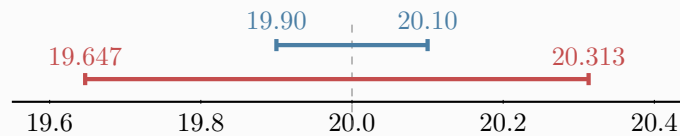
According to the setup, the allowed range for the average length is:

$$[19.90, 20.10].$$

But the 95% confidence interval that we found is

$$[19.647, 20.313].$$

This interval is much wider than the required engineering tolerance:

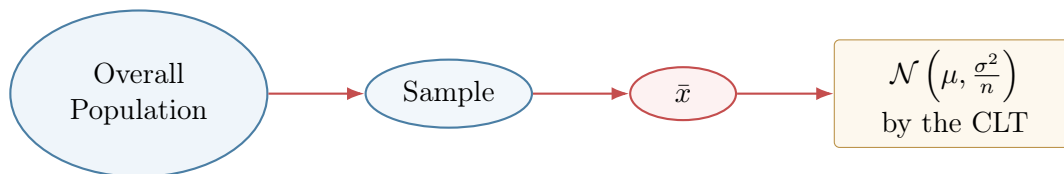


So, the sample does *not* provide strong evidence that the true mean is within the required tolerance. However, this shouldn't mean that we automatically shut down the factory. Instead, we have to accept that the sample data is not convincing enough either way: we don't have enough evidence to conclude that the true mean is outside of the desired $20 \text{ mm} \pm 0.1 \text{ mm}$, and we also don't have enough evidence to conclude that the true mean is inside those bounds either.

13.2 Estimation with Unknown σ

13.2.1 Where the known σ method breaks

In the previous example, we used the following workflow:



From there, we estimated μ via a margin of error calculated using:

$$E = z_c \frac{\sigma}{\sqrt{n}}.$$

However, in many real-world situations we simply *don't know* σ . This means that we cannot plug that number into the formula above, and therefore we can no longer estimate μ using this method. Instead, we need to replace σ with something that we do know, and then evaluate the consequences of this replacement.

13.2.2 Replacing σ with s

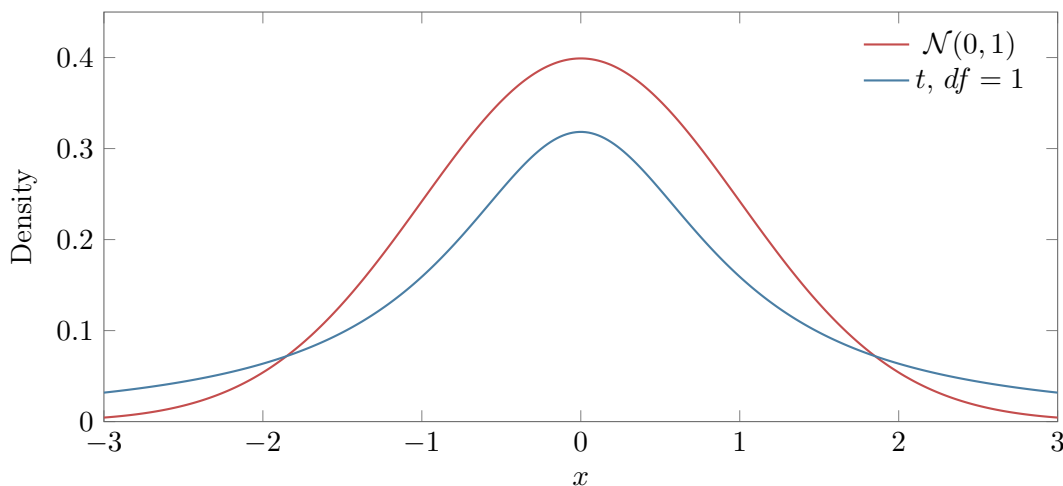
If we don't know what the true population standard deviation σ is, then it seems that the next best thing is to replace it with the sample standard deviation:

$$s = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n - 1}}.$$

However, if we make this replacement, then we are introducing more uncertainty than we had before. Therefore, we have to be careful about which probability distribution we use to derive our estimates. As a matter of fact, it is no longer appropriate to use the standard normal distribution. Instead, we need to use a modification known as Student's t -distribution. We will now explore exactly what these t -distributions look like.

13.2.3 Student's t -distributions

Student's t -distributions are related to the standard normal distribution. The only major difference is that more of the area is distributed in the tails. A simple comparison is depicted below:



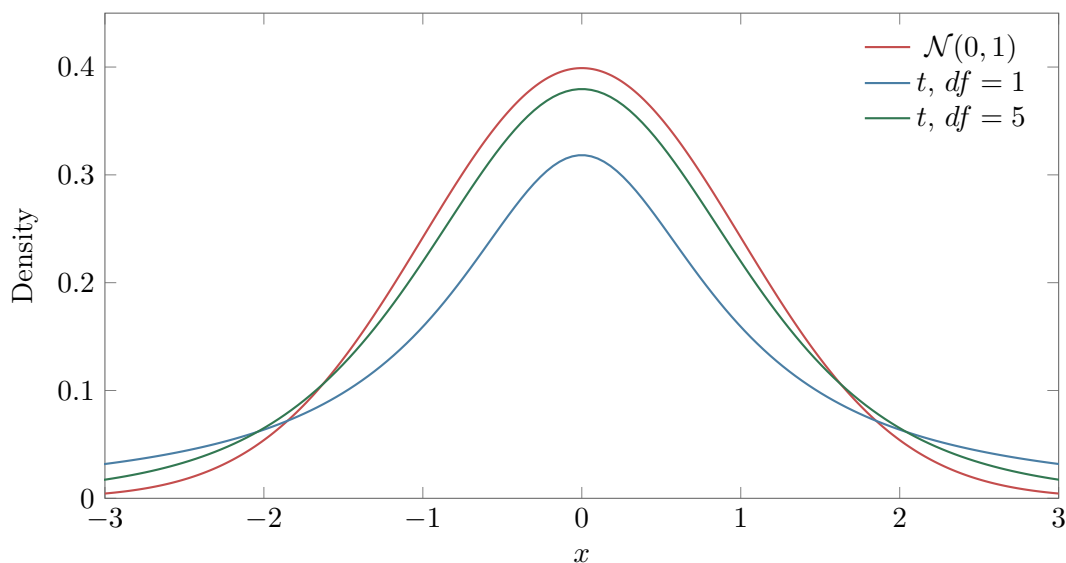
Here the standard normal distribution is in red, and the t -distribution is in blue. Notice that the blue curve is still symmetric and still centered at 0. In fact, the only clear difference is in the tails – they are *fatter*.¹ This is for a good reason: when σ is unknown we have more uncertainty, and therefore confidence intervals built off of this distribution should be wider than their normal-based relatives.

Notice in the graph above that there is an extra term “ df ” in the label of the t -distribution. In fact, this is an extra input that is required for the t -distribution to be fully described. This extra piece of information is known as the *degrees of freedom* (hence the letters df), and it is defined simply as:

$$df = n - 1.$$

¹The technical term for this is *kurtosis*, which is a measure of how “heavy” a distribution’s tails are.

The idea here is that this number will increase as the sample size does. As a sample size increases, we would generally expect that the sample standard deviation s becomes less influenced by random noise, and therefore it becomes a better estimate of the true population parameter σ . If s gets closer to σ as n increases, then we would expect that the uncertainty built into the tails of the t -distribution should decrease. As n gets very large, we should have a sample standard deviation s very close to σ , and therefore we expect a t -distribution that looks very similar to the standard normal distribution $\mathcal{N}(0, 1)$. In fact, that is exactly what happens: as df increases, a t -distribution looks more like the standard normal distribution:



As df gets bigger and bigger, the resulting t -distribution will look more and more like a normal distribution. This gives us another way to describe the standard normal distribution: it is a limit of standard t -distributions as $df \rightarrow \infty$.

Properties of Student's t -distributions

Student's t -distributions have the following properties.

1. They are symmetric about 0.
2. The total area under the curve is 1.
3. They have larger tails than $\mathcal{N}(0, 1)$.
4. They require one extra input:

$$df = n - 1.$$

5. As df gets larger, the t -distribution gets closer to $\mathcal{N}(0, 1)$.

13.2.4 Critical values for t

Student's t -statistic

Assume that the original population is normally distributed with mean μ . For samples of size n , with sample mean $\bar{X}$ and sample standard deviation S ,^a the statistic

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has a Student's t -distribution with

$$df = n - 1$$

degrees of freedom.

^aHere we have upgraded the sample standard deviation into a random variable in its own right.

The expression looks almost exactly like the z -score formula, except that we have replaced σ with s . Once have an observed sample, the formula changes:

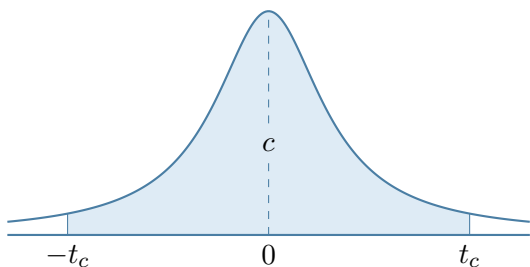
$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \xrightarrow{\text{Unknown } \sigma} t = \frac{\bar{x} - \mu}{s/\sqrt{n}}.$$

Recall that when σ is known, we used a critical value z_c from the standard normal distribution. Geometrically, this corresponded to the area in the standard normal distribution that was exactly c . We then constructed confidence intervals to estimate μ by mapping this region back into the sampling distribution $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ using the z -score formula.

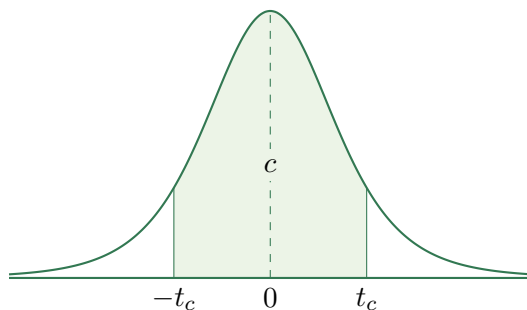
In contrast, when σ is unknown, we instead use a critical value from Student's t -distribution. In analogy to the critical value z_c , we will write the critical values in the t -distribution as t_c . Geometrically, these critical values play exactly the same role: they provide the endpoints of a region in the t -distribution that bounds an area of exactly c , centered at the origin. Since t -distributions depend on the degrees of freedom df , the exact value of t_c will depend on two pieces of information:

1. the confidence level c , and
2. the degrees of freedom $df = n - 1$.

For instance, consider the t -distributions for $df = 1$ and $df = 5$ drawn previously. Here, since the two curves have different shapes, the critical values t_c will change:



The area c when $df = 1$



The area c when $df = 5$

As you can see, the critical values for t_c will be larger when df is smaller. A table for common values of t_c is written below. Notice that it is much more complicated than the table for its counterpart z_c :

df	80%	90%	95%	98%	99%	99.5%
1	3.078	6.314	12.706	31.821	63.657	127.321
2	1.886	2.920	4.303	6.965	9.925	14.089
3	1.638	2.353	3.182	4.541	5.841	7.453
4	1.533	2.132	2.776	3.747	4.604	5.598
5	1.476	2.015	2.571	3.365	4.032	4.773
6	1.440	1.943	2.447	3.143	3.707	4.317
7	1.415	1.895	2.365	2.998	3.499	4.029
8	1.397	1.860	2.306	2.896	3.355	3.833
9	1.383	1.833	2.262	2.821	3.250	3.690
10	1.372	1.812	2.228	2.764	3.169	3.581
12	1.356	1.782	2.179	2.681	3.055	3.428
15	1.341	1.753	2.131	2.602	2.947	3.286
20	1.325	1.725	2.086	2.528	2.845	3.153
25	1.316	1.708	2.060	2.485	2.787	3.078
30	1.310	1.697	2.042	2.457	2.750	3.030
40	1.303	1.684	2.021	2.423	2.704	2.971
60	1.296	1.671	2.000	2.390	2.660	2.915
120	1.289	1.658	1.980	2.358	2.617	2.860
∞	1.282	1.645	1.960	2.326	2.576	2.807

Here, each column corresponds to a different common value of the confidence level c . Each row gives the value of t_c depending on some common degrees of freedom. The bottom row represents the case in which df gets arbitrarily large. In this case, the value of t_c will get closer and closer to the value of z_c , eventually approaching it in the limit.

13.2.5 Confidence interval with unknown σ

Now that we have access to critical values t_c , we follow the usual derivation of the margin of error to describe confidence intervals around observed values $\bar{x}$. In this case we have to use the t -statistic:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

instead of the usual z -score formula. Using the critical values $\pm t_c$ and solving gives the margin of error:

$$E = t_c \frac{s}{\sqrt{n}}.$$

Notice the clear similarity to the case of known σ – here we have simply replaced σ with s and have used a t -distribution's critical value instead of the usual z_c . We can describe confidence intervals around observed sample means $\bar{x}$ as follows.

Confidence interval for μ with unknown σ

If the population standard deviation σ is unknown, we use the sample standard deviation s . The confidence interval for μ becomes:

$$[\bar{x} - E, \bar{x} + E] = \left[\bar{x} - t_c \frac{s}{\sqrt{n}}, \bar{x} + t_c \frac{s}{\sqrt{n}} \right],$$

where $df = n - 1$.

The key differences between the known and unknown cases are summarized below.

	σ known	σ unknown
Score	$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$	$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$
Critical value	z_c	t_c
Margin of error	$E = z_c \frac{\sigma}{\sqrt{n}}$	$E = t_c \frac{s}{\sqrt{n}}$
Confidence interval	$[\bar{x} - E, \bar{x} + E]$	$[\bar{x} - E, \bar{x} + E]$

13.2.6 Worked example: screws with unknown σ

Let's return to the example of the factory from Section 1. Suppose now that the factory does *not* know the true population standard deviation σ . Instead, a sample of 50 screws gives:

$$n = 50, \quad \bar{x} = 19.98, \quad s = 1.1 \text{ mm.}$$

If we now want to construct a 95% confidence interval for the true mean length μ , we must use s instead of σ , and we must use the t -statistic instead of the z -score. Since $n = 50$, the degrees of freedom are $df = n - 1 = 49$. From a t -table, for a 95% confidence interval with $df = 49$, the critical value is approximately:

$$t_c \approx 2.010.$$

Using $s = 1.1$ and the above, we compute the margin of error to be:

$$E = t_c \frac{s}{\sqrt{n}} = 2.010 \cdot \frac{1.1}{\sqrt{50}} \approx 0.313.$$

Therefore, the 95% confidence interval centered around our observed value of $\bar{x}$ will be:

$$\text{CI} = [19.98 - 0.313, 19.98 + 0.313] = [19.667, 20.293].$$

Here, we are met with the same interpretation as the previous example.

13.3 Estimating a Binomial Proportion

13.3.1 Defining a Sample Proportion

Suppose that we witness part of a binomial experiment. We happen to observe n trials, and we count r successes. Because of random variation in the sample, this is not enough information to deduce the true value of p , the probability of success. However, we can still use this information to estimate the value of p . As with the previous section and Lecture 12, we will create a confidence interval around sample data, which will give us an appropriate error bound around our estimate. Our first goal is to obtain a meaningful sample statistic.

Sample proportion

If we observe r successes in n binomial trials, then the sample proportion is defined as:

$$\hat{p} = \frac{r}{n}.$$

From this, we also define $\hat{q} = 1 - \hat{p}$.

The quantity $\hat{p}$ is going to be our point estimate for the true success probability p .

13.3.2 Defining a Sampling Distribution

When discussing the normal approximation to a binomial distribution in Lecture 11, we mentioned that the usual binomial random variable X that counts the number of successes within a binomial experiment can be broken down into a sum of Y_i , where Y_i is a random variable defined for each individual trial of the experiment:

$$Y_i = \begin{cases} 1 & \text{if trial } i \text{ is a success,} \\ 0 & \text{if trial } i \text{ is a failure.} \end{cases}$$

In symbols, we saw in Lecture 11 that:

$$X = Y_1 + Y_2 + \cdots + Y_n$$

and therefore dividing X by n gave a quantity that acts like a sample mean. We will now put this together with our sample proportion: we can introduce a random variable, written $\hat{P}$, if we define it as:

$$\hat{P} = \frac{X}{n} = \frac{Y_1 + Y_2 + \cdots + Y_n}{n}.$$

In fact, this is exactly what we did in Lecture 11 without giving it a name. Written this way, we can see that $\hat{P}$ is a random variable whose observed values $\hat{p}$ act like sample means taken from a sample of size n – the observed value r above is the sum of the outcomes of the n individual random variables Y_i , and then we divide by n to get the mean.

We saw in Lecture 11 that each one of these individual random variables Y_i has:

$$E(Y_i) = p \quad \text{and} \quad \text{Var}(Y_i) = pq.$$

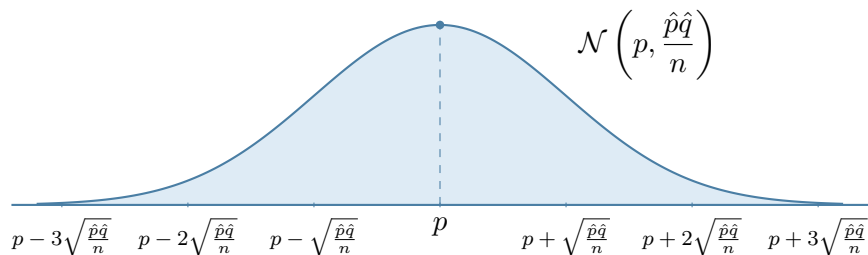
The Central Limit Theorem tells us that the collection of all sample means follows an approximate normal distribution in which the center is the mean of the original variable from which we form the sample, and the new variance is the variance of the original variable divided by n , the sample size. In our case, the Central Limit Theorem tells us that $\hat{P}$ follows an approximate normal distribution with parameters:

$$\mathcal{N}\left(p, \frac{pq}{n}\right).$$

This appears to give us precisely what we need: we have just confirmed that our sample proportion has an approximate normal distribution, and the true mean is at the center. From here, it seems that we can start to create some confidence intervals. However, there is one fundamental issue: we are trying to estimate p , the true probability of success. If we don't know what the exact value

of p is, then we *also* don't know what the value of $\frac{pq}{n}$ is either! In other words, if we were to try and continue, and create some sort of margin of error, how would we actually calculate it, without already knowing the value of p ?

Here, we have to accept that the true sampling distribution for $\hat{P}$ cannot serve our needs. Instead, we have to do the next-best thing, which is to replace the variance term with point estimates. In practice, we replace p and q by the observed values $\hat{p}$ and $\hat{q}$, giving the following approximation:



This will be the setting of our estimation.

Sampling distribution for $\hat{P}$

For a binomial experiment, the sample proportion is approximately normally distributed:

$$\hat{P} \approx \mathcal{N}\left(p, \frac{\hat{p}\hat{q}}{n}\right).$$

This approximation is reasonable when

$$n\hat{p} > 5 \quad \text{and} \quad n\hat{q} > 5.$$

13.3.3 Confidence intervals for p

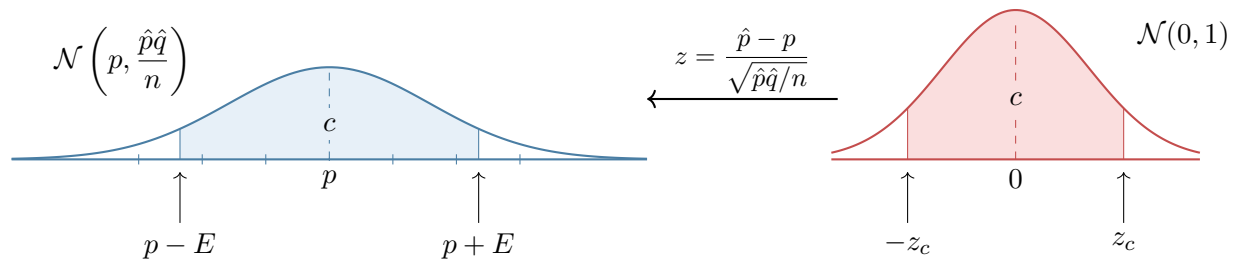
Now we construct confidence intervals exactly as before. The estimated standard error is

$$\sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

As such, the formula for the z -score becomes:

$$z = \frac{\hat{p} - p}{\sqrt{\hat{p}\hat{q}/n}}.$$

To derive a confidence interval with confidence level c , we find the critical value z_c in $\mathcal{N}(0, 1)$ that gives $P(-z_c < Z < z_c) = c$. Using the z -score formula in reverse then maps this region back into the sampling distribution:



The margin of error E is given by z_c multiplied by the standard error. In this case:

$$E = z_c \sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

Therefore, confidence intervals around observed values $\hat{p}$ can be defined as follows.

Confidence interval for a binomial proportion

Given r successes in n trials,

$$\hat{p} = \frac{r}{n}, \quad \hat{q} = 1 - \hat{p}.$$

The confidence interval for the true success probability p is

$$\left[\hat{p} - z_c \sqrt{\frac{\hat{p}\hat{q}}{n}}, \hat{p} + z_c \sqrt{\frac{\hat{p}\hat{q}}{n}} \right].$$

13.3.4 The method

To estimate a binomial success probability p :

Step 1: Calculate

$$\hat{p} = \frac{r}{n}, \quad \hat{q} = 1 - \hat{p}.$$

Step 2: Check that the normal approximation is reasonable:

$$n\hat{p} > 5 \quad \text{and} \quad n\hat{q} > 5.$$

Step 3: Find the critical value z_c for the chosen confidence level.

Step 4: Calculate the margin of error:

$$E = z_c \sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

Step 5: Write the confidence interval:

$$[\hat{p} - E, \hat{p} + E].$$

13.4 Worked Example: Basketball Scouting

Suppose that a basketball team only wants to hire players whose true free-throw success rate is at least 70%. In symbols, the team wants evidence that $p \geq 0.70$.

You scout a player and watch him shoot some free throws. You cannot observe him all day, so you take a sample of his performance and decide to compute a 95% confidence interval to estimate his true shooting success rate p .

13.4.1 Example 1: 65 baskets in 95 attempts

Suppose the player takes $n = 95$ shots, and scores $r = 65$ of them. Then our point estimate of the success probability will be

$$\hat{p} = \frac{65}{95} \approx 0.684,$$

and therefore $\hat{q} = 1 - \hat{p} \approx 0.316$. Before going any further, we should first confirm the approximation conditions. We have:

$$n\hat{p} = 95(0.684) \approx 65, \quad \text{and} \quad n\hat{q} = 95(0.316) \approx 30.$$

These are both greater than 5, so the normal approximation is reasonable. For a 95% confidence interval, we have that $z_c = 1.96$. Therefore, we calculate the margin of error to be:

$$E = z_c \sqrt{\frac{\hat{p}\hat{q}}{n}} = 1.96 \sqrt{\frac{(0.684)(0.316)}{95}} \approx 0.093.$$

The 95% confidence interval around our observed value $\hat{p} = 0.684$ is therefore:

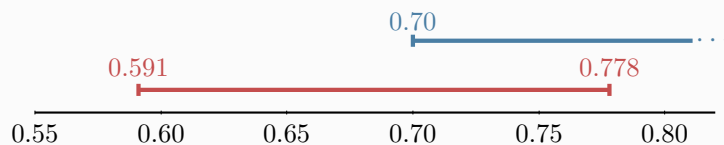
$$\text{CI} \approx [0.684 - 0.093, 0.684 + 0.093] = [0.591, 0.778].$$

Interpretation

According to the setup, we are looking for a desired threshold of $p = 0.70$. However, our confidence interval is

$$[0.591, 0.778].$$

This interval straddles the threshold:



Therefore, we have a similar situation to that of Section 1.1 – we cannot conclusively say that the player's true shooting rate is above or below 70%. Simply put, this sample does not provide enough convincing evidence either way.

13.4.2 Example 2: 773 baskets in 1000 attempts

After finding that your first observation was inconclusive, you go back to observe the player for a second time, this time for much longer. Suppose that now, you see the player take $n = 1000$ shots, and he scores $r = 773$ of them. Now, our point estimate of the success probability will be

$$\hat{p} = \frac{773}{1000} = 0.773,$$

and therefore $\hat{q} = 1 - \hat{p} = 0.227$. Again, we should first double-check that the approximation conditions hold. We have:

$$n\hat{p} = 1000(0.773) = 773, \quad \text{and} \quad n\hat{q} = 1000(0.227) = 227.$$

These are both greater than 5, so the normal approximation is reasonable. Again using the critical value $z_c = 1.96$, we calculate the margin of error:

$$E = z_c \sqrt{\frac{\hat{p}\hat{q}}{n}} = 1.96 \sqrt{\frac{(0.773)(0.227)}{1000}} \approx 0.026.$$

The 95% confidence interval around our observed value $\hat{p} = 0.773$ is then:

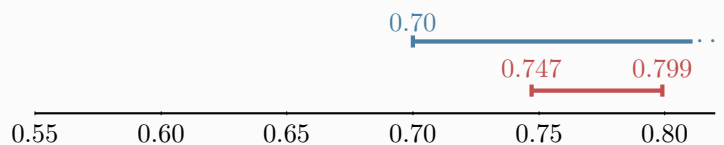
$$\text{CI} = [0.773 - 0.026, 0.773 + 0.026] = [0.747, 0.799].$$

Interpretation

In this case, our confidence interval is

$$[0.747, 0.799].$$

This entire interval now sits well above the desired threshold of 0.70:



Since this method captures the true value of p about 95% of the time, we have convincing evidence that the true value of p is above 0.70. So, this sample provides convincing evidence that the player meets the team's standard.

Part IV

Hypothesis Testing

Lecture 14

Introduction to Hypothesis Testing

In our recent lectures we have studied estimation, which uses sample data to guess roughly where a population parameter is. We will now continue our discussion of inferential statistics by performing *tests of claims*. Instead of estimating a parameter, we will now evaluate claims of a parameter's true value. Such a process is called a *hypothesis test*, and it is a powerful tool in inferential statistics. Throughout the next four lectures we will discuss this technique in detail. Today we will simply introduce the overall structure of a hypothesis test.

14.1 Introduction

Suppose that you are on a walk, and you come across two groups of people. One of these groups looks dangerous and suspicious, and the other is a group of trustworthy-looking nuns and children:



A suspicious group



A trustworthy group

You walk over to the shady-looking group first. They hand you a coin and propose the following game:

Keep flipping the coin. Every time you land on heads, we will give you \$2. But for every tails you land, you have to give us \$1.

At first this feels generous, so you decide to play. But after 30 flips, you notice that you only landed on heads 3 times, and got 27 tails. The net total is that you have to pay out \$21, and you go about your business.

Feeling suspicious, you decide to do some calculations: since coin flips are a binomial experiment, you use the formula:

$$P(X = 3) = C_{30,3}(0.5)^3(0.5)^{27} \approx 3.78 \times 10^{-6},$$

where here you assume that the coin is fair with $p = q = 0.5$. According to the above, you notice that the probability of getting 3 heads in 30 attempts is awfully small, and therefore you conclude that probably the coin was a scam, with p much lower than 50%.

You now walk over to the trustworthy-looking group, and start talking to them. Weirdly, this second group offers you the exact same game:

Keep flipping the coin. For every head, we will give you \$2. But for every tail, you have to give us \$1.

Against your better judgement, you decide to play again for another 30 coin flips. In this case, you manage to get 12 heads and 18 tails, and therefore you win \$6. Again assuming that the coin is fair, you calculate:

$$P(X = 12) = C_{30,12}(0.5)^{12}(0.5)^{18} = C_{30,12}(0.5)^{30} \approx 0.081.$$

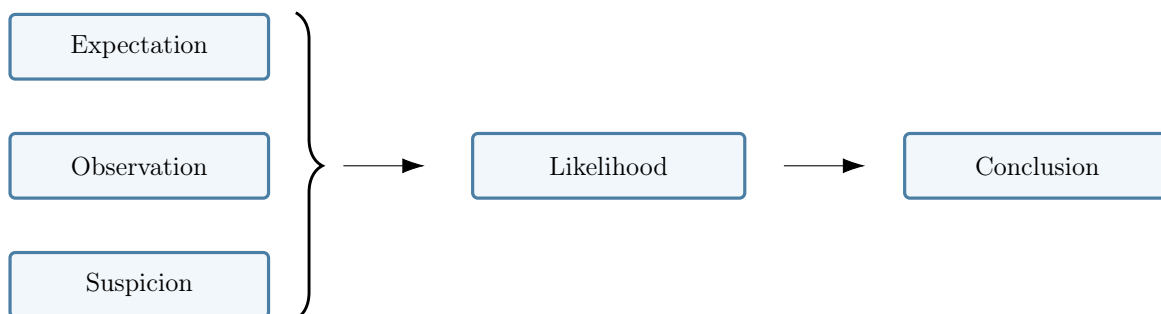
In fact, this is not too unreasonable, given that the chance of getting 15 heads in 30 flips is only about 14%. From this information alone, there is no good reason to think that the second coin was unfair.

14.1.1 The informal structure of a statistical test

We have just seen the informal idea behind a very important statistical procedure known as *hypothesis testing*. In both situations we were presented with a system and challenged an assumption that we were working with a fair coin, based on some observations that we were able to make. In both cases we made some sort of probability calculation, and drew two different conclusions:

1. in the case of the suspicious coin, we concluded that our random observation was *extremely unlikely*, which cast doubt on the expectation that $p = 0.5$, causing us to reject the idea of a fair coin, whereas
2. in the case of the nun's coin, our random observation was consistent with the probability $p = 0.5$, therefore giving us no reason to reject the fairness of the coin.

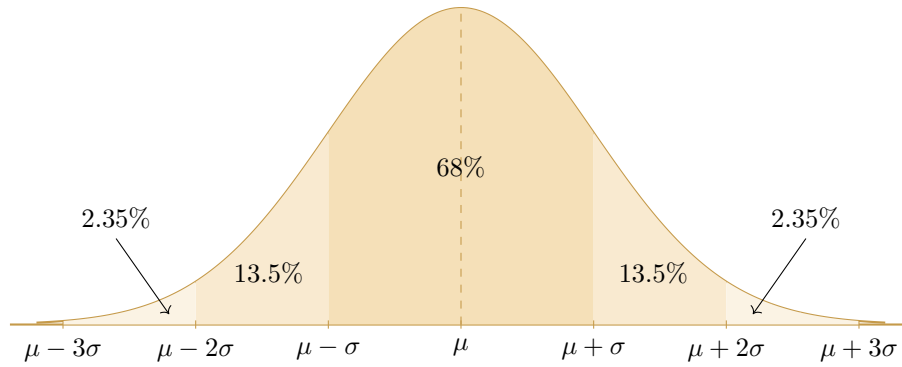
Lots of hypothesis tests follow this general pattern: we have an expectation, we make an observation which drives a suspicion, and then we perform a probability calculation that informs us of how reasonable it would be to reject our original assumption. This conceptual workflow can be summarized as follows:



Throughout the next few lectures, we will explore this structure in various contexts. For now, we will simply describe the general structure. In order to do that, we will first recall some basic properties of the normal distribution.

14.1.2 A Brief Review of the Normal Distribution

Recall that a normal distribution is defined by two parameters: the population mean μ and the population variance σ^2 . When we write $X \sim \mathcal{N}(\mu, \sigma^2)$, the x -axis of the distribution represents the possible values of the random variable X . Since this is a continuous random variable, the probability of finding X in some region is given by the area under the graph. The shape of the graph therefore encodes the likelihood of finding X nearby. Since a normal distribution will peak around its centre and then run off quickly into its tails, we would expect that randomly selecting a value of X will probably land us close to μ . In fact, based on our discussion from Lecture 10, there is about a 68% chance of landing within one standard deviation of μ :

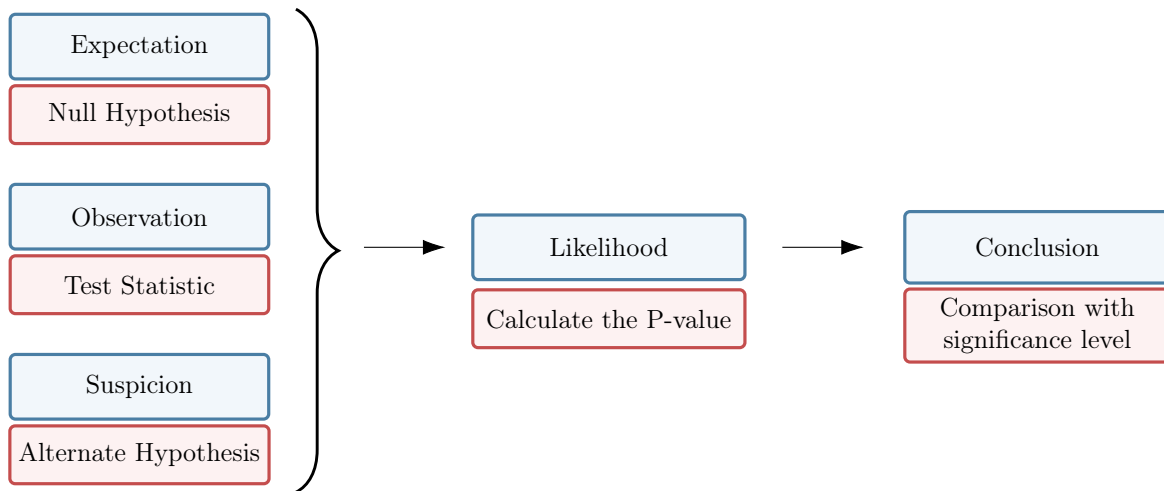


In other words, if we were to keep randomly selecting a value of X , then this value would be in the interval $[\mu - \sigma, \mu + \sigma]$ about 68% of the time. In contrast, if we were to randomly select a value of X , then it would be very unlikely that the value is far from the centre. This observation underpins the concept of hypothesis tests.

14.2 Hypothesis Testing

14.2.1 The general layout of a statistical test

In the introduction, we described a general workflow for performing a statistical test. Written below is the same workflow, decorated with the correct technical terms.



We will now explain these terms one-by-one.

Important Terminology

- **Null hypothesis (H_0):** the statement that we want to test.
- **Alternate hypothesis (H_1):** the competing statement that reflects our suspicion.
- **Test statistic:** the numerical information obtained from sample data.
- **P-value:** the probability of getting a result at least as extreme as the observation, assuming H_0 is true.
- **Significance level (α):** the cut-off probability that tells us what counts as “too unlikely”.

14.2.2 The P-value

According to our definition above, the P-value of a hypothesis test does not simply calculate the probability of an outcome happening. Instead, the P-value calculates the chances of observing our sample data *or worse* whilst assuming the null hypothesis is true. We saw prototypes of this idea in Section 1, where we performed binomial calculations for the two coins. In fact, these probability calculations do not yet represent P-values of those hypothesis tests. Instead, we would need to make more detailed calculations.

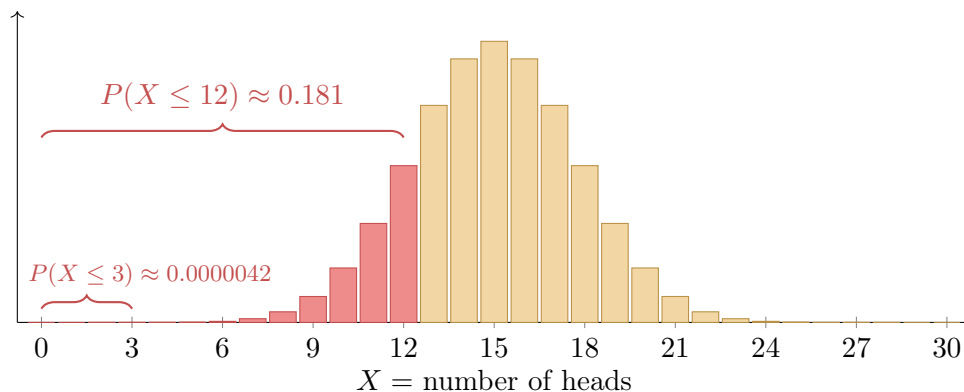
For the suspicious coin, the exact observation was $X = 3$, i.e. 3 heads. Here, the P-value of this observation assumes that the coin is fair, i.e. $p = 0.5$, and then calculates the chance of getting 3 heads *or less*:

$$P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) \approx 0.0000042.$$

For the second coin, we observed $X = 12$. Again, the P-value is calculated by computing the probability of getting 12 *or less* heads:

$$P(X \leq 12) = P(X = 0) + P(X = 1) + \dots + P(X = 12) \approx 0.181.$$

Visually, we are looking to compute the probability found in the left tail of the binomial distribution built from $n = 30$ and $p = 0.5$:



14.2.3 The Significance Level

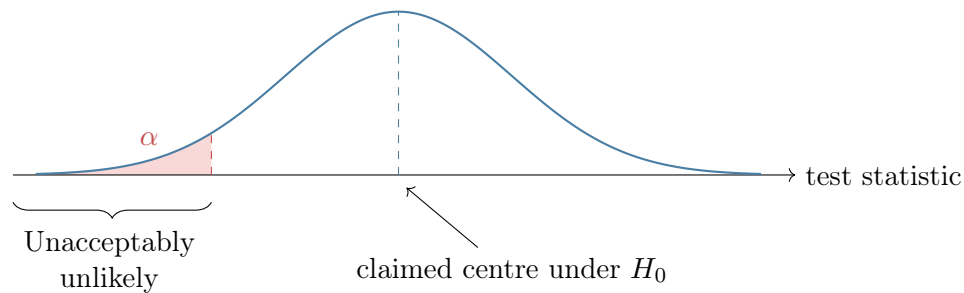
Hypothesis tests assume that the null hypothesis is actually *true*, and then they explore the consequences of this assumption. As we mentioned in Section 1.2, under most circumstances, we would

expect a randomly selected test statistic to land relatively close to the centre of its probability distribution – especially when it is normally distributed. If we were to obtain an observation very far into a tail, then this observation has a very low probability under the original model.

Of course, by the nature of randomness it is always *possible* to obtain an unusually extreme result by accident. Really, we need to ask:

How unlucky do we need to be before we start to think that something is wrong with our null hypothesis?

The significance level α serves to answer this question by giving a “cut-off” value for luck:



Once we have sample data in hand, the P-value calculates the chance of this particular sample *or worse* happening, assuming that the true centre of the data is given by H_0 . If this probability is unacceptably low, i.e. less than α , then we conclude that the value of the population parameter claimed by H_0 is probably wrong. The final judgement can be summarised as follows.

Decision Rule

- If the calculated P-value is **less than or equal to** α , then the observation is too unlikely under H_0 , so we **reject** H_0 .
- If the calculated P-value is **greater than** α , then the observation is not unlikely enough, so we **do not reject** H_0 .

In practice, we choose the significance level at the start of the test. Common values are

$$\alpha = 0.01, \quad \alpha = 0.02, \quad \alpha = 0.05.$$

Importantly, “do not reject H_0 ” does *not* mean that we have proved H_0 is true. It only means that the sample does not provide strong enough evidence against it.

14.2.4 Different types of tests

The alternate hypothesis H_1 is always a statement to the contrary of the null hypothesis. As a matter of fact, there are many ways to disagree with an “is” statement. Typically, the null hypothesis will say that the true value of a population parameter equals a particular claimed value. There are three ways that this could be wrong:

- the true value could be less than the claimed value,

- the true value could be more than the claimed value, or
- the true value could be simply *different from* the claimed value.

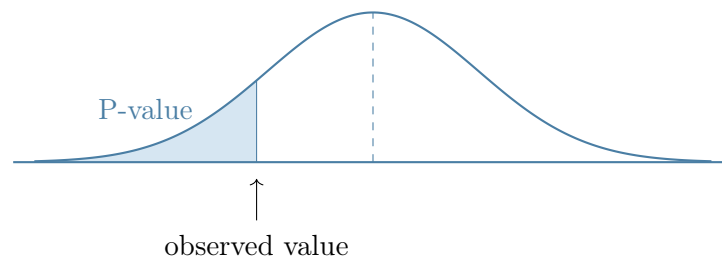
These are called left-tailed, right-tailed and two-tailed tests, respectively. We will now demonstrate how the alternate hypothesis tells us how to compute the correct P-value for a given test.

Left-tailed tests

A left-tailed test asks whether the true value is less than the value claimed by the null hypothesis. Formally, the test would start by writing something like:

$$H_0 : \text{true value} = \text{claimed value}, \quad H_1 : \text{true value} < \text{claimed value}.$$

As the name may suggest, in a left-tailed test we calculate the P-value of the observed data by taking the area to the *left*:

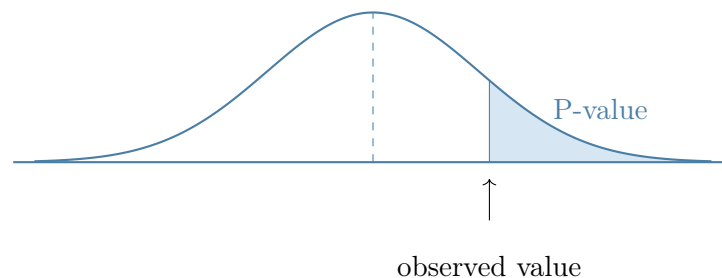


Right-tailed tests

In a right-tailed test, the alternate hypothesis asks whether the true value is greater than the value claimed by the null hypothesis:

$$H_0 : \text{true value} = \text{claimed value}, \quad H_1 : \text{true value} > \text{claimed value}.$$

In this case, the P-value is the area to the *right* of the observed sample's statistic:

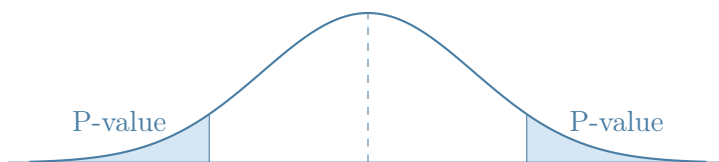


Two-tailed tests

For a two-tailed test, we ask whether the true value is merely different from the value claimed by the null hypothesis:

$$H_0 : \text{true value} = \text{claimed value}, \quad H_1 : \text{true value} \neq \text{claimed value}.$$

Recall that if two numbers a and b are different from each other, then either $a < b$ or $b < a$. As such, a two-tailed test computes the P-value by combining the probabilities from both tails at the same time:



14.2.5 Type 1 and Type 2 errors

By their very nature, hypothesis tests always contain some amount of uncertainty – if we find a P-value lower than the significance level, then this does not *prove* that the null hypothesis is false. Instead, it only means that the sample provides evidence against the null hypothesis. Similarly, if we do not reject the null hypothesis, this does not prove that it is genuinely true. Simply, it means that there is not enough evidence to reject it.

Because of this inherent uncertainty, it is always possible to draw the wrong conclusion by accident. There are two main ways that we could make a mistake when performing a hypothesis test. These are known as Type 1 and Type 2 errors.

Type 1 and Type 2 Errors

- A **Type 1 Error** happens when we reject H_0 even though H_0 is actually true.
- A **Type 2 Error** happens when we do not reject H_0 even though H_0 is actually false.

Type 1 errors can occur if we happen to find a statistically unlikely sample on the first try. For example, technically it is not *impossible* to flip a coin 30 times and to get 27 tails and 3 heads. However, it is extremely *unlikely*, and therefore we are compelled to reject the claim that such a coin is fair. However, when doing so, there is no guarantee that we are simply not living in an unlikely world. Rejecting the null hypothesis in a situation like that could be an example of a Type 1 error.

Type 2 errors occur when we take sample data that is consistent with the null hypothesis. In such a situation, we have no good reason to reject the null hypothesis, even if it might be genuinely false in reality. This can happen if our sample observation gives a P-value above α . To illustrate this, recall that the nun's coin gave 12 heads in 30 attempts. We concluded that this observation is reasonably consistent with a fair coin and therefore there is no reason to reject the claim that the coin is fair. However, what if the coin happened to have a probability of $P(H) = 0.4$ in reality? Our sample data would not exclude the possibility of a fair coin, even though it could be false.

There are four possible outcomes that can result from a hypothesis test:

	Do not reject H_0	Reject H_0
H_0 is true	Correct decision	Type 1 Error
H_0 is false	Type 2 Error	Correct decision

The significance level α controls the probability of making a Type 1 Error. For example, using $\alpha = 0.05$ means that the test is designed to reject a true null hypothesis about 5% of the time in repeated use.

14.3 A Sample Test: The Two Coins

We will now demonstrate how to express our coin flip examples as genuine hypothesis tests. For future convenience, we will actually phrase the experiments in terms of normal approximations using the random variable $\hat{P}$, which we saw in previous lectures when creating estimates for the true binomial probability p . In what follows we will simply state the P-values associated to our sample experiments, and we will defer the actual calculations of P-values to next lecture. For now, it is enough to recall that the random variable $\hat{P}$ follows an approximate normal distribution with parameters:

$$\hat{P} \sim \mathcal{N}\left(p, \frac{pq}{n}\right).$$

Importantly, the true binomial probability p lies at the centre of this distribution, and therefore we can visualize our null hypotheses as claims about the middle of this normal distribution. Our observed sample data will then eventually manifest as points on the x -axis of this distribution. In other words, the above normal distribution tells us how to *structure our luck*.

Before getting to visualizations of our observed data, we will first demonstrate how to properly phrase our tests.

14.3.1 The suspicious coin

We now return to the first coin.

Step 1: state the hypotheses.

We begin by assuming that the coin is fair, but we suspect that the probability of landing on Heads is actually lower than 0.5. Therefore, our null and alternate hypotheses become:

$$H_0 : p = 0.5, \quad \text{and} \quad H_1 : p < 0.5,$$

respectively. Notice that according to our alternate hypothesis, this is a left-tailed test.

Step 2: record the test statistic.

According to our sample data, we observed $r = 3$ heads in $n = 30$ coin flips. This gives a sample proportion of:

$$\hat{p} = \frac{r}{n} = \frac{3}{30} = 0.1.$$

Step 3: calculate the P-value.

Ordinarily here is where we could calculate the P-value directly. However, we will simply state the calculation according to our normal approximation:

$$\hat{P} \sim \mathcal{N}\left(0.5, \frac{(0.5)(0.5)}{30}\right).$$

Therefore,

$$P(\hat{P} \leq 0.1) \approx 0.0000134.$$

Note that we compute the values to the left of the observed $\hat{p}$, because we are working with a left-tailed test.

Step 4: compare with the significance level.

Using a level of significance $\alpha = 0.05$, we have, quite obviously:

$$0.0000059 < 0.05.$$

Therefore, we are compelled to reject the null hypothesis H_0 .

Interpretation

Our sample data of 3 heads in 30 coin flips provides extremely strong evidence against the hypothesis that the coin is fair. Instead, the data supports the suspicion that the probability of heads is less than 0.5.

14.3.2 The nun's coin

We will now perform the same test for the second coin.

Step 1: state the hypotheses.

Again, we are going to test whether the coin is biased against heads:

$$H_0 : p = 0.5, \quad \text{and} \quad H_1 : p < 0.5.$$

Step 2: record the test statistic.

In this case, we observed $r = 12$ heads in $n = 30$ flips, so

$$\hat{p} = \frac{r}{n} = \frac{12}{30} = 0.4.$$

Step 3: calculate the P-value.

Under the null hypothesis, we apply a continuity correction and calculate the P-value to be approximately

$$P(X \leq 12) \approx P(Z \leq -0.913) \approx 0.181.$$

Step 4: compare with the significance level.

Again using a significance level of $\alpha = 0.05$, we now see that:

$$0.181 > 0.05.$$

Therefore, we do not reject H_0 .

Interpretation

Our sample of 12 heads in 30 coin flips does not provide strong enough evidence to conclude that the true probability of heads is less than 0.5. This does not prove that the coin is fair; it only means that the observation is reasonably consistent with a fair coin.

14.3.3 The tests visualized

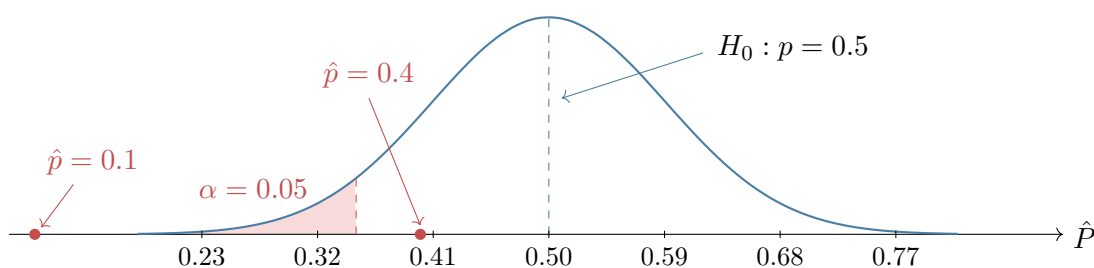
Under the null hypothesis, the sampling distribution of the observed proportion is approximately

$$\hat{P} \sim \mathcal{N}\left(0.5, \frac{(0.5)(0.5)}{30}\right).$$

The standard deviation of this distribution is

$$\sqrt{\frac{(0.5)(0.5)}{30}} \approx 0.091.$$

The suspicious coin produced $\hat{p} = 0.1$, which is extremely far into the left tail. The nun's coin produced $\hat{p} = 0.4$, which is below the centre but not especially extreme.



As you can see, the left-tailed P-value associated to the nun's coin will give an area greater than the threshold $\alpha = 0.05$, and therefore we have no good reason to reject the hypothesis that that coin is fair. However, in the case of the suspicious coin, the observed sample proportion $\hat{p} = 0.1$ is far into the left-tail of the distribution, meaning that it is extremely unlikely to occur on our first attempt at getting sample data. The area to the left of 0.1 is far smaller than 0.05, and therefore we rejected the hypothesis that the suspicious coin was fair.

Lecture 15

Hypothesis Testing the Mean

In Lecture 14, we introduced the general idea of *hypothesis testing*. We learned how to state a null hypothesis and an alternate hypothesis, how to interpret a P-value, and how to compare that P-value with a significance level. In this lecture, we will focus on the probability calculation at the centre of the test. We will study three important situations:

1. testing the population mean μ when σ is known,
2. testing the population mean μ when σ is unknown, and
3. testing the binomial success probability p .

As we will see, these three situations follow the same general pattern, albeit with a different probability calculation at their cores.

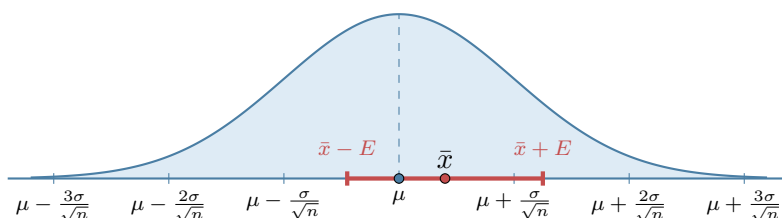
15.1 A Review of Hypothesis Testing

15.1.1 Estimation and hypothesis testing

When estimating an unknown population parameter, we take a sample and construct an interval of reasonable values around an estimate of it. For example, when estimating a population mean, we might construct a confidence interval centred at the observed sample mean:

$$[\bar{x} - E, \bar{x} + E].$$

Here, the true population mean μ is unknown. The idea is that the size of the interval is calculated so that the method captures μ a chosen percentage of the time: if we repeatedly took samples of size n and computed all of the confidence intervals around their sample means, then a certain proportion of these intervals would contain the actual μ . Overall, this construction uses the Central Limit Theorem, which says that the sample mean $\bar{X}$ admits its own probability distribution with mean μ and variance $\frac{\sigma^2}{n}$.¹ The $\bar{x}$ and its confidence interval then live in this distribution:



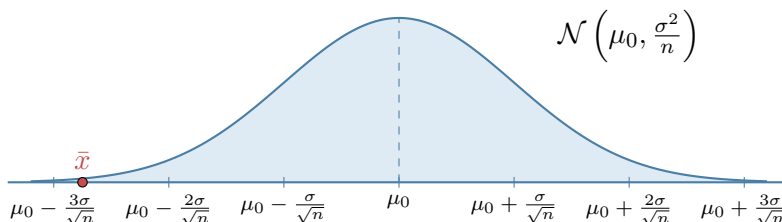
When performing hypothesis tests, we completely reverse our thought process. We begin by *assuming* a particular value of the population parameter and then ask whether the observed sample data would be reasonable under that assumption. For example, suppose that the null hypothesis claims

$$H_0 : \mu = \mu_0.$$

We *assume* that the real population mean is actually μ_0 , and then explore the consequences of this assumption. If H_0 were true, then the Central Limit Theorem says that the sampling distribution should be:

$$\bar{X} \approx \mathcal{N}\left(\mu_0, \frac{\sigma^2}{n}\right).$$

Therefore, any observed sample mean $\bar{x}$ should appear somewhere within this distribution. If we take a sample of real-world data, and the sample mean $\bar{x}$ lands extremely far into a tail, then the data is unlikely under H_0 :



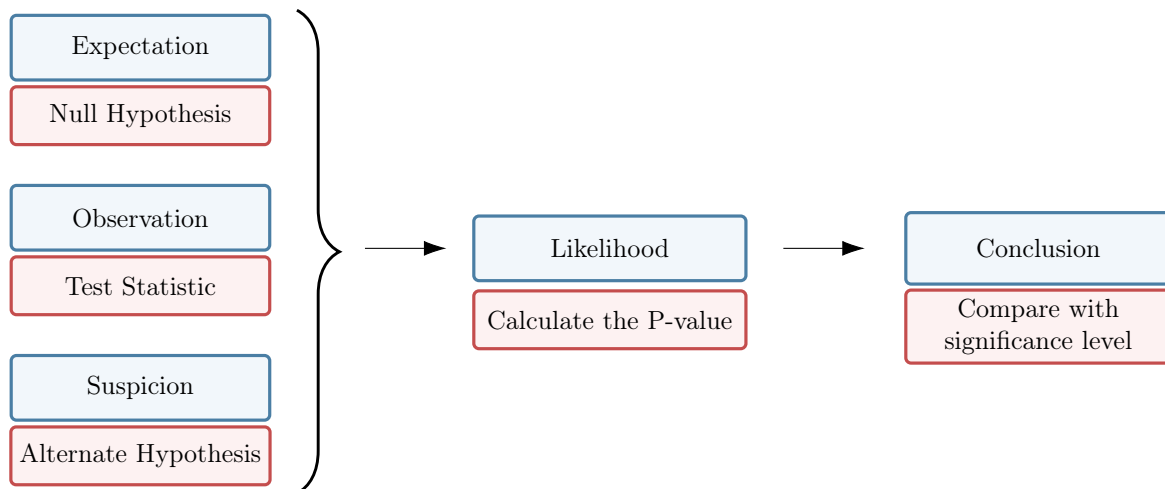
¹Technically, n should be sufficiently large unless the population is normally distributed. Usually, we take samples of size 30 or larger.

Although these two methods appear to be very similar, there are key differences:

- in **estimation**, the population parameter is unknown and we construct an interval of plausible values;
- in **hypothesis testing**, we assume a particular value of the population parameter and ask whether the observed data is plausible under that assumption.

15.1.2 The general layout of a statistical test

Recall the general layout of a statistical test from Lecture 14:



The first stage involves reading and interpreting the situation. We identify:

- the claim being tested;
- the suspicion against that claim;
- the relevant sample information.

The second stage involves the main probability calculation. We then take this value (called the *P*-value), and make a judgement: we compare the resulting P-value with the chosen significance level and decide whether the sample provides enough evidence to reject the null hypothesis.

Important Definitions

- **Null hypothesis (H_0):** the statement that we want to test.
- **Alternate hypothesis (H_1):** the competing statement that reflects our suspicion.
- **Test statistic:** the numerical information obtained from the sample data.
- **P-value:** the probability of obtaining the observed test statistic, or something more extreme, assuming that H_0 is true.
- **Significance level (α):** the cut-off probability that tells us what counts as unacceptably unlikely.

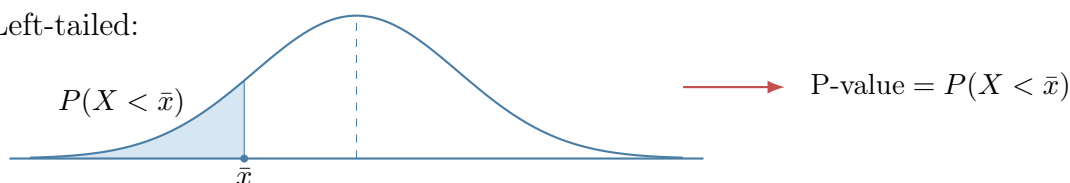
15.1.3 Different tails, different P-values

It is common for students to think that the P-value somehow tells us how likely the null hypothesis is. In fact, this is **not** true – instead, the P-value measures how unusual the observed sample (or something worse) would be *if* the null hypothesis were true. The alternate hypothesis tells us which outcomes count as “worse” than the observed sample data. There are three common forms of alternate hypotheses:

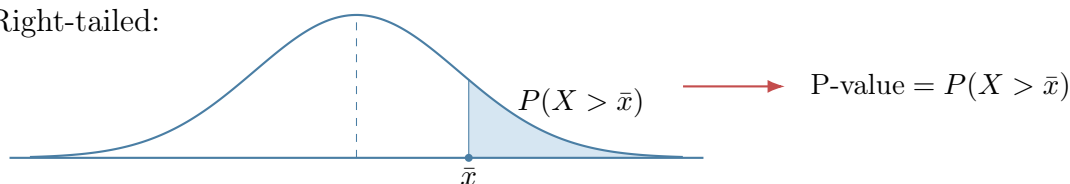
$$H_1 : \begin{cases} \text{true value} < \text{claimed value,} & \text{left-tailed test,} \\ \text{true value} > \text{claimed value,} & \text{right-tailed test,} \\ \text{true value} \neq \text{claimed value,} & \text{two-tailed test.} \end{cases}$$

According to the statement of the alternate hypothesis, the calculation of the P-value will change depending on which tail we are considering:

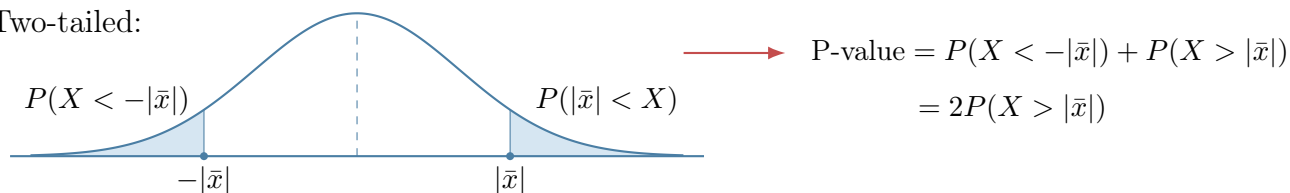
Left-tailed:



Right-tailed:



Two-tailed:



15.2 Testing μ with Known σ

15.2.1 The general idea

Suppose that we want to test a claim about the population mean μ , and suppose that the population standard deviation σ happens to be known. In this situation, the null hypothesis will take the form:

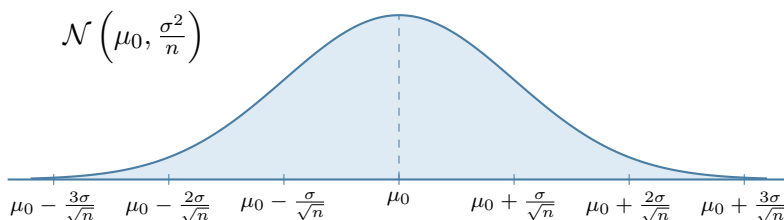
$$H_0 : \mu = \mu_0,$$

where here μ_0 is the claimed value of μ . Sample data will typically be provided, meaning that we have a definite value of the sample size n and the sample mean $\bar{x}$.

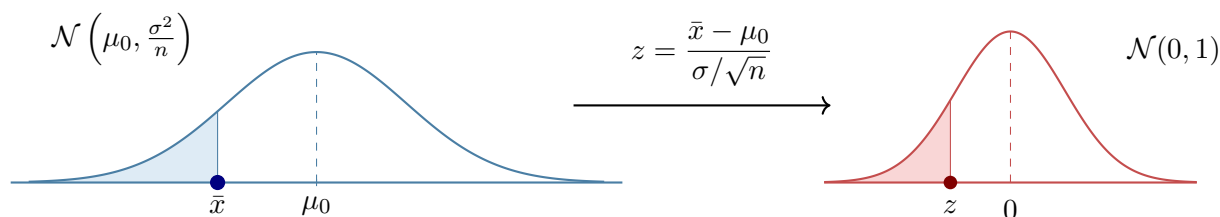
In order to calculate the relevant P-value, we assume that the null hypothesis is true, that is, we assume that the population mean is actually μ_0 . Under this assumption, the Central Limit Theorem tells us that the sampling distribution for the sample means will have the parameters:

$$\bar{X} \approx \mathcal{N}\left(\mu_0, \frac{\sigma^2}{n}\right),$$

meaning that the centre of this normal distribution is the claimed mean μ_0 , and the standard error is $\frac{\sigma}{\sqrt{n}}$.



If we wanted to calculate the probability of a given $\bar{x}$ (or worse), then we would need to convert the area into an area of the standard normal distribution. The connection between the two distributions is the familiar z -score formula. For example, for a left-tailed test, the P-value associated with an observation $\bar{x}$ might be an area like this:



Since the centre of our sampling distribution is assumed to be μ_0 and the standard error is $\frac{\sigma}{\sqrt{n}}$, we can use the formula:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}.$$

The exact calculation of the P-value will then depend on what type of test we are performing. According to our discussion in Section 1, we either need to calculate:

- $P\left(Z < \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right)$ if it is a left-tailed test,
- $P\left(Z > \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right) = 1 - P\left(Z < \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right)$ if it is a right-tailed test, or perhaps
- $2P\left(Z > \left|\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right|\right)$ if it is a two-tailed test.

Testing μ with known σ

When the population standard deviation σ is known, use the test statistic

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}.$$

The probability of the resulting value can be calculated using the z -table for the standard normal distribution $Z \sim \mathcal{N}(0, 1)$.

15.2.2 Example 1: Coffee bags

Suppose that you work at a company that produces bags of coffee beans. The bags are supposed to have an average weight of 150 g, but you suspect that the true average is actually lower than this.

To test your suspicion, you take a simple random sample of 30 bags and find a sample mean of $\bar{x} = 145$ g. From historical data, you know that the population standard deviation happens to be $\sigma = 15$ g. We will now perform the test using a significance level of $\alpha = 0.01$.

Step 1: Identifying the information

According to our setup, the important values are:

$$\mu_0 = 150, \quad n = 30, \quad \bar{x} = 145, \quad \sigma = 15, \quad \alpha = 0.01.$$

The company claims that the population mean is 150 g, while we suspect that it is less than 150 g. Therefore the null and alternate hypotheses are:

$$H_0 : \mu = 150, \quad \text{and} \quad H_1 : \mu < 150,$$

respectively. Since the alternate hypothesis is a “less than” statement, we are performing a left-tailed test.

Step 2: Calculate the P-value

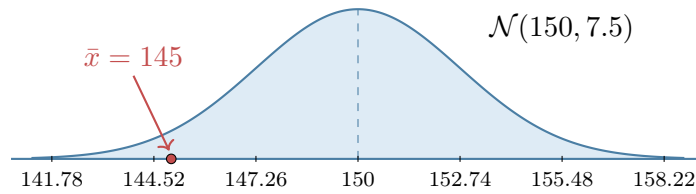
If we assume that the true mean is 150 g, the Central Limit Theorem tells us that the sampling distribution is approximately normal, with parameters:

$$\bar{X} \approx \mathcal{N}\left(150, \frac{15^2}{30}\right) = \mathcal{N}(150, 7.5).$$

Here, the standard error is

$$\frac{\sigma}{\sqrt{n}} = \sqrt{7.5} \approx 2.74.$$

Therefore, the horizontal axis of the null sampling distribution may be labelled approximately as follows:



We now calculate the z -score:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{145 - 150}{15/\sqrt{30}} \approx -1.83.$$

Therefore, we can compute the probability of $\bar{X}$ being less than 145 g by finding the probability of Z being less than -1.83 . Using a z -table, we have:

$$P(\bar{X} < 145) = P(Z < -1.83) \approx 0.0336.$$

Since we are performing a left-tailed test, this is our P-value. In other words, if the true average weight were really 150 g, there would be about a 3.36% chance of obtaining a sample mean of 145 g or less from a sample of this size.

Step 3: Compare the P-value with α

A direct comparison between the P-value and α gives:

$$0.0336 > 0.01.$$

Therefore, at a significance level of 1%, we do not reject H_0 .

Step 4: Give an interpretation

One example interpretation could be as follows.

Interpretation

At the 1% significance level, our sample does not provide convincing evidence that the true average weight of the coffee bags is less than 150 g.

15.3 Testing μ with Unknown σ

15.3.1 Where the previous method breaks

Suppose again that we want to test

$$H_0 : \mu = \mu_0,$$

but this time the population standard deviation σ is unknown. In the previous section, we used the z -score formula

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

to calculate the appropriate P-value. However, if σ is unknown then we can't actually compute this number. Instead, we have to proceed as in Lecture 13 and estimate σ with the *sample* standard deviation s . We can simply replace σ with s in the formula above to obtain a different quantity:

$$\frac{\bar{x} - \mu_0}{s/\sqrt{n}}.$$

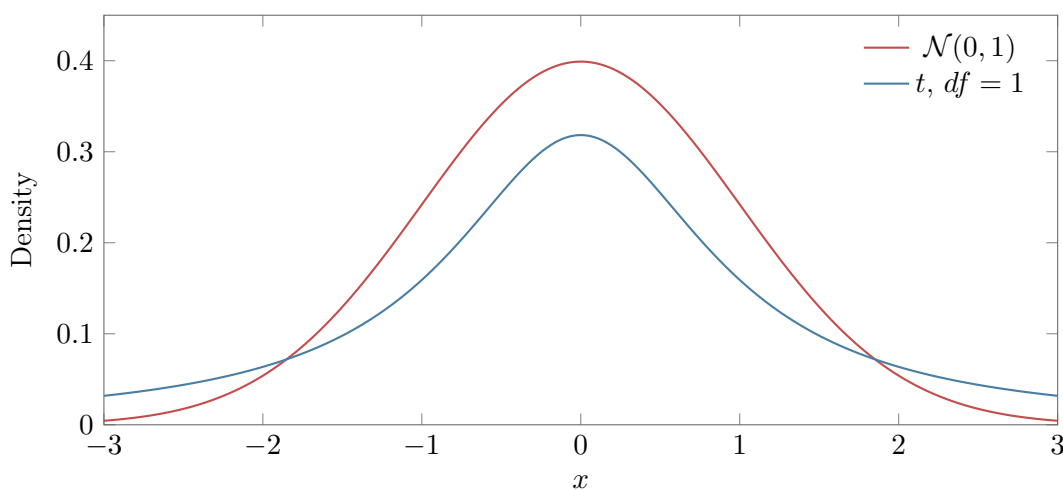
However, as we saw in Lecture 13, this quantity is no longer normally distributed. Instead, we have introduced an extra layer of uncertainty when estimating σ , and therefore we need to use a Student's t -distribution. We will now briefly review this.

15.3.2 A review of Student's t -distribution

Assume that the original population is normally distributed with mean μ . For a sample of size n , with sample mean $\bar{X}$ and sample standard deviation S (now viewed as a random variable in its own right), the quantity

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

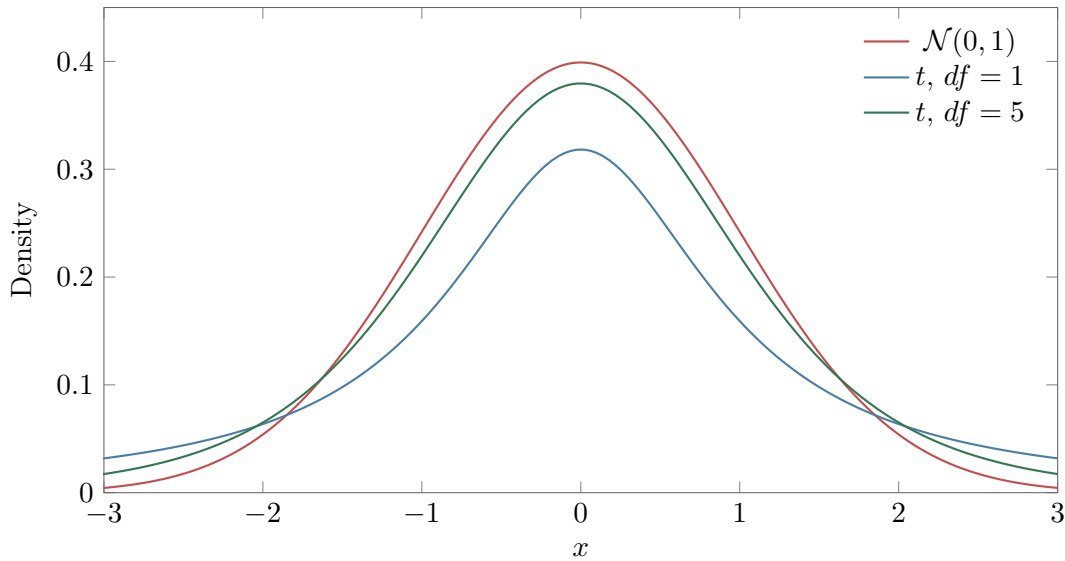
has Student's t -distribution with $df = n - 1$. Student's t -distributions somewhat resemble the standard normal distribution, except that they have heavier tails:



In particular, these distributions still satisfy lots of the same properties as the standard normal distribution:

- they are still symmetric about 0,
- the total area under the curve is still 1, and
- probabilities still correspond to areas under the curve.

However, there are also some important differences. Firstly, there is more than one t -distribution. Instead, there is an entire *family* of these curves, depending on the different values of df . The idea here is that as we increase the size of our sample, the initial random noise in observations should ultimately cancel out, giving us a more accurate estimate of the true value σ . As expected, when we increase the value of n , the value of s tends to get closer to σ , and therefore the t -distribution looks more and more like $\mathcal{N}(0,1)$:



15.3.3 Testing μ with unknown σ

Instead of using the z -score formula from Section 2, in the absence of σ we need to use the t -statistic:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}},$$

where again we use μ_0 as the assumed value of the population mean μ . This formula allows us to calculate the P-value associated with a sample observation.

Testing μ with unknown σ

When the population standard deviation σ is unknown, use

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}.$$

The resulting value allows us to compute the P-value of an observed sample by computing probabilities under the Student's t -distribution with $df = n - 1$.

Again, the exact P-value will depend on what the alternate hypothesis says to do:

$$H_1 : \mu < \mu_0 \implies \text{P-value} = P(T < t),$$

$$H_1 : \mu > \mu_0 \implies \text{P-value} = P(T > t),$$

$$H_1 : \mu \neq \mu_0 \implies \text{P-value} = 2P(T > |t|).$$

Of course, the obvious question at this stage is: how do we actually compute these probabilities in the first place? In the past, if we wanted to compute the value of a probability in the normal distribution then we would compute a z -score and then read the probability $P(Z < z)$ from a table. Similarly, there is a table associated with the Student's t -distributions. In the Appendix below, you will see such a table. The idea is essentially the same as the z -table from Lecture 10: we compute a value of t , and then read the associated probability $P(T < t)$ from the table. Whenever we want

a probability such as $P(T > t)$ or $P(t_1 < T < t_2)$, we can use the same tricks as in the normal distribution:

How to find other areas in a t -distribution

- **Areas to the left of t :** read off $P(T < t)$ directly.
- **Areas to the right of t :** use $P(T > t) = 1 - P(T < t)$.
- **Areas between two values:** use $P(t_1 < T < t_2) = P(T < t_2) - P(T < t_1)$.
- **Negative values:** by symmetry, for $t > 0$,

$$P(T < -t) = 1 - P(T < t).$$

The only major difference in the tables is the precision. In the z -table, we can look up z -values to two decimal places, but in the t -table, we can only look up t -values to one decimal place. The other information in the table provides answers for the different values of df , up to 30.

Exercise

Using the supplied left-tail t -table in the Appendix, find:

- $P(T < 1.3)$ when $df = 18$;
- $P(T < -2.0)$ when $df = 7$;
- $P(T > 1.8)$ when $df = 12$.

Solution

(a)

$$P(T < 1.3) \approx 0.8950.$$

(b)

$$P(T < -2.0) \approx 0.0428.$$

(c) From the table,

$$P(T < 1.8) \approx 0.9515.$$

Therefore,

$$P(T > 1.8) = 1 - P(T < 1.8) \approx 1 - 0.9515 = 0.0485.$$

15.3.4 Example 2: Underweight onigiri

Suppose that every day you buy the same advertised 200 g onigiri from a convenience store. However, you suspect that the company is misleading its customers, and that the true average weight is actually lower than 200 g.

To test your suspicion, you weigh one randomly selected onigiri on each of the 31 days in July. From this sample, you calculate a sample mean of $\bar{x} = 197$ g and a sample standard deviation of $s = 9$ g.

We will now perform the test using a significance level of $\alpha = 0.05$.

Step 1: Identifying the information

According to our setup, the important values are:

$$\mu_0 = 200, \quad n = 31, \quad \bar{x} = 197, \quad s = 9, \quad \alpha = 0.05.$$

The company claims that the population mean is 200 g, while we suspect that it is less than 200 g. Therefore the null and alternate hypotheses are:

$$H_0 : \mu = 200, \quad \text{and} \quad H_1 : \mu < 200,$$

respectively. Since the alternate hypothesis is a “less than” statement, we are performing a left-tailed test.

Step 2: Calculate the P-value

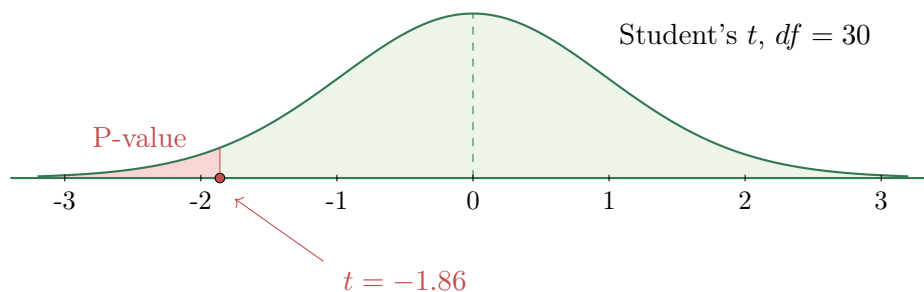
In the previous example, we knew the true population standard deviation σ , and therefore we were able to use the standard normal distribution. In this example, however, we only know the sample standard deviation $s = 9$. Therefore, we need to calculate a t -statistic instead:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{197 - 200}{9/\sqrt{31}} \approx -1.86.$$

Since the sample size is $n = 31$, the number of degrees of freedom is:

$$df = n - 1 = 31 - 1 = 30.$$

Therefore, assuming that the null hypothesis is true, the calculated value $t = -1.86$ lives in a Student's t -distribution with $df = 30$:



Since we are performing a left-tailed test, the P-value is the probability of obtaining a t -statistic of -1.86 or less:

$$\text{P-value} = P(T < -1.86), \quad df = 30.$$

Our t -table only lists values to one decimal place, so -1.86 lies between the listed values -1.9 and -1.8 . From the table, we have:

$$P(T < -1.9) \approx 0.03354$$

and

$$P(T < -1.8) \approx 0.04096.$$

Since -1.86 lies between these two values, we estimate the P-value by taking their midpoint:

$$\text{P-value} \approx \frac{0.03354 + 0.04096}{2} = 0.03725.$$

In other words, if the true average weight were really 200 g, there would be about a 3.7% chance of obtaining a t -statistic of -1.86 or less from a sample of this size.

Step 3: Compare the P-value with α

A direct comparison between the P-value and α gives:

$$0.03725 < 0.05.$$

Therefore, at a significance level of 5%, we reject H_0 .

Step 4: Give an interpretation

Interpretation

At the 5% significance level, our sample provides convincing evidence that the true average weight of the onigiri is less than the advertised 200 g.

15.4 Testing the Binomial Success Probability p

15.4.1 The sampling distribution of the sample proportion

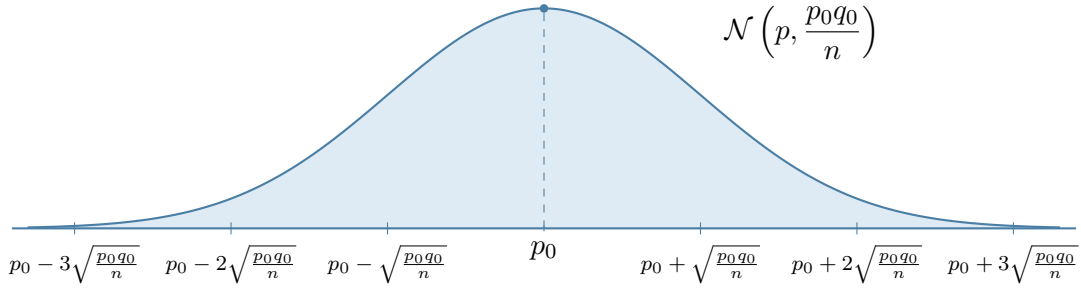
Suppose now that we are working with a binomial experiment with an unknown success probability p . As a sample, we observe n trials of the experiment, and we witness r successes and $n - r$ failures. From this, we can test a claim about the true probability p , that is, a null hypothesis of the form:

$$H_0 : p = p_0.$$

Using the sample data, we can calculate the sample proportion $\hat{p} = \frac{r}{n}$. Assuming that the null hypothesis is true, the normal approximation to this binomial experiment will be the sampling distribution of $\hat{P}$, arranged as a normal distribution with parameters

$$\hat{P} \approx \mathcal{N}\left(p_0, \frac{p_0 q_0}{n}\right),$$

where here $q_0 = 1 - p_0$. Geometrically, the centre of this distribution is the claimed value p_0 , and its standard error is the term $\sqrt{\frac{p_0 q_0}{n}}$, meaning that our normal distribution counts out in terms of the latter quantity:



Therefore, the corresponding z -score is

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}.$$

Testing the binomial success probability p

When testing the true value of a binomial success probability p , use the test statistic

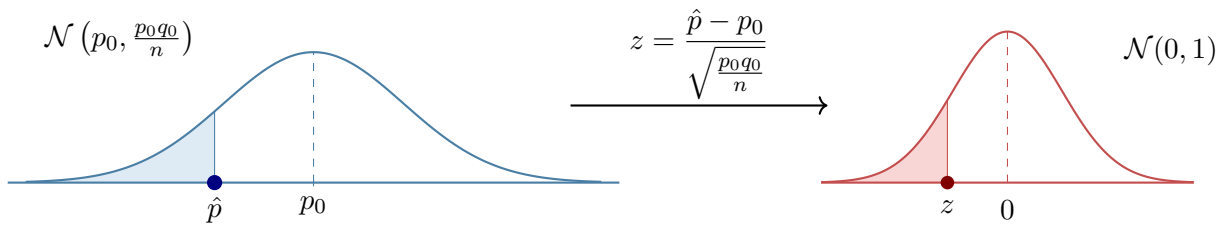
$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}},$$

where here

$$\hat{p} = \frac{r}{n} \quad \text{and} \quad q_0 = 1 - p_0.$$

The probability of the resulting value can be calculated using the z -table for the standard normal distribution $Z \sim \mathcal{N}(0, 1)$.

As with Section 2, geometrically the formula transforms our probability calculation into an area of the normal distribution, which we can determine from the z -table. The formula in the above box provides the link to the standard normal distribution, so that areas can be calculated easily:



Before applying the approximation, we should always check that $np_0 \geq 5$ and $nq_0 \geq 5$ – this suggests that the normal approximation is reasonable. Also, notice that for a hypothesis test the standard error is calculated using the null value p_0 :

$$\sqrt{\frac{p_0 q_0}{n}}.$$

This differs from a confidence interval for p , where we used the observed values $\hat{p}$ and $\hat{q}$:

$$\sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

15.4.2 Example 3: Penalty kicks

Suppose that your mate John claims that he is as good as Lionel Messi at taking penalty kicks. Messi's career penalty success rate is 78%. So, you watch John practice, and amazingly he scores 52 penalties out of 55 attempts. After witnessing John's greatness, you suspect that his true accuracy may be even better than Lionel Messi's.

We will now perform the hypothesis test using a significance level of $\alpha = 0.01$.

Step 1: Identifying the information

According to our setup, the important values are:

$$p_0 = 0.78, \quad q_0 = 1 - p_0 = 0.22, \quad n = 55, \quad r = 52, \quad \alpha = 0.01.$$

John's initial claim corresponds to a success probability of 0.78, while we suspect that his true probability is greater than 0.78. Therefore the null and alternate hypotheses are:

$$H_0 : p = 0.78, \quad \text{and} \quad H_1 : p > 0.78,$$

respectively. Since the alternate hypothesis is a "greater than" statement, we are now performing a *right*-tailed test.

Step 2: Calculate the P-value

Before using the normal approximation, we should first check the approximation conditions. We have:

$$np_0 = 55(0.78) = 42.9 \quad \text{and} \quad nq_0 = 55(0.22) = 12.1.$$

Both of these values are greater than 5, so indeed a normal approximation is reasonable.

To calculate the P-value, we first need to find the observed sample proportion:

$$\hat{p} = \frac{r}{n} = \frac{52}{55} \approx 0.945.$$

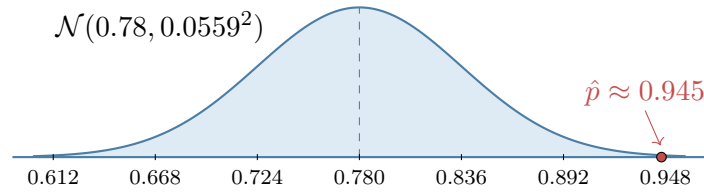
If we assume that the true success probability is $p_0 = 0.78$, then the normal approximation tells us that the sampling distribution of $\hat{P}$ is approximately normal, with parameters:

$$\hat{P} \approx \mathcal{N}\left(0.78, \frac{(0.78)(0.22)}{55}\right).$$

In particular, this distribution is centred at our claimed p_0 . Here, the standard error is:

$$\sqrt{\frac{p_0 q_0}{n}} = \sqrt{\frac{(0.78)(0.22)}{55}} \approx 0.0559.$$

Therefore, the horizontal axis of the sampling distribution may be labelled approximately as follows:



Just by inspecting the graph, we see that $\hat{p}$ is extremely far from the center, and therefore extremely unlikely. Nonetheless, we may determine this probability by calculating the relevant z -score:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} = \frac{0.945 - 0.78}{\sqrt{\frac{(0.78)(0.22)}{55}}} \approx 2.96.$$

Therefore, we can compute the probability of $\hat{P}$ being greater than approximately 0.945 by finding the probability of Z being greater than 2.96. Using a z -table, we first find that:

$$P(Z < 2.96) = 0.99846.$$

This value is the probability to the *left* of 2.96. We are performing a right-tailed test, so we would like to find the area to the right instead. We have:

$$P(Z > 2.96) = 1 - P(Z < 2.96) = 1 - 0.99846 = 0.00154.$$

This is our P-value. In other words, if John's true probability of scoring a penalty kick were really 0.78, there would be about a 0.154% chance of observing a sample proportion of approximately 0.945 or greater from 55 attempts.

Step 3: Compare the P-value with α

A direct comparison between the P-value and α gives:

$$0.00154 < 0.01.$$

Therefore, at a significance level of 1%, we reject H_0 .

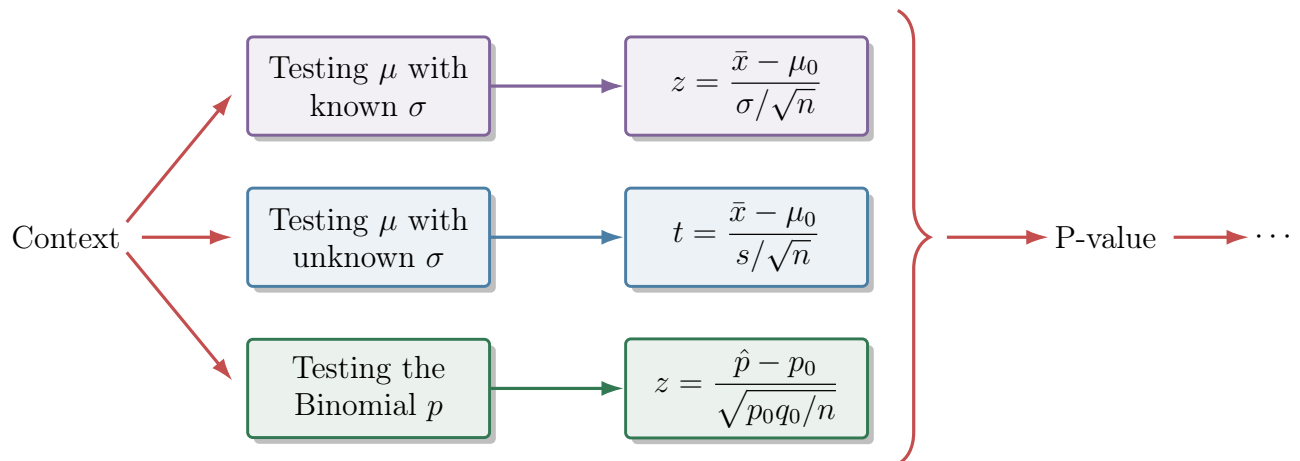
Step 4: Give an interpretation

Interpretation

At the 1% significance level, John's results provide convincing evidence that his true probability of scoring a penalty kick is greater than 0.78.

15.5 Summary

Throughout this lecture, we have seen three unique settings in which we can perform a hypothesis test on a population parameter. Generally speaking, it is our job to correctly identify the context of any given problem, and then to know which test statistic to compute in each situation. Roughly, we have the following flow:



Regardless of the context, the overall workflow of a hypothesis test still remains the same:

1. we state H_0 and H_1 , and record some sample data, then
2. we calculate the appropriate test statistic and P-value, then
3. we compare the P-value with α and decide whether to reject H_0 , and finally
4. we give a conclusion in the context of the problem.

15.6 Appendix: Tables for Student's t -Distributions

The tables below provide approximate answers to the area under students t -distributions at various degrees of freedom. In distinction to the z -table we have seen previously, here the tables are only accurate to one decimal place. In the event that t -statistics are obtained to two decimal places, there are three options:

- if the second decimal is a 1, 2 or 3, then we may round *down* to the first decimal place,
- if the second decimal is a 7, 8 or 9, then we may round *up* to the first decimal place, and
- if the second decimal is a 4, 5 or 6, then we may take the midpoint of the two nearest answers on the table.

Replacing the second decimal are rows that correspond to areas under the curve for values of df ranging from 1 up to 30.

Lecture 16

Hypothesis Tests of Paired Data

In the previous lecture, we learned how to test claims about a single population parameter. We saw tests of two parameters: the population mean μ with either known or unknown σ , and the binomial success probability p . In this lecture and the next, we will discuss the prospect of testing whether or not *two* populations have the same mean. In fact, this is generally quite a complicated problem, so today we will assume for simplicity that our two populations are paired in some natural way. As we will see, this assumption can be leveraged to create an appropriate sampling distribution from which we can test whether or not two population means are the same.

16.1 A Review of Hypothesis Tests for the Population Mean

16.1.1 The workflow of a test

Generally speaking, the goal of a hypothesis test is to make a claim and then test the evidence for the claim. The baseline claim is called the *null hypothesis*, and is written H_0 . We then test this statement against something to the contrary by formulating some kind of *alternate hypothesis*, written H_1 . From there, we take a sample, and then answer the following question.

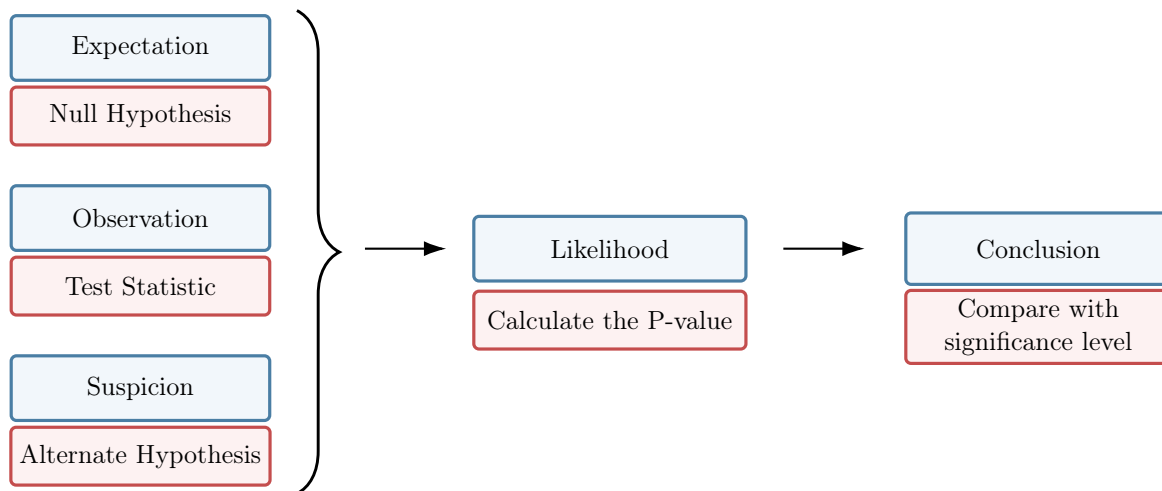
The main question of hypothesis tests

Every hypothesis test asks the same basic question:

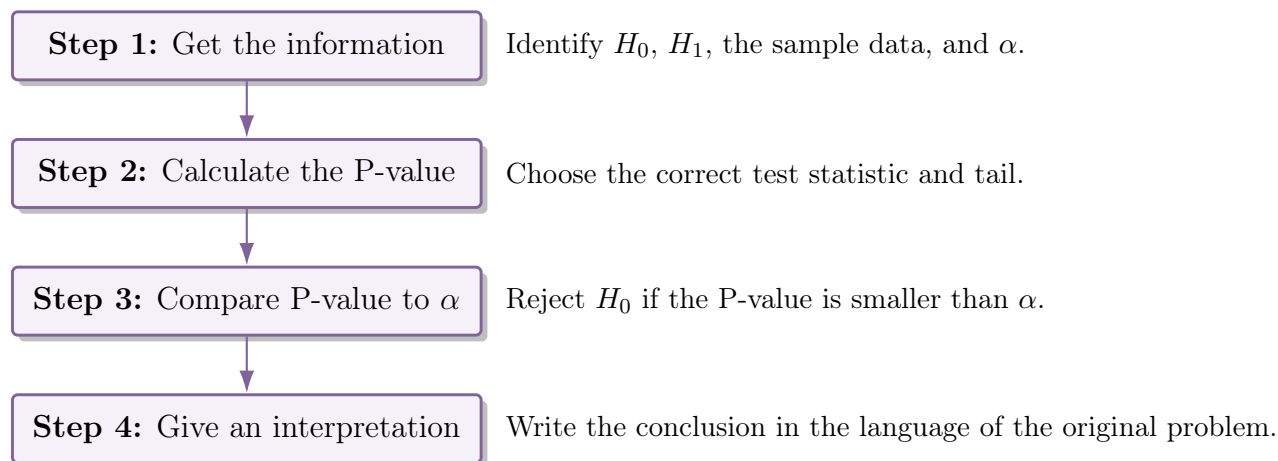
Assuming H_0 is true, how unusual is the observed sample?

The probability calculation will generally change based on the setting of the problem.

Generally, hypothesis tests follow the same overall workflow:



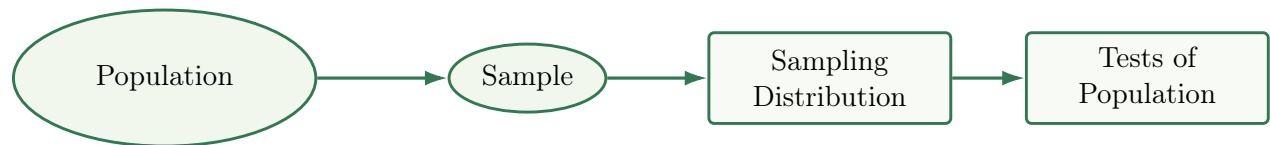
Alternatively, the process of a hypothesis test can be broken down into four major steps:



16.2 What are Paired Populations?

16.2.1 The motivation

So far, our statistical tests have mostly followed this pattern:

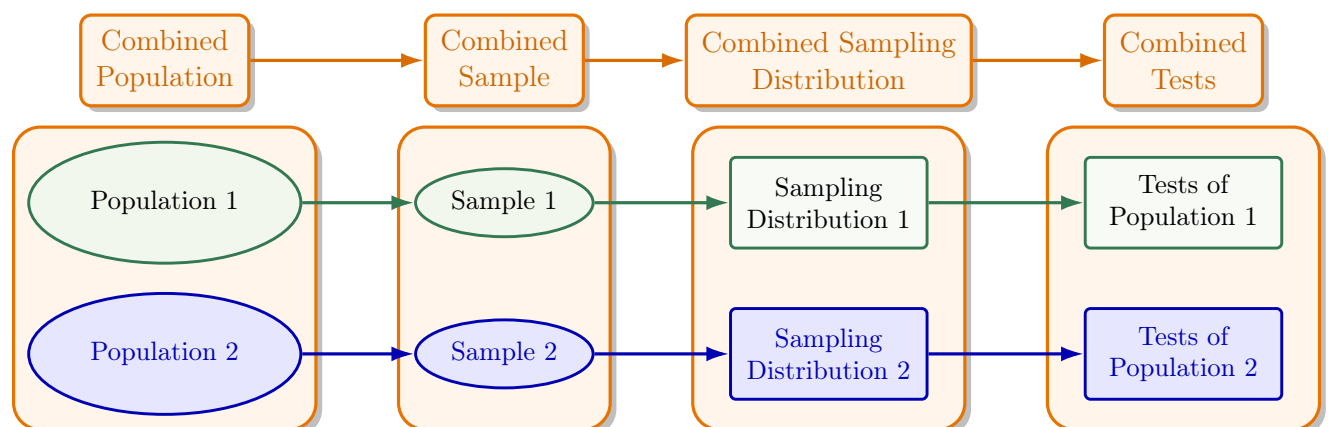


That is, we make a claim about some unknown population parameter of a given population, and then test the legitimacy of that claim by taking a random sample and evaluating the likelihood of our observation.

We will now explore the prospect of comparing two different populations via a hypothesis test. The main claim that we will try to test is:

Are these two populations the same, on average?

In symbols, if the first population has mean μ_1 and the second population has mean μ_2 , we are going to try and test whether or not $\mu_1 = \mu_2$. In order to do so, we are going to combine two populations into a third population, take a sample, and then derive the relevant sampling distribution. Schematically, it will look something like this:



As a matter of fact, the act of combining two populations into a third is rather complicated. So, to make our lives easier, we will assume that our populations are naturally paired together in some meaningful way.

16.2.2 Defining Paired Populations

Throughout the remainder of this lecture, we will only consider populations that are *paired*. Heuristically speaking, populations are paired when there is a natural connection between points in one population and points in the other.

Paired populations

We say that two populations are **paired** when every observation in one data set has a unique counterpart in the other data set.

There are various situations in which we may want to study paired data. For example, we may find paired data in any of the following contexts:

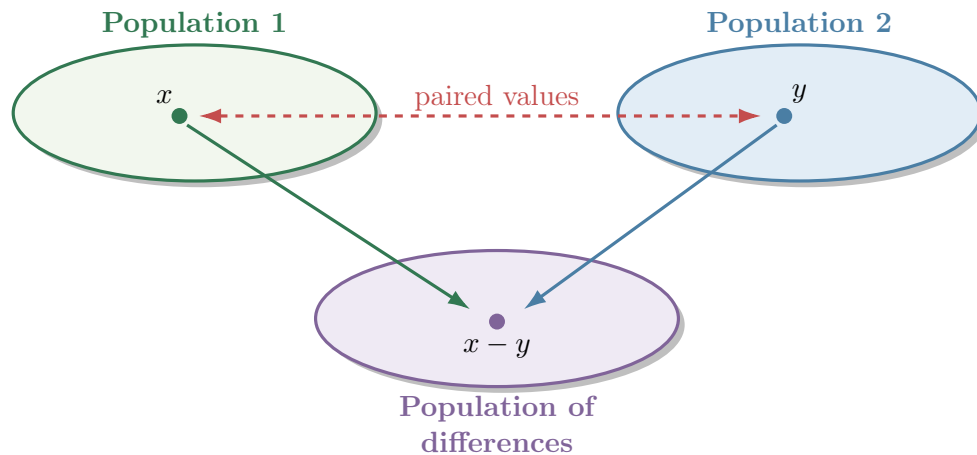
- studying the effects of an action on the same subjects, for instance the effects of caffeine on reaction time.
- studying the accuracy of multiple measuring devices, for instance the accuracy of two different heart rate monitors used on the same person.
- studying groups of naturally matched objects, for instance twins, or left/right feet on the same person.

16.2.3 Combining paired populations

Suppose that our two paired populations are defined by random variables X and Y . Because the populations are paired, any observation x from X has a unique partner y in Y . That means that we can consider the quantity $x - y$, and this will be some meaningful number that represents the difference between the value of X and the value of Y . As a shorthand, we will write:

$$D = X - Y.$$

We can interpret this as a new random variable – the difference of X and Y .¹ Schematically, we are doing something like the following.



16.3 The Sampling Distribution for Differences of Pairs

In order to perform hypothesis tests on the difference $X - Y$, we are going to describe the sampling distribution of $\bar{D}$, the random variable of sample means taken from the population D . At the

¹Hence the letter “ D ” for “difference”.

very least, we know from the Central Limit Theorem that, for sufficiently large n , $\bar{D}$ should be approximately normally distributed with parameters given by:

$$\bar{D} \approx \mathcal{N}\left(\mu_D, \frac{\sigma_D^2}{n}\right).$$

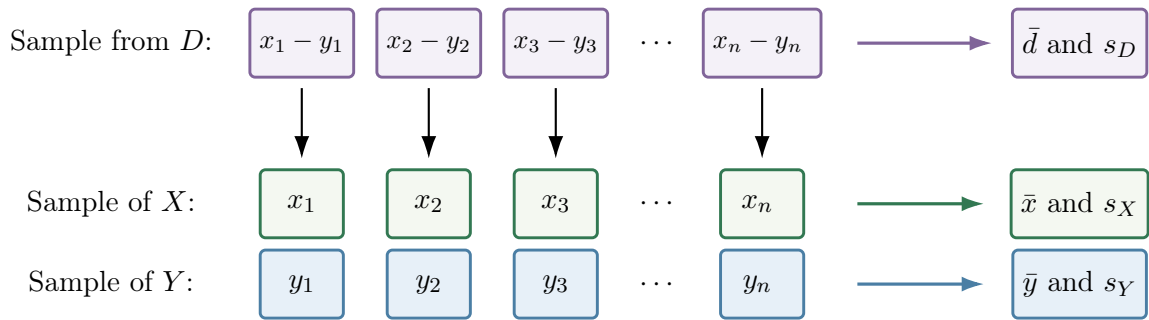
To understand this distribution, we need to understand both the mean μ_D and the variance σ_D^2 of the random variable $D = X - Y$.

16.3.1 Samples of paired data

Suppose we take a sample of n observations from D , say $d_1, \dots, d_n$. Since each of the observations is a difference of the form $x - y$, taking n of these will look like:

$$d_1 = x_1 - y_1, \quad d_2 = x_2 - y_2, \quad \dots, \quad d_n = x_n - y_n.$$

These differences can be broken apart to recover samples of X and Y of size n :



It would be helpful if we could somehow describe the value of the sample statistics $\bar{d}$ and s_D in terms of the information coming from x_i and y_i . At the very least, this should give us clues about how averages and variances interact between X , Y and D .

16.3.2 The Mean of $D = X - Y$

If we wanted to compute the sample mean of our sample $d_1, \dots, d_n$, we can use the fact that each $d_i = x_i - y_i$ and break apart the formula into two separate pieces:

$$\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i = \frac{1}{n} \sum_{i=1}^n (x_i - y_i) \stackrel{(*)}{=} \frac{1}{n} \sum_{i=1}^n x_i - \frac{1}{n} \sum_{i=1}^n y_i = \bar{x} - \bar{y}.$$

Here, in the step labelled $(*)$, we have broken apart and rearranged the sum using the property:

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n (x_i - y_i) &= \frac{1}{n} \left((x_1 - y_1) + (x_2 - y_2) + \dots + (x_n - y_n) \right) \\
&= \frac{1}{n} \left(x_1 + x_2 + \dots + x_n - y_1 - y_2 - \dots - y_n \right) \\
&= \frac{1}{n} \left((x_1 + x_2 + \dots + x_n) - (y_1 + y_2 + \dots + y_n) \right) \\
&= \frac{1}{n} (x_1 + x_2 + \dots + x_n) - \frac{1}{n} (y_1 + y_2 + \dots + y_n) = \frac{1}{n} \sum_{i=1}^n x_i - \frac{1}{n} \sum_{i=1}^n y_i.
\end{aligned}$$

Sample mean of the differences

For paired data,

$$\bar{d} = \bar{x} - \bar{y}.$$

In other words, the sample mean of the differences is the difference of the sample means.

As a matter of fact, the same phenomenon happens at the population level. Recall that the population mean of a random variable is really given by the *expected value*. In the discrete case, we saw that the formula for the expected value is the weighted average. In the case of a continuous random variable, the definition is a little bit complicated and involves some calculus that we have not seen before. What is important for us is the following property:

$$E(X - Y) = E(X) - E(Y),$$

which holds for all random variables with finite expected values, continuous or otherwise. We can use this property to describe the population mean μ_D in terms of μ_X and μ_Y :

$$\mu_D = E(D) = E(X - Y) = E(X) - E(Y) = \mu_X - \mu_Y.$$

As with the sample statistic $\bar{d}$, we see that the mean of the differences is the difference of the means.

Population mean of the differences

If $D = X - Y$, then

$$\mu_D = \mu_X - \mu_Y.$$

16.3.3 The Variance of $D = X - Y$

Unfortunately, the variance of a difference is not as well-behaved as the mean. According to the formula for the sample variance, we have:

$$s_D^2 = \frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n - 1}.$$

Now, we know from the previous section that $\bar{d} = \bar{x} - \bar{y}$. This means that we can rewrite the terms in the sum as:

$$(d_i - \bar{d})^2 = \left((x_i - y_i) - (\bar{x} - \bar{y}) \right)^2 = \left((x_i - \bar{x}) - (y_i - \bar{y}) \right)^2.$$

Unfortunately, when we expand out this bracket, we gain an extra cross-term:

$$\begin{aligned} (d_i - \bar{d})^2 &= \left((x_i - \bar{x}) - (y_i - \bar{y}) \right)^2 \\ &= (x_i - \bar{x})^2 - 2(x_i - \bar{x})(y_i - \bar{y}) + (y_i - \bar{y})^2. \end{aligned}$$

This means that if we were to try to break down the formula for s_D^2 and rearrange it into smaller parts, we would obtain:

$$\begin{aligned} s_D^2 &= \frac{1}{n-1} \sum_{i=1}^n \left((x_i - \bar{x})^2 - 2(x_i - \bar{x})(y_i - \bar{y}) + (y_i - \bar{y})^2 \right) \\ &= \underbrace{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}_{s_X^2} + \underbrace{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}_{s_Y^2} - \frac{2}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}). \end{aligned}$$

Although our algebra has uncovered the variances s_X^2 and s_Y^2 , we also have an extra quantity. This extra piece is known as the *sample covariance*, and is often written either s_{XY} or $\text{cov}(x, y)$.

The Sample Variance of the Differences

For paired data, the sample variance can be written

$$s_D^2 = s_X^2 + s_Y^2 - 2 \text{cov}(x, y).$$

Compute the differences first:

$$d_i = x_i - y_i,$$

and then compute the sample standard deviation s_D from the list of differences.

At the population level, the same issue appears. We have, in general,

$$\text{Var}(X - Y) \neq \text{Var}(X) - \text{Var}(Y).$$

This can be verified using the definition of the variance Var of a random variable. Again using that the expected value E behaves well with sums and differences, we find that

$$\begin{aligned} \text{Var}(X - Y) &= E \left[((X - Y) - E(X - Y))^2 \right] \\ &= E \left[((X - Y) - (E(X) - E(Y)))^2 \right] \\ &= E \left[((X - E(X)) - (Y - E(Y)))^2 \right] \\ &= E \left[(X - E(X))^2 \right] + E \left[(Y - E(Y))^2 \right] - 2E \left[(X - E(X))(Y - E(Y)) \right] \\ &= \text{Var}(X) + \text{Var}(Y) - 2 \text{Cov}(X, Y), \end{aligned}$$

which directly mimics the result for sample variances.

Roughly, the population covariance tells us how X and Y are linearly related to each other – we will explore this in depth during Lecture 18. For now, we summarise our discussion with regard to the population variance σ_D^2 .

The Population Variance of the Differences

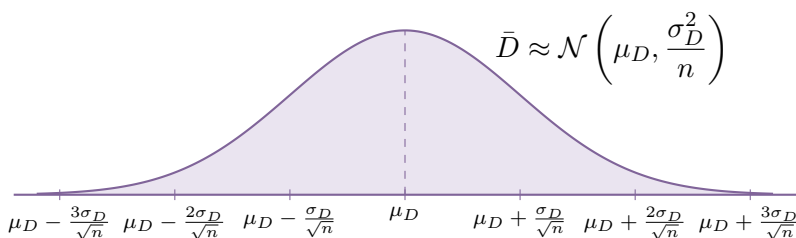
For paired random variables, the population variance can be written

$$\sigma_D^2 = \sigma_X^2 + \sigma_Y^2 - 2 \text{Cov}(X, Y).$$

In particular, it is generally not the case that $\sigma_D^2 = \sigma_X^2 + \sigma_Y^2$.

16.3.4 The sampling distribution of differences

If we take a random sample of n independent pairs and compute their differences, then the Central Limit Theorem tells us that the sampling distribution should be approximately normal:



Using the results of the previous section, we now see that if X has mean μ_X and variance σ_X^2 and Y has mean μ_Y and variance σ_Y^2 , then the difference $D = X - Y$ has parameters $\mu_D = \mu_X - \mu_Y$ and $\sigma_D^2 = \sigma_X^2 + \sigma_Y^2 - 2 \text{Cov}(X, Y)$.

However, there is one problem: in practice, we almost never know the value of the extra covariance term. This means that, even if we happened to know the variances σ_X^2 and σ_Y^2 , it is not enough information to calculate the value of σ_D^2 . Therefore, we replace σ_D with the *sample* standard deviation s_D , and use a Student's t -distribution.

$$z = \frac{\bar{d} - \mu_D}{\sigma_D / \sqrt{n}} \quad \xrightarrow{\text{replace } \sigma_D \text{ by } s_D} \quad t = \frac{\bar{d} - \mu_D}{s_D / \sqrt{n}}$$

σ_D is unknown Student's t , $df = n - 1$

When performing hypothesis tests, the latter becomes the relevant test statistic.

16.4 Hypothesis Testing Paired Data

16.4.1 The null hypothesis

Usually, when testing paired data, we are trying to answer:

Are these two populations the same, on average?

If μ_X is the mean of Population 1 and μ_Y is the mean of Population 2, mathematically we can rewrite this as:

$$\mu_X = \mu_Y \quad \text{or equivalently, } \mu_X - \mu_Y = 0.$$

Since $\mu_D = \mu_X - \mu_Y$, the null hypothesis therefore becomes

$$H_0 : \mu_X - \mu_Y = 0.$$

Whenever we test whether two means are equal, this is always the correct form of the null hypothesis. The alternate hypothesis depends on the suspicion:

$$H_1 : \begin{cases} \mu_X - \mu_Y < 0, & \text{left-tailed test,} \\ \mu_X - \mu_Y > 0, & \text{right-tailed test,} \\ \mu_X - \mu_Y \neq 0, & \text{two-tailed test.} \end{cases}$$

In fact, selecting the correct alternate hypothesis is a challenge.

Choosing the correct alternate hypothesis

If you suspect that the average of Population 1 is *less* than the average of Population 2, then:

$$\mu_X < \mu_Y \quad \implies \quad \mu_X - \mu_Y < 0$$

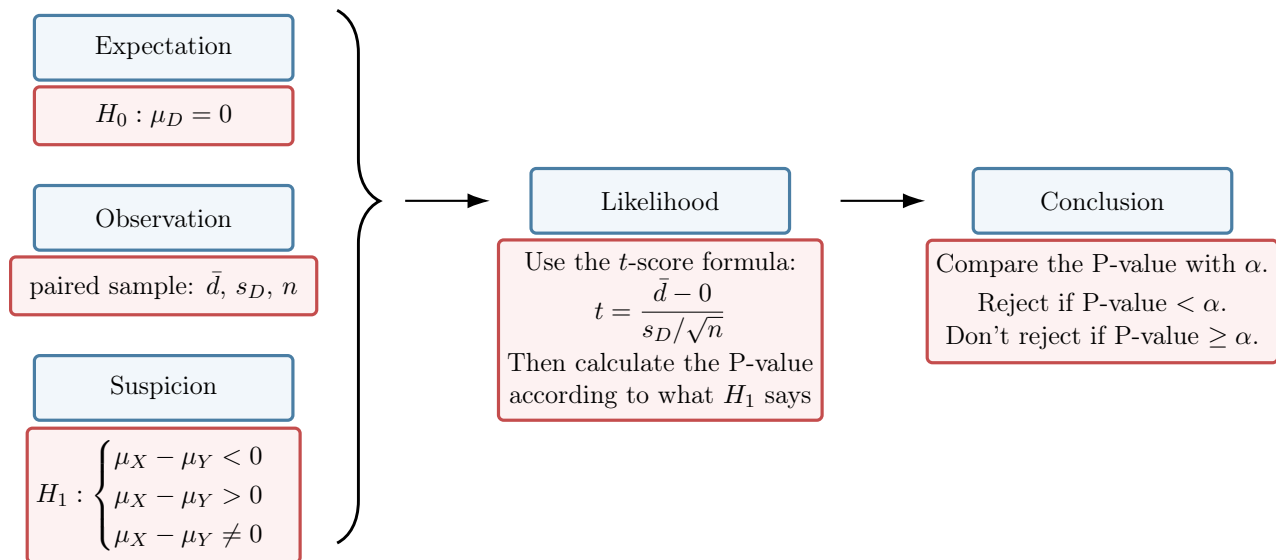
and the alternate hypothesis should read $H_1 : \mu_X - \mu_Y < 0$. Conversely, if you suspect that the average of Population 1 is *greater* than the average of Population 2, then:

$$\mu_X > \mu_Y \quad \implies \quad \mu_X - \mu_Y > 0$$

and the alternate hypothesis should read: $H_1 : \mu_X - \mu_Y > 0$.

16.4.2 The paired-data workflow

According to our discussion in Section 3.4, we will generally not have access to the covariance of the two populations, and therefore we treat σ_D^2 as unknown. This means that, under the null hypothesis, the test statistic follows a Student's t-distribution instead of a normal distribution. Putting this all together, the total workflow for a hypothesis test for paired data is:



16.5 Example: Movie Ratings

Suppose that Dr. O'Connell wants to test whether the public think that *Twilight* and *The Twilight Saga: New Moon* are equally good. Let

μ_X = the average rating of *Twilight* out of 10, and

μ_Y = the average rating of *The Twilight Saga: New Moon* out of 10.

After watching both movies himself, he suspects that, on average, people think *Twilight* is probably better than *The Twilight Saga: New Moon*. To test this claim, he takes a simple random sample of 30 people and asks each person to rate both movies. For each person, he calculates the difference $d_i = x_i - y_i$ of their ratings, where x_i is the rating for *Twilight* and y_i is the rating for *The Twilight Saga: New Moon*. From the paired data, he finds

$$\bar{d} = 0.6 \quad \text{and} \quad s_D = 1.5.$$

We will perform a test with significance level $\alpha = 0.03$.

Step 1: Extracting the information

According to the setup of the test, the important values are:

$$\mu_{D,0} = 0, \quad n = 30, \quad \bar{d} = 0.6, \quad s_D = 1.5, \quad \alpha = 0.03.$$

The null hypothesis says that the two movies have the same average rating:

$$H_0 : \mu_X - \mu_Y = 0.$$

Since we defined $D = X - Y$, the suspicion that *Twilight* is rated higher than *The Twilight Saga: New Moon* becomes

$$H_1 : \mu_X - \mu_Y > 0.$$

So, this is a right-tailed test.

Step 2: Calculate the P-value

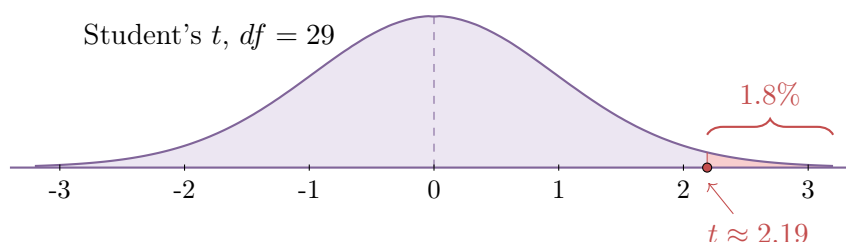
Since we don't know σ_D , we cannot use a z -score. Instead, we use the test statistic:

$$t = \frac{\bar{d} - 0}{s_D/\sqrt{n}} = \frac{0.6 - 0}{1.5/\sqrt{30}} \approx 2.19.$$

Since $n = 30$, the degrees of freedom are $df = n - 1 = 29$. Therefore, the P-value is

$$\text{P-value} = P(T > 2.19) \approx P(T > 2.2) = 1 - P(T < 2.2) \approx 1 - 0.98204 = 0.01796.$$

Our usual table only gives values to one decimal place, so in the above approximation, we have used $t = 2.2$ instead. Graphically, the P-value is the shaded area below:



Put differently, if the two movies truly had equal average ratings, then there would be about a 1.8% chance of getting a t -statistic this large (or larger) from a random sample of 30 people.

Step 3: Compare the P-value to α

In this example, we are asked to perform the test with a significance level of $\alpha = 0.03$. Our P-value is approximately 0.018, which is less than 0.03. Therefore, we reject the null hypothesis.

Step 4: Give an interpretation

Interpretation

At the 3% significance level, the sample provides convincing evidence to suggest that the public rate *Twilight* higher than *The Twilight Saga: New Moon*, on average.

Here, we do not make any separate claims about the average rating of either film. Instead, we tested the difference between their average ratings.

Lecture 17

Hypothesis Tests of Independent Data

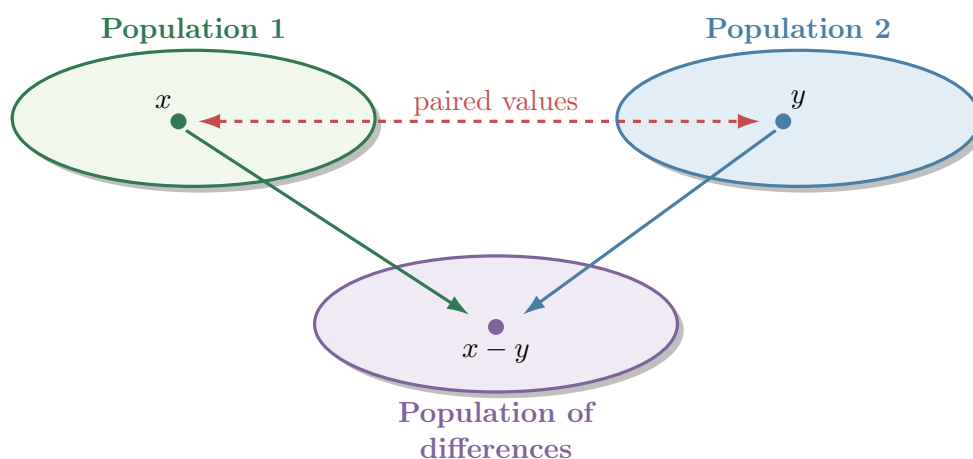
In Lecture 16, we compared two populations whose observations were naturally paired. The pairing allowed us to subtract inside each pair, form a new population of differences $D = X - Y$, and reduce the problem to a one-sample t -test. In this lecture, we will remove that assumption and instead take our populations to be *independent*. As we will see, this new assumption allows us to again perform hypothesis tests of two populations simultaneously. This time, however, we can no longer construct differences $d = x - y$. Instead, as we will see, we need to work around this loss of convenience in another way. In the end, we will use our new techniques to test the difference of population means, and also to test the difference between the success probabilities of two independent binomial experiments.

17.1 A Review of Paired Data

In this lecture, we will continue to discuss hypothesis tests that compare the parameters of two separate populations. Given two populations, the central question here is:

Are these two populations the same, on average?

In principle, this can be answered by comparing the true means of the two populations and testing whether or not they are equal. In the process of performing a test like this, we need to create a sampling distribution that combines sample data from the two populations. In Lecture 16, we noted that, generally speaking, this is a difficult task. Therefore, we made our lives easier by assuming that our populations were naturally *paired*. Since each individual observation in one population had a unique counterpart in the other, this assumption allowed us to create a third population made up of all of the differences between these paired values:



Symbolically, we used the letter D to represent this “population of differences”. Formally, D is nothing but $X - Y$, the difference of the random variables used to define the two populations.

Using some standard facts about expected values and variances, we saw that the mean and the variance of this random variable D were:

$$\mu_D = \mu_X - \mu_Y, \quad \sigma_D^2 = \sigma_X^2 + \sigma_Y^2 - 2 \text{Cov}(X, Y).$$

Although the mean takes a nice form, the variance carries an extra term known as *covariance*. Roughly speaking, this extra covariance term measures how the paired values move together. Since σ_D is generally unknown, we had to estimate the standard deviation of the differences using s_D and use a Student’s t -distribution, giving the following test statistic.

Paired-data test statistic

For paired data, first compute the differences $d_i = x_i - y_i$. Then use

$$t = \frac{\bar{d} - 0}{s_D / \sqrt{n}}, \quad df = n - 1.$$

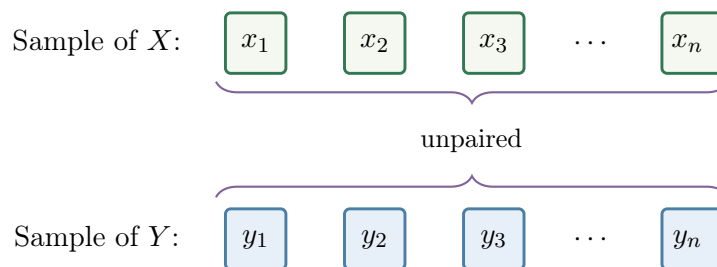
17.2 Independent Populations

17.2.1 Changing our assumptions

When performing hypothesis tests involving two populations, there is a major technical hurdle when attempting to describe the appropriate sampling distribution. In order to make things easier for ourselves, in the last lecture we assumed that the two populations were naturally paired. This allowed us to take samples from the population D and break them apart into corresponding samples of X and Y , thereby giving us a description of the sampling distribution of $\bar{D}$.

In this lecture, we will no longer assume that our populations are paired. Instead, we will assume that X and Y are *independent*, meaning that there is no probabilistic dependence between the two. In fact, we saw this term when discussing probability earlier in the course: two events were called “independent” if the occurrence of one did not change the probability of the other. In a similar spirit, we call X and Y independent whenever knowledge of one does not change the probability distribution of the other. Mathematically, assuming that X and Y are independent implies that there is *no* interaction between the two, and therefore the covariance $\text{Cov}(X, Y)$ vanishes.

Although this looks promising, we also have to admit that we have lost something in the process of relaxing the assumption of paired data. When observations of X and Y were paired, we were able to form a third population governed by $D = X - Y$, where data points were of the form $d = x - y$. This allowed us to take samples of the population D and break them apart into samples of X and Y individually. If we drop the assumption that X and Y are paired, then we can no longer perform this trick. In other words, even if we picked two simple random samples from X and Y of the same size, there is no natural way to pair them up into some kind of sample of differences:



This means that there is no natural way to create a sample of individual differences for testing the difference $\mu_X - \mu_Y$.

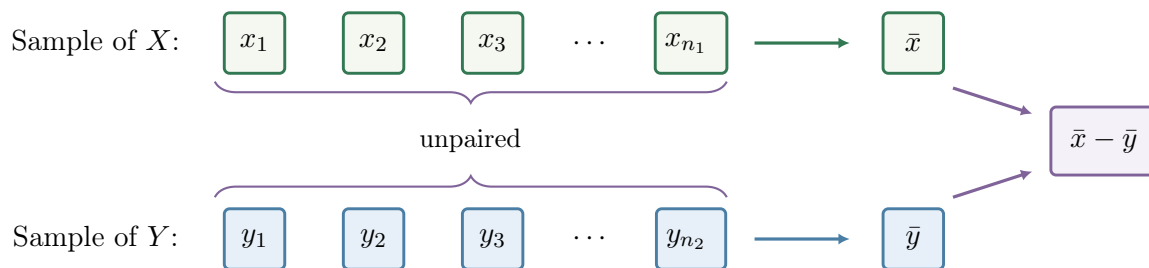
17.2.2 The difference of sample means

Suppose that we take two independent random samples from X and Y , respectively. Since the samples are independent, there is no need to require that these two samples are of the same size. So, let n_1 denote the size of the sample from X and let n_2 denote the size of the sample from Y . Then the sample means $\bar{x}$ and $\bar{y}$ can be calculated individually:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_{n_1}}{n_1} \quad \text{and} \quad \bar{y} = \frac{y_1 + y_2 + \dots + y_{n_2}}{n_2}.$$

Ultimately, our goal is to describe a sampling distribution centred at $\mu_X - \mu_Y$, without being able

to directly appeal to some population of differences D . According to the sample data that we have available to us, we can create a difference of averages by considering the quantity $\bar{x} - \bar{y}$:



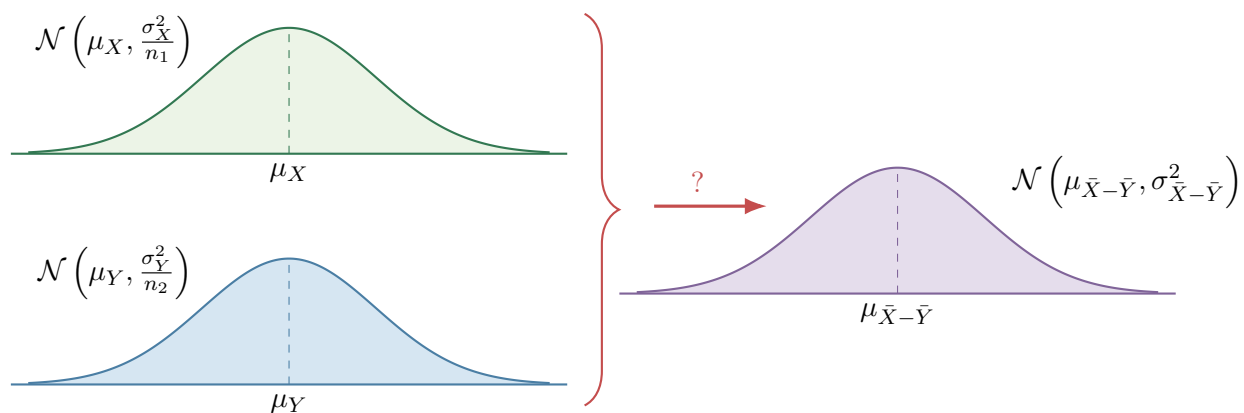
If we wanted to perform a hypothesis test by evaluating the chances of this particular observed value $\bar{x} - \bar{y}$, then we need to consider a probability distribution containing the collection of all possible randomly obtained differences $\bar{x} - \bar{y}$. Therefore, we will study the random variable $\bar{X} - \bar{Y}$.

For sufficiently large samples, the Central Limit Theorem tells us that the individual sampling distributions of $\bar{X}$ and $\bar{Y}$ are approximately normal, with parameters:

$$\bar{X} \approx \mathcal{N}\left(\mu_X, \frac{\sigma_X^2}{n_1}\right) \quad \text{and} \quad \bar{Y} \approx \mathcal{N}\left(\mu_Y, \frac{\sigma_Y^2}{n_2}\right).$$

We emphasize that, in contrast to the case of paired data, the two samples need not have the same size, i.e. it does not need to be the case that $n_1 = n_2$.

We are now going to use these two normal distributions to describe the overall distribution of the random variable $\bar{X} - \bar{Y}$. Schematically, we are looking for the following:



17.2.3 Constructing the sampling distribution for $\bar{X} - \bar{Y}$

In the previous lecture, we saw that the difference of two random variables has an expected value equal to the difference of the expected values of the individual variables. In symbols, for $\bar{X} - \bar{Y}$, we have:

$$E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y}).$$

Since $E(\bar{X})$ is simply the mean of $\bar{X}$, which is μ_X , and similarly for $\bar{Y}$, we may conclude that:

$$\mu_{\bar{X}-\bar{Y}} = E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y}) = \mu_X - \mu_Y.$$

Conveniently, this is the same centre that appeared for paired data: the expected difference between two sample means is the difference between the two population means.

As for the variance, in fact we may again use the same fact as in Lecture 16 to conclude that

$$\text{Var}(\bar{X} - \bar{Y}) = \text{Var}(\bar{X}) + \text{Var}(\bar{Y}) - 2 \text{Cov}(\bar{X}, \bar{Y}),$$

where $\text{Cov}(\bar{X}, \bar{Y})$ is the covariance of $\bar{X}$ and $\bar{Y}$. This is where the magic happens: because the samples are assumed to be independent, $\bar{X}$ and $\bar{Y}$ are also independent, and therefore this extra covariance term vanishes: $\text{Cov}(\bar{X}, \bar{Y}) = 0$. Therefore, the variance formula above can be conveniently simplified to:

$$\text{Var}(\bar{X} - \bar{Y}) = \text{Var}(\bar{X}) + \text{Var}(\bar{Y}).$$

The variance of $\bar{X}$ is simply the second parameter in its normal distribution, that is, the square of the standard error term for $\bar{X}$, namely $\frac{\sigma_X^2}{n_1}$. Similarly, the variance of $\bar{Y}$ is given by $\frac{\sigma_Y^2}{n_2}$. Therefore, we may conclude that for independent populations, the variance of the distribution of $\bar{X} - \bar{Y}$ will be:

$$\sigma_{\bar{X} - \bar{Y}}^2 = \text{Var}(\bar{X} - \bar{Y}) = \text{Var}(\bar{X}) + \text{Var}(\bar{Y}) = \frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}.$$

Putting this all together, we now have a full description of the normal distribution that approximately describes $\bar{X} - \bar{Y}$.

Sampling distribution of the difference of independent means

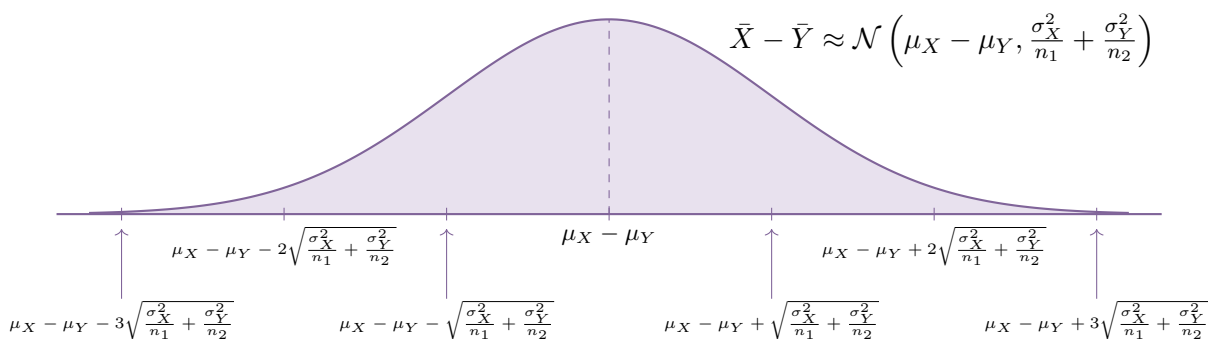
For independent samples,

$$\bar{X} - \bar{Y} \approx \mathcal{N}\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}\right).$$

Our normal distribution will be centred at the term $\mu_X - \mu_Y$, and it will “count out” in terms of the standard error, which is the square root of the variance term above:

$$\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}.$$

Geometrically, we obtain the following image:



Although this is complicated, the important point is that this is entirely *tractable* – there will be a formula that we can use to compute relevant probabilities.

17.3 Hypothesis Tests for Independent Means

17.3.1 The null and alternate hypotheses

As with our discussion in the previous lecture, when we want to test whether or not the two independent populations are equal on average, the null hypothesis is given by

$$H_0 : \mu_X - \mu_Y = 0.$$

We then take two independent samples from X and Y of potentially different sizes. Assuming the null hypothesis, the sampling distribution becomes

$$\bar{X} - \bar{Y} \approx \mathcal{N}\left(0, \frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}\right).$$

The two samples from X and Y have their respective sample means $\bar{x}$ and $\bar{y}$, and these have a difference $\bar{x} - \bar{y}$ that will land somewhere on the normal distribution described above. As with the other hypothesis tests that we have seen previously, the farther the observed value $\bar{x} - \bar{y}$ lies from the centre, the smaller the P-value becomes.

As with the paired populations of Lecture 16, if we wanted to test H_0 against a statement to the contrary, then we can use different statements depending on our particular suspicion:

$$H_1 : \begin{cases} \mu_X - \mu_Y < 0, & \text{says that the population for } X \text{ is smaller on average,} \\ \mu_X - \mu_Y > 0, & \text{says that the population for } X \text{ is larger on average,} \\ \mu_X - \mu_Y \neq 0, & \text{merely says that the averages are different.} \end{cases}$$

Choosing the direction

The order of subtraction must match the alternate hypothesis. If the setup uses the sample difference $\bar{x} - \bar{y}$, then:

- $\mu_X < \mu_Y$ becomes $H_1 : \mu_X - \mu_Y < 0$;
- $\mu_X > \mu_Y$ becomes $H_1 : \mu_X - \mu_Y > 0$.

17.3.2 Test Statistics with known and unknown σ

Suppose that both population standard deviations, σ_X and σ_Y , happen to be known. In this case, the distribution for $\bar{X} - \bar{Y}$ has standard error given by

$$\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}.$$

When we convert the observed difference into a z -score, we schematically take the quotient:

$$z = \frac{\text{observed value} - \text{mean}}{\text{standard error}}.$$

In our case, we are assuming the null hypothesis, which says that $\mu_X - \mu_Y = 0$. The standard error is given above, and can be calculated directly from the information available. Therefore, the observed difference $\bar{x} - \bar{y}$ is converted into the z -score via the formula:

$$z = \frac{(\bar{x} - \bar{y}) - 0}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}}.$$

Independent means with known standard deviations

To test $H_0 : \mu_X - \mu_Y = 0$ when σ_X and σ_Y are known, we use the formula:

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}}.$$

Suppose now that either σ_X or σ_Y is unknown. In this case, we don't have the values to substitute into the previous formula. Therefore, the next best thing we can do is replace both population standard deviations with their sample estimates, and proceed with an approximate Student's t -distribution instead:

$$\begin{array}{ccc} z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}} & \xrightarrow{\text{replace } \sigma_X \text{ and } \sigma_Y \text{ by } s_X \text{ and } s_Y} & t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}}}. \\ \text{either } \sigma_X \text{ or } \sigma_Y \text{ is unknown} & & \text{Student's } t, \\ & & df = \min(n_1 - 1, n_2 - 1) \end{array}$$

As a convention, in this course we will use the smaller of the two available degrees of freedom.

Independent means with unknown standard deviations

To test $H_0 : \mu_X - \mu_Y = 0$ when either population standard deviation is unknown, use the formula

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}}},$$

where $df = \min(n_1 - 1, n_2 - 1)$.

After this, the appropriate P-value can be calculated using the corresponding z -table or t -table.

17.4 Testing Two Binomial Probabilities

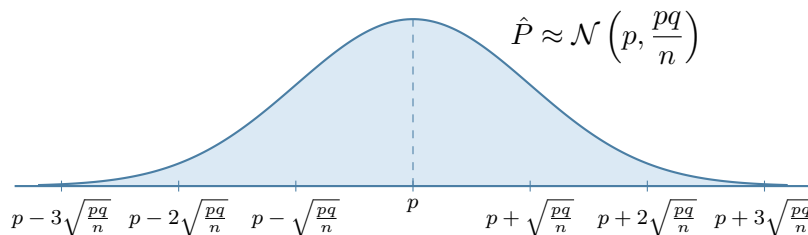
17.4.1 A reminder about one-sample proportions

Suppose that we have a single binomial experiment, with success probability p . If we were to take n trials and witness r successes, then we can calculate an estimate of p from this sample data. This estimate is known as the sample proportion, and it can be calculated simply as:

$$\hat{p} = \frac{r}{n}.$$

As we saw in Lectures 13 and 15, in this setting the Central Limit Theorem gives the normal approximation

$$\hat{P} \approx \mathcal{N}\left(p, \frac{pq}{n}\right), \quad q = 1 - p.$$



When working in a situation like this, we must always check that the sample is large enough for the normal approximation to be reasonable. In our course, we use the convention of $np > 5$ and $nq > 5$.

17.4.2 The difference of two independent sample proportions

Suppose that we now have two binomial experiments:

- Experiment 1 has true success probability p_1 , sample size n_1 , and sample proportion $\hat{p}_1$;
- Experiment 2 has true success probability p_2 , sample size n_2 , and sample proportion $\hat{p}_2$.

Individually,

$$\hat{P}_1 \approx \mathcal{N}\left(p_1, \frac{p_1 q_1}{n_1}\right) \quad \text{and} \quad \hat{P}_2 \approx \mathcal{N}\left(p_2, \frac{p_2 q_2}{n_2}\right),$$

where here $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$. The sample proportions play the same role as the sample means, and we can take their difference to get a new variable $\hat{P}_1 - \hat{P}_2$. If we assume that these two experiments are independent, then the covariance term will again vanish and we are left with a normal distribution of the form:

$$\hat{P}_1 - \hat{P}_2 \approx \mathcal{N}\left(p_1 - p_2, \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}\right).$$

Using this as motivation, we will now describe a method for testing whether or not the two experiments have the same success probability.

17.4.3 Hypothesis tests for differences of success probabilities

To test whether the two success probabilities are equal, we start with the null hypothesis:

$$H_0 : p_1 - p_2 = 0.$$

Under this assumption, the normal distribution mentioned above will be centred at 0:

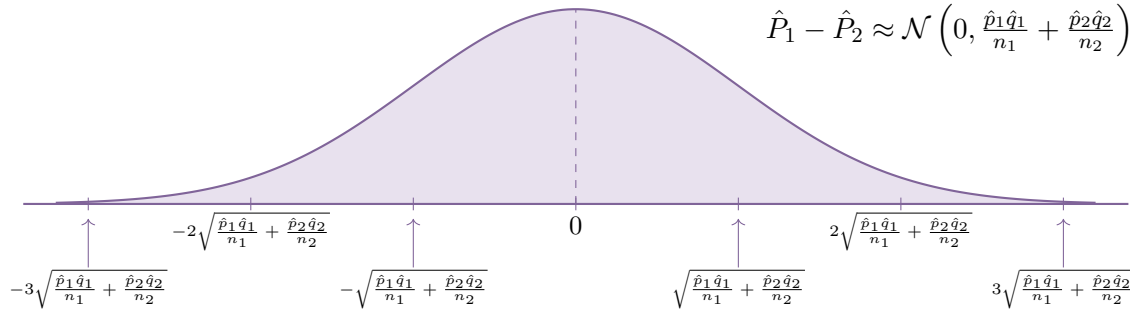
$$\mathcal{N}\left(0, \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}\right).$$

However, we cannot directly use this distribution in testing, because we don't actually know the values of p_i and q_i . Why not? Because the null hypothesis only makes a statement about the

difference $p_1 - p_2$, not about their precise value. Therefore, we have to treat p_1 and p_2 as unknown, and we need to instead estimate their value with the sample proportions instead. This means that our normal distribution will be:

$$\hat{P}_1 - \hat{P}_2 \approx \mathcal{N}\left(0, \frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}\right),$$

where $\hat{q}_1 = 1 - \hat{p}_1$ and $\hat{q}_2 = 1 - \hat{p}_2$. Geometrically, this is again a bell curve centred at 0, but this time we are counting out in terms of the standard error $\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$:



The corresponding z -score starts counting from the centre at 0 in terms of the standard error term above, meaning that the formula will be:

$$z = \frac{\text{observed value} - \text{mean}}{\text{standard error}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}}.$$

This will be the formula to use to compute the appropriate P-value.

Testing two independent binomial probabilities

To test $H_0 : p_1 - p_2 = 0$, use

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}}.$$

The P-value is calculated using $Z \sim \mathcal{N}(0, 1)$.

As with the normal distribution of Section 17.4.1, we should always confirm that the two binomial experiments can be meaningfully approximated by normal distributions in the first place. In our setting, we must always confirm that the four quantities

$$n_1 \hat{p}_1, \quad n_1 \hat{q}_1, \quad n_2 \hat{p}_2, \quad n_2 \hat{q}_2$$

are greater than 5 before proceeding with the approximation.

17.5 Example: Orange Skittles

Suppose that a professor gives bags of Skittles to two classes. He would like to test whether the proportion of orange Skittles in Class 1 is less than the proportion of orange Skittles in Class 2.

The recorded data is:

- Class 1 counted 740 Skittles, of which 139 were orange;
- Class 2 counted 741 Skittles, of which 150 were orange.

We perform the test using a significance level of $\alpha = 0.05$.

Step 1: Gathering the information

The important values are

$$n_1 = 740, \quad r_1 = 139, \quad n_2 = 741, \quad r_2 = 150, \quad \alpha = 0.05.$$

Let p_1 be the true proportion of orange Skittles for Class 1, and let p_2 be the true proportion for Class 2. The null hypothesis says that the proportions are equal, while the suspicion says that the first is smaller:

$$H_0 : p_1 - p_2 = 0, \quad H_1 : p_1 - p_2 < 0.$$

This is a left-tailed test.

Step 2: Calculate the P-value

First compute the two sample proportions:

$$\hat{p}_1 = \frac{r_1}{n_1} = \frac{139}{740} \approx 0.188, \quad \hat{p}_2 = \frac{r_2}{n_2} = \frac{150}{741} \approx 0.202.$$

Therefore,

$$\hat{q}_1 = 1 - \hat{p}_1 \approx 0.812, \quad \hat{q}_2 = 1 - \hat{p}_2 \approx 0.798.$$

Next check the normal approximation conditions:

$$\begin{aligned} n_1 \hat{p}_1 &\approx 740(0.188) = 139 > 5, \\ n_1 \hat{q}_1 &\approx 740(0.812) = 601 > 5, \\ n_2 \hat{p}_2 &\approx 741(0.202) = 150 > 5, \\ n_2 \hat{q}_2 &\approx 741(0.798) = 591 > 5. \end{aligned}$$

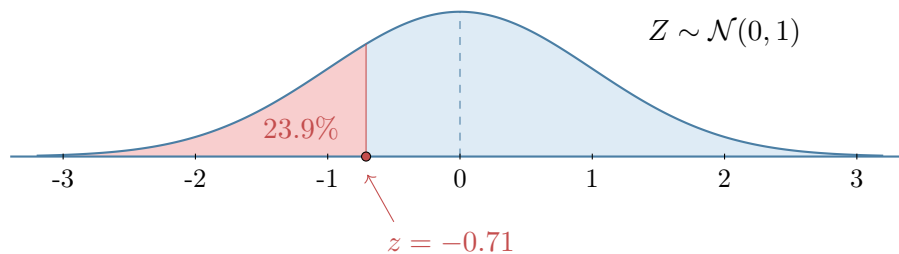
All four values are greater than 5, so the normal approximation is reasonable. Now we can use our information to calculate the z -score:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}} \approx \frac{0.188 - 0.202}{\sqrt{\frac{(0.188)(0.812)}{740} + \frac{(0.202)(0.798)}{741}}} \approx -0.71.$$

Since this is a left-tailed test, the P-value can be read directly from the z -table:

$$\text{P-value} = P(\hat{P}_1 - \hat{P}_2 \leq -0.0146) \approx P(Z \leq -0.71) \approx 0.2389.$$

Geometrically, we have just computed an area in the left tail that is around 23.9% of the total graph:



By construction, this area is the same as the area to the left of our sample data $\hat{p}_1 - \hat{p}_2$ in the sampling distribution built according to the null hypothesis. In words: if the two classes truly had the same proportion of orange Skittles, then there would be about a 23.9% chance of obtaining a difference at least this far in the direction $\hat{p}_1 < \hat{p}_2$.

Step 3: Compare the P-value with α

A comparison gives:

$$0.2389 > 0.05.$$

Therefore, we do not reject H_0 .

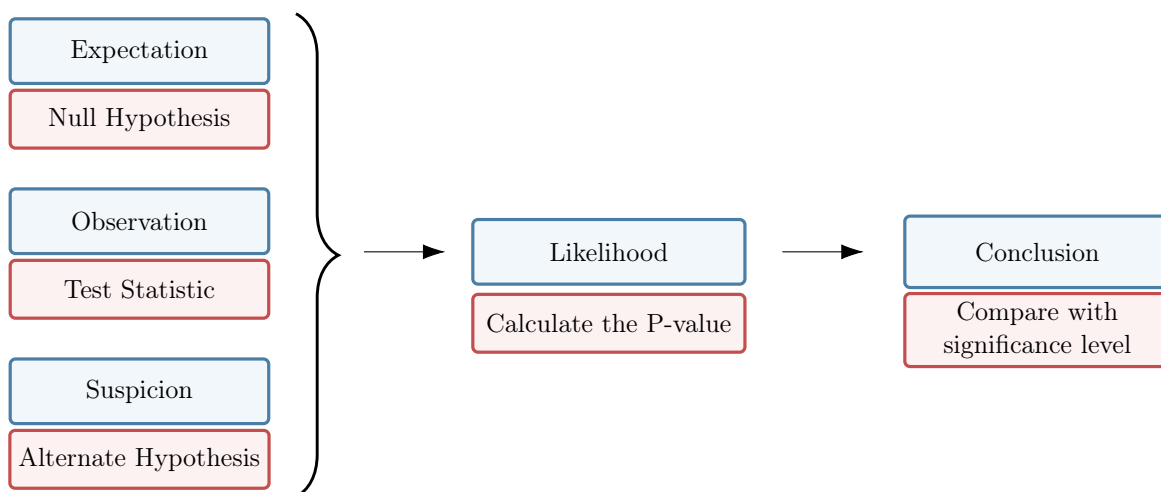
Step 4: Give an interpretation

Interpretation

At the 5% significance level, the samples do not provide convincing evidence that the proportion of orange Skittles in Class 1 is lower than the proportion in Class 2.

17.6 A Summary of Hypothesis Tests

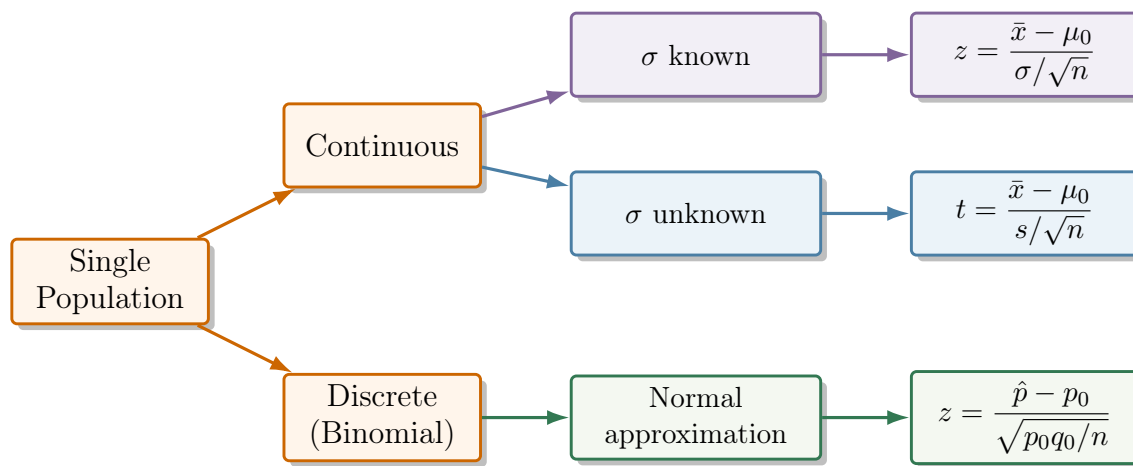
All of the hypothesis tests we have seen follow the same overall pattern:



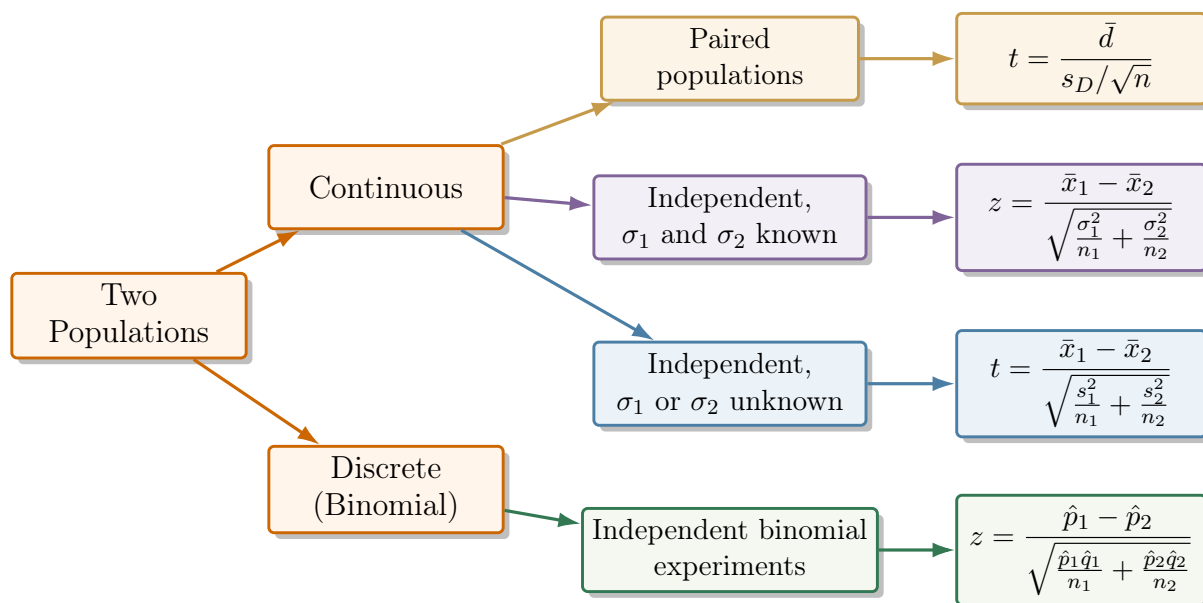
So far in the course, we have studied seven major types of hypothesis test. These are:

1. testing a single population mean μ with known σ ,
2. testing a single population mean μ with unknown σ ,
3. testing a single binomial success probability p ,
4. testing $\mu_X - \mu_Y$ for paired data,
5. testing $\mu_X - \mu_Y$ for independent data with known population standard deviations,
6. testing $\mu_X - \mu_Y$ for independent data with unknown population standard deviations, and
7. testing the difference $p_1 - p_2$ for two independent binomial experiments.

These can be divided into two main camps: tests of a single population and tests of two populations simultaneously. In the case of a single population, the following flow chart organises the different tests.



When testing two populations at the same time, we are met with the following flow chart:



Part V

Correlation and Regression

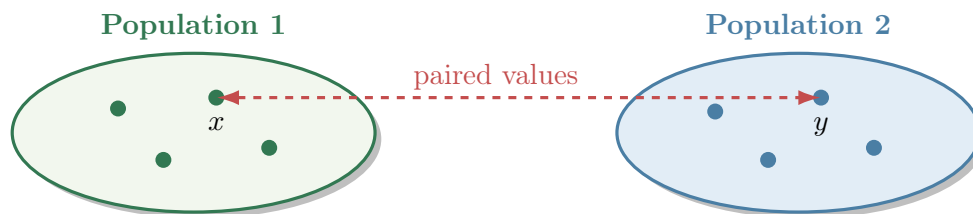
Lecture 18

Linear Correlation

In the previous two lectures, we performed hypothesis tests on two populations simultaneously. In this lecture, we will use paired data in a completely different way: instead of testing population parameters against each other, we look at descriptive statistics of the paired data. We will concern ourselves with *linear correlation*, which asks whether or not two data sets display a pattern resembling a line. We will start our discussion intuitively (that is, visually) and then we will look at ways to describe linear relationships mathematically.

18.1 What is Correlation?

Recall that in Lecture 16 we discussed the notion of *paired* populations. In short, these are populations in which each value in one population has a natural counterpart or partner in the other:



Paired data can arise in several ways:

- **successive experiments on the same group**, such as reaction times before and after caffeine;
- **different experiments on the same object**, such as heart-rate readings from a chest strap and an Apple Watch during the same run;
- **naturally mirrored or matched objects**, such as first-born and second-born twins, or the left and right feet of the same person.

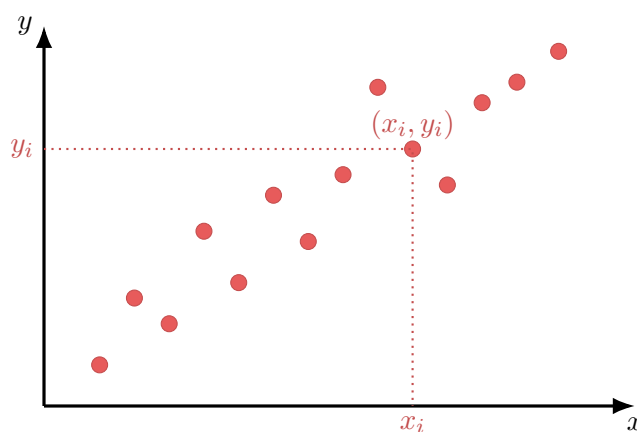
Previously, we used the pairing to form differences of the random variables, which we wrote $D = X - Y$. We then used this to test whether or not X and Y have different means. Today, we are simply going to study *descriptions* of paired data. We start with a visual description known as a scatter diagram.

18.1.1 Scatter diagrams

Suppose that we take a sample of n pairs of data. This can be written as a list:

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n).$$

A *scatter diagram* is a visual representation of this data in which every pair (x_i, y_i) is plotted as an individual point in two-dimensional space. The horizontal axis represents the x -variable and the vertical axis represents the y -variable, so that each point represents a data pair:



In reality, data of this type will occur whenever we measure the same object in two different ways. The hope is that by plotting the paired observations, we can visually inspect the data to see if there are any patterns. As a mathematical setup, we often try to explain the variable Y in terms of the variable X . For that reason, we call x the *explanatory variable*, and we call y the *response variable*.¹

Scatter diagrams

A **scatter diagram** is a graph of paired quantitative data in which each pair (x_i, y_i) is plotted as one point.

The variable x is called the **explanatory variable**, while y is called the **response variable**.

The terminology suggests the type of question we are asking. We suspect that changes in x may help to explain why y changes. For example, suppose x is the speed of a car and y is its stopping distance. It is reasonable to suspect that the speed helps explain the stopping distance, and that the stopping distance responds to changes in speed according to some underlying rule.

18.1.2 Linear correlation

The word “correlation” means that two variables display some kind of pattern together. In principle, the pattern between two variables could have many different shapes, i.e. two sets of data may be correlated in many different ways. In this lecture, we will study the simplest possibility: *linear* correlation. Roughly speaking, the data has linear correlation when the pattern of points approximately follows a straight line.

Linear correlation

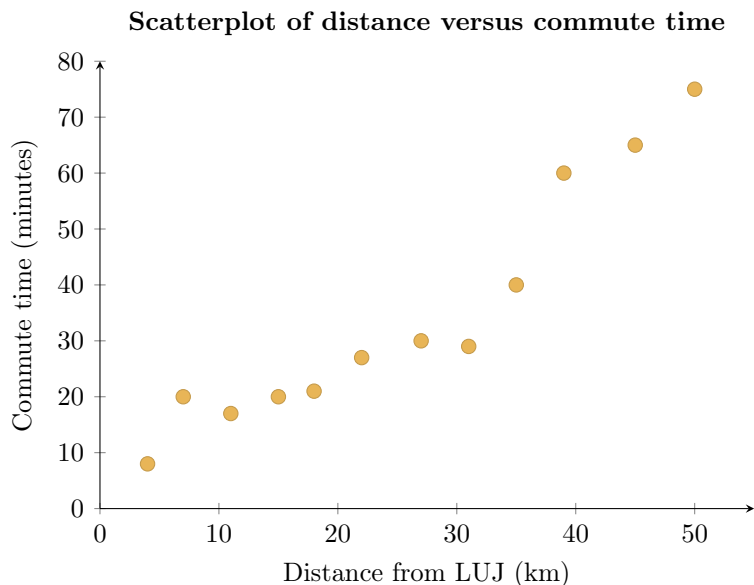
Paired data displays a **linear correlation** when the scatter diagram displays an overall pattern that looks like a straight line.

Linear relationships are extremely common. Examples include the population and GDP of geographical regions, the force required to extend a spring, and oxygen consumption compared with energy expenditure. Later, in Lecture 19, we will introduce *linear regression*: a method for finding the line that best approximates a linear pattern.

As an example, suppose that we took a survey of 12 students, and recorded the distance from their home to campus, and also the time it takes them to commute here. Of course, this data is naturally paired: both values are connected due to the fact that they relate to the same people. The data and the corresponding scatter diagram are depicted below.

¹There are other common names for these two variables in the literature. Sometimes, x is known as the *independent* variable and y is known as the *dependent* variable.

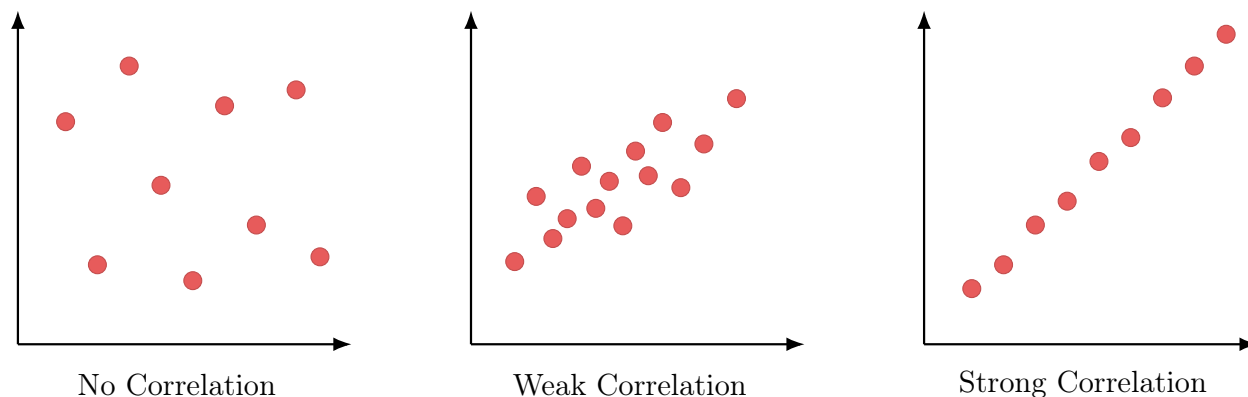
Student	Distance (km)	Time (min)
1	4	8
2	7	20
3	11	17
4	15	20
5	18	21
6	22	27
7	27	30
8	31	29
9	35	40
10	39	60
11	45	65
12	50	75



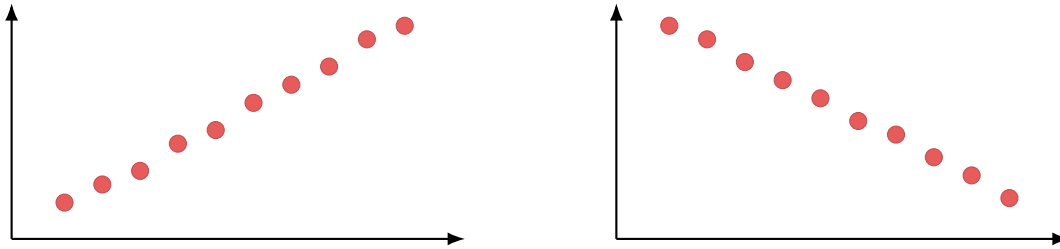
As you can see, the data points display a clear linear pattern: those students who live close to campus have a short travel time, and those students who live far away have a longer travel time. Of course, this data does not arrange itself as a perfect straight line, and this is to be expected: there are many small factors that influence travel time, such as train lines, travel methods and so on. However, based on the data, we can still see a general pattern.

18.1.3 The strength and sign of a correlation

Generally speaking, some scatter diagrams may contain no obvious patterns at all, whereas others very much resemble a line. In natural language, we describe these differences using the “strength” of the correlation. Roughly speaking, a correlation is called “strong” if the pattern is quite clear, and a correlation is called “weak” if there is still *some* pattern, but not very much:



In addition to this, there is another important feature of linear correlation worth mentioning. Consider the two images below:



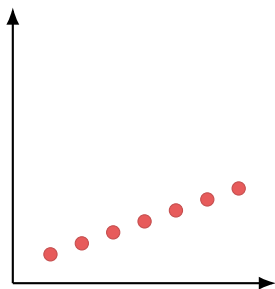
Both of these display a strong linear correlation, yet they are clearly different. In the graph on the left, x and y increase together: when x is low so is y , and when x is high so is y . In contrast, in the graph on the right, the relationship is flipped: when x is low y is high, and when x is high, y is low. In other words, the lines that fit these data sets have gradients of different *signs*. This observation motivates the following terminology.

The sign of a correlation

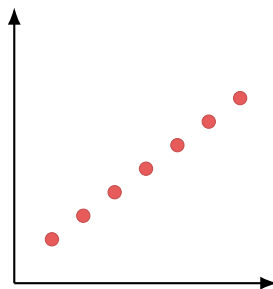
The *sign* of a correlation describes the direction of the general trend:

- **positive correlation:** x and y tend to increase together, so the general trend has a positive gradient;
- **negative correlation:** as x increases, y tends to decrease, so the general trend has a negative gradient.

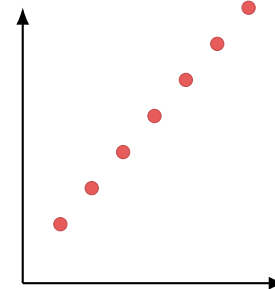
Note that here, the words “positive” and “negative” only refer to the *sign* of the pattern, not to the gradient itself. For instance, all three graphs below display a positive linear correlation, despite having different gradients:



Perfect positive correlation



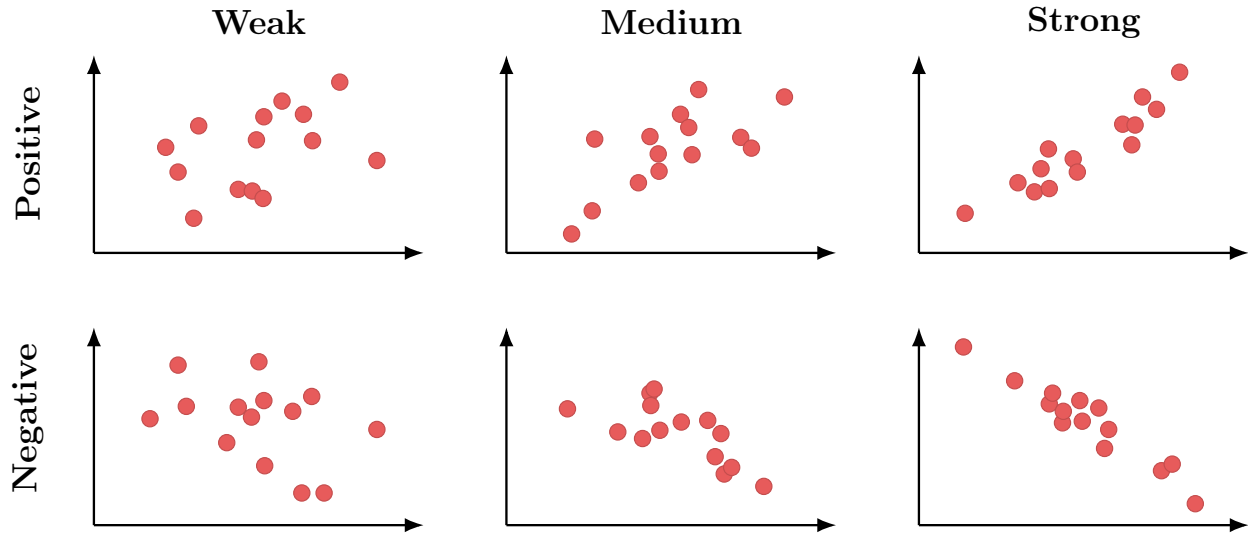
Perfect positive correlation



Perfect positive correlation

18.1.4 Sign versus strength

Sign and strength are simply two different features of linear correlation. As the picture below summarizes, we can find weak, medium, or strong patterns of either sign.



18.2 The Correlation Coefficient

18.2.1 The sample correlation coefficient

Although intuitive, natural language terms like “weak” and “strong” are somewhat vague. To resolve this, we will now introduce a number that quantifies how strong a correlation is. Simply put, we are now going to define a number that measures how strong a correlation is: the stronger the correlation, the higher the absolute value of this number, and the weaker the correlation, the lower the absolute value. This number is typically known as the *sample correlation coefficient*, and is written r . By design, it measures both the strength and the sign of a linear relationship between the variables x and y .

The sample correlation coefficient

The **sample correlation coefficient** r is a unitless number satisfying

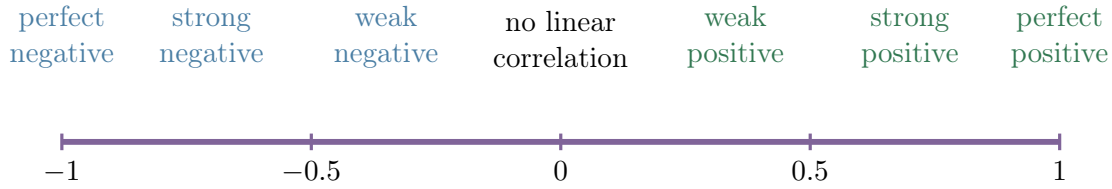
$$-1 \leq r \leq 1.$$

Its sign records whether the linear correlation is positive or negative, while the absolute size of r measures the strength of the linear relationship.

The term “sample correlation coefficient” can be broken down into its three pieces:

- the term “coefficient” tells us that r is a number,
- the term “correlation” refers to the pattern between x and y , and
- the term “sample” refers to the fact that the number r is built from our sample data (x_i, y_i) .

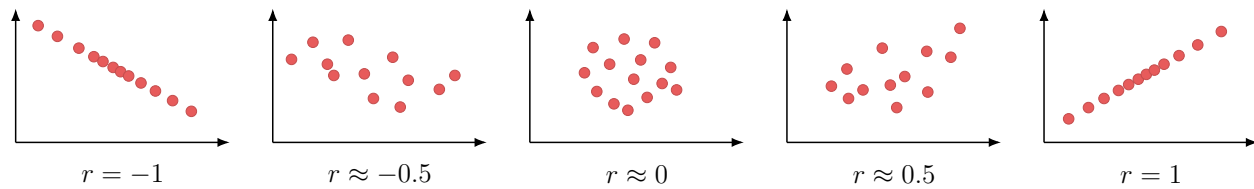
Therefore, the term “sample correlation coefficient” really just means “a number that measures how much our sample follows the pattern of a line”. Using the number r , we can be more precise about the strength of a linear correlation:



In words, the important reference values are:

- $r = 1$ means that the data is arranged perfectly on a straight line with a positive gradient,
- $0 < r < 1$ describes some positive linear correlation,
- $r = 0$ says that the data has no linear correlation at all,
- $-1 < r < 0$ describes some negative linear correlation in the data, and
- $r = -1$ indicates perfect negative linear correlation.

The following diagrams illustrate data with several different approximate values of r :



Even if there is a precise mathematical rule relating x and y , real-world data will rarely follow that rule perfectly. Measurements contain noise, individual differences, and small errors. Perfect linear correlation is therefore best viewed as an idealized theoretical limit.

Although we will not discuss it directly in this course, it is worth mentioning that there is also a population parameter that measures linear correlation. As with the mean or the standard deviation, we denote the population parameter by a Greek letter, this time we use the letter ρ :

	Sample	Population
Mean	$\bar{x}$	μ
Standard deviation	s	σ
Correlation coefficient	r	ρ

The sample correlation coefficient r measures the linear relationship between the pairs of points in a given sample. In contrast, the population correlation coefficient ρ describes the true correlation between all pairs in the full population.

18.2.2 The formula for r

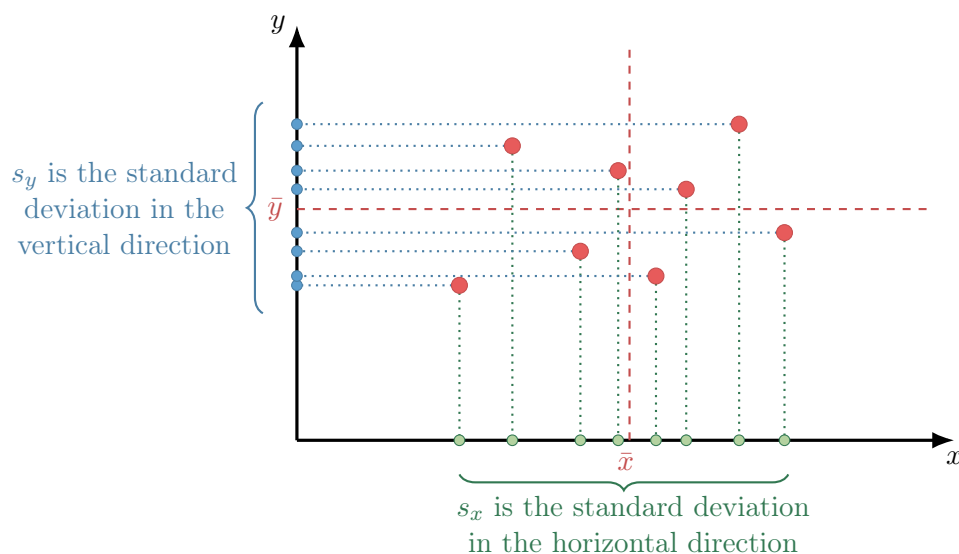
The sample correlation coefficient is defined using the formula:

$$r = \frac{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}}.$$

At first glance, this formula looks overwhelmingly complicated. To better explain what is going on here, we will break down the formula step by step. Notice first that the denominator of r has two sample standard deviations:

$$s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{and} \quad s_y = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}.$$

As a reminder: here we are working with *paired* data, meaning that every data point in our sample is of the form (x_i, y_i) . Each pair can be separated into its two coordinates and considered individually:



So, in the formula for r we are somehow dividing by the individual spreads of x and y . In the numerator of r , we have a new term, known as the *covariance*:

$$\text{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

In fact, this is the sample version of the covariance we mentioned in Lecture 16 when discussing the variance of paired samples. Back then, we merely commented on its existence as a measure of the interrelatedness of x and y . We will now talk at length about what this really means.

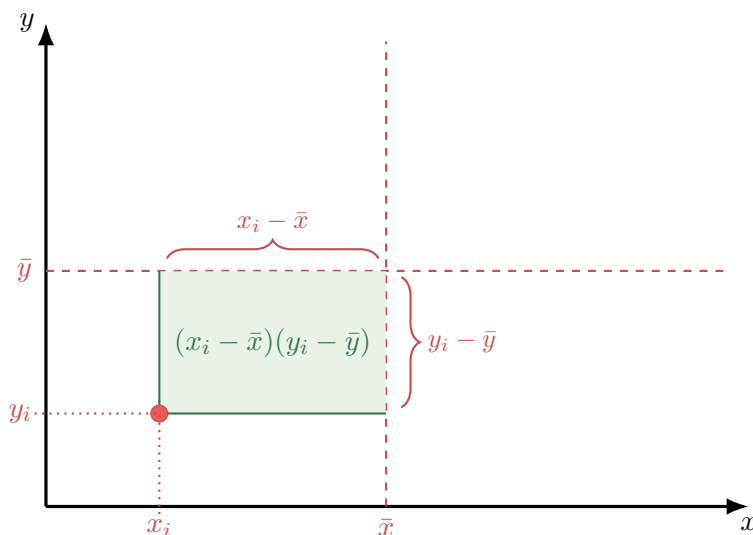
18.2.3 Understanding covariance

Let's examine the formula for covariance:

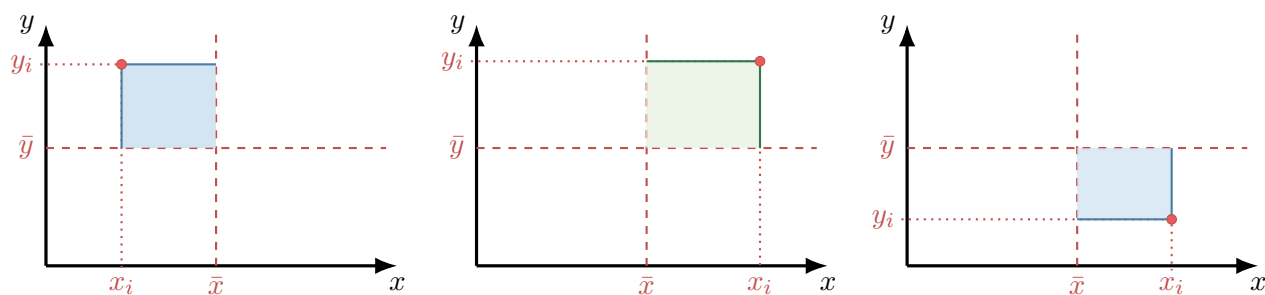
$$\text{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

What's going on here? Firstly, we may notice that on the very outside we are performing a sum over many quantities, and then we divide by $n-1$. This part of the formula looks similar to that of sample variance: in fact, it resembles a sort-of average. So, what are these quantities, exactly?

The individual parts $(x_i - \bar{x})$ and $(y_i - \bar{y})$ are actually measuring distances between points: they measure how far x_i is from $\bar{x}$ and how far y_i is from $\bar{y}$. These are two lengths, and multiplying them together forms an *area*:



In the picture above, we have deliberately drawn our data point (x_i, y_i) in the bottom left quadrant of the chart. In this situation, x_i is to the *left* of $\bar{x}$, meaning that $x_i < \bar{x}$ and therefore $x_i - \bar{x} < 0$. Similarly, the component y_i is *below* the mean $\bar{y}$, meaning that $y_i < \bar{y}$ and therefore $y_i - \bar{y} < 0$. Therefore our expressions $(x_i - \bar{x})$ and $(y_i - \bar{y})$ are distances that carry a *sign*.² If both of these signed distances are negative, then multiplying them together gives an overall positive answer. As for the other three quadrants, we have different overall signs of the areas:



The sign of each rectangle depends only on which side of the two means the point lies:

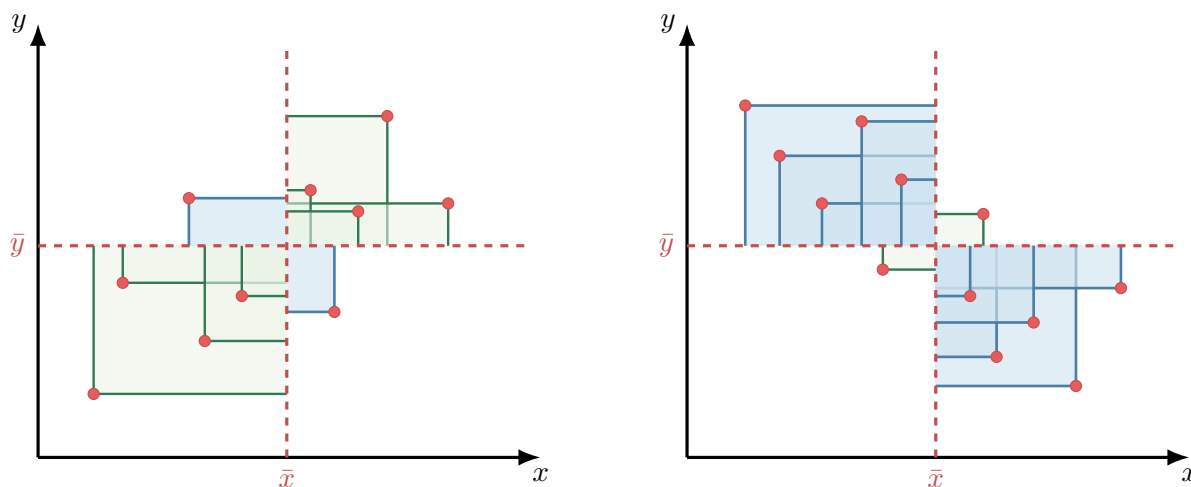
Location of (x_i, y_i)	$(x_i - \bar{x})(y_i - \bar{y})$	Signed Area
$x_i < \bar{x}, y_i < \bar{y}$	$(-)(-)$	positive
$x_i < \bar{x}, y_i > \bar{y}$	$(-)(+)$	negative
$x_i > \bar{x}, y_i > \bar{y}$	$(+)(+)$	positive
$x_i > \bar{x}, y_i < \bar{y}$	$(+)(-)$	negative

If $x_i = \bar{x}$ or $y_i = \bar{y}$, then the signed area is 0.

²These are often called *signed deviations* from the mean.

The covariance adds together the signed areas of all of these rectangles, and then divides by $n - 1$. In other words, covariance is an average measure of whether the two variables lie on the same side of their means, and how far they lie from those means.

In fact, this “sort-of average of signed area” allows us to recover the sign of a possible correlation. In the diagram below, we have created the signed rectangular areas for each of the data points. Here we represent the areas with a positive sign using green, and the areas with a negative sign using blue.



Notice that the diagram on the left is somewhat positively correlated: most of the data lies in the bottom left and top right quadrants of the diagram. Visually, there is more green in the diagram than there is blue, and therefore the covariance should be positive. In contrast, the diagram on the right is negatively correlated: most of the data points are in the top left or bottom right of the diagram. Visually, the signed rectangles are mostly blue, meaning that the covariance will be negative. To summarize:

- if the diagram contains more green area than blue, then the sum of the signed areas is positive, and $\text{cov}(x, y) > 0$,
- if the diagram contains more blue area than green, then the sum of the signed areas is negative and $\text{cov}(x, y) < 0$, and
- if the positive and negative areas largely cancel in the sum, then the covariance is close to 0.

Going back to the formula, each symbol can be broken down as follows:

$$\text{cov}(x, y) = \frac{1}{n-1} \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

Horizontal signed distance between point and $\bar{x}$ Vertical signed distance between point and $\bar{y}$

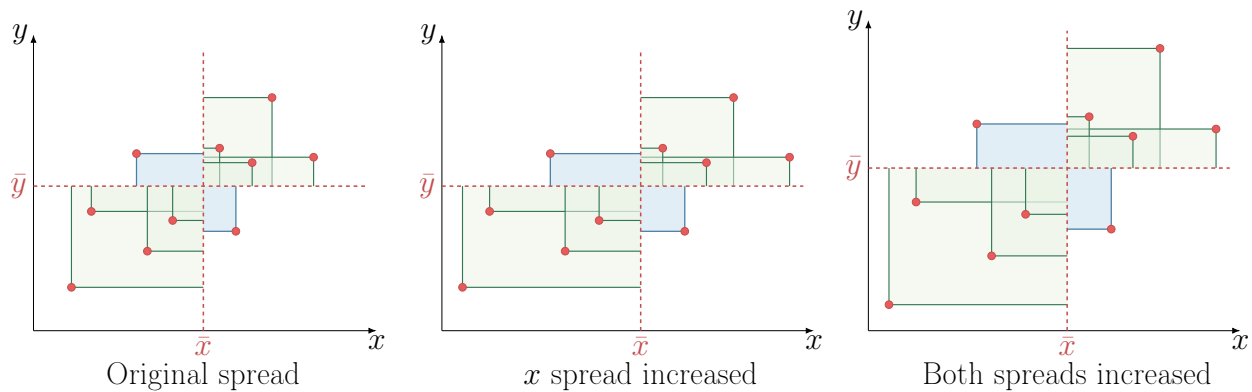
Average over all areas Signed area of a rectangle

What covariance measures

Covariance measures whether the two variables tend to lie on the same side of their respective means.

18.2.4 Understanding the formula for r

By construction, the formula for covariance cleverly uses areas of rectangles to keep track of the sign of any (potential) linear correlation. However, covariance on its own still depends on the overall spread of the data in question. To see this, let's take a data set and manually increase the spread in both the x and y directions:



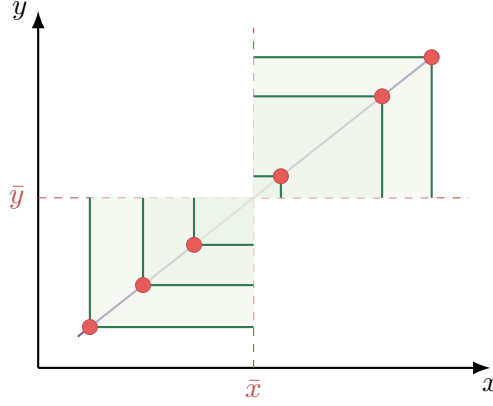
Notice that as we increase the spread of the data, the average distance of each point from either $\bar{x}$ or $\bar{y}$ will also increase, meaning that the average area of each rectangle will also increase. However, generally speaking, increasing the spread in this way doesn't turn a green rectangle into a blue one, or vice versa. If we wanted to care only about the interrelatedness of green and blue rectangles, rather than their raw size, then we could divide by the spread of the data to get rid of their absolute sizes. This is the formula for r :

$$r = \frac{\text{cov}(x, y)}{s_x s_y}.$$

Roughly speaking, the standard deviations s_x and s_y measure the spread of the data in the horizontal and vertical directions, respectively. If we divide the covariance by these two terms, then we get rid of the overall scale of the diagram above, so that increasing the spread of the data doesn't actually change the answer. Instead, all that matters is the relative pattern of the areas formed from the data.

18.2.5 Recovering $r = 1$ for Perfect Positive Correlation

Suppose now that we have data that is perfectly arranged on a line. We will now demonstrate how our formula for r gives a value of $+1$. In the image below, we have plotted six points on a straight line with positive gradient, and have drawn all of their signed areas to the side.



Geometrically, all of these rectangles have the same *aspect ratio*:

$$\bar{y} - y_1 \left\{ \underbrace{\hspace{1.5cm}}_{\bar{x} - x_1} \right\} \sim \underbrace{\hspace{1.5cm}}_{\bar{x} - x_2} \bar{y} - y_2 \Rightarrow \frac{y_1 - \bar{y}}{x_1 - \bar{x}} = \frac{y_2 - \bar{y}}{x_2 - \bar{x}}$$

We will call this aspect ratio k . If all the points (x_i, y_i) lie on the same line, then each of these ratios equals k .³ Rearranging, for all $i = 1, \dots, n$ we have:

$$\frac{y_i - \bar{y}}{x_i - \bar{x}} = k \Rightarrow y_i - \bar{y} = k(x_i - \bar{x}).$$

If we substitute this into the formula for covariance, we happen to recover an expression containing the sample variance for x :

$$\text{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})k(x_i - \bar{x}) = \frac{k}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = ks_x^2.$$

Similarly, replacing all occurrences of $y_i - \bar{y}$ with $k(x_i - \bar{x})$ in the formula for s_y happens to give:

$$s_y = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n k^2 (x_i - \bar{x})^2} = ks_x,$$

where the last equality uses $k > 0$. Substituting into the formula for r gives

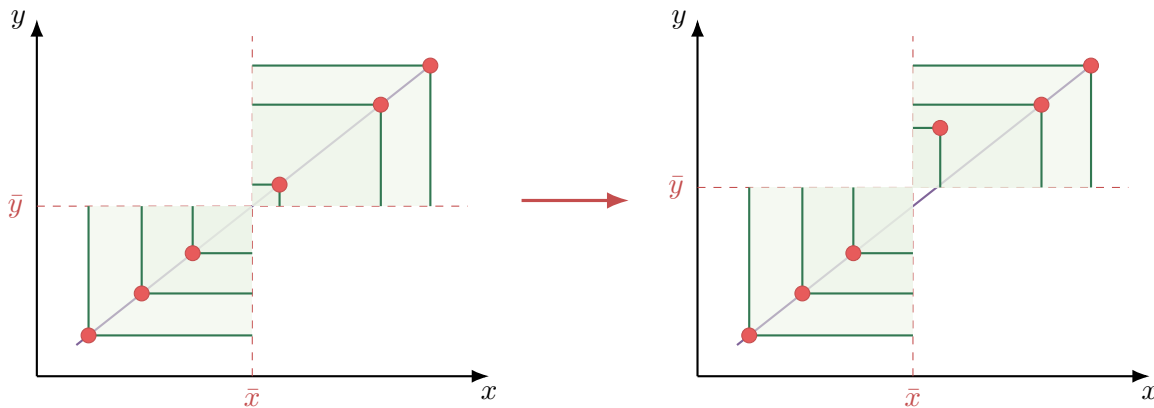
$$r = \frac{ks_x^2}{s_x(ks_x)} = 1,$$

as expected.

³Technically, this does not work whenever the point has either $x_i = \bar{x}$ or $y_i = \bar{y}$. In those cases, the “areas” would collapse to a one-dimensional line.

What happens when one point moves?

To get a better idea of what r is doing geometrically, let's take a point and shift it up a bit, so that the data is no longer linear:



We would expect that doing this would decrease r from 1 to something slightly smaller. Visually, we can see that manually moving one point off the line vertically will drag up the overall mean $\bar{y}$. Therefore, all of the distances $y_i - \bar{y}$ will also change. In the process, these rectangles will no longer have the same aspect ratio: in the image above, the rectangles now have different shapes.

Algebraically, shifting a single point vertically adds a small piece to the y component of some fixed point, e.g. $(x_4, y_4) \rightarrow (x_4, y_4 + \varepsilon)$, where ε is some positive number. Roughly speaking, making this change to one point will make the covariance grow a little bit, but the formula for the standard deviation s_y contains lengths *squared*, i.e. pieces of the form $(y_i - \bar{y})^2$. When we shift a single point by ε , this overall small change cascades into a relatively big change because of the exponent. This causes the s_y term to grow faster than the $cov(x, y)$ term, and therefore the ratio r shrinks to be smaller than 1.

18.2.6 Example: calculating r

We now return to the commute-time data from Section 1. The sample means are

$$\bar{x} \approx 25.333 \quad \text{and} \quad \bar{y} \approx 34.333.$$

In order to compute r , we need to sum over several types of terms. The table below records the relevant information:

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
4	8	-21.333	-26.333	455.111	693.444	561.778
7	20	-18.333	-14.333	336.111	205.444	262.778
11	17	-14.333	-17.333	205.444	300.444	248.444
15	20	-10.333	-14.333	106.778	205.444	148.111
18	21	-7.333	-13.333	53.778	177.778	97.778
22	27	-3.333	-7.333	11.111	53.778	24.444
27	30	1.667	-4.333	2.778	18.778	-7.222
31	29	5.667	-5.333	32.111	28.444	-30.222
35	40	9.667	5.667	93.444	32.111	54.778
39	60	13.667	25.667	186.778	658.778	350.778
45	65	19.667	30.667	386.778	940.444	603.111
50	75	24.667	40.667	608.444	1653.778	1003.111
Total		0	0	2478.667	4968.667	3317.667

Using the simplified computational formula,

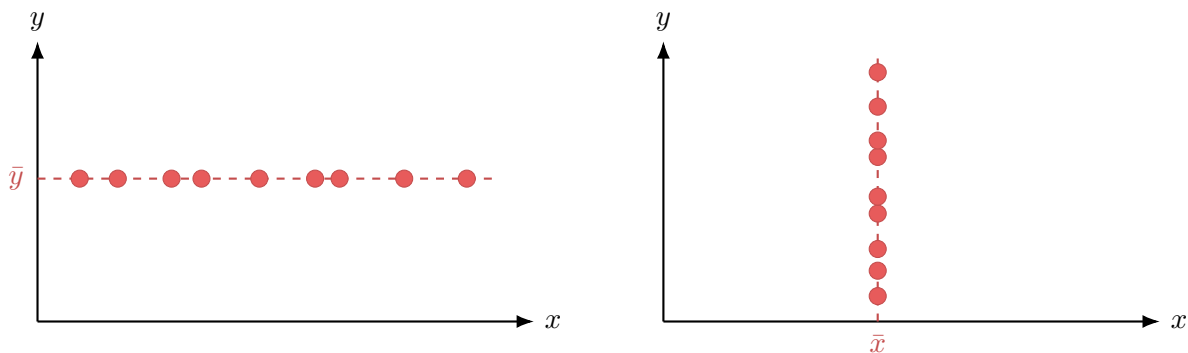
$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{3317.667}{\sqrt{2478.667} \sqrt{4968.667}} \approx 0.945.$$

This is a strong positive linear correlation, meaning that students who live farther from LUJ tend to have longer commute times.

18.3 Limitations of Correlation

18.3.1 Cases where r is undefined

Consider the following two data sets:



Clearly, these data sets both lie on a straight line. However, it turns out that the sample correlation coefficient r is undefined. In either case, one of the two directions has *zero* spread, meaning that

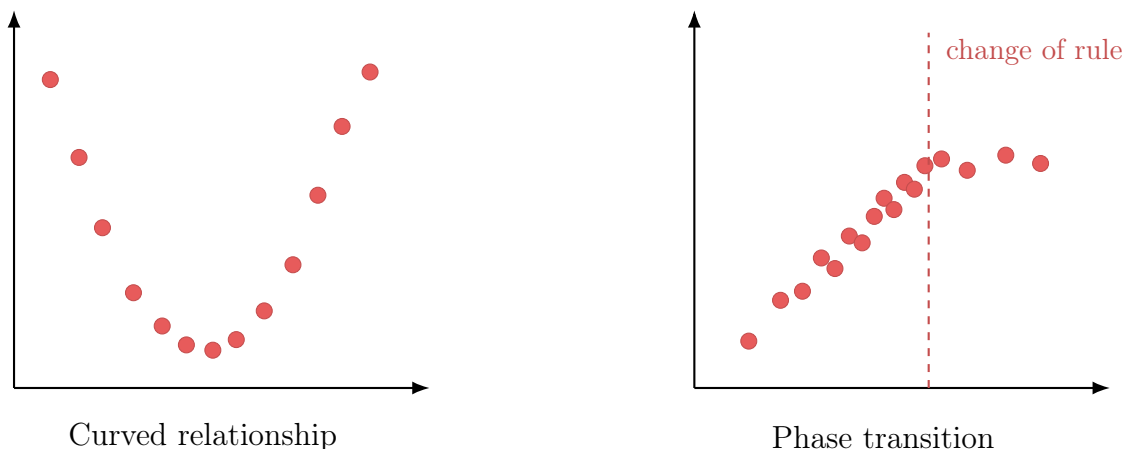
either every y_i equals $\bar{y}$ or every x_i equals $\bar{x}$. This causes the covariance $\text{cov}(x, y)$ and either s_y or s_x to vanish, giving:

$$r = \frac{\text{cov}(x, y)}{s_x s_y} = \frac{0}{0},$$

which is undefined.

18.3.2 Non-linear correlation

The correlation coefficient r only measures *linear* correlation. In particular, some data sets may be correlated in other ways, yet still have a small value of r . In other cases, it may be that the relationship is linear for a while, yet then changes to some other type of relationship:



In scientific applications, a relationship may be approximately linear only up to a certain point, after which a new rule takes over. In situations like this, the correlation coefficient may return a high value anyway, even though it is not technically representative of the full system.

18.3.3 Correlation without causation

Just because there is an apparent correlation in data does not mean that there is necessarily a causal link between the two variables. For instance, if X and Y are correlated, then

- X may cause Y , or
- Y may cause X , or
- X and Y may be correlated by pure coincidence, or
- there may be a third variable that influences both X and Y .

The slogan here is:

Correlation does not imply causation.

As an example, ice cream sales are often correlated with suncream sales. This does not mean that ice cream sales cause suncream sales or vice versa; instead, the weather is a third variable that influences both.

Correlation does not imply causation

The value of r can describe the existence, sign, and strength of a linear pattern. However, it cannot give that pattern a causal interpretation.

Lecture 19

Linear Regression

In Lecture 18, we studied paired quantitative data by plotting the observations (x_i, y_i) on a scatter diagram. We then used the sample correlation coefficient r to describe the sign and strength of a linear relationship. Correlation itself merely tells us whether the points approximately fit a straight line. But, it does not actually construct the line itself. Today we will build this line, known as the *least-squares regression line*. We will start by identifying the key features of the regression line, and then we will use it to make predictions. Finally, we will introduce the *coefficient of determination*, which measures the quality of the regression line as a predictor.

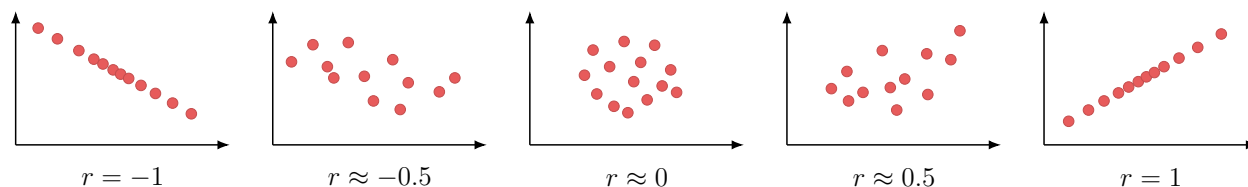
19.1 Linear Regression

19.1.1 The correlation coefficient

Suppose that we have n paired observations:

$$(x_1, y_1), \quad (x_2, y_2), \quad \dots, \quad (x_n, y_n).$$

In the previous lecture, we started to study the descriptive statistics of paired data. In particular, we looked at *linear correlation*, which described how closely the paired data followed a straight line. In the process, we introduced the *correlation coefficient*, which was a number that quantified the linear correlation between x and y . This number, which we denoted by r , ranged from -1 to 1 , and the value of r dictated the strength and sign of the linear correlation present:



Algebraically, the sample correlation coefficient can be calculated using the formula:

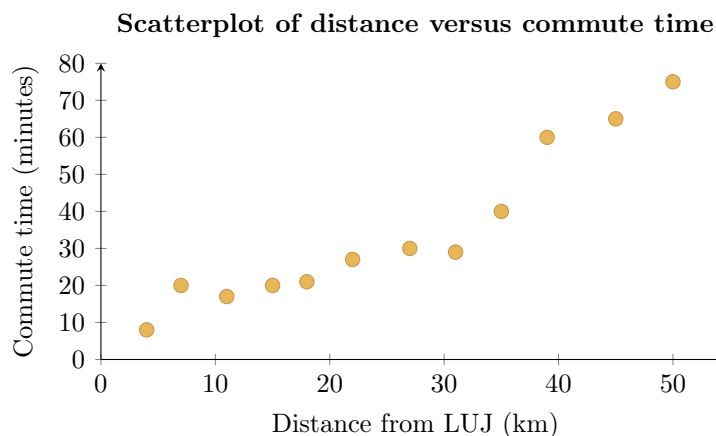
$$r = \frac{\text{cov}(x, y)}{s_x s_y},$$

where here the numerator is the covariance (a measure of the interrelatedness of the two variables) and s_x and s_y are the sample standard deviations in the x and y directions.

19.1.2 Example: distance and commute time

Throughout this lecture, we will use the following data set as a regular example. Suppose that twelve students record their distance from LUJ and their commute time to campus.

Student	Distance (km)	Time (min)
1	4	8
2	7	20
3	11	17
4	15	20
5	18	21
6	22	27
7	27	30
8	31	29
9	35	40
10	39	60
11	45	65
12	50	75



For this data, we saw last time that

$$\bar{x} \approx 25.333, \quad \bar{y} \approx 34.333, \quad s_x \approx 15.011, \quad s_y \approx 21.253, \quad \text{and} \quad \text{cov}(x, y) \approx 301.606.$$

Therefore,

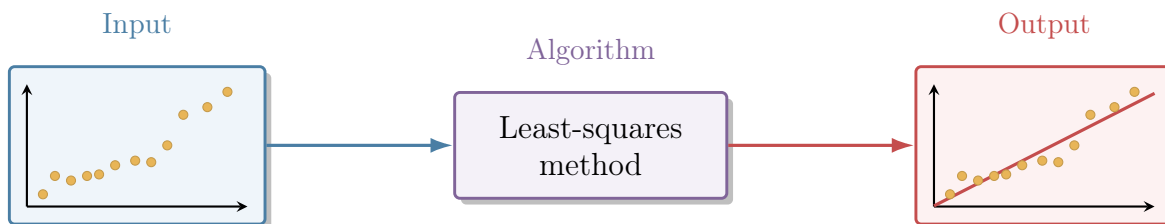
$$r = \frac{301.606}{(15.011)(21.253)} \approx 0.9454,$$

which indicates a strong, positive linear relationship between the two variables.

19.2 The Least-Squares Method

19.2.1 Regression as an algorithm

An *algorithm* is a fixed procedure that takes an input and reliably produces an output. In today's lecture, we will present an algorithm that takes paired data as an input, and produces the “line of best fit” as an output. For reasons that will be clear later, we will call this the *least-squares method*. Schematically, we have:



Recall that any non-vertical straight line can be fully described using two numbers:

- the *gradient*, which describes the rate of change of the line (i.e. the steepness), and
- the *y-intercept*, which is the value of the line when $x = 0$.

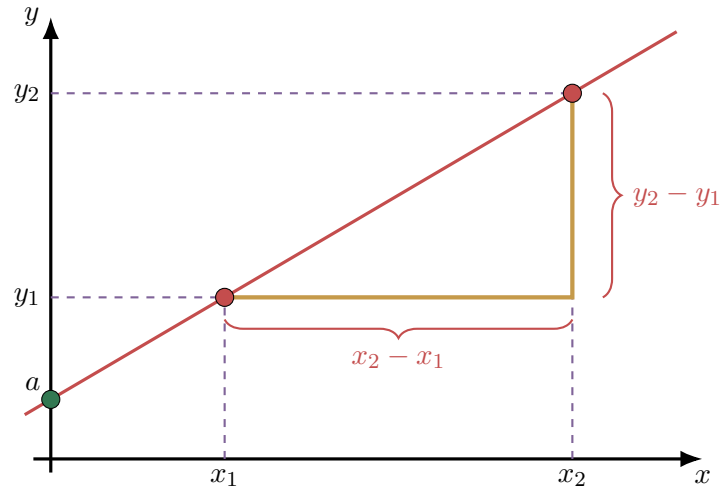
We will write our line using the notation:

$$\hat{y} = a + bx,$$

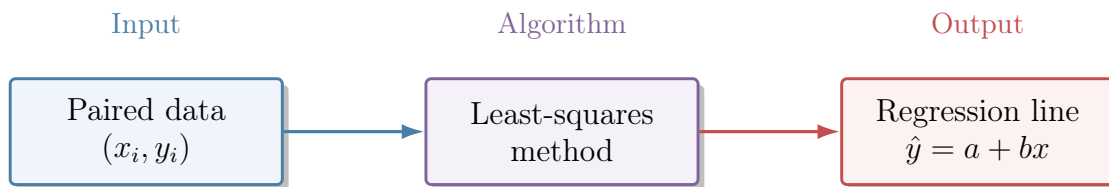
where the hat notation for the y will make sense later. Here, a is the value of the y -intercept, and b is the gradient of the line. The gradient can be calculated by picking two points on the line and using the formula:

$$b = \frac{y_2 - y_1}{x_2 - x_1},$$

that is, we take the vertical change and divide it by the horizontal change.

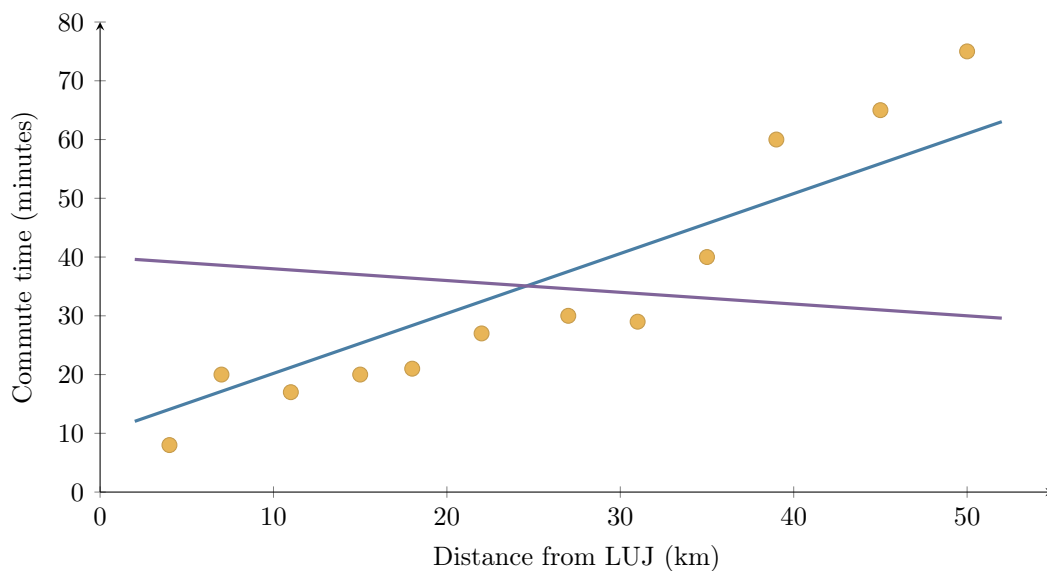


With this notation in mind, the output of the least-squares method is actually a pair of numbers a and b , so that the resulting line best fits the input data. Such a line is called a *regression line*. So, a better picture of our algorithm might be:



19.2.2 The least-squares criterion

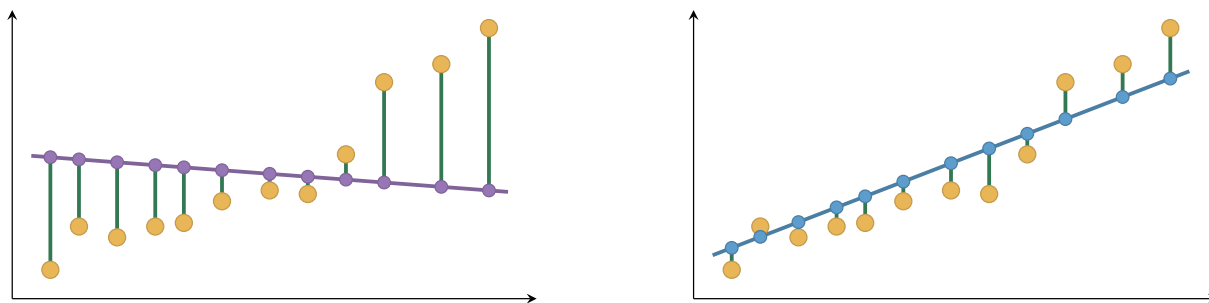
Using the data from Section 1, let's consider two lines:



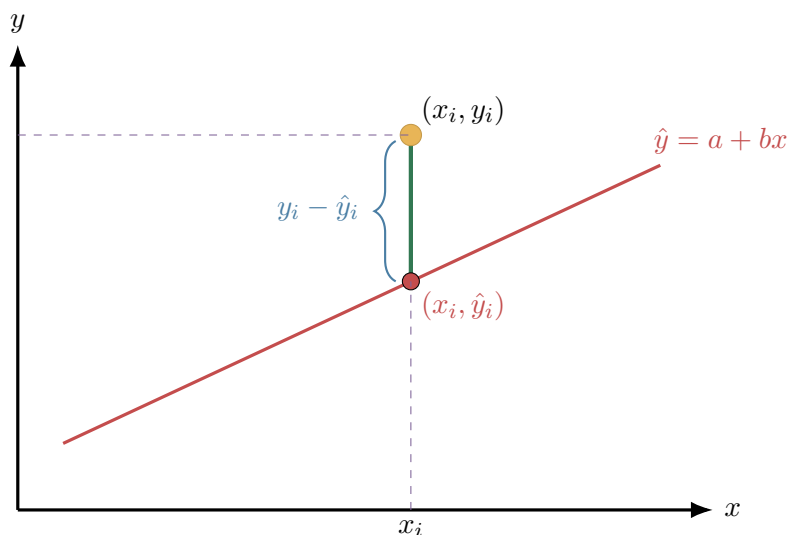
Clearly the blue line fits the data better than the purple line. But, mathematically speaking, why is this the case? In other words, what is it about the blue line that is better than the purple? One answer to this question can be found by looking at the vertical distances between the data and the regression line. As a reminder: each data point is of the form (x_i, y_i) . Therefore, to describe the vertical distance from each of these points to the line, we can input the value x_i into the equation for the regression line $\hat{y}$ to find the output of the line. The value on the regression line above x_i is given by the number:

$$\hat{y}_i = a + bx_i.$$

Taken together, this defines a point $(x_i, \hat{y}_i)$ sitting on the line that is directly above x_i . Ideally, this point should be close to the observed data point (x_i, y_i) . As we can see, in the case of the purple line, these vertical distances to the real data are much larger than those for the blue line:



It is in this sense that we intuitively think of the blue line as a better fit: the total of all of the lengths of these vertical lines is smaller than that of the purple line. Generally, these signed differences are called *residuals*, and they are simply the value of the difference $y_i - \hat{y}_i$:



Notice that values y_i lying above or below the line will cause $y_i - \hat{y}_i$ to have either a positive or negative sign, respectively. Therefore, if we were to naively add up all of these residuals to try and create some kind of “total error”, then there is a risk that they may cancel each other out,

jeopardizing our measure of “good fit”. To repair this issue, we first square all of the quantities $y_i - \hat{y}_i$ before adding them up.¹ As such, the correct measure of “total error” is actually the total *squared* error:

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (a + bx_i))^2.$$

Different values of a and b produce different lines, and therefore different values of this sum. In the example above, the purple line’s total squared error happens to be much larger than the blue line’s. With this in mind, we can define the “line of best fit” as the line which minimizes this total squared error.

The least-squares criterion

The least-squares regression line $\hat{y} = a + bx$ is chosen so that

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2$$

is as small as possible among all straight lines.

19.2.3 The formulas for the regression line

The exact derivation of the values of a and b requires the use of calculus. Since this is just an introductory course, we will not spend time deriving the regression line here.² Instead, we will merely accept that the resulting formulas for a and b are:

$$b = \frac{\text{cov}(x, y)}{s_x^2} \quad \text{and} \quad a = \bar{y} - b\bar{x}.$$

These two formulas allow us to write the regression line $\hat{y} = a + bx$.

Computing the regression line

Given paired data (x_i, y_i) , calculate

$$b = \frac{\text{cov}(x, y)}{s_x^2} \quad \text{and} \quad a = \bar{y} - b\bar{x}.$$

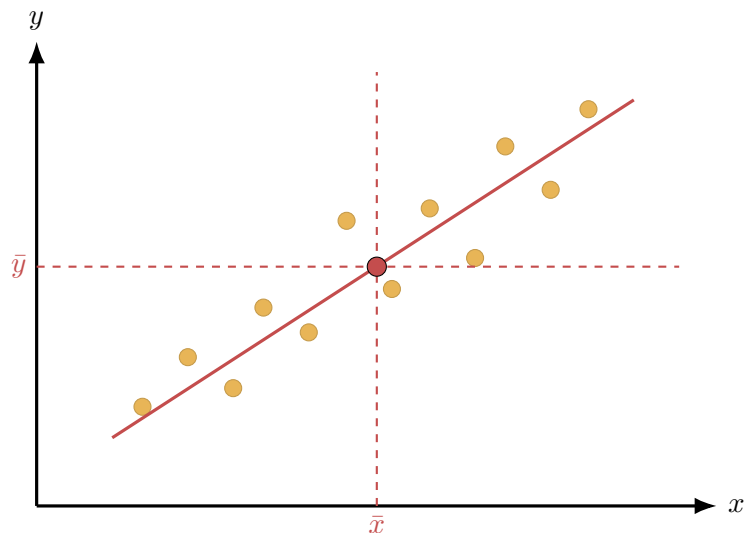
Then write the least-squares regression line as

$$\hat{y} = a + bx.$$

As a small corollary, we notice that the point $(\bar{x}, \bar{y})$ always lies on the least-squares regression line:

¹This is the same reason that we square the quantities $y_i - \bar{y}$ in the formula for the sample variance s_y^2 .

²For the curious: we can take the expression $\sum_{i=1}^n (y_i - (a + bx_i))^2$ and differentiate it with respect to a and b . To find the minimum, we set these derivatives equal to zero and solve the pair of resulting linear equations.



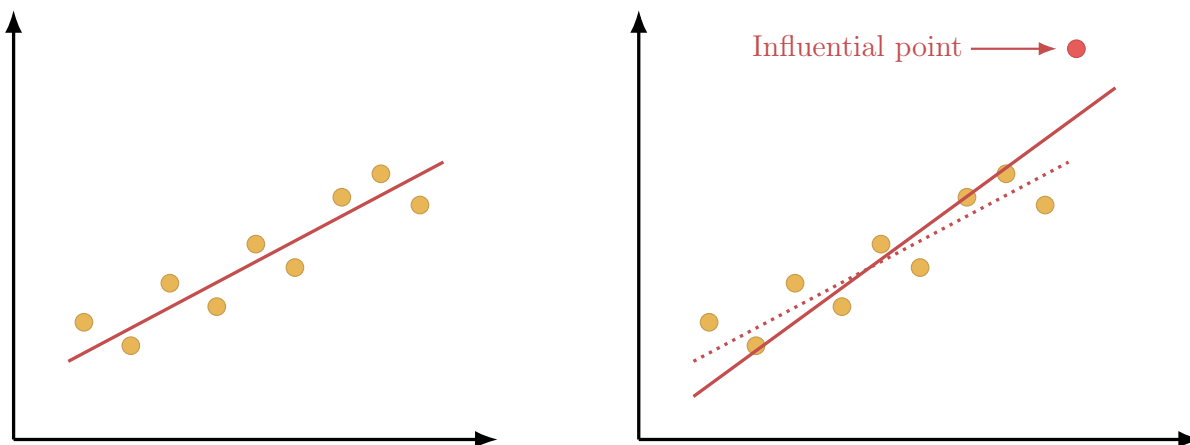
To check this, we may substitute $x = \bar{x}$ into the equation and see:

$$\hat{y} = a + b\bar{x} = (\bar{y} - b\bar{x}) + b\bar{x} = \bar{y}.$$

Note that it doesn't matter which data we start with – it is simply a consequence of the construction of the regression line.

19.2.4 Influential points

A point can have a large effect on the regression line when it lies far from the main cloud of data, especially in the horizontal direction. Such an observation is called an *influential point*.



An influential point is not necessarily an issue. However, because it can radically change a regression line, the inclusion of influential points may affect any predictions that we may want to make using our regression line as a model. With this in mind, we should always be cautious when dealing with points of this type.

19.2.5 Example: computing the regression line

We now calculate the least-squares regression line for the commuting data of Section 1. From Lecture 18, we already know

$$\bar{x} \approx 25.333, \quad \bar{y} \approx 34.333, \quad s_x \approx 15.011, \quad \text{cov}(x, y) \approx 301.606.$$

We first calculate the gradient b :

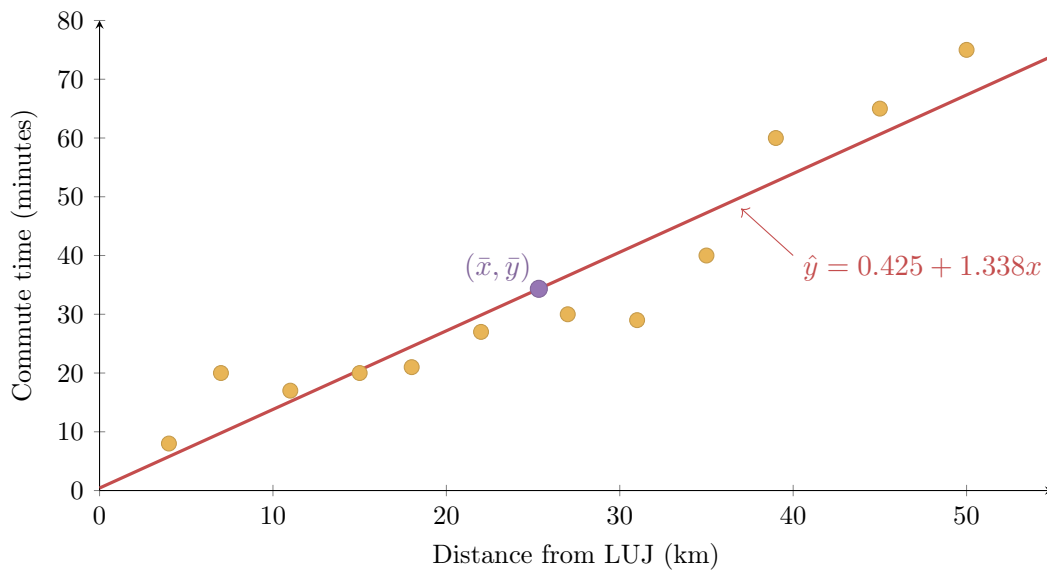
$$b = \frac{\text{cov}(x, y)}{s_x^2} = \frac{301.606}{(15.011)^2} \approx 1.33849.$$

We may now use this value to calculate the y -intercept a :

$$a = \bar{y} - b\bar{x} = 34.333 - (1.33849)(25.333) \approx 0.425.$$

Putting these together, the regression line becomes:

$$\hat{y} = 0.425 + 1.338x.$$



Interpretation of the gradient

The gradient is approximately 1.338. According to the fitted linear model, each additional kilometre of distance from LUJ is associated with an increase of approximately 1.34 minutes in predicted commute time.

The intercept is approximately 0.425 minutes. In this example, the intercept is a theoretical requirement that is mainly needed to position the line correctly, and it should not automatically be given a real-world interpretation.

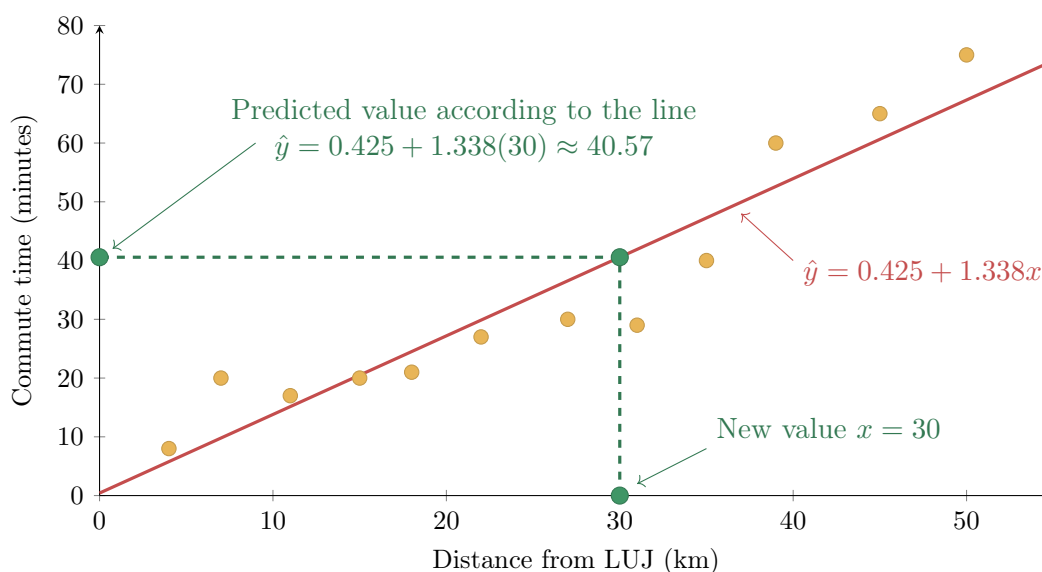
19.3 Using the Regression Line for Predictions

19.3.1 Making a prediction

Aside from visual summaries, one key benefit of a regression line is that it can act as a model that allows us to make predictions. Using the example of commute times and distance travelled, let's suppose that a student moves and their new apartment is 30 km away from campus. In this case, we can use the regression line to predict the approximate commute time by plugging the value $x = 30$ into the equation:

$$\hat{y} = 0.425 + 1.338(30) \approx 40.57.$$

Thus, the model predicts a commute time of approximately 41 minutes. Visually, our prediction corresponds to the following:

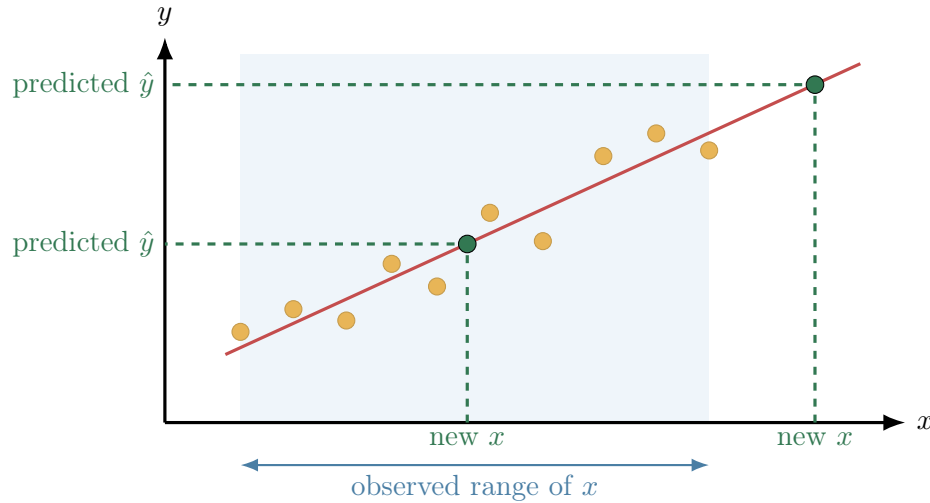


19.3.2 Interpolation and extrapolation

There are two fundamentally different types of prediction, known as *interpolation* and *extrapolation*.

Interpolation and Extrapolation

Interpolation makes predictions based on new values that lie within the observed range of x , whereas **extrapolation** makes predictions based on new values of x that are outside of the range of the observed values.



Typically, interpolation is usually the safer use of a regression line, because the model is being applied in a region where the data was actually observed. In contrast, extrapolation carries the risk that we use a model outside an appropriate regime. There may be other variables that affect our data at extreme values of x , and this is not accounted for in the regression line alone.

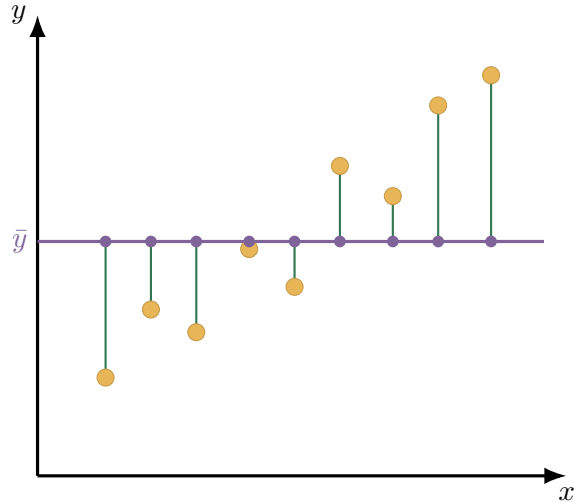
Warning about extrapolation

A linear pattern observed over one range of x may not continue outside that range. Extrapolated predictions can therefore be unreliable, even when the original correlation is strong.

19.3.3 The coefficient of determination

The regression line summarizes the scatter diagram and allows us to predict values of y based on new values of x . We will now introduce a number that quantifies how good this fit is. In order to do so, we first need to introduce a baseline predictor.

Suppose that we ignored the data in x completely. If we were to do this, the best general prediction for any value of y would be to use the sample mean $\bar{y}$. Of course, this wouldn't be correct most of the time, but *on average* it would be the best thing that we can do. Using the mean $\bar{y}$ to predict every value of y would be like picking a horizontal line $y = \bar{y}$:

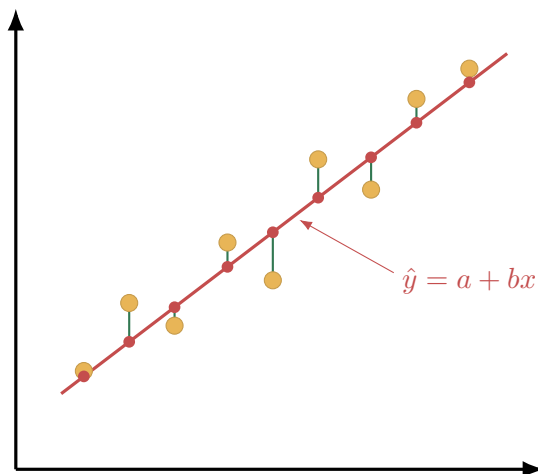


We can measure how good this prediction method is by looking at the total amount of error created between the line and the observed data points. Again, making sure to square the values first so as to avoid accidental cancellation, the total squared error coming from our naive prediction will be:

$$\sum_{i=1}^n (y_i - \bar{y})^2.$$

This quantity is known as the *total squared error*, and it will serve as our baseline.

We are now going to measure how good the regression line is at predicting by comparing its error with the naive prediction. Suppose now that instead of simply guessing $\bar{y}$ for every value, we use the paired input of x_i together with the regression line. This will again give some errors:



The total squared error obtained from using the regression line as a predictor is given by the quantity

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

By construction, this is the quantity that is minimised by the least-squares criterion. Therefore, it will always be the case that

$$\sum_{i=1}^n (y_i - \bar{y})^2 \geq \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

If we wanted to compare the relative size of these two numbers, we can always take their ratio:

$$\frac{\text{total squared error using regression line}}{\text{total squared error using the sample mean}} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

If the regression line is a good predictor, then it should produce a small error compared to the error from using the sample mean. Therefore, we would expect the above ratio to be *small*. Conversely, if the regression line is bad as a predictor, then it would generate as much error as simply using the mean, implying that the above ratio would be close to 1. If we wanted the number to be large when the predictor is good and small when the predictor is bad, then we could subtract the number from 1:

$$1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

This is known as the *coefficient of determination*.

Coefficient of determination

The **coefficient of determination** is

$$1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

It measures one minus the ratio of the error produced by the regression line to the error produced by the naive guess using the sample mean. When it is defined, the number is between 0 and 1, with larger numbers implying that the regression model is a better fit.

19.3.4 Other forms of the coefficient of determination

We can rearrange the coefficient of determination in order to find other ways to talk about this number. One way to do that is to insert a copy of $\hat{y}_i - \bar{y}$ into the expression $y_i - \bar{y}$ to obtain:

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}).$$

Conveniently, the regression line has the useful property that:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n \left((y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \right)^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2.$$

Usually we would expect there to be cross-terms of the form $2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$, but they happen to vanish when summed over, as a consequence of the calculus used to construct the regression line.

With some basic algebra, we now see:

$$\begin{aligned} 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} &= \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \frac{\sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}. \end{aligned}$$

We can divide both the numerator and denominator by $n - 1$ to obtain:

$$1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\frac{1}{n-1} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}.$$

On the bottom of this expression is the variance for y , i.e. the spread of the paired data in the vertical direction. On the top of the fraction we have an expression which calculates a sort-of average squared distance from the predicted values to $\bar{y}$. This is again a type of variance, but this time we are calculating the spread of the predicted values $\hat{y}_i$ around $\bar{y}$. We may therefore think of the coefficient of determination in these terms.

Alternate Expression of the Coefficient of determination

The **coefficient of determination** can be equivalently expressed as

$$1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\frac{1}{n-1} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}.$$

Therefore, we may understand the coefficient of determination as measuring how much of the variance in y is explained by the regression line $\hat{y}$.

19.3.5 The coefficient of determination as r^2

Consider again the alternate form of the coefficient of determination. For convenience, we are going to insert copies of $\frac{1}{n-1}$ into the numerator and denominator:

$$\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\frac{1}{n-1} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}.$$

Therefore, the denominator becomes s_y^2 , the sample variance of y . Since the regression line is given by $\hat{y}_i = a + bx_i$ where $a = \bar{y} - b\bar{x}$, we see that:

$$\hat{y}_i - \bar{y} = a + bx_i - \bar{y} = (\bar{y} - b\bar{x}) + bx_i - \bar{y} = b(x_i - \bar{x}).$$

Therefore, the numerator of the coefficient of determination becomes:

$$\frac{1}{n-1} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \frac{1}{n-1} \sum_{i=1}^n (b(x_i - \bar{x}))^2 = b^2 \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = b^2 s_x^2.$$

Putting this together, we have:

$$\frac{\frac{1}{n-1} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2} = \frac{b^2 s_x^2}{s_y^2}.$$

Using the definition of b , we see that

$$b = \frac{\text{cov}(x, y)}{s_x^2} = \frac{\text{cov}(x, y)}{s_x^2} \cdot \frac{s_y^2}{s_y^2} = \frac{\text{cov}(x, y)}{s_x s_y} \cdot \frac{s_y^2}{s_x s_y} = r \cdot \frac{s_y}{s_x}.$$

Therefore, the coefficient of determination simplifies heavily to:

$$\frac{\frac{1}{n-1} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2} = b^2 s_x^2 \div s_y^2 = \frac{\left(r \frac{s_y}{s_x}\right)^2 s_x^2}{s_y^2} = r^2 \frac{s_y^2}{s_x^2} s_x^2 \div s_y^2 = r^2 s_y^2 \div s_y^2 = r^2.$$

Indeed, for simple linear regression with one explanatory variable and an intercept, the coefficient of determination is exactly the square of the sample correlation coefficient:

$$r^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\frac{1}{n-1} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}.$$

This has several immediate consequences:

- the sign of r disappears when it is squared;
- positive and negative correlations of equal strength produce the same r^2 ;
- a strong linear correlation produces a large coefficient of determination;
- a weak linear correlation produces a small coefficient of determination.

19.3.6 Example: coefficient of determination for the commute data

For our commuting data, we calculated previously that $r \approx 0.9454$. Therefore,

$$r^2 \approx (0.9454)^2 \approx 0.894.$$

Thus, approximately 89.4% of the observed variation in commute times is accounted for by the linear regression on distance from LUJ in this sample. This unexplained variation may reflect other variables influencing the commute time, such as method of transport, traffic, transfers, walking time, measurement noise, and so on.