MA22004 - Statistics II

Dr Eric Hall • ehall001@dundee.ac.uk

2024-09-02

University of Dundee

Table of contents

Welcome

Welcome to MA22004 at the University of Dundee.

This module covers the basics of statistical inference. This document contains the content. The appendix contains a list of curated content for your to investigate.

About your instructor

Hi, folks!

I’m Eric—your instructor for MA22004 this semester. I am a new Baxter Fellow in Applied Mathematics at Dundee, and my research focuses on uncertainty quantification and predictive modelling.

Originally from the US, I graduated from the University of Pennsylvania with a BA in Mathematics. I wrote my PhD in Probability and Stochastic Analysis at the University of Edinburgh. Math and stats have opened up some exciting doors for me, and I’ve had the opportunity to undertake postdoctoral work at KTH Stockholm, the University of Massachusetts Amherst, and RWTH Aachen University. I’m very excited to be at Dundee and back in Scotland. I’m even more excited to be teaching you statistics this semester!

These notes are available at dundeemath.github.io/MA22004/ and also as a PDF (visit the page and click on the PDF icon to download).

Licence

Part I

Module Introduction

Lab Guide {.unnumbered -#labs="" "}

You will learn about the statistical programming language `R` and the software RStudio by working through seven interactive lab tutorials and completing lab reports. The lab reports should answer the exercise questions at the end of each tutorial.

Tutorials and all associated materials (templates, data sets, further instructions, etc.) are available as an `R` package at the GitHub repository `dundeemath/MA22004labs` (i.e., https://github.com/dundeemath/MA22004labs).

Instructions on how to install and access the interactive lab tutorials can be found at:

- https://dundeemath.github.io/MA22004labs/.

The following section contains details about lab reports.

Writing Lab Reports {.unnumbered -#writing=“ “}

Assessment Criteria

There are seven interactive lab tutorials with accompanying exercises. Each lab tutorial specifies how marks are allocated across the exercises (a maximum of 20 marks available for each lab report).

> ❗ Marks are awarded for both **content** and **presentation**.

Content

Please work through the interactive tutorial for each lab. Your lab report should answer the exercises found at the end of each tutorial.

Presentation

Please use R Markdown to create your lab report. Further instructions on using R Markdown for creating *reproducible* lab reports that combine data analysis and text can be found in Lab 1.

Plots

Plots should be neat and legible, with appropriate aesthetic elements. Please use `ggplot` for creating plots and visualisations. Each plot should be annotated with titles, axis labels, and legends as appropriate. Plot aesthetics should be distinguished, e.g. using colours or line styles that are identified using a legend. Important data points and coordinates should be annotated using labels.

Mathematical formulas

Mathematical formulas should follow the same style rules as the lecture notes. Formulas can be included in R Markdown documents using LaTeX syntax. There should be appropriate spacing around operators and equals signs, e.g. $a + b = c$. For punctuation, formulas are treated as part of the text, so they often need to end with a full stop or comma. Important formulas can appear “displayed” on their own line (with line spacing above and below them), e.g. [A=πr^2,.]

Structure

Structure should be logical and clear. Organise your writing with suitable headings and sub-headings. For example, provide a solution to each exercise under its own heading.

Writing

Writing should follow the usual rules of good written English, including writing complete sentences and paragraphs that get to the point quickly. Your tone and language should be similar to lecture notes or scientific journal articles. Formal writing does not require unnecessary words, long words or monotonous use of passive voice. I will reward concise and clear communication, so please do not write, “Upon carefully analysing the aforementioned equations, the following mathematical solution was found,” when “The solution is” conveys the same thing.

Formatting

Formatting should rely on the *MA22004 Lab Report* template. This is available in the `MA22004labs` package, and further instructions can be found in Lab 1.

Part II

Lecture Notes

Preliminaries

This section contains a list of abbreviations, comment on notation, and a (very quick) review of probability.

Abbreviations

In Table 1 we list abbreviations used throughout these lecture notes. These abbreviations are pretty standdard and you might encounter them outside the module in other references.

Table 1: Commonly used abbreviations.

Abbreviation	Expanded
pdf	probability density function
cdf	cumulative distribution function
rv	random variable
iid	independent and identically distributed
obs	observations
CI	confidence interval
df	degrees of freedom

Notation

Uppercase roman letters, e.g., X, will typically denote random variables (rvs); lower case letters, e.g., x, will represent a particular value (observation) of a rv. Rvs have probability distributions. Distributions are typically characterised by *parameters* that describe population characteristics. In the present module, we will adopt the (frequentists) view that parameters are fixed real numbers that are often unknown and must be estimated from data. Statistical inference is a tool that will help us to do this.

> ⚠ Warning
>
> Statistical models comprise both rvs and parameters. Be careful not to confuse them!

For a random variable X that has a distribution F depending on a parameter θ, we will write $X \sim F(\theta)$.

> ⚠ Warning
>
> We write $X \sim F$ to indicate X has distribution F, not "X is approximately F"!

Sample space, events, probabilities

A *sample space* Ω is a set of possible outcomes of an experiment. Points $\omega \in \Omega$ are *sample outcomes* or realizations. Subsets $A \subset \Omega$ are called *events*.

Example 0.1 (Sample space)**.** Consider an experiment where we measure the petal widths from a randomly sampled cyclamen flowers. Before we observe the petal width, there is uncertainty that we can model using a sample space of events. The sample space is $\Omega = (0, \infty)$, since measurements of length should be positive (practically, the lengths will have a finite size, too). Each $\omega \in \Omega$ is a measurement of petal width for a cyclamen flower. Consider an event $A = (5, 12]$; this is the event that the petal width is larger than 5 but less than or equal to 12. Remember, we use probability to model uncertainty *before* we observe the petal width — after we take a measurement, the petal width is no longer uncertain (we have collected a statistic).

As sample spaces and events are described using sets, we recall the following notations, definitions, and laws about set theory. Let A, B, and $A_1, A_2, \ldots$ be events in a sample space Ω.

- complement: $A^c = \{\omega \in \Omega : \omega \notin A\}$.
- null event: $\varnothing = \Omega^c$.
- intersection: $A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}$. In particular, for $A_1, A_2, \ldots$, then

$$\bigcap_{i=1}^{\infty} A_i = \{\omega \in \Omega : \omega \in A_i \text{ for all } i\}\,.$$

- difference: $A \setminus B = \{\omega \in \Omega : \omega \in A, \omega \notin B\}$.
- size: $|A|$ denotes the number of elements in A.
- disjoint: $A_i \cap A_j = \varnothing$, for $i \neq j$.
- partition: disjoint $A_1, A_2, \ldots$ such that $\bigcup_{i=1}^{\infty} A_i = \Omega$.
- indicator: $I_A(\omega) = I(\omega \in A) = \{1 \text{ if } \omega \in A; 0 \text{ if } \omega \notin A\}$.
- monotone increasing: $A_1 \subset A_2 \subset \ldots$ and define limit

$$\lim_{n\to\infty} A_n = \bigcup_{i=1}^{\infty} A_i\,.$$

- monotone decreasing: $A_1 \supset A_2 \supset \ldots$ and define limit

$$\lim_{n\to\infty} A_n = \bigcap_{i=1}^{\infty} A_i\,.$$

- distributive laws:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)\,,$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)\,.$$

- De Morgan's laws:

$$(A \cap B)^c = A^c \cup B^c\,,$$
$$(A \cup B)^c = A^c \cap B^c\,.$$

We assign probabilities to events in our sample space.

Definition 0.1 (Probability distribution)**.** A probability distribution is a function $P : \Omega \to \mathbf{R}$ satisfying three axioms: 1. $P(A) \geq 0$ for every $A \subset \Omega$ (positivity), 2. $P(\Omega) = 1$ (totality), 3. if $A_1, A_2, \ldots$ are disjoint subsets of Ω, then

$$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)\,.$$

Tip

We can interpret $P(A)$ as representing:

- **frequency**, i.e., the long-run proportion of times A is true (the *frequentist perspective*),
- **degrees of belief**, i.e, as a measure of the observer's strength of belief that A is true (the *Bayesian perspective*).

Theorem 0.1 (PIE)**.** *The principal of inclusion-exclusion,*

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

Theorem 0.1 follows from the definition of a probability distributions and facts about set theory.

Definition 0.2 (Probability of an event)**.** For events A from finite sample spaces Ω, we assign probabilities according to:

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

For finite sample spaces, we assign probabilities according to their long-run frequency of occurring. For an event A, this is the ratio of the size of A (number of ways A can happen) to the size of Ω (number of total outcomes).

Definition 0.3 (Independent events)**.** Events A and B are independent, i.e., $A \perp\!\!\!\perp B$, iff $P(A \cap B) = P(A)P(B)$.

That is, events A and B are independent if and only if the probability of A and B occurring is equal to the the probability A occurring times the probability of B occurring.

Definition 0.4 (Conditional probability)**.** If $P(B) > 0$, then

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

Note that:

- $P(\cdot \mid B)$ satisfies the axioms of probability, for fixed B,
- in general, $P(A \mid \cdot)$ is not a probability for fixed A, and,
- in general, $P(A \mid B) \neq P(B \mid A)$.

Theorem 0.2 (Bayes Theorem)**.** *Let events $A_1, \dots, A_k$ partition Ω, with $P(A_i) > 0$.*

If $P(B) > 0$, then

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{\sum_j P(B \mid A_j)P(A_j)}\,.$$

Generally, it is not feasible to assign probabilities to *all* subsets of Ω (e.g., if □is infinite). In that case, we restrict to our attention to a σ-algebra $\mathscr{A}$ (also called, σ-field), which is a collection of sets satisfying: 1. $\mathbf{E}mptyset \in \mathscr{A}$, 2. if $A_1, A_2, \dots, \in \mathscr{A}$ then $\cup_{i=1}^{\infty} A_i \in \mathscr{A}$, 3. $A \in \mathscr{A} \implies A^c \in \mathscr{A}$.

Sets in $\mathscr{A}$ are said to be *measurable* and $(\Omega, \mathscr{A})$ is a measure space. If P is a probability defined on $\mathscr{A}$, then $(\Omega, \mathscr{A}, P)$ is called a *probability space*.

E.g., when $\Omega\mathbf{E}quiv\mathbf{R}$, we take $\mathscr{A}$ to be the smallest σ-field containing all open subsets of $\mathbf{R}$, which is called the Borel σ-field. If you find these details interesting, take: MA42008 Mathematical Statistics!

Random variables

Note

How do we link sample spaces and events to data?

We use random variables to link sample spaces and events to data.

Definition 0.5 (Random variables)**.** A random variable (rv) is a mapping $X : \Omega \to \mathbf{R}$ that maps $\omega \in \Omega \mapsto X(\omega)$.

Example 0.2. Consider a coin flipping experiment where you flip a fair coin eight times. Let X be the number of heads in the sequence. If three heads occur, e.g., $\omega = HTTTTTHH$, then $X(\omega) = 3$.

Example 0.3. Consider an experiment where you draw a point a random from the unit disk. Then $\Omega = \{(x, y) : x^2 + y^2 \leq 1\}$ and a typical outcome will be the pair $\omega = (x, y)$. Some random variables to consider are $X(\omega) = x$, $Y(\omega) = y$, $Z(\omega) = x + y$, and $W(\omega) = \sqrt{x^2 + y^2}$.

Definition 0.6 (Assigning probabilities to rvs)**.** Given X and $A \subset \mathbf{R}$, we define

$$X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\}$$

and let

$$P(X \in A) = P(X^{-1}(A)) = P(\{\omega \in \Omega : X(\omega) \in A\}),$$

e.g., $P(X = x) = P(X^{-1}(x)) = P(\{\omega \in \Omega : X(\omega) = x\})$.

> ⚠ Warning
>
> X denotes a rv and x denotes a particular value of X. You would never write $P(X)$, would you!?

Example 0.4. Consider a coin flipping experiment where you flip a fair coin twice. Let X be the number of heads. Then

$$P(X = 0) = P(\{TT\}) = \frac{1}{4},$$

$$P(X = 1) = P(\{HT\} \cup \{TH\}) = P(\{HT\}) + P(\{TH\}) = \frac{1}{2},$$

$$P(X = 2) = P(\{HH\}) = \frac{1}{4}.$$

Definition 0.7. ## Cdf

The cumulative distribution function (cdf), $F_X : \mathbf{R} \to [0, 1]$, is defined by $F_X(x) = P(X \leq x)$.

```r
x <- c(0, 0, 1, 1, 2, 2)
y <- c(0, 0.25, 0.25, 0.75, 0.75, 1)
b <- c("o", "c", "o", "c", "o", "c")
dat <- data.frame(x, y, b)
dat$b <- factor(dat$b)
dat |> ggplot(aes(x, y, shape = b)) +
  geom_point(size = lsz + lsz*2, color = "#E41A1C") +
  geom_segment(aes(x = 0, y = 0, xend = -1, yend = 0), arrow = arrow(length = unit(0.25*lsz, "cm")), li
  geom_segment(aes(x = 0, y = 0.25, xend = 1, yend = 0.25), linewidth = lsz, color = "#E41A1C") +
  geom_segment(aes(x = 1, y = 0.75, xend = 2, yend = 0.75), linewidth = lsz, color = "#E41A1C") +
  geom_segment(aes(x = 2, y = 1, xend = 3, yend = 1), arrow = arrow(length = unit(0.25*lsz, "cm")), lin
  geom_vline(aes(xintercept = 0)) +
  geom_hline(aes(yintercept = 0)) +
 scale_shape_manual(values = c(19, 1)) +
  theme(axis.line.x = element_blank(),
        axis.text.x = element_blank(),
        axis.ticks.x = element_blank()) +
 guides(shape = "none") +
 labs(x = TeX("$x$"), y = TeX("$F_X(x)$")) + theme_ur
```

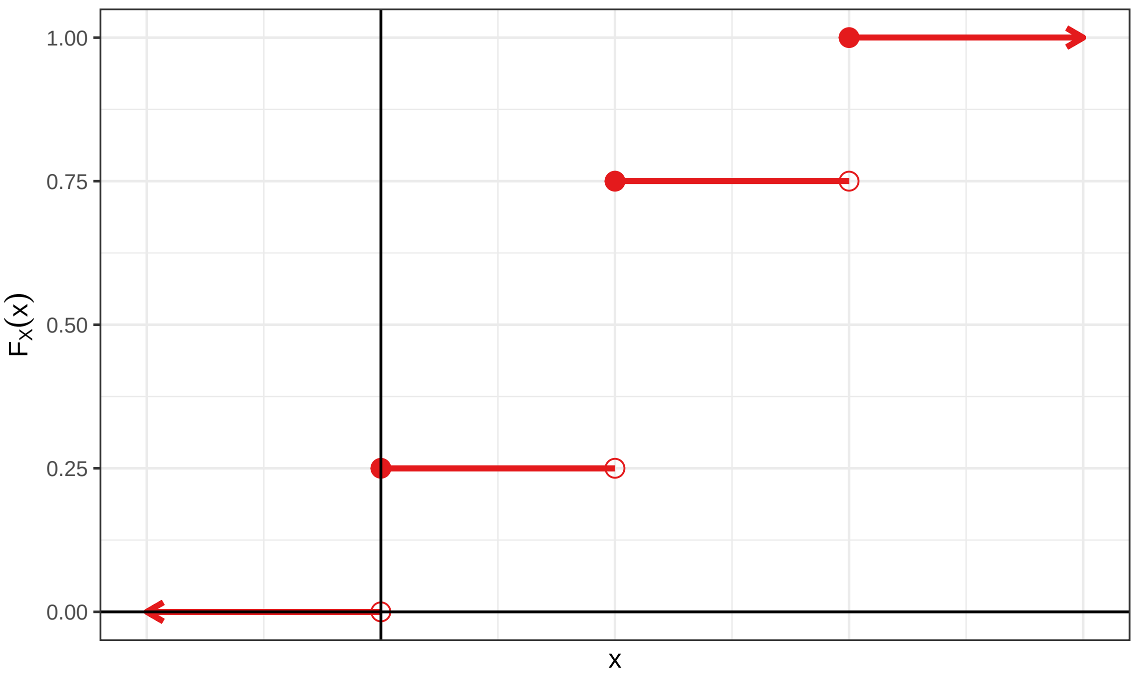

Figure 1: The cdf for the two coin flip example.

Note that a cdf completely determines the distribution of a random variable.

Theorem 0.3. *Let X have cdf F and Y have cdf G. If $F(x) = G(x)$ for all x, then $P(X \in A) = P(Y \in A) \forall A \in \mathbf{R}$.*

Theorem 0.4 (Properties of cdfs)**.** *$F : \mathbf{R} \to [0,1]$ is a cdf for some P iff,*

1. *F is nondecreasing (i.e., $x_1 < x_2 \implies F(x_1) \leq F(x_2)$),*
2. *F is normalized to $[0,1]$ (i.e., $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$),*
3. *F is right-continuous (i.e., $F(x) = F(x^*) \forall x$ where $F(x^*) = \lim_{y > x; y \to x} F(y)$).*

For a rv X we say X is *discrete* if it assumes at most a *countable* number of (discrete) values. For a discrete sample space, the collection of all probabilities of $X(\omega)$ gives us a probability distribution.

Definition 0.8 (Pmf)**.** A pdf for a discrete rv X is $f_X(x) = P(X = x)$. Since this density function places a "point mass" at each x, it is sometimes referred to as a probability mass function (pmf).

```
x <- c(0, 1, 2)
y <- c(0.25, 0.50, 0.25)
b <- c("a", "a", "a")
dat <- data.frame(x, y, b)
dat$b <- factor(dat$b)
dat |> ggplot(aes(x, y, color = b, fill = b)) +
  geom_bar(stat = "identity", linewidth =lsz) +
  geom_hline(yintercept = 0) + geom_vline(xintercept = -1) +
 guides(fill = "none", color = "none") +
 labs(x = TeX("$x_i$"), y = TeX("$f_X(x_i) = P(X(\\omega) = x_i)$")) + theme_ur
```

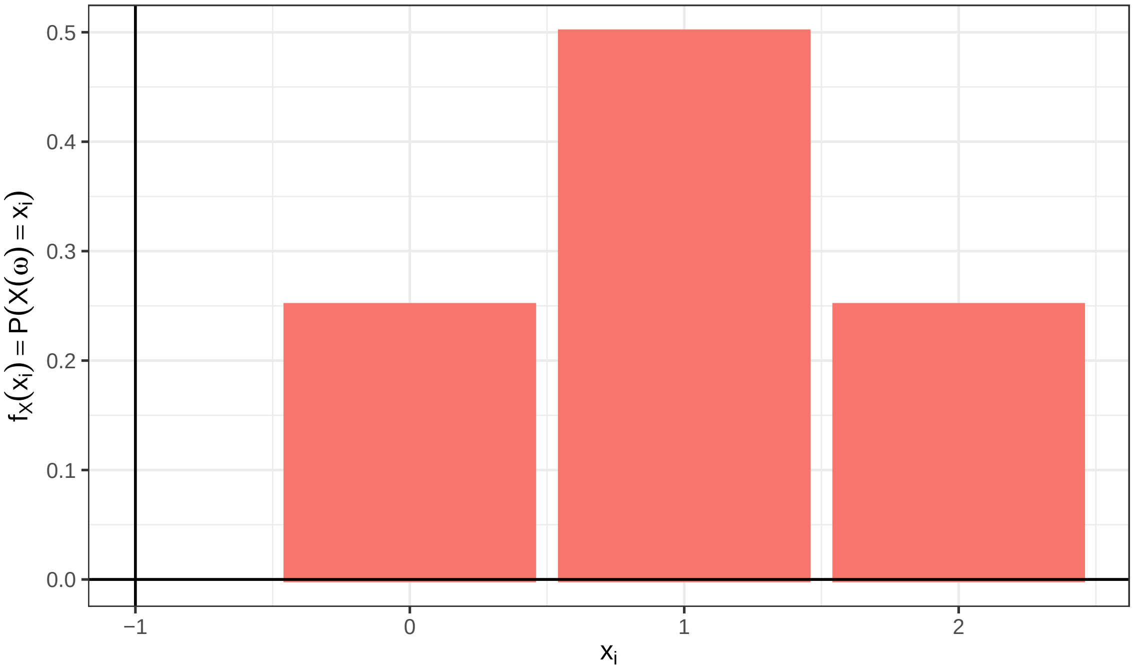

Figure 2: The histogram (pmf) for the two coin flip example.

Note, from the axioms of probability, that the pdf for a discrete random variable therefore satisfies $f(x) \geq 0, \forall x \in \mathbf{R}$ and $\sum_i f(x_i) = 1$.

A rv X is *continuous* if there exists a continuous function f_X such that, 1. $f_X(x) \geq 0 \forall x$, 2. $\int_{-\infty}^{\infty} f_X(x)dx = 1$ and 3. $P(a < X < b) = \int_a^b f_X(x)dx$, for $a \leq b$.

Definition 0.9 (Pdf)**.** A f_X satisfying the three properties above is a pdf for the continous rv X.

⚠ Warning

If X is continuous, then $P(X = x) = 0$ for every x. That is,

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

The cdf is related to the pdf by the derivative (difference). If X is continuous:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt$$

and $f_X(x) = F_X'(x)$ at all x at which F_X is differentiable. (Likewise, if X is discrete, then we replace the integral with a sum $F_X(x) = P(X \leq x) = \sum_{x_i \leq x} f_X(x_i)$.)

Definition 0.10 (Quantile function)**.** Let X be a rv with cdf F. The inverse cdf, or quantile function, is defined by

$$F^{-1}(q) = \inf\{x : F(x) > q\}$$

for $q \in [0, 1]$. If F is monotonic increasing and continuous then $F^{-1}(q)$ is the unique real number x such that $F(x) = q$.

Some quantiles get used more than others (and therefore get names). Important quantiles include, $F^{-1}(\frac{1}{4})$ is the first quantile, $F^{-1}(\frac{1}{2})$ is the median, and $F^{-1}(\frac{3}{4})$ is the third quantile.

Definition 0.11 (Equality in distribution)**.** We say X and Y are equal in distribution, $X \equiv Y$, if $F_X(x) = F_Y(x)$ for all x.

! Important

Note that equality in distribution does not mean that the random variables are the same. Rather, probability statements are the same.
Suppose $P(X = 1) = P(X = -1) = \frac{1}{2}$. Let $Y = -X$. Then $P(Y = 1) = P(Y = -1) = \frac{1}{2}$. Thus, $X\mathbf{E}quivY$, but X and Y are not equal; in fact, $P(X = Y) = 0$.

We sometimes consider more than one random variable, taken to together. This leads to the concept of a joint and marginal densities.

Definition 0.12 (Joint pdf)**.** A joint pdf for (X, Y) satisfies

1. $f(x, y) \geq 0 \; \forall x, y$,
2. $\iint_{-\infty}^{\infty} f(x, y)dxdy = 1$,
3. for $A \in \mathbf{R} \times \mathbf{R}$, $P((X, Y) \in A) = \iint_A f(x, y)dxdy$.

Definition 0.13 (Joint cdf)**.** A joint cdf is given by $F(x, y) = P(X \leq x, Y \leq y)$.

Definition 0.14 (marginal pdf)**.** For X, Y with joint pdf $f(x, y)$, we define the marginals for X and Y as $f_X(x) \int f(x, y)dy$ and $f_Y(y) = \int f(x, y)dx$, respectively.

We also have a notion of independence for rvs.

Definition 0.15 (Independence of rvs)**.** Rvs X and Y are independent if $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$.

Theorem 0.5. *Let X, Y have joint f_{XY}. Then X and Y are independent iff $f_{XY} = f_X \cdot f_Y$ for all x, y.*

If $X_1, \dots X_n$ are independent and each as the same marginal distribution with cdf F, we say $X_1, \dots, X_n$ are iid and write $X_1, \dots, X_n \sim F$ iid. We also write $X_1, \dots, X_n \sim f$ if F has corresponding density f, when no confusion arises.

Definition 0.16 (Random sample)**.** $X_1, \dots, X_n \sim F$ iid is a random sample of size n from a distribution F.

We also consider the expected value of a rv.

Definition 0.17 (Expectation)**.** For a discrete rv X with possible outcomes $x_1, x_2, \dots$ and corresponding probabilities $p_1, p_2, \dots$, the expectation is defined by

$$\mathbf{E}[X] = \sum_{i=1}^{\infty} x_i p_i \,.$$

For a continuous rv X with pdf f, the expectation is defined by

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

For both discrete and continuous rvs, we refer to various statistics relating to expected values as moments of the distribution.

Definition 0.18 (n-th raw moment)**.** For a rv X, the n-th raw moment is given by $\mathbf{E}[X^n]$.

Definition 0.19 (n-th central moment)**.** For a rv X with $\mu = \mathbf{E}[X]$, the n-th central moment is defined as $\mathbf{E}[(X - \mu)^n]$.

The *mean* of a distribution is the first raw moment. The *variance* of a distribution is the second central moment.

Table 2: First three moments for a rv X with mean $\mu = \mathbf{E}[X]$.

(a) Raw moments

Expression	Name
$\mathbf{E}[X]$	first (raw) moment
$\mathbf{E}[X^2]$	second (raw) moment
$\mathbf{E}[X^3]$	third (raw) moment

(b) Central moments

Expression	Name
$\mathbf{E}[(X - \mu)]$	first central moment
$\mathbf{E}[(X - \mu)^2]$	second central moment
$\mathbf{E}[(X - \mu)^3]$	third central moment

Topic 1

Sampling distributions

A **statistic** is a quantity that can be calculated from sample data. Before observing data, a statistic is an unknown quantity and is, therefore, a rv.

Definition 1.1 (statistic)**.** Let $X_1, \ldots, X_n$ be observable rvs and let g be an arbitrary real-valued function of n random variables. The rv $T = g(X_1, \ldots, X_n)$ is a **statistic**.

We refer to the probability distribution for a statistic as a **sampling distribution**. The sampling distribution illustrates how the statistic will vary across *possible* sample data. The sampling distribution contains information about the values a statistic is likely to assume and how likely it is to assume those values *prior to observing data*.

Definition 1.2 (sampling distribution)**.** Suppose rvs $X_1, \ldots, X_n$ are a random sample from $F(\theta)$, a distribution depending a parameter θ whose value is uknown. Let the rv $T = g(X_1, \ldots, X_n, \theta)$ be a function of $X_1, \ldots, X_n$ and (possibly) θ. The distribution of T (given θ) is the **sampling distribution** of T.

The sampling distribution of T is derived from the distribution of the random sample. Often we will be interested in a statistic T that is an estimator for a parameter θ (that is, T will not depend on θ).

In what follows, we review several special families of distributions that are widely used in probability and statistics. These special families of distributions will be indexed by one or parameters.

1.1 Uniform distribution

The uniform distribution places equal on uniform weight on the items being sampled.

Definition 1.3 (uniform distribution)**.** A continuous rv X has a **uniform distribution** on $[a, b]$ with $a < b$, if X has pdf

$$f(x; a, b) = \frac{1}{b-a}, \quad a < x < b,$$

or zero otherwise. We write $X \sim \mathsf{Unif}(a, b)$.

Note a and b are parameters.

Exercise 1.1. As an exercise, derive the cdf using the definition. Derive a formula for the mean and variance in terms of the parameters a and b.

1.2 Normal distribution

Normal distributions play an important role in probability and statistics as they describe many natural phenomena. For instance, the Central Limit Theorem tells us that the sample mean of a large random sample (size m) of rvs with mean μ and variance σ^2 is approximately normal in distribution with mean μ and variance σ^2/m.

Definition 1.4 (Normal or Gaussian distribution)**.** A continuous rv X has a **normal distribution** with parameters μ and σ^2, where $-\infty < \mu < \infty$ and $\sigma > 0$, if X has pdf

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}, \quad -\infty < x < \infty .$$

We write $X \sim \mathsf{N}(\mu, \sigma^2)$.

For $X \sim \mathsf{N}(\mu, \sigma^2)$, it can be shown that $\mathbf{E}(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$, that is, μ is the *mean* and σ^2 is the *variance* of X. The pdf forms a bell-shaped curve that is symmetric about μ. The value σ (*standard deviation*) is the distance from μ to the inflection points of the curve. Thus, the distribution's position (location) and spread depend on μ and σ.

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```

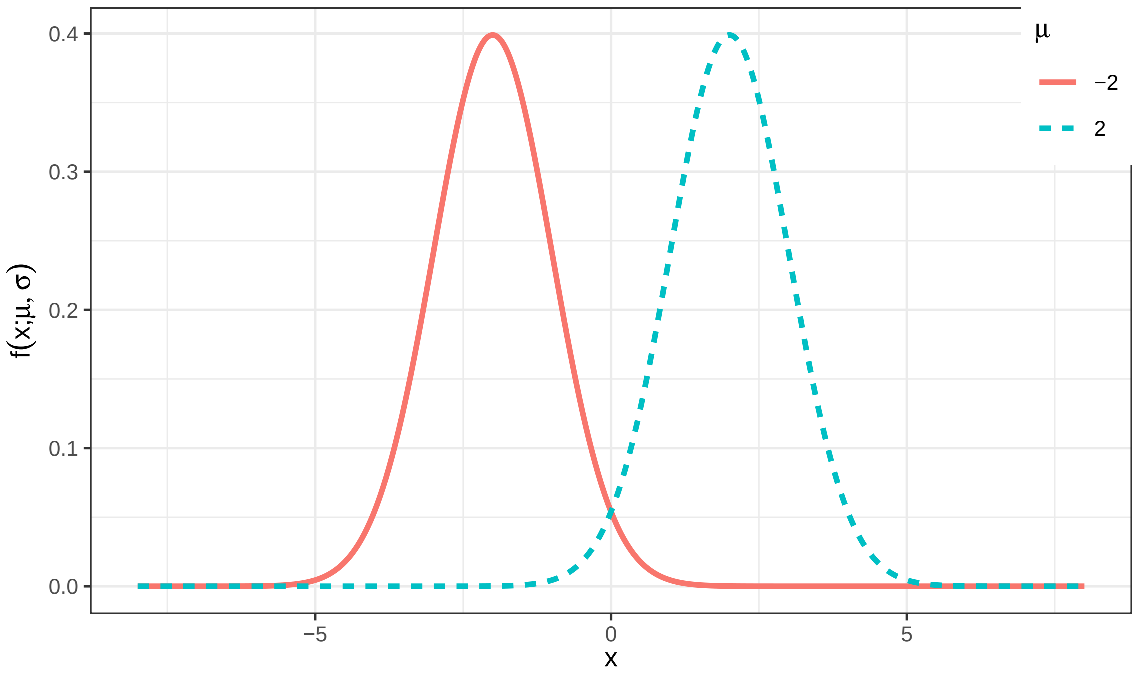

Figure 1.1: The pdfs of two normal rvs, $X_1 \sim \mathsf{N}(-2, 1)$ and $X_2 \sim \mathsf{N}(2, 1)$, with *different means* and the same standard deviations.

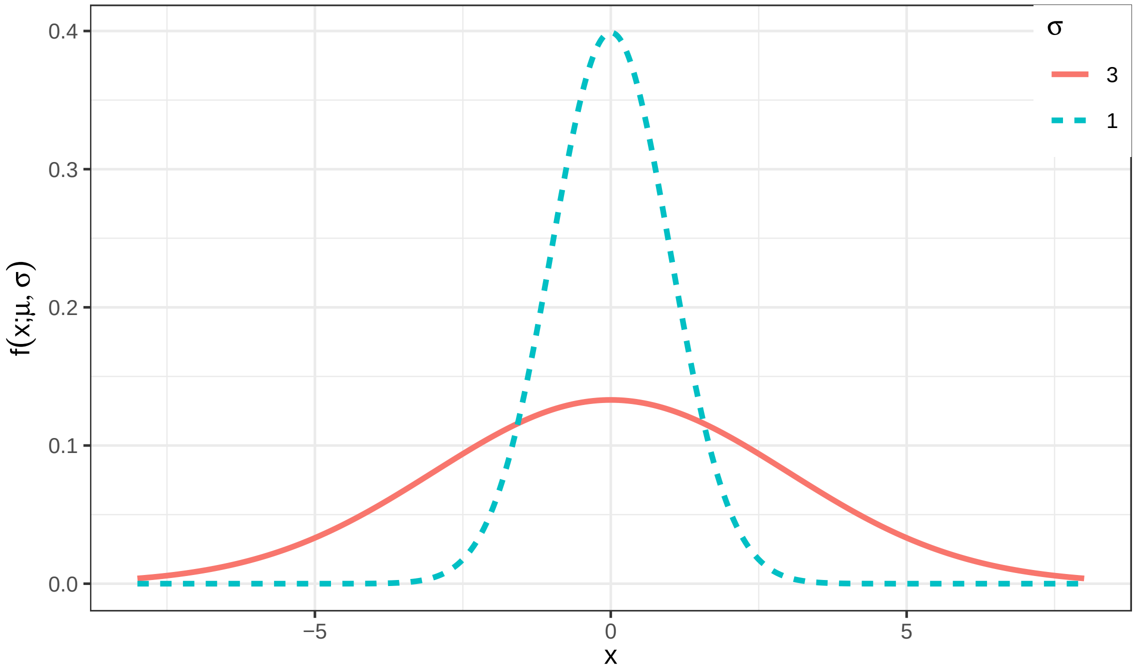

Figure 1.2: The pdfs of two normal rvs, $X_1 \sim \mathsf{N}(0,9)$ and $X_2 \sim \mathsf{N}(0,1)$, with the same means and *different standard deviations*.

Definition 1.5 (Standard Normal distribution)**.** We say that X has a **standard normal distribution** if $\mu = 0$ and $\sigma = 1$ and we will usually denote standard Normal rvs by $Z \sim \mathsf{N}(0,1)$ (why Z? tradition![1]). We denote the cumulative distribution function of the standard normal by $\Phi(z) = P(Z \leq z)$ and write $\varphi = \Phi'$ for its density function.

1.2.1 Some useful facts about normal variates

Here are some useful facts about how to manipulate Normal rvs.

1. If $X \sim \mathsf{N}(\mu, \sigma^2)$, then $Z = (X - \mu)/\sigma \sim \mathsf{N}(0,1)$.
2. If $Z \sim \mathsf{N}(0,1)$, then $X = \mu + \sigma Z \sim \mathsf{N}(\mu, \sigma^2)$.
3. If $X_i \sim \mathsf{N}(\mu_i, \sigma_i^2)$ for $i = 1, \ldots, n$ are independent rvs, then

$$\sum_{i=1}^{n} X_i \sim \mathsf{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right).$$

In particular, we note that for differences of independent rvs $X_1 \sim \mathsf{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathsf{N}(\mu_2, \sigma_2^2)$ then the variances also add:

$$X_1 - X_2 \sim \mathsf{N}(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2).$$

Probabilities $P(a \leq X \leq b)$ are found by converting the problem in $X \sim \mathsf{N}(\mu, \sigma^2)$ to the *standard normal* distribution $Z \sim \mathsf{N}(0,1)$ whose probability values $\Phi(z) = P(Z \leq z)$ can then be looked up in a table. From (1.) above,

$$\begin{aligned} P(a < X < b) &= P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right). \end{aligned}$$

This process is often referred to as *standardising* (the normal rv).

[1] "Traditions, traditions… Without our traditions, our lives would be as shaky as a fiddler on the roof!" [https://www.youtube.com/watch?v=gRdfX7ut8gw].

Example 1.1. Let $X \sim \mathsf{N}(5, 9)$ and find $P(X \geq 5.5)$.

$$\begin{aligned} P(X \geq 5.5) &= P\left(Z \geq \frac{5.5-5}{3}\right) \\ &= P(Z \geq 0.1667) \\ &= 1 - P(Z \leq 0.1667) \\ &= 1 - \Phi(0.1667) \\ &= 1 - 0.5662 \\ &= 0.4338\,, \end{aligned}$$

where we look up the value of $\Phi(z) = P(Z \leq z)$ in a table of standard normal curve areas.

The probability corresponds to the shaded area under the normal density $\varphi(x) = \Phi'(x)$ corresponding to $x \geq 5.5$.

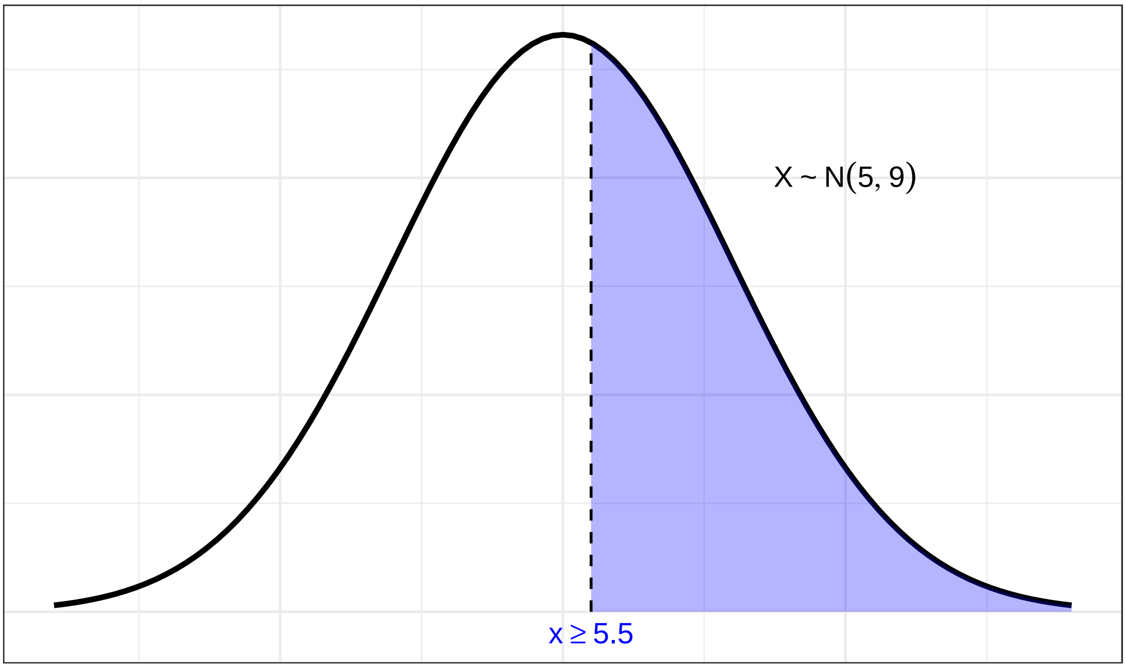

Figure 1.3: The normal density from the exercise with the (one-sided) interval shaded in blue.

Alternatively, we can use the `r` code: `pnorm(5.5, mean = 5, sd = 3, lower.tail = FALSE)`.

Example 1.2. Let $X \sim \mathsf{N}(5, 9)$ and find $P(4 \leq X \leq 5.25)$.

$$\begin{aligned} P(4 \leq X \leq 5.25) &= P\left(\frac{4-5}{3} \leq Z \leq \frac{5.25-5}{3}\right) \\ &= P(-0.3333 \leq Z \leq 0.0833) \\ &= \Phi(0.0833) - \Phi(-0.3333) \\ &= 0.5332 - 0.3694 \\ &= 0.1638\,. \end{aligned}$$

where we look up the value of $\Phi(z) = P(Z \leq z)$ in a table of standard normal curve areas.

The probability corresponds to the shaded area under the normal density $\varphi(x) = \Phi'(x)$ corresponding to $4 \leq x \leq 5.25$.

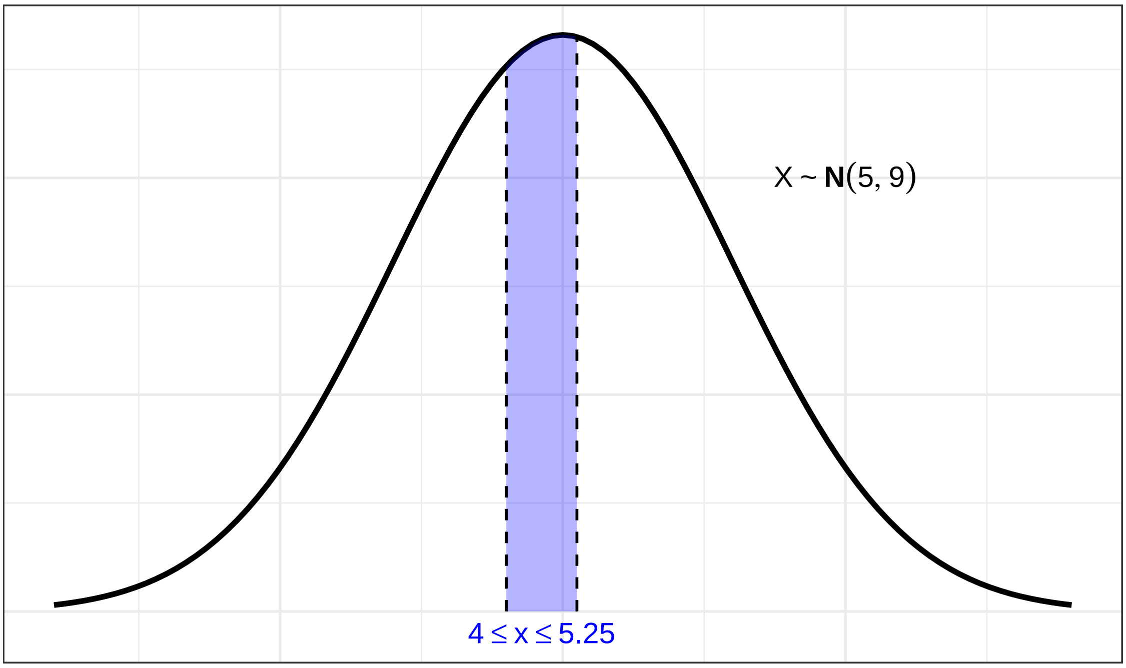

Figure 1.4: The normal density from the exercise with the interval shaded in blue.

Alternatively, we can use the `r` code: `pnorm(5.25, mean = 5, sd = 3) - pnorm(4, mean = 5, sd = 3)`. ◊

1.2.2 Empirical rule (68 – 95 – 99.7 rule)

For samples from a normal distribution, the percentage of values that lie within one, two, and three standard deviations of the mean are 68.27%, 95.45%, and 99.73%, respectively. That is, for $X \sim \mathsf{N}(\mu, \sigma^2)$,

$$P(\mu - 1\sigma \le X \le \mu + 1\sigma) \approx 0.6827\,,$$

$$P(\mu - 2\sigma \le X \le \mu + 2\sigma) \approx 0.9545\,,$$

$$P(\mu - 3\sigma \le X \le \mu + 3\sigma) \approx 0.9973\,.$$

For a normal population, nearly all the values lie within "three sigmas" of the mean.

1.3 t and Cauchy distribution

Student's t distribution gets its peculiar name as it was first published under the pseudonym "Student".[2] This bit of obfuscation was to protect the identity of his employer,[3] and thereby vital trade secrets, in a highly competitive and lucrative industry.

Definition 1.6 (Student's t-distribution)**.** A continuous rv X has a t **distribution** with parameter $\nu > 0$, if X has pdf

$$f(x;\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}\,, \quad -\infty < x < \infty\,.$$

We write $X \sim \mathsf{t}(\nu)$. Note Γ is the standard gamma function.[4]

[2] William Sealy Gosset (1876–1937) wrote under the pseudonym "Student" [https://mathshistory.st-andrews.ac.uk/Biographies/Gosset/].

[3] Gosset invented the t-test to handle small samples for quality control in brewing, specifically for the Guinness brewery in Dublin [https://www.wikiwand.com/en/Guinness_Brewery].

[4] The gamma function is defined by $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx$ when the real part of z is positive. For any positive integer n, $\Gamma(n) = (n-1)!$ and for half-integers $\Gamma(\frac{1}{2}+n) = \frac{(2n)!}{4^n n!}\sqrt{\pi}$.

The density for $\mathrm{t}(\nu)$ for several values of ν are plotted below.

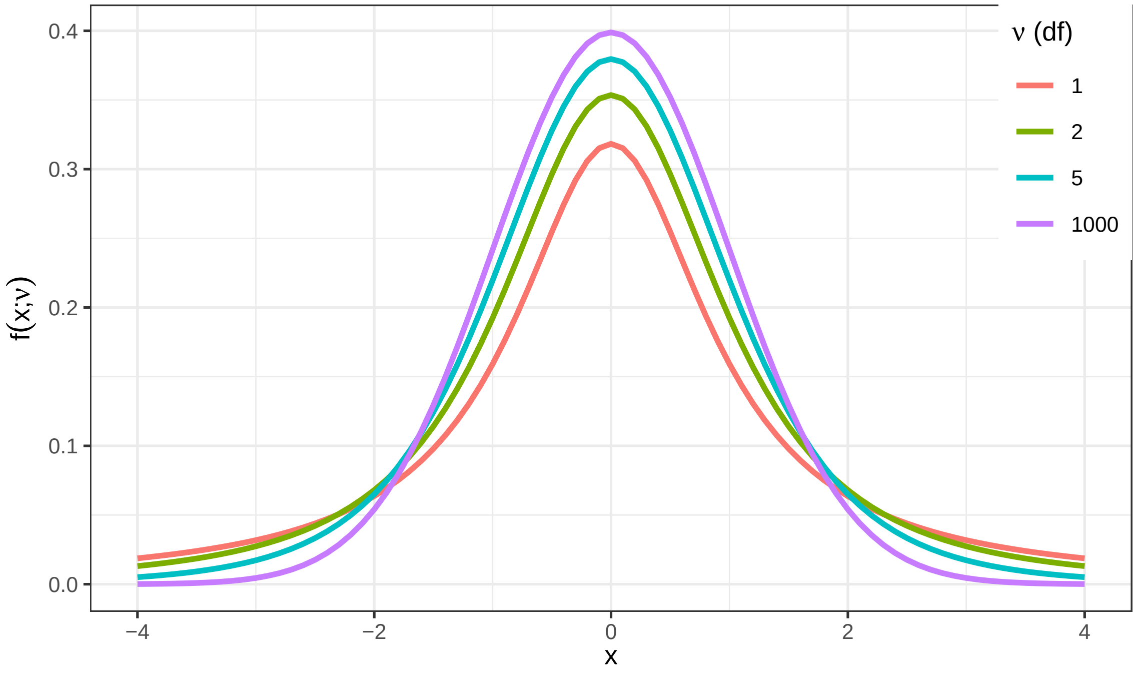

Figure 1.5: The density for $\mathrm{t}(\nu)$ for several values of ν (df).

> A t distributions with $\nu = 1$ has pdf
>
> $$f(x) = \frac{1}{\pi(1 + x^2)},$$
>
> and we call this the Cauchy distribution.

1.3.1 Properties of t distributions

1. The density for $\mathrm{t}(\nu)$ is a bell-shaped curve centred at 0.
2. The density for $\mathrm{t}(\nu)$ is more spread out than the standard normal density (i.e., it has "fatter tails" than the normal).
3. As $\nu \to \infty$, the spread of the corresponding $\mathrm{t}(\nu)$ density converges to the standard normal density (i.e., the spread of the $\mathrm{t}(\nu)$ density decreases relative to the standard normal).

If $X \sim \mathrm{t}(\nu)$, then $\mathbf{E}[X] = 0$ for $\nu > 1$ (otherwise the mean is undefined).

1.4 χ^2 distribution

The χ^2 distribution arises as the distribution of a sum of the squares of ν independent standard normal rvs.

Definition 1.7 (χ^2 distribution)**.** A continuous rv X has a χ^2 **distribution** with parameter $\nu \in \mathbf{N}_>$, if X has pdf

$$f(x;\nu) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{(\nu/2)-1} e^{-x/2},$$

with support $x \in (0, \infty)$ if $\nu = 1$, otherwise $x \in [0, \infty)$. We write $X \sim \chi^2(\nu)$.

The pdf $f(x;\nu)$ of the $\chi^2(\nu)$ distribution depends on a positive integer ν referred to as the df. The densities for several values of ν are plotted below.

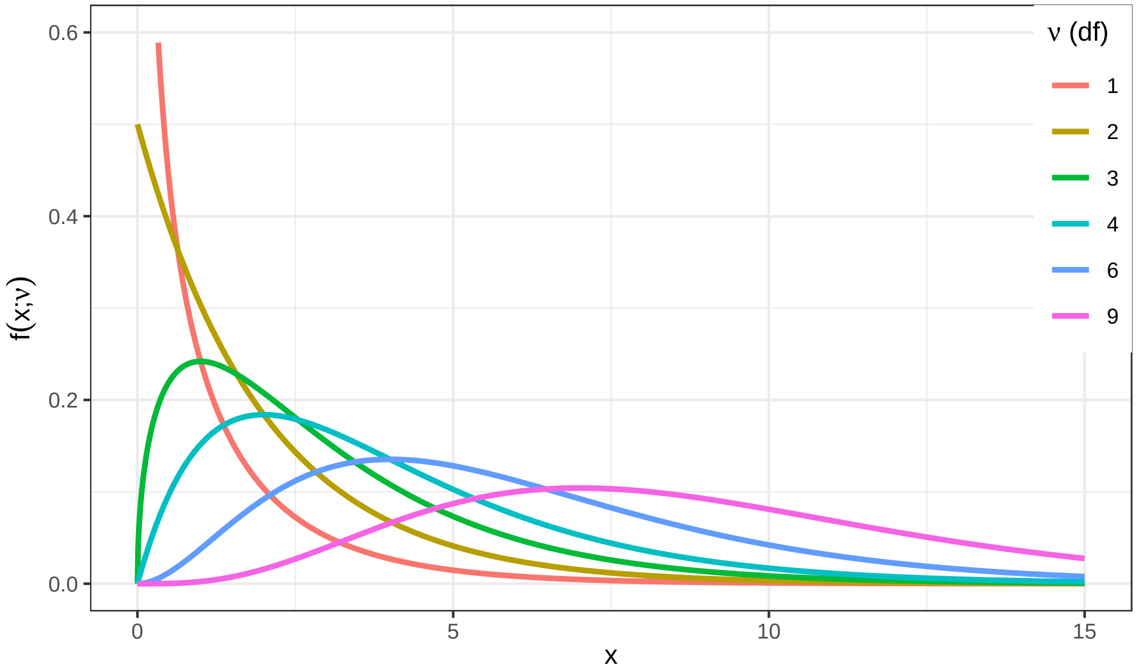

Figure 1.6: The density for $\chi^2(\nu)$ for several values of ν (df).

The density $f(x;\nu)$ is positively skewed, i.e., the right tail is longer, so the mass is concentrated to the figure's left. The distribution becomes more symmetric as ν increases. We denote critical values of the $\chi^2(\nu)$ distribution by $\chi^2_{\alpha,\nu}$.

> ⚠ Unlike the normal and t distributions, the χ^2 distribution is not symmetric. This means that the critical values e.g. $\chi^2_{.99,\nu}$ and $\chi^2_{0.01,\nu}$ are **not** equal. Hence, it will be necessary to look up both values for CIs based on χ^2 critical values.

If $X \sim \chi^2(\nu)$, then $\mathbf{E}[X] = \nu$ and $\mathrm{Var}[X] = 2\nu$.

1.5 F distribution

The F distribution ("F" for Fisher) arises as a test statistic when comparing population variances and in the analysis of variance (see @ref(anova)).

Definition 1.8 (F distribution)**.** A continuous rv X has an F **distribution** with df parameters ν_1 and ν_2, if X has pdf

$$f(x;\nu_1,\nu_2) = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)\nu_1^{\nu_1/2}\nu_2^{\nu_2/2}}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \frac{x^{\nu_1/2-1}}{(\nu_2+\nu_1 x)^{(\nu_1+\nu_2)/2}} .$$

The pdf $f(x;\nu_1,\nu_2)$ of the $\mathsf{F}(\nu_1,\nu_2)$ distribution depends on two positive integers ν_1 and ν_2 referred to, respectively, as the numerator and denominator df. The density is plotted below for several combinations of (ν_1,ν_2).

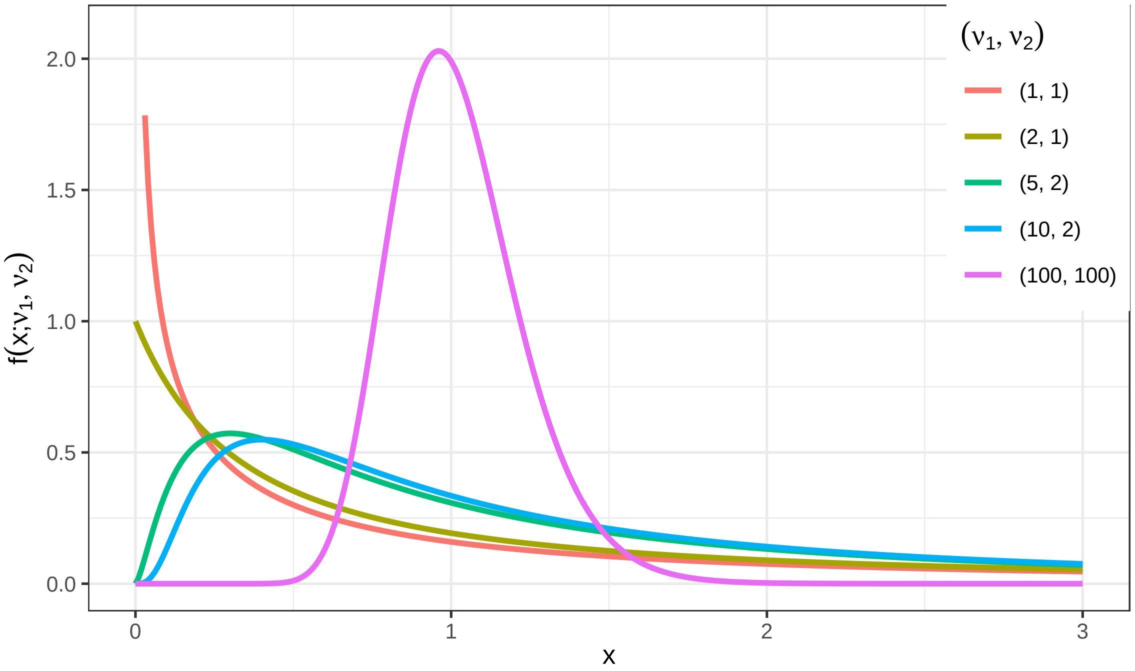

Figure 1.7: The density for $\mathsf{F}(\nu_1, \nu_2)$ for several combinations of (ν_1, ν_2).

Where do the terms numerator and denominator df come from? The F distribution is related to ratios of χ^2 rvs.

Theorem 1.1 (Ratio of χ^2 rvs)**.** *If $X_1 \sim \chi^2(\nu_1)$ and $X_2 \sim \chi^2(\nu_2)$ are independent rvs, then the rv*

$$F = \frac{X_1/\nu_1}{X_2/\nu_2} \quad \sim \mathsf{F}(\nu_1, \nu_2)\,,$$

that comprises the ratio of two χ^2 rvs divided by their respective df has an $\mathsf{F}(\nu_1, \nu_2)$ distribution.

Curated Content

Below we provide links to supplementary online material. Hopefully, some of the items will inspire you to view the module material in a broader context and lead to further investigations.

Investigation 0

What is Statistics?

- **Cambridge Ideas - Professor Risk**
  https://www.youtube.com/watch?v=a1PtQ67urG4
  Prof David Spiegelhalter (Cambridge University) discusses public understanding of risk. You may also be interested in reading (Spiegelhalter 2020).
- **The Joy of Statistics**
  https://www.youtube.com/watch?v=jbkSRLYSojo
  Prof Hans Rosling (Karolinska Institute and Gapminder Foundation) analyses data from 200 Countries over 200 Years in 4 Minutes - The Joy of Stats - BBC Four.
- **Teach statistics before calculus!**
  https://www.ted.com/talks/arthur_benjamin_teach_statistics_before_calculus
  Prof Arthur Benjamin (Harvey Mudd College) argues that the pinnacle of math education is probability and statistics — not calculus.
- **Kaggle**
  https://www.kaggle.com/
  Towards data science.
  https://www.youtube.com/watch?v=TNzDMOg_zsw
  What's Kaggle?

Investigation 1

Defence against the dark arts.

- **Three ways to spot bad statistics**
  https://www.ted.com/talks/mona_chalabi_3_ways_to_spot_a_bad_statistic
  Mona Chalabi (Data Journalist) discusses three ways to spot bad statistics.
- **Statistics Done Wrong**
  https://www.statisticsdonewrong.com/
  A book by Dr Alex Reinhart (Carnegie Mellon University).
- **How to defend yourself against misleading statistics in the news**
  https://www.youtube.com/watch?v=mJ63-bQc9Xg
  Sanne Blauw (Journalist) discusses how the presentation of statistics can mislead.

Investigation 2

Data analysis and visualisation.

- **The Grammar of Graphics**
  https://www.youtube.com/watch?v=h-62NwWUI5c
  What Makes A Good Visualisation? Rhys Jackson from RocketMill, a UK Digital Marketing Agency, gives a perspective on visualising data from a marketing perspective.
  https://www.youtube.com/watch?v=kepKM7Z2O54
  David Keyes (RStudio) discusses how the grammar of graphics underpins the `ggplot2` data visualization package in `R`.

- **Same Stats, Different Graphs**
  https://www.autodeskresearch.com/publications/samestats
  Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing (ACM SIGCHI Conference on Human Factors in Computing Systems) by Justin Matejka, George Fitzmaurice.

- **Why do we so often use 0.05 for hypothesis testing?**
  https://www.openintro.org/book/stat/why05/
  In this online exercise, you will gain an improved understanding of what a significance level is, and why a value in the neighbourhood of 0.05 is reasonable as a default.

- **Data visualisations**
  https://flowingdata.com/
  FlowingData blog by Nathan Yau.
  https://fivethirtyeight.com/
  FiveThirtyEight blog by Nate Silver.

- **Storytelling with data**
  http://www.storytellingwithdata.com/blog
  Blog with nice hints and tips for how to present data in tables, graphics, and visualisations.
  https://community.storytellingwithdata.com/challenges
  Monthly challenge.

Investigation 3

Statistical paradoxes.

- **How statistics can be misleading (TED-Ed)**
  https://www.ted.com/talks/mark_liddell_how_statistics_can_be_misleading
  Mark Liddell (Educator) discusses Simpson's Paradox in this TED-Ed animation.

- **Low birth-weight paradox**
  https://www.wikiwand.com/en/Low_birth-weight_paradox

- **Gambler's Fallacy**
  https://www.youtube.com/watch?v=4eVluL-idkM
  Prof Kelly Shue (Chicago Booth) discusses the gambler's fallacy.

Investigation 4

The law and interpreting statistics.

- **How stats fool juries.**
  https://youtu.be/kLmzxmRcUTo
  Prof Peter Donnelly (Oxford University) discusses common mistakes in interpreting statistics.

- **Better Data in Forensic Science**
  https://www.dundee.ac.uk/leverhulme/projects/details/better-data-in-forensic-science.php

Dr Christian Cole (Dundee) is leading a data-focused project as part of the Leverhulme Research Centre for Forensic Science right here at Dundee.

- **Prosecutor's fallacy**
  https://www.wikiwand.com/en/Prosecutor%27s_fallacy
  A fallacy of statistical reasoning, typically used by a prosecutor to exaggerate the likelihood of guilt: because $P(\text{hypothesis} \mid \text{evidence}) \neq P(\text{evidence} \mid \text{hypothesis})$!

Investigation 5

Data-driven decision making in epidemiology.

- **Project Tycho**
  https://www.tycho.pitt.edu/
  Digitized archival epidemiological data for the United States and the world.
  https://www.youtube.com/watch?v=Kn9OJy1BPDo
  An overview of the origins of project Tycho.
- **Public Health Scotland COVID-19 Dashboard**
  https://public.tableau.com/app/profile/phs.covid.19/viz/COVID-19DailyDashboard_15960160643010/Dailyupdate
  The official COVID-19 dashboard of Public Health Scotland.
- **Our World in Data**
  https://ourworldindata.org/
  A project of the Oxford Martin School to make public health data, including progress in UN Sustainable Development Goals, available and accessible.
- **Demographic Party Trick**
  https://www.youtube.com/watch?v=2nDh8MQuS-Y
  Prof Hans Rosling (Karolinska Institute and Gapminder Foundation) and Bill Gates seek to shed light on the true statistics of childhood vaccinations.

Investigation 6

Spurious correlations!

- **The danger of mixing up causality and correlation**
  https://www.youtube.com/watch?v=8B271L3NtAw
  Prov Ionica Smeets (University of Leiden) discusses causality and correlation.
- **Spurious correlations**
  https://tylervigen.com/spurious-correlations
  Tyler Vigen's site dedicated to spurious correlations.
- **Cause & Effect**
  https://www.youtube.com/watch?v=lbODqslc4Tg
  Correlation vs. causality from the Clip from the 2010 documentary "Freakonomics: The Movie".

Investigation 7

Data and Society: can data-driven and predictive modelling lead to a better world? What are the ethics of mass data collection?

- **Science behind the news: Predictive Policing**
  https://www.youtube.com/watch?v=74_jreara3w
  The Los Angeles Police Department is using a new tactic in their fight against crime called "predictive

policing.” It’s a computer program originally developed by a team at UCLA, including mathematician Andrea Bertozzi and anthropologist Jeff Brantingham. “Science Behind the News” is produced in partnership with NBC Learn. (Provided by the National Science Foundation & NBC Learn)

- **You should get paid for your data**
  https://www.nytimes.com/video/opinion/100000006678020/data-privacy-jaron-lanier-2.html
  Jaron Lanier (Computer Scientist and Author) discusses a compensation plan and data dignity.
  https://www.ted.com/talks/jennifer_zhu_scott_why_you_should_get_paid_for_your_data
  Jennifer Zhu Scott (Computer Scientist) also thinks you should get paid for your data.
- **How tech companies deceive you into giving up your data and privacy**
  https://www.ted.com/talks/finn_lutzow_holm_myrstad_how_tech_companies_deceive_you_into_giving_up_your_data_and_privacy
  Finn Lützow-Holm Myrstad (Norwegian Consumer Council) discusses consumer protections and data collection.
- **Your company’s data could help end world hunger**
  https://www.ted.com/talks/mallory_freeman_your_company_s_data_could_help_end_world_hunger
  Mallory Freeman (Data Scientist) discusses how to do the most good with data.

Investigation 8

Machine learning / big data.

- **What is Machine Learning?**
  https://www.youtube.com/watch?v=f_uwKZIAeM0
  OxfordSparks discusses the topic of supervised learning algorithms and how machine learning is used all around us.
- **Big Data (TED-Ed)**
  https://www.youtube.com/watch?v=j-0cUmUyb-Y
  Tim Smith (educator) discusses the historical arc of big data in this TED-Ed animation.
- **The human insights missing from big data**
  https://www.ted.com/talks/tricia_wang_the_human_insights_missing_from_big_data
  Tricia Wang (Ethnographer) discusses the human insights missing from big data.
- **How we can find ourselves in data**
  https://www.ted.com/talks/giorgia_lupi_how_we_can_find_ourselves_in_data
  Giorgia Lupi (Designer) discusses a humanistic approach to data and data visualization.

Spiegelhalter, David J. 2020. *The Art of Statistics: Learning from Data*. London: Pelican Books.