Introduction and Chi-square Distribution

Haziq Jamil SM-4331 Students :)

Table of contents

Preface

These are the notes for SM-4331 Advanced Statistics.

What is statistics?

Statistics is a scientific subject on collecting and analysing data.

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

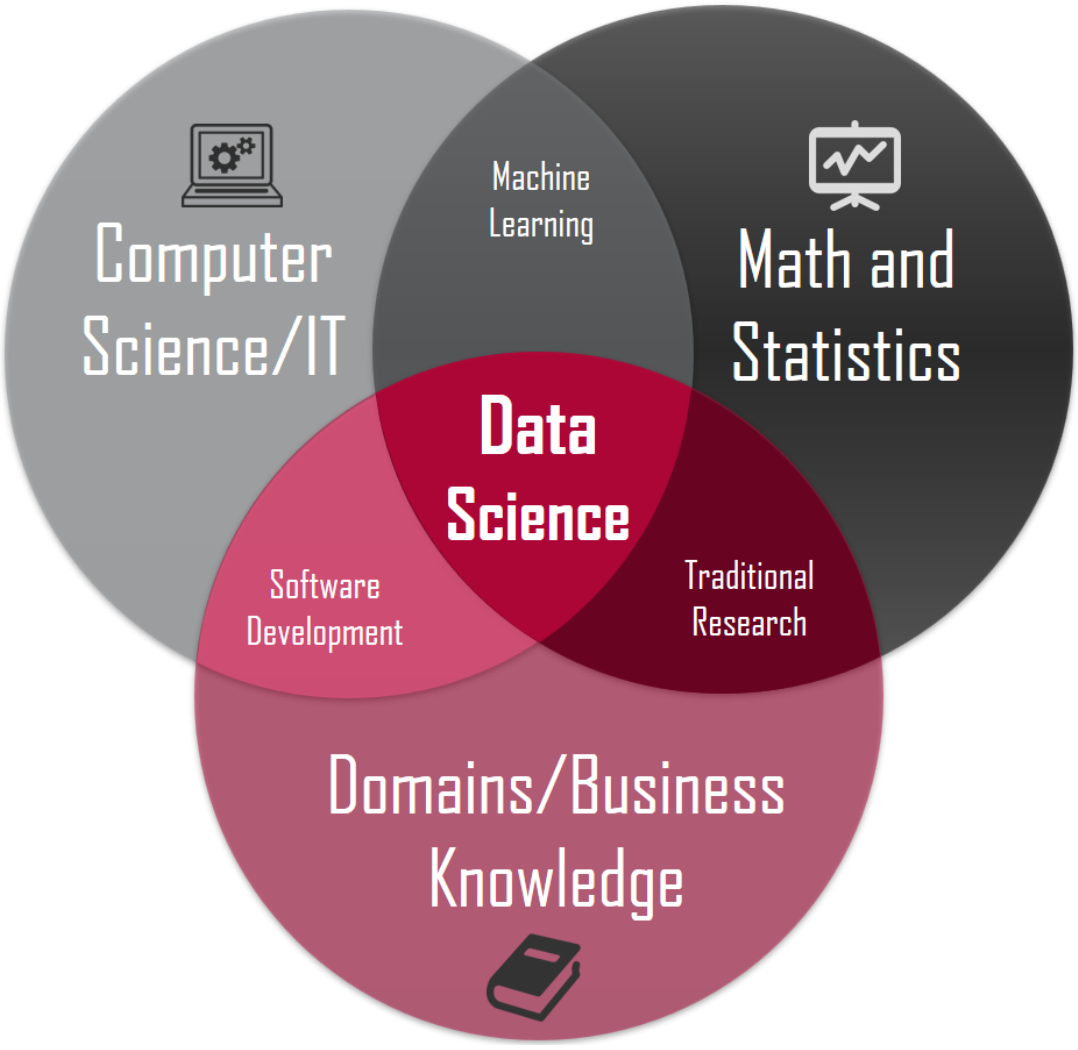


Figure 1

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

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

There are three aspects to learning statistics:

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

In this course, there is an emphasis on the **theory** aspect of statistics.

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

A probability example

In Brunei, boba tea is a sensation, with countless shops vying for the title of the "best" through customer satisfaction ratings. *Tapioca Treasure*, one of the crowd favorites, boasts an impressive 4.3 out of 5-star rating—based on a simple question: *Do you like their bubble tea?*

Intrigued, you and a friend decide to visit and sample their offerings. This raises some interesting **probability**-based questions:

Exercise 0.0.1.

a. What is the probability that you like the bubble tea?

b. What is the probability that at least one of you likes the bubble tea?

c. Suppose you invite your entire family of size (n). What is the expected number of people who will enjoy the bubble tea?

Figure 2: Credits: Unsplash @Snappr.

[Of course, these questions involve certain assumptions, which we'll discuss in a bit.]

Solution 0.0.1. Let X represent whether a person likes the bubble tea at *Tapioca Treasure*. We may write

$$X = \begin{cases} 1 & \text{likes tea} \\ 0 & \text{does not like tea} \end{cases}$$

which is a convenient way of "coding" the qualitative response of like / do not like into numbers (1/0). Suppose that X is a random variable, then we may also write

$$\mathrm{P}(X = 1) := \theta = 0.86,$$

based on the shop's 4.3/5 star rating, interpreted as an 86% likelihood of liking the bubble tea. Since X is a binary random variable (i.e. takes only two outcomes), X is said to follow a *Bernoulli* distribution[1].

a. Since $X \sim \text{Bernoulli}(\theta)$, the probability that you like the bubble tea is simply:

$$P(X = 1) = \theta = 0.86.$$

b. Let Y denote whether you or your friend like the bubble tea. Assume the two events are independent. The probability that at least one of you likes the bubble tea is given by:

$$\begin{aligned} \mathrm{P}(Y \geq 1) &= 1 - P(\text{neither likes it}) \\ &= 1 - P(X_1 = 0)P(X_2 = 0), \end{aligned}$$

where X_1 and X_2 represent your and your friend's preferences, respectively. Substituting the values:

$$\begin{aligned} P(Y \geq 1) &= 1 - (1 - \theta)(1 - \theta) \\ &= 1 - (1 - 0.86)^2 \\ &= 1 - 0.14^2 \\ &= 1 - 0.0196 \\ &= 0.9804. \end{aligned}$$

Thus, there is approximately a 98.04% chance that at least one of you will like the bubble tea.

c. If you invite your entire family of size n, the number of people who like the bubble tea S is known to follow a Binomial distribution:

$$S \sim \text{Binomial}(n, \theta).$$

Using properties of the Binomial distribution, the expected value of a Binomial random variable is given by:

$$E(S) = n\theta.$$

Substituting the values:

$$E(S) = 0.86n.$$

[1]Which is a special case of the binomial distribution which you might be more familiar with ($n = 1$).

⚠ Assumptions

In part (b) above, we came to the solution by assuming that you and your friend have the same probability of liking the tea. In other words, your preferences are independent of each other[2].
In fact, for the binomial distribution, this same assumption must be met for all your family members.

A statistical example

Suppose you're hired by a new boba tea shop, *Pearl Paradise*, to determine whether their signature drink is as good as their rival's, *Tapioca Treasure*.

(a) Pearl Paradise (b) Tapioca Treasure

Figure 3: Boba tea companies.

The questions we have to answer are:

Exercise 0.0.1.

a. What is the rating (θ) for Pearl Paradise?
b. Is $\theta = 0.86$ or not?
c. How confident are we in our estimate?

Notice how these questions are fundamentally different from the previous questions. Previous calculations are "straightforward" if you know probabilities and distribution theory,

[2]Non-independence could be you liking whatever your friend likes! You know, becase you're BFFs.

since the θ value is given. Here, you're dealing with the fact that the θ values is **unknown**, and somehow is the focus of attention.

The other thing you might realise is that there is no way of answering the questions *without having data points* to infer from. This is the difference between statistics and probability. The above three questions implicitly describe the three main activities concerning statistical inference: 1. Point estimation; 2. Hypothesis testing; and 3. Interval estimation.

For now, let us assume that we may collect some data, to at least answer the first part (a). You conduct a survey of 10 random individuals, and ask them the question "Do you like the bubble tea from Pearl Paradise?". Here are the responses to the survey:

```
n <- 10
theta_pp <- 0.8  # PRETEND YOU DON'T KNOW THIS

X <- rbinom(n = n, size = 1, prob = theta_pp)
X
```

```
 [1] 1 1 1 1 1 1 0 1 1 1
```

Let us denote X_i to be the response for individual i to the survey. So in the above, $X_1 = 1$, $X_2 = 1$, and so on. The assumption we make here, similar to the probability question above, is that $X_i \sim \text{Bern}(\theta)$ independently. Let's now work through the solutions:

Solution 0.0.1.

a. If θ represents the proportion of people who like the bubble tea from Pearl Paradise, it would make sense to count the number of people who like the bubble tea, and divide by the total number of responses. Mathematically,

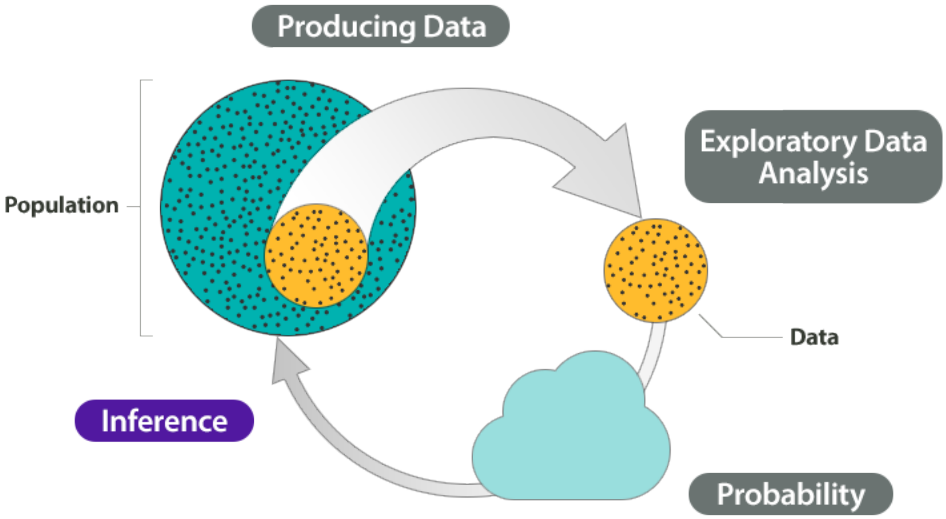
$$\frac{1}{n}(X_1 + \cdots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n$$

So we plug in the numbers and get $\bar{X}_n = 9/10 = 0.9$.

We often denote the *estimate* of θ by its hat version, so we often write

$$\hat{\theta} = 0.9.$$

Variability in your answer



I am sure you notice, that the estimate $\hat{\theta}$ will depend on *who you ask* in the survey. In the context of Pearl Paradise, the population represents all potential customers who could provide their opinion about the boba tea, while the sample refers to the subset of individuals surveyed.

Consider a *random sample* of size n (in the example above, $n = 10$):

$$\mathcal{S} = \{X_1, \ldots, X_n\}.$$

Here X_i are concrete numbers or data regarding the population which is pertinent to answer your statistical question. In statistical inference, we use the sample data to estimate characteristics of the population, such as the proportion of customers who like Pearl Paradise's boba tea. Since it is often impractical to survey the entire population, we rely on the sample to draw conclusions, recognizing that there is uncertainty due to sampling variability.

Of course, a lot of things can influence this, such as demographics, time period, circumstances, non-response rates, etc. – and there's a lot of work to ensure "representativeness" of a sample, but our course does not deal with this.

But even if all things perfect, there is inherent randomness due to sampling itself. We can simulate this by repeatedly conducting the survey of 10 people. Here are the results:

```
B <- 20  # number of repeated surveys
res <- list()
for (i in 1:B) {
  res[[i]] <- rbinom(n = n, size = 1, prob = theta_pp)
}

tab <-
```

Survey no.	X	$\hat{\theta}$
1	0,1,1,0,1,1,0,1,1,0	0.6
2	1,1,1,1,1,1,1,1,1,0	0.9
3	1,1,1,1,0,0,0,1,0,1	0.6
4	1,0,0,1,0,1,1,1,1,0	0.6
5	1,0,1,1,1,1,0,0,1,1	0.7
6	1,1,1,1,1,1,0,1,1,0	0.8
7	1,1,1,0,1,1,1,1,1,1	0.9
8	1,1,1,1,1,1,1,1,1,1	1
9	1,1,1,1,1,0,1,1,1,1	0.9
10	1,1,1,1,0,0,0,1,1,1	0.7
11	1,1,1,1,1,1,0,1,1,1	0.9
12	1,1,1,0,1,1,1,1,1,0	0.8
13	1,0,1,1,1,1,1,1,1,1	0.9
14	1,0,1,1,1,1,1,0,0,1	0.7
15	1,1,1,1,1,1,1,1,1,1	1
16	1,1,0,0,0,0,1,1,1,1	0.6
17	1,1,0,1,1,1,1,0,1,1	0.8
18	0,1,1,1,0,0,0,1,1,1	0.6
19	1,1,0,1,1,1,1,1,1,0	0.8
20	0,1,1,0,0,1,1,1,1,1	0.7

```
  tibble(
    survey = 1:B,
    X = res
  ) |>
  mutate(
    theta_hat = sapply(X, mean),
    X = sapply(X, \(x) paste0(x, collapse = ","))
  )

gt(tab) |>
  fmt_markdown() |>
  cols_label(
    survey ~ "Survey no.",
    theta_hat ~ gt::md("$\\hat\\theta$")
  ) |>
  tab_options(quarto.disable_processing = TRUE)
```

Suppose we conducted more and more of the surveys, we could even get more information regarding *the variability of the estimator*. Ideally, we could even plot a histogram to show the *distribution* of the estimator.

```r
B <- 100000  # number of repeated surveys
theta_hats <- c()
for (i in 1:B) {
  theta_hats[i] <- mean(rbinom(n = n, size = 1, prob = theta_pp))
}

tibble(
  theta_hat = theta_hats
) |>
  ggplot(aes(x = theta_hat)) +
  geom_histogram(fill = "lightblue", col = "black", binwidth = 0.05) +
  scale_x_continuous(breaks = seq(0, 1, by = 0.1))
```

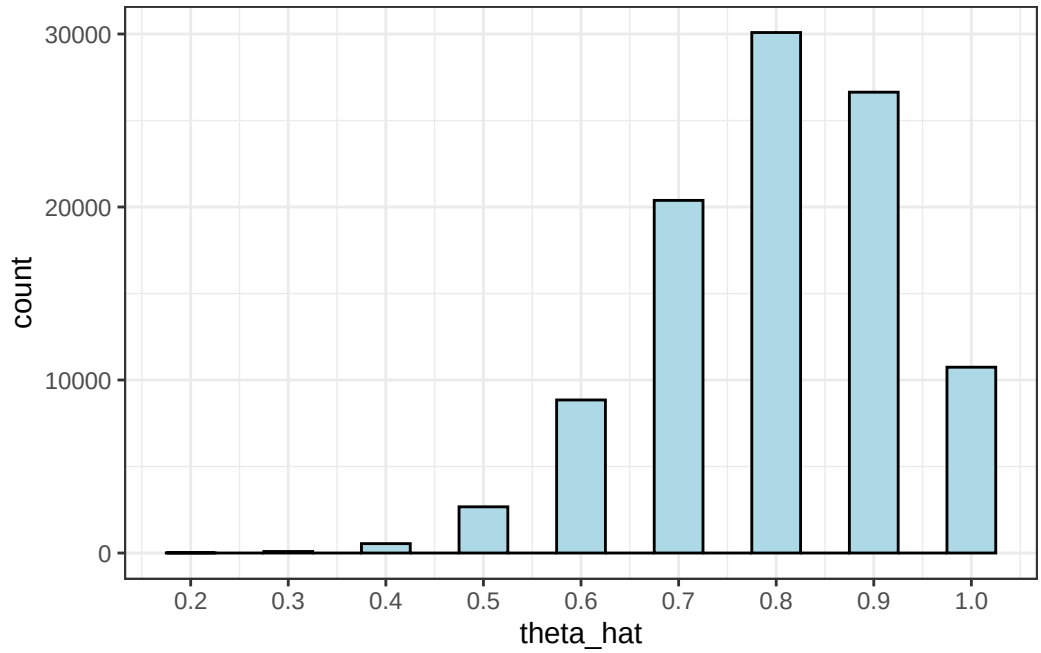


Figure 4: Histogram of estimated $\hat{\theta}$ values in repeated sampling of 100,000 surveys.

In Figure 4 above, it is really evident that under repeated sampling, it is obvious to pick "the best" estimator value $\hat{\theta}$, say, by choosing the value corresponding to the highest bar.

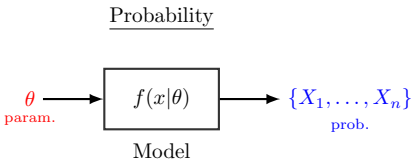
But we cannot "do" repeated sampling most of the time. It is too tiresome, expensive, and just not feasible or impossible sometimes! (Think clinical trials… can you repeat the trial 10,000 times?!). Can we still do something about it then? Yes! By studying statistics from a mathematical angle, we can come up with neat results to come up with reliable statements about the variability of estimators.

Take aways

1. Estimators (such as $\hat{\theta} = \bar{X}_n$) are functions of random variables, and therefore **are themselves random** (i.e. have variability).
2. Figuring out the distribution of estimators is the central idea around statistics.
3. From the distribution of $\hat{\theta}$, we can know
    a. $\mathrm{E}(\hat{\theta})$ – the average value to expect.
    b. $\mathrm{Var}(\hat{\theta})$ – how variable is my estimator?
    c. $\mathrm{P}(a < \hat{\theta} b)$ – how confident am I that my estimator lies in a certain interval?

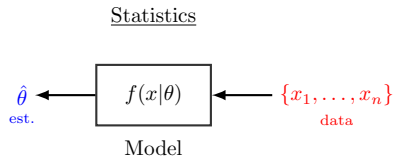
Probability vs statistics

```
\usetikzlibrary{fit,positioning,shapes.geometric,decorations.pathreplacing,calc}
\begin{tikzpicture}[scale=0.8, transform shape]
\tikzstyle{obsvar}=[rectangle, thick, minimum size = 10mm,draw =black!80, node distance =
\tikzstyle{connect}=[-latex, thick]

\node[obsvar] (fx) [] {$\hspace{1em}f(x|\theta)\hspace{1em}$};
\node (xx) [right=of fx] {\textcolor{blue}{$\{X_1,\dots,X_n\}$}};
\node (theta) [left=of fx] {\textcolor{red}{$\theta$}};
\node (d1) [below=of fx,yshift=9mm] {Model};
\node (d2) [below=of xx,yshift=11mm] {\scriptsize \textcolor{blue}{prob.}};
\node (d3) [below=of theta,yshift=11mm] {\scriptsize \textcolor{red}{param.}};
\node (d1) [above=of fx,yshift=-5mm] {\underline{Probability}};

\path (fx) edge [connect] (xx)
      (theta) edge [connect] (fx);

\end{tikzpicture}
```

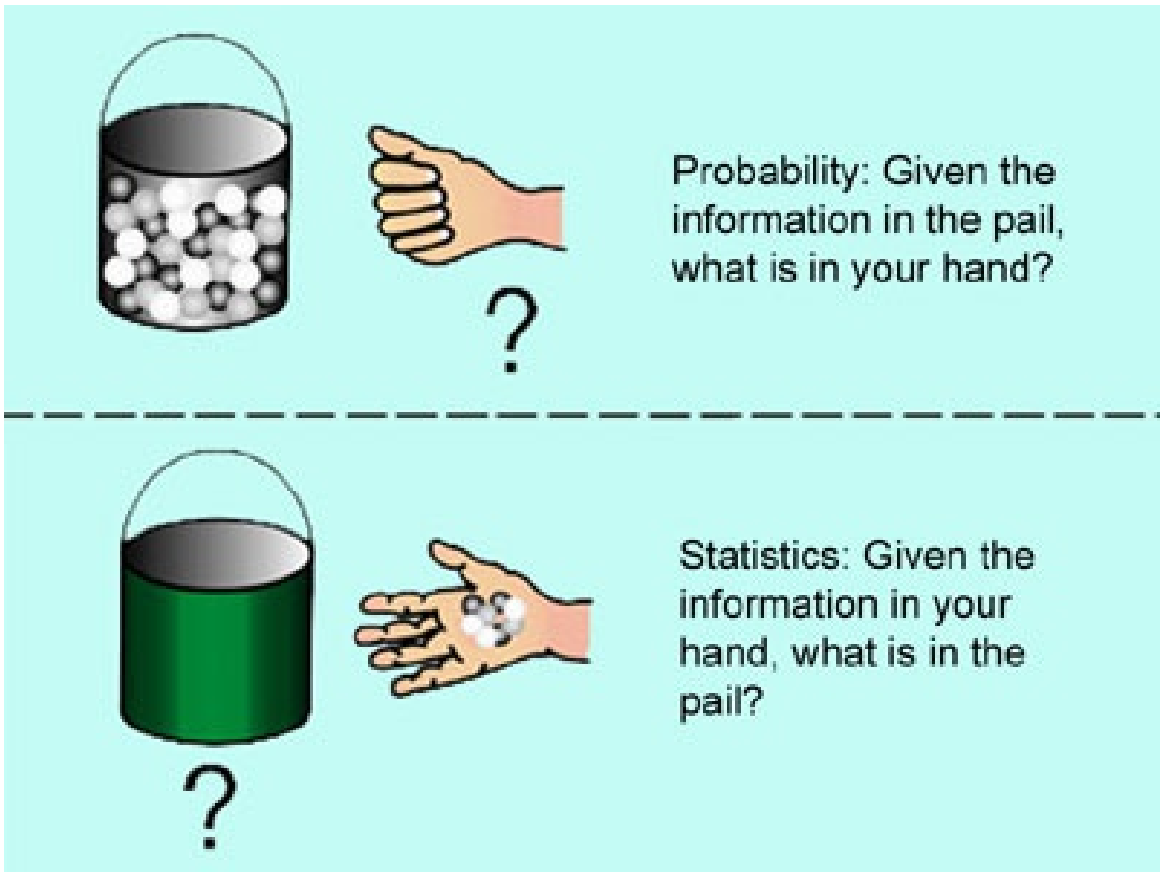


- What is $\mathrm{E}(X)$?
- What is $\mathrm{P}(X > a)$?

```
\usetikzlibrary{fit,positioning,shapes.geometric,decorations.pathreplacing,calc}
\begin{tikzpicture}[scale=0.8, transform shape]
\tikzstyle{obsvar}=[rectangle, thick, minimum size = 10mm,draw =black!80, node distance =
\tikzstyle{connect}=[-latex, thick]
\node[obsvar] (fx) [] {$\hspace{1em}f(x|\theta)\hspace{1em}$};
\node (xx) [right=of fx] {\textcolor{red}{$\{x_1,\dots,x_n\}$}};
\node (theta) [left=of fx] {\textcolor{blue}{$\hat\theta$}};
\node (d1) [below=of fx,yshift=9mm] {Model};
\node (d2) [below=of xx,yshift=11mm] {\scriptsize \textcolor{red}{data}};
\node (d3) [below=of theta,yshift=11mm] {\scriptsize \textcolor{blue}{est.}};
\node (d1) [above=of fx,yshift=-5mm] {\underline{Statistics}};

\path (fx) edge [connect] (theta)
      (xx) edge [connect] (fx);

\end{tikzpicture}
```



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


Probability: Given the information in the pail, what is in your hand?
?
Statistics: Given the information in your hand, what is in the pail?
?

1 Probability Theory Primer

Hello, Students!

Students, write your notes in the corresponding `.qmd` files under the `students/` folder. The Editor will then merge everything into one cohesive notes.
BTW, for more callout blocks, look here: https://quarto.org/docs/authoring/callouts.html

2 Random Sampling

When experimenting, the data collected can be represented or modeled as a set of **random variables** that describe the observed values.

We can model this by assuming $X = X_1, \dots, X_n$ sampled from a population whose pdf or pmf is $f_x(x)$

We can write this as

$$X_1, \dots, X_n \sim f_x(x)$$

Key Assumption : The distribution f_x is imposed on data as a model.

2.1 Independent and Identically Distributed (IID) Samples

The samples are assumed to be:

- Independent: No relationship between observations.
- Identical Distribution: Each sample follows the same probability distribution.

As a result, the **joint probability density function (pdf)** can be written as:

$$f_X(X_1, \dots, X_n) = \prod_{i=1}^{n} f(X_i)$$

3 Definition 1 (Statistics)

A statistic is any function :::

Let

$$X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2).$$

What is the distribution of the sum of squares?

Check out https://quarto.org/docs/authoring/cross-references.html#theorems-and-proofs for more div types (like theorem, lemmas, proofs, etc.).

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis sagittis posuere ligula sit amet lacinia. Duis dignissim pellentesque magna, rhoncus congue sapien finibus mollis. Ut eu sem laoreet, vehicula ipsum in, convallis erat. Vestibulum magna sem, blandit pulvinar augue sit amet, auctor malesuada sapien. Nullam faucibus leo eget eros hendrerit, non laoreet ipsum lacinia. Curabitur cursus diam elit, non tempus ante volutpat a. Quisque hendrerit blandit purus non fringilla. Integer sit amet elit viverra ante dapibus semper. Vestibulum viverra rutrum enim, at luctus enim posuere eu. Orci varius natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus.

Figure 3.1: Caption for the image.

Figure 3.1 shows an image.

4 Definition 5

Let $Z_1, ..., Z_k$

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis sagittis posuere ligula sit amet lacinia. Duis dignissim pellentesque magna, rhoncus congue sapien finibus mollis. Ut eu sem laoreet, vehicula ipsum in, convallis erat. Vestibulum magna sem, blandit pulvinar augue sit amet, auctor malesuada sapien. Nullam faucibus leo eget eros hendrerit, non laoreet ipsum lacinia. Curabitur cursus diam elit, non tempus ante volutpat a. Quisque hendrerit blandit purus non fringilla. Integer sit amet elit viverra ante dapibus semper. Vestibulum viverra rutrum enim, at luctus enim posuere eu. Orci varius natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus.

Write R code like this:

````
```{r}
1 + 1
```
````

Which will be rendered as:

```
1 + 1
```

```
[1] 2
```

> **Tip**
>
> Feel free to rearrange the sections however you wish. Below, you will find some extra bits that you may want to include as part of your notes, or just leave them as the appendix.

4.1 Algebras of sets

This is the mathematical structure that allows us to *observe* and *measure* random events. Logically,

1. If an event A can be observed, then its complement can be too. I.e. $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$.
2. At least one outcome can be observed, i.e. $\Omega \in \mathcal{F}$.
3. If two or more events are observed, then at least one of them (or both) can be observed, i.e.
$$A, B \in \mathcal{F} \Rightarrow A \cup B \in \mathcal{F}$$

If 1–3 holds, then $\mathcal{F}$ is said to be an *algebra* over Ω. In addition, if you can "add" up infinitely many countable things, $\mathcal{F}$ is called a *σ-algebra*.

$$A_1, A_2, \cdots \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$$

4.2 Why σ-algebra?

In probability theory and statistics, an experiment (or trial) is a procedure that can be repeated and has a well-defined set of possible outcomes (known as the sample space Ω). Events are thought of as being subsets of Ω, while probabilities are merely a mapping from some *event space* $\mathcal{F}$ to $[0, 1]$.

To make this idea concrete, for the die roll example, $\Omega = \{1, \ldots, 6\}$, while an event could be $E = \{2, 4, 6\} \subset \Omega$ (getting an even number). The probability of the event E occurring is $\mathrm{P}(E) = \frac{1}{2}$–so it indeed behaves like a function, taking input some event and spitting out a number between 0 and 1.

Note here that $\mathcal{F}$ is not Ω–it has to be bigger than Ω as we're not just interested in singleton outcomes. A good starting point would be $\mathcal{F} = \mathcal{P}(\Omega)$, the set of all subsets of Ω, which should contain all possible events constructed from the set of outcomes.

4.2.1 Rules of probability

Having abstracted the notion of Ω and $\mathcal{F}$, we should also define some rules that the probability function $\mathrm{P} : \mathcal{P}(\Omega) \to [0, 1]$ must follow. Let us list down a few:

i. $\mathrm{P}(E) \geq 0, \forall E$;
ii. $\mathrm{P}(\emptyset) = 0$ and $\mathrm{P}(\Omega) = 1$;[1]
iii. If $E_1 \cap E_2 = \emptyset$, then $\mathrm{P}(E_1 \cup E_2) = \mathrm{P}(E_1) + \mathrm{P}(E_2)$; and
iv. If $E_1, E_2, \ldots$ are mutually disjoint events, then $\mathrm{P}\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mathrm{P}(E_i)$.

[1] $\emptyset = \{\}$ is the empty set.

Indeed, these are quite logical impositions that ensure we don't end up with nonsensical probabilities. For instance, by ii. and iii., modelling a (biased) coin toss by $\mathrm{P}(H) = 0.7$ necessitates $\mathrm{P}(T) = 0.3$ and not anything else, e.g. $\mathrm{P}(T) = 0.5$.

4.2.2 The need for measure theory

We've managed to come up with probability rules so far without the need for measure theory, so what's the problem? The problem is that in the way that we've described it, this is actually too much to ask! There will be instances where this whole framework fails and we can't assign probabilities properly, especially when we need it the most.

Consider that, with all these demands, we can't even define the uniform random variable on $\Omega = [0, 1]$! That is, no mapping $\mathrm{P} : \mathcal{P}([0, 1]) \to [0, 1]$ exists such that $\mathrm{P}([a, b]) = b - a$ for $0 \leq a \leq b \leq 1$ which satisfies all of the rules i. to iv. listed above. For a proof, see the appendix. Evidently some concession has to be made (which one?), and the probability map must be constructed more carefully. The answer lies in measure theory.

4.3 An unmeasurable set

As mentioned, we are unable to define a uniform probability measure on the unit interval, given by

$$\mathrm{P}([a, b]) = b - a$$

that satisfies all the probability rules listed in i. to iv. earlier. On the face of it, all the rules themselves are satisfied: $\mathrm{P}(\Omega) = \mathrm{P}([0, 1]) = 1$, $\mathrm{P}(\emptyset) = \mathrm{P}([a, a]) = 0$ (for any $a \in [0, 1]$), and certainly probabilities of disjoint subsets of $[0, 1]$ are just the sum of the lengths of the intervals.

These are all great properties to have, so we must concede instead on the domain of the probability function, i.e. the event space. The proof of the proposition below in instructive, in that it illustrates the existence of a "non-measurable" set. That is, there are such events (subsets in $[0, 1]$) for which we are unable to assign probabilities to.

Proposition 4.3.1. *There does not exist a definition of* $\mathrm{P} : \mathcal{P}([0, 1]) \to [0, 1]$ *satisfying* $\mathrm{P}([a, b]) = b - a$ *and i. to iv. (as listed earlier).*

Proof. All we need to show is the existence of one such subset of $[0, 1]$ whose measure is undefined. The set we are about to construct is called the Vitali set[2], after Giuseppe Vitali who described it in 1905.

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

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

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

Then we should have

$$\mathrm{P}(A \oplus r) = \mathrm{P}(A).$$

For example, $\mathrm{P}([0.7, 0.9] \oplus 0.2) = \mathrm{P}([0.9, 1] \cup [0, 0.1]) = 0.2$.

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

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

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

- The equivalence class of 0 is the set of real numbers x such that $x \sim 0$, i.e. $[0] = \{y \in [0, 1] \mid y - 0 \in \mathbb{Q}\}$, which is the set of all rational numbers in $[0, 1]$.
- The equivalence class of an irrational number $z_1 \in [0, 1]$ is clearly not in $[0]$, thus would represent a different equivalence class $[z_1] = \{y \in [0, 1] \mid y - z_1 \in \mathbb{Q}\}$.
- Yet another irrational number $z_2 \notin [z_1]$ would exist, i.e. a number $z_2 \in [0, 1]$ such that $z_2 - z_1 \notin \mathbb{Q}$, and thus would represent a different equivalence class $[z_2]$.
- And so on…

The equivalence classes may therefore be represented by $[0], [z_1], [z_2], \ldots$ where z_i are all irrational numbers that differ by an irrational number, and there are uncountably many such numbers, and therefore classes.

Construct the Vitali set V as follows: Take precisely one element from each equivalent class, and put it in V. As a remark, such a V must surely exist by the Axiom of Choice[3].

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

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

As a reminder, the set of rational numbers is countably infinite[4]. We make two observations:

1. **The equivalence relation partitions the interval $[0, 1]$ into a disjoint union of equivalence classes.** In other words, the sets $(V \oplus r)$ and $(V \oplus s)$ are disjoint for any rationals $r \neq s$, such that $r, s \in [0, 1]$. If they were not disjoint, this would mean that there exists some $x, y \in [0, 1]$ with $x + r \in (V \oplus r)$ and $y + s \in (V \oplus s)$ such that $x + r = y + s$. But then this means that $x - y = s - r \in \mathbb{Q}$ so x and y are in the same equivalent class, and this is a contradiction. Importantly,

$$\mathrm{P}\left(\bigcup_r (V \oplus r)\right) = \sum_r \mathrm{P}(V \oplus r) = \sum_r \mathrm{P}(V) \tag{4.1}$$

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

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

2. **Every point in $[0,1]$ is contained in the union $\bigcup_r (V \oplus r)$.** To see this, fix a point x in $[0,1]$. Note that this point belongs to some equivalent class of x, and in this equivalence class there exists some point α which belongs to V as well by construction. Hence, $\alpha \sim x$, and thus $x - \alpha = r \in \mathbb{Q}$, implying that x is a point in the Vitali set V shifted by r. Therefore,

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

and we may write

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

since the measure of any set contained in another must have smaller or equal measure (a relation implied by property iii.[5]) as well as all probabilities are less than equal to 1[6]. We see that

$$\mathrm{P}\left(\bigcup_r (V \oplus r)\right) = 1. \tag{4.2}$$

Equating (4.1) and (4.2) together, we find a contradiction: A countably infinite sum of a constant value can only equal 0, $+\infty$ or $-\infty$, but never 1. □

4.4 Conditional probability

> The latest estimate puts the proportion of geology students at FOS to be 5%. A randomly selected student from FOS, Nafeesah, is described by her peers as someone who loves the outdoors and gets overly excited when shown something that is related to rocks.

Which statement is more likely?

A. Nafeesah is undertaking a BSc Geology programme.
B. Nafeesah is not undertaking a BSc Geology programme.

Let

- E be the 'evidence'
- G be the event that a student takes Geology

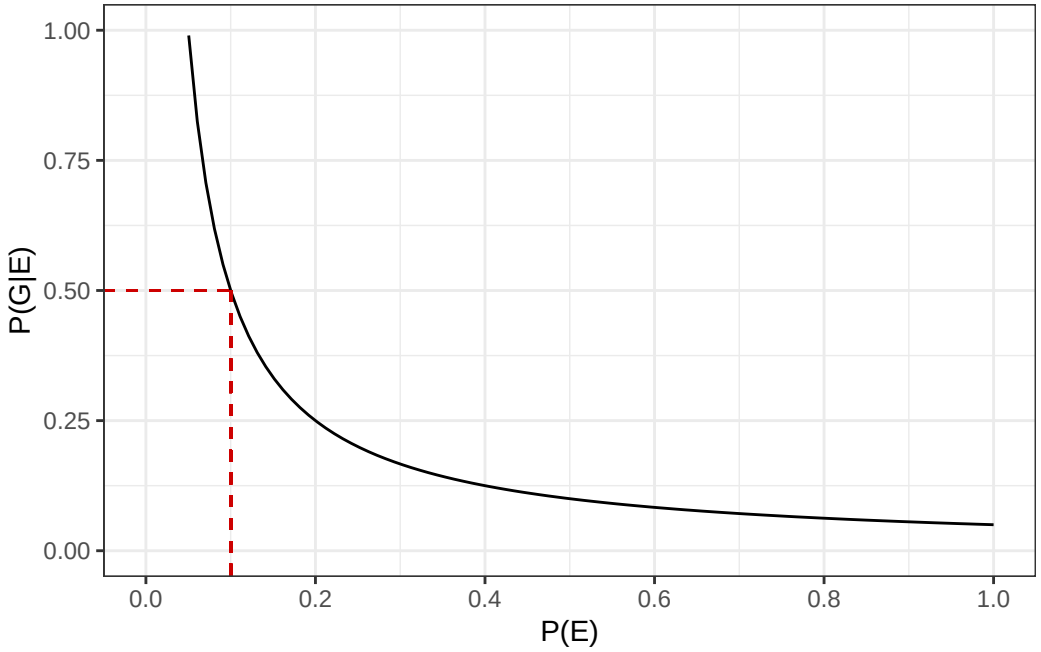
Then

$$\mathrm{P}(G|E) = \frac{\mathrm{P}(E|G)\,\mathrm{P}(G)}{\mathrm{P}(E)} \approx \frac{0.05}{\mathrm{P}(E)}$$

[5] Let A and B be such that $A \subseteq B$. Then we may write $B = A \cup (B \setminus A)$ where the sets A and $B \setminus A$ are disjoint. Hence, $\mathrm{P}(B) = \mathrm{P}(A) + \mathrm{P}(B \setminus A)$, and since probabilities are non-negative, we have that $\mathrm{P}(B) \geq \mathrm{P}(A)$.

[6] For any A, $\mathrm{P}(\Omega) = \mathrm{P}(A \cup A^c) = \mathrm{P}(A) + \mathrm{P}(A^c) = 1$, so $\mathrm{P}(A) \leq 1$.

```r
x <- seq(1e-10, 1, length = 100)
plot_df <- tibble(
  x = x,
  y = 0.05 / x
)

ggplot(plot_df, aes(x, y)) +
  geom_line() +
  geom_segment(x = -Inf, xend = 0.05 / 0.5, y = 0.5,yend = 0.5,
               linetype = "dashed", col = "red3") +
  geom_segment(x = 0.05 / 0.5, xend = 0.05 / 0.5, y = 0.5, yend = -Inf,
               linetype = "dashed", col = "red3") +
  scale_y_continuous(limits = c(0, 1)) +
  scale_x_continuous(breaks = seq(0, 1, by = 0.2)) +
  labs(x = "P(E)", y = "P(G|E)")
```



4.5 Bayesian statistics

Sometime between 1746 and 1749, Rev. Thomas Bayes conducted this experiment.

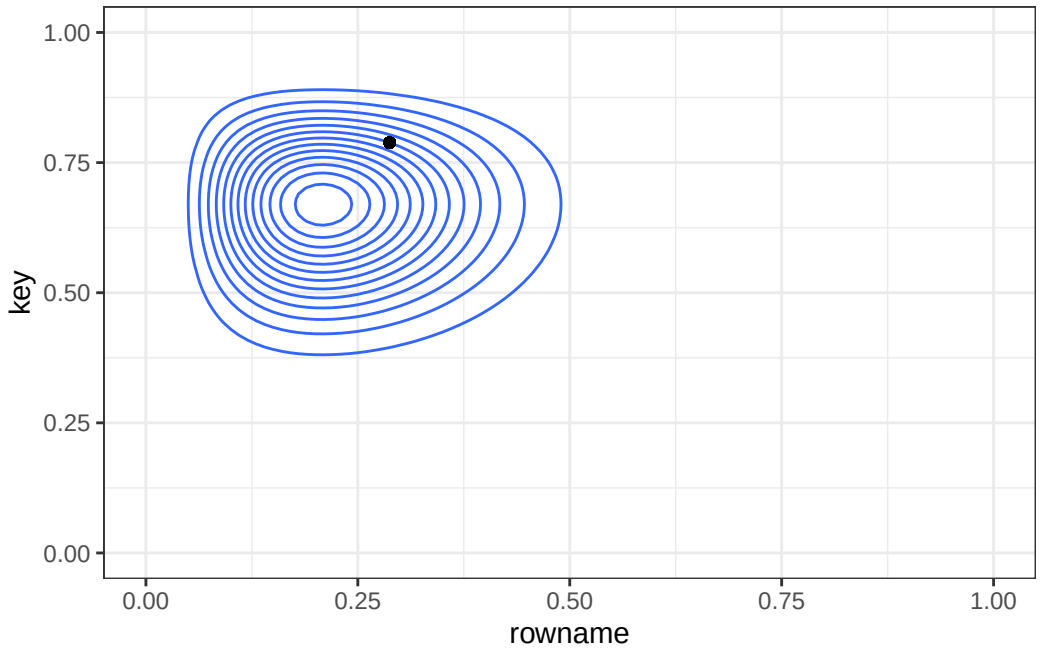
> Imagine a square, flat table. You throw a marker (e.g. a coin) but do not know where it lands. You ask an assistant to randomly throw a ball on the table. The assistant informs you whether it stopped to the left or right from the first ball. How to use this information to better estimate where your marker landed?

The Bayesian principle is about updating beliefs.

- Let $X \in [0, 1]$ be the location of the ball on a horizontal axis.
- Before any new information, any position X is possible, say $X \sim \text{Unif}(0, 1)$.
- Let Y be the number of times the assistant's ball landed left of the marker after n throws. Then $Y|X \sim \text{Bin}(n, X)$.
- What we want is information regarding $X|Y$, which is obtained using Bayes Theorem

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

```r
# https://dosreislab.github.io/2019/01/27/ballntable.html
set.seed(123)
n <- 15
xy <- runif(2) # position of coin after intial throw
xy.2 <- matrix(runif(2 * n), ncol=2) # additional n throws of the ball

pos <- numeric(2)
pos[1] <- sum(xy.2[,1] < xy[1])
pos[2] <- sum(xy.2[,2] < xy[2])

jointf <- function(pos = pos, n = n, N=100) {
  a <- pos[1]; b <- pos[2]
  x <- y <- seq(from=0, to=1, len=N)
  xf <- x^a * (1-x)^(n-a)
  yf <- y^b * (1-y)^(n-b)
  z <- xf %o% yf
}

Cf <- function(x, y, n) {
  ( factorial(n+1) )^2 /
  ( factorial(x) * factorial(n-x) * factorial(y) * factorial(n-y) )
}

z <- jointf(pos, n) * Cf(pos[1], pos[2], n)

as.data.frame(z) %>%
  `colnames<-`(1:100 / 100) %>%
  rownames_to_column() %>%
  gather(key, value, -rowname) %>%
  mutate(rowname = as.numeric(rowname) / 100,
         key = as.numeric(key)) %>%
  ggplot(aes(rowname, key, z = value)) +
  geom_contour() +
  geom_point(x = xy[1], y = xy[2]) +
  lims(x = c(0, 1), y = c(0, 1))
```



4.6 Probability integral transform

Theorem 4.6.1. *Let X have continuous cdf $F_X(x)$ and define the random variable Y as $Y = F_X(X)$. Then Y is uniformly distributed on $(0,1)$, that is $f_Y(y) = 1\ \forall y \in [0,1]$ with $\mathrm{P}(Y \leq y) = y$.*

Proof.

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

□

The PIT is a special kind of transformation, useful for various statistical purposes. Suppose we wish to generate $X \sim F_X$–this is done via $X = F_X^{-1}(U)$ where $U \sim \mathrm{Unif}(0,1)$.

5 Commonly used probability models

5.1 Poisson-Binomial relationship

The Poisson distribution plays a useful approximation role for the binomial:

$$X \sim \text{Bin}(n, p) \quad \Rightarrow \quad X \approx \text{Poi}(np)$$

when n is large ($n > 20$) and np is small ($np < 5$). The reason is the Poisson can be seen as the *limiting case* to the binomial as $n \to \infty$ while $\text{E}(X) = np$ remains fixed.

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

The reason is that the Poisson can be seen as the *limiting case* to the binomial as $n \to \infty$ while $\text{E}(X) = np$ remains fixed.

```
library(tidyverse)
poibin_df <- function(n, p, x = 0:10) {
  lambda <- n * p
  the_title <- paste0("n = ", n, ", p = ", p)

  tibble(
    x = x,
    bin = dbinom(x, size = n, prob = p),
    poi = dpois(x, lambda = n * p)
  ) %>%
    pivot_longer(-x) %>%
    mutate(title = the_title)
}

plot_df <- bind_rows(
```

```
  poibin_df(20, 0.05),
  poibin_df(10, 0.3),
  poibin_df(100, 0.3, 20:30),
  poibin_df(1000, 0.01)
)
mylevels <- unique(plot_df$title)
plot_df$title <- factor(plot_df$title, levels = mylevels)
# levels(plot_df$title) <- mylevels

ggplot(plot_df, aes(x, value, fill = name)) +
  geom_bar(stat = "identity", position = "dodge", alpha = 0.7) +
  facet_wrap(. ~ title, ncol = 2, scales = "free") +
  scale_x_continuous(breaks = 0:100) +
  # scale_fill_manual(values = c(palgreen, palred)) +
  labs(y = "P(X=x)", col = NULL, fill = NULL) +
  theme(legend.position = "top")
```

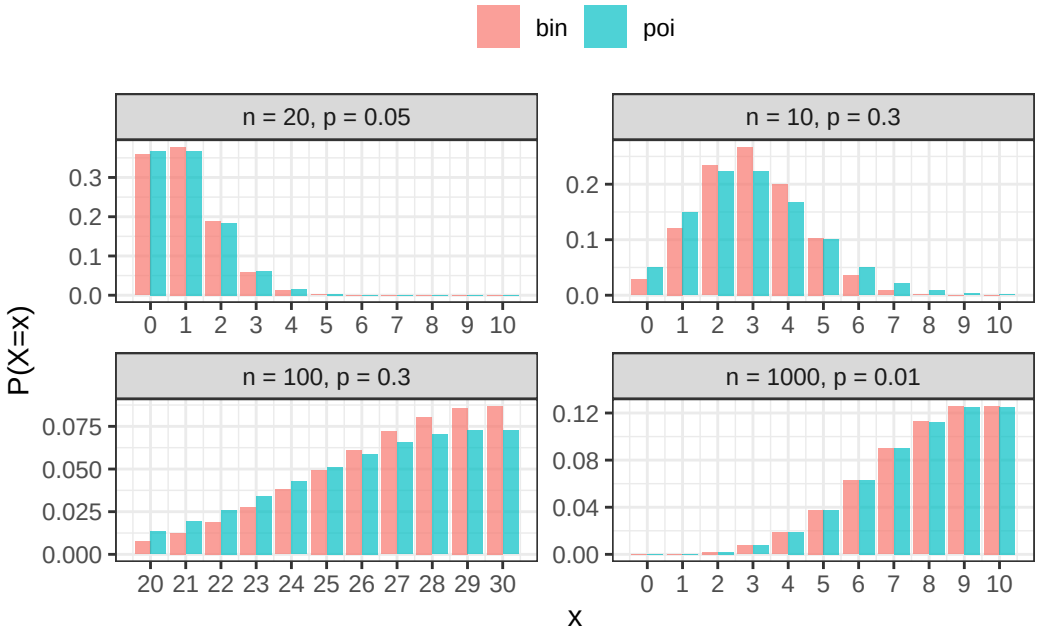


Figure 5.1

5.2 Memoryless property

X is a positive rv and memoryless, in the sense that for all $t > s > 0$,

$$\mathrm{P}(X > t + s \mid X > s) = \mathrm{P}(X > t)$$

if and only if it is exponentially distributed[1].

> Given that we have been waiting for s units of time, the probability that we wait a further t units of time is independent to the first fact!

Example 5.2.1. Assume that bus waiting times are exponentially distributed, and you are concerned about the event $A =$ a bus arrives in the next minute. Let $p_i = \mathrm{P}(A|B_i)$ where

i. $B_1 =$ you just arrived to the station; and
ii. $B_2 =$ you've been sitting there for 20 minutes already.

Then $p_1 = p_2$.

5.3 Relationships

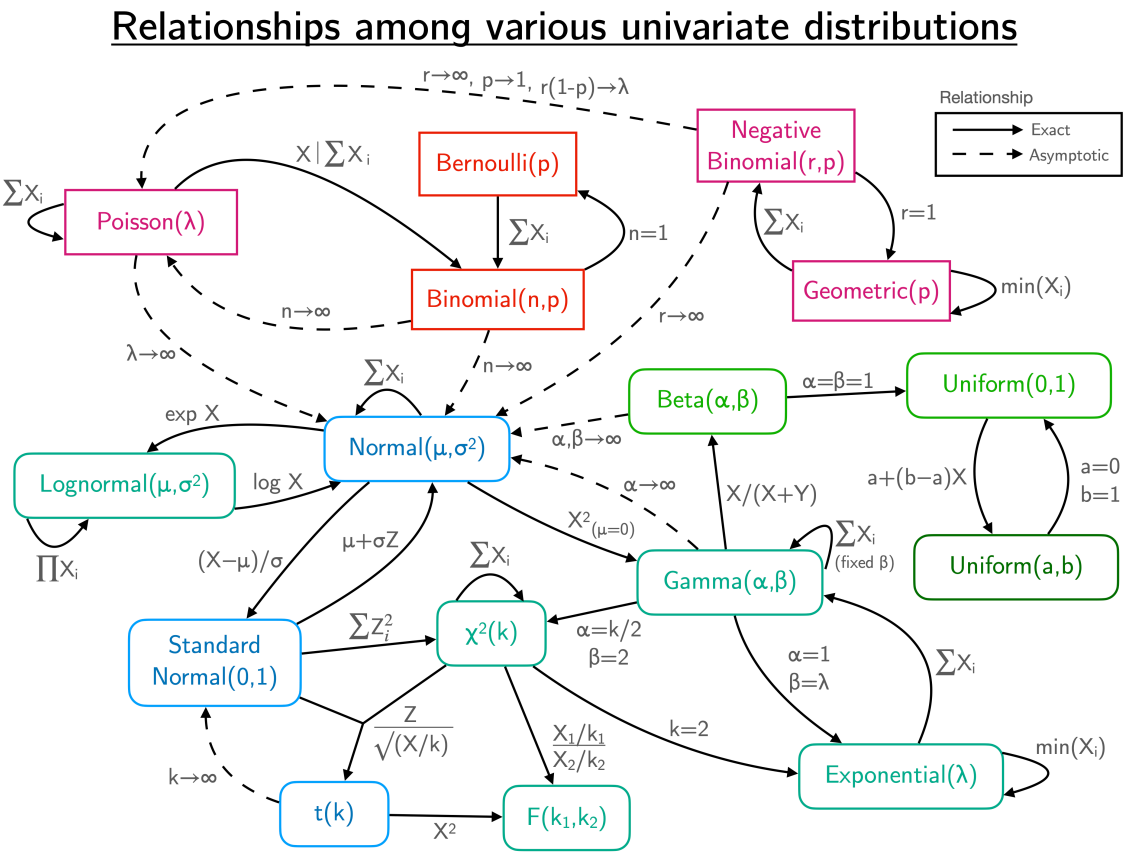


Figure 5.2: Relationships among various univariate distributions.

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

6 Sampling from the normal distribution

6.1 An example

Example 6.1.1. Consider the time to failure (in years) for circuit boards modelled by $\text{Exp}(\lambda)$. An independent random sample $X_1, \dots, X_n$ was collected. What is the probability that the minimum time lasted is more than 2 years?

$$\begin{aligned}
\text{P}(\min\{X_1, \dots, X_n\} > 2) &= \text{P}(X_1 > 2, \dots, X_n > 2) \\
&= \text{P}(X_1 > 2) \cdots \text{P}(X_n > 2) \\
&= [\text{P}(X_1 > 2)]^n \\
&= [e^{-2/\lambda}]^n \\
&= e^{-2n/\lambda}
\end{aligned}$$

A reasonable estimator for $1/\lambda$ is $T_n = \min\{X_1, \dots, X_n\}$.

6.2 Finite vs inifinite population

Infinite population

Implicitly iid: "Removing" $X_1 = x_1$ from the population does not affect the probability distribution for the subsequent samples. Why "infinite"? In scenarios where the exact population size is either unknown, uncountable, or effectively limitless, it is simpler to treat it as infinite.

Finite population

Not necessarily iid, depending on the sampling method:

- Sampling without replacement
- Cluster sampling
- Stratified sampling
- etc.

Standard errors calculations are affected here. Check out Finite Population Correction factors if interested.

Example 6.2.1. Estimate the average height of goalkeepers. Which ones? Presumably all of them–past, present, and future–in all leagues. For all intents and purposes, this is an infinite population.

6.3 An experiment

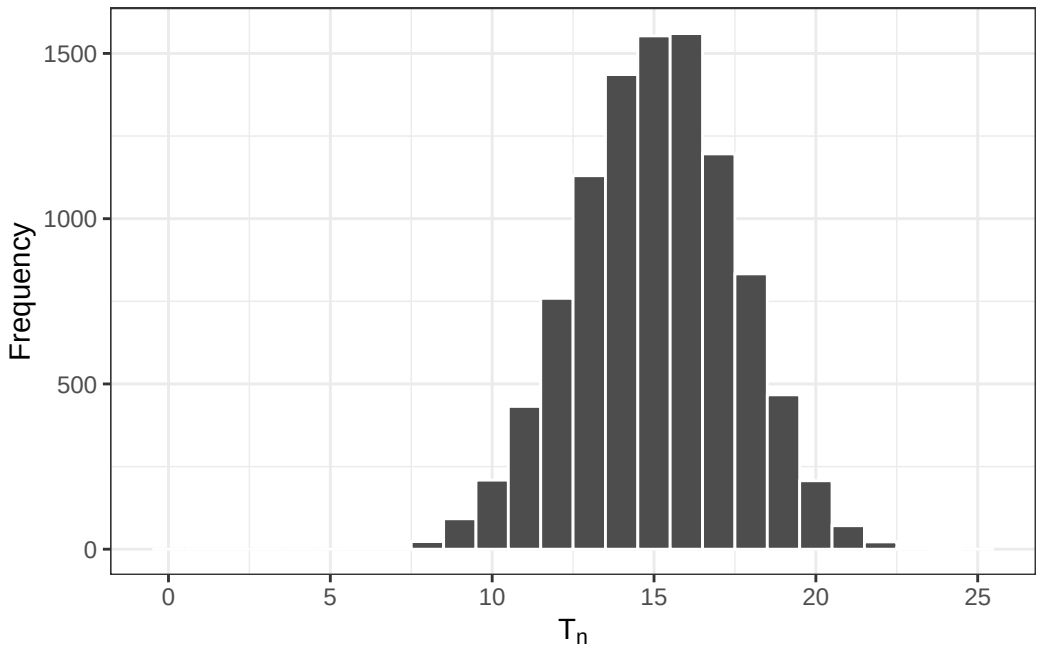
Using R, we can draw multiple instances of the statistic T_n. Let $n = 25$ and $p = 0.6$ (true value).

1. Draw $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bern}(p)$
2. Compute $T_n = \sum_{i=1}^{n} X_i$
3. Repeat steps 1–2 a total of $B = 10000$ times

```
n <- 25
p <- 0.6
B <- 10000
x <- rbinom(B, n, p)

# First 10 values
head(x, 10)
```

```
 [1] 19 14 15 15 16 10 14 17 16 14
```

```
ggplot() +
  geom_histogram(aes(x), fill = "gray30", col = "white",
                 breaks = seq(-0.5, n + 0.5, by = 1)) +
  scale_x_continuous(breaks = seq(0, n, by = 5)) +
  labs(x = expression(T[n]), y = "Frequency")
```



7 Large sample approximations

7.1 Illustration of convergence in probability

While there is no guarantee that the points will eventually stay inside the ϵ-band, the probability of it being outside the band tends to 0.

```
set.seed(123)
eps <- 0.15
plot_df <- tibble(
  x = 1:25,
  y = 1 / x ^ {1/1.2} + rnorm(25, sd = 0.1)) %>%
  mutate(
    y = case_when(y > 0.45 & x > 2 ~ 0.45, TRUE ~ y),
    up = y +  8.5 * eps / (x ^ 1),
    lo = y -  8.5 * eps / (x ^ 1),
    dist = up - lo,
    longer = dist > (2 * eps)
  )  %>%
  filter(x > 2)

ggplot(plot_df, aes(x, y, col = longer)) +
  geom_hline(yintercept = 0, linetype = "dashed", col = "grey60") +
  geom_hline(yintercept = c(eps, -eps), col = "gray") +
  annotate("rect", xmax = Inf, xmin = -Inf, ymax = eps, ymin = -eps, alpha = 0.1) +
  geom_point() +
  geom_errorbar(aes(ymin = lo, ymax = up), width = 0.4) +
  coord_cartesian(ylim = c(-0.3, 0.6), xlim = c(3, 25)) +
  scale_y_continuous(breaks = c(-eps, 0, eps),
                     labels = c(expression(X - epsilon), "X",
                                expression(X + epsilon))) +
  scale_colour_brewer(palette = "Set1") +
  labs(y = expression(X[n]), x = "n") +
  theme(axis.text.x = element_blank(), axis.ticks.x = element_blank()) +
  guides(col = "none") +
  geom_errorbar(aes(x = 2.05, y = 0, ymax = 0.05, ymin = -0.05), col = "black",
                width = 0.4, linewidth = 1)
```

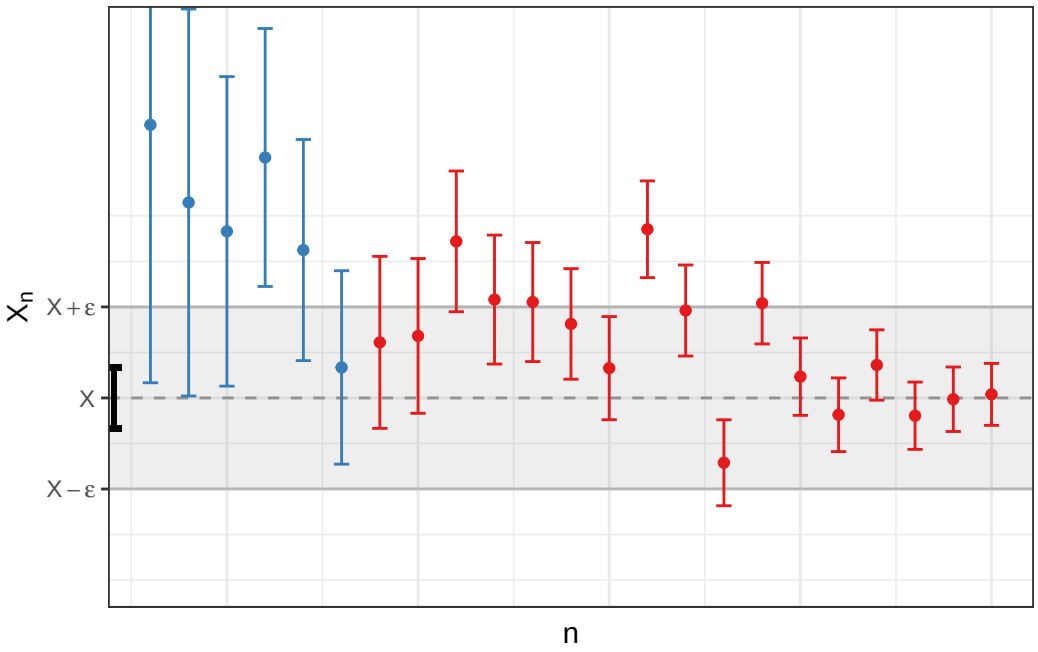


Figure 7.1

7.2 Almost sure convergence

Definition 7.2.1 (Almost sure convergence)**.** X_n converges to X in *almost surely* if for every $\epsilon > 0$,

$$\mathrm{P}\left(\lim_{n\to\infty} |X_n - X| \geq \epsilon\right) = 0.$$

We write $X_n \xrightarrow{\text{a.s.}} X$.

That is, $X_n(\omega) \to X(\omega)$ for all outcomes $\omega \in \Omega$, except perhaps for a collection of outcomes $\omega \in A$ with $\mathrm{P}(A) = 0$. This is stronger than (i.e. it implies, but is not implied by) convergence in probability. There is no relationship between convergence in mean square and convergence almost surely.

7.3 The Strong Law of Large Numbers

With the same setup as for WLLN, a different argument leads to the stronger conclusion as per the result below.

Theorem 7.3.1 (Strong Law of Large Numbers)**.** *Let $X_1, X_2, \ldots$ be iid rvs with mean μ and variance σ^2. Let $\bar{X}_n$ denote the sample mean, i.e.*

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

Then, $\bar{X}_n \xrightarrow{a.s.} \mu$ *as* $n \to \infty$*, i.e.*

$$\mathrm{P}\left(\lim_{n\to\infty} |\bar{X}_n - \mu| < \epsilon\right) = 1.$$

Proof is outside the scope of this module. It's satisfying to know that the SLLN exists, but for our purposes, WLLN suffices!

8 Likelihood theory

8.1 Finding the MLE numerically

Here's how we simulation $n = 100$ random sample from a normal distribution with mean $\mu = 8$ and $\sigma = 1$.

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

```
[1] 7.985771
```

The mean is found to be 7.99. Here's a plot of the log-likelihood function (μ against $\ell(\mu)$):

```
tibble(
  x = mean(X) + seq(-1, 1, length = 100)
) |>
  rowwise() |>
  mutate(y = sum(dnorm(X, mean = x, log = TRUE))) |>
  ggplot(aes(x, y)) +
  geom_line() +
  geom_segment(linetype = "dashed", x = mean(X), xend = mean(X), y = -Inf,
               yend = sum(dnorm(unlist(X), mean = mean(X), log = TRUE)),
               size = 0.4, col = "gray") +
  labs(x = expression(mu), y = expression(l(mu)))
```

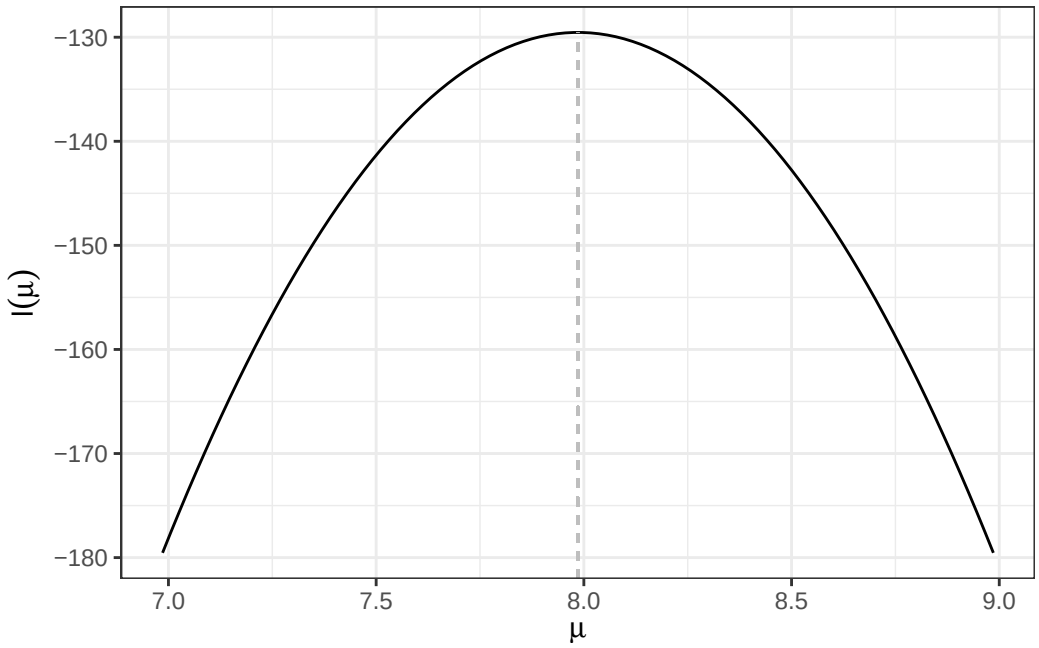


Figure 8.1: Log-likelihood function of the normal mean.

Here's how to optimise the (log-)likelihood function.

```
neg_loglik <- function(theta, data = X) {
  -1 * sum(dnorm(x = data, mean = theta, log = TRUE))
}

res <- nlminb(
  start = 1,  # starting value
  objective = neg_loglik,
  control = list(
    trace = 1  # trace the progress of the optimiser
  ))
```

```
  0:     2569.5905:  1.00000
  1:     575.28195:  5.00000
  2:     129.54040:  7.98579
  3:     129.54040:  7.98577
  4:     129.54040:  7.98577
```

```
glimpse(res)
```

```
List of 6
 $ par        : num 7.99
```

```
$ objective  : num 130
$ convergence: int 0
$ iterations : int 4
$ evaluations: Named int [1:2] 6 7
 ..- attr(*, "names")= chr [1:2] "function" "gradient"
$ message    : chr "relative convergence (4)"
```

8.2 Variance reduction: *Rao-Blackwell*isation

It is possible to reduce the variance of an unbiased estimator by conditioning on a sufficient statistic.

Theorem 8.2.1 (Rao-Blackwell)**.** *Suppose that $\hat{\theta}(\mathbf{X})$ is unbiased for θ, and $S(\mathbf{X})$ is sufficient for θ. Then the function of S defined by*

$$\phi(S) = \mathrm{E}_\theta(\hat{\theta}|S)$$

i. is a statistic, i.e. $\phi(S)$ does not involve θ;
ii. is an unbiased statistic, i.e. $\mathrm{E}(\phi(S)) = \theta$; and
iii. has $\mathrm{Var}_\theta(\phi(S)) \leq \mathrm{Var}_\theta(\hat{\theta})$, with equality iff $\hat{\theta}$ itself is a function of S.

In other words, $\phi(S)$ is a uniformly better unbiased estimator for θ. Thus the Rao-Blackwell theorem provides a systematic method of variance reduction for an estimator that is not a function of the sufficient statistic.

Proof.

i. Since S is sufficient, the distribution of $\mathbf{X}$ given S does not involve θ, and hence $\mathrm{E}_\theta(\hat{\theta}(\mathbf{X})|S)$ does not involve θ.

ii. $\mathrm{E}(\phi(S)) = \mathrm{E}\left[\mathrm{E}(\hat{\theta}|S)\right] = \mathrm{E}(\hat{\theta}) = \theta$.

iii. Using the law of total variance,

$$\begin{aligned}\mathrm{Var}(\hat{\theta}) &= \mathrm{E}\left[\mathrm{Var}(\hat{\theta}|S)\right] + \mathrm{Var}\left[\mathrm{E}(\hat{\theta}|S)\right] \\ &= \mathrm{E}\left[\mathrm{Var}(\hat{\theta}|S)\right] + \mathrm{Var}(\phi(S)) \\ &\geq \mathrm{Var}(\phi(S)),\end{aligned}$$

with equality iff $\mathrm{Var}(\hat{\theta}|S) = 0$, i.e. iff $\hat{\theta}$ is a function of S.

□

Example 8.2.1. Suppose we have data $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Poi}(\lambda)$ pertaining to the number of road accidents per day, and we want to estimate the probability of having no accidents $\theta = e^{-\lambda} = \text{P}(X_i = 0)$.

An unbiased estimator of θ is

$$\hat{\theta}(\mathbf{X}) = \begin{cases} 1 & X_1 = 0 \\ 0 & \text{otherwise,} \end{cases}$$

as $\text{E}\,\hat{\theta}(\mathbf{X}) = 1 \cdot \text{P}(X_1 = 0) = e^{-\lambda} = \theta$. But this is likely to be a poor estimator, since it ignores the rest of the sample $X_2, X_3, \dots, X_n$.

We can see that $S(\mathbf{X}) = \sum_{i=1}^n X_i$ is sufficient since the joint pdf can be expressed as

$$f(\mathbf{x}|\lambda) = \frac{1}{x_1! \cdots x_n!} \cdot e^{-n\lambda} \lambda^{\sum_{i=1}^n x_i}.$$

Now apply the Rao-Blackwell theorem:

$$\begin{aligned} \phi(S) = \text{E}(\hat{\theta}|S) = \text{E}\left(\hat{\theta} \,\middle|\, \sum_{i=1}^n X_i = S\right) &= \text{P}\left(X_1 = 0 \,\middle|\, \sum_{i=1}^n X_i = S\right) \\ &= \left(1 - \frac{1}{n}\right)^S, \end{aligned}$$

where the conditional probability in the last step comes from the Poisson-binomial relationship (Refer Ex. Sheet 2: Suppose $X_i \overset{\text{iid}}{\sim} \text{Poi}(\lambda_i)$, then $X_1 | (\sum_{i=1}^n X_i = m) \sim \text{Bin}(m, \pi)$, where $\pi = \lambda_1 / \sum_{i=1}^n \lambda_i$).

By the Rao-Blackwell theorem, $\text{Var}(\phi) < \text{Var}(\hat{\theta}(\mathbf{X}))$ (strict inequality since $\hat{\theta}(\mathbf{X})$ is not a function of S), so prefer $\phi(S)$ over $\hat{\theta}(\mathbf{X})$ as an estimator.

```
lambda <- 2
n <- 25
(theta <- dpois(x = 0, lambda = lambda))
```

```
[1] 0.1353353
```

```
B <- 1000
X <- matrix(rpois(n * B, lambda = lambda), ncol = B)
theta_hat <- apply(X, 2, function(x) as.numeric(x[1] == 0))
phi_hat <- apply(X, 2, function(x) (1-1/n)^(sum(x)))

tibble(theta_hat, phi_hat) %>%
  pivot_longer(everything(), names_to = "Estimator", values_to = "theta_hat") %>%
  ggplot() +
  geom_density(aes(theta_hat, col = Estimator, fill = Estimator), alpha = 0.6) +
```

```
# # geom_vline(xintercept = theta, linetype = "dashed") +
# geom_vline(data = tibble(
#   x = c(mean(MLE), mean(MOM)),
#   Estimator = c("MLE", "MOM")
# ), aes(xintercept = x), linetype = "dashed") +
facet_grid(. ~ Estimator) +
# labs(x = expression(hat(theta)), y = expression(f~(hat(theta)))) +
# scale_x_continuous(breaks = seq(2, 14, by = 2)) +
theme(legend.position = "none")
# geom_text(data = tibble(x = c(4.9, 5.85), y = c(0.45, 0.45),
#                         Estimator = c("MLE", "MOM"),
#                         label = c("E(hat(theta)[ML])", "E(hat(theta)[MOM])")),
#           aes(x, y, label = label), parse = TRUE)
```

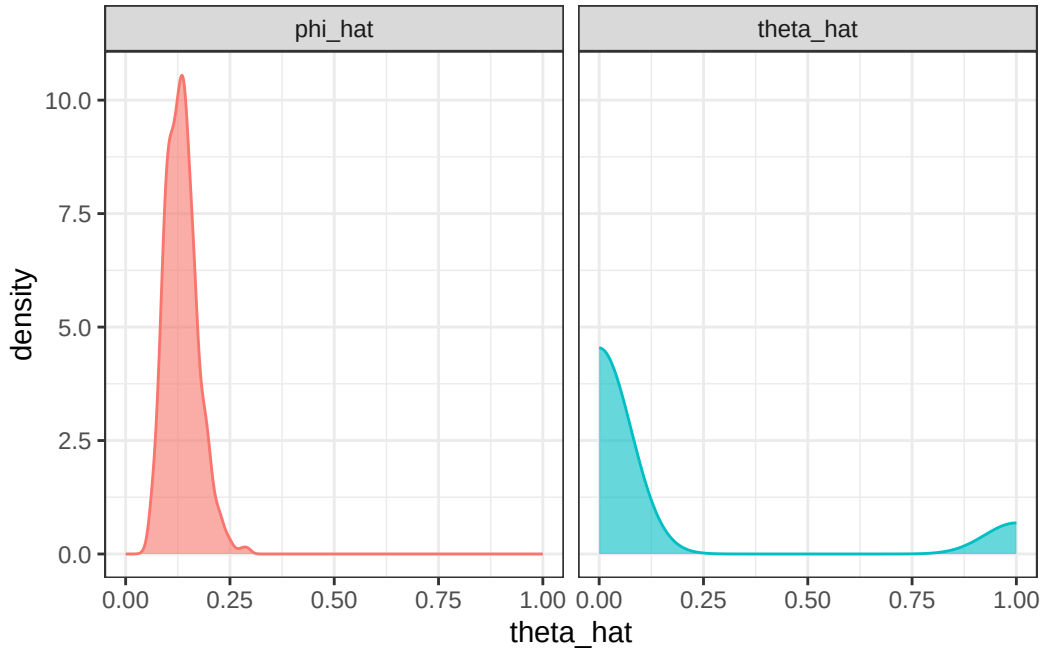


Figure 8.2

But is $\phi(S) = (1 - 1/n)^S$ unbiased? This is guaranteed by the RB theorem. Check: Since $S = \sum_{i=1}^{n} X_i \sim \text{Poi}(n\lambda)$, we get

$$\begin{aligned}\text{E}(\phi(S)) &= \sum_{s=0}^{\infty} \left(1 - \frac{1}{n}\right)^s \frac{e^{-n\lambda}(n\lambda)^s}{s!} \times e^{-\lambda} e^{\lambda} \\ &= e^{-\lambda} \sum_{s=0}^{\infty} \underbrace{\frac{e^{-\lambda(n-1)}[\lambda(n-1)]^s}{s!}}_{\text{pmf of Poi}(\lambda(n-1))} = e^{-\lambda}.\end{aligned}$$

A similar calculation can give us the variance of this estimator.

8.3 Continuity

A continuous function $\psi(x)$ is a function such that a continuous variation of x induces a continuous variation of $\psi(x)$–i.e. no jumps allowed. ψ is continuous at c if $\forall \epsilon > 0$, $\exists \delta > 0$ s.t. $|x - c| < \delta \Rightarrow |\psi(x) - \psi(c)| < \epsilon$.

```
tibble(
  x = seq(1, 3, length = 1000),
  y = 16 - x^2
) -> plot_df

mycols <- grDevices::palette.colors(3, palette = "Set1")

ggplot() +
  annotate("rect", xmin = 2 - 0.25, xmax = 2 + 0.25, ymin = -Inf, ymax = Inf,
           fill = mycols[1], alpha = 0.3) +
  annotate("rect", xmin = -Inf, xmax = Inf, ymin = 12 - 2, ymax = 12 + 2,
           fill = mycols[2], alpha = 0.3)  +
  geom_line(data = plot_df, aes(x, y)) +
  geom_line(data = plot_df %>% filter(x >= 2 - 0.25, x <= 2 + 0.25), aes(x, y),
            col = mycols[3], linewidth = 2) +
  # geom_segment(aes(x = 2, xend = 2, y = 12, yend = -Inf), linetype = "dashed",
  #              size = 0.4)
  geom_hline(yintercept = 12, linetype = "dashed") +
  geom_vline(xintercept = 2, linetype = "dashed") +
  scale_x_continuous(breaks = 2 + c(-0.25, 0, 0.25),
                     labels = c(expression("c-"*delta),
                                "c",
                                expression("c+"*delta))) +
  scale_y_continuous(breaks = 12 + c(-2, 0, 2),
                     labels = c(expression("f(c)-"*epsilon),
                                "f(c)",
                                expression("f(c)+"*epsilon)))
```

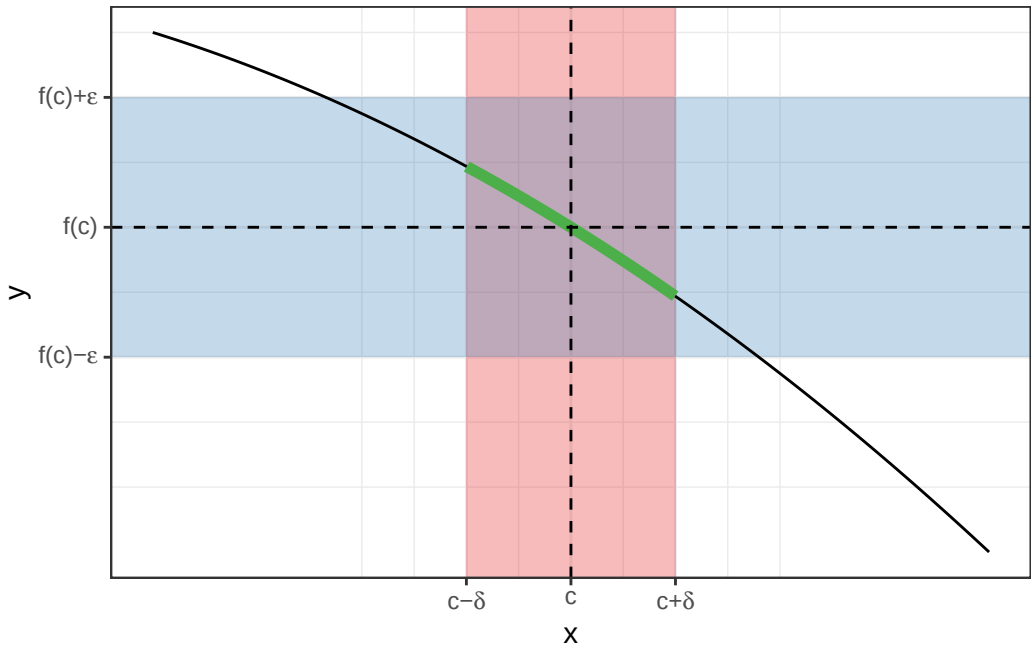


Figure 8.3

9 Hypothesis testing

9.1 A fuzzy and cute example

On a farm there are 499 white bunnies, and 1 brown bunny. One of the bunnies ravaged through the carrot farm, leaving the farmer furious.

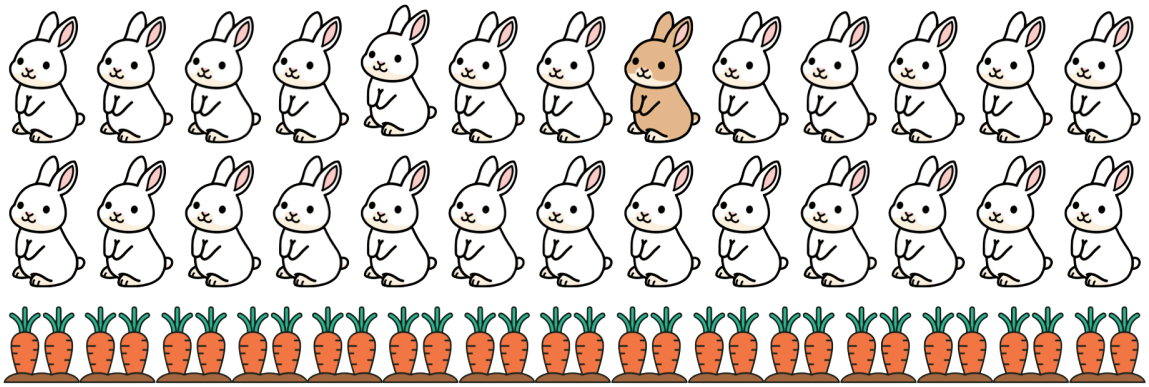


Figure 9.1

> **Question**
>
> Can we say that we know, or reasonably believe with confidence, that it was a white bunny that caused the problem? What's your proof?[1]

Assume colour difference is not associated with behavioural differences in rabbits. If we believe that a white rabbit indeed was at fault, the error rate is $1/500 = 0.2\%$.

Suppose there was a witness that claimed the brown rabbit did it. The witness performed a colour identification test, reporting the right colour 95% of the time. Given the evidence, the probability that a brown rabbit was at fault is

$$\begin{aligned} \mathrm{P}(B \mid E) &= \frac{\mathrm{P}(E \mid B)\,\mathrm{P}(B)}{\mathrm{P}(E \mid B)\,\mathrm{P}(B) + \mathrm{P}(E \mid B^c)\,\mathrm{P}(B^c)} \\ &= \frac{0.95 \times 0.002}{0.95 \times 0.002 + 0.05 \times 0.998} \\ &\approx 3.6\% \end{aligned}$$

[1] Example adapted from Schoeman, F. (1987). Statistical vs. direct evidence. Noûs, 179-198.

giving an error rate of 96.4%!

9.2 Fisherian view

The p-value is interpreted as a *continuous measure of evidence* against some null hypothesis– there is no point at which the results become 'significant'.

Remark

Statistical evidence differs from direct evidence (e.g. having CCTV recording in the house). We may **never know** what exactly happened. The best we can do is to base decisions based on the *likelihood of the evidence* materialising.

Figure 9.2: Credits: https://xkcd.com/892/.

9.3 Uniformity of p-values

> **Question**
>
> Since $p(\mathbf{X})$ is a statistic, it is a rv. What is its distribution?

Theorem 9.3.1 (Uniformity of p-values)**.** *If θ_0 is a point null hypothesis for the parameter of continuous $\mathbf{X}$, then a correctly calculated p-value $p_T(\mathbf{X})$ based on any test statistic W, is such that*

$$p_T(\mathbf{X}) \sim \text{Unif}(0,1)$$

in repeated sampling under H_0.

Proof. This is a consequence of the *probability integral transform*: Suppose that a continuous rv T has cdf $F_T(t), \forall t$. Then the rv $Y = F_T(T) \sim \text{Unif}(0,1)$ because:

$$F_Y(y) = \text{P}(\overbrace{F_T(T)}^{Y} \leq y) = \text{P}\left(T \leq F_T^{-1}(y)\right) = F_T\left(F_T^{-1}(y)\right) = y,$$

which is the cdf of a Unif$(0,1)$ distribution.

Now for any data $\mathbf{x}$,

$$p_T(\mathbf{x}) = \text{P}_{\theta_0}\left(T(\mathbf{X}) \geq T(\mathbf{x})\right) = 1 - F(T(\mathbf{x})),$$

where F is the cdf (under H_0) of $T(\mathbf{X})$. Hence, $p_T(\mathbf{x}) = 1 - Y$ where $Y \sim \text{Unif}(0,1)$ by the probability integral transform. But clearly if $Y \sim \text{Unif}(0,1)$, then so is $1 - Y$. □

This result is useful especially for *checking the validity* of a complicated p-value calculation:

1. Simulate several new data sets from the null distribution.
2. For each simulated data set, apply the p-value calculation.
3. Assess the collection of resulting p-values–do they seem to be uniformly distributed?

Suppose we are testing $H_0 : \mu = 0$ on a random sample of $X_1, \ldots, X_n$ assumed to be normally distributed with mean μ and variance $\sigma^2 = 4.3^2$. Let's do this experiment:

1. Draw $X_1, \ldots, X_{10} \overset{\text{iid}}{\sim} \text{N}(0, 4.3^2)$ (the distribution under H_0)
2. Compute the p-value $p(\mathbf{x})$ based on the simulated data
3. Repeat 1–2 a total of $B = 100000$ times to get $p_1, \ldots, p_B \in (0,1)$

Plotting a histogram of the simulated p-values yields:

```r
n <- 10
sigma <- 4.3
B <- 100000
res <- rep(NA, B)
for (i in 1:B) {
  x <- rnorm(n, sd = sigma)
  res[i] <- 2 * (pnorm((sqrt(n) * abs(mean(x)) / sigma), lower.tail = FALSE))
}
ggplot() +
  geom_histogram(aes(res, ..density..), breaks = seq(0, 1, by = 0.05),
                 col = "white") +
  geom_hline(yintercept = 1, linetype = "dashed") +
  scale_y_continuous(breaks = seq(0, 1, by = 0.25)) +
  scale_x_continuous(breaks = seq(0, 1, by = 0.1)) +
  labs(x = "p", y = "Density")
```

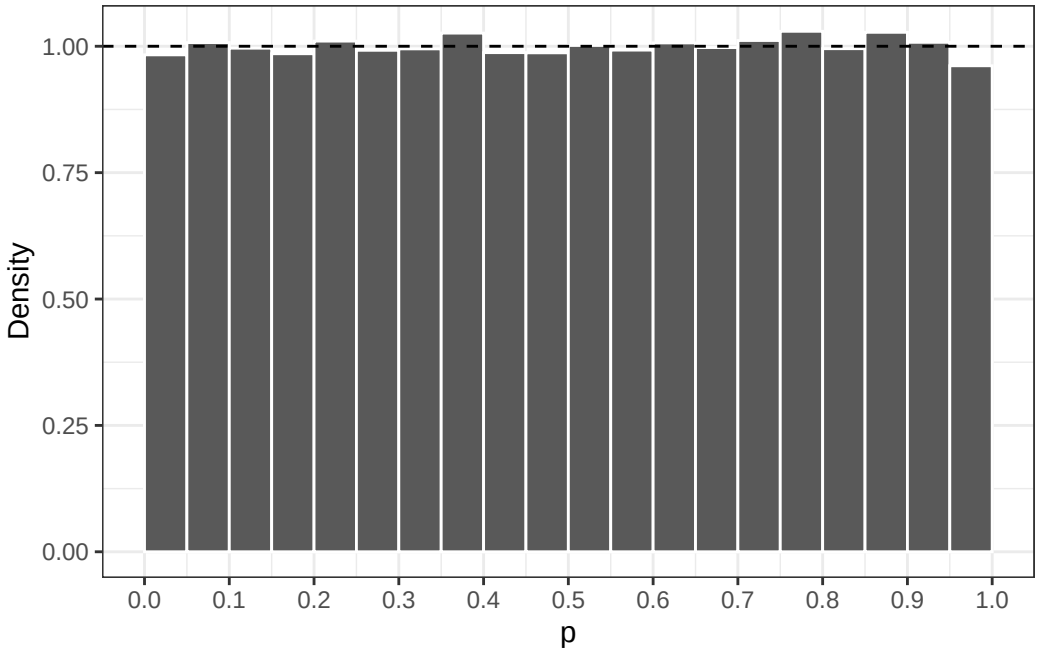


Figure 9.3

Why is this? Assume that H_0 is true. In the Neyman-Pearson approach, α is the rate of false positives, i.e. the rate at which the null hypothesis is rejected given that H_0 is true. This rate is fixed. On the other hand, $p = p(\mathbf{X})$ is a random variable.

For any value α, the null is rejected when the observed $p < \alpha$. This happens, by definition, with probability α! The only way that this happens is when the p-value comes from a uniform distribution, since $\mathrm{P}(U \leq u) = u$. I.e., under the null

- p has a 5% chance of being less than $\alpha = 0.05$;

- p has a 10% chance of being less than $\alpha = 0.1$;
- etc.

So, as a consequence, if H_0 is false, then (hopefully) the p-values are biased towards 0.

See http://varianceexplained.org/statistics/interpreting-pvalue-histogram/.

9.4 One-sided tests

All of the tests thus far are called *two-sided* tests. Sometimes we wish to measure the evidence (against H_0) in one direction only.

Example 9.4.1. Suppose $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{N}(\mu, \sigma^2)$ with σ^2 known. Consider testing $H_0 : \mu \leq 0$. The unrestricted MLE remains $\hat{\mu} = \bar{X}$, but the restricted MLE under H_0 is a bit tricky. With a little bit of reasoning,

$$\tilde{\mu} = \begin{cases} \bar{X} & \bar{X} \leq 0 \\ 0 & \bar{X} > 0 \end{cases}$$

```
set.seed(123)
n <- 10
sigma <- 4.3
tibble(
  mu = seq(-6, -2, length = 100),
  ll = dnorm(-4, mean = mu, sd = sigma, log = TRUE)
) %>%
  mutate() %>%
  ggplot() +
  annotate("rect", xmin = -Inf,  xmax = 0, ymin = -Inf, ymax = Inf,
           alpha = 0.2) +
  geom_line(aes(mu, ll, col = "a")) +
  geom_line(aes(mu + 5, ll, col = "b")) +
  geom_vline(xintercept = 0, linetype = "dashed") +

  scale_y_continuous(breaks = NULL, name = expression(log~L(mu))) +
  scale_x_continuous(breaks = 0, name = expression(mu)) +
  guides(col = "none")
```

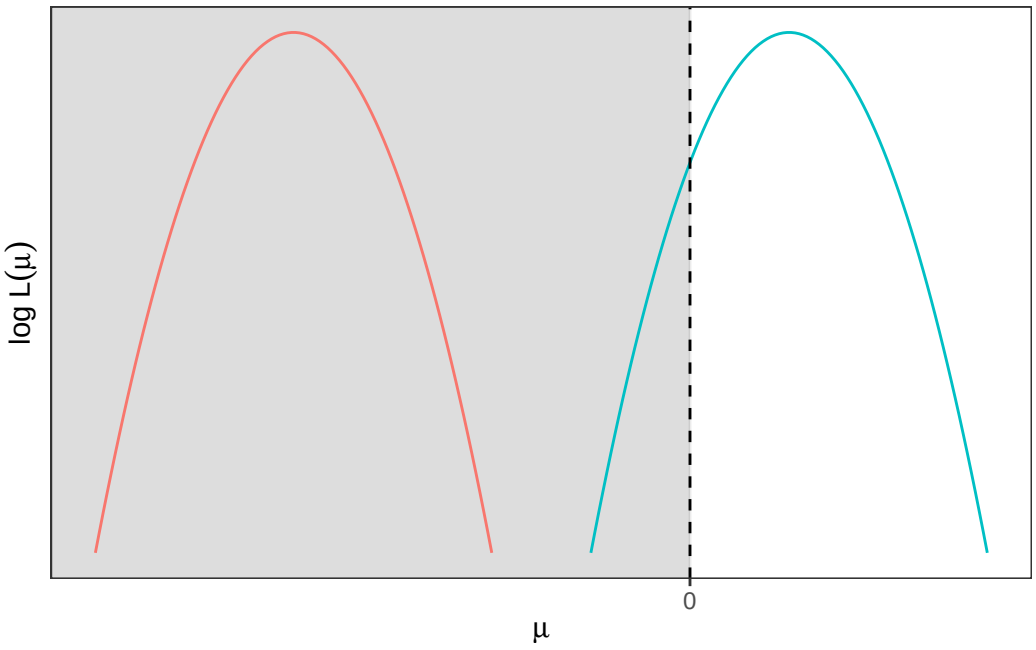


Figure 9.4

Therefore, the log LR statistic depends on the value of $\bar{X}$:

$$\log W_{LR} = \ell(\hat{\mu}|\mathbf{X}) - \ell(\tilde{\mu}|\mathbf{X}) = \begin{cases} 0 & \bar{X} \leq 0 \\ \frac{n\bar{X}^2}{2\sigma^2} & \bar{X} > 0 \end{cases}$$

(the second case when $\bar{X} > 0$ is as before). The p-value from data $\mathbf{x}$, using the monotonicity of $\bar{X}$ in the LRT statistic, is

$$p(\mathbf{x}) = \begin{cases} 1 & \bar{x} \leq 0 \\ \mathrm{P}(\bar{X} > \bar{x}) = 1 - \Phi(\sqrt{n}\bar{x}/\sigma) & \bar{x} > 0 \end{cases}$$

Hence, relative to the 'two-sided' test that we saw previously, the p-value is *halved* if $\bar{x} > 0$, and ignores the precise value of $\bar{x}$ if $\bar{x} \leq 0$.

Further remarks:

1. Performing a one-sided test instead of a two-sided test thus makes any apparent evidence against H_0 seem stronger (since the p-value is halved).
2. In practice there are rather few situations where performing a one-sided test, which assumes that we know in advance that departures from H_0 are in one direction only, can be justified. When assessing the effect of a new drug, for example, the convention is to assess evidence for an effect in either direction, positive or negative.
3. The two-sided test is said to be more *conservative* than the one-sided test: The one-sided test risks over-stating the strength of evidence against H_0 if the underlying assumption–that evidence against H_0 counts in one direction only–is actually false.

9.5 "Failing to reject the null hypothesis"

Absence of proof is not proof of absence. You are not able prove a negative.

1. Australian Tree Lobsters were assumed to be extinct. There was no evidence that any were still living because no one had seen them for decades. Yet in 1960, scientists observed them.
2. In criminal trial, we start with the assumption that the defendant is innocent until proven guilty. If the prosecutor fails to meet a an evidentiary standard, it does not mean the defendant is innocent.

> ⚠ Accepting the null hypothesis
>
> Accepting the null hypothesis indicates that you have proven that an effect does not exist. Maybe, this is what you mean?[2]

9.6 Asymptotic distribution of LRT: An experiment

Let's try to "verify" the distribution of the test statistic $2 \log \lambda(\mathbf{X})$.

1. Draw $X_1, \ldots, X_{10} \sim \mathrm{N}(8, 1)$
2. Compute $T(\mathbf{X}) = 2 \log \lambda(\mathbf{X}) = \sum_{i=1}^{n} (X_i - \bar{X})^2$
3. Repeat steps 1–2 $B = 10000$ number of times to get $T_1, \ldots, T_B$

We can plot the histogram of the observed test statistic, and overlay a χ^2_9 density over it. As can be seen, it is a good fit.

```
B <- 10000
res <- rep(NA, B)
for (i in 1:B) {
  X <- rnorm(10, mean = 8)
  res[i] <- sum((X - mean(X)) ^ 2)
}
ggplot() +
  geom_histogram(aes(x = res, y = ..density..), col = "white") +
  geom_line(data = tibble(x = seq(0, 35, length = 100),
                          y = dchisq(x, 10 - 1)),
            aes(x, y), col = "red3", size = 1) +
  scale_y_continuous(breaks = NULL) +
  labs(x = expression(2~log~lambda(X)), y = "Density")
```

[2] https://statisticsbyjim.com/hypothesis-testing/failing-reject-null-hypothesis/

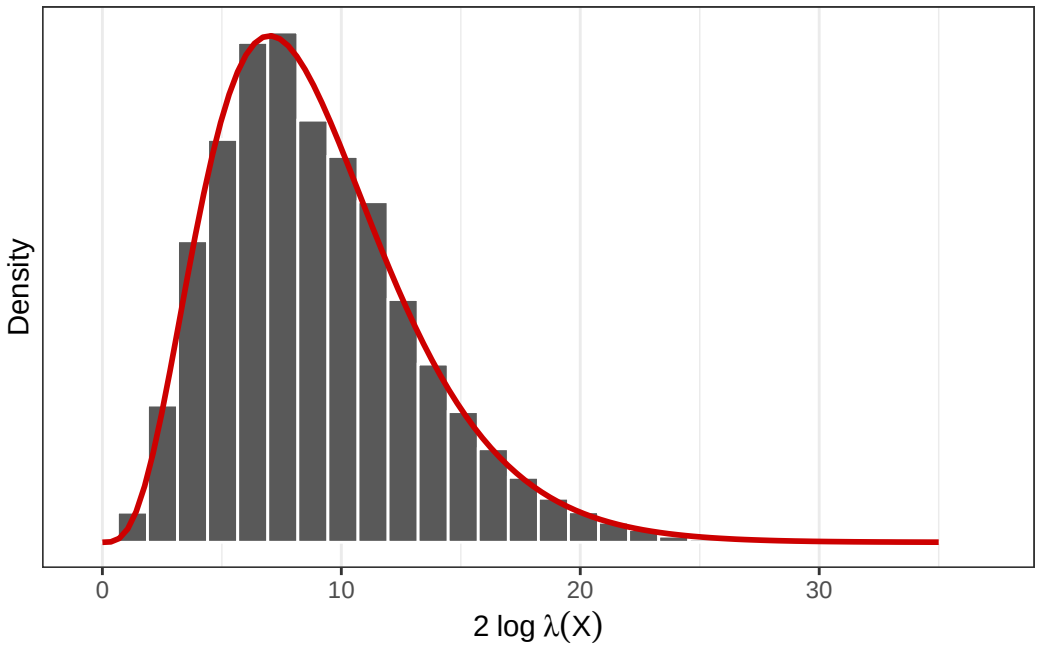


Figure 9.5

Actually, in this particular case, the distribution of $2\log\lambda(\mathbf{X})$ is **exact**. Note that

$$2\log\lambda(\mathbf{X}) = \frac{n-1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = (n-1)S^2$$

which is the sample variance. We've seen previously that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Thus, $2\log\lambda(\mathbf{X})$ is merely a *scaled* χ^2 distribution (but in this case $\sigma^2 = 1$).

10 Interval estimation

Add your notes here.

11 Using `nlminb()` for Maximum Likelihood Estimation

Consider Q2 in the Exercise Sheet 5 (Hypothesis Testing). A random sample $X_1, \ldots, X_n$ is drawn from a Pareto distribution with pdf

$$f(x \mid \alpha, \nu) = \frac{\alpha \nu^{\alpha}}{x^{\alpha+1}} \quad \text{for } x > \nu,\ \alpha > 0,\ \nu > 0$$

The Pareto distribution is frequently used in economics to model income and wealth distributions, especially the upper tail–where a small fraction of the population holds a disproportionately large share of income or wealth. This fits the famous Pareto Principle or 80/20 rule.

The two parameters in the Pareto distribution:

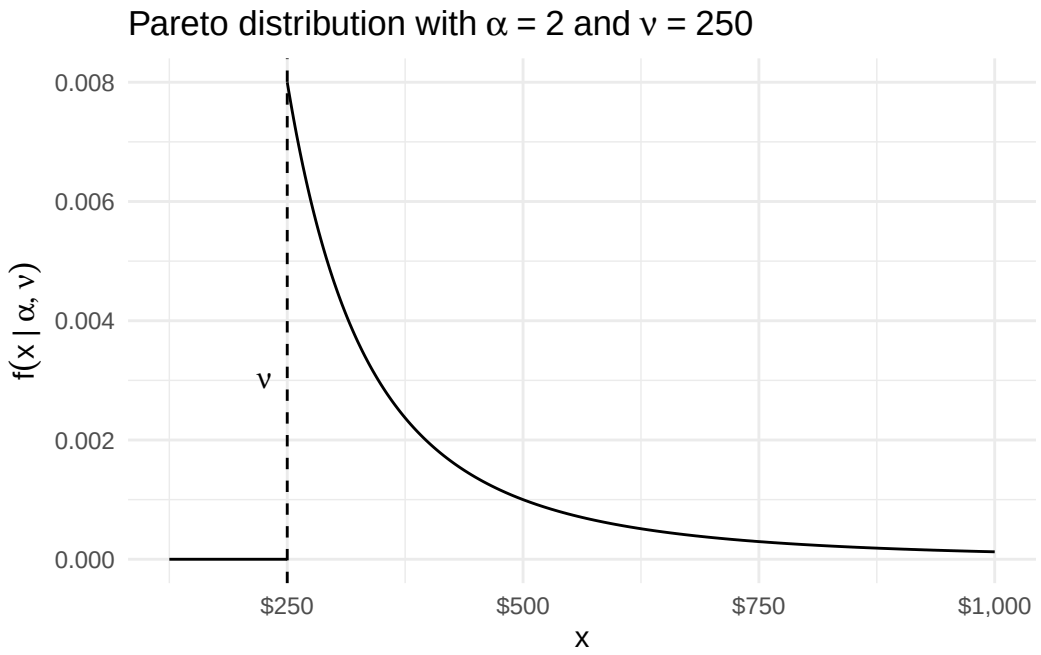
- ν (scale parameter): the minimum possible income/wealth (i.e., the distribution starts at this value),
- α (shape parameter): controls the "fatness" of the tail, smaller α means fatter tails and greater inequality. In wealth modelling, this is the so-called *Pareto index*.

An example from empirical literature (e.g. Atkinson and Piketty, 2007) suggests that α for income in the U.S. top 1% is around 1.5-2.5, depending on the year and method, while ν varies depending on the income bracket analyzed, typically \$100k to \$500k for high earners.

Here is what the pdf looks like:

```
# True values
alpha <- 2  # shape parameter
nu <- 250  # thousands of dollars, say

tibble(
  x = seq(nu, nu*4, length.out = 1000),
  f = case_when(
    x < alpha ~ 0,
    TRUE ~ alpha * nu^alpha / x^(alpha + 1)
)) |>
  ggplot(aes(x, f)) +
  geom_line() +
  geom_vline(xintercept = nu, linetype = "dashed") +
```

```
  annotate("segment", x = nu, xend = nu / 2, y = 0, yend = 0) +
  annotate("text", x = nu, y = 3e-3, label = expression(nu), hjust = 2) +
  labs(
    title = expression("Pareto distribution with"~alpha~"="~2~"and"~nu~"="~250),
    x = "x",
    y = expression(f(x~"|"~alpha, nu))
  ) +
  scale_x_continuous(labels = scales::dollar) +
  theme_minimal()
```



The following code generates a random sample of size n from the Pareto distribution with parameters $\alpha = 2$ and $\nu = 250$:

```
library(VGAM)
set.seed(1)

n <- 250  # sample size
X <- rpareto(n, scale = nu, shape = alpha)
head(X, 10)
```

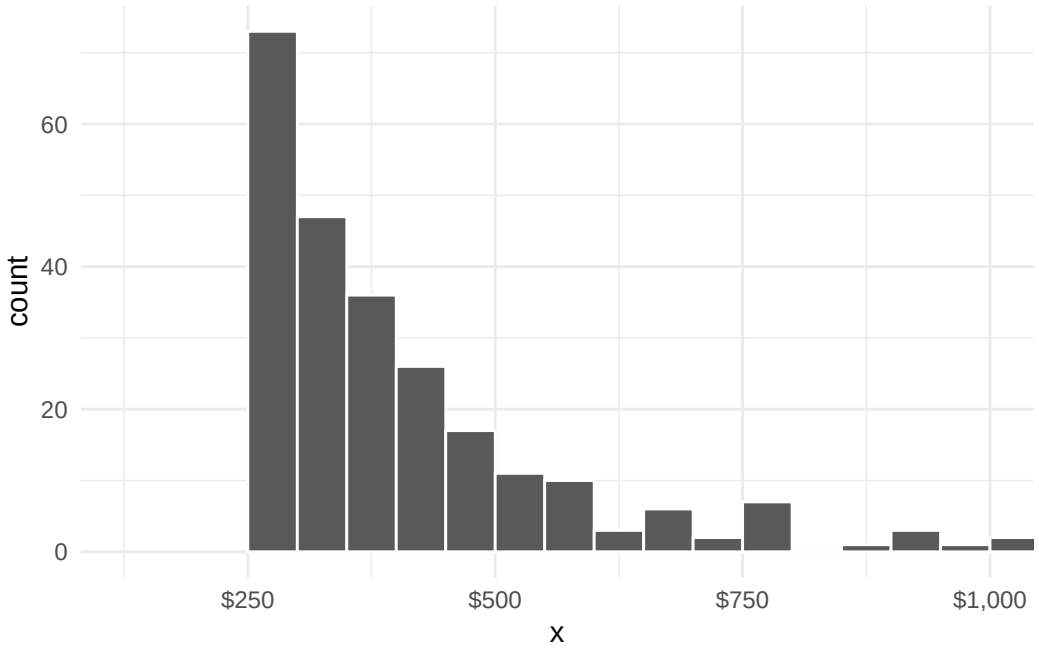
```
 [1]  485.1775  409.8229  330.3074  262.3297  556.6811  263.7592  257.2164
 [8]  307.5429  315.1921 1005.7592
```

```
summary(X)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  250.9   292.4   354.7   428.1   469.4  2186.1
```

And suppose we were to plot a histogram of the sample:

```
ggplot(data.frame(x = X), aes(x)) +
  geom_histogram(col = "white", binwidth = 50, boundary = nu) +
  scale_x_continuous(labels = scales::dollar) +
  coord_cartesian(xlim = c(nu / 2, 1000)) +
  theme_minimal()
```



Think

When we "draw" samples from a particular pdf, we expect the distribution of the sample (i.e., the histogram) to resemble the theoretical pdf. Do you see any resemblance? Looking ahead to parameter estimation, suppose the true value of ν was not known. What value would you guess for ν based on the data?

11.1 Paremeter estimation using MLE

In class, we solved for the MLE of α and ν in the usual way using derivatives and sketching the likelihood function. Recall that

$$\hat{\alpha} = \frac{n}{\sum_{i=1}^{n} \log(X_i/\hat{\nu})} \quad \text{and} \quad \hat{\nu} = \min(X_i).$$

If we plug in the data to compute the MLE, we get:

```
nu_hat <- min(X)
alpha_hat <- n / sum(log(X / nu_hat))
cat("nu_hat =", nu_hat, "\nalpha_hat =", alpha_hat)
```

```
nu_hat = 250.9195
alpha_hat = 2.267916
```

We can also let the computer do the work for us using the `nlminb()` function in R. This function is a general-purpose optimization function that can be used to find the maximum likelihood estimates of parameters in a statistical model. What we need is to first code the likelihood function, and then use `nlminb()` to find the values of α and ν that maximize the likelihood function.

```
# The pdf function
fx <- function(x, alpha, nu) {
  alpha * nu^alpha / x^(alpha + 1)
}
fx(X[1:10], alpha, nu)
```

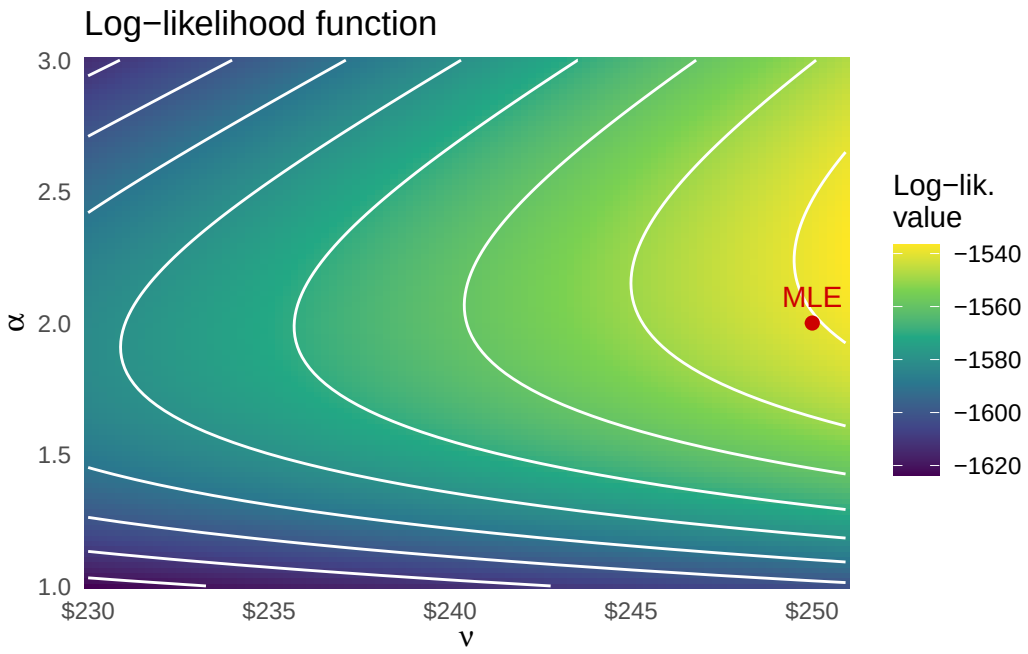
```
 [1] 0.0010944805 0.0018160229 0.0034686077 0.0069241716 0.0007245869
 [6] 0.0068121958 0.0073453722 0.0042972722 0.0039919405 0.0001228649
```

```
# The log-likelihood function
ll <- function(theta) {
  alpha <- theta[1]
  nu <- theta[2]

  # Return really small value if support condition is violated
  if (alpha <= 0 | nu <= 0 | any(X < nu)) return(-1e10)

  sum(log(fx(X, alpha, nu)))
}
ll(theta = c(2, 250))
```

```
[1] -1540.532
```

Here's a plot of the 2-dimensional log-likelihood function based on the data:

```r
expand_grid(
  nu = seq(230, min(X), length.out = 100),
  alpha = seq(1, 3, length.out = 100)
) |>
  mutate(ll = purrr::map2_dbl(alpha, nu, ~ll(c(.x, .y)))) |>
  filter(ll > -1e10) |>
  ggplot(aes(nu, alpha, z = ll)) +
  geom_raster(aes(fill = ll)) +
  geom_contour(color = "white") +
  scale_fill_viridis_c() +
  annotate("point", x = nu, y = alpha, color = "red3", size = 2) +
  annotate("text", x = nu, y = alpha + 0.1, label = "MLE", color = "red3") +
  scale_x_continuous(labels = scales::dollar, expand = c(0, 0)) +
  scale_y_continuous(expand = c(0, 0)) +
  labs(
    title = "Log-likelihood function",
    x = expression(nu),
    y = expression(alpha),
    fill = "Log-lik.\nvalue"
  ) +
  theme_minimal()
```

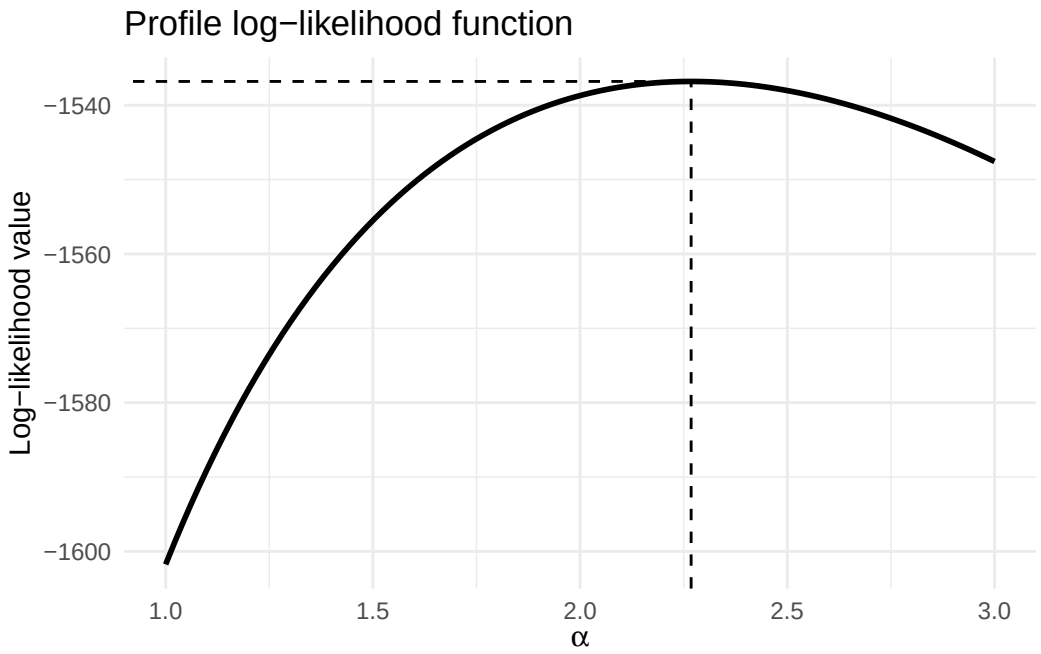


The *profile log-likelihood function*

$$f(\alpha) = \max_{\nu} \ell(\alpha, \nu) = \ell(\alpha \mid \hat{\nu})$$

can be sketched as follows:

```r
tibble(
  alpha = seq(1, 3, length.out = 100),
  ll = map_dbl(alpha, ~ll(c(.x, nu_hat)))
) |>
  ggplot(aes(alpha, ll)) +
  geom_line(linewidth = 1) +
  geom_segment(
    data = tibble(
      x = c(alpha_hat, alpha_hat),
      y = c(-Inf, ll(c(alpha_hat, nu_hat))),
      xend = c(alpha_hat, -Inf),
      yend = rep(ll(c(alpha_hat, nu_hat)), 2)
    ),
    aes(x = x, y = y, xend = xend, yend = yend),
    linetype = "dashed",
  ) +

  theme_minimal() +
  labs(
    title = "Profile log-likelihood function",
    x = expression(alpha),
    y = "Log-likelihood value"
  )
```



Now, we use `nlminb()` to find the MLE of α and ν.

```r
res <- nlminb(
  start = c(alpha, nu),  # initial "guess"
  objective = function(theta) -1 * ll(theta),  # negative log-likelihood
  lower = 0,
  upper = c(Inf, min(X))
)
print(res)
```

```
$par
[1]   2.267916 250.919541

$objective
[1] 1536.801

$convergence
[1] 0

$iterations
[1] 6

$evaluations
function gradient
       9       16

$message
[1] "both X-convergence and relative convergence (5)"
```

```r
# Compare nlminb to direct calculations. They are identical!
cat("nu_hat (calculation) =", nu_hat, "vs. nu_hat (MLE) =", res$par[2],
    "\nalpha_hat (calculation) =", alpha_hat, "vs. alpha_hat (MLE) =", res$par[1], "\n")
```

```
nu_hat (calculation) = 250.9195 vs. nu_hat (MLE) = 250.9195
alpha_hat (calculation) = 2.267916 vs. alpha_hat (MLE) = 2.267916
```

We can also check that the gradients are close to zero at the MLE. But only for the α parameter, since the log-likelihood is **not differentiable** at ν!

```r
# Gradient at MLE
numDeriv::grad(
  function(theta) -1 * ll(theta),
  res$par
)
```

```
[1] 1.654437e-06 1.937072e+12
```

The Hessian (observed Fisher information matrix) can also be obtained as follows:

```
J <- -1 * numDeriv::hessian(ll, res$par)
print(J)
```

```
           [,1]          [,2]
[1,] 48.6055593 -6.093646e-01
[2,] -0.6093646  1.350050e+09
```

```
solve(J)  # To get asymptotic variance
```

```
             [,1]         [,2]
[1,] 2.057378e-02 9.286271e-12
[2,] 9.286271e-12 7.407132e-10
```

```
# Standard errors
se <- sqrt(diag(solve(J)))
print(se)
```

```
[1] 1.434356e-01 2.721605e-05
```

References