

SM-4331 Advanced Statistics

Haziq Jamil

SM-4331 Students :)

Table of contents

Preface	4
1 Probability Theory Primer	5
1.1 Algebras of sets	5
1.2 Why σ -algebra?	5
1.2.1 Rules of probability	6
1.2.2 The need for measure theory	6
1.3 An unmeasurable set	6
1.4 Conditional probability	8
1.5 Bayesian statistics	10
1.6 Probability integral transform	11
2 Commonly used probability models	12
2.1 Poisson-Binomial relationship	12
2.2 Memoryless property	13
2.3 Relationships	14
3 Sampling from the normal distribution	15
3.1 An example	15
3.2 Finite vs infinite population	15
3.3 An experiment	16
4 Large sample approximations	17
4.1 Illustration of convergence in probability	17
4.2 Almost sure convergence	18
4.3 The Strong Law of Large Numbers	18
5 Likelihood theory	20
5.1 Finding the MLE numerically	20
5.2 Variance reduction: <i>Rao-Blackwellisation</i>	22
5.3 Continuity	25
6 Hypothesis testing	27
6.1 A fuzzy and cute example	27
6.2 Fisherian view	28
6.3 Uniformity of p -values	29
6.4 One-sided tests	31
6.5 “Failing to reject the null hypothesis”	33
6.6 Asymptotic distribution of LRT: An experiment	33

7 Interval estimation	35
References	36

Preface

These are the notes for SM-4331 Advanced Statistics.

1 Probability Theory Primer

1.1 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*.

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

1.2 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, $\Omega = \{1, \dots, 6\}$, while an event could be $E = \{2, 4, 6\} \subset \Omega$ (getting an even number). The probability of the event E occurring is $P(E) = \frac{1}{2}$ —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.

1.2.1 Rules of probability

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

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

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 $P(H) = 0.7$ necessitates $P(T) = 0.3$ and not anything else, e.g. $P(T) = 0.5$.

1.2.2 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 $P : \mathcal{P}([0, 1]) \rightarrow [0, 1]$ exists such that $P([a, b]) = b - a$ 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.

1.3 An unmeasurable set

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

$$P([a, b]) = b - a$$

that satisfies all the probability rules listed in i. to iv. earlier. On the face of it, all the rules themselves are satisfied: $P(\Omega) = P([0, 1]) = 1$, $P(\emptyset) = P([a, a]) = 0$ (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 is 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.

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

Proposition 1.1. *There does not exist a definition of $P : \mathcal{P}([0, 1]) \rightarrow [0, 1]$ satisfying $P([a, b]) = b - a$ 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 set², 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

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

Then we should have

$$P(A \oplus r) = P(A).$$

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

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 $[0], [z_1], [z_2], \dots$ 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 Choice³.

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 infinite⁴. We make two observations:

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

³Given 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>

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,

$$P\left(\bigcup_r (V \oplus r)\right) = \sum_r P(V \oplus r) = \sum_r P(V) \quad (1.1)$$

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

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

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 1⁶. We see that

$$P\left(\bigcup_r (V \oplus r)\right) = 1. \quad (1.2)$$

Equating (1.1) and (1.2) together, we find a contradiction: A countably infinite sum of a constant value can only equal 0, $+\infty$ or $-\infty$, but never 1. \square

1.4 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.

⁵Let 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, $P(B) = P(A) + P(B \setminus A)$, and since probabilities are non-negative, we have that $P(B) \geq P(A)$.

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

Let

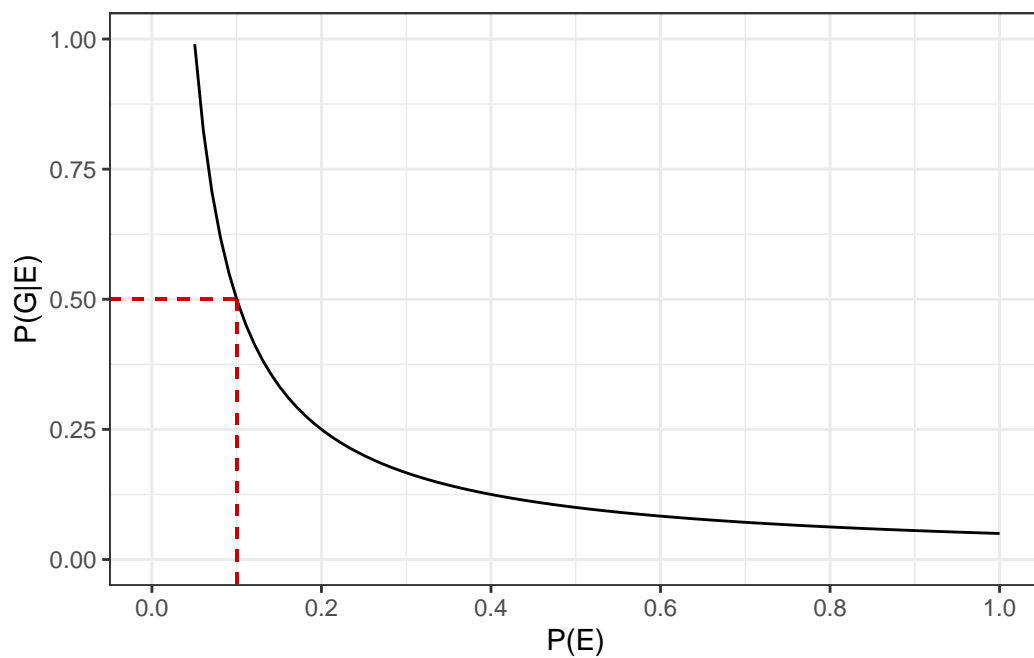
- E be the ‘evidence’
- G be the event that a student takes Geology

Then

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

```
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)")
```



1.5 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

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

```
# https://dosreislabs.github.io/2019/01/27/ballntable.html
set.seed(123)
n <- 15
xy <- runif(2) # position of coin after initial 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) )
}

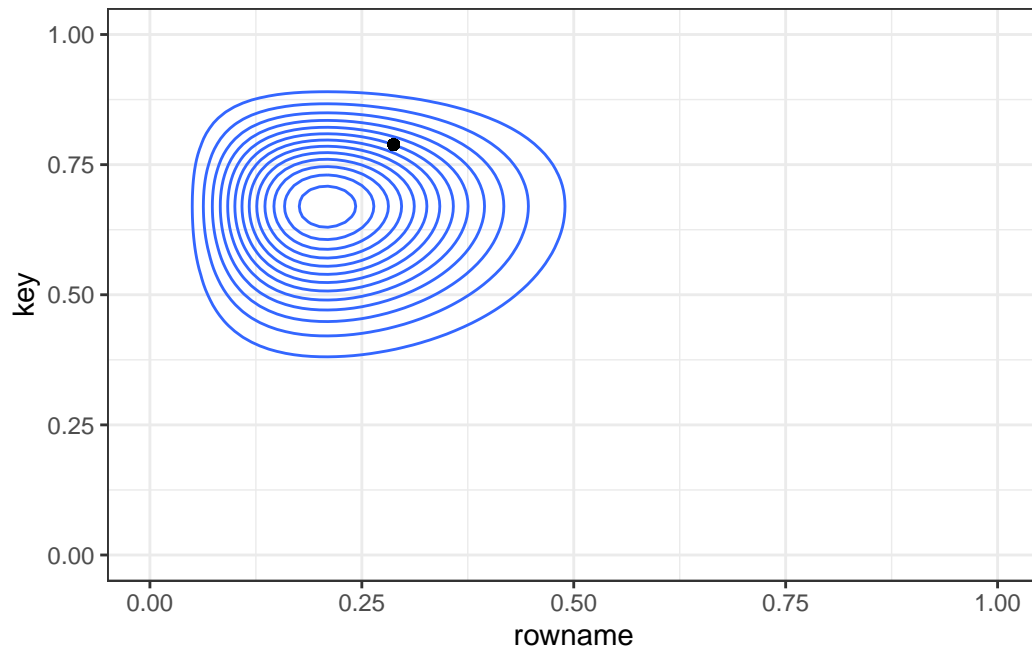
z <- jointf(pos, n) * Cf(pos[1], pos[2], n)

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))

```



1.6 Probability integral transform

Theorem 1.1. *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 $P(Y \leq y) = y$.*

Proof.

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

□

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 $U \sim \text{Unif}(0, 1)$.

2 Commonly used probability models

2.1 Poisson-Binomial relationship

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

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

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 $n \rightarrow \infty$ while $E(X) = np$ remains fixed.

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

The reason is that the Poisson can be seen as the *limiting case* to the binomial as $n \rightarrow \infty$ while $E(X) = np$ 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")

```

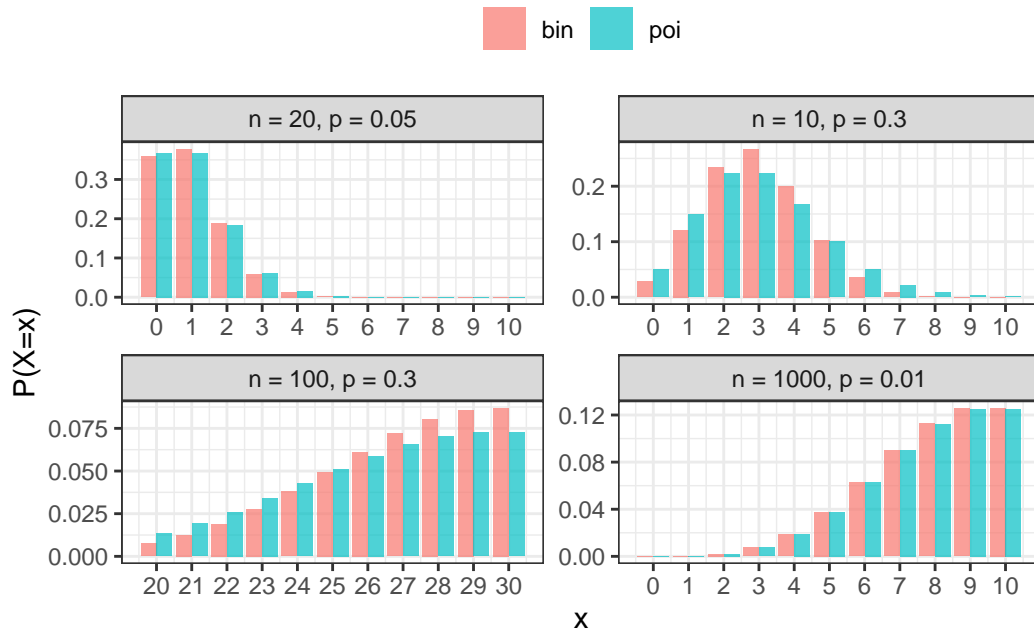


Figure 2.1

2.2 Memoryless property

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

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

3 Sampling from the normal distribution

3.1 An example

Example 3.1. 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?

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

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

3.2 Finite vs infinite 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 3.2. 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.

3.3 An experiment

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

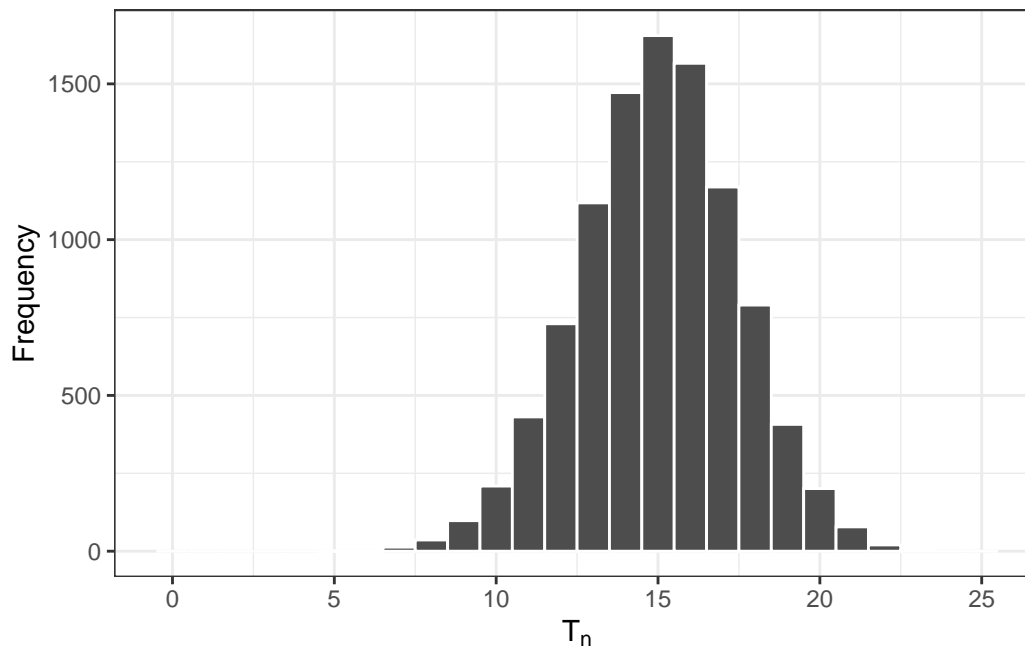
1. Draw $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Bern}(p)$
2. Compute $T_n = \sum_{i=1}^n X_i$
3. Repeat steps 1–2 a total of $B = 10000$ times

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

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

```
[1] 12 17 12 14 14 16 12 12 11 12
```

```
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")
```



4 Large sample approximations

4.1 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)
```

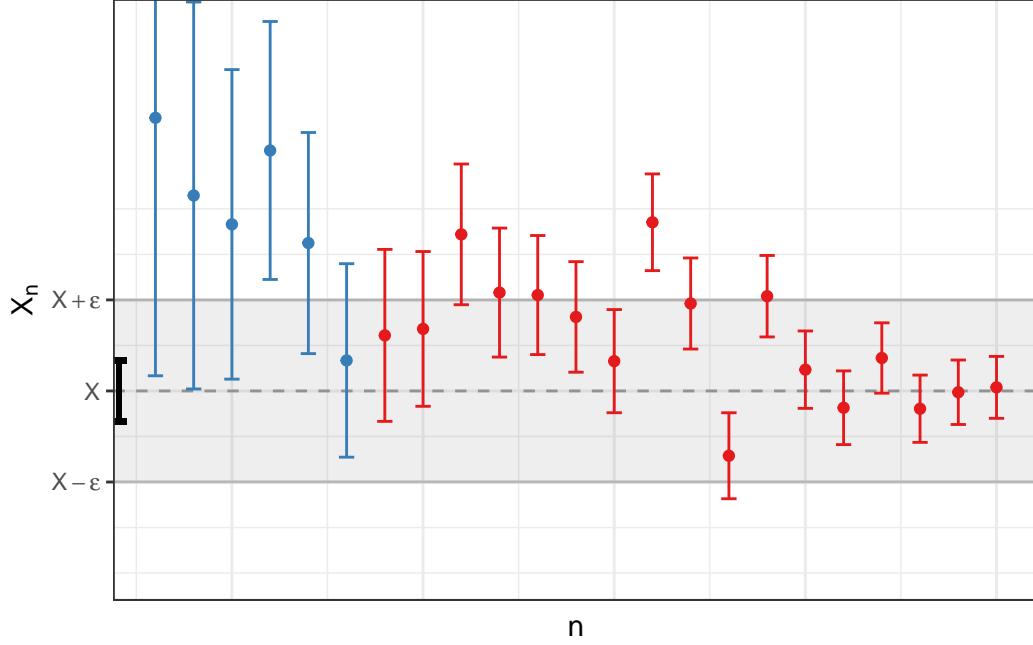


Figure 4.1

4.2 Almost sure convergence

Definition 4.1 (Almost sure convergence). X_n converges to X in *almost surely* if for every $\epsilon > 0$,

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

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

That is, $X_n(\omega) \rightarrow X(\omega)$ for all outcomes $\omega \in \Omega$, except perhaps for a collection of outcomes $\omega \in A$ with $P(A) = 0$. 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.

4.3 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 4.1 (Strong Law of Large Numbers). *Let X_1, X_2, \dots 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 $n \rightarrow \infty$, i.e.

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

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

5 Likelihood theory

5.1 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] 7.91364
```

The mean is found to be 7.91. 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)))
```

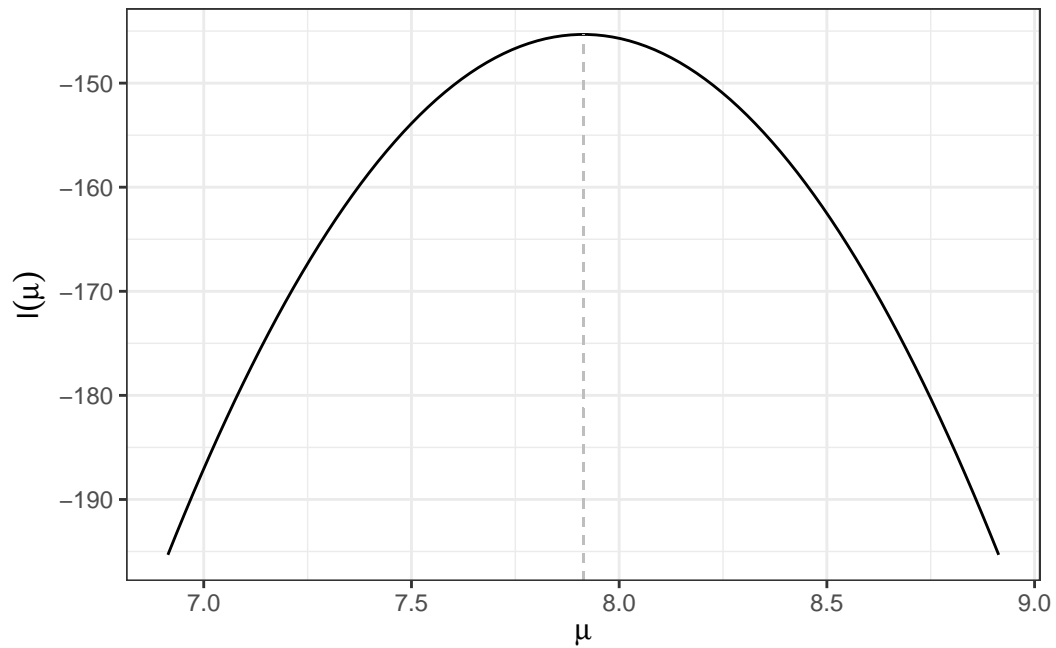


Figure 5.1: Log-likelihood function of the normal mean.

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:    2535.2413:  1.00000
1:    569.78526:  5.00000
2:    145.32034:  7.91366
3:    145.32034:  7.91364
4:    145.32034:  7.91364
```

```
glimpse(res)
```

```
List of 6
 $ par      : num 7.91
```

```

$ objective : num 145
$ 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)"

```

5.2 Variance reduction: *Rao-Blackwellisation*

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

Theorem 5.1 (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*

$$\phi(S) = E_{\theta}(\hat{\theta}|S)$$

- i. is a statistic, i.e. $\phi(S)$ does not involve θ ;*
- ii. is an unbiased statistic, i.e. $E(\phi(S)) = \theta$; and*
- iii. has $\text{Var}_{\theta}(\phi(S)) \leq \text{Var}_{\theta}(\hat{\theta})$, 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 $E_{\theta}(\hat{\theta}(\mathbf{X})|S)$ does not involve θ .
- ii. $E(\phi(S)) = E[E(\hat{\theta}|S)] = E(\hat{\theta}) = \theta$.
- iii. Using the law of total variance,

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

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

□

Example 5.1. Suppose we have data $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Poi}(\lambda)$ pertaining to the number of road accidents per day, and we want to estimate the probability of having no accidents $\theta = e^{-\lambda} = P(X_i = 0)$.

An unbiased estimator of θ is

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

as $E\hat{\theta}(\mathbf{X}) = 1 \cdot P(X_1 = 0) = e^{-\lambda} = \theta$. 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:

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

where the conditional probability in the last step comes from the Poisson-binomial relationship (Refer Ex. Sheet 2: Suppose $X_i \stackrel{\text{iid}}{\sim} \text{Poi}(\lambda_i)$, 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)

```

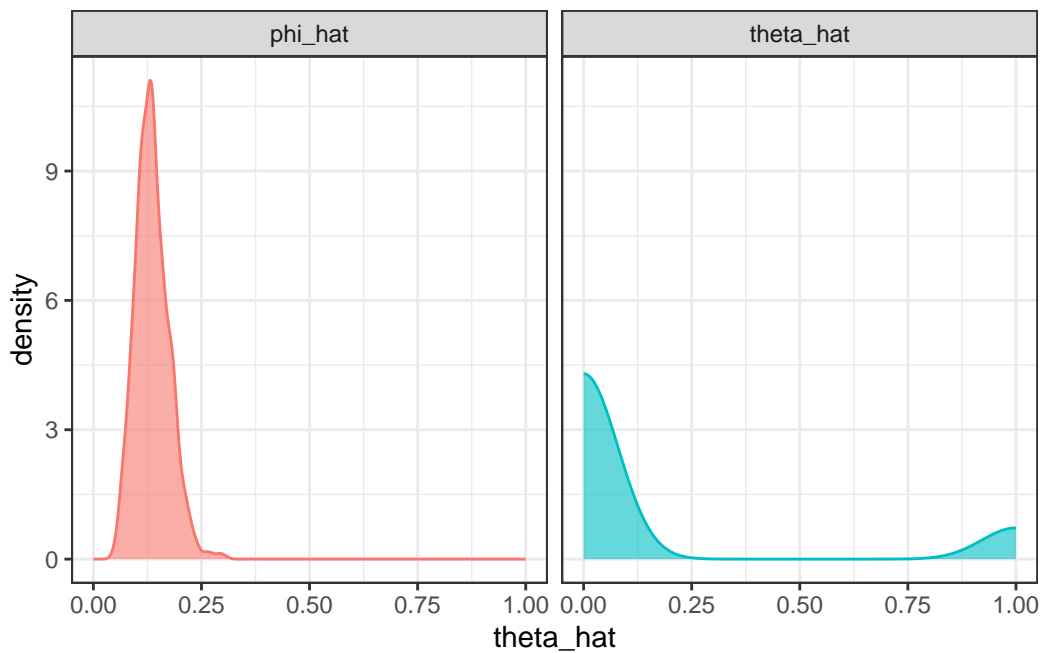


Figure 5.2

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

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

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

5.3 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")))
```

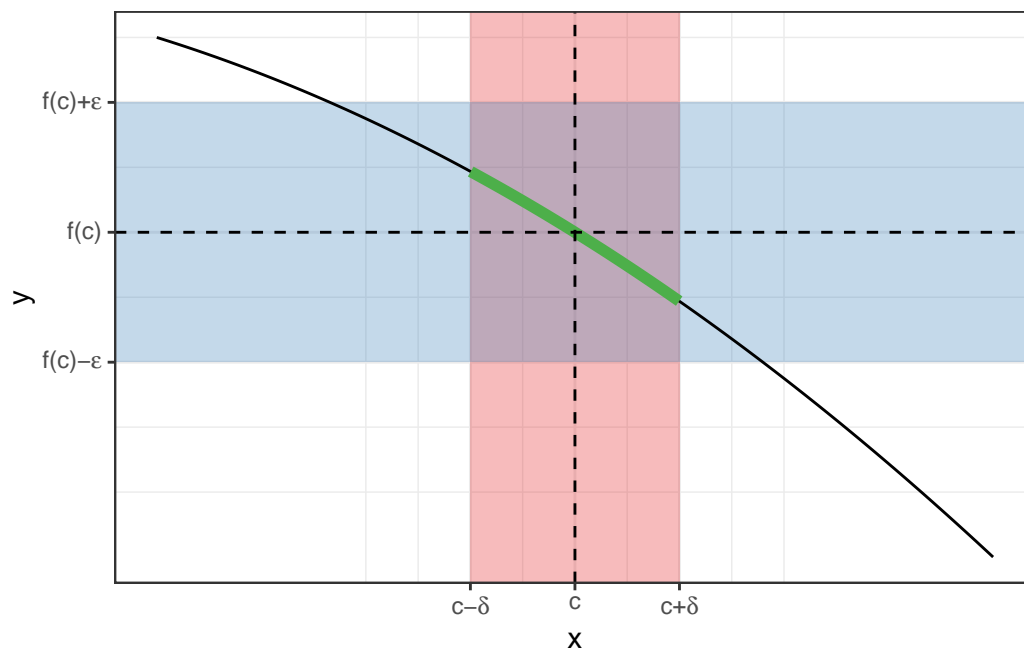


Figure 5.3

6 Hypothesis testing

6.1 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.

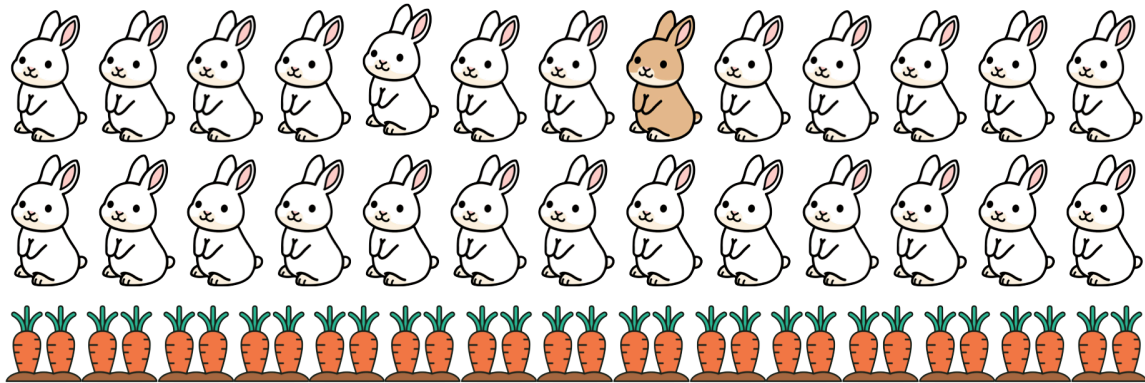


Figure 6.1

i 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

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

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

giving an error rate of 96.4%!

6.2 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 6.2: Credits: <https://xkcd.com/892/>.

6.3 Uniformity of p -values

i Question

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

Theorem 6.1 (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:

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

which is the cdf of a $\text{Unif}(0, 1)$ distribution.

Now for any data \mathbf{x} ,

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

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 $1 - Y$. \square

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, \dots, X_n assumed to be normally distributed with mean μ and variance $\sigma^2 = 4.3^2$. Let's do this experiment:

1. Draw $X_1, \dots, X_{10} \stackrel{\text{iid}}{\sim} \text{N}(0, 4.3^2)$ (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 $p_1, \dots, p_B \in (0, 1)$

Plotting a histogram of the simulated p -values yields:

```

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")

```

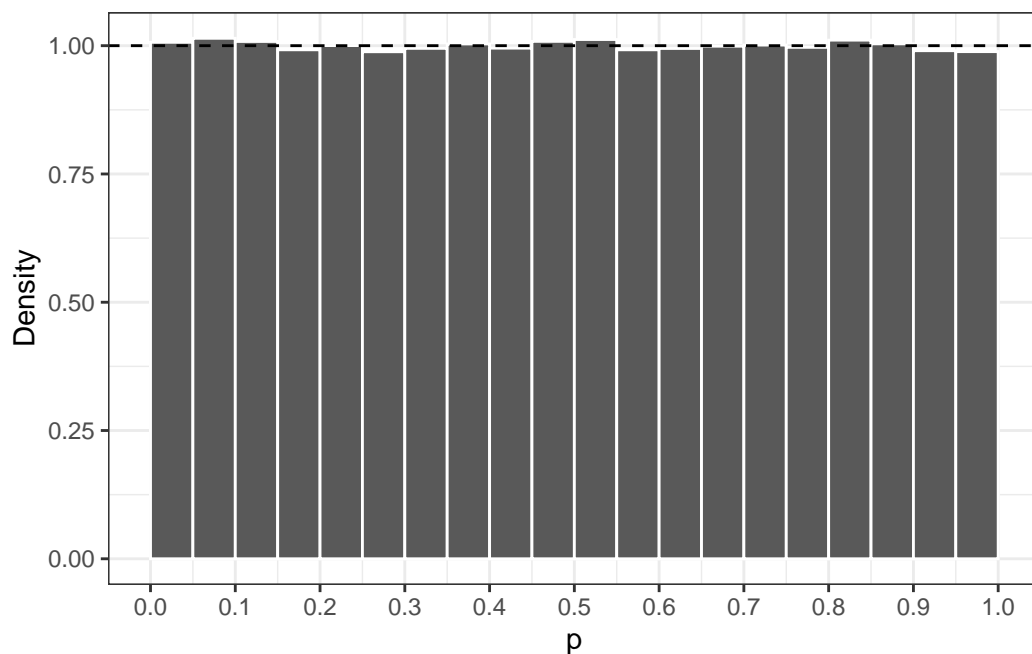


Figure 6.3

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 $P(U \leq u) = u$. 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/>.

6.4 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 6.1. Suppose $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$ 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")
```

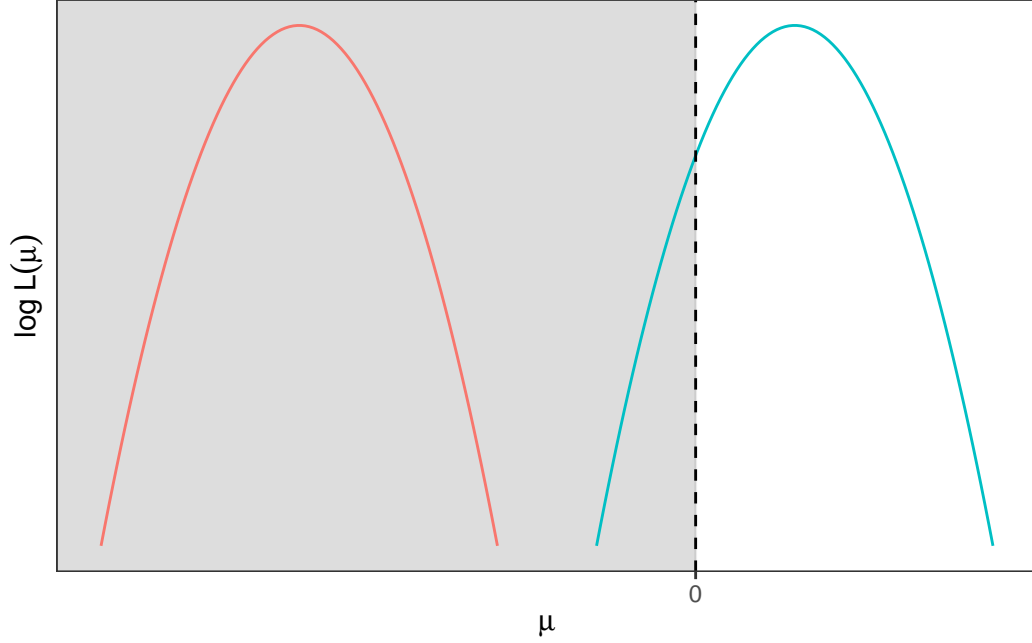


Figure 6.4

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

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

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.

6.5 “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

6.6 Asymptotic distribution of LRT: An experiment

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

1. Draw $X_1, \dots, X_{10} \sim N(8, 1)$
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, \dots, T_B

We can plot the histogram of the observed test statistic, and overlay a χ_9^2 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/>

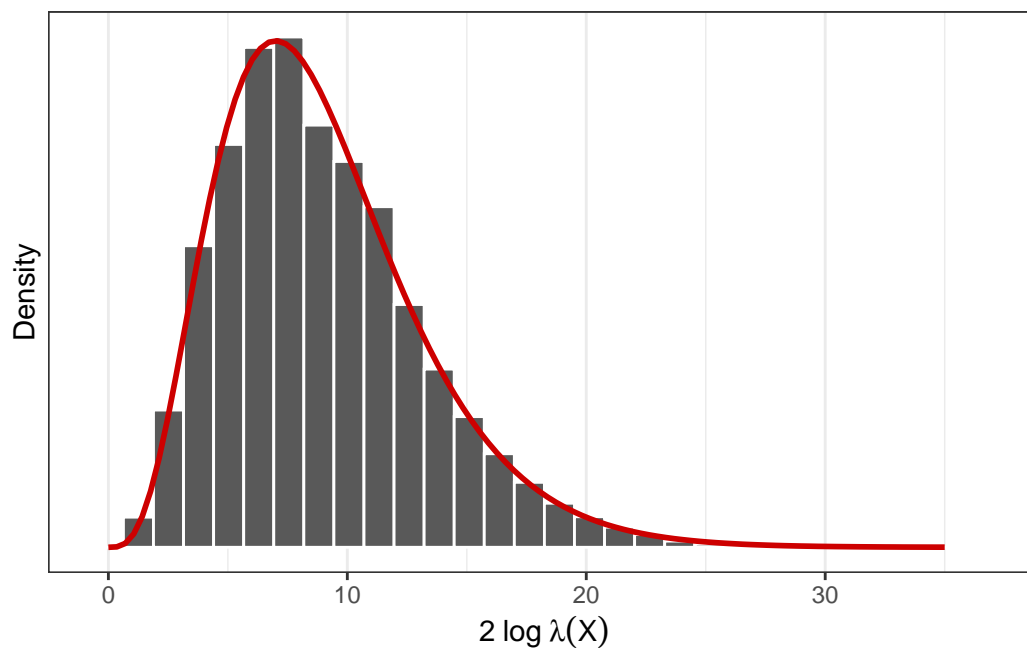


Figure 6.5

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

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

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

7 Interval estimation

Add your notes here.

References