Open Science

replicability, reproducibility, and robustness

Dr. Blazej M. Baczkowski 

February 11, 2025

Table of contents

Welcome

In recent years, the scientific community has faced a significant challenge where many key research findings failed to replicate, undermining the trustworthiness of scientific research (so called *replication crisis*). In response, a transformative shift has emerged (*credibility revolution*) focused on promoting transparency, reproducibility, and robustness through Open Science practices. This course will explore how these practices are restoring trust and enhancing the quality of research. Through a range of activities (presentations, journal-club discussions, and hands-on exercises) students gain insight into how Open Science is reshaping scientific practices. In this course, students develop practical skills for navigating, contributing, and assessing the evolving academic standards. This seminar is designed to benfit everyone regardless whether you plan to pursue a career in academia or beyond – as recipients of science, we are all impacted by its quality and advancements.

Target audiance

- Students from BSc / MSc level (with a background in psychology / social sciences)
- Background in statistical concepts and analysis software (e.g., `R`) is helful but not necessary
- Interest in research quality, transparency, and communication

Course objectives

The primary goal of the seminar is to familiarize students with the tools and practices of Open Science, equipping them with the knowledge and skills to integrate these principles into their own academic activity but also to critically evaluate the work of others, fostering scientific rigor. To name a few specific goals:

- Students can identify several key examples that contributed to the so-called replication crisis, as well as the practices that have since emerged to mitigate such issues in the future.
- Students understand the rigor and current trends in academic standards that make research trustworthy. For example, students recognize the importance of transparency in research and can differentiate between high-quality, reproducible research and studies with potential biases or methodological flaws.
- Students can evaluate the benefits and challenges of Open Science practicies in their field of study. For example, students can assess how *open access* policies improve the dissemination, accessibility, and quality of research with their limitations with respect to legal and ethical considerations.
- Students have practical skills in using Open Science tools, for example, how to write reproducible analysis code and openly share their study materials in appropriate repositories.
- Students have a good overview of what makes scientific research credible.

The course material encourages students to develop a mid-set for seeing their studies (and beyond) as an **ongoing epistemic activity** rather than a *fixed* body of knowledge that needs to be consumed.

Citation

Kellie, D., Kar, F., Balasubramaniam, S., Schneider, M., Schwenke, A., Torresan, O., Waite, C., Fenker, J., Westgate,

M. (2024). *Cleaning Biodiversity Data in R*. (Version 1.0.0). https://doi.org/10.54102/ala.77009.

Acknowledgment

Whenever possible, I tried to acknowledge the sources of the included information. But I am not perfect. If there is some material that is yours and you were not properly acknolwedged, or wish to remove your content, shoot me an email so that I can correct myself.

Funding

The funding to develop this course was provided by the University of Tuebingen in the framework of the prestigious Humboldt Professorship awarded to Prof. Kou Murayama.

(Main) References

Schönbrodt, F. D., & Ihle, M. (2024, October 1). Open Science Workshop Materials of the LMU Open Science Center. Retrieved from osf.io/zjrhu

Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L, Teal TK (2017) Good enough practices in scientific computing. PLoS Comput Biol 13(6): e1005510. https://doi.org/10.1371/journal.pcbi.1005510

The Turing Way Community. (2022). The Turing Way: A handbook for reproducible, ethical and collaborative research (1.0.2). Zenodo. https://doi.org/10.5281/zenodo.7625728

The Turing Way Community, & Scriberia. (2021, May 29). Illustrations from the Turing Way book dashes. Zenodo. https://doi.org/10.5281/zenodo.4906004

Recommended reading

Ioannidis, J. P. A. (2019). Why Most Published Research Findings Are False. CHANCE, 32(1), 4–13. https://doi.org/10.1080/09332480.2019.1579573

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, Robustness, and Reproducibility in Psychological Science. In Annual Review of Psychology (Vol. 73, Issue 1, pp. 719–748). Annual Reviews. https://doi.org/10.1146/annurev-psych-020821-114157

Peikert, A., & Brandmaier, A. M. (2021). A Reproducible Data Analysis Workflow. Quantitative and Computational Methods in Behavioral Sciences, 1, Article e3763. https://doi.org/10.5964/qcmb.3763

Wiebels K, Moreau D. Leveraging Containers for Reproducible Psychological Research. Advances in Methods and Practices in Psychological Science. 2021;4(2). https://doi:10.1177/25152459211017853

Yarkoni T. (2020). The generalizability crisis. The Behavioral and brain sciences, 45, e1. https://doi.org/10.1017/S0140525X20001685

Oberauer, K., Lewandowsky, S. Addressing the theory crisis in psychology. Psychon Bull Rev 26, 1596–1618 (2019). https://doi.org/10.3758/s13423-019-01645-2

Vazire, S. (2018). Implications of the Credibility Revolution for Productivity, Creativity, and Progress. Perspectives on Psychological Science, 13(4), 411-417. https://doi.org/10.1177/1745691617751884

Useful Links

The psyTeachR team. (2024, October 16). Teaching Reproducible Psychology. psyteachr. https://psyteachr.github.io/

Frank, M. C., Braginsky, M., Cachia, J., Coles, N. A., Hardwicke, T. E., Hawkins, R. D., Mathur, M. B., & Williams,

R. 2025. Experimentology: An Open Science Approach to Experimental Psychology Methods. Stanford University. https://doi.org/10.25936/3JP6-5M50.

Leipzig, J (2020). Awesome Reproducible Research: A curated list of reproducible research case studies, projects, tutorials, and media. leipzig/awesome-reproducible-research

Part I

Credibility revolution

1 Introduction

Part II

Statistical reasoning

2 Invention and test

3 The likelihood paradigm

Researchers often state, "We found (strong) evidence for…" But what exactly qualifies as *evidence* in this context? The short answer is **data.** The longer answer is that data become evidence when they are more likely under one hypothesis than another. In other words, data alone do not constitute evidence – they gain evidential value through likelihood comparisons. This concept is grounded in the **likelihood paradigm** (Royall, 1997), which provides the statistical foundation for such claims.

3.1 What is meant by likelihood?

3.1.1 Starting with a toy example: a coin tossing experiment

Imagine that we flip a coin and record its outcome. To simplify, we ignore the angle the coin was thrown at, physical properties of the throw, and other things like that… Because of our ignorance, we cannot perfectly predict the behavior of the coin (i.e., we deal with an uncertain situation). Therefore, our description (i.e., idealization) of a coin flip is *probabilistic* rather than *deterministic*.

How shall we describe our experiment? The outcome of a coin flip can result in only one of two events: landing heads up or tails up. Let us assume that our coin is fair, i.e., landing heads or tails up is equally likely. Hence, the probability of a coin landing heads up is $1/2$. Next, we flip a coin twice and assume that the result of the first flip does not affect the result of the second flip (which seems rational) – the results of coin flips are *independent*. Additionally, each coin toss is assumed to be generated by the same underlying process, which we describe by saying that the flips are *identically distributed*. Hence, together

the the flips are independent and identically distributed (i.i.d.). Alternatively, we say that we have a *random sample*. Most importantly, however, our description constitutes a valid probabilistic model[1], because the probabilities of individual events are non-negative and they sum up to 1.

What is the probability of getting two heads in a row?

The probability of flipping heads in a single coin toss given a parameter θ can be denoted as:

$$Pr(H|\theta)$$

If we assume the flips are independent events, then we multiply the probability of getting heads on the first flip by the probability of getting heads on the second flip (i.e., we obtained heads on the first flip *and* heads on the second flip). Since both flips are independent and were generated by the same underlying process, we have:

$$Pr(H, H|\theta) = Pr(H|\theta) \times Pr(H|\theta)$$

In the case of a fair coin, where $\theta = 1/2$, the equation would give:

$$Pr(H, H|\theta = 1/2) = 1/2 \times 1/2 = 1/4$$

[1] We will provide a more detailed description of what a probabilistic model entails in the next lecture.

What are all the possible outcomes of an experiment with two coin flips?

So now, we intend to perform the experiment – to toss a fair coin twice. What is the list of all possible outcomes we may expect before we flip the coin?

- $Pr(H, H|\theta = 1/2) = 1/4$
- $Pr(T, T|\theta = 1/2) = 1/4$
- $Pr(T, H|\theta = 1/2) = 1/4$
- $Pr(H, T|\theta = 1/2) = 1/4$

When we ignore the order of outcomes, we can simplify:

- $Pr(0H|\theta = 1/2) = 1/4$
- $Pr(1H|\theta = 1/2) = 1/2 = Pr(H, T|\theta = 1/2) + Pr(T, H|\theta = 1/2)$
- $Pr(2H|\theta = 1/2) = 1/4$

3.1.2 Why to use likelihood rather than probability?

We have established above that when we specify the *fairness* of a coin with a θ parameter, we obtain a valid probabilistic model. In other words, when we hold the parameters of our model fixed (e.g., $\theta = 1/2$), the resultant distribution of possible data is a probability distribution that sums up to 1. However, when the data are **fixed** and the parameters **vary**, the obtained probability distribution of our parameters do no sum up to 1.

Let us illustrate the principle by varying the data on one hand and varying the parameter θ on the other. In the table Table 3.1, we make θ to take on one of 6 discrete values $\theta \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. And generate data by flipping the coin two times, which may result in one of the three possible number of heads: $\{0H, 1H, 2H\}$. Moving along each row per column (i.e., holding the parameter fixed), the values sum up to 1. In

contrast, moving along each column per row (i.e., holding the data fixed), the values do not sum up to 1.

Table 3.1: Probability vs Likelihood

	Theta = 0	Theta = 0.2	Theta = 0.4	Theta = 0.6	Theta = 0.8	Theta = 1
Outcome = 0H	1	0.64	0.36	0.16	0.04	0
Outcome = 1H	0	0.32	0.48	0.48	0.32	0
Outcome = 2H	0	0.04	0.16	0.36	0.64	1

How is this relevant? Because in our daily statistical life, the **data are fixed** – the obtained data are a realization of an underlying generative process. Hence, we use the term *likelihood* to expresses the probability of observing the given data as a function of the parameters.

! Equivalence relation

Given a probability density function f such that

$$x \mapsto f(x|\theta),$$

where x represents data and θ represents parameters, the likelihood function is

$$\theta \mapsto f(x|\theta).$$

In other words, when $f(x|\theta)$ is viewed as a function of x with θ fixed, it is a probability density function, and when $f(x|\theta)$ is viewed as a function of θ with fixed x, it is a likelihood function.[2]
To emphasize the distinction, we often write:

$$\mathcal{L}(\theta|x) = p(x|\theta).$$

[2]In the frequentist paradigm, the notation $f(x|\theta)$ is often avoided and instead $f(x;\theta)$ or $f(x,\theta)$ are used to indicate that θ is regarded as a

3.1.3 Introducing the binomial likelihood function

What kind of a likelihood function did we just use to describe our coin tossing experiment? What kind of a probabilistic model did generate the data?

Let us revisit our toy example step by step again but in a more formal way. The status of an individual coin can be denoted by a binary random variable X[3], which takes on one of the two values:

$$X = \begin{cases} 0 & \text{when tail,} \\ 1 & \text{when head.} \end{cases} \tag{3.1}$$

Therefore, the probability of an outcome for an individual coin described by the parameter θ is:

$$\begin{aligned} Pr(X = 1|\theta) &= \theta, \\ Pr(X = 0|\theta) &= 1 - \theta. \end{aligned} \tag{3.2}$$

If we write these two expressions in a single rule, we obtain the following:

$$Pr(X = k|\theta) = \theta^k(1 - \theta)^{1-k} \tag{3.3}$$

where $k \in \{0, 1\}$. The Equation 3.3 is known as *Bernoulli probability density.*

If a random experiment has exactly two possible outcomes (typically referred to as *success* and *failure*), which probability is the same every time the experiment is conducted, then we refer to such a random process as a **Bernoulli trial**[4]. In our toy example, we perform two Bernoulli trials X_1 and X_2 that result

> **Bernoulli distribution**
>
> If X is a random variable with a Bernoulli distribution, then:
>
> $$\begin{aligned} &Pr(X = 1) = \theta, \\ &\theta = 1 - Pr(X = 0). \end{aligned}$$
>
> The probability density function p of this distribution over possible outcomes $k \in \{0, 1\}$ is:
>
> $$p(k; \theta) = \theta^k(1 - \theta)^{1-k}.$$
>
> We can also informally write that the random variable X comes from Bernoulli trials:
>
> $$X \sim Bernoulli(\theta).$$

fixed unknown quantity rather than an outcome of a random process which x is conditioned on. More on that in the next lecture.

[3] A random variable is a function that maps outcomes of an experiment to numerical values.

[4] Another example of a Bernoulli trial would be rolling a die and checking if you get a "6" or not. Getting a "6" is a success, and any other number is a failure.

in two independent and identically distributed observations k_1 and k_2:

$$Pr(X_1 = k_1, X_2 = k_2|\theta) = \theta^{k_1}(1-\theta)^{1-k_1} \times \theta^{k_2}(1-\theta)^{1-k_2} \tag{3.4}$$

Note that the Equation 3.4 can be simplified by rearranging exponent expressions, which yields:

$$Pr(X_1 = k_1, X_2 = k_2|\theta) = \theta^{k_1+k_2}(1-\theta)^{2-k_1-k_2} \tag{3.5}$$

Instead of writing out the outcomes of individual trials, we can also write out the number of heads we obtain, essentially introducing a new random variable K. For example, such an equation for obtaining exactly two heads is:

$$\begin{aligned} Pr(K = 2 \mid \theta) &= Pr(X_1 = 1, X_2 = 1 \mid \theta) \\ &= \theta^{1+1}(1-\theta)^{2-1-1} \\ &= \theta^2(1-\theta)^0 \\ &= \theta^2 \end{aligned} \tag{3.6}$$

and for obtaining exactly 1 heads is:

$$\begin{aligned} Pr(K = 1|\theta) &= Pr(X_1 = 0, X_2 = 1|\theta) + Pr(X_1 = 1, X_2 = 0|\theta) \\ &= \theta^1(1-\theta)^1 + \theta^1(1-\theta)^1 \\ &= 2 \times \theta^1(1-\theta)^1 \end{aligned} \tag{3.7}$$

Note that in the Equation 3.7, there are 2 ways in which the number of heads is 1: $Pr(X_1 = 0, X_2 = 1|\theta)$ or $Pr(X_1 = 1, X_2 = 0|\theta)$. To determine the number of ways to choose k items from a set of n items, without regard to the order of selection, we can use the so-called "n choose k" equation:

$$\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!} \tag{3.8}$$

Now, we are ready to generalise and establish the equation for obtaining k heads in n trials:

$$Pr(K = k|n, \theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k} \tag{3.9}$$

Binomial distribution

The probability of getting exactly k successes in n independent Bernoulli trials with the same probability θ is given by the probability density function p such that:

$$p(k, n, \theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}.$$

We can also informally write that the random variable K comes from binomal distribution:

$$K \sim Binomial(n, \theta).$$

This is called the *binomial probability distribution*. And this is our *likelihood function* for the data from the coin tossing example!

3.1.4 The binomial likelihood function in action

Alright, let us put the binomial likelihood function into practice. For example, suppose that we perform a coin tossing experiment with n trials (coin tosses) and k successes (heads). What is the probability of obtaining $k = 7$ heads when there are $n = 10$ trials and the probability of success is $\theta = 0.7$? To this end, we use the binomial probability density function from the Equation 3.9, substitute the values, and calculate:

$$Pr(k = 7|n = 10, \theta = 0.7) = \binom{10}{7} 0.7^7 \times (1 - 0.7)^{10-7} = 0.27 \tag{3.10}$$

Recall from the Table 3.1 that when we keep the probability of heads fixed, we obtain a valid probability distribution, which is illustrated in the Figure 3.1. And what happens when instead we fix the data, for example, $k = 7$, and make the parameter θ free to vary between 0 and 1? The likelihood function is:

$$\mathcal{L}(\theta|k = 7, n = 10) = \binom{10}{7} \theta^7 (1 - \theta)^{10-7} \tag{3.11}$$

and its shape can be easily obtained in `R` using a `dbinom` function (Figure 3.2).[5]

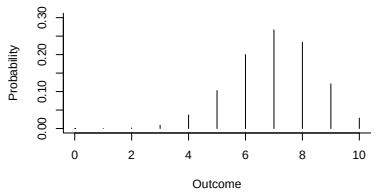


Figure 3.1: Probability distribution of all possible outcomes $k \in [0, 10]$ when $n = 10$ trials and the probability of heads is $\theta = 0.7$.

```
# Define a sequence of theta values from 0 to 1, with a step size of 0.01
thetas = seq(0, 1, by=.01)

# Calculate the likelihood for observing k=7 successes in n=10 trials for each value of theta
liks = dbinom(x=7, size=10, prob=thetas)

# Plot the likelihood function for different values of theta
plot(thetas,                    # x-axis: values of theta
```

[5] Note the striking difference from the Figure 3.1, where the probability values are discrete and sum up to 1.

```
     liks,                          # y-axis: corresponding likelihoods
     xlab=bquote(theta),                 # Label for x-axis
     ylab="Likelihood", # Label for y-axis
     type="l",                      # Line plot to show the likelihood curve
     bty="l",                       # Box type for the plot (L-shaped)
     #main=bquote("L("*theta*" | k=7, n=10)"),  # Main title with LaTeX-style formatting for tl
     xlim=c(0, 1),                  # Set the x-axis range (theta from 0 to 1)
     ylim=c(0, max(liks) + .05)   # Set the y-axis range, slightly above the maximum likelihood
)
```

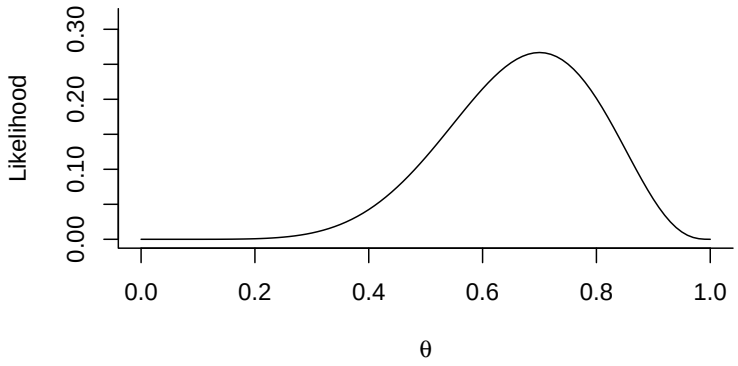


Figure 3.2: The binomial likelihood function when $k = 7$ and $n = 10$.

What is the meaning of this curve? In a nutshell, the function describes how likely different values of θ are, given the observed data $k = 7$ heads in $n = 10$ trials. It clearly shows that the values close to 0.7 are more likely than those close to either 0 or 1. In other words, likelihood functions are plotted to provide a visual impression of the evidence over the parameter space.

We are now well equipped to tackle the basic tenets of the likelihood paradigm. Let's do it!

For presentation purposes, likelihood functions may be standardized by an arbitrary value (a constant), typically by their maximum value which scales its range to be between 0 and 1.

3.2 The law of likelihood

Now that we have established the likelihood function, we can apply it to compare the likelihood of different parameters that correspond to our specific hypotheses. Essentially, we use the likelihood function to assess which hypothesis is best supported by the data. To this end, we introduce the Law of Likelihood.

> If hypothesis A implies that the probability that a random variable X takes the value x is $p_a(x)$, while hypothesis B implies that the probability is $p_b(x)$, then the observation $X = x$ is evidence supporting A over B if and only if $p_a(x) > p_b(x)$, and the likelihood ratio, $p_a(x)/p_b(x)$, measures the strength of that evidence (Hacking, 1965).

Here $p_a(x) = p(x|H_a)$ is the probability of observing x given that hypothesis A is true and $p_b(x) = p(x|H_b)$ is the probability of observing x given that hypothesis B is true. The ratio of these conditional probabilities, $p(x|H_a)/p(x|H_b)$, is the likelihood ratio.[6] In other words, if an event is more probable under hypothesis A than hypothesis B, then the occurrence of that event is the evidence supporting A over B. The *degree* to which the occurrence of the event supports A over B is quantified by the ratio of the two probabilities:

$$\Lambda(\theta) = \frac{\mathcal{L}(\theta_a|X = x)}{\mathcal{L}(\theta_b|X = x)} = \frac{p(X = x|\theta_a)}{p(X = x|\theta_b)} \tag{3.12}$$

> **!** Simple vs composite hypothesis
>
> The **Law of Likelihood** is applied in contexts where hypotheses are **simple**, meaning they specify a single value for each parameter of interest. A **composite hypoth-**

[6] The concept of statistical evidence is essentially relative in nature; the data represent evidence for one hypothesis in relation to another. The data do not represent evidence for, or against, a single hypothesis. Why would it be wrong to say that the data represent evidence against H_a if $p(x|H_a)$ is small? The reason is $p(x|H_a)$, while small, might be the largest among the hypotheses under consideration, making H_a the hypothesis best supported by the data.

esis, in contrast, is one that does not specify a single value for the parameter but instead includes a range or set of possible values. For example, in a coin-tossing experiment: a *simple hypothesis* is $\theta = 0.5$ while a *composite hypothesis* is $\theta \geq 0.5$ which allows for multiple values of θ. However, this does not mean that likelihood-based reasoning is restricted to single-parameter models. Instead, it means that each hypothesis being compared corresponds to a **specific likelihood function** rather than a range of possible parameter values. In *multi-parameter models*, a simple hypothesis fully specifies all parameters. For example, a *simple hypothesis* is $\mu = 0$ and $\sigma = 0.5$ while a *composite hypothesis* is $\mu > 0$ and $\sigma = 0.5$ since μ is not fully specified.

The likelihood ratio directly reflects the strength of the evidence. For the purpose of interpreting and communicating the strength of evidence, it is useful to divide the *continuous scale* of the likelihood ratio into descriptive categories (Table 3.2). Note that while a likelihood ratio of 8 represents fairly moderate evidence, so does a likelihood ratio of 7.5 or 10 (albeit to a lesser or greater degree).

Table 3.2: Likelihood ratio interpretation.

Likelihood ratio	Evidence strength
1	neutral
1-8	weak
8-32	moderate
32+	strong

For example, given the data from the previous section in the coin tossing experiment, $k = 7$ and $n = 10$, we have only weak evidence that the coin is biased (Figure 3.3). The likelihood ratio between two hypotheses $\theta_a = 0.7$ vs $\theta_b = 0.5$ is:

$$\frac{\mathcal{L}(\theta = 0.7|k = 7, n = 10)}{\mathcal{L}(\theta = 0.5|k = 7, n = 10)} = 2.3.$$

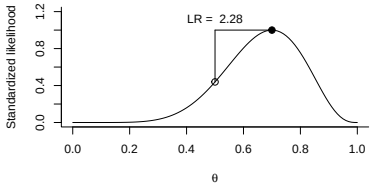


Figure 3.3: The standardized likelihood function indicating the ratio between $\theta = 0.5$ and $\theta = 0.7$.

3.3 Support intervals

Once we have observed the data, can we determine a range of parameter values that remain plausible at a given level of support? Within the likelihood framework, one approach is to construct a *support interval*, which consists of all parameter values for which the likelihood exceeds a specified fraction of its maximum. This interval directly reflects the relative strength of evidence provided by the data.

> A support interval is an interval of parameter values for which the likelihood remains above a specified threshold relative to its maximum, indicating the level of support provided by the data.

For example, the values of θ that are most consistent with the data correspond to those near the peak of the likelihood function. A $1/\lambda$ likelihood support interval is then defined as the set of θ values for which the likelihood remains above $1/\lambda$ times its maximum value. Formally, we can express the support interval as:

$$\left\{\text{all } \theta \text{ where } \frac{\mathcal{L}(\theta)}{max_\theta \mathcal{L}(\theta)} \geq \frac{1}{\lambda}\right\} \quad = \quad \left\{\text{all } \theta \text{ where } \frac{\mathcal{L}(\hat{\theta})}{\mathcal{L}(\theta)} \leq k\right\} \tag{3.13}$$

In words, the interval is determined by comparing the likelihood at each possible value of the parameter to the maximum likelihood. In this framework, the support interval doesn't directly correspond to the probability of the parameter lying in the interval, but rather reflects the range of values for which the data provides reasonable support under the likelihood function. Any θ within the $1/\lambda$ interval is supported by the data because the best supported hypothesis, $\hat{\theta}$, is only better supported by a factor of λ or less.

To illustrate the implications of a support interval, let's consider a concrete example. We revisit the coin-tossing experiment (i.e., a binomial model) and examine two scenarios: one with 10 trials and another with 10 times as many. In both cases, the observed proportion of heads is the same, with $\frac{k_1=7}{n_1=10}$ in the first scenario and $\frac{k_2=70}{n_2=100}$ in the second.

```r
# Create a sequence of possible theta values from 0 to 1 with a step size of 0.01
thetas <- seq(0, 1, .01)

# Define the number of heads and tosses for the first scenario
n_heads <- 7
n_tosses <- 10

# Calculate the likelihood for each theta given the observed number of heads (n_heads) and toss
lik <- dbinom(n_heads, size=n_tosses, prob=thetas)

# Scale the likelihood by dividing each value by the sum of all likelihoods
lik <- lik / max(lik)

# Set up the plot with an empty frame, specify limits and labels
plot(x=.5,                          # Initial x position (not used here as it's an empty plot)
     type='n',                      # Type 'n' creates an empty plot
     xlim=c(0, 1),                  # x-axis limits (theta from 0 to 1)
     ylim=c(0, 1.3),                  # y-axis limits for the scaled likelihood
     bty="l",                       # Box type 'L' for the plot border
     #main=bquote("L("*theta*" | k=7, n=10) vs L("*theta*" | k=70, n=100)"),  # Main title witl
     xlab=bquote(theta),                 # x-axis label
     ylab='Normalised likelihood'   # y-axis label
)
# Add a custom x-axis label at theta = 0.5
axis(1, at=.5, labels=paste(.5))

# Plot the first likelihood curve (scaled likelihood for n_tosses = 10)
lines(thetas, lik, type='l', lty=5)  # Adds a line plot for the first likelihood

# Define the second scenario with more tosses (10 times the original)
n_tosses_more <- n_tosses * 10
n_heads_more <- n_heads * 10

# Calculate the likelihood for the second scenario (scaled likelihood for n_tosses = 100)
lik_100 <- dbinom(n_heads_more, size=n_tosses_more, prob=thetas)

# Scale the likelihood for the second scenario
lik_100 <- lik_100 / max(lik_100)

# Plot the second likelihood curve (scaled likelihood for n_tosses = 100)
```

```r
lines(thetas, lik_100, type='l', lty=3)  # Adds a dashed line for the second likelihood

# Mark specific points on the plots for comparison at theta = 0.5 and theta = 0.7
points(.5, lik[51], pch=21, cex=1.2)  # Point for theta = 0.5 in the first plot
points(.5, lik_100[51], pch=21, cex=1.2)  # Point for theta = 0.5 in the second plot
points(.7, lik[71], col='black', pch=19, cex=1.2)  # Point for theta = 0.7 in the first plot

# Calculate the likelihood ratios for the two scenarios at theta = 0.7 vs theta = 0.5
lik_ratio1 <- round(lik[71] / lik[51], 1)  # Ratio for the first plot (n_tosses = 10)
lik_ratio2 <- round(lik_100[71] / lik_100[51], 1)  # Ratio for the second plot (n_tosses = 100)

# Annotate the plot with the likelihood ratios
t1_x <- 0.4
t1_y <- 0.7
t2_x <- 0.3
t2_y <- 0.4
text(t1_x, t1_y, paste("LR1 = ", lik_ratio1), adj=0, pos=2, cex=.9, col='grey')  # Annotation f
text(t2_x, t2_y, paste("LR2 = ", lik_ratio2), adj=0, pos=2, cex=.9, col='grey')  # Annotation f
segments(t1_x, t1_y, 0.5, lik[51], col='grey')
segments(t2_x, t2_y, 0.5, lik_100[51], col='grey')

# Add a horizontal line at the 1/8 likelihood level
abline(h=1/8, col="black", lty=1)
text(0.1, 1/8, "1/8 bound", pos=3, col="black")


legend("topright",                      # Position of the legend
       legend = c("n1", "n2"), # Labels
       col = c("black", "black"),         # Colors of the lines
       lty = c(5, 3),                        # Line types
       lwd = 2,
       bty = "n"
       )
```

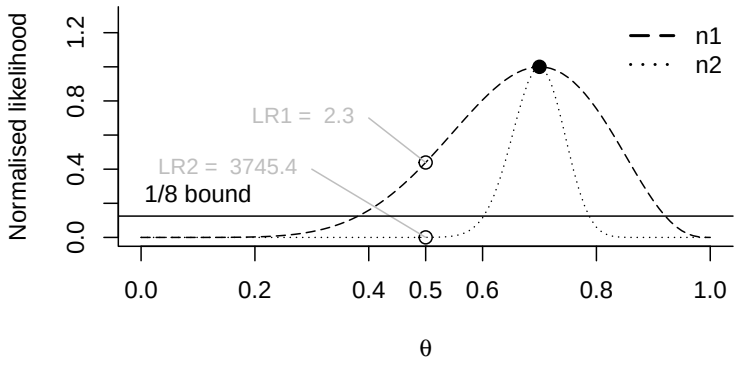


Figure 3.4: The standardized likelihood function for a binomial model with two different sample sizes $n_1 = 10$ trials and $n_2 = 100$ trials.

3.4 Probability of obtaining weak and misleading evidence

> Evidence is considered misleading if it strongly supports the wrong hypothesis while weak evidence refers to a situation where the evidence does not clearly favor either hypothesis.

Once the data are collected, the strength of the evidence is assessed using the likelihood ratio, which quantifies whether the evidence is weak or strong. Weak evidence is merely uninformative and does not contribute meaningfully to conclusions. To prevent such outcomes, we can evaluate whether a given study design is likely to produce weak evidence. Similarly, we can assess whether a study design might generate misleading evidence. Unlike weak evidence, misleading evidence is particularly problematic because it strongly supports the wrong hypothesis, leading to incorrect conclusions.[7]

[7]It is critical to distinguish between (a) the probability of observing misleading evidence (future) vs (b) the probability that the observed evi-

The probability of observing weak evidence favoring either hypothesis θ_0 or θ_1 when θ_0 is the correct hypothesis depends on the strength of the evidence λ and the sample size n:

$$p\Big(1/\lambda < \frac{\mathcal{L}_n(\theta_1)}{\mathcal{L}_n(\theta_0)} < \lambda\Big) \quad = \quad W(n, \lambda). \tag{3.14}$$

Likewise, the probability of observing misleading evidence for θ_1 over θ_0 depends on the strength of the evidence λ and the sample size n:

$$p\Big(\frac{\mathcal{L}_n(\theta_1)}{\mathcal{L}_n(\theta_0)} \geq \lambda\Big) \quad = \quad M(n, \lambda). \tag{3.15}$$

To better understand these principles, let us simulate some data. To this end, we will use a custom written functions in `R`. The Listing 3.1 and Listing 3.2 include the functions to calculate the probability of weak and misleading evidence, respectively. The outcome of the simulation is plotted in the Figure 3.5.

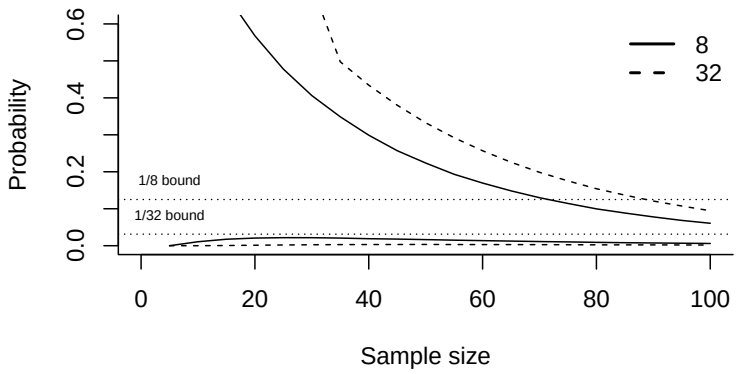


Figure 3.5: The probability of observing weak and misleading evidence in a binomial model for a given strength of evidence (8 vs 32) and sample size.

The outcome of the simulation is quite striking! With increasing sample size, the probability of observing weak evidence is

dence is misleading (past).

clearly decreasing. Second, the frequency with which misleading evidence is observed is, in general, low. In fact, for any fixed sample size and any pair of probability distributions, the probability of observing misleading evidence of strength λ or greater is always less than or equal to $1/\lambda$.

! The universal bound

One nice feature of likelihood ratios is that they are seldom misleading. The probability of observing misleading evidence of λ-strength is always bounded by $1/\lambda$, the so-called **universal bound**.
Mathematically, if both $f(x)$ and $g(x)$ are probability density functions and x is distributed according to $f(x)$ then:

$$p_f\left(\frac{g(x)}{f(x)} \geq \lambda\right) \quad \leq \quad \frac{1}{\lambda}$$

3.4.1 Sequential designs

The universal bound applies to both fixed sample size designs and sequential study designs. In a sequential study, the probability of observing misleading evidence does increase with the number of looks at the data. However, multiple looks at the data do not affect the likelihood function, and the amount by which misleading evidence increases shrinks to zero as the sample size grows. As a result, the overall probability can remain bounded and is less than $1/\lambda$. Thus, an experimenter who plans to examine the data with each new observation, stopping only when the data support the alternative over the null, will be frustrated with probability at least $1 - 1/\lambda$. Practically speaking, one will be frustrated quite often. This does not seem obvious, so let us show the principle using simulations.

i The universal bound in sequntial designs

The probability of observing misleading evidence in a fixed sample size study is less than the corresponding probability in a sequential study, but both are less than $1/\lambda$.

Mathematically, if both $f(x)$ and $g(x)$ are probability density functions and x is distributed according to $f(x)$ then:

$$p_f\Big(\frac{g(x_n)}{f(x_n)} \geq \lambda; \text{ for any } n = 1, 2, ...\Big) \quad \leq \quad \frac{1}{\lambda}$$

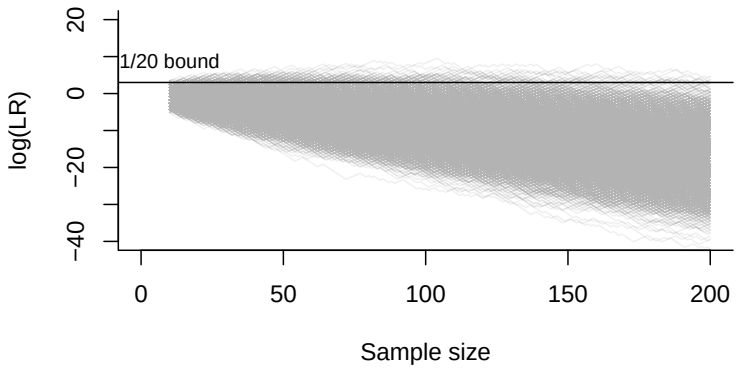


Figure 3.6: Likelihood ratio in a sequential design. Each grey line indicates a single simulated study.

The Figure 3.6 depicts a sequential design under simulated repeated hypothesis testing with each new observation. The design sets the likelihood ratio at the level $\lambda = 20$, and tests between two hypotheses $\theta_0 = 0.5$ and $\theta_1 = 0.7$. The data speak for themselves. The probability of observing misleading evidence of λ-strength or greater remains bounded by $1/\lambda$ for any number of looks under very general conditions. Indeed, the simulation shows that the probability of observing misleading evidence is 0.044, and hence below the bound of $1/\lambda = 1/20$. It is also clear that as the amount of data grows, the strength of evidence against the alternative hypothesis increase.

Tip

Note that what we did here is we repeatedly (every sample) peeked at data to check the evidence strength between

two pre-specified (*a priori*) single-valued hypotheses (null and alternative). We did not look for the evidence between a null and **any** alternative hypothesis. This is not a subtle distinction. Do you know why?

Let's explore what happens when we examine the data after every sample and treat the parameter value with the maximum likelihood as our alternative hypothesis. In other words, with each new data point, we update our alternative hypothesis to reflect the observed results, effectively selecting the hypothesis that best fits the data at that particular moment.

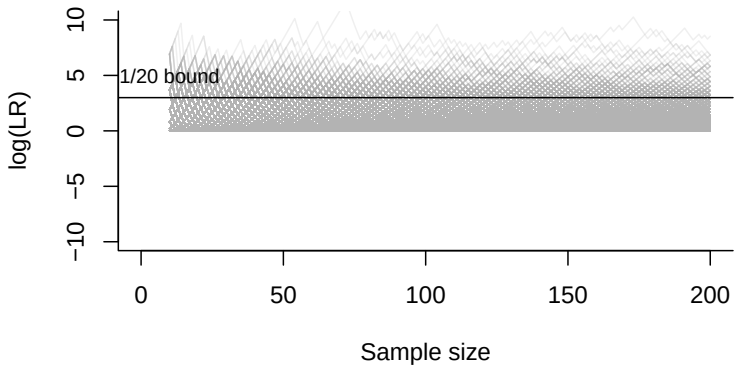


Figure 3.7: Likelihood ratio in a sequential design, where the alternative hypothesis is continuously updated to match the value of maximum likelihood.

We clearly observe a different pattern. As the amount of data grows, the strength of evidence against the alternative hypothesis remains stable. Moreover, the simulation shows that the probability of observing misleading evidence is 0.138, and hence above the bound of $1/\lambda = 1/20$.

> **Warning**
>
> Switching hypotheses as we collect data increases the probability of being misled more often than we would like. In other words, the probability of observing misleading evidence of λ-strength or greater is not guaranteed to be bounded by $1/\lambda$ threshold.

3.5 Is the observed evidence misleading?

We have established that once the data are collected, the strength of the evidence is determined by the likelihood ratio. The probability of obtaining *another* set of data that might be misleading is no longer relevant; the question of what *could* have happened becomes insignificant. What matters is whether the *observed* data are misleading or not. Unfortunately, we will never be certain if the observed result is misleading – sorry to disappoint you! However, it is sometimes possible to assess the probability that the observed results could be misleading. So, how should we approach this challenge?

Let us consider a hypothethical example based on a medical test T for a disease D. An observed positive test for the disease D is wrong *if and only if* the patient does not have the disease. Knowing the disease prevalence in the general population give us a clue whether the patient has a disease or not. Let us assume that the prevalence of the disease is 5% in the general population, thus $P(D+) = 0.05$ and by symmetry $P(D-) = 1 - 0.05$. Given the parameters of a specific test such as its sensitivity (true positives) expressed as $P(T+|D+) = .94$ and specificity (true negatives) expressed as $P(T+|D-) = .02$, we can compute the strength of evidence, $\lambda = P(T+|D+)/P(T+|D-) = .94/.02 = 49$. So we have $\lambda = 49$ and prior probability of the disease $P(D+) = 0.05$. What remains is to compute the probability of the patient to have the disease given the positive test, which is $P(D+|T+)$. We can compute that probability using so called *Bayes theorem*. We will explain the theorem in the next chapter, but for now simply realize based on daily experience that having known the

test came out positive, we would change our belief about having the disease. If the test is positive, the belief would increase. Otherwise, it would decrease. In other words, having obtained the test results, we "scale" our beliefs about the odds of having the disease (i.e., the chance of having the disease vs not-having the disease). This intuitive reasoning can be expressed as:

$$\lambda \times \text{prior odds} = \text{posterior odds} \tag{3.16}$$

Expressed more formally:

$$\lambda \times \frac{P(D+)}{1 - P(D+)} = \frac{P(D+|T+)}{1 - P(D+|T+)} \tag{3.17}$$

The left-hand side is our beliefs about the disease before seeing the result and scaled by the evidence. The right-hand side is our updated beliefs due to evidence.

If we re-arrange the Equation 3.17, we get:

$$P(D+|T+) = \left[1 + \frac{1 - P(D+)}{\lambda \times P(D+)}\right]^{-1} \tag{3.18}$$

! The probability of the evidence being misleading

$$P(\theta_0|data) = \left[1 + \frac{\lambda \times (1 - P(\theta_0))}{P(\theta_0)}\right]^{-1}$$

Let us plot how the probability of misleading evidence changes as a function of prior probability of the hypothesis to be false.

In short, the probability of misleading evidence depends on the prior odds of our hypotheses. This answers the question why the same statistical evidence may convince one researcher but not the other. This is because they may have different prior probabilities for the hypotheses depending on their own experience and beliefs.

You strongly believe that a coin is fair. Now consider a scenario where the coin is tossed, and for every 8 heads, there is 1 tail (LR=8). How strongly do you belive now that the coin is fair?

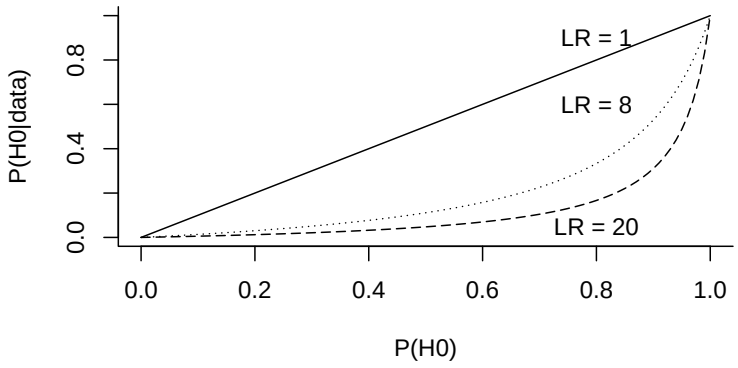


Figure 3.8: Probability of observed evidence to be misleading as a function of likelihood ratio.

! Important

The likelihood ratio quantifies statistical evidence that uniformly modify prior beliefs:

$$\frac{p(\theta_1|x)}{p(\theta_0|x)} = \lambda \times \frac{p(\theta_1)}{p(\theta_0)} \tag{3.19}$$

where $\lambda = p(x|\theta_1)/p(x|\theta_0)$ is the likelihood ratio, $p(\theta_1)/p(\theta_0)$ is prior probability ratio, and $p(\theta_1|x)/p(\theta_0|x)$ is the posterior probability ratio.

3.6 Summary: Evidential metrics

The foundation of statistical inference is the interpretation of data as evidence. There are three essential quantities for assessing and interpreting the strength of statistical evidence in data (Table 3.3).

Table 3.3: Evidential metric.

	Metric	What it measures
1	Likelihood ratio (LR)	The strength of the evidence
2	Pr(LR > | H0) or P(LR < 1/ | H1)	The probability that a particular study design will generate misleading evidence
3	P(H0 | LR =) or P(H1 | LR = 1/)	The probability that the observed evidence is misleading

3.7 Exercises

1) Modify the value of the alternative hypothesis to check how it affects the probability of misleading and weak evidence. How does it change when the difference between theta 1 and 0 grows?
2) Modify the code and check how the probability of misleading evidence change in a sequential design when data are inspected after every 10th observations. Does the probability increase or decrease?

3.8 Relation to Open Science

The likelihood paradigm (Royall, 1997) is highly relevant for **open science** in several key ways:

1. Transparent Evidence Quantification. The likelihood paradigm emphasizes *direct comparison of evidence* via likelihood ratios (rather than relying on p-values). Open science promotes *transparency in statistical inference*, and likelihood-based methods align well with this by offering a clear and interpretable measure of evidence.

2. Reproducibility and Robustness. The likelihood approach encourages *reporting full likelihood functions*,

which allows other researchers to reanalyze data, incorporate new data, and update inferences. This is crucial for *reproducibility*, a core principle of open science.

3. Avoiding Dichotomous Thinking. Classical hypothesis testing often leads to *binary "significant vs. not significant" thinking*, which can be misleading. Likelihood ratios provide *graded evidence strength*, which fits well with open science's push for *nuanced, data-driven interpretations* rather than rigid thresholds.

4. Supports Bayesian and Frequentist Paradigms. The notion of *likelihood* is present in both bayesian and frequentist frameworks. Proper understanding of the likelihood is necessary for the proper understanding of the other approaches. Hence, supporting rigourous research.

5. Combatting P-Hacking and Publication Bias. Since likelihood methods don't rely on *arbitrary significance thresholds*, they reduce incentives for *p-hacking and selective reporting*, which are major concerns in open science.

3.9 References

https://www.statisticalevidence.com/

Blume J. D. (2002). Likelihood methods for measuring statistical evidence. Statistics in medicine, 21(17), 2563–2599. https://doi.org/10.1002/sim.1216

Blume, J. D., & Choi, L. (2017). Likelihood Based Study Designs for Time-to-Event Endpoints (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1711.01527

Blume, J. D. (2011). Likelihood and its Evidential Framework. In Philosophy of Statistics (pp. 493–511). Elsevier. https://doi.org/10.1016/b978-0-444-51862-0.50014-9

Hacking, I. (1965). Logic of Statistical Inference. Cambridge Univ. Press.

Lambert, B. (2018). A Student's Guide to Bayesian Statistics. SAGE Publications Ltd., London, UK.

Morey, R. D. (2018). Statistical Inference. In Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience (pp. 1–42). Wiley. https://doi.org/10.1002/9781119170174.epcn504

Royall (1997). Statistical Evidence: A Likelihood Paradigm. Chapman and Hall

Listing 3.1 Function written in R to calculate the probability of observing weak evidence in a binomial model.

```r
get_p_weak <- function(
  nobs = 14,   # Number of observations (sample size)
  k = 32,      # Likelihood ratio threshold for weak evidence
  theta1 = 0.8, # Alternative hypothesis parameter
  theta0 = 0.5, # Null hypothesis parameter
  nsim = 1e+2  # Number of simulations
  ){

  # Simulate data under the null hypothesis
  dat <- rbinom(n=nsim, size=nobs, prob=theta0)

  # Compute likelihood of observed data under the alternative hypothesis
  lik1 <- dbinom(x=dat, size=nobs, prob=theta1)

  # Compute likelihood of observed data under the null hypothesis
  lik0 <- dbinom(x=dat, size=nobs, prob=theta0)

  # Compute the likelihood ratio (LR)
  LR <- lik1 / lik0

  # Calculate the proportion of cases
  # where the likelihood ratio is between 1/k and k
  # This represents the probability of weak evidence
  w <- sum(1/k < LR & LR < k) / nsim

  # Return the probability of weak evidence
  return(w)
}
```

Listing 3.2 Function written in R to calculate the probability of observing misleading evidence in a binomial model.

```r
get_p_misleading <- function(
  nobs = 14,   # Number of observations (sample size)
  k = 8,       # Likelihood ratio threshold for misleading evidence
  theta1 = 0.7, # Alternative hypothesis parameter
  theta0 = 0.5, # Null hypothesis parameter
  nsim = 1e+2  # Number of simulations
  ){

  # Simulate data under the null hypothesis
  dat <- rbinom(n=nsim, size=nobs, prob=theta0)

  # Compute likelihood of observed data under the alternative hypothesis
  lik1 <- dbinom(x=dat, size=nobs, prob=theta1)

  # Compute likelihood of observed data under the null hypothesis
  lik0 <- dbinom(x=dat, size=nobs, prob=theta0)

  # Compute the likelihood ratio (LR)
  LR <- lik1 / lik0

  # Calculate the proportion of cases where LR exceeds the threshold 'k'
  # This represents the probability of misleading evidence
  m <- sum(LR >= k) / nsim

  # Return the probability of misleading evidence
  return(m)
}
```

4 The Bayesian paradigm

In the previous chapter, we established how to evaluate evidence from the observed data. Once we have seen the data, the next question arises: "What should we *believe*?" This question moves beyond simple analysis and asks how we can incorporate the evidence we've gathered into our knowledge. **Bayesian inference** provides a powerful framework for answering this by allowing us to continuously refine our knowledge as new information becomes available.

At its core, Bayesian inference begins with the concept of the *prior* – our initial set of beliefs before encountering data. As we observe new information, we use the *likelihood* to measure how probable the observed data are under different assumptions. By applying **Bayes' theorem**, we use the prior and likelihood combined to form the *posterior* – the updated set of beliefs. Essentially, Bayesian inference offers a systematic approach to learning from data, helping us navigate through uncertainty by constantly adjusting our predictions based on the latest evidence. As beautifully put by John Kruschke *Bayesian inference is the re-allocation of credibility across possibilities.* Let's dive in to unpack it!

4.1 Conditional probability

At the heart of Bayesian inference lies the notion of **conditional probability.** To form an intuitive understanding of the concept, consider the following two scenarios. In one scenario, you randomly pick a person in a supermarket. What's the probability that they own a house? In the second scenario, you pick a person in a supermarket but are told this person is a student. What's the probability that this person owns a house? In the former, you consider everyone (i.e., population of a city).

In the latter, you only consider the subset of students, which changes the probability of owning a house based on the extra information. In other words, the new information about the person being a student changes your belief about them owning a house.[1]

4.1.1 Elementary, yet powerful ideas of probability theory

> A probabilistic model is a mathematical description of an uncertain situation.

We have already formalised and used a probabilistic model in the previous chapter. Now, we provide a more detailed description of what a probabilistic model entails.

Creating a **probabilistic model** involves two steps:

1) describing possible outcomes (**sample space**),
2) describing our beliefs about the likelihood of possible outcomes (**probability law**).

Yes, this it – the Figure 4.1 depicts the main ingredients of a probabilistic model.

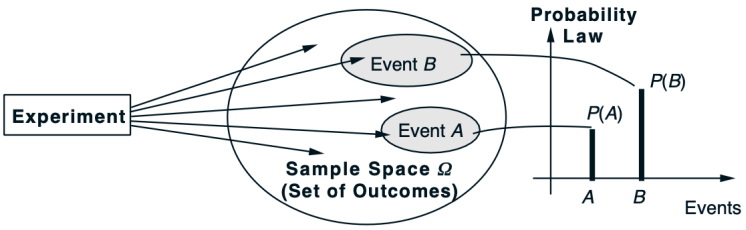


Figure 4.1: Ingridients of a probabilistic model. From Bertsekas and Tsitsiklis (2000) used under terms of copyright for instructional purposes.

[1] Another example: A motion sensor detects movement in a room. How likely is it that the movement corresponds to a person rather than an animal?

Sample space is a set of all possible outcomes Ω. The elements of that set are mutually exclusive and collectively exhausting. Sets can be continuous or discrete / finite, and so do sample spaces. For example, flipping a single coin results in one of two possible outcomes: heads or tails up. Together, these two outcomes exhaust all possibilities and are mutually exclusive (it is either heads or tails, never both at the same time). Moreover, it is a discrete / finite set. When we toss a coin and observe an outcome, an **event** has occured. An event is a subset of a sample space. For example, an event A is a coin that landed heads up while an event B is a coin that landed tails up.[2] We may ask: what is the probability of a coin landing heads up? To answer the question, we need to specify **the probability law** that assigns probabilities to events. These probabilities[3] must be non-negative, i.e., $P(A) \geq 0$, follow the additivity rule: if A and B are two disjoint events, then the probability that one of them happens is their sum $P(a) + P(B)$, and sum up to 1, i.e., $P(\Omega) = 1$. For example, the probability of a coin landing heads up is $P(A) = .4$ while the probability of a coin landing tails up is $P(B) = 1 - P(A)$. We have just created a valid probabilistic model .

Discrete Uniform Probability Law

If the sample space consists of n possible outcomes which are equally likely (i.e., all single-element events have the same probability), then the probability of any event $Pr(A)$ is given by:

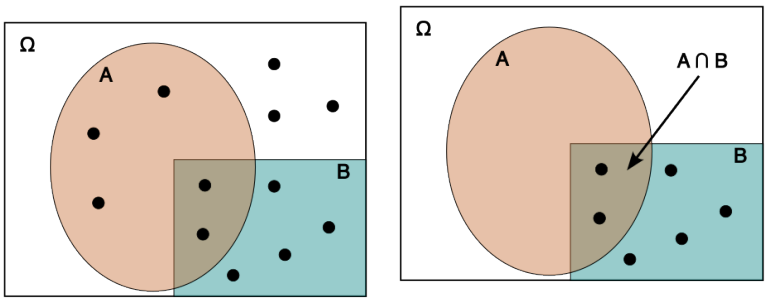
$$Pr(A) = \frac{\text{number of elements of } A}{n}$$

[2] In theory, we may also specify an event that corresponds to a coin landing on any side. But this model would not be very practical, as the same event would always occur.

[3] These are the three probability axioms:
1. Non-negativity
2. Additivity
3. Normalisation.

4.1.2 Using new information to revise a probabilistic model

Let us consider the following probabilistic model depicted in the Figure 4.2. There are 12 equally likely outcomes. Hence, their individual probabilities are $\frac{1}{12}$ each. We focus on two events (two subsets of the sample space) such that an event A has 5 elements so that its probability is $P(A) = \frac{5}{12}$, and the event B that has 6 elements so that its probability is $P(B) = \frac{6}{12}$.



(a) Probabilistic model with two events A and B.

(b) Revising the model based on partial information that the event B occurred.

Figure 4.2: Illustration of conditional probability. Based on Bertsekas and Tsitsiklis (2000) used under terms of copyright for instructional purposes.

Suppose now, that someone told you that an event B has occurred. How should the model change? First, those outcomes that are outside the event B are no longer possible. We can eliminate them or assign 0 probability to them. Second, we are told that one of the outcome of B has occurred. The individual probabilities of these outcomes did not change – they are still all equally likely. Hence, the probability of each outcome of an event B is $\frac{1}{6}$. And the probability of an event B is $\frac{6}{6}$ so 1. Well, we were told that the event B has occurred. We denote this situation as $P(B|B) = 1$. But what about the probability of an event A? Well, there are two possible outcomes of an event A that are also in B. Hence, the conditional probability of A given that B occurred is $P(A|B) = \frac{2}{6}$. Once we are told that an event B occurred, the probability of an event A has changed from $\frac{5}{12}$

Conditional probability when all outcomes are equally likely is given by:

$$P(A|B) = \frac{\text{number of elements of } A \cap B}{\text{number of elements of } B}$$

to $\frac{1}{3}$. Note that what happened is that we simply changed the relevant sample space from Ω to B. The probability of individual outcomes of A has not changed $P(A) = \frac{1}{6} + \frac{1}{6}$.

How shall we express the reasoning in a formula. Given the total probability assigned to event B, we want to determine what fraction of this probability $P(A \cap B)$ is assigned to the outcomes where event A also occurs. As a result, conditional probability $P(A|B)$ provides us with a way to reason about the outcome of an experiment based on new information.

! Conditional probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \text{ when } P(B) > 0$$

In words, out of the total probability of the elements of B, $P(A \cap B)$ is the fraction that is assigned to possible outcomes that also belong to A.

4.1.2.1 Example: Spam filter

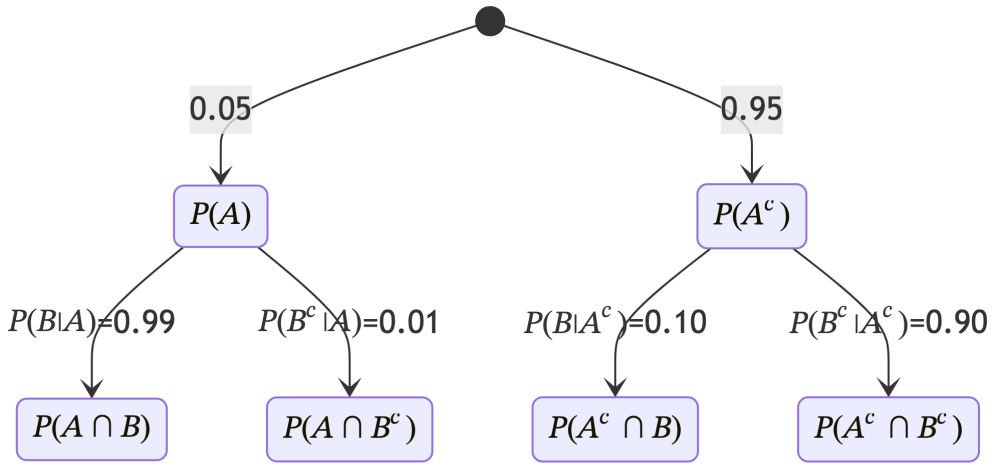


Figure 4.3: Sequential description of the sample space for the spam detection problem.

Conditional probabilities are used to revise a probabilistic model when we obtain new information. But we can also use

conditional probabilities to build a multi-stage probabilistic model. Let us consider an example that involves the detection of a spam email. Let A and B be the events

- $A = \{\text{a spam message is received}\}$,
- $B = \{\text{a spam filter registers a spam message}\}$,

and their complements are

- $A^c = \{\text{a regular message is received}\}$,
- $B^c = \{\text{a spam filter does not register a spam message}\}$.

The given probabilities are recorded along the corresponding branches of the tree describing the sample space, as shown in the Figure 4.3. Each event of interest corresponds to a leaf of the tree and its probability is equal to the product of the probabilities associated with the branches in a path from the root to the corresponding leaf.

A message is sent to an inbox. It is either a spam, i.e., event A occurs, or a regular message, i.e., event A does not occur. The probabilities of these events are as follows $P(A) = 0.05$ and $P(A^c) = 0.95$, respectively. Then a spam filter registers the message as a spam, i.e., event B occurs. Alternatively, the filter does not register the spam, i.e., event B does not occur. How good is the filter? Its specs are essentially formulated as conditional probabilities, e.g., the sensitivity of the filter (the filter correctly identifies a spam message) is $P(B|A)$ while the specificity of the filter (the filter correctly identifies a regular message) is $P(B^c|A^c)$. As a result, we have various scenarios. For example, it is possible that a spam message is received but the filter did not register it (i.e., the so-called "miss"). This case would be denoted as $P(A \cap B^c)$. Conversely, a regular message is received but the filter does register it as a spam (i.e., the so-called false alarm). This case is denoted as $P(A^c \cap B)$.

Given the assigned probabilities depicted in the figure, let us, for example, calculate the probability of receiving a spam message that is correctly registered:

$$P(A \cap B) = P(A) \times P(B|A) = 0.05 \times 0.99 = 0.0495 \quad (4.1)$$

Likewise, let us calculate the probability of registering a message by the filter as a spam:

$$P(B) = P(A \cap B) + P(A^c \cap B) = 0.05 \times 0.99 + 0.95 \times 0.10 = 0.1445 \tag{4.2}$$

And now, perhaps the most interesting case. Suppose that the filter registers something. What is the probability that it is a spam?

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} = \frac{P(A) \times P(B|A)}{P(B)} \\ &= \frac{0.05 \times 0.99}{0.1445} = 0.34 \end{aligned} \tag{4.3}$$

Even though we have a very good spam filter (99% reliability), the probability of a spam when flagged by the filter is somehow low.

The answer to the problem above is an example of Bayesian inference that has the following structure of reasoning:

1) we first specify our inital beliefs about each scenario of a problem at hand $P(A_i)$ (e.g., the probability of a spam message being received),
2) we consult our model of the world $P(B|A_i)$ which describes the probability of an event B under each scenario A_i (e.g., the probability of a filter to register a spam when the message is indeed a spam)

$$A_i \overset{\text{model}}{\underset{P(B|A_i)}{\longrightarrow}} B, \tag{4.4}$$

3) then, if we obsere that the event B actually occurs, we revise our inial beliefs about the probability of each scenario A_i. In other words, knowing that B occurred, we are willing to say which scenario is more or less likely:

$$B \overset{\text{inference}}{\underset{P(A_i|B)}{\longrightarrow}} A_i. \tag{4.5}$$

To ensure that the reasoning is valid, we use the Bayes' rule.

Bayes' rule

Let $A_1, A_2, \ldots, A_n$ be disjoint events that form a partition of the sample space, and assume that $P(A_i) > 0$, for all i. Then, for any event B such that $P(B) > 0$, we have:

$$\begin{aligned} P(A_i|B) &= \frac{P(A_i)P(B|A_i)}{P(B)} \\ &= \frac{P(A_i)P(B|A_i)}{P(A_1)P(B|A_2) + \ldots + P(A_n)P(B|A_n)} \end{aligned}$$

Total Probability Theorem

Let $A_1, \ldots, A_n$ be disjoint events that form a partition of the sample space (each possible outcome is included in one and only one of the events $A_1, \ldots, A_n$) and assume that $P(A_i) > 0$, for all $i = 1, \ldots, n$. Then, for any event B, we have:

$$\begin{aligned} P(B) &= P(A_1 \cap B) + \cdots + P(A_n \cap B) \\ &= P(A_1)P(B|A_1) + \cdots + P(A_n)P(B|A_n). \end{aligned}$$

4.2 Bayesian data analysis

You may now wonder: how to apply Bayes' rule to data analysis? Let us rewrite the expression 1 in terms of data and parameters:

$$\underbrace{p(\Theta|y)}_{posterior} = \left(\underbrace{p(y|\Theta)}_{likelihood} \times \underbrace{p(\Theta)}_{prior} \right) / \underbrace{p(y)}_{evidence} \tag{4.6}$$

Given a vector of data y, Bayes' rule allows us to work out the posterior distributions of the parameters of interest, which we can represent as the vector of parameters Θ. What is different here from the previous sections is that Bayes' rule is written in terms of probability distributions $p(\cdot)$ and not discrete events. More specifically:

- $p(\Theta)$ is the probability distribution of the parameter(s) of our model that reflects our knowledge before we collect data. We either know or make plausible assumptions about the values of these paramters. Crucially, those assumptions are embodied in the type of a probability distribution we believe reflects the structure of our problem at hand.
- $p(y|\Theta)$ is the likelihood. Having collected data, y, we can now examine the probability of having obtained a particular outcome in light of the prior values of the parameters Θ.[4]
- $p(\Theta|y)$ is the posterior probability of the parameters, Θ, after seeing the data that is the result of the application of Bayes' rule.
- $p(y)$ represents the overall probability of the data, irrespective of the values of the parameters. For now all we need to know is that it serves as a normalizing factor that assures $p(\Theta|y)$ is scaled to the range 0-1 to be a valid probability distribution.

Together, to arrive at updated knowledge about the parameter values of our model representing the problem at hand, we combine our prior knowledge about the parameters with the obtained data. T see how this works in practice, we refer to our example from the previous chapter – flipping a coin two times.

4.2.1 Likelihood

What is the likelihood function of the generative process underlying the outcome of a double coin flip? Last time, we established that the outcome of our experiment follows the binomial probability distribution. That is, the number of heads k in

[4]It might appear confusing to see that the probability $P(y|\Theta)$ is called the "likelihood," whereas in the previous chapter we used "likelihood" to refer to $\mathcal{L}(\Theta|y)$. Those quantities are the same and differ only in what is considered to be *given* and what is considered to be *variable*.

n tosses of a coin are described with the binomial likelihood function:

$$p(K = k|n, \theta) = \binom{n}{k} \theta^k (1 - \theta)^{n-k} \tag{4.7}$$

In the experiment that we carry out, we obtain two heads. Hence, the only variable in the equation that we do not know is θ:

$$p(K = 2|n = 2, \theta) = \binom{2}{2} \theta^2 (1 - \theta)^{2-2} \tag{4.8}$$

The above function is now a continuous function of the θ, values of which range from 0 to 1 as depicted in the Figure 4.4. Our main goal is to find out, using the Bayes' rule, the posterior distribution of θ given the observed data: $p(\theta|n, k)$. To this end, we need to specify our prior beliefs about the values of θ.

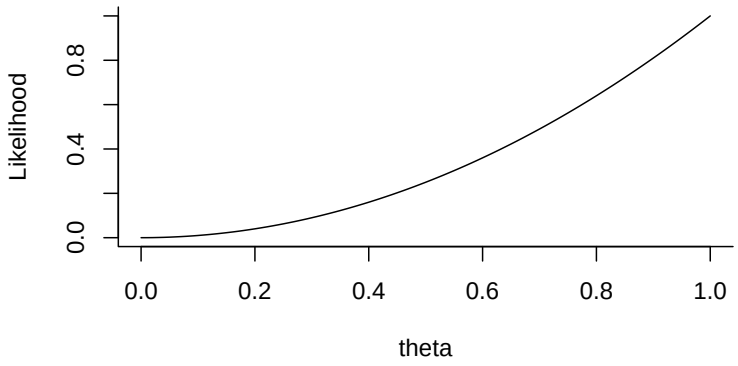


Figure 4.4: The binomial likelihood function when two flips of a coin yield two heads.

4.2.2 Priors

To select a prior for θ in the binomial distribution, we must assume that θ is a random variable with a probability distribution whose support lies within the interval $[0, 1]$, the range over

which θ can vary. The *beta distribution*, which is a probability density function for a continuous random variable, is commonly used as prior for parameters representing probabilities.

$$p(\theta|a,b) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1} \tag{4.9}$$

The term $B(a,b)$ is a normalizing constant to ensure that the area under the curve sums to one. We will ignore it for now. The beta distribution's parameters a and b can be interpreted as expressing our prior beliefs about the probability of success where a represents the number of "successes" (i.e., number of heads), and b the number of "failures" (i.e., number of tails). Possible shapes of the beta distribution are illustrated in the Figure 4.5.

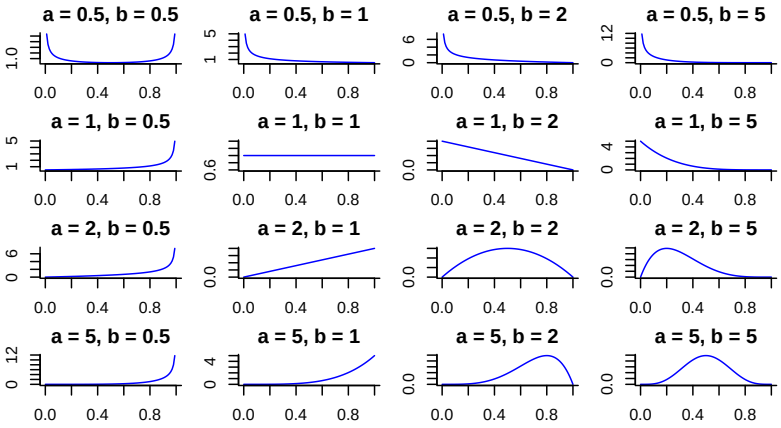


Figure 4.5: Examples of beta distributions.

What values of a and b shall we choose? Let's say we have no idea how a coin may behave and, hence, choose $a = 1$ and $b = 1$. We are therefore ignorant:

$$p(\theta) = \frac{1}{B(1,1)}\theta^{1-1}(1-\theta)^{1-1} = \frac{1}{B(1,1)} = 1 \tag{4.10}$$

4.2.3 Posterior

Having specified the likelihood and the prior, we will now use Bayes' rule to calculate $p(\theta|n,k)$. To this end, we simply replace the likelihood and the prior we defined above:

$$p(\theta|k=2,n=2) = \frac{\left[\binom{2}{2}\theta^2(1-\theta)^{2-2}\right] \times \left[\frac{1}{B(1,1)}\theta^{1-1}(1-\theta)^{1-1}\right]}{p(k=2)} \tag{4.11}$$

If we ignore the constant terms, we can simplify and obtain the equation:

$$p(\theta|k=2,n=2) \propto \theta^2(1-\theta)^{2-2} \times \theta^{1-1}(1-\theta)^{1-1} \tag{4.12}$$

Resolving the right-hand side now simply involves adding up the exponents! In this example, computing the posterior really does boil down to this simple addition operation on the exponents.

$$p(\theta|k=2,n=2) \propto \theta^2(1-\theta)^0 = \theta^2 \tag{4.13}$$

In our special case, when the prior was 1, we see that the posterior is proportional to the likelihood (Figure 4.6). But we may ask what if we had chosen a different prior.

Given some data and a likelihood function, the tighter the prior, the greater the extent to which the posterior orients itself towards the prior (Figure 4.7). In general, we can say the following about the likelihood-prior-posterior relationship:

- The posterior distribution of a parameter is a compromise between the prior and the likelihood.
- For a given set of data, the greater the certainty in the prior, the more heavily will the posterior be influenced by the prior mean.
- Conversely, for a given set of data, the greater the uncertainty in the prior, the more heavily will the posterior be influenced by the likelihood.

Simply put: the posterior distribution is a compromise between the prior and the likelihood.

The fact that the posterior distribution is a compromise between the prior and the likelihood has important implications. Whenever we do a Bayesian analysis, it is good practice to check whether the parameter of interest is sensitive to the prior specification. Such a robustness check is called a *sensitivity analysis*.

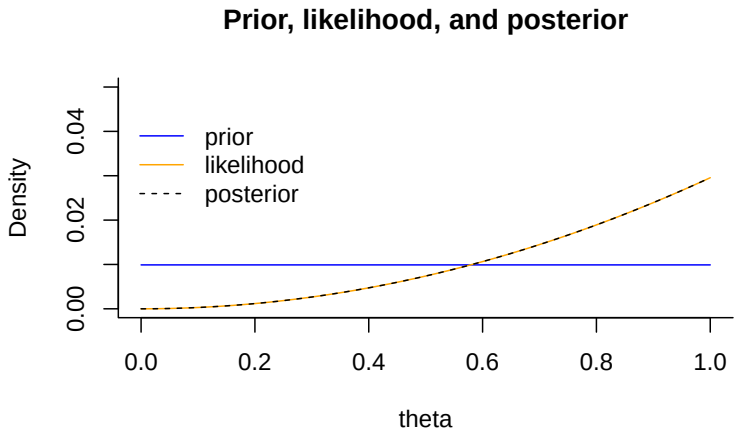


Figure 4.6: Prior, likelihood, and posterior.

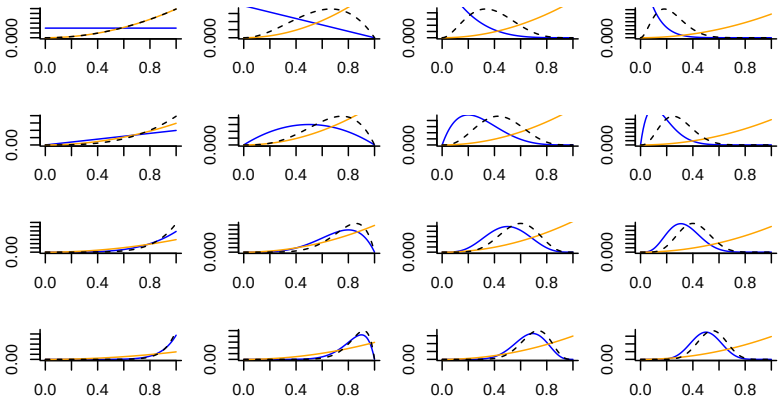


Figure 4.7: The posterior distribution is a compromise between the prior and the likelihood.

4.3 Accumulating evidence over time

Bayes' rule provides a powerful framework for updating our beliefs as new evidence are collected. In the context of data analysis, we've shown how to apply this rule effectively. By adjusting the prior probabilities of model parameters with the likelihood of new data, we've been able to reduce uncertainty and refine our understanding of the most likely parameter values. This approach is particularly valuable in situations where we have limited information or when we need to continuously adjust our predictions as new observations are made. Referring to our coin tossing example, we can now ask: what would happen if we had collected more coin flips?

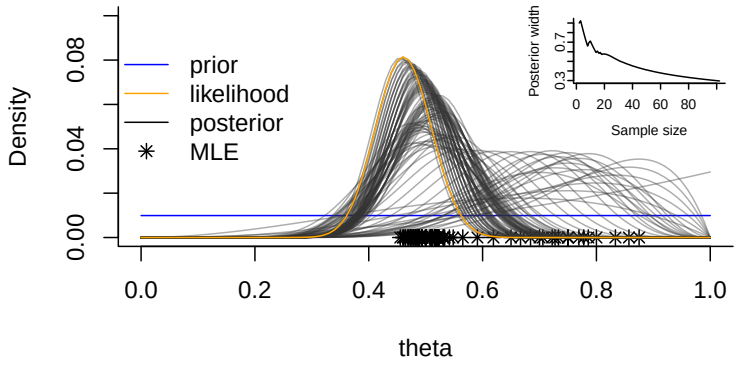


Figure 4.8: Narrowing the posterior with each data point.

Let us simulate. We are assuming flat prior as before $a = 1$ and $b = 1$. Our starting point is the same data $k = 2$ heads in $n = 2$ tosses. But we now continue tossing the coin until $n = 100$ trials. The results are illustrated in the Figure 4.8. As we incrementally gain information, the posterior becomes narrower, essentially reflecting our increased confidence in the best parameter values for predicting the data. This avantage is still there even when our prior we start with is biased. This scenario is illustrated in the Figure 4.9. Even if we start with a biased prior, over time the data (evidence) start to dominate,

and eventually the influence of our initial prior beliefs is negligible.

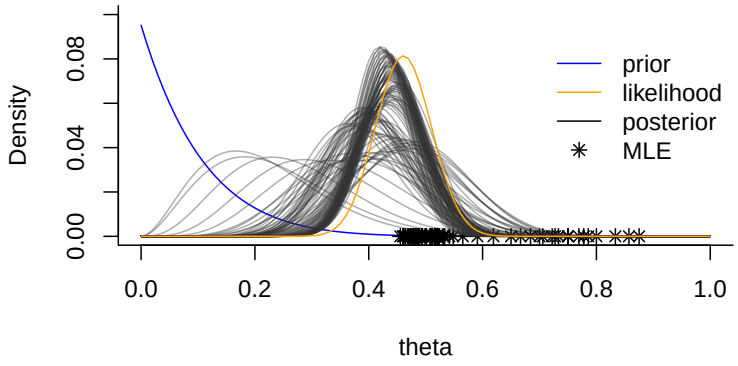


Figure 4.9: Narrowing the posterior with each data point when the inital prior is biased.

This toy example illustrates an important point that has great practical importance. One can use information from previous studies to derive a prior that then can be used for the next study. This approach allows us to build on the information available from previous work and accumuate knolwedge over time.

Warning

Using the posterior from one study as a prior for another study is a great advantage in Bayesian statistics, but it comes with several potential risks and concerns. For example:

- If the two studies are based on different models or have different underlying assumptions, the posterior from the first study may not accurately reflect the relevant information for the second study.
- If the posterior from the first study contains errors or biases, those errors can propagate and affect the

second study. If the first study was not properly designed, or had flaws in data collection or analysis, these issues will carry over to the second study.

- Ideally, priors should be informed not only by data but by a solid understanding of the theory and context. Simply using a posterior from another study may ignore important theoretical considerations or lead to a misinterpretation of the evidence, especially if the second study aims to test hypotheses that diverge from the first.
- Relying on the posterior from one study as a prior might limit the exploration of other potential priors. This can be problematic if there's uncertainty or if the previous study's posterior was based on assumptions that are not robust across different contexts.

4.4 Summarizing the posterior

Once we arrive at the posterior distribution, the next crucial step is to summarize and interpret it. Exactly how it is summarized depends upon our objective. Essentially, we can divide the ways to summarise the pestorior into three classes of objects:

- to specify the intervals of defined boundaries,
- to specify the intervals of defined probability mass, and
- to specify its point estimates.

To better understand the differences, let us look at a few examples.

To specify the interval of defined boundaries, we ask how much posterior probability lies between, for example, $\theta = 0.5$ and $\theta = 0.75$.

```r
# To calculate how much posterior probability lies between 0.5 and 0.75
# for a posterior distribution modeled by a Beta distribution,
# we can use the cumulative distribution function (CDF) of the Beta distribution.
# The CDF will give the probability that a random variable
# from that distribution is less than or equal to a specific value.


# Assume the data from the last simulation
shape1 = a + k
shape2 = b + n - k
p_075 <- pbeta(0.75, shape1, shape2)
p_050 <- pbeta(0.50, shape1, shape2)

prob = p_075 - p_050
print(prob)
```

```
[1] 0.05389114
```

Likewise, to specify the intervals of defined mass, we ask, for example what the boundary of the upper 20% posterior probability is.

```r
# To find the boundary of the upper 20% of the posterior probability for a Beta distribution,
# we're looking for the 80th percentile of the distribution.
# In R we can use qbeta function for this purpose

lb <- qbeta(0.80, shape1, shape2)
ub <- qbeta(1, shape1, shape2)
print(c(lb, ub))
```

```
[1] 0.4638061 1.0000000
```

Intervals of this sort, which assign equal probability mass to each tail, are very common in the scientific literature. We can call them **percentile intervals**. In contrast, the **highest posterior density interval (HPDI)** is the narrowest interval containing the specified probability mass. For example, we may ask: what is the 89% HPDI for the posterior distribution? To

ask about the 89% HPDI, we're referring to the interval where the central 89% of the posterior probability lies, with the remaining 11% spread out outside of it. The HPDI is typically used to summarize the most likely values of a parameter.

```
lb <- qbeta(0.055, shape1, shape2)
ub <- qbeta(0.945, shape1, shape2)
print(c(lb, ub))
```

```
[1] 0.3514191 0.4995254
```

> **i** Region of Practical Equivalence
>
> The description of the posterior distribution allows to draw conclusions from Bayesian analyses. We have just demonstrated that the posterior distribution is often summarized by using intervals. A particularly relevant "interval" for hypothesis testing is the **Region of Practical Equivalence** (ROPE). ROPE defines a small interval around a null value (e.g., 0) within which differences are considered practically negligible (Figure 4.10). In other words, the null hypothesis is re-defined from a point-null to a range of values considered negligible or too small to be of any practical relevance (Kruschke, 2014; Lakens, 2017; Lakens et al., 2018), usually spread equally around the null point value (e.g., [−0.1; 0.1]). The idea behind ROPE is that an effect is almost never exactly zero, but instead can be very tiny, with no practical relevance.[5] This perspective unites *significance testing* with the focus on effect size. The typical recommendation is to check how much of the posterior density falls inside (or outside) the ROPE. If a posterior distribution falls mostly within the ROPE, the effect is considered practically equivalent to the null. Likewise, if the posterior falls outside the ROPE, then it is considered as practically significant. The Figure 4.11 illustrates potential error in decision making depending on the true value (effect size) and sample size.

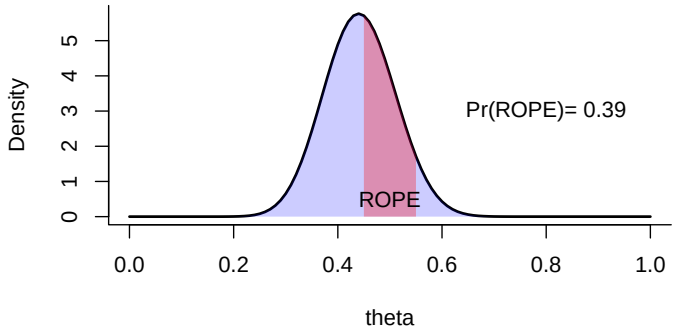


Figure 4.10: The percentage of the posterior distribution falling inside ROPE [.45, .55].

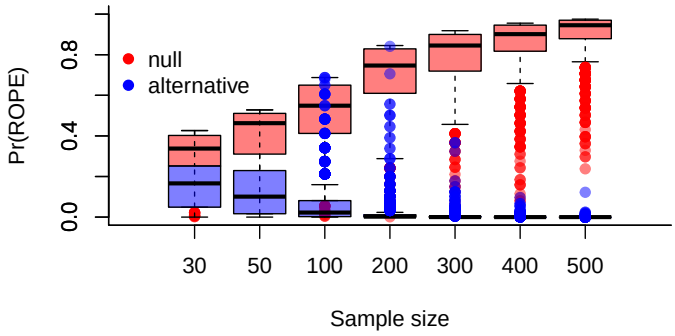


Figure 4.11: Proportion of the posterior falling inside ROPE [.45, .55] when the null (0.5) and alternative (0.65) parameter is true.

Probability of making a wrong decision when null is true at the level of Pr(ROPE) is 5% for each sample size:

30	50	100	200	300	400	500
0.052	0.015	0.009	0.001	0	0.001	0

And converesely, the probability of correct decision when the null is true at the level of Pr(ROPE) is 95% for each sample size:

30	50	100	200	300	400	500
0	0	0	0	0	0.201	0.45

In other words, chasing the null is hard, even when using Bayesian statistics.

> **!** Important
>
> It is important to note that these calculations are done for a one-paramter model. In a more complex models, that we typically use, these values may look dramatically different. It is important to run such simulations specific for the planned design.

One of the core features of the Bayesian data anlysis is that paramter estimates are expressed in the entire posterior distribution, which is not a single number, but instead a function that maps each unique parameter value onto a plausibility value. So in fact, providing a point estimate is not the priority in Bayesian analysis because it discards information. Nevertheless, if you must produce a single point to summarize the posterior, it is very common to report the parameter value with highest posterior probability – the **maximum a posteriori (MAP) estimate.** To report the MAP estimate for a

[5] Researchers often incorrectly use credible intervals for null hypothesis testing. Often, researchers test whether a value of interest (usually 0) is included in the 95% credible interval. If it is, then the null hypothesis that the effect is zero is accepted; and if zero is outside the interval, then the null is rejected. The reasoning is that if zero is outside the interval, it must have a low probability density. This is true, but it's meaningless: any point value have a probability mass of exactly zero for a contionous probability distribution.

posterior distribution, we obtain the mode of the posterior distribution, which is the value of the parameter that maximizes the posterior probability.

```
# For a Beta distribution, the MAP estimate can be calculated
# using the formula for the mode:

MAP = (shape1 - 1) / (shape1 + shape2 - 2)
print(MAP)
```

```
[1] 0.44
```

Alternative, one can simply report the mean:

```
print(shape1 / (shape1 + shape2))
```

```
[1] 0.4423077
```

Which estimate is considered optimal can be approached in principled way by using a **loss function.** A loss function helps determine which single estimate from the posterior is the "best" one. It quantifies the cost of choosing a particular estimate when the true value might be different. Different loss functions lead to different optimal estimates. The two most common examples of a loss function are the absolute loss $abs(d-p)$, which leads to the median as the point estimate, and the quadratic loss $(d-p)^2$, which leads to the posterior mean as the point estimate (Figure 4.12).

4.5 The Savage–Dickey (pointwise) density ratio

The Bayes factor is a measure of relative evidence, the comparison of the predictive performance of one model against another one. This comparison is a ratio of marginal likelihoods:

$$BF_{10} = \frac{p(y|M_1)}{p(y|M_0)} \tag{4.14}$$

> **i** Probability vs density
>
> For a continuous distribution (e.g., normal distribution), the probability of observing any single point is zero, but the probability density at that point is not necessarily zero. The Savage-Dickey density ratio is meaningful because it compares densities, not probabilities. Likewise, in the context of a likelihood function, we compare the densities at two different parameter values and correctly obtain the likeli-

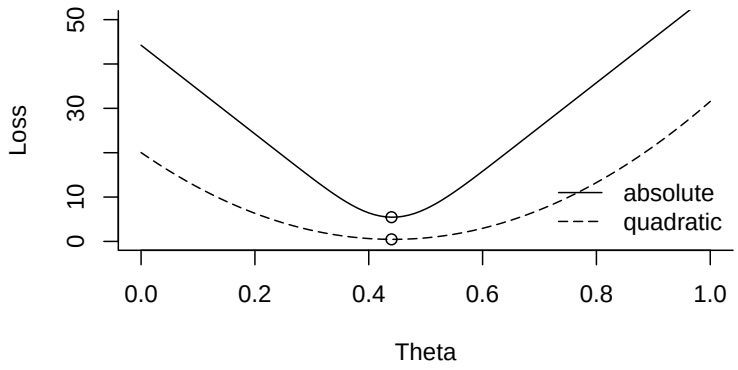


Figure 4.12: Effect of different loss functions on point estimate of the posterior distribution.

BF_{10} indicates the extent to which the data are more likely under M_1 over M_0, or in other words, the relative evidence that we have for M_1 over M_0. In other words, Bayes Factor quantifies on a continuous scale the change in belief that the data bring about for the two models under consideration. The interpretation of the Bayes Factor values are depicted in the Table 4.3. For example, values close 1 indicate that the evidence is inconclusive. This model comparison does not depend on a specific parameter value. Instead, all possible prior parameter values are taken into account simultaneously.

Table 4.3: Jeffreys' scale for Bayes Factors.

Bayes Factor BF_{10}	Interpretation
1 - 3	Weak evidence for M_1
3 - 10	Moderate evidence for M_1
10 - 30	Strong evidence for M_1
30 - 100	Very strong evidence for M_1
$>$ **100**	Decisive evidence for M_1

The Bayes Factor provides a powerful framework for model comparison, but calculating the marginal likelihoods $p(y|M_1)$

and $p(y|M_0)$ can be computationally intensive, especially when dealing with complex models or large datasets. In some cases, however, we are interested in comparing models that are *nested*, where one model is a special case of another (i.e., when the null hypothesis is a special case of the alternative hypothesis). For such cases, the **Savage-Dickey method** offers an elegant and computationally efficient way to compute the Bayes Factor. Instead of integrating over the entire parameter space, the Savage-Dickey method leverages the **posterior-to-prior density ratio** at a specific parameter value, significantly simplifying the calculation when both the prior and posterior distributions are of the same family (e.g., Beta or normal distributions). This approach allows for an efficient evaluation of the Bayes Factor, particularly in the context of hypotheses testing or model comparison, without the need for extensive numerical integration.

The Savage-Dickey density ratio is calculated as:

$$BF_{10} = \frac{p(\theta_0|y)}{p(\theta_0)} \tag{4.15}$$

where $p(\theta_0)$ is prior density at θ_0 and $p(\theta_0|y)$ is the posterior density at θ_0 given the data y.

Note that this is in contrast with the likelihood ratio we talked about in the previous chapter:

$$LR = \frac{p(y|\theta_1)}{p(y|\theta_0)} \tag{4.16}$$

where $p(y|\theta_1)$ is the likelihood of the data at θ_1 and $p(y|\theta_0)$ is the likelihood of the data at θ_0.

Their direct comparison is summarised in the Table 4.4. Both express how much more the data favor one hypothesis over another. But Bayes Factor depends on the prior distribution while the likelihood ratio is purely data-driven.

Table 4.4: Key differences between Bayes Factor (Savage-Dickey) vs Likelihood Ratio (Royall).

Feature	Bayes Factor (Savage-Dickey)	Likelihood Ratio (Royall)
Comparison	Prior vs. posterior density at θ_0	Likelihood density at θ_0 vs θ_1
Interpretation	How much the data update belief at θ_0	How much more likely one parameter value is compared to another
Incorporates Priors?	Yes	No (purely data-driven)

i Interpreting Bayes Factor with Savage-Dickey

The Bayes Factor is a ratio used to quantify the relative support for one hypothesis (or parameter value) compared to another based on the observed data. This is typically done as a ratio of marginal likelihoods of two competing statistical models.
In contrast, when testing a *specific parameter value* θ_0, the standard Bayes Factor is defined as the ratio of the likelihood of the data given the parameter value to the marginal likelihood (also called evidence). Similarly, the Savage-Dickey density ratio is a special case where it simplifies to the *posterior-to-prior* ratio at a specific parameter value. The key assumption in the method is that it holds when null hypothesis is nested under the alternative hypothesis, i.e., the null can be obtained from the alternative by setting θ equal to θ_0. Why is it so?
The Bayes' rule formula in the context of data analysis and a specific parameter value of the model is:

$$p(\theta_0|y) = \frac{p(y|\theta_0) \times p(\theta_0)}{p(y)}$$

If we re-write the equation but emphasise that the equation gives a relation between the unconditional probability of a parameter $p(\theta_0)$ and the probability of the same pa-

rameter conditioned on data, $p(\theta_0|y)$, we then obtain:

$$p(\theta_0|y) = \frac{p(y|\theta_0)}{p(y)} \times p(\theta_0)$$

We further re-arrange the terms so that they match the formula of the Savage-Dickey method, and obtain:

$$\underbrace{\frac{p(y|\theta_0)}{p(y)}}_{\text{likelihood-to-marginal likelihood}} = \underbrace{\frac{p(\theta_0|y)}{p(\theta_0)}}_{\text{posterior-to-prior}}$$

The Savage-Dickey posterior-to-prior ratio is mathematically equivalent to the standard Bayes Factor (the likelihood-to-marginal likelihood ratio) when evaluated at a specific point hypothesis. The key consequence or interpretation of this equivalence is that the Bayes Factor can be understood in two ways:

- The interpretation of a shift in belief: As a ratio of posterior to prior probability at a specific value of the parameter, i.e, as a way of comparing how likely a parameter value is given the data (posterior) relative to how likely it was before seeing the data (prior).
- The interpretation of a likelihood-based comparison: As a ratio of likelihood to marginal likelihood (evidence), i.e., as a way of assessing how much more likely the data is under a specific hypothesis (parameter value) compared to the total evidence, i.e., when considering all possible hypotheses (all parameter values). In other words, the marginal likelihood tells us how likely the data is overall, considering all possible parameter values weighted by how probable each value is according to the prior.

```r
set.seed(1234)
data = rbinom(100, 1, prob=.5)

a = 2
b = 5
k = sum(data)
n = length(data)

# Fine resolution for calculation
thetas <- seq(0, 1, length.out = 10000)
prior = dbeta(thetas, shape1 = a, shape2 = b)
lik = dbinom(x=k, size=n, prob=thetas)
posterior = dbeta(thetas, shape1 = a + k, shape2 = b + n - k)
idx <- which.min(abs(thetas - 0.5))  # Find closest value to 0.5
marginal_lik = integrate(function(theta) dbinom(k, size = n, prob = theta) * dbeta(theta, a,
                         lower = 0, upper = 1, rel.tol = 1e-8)$value

BF1 = posterior[idx] / prior[idx]
BF2 = lik[idx] / marginal_lik
```

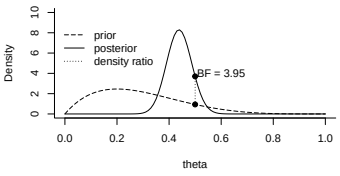


Figure 4.13: Bayes Factor based on Savage-Dickey posterior-to-prior density ratio.

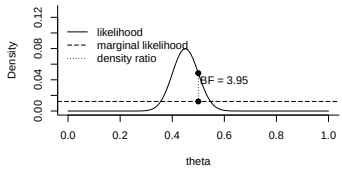


Figure 4.14: Bayes Factor based on likelihood-to-marginal likelihood ratio.

It is important to remember that this holds only when some assumptions are met. For example:

- the hypothesis put to the test is expressed as a spe-

cific parameter value;

- the prior and posterior should both be of the same family;
- the prior, likelihood, and posterior should be smooth at the parameter of interest;
- for numerical methods, grid resolution and integration should be sufficiently precise.

4.6 Summary

Bayesian data analysis is a statistical approach that applies Bayes' rule to update beliefs[6] based on new data. It treats parameters as probability distributions that are refined by evidence. This approach allows for incorporating prior knowledge, quantifying uncertainty, and making probabilistic predictions.

i The basic tenets of Bayesian inference

- Our degree of belief is quantified by probability – we can express varying levels of uncertainty that can vary between 0 and 1.
- Our prior beliefs are updated by evidence according to the Bayes' rule which ensures coherence and adherence of our propositions with basic logical axioms.
- We treat the parameters of our probabilistic models as random variables. This allows us to capture the uncertainty about their true values, resulting in a posterior distribution that reflects both prior beliefs and the observed data.

[6] The notion of belief plays a central role in Bayesian statistics, but these beliefs need not be the literal beliefs of any person. In Bayesian inference, belief refers to the **degree of confidence** in a particular hypothesis, represented mathematically as a **probability distribution**.

4.7 Exercises

1. Pick a value of a θ parameter. Simulate some data by repeatedly tossing a coin 100 times. Pick a prior using beta distribution. Now, calculate the posterior distribution when a) one assumes all data are collected at once, b) when one assumes data are collected sequentially. Do they differ? Explain why.
2. Using the simulation from exercise 1, quantify your certainty that the simulated data indeed come from the distribution you assumed when they were generated. There are multiple ways to approach this problem.
3. Consider how the Bayes Factor interpretation may change when the prior is flat (e.g., Beta(1,1)).

4.8 Relation to Open Science

The bayesian paradigm is highly relevant for **open science** in several key ways:

1. Transparency in Modeling: Bayesian statistics encourages the explicit specification of prior beliefs, likelihood functions, and the overall model. This makes it easier for researchers to document their assumptions, which is important for transparency. In open science, sharing these models allows others to critique, reproduce, and build on the research.
2. Flexibility and Updating Models: One of the key features of Bayesian statistics is the ability to update beliefs as new data becomes available. This iterative approach reflects the open science goal of adapting to new evidence in a flexible, transparent manner. In an open science context, researchers can share their ongoing updates to models and hypotheses, allowing others to contribute or modify them.
3. Better Communication of Uncertainty: Bayesian methods explicitly quantify uncertainty in the form of probability

distributions. This focus on uncertainty is important for open science because it helps communicate the limitations of findings and fosters more honest, nuanced interpretations of research results.

Bayesian data analysis strengthens open science by ensuring transparency, improving reproducibility, better quantifying uncertainty, and reducing biases in inference. As open science continues to evolve, Bayesian approaches provide a rigorous and flexible framework for making scientific knowledge cumulative and accessible.

4.9 Recommendations

If you are interested in learning about probability theory, the open course at MIT by Prof. J. Tsitsiklis is absolute gold Introduction to probability.

4.10 References

Bertsekas D.P. & Tsitsiklis, J. N. (2000). Introduction to Probability, MIT Cambridge, Massachusetts.

Lambert, B. (2018). A Student's Guide to Bayesian Statistics. SAGE Publications Ltd., London, UK.

McElreath, R. (2016). Statistical Rethinking: A Bayesian Course with Examples in R and Stan (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315372495

Morey, R. D. (2018). Statistical Inference. In Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience (pp. 1–42). Wiley. https://doi.org/10.1002/9781119170174.epcn504

Nicenboim, B., Schad, D., & Vasishth, S. (2021). An introduction to Bayesian data analysis for cognitive science. Under contract with Chapman and Hall/CRC statistics in the social and behavioral sciences series. https://bruno.nicenboim.me/bayescogsci/index.html

5 The frequentist paradigm

In Bayesian inference, we quantify uncertainty about parameters using probability distributions, combining prior beliefs with data to update our knowledge. This approach treats parameters as random variables with probability distributions reflecting our uncertainty.

Frequentist inference, in contrast, takes a different perspective. It does not assign probabilities to parameters but instead considers them as fixed, unknown quantities. Uncertainty comes from the randomness of data, not the parameters themselves. In this view, probability represents long-run frequencies of events, and inference is based on procedures, which control error rates over repeated (hypothetical) samples (experiments).

Interestingly, the likelihood function is often the only thing a Bayesian and frequentist have in common. Yet, they have fundamentally different interpretation of what is meant by probability.

5.1 Notion of probability – foundation and interpretation

Intuitevely, we all have a good sense of what the notion of probability is. But it may be interpreted in different ways. For example, consider the following sentence from a medical leaflet: "Approximately 1 in 1,000 people may experience a severe allergic reaction (anaphylaxis) while taking this medication." Or this one "Our analytic team says that there's a 70% probability that the new marketing strategy will increase sales." These sentences reflect two distinct ways of how to think about probability.

5.1.1 Philosophical aspects

1) **Epistemic probability** (De Morgan, Boole, Carnap, Savage). Probabilities represent *degrees of belief* about events (the occurrence of the state of affairs) or hypotheses. Example: we may strongly believe that a coin is fair, i.e., we belive there is 50% chance that it lands heads up – this is our belief, not an inherent property of the coin. Our belief can change with new data (e.g., if we toss the coin many times and see heads more often than expected, we may revise our belief about its fairness).

2) **Physical probability** (Venn, Maxwell). Probabilities represent *relative frequencies* in repeated experiments or *tendencies* in physical systems. Example: A coin landing heads up has a 50% probability if, in a long sequence of tosses, it lands heads about half the time. Alternatively, probability can be seen as an inherent tendency in the system (e.g., the physical properties of the coin and how it is flipped determine its fairness). This interpretation does not apply to hypotheses because a hypothesis is not a repeatable (physical) event.

Epistemic probability is about what we *believe* (subjective, can change with new information) while physical probability is about what *actually* happens in repeated trials (objective, based on observed frequencies or inherent tendencies). We will return later to the issue of subjectivity / objectivity.

5.1.2 Consequences for data analysis

These two interpretations of probability have practical consequences for data analysis.

Bayesian inference (epistemic probability). We have parameters Θ and observations y which are both treated as random variables reflecting our prior beliefs $p(\Theta)$ and likelihood $p(y|\Theta)$, respectively. Our goal is to update our beliefs using Bayes' rule, i.e., to find $p(\Theta|y)$. That's all!

Frequentist inference (physical probability). The key idea is that some unknown quantities – such as the mass of

an electron – are not random in reality. If they are not random, then we cannot assign them a probability distribution. Instead, they are **fixed but unknown.** More specifically, the reasoning goes as follows.

- There is a constant true but unknown value θ (e.g., the mass of an electron).
- It happens so that some observations x are obtained from a probability distribution of a random variable X such that $p_X(x;\theta)$. This is an ordinary probability distribution (not a conditional probability). It is just affected by the value of θ, which is constant. As a result, while θ is fixed (and unknown), our observations x are random. In other words, when we perform an experiment, one time we get one set of values x_1 and the other time another set of values x_2, and so on…
- Given this setup, we take the specific set of data x and we process them through a function $g(x)$ which is called an *estimator.* The output of that function is an *estimate*, $\hat{\theta}$. Note that once we have some specific observations x, we have a specific estimate $\hat{\theta}$. While the process of mapping a specific x to a specific $\hat{\theta}$ is deterministic, estimates are realizations of the random variable $\hat{\Theta}$ because they depend on the outcomes of the random variable X. In other words, since X is a random variable (data are random), then $\hat{\Theta}$ must also be a random variable (our estimates are random).

$$\theta \xrightarrow[\text{rv. X}]{p_X(x;\theta)} x \xrightarrow[\text{rv. } \hat{\Theta}]{g(x)} \hat{\theta} \tag{5.1}$$

> **Note**
>
> Coming up with good estimators is a bit of an art. Good estimators aim at the error $\hat{\theta} - \theta$ to be "small". For example, an unbiased estimator of the population variance σ^2 is sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Assuming that $X_1, ..., X_n$ are independent and identically distributed according to a normal distribution. In practice, this means

that estimators can be designed in many ways – they need to adhere to certain properties but to some extent they are arbitrary. While Bayes' rule is a single unambiguous method to draw inference, classical statistics allows for some ambiguity in this respect.

5.2 The sampling distribution

To better understand the structure of reasoning in the frequentist paradigm, let us consider our toy example of a coin tossing experiment. So, we start with the assumption that there is a true but unknown θ parameter that is constant. This is the true value of a coin "fairness", i.e., its tendency to lend heads up. We assume that the coin tosses are generated from a Binomial probability distribution such that $k \sim Binomial(n = n, \theta = 0.5)$. In other words, we assume that the data of the coin tossing experiment are random and come from a specific probability distribution, which depends on the *theta* parameter. How do we proceed to infer the value of θ.

The recipe of the frequentist inference tells us that we first think of hypothetical data which come from **repeated sampling**. In other words, we repeat the hypothetical process of data collection from the same generative model many many many times. Let us simulate some data through repeated sampling of an experiment with the sample size of $n = 50$.

Tip

Frequentist inference in practice means that there are many candidate models that generated the data x, one for each possible value of θ. The goal is to infer the best possible model. For example, we choose between two candidate models: $\theta = 0.5$ vs. $\theta = 0.75$. Or we choose between $\theta = 0.5$ vs. $\theta \neq 0.5$. Note that in the latter case, we do a comparison between a single model and all models that can take on any value of θ except 0.5.

```
set.seed(12434)

# generate data assuming exact repetition of an experiment
get_binom_data <- function(theta_true, n){
  dat <- rbinom(n, 1, prob=theta_true)
  return(dat)
}

n_sim <- 1e+5
theta_true <- 0.5
```

```r
n <- 50
data <- replicate(n_sim, get_binom_data(theta_true=theta_true, n=n), simplify=T)
```

These are the data obtained from an exemplary hypothethical experiment:

```r
print(data[,4])
```

```
 [1] 1 1 1 0 0 1 0 1 1 0 1 1 1 0 0 1 1 1 1 1 0 0 1 0 0 0 0 0 0 1 1 1 0 1 1 1 1 1
[39] 1 1 0 1 1 1 1 0 0 1 0 0
```

What may be a good description (i.e., statistic) of the data generated by each experiment? Perhaps its maximum likelihood? In case of a binomial model, the maximum likelihood estimator is $\hat{\theta} = \frac{k}{n}$. Hence, let us calculate the estimate $\hat{\theta}$ for every experiment.

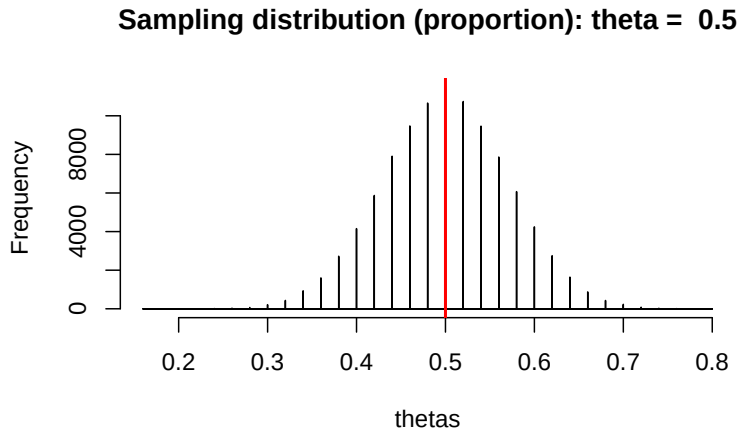


Figure 5.1: Sampling distribution of $\hat{\theta}$ (i.e, the proportion of heads).

By repeatedly taking samples and calculating a statistic (i.e., proportion), we have built a distribution, which describes the variability of that statistic across different samples. This distribution is called **the sampling distribution.** In other words,

we obtained a distribution of $\hat{\theta}$. Confusingly, the standard deviation of the sampling distribution is called **the standard error**, which helps us estimate how much a sample result might differ from the population parameter. In other words, the variability of a sample statistic across repeated samples can be quantified using the standard error, which decreases as sample size increases.

Second, we see that given our large number of samples 1e+5 the sample statistic tends to converge to the true population parameter θ. This is called **the law of large numbers.**

Third, we see that the obtained distribution resembles a normal distribution. This is quite cool – even though the data do not come from a normal distribution, the distribution of their means do! In fact, if we take enough repeated samples of a certain size, the sampling distribution of the sample mean tends to follow a normal distribution, regardless of the shape of the population distribution (assuming finite variance). This is called **the central limit theorem**.

In sum, our simulations reveal quite a powerful insight – sample statistics (e.g., sample mean, sample proportion) vary from sample to sample, but they follow a predictable pattern known as a sampling distribution.

! Central Limit Theorem under independence

If $\{X_1, X_2, \ldots, X_n\}$ are independent and identically distributed (i.i.d.) random variables, each representing an observation from a population with expected value μ and finite variance σ^2, and let $\overline{X}$ denote the sample mean:

$$\overline{X} \equiv \frac{X_1 + X_2 + \cdots + X_n}{n},$$

then as $n \to \infty$, by the Law of Large Numbers, the distribution of the random variable $\overline{X}$ approaches:

$$\overline{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right),$$

which means that the distribution of the sample mean $\overline{X}$ will increasingly resemble a normal distribution, regardless of the original population distribution!

In other words, with a sufficiently large sample, the average of that sample will follow a normal (bell-shaped) curve, even if the original data does not. The larger the sample size n, the closer the sample mean $\overline{X}$ is to the true population mean μ. The variability of the sample means (the standard deviation) decreases as n increases, shrinking at a rate of $\frac{\sigma}{\sqrt{n}}$.
To make comparisons across different data sets, we can "standardize" the sample mean by converting it into a new random variable Z:

$$Z \equiv \frac{\overline{X} - \mu}{\sigma/\sqrt{n}},$$

which follows a standard normal distribution:

$$Z \sim N(0, 1).$$

Why does Z have variance 1 even though the sample mean's variance shrinks?
Even though the variance of $\overline{X}$ decreases as n increases, we divide by its standard deviation. Standardizing means dividing by the square root of the variance, which cancels out the shrinking effect, ensuring that Z always has a variance of 1.
You may now ask how to standardize the sample mean $\overline{X}$ when we do not know the population variance σ^2. We estimate it using the sample variance:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

As a result, we obtain a standardized sample mean measured in units of the estimated standard error (using the sample standard deviation s) such that:

$$T = \frac{\overline{X} - \mu}{s/\sqrt{n}}.$$

Since the sample variance s^2 is a random variable itself, it introduces additional variability in the denominator. This

extra uncertainty causes the distribution of the standardized mean T to have heavier tails compared to the normal distribution. Therefore, our standardization is no longer exactly normal, but follows a Student's t-distribution with $\nu = n - 1$ degrees of freedom:

$$T \sim t(\nu).$$

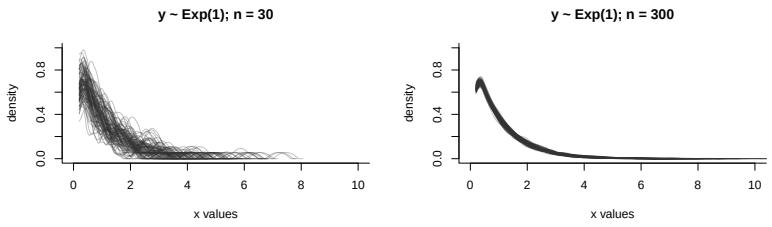
The t-distribution looks similar to the normal distribution but has heavier tails (more spread out), especially for small n. As $n \to \infty$, the t-distribution approaches the standard normal distribution.

5.2.1 The Law of Large Numbers and Central Limit Theorem

Let us explain the central limit theorem with simulations.

Let's assume that the hypothetical data come from a model with an exponential probability distribution, $X \sim Exp(\lambda = 1)$. First, let us sample many many experiments with the sample size $n = 30$ and $n = 3000$ (Figure 5.2).

```
set.seed(1234)
exp_dat1 <- replicate(n_sim, rexp(30, 1))
exp_dat2 <- replicate(n_sim, rexp(1000, 1))
```



(a) d (b) f

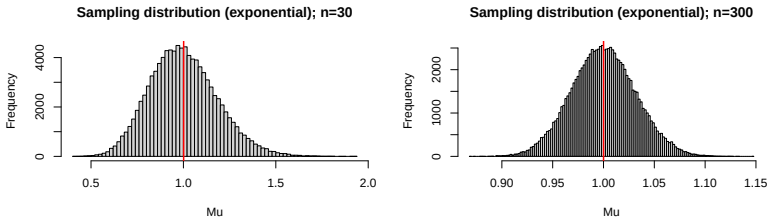
Figure 5.2: Sampling distribution.

And plot the distribution of their means (Figure 5.3).

```r
exp_m1 <-colMeans(exp_dat1)
exp_sd1 <- apply(exp_dat1, 2, function(x){sd(x)})
exp_m2 <-colMeans(exp_dat2)
exp_sd2 <- apply(exp_dat2, 2, function(x){sd(x)})

cat("N=30", "\n",
    "True mean: ", 1, "\n",
    "Obtained mean: ", round(mean(exp_m1), 3), "\n",
    "True sigma: ", round(sqrt(1/30), 3), "\n",
    "Obtained sigma: ", round(sd(exp_m1), 3), "\n",
    "N=1000", "\n",
    "True mean: ", 1, "\n",
    "Obtained mean: ", round(mean(exp_m2), 3), "\n",
    "True sigma: ", round(sqrt(1/1000), 3), "\n",
    "Obtained sigma: ", round(sd(exp_m2), 3)
    )
```

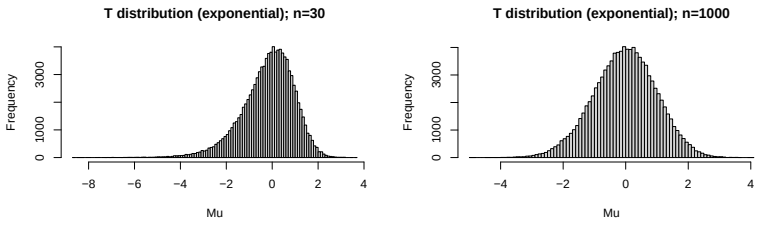
```
N=30
 True mean:  1
 Obtained mean:  1.001
 True sigma:  0.183
 Obtained sigma:  0.182
 N=1000
 True mean:  1
 Obtained mean:  1
 True sigma:  0.032
 Obtained sigma:  0.032
```



(a) d (b) f

Figure 5.3: Sampling distribution.

We see that indeed the sampling distibution of the means resembles a normal distribution $N\left(\mu, \frac{\sigma^2}{n}\right)$. Let us now standardize the means and plot their distribution (Figure 5.4).



(a) d (b) f

Figure 5.4: Sampling distribution.

```
v1 <- 30-1
v2 <- 1000-1
cat("N=30", "\n",
    "True mean: ", 1, "\n",
    "Obtained mean: ", round(mean(exp_t1), 3), "\n",
    "True sigma: ", round(sqrt(v1/(v1-2)), 3), "\n",
    "Obtained sigma: ", round(sd(exp_t1), 3), "\n",
    "N=1000", "\n",
    "True mean: ", 1, "\n",
    "Obtained mean: ", round(mean(exp_t2), 3), "\n",
    "True sigma: ", round(sqrt(v2/(v2-2)), 3), "\n",
    "Obtained sigma: ", round(sd(exp_t2), 3)
    )
```

```
N=30
 True mean:  1
 Obtained mean:  -0.19
 True sigma:  1.036
 Obtained sigma:  1.141
 N=1000
 True mean:  1
 Obtained mean:  -0.03
 True sigma:  1.001
 Obtained sigma:  1.006
```

These new distributions of the standardized means also resemble the theoretical standard normal distribution as long the sample size is large. We demonstarted that the assumptions hold true. But we are still faced with the problem of how to approximate the sampling distribution from a single study.

5.3 Single study

We have demonstrated that with repeated sampling we eventually converge on a true but unknown parameter value. Yet, we are still faced with the problem – how to infer the population parameter from a single sample, i.e., data from a single experiment. The solution is to approximate the sampling distribution from the observed data using estimators. In brief, we take the obtained data and process them through an estimator (function) to arrive at the best estimate of the population parameter.

To demonstrate the idea, let us revisit our coin tossing example. Yet, instead of using pre-defined estimators, let's try to see if we can arrive at the same results through simulations. A method used for this purpose is called **bootstrapping** - a resampling method used to estimate the distribution of a statistic (e.g., the mean) by repeatedly sampling with replacement from the original dataset.

The method works like this:

- Given a dataset of size n, create multiple new datasets (called bootstrap samples) by randomly sampling with replacement from the original dataset. Each bootstrap sample is also of size n.
- Compute the statistic of interest for each bootstrap sample.
- Repeat this process many times (typically 1,000 or more iterations).

To demonstrate the method, we can use the data from one sample generated from the previous code of the coin tossing experiment.

```r
id <- 4
data_obs <- data[, id]
n_size <- length(data_obs)
print(data_obs)
```

```
 [1] 1 1 1 0 0 1 0 1 1 0 1 1 1 0 0 1 1 1 1 1 0 0 1 0 0 0 0 0 0 1 1 1 0 1 1 1 1 1
[39] 1 1 0 1 1 1 1 0 0 1 0 0
```

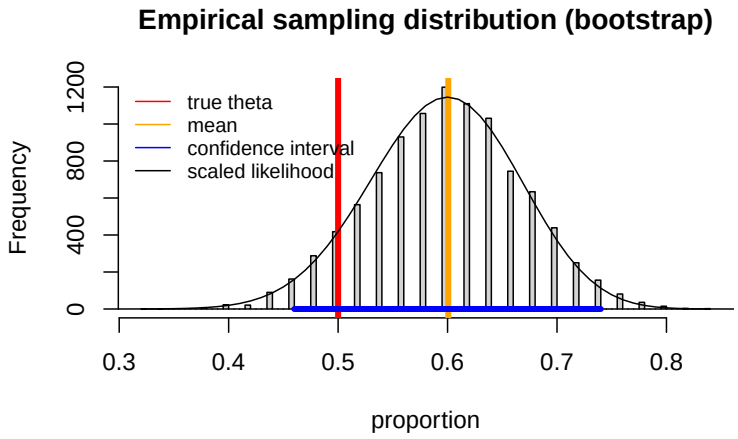


Figure 5.5: Sampling distribution through bootstrapping.

Indeed, we obtained a sampling distribution that is normal. To check if we did everything correct, let us compare the estimates obtained from bootstrapping vs estimators. The estimator of the mean and standard error from a binomial distribution is $\hat{\theta} = \frac{k}{n}$ and $SE(\hat{\theta}) = \sqrt{\frac{\hat{\theta} \times (1-\hat{\theta})}{n}}$, respectively.

```r
m <- sum(data_obs) / n_size
se <- sqrt((m*(1-m))/n_size)
ci <- c(m - 1.96 * se, m + 1.96 * se)
cat(paste(
  "Bootstrapped mean: ", prop_boot_m, "\n",
  "Estimator mean: ", sum(data_obs) / n_size, "\n",
  "Bootstrapped CI: ", prop_boot_CI[1], "\n",
  "Estimator CI: ", round(ci[1], 2), "\n",
```

```
  "Bootstrapped CI: ", prop_boot_CI[2], "\n",
  "Estimator CI: ", round(ci[2], 2)
)
)
```

```
Bootstrapped mean:  0.60066
 Estimator mean:  0.6
 Bootstrapped CI:  0.46
 Estimator CI:  0.46
 Bootstrapped CI:  0.74
 Estimator CI:  0.74
```

They both align. Instead of standard error, I used something called the **confidence interval**, which is approximately defined by the range $\hat{\theta} \pm SE * 1.96$. Why to define the confidence interval?

5.4 Confidence procedures and intervals

In general, (estimates) intervals are constructed to account for measurement or sampling uncertainty by yielding a range of values for a parameter instead of a single value (consider 1/8 likelihood interval or HPDI). Confidence interval (CI) is another kind of interval estimates. The meaning of the CI comes from confidence interval theory formalized by Jerzy Neyman. According to the theory (after Morey…), when a researcher is interested in estimating a parameter θ, they perform three steps:

- Collect relevant data.
- Compute two numbers – the smaller of which we can call L, the greater U – forming an interval (L, U) according to a specified procedure.
- State that $L < \theta < U$ – that is, that θ is in the interval.

We may choose any procedure in step 2 such that in the long run the claim in step 3 will be correct, on average, X% of time. A confidence interval is any interval computed using such a procedure.

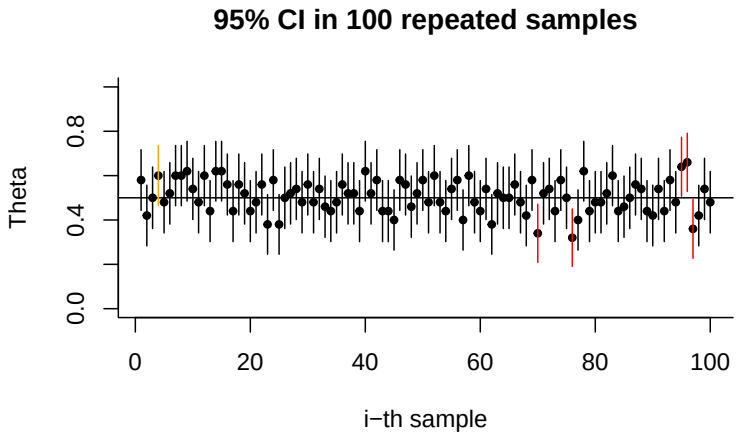
Note that what happens in step 3 is not a belief, any reasoning from the data, or any uncertainty about θ. It is merely a dichotomous statement that is meant to have a specified probability of being true in the long run.

The basic definition of CI is: CI is an interval generated by a procedure that, on repeated sampling, has a fixed probability of containing the parameter. If the probability that the process generates an interval including θ is .5, it is a 50 % CI; likewise, the probability is .95 for a 95 % CI. In other words, these are the probabilities that reflect the accuracy of our procedure.

Let us simulate! The confidence intervals we will construct are two-sided or two-tailed intervals that are symmetrical around the sample statistic, and give us 95% confidence in our procedure. In other words, we will construct very popular 95% CI.

```
plot(0,0,type='n', xlim=c(1, 100), ylim=c(0, 1), bty='l', ylab="Theta", xlab='i-th sample',
     main="95% CI in 100 repeated samples")
abline(h=theta_true)
for (i in 1:100){
  dd <- data[, i]
  n_size <- length(dd)
  m <- sum(dd) / n_size
  se <- sqrt((m*(1-m))/n_size)
  ci <- c(m - 1.96 * se, m + 1.96 * se)
  col = 'black'
  if (ci[1]> theta_true){
    col = 'red'
  } else if (ci[2] < theta_true){
    col= 'red'
  } else if (i == id) {
    col = 'orange'
  }
  points(i, m, pch=20)
```

```
  segments(i, ci[1], i, ci[2], col=col)
}
```



Our procedure is indeed reliable. Constructing 95% CIs leads to the observation that 5% of these intervals do not include the true parameter value.

Note that it is unlikely that two samples from a given population will yield identical confidence intervals. But, over time a large proportion of the confidence intervals constructed from the same population will contain the parameter. It is a misconception to claim that the first sample has an interval including the true value with 95% probability – that there is a 95% probability that future samples would obtain statistics within the first confidence interval. This would only be true if the initial estimate is exactly equal to the true parameter. The confusion comes from thinking that the interval is a "range of possible values" for the true parameter, which is not the case. The true parameter value is fixed but unknown.

Instead, we should think about the CI in the following way. Before the data is collected, we can say that if we were to repeat the sampling process many times, 95% of the intervals we construct (from different samples) would contain the true population parameter. This reflects the reliability of the process, not the interval itself. After the data is collected and the interval is computed, the true population parameter either is or

is not in the interval. The confidence level (like 95%) refers to how often this method will correctly capture the true value if we repeat the process with many samples. It doesn't apply to any one specific interval. The correct interpretation of the interval indicates that we quantify our uncertainty not in the true parameter value but in our behaviour. Namely, how often we are mistaken given a particular procedure.

5.4.1 Confidence fallacy

> **!** The Fundamental Confidence Fallacy
>
> If the probability that a random interval contains the true value is X%, then the plausibility or probability that a particular observed interval contains the true value is also X%; or, alternatively, we can have X% confidence that the observed interval contains the true value.

To demonstrate the fundamental fallacy, we use an example taken from Morey that is based on the confidence interval literature.

> A 10-meter-long research submersible with several people on board has lost contact with its surface support vessel. The submersible has a rescue hatch exactly halfway along its length, to which the support vessel will drop a rescue line. Because the rescuers only get one rescue attempt, it is crucial that when the line is dropped to the craft in the deep water that the line be as close as possible to this hatch. The researchers on the support vessel do not know where the submersible is, but they do know that it forms two distinctive bubbles. These bubbles could form anywhere along the craft's length, independently, with equal probability, and float to the surface where they can be seen by the support vessel.

The rescue hatch is the unknown location θ, and the bubbles can rise from anywhere with uniform probability between $\theta - 5$

meters (the bow of the submersible) to $\theta + 5$ meters (the stern of the submersible). The rescuers want to use these bubbles to infer where the hatch is located.

$$y_i \sim Uniform(\theta - 5, \theta + 5)$$

We will denote the first and second bubble observed by y_1 and y_2, respectively (y_1 always denotes the smaller location of the two). We denote their difference as d.

The rescuers first note that from observing two bubbles, it is easy to rule out all values except those within five meters of both bubbles because no bubble can occur further than 5 meters from the hatch. If the two bubble locations were $y_1 = 4$ and $y_2 = 6$, then the possible locations of the hatch are between 1 and 9, because only these locations are within 5 meters of both bubbles. This constraint is formally captured in the *likelihood*, which is the joint probability density of the observed data for all possible values of θ. In this case, because the observations are independent, the joint probability density is:

$$p(y_1, y_2; \theta) = p_y(y_1; \theta) \times p_y(y_2; \theta).$$

The density for each bubble p_y is uniform across the submersible's 10 meter length, which means the joint density must be $1/10 \times 1/10 = 1/100$. If the lesser of y_1 and y_2 (which we denote x_1) is greater than $\theta - 5$, then obviously both y_1 and y_2 must be greater than $\theta - 5$. This means that the density, written in terms of constraints on x_1 and x_2, is:

$$p(y_1, y_2; \theta) = \begin{cases} 1/100 & \text{if } x_1 > \theta - 5 \text{ and } x_2 < \theta + 5, \\ 0 & \text{otherwise.} \end{cases} \tag{5.2}$$

If we write the Equation 5.2 as a function of the unknown parameter θ for fixed, observed data, we get the likelihood, which indexes the information provided by the data about the parameter. In this case, it is positive only when a value θ is possible given the observed bubbles

$$p(\theta; y_1, y_2) = \begin{cases} 1 & \theta > x_2 - 5 \text{ and } \theta < x_1 + 5, \\ 0 & \text{otherwise.} \end{cases}$$

We replaced 1/100 with 1 because the particular values of the likelihood do not matter, only their relative values. Writing the likelihood in terms of $\bar{x}$ and the difference between the bubbles $d = x_2 - x_1$, we get an interval:

$$p(\theta; y_1, y_2) = \begin{cases} 1 & \bar{x} - (5 - d/2) < \theta \leq \bar{x} + (5 - d/2), \\ 0 & \text{otherwise.} \end{cases}$$

If the likelihood is positive, the value θ is possible; if it is 0, that value of θ is impossible. Expressing the likelihood as in Eq. 2 allows us to see several important things. First, the likelihood is centered around a reasonable point estimate for θ, $\bar{x}$. Second, the width of the likelihood 10−d, which here is an index of the uncertainty of the estimate, is larger when the difference between the bubbles d is smaller. When the bubbles are close together, we have little information about θ compared to when the bubbles are far apart. Keeping in mind the likelihood as the information in the data, we can define our confidence procedure based on the sampling distribution of the mean $\bar{x}$.

The sampling distribution of $\bar{x}$ has a known triangular distribution with θ as the mean (see Figure 5.6). With this sampling distribution, there is a 50 % probability that $\bar{x}$ will differ from θ by less than $5 - 5/\sqrt{2}$. We can thus use $\bar{x} - \theta$ by noting that there is a 50 % probability that θ is within this same distance of $\bar{x}$ in repeated samples. This leads to the confidence procedure:

$$\bar{x} \pm (5 - 5/\sqrt{2})$$

This procedure also has the familiar form $\bar{x} \pm C \times SE$. Consider Figure 5.7, which shows the resulting likelihood, Bayesian posterior, and confidence intervals when y_1 and y_2 are either close to one another (scenario 1) or far apart (scenario 2). Note the relation between the confidence interval and the likelihood in both scenarios. In the scenario 1, the CI is nested in the likelihood while in the scenario 2 the opposite is true. When the bubbles are far apart (scenario 2), the hatch can be localized very precisely: the bubbles are far enough apart that they must have come from the bow and stern of the submersible. In other words, we are 100% sure that the hatch must be located between these bubbles. This is reflected in the likelihood function and also in the Bayesian posterior when assuming uniform prior. When the data is clear (i.e., the bubbles are far

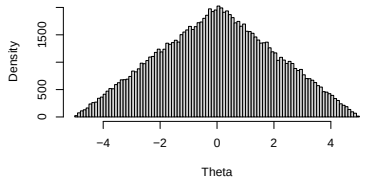
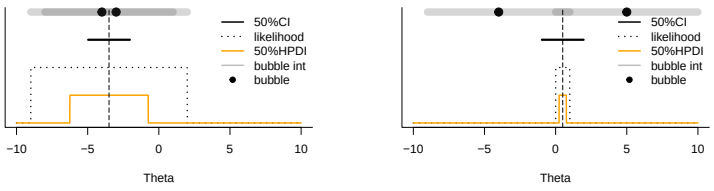


Figure 5.6: Sampling distribution of θ.

apart), the posterior distribution is narrow and more concentrated around the true parameter (the hatch's location). When the bubbles are close together, the posterior distribution is more spread out, reflecting the greater uncertainty about the hatch's location. In contrast, the confidence interval cannot represent 50% probability of containing the true parameter. The fallacy occurs when one misinterprets confidence levels as direct probabilities in a specific case. The confidence level (e.g., 50%) applies to the procedure, not any one particular interval. If we were to repeat the experiment many times, 50% of the intervals generated using this method would contain the true parameter, but we cannot assign a probability to any single interval after we have observed it.



(a) Scenario 1 (b) Scenario 2

Figure 5.7: Confidence and posterior interval to submersible rescue attempts.

5.5 Significance testing (R. Fisher)

P-value is the probability of obtaining test results at least as extreme as the result actually observed given the parameter values of the model under consideration.

$$Pr(T(D_{sim}) \geq T(D_{obs}); \text{mdl, int})$$

where $T(D_{sim})$ is a descriptive summary value of data that were sampled from a hypothetical population characterized by (fixed) parameters of a statistical model *mdl* according to stopping and testing

intentions int while $T(D_{obs})$ is a descriptive summary value of data that were observed according to the same model as well as the stopping and testing intentions.

The concept of the **p-value** was introduced by **Ronald A. Fisher** as a way to measure the strength of evidence against a null hypothesis. According to Fisher, the **p-value** (or **probability value**) is the probability of obtaining a test statistic at least as extreme as the one observed, assuming that the null hypothesis is true. For example, in our coin tossing experiment, we have:

- stopping and testing intentions such that we stop the experiment once 50 tosses is obtained;
- a statistical model generating tosses from a fair coin, i.e., the parameter value of coin landing heads is fixed at $\theta = 0.5$;
- a descriptive summary of coin tosses (sufficient statistic) which is the number of heads.

Again, our observed data have 30 heads. Now we need to build the probability distribution if we were sampled many experiments under these conditions. We already simulated such imaginary data but were considering the proportion of tosses. Now we simply count the number of heads per sample, and compute the p-value (Figure 6.1).

i Fisher's Interpretation of the p-value

1. **Measure of Evidence** – Fisher saw the p-value as a tool to quantify how unusual the observed data is under the null hypothesis.
2. **Continuous Measure** – Unlike strict decision-making rules, Fisher believed that smaller p-values provided stronger evidence against the null hypothesis.
3. **No Fixed Threshold** – Fisher did not advocate for a strict cutoff (such as 0.05). Instead, he suggested that p-values should be interpreted in context.

4. **Significance Testing** – He proposed that if the p-value is **"small enough"**, the data provide reason to doubt the null hypothesis.

The Fisherian approach (by Ronald Fisher) emphasizes p-values and evidence against H_0 without strict decision-making. Interpretation of p-value as strength of evidence.

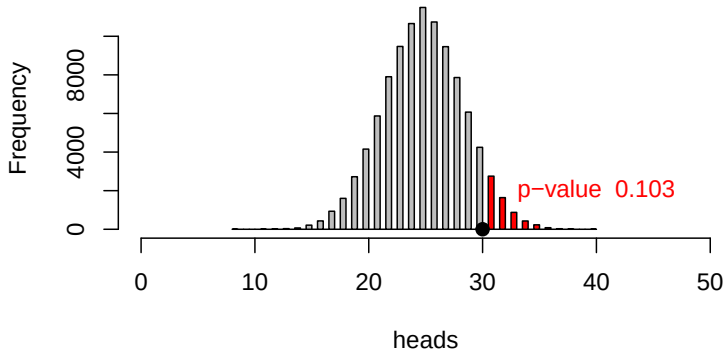


Figure 5.8: Sampling plan with the stopping rule Stop when 50 trials.

Computing p-value using exact binomial test – an exact test of a simple null hypothesis about the probability of success in a Bernoulli experiment. We obtain the same p-value as the one from our simulations.

5.5.1 Relevance of stopping rules

The sampling distribution depends on the particular experimental scheme chosen. Previously, we set the sample size to be fixed at $n = 50$, and obtained $k = 30$ heads. What if there is another sample plan but the obtained data are the same? For example, we could also sample until we obtained $n - k$ failures. Let us simulate such a scenario. The sampling distribution under the new sampling plan is depicted in the Figure 5.9. As a consequence of the sampling plan, we obtain a different p-value despite collected exactly the same data. In other words, the intensions before data collection affect our inference. This

striklingly different than what the frequentist interpretation of probability promises.

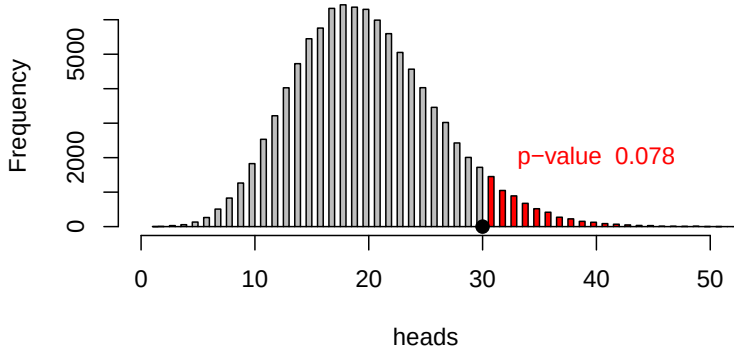


Figure 5.9: Stop when 50-30 failures.

Sampling plan and likelihood paradigm

How does the sampling plan affect the inference according to the likelihood paradigm? The Figure 5.10 depicts two likelihood functions: one under the fixed sampling plan while the other under $n-k$ failures plan. To a likelihoodist, the stopping rule is irrelevant, because the likelihoods are proportional (i.e., the likleihood ratio is the same). We may also say that p-values violate the *likelihood principle* that states the two studies yielding the same data and using the same probabilistic model must have equivalent measurements of the strength of the evidence in those data. This lack of dependence on the stopping rule chosen is seen by frequentists as a weakness, because it implies that procedures have no frequentist error guarantees. By likelihoodists, however, this lack of dependence is seen as a strength, because inferences will be invariant to seemingly irrelevant considerations like why an experimenter chose to end an experiment.

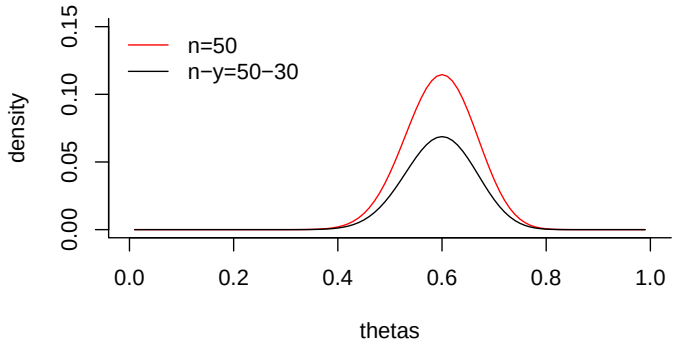


Figure 5.10: Likelihood functions under two sampling plans are proportional.

5.6 Hypothesis testing (J. Neyman + E. Pearson)

Jerzy Neyman and Egon Pearson introduced hypothesis testing as an alternative to Fisher's significance testing because they wanted a more structured and repeatable decision-making framework for statistical inference. Neyman and Pearson found Fisher's approach unsatisfactory because it lacked a formal decision-making framework and control over long-term error rates. Their approach focused on decision rules rather than just measuring evidence (see the comparison in the Table 5.1).

i Neyman-Pearson concerns over p-values

- *No decision rule.* Fisher viewed the p-value as a continuous measure of evidence against the null hypothesis, but he did not define a strict cutoff for decision-making. Scientists often had to subjectively decide whether a p-value was "small enough". Neyman and Pearson argued that this approach was too ambiguous and lacked a repeatable rule for making

conclusions.

- *No alternative hypothesis.* Fisher only focused on testing whether H_0 was likely or not but didn't explicitly consider an alternative hypothesis. Neyman and Pearson argued that real-world decisions require comparing two competing claims.

- *No systematic error control.* Fisher's p-value does not control the probability of making wrong decisions over many repeated experiments. Neyman and Pearson wanted a method that systematically controlled errors (false positives and false negatives). They emphasized statistical power to minimize false negatives. Neyman and Pearson were focused on practical applications, like quality control in manufacturing or medical trials. Fisher's approach worked well for one-time scientific investigations but wasn't ideal for situations where decisions had to be made repeatedly. They wanted a method that could provide reliable long-term decision-making, minimizing false conclusions over multiple trials.

Together, Fisher's method was designed more for exploratory research, where scientists assess how much evidence data provides. Neyman and Pearson wanted a method that worked consistently for decision-making, especially in industry and applied sciences.

The Neyman-Pearson framework of hypothesis testing is a formal and rigorous approach to statistical decision-making. In hypothesis testing, we start with two competing claims or hypotheses:

- Null Hypothesis (H_0): This represents the default assumption, usually stating that there's no effect, no difference, or no relationship. For example, "the coin is fair."

- Alternative Hypothesis (H_1): This represents the claim that contradicts the null hypothesis, typically suggesting

that there's some effect, difference, or relationship. For example, "the coin is biased."

For each hypothesis, we want to understand what kind of data (or sample statistics) we would expect to observe if the hypothesis were true. This leads to the idea of comparing data under two sampling distributions:

- Sampling Distribution under H_0: When we assume that the null hypothesis H_0 is true, the distribution of the test statistic (such as the sample mean, sample variance, or some other statistic) follows a certain distribution. This is the null distribution.
- Sampling Distribution under H_1: Similarly, if we assume the alternative hypothesis H_1 is true, the distribution of the same test statistic will differ in some way. This is the alternative distribution.

These distributions are different because the underlying assumptions about the data are different depending on which hypothesis is true. For example, under H_0, the data might come from a normal distribution with a certain mean and variance while under H_1 the data might come from a normal distribution with a different mean or variance (or even a completely different type of distribution). The two distributions help us assess the likelihood of the observed data under each hypothesis, and the goal is to decide which hypothesis is more consistent with the data. To account for the error in the decision making, we specify two error rates that define a region of rejection for the two hypotheses:

- Type I error (false positive rate) – "an innocent person is convicted" – accepting H_1 when H_0 is true
- Type II error (false negative rate) – "a guilty person is not convicted" – accepting H_0 when H_1 is true.

Nowadays this interpreted as rejecting H_0 when it is actually true and failing to reject H_0 when H_1 is actually true.

Consider again our coin tossing example. We specify the null hypothesis as $\theta = 0.5$ and the alternative hypothesis as $\theta = 0.65$. We specify $\alpha = 0.05$ and samples size $n = 50$, which gives

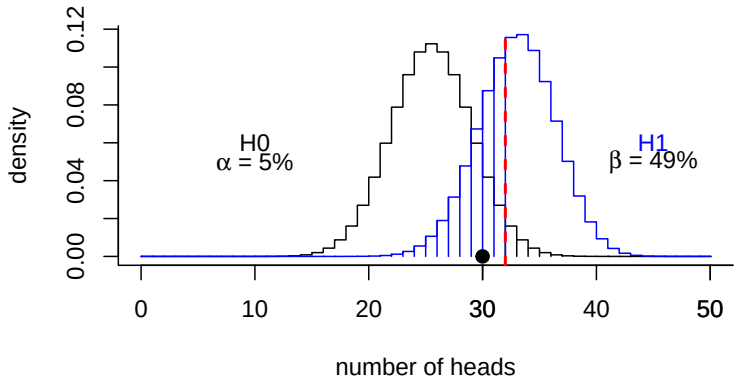


Figure 5.11: Something

us $\beta = 0.49$ so that the power of the test is 51%. This is extremely low, but for the demonstration purpose, we ignore it for now. We will discuss power / study planning in the next lecture. Given our set-up, we are prepared to make a decision: *If the data falls into the rejection region of H_0, we accept H_1; otherwise accept H_0*. Importantly, accepting a hypothesis does not mean that we believe in it, but only that we **ACT AS IF** it were true. Referring to our example: *if I act AS IF H_0 was true, then I will be wrong no more than 49% of the time in the long run, given than H_1 is true.*

This means that we can be either wrong or not when accepting a hypothesis. Likewise, the hypothesis itself can be either true or false. The specified error rates control how often we are wrong in relation to the pre-specified null and alternative hypotheses (not in general). They say nothing about the hypothesis itself. The probability a hypothesis being true can only be derived from Bayes' Rule. This was was unsatisfactory to both the Fisher and Neyman–Pearson camps due to the explicit use of subjectivity in the form of the prior probability. Fisher's strategy is to sidestep this with the p-value (an objective index based on the data alone) followed by *inductive inference*, while Neyman–Pearson devised their approach of *inductive behaviour*.

Table 5.1: Fisher vs. Neyman-Pearson: Key Differences.

Feature	Fisher's Approach	Neyman-Pearson Approach
Purpose	Measures evidence against H_0	Makes a decision (accept/reject)
Hypotheses	Only H_0 (no H_1)	Both H_0 and H_1
p-value	Used as a measure of strength	Used to control false positive rate
Error Consideration	No explicit error control	Controls Type I and II errors
Decision Rule	Flexible	Pre-determined cutoffs

$$f : (\delta, n, \alpha) \mapsto (1 - \beta) \tag{5.3}$$

Statistical power is a function of effect size, sample size, and significance level.

Types of power when assuming fixed significance level:

- a prior power analysis $n = f(\delta, (1-\beta), \alpha)$ – aims to quantify sample size when assuming true effect size and desired power.
- a (post-hoc) sensitivity power analysis $\delta = f(n, (1-\beta), \alpha)$ – aims to quantify effect size when assuming sample size and power. For example, "this sample size was sufficient to detect the effect size X with sufficient power (e.g., 80%)".
- a post-hoc power analysis $(1 - \beta) = f(\hat{\delta}, n, \alpha)$ – aims to quantify power once effect size is estimated and sample size is fixed. It is equivalent with the p-value (no new information).

When high-powered study is unfeasible?

The **smallest effect size of interest (SESOI)** is the minimum magnitude of an effect that is considered practically or scientifically meaningful in a given context. Ideally, if outcomes were measured on the same real scales in all studies. If this is not possible, then standardized effect sizes are often the second best option.

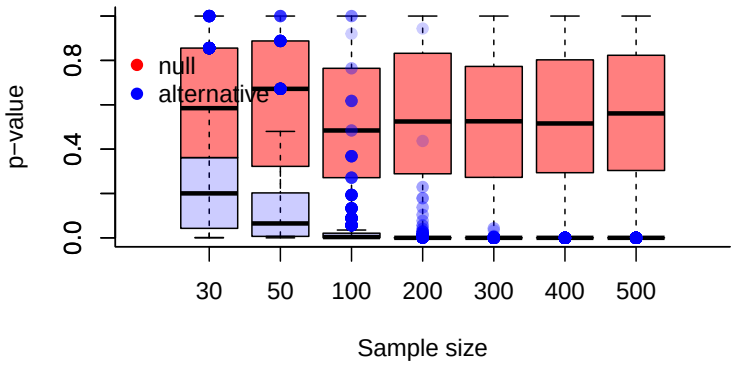


Figure 5.12: som

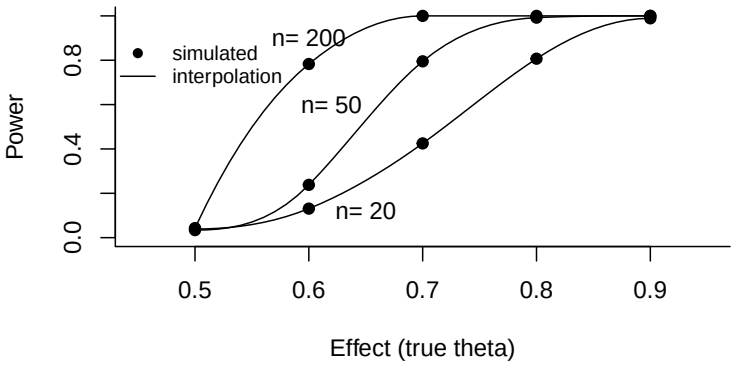


Figure 5.13: som

5.6.1 Type M and S error

There are two other types of error beyond Type I and II error. These are **Type M(agnitude)** and **S(ign) error**, which are both closely related to statistical power.

- Type S error: the probability that the sign of the effect is incorrect (e.g., positive instead of negative), given that the result is statistically significant.
- Type M error: the probability that the effect size estimate from a study is larger in magnitude than the true effect size, given that the result is statistically significant. This error is more likely when statistical power is low, leading to an overestimation of effects in published research.

When power is low, even statistically significant results might be misleading, especially when the significant effect is an overestimate (Type M error). This is important in the context of *publication bias* – where the original published estimates based on low-powered studies do not replicate. A useful way to inspect small-study effects is through **funnel plots** (in meta-analysis[1]). A funnel plot is a scatter plot of the studies' observed effect sizes on the x-axis against a measure of their power (or standard error) on the y-axis.[2] Estimates obtained from low-powered studies tend to be overestimated (the lower part of the funnel), and as power goes up (or standard error goes down), the effect estimates start to cluster tightly around the true value of the effect. Publication bias leads to an asymmetrical funnel plot, where small studies with non-significant results are missing (**?@fig-funnel**). When smaller studies report only large, significant effects, it increases the likelihood of Type M errors, inflating

[1] Meta-analysis is a statistical technique used to combine and analyze results from multiple independent studies on the same topic to estimate an overall effect size. It helps increase statistical power, improve precision, and identify patterns or inconsistencies across studies. By systematically reviewing and synthesizing data, meta-analysis provides a more reliable and generalizable conclusion than individual studies alone. Common tools in meta-analysis include forest plots (to visualize effect sizes) and funnel plots (to check for publication bias).

[2] Usually, the y-axis in funnel plots is inverted (meaning that "higher" values on the y-axis represent lower standard errors).

the perceived effect size in the literature. A symmetrical funnel plot suggests lower risk of Type M errors, while asymmetry signals potential bias and overestimation.

P-values in the context of evidential framework

The concept of a null hypothesis is used differently in two approaches to statistical inference presented by Fisher vs Neyman-Pearson. Acccording to Fisher in the significance testing framework, a null hypothesis is rejected if the observed data are significantly unlikely to have occurred if the null hypothesis were true. In this context, p-value represents the strength of evidence. In the hypothesis testing approach of Neyman and Pearson, a null hypothesis is contrasted with an alternative hypothesis, and the two hypotheses are distinguished on the basis of data, with certain error rates. In this context, p-value represents the probability that a particular study design will generate misleading evidence (i.e, the type I error) in relation to the alternative hypothesis. Importantly, there is no strength of evidence in the hypothesis testing framework and converesely there is no probability of observing misleading evidence in the significance testing framework. As a consequence, very often the researchers merge these two approaches when interpreting p-values.

Significance testing and Bayes' rule

What is the probability of a hypothesis to be true when a significant result is obtained? To calculate that we refer to the Bayes' rule:

$$
\begin{aligned}
Pr(true|+) &= \frac{Pr(+|true) \times Pr(true)}{Pr(+)} \\
&= \frac{Pr(+|true) \times Pr(true)}{Pr(+|true) \times Pr(true) + Pr(+|false) \times Pr(false)}
\end{aligned}
$$

Let us now assume we perform hypothesis testing with $\alpha = 0.05$ and $1 - \beta = 0.80$. We do not know the base rate of a hypothesis to be true. Nevertheless, it must

be rather low. For the demonstration purpose, let's say $Pr(true) = 10$. With these assumptions, the $Pr(true|+)$ is:

```
# power of 80% with low base rate
fp_rate <- 0.05
tp_rate <- 0.8
true_base <- .1
false_base <- 1 - true_base

round(tp_rate * true_base / (tp_rate * true_base + fp_rate * false_base), 2)
```

```
[1] 0.64
```

If we increase $1 - \beta = 0.95$, then the $Pr(true|+)$ is:

```
# power of 95% with low base rate
fp_rate <- 0.05
tp_rate <- 0.95
true_base <- .1
false_base <- 1 - true_base

round(tp_rate * true_base / (tp_rate * true_base + fp_rate * false_base), 2)
```

```
[1] 0.68
```

The difference is barely noticible. In other words, well-powered study won't tell us much if the hypothesis is poor.

5.6.2 Neyman-Pearson Lemma

The Neyman-Pearson Lemma is a fundamental result in statistical hypothesis testing that provides a method for constructing the most powerful test for a given size (false positive rate). Specifically, it applies to simple hypothesis tests where both the null hypothesis H_0 and the alternative hypothesis H_1 are completely specified.

Formally, suppose we have observations X with a probability

density function (PDF) under two hypotheses:

- Null hypothesis (H_0): $f_0(X)$;
- Alternative hypothesis (H_1): $f_1(X)$.

The Neyman-Pearson Lemma states that the most powerful test of size α rejects H_0 in favour of H_1 if the likelihood ratio exceeds a certain threshold k:

$$\Lambda(X) = \frac{f_1(X)}{f_0(X)} > k,$$

where k is chosen such that the test has a pre-specified significance level α:

$$P(\Lambda(X) > k|H_0) = \alpha$$

This means that among all tests with the same false positive rate α the likelihood ratio test (LRT) is the most powerful, i.e., it maximizes the probability of correctly detecting H_1 when it is true.

The Neyman-Pearson Lemma applies directly to simple hypotheses, but in many practical scenarios, at least one of the hypotheses is composite (i.e., it contains multiple possible parameter values). This leads to the **Likelihood Ratio Test (LRT)**, which extends this idea to composite hypotheses by using maximum likelihood estimation. In such cases, a frequentist approach of LRT is used for hypothesis testing.

> **Note**
>
> The (generalized) likelihood ratio test (GLRT) extends the Neyman-Pearson framework by comparing the *maximum likelihood estimates (MLEs)* under the null and alternative hypotheses. It is defined as:
>
> $$\Lambda(X) = \frac{\sup_{\theta \in \Theta_1} L(\theta \mid X)}{\sup_{\theta \in \Theta_0} L(\theta \mid X)}$$
>
> where $L(\theta \mid X)$ is the likelihood function of the data given parameter θ, Θ_0 represents the parameter space under H_0, Θ_1 represents the parameter space under H_1. Add that models must be nested.
> The decision rule is:

$$\Lambda(X) > k \quad \Rightarrow \quad \text{Reject } H_0$$

where the threshold k is chosen based on the desired significance level α.
Under regularity conditions, Wilks' theorem states that for large samples, the statistic:

$$-2 \log \Lambda(X)$$

follows approximately a chi-square (χ^2) distribution with degrees of freedom equal to the difference in the number of free parameters between H_0 and H_1.
This framework is widely used in frequentist hypothesis testing across various statistical models.

5.7 Relevance to Open Science

While frequentist statistics has many strengths, some aspects of its application have contributed to the **replicability crisis**. Here are a few criticisms, particularly from the perspective of Open Science:

1. **P-Hacking and Selective Reporting**:
    - One of the most significant criticisms of frequentist statistics is **p-hacking**, where researchers may manipulate their analysis to achieve a statistically significant p-value (typically less than 0.05). This might involve selective reporting of data, changing statistical methods until they yield the desired results, or even stopping data collection early when p-values appear significant. These practices are facilitated by a focus on achieving "statistical significance" rather than reporting full, nuanced findings.
    - In open science, there's an emphasis on **preregistration of studies** (i.e., committing to a methodology before conducting the research) and sharing raw data. This helps prevent such behaviors, but the pressure for significant results in frequentist

analyses has historically led to questionable research practices, contributing to the reproducibility crisis.

2. **Over-reliance on P-values**:

   - Frequentist statistics often centers on p-values as a measure of evidence. However, this focus on **arbitrary thresholds** (e.g., $p < 0.05$) has been criticized for being misleading. Researchers may treat p-values as a binary "pass/fail" criterion for the truth of a hypothesis, rather than as part of a more nuanced understanding of the data. This leads to a **misinterpretation of results**, for example, a p-value of 0.049 might be treated as significant, while a p-value of 0.051 is dismissed, even though the difference is small and possibly not meaningful.
   - This reliance on p-values can distort scientific findings, especially when replication studies, which often yield slightly different results, fail to reproduce original findings because they are outside of this arbitrary threshold.

3. **Publication Bias**:

   - **Publication bias** is the tendency for journals to favor publishing studies with statistically significant findings. Frequentist statistics, due to its focus on p-values, has contributed to this bias by emphasizing "significant" results over null findings (i.e., results that do not show an effect). As a result, many non-significant studies are not published, and those that are may be subject to selective reporting (e.g., only publishing certain variables or outcomes).
   - Open science has tried to address this issue by advocating for **open access**, **preprints**, and the sharing of all research outputs, including null results. However, the inherent biases in frequentist methods, particularly in how results are treated or discarded based on significance, have played a role in perpetuating the reproducibility crisis.

4. **Misuse of Statistical Significance**:

- There is often a misunderstanding of **statistical significance** as implying practical or scientific significance. Frequentist methods, when not properly interpreted, can suggest that a statistically significant result is meaningful or important, even if the effect size is trivial or the study lacks real-world relevance.
- This contributes to the reproducibility crisis because many published studies with "statistically significant" results are later found to have little to no meaningful impact when reproduced, raising questions about the relevance of statistical significance in some fields.

5. **Underpowered Studies**:

   - Many frequentist studies suffer from being **underpowered**, meaning they don't have a large enough sample size to reliably detect an effect if one exists. This is often due to poor planning or resource constraints, and the resulting studies may yield false positives or fail to detect meaningful effects. Underpowered studies are more likely to produce results that don't replicate, contributing to the growing concerns around reproducibility.
   - Open science initiatives like **preregistration** and the emphasis on sample size planning can help mitigate this issue, but it remains a concern for frequentist statistics, especially when researchers conduct studies without adequate power analysis or rely on convenience sampling.

6. **"Null Hypothesis Significance Testing" (NHST) Problems**:

   - The **Null Hypothesis Significance Testing (NHST)** framework, central to frequentist statistics, is criticized for fostering a **binary thinking** approach (i.e., results are either "significant" or "not significant"). This oversimplifies complex scientific phenomena and encourages dichotomous thinking about hypotheses. As a result, findings that don't fit neatly into this framework are often

dismissed or ignored, reducing the overall quality of scientific understanding.
- The NHST approach often obscures the underlying **uncertainty** about estimates, and researchers may report misleading results, thinking that a rejection of the null hypothesis proves a theory true.

5.8 Summary

Frequentist statistics focuses on decision-making and error, emphasizing the long-run behavior of estimators and tests under repeated sampling. It relies on the concept of the sampling distribution, without assigning probabilities to parameters themselves, and aims to minimize errors like Type I and Type II. In contrast, Bayesian statistics centers on what to believe, incorporating prior knowledge and updating beliefs with new data via Bayes' rule.[3] It might be more effective to frame the distinction by stating that Bayesian inference incorporates prior information, while frequentist inference incorporates the information about the sample space (which may depends on subjective intentions). Finally, likelihood is focused on the strength of evidence, quantifying how well different hypotheses explain the observed data without incorporating prior beliefs or decision-making strategies. Each approach serves a distinct purpose in understanding and analyzing data.

5.9 Recommendations

If you are interested in learning about probability theory, the open course at MIT by Prof. J. Tsitsiklis is absolute gold Introduction to probability.

[3]The Bayesian Decision Rule is a decision-making approach that uses probabilities to guide decisions under uncertainty. It incorporates Bayes' theorem to update the probability of different outcomes based on new evidence or information, and it aims to minimize decision-making risk or cost.

5.10 References

Philosophy of statistics

Bertsekas D.P. & Tsitsiklis, J. N. (2000). Introduction to Probability, MIT Cambridge, Massachusetts.

Gigerenzer, G. (2004). Mindless statistics. In The Journal of Socio-Economics (Vol. 33, Issue 5, pp. 587–606). Elsevier BV. https://doi.org/10.1016/j.socec.2004.09.033

Halpin, P. F., & Stam, H. J. (2006). Inductive Inference or Inductive Behavior: Fisher and Neyman: Pearson Approaches to Statistical Testing in Psychological Research (1940-1960). *The American Journal of Psychology*, *119*(4), 625–653. https://doi.org/10.2307/20445367

Morey, R. D. (2018). Statistical Inference. Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience, 5, 1-42.

Kruschke, J. K., & Liddell, T. M. (2017). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. In Psychonomic Bulletin & Review (Vol. 25, Issue 1, pp. 178–206). Springer Science and Business Media LLC. https://doi.org/10.3758/s13423-016-1221-4

Kruschke, J. K., & Liddell, T. M. (2017). Bayesian data analysis for newcomers. In Psychonomic Bulletin & Review (Vol. 25, Issue 1, pp. 155–177). Springer Science and Business Media LLC. https://doi.org/10.3758/s13423-017-1272-1

6 Example

In this section, we apply the three paradigms to a hypothetical drug testing scenario in order to highlight their unique aspects and similarities.

> A medical researcher is studying whether a new drug improves recovery rates for a certain illness compared to a placebo. The trial includes 200 participants, randomly split into two groups of 100: one group receives the drug, and the other receives a placebo. After the treatment period, the recovery rates are recorded for both groups.

6.1 Likelihood

The likelihood of observing $X_1 = k_1$ recoveries in $n_1$trials, assuming a probability of success θ_1 follows the binomial probability mass function:

$$p(X_1 = k_1 \mid \theta_1) = \binom{n_1}{k_1} \theta_1^{k_1} (1 - \theta_1)^{n_1 - k_1} \tag{6.1}$$

Similarly, for the placebo group, the likelihood of observing $X_2 = k_2$ recoveries in n_2 trials, with success probability θ_2, is:

$$p(X_2 = k_2 \mid \theta_2) = \binom{n_2}{k_2} \theta_2^{k_2} (1 - \theta_2)^{n_2 - k_2} \tag{6.2}$$

Since the two groups are independent, the joint likelihood function is simply the product of these two binomial likelihoods:

$$p(k_1, k_2 \mid \theta_1, \theta_2) = \binom{n_1}{k_1} \theta_1^{k_1} (1 - \theta_1)^{n_1 - k_1} \times \binom{n_2}{k_2} \theta_2^{k_2} (1 - \theta_2)^{n_2 - k_2} \tag{6.3}$$

This defines a two-parameter model, leading to a two-dimensional likelihood function.

To simplify the model, we can condition on the total number of observed recoveries, $Z = X_1 + X_2$, which allows us to express the problem in terms of the *odds ratio* between the two groups. Given that Z is fixed, the conditional distribution of $X_1 \mid Z$ follows a hypergeometric-like structure, but it can be rewritten in binomial form using the odds ratio:

$$\psi = \frac{\theta_1/(1-\theta_1)}{\theta_2/(1-\theta_2)} \tag{6.4}$$

This leads to a new binomial probability function:

$$p(k_1 \mid z, n_1, n_2, \psi) = \binom{z}{k_1} \left(\frac{n_1\psi}{n_1\psi + n_2}\right)^{k_1} \left(1 - \frac{n_1\psi}{n_1\psi + n_2}\right)^{z-k_1} \tag{6.5}$$

Suppose we observe $k_1 = 45$ recoveries in the drug group and $k_2 = 30$ in the placebo group, with both groups having $n_1 = n_2 = 100$ participants. The estimated odds ratio is:

$$\begin{aligned} OR &= \frac{\hat{\theta}_1/(1-\hat{\theta}_1)}{\hat{\theta}_2/(1-\hat{\theta}_2)} \\ &= \frac{0.45/(1-0.45)}{0.30/(1-0.30)} \\ &= 1.90 \end{aligned} \tag{6.6}$$

Finally, we visualize the likelihood function and compare two hypotheses: the null hypothesis $H_0 : \psi = 1$ versus an alternative $H_1 : \psi = 1.5$ (see Figure 6.1).

Computing the values, we obtain the likelihood ratio of 4.53, which suggests rather weak evidence for the data being supported by the alternative hypothesis. The odds of recovering in the drug groups is 1.5 times the odds of recovering in the placebo group. In our interpretation we should also consider 1/8 support interval, which in this case overlaps with the regions of the odds ratio < 1.

! Conditional likelihood

When you condition on the sum $Z = X + Y$, the distribution of X given Z will depend on the total Z as well as the relative relationship between X and Y. The odds ratio will now reflect how X and Y are distributed given their fixed sum, and it will no longer be a simple comparison between their raw (independent) odds. In our case, the distribution $p(X \mid Z)$ depends on the true odds ratio but also on the total Z. So the relationship between X and Y becomes somewhat "compressed" or constrained when conditioned on the sum (i.e., the sample space is altered), leading to the reduction in the odds ratio from 1.9 to 1.5.

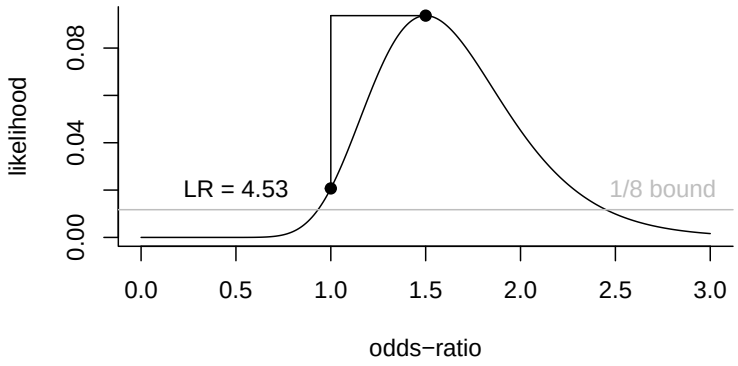


Figure 6.1: Conditional likelihood function

6.2 Frequentists

To apply frequentist inference, we need the sampling distribution of the *odds ratio (OR)* before collecting data. We simulate the null distribution under the assumption that there is no difference between the two binomial proportions, i.e., $\theta_1 = \theta_2$. We consider two cases:

- $\theta_1 = \theta_2 = 0.5$ and
- $\theta_1 = \theta_2 = 0.3$

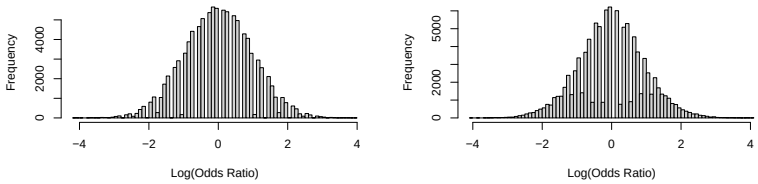
to explore how the sampling distribution behaves under these two assumptions.

The odds ratio (OR) is computed as:

$$OR = \frac{\hat{\theta_1}/(1-\hat{\theta_1})}{\hat{\theta_2}/(1-\hat{\theta_2})} \tag{6.7}$$

To normalize the distribution and center it at 0, we compute the *log odds* and divide it by the *standard error (SE)*:

$$SE = \sqrt{\frac{1}{k_1} + \frac{1}{n_1 - k_1} + \frac{1}{k_2} + \frac{1}{n_2 - k_2}} \tag{6.8}$$



(a) Assuming the proportion of the two groups is 0.5.

(b) Assuming the proportion of the two groups is 0.3.

Figure 6.2: The null distribution of odds ratio on a log scale obtained with simulations.

The sampling distribution of $\log(OR)$ approximates a standard normal distribution:

$$log(OR) \sim N(0, 1) \tag{6.9}$$

Alternatively, it can be approximated by a *t-distribution*:

$$log(OR) \sim t(\nu) \tag{6.10}$$

where the degrees of freedom (ν) are determined based on the *Wald test*:

$$\nu = n_1 + n_2 - 2 \tag{6.11}$$

We now compute the *p-value* for our data using simulated null distribution (Listing 6.1) and compare it with the result from R's built-in `prop.test` function (Listing 6.2).

```
p-value (z dist) 0.029295
 p-value (t dist) 0.03047505
 p-value (sim 0.5) 0.02827
 p-value (sim 0.3) 0.02857
```

```
    2-sample test for equality of proportions without continuity correction
```

Listing 6.1 Calculating p-value based on simulated null distribution.

```
k2 <- 30 # assuming control group
k1 <- 1.5 * k2 # assuming experimental new group
n1 <- 100
n2 <- 100

# Calculate observed log odds ratio and its standard error
p_hat1 <- k1 / n1
p_hat2 <- k2 / n2
odds1 <- p_hat1 / (1 - p_hat1)
odds2 <- p_hat2 / (1 - p_hat2)
or <- odds1 / odds2
lor <- log(or)
se <- sqrt((1 / k1) + (1 / k2) + (1 / (n1 - k1)) + (1 / (n2 - k2)))
t_lor <- lor / se
p_value_z <- 2 * (1 - pnorm(abs(t_lor), 0, 1))
p_value_t <- 2 * (1 - pt(abs(t_lor), n1+n2-2))
p_value_sim_50 <- sum(abs(null_dist_50) >= abs(t_lor)) / length(null_dist_50) # Two-tailed test
p_value_sim_30 <- sum(abs(null_dist_30) >= abs(t_lor)) / length(null_dist_30) # Two-tailed test

cat("p-value (z dist)", p_value_z, "\n",
    "p-value (t dist)", p_value_t, "\n",
    "p-value (sim 0.5)", p_value_sim_50, "\n",
    "p-value (sim 0.3)", p_value_sim_30, "\n"
    )
```

```
data:  c(k1, k2) out of c(n1, n2)
X-squared = 4.8, df = 1, p-value = 0.02846
alternative hypothesis: two.sided
95 percent confidence interval:
 0.01743049 0.28256951
sample estimates:
prop 1 prop 2
  0.45   0.30
```

Our simulated *p-value* closely matches the result from `prop.test`, confirming the validity of our approach.

Listing 6.2 Test of equal proportions as implemented in R.

```
# Test of Equal or Given Proportions
# to check if we obtain the same p-value
# normal approximation to the binomial distribution (a z-test)
prop.test(x = c(k1, k2), n = c(n1, n2), correct = FALSE, alternative="two.sided")
```

6.2.1 Sensitivity power analysis

Normally, we would perform a *prior power analysis* to determine the required sample size before the study begins. However, this step was not performed in this case.

What we can do now is conduct a *post-hoc sensitivity power analysis*. The *effect size* obtained is represented by the *log-odds ratio* (0.65), which reflects the difference between the proportions $\hat{p_1}$ and $\hat{p_2}$.

Based on this effect size, the current sample size was sufficient to detect the observed effect with a minimum power of 58.41%.

In hindsight, the study was *underpowered*, indicating that we should have designed it differently to ensure adequate statistical power.

6.2.2 Obtaining Maximum Likelihood Estimate using `optim`

An alternative approach to finding the maximum likelihood estimate (MLE) is to use numerical optimization. By utilizing optimization techniques, we can numerically maximize the likelihood function to estimate the parameters of our model. In this case, we can leverage the `optim` function to perform this task efficiently (Listing 6.3).

```
Optimized theta1: 0.4500001


Optimized theta2: 0.3000006
```

Listing 6.3 Obtaining Maximum Likelihood Estimate using optim.

```
# Define the log-likelihood function (to minimize negative log-likelihood)
loglik_fun <- function(par) {
  # Extract parameters from the vector
  k1 <- 45  # Fixed data points for k1
  k2 <- 30  # Fixed data points for k2
  n1 <- 100 # Fixed data points for n1
  n2 <- 100 # Fixed data points for n2
  theta1 <- par[1]  # First parameter (theta1)
  theta2 <- par[2]  # Second parameter (theta2)

  # Calculate log-likelihood for x and y
  loglik <- dbinom(k1, n1, theta1, log = TRUE) +
    dbinom(k2, n2, theta2, log = TRUE)

  # Return the negative log-likelihood (since optim minimizes, we negate it)
  return(-loglik)
}

# Set initial guesses for the parameters
par <- c(theta1 = 0.5, theta2 = 0.5)

# Use optim to minimize the negative log-likelihood
result <- optim(par = par, fn = loglik_fun, method = "L-BFGS-B",
                lower = c(0.01, 0.01), upper = c(0.99, 0.99), hessian = TRUE)
```

Once the maximum likelihood estimates are obtained, we can compute the *variance-covariance matrix*. This matrix is crucial for assessing the precision and correlation between the estimated parameters. The variance-covariance matrix can be approximated by the inverse of the *Hessian matrix*, which represents the second derivative of the log-likelihood function. The Hessian matrix is often computed as part of the optimization procedure, and its inverse provides an estimate of the parameter uncertainties.

```
          theta1    theta2
```

```
theta1 404.0438   0.0000
theta2   0.0000 476.1973
```

Once we have the variance-covariance matrix, we can use it to compute the 95% confidence interval for the parameter estimates.

```
Odds Ratio: 1.909086
```

```
95% CI for Odds Ratio: 1.067289 to 3.414828
```

6.2.3 Simulating confidence interval with bootstrapping

In addition to using the variance-covariance matrix for confidence intervals, we can also simulate confidence intervals using the bootstrapping method. Bootstrapping is a powerful resampling technique that allows us to estimate the distribution of a statistic by repeatedly sampling from the observed data with replacement. The advantage of bootstrapping is that it does not rely on parametric assumptions, making it a non-parametric method for estimating the uncertainty of a statistic. It is particularly useful when the underlying distribution of the data is unknown or when the model is complex. By simulating confidence intervals using bootstrapping, we can obtain more robust estimates of parameter uncertainty, especially in cases where traditional methods might be challenging to apply. The bootstrapped confidence interval for our hypothetical scenario is depicted in the Figure 6.3.

6.3 Bayesian

In the Bayesian analysis, we begin by specifying the prior distribution of the recovery rates for each group independently. To avoid introducing bias, we assume the same prior distribution for both groups.

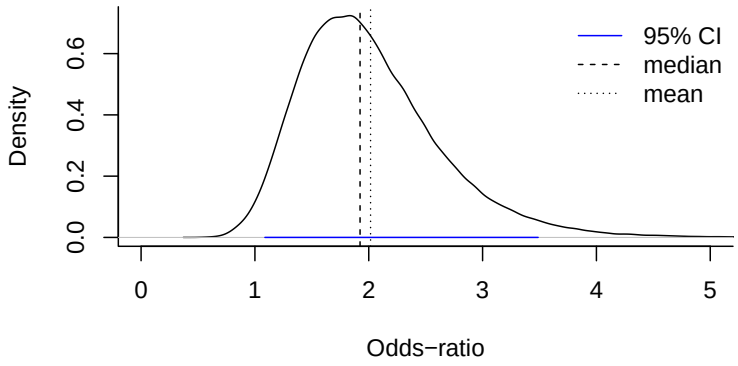
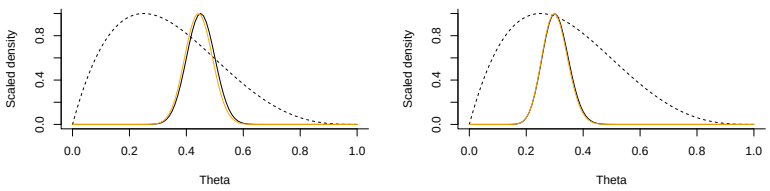


Figure 6.3: Bootstrapped odds ratio

Next, we obtain the posterior distribution of the recovery rates for each group (see Figure 6.4). Using this posterior distribution, we apply a sampling technique to estimate the posterior distribution of the odds ratio.

To summarize the posterior distribution, we use the 95% highest posterior density interval (HPDI). To test our hypothesis, we calculate the Bayes Factor at the odds ratio of 1 and determine the proportion of the posterior interval that lies within the Region of Practical Equivalence (ROPE), defined as [0.84, 1.2].



(a) Experimental group (b) Control group

Figure 6.4: Bayesian analysis of group proportion.

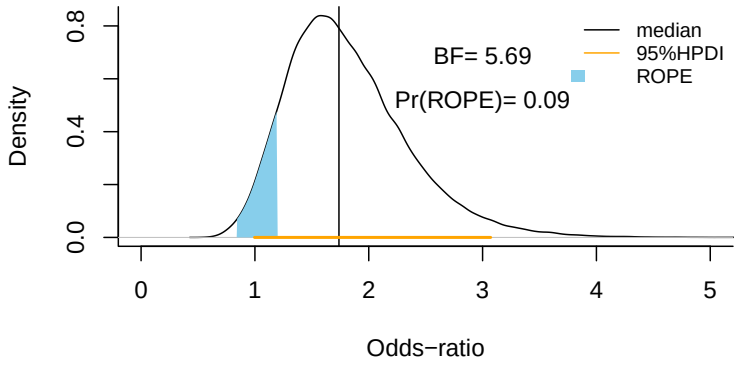


Figure 6.5: Samples of the posterior odds-ratio distribution.

6.4 Summary

Table 6.1: Two conceptual distinctions in the practice of data analysis.

	Frequentist	**Bayesian**
Hypothesis test	p-value (null hypothesis significance testing)	Bayes Factor
Estimation with uncertainty	MLE with confidence interval	Posterior distribution with highest density interval

Part III

Reproducibility

7 Data management

8 Version control

9 Reproducible environments

10 GNU Make

Part IV

Communication

11 Data visualisation

12 Open data and materials

13 Publishing

References