Table of Contents

Statistics: Inferential Statistics

Signposts Throughout the Course Book

Welcome

This course book contains the core content for this course. Additional learning materials can be found on the learning platform, but this course book should form the basis for your learning.

The content of this course book is divided into units, which are divided further into sections. Each section contains only one new key concept to allow you to quickly and efficiently add new learning material to your existing knowledge.

At the end of each section of the digital course book, you will find self-check questions. These questions are designed to help you check whether you have understood the concepts in each section.

For all modules with a final exam, you must complete the knowledge tests on the learning platform. You will pass the knowledge test for each unit when you answer at least 80% of the questions correctly.

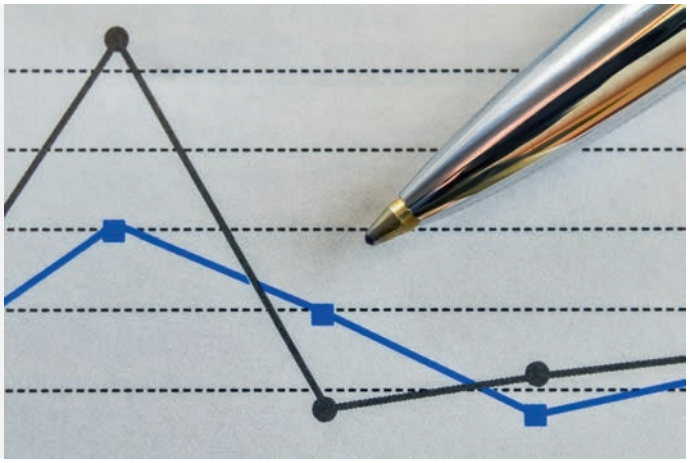
When you have passed the knowledge tests for all the units, the course is considered finished and you will be able to register for the final assessment. Please ensure that you complete the evaluation prior to registering for the assessment.

Good luck!

Learning Objectives

Statistical analysis and understanding are the foundations of data-driven methods and machine learning approaches. **“Statistics: Inferential Statistics”** gives a thorough introduction to point estimators and discusses various techniques to estimate and optimize parameters. Special focus is given to a detailed discussion of both statistical and systematic uncertainties as well as propagation of uncertainties. Bayesian statistics is fundamental to data-driven approaches, and this course takes a close look at Bayesian techniques such as Bayesian parameter estimation and prior probability functions. Furthermore, this course gives an in-depth overview of statistical testing and decision theory, focusing on aspects such as A/B testing, hypothesis testing, p-values, and multiple testing, which are fundamental to statistical analysis approaches in a broad range of practical applications.

The contents of this course book will teach you to understand point estimation methods, apply maximum likelihood and ordinary least squares method to estimate parameters, comprehend the concept of statistical and systematic errors, employ error propagation methods, utilize Bayesian inference and non-parametric techniques, evaluate statistical tests, and grasp the fundamentals of statistical decision theory.



Unit 1

Point Estimation

STUDY GOALS

On completion of this unit, you will have learned...

- ... how to estimate parameters using the method of moments, maximum likelihood, ordinary least squares, and re-sampling techniques.
- ... how to determine if a statistic is sufficient for estimating a parameter.
- ... the definitions and interpretations of the likelihood, log-likelihood, and negative log-likelihood functions.
- ... the assumptions required to determine the ordinary-least-squares estimates for model/function parameters.
- ... how to use the bootstrap and jackknife techniques to point estimates.
- ... how to estimate the uncertainties using bootstrap and jackknife techniques.

1. Point Estimation

Introduction

Suppose we have a sample of 100 numbers that came from an exponential distribution. Remember that the exponential distribution is characterized by its rate parameter λ. In other words, once we know this parameter, we know everything there is to know about the distribution. How can we use the sample to find an estimate for λ? Once we find the estimate, how can we evaluate the quality of the method (estimator) we used? Is there a lot of uncertainty associated with this method? Can we throw away the sample once we have our estimate, or will the individual data points still provide some more information? This unit will help us answer such questions.

Statistical inference is about using information contained from an observed sample to draw inference about the population from which the sample was taken. Populations are characterized by numerical measures called parameters. The objective of point estimation is to estimate the relevant parameters. Many results from probability play an important role in the tools we develop and use in statistical inference. As such, we will review and remind you of the relevant results from probability where it is appropriate.

In the first section, we will learn how to use the the method of moments to find point estimates of parameters of interests. In this method, we relate the parameter of interest to the moments of the underlying distribution and then use the associated sample moments to build the estimator. In the next section, we learn how to figure out if an aggregate quantity (a statistic), based on the data, captured all the relevant information about the parameter of interest we are aiming to estimate. Such a quantity will be called a sufficient statistic.

In section 1.3, we develop an alternative method to estimate parameters of interest: the method of maximum likelihood. At a high-level, this method aims to find an estimate of the parameter of interest by maximizing the likelihood of observing the given data. We explore one of the most commonly used functions in statistics, the likelihood function, and how this quantifies the strategy of finding the point estimate.

In the next section, we introduce the general idea behind ordinary least squares. You may have seen this method applied to simple regression. The advantage this method has over maximum likelihood is that we don't need to know anything about the distribution that generated the given data. Instead, we just need to know the functional or model dependence.

Finally, in section 1.5, we explore two popular re-sampling techniques: the bootstrap and the jackknife. Although their application is vast, we focus on how to make use of the given data in a number of ways to come up with estimates for the parameters of interest without knowing the underlying distribution. The two advantages that come with using these techniques is that (i) they work quite well with small samples sizes where the central limit theorem may not apply and (ii) it provides estimates of the

uncertainties associated with the estimates. In summary, this unit will provide a variety of ways you can use sample data to compute point (single number) estimates of unknown quantities which describe the data.

1.1 Method of Moments

The method of moments is one of the simplest techniques for deriving point estimators. As the name suggests, it is related to the moments of a random variable. Let X be a random variable; its first moment is just its expectation $\mu = \mu^{(1)} = \mathbb{E}\left[X\right]$. Its second moment is the expectation of its square $\mu^{(2)} = \mathbb{E}\left[X^2\right]$. In general, the k^{th} moment is given by

$$\mu^{(k)} = \mathbb{E}\left[X^k\right], k = 1, 2, ...$$

Let's start with the first moment and replace the expectation $\mathbb{E}$ with the average as follows. Take n copies of the random variable X and denote these copies by $X_1, X_2, ..., X_n$. The (first) sample moment is then

$$\widetilde{m}^{(1)} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

This equation is an estimator of the first moment. An estimator is a random variable that aims to estimate an unknown but non-random parameter. If we have observed n realizations of X, a random sample, given by $x_1, x_2, ..., x_n$, then we can compute the estimate of the first moment based on this data by replacing each X_i with x_i to get the following equation:

$$\widehat{m}^{(1)} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Notice that while the estimator $\widetilde{m}^{(1)}$ is a random variable, the estimate $\widehat{m}^{(1)}$ is non-random. We can get analogous equations for the estimators and estimates of higher moments. The sample k^{th} moment estimator is

$$\widetilde{m}^{(k)} = \frac{1}{n}\sum_{i=1}^{n} {X_i}^k$$

and the corresponding estimate based on the observed data is

$$\widehat{m}^{(k)} = \frac{1}{n}\sum_{i=1}^{n} {x_i}^k$$

Example 1.1.1
Compute the first and second sample moments of the observed data shown in the table below.

Sample Data for Example 1.1.1									
1	2	3	4	5	6	7	8	9	10

Solution
We will use equations for $\widehat{\mathrm{m}}^{(1)}$ and $\widehat{\mathrm{m}}^{(\mathrm{k})}$ with this data. The first sample moment is

$$\widehat{\mathrm{m}}^{(1)} = \frac{1}{10}[1 + 2 + \cdots + 10] = \frac{55}{10} = 5\,.\,5$$

The second sample moment is

$$\widehat{\mathrm{m}}^{(2)} = \frac{1}{10}\left[1^2 + 2^2 + \cdots + 10^2\right] = \frac{385}{10} = 38\,.\,5$$

One of the primary tasks in this section is to estimate one or more parameters of a distribution. We need to find a way to relate the unknown parameters to one or more moments. Once this is done, we can use the sample moments to estimate the unknown parameter(s). To this end, let's start with a simple example.

Example 1.1.2
Let $\mathrm{X}_1, \mathrm{X}_2, ..., \mathrm{X}_\mathrm{n}$ be a random sample (independent variables) from the uniform distribution $\mathcal{U}[0, \theta]$, where θ is unknown. Use the method of moments to find an estimator of θ. Next, use the data given below to estimate θ using the estimator you found.

Sample Data for Example 1.1.2									
0.6	0.2	3.9	3.1	3.8	1.6	1.0	1.3	3.0	3.2

Solution
Recall that for $\mathrm{X} \sim \mathcal{U}[\mathrm{a}, \mathrm{b}]$, the first moment is $\mu^{(1)} = \mathbb{E}\left[\mathrm{X}\right] = \frac{\mathrm{a}+\mathrm{b}}{2}$. In this case, we have $\mu^{(1)} = \frac{\theta}{2}$ and replacing $\mu^{(1)} \Longleftrightarrow \widehat{\mathrm{m}}^{(1)}$ gives $\widetilde{\theta} = 2\widehat{\mathrm{m}}$. Therefore, the estimator we get is

$$\widetilde{\theta} = \frac{2}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{X}_\mathrm{i}$$

Next, given the observed data, we can compute the estimate by

$$\widehat{\theta} = \frac{2}{10}[0.6 + 0.2 + \cdots + 3.2] = 4.34$$

We want to explore the performance of the method of moments estimator for various sample sizes:

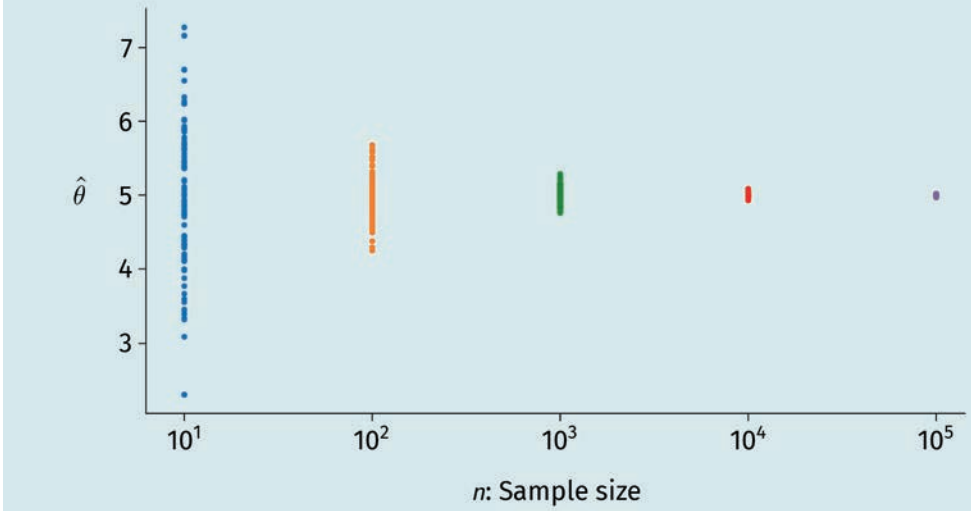
$$\{10, 100, 1,000, 10,000, 100,000\}$$

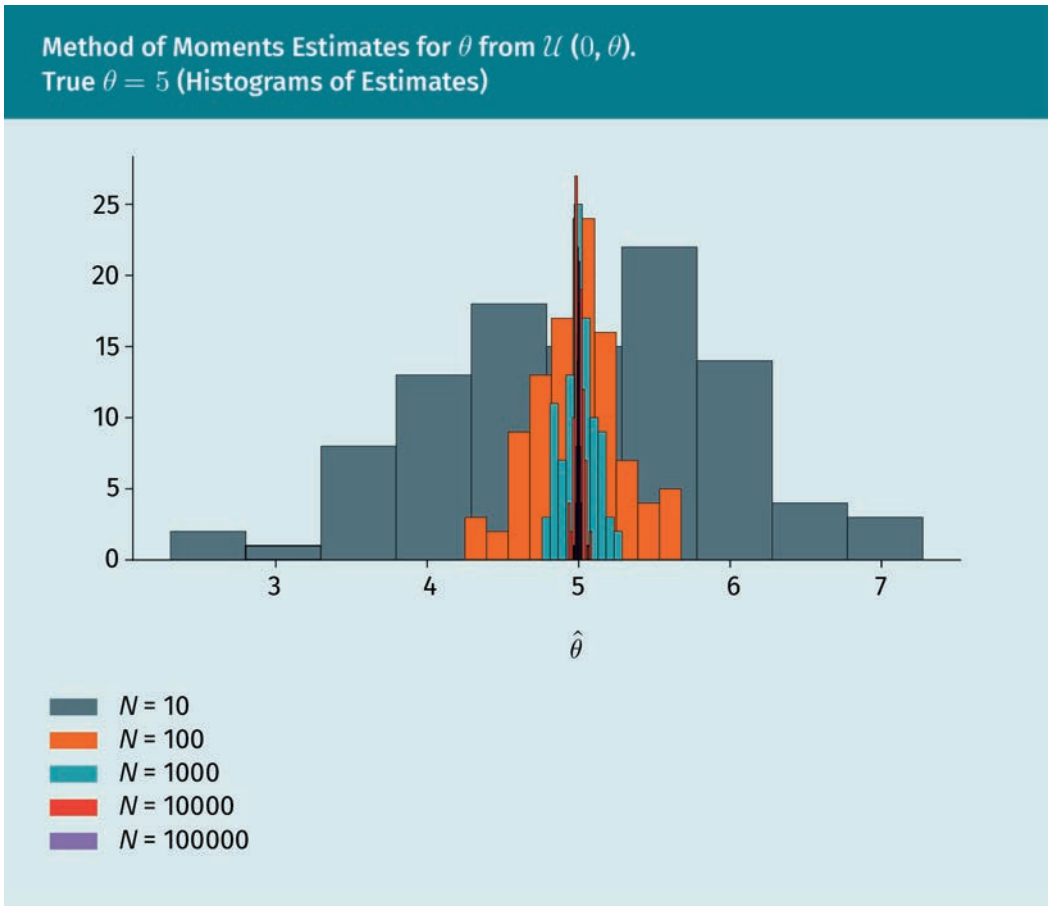
We simulate 100 samples from $\mathcal{U}(0, 5)$ corresponding to each of these samples and compute the method of moments estimate for θ from example 1.1.2. Explore the figure below and note the variation of the estimates for various sample sizes. In the first plot, each point is a method of moments estimate for θ, with the value of this estimate on the vertical axis and the size of the sample used to estimate it on the horizontal axis. The various sample sizes are color coded. Note how the points are scattered for $N = 10$, less scattered for $N = 100$, and eventually, as N is increased, the 100 estimates cluster very tightly around the true value of 5.

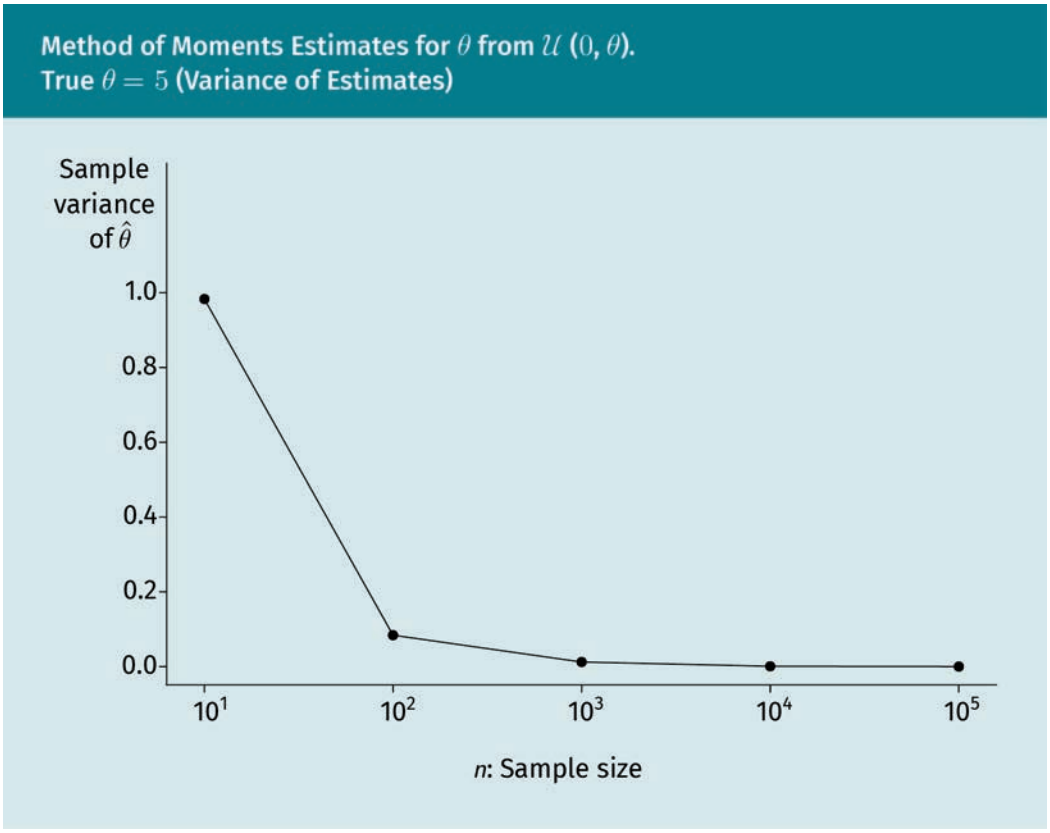
In the middle plot, we have histograms of these points. The value of the estimate is on the horizontal axis and the frequency of the bins are on the vertical axis. Note that values far from the center are less likely for larger values of N, and the shape of the distribution is symmetrical and resembles the Gaussian distribution.

In the final plot, we calculate the sample variance of each of the 100 estimates from each of the samples sizes. In other words, each point represents the sample variance of 100 estimates of θ with the vertical axis showing the values of the sample variance and the horizontal axis showing the number of points used to generate each of the 100 estimates. As you can see, when N is large, the sample variance is small, confirming the clustering behavior we observed in the two plots above.

Method of Moments Estimates for θ from $\mathcal{U}(0, \theta)$.
True $\theta = 5$ (Scatter Plot of Estimates)




Method of Moments Estimates for θ from $\mathcal{U}(0, \theta)$.
True $\theta = 5$ (Histograms of Estimates)
25
20
15
10
5
0
3
4
5
6
7
$\hat{\theta}$
N = 10
N = 100
N = 1000
N = 10000
N = 100000



The next example is the geometric distribution. Recall that the geometric distribution models the number of failures before the first success occurs. If $\mathrm{X}\sim\mathrm{Geometric}(\mathrm{p})$, then its expectation is $\mathrm{E}\left[\mathrm{X}\right] = \frac{1-\mathrm{p}}{\mathrm{p}}$.

Example 1.1.3
Let $\mathrm{X}_1, \mathrm{X}_2, ..., \mathrm{X}_\mathrm{n}$ be iid from $\mathrm{Geometric}(\mathrm{p})$. Find the method of moments estimator for p. Use the data below to find the estimate for $\widehat{\mathrm{p}}$. Recall that iid stands for "independently and identically distributed."

Sample Data for Example 1.1.3									
0	2	5	1	7	4	1	0	3	0

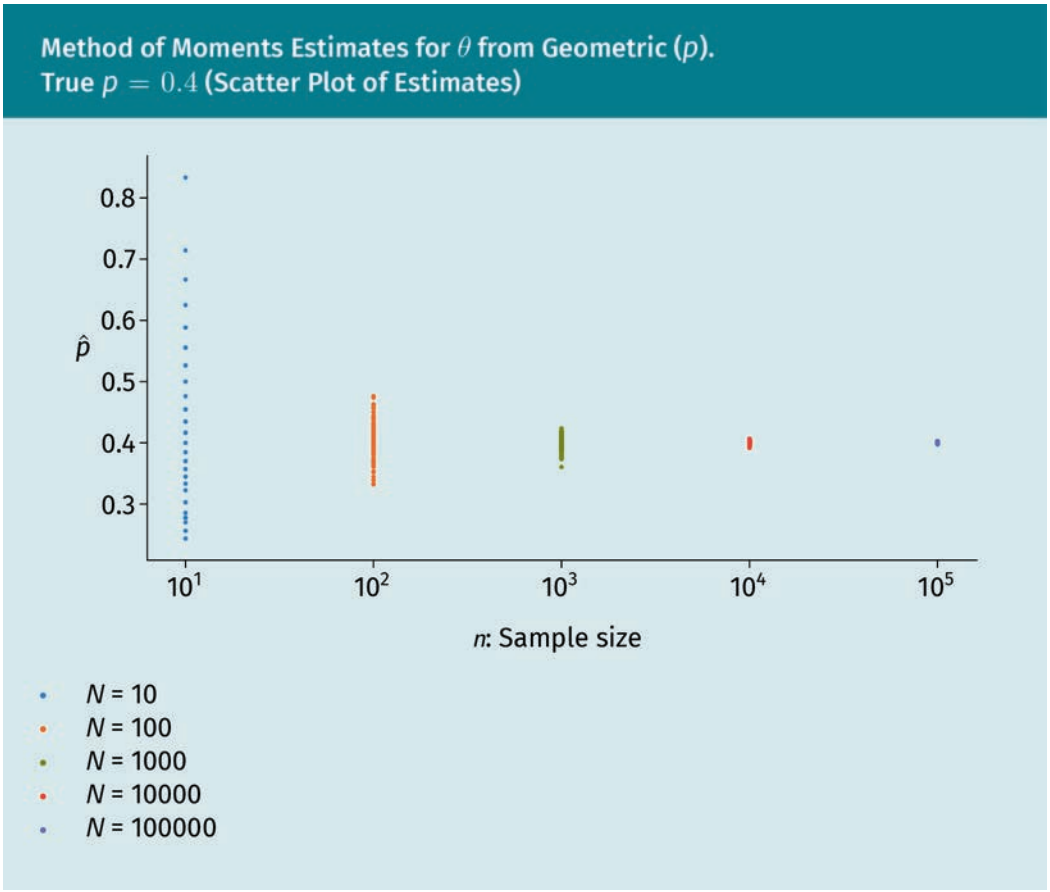
Solution
As stated above, $\mathbb{E}\left[\mathrm{X}\right] = \frac{1-\mathrm{p}}{\mathrm{p}}$ for $\mathrm{X}\sim\mathrm{Geometric}(\mathrm{p})$. Thus, the sample moment estimator gives

$$\widetilde{\mathrm{m}}^{(1)} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{X}_\mathrm{i} = \frac{1-\mathrm{p}}{\mathrm{p}} \Rightarrow \widetilde{\mathrm{p}} = \frac{1}{1+\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{X}_\mathrm{i}}$$

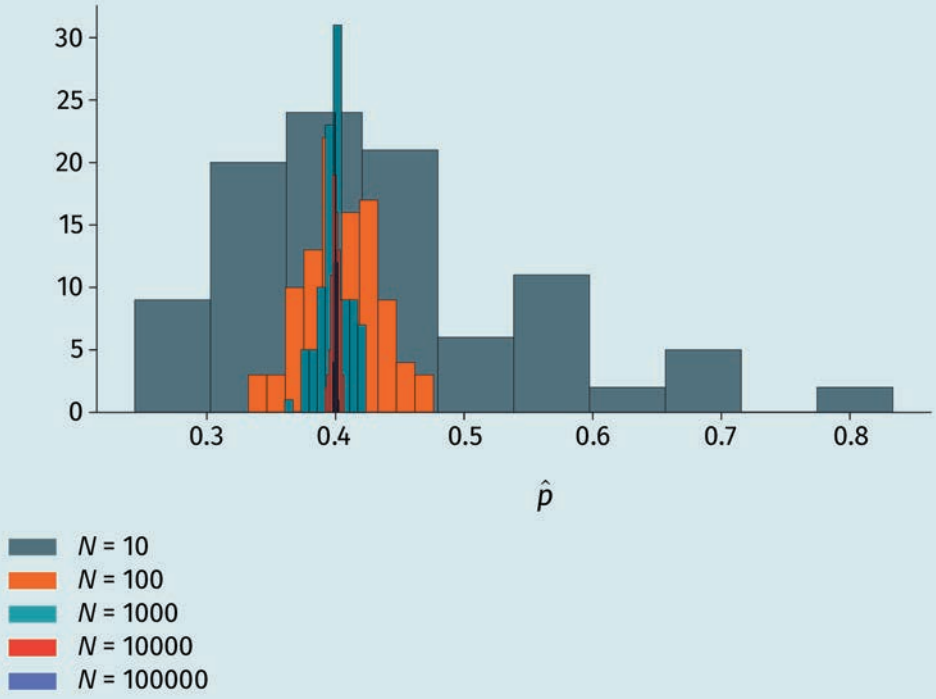
Using the given data, the estimate for p is given by

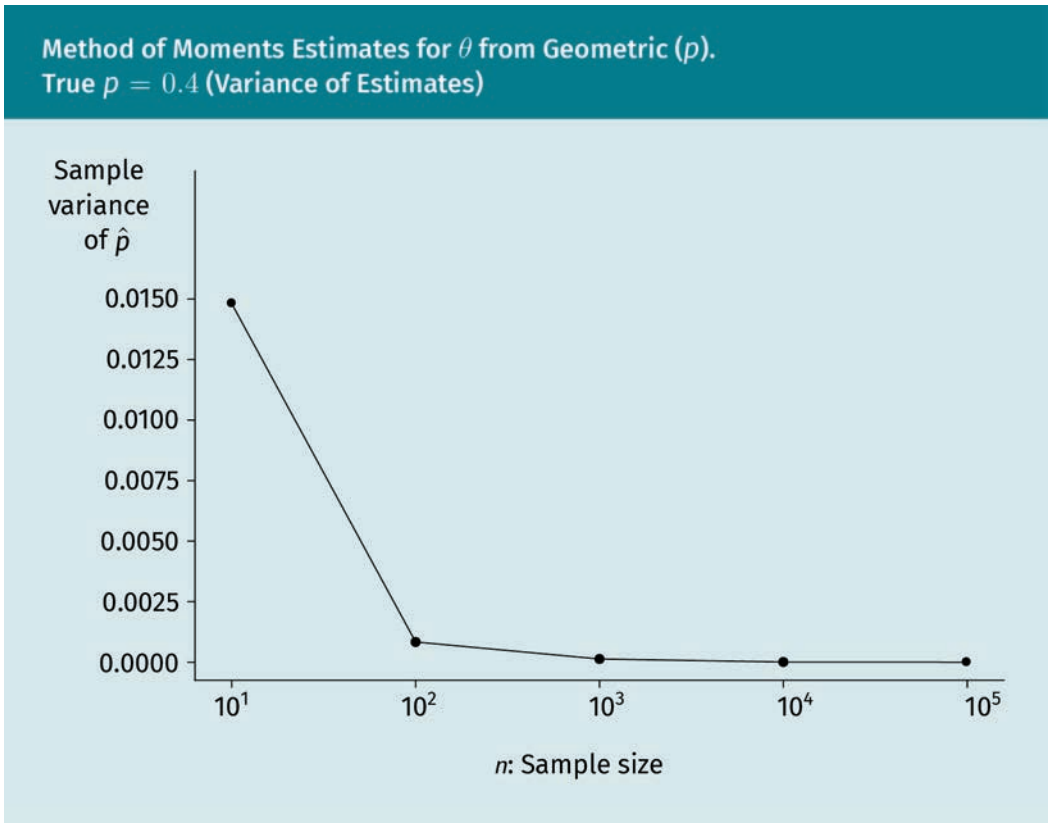
$$\widehat{\mathrm{p}} = \frac{1}{1 + \frac{1}{10}[0 + 2 + \cdots 3 + 0]} = \frac{1}{1 + \frac{23}{10}} = \frac{10}{33} \approx 0.303$$

Similar to the simulation for example 1.1.2, we simulate sets of 100 samples of various sample sizes and compute the method of moment estimates from example 1.1.3. Explore the figure below and note the characteristics of the method of moment estimates for various sample sizes.



Method of Moments Estimates for θ from Geometric (p).
True $p = 0.4$ (Histograms of Estimates)





In all the examples discussed thus far, the first moment was sufficient to obtain an estimator for the unknown parameter. We now discuss an example where we must use the second moment. Recall that the variance of a random variable is $\mathbb{V}[X] = \mathbb{E}[X - \mathbb{E}[X]] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$. If X follows a Gaussian distribution, $X \sim \mathcal{N}(\mu, \sigma)$, then $\mathbb{E}[X] = \mu$ and $\mathbb{V}[X] = \sigma^2$.

Example 1.1.4
Let $X_1, \ldots, X_n$ be iid from $\mathcal{N}(0, \sigma)$. Find the method of moments estimator for σ^2.

Solution
Since $\mu = 0$, we have $\sigma^2 = E[X^2]$; therefore, the method of moments estimator for σ is

$$\widetilde{\sigma}^2 = \widetilde{m}^{(2)} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

Example 1.1.5
Let $X_1, \ldots, X_n$ be iid from $\mathcal{N}(\mu, \sigma)$ with unkown μ and unknown σ. Find the method of moments estimator for μ and σ.

Solution
We know that for $X \sim \mathcal{N}(\mu, \sigma)$, $\mathbb{E}[X] = \mu$, and $\mathbb{V}[X] = \sigma^2$. Therefore, the method of moments estimators are given by

$$\widetilde{\mu} = \widetilde{m}^{(1)} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\widetilde{\sigma}^2 = \widetilde{m}^{(2)} - \left(m_X{}^{(1)}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i{}^2 - (\widetilde{\mu})^2$$

Unbiasedness

Unbiased estimator
An estimator is said to be unbiased if and only if its expected value agrees with the the target parameter.

In addition to providing a way to estimate unknown parameters, point estimates have certain properties with which we evaluate their quality. One of the dimensions along which we evaluate a point estimator is whether they are **unbiased**. The estimator for θ in $\mathcal{U}[0,\theta]$ from example 1.1.2 was given by

$$\widetilde{\theta} = \frac{2}{n}\sum_{i=1}^{n} X_i$$

The expectation of this estimator is

$$\mathbb{E}\left[\widetilde{\theta}\right] = \frac{2}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] = \frac{2}{n}\sum_{i=1}^{n} \frac{\theta}{2} = \frac{2}{n} \cdot n \cdot \theta = \theta$$

Therefore, this estimator is unbiased.

Example 1.1.6
Show that the estimator for μ from example 1.1.5 is unbiased.

Solution
Once again, we just need to compute the expectation of the estimator

$$\mathbb{E}[\widetilde{\mu}] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] = \frac{1}{n} \cdot n\mu = \mu$$

Therefore, the estimator is unbiased.

The statistic $\frac{1}{n}\sum_{i=1}^{n} X_i$ comes up frequently. This statistic is called the **sample mean estimator** and will be denoted by $\overline{X}$ or $\overline{X}_n$. The latter notation is used to explicitly show the size of the sample. In short, we have defined

$$\overline{X} = \overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$$

To summarize, we have seen that $\overline{X}$ is an unbiased estimator for μ for the Gaussian distribution.

Some estimators that come from the method of moments are unbiased, but some are not. Before we give an example of a biased estimator, we need to discuss some preliminary results from probability.

Sample mean estimator
This random variable estimates the mean of the distribution from which a sample is drawn by computing the quotient between the sum of the variables and the sample size.

The variance of a sum of independent random variables is the sum of the variances $\mathbb{V}[X_1 + \cdots X_n] = \mathbb{V}[X_1] + \cdots + \mathbb{V}[X_n]$. Also, the variance of non-random multiples of a random variable is the square of the multiple times the variance of the random variable $\mathbb{V}[c \cdot X] = c^2\mathbb{V}[X]$. Using these two facts, we can compute the variance of $\overline{X}_n$.

$$\begin{aligned}\mathbb{V}[\overline{X}] &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}[X_i] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 \\ &= \frac{1}{n^2}\cdot n\cdot\sigma^2 \\ &= \frac{\sigma^2}{n}\end{aligned}$$

Now recall that $\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$. Therefore, $\mathbb{E}[X^2] = \mathbb{V}[X] + \mathbb{E}[X]^2$. Using this fact, the second moment of $\overline{X}$ is computed as follows

$$\begin{aligned}\mathbb{E}\left[\overline{X}^2\right] &= \mathbb{V}[\overline{X}] + \mathbb{E}[\overline{X}]^2 \\ &= \frac{\sigma^2}{n} + \mu^2\end{aligned}$$

Example 1.1.7
Show that the estimator for σ^2 from example 1.1.5 is biased.

Solution
We will compute the expectation of $\widetilde{\sigma}^2$:

$$\begin{aligned}\mathbb{E}\left[\widetilde{\sigma}^2\right] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i^2\right] - \mathbb{E}\left[\overline{X}^2\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{V}[X_i] + \mathbb{E}[X_i]^2\right) - \left(\frac{\sigma^2}{n} + \mu^2\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\sigma^2 + \mu^2\right) - \left(\frac{\sigma^2}{n} + \mu^2\right) \\ &= \frac{1}{n}\cdot n\cdot\left(\sigma^2 + \mu^2\right) - \left(\frac{\sigma^2}{n} + \mu^2\right) \\ &= \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2 \\ &= \frac{n-1}{n}\sigma^2 \left(\neq \sigma^2 \text{ for } n > 1\right)\end{aligned}$$

We have established that the estimator is biased.

The factor $\frac{n-1}{n}$ is what makes the estimator biased. Therefore, if we multiply the estimator $\frac{n}{n-1}$, the resulting estimator will be unbiased:

$$S_n^2 := \frac{1}{n-1}\sum_{i=1}^{n} \mathbb{E}[X_i]^2 - \overline{X}^2$$

Sample variance
This random variable acts as an estimator of the variance of a distribution from which a random sample is drawn.

This unbiased estimator, S_n^2 for σ^2 is called the **sample variance**. In fact, the estimator $\overline{X}_n$ and S_n^2 are unbiased estimators of the population mean and variance respectively, whether or not the distribution is Gaussian.

1.2 Sufficient Statistics

In the previous section, we defined two important estimators, the sample mean ($\overline{X}$) and the sample variance (S_n^2). These two quantities summarize all the information about the respective parameters (population mean and variance) that the sample contains. In other words, once we have these quantities, the individual values from the sample data play no role in providing additional information about the parameters of interest. In this sense, the statistics $\overline{X}$ and S_n^2 are said to be sufficient. Before defining what is a sufficient statistic, let's recall the definition of a statistic.

Statistic
This random variable is defined by a function of a random sample, usually to estimate an unknown parameter of interest.

Let $X_1, ..., X_n$ be a sequence of random variables. A **statistic** of this sequence is a function $g: \mathbb{R}^n \to \mathbb{R}$ of $X_1, ..., X_n$. In other words, if U is a statistic of this sequence, then $U = g(X_1, ..., X_n)$.

Given a sample $X_1, \ldots, X_n$, the sample mean, $\overline{X}_n$ and the sample variance S_n^2 are two examples of statistics of this sample. We can also define other statistics. One example is the maximum X_{max}, another is the minimum X_{min}, and yet another is the median X_{mid}. All of these are functions of the random sample and therefore statistics. Depending on the parameter of interest, some of these quantities are sufficient and some are not. We are now ready to define sufficient statistics (Hogg et al., 2019).

Sufficient statistic
This statistic contains all the information a random sample provides with respect to estimating an unknown parameter of interest.

Let $X_1, ..., X_n$ be iid from a probability distribution characterized by an unknown parameter θ. A statistic U is said to be **sufficient** for θ if the conditional distribution of $X_1, ..., X_n$, given U, is independent of θ.

Example 1.2.1
Suppose that we have a small sample of just two: X_1, X_2 iid from $\text{Bernoulli}(p)$ with unknown p. Show that the sample mean $U = X_1 + X_2$ is a sufficient statistic for p.

Solution
Since the variables are independent, the joint PMF of X_1, X_2 is just the product of the individual densities:

$$\begin{aligned} f_{X_1, X_2}(x_1, x_2) &= p^{x_1}(1-p)^{1-x_1} \cdot p^{x_2}(1-p)^{1-x_2} \\ &= p^{x_1+x_2}(1-p)^{2-(x_1+x_2)} \end{aligned}$$

Now for the PMF of U, note that $U \sim \text{Binomial}(2, p)$:

$$f_U(u) = \binom{2}{u} p^u (1-p)^{2-u}$$

The joint conditional density of $(X_1, X_2) \,|\, U = u$ is given by

$$\begin{aligned} f(x_1, x_2 \,|\, U = u) &= \frac{f_{X_1, X_2}(x_1, x_2)}{f_U(u)} \\ &= \frac{p^{x_1+x_2}(1-p)^{2-(x_1+x_2)}}{\binom{2}{u} p^u (1-p)^{2-u}} \\ &= \frac{p^u (1-p)^{2-(u)}}{\binom{2}{u} p^u (1-p)^{2-u}} \\ &= \frac{1}{\binom{2}{u}} \end{aligned}$$

Since this joint conditional density doesn't depend on p, the statistic $U = X_1 + X_2$ is a sufficient statistic for this parameter (p).

It may be quite difficult to find the distribution of a statistic U. This makes it difficult to know if U is a sufficient statistic for estimating some unknown parameter θ. To circumvent this challenge, we will introduce a result that will help us not only determine whether a given statistic is sufficient for estimating a parameter but may help us determine a sufficient statistic. We need one more concept before introducing this result.

The joint distribution of a sequence of random variables plays a key role in many statistical inference applications. When this joint distribution depends on a(n) unknown parameter(s), it can be viewed as a function of the parameter(s). This function is the likelihood of observing the data as a function of the parameter(s). Here is the formal definition:

Let $x_1, ..., x_n$ be a sample of observations from the sequence of random variables $X_1, ..., X_n$. Suppose that the distribution of each X_i (for $i = 1, ..., n$) depends on some unknown parameter(s) θ. The **likelihood** of the sample data is the joint density (PMF) of $X_1, ..., X_n$ evaluated at $x_1, ..., X_n$. The **likelihood function** is given by

$$\ell(\theta) = \ell(x_1, ..., x_n \,|\, \theta) = f_{X_1, ..., X_n}(x_1, ..., x_n \,|\, \theta)$$

where f denotes the joint density.

Likelihood
This is a quantified value of how likely a given sample is to be observed.

Likelihood function
This function com-

putes how likely a given sample is to be observed based on a value of the parameter(s) of interest.

Suppose that the random variables $X_1, \ldots, X_n$ are independent and the distribution of each depends on unknown parameter(s) θ_i, $i = 1, ..., n$. Denote the density of X_i by $f_i(x|\theta_i)$. Then, the likelihood function of the sample data $x_1, ..., x_n$ observed from the X_i's is again the joint density. Since the X_i's are independent, the joint density is just the product of the marginals. Therefore, the likelihood function is given by

$$\ell(\theta_1, ..., \theta_n) = \ell(x_1, ..., x_n|\theta_1, ..., \theta_n) = \prod_{i=1}^{n} f_i(x_i|\theta_i)$$

In practice, it is often the case when the observed data are independent and from identically distributed random variables. In other words, we have a random sample $x_1, ..., x_n$ observed from $X_1, ..., X_n$, iid. The distribution of each X_i, $i = 1, ..., n$, is the same and the density (PMF) depends on a(n) unknown parameter(s) θ. To write the likelihood function in this simple case, we can just use the the result of the equation above without the need for subscripts. This is because $f_i \equiv f_j$ for all i, j. To summarize, the likelihood function for sample data observed from an independent and identically distributed random variables is given by

$$\ell(\theta) = \ell(x_1, ..., x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

Example 1.2.2
Write down the likelihood function of the iid sample $\{2, 3, 2, 1\}$ from a Poisson distribution with unknown parameter λ. Recall that the PMF of $X \sim \text{Poisson}(\lambda)$ is given by

$$f_X(x|\lambda) = \begin{cases} \dfrac{e^{-\lambda}\lambda^x}{x!}, & x = 0, 1, 2, ... \\ 0, & \text{otherwise} \end{cases}$$

Solution
Since the sample is iid, we can use the equation above to write the likelihood function:

$$\begin{aligned} \ell(\lambda) &= \ell(x_1, ..., x_4|\lambda) \\ &= \prod_{i=1}^{4} f(x_i|\lambda) \\ &= \prod_{i=1}^{4} \frac{e^{-\lambda}\lambda^{x_i}}{x!} \\ &= \frac{e^{-4\lambda}\lambda^{x_1 + x_2 + x_3 + x_4}}{x_1! \cdot x_2! \cdot x_3! \cdot x_4!} \\ &= \frac{e^{-4\lambda}\lambda^8}{24} \end{aligned}$$

Now consider the statistic $U = X_1 + X_2 + X_3 + X_4$ for the previous example. Is this statistic sufficient for estimating λ? As mentioned above, applying the definition of sufficient statistic involves finding the distribution of U, which may be difficult. Now that we are equipped with some knowledge about the likelihood function, we can use an important result based on this tool. This result makes the connection of the likelihood function and whether or not a statistic is sufficient to estimate an unknown parameter. The following can be found in Hogg et al. (2019).

Theorem 1.2.1
Suppose $X_1, ..., X_n$ is a random sample, $x_1, ..., x_n$ of the corresponding observed data, and U a statistic. Let $\ell(\theta) = \ell(\theta | x_1, ..., x_n)$ be the likelihood function. U is a sufficient statistic for estimating θ if and only if $\ell(\cdot)$ can be factored into a product of two non-negative functions as

$$\ell(\theta) = g(u, \theta) \cdot h(x_1, ..., x_n)$$

where $g(u, \theta)$ doesn't depend on the observed data, and h doesn't dependent on θ.

Let's now return to our original question: is the statistic $U = X_1 + X_2 + X_3 + X + 4$ sufficient for estimating λ in example 1.2.1? Based on the previous result, we have to check whether the likelihood function enjoys a specific form of factorization:

$$\ell(\lambda) = \frac{e^{-4\lambda}\lambda^{x_1 + x_2 + x_3 + x_4}}{x_1! \cdot x_2! \cdot x_3! \cdot x_4!} = \underbrace{e^{-4\lambda}\lambda^{u}}_{g(u, \lambda)} \cdot \underbrace{\frac{1}{x_1! \cdot x_2! \cdot x_3! \cdot x_4!}}_{h(x_1, x_2, x_3, x_4)}$$

Indeed, g doesn't depend on the data except via u, and h doesn't depend on λ. According to this theorem, the U is a sufficient statistic for estimating λ. As a matter of fact, we can extend this result for an arbitrary sample size. Let $X_1, ..., X_n$ denote a random sample from a Poisson distribution with parameter λ. $U = \sum_{i=1}^{n} X_i$ is a sufficient statistic for λ.

Example 1.2.3
Let $X_1, ..., X_n$ be a random sample from a Gaussian distribution with unknown mean μ and known variance σ^2. Show that the sample mean estimator $U = \overline{X}$ is a sufficient statistic for estimating μ. Recall that the density of a Gaussian $X \sim \mathcal{N}(\mu, \sigma)$ is given by

$$f_X(x | \mu, \sigma) = \frac{1}{\sqrt{2\sigma^2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

Solution
We start by writing the likelihood function for an observed sample $x_1, ..., x_n$ corresponding to $X_1, ..., X_n$:

$$
\begin{aligned}
\ell(\mu) &= \ell(x_1, ..., x_n, \sigma \,|\, \mu) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\sigma^2\pi}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2} \right) \\
&= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right) \\
&= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i^2 - 2x_i\mu + \mu^2) \right) \\
&= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2 + \frac{\mu}{\sigma^2} \sum_{i=1}^{n} x_i - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \mu^2 \right)
\end{aligned}
$$

Now set $u = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, so that $nu = \sum_{i=1}^{n} X_i$. The likelihood function can be written as

$$
\begin{aligned}
\ell(\mu) &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2 + \frac{\mu}{\sigma^2} \cdot nu - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \mu^2 \right) \\
&= \underbrace{(2\sigma^2\pi)^{-n/2} \exp\left(\frac{\mu}{\sigma^2} \cdot nu - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \mu^2 \right)}_{g(u,\mu)} \cdot \underbrace{\exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2 \right)}_{h(x_1, ..., x_n)}
\end{aligned}
$$

Therefore, according to the factorization criterion, we know that $U = \overline{X}$ is a sufficient statistic for estimating μ when σ^2 is known.

With a similar computation, it can be shown that $U = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \mu)^2$ is a sufficient statistic for estimating σ^2 when μ is known. Finally, $U_1 = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $U_2 = {S_n}^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \overline{X})^2$ are jointly sufficient statistic for estimating μ and σ^2. To use the factorization criterion to establish joint sufficient statistics, U_1 and U_2 for estimating parameters θ_1 and θ_2, we just replace $g(u, \theta)$ with $g(u_1, u_2, \theta_1, \theta_2)$.

1.3 Maximum Likelihood

The likelihood function measures the likelihood of observing a given sample as a function of a (possibly unknown) parameter. Suppose we are given observed data $x_1, ..., x_n$ corresponding to $X_1, ..., X_n$ whose distribution depend on an unknown parameter θ. The likelihood function gives us a way of determining which value of θ among a set of possible values, best suits our observed sample. In other words, if we have two estimates of θ, $\widehat{\theta}_1$ and $\widehat{\theta}_2$, such that $\ell(\widehat{\theta}_1) > \ell(\widehat{\theta}_2)$, then the given data has a higher likelihood of being observed from a distribution with parameter $\theta = \widehat{\theta}_1$ than from a distribution with parameter $\theta = \widehat{\theta}_2$. Therefore, if these were our only choices for θ, we would choose $\widehat{\theta}_1$ as a point estimate for θ. In this way, we would have used the method of maximum likelihood to determine the point estimate for θ. The estimator $\widetilde{\theta}^{(MLE)} = \widetilde{\theta}^{(MLE)}(X_1, ..., X_n)$ is called the **maximum likelihood estimator** for θ. Analogously, $\widehat{\theta}^{(MLE)} = \widehat{\theta}^{(MLE)}(x_1, ..., x_n)$ is called the **maximum likelihood estimate** variable.

Example 1.3.1
Suppose that the sample given below is iid from Poisson distribution with unknown parameter λ. Write down the likelihood function for this sample. Determine which of the two values $\{3, 4\}$ makes the observed data more likely.

Sample Data for Example 1.3.1									
6	5	6	1	3	6	3	3	2	2

Solution
The likelihood function is

$$\ell(\lambda) = \frac{e^{-10\lambda}\lambda^{37}}{3.87 \times 10^{13}}$$

We compute the likelihood at these values: $\ell(3) = 1.09 \times 10^{-9}$ and $\ell(4) = 2.07 \times 10^{-9}$. Since $\ell(4) > \ell(3)$, $\widehat{\lambda} = 4$ is a better likelihood estimate for λ than $\widehat{\lambda} = 3$.

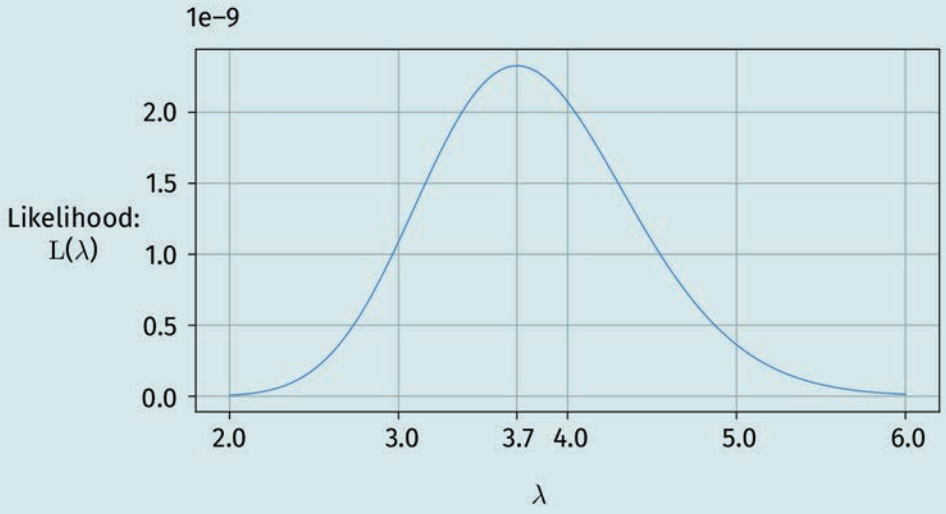
Continuing from the solution of example 1.3.1, in practice, there aren't only two choices for the parameter, there is a continuous range of choices. We definitely can't try all possible values! Instead, we use calculus to find the maximum of the likelihood function. This value is called the maximum likelihood estimate for λ. The figure below shows the graph of the likelihood function $\ell(\lambda)$ from example 1.3.1. It has the values of the parameter λ on the horizontal axis and the likelihood values on the vertical axis. Our goal is to locate the peak (maximum) of this function and then read where this peak occurs on the horizontal axis. In our case, this occurs at 3.7. For this reason, we can denote this as our maximum likelihood estimate: $\widehat{\lambda}^{(\mathrm{MLE})} =$.

Maximum likelihood estimator
This function of a random sample maximizes the likelihood function. Note that this is a random variable.

Maximum Likelihood Estimator
A function of an observed sample which maximizes the likelihood function. Note that this is a non-random quantity.

Maximum likelihood estimate
This function of an observed sample maximizes the likelihood function. Note that this is a non-random quantity.

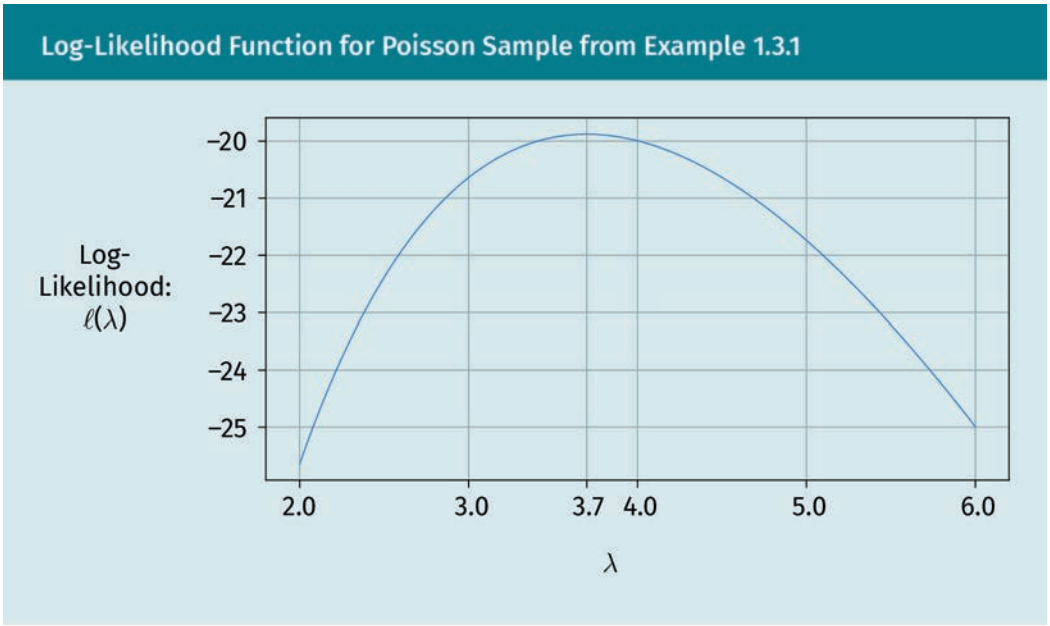
Likelihood Function for Poisson Sample from Example 1.3.1



Log-Likelihood
The (natural) logarithm of the likelihood function admits the same maximizer as the likelihood function but is easier to work with.

Recall that the map $z \mapsto \log z$ is non-decreasing for $z > 0$. Therefore, the maximizer of the log-likelihood, $\ell\ell(\lambda) = \log\ell(\lambda)$ is the same as that of the likelihood $\ell(\lambda)$. Moreover, the **log-likelihood** is easier to work with because it converts products into sums of logarithms; differentiating sums is much easier than differentiating products. In practice, we use computer algorithms to find the maximizer of functions. The likelihood function deals with very small numbers (in magnitude), and we might run into underflow problems where the values we work with are smaller than the smallest number that a computer can represent. The log-likelihood values, on the other hand, avoid small magnitude numbers and are better suited to numerical schemes.

Take a look at the figure below, which shows the log-likelihood function for this Poisson example. Similar to the graph of the likelihood function, the values of the parameter (λ) are on the horizontal axis. The values on the vertical axis have the values of the log-likelihood function $\ell\ell(\lambda)$. As with the likelihood function, we are interested in locating the value on the x-axis where the maximum occurs. This will be the same value as we get from the likelihood graph.



The log-likelihood function corresponding to likelihood from the solution of example 1.3.1 is given by

$$\ell\ell(\lambda) = \log \quad \ell \quad (\lambda) = \; -10\lambda + 37\log\lambda - 13\log(3\,.\,87)$$

Its first and second derivatives are

$$\ell(\lambda)' = \; -10 + \frac{37}{\lambda} \text{ and } \ell\ell(\lambda)'' = \; -\frac{37}{\lambda^2}$$

respectively. Since $\ell\ell''(\lambda) < 0$ for every λ, we know that any zero of $\ell'()$ is a local maximizer. In this instance, $\ell\ell'()$ admits exactly one zero:

$$0 = \; -10 + \frac{37}{\lambda} \Leftrightarrow \lambda = 3\,.\,7$$

Therefore, $\widehat{\lambda}^{(\mathrm{MLE})} = 3\,.\,7$ is the (global) maximizer of the log-likelihood and therefore of the likelihood. For this reason, we have denoted it as the maximum likelihood estimate for λ. Any maximizer of a function ϕ is the minimizer of $-\phi$. Since most optimization algorithms are written for minimizing functions, in practice we often work with the **negative log-likelihood** function, which is

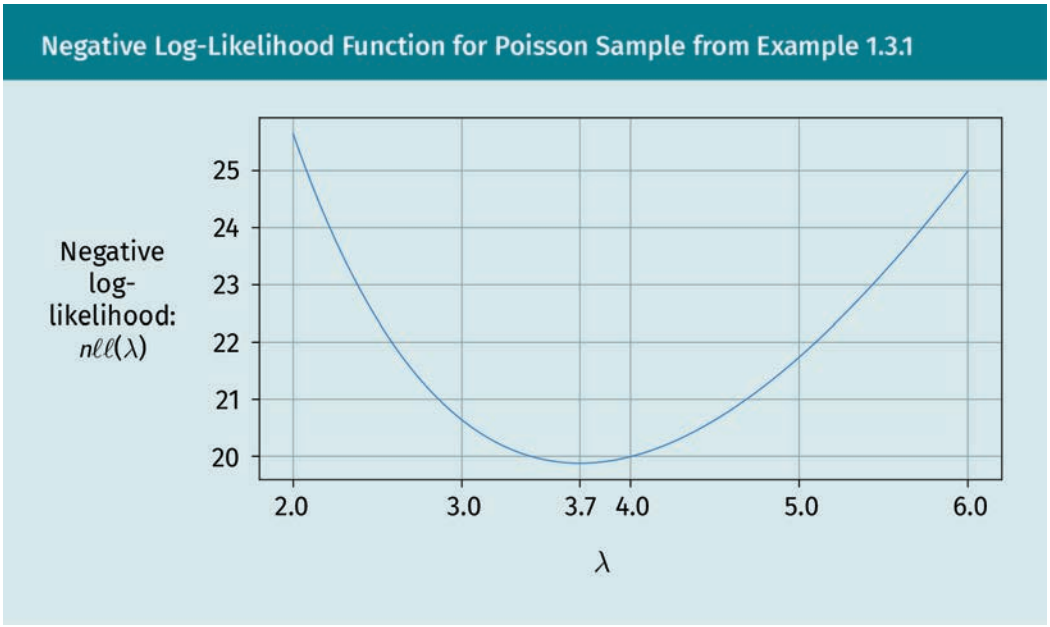
$$n\ell\ell(\theta) = \; -\ell\ell(\theta) = \; -\log \quad \ell(\theta)$$

Negative log-likelihood

With the negative of the log-likelihood function, numerical schemes are written to minimize objective function, as

such, the negative log-likelihood is used in practice. The minimizer of this function is the same as the maximizer of the log-likelihood (and likelihood) function.

The figure below shows the graph of the negative log-likelihood function for the Poisson example. The shape of this graph is the inverted shape of the log-likelihood function. Therefore, using this function, we can locate the MLE by finding the minimum. Notice that 3.7 is the minimizer of this function. For the rest of this section, we will find MLE point estimator/estimates by working with the negative log-likelihood.



Example 1.3.2
Consider the random sample $X_1, ..., X_n$ iid from the exponential distribution with unknown rate λ. Find the MLE estimator by minimizing the negative log-likelihood function. Recall that the density of $X \sim Exp(\lambda)$ is given by

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

Solution
The likelihood function corresponding to an observed sample of $\mathcal{D} = \{x_1, ..., x_n\}$ is given by

$$\ell(\lambda) = \ell(\mathcal{D}|\lambda) = \prod_{i=1}^{n} \left[\lambda e^{-\lambda x_i}\right] = \lambda^n exp\left(-\lambda \sum_{i=1}^{n} x_i\right)$$

The negative log-likelihood function is

$$n\ell\ell(\lambda) = -\log \quad \ell \quad (\lambda) = -n\log\lambda + \lambda \sum_{i=1}^{n} x_i$$

As before, we find the zero(s) of the first derivative:

$$0 = -\frac{n}{\lambda} + \sum_{i=1}^{n} x_i \Leftrightarrow \frac{1}{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ or } \widehat{\lambda}^{(MLE)} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} x_i}$$

The second derivative will confirm that this is the minimizer and therefore, the MLE estimate is indeed the expression above. The corresponding MLE estimator for the rate of an exponential sample is

$$\widetilde{\lambda}^{(MLE)} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} X_i}$$

Example 1.3.3
Suppose that we want to estimate the probability of heads, p, of a coin. We toss a coin n times and denote $X_i = 1$ if the coin turns up heads and $X_i = 0$ if the coin turns up tails. Thus we have a sequence $X_1, ..., X_n$. Let $x_1, ..., x_n$ denote observed values. Find the MLE estimate and estimator for p.

Solution
Assuming coin tosses are independent, and since we are using the same coin, we have $X_1, ..., X_n$ are iid from $\text{Bernoulli}(p)$. Recall that the PMF for $X \sim \text{Bernoulli}(p)$ is given by

$$f_X(x) = \begin{cases} p^x(1-p)^{1-x}, & x = 0, 1 \\ 0, & \text{otherwise} \end{cases}$$

The likelihood of the sample is

$$\ell(p) = \prod_{i=1}^{n} p^{x_i}\left(1-p\right)^{1-x_i}$$

The negative log-likelihood is

$$\begin{aligned} n\ell\ell(p) &= -\sum_{i=1}^{n} (x_i \log p + (1-x_i)\log(1-p)) \\ &= n\log\left(1-p\right) + (\log p - \log(1-p))\sum_{i=1}^{n} x_i \\ &= n\log(1-p) + n(\log p - \log(1-p))\overline{x} \end{aligned}$$

where we have set $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Next, we find the zero(s) of the negative log-likelihood function:

$$0 = -\frac{n}{1-p} + n\left(\frac{1}{p} + \frac{1}{1-p}\right)\overline{x} \Leftrightarrow \widehat{p}^{(MLE)} = \overline{x}$$

Computing the second derivative of the $n\ell\ell(p)$ will confirm that this is indeed the minimizer we are looking for and therefore the maximum likelihood estimate. The corresponding maximum likelihood estimator is

$$\widetilde{p}^{(MLE)} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Uncertainties of MLE

To evaluate the quality of an estimator $\widetilde{\theta}^{(MLE)}$, we need to look at their associated uncertainty: $\mathbb{V}\left[\widetilde{\theta}^{(MLE)}\right]$. Recall that the variance of $X \sim \text{Bernoulli}(p)$ is given by $\mathbb{V}[X] = p(1-p)$. Since the X_i's from the coin toss example are independent, we have

$$\begin{aligned}
\mathbb{V}\left[\widetilde{p}^{(MLE)}\right] &= \frac{1}{n^2}\sum_{i=1}^{n} V[X_i] \\
&= \frac{1}{n^2}\sum_{i=1}^{n} p(1-p) \\
&= \frac{1}{n^2} \cdot np(1-p) \\
&= \frac{p(1-p)}{n}
\end{aligned}$$

Since we don't actually know p, we can estimate the variance with

$$\mathbb{V}\left[\widetilde{p}^{(MLE)}\right] \approx \frac{\widetilde{p}^{(MLE)}\left(1-\widetilde{p}^{(MLE)}\right)}{n}$$

As it turns out, we don't need to compute the uncertainty directly this way. The uncertainty associated with an MLE is the reciprocal of the expected value of the second derivative of the negative log-likelihood evaluated at the MLE:

$$\mathbb{V}\left[\widetilde{\theta}^{MLE}\right] \approx \mathbb{E}\left[\frac{1}{n\ell\ell''(\theta)|_{\widetilde{\theta}^{(MLE)}}}\right]$$

Additionally, this approximation gets better and better for larger values of n. Let's apply this to our coin-toss example. The second derivative of the negative log-likelihood for the coin toss example is

$$n\ell\ell''(p) = -\frac{n}{(1-p)^2} + n\left(-\frac{1}{p^2} + \frac{1}{(1-p)^2}\right)\overline{X}$$

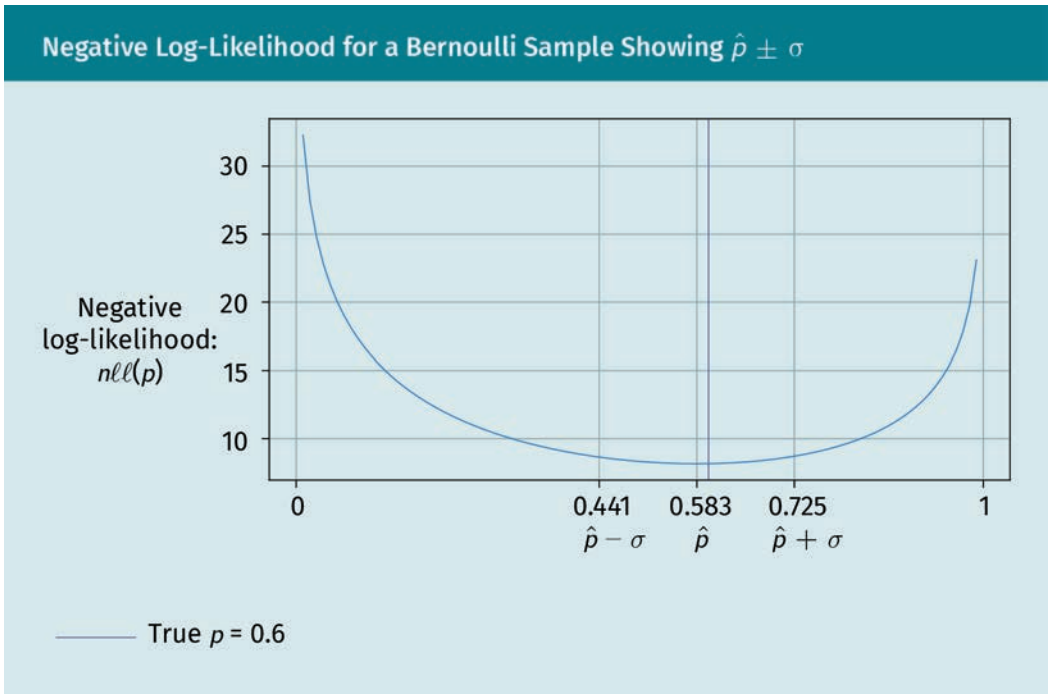
Taking the expectations on both sides (recall that $\mathbb{E}[\overline{\mathrm{X}}] = \mathrm{p}$),

$$\mathbb{E}[n\ell\ell''(\mathrm{p})] = -\frac{\mathrm{n}}{(1-\mathrm{p})^2} + \mathrm{n}\left(-\frac{1}{\mathrm{p}^2} + \frac{1}{(1-\mathrm{p})^2}\right)\mathrm{p} = \frac{\mathrm{n}}{\mathrm{p}(1-\mathrm{p})}$$

Finally, using equation for computing the variance of the MLE, we have

$$\mathbb{V}\left[\widetilde{\mathrm{p}}^{(\mathrm{MLE})}\right] \approx \frac{\widetilde{\mathrm{p}}^{(\mathrm{MLE})}\left(1-\widetilde{\mathrm{p}}^{(\mathrm{MLE})}\right)}{\mathrm{n}}$$

The corresponding standard deviation can be estimated as $\sigma = \sqrt{\frac{\widetilde{\mathrm{p}}^{(\mathrm{MLE})}\left(1-\widetilde{\mathrm{p}}^{(\mathrm{MLE})}\right)}{\mathrm{n}}}$. In the figure below, we plot the negative log-likelihood function for a random sample from $\mathrm{Bernoulli}(\mathrm{p})$ and highlight the estimate $\widehat{\mathrm{p}}$ using the result of example 1.3.2 as well as the estimates $\widehat{\mathrm{p}} \pm \sigma$. Once again, the MLE estimate of p is the minimizer of this function.



Example 1.3.4
Use the formula for the variance of MLE to approximate the uncertainty associated with the MLE estimate for λ from example 1.3.1.

Solution
We start by writing down the negative log-likelihood function using the result from the equation in example 1.3.1,

$$n\ell\ell(\lambda) = 10\lambda - 37\log\lambda + 13\log(3.87)$$

The derivatives are

$$n\ell\ell'(\lambda) = -\frac{37}{\lambda}$$

and

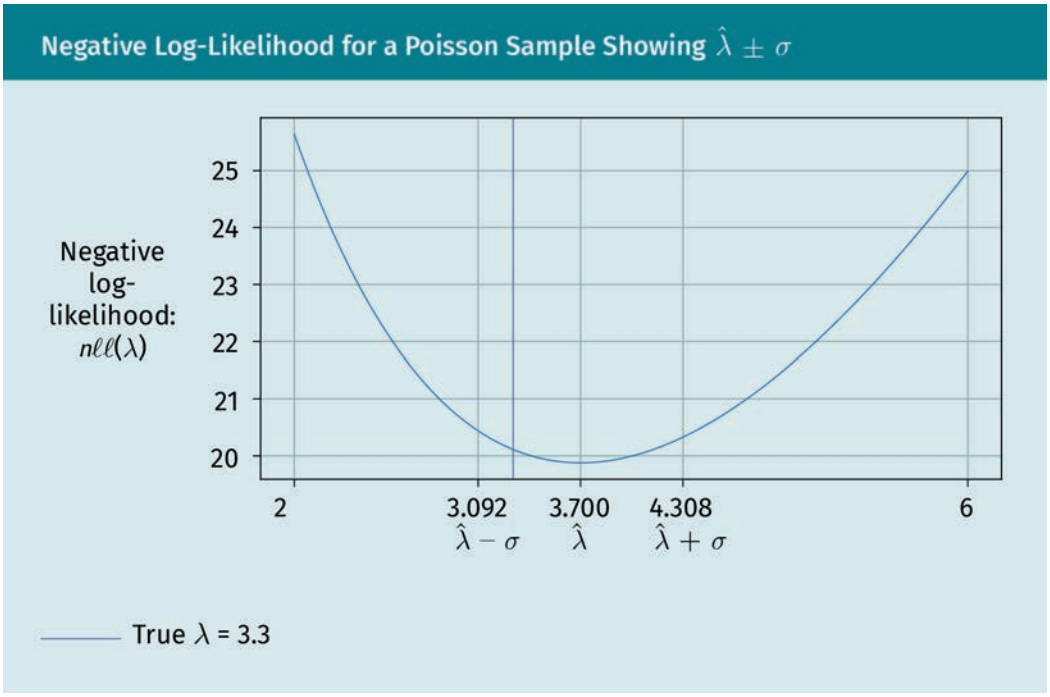
$$n\ell\ell''\lambda = \frac{37}{\lambda^2}$$

Now, applying the equation for the variance of MLE, we have

$$\mathbb{V}\left[\tilde{\lambda}^{(\mathrm{MLE})}\right] \approx \frac{1}{37/3.7^2} = 0.37$$

Therefore, $\sigma = \sqrt{3.7} \approx 0.608$

The figure below shows the negative log-likelihood plot. The MLE estimate is the minimizer of this function. The figure also shows a range over estimates within one standard deviation of the MLE.



1.4 Ordinary Least Squares

The method of maximum likelihood requires knowledge about the underlying distribution that generated the sample data. This distribution is used to find the maximum likelihood estimate of the unknown parameter of interest. In contrast, the ordinary least squares method doesn't require that we know this underlying distribution; instead, we need to know the model that generated the data.

A manufacturing facility fabricates aluminum cans. A new camera based measurement system is to be deployed to measure the diameter of the produced cans for quality assurance. Before deploying this system into production, a test is designed to measure the accuracy of the camera (image)-based measurement system. A single aluminum can with a diameter of five centimenters is used during this test. The measurement system is used to measure the diameter of this can in various conditions.

Ten measurements from the camera system are recorded $\{y_1, ..., y_{10}\} = \{5.13, 5.07, 4.85, 5.00, 5.06, 4.93, 5.03, 5.01, 5.00, 4.98\}$. In order to assess the quality of this system, we compute the difference (residuals) between these measurements and the known true value of $y = 5$. To penalize larger residuals more than smaller residuals, we compute the squares of the residuals: $(y - y_i)^2$. Finally, we sum these squared residuals to get a measure of the quality of the camera-based measurement system. The computations are shown in the table below.

Observed Measurements, Residuals, and Squared-Residuals		
y_i	$y - y_i$	$(y - y_i)^2$
5.13	−0.13	0.0169
5.07	−0.07	0.0049
4.85	0.15	0.0225
5.00	0.00	0.0000
5.06	−0.06	0.0036
4.93	0.07	0.0049
5.03	−0.03	0.0009
5.01	−0.01	0.0001

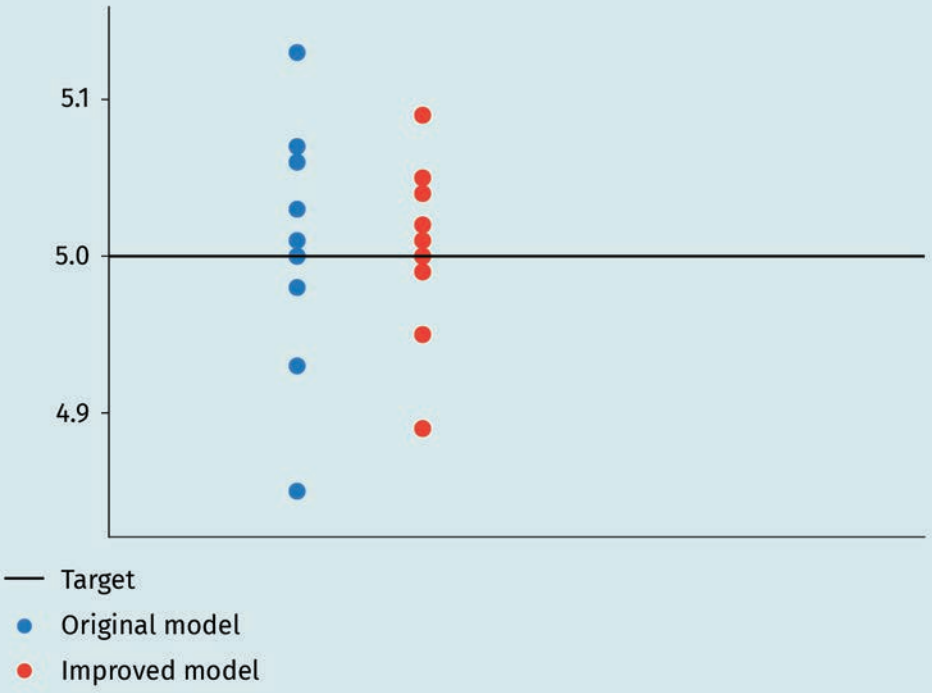
Observed Measurements, Residuals, and Squared-Residuals		
5.00	0.00	0.0000
4.98	0.02	0.0004

The sum of squared residuals is $\sum_{i=1}^{10}(y-y_i)^2 = 0.0542$. In the method of ordinary least squares, our aim to is tweak the model over and over again, until y_i's give us a lowest sum-of-squares. In practice, the model is based on parameters that we need to estimate. The way we tweak our model is to adjust the parameters in a way that reduces the sum-of-squares of the residuals.

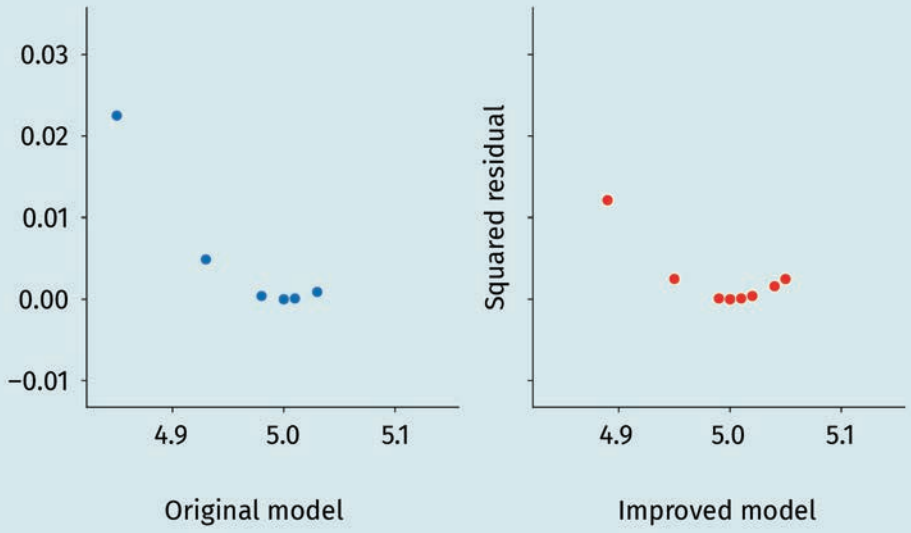
Let's say that for our camera example we tweak some of these parameters and get new measurements$y_1^{(new)}, ..., y_n^{(new)} = \{5.095.054.895.5.044.955.025.015.4.99\}$. If we repeat the computations with this new set of measurements, we get $\sum_{i=1}^{10}\left(y-y_i^{(new)}\right)^2 = 0.0274$. Since the new sum of squares is lower than the original one, we have improved our model. If this is the best this model can achieve, in other words, if any change to the model's parameters will always result in a higher sum-of-squares, then this model is said to be the least-squares solution.

Explore the the two figures below which show the measurements and squared-residuals, respectively. In the former figure, each point represents a prediction (estimate) for the diameter of the detected can. The values of the estimates are on the vertical axis. Estimates are plots in two different colors, where each color is a different model. Notice that the “new” model has estimates that are less scattered around the target value of 5.0 compared to the other model. In the latter figure, each point from the graphs represents a prediction where the vertical axis measures the squared residual and the horizontal axis the predicted value. The “new” model has smaller squared residuals, tightly clustered near zero, compared to the other model.

Measurements from Two Models



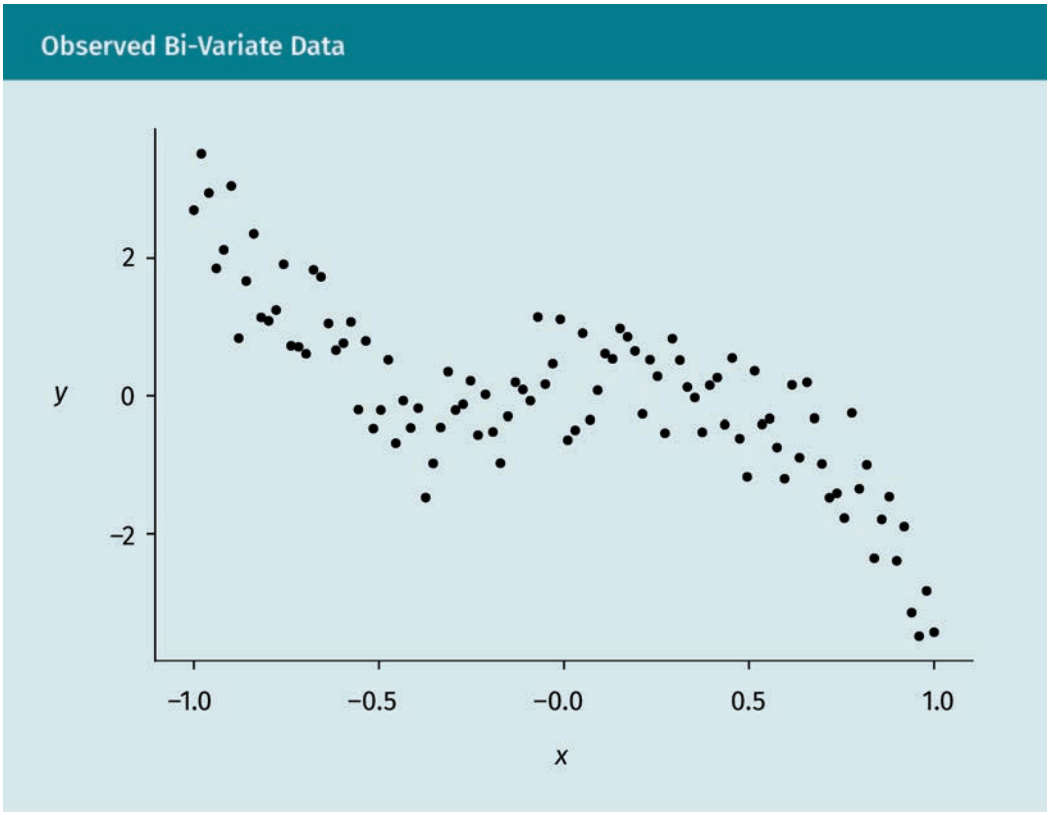
Squared-Residuals from Two Models



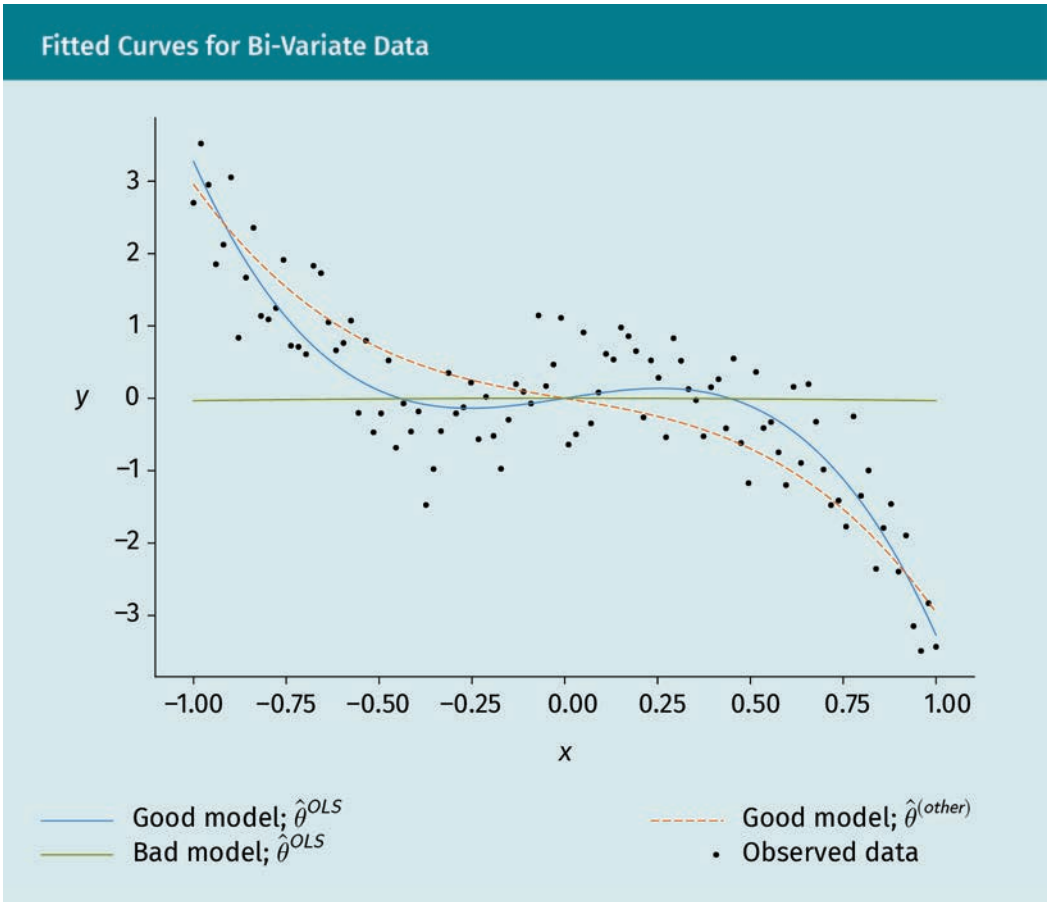
Now suppose that we have observed $\{(x_1, y_1), ..., (x_n, y_n)\}$. Assume that a functional dependence between x_i and y_i exists such that $y_i = f(x_i|\theta)$, where θ is/are the parameter(s) that completely determines this function. The least-squares estimate $\widehat{\theta}^{(OLS)}$ minimizes the sum-of-square residuals:

$$C(\theta) = \sum_{i=1}^{n} (y_i - f(x_i|\theta))^2$$

In other words, $\widehat{\theta}^{(OLS)} = \min_\theta C(\theta)$. The OLS method will produce a good result assuming that the functional dependence or the model being used is appropriate for the data. If so, then the OLS estimate of the parameter, which determines the best function or the best model, would serve the purposes well. However, if the functional or model dependence is wrong to begin with, then even the OLS estimate for the parameters would produce a useless result. The figure entitled "Observed Bi-Variate Data" shows an observed dataset of 100 points. The observed data has two variables, x and y, plotted on the graph as the horizontal and vertical values respectively. The goal is to fit a curve to this data. In the figure entitled "Cost Functions for Two Models," we show two graphs from two cost functions, one from a good model assumption and the other from a bad model assumption. Notice that the cost function from the good model assumption has values lower than the lowest cost value from the bad model assumption. The OLS fit for each of these models comes from finding the value of the parameter θ (on the horizontal axis), which minimizes the cost $C(\theta)$ (on the vertical axis). Finally, in the figure entitled "Fitted Curves for Bi-Variate Data," we show three curves fitted to the data. The best fitted curve comes from a good model assumption and the OLS estimate of θ. The second best model comes from the good model and some other estimate of θ (not the OLS). The worst fitted curve comes from a bad model assumption, even though we chose the OLS estimate for θ. Therefore, model assumption is very important when choosing to use OLS to find the point estimates of unknown parameters.


Observed Bi-Variate Data
y
2
0
−2
−1.0
−0.5
−0.0
0.5
1.0
x

Cost Functions for Two Models



1.5 Re-Sampling Techniques

In all of the previous sections, we learned about different ways to estimate an unknown parameter of the population using an observed sample. When the sample size is large, the central limit theorem provides a way to measure the uncertainty associated with a parameter estimate. If the sample size is not large, but the distribution of the population from which the sample was drawn is known (except for the unknown parameter), then we can use properties of that distribution to measure the uncertainty associated with our point estimate. But what if the sample size is too small to apply the central limit theorem, and we have no idea about the distribution family of the parent population? Re-sampling techniques were introduced to address this scenario.

Re-sampling techniques are parameter estimation techniques that use sub-samples of an observed sample. In this section, we will discuss the bootstrap and jackknife methods in some detail.

The Bootstrap

Denote an observed sample by $\mathbf{x} = (x_1, ..., x_n)$. The bootstrap re-sampling procedure chooses n numbers for the observed sample one after the other and with replacement. For example, if the observed sample is $\{1, 2, 3, 4, 5\}$, then a bootstrap sample could be $\{5, 1, 1, 3, 2\}$. This procedure is repeated a fixed number of times, which is set in advance. The table below shows an observed sample along with ten bootstrap samples.

A Random Sample of Five Numbers and Ten Bootstrap Samples					
$\mathbf{x}$	7.0	2.9	2.3	5.5	7.2
$\mathbf{x}^{(1)}$	2.3	7.2	2.3	2.9	5.5
$\mathbf{x}^{(2)}$	2.3	5.5	2.9	2.9	7.0
$\mathbf{x}^{(3)}$	2.9	2.9	7.0	7.0	2.9
$\mathbf{x}^{(4)}$	5.5	7.2	7.0	7.0	7.2
$\mathbf{x}^{(5)}$	2.9	5.5	2.3	7.2	2.3
$\mathbf{x}^{(6)}$	7.2	7.0	7.0	2.9	5.5
$\mathbf{x}^{(7)}$	7.2	7.2	7.2	2.9	5.5
$\mathbf{x}^{(8)}$	2.3	2.9	7.2	7.0	5.5
$\mathbf{x}^{(9)}$	2.3	7.0	5.5	2.3	2.3
$\mathbf{x}^{(10)}$	2.3	2.3	7.2	5.5	5.5

Say we want to compute the mean of the population from which the data were observed. One point estimate would be the sample mean using the observed sample

$$\overline{x} = \frac{7.0+2.9+2.3+5.5+7.2}{5} = 4.98$$

To estimate using the bootstrap method, we would compute the mean of each of the bootstrap samples (results shown in table below) and then compute the mean of these means.

$$\overline{\mathrm{x}}^{(\mathrm{boot})} = \frac{1}{10}\sum_{\mathrm{b}=1}^{10}\overline{\mathrm{x}}^{(\mathrm{b})} = \frac{4.04+4.12+\cdots+4.56}{10} = 4.886$$

Sample Means of the Bootstrap Samples from the Previous Table

$\overline{\mathrm{x}}^{(1)}$	$\overline{\mathrm{x}}^{(2)}$	$\overline{\mathrm{x}}^{(3)}$	$\overline{\mathrm{x}}^{(4)}$	$\overline{\mathrm{x}}^{(5)}$	$\overline{\mathrm{x}}^{(6)}$	$\overline{\mathrm{x}}^{(7)}$	$\overline{\mathrm{x}}^{(8)}$	$\overline{\mathrm{x}}^{(9)}$	$\overline{\mathrm{x}}^{(10)}$
4.04	4.12	4.54	6.78	4.04	5.92	6.00	4.98	3.88	4.56

The benefit of having these (ten) estimates, is that we can use them to figure out the uncertainty associated with the final estimate $\overline{\mathrm{x}}^{(\mathrm{boot})}$. In our case, we compute the squared standard error of the bootstrap estimate as follows:

$$\begin{aligned}\mathrm{SE}(\overline{\mathrm{x}})^2_{\mathrm{boot}} &= \frac{1}{10}\sum_{\mathrm{b}=1}^{10}\left(\overline{\mathrm{x}}^{(\mathrm{b})}-\overline{\mathrm{x}}\right)^2 \\ &= \frac{(4.04-4.98)^2+(4.12-4.98)^2+\cdots(4.56-4.98)^2}{10} \\ &= \ = 0.925\end{aligned}$$

equivalently, we have $\mathrm{SE}(\overline{\mathrm{x}})_{\mathrm{boot}} = \sqrt{0.925} = 0.9618$.

Here is a summary of the general approach using an observed sample given by $\mathbf{x} = \{\mathrm{x}_1, ..., \mathrm{x}_\mathrm{n}\}$ using B bootstrap samples. First, compute the point estimate of the target parameter θ and denote it by $\widehat{\theta}$. Next, for each b from 1 to B, compute the bootstrap point estimates $\widehat{\theta}^{(\mathrm{b})}$ from $\mathbf{x}^{(\mathrm{b})} = \left\{\mathrm{x}_{\mathrm{i}_1}, \mathrm{x}_{\mathrm{i}_2}, ..., \mathrm{x}_{\mathrm{i}_\mathrm{n}}\right\}$ where i_j are drawn from $\{1, ..., \mathrm{n}\}$ with equal probability and with replacement. Next, compute the uncertainty of the original point estimate using the bootstrap estimates as the standard error using the formula

$$\mathrm{SE}\left(\widehat{\theta}\right)_{\mathrm{boot}} = \sqrt{\frac{1}{\mathrm{B}}\sum_{\mathrm{b}=1}^{\mathrm{B}}\left(\widehat{\theta}^{(\mathrm{b})}-\widehat{\theta}\right)^2}$$

For a random sample of $\mathrm{X}_1, ..., \mathrm{X}_{10}$, the **ordered statistics** are given by $\mathrm{X}_{(1)}, ..., \mathrm{X}_{(10)}$ where $\mathrm{X}_{(1)} \leq \mathrm{X}_{(2)} \leq \cdots \leq \mathrm{X}_{(10)}$. The maximum is thus, $\mathrm{X}_{\max} = \mathrm{X}_{(10)}$, and the minimum is $\mathrm{X}_{\min} = \mathrm{X}_{(1)}$.

Ordered statistics
This is an ordering of random variables in non-descending order.

Example 1.5.1
Use the observed sample together with the ten bootstrap samples to find the bootstrap estimate of the maximum, $\mathrm{X}_{\max} = \mathrm{X}_{(5)}$.

Solution
First we compute the point estimate from the observed sample: $\widehat{\mathrm{x}}_{\max} = 7.2$. Next, we compute the point estimates from the ten bootstrap estimates. These are just the maximums from each of the samples (see table below).

Sample Max of the Bootstrap Samples									
7.20	7.00	7.00	7.20	7.20	7.20	7.20	7.20	7.00	7.20

The standard error via bootstrap is given by the equation below:

$$\begin{aligned}\mathrm{SE}(\widehat{\mathrm{x}}_{\max})_{\mathrm{boot}} &= \sqrt{\frac{1}{10}\sum_{\mathrm{b}=1}^{10}\left(\widehat{\mathrm{x}}_{\max}^{(\mathrm{b})} - \widehat{\mathrm{x}}_{\max}\right)^2} \\ &= \sqrt{\frac{1}{10}\left[(7.20-7.20)^2 + (7.00-7.20)^2 + \cdots + (7.20-7.20)^2\right]} \\ &= 0.7733\end{aligned}$$

In practice, $\mathrm{B} = 10$ bootstrap samples might not be sufficient. On the other hand, a sample of size 12 would have a total of about 1.3 million bootstrap sub-samples which makes using all possible bootstrap samples computationally infeasible. Usually, between $\mathrm{B} = 30$ to $\mathrm{B} = 100$ suffice for practical applications.

The Jackknife

Delete-1 jackknife replicate
This sub-sample is obtained by removed one value from a given sample.

Given an observed sample $\mathbf{x} = \{\mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_\mathrm{n}\}$, the **delete-1 jackknife replicate** is a sub-sample of the observed data with one data point deleted. For example, $\mathbf{x}_{(-1)} = \{\mathrm{x}_2, ..., \mathrm{x}_\mathrm{n}\}$ is one of these jackknife replicates and $\mathbf{x}_{(-3)} = \{\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_4, ..., \mathrm{x}_\mathrm{n}\}$ is another one. There are a total of n jackknife replicates:

$$\mathbf{x}_{(-\mathrm{i})} = \{\mathrm{x}_1, ..., \mathrm{x}_{\mathrm{i}-1}, \mathrm{x}_{\mathrm{i}+1}, ..., \mathrm{x}_\mathrm{n}\}, \text{ for } \mathrm{i} = 1, ..., \mathrm{n}$$

Suppose that $\mathrm{g}: \mathbb{R}^\mathrm{n} \to \mathbb{R}$ is a function that computes a statistic $\widehat{\theta} = \mathrm{g}(\mathbf{x}) = \mathrm{g}(\mathrm{x}_1, ..., \mathrm{x}_\mathrm{n})$ to estimate an an unknown parameter θ. The corresponding function $\widetilde{\mathrm{g}}: \mathbb{R}^{\mathrm{n}-1} \to \mathbb{R}$ computes an estimate $\widehat{\theta}_{(-\mathrm{i})}$ by applying $\widetilde{\mathrm{g}}$ to a jackknife replicate $\mathbf{x}_{(-\mathrm{i})}$; $\widehat{\theta}_{(-\mathrm{i})} = \widetilde{\mathrm{g}}(\mathbf{x}_{(-\mathrm{i})}) = \widetilde{\mathrm{g}}(\mathrm{x}_1, ..., \mathrm{x}_{\mathrm{i}-1}, \mathrm{x}_{\mathrm{i}+1}, ..., \mathrm{x}_\mathrm{n})$.

Suppose that $\mathrm{g}(\cdot)$ computes the sample mean, then $\widehat{\theta} = \frac{1}{\mathrm{n}}(\mathrm{x}_1 + ..., \mathrm{x}_\mathrm{n})$. The corresponding function $\widetilde{\mathrm{g}}(\cdot)$ computes the sample mean on the $\mathrm{n}-1$ points from a jackknife replicate: $\theta_{(-1)} = \widetilde{\mathrm{g}}(\mathbf{x}_{(-1)}) = \frac{1}{\mathrm{n}-1}(\mathrm{x}_2 + ..., \mathrm{x}_\mathrm{n})$. The jackknife estimate of θ is the sample mean of all the jackknife estimates:

$$\widehat{\theta}_{(\,\cdot\,)} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\theta}_{(-i)}$$

Example 1.5.2
Given the observed data $\{3, 7, 1, 0, 4\}$ from $X_1, ..., X_5$ iid copies of X, estimate $\theta = \mathbb{E}\left[X^2\right]$ with the jackknife method.

Solution
Set $\mathbf{x} = \{3, 7, 1, 0, 4\}$. The sample moment via this sample gives an estimate of θ:

$$\widehat{\theta} = \frac{1}{5}\left(3^2 + 7^2 + 1^2 + 0^2 + 4^2\right) = 15$$

The five jackknife replicates and their corresponding estimates are

$$\begin{aligned}
\mathbf{x}_{(-1)} &= \Big\{7, 1, 0, 4\Big\}\widehat{\theta}_{(-1)} = \frac{1}{4}\left(7^2 + 1^2 + 0^2 + 4^2\right) = 13.2\\
\mathbf{x}_{(-2)} &= \Big\{3, 1, 0, 4\Big\}\widehat{\theta}_{(-2)} = \frac{1}{4}\left(3^2 + 1^2 + 0^2 + 4^2\right) = 5.2\\
\mathbf{x}_{(-3)} &= \Big\{3, 7, 0, 4\Big\}\widehat{\theta}_{(-3)} = \frac{1}{4}\left(3^2 + 7^2 + 0^2 + 4^2\right) = 14.8\\
\mathbf{x}_{(-4)} &= \Big\{3, 7, 1, 4\Big\}\widehat{\theta}_{(-4)} = \frac{1}{4}\left(3^2 + 7^2 + 1^2 + 4^2\right) = 15\\
\mathbf{x}_{(-5)} &= \Big\{3, 7, 1, 0\Big\}\widehat{\theta}_{(-5)} = \frac{1}{4}\left(3^2 + 7^2 + 1^2 + 0^2\right) = 11.8
\end{aligned}$$

Finally, the jackknife estimate is given by

$$\begin{aligned}
&\widehat{\theta}_{(\,\cdot\,)} = \frac{1}{5}\left(\widehat{\theta}_{(-1)} + ... + \widehat{\theta}_{(-5)}\right)\\
&= \frac{1}{5}(13.2{+}5.2 + 14.8 + 15 + 11.8) = 12
\end{aligned}$$

The benefit of the jackknife re-sampling method is that we are able to obtain an estimate of the uncertainty of the jackknife estimate, something we aren't able to do without more knowledge of the distribution from which the data was sampled. We measure the uncertainty associated with the jackknife estimate, $\widehat{\theta}_{(\mathrm{cot})}$ using the jackknife standard error given by

$$\mathrm{SE}\big(\widehat{\theta}\big)_{\mathrm{jack}} = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}\left(\widehat{\theta}_{(-i)} - \widehat{\theta}_{(\,\cdot\,)}\right)^2}$$

Example 1.5.3
Use the formula above for the jackknife standard error to compute the standard error of the jackknife estimate found in example 1.5.2.

Solution
Here we have $n = 5$, $\widehat{\theta}_{(-1)} = 13.2, \widehat{\theta}_{(-2)} = 5.2, \widehat{\theta}_{(-3)} = 14.8, \widehat{\theta}_{(-4)} = 15, \widehat{\theta}_{(-5)} = 11.8$, and $\widehat{\theta}_{(\cdot)} = 12$:

$$\mathrm{SE}\big(\widehat{\theta}\big)_{\mathrm{jack}} = \sqrt{\frac{4}{5}\big[(13.2-12)^2 + \cdots + (11.8-12)^2\big]}$$

we give $\mathrm{SE}\big(\widehat{\theta}\big)_{\mathrm{jack}} = 7.187$

Summary

In this unit we introduced different methods for estimating parameters: method of moments, method of maximum likelihood, ordinary least squares (OLS), and re-sampling methods. The method of moments method relates the moments of a distribution with the parameter of interest and then uses the sample moments to find an estimator. Recall that the k^{th} moment of a random variable X is $\mu^{(k)} = \mathbb{E}\big[X^k\big]$ and when we have a sequence of copies, $X_1, ..., X_n$, the sample moment estimator is $\widetilde{m}^{(k)} = \frac{1}{n}\sum_{i=1}^{n} X_i^k$. The corresponding estimate with observed data $x_1, ..., x_n$ is $\widehat{m}^{(k)} = \frac{1}{n}\sum_{i=1}^{n} x_i^k$. Suppose that we want to estimate an unknown parameter θ using the method of moments. We find a function f which relates the moments $\mu^{(k)}$ to θ, $\theta = f\big(\mu^{(1)}, \mu^{(2)}, ..., \mu^{(K)}\big)$ and then the estimator is given by $\widetilde{\theta} = f\big(\widetilde{m}^{(1)}, \widetilde{m}^{(2)}, ..., \widetilde{m}^{(K)}\big)$. Finally, the estimate based on observed data is given by $\widehat{\theta} = f\big(\widehat{m}^{(1)}, \widehat{m}^{(2)}, ..., \widehat{m}^{(K)}\big)$.

One important property of any point estimator is whether or not it is unbiased: $\mathbb{E}\big[\widetilde{\theta}\big] = \theta$. We showed that some method of moment estimates are unbiased while others aren't. For example, the sample mean estimator $\overline{X}$ is an unbiased estimator for the mean (μ) of a Gaussian distribution. But, the method of moments estimator for the variance σ^2 of a Gaussian distribution is not unbiased when the the mean is unknown.

A statistic is a random variable and is a function of a random sample $X_1, ..., X_n$. In our context, a statistic is an estimator for an unknown parameter. If this statistic encompasses all the information that the random sample has to offer for estimating the unknown parameter, it is called a sufficient statistic. We looked at the definition of a sufficient statistic based on the joint distribution of the sample conditioned on the statistic being free of the parameter. A more practical way of determining sufficiency was through the likelihood factorization criterion.

The method of maximum likelihood uses the distribution family of a random sample to derive a point estimator as the maximizer of the likelihood function. When the data are independent, the likelihood function is the product of the densities (or PMS) of the random variables in the sample. This is often the case in practice. The log-likelihood is the function that is used because it provides algebraic and computational efficiency. Finally, the negative log-likelihood is used when using numer-

ical (computer) algorithms to find the MLE estimates. As for the method of moments, some MLE estimators are unbiased while others aren't. One of the nice properties of MLE estimators (in contrast to method of moment estimators) is that they are always sufficient statistics or functions of all sufficient statistics. In other words, when we get an MLE, we know we have summarized all the information the random sample has to offer with regards to the unknown parameter of interest.

In contrast with maximum likelihood, the ordinary least squares (OLS) method of estimating parameters doesn't require knowledge about the underlying distribution that produced the random sample. The only requirement for OLS is the functional/model dependence within the random sample. The OLS estimates of the parameters of interest is obtained by minimizing the sum of squared-residuals.

In the last section, we discussed two re-sampling methods: the bootstrap and the jackknife. Both of these methods introduce a way of computing uncertainty of a point estimate when the sample size is not sufficient to use the large-sample theory using the central limit theorem or when the distribution of the underlying (population) is unknown.

Knowledge Check

Did you understand this unit?

You can check your understanding by completing the questions for this unit on the learning platform.

Good luck!

Unit 2

Uncertainties

STUDY GOALS

On completion of this unit, you will have learned...

- ... how to classify types of uncertainty as statistical or systematic.
- ... the two common measures of uncertainty: variance and standard deviation.
- ... how to compute the variance of linear functions of random variables.
- ... how to approximate the variance of nonlinear functions of random variables.

2. Uncertainties

Introduction

A table top is to be manufactured with target dimensions $300\ \mathrm{cm}$ long by $100\ \mathrm{cm}$ wide. The carpenter will glue together five boards of various lengths each of which are X wide with $\mathbb{E}[\mathrm{X}] = 20\ \mathrm{cm}$ and $\mathbb{V}[\mathrm{X}] = 0.5\mathrm{cm}^2$. After the table top cures, they will cut the two ends of the table top so that the final length is Y with $\mathbb{E}[\mathrm{Y}] = 300\ \mathrm{cm}$ and $\mathbb{V}[\mathrm{Y}] = 1\ \mathrm{cm}^2$. Suppose that we are interested in modeling the area of a table. We can assume that each plank has length $\mathrm{X}_1, \mathrm{X}_2, ..., \mathrm{X}_5$ iid with $\mathbb{E}[\mathrm{X_i}] = 20$ and $\mathbb{V}[\mathrm{X_i}] = 0.5$ for $\mathrm{i} = 1, ..., 5$ and length Y, independent of $\mathrm{X_i}$. (Recall that iid stands for "independently and identically distributed.") Thus, the area $\mathrm{Z} = (\mathrm{X}_1 + \mathrm{X}_2 + \cdots + \mathrm{X}_5)\mathrm{Y}$. The expected area is $\mathbb{E}[\mathrm{Z}] = 30,000\mathrm{cm}^2$. What about the variance? The uncertainties (variances) of the $\mathrm{X_i}$'s and of Y will certainly have something to contribute to the uncertainty in the area. In this unit, we will learn how to quantify uncertainties and how they propagate through functions of random variables.

Additionally, we learn how to quantify uncertainties associated with individual quantities using variance and standard deviation. In the first section, we will learn to identify uncertainties as statistical or systematic. We will discuss examples of each and also how to report them. In particular, we will highlight the difference between a systematic mistake and a systematic uncertainty.

In section 2.2, we will evaluate the properties of variance and use these properties to learn uncertainty propagation formulas in some simple cases (i.e., linear functions of random variables). We will see that when the underlying variables are independent, the propagation of errors of the random variables to the linear functions is quite straightforward. However, in practice, the underlying random variables we work with may not be independent, they may be correlated. We will revisit the definition of covariance, which measures how two random variables are dependent, and use this quantity together with the individual uncertainties in the variables to propagate the uncertainties to a linear function. In the case that the quantity of interest is a nonlinear function of random variables, simple formulas are no longer possible. Furthermore, even if we opt for complicated formulas, we won't be able to use them in practice using only the variance and covariance quantities of the underlying random variables. Therefore, we will learn how to approximate the uncertainties associated with functions of random variables.

In statistical inference, we will use statistics collected from a finite sample to inform us about claims made about the data. The formulas we use to compute these statistics are functions of random variables. We need to understand how much uncertainty is associated with the statistics we compute without having to repeat the experiment many times. The results in this unit will inform such uncertainties of the statistics.

2.1 Statistical and Systematic Uncertainties

Statistical inference involves data, which come about various methods such as counting unique website visitors, weighing a piece of jewelry, or measuring the length of a specimen using a steel ruler. When repeated measurements are taken, in an ideal world, we expect the same result. In practice, however, this isn't the case. When measurements are repeated, we often get different results. If a specimen is stored properly and a researcher measures the length using a steel ruler in the year 2008, and the same specimen is measured by the same researcher using the same ruler in 2010, they may get different results due to the expansion of the ruler due to changes in temperature. If the temperatures weren't recorded when the original measurements took place, there would be no way to correct the measurement error. This is an example where the measurement has **systematic uncertainty**. For example, if there is expansion in the ruler, then the new measurements taken would be systematically shorter than their true length. If we can correct errors due to this effect, we would have no systematic uncertainty, in practice it is very difficult to detect systematic uncertainties.

Systematic uncertainty
These are uncertainties in measurement estimates that are not statistical.

There are, however, instances where the errors are due to systematic mistakes rather than systematic uncertainty. Suppose we have a measurement device that was calibrated at a specific temperature because it is known that changes in temperature will affect the measurement of a specimen. Also assume that there is no way to re-calibrate this device. If we use this device in an environment where the temperature is drastically different, and either don't know of this effect or just ignore the effect, then the measurements we obtain will be systematically different (e.g., higher) than the true values. This is an example of a systematic mistake. On the other hand, if we are aware of the effect and adjust the measurements (e.g., by subtracting some fixed or proportional quantity) based on the effect predicted by the laws of nature, then we would have corrected the systematic mistake and there would be no systematic uncertainty (or mistake). Yet another possibility is that we are aware of the effect but aren't certain of the temperature at which the device was originally calibrated. Since we have no way of calibrating, we can't correct for the systematic differences in the measurements we obtain. In this third scenario, we would have a systematic uncertainty associated with the measurements.

Now suppose that we are measuring the weight of a piece of jewelry, say a gold ring. We place it on our digital scale, correctly calibrated (so as to avoid systematic uncertainty), but with three repeated measurements we get 14.25, 14.23, 14.28 grams respectively. We can quantify the uncertainty by the sample standard variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n-1}(x_i - \overline{x})^2 = 0.000633333$$

The uncertainty is about 0.0006 squared-grams. The sample standard deviation is an equivalent measure, but in the same units as the measurements (data)

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} = 0.0251661$$

Thus, the uncertainty is about 0.0252 grams. It is common to write $\overline{x} \pm s = 14.2533 \pm 0.0252$. Unlike systematic errors, statistical errors can be reduced by collecting more data (measuring the ring more times). If a Gaussian (normal distribution) can be used, the uncertainty (standard deviation) is reduced proportionally to $1/\sqrt{n}$. The different values we got for the weight of this ring are due to the inherent randomness associated with the ring or the scale. Perhaps small particles that have some weight are on top of the scale or attached to the ring. Uncertainties due to inherent randomness in the data itself is called **statistical uncertainty**.

Statistical uncertainty
This type of uncertainty is associated with the underlying randomness in the measurement process or the quantity being measured. It can be reduce by collecting more data.

If the digital scale was not properly calibrated, then the experiment of repeatedly measuring the ring would be subject to both systematic and statistical uncertainties. In this case, we could summarize our estimate as follows:

$$\text{estimate} \pm (\text{statistical uncertainty}) + \pm (\text{systematic uncertainty})$$

In general, systematic uncertainties can be reduced by acquiring more knowledge and incorporating this knowledge in the measurement process. Statistical uncertainties can be reduced by acquiring more data.

Statistical uncertainties are defined by the variance (and standard deviation). Let X be a random variable, its variance and standard deviation are defined by

$$\mathbb{V}[X] := \mathbb{E}\Big[(X - \mathbb{E}[X])^2\Big] = \mathbb{E}\Big[X^2\Big] - \mathbb{E}[X]^2$$

$$\mathrm{SD}[X] := = \sqrt{\mathbb{V}[X]}$$

In the next section, we will discuss uncertainties of random variables that are based on two or more other random variables. If the latter random variables have some dependence we need to take this dependence into account. The standard way of measuring dependence of two random variables is by their covariance. Let X_1 and X_2 be two random variables, whose covariance is defined by

$$\mathrm{Cov}(X_1, X_2) := \mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_2 - \mathbb{E}[X_2])] = \mathrm{Cov}(X_2, X_1)$$

2.2 Propagation of Uncertainties

Suppose that a random variable Y represents the quantity we want to measure but can't. Instead, we can measure X and know how X and Y are related. So if we have a value of X, then we can calculate the corresponding value for Y. Since they are associated with X, there will be uncertainties associated with Y. In this section, we want to understand how these uncertainties are propagated from X to Y.

Suppose long railing is manufactured in two parts and is to be joined on-site. Let X_1 and X_2 denoted the lengths of the two parts. $Y = X_1 + X_2$ will be the total length of the railing after installation. We would like to estimate the uncertainty of Y during the manufacturing process. As such, we will take measurements of the two parts, compute the uncertainties associated with them, and then figure out how these uncertainties propagate to Y, the total length.

Linear Functions

Let's start with some basic notation. Let X be a random variable. We will denote the variance and standard deviation of X by $\sigma_X^2 = \mathbb{V}[X]$ and $\sigma_X = \sqrt{\mathbb{V}[X]}$ respectively. Adding a non-random (constant) quantity to a random variable doesn't change its variance: $\mathbb{V}[X + c] = \mathbb{V}[X]$.

For two random variables X_1 and X_2, their covariance is denoted by $\sigma_{12} = \mathrm{Cov}(X_1, X_2)$. Note that $\sigma_{12} = \sigma_{21} = \mathrm{Cov}(X_2, X_1)$. Recall the following formulas from probability theory for the variance of the sum/difference of two random variables X_1 and X_2:

$$\mathbb{V}[X_1 \pm X_2] = \sigma_1^2 + \sigma_2^2 \pm 2\sigma_{12}$$

If the covariance is zero, which happens when the variables are independent, then the formula reduces to

$$\mathbb{V}[X_1 \pm X_2] = \sigma_1^2 + \sigma_2^2$$

Now recall the formula for the variance of a non-random multiple (a) of a random variable X:

$$\mathbb{V}[aX] = a^2\mathbb{V}[X] = a^2\sigma_X^2$$

The covariance is linear in both of its arguments; in other words, for random variables X_1 and X_2 and scalars a and b, we have

$$\mathrm{Cov}(aX_1, bX_2) = ab\mathrm{Cov}(X_1, X_2) = ab\sigma_{12}$$

If we combine the results of the equations above, we get the following formula for a variance of a linear combination of two random variables X_1 and X_2, where a_1 and a_2 are non-random scalars:

$$\mathbb{V}[a_1X_1 + a_2X_2] = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + a_1a_2\sigma_{12}$$

Once again, if the covariance is zero, $\sigma_{12} = 0$, then the formula above reduces to

$$\mathbb{V}[a_1X_1 + a_2X_2] = a_1^2\sigma_1^2 + a_2^2\sigma_2^2.$$

Example 2.2.1
Given random variables X_1 and X_2 with $\sigma_1^2 = 2$, $\sigma_2^2 = 3$, and $\sigma_{12} = -1$, compute the following variances:

1. $\mathbb{V}[X_1 + X_2]$
2. $\mathbb{V}[10X_1]$
3. $\mathbb{V}[2X_1 - 3X_2]$

Solution

1. Using a formula from above, we get $\mathbb{V}[X_1 + X_2] = \sigma_1^2 + \sigma_2^2 + \sigma_{12} = 2 + 3 - 1 = 4$
2. Using a formula from above, we get $\mathbb{V}[10X_1] = 10^2\sigma_1^2 = 100 \cdot 2 = 200$
3. Using a formula from above, we get

$$\mathbb{V}[2X_1 - 3X_2] = \mathbb{V}[2X_1 + (-3)X_2] = 2^2\sigma_1^2 + (-3)^2\sigma_2^2 + (2)(-3)\sigma_{12} = 4 \cdot 2 + 9 \cdot 3 + (-6)(-1) = 29$$

Example 2.2.2
Let $X_1, ..., X_n$ be independent random variables. Find the variance of the sample mean, $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Solution
As before, let's set $\sigma_i = \mathbb{V}[X_i]$ for $i = 1, ..., n$. First, we can use the formula for the variance of a non-random multiple from above with $a = 1/n$

$$\mathbb{V}[\overline{X}] = \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \left(\frac{1}{n}\right)^2\mathbb{V}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\mathbb{V}\left[\sum_{i=1}^{n} X_i\right]$$

Now recall that since the variables are independent, we have $\mathrm{Cov}(X_i, X_j) = 0$ for $i \neq j$. Therefore, to find the variance of the sum, we can just extend the formula from the variance formula for the sum of two random variables

$$\mathbb{V}[\overline{X}] = \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}[X_i] = \frac{1}{n^2}\sum_{i=1}^{n} \sigma_i^2$$

Example 2.2.3
Let $X_1, ..., X_n$ be iid with $\mathbb{V}[X] = \sigma^2$, i.e., they all have the same variance. What is the variance of the sample mean, $\mathbb{V}[\overline{X}]$?

Solution
We can use the result of the previous example with $\sigma_i^2 = \sigma^2$ for $i = 1, ..., n$ to get

$$\mathbb{V}[\overline{X}] = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}$$

Recall that the sample mean $\overline{X}$ is an unbiased estimator of the population mean $\mu^{(1)} = \mu$. When viewed as an estimator, the standard deviation of $\overline{X}$ is called the standard error of the sample mean. Using the result from equation variance of the sample mean estimator, we have

$$SE(\overline{X}) = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

The standard error of an estimator is the standard way of reporting the uncertainty associated with it. When the population standard deviation (σ) is unknown, we can replace σ^2 with s_n^2, the sample variance

$$SE(\overline{X}) \approx \widehat{SE}(\overline{X}) = \frac{s_n}{\sqrt{n}}$$

Nonlinear Functions

So far, we have worked with random variables $X_1, ..., X_n$, for which the uncertainties are known and computed the uncertainties of linear functions of these, i.e., $Y = X_1 + X_2$, $Y = 10X_1$, $Y = 2X_1 - 3X_2$, $Y = \overline{X}$. Often, the quantity we are interested in may not be a linear function, i.e., $Y = f(X_1, X_2)$, where f is nonlinear. Computing the uncertainties of Y in such cases is extremely difficult. Instead, we use formulas that aim to approximate $\mathbb{V}[Y]$. To illustrate this difficulty, we will compute the variance of a product of two independent random variables and note the usual approximation formula used.

Example 2.2.4
Let X_1 and X_2 be independent random variables with variances σ_1^2 and σ_2^2, respectively. What is the variance of $Y = X_1X_2$?

Solution
From the definition of variance, we have

$$\begin{aligned}\mathbb{V}[Y] &= \mathbb{E}\left[(X_1X_2)^2\right] - \mathbb{E}[X_1X_2]^2 \\ &= \mathbb{E}\left[X_1^2\right]\mathbb{E}\left[X_2^2\right] - \mathbb{E}[X_1]^2\mathbb{E}[X_2]^2\end{aligned}$$

where the second equality comes from $\mathbb{E}[X_1X_2] = \mathbb{E}[X_1]\mathbb{E}[X_2]$, which holds since the variables are independent. Let's denote $\mu_1 = \mathbb{E}[X_1]$ and $\mu_2 = \mathbb{E}[X_2]$. Now we have $\mathbb{E}\left[X_1^2\right] = \mathbb{V}[X_1] + \mu_1^2 = \sigma_1^2 + \mu_1^2$ and $\mathbb{E}\left[X_2^2\right] = \mathbb{V}[X_1] + \mu_2^2 = \sigma_2^2 + \mu_2^2$. Substituting these results we have

$$\begin{aligned}\mathbb{V}[\mathrm{Y}] &= (\sigma_1^2+\mu_1^2)(\sigma_2^2+\mu_2^2)-\mu_1^2\mu_2^2\\ &= \sigma_1^2\sigma_2^2+\sigma_1^2\mu_2^2+\sigma_2^2\mu_1^2\end{aligned}$$

The formula for approximating the uncertainty of $\mathrm{Y}=\mathrm{X}_1\mathrm{X}_2$, where X_1 and X_2 are independent is given by

$$\mathbb{V}[\mathrm{Y}]=\mathbb{V}[\mathrm{X}_1\mathrm{X}_2]\approx\mu_2^2\sigma_1^2+\mu_1^2\sigma_2^2\ .$$

Example 2.2.5
Let X_1 and X_2 be independent with $\mu_1=2$, $\mu_2=3$, $\sigma_1=0.2$, and $\sigma_2=0.4$. Compute the variance of $\mathrm{Y}=\mathrm{X}_1\mathrm{X}_2$ using the exact formula and the approximate formula. Compare both using the relative approximation error.

Solution
We will start by computing the exact variance from the result of example 2.2.4:

$$\sigma_{\mathrm{Y}}=\mathbb{V}[\mathrm{Y}]=\big(0.2\big)^2\big(0.4\big)^2+\big(0.2^2\big)\big(3^2\big)+\big(0.4\big)^2\big(2\big)^2=1.0064$$

Using the approximating formula, which approximates the variance of a product of independent random variables, we get

$$\sigma_{\mathrm{Y}}^2\approx\big(0.2^2\big)\big(3^2\big)+\big(0.4\big)^2\big(2\big)^2=1$$

The error is $|1.0064-1|=0.0064$, and the relative error is $0.0064/1.0064\approx0.636\,\%$, which is quite small.

The general formula for approximating the variance of a product to two random variables $\mathrm{Y}=\mathrm{X}_1\mathrm{X}_2$ is

$$\sigma_{\mathrm{Y}}^2=\mathbb{V}[\mathrm{Y}]\approx\mu_2^2\sigma_1^2+\mu_1^2\sigma_2^2+2\mu_1\mu_2\sigma_{12}$$

Coefficient of variation
A quantity used to measure relative uncertainty, this is the squared ratio of the standard deviation with respect to the random variable.

In practice, the interpretation of the variance of a quantity alone is not sufficient. We want to see how it compares to the measured value. As such, instead of just examining the variance of a random variable Y using σ_{Y}^2, we want to examine the **coefficient of variation** associated with the random variable. This quantity is the ratio $(\sigma_{\mathrm{Y}}/\mathrm{Y})^2$. As such, the following formulas for various nonlinear functions of X_1 and X_2, are given in terms of the coefficient of determination:

$$\mathrm{Y}=\mathrm{X}_1\mathrm{X}_2\rightarrow\left(\frac{\sigma_{\mathrm{Y}}}{\mathrm{Y}}\right)^2\approx\left(\frac{\sigma_1}{\mathrm{X}_1}\right)^2+\left(\frac{\sigma_2}{\mathrm{X}_2}\right)^2+2\frac{\sigma_{12}}{\mathrm{X}_1\mathrm{X}_2}$$

$$\mathrm{Y}=\frac{\mathrm{X}_1}{\mathrm{X}_2}\rightarrow\ \left(\frac{\sigma_{\mathrm{Y}}}{\mathrm{Y}}\right)^2\approx\left(\frac{\sigma_1}{\mathrm{X}_1}\right)^2+\left(\frac{\sigma_2}{\mathrm{X}_2}\right)^2-2\frac{\sigma_{12}}{\mathrm{X}_1\mathrm{X}_2}$$

Of course, in practice the nonlinear functions we encounter are more diverse than just products or quotients. As such, we need a general method of propagating uncertainties. The formulas we have introduced so far use a linearization of the function given by the Taylor series. Recall that for $\mathrm{Y} = \mathrm{f}(\mathrm{X})$, the first-order (linear) Taylor approximation centered at a is given by $\mathrm{Y} \approx \mathrm{f}(\mathrm{a}) + \mathrm{f}'(\mathrm{a})(\mathrm{X} - \mathrm{a})$. For our task, we will choose $\mathrm{a} = \mu = \mathbb{E}[\mathrm{X}]$ so that $\mathrm{Y} = \mathrm{f}(\mathrm{X}) \approx \mathrm{f}(\mu) + \mathrm{f}'(\mu)(\mathrm{X} - \mu)$. Next, using the formula for the variance of a non-random multiple of a random variable, we get

$$\sigma_{\mathrm{Y}}^2 = \mathbb{V}[\mathrm{Y}] \approx (\mathrm{f}'(\mu))^2 \sigma_{\mathrm{X}}^2$$

Example 2.2.6
Let X be a random variable with mean $\mu = \mathbb{E}[\mathrm{X}]$ and variance $\sigma^2 = \sigma_{\mathrm{X}}^2$. Use the result of linearization to approximate the variance of $\mathrm{Y} = \log \mathrm{X}$.

Solution
The derivative is $\mathrm{f}'\Big(\mathrm{X}\Big) = \frac{1}{\mathrm{X}}$, so $\mathrm{f}'\Big(\mu\Big) = \frac{1}{\mu}$. Therefore, $\sigma_{\mathrm{Y}}^2 = \mathbb{V}[\mathrm{Y}] \approx \frac{\sigma^2}{\mu^2}$

Now, we consider a function of two (random) variables, $\mathrm{Y} = \mathrm{f}(\mathrm{X}_1, \mathrm{X}_2)$. The first-order (linear) Taylor approximation of f centered at the mean $(\mu_1, \mu_2) = (\mathbb{E}[\mathrm{X}_1], \mathbb{E}[\mathrm{X}_2])$, is given by

$$\mathrm{Y} = \mathrm{f}\Big(\mathrm{X}_1, \mathrm{X}_2\Big) \approx \mathrm{f}\Big(\mu_1, \mu_2\Big) + \left[\frac{\partial \mathrm{f}}{\partial \mathrm{X}_1}\Big(\mu_1, \mu_2\Big)\right]\Big(\mathrm{X}_1 - \mu_1\Big) + \left[\frac{\partial \mathrm{f}}{\partial \mathrm{X}_2}\Big(\mu_1, \mu_2\Big)\right]\Big(\mathrm{X}_2 - \mu_2\Big) \quad .$$

For the sake of brevity, we will use the shorthand

$$\partial_{\mathrm{X}_1} \colon = \frac{\partial \mathrm{f}}{\partial \mathrm{X}_1}\Big(\mu_1, \mu_2\Big)$$

and

$$\partial_{\mathrm{X}_2} \colon = \frac{\partial \mathrm{f}}{\partial \mathrm{X}_2}\Big(\mu_1, \mu_2\Big)$$

With this notation, we have

$$\mathrm{Y} = \mathrm{f}\Big(\mathrm{X}_1, \mathrm{X}_2\Big) \approx \mathrm{f}\Big(\mu_1, \mu_2\Big) + \partial_{\mathrm{X}_1}\Big(\mathrm{X}_1 - \mu_1\Big) + \partial_{\mathrm{X}_2}\Big(\mathrm{X}_2 - \mu_2\Big)$$

Using the formula for the sum of two random variables, we have

$$\sigma_{\mathrm{Y}}^2 = \mathbb{V}[\mathrm{Y}] \approx \Big[\partial_{\mathrm{X}_1}\Big]^2 \sigma_1^2 + \Big[\partial_{\mathrm{X}_2}\Big]^2 \sigma_2^2 + 2\Big[\partial_{\mathrm{X}_1}\partial_{\mathrm{X}_2}\Big]\sigma_{12}$$

Example 2.2.7
Let X_1, X_2, and $Y = f(X_1, X_2) = X_1X_2$. Use the formula from above to approximate the variance of Y.

Solution
We start by computing the partial derivatives at the mean (μ_1, μ_2)

$$\partial_{X_1} = \mu_2 \text{ and } \partial_{X_2} = \mu_1$$

Now, applying the formula, we get

$$\sigma_Y^2 = V[Y] \approx \mu_2^2\sigma_1^2 + \mu_1^2\sigma_2^2 + 2\mu_1\mu_2\sigma_{12} = \mu_1^2\mu_2^2\left[\frac{\sigma_1^2}{\mu_1^2} + \frac{\sigma_2^2}{\mu_2^2} + 2\frac{\sigma_{12}}{\mu_1\mu_2}\right]$$

Example 2.2.8
Let X_1, X_2, and $Y = f\left(X_1, X_2\right) = \frac{X_1}{X_2}$. Additionally, suppose that $\mu_2 = \mathbb{E}[X_2] > 0$. Use the formula that gives the approximate variance of a function of two random variables to approximate the variance of Y.

Solution
We start by computing the partial derivatives at the mean (μ_1, μ_2)

$$\partial_{X_1} = \frac{1}{\mu_2}$$

and

$$\partial_{X_2} = -\frac{\mu_1}{\mu_2^2}$$

Now, applying the formula, we get

$$\sigma_Y^2 = \mathbb{V}[Y] \approx \frac{1}{\mu_2^2}\sigma_1^2 + \frac{\mu_1^2}{\mu_2^4}\sigma_2^2 - 2\frac{\mu_1}{\mu_2^3}\sigma_{12} = \frac{\mu_1^2}{\mu_2^2}\left[\frac{\sigma_1^2}{\mu_1^2} + \frac{\sigma_2^2}{\mu_2^2} - 2\frac{\sigma_{12}}{\mu_1\mu_2}\right]$$

General Formulas

Let $X_1, ..., X_n$ be random variables with known variances and covariances. We denote the random vector $\mathbf{X} = (X_1, ..., X_n)$. The uncertainties can be summarized by the variance-covariance matrix given by

$$\mathbb{V}[\mathbf{X}] = \mathrm{V}_{\mathbf{X}} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \cdots & \sigma_{2n} \\ \vdots & & \ddots & & \\ \sigma_{n1} & \sigma_{n2} & \sigma_{n3} & \cdots & \sigma_n^2 \end{bmatrix}$$

The diagonal elements of this matrix contain the variances of each of the random variables and the off-diagonal elements contain the relevant covariances $\sigma_{ij} = \mathrm{Cov}(\mathrm{X_i}, \mathrm{X_j})$. Let Y be a linear function (transformation) of the X's: $\mathrm{Y} = \mathrm{a}_1\mathrm{X}_1 + \cdots \mathrm{a}_2\mathrm{X}_n$. Then, define a (row-)matrix A by

$$\mathrm{A} = \begin{bmatrix} \mathrm{a}_1 & \cdots & \mathrm{a}_n \end{bmatrix}$$

and get $\mathrm{Y} = \mathrm{A}\mathbf{X}$. The uncertainties of Y can be computed with the formula

$$\sigma_{\mathrm{Y}}^2 = \mathbb{V}[\mathrm{Y}] = \mathrm{A}\mathrm{V}_{\mathbf{X}}\mathrm{A}^{\mathrm{T}}.$$

Example 2.2.9
Suppose $\mathrm{Y} = 2\mathrm{X}_2 + 3\mathrm{X}_2$ with $\mathbf{X} = (\mathrm{X}_1, \mathrm{X}_2)$, having the variance-covariance matrix

$$\mathrm{V}_{\mathbf{X}} = \begin{bmatrix} 3 & 2 \\ 2 & 4 \end{bmatrix}$$

Use the formula above to compute $\mathbb{V}[\mathrm{Y}]$.

Solution
First, we note the matrix that gives Y:

$$\mathrm{A} = \begin{bmatrix} 2 & 3 \end{bmatrix}$$

Next, using the formula, we get

$$\mathbb{V}[\mathrm{Y}] = \begin{bmatrix} 2 & 3 \end{bmatrix}\begin{bmatrix} 3 & 2 \\ 2 & 4 \end{bmatrix}\begin{bmatrix} 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 2 & 3 \end{bmatrix}\begin{bmatrix} 12 \\ 16 \end{bmatrix} = 72$$

Sometimes, we have more than one measurement for which we need to find uncertainties. For example, if we have $\mathbf{Y} = (\mathrm{Y}_1, ..., \mathrm{Y}_m)$ and m measurements, each of which are linear functions of the $\mathrm{X_i}$'s, say

$$\begin{aligned} \mathrm{Y}_1 &= \mathrm{a}_{11}\mathrm{X}_1 + \mathrm{a}_{12}\mathrm{X}_2 + \cdots + \mathrm{a}_{1n}\mathrm{X}_n \\ \mathrm{Y}_2 &= \mathrm{a}_{21}\mathrm{X}_1 + \mathrm{a}_{22}\mathrm{X}_2 + \cdots + \mathrm{a}_{2n}\mathrm{X}_n \\ &\vdots \\ \mathrm{Y}_m &= \mathrm{a}_{m1}\mathrm{X}_1 + \mathrm{a}_{m2}\mathrm{X}_2 + \cdots + \mathrm{a}_{mn}\mathrm{X}_n \end{aligned}$$

then, we can define the matrix A by

$$\mathrm{A} = \begin{bmatrix} \mathrm{a}_{11} & \mathrm{a}_{12} & \cdots & \mathrm{a}_{1\mathrm{n}} \\ \mathrm{a}_{21} & \mathrm{a}_{22} & \cdots & \mathrm{a}_{2\mathrm{n}} \\ & & \vdots & \\ \mathrm{a}_{\mathrm{m}1} & \mathrm{a}_{\mathrm{m}2} & \cdots & \mathrm{a}_{\mathrm{mn}} \end{bmatrix}$$

and write the relation as $\mathbf{Y} = \mathrm{A}\mathbf{X}$. The uncertainties contained in Y can then be summarize by its own variance covariance matrix:$\mathbb{V}[\mathbf{Y}] = \mathrm{V}_{\mathbf{Y}} = \mathrm{A}\mathrm{V}_{\mathbf{X}}\mathrm{A}^{\mathrm{T}}$. Although the formula looks the same as before, the result is not a number but a matrix with m rows and m columns.

Example 2.2.10
Let $\mathrm{Y}_1 = \mathrm{X}_1 + 2\mathrm{X}_2$ and $\mathrm{Y}_2 = 2\mathrm{X}_1 + \mathrm{X}_2$, with $\mathbf{X} = (\mathrm{X}_1, \mathrm{X}_2)$ having the variance-covariance matrix

$$\mathrm{V}_{\mathbf{X}} = \begin{bmatrix} 10 & 4 \\ 4 & 2 \end{bmatrix}$$

Use the matrix to compute $\mathrm{V}_{\mathbf{Y}} = \mathbb{V}[\mathrm{Y}]$ for $\mathbf{Y} = (\mathrm{Y}_1, \mathrm{Y}_2)$.

Solution
As before, we define the matrix A

$$\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$$

Next, we apply the formula

$$\mathrm{V}_{\mathbf{Y}} = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}\begin{bmatrix} 10 & 4 \\ 4 & 2 \end{bmatrix}\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}\begin{bmatrix} 2 & 16 \\ 0 & -6 \end{bmatrix} = \begin{bmatrix} 2 & 4 \\ 4 & 26 \end{bmatrix}$$

In the example above, the formula produced all the uncertainties we might want to know about $\mathbf{Y}$. For instance, $\mathbb{V}[\mathrm{Y}_1] = 2$, $\mathbb{V}[\mathrm{Y}_2] = 26$, and $\mathrm{Cov}(\mathrm{Y}_1, \mathrm{Y}_2) = 4$.

Linearization of Nonlinear Functions

In the most general case, each of the random variables in $\mathbf{Y} = (\mathrm{Y}_1, ..., \mathrm{Y}_{\mathrm{m}})$ are (possibly) nonlinear functions in $\mathbf{X} = (\mathrm{X}_1, ..., \mathrm{X}_{\mathrm{n}})$. In this case, we would use the linear (first-order) Taylor approximations centered at the means $\boldsymbol{\mu} = (\mu_1, ..., \mu_{\mathrm{n}})$ of the function in $\mathbf{Y}$ to obtain the matrix

$$\mathrm{B} = \begin{bmatrix} \partial_{11} & \partial 12 & \partial_{13} & \cdots & \partial_{1\mathrm{n}} \\ \partial_{21} & \partial_{22} & \partial_{23} & \cdots & \partial_{2\mathrm{n}} \\ \vdots & & \ddots & & \\ \partial_{\mathrm{m}1} & \partial_{\mathrm{m}2} & \partial_{\mathrm{m}3} & \cdots & \partial_{\mathrm{mn}} \end{bmatrix}$$

where $\partial_{ij} := \left.\frac{\partial Y_i}{\partial X_j}\right|_{\boldsymbol{\mu}}$ for $i = 1, ..., m$ and $j = 1, ..., n$. Then, we can approximate the variance-covariance matrix $V_{\mathbf{Y}}$ with

$$V_{\mathbf{Y}} \approx B V_{\mathbf{X}} B^T$$

Example 2.2.11
Let $\mathbf{Y} = (X_1X_2, X_1/X_2)$. Given that $V_{\mathbf{X}} = \begin{bmatrix} 20 & -10 \\ -10 & 10 \end{bmatrix}$; and $\boldsymbol{\mu} = (1, 2)$ for $\mathbf{X} = (X_1, X_2)$, use the approximation from the formula above to approximate $V_{\mathbf{Y}}$.

Solution
First, we compute the elements in B: $\partial_{11} = \mu_2 = 2$, $\partial_{12} = \mu_1 = 1$, $\partial_{21} = 1/\mu_2 = 1/2$, and $\partial_{22} = -1/4$. Next, we use the approximation formula

$$V_{\mathbf{Y}} \approx \begin{bmatrix} 2 & 1 \\ \frac{1}{2} & -\frac{1}{4} \end{bmatrix} \begin{bmatrix} 20 & -10 \\ -10 & 10 \end{bmatrix} \begin{bmatrix} 2 & \frac{1}{2} \\ 1 & -\frac{1}{4} \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ \frac{1}{2} & -\frac{1}{4} \end{bmatrix} \begin{bmatrix} 30 & \frac{25}{2} \\ -10 & -\frac{15}{2} \end{bmatrix} = \begin{bmatrix} 50 & \frac{35}{2} \\ \frac{35}{2} & \frac{65}{8} \end{bmatrix}$$

Summary

In this unit, we defined and discussed two primary types of uncertainties: statistical and systematic. We learned that systematic uncertainties arise from uncorrected errors when obtaining measurements and cannot by reduced by collecting more data. Statistical uncertainties arise from the inherent randomness associated with the measured quantity and can be reduced by collecting more data (making more measurements).

Uncertainties are quantified using variance (or standard deviation) and relationships between two variables, quantified by their covariance, also plays a part. If X is a random variable, then its variance and standard deviation are defined by

$$\mathbb{V}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 . \mathrm{SD}[X] = \sqrt{\mathbb{V}[X]}$$

For two random variables X_1 and X_2, the covariance $\mathrm{Cov}(X_1, X_2)$ is defined by

$$\mathrm{Cov}(X_1, X_2) = \mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_2 - \mathbb{E}[X_2])]$$

Uncertainty propagation is of crucial importance in studying uncertainties. We are often interested in a transformed (aggregate) quantity based on measurable quantities. We can measure the uncertainties associated with the measured quantities but not the uncertainty of the quantity of interest.

In section 4.2, we discussed some formulas that compute how uncertainties in the quantities of interest (transformed quantities) are based on the uncertainties of the underlying measured quantities. In the case where the quantity of interest was a linear transformation of the underling quantities, we gave exact formulas for the uncertainty of the quantity of interest. In the case where the transformation was nonlinear, we gave a similar formula that approximates the uncertainty of the quantity of interest. Let $\mathrm{X}_1, ..., \mathrm{X}_\mathrm{n}$ be n random variables. We summarize their uncertainties and dependencies via the variance-covariance matrix V_X, where $\mathbf{X} = (\mathrm{X}_1, ..., \mathrm{X}_\mathrm{n})$. This matrix is $\mathrm{n} \times \mathrm{n}$ whose diagonal elements are the variances of the random variables and off diagonal entries contain the covariance of the respective pair (row/column number) of the random variables. For any $\mathrm{m} \times \mathrm{n}$ matrix A, the variance-covariance matrix of $\mathrm{Y} = \mathrm{AX}$, which summarizes the uncertainties and dependencies of the quantities of interest in the random vector $\mathbf{Y} = (\mathrm{Y}_1, ..., \mathrm{Y}_\mathrm{m})$, is given by

$$\mathrm{V}_\mathbf{Y} = \mathrm{A}\mathrm{V}_\mathbf{X}\mathrm{A}^\mathrm{T}$$

This is the formula for a linear transformation. If $\mathbf{Y} = \mathbf{f}(\mathbf{X})$ where $\mathrm{f}: \mathbb{R}^\mathrm{n} \to \mathbb{R}^\mathrm{m}$ is a (possibly) nonlinear function, then we define the matrix B which is the Jacobian matrix (containing the partial derivatives) evaluated at the mean vector $\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}] = (\mathbb{E}[\mathrm{X}]_1, ..., \mathbb{E}[\mathrm{X}]_\mathrm{n})$. The uncertainties in $\mathbf{Y}$ can be approximated with the equation

$$\mathrm{V}_\mathbf{Y} \approx \mathrm{B}\mathrm{V}_\mathbf{X}\mathrm{B}^\mathrm{T}$$

Knowledge Check

Did you understand this unit?

You can check your understanding by completing the questions for this unit on the learning platform.

Good luck!

Unit 3

Bayesian Inference and Non-Parametric Techniques

STUDY GOALS

On completion of this unit, you will have learned...

- ... about prior distributions, posterior distributions, and conjugate priors and use these to find the Bayes estimate of a parameter of interest.
- ... how to distinguish between object, weakly informative, and informative priors.
- ... Jeffrey's rule for determining weakly informative, transformation invariant, and priors distributions.
- ... how to use Parzen windows to approximate the density of a distribution using observed data.
- ... the k-nearest neighbors in both classification and density estimation contexts.

3. Bayesian Inference and Non-Parametric Techniques

Introduction

Suppose we have a (possibly biased) coin that lands on heads with probability π, which is unknown. The frequentist's approach to estimating π is to collect data, toss the coin a fixed number of times, and compute the proportion of times the coin lands on heads. This is the frequentist (maximum likelihood) estimate of π. In contrast, Bayesian statistics treats π as a random variable and, based on prior experience or knowledge, assigns a probability distribution to the values of π. Next, observations are recorded (coin tosses). The prior distribution, together with the observations, inform the Bayesian estimate of π.

The two overarching objectives of this unit are (i) to give a basic introduction to Bayesian statistics and (ii) to discuss some common non-parametric techniques. In frequentist statistics, we assume that a parameter of interest (e.g., the mean of a Gaussian distribution) is unknown by deterministic (not random). In Bayesian statistics, we assume that a parameter parameter of interest is a random variable. In summary, suppose that we want to estimate the mean, μ, from a Gaussian distribution, $\mathcal{N}(\mu, \sigma)$, where σ is known (say $\sigma = 1$). In frequentist statistics, we assume that μ is unknown but deterministic. In the Bayesian approach, we treat μ as a random variable. Next, based on previous knowledge or domain expertise, we choose a distribution of μ that doesn't depend on the current observed data. This distribution is called the prior distribution. We will discuss this in more detail in sections 3.1 and 3.2.

The choice of the prior is prone to controversy. We can either be objective, in which case we just agree with the MLE estimate, or be subjective, in which case we incorporate some information about the parameter of interest based on experience. For example, if we want to be completely objective about π, then we can say that it can take on any value between 0 and 1 with equal probability. If we want to be subjective, and almost all coins we have seen in our experience have been unbiased, we can use a distribution of π that has a sharp peak at 0.5. In section 3.2, we will learn about different types of priors and demonstrate one important (semi)-objective prior known as Jeffrey's prior.

Whenever faced with determining a parameter of interest from a predetermined family of distributions (Gaussian, Bernoulli, etc.), we use methods such as method of moments, maximum likelihood, and Bayesian methods, to estimate this unknown parameter. We may have some observed data, but we cannot assume that it came from any specific family of distributions (e.g. it can't be assumed Gaussian). Perhaps the histogram reveals a distribution whose shape is not in a standard family of distributions. We cannot assume a family, and using parametric methods to analyze this data doesn't make sense. In such situations, we try to approximate the density of the distribution directly from the data. When we aren't making any assumptions about the density and instead directly trying to estimate a density, we are in the realm of non-parametric techniques. To this end, in section 3.3, we discuss Parzen Windows. This is a method for estimating the density (PDF) of the distribution from which the observed data may have been drawn from without making any assumptions about the distribution family.

Another common density estimation technique is called k-nearest neighbors (*k*-NN). The default application of this technique is to classify unobserved data points to one class or another by asking its k nearest data points from the observed data to vote based on their own class. In section 3.4, we illustrate the *k*-NN approach to classification and show how this method can be adjusted to estimate the density of an observed sample.

3.1 Bayesian Parameter Estimation

Bayes law is at the heart of Bayesian statistics. Let's take a moment and restate this formula as it lays the foundation of Bayesian parameter estimation. Let A and B recall that the conditional probability $\mathbb{P}(\mathrm{A}\,|\,\mathrm{B})$ read "the probability of event A given that B has occured," is defined by

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

In words, the conditional probability is the joint probability divided by the probability of the conditional event. We can rewrite this formula as

$$\mathbb{P}(\mathrm{A}\cap\mathrm{B}) = \mathbb{P}(\mathrm{B}\,|\,\mathrm{A})\cdot\mathbb{P}(\mathrm{A})$$

In words, the joint probability is the product of the conditional probability times the marginal probability (of the conditioned event). We are now ready to state Bayes' formula:

$$\mathbb{P}(\mathrm{A}\,|\,\mathrm{B}) = \frac{\mathbb{P}(\mathrm{B}\,|\,\mathrm{A})\cdot\mathrm{P}(\mathrm{A})}{\mathbb{P}(\mathrm{B})}$$

Next, if we use A^{c} to denote the event that is complementary to A, we have $\mathrm{B} = (\mathrm{B}\cap\mathrm{A})\cup(\mathrm{B}\cap\mathrm{A}^{\mathrm{c}})$, furthermore, $(\mathrm{B}\cap\mathrm{A})\cap(\mathrm{B}\cap\mathrm{A}^{\mathrm{c}}) = \emptyset$. Therefore,

$$\mathbb{P}(\mathrm{B}) = \mathbb{P}(\mathrm{B}\cap\mathrm{A}) + \mathbb{P}(\mathrm{B}\cap\mathrm{A}^{\mathrm{c}})$$

Finally, using the relationship between joint probabilities and conditional probabilities, we have the total law of probability.

$$\mathbb{P}(\mathrm{B}) = \mathbb{P}(\mathrm{B}\,|\,\mathrm{A})\mathbb{P}(\mathrm{A}) + \mathbb{P}(\mathrm{B}\,|\,\mathrm{A}^{\mathrm{c}})\mathbb{P}(\mathrm{A}^{\mathrm{c}})$$

Now Bayes' formula can be written in its usual form:

$$\mathbb{P}(\mathrm{A}\,|\,\mathrm{B}) = \frac{\mathbb{P}(\mathrm{B}\,|\,\mathrm{A})\cdot\mathrm{P}(\mathrm{A})}{\mathbb{P}(\mathrm{B}\,|\,\mathrm{A})\mathbb{P}(\mathrm{A}) + \mathbb{P}(\mathrm{B}\,|\,\mathrm{A}^{\mathrm{c}})\mathbb{P}(\mathrm{A}^{\mathrm{c}})}$$

In fact, if we have a sequence of events $A_1, A_2, \ldots A_k$ which are pairwise disjoint ($A_i \cap A_j = \emptyset$ for $i \neq j$) and whose union contains all the possible outcomes, then we can write Bayes' formula in its most general form (Downey, 2016):

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i) \cdot P(A_i)}{\mathbb{P}(B|A_1)\mathbb{P}(A_1) + \mathbb{P}(B|A_2)\mathbb{P}(A_2) + \cdots + \mathbb{P}(B|A_k)\mathbb{P}(A_k)}$$

We are now ready to talk about Bayesian parametric estimation. Suppose that there are two urns, each of which contain balls that are identical except for their color. The possible colors are blue, orange, and green. These urns are in a machine that returns a ball from one of the urns when activated. In other words, it chooses one of the urns at random, and then chooses a ball from that chosen urn at random. The first urn contains three blue, five orange, and two green balls. The second urn contains three blue, five orange, and ten green balls. We don't know how the machine chooses the urns. Perhaps it chooses the first urn with 50 percent probability and the second urn with 50 percent probability. Or perhaps it chooses the first urn with 30 percent probability and the second urn with 70 percent probability. In fact, if we can figure out the probability with which it chooses the first urn, we can easily deduce the probability it chooses the second urn. Let θ denote the probability that the machine chooses the first urn. Then, $1 - \theta$ is the probability it chooses the second urn. Before we observe anything this machine produces, we can say that the θ is a random variable that is equally likely to be any number from the set $\{0.2, 0.5, 0.7\}$. In other words, $\mathbb{P}(\theta = 0.2) = \mathbb{P}(\theta = 0.5) = \mathbb{P}(\theta = 0.7) = 1/3$. In Bayesian statistics, this distribution is called a **prior distribution**. We run the machine once and observe a green ball. Using this observation (data), we want to update our distribution about θ. We do this using Bayes' formula. The following expression computes the probability that $\theta = 0.2$:

Prior distribution
This distribution is the target parameter of interest set before observing any data. It encodes our belief about the parameter of interest.

$$\mathbb{P}(\theta = 0.2|G) = \frac{\mathbb{P}(G|\theta = 0.2)\mathbb{P}(\theta = 0.2)}{\mathbb{P}(G|\theta = 0.2)\mathbb{P}(\theta = 0.2) + \mathbb{P}(G|\theta = 0.5)\mathbb{P}(\theta = 0.5) + \mathbb{P}(G|\theta = 0.7)\mathbb{P}(\theta = 0.7)}$$

Next, to compute the probability $\mathbb{P}(G|\theta = 0.2)$ considers two cases: (i) that the ball came from the first urn and (ii) that the ball came from the second urn.

$$\mathbb{P}(G|\theta = 0.2) = \frac{2}{10} \cdot 0.2 + \frac{10}{18} \cdot 0.8 \approx 0.4844$$

Similarly,

$$\mathbb{P}(G|\theta = 0.5) = \frac{2}{10} \cdot 0.5 + \frac{10}{18} \cdot 0.5 \approx 0.3778$$
$$\mathbb{P}(G|\theta = 0.7) = \frac{2}{10} \cdot 0.7 + \frac{10}{18} \cdot 0.3 \approx 0.3067$$

In each of these computations, the probability quantity occurring in the denominator of Bayes' formula is called the **evidence**. Plugging these values into Bayes formula, we have

Evidence
The (unconditional) probability of observing the data is the evidence.

$$\mathbb{P}(\theta = 0.2 \,|\, \mathrm{G}) = \frac{0.4844 \cdot 1/3}{0.4844 \cdot 1/3 + 0.3778 \cdot 1/3 + 0.3067 \cdot 1\,.\,3} \approx 0.4144$$

Similarly, we can compute the following:

$$\begin{array}{l} \mathbb{P}(\theta = 0.5 \,|\, \mathrm{G}) \approx 0.3232 \\ \mathbb{P}(\theta = 0.7 \,|\, \mathrm{G}) \approx 0.2624 \end{array}$$

Now we have a new distribution for θ, this new distribution, which takes into account the prior distribution and the new observation, is called the posterior distribution. If we stop here, the Bayes estimate of θ can be computed either by considering the mode, the value of θ with the highest posterior probability, $\theta^{(\mathrm{Bayes,mode})} = 0\,.\,2$ or the mean,

$$\theta^{(\mathrm{Bayes,\,mean})} = 0.2 \cdot 0.4144 + 0.5 \cdot 0.3232 + 0.7 \cdot 0.2624 = 0.4282$$

Either way, we see that if we only consider this one observation, it is more likely that the machine chooses the second urn with higher probability. Let's continue with this idea but make a couple of adjustments. As a prior distribution, we assume that θ is equally likely to be any value between 0 and 1, that is $\theta \sim \mathrm{U}(0, 1)$. Furthermore, assume that each of the urns contains an infinite number of balls of either blue or green. The first urn contains blue and green with proportions 0.4 and 0.6 respectively. The second urn contains blue and green with proportions 0.7 and 0.3. Next, let's run the machine ten times to get ten balls. We observed four blue and six green balls. We would like to estimate the probability θ that this machine chooses the first urn. Let's collect our observations into an event $\mathrm{D} = \{\mathrm{B}, \mathrm{B}, \mathrm{B}, \mathrm{B}, \mathrm{G}, \mathrm{G}, \mathrm{G}, \mathrm{G}, \mathrm{G}, \mathrm{G}\}$, where B indicates a blue ball and G indicates a green ball.

Let's calculate $\mathbb{P}(\mathrm{D} \,|\, \theta = \mathrm{t})$ using Bayes formula. For each ball, it either came from the first urn or the second urn independently. Therefore,

$$\mathbb{P}(\mathrm{D} \,|\, \theta = \mathrm{t}) = (0\,.\,4\mathrm{t} + 0\,.\,7(1 - \mathrm{t}))^4 (0\,.\,6\mathrm{t} + 0\,.\,3(1 - \mathrm{t}))^6$$

This quantity is called the likelihood. The numerator of the Bayes' formula is the likelihood times the prior distribution. Since the prior is uniform on (0, 1), its probability density function is constant of 1. Since θ is assumed to be a continuous random variable, the evidence is the integral of this quantity.

$$\begin{aligned}
\text{evidence} &= \int_0^1 \text{likelihood} \times \text{prior} \\
&= \int_0^1 \mathbb{P}(\mathrm{D}\,|\,\theta = \mathrm{t}) \cdot 1\,\mathrm{dt} \\
&= \int_0^1 (0.4\mathrm{t} + 0.7(1-\mathrm{t}))^4 (0.6\mathrm{t} + 0.3(1-\mathrm{t}))^6 \cdot 1\,\mathrm{dt} \\
&\approx 0.000737597
\end{aligned}$$

Finally, Bayes' formula computes the posterior distribution,

$$\begin{aligned}
\text{post}_\theta(\mathrm{t}) &= \mathbb{P}(\theta = \mathrm{t}\,|\,\mathrm{D}) \\
&= \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} \\
&= \frac{(0.4\mathrm{t} + 0.7(1-\mathrm{t}))^4 (0.6\mathrm{t} + 0.3(1-\mathrm{t}))^6}{0.000737597}
\end{aligned}$$

Note

Your focus here should be on the formulas and concepts, and not necessarily how to compute the integral in the evidence as you are not responsible for tedious computations involving integrals.

Let's summarize the main ideas we just discussed. For any parametric inference, we need an observed data set, $\mathcal{D}$, and the distribution from which this data is drawn $\mathrm{f}(\,\cdot\,|\,\theta)$ based on one or more unknown parameters of interest, θ. Our goal is to estimate θ. The frequentist approach treats θ as an unknown but deterministic (non-random) quantity and uses the maximum likelihood method. In Bayesian inference, we treat θ as a random variable and encode our prior belief by specifying a distribution called the prior: $\theta \sim \text{prior}(\theta)$. Next, using Bayes' formula, we derive the posterior distribution, $\text{post}(\theta)$ as (Hogg et al., 2019)

$$\text{post}(\theta) = \frac{\ell(\mathcal{D}\,|\,\theta) \cdot \text{prior}(\theta)}{\mathbb{P}(\mathcal{D})}$$

If θ is discrete, then

$$\mathbb{P}(\mathcal{D}) = \sum_\theta \ell(\mathcal{D}\,|\,\theta)\text{prior}(\theta)$$

If it is continuous, then

$$\mathbb{P}(\mathcal{D}) = \int_{-\infty}^{\infty} \ell(\mathcal{D}\,|\,\theta)\text{prior}(\theta)\mathrm{d}\theta$$

Finally, the Bayes (mean) estimate is $\mathbb{E}_{\text{post}}[\theta]$. For discrete θ, we have

$$\widehat{\theta}_{\text{Bayes}} = \mathbb{E}_{\text{post}}[\theta] = \sum_{\theta} \theta \text{post}(\theta)$$

and for continuous θ we have

$$\widehat{\theta}_{\text{Bayes}} = \mathbb{E}_{\text{post}}[\theta] = \int_{-\infty}^{\infty} \theta \text{post}(\theta) d\theta$$

Computationally, the most difficult part of this process is to compute the evidence. In practice, we don't need to compute this directly. The evidence is just a normalizing constant to make sure that the posterior is a valid distribution. As such, in our computations, we just want to find the form of the posterior distribution and therefore we can ignore the evidence term and only consider quantities that involve the parameter of interest. Equality is replaced by proportionality, so we have

$$\begin{array}{rcl} \text{post} & \propto & \ell(\mathcal{D}|\theta) \cdot \text{prior}(\theta) \\ \text{posterior} & \propto & \text{likelihood} \cdot \text{prior} \end{array}$$

Once we have the form of the posterior, we can choose the constant that makes the posterior a valid PMF (PDF). We will need to use the Beta distribution in the next example. Recall that if $X \sim \text{Beta}(\alpha, \beta)$ then its PDF is given by

$$f_X(x) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, & 0 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

The mean of the Beta distribution is given by $\frac{\alpha}{\alpha+\beta}$.

Example 3.1.1

Our goal in this example is to find an estimate for the probability of heads, θ, of a (possibly biased) coin. We toss a coin five times and observe $\{H, H, T, H, T\}$. Assume a prior distribution on θ to be $\theta \sim \text{Beta}(2, 2)$. Find the posterior distribution and the Bayes (mean) estimate, $\widehat{\theta}_{\text{Bayes}}$.

Solution

We start by writing down the likelihood. We can encode the observed data as $X = 1$ for heads and $X = 0$ for tails. As such, $X \sim \text{Bernoulli}(\theta)$ and its PMF is $f_X(x) = \theta^x(1-\theta)^{1-x}$, for $x = 0, 1$ and zero otherwise. Therefore, our data is $\mathcal{D} = \{x_1, ..., x_5\} = \{1, 1, 0, 1, 0\}$ and the likelihood is given by

$$\begin{aligned}\ell(\mathcal{D}|\theta) &= \prod_{i=1}^{5} f(x_i|\theta) \\ &= \prod_{i=1}^{5} \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \theta^{\sum_i x_i}(1-\theta)^{1-\sum_i x_i} \\ &= \theta^3(1-\theta)^2\end{aligned}$$

Next, we know that $\theta \sim \text{Beta}(2,2)$, therefore the prior distribution is given by

$$\text{prior}(\theta) \propto \begin{cases}\theta(1-\theta), & 0 \leq \theta \leq 1 \\ 0, & \text{otherwise}\end{cases}$$

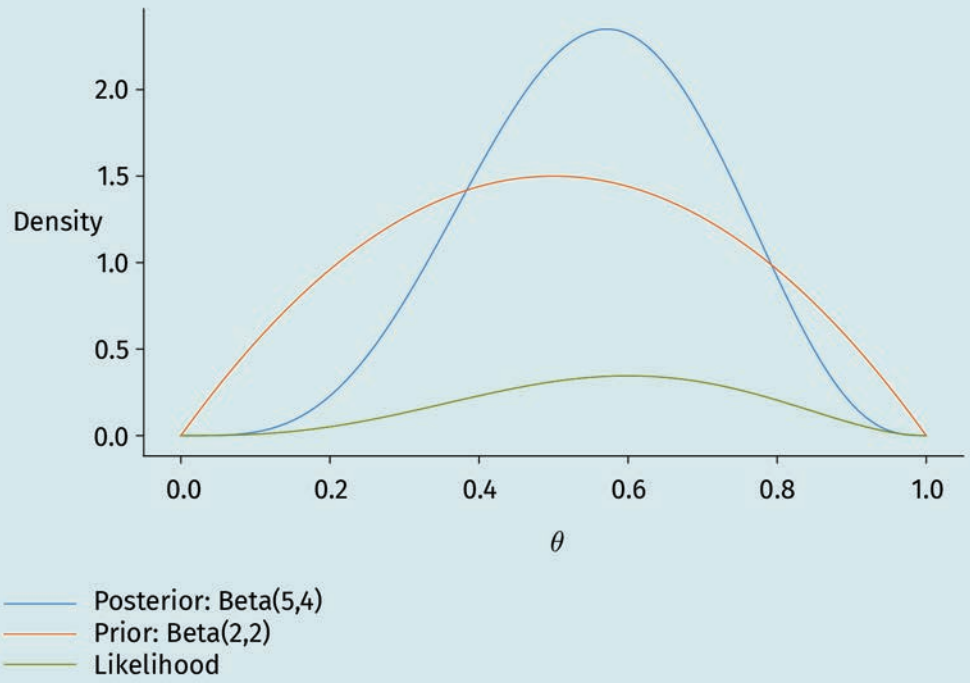
The posterior is given by

$$\text{post}(\theta) \propto \theta^3(1-\theta)^2 \cdot \theta(1-\theta) = \theta^4(1-\theta)^3$$

Notice that the posterior takes the form of a Beta distribution. We can infer the parameters via pattern machine $\theta^{5-1}(1-\theta)^{4-1}$ to see that the posterior distribution of θ is $\text{Beta}(5,\ 4)$. Explore the figure below for the likelihood, the prior, and the posterior distributions. Now, using this posterior, we can write the Bayes' mean estimate as

$$\widehat{\theta}_{\text{Bayes}} = \frac{\alpha}{\alpha+\beta} = \frac{5}{9} \approx 0.5556$$

Graphs of the Likelihood, Prior, and Posterior Distributions from Example 3.1.1



For the example above, the frequentist approach would have been to estimate θ via the MLE, which is $\widehat{\theta}_{\text{MLE}} = \frac{3}{5} = 0.6$. The prior estimate is the mean of the prior distribution which is $\widehat{\theta}_{\text{prior}} = \frac{2}{2+2} = \frac{1}{2} = 0.5$. Notice that the Bayes estimate is between these two estimates. In fact, it is their weighted average:

$$\widehat{\theta}_{\text{Bayes}} = \frac{5}{9} = \frac{5}{9} \cdot \frac{3}{5} + \frac{4}{9} \cdot \frac{1}{2} = \frac{5}{9}\widehat{\theta}_{\text{MLE}} + \frac{4}{9}\widehat{\theta}_{\text{prior}}$$

This is not a coincidence. Working out example 3.1.1 in a more general setting, say our observed dataset is

$$\mathcal{D} = \left\{ \underbrace{1, ..., 1}_{\text{h times}}, \underbrace{0, ..., 0}_{(\text{n}-\text{h}) \text{ times}} \right\}$$

and the prior on θ is $\text{Beta}(\alpha, \beta)$. The likelihood is

$$\ell(\mathcal{D}\,|\,\theta) = \theta^{\text{h}}(1-\theta)^{\text{n}-\text{H}}$$

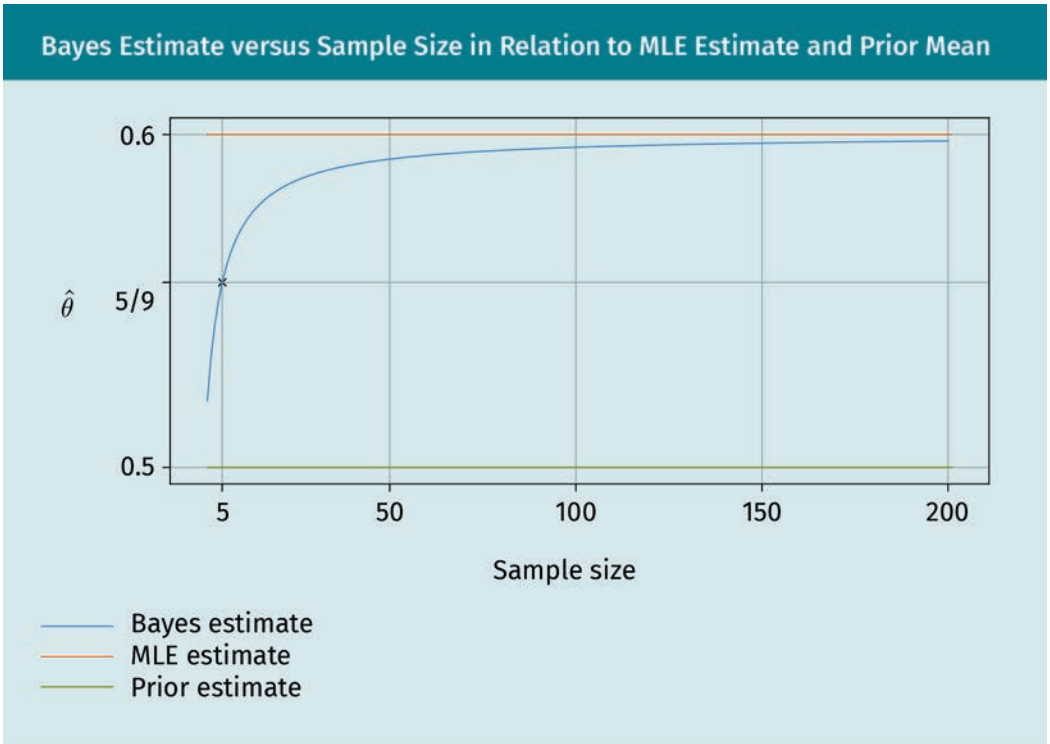
and the prior is

$$\text{prior}(\theta) \propto \begin{cases} \theta^{\alpha-1}(1-\theta)^{\beta-1}, & 0 \le \theta \le 1 \\ 0, & \text{otherwise} \end{cases}$$

Therefore, as expected, the posterior follows $\text{Beta}(h+\alpha, n-h+\beta)$. The Bayes (mean) estimate is $\widehat{\theta}_{\text{Bayes}} = \frac{h+\alpha}{n+\alpha+\beta}$. The MLE estimate is $\widehat{\theta}_{\text{MLE}} = \frac{h}{n}$, and the prior estimate is $\widehat{\theta}_{\text{prior}} = \frac{\alpha}{\alpha+\beta}$. Finally, we can write the Bayes estimate as a weighted sum of the MLE and prior estimates:

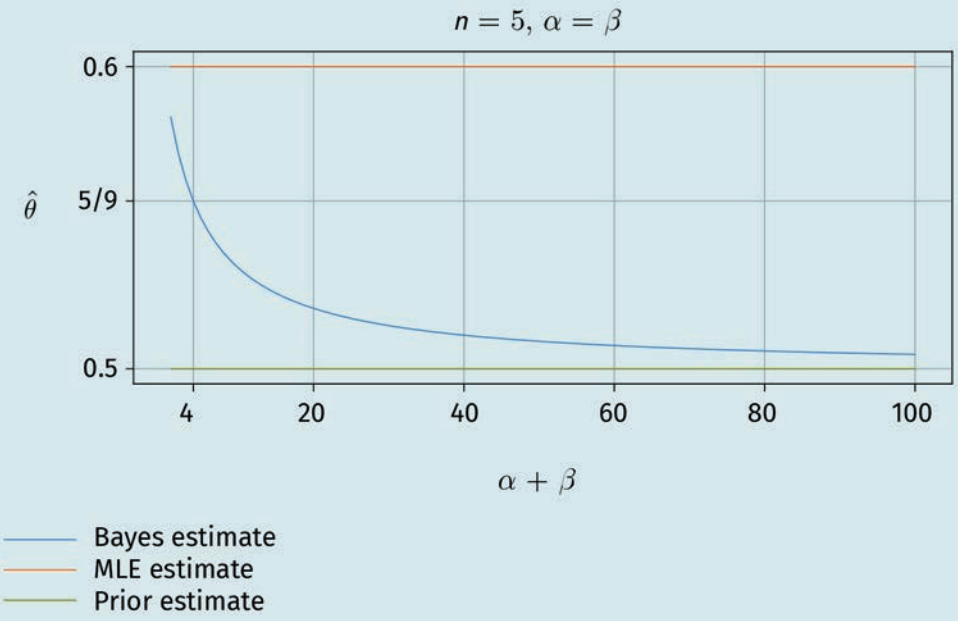
$$\widehat{\theta}_{\text{Bayes}} = \underbrace{\frac{n}{n+\alpha+\beta}}_{\lambda} \cdot \frac{h}{n} + \underbrace{\frac{\alpha+\beta}{n+\alpha+\beta}}_{1-\lambda} \cdot \frac{\alpha}{\alpha+\beta} = \lambda\widehat{\theta}_{\text{MLE}} + (1-\lambda)\widehat{\theta}_{\text{prior}}$$

For $\alpha,\ \beta,\ n > 0$, we have $0 < \lambda, 1-\lambda < 1$. Therefore, indeed the Bayes estimate is between the MLE and prior estimates. This illuminates and important characteristic. If n is large (we have a lot of data), then λ is close to 1 and $1 - \lambda$ is close to zero, so the Bayes estimate is close to the MLE. On the other hand, if we don't have much data (n is small relative to $\alpha + \beta$), then λ is small, and $1 - \lambda$ is close to 1, and the Bayes estimate is close to the prior. The figure below shows the MLE and prior estimates together with the Bayes (mean) estimate for various sample sizes. The computed value from example 3.1.1 is shown as well.



One the other hand, for a fixed sample size, the parameters of the prior also affect the Bayes estimate. The figure below shows the Bayes estimate versus $\alpha + \beta$ with $\alpha = \beta$ for a fixed sample of size $n = 5$ and sum of observed values $\sum x_i = 3$, as in example 3.1.

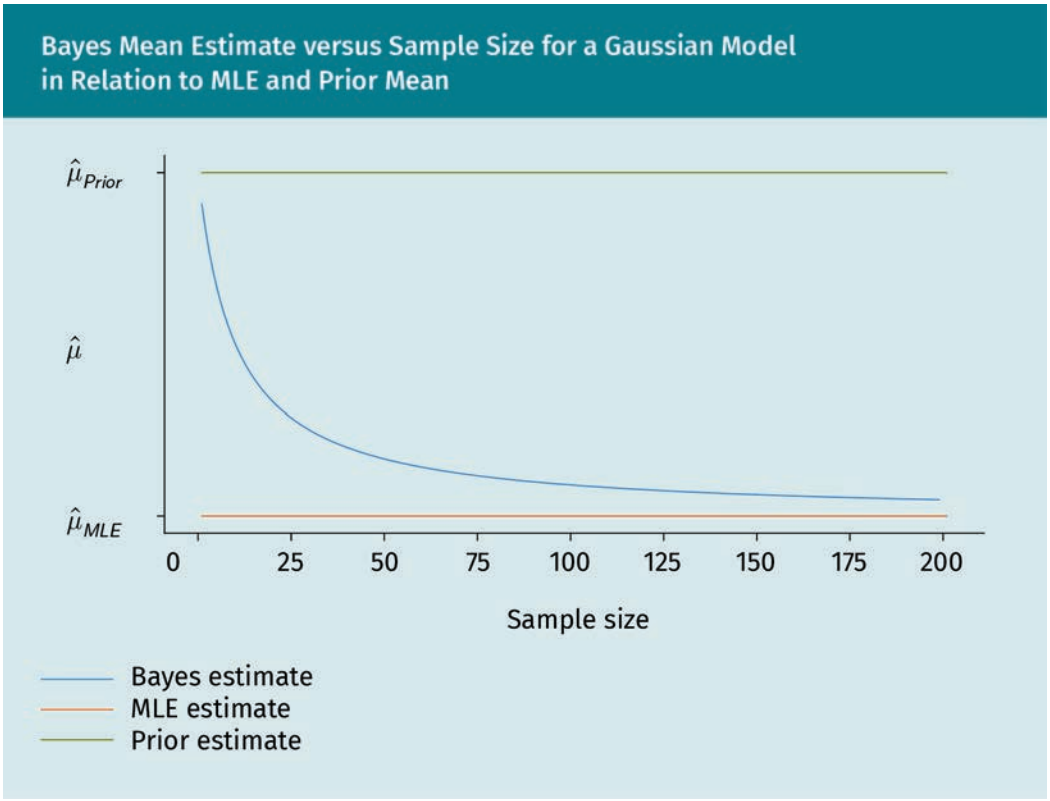
Bayes Estimate for a Fixed Sample Size versus Parameters of the Prior in Relation to MLE Estimate and Prior Mean



This decomposition of the Bayes estimate is not unique to the Bernoulli-Beta combination. Suppose $\mathcal{D} = \{x_1, ..., x_n\}$ is observed from a Gaussian distribution $\mathcal{N}(\mu, \sigma)$ with unknown μ and known σ. Based on our beliefs, we set the prior for $\mu \sim \mathcal{N}(\mu_0, 1)$. The posterior distribution of μ is also Gaussian with $\left[\mu \middle| \mathcal{D}\right] \sim \mathcal{N}\left(\frac{n\bar{x} + \mu_0\sigma^2}{n + \sigma^2}, \sigma^2(1 + \sigma^2/n)\right)$. Therefore, the Bayes estimate is $\widehat{\mu}_{\text{Bayes}} = \frac{n\bar{x} + \mu_0\sigma^2}{n + \sigma^2}$. The MLE estimate for μ is the sample mean, $\widehat{\mu}_{\text{MLE}} = \bar{x}$ and the prior mean is $\widehat{\mu}_{\text{prior}} = \mu_0$. As before, we can write the Bayes estimate as a weighted sum of the MLE and the prior estimates:

$$\widehat{\mu}_{\text{Bayes}} = \frac{n\bar{x} + \mu_0\sigma^2}{n + \sigma^2} = \underbrace{\frac{n}{n + \sigma^2}}_{\lambda} \cdot \bar{x} + \underbrace{\frac{\sigma^2}{n + \sigma^2}}_{1 - \lambda} \cdot \mu_0 = \lambda\widehat{\mu}_{\text{MLE}} + (1 - \lambda)\widehat{\mu}_{\text{prior}}$$

The figure below shows the relationship between the Bayes estimate of μ versus the sample size in relation to the MLE estimate and the prior mean for a Gaussian model.



Maximum a posteriori estimate
This is the value of the target parameter at which the posterior distribution achieves its global maximum; essentially the mode of the posterior distribution.

So far, our Bayes estimate has been the posterior mean. In practice, some researchers also use another estimate: **maximum a posteriori estimate** (MAP). The MAP estimate is the maximizer of the posterior distribution:

$$\widehat{\theta}_{\text{MAP}} = \arg \max_{\theta} \text{post}(\theta)$$

The maximizer of $\text{Beta}(\theta|\alpha,\beta)$ is $\frac{\alpha-1}{\alpha+\beta-2}$ for $\alpha,\ \beta > 1$. The posterior distribution in example 3.1.1 was $\text{Beta}(5,\ 4)$. Therefore, the MAP estimate for that example is $\widehat{\theta}_{\text{MAP}} = \frac{4}{7} \approx 0.5714$. If the posterior is Gaussian, the Bayes mean estimate is the same as the MAP estimate. This is because a Gaussian distribution attains its maximum at its mean.

Bayesian parameter estimation treats an unknown parameter of interest as a random variable. We first set a prior distribution for our parameter of interest. Next, we use Bayes' formula to find the posterior distribution using this prior and observed data. Finally, we use the posterior distribution to compute a Bayes estimate for the parameter of interest. The two common Bayes estimates are the posterior mean, which is the expected value of the posterior distribution and the maximizer (mode) of the posterior distribution.

3.2 Prior Probability Functions

The prior distribution of a parameter of interest encodes the beliefs of said parameter. In section 3.1, we used a Beta prior for the parameter π of the Bernoulli distribution. The resulting posterior distribution of π also turned out to be Beta (although with different parameters). Therefore, the prior of Beta is called a **conjugate prior**. Conjugate priors are desirable because they allow us to investigate the posterior analytically, with formulas. Before the popularization of computing, having a conjugate prior was the only way to investigate posteriors. Today, we no longer have this restriction and are free to choose any prior distribution that we think appropriate to encode our beliefs of the parameter of interest. The table below shows some of conjugate priors for certain likelihoods.

Conjugate prior
In general, when a prior distribution results in a posterior distribution of the same functional form with different parameters, it is called a conjugate prior.

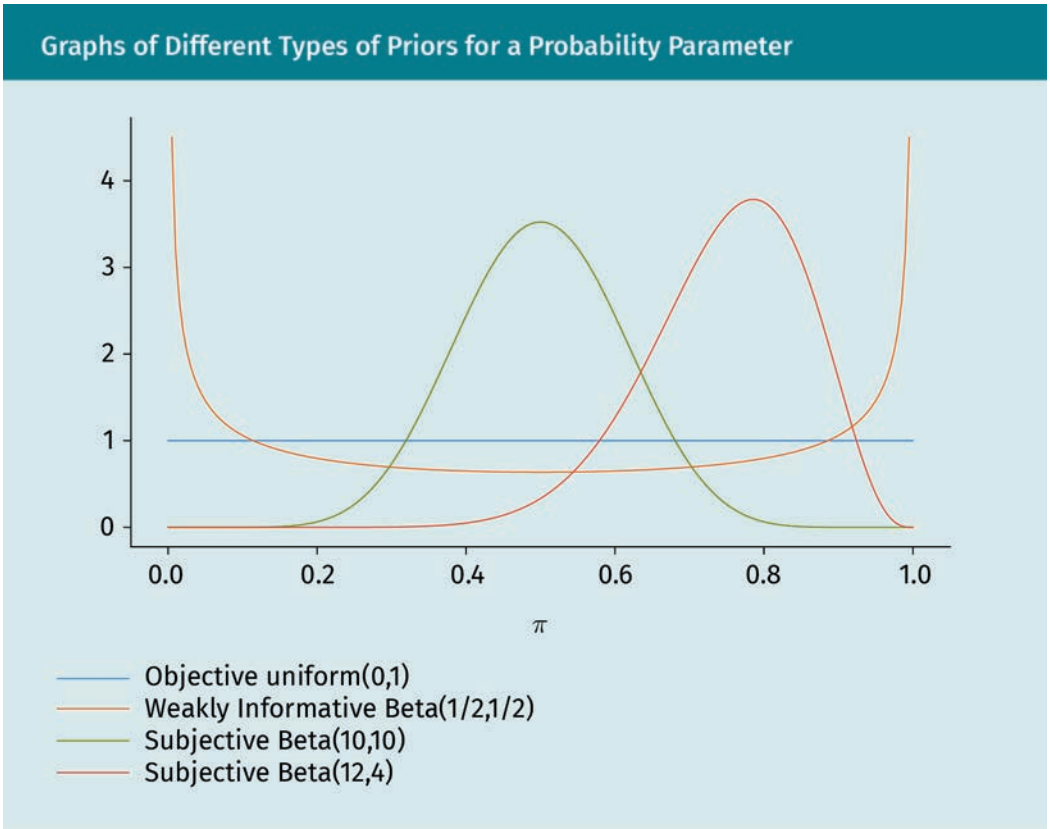
Conjugate Priors for Some Discrete and Continuous Likelihoods		
Likelihood	Target parameter	Conjugate prior
$\text{Bernoulli} \quad (\pi)$	π (probability)	Beta
$\text{Binomial}(n, \pi)$	π (probability)	Beta
$\text{Poisson}(\lambda)$	λ (rate)	Gamma
$\text{Geometric}(\pi)$	π (probability)	Beta
$\mathcal{N}(\mu, \sigma)$	μ (mean)	Gaussian
$\text{Exponential}(\lambda)$	λ (rate)	Gamma
$\text{Gamma}(\alpha, \beta)$	β (rate)	Gamma

In general, we can classify priors into three categories:

1. Objective priors
2. Weakly informative priors
3. Subjective (highly informative) priors

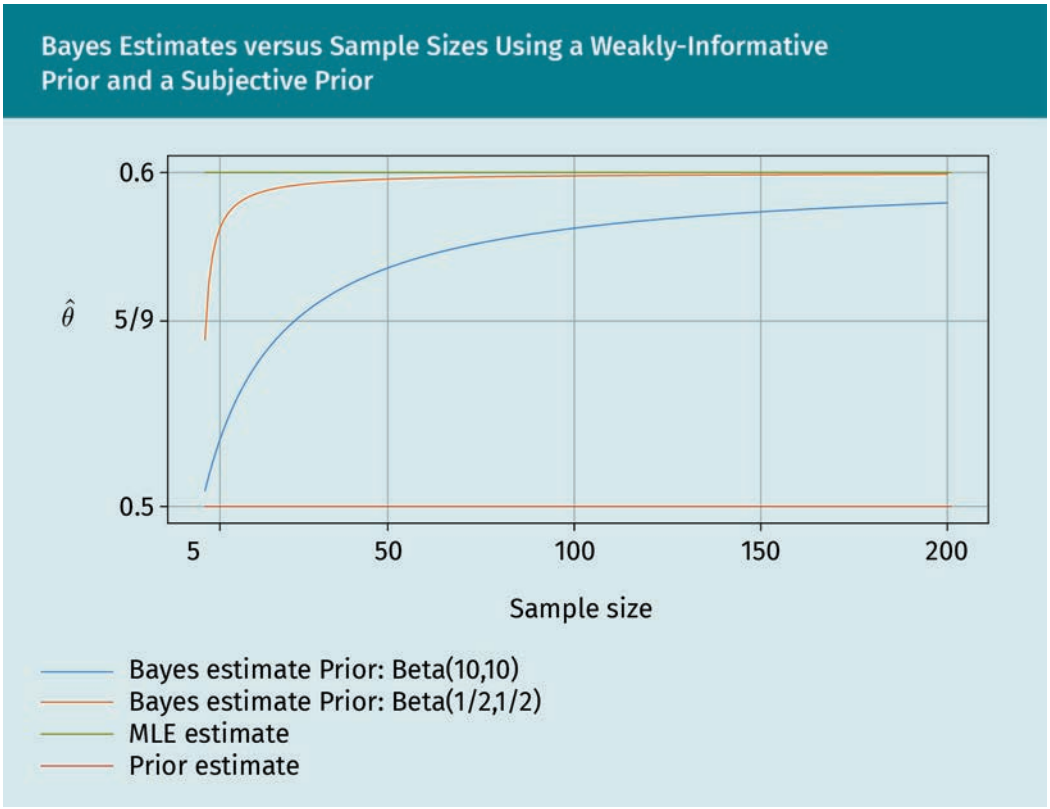
Objective priors only enforce domain restrictions on the target parameter. Consider a likelihood following $\text{Bernoulli}(\pi)$. An objective prior is given by $\text{prior}(\pi) \sim \mathcal{U}(0,1)$, the uniform distribution on the interval $[0,1]$. This prior forces the parameter to lie between 0 and 1; which is required for a probability parameter such as π. A weakly informative prior is given by $\text{prior}(\pi) = \text{Beta}(\pi \mid 1/2, 1/2) \propto \frac{1}{\sqrt{\pi(1-\pi)}}$. This prior forces π to live between 0 and 1 exclusively and slightly favors extreme values (near 0 or 1) and is otherwise almost uniform. An example of a subjective prior might be $\text{prior}(\pi) = \text{Beta}(\pi \mid 10, 10) \propto \pi^9(1-\pi)^9$. This distribution highly favors values near $1/2$. The figure below shows graphs of these three prior distributions.

Graphs of Different Types of Priors for a Probability Parameter



The choice of priors greatly impacts the estimates obtained from Bayesian methods. As such, choosing an appropriate prior is very important. Choosing a bad prior might render the estimates obtained useless. Generally, there are two schools of thought regarding the choice of priors: subjective Bayesians and objective Bayesians. Subjective Bayesians subscribe to the opinion that the prior should encode all available information (e.g., prior experience). Objective Bayesians argue that scientific thought must be as objective as possible, and injecting information into the model other than the given data violates this objectivity. As a concrete example, the prior $\text{Beta}(10,10)$ strongly favors values near 0.5 and is used to express the belief that the probability parameter is near that value. In the real world, this might mean that the researcher has almost always encountered unbiased coins and wants to use this knowledge in their model.

The figure below shows a graph of the Bayes (mean) estimate versus sample sizes for the weakly-informative prior ($\mathrm{Beta}(1/2, 1/2)$), a subjective prior ($\mathrm{Beta}(10, 10)$), and their relative relation to one another and to the MLE and and Prior mean.



The question remains, however, how does one choose an objective prior? Intuitively, it seems that the most objective choice would be to choose a constant prior. For example, in the case of a probability parameter, the uniform distribution on $[0, 1]$; basically, $\mathrm{prior} \propto \mathrm{c}$, a constant. Now consider that we want to estimate the mean μ of a Gaussian $\mathcal{N}(\mu, \sigma)$ with known σ. We want to choose a prior on μ that is objective. Following the same reasoning, we choose $\mathrm{post}(()\mu) \propto \mathrm{c}$. However, $\int_0^\infty = \infty$, and so this is not a valid density. Such priors are called improper priors. It may be the case that even if the prior is improper, the posterior is still a valid density.

Given observed data $\mathcal{D} = \{\mathrm{x}_1, ..., \mathrm{x}_\mathrm{n}\}$ from $\mathcal{N}(\mu, \sigma)$ with known σ and unknown μ. The likelihood,

$$\ell(\mathcal{D}\,|\,\theta) \propto \exp\left[-\frac{1}{2\left(\frac{\sigma}{\sqrt{\mathrm{n}}}\right)^2}(\overline{\mathrm{x}} - \mu)^2\right]$$

Choosing a uniform prior on μ, $\mathrm{prior}(\mu) = 1$, the posterior is

$$\text{post}(\mu) \propto \exp\left[-\frac{1}{2\left(\frac{\sigma}{\sqrt{n}}\right)^2}(\bar{x}-\mu)^2\right]$$

Therefore, $\text{post}\left(\theta\right) = \mathcal{N}\left(\mu \,\middle|\, \bar{x}, \frac{\sigma}{\sqrt{n}}\right)$. In other words, the posterior is Gaussian with mean $\bar{x}$ and variance σ^2/n; a valid density! The Bayes (mean) estimate is $\widehat{\mu}_{\text{Bayes}} = \bar{x}$, which is the same as the MLE estimate. Using a uniform (objective) prior resulted in the Bayes estimate agreeing with the frequentist estimate. This shouldn't surprise us. The only source of information with a uniform prior is the data.

Everything worked out fine with the uniform prior for the mean in the Gaussian model. But things might go wrong. Let's return to the $\textbf{Bernoulli}(\pi)$ with unknown π. If we use a uniform prior on π, we get $\text{post}(\pi) = \text{Beta}(\pi \,|\, \sum x_i + 1, n + 1)$ and the Bayes (mean) estimate is $\widehat{\pi}_{\text{Bayes}} = \frac{\sum x_i + 1}{n+1}$, not too different from the MLE estimate, $\widehat{\pi}_{\text{MLE}} = \frac{\sum x_i}{n}$, of the frequentist approach.

A closely related quantity associated with a probability π are the odds: $\frac{\pi}{1-\pi}$. This compares the probability of success relative to the probability of failure. The log-odds is the natural logarithm of this quantify: $\log\left(\frac{\pi}{1-\pi}\right)$. Note that the transformation $\phi = \log\frac{\pi}{1-\pi}$ is one-to-one. If we want to be objective about π, then we should also be objective about the log-odds parameter ϕ. However, using a flat prior, $\text{prior}(\pi) = 1$, makes results in a density for ϕ that is not flat:

$$\text{prior}(\phi) = \frac{e^{\phi}}{\left(1+e^{\phi}\right)^2}$$

This is a contradiction! If we assume that we have no information about the parameter of interest and choose a flat prior, then any transformation of that parameter must also have a flat prior. Otherwise, the transformation erroneously encodes some information. Said another way, if we choose a prior for a parameter that has some degree of objectivity, then we would like any transformed parameter to have the same degree of objectivity, i.e., the same distribution. This property is called transformation invariance. Harold Jeffrey came up with a rule that produces weakly informative (almost objective) priors that are transformation invariant. This rule is based on the Fisher information, a way to quantify the information a random variable provides about a parameter of interest. Formally, the Fisher information of a parameter, θ, is denoted by $I(\theta)$ and is defined by

$$I(\theta) = -\mathbb{E}[\ell\ell''(X \,|\, \theta)|\theta]$$

where the derivatives are with respect to the parameter θ and the expectation is with respect to the distribution of the data. Jeffrey's prior is defined by $\text{prior}_J\left(\theta\right) = \sqrt{|I(\theta)|}$.

Example 3.2.1
Let $X \sim \text{Bernoulli}(\pi)$ with unknown π. Use the log-likelihood function

$$\ell\ell(\mathrm{X}|\pi) = \mathrm{X} \quad \log \quad \pi + (1-\mathrm{X})\log(1-\pi)$$

to find Jeffrey's prior.

Solution

We start by computing the first and second derivatives with respect to π. First, recall that the derivative of the (natural) logarithm is the reciprocal function. In other words, $\frac{\mathrm{d}}{\mathrm{du}}\log\mathrm{u} = \frac{1}{\mathrm{u}}$. Using these facts, the derivative of $\mathrm{X}\log\pi$ is $\frac{\mathrm{d}}{\mathrm{d}\pi}(\mathrm{X}\log\pi) = \frac{\mathrm{X}}{\pi}$. Also, if the argument of the logarithm is itself a function of the independent variable, then its derivative is $\frac{\mathrm{d}}{\mathrm{du}}(\log(\mathrm{g}(\mathrm{u})) = \frac{\mathrm{g}'(\mathrm{u})}{\mathrm{g}(\mathrm{u})}$. Using this fact, we have $\frac{\mathrm{d}}{\mathrm{d}\pi}((1-\mathrm{X})\log(1-\pi)) = \frac{1-\mathrm{X}}{1-\pi}\cdot\Big(-1\Big)$. Using these results, we have the derivative of the log-likelihood

$$\ell\ell'(\mathrm{X}|\pi) = \frac{\mathrm{X}}{\pi} - \frac{1-\mathrm{X}}{1-\pi}$$

Next, recall that the derivative of a power function uses the power rule. If $\mathrm{n} \neq 1$, then $\frac{\mathrm{d}}{\mathrm{du}}\mathrm{u}^{\mathrm{n}} = \mathrm{nu}^{\mathrm{n}-1}$. Therefore, we have $\frac{\mathrm{d}}{\mathrm{d}\pi}\Big(\frac{\mathrm{X}}{\pi}\Big) = -\frac{\mathrm{X}}{\pi^2}$. Next, if the factor in a power function is itself a function of the independent variable, then we need to use the chain rule: $\frac{\mathrm{d}}{\mathrm{du}}\mathrm{g}\Big(\mathrm{u}\Big)^{\mathrm{n}-1} = \mathrm{ng}\Big(\mathrm{u}\Big)^{\mathrm{n}-1}\cdot\frac{\mathrm{dg}}{\mathrm{du}}$. Therefore, $\frac{\mathrm{d}}{\mathrm{d}\pi}\Big(\frac{1-\mathrm{X}}{1-\pi}\Big) = \Big(-1\Big)\cdot\frac{1-\mathrm{X}}{(1-\pi)^2}$. With these results, the second derivative of the log-likelihood function is

$$\ell\ell''(\mathrm{X}|\pi) = -\frac{\mathrm{X}}{\pi^2} - \frac{1-\mathrm{X}}{(1-\pi)^2}$$

The Fisher information is

$$\begin{aligned}
\mathrm{I}(\pi) &= -\mathbb{E}\left[-\frac{\mathrm{X}}{\pi^2} - \frac{1-\mathrm{X}}{(1-\pi)^2}\middle|\pi\right] \\
&= -\left(-\frac{\pi}{\pi^2} - \frac{1-\pi}{(1-\pi)^2}\right) \\
&= \frac{1}{\pi} + \frac{1}{1-\pi} \\
&= \frac{1}{\pi(1-\pi)}
\end{aligned}$$

Finally, using Jeffrey's rule, Jeffrey's prior is given by

$$\mathrm{prior}_{\mathrm{J}}(\pi) = \sqrt{|\mathrm{I}(\pi)|} = \frac{1}{\sqrt{\pi(1-\pi)}}$$

In this section, we have discussed how the choice of the prior plays an important role in the Bayes estimate of a parameter of interest. In particular, we have seen that priors can be uninformative (contain no prior information about the parameter of interest), weakly informative (contain some information), or informative. Furthermore, we learned about Jeffrey's rule in determining an appropriate weakly informative prior distribution that has the desirable property of being transformation invariant. That is, the

information imputed by the Jeffrey's prior is the same as the information imputed by corresponding distribution of the transformed parameter. Finally, we also discussed conjugate priors. These are prior distributions that allow the posterior to belong to the same family. Such distributions are only chosen because the resulting mathematics is easy. Today, with the popularity of software packages, we are no longer bound by purely analytical results and can choose any prior that imputes the information about the parameter we want.

3.3 Parzen Windows

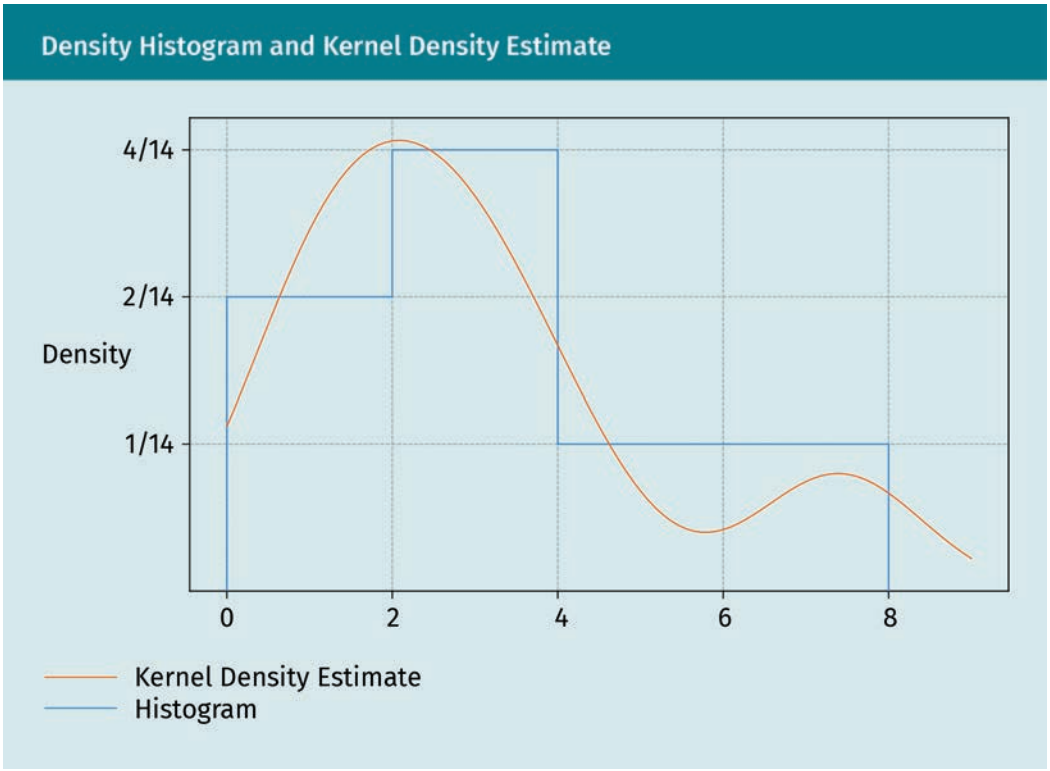
Parzen windows, also called Parzen-Rosenblatt windows, is a method for estimating the PDF of a distribution from a finite sample. It is a very flexible method since it requires no knowledge of the distribution nor any knowledge of its parameters. One of the simplest visual tools we can use to estimate the shape of a distribution from a finite sample is to look at a histogram. Since this method is a generalization for creating histograms, we will start our discussion here.

Given an observed sample $\{x_1, ..., x_n\}$, we would like to organize our data into "bins" of a fixed size. One way to do this is to choose the bin width, h, as well as the smallest value that the first bin contains. Consider the observed sample $\{1, 1, 2, 2.5, 3, 4, 7.4\}$. Suppose we choose a bin size of $h = 2$. The first bin should be chosen to include the smallest value of 1, so we set $[0, 2)$ as the first bin. The next bin will be $[2, 4)$, the next bin will be $[4, 6)$, and the final bin will be $[6, 8)$. We will stop here because the last bin contains the largest value in our sample. Next, we count the number of data points that fall in the respective bins. From these count values, we can compute the proportion of the total or the observed probability by dividing each count by the sample size; in this case the sample size is 7. Finally, to get a density, we divide the probability by the bin size. The table below shows these values for our observed sample.

Computing a Density Histogram

Bin	Coun	Probability	Density
[0,2)	2	1/7	2/14
[2,4)	3	3/7	3/14
[4,6)	1	1/7	1/14
[6,8)	1	1/7	1/14

The Parzen window method can be thought of as a way to construct a "smooth" histogram. The figure below shows the histogram we computed together with a density estimate using the Parzen window (kernel density estimate).



The Parzen window method approximates the density of the distribution using a finite sample $\{x_1, ..., x_n\}$ by the formula

$$\widehat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{x - x_i}{h}\Big)$$

where $K(\,\cdot\,)$ is a non-negative function called the kernel and $h > 0$ is a number called the bandwidth or window-size. In practice, the kernel function K is chosen to be a valid PDF. A common choice is the standard Gaussian PDF, $K = \phi$ where

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

With this kernel, the estimate of the density from the sample data is given by

$$\widehat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} \phi\Big(\frac{x - x_i}{h}\Big) = \frac{1}{nh}\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x - x_i)^2}{2h^2}}$$

Example 3.3.1
Give the kernel density estimate of the sample data $\{1, 1, 2, 2.5, 3, 4, 7.4\}$ using the Gaussian kernel with window size $h = 1$.

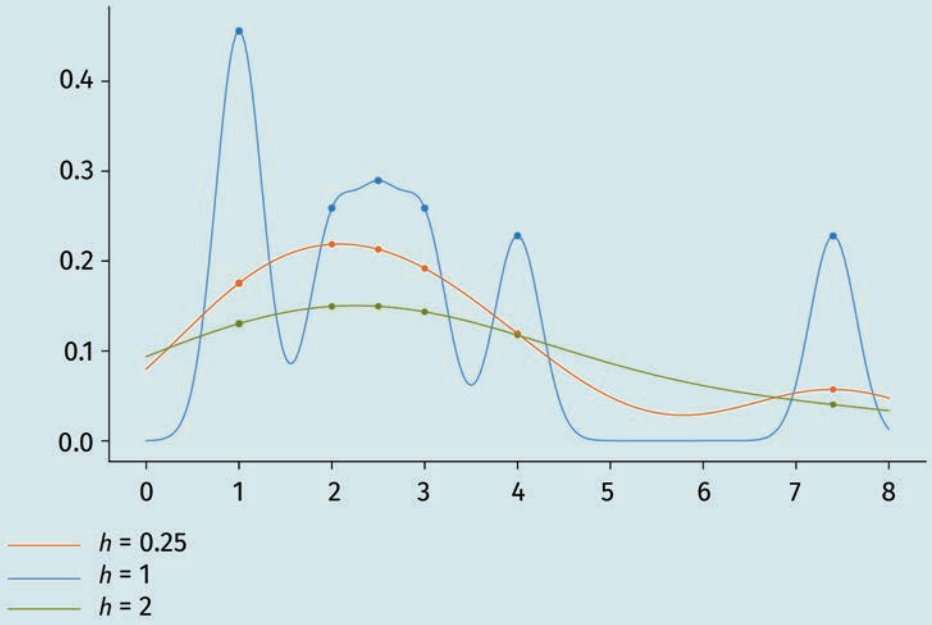
Solution
Following the equation above, we have

$$\begin{aligned}\widehat{f}(x) &= \frac{1}{7 \cdot 1} \sum_{i=1}^{7} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x - x_i)^2}{2 \cdot 1^2}} \\ &= \frac{1}{7\sqrt{2\pi}} \left[e^{-\frac{(x-1)^2}{2}} + e^{-\frac{(x-1)^2}{2}} + e^{-\frac{(x-2)^2}{2}} + e^{-\frac{(x-2.5)^2}{2}} + e^{-\frac{(x-3)^2}{2}} + \right. \\ &\qquad \left. e^{-\frac{(x-4)^2}{2}} + e^{-\frac{(x-7.4)^2}{2}} \right]\end{aligned}$$

The "Density Histogram and Kernel Density Estimate" figure contains the graph of this function. Using this method, we get an approximate density function from the sample data. Furthermore, if the kernel used is a valid PDF, then this density function is a valid PDF, i.e., it is normalized: $\int_{\mathbb{R}} \widehat{f}(x) dx = 1$. Once we have this approximate PDF, we can compute everything we would compute with any other PDF. For instance we can compute probabilities such as $\widehat{\mathbb{P}(X < 3)} = \int_{-\infty}^{3} \widehat{f}(x) dx$. We could also compute moments, for example, the mean $\widehat{\mathbb{E}[X]} = \int_{\mathbb{R}} x \widehat{f}(x) dx$.

The window size (bandwidth), h, affects the variability of the estimated density function. When h is small, the $\widehat{f}$ has more variability. When h is large, the $\widehat{f}$ is smoother and has less variability. Explore the figure below, which estimates the density function using a Gaussian kernel function for different window sizes.

Gaussian Kernel Density Estimate with Various Window Sizes

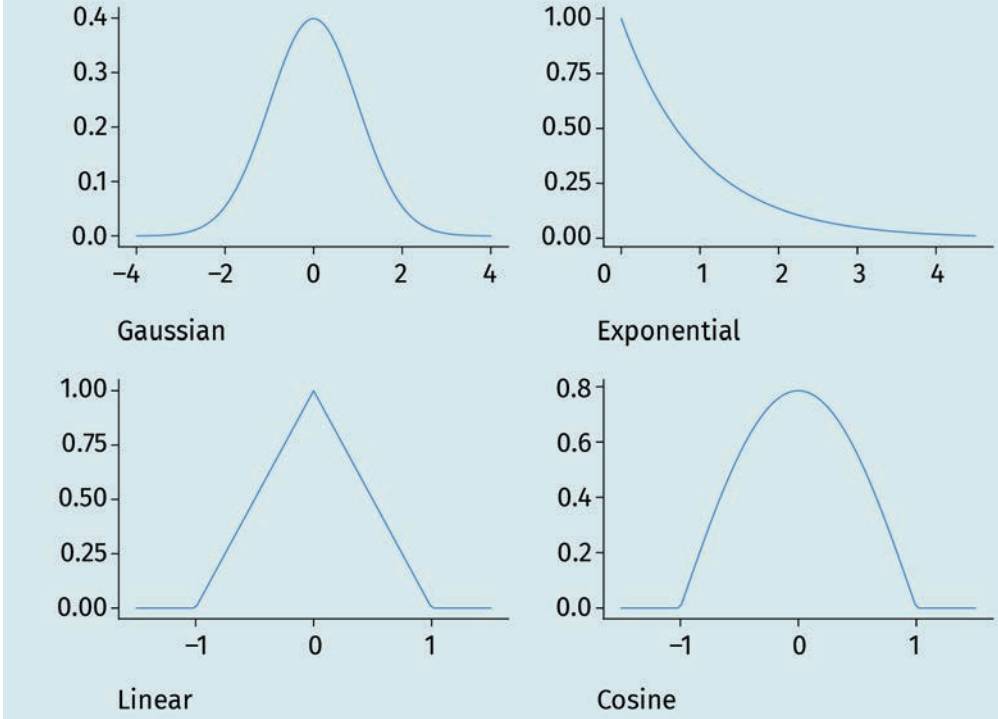


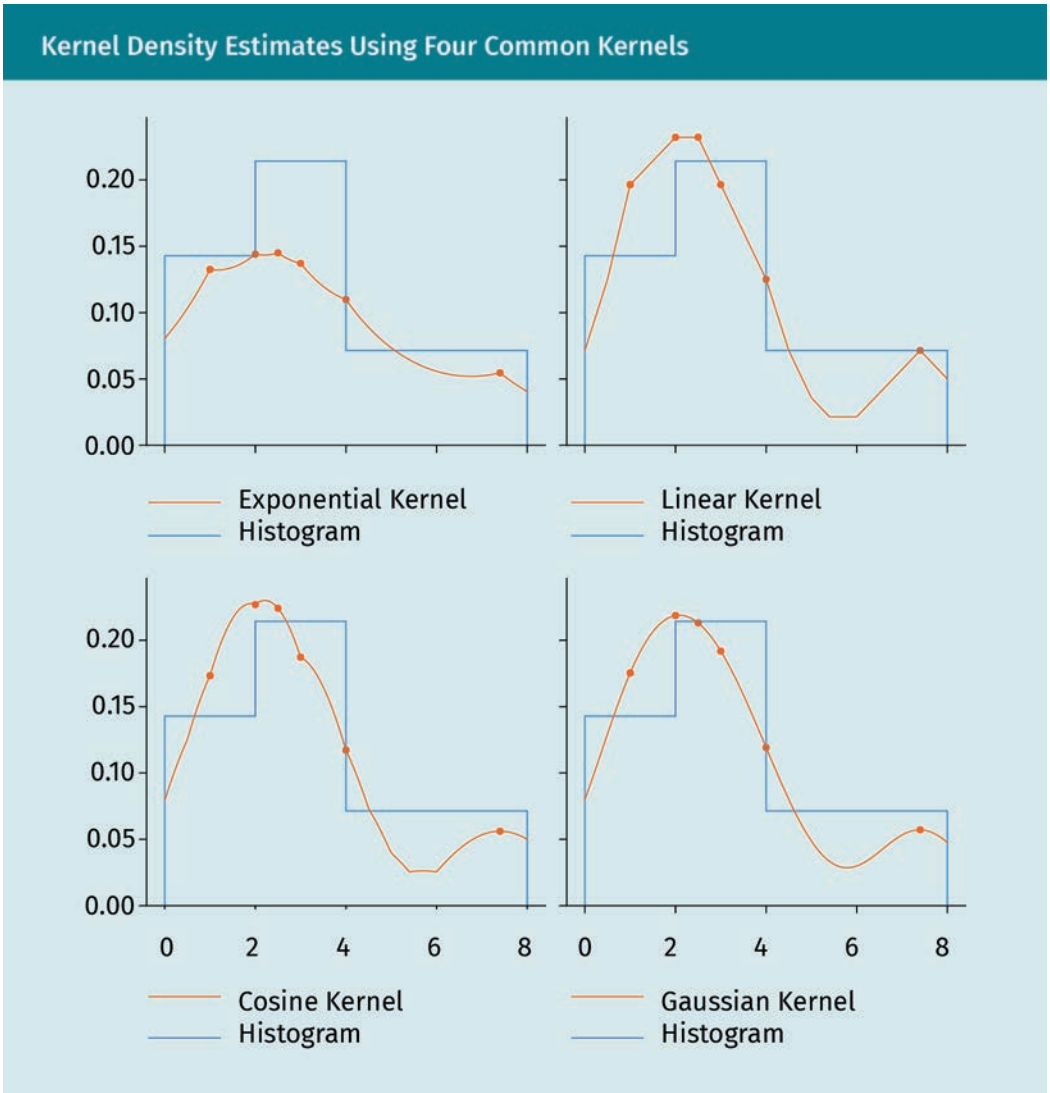
The Gaussian kernel is only one option in kernel density estimation. Although it is quite popular, there are other commonly used kernels. Among them are the exponential, linear, and cosine kernels. Including the Gaussian, the four are summarized in the table and figure below.

Four Kernels	
Name	Function
Gaussian	$\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}, -\infty < x < \infty$
Exponential	$e^{-x}, x \geq 0$
Linear	$1 - \lvert x \rvert, -1 \leq x \leq 1$
Cosine	$\frac{\pi}{4}\cos\left(\frac{\pi x}{2}\right), -1 \leq x \leq 1$

Typically, the choice of the kernel as well as the window size to use for a particular problem depends on the data set. The figures below show our kernel density estimates using the four common kernel functions.

Graphs of the Four Common Kernels





As our first introduction to non-parametric techniques, we have discussed Parzen windows. This techniques allows us to estimate the probability density function using only the data. The quality of the resulting approximate probability density is highly dependent on the choice of kernel (linear, exponential, Guassian, etc.) as well as on the choice of the window size.

3.4 K-Nearest-Neighbors

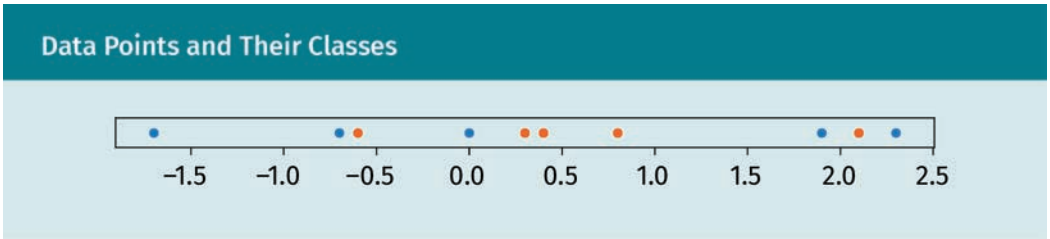
K-nearest neighbors (k-NN) has two main applications. In this section, we will discuss both. As a preview, the first application is in classification. The k-NN approach to classify a new unclassified data point is to find the k closest data points from the seen (classified) data (the neighbors) and then choose the class that is the most popular.

The second application of *k*-NN is as a tool to approximate the probability density directly from the data. As such, it is a non-parametric technique very much like the Parzen windows from section 3.3.

The problem we will discuss first is how to classify a new (unseen) data point, x, as one of two classes, c_1 or c_2, based on other data points whose classes we already know. The table below shows ten data points and their respective classes.

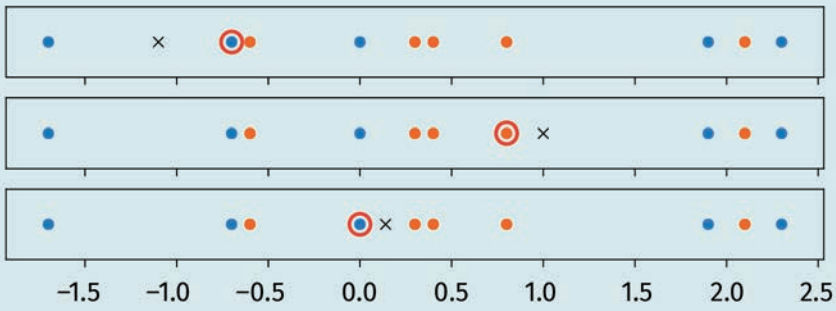
Data Points and Classes										
Point	-1.7	-0.7	-0.6	0.0	0.3	0.4	0.8	1.9	2.1	2.3
Class	1	1	2	1	2	2	2	1	2	1

The figure below shows a plot of this data. Points from different classes have different colors.



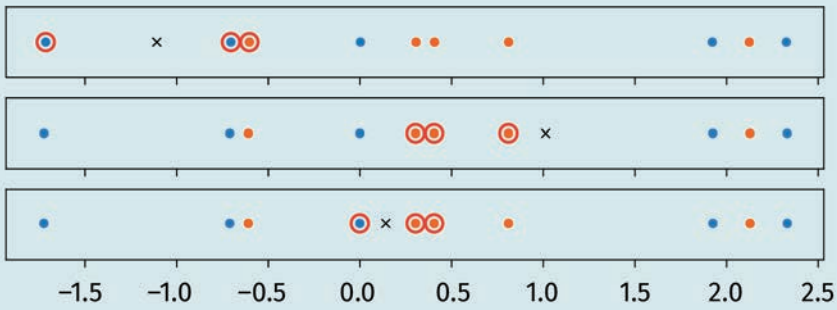
Suppose we are given a new data point: $x = -1.1$. Should we classify it as class 1 or class 2? One way to do this is to look at its "closest neighbor," i.e., the point in the data set that is closest to it. If we calculate the distance from $x = -1.1$ to every point in the data set, the one with the smallest distance (the point closest to $x = -1.1$) is -1.7 and belongs to class 1. Therefore, based on this 1 closest neighbor, we classify our point as class 1. Now consider a different point, $x = 1$, the closest neighbor to this point is 0.3, and this point belongs to class 2; thus, we would assign the new point $x = 1$ to class 2. Similarly, we would assign $x = 0.14$ to class 2; its closest neighbor, 0, belongs to class 2. What we have just described is the method of classification using 1-NN, one-nearest-neighbor. The figure below shows three plots, each of which contain our data set along with the three "test" points we wanted to classify. The circled data point is the one closest to the test point.

An Example of a Two-Class Classification with 1-NN with Scalar Data Points



Instead of using only one closest neighbor, let's now use the three closest neighbors, 3-NN. The three closest points for $\mathrm{x} = -1.1$ are $\{-.7, -0.6, -1.7\}$ and all these points belong to class 1, so if we take these 3 neighbors into account they all "vote" for class 1, and we would classify our test point to this class. Next, let's consider the second test point, $\mathrm{x} = 1$, the three closest points (neighbors) are $\{0.4, 0.3, 0.8\}$. The first two would "vote" for their class, class 2, and the last point for class 1. If we consider a majority vote, then this data point $\mathrm{x} = 1$ would be classified as class 2. The figure below shows three plots, each of which contains our data set along with the three "test" points we wanted to classify. The three circled data points in each graph are the closest to the test point.

An Example of a Two-Class Classification with 3-NN with Scalar Data Points



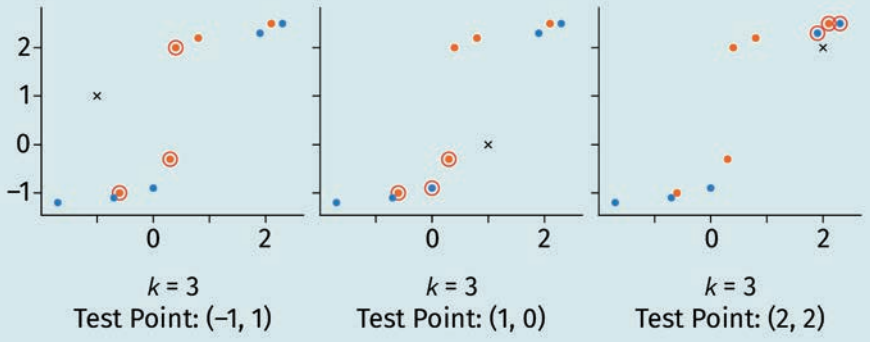
In an extreme case, we can look at the 10-NN classifier, the one that uses all the data points. Since 6 out of the 10 points belong to class 2, the 10-NN classifier would always assign every new test point to class 2. We consider the other extreme case of the 1-NN classifier. These two extremes are not very good in applications. Using too many neighbors starts using information for data points that may be too far, diminishing the idea of localized information. On the other hand, using too few neighbors is also not ideal since it uses only a very small part of the neighborhood of the data points. In practice,

there are some standard ways of choosing the ideal number of neighbors, the k in k-NN. We will discuss one way of doing this: cross-validation. Before discussing cross-validation, let's consider a two-dimensional data set, where each point is a pair of numbers, each pair belonging to one of two classes. The table below shows the data set we will use.

Two Dimensional Data in Two Classes	
Point	Class
$x_1 = (-1.7, -1.2)$	1
$x_2 = (-0.7, -1.1)$	1
$x_3 = (-0.6, -1.0)$	2
$x_4 = (0.0, -0.9)$	1
$x_5 = (0.3, -0.3)$	2
$x_6 = (0.4, 2.0)$	2
$x_7 = (0.8, 2.2)$	2
$x_8 = (1.9, 2.3)$	1
$x_9 = (2.1, 2.5)$	2
$x_{10} = (2.3, 2.5)$	1

Given a test point $x = (-1, 1)$, we compute the distances to all the data points by $d_i = ||x - x_i||$ for $i = 1, ..., 10$ and then order from least to greatest: $d_{(1)} \leq d_{(2)} \leq \cdots \leq d_{(10)}$. The 1-NN classifier would choose the class of $x_{(1)}$. The 3-NN classifier would choose the majority of the class labels among $\{x_{(1)}, x_{(2)}, x_{(3)}\}$. For the given test point, the three closest distance are $d_{(1)} = 1.72$ corresponding to $x_{(1)} = x_6 = (0.4, 2)$, $d_{(2)} = 1.84$ corresponding to $x(2) = x_5 = (0.3, -0.3)$, and $d_{(3)} = 2.04$ corresponding to $x_{(3)} = x_3 = (-0.6, -1.0)$. The three data points belong to class 2. Therefore, the 3-NN (and 1-NN) would assign this data point to class 2. The figure below shows three graphs each containing the data set, a test point, and the three closest neighbors.

An Example of a Two-Class Classification with 3-NN with Two-Dimensional Data Points



Now that we have seen some simple examples of *k*-NNs, let's introduce some notation to formalize our discussion. The given data set will be denoted by $\mathcal{D} = \{(x_1, y_1)..., (x_n, y_n)\}$ where $x_i \in \mathbb{R}^d$ and $y_i \in \{1, ..., C\}$ for $i = 1, ...n$. In this setup, each data point is a pair consisting of a point in d-dimensional space and an integer denoting the class of that data point. In the most general setting; we consider C classes. In the discussion above, we were working with a binary classification problem where $C = 2$, in which case $y_i \in \{1, 2\}$. Given a test point $x \in \mathbb{R}^d$, the *k*-NN classifier assigns the most probable (majority vote) class among the k closest neighbors of x. We will denote the k closest neighbors of x by $N(x)$. Next, we denote by $N_1(x)$ the data points in $N(x)$ which belong to class 1 and $N_2(x)$, the data points which belong to class 2. Using this notation, the class assigned to x, $y(x)$ is given by

$$y(x) := \arg\max_c \frac{|N_c(x)|}{|N(x)|} = \arg\max_c \frac{|N_c(x)|}{k}$$

where $|N_c(x)|$ is the number of elements in the set $N_c(x)$. In other words, $|N_1(x)|$ gives the number elements in the k closest data points that are assigned to class 1. This definition of $y(x)$ is exactly the same thing we did for the numerical examples at the beginning of this section. The class assigned to a data point x is the one that has the highest proportion among the neighbors considered.

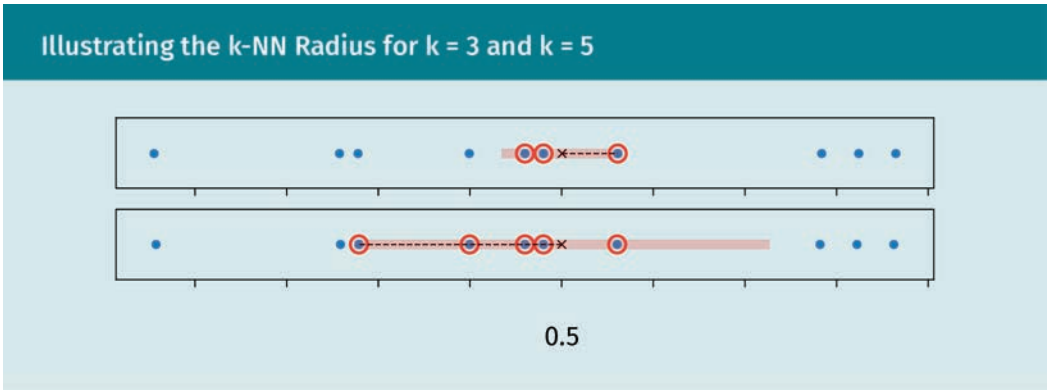
Another application of the ideas presented here can be used to approximate the density of the distribution from which a random sample is drawn. Given a random sample $x_1, ..., x_n$ of size n, we define the *k*-NN radius, $r_k(x)$ of a point x to be the k^{th} largest distance among the distances between x and each of the sample points. As before, let $d_i = |x - x_i|$, be the distance between x_i and x for $i = 1, ..., n$. Order these distance in non-decreasing order: $d_{(1)} \leq d_{(2)} \leq \cdots \leq d_{(n)}$. With this notation, the *k*-NN radius is $r_k(x) = d_{(k)}$. The estimated density of x is given by

$$\widehat{f}_k(x) = \frac{k}{2nr_k(x)}$$

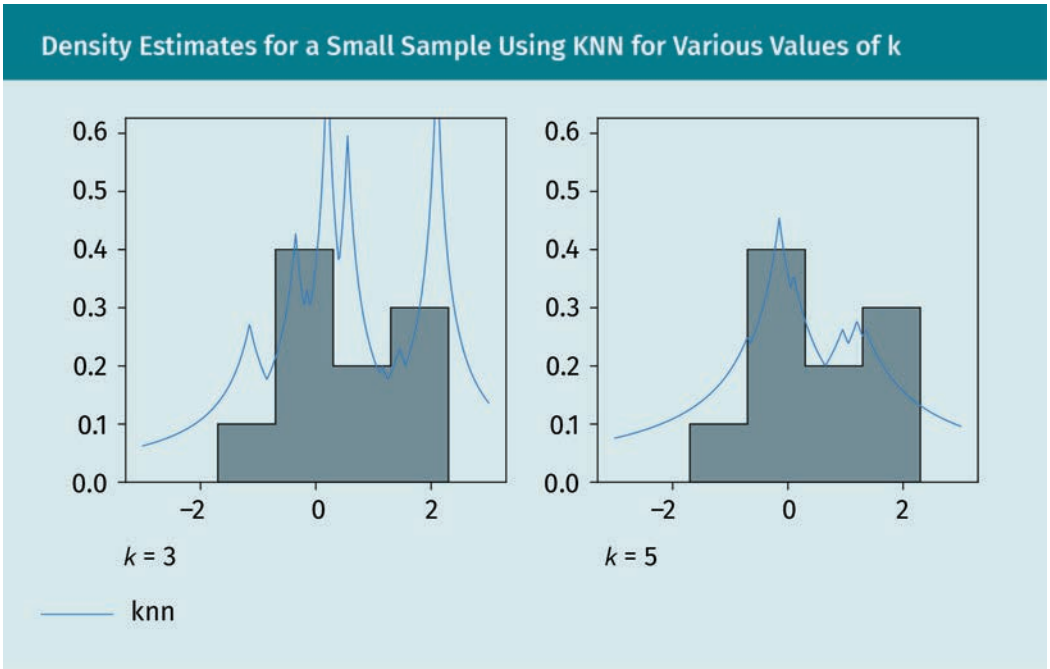
The motivation of this formula is that $2r_k(x)$ is the volume of the 1-dimensional ball centered at x, with radius $r_k(x)$. This ball contains k out of the n sample points, thus the presence of the factor k/n. The figure below shows the sample data along with the distances to the point $x = 0.5$.

Sample Data and Distance to a Given Point: 0.5										
Sample	-1.7	-0.7	-0.6	0.0	0.3	0.4	0.8	1.9	2.1	2.3
Distances	2.2	1.2	1.1	0.5	0.2	0.1	0.3	1.4	1.6	1.8

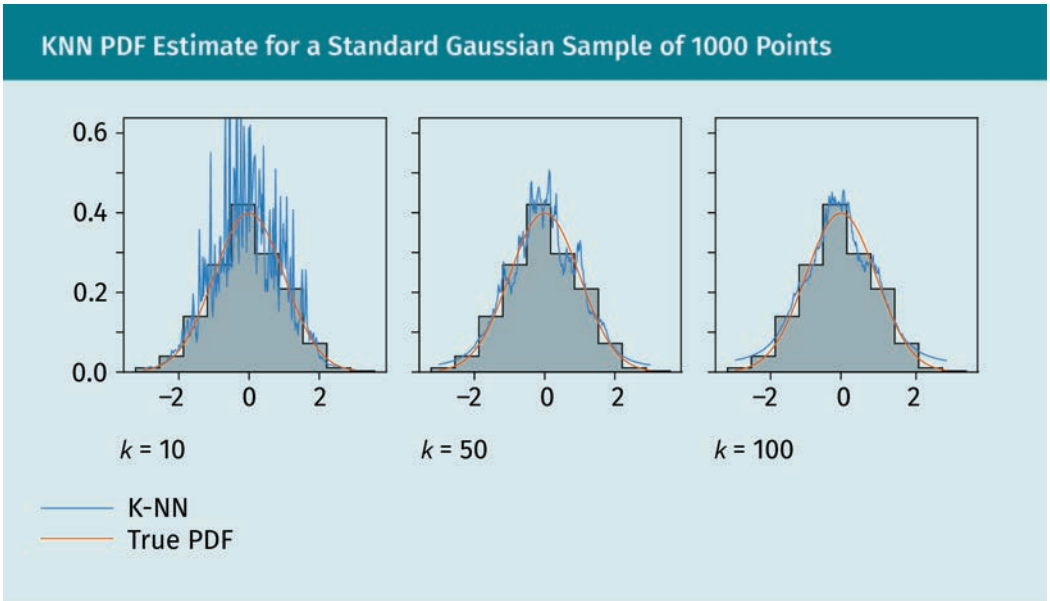
If we order the distances; we have $\{0.1, 0.2, 0.3, 0.5, 1.1, 1.2, 1.4, 1.6, 1.8, 2.2\}$. Now, using $k = 3$, we have $r_k(x) = r_3(0.5) = 0.3$, the third largest distance. As such, the estimate for the density is given by $\widehat{f}_k(x) = \widehat{f}_3(0.5) = \frac{3}{2 \cdot 10 \cdot 0.3} = 0.5$. Now let's see what we get with a different value of k. With $k = 5$, we have $r_5(0.5) = 1.1$, so $\widehat{f}_5(0.5) = \frac{5}{2 \cdot 10 \cdot 1.1} \approx 0.227$. In the figure below, we illustrate $r_k(x)$ for this sample data with $x = 0.5$ and $k = 3, 5$. The semi-transparent (pink) solid line shows the 1-dimensional ball centered at x with radius $r_k(x)$. The radius is illustrated as the dashed line connecting the point x to the k^{th} farthest point. The circled data points fall in the ball.



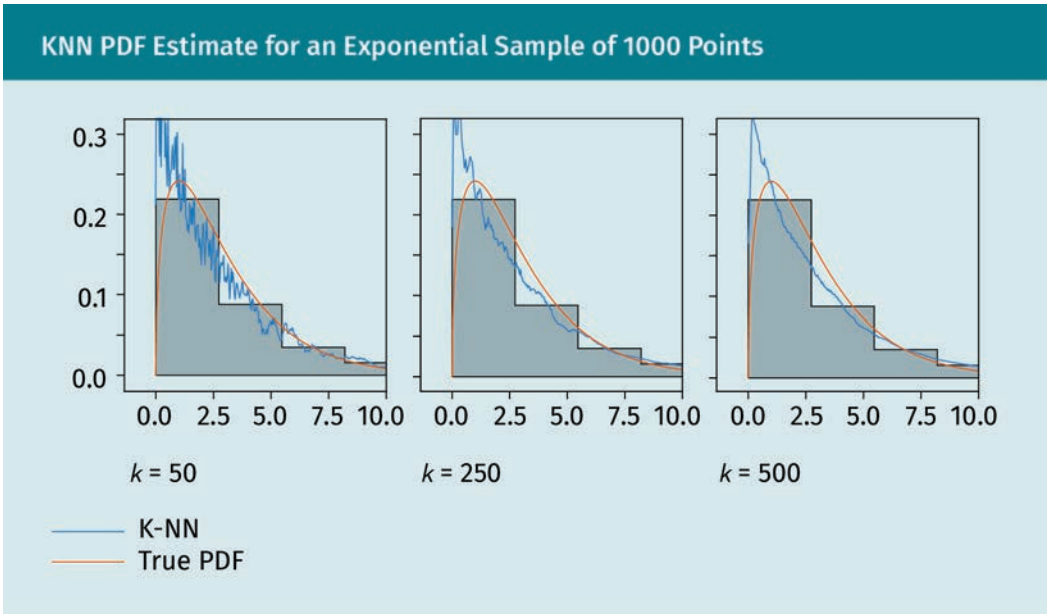
The figure below shows estimates of the density estimated using the *k*-NN approach with various values of k together with the histogram of the data.



Finally, we will show how the *k*-NN estimates of the PDF work with larger samples. We generate $\mathrm{n} = 1000$ observations from a standard Gaussian distribution $\mathcal{N}(0, 1)$ and compute the *k*-NN estimate $\widehat{\mathrm{f}}_{\mathrm{k}}(\mathrm{x})$ for $-3 \leq \mathrm{x} \leq 3$ for $\mathrm{k} = 10, 50,$ and 100. The figure below shows the graphs of the true PDF, a histogram of the data, and the *k*-NN estimate for these various *k*.



For another demonstration, we generate $\mathrm{n} = 5000$ observations from an exponential distribution with rate $\lambda = 1/3$, that is $\mathrm{Exponential}(1/3)$. The *k*-NN estimate $\widehat{\mathrm{f}}_{\mathrm{k}}(\mathrm{x})$ are computed for $0 \leq \mathrm{x} \leq 10$ for $\mathrm{k} = 50, \quad 250$, and 500. The figure below shows the graphs of the true PDF, a histogram of the data, and the *k*-NN estimate for these various k.



For both large sample estimates of the PDF using *k*-NN, notice that $\widehat{\mathrm{f}}_{\mathrm{k}}$ is quite noisy for small values of k and less so for larger k.

One advantage of this approach to estimate a PDF is that it can be easily extended to higher dimensional data. For example, if our data is two-dimensional, then by replacing the factor $2\mathrm{r_k(x)}$ by $\pi \mathrm{r_k(x,y)}^2$, we get the estimate $\widehat{\mathrm{f}}_\mathrm{k}(\mathrm{x},\mathrm{y})$ for the joint PDF $\mathrm{f(x,y)}$. For three dimensions, we can replace the factor $2\mathrm{r_k(x)}$ with $\frac{4\pi}{3}\mathrm{r_k(x,y,z^3)}$ and use it to get the estimate $\widehat{\mathrm{f}}_\mathrm{k}(\mathrm{x},\mathrm{y},\mathrm{z})$ of the joint PDF $\mathrm{f(x,y,z)}$.

In this section, we have discussed the concept of using k-nearest neighbors as a classification technique and as a density estimation technique. For the former, we have seen how new unclassified data can be classified using information from the closest data points whose classes we know. Adapting this idea to density estimation, we have seen how to apply the idea of *k*-NN to estimate the density, point-wise, from observed data. Just as with Parzen windows, the *k*-NN is a non-parametric technique. It doesn't try to estimate parameters with some pre-determined family of distributions. Unlike Parzen windows, the method enables a non-fixed window size, as this is determined by the radius, $\mathrm{r_k}$ from the data. In this sense, *k*-NN is a more flexible technique.

Summary

In section 3.1, upon review of relevant concepts from probability, we used Bayes' formula to develop a Bayesian treatment of parameter estimation. The key concepts in this section were the prior distribution, the likelihood, evidence, and the posterior distribution. The key difference between frequentist statistics and Bayesian statistics is that in the latter the unknown parameter of interest is treated as a random variable. Our beliefs about this parameter are encoded in a chosen prior distribution. Given some observed data, we can update this distribution using the the relation $\mathrm{posterior} = (\mathrm{likelihood} \times \mathrm{prior})/\mathrm{evidence}$ to get a posterior distribution of the parameter. From this distribution, we can use the mean or mode to obtain a Bayes estimate of the target parameter.

The Bayes estimate is sensitive to the chosen prior distribution. In section 3.2, we discussed some ways in which the prior distribution can be chosen. In particular, we distinguished between uninformative, weakly informative, and informative priors. A key type of prior is that which comes from Jeffrey's rule. A rule that enables us to find an uninformative or weakly informative prior which is invariant to a (reversible) transformation of the parameter.

In section 3.3 and 3.4, we introduced two non-parametric techniques. Parametric techniques require that a distribution family is assumed. Our goal is to approximate the parameter (or parameters) which best fit the observed data. In contrast, non-parametric techniques don't make any assumptions about the underlying distribution from which the data was generated. Instead, we seek to directly model the probability density function purely based on the data. In section 3.3, we discussed Parzen windows. Here, we have to choose the kernel and the window size. We discussed how different choices of the kernel and window size affects the character of the resulting approximate PDF. In section 3.4, we briefly introduced k-nearest neighbours first as a technique to address classification problems. The section ended

with an adaptation of k-NN for approximating PDFs. Unlike Parzen windows, we don't need to choose a kernel, and the technique of k-NN automatically chooses a varying window size by computing the relevant “radius” quantity directly from the data.

Knowledge Check

Did you understand this unit?

You can check your understanding by completing the questions for this unit on the learning platform.

Good luck!

Unit 4

Statistical Testing

STUDY GOALS

On completion of this unit, you will have learned...

- ... the general framework and how to interpret the result of hypothesis testing.
- ... how to conduct and interpret some common non-parametric hypothesis tests.
- ... how to conduct and interpret some common parametric one- and two-sample tests.
- ... how to define and interpret p-values.
- ... how to construct and interpret confidence intervals.
- ... how to control two different measures of error related to multiple hypothesis testing.

4. Statistical Testing

Introduction

Statistical testing lies at the core of statistical inference. Also called hypothesis testing, statistical testing involves structured paradigms which use results from probability theory to quantify the evidence, or lack of evidence, that an observed sample provides towards a claim. A diverse set of claims can be addressed with statistical testing, as such, it is used in almost every industry. Applications include marketing, finance, medicine, and psychology. In this unit, we will discuss the most common statistical tests, learn how to conduct them, interpret their results, and, most importantly, understand their limitations and the assumptions under which these tests can be applied. The tests we have chosen to include here are both popular and require only the fundamental mathematical knowledge expected in this course.

Suppose that the prevalence of breast cancer for females aged 54—65 is two percent We are interested in the prevalence of breast cancer for females aged 54—65 whose mothers were diagnosed with breast cancer. To this end, we sample 10,000 females who fit this criteria. Out of this sample, 400 have developed breast cancer. Thus, the sample prevalence of breast cancer is $400/10,000 = 0.04 = 4\%$. We want to test whether the true prevalence of breast cancer among females whose mothers had breast cancer is different than the prevalence among all females aged 54—65. The statistical test that enables us to examine such claims will be the subject of section 4.1. This section will introduce the idea of hypothesis testing with regards to claims made about an unknown parameter for a single population of interest.

Many of the hypothesis tests we discuss in this unit are parametric tests, statistics tests about parameters of a distribution of the underlying population(s). In section 4.2, we discuss some common parametric tests. Many statistical models assume that the underlying population has a Gaussian distribution. To this end, we present the Kolmogorov-Smirnov test of normality, a non-parametric test that determines how well the sample data fit a pre-determined Gaussian distribution. Another goodness-of-fit test uses the χ^2 distribution to assess how well count data (frequency distribution) fits a proposed categorical distribution. Finally, another χ^2 test is discussed, which helps evaluate whether two categorical variables are dependent.

At a high-level, sections 4.1 and 4.2 discuss hypothesis tests involving a claim about data that is generated from a single population, “one-sample tests.” In section 4.3, we introduce hypothesis tests which compare parameters from two populations. For example, given two samples, one from each of two independent samples, we will learn how to conduct a hypothesis test that compares the true means of the populations. As an example from user interface (UI) and user experience (UX), as applied to a business setting, let's consider the example below.

The click-through rate (CTR) is defined as the proportion of users who click on a call-to-action link on a Web page relative to the total number of users who viewed it. In an effort to increase the CTR of a certain page, a Web designer decides to test a new design for the call-to-action button. The current background color of the button is light

blue. Let's call the current design A. The proposed design of the call-to-action button is to make the background color light green. This proposed design will be denoted by B. To avoid confounding the experiment with other variables, everything on the Web page, besides the background of the call-to-action button, is kept the same. The page will now serve one of A or B at random to each visitor. In other words, about 50 percent of the visitors will see design A while others will see design B. After some time, the CTR will be computed for all the visitors who experienced each of the designs; let's call these proportions $\widehat{\pi}_{\mathrm{A}}$ and $\widehat{\pi}_{\mathrm{B}}$. These two will serve as estimates to the true proportion of CTRs from the respective designs π_{A} and π_{B}. We want to find out if the collected data provides evidence statistically significant to support the claim that $\pi_{\mathrm{A}} < \pi_{\mathrm{B}}$. In other words, should we expect that the new design will perform better with respect to CTR? Section 4.3 will allow us to evaluate this scenario using A/B testing, a marketing term meaning "two sample hypothesis testing." In general, this section will help us use hypothesis testing with regards to claims made about how analogous unknown parameters of interest are related to one another.

In section 4.4, we continue our discussion of hypothesis testing. We discuss in detail the trade-off between type I and type II errors and define the power of a test, which is a commonly overlooked but important consideration. In this section we also introduce the idea of p-values, an important but commonly misinterpreted value. We learn how to correctly compute and interpret p-values. Additionally, in this section, we learn about interval estimation as superior alternative to point estimates. While a point estimate gives us a single value as an estimate for an unknown parameter of interest, an interval estimate is a range of values, taking into account the uncertainty associated with the estimation process. We also discuss how confidence intervals may be used to evaluate hypotheses exactly like the ones we address in units 4.1 and 4.2.

In the final section, section 4.5, we discuss multiple hypothesis testing. Suppose that we want to grow tomatoes in a field and have four different soil hydration schedules. We would like to know if the average yield over a period of time, measured in kilograms, differs at a statistically significant level across the different hydration schedules. In this scenario, we have many hypotheses. Let's label $\mu_1, \mu_2, \mu_3, \mu_4$ as the true average yields from the four different hydration schedules. The status quo, the null hypothesis, would state that all the means are the same: no effect. The null hypothesis contains the statement H_0:$\mu_1 = \mu_2$, $\mu_1 = \mu_3$, $\mu_1 = \mu_4$, $\mu_2 = \mu_3$, $\mu_2 = \mu_4$, and $\mu_3 = \mu_4$. The alternative hypothesis, which looks for an effect, may be found by replacing $=$ with $\neq$ in any of these statements. If we accept a five percent rate of error for rejecting one true null hypothesis, then the probability of rejecting at least one true null hypothesis is much higher, more than 20 percent! In this section, we will learn two methods to control errors of this nature.

4.1 Hypothesis Tests and Test Statistics

Hypothesis testing is a four-part paradigm used to evaluate the statistical significance of observed data with respect to a pair of competing hypotheses (claims). The first part of this paradigm is to determine the hypotheses. The status quo, that is, the statement

of no effect or no change, is summarized as the null hypothesis and is denoted by H_0. The test hypothesis, the statement indicating the presence of an effect or a change, is summarized as the alternative hypothesis and is denoted by H_1. In the example from the introduction about the prevalence of breast cancer, the null hypothesis states that the prevalence of breast cancer for women aged 54—65 whose mothers had breast cancer is the same as all women aged 54—65 (2%). The alternative hypothesis states that the prevalence is different. Let π denote the prevalence of breast cancer for our population of interest. We can write the two hypotheses as

$$\mathrm{H}_0{:}\pi = 0\,.02; \quad \mathrm{H}_1{:}\pi \neq 0\,.02$$

In general, testing claims about a population proportion against a known value π_0 is

$$\mathrm{H}_0{:}\pi = \pi_0; \quad \mathrm{H}_1{:}\pi \neq \pi_0$$

In our case, $\pi_0 = 0\,.02$. Once a sample is obtained and the sample prevalence (proportion) is computed, due to randomness, it is unlikely to get exactly 0.02, even if the true prevalence in the population of interest is in fact 0.02. The sample prevalence may turn out to be 0.03, 0.021, 0.04, theoretically any number between 0 and 1. Of course if the true prevalence was in fact 0.02, it is more likely to observe a sample proportion near 0.02 than far from it. In fact, for a given sample size, the farther the sample prevalence is from 0.02, the more statistically significant the evidence is against our assumption that $\mathrm{H}_0{:}\pi = 0\,.02$. In other words, for a given sample size, the difference between the sample proportion and the assumed proportion, $\widehat{\pi} - \pi_0$ is an indicator of evidence against $\mathrm{H}_0{:}\pi = \pi_0$. The larger this difference, the more likely our assumption (the null hypothesis) is false. But what is the cutoff value? This brings us to the second part of the hypothesis testing paradigm: the significance level.

The significance level of a hypothesis test is the highest probability of incorrectly rejecting the null hypothesis. That is, the highest risk we are willing to take if we decide to reject the null hypothesis. Formally, the significance level α is the probability of rejecting a true null hypothesis

$$\alpha = \mathbb{P}(\text{reject } \mathrm{H}_0 \,|\, \mathrm{H}_0(\text{is true}))$$

Type I error
This error occurs from rejecting a true null hypothesis.

Type II Error
This error occurs from failing to reject a false null hypothesis.

The decision to reject a true null hypothesis is certainly an error. In statistical testing, this error is called a **type I error**. The cutoff value for the difference between the sample proportion and the assumed proportion that results in a decision to reject the null hypothesis comes from the significance level, α. This controls the probability of committing a type I error. Another error is also possible when conducting a hypothesis test: failing to reject a false null hypothesis. This error is known as a **type II error**, and its associated probability is denoted by β. The table below summarizes these two types of errors. The tricky part is that α and β are inversely related. If one is set to a low value, the other tends to be high. We will discuss the trade-off between α and β in a later section.

Type I and Type II Errors and (Probabilities)		
	True H_0	False H_0
Reject H_0	Type I error (α)	No error
Fail to reject H_0	No error	Type II error (β)

In practice, instead of working with the observed difference $\widehat{\pi} - \pi_0$, and trying to come up with a cutoff value, we work with a standardized quantity for which we know the (approximate) distribution. This standardized value is called the **test statistic**. The central limit theorem tells us that if the sample size is large, the quantity

Test statistic
This random variable standardizes the quantity measuring a departure from the null hypothesis. Usually, such a random variable is chosen so that its distribution is known, at least approximately.

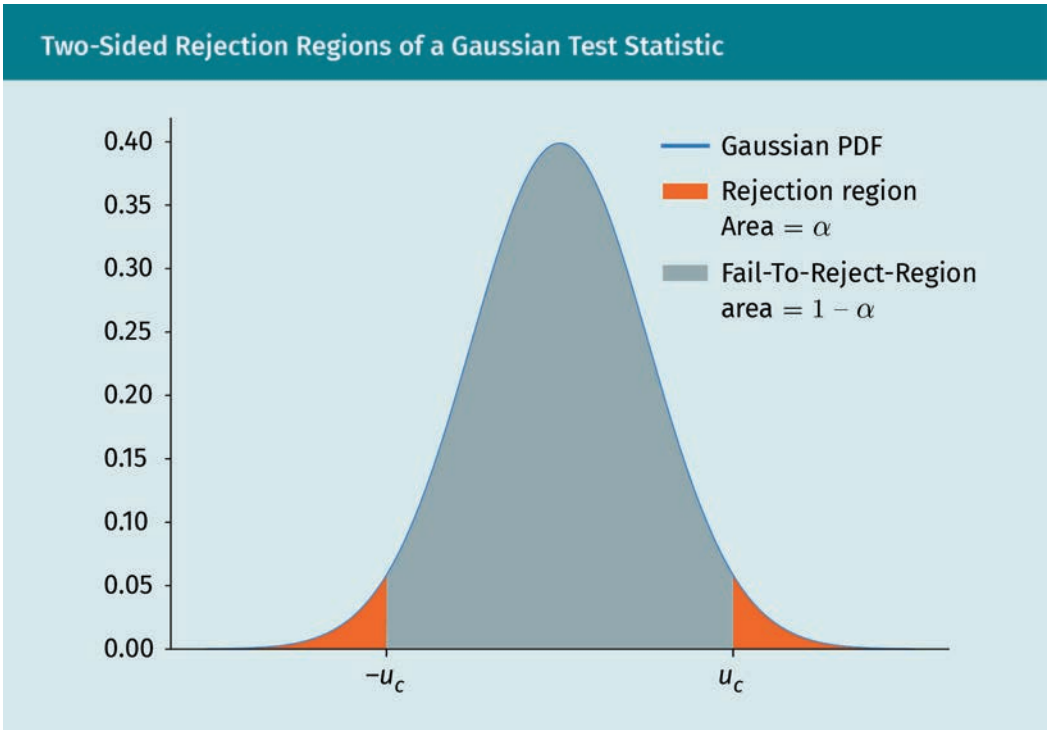
$$U = \frac{\widetilde{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

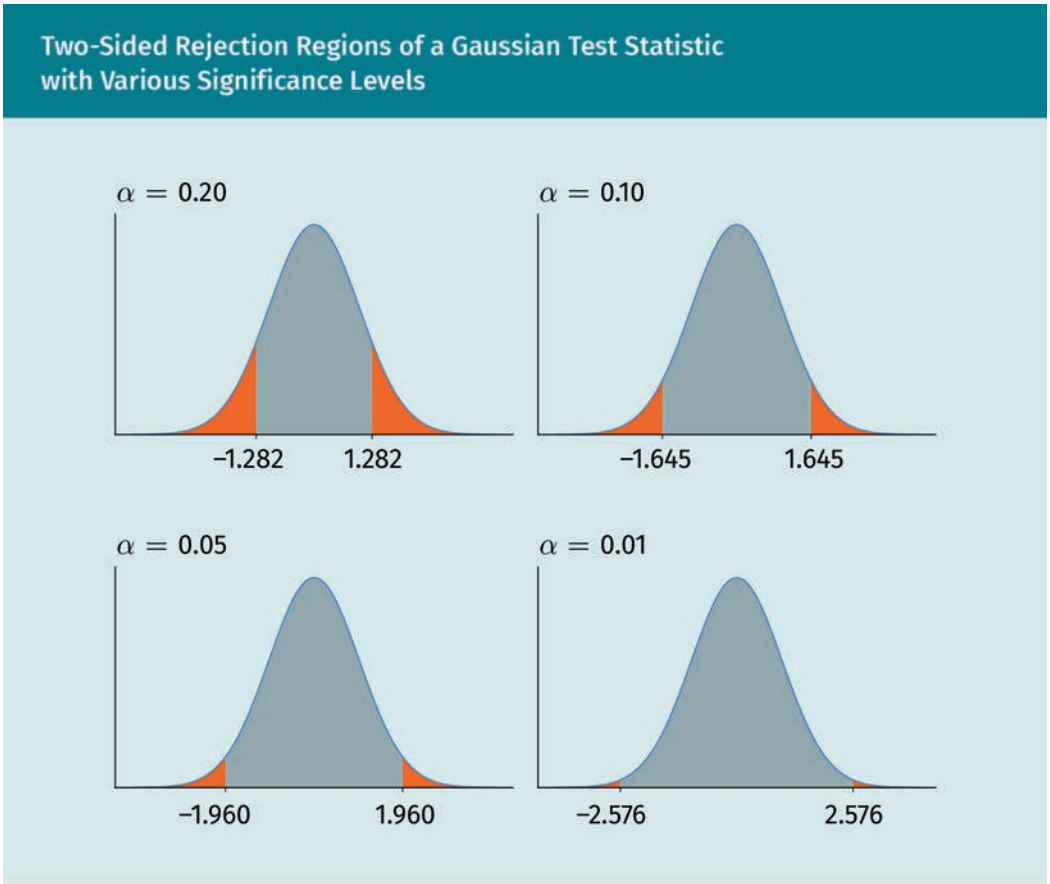
(given $\pi = \pi_0$), approximately follows a standard Gaussian distribution (Hogg et al., 2019), $U|H_0 \sim \mathcal{N}(0,1)$. Here $\widetilde{\pi} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the estimator of the sample proportion where $X_i = 1$ if the i^{th} subject in our sample has breast cancer and zero otherwise. Instead of coming up with a cutoff value for the difference $\widehat{\pi} - \pi_0$, we come up with a cutoff value of u such that rejecting H_0 when H_0 is true is α. In our case, when $|U|$ is large, we will tend to reject H_0. Our cutoff value(s) corresponds to u_c where

$$P(|U| > |u_c||H_0) = \alpha$$

Viewed a different way, $\mathbb{P}(U < u_c|H_0) = \alpha/2$ or $\Phi(u_c) = \alpha/2$ where Φ is the CDF of the Gaussian distribution. Once this cutoff value is determined, the **rejection region** is a set of values u such that $|u| > |u_c|$. The first figure below shows the Gaussian distribution together with the rejection region (in orange). The second figure below shows various cutoff values and corresponding rejection regions for various values of α

Rejection region
This set of values of the test statistic is associated with a total probability equal to the significance level and indicates departure from the null hypothesis in the direction(s) of the alternative hypothesis.


Two-Sided Rejection Regions of a Gaussian Test Statistic
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
$-u_c$
u_c
Gaussian PDF
Rejection region
Area $= \alpha$
Fail-To-Reject-Region
area $= 1 - \alpha$



As you might expect, the smaller the significance level, α, the smaller the rejection region. It is very important to note that the hypotheses as well as the significance level (and thus the rejection region) are determined before any data ar collected. Otherwise, the data may bias our choice of hypotheses as well as the significance level in order to favor one decision over another. In other words, choosing the hypotheses and/or the significance level after analyzing the observed data completely invalidates the integrity of the statistical test. In practice, it is unjustifiable to acquire/analyze data before these choices are made! Our choice of the hypotheses was already made; we will choose a one percent level of significance, i.e., $\alpha = 0.01$. The third part of statistical testing is to acquire data and compute the observed values of the quantities of interest. Ultimately, the observed value of the test statistic, u_{obs}.

Suppose that in a sample of $n = 10,000$ women aged 54—65 whose mothers had breast cancer, we find that 400 of them have breast cancer. The sample prevalence is $\widehat{\pi} = 400/10,000 = 0.04$. The observed value of the test statistic is

$$u_{obs} = \frac{0.04 - 0.02}{\sqrt{\frac{0.02(1-0.02)}{10000}}} \approx 14.29$$

From the figure above, which shows various two-sided rejection regions, this observed value is in the rejection region corresponding to $\alpha = 0.01$. This brings us to the fourth and final part: the decision. Since the observed value of the test statistic is in the rejection region, we reject the null hypothesis at the five percent level of significance. This means that the data provides evidence at the five percent significance level that the null hypothesis is false.

Let's summarize the four parts for this example:

1. Establish hypotheses: $H_0: \pi = \pi_0 = 0.02$ versus $H_1: \pi \neq \pi_0 = 0.02$
2. Set $\alpha = 0.01$ and the test statistic

$$U = \frac{\widetilde{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \overbrace{\sim}^{\text{approx.}} \mathcal{N}(0,1)$$

   The rejection region is $|u| > u_c = 2.576$

3. Compute the observed value of the test statistic from the sample data: $u_{obs} = 14.29$
4. The observed value, u_{obs} falls in the rejection region and therefore we reject H_0 at the five percent level of significance.

Two-sided test
In this statistical test, the null hypothesis indicates that the true parameter is different than what is claimed, yielding a rejection region composed of intervals extending in opposite directions.

The statistical test we performed on the breast cancer scenario was a **two-sided test**. This is because the alternative hypothesis, $\pi \neq \pi_0$, is equivalent to the statement $\pi < \pi_0$ or $\pi > \pi_0$. Viewed another way, the rejection region, $|u| > |u_c|$, was composed of the two regions $u < |u_c|$ and $u > |u_c|$. Sometimes, the statistical test incorporates some prior notion of the direction of the effect. In this case, a researcher may want to use a **one-sided test**. There are two possible one-sided tests: a left-tailed test in which the alternative hypothesis takes the form $\theta < \theta_0$, and a right-tailed test in which $H_1: \theta > \theta_0$. As we mentioned before, the choice of the hypotheses must be made without acquiring or analyzing the data. The rejection regions associated with a one-sided test is a single interval of values extending in only one direction. The rejection region of a left-tailed test extends to the left and that of a right-tailed test extends to the right. The figure below shows the rejection regions of the three types of alternative hypotheses together with the significance level and the cut-off value(s).

One-sided test
In this statistical test, the alternative hypothesis indicates a directional difference (less than or more than).

Example 4.1.1
Let $X_1, ..., X_n$ be independent with unknown mean μ and known standard deviation σ. If n is large, the central limit theorem assures us that test statistic defined by

$$U = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

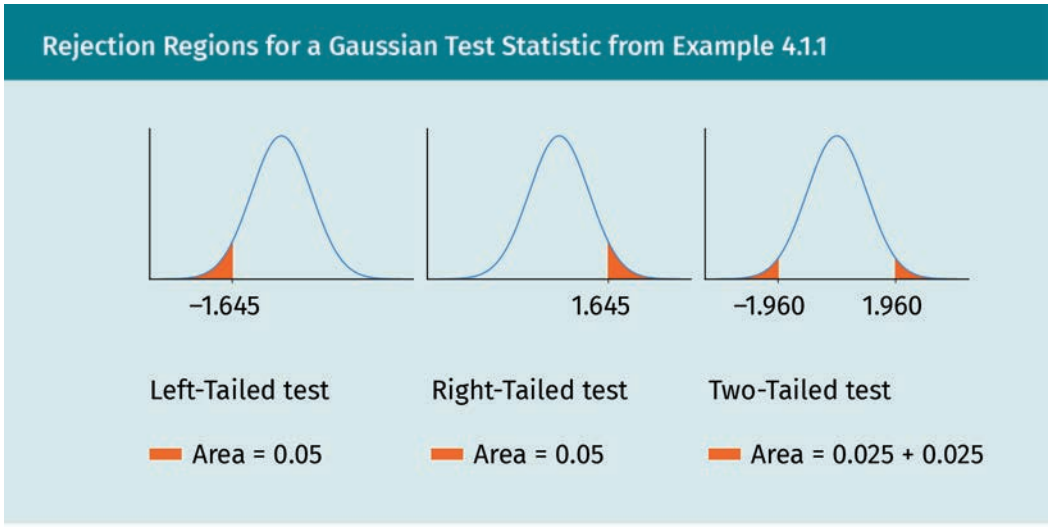
approximately follows the standard Gaussian distribution, that is, $U \overbrace{\sim}^{\text{approx.}} \mathcal{N}(0,1)$. Using $\alpha = 0.05$, find the rejection region corresponding to each of the following alternative hypotheses under the assumption that $H_0: \mu = \mu_0$ is true.

1. $H_1: \mu > \mu_0$
2. $H_1: \mu < \mu_0$
3. $H_1: \mu \neq \mu_0$

Solution

Let z_α denote the $(1-\alpha)100$ quantile of the Gaussian distribution. In other words, $\Phi(z_\alpha) = 1-\alpha$ or equivalently, $z_\alpha = \Phi^{-1}(1-\alpha)$. First, the cutoff value u_c must satisfy $\mathbb{P}(U > u_c) = \alpha = 0.05$ so $u_c = z_{0.05} = \Phi^{-1}(0.95) = 1.645$. The rejection region is $(1.645, \infty)$. The cutoff value u_c must satisfy $\mathbb{P}(U < u_c) = \alpha = 0.05$. By symmetry, we have $u_c = -z_{0.05} = -1.645$. The rejection region is $(-\infty, -1.645)$. This is a two-tailed test. The cutoff values u_c must satisfy $\mathbb{P}(|U| > |u_c|) = \alpha = 0.05$. In other words, $\mathbb{P}(U < u_{c_L}) = \alpha/2 = 0.025$ so $u_{c_L} = -z_{0.025} = -\Phi^{-1}(0.975) = -1.96$. Analogously, $u_{c_R} = 1.96$. The rejection region is therefore $(-\infty, -1.96) \cup (1.96, \infty)$.

The figure below shows the three rejection regions found in the previous example.



Example 4.1.2

Let $X_1, \ldots, X_n$ be independent with unknown mean μ and unknown standard deviation σ. The test statistic defined by

$$U = \frac{\overline{X} - \mu_0}{\frac{S}{\sqrt{n}}}$$

follows the student T distribution with $n-1$ degrees of freedom (Hogg et al., 2019), that is, $U \sim T(n-1)$, provided that either n is large enough to use the central limit theorem, or the underlying distribution is approximately Gaussian. Here S^2 is the sample variance (the unbiased estimator of variance). Using $\alpha = 0.05$ and a sample size of $n = 10$, find the rejection region corresponding to each of the following alternative hypotheses under the assumption that $H_0: \mu = \mu_0$ is true.

- $H_1: \mu > \mu_0$
- $H_1: \mu < \mu_0$
- $H_1: \mu \neq \mu_0$

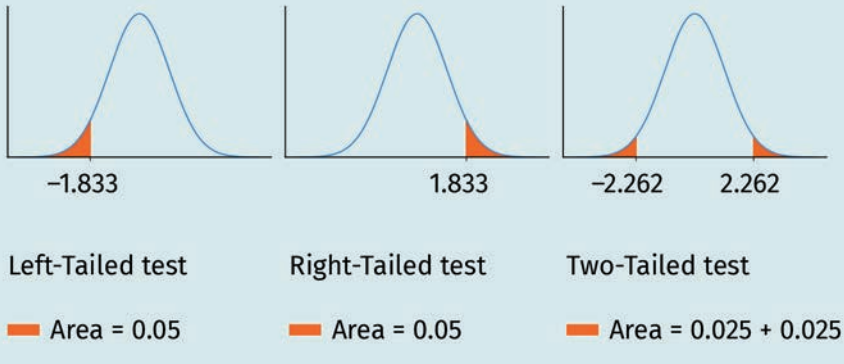
Solution

Let $t_{n-1,\alpha}$ denote the $(1-\alpha)100$ quantile of the T distribution with $n-1$ degrees of freedom. In other words, $\mathbb{P}\left(T(n-1) < t_{n-1,\alpha}\right) = 1-\alpha$. In this problem we have $n = 10$ so the distribution we will work with is $T(9)$.

1. The cutoff value u_c must satisfy $\mathbb{P}(U > u_c) = \alpha = 0.05$ so $u_c = t_{9,0.05} = 1.833$. The rejection region is $(1.833, \infty)$.
2. The cutoff value u_c must satisfy $\mathbb{P}(U < u_c) = \alpha = 0.05$. By symmetry, we have $u_c = -t_{9,0.05} = -1.833$. The rejection region is $(-\infty, -1.833)$.
3. This is a two-tailed test. The cutoff values u_c must satisfy $\mathbb{P}(|U| > |u_c|) = \alpha = 0.05$. In other words, $\mathbb{P}\left(U < u_{c_L}\right) = \alpha/2 = 0.025$ so $u_{c_L} = -t_{9,0.025} = -2.262$. Analogously, $u_{c_R} = 2.262$. The rejection region is therefore $(-\infty, -2.262) \cup (2.262, \infty)$.

The figure below shows the rejection regions from example 4.1.2.

Rejection Regions for a T(9) Test Statistic from Example 4.1.2



So far, we have seen three test statistics for conducting hypothesis tests. The first one was used for a statistical test involving π, the true proportion of a population of interest. The other two were used for statistical tests involving μ, the true mean of a population of interest. Examine the table below summarizing these three test statistics and when it is appropriate to use them.

Common Test Statistics for Parameters of Interest		
Parameter	Test Statistic	Assumptions

Common Test Statistics for Parameters of Interest		
π (true proportion)	$U = \frac{\widetilde{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \sim \mathcal{N}(0,1)$	$n \quad p > 10$ and $n \quad (1-p) > 10$
μ (true mean)	$U = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$	σ is known, and either n is sufficiently large for CLT, or the underlying distribution is Gaussian.
μ (true mean)	$U = \frac{\overline{X} - \mu_0}{\frac{S}{\sqrt{n}}} \sim T(n-1)$	S^2 is the sample variance and either n is sufficiently large for CLT, or the underlying distribution is Gaussian.

Example 4.1.3
A factory is accused of improperly releasing industrial waste into water reservoirs used by local farms. It is believed that the chemicals in this industrial waste causes birth defects, including low birth-weight. An independent council is charged with assessing the claim that average weight of newborn babies in this town is less than the national average of 3480 grams. The significance level is set at $\alpha = 0.01$. The birth weights of twelve newborn babies are sampled at random from the only hospital in town. The sample average of these weights is 3250 grams with a sample standard deviation of 250 grams. Use this data together with $t_{11,\,0.01} = -2.718$ to perform the appropriate statistical test. Assume that the weights approximately follow a Gaussian distribution.

Solution
Let μ denote the true average birth weight of babies in this town. Set $\mu_0 = 3480$ as the true average birth-weight of babies in the country to which we would like to compare. The null hypothesis is $H_0{:}\mu = \mu_0 = 3480$. Since the council is charged with assessing the claim that the true birth weight in this town is less than the population average, we have a one-tailed (left-tailed) alternative hypothesis: $H_1{:}\mu < \mu_0 = 3480$. Since the standard deviation of the birth weight in the country is not given, we will assume it is unknown and instead use $S = 250$ grams, the sample standard deviation. From the table above, the appropriate test statistic is

$$U = \frac{\overline{X} - \mu_0}{\frac{S}{\sqrt{n}}}$$

and its observed value is

$$u_{obs} = \frac{3250 - 3480}{\frac{250}{\sqrt{12}}} \approx -3.187$$

The rejection region is $u < u_c = t_{11,\,0.01} = -2.718$. Since the observed value of the test statistic falls in the rejection region, we reject the null hypothesis at a one percent level of significance that the average birth weight of newborns in this town is the same as the national average. In other words, the data provides evidence in favor of the alternative hypothesis.

Statistical tests help us make objective conclusions about the statistical significance of the data. However, it is not appropriate, solely based on these conclusions, to make recommendations and/or take action (or not to take action) in a real-world problem. In real-world applications, we must also consider the scientific and practical significance of the observed data. From the previous example, a obstretician should probably be consulted about whether the sample average is within the "normal" birth weight limits. Additionally, using the sample standard deviation, the doctor may ask for an interval estimate of the average rather than a point estimate and decide whether this interval estimate is entirely within normal birth weight limits. These are just some considerations that are beyond the scope of what a statistical test can ascertain.

So far we have discussed the following three statistical tests:

1. One-sample Gaussian test for testing claims about the true proportion π of a population against a baseline.
2. One-sample Gaussian test for testing claims about the true mean μ of a population against a baseline when the standard deviation, σ, of the population is known. This test assumes that the distribution of the population is (approximately) Gaussian. If this assumption is not valid, the sample size must be large enough for the central limit theorem to provide guarantees of the normality of the distribution of the sample mean.
3. One-sample T-test for testing claims about the true mean of μ of a population against a baseline when the standard deviation is unknown and replaced with the sample standard deviation. This test requires the same assumptions as the previous test.

4.2 Some Common Non-Parametric Tests

The statistical tests discussed in the previous section are called parametric tests because the hypotheses involve statements about parameters of distributions. In this section, we will discuss some common non-parametric tests. These are tests in which the hypothesis doesn't involve statements about population parameters.

The Chi-Square Goodness-Of-Fit Test

Suppose that a company is interested in the distribution of employee absences over work days: Monday to Friday. To this end, they want to test the hypothesis that the distribution of absences across the five working days varies against the null hypothesis that the distribution is uniform. The null hypothesis states that the distribution is uniform

$$\mathrm{H}_0 : \mathbb{P}(\mathrm{A}_{\mathrm{Mon}}) = \cdots = \mathbb{P}(\mathrm{A}_{\mathrm{Fri}}) = \frac{1}{5}$$

The alternative hypothesis, H_1, states that the distribution is different. We will set a significance level of $\alpha = 0.05$ and then collect some data. We collect data on expected and observed absences. The table below shows a summary of this data.

Observed and Expected Absences on Different Days of the Week					
	Weekday				
Number of absences	Monday	Tuesday	Wednesday	Thursday	Friday
	140	158	152	150	200
	(160)	(160)	(160)	(160)	(160)

Each of the five cells in this table contain the observed and expected counts. The total number of observed absences is $140 + 158 + \cdots + 200 = 800$. Therefore, under the null hypothesis the expected number of counts for Monday is $\frac{1}{5} \cdot 800 = \mathrm{E}_{\mathrm{Mon}}$. The observed number of absences on Monday is $\mathrm{O}_{\mathrm{Mon}} = 140$. In fact, the expected counts are all equal under the null hypothesis: $\mathrm{E}_{\mathrm{Mon}} = \cdots \mathrm{E}_{\mathrm{Fri}}$. None of the observed counts match the expected counts, however, this could still happen because of randomness even if the null hypothesis were true. We want to know if departure of the observed counts from the expected counts are statistically significant. To this end, we will use a standardized quantity, a test statistic, that quantifies the overall departure of the observed counts from the expected counts. This test statistic is defined by

$$\mathrm{U} = \sum_{\mathrm{i} \in \{\mathrm{Mon}, \ldots, \mathrm{Fri}\}} \frac{(\mathrm{O}_\mathrm{i} - \mathrm{E}_\mathrm{i})^2}{\mathrm{E}_\mathrm{i}}$$

This quantity is, of course, random because it depends on the observed counts O_i. The distribution of U is $\chi^2(4)$: a chi-square distribution with $\nu = 5 - 1 = 4$ degrees of freedom. The larger the difference between the observed and expected counts, the larger

the observed value of this test statistic. Thus, large values of U provide evidence against the null hypothesis. The $1-\alpha = 1-0.05 = 0.95$ quantile of $U \sim \chi^2(4)$ is $u_c = 9.488$. If the observed value u_{obs} is larger than this value, we will reject the null hypothesis at $\alpha = 0.05$ level of significance. Otherwise, we will fail to reject the null hypothesis. The observed value is calculated by

$$u_{obs} = \frac{(140-160)^2}{160} + \cdots + \frac{(200-160)^2}{160} \approx 13.55$$

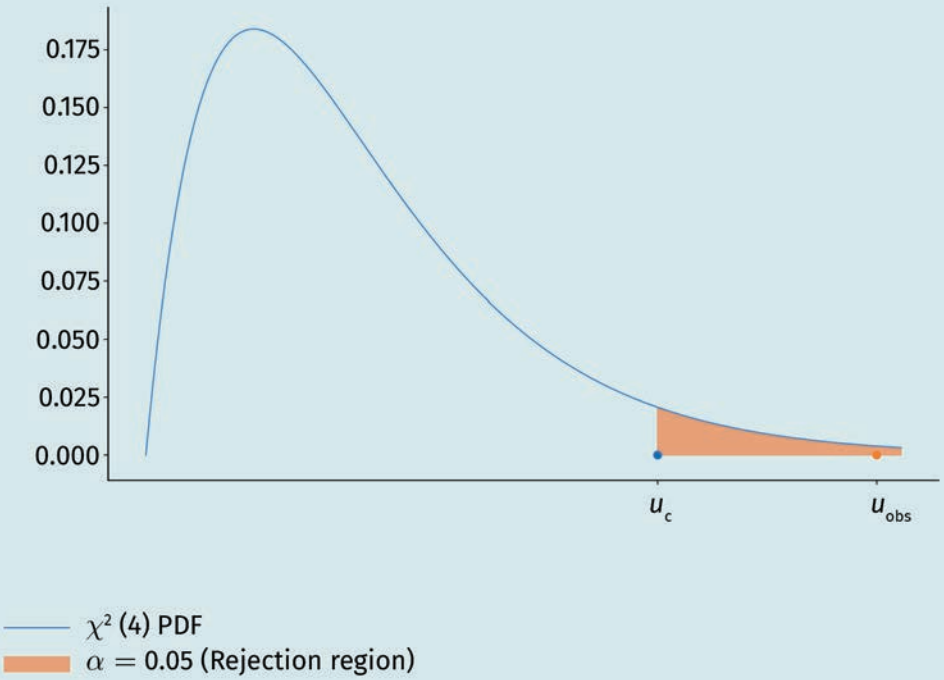
Thus, $u_{obs} > u_c = 9.488$ and we can conclude that the data provide evidence against the null hypothesis. We will reject the hypothesis that the distribution of absences across the days of the week are uniform in favor of the alternative hypothesis that the distribution is not uniform (at the five percent level of significance).

A χ^2 distribution is characterized by one parameter, ν, called the degrees-of-freedom. Although we will not use the PDF directly, it is given below as a reference. The most common statistical packages (Excel, R, Python (SciPy)), have numeric implementations of the PDF, CDF, and Inverse CDF of the χ^2 distribution. Note that the the support of the χ^2 distribution is non-negative and is used to model random variables that are non-negative.

$$f(x) = \begin{cases} \dfrac{1}{2^{\nu/2}\Gamma\left(\nu/2\right)} x^{\nu/2-1} e^{-x/2} & x > 0 \text{ and } \nu = 1 \text{ or } x \geq 0 \text{ and } \nu > 1 \\ 0 & \text{otherwise} \end{cases}$$

The figure below shows a graph of the PDF of $\chi^2(4)$ used above as well as the values of u_c and u_{obs}.

PDF of the Chi-Square Distribution with Four Degrees of Freedom and a Right-Rejection Region at a Five Percent Level of Significance



The χ^2 goodness-of-fit test uses observed counts that fall into pre-defined and exhaustive categories. To compare the empirical distribution to the theoretical distribution defined in the null hypothesis, let's suppose that a categorical variable has K possible values (classes), denoted as $\mathrm{k} = 1, 2, ..., \mathrm{K}$. The null hypothesis specifies the (discrete) categorical distribution $\mathrm{p}_1 = \mathbb{P}(\mathrm{A}_1)$, the probability that the variable belongs to class $\mathrm{k} = 1$, $\mathrm{p}_2 = \mathbb{P}(\mathrm{A}_2)$, the probability that the variable belongs to class $\mathrm{k} = 2$, and so on until $\mathrm{p}_\mathrm{k} = \mathbb{P}(\mathrm{A}_\mathrm{k})$. In the previous discussion, the categories were the days of the week. We could have equivalently re-coded the classes to be $\mathrm{k} = 1$ for Monday, $\mathrm{k} = 2$ for Tuesday, etc. The observed counts are denoted by O_k for $\mathrm{k} = 1, 2, ...\mathrm{K}$. The expected counts are computed using the total counts $\mathrm{n} = \sum_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{O}_\mathrm{k}$ and the distribution specified in the null hypothesis. The expected count for class k is given by $\mathrm{E}_\mathrm{k} = \mathrm{np}_\mathrm{k}$, $\mathrm{k} = 1, 2, ..., \mathrm{K}$.

The test statistic given below is used to compare the empirical distribution (observed counts) to the theoretical distribution in the null hypothesis (expected counts):

$$\mathrm{U} = \sum_{\mathrm{k}=1}^{\mathrm{K}} \frac{(\mathrm{O}_\mathrm{k} - \mathrm{E}_\mathrm{k})^2}{\mathrm{E}_\mathrm{k}} \sim \chi^2(\mathrm{K} - 1)$$

In other words, U follows a χ^2 distribution with $K-1$ degrees of freedom. In the days-of-the-week absences discussion, we have $K=5$ classes (working days) so the test statistic follows a $\chi^2(4)$ distribution. If we want to test at a α significance level, the rejection region is defined to be values larger than u_c, the value that separates the top five percent of the $\chi^2(K-1)$ distribution from the rest. In other words, the χ^2 goodness-of-fit test is always a right-tailed test. If the observed value of the test statistic u_{obs} is larger than the critical value u_c, $u_{obs} > u_c$, then we reject the null hypothesis at an α level of significance. Otherwise, we would say that the data does not provide evidence for a departure from the proposed distribution.

Example 4.2.1
Suppose in the German population, the distribution of the blood phenotype AB, B, O, and A are 5, 11, 41, and 43 percent respectively. Five-hundred randomly chosen university students report their blood types. The counts are summarized in the table below. Conduct a χ^2 goodness-of-fit test to test the claim that the proportions of blood phenotype differ from the baseline rates of the German population. Use $\alpha = 0.05$ level of significance and the $1-\alpha = 0.95$ quantile value of $\chi^2(3)$, $u_c = u_{3,0.95} = 7.815$.

Blood Phenotype of 500 College Students			
AB	B	O	A
29	51	207	213

Solution
The null hypothesis states that the distribution of the blood types of the college student population is the same as the German population: $H_0: \pi_{AB} = 0.05, \pi_B = 0.11, \pi_O = 0.41, \pi_A = 0.43$. The alternative hypothesis states that the distribution is different. We first compute the expected counts of each of the blood types for these students.

$$\begin{aligned} E_{AB} &= 0.05 \cdot 500 = 25 \\ E_B &= 0.11 \cdot 500 = 55 \\ E_O &= 0.41 \cdot 500 = 205 \\ E_A &= 0.43 \cdot 500 = 215 \end{aligned}$$

Next, we compute the observed value of the test statistic $U \sim \chi^2(3)$

$$u_{obs} = \frac{(29-25)^2}{25} + \frac{(51-55)^2}{55} + \frac{(207-205)^2}{205} + \frac{(213-215)^2}{215} \approx 0.969$$

Since the observed value is not more than the critical value $u_c = 7.815$, we fail to reject the null hypothesis at a 0.05 level of significance. In other words, the data do not provide evidence of a departure from the distribution specified in the null hypothesis.

Contextually, in this problem, we interpret this as: "the data do not provide statistically significant evidence, at the five percent level, that the distribution of blood type from college students differs from the baseline distribution of the German population."

The Chi-Square Test of Independence

A researcher would like to know if there is an association between frequency of church attendance and political affiliation in a predominantly Christian neighborhood the United States. To this end, they collect data on 500 individuals. The data are the answers to two questions: (i) How often do you attend church service per year? (ii) What is your political party affiliation? The answers to the first question are grouped into 4 categories: C_1: less than three times per year, C_2: between 4 and 8 times per year, C_3: between 9 and 12 times per year, and C_4: more than 12 times per year. The answers choices for the second questions are P_1: Republican Party, P_2: Democratic Party, P_3: Independent. The results of their survey are summarized in the table below.

Observed Counts of Political Party Affiliation versus Church Attendance					
		Political party affiliation			
Church attendance		Republican	Democratic	Independent	Total
	Less than 3	40	90	110	240
	Between 4 and 8	60	50	30	140
	Between 9 and 12	30	20	15	65
	More than 12	25	20	10	55
	Total	155	180	165	500

The status quo is to assume that the two variables are independent. This is the null hypothesis, H_0. From probability theory, we know that if two events are independent, then the joint probability is the product of the marginal probabilities. Let's use the observed data to compute the probability that a randomly chosen person is a Demo-

crat and attends church between four and eight times, assuming political party affiliation and church attendance are independent. The marginals are $\mathbb{P}(\text{democrat}) = \mathbb{P}(\mathrm{C}_2) = \frac{180}{500}$ and $\mathbb{P}(\text{b/w 8 and 8}) = \mathbb{P}(\mathrm{P} = \mathrm{P}_2) = \frac{140}{500}$

$$\mathbb{P}(\mathrm{P} = \mathrm{P}_2, \mathrm{C} = \mathrm{C}_2 | \mathrm{P} \text{ indep. } \mathrm{C}) = \mathbb{P}(\mathrm{P} = \mathrm{P}_2) \cdot \mathbb{P}(\mathrm{C} = \mathrm{C}_2) = \frac{180}{500} \cdot \frac{140}{500} = 0\,.1008$$

If we look at the observed (unconditional) probability, we have

$$\mathbb{P}(\mathrm{P} = \mathrm{P}_2, \mathrm{C} = \mathrm{C}_2) = \frac{50}{500} = 0\,.1000$$

Clearly, these two probabilities are difference, but is this different significant? What about the other cells. You can verify that $\mathbb{P}(\mathrm{P}_1, \mathrm{C}_3) = 0\,.0800$ and $\mathbb{P}(\mathrm{P}_1, \mathrm{C}_3 | \mathrm{H}_0) = 0\,.14880$, a greater difference. If we compute analogous probabilities for the other cells in the table, we will see that the joint conditional probability is not the same as the joint (unconditional) probability. We would like to test whether the overall difference between what is observed (unconditional probability) and what is expected (under the condition of independence) is statistically significant. Similar to the χ^2 goodness-of-fit test, we will use counts instead of probabilities.

We will first compute the expected counts, E_{ij}, for all the cells under H_0. For example,

$$\begin{aligned}
\mathrm{E}_{11} &= 500 \cdot \mathbb{P}(\mathrm{C}_1, \mathrm{P}_1 | \mathrm{H}_0) \\
&= 500 \cdot \mathbb{P}(\mathrm{C}_1 | \mathrm{H}_0) \cdot \mathbb{P}(\mathrm{P}_1 | \mathrm{H}_0) \\
&= 500 \cdot \frac{240}{500} \cdot \frac{155}{500} \\
&= \frac{240 \cdot 155}{500} \\
&= 74\,.4
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathrm{E}_{12} &= 500 \cdot \mathbb{P}(\mathrm{C}_1, \mathrm{P}_2 | \mathrm{H}_0) \\
&= 500 \cdot \mathbb{P}(\mathrm{C}_1 | \mathrm{H}_0) \cdot \mathbb{P}(\mathrm{P}_2 | \mathrm{H}_0) \\
&= 500 \cdot \frac{240}{500} \cdot \frac{180}{500} \\
&= \frac{240 \cdot 180}{500} \\
&= 86\,.4
\end{aligned}$$

Let O_{ij} denote the observed count for the cell in the i^{th} row and j^{th} column (with $\mathrm{i} = 1, ..., 4$ and $\mathrm{j} = 1, \cdots 3$). Then, the total number of observations is $\mathrm{n} = \sum_{\mathrm{ij}} \mathrm{O}_{\mathrm{ij}} = 500$. Let $\mathrm{O}_{\mathrm{i}:} = \sum_{\mathrm{j}} \mathrm{O}_{\mathrm{ij}}$ denote the row total of the i^{th} row and $\mathrm{O}_{:\mathrm{j}} = \sum_{\mathrm{i}} \mathrm{O}_{\mathrm{ij}}$ denote the column total of column j. From our calculations above, the expected count, E_{ij} in the i^{th} row and j^{th} column is given by

$$E_{ij} = \frac{O_{i:} \cdot O_{:j}}{n}$$

It may be helpful to remember this formula in words:

$$\text{expected count in cell ij} = \frac{(\text{row i total}) \cdot (\text{column j total})}{\text{grand total}}$$

Using this formula, we can compute the expected counts of all 12 cells. The table below consolidates both the observed counts and expected counts (which appear in parentheses).

Observed and Expected Counts of Political Party Affiliation versus Church Attendance

		Political party affiliation			
Church attendance		Republican	Democratic	Independent	Total
	Less than 3	40 (74.40)	90 (86.40)	110 (79.20)	240
	Between 4 and 8	60 (43.40)	50 (50.40)	30 (46.20)	140
	Between 9 and 12	30 (20.15)	20 (23.40)	15 (21.45)	65
	More than 12	25 (17.05)	20 (19.80)	10 (18.15)	55
	Total	155	180	165	500

As with the χ^2 goodness-of-fit test, we will compare the differences between the observed counts, O_{ij}, and the expected counts, E_{ij}, and aggregate all the differences to form our standardized quantity to serve as our test statistic.

$$U = \sum_{i=1}^{4} \sum_{j=1}^{3} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \sim \chi^2(6)$$

In general, if we have I rows and J columns, the test statistic is

$$U = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \sim \chi^2((I-1)(J-1))$$

In words, this test statistic follows a χ^2 distribution, and the degrees of freedom is the product of the number of row minus one times the number of columns minus one. In our example, we have $I = 4$ rows and $J = 3$, columns so the degrees of freedom is $(4-1)(3-1) = 6$.

As with the goodness-of-fit test, the higher the value of U, the more evidence the data provides against H_0. Therefore, we have a right-tailed test with the rejection region being in the right tail. Using a one percent level of significance, the critical value (from a χ^2 table or using a computer software) gives $u_c = 16.812$. The observed value is computed as given in the equation above and, more concretely,

$$\begin{aligned} u_{obs} &= \frac{(40-74.40)^2}{74.40} + \frac{(90-86.40)^2}{86.40} + \frac{(110-79.20)^2}{79.20} \\ &+ \frac{(60-43.40)^2}{43.40} + \frac{(50-50.40)^2}{50.40} + \frac{(30-46.20)^2}{46.20} \\ &+ \frac{(25-17.05)^2}{17.05} + \frac{(20-19.80)^2}{19.80} + \frac{(10-18.15)^2}{18.15} \\ &\approx 54.683 \end{aligned}$$

Since $u_{obs} > u_c$, we reject the null hypothesis at a one percent level of significance. In other words, the data provide evidence at this significance level that political affiliation and frequency of church attendance are dependent.

The table below contains the commands in various common software programs to compute the critical value of the χ^2 distribution with `df` degrees of freedom and significance level.

Commands for Computing the Critical Value of a Chi-Square Distribution in Various Software Packages	
Software Package	Command
Excel	`CHISQ.INV(1-a,df)`
R	`qchisq(1-a,df)`
Python + SciPy	`scipy.stats.chi2(df).ppf(1-a)`
Matlab/Octave	`icdf('chi2',1-a,df)`

In both of the χ^2 tests we have studied, we should note that the distribution of the test statistic is not exactly the χ^2 distribution. Instead, it is, asymptotically, the χ^2 distribution, i.e., for a large number of observations. The appropriateness of the test may break down if almost all the cells (more than 80 percent) contain expected counts fewer than five. In such cases, other tests are more appropriate.

Kolmogorov-Smirnov Test Of Normality

In this section, we will discuss how to perform goodness-of-fit test specifically for the Gaussian distribution. As usual, the null hypothesis states that the distribution that the data are drawn from follow a Gaussian distribution with known mean, μ, and known standard deviation, σ. That is $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma)$. The alternative hypothesis states that this is not the case. The test statistic used to evaluate the extent to which the distribution of the data differs from the stated normal distribution considers the maximum observed difference between the (theoretical) CDF $\Phi(\cdot)$ and the empirical CDF $F(\cdot)$ evaluated at the random sample:

$$D = \max_x |\Phi(x) - F(x)|$$

At first view, it is not clear where the random sample is used in the test statistic because the dependence on the X_i's are an explicit. Note that the maximum is over all possible values of x, not only on the observed values. However, as with all test statistics, D indeed depends on the random sample via the empirical CDF,F , which is defined by

$$F(x) := \frac{1}{n} \sum_{i=1}^{n} \# \left\{ X_{(i)} \leq x \right\}$$

where $X_{(i)}$ is the i^{th} largest value in the random sample (ordered from least to greatest) and $\# \left\{ X_{(i)} \leq x \right\}$ counts the number of values in the random sample that are, at most, as large as x. Once we have observed values of the random sample $x_1, ..., x_n$, the empirical CDF is $F_{obs}(x)$ where $X_{(i)}$ is replaced with $x_{(i)}$ for each $i = 1, ...n$. Therefore, in practice we will compute the value $D_{obs} = \max_x |\Phi(x) - F_{obs}(x)|$. The table below gives the critical (cut-off) values for various values of α for $n = 10$ (to use with samples of size 10).

Critical Values for the Kolmogorov-Smirnov Test of Normality for Various Significance Levels and Samples of Size Ten

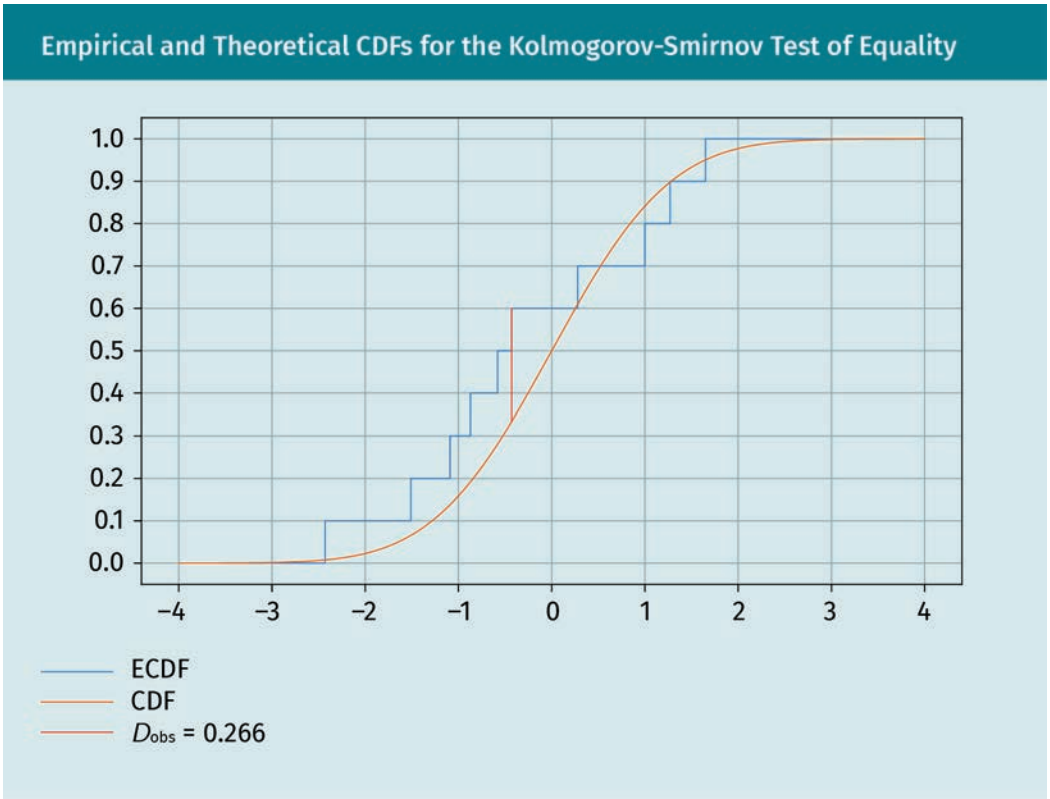
α	D_c
0.001	0.58

Critical Values for the Kolmogorov-Smirnov Test of Normality for Various Significance Levels and Samples of Size Ten	
0.01	0.49
0.02	0.46
0.05	0.41
0.10	0.37
0.15	0.34
0.20	0.32

Although we cannot compute the observed value of the test statistic by hand, to make the ideas behind this test concrete, let's consider a data set of size ten: $\{-2.43, -1.51, -1.09, -0.87, -0.58, -0.43, 0.28, 1.00, 1.27, 1.65\}$. We want to see if this sample came from a standard Gaussian distribution: $\mathcal{N}(0,1)$. The first step is to write down the empirical CDF corresponding to this sample: $\mathrm{F}_{\mathrm{obs}}(\mathrm{x})$. This piece-wise function is given by

$$\mathrm{F}_{\mathrm{obs}}(\mathrm{x}) = \begin{cases} 0, & \mathrm{x} < -2.43 \\ 0.1, & -2.43 \leq \mathrm{x} < -1.51 \\ 0.2, & -1.51 \leq \mathrm{x} < -1.09 \\ \vdots & \vdots \\ 0.9, & 1.27 \leq \mathrm{x} < 1.65 \\ 1.0, & \mathrm{x} \geq 1.65 \end{cases}$$

The observed value of the test statistic is $\max_{\mathrm{x}} |\Phi(\mathrm{x}) - \mathrm{F}_{\mathrm{obs}}(\mathrm{x})|$; graphically, this is the maximum vertical distance between the graphs of Φ and $\mathrm{F}_{\mathrm{obs}}$. The figure below shows a plot of the two graphs and highlights the maximum value as $\mathrm{D}_{\mathrm{obs}} = 0.266$. According to the table, the critical value, the $\alpha = 0.05$, correponds to $\mathrm{D}_{\mathrm{c}} = 0.41$. Since the observed value is not greater than the critical value, we fail to reject the null hypothesis at $\alpha = 0.05$ level of significance. In other words, the data does not provide evidence at this significance level that the data came from a distribution different than the standard normal distribution.



In our discussion above, we used the standard Gaussian distribution. However, the Kolmogorov-Smirnov test works with any arbitrary normal distribution, as long as it is fully specified. In other words, we must pre-determine the mean and standard deviation of the normal distribution we intend to test for, without using the data. It is tempting to use the sample mean and sample standard deviation from the data for these parameter values, but the test is then invalidated. If the parameters of the target distribution are not known, a different test should be applied. The Kolmogorov-Smirnov test can also be used to perform a goodness-of-fit test of continuous distributions other than the Gaussian. We would just choose the appropriate CDF. In practice, many statistical inference and statistical learning algorithms assume that the underlying distribution is Gaussian. It is for this reason that we focus our presentation on the Kolmogorov-Smirnov test for this distribution.

4.3 Two-Sample Tests

In the previous section, we learned how to conduct some statistical tests comparing a parameter of a population of interest against the known respective parameter from a baseline population. In this section, we will learn how to perform statistical tests where we compare parameters from two populations of interest whose true parameters are both unknown. Such statistical tests require two samples, one from each population. Note that in this section, we assume that the two populations are independent of one another.

Comparing Two Proportions: A Z-Test

Note that A/B testing is sometimes used in place of two-sample tests. Either way, we are talking about a two-sample statistical test where each sample is (ideally) identical in every way except in how they are "treated." In such tests, we try to ascertain whether the parameter of interest (proportion or mean) is associated with the type of treatment.

Consider the CTR scenario discussed in the introduction. Let π_A and π_B denote the true CTR for the designs A and B respectively. As in the case of single-sample tests, the null hypothesis makes the statement of "no effect"; the CTRs are the same for both designs or $H_0: \pi_A = \pi_B$. As with the one-sample tests, the alternative hypothesis is a statement about a change in the CTR. It can take one of three forms: (i) design A has a lower CTR than design B, (ii) design A has a higher CTR than design B, (iii) the two designs have different CTRs. These three possible alternative hypotheses can be represented respectively as

1. $H_1: \pi_A < \pi_B$
2. $H_1: \pi_A > \pi_B$
3. $H_1: \pi_A \neq \pi_B$

Similar to one-sample tests, the first two are for one-sided tests and the last one is for a two-sided test. The choice of which of the three alternative hypotheses to use must be made before collecting or viewing any data. In our case, we want to see if the new design, B, yields a higher click-through-rate. Therefore, we will choose $H_1: \pi_A < \pi_B$ as our alternative hypothesis.

Let's clarify these symbols further. For our scenario, we have two independent populations, A and B. Population A represents all the visitors who would be exposed to design A, and population B the visitors who would be exposed to design B. The outcome of each member of the populations can be represented as a random variable $X \sim \text{Bernoulli}(\pi_A)$ and $Y \sim \text{Bernoulli}(\pi_B)$ where $X = 1$ denotes the event that a user from the first population clicked on the target button and similarly for $Y = 1$.

Every statistical test requires data. For A/B testing, it is essential that we acquire data from a randomized experiment. That is, for every visitor who loads the target Web page, we randomly choose (with equal probability) to show them either design A or B. Depending on which page they view, the visitor will represent a sample point from population A or B. As such, we collect two samples, one from each of the populations. Let $X_1, ..., X_m$ denote a random sample from population A and $Y_1, ..., Y_n$ denote a random sample from population B. Note that the two samples are independent of one another.

Once we have the observed values of these samples $x_1, ..., x_m$ and $y_1, ..., y_n$, we can compute the estimates of the CTRs: $\widehat{\pi}_A = \frac{1}{m}\sum_{i=1}^{m} x_i$ and $\widehat{\pi}_B = \frac{1}{n}\sum_{j=1}^{n} y_j$. Theoretically, there are two possible extreme cases that may be observed. On the one hand, we may get $\widehat{\pi}_A - \widehat{\pi}_B = 0$ in which case there would be no reason to reject H_0, and on the other hand, we may get $|\widehat{\pi}_A - \widehat{\pi}_B| = 1$, in which case we could certainly argue for rejecting H_0.

Even if H_0 were true, due to the inherent randomness, it is unlikely that we will see $\widehat{\pi}_A - \widehat{\pi}_B = 0$. Similarly, if H_0 were false, it is unlikely that we observe a specific difference in the sample proportions: $\widehat{\pi}_A - \widehat{\pi}_B = 0.1$. It is more appropriate to compute probabilities of a range of values.

If H_0 were true, how likely is it that we observe $\widehat{\pi}_A - \widehat{\pi}_B \leq d$? We can ask this question using the estimators $\widetilde{\pi}_A = \frac{1}{m}\sum_{i=1}^{m} X_i$ and $\widetilde{\pi}_B = \frac{1}{n}\sum_{j=1}^{n} Y_j$ as follows

$$\mathbb{P}(\widetilde{\pi}_A - \widetilde{\pi}_B \leq d \,|\, \pi_A = \pi_B)$$

To compute such a probability, we need to know the distribution of $\widetilde{\pi}_A - \widetilde{\pi}_B$. It turns out that if the samples are large (m, n are large), the central limit theorem tells us that the standardized difference of sample proportions is approximately normally distributed (conditioned on $H_0: \pi_A = \pi_B$). Symbolically,

$$U = \frac{\widetilde{\pi}_A - \widetilde{\pi}_B}{SE(\widetilde{\pi}_A - \widetilde{\pi}_B)} \overset{\text{approx.}}{\sim} \mathcal{N}(0,1)$$

Recall that $\mathbb{V}[\widetilde{\pi}_A - \widetilde{\pi}_B] = \mathbb{V}[\widetilde{\pi}_A] + \mathbb{V}[\widetilde{\pi}_B] = \frac{\pi_A(1-\pi_A)}{m} + \frac{\pi_B(1-\pi_B)}{n}$. Since we don't know the true proportions, we can approximate the variance by replacing these with their estimators: $\mathbb{V}[\widetilde{\pi}_A - \widetilde{\pi}_B] \approx \frac{\widehat{\pi}_A(1-\widehat{\pi}_A)}{m} + \frac{\widehat{\pi}_B(1-\widehat{\pi}_B)}{n}$. Therefore, we have

$$SE(\widetilde{\pi}_A - \widetilde{\pi}_B) \approx \widehat{SE}(\widetilde{\pi}_A - \widetilde{\pi}_B) = \sqrt{\frac{\widehat{\pi}_A(1-\widehat{\pi}_A)}{m} + \frac{\widehat{\pi}_B(1-\widehat{\pi}_B)}{n}}$$

Now, the probability from $\mathbb{P}(\widetilde{\pi}_A - \widetilde{\pi}_B \leq d \,|\, \pi_A = \pi_B)$ can be written as

$$\mathbb{P}\left(U \leq \frac{d}{\widehat{SE}(\widehat{\pi}_A - \widehat{\pi}_B)}\right) = \mathbb{P}(U \leq u) = \Phi(u)$$

where $\Phi(\,\cdot\,)$ is the CDF of the standard Gaussian distribution.

This brings us to the second part of hypothesis testing: the test statistic (U), and the rejection region. The probability measures how likely we are to observe a data set which corresponds to a difference in proportions of at least d, given that the null hypothesis is true. Intuitively, if the null hypothesis were true (the true proportions are the same), the probability of observing a large (negative) difference in the proportions will be small. Analogously, the probability of observing small differences will be large.

The significance level α is the maximum probability of type I error we want to be subject to. In other words, given that the proportions are the same, we want to reject the null hypothesis (incorrectly) at most $100\alpha\,\%$ of the time. If $\alpha = 0.01$, this means that the probability of making a type I error should be no more than one percent. With this in mind, we decide to reject if the probability is less than $\alpha = 0.01$. The set of values of u for which $\mathbb{P}(U < u) < \alpha$ is called the rejection region. The cutoff value is called the critical value, denoted by u_{cr} which comes from $\mathbb{P}(U = u_{cr}) = \alpha$. In our case, with

$\alpha = 0.01$, u_{cr} is just the quantile of the Gaussian distribution corresponding to one percent, $\phi(u_c) = 0.01$ or $u_c \approx -2.33$. Therefore, the rejection region corresponding to a significance level of $\alpha = 0.01$ is the values in the set $RR = \{u | u < 2.33\}$. If we observed data that corresponds to a test statistic value (u) of less than 2.33, then our decision will be to reject the null hypothesis at the one percent level of significance. This would indicate that it may be the case that the design B would result in a higher CTR than design A.

Example 4.3.1
Continuing from our discussion about comparing CTRs, suppose we observed 567 clicks out of 900 and 650 clicks out of 950 from samples obtained from designs A and B, respectively. Compute the observed test statistic u_{obs} of U and write the decision of the A/B test at the one percent level of significance.

Solution
The sample proportions are given by $\widehat{\pi}_A = \frac{567}{900} = 0.6300$ and $\widehat{\pi}_B = \frac{650}{950}a \approx 0.6842$. The standard error (standard deviation of the difference in sample proportions) is computed by

$$\widehat{SE}(\widehat{\pi}_A - \widehat{\pi}_B) = \sqrt{\frac{\frac{567}{900} \cdot \frac{333}{900}}{900} + \frac{\frac{650}{950} \cdot \frac{300}{950}}{950}} \approx 0.0221$$

Putting these numbers together, we can compute the observed value of the test statistic from the equation for U above as

$$u_{obs} = \frac{0.6300 - 0.6842}{0.0221} \approx -2.4525$$

As mentioned in the discussion, this is a left-tailed test with rejection region $RR = \{u | u < -2.33\}$. Since the u_{obs} falls in the rejection region, we will reject the null hypothesis at a one percent level of significance. Therefore, the data provides evidence in favor of the alternative $H_1 : \pi_A < \pi_B$; design B may have a higher CTR.

Changing designs or making recommendations cannot happen with the result of a statistical test alone. As discussed in the first section of this unit, other considerations, such as the practical significance must be analyzed. Here the observed effect is $\widehat{\pi}_A - \widehat{\pi}_B \approx 0.6300 - 0.6842 = -0.0542$. In other words, the change in CTR favors design B by about a five percent margin. A business analyst, together with the design team, may need to decide whether such an increase in CTR warrants the deployment of additional resources in order to make the change permanent. This consideration is outside the scope of the statistical test we have conducted and is left to the relevant experts.

Comparing Two Means

In section 4.1, we discussed how to test a claim involving the true mean of a population of interest with a baseline value. In this part, we will learn how to test claims about how the true means of two independent populations compare. The null hypothesis, the status quo, states that the two means are equal: $\mathrm{H}_0: \mu_1 = \mu_2$, while the alternative can take different forms:

- two-sided test (two-tailed rejection region): $\mathrm{H}_1: \mu_1 \neq \mu_2$
- one-sided test
- left-tailed rejection region: $\mathrm{H}_1: \quad \mu_1 < \mu_2$
- right-tailed rejection region:$\mathrm{H}_1: \quad \mu_1 > \mu_2$

As before, the alternative hypothesis, as well as the significance level, α, must be specified before analyzing any data. The framework for comparing two means from independent samples is similar to any other hypothesis test. There are four parts:

1. Hypotheses
2. α and choice of test-statistic
3. Observed value of test-statistic and defining the rejection region
4. Decision

To discuss the three most common cases of comparing means from independent populations, the overall framework is the same. The only thing that changes is the test-statistic and its associated distribution. Let μ_1 and μ_2 denote the true but unknown means of the (independent) populations of interest and let $\mathrm{X}_1, \ldots, \mathrm{X}_{\mathrm{n}_1}$ and $\mathrm{Y}_1, \ldots, \mathrm{Y}_{\mathrm{n}_2}$ denote two random samples, one from each population. Note that the sample sizes are denoted by n_1 and n_2, respectively. The observed values of these samples are denoted by $\mathrm{x}_1, \ldots, \mathrm{x}_{\mathrm{n}_1}$ and $\mathrm{y}_1, \ldots, \mathrm{y}_{\mathrm{n}_2}$ respectively. The one assumption shared by each of the three cases we discuss below is that either both n_1 and n_2 are large enough to allow the central limit theorem's approximation of the normal distribution of the sample means $\overline{\mathrm{X}}$ and $\overline{\mathrm{Y}}$, or, the underlying distribution from which both samples are drawn are (at least approximately) Gaussian. If neither of these conditions are satisfied, then the tests may not be valid. The three cases we discuss are based on information about the variances (standard deviations) of the underlying distributions:

1. The variances of the two populations, σ_1^2 and σ_2, are known.
2. The variances of the two populations are unknown but assumed to be equal.
3. The variances of the two populations are unknown and assumed to be unequal.

In the first case, we will use a Z-test, that is, a test based on the Gaussian distribution. In the other two cases, we will use a T-test, that is, a test based on the Students' T distribution.

Known Variances: A Z-Test

The null hypothesis, $H_0{:}\mu_1 = \mu_2$ can also be written as $H_0{:}\mu_1 - \mu_2 = 0$. Written this way, it motivates us to compute the difference of the sample means, $\overline{X} - \overline{Y}$ and compare this difference to zero. Indeed if the difference is far from zero, we will tend to reject the null hypothesis. As before, instead of working with this difference, we would prefer to work with a standardized value. If the samples are drawn from independent Gaussian distributions, then

$$(\overline{X} - \overline{Y}) \sim \mathcal{N}\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right).$$

Under the null hypothesis, we know that $\mu_1 - \mu_2 = 0$, so

$$(\overline{X} - \overline{Y}) \sim \mathcal{N}\left(0, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right).$$

Therefore, the quantity

$$U = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0,1)$$

This random variable is our test statistic. The observed value of this test statistic replaces $\overline{X}$ and $\overline{Y}$ with the observed sample means, $\overline{x}$ and $\overline{y}$, respectively. Finally, the rejection region is developed in exactly the same way as with any other Z-test.

Example 4.3.2

The hourly wages of two companies, A and B, are normal distribution. The variances of hourly wages are $\sigma_1^2 = 20$ and $\sigma_2^2 = 18$ for companies A and B, respectively. A researcher wants to know if the average hourly wage of the two companies are different. To this end, they collect the hourly wages of 20 workers, 10 from each of the two companies. The sample means are $\overline{x}$ =15.23€ and $\overline{y}$ =14.15€, for companies A and B, respectively. Conduct a hypothesis test at a five percent level of significance to address the researcher's interest.

Solution

The distribution of the hourly wages are Gaussian for both companies and the variances are known. Therefore, we can conduct a Z-test.

1. $H_0{:}\mu_1 = \mu_2$, the mean hourly wages of the two companies are the same. $H_1{:}\mu_1 \neq \mu_2$, the mean hourly wages of the two companies are different.
2. $\alpha = 0.05$ and

$$U = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0,1)$$

3. The cutoff values are $u_c = \pm z_{0.025} = \pm 1.96$, so the rejection region is $RR = \{u | u < -1.96 \text{ or } u > 1.96\}$. The observed value of the test statistic is

$$u_{obs} = \frac{15.23 - 14.15}{\sqrt{\frac{20}{10} + \frac{18}{10}}} \approx 0.55$$

4. Since u_{obs} is not in the rejection region, we fail to reject the null hypothesis at a five percent level of significance.

Following our decision statement, we can say that the data do not provide evidence of differing average wages between the two companies (at the given level of significance).

Unknown/Equal Variances: A T-Test

The second case in testing for equality of means arises when the variances of the populations are unknown but assumed to be equal. In this case, we need to adjust the test statistic because we don't know the values of σ_1^2 and σ_2^2. As such, we need to replace these quantities by an estimate of a common sample variance, this quantity is called the pooled variance, S_p^2 , and is computed as

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

where S_1^2 and S_2^2 are the sample variance of the X_i's and Y_j's respectively ($i = 1, ..., n_1$ and $j = 1, ..., n_2$). The associated test statistic follows a T distribution:

$$U = \frac{\overline{X} - \overline{Y}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim T(n_1 + n_2 - 2)$$

The quantity $\nu = n_1 + n_2 - 2$ is the degrees of freedom of the T distribution. The observed value of this test statistic comes from replacing $\overline{X}$, $\overline{Y}$, S_1^2, and S_2^2 with the observed values $\overline{x}$, $\overline{y}$, s_1^2, and s_2^2, respectively.

Example 4.3.3

A researcher wants to see if the average (yearly) salary of male managers at a certain company is larger than the average salary of female managers. They assume that the salaries of male managers and female managers are approximately Gaussian with equal (but unknown) variances. The researcher collects a random sample of 32 managers, 16 males and 16 females. The table below summarizes the data.

Data Summary for Example 4.3.3			
Males	$\overline{x} = 150,000$	$s_1 = 3100$	$n_1 = 16$
Females	$\overline{x} = 125,000$	$s_1 = 2900$	$n_2 = 16$

Test the researcher's claim at a $\alpha = 0.01$ level of significance.

Solution
Since the salaries of the managers of the males and females are normal distribution, and the variances are unknown but assumed equal, we will use a T-test with the pooled variance version of the test statistic. Let μ_1 and μ_2 denote the average salaries of the male managers and female managers, respectively.

1. $H_0: \mu_1 = \mu_2$, the average salary of the male and female managers are the same.
   $H_1: \mu_1 > \mu_2$, the average salary of male managers is more than the average salary of the female managers.
2. $\alpha = 0.01$ and

$$U = \frac{\overline{X} - \overline{Y}}{S_p\sqrt{\frac{1}{16} + \frac{1}{16}}} \sim T(30)$$

3. The cutoff value is $u_c = t_{30,0.01} = 2.457$ and we have a right-tailed rejection region: $RR = \{u \mid u > 2.457\}$. The observed values of the pooled variance and pooled standard deviation are

$$s_p^2 = \frac{15 \cdot 3100^2 + 15 \cdot 2900^2}{30} = 9,010,000$$

$$s_p = \sqrt{9,010,000} \approx 3001.67$$

   Next, the observed value of the test statistic is

$$u_{obs} = \frac{15,000 - 125,000}{3001.67\sqrt{\frac{1}{16} + \frac{1}{16}}} \approx 2.945$$

4. Since the observed value is in the rejection region, we reject the null hypothesis at a one percent level of significance.

Following our decision, we can say that the data provide evidence that the average salary of male managers is higher than the average salary of female managers (at the given level of significance).

Unknown/Unequal Variances: A T-Test

The final case we will discuss arises when the variance of the populations are unknown and cannot be assumed equal. In this case, we can replace σ_1^2 and σ_2^2, from the test statistic of the Z-test (known variances) with S_1^2 and S_2^2 respectively. The distribution of the resulting test statistic is approximately T (Welch, 1947):

$$\mathrm{U} = \frac{\overline{\mathrm{X}} - \overline{\mathrm{Y}}}{\sqrt{\frac{\mathrm{S}_1^2}{\mathrm{n}_1} + \frac{\mathrm{S}_2^2}{\mathrm{n}_2}}} \sim \mathrm{T}\left(\nu\right)$$

where the degrees of freedom, ν, is given by

$$\nu_{\mathrm{W}} = \frac{\left(\frac{\mathrm{S}_1^2}{\mathrm{n}_1} + \frac{\mathrm{S}_2^2}{\mathrm{n}_2}\right)^2}{\frac{\mathrm{S}_1^4}{\mathrm{n}_1^2(\mathrm{n}_1 - 1)} + \frac{\mathrm{S}_2^4}{\mathrm{n}_2^2(\mathrm{n}_2 - 1)}}$$

Although the computation of ν_{W}, the degrees of freedom, is quite involved, almost all common statistical packages have implementations to perform this T-test. The table below shows how to conduct the T-test with two independent samples, `x` and `y`, assuming the variances are unknown and unequal. All commands use a two-tailed alternative hypothesis by default. This test is sometimes called Welch's test after the famous statistician who developed it.

Commands for A T-Test with Unknown and Unequal Variances in Various Software Packages

Software Package	Command
Excel	`=T.TEST(x,y,2,3)`
R	`t.test(x,y)`
Python + SciPy	`scipy.stats.ttest_ind(x,y,equal_var=False)`
Matlab/Octave	`ttest2(x,y,'Vartype','unequal')`

4.4 Power, P-Values, and Confidence Intervals

The Power of a Test

In section 4.1, we discussed that when conducting a hypothesis test, depending on the decision, there are two possible errors that could be made: (i) A type I error where we reject a true null hypothesis and (ii) A type II error where we fail to reject a false null hypothesis. The probabilities of committing such errors are denoted by α and β, respectively. Namely,

$$\alpha = \mathbb{P}(\text{reject } H_0 | H_0 \text{ is true})$$

and

$$\beta = \mathbb{P}(\text{fail to reject } H_0 | H_0 \text{ is false})$$

A closely related quantity is the power of a test. Informally, this quantity is the probability of detecting an effect when there actually is an effect. In other words, the power of a test is the complement of the type II error:

$$\text{power} = 1 - \beta$$

If we can compute one of these, β or power, then the other quantity is just the complement. For general statistical tests, computing these probabilities is very difficult. Consequently, we focus on a one-sample Z-test.

We recall the setting for a one-sample Z-test. $X_1, ..., X_n$ are independent Gaussian $\mathcal{N}(\mu, \sigma)$ with σ known. The probability of a type I error, α, is set and the null hypothesis is $H_0 : \mu = \mu_0$ (for some value of μ_0. Let's consider a right-tailed alternative, $H_1 : \mu > \mu_0$. The associated test statistic is

$$U = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

The critical value is given by $u_c = z_\alpha$ and the rejection region is $RR = \{u | u > u_c\}$. The rejection region can also be specified in terms of the sample mean: $RR = \{\overline{x} | \overline{x} > x_c\}$ with $x_c = \mu_0 + u_c \cdot \frac{\sigma}{\sqrt{n}}$

The probability of a type II error as well as the power of a test are computed for a fixed alternative value $\mu = \mu_1$ with $\mu_1 > \mu_0$:

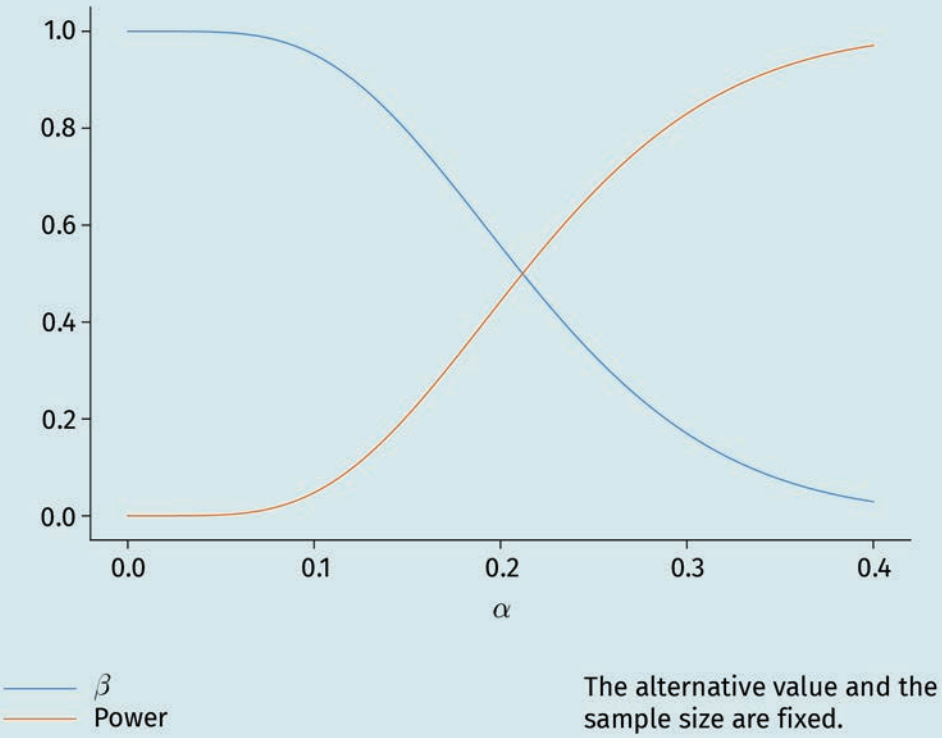
$$\beta(\mu_1) = \mathbb{P}(\overline{X} \notin RR | \mu = \mu_1) = \mathbb{P}(\overline{X} \leq x_c | \mu = \mu_1)$$

We have made the dependence on μ_1 explicit by using the notation $\beta(\mu_1)$ instead of β. Similarly, the power of the test for this alternative is given by

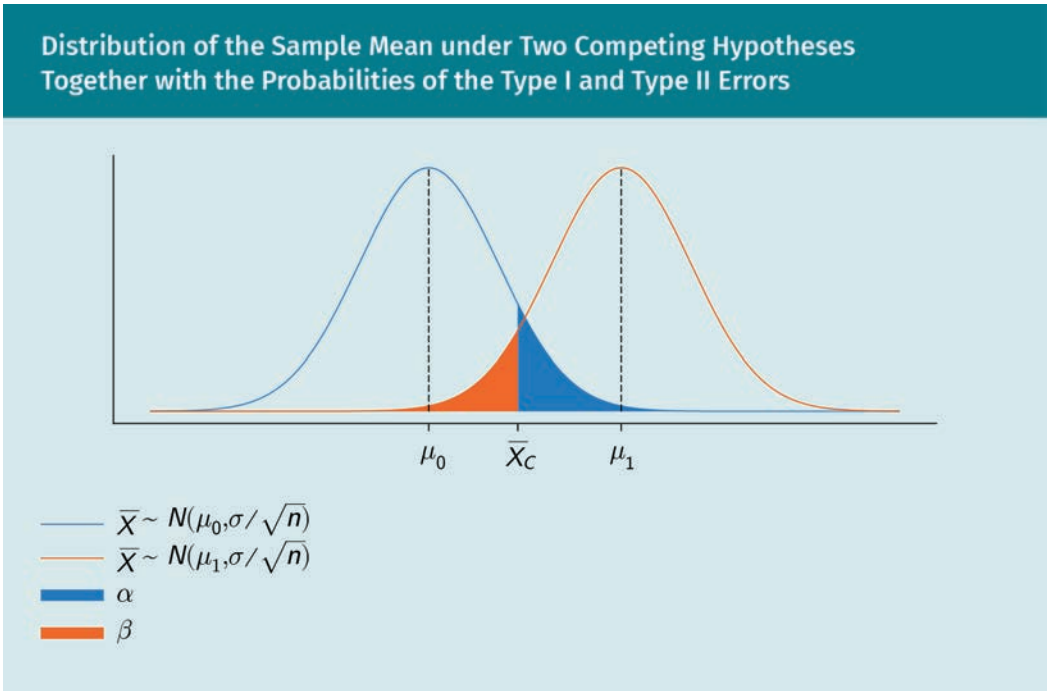
$$\text{power}(\mu_1) = \mathbb{P}(\overline{X} > x_c | \mu = \mu_1)$$

The figure below shows the two distributions of $\overline{X}$, one under $H_0: \mu = \mu_0$ and one under $H_1': \mu = \mu_1$. Additionally, we have highlighted the areas that correspond to the type I and type II errors, α and β, respectively.

Type II Error and the Power of a Test Plotted Against the Probability of a Type I Error



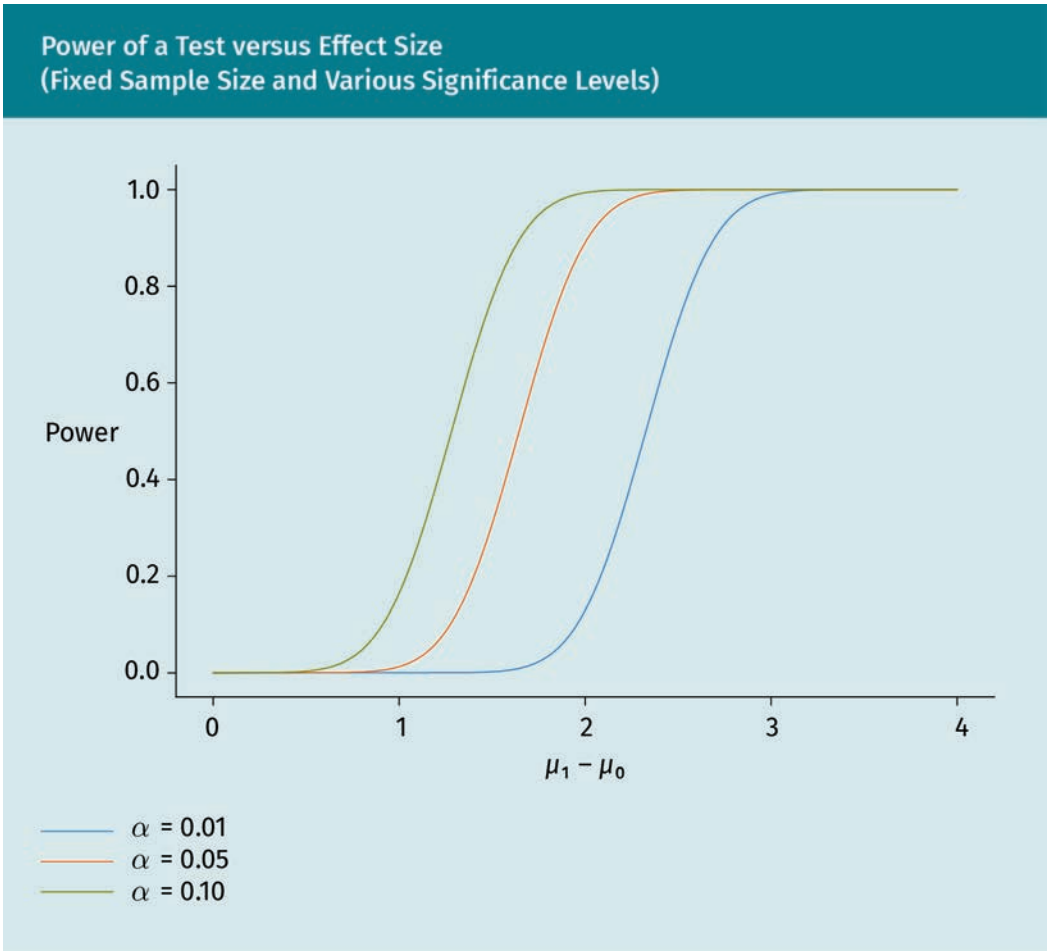
For a fixed value of the alternative, $\mu = \mu_1$ (equivalently, a fixed effect size $\mu_1 - \mu_0$), there is a trade-off between β and α. If we make α small, then the rejection region gets smaller and $\beta(\mu_1) = \mathbb{P}(\overline{X} \notin RR | \mu = \mu_1)$ gets larger. This trade-off is demonstrated in the figure below, which shows the graphs of $\beta(\mu_1)$ and $\text{power}(\mu_1)$ against α, for fixed μ_1 (and fixed sample size, n).



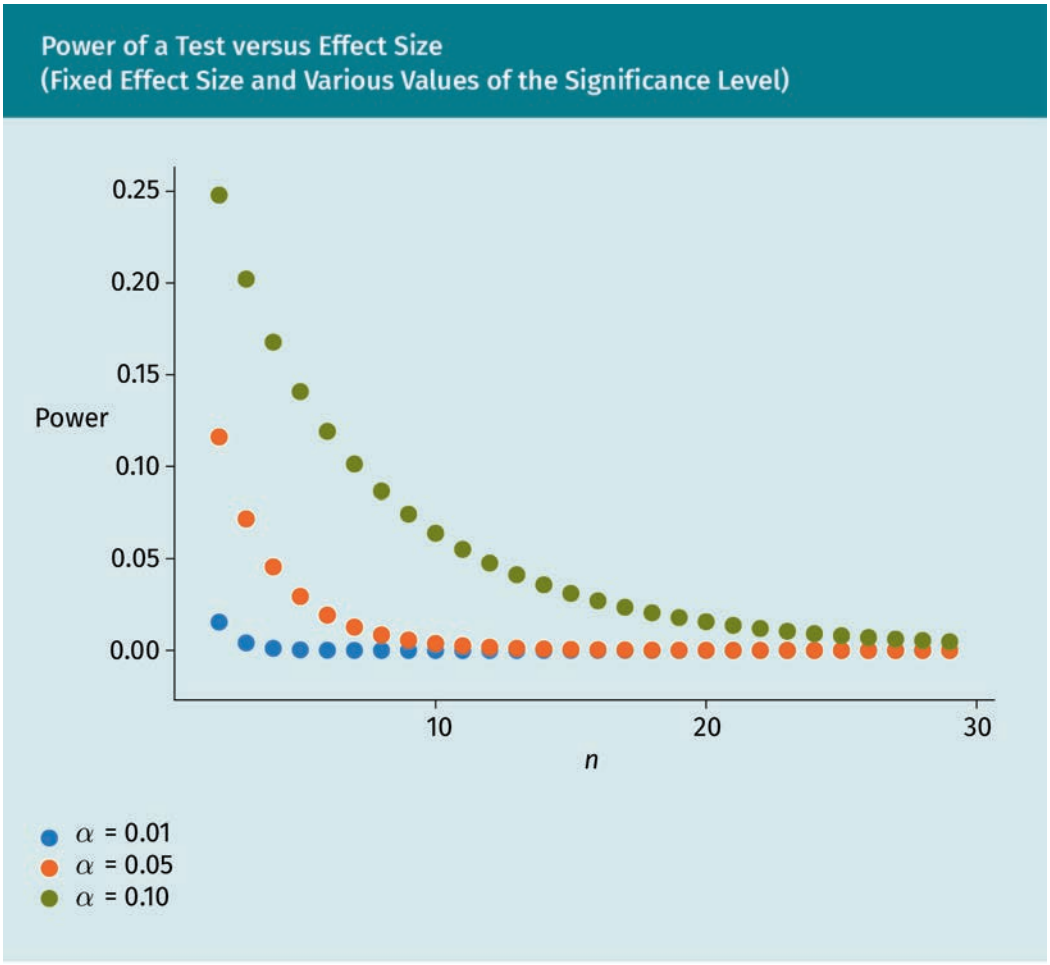
We certainly want the power of a test to be as large as possible. Therefore, it is important to also understand its relationship with the effect size $\mu_1 - \mu_0$ as well as with the sample size n. For a fixed α we have

$$\begin{aligned}
\text{power}(\mu_1) &= \mathbb{P}\left(\overline{X} > \mu_0 + z_\alpha \cdot \frac{\sigma}{\sqrt{n}} \middle| \mu = \mu_1\right) \\
&= \mathbb{P}\left(\frac{\overline{X} - \mu_1}{\frac{\sigma}{\sqrt{n}}} > \frac{\mu_0 - \mu_1 + z_\alpha \cdot \frac{\sigma}{\sqrt{n}}}{\frac{\sigma}{\sqrt{n}}} \middle| \mu = \mu_1\right) \\
&= \mathbb{P}\left(Z > \frac{\mu_0 - \mu_1 + z_\alpha \cdot \frac{\sigma}{\sqrt{n}}}{\frac{\sigma}{\sqrt{n}}}\right) \\
&= \mathbb{P}\left(Z > \frac{\mu_0 - \mu_1}{\frac{\sigma}{\sqrt{n}}} + z_\alpha\right), \quad \text{where } Z \sim \mathcal{N}(0,1)
\end{aligned}$$

For a fixed sample size, n, we can see that for a larger effect size, when μ_1 becomes larger and larger than μ_0, the probability increases. This is because the right-hand side of the inequality describing the event becomes a larger negative value and so the event $[Z > ...]$ is larger. In other words, the larger the effect size (with α and n fixed), the higher the power of the test. The graph below shows the relationship between the power of a test versus effect size for various values of α.



In contrast, if we fix the effect size (i.e., fix μ_1), then for a given α, the larger the sample size, the smaller the power. The following figure shows the relationship between the power of a test and the sample size.



During the planning phase of conducting a statistical test, we have various choices to make. We need to determine the direction of the alternative hypothesis (two-tailed, left-tailed, right-tailed), our tolerance for a type I error, α, and, for a desired effect size, our desired power. As such, we can aim to satisfy our requirements by choosing the correct sample size. To demonstrate how this is done, let's fix $H_0{:}\mu = \mu_0$, and plan to detect an effect size corresponding to $\mu = \mu_1$ for a right tailed test, that is $\mu_1 > \mu_0$. We want to have a type I error of α and type II error of β ($\text{power} = 1 - \beta$). We have

$$\beta = \mathbb{P}\left(Z \leq \frac{\mu_0 - \mu_1 + z_\alpha \cdot \frac{\sigma}{\sqrt{n}}}{\frac{\sigma}{\sqrt{n}}}\right) = \mathbb{P}\left(Z \leq \frac{\mu_0 - \mu_1}{\frac{\sigma}{\sqrt{n}}} + z_\alpha\right) = \mathbb{P}\left(Z \leq -z_\beta\right)$$

where $Z \sim \mathcal{N}(0,1)$. From these equations above, we have

$$z_\alpha + z_\beta = \frac{\mu_1 - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

or equivalently,

$$n = \frac{\left(z_\alpha + z_\beta\right)^2}{\left(\mu_1 - \mu_0\right)^2 \sigma^2}$$

Example 4.4.1
A researcher wants to conduct a hypothesis test with $H_0{:}\mu = \mu_0 = 0$ and $H_1{:}\mu > 0$ for a quantity that is known to follow a Gaussian distribution with known standard deviation $\sigma = 1$. The researcher wants to be able to detect an effect corresponding to $\mu_1 = 0.5$ and their tolerance for the type I error is $\alpha = 0.10$ and desired power is $\text{power} = 0.80$. What is the sample size they should collect?

Solution
Following the formula above, we have $z_\alpha = z_{0.10} = 1.28$. Since $\text{power} = 0.80$, we have $\beta = 1 - 0.80 = 0.20$, and $z_\beta = z_{0.20} = 0.84$. Using $\sigma = 1$ and $\mu_1 = 0.5$, we have

$$n = \frac{(1.28 + 0.84)^2}{(0.5 - 0)^2 (1)^2} \approx 18.14$$

Therefore, with a sample size of $n = 18$, the test will satisfy $\alpha \approx 0.05$ and $\text{power} \approx 0.80$.

P-Values

In practice, the value of α chosen by researchers is somewhat arbitrary. Therefore, two researchers with the same data may have opposing conclusions of the same hypothesis test. It may be that one researcher chooses $\alpha = 0.05$ and another chooses $\alpha = 0.01$. The former may reject the null hypothesis while the latter may fail to reject it. Technically, as discussed above, the choice of α must be made carefully with respect to β/power. Many publications require that researchers report the smallest value of α that would lead to a rejection of the null hypothesis, also known as the p-value. However, as we have mentioned frequently in this unit, since α must be chosen, before looking at the data, the interpretation of the p-value is prone to misinterpretation.

For a test statistic U, the p-value is the smallest value of α for which the observed data suggest that the null hypothesis be rejected. In this setting, the smaller the p-value, the more likely that the null hypothesis is to be rejected. To understand the intricacies involved, let's consider a test with hypotheses $H_0{:}\mu = \mu_0$ and $H_1{:}\mu > \mu_0$. Let U denote the associated test statistic and u_{obs} the associated observed value. The p-value is the probability of observing a value of U that is at least as extreme as the observed value, where the probability is computed under the assumption that the null hypothesis is true. Formally,

$$p\text{-value} = \mathbb{P}(U > u_{obs} | \mu = \mu_0)$$

Suppose the U follows a T distribution with 30 degrees of freedom: $U \sim T(30)$. If $u_{obs} = 2.21$, then

$$\text{p-value} = \mathbb{P}(\text{U} > 2\,.\,21|\mu = \mu_0) \approx 0\,.\,0174$$

In other words, $0\,.\,0174$ or $1\,.\,74\,\%$ is the smallest value of α that would lead to a rejection of the null hypothesis.

If the alternative hypothesis was two-sided, $\text{H}_1{:}\mu \neq \mu_0$, the statement "at least as extreme" means that we have to consider both directions of the extreme values, extreme and above and extreme and below the opposite. In other words, we have

$$\text{p-value} = \mathbb{P}(\text{U} > 2\,.\,21 \text{ or } \text{U} < -2\,.\,21|\mu = \mu_0) \approx 0\,.\,0279$$

Since T is a symmetric distribution we could have also computed this as

$$\text{p-value} = 2\mathbb{P}(\text{U} > 2\,.\,21|\mu = \mu_0) \approx 0\,.\,0279$$

The key here is to understand that u_{obs} comes from a random sample. As such, it is a random quantity and makes the p-value a random variable. In fact the p-value $\sim\text{Uniform}(0, 1)$, i.e. it follows a uniform distribution on the interval $[0, 1]$. Due to this fact, the p-value has very high variation.

A common misconception is that the p-value can be used to assert that the alternative hypothesis is true. Instead, a small p-value casts doubt on the null hypothesis and encourages a researcher to continue in the direction of the research. Another misconception is that a large p-value indicates that the observed effect is due to random chance. Instead, it is important to note that the alternative hypothesis may still be true, even if a large p-value was obtained. This is because a large p-value just means there is no evidence against the null hypothesis, and we should not interpret this as evidence of absence of an effect.

What happens if two studies that are replications of the same test, with different data, report the same small p-value? Does this mean that the null hypothesis is false? Not necessarily, and we cannot make this conclusion alone. We must consider the effect size reported by these researchers. If they are not comparable, then their word doesn't lead to the same conclusion. Additionally, even if the data are not too different, due to randomness, one study may find a significant effect while the replication doesn't. Perhaps the p-values are so close to the threshold that one is just under and the other is just above.

There is both an advantage and disadvantage of reporting p-values. The advantage is that instead of reporting the decision of whether to reject the null, the decision can be left to the reader. The disadvantage is that p-values are frequently misinterpreted in practice. Therefore, decisions based only on p-values are meaningless. Additionally, a small p-value doesn't indicate an effect size that is practically significant.

Probably the single most important quality of any scientific study is its replicability. This quality forms the basis of how much (or how little) the p-value informs our decision to reject or fail to reject the null hypothesis. The reason replicability is important here comes from the right-hand side of the probability statement. The event

$[\mathrm{U} > \mathrm{u_{obs}}|\mathrm{H_0}]$ assumes that it is theoretically possible to repeatedly sample data, from the same population, and with the sample size. If this is the case, the p-value is the long-term proportion of those samples that results in a test statistic that it at least as extreme as the one we observed from our single sample.

The issues that exist in the scientific community with regards to computing and interpreting p-values lie in the assumption of replicability. For example, if a researcher collected a sample and the result p-value was not as small as they desired, they may sample more and more data until the p-value was small enough for the publication (this is known as "p-value hacking"). There are two issues with this: it violates the assumption of replicability with the sample size, since the sample size was not fixed in advance and the power decreases! It is worth noting that a test statistic such as the one given below tends to increase with larger values of n, which in turn increases $\mathrm{u_{obs}}$ decreasing the p-value:

$$\mathrm{U} = \frac{\overline{\mathrm{X}} - \mu_0}{\frac{\sigma}{\sqrt{\mathrm{n}}}}$$

The example below illustrates one of the issues with p-values.

Example 4.4.2
A student answers ten (factual) true/false questions and gets seven correct and three wrong. The null hypothesis states that the student guessed on every question. Compute the p-value under two scenarios:

1. The number of questions, ten, was predetermined and fixed.
2. The student is asked questions repeatedly until they get three wrong.

Solution

1. The p-value is the probability of observing a value as extreme as the one we have observed. In this context, this is the probability of observing at least seven correct answers. The binomial distribution is appropriate here. Let $\mathrm{U}\sim\mathrm{Binomial}(10, 0.5)$. We can use the binomial distribution because (i) the number of questions (trials) is fixed, (ii) the questions (trials) are independent, and (iii) the probability of getting a question right is fixed (50%) from one trial to the next. The p-value is

$$\mathbb{P}(\mathrm{U} \geq 7) = \mathbb{P}(\mathrm{U} = 7) + \mathbb{P}(\mathrm{U} = 8) + \mathbb{P}(\mathrm{U} = 9) + \mathbb{P}(\mathrm{U} = 10) \approx 0.055$$

2. In this scenario, the test statistic U follows a negative binomial distribution $\mathrm{U}\sim\mathrm{Neg\text{-}Binomial}(3, 0.5)$, which models the number of questions until three incorrect questions are observed. The p-value is computed as

$$\mathbb{P}(\mathrm{U} \geq 7) = \mathbb{P}(\mathrm{U} = 7) + \mathbb{P}(\mathrm{U} = 8) + \cdots \approx 0.011$$

If we were to use a threshold of $\alpha = 0.05$, the first scenario doesn't cast much doubt on the null hypothesis, while in the second scenario it does. This is an important consideration since we used the same observed data in both scenarios!

Another issue with very small p-values is that even if all assumptions of a test are satisfied, we must be careful to assess the practical significance of the observed effect size against the statistical significance. We mentioned this for rejection region-based decisions, but it is even more important here. Suppose that a certain procedure extends the life of cancer patients by one week on average. Suppose further that this effect was statistically significant with a very small p-value. Before recommending such a treatment, it would be wise to consider whether it is practically worthwhile. Statistical theory (even very small p-values) is not equipped to make such determinations.

Example 4.4.3
Compute the p-value for example 4.3.3 about the average salary of male managers versus female managers.

Solution
Recall that this example had hypotheses: $H_0: \mu_1 = \mu_2$ and $H_1: \mu_1 > \mu_2$. The test statistic is $U = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{16} + \frac{1}{16}}} \sim T(30)$ and $u_{obs} = 2.945$. The p-value is given by

$$\text{p-value} = \mathbb{P}(U \geq u_{obs}) = \mathbb{P}(U \geq 2.945) = 0.0031$$

Confidence Intervals

Confidence intervals provide a range of values to estimate a parameter of interest. In contrast to a point estimate (a single value), an interval estimate also considers the uncertainty associated with the estimation and is therefore superior to a point estimate. Additionally, as we see below, confidence intervals have the additional strength of helping us make decisions regarding hypothesis tests as an alternative to rejection regions or p-values.

If the point estimator of the parameter of interest follows a symmetric distribution (Gaussian or T distribution), then the associated confidence interval is a random interval of the form

$$\widetilde{\theta} \pm ME$$

where $\widetilde{\theta}$ is the estimator of the parameter of interest θ, and ME is a positive value that determines the width of the confidence interval. The estimator $\widetilde{\theta}$ is based on a random sample and is a random variable. The margin of error (ME) contains information about the uncertainty of this estimator. The probability that a confidence interval contains the true value of the target parameter is called the confidence level (CL). The relationship between CL and the significance level α are related by $CL = 1 - \alpha$ for two-sided hypothesis tests. The confidence level also affects the margin of error. For a given uncertainty, the width of the confidence interval, via ME, is determined by the confi-

dence level. The larger the confidence level, the larger the margin of error, and vice versa. Intuitively, the larger the basket (confidence interval), the more likely it will catch the ball (true value of the target parameter). Similarly, for a given confidence level, the larger the uncertainty, the larger the margin of error.

Suppose that $X_1, ..., X_n$ is an independent sample from a Gaussian distribution $\mathcal{N}(\mu, \sigma)$ with unknown μ, and known σ. We want to find a $(1-\alpha)\,\%$ confidence interval for μ. We use the sample mean as our point estimate. $\widetilde{\theta} = \overline{X}$, which is also Gaussian with the same mean but scaled standard deviation: $\overline{X} \sim \mathcal{N}\left(\mu, \sigma/\sqrt{n}\right)$. We want our confidence interval to contain μ with probability $1-\alpha$. In other words,

$$1 - \alpha = \mathbb{P}(\overline{X} - ME \leq \mu \leq \overline{X} + ME)$$

This is equivalent to

$$\alpha = \mathbb{P}\left(\left|\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}\right| > \frac{ME}{\frac{\sigma}{\sqrt{n}}}\right) = \mathbb{P}\left(|Z| > \frac{ME}{\frac{\sigma}{\sqrt{n}}}\right)$$

or

$$\frac{\alpha}{2} = \mathbb{P}\left(Z > \frac{ME}{\frac{\sigma}{\sqrt{n}}}\right)$$

Therefore, $\frac{ME}{\frac{\sigma}{\sqrt{n}}} = z_{\alpha/2}$. Finally, the margin of error is given by

$$ME = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Therefore, under the conditions on the random sample stated above, the confidence interval is given by

$$\overline{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Once we have the observed values of the random sample, $x_1, ..., x_n$, we can report the observed confidence interval via by just replacing $\overline{X}$ with $\overline{x}$:

$$\overline{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

If we are asked to find a 95 percent confidence interval, we would compute the associated critical z value, $z_{0.05/2} = z_{0.025} = 1.96$, and our confidence interval would be

$$\overline{\mathrm{x}} - 1.96\frac{\sigma}{\sqrt{\mathrm{n}}} \leq \mu \leq \overline{\mathrm{x}} + 1.96\frac{\sigma}{\sqrt{\mathrm{n}}}$$

For a 90 percent confidence interval, we have $\mathrm{z}_{0.01/2} = \mathrm{z}_{0.005} = 2.58$ so that our confidence interval would be

$$\overline{\mathrm{x}} - 2.58\frac{\sigma}{\sqrt{\mathrm{n}}} \leq \mu \leq \overline{\mathrm{x}} + 2.58\frac{\sigma}{\sqrt{\mathrm{n}}}$$

Now suppose that the conditions on the random sample are the same, except that the standard deviation, σ, is unknown. We can replace σ^2 with S^2, the sample variance estimator, or σ with $\mathrm{S} = \sqrt{\mathrm{S}^2}$ in which case the sample mean would follow a T distribution with $\nu = \mathrm{n} - 1$ degrees of freedom: $\overline{\mathrm{X}} \sim \mathrm{T}(\mathrm{n} - 1)$. The rest of the analysis is similar with $\mathrm{z}_{\alpha/2}$ replaced with $\mathrm{t}_{\mathrm{n}-1,\,\alpha/2}$. In this case, the (observed) confidence interval would be

$$\overline{\mathrm{x}} - \mathrm{t}_{\mathrm{n}-1,\,\alpha/2} \cdot \frac{\mathrm{s}}{\sqrt{\mathrm{n}}} \leq \mu \leq \overline{\mathrm{x}} + \mathrm{t}_{\mathrm{n}-1,\,\alpha/2} \cdot \frac{\mathrm{s}}{\sqrt{\mathrm{n}}}$$

For example, given a sample of size $\mathrm{n} = 20$, the 95 percent confidence interval would use the T critical value $\mathrm{t}_{19,\,0.025} = 2.09$ so that the confidence interval is

$$\overline{\mathrm{x}} - 2.09 \cdot \frac{\mathrm{s}}{\sqrt{\mathrm{n}}} \leq \mu \leq \overline{\mathrm{x}} + 2.09 \cdot \frac{\mathrm{s}}{\sqrt{\mathrm{n}}}$$

Example 4.4.4
Take the scenario from example 4.1.3 about the birth weight in a town. The sample of 12 newborns gave a mean birth weight of 3250 grams and a standard deviation of 250 grams. Construct a 99 percent confidence interval for the true mean of newborns. Use $\mathrm{t}_{11,\,0.005} = 3.11$.

Solution
A $(1-\alpha)100\,\%$ confidence interval for the true mean μ, is given by

$$\overline{\mathrm{x}} - \mathrm{t}_{\mathrm{n}-1,\,\alpha/2} \cdot \frac{\mathrm{s}}{\sqrt{\mathrm{n}}} \leq \mu \leq \overline{\mathrm{x}} + \mathrm{t}_{\mathrm{n}-1,\,\alpha/2} \cdot \frac{\mathrm{s}}{\sqrt{\mathrm{n}}}$$

where $\mathrm{n} = 12$, $\alpha = 0.01$, $\mathrm{s} = 250$ and $\overline{\mathrm{x}} = 3250$. Therefore,

$$3250 - 3.11 \cdot \frac{250}{\sqrt{12}} \leq \mu \leq 3250 + 3.11 \cdot \frac{250}{\sqrt{12}}$$

or $3025.56 \leq \mu \leq 3474.44$.

On their own, confidence intervals are very useful to describe the precision of the point estimate, however, they can also be used to make decisions about the null hypothesis: $\mathrm{H}_0 : \theta = \theta_0$. If we are conducting a two-sided test, $\mathrm{H}_1 : \theta \neq \theta_0$, then we can reject the null hypothesis, at a α level of significance, if θ_0 is not contained in the $(1-\alpha)100\,\%$ confidence interval. If we are conducting a left-tailed test, $\mathrm{H}_1 : \theta < \theta_0$, then we reject the null

hypothesis if θ_0 is greater than the upper bound of the confidence interval. Finally, for a right-tailed test, $H_1: \theta > \theta_0$, we reject the null hypothesis if θ_0 is lower than the lower bound of the confidence interval.

Revisiting example 4.1.3, we had $H_0: \mu = \mu_0 = 3480$ and $H_1: \mu < \mu_0 = 3480$. Since this is a left-tailed test, we reject the null hypothesis if $\mu_0 = 3480$ is greater than the upper bound of the confidence interval. Indeed, from the result of example 4.4.1, this is the case as the upper bound of the confidence interval is 3474.44. Since the confidence interval was constructed using a 99 percent confidence level, the decision is made at the $\alpha = 1\,\%$ level.

So far we have discussed confidence intervals for unknown parameters of a single population. We can also construct confidence intervals for the difference of two parameters from independent populations. The derivation of these results are the same for the one-sample case. We summarize the formulas in the following table.

Confidence Intervals for the Difference of Parameters from Two Independent Populations		
Assumptions	Difference	Margin of error (ME)
Large samples	$\pi_1 - \pi_2$	$z_{\alpha/2}\sqrt{\frac{\widehat{\pi}_1(1-\widehat{\pi}_1)}{n_1} + \frac{\widehat{\pi}_2(1-\widehat{\pi}_2)}{n_2}}$
Gaussian (σ_1, σ_2 known)	$\mu_1 - \mu_2$	$z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
Gaussian ($\sigma_1 = \sigma_2$ unknown)	$\mu_1 - \mu_2$	$t_{\nu,\,\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$, $\nu = n_1 + n_2 - 2$
Gaussian ($\sigma_1 \neq \sigma_2$ unknown)	$\mu_1 - \mu_2$	$t_{\nu_W,\,\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

The quantity s_p is the pooled standard deviation we discussed in section 4.3 and ν_W is the degrees of freedom from Welch's test.

Example 4.4.5

Referring to example 4.3.3, find a 99 percent confidence interval for the difference in means, $\mu_1 - \mu_2$, where μ_1 is the average yearly salary of male managers and μ_2 is the same for female managers. Use the confidence interval to make a decision about the hypotheses $H_0: \mu_1 = \mu_2$ and $H_1: \mu_1 > \mu_2$.

Solution
In example 4.3, we assumed that the unknown variances of the two populations were equal. Therefore, the appropriate confidence interval is

$$(\overline{x} - \overline{y}) \pm t_{\nu,\, \alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \_\nu = n_1 + n_2 - 2$$

From the solution of example 4.3.3, we have $s_p \approx 3001.67$. We are given $\overline{x} = 150,000$, and $\overline{y} = 125,000$, so $\overline{x} - \overline{y} = 25,000$. Also, $n_1 = n_2 = 16$ and so $\nu = 16 + 16 - 2 = 30$. Since we want to compute a 99 percent confidence interval, we have $\alpha = 1 - 0.99 = 0.01$ so $\alpha/2 = 0.005$ and $t_{30,\, 0.005} = 2.75$. Putting this all together, we have

$$\text{ME} = 2.75 \cdot 3001.67 \sqrt{\frac{1}{16} + \frac{1}{16}} \approx 2918.44$$

so the confidence interval is

$$22081.56 \leq \mu_1 - \mu_2 \leq 27918.44$$

To inform our decision regarding the hypothesis test, the difference under the null hypothesis $\mu_1 - \mu_2 = 0$ is less than the lower bound and, therefore, we reject the null hypothesis at a one percent level of significance.

4.5 Multiple Testing

Multiple testing is about testing two or more null hypotheses. When we want to test each of the many null hypotheses separately, the probability of committing a type I error is amplified, and we need to find ways to control it. In this section, we first detail why the issue of amplification of the type I error comes. Next, we discuss two error control measures and how to apply them. Let's start with an example.

Suppose that 100 students take a ten-question factual true/false quiz. We want to see if the students guessed on the quiz. In this context, we perform 100 hypothesis tests, one for each student. Let's say that for each hypothesis test we choose a five percent level of significance, that is $\alpha = 0.05$. The null hypotheses, $H_0^{(1)}, H_0^{(2)}, ..., H_0^{(100)}$ each say that the respective student guessed randomly on the quiz. The probability that at least one true null hypothesis is rejected is $1 - 0.95^{100} = 0.994$, which is quite high. This is not very promising! What value of α should we choose for the individual tests so that the probability of rejecting at least one true null hypothesis is small, say 0.05? In our case, with 100 hypotheses, we should put $\alpha = 0.0005$:

$$1 - \left(1 - \alpha\right)^{100} < 0.05 \Rightarrow \left(1 - \alpha\right)^{100} > 0.95 \Rightarrow 1 - \alpha = 0.95^{1/100} \approx 0.9995 \Rightarrow \alpha \leq 0.0005.$$

This type of error is called a family-wise error rate (FWE). The typical method of controlling this type of error is the Bonferroni method, which rejects a specific null hypothesis if its corresponding p-value is less than α/m, where m is the number of hypotheses. In our case, $\alpha/\mathrm{m} = 0.05/100 = 0.0005$. That we obtained this value from our computation above should be surprising because when α is small and m is large, $1-(1-\alpha)^{1/\mathrm{m}} \approx \alpha/\mathrm{m}$

The issue with this type of control is that the multiple testing procedures might result in low power. Recall that the type I error (in this case α/m) and type II error (β) have an inverse relationship. So when α/m is too small, β might be unacceptably large, in which case the power of the test is too low. Therefore, we need an alternative error measure to control.

When we have many hypothesis tests, it makes sense to allow a small proportion of true null hypotheses to be rejected. This way, the power won't be too low. To this end, we discuss a measure that controls the proportion of true null rejected hypotheses. We will need some preliminary quantities:

- FP (false positives) is the number of true null hypotheses that are rejected.
- TP (true positives) is the number of false null hypotheses that are rejected.
- TN (true negatives) is the number of true null hypotheses that are not rejected.
- FN (false negatives) is the number of false null hypotheses that are not rejected.
- R is the total number of rejected null hypotheses: $\mathrm{R} = \mathrm{FP} + \mathrm{TP}$
- m is the total number of null hypotheses.
- $\mathrm{m} - \mathrm{R}$ is the total number of null hypotheses that are not rejected.
- m_0 is the total number of true null hypotheses.
- $\mathrm{m} - \mathrm{m}_0$ is the total number of false null hypotheses.

A two-way table is a good way to remember these quantities.

A Summary of Quantities from the Result of Multiple Hypothesis Tests			
	True H_0	False H_0	Total
Reject H_0	FP	TP	R
Do not Reject H_0	TN	FN	$\mathrm{m} - \mathrm{R}$
Total	m_0	$\mathrm{m} - \mathrm{m}_0$	m

In a (single) hypothesis test, we like to have a small probability of rejecting a true null hypothesis. In other words, we want a small chance of a false positive or a false discovery. If we take this interpretation and apply multiple hypothesis testing, we can require a small false positive rate or false discovery rate (FDR). The FDR is defined as the expected proportion of false positives with respect to all positives:

$$\mathrm{FDR} = \mathbb{E}\left[\frac{\mathrm{FP}}{\mathrm{R}}\right] = \mathbb{E}\left[\frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TP}}\right]$$

The downside of this quantity is that since the random variables FP and TP are not observable, the expectation is impossible to compute. A workaround (for uncorrelated or positive correlated tests) was proposed using marginal p-values, the p-values associated with each null hypothesis. We reorder the hypotheses in increasing order with respect to their p-values:

$$\mathrm{H}_0^{(\mathrm{i}_1)}, \mathrm{H}_0^{(\mathrm{i}_2)}, ..., \mathrm{H}_0^{(\mathrm{i}_\mathrm{m})}$$

where the p-values are in non-decreasing order

$$\mathrm{p}_{\mathrm{i}_1} \leq \mathrm{p}_{\mathrm{i}_2} \leq \cdots \leq \mathrm{p}_{\mathrm{i}_\mathrm{m}}$$

Next, we choose the largest positive integer k, such that

$$\mathrm{p}_{\mathrm{i}_\mathrm{k}} \leq \frac{\mathrm{k}}{\mathrm{m}} \cdot \alpha$$

Then, the hypotheses $\mathrm{H}_0^{(\mathrm{i}_1)}, \mathrm{H}_0^{(\mathrm{i}_2)}, ..., \mathrm{H}_0^{(\mathrm{i}_\mathrm{k})}$ (with the new ordering) are rejected, indicating a statistically significant effect.

Summary

In section 4.1, we discussed the general framework of hypothesis testing as a way to address claims about the population mean or proportion from a single population using an observed sample. Section 4.2 uses the general framework of hypothesis testing and applies it to a non-parametric setting.

In section 4.3, we learned how to apply the framework of hypothesis testing to the cases where we have to compare analogous parameters of two populations. The interpretation of these test as well as their limitations still abide by those of the one-sample tests. It is important to note the underlying assumptions when using these tests. The quality of a test is measured by various quantities, type I errors, type II errors, and the power of a test. The p-value is intimately related to these quantities. It has a place in interpretation but it is often misinterpreted. Therefore, reporting the relevant confidence interval is important upon conducting a hypothesis test. These qualities, the p-value and confidence intervals, are the discussion points of section 4.4.

Finally, in section 4.5, we outlined why we might need to conduct multiple tests and the challenges that come with them. We learned how to address some of these challenges using two different measures of error control. Throughout this unit we

have stressed this takeaway: even if we come to a decision to reject the null hypothesis (which means we have detected an effect), this alone cannot inform policy or recommendation or action. In real-world applications, we need to follow up with replication of the tests, and then determine if the detected effect sizes are of practical importance.

Knowledge Check

Did you understand this unit?

You can check your understanding by completing the questions for this unit on the learning platform.

Good luck!

Unit 5

Statistical Decision Theory

STUDY GOALS

On completion of this unit, you will have learned...

- ... the basic elements of statistical decision theory including loss function, decision function, and risk function.
- ... the definitions of minimax risk, Bayes risk, minimax decision functions, and Bayes decision functions.
- ... the definition of admissibility of a decision function.
- ... about Stein's Paradox together with an illustration using the James-Stein estimator for multiple means.

5. Statistical Decision Theory

Introduction

We are interested in building a spam filter that automatically classifies incoming emails as “spam” or “not spam.” When this filter makes an error, the user incurs a loss (wasted time). If the filter lets a spam email go through, then the user has to spend time deleting the email (one second). If the filter blocks a non-spam email, then the user has to dig through the spam folder to find the “good” email and loses a lot of time (100 seconds). Applied to statistical decision theory, the spam filter is called a decision function which tries to guess the true state (spam or not spam) of a previously unseen (random) observation (an email). In this unit, we will learn how to measure the quality of such decision functions.

In section 1, we introduce the basic elements of statistical decision theory including decision, loss, and risk functions. We will see how to use these elements to frame questions about how well our decision function (estimator) is performing. In section 5.2, we consider some ways of defining an optimal decision function for a target parameter of interest. In particular, we consider the mini-max risk, together with the associated mini-max decision function (estimator). We also define and discuss a Bayesian treatment where we define the Bayes risk and the corresponding Bayes decision function (Bayes estimator), which minimizes this Bayes risk. Finally, in section 5.3, we review an important concept, admissibility, which is a desirably quality of a decision function. Contrary to intuition, when we aim to estimate three or more independent target parameters simultaneously, we will see how the “usual best” estimator for each parameter is not the best option. This rather paradoxical result is called Stein's paradox. Section 5.3 ends with the James-Stein estimator (decision function) for estimating three or more means of independent Gaussian distribution. You may be surprised by the result.

5.1 The Risk Function

Consider the problem of determining whether an email is spam. We aim to do this by making a decision δ based on an observation $\mathbf{x}$. In this case, $\mathbf{x}$ contains information about the email such as the sender, subject line, or text. The state of the email is the true classification, which we may encode in the variable y. We will choose $\mathrm{y} = 0$ to indicate “spam” and $\mathrm{y} = 1$ to indicate “good.” Our task is to come up with a decision function δ which assigns a zero (“spam”) or 1 (“good”) to a given email represented by $\mathbf{x}$. If we think the email is “spam,” then $\delta(\mathbf{x}) = 0$. A correct decision matches the (true) state. This can happen in two cases:

1. The email is spam, and the decision function predicts $\mathrm{y} = \delta(\mathbf{x}) = 0$
2. The email is good, and the decision function predicts $\mathrm{y} = \delta(\mathbf{x}) = 1$

An incorrect decision also has two cases:

1. The email is spam, but the decision function predicts it as good: $0 = \mathrm{y} \neq \delta(\mathbf{x}) = 1$
2. The email is good, but the decision function predicts it as spam: $1 = \mathrm{y} \neq \delta(\mathbf{x}) = 0$

Once we have a decision function, we want to evaluate how well it performs. To this end, we want to assign a "loss" to each pair (y, δ) containing the true state and the decision. If the values in the pair match, we will incur no loss. If the values in the pair don't match, we have an incorrect decision and incur a (positive) loss. A **loss function**, L, takes such a pair and returns a (non-negative) real number. As desired, if $\mathrm{y} = \delta$, $\mathrm{L}(\mathrm{y}, \delta) = 0$, and if $\mathrm{y} \neq \delta$, then $\mathrm{L}(\mathrm{y}, \delta) > 0$. In the email classification example, a common loss function is the zero-one (0-1) loss defined as follows:

Loss function
This function measures the quality of a decision against the true state. It is a non-negative function. Since it is based on a random observation, the loss function is a random variable.

$$\mathrm{L}(\mathrm{y}, \delta) = \begin{cases} 0, & \mathrm{y} = \delta \\ 1, & \mathrm{y} \neq \delta \end{cases}$$

For our example, both the state (y) a well as the decision (δ) can each take on a value of 0 (spam) or 1 (good). The zero-one loss function can be computed for every possible pair:

$$\mathrm{L}(0,0) = \mathrm{L}(1,1) = 0 \text{ and } \mathrm{L}(0,1) = \mathrm{L}(1,0) = 1$$

It is sometimes useful to summarize these values in a loss matrix:

The Zero-One Loss Matrix for Binary Classification

	$\mathrm{y} = 0$	$\mathrm{y} = 1$
$\delta = 0$	0	1
$\delta = 1$	1	0

The loss matrix is the analogous object of a confusion matrix, which is used for classification problems. Each cell in this matrix has the value of the loss incurred by the value of the chosen decision function (rows) against the true state (columns). When these match, the loss is zero. When they don't, the loss is positive.

In this loss function, we are assuming that classifying an email as spam when it is good is equally as bad as classifying an email as good when it is actually spam. This may or may not be desirable; perhaps classifying a good email as spam' may be more costly than the inverse. Perhaps you were expecting an email that goes to the spam folder, and you spend more time looking for it. The inverse is annoying, but you may as well just delete it, and it doesn't take up too much time. To model such an unbalanced (weighted) loss, consider the following loss function which assigns ten times more weight to classifying a good email as spam versus vice versa:

$$\mathrm{L}(\mathrm{y}, \delta) = \begin{cases} 0, & \mathrm{y} = \delta \\ 1, & 0 = \mathrm{y} \neq \delta = 1 \\ 10, & 1 = \mathrm{y} \neq \delta = 0 \end{cases}$$

The corresponding loss matrix is given by

A Weighted Loss Matrix for Binary Classification		
	$\mathrm{y} = 0$	$\mathrm{y} = 1$
$\delta = 0$	0	10
$\delta = 1$	1	0

Recall that in our setup, $\mathbf{x}$ is an email (observation). Since we are building our decision function for any email that may come in, $\mathbf{x}$ is a realization of a random process (variable) $\mathbf{X}$. As such, this randomness makes the decision function δ random as well, albeit, it must have some pattern (otherwise we would just be guessing randomly). This randomness further affects the loss function. Indeed, writing the loss function explicitly, $\mathrm{L}(\mathrm{y}, \delta(\mathbf{X}))$ makes it clear that it is a random variable. As such, we cannot evaluate the loss function, L, for all inputs $\mathbf{X}$. Therefore, we can quantify the "goodness" of a decision function by analyzing the expected value of the loss function: $\mathbb{E}[\mathrm{L}(\mathrm{y}, \delta(\mathbf{X})]$. This expected value is exactly the definition of the **risk function**. In other words, given a loss function, L, a decision function δ, and a random variable $\mathbf{X}$ whose values we will observe, the risk function is given by (Jiao et al., 2013)

Risk function
This is the expected loss for a given loss function and a given decision function.

$$\mathrm{R}(\mathrm{y}; \delta, \mathrm{L}) = \mathbb{E}[\mathrm{L}(\mathrm{y}, \delta(\mathbf{X}))]$$

Finally, since the expectation absorbs the randomness from $\mathbf{X}$, the risk function is a deterministic function of the state, y. If $\mathbf{X}$ has a discrete distribution, then

$$\mathrm{R}(\mathrm{y}; \mathrm{L}, \delta) = \sum_{\mathbf{x}} \mathrm{L}(\mathrm{y}, \delta(\mathbf{x}))\mathbb{P}(\mathbf{X} = \mathrm{x})$$

If $\mathbf{X}$ has a continuous distribution, then

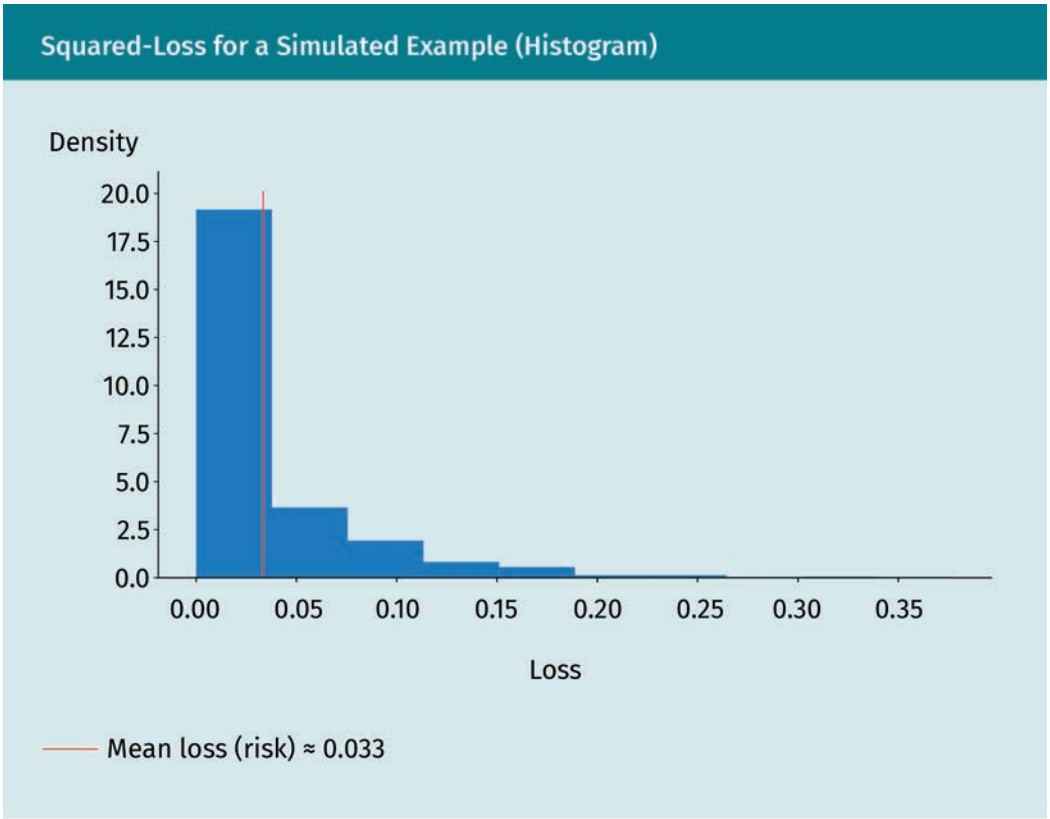
$$\mathrm{R}(\mathrm{y}; \mathrm{L}, \delta) = \int_{\mathbb{R}^{\dim \mathbf{x}}} \mathrm{L}(\mathrm{y}, \delta(\mathbf{x}))\mathrm{f}(\mathbf{x})\mathrm{d}\mathbf{x}$$

where f is the probability density function (PDF) of $\mathbf{X}$. You don't have to worry too much about this (possibly multi-dimensional) integral. We will not concern ourselves with evaluating such quantities in this unit. Instead, our focus will be in understanding how the risk function is defined, quantified, and in the next section, some ways it is opti-

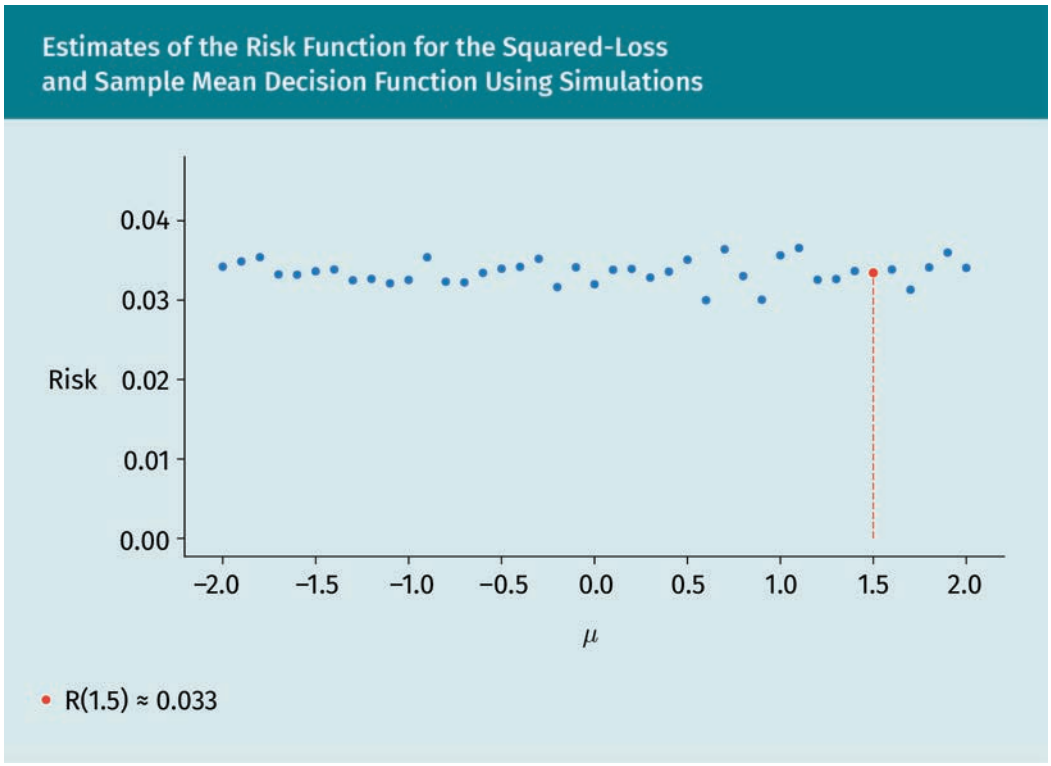
mized. In summary, given a loss function and the distribution of the observations for which we want to predict their state, we quantify the quality of our decision function by the risk function.

We will now consider a more familiar example and explore how the key concepts in this section are applied. Suppose that $\mathbf{X}$ follows a Gaussian distribution $\mathcal{N}(\mu, 1)$ with unknown mean μ and unit standard deviation $\sigma = 1$. We assume that our observation takes the form of n independent observations from this distribution: $\mathbf{x} = (x_1, ..., x_n)$ from $\mathbf{X} = (X_1, ..., X_n)$ iid from $\mathcal{N}(\mu, 1)$. (Recall that iid stands for "independently and identically distributed.") In this case, the true state is the true mean, μ. One obvious way of estimating this mean via a random sample is to use the sample mean. For the observation, we write $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and for the random sample we write $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. In other words, one option for the decision function is the sample mean. $\delta(\mathbf{x}) = \overline{x}$ for a specific observation, and $\delta(\mathbf{X}) = \overline{X}$ for the random sample. A loss function inspired by ordinary least squares is just the square difference between the true mean and the decision: $L\Big(\mu, \delta\Big) = (\mu - \delta)^2$. When we are calculating the loss for a specific observation $\mathbf{x}$, the quantity $L\Big(\mu, \delta\Big) = (\mu - \delta(\mathbf{x}))^2$ is a number. But when we consider the loss of a random sample, the quantity $L\Big(\mu, \delta\Big) = (\mu - \delta(\mathbf{X}))^2$ is a random variable.

The former version is quite straightforward. Suppose we observe $\mathbf{x} = (1, 3, 3, 2, 1)$ (here $n = 5$). Then $\delta\Big(\mathbf{x}\Big) = \overline{x} = \frac{1+3+3+2+1}{5} = 2$. If the true state is $\mu = 1.5$, then the loss is $L(1.5, 2) = (1.5 - 2)^2 = 0.25$. For the random variable L, let's explore the randomness using a simulation. We fix the sample size to $n = 30$ and assume that the true state is $\mu = 1.5$. As such, we generate 1000 random observations (each having 30 numbers) from $\mathcal{N}(1.5, 1)$. Next, for each observation, we compute the decision, δ, which is just the sample mean. At this point, we have 1000 sample means. Next, for each sample mean, we compute the loss just as we did for the (small) sample at the beginning of this paragraph. At this point, we have 1000 loss values. A histogram of these loss values is shown in the figure below. Additionally, we indicate the sample mean loss, an estimate of the value of the risk function $R(1.5, \delta)$, with a red vertical line.



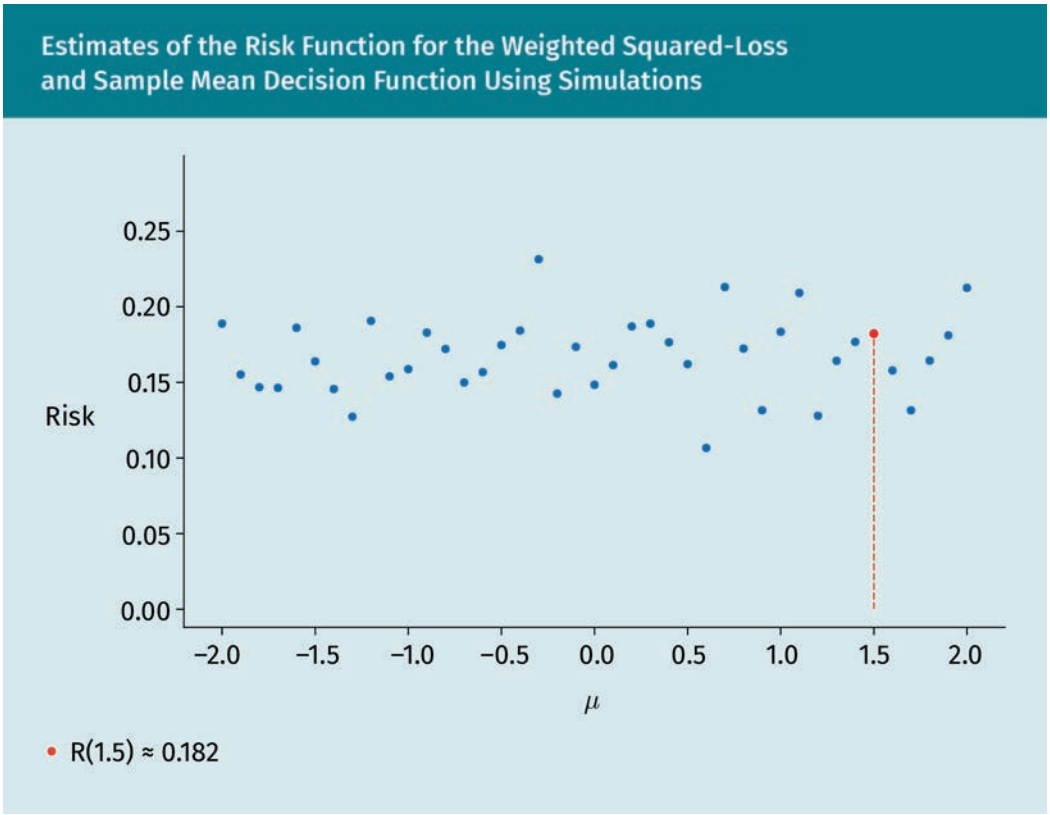
The value 0.033 is an estimate for the risk value at the state $\mu = 1.5$. We can repeat this experiment for many states and plot the risk as a function of the state. The figure below shows the risk function for various values of μ. This is the primary way in which the risk function must be explored, that is, as a function of the state.



When the decision δ doesn't match the the true mean, μ, it is because either our estimate underestimates the mean, $\delta < \mu$ or because it overestimates it, $\delta > \mu$. In certain situations, one of these may be worse than the other. Similar to the weighted loss we discussed for the email classification problem, we can define a loss function which models these two scenarios differently. Suppose that underestimating the mean is 100 times worse than overestimating it by the same amount. The loss function, which encodes this imbalance, can be defined as follows:

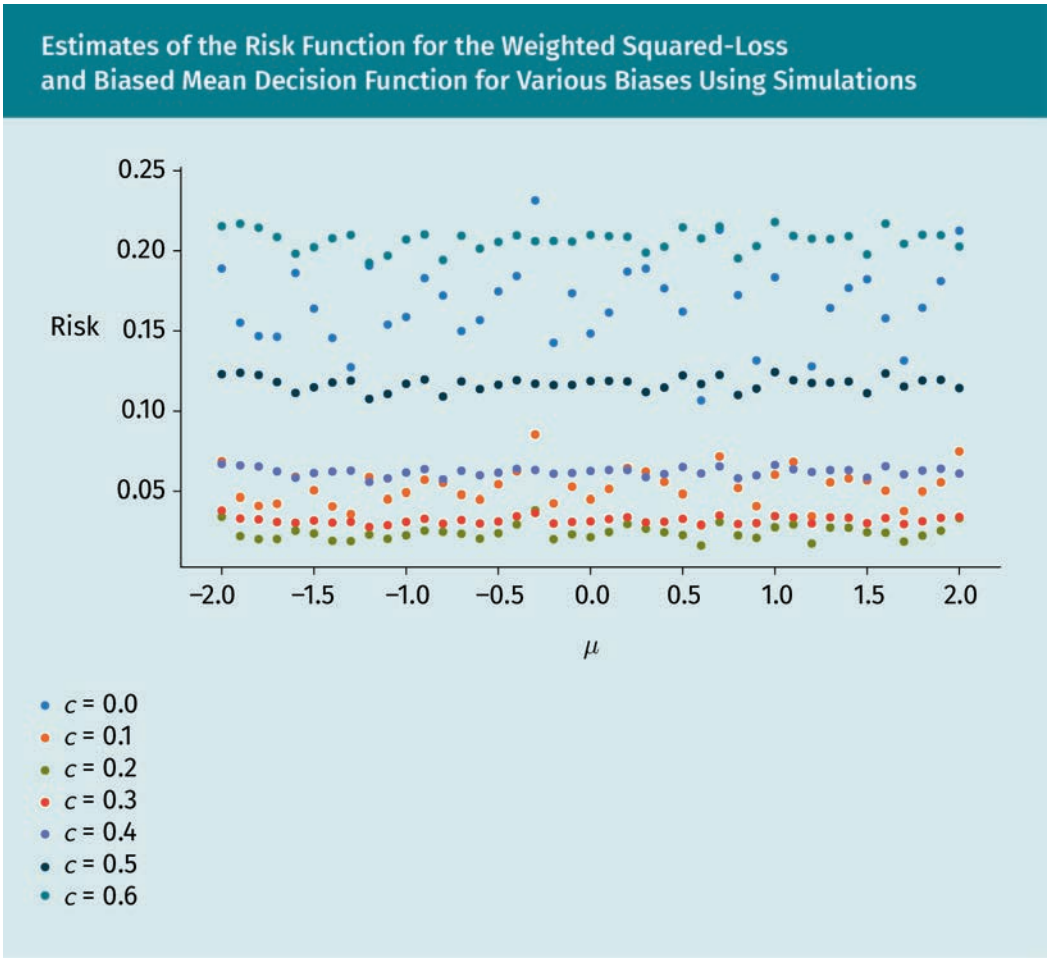
$$L(\mu, \delta) = \begin{cases} 100(\mu - \delta)^2, & \mu \geq \delta \\ (\mu - \delta)^2, & \mu \leq \delta \end{cases}$$

Using this loss function, together with the decision function as before (sample mean), the risk function can be estimated via simulation. We perform this simulation as before and summarize the result in the figure below.



As you can see, the risk function for the weighted squared-loss have much higher values than those of the (ordinary) squared-loss. This suggests that we might want to choose a decision function which overestimates the mean. As such, we might choose $\delta(\mathbf{x}) = \overline{\mathrm{x}} + \mathrm{c}$ for some positive value c. The figure below shows similar simulations for various values of c.

The horizontal axis has the true mean of the Gaussian distribution, and the vertical axis has the risk values (expected losses). Each point in the plot represents an observed sample and its associated risk. The different colors represent different decision functions, parametrized by c.



As you can see, some biased estimators $(c > 0)$ perform better than the unbiased estimator $(c = 0)$. Among the ones we have tried, $c = 0.2$ seems to be the best. Also, too much bias $(c = 0.6)$ actually does worse than the unbiased estimator.

For the mean estimation problem for a Gaussian distribution with unknown standard deviation, we have seen how the choice of the loss function dictates the use of one decision function over another. Before wrapping up this section, let's consider one last loss function. Sometimes, the loss incurred in making a decision that is different than the true state also depends on the magnitude of the true state. Here is an example loss function which models such a case:

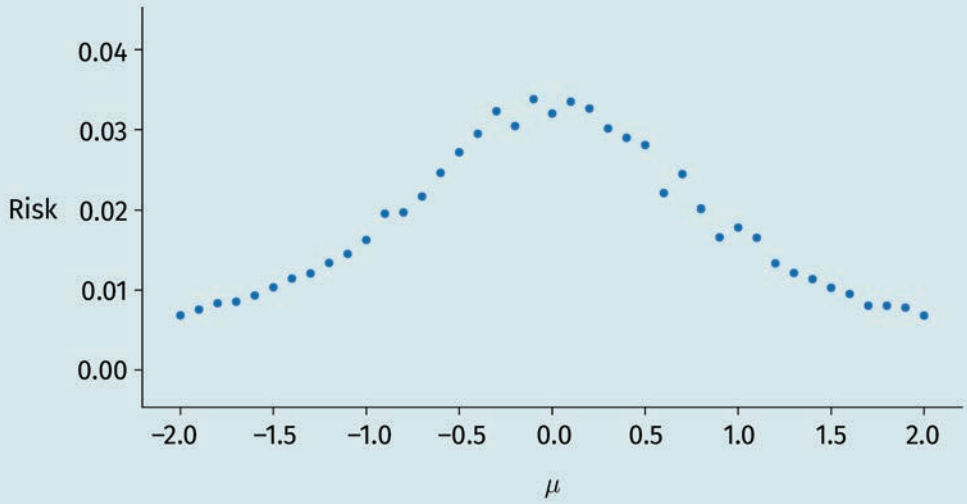
$$L(\mu, \delta) = \frac{(\mu - \delta)^2}{\mu^2 + 1}$$

In such a loss function, we are assuming that the loss incurred for a difference of say $\mu - \delta = 1$ is larger if μ is small (in absolute-value) than if μ is large (in absolute value). Here are some numbers for illustration. Suppose the true mean is $\mu = 1$ and our decision is $\delta = 1.5$, that is, it is off by 0.5. Then, the loss will be

$L(1, 1.5) = \frac{(1-1.5)^2}{1^2+1} = 0.125$. Now suppose the true state is $\mu = 10$, and we are off by 0.5, i.e., our decision is $\delta = 10.5$. Then, the loss will be $L(10, 10.5) = \frac{(10-10.5)^2}{10^2+1} \approx 0.003$. As before, let's explore the risk function using a simulation. The figure below shows the results.

On the horizontal axis, we have the values of the true-state, the true mean of the Gaussian distribution. On the vertical axis, we have the risk values, i.e., the expected loss. Each point in the plot represents one observed sample. As you can see, higher risk is associated with smaller (absolute) values of the mean.

Estimates of the Risk Function for the Squared-Loss Weighted by Magnitude of the Mean and the Sample Mean as the Decision Function Using Simulations

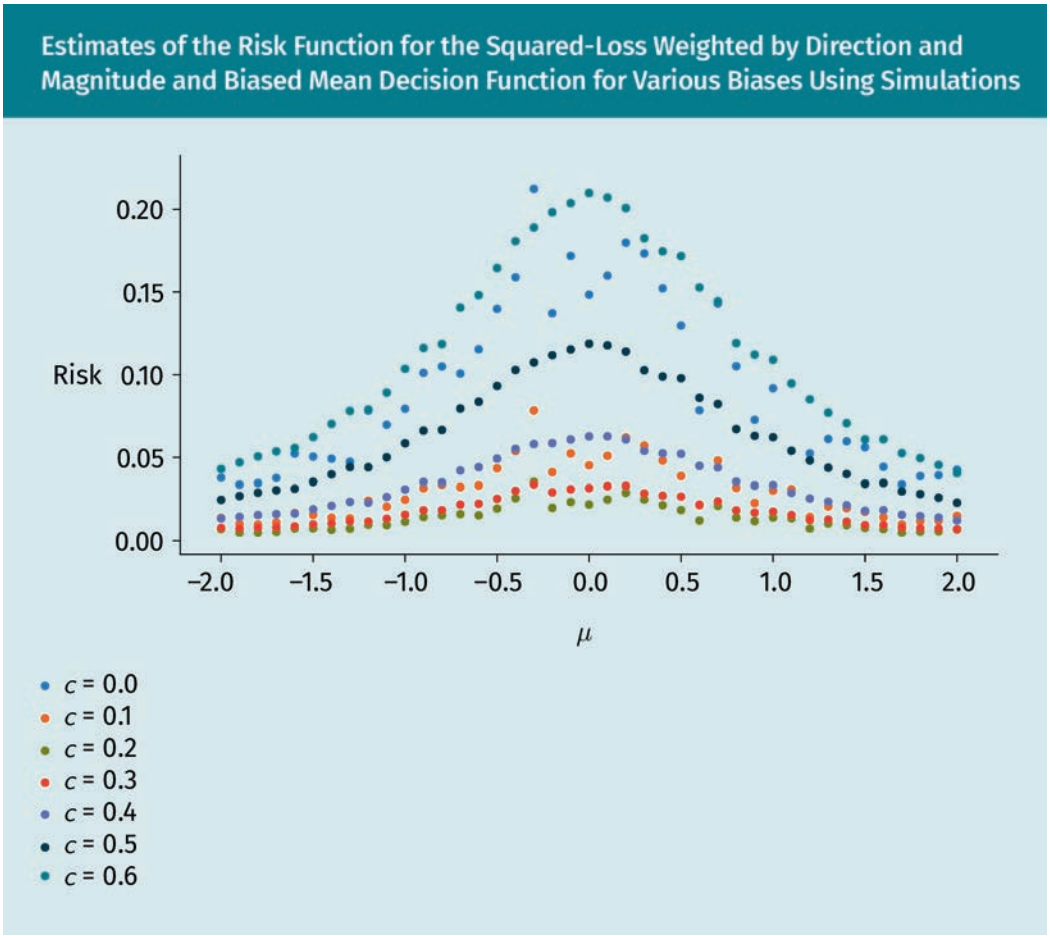


The risk functions prior to this appeared to be more or less constant as functions of the mean. This risk function, on the other hand, is curved. Finally, we can combine the ideas of penalizing underestimates and overestimates differently together with the magnitude of the mean in our loss function:

$$L(\mu, \delta) = \begin{cases} \frac{100(\mu-\delta)^2}{\mu^2+1}, & \delta \leq \mu \\ \frac{(\mu-\delta)^2}{\mu^2+1}, & \delta \geq \mu \end{cases}$$

As before, motivated by the loss function, we prefer to overestimate than underestimate and might consider a decision function which is a (possibly) biased estimate of the mean: $\delta(\mathbf{x}) = \bar{x} + c$, $c \geq 0$. Once again, we will use simulations to explore the risk functions for various biases, c.

This figure is set up as in the previous figure, except that it now contains the estimates of risk (vertical axis) values (expected losses) for various true states (horizontal axis) for different decision functions. The points corresponding to different decision functions are colored differently. As before, each point represents one observed sample.



The core elements of statistical decision theory are as follows:

- the quantity we want to identify, i.e., the state θ among all possible states Θ
- $\mathbf{X}$ a random variable corresponding to the information we will observe and $\mathbf{x}$ the observed value of this random variable
- a set of decision functions, $\delta \in \Delta$ where each decision function δ assigns a decision $\delta(\mathbf{X})$ or $\delta(\mathbf{x})$ which takes on values in the state space Θ
- a loss function L, which evaluates the quality of a decision function (a random variable)
- a risk function, which reduces the loss function to a single number by computing its expected value

Let's consider each of these elements from the last example of this section, i.e., the estimation of the unknown mean of a Gaussian distribution. The state space $\Theta = (-\infty, \infty)$, all real numbers. The true state is some number

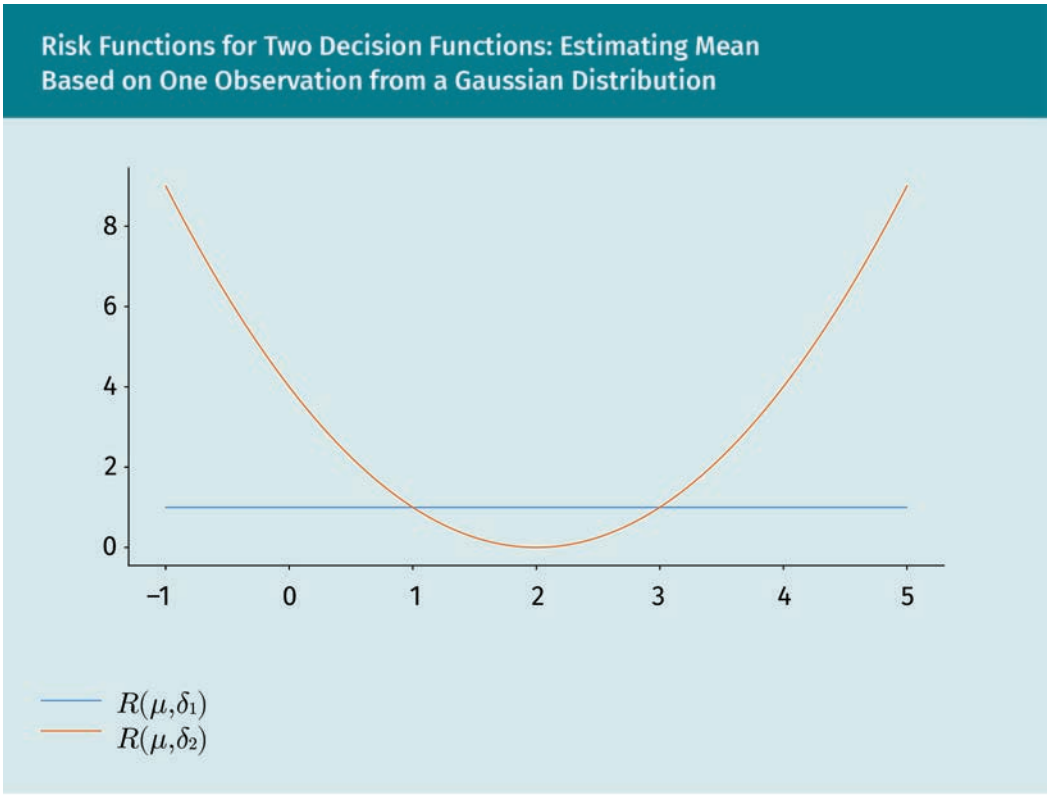
$\theta = \mu \in (-\infty, \infty)$. A random observation $\mathbf{X} = (X_1, X_2, ..., X_n)$ is a random sample of size n, and the observed values were denoted by $\mathbf{x} = (x_1, x_2, ..., x_n)$. Next, we consider the sample mean estimator as a possible decision function: $\delta(\mathbf{X}) = \overline{X}$. An example of a loss function we considered was the the squared-loss: $L\Big(\mu, \delta\Big) = (\mu - \delta(\mathbf{X}))^2 = (\mu - \overline{\mathbf{X}})^2$. Note that this is a random variable. Finally, the risk function, R is the expected loss.

5.2 Maximum Likelihood, Minimax, and Bayes

Maximum likelihood, minimax, and Bayes δ function for a given loss function L. However, as we will see, this process is not straightforward and there are a few different approaches. We will start with some basic examples and give a summary for the general case.

Suppose that we want to estimate the mean of a Gaussian distribution $N(\mu, 1)$. In other words, we have a Gaussian distribution with unknown mean μ and known standard deviation $\sigma = 1$. To simplify the matter, say we only consider a sample with one number, $X \sim \mathcal{N}(\mu, 1)$ with its observed value denoted by x. Also, we will use the standard squared-loss function $L(\mu, \delta) = (\mu - \delta)^2$. Consider the two-decision function $\delta_1(X) = X \quad (\delta_1(x) = x)$ and $\delta_2(X) = 2 \quad (\delta_2(x) = 2)$. The risk functions are $R\Big(\mu, \delta_1\Big) = \mathbb{E}\Big[(\mu - X)^2\Big] = Var\Big[X\Big] = 1$ and $R\Big(\mu, \delta_2\Big) = \mathbb{E}\Big[(\mu - 2)^2\Big] = \Big(\mu - 2\Big)^2$. To compare these two decision functions, we compare their corresponding risk functions. We like to use the decision function that has a lower risk. However, the answer depends on the true value of μ. If $1 < \mu < 3$, then the δ_2 is a better decision function because it has lower risk. But if $\mu < 1$ or $\mu > 3$, then δ_1 is a better decision function because its risk is lower. Finally, if $\mu = 1$ or $\mu = 3$ the two decision functions have the same risk, so we can choose either one. The figure below shows a graph of the two risk functions.

On the horizontal axis, we plot the values of the true state, i.e., the true mean of the Gaussian distribution. On the vertical axis, we plot the risk, the expected loss. For example, when the true mean is zero, the expected loss (risk) for the first decision function, δ_1, is 1 and the expected loss of the second decision function, δ_2, is about 5.



As you can see, neither risk function is entirely below the other one. In other words, neither risk function is uniformly (for all values of μ) lower. To help clarify things, we need a single number for the risk function corresponding to a certain decision function. One option is the **maximum risk** (Kasy, 2014): $\overline{\mathrm{R}}(\mu, \delta) = \max_{\mu} \mathrm{R}(\mu, \delta)$. In our case, $\overline{\mathrm{R}}(\mu, \delta_1) = 1$ and $\overline{\mathrm{R}}(\mu, \delta_2) = \infty$. Therefore, based on this (one-number) summary of the risk function, we will choose δ_1 as the better decision function.

Maximum risk
This is the maximum of the risk function over all possible values of the true state.

Another way to get a single number for the risk function of a decision function is to compute the **Bayes risk**. As you may remember, a Bayesian formulation for estimating an unknown parameter (in this case μ), is to choose a prior distribution for this parameter. Let's say we choose the prior $\mu \sim \mathrm{prior}(\mu)$. Then, the Bayes risk is defined by (Kasy, 2014)

$$\mathrm{R}_{\mathrm{B}}(\delta) = \mathbb{E}_{\mathrm{prior}(\mu)}[\mathrm{R}(\mu, \delta)] = \int_{-\infty}^{\infty} \mathrm{R}(\mu, \delta)\mathrm{prior}(\mu)\mathrm{d}\mu$$

Bayes risk
For this expected risk, the true state is treated as a random variable with a given prior distribution.

If the distribution of the prior is discrete, we would replace the integral with a sum. In other words, the Bayes risk treats the unknown parameter as a random variable, gives it a prior distribution, then computes the expected risk using this prior distribution. For our example above, let's put a standard Gaussian prior: $\mu \sim \mathrm{prior} = \mathcal{N}(0, 1)$. We are now ready to compute the Bayes risk for each of the two estimators:

$$
\begin{aligned}
R_B(\delta_1) &= \mathbb{E}_{\text{prior}(\mu)}[R(\mu, \delta_1)] = \mathbb{E}[1] = 1 \\
R_B(\delta_2) &= \mathbb{E}_{\text{prior}(\mu)}[R(\mu, \delta_2)] = \mathbb{E}\left[(\mu - 2)^2\right] \\
&= \mathbb{E}[\mu^2] - 4\mathbb{E}[\mu] + \mathbb{E}[4] = 1 - 4 \cdot 0 + 4 = 5
\end{aligned}
$$

Now, as you can see, the Bayes risk is lower for δ_1, and we would choose this as our decision function of choice.

Let's consider another example. We want to estimate the (unknown) probability of success from a Bernoulli distribution: $\text{Bernoulli}(p)$. Recall that a variable, X, follows the Bernoulli distribution, if it can take on exactly two values: 0 with probability $1\text{-}p$ and 1 with probability p. We use an observation containing n independent and identically distribution variables: $\mathbf{X} = (X_1, ..., X_n) \overset{\text{i.i.d}}{\sim} \text{Bernoulli}(p)$. The realized values are $\mathbf{x} = (x_1, ..., x_n)$. We will apply the two decision functions given below adapted from an example found in Wasserman (2004).

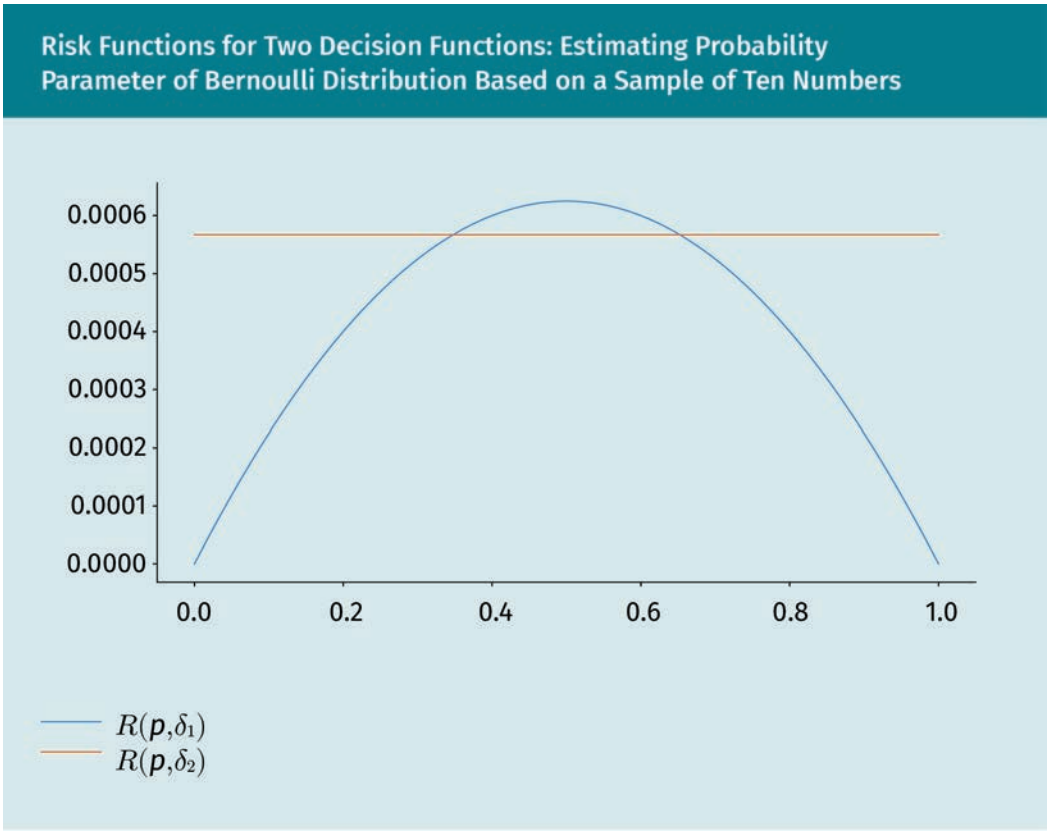
$$
\begin{aligned}
\delta_1(\mathbf{X}) &= \overline{X} = \frac{\sum_i X_i}{n} \\
\delta_2(\mathbf{X}) &= \frac{\sum_i X_i + \sqrt{n/4}}{n + \sqrt{n}}
\end{aligned}
$$

Let's compute the risk functions for each of these decision functions so we can start comparing them:

$$
\begin{aligned}
R(p, \delta_1) &= \mathbb{E}\left[(\overline{X} - p)^2\right] = \text{Var}\overline{X} = \frac{p(1-p)}{n} \\
R(p, \delta_2) &= \mathbb{E}\left[\left(\frac{\sum_i X_i + \sqrt{n/4}}{n + \sqrt{n}} - p\right)^2\right] \\
&= \frac{1}{4\left(1 + \sqrt{n}\right)^2}
\end{aligned}
$$

Let's suppose we are using an observation containing $n{=}400$ numbers. The figure below shows the graph of two risk functions.

The horizontal axis contains the true values of the true state (the probability p of success of a Bernoulli distribution). The vertical axis contains the risk (expected loss) corresponding to each of those true states. For example, when the true state, p, is small or large (away from 0.5), we see that the risk associated with the first decision function, δ_1 is smaller. But, for value of p close to 0.5, the second decision function, δ_2 has lower risk.



Notice that neither one is uniformly below the other. Therefore, considering all the values, we cannot say which one is better. Following the guidance of the previous example, we can calculate a one-number summary of each risk function and choose the one with the lower value. First, let's compute the maximum risk for each of the two decision functions:

$$\overline{\mathrm{R}}(\mathrm{p},\delta_1) = \max_{\mathrm{p}} \frac{\mathrm{p}(1-\mathrm{p})}{\mathrm{n}} = \frac{1}{4\mathrm{n}} = \frac{1}{1600} = 0.000625$$

$$\overline{\mathrm{R}}(\mathrm{p},\delta_2) = \max_{\mathrm{p}} \frac{1}{4\left(1+\sqrt{\mathrm{n}}\right)^2} = \frac{1}{4\left(1+\sqrt{\mathrm{n}}\right)^2} = \frac{1}{4\left(1+\sqrt{10}\right)^2} \approx 0.000567$$

Therefore, the better decision function (the one with a lower maximum risk) is δ_2.

Another way to compute a one-number summary is to calculate the Bayes risk. For this, we need to choose a prior distribution for p. Let's choose $\mathrm{prior}(\mathrm{p}) = \mathrm{Beta}(\mathrm{p};2,2)$, a Beta distribution. Next, treating p as a random variable (following this prior distribution), we compute the expected value of the risk functions under this prior:

$$R_B(\delta_1) = \mathbb{E}\left[\frac{p(1-p)}{n}\right] = \frac{1}{5n} = \frac{1}{2000} = 0.0005$$

$$R_B(\delta_1) = \mathbb{E}\left[\frac{1}{4\left(1+\sqrt{n}\right)^2}\right] = \frac{1}{4\left(1+\sqrt{n}\right)^2} = \frac{1}{4\left(1+\sqrt{400}\right)^2} \approx 0.000567$$

Thus, the first decision function has (marginally) better Bayes risk.

We are finally ready to summarize our observations. In the two examples considered above, we essentially had only two decision functions the were competing. Now suppose that Δ contains all the possible decision functions to choose from. So each decision function δ which is consider is an element of Δ. Next, denote the parameter we are estimating (the true state) as θ. In the first example we had $\theta = \mu$ and in the second example, we have $\theta = p$. Next, consider the collection of all possible values of θ as Θ. In the case of an unknown mean for a Gaussian, $\Theta = (-\infty, \infty)$ and in the case of the unknown probability parameter p for the Bernoulli, $\Theta = [0, 1]$. For a fixed loss function, we know that the risk is based on the true value of θ and the decision function δ:

$$R = R(\theta, \delta)$$

When we have only two risk functions, we found the maximum of the risk function by varying the true parameter. Then, among all the maximum risk values, we chose the smaller one. If we want to do this for all $\delta \in \Delta$, then we would have the **minimax risk** defined by

Minimax risk
For each decision function, we compute the worst (max) risk across all values of the true state. Then, among all these maximum risks, we take the smallest one across the different decision functions.

$$R_{\text{minimax}} = \min_{\delta \in \Delta} \max_{\theta \in \Theta} R(\mu, \delta)$$

You can probably tell why this is called minimax! Once again, in our examples, we computed the maximum risks for each of the decision functions: $\overline{R}(\delta_i) = \max_{\theta \in \Theta} R(\theta, \delta_i)$ for $i = 1, 2$. Then we chose the smaller of the two risks: $\min_{\delta_i \in \Delta} \overline{R}(\delta_i)$ where $\Delta = \{\delta_1, \delta_2\}$. Another way to look at this is the "best (min) worst (max) risk." Finally, the decision function which attains this minimax value is called the minimax decision function (estimator):

$$\delta_{\text{minimax}} = \arg \min_{\delta \in \Delta} \max_{\theta \in \Theta} R(\theta, \delta)\,.$$

In other words,

$$\overline{R}(\delta_{\text{minimax}}) = \max_{\theta \in \Theta} R(\theta, \delta_{\text{minimax}}) = \min_{\delta \in \Delta} \max_{\theta \in \Theta} R(\mu, \delta)$$

The Bayes risk was defined as the expected value of the risk with respect to a prior distribution of the unknown parameter. For our examples, we computed the Bayes risk for each of the two decision functions and then chose the one with the lower Bayes risk. When we have many decision functions in Δ, we essentially do the same thing. We want

$$\mathrm{R}_{\mathrm{Bayes}} = \min_{\delta \in \Theta} \mathrm{R}_{\mathrm{B}}(\delta) = \min_{\delta \in \Delta} \mathbb{E}_{\mathrm{prior}(\theta)}[\mathrm{R}(\theta, \delta)]$$

The decision function that attains this minimum is called the **Bayes decision function** (or Bayes estimator). If we denote this Bayes decision function by δ_{Bayes}, then

Bayes decision function
This decision function minimizes the Bayes risk.

$$\delta_{\mathrm{Bayes}} = \arg\min_{\delta \in \Delta} \mathrm{R}_{\mathrm{B}}(\delta)\,.$$

In other words,

$$\mathrm{R}\big(\delta_{\mathrm{Bayes}}\big) = \min_{\delta \in \Delta} \mathrm{R}_{\mathrm{B}}(\delta)\,.$$

5.3 Admissibility and Stein's Paradox

Given two decision functions (estimators), δ_1 and δ_2, for an unknown parameter, θ, we say that δ_1 is admissible over δ_2 if the risk function associated with δ_1 is uniformly less than (or equal to) the risk function associated with δ_2. This means that $\mathrm{R}(\theta, \delta_1) \leq \mathrm{R}(\theta, \delta_2)$ for every possible value of θ. We revisit the graph of the risk functions of two decision functions from the last section (last figure of section 5.2 above). Notice that neither of the risk functions have their plots entirely below the other. The risk function corresponding to δ_1 is lower than that of δ_2 only for values of $\theta = \mathrm{p}$ far from $\mathrm{p} = 1/2$. Next, consider the various risk functions in the last figure of section 5.1.

It appears that the decision function corresponding to $\mathrm{c} = 0\,.2$ is below every other risk function. Therefore, $\delta_{0\,.2}(\mathbf{X}) = \overline{\mathrm{X}} + 0\,.2$ is said to be **admissible** to every other risk function. Its graph lies entirely below every other risk function. In other words, the risk associated with this decision function is lower than other risk functions for every value of the parameter $\theta = \mu$. One of the nice properties of Bayes estimators is the following fact: given a prior distribution for the unknown parameter θ, if δ is a uniquely determined Bayes estimator, then it is admissible. In other words, a unique Bayes estimator will have its risk function not larger than any other risk function (corresponding to any other decision function) for every value of the parameter θ! We will not indulge in the proof of this fact, but it is an important fact to remember.

Admissible decision function
This decision function's risk function is dominated by every other decision function. Equivalently, its risk (expected loss) is lower than every other decision function for every possible value of the true state.

Suppose we have three unrelated quantities, which are known to follow (independent) Gaussian distributions with known (unit) variance but unknown means. In other words, we have $\mathrm{X}_{\mathrm{i}} \sim \mathcal{N}(\mu_{\mathrm{i}}, 1)$ for $\mathrm{i} = 1, 2, 3$ and $\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3$ are independent. We observe one realization from each Gaussian and want to use this realization to estimate the three

unknown means. The most natural estimate is the value of the observation itself. The natural decision function is $\delta(X_1, X_2, X_3) = (X_1, X_2, X_3)$. In other words, our best guess for the (unknown) mean, μ_1, for the first Gaussian is X_1 and similarly for the other two. To evaluate this decision function, let's use the standard squared loss:

$$L\Big((\mu_1, \mu_2, \mu_3), \delta) = (\mu_1 - X_1)^2 + (\mu_2 - X_2)^2 + (\mu_3 - X_3)^2$$

The associated risk function, the expectation of this loss with respect to the joint distribution of (X_1, X_2, X_3), can be computed term-by-term because of independence:

$$R((\mu_1, \mu_2, \mu_3), \delta) = \mathbb{E}L = \mathrm{Var}[X_1] + \mathrm{Var}[X_2] + \mathrm{Var}[X_3] = 1 + 1 + 1 = 3$$

One would think that this is a pretty good decision function. Now consider the (James-Stein) decision function given by $\delta'(X_1, X_2, X_3) = \left(1 - \frac{1}{S^2}\right)(X_1, X_2, X_3)$, where $S^2 = X_1^2 + X_2^2 + X_3^2$. In other words, this decision function estimates the first mean by $\left(1 - \frac{1}{S^2}\right)X_1$ and the other means similarly. It turns out that the risk associated with this decision function is

$$R((\mu_1, \mu_2, \mu_3), \delta') = 3 - \mathbb{E}\frac{1}{S^2}$$

Note that $\frac{1}{S^2} > 0$ with probability 1. Therefore, $\mathbb{E}\left[\frac{1}{S^2}\right] > 0$ and so $R(\boldsymbol{\mu}, \delta') < 3$. Finally, since this is true for all values of $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)$, we have that the James-Stein decision function (estimate) for the three means is in fact admissible! The result is quite paradoxical for the following reason. To estimate μ_1, our decision function uses the estimator

$$\left(1 - \frac{1}{X_1^2 + X_2^2 + X_3^2}\right)X_1$$

which depends on X_2 and X_3. However, X_1 is independent of X_2 and X_3! In a concrete example, μ_1 is is the average price of tea in China, μ_2 is the average temperature on the surface of Mars, and μ_3 is the average kilograms of food consumed by a polar bear. What Stein's paradox implies is this: if we want to simultaneously estimate these three averages, then we must, for example, use the information about the temperatures on Mars and the consumption of polar bears to inform our estimate about the average price of tea in China! Doing so reduces the risk (the expected squared-loss). The resulting estimate will perform better, on average, then a decision function (estimate) whose estimates for the three means each only depend on their respective values. An intuitive way to think about why this is the case is to consider that although the three quantities have nothing to do with one another, if we end up with a bad estimate for one of the means, we want to make up for it in the other variables. Take a look at the loss function again. For the James-Stein decision function, the associated loss is

$$L((\mu_1, \mu_2, \mu_3), \delta') = \left[\mu_1 - \left(1 - \frac{1}{S^2}\right)X_1\right]^2 + \left[\mu_2 - \left(1 - \frac{1}{S^2}\right)X_2\right]^2 + \left[\mu_3 - \left(1 - \frac{1}{S^2}\right)X_3\right]^2$$

Therefore, each squared term depends on the information from all three variables. In a way, as mentioned above, we simultaneously want a good estimate for all three means.

Summary

This unit serves as general introduction to the core concepts in statistical decision theory. The main elements of statistical decision theory were introduced in section 5.1. The key takeaways from this section wer the relevant definitions of the state space, decision functions, loss functions, and finally the risk function. These are the elements with respect to which problems of statistical decision theory are framed, analyzed, and evaluated.

Taking a risk function and consolidating it to a single number is not a very straightforward task. However, it is an ideal towards which we strive in order to choose the best decision function for a particular problem. To this end, we discussed minimax and Bayes risks and their associated decision functions.

Finally, we discussed the notion of admissibility. This notion is a quality of a decision function that is desirable. The concept of admissibility was discussed with a few examples. Stein's paradox was the final topic of section 5.3. We gave an example of a James-Stein estimator (decision function) that was admissible.

Knowledge Check

Did you understand this unit?

You can check your understanding by completing the questions for this unit on the learning platform.

Good luck!

Congratulations!

You have now completed the course. After you have completed the knowledge tests on the learning platform, please carry out the evaluation for this course. You will then be eligible to complete your final assessment. Good luck!

Appendix 1

List of References

List of References

Downey, A. B. (2016). *Think Bayes*. O'Reilly.

Hogg, R. V., McKean, J., & Craig, A. T. (2019). *Introduction to mathematical statistics: Global edition* (8th ed.). Pearson Education Limited.

Jiao, J., Weissman, T., Miller, J., & Nayebi, A. (2013). Basic concepts of statistical decision theory. *EE378A Statistical Signal Processing*. https://web.stanford.edu/class/ee378a/lecture-notes/lecture2.pdf

Kasy, M. (2014, March 10). Lecture notes on statistical decision theory econ 2110. https://scholar.harvard.edu/files/kasy/files/decisiontheory_0.pdf

Wasserman, L. (2004). *All of statistics*. Springer.

Welch, B. L. (1947). The generalization of 'student's' problem when several different population variances are involved. *Biometrika*, *34*(1/2), 28—35.

Appendix 2

List of Tables and Figures

List of Tables and Figures

List of Tables and Figures

Sample Max of the Bootstrap Samples
Source: Author.

Graphs of the Likelihood, Prior, and Posterior Distributions from Example 3.1.1
Source: Author.

Bayes Estimate Versus Sample Size in Relation to MLE Estimate and Prior Mean
Source: Author.

Bayes Estimate for a Fixed Sample Size Versus Parameters of the Prior in Relation to MLE Estimate and Prior Mean
Source: Author.

Bayes Mean Estimate versus Sample Size for a Gaussian Model in Relation to MLE and Prior Mean
Source: Author.

Conjugate Priors for Some Discrete and Continuous Likelihoods
Source: Author.

Graphs of Different Types of Priors for a Probability Parameter
Source: Author.

Bayes Estimates Versus Sample Sizes Using a Weakly-Informative Prior and a Subjective Prior
Source: Author.

Computing a Density Histogram
Source: Author.

Density Histogram and Kernel Density Estimate
Source: Author.

Gaussian Kernel Density Estimate with Various Window Sizes
Source: Author.

List of Tables and Figures

KNN PDF Estimate for a Standard Gaussian Sample of 1000 Points
Source: Author.

KNN PDF Estimate for a Exponential Sample of 1000 Points
Source: Author.

Type I and Type II Errors and (Probabilities)
Source: Author.

Two-Sided Rejection Regions of a Gaussian Test Statistic
Source: Author.

Two-Sided Rejection Regions of a Gaussian Test Statistic with Various Significance Levels
Source: Author.

Rejection Regions for a Gaussian Test Statistic from Example 4.1.1
Source: Author.

Rejection Regions for a $\mathrm{T}(9)$ Test Statistic from Example 4.1.2
Source: Author.

Common Test Statistics for Parameters of Interest
Source: Author.

Observed and Expected Absences on Different Days of the Week
Source: Author.

PDF of the Chi-Square Distribution with Four Degrees of Freedom and a Right-Rejection Region at a Five Percent Level of Significance
Source: Author.

Blood Phenotype of 500 College Students
Source: Author.

Observed Counts of Political Party Affiliation versus Church Attendance
Source: Author.

List of Tables and Figures

Observed and Expected Counts of Political Party Affiliation versus Church Attendance
Source: Author.

Commands for Computing the Critical Value of a Chi-Square Distribution in Various Software Packages
Source: Author.

Critical Values for the Kolmogorov-Smirnov Test of Normality for Various Significance Levels and Samples of Size Ten
Source: Author.

Empirical and Theoretical CDFs for the Kolmogorov-Smirnov Test of Equality
Source: Author.

Data Summary for Example 4.3.3
Source: Author.

Commands for A T-Test with Unknown and Unequal Variances in Various Software Packages
Source: Author.

Type II Error and the Power of a Test Plotted against the Probability of a Type I Error
Source: Author.

Distribution of the Sample Mean under Two Competing Hypotheses Together with the Probabilities of the Type I and Type II Errors
Source: Author.

Power of a Test versus the Effect Size (Fixed Sample Size and Various Significance Levels)
Source: Author.

Power of a Test versus the Effect Size (Fixed Effect Size and Various Values of the Significance Level)
Source: Author.

Confidence Intervals for the Difference of Parameters from Two Independent Populations
Source: Author.

A Summary of Quantities from the Result of Multiple Hypothesis Tests
Source: Author.

The Zero-One Loss Matrix for Binary Classification
Source: Author.

A Weighted Loss Function for Binary Classification
Source: Author.

Squared-Loss Function for a Simulated Example (Histogram)
Source: Author.

Estimates of the Risk Function for the Squared-Loss and Sample Mean Decision Function Using Simulations
Source: Author.

Estimates of the Risk Function for the Weighted Squared-Loss and Sample Mean Decision Function Using Simulations
Source: Author.

Estimates of the Risk Function for the Weighted Squared-Loss and Biased Mean Decision Function for Various Biases Using Simulations
Source: Author.

Estimates of the Risk Function for the Squared-Loss Weighted by Magnitude of the Mean and the Sample Mean as the Decision Function Using Simulations
Source: Author.

Estimates of the Risk Function for the Squared-Loss Weighted by Direction and Magnitude and Biased Mean Decision Function for Various Biases Using Simulations
Source: Author.

List of Tables and Figures