Course Book

Advanced Statistics

DLMDSAS01

INTERNATIONAL
UNIVERSITY OF
APPLIED SCIENCES

Course Book

ADVANCED STATISTICS

DLMDSAS01

INTERNATIONAL
UNIVERSITY OF
APPLIED SCIENCES

Publisher:

IU Internationale Hochschule GmbH
IU International University of Applied Sciences
Juri-Gagarin-Ring 152
D-99084 Erfurt

Mailing address: Kaiserplatz 1
D-83435 Bad Reichenhall

media@iubh.de
www.iubh.de

DLMDSAS01
Version No.: 001-2021-1026

Contents

Basic Reading

Hogg, R. V., McKean, J., & Craig, A. T. (2005). Introduction to mathematical statistics. Pearson Education.

Illowsky, B., & Dean, S. (2018). Introductory statistics. online available at `https://openstax.org/details/introductory-statistics` (last accessed 2020-Nov-30).

Required Reading

Unit 1

Illowsky, B., & Dean, S. (2018). Introductory statistics. online available at `https://openstax.org/details/introductory-statistics` (last accessed 2020-Nov-30). Chapter 1: Sampling and Data

Hogg, R. V., McKean, J., & Craig, A. T. (2005). Introduction to mathematical statistics. Pearson Education. Chapters: 1.6 (discrete random variables), 1.7 (continuous random variables)

Unit 2

Illowsky, B., & Dean, S. (2018). Introductory statistics. online available at `https://openstax.org/details/introductory-statistics` (last accessed 2020-Nov-30). Chapter 2.3, 2.5-2.8

Unit 3

Hogg, R. V., McKean, J., & Craig, A. T. (2005). Introduction to mathematical statistics. Pearson Education. Chapters: 3.1 (binomial distribution), 3.2 (Poisson distribution, 3.4 (Gaussian distribution)

Unit 4

Hogg, R. V., McKean, J., & Craig, A. T. (2005). Introduction to mathematical statistics. Pearson Education. Chapters: 11.1.1 (prior and posterior)

MacKay, D. J., & Mac Kay, D. J. (2003). Information theory, inference and learning algorithms. Cambridge university press. Chapter 35.1, Benford's Law

Unit 5

Bruce, P., Bruce, A., & Gedeck, P. (2020). Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python. O'Reilly Media. Chapter 1: Exploring the Data Distribution, Exploring Two or More Variables

Unit 6

Hogg, R. V., McKean, J., & Craig, A. T. (2005). Introduction to mathematical statistics. Pearson Education. Chapter: 6.1: Maximum Likelihood Estimation.

Unit 7

Hogg, R. V., McKean, J., & Craig, A. T. (2005). Introduction to mathematical statistics. Pearson Education. Chapter: 4.5: Introduction to Hypothesis Testing

Bruce, P., Bruce, A., & Gedeck, P. (2020). Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python. O'Reilly Media. Chapter 3: Statistical Significance and P-Values

Further Reading

Unit 1

Bishop, C. M. (2006). Pattern recognition and machine learning. Springer. Chapter 1.4: The Curse of Dimensionality

Hogg, R. V., McKean, J., & Craig, A. T. (2005). Introduction to mathematical statistics. Pearson Education. Chapters: 7.2, 7.3 (sufficient statistics)

Unit 2

—

Unit 3

Hogg, R. V., McKean, J., & Craig, A. T. (2005). Introduction to mathematical statistics. Pearson Education. Chapter 3.5, 3.5-3.7

Unit 4

Liu, Y., & Abeyratne, A. I. (2019). Practical Applications of Bayesian Reliability. John Wiley & Sons. Appendix D (Jeffrey's Prior)

Unit 5

Bruce, P., Bruce, A., & Gedeck, P. (2020). Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python. O'Reilly Media. Chapter 1: Exploring Binary or Categorical Variables

Unit 6

Deisenroth, M. P., Faisal, A. A., & Ong, C. S. (2020). Mathematics for machine learning. Cambridge University Press. online available at `https://mml-book.com` (last accessed 2020-Nov-30). Chapter 11.1 - 11.3.

Hogg, R. V., McKean, J., & Craig, A. T. (2005). Introduction to mathematical statistics. Pearson Education. Chapters: 6.6: The EM Algorithm

Bishop, C. M. (2006). Pattern recognition and machine learning. Springer. Chapter 3.1.1: Maximum likelihood and least squares

Unit 7

Illowsky, B., & Dean, S. (2018). Introductory statistics. online available at `https://openstax.org/details/introductory-statistics` (last accessed 2020-Nov-30). Chapter 9, 10

Benjamini, Y., & Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics.Journal of educational and Behavioral Statistics,25(1), 60–83.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency.Annals of statistics, 1165–1188.

1 Introduction to Statistics

Study Goals

After completing this unit you will have learned:

- The fundamental aspects of statistics.
- What the Kolmogorov Axioms are.
- What the problems of dimensionality are.
- What the principal component and linear discriminant analysis are.

Introduction

"Statistics" can be understood in several different ways:

- Statistics as the science of collecting, presenting, analyzing and interpreting of facts and data.
- Statistics as the plural of "a statistic". A statistic is a single measure of a sample and will be explained later in more detail.

Statistics as a science is commonly subdivided into two main branches:

- Descriptive Statistics. This comprises presenting and analyzing of observed data. Descriptive statistics strictly limits itself to collected data from a specific group of observed units (such as people). It is not assumed that data come from a larger population, nor is any inference made about non-observed data.

- Inferential Statistics. Using statistical methods, data from a subset of a population (sample) is used to estimate corresponding data of a larger population. For example, using the arithmetic mean of a sample, the arithmetic mean of the whole population is estimated. Estimates take the form of points or intervals.

The theory of probability is used to bridge these two branches.

Why study statistics? Since our world is not inherently deterministic we have to resort to statistical methods to describe observations. We also have to consider the discussion of a large collection of data. Even if we are able to describe the behavior of a single unit in fine details, statistical methods are required to do so for the whole population. Note that the term "population" is used not only for people, but also in a wider sense to include any measurable objects. Moving on to the term "measurement": this includes "observation" in general, such as observing the color of objects, the quality of room service in a hotel, learning outcomes, and so forth. To allow the measurement of different types of objects, different levels of measurement are distinguished. The four commonly used ones are:

- Nominal scale. For values like country of residence or type of tree, labels are assigned to denote the possible values of objects. It is obvious that labels cannot be used for calculations. It is meaningless to try to calculate an average of the labels "France," "Germany," "Italy." This is also true if we assign numerical values to the labels, like "1" to "France", etc. We can, however, say that two people differ (or do not) in hair color. It is not meaningful to sort nominally scaled objects by value: We cannot say that black is of higher or lower value than brown.

- Ordinal scale. Some categories of objects can be arranged in an order. These are often rankings: For example, unsatisfactory, neutral, satisfactory. As with the nominal scale, values are typically denoted as labels, although numerical values can be used as well. An inherent problem with values on the ordinal scale is that the difference between two values is not well-defined. Although we may hope that the difference between "unsatisfactory" and "neutral" is the same as between "neutral" and "satisfactory", this is by no means guaranteed to be the case. Therefore, even if we use numerical values instead of labels, calculation of differences and ratios is not appropriate.

- Interval scale. On this scale, the distances between individual units are equal. The length of 1 cm is the same, regardless where on the ruler it is taken. The difference between 50 kgs and 55 kgs is the same as the difference between 65 kgs and 70 kgs. The value point 0 (zero) is simply one of the possible points on the scale, as both negative and positive values are possible. This means that it is meaningless to state that yesterday's temperature of $+5^\circ$ Celsius was twice as high as today's -5° Celsius.

- Ratio scale. The scale is different to the interval scale in only one, but significant, aspect: The point 0 (zero) on the scale is the lowest possible one. Typically measurement scales are of this type, e.g., length, weight. It is therefore permissible to state that that one person's weight is one third of someone else's weight. Their ratio is 1:3 in respect of weight.

1.1 Probability theory

Very generally, probability theory is concerned with events whose outcomes are uncertain. This is in contrast with deterministic events, where models or sets of equations predict the outcome of an event. Probability theory assists us in dealing with uncertainty, such as found in random experiments. Axioms provided by probability theory help us to quantify the outcome of uncertain (random) events. It is useful to first clarify some commonly used terms in probability theory.

Random experiment

An experiment for which the outcome cannot be predicted with certainty is called a random experiment. Obvious examples are a coin toss or a dice roll. However, we also speak of a random experiment in the case of the life span of an electric bulb or of a car engine. In both cases we assume that the life spans are non-deterministic and random. We intuitively understand that in each case more than one outcome is possible and that each electric bulb and each car engine is likely to have a different life span.

Sample Space, Event, Outcome

These basic terms are commonly used in the discussion of probability theory and are therefore briefly introduced here. They will be referred to throughout this unit, using practical examples for better understanding.

The set of all possible outcomes of a random experiment are referred to as sample space, denoted $\mathcal{S}$. A set is a defined collection of distinct elements. For example, $A = \{1, 2, 3\}$ is the set of the numbers 1, 2, and 3; or the set of all possible road-trips with a particular car. The order of the elements within a set does not matter—we could, for example, also have written $A = \{2, 1, 3\}$ or $A = \{3, 2, 1\}$. A probability measure P assigns probabilities to events, with each event containing zero or more outcomes. An outcome is the result of a random experiment. Individual outcomes are often grouped to more complex events. An impossible event (such as rolling a 7 with one die) has the probability of 0 (zero); an absolutely certain event (such as rolling any of the six numbers with one die) has the probability of 1.

Random Variable

A random variable is a function that assigns a unique numerical value to the outcome of a random experiment.

A variable whose value depends on the outcome of a random experiment is called a **random variable**. As with other variables, a random variable can be discrete or continuous. As the set of values (all possible outcomes) of a coin toss is well-defined, such random experiments will result in a discrete random variable. If, on the other hand, the outcome of a random experiment can cover a broad set of real values, the associated random variable would be of the continuous type.

The following table presents examples of both types of random variables

Example	Type of Random Variable
Count of broken eggs in each carton of eggs.	Discrete
Quantity/measure of ozone in samples of air.	Continuous
Length of time a customer spends in a store.	Continuous
Count of gas pumps in use.	Discrete
Annual sales/revenue for a company.	Continuous
A company's number of products sold.	Discrete

Expectation Value

For discrete random variables the expectation value (also: expected value) is the probability-weighted mean of all its possible values. This is easily illustrated by rolling a die. Since there are six equally-weighted possibilities in a single roll of a die, each possible outcome will have the probability of 1/6. The expectation value is therefore:

$$1 \cdot 1/6 + 2 \cdot 1/6 + 3 \cdot 1/6 + 4 \cdot 1/6 + 5 \cdot 1/6 + 6 \cdot 1/6 = 21/6 = 3.5$$

The same concept applies to continuous random variables, except that an integral of the variable with respect to its probability replaces the sum.

Union, Intersection, Complement

A union B defines elements in A alone and B alone and elements in the intersection of both A and B. A intersect B consists of elements common to both A and B.

For any two events A and B, the **union** $A \cup B$ consists of all outcomes that are either in A or in B. Therefore the event $A \cup B$ is realized if either A or B occurs.

For any two events A and B, the **intersection** $A \cap B$ consists of all outcomes that are both in A and B. Therefore the event $A \cap B$ is realized if both A and B occur.

The events A and B are said to be mutually exclusive if $A \cap B = \varnothing$. Therefore the events A and B cannot both occur at the same time.

For any event A, the event $\bar{A}$ is defined. $\bar{A}$ refers to the complement of A and consists of all outcomes in the sample space $\mathcal{S}$ that are not in A. Often, the complement is also written as A^c in the literature. Therefore the event $\bar{A}$ is realized if A does not occur. Note that $A \cap \bar{A}$ is $\varnothing$ and $A \cup \bar{A}$ is $\mathcal{S}$. A summary of the notations is shown in Tab. 1.1.

Venn Diagrams

Venn diagrams are useful to illustrate the concepts of union, intersection, and complement of sets. They were introduced by John Venn in the 19th century and graphically visualize relationships among finite sets. In prob-

Notation	Description
$x \in A$	Element x is contained in the event A.
$S \subset A$	All elements of sample space $\mathcal{S}$ are contained in event A.
$S \subseteq A$	All elements of sample space $\mathcal{S}$ are contained in subset$A$, or $\mathcal{S}$ equal to, the event A.
$V = A \cup B$	Event V contains all elements of the union of event A and event B.
$V = A \cap B$	Event V contains all elements in the intersection of the event A and event B.
$\|A\|$	Count of elements in event A.
$A \cap B = \varnothing$	Event A and event B are mutually exclusive events (disjoint events).
$\bar{A}$	Event $\bar{A}$ is the complement of event A, such that $A \cap A^c = \varnothing$.

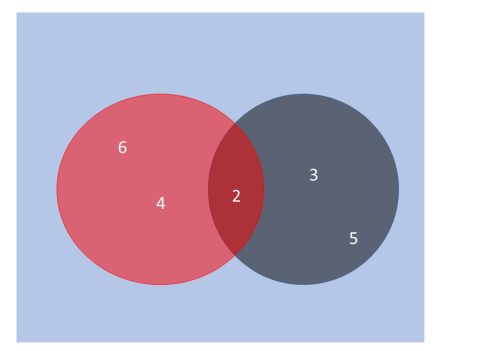
Table 1.1: Summary of the notation for union, intersection and complement.

ability theory, events are subsets of the sample space; we can therefore also use Venn diagrams to visualize properties and operations on events.

The basic types are illustrated in Fig. 1.1, where event A and B intersect (part a), are mutually inclusive (part b), or A is fully contained in B (part c).

The event A and its complement A^c is illustrated using Venn diagrams in Fig. 1.2.

Example



In the random experiment "Roll of a fair die", the possible outcomes

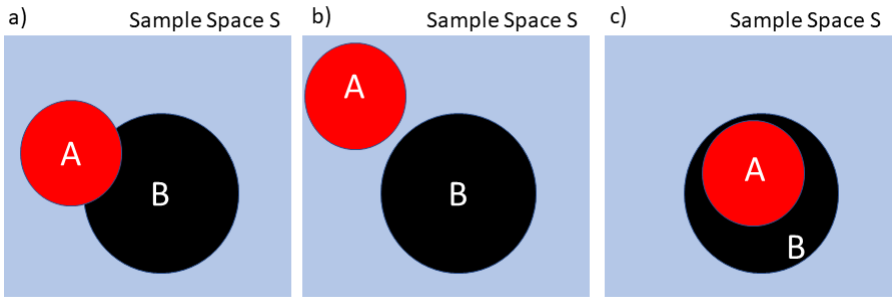


Figure 1.1: : Three basic types of Venn diagrams. a) intersecting events A and B, b) mutually exclusive events A and B, c) A is fully contained in B.

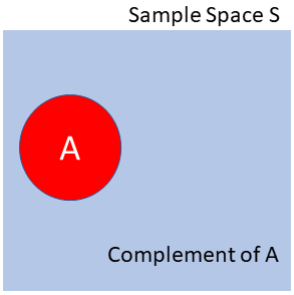


Figure 1.2: : Illustration of event A and complement $\bar{A}$.

of the event "Even numbers" are $\{2, 4, 6\}$, while the outcomes of the event "Prime numbers" are $\{2, 3, 5\}$. The intersection of both events consists of the outcome $\{2\}$. The union of both events corresponds to all possible outcomes of the sample space $\mathcal{S}$. The Venn diagram below demonstrates the relationships among these events.

Self-Check Questions

1. Using the notation C = continuous and D = discrete, indicate whether each of the random variables are discrete or continuous:

   a) The number of sentences in a short story ....

   b) The average oven temperature during the cooking of bread ...

   c) The number of lightning strikes during a thunderstorm ...

   d) The atmospheric pressure at midnight ....

2. A random variable is continuous if the set of possible values includes an entire interval on the number line.

3. Complete the possible outcomes of the sample space $\mathcal{S}$ for the following random experiments:

   a) Flip a coin once. $\mathcal{S} = \{H, \ldots\}$

   b) Flip a coin twice. $\mathcal{S} = \{HH, \ldots\}$

   c) Roll a die once. $\mathcal{S} = \{1, 2, \ldots\}$

Solutions

1. D, C, D, C.

2. True

3. a) $\mathcal{S} = \{H, T\}$

   b) $\mathcal{S} = \{HH, HT, TH, TT\}$

   c) $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$

1.2 Kolmogorov Axioms

The probability axioms introduced by Andrey Kolmogorov are central to probability theory. They have since become known as the Kolmogorov Axioms. In his original work Kolmogorov (1956), lists five axioms, which are commonly combined into three.

Kolmogorov axioms

1. Positivity. The probability P of an event E is a non-negative real number: $P(E) \geqslant 0, P(E) \in \mathbb{R}$.
2. Normalization. The probability that at least one event of the sample space $\mathcal{S}$ occurs is 1: $P(\mathcal{S}) = 1$.
3. Additivity. If two events A and B are mutually exclusive, then the probability of either A or B occurring is the sum of the probabilities of A and B: $P(A + B) = P(A) + (B)$. This also applies for a sequence of mutually exclusive events, such that $A_i \cap A_j = \varnothing \; \forall i, j$, then $P(A_1 \cup A_2 \cup \cdots \cup A_j \cup \cdots) = P(A_1) + P(A_2) + \cdots + P(A_j) + \cdots$

(Kolmogorov, 1956)

The first two axioms are intuitive: Like percentage (or relative frequency) of absolute frequency, probabilities can neither be negative nor greater than one. "Mutually exclusive" in the third axiom means that events A and B have no element in common.

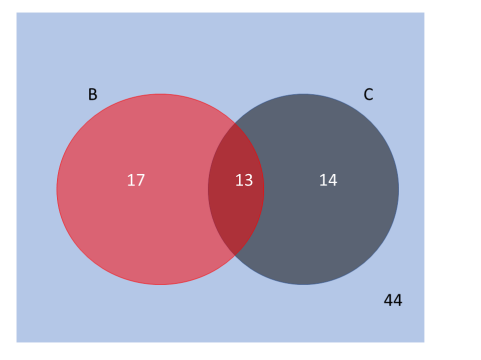
We can understand this a bit better if we imagine rolling a die. Looking at Fig. 1.1, part b) of the figure depicts two events A and B that are independent of each other. Since they are the only ones in the sample

space, their combined probabilities must result in one. This would be the case of rolling one die, with event A specifying the outcome of 1 and 5 and event B with 2, 3, 4 and 6: $P(A) = 2/6 + P(B) = 4/6 = 6/6 = 1$. Events A and B have no common elements, therefore the condition of the third axiom is fulfilled.

However, in practice events can intersect each other. Let event A specify the outcome of rolling 1 and 5, and event B rolling 2, 3, 4, 5 and 6. Their individual probabilities are 2/6 and 5/6, respectively, adding up to 7/6. Such a probability is clearly not permissible, as it contravenes the second axiom (normativity). The reason, of course, is because we counted the single event "rolling a 5" twice. This particular event is in the intersection of A and B, and the solution is to deduct its probability to arrive at the correct total: $P(A + B) = P(A) + P(B) - P(A \cap B)$.

Example

The Venn diagram below shows that 17 students study only biology only and another 13 study biology and chemistry, for a total of 30 biology students.



The diagram further shows that 14 students study only chemistry. Together with the 13 students who study both subjects there are 27 students of chemistry. The total number of students is 44. The relative frequency of those studying biology is the same as the probability of a randomly selected student studying biology: 30/44. Likewise, for students of chemistry we arrive at 27/44. It is obvious that no negative number of students is possible, which is what the first axiom states.
The second axiom states that the probability of the sample space Ω

must be one. The sets of students are not mutually exclusive (or: the sets are not disjoint), we therefore need to prove that the combined probabilities of 30/44 (biology students) and 27/44 (chemistry students) indeed result in a probability of $44/44 = 1$.
We now calculate the combined probability of events B and C as follows:

$$\begin{aligned} P(A) &= 30/44 \\ P(B) &= 27/44 \\ P(A \cap B) &= 13/44 \\ P(A + B) &= 30/44 + 27/44 - 13/44 \\ &= 44/44 \\ &= 1. \end{aligned}$$

Since the students of both subjects constitute the total sample space, this demonstrates that by making use of the third axiom we indeed arrive at what the second axiom postulates: The probability of the sample space Ω equals one.

Conditional Probability

Conditional probability incorporates the knowledge that the occurrence of one event may give more information about the assessment of a current event. Suppose we have two events, A and B. If we know (or assume) that B has already occurred, we want to express the probability that event A now also occurs. This is the conditional probability defined by:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{1.1}$$

The quantity $P(A|B)$, pronounced "probability of A, given B", is the conditional probability that event A occurs if B has already happened (or is assumed). If the events A and B are independent, then $P(A|B) = P(A)$ because then the events A and B can happen without any influence on each other.

We can use the conditional probability to decompose the probability of

event A into constituents:

$$P(A) = \sum_i P(A|B_i)P(B_i) \tag{1.2}$$

This is known as the "total law of probabilities". The advantage of this decomposition is that in many cases the smaller parts are easier to quantify than the total probability for A. For example, consider a machine. It is quite difficult to come up with the total probability of the machine failing. However, we can probably imagine a number of particular incidents B_i that will lead to the failure of the machine, and by using Eqn. (1.2) we can express the total failure probability A as a sum of all (known) ways the machine can fail, weighted by the probability at which they occur.

In the case of a random variable, we say that we "condition on" this random variable if we consider that this variable takes a specific value. For example, if we consider the random variable X which takes value x, i.e. $X = x$, the probability of event A occurring given $X = x$ is given by $P(A|X = x)$. We then say that this is the probability of A, conditional on $X = x$.

Example

Suppose there is a situation where it is known that, say, 0.1% of the individuals in a population carry a certain disease. We further assume that the presence of the disease can only be detected using specialized equipment. Let D be the event that a random individual tests positive for the said disease: $P(D) = 0.001$.
When actually testing individuals for the said disease we have to allow for the fact that the testing equipment is not completely accurate in detecting the disease. This fact leads to the following possibilities:

1. The test result is positive, and correctly so (true positive; TP).
2. The test result is positive, and wrongly so (false positive; FP).
3. The test result is negative, and correctly so (true negative; TN).
4. The test result is positive, and wrongly so (false negative; FN).

We can also express this using conditional probabilities: Given that the person has the disease, what is the probability of the test being

positive or negative, i.e., $P(T+|D+)$, these are the true positives, or $P(T-|D+)$, which are the false positives.
Using the following table, we can illustrate the concepts of TP, TN, FP, and FN:

	D+	D-	
Test result: positive	9 TP	198 FP	207 P
Test result: negative	1 FN	792 TN	793 N
	10	990	1000

where: D+ means "disease present" and D- means "disease absent". This table can be interpreted as follows:

- Of 1000 people, 10 carry the disease. Out of the 10 carriers, testing showed 9 positive results (true positives; TP) and 1 negative result (false negative; FN).
- Of 1000 people, 990 do not carry the disease. Out of the 990 non-carriers, testing showed 792 negative results (true negatives; TN) and 198 positive results (false positives; FP).

Obviously, in reality, test results are not 100% accurate. To describe how well the test discriminates between people with and without a disease, we calculate sensitivity and specificity.
Using the numbers from the example above, we can define: Sensitivity is the probability of a positive test result, under the condition of a present disease: $P(P|D+) = \#TP/(\#TP + \#FN) = 207/(207 + 1) = 0.995$.
Specificity is the probability of a negative test result, under the condition of an absent disease: $P(N|D-) = \#TN/(\#TN + \#FP) = 793/(793 + 198) = 0.800$.
To find out how well the test rules in disease, we look at the predictive value of a positive test, which is the proportion of people with positive tests who have disease. This is the same as the post-test probability of disease given a positive test: $P(D+|P) = \#TP/(\#TP + \#FP) = 207/(207+198) = 0.511$. To find out how well the test rules out disease, we look at the predictive value of a negative test, which is the proportion of patients with negative tests who do not have disease: $P(D-|N) = \#TN/(\#TN + \#FN) = 793/(793 + 1) = 0.999$.

In generalized terms, the four types can be summarized as follows:

	Reality: True	Reality: False
Measurement: True	Correct	Type I Error (False Positive)
Measurement: False	Type II Error (False Negative)	Correct

Self-Check Questions

1. True or False: Pr(Sample space) = 1
2. What does the first Kolmogorov Axiom state?
3. What does the third Kolmogorov Axiom state?

Solutions

1. True
2. The probability P of an event E is a non-negative real number: $P(E) \geqslant 0, P(E) \in \mathbb{R}$.
3. Additivity. If two events A and B are mutually exclusive, then the probability of either A or B occurring is the sum of the probabilities of A and B: $P(A + B) = P(A) + (B)$. This applies also for a sequence of mutually exclusive events, such that $A_i \cap A_j = \varnothing \; \forall i, j$, then $P(A_1 \cup A_2 \cup \cdots \cup A_j \cup \cdots) = P(A_1) + P(A_2) + \cdots + P(A_j) + \cdots$

1.3 Probability distributions

We have previously introduced the concept of a random variable as a quantity where the numerical value depends on the outcome of a (potentially future) random experiment. In mathematical notation, "X is a random variable is" written as $X : \mathcal{S} \longrightarrow \mathbb{R}$. This means that we cannot predict

the individual value such a random variable will take when we observe the next outcome.

However, this does not mean that we cannot describe the general properties of a random system. The word random—contrary to our everyday experience—does not necessarily imply that each outcome is equally likely, but that some outcomes, or values of a random variable, may be more frequent than others. For example, in our everyday experience we refer to the drawing of lottery numbers, the result of rolling a fair die, or tossing a fair coin as random. In these examples, each outcome is equally likely: We may roll any number between one and six on a die or the flipped coin may show heads or tails with equal probability. These **stochastic processes** are indeed random experiments - but with the additional requirement that each outcome is equally likely. However, we can easily imagine a situation where this requirement is not fulfilled (for example, a coin where one side is heavier than the other so it tends to fall on this side or a die that has an additional weight at a specific number). This means that while the outcome of each experiment, such as flipping the biased coin or rolling the biased die, still cannot be predicted for each individual experiment, now one outcome is much more likely than the others.

A stochastic process is a system described by random variables.

Probability distributions can either discrete, meaning that the variable can only take specific (generally integer) values, or continuous (the variable can take any value).

A probability distribution is often denoted by a capital letter, e.g., B and we use the $\sim$ operator to express that this distribution follows a specific mode. For example, the notation $X \sim B(\frac{1}{3})$ expresses that the variable X follows a model B that in turn depends on some parameter that, in this example, takes the value $\frac{1}{3}$. Individual values of a probability distribution are often denoted by the same letter written in lowercase, e.g., x. It is important to differentiate X (a mapping from the sample space to values) from individual values (a single real number), e.g., x or x_i. More formally, we can say that this is a map from $\mathcal{S}$ to $\mathbb{R}$.

Mappings

The fundamental building blocks of mathematics are sets which are rather generic collections. Sets are the object of study of set-theory.

Examples of sets include the sample space (a rather unknown set), the set of real plants, or the set of points in a space.
A mapping or map between two sets A and B is a set of pairs (a, b) with exactly one pair for each a. One says that a is mapped, or associated, or in relation to b. A mapping is generally noted with a small Roman letter, such as f or h, has a domain of definition (A in our case) and a destination set (B in our case); it is common to use the transformation (or "machine") notation for a mapping f, one writes $b = f(a)$ to mean that the pair (a, b) is in the mapping; this allows, for example, to speak of $f(a) + x$.
Examples of mappings include:

- real functions such as $f : \mathbb{R} \to \mathbb{R}$ defined by (x, x^2) and generally written as $f(x) = x^2$
- parametric equations of curves in the space such as that of the spiral $f : \mathbb{R} \to \mathbb{R}^3$ defined by $f(t) = (\cos(t), \sin(t), t)$
- sequences of numbers such as the harmonic sequence $x : \mathbb{N} \to \mathbb{R}$ defined by $x_n = \frac{1}{3+5 \cdot n}$,
- real functions with a slightly irregular definition domains such as the inverse of the remainder function: $f : \mathbb{R} - \mathbb{N} \to \mathbb{R}$ (where $\mathbb{R} - \mathbb{N}$ is the set of non-integer real numbers) defined by $f(x) = \frac{1}{x - floor(x)}$,
- and, finally, random variables (also called distributions): they are mappings $X : \mathcal{S} \to \mathbb{R}$ with a discrete or continuous destination set.

A discrete probability distribution can be characterized by its probability mass function, which assigns a probability to each possible value.

$$P(X = x) = f_X(X) \tag{1.3}$$

A continuous probability distribution can be characterized by its cumulative distribution function $P(X \leqslant x)$.Because it is often simpler to compute with, the probability density function $f_X(x)$ is generally used. It is defined

to be the function f_X such that

$$P(X \leqslant x) = \int_{-\infty}^{x} f_X(t)dt \tag{1.4}$$

Often, the cumulative distribution is denoted by the capital letter corresponding to the probability density function. For example, if the probability density distribution is given by $f(.) = \ldots$, the corresponding formula for the cumulative distribution is written as $F(.) = \ldots$.

To make the notation more tangible, we use the following example: the distribution of the random variable X follows (or: is distributed according to) a Gaussian (or Normal) distribution and we write this as:

$$X \sim \mathcal{N}(\mu, \sigma) \tag{1.5}$$

The density of the distribution is given by (see Tab. 1.3):

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1.6}$$

and depends on the parameters μ for the mean of the Normal distribution and σ for its standard deviation. The values the random variable X takes are then denoted by a lowercase letter, such as x. Therefore the set of numbers x of variable X are distributed according to some probability distribution, such as the Gaussian distribution in the example above.

Often, a subscript is used to indicate the random variable which we refer to in order to make this more explicit. In the above example, we could therefore write $f_X(x)$ to express that the probability density function f is used to describe the behavior of variable X. Note that this notation is very similar to, but should not be confused with, the notation for partial derivatives. In partial derivatives, subscript is used to indicate the variable we are taking the partial derivative of. Here, we use the subscript to express that the function describes the behavior of a variable. Furthermore, the parameters are often included in the description of the density. In the example of Eqn. (1.6) above, we did not include the parameters explicitly. Often, we can find the following notation in the literature as well: $f(x; \lambda_1, \lambda_2, \ldots)$ where we indicate the parameters explicitly. Here, the independent variable (x) is separated from the parameters ($\lambda_1, \lambda_2, \ldots$) by a semi-colon to express that the parameters are determined outside the

evaluation of the function. In our example above, instead of Eqn. (1.6) we could write:

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1.7}$$

This way, the parameters μ and σ that characterize the distribution $f(.)$ are made explicit and we have to know the value of μ and σ if we want to evaluate the function $f(.)$ at specific values x_i of the variable X.

Because of the Kolmogorov axioms earlier, we note that probability distributions have the following characteristics:

- Positivity: The value of the probability distribution $f_X(x)$ is always positive semi-definite, i.e., $f_X(x) \geqslant 0 \; \forall x$.
- Normalization: The probability distribution is always normalized, i.e., $\sum_i f(x_i) = 1$ for discrete probability distributions and $\int_{-\infty}^{\infty} f(x)dx = 1$ for continuous distributions. That means that "something" has to happen, i.e., one of the events described by the distribution has to occur.

For the case of continuous probability distributions we note that the probability is always assigned for a range and not for a specific value. This means that we do not specify the probability that the random variable x has the exact value of, say $x = 3.14$ but is in a given range $a \leqslant x < b$:

$$P(a \leqslant x < b) = \int_a^b f(x)dx \tag{1.8}$$

This also implies that the probability for a specific value $a = b$ is always zero.

We summarize the most important discrete probability distributions in Tab. 1.2 and continuous distributions in Tab. 1.3. We will discuss some of these in more detail later. We refer to the way we express that a variable follows a given distribution in the column "Notation," and the column "pmf" lists the functional form the distribution has. In the table, we always assume that we refer to a variable X and therefore use the sub-script $f_X(x)$. For simplicity, we do not include the parameters of the distribution in the definition itself, i.e., we will write $f_X(x)$ and not $f_X(x;\lambda_1,\lambda_2,\ldots)$.

The probability distributions we have discussed so far are similar in that they all represent one random variable. We can extend the concept and

Table 1.2: Important discrete probability distributions

Name	Notation	pmf
Binomial	$X \sim B(n,p)$	$f_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$
Negative Binomial	$X \sim NB(r,p)$	$f_X(k) = \binom{k+r-1}{k} p^k (1-p)^r$
Poisson	$X \sim P(\mu)$	$f_X(k) = \dfrac{\mu^k e^{-\mu}}{k!}$

Table 1.3: Important continuous probability distributions

Name	Notation	pdf
Gauss (Normal)	$X \sim N(\mu,\sigma)$	$f_X(x) = \dfrac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
Exponential	$X \sim Exp(\lambda)$	$f_X(x) = \lambda e^{-\lambda x}$ for $x > 0$
Gamma	$X \sim \Gamma(k,\theta)$	$f_X(x) = \dfrac{x^{k-1} e^{-x/\theta}}{\Gamma(k)\theta^k}$
Cauchy	X a Cauchy distribution	$f_X(x) = \dfrac{1}{\pi} \dfrac{1}{1+x^2}$
Student t	T a student-t of ν degrees of freedom	$f_T(t) = \dfrac{1}{\sqrt{\pi\nu}} \dfrac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} \left(1 + \dfrac{t^2}{\nu}\right)^{-(\nu+1)/2}$
Weibull	$X \sim W(\lambda,k)$	$f_X(x) = \begin{cases} \frac{k}{\lambda} \left(\frac{x}{\lambda}\right)^{(k-1)} e^{-(x/\lambda)^k} & x \geqslant 0 \\ 0 & x < 0 \end{cases}$

define probability distributions for two or more variables. This is called the joint probability distribution. In the case of two variables, X and Y, the probability distribution is then denoted by $f_{X,Y}(x, y)$ to indicate the functional dependency on x and y.

We can define the marginal distribution which is given by the following integrals in case of continuous distributions:

$$f_X(x) = \int_{y=-\infty}^{\infty} f_{X,Y}(x, y)dy \tag{1.9}$$

$$f_Y(y) = \int_{x=-\infty}^{\infty} f_{X,Y}(x, y)dx \tag{1.10}$$

Intuitively, the marginal distributions mean that we look at the projection of the joint distribution such that we only keep the variation along one axis and "integrate out" (or ... sum) the other(s). Again, note that in this case the subscript indicates the variable and not a partial derivative. Therefore, the X, Y in $f_{X,Y}(x, y)$ means that we look at variables X and Y and the function $f(.)$ describes their behavior and not that we look at a two-dimensional function $f(x, y)$ where we take partial derivatives with respect to x and y. The notation is almost the same and therefore very confusing and mostly only clear from the context. However, in most cases, partial derivatives are indicated with lowercase letters. In this course, we shall express partial derivatives using symbols such as $\frac{\partial f}{\partial x}$.

We can then extend the two-dimensional case to more random variables, i.e., instead of x and y we use the variables $x_1, x_2, \ldots, x_n$.

For example, the multivariate Gaussian or Normal distribution of a vector $\vec{x} = (x_1, x_2, \ldots, x_n)$ of random variables is given by:

$$f(\vec{x}; \vec{\mu}, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\left(\vec{x} - \vec{\mu}\right)^T \Sigma^{-1} \left(\vec{x} - \vec{\mu}\right)\right) \tag{1.11}$$

where $\vec{\mu}$ denotes the vector of the means and Σ is the covariance matrix defining the Gaussian distribution. The distribution for

$$\vec{\mu} = (0.5, -0.2)$$

$$\Sigma = \begin{bmatrix} 1.0 & 3/5 \\ 3/5 & 2 \end{bmatrix}$$

is shown in Fig. 1.3.

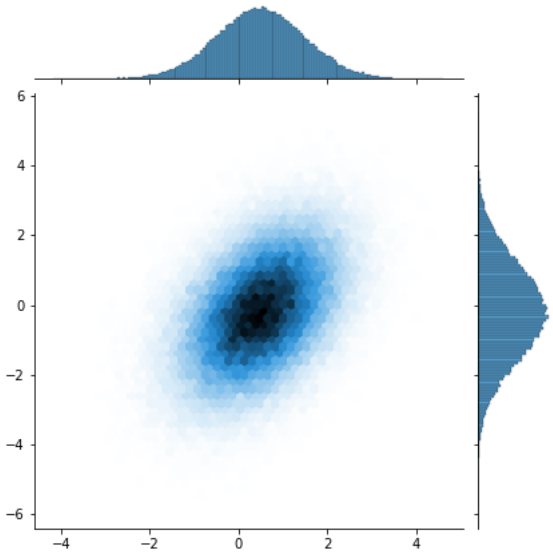


Figure 1.3: An example multivariate Gaussian distribution in two dimensions.

Self-Check Questions

1. What is a stochastic process?
2. True or False: A probability distribution can take negative values.
3. What is the normalization requirement for continuous probability distributions?
4. How is the marginal distribution $f_X(x)$ defined?

Solutions

1. A stochastic process is a system described by random variables.
2. False, $f(x) \geqslant 0 \; \forall x$
3. $\int_{-\infty}^{\infty} f(x)dx = 1$

4. $f_X(x) = \int f_{XY}(x, y)dy$

1.4 Samples and Statistics

We have already established that when we talk about a population in statistics, we do not refer to people as such but to any measurable objects. We also understand a population to include all the elements of a set of data, as opposed to a sample, which consists of only a few observations drawn from the population.

The main reason why we draw a sample is because often the size of a population is too large to measure each individual object. Before measuring an object we need to be clear about the attribute(s) of that object we are interested in. This may be the weight of a person, or the temperature at which a production process takes place, taking into consideration the measurement scales which were introduced earlier.

From the data items derived from the sample we may calculate measures such as the arithmetic mean or the standard deviation. If we have a model of an underlying probability distribution, we can also use them to estimate its parameters. Measures derived from a population are also referred to as parameters, commonly denoted with a Greek letter. Measures from a sample are referred to as sample statistics, commonly referred to with a Latin letter. An exception is $\bar{x}$, which is the arithmetic mean of a sample.

Order Statistics

Order statistics describe the sample in an ascending order and help us describe the sample in a structured manner. Order statistics of a variable Y are denoted as $Y_{(1)} \leqslant Y_{(2)} \leqslant Y_{(3)} \leqslant \ldots \leqslant Y_{(n)}$. The sample observation with the lowest numerical value is known as the first order statistic. For a sample size of n, the n-th order statistic is the maximum, i.e., $Y_{(n)} = \max\left\{Y_{(1)}, \ldots, Y_{(n)}\right\}$

For example, we may wish to divide the number of observations in an ordered sample into four equal parts, also known as quartiles. The first quartile comprises the first quarter up to a point known as Q_1, the second

quartile includes the second quarter up to Q_2, and so forth. The point Q_2 is also known as the median: below and above the median are 50% of the data items, respectively.

Example

For the data set {3, 7, 8, 5, 12, 14, 21, 13, 18}, to find the order statistics, we first order the observations, either ascending or descending, as 3, 5, 7, 8, 12, 13, 14, 18, 21. Sample size n=9. $X(1) = 3$, $X(9) = 21$, the median = 12, the range $R = 21 - 3 = 19$.

An estimate of a population parameter, otherwise known as a sample statistic, is sufficient if no other statistic that provides additional information about the sample and the population parameter can be computed from a sample within the population (Fisher, 1922). That is, the sample statistic computed from a single sample is just as good as a sample statistic from all possible samples from the population. If for a provided sample we had observations 1, 2, 3, 4, 5, then from quick calculations the sample mean is 3. In the event that we knew the sample mean without having the actual sample data, then we would not lose anything in estimating the population mean since we already have the sample mean (even though we had no sample). Another example is order statistics which are sufficient statistics for the sample. Sufficiency in statistics is valuable because it allows us to accomplish data reduction without losing important details. We also call this process lossless data compression.

Thus, having a sufficient statistic implies that we do not need more information about a sample to estimate the population parameter after a sufficient sample statistic is is available. If in an experiment is to be modelled after a single trial ($P(head) = p$ and $P(tail) = 1 - p$) and there were 100 coin tosses, then finding out that the number of heads is, say 53, is sufficient to estimate the parameter p.

Self-Check Questions

1. What is the definition of sufficient statistics?

2. The median is the ... quartile in order statistics.

Solutions

1. A statistic is called sufficient if "no other statistic that can be calculated from the same sample provides any additional information as to the value of the parameter" (Fisher, 1922)

2. second

1.5 Problems of Dimensionality

To illustrate the problems of dimensionality (sometimes also called the "Curse of Dimensionality", a term first coined by Bellman (1957, p. ix)), we use an example with a set of images depicting either a cat or a dog.

In this example we create a classifier that is able to distinguish dogs from cats automatically. To do so, we first need to think about a descriptor for each object class (cat, dog) that can be expressed by numbers to recognize the object. We could, for instance, argue that cats and dogs generally differ in color. A possible descriptor that discriminates these two classes could then consist of three numbers: the average red color, the average green color and the average blue color of the image. A simple classifier could combine these features to decide on the class label, as shown in this pseudo-code:

If 0.5*red + 0.3*green + 0.2*blue is greater than 0.75 return cat; else return dog;

However, these three color-describing numbers (features) will obviously not suffice to obtain a good classification. Therefore, we should add some features that describe the texture of the image, for instance by calculating the average gradient intensity in both the x and y direction. We now have five features that, in combination, could be used by a classification algorithm to distinguish cats from dogs. We are not yet satisfied and continue to add more features. Perhaps we can obtain a perfect classification by carefully designing a few hundred features? The answer to this question may sound counter-intuitive: No, we cannot! In fact, after a certain point, increasing the dimensionality by adding new features would actually degrade the performance of our classifier as illustrated.

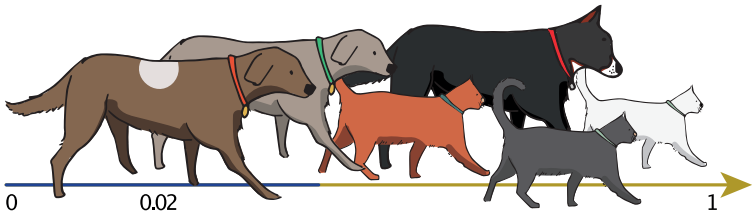


Figure 1.4: Separating cats and dogs using a single feature.

As the dimensionality increases, the classifier's performance increases until the optimal number of features is reached. We will now return to the example of cats and dogs. We assume there is an infinite number of cats and dogs living on our planet. However, due to our limited time and processing power, we were only able to obtain 10 pictures of cats and dogs. The final goal in classification is then to train a classifier, using thes ten training instances, that is able to correctly classify the infinite number of dog and cat instances which we do not have any information about. We use a simple linear classifier and try to obtain a good classification. We can start with a single feature, e.g., the average "red color" as shown in Fig. 1.4.

We note that a single feature does not result in good separation of our training data. We therefore add the feature average "green color" as a second feature shown in Fig. 1.5.

Adding a second feature still does not result in a linearly separable classification problem: No single line can separate all cats from all dogs. Finally, we decide to add a third feature: the average "blue color" in the image, yielding a three-dimensional feature space.

Adding a third feature results in a linearly separable classification problem in our example. A plane exists that separates dogs from cats very well as shown in Fig. 1.6. This means that a linear combination of the three features can be used to obtain good classification results on our training data of 10 images.

These illustrations might seem to suggest that increasing the number of

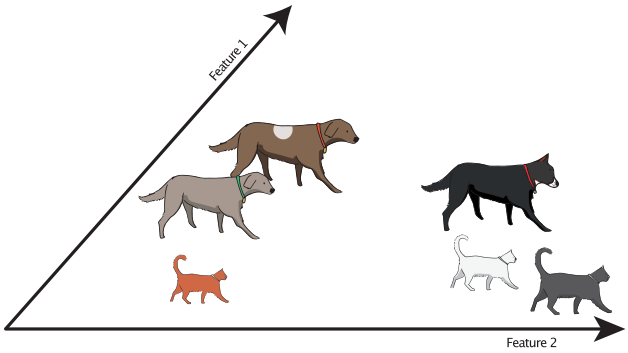


Figure 1.5: Separating cats and dogs using two features.

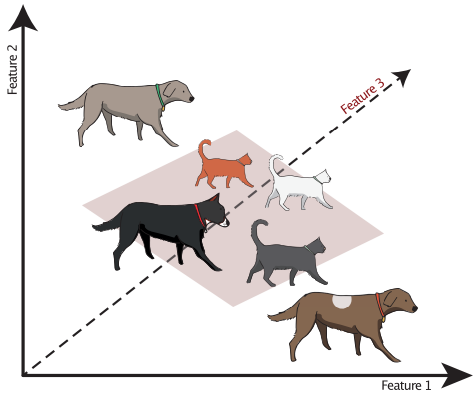


Figure 1.6: Separating cats and dogs using three features.

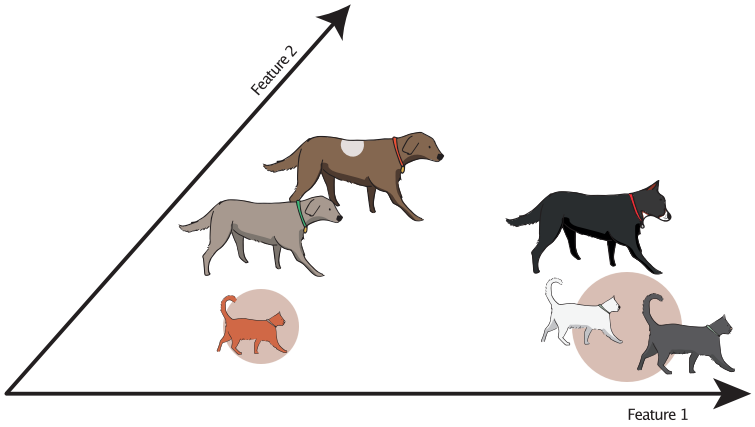


Figure 1.7: Projection of the hyperplane linearly separating between cats and dogs onto a two-dimensional plane illustrating overfitting.

features until perfect classification results are obtained is the best way to train a classifier. However, in the introduction we argued that this is not the case. Note how the density of the training samples decreased exponentially when we increased the dimensionality of the problem. In the case of one feature, 10 training instances covered the complete one-dimensional feature space, the width of which was five unit intervals. Therefore, the sample density was 10 / 5 = 2 samples per interval.

In the two-dimensional case we still had ten training instances, which now cover feature space with an area of 5 x 5 = 25 unit squares. Therefore, the sample density was 10 / 25 = 0.4 samples per interval. Finally, in the three-dimensional case, the ten samples had to cover a feature space volume of 5 x 5 x 5 = unit cubes. Therefore, the sample density was 10 / 125 = 0.08 samples per interval. Adding features means that the dimensionality of the feature space grows and becomes more and more sparse.

A hyperplane is a plane in a higher-dimensional vector space.

Due to this sparsity, it becomes much easier to find a separating **hyperplane**. This is because the likelihood that a training sample lies on the wrong side of the best hyperplane becomes infinitely small when the number of features becomes infinitely large.

However, if we project the highly dimensional classification back to a lower dimensional space a serious problem becomes evident as shown in Fig. 1.7.

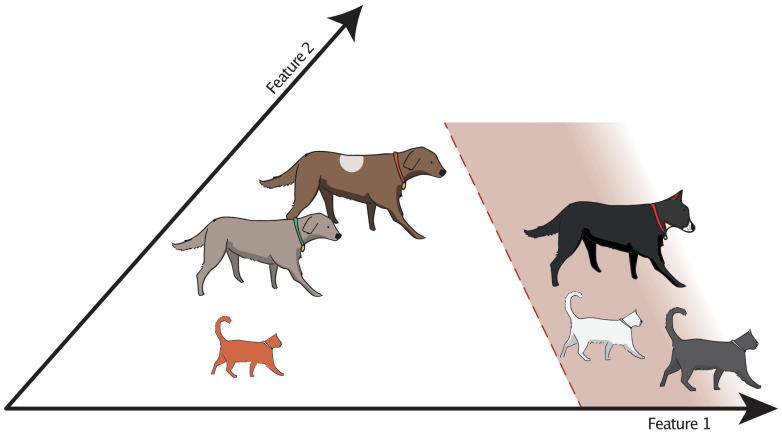


Figure 1.8: Behavior of a classifier using two features.

Using too many features results in overfitting. The classifier starts learning specific details that are specific to the training data. While data was linearly separable in the three-dimensional space, this is not the case in the two-dimensional space. In fact, adding the third dimension to obtain better classification results simply corresponds to using a complicated non-linear classifier in the lower dimensional feature space. As a result, the classifier learns the appearance of specific instances and specific details of our training data set. Because of this, the resulting classifier would fail on real-world data consisting of an infinite amount of unseen cats and dogs that often do not adhere to these specific details. It is a direct result of the curse of dimensionality.

Training on only two instead of three features, the resulting classifier behaves quite differently as shown in Fig. 1.8. Although the simple linear two-dimensional classifier seems to perform worse than the non-linear classifier above, this simple classifier **generalizes** much better to unseen data because it did not learn specific exceptions that were only in our training data by coincidence. In other words, by using fewer features, the curse of dimensionality was avoided such that the classifier did not overfit the training data.

Generalization refers to the ability of a classifier to perform well on unseen data, even if that data is not exactly the same as the training data.

We now illustrate this concept in a different manner as shown by Fig. 1.9. We assume we want to train a classifier using only a single feature whose

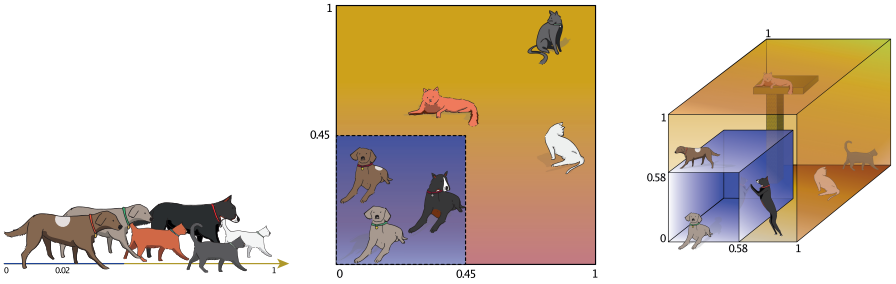


Figure 1.9: As the dimensions increase, a fixed number of samples can cover less and less of the sample space. Even if well-covered in one dimension (left), gaps emerge in two dimensions (middle), and almost all of the sample space is empty in three dimensions (right).

value ranges from zero to one. We also assume that this feature is unique for each cat and dog. If we want our training data to cover 20% of this range, then the amount of training data needed is 20% of the complete population of cats and dogs. If we add another feature, resulting in a two-dimensional feature space, things change: To cover 20% of the two-dimensional range, we now need to obtain 45% of the complete population of cats and dogs in each dimension, since $0.45^2 = 0.2$ (rounded). In the three-dimensional space it gets even worse: To cover 20% of the three-dimensional feature range, we need to obtain 58% of the population in each dimension, since $0.58^3 = 0.2$ (rounded). This illustrates the fact that the amount of training data needed to cover 20% of the feature range grows exponentially with the number of dimensions.

We showed that increasing dimensionality introduces sparseness of the training data. The more features we use, the more sparse the data becomes such that accurate estimation of the classifier's parameters (i.e., its decision boundaries) becomes more difficult. Another effect is that this sparseness is not uniformly distributed over the search space. In fact, data around the origin (at the center of the **hypercube**) is much more sparse than data in the corners of the search space. This can be understood as follows: Imagine a unit square that represent the two-dimensional space. The average of the feature space is the centre of this unit square, and all points within unit distance from this center, are inside a unit circle that inscribes the unit square. The training samples that do not fall within this unit circle are closer to the corners of the search space than

A hypercube is the generalization of a cube into more than 3 dimensions.

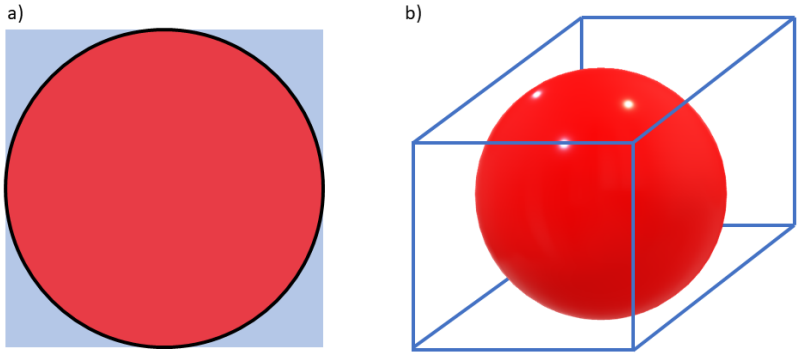


Figure 1.10: Illustration that most training data are outside the central hypersphere in a two-dimensional square and a three-dimensional cube.

to its center. These samples are difficult to classify because their feature values greatly differ (e.g., samples in opposite corners of the unit square). Therefore, classification is easier if most samples fall inside the inscribed unit circle.

Training samples that fall outside the unit circle are in the corners of the feature space and are more difficult to classify than samples near the center of the feature space. The volume of the hypersphere tends to zero as the dimensionality tends to infinity, whereas the volume of the surrounding hypercube remains constant. This surprising and rather counter-intuitive observation partially explains the problems associated with the curse of dimensionality in classification: In high dimensional spaces, most of the training data reside in the corners of the hypercube, defining the feature space as shown in Fig. 1.10. As mentioned before, instances in the corners of the feature space are much more difficult to classify than instances around the centroid of the hypersphere. For an eight-dimensional hypercube, about 98% of the data is concentrated close to its 256 corners.

We have shown that the performance of a classifier decreases when the dimensionality of the problem becomes too large. The question then is what "too large" means, and how overfitting can be avoided. Regrettably there is no fixed rule that defines how many features should be used in a classification problem. In fact, this depends on the amount of training data available, the complexity of the decision boundaries, and the type of classifier used.

Self-Check Questions

1. What is the problem of dimensionality?

2. True or false? As the number of features or dimensions grows, the amount of data we need to generalize accurately grows exponentially.

3. True or false? High dimensional feature spaces cause issues such as losing accuracy of classification due to issues with finding the most representative sample.

Solutions

1. The problem of dimensionality states that the more features we consider, the sparser the data become and the more data we require to avoid overfitting.

2. True

3. True

1.6 Principal Component Analysis and Discriminant Analysis

Principal Component Analysis vs. Discriminant Analysis

Real-world data are often structured in a complex manner. This is especially true for pattern-classification and machine-learning applications. The challenge is to reduce the dimensions of the data sets with minimal loss of information.

There are two commonly-used techniques to achieve this: Principal Component Analysis (PCA) and Discriminant Analysis (DA). One of the earliest references to principal analysis is from Pearson (1901) and Hotelling (1933); for more details about the historic and recent developments refer to, for

example, Jolliffe and Cadima (2016). Discriminant analysis was originally established by Fisher (1936) for two classes and then later extended for use with multiple classes (Rao, 1948).

To illustrate both techniques we use the Iris dataset first established by Fisher (1936). The dataset consists of a sample of the size $n = 150$, containing three classes (types of iris flower), each with four flower features (sepal and petal lengths or widths). Each class has a sub-sample of the size $n = 50$.

Both Principal Component Analysis and Discriminant Analysis are linear transformation methods and are closely related to each other. When using PCA we are interested in finding the components (directions) that maximize the variance in our dataset. With DA we are additionally interested in finding the components (directions) that maximize the separation (discrimination) between different classes. In DA, classes are expressed with class labels. In contrast, PCA ignores class labels. In pattern recognition problems a PCA is often followed by a DA. The difference between the two techniques is summarized Tab. 1.4.

PCA	DA
Projection of the whole data set (without class labels) onto a different subspace.	Identification of a suitable subspace to distinguish between patterns belong to different classes (with class labels).
Whole data set is treated as one class.	Classes in data set are retained.
Identification of the axes with maximum variances where data are most spread	Identification of components that maximize the spread between classes

Table 1.4: Comparison of PCA and DA.

To demonstrate these techniques we use the Iris data set. The flower colors are varied (The flowers, which vary in color, are appropriately named after Iris, the ancient Greek goddess of the rainbow.)

It contains only four variables, measured in centimeters: sepal length, sepal width, petal length, and petal width. There also only three classes: Iris Setosa (Beachhead Iris), Iris Versicolour (Larger Blue Flag or Harlequin

Blue Flag), and Iris Virginica (Virginia Iris). The data set is also known as Anderson's Iris data, since it was Edgar Anderson who collected the data to quantify the morphologic variation of three related Iris species. R. Fisher prepared the multivariate data set and developed a linear discriminant model to distinguish the species from each other.

Even though it is a very simple data set, it becomes difficult to visualize the three classes along all dimensions (variables). We can use the visualization shown in Fig. 1.11. On the diagonal, we see the distribution of each feature variable (as a "histogram") where the color indicates the three species. The off-diagonal elements show the distribution of the data points looking at the combination of two of the feature variables at a time (as a "scatter plot"). We notice that the distribution of sepal width and sepal length is overlapping, therefore we cannot separate one species from another. However, if we look at the Iris Setosa alone, we can see that the variables for petal length and width show a distinct difference from the other two species.

Principal Component Analysis

In Principal Component Analysis we are interested in finding the components (directions) that maximize the separation (discrimination) between different classes. The basic principle is to transform the data into a subspace that summarizes properties of the whole data set with a reduced number of dimensions. We can then use these newly formed dimensions to visualize the data.

The new dimensions are called principal components. The first principal components capture most of the variation in the data. Therefore, along the first principal component are the data that express most of its variability, followed by the second principal component, and so forth. Principal components are orthogonal to each other and therefore not correlated.

Example: Maximizing variance

Instead of the original variables x_1 and x_2, we can choose two new variables that are aligned along the main axes of the ellipse enclosing the data. This means that they are chosen in such a way that these

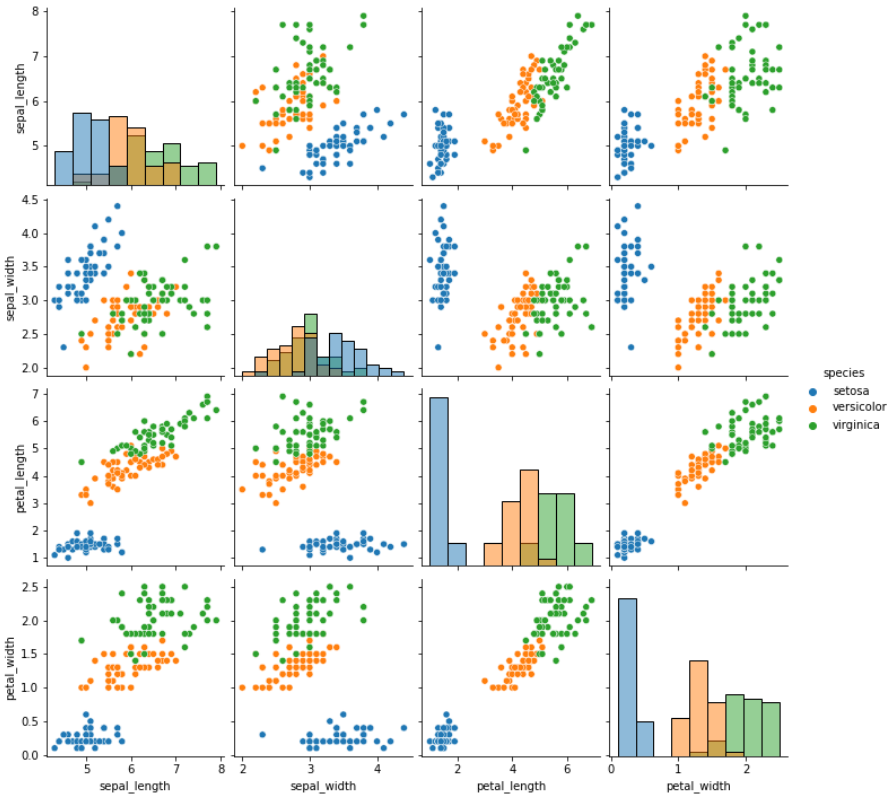
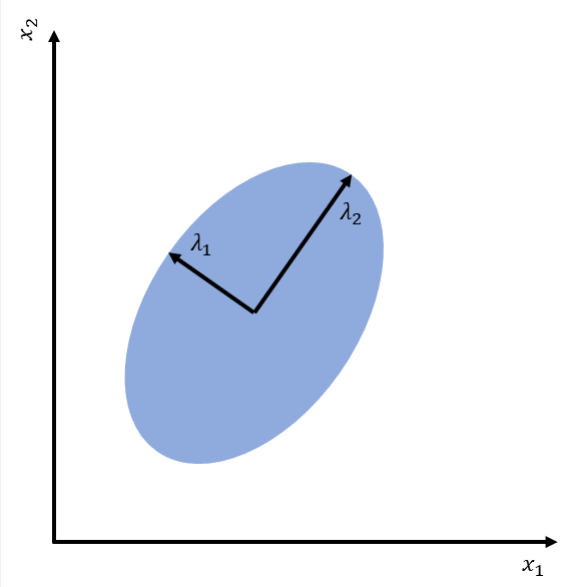


Figure 1.11: Visualization of all feature variables in the iris dataset.

new orthogonal axes λ_1 and λ_2 maximize the variance.



If we look at the off-diagonal elements in Fig. 1.11, (the scatter plots) we see that the variables petal length and width clearly separate the variability of the three classes. In most real-life data sets this would normally not be the case, but the aim of the PCA is to determine new variables such that these new variables explain the variability observed in the data better. Then, these new variables are better suited to separate the classes found in the data.

The steps of the PCA can be summarized as (Raschka, 2014a):

1. Start from the original sample (without class labels)
2. Compute the mean for each variable
3. Compute the covariance matrix between all variables.
4. Determine the Eigenvectors $\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_n$ and the Eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ of the covariance matrix.
5. Sort the Eigenvalues and corresponding Eigenvectors starting from the highest Eigenvalue and place the Eigenvectors in a corresponding

matrix. Choose a suitable cut-off such that only $k < n$ Eigenvalues and Eigenvectors remain.

6. Transform the data using $\vec{y} = \mathbf{W}^T \vec{x}$, where $\vec{x}$ describes the original dataset and $\vec{y}$ the transformed. Here we transform each part of the data-set individually, i.e., one "row" of the data at the time.

The covariance between two discrete variables X and Y is given by:

$$cov_{X,Y} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}), \tag{1.12}$$

where $\bar{x}$ and $\bar{y}$ are the mean of the data samples $\{x_i\}$ and $\{y_i\}$, i.e, the set of numbers we have measured. In our case here, we consider a fixed data sample of distinct values and can therefore use the above formula to calculate the covariance. In most cases, we will have more than two variables, i.e., instead of X and Y we have many variables $X_1, X_2, \ldots$. The covariance matrix captures the covariances between all combinations of all variables. Note that, following the definition above, this matrix is symmetric.

If we perform the PCA using, e.g., `scikit-learn` using Python (Pedregosa et al., 2011), we can see how much of the variance (in percent) is explained by each new variable:

PC1	PC2	PC3	PC4
72.96	95.81	99.48	100.

This means that the first new variable, PC1, retains about 73% of the total variance, and using the first two variables PC1 and PC2 we can explain about 96% of the variance observed in the data, etc. Note that the new variables do not have intuitive names. Essentially, we can understand PCA as a method to find the best linear combination of the original features or variables such that the new variables are ordered by retaining the maximum variance found in the data. Since, ideally, the first few variables retain most of the variance observed in the data, we can then continue with these variables after defining a suitable cut-off of how much variance we would like to retain. This has the advantage that we can limit ourselves to a much smaller list of these new variables. Most real-life data sets have

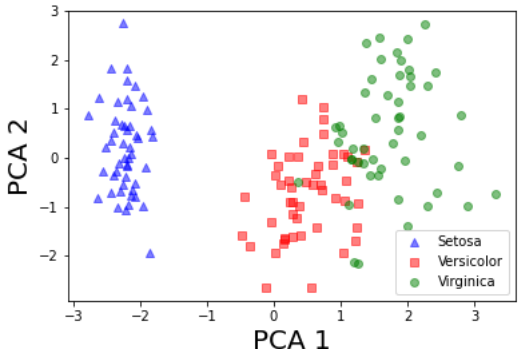


Figure 1.12: Example of the PCA for the first two components of the Iris dataset.

many more features or variables than the Iris data we have used for this example. Reducing a list of, say, several hundred features to maybe twenty has significant advantages computationally.

The result of the transformation for the first two components using a PCA is shown in Fig. 1.12.

For further details about PCA see also, e.g., Raschka (2014a).

Discriminant Analysis

Linear Discriminant Analysis (LDA) is most commonly used as a dimensionality reduction technique in the pre-processing step for pattern-classification and machine learning applications. The goal is to project a dataset onto a lower-dimensional space with good class-separability in order to avoid overfitting ("curse of dimensionality") and also to reduce computational cost.

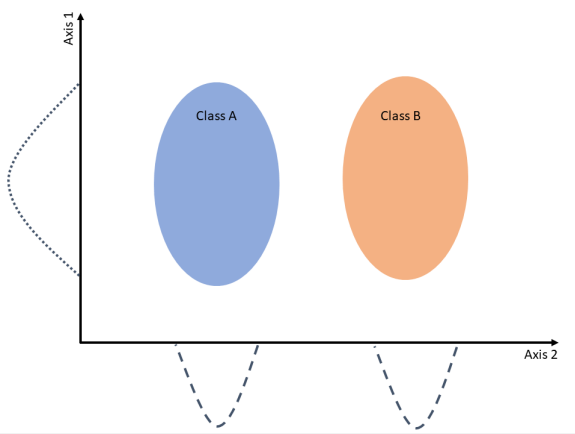
The general LDA approach is very similar to a PCA, but in addition to finding the component axes that maximize the variance of our data (PCA), we are additionally interested in the axes that maximize the separation between multiple classes (LDA). A detailed discussion of the two approaches can be found, for example, in Martinez and Kak (2001).

The main goal of an LDA is therefore to project a feature space (an n-

dimensional dataset) onto a smaller subspace k (where $k \leqslant n-1$) while maintaining the class-discriminatory information.

Example:Maximizing two-dimensional component axes for class-separation.

Define component axis for a dataset with two features or variables that maximize the class separation



Choosing a suitable component axis for class separation in LDA: Choosing "Axis 1" does not allow us to discriminate between class A and B, choosing "Axis 2" separates the two classes.

The LDA can be summarized in the following steps (Raschka, 2014b):

1. Compute the means for each class in the original dataset with n elements and d variables (or dimensions).

2. Compute the scatter matrix both for each class as well as between the class

3. Determine the Eigenvectors $\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_n$ and the Eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ of the scatter matrix.

4. Sort the Eigenvalues and corresponding Eigenvectors starting from the highest Eigenvalue and place the Eigenvectors in a corresponding matrix. Choose a suitable cut-off such that only $k < n$ Eigenvalues

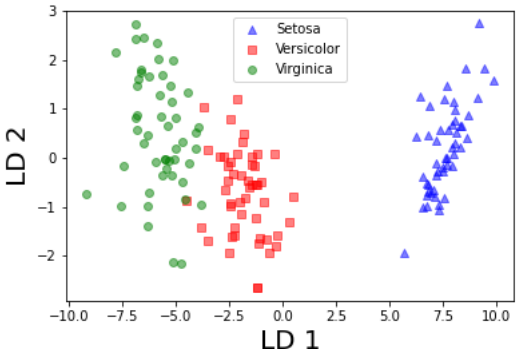


Figure 1.13: Example of the LDA for the first two components of the Iris dataset.

and Eigenvectors remain.

5. Transform the data into a new subspace using the $d \times k$ matrix W. We can write this as a matrix multiplication: $Y = W \times X$ where X represents the whole original dataset (i.e., a matrix of dimension $n \times d$) and Y the whole new dataset (i.e., a matrix of dimension $n \times k$.

The result of the transformation for the first two components using a linear discriminant analysis is shown in Fig. 1.13.

For further details about the discriminant analysis see also, for example, Raschka (2014b).

Self-Check Questions

1. What do the axes in the PCA maximize?
2. Which approach retains the classes in the data, PCA or DA?
3. What is the aim of discriminant analysis?

Solutions

1. The variance in the observed data.
2. DA
3. Identification of components that maximize the spread between classes.

Summary

Statistics is an interdisciplinary art and science: The art of selecting the best methodological approach to reach the study objective. It is important to understand statistical methods in order to understand previous research, as well as to conduct new research in an unbiased, effective, and efficient manner. Randomness is measured in probability. Because they are outcomes of random experiments, random variables are different than deterministic variables. There are two types of random variables: discrete and continuous. Another way to capture probability is by using Venn diagrams. The probability model is concerned with evaluating the likeliness of events. However often when evaluating the likeliness of events, researchers need to incorporate some prior information, such as the conditional probability, independence, and identical distributions. Kolmogorov axioms of domain and range, non-negativity, normalization, and additivity add constraints and help us calculate probability. A probability distribution indicates a description of the probabilities associated with the possible values of the random variable. There are discrete (e.g., binomial) and continuous probability distributions (e.g., Normal). A parameter is a random variable of the population, whereas a statistic is an estimated quantity of the sample. Sample statistics have Roman letters whereas the population parameters take Greek letters. Order statistics describe the sample in an ascending order and they help us describe the sample in a structured manner. A statistic is sufficient if there is no other statistic that can be calculated from the same sample that provides any additional information about the value of the parameter. It can be calculated using conditional probability. As the number of features or dimensions grows, the amount of data we need to generalize accurately grows exponentially. This problem of dimensionality can be alleviated using data reduction techniques such as PCA and DA. Both approaches rely on linear transformation and have their differences and similarities.

2 Descriptive Statistics

Study Goals

After completing this unit you will have learned:

- how to summarize a dataset
- how to compute and interpret location parameters such as mean, median, mode and quantiles of probability distributions
- how to compute and interpret dispersion parameter such as variance, skewness and kurtosis of probability distributions

Introduction

In any data analysis, we start by collecting the raw data. For example, we could read out sensor values, evaluate questionnaires, take measurements from an experiment, etc. Say, in an experiment we would measure the following numbers: 12.48, 9.55, 6.01, 6.72, 4.58, 5.35, 3.68, 13.71, 4.43 and we can visualize these data like in Fig. 2.1. We shall denote the vector of numbers as $\boldsymbol{x}$ and their individual values $x_1, \ldots, x_n$.

We assume for now that we do not have any issues regarding the quality of the data – making sure that we can trust the data we record is a very important part of working with the data, but a separate issue from what we want to discuss now. As we have access to the raw data, the data-points themselves do not tell us too much – if we want to describe it or compare it to other datasets, it is helpful to use a handful of numbers that can help us to summarize the dataset and describe its main characteristics, this is

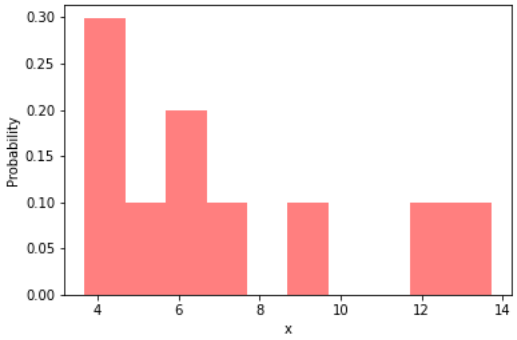


Figure 2.1: Visualization of a data sample

called "descriptive statistics". From our everyday experience, we already know the average value, also called the arithmetic mean, which we obtain by summing all data-points and divide by the number of data-points we have: $\bar{\boldsymbol{x}} = \frac{1}{N}\sum_i x_i = 7.26$. We can already see that this is not ideal for our data sample as, even though we only have 10 data-points, most of them are below the average value of 7.26 and we will need to look for other measures that allow us to describe the data better. We will also need a dispersion metric to quantify how "wide" the data are spread out, as well as metrics that quantify if the distribution of the data is symmetric or not.

However, as we will demonstrate below, we need to keep in mind that if we reduce the description of the data to a few key metrics, we may lose a lot of details that may be crucial to our understanding of the data. It is therefore important to not just rely on metrics defined by the descriptive statistics alone but also to understand and visualize the data we work with.

The data-points we have measured or obtained otherwise form our data sample. Therefore, we call the descriptive statistics that we use to describe these samples "sample statistics" because the computed numbers directly refer to the sample we have recorded. Ultimately, however, we want to infer general properties from the data we observe. In the language of statistics, we want to infer the population metrics from the sample, keeping in mind that we in general have no way of measuring the properties of the whole population. Generally, we assume that the behavior of the population is described by a specific probability distribution (or maybe by a combination of several distributions) and the data-points we observe are a concrete realization of the random variable described by this probabil-

ity distribution. The probability distribution can be either continuous or discrete.

In our example above, the data-points were created using a Moyal distribution (Moyal, 1955) that can be used to model the energy loss of a particle due to ionization. Overlaying the small sample over the underlying probability distribution, we see in the left part of Fig. 2.2 that a small sample does not describe the probability distribution describing the whole population well. However, if we increase the number of data-points in our sample, the description becomes a lot better as shown in the right part of Fig. 2.2. Hence, if we have a large sample, we typically assume that the sample statistics reflect the behavior of the population. However, in many cases we know, infer or model the underlying probability distribution and need to characterize this to describe its behavior and compare it to, e.g., the observed data.

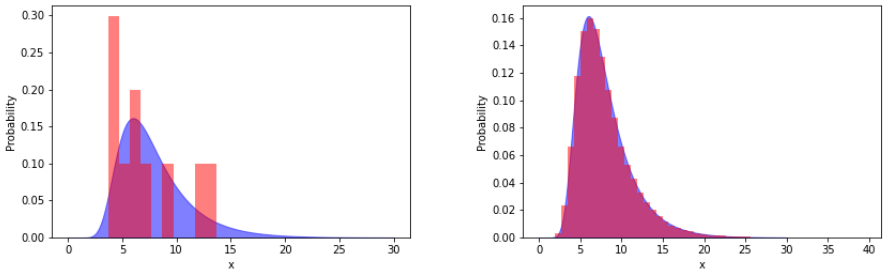


Figure 2.2: Moyal distribution and data sample

2.1 Mean, Median, Mode, Quantiles

Mean

A key part of describing a data sample in sample statistics or a continuous or discrete probability distribution is to define a location parameter that describes where on the axis we can find at least the bulk of the values. From our everyday experience we are already familiar with the average value, this is also called the arithmetic mean which we generally understand to refer to the sample statistics. Given a vector of data-points $\boldsymbol{x} = x_1, x_2, \ldots, x_n$, the arithmetic mean is given by

$$\bar{\boldsymbol{x}} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{2.1}$$

This means that the arithmetic mean is given by the sum of all data-points in our sample, divided by the number of data-points we have. This is not the only way we can define a mean, although this is the most common. Another option is the geometric mean defined as

$$\bar{\boldsymbol{x}}_g = \left(\prod_{i=1}^{N} x_i\right)^{\frac{1}{N}} \tag{2.2}$$

The geometric mean is often better suited to describe growth or growth rates of a given quantity than the arithmetic mean. A further option is the harmonic mean given by

$$\bar{\boldsymbol{x}}_h = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}} \tag{2.3}$$

which is often more appropriate to use when we describe rates or ratios. Finally, we can also define the root mean square given by

$$\bar{\boldsymbol{x}}_{\text{rms}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2} \tag{2.4}$$

This quantity is often used in electrical engineering or to compare a model prediction to observed values. As mentioned above, these definitions are used when we work with a concrete sample of data-points. In other cases however, we will work with either discrete or continuous probability distributions that describe the underlying process of the system we consider and we interpret the data-points we observe as the realizations of a random variable that is described by a probability distribution. An example for a discrete probability distribution is given in Fig. 2.3 The mean value of a discrete random variable X with a probability mass function $P\left(X = x\right)$ is given by:

$$\langle X \rangle = \sum_{i=1}^{N} x_i \cdot P(X = x_i) \tag{2.5}$$

Intuitively, we multiply each value x_i the probability distribution can take with the probability to observe this value $P\left(X = x_i\right)$.

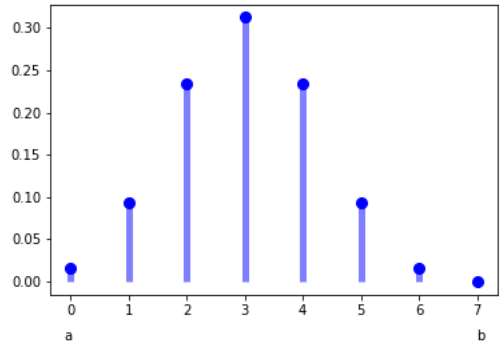


Figure 2.3: The probability mass function of a discrete probability distribution

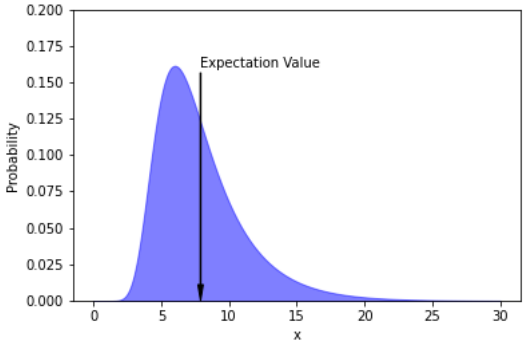


Figure 2.4: Mean or expectation value of a continuous probability distribution drawn on the graph of its density function.

The concept can be brought to continuous probability distributions: Consider a random variable X whose density is $f\,(x)$, we define the expectation, mean, or expected value as

$$\langle X \rangle = \int_{-\infty}^{\infty} x f(x) dx \tag{2.6}$$

An example expectation is shown in Fig. 2.4.

In many cases it is useful to calculate the expected value of the transformed value of the random variable along some function $h\,(x)$: The transformed random variable $h(X)$ which assigns to each event e the value $h(X(e))$ is a distribution and it can be proven using integration by substitution that

the expectation is:

$$E[h(X)] = \int_{-\infty}^{\infty} h(x) f(x) dx \tag{2.7}$$

As a remark, a strict notation for the transformed random variable should rather be $h \circ X$ as it is the composition of mappings (from the sample space to $\mathbb{R}$ then from $\mathbb{R}$ to $\mathbb{R}$), however this abuse of notation is common and practical.

This is important in many practical applications, since – if we have a sufficiently large set of data-points – we can use the sample mean to approximate the expectation value of a probability function, i.e., $\bar{\boldsymbol{x}} \approx E[x]$.

The expectation value is a linear operator: $E[a \cdot g(x) + b \cdot h(x)] = a \cdot E[g(x)] + b \cdot E[h(x)]$.

Example: Expectation of the exponential distribution

The exponential distribution has positive real values and has the probability density function $f(x) = \frac{1}{\lambda} e^{-x/\lambda}$, where $0 \leqslant x < \infty$ and $f(x) = 0$ for $x < 0$ and where $\lambda > 0$. Show that the expectation value is given by λ.

Solution: The expectation value is given by:

$$E[x] = \int_0^{\infty} \frac{1}{\lambda} x e^{-x/\lambda} dx \tag{2.8}$$

where we note that the lower bound of the integral can be set to zero instead of $-\infty$ because the integrand is zero otherwise. In the next step, we integrate the equation by parts. Remember that the rule for integration by parts is given by $\int u dv = uv - \int v du$. Hence:

$$E[x] = \left. -x e^{-x/\lambda} \right|_0^{\infty} + \int_0^{\infty} e^{-x/\lambda} dx \tag{2.9}$$

To evaluate the first part, we have to take the limit $\lim_{x \to \infty} x e^{-x} = 0$, since, intuitively, the exponential function falls faster than the polynomial x. Hence we have to evaluate the second integral $\int_0^{\infty} e^{-x/\lambda} dx$. We remember that $\int e^{ax} dx = \frac{1}{a} e^{ax}$ and hence $\int_0^{\infty} e^{-x/\lambda} dx = \lambda$.

Median

If we go back to our first example and consider the data-points again: $12.48, 9.55, 6.01, 6.72, 4.58, 5.35, 3.68, 13.71, 4.43$ we saw already that the arithmetic (sample) mean $\bar{\boldsymbol{x}} = \frac{1}{N}\sum_i x_i = 7.26$ does not give a very good description of the sample and similarly, if we compare the value of the mean or expected value of the continuous probability distribution above, it does not quite capture how the distribution behaves: The bulk of the distribution is below the mean or expected value but the distribution has a long tail to the right side that influences the mean value. A more robust way to define the location of both a sample and a probability distribution is to define the mid-way point where 50% of the sample or distribution are below this point and 50% are above. This is called the median. To compute the median in case of sample statistics we first order the values of our data points from low to high. In the example above we would then obtain: $3.68, 4.43, 4.58, 5.35, 6.01, 6.72, 9.55, 12.48, 13.71$. The median is then the number that cuts the sample in half, in this case, the corresponding value is 6.01. More generally, if we have N data-points, the median m is given by:

$$m = \begin{cases} x_{(N+1)/2} & \text{if N is odd} \\ \frac{1}{2}\left(x_{N/2} + \ x_{(N/2)+1}\right) & \text{if N is even} \end{cases} \tag{2.10}$$

For discrete distributions, the median $x_{0.5}$ is defined as the following number:

$$P(x \leqslant x_{0.5}) \geqslant \frac{1}{2} \ \text{ and } \ P(x \geqslant x_{0.5}) \geqslant \frac{1}{2} \tag{2.11}$$

In the same way, we define the median of a continuous probability distribution as the point where 50% are below this point and 50% are above. As probability distributions are by definition normalized to one, we can express this using the integral

$$\int_{-\infty}^{x_{0.5}} f(x)dx = 0.5 \tag{2.12}$$

This means that if we take a point at random, it has a 50% chance to fall on the left side of the median $x_{0.5}$ and a 50% chance to fall on the right side.

If we compare mean or the expectation value and median in our previous example as shown in Fig. 2.5 , we find that they are close but not identical.

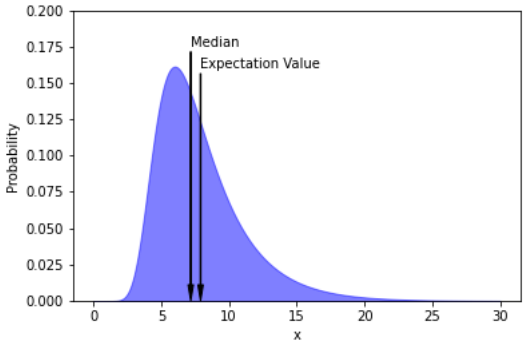


Figure 2.5: Median and Expectation of a continuous probability distribution displayed on the graph of its density function.

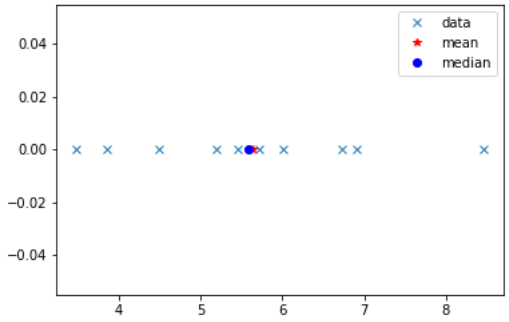


Figure 2.6: Sample of data points

In general, mean and median are the same for symmetric distributions but deviate if the probability distribution is asymmetric. In our case, we have a longer tail to the right side compared to the left side which pulls the expectation value towards the right.

As we have said above, we expect that the median is a bit more robust compared to the mean, since the median defines the mid-way point of our sample or distribution. By robust we mean that the value of this quantity is less sensitive to outliers or the behavior in the tails of the distribution. We can illustrate this in the following way: Consider a sample of data points shown in Fig. 2.6 where all points are relatively close together and symmetric around the value of five. Both mean and median have almost the same value are centered in the middle of the sample. However, if we

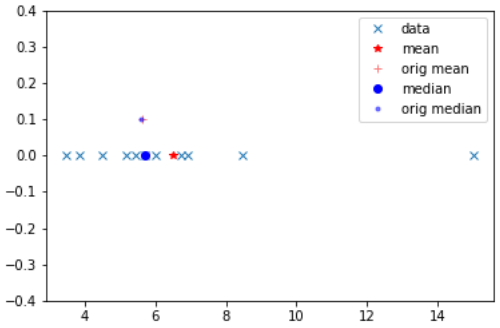


Figure 2.7: Sample of data points with outlier

now add an outlier at 15, we can observe that the mean is affected much more than the median as shown in This is because the mean is calculated from all data-points, including the outlier, whereas the median is defined as the point that cuts the sample in half and adding an outlier does not change this by much.

Quantiles

We have defined above that the median is the point in a probability distribution that splits the distribution in two equal halves: 50% of the values are below this value, 50% are above. Although this point is a convenient choice to describe the localization of a probability function, there is nothing special about the 50% mark per se. We could just as well define a number such that 10% of the distribution are below that point and 90% above, etc. In general, the quantile x_q of a distribution is the point that splits the distribution such that q% are below this point and $(1-q)$% are above. We can therefore define the quantile as:

$$\int_{-\infty}^{x_q} f(x)dx = q \tag{2.13}$$

Using this more general definition, we note that the median is the quantile for which we set $q = 0.5$. We can use the quantiles in general to describe the overall shape of the distribution. For example, if we calculate 100 quantiles, i.e., $q = 0.0, 0.01, 0.02, \ldots, 0.99, 1.0$ we can approximate most

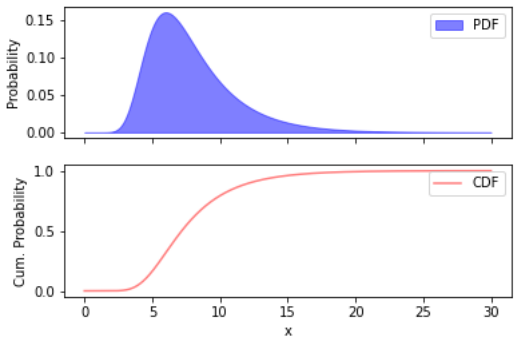


Figure 2.8: Probability Density Function (PDF) and Cumulative Distribution Function (CDF) for a continuous distribution.

functions fairly accurately by plotting these quantiles and interpolating between them.

The cumulative distribution function (CDF) $F(x)$ for a probability distribution $f(x)$ is defined as: $F(x) = \int_{-\infty}^{x} f(x)dx$

If we want to determine a specific quantile, we can also use the **cumulative distribution function** (CDF). Using our previous example, we plot both the probability distribution along with the cumulative distribution as shown in Fig. 2.8

If we want to find the value of a specific quantile, we can use the plot of the CDF, draw a horizontal line for the required quantile and then a vertical line at the point where the horizontal line crosses the plot of the CDF. For example, for the median, we would draw a horizontal line in the lower plot such that it goes through the value of 0.5 on the y – axis and then we draw a vertical line where it crosses the graph. More formally, if the function $F(x)$ is the cumulative distribution for the probability distribution $f(x)$, then the median is given by:

$$x_{0.5} = F^{-1}(1/2) \tag{2.14}$$

where $F^{-1}(.)$ is the inverse of the cumulative distribution. Note that the inverse is not always defined but it can be proven that it exists when the density function is continuous and its domain of definition is an interval.

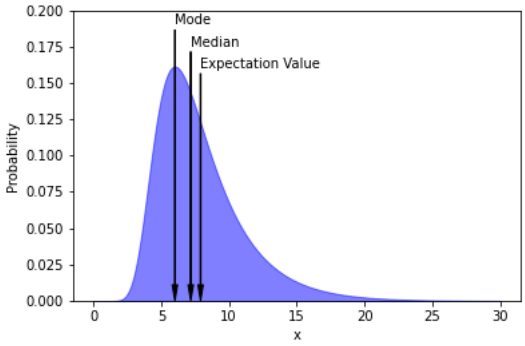


Figure 2.9: Mean, Median, Mode

Mode

The mode specifies the highest point of a probability distribution, meaning that this is the most likely value. Formally, we can define the mode as:

$$\text{mode} = \underset{x}{\text{argmax}}\, f(x) \tag{2.15}$$

Using our previous example, we can compare the mode to the mean (or expectation value) and the median. Each of these parameters describes the location of the distribution in different ways as illustrated by Fig. 2.9. We should note that the mode is not a stable location parameter, as even small details in the distribution will shift the mode noticeably, even if a more robust metric such as the median is hardly affected. Because of this, the mode is rarely used in practice to describe a distribution.

Distributions can also have multiple modes. The example in Fig. 2.10 shows the density function of a distribution with two modes of equal heights. In the strict sense we would say that a distribution has two or more modes if these are of the same height, however, we can also include peaks that are locally higher than their surrounding values in a more general definition.

Self-Check Questions

1. Calculate the mean of the following sample: 7.69, 6.51, 9.01, 9.74, 10.48, 6.01, 7.05, 6.17, 7.28.

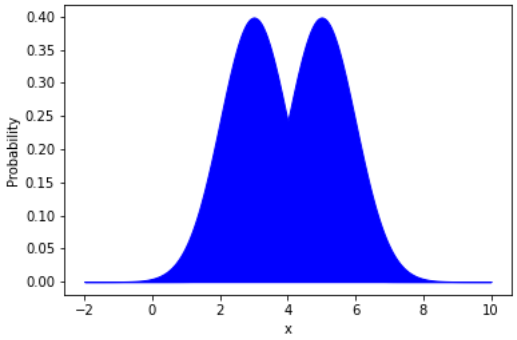


Figure 2.10: The probability density function of a bimodal Distribution

2. Which of the two location parameters, mean or median, is considered to be more robust?

3. A set of 5 data-points has the sample mean of $\bar{\boldsymbol{x}} = 3$. Now the number 4 is added as an additional data point. What is the mean of the expanded set?

Solutions

1. The sample mean is: $\bar{\boldsymbol{x}} = 7.70$.

2. Median

3. We can construct a simple dataset containing 5 data-points with a sample mean $\bar{\boldsymbol{x}} = 3$ like this: $3, 3, 3, 3, 3$. Now we add a new data-point with value 4, i.e. out set is now $3, 3, 3, 3, 3, 4$. Hence the new mean is $\bar{\boldsymbol{x}} = \frac{1}{6}(5 \cdot 3 + 4) = \frac{19}{6}$.

2.2 Variance, Skewness, Kurtosis

Our discussion so far was focused on defining location parameters such as mean, median and mode. However, these metrics do not tell us much about the shape of the distribution. In many cases, we want to know how

"wide" or symmetric a distribution is.

Variance

The variance is a dispersion parameter and measures how much the values fluctuate around the mean.

In sample statistics, where we describe a fixed set of data-points, we define the sample variance as

$$s^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{\boldsymbol{x}})^2 \tag{2.16}$$

And the sample standard deviation as the positive square root of the sample variance:

$$s = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{\boldsymbol{x}})^2} \tag{2.17}$$

The above definition has the disadvantage that we first need to calculate the sample mean $\bar{\boldsymbol{x}}$ before we can determine the sample variance. Using the following alternative definition, we can avoid this and calculate both at the same time:

$$\begin{aligned}
s^2 &= \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{\boldsymbol{x}})^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}x_i^2 - \frac{1}{N}\sum_{i=1}^{N}2x_i\bar{\boldsymbol{x}} + \frac{1}{N}\sum_{i=1}^{N}\bar{\boldsymbol{x}}^2 \\
&= \overline{\boldsymbol{x^2}} - 2\bar{\boldsymbol{x}}^2 + \bar{\boldsymbol{x}}^2 \\
&= \overline{\boldsymbol{x^2}} - \bar{\boldsymbol{x}}^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}x_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N}x_i\right)^2
\end{aligned}$$

Using these definitions, we can illustrate the mean and the variance of a set of data-point as illustrated by Fig. 2.11.

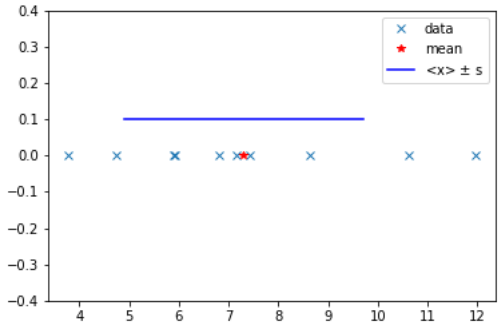


Figure 2.11: Mean and Variance of a set of Data-Points

The sample variance and sample standard deviation give us an indication about how wide the data-points are distributed around the (arithmetic) mean.

When we consider continuous probability distributions, we define the variance as

$$V[X] = E\left[(X - \langle X \rangle)^2\right] = \int_{-\infty}^{\infty} (x - \langle X \rangle)^2 f(x) dx \tag{2.18}$$

in the same way as for the sample variance and accordingly for discrete probability distributions:

$$V[X] = \sum_{i=1}^{N} (x_i - \langle X \rangle)^2 f(x_i) \tag{2.19}$$

Like the sample variance, we can express the variance for probability distributions in a different way:

$$\begin{aligned} V[X] &= E\left[(X - \langle X \rangle)^2\right] \\ &= E\left[X^2 - 2X\langle X \rangle + \langle X \rangle^2\right] \\ &= E\left[X^2\right] - 2E\left[X\right]\langle X \rangle + \langle X \rangle^2 \\ &= E\left[X^2\right] - 2\left(E\left[X\right]\right)^2 \\ &= E\left[X^2\right] - \langle X \rangle^2 \end{aligned}$$

The standard deviation is then again defined as the positive square root,

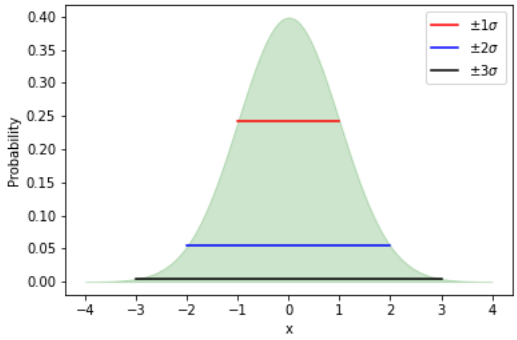


Figure 2.12: Positions of $\pm 1\sigma, \pm 2\sigma, \pm 2\sigma$ for a symmetric distribution

i.e.

$$\sigma = \sqrt{V[X]} \tag{2.20}$$

and is illustrated in Fig. 2.12 for a continuous distribution.

Moments

When defining the mean and variance for probability distributions, we have already encountered the expectation value transformed through some function h. When defining the mean, we took $h(x) = x$, and if we look at the definition of the variance we can see that this definition uses $h(x) = (x - \langle X \rangle)^2$.

We can generalize this and define the algebraic moment of order n as:

$$\mu_n = \int_{-\infty}^{\infty} x^n f(x) dx \tag{2.21}$$

where we take $h(x) = x^n$. We can also define the central moment of order n as:

$$\mu'_n = \int_{-\infty}^{\infty} (x - \langle X \rangle)^n f(x) dx \tag{2.22}$$

i.e., $h(x) = (x - \langle X \rangle)^n$.

With these definitions, the mean is the same as the first algebraic moment, i.e., $\mu = \mu_1 = E[X] = \langle X \rangle$ and the variance is the second central moment.

An important application of the moments is that a probability distribution is defined by all its moments, meaning that if we know all moments of a distribution, we can re-create the distribution from them. In practical application one can approximate a distribution by calculating the first few moments and then stop once the desired precision is obtained.

Proof: A probability distribution is defined by its moments

Consider the density functions of probability distributions $f_1(x)$ and $f_2(x)$ and expand the difference between them into a polynomial, i.e.

$$f_1(x) - f_2(x) = c_0 + c_1 x + c_2 x^2 + \cdots$$

We then evaluate the following integral

$$\begin{aligned} &\int_{-\infty}^{\infty} \left(f_1(x) - f_2(x)\right)^2 dx \\ &= \int_{-\infty}^{\infty} \left(f_1(x) - f_2(x)\right) \left(c_0 + c_1 x + c_2 x^2 + \cdots\right) dx \end{aligned}$$

which is always greater or equal to zero. Using the definition of the moments, we can write this as

$$c_0(1-1) + c_1(\mu_1(1) - \mu_1(2)) + c_2(\mu_2(1) - \mu_2(2)) + \cdots \geqslant 0$$

where $\mu_1(1)$ are the moments for the first function $f_1(x)$ and correspondingly $\mu_n(2)$ are the moments for the second function. If all moments are identical, i.e., $\mu_n(1) = \mu_n(2)$, then the integral has to be zero since the quantity $\left(f_1(x) - f_2(x)\right)^2$ is always positive and hence the two functions are identical (Blobel & Lohrmann, 2012).

Skewness and Kurtosis

The skewness is a measure how symmetric a distribution is: a fully symmetric distribution has a skewness of zero. If a distribution has a tail to the left, the skewness is negative, if the distribution has a tail to the right

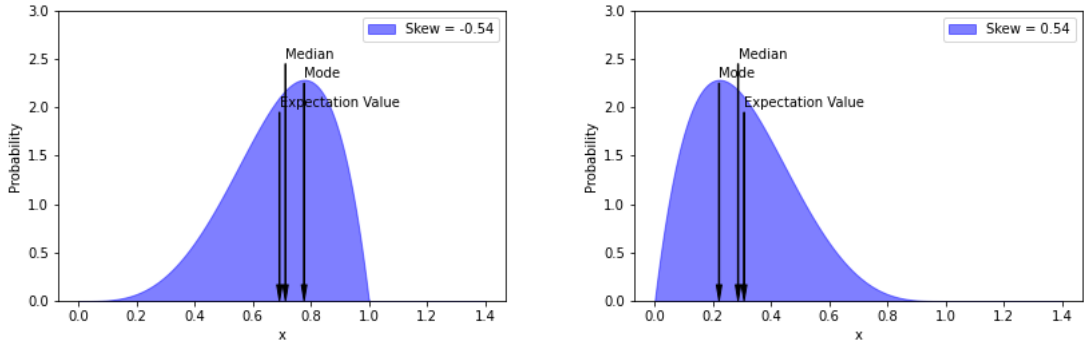


Figure 2.13: Positively and negatively skewed distribution

on the other hand, the skewness is positive. The skewness is defined as:

$$g_3 = \frac{\mu_3'}{\mu_2'^{3/2}} = \frac{\mu_3'}{\sigma^3} = E\left[\left(\frac{X - \langle X \rangle}{\sigma}\right)^3\right] \tag{2.23}$$

An example for a a left- and a right-skewed distribution is shown in Fig. 2.13.

We note that if the skewness is negative, the mean (expectation value) and median are below the mode. If the distribution is positively skewed, mean and median are above the mode. In case of a symmetric distribution with zero skewness, mean, median and mode have the same value.

To calculate the sample skewness, we replace the integrals by the appropriate sums:

$$g_3 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{\boldsymbol{x}})^3}{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{\boldsymbol{x}})^2\right)^{3/2}} \tag{2.24}$$

The kurtosis on the other hand is a measure of how pronounced the tails of a distribution are. It is defined as

$$\kappa = g_4 = \frac{\mu_4'}{\mu_2'^{4/2}} = \frac{\mu_4'}{\sigma^4} = E\left[\left(\frac{X - \langle X \rangle}{\sigma}\right)^4\right] \tag{2.25}$$

In practice, the kurtosis is typically compared to the case of a standardized normal distribution for which $\kappa = 3$. Normalized to this value, the remaining kurtosis is often called the excess or excess kurtosis. As illustrated in

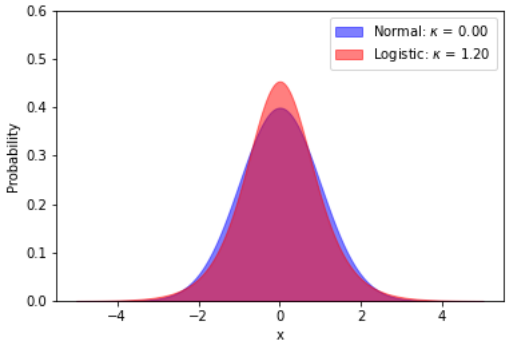


Figure 2.14: Kurtosis for normal and logistic distribution

Fig. 2.14, for example the logistic distribution has longer tails compared to a normal distribution and has hence a positive excess kurtosis.

If we compute the kurtosis from a sample, replacing the integral of the moments with the respective sums, the value is a measure of the outliers in the sample, i.e., a high value for the kurtosis indicates the existence of outliers in the sample.

$$\kappa = g_4 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{\boldsymbol{x}})^4}{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{\boldsymbol{x}})^2\right)^2} \tag{2.26}$$

To calculate the excess kurtosis, we again subtract the value three from the above formula.

We can define higher orders of this dimensionless quantity as

$$g_k = \frac{\mu'_k}{\mu'^{k/2}_2} = \frac{\mu'_k}{\sigma^k} \tag{2.27}$$

although beyond skewness and kurtosis, these are rarely used in practice.

Descriptive Statistics and Distributions

The location and dispersion parameters we have discussed so far can help us to describe both a sample of data-points as well as a discrete or continuous probability distribution. As we have indicated in the introduction earlier,

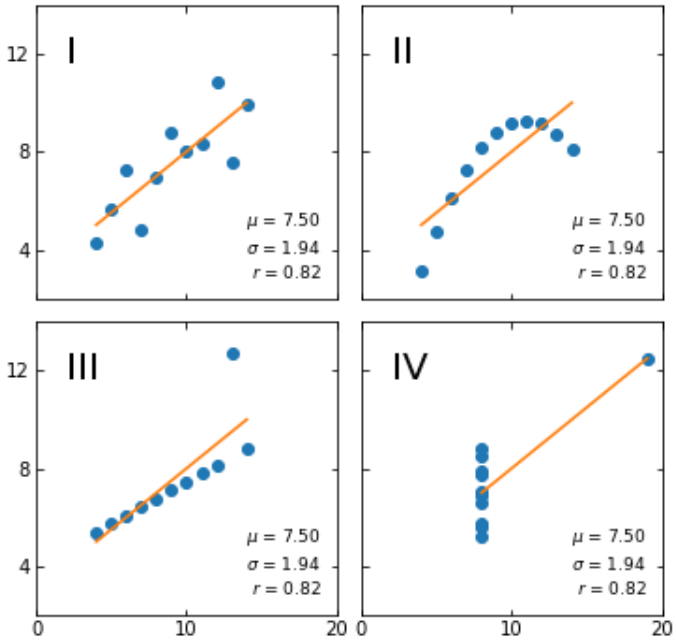


Figure 2.15: Anscombe's Quartet

these metrics are very helpful to get more insights into the behavior of the sample or distribution but we should be careful to rely just on these metrics as they inevitably neglect many details. In the following we want to look a bit more closely why we always need to understand the sample or probability distribution more fully and ook at all the data we have rather than just relying on mean, median, variance or other metrics.

We first look at "Anscome's Quartet" (Anscombe, 1973) shown in Fig. 2.15. The quartet consists of four sets of data-points of equal length. Each set has the same sample mean and sample variance. If we add a regression line, the fitted parameters, as well as a metric for the quality of the fit, are also identical. Therefore, just by looking at these metrics of descriptive statistics, we would not be able to tell the four data-sets apart, even though they are visually very different.

A further and more fun dataset was developed by Mateika and Fitzmau-

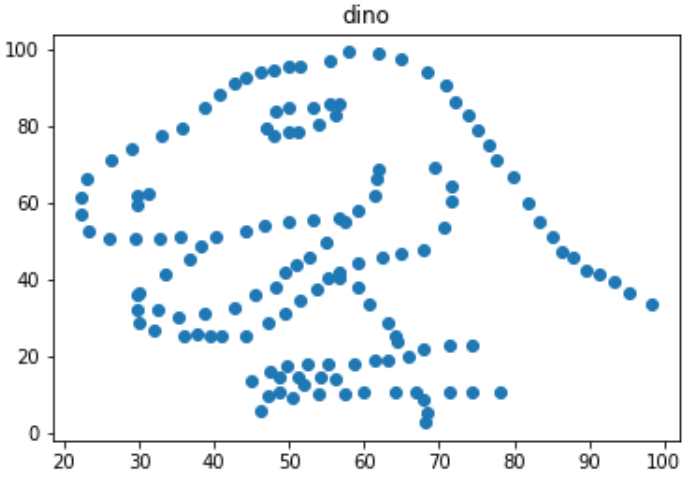


Figure 2.16: The Datasaurus

rice (2017) who also make the corresponding data available (Mateika & Fitzmaurice, n.d.). They created multiple dataset that all have the same mean and standard deviation in both x and y direction. These datasets are called the "Datasaurus Dozen" as they start to morph a dataset showing a dinosaur shown in Fig. 2.16 into different shapes shown in Fig. 2.17.

Self-Check Questions

1. Calculate the variance, skewness and excess kurtosis of the following sample: $7.69, 6.51, 9.01, 9.74, 10.48, 6.01, 7.05, 6.17, 7.28$

2. The kurtosis of a dataset A is 5, the one of another dataset B is 8. Which dataset is more prone to outliers?

3. If the skewness of a distribution is negative, to which side does the distribution has more pronounced tails?

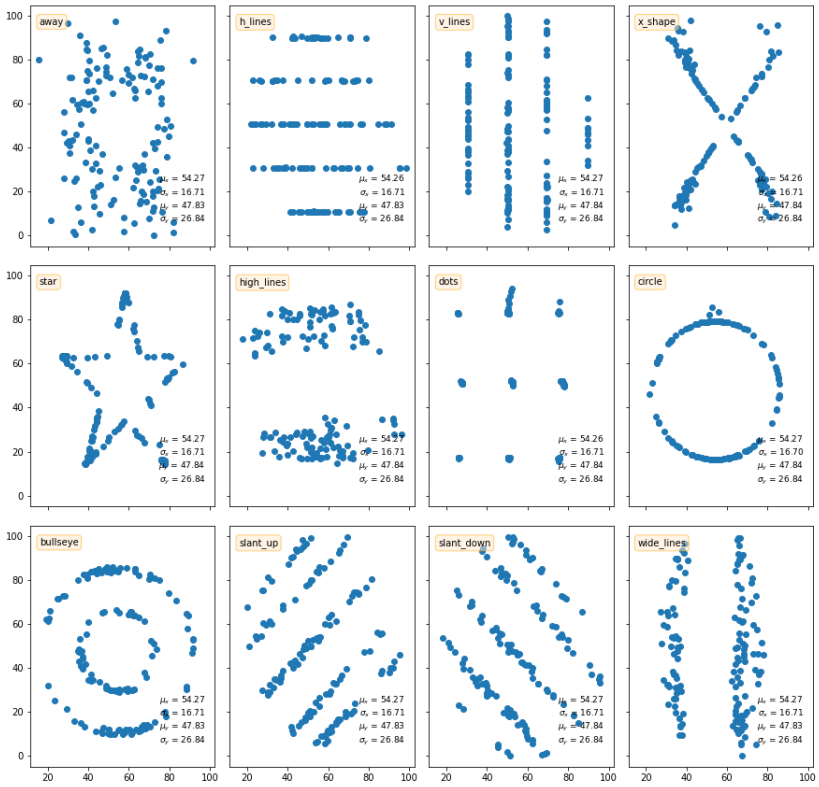


Figure 2.17: The Datasaurus Dozen

Solutions

1. Variance: $s^2 = 2.31$, Skewness: 0.55, Kurtosis = -1.12.
2. Dataset B due to its higher kurtosis.
3. To the left.

Summary

Descriptive statistics is a very valuable tool to describe a dataset or a probability distribution with a few numbers. We generally consider either sample statistics, where we focus on a fixed set of numbers, our data sample, or on metrics that describe the behavior of probability distributions.
We can use location parameters such as the mean, median or mode to describe where on the value axis the sample or distribution is located. The mean is generally the more robust metric as it is less affected by, e.g., outliers than the mean.
The range of a sample or distribution is described by a dispersion parameter such as the variance or standard deviation, whereas the skewness measures the symmetry and the kurtosis how strong the tails of the distribution or data sample are.
More generally, we can define a probability distribution according to its moments, where we distinguish between central and algebraic moments. The mean is the first algebraic moment and the variance the second central moment.
The expectation value of a variable transformed by some function $h(x)$ is an important concept to describe the behavior of probability distributions. The expected value of the transformed variable $h(X)$ can easily be computed. If we have access to a sufficiently large sample of data-points, we can in practice often approximate the expected value with the sample mean, i.e., $\langle \boldsymbol{x} \rangle \approx E[X]$.

3 Important Probability Distributions and their Applications

Study Goals

After completing this unit you will have learned:

- The most important probability distributions
- Where these distributions are used in practice
- The key characteristics of each distribution

Introduction

In our everyday life, we perceive the world as mainly deterministic: Everything seems to follow a strict set of rules and we tend to assume that every action has a distinct outcome or consequence. However, even after a cursory glance, we find that this is not the case: While there are indeed deterministic systems such as a pendulum or grandfather clock (neglecting air resistance or friction), we find that we cannot describe even simple systems this way: If we toss a coin, will it show head or tail? How many cars exactly will drive on a given road in the next hour? How many customers will enter a shop? When will a given component fail?

When we look at many of these and other events, we find that while we cannot describe the outcome in a specific situation, the observed events

follow a specific pattern. This means that the occurrence of events is not fully random as in the case of a lottery draw where we cannot predict the winning numbers at all, but we can describe the pattern of events fairly accurately. For example, while we cannot predict the exact number of customers who will enter a shop in the next hour, we can describe the underlying behavior in terms of a distribution: Each outcome, say, that we observe 10, 11, 12, or more customers in the next hour, can be associated with a probability. We can also describe these distributions in terms of their descriptive statistics such as the expected value (mean) or the variance.

More formally, we say that a given system or specific events are described by a random variable x that follows a specific probability distribution and the values of the random variable represent measurable outcomes. Random variables can be either discrete or continuous, and, correspondingly, probability distributions are also either discrete or continuous. The number facing upwards if we roll a die, is one of 1, 2, 3, 4, 5, 6. This is an example of a discrete probability distribution. The distribution of the height of people in a given country, on the other hand, is a continuous number as humans can, within a given range, grow to any height.

We have encountered the basic concepts of probability distributions in the first unit. In this unit, we discuss some of the most important of these probability distributions that are frequently used in practical applications.

3.1 Binomial and Negative Binomial Distribution

Bernoulli Trials

Let us suppose we toss a coin into the air - which side will face up when we catch it? Head or tail? What will happen if we toss the coin 10 times or 100 times? How many heads will we count and often will the coin face up with tail? Or suppose we own a e-commerce and run a web-shop. Our objective is, in the end, to motivate the customer to buy our goods. Where should we place the icon for the check-out? We can run two different versions of the web-page: In one setting the icon is on the top right, in the other the top left - which one is more successful? This is called an A/B-test. We

can also imagine that we show customers an advertisement on a web-page. Will they click on the corresponding link or not?

What all of these examples have in common is that there are two outcomes: The coin lands head or tail up, a customer clicks on the link or not, etc. In the language of statistics, this is called a Bernoulli trial and we can call these outcomes A and B. Quite often, the outcomes are labelled "success" or "failure" as we typically wish to achieve something: If we place an advertisement on a web-page, we hope the customers will click on it and our ad campaign is successful. If the probability to observe the outcome A is given by $P(A) = p$, the corresponding probability for B is $P(B) = q = 1 - p$. This is because we only have two options, A or B and hence, if we observe A we cannot observe B and vice versa.

When a random variable X describes a Bernoulli trial with probability of success p, one generally writes $X \sim B(p)$.

Binomial Distribution

The Binomial distribution then describes what happens if we perform n independent Bernoulli trials. For example, we toss the coin n times and each event where we toss the coin is independent from all others, meaning that the result does not depend on the sequence of previous observations. The random variable X then describes the probability that we observe the event A, the "success" m times in our n trials. Consequently, X is a discrete random variable that can take the values: $0, 1, 2, \ldots, n$. After we completed our trials, we have X "successes" (or occurrences of event A), and $n - k$ " failures". With $P(A) = p$ and $P(B) = q = 1 - p$, the overall probability is given by: $p^x q^{n-x}$. However, we also assumed that the individual events or outcomes are independent. This means that the sequence $ABAB$ has the same probability as $AABB$ or $BBAA$ or $BABA$: In each case there are two "successes" and two "failures". The number of these permutation is given by the binomial coefficient:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{3.1}$$

for $0 \leqslant k \leqslant n$, that describes the number of ways we can arrange the outcomes if the order or sequence of events does not matter. The binomial

distribution describes the probability to observe k "successes" in n trials where p is the probability to observe a "success" in a single trial or event:

$$X \sim B(n,p) \Longleftrightarrow P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{3.2}$$

The mean or expectation value is given by:

$$\langle X \rangle = E[X] = \sum_{k=0}^{n} kP(k) = np \tag{3.3}$$

This can be understood intuitively by the fact that a Binomial distribution describes a sequence of n Bernoulli trials. Each Bernoulli trial has an expectation value of p for the "success" and since we assume the events to be independent, the expectation value for the series is np.

The variance is given by:

$$V[X] = \sigma^2 = \sum_{k=0}^{n} (k - \langle X \rangle)^2 P(k) = np(1-p) \tag{3.4}$$

Fig. 3.1 shows the behavior of the distribution for $n = 10$ trials with $p = 0.1, 0.2, 0.5$ and $p = 0.9$

Example

What is the probability of a observing five heads in 10 tosses of a fair coin?

We assume the coin to be fair, i.e., the probability for head (or "success") is: $p = 0.5$ Hence, if K is the random variable counting the number of heads:

$$P(K=5) = \binom{10}{5} 0.5^5 (1-0.5)^{10-5} = 0.25$$

In practical applications the following recurrence formula is often helpful:

$$P(X=k+1) = \frac{p}{1-p} \binom{n-k}{k+1} P(X=k) \tag{3.5}$$

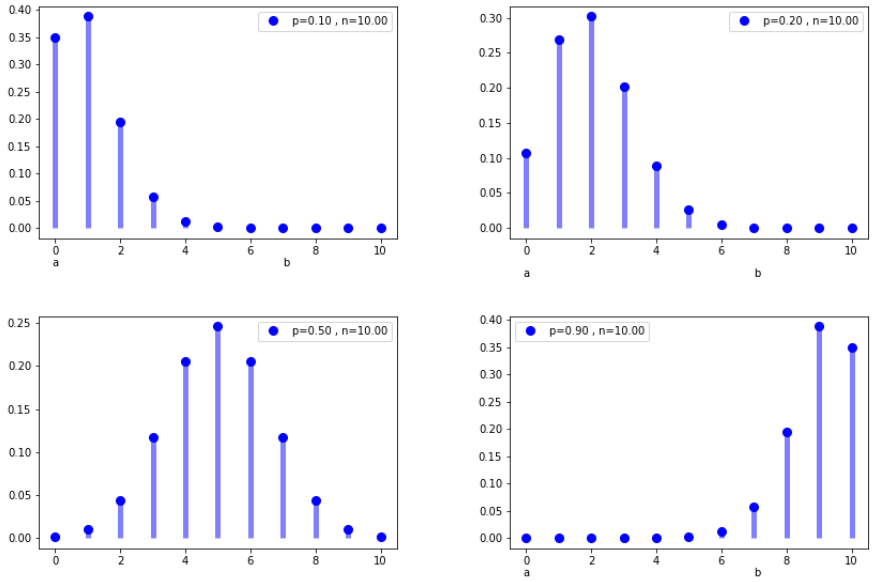


Figure 3.1: Binomial distribution for $n = 10$ trials with $p = 0.1, 0.2, 0.5$ and $p = 0.9$

Negative Binomial Distribution

In the discussion of the binomial distribution so far, we have assumed that we observe the number of successes in a fixed number of trials. For example, we decide to toss the coin 10 times and then observe the number of head and tails.

We could alternatively also ask: How many failures do we observe before we observe the k^{th} success? One way to obtain this result is the following: We first observe $k-1$ successes and then observe m failures, before observing the final success. Again, the probability to observe a success is p and the probability for failure is $q = 1-p$, hence the probability of this sequence is $\prod_{i=1}^{k-1} p_i \cdot \prod_{j=1}^{m} q_j \cdot p = p^k q^m$. As before, this particular sequence is just one permutation and we need to keep in mind that the events are independent from each other. We can express this as before with the binomial coefficient and the negative binomial distribution is then given by:

$$M \sim NB(k,p) \Longleftrightarrow P(M=m) = \binom{m+k-1}{k-1} p^k (1-p)^m \tag{3.6}$$

where m is the number of failures before the k^{th} success, k is the number of successes and p is the probability to observe a success.

The distribution takes its name from the relationship:

$$\binom{m+k-1}{k-1} = (-1)^m \binom{-k}{m} \tag{3.7}$$

which defines the binomial coefficient for negative integers.

The mean and the variance of the negative binomial distribution are given by

$$E[M] = \frac{k(1-p)}{p} \quad \text{and} \quad V[M] = \frac{k(1-p)}{p^2} \tag{3.8}$$

Self-Check Questions

1. We are given a set of observations: $ABABABAB$, where A denotes "success" and B "failure". Do we describe this with the binomial or negative binomial distribution?

2. If we toss a die six times, what is the probability of observing the number three four times?

Solutions

1. We cannot answer this question without further information. We would have to know if the number of trials was fixed to obtain the data, then we would use the binomial distribution. If the number of successes was fixed, we would use the negative binomial distribution.

2. We use the binomial distribution as the number of trials is fixed to $n = 6$. The probability of "success", in our case, the observe the number three is $p = 1/6$ and we want to know the probability of $k = 4$. The probability is $P(k = 4) \cong 0.008$.

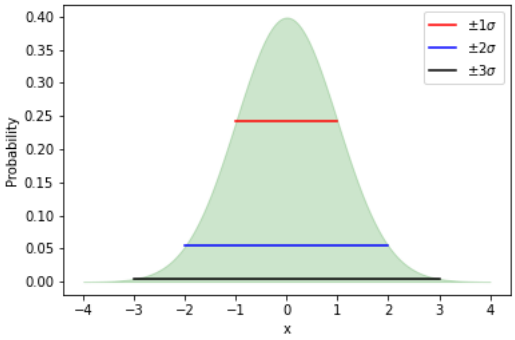


Figure 3.2: Gaussian or normal distribution

3.2 Gauss or Normal Distribution

The Gauss or normal distribution is one of the most important distributions in statistics. It is attributed to Gauss (1877), although, unsurprisingly, the historical developments are more complex. You can find a good overview over the history in, for example, Stahl (2006). The normal distribution appears in all aspects of our everyday life, science and engineering.

A normal random variable X for $x \in (-\infty, \infty)$ has the probability density function:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3.9}$$

The density depends on two parameters μ and σ. The mean or expectation value of the distribution is given by μ and the standard deviation by σ. If we set $\mu = 0$ and $\sigma = 1$, we obtain the "standard" normal distribution whose density function is shown in Fig. 3.2. The normal distribution is often indicated by the notation $\mathcal{N}(\mu, \sigma^2)$.

The reason the normal distribution is of such high importance is because of the central limit theorem (which we provide without proof):

Central Limit Theorem

The probability distribution of a sum $W = \sum_{i=1}^{n} X_i$ of n independent random variables X_i that follow some probability distribution with mean $\langle X \rangle$ and variance σ^2, will converge towards a normal distribu-

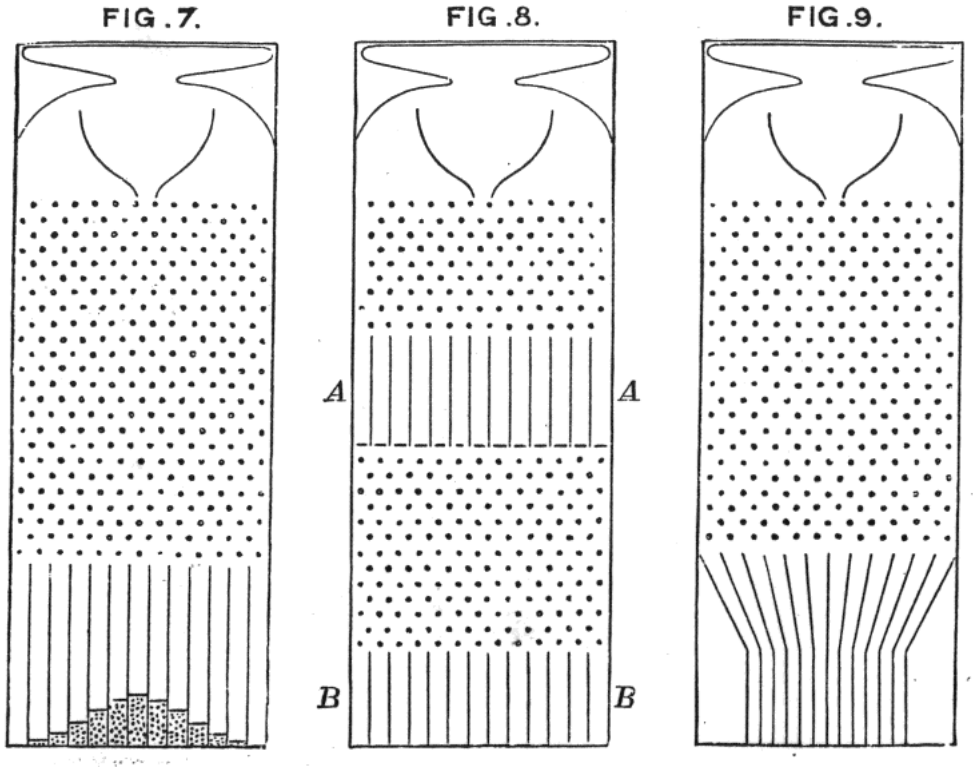


Figure 3.3: Galton board

tion with mean $\langle W \rangle = n\langle X \rangle$ and variance $V[W] = n\sigma^2$.

This means that if we take the sum of the values x_i from n random variables X_i, the sum will approximate a normal distribution. The random variables X_i are in general required to be identically and independently distributed (i.i.d.). A proof is provided in Hogg, McKean, and Craig (2020). Extensions of the central limit theorem exist for cases the X_i do not follow the same distribution but a particular convergence criterion is assumed.

To understand this a bit better, we start with the famous Galton board (Galton, 1894) shown in the left part of Fig. 3.3. The board consists of a reservoir at the top in which we put beans or marbles - this is also why the board is often called a "bean machine". The funnel directs the beans to a board with nails in a regular grid, and the beans are then collected in a set of slots at the bottom. As the beans fall through the nails, they

A Bernoulli trial is an experiment with two distinct outcomes.

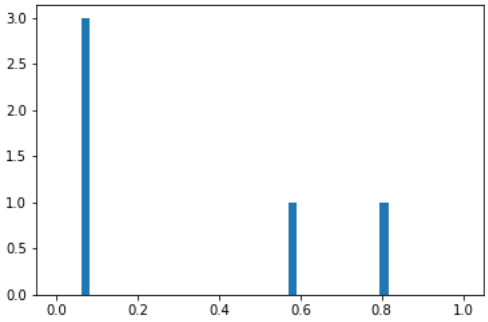


Figure 3.4: Sample of five random numbers following a uniform distribution

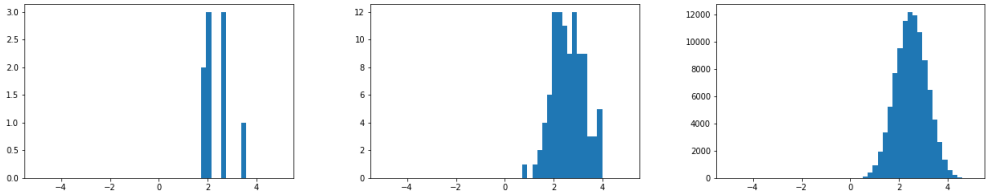


Figure 3.5: Distribution of 10, 100, 100000 sums of 5 random numbers

get deflected either left or right, meaning that at each nail, a **Bernoulli trial** is performed where we can define "success" as the bean goes to the left and "failure" the bean goes to the right. As the bean falls through the lines of nails, this Bernoulli trial is repeated line by line and in the end we observe the sum of the Bernoulli trials for each bean. As we can see from the left part in Fig. 3.3, the shape of the resulting distribution is given by a normal distribution.

We can also illustrate this in another way: Fig. 3.4 shows a sample of five random numbers that are distributed according to a uniform distribution in the interval $(0, 1)$, meaning that the occurrence of each value in this interval is equally likely. As shown in Fig. 3.5, if we repeat calculating this sum, the distribution of this new random variable, the sum of the individual values we obtained from the uniform distribution, approximates a normal distribution.

Because of its enormous importance in practical applications, we typically

assume a normal distribution when we report measured or determined values as well as a dispersion relation. For example, if we say that the average body mass index (BMI) in young men aged 18-30 in Italy is 23.05 ± 2.83 (Krul, Daanen, & Choi, 2010), we understand that the distribution of the BMI follows a normal distribution with $\mu = 23.05$ and $\sigma = 2.83$. An important detail that often gets forgotten in practice is that only $\approx 68\%$ of the distribution are within $\pm 1\sigma$ and 32% are outside. For example, if we were to measure a BMI of 26.38, the value is outside the $\pm 1\sigma$ interval - but only just and we would expect this to happen in 32% of the cases. If we observe less than 32% of the values outside the range indicated by the uncertainties, this should raise suspicion if the results are correctly reported. The percentages outside the first three standard deviations are:

$$\begin{array}{ll} |x-\mu| \geqslant \sigma & 31.74\% \\ |x-\mu| \geqslant 2\sigma & 4.55\% \\ |x-\mu| \geqslant 3\sigma & 0.27\% \end{array}$$

Self-Check Questions

1. What fraction of the events are inside $|x - \mu| \leqslant \sigma$

2. According to the central limit theorem, the sum of random variables will converge to a normal distribution ... of the (single) distribution of the random variable.

Solutions

1. 68%

2. regardless

3.3 Poisson, Gamma-Poisson and Exponential Distribution

If you drop a small handful of rice on a chess board, how many grains will fall on each square? Admittedly, this is not the most obvious question we might ask ourselves in our daily lives, but related questions appear in many different applications. For example, how many lighting bolts will strike the ground in a grid of one square-kilometer areas? How many of a specific product will be sold on any given day in a specific supermarket location? In the second world war, the German bombers were attacking London in the UK during the Blitz and the bombs seem to hit some areas in clusters compared to others - did the German military have special intelligence? In these and similar situations we want to investigate how many items we can count in a given situation such as a supermarket store or a square of land, etc. and how the number of items we count differs from what we would expect if the process was entirely random.

Poisson Distribution

The Poisson distribution describes the count data, i.e. the number of items we observe per defined unit time interval, for a fully random process. The distribution $X \sim P(\lambda)$ is defined as

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{3.10}$$

The Poisson distribution is a discrete distribution and $P(X = k)$ is the probability of observing k events if we observe on average per unit time λ events. For example, if a given product is sold 5 times per day in a specific supermarket on average, $P(X = k) = \frac{\lambda^k}{k!}e^{-\lambda}$ for $k = 0, 1, 2, 3, \ldots$ is the probability of observing the sales of $0, 1, 2, 3, \ldots$ items on any given day.

The Poisson distribution emerges as the limit if we take many trials n while keeping the mean $E[X] = \lambda = np$ fixed.

Some examples of the Poisson distribution for $\lambda = 0.1, 1.5$ and $\lambda = 15$ are shown in Fig. 3.6. Notably, as the parameter λ increases, the Poisson distribution approximates a Gaussian or normal distribution.

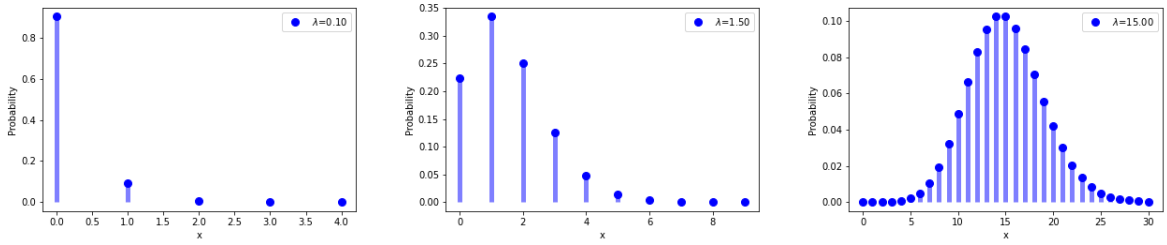


Figure 3.6: Example of a Poisson distribution with $\lambda = 0.1, 1.5$ and $\lambda = 15$.

Proof

Show that the Poisson distribution $X \sim P(\lambda)$ emerges as the limit for $\lim_{n\to\infty} X_n$ with fixed mean $\lambda = np$ of the Binomial distributions $X_n \sim B(n, p)$.

The probability to get $X_n = k$:

$$P(X_n = k) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

We now take the limit $n \to \infty$:

$$\begin{aligned}
&\lim_{n\to\infty} \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\
&\lim_{n\to\infty} \frac{n!}{k!(n-k)!} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} \\
&\lim_{n\to\infty} \frac{n!}{k!(n-k)!} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n} \left(1-\frac{\lambda}{n}\right)^{-k} \\
&\lim_{n\to\infty} \frac{n!}{k!(n-k)!} \frac{\lambda^n}{n^k} \left(1-\frac{\lambda}{n}\right)^{n} \left(1-\frac{\lambda}{n}\right)^{-k} \\
&\lim_{n\to\infty} \underbrace{\frac{n!}{n^k(n-k)!}}_{\approx 1} \frac{\lambda^n}{k!} \underbrace{\left(1-\frac{\lambda}{n}\right)^{n}}_{\approx e^{-\lambda}} \underbrace{\left(1-\frac{\lambda}{n}\right)^{-k}}_{\approx 1} \\
&\approx \frac{\lambda^k e^{-\lambda}}{k!}
\end{aligned}$$

The mean or expectation value of the Poisson distribution is given by λ

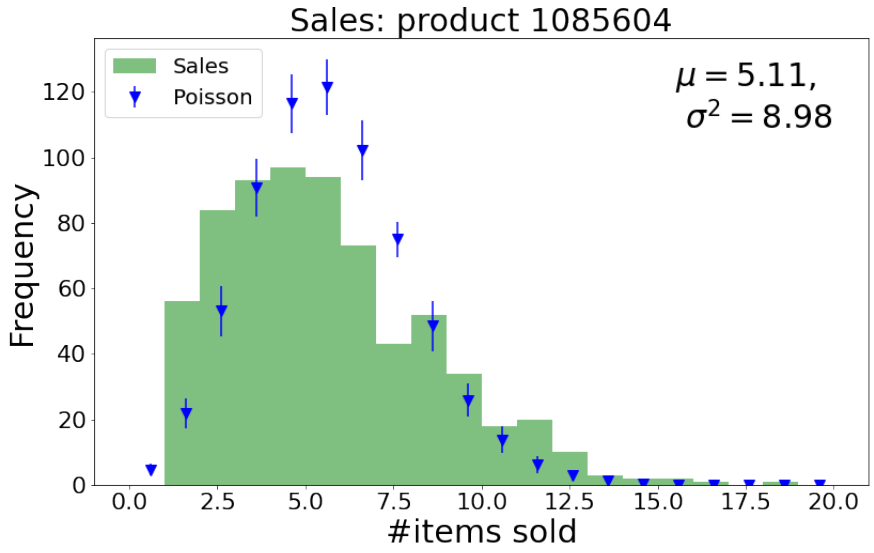


Figure 3.7: Count of sales of a specific supermarket overlaid with a Poisson distribution using data from Venkatesan (2014).

and equal to the variance. This implies that the Poisson distribution is described by only one free parameter, i.e.

$$E[X] = Var[X] = \lambda = \sigma^2$$

A detailed analysis of the bombs on London showed, that they closely follow a random Poisson process (Clarke, 1946), although a recent more detailed study showed that there are more subtle effects to consider (Shaw & Shaw, 2019).

If we look at sales data from a UK supermarket chain (Venkatesan, 2014), we can see in Fig. 3.7 that at least some products sales can be described reasonably well by a Poisson distribution. The data show that the mean is not the same as the variance, hence the assumption of Poisson distribution does not quite hold, however, if we overlay the expected behavior with the observed data, we can see that the description points us in the right direction.

Sum of independent Poisson numbers is a Poisson

Show that the sum of independent numbers distributed according to a Poisson distribution is again distributed according to a Poisson distribution.

We start from two random variables X_1 and X_2 that follow a Poisson distribution with parameters λ_1 and λ_2. We need to show that $Z = X_1 + X_2$ also follows a Poisson distribution. The starting point is

$$P(Z = z) = \sum_{j=0}^{z} P(X_1 = j \;\&\; X_2 = z - j)$$

because $Z = X_1 + X_2$, i.e., we need to choose x_1 and x_2 so that we obtain all ways in which we can can combine the two to obtain the sum. Because we assume that x_1 and x_2 are independent from each other, the joint probability factorizes in to the product of the individual probabilities: $P(X_1 = j \;\&\; X_2 = z - j) = P(X_1 = j)P(X_2 = z - j)$. With this we can write:

$$\begin{aligned} p(z) &= \sum_{j=0}^{z} P(X_1 = j \;\&\; X_2 = z - j) \\ &= \sum_{j=0}^{z} P(X_1 = j)P(X_2 = z - j) \\ &= \sum_{j=0}^{z} \frac{\lambda_1^j e^{-\lambda_1}}{j!} \cdot \frac{\lambda_2^{z-j} e^{-\lambda_2}}{(z-j)!} \\ &= \frac{1}{j!(z-j)!} \lambda_1^j e^{-\lambda_1} \lambda_2^{z-j} e^{-\lambda_2} \end{aligned}$$

In the next step, we multiply numerator and denominator by $z!$. Since we do this both in numerator and denominator, we effectively

multiply by one.

$$\begin{aligned} p(z) &= \sum_{j=0}^{z} \frac{z!}{j!(z-j)!} \frac{\lambda_1^j e^{-\lambda_1} \lambda_2^{z-j} e^{-\lambda_2}}{z!} \\ &= \sum_{j=0}^{z} \frac{z!}{j!(z-j)!} \frac{e^{-(\lambda_1+\lambda_2)} \lambda_1^j \lambda_2^{z-j}}{z!} \\ &= \frac{e^{-\lambda}}{z!} \sum_{j=0}^{z} \frac{z!}{j!(z-j)!} \lambda_1^j \lambda_2^{z-j} \end{aligned}$$

There we have factored out $e^{-(\lambda_1+\lambda_2)} = e^{-\lambda}$ as well as the denominator $z!$. Using the definition of the binomial coefficient

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

we can write the first term inside the sum more conveniently:

$$p(z) = \frac{e^{-\lambda}}{z!} \sum_{j=0}^{z} \binom{z}{j} \lambda_1^j \lambda_2^{z-j}$$

We then make use of the binomial identity:

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k$$

Comparing this to our formula above, we note that

$$\sum_{j=0}^{z} \binom{z}{j} \lambda_1^j \lambda_2^{z-j} = (\lambda_1 + \lambda_2)^z = \lambda^z$$

Putting this back in, we see that:

$$p(z) = \frac{e^{-\lambda} \lambda^z}{z!}$$

Which is again a Poisson distribution with parameter $\lambda = \lambda_1 + \lambda_2$. We can apply the same approach to more numbers that follow a Poisson distribution. The sum $x_1 + x_2 + \ldots + x_n$ will also follow a Poisson distribution with parameter $\lambda = \lambda_1 + \lambda_2 + \ldots \lambda_n$. The larger the parameter λ becomes, the more the resulting Poisson distribution resembles a normal distribution, as we would expect from the central limit theorem.

Gamma-Poisson Mixture Model

The main feature of the Poisson distribution, namely, that the mean is the same as the variance, also is it's limitation: We cannot use the Poisson distribution to describe over-dispersed count data where $\sigma^2 > \lambda$. As already seen in Fig. 3.7, even if we can see that the underlying process is Poisson-like, we cannot use the Poisson distribution to describe it. We therefore need a more general approach to describe discrete count data, an early investigation can be found in Greenwood and Yule (1920). The Poisson distribution has one free parameter λ which defines the mean of the distribution. However, this implies that we know the value of the mean without uncertainties. In practical applications, the mean itself is typically a random variable and therefore needs to be described by a probability distribution which can be done in Bayesian statistics; in this field, the information of a sample and the prior assumption allow to compute the modelled distribution. The Gamma family of distribution is a **conjugate prior** for the Poisson distribution (Gelman et al., 2013, p.44) and we therefore choose the Gamma distribution as prior for the Poisson parameter λ. This means that the random numbers $n_i = P(X = k)$ are given by a Poisson distribution where λ is not a fixed single number but follows a Gamma distribution. This means that for a random variable X:

For a conjugate prior, the prior is of the same family of distributions as the posterior.

$$X \sim \text{Poisson}(\lambda) \quad \text{where} \quad \lambda \sim \text{Gamma}(r, p) \tag{3.11}$$

This is typically parameterized as

$$\Lambda \sim \text{Gamma}\left(r, \frac{p}{1-p}\right) \Longleftrightarrow P(\Lambda < t) = \int_{-\infty}^{t} \frac{(\frac{p}{1-p})^r}{\Gamma(r)} \lambda^{r-1} e^{-\frac{p}{1-p}\lambda} d\lambda \tag{3.12}$$

where r is a form parameter and the rate $\frac{p}{1-p}$ is chosen from the binomial distribution, indicating that this prior describes the the total count of $r-1$ in $\frac{p}{1-p}$ prior observations (Gelman et al., 2013, p.44). We can also express this by the following convolution integral:

$$P(\gamma < c|r, p) = \int_0^{\infty} f_{\text{Poisson}(\lambda)}(c) \cdot f_{\text{Gamma}\left(r, \frac{p}{1-p}\right)}(\lambda) d\lambda \tag{3.13}$$

Gamma-Poisson Mixture Model and Negative Binomial Distribution

To show that the Gamma-Poisson mixture model can be expressed as the negative binomial distribution, we start from the convolution integral in Eqn. (3.13) and insert the definition of the distributions. The density function of the Poisson distribution is given by Eqn. (3.10), and the Gamma distribution by

$$f_{\alpha,\beta}(x) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} \tag{3.14}$$

where α is a shape parameter and β an inverse scale parameter. The Gamma function arises from the extension of the factorial to non-integer numbers via

$$x! = \int_0^\infty y^x e^{-y} dy \tag{3.15}$$

and

$$\Gamma(x+1) = x! \tag{3.16}$$

Inserting this into Eqn. (3.13), using $\alpha = r$ and $\beta = \left(\frac{p}{1-p}\right)^{-1}$ (because of the inverse scale parameter) we obtain

$$P(\gamma < c|r,p) = \int_0^\infty f_{\text{Poisson}(\lambda)}(c) \cdot f_{\text{Gamma}\left(r,\frac{p}{1-p}\right)}(\lambda) d\lambda \tag{3.17}$$

$$= \int_0^\infty \left[\frac{\lambda^k e^{-\lambda}}{k!}\right] \left[\frac{1}{\Gamma(r)\left(\frac{p}{1-p}\right)^r} \lambda^{r-1} e^{-\lambda(1-p)/p}\right] d\lambda \tag{3.18}$$

$$= \frac{(1-p)^r p^{-r}}{k!\Gamma(r)} \int_0^\infty \lambda^{r-1+\lambda} e^{-\lambda/p} d\lambda \tag{3.19}$$

Using the identity

$$\int_0^\infty y^b e^{-ay} dy = \frac{\Gamma(b+1)}{a^{b+1}} \tag{3.20}$$

from Eqn. (3.15) with Eqn. (3.16), we can evaluate the integral

$$\int_0^\infty \lambda^{r-1+\lambda} e^{-\lambda/p} d\lambda = p^{r+k}\Gamma(r+k) \tag{3.21}$$

with Eqn. (3.16), we can then write:

$$P(k|r,p) = \frac{(1-p)^r p^{-r}}{k!\Gamma(r)} p^{r+k}\Gamma(r+k) \tag{3.22}$$

$$= \frac{\Gamma(r+k)}{\Gamma(r)k!} p^k(1-p)^r \tag{3.23}$$

$$= \frac{(r+k-1)!}{(r-1)!k!} p^k(1-p)^r \tag{3.24}$$

$$= \binom{k+r-1}{k} p^k(1-p)^r \tag{3.25}$$

$$= P(N=k) \text{ where } N \sim \mathrm{NB}(r,p) \tag{3.26}$$

This means that we can describe a Poisson-process where we treat the parameter as a random variable following a Gamma function as prior using a negative binomial distribution.

The advantage of the Gamma-Poisson mixture model is that it describes random events that are overdispersed with respect to a Poisson distribution, because the variance can be greater than the mean. Some examples are shown in Fig. 3.8 for the case $\mu = 0.1, \sigma^2 = 0.5$, $\mu = 1.5, \sigma^2 = 3.5$ and $\mu = 15, \sigma^2 = 30$ and Fig. 3.9 shows how the negative binomial is overdispersed with respect to the Poisson distribution for the same mean but $\sigma^2 > \mu$.

We can use the following relationship between the mean and variance of the distribution with the parameters of the negative-binomial distribution:

$$p = \frac{\mu}{\sigma^2} \tag{3.27}$$

$$r = \frac{\mu^2}{\sigma^2 - \mu} \tag{3.28}$$

where we assume $\sigma^2 > \mu$. For the case $\sigma^2 = \mu$ we will use the Poisson distribution.

Returning to the example of the supermarket from above, we can use the Gamma-Poisson mixture model or negative binomial distribution to describe the data as shown in Fig. 3.10. Its use to describe sales data is well established in the theory of Operations Research, see, for example,

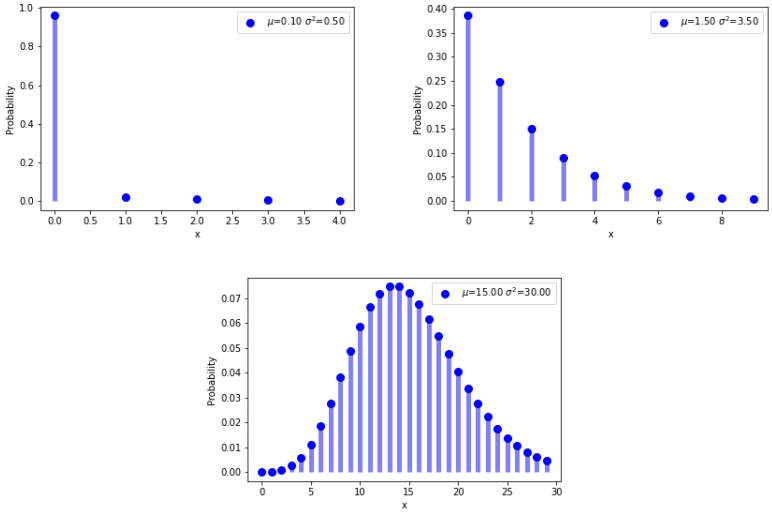


Figure 3.8: Negative binomial distribution with $\mu = 0.1, \sigma^2 = 0.5$, $\mu = 1.5, \sigma^2 = 3.5$ and $\mu = 15, \sigma^2 = 30$.

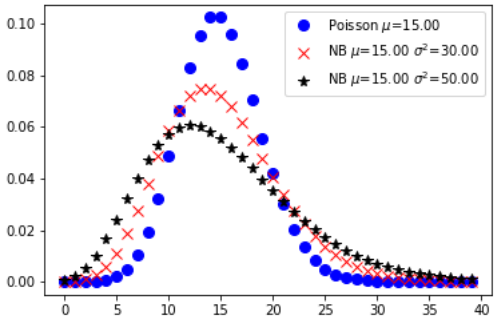


Figure 3.9: Comparison of Poisson and negative binomial distribution for the same mean μ and varying variance $\sigma^2 = 30$ and $\sigma^2 = 50$.

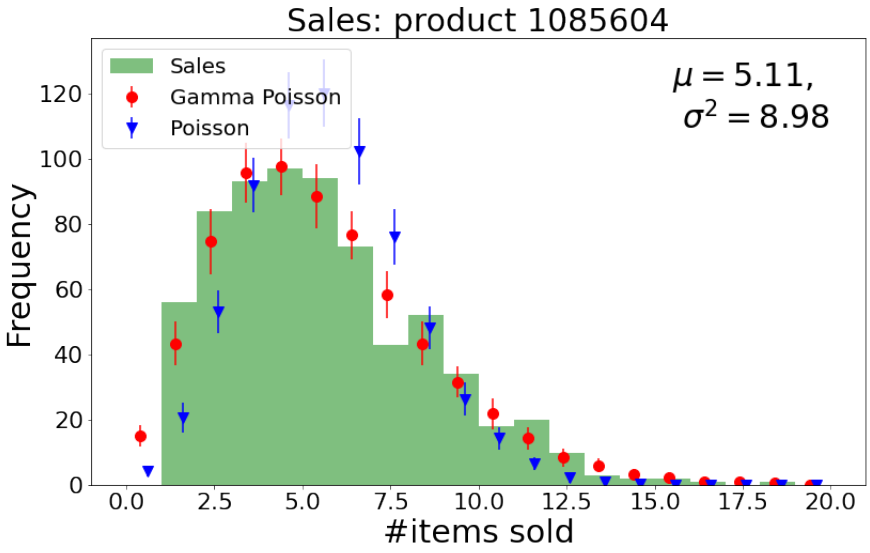


Figure 3.10: Count of sales of a specific supermarket overlaid with a Gamma-Poisson distribution using data from Venkatesan (2014).

A. S. C. Ehrenberg (1959); Goodhardt and Ehrenberg (1967); A. Ehrenberg (1972); Chatfield and Goodhardt (1973); Schmittlein, Bemmaor, and Morrison (1985).

Exponential Distribution

In our discussion so far we focused on describing count data, meaning that we described how many items we can observe in a given unit time. For example, how many of a given product would be sold in a day in a supermarket or how many bolts of lighting we can count per square kilometer per month, etc.

Alternatively, we can ask a related question: How long do we have to wait until we observe the next event? In this case, we want to describe the time between two count events, this is called the "arrival time". We have said earlier, that the Poisson distribution describes random events where we only know how many events happen on average per unit time, i.e., with a

`The rate and mean are connected via` $\mu = \lambda t$.

fixed rate λ. We can rewrite the Poisson distribution for this case as:

$$P(k \text{ events in time } t) = \frac{(\lambda t)^k e^{-\lambda t}}{k!} \tag{3.29}$$

We now want to determine the arrival time until the next event for such a process.

Let Y_n denote the time from the beginning of our measurement, i.e., when we start the clock, until we observe the n-th measurement. $X(t)$ denotes the number of events we observe until time t, i.e. the corresponding count data. The two quantities are related in the following way:

$$Y_n \leqslant 1 \;\Leftrightarrow\; X(t) \geqslant n \tag{3.30}$$

If we denote the distribution of the count data as $C_n(t)$, we can express this using the following relationship:

$$\begin{aligned} C_n(t) &= P(X(t) = n) && (3.31)\\ &= P(X(t) \geqslant n) - P(X(t) \geqslant n+1 && (3.32)\\ &= P(Y_n \leqslant t) - P(Y_{n+1 \leqslant t}) && (3.33) \end{aligned}$$

If $F_n(t)$ is the cumulative probability distribution of $Y_n(t)$, we can express this as:

$$C_n(t) = F_n(t) - F_{n+1}(t) \tag{3.34}$$

If we consider a pure Poisson process, we observe on average $\mu = \lambda$ events per unit time. Alternatively, we can say that we observe the occurrence of events with a fixed rate λ. If we denote the time until some event (for example, the first) is observed with L, we can express the probability that the time until the event occurs is longer than t as:

$$P(L > t) = P(\text{no event until time } t) = \frac{(\lambda t)^0 e^{-\lambda t}}{0!} \tag{3.35}$$

using the definition of the Poisson distribution for a fixed rate λ. Consequently, if $P(L > t) = e^{-\lambda t}$, the distribution for $P(L \leqslant t)$ is given by:

$$P(L \leqslant t) = 1 - e^{-\lambda t} \tag{3.36}$$

As this is the cumulative distribution, we obtain the probability distribution by taking the derivative with respect to t and obtain:

$$f(t) = \lambda e^{-\lambda t} \;\; t > 0 \tag{3.37}$$

This is the exponential distribution and our derivation above has shown that the exponential distribution can describe the time between two events that happen with a fixed rate but are otherwise completely random. Some examples are shown in Fig. 3.11.

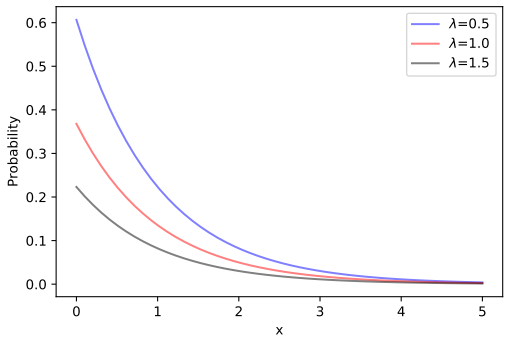


Figure 3.11: Graphs of the density of examples of the exponential distribution for $\lambda = 0.5, 1.0$ and $\lambda = 1.5$

Self-Check Questions

1. Which kind of observable can be described using the Poisson or Gamma-Poisson distribution?

2. The Gamma-Poisson distribution is ... compared to the Poisson distribution.

3. How is the exponential distribution related to the Poisson distribution?

Solutions

1. The Poisson and Gamma-Poisson distribution describe count data, i.e. the distribution of the number of observed events, given the parameters.

2. overdispersed

3. The exponential distribution describes the arrival time between the observed events that are described by a Poisson distribution as their count distribution.

3.4 Weibull Distribution

We have previously encountered the exponential distribution when we discussed the time between two events for a random Poisson process that is described by an average rate. This means that the events occur at some specific and fixed rate per unit time, and the exponential distribution describes when it is most likely that the next event occurs. Given the nature of the exponential distribution, the most likely moment for the next event is always "now", since the exponential distribution is falling monotonously.

However, in many cases this is not what we want - in fact, a lot of effort in engineering is spent making sure that things do not break at random times. For example, imagine a turbine of a plane - for safety reasons we want to make sure that it does not fail "now" but we want to be able to shape the probability distribution for the event "turbine fails" such that it is very, very unlikely that the event occurs before the next maintenance cycle is scheduled. In practice, most maintenance cycles follow a rigid pattern at the moment but the idea of "predictive maintenance" is to estimate the time to the next failure using operational data from a machine using machine learning.

The Weibull distribution (Weibull, 1939) was originally developed in the context of materials science to describe the failure of materials under stress, but has also numerous other applications such as reliability engineering, for example, Yong (2004), failure analysis, Abernethy, Breneman, Medlin, and Reinman (1983); Zhang and Xie (2007); Hall and Strutt (2003) and even the description of wind speeds Mahmood, Resen, and Khamees (2020); Bowden, Barker, Shestopal, and Twidell (1983); Pérez, Sánchez, and García (2007).

The Weibull distribution is given by:

$$P(X = x; \lambda, k) = \begin{cases} \frac{k}{\lambda} \left(\frac{x}{\lambda}\right)^{(k-1)} e^{-(x/\lambda)^k} & x \geqslant 0 \\ 0 & x < 0 \end{cases} \tag{3.38}$$

The parameter $k > 0$ is called the "shape parameter" or **Weibull modulus** and $\lambda > 0$ is called the "scale parameter". In the case of $k = 1$, the Weibull distribution reduces to the exponential distribution. We can therefore discuss three different regimes for the shape parameter k:

In materials science,the Weibull modulus is used to characterize brittle materials.

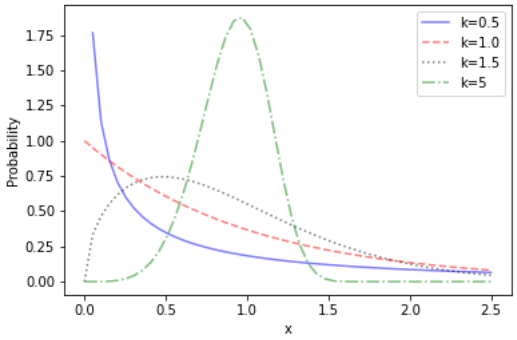


Figure 3.12: Weibull distribution for $k = 0.5$, $k = 1$ and $k = 1.5$ and $k = 5$.

- $k < 1$: The failure rate decreases with time, meaning that most failures occur early on ("infant mortality").
- $k = 1$: The Weibull distribution becomes the exponential distribution. Failures occur randomly with a fixed rate over time.
- $k > 1$: The failure rate increases with time.This means that there are few to no failures early on but then failures occur at later times, for example, due to aging.

These three cases are illustrated in Fig. 3.12. Notably, the distribution becomes more symmetric and localized for values $k >> 1$. This means that, for example, the probability of failure is localized in a specific region. If the x-axis represents time, the probability of failure is very low initially and then rises after some time. If we are to schedule the maintenance for a specific component, we could, for example, define a threshold where the risk of failure is acceptable (this depends on the exact use-case) and then schedule the maintenance once the probability if failure approaches this threshold. In predictive maintenance, we can use observational data from the component to predict the parameters of the Weibull distribution describing this component and then schedule the maintenance according to this prediction.

Self-Check Questions

1. For $k = 1$, the Weibull distribution reduces to the ... distribution, meaning that failure can occur at a fixed ... randomly per unit-time.

2. True or false: For $k >> 1$, most failures occur early on.

Solutions

1. Exponential, rate

2. False.

3.5 Transformed Random Variables

Suppose that you know that a random experiment is determined by a system that yields points in the up-right quadrant (so it can be described by X and Y two random variables both with values in $\mathbb{R}^+$. Let us suppose, for example, that X and Y are independent and follow a $\Gamma(2, 5)$ distribution $X \sim \Gamma(2, 5)$ and $Y \sim \Gamma(2, 5)$. What can be said about the difference $T = X - Y$? The difference is, again, a random variable, i.e. a mapping from the sample space to the real numbers $T : \mathcal{S} \longrightarrow \mathbb{R}$ with any real values. Is it possible to compute the density of T? The answer is yes and it is provided by the following general theorem which requires to reason over multivariate random variables (so, to see random variables as having values with $\mathbb{R}^n$).

Transformed Variables

Recall that a multivariate continuous random variable is a mapping from the sample space $\mathcal{S}$ to the set of vectors in $\mathbb{R}^n$: $X : \mathcal{S} \to \mathbb{R}^n$; let us call D the set of possible values of $X(s)$. In the case of the X and Y both Gamma as above, D is the set of (x, y) with $x > 0$ and $y > 0$.

Suppose that we have a mapping $g : D \to E$ where E is a subset of $\mathbb{R}^n$:

we call the transformed random variable $g \circ X$, often written as $g(X)$, the random variable whose values are $g(X(s))$ for each outcome s in the sample space.

This is the same as saying: suppose that we have a series of continuous random variables X_i $(0 \leqslant i \leqslant n)$ and a series of functions $g_i : \mathbb{R}^n \longrightarrow \mathbb{R}$ we consider the transformed variable $g(X)$ defined by:

$$g(X)(s) = (g_1(X_1(s), \ldots, X_n(s)), \ldots, g_n(X_1(s), \ldots, X_n(s))).$$

In the case of the example above, we have $X_1 = X$ and $X_2 = Y$. We could take, as transformation the mapping $g(x, y) = (x - y, x + y)$ from $D = \mathbb{R}^+ \times \mathbb{R}^+$.

Suppose that g is a differentiable transformation (which means: $g_i(x_1, ..., x_n)$ is differentiable, i.e. $\frac{\partial g_i}{\partial x_j}$ exists for each $0 \leqslant i, j \leqslant n$) and that the inverse $g^{-1} : E \longrightarrow D$ of g exists. The transformation theorem, proved in Hogg et al. (2020, 2.7), says that:

- $g(X)$ It is a continuous random variable.
- if X has a probability density function (pdf) $f_X : D \to \mathbb{R}$ then the pdf of $g(X)$, noted $f_{g(X)}$, is equal to the following, $\forall \mathbf{e} \in E$:

$$f_{g(x)}(\mathbf{e}) = f_X(g^{-1}(\mathbf{e})) \cdot |J_{g^{-1}}(\mathbf{e})| \tag{3.39}$$

Where $J_{g^{-1}}(\mathbf{e})$ is the Jacobian of the transformation g^{-1} and is non-zero at least in on $\mathbf{e}$: The Jacobian is the determinant of the matrix of each partial derivative of g^{-1}:

$$J_{g^{-1}}(\mathbf{e}) = \begin{vmatrix} \frac{\partial g_1}{\partial e_1}(\mathbf{e}) & \frac{\partial g_1}{\partial e_2}(\mathbf{e}) & \cdots & \frac{\partial g_1}{\partial e_n}(\mathbf{e}) \\ \frac{\partial g_2}{\partial e_1}(\mathbf{e}) & \frac{\partial g_2}{\partial e_2}(\mathbf{e}) & \cdots & \frac{\partial g_2}{\partial e_n}(\mathbf{e}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_n}{\partial e_1}(\mathbf{e}) & \frac{\partial g_n}{\partial e_2}(\mathbf{e}) & \cdots & \frac{\partial g_n}{\partial e_n}(\mathbf{e}) \end{vmatrix}$$

Said simply, the transformation theorem allows us to calculate the probability density function of a transformed random variable by combining the density with the inverse of the transformation and multiplying by the Jacobian of the inverse of the transformation. If X has a single dimension (and thus g can be written as $g(x) = u$:

$$f_{g(X)}(u) = f_X(g^{-1}(u)) \cdot |g^{-1\prime}(u)|$$

It is important to note that D (the value-set of X and source set of the transformation) and E (the value-set of the transformation) are rarely equal to complete $\mathbb{R}^n$. E.g. It could be a subset of a plane where the lowest x depends on y as we shall see in the example.

Example: Difference of two Gamma-variables

Coming back to our example random variable $T(s) = X(s) - Y(s)$ with $X \sim \Gamma(2, 5)$ and $Y \sim \Gamma(2, 5)$ independent of each other. We consider the random variable couple $C(s) = (X(s), Y(s))$ which gives a point for each possible random outcome (the points are in the upper-half quadrant $\{(x, y)|x > 0 \text{ and } y > 0\}$).
To calculate the density of the $T = X - Y$, we can simply use a transformation where $x - y$ is one of the components. For example, we could choose the following mapping:

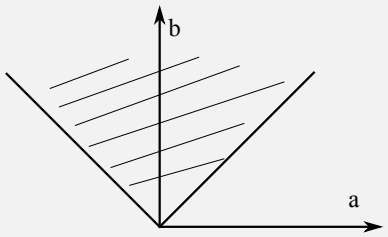
$$\begin{array}{rccl} g: & \mathbb{R}^+ \times \mathbb{R}^+ & \longrightarrow & E \\ & (x, y) & \mapsto & (a, b) = g(x, y) = (x - y, x + y) \end{array}$$

What is the destination subset $E \subset \mathbb{R} \times \mathbb{R}$ of the transformation g? It is the set of $(a, b) = g(x, y)$. That is, it is the set of (a, b) such that we can find $(x, y) \in D$ with $x - y = a$ and $x + y = b$. We can calculate this set by solving the equation to obtain x and y from a and b:

$$\left\{ \begin{array}{l} a = x - y \\ b = x + y \end{array} \right. \overset{\text{add first line}}{\underset{\text{to second}}{\Longleftrightarrow}} \left\{ \begin{array}{l} a = x - y \\ a + b = 2x \end{array} \right.$$

$$\begin{array}{c}\text{express x from}\\ \text{second line}\\ \Longleftrightarrow\\ \text{use expression of x}\end{array}\quad\begin{cases} x = \frac{1}{2}(a+b) \\ y = \frac{1}{2}(a+b) - a \end{cases}\quad\begin{array}{c}\text{simplify}\\ \Longleftrightarrow\\ \cdot\cdot\end{array}\quad\begin{cases} y = \frac{1}{2}(b-a) \\ x = \frac{1}{2}(a+b) \end{cases} \tag{3.40}$$

Thus, E is the set of pairs $(a, b) \in \mathbb{R} \times \mathbb{R}$ such that $b - a > 0$ and $b + a > 0$ which is the hatched region in the figure below.



The equation solution (Eqn 3.40) let us calculate the inverse transformation g^{-1}:

$$(x, y) = g^{-1}(a, b) = \left(\frac{1}{2}(a+b), \frac{1}{2}(b-a)\right)$$

We now apply the transformation theorem's formula 3.39 with the transformation g so as to get the density of $g(C)$. This will give us the joint density of $(X - Y, X + Y)$ from which we shall be able to deduce the density of $X - Y$ as intended.
To apply the theorem, we need to calculate the Jacobian of g^{-1}:

$$J_{g^{-1}}(a, b) = \begin{vmatrix} \frac{\partial(\frac{1}{2}(b-a))}{\partial a} & \frac{\partial(\frac{1}{2}(b-a))}{\partial b} \\ \frac{\partial(\frac{1}{2}(b+a))}{\partial a} & \frac{\partial(\frac{1}{2}(b+a))}{\partial b} \end{vmatrix} = \frac{1}{2} \cdot \begin{vmatrix} -1 & 1 \\ 1 & 1 \end{vmatrix} = \frac{1}{2} \cdot (-2) = -1$$

Let us remind that the density of the X and Y (both $\sim \Gamma(2, 5)$) is given by the function $f(t) = \frac{25}{1} \cdot t^{2-1} \cdot e^{-5t} = 25te^{-5t}$ and they are independent thus the joint density of (X, Y) is:

$$f_{(X,Y)}(x, y) = 25xe^{-5x} \cdot 25ye^{-5y} = 625 \cdot x \cdot y \cdot e^{-5(x+y)}$$

The theorem thus gives us the joint density of $g(X,Y)$ is the following for $(a,b) \in E$ and is 0 otherwise:

$$f_{g(X,Y)}(a,b) = f_{(X,Y)}(g^{-1}(a,b)) \cdot |J_{g^{-1}}(a,b)| =$$

$$= \frac{625}{4} \cdot (b-a) \cdot (a+b) \cdot e^{-5 \cdot 2 \cdot b} = \frac{625}{4}(b^2 - a^2)e^{-10b}$$

To calculate the density of $X - Y$, we simply have to calculate the density of the first component a of $g(X,Y)$ since $g(x,y) = (x-y, x+y)$. This can be done by calculating the marginal density of $g(X,Y)$:

$$f_{X-Y}(a) = \int_{-\infty}^{\infty} f_{g(X,Y)}(a,b)db = \int_{b=|a|}^{\infty} \frac{625}{4}(b^2 - a^2)e^{-10b}db$$

as $f_{g(X,Y)}(a,b)$ is 0 outside of E. Thus, as calculated by Wolfram Alpha:

$$f_{X-Y}(a) = \frac{5}{2}(10|a| + 1)e^{-10|a|}$$

Self-Check Questions

1. True or false: The probability density function of $X + Y$ is $f_X + f_Y$ if X and Y are two continuous random variable. Justify your answer.

2. True or false: if X is a continuous real random variable with probability density function f_X then $666 \cdot X$ is a random variable with density function $f_{666 \cdot X}(u) = f_X(\frac{1}{666} \cdot u)$. Justify your answer.

Solutions

1. False. Among other justifications: $f_X + f_Y$ is not even a probability density function. Extra knowledge: There is, however, a corollary of the transformation theorem which states that the density function of the sum of two continuous random variables is the convolution of the density functions.

2. True as the transformation is, then, $g : x \mapsto u = 666 \cdot x$ thus the Jacobian of the inverse transformation is $\frac{1}{666}$.

Summary

In many processes, specific probability distributions play a major role to describe the observed data. We therefore say that the observed data are concrete realizations of a random variable following a specific probability distribution.
The binomial and negative binomial distributions are discrete probability distributions that describe the number of successes and failures in repeated trials with two outcomes, such as, e.g., tossing a coin.
The normal or Gaussian distribution is one of the most important probability distributions due to the central limit theorem that states that the sum of random numbers that follow any probability distribution converge to a normal or Gaussian distribution. The normal distribution is often denoted as $\mathcal{N}(\mu, \sigma)$ and is described by two parameters, the mean μ and the standard deviation σ.
The discrete Poisson distribution describes the observed count data of random events that occur with a fixed rate and is characterized by one parameter describing the rate or mean. The variance of the Poisson distribution is equal to the mean. This distribution is linked to the exponential distribution that describes the arrival time between Poisson events. The Gamma-Poisson distribution describes count data that are overdispersed with respect to a pure Poisson distribution where the variance is greater than the mean.
The Weibull distribution plays a major role in particular in describing the failure of materials or components. The value of its shape parameter k determines if failure occurs mainly early on, randomly with a fixed rate or is more localized at a specific time interval.

4 Bayesian Statistics

Study Goals

After completing this unit you will have learned:

- What the Bayes' formula is.
- How to calculate the *a posteriori* probability of events in an Bayesian approach.
- How to choose priors.
- How to understand the differences between Frequentist and Bayesian statistics.

Introduction

Intuitively, we have a good grasp on probabilities. For example, if we toss a coin many times and count the number of head or tails, we will observe that both occur with the same frequency. If we roll a die, we determine that each number comes up 1/6th of the time and we would agree with the statement "the probability for each number in a fair die is 1/6".

Or do we? Kurt (2019) gives a range of intuitive examples where we assume a specific behavior or outcome because it is consistent with what we observe in our world. For example, if we approach a traffic intersection and the traffic light in our direction shows "green", we assume that the traffic light in the other direction crossing our path will show "red" and we can safely cross the road. We do not stop to check if the other traffic

light does indeed show a red stop signal, we assume that this is the case - because based on all our experience, traffic lights are made such that they only show "green" in one direction and stopping the other, thus preventing accidents. In other words, we hold a prior belief that if our traffic light shows "green", we can safely proceed as the others will wait.

The famous quote "When you hear hoof beats, think of horses, not zebras" is attributed to Th. Woodward (see, for example, Dickinson (2016)) who coined the term in the 1940s to teach medical students to look for common causes of an illness first before starting a detailed investigation in possible, but unlikely causes. Again, incorporating prior knowledge plays a major role in decision making and, importantly, directly affects the outcome and the decisions we take.

This way of using and interpreting probabilities is called "Bayesian statistics", named after Rvd. Thomas Bayes (1701 - 1761) (Bayes, 1763).

4.1 Bayes' Rule

Bayesian statistics formalizes the way we can express the probability of events incorporating data as well as prior beliefs or knowledge.

Central to Bayesian analysis is the definition of conditional probabilities:

$$P(B|A) = \frac{P(B \wedge A)}{P(A)} \tag{4.1}$$

where we read $P(B|A)$ as "$P(B$ given $A)$". This conditional probability describes the probability to observe event B when event A has already occurred. This conditional probability is the probability of observing both A and B, divided by the probability of observing A. We could have, of course, also started from the probability to observe event A, given we have observed event B, i.e., $P(A|B)$. Bayes' theorem connects these two probabilities:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \tag{4.2}$$

The key point of Bayes' formula is that $P(B|A)$ is not the same as $P(A|B)$ but the two are connected with the respective probability of observing event A and B.

Derivation of Bayes' Formula

We start from the definition of the conditional probabilities:

$$P(B|A) = \frac{P(B \wedge A)}{P(A)} \quad \text{and} \quad P(A|B) = \frac{P(A \wedge B)}{P(B)}$$

Noting that $P(A \wedge B) = P(B \wedge A)$ because here we observe the events A and B at the same time, we can substitute $P(B \wedge A)$ in the formula we start from and obtain:

$$\begin{aligned} P(B|A) &= \frac{P(B \wedge A)}{P(A)} \\ &= \frac{P(A|B)\,P(B)}{P(A)} \end{aligned}$$

In practice, Bayes' formula is essential if we want to express how well our beliefs or hypotheses are matched with the data. We use H to denote our hypothesis and D (or sometimes E) the data or "evidence". Then Bayes' theorem becomes

$$P(H|D) = \frac{P(D|H)\,P(H)}{P(D)} \tag{4.3}$$

The components have the following distinct meanings:

- $P(H|D)$ is the *a posteriori* probability. This is the quantity we want to know: What is the probability of our prior belief or hypothesis after we look at the data?

- $P(H)$ is called the "prior" and reflects our prior knowledge or hypothesis before we look at the data.

- $P(D|H)$ is called the "likelihood". For a given hypothesis, this quantity describes the probability of observing the data D we have recorded.

- $P(D)$ is called the evidence. In a way, this is a normalization constant and incorporates the probability to observe the data, i.e., it includes all ways we can observe the data. Note that the evidence does not depend on the hypothesis H and is therefore the same for all hypotheses we might want to test.

Therefore, in short, the posterior (the quantity we want to know) is given by the likelihood times the prior, normalized to the evidence.

We can get a better intuitive understanding of how the Bayes' formula works with this simple example. A further nice example using Lego bricks is given in Kurt (2019, p. 66ff).

Example: House ablaze in flames

What we want to know is the probability that if we observe smoke, a house is ablaze and we need to call the firefighters quickly to prevent damage. To apply Bayes' theorem, we need to know the other components: P(Smoke) is the probability to observe smoke. If we use our everyday experience, we would maybe conclude that observing smoke is not very common but not rare either: In the summer, there are often barbeques, or people have a fireplace in their home or garden, etc. Let us say, the probability of observing smoke is 20%. We also need to define the likelihood, i.e., given that there is a fire, what is the probability of observing smoke: P(Smoke|Fire). Most fires will produce smoke, especially if we burn wood or coal. However, some fires do not, for example, if we burn hydrogen and oxygen or a very clean gas also does not produce much smoke. We could say that the likelihood is 90%: Most, but not all fires produce smoke. Finally, we need to specify the prior, in our case, the probability that a house is ablaze P(Fire). In most of our experiences, we will not have encountered a house ablaze (unless we are fire fighters) - but we do know that this happens, since our town does have a fire department. Say, the probability is maybe P(Fire)=0.1%. Putting this all together we have:

$$\begin{aligned} P(\text{Fire}|\text{Smoke}) &= \frac{P(\text{Smoke}|\text{Fire})P(\text{Fire})}{P(\text{Smoke})} \\ &= \frac{0.1\%\,90\%}{20\%} \\ &= 0.45\% \end{aligned}$$

Hence, if we observe smoke, we would say that there is less than half of a percent chance that this is due to a house burning down.

In this simple example, we have just taken a specific value for the evidence. In many cases, we do not have have access to this quantity directly, but we need to calculate it. To do so, we can decompose the evidence using the total law of probabilities:

$$P(A) = \sum_i P(A|B_i)P(B_i) \tag{4.4}$$

and express the Bayes' theorem as:

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_i P(B|A_i)P(A_i)} \tag{4.5}$$

For example, if our single hypothesis H can be either "true" or "false", we can write $P(D) = P(D|H)P(H) + P(D|\neg H)P(\neg H)$.

Self-Check Questions

1. What does the prior in Bayesian statistics describe?

2. What is the *a posteriori* probability?

Solutions

1. The prior reflects our knowledge of the system we want to describe before we look at the data.

2. The *a posteriori* is the quantity we want to know: given the observed data, how likely is or hypothesis?

4.2 Estimating the Prior, Benford's Law and Jeffrey's Rule

The Role of the Prior

When defining the Bayes' rule, we have already encountered the prior. In the previous simple example, we just "made up" the numbers to see how

the formula works.

Now we investigate the important rule the prior plays in more detail. We start with an example:

Example: AIDS Test

Imagine an AIDS test is performed on a male patient who does not belong to high risk groups. The test is very reliable: If the subject has AIDS, the test will be positive with 99.9% certainty. The test also has a low false-positive rate of 0.5%, i.e., in 0.5% of the cases, the test is positive, even though the subject does not have AIDS. What is the probability of the patient to have AIDS if the test returns positive?

We first translate the probabilities regarding the test into the language of Bayesian statistics:

- $P(\text{pos}|\text{AIDS}) = 0.999$, from the accuracy of the test.
- $P(\text{neg}|\text{AIDS}) = 0.001$, because the probabilities need to be normalized: if the test is 99.9% accurate, there is a 0.1% probability to go wrong.
- $P(\text{pos}|\neg\text{AIDS}) = 0.05$, the false-positive rate of the test.

However, to apply Bayes' formula, we need to know the prior, this is one of the main challenges in Bayesian statistics. For this example, we can base the further evaluation on Germany and obtain the number from the epidemiological report of the Robert Koch Institute, the German authority for infectious diseases. For other countries, we would of course have to refer to similar resources in the respective country. According to Marcus, Gunsenheimer-Bartmeyer, Kollan, and Bremer (2019), including data up to the end of 2018, there are 76,600 men affected. We subtract the main risk factors listed in the table: sex between men (54,200 cases) and i.v. drug abuse (8,200 cases) and estimate that there are then 76,600-54,200-8,200=14,200 cases in Germany. Using the official German census (Bundesamt, n.d.-b) from 2011, adjusted for the progression until the end of 2018 (Bundesamt, n.d.-a), there were 83,019 people in Germany, of which

40,966 were male. Hence our prior is P(AIDS) = 14,200/40,966,000 = 0.00035.
We then need to calculate the probability of the evidence $P(D)$ for our normalization: $P(D) = P(D|H)P(H) + P(D|\neg H)P(\neg H) = 0.999 \cdot 0.00035 + 0.005 \cdot (1 - 0.00035) = 0.05$.
Putting this all together, we obtain:

$$\begin{aligned} P(H|D) =& \frac{P(D|H)P(H)}{P(D|H)P(H) + P(D|\neg H)P(\neg H)} \\ =& \frac{0.999 \cdot 0.00035}{0.05} \\ =& 0.069 \\ \approx& 7\% \end{aligned}$$

This means that even though the very accurate and reliable test is positive, the *a posteriori* probability, the quantity that we want to know and on which we base our actions and decisions, is only around 7%.

The example above shows the important role of the prior: Even with a highly accurate result, the quantity we want to know, the *a posteriori* probability, is quite low. Similar tests have been made with physicians and medical professionals, although not many, sadly, will get the answer right, see, for example, Casscells, Schoenberger, and Graboys (1978); Eddy (1982); Gigerenzer and Hoffrage (1995).

We have so far assumed that the male patient does not belong to a risk group - what happens if he does? Then we need to account for this and, depending on exactly how we define this, use a different number. For simplicity, we just take all affected men. Then the prior is P(AIDS) = 70,600/40,966,000 = 0.0017, which is almost a factor 10 higher. The evidence is then $P(D) = P(D|H)P(H) + P(D|\neg H)P(\neg H) = 0.999 \cdot 0.0017 + 0.005 \cdot (1 - 0.0017) = 0.0067$, and the posterior $P(H|D) = 0.25$ or 25%. Note that the data, namely, the result of the test, is the same in both cases, yet the result is very different. Which result is correct? The key point is that both results are correct - but both depend on the assumptions and the data we have for the prior. The advantage of this approach is that we need to define the prior and by doing so we explicitly state the assumptions or conditions under which our result is valid. However, in practice this is

easier said than done because defining the prior can be very difficult. In our previous example, we made the prior explicit by looking at the report from the official state authority and then place the subject in one or the other category. However, what if we cannot do this, for example, the patient is unconscious and we cannot ask him or his relatives? Or if there is no report, we can look up?

Benford's Law

Obtaining the prior is one of the most difficult aspects in Bayesian analysis. It is often suggested that this could be avoided if we used a uniform or flat prior, meaning that we use a uniform distribution where all values of the prior are equally likely. In fact, this is one of the most common mistakes. At first glance, the argument seems to make sense: If we do not want our results to be biased by the wrong choice of prior, why not choose one that does not favour any particular setting?

However, if we look at many systems we find that the distribution of numbers is not uniform. This was first noted by Newcomb (1881) working with logarithm tables for his calculations before the invention of pocket calculators or computers, where he noticed that the pages containing logarithms that start with numbers one or two are more worn out than the others. He described this with the empirical formula:

$$P(d) = \log_{10}\left(1 + \frac{1}{d}\right) \tag{4.6}$$

where $d = 1, 2, \ldots, 9$ is the leading digit and $P(d)$ is the probability to observe this digit. Later, Benford (1938) revisited this and did a systematic study across many thousand observations from various sources and found that they also follow this distribution. Even though it was originally proposed by Newcomb, Eqn. (4.6) is commonly known as "Benford's law".

For example, if we look at the first digit of all physical constants as shown in Fig. 4.1, we find that the distribution of digits is well approximated by Benford's law. The figure was created using the software "Benford for Python" (Milcent, 2014). This relationship is also found in many other applications, for example, in the discovering of financial fraud, see, for example, Tam Cho and Gaines (2007); Asllani and Naco (2015); Barabesi,

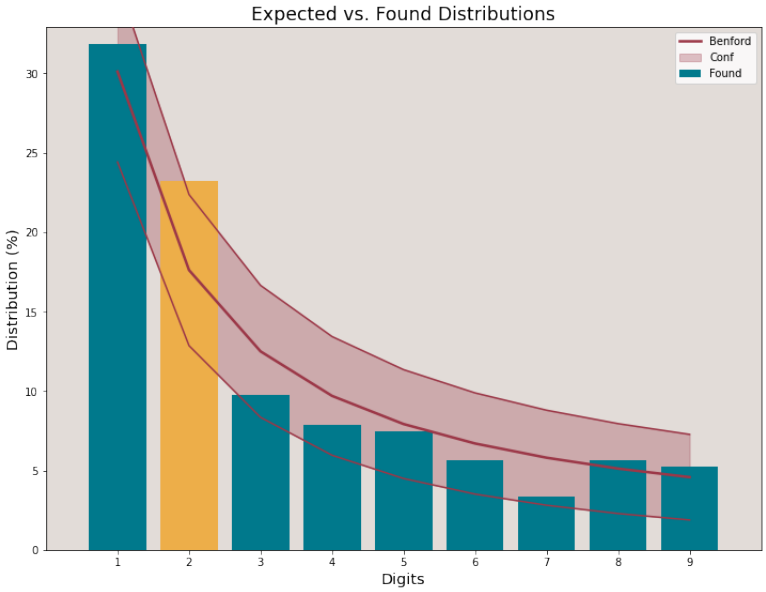


Figure 4.1: Benford's law shown using a wide range of the leading digit from physical constants including a 95% confidence interval.

Cerasa, Cerioli, and Perrotta (2018); Diekmann and Jann (2010), phenomena in physics, such as Pain (2008); Pietronero, Tosatti, Tosatti, and Vespignani (2001); Buck, Merchant, and Perez (1993), and even the answers to textbook exercises(Slepkov, Ironside, & DiBattista, 2015). A list of many publications about this topic can for example be found in Berger, Hill, and Rogers (2009).

The key point in understanding the emergence of this phenomenon is that we combine many different numbers from many different sources. For example, if we look at a long list of physical constants, they describe a wide range of phenomena and since these phenomena are very different to each other, the range of values the constants can take is also very different, spanning many orders of magnitude. This means that if these numbers indeed follow Benford's law, as we can observe empirically, this law must be scale invariant. After all, there is no "right" scale of units that would describe everything from sub-atomic to cosmic phenomena.

Intuitively, we would expect that each leading digit 1,2,..., 9 occurs with the same probability or frequency. However, if we demand that this is scale-invariant, this must also hold if we multiply each constant with a common

factor indicating a new scale or set of measurement units. As an example, we can choose a factor two - say, instead of measuring everything in metres, we measure everything in units of 50 cm, so what was "1 meter" before is now "2 new-meters", because the physical distance does not change, whether or not we describe it in meters or "new-meters". Then, if our original number started with one, it now starts with two after we change to the new scale, i.e., after we multiply by two. If we start with two, we obtain four and so on. However, if our original number starts with a digit between five and nine, the new leading digit will be one as we cross the mark "ten". Therefore, it is more likely that we will observe a "one" as the leading digit than any other number.

A detailed proof and derivation can be found in Hill (1995) with a more detailed exposition in Berger and Hill (2015), Jamain (2001) gives a good overview and more recent further investigations can be found in, for example, Whyman, Shulzinger, and Bormashenko (2016); Ryder (2009).

Jeffrey's Rule

In many cases we do not know much about the system we want to analyze using Bayesian statistics. In these cases we would like to avoid specifying the prior, although Bayes' formula demands that we do so. Intuitively, using a uniform prior seems the most natural candidate - but as we have just seen, when discussing Benford's law, many numbers are not uniformly distributed.

We can also illustrate this in a different way: What we want to achieve if we were to use a uniform distribution as a prior is to express that we do not know the value the parameter should take and we do not want to impose any constraints. However, consider that a parameter of family of random variables θ is assumed to follow a uniform distribution in the interval $(0, 1)$ and we hope that this expresses that we do not know anything about θ. However, when setting up our model, there is no "single best" parametrization. We could, for example, also have expressed θ in terms of the logit function and apply the transformation:

$$\theta' = \log\left(\frac{\theta}{1-\theta}\right) \tag{4.7}$$

where θ' is now in the interval $(-\infty, \infty)$. Now, the transformed parameter

θ' is no longer uniformly distributed and our seemingly non-informative uniform prior has become the opposite. More formally, we can see this when applying the transformation rule for probability distributions. We start from a probability distribution expressed in the parameter θ. We then transform this using a transformation g to a new variable ϕ, where $\phi = g(\theta)$, i.e., the function $g(\cdot)$ transforms the parameter θ into ϕ. Remember that the transformation rule is defined as:

$$f(\phi) = f(g^{-1}(\phi)) \left| \frac{dg^{-1}(\phi)}{d\phi} \right| \tag{4.8}$$

In this case, $f(g^{-1}(\phi)) = f(\theta) = \text{const.}$, because we have assumed that θ follows a uniform distribution. The second term is only constant if $g(\cdot)$ is a linear function - but not in general. However, this is contradicting our previous assumption: We wanted to express that we have little to no knowledge about the prior and applying a transformation should not make a difference because there is no *a priori* right way to parametrize the parameter θ. Therefore, using a uniform distribution as a prior is not a suitable choice.

The Jeffrey's prior (Jeffreys, 1946) defines a prior for a parametrized random variable that is invariant under transformations and is defined by the probability density function:

$$f(\theta) \propto \sqrt{J(\theta)} \tag{4.9}$$

where $J(\theta)$ is the expected Fisher information of the parameter θ for the parametrized random variable. The symbol "$\propto$" means "proportional to". This means that there is a constant c such that $f(\theta) = c \cdot \sqrt{J(\theta)}$. Furthermore, $f(\theta)$ is unique because the integral over all real numbers is 1.

The **Fisher information** measures the amount of information about the parameters θ and is given by the negative of the second derivative of the log-likelihood function:

$$I(\theta) = -\frac{d^2 \text{LogL}(\theta)}{d\theta^2} \tag{4.10}$$

The Fisher information is a metric that measures the amount of information that a random variable X contains about the parameter θ given an observed sample $x_1, \ldots, x_n$.

(Log) Likelihood Function

The likelihood function $\mathcal{L}$ measures, essentially, the probability—or likelihood, hence the name—of observing the current data given a specific model that depends on one or more parameters θ:

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

Here, x is the concrete realization of a random variable or a vector of random variables X, i.e., the values we observe or measure. Crucially, we assume that we know the density of the underlying probability distribution $f_X(.|\theta)$ except the values of the parameters θ. For example, we may know or assume that the values x of the random variable X are distributed according to a Gaussian distribution: $X \sim \mathcal{N}$ but we do not know the values $\theta = (\mu, \sigma)$ of the parameters. In the simplest case we have only one probability distribution for all realizations and each realization is independent from the others. In this case the likelihood function is given by the product of the individual factors associated with each observed value, i.e.

$$\mathcal{L}(\theta) = f_X(x|\theta) = \prod_{i=1}^{n} f_X(x_i|\theta)$$

where $i = 1, \ldots, n$ is the index that identifies the individual n observations.
For practical reasons, we often use the log-likelihood function given by $\mathrm{LogL} = \ln(\mathcal{L})$.
For further details see, for example, Held (2008, chap. 2)

The first derivative of the log-likelihood function is also called the "score function" $S(\theta)$:

$$S(\theta) = \frac{d\mathrm{LogL}(\theta)}{d\theta} \tag{4.11}$$

The Fisher information can then be written as:

$$I(\theta) = -\frac{d^2\mathrm{LogL}(\theta)}{d\theta^2} = -\frac{dS(\theta)}{d\theta} \tag{4.12}$$

The expected Fisher information is then the expectation value of $I(\theta)$, i.e.,

$$J(\theta) = E[I(\theta)] \tag{4.13}$$

Under the assumption that we can change the order of differentiation and integration (regularization assumption), we can show that (Held, 2008, p. 66):

$$E[S(\theta]) = 0 \tag{4.14}$$
$$Var[S(\theta)] = E[S(\theta)^2] = J(\theta) \tag{4.15}$$

For further information see also Gelman et al. (2013, p. 52ff) or Liu and Abeyratne (2019, App. 4).

Jeffrey's Prior is invariant under bijective transformations

Show that the Jeffrey prior is invariant under bijective transformations.

We define the Jeffrey's prior for the parameter θ as: $f(\theta) \propto \sqrt{J(\theta)}$ according to Eqn. (4.9). Then we use the rule for the transformation of probability distributions in Eqn. (4.8):

$$\begin{aligned}
f(\phi) &\propto f(g^{-1}(\phi)) \left| \frac{dg^{-1}(\phi)}{d\phi} \right| \\
&\propto f(\theta) \left| \frac{dg^{-1}(\phi)}{d\phi} \right| \quad \text{with } f(\theta) = f(g^{-1}(\phi)) \\
&\propto \sqrt{J(\theta)} \left| \frac{dg^{-1}(\phi)}{d\phi} \right| \quad \text{with } f(\theta) \propto \sqrt{J(\theta)} \\
&= \sqrt{J(\theta) \left| \frac{dg^{-1}(\phi)}{d\phi} \right|^2} \\
&= \sqrt{J(\phi)}
\end{aligned}$$

Hence, if we express the prior $f(\theta)$ according to Jeffrey's rule and then transform $\theta \rightarrow \phi$, the resulting prior using the transformed variable also follows Jeffrey's rule (Held, 2008, p. 152).

Example: Jeffrey prior for Poisson distribution

Calculate the Jeffrey prior for the Poisson family of distributions given a sample x_1, ..., x_n. The Poisson family of distribution

means that we refer to all possible values of the parameter in the parametrization of the Poisson density.

A variable X that follows a Poisson distribution is given by:

$$P(X = k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{4.16}$$

for integer $k = 0, 1, 2, \ldots$. The rate parameter λ defines the average rate of events. When one says that X follows a Poisson distribution, it means that its probability mass function is as in the equation (Eqn. 4.16). If λ is unknown, X is one of the members of a family of Poisson distributions; the family is parametrized by λ and finding an appropriate λ is going to be an objective.
The score function is according to Eqn. (4.11) given by:

$$\begin{aligned} S(\lambda) &= \frac{d}{d\lambda} \ln P(X = k|\lambda) \\ &= \frac{d}{d\lambda} \left[\ln \left(\frac{\lambda^k e^{-\lambda}}{k!} \right) \right] \\ &= \frac{d}{d\lambda} \left[k \ln(\lambda) - \ln(k!) - \lambda \right] \\ &= \frac{k}{\lambda} - 1 \\ &= \frac{k - \lambda}{\lambda} \end{aligned}$$

where we have also used $\ln(x^b) = b \ln(x)$. The expected Fisher information is given by Eqn. (4.13), and using Eqn. (4.15), we can write:

$$\begin{aligned} J(\lambda) &= E[I(\lambda)] \\ &= Var[S(\lambda)] \\ &= E[S(\lambda)^2] \\ &= E\left[\left(\frac{k - \lambda}{\lambda} \right)^2 \right] \end{aligned}$$

To solve this, we remember that the expectation value is a linear operator, meaning that $E[ax + by] = aE[x] + bE[y]$. Additionally,

we need a convenient equation for the Poisson distribution that links the expectation value to the rate parameter (Pitman, 1997, Eqn. 14):

$$E[X_\lambda^n] = \sum_{k=1}^{n} \left\{ {n \atop k} \right\} \lambda^k \quad (n = 1, 2, \ldots) \tag{4.17}$$

where $\left\{ {n \atop k} \right\}$ are the Stirling numbers of the second kind. For the second algebraic moment we will need $\left\{ {2 \atop 1} \right\} = 1$ and $\left\{ {2 \atop 2} \right\} = 1$. Hence for the second algebraic moment:

$$\mu^2 = \left\{ {2 \atop 1} \right\} \lambda + \left\{ {2 \atop 2} \right\} \lambda^2 = \lambda(1 + \lambda) \tag{4.18}$$

We can then evaluate the expected Fisher information further:

$$\begin{aligned} J(\lambda) &= E\left[\left(\frac{k - \lambda}{\lambda} \right)^2 \right] \\ &= E\left[\frac{k^2}{\lambda^2} - 2\frac{k}{\lambda} + 1 \right] \\ &= E\left[\frac{k^2}{\lambda^2} \right] - \frac{2}{\lambda} E[k] + E[1] \\ &= \frac{\lambda(1 + \lambda)}{\lambda^2} - 2\frac{\lambda}{\lambda} + 1 \\ &= \frac{1}{\lambda} + \frac{\lambda^2}{\lambda^2} - 2\frac{\lambda}{\lambda} + 1 \\ &= \frac{1}{\lambda} \end{aligned}$$

Hence, the Jeffrey's prior for the Poisson family of distributions is given by:

$$f(\lambda) \propto \sqrt{J(\lambda)} = \sqrt{\frac{1}{\lambda}}. \tag{4.19}$$

Note that this is, strictly speaking, an improper prior, meaning that the integral over the prior does not need to be one (or even be finite). This needs to be kept in mind when calculating the posterior distribution, whether or not this causes an issue for the concrete problem.

Other Approaches

In our previous discussion we focused on the situation where we wanted to limit the influence of the prior, for example, because we do not know much about the system. However, in many cases we do know many details about the system we want to analyze. Going back to the medical example in the beginning, if we know the patient is male and does not belong to a risk group, we can define the prior quite precisely.

The same holds in other situations: In many cases, relevant data are available, for example in the form of a census, statistics or other analyses which we can refer to. In a way, we can interpret the training data we use in a machine learning approach as the prior knowledge as well: Making sure that the data represents our system well, these data describe all our knowledge we have, at least implicitly, about the system we want to apply the machine learning model to. This also highlights the role of data quality: If the data are faulty or biased, this can, at least potentially, have a significant impact on the output.

In some other situations, we may have expert knowledge and using the prior, we can include our knowledge in the further calculations. In the same way, we can add constraints from physical processes or engineering into the model, for example, if we know that a system is constrained to operate within a given set of parameter values.

In any case, the Bayesian approach to statistics forces us to think about the prior we want to use and make the choice explicit.

Self-Check Questions

1. The Jeffrey's prior is invariant under ...
2. Bayes rule can be summarized in words as: ...
3. Benford's law is given by: ...

Solutions

1. Bijective transformations of the parameter.

2. Posterior equals likelihood times prior divided by evidence

3. Benford's law is given by:

$$P(d) = \log_{10}\left(1 + \frac{1}{d}\right)$$

4.3 Conjugate Priors

The prior plays a central role in Bayesian statistics. We have previously looked at how we can use the prior to include our *a priori* knowledge of the system we want to consider - and what to do if we want to encode as little information as possible.

However, if we look at the Bayes formula again, we also notice another crucial role the prior plays. Looking back to Eqn. 4.5, the *a posteriori* probability can be written as:

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_i P(B|A_i)P(A_i)}$$

In the case of continuous values, we need to replace the sum with an integral, for example, if we do not consider discrete events but a continuous values of a probability distribution:

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta} \tag{4.20}$$

Here, θ is the parameter of the distribution that we need to determine, $f(\theta)$ is the prior that depends on this parameter, x are the observed data, the likelihood is given by $f(x|\theta)$ and the *a posteriori* distribution is given by $f(\theta|x)$.

This means that we have to sum (or integrate) over all possible values of the likelihood times the prior to evaluate the normalization of the denominator. Similarly, if we want to use the *a posteriori* distribution to determine,

for example, the most likely value, we would have to "integrate out" the parameter, i.e. sum or integrate over the Bayes' formula which again leads us to a sum or integral over the likelihood times the prior.

The likelihood formalizes the description of the observed data.

We cannot influence the **likelihood** too much. However, we can influence the prior: Because we need to perform the sum or integral over the likelihood times prior, we can choose the parametrization of the prior such that this sum or integral becomes easier. Before the widespread use of computers, this was crucial because executing such an integral becomes intractable very quickly and even using computers, the complexity of a numerical evaluation, in particular, if we need to consider multiple parameters, becomes very difficult very quickly.

As a remedy, we can choose the prior such that when combined with the likelihood, the resulting combination is again a probability distribution that we can use easily, in particular, one that can be expressed in the closed form of a commonly used probability distribution. We call these choices of priors "conjugate priors":

Conjugate Prior

A class of priors is called a conjugate prior with respect to a given likelihood function, if the *a posteriori* distribution is of the same family of probability distributions as the prior.

The theory of conjugate priors was first developed in Raiffa and Schlaifer (1961). It is important to keep in mind that, ultimately, choosing a conjugate prior is a convenience - if we can describe our *a priori* knowledge in terms of a conjugate prior, then we can make the further handling of the Bayes' formula easier. However, if we cannot make such a choice, then we should not try to "force" it.

A brief list of conjugate priors is given below (Held, 2008, p. 148):

likelihood	conjugate prior	prior hyper-parameter
Binomial, Bernoulli	Beta	α, β
Negative Binomial	Beta	α, β
Poisson	Gamma	α, β
Exponential	Gamma	α, β
Normal (σ^2 known)	Normal	μ, σ^2
Normal (μ known)	Inverse Gamma	α, β

A more detailed list can be found, for example, in Fink (1997) or Gelman et al. (2013, p. 35ff).

Example: Conjugate Beta Prior

Imagine we flip a fair coin multiple times or we perform an A/B Test. For example, we want to investigate the case where "head" comes up five times in a row when we toss the coin five times in total. We can describe this as a series of Bernoulli trials, described by the binomial distribution. Hence, the likelihood is given by:

$$P(X = k|n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

where k is the number of "heads-up" we observe in n trials (e.g. $n = k = 5$ in our case) and p is the probability to observe "heads up", e.g., $p = 0.5$ for a fair coin.
To compute the *a posteriori* distribution, for example, the probability of a fair coin to show five times "head up" in a row, we need to calculate the following quantity: $p(\theta|k) = p(k|\theta)p(\theta)$, ignoring the denominator that acts as constant normalization, for now.
We now choose the Beta distribution as prior given by:

$$f_{\alpha,\beta}(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{4.21}$$

and that depends on two parameters α and β. Putting this together

we obtain:

$$\begin{aligned} p(\theta|k) &= p(k|\theta)p(\theta) \\ &= \underbrace{\binom{n}{\theta}\theta^k\,(1-\theta)^{n-k}}_{\text{likelihood}}\,\underbrace{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}}_{\text{prior}} \\ &\propto \theta^k\,(1-\theta)^{n-k}\cdot\theta^{\alpha-1}(1-\theta)^{\beta-1} \end{aligned}$$

where we have dropped the constant parts in the last step to make the exposition easier to follow. We can rearrange this as:

$$p(\theta|k) = p(k|\theta)p(\theta) \propto \theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1}$$

Comparing this with Eqn. (4.21), we find that the expression above is again a Beta distribution with parameters: Beta$(k+\alpha, n-k+\beta)$. This means that the *a posteriori* distribution is of the same family of distributions as the prior. In our case, both prior and posterior are both a Beta distribution. Hence, the beta distribution is the conjugate prior with respect to the binomial distribution used in the likelihood.

Self-Check Questions

1. A prior is called the conjugate prior if the ... belongs to the same ... of distributions.
2. The conjugate prior to the Poisson distribution is the ... distribution.
3. The conjugate prior to the Normal distribution with known variance is the ... distribution.

Solutions

1. posterior, family
2. Gamma

3. Normal

4.4 Bayesian and Frequentist Approach

When thinking about probabilities, we can approach this in two different ways. For example, if we keep rolling a die or tossing a coin, we observe the outcome. In the case of the die, we observe the frequency with which each number comes face up. In the case of a coin, we can count the number of "head" or "tails" we observe.

How would we then define the probability of observing "head"? We could repeat tossing a coin over and over and say: The probability for "head" is the number of times we observe the head of the coin face up, divided by the number of the number of tosses of the coin for the limit of very many coin tosses:

$$P(E) = \lim_{n \to \infty} \frac{k}{n} \tag{4.22}$$

In this way of reasoning, we understand the probability as the frequency of events in long running processes or experiments or as a random sample of a - potentially fictitious - population. This is called the "Frequentist's" approach to statistics. Within this school of thought we say that we observe the data as a concrete realization of a random but fixed process. In the example of the coin, we assume that we toss the coin very often or, alternatively, we have a population of very many exactly the same coins we can toss. This implies that the parameters of the process or experiment we base our definition of probabilities on is fixed. For example, the coin is fair and has a 50% probability to show either "head" or "tail". This number, the 50% is fixed and there is no uncertainty. In this interpretation, the data we observe are a random sample: We imagine that there is such a process of the coin being tossed very many times and we then compare the data we observe to this model and evaluate if the observed data are compatible with what we expect from the model or not.

On the other hand, this is not how we use probabilities in our day-to-day live. We have come across the quote "When you hear hoof beats, think of horses, not zebras" already (see, for example, Dickinson (2016)): When we see a light in the sky at night, we would say that it is far more likely that this is due to a plane or a satellite rather than an alien spaceship. This

is the Bayesian way of thinking about probabilities: We assign a degree of plausibility or belief to what we observe. Coming back to the example of the coin, we do not know *a priori* if this is a fair coin - we may be inclined to believe so to start with but we are not sure. In this way of reasoning, we would assign a prior for the probability of the coin to show "head" or "tail" that may have its highest value at $p = 0.5$ but is fairly broad. As we observe the coin tosses, we can "update" our belief and adapt the prior so that it matches our observations, for example, that the coin is fair.

Comparing this to the Frequentist's approach above we notice two major differences: Firstly, in the Bayesian approach we need to define a prior that expresses our *a priori* knowledge or belief. In the case of the light in the sky, we would say that $P(\text{aliens})$ is much lower than, say, $P(\text{plane})$. This prior does not exist in the Frequentist's definition. Furthermore, we treat the parameters of the model as variable: In the Frequentist's view, the coin is fixed, its probability of showing "head" is fixed but maybe unknown. Unlike in the Frequentist's approach. the parameter of the model, in this case, the probability of showing "head" is a random variable in the Bayesian method that is described using a probability distribution. The data, on the other hand, are fixed: They are what we observe.

The main difference between the two approaches reduces to the understanding of probability: In the Frequentist's view, probabilities are the frequencies of the occurrence of events in long running experiments or a population, in Bayesian statistics, they represent our degree of "belief" and are modelled as a random variable using a probability distribution.

Much of the debate between Frequentist's and Bayesians center around the understanding and treatment of probabilities. In the Frequentist's view one can say that using a prior introduces a subjective point of view into the statistical analysis: Since there is no single way to determine the prior and, indeed, we may use it to incorporate expert knowledge, the answer depends on this subjective choice. On the other hand, Bayesians might argue, that this is exactly the point: We make our assumptions explicit and can calculate probabilities even for single events. For example, what is the probability that a given candidate will win an election? From a Frequentist's point of view, this question does not make much sense: We cannot repeat the election infinitely often with exactly the same settings nor do we have a population of exact replicas of the country in question in which we can observe the outcomes and determine the probabilities.

In Bayesian statistics, we can model our assumptions in the prior and calculate the *a posteriori* probability.

Self-Check Questions

1. In Bayesian statistics, the data are treated as ....
2. Which element in Bayesian statistics object Frequentists the most to?

Solutions

1. fixed
2. The prior

Summary

Bayesian statistics allows us to make make inferences using the data we observe. Central to Bayesian statistics is Bayes theorem that links the likelihood that we use to describe the data we observe using probability distributions with a prior to compute the posterior which expresses all our knowledge about a particular system, given the data. The prior is the most important aspect of Bayesian statistics as it encodes our prior knowledge, expert opinion or even our subjective degree of belief in a certain hypothesis. Choosing a suitable prior is a challenging task and there is no single way to do so. In some cases, we may wish to encode as little information the prior as we can and Jeffrey's rule allows us to construct such a prior. In other cases we use the prior to make the knowledge we have explicit. To ease the computation of the posterior, we can also choose conjugate priors that have the property that they are of the same family of distributions as the posterior. Although, in a way, using priors corresponds to our natural way of thinking, the use of priors is also the most controversial aspect - the critics point out that this use of subjective information makes the results difficult to generalize or transfer from person to person.

5 Data Visualization

Study Goals

After completing this unit you will have learned:

- The importance of visualizing data.
- The general principles of visualizing data.
- The most important chart types.
- How to construct histograms, scatter plots and profile histograms.

Introduction

The visualization of data plays a vital role in using data and we can use it throughout the process of understanding data and using data to extract information or making inferences.

For example, before we attempt to build a statistical model that describes our data, we can use a process called "exploratory data analysis" (EDA) to gain a first insight into the data: Visualizing data helps us to get a first idea about the distribution and variability of the variables in our dataset, appropriate charts allow us to explore correlations or dependencies between two or more variables.

While building a model, visualizing the data can help us to make the model more tangible and make the interpretation easier. For example, if the model makes a specific prediction, we can compare the model to the data visually and illustrate the results of the modeling stage.

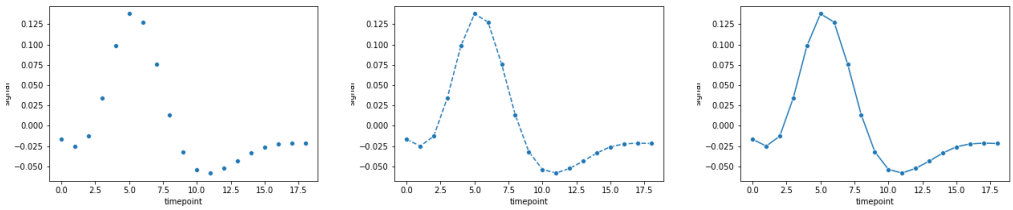


Figure 5.1: Showing the same data where the data points are not connected, connected with a dashed line or a solid line.

Finally, we can use data visualization techniques to tell a "story" about and with the data: Using appropriate charts, we can guide the audience through understanding the data, how our model works and which conclusions and predictions we can make using the data.

We begin this unit with some general considerations when building data visualizations and then focus on specific ways of visualizing data, such as histograms, box and violin plots, bar charts and others.

5.1 General Principles

When we visualize data, we not only have to choose the appropriate chart type or technique to do so, but within these choice we also have to choose the design elements that are used to create the visualization.

One of the earliest reference work is Haskell (1919), outlining how we should create graphics. One key aspect to keep in mind is that the visualization is not truly objective: By choosing a specific visualization type and visualization style, we can suggest associations the audience should make. This can lead to cognitive biases - in some cases, this is what we intend to do: The audience should follow our "narrative" about the data. In other cases, this is an unintended consequence and we may bias the audience into thinking in a specific way.

We can illustrate this with the example shown in Fig. 5.1 which has been created using the "seaborn" package (M. Waskom, 2020) using the test datasets included in the package. In this particular example, we illustrate some data-points that may have been taken from a sensor reading. We

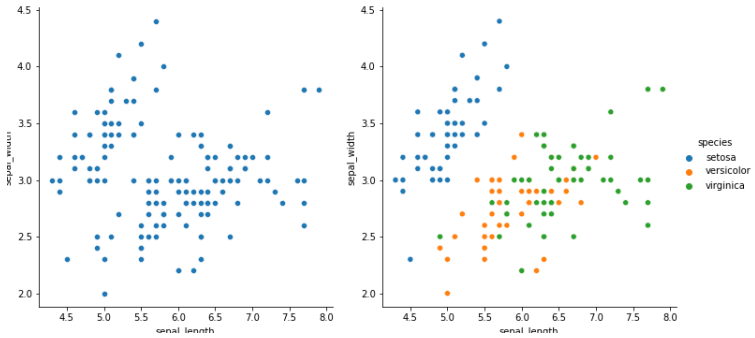


Figure 5.2: Using color to highlight semantic classes in the data.

could imagine that the x-axis represents time and the y-axis, for example, the temperature of a given system. The sensor, our thermometer, takes regular readings. Depending on how we visualize the data, we may induce a specific conclusion in the audience. For example, if we just draw the data-points, we may want to indicate that each point represents an independent measurement. Presenting the data this way we also express that we do not know anything about the system between the data-points: Since there is no measurement, we do not have information about the state of the system between the measurements - we may infer the behavior, but this requires a model with its corresponding assumptions. However, if we choose to connect the data-points with either a dashed or a solid line, we can induce the interpretation that the points are not independent but that there is a linear relationship between one point and the next. Not only do we indicate that the measured values are connected to each other, we also imply that there is a linear relationship that we can use to derive values between measurements. When we choose the visualization style, we need to think about which associations we want the audience to make, choosing a particular style can have implications beyond making the plot "look nice".

Changing the color or **hue** of the visualization can also help to make the plot more accessible. Using the famous "iris" dataset (Fisher, 1936), we can see in Fig. 5.2 that using color information allows us to visually identify, analyse and interpret semantic classes in the data. However, if we use this to try to display too many classes as shown in Fig. 5.3, our eye is drawn more to the solid lines visualized in stronger colors and we intuitively interpret the softer colors and dashed lines as less important. Depending on the data, this may be what we want to show, but it may also induce a cognitive bias in our audience. We always need to consider the impact on

In the standard CIECAM02, hue is one of the six dimensions that describe the color appearance.

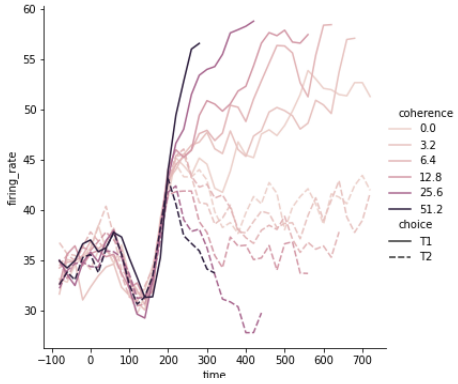


Figure 5.3: Using color and hue in data visualization

the audience when we choose a visualization style and we should get into the habit of making each decision deliberately, rather than relying on the defaults of the program we use to create the visualization.

Based on the key concepts by Stephen Few, Parmenter suggests the following best practices (Parmenter, 2015, p. 222)

- Keep within the boundaries of a single screen: Instead of providing many options of unfolding detailed charts and data representations, think carefully what the intended audience should see and how the information should be presented.

- Provide sufficient context: Indicate whether the numbers are within a "good" or "bad" range. This range has to be decided beforehand including the advice of experts.

- Provide adequate level of detail or precision: If you add numbers to the visualization, the level of detail in charts or the precision of the numbers shown should reflect the overall message of the visualization.

- Start scales at zero: It is often tempting to start the scale of graphs at some other point than zero. However, this introduces a cognitive bias and distorts the magnitude of differences.

- Keep a consistent colour scheme: All visualisations should have the same colour scheme, for example, low numbers indicated by blue, high numbers by red. In addition, the colour scheme should use as

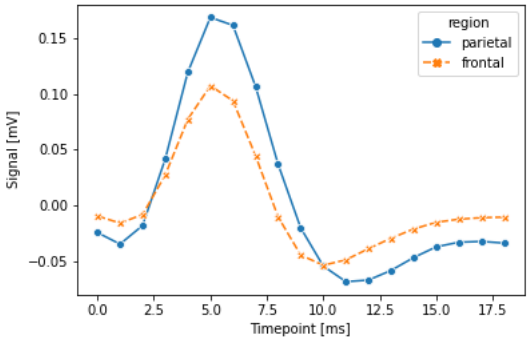


Figure 5.4: Graph with multiple elements and labels

few colours as possible but as many as necessary to avoid clutter and keep the visualisations simple.

- Avoid decorations: Additional decorations and graphical elements without functionality mostly just clutter the visualization and should be avoided unless they provide additional context.

Additionally, we should follow these points as illustrated in Fig. 5.4:

- All plots should be clearly labelled: Each axis should have a relevant description, the units should be added, if there is more than one dataset visualized in a plot, a legend should be used to indicate which graph shows what.

- All labels should be large enough. A good test is to imagine we give a presentation and the persons in the last row of the large venue still needs to be able to read and interpret the plot.

- Be inclusive in the choice of colors. For example, a number of people cannot distinguish between red and green, so we should avoid these colors as much as possible, at least their combination. A good way to test this is to reproduce the figure in black and white and check that it is still usable. In addition to color, use different marker and line styles to distinguish between multiple elements in the graph so that there are at least two different ways of obtaining the same information.

A nice “cheat sheet” how to choose a particular visualization type and style

can be found in Franconeri (2019).

Self-Check Questions

1. Which combination of colors is problematic?
2. Why do we need to think about cognitive biases when we create any visualization?
3. Axis in plots should always show: ...

Solutions

1. Red and green because quite a few people cannot tell them apart.
2. The way we create the visualization can already influence the interpretation by the audience. We should always look at a visualization from another angle and ask ourselves which meaning we convey with it.
3. Scale, units and label.

5.2 One- and Two-Dimensional Histograms

One-dimensional Histograms

When analyzing a small dataset, we can easily look at each individual element of our dataset. However, once we need to analyze a large dataset, it becomes impossible to look at individual values of variables, in particular in the case of continuous values. An efficient way to visualize and work with large numbers of continuous data are histograms: Instead of looking at individual values, we split the x-axis in J intervals. This implies that we turn the continuous range of values observed in the data into **bins**. This is called a "histogram" and was first introduced by Pearson (1895).

Bins are a discrete set of intervals along an axis.

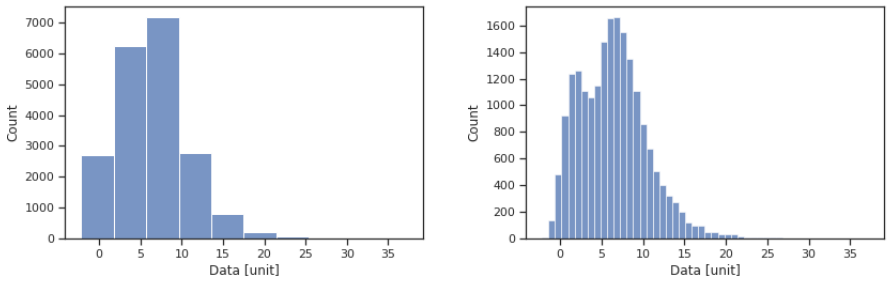


Figure 5.5: Visualizing the same data with 10 or 50 bins

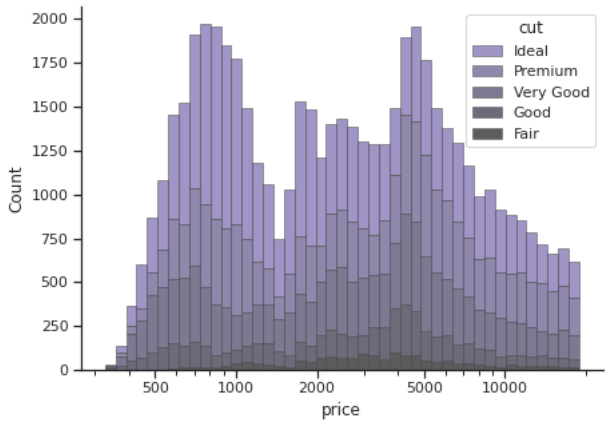


Figure 5.6: Stacked histogram with logarithmic x axis.

Choosing the "right" number of bins has a major impact not only on how we visualize the data but also how we can use the resulting histogram in a further analysis. In Fig. 5.5, the same data of some continuous variable is shown as a histogram twice: In one case, we only use 10 bins and come to the conclusion that the data are distributed almost symmetrically around a single center. However, if we choose 50 bins, we find that the data contain two (instead of just one) peaks and the tail on the right is much more pronounced than on the left. Clearly, choosing the right binning has a major impact on our further understanding and analysis, which we will discuss in more detail below. Depending on the values, we may choose to use a logarithmic scale on the x axis. If we want to compare multiple histograms, we can, for example, use stacked histograms as shown in Fig. 5.6.

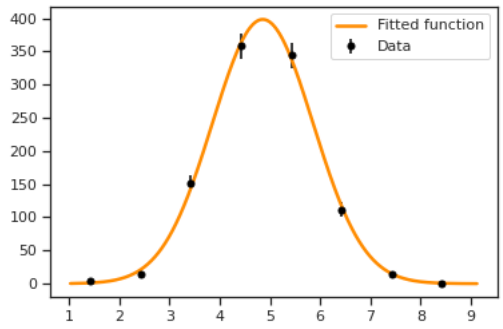


Figure 5.7: Histogram with error bars and overlaid fit function

Another use-case for histograms is the determination of the parameters of an empirical or theoretical model from the observed data. In an ideal case, we would use each data point in this determination, for example, in an unbinned maximum likelihood fit. However, if the number of data-points gets very large, this becomes numerically very challenging and we need to resort to a binned determination. Essentially, we fill the data into a histogram and then use these binned values in the determination of the parameters of the model. This is illustrated in Fig. 5.7 that shows a dataset generated by drawing random numbers from a Gaussian distribution and then fit a Gaussian distribution to the binned data, the histogram. In reality, we would of course not generate these data but measure them in an experiment. Unlike the histograms so far, the data are now represented by points with error bars. We can understand this in the following way: We discretisize the data by defining J bins and place the data into the respective bin. In each bin j, we observe n_j entries with $\sum_{j=1}^{J} n_j = n$, i.e., the sum of all entries in the data corresponds to the number of data points observed. We can then interpret the number of entries in each bin of the histogram as a random number. More precisely, we interpret the number n_j as the expectation value of the probability distribution that describes the underlying process:

$$\mu_j(\theta) = n \int_j f(x|\theta)dx \approx n \cdot f(x_c|\theta) \cdot \Delta x \tag{5.1}$$

In this equation, μ_j is the expected number of entries in bin j and $f(x|\theta)$ is the underlying probability distribution for our process with parameter θ. Since we choose the bins such that the variation of the probability

distribution is not too large from one bin to the next, we can replace the integral by the approximation that μ_j is given by the value of the distribution in the centre of the bin times the bin width Δx, multiplied with the total number of events. We can then interpret the number of events in each bin j as a random variable that follows a Poisson distribution (Blobel & Lohrmann, 2012, p. 138):

$$P(n_j|\mu_j) = \frac{\mu_j^{n_j} e^{n_j}}{n_j!} \tag{5.2}$$

The error bars indicating the statistical uncertainty in each bin of the histogram are then obtained from the definition of the standard deviation of the Poisson distribution:

$$\sigma_j = \sqrt{n_j} \tag{5.3}$$

This treatment works well in most practical situations. However, we have to take care to consider a few subtleties. For example, if a bin has no entry, the expected uncertainty is also zero, which is clearly not the case. The key point to remember is although that the observed data do not have an associated uncertainty with them - they are what we observe. What the error-bars indicate is the uncertainty we would expect if we were to repeat the same experiment repeatedly: Then the number of data points in each bin would vary according to a Poisson distribution and it is this associated uncertainty we add to the data points. For further details see also, for example, Aggarwal and Caldwell (2012).

Choosing the Right Binning

As our previous discussion indicated, choosing the right binning for a histogram is crucial for further understanding and analysis of the histogram: If we choose too few bins, we will miss important features in the distribution of the data, if we have too many bins, we will both lose the advantage of using a histogram for a large dataset and likely observe artifacts where some bins are empty and others are not due to an unfortunate choice of binning.

Unfortunately, it turns out that the best choice of binning is not straight forward to answer and there is no single answer. Looking at the academic literature, the optimal bin size according to Scott (1979):

$$W = 3.49\sigma N^{-1/3} \tag{5.4}$$

where W is the width of the bin, σ the standard deviation of the distribution and N the total number of observed data points in our sample. Izenman (1991) derive a similar equation:

$$W = 2(IQR)N^{-1/3} \tag{5.5}$$

where IQR is the interquartile range, i.e., the 75% percentile minus the 25% percentile. Sturges (1926) on the other hand suggests for the number of bins k:

$$k = \lceil \log_2 n \rceil + 1 \tag{5.6}$$

Further details can also be found in, for example, Gholamy and Kreinovich (2017); He and Meeden (1997).

Since there is no single rule we can adhere to, we can look at the choice of binning from a broader perspective. Generally, we have two choices for binning: We can choose equidistant bins where all bins have the same width or variable bins where the width of the bin is not fixed. In the case of fixed or equidistant binning, we split the range of the x-axis into J intervals of equal length. We can use formulae above as a guidance to determine the width or the number of bins in this case. However, these formulae do not take any experimental setup into account but assume that we can measure the data with unlimited precision. In a realistic setting, the data are always acquired using some measurement device. In some cases, the resolution of this device may be very high and we may not have to worry about it unless we choose many bins. However, in many realistic scenarios, the finite resolution of our sensors or measurement devices limits the width of the bins we can choose and the precision with which we can draw conclusions from the data. For example, if the width of the bins is lower than the resolution of the sensor, we cannot make a firm statement if a given observation should be attributed to bin $j-1, j$ or $j+1$. Due to the resolution, the "true" event could be in either - but we do not know since the sensor does not allow us to look in more detail. Consequently, the bins will be highly correlated: Imagine, for example, we measure the occurrence of an event, in one measurement we put the event into bin j, in the next into bin $j+1$, etc. - but in reality the events are very similar and should belong to the same bin. Hence, the width of the bins should at most be matched by the resolution of the sensors or measurement devices, and we should aim to minimize these bin-to-bin migrations. This is particularly important because we have assumed earlier that the bins are independent from each other when we treated the number of entries in

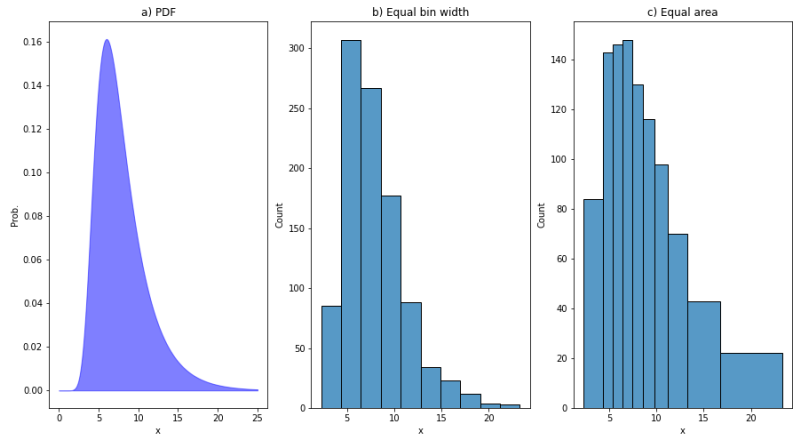


Figure 5.8: Effect of fixed and variable bin widths. a) PDF according to which the data are generated, b) equal bin width, c) equal area in each bin.

each bin as a random variable following a Poisson distribution. With the same argument, we should also aim for a binning where each bin has at least a few entries to avoid the ambiguities that arise for the uncertainties if bins have no entries.

Alternatively, we can choose variable bin widths where the width of the bin can vary from one bin to the next. Again, there are several choices how we can do this: The most common approach is to choose bins such that each bin has the same number of entries and each bin border is chosen such that it lies between two neighboring observations. In this case, the number of entries in each bin is fixed, but the distribution of bin borders is a random variable. For further details see, for example, Sulewski (2020). This way of constructing histograms has the advantage to avoid empty bins or bins with just one or two entries, however, depending on the range of the data, the bins may become very wide. The bulk of the data is, however, captured in narrow bins. As a variant, we can construct the bin edges such that each bin has the same area. The effect is illustrated in Fig. 5.8. Part a) of the figure shows the asymmetric distribution that is used to generate a sample of 1000 data points. Part b) shows a histogram with equidistant bins, the bins in part c) are chosen such that each bin has the same area.

This choice of binning again does not take the measurement process into account: If our data are captured using sensors or a measurement device with a given resolution, we can choose the binning such that the bin edges match our resolution. In our previous discussion where we considered the finite resolution, we implicitly assumed that the resolution is fixed across the range of measurements. However, this does not necessarily need to apply - the sensor may be very precise in one range of measurements but have a lower resolution in others. Adjusting the bin widths accordingly again limits bin-to-bin migration and correlated histogram bins.

Kernel Density Estimation

In our discussion so far we have encountered histograms as a way to visualize the data. In many cases, we assume that the data are concrete realizations of a random variable for the process we observe that in turn is described by a probability distribution. In this sense, the histogram allows us to get an idea of the underlying probability distribution. However, as we have discussed above, the choice of binning has a major impact on the visualization and the further analysis of the data. Kernel Density Estimation (KDE) aims to tackle this challenge from a different angle. Instead of treating each measurement as a single "building block" from which we create the histogram, we treat each measurement as the position of a suitable "kernel". In most cases, we will use a Gaussian or normal distribution as the kernel, although there are other choices. This way, each measurement is taken as the central value of the Gaussian kernel and the size of the kernel is controlled by a further hyper-parameter h, the "kernel bandwidth".

More formally, the aim of kernel density estimation is to approximate a probability density function $f(\cdot)$ of a random variable X. We observe n concrete realizations of this variable, i.e., we have n data-points $x_1, \ldots, x_n$. The kernel density estimator is then defined as:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) \tag{5.7}$$

The factor $1/nh$ is a normalization and the function $K(\cdot)$ is the kernel. Typically, a Gaussian function is used as the kernel, although other choices are possible. Effectively, the kernel density estimator places a kernel at each observation and the sum of all these kernels then gives a smooth

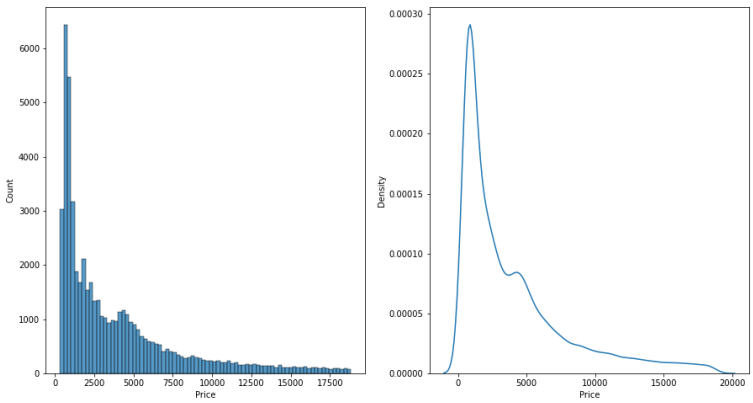


Figure 5.9: Comparison between histogram (left) and KDE (right).

approximation of the original probability density function. The parameter h controls how "wide" each kernel is, and controls the smoothness of the resulting approximation.

Using these kernels instead of single measurements gives a much smoother estimation of the shape of the distribution as illustrated in Fig. 5.9.

This approach has several advantages: The resulting distribution is much smoother and less prone to small differences in measured data points. Furthermore, we can use this as a method to build an empirical probability distribution without relying on a theoretical model. On the other hand, a KDE plot does no longer show where the data are in the visualization, we cannot use it to fit a theoretical curve, nor can we include the finite resolution of a sensor or measurement device.

For further details see, for example, VanderPlas (2016, chap. 5.13) or Pedregosa et al. (2011, chap. 2.8.2).

Two-dimensional Histograms

Our previous discussion centered on one-dimensional histograms where the x-axis depicts the variable we are interested in and the y-axis shows the frequency at which we measure each value. We can extend this concept to

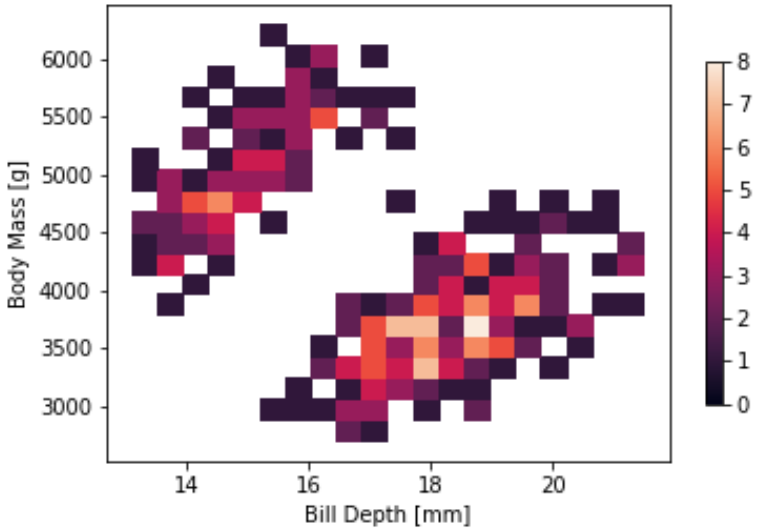


Figure 5.10: Two dimensional histogram. Data are taken from M. Waskom (2020).

two-dimensional histograms as illustrated by Fig. 5.10. Each axis is used to depict a variable and the entries in the corresponding bin are increased if the value of the variable along the x-axis and the value of the variable along the y-axis fall within the respective interval. The same considerations for the choice of binning as in the one- dimensional case apply here as well.

As with the one-dimensional histogram, a two-dimensional histogram can be used more efficiently, if, for example, the number of data points are very large as shown in Fig. 5.11.

Self-Check Questions

1. In a histogram with equidistant bins, the entries are a random variable that are distributed according to which probability distribution?

2. If the data visualized in a histogram are taken a measurement, what limits the bin width?

3. What is the formula for the bin width W according to Scott's rule?

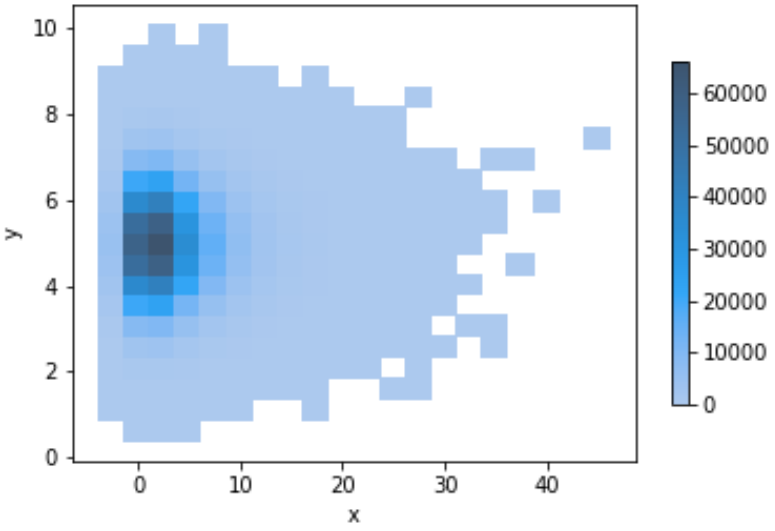


Figure 5.11: Two-dimensional histogram with many entries.

Solutions

1. Poisson
2. The width of the histogram should not be smaller than the experimental resolution.
3. The bin width is given by: $W = 3.49\sigma N^{-1/3}$.

5.3 Box and Violin Plots

The visualization techniques we have encountered so far, in particular the histogram, allow us to visualize the data and investigate their shape. However, they are not convenient if we want to summarize the main characteristics of a set of data-points.

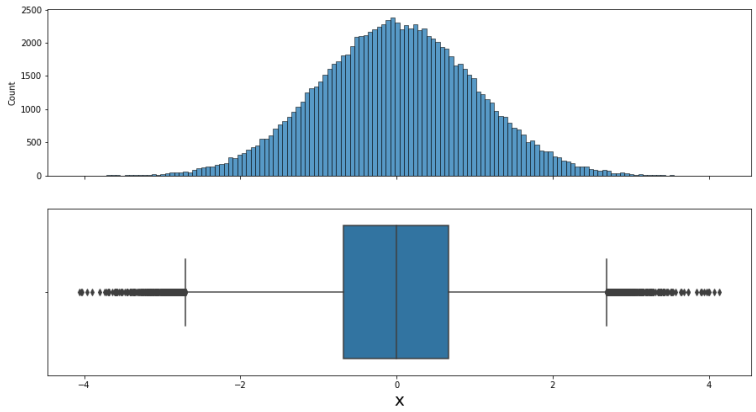


Figure 5.12: Histogram and box plot for a Gaussian distribution.

Box Plots

One way to include more quantitative information in the visualization is the "box plot". This approach was originally developed as a range bar (Spear, 1952, 1969) and then later augmented (Tukey, 1970, 1977). The history of the box plot is discussed in more detail in Wickham and Stryjewski (2011).

The box plot visualizes the most common metrics we can use in descriptive statistics and typically consists of the following elements:

- The minimal value in the data excluding outliers.
- The maximum value in the data excluding outliers.
- The median (50% quantile).
- The first and third quartiles (25% and 75% quantile).
- **IQR** = Q_3 - Q_1

`IQR is short for interquartile range`

Although there is clearly no single definition of an outlier, a common choice is to consider all points outside the interval $[Q_1 - 1.5 \cdot \text{IQR}, Q_3 + 1.5 \cdot \text{IQR}]$ as potential outliers (Tukey, 1977).

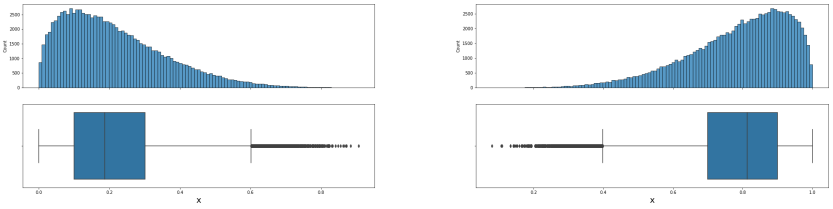


Figure 5.13: Histogram and box plot for a positively and negatively skewed distribution with tails on one side.

A simple example is shown in Fig. 5.12 for a Gaussian or normal distribution with $x \sim \mathcal{N}(0, 1)$. The box itself shows the interquartile range (IQR), the vertical line the median, the whiskers the minimum and maximum value except the outliers that in turn are shown as points.

In the case of an asymmetric distribution, the box plot can help to make the most important metrics more accessible as shown in Fig. 5.13. Depending on the skewness of the distribution, the median will be to the right or the left of the IQR box and the whiskers quantify the range of the tail more precisely. However, as we can see from the figure, for heavily skewed distributions the standard definition of the whiskers may be sub-optimal can be improved (Hubert & Vandervieren, 2008).

Violin Plots

Box plots are very useful to help us understand the characteristics of a distribution in a more quantitative way. However, since they do not visualize the distribution and the shape of the data themselves.

Violin plots (Hintze & Nelson, 1998) are a variant of the box plot that includes a visualization of the data themselves. The quantitative elements of the box plots, namely the interquartile range and the median, are shown with a central line. Instead of the box, a kernel density estimate similar to the one we have encountered earlier when we discussed histograms, is used to illustrate the shape of the data distribution as shown in Fig. 5.14. The median is shown as a white dot, the interquartile range as a thick black line.

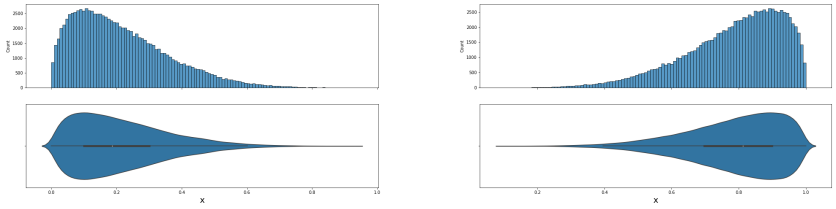


Figure 5.14: Histogram and violin plot for a positively and negatively skewed distribution with tails on one side.

As we can see from the figure, the more elements we add to the visualization, the more difficult it becomes to make the graph easily accessible. When choosing a particular visualization, we always need to consider if this is a suitable way to convey the relevant information.

Self-Check Questions

1. What is a common definition of an outlier?
2. In box or violin plots, which metric is used to indicate the central point of the distribution?

Solutions

1. A common definition is: $[Q_1 - 1.5 \cdot \text{IQR}, Q_3 + 1.5 \cdot \text{IQR}]$
2. The Median.

5.4 Scatter and Profile Plots

Most of the visualization elements we have encountered so far allow us to get a deeper understanding of a single variable. In many cases we also want to learn more about the relationship between two variables.

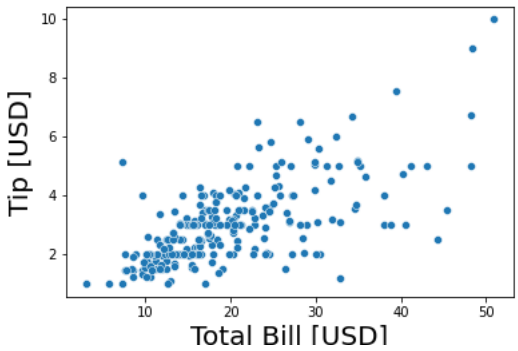


Figure 5.15: Scatter plot of the total bill vs. the tips in a restaurant visit. Data are taken from M. Waskom (2020).

Scatter Plot

For example, using the "tips" dataset in seaborn (M. Waskom, 2020), we might be interested in the relationship between the total bill in a restaurant and the tip the patrons left as illustrated in Fig. 5.15. This is called a "scatter plot". For more details about its historical origins, refer to, for example, Friendly and Denis (2005).

Pearson correlation coefficient: $\rho = \frac{\overline{(x-\mu_x)(y-\mu_y)}}{\sigma_x \sigma_y}$.

The data show that there is some dependency that seems to indicate a **correlation** between the the total bill and the tip. We might want to add some further details to the scatter plot, for example, whether or not there were smokers present in the group of patrons. Using color, we can highlight these cases as shown in Fig. 5.16.

As a variant, we can instead, or in addition, vary the size of the marker, this is called a "bubble plot" and an example is shown in Fig. 5.17. As with other approaches, we need to be careful to avoid adding too many details into one visualization as they quickly become hard to read and understand.

If there are not too many variables we want to investigate for our further analysis, we often use a scatter plot matrix as shown in Fig. 5.18. This arrangements shows the behavior for the combination of any two variables. The diagonal elements are visualizations of the variable itself, for example, as a histogram or density plot. The example uses the iris dataset (Fisher, 1936) which contains four feature variables for three types of iris flowers.

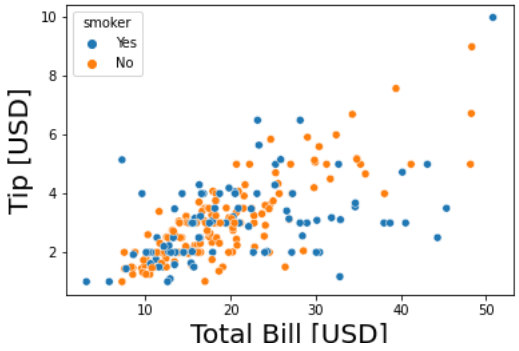


Figure 5.16: Scatter plot of the total bill vs. the tips in a restaurant visit, using color to indicate whether or not smokers where present. Data are taken from M. Waskom (2020).

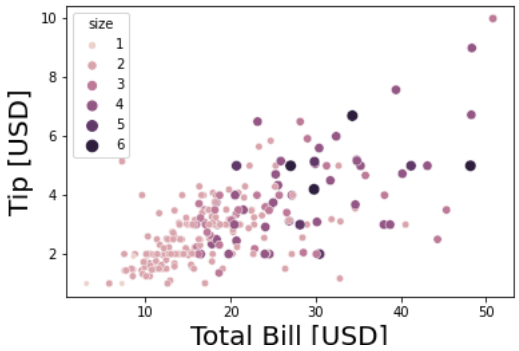


Figure 5.17: Bubble plot of the total bill vs. the tips in a restaurant visit, using color and the marker size to indicate the size of the group of patrons. Data are taken from M. Waskom (2020).

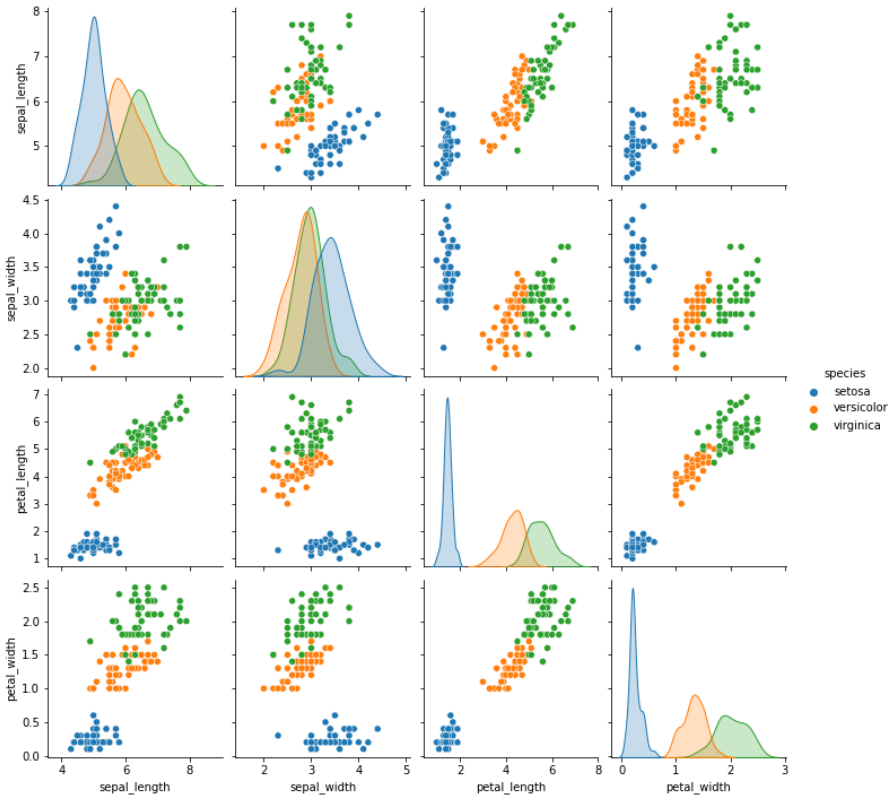


Figure 5.18: Matrix of scatter plots using the iris dataset Fisher (1936).

The types of flower are setosa, versicolor and virginica and the feature variables are the length and width of the petal and sepal of the flower. Each element of the scatter plot matrix shows the behavior of the respective variable(s). We can already see at first glance that we can tell the species "setosa" easily apart from the others.

Although scatter plots seem very useful at first glance, they become more challenging to use if we have to analyze many data points. Imagine the situation that we need to understand a dataset that contains a number of variables, such as, for example Y and X. In many situations, we will know what these variables are and our domain knowledge will guide us to some expectation how these variables behave and how they are related to each other. However, in practice we will quite often encounter the situation where we do not know anything about these variables and our first task is then to understand more about the behaviour of the variables

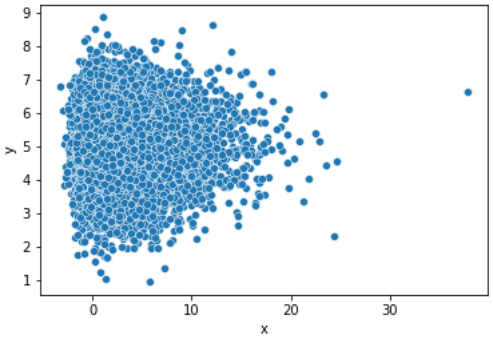


Figure 5.19: Scatter plot with 10,000 entries.

and their relationship. As an example, let us consider an unknown dataset containing 10,000 observations of X and Y and look at the scatter plot in Fig. 5.19. We notice immediately that the scatter plot is much less useful than in the earlier examples where we only had a few hundred data points. Instead of a structure, the main area of the distribution is covered in a seemingly uniform structure. This is because each data point is added individually. This data sample only has 10,000 entries - in modern Big Data environments, we can imagine that we have easily hundreds of thousands or more data points. Therefore, we need to look at alternative ways of visualizing the data.

There are multiple approaches that we can take to improve the visualization in these cases, two examples are shown in Fig. 5.20. The left part of the figure shows the same scatter plot as in Fig. 5.19, however, we have added the marginal distribution to the scatter plot. The marginal distribution is obtained by projecting the data on the respective axis. For example, the marginal distribution along the y-axis is a Gaussian or normal distribution. We can imagine that we obtain this distribution by looking along the axis, effectively ignoring the distribution of the data along the other axis. This does not improve the scatter plot per se, but we gain a better understanding of how the data are distributed. Additionally, we can replace the scatter plot by a histogram as shown in the right part of the figure. We have discussed two-dimensional histograms earlier, in this example we use a variant that uses hexagonal bins instead of squares in x and y direction. While this may be better suited to the two-dimensional structure of the data, this choice has the additional challenge that defining the bin sizes is

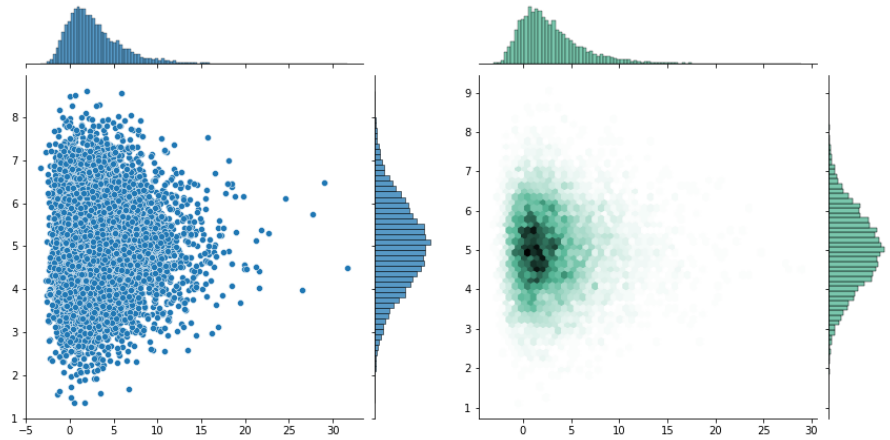


Figure 5.20: Scatter plot with 10,000 entries showing the marginal distribution (left) or a histogram using hexagonal bins (right).

even more complex compared to the case we have encountered so far.

Profile Plot

In many cases we are interested in analyzing how one variable behaves as a function of another. Scatter plots and two-dimensional histograms can help to illustrate this relationship, but both have shortcomings: If our dataset consists of many measurements, scatter plots cannot be used to visualize finer details. Two-dimensional histograms can illustrate the behavior of the two variables better, even in the case of high statistics - but neither approach can be used to quantify the relationship further. For example, if we want to express the relationship between the variables in terms of a fitted function or a theoretical model, neither approach provides a straightforward path to do so. This is particularly challenging in the case of many data points where taking each measurement into account may lead to significant computational challenges.

The profile plot is a one-dimensional representation of the two-dimensional data and constructed in the following way: First, the x-axis representing one variable is discretized into n bins. In each bin of of x, the bin borders are applied as constraints or selection criteria on the other variable for the y-axis. For example, for bin j, we then analyze all values of y for

which $\text{bin} - \text{border}_j \leqslant x < \text{bin} - \text{border}_{j+1}$. Within each of these bins, we compute a localization parameter, typically the sample mean, as well as a **dispersion parameter**. There are several choices for the dispersion parameter. The most common ones are the standard deviation of the distribution of y in each bin of x, the root mean square (RMS) of this distribution, or the error on the mean. The root mean square is defined as

The dispersion parameter is a metric for the variance or volatility of a distribution.

$$x_{\text{RMS}} = \sqrt{\frac{1}{n}\left(x_1^2 + x_2^2 + \ldots + x_n^2\right)} \tag{5.8}$$

where x_i are the observations of a variable X. Choosing the standard deviation or RMS puts more emphasis on the width of the distribution of y, whereas choosing the error on the mean emphasizes how well we can compute the location parameter given the data.

The profile plot is then constructed from these quantities in the following way: For each bin in x, the marker is placed in the middle of the bin such that its position in y corresponds to the mean we have computed as the localization parameter earlier. The error bars in the x direction indicate the width of the bin, the error bars in the y direction the dispersion parameter. The profile plot is illustrated in Fig. 5.21, both for the choice of the standard deviation and the error on the mean as dispersion parameter. Comparing it with the scatter plot for the same data, we can see that the profile plot reveals the structure of the data more clearly. As the profile plot is now a one-dimensional histogram, we can use the corresponding data points to extract the parameters of an empirical or theoretic model using a suitable fit to the data.

Self-Check Questions

1. Which localization parameter do we typically use in profile plots?
2. Name three choices for the dispersion parameter used in profile plots.

Solutions

1. The sample mean

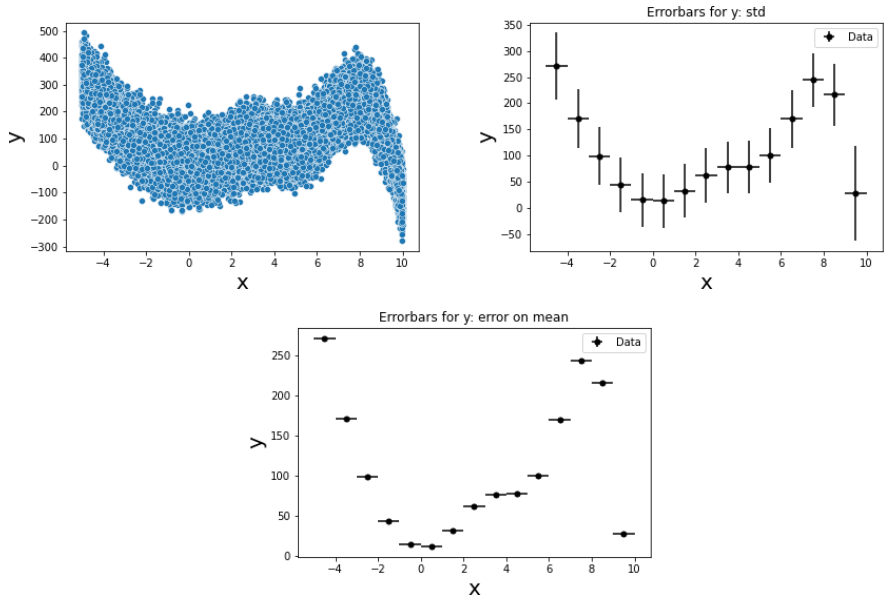


Figure 5.21: Dependency of two variables for a dataset with 100,000 data points: As a scatter plot, as a profile plot where the uncertainty on y is given by the standard deviation (std) and the error on the mean (yerr).

2. Standard deviation, RMS, error of the mean.

5.5 Bar and Pie Charts

Bar Plots

The visualisation approaches we have discussed so far mainly focus on the analysis of continuous variables.

In many cases, we need to work with categorical data as well, namely, data that can be grouped in fixed categories.

As an example, we look at the tips left by patrons in a restaurant, the data are available as part of the seaborn package (M. Waskom, 2020) (the `tips` dataset).

total bill	tip	sex	smoker	day	time	size
16.99	1.01	Female	No	Sun	Dinner	2
10.34	1.66	Male	No	Sun	Dinner	3
21.01	3.50	Male	No	Sun	Dinner	3
23.68	3.31	Male	No	Sun	Dinner	2
24.59	3.61	Female	No	Sun	Dinner	4
...						

While the variables "total bill" and "tip" are continuous, the remaining variables are categorical. If we want to look at how the tips or the total bill vary as according to the weekday at which the patrons were at the restaurant, we should not choose a scatter plot since this implicitly assumes that we use continuous variables and will show artifacts due to the discrete nature of the variable "day". We can, however, use a bar plot to illustrate the behavior as shown in Fig. 5.22. The bar plot is one of the longest used visualization technique, a summary of the historical developments can, for example, be found in Beniger and Robyn (1978). If we want to illustrate more than one category, we typically place the corresponding bars close to each other such that they are well separated between the categories.

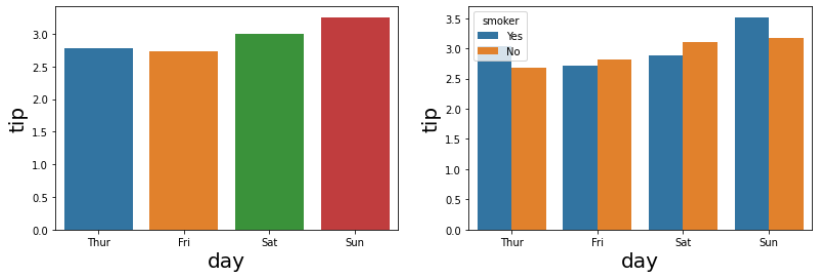


Figure 5.22: Bar plot illustrating the behavior of a continuous and a categorical variable. Data are taken from M. Waskom (2020).

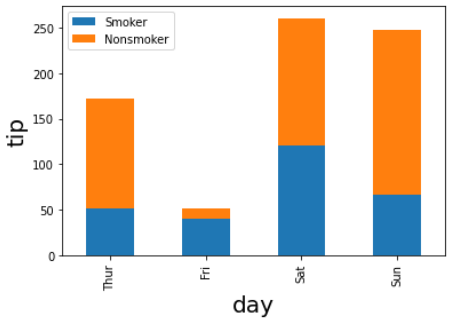


Figure 5.23: A stacked bar plot illustrating the behavior of a continuous and a categorical variable. Data are taken from M. Waskom (2020).

If we want to highlight the relative fraction within categories, we can use a stacked bar plot. The example in Fig. 5.23 shows the sum of all tips on the y-axis and then uses the category "smoker" to illustrate the fraction of patrons where at least one smoker was present.

Pie Chart

In many cases we want to illustrate the composition of a sample where we want to highlight the fraction with which each element contributes to the total. A convenient way to show this relationship is the pie chart which was first used in Playfair (1801).

For example, if we want to show which pet people keep, we can use a pie chart as shown in Fig. 5.24: The sum of all pets defines the whole "pie"

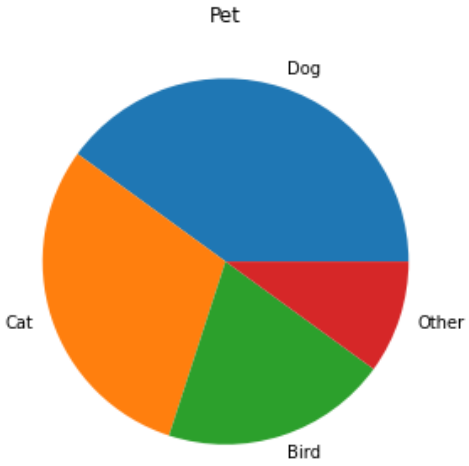


Figure 5.24: A pie chart showing the relative fraction of pets (fictitious data).

and each type of pet is symbolized by a "slice". However, we should keep in mind that this visualization works best if there are only a few categories. Since humans can judge angles only up to a point, using a pie chart for many categories does not aid the understanding of the data as illustrated by Fig. 5.25

Self-Check Questions

1. Which visualization is suited best to illustrate relative fractions?
2. What do we use the stacked bar chart for?

Solutions

1. A pie chart
2. We use a stacked bar chart if we want to highlight relative fractions

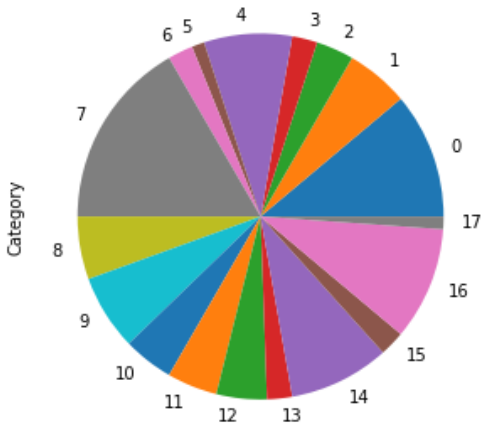


Figure 5.25: Pie Charts do not work well for many categories.

within categories.

Summary

Visualizing data is a powerful way both to explore data as well as summarize and present the findings of a detailed analysis. When we compose a particular visualization, we should adhere to best practices, in particular avoid "visual clutter", add labels and a legend to the graph where appropriate. We should also choose a consistent color scheme and complement the use of color with different drawing styles to include a wide range of audiences.

Histograms are frequently used to visualize continuous variables, in particular, when the available data set is very large. The variables are binned into fixed intervals and the choice of binning place a major role in the further analysis and the conclusions that can be drawn from a histogram. Box and violin plots are used to add quantitative metrics to the visualization and can be used, for example, in addition to a histogram.

Scatter plots allow the visual inspection of the behavior of one continuous variable as a function of another. However, for large number of data points, two-dimensional histograms or profile plots are more suitable.

Bar charts are frequently used to visualize categorical data which cannot be displayed well using other means. Pie chart on the other hand allow the visual representation of fractions as part of a whole. Both visualizations are best used if few categories or fractions are shown as they become difficult to read when used to visualized many categories or fractions.

6 Parameter Estimation

Study Goals

On completion of this unit, you will have learned:

- How to find maximum likelihood estimates by directly maximizing the log-likelihood function and the ordinary least squares (OLS) approach to parameter estimation.
- The definition of the OLS cost function and have understood how OLS is implemented to find parameter estimates for linear regression.
- How to use the expectation maximization (EM) algorithm.
- Gaussian Mixture Models and how the EM algorithm is applied to estimate its parameters with incomplete data.
- The role regularization plays in parameter estimation, how ridge and lasso perform regularization, and the specific penalty terms they add to the cost function.
- How uncertainties are encoded for a vector of random variables and be able to compute the uncertainties of a transformation (linear and non-linear) of a vector of random variables.

Introduction

One of the primary goals of statistical inference is to use the observed data to make inferences about one or more of the following:

- How the data were generated,
- The relationship between two or more variables of the observed data and theoretical values
- The relationship of two or more variables of the observed data

In each of these scenarios, an assumption is made as to either the underlying probability distribution that generated the data, or the functional form and the family of models that describe the relationship of two or more variables of the observed data. To figure out the underlying generating processes or relationships, we typically use some domain knowledge to assume the probability distribution or the model. In any of these cases, the assumptions contain unknown parameters. For example, if we have a set of data we believe came from a Gaussian distribution, we would not know the mean and standard deviation of this distribution that generated the data. If we assume that the model that relates two variables x and y is $y = t_0 + t_1 x$, we still have to determine the parameters of the model t_0 and t_1. In this unit, we present tools to estimate parameters. The different methods are used in different situations and it is important to know when each method is appropriate to apply. However, whichever tools are used for the specific problem at hand, there are some properties that we wish our estimates to strive toward. In practice, we cannot have it all, but we would like our parameter estimates to be

- unbiased. If the true parameter is denoted by a_0, the expected value of the estimated value $\hat{a}$ should be the true value, i.e. $E[\hat{a}] = a_0$, i.e., the method we use to estimate the value of the parameter we are interested in from the data we have should not introduce an additional bias. However, we should keep in mind that the variance of the estimate may be quite large, especially for a small number of data points from which we derive our estimate.
- consistent. As we add more data to the sample, the estimate should converge to the "true" value, i.e., $\lim_{n\to\infty} \hat{a} = a_0$.
- effective. The variance of the estimated parameter should be as small as possible.
- robust. The estimator should be insensitive to wrong data or any assumptions we may make.

- sufficient statistics. The estimator includes all the information in the observed data relevant to the parameter.

As we can guess from this list of desirable quantities, it will be difficult to find one method to estimate the value of a given parameter or a set of parameters that fulfills all our wishes. Instead, different approaches have their own strengths and weaknesses as well as assumptions and we need to choose the one that is most sensible in the specific case we are interested in.

In the following, we will discuss two popular methods to estimate the values of parameters, namely the maximum likelihood approach as well as the method of least squares.

6.1 Maximum Likelihood

The first method of parameter estimation we will discuss in this unit is the method of maximum likelihood. If we believe that our observed data was generated from a distribution for a variable X with probability density function $f(\cdot|\theta)$ parametrized by an unknown parameter (or a set of parameters) θ, the likelihood is the probability (distribution) of observing this data for a specific value of this parameter: $\mathcal{L}(\theta) = P(x_1, \ldots, x_N|\theta)$ where $x_1, \ldots, x_N$ represents the given data. The **likelihood** is the joint probability distribution of the observed data and describes the probability of observing the data given a specific value of θ for a specific choice of probability density function $f(\cdot|\theta)$. Note that the likelihood depends on the choice of this (unknown) parameter. If we pick a good value of this parameter, the likelihood value will be (relatively) large, and if we pick a bad value, the likelihood value will be (relatively) small. There is an optimal value: a value of $\hat{\theta}$ that maximizes the likelihood of observing the data. This optimal value of $\hat{\theta}$ is called the maximum likelihood estimate.

The Likelihood is the probability of observing a given dataset or the joint distribution of the dataset evaluated at the given data.

This means that if we want to use this method, we need to fulfill the following prerequisites:

- We need a sample of n measurements of some random variable $x_1, \ldots, x_N$, where x_i can either be a single variable or a vector of variables.

- We assume that we know the underlying probability density distribution $f(\cdot|\theta)$ but not the value of θ. This means that there is a data-generating process that can be described using a probability density distribution (PDF) or probability mass function (PMF) for discrete variables $f(\cdot|\theta)$ which maps measurements x to a number yielding the probability of sets of possible values. This function describes how the values of the measurements are distributed, and each measurement is a so-called "realization" of this PDF. The functional form of the density function depends on some parameter θ. In the maximum likelihood approach, we estimate the best numerical value of the parameter θ that maximises the probability to observe the data, but we assume that the choice of the underlying PDF $f(x|\theta)$ is correct. This implies that if we make the wrong assumption about it, meaning that we choose the wrong type of probability distribution, the result will also be wrong, even if all subsequent numerical steps and estimates of the parameter are done correctly.

Before we dive in to a more complete and formal discussion, let us look at a simple example with some data. Suppose that the number of daily accidents in a small town are independent from one day to another and follow the same Poisson distribution with unknown mean λ. We use the Poisson distribution when we want to model events that are independent from one another and all we know is the average number of events per unit time, for example, the average number of accidents per day. We observe the number of accidents from 10 days: 6, 5, 6, 1, 3, 6, 3, 3, 2, 2.

Therefore, this dataset consists of 10 realizations (or measurements) that all follow the Poisson distribution: $X_1, X_2, \ldots, X_{10} \sim \text{Poisson}(\lambda)$:

$$f(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, 2, ... \tag{6.1}$$

Since the variables are independent and originate from the same distribution, we say that they are independent and identically distributed (i.i.d.). This means that the joint PMF describing the entire data is just the product of the individual PMFs:

$$
\begin{aligned}
f(x_1, x_2, \ldots, x_{10}) &= f(x_1) f(x_2) \cdots f(x_{10}) \\
&= \prod_i \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \\
&= \frac{e^{-10\lambda} \lambda^{\sum_i x_i}}{x_1! x_2! \cdots x_{10}!} \\
&= \frac{e^{-10\lambda} \lambda^{37}}{\text{constant}}
\end{aligned}
$$

Note that this quantity depends on λ, which is the unknown average rate of accidents in this example that we want to determine. The likelihood for these data-points is given by:

$$
\mathcal{L}(\lambda) = f(x_1, x_2, ..., x_{10}|\lambda) = \frac{e^{-10\lambda} \lambda^{37}}{\text{constant}} \tag{6.2}
$$

Let us take a couple of guesses. For $\lambda = 3$ we would obtain $\mathcal{L}(3) = 1.08 \cdot 10^{-9}$. For $\lambda = 4$ we would obtain $\mathcal{L}(3) = 2.07 \cdot 10^{-9}$. Clearly,the second guess is better, it is more likely the data came from a Poisson distribution with parameter $\lambda = 4$ than the one with parameter $\lambda = 3$. After applying the full procedure, illustrated by Fig. 6.1, we will find that the optimal value of λ is 3.7:

$$
\mathcal{L}(3.7) = 2.33 \cdot 10^{-9} \tag{6.3}
$$

This optimal value is called the maximum likelihood estimate (MLE) of the parameter λ: $\hat{\lambda}^{\text{MLE}} = 3.7$.

We highlight again that this approach requires us to use the correct underlying probability distribution. In the case of this example, we used the Poisson distribution. The data themselves did not "tell" us this – we have to know this from some external domain knowledge. Note that we would be able to compute the likelihood function even if we were to use the wrong probability distribution – but the result would be wrong.

We now look at the maximum likelihood method a bit more formally. We start by defining the likelihood function: Let $(x_1, \ldots, x_N)$ be a realization of the random variables $X_1, \ldots, X_N$. Suppose that each random variable

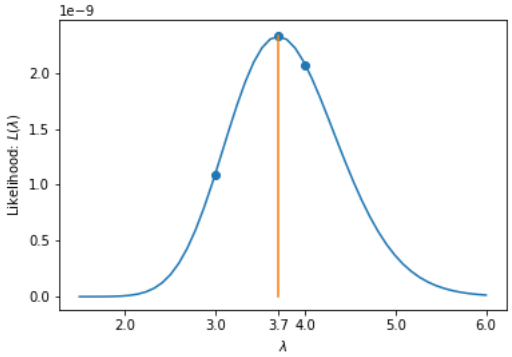


Figure 6.1: Likelihood function from a Poisson distribution, highlighting three values of the parameter λ.

X_i has PDF (PMF if they are discrete) given by $f_i(\cdot|\theta_j)$ for $i = 1, \ldots, N$. The parameters θ_j are the parameters of the distribution, each of which may be a scalar or a vector. For example, in the case of the Poisson distribution, it would be a scalar since this distribution has one parameter. In the case of the normal distribution, we might have two parameters if both the mean and standard deviation are unknown and one parameter if we know, say, the mean but not the variance (or vice versa).

The likelihood function is a function mapping the parameter to the probability of observing the given dataset.

The **likelihood function** $\mathcal{L}(\theta_1, \ldots, \theta_k)$ is defined as the mapping from the parameters to the joint density evaluated on the observed data $(x_1, x_2, \ldots, x_N)$. It describes the probability of observing the data given the parameter(s) $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$. Let $f(\cdot|\theta)$ be the joint density of the random variables $X_1, \ldots, X_N$, where $\boldsymbol{\theta}$ denotes the tuple $(\theta_i)_{i=1}^k$, i.e., all our parameters, either scalar or vector. Then the likelihood function is given by:

$$\mathcal{L}(\boldsymbol{\theta}) = f(x_1, \ldots, x_N|\boldsymbol{\theta}) \tag{6.4}$$

Assuming that the random variables $X_1, \ldots, X_N$ are independent, the joint density is given by the product and, the likelihood function in this case reduces to:

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} f_i(x_i|\theta) \tag{6.5}$$

Typically, we assume that there is only one underlying probability distribution that describes the data. Then the data are independent and identically

distributed (i.i.d.) realizations of the random variable and the likelihood simplifies further to:

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} f(x_i|\theta) \tag{6.6}$$

This means that the functions $f_i(x_i|\theta)$ are all the same so that we can drop the index i for the functions. The parameters θ can either be a single parameter or a set of parameters $\theta_1, \theta_2, \ldots, \theta_k$.

Remember

The likelihood function gives the probability of observing the given data for the parameter(s) θ, i.e., $P(x|\theta)$. It is not the probability of θ given the observed data given by $P(\theta|x)$. This distinction is crucial and has already been discussed in the context of Bayes' theorem.

In the maximum likelihood method, our goal is to find the parameter(s) θ that maximizes $\mathcal{L}(\theta)$, i.e., to maximize the probability of observing the given data by adjusting the value of the parameter(s) θ for an assumed choice of a probability distribution.

An important detail to note is that we work with a specific probability function $f(x|\theta)$ when we use the maximum likelihood approach to find the best value of the parameter θ. This implies that $f(x|\theta)$ has to be normalized at all stages of the procedure, i.e., $\int f(x|\theta)dx = 1$ for all values of θ. Although this may sound trivial, a sizeable part of the numerical implementation of maximum likelihood estimation is concerned with making sure that this normalization is fulfilled at all times.

In most practical applications, we do not work with the likelihood function directly but use the log-likelihood function. Composing a function with a logarithm allows to find the same maximum as the logarithm is monotonic and increasing. Also, the logarithm of a product is the sum of the logarithms and often easier to handle in practice. Additionally, the products of densities (or probabilities) can get very small, smaller than the minimum value of the data type that a computer can handle. Recall that the natural logarithm maps the values $x \mapsto \log(x)$ one-to-one, meaning that for each value of x there is a unique value of $\log(x)$. Furthermore, the logarithm is monotonously increasing for positive numbers x. Therefore, the maximizer(s) of the likelihood function is(are) also the maximizer(s) of

The Log-likelihood function is used for analytic and computational efficiencies over the likelihood function, this function is a logarithm of the likelihood function.

the **log-likelihood** function given by

$$\ell(\boldsymbol{\theta}) = \log \mathcal{L}(\boldsymbol{\theta}). \tag{6.7}$$

In the case of independent random variables, this reduces to

$$\ell(\theta) = \sum_{i=1}^{N} \log f_i(x_i|\boldsymbol{\theta}). \tag{6.8}$$

and, again, for i.i.d. distributed variables we can drop the index at the function, i.e. $f(x_i|\boldsymbol{\theta})$.

Closely related to this quantity is the **negative log-likelihood function** given by

$$n\ell\ell(\boldsymbol{\theta}) = -\sum_{i=1}^{N} \log f_i(x_i|\boldsymbol{\theta}). \tag{6.9}$$

The negative log-likelihood function is used over the likelihood because optimization schemes tend to be programmed to minimize rather than maximize.

It is quite straightforward to see that any maximizer of the likelihood (log-likelihood) function is a minimizer of the negative log-likelihood function. Since most numerical algorithms for optimization are designed to minimize functions, this is the quantity used in practice.

Let us now assume that in addition to independence, the random variables from the observations assumed to be drawn from are identically distributed, i.e., $X_1, \ldots, X_N$ are i.i.d. with common a PDF (or PMF) given by $f(\cdot|\theta) = f_1(\cdot|\theta) = \ldots = f_N(\cdot|\theta)$. Note that in this setting, θ is just one set of parameters since we only have a single distribution. Consequently, we can drop the subscript for the distributions. In this setting, the likelihood, log-likelihood, and negative log-likelihood functions are given by

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{N} f_i(x_i|\boldsymbol{\theta}) \tag{6.10}$$

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log f(x_i|\boldsymbol{\theta}) \tag{6.11}$$

$$n\ell\ell(\boldsymbol{\theta}) = -\sum_{i=1}^{N} \log f(x_i|\boldsymbol{\theta}) \tag{6.12}$$

In Fig. 6.2 we illustrate the behavior of the log-likelihood and the negative log-likelihood functions from the example of the accidents per day discussed

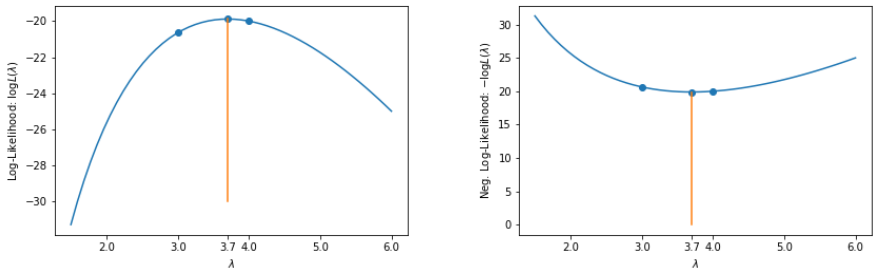


Figure 6.2: Log-likelihood and negative log-likelihood for the Poisson example.

earlier in Fig. 6.1. Note that the maximizer of the likelihood ($\lambda = 3.7$) is also the maximizer of the log-likelihood as well as the minimizer of the negative log-likelihood.

Example: MLE of a Coin Toss

Suppose that a (possibly biased) coin is tossed 10 times and we observe seven heads. What is the maximum likelihood estimate of the probability of heads?

Solution:
Each coin toss can be seen as a Bernoulli trial, hence we have 10 independent observations from a Bernoulli distribution, where success is defined to be the event of observing heads. Therefore, we have $X_1, \ldots, X_N \sim \text{Bernoulli}(\theta)$ with observed data $1, 1, 1, 1, 1, 1, 1, 0, 0, 0$ (or some permutation of these). Recall that the PMF is given by

$$f(x; \theta) = \theta^x (1 - \theta)^{1-x}$$

for $x = 0, 1$ and zero otherwise. Therefore, the likelihood of the data is

$$\begin{aligned} \mathcal{L}(\theta) &= \prod_{i=1}^{10} \theta^{x_i} (1 - \theta)^{1-x_i} \\ &= \theta^{\sum_{i=1}^{N} x_i} (1 - \theta)^{10 - \sum_{i=1}^{10} x_i} \\ &= \theta^7 (1 - \theta)^3 \end{aligned}$$

Since this function is continuous for $\theta \in [0, 1]$, the extreme value theorem guarantees a maximizer. Furthermore, since this function is

differentiable on (0,1), we can use derivatives to find the maximizer

$$\begin{aligned}\frac{d\mathcal{L}}{d\theta} &= 7\theta^6(1-\theta)^3 - 3\theta^7(1-\theta)^2 \\ &= \theta^6(1-\theta)^2[7(1-\theta) - 3\theta] \\ &= \theta^6(1-\theta)^2(7-10\theta)\end{aligned}$$

This equation has a single solution in (0,1) at $\hat{\theta}^{MLE} = \frac{7}{10}$. This is indeed a maximizer of the likelihood function (e.g., the second derivative is negative here).

Note that we do not consider the values $\theta = 0$ and $\theta = 1$. While they also lead to zero in the first derivative, either value would mean that we always observe head or tail and this formal solution is clearly not what we observe in the data. The result should not be surprising. In fact, frequentist statistics would tell us that since we have observed seven heads out of 10 tosses, the probability of heads must be 7/10.

For computational demonstration, let us compute the maximizer using the log-likelihood function

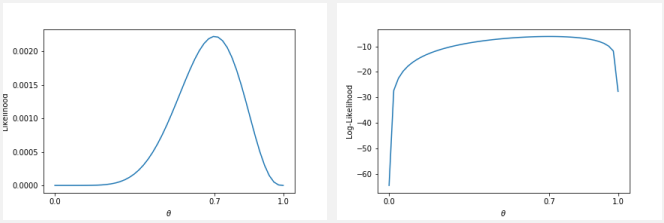
$$\ell(\theta) = 7\log\theta + 3\log(1-\theta)$$

Differentiating this equation gives

$$\ell'(\theta) = \frac{7}{\theta} - \frac{3}{1-\theta} = \frac{7-10\theta}{\theta(1-\theta)}.$$

As before, the only solution is, of course, $\hat{\theta}^{MLE} = \frac{7}{10}$. Even in this very simple example, working with the log-likelihood is much easier than working directly with the likelihood function.

The values of the likelihood and log-likelihood functions can be visualized like this:



Both the likelihood and log-likelihood functions are both are maximized at the MLE estimate. Furthermore, notice that the log-likelihood function is strictly concave, meaning that there is only one global maximum.

The fact that the likelihood and log-likelihood functions in the example

above were concave is an important property, in particular when numerical algorithms are used to optimize a function. If a maximizer is required, we would like the objective function to be concave, otherwise, the algorithm might converge to a local maximum instead of the global one. In practice, if we do not have this shape for the log-likelihood, we would apply a transformation to the observed data and start over. A function is concave if its second derivative is negative.

Before looking at another example, we should note that the likelihood function, and therefore the log-likelihood as well as its maximizer, are random variables. This is because they are functions of the observed data, which are themselves random variables. Therefore, the MLE estimator, θ^{MLE} is a function of the random variables $X_1, X_2, \ldots, X_N$,

For large samples (large N), this MLE estimator follows an approximately Gaussian distribution with mean θ^{true}, the true value of the parameter, and variance-covariance matrix given by

$$[I(\theta^{\text{true}})]^{-1} \tag{6.13}$$

where $I(\theta^{\text{true}})$, the information matrix, is the expected value of the matrix of second order derivatives (Hessian) of the negative log-likelihood function:

$$I(\theta^{\text{true}}) = E\begin{bmatrix} \frac{\partial^2 n\ell\ell(\boldsymbol{\theta})}{\partial\theta_1^2} & \frac{\partial^2 n\ell\ell(\boldsymbol{\theta})}{\partial\theta_1\,\partial\theta_2} & \cdots & \frac{\partial^2 n\ell\ell(\boldsymbol{\theta})}{\partial\theta_1\,\partial\theta_n} \\ \frac{\partial^2 n\ell\ell(\boldsymbol{\theta})}{\partial\theta_2\,\partial\theta_1} & \frac{\partial^2 n\ell\ell(\boldsymbol{\theta})}{\partial\theta_2^2} & \cdots & \frac{\partial^2 n\ell\ell(\boldsymbol{\theta})}{\partial\theta_2\,\partial\theta_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 n\ell\ell(\boldsymbol{\theta})}{\partial\theta_n\,\partial\theta_1} & \frac{\partial^2 n\ell\ell(\boldsymbol{\theta})}{\partial\theta_n\,\partial\theta_2} & \cdots & \frac{\partial^2 n\ell\ell(\boldsymbol{\theta})}{\partial\theta_n^2} \end{bmatrix} \tag{6.14}$$

The information matrix captures information about the curvature of the negative log-likelihood function. Conceptually, the sharper the low point, the less variance the MLE estimator has. Consequently, the more variance in the MLE estimator results in a flatter low point. In the case where there is only one parameter, the variance of the estimator reduces to

$$\sigma^2_{\theta^{\text{MLE}}} = \frac{1}{\frac{\partial^2 n\ell\ell}{\partial\theta^2}\big|_{\theta^{\text{MLE}}}} \tag{6.15}$$

We can understand this more intuitively if we consider only one parameter θ. The negative log-likelihood function is then given by $n\ell\ell(\theta)$ and the

optimal parameter is given by $\hat{\theta}$. Expanding the negative log-likelihood around the optimal parameter yields:

$$n\ell\ell(\theta) = n\ell\ell(\hat{\theta}) + \frac{1}{2}\frac{d^2 n\ell\ell(\theta)}{d\theta^2}(\theta - \hat{\theta})^2 + \cdots \tag{6.16}$$

The negative log-likelihood function approximates the shape of a parabola, Equivalently, the likelihood function approximates a normal distribution:

$$\mathcal{L} \approx \text{const.} \cdot \exp\left[-\frac{1}{2}\frac{d^2 n\ell\ell(\theta)}{d\theta^2}(\theta - \hat{\theta})^2\right] \tag{6.17}$$

$$= \text{const.} \cdot \exp\left[-\frac{(\theta - \hat{\theta})^2}{2\sigma^2}\right] \tag{6.18}$$

where we identify (as in Eqn. (6.15))

$$\sigma = \left[\frac{d^2 n\ell\ell(\theta)}{d\theta^2}\bigg|_{\theta=\hat{\theta}}\right]^{-1/2} \tag{6.19}$$

We can then use this to define the uncertainty of the on the optimal parameter $\hat{\theta}$ in analogy to the variance of the normal distribution. At the minimum: $\theta = \hat{\theta} \pm n \cdot \sigma$ and for the negative negative log-likelihood function: $n\ell\ell(\theta) = n\ell\ell(\hat{\theta}) + \frac{1}{2}n^2$. Note that if the log-likelihood function does not approximate a parabola, we can usually find a suitable variable transformation. The case of multiple parameters is analogous to the above and we need to use the information matrix defined in Eqn. (6.14) instead of the second derivative.

Let us return to our Poisson example and compute the variance. Recall that the likelihood is:

$$\mathcal{L}(\lambda) = f(x_1, x_2, ..., x_{10}|\lambda) = \frac{e^{-10\lambda}\lambda^{37}}{\text{constant}}. \tag{6.20}$$

The negative log-likelihood is given by:

$$n\ell\ell(\lambda) = 10\lambda - 37\log\lambda + \log(\text{large number}). \tag{6.21}$$

Its derivatives are

$$n\ell\ell'(\lambda) = 10 - \frac{37}{\lambda} \quad \text{and} \quad n\ell\ell''(\lambda) = \frac{37}{\lambda^2} \tag{6.22}$$

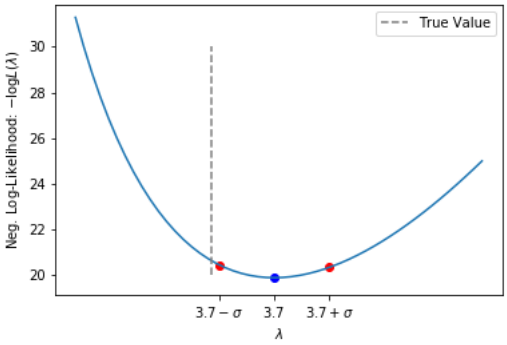


Figure 6.3: Negative Log-Likelihood and $1 \pm \sigma$ intervals (Poisson Example).

Therefore, the variance of the estimator is given by:

$$\sigma^2 = \frac{1}{\frac{37}{3.7^2}} = 0.37, \tag{6.23}$$

and the standard deviation is $\sigma = 0.61$ as illustrated in Fig. 6.3

MLE for the mean of a Gaussian Function

Suppose that $x_1, \ldots, x_N$ are i.i.d. observations from a normal distribution with unknown mean μ and unknown standard deviation σ. Find the maximum likelihood estimate of mean μ by minimizing the negative log-likelihood function.

Solution
Recall the density of the normal distribution $X \sim \mathcal{N}(\mu, \sigma)$ is

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

Therefore, the log-likelihood function for the observed data is

$$\begin{aligned}
\ell(\mu,\sigma) &= \sum_{i=1}^{N} \log\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma}\right)\right] \\
&= \sum_{i=1}^{N}\left[\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \log\left(\exp\left(-\frac{(x-\mu)^2}{2\sigma}\right)\right)\right] \\
&= \sum_{i=1}^{N}\left[\log(1) - \log\left(\sqrt{2\pi\sigma^2}\right) + \log\left(\exp\left(-\frac{(x-\mu)^2}{2\sigma}\right)\right)\right] \\
&= -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2
\end{aligned}$$

The negative log-likelihood function is

$$n\ell\ell(\mu,\sigma) = K + \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2,$$

where the term not including the variable μ is a constant (denoted by K). Differentiating this with respect to μ gives

$$\begin{aligned}
\frac{\partial n\ell\ell(\mu,\sigma)}{\partial\mu} &= -\frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i-\mu) \\
&= -\frac{1}{\sigma^2}\left[\sum_{i=1}^{N}x_i - \sum_{i=1}^{N}\mu\right] \\
&= -\frac{1}{\sigma^2}\left[\sum_{i=1}^{N}x_i - N\mu\right]
\end{aligned}$$

Remember that when calculating the partial derivative for μ we treat the other variable, in our case σ as a constant. In order to find the maximum likelihood estimate, we require the first derivative to vanish and the result is just the sample mean:

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N}x_i.$$

This estimator is consistent. To understand this a bit better we use the weak law of large numbers. This states if we have a sequence of i.i.d. random variables $X_1, X_2, \ldots, X_n$, each of which has the mean $\bar{X}_i = E[X_i] = \mu$, and we define a new random variable

$$X = \frac{X_1 + X_2 + \cdots + X_n}{n} \tag{6.24}$$

then for $n \to \infty$, the sample mean $\bar{x}$ approaches the population mean, i.e., $\bar{X} = \mu$. We denote $\bar{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$ We know that $E[\bar{X}_N] = \mu$ and for any positive ϵ, the weak law of large numbers assures us that

$$\lim_{n\to\infty} P(|\bar{X}_N - \mu| > \epsilon) = 1. \tag{6.25}$$

Then for any $t > 0$ (no matter how small) (Khinchina, 1929),

$$\lim_{n\to\infty} P(|\bar{X} - \mu| > t) = 0. \tag{6.26}$$

We can use the same argument for the variance. Remembering that $V[ax] = a^2V[x]$

$$\begin{aligned} V[\bar{X}] &= \frac{V[X_1 + X_2 + \cdots + X_n]}{n^2} \\ &= \frac{V[X_1] + V[X_n] + \cdots + V[X_n]}{n^2} \quad X_i \text{ are independent} \\ &= \frac{nV[X]}{n^2} \\ &= \frac{V[X]}{n} \end{aligned} \tag{6.27}$$

It is important to remember that although the MLE estimate of the mean from the example above turned out to be unbiased, this is not an intrinsic property of MLE estimators in general.

Note

It is important to understand that $\hat{\mu}^{\text{MLE}}$ which is a (non-random) quantity based on the observed data $(x_i)_{i=1}^N$ while the corresponding quantity $\bar{X}_N$ is a random variable based on the sequence of random variable $(X_i)_{i=1}^N$. To explore the properties of an MLE estimator, we

use the latter. To compute the estimate for an observed sample, we use the former.

MLE for the variance of a Gaussian Function

Suppose $x_1, \ldots, x_N$ are i.i.d. observations from the normal distribution with unknown mean μ and unknown standard deviation σ. Find the MLE estimator of σ by minimizing the negative log-likelihood function of the precision $\beta = \frac{1}{\sigma^2}$

Solution:
We have already calculated the log-likelihood function for a Gaussian distribution in the previous example where we looked at the mean:

$$\ell(\mu, \sigma) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2$$

We can express this in terms of β instead of σ as:

$$\ell(\mu, \sigma) = -\frac{N}{2}\log(2\pi) + \frac{N}{2}\log\beta - \frac{\beta}{2}\sum_{i=1}^{N}(x_i - \mu)^2$$

where we remember that $\log(ab) = \log(a) + \log(b)$. The negative log-likelihood in terms of μ and β is given by

$$n\ell\ell(\mu, \beta) = K - \frac{N}{2}\log\beta + \frac{\beta}{2}\sum_{i=1}^{N}(x_i - \mu)^2$$

We now compute the partial derivative with respect to β (treating μ as a constant) and obtain:

$$\frac{\partial n\ell\ell(\mu, \beta)}{\partial\beta} = -\frac{N}{2\beta} + \frac{1}{2}\sum_{i=1}^{N}(x_i - \mu)^2.$$

As before, we require this first derivative to vanish for the maximum-likelihood estimator. Therefore, the result is given by:

$$\hat{\sigma}^2 = \frac{1}{\hat{\beta}} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})^2.$$

As in the case of the mean, the maximum likelihood estimator of the standard deviation is the sample standard deviation for a Gaussian distribution with respect to the maximum likelihood estimator for the sample mean $\hat{\mu}$.

However, unlike the estimator for the mean, the maximum likelihood estimator for the variance is biased. We can construct an unbiased estimator as (Bishop, 2006, p. 27):

$$\tilde{\sigma}^2 = \frac{N}{N-1}\hat{\sigma}^2_{\text{MLE}} = \frac{N}{N-1}\sum_{i=1}^{N}(x_i - \hat{\mu})^2. \tag{6.28}$$

Self-Check Questions

1. True or False. The likelihood function is the probability of observing the parameter from a given sample.

2. True or False. Under the assumption that the observed data are independent, the likelihood function is the product of the densities (or PMFs).

3. Fill in the blanks. The maximizer of the likelihood function is the ... of the log-likelihood function and the ... of the negative log-likelihood function.

4. Fill in the blanks. For independent normally distributed data, the maximum likelihood estimator for the sample mean is ... and the maximum likelihood estimator with respect to the maximum likelihood estimator of the mean for the variance is ....

Solutions

1. False

2. True

3. Maximizer, minimizer

4. Unbiased, biased

6.2 Ordinary Least Squares (OLS)

In the case of maximum likelihood discussed earlier, one of the crucial assumptions we made was that we know the underlying probability distribution $f(x|\theta)$, from which all observed data points originate, i.e., the data are concrete realizations of a random variable described by $f(x|\theta)$. However, in many practical cases we do not know $f(x|\theta)$ – but we may have a model that can be used to describe the data. This can be an empirical model or a theoretic description of some stochastic process we wish to understand. The crucial point is that while we do not know the underlying probability distribution, we do have a method to generate or predict values from some model that we can compare to the observed data. In fact, we do not need to be able to make such predictions for any data point, just the ones we observe in our measurements. Since such a model will in general also depend on some parameter(s) θ, we need to determine the value of these parameters from the data to tune our model or prediction. This can be done for example using the method of ordinary least squares (OLS) that requires only the functional dependence of two observed variables: the independent and dependent variables. The independent variables are those that are, for example, related to measurements or experimental quantities that we can manipulate. The dependent variables depend on those variables and we expect them to change if we manipulate the independent variables. Typically, we denote the independent variables as X and the dependent variable as Y.

Suppose that we have observed the data $(\boldsymbol{X}, \boldsymbol{Y}) = ((x_1, y_1), \ldots, (x_N, y_N))$ and assume that

$$y_i = h_\theta(x_i) \tag{6.29}$$

for $i \in 1, \ldots, N$. Here, the x_i are the values of the observed data points for the independent variable. The values of y_i are the observed outcomes for the dependent variable we wish to model. The function $h_\theta(x_i)$ is the model that depends on one or more parameters θ. For a fixed value of the parameter(s) $\theta = \hat{\theta}$, we define the prediction as $\hat{y}_i = h(x_i|\hat{\theta})$ and follow the

convention that estimates or predictions are denoted with a little "hat". The error or residual of the i^{th} prediction is

$$r_i = y_i - \hat{y}_i \tag{6.30}$$

The method of ordinary least squares minimizes the sum of squares of the residuals:

$$\tilde{C}(\hat{\theta}) = \sum_{i=1}^{N} r_i^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2. \tag{6.31}$$

In case we need to compare the cost with various sample sizes, it is typically better to use the following "average" cost function:

$$C(\hat{\theta}) = \frac{1}{2N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2. \tag{6.32}$$

The uncertainty associated with a value quantifies our knowledge about its precision.

Even if we assume that the model is correct, we must account for measurement errors. We introduce this **uncertainty** in the simplest case by

$$\text{Var}[Y_i|X_i] = \sigma^2 \tag{6.33}$$

for $i \in 1, \ldots, N$. In terms of random variables, $Y_1|X_1, \ldots, Y_N|X_N$ all have the same unknown distribution and

$$\boldsymbol{Y} = h_\theta(\boldsymbol{X}) + \boldsymbol{\epsilon} \tag{6.34}$$

where $E[\epsilon] = 0$ and $Var[\epsilon] = \sigma^2$.

When each data point has a different associated uncertainty, say $\text{Var}[X_i] = \sigma_i^2$, the least squares cost function becomes

$$\sum_{i=1}^{N} \frac{(y_i - \hat{y}_i)^2}{\sigma_i^2} \tag{6.35}$$

Additionally, if the measurements are correlated, then the cost function is adjusted to contain the covariance matrix to account for the correlation:

$$(\boldsymbol{y} - \hat{\boldsymbol{y}})^{\boldsymbol{T}} [\mathbf{Cov}(\boldsymbol{X})]^{-\mathbf{1}} (\boldsymbol{y} - \hat{\boldsymbol{y}}). \tag{6.36}$$

Note that Eqn. (6.36) reduces to Eqn. (6.35) for uncorrelated measurements since then the (inverse) covariance matrix is diagonal. In general, our model may have an uncertainty as well that needs to be included in the same way as the uncertainties from the observed data, for example, to account for uncertainties in the theoretical description.

Linear Regressions and OLS

To demonstrate how the OLS estimates the parameters of the assumed model, we will explore this in the specific case of linear regression. In this case, the function $f(\cdot|\theta)$ is linear in the parameters θ. When the random variable X has a single dimension, the problem is known as simple linear regression. When the random variable X is a vector, the problem is known as multiple linear regression.

Suppose that we have the set of observed data $((x_i, y_i))_{i=1}^N$ and assume the functional family $y_i = f(c_i|\theta) = \theta \cdot x_i$ to model this data. We will find the (ordinary least squares) OLS estimate of θ. In this setting, the predictions are $\hat{y}_i = \theta x_i$, so the cost function to minimize is:

$$C(\theta) = \frac{1}{2N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = \frac{1}{2N}\sum_{i=1}^{N}(y_i - \theta x_i)^2. \tag{6.37}$$

Since we want to find the minimum of the cost function, we differentiate Eqn. (6.37) and obtain:

$$\begin{aligned}\frac{dC}{d\theta} &= \frac{1}{N}\sum_{i=1}^{N}(y_i - \theta x_i)(-x_i) \\ &= -\frac{1}{N}\sum_{i=1}^{N} x_i y_i + \frac{\theta}{N}\sum_{i=1}^{N} x_i^2.\end{aligned} \tag{6.38}$$

To find the minimum, we require that the first derivative vanishes and obtain:

$$\hat{\theta}^{\mathrm{OLS}} = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2} \tag{6.39}$$

To evaluate this estimator, we look at the corresponding random variable

$$\theta = \frac{\sum_{i=1}^{N} X_i Y_i}{\sum_{i=1}^{N} X_i^2} \tag{6.40}$$

Given the observations from X, we have that

$$E[\Theta|X_1, \ldots, X_N] = \frac{\sum_{i=1}^{N} X_i E[Y_i]}{\sum_{i=1}^{N} X_i^2} = \frac{\sum_{i=1}^{N} X_i \theta X_i}{\sum_{i=1}^{N} X_i^2} = \theta \tag{6.41}$$

Therefore, the estimator is unbiased. Next, let us look at the variance of this estimator:

$$Var[\Theta|X_1, \ldots, X_N] = \frac{\sum_{i=1}^{N} X_i^2 \text{Var}[Y_i]}{\left(\sum_{i=1}^{N} X_i^2\right)^2} = \frac{\sigma^2}{\sum_{i=1}^{N} X_i^2} \tag{6.42}$$

The approach given above can be extended if we want to consider more than one independent variable. We can use the same approach but replace x with $\boldsymbol{x}$ and y with $\boldsymbol{y}$ to indicate that we now analyze several variables. For further details, see, for example, Wasserman (2013, p. 254ff.).

Self-Check Questions

1. True or False. Ordinary least squares can only be applied to linear regression problems.
2. Fill in the blank. The OLS estimate for the parameters of a linear regression model is ....
3. True or False. The method of OLS requires that the distribution of the variables be specified.
4. True or False. The cost function associated with OLS can always be solved using analytical methods.

Solutions

1. False
2. Unbiased
3. False
4. False

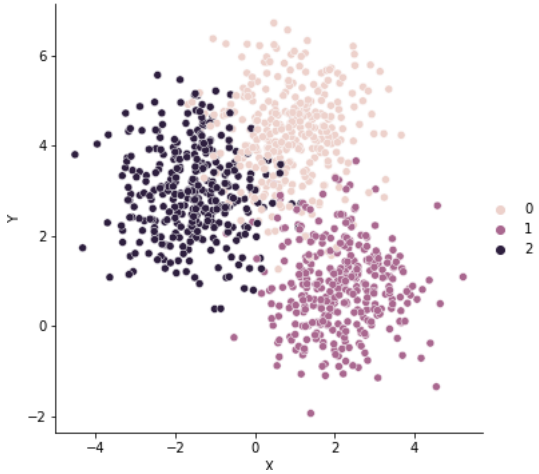


Figure 6.4: Data generated according to three Gaussian distributions. The color indicates the true function the point originates from.

6.3 Expectation Maximization (EM)

Gaussian Mixture Models

In many situation we want to represent data in a meaningful way. If we know the underlying probability distributions, we can use these to describe the data if we can determine the relevant parameters.

However, if we look at Fig. 6.4, we can see that already in this simple case it is not obvious how to parameterize the data. In this figure, we generated the data using three different Gaussian or normal distributions and used the color to highlight which data-point comes from which function - however, in general we will neither know that the underlying distribution is a Gaussian nor which part of the data is from which distribution.

We can, however, build an empirical model that describes the data, these are called "mixture models". Intuitively, we create a distribution with density function $p(x)$ by adding multiple density functions together (Deisenroth, Faisal, & Ong, 2020, p. 349):

$$p(x) = \sum_{k=1}^{K} \pi_k p_k(x) \tag{6.43}$$

where each component $p_k(x)$ contributes according to some factor π_k called "mixture weights". These mixture weights lie in the interval $0 \leqslant \pi_k \leqslant 1$ and are normalized, i.e., $\sum_{k=1}^{N} \pi_k = 1$. The components $p_k(x)$ can be any distribution, however, in most cases we will use a Gaussian (or normal) distribution and limit the further discussion to this case. Such a model is therefore called a "Gaussian mixture model":

$$p(x|\theta) = \sum_{k=1}^{K} \pi_k f_{\mu_k,\sigma_k}(x) \tag{6.44}$$

The quantity $\theta = \{(\mu_k, \sigma_k, \pi_k) : k = 1, \ldots, K\}$ refers to the collection of all the parameters we use in the model, i.e., the K means μ_k and covariances σ_k of the normal distribution $\sim \mathcal{N}(\mu_k, \sigma_k)$, as well as the mixture weights π_k. If we want to determine the best way to describe some given dataset, we need to estimate these parameters to define our Gaussian mixture model.

We denote the dataset as $X = (x_1, \ldots, x_n)$ with $n = 1, \ldots, N$ and assume that these are i.i.d. according to some function $p(x)$ (Deisenroth et al., 2020, p. 350). Following our previous discussion, we use the maximum likelihood method to estimate the parameters. Since the data are independent of each other, the likelihood function is given by the product of the individual components:

$$p(X|\theta) = \prod_{n=1}^{N} p(x_n|\theta)$$
$$p(x_n|\theta) = \sum_{k=1}^{K} \pi_k f_{\mu_k,\sigma_k}(x_n)$$

This means that each data point is described by the distribution $p(x_n|\theta)$, which in turn is a sum of K Gaussian distributions, the Gaussian mixture. The log-likelihood is then given by (Deisenroth et al., 2020, p. 351):

$$\mathcal{L} = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k f_{\mu_k,\sigma_k}(x) \tag{6.45}$$

In the examples we have encountered so far, we could find an analytic solution to the best maximum likelihood estimator: We computed the derivative of the log-likelihood method, set it to zero and then solved for the parameter(s). Unfortunately, in the case of the mixture model, we cannot do this easily, because the second summation is inside the logarithm.

The intuitive idea behind the expectation-maximization (EM) algorithm is to use an iterative approach to find the best maximum likelihood estimate for the mean, the covariance and the mixture weights. The update for the mean $\mu_k, k = 1, \ldots, K$ is given by:

$$\mu_k^{\text{new}} = \frac{\sum_{n=1}^{N} r_{nk} x_n}{\sum_{n=1}^{N} r_k} \tag{6.46}$$

The proof can be found in Deisenroth et al. (2020, p. 353ff). The factors r_{nk} are called "responsibilities" and are defined as (Deisenroth et al., 2020, p. 350):

$$r_{nk} = \frac{\pi_k f_{\mu_k, \sigma_k}(x_n)}{\sum_{j=1}^{K} \pi_j f_{\mu_k, \sigma_k}(x_n)} \tag{6.47}$$

Essentially, the responsibilities describe the likelihood that a given data point originates from a specific component in the mixture model. The responsibilities are normalized due to the choice of the denominator, i.e. $\sum_k r_{nk} = 1$ and $r_{nk} \geqslant 0$. The update formula for the mean Eqn. (6.46) "hides" the complexity that we cannot compute the updated values for the means directly. This is because the responsibilities r_{nk} depend on all means, covariances and mixture weights of the complete model.

Similarly, the update rule for the covariances $\sigma_k, k = 1, \ldots, K$ is given by:

$$\sigma_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (x_n - \mu_k)(x_n - \mu_k)^T \tag{6.48}$$

where r_{nk} are again the responsibilities and N_k is defined as

$$N_k = \sum_{n=1}^{K} r_{nk} \tag{6.49}$$

The proof can be found in Deisenroth et al. (2020, p. 356ff). Finally, we need the update rule for the mixture weights:

$$\pi_k^{\text{new}} = \frac{N_k}{N} \quad k = 1, \ldots, K \tag{6.50}$$

The proof can be found in Deisenroth et al. (2020, p. 358).

Expectation Maximization Algorithm

As we have stated earlier, we cannot determine the updated values of the parameters of our mixture models directly, namely the mean, covariances and mixture weights. However, we can do this iteratively using the Expectation Maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977). A proof of the convergence properties of the algorithm can be found in Wu (1983). This approach consists of two steps (Deisenroth et al., 2020, p.361):

- E-step: Evaluate the responsibilities r_{nk} that determine the probability that a given data point belongs to component k.
- M-step: Use these updated responsibilities to estimate the model parameters μ_k, σ_k, π_k.

The algorithm starts by choosing some initial values for μ_k, σ_k, π_k and then alternates between the E-step and the M-step until the procedure converges. Our description of the expectation maximisation is done on an example: consider the data shown in Fig. 6.4, The data are created using random numbers and we generate three different but overlapping distributions that are shown in the right part of the figure. The colors indicate which cluster each generated data point belongs to. Using the EM algorithm we obtain the assignments shown in the left part of Fig. 6.5, if we assume that the data can be described by three components. Comparing this to the right part of the figure, we can see that we obtain a reasonable description of our generated dataset.

However, in reality we will not know that there were three Gaussian components to begin with. For example, we might guess that there were four components instead as shown in Fig. 6.6. In this case, the algorithm still describes the data well, but we now have four components. The number of components is therefore a parameter that we need to tune if we use the EM algorithm with a Gaussian mixture model.

Self-Check Questions

1. True or False: Convergence is guaranteed for the EM algorithm.

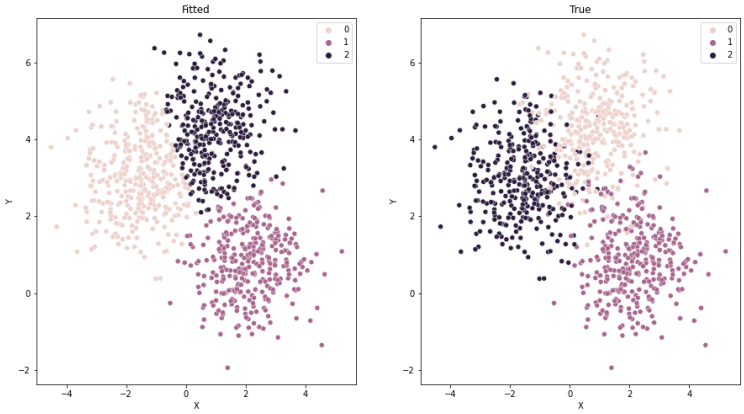


Figure 6.5: Gaussian mixture model with three Gaussian components. Left: fitted distribution after the EM Algorithm converges, right: true origins of the data points.

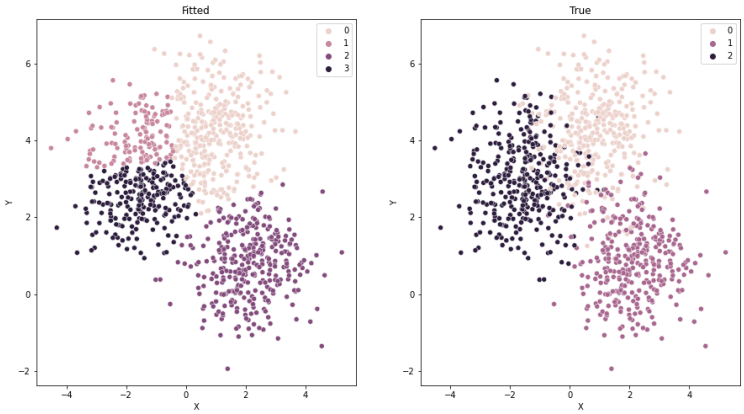


Figure 6.6: Gaussian mixture model with four Gaussian components. Left: fitted distribution after the EM Algorithm converges, right: true origins of the data points.

2. What do the responsibilities express?

3. True or False: The mixture model can only be used with Gaussian distributions.

Solutions

1. False

2. Essentially, the responsibilities describe the likelihood that a given data point originates from a specific component in the mixture model.

3. False, Gaussian distributions are the most commonly used ones but we can use others.

6.4 Lasso and Ridge Regularization

When we discussed the least squares method before, we made no assumption on the model we use in the method and only looked at the residuals between the prediction of the model and the corresponding data point. The least squares method then worked by minimizing the sum of all the squared residuals. While the method may be able to reproduce the observed data perfectly, we may end up in a situation where we do not trust the model in the sense that it can reproduce the data but is not a good description of the data-generating process.

This can be illustrated with the example shown in Fig. 6.7: Suppose we have 10 data points and choose to fit an 11-degree polynomial as our model to the data. Using OLS, without any further constraints, we determine the coefficients of each term in the model. After completing the procedure, we find that each data point is described perfectly, but the resulting curve is quite "wiggly" in the sense that it has high variance. We would therefore have some doubts that this model really describes the data generating process and intuitively we would expect that the resulting curves are more smooth (with a lower variance), even if that means that we cannot describe each measured data point perfectly anymore. In reality, each measurement is associated with an uncertainty, and we therefore only expect that the

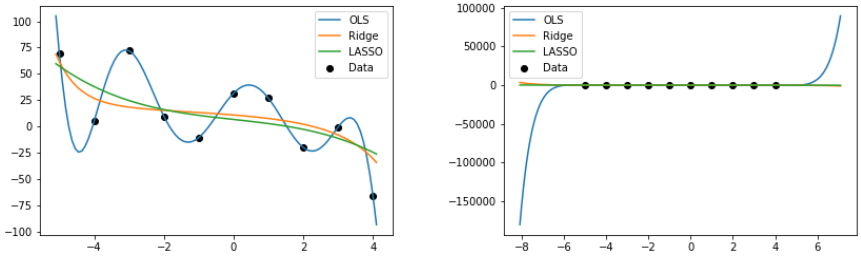


Figure 6.7: Dataset fitted to an 11-Degree Polynomial Using OLS without regularization and with regularization (ridge and lasso). Note the different scale on the y-axis for the plots.

observed data are compatible with the model to some level of the associated uncertainty.

One way to achieve this is to choose a different model, but we can also add a penalty term to the optimization procedure that prevents the "wiggly" behavior we have noted above. This is called "regularization" and can be used to make a model more robust and help to prevent overfitting. We can see in Fig. 6.7 that these regularization techniques can make the curve much smoother and the result is much closer to what we expect than with the unregularized least squares fit. In the following, we will discuss two commonly used regularization techniques, lasso (least absolute shrinkage and selection operator) (Tibshirani, 1996) and ridge (Hoerl & Kennard, 1970) regularization.

Although we will apply these regularization methods to the linear regression in the following the general approach can be extended to many other applications. In linear regression we aim to find the best parameter such that the linear model $\hat{y} = \sum_i a_i x_i$ describes the observed data well. Using the method of least squares, we can write this as:

$$\hat{\boldsymbol{a}} = \arg\min_{\boldsymbol{a}} \left[\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \right] \tag{6.51}$$

$$= \arg\min_{\boldsymbol{a}} \left[\sum_{i=1}^{N} \left(y_i - \sum_{k=1}^{K} a_k x_{i,k} \right)^2 \right] \tag{6.52}$$

where the index i runs over all data points in our data set and the index j over all coefficients in our regression model, up to an order K. The term

x	-5	-4	-3	-2	-1	0	1	2	3	4
y	69.1	5.3	72.0	9.6	-10.6	31.6	26.8	-20.1	-0.5	-66.3

Table 6.1: Dataset used for the linear regression example.

$\arg\min_a$ indicates the least squares method formulated as an optimization problem: we seek the values of the parameter(s) a that minimize the least squares cost function, namely the square of the residuals between the observed data and the prediction of the linear regression model. The intuition behind the regularization approach discussed here is to add a penalty term such that the optimization task is now:

$$\hat{\boldsymbol{a}} = \arg\min_{\boldsymbol{a}} \left[\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \text{penalty}(\boldsymbol{a}) \right] \tag{6.53}$$

Intuitively, this means that we no longer seek the parameters a such that the model reproduces the data well but we aim to find the optimal parameters such that the data is described well and the penalty term is considered. In other words, we minimize the joint cost function of the model and the penalty. If the penalty is very weak, the result will be close to the original least squares result without regularization, if the regularization is very strong, the influence the data have on the model will vanish. Choosing the "right" level of regularization is therefore critical to obtaining the best result. The two commonly used regularizations are:

$$\text{penalty}(a) = \lambda \sum_{k=1}^{K} a_k^2 \quad \text{(Ridge)} \tag{6.54}$$

$$\text{penalty}(a) = \lambda \sum_{k=1}^{K} |a_k| \quad \text{(Lasso)} \tag{6.55}$$

Each regularization scheme introduces a new free parameter λ which determines the strength of the regularization. For $\lambda = 0$ we obtain the unregularized case, if λ is very large, the penalty term dominates over the actual model we want to use to describe the data. We can also combine both penalties, this approach is commonly called "elastic net".

As a simple example we can look at the effect of the regularization using the data in Tab. 6.1. Computing the coefficients a_j for the regularized and unregularized case, we obtain the result summarized in Tab. 6.2. By

	OLS	Ridge	Lasso
1	31.61	10.86	6.62
x	33.53	-2.69	-3.67
x^2	-30.00	-0.47	0
x^3	-16.58	-0.13	-0.26
x^4	6.92	-1.51E-02	0
x^5	1.80	-4.52E-03	0
x^6	-0.46	-3.42E-04	0
x^7	-0.05	-1.53E-04	0
x^8	6.08E-03	-3.03E-06	0
x^9	-1.63E-03	-5.42E-06	0
x^{10}	1.13E-04	2.08E-07	0
x^{11}	6.40E-05	-2.00E-07	0

Table 6.2: Linear regression coefficients for the unregularized and regularized case.

looking at the table of the coefficients, we notice that the first few numerical values of the OLS estimates are quite large and have opposite signs, trying to cancel each other out. The regularized coefficients on the other hand, quickly drop in magnitude. We also note that the ridge regression forces the values to be small while the lasso approach leads to coefficients that are exactly zero. This means that the lasso method leads to sparse models and reduces the model complexity by making the model itself smaller.

We can understand this intuitively by looking at the two-dimensional case, namely, when we have two coefficients with their penalty. In the case of the ridge regularization, the penalty is of the form $a_1^2 + a_2^2$, whose level sets are circles Fig. 6.8. Given a function $h : \mathbb{R}^n \to R$, level-sets are sets of $(x_1, ..., x_n)$ where $f(x_1, ..., x_n) = k$ for any k (e.g. in the case of the map of terrain and the height function, those are the altitude curves). The lasso regularization has the functional form $|a_1| + |a_2|$ whose level sets are squares in the two-dimensional plane as shown in the left part of Fig. 6.8. By looking at these visualizations we realize that the "cornerness" of the lasso regularizaion offers more opportunities to set some coefficients to zero as the corners of the square are "special" points of the square whereas the circle for the ridge regression has no such feature.

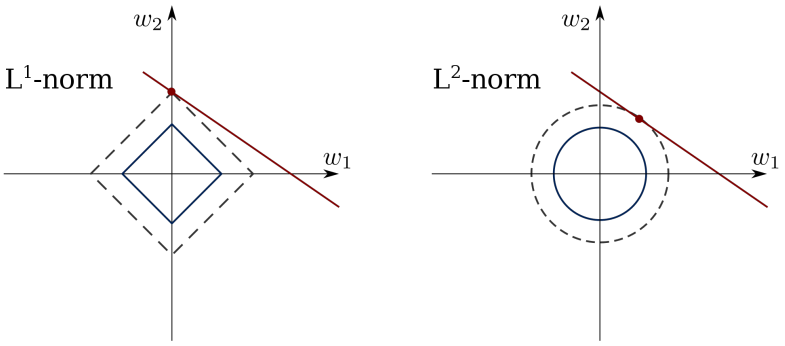


Figure 6.8: Comparison of impact at lasso and ridge penalty. Nicoguarro, BY-CC 4.0

Self-Check Questions

1. Which regularization scheme promotes sparse models?
2. True or False: Lasso and ridge regularization can only be used for linear regression.
3. What is the form of the penalty for ridge regularization?
4. What is the form of the penalty for lasso regularization?

Solutions

1. Lasso
2. False
3. Ridge: $\lambda \sum_{k=1}^{K} a_k^2$
4. Lasso: $\lambda \sum_{k=1}^{K} |a_k|$

6.5 Propagation of Uncertainties

Statistical and Systematic Uncertainties

Whenever we talk about "real" systems or data, we need to consider that both the data we may record, theoretic calculations and predictions about such systems, as well as the instruments we use for observing such systems are not ideal mathematical abstractions. Here we interpret "real" in the sense that they refer to physical systems. This implies that, for example, a measurement we obtain is associated with an intrinsic uncertainty. This means that we do not know the values with infinite precision but only up to a point–within its associated uncertainty. Each value that is related to a physical quantity—either because it is a measured value or a theoretical prediction for a "real" system— is associated with such an uncertainty that quantifies how well we know that number. This is because on one hand, we do not have perfect measurement devices but each sensor has, for example, a specific resolution or range where it can be used. For example, consider a thermometer: A household digital thermometer will maybe have a resolution of one or two degrees and can be used in the range of, say, -20 degree Celsius to + 120 degree Celsius. A fever thermometer will have a higher resolution of one tenth of a degree, but only in the range of 30-45 degree Celsius. A thermometer for scientific experiments will have a much higher resolution and operate in different regimes. However, whatever thermometer we may use, it will not be possible to have infinite resolution. On the other hand, almost all physical processes are stochastic in nature. This means that if we had a measurement device or sensor with very high precision, subsequent measurements would not result in the same value as the measured variable is a random variable and subsequent measurements would sample from the underlying probability density distribution.

We can also understand uncertainties in a different way: Suppose we know that a specific random variable X is distributed according to some probability density function. For example, the sales of goods in supermarkets can typically be described using a Gamma-Poisson or negative binomial distribution. The simplest assumption we could make is that these sales are a described by a random variable and are identically and independently distributed (i.i.d.). Unfortunately, this is not the case in reality (the samples are not i.i.d.) but if we made this assumption, we could measure the value of the parameters of the negative-binomial distribution from previously

observed sales data and then use this to infer the future sales. Necessarily, the sample from which we determine the parameters is finite, therefore, we can determine the values only up to a certain point or with a given precision. If we were to use a different sample, we would get slightly different values for the parameters. Since we use the (unrealistic) assumption that the sample is i.i.d., these values will be close to each other—but not the same. If we use relatively little data to measure the parameters, these fluctuations between datasets will be quite large, if we use a lot of data, the variations will be much smaller, since we can use more data for an accurate determination. These variations are the statistical uncertainties that arise because we rely on a given data-sample and because, fundamentally, our world is non-deterministic. Each of these samples will lead to a (slightly) different value for the parameters we extract from them.

In addition, our experimental setup or the sensor may introduce a bias or we need to consider external effects that influence our measurement. We group the uncertainties into the following categories: **systematic** and statistical uncertainties. Systematic uncertainties may arise from a variety of factors including the measurement setup, the tools used to obtain the measurements, and other natural factors. For example, if the scale used is not calibrated correctly, then the weight of objects will be consistently lower (or higher). Statistical uncertainties are due to the inherent randomness of the underlying process that we are trying to measure.

Systematic uncertainties are due to experimental setup, limited resolution, etc. This type of uncertainty cannot be reduced by adding more data.

We can express the value of the parameter of interest (say, the parameter of a Poisson distribution λ that describes mean and variance of the distribution) as:

$$\lambda = 3.14^{+0.14}_{-0.20}\ (\text{stat.})\ ^{+0.30}_{-0.15}\ (\text{syst.}) \tag{6.56}$$

The numerical values in this examples are "made-up" and just serve to illustrate how the numbers are reported. First, we have the best estimator that we may have obtained, for example, using a regularized least-squares approach, then we indicate the statistical uncertainty and then the systematic uncertainty. Note that these uncertainties may be asymmetric. Often, uncertainties that arise from theoretic predictions or calculations are reported separately because they may have a different behavior than uncertainties from experimental sources. Furthermore, we typically need to consider many different sources of systematic uncertainties in our experiments and we will discuss how to aggregate these contributions in the following.

Uncertainties in Simple Variable Transformations

Suppose that a random vector X represents the quantities we want to measure. Let Y be the vector of quantities we are ultimately interested in, but we cannot directly measure. Instead, we define a transformation that takes us from X to Y. The uncertainties in X affect the uncertainties of Y. In this section, we want to understand how these uncertainties are propagated from X to Y. In the simplest case, the transformation from X to Y is linear; there is a matrix B such that

$$Y = BX \tag{6.57}$$

In this setting, the uncertainties of Y, given by the covariance matrix $V[Y]$ can be obtained from the uncertainties of X, given by $V[X]$ as follows:

$$V[Y] = B\,V[X]\,B^T. \tag{6.58}$$

Note that $\sigma_1^2 = V[X_1]$ and analogously for σ_2^2, and $\sigma_{12} = V[X_1, X_2] = V[X_2, X_1] = \sigma_{21}$.

To illustrate this formula, let us consider the following example.

Uncertainties of a Transformed Variable

Let $X = (X_1, X_2)$ and $Y = (X_1 + X_2)/2$.
Use the formula to find $V[Y]$.

Solution
Note that we can write $Y = BX$ where $B = [1/2\ 1/2]$. The covariance of X is

$$V[X] = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} \tag{6.59}$$

The formula gives

$$V[Y] = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix} = \frac{1}{4}\sigma_1^2 + \frac{1}{4}\sigma_2^2 + \frac{1}{2}\sigma_{12} \tag{6.60}$$

Now suppose that $X = (X_1, X_2, \ldots, X_N)$ where the individual X_i are i.i.d. with $V[X_i] = 1$. Let Y be the sample mean: $Y = (X_1 + X_2 + \ldots + X_N)/N$. The covariance of X is given by the identity matrix:

$$V[X] = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} \tag{6.61}$$

As before, $Y = BX$ where

$$B = \begin{bmatrix} 1/N & \cdots & 1/N \end{bmatrix} \tag{6.62}$$

Then, the formula gives the variance of Y:

$$V[Y] = \begin{bmatrix} \frac{1}{N} & \frac{1}{N} \end{bmatrix} \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{N} \\ \vdots \\ \frac{1}{N} \end{bmatrix} = \frac{1}{N} \tag{6.63}$$

In both of these illustrations, Y is a scalar. Let us take a look at a simple example where Y is a vector. Let $X = (X_1, X_2)$ with X_1 independent of X_2 and all uncertainties are the same, i.e., $V[X_1] = V[X_2] = \sigma^2$. As an example, we can define Y as

$$Y = \begin{bmatrix} X_1 + 2X_2 \\ X_1 - 3X_2 \end{bmatrix} \tag{6.64}$$

Or, alternatively, in matrix notation, where we have $Y = BX$:

$$B = \begin{bmatrix} 1 & 2 \\ 1 & -3 \end{bmatrix} \tag{6.65}$$

The covariance of X is

$$V[X] = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix} \tag{6.66}$$

because the X_1 and X_2 are independent and have the same standard deviation. The covariance of Y is then given by:

$$\begin{aligned} V[Y] &= BV[X]B^T \\ &= \begin{bmatrix} 1 & 2 \\ 1 & -3 \end{bmatrix} \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 2 & -3 \end{bmatrix} \\ &= \begin{bmatrix} 5\sigma^2 & -5\sigma^2 \\ -5\sigma^2 & 10\sigma^2 \end{bmatrix} \end{aligned}$$

Propagation of Uncertainties for Variable Transformations

We now consider more general variable transformations for a vector of variables $\boldsymbol{x}$ to a new vector of variables $\boldsymbol{y}$. In the original base, the vector $\boldsymbol{x}$ has the means $\mu_{\boldsymbol{x}}$ and the covariance matrix $V[\boldsymbol{x}]$. Using a matrix B, the variables are transformed as $\boldsymbol{y} = B\boldsymbol{x}$. Note that in general the vector $\boldsymbol{x}$ has the dimension n and the vector $\boldsymbol{y}$ has the dimension m, where m can be different from n. This implies that the matrix B has the dimensions $m \times n$. The matrix B is given by

$$B_{ik} = \frac{\partial y_i}{\partial x_k} \tag{6.67}$$

and is composed of all combinations of partial derivatives, defined by the transformation. If we go back to our earlier examples, we find that we have implicitly used this already above. The mean of the new vector $\boldsymbol{y}$ is then given by:

$$\begin{aligned} \mu_{\boldsymbol{y}} &= E[\boldsymbol{y}] \\ &= E[B\boldsymbol{x}] \\ &= BE[\boldsymbol{x}] \\ &= B\mu_{\boldsymbol{x}} \end{aligned} \tag{6.68}$$

The variance is given by:

$$\begin{aligned} V[\boldsymbol{y}] &= E\left[(\boldsymbol{y} - \mu_{\boldsymbol{y}})\,(\boldsymbol{y} - \mu_{\boldsymbol{y}})^T\right] \\ &= E\left[(B\boldsymbol{x} - B\mu_{\boldsymbol{x}})\,(B\boldsymbol{x} - B\mu_{\boldsymbol{x}})^T\right] \\ &= E\left[B\,(\boldsymbol{x} - \mu_{\boldsymbol{x}})\,(\boldsymbol{x} - \mu_{\boldsymbol{x}})^T\,B^T\right] \\ &= BE\left[(\boldsymbol{x} - \mu_{\boldsymbol{x}})\,(\boldsymbol{x} - \mu_{\boldsymbol{x}})^T\right]B^T \\ &= BV[x]B^T \end{aligned} \tag{6.69}$$

We now revisit out earlier example where the new variable y is the sum of two uncorrelated variables x_1 and x_2:

Error Propagation for the Sum of Two Variables

What is the propagated uncertainty under the transformation $y = x_1 + x_2$?

To calculate the variance we need to determine $V[y]$:

$$\begin{aligned} V[y] &= BV[x]B^T \\ &= \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \end{bmatrix} \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \end{bmatrix}^T \\ &= \left(\frac{\partial y}{\partial x_1} \right)^2 \sigma_1^2 + \left(\frac{\partial y}{\partial x_2} \right)^2 \sigma_2^2 \\ &= \sigma_1^2 + \sigma_2^2 \end{aligned}$$

As a rule of thumb we remember that if we add variables, we add the variances or, in terms of the standard deviation, the standard deviation of the sum is the square root of the sum of the squared standard deviations. Colloquially: "Add the errors squared".

Another common transformation is the product of two variables, namely $y = x_1 x_2$.

Error Propagation for the Product of Two Variables

What is the propagated uncertainty under the transformation $y = x_1 \cdot x_2$?

To calculate the variance we need to determine $V[y]$:

$$\begin{aligned} V[y] &= BV[x]B^T \\ &= \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \end{bmatrix} \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \end{bmatrix}^T \\ &= \begin{bmatrix} x_2 \\ x_1 \end{bmatrix} \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \begin{bmatrix} x_2 \\ x_1 \end{bmatrix}^T \\ &= x_2^2 \sigma_1^2 + x_1^2 \sigma_2^2 \end{aligned}$$

This is easier to remember if we write it in terms of the relative error:

$$\left(\frac{\sigma_y}{y}\right)^2 = \left(\frac{\sigma_1}{x_1}\right)^2 + \left(\frac{\sigma_2}{x_2}\right)^2 \tag{6.70}$$

In the case that the variables are correlated, the rules for transformation of two variables as sum, difference, product and division are:

Transformation	Propagation of uncertainties
$y = x_1 \pm x_2$	$\sigma_y^2 = \sigma_1^2 + \sigma_2^2 + 2\rho_{x_1 x_2}$
$y = x_1 \cdot x_2$	$\left(\frac{\sigma_y}{y}\right)^2 = \left(\frac{\sigma_1}{x_1}\right)^2 + \left(\frac{\sigma_2}{x_2}\right)^2 + 2\frac{\sigma_1\sigma_2}{x_1 x_2}\rho_{x_1 x_2}$
$y = x_1/x_2$	$\left(\frac{\sigma_y}{y}\right)^2 = \left(\frac{\sigma_1}{x_1}\right)^2 + \left(\frac{\sigma_2}{x_2}\right)^2 - 2\frac{\sigma_1\sigma_2}{x_1 x_2}\rho_{x_1 x_2}$

Self-check questions

1. The transformation matrix B that transforms the variables from x to y via $y = Bx$ is defined as ....
2. What is the propagated uncertainty for $y = x_1 + x_2$ if x_1 and x_2 are uncorrelated?
3. True or False: Systematic uncertainties can be reduced by adding more data.
4. True or False: Statistical uncertainties can be reduced by adding more data.

Solutions

1. $B_{ik} = \frac{\partial y_i}{\partial x_k}$
2. $\sigma_y^2 = \sigma_1^2 + \sigma_2^2$
3. False
4. True

Summary

In this unit, we explored maximum likelihood and ordinary least squares (OLS) as approaches to parameter estimation. Maximum likelihood estimation requires that we specify the distribution from which the data are drawn. OLS requires that we specify a model that describes the relationship between two variables.

In the cases where we cannot determine the likelihood function, the expectation maximization algorithm provides a method to estimate the likelihood iteratively. This is commonly used in Gaussian mixture models that we can use, for example, to empirically describe a data set.

We discussed two types of regularization, ridge and lasso. The ridge regularization adds a penalty term which is a multiple of the squared L_2-norm, i.e., the sum of the squares of the parameters. The lasso regularization adds a penalty term which is a multiple of the L_1-norm, i.e., the sum of the absolute values of the parameters. Each method works by promoting smaller values of the model parameters. Lasso can be interpreted as a model selection approach; regularization is commonly applied to regression problems.

Finally, we discussed the propagation of uncertainties. The parameter estimators are functions of random variables, and the uncertainties of the random variables (sample) propagate to the parameter estimators. To evaluate certain properties of the estimator, it is important to note how these uncertainties are propagated. The general rule for the propagation of uncertainties is $V[y] = BV[x]B^T$.

7 Hypothesis Testing

Study Goals

On completion of this unit, you will have learned:

- the basics of a hypothesis test including the null and alternative hypotheses and test statistics.
- how to define type I and type II errors and how to compute a cut-off value using type I error rate.
- how to define and compute type II errors and how to compute the power of a test.
- how to determine a rejection region that maximizes the power of a test.
- how to compute and interpret p-values, and some common misconceptions about them.
- the context of multiple hypothesis testing and how to define a family-wise error rate and how the Bonferroni method offers a solution.
- how to define and control a False Discovery Rate.

Introduction

In many scenarios, we are interested in making a statement about an entire population, for example "is the medicine effective to treat a particular medical condition". Unfortunately, in almost all practical scenarios, we

cannot examine the entire population. In the example of medicine, it is not workable to give the medicine to every person on earth or every person who is with a given medical condition. Instead, we need to define a suitable sample that we can use to make a statement about a population. The exact details how to obtain a sample can be exceedingly difficult to avoid biases and make sure that the sample is indeed representative of the population. For the sake of the further discussion, we assume that we can obtain such a sample.

What we then want to do is to test a hypothesis, or rather, we want to decide, based on a sample, if a hypothesis is true (Casella & Berger, 2002, p.373). The **Null hypothesis** H_0 typically denotes the absence of an effect and the alternative hypothesis H_1 is the presence of the effect we want to establish. To do so, we employ knowledge about the situation, which allows us to propose a mathematical model (e.g. random variables with a particular family of distributions along some parameters) assuming some conditions (e.g. independence of the sample). This model is the core of the test which also provides a decision rule to decide by a calculation when a hypothesis is probable.

The Null hypothesis is a statement that represents no effect, i.e., the status quo.

For example, suppose that we model by the random variable θ the average change in cholesterol after taking a drug, we would be interested in discriminating between the null hypothesis $H_0 : \theta = 0$ (i.e. the drug has no effect) vs. $H_1 : \theta \neq 0$ (the drug causes a change in cholesterol). In this particular case, we would be more interested in the case $\theta < 0$, i.e., the drug causes a decrease in cholesterol rather than an increase, so we might want to formulate the discrimination between the null and test hypothesis as "the level of cholesterol remains the same or increases" vs. "the level of cholesterol decreases". The important part is that we define two disjoint regions of some parameter θ that is representative of the population we want to investigate.

The hypothesis test is defined as a rule that we can use to (Casella & Berger, 2002, p. 374):

- fail to reject the null hypothesis H_0 or
- reject the null hypothesis H_0 and accept the alternative hypothesis H_1 as true.

In the case of the cholesterol drug, there are several ways in which we can

test its effectiveness. For example, we could compare how the levels of cholesterol change if we switch a cohort of patient from one drug to the next. In this example, we would have the same number of patients and we would take at least two measurements from each patient: One when the old drug was taken and then sometime later when the new drug was taken. This is called a "paired test" since we have at least two measurements for the same person, i.e., the earlier measurement can be "paired" to the later one.

We could also design a study using two independent groups that are evaluated at the same time: One group receives a placebo and the other group the drug. To avoid any biases, we would assign the study participants randomly to each group and ideally also do not tell the persons administering the medication whether they hand out the old or new drug. This is called a randomized controlled trial. Unlike the above case where the data was "paired" (two measurements for the same person), we consider two independent statistical samples, one for each group of patients. The number of individuals in each group may or may not be the same.

When we perform the measurements, we obtain a large dataset of measured values $x_1, x_2, x_3, \ldots$ we take from each individual in the sample. Hence, for each group we obtain a distribution of values, for example, one distribution of measured values for the group of patients that have taken the placebo and another distribution of measurements for the group of individuals that have taken the drug. To perform the hypothesis test to decide whether to accept or reject the null hypothesis, we need to construct a test statistic $t = f(x_1, x_2, x_3, \ldots) = f(\mathbf{x})$ that we can use to derive this decision formally.

The simplest way we can discriminate between the null and the alternative hypothesis is to compare the mean of the two groups, approximated by the sample mean from the measured values. Recall that due to the central limit theorem, any sum of identically distributed random numbers will approximate a normal distribution, regardless of the distribution the individual random numbers follow. Hence, if the number of individuals in each sample is sufficiently large (say, $n > 30$), the sample means $\langle \mathbf{x_1} \rangle$ and $\langle \mathbf{x_2} \rangle$ of the two samples will follow an approximate normal distribution.

To understand the discussion how to compare the two groups better, we first limit ourselves to the case where we have only one group and want to determine whether this group is compatible with a given hypothesis. For

example, we can investigate whether the height of all men in a certain age range is compatible with an (assumed) true average height μ.

The z-score for a statistic t is defined as $z = \frac{t-\mu}{\sigma}$ where μ is the mean and σ is the standard deviation of the population.

We therefore define the z-**score** of the mean $\langle \mathbf{x} \rangle$ as

$$z = \frac{\langle \mathbf{x} \rangle - \mu}{\sigma/\sqrt{n}} \tag{7.1}$$

where $\langle \mathbf{x} \rangle$ is the sample mean, σ is the variance of the population and n the number of men we look at. In this example we assume that we know both μ and σ, i.e. the only unknown is the sample mean $\langle \mathbf{x} \rangle$. This distribution follows (for a sufficiently large number of n) an approximate normal distribution since the sample means follow a normal distribution for large enough samples and the number of elements in each sample and the variances σ^2 are fixed. If the measurements are taken from a normal distributed sample, z will also follow a normal distribution, without relying on the properties of the central limit theorem.

However, in most cases we will not now know the population variance σ and need to replace it with the sample standard deviation s. We can then write:

$$t = \frac{\langle \mathbf{x} \rangle - \mu}{s/\sqrt{n}} \tag{7.2}$$

Recall that the sample variance defined as:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \langle \mathbf{x} \rangle^2) \tag{7.3}$$

is an unbiased estimator due to the correction factor $\frac{1}{n-1}$. Note that now we have two random numbers in the definition of t, the sample mean and the sample variance s.

The distribution of the t values does not follow a normal distribution, essentially because we use the sample variance instead of the (unknown) population variance σ. Instead, if the variable X follows a normal distribution (which we might know by previous knowledge) t follows the Student's t-distribution which has a single parameter, the number of degrees of freedom. In the above case, the number of degrees of freedom is $n - 1$. The larger the number of elements in the sample is, the closer the distribution approximates a normal distribution.

The Student's t Distribution

This distribution behaves very much like the standard normal distribution, but the tails are more pronounced compared to the standard normal distribution. The mean (center) of the t distribution is always zero, but the standard deviation changes with the degrees of freedom parameter, n. If a variable X follows a Student t distribution with n degrees of freedom, then we write $X \sim t(n)$. The PDF of X is given by

$$f_n(t) = \frac{1}{\sqrt{\pi n}} \frac{\Gamma((n+1)/2)}{\Gamma(n/2)} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2} \tag{7.4}$$

We now return to the case of the two groups: We can construct a test statistic similar to the the z-score above as:

$$t = \frac{(\langle \mathbf{x_1} \rangle - \langle \mathbf{x_2} \rangle) - \delta}{SE\left[\langle \mathbf{x_1} \rangle - \langle \mathbf{x_2} \rangle\right]} \tag{7.5}$$

where $(\langle \mathbf{x_1} \rangle - \langle \mathbf{x_2} \rangle)$ is the observed difference between the sample means of group one and group two, the constant δ is the known difference between the population means under the null hypothesis and $SE[\langle \mathbf{x_1} \rangle - \langle \mathbf{x_2} \rangle]$ is an estimate of the standard error of the difference between the two samples. In general, we do not know the difference δ unless we have some further or external knowledge about the populations, and we therefore set $\delta = 0$ in most practical applications. The test statistics then becomes:

$$t = \frac{\langle \mathbf{x_1} \rangle - \langle \mathbf{x_2} \rangle}{SE\left[\langle \mathbf{x_1} \rangle - \langle \mathbf{x_2} \rangle\right]} \tag{7.6}$$

From the propagation of uncertainties we remember that the standard error of the difference is given by

$$SE\left[\langle \mathbf{x_1} \rangle - \langle \mathbf{x_2} \rangle\right] = \sqrt{V[\mathbf{x_1}] + V[\mathbf{x_2}]} \tag{7.7}$$

If we know the variances of the population, the above equation becomes

$$t = \frac{\langle \mathbf{x_1} \rangle - \langle \mathbf{x_2} \rangle}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{7.8}$$

As in the single sample case discussed earlier, we generally do not know the population variances and need to estimate them from the sample variances.

In this case we obtain:

$$t = \frac{\langle \mathbf{x_1} \rangle - \langle \mathbf{x_2} \rangle}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{7.9}$$

Again, the values of t follow a Student's t-distribution as proven by Hogg et al. (2020, theorem 3.6.1). However, in this case where neither the sample variances nor the number of elements in the sample are the same, the number of degrees of freedom (d.o.f.) is difficult to determine and calculated using the Welch-Satterthwaite formula (Satterthwaite, 1946; Welch, 1946). The value of the number of degrees of freedom is between the limits $\min(n_1 - 1, n_2 - 1) \leqslant \text{d.o.f} \leqslant n_1 + n_2 - 2$. The number of degrees of freedom is close to the upper limit if the variances are similar and close to the lower limit if the variances in the two samples are very different. A reasonable approximation is the harmonic mean:

$$\text{d.o.f.} = \frac{2}{\frac{1}{n_1} + \frac{1}{n_2}} \tag{7.10}$$

If we assume that the variances in the two groups are the same, then this is called the "two-sample Student t-test", if we allow the variances to be different from each other, it is called the "Welch's test".

The case of the test for paired samples can be expressed as a special case of the one-sample t-test: In this case, we have two measurements for each individual in the sample and we can look at the difference in the measurements: $d_i = x_i - y_i$ between the two samples. We then define the test statistics as

$$t = \frac{\langle d \rangle - \mu}{s_d / \sqrt{n}} \tag{7.11}$$

where $\langle d \rangle$ and s_d are the sample average and sample standard deviation of the difference between all pairs and μ is the hypothesis we want to test, e.g, $\mu = 0$ if the null hypothesis is that there is no effect. The number of degrees of freedom is $n - 1$, where n is the number of paired measurements.

Given that the null hypothesis is true, we can use the distribution of this test statistic to decide whether the observed value of the test statistic is unlikely or not. The observed value of the test statistic comes from

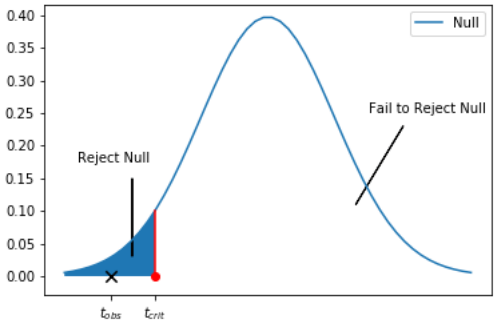


Figure 7.1: Distribution of the Test Statistic Under Null] Hypothesis and the Cutoff Value Sample

replacing all the estimators by their observed values (estimates):

$$t_{\text{observed}} = \frac{\langle x_1 \rangle - \langle x_2 \rangle}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_1^2}{n_2}}} \tag{7.12}$$

To decide whether this observed test statistic is statistically significant, we need to see where it falls in the distribution. The closer this value is to the tails of the distribution, the less likely the data is under the null hypothesis. The farther from the tails, the more it is in the observed data. To decide, we need to set a cutoff value, t_c.

This is illustrated in Fig. 7.1 that shows the distribution of our test statistic together with the cutoff value (in red). If the observed value is in the "rejection region", it means that the data is more unlikely than we want to tolerate. Otherwise, it is in the "fail to reject region". What remains to figure out is the choice of this cutoff value.

Note that "significant" is not the same as "large effect": We can observe a significant difference between two samples, meaning that we can distinguish with high confidence that the two samples are different, even if the size of the difference is very small – maybe even too small to be of any practical consequence.

So far, in our discussion, we have generally assumed that the sample means follow a normal distribution, meaning that if we take different samples from a population, the distribution of these sample means follows a normal

distribution. When the number of measurements is large enough (say, $n > 30$), this is fulfilled due to the Central Limit Theorem

However, whenever we do an analysis, we need to check if the distribution actually follows a normal distribution. Generally, the t-test approach is fairly robust if the distributions are non-normal but symmetric but it tends to work less well if the distributions are asymmetric. Therefore, in practice, if we have to deal with a situation where the assumption that the underlying distribution is a normal distribution is not fulfilled, we have to take this into consideration as a source of systematic uncertainty and quantify how big the effect of this "non-normality" is on our final result.

7.1 Type I and Type II Errors

In our discussion above, we focused on the special case where we constructed a test statistic such that we could compare the sample means of two samples to make a statement about the null and alternative hypothesis of the population. This is not the only way in which we can construct a test statistic – as we noted above, to decide between the hypothesis we need to construct some function of the measurements that – intuitively speaking – maps the measured values into a number: $t = f(x_1, x_2, x_3, ...) = f(\boldsymbol{x})$. This function can in principle be anything, even a sophisticated machine learning algorithm. Using this function, we can show the distribution of this test statistic for the null and the alternative hypothesis as shown below. To decide whether to reject the null hypothesis, we need to define a critical value t_c as a cut-off: For any value of the test statistic below this value, we accept the null hypothesis, for any value above we reject the null hypothesis and accept the alternative hypothesis. Unfortunately, however, in most realistic cases we cannot distinguish between the null and alternative hypothesis perfectly, i.e., the distributions of the test statistics for the two cases will overlap. This means we can make two types of errors as illustrated in Fig. 7.2:

- Error of the first kind (or type I error): The null hypothesis is true, but the value of the test statistics we compute from the observed sample is beyond the cut-off. Hence, we reject the null hypothesis and accept the alternative hypothesis, even though it is wrong. We call α the probability for this.

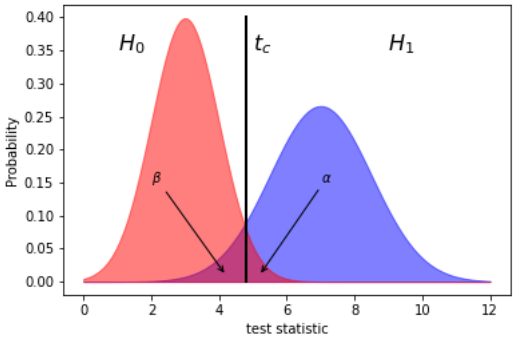


Figure 7.2: Two Types of Errors and Associated Probabilities

- Error of the second kind (or type II error): The alternative hypothesis is true, but the value of the test statistics is lower than the cut-off. Hence, we did not reject the null-hypothesis even though it is wrong. We call β the probability for this to happen.

We can summarize this in the following table:

	H_0 is True	H_0 is False
Accept H_0	correct inference	Type II Error (β)
Reject H_0	Type I Error (α)	correct inference

In other words: $\alpha = P(\text{reject } H_0 | H_0 \text{ is true})$ and $\beta = P(\text{fail to reject } H_0 | H_0 \text{ is false})$.

The cut-off value t_c affects these two probabilities, and it is important to consider the trade-off between α and β when deciding which value of t_c to choose. In many practical applications, α is chosen in advance, and this directly determines the cut-off value t_c as well as β. The reason we like to choose α in advance is to find, without looking at the data, what kind of type I error we are willing to tolerate. Once we have seen the data, then our choice for β will be tainted because if the data does not fit the alternative hypothesis, we might decide to use a larger α so the test is less stringent. If the data seem to fit the alternative hypothesis well, we might choose a small α to give a stringent impression.

Power of the test

Since α and β have an inverse relationship (i.e., as one decreases, the other increases), we need a strategy for defining α. We will discuss this below. Note that the cut-off value automatically determines the rejection region (RR): the set of values of the statistic t that are at least as extreme as the cut-off value. The chosen value of α is a naïve metric for evaluating the performance of the hypothesis test. However, another more useful metric is the **power of the test**. Informally, the power of a test is the probability that the hypothesis test will yield a decision to reject H_0: Power $= P(\text{reject } H_0 | H_1 \text{ is true})$ and describes the probability to detect an effect if it is indeed there.

The power of a test is a measure of the quality of a test with respect to the probability that the testing process will detect an effect.

Suppose that the hypothesis test is about the parameter θ. We have $H_0 : \theta = \theta_0$ and $H_1 : \theta \in W \subseteq \mathbb{R} \backslash \{\theta_0\}$. Given $\theta_1 \in W$, the test statistic T, and the rejection region (RR), the power of the test when the actual parameter is $\theta = \theta_1$ is

$$\text{Power}(\theta_1) = P(T \in \text{RR} | \theta = \theta_1) = P(\text{rejecting} H_0 | \theta = \theta_1). \tag{7.13}$$

Returning to the probability of the type II error, we say that $\beta(\theta_1)$ is the probability of failing to reject H_0 when the true value of the parameter is $\theta = \theta_1$. Then

$$\text{Power}(\theta_1) = 1 - \beta(\theta_1). \tag{7.14}$$

If the population standard deviation is $\sigma = 3$, we want to test the claim that the population mean is at least 15 with a confidence level of $\alpha = 0.05$. The hypotheses to test are $H_0 : \mu = 15$ versus $H_1 : \mu > 15$. The test statistic, the one-sample Student's t-test or the paired t-test, is given by:

$$t(x | \mu = 15) = \frac{\langle x \rangle - 15}{\sigma / \sqrt{N}} \tag{7.15}$$

where the sample mean $\langle x \rangle$ is defined as $\langle X \rangle = \frac{1}{N} \sum_{i=1}^{N} x_i$ and σ is the population standard deviation. We know from the central limit theorem that if N is large enough, the distribution of t follows a normal distribution if we know the population standard deviation σ. If we need to estimate σ using the sample standard deviation, and the x_i can be considered distributed according to a normal distribution, the distribution of t will follow

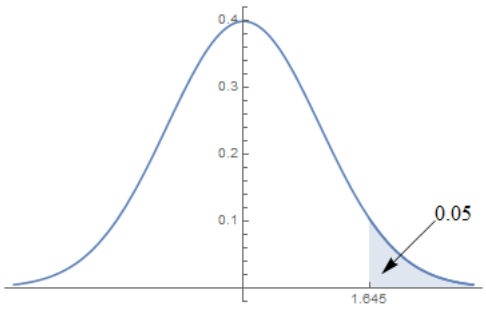


Figure 7.3: Standard Normal Distribution with an Upper Tail Cut-Off of 0.05.

a Student's t distribution with $N-1$ degrees of freedom. Given our choice of $\alpha = 0.05$, the cutoff value comes from the relationship $P(Z > z_c) = 0.05$.

In the case where we know the population variance, we can use, instead of the test statistic t, the z-score which follows a standard normal distribution. If we choose to perform the test at the $\alpha = 0.05$ level, then the cut-off value is $z_c = 1.645$ (since $P(Z > 1.645) = 0.05$ as illustrated in Fig. 7.3). The corresponding rejection region is defined by $z > 1.645$, which corresponds to the region defined by:

$$\mathrm{RR}_N = \left\{ \langle x \rangle > 15 + 1.645 \cdot \frac{3}{\sqrt{N}} \right\} \tag{7.16}$$

Suppose that the true value of the population mean is $\mu = 16$. Let us compute the probability of the type II error, $\beta(16)$ and the power of the test power(16).

$$\begin{aligned} \beta(16) &= P\left(\frac{t(X|\mu = 16)}{3/\sqrt{N}} \leqslant \frac{15 + 1.645 \cdot \frac{3}{\sqrt{N}} - 16}{3/\sqrt{N}} \right) \\ &= P\left(Z \leqslant 1.645 - \frac{\sqrt{N}}{3} \right) \end{aligned} \tag{7.17}$$

The power of the test is

$$\begin{aligned} \mathrm{power}(16) &= 1 - P\left(Z \leqslant 1.645 - \frac{\sqrt{N}}{3} \right) \\ &= P\left(Z > 1.645 - \frac{\sqrt{N}}{3} \right) \end{aligned} \tag{7.18}$$

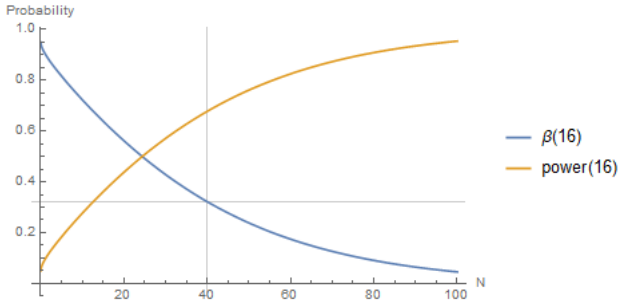


Figure 7.4: The Probability of Type II Error and Power for Various Values of N

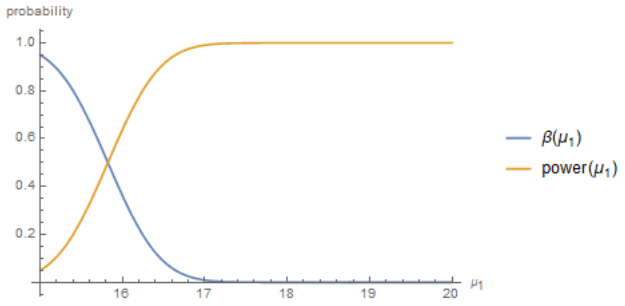


Figure 7.5: The Type II Error Curve and the Power Curve for a One-Sided Test

As shown in the Fig. 7.4, the probability of type II error decreases as the sample size increases and the power of the test increases with the sample size. If $N = 36$, the probability of the type II error is about 36 percent. What this means is that the probability of failing to reject the null hypothesis, when the null hypothesis is false because the actual parameter value is 1 more than the assumed value, is 32 percent. This is quite high. If detecting a 1-unit change is practically significant for us, then we should collect more data to decrease the type II error.

Let us now fix the sample size to $N = 36$ and compute the power of the test for various values of μ in the rejection region RR as illustrated in Fig. 7.5.

$$\begin{aligned} \beta(\mu_1) &= P\left(\frac{t(x|\mu = \mu_1)}{3/\sqrt{36}} \leqslant \frac{15 + 1.645 \cdot \frac{3}{\sqrt{36}} - \mu_1}{3/\sqrt{36}}\right) \\ &= P(Z \leqslant 31.645 - 2\mu_1) \qquad (7.19) \\ \text{power}(\mu_1) &= P(Z > 31.645 - 2\mu_1) \qquad (7.20) \end{aligned}$$

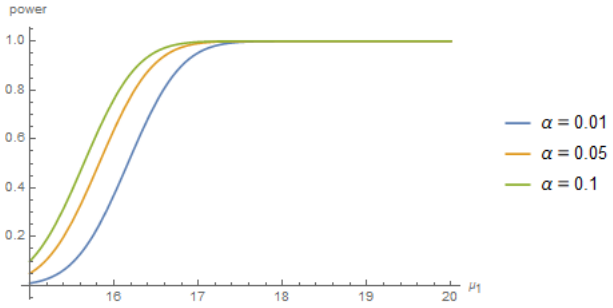


Figure 7.6: Power Curves for Various Values of α.

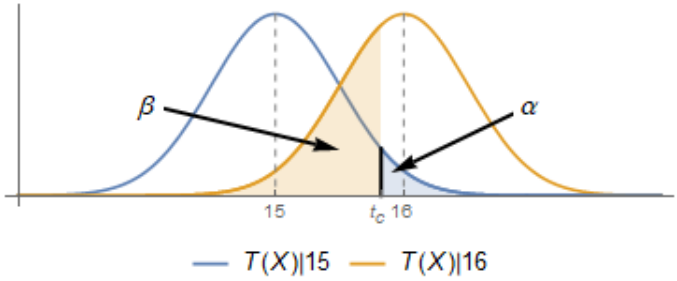


Figure 7.7: Distributions of the Test Statistic Together with α and β.

When performing a hypothesis test, it is important to consider that the probability of a type I error and the power of a test have an inverse relationship. We would like the power to be high while the probability of committing a type I error to be small. Exploring the power curves for various values of α in Fig. 7.6 explores this relationship.

Let us consider the same example above, and we fix $\mu_1 = 16, N = 36$, and $\alpha = 0.05$. The two distributions of $t(X|\mu = 15)$ and $t(X|\mu = 16)$ respectively are shown in Fig. 7.7, together with the fail-to-reject and rejection regions.

The Neyman-Pearson Lemma

As we have seen earlier, choosing the cut-off value t_c is vital in the construction of the hypothesis test. We have discussed above how we can use the error of the first and second kind, as well as the power of the test to determine t_c. However, we can also look at it from a different perspec-

tive and consider the graph in Fig. 7.2 showing the distribution of the test statistic for the null and alternative hypothesis. We might be tempted to choose the cut-off value t_c where the distributions intersect (as shown in Fig. 7.2). This means that left of the cut-off we consider the cases where the probability to accept the null hypothesis is larger than accepting the alternative hypothesis and correspondingly for values larger than t_c. However, this choice is not optimal if the a priori probabilities of the null and alternative hypothesis are very different. Using the Bayesian priors for the null hypothesis $P(H_0)$ and the alternative hypothesis $P(H_1)$, we can define the cut-off as the biggest t such that

$$P(H_0)f_0(t) \geqslant P(H_1)f_1(t) \tag{7.21}$$

Where $f_i(t)$ is the distribution of the test statistic for the null and alternative hypothesis. The best choice can be determined by the Neyman – Pearson lemma Hogg et al. (2020, Theorem 8.11):

Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ denote a random sample from a distribution with a single unknown parameter θ. Consider an α-level hypothesis test with $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$. Denote the likelihood of the sample by $\mathcal{L}(\theta)$. Then the most powerful test has a rejection region given by

$$\frac{\mathcal{L}(\theta_0)}{\mathcal{L}(\theta_1)} > k \tag{7.22}$$

where k is chosen so that the probability of the type I error is α.

As an example, suppose that $x_1, x_2, \ldots, x_n$ are n independent observations from a Poisson distribution with unknown mean rate μ. Recall that the Poisson distribution is a discrete probability distribution for positive integer numbers:

$$f(k;\mu) = \frac{e^{-\mu}\mu^k}{k!} \quad k = 0, 1, 2, \ldots \tag{7.23}$$

As an example we want to test the simple hypothesis $H_1 : \mu = 1$ versus $H_0 : \mu = 2$. Since we assume that the observations are independent, the likelihood factorizes and we can exploit the fact that $e^a \cdot e^b = e^{a+b}$:

$$\mathcal{L}(\mu) = \prod_{i=1}^{n} \frac{\mu^{k_i} e^{-\mu}}{k_i!} \tag{7.24}$$

$$= \frac{\mu^{\sum_{i=1}^{n} k_i} e^{-n\mu}}{\prod_{i=1}^{m} k_i!} \tag{7.25}$$

With the abbreviation $u = \sum_{i=1}^{n} k_i$, we can write the likelihood ration as:

$$\frac{\mathcal{L}(\mu_1)}{\mathcal{L}(\mu_2)} = \frac{\mu_1^u e^{-n\mu_1} / \prod_{i=1}^{m} k_i!}{\mu_2^u e^{-n\mu_2} / \prod_{i=1}^{m} k_i!} = \frac{\mu_1^u e^{-n\mu_1}}{\mu_2^u e^{-n\mu_2}} \tag{7.26}$$

In our case, $\mu_1 = 1$ and $\mu_2 = 2$, hence

$$\frac{\mathcal{L}(1)}{\mathcal{L}(2)} = \frac{1^u e^{-n\cdot 1}}{2^u e^{-n\cdot 2}} = \frac{1}{2^u e^{-n}} \tag{7.27}$$

According the Neyman-Pearson lemma, the rejection region for the most powerful alpha-level test is given by a suitable constant k such that

$$P\left(\frac{1}{2^u e^{-n}} < k \middle| \mu = 1\right) = \alpha \tag{7.28}$$

We can rewrite this in the following way:

$$\begin{aligned}
\frac{1}{2^u e^{-n}} < k \quad &\Leftrightarrow \quad 2^u e^{-n} > \frac{1}{k} \\
&\Leftrightarrow \quad 2^u > \frac{e^n}{k} \\
&\Leftrightarrow \quad \ln(2^u) > \ln\left(\frac{e^n}{k}\right) \\
&\Leftrightarrow \quad u\ln(2) > n - \ln(k) \\
&\Leftrightarrow \quad u > \frac{n - \ln(k)}{\ln(2)}
\end{aligned}$$

If we introduce a new constant $k' = \frac{n-\ln(k)}{\ln(2)}$ to replace our previous constant k, we can rewrite Eqn. (7.28) as:

$$P(u > k' | \mu = 1) = \alpha \tag{7.29}$$

Since a sum of Poisson numbers is again a Poisson with parameter $\lambda = \sum_i \lambda_i$, the sum u we have introduced earlier follows again a Poisson distribution. Now, let us assume we have $n = 10$ observations. If $\mu_1 = 1$ is true, the sum of all rate parameters is 10 and hence:

$$u = \sum_{i=1}^{10} y_i \ \sim \ \text{Poisson}(10) \tag{7.30}$$

m	α
14	0.1355356
15	0.08345847
16	0.0487404
17	0.02704161
18	0.01427761

Table 7.1: Closest confidence level for Poisson Example.

Unfortunately, the constant k' is not an integer, so we cannot evaluate Eqn. (7.29) directly. Instead we define an integer $m = \lceil k' \rceil + 1$, where $\lceil k' \rceil$ is the least integer greater or equal to k' (the "ceiling" of k'). With this, we evaluate:

$$P(u > m|\mu = 1) = \alpha \tag{7.31}$$

to determine the closest value of α numerically. The values are shown in Tab. 7.1. The closest we can get to $\alpha = 0.05$ is for the value $m = 16$, hence the rejection region is given by RR $= \{u \geqslant 16\}$.

Self-Check Questions

1. The power is connected to the type II error via ....
2. The probability of rejecting the null hypothesis H_0 if the null hypothesis is true is called?
3. If the null hypothesis is true, the probability to fail to reject the null hypothesis is called?
4. A type I error has a probability of... ?

Solutions

1. Power $= 1 - \beta$
2. α
3. β

4. α

7.2 p-Values

In the discussion so far, we focused on how we can define a hypothesis and a test statistic such that we can discriminate between two groups and determine if any observed difference between them is "real" in the sense that we can tell the two groups apart and we either accept the null hypothesis (e.g. "There is no difference") or "reject the null" and accept the alternative hypothesis (e.g. "There is a difference"). We have also discussed how we need to define a cut-off value that allows us to make the distinction and looked at errors of the first and second kind and which role the values of α and β play in this.

We now discuss in more detail how we can quantify the significance of the result of the hypothesis test. The key question can be phrased for example like this: As we have observed a difference between the groups and, say, rejected the null hypothesis and accepted the alternative hypothesis, how sure are we, that this is not due to random chance? For example, if we were to repeat the same experiment many times by using different samples, how sure are we that we would always observe this difference between the groups and we are not "lucky" that we have found one specific sample where the difference is observable between the groups but this is just a statistical fluke?

For example, if we are studying the effect of a medicine and we give one group of patients the medicine and one group obtains the placebo, we would use the test statistic discussed earlier to assess whether or not we can reject the null hypothesis and accept the alternative, i.e., the medicine works. We could then declare victory if we observed the result that the group that obtains the medicine fairs better than the group with the placebo – but how sure are we that, if we were to do a new study with different participants, we would observe the same result? And again if we did another study? Or after 100 studies? Or 1000?

We therefore need a metric that allows us to make a statement that an observed outcome, e.g., the difference of the sample mean of two groups, is larger than we would expect it from random variation and chance alone. If

we are performing multiple studies with different samples (such as participants in a study), the exact outcome will always depend on the composition of the sample and hence we always expect some variation in the difference of the sample mean. However, for a result to be "statistically significant", we require that the expected variation due to random fluctuation is smaller than what we see.

This is expressed by the p-value, which is defined, following the statement of the American Statistical Association, as:

> "Informally, a p-value is the probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed value." (Wasserstein & Lazar, 2016).

This means that the p-value is a probability that - if we assume a specific hypothesis-, the observed value of the test statistics would have this (or a more extreme) value. In other words, for a test statistic T, the p-value is the smallest value of α for which the observed data suggest that the null hypothesis is rejected. In this setting, the smaller the p-value, the more unlikely it is that the data comes from the distribution specified by the null hypothesis.

In many scientific fields, a threshold of $p < 0.05$ is chosen. This means that all p-values that have been determined in the analysis that are smaller than this threshold are considered to originate from a statistically significant effect. Here we need to make two important observations: First, the threshold of 0.05 is arbitrary. The value is fairly common in several scientific fields – but that does not mean that a value of 0.004 or 0.1 is better or worse per se. A probability of 0.05 that an effect is significant, i.e., not due to random fluctuations, also means that in 5% of the cases (i.e., one out of 20) the effect is not significant and due to random fluctuations. We can set higher or lower thresholds of what we consider good practice to claim an effect. For example, in particle research at laboratories such as CERN, a threshold of about $p < 0.0000003$ (the "five sigma level") is required to claim the observation of a new effect. In many other scientific fields, this would not be workable: For example, a medical study would need many tens of thousands of participants to reach such a high level of precision.

Consequently, the threshold is set at a different level, accepting the risk that the results of one in 20 studies are wrong, even though all steps in the scientific work were followed correctly.

Second, in our everyday language we relate the word "significant" to something important and we are tempted to relate a significant finding to a big effect. For example, if we say that a medicine has a significant effect to lower blood pressure, we would think that by taking such a medicine we would observe a sizeable drop in our blood pressure. However, significance is not the same as effect size – by performing a very precise measurement, we can be certain that the observed effect is not due to random chance, but it can still be a very small effect. In the example of the medicine, we may well observe that the medicine causes a significant reduction in blood pressure in patients who take this medicine – but the change in blood pressure, the effect size, is so small that it is of no practical value.

As the above definition indicates, the p-value and its meaning are not easily grasped, in fact, in the same statement the authors say:

> "While the p-value can be a useful statistical measure, it is commonly misused and misinterpreted. This has led to some scientific journals discouraging the use of p-values, and some scientists and statisticians recommending their abandonment, with some arguments essentially unchanged since p-values were first introduced." (Wasserstein & Lazar, 2016)

The p-value is calculated in the following way. We assume we have two hypotheses, the null hypothesis H_0 and the alternative hypothesis H_1. Let T denote the associated test statistic and T_{observed}, its observed value. The p-value is the probability of observing a value for T that is at least as extreme as the observed value (under the assumption that H_0 is true). In this hypothesis test, we have

$$p - \text{value} = P(T > T_{\text{observed}}|H_0) \tag{7.32}$$

For example, if T follows a student's t distribution with 30 degrees of freedom and $T_{observed} = 2.3$, then

$$p - \text{value} = P(T > 2.3|H_0) = 0.014. \tag{7.33}$$

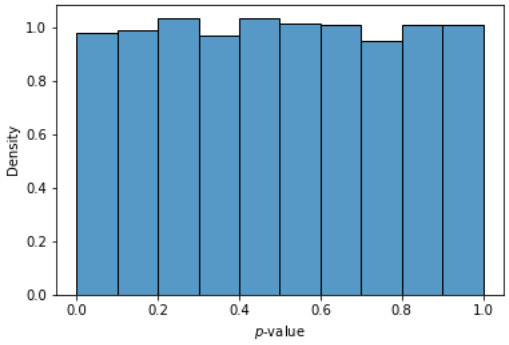


Figure 7.8: Distribution of the p-value for the case that the Null hypothesis is true.

Remember that the p-value is based on a random sample. This means that the p-value itself is a random variable distributed on the interval [0,1].

As a matter of fact, if the null hypothesis is true, the distribution is uniform, which means that we can obtain any value for the p-value with equal probability (Murdoch, Tsai, & Adcock, 2008). We can illustrate this with a small simulation where we generate two sets of random numbers according to a normal distribution with the same mean and variance. Hence the null hypothesis is that the difference of sample means is zero – which is by construction the case here. Then we calculate the p-value for this sample and fill it into a histogram. By repeating this process 10.000 times, we obtain the distribution of p-values under the null, i.e. for the case that the null hypothesis is correct. As expected, this is a uniform distribution where each p-value is equally likely as shown in Fig. 7.8.

Since there is much confusion about the p-value, we summarize some key concepts below:

- The p-value is a statistical measure to indicate how incompatible the data are with a given hypothesis. In many cases, we challenge the null hypothesis (e.g., there is no difference between two groups as measured by their sample means) and want to assess how likely it is that any observed difference is due to random fluctuations. In this sense, the p-value is a measure to cast doubt – the smaller the p-value, the less likely it is that the data are compatible with the assumption that the null hypothesis is true and we need to investigate further.

- The p-value is a measure of the evidence of an effect, not its size. For example, a p-value of $p = 0.001$ does not indicate a larger effect than, say, $p = 0.1$.

- The p-value is a random variable. P-values between experiments and different samples vary, one study may find a significant effect determined by $p < 0.05$, the other does not – but this does not mean that the effect is not there or that either study did something wrong.

- The threshold $p < 0.05$ is arbitrary and to some extent a convention. It implies that on average one in twenty results published with this threshold are wrong.

- If the null hypothesis is true, the distribution of p-values is uniform, i.e,. each value is equally likely.

What the p-value is not or what it cannot do:

- The p-value does not determine that the (alternative) hypothesis studied is true.

- The p-value does not determine that the data was produced by random chance alone.

- Statistical significance determined by the p-value (e.g., $p < 0.05$) does not imply effect size or importance: The finding may be statistically significant but irrelevant in practice because the effect size is too small.

- A non-significant p-value (e.g., $p < 0.05$) does not imply that there is no effect, just that the null hypothesis is also compatible with the data – you may need more data or larger samples to establish the effect. In other words: absence of evidence is not the same as evidence of absence.

Due to the confusing nature of the p-value, there are also many cognitive biases regarding published scientific work that need to be taken into account. In many studies, only "significant" results where the p-value is below a threshold (e.g., $p < 0.05$) are reported, but not all the hypotheses or tests that did not match these thresholds are included in the report as well. This can lead to a bias since the p-value is a random variable and

at some point, the researcher may be "lucky"to obtain a result past the threshold. When all the experiments that were not deemed significant are not reported, we do not know if the final published study is just "lucky". Even studies published in one of the highest ranking scientific journals can be quite problematic when it comes to the discussion of statistical significance. In the paper, Weber, Hoang Do, and Chung (2018) report: "[...] we continuously increased the number of animals until statistical significance was reached to support our conclusions." A detailed analysis on the distribution of p-values reported in scientific publications found that the values are all below the threshold $p < 0.005$, meaning that no author in the publication they analysed reported a non-significant finding (or that the editors did not approve the publication of such results). Furthermore, the analysis showed that the distribution of all p-values shows sharp peaks at round values like $p = 0.1, p = 0.2, \ldots$ with the highest peak at $p = 0.05$ (Chavalarias, Wallach, Li, & Ioannidis, 2016).

Conclusions are Significant and not Significant at the same time

The same data can also lead to different interpretations and the result can be either significant or not, depending on the assumptions we make. An illustrative example focuses on whether students cheat in exams (Wagenmakers, 2007): A student is asked 12 factual true/false statements and the sequence of answers may be C,C,C,E,E,C,C,C,C,C,C,E where C indicates a correct response and E and error or wrong answer. Hence, we have nine correct answers and three wrong ones.

As a null hypothesis, we assume that the student does not know the answer to any question and just guesses. Since we have a binary response (yes/no), we can model this as a sequence of 12 Bernoulli trials and use the binomial distribution, assuming that random guessing will result in a 50%/50% chance of obtaining the correct result. Using the binomial distribution, the probability of obtaining 9 correct answers out of 12 is 0.054 and the corresponding p-value is $p = 0.073$. Since this is greater than the common threshold of $p < 0.05$, we would conclude that the data are compatible with the null hypothesis, i.e., the student guessed the answer (or at least we cannot say they did not guess).

However, we never specified that we wanted to ask 12 questions – in the discussion so far we only said that there were 12 questions and

we assumed that in this study a set of 12 questions was asked and we observed 9 correct answers (out of 12). We could equally say that we would keep on asking the student questions until they got three questions wrong, and in the observed data 12 questions are what it took to obtain three wrong ones. If we want to determine the p-value for this setting, we cannot use the binomial distribution where we fix the number of trials but need to use the negative-binomial distribution where the number of mistakes or failures is fixed. If we calculate the p-value under this assumption, we obtain $p = 0.0330.05$, i.e., we would conclude that this result is significant and unlikely to be due to random guessing.
However, the data are exactly the same (12 questions, nine correct answers, three wrong answers). This illustrates that we need to be cautious not only to report the data and the hypothesis we want to test (random guessing or not), but also the setting and assumptions we have made when we collected the data and performed the experiment, in other words we also need to know the data-generating process.

Self-Check Questions

1. True/False. The p-value gives the probability that the null hypothesis is true.

2. True/False. The p-value gives the probability of observing data that is at least as extreme as the data I have already observed.

3. True/False. The p-value gives the probability that you are making a wrong decision if your reject the null hypothesis.

4. True/False. The p-value is a random variable.

Solutions

1. False

2. True

3. False

4. True

7.3 Multiple Hypothesis Testing

We have already seen in the discussion above that just reporting or publishing a "significant" result based on the p-value can be very misleading if we do not describe the way we test the hypothesis, as well as report all those tests that did not yield a significant result. This can be illustrated intuitively with the following example (Munroe, n.d.): Suppose we read in a publication that jellybeans cause acne and further "analysis" revealed that green jelly beans are found to cause acne with a level of significance of $p < 0.05$. Does this mean that we should avoid green jellybeans? Quite commonly a scenario similar to the following happens: Jellybeans come in many different colors and each color was tested separately, meaning that we analyzed the hypothesis "people who only eat purple jelly beans do not get acne", as well as "people who only eat blue jelly beans do not get acne" and so on for each of the many colors, including green. Since there are many colors, it so happens that the threshold of $p < 0.05$ is crossed for green jellybeans. Recall that we have seen earlier that this threshold value means that one in 20 results will be wrong. By testing many colors, we expect that some results will cross this threshold just because the p-value is a random variable and the threshold is arbitrary and chosen by convention. Does that mean that green jellybeans cause acne? After all, we have observed a significant result when we disregard all the other hypotheses we have tested. If we do not report that we have tested tens, maybe hundreds of different colors of jellybeans, we would maybe get away with claiming that we observed a statistically significant result "proving" that green jellybeans cause acne. However, if we also report that we tested 20 different hypotheses (one per color), reporting one significant result is exactly what we expect to happen. If we were to test 100 or 1000 different shades of jellybean, we would expect even more statistically significant results that are due to random chance alone. In other words, if we perform multiple hypothesis tests, we need a mechanism that allows to prevent us from obtaining statistically significant results by testing many hypotheses

and "cherry pick" the ones that give us promising results.

Suppose that 100 students take the 12-question factual true/false quiz we have discussed earlier. We want to see whether any of the students did not guess on the quiz. In this context, we are performing 100 hypothesis tests, one for each student. Let us say that for each hypothesis test, we choose a five percent significance level, that is $\alpha = 0.05$. The null hypotheses, $H_{(1)}, H_{(2)}, \ldots, H_{(100)}$, 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. The question is, which value of α should we choose so that the probability that at least one true null hypothesis is rejected is small, say 0.05? In our case, with 100 hypotheses, choosing $\alpha = 0.0005$ suffices:

$$\begin{aligned} 1 - (1-\alpha)^{100} &< 0.05 \\ \Rightarrow (1-\alpha)^{100} &> 0.95 \\ \Rightarrow 1 - \alpha &> (0.95)^{1/100} \approx 0.9995 \\ \Rightarrow \alpha &\leqslant 0.005 \end{aligned}$$

The Familywise Error Rate aims at reducing the overall probability of rejecting true null hypothesis in multiple hypothesis testing.

Controlling this type of error measure is called **Familywise Error Rate (FWE)**. The typical method to control FWE is the Bonferroni method (Bonferroni, 1936), 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 example: $\alpha/m = 0.05/100 = 0.0005$. This is exactly the same value we obtained above, in fact, when p is small, $1 - (1-p)^{1/100} \approx p/100$, so $1 - (0.95)^{1/100} = 1 - (1 - 0.05)^{1/100} \approx 0.05/100 = 0.0005$

The issue with this type of control is that the multiple testing procedure might result in low power. Recall that power is the ability to reject a false null hypothesis. It is very strict to not even allow one true hypothesis to be rejected. We need an alternative measure.

When we have many hypothesis tests, it makes sense to allow a small proportion of true null hypotheses to be rejected, meaning that we accept a fraction of "false discoveries" as long as we are able to quantify the level and can introduce a control method that allows us to specify the level at which we accept these false discoveries. This way, the statistical power of our tests will not be too low. To this end, we will define a measure that controls the proportion of true null hypotheses rejected. Before defining this measure formula, let us introduce some quantities we will need:

	H is accepted	H is rejected	Total
H is true	FN	FD	T_0
H is false	FN	TD	T_1
Total	N	D	m

Where each entry is the following

- H is the (null) hypothesis to be tested
- m is the total number of hypotheses to be tested
- T/F indicates True or False
- D: discovery, i.e., we reject the null hypothesis and observe a statistically significant effect
- N: non-discovery, i.e., we cannot reject the null hypothesis

Hence, "TD" is a true discovery where the null hypothesis is false and we reject it. This is what we want to achieve. consequently, "FD", a false discovery, is not what we want to do as the null hypothesis is true, but we fail to reject it.

Unfortunately, we do not know if the null hypothesis is true or not, and therefore all the quantities with a "T" or a "F" are not accessible. What we can measure is the number of experiments (or hypothesis tests) we do, the number of discoveries "D" where we exceed the threshold and the number of "N" where we do not exceed the threshold.

In a single hypothesis test, we want to have a small type I error, the probability of rejecting a true null hypothesis. In other words, we want to control the probability of a false positive (detecting an effect when there is none). The corresponding quantity for multiple hypothesis test is the false discovery rate (FDR), defined as the expected proportion of the false positive with respect to all positives:

$$FDR = E\left(\frac{FD}{D}\right) = E\left(\frac{FD}{FD+TD}\right) \tag{7.34}$$

We need to compute the expectation value, because the quantities are random variables, however, unfortunately, the expectation value is impossible

to compute because the random variables FD and TD are not observable. For uncorrelated or positively correlated test we can use the following approach (Benjamini & Hochberg, 2000): For each test, we define a null hypothesis $H_1, H_2, \ldots, H_m$, each with the associated p-value. Then we reorder the hypotheses in increasing order with respect to their p-values:

$$H_{(1)}, H_{(2)}, \ldots, H_{(m)} \tag{7.35}$$

where the indices denote the ordering

$$p_{(1)} \leqslant p_{(2)} \leqslant \ldots \leqslant p_{(m)} \tag{7.36}$$

Next, we choose the largest k, such that

$$p_{(k)} \leqslant \frac{k}{m} \cdot \alpha \tag{7.37}$$

Then, the hypotheses $H_{(1)}, \ldots, H_{(k)}$ (with the new ordering) are rejected, and we say that we control the FDR at level α. Additionally, we choose the value of α such that the number of rejections is maximised (Benjamini & Hochberg, 2000).

In case that the hypotheses are correlated between the tests, we modify the procedure such that

$$P_{(k)} = \frac{k}{m \cdot c(m)}\alpha \tag{7.38}$$

where $c(m) = 1$ for uncorrelated or positively correlated hypothesis (this is then the same as above) or $c(m) = \sum_{i=1}^{m} \frac{1}{i}$ (Benjamini & Yekutieli, 2001).

Self-Check Questions

1. If 10 hypotheses are tested simultaneously, each at $\alpha = 0.1$, what is the probability of rejecting a true null hypothesis?

2. What measure does the Bonferroni method control and how?

3. What is a downside of the Bonferroni method?

4. What is the false discovery rate ?

Solutions

1. The probability of rejecting a true null hypothesis is $1 - (1 - \alpha)^n = 1 - (1 - 0.1)^{10} = 1 - 0.9^{10} = 0.65$, i.e., 65%.

2. The Bonferroni method controls a family wise error, specifically the probability of rejecting at least one true null hypothesis. The correction to the significance level is α/m, where m is the number of null hypotheses and α is the significance level for each hypothesis.

3. A downside of the Bonferroni method is that controlling a family wise error is too strict and will often result in low power.

4. The FDR is the proportion of true null hypotheses rejected with respect all null hypotheses rejected.

Summary

Hypothesis testing provides a framework for observed data to research hypotheses. A null hypothesis, representing the status quo, and a test hypothesis, representing an effect to detect, are designed. The random number Test Statistic aggregates the observed data. With its distribution and observed value, we have some evidence to shape our hypotheses.

Whether we reject or fail to reject the null hypothesis, hypothesis testing is prone to two types of errors: type I and type II. Type I error occurs when we reject a true null hypothesis. Type II error occurs when we fail to reject a false null hypothesis. T here are considerations with respect to these errors when determining a threshold for the test statistic. It is important that we understand the repercussions of setting a probability of type I error (α) and how this affects other factors in our hypothesis test such as the power.

The p-values offer a different way to report the results of a hypothesis test. They are meant to give information beyond whether the observed test statistic falls on one side of the cutoff of or not. However, since the misconception about p-values and their interpretation is prominent in the research community, great care must be taken to understand its significance. In this section, we summarized some of these pitfalls and hints on how to properly interpret p-values.

Finally, we discussed testing multiple sets of hypotheses simultaneously. The main issue that comes up with multiple hypothesis testing is how the probability of type I error increases dramatically with more hypotheses. Therefore, we discussed two general ways of controlling reject true null hypotheses: a family wise error control and a false discovery rate control.

References

Abernethy, R. B., Breneman, J., Medlin, C., & Reinman, G. L. (1983). *Weibull analysis handbook* (Tech. Rep.). West Palm Beach FL 33482, USA: Pratt and Whitney Aircraft.

Aggarwal, R., & Caldwell, A. (2012). Error bars for distributions of numbers of events. *The European Physical Journal Plus*, *127*(2), 24.

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, *27*(1), 17–21. doi: \url{10.1080/00031305.1973.10478966}

Asllani, A., & Naco, M. e. *al.*. (2015). Using benford¡¯ s law for fraud detection in accounting practices. *Journal of social science studies*, *2*(1), 129–143.

Barabesi, L., Cerasa, A., Cerioli, A., & Perrotta, D. (2018). Goodness-of-fit testing for the newcomb-benford law with application to the detection of customs fraud. *Journal of Business & Economic Statistics*, *36*(2), 346–358.

Bayes, T. (1763). Lii. an essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price, in a letter to john canton, amfr s. *Philosophical transactions of the Royal Society of London*(53), 370–418.

Bellman, R. (1957). *Dynamic programming* (1st ed.). Princeton, NJ, USA: Princeton University Press.

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American philosophical society*, 551–572.

Beniger, J. R., & Robyn, D. L. (1978). Quantitative graphics in statistics: A brief history. *The American Statistician*, *32*(1), 1–11.

Benjamini, Y., & Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of educational and Behavioral Statistics*, *25*(1), 60–83.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 1165–1188.

Berger, A., & Hill, T. P. (2015). *An introduction to benford's law*. Princeton University Press. Retrieved from `http://www.jstor.org/stable/j.ctt1dr35m0`

Berger, A., Hill, T. P., & Rogers, E. (2009). *Benford online bibliography*. http://www.benfordonline.net. (Accessed: 2020-11-11)

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Blobel, V., & Lohrmann, E. (2012). *Statistische und numerische methoden der datenanalyse* ([2. Aufl] ed.). Hamburg: V. Blobel.

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, *8*, 3–62.

Bowden, G., Barker, P., Shestopal, V., & Twidell, J. (1983). The weibull distribution function and wind power statistics. *Wind Engineering*, 85–98.

Buck, B., Merchant, A., & Perez, S. (1993). An illustration of benford's first digit law using alpha decay half lives. *European Journal of Physics*, *14*(2), 59.

Bundesamt, S. (n.d.-a). *Bevölkerung nach geschlecht und staatsangehörigkeit.* https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Bevoelkerungsstand/Tabellen/liste-zensus-geschlecht-staatsangehoerigkeit.html. (Accessed: 2020-11-11)

Bundesamt, S. (n.d.-b). *Zensus 2011: Bereitstellung der daten über die forschungsdatenzentren der statistischen ämter des bundes und der länder.*

Casella, G., & Berger, R. L. (2002). *Statistical inference* (No. 2). Duxbury.

Casscells, W., Schoenberger, A., & Graboys, T. B. (1978). Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine*, *299*(18), 999–1001.

Chatfield, C., & Goodhardt, G. J. (1973). A consumer purchasing model with erlang inter-purchase time. *Journal of the American Statistical Association*, *68*(344), 828–835. Retrieved from `http://www.jstor.org/stable/2284508`

Chavalarias, D., Wallach, J., Li, A., & Ioannidis, J. (2016). Evolution of reporting p values in the biomedical literature. *JAMA*, *315*(11), 1141–1148.

Clarke, R. (1946). An application of the poisson distribution. *Journal of the Institute of Actuaries*, *72*(3), 481–481.

Deisenroth, M. P., Faisal, A. A., & Ong, C. S. (2020). *Mathematics for machine learning.* Cambridge University Press.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

Dickinson, J. A. (2016). Lesser-spotted zebras: Their care and feeding. *Canadian Family Physician*, *62*(8), 620–621.

Diekmann, A., & Jann, B. (2010). Benford's law and fraud detection: Facts and legends. *German economic review*, *11*(3), 397–401.

Eddy, D. (1982). Judgment under uncertainty: Heuristics and biases. *Judgment under uncertainty: Heuristics and biases*.

Ehrenberg, A. (1972). *Repeat-buying; theory and applications.* North-Holland Pub. Co.

Ehrenberg, A. S. C. (1959). The pattern of consumer purchases. *Journal of the Royal Statistical Society Series C*(1), 26–41. Retrieved from `http://search.ebscohost.com.pxz.iubh.de:8080/login.aspx?direct=true&db=edsrep&AN=edsrep.a.bla.jorssc`

.v8y1959i1p26.41&site=eds-live&scope=site
Fink, D. (1997). A compendium of conjugate priors. *See http://www. people. cornell. edu/pages/df36/CONJINTRnew% 20TEX. pdf*, *46*.
Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *222*(594-604), 309–368.
Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, *7*(2), 179–188.
Franconeri, S. (2019). *Which visualization? a quick reference.* http://experception.net. (Accessed: 2020-11-17)
Friendly, M., & Denis, D. (2005). The early origins and development of the scatterplot. *Journal of the History of the Behavioral Sciences*, *41*(2), 103–130.
Galton, F. (1894). *Natural inheritance*. Macmillan and Company.
Gauss, C. F. (1877). *Theoria motus corporum coelestium in sectionibus conicis solem ambientium* (Vol. 7). FA Perthes.
Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.
Gholamy, A., & Kreinovich, V. (2017). What is the optimal bin size of a histogram: an informal description.
Gigerenzer, G., & Hoffrage, U. (1995). How to improve bayesian reasoning without instruction: frequency formats. *Psychological review*, *102*(4), 684.
Goodhardt, G. J., & Ehrenberg, A. (1967, 5). Conditional trend analysis: A breakdown by initial purchasing level. *Journal of Marketing Research*, *4*, 155–161.
Greenwood, M., & Yule, G. U. (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of the Royal statistical society*, *83*(2), 255–279.
Hall, P., & Strutt, J. (2003). Probabilistic physics-of-failure models for component reliabilities using monte carlo simulation and weibull analysis: a parametric study. *Reliability Engineering & System Safety*, *80*(3), 233–242.
Haskell, A. C. (1919). *How to make and use graphic charts*. Codex book Company Incorporated.
He, K., & Meeden, G. (1997). Selecting the number of bins in a histogram: A decision theoretic approach. *Journal of Statistical Planning and*

inference, *61*(1), 49–59.
Held, L. (2008). *Methoden der statistischen inferenz : Likelihood und bayes*. Heidelberg: Spektrum Akademischer Verlag.
Hill, T. P. (1995). A statistical derivation of the significant-digit law. *Statistical science*, *10*(4), 354–363.
Hintze, J. L., & Nelson, R. D. (1998). Violin plots: a box plot-density trace synergism. *The American Statistician*, *52*(2), 181–184.
Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.
Hogg, R., McKean, J., & Craig, A. (2020). *Introduction to mathematical statistics* (eigth edition (global) ed.). Pearson Education.
Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, *24*(6), 417.
Hubert, M., & Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational statistics & data analysis*, *52*(12), 5186–5201.
Izenman, A. J. (1991). Review papers: Recent developments in nonparametric density estimation. *Journal of the American Statistical Association*, *86*(413), 205–224.
Jamain, A. (2001). Benford's law. *Unpublished Dissertation Report, Department of Mathematics, Imperial College, London*.
Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, *186*(1007), 453–461.
Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202.
Khinchina, A. (1929). On the law of large numbers. *Reports of the Acad 'e mie des Sciences*, *189*, 477–479.
Kolmogorov, A. (1956). *Foundations of the theory of probability .(trad. n. morrison)*. New York: Chelsea Publishing Company.(Original en alemán, 1933).
Krul, A. J., Daanen, H. A. M., & Choi, H. (2010, 01). Self-reported and measured weight, height and body mass index (BMI) in Italy, the Netherlands and North America. *European Journal of Public Health*, *21*(4), 414-419. Retrieved from `https://doi.org/10.1093/eurpub/ckp228` doi: 10.1093/eurpub/ckp228
Kurt, W. (2019). *Bayesian statistics the fun way: Understanding statistics and probability with star wars, lego, and rubber ducks*. No Starch

Press.

Liu, Y., & Abeyratne, A. I. (2019). *Practical applications of bayesian reliability*. Wiley. doi: 10.1002/9781119287995

Mahmood, F. H., Resen, A. K., & Khamees, A. B. (2020). Wind characteristic analysis based on weibull distribution of al-salman site, iraq. *Energy Reports*, *6*, 79–87.

Marcus, U., Gunsenheimer-Bartmeyer, B., Kollan, C., & Bremer, V. (2019). Hiv-jahresbericht 2017/2018. *Epid Bull*, *46*, 493–501.

Martinez, A. M., & Kak, A. C. (2001). Pca versus lda. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(2), 228-233. doi: 10.1109/34.908974

Mateika, J., & Fitzmaurice, G. (n.d.). *Same stats - data and images.* Retrieved 04/11/2020, from `https://www.autodesk.com/content/dam/autodesk/www/autodesk-reasearch/Publications/pdf/SameStatsDataAndImages.zip`

Mateika, J., & Fitzmaurice, G. (2017). Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 1290–1294.

Milcent, M. (2014). *Benford for python.* https://github.com/milcent/benford_py. (Accessed: 2020-11-13)

Moyal, J. E. (1955). Xxx. theory of ionization fluctuations. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *46*(374), 263–280. doi: \url{10.1080/14786440308521076}

Munroe, R. (n.d.). *Signifikant.* `https://xkcd.com/882/`. (Accessed: 2020-11-25)

Murdoch, D. J., Tsai, Y.-L., & Adcock, J. (2008). P-values are random variables. *The American Statistician*, *62*(3), 242–245.

M. Waskom, t. (2020, September). *mwaskom/seaborn.* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.592845` doi: 10.5281/zenodo.592845

Newcomb, S. (1881). Note on the frequency of use of the different digits in natural numbers. *American Journal of mathematics*, *4*(1), 39–40.

Pain, J.-C. (2008). Benford's law and complex atomic spectra. *Physical Review E*, *77*(1), 012102.

Parmenter, D. (2015). *Key performance indicators: developing, implementing, and using winning kpis*. John Wiley & Sons.

Pearson, K. (1895). X. contributions to the mathematical theory of evolution.—ii. skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London.(A.)*(186), 343–414.

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559–572.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pérez, I. A., Sánchez, M. L., & García, M. Á. (2007). Weibull wind speed distribution: Numerical considerations and use with sodar data. *Journal of Geophysical Research: Atmospheres*, *112*(D20).

Pietronero, L., Tosatti, E., Tosatti, V., & Vespignani, A. (2001). Explaining the uneven distribution of numbers in nature: the laws of benford and zipf. *Physica A: Statistical Mechanics and its Applications*, *293*(1-2), 297–304.

Pitman, J. (1997). Some probabilistic aspects of set partitions. *The American mathematical monthly*, *104*(3), 201–209.

Playfair, W. (1801). *The statistical breviary; shewing the resources of every state and kingdom in europe.* J. Wallis.

Raiffa, H., & Schlaifer, R. (1961). Applied statistical decision theory.

Rao, C. R. (1948). The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, *10*(2), 159–203.

Raschka, R. (2014a). *Implementing a principal component analysis (pca).* `https://sebastianraschka.com/Articles/2014_pca_step_by_step.html`. (Accessed: 2020-11-25)

Raschka, R. (2014b). *Linear discriminant analysis– bit by bit.* `https://sebastianraschka.com/Articles/2014_python_lda.html`. (Accessed: 2020-11-25)

Ryder, P. (2009). Multiple origins of the newcomb-benford law: rational numbers, exponential growth and random fragmentation.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, *2*(6), 110–114.

Schmittlein, D. C., Bemmaor, A. C., & Morrison, D. G. (1985). Technical note - why does the nbd model work? robustness in representing product purchases, brand purchases and imperfectly recorded purchases. *Marketing Science*, *4*(3), 255–266. Retrieved from `http://dx.doi.org/10.1287/mksc.4.3.255` doi: 10.1287/mksc.4.3.255

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika*, *66*(3), 605–610.

Shaw, L. P., & Shaw, L. F. (2019). The flying bomb and the actuary. *Significance*, *16*(5), 12–17.

Slepkov, A. D., Ironside, K. B., & DiBattista, D. (2015). Benford's law: Textbook exercises and multiple-choice testbanks. *PLoS One*, *10*(2), e0117972.
Spear, M. E. (1952). Charting statistics.
Spear, M. E. (1969). Practical charting techniques.
Stahl, S. (2006). The evolution of the normal distribution. *Mathematics magazine*, *79*(2), 96–113.
Sturges, H. A. (1926). The choice of a class interval. *Journal of the american statistical association*, *21*(153), 65–66.
Sulewski, P. (2020). Equal-bin-width histogram versus equal-bin-count histogram. *Journal of Applied Statistics*, 1–20.
Tam Cho, W. K., & Gaines, B. J. (2007). Breaking the (benford) law: Statistical fraud detection in campaign finance. *The american statistician*, *61*(3), 218–223.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.
Tukey, J. W. (1970). *Limited preliminary edition* (Vol. 1). Addison-Wesley Publishing Co.
Tukey, J. W. (1977). Box-and-whisker plots. *Exploratory data analysis*, 39–43.
VanderPlas, J. (2016). *Python data science handbook: Essential tools for working with data.* " O'Reilly Media, Inc.".
Venkatesan, R. (2014). *The complete journey.* https://www.dunnhumby.com/careers/engineering/sourcefiles. (Visited on 2019-Sep-10)
Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems ofp values. *Psychonomic bulletin & review*, *14*(5), 779–804.
Wasserman, L. (2013). *All of statistics: a concise course in statistical inference.* Springer Science & Business Media.
Wasserstein, R. L., & Lazar, N. A. (2016). The asa statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133.
Weber, F., Hoang Do, J., & Chung, S. e. a. (2018). Regulation of rem and non-rem sleep by periaqueductal gabaergic neurons. *Nature Communications*, *9*(354).
Weibull, W. (1939). A statistical theory of the strength of materials. *Swed. Inst. Eng. Res*, *151*(545), 7.
Welch, B. L. (1946). The generalization of "student's" problem when several different population variances are involved. *Biometrika*, *34*,

28–35.
Whyman, G., Shulzinger, E., & Bormashenko, E. (2016). Intuitive considerations clarifying the origin and applicability of the benford law. *Results in Physics*, *6*, 3–6.
Wickham, H., & Stryjewski, L. (2011). 40 years of boxplots. *Am. Statistician*.
Wu, C. F. J. (1983). On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, *11*(1), 95 – 103. doi: 10.1214/aos/1176346060
Yong, T. (2004). *Extended weibull distributions in reliability engineering* (Unpublished doctoral dissertation). National University of Singapore.
Zhang, T., & Xie, M. (2007). Failure data analysis with extended weibull distribution. *Communications in Statistics—Simulation and Computation®*, *36*(3), 579–592.

List of Figures