Lecture notes for MA52112 (Statistics for Data Science)

Karim Anaya-Izquierdo (based on notes by Vangelis Evangelou)

2025-11-09

Table of contents

Overview of Statistics for Data Science

Synopsis

In this unit you will develop your understanding of the basic theory of probability and statistics and recognise when this theory can be applied in practice.

Learning outcomes

By the end of the unit you will be able to:

- perform elementary mathematical operations in probability and statistics
- translate real-world problems into a probabilistic or statistical framework
- solve statistical problems in abstract form
- critically interpret the outcomes of statistical analysis in a real-world context
- relate underlying theory to requirements in practical data science

Content

The laws of probability. Discrete and continuous random variables. Expectation, variance and correlation. Conditional and marginal distributions. Common distributions including the normal, binomial and Poisson. Statistical estimation including maximum likelihood. Hypothesis testing and confidence intervals.

Summative assessment

- **Exam:** 100% of unit mark.

Moodle page

Please see the Moodle page for this unit for more a more detailed overview on the organisation and expectations for Statistics for Data Science this year.

1 Probability Theory for Data Scientists

1.1 Set theory Concepts

Definition: Sample Space, Event, and Empty Set

Definition 1.1. Consider an uncertain scenario. This include a random experiment, a data-generating process or simply the future. We define the following concepts:

[1]

- **Sample Space (Ω)**: The set of all possible outcomes or results from the scenario. Sample spaces can be either countable or uncountable. If the elements of a sample space can be put into one-to-one correspondence with the set of integers, the sample space is countable. If the sample space contains only a finite number of elements, it is also countable. Otherwise is uncountable.
- **Event**: A subset of the sample space. It represents a specific outcome or a collection of outcomes of interest.
- **Empty Set ($\emptyset$)**: A set containing no elements. It represents an impossible event.

Example 1.1. If we flip a coin twice then the sample space can be written as:

$$\Omega = \{HH, HT, TH, TT\}$$

where H represents *heads* and T *tails*. This sample space is finite. An event (say A) could be *at least one head appears*, that is

$$A = \{HT, TH, HH\} \subset \Omega$$

Example 1.2. If we are analyzing customer purchase behavior for a single online transaction, the sample space could be the set of all possible combinations of items a customer might select from the store's catalog. This sample space is in principle finite and therefore countable. However, if the catalog is very large, the sample space can be considered uncontably large for practical purposes. More on this later.

An event could be "customer buys at least one item from category X", or "customer buys product Y".

Example 1.3. We measure the time (in seconds) it takes for a user to complete a task on a website. The time limit is predefined at 5 minutes. Then the sample space is $\Omega = \{0, 1, 2, 3, \ldots, 300\}$ which is finite. If, however, we measure the time with arbitrary precision, then the sample space is the interval $(0, 300)$ of real numbers. This sample space is uncountable.

An event could be "user completes the task in under 2 minutes". In the former case, this corresponds to the set $A = \{1, 2 \ldots, 119\}$. Int he latter case is the real interval $A = (0, 120)$.

Events can be described in many different ways. We will use set theory and notation to describe events and operations on events. This can help later in the computation of probabilities.

Basic Set Operations

Definition 1.2. Given events A, B, C in the sample space Ω:

- **Union ($A \cup B$)**: The event that A occurs, or B occurs, or both occur.
- **Intersection ($A \cap B$)**: The event that both A and B occur.
- **Complement (A^c)**: The event that A does not occur. It is the set of all outcomes in Ω that are not in A.

The following propertties hold for any events A, B, C:

- **Commutativity**:
  - Union: $A \cup B = B \cup A$
  - Intersection: $A \cap B = B \cap A$
- **Associativity**:
  - Union: $(A \cup B) \cup C = A \cup (B \cup C)$
  - Intersection: $(A \cap B) \cap C = A \cap (B \cap C)$
- **Distributive Laws**:
  - Intersection over Union: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
  - Union over Intersection: $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
- **De Morgan's Laws**:
  - $(A \cup B)^c = A^c \cap B^c$
  - $(A \cap B)^c = A^c \cup B^c$

Disjoint Sets and Partitions of Sample Space

Definition 1.3.

- **Disjoint Sets (Mutually Exclusive Events)**: Two sets A and B are disjoint if they have no elements in common, i.e., $A \cap B = \emptyset$.
- **Partition**: A collection of non-empty, disjoint subsets (events) of Ω whose union is Ω. That is $A_1, A_2, \ldots$ is a partition if

$$\bigcup_i A_i = \Omega \quad \text{and} \quad A_i \cap A_j = \emptyset \text{ for } i \neq j$$

Representation of events using set operations

Example 1.4. When we flip a coin twice, the event A "at least one head appears" can be written in various ways. These include:

- the union of three events $A = \{HT\} \cup \{TH\} \cup \{HH\}$. That is, A occurs if we get heads on the first flip and tails on the second flip, or tails on the first flip and heads on the second flip, or heads on both flips. Note that these three events are disjoint as they do not share any outcomes.
- the union $A = A_1 \cup A_2$ where $A_1 = \{HT, HH\}$ is the event "head on first flip" and $A_2 = \{TH, HH\}$ is the event "head on second flip". Note that A_1 and A_2 are not disjoint as they both contain the outcome HH.
- the complement $A = B^c$ where $B = \{TT\}$ is the event "no heads appear".

Three different partitions of the sample space are given by:

- The trivial partition where each event contains a single outcome:

$$\mathcal{P}_1 = \{\{HT\}, \{TH\}, \{HH\}, \{TT\}\}$$

- The partition:

$$\mathcal{P}_{equal} = \{\{HH, TT\}, \{HT, TH\}\}$$

  that is, when we flip the coin twice, either we get the same results in both throws OR different ones.

- The partition where we group the outcomes based on the number of heads:

$$\mathcal{P}_{heads} = \{\{TT\}, \{HT, TH\}, \{HH\}\}$$

that is, when we flip the coin twice, we can get no heads, one head or two heads.

Sigma Algebra

Definition 1.4. A collection $\mathcal{F}$ of subsets of Ω is a **sigma algebra** (or σ-algebra) if it satisfies the following properties:

1. $\Omega \in \mathcal{F}$ (The sample space is in the collection).
2. If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$ (The collection is closed under complementation).
3. If $A_1, A_2, \ldots$ are in $\mathcal{F}$, then
$$\bigcup_i A_i \in \mathcal{F}$$

that is, the collection is closed under arbitray number of unions.

Note the definition of sigma-algebra does not explicitly require that the intersection of two sets in $\mathcal{F}$ is also in $\mathcal{F}$. However, this property follows from the other properties and De Morgan's laws. If $A, B \in \mathcal{F}$, then

$$A \cup B \in \mathcal{F} \implies (A \cup B)^c = A^c \cap B^c \in \mathcal{F} \implies (A^c \cap B^c)^c = A \cup B \in \mathcal{F}$$

$$A^c \in \mathcal{F}, B^c \in \mathcal{F} \implies A^c \cup B^c \in \mathcal{F} \implies (A^c \cup B^c)^c = A \cap B \in \mathcal{F}$$

(corrected from previous version)

Examples of sigma-algebras

Example 1.5. The trivial sigma algebra is clearly $\mathcal{F}_0 = \{\emptyset, \Omega\}$ which does not seem very useful.
The partition $\mathcal{P}_{equal} = \{\{HH, TT\}, \{HT, TH\}\}$ above, is not a sigma-algebra as it does not contain the empty set. If we add the empty set, then is still not a sigma algebra as it is not closed under union. The union of the only two elements is Ω. If we include Ω then we have the sigma algebra:

$$\mathcal{F}_{equal} = \{\emptyset, \Omega, \{HH, TT\}, \{HT, TH\}\}$$

The partition $\mathcal{P}_{heads}$ above is also not a sigma algebra but if we add all possible unions then we obtain the sigma algebra:

$$\begin{aligned}\mathcal{F}_{heads} = \{&\emptyset, \Omega, \{TT\}, \{HT, TH\}, \{HH\}, \{HT, TH, HH\},\\ &\{HT, TH, TT\}, \{HH, TT\}\}\end{aligned}$$

The set

$$\mathcal{G} = \{\emptyset, \Omega, \{HT\}, \{TH\}, \{HH\}, \{TT\}\}$$

obatined from $\mathcal{P}_1$ is neither a partition nor a sigma algebra as it is not closed under union. For example, $\{HT\} \cup \{TH\} = \{HT, TH\} \notin \mathcal{G}$. However, if we add all possible unions of the elements of $\mathcal{G}$ we obtain the **power set** of Ω, that is the set of all subsets of Ω:

$$\begin{aligned}\mathcal{F}_{max} = \{&\emptyset, \{HT\}, \{TH\}, \{HH\}, \{TT\},\\ &\{HT, TH\}, \{HT, HH\}, \{HT, TT\}, \{TH, HH\}, \{TH, TT\}, \{HH, TT\}\\ &\{HT, TH, HH\}, \{HT, TH, TT\}, \{HT, HH, TT\}, \{TH, HH, TT\}, \Omega\}\end{aligned}$$

This is the largest possible sigma-algebra for this sample space. It has $2^4 = 16$ elements since the sample space has 4 elements. In general, if the sample space has n elements, then its power set has 2^n elements.
Also generally, if we have a finite partition of Ω then the collection of all unions of sets in the partition (including the empty set) is a sigma-algebra.
Note that different sigma algebras serve for different purposes. For example, the sigma algebra $\mathcal{F}_{equal}$ is useful if we are only interested in whether the two coin flips are the same or different. The sigma algebra $\mathcal{F}_{heads}$ is useful if we areinterested in the number of heads. The power set $\mathcal{F}_{max}$ is a sigma algebra that me be more useful if we are interested in all possible events.
In this example we also observe that:

$$\mathcal{F}_0 \subset \mathcal{F}_{equal} \subset \mathcal{F}_{heads} \subset \mathcal{F}_{max}$$

so that $\mathcal{F}_0$ and $\mathcal{F}_{max}$ are the smallest and largest sigma algebras possible for this sample space.

1.2 Probability

We will start by defining probability in an intuitive way. Later we will give a more formal mathematical definition .

1.2.1 Types of Probability

There are several ways to think about probability. These include

- **Classical Probability:** Assumes all possible outcomes in a finite sample space are equally likely. That is, for any event A with $n(A)$ outcomes in a sample space Ω with $n(\Omega)$ equally likely outcomes, the probability of A is:

$$P(A) = \frac{n(A)}{n(\Omega)}$$

Example 1.6. Under this framework, the probability of rolling an even number on a die is assigned to be $P(\text{rolling an even number}) = \frac{3}{6}$. More, generally this is equivalent to say the die is fair. Another example is when we assign the probability of rain tomorrow, locally at 10 AM, to be 1/2 as there are only two possible outcomes: rain or no rain.

- **Empirical (or Frequentist) Probability:** Based on observed frequencies from repeated experiments. As the number N of experiment repetitions increases, the probability of an event A approaches the true probability:

$$P(A) \approx \frac{\text{Number of times A occurred}}{N}$$

Example 1.7. If we do not what the probability of heads when flipping a coin is. We can we flip the coin 1000 times and if it lands heads 537 times, we would say the empirical probability of heads is 0.537. Furthermore we might say that the true probability of heads is ≈ 0.537 and the important aspect of thios framework is that, in theory, the more times we flip the coin the closer the empirical proportion will be to the true probability. Finally, according to historical data for our location, it has rained 33.6% of the days out of the last 10 years. The empirical probability of rain tomorrow locally at 10 AM is 0.336.

- **Subjective Probability:** Based on personal belief or judgment, often used when objective data is scarce.

Example 1.8. I had a look through the window and is a bit overcast, then I believe the probability of rain tomorrow locally at 10 AM is 0.7. On the other hand, if I am a weather expert from the point of atmospheric physics, I might believe the probability of rain tomorrow locally at 10 AM is 0.9.

1.2.2 Formal definition of probability

After we have chosen a sigma algebra $\mathcal{F}$ that contains events we are interested in, we can define probabilities for all the events in a more formal way.

Probability Measure (Kolmogorov's Axioms)

Definition 1.5. A **probability measure** P on a sample space Ω with a σ-algebra $\mathcal{F}$ is a function $P : \mathcal{F} \to [0, 1]$ that assigns a probability to each event in $\mathcal{F}$ and satisfies the following three axioms:

1. **Non-negativity**: For any event $A \in \mathcal{F}$, $P(A) \geq 0$. The probability of any event is non-negative.
2. **Normalization**: $P(\Omega) = 1$. The probability of the entire sample space (the certain event) is 1.
3. **Additivity (for disjoint events)**: If $A_1, A_2, \ldots, A_n$ are disjoint events in $\mathcal{F}$ (i.e.,

$A_i \cap A_j = \emptyset$ for $i \neq j$), then

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$$

For a countably infinite sequence of disjoint events, this extends to:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

The probability of the union of disjoint events is the sum of their individual probabilities.

Probability measure for the equality of two coin flips

Example 1.9. For the sigma algebra $\mathcal{F}_{equal} = \{\emptyset, \Omega, \{HH, TT\}, \{HT, TH\}\}$ we can define a probability measure simply by specifying:

- $P(\emptyset) = 0$
- $P(\{HH, TT\}) = 0.4$

Note we can compute the probability of the other two events in $\mathcal{F}_{equal}$ using the axioms:

- $P(\Omega) = 1$ (by axiom 2)
- $P(\{HT, TH\}) = P(\Omega) - P(\{HH, TT\}) = 1 - 0.4 = 0.6$ (by axiom 3)

The assignment of probability of $\{HH, TT\}$ to be 0.4 maybe frequentist or subjective but regardless of this, it generates is a valid probability measure as it satisfies all three axioms.

Probability measure for the number of heads in two coin flips

Example 1.10. For the sigma algebra $\mathcal{F}_{heads}$ we can define a probability measure simply by specifying:

- $P(\{HT, TH\}) = 0.5$
- $P(\{TT\}) = 0.1$

The probabilities for the rest of the event in $\mathcal{F}_{heads}$ can be computed using axiom 3 as follows:

- $P(\{HH\}) = 1 - 0.1 - 0.5 = 0.4$

- $P(\{HT, TH, HH\}) = P(\{HT, TH\}) + P(\{HH\}) = 0.5 + 0.4 = 0.9$
- $P(\{HT, TH, TT\}) = P(\{HT, TH\}) + P(\{TT\}) = 0.5 + 0.1 = 0.6$
- $P(\{HH, TT\}) = 0.1 + 0.4 = 0.5$
- $P(\Omega) = 1$ (Trivial but good to double check in practice)
- $P_{heads}(\emptyset) = 1 - 1 = 0$ (Trivial, always true)

As before the probability assignment maybe frequentist or subjective but regardless of this, it generates is a valid probability measure as it satisfies all three axioms.

Probability measure for power set

Example 1.11. For the largest sigma algebra $\mathcal{F}_{max}$ we can define a probability measure simply by specifying probabilities for the four singletons or atoms:

- $P(\{HH\}) = 0.3$
- $P(\{HT\}) = 0.2$
- $P(\{TH\}) = 0.4$

The probabilities for the rest of the events in $\mathcal{F}_{max}$ can be computed using the axioms as follows:

- $P(\{TT\}) = 1 - 0.3 - 0.2 - 0.4 = 0.1$
- $P(\{HT, TH\}) = 0.2 + 0.4 = 0.6$
- $P(\{HT, HH\}) = 0.2 + 0.3 = 0.5$
- $P(\{HT, TT\}) = 0.2 + 0.1 = 0.3$
- $P(\{TH, HH\}) = 0.4 + 0.3 = 0.7$
- $P(\{TH, TT\}) = 0.4 + 0.1 = 0.5$
- $P(\{HH, TT\}) = 0.3 + 0.1 = 0.4$
- $P(\{HT, TH, HH\}) = 0.2 + 0.4 + 0.3 = 0.9$
- $P(\{HT, TH, TT\}) = 0.2 + 0.4 + 0.1 = 0.7$
- $P(\{HT, HH, TT\}) = 0.2 + 0.3 + 0.1 = 0.6$
- $P(\{TH, HH, TT\}) = 0.4 + 0.3 + 0.1 = 0.8$

As before the probability assignment maybe frequentist or subjective but regardless of this, it generates is a valid probability measure as it satisfies all three axioms.

Simple Probability Operations

Proposition 1.1. *From the axioms, we can derive several useful properties:*

- ***Probability of the Complement**: For any event $A \in \mathcal{F}$,*

$$P(A^c) = 1 - P(A)$$

- ***Probability of the empty set***: $P(\emptyset) = 0$.
- ***Probability of the Union of Two Events (General)***: *For any two events* $A, B \in \mathcal{F}$:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

This is known as the addition rule. It accounts for the overlap between events.

Probability of the union

Example 1.12.

$$\begin{aligned} P(\{HT, TH, TT\} \cup \{HH, TT\}) &= P(\{HT, TH, TT\}) + P(\{HH, TT\}) - P(\{TT\}) \\ &= 0.6 + 0.5 - 0.1 \\ &= 1 \end{aligned}$$

clearly this is correct as the union of these two events is Ω.

Boole and Bonferroni inequalities

Theorem 1.1.

- ***Boole's inequality*** *For any events* $A_1, A_2, \ldots, A_n$ *in* $\mathcal{F}$:

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i)$$

This inequality provides an upper bound for the probability of the union of events.

Bonferroni Inequality: *For any events* $A_1, A_2, \ldots, A_n$ *in* $\mathcal{F}$:

$$P\left(\bigcap_{i=1}^{n} A_i\right) \geq 1 - \sum_{i=1}^{n} P(A_i^c)$$

This inequality provides a lower bound for the probability of the intersection of events.

These inequalities, specially Bonferroni's will be useful later. Booles inequality can be proved by induction and Bonferroni's inequality follows from Booles inequality and the properties of complements. These facts can be verified by the reader.

1.3 Conditional Probability

Conditional Probability

Definition 1.6. The **conditional probability** of event A given that event B has occurred, denoted $P(A|B)$, is defined as:

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

provided that $P(B) > 0$. This measures the probability of event A occurring, knowing that event B has already happened.

Example of Conditional Probability

Example 1.13. What is the probability of getting heads on the first coin flip GIVEN that at least one head appears in two flips? This can be expressed as $P(A|B)$ where $A = \{HT, HH\}$ is "head on first flip" and $B = \{HT, TH, HH\}$ is "at least one head appears". We have:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)} = \frac{P(\{HT, HH\})}{P(\{HT, TH, HH\})}$$

since $A \subset B$ in this case. We notice a subtlety here. The event $A = \{HT, HH\}$ (head on the first flip) is not a member of the sigma-algebra $\mathcal{F}_{heads}$. So cannot use the probability measure P from Example 1.10 to compute this conditional probability. However, it is a member of the sigma algebra (the power set) $\mathcal{F}_{max}$ so we might need to use the probabilities using such sigma algebra (see Example 1.11) as follows:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(\{HT, HH\})}{P(\{HT, TH, HH\})} = \frac{0.5}{0.9} \approx 0.556$$

On a more practical situation, if A is "a user makes a purchase" and B is "a user clicks on an advertisement", then $P(A|B)$ is the probability that a user makes a purchase GIVEN that they clicked on the advertisement. This is a key metric for evaluating ad campaign effectiveness.

Two very useful consequences of the above are: the law of total probability that combines the notion of partition with that of conditional probability and Bayes rule that allows us to reverse conditional probabilities.

Law of Total Probability

Proposition 1.2. *Let $A_1, A_2, \ldots$ be a partition of the sample space Ω (recall Definition 1.3). Then for any event $B \in \mathcal{F}$:*

$$P(B) = \sum_i P(B|A_i)P(A_i)$$

Bayes' Rule

Proposition 1.3. *For events A and B where $P(B) > 0$:*

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

If $A_1, \ldots, A_n$ form a partition of Ω, and $P(A_i) > 0$ for all i, then Bayes' Rule can be written using the Law of Total Probability for the denominator:

$$P(A|B) = \frac{P(B|A)P(A)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}$$

The proof of this result is staightforward and left to the reader.

Example of Law of Total Probability and Bayes' Rule

Example 1.14 (Medical Testing)**.** Suppose a rare disease affects 1 in 10,000 people. A test for this disease is 99% accurate:

- If a person has the disease, the test correctly identifies it 99% of the time (True Positive).
- If a person does not have the disease, the test correctly identifies it 99% of the time (True Negative).

Let D be the event that a person has the disease, and T^+ be the event that the test is positive. The probabilities we know are:

- $P(D) = \frac{1}{10000} = 0.0001$ (Prevalence)
- $P(T^+|D) = 0.99$ (Sensitivity - True Positive Rate)
- $P(T^-|D^c) = 0.99$ (Specificity - True Negative Rate)

Before we proceed we note the probability specifications above are emprical.
Suppose we want to find $P(D|T^+)$, the probability that a person actually has the disease given a positive test result.
First, we need $P(T^+)$. A positive test can occur in two ways:

- $(D \cap T^+)$ or
- $(D^c \cap T^+)$

e.g. a partition of A. We also have:

- $P(T^+|D^c) = 1 - P(T^-|D^c) = 1 - 0.99 = 0.01$ (False Positive Rate)
- $P(D^c) = 1 - P(D) = 1 - 0.0001 = 0.9999$

Using the law of total probability:

$$\begin{aligned} P(T^+) &= P(T^+ \cap D) + P(T^+ \cap D^c) \\ &= P(T^+|D)P(D) + P(T^+|D^c)P(D^c) \\ &= (0.99)(0.0001) + (0.01)(0.9999) \\ &= 0.000099 + 0.009999 = 0.010098 \end{aligned}$$

Now, using Bayes' Theorem:

$$\begin{aligned} P(D|T^+) &= \frac{P(T^+|D)P(D)}{P(T^+)} \\ &= \frac{(0.99)(0.0001)}{0.010098} \approx 0.0098 \end{aligned}$$

This may look counter-intuitive. Even with a positive test, there's only about a 0.98% (less than 1%) chance the person actually has the disease! In particular it is a rare disease. This highlights the importance of understanding base rates and conditional probabilities in interpreting results.

Lets now code the previous example in Python. We code a function that returns $P(D|T^+)$ given the prevalence of the disease, the sensitivity and the specificity of the test. WE vary the prevalence to see how it affects the result.

```
import numpy as np
def bayes_medical_test(prevalence, sensitivity, specificity):
    P_D = prevalence  # Prevalence of the disease
    P_T_given_D = sensitivity  # Sensitivity (True Positive Rate)
    P_T_given_not_D = 1 - specificity  # False Positive Rate
    P_not_D = 1 - P_D  # Probability of not having the disease
    # Calculate P(T+)
    P_T = (P_T_given_D * P_D) + (P_T_given_not_D * P_not_D)
    # Calculate P(D|T+) using Bayes' Theorem
    P_D_given_T = (P_T_given_D * P_D) / P_T
    return P_D_given_T
```

```python
# Example usage
prevalence = 1 / 10000  # 1 in 10,000
sensitivity = 0.99  # 99% sensitivity
specificity = 0.99  # 99% specificity
result_10k = bayes_medical_test(prevalence, sensitivity, specificity)
print(f"P(D|T+) = {result_10k:.4f}")

prevalence = 1 / 1000  # 1 in 1,000

result_1k = bayes_medical_test(prevalence, sensitivity, specificity)
print(f"P(D|T+) = {result_1k:.4f}")

prevalence = 1 / 100  # 1 in 100

result_100 = bayes_medical_test(prevalence, sensitivity, specificity)
print(f"P(D|T+) = {result_100:.4f}")
```

```
P(D|T+) = 0.0098
P(D|T+) = 0.0902
P(D|T+) = 0.5000
```

We can see how the prevalence of the disease affects the probability $P(D|T^+)$ significantly. As the disease becomes more common, the probability that a person actually has the disease given a positive test result increases.

The Law of Total probability allows us to calculate the probability of an event A by considering the different ways it can occur through the events in a partition.

Customer churn

Example 1.15. Suppose we have three models, M_1, M_2, and M_3, that are used to predict customer churn. Let $P(M_1) = 0.5$, $P(M_2) = 0.3$, $P(M_3) = 0.2$ be the probabilities that each model is the "best" for a given customer. Let A be the event "customer churns". If we know the probability of churn given each best model (e.g., $P(A|M_1) = 0.1$, $P(A|M_2) = 0.2$, $P(A|M_3) = 0.15$), the Law of Total Probability allows us to find the overall probability of churn:

$$\begin{aligned} P(A) &= P(A|M_1)P(M_1) + P(A|M_2)P(M_2) + P(A|M_3)P(M_3) \\ &= (0.1)(0.5) + (0.2)(0.3) + (0.15)(0.2) \\ &= 0.05 + 0.06 + 0.03 = 0.14 \end{aligned}$$

1.4 Independence

1.4.1 Independence of events

First intuitively, two events, A and B, are considered **independent** if the occurrence of one event does not affect the probability of the other event occurring. Formally,

Independent Events

Definition 1.7. Tvents A and B are independent if:

$$P(A \cap B) = P(A) \times P(B)$$

or equivalently, if either of the following conditions hold:

- $P(A|B) = P(A)$
- $P(B|A) = P(B)$

This means that knowing that event B has occurred gives us no new information about the probability of event A occurring, and vice versa.

Independent event when flipping a coin twice

Consider the following events when flipping a coin twice:

- $A = \{HT, HH\}$ the first flip is heads
- $B = \{TT, HH\}$ the two flips are the same

Then using the probabilities in Example 1.11 we have:

$$P(A \cap B) = P(\{HH\}) = 0.3 \neq P(\{HT, HH\})P(\{TT, HH\}) = 0.5 \times 0.4 = 0.2$$

Therefore these two events are not independent. Of course, the assignment of probabilities here play a role. In this way, if we had assigned $P(\{HH\}) = 0.2$ then the events would have been independent.

An obvious consequence of Definition 1.6 of conditional probability is the so-called multiplication rule.

Multiplication Rule

Proposition 1.4. *For any two events A and B*

$$P(A \cap B) = P(A) \times P(B|A) = P(B) \times P(A|B)$$

Drawing Cards without Replacement

Example 1.16. Imagine drawing two cards from a standard deck without replacement. Let A be the event that the first card is a Heart. $P(A) = \frac{13}{52}$. Let B be the event that the second card is a Heart. Since the first card is not replaced, these events are dependent. The probability of the second card being a Heart *depends* on the first card drawn. $P(B|A)$ (the probability the second card is a Heart, given the first was a Heart) is $\frac{12}{51}$ (as there are 12 Hearts left and 51 total cards). So, the probability of drawing two Hearts in a row is $P(A \cap B) = P(A) \times P(B|A) = \frac{13}{52} \times \frac{12}{51}$.

Note

The outcome of flipping a coin maybe independent of the outcome of any previous coin flips. If you flip a coin and get heads, the probability of getting heads on the next flip should remain as before. Of course, this a simplifying assumption that may not hold in practice. In this course we will make these kind of assumption specially when it involves sequences of events. Not assuming independence for sequences of events make things more complicated for what we want to achieve in this course.

Independence of many events

Definition 1.8. A collection of events $A_1, A_2, \ldots, A_n$ are **mutually independent** if for every subset of size k, e.g. $\{A_{i_1}, A_{i_2}, \ldots, A_{i_k}\}$ (k such that $2 \leq k \leq n$) we have that:

$$P(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}) = P(A_{i_1}) \times P(A_{i_2}) \times \ldots \times P(A_{i_k})$$

For example, three events A, B, and C are mutually independent if all the following conditions hold:

- $P(A \cap B) = P(A)P(B)$
- $P(A \cap C) = P(A)P(C)$
- $P(B \cap C) = P(B)P(C)$
- $P(A \cap B \cap C) = P(A)P(B)P(C)$

1.5 Random Variables

So far we have talked about events, which are subsets of the sample space. In many applications, especially in data science, we are interested in quantifying outcomes numerically. This is where random variables come into play.

Random Variable

Definition 1.9. A **random variable** X is a function that maps outcomes from the sample space Ω to real numbers. That is, $X : \Omega \to \mathbb{R}$. It quantifies the outcomes of a random phenomenon numerically.
The set of all possible values that X can take is called the **image** or **range** of the random variable, denoted as $X(\Omega)$.

Random variable examples

Example 1.17. If Ω is the set of all possible customer orders, a random variable X could be "the total dollar amount spent in an order". For each order (an outcome in Ω), X assigns a specific monetary value. As another example: for a user's session on a website, X could be "the number of pages visited" or the "overall time spent in the website".

Note a random variable is a function defined on the sample space Ω rather than on a sigma algebra, so for each outcome $\omega \in \Omega$, there is a corresponding real number $X(\omega)$. However, events in a sigma algebra can be defined in terms of random variables. For example, the event "the total amount spent in an order is greater than 50 dollars" can be expressed as $\{X > 50\}$.

Random variable: Number of equal coin flips

Example 1.18. When we flip a coin twice, the sample space is $\Omega = \{TT, HT, TH, HH\}$. We can define a very simple random variable X as the "number of times the flips are the same". The mapping would be:

- $X(\{TT\}) = 1$ (both flips are the same)
- $X(\{HT\}) = 0$ (flips are different)
- $X(\{TH\}) = 0$ (flips are different)
- $X(\{HH\}) = 1$ (both flips are the same)

The possible values of X are $X(\Omega) = \{0, 1\}$.

Random variable: Number of heads in two coin flips

Example 1.19. When we flip a coin twice, the sample space is $\Omega = \{TT, HT, TH, HH\}$. We can define a random variable X as the "number of heads" in the two flips. The mapping would be:

- $X(\{TT\}) = 0$ (no heads)
- $X(\{HT\}) = 1$ (one head)
- $X(\{TH\}) = 1$ (one head)

- $X(\{HH\}) = 2$ (two heads)

The possible values of X are $\{0, 1, 2\}$. This random variable quantifies the outcome of the coin flips in terms of the number of heads observed. Also note the order in which the heads appear does not matter for this random variable.

The above two random variables are discrete random variables as they take on a finite or countable number of values, that is $X(\Omega)$ is finite or countable. There are also continuous random variables that can take on any value in a continuous range.

Continuous vs Discrete Random Variables

Example 1.20. Going back to Example 1.3 we have already defined a random variable X as the time to complete a task in a website with a limit of 5 minutes. If we round to the nearest second, then the possible values of X are $\{0, 1, 2, 3, \ldots, 300\}$ and X is a discrete random variable. However, if we do not round then X can take any value in the interval $(0, 300)$ and X is a continuous random variable.

The definition of continuous random variables requires a bit more than simply having an uncountably infinite image set $X(\Omega)$. The definition is a bit thechnical as it involves the notion of probability density function.

Discrete and continuous random variables

Definition 1.10. We say a random variable X is

- **discrete** if it takes on a finite or countably infinite number of distinct values. That is if the image set $X(\Omega)$ is either finite or countably infinite.

The function:

$$f_X(x) = P(X = x) := P(\{\omega : X(\omega) = x\}) \quad \text{for } x \in X(\Omega)$$

is called the **probability mass function (PMF)** of the discrete random variable X. The PMF satisfies:

- $f_X(x) \geq 0$ for all $x \in X(\Omega)$.
- $\sum_{x \in X(\Omega)} f_X(x) = 1$
- **continuous** if there exists a function $f_X(x)$ such that for any two numbers a and b with $a < b$:

$$P(a \leq X \leq b) := P(\{\omega : a \leq X(\omega) \leq b\}) = \int_a^b f_X(x)dx$$

where

- $f_X(x) \geq 0$ for all x and
- $\int_{-\infty}^{\infty} f_X(x)dx = 1$.

The function $f_X(x)$ is called the **probability density function (PDF)** of the random variable X.

The idea is that there are no "gaps", which would correspond to real numbers which have a finite probability of occurring. Instead, continuous random variables never take an exact prescribed value, that is $P(X = x) = 0$ for all x but there is a positive probability that its value will lie in particular intervals which can be arbitrarily small.

Cumulative Distribution Function (CDF)

Definition 1.11. The **cumulative distribution function (CDF)** of a random variable X, denoted by $F_X(x)$, is the function $F : \mathcal{R} \to [0, 1]$ defined by

$$F_X(x) = P(X \leq x) := P(\{\omega \,:\, X(\omega) \leq x\})$$

for any real number x. The CDF gives the probability that the random variable X takes on a value less than or equal to x.

We note that

$$\begin{aligned} P(a \leq X < b) &= F_X(b) - F_X(a) \\ &= \int_a^b f_X(x)dx \end{aligned}$$

so that

$$f_X(x) = \frac{d}{dx} F_X(x) \qquad \forall x \in \mathbb{R}$$

CDF of a discrete random variable

Example 1.21. Consider a discrete random variable X with possible values in the set

$\{0, 1, 2, 3\}$. Assume we probability mass function (pmf) is given by:

$$f_X(x) = P(X = x) = \begin{cases} 0.1 & \text{if } x = 0 \\ 0.3 & \text{if } x = 1 \\ 0.4 & \text{if } x = 2 \\ 0.2 & \text{if } x = 3 \\ 0 & \text{otherwise} \end{cases}$$

then the CDF is given by

$$F_X(x) = P(X \leq x) = \begin{cases} 0 & \text{if } x < 0 \\ 0.1 & \text{if } 0 \leq x < 1 \\ 0.4 & \text{if } 1 \leq x < 2 \\ 0.8 & \text{if } 2 \leq x < 3 \\ 1.0 & \text{if } x \geq 3 \end{cases}$$

We can plot the CDF in Python as follows:

```
import numpy as np
import matplotlib.pyplot as plt

x = np.linspace(-1, 4, 1000)
y = np.piecewise(x, [x < 0, (x >= 0) & (x < 1), (x >= 1) & (x < 2), (x >= 2) & (x < 3), x
plt.step(x, y, where='post')
# emphasize the continuity from the right
plt.scatter([0, 1, 2, 3], [0.1, 0.4, 0.8, 1.0], color='blue')  # filled circles
plt.scatter([0, 1, 2,3], [0, 0.1, 0.4,0.8], color='white', edgecolor='blue')  # open circl
plt.title('CDF of Discrete Random Variable')
plt.xlabel('x')
plt.ylabel('F(x)')
plt.grid()
plt.yticks(np.array([0,0.1, 0.4, 0.8, 1.0]))
plt.show()
```

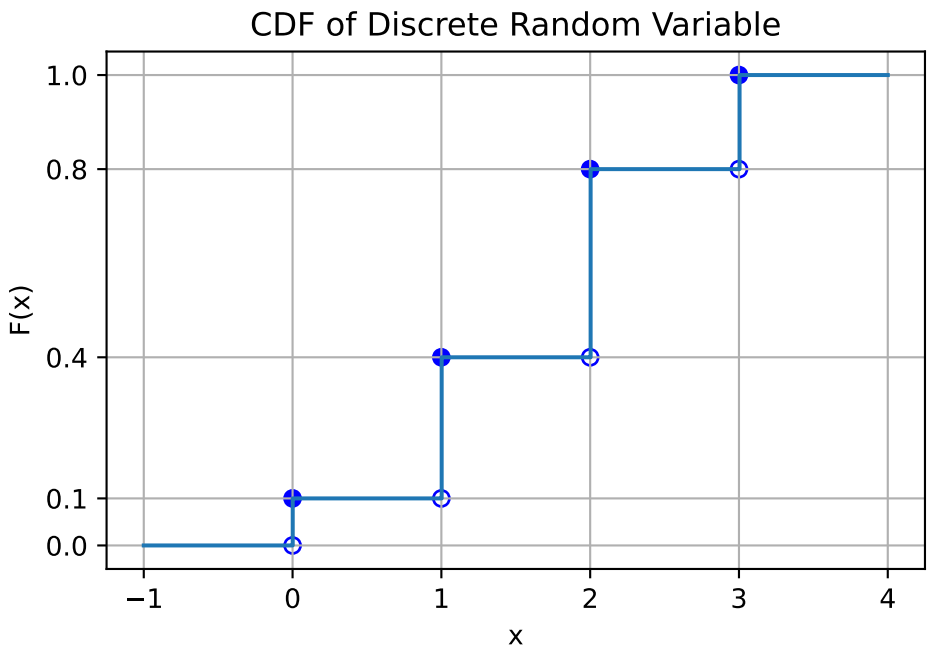


Figure 1.1: CDF of a discrete random variable. Note the function is defined over all real numbers

Figure 1.1 show the CDF is a step function with jumps at the points where the random variable takes values and is continuos from the right

Properties of Cumulative Distribution Functions

For any random variable X, its CDF $F_X(x)$ has the following properties:

1. **Monotonicity**: $F_X(x)$ is non-decreasing. For any $x_1 < x_2$, $F_X(x_1) \leq F_X(x_2)$.
2. **Limits**: $\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to\infty} F_X(x) = 1$.
3. **Right-continuity**: $F_X(x)$ is right-continuous, meaning $\lim_{h\to0^+} F_X(x+h) = F_X(x)$ for all x.

The properties above hold for both discrete and continuous random variables. For discrete random variables, the CDF is a step function (continuous from the right), while for continuous random variables, the CDF is a continuous function.

CDF of a continuous random variable

Example 1.22. Consider a random variable X representing the time (in hours) a server remains operational before crashing. X can take any non-negative real value. Assume the probability density function (pdf) is given by:

$$f_X(x) = \begin{cases} \frac{1}{100}e^{-x/100} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

This probability distribution is called an **exponential distribution with a mean of 100 hours**. We will define and talk about mean later. The CDF is computed as follows:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt = \begin{cases} 0 & \text{if } x < 0 \\ 1 - e^{-x/100} & \text{if } x \geq 0 \end{cases}$$

The CDF $F_Y(y) = P(Y \leq y)$ would give the probability that the server operates for at most y hours. For instance, $F_Y(10)$ would be the probability the server fails within the first 10 hours.

We can plot the CDF and the PDF in Python as follows:

```
import numpy as np
import matplotlib.pyplot as plt
x = np.linspace(-50, 400, 1000)
pdf = np.piecewise(x, [x < 0, x >= 0], [0, lambda x: (1/100) * np.exp(-x/100)])
cdf = np.piecewise(x, [x < 0, x >= 0], [0, lambda x: 1 - np.exp(-x/100)])
plt.subplots(2,1, sharex=True)
plt.subplot(2, 1, 1)
plt.plot(x, pdf, label='PDF', color='blue')
plt.title('PDF')
plt.ylabel('f(x)')
plt.grid()
plt.subplot(2, 1, 2)
plt.plot(x, cdf, label='CDF', color='orange')
plt.title('CDF')
plt.xlabel('x')
plt.ylabel('F(x)')
plt.grid()
plt.show()
```

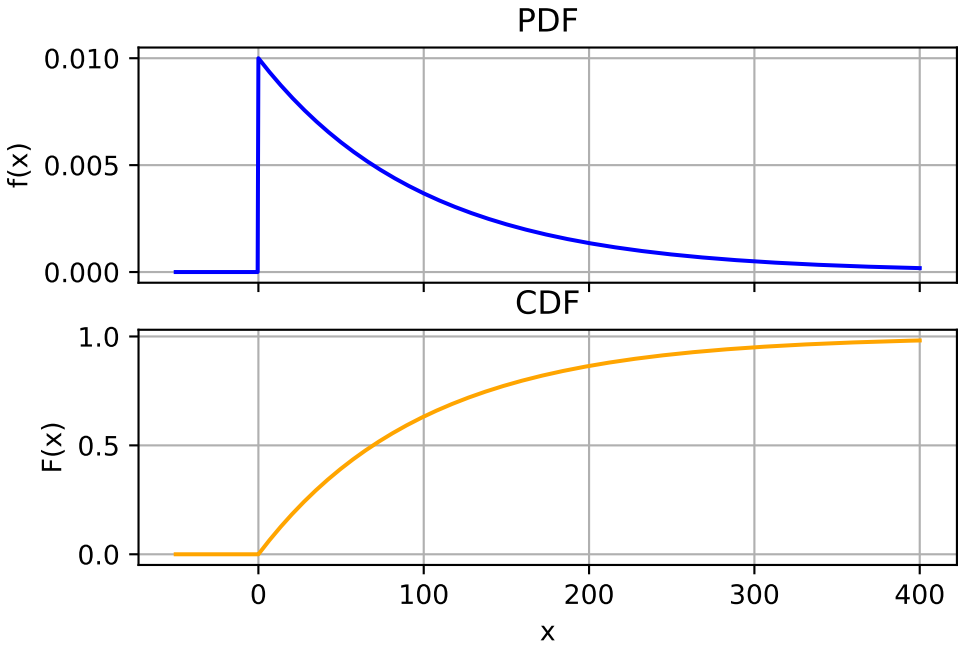


Figure 1.2: PDF and CDF of of an Exponential Random Variable

Support of a Random Variable

Definition 1.12. The support of a random variable is the set of values where its probability distribution is non-zero.

- For a discrete random variable, the support is the set of values x such that $f_X(x) > 0$.
- For a continuous random variable, the support is the set of values x where the PDF $f_X(x) > 0$.

In terms of the sample space Ω, the support is equal to $X(\Omega)$

1.6 Common Probability Distributions

1.6.1 Discrete distributions

Some common discrete probability distributions include:

Discrete Uniform Random Variable

Definition 1.13. Let X be a random variable that can take on any of k equally likely values. The PMF is given by:

$$f_X(x|k) = \begin{cases} \frac{1}{k} & \text{if } x \in \{1, 2, \ldots, k\} \\ 0 & \text{otherwise} \end{cases}$$

We only need to specify k to specify the distribution, hence the notations $f_X(x|k)$ for the PMF.
Note the choice of support is arbitrary. We could have chosen any k distinct values for the support, we choose the first k integers for simplicity/convenience.
Examples of random variables modelled this way is the outcome of rolling a fair $k = 6$-sided die or the outcome of randomly selecting one item from a set of k distinct items. The notion of *classical probability* in Section 1.2.1 is equivalent to assuming a discrete uniform distribution to the random variable that assigns a real number to each ouctome.

General discrete random variable

Definition 1.14. Let X be a discrete random variable with possible values in the set $\{x_1, x_2, \ldots, x_k\}$. The PMF is given by:

$$f_X(x|p_1, \ldots, p_{k-1}) = P(X = x) = \begin{cases} p_i & \text{if } x = x_i \\ 0 & \text{otherwise} \end{cases}$$

where $p_i > 0$ for all i and $\sum_i p_i = 1$.
We only need to specify $k - 1$ probabilities as the last one is determined by the fact that the probabilities must sum to 1. Hence the notation $f_X(x|p_1, \ldots, p_{k-1})$ for the PMF.

Bernoulli and Binomial Random Variables

Definition 1.15. Bernoulli Distribution: Models a single binary outcome (success/failure) with parameter p (probability of success). The PMF is given by:

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

We only need to specify p to specify the distribution. We could have chosen any two distinct values instead of 0/1, we choose 0/1 for simplicity/convenience.
Binomial Distribution: If n identical Bernoulli trials are performed, define the events:

- A_i: the i-th trial is a success (for $i = 1, 2, \ldots, n$) with $P(A_i) = p$.

If we **assume the events** $A_1, \ldots, A_n$ **are mutually independent** (as in Definition 1.8), then the random variable Y defined as the number of successes in the n trials follows a Binomial distribution.
The event $\{Y = y\}$ will occur only if, out of the events $A_1, \ldots, A_n$, exactly y of them occur, and necessarily $n - y$ of them do not occur. For example, when $y = 2$, one particular outcome (one particular ordering of occurrences and nonoccurrences) of the n Bernoulli trials might be:

$$A_1, A_2, A_3^c, A_4^c, \ldots, A_n^c$$

which has probability

$$p^2(1-p)^{n-2}$$

However, there are many such orderings that lead to the same event $\{Y = 2\}$, for example:

$$A_2, A_5, A_1^c, A_3^c, A_4^c, \ldots, A_n^c$$

which also has probability $p^2(1-p)^{n-2}$. The number of such orderings is the number of ways of choosing 2 successes from n trials, which is given by the binomial coefficient $\binom{n}{2}$. In general, the number of ways of choosing y successes from n trials is given by the binomial coefficient $\binom{n}{y}$. Therefore, the PMF of the Binomial distribution is given by:

$$f_Y(y|n,p) = P(Y = y) = \binom{n}{y} p^y (1-p)^{n-y} \quad \text{for } y = 0, 1, \ldots, n$$

We only need to specify n and p to specify the distribution, hence the notation $f_Y(y|n,p)$ for the PMF.

Hypergeometric Random Variable

Definition 1.16. The hypergeometric distribution models the number of successes in a sequence of n draws from a finite population **without replacement**. So is similar to the Binomial distribution but without the assumption of independence of the Bernoulli trials.
It is easy to describe this distribution with a concrete example. Suppose we have an urn with:

- a total of n balls
- m balls are red and
- $n - m$ are green.

We select k balls at random (the k balls are taken all at once, a case of sampling without replacement) so that $k \leq n$. What is the probability that exactly y of the balls are red? The corresponding random variable Y is the number of red balls in the sample of size k. The support of the random variable Y can be obtained using the following reasoning:

- To obtain the minimum number of red balls, there are two cases:
  - if there more green balls than those we can choose($k \leq n - m$) and if we happen to choose all green balls then the number of red balls is 0, e.g. $Y = 0$ and this is the smallest it can be.
  - if $k > n - m$ and we choose all green balls then all the remaining $k - (n - m)$ balls are necessarily red so $Y = k - (n - m) > 0$ and this is the smallest it can be.

Therefore, the minimum number of red balls is $\max\{0, k - (n - m)\}$.

- For the maximum number of red balls, there are two cases:
  - we cannot choose more red balls than there are in the urn (e.g. $m \leq k$) so that the maximum number of red balls is m.
  - we cannot choose more red balls than the total number k of balls we are choosing (e.g. $k > m$) so that the maximum number of red balls is k

Therefore, the maximum number of red balls is $\min\{m, k\}$.
We also have that:

- the number of ways of choosing k balls from n is $\binom{n}{k}$
- the number of ways of choosing y red balls from the m red balls is $\binom{m}{y}$ and
- the number of ways of choosing the remaining $k - y$ balls from the $n - m$ green balls is $\binom{n-m}{k-y}$. Therefore, the PMF of the Hypergeometric distribution is given by:

$$f_Y(y|n, m, k) = P(Y = y) = \begin{cases} \dfrac{\binom{m}{y}\binom{n-m}{k-y}}{\binom{n}{k}} & \text{for } y = \max\{0, k - (n - m)\}, \ldots, \min\{m, k\} \\ \\ \text{otherwise} \end{cases}$$

The equally likely implicit assumption can be justified if we can guarantee the balls are randomly choosen.
We only need to specify n, m and k to specify the distribution, hence the notation $f_Y(y|n, m, k)$ for the PMF.

Poisson Random Variable

Definition 1.17. This random variable is relevant when we are modeling the ocurrences of an event in time or space. For example

- waiting for a bus to arrive,
- waiting for customers to arrive in a bank,
- number of damaged trees in a given area of a forest
- the number of defects in a given length of communications cable

The number of occurrences in a given interval or area can sometimes be modeled by the Poisson distribution.
The Poisson distribution is based on the following assumptions:

- for small time intervals, the probability of an event is proportional to the length of waiting time or area size
- The number of events in disjoint time intervals or disjoint areas are independent.
- the intensity λ (average rate of occurrence) is constant over time or space.

So let define random variable X that models the number of events occurring in a fixed interval of time or area of space, given an intensity parameter λ (which also relative to the length of time or area size). The PMF is given by:

$$f_X(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \quad \text{for } x = 0, 1, 2, \dots$$

We only need to specify the rate λ to specify the distribution, hence the notation $f_X(x|\lambda)$ for the PMF.

Negative Binomial Random Variable

Definition 1.18. The negative binomial distribution models the number of trials needed to achieve a fixed number of successes in a sequence of independent Bernoulli trials, each with the same probability of success p.
Let X be the random variable representing the number of trials needed to achieve r successes. The PMF is given by:

$$f_X(x|r,p) = \binom{x-1}{r-1} p^r (1-p)^{x-r} \quad \text{for } x = r, r+1, r+2, \dots$$

This is because the r-th success must occur on the x-th trial, and the previous $x-1$ trials

must contain exactly $r-1$ successes. The probability of any such sequence of trials is $p^r(1-p)^{x-r}$. The number of ways to choose which $r-1$ trials out of the first $x-1$ are successes is given by the binomial coefficient $\binom{x-1}{r-1}$.
We only need to specify r and p to specify the distribution, hence the notation $f_X(x|r,p)$ for the PMF.
The specific case where $r=1$ is called the **geometric distribution**, which models the number of trials until the first success. The PMF for the geometric distribution is given by:

$$f_X(x|p) = (1-p)^{x-1}p \quad \text{for } x = 1, 2, 3, \ldots$$

There are alternative definitions of the Negative Binomial distribution. For example, the random variable Y is defined as the number of failures before the r-th success. The PMF in this case is given by:

$$f_Y(y|r,p) = \binom{y+r-1}{r-1}p^r(1-p)^y = \binom{y+r-1}{y}p^r(1-p)^y \quad \text{for } y = 0, 1, 2, \ldots$$

Clearly, the two random variables are related in that $Y = X - r$.

1.6.2 Continuous distributions

Some common continuous probability distributions include:

Continuous Uniform

Definition 1.19. Models a continuous random variable such that intervals of the same length are equally likely. The support is the interval (a,b) for $a < b$. The probability density fucntion (PDF) is given by:

$$f_X(x|a,b) = \begin{cases} \frac{1}{b-a} & \text{if } a < x < b \\ 0 & \text{otherwise} \end{cases}$$

We only need to specify a and b to specify the distribution. The CDF for this distribution is given by:

$$F_X(x|a,b) = \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } a \leq x \leq b \\ 1 & \text{if } x > b \end{cases}$$

Normal (Gaussian) Random Variable

Definition 1.20. Models a continuous random variable with a bell-shaped curve, characterized by its mean μ and standard deviation σ. The PDF is given by:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{for } x \in \mathbb{R}$$

The CDF for this distribution is given by:

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\, du = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$$

where erf is the error function. Usually, we will express this CDF in terms of the CDF of the standard normal distribution (mean 0 and standard deviation 1) that we will denote by $\Phi(x)$. Then we can write:

$$F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

where $\Phi(x)$ is the CDF of the standard normal distribution, that is

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}\, du$$

Exponential Random Variable

Definition 1.21. THis random variable can be used to model the time (continuouslye.g. infinite precision) to the ocurrence of an event of interest or the time in between events of interest. It is fully characterised by the rate parameter λ. The PDF is given by:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}$$

The CDF for this distribution is given by:

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - e^{-\lambda x} & \text{if } x \geq 0 \end{cases}$$

1.6.3 Joint Distributions

We can define more than one random variable on the same sample space.

Random Vectors

Definition 1.22. A random vector is a vector whose components are random variables defined on the same probability space. If we have k random variables $X_1, X_2, \ldots, X_k$ defined on the same sample space Ω, we can define a random vector $\mathbf{X}$ as the function from Ω to $\mathbb{R}^k$ given by:

$$\mathbf{X}(\omega) := (X_1(\omega), X_2(\omega), \ldots, X_n(\omega))$$

Joint Distribution of two random variables

Definition 1.23. The joint distribution of two random variables X and Y describes the probability distribution of their combined outcomes. It can be easily represented by the joint cumulative distribution function (CDF) defined as:

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y)$$

for any real numbers x and y. This definition is irrespective of whether the random variables are discrete or continuous.
The joint distribution can also represented by the joint probability mass function (PMF) for discrete random variables or the joint probability density function (PDF) for continuous random variables.
For two discrete random variables X and Y, the joint PMF is defined as:

$$f_{X,Y}(x, y) = P(X = x, Y = y)$$

for all possible joint values (x, y) in the image set $(X, Y)(\Omega)$.
For two continuous random variables X and Y, the joint PDF can be defined if there exists a function $f_{X,Y}(x, y)$ such that for any two numbers any subset $A \subset \mathbb{R}^2$:

$$P((X, Y) \in A) = \int_A \int f_{X,Y}(x, y)\, dx\, dy$$

The joint PDF can also be defined as the partial cross-derivative of their joint cumulative distribution function (CDF):

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} P(X \leq x, Y \leq y)$$

The joint PDF satisfies:

- $f_{X,Y}(x, y) \geq 0$ for all $x, y \in \mathbb{R}$.
- $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx\, dy = 1$.

Marginal Distributions

Definition 1.24. The marginal distribution of a random variable is the probability distribution of that variable when considered independently of other variables. It is obtained by summing (for discrete variables) or integrating (for continuous variables) the joint distribution over the values of the other variables. For two discrete random variables X and Y with joint PMF $f_{X,Y}(x, y)$, the marginal PMFs are given by:

$$f_X(x) = \sum_y f_{X,Y}(x, y)$$

$$f_Y(y) = \sum_x f_{X,Y}(x, y)$$

For two continuous random variables X and Y with joint PDF $f_{X,Y}(x, y)$, the marginal PDFs are given by:

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

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

We can define marginal CDFs in a similar manner. For two random variables X and Y with joint CDF $F_{X,Y}(x, y)$, the marginal CDFs are given by:

$$F_X(x) = \lim_{y\to\infty} F_{X,Y}(x, y)$$

$$F_Y(y) = \lim_{x\to\infty} F_{X,Y}(x, y)$$

Joint Distribution in the case of two coin flips

Example 1.23. When flipping a coin twice, we can define two random variables:

- X_1: outcome of the first flip (1 for heads, 0 for tails)
- X_2: outcome of the second flip (1 for heads, 0 for tails)

Following from Example 1.11 The joint distribution of X_1 and X_2 can be represented in a table:

$X_1 \backslash X_2$	0 (Tails)	1 (Heads)
0 (Tails)	0.1	0.4
1 (Heads)	0.2	0.3

The joint PMF is given by:

- $f_{X_1,X_2}(0,0) = P(X_1 = 0, X_2 = 0) = P(\{TT\}) = 0.1$
- $f_{X_1,X_2}(0,1) = P(X_1 = 0, X_2 = 1) = P(\{TH\}) = 0.4$
- $f_{X_1,X_2}(1,0) = P(X_1 = 1, X_2 = 0) = P(\{HT\}) = 0.2$
- $f_{X_1,X_2}(1,1) = P(X_1 = 1, X_2 = 1) = P(\{HH\}) = 0.3$

The marginal distributions of X_1 and X_2 can be obtained by summing over the rows and columns respectively:

$$f_{X_1}(x_1) = \sum_{x_2 \in \{0,1\}} f_{X_1,X_2}(x_1, x_2) = \begin{cases} f_{X_1,X_2}(0,0) + f_{X_1,X_2}(0,1) = 0.1 + 0.4 = 0.5 & \text{if } x_1 = 0 \\ f_{X_1,X_2}(1,0) + f_{X_1,X_2}(1,1) = 0.2 + 0.3 = 0.5 & \text{if } x_1 = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$f_{X_2}(x_2) = \sum_{x_1 \in \{0,1\}} f_{X_1,X_2}(x_1, x_2) = \begin{cases} f_{X_1,X_2}(0,0) + f_{X_1,X_2}(1,0) = 0.1 + 0.2 = 0.3 & \text{if } x_2 = 0 \\ f_{X_1,X_2}(0,1) + f_{X_1,X_2}(1,1) = 0.4 + 0.3 = 0.7 & \text{if } x_2 = 1 \\ 0 & \text{otherwise} \end{cases}$$

Now consider another random variable Z be the indicator that the two flips are the same, that is $Z = 1$ if $X_1 = X_2$ and $Z = 0$ otherwise. The joint distribution of X_1 and Z can be represented in a table:

$X_1 \backslash Z$	0 (Different)	1 (Same)
0 (Tails)	0.4	0.1
1 (Heads)	0.2	0.3

The joint PMF is given by:

- $f_{X_1,Z}(0,0) = P(X_1 = 0, Z = 0) = P(\{TH\}) = 0.4$
- $f_{X_1,Z}(0,1) = P(X_1 = 0, Z = 1) = P(\{TT\}) = 0.1$
- $f_{X_1,Z}(1,0) = P(X_1 = 1, Z = 0) = P(\{HT\}) = 0.2$
- $f_{X_1,Z}(1,1) = P(X_1 = 1, Z = 1) = P(\{HH\}) = 0.3$

The marginal distributions of X_1 and Z can be obtained by summing over the rows and columns respectively:

$$f_{X_1}(x_1) = \sum_{z \in \{0,1\}} f_{X_1,Z}(x_1, z) = \begin{cases} f_{X_1,Z}(0,0) + f_{X_1,Z}(0,1) = 0.4 + 0.1 = 0.5 & \text{if } x_1 = 0 \\ f_{X_1,Z}(1,0) + f_{X_1,Z}(1,1) = 0.2 + 0.3 = 0.5 & \text{if } x_1 = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$f_Z(z) = \sum_{x_1 \in \{0,1\}} f_{X_1,Z}(x_1, z) = \begin{cases} f_{X_1,Z}(0,0) + f_{X_1,Z}(1,0) = 0.4 + 0.2 = 0.6 & \text{if } z = 0 \\ f_{X_1,Z}(0,1) + f_{X_1,Z}(1,1) = 0.1 + 0.3 = 0.4 & \text{if } z = 1 \\ 0 & \text{otherwise} \end{cases}$$

Examples of joint continuous Distribution

Example 1.24. Consider two continuous random variables X and Y with the joint PDF given by:

$$f_{X,Y}(x,y) = \begin{cases} 2 & \text{if } 0 < x < 1 \text{ and } 0 < y < x \\ 0 & \text{otherwise} \end{cases}$$

We can verify that this is a valid joint PDF by checking that it is non-negative and integrates to 1 over the entire plane:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dy\, dx = \int_0^1 \int_0^x 2\, dy\, dx = \int_0^1 2x\, dx = 1$$

The support of the joint distribution is the triangular region in the xy-plane where $0 < x < 1$ and $0 < y < x$.
The marginal distributions can be computed as follows:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dy = \int_0^x 2\, dy = 2x \quad \text{for } 0 < x < 1$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dx = \int_y^1 2\, dx = 2(1-y) \quad \text{for } 0 < y < 1$$

We can also obtain the joint cumulative distribution function (CDF) as follows. For any $0 < y < x < 1$

$$\begin{aligned} F_{X,Y}(x,y) &= P(X \le x, Y \le y) \\ &= \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(u,v)\,dv\,du \\ &= \int_{0}^{y}\int_{0}^{u} f_{X,Y}(u,v)\,dv\,du + \int_{y}^{x}\int_{0}^{y} f_{X,Y}(u,v)\,dv\,du \\ &= \int_{0}^{y}\int_{0}^{u} 2\,dv\,du + \int_{y}^{x}\int_{0}^{y} 2\,dv\,du \\ &= \int_{0}^{y} 2u\,du + \int_{y}^{x} 2y\,du \\ &= y^2 + 2y(x-y) \end{aligned}$$

The marginals CDFs can be computed as follows:

$$F_X(x) = \lim_{y\to\infty} F_{X,Y}(x,y) = F_{X,Y}(x,y) = x^2 + 2x(x-x) = x^2 \quad \text{for } 0 < x < 1$$

$$F_Y(y) = \lim_{x\to\infty} F_{X,Y}(x,y) = F_{X,Y}(1,y) = y^2 + 2y(1-y) = 2y - y^2 \quad \text{for } 0 < y < 1$$

or, alternatively, we can compute the marginal CDFs directly from the marginal PDFs as follows:

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\,du = \int_{0}^{x} 2u\,du = x^2 \quad \text{for } 0 < x < 1$$

$$F_Y(y) = \int_{-\infty}^{y} f_Y(v)\,dv = \int_{0}^{y} 2(1-v)\,dv = 2y - y^2 \quad \text{for } 0 < y < 1$$

Now consider another joint PDF given by:

$$f_{X,Y}(x,y) = \begin{cases} 6xy^2 & \text{if } 0 < x < 1 \text{ and } 0 < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

We can verify that this is a valid joint PDF by checking that it is non-negative and integrates to 1 over the entire plane:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy\,dx = \int_{0}^{1}\int_{0}^{1} 6xy^2\,dy\,dx = \int_{0}^{1} 2x\,dx = 1$$

The support of the joint distribution is the unit square in the xy-plane where $0 < x < 1$ and $0 < y < 1$.
The marginal distributions can be computed as follows:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy = \int_0^1 6xy^2\,dy = 2x \quad \text{for } 0 < x < 1$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx = \int_0^1 6xy^2\,dx = 3y^2 \quad \text{for } 0 < y < 1$$

Then we have

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

so the random variables X and Y are independent. THis also implies that the joint cumulative distribution function (CDF) is given by:

$$F_{X,Y}(x,y) = F_X(x)F_Y(y)$$

where $F_X(x)$ and $F_Y(y)$ are the marginal CDFs of X and Y, respectively. We can compute these marginal CDFs as follows:

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\,du = \int_0^x 2u\,du = x^2 \quad \text{for } 0 < x < 1$$

$$F_Y(y) = \int_{-\infty}^{y} f_Y(v)\,dv = \int_0^y 3v^2\,dv = y^3 \quad \text{for } 0 < y < 1$$

Therefore, the joint CDF is given by:

$$F_{X,Y}(x,y) = F_X(x)F_Y(y) = x^2y^3 \quad \text{for } 0 < x < 1 \text{ and } 0 < y < 1$$

Independent Random Variables

Definition 1.25. Two discrete random variables X and Y are independent if for all x and y in their respective image sets:

$$P(X = x, Y = y) = P(X = x) \times P(Y = y)$$

for all x and y. Equivalently, if either of the following conditions hold for all x and y:

- $P(X = x|Y = y) = P(X = x)$
- $P(Y = y|X = x) = P(Y = y)$

for all x and y.
For the case of continuous random variables, X and Y are independent if for all x and y in their respective image sets:

$$f_{X,Y}(x,y) = f_X(x) \times f_Y(y)$$

where $f_{X,Y}(x,y)$ is the joint probability density function of X and Y, and $f_X(x)$ and $f_Y(y)$ are the marginal probability density functions of X and Y, respectively.

Example of Independent Random Variables

Example 1.25. We can check if X_1 and X_2 define d in Example 1.23 are independent random variables. We have:

$$P(X_1 = 0, X_2 = 0) = 0.1 \neq P(X_1 = 0)P(X_2 = 0) = 0.5 \times 0.4 = 0.2$$

so they are not independent.
We can check if X_1 and Z defined in Example 1.23 are independent random variables. We have:

$$P(X_1 = 0, Z = 0) = 0.4 = P(X_1 = 0)P(Z = 0) = 0.5 \times 0.6 = 0.3$$

so they are not independent.

1.7 Conditional distributions

Conditional Distribution of two random variables

Definition 1.26. The conditional distribution of a random variable X given another random variable Y describes the probability distribution of X when the value of Y is known. It is represented by the conditional probability mass function (PMF) for discrete random variables or the conditional probability density function (PDF) for continuous random variables. For two discrete random variables X and Y, the conditional PMF of X given $Y = y$ is defined as:

$$f_{X|Y}(x|y) = P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

for all possible values x in the image set of X and for all y such that $P(Y = y) > 0$. For two continuous random variables X and Y, the conditional PDF of X given $Y = y$ is defined as:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

for all possible values x in the image set of X and for all y such that $f_Y(y) > 0$.

Example of Discrete Conditional Distribution

Example 1.26. We can compute the conditional distribution of X_1 given Z in Exam-

ple 1.23. We have:

$$f_{X_1|Z}(x|z) = \frac{f_{X_1,Z}(x,z)}{f_Z(z)}$$

for all possible values x in the image set of X_1 and for all z such that $f_Z(z) > 0$. We have:

- $f_{X_1|Z}(0|0) = \frac{f_{X_1,Z}(0,0)}{f_Z(0)} = \frac{0.4}{0.6} = \frac{2}{3}$
- $f_{X_1|Z}(1|0) = \frac{f_{X_1,Z}(1,0)}{f_Z(0)} = \frac{0.2}{0.6} = \frac{1}{3}$

the other conditional PMF is:

- $f_{X_1|Z}(0|1) = \frac{f_{X_1,Z}(0,1)}{f_Z(1)} = \frac{0.1}{0.4} = \frac{1}{4}$
- $f_{X_1|Z}(1|1) = \frac{f_{X_1,Z}(1,1)}{f_Z(1)} = \frac{0.3}{0.4} = \frac{3}{4}$

Example of Continuous Conditional Distribution

Example 1.27. Consider the joint PDF given in Example 1.24:

$$f_{X,Y}(x,y) = \begin{cases} 2 & \text{if } 0 < x < 1 \text{ and } 0 < y < x \\ 0 & \text{otherwise} \end{cases}$$

We can compute the conditional distribution of Y given $X = x$. We have:

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

for all possible values y in the image set of Y and for all x such that $f_X(x) > 0$. We have:

$$f_{Y|X}(y|x) = \frac{2}{2x} = \frac{1}{x} \quad \text{for } 0 < y < x$$

and 0 otherwise.
The other conditional distribution is:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

for all possible values x in the image set of X and for all y such that $f_Y(y) > 0$. We have:

$$f_{X|Y}(x|y) = \frac{2}{2(1-y)} = \frac{1}{1-y} \quad \text{for } y < x < 1$$

and 0 otherwise.

1.8 Moments, Variance, covariance and correlation

EXpectation and Variance of a Random Variable

Definition 1.27. The **expectation**, **expected value** or **mean** of a random variable X, denoted by $E[X]$ or μ_X, is defined as:

- For a discrete random variable:

$$E[X] = \sum_{x \in X(\Omega)} x \cdot P(X = x) = \sum_{x \in X(\Omega)} x \cdot f_X(x)$$

- For a continuous random variable:

$$E[X] = \int_{-\infty}^{\infty} x \cdot f_X(x)\, dx$$

The **variance** of a random variable X, denoted by $Var(X)$ is defined as:

$$Var(X) = E[(X - E[X])^2]$$

It is easy to show that:

$$Var(X) = E[X^2] - (E[X])^2$$

The standard deviation of X, is the square root of the variance: $\sigma_X = \sqrt{Var(X)}$

The variance measures the spread or dispersion of the random variable around its mean. A higher variance indicates that the values of the random variable are more spread out, while a lower variance indicates that they are more concentrated around the mean.

Examples of Expectation and Variance

Example 1.28.

- For a discrete uniform random variable with parameter k:
  - $E[X] = \frac{k+1}{2}$
  - $Var(X) = \frac{k^2-1}{12}$
- For a bernoulli random variable with parameter p:
  - $E[X] = p$
  - $Var(X) = p(1-p)$
- For a binomial random variable with parameters n and p:

  - $E[X] = np$
  - $Var(X) = np(1-p)$
- For a hypergeometric random variable with parameters n, m and k:
  - $E[X] = k\frac{m}{n}$
  - $Var(X) = k\frac{m}{n}\frac{n-m}{n}\frac{n-k}{n-1}$
- For a Poisson random variable with parameter λ:
  - $E[X] = \lambda$
  - $Var(X) = \lambda$
- For the negative binomial random variable with parameters r and p:
  - $E[X] = \frac{r}{p}$
  - $Var(X) = \frac{r(1-p)}{p^2}$
  - $E[Y] = \frac{r(1-p)}{p}$
  - $Var(X) = \frac{r(1-p)}{p^2}$
- For a uniform random variable on the interval $[a, b]$:
  - $E[X] = \frac{a+b}{2}$
  - $Var(X) = \frac{(b-a)^2}{12}$
- For a normal random variable with mean μ and standard deviation σ:
  - $E[X] = \mu$
  - $Var(X) = \sigma^2$
- For an exponential random variable with rate parameter λ:
  - $E[X] = \frac{1}{\lambda}$
  - $Var(X) = \frac{1}{\lambda^2}$

These can be derived from the definitions above and the corresponding probability mass or density functions.

Expected values of functions of random variables

Definition 1.28. The expected value of a function $g(X)$ of a random variable X is given by:

- For a discrete random variable:

$$E[g(X)] = \sum_{x \in X(\Omega)} g(x) \cdot P(X = x) = \sum_{x \in X(\Omega)} g(x) \cdot f_X(x)$$

- For a continuous random variable:

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot f_X(x)\, dx$$

For joint random variables X and Y, the expected value of a function $g : \mathbb{R}^2 \to \mathbb{R}$ is given by:

- For discrete random variables:

$$\begin{aligned} E[g(X,Y)] &= \sum_{(x,y) \in (X,Y)(\Omega)} \sum g(x,y) \cdot P(X = x, Y = y) \\ &= \sum_{(x,y) \in (X,Y)(\Omega)} \sum g(x,y) \cdot f_{X,Y}(x,y) \end{aligned}$$

- For continuous random variables:

$$E[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) \cdot f_{X,Y}(x,y)\, dx\, dy$$

HIgher order moments and moment generating function

Definition 1.29. The n-th moment of a random variable X is defined as:

$$E[X^n] = \begin{cases} \sum_{x \in X(\Omega)} x^n \cdot P(X = x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x^n \cdot f_X(x)\, dx & \text{if } X \text{ is continuous} \end{cases}$$

The moment generating function (MGF) of a random variable X is defined as:

$$M_X(t) = E[e^{tX}] = \begin{cases} \sum_{x \in X(\Omega)} e^{tx} \cdot P(X = x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} e^{tx} \cdot f_X(x)\, dx & \text{if } X \text{ is continuous} \end{cases}$$

The MGF can be used to compute the moments of a random variable. The n -th moment of X can be obtained by taking the n-th derivative of the MGF and evaluating it at $t = 0$:

$$E[X^n] = M_X^{(n)}(0) = \frac{d^n}{dt^n} M_X(t) \Big|_{t=0}$$

Notes:

- We are assuming here that the MGF exists in a neighborhood of $t = 0$.
- The derivative formula above does not depend on whether X is discrete or continuous.
- The first moment is the mean: $E[X] = M'_X(0)$
- The second moment is $E[X^2] = M''_X(0)$
- The variance can be computed as: $Var(X) = M''_X(0) - (M'_X(0))^2$
- The MGF uniquely determines the distribution of a random variable, if it exists in a neighborhood of $t = 0$.

Bernoulli MGF and Moments

Example 1.29. For a Bernoulli random variable with parameter p. The MGF is given by

$$M_X(t) = 1 - p + pe^t$$

Differentiating and evaluating at $t = 0$, we have:

- $M'_X(t) = pe^t$ so $E[X] = M'_X(0) = p$
- $M''_X(t) = pe^t$ so $E[X^2] = M''_X(0) = p$
- $Var(X) = M''_X(0) - (M'_X(0))^2 = p - p^2 = p(1 - p)$

Poisson MGF and Moments

Example 1.30. For a Poisson random variable with parameter λ. The MGF is given by

$$M_X(t) = e^{\lambda(e^t-1)}$$

Differentiating and evaluating at $t = 0$, we have: * $M'_X(t) = \lambda e^t e^{\lambda(e^t-1)}$ so $E[X] = M'_X(0) = \lambda$ * $M''_X(t) = \lambda e^t e^{\lambda(e^t-1)} + \lambda^2 e^{2t} e^{\lambda(e^t-1)}$ so $E[X^2] = M''_X(0) = \lambda + \lambda^2$ * $Var(X) = M''_X(0) - (M'_X(0))^2 = \lambda + \lambda^2 - \lambda^2 = \lambda$

MFG and Moments example

Example 1.31. A random variable X has the following MFG:

$$M_X(t) = \frac{3}{3-t}, \qquad t < 3$$

We can obtain the moments without knowledge of the PDF or CDF as follows:

- $M'_X(t) = \frac{3}{(3-t)^2}$ so $E[X] = M'_X(0) = \frac{1}{3}$
- $M''_X(t) = \frac{6}{(3-t)^3}$ so $E[X^2] = M''_X(0) = \frac{6}{27}$
- $Var(X) = M''_X(0) - (M'_X(0))^2 = \frac{6}{27} - \frac{1}{9} = \frac{6}{27} - \frac{3}{27} = \frac{1}{9}$

Now, let X be an exponential random variable with rate parameter λ. The MGF is given by

$$M_X(t) = \int_0^\infty e^{tx} \lambda e^{-\lambda x}\, dx = \frac{\lambda}{\lambda - t}$$

for $t < \lambda$. So the only distribution with the MGF given above is an exponential distribution with rate parameter $\lambda = 3$.

Expectation and Independence

Proposition 1.5. *If X and Y are independent random variables, then:*

- $E[XY] = E[X]E[Y]$
- *More generally, if g and h are functions, then $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$*
- *In particular,*

$$M_{X+Y}(t) = E[e^{t(X+Y)}] = E[e^{tX}e^{tY}] = E[e^{tX}]E[e^{tY}] = M_X(t)M_Y(t)$$

this means that the MGF of the sum of independent random variables is the product of their MGFs.

Sum of Bernoulli Random Variables

Example 1.32. Let $X_1, X_2, \ldots, X_n$ be independent Bernoulli random variables with parameter p. Let $S_n = X_1 + X_2 + \cdots + X_n$ be their sum. We can compute the MGF of S_n as follows:

$$M_{S_n}(t) = E[e^{tS_n}] = E[e^{t(X_1+X_2+\cdots+X_n)}] = E[e^{tX_1}e^{tX_2}\cdots e^{tX_n}] = E[e^{tX_1}]E[e^{tX_2}]\cdots E[e^{tX_n}] = (M_{X_1}(t))^n$$

where we used the independence of the X_i's. Since each X_i is a Bernoulli random variable with parameter p, we have:

$$M_{X_i}(t) = 1 - p + pe^t$$

for all i. Therefore, we have:

$$M_{S_n}(t) = (1 - p + pe^t)^n$$

which is the MGF of a Binomial random variable with parameters n and p. Therefore, we conclude that S_n follows a Binomial distribution with parameters n and p.

Sum of Poisson Random Variables

Example 1.33. Let $X_1, X_2, \ldots, X_n$ be independent Poisson random variables with parameters $\lambda_1, \lambda_2, \ldots, \lambda_n$, respectively. Let $S_n = X_1 + X_2 + \cdots + X_n$ be their sum. We can compute the MGF of S_n as follows:

$$M_{S_n}(t) = E[e^{tS_n}] = E[e^{t(X_1+X_2+\cdots+X_n)}] = E[e^{tX_1}e^{tX_2}\cdots e^{tX_n}] = E[e^{tX_1}]E[e^{tX_2}]\cdots E[e^{tX_n}] = \prod_{i=1}^{n} M_{X_i}(t)$$

where we used the independence of the X_i's. Since each X_i is a Poisson random variable with parameter λ_i, we have:

$$M_{X_i}(t) = e^{\lambda_i(e^t-1)}$$

for all i. Therefore, we have:

$$M_{S_n}(t) = \prod_{i=1}^{n} e^{\lambda_i(e^t-1)} = e^{(\sum_{i=1}^{n}\lambda_i)(e^t-1)}$$

which is the MGF of a Poisson random variable with parameter $\sum_{i=1}^{n}\lambda_i$. Therefore, we conclude that S_n follows a Poisson distribution with parameter $\sum_{i=1}^{n}\lambda_i$.

Sum of Exponential Random Variables

Example 1.34. Let $X_1, X_2, \ldots, X_n$ be independent exponential random variables with rate parameter λ. Let $S_n = X_1 + X_2 + \cdots + X_n$ be their sum. We can compute the MGF of S_n as follows:

$$M_{S_n}(t) = E[e^{tS_n}] = E[e^{t(X_1+X_2+\cdots+X_n)}] = E[e^{tX_1}e^{tX_2}\cdots e^{tX_n}] = E[e^{tX_1}]E[e^{tX_2}]\cdots E[e^{tX_n}] = (M_{X_1}(t))^n$$

where we used the independence of the X_i's. Since each X_i is an exponential random variable with rate parameter λ, we have:

$$M_{X_i}(t) = \frac{\lambda}{\lambda - t}$$

for all i and for $t < \lambda$. Therefore, we have:

$$M_{S_n}(t) = \left(\frac{\lambda}{\lambda - t}\right)^n$$

which is the MGF of a Gamma random variable with shape parameter n and rate parameter λ. Therefore, we conclude that S_n follows a Gamma distribution with shape parameter n and rate parameter λ.
We have not defined the Gamma distribution yet. WE do this here for completeness. The Gamma distribution with shape parameter k and rate parameter θ has the following

PDF:

$$f_X(x) = \begin{cases} \frac{\theta^k}{\Gamma(k)} x^{k-1} e^{-\theta x} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

§ where $\Gamma(k)$ is the Gamma function defined as:

$$\Gamma(k) = \int_0^\infty x^{k-1} e^{-x}\, dx$$

The mean and variance of a Gamma random variable are given by: * $E[X] = \frac{k}{\theta}$ * $Var(X) = \frac{k}{\theta^2}$

Covariance and Correlation

Definition 1.30. The **covariance** between two random variables X and Y, denoted by $Cov(X, Y)$ is defined as:

$$Cov(X, Y) = E[(X - E[X])(Y - E[Y])]$$

It is easy to show that:

$$Cov(X, Y) = E[XY] - E[X]E[Y]$$

The covariance measures the linear relationship between two random variables. A positive covariance indicates that the variables tend to increase or decrease together, while a negative covariance indicates that one variable tends to increase when the other decreases. The **correlation coefficient** between two random variables X and Y, denoted by $\rho_{X,Y}$ or $Corr(X, Y)$, is defined as:

$$Corr(X, Y) = \rho_{X,Y} := \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

where σ_X and σ_Y are the standard deviations of X and Y, respectively.
The correlation coefficient measures the strength and direction of the linear relationship between two random variables. It ranges from -1 to 1, where 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship.

Example of Covariance and Correlation

Example 1.35. Using the joint distribution of X_1 and Z in Example 1.23, we can compute the covariance and correlation between X_1 and Z. We have:

- $E[X_1] = E[X_1^2] = P(X_1 = 1) = 0.5$

- $E[Z] = E[Z^2] = P(Z = 1) = 0.4$
- $E[X_1 Z] = P(X_1 = 1, Z_1 = 1) = 0.3$

Then we have:

- $Cov(X_1, Z) = E[X_1 Z] - E[X_1]E[Z] = 0.3 - 0.5 \times 0.4 = 0.1$
- $Var(X_1) = E[X_1^2] - (E[X_1])^2 = 0.5(1 - 0.5) = 0.25$
- $Var(Z) = E[Z^2] - (E[Z])^2 = 0.4(1 - 0.4) = 0.24$
- $Corr(X_1, Z) = \rho_{X_1,Z} = \frac{Cov(X_1,Z)}{\sqrt{Var(X_1)Var(Z)}} = \frac{0.1}{\sqrt{0.25 \times 0.24}} \approx 0.408$

We can doublecheck the correlation result above by simulating a large number of coin flips and computing the sample correlation between X_1 and Z.

```
import numpy as np
n = 10000
x1 = np.random.binomial(1, 0.5, n)
x2 = np.random.binomial(1, 0.7, n)
z = (x1 == x2).astype(int)
print("Sample correlation between X1 and Z:", np.corrcoef(x1, z)[0, 1])
```

```
Sample correlation between X1 and Z: 0.39623892451937126
```

Example of Covariance and Correlation

Example 1.36. Now using the joint distribution of X and Y in Example 1.24

$$f_{X,Y}(x, y) = \begin{cases} 2 & \text{if } 0 < x < 1 \text{ and } 0 < y < x \\ 0 & \text{otherwise} \end{cases}$$

we can compute the covariance and correlation between X and Y. We have:

- $E[X] = \int_0^1 x \cdot 2x \, dx = \frac{2}{3}$
- $E[Y] = \int_0^1 y \cdot 2(1 - y) \, dy = 1 - 2/3 = 1/3$
- $E[X^2] = \int_0^1 x^2 \cdot 2x \, dx = \frac{1}{2}$
- $E[Y^2] = \int_0^1 y^2 \cdot 2(1 - y) \, dy = \frac{2}{3} - \frac{1}{2} = \frac{1}{6}$
- $Var(X) = E[X^2] - (E[X])^2 = \frac{1}{2} - \left(\frac{2}{3}\right)^2 = \frac{1}{18}$
- $Var(Y) = E[Y^2] - (E[Y])^2 = \frac{1}{6} - \left(\frac{1}{3}\right)^2 = \frac{1}{18}$
- $E[XY] = \int_0^1 \int_0^x xy \cdot 2 \, dy \, dx = \int_0^1 x^3 \, dx = \frac{1}{4}$

Then we have:

- $Cov(X,Y) = E[XY] - E[X]E[Y] = \frac{1}{4} - \frac{2}{3} \times \frac{1}{3} = \frac{1}{36}$
- $Corr(X,Y) = \rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\frac{1}{36}}{\sqrt{\frac{1}{18} \times \frac{1}{18}}} = \frac{1}{2}$

Now assume we have two independent random variables X and Y uniformly distributed on the interval $[0,1]$.

Then for any x and y in the interval $[0,1]$ such that $x > y$, we have:

$$\begin{aligned}
F_{X,Y}(x,y|X>Y) &= P(X \le x, Y \le y | X > Y) \\
&= \frac{P(X \le x, Y \le y, X > Y)}{P(X > Y)} \\
&= \frac{\int_0^y \int_0^u f_X(u) f_Y(v) dvdu + \int_y^x \int_0^y f_X(u) f_Y(v) dvdu}{\int_0^1 \int_0^u f_X(u) f_Y(v) dvdu} \\
&= \frac{\int_0^y \int_0^u dvdu + \int_y^x \int_0^y dvdu}{\int_0^1 \int_0^u dvdu} \\
&= \frac{\frac{y^2}{2} + y(x-y)}{1/2} \\
&= y^2 + 2y(x-y)\,, \quad 0 < y < x < 1
\end{aligned}$$

which is the same cdf in Example 1.24.

This gives us a way to simulate the joint distribution of X and Y given $X > Y$. We can first simulate two independent uniform random variables X and Y on the interval $[0,1]$, and then keep only the pairs (X,Y) such that $X > Y$. Wer show below a Python code to do this and in passing we double check the correlation result above.

```
import numpy as np
import matplotlib.pyplot as plt

n = 10000
x = np.random.uniform(0, 1, n)
y = np.random.uniform(0, 1, n)
z = x>y
```

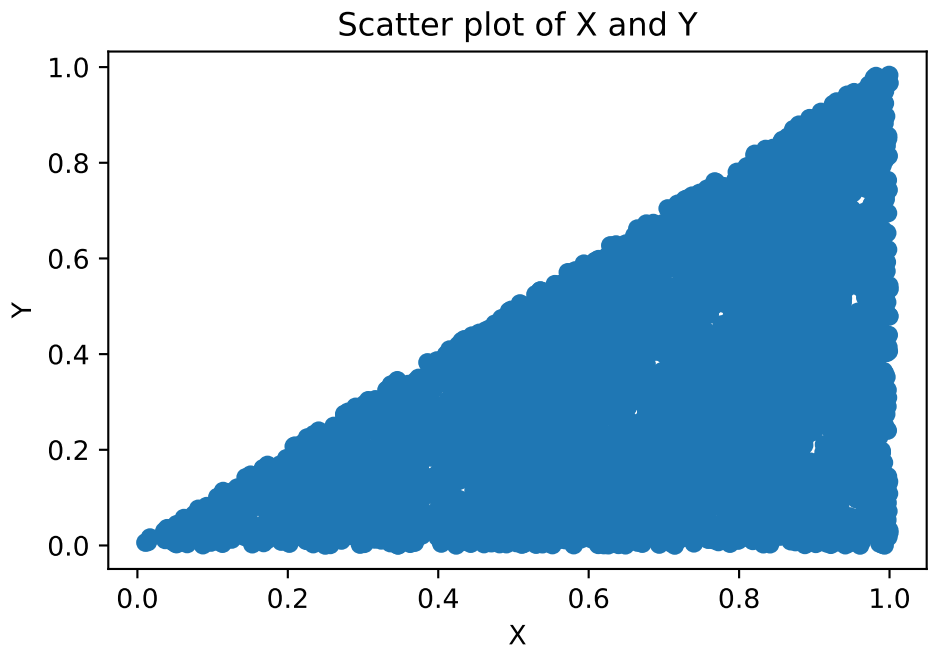
```
y = y[z]
x = x[z]

# can compute the mean of each variable

mean_x = np.mean(x)
mean_y = np.mean(y)
var_x = np.var(x)
var_y = np.var(y)
# display the means and variances
print("Mean of X:", mean_x)
print("Mean of Y:", mean_y)
print("Variance of X:", var_x)
print("Variance of Y:", var_y)
print("Correlation:", np.corrcoef(x, y)[0, 1])


plt.scatter(x, y)
plt.xlabel('X')
plt.ylabel('Y')
plt.title('Scatter plot of X and Y')
plt.show()
```

```
Mean of X: 0.6726852516343149
Mean of Y: 0.3391201776985895
Variance of X: 0.054949285110713585
Variance of Y: 0.05588954650546003
Correlation: 0.4933618055224526
```



The scatter plot shows a negative correlation between X and Y, which is consistent with our calculation of the correlation coefficient.

Covariance and correlation under independence

Proposition 1.6. *If X and Y are independent random variables, then $E[XY] = E[X]E[Y]$ and therefore*

- $Cov(X, Y) = 0$
- $\rho_{X,Y} = 0$

2 Random Sample and Sampling Distributions

2.1 Random sample

Statistics is the science of collecting, analysing, and interpreting data. The earliest applications of statistics were on demographic and economic measures and were driven by the state (from where the name "statistics" comes).

We use statistics when we want to draw conclusions about a set of individuals which we are unable to examine in its entirety. We then define the **population** as the set of individuals that we want to draw conclusions about while the **sample** is defined as the portion of the population that we actually examine. The number of individuals in the sample corresponds to the **sample size**. The measured characteristic from each individual in the sample is a random variable and the collection of characteristics from all individuals in the sample is called a **random sample**. Each element in the random sample is an observation from the same population. The set of all possible values of these random variables is called the **sample space**.

Often we wish to measure some unknown characteristic of the population. A characteristic of the population is called a **parameter**. The set of all possible values of the parameters is called the **parameter space**. We use the sample to infer the value of the parameter. Any quantity calculated from the sample is called a **statistic**. A statistic is therefore a random variable and its distribution is called the **sampling distribution**.

Example 2.1. Market research organisations conduct opinion polls regularly. Figure 1.1 shows the result of such a poll. This poll was conducted by the firm YouGov on **15 October 2021**. The question asked participants to state whether they felt older, the same, or younger than their real age. We can see that **4621** adults from Great Britain responded to this question, so the sample size is $n = 4621$. This number is significantly lower than the adult population of Great Britain, but it would have been impractical for YouGov to poll every adult.

From the results of the poll we can see that **16%** of the responders feel older than their real age, **32%** feel the same, and **47%** feel younger. The remaining **5%** said they don't know. Although these results are derived from the sample, if we assume that the sample is properly chosen, then we can claim that the corresponding proportions in the whole population would be similar.

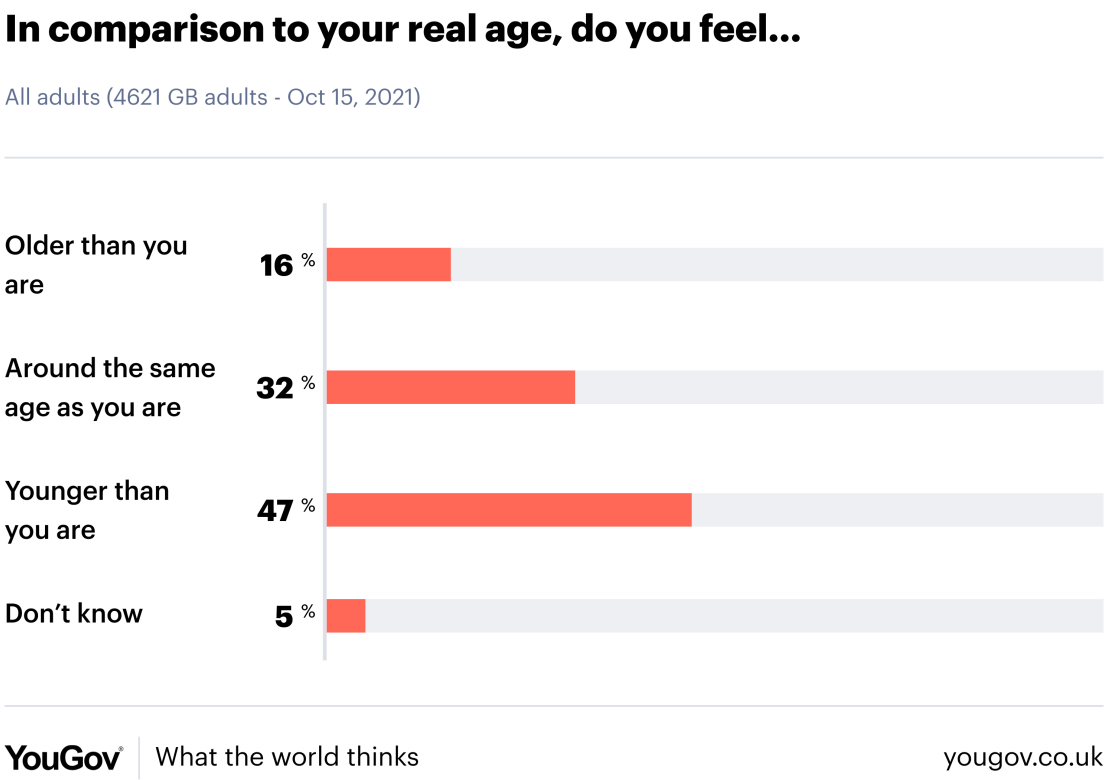


Figure 2.1: An example of a poll Source: YouGov

Example 2.2. The Office for National Statistics (ONS) wishes to measure the unemployment rate in the UK. To that end, it chooses people of working age within the UK and asks them whether they are employed or seeking employment. The proportion among those asked who are seeking employment can be used to estimate the unemployment rate. Figure 1.2 shows a typical warning appearing on ONS's webpage regarding uncertainty in their estimates of population measures.

In this example the population consists of all individuals able to work in the UK. The parameter we wish to estimate is the unemployment rate p which is a proportion so the parameter space is the set $[0, 1]$. Because the ONS cannot ask every individual, it asks a subset of the population. The individuals asked consist of the sample. The proportion in the sample seeking employment is a statistic because it is calculated from the sample and not the whole population.

Suppose n individuals were asked and let X_i denote the response of the ith individual, $i = 1, \ldots, n$. We let $X_i = 1$ if the ith individual is seeking employment and 0 if not so in this case the sample space is the set $\{0, 1\}$. The random sample is the set $\{X_1, \ldots, X_n\}$. The proportion in the sample is also the mean of the X_i's, denoted by $\bar{X}$. Each X_i is distributed as $X_i \sim \text{Bernoulli}(p)$, so the sampling distribution of $\bar{X}$ is the distribution of the sample proportion, $\text{Bin}(n, p)/n$.

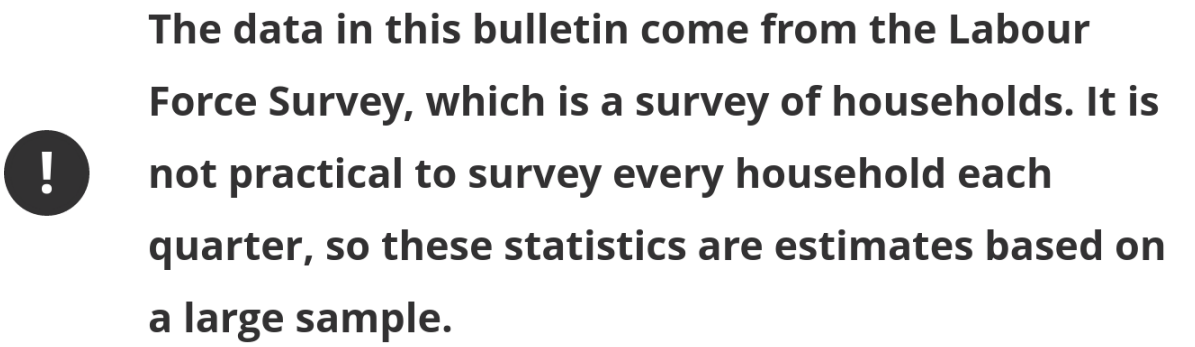


Figure 2.2: ONS uncertainty note (Source: www.ons.gov.uk)

Random sample

Definition 2.1. The random variables $X_1, \ldots, X_n$ are called a **random sample** of size n from the population $f(x \mid \theta)$ depending on a parameter θ if $X_1, \ldots, X_n$ are mutually independent random variables and the probability density/mass function (pdf/pmf) of each X_i is the same function $f(x \mid \theta)$. The variables $X_1, \ldots, X_n$ are also called **independent and identically distributed (iid) random variables**. We write $X_1, \ldots, X_n$ iid $\sim f(x \mid \theta)$.

Often we are interested in the joint distribution of our sample. Let $X_1, \ldots, X_n$ iid $\sim f(x \mid \theta)$.

Then the joint pdf/pmf of $X_1, \ldots, X_n$ is

$$f(x_1, \ldots, x_n \mid \theta) = f(x_1 \mid \theta) \times \cdots \times f(x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta),$$

where the first equality is true because the random variables are mutually independent.

Example 2.3. Let $X_1, \ldots, X_n$ iid $\sim$ Exponential(μ), where μ denotes the mean of the distribution. For example $X_1, \ldots, X_n$ may correspond to the failure times (measured in years) for n identical circuit boards that are put to test and used until they fail and μ denotes the average lifetime. Note that with this notation, the rate parameter is $\lambda = 1/\mu$.

Each X_i has pdf $f(x \mid \mu) = \frac{1}{\mu} \exp\left(-\frac{x}{\mu}\right)$, so the joint pdf of the sample is

$$f(x_1, \ldots, x_n \mid \mu) = \prod_{i=1}^{n} f(x_i \mid \mu) = \frac{1}{\mu^n} \exp\left(-\frac{1}{\mu} \sum_{i=1}^{n} x_i\right).$$

2.2 Statistics and their sampling distributions

In statistical inference, we are interested in describing the distribution of the population. In most cases, a suitable calculation using the sampled values can help.

Statistic and its sampling distribution

Definition 2.2. Let $X_1, \ldots, X_n$ iid $\sim f(x \mid \theta)$. A function $T = T(X_1, \ldots, X_n)$ of the variables $X_1, \ldots, X_n$, which does not depend on θ, is called a **statistic**. The statistic is itself a random variable. The probability distribution of T is called its **sampling distribution**.

In other words, any quantity that is calculated using the sample is a statistic. Another way to think of the sampling distribution is as the distribution of all possible values of T for all possible random samples of size n from the population $f(x \mid \theta)$.

Example 2.4. Let $X_1, \ldots, X_n$ be a random sample of size n. Two of the most frequently used statistics are the sample mean, $\bar{X}$, and the sample variance S^2 defined by

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i, \qquad S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

Suppose $X_1, \ldots, X_n$ iid $\sim \mathcal{N}(\mu, \sigma^2)$. Then, the sampling distribution of $\bar{X}$ is $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$, i.e., the normal distribution with mean μ and variance σ^2/n, and the sampling distribution of S^2 is $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$, i.e., the chi-squared distribution with $n-1$ degrees of freedom times the constant $\sigma^2/(n-1)$. Moreover, $\bar{X}$ and S^2 are independent in the case of normal populations.

The sampling distribution is not always easy to derive, either because the distribution of the population is unknown or because the statistic does not have a straightforward expression. Sometimes we can state asymptotic results as the sample size increases.

Law of large numbers

Theorem 2.1. *Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and variance $\sigma^2 < \infty$. Then, the sample mean $\bar{X}$ approximates the population mean μ when the sample size n is large.*
Formally, for any small error $\varepsilon > 0$,

$$\Pr\Big(|\bar{X} - \mu| \geq \varepsilon\Big) \to 0 \quad \textit{as } n \to \infty.$$

Not examinable

Proof. This is easily proved by Chebyshev's inequality: for any random variable Y with variance, $\Pr\{|Y| \geq r\} \leq \frac{\mathrm{Var}(Y)}{r^2}$ for all $r > 0$. Hence,

$$\Pr\{\,|\bar{X} - \mu| \geq \varepsilon\,\} \leq \frac{\mathrm{Var}(\bar{X})}{\varepsilon^2} = \frac{\sigma^2/n}{\varepsilon^2} = \frac{\sigma^2}{n\,\varepsilon^2} \to 0 \quad \text{as } n \to \infty.$$

The law of large numbers simply states that the probability of small deviations of the sample mean from the population mean can be made very small if we choose a large enough sample size. □

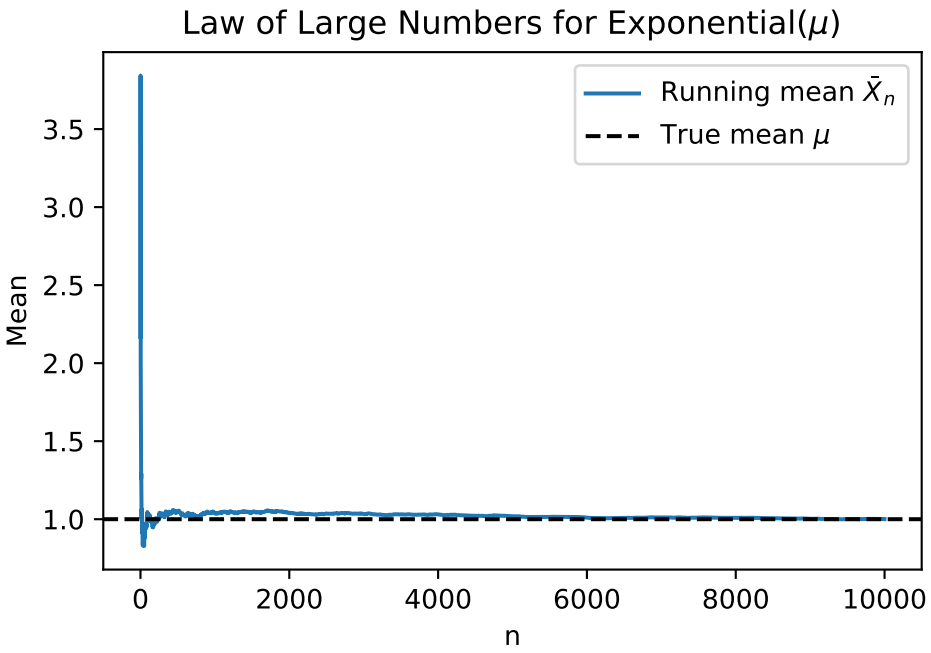
Example 2.5. Suppose $X_1, \ldots, X_n$ iid $\sim$ Exponential(μ). Then $\mathbb{E}[X_i] = \mu$, therefore $\mathbb{E}[\bar{X}] = \mu$. The law of large numbers says that the probability that $|\bar{X} - \mu|$ exceeds a small number ε can become arbitrarily small by increasing the sample size n. This is illustrated by the following Python code.

```
import matplotlib.pyplot as plt
import scipy
import scipy.stats as st
import numpy as np

N = 10000  # Max sample size
mu = 1     # The mean (scale) parameter
x = st.expon.rvs(size=N, scale=mu)
xbar = (x.cumsum()) / (np.arange(1, N+1))

plt.plot(xbar, label='Running mean $\\bar X_n$')  # xbar at n = 1,2,...,N
plt.axhline(mu, ls='--', color='k', label='True mean $\\mu$')
```

```
plt.xlabel('n')
plt.ylabel('Mean')
plt.legend()
plt.title('Law of Large Numbers for Exponential($\\mu$)')
plt.show()
```



Example 2.6. The game of roulette — and why the house always wins. In the game of roulette, a wheel consisting of 37 pockets, numbered 0 to 36, is spun and a ball is dropped onto it (see Figure 1.3). The ball will eventually come to rest in one of the numbered pockets. Players can bet money on the outcome of the spin and win money if they guess correctly.

Suppose a player bets £1 on a specific number x. This player will win £35 if the ball lands in x, otherwise, they lose their bet of £1. In other words, their profit is $+35$ if they win the bet and -1 if they lose the bet. Let X denote the outcome of the wheel spin, and let W denote the player's winnings after one bet. The expected value of W is

$$\mathbb{E}[W] = 35 \Pr\{X = x\} - 1 \cdot \Pr\{X \neq x\} = 35 \cdot \frac{1}{37} - 1 \cdot \frac{36}{37} = \frac{35 - 36}{37} = -\frac{1}{37} \approx -0.027.$$

We observe that the expected winnings, *from a player's point of view*, are negative. This does not mean that a player loses money at every bet, and in fact, it is possible that any one player will win big. However, in a typical day, there are thousands of bets taking place. The average winnings from these bets will converge to the distribution mean of -0.027, so *collectively* every player loses about 2.7 pence per £1 bet on average.

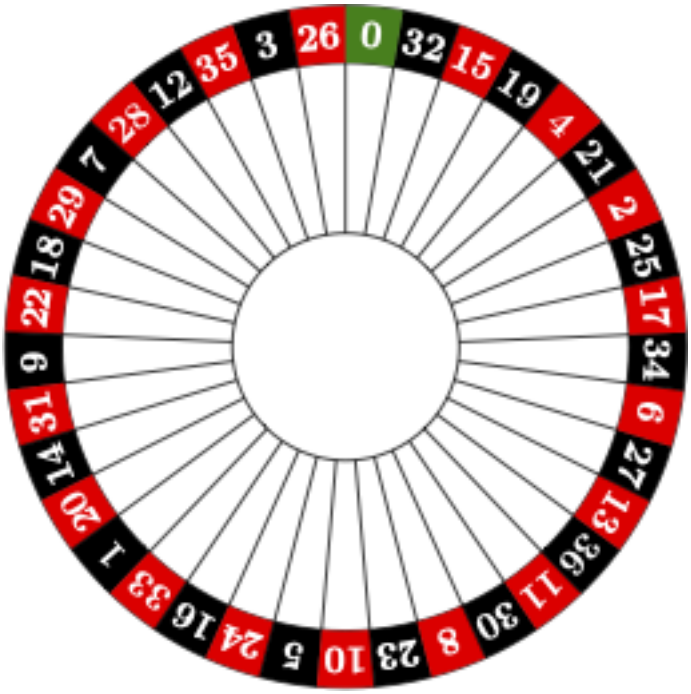


Figure 2.3: Roulette wheel

The sample mean is ubiquitous in statistics and it is important to know its sampling distribution. The next theorem summarises the large-sample behaviour of the sample mean.

Central limit theorem

Theorem 2.2. *Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and variance $\sigma^2 < \infty$. Then, the sampling distribution of the sample mean $\bar{X}$ can be approximated by the normal distribution with mean μ and variance σ^2/n, i.e., $\mathcal{N}(\mu, \sigma^2/n)$, for large sample size n.*
Formally, let $Z_n = \sqrt{n}(\bar{X} - \mu)/\sigma$. Then, for any $z \in \mathbb{R}$,

$$\Pr\{Z_n < z\} \to \Phi(z) \quad \text{as } n \to \infty,$$

where Φ denotes the CDF of the $\mathcal{N}(0,1)$ distribution.

In other words, the central limit theorem says that the CDF of $\bar{X}$ and the CDF of $\mathcal{N}(\mu, \sigma^2/n)$ are visually indistinguishable for large sample size. Since in many cases we cannot come up with the sampling distribution of the sample mean, the approximate normal distribution can be used assuming that the sample size is large.

Not examinable

Proof. We will prove this theorem by showing that the moment generating function (mgf) of Z_n, $M_n(t)$, converges, as $n \to \infty$, to the moment generating function of $\mathcal{N}(0,1)$. Since the mgf determines the distribution of the random variable uniquely, it follows that the limiting distribution of Z_n is $\mathcal{N}(0,1)$. Without loss of generality, we can assume $\mu = 0$. In this case $Z_n = \sqrt{n}\,\bar{X}/\sigma = \sum X_i/(\sqrt{n}\,\sigma)$. If $\mu \neq 0$, we can apply the theorem to the random variables $Y_i = X_i - \mu$ and then substitute $\bar{Y}$ with $\bar{X} - \mu$.
Let $M_X(t)$ denote the mgf of X_i, i.e., $M_X(t) = \mathbb{E}[e^{tX_i}]$. By the properties of the mgf, the mgf of Z_n is

$$M_n(t) = \mathbb{E}[e^{tZ_n}] = \prod_{i=1}^{n} M_X\left(\tfrac{t}{\sqrt{n}\,\sigma}\right) = \left\{M_X\left(\tfrac{t}{\sqrt{n}\,\sigma}\right)\right\}^n.$$

The mgf of $\mathcal{N}(0,1)$ is $M(t) = \exp(t^2/2)$. We will show that $\lim_{n\to\infty} \log M_n(t) = \log M(t)$, i.e.,

$$\lim_{n\to\infty} n \log M_X\left(\tfrac{t}{\sqrt{n}\,\sigma}\right) = \tfrac{t^2}{2}.$$

Let $u = 1/\sqrt{n}$ and consider the limit $\lim_{u\to 0} \dfrac{\log M_X\left(\frac{tu}{\sigma}\right)}{u^2}$. Using L'Hôpital's rule twice (and the facts $M_X(0) = 1$, $M_X'(0) = \mathbb{E}[X_i] = 0$, $M_X''(0) = \mathbb{E}[X_i^2] = \sigma^2$), we obtain the desired limit $t^2/2$. □

Example 2.7. Suppose $X_1, \ldots, X_n$ iid $\sim$ Exponential(μ). Then $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \mu^2$. By the central limit theorem, the distribution of $\bar{X}$ is approximately $\mathcal{N}(\mu, \mu^2/n)$ for large n. This is illustrated by the following Python code.

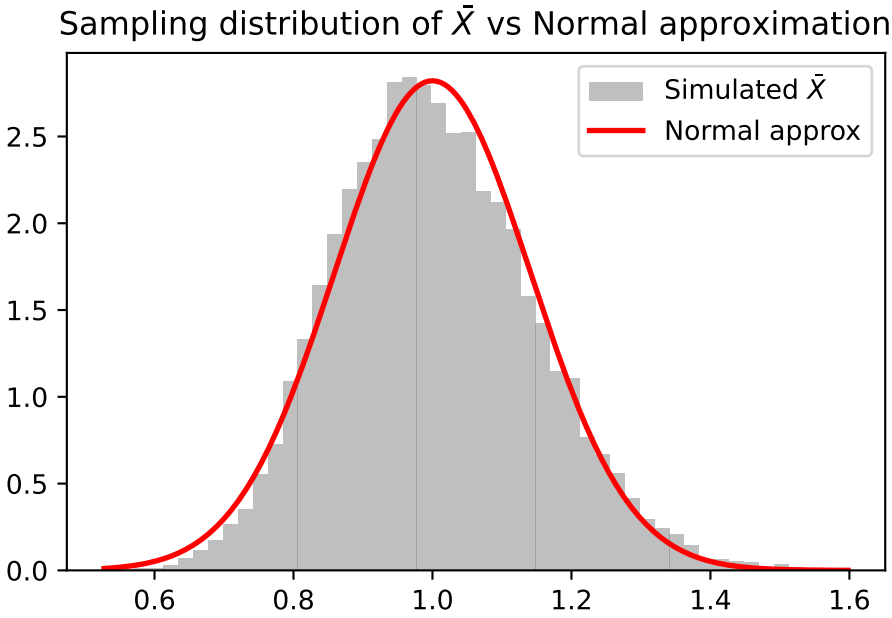
```
import matplotlib.pyplot as plt
import numpy as np
import scipy.stats as st

N = 10000  # Number of repetitions
n = 50     # Sample size for each repetition
mu = 1.0   # The mean (scale) parameter

x = st.expon.rvs(size=(N, n), scale=mu)
xbar = x.mean(axis=1)  # Sample mean across rows

xx = np.linspace(xbar.min(), xbar.max(), 200)
plt.hist(xbar, density=True, bins=50, alpha=0.5, facecolor='gray', label='Simulated $\\bar
plt.plot(xx, st.norm.pdf(xx, mu, mu/np.sqrt(n)), 'r-', lw=2, label='Normal approx')
plt.legend()
plt.title('Sampling distribution of $\\bar X$ vs Normal approximation')
```

```
17 plt.show()
```



2.3 Exercises

1. A coffee shop buys roasted coffee from a supplier. In order to assess the quality of the supplied coffee, the manager of the shop conducts a tasting experiment where she selects a small portion of coffee beans from different batches and tastes the coffee from each portion. For each portion she gives a score in the scale $1, 2, \ldots, 10$ with 10 corresponding to coffee of the best taste and uses the results to assess the quality of the coffee. Identify the population, parameter, and statistic.

2. Read the abstract of the article: Dietary Intake of Marine n-3 Fatty Acids, Fish Intake, and the Risk of Coronary Disease among Men by Ascherio and others published in *The New England Journal of Medicine* in 1995. Identify the population, parameter, sample, and statistic.

3. Let $X_1, \ldots, X_n$ iid $\sim \mathcal{N}(\mu, \sigma^2)$. Derive the sampling distribution of $\bar{X}$ given in Example 1.4.

4. Let $X_1, \ldots, X_n$ iid $\sim$ Bernoulli(p).

   a) Derive the sampling distribution of $\bar{X}$, i.e., for $x \in \{0/n, 1/n, 2/n, \ldots, n/n\}$ find the probability $\Pr\{\bar{X} = x\}$.

Hint. Let $W = \sum_{i=1}^{n} X_i$ so that $\bar{X} = W/n$. First find the distribution of W and use that to find $\Pr\{\bar{X} = x\}$.

b) Derive the asymptotic distribution of $\bar{X}$ from the central limit theorem.

c) Draw a graph of the exact and approximate CDFs when $n = 20$ and $p = 0.4$.

3 Decision Theory

Decision theory is the branch of statistics and probability concerned with making decisions based on data. In many aspects of real life, we are asked to make a decision, which will impact our future in some way, with limited information. This chapter is about putting a mathematical framework around this concept and using that to analyse decisions.

Example 3.1. Local councils in the UK are responsible for maintaining about 225,000 miles of road in total. In winter months this means spreading salt on the roads (gritting) to prevent frost and keep them safe for driving. During the 2011 winter, councils spread 1.2 million tonnes of salt at the cost of £30–40 per tonne. Each winter evening local councils must decide whether to grit or not, based on the available information. To help with their decision, Winter Duty Managers use the national weather forecast as well as sensors embedded in roads which measure road and air temperatures, rain, dew and salt levels (see here and Figure 3.1). Due to the high cost of gritting, it is important that it is done only when necessary, however, it is impossible to know with certainty when to do it.

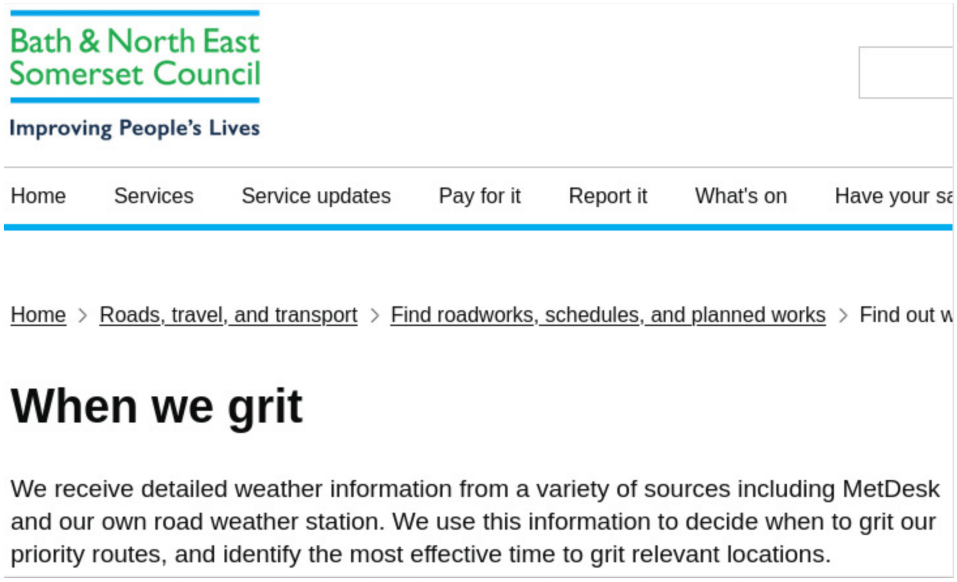


Figure 3.1: The Bath and North-East Somerset council webpage detailing what information they use to decide whether to grit the roads.

3.1 Mathematical formulation

In decision theory, the person making the decision, the decision-maker, is given a set of possible actions, $\mathcal{A}$, and data, $\mathbf{x}$. It is rarely the case that we can make decisions having complete information. The unknown state of nature is represented by a parameter, $\theta \in \Theta$, and the task is to choose the best possible action $a \in \mathcal{A}$, according to some loss function $L(\theta, a)$.

Example 3.2. Below are some examples of decision problems in various areas.

1. When building flood defences around rivers, the decision is how high to build them. Higher defences protect against extreme rainfalls, however, they cost more. The decision in this case is the height of the barrier, the loss is a function of the construction cost and the economic damage in the event of barrier failure, while the state of nature would be the probability of flooding in any year. The data in this case consist of river heights in previous years.
2. When playing poker, the decisions are whether to fold, call, or raise the bid, and if so, by how much (which is a number between 0 and our current money). The unknown state of nature is the probability of winning. The data are the player's hand and the bids, while the uncertainty comes from not knowing the opponents' hands. The loss in this case is the money that we will lose from playing the game.
3. A football coach must decide who and how they should play. The uncertainty comes from how the opposing team plays. The unknown state of nature is the probabilities that each team scores a goal. The data in this case consist of the performances of the teams in previous matches and the loss function represents the final score of the match.
4. Facing a pandemic, the government must decide what measures are appropriate ranging from complete indifference to total lockdown. The data consist of the daily infection numbers and advice from scientific advisors, while the uncertain parameters are the reproduction rate of the disease. The loss in this case can be measured in terms of the number of deaths combined with the impact of the measures to the economy.

To set the mathematical framework, we first define the terms that comprise a decision problem.

Decision problem

Definition 3.1. A decision problem consist of the following elements.

- A **parameter** $\theta \in \Theta$, where Θ is the **parameter space**. The parameter represents the unknown state of nature.
- A set of **data** $\mathbf{x} \in \mathcal{X}$, where $\mathcal{X}$ is the **sample space**. The data are assumed to be a random sample from a population depended on the unknown parameter θ, with

distribution $f(\mathbf{x}|\theta)$.

- An **action** $a \in \mathcal{A}$, where $\mathcal{A}$ is the **action space**, i.e., the set of possible actions that we can take.
- A **loss function** $L : \Theta \times \mathcal{A} \mapsto \mathbb{R}$, such that $L(\theta, a)$ denotes the loss when the true parameter value is θ and the decision-maker chooses action a. We prefer lower values of L.

Example 3.3. The three classical examples of loss functions are the quadratic, absolute, and 0-1 loss. The first two are mainly used in the context of parameter estimation, while the latter in hypothesis testing.

1. The quadratic loss is defined by
$$L(\theta, a) = (\theta - a)^2.$$
2. The absolute loss is defined by
$$L(\theta, a) = |\theta - a|.$$
3. The 0-1 loss is defined by
$$L(\theta, a) = \begin{cases} 0 & \text{if } a = \theta, \\ 1 & \text{if } a \neq \theta. \end{cases}$$

Of course, other loss functions are possible, depending on the scenario.

i Note

Often in life, we evaluate the wisdom of our actions only after observing a single outcome. The loss function should *not* capture the loss incurred from a single event; instead, it should reflect the average loss across multiple occurrences of such events. This means the loss function is designed to measure the overall success of an action by taking into account repeated outcomes.

3.2 Decision rule

In practice, the true state of nature, θ, is unknown. The data, $\mathbf{x}$ provide some information about θ that we want to utilise to inform about our action. The solution to the decision problem is obtained by finding a decision rule $d : \mathcal{X} \mapsto \mathcal{A}$, such that $d(\mathbf{x})$ incorporates the data in some way to determine the appropriate action to take. You can think of the decision rule as the strategy for choosing an action, given data $\mathbf{x}$.

Example 3.4. A traveller buying a flight ticket is considering whether to also buy travel insurance that pays up £1000 in the event of a flight cancellation. The travel insurance costs £50. In the event that the flight is cancelled, she will loose her hotel deposit which is £500. The available actions in this case are $\mathcal{A} = \{0, 1\}$ with 1 representing "buy travel insurance" and 0 being "don't buy travel insurance".

To assess the probability of her flight being cancelled, θ, she decides to look into how many times, in the past 10 years, a similar flight was cancelled. Let x be the proportion of times a flight was cancelled. A decision rule $d(x)$ may be

$$d(x) = \begin{cases} 1 & \text{if } x \geq 0.10, \\ 0 & \text{if } x < 0.10, \end{cases} \tag{3.1}$$

in other words, the traveller's strategy is to buy travel insurance if they find that at least 10% of the past flights were cancelled, and not buy travel insurance otherwise.

Let y denote the future event that the flight be cancelled. We set $y = 1$ if the flight is cancelled, and $y = 0$ if the flight is not cancelled. We define the function $l(y, a)$ to denote the loss when we take action a and the event y occurs. Then, if we do buy insurance ($a = 1$), our loss is 50 if the flight is not cancelled (the cost of the insurance), and $50 + 500 - 1000 = -450$ if the flight is cancelled (the cost of the insurance plus the hotel deposit, but we receive a payment of 1000). On the other hand, if we do not buy insurance ($a = 0$), and our flight is not cancelled, our loss is 0, however if the flight is cancelled our loss is 500 (the hotel deposit). Putting these together gives

$$l(y, a = 0) = \begin{cases} 0 & \text{if } y = 0, \\ 500 & \text{if } y = 1, \end{cases} \quad l(y, a = 1) = \begin{cases} 50 & \text{if } y = 0, \\ -450 & \text{if } y = 1. \end{cases}$$

According to our problem, $\mathbb{P}P(y = 1) = \theta$ and $\mathbb{P}P(y = 0) = 1 - \theta$, so the loss function for this problem is computed as the expected value of $l(y, a)$ over the distribution of y:

$$L(\theta, 0) = \mathsf{E}\, l(y, 0) = 0 \times \mathbb{P}P(y = 0) + 500 \times \mathbb{P}P(y = 1) = 0 \times (1 - \theta) + 500 \times \theta = 500\theta$$
$$L(\theta, 1) = \mathsf{E}\, l(y, 1) = 50 \times \mathbb{P}P(y = 0) - 450 \times \mathbb{P}P(y = 1) = 50 \times (1 - \theta) - 450 \times \theta = 50 - 500\theta.$$

This can be combined as

$$L(\theta, a) = \begin{cases} 500\theta & \text{if } a = 0, \\ 50 - 500\theta & \text{if } a = 1. \end{cases} \tag{3.2}$$

In other words, if she does not buy insurance ($a = 0$), the potential loss is $L(\theta, a = 0) = 500 \times \theta$, so the hotel cost times the probability of the flight being cancelled. Note that the actual loss

is going to be 500 if the flight is cancelled, and 0 if the flight is not cancelled, but at this point we don't know whether the flight will be cancelled or not, so $500 \times \theta$ is in fact the expected loss under a future cancellation event. Similarly, if she does buy insurance ($a = 1$), then the loss is $L(\theta, a = 1) = 50 + (500 - 1000) \times \theta = 50 - 500 \times \theta$, where 50 is the insurance cost and will have to pay $500 \times \theta$ for the hotel cost but receive a payment of $1000 \times \theta$ from the insurance. The loss functions for the two decisions are shown in Figure 3.2 .

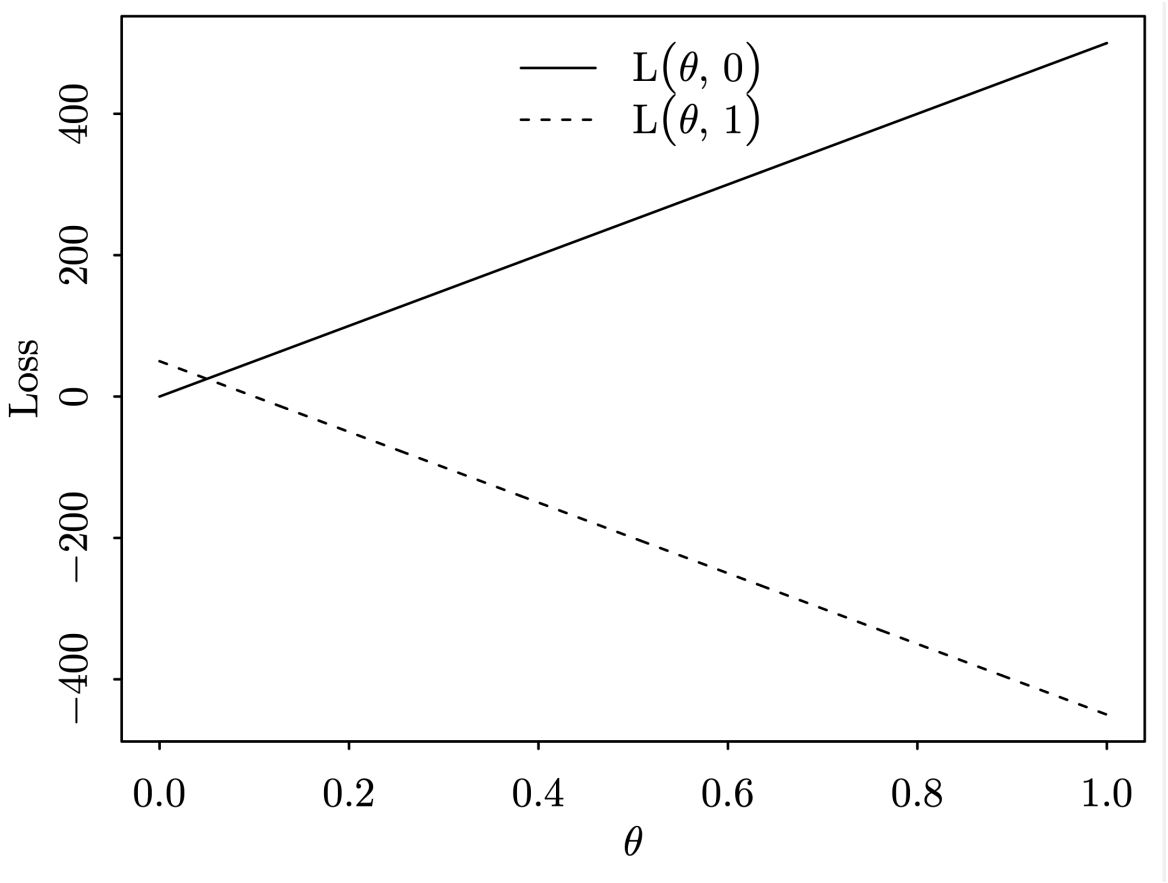


Figure 3.2: Losses for Example 3.4. The two lines intersect at $\theta = 0.05$. We can see that, at $\theta > 0.05$, the losses from not buying insurance are greater than the losses from buying, so if we believe that there is a greater than 5% chance that the flight is cancelled, we will want to buy insurance because in this case we minimise our losses. If we believe that $\theta < 0.05$, then it is the other way around.

It is clear that as $\theta \to 1$, i.e., there is increasing chance that the flight will be cancelled, then $L(\theta, a = 0) \to 500$ and $L(\theta, a = 1) \to -450$, so the loss in this case is minimised when $a = 1$. Similarly, as $\theta \to 0$, $L(\theta, a = 0) \to 0$ and $L(\theta, a = 1) \to 50$, so in this case the loss is minimised when $a = 0$.

It is also possible for a decision rule to ignore the data, or not use any data. One such rule in the context of Example 3.4 may be $d(\mathbf{x}) = 0$, i.e. don't buy travel insurance no matter what proportion of flights were cancelled in the past.

3.2.1 Deterministic and randomised decision rules

A decision rule is called *deterministic* if it specifies a single action to take for given data. The decision rule in Example 3.4 is deterministic because we know which one action the traveller will take when $x \geq 0.10$ and which one action she will take when $x < 0.10$. However, not all decision rules need to do that. A decision rule may specify multiple actions for given data, with corresponding probabilities for each action. In this case, we call the decision rule *randomised*. An example of a randomised decision rule for Example 3.4 is

$$d(x) = \begin{cases} 0 \text{ with probability } 1/5 \text{ and } 1 \text{ otherwise} & \text{if } x \geq 0.10, \\ 0 \text{ with probability } 3/4 \text{ and } 1 \text{ otherwise} & \text{if } x < 0.10. \end{cases} \tag{3.3}$$

In other words, if the observed proportion of cancelled flights, x, turns out to be 0.15, i.e., we are in the case $x \geq 0.10$, then the traveller will pick a random integer between 1 and 5, and if this integer is 1, then she will not buy insurance, but if the integer is 2, 3, 4, or 5, she will.

3.3 Risk

A decision rule uses the observed data to choose an action. To evaluate different decision rules, we want to evaluate how well they fare if different data were observed. So we consider the hypothetical scenario where new data $\mathbf{x}$ are obtained, from the same distribution as our observed data. In this case, we compare them in terms of their expected loss over repeated observations $\mathbf{x}$, which we call the **risk**.

Risk

Definition 3.2. The **risk** of a decision rule d for a parameter value θ, $R(\theta, d)$, based on data $\mathbf{x} \sim f(\mathbf{x}|\theta)$ is defined as

$$R(\theta, d) = \mathop{\mathsf{E}}_{\theta} L(\theta, d(\mathbf{x})),$$

i.e., the expected loss under the action obtained by following the decision rule d.

The subscript in E_θ indicates that the expectation is taken with respect to $\mathbf{x} \sim f(\mathbf{x}|\theta)$. If the decision rule $d(\mathbf{x})$ does not depend on the data $\mathbf{x}$, i.e., it ignores the data $\mathbf{x}$, then $R(\theta, d) = L(\theta, d(\mathbf{x}))$. For example, if the traveller of Example 3.4 is going on a business trip, company

policy might dictate that the traveller should buy travel insurance regardless of how likely it is for the flight to be cancelled. In this case, the traveller's action is $d(\mathbf{x}) = 1$ for all $\mathbf{x} \in \mathcal{X}$.

Example 3.5. (Example 3.4 continued) Suppose there were $n = 900$ flights in the past 10 years that were examined by the traveller. According to our model, each flight has a probability θ of being cancelled. Assuming that the event that a flight be cancelled is independent of whether previous flights were cancelled, the central limit theorem says that the proportion x of cancelled flights is distributed asymptotically as

$$x \sim N\left(\theta, \frac{\theta(1-\theta)}{n}\right).$$

Therefore,

$$\mathbb{P}P(x < 0.10) \approx \Phi\left(\frac{0.10-\theta}{\sqrt{\frac{\theta(1-\theta)}{n}}}\right) = \Phi\left(\frac{30(0.10-\theta)}{\sqrt{\theta(1-\theta)}}\right)$$
$$\mathbb{P}P(x \geq 0.10) \approx 1 - \Phi\left(\frac{30(0.10-\theta)}{\sqrt{\theta(1-\theta)}}\right) = \Phi\left(\frac{30(\theta-0.10)}{\sqrt{\theta(1-\theta)}}\right)$$

Let d_1 denote the decision rule (4.1). According to this rule, the traveller will choose to buy insurance ($a = 1$) with probability $\Phi\left(\frac{30(\theta-0.10)}{\sqrt{\theta(1-\theta)}}\right)$ and not buy ($a = 0$) with probability $\Phi\left(\frac{30(0.10-\theta)}{\sqrt{\theta(1-\theta)}}\right)$. Therefore, the risk of (4.1) according to the loss function (3.2) is

$$\begin{aligned} R(\theta, d_1) &= L(\theta, 0) \times \mathbb{P}P(x < 0.10) + L(\theta, 1) \times \mathbb{P}P(x \geq 0.10) \\ &= (500\theta) \times \Phi\left(\frac{30(0.10-\theta)}{\sqrt{\theta(1-\theta)}}\right) + (50 - 500\theta) \times \Phi\left(\frac{30(\theta-0.10)}{\sqrt{\theta(1-\theta)}}\right) \end{aligned}$$

Now let d_2 denote the randomised rule (3.3). According to this rule, the traveller will choose to buy insurance with probability $4/5$ if $x \geq 0.10$ and with probability $1/4$ if $x < 0.10$. So the risk of this rule is

$$\begin{aligned} R(\theta, d_2) =& L(\theta, 0) \times \left\{\frac{3}{4}\mathbb{P}P(x < 0.10) + \frac{1}{5}\mathbb{P}P(x \geq 0.10)\right\} \\ &+ L(\theta, 1) \times \left\{\frac{1}{4}\mathbb{P}P(x < 0.10) + \frac{4}{5}\mathbb{P}P(x \geq 0.10)\right\} \\ =& (500\theta) \times \left\{\frac{3}{4}\Phi\left(\frac{30(0.10-\theta)}{\sqrt{\theta(1-\theta)}}\right) + \frac{1}{5}\Phi\left(\frac{30(\theta-0.10)}{\sqrt{\theta(1-\theta)}}\right)\right\} \\ &+ (50 - 500\theta) \times \left\{\frac{1}{4}\Phi\left(\frac{30(0.10-\theta)}{\sqrt{\theta(1-\theta)}}\right) + \frac{4}{5}\Phi\left(\frac{30(\theta-0.10)}{\sqrt{\theta(1-\theta)}}\right)\right\}. \end{aligned}$$

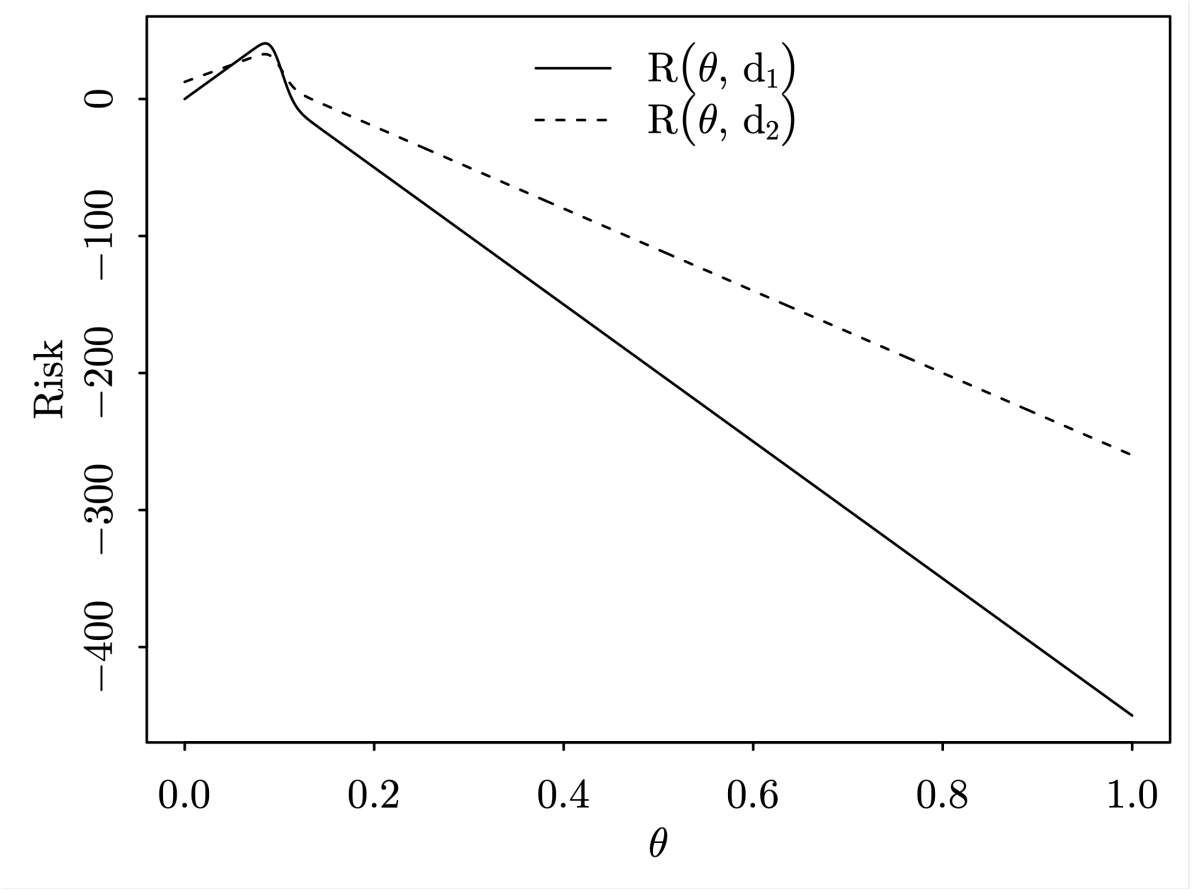


Figure 3.3: Risks for the two decision rules considered in Example 3.5.

The risks of the two decision rules for different θ values are shown in Figure 3.3. Both risks decrease as it becomes more certain that the flight will be cancelled. It can be seen that for most values of θ, d_1 has lower risk than d_2, while d_2 is better for values of θ between 0.05 and 0.10. In fact, both rules have the most risk when $\theta = 0.085$, and in this case d_1 has the highest risk.

3.4 Criteria for choosing a good decision rule

Based on the discussion above, it is clear that we want to choose a decision rule that has low risk among a choice of different decision rules. Let $\mathcal{D}$ denote the set of decision rules that we are considering. For Example 3.4, there is no reason why we should only consider the decision rule (4.1) with a fixed threshold 0.10. We could consider a family of decision rules of the form $\mathcal{D} = \{d_t(x) = (1 \text{ if } x \geq t, \, 0 \text{ otherwise})\}$, $t \in [0, 1]$. The problem then reduces to finding the optimal threshold according to some criterion.

It is apparent from Definition 3.2 and Example 3.5 that the risk of a decision rule is a function of the parameter θ. It is possible that we are not able to find a decision rule that uniformly dominates all other decision rules for all values of θ. On the other hand, we want to choose a decision rule that is optimal regardless of the true value of θ. We present below two criteria that can be used for this purpose.

3.4.1 Minimax criterion

The idea behind the minimax criterion is to safeguard against the worst possible situation. Consider, for example, an aeroplane manufacturer considering various aeroplane designs. The manufacturer wants to choose the design that is safest under the worst possible weather conditions.

For a decision rule d, with risk $R(\theta, d)$, the maximum possible risk, $\bar{R}(d)$, is given by

$$\bar{R}(d) = \max_{\theta \in \Theta} R(\theta, d).$$

The value of θ that maximises $R(\theta, d)$ is the worst possible situation for the decision rule d. Thus, we want to choose the decision rule among all those considered in the set $\mathcal{D}$ that is best under the worst possible conditions, i.e., we want to choose the decision rule with the lowest $\bar{R}(d)$. Such decision rule is called *minimax*, and is given by

$$d_{\text{MM}} = \operatorname*{argmin}_{d \in \mathcal{D}} \bar{R}(d) = \operatorname*{argmin}_{d \in \mathcal{D}} \max_{\theta \in \Theta} R(\theta, d).$$

The notation $\operatorname*{argmin}_{d \in \mathcal{D}}$ reads "the argument that minimises over $d \in \mathcal{D}$", meaning "search over all decision rules $d \in \mathcal{D}$ and pick the one that gives the smallest $\bar{R}(d)$". It is clear by looking at

Figure 3.3 that if we have to choose only between d_1 and d_2 in Example 3.5, then, according to the minimax criterion, we would choose d_2, because it has a lower maximum risk than d_1. Note that it does not matter whether the maximum risk is attained at the same θ value.

Example 3.6. Suppose we consider a family $\mathcal{D}$ of decision rules of the form

$$d_t(x) = \begin{cases} 1 & \text{if } x \geq t, \\ 0 & \text{if } x < t, \end{cases} \tag{3.4}$$

for $t \in [0, 1]$. In other words, we want to find the optimum threshold t such that the traveller decides to buy insurance if the proportion of cancelled flights, x, exceeds that threshold, and not buy otherwise. Then, repeating the calculations from Example 3.5, with an arbitrary threshold t, we have, by Equation 3.4,

$$R(\theta, d_t) = (500\theta) \times \Phi\left(\frac{30(t-\theta)}{\sqrt{\theta(1-\theta)}}\right) + (50 - 500\theta) \times \Phi\left(\frac{30(\theta - t)}{\sqrt{\theta(1-\theta)}}\right). \tag{3.5}$$

Then, $\bar{R}(d_t) = \max_\theta R(\theta, d_t)$.

Finding $\bar{R}(d_t)$ is closed-form is not possible, but we can compute it numerically. A plot of $\bar{R}(d_t)$ for different choices of the threshold t is shown in Figure 3.4. It can be seen that the minimum is attained at $t = 0.05$. For comparison, the risks of the minimax decision rule ($t = 0.05$), and the original decision rule ($t = 0.10$) are shown in Figure 3.5. It can be seen that both rules have similar risks, however, in the region of θ between 0.05 and 0.10 the minimax rule is better.

3.5 Exercises

1. A patient is considering a number of treatment options available through her general practitioner (GP) between receiving medication or having a surgery. The costs of the different treatments vary as well as their likelihood of success. The GP has discussed with the patient the success rates of each treatment when used in other patients.

   Describe the parameter, data, actions, and loss function for this problem.

2. An investor is considering whether or not to buy certain risky bonds. If he buys the bonds, they can be redeemed at maturity for a net gain of £500. There is probability θ that there will be a default on the bonds, in which case the investor is set to lose his investment of £1000. If the investor instead puts his money in a "safe" investment, he will receive a net gain of £300 over the same period.

   a. Define appropriate actions, parameter, and parameter space for the problem.

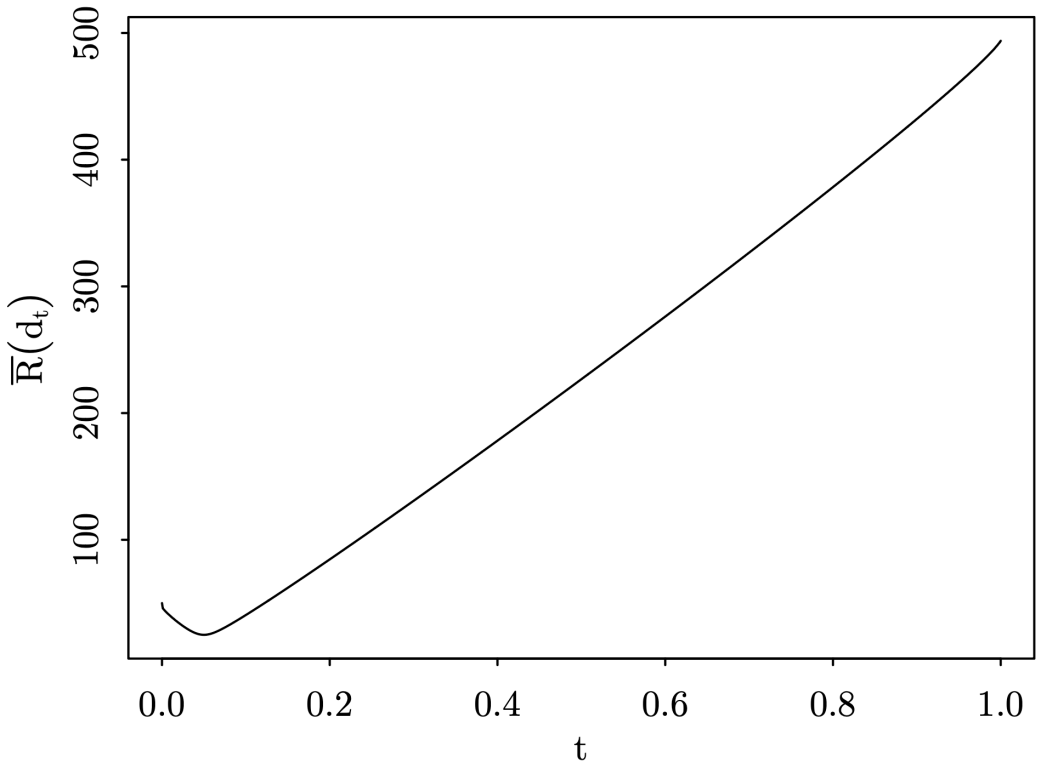


Figure 3.4: Maximum risk for the varying threshold of the decision rules in Example 3.6

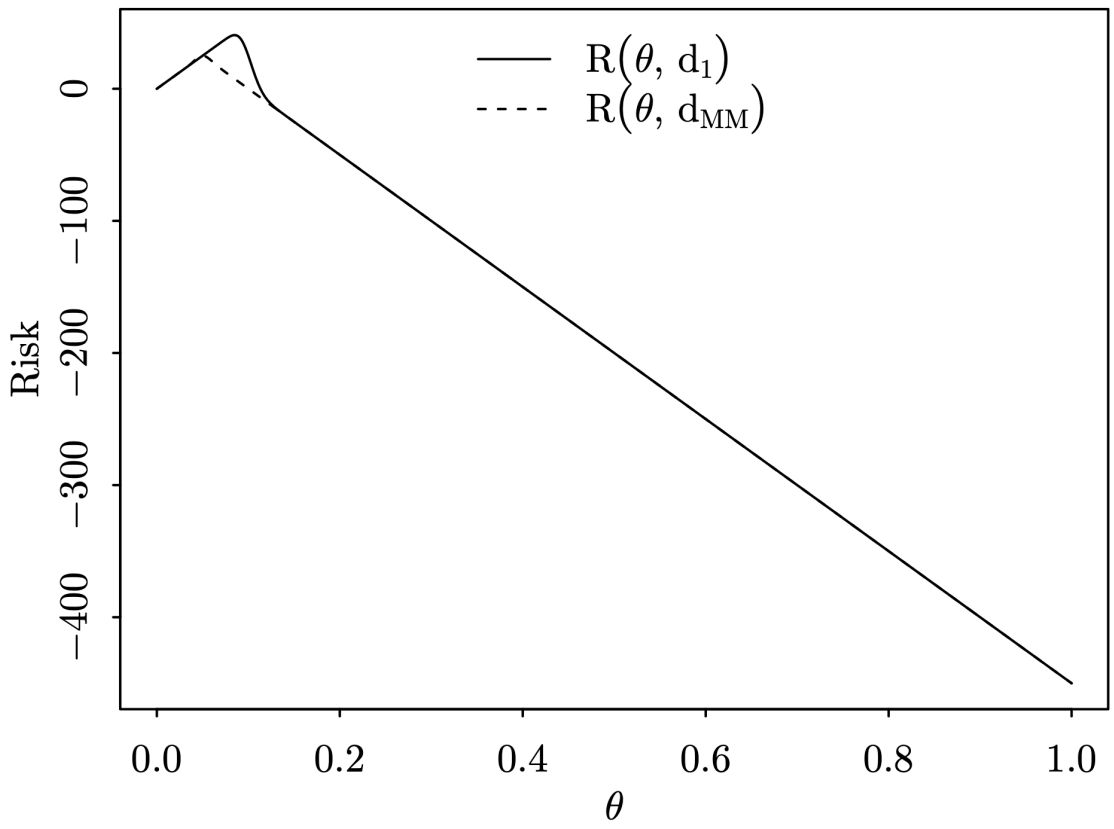


Figure 3.5: Risks of the decision rules for thresholds $t = 0.10$ and $t = 0.05$ (minimax) in Example 3.6.

b. Derive the loss function for the problem.

c. Describe all randomised decision rules and find the minimax decision among them.

3. A coin has probability $\theta \in [0, 1]$ of coming up heads ($y = 1$), and $1 - \theta$ of coming up tails ($y = 0$). You are playing a game where if you guess the outcome of a coin flip correctly you receive a payment of £1, but if you guess wrongly, you loose £1.

   a. What are the parameter and parameter space for this problem?

   b. What is the action space for this problem?

   c. Show that the loss function, $L(\theta, a)$, for this problem is given by

$$L(\theta, a) = \begin{cases} 2\theta - 1 & \text{if guessing "tails",} \\ 1 - 2\theta & \text{if guessing "heads".} \end{cases}$$

   4. Let x be the outcome the coin flip from an earlier game. Consider the following two strategies for guessing the outcome of a future coin flip:

      - **Strategy 1:** Guess the same as the outcome of the earlier coin flip.
      - **Strategy 2:** Guess "heads" regardless of the outcome of the earlier coin flip.

      a. Write a mathematical expression for the decision rules corresponding to these two strategies.

      b. Between the two strategies, which one is the minimax decision rule?

4 Parameter Estimation

In statistical inference we are interested in making conclusions about the population of interest. Often this means making a statement about an unknown parameter describing the population. In this chapter we discuss methods using data that can infer the value of the unknown parameter.

4.1 Point estimation

Suppose we are given a random sample from a population $f(x|\theta)$ depending on an unknown parameter θ with values in the parameter space Θ, i.e., $\theta \in \Theta$. We wish to use the sample to infer the value of θ within Θ. Any function of the sample which can be used for this purpose is an estimator for θ.

Estimator

Definition 4.1. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x|\theta)$ be a random sample from a population which depends on a parameter $\theta \in \Theta$. Any statistic $T = T(X_1, \ldots, X_n)$ taking values in a subset of Θ, i.e., $T \in \Theta$, is called an **estimator** for the parameter θ. Suppose we observe $X_1 = x_1, \ldots, X_n = x_n$ and evaluate $t = T(x_1, \ldots, x_n)$. The value t corresponding to the observed values $x_1, \ldots, x_n$ is called an **estimate** of θ.

Note that an estimator, being a function of the random sample, is itself a random variable. We can therefore talk about the distribution of this estimator. We can potentially come up with several estimators so using their distribution we can evaluate their performance. Two of the most commonly used criteria for evaluating estimators are the **bias** and the **mean squared error** which we define below.

Bias

Definition 4.2. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x|\theta)$ and let $T = T(X_1, \ldots, X_n)$ be an estimator for θ. The difference

$$\text{Bias}_\theta(T) = \mathsf{E}\, T - \theta,$$

is called the **bias** of the estimator T for the parameter θ. If $\text{Bias}_\theta(T) = 0$, then the estimator T is called **unbiased** for θ, otherwise it is called **biased** for θ.

A desirable property for an estimator is to be unbiased.

Mean squared error (MSE)

Definition 4.3. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x|\theta)$ and let $T = T(X_1, \ldots, X_n)$ be an estimator for θ. The **mean squared error (MSE)** of the estimator T for the parameter θ is defined by

$$\text{MSE}_\theta(T) = \mathsf{E}\left\{(T-\theta)^2\right\}.$$

The MSE is always non-negative. It is desirable that the MSE be small. Figure 4.1 illustrates the bias and variability of four different estimators. The estimator with low bias and variability is in fact the one with the lowest MSE as the following lemma tells us.

Mean squared error decomposition

Proposition 4.1. *Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x|\theta)$ and let $T = T(X_1, \ldots, X_n)$ be an estimator for θ. Then*

$$\text{MSE}_\theta(T) = \mathsf{Var}\, T + (\text{Bias}_\theta(T))^2$$

Proof. By the definition of MSE, add and subtract $\mathsf{E}\, T$ in the brackets, and note that $\mathsf{E}\, T$ and θ are not random,

$$\begin{aligned}
\text{MSE}_\theta(T) &= \mathsf{E}\left\{(T-\theta)^2\right\} \\
&= \mathsf{E}\left\{(T-\mathsf{E}\, T+\mathsf{E}\, T-\theta)^2\right\} \\
&= \mathsf{E}\left\{(T-\mathsf{E}\, T)^2 + 2(T-\mathsf{E}\, T)(\mathsf{E}\, T-\theta) + (\mathsf{E}\, T-\theta)^2\right\} \\
&= \mathsf{E}\left\{(T-\mathsf{E}\, T)^2\right\} + 2\,\mathsf{E}\left\{(T-\mathsf{E}\, T)(\mathsf{E}\, T-\theta)\right\} + \mathsf{E}\left\{(\mathsf{E}\, T-\theta)^2\right\} \\
&= \mathsf{E}\left\{(T-\mathsf{E}\, T)^2\right\} + 2(\mathsf{E}\, T-\theta)\,\mathsf{E}\left\{(T-\mathsf{E}\, T)\right\} + (\mathsf{E}\, T-\theta)^2 \\
&= \mathsf{Var}\, T + 0 + (\text{Bias}_\theta(T))^2.
\end{aligned}$$

□

According to Proposition 4.1 , the MSE incorporates two components, one measuring the variability of the estimator and the other measuring its bias (accuracy). An estimator with low MSE has low combined variance and bias.

Example 4.1. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x|\theta) \overset{\text{iid}}{\sim} \text{N}(\mu, \sigma^2)$. The parameter space for μ is $\mathbb{R}$ and for σ^2 is $[0, \infty)$. Then $\bar{X}$ is an estimator for μ because $\bar{X} \in \mathbb{R}$. Its bias is $\text{Bias}_\mu(\bar{X}) = \mathsf{E}\,\bar{X} - \mu = \mu - \mu = 0$ and its variance is $\mathsf{Var}\,\bar{X} = \sigma^2/n$. Therefore, its MSE is $\text{MSE}_\mu(\bar{X}) = \mathsf{Var}\,\bar{X} + \text{Bias}_\mu(\bar{X})^2 = \sigma^2/n$.

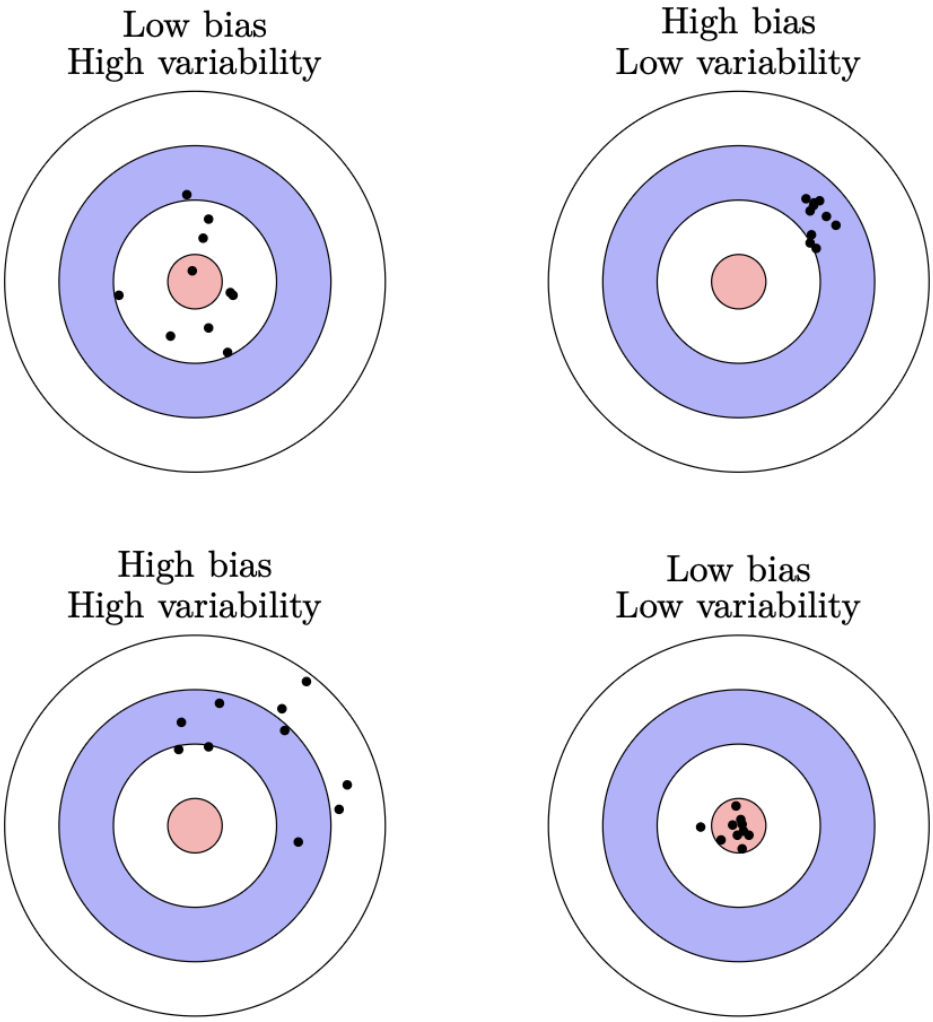


Figure 4.1: Illustration of the bias and variability of an estimator.

Example 4.2. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$. The parameter space for p is $[0,1]$. Then $\bar{X}$ is an estimator for p because $\bar{X} \in \{0, 1/n, 2/n, \ldots, 1\} \subset [0,1]$. Its bias is $\text{Bias}_p(\bar{X}) = \mathsf{E}\,\bar{X} - p = p - p = 0$ and its variance is $\mathsf{Var}\,\bar{X} = p(1-p)/n$. Therefore, its MSE is $\text{MSE}_p(\bar{X}) = p(1-p)/n$. Because $p(1-p) \in [0, \frac{1}{4}]$, with the lower bound attained when $p = 0$ or 1 and the upper bound attained when $p = \frac{1}{2}$, $\text{MSE}_p(\bar{X}) \in [0, \frac{1}{4n}]$.

Consider a different estimator given by $T = \dfrac{2\sum X_i + \sqrt{n}}{2n + 2\sqrt{n}}$. Because $\sum X_i \in \{0, 1, \ldots, n\}$, $T \in [0,1]$ so it is an estimator for p. Its bias is $\text{Bias}_p(T) = \dfrac{2np + \sqrt{n}}{2n + 2\sqrt{n}} - p = (1-2p)\dfrac{\sqrt{n}}{2n + 2\sqrt{n}} = \dfrac{\frac{1}{2} - p}{\sqrt{n} + 1}$. So this estimator has no bias if $p = \frac{1}{2}$ but has positive bias (overestimates p) if $p < \frac{1}{2}$ and negative bias (underestimates p) if $p > \frac{1}{2}$. The variance of this estimator is $\mathsf{Var}\,T = \dfrac{np(1-p)}{(n+\sqrt{n})^2} = \dfrac{p(1-p)}{(\sqrt{n}+1)^2}$ so $\text{MSE}_p(T) = \dfrac{p(1-p) + (\frac{1}{2} - p)^2}{(\sqrt{n}+1)^2} = \dfrac{\frac{1}{4}}{(\sqrt{n}+1)^2}$.

If we wish to choose between $\bar{X}$ and T in terms of their MSE, we see that for all $n \geq 1$, $0 < \text{MSE}_p(T) < \frac{1}{4n}$ so it falls between the values of $\text{MSE}_p(\bar{X})$. In particular, if in reality $p = \frac{1}{2}$, then T will always have lower MSE than $\bar{X}$. For some other value of p, say $p = \frac{1}{5}$ then T has lower MSE if $n \leq 16$ but otherwise $\bar{X}$ has lower MSE (see Figure 4.2).

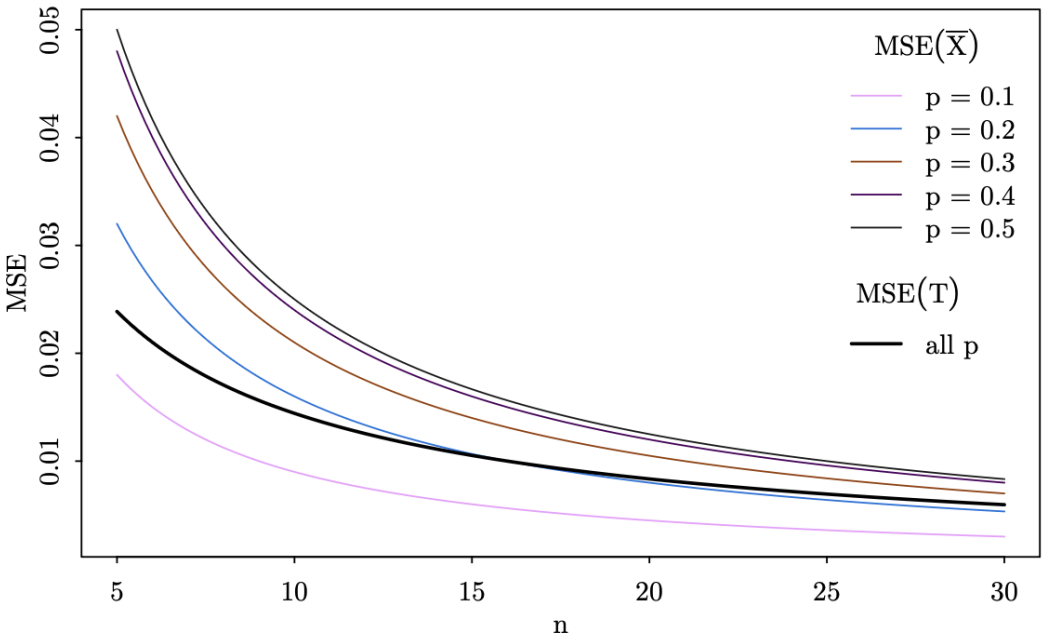


Figure 4.2: Comparison of $MSE_p(\bar{X})$ and $MSE_p(T)$ for Example 4.2 for different values of the parameter p.

We discuss next a few classical estimation methods.

4.1.1 Method of moments estimator

The method of moments estimation is the simplest method for finding estimators. Consider a population $f(x|\theta)$, $\theta \in \Theta$ and define the rth moment by

$$\mu_r = \mathsf{E}(X^r), \text{ for } r = 1, 2, \ldots,$$

i.e., the expectation of the rth power of X. In the case $r = 1$, $\mu_1 = \mathsf{E}\, X$ corresponds to the mean of the population, while for $r = 2$, $\mu_2 = \mathsf{E}\, X^2$, so $\mathsf{Var}\, X = \mathsf{E}\, X^2 - (\mathsf{E}\, X)^2 = \mu_2 - \mu_1^2$.

A convenient method for computing the moments is through the moment generating function (mgf). Recall $M_X(t) = \mathsf{E} \exp(tX)$ and

$$\mu_r = \frac{d^r}{dt^r} M_X(t)|_{t=0}.$$

Example 4.3. Let $X \sim \text{Exponential}(\mu)$, i.e., $f(x|\mu) = (1/\mu)\exp(-x/\mu)$. Then

$$\begin{aligned}
M_X(t) &= \int_0^\infty e^{tx} \frac{1}{\mu} e^{-\frac{x}{\mu}}\, \mathrm{d}x \\
&= \frac{1}{\mu} \int_0^\infty e^{-x(\frac{1}{\mu} - t)}\, \mathrm{d}x \\
&= \frac{1}{\mu} (\frac{1}{\mu} - t)^{-1} \left[-e^{-x(\frac{1}{\mu} - t)} \right]_0^\infty \\
&= (1 - t\mu)^{-1}, \text{ assuming } t < 1/\mu. \\
\Rightarrow M_X^{(1)}(t) &= \frac{d}{dt} M_X(t) = \mu(1 - t\mu)^{-2} \\
\Rightarrow \mu_1 &= M_X^{(1)}(0) = \mu \\
\Rightarrow M_X^{(2)}(t) &= \frac{d^2}{dt^2} M_X(t) = 2\mu^2(1 - t\mu)^{-3} \\
\Rightarrow \mu_2 &= M_X^{(2)}(0) = 2\mu^2
\end{aligned}$$

It is apparent that the rth moment is a function of the parameter θ which we write as $\mu_r(\theta)$ to make this dependence explicit. Note that θ may be a scalar or a κ-dimensional vector $\theta = (\theta_1, \ldots, \theta_\kappa)$.

Now suppose $X_1, \ldots, X_n \overset{\mathsf{iid}}{\sim} f(x|\theta)$ and consider the rth *sample* moment

$$m_r = \frac{1}{n} \sum_{i=1}^n X_i^r, \text{ for } r = 1, 2, \ldots.$$

In particular $m_1 = \bar{X}$ and $m_2 = \frac{1}{n} \sum X_i^2$. The sample moments are functions of the sample $\mathbf{X} = \{X_1, \ldots, X_n\}$ and we write $m_r(\mathbf{X})$ to make this dependence explicit.

Method of moments estimator (MoM)

Definition 4.4. The method of moments estimates the rth moment by the corresponding sample moment, i.e., the method of moments estimator (MoM) for θ, which we denote by $\hat{\theta}$, is given by the solution of the following system of equations,

$$\mu_r(\hat{\theta}) = m_r(\mathbf{X}), \text{ for } r = 1, 2, \ldots.$$

Because there are κ unknown parameters, we need κ equations to be able to identify $\hat{\theta}$ uniquely. These are selected among those equations corresponding to the lowest moments up to as many as needed to be able to solve for $\hat{\theta}$.

Example 4.4. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exponential}(\mu)$, $\mu > 0$. Then $\mu_1 = \mu$ so $\hat{\mu} = \bar{X}$ is the method of moments estimator.

Example 4.5. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$, i.e., normal with known mean 0 and unknown variance $\sigma^2 > 0$. Then $\mu_1 = 0$ and $\mu_2 = \sigma^2$. Note that the first moment does not depend on the parameter so the first equation, $\mu_1 = \bar{X}$, is not helpful for estimating σ^2. Using the second equation we have $\hat{\sigma}^2 = m_2$.

Example 4.6. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{U}(0, \theta)$, i.e., uniform with known lower bound 0 and unknown upper bound $\theta > 0$. The pdf is $f(x|\theta) = \theta^{-1}$, $x \in (0, \theta)$, so $\mu_1 = \int_0^\theta x\theta^{-1}\,\mathrm{d}x = \theta^{-1}\left[\frac{x^2}{2}\right]_0^\theta = \theta/2$. Using the first moment equation we have $\hat{\theta}/2 = \bar{X} \Rightarrow \hat{\theta} = 2\bar{X}$.

Example 4.7. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \beta)$, i.e., gamma with shape $\alpha > 0$ and rate $\beta > 0$ (this distribution was defined in Example 1.34). In this case there are two parameters to estimate, i.e., $\kappa = 2$. The mgf of this distribution is given by $M_X(t) = (1 - \frac{t}{\beta})^{-\alpha}$. Then $\mu_1 = \alpha/\beta$ and $\mu_2 = \alpha/\beta^2 + \alpha^2/\beta^2$. This leads to the following system of equations

$$\hat{\alpha}/\hat{\beta} = m_1, \qquad \hat{\alpha}/\hat{\beta}^2 + \hat{\alpha}^2/\hat{\beta}^2 = m_2.$$

By substituting $\hat{\alpha} = \hat{\beta}m_1$ from the first equation into the second, we have $m_1/\hat{\beta} + m_1^2 = m_2 \Rightarrow \hat{\beta} = m_1/(m_2 - m_1^2)$ and $\hat{\alpha} = m_1^2/(m_2 - m_1^2)$.

An obvious question to ask is are these actual estimators? In other words, are $\hat{\alpha}$ and $\hat{\beta} > 0$? To check this we need to check whether $m_2 > m_1^2$ for all possible samples. But $0 < \sum(X_i - \bar{X})^2 = \sum(X_i^2 - 2X_i\bar{X} + \bar{X}^2) = \sum X_i^2 - 2\bar{X}\sum X_i + n\bar{X}^2 = \sum X_i^2 - n\bar{X}^2$. So $\sum X_i^2 > n\bar{X}^2 \Rightarrow \frac{1}{n}\sum X_i^2 > n\bar{X}^2 \Rightarrow m_2 > m_1^2$ as required.

4.1.2 Maximum likelihood estimator

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x|\theta)$, $\theta \in \Theta$. Then, the joint density/mass function of $\mathbf{X} = (X_1, \ldots, X_n)$ is given by

$$f(\mathbf{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta). \tag{4.1}$$

[1]

In (4.1), we see the parameter θ as fixed and evaluate the function at a given $\mathbf{x}$. If instead(4.1) is viewed as a function of θ for a given sample $\mathbf{x}$, then it is called a **likelihood function** and is denoted by $L(\theta|\mathbf{x})$. We have the following definition.

Likelihood function

Definition 4.5. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x|\theta)$, $\theta \in \Theta$. Suppose we observe data $\mathbf{x} = (x_1, \ldots, x_n)$. Then, the likelihood function for θ is

$$L(\theta|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|\theta).$$

Intuitively, the likelihood function tells us how likely the observed data are for that value of θ. Therefore it makes sense to estimate θ by that value which makes the observed data appear more likely. Therefore we define the **maximum likelihood estimator** for the parameter θ, the value $\hat{\theta}$ for which $L(\theta|\mathbf{x})$ is maximised, i.e.,

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, L(\theta|\mathbf{x}).$$

In practice, it is usually easier to maximise the logarithm of the likelihood function instead of the likelihood function itself. We define the **log-likelihood function**, $\ell(\theta|\mathbf{x}) = \log L(\theta|\mathbf{x})$. In this case the **maximum likelihood estimator** (MLE) becomes

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, \ell(\theta|\mathbf{x}).$$

Note that θ could be a vector, i.e., $\theta = (\theta_1, \ldots, \theta_\kappa)$. In some cases (but not always) the MLE can be obtained by solving a system of equations

$$\frac{\partial}{\partial \theta_r} \ell(\theta|\mathbf{x}) = 0, \;\; r = 1, \ldots, \kappa.$$

Note

It is custom when writing the log-likelihood function to omit additive constants which do not depend on the parameters. This makes the expression for the log-likelihood brief and does not affect the MLE. It is important however to remain consistent throughout our calculations to avoid errors.

Example 4.8. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exponential}(\mu)$, $\mu > 0$. Then $f(x|\mu) = (1/\mu)\exp(-x/\mu)$ so $L(\mu|\mathbf{x}) = \prod\{(1/\mu)\exp(-x_i/\mu)\} = (1/\mu^n)\exp(-\sum x_i/\mu)$ and $\ell(\mu|\mathbf{x}) = -n\log\mu - \sum x_i/\mu$.

In this case we can find the MLE by solving $\frac{d\ell}{d\mu} = 0$:
$\frac{d\ell}{d\mu} = -n/\hat{\mu} + \sum x_i/\hat{\mu}^2 = 0 \Rightarrow -n + \sum x_i/\hat{\mu} = 0 \Rightarrow \hat{\mu} = \sum x_i/n = \bar{x}$. Note that this is identical to the MoM estimator in Example 4.4.

Example 4.9. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$, i.e., normal with known mean 0 and unknown variance $\sigma^2 > 0$. Then
$L(\sigma^2|\mathbf{x}) = \prod(2\pi\sigma^2)^{-\frac{1}{2}}\exp(-\frac{x_i^2}{2\sigma^2}) = (2\pi\sigma^2)^{-\frac{n}{2}}\exp(-\frac{1}{2\sigma^2}\sum x_i^2)$, so
$\ell(\sigma^2|\mathbf{x}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum x_i^2$.

Again we solve for $\frac{d\ell}{d\sigma^2} = 0$:
$\frac{d\ell}{d\sigma^2} = -\frac{n}{2}\frac{1}{\hat{\sigma}^2} + \frac{1}{2(\hat{\sigma}^2)^2}\sum x_i^2 = 0 \Rightarrow -n + \frac{1}{\hat{\sigma}^2}\sum x_i^2 = 0 \Rightarrow \hat{\sigma}^2 = \sum x_i^2/n$. Note that this is identical to the MoM estimator in Example 4.5.

Example 4.10. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{U}(0, \theta)$, i.e., uniform with known lower bound 0 and unknown upper bound $\theta > 0$. The pdf is $f(x|\theta) = \theta^{-1}$ for $x \in (0, \theta)$, i.e.,

$$f(x|\theta) = \begin{cases} \theta^{-1} & 0 < x < \theta, \\ 0 & \text{otherwise.} \end{cases}$$

Then,

$$\begin{aligned} L(\theta|\mathbf{x}) &= \prod_{i=1}^{n} f(x_i|\theta) \\ &= \begin{cases} \theta^{-n} & 0 < x_1, \ldots, x_n < \theta \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} \theta^{-n} & 0 < x_{(n)} < \theta \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} 0 & \text{if } \theta < x_{(n)}, \\ \theta^{-n} & \text{if } \theta \geq x_{(n)}, \end{cases} \end{aligned}$$

where $x_{(n)} = \max\{x_1, \ldots, x_n\}$. In other words, the likelihood is 0 if at least one of the x_i's falls outside the interval $(0, \theta)$. If all of the x_i's fall within $(0, \theta)$, then the likelihood is θ^{-n}. The statement "all of the x_i's fall within $(0, \theta)$" is equivalent to "the largest of the x_i's falls within $(0, \theta)$". Because θ^{-n} is a decreasing function in θ, the likelihood is then maximised when $\hat{\theta} = x_{(n)}$. This can be verified by the plot in Figure 4.3

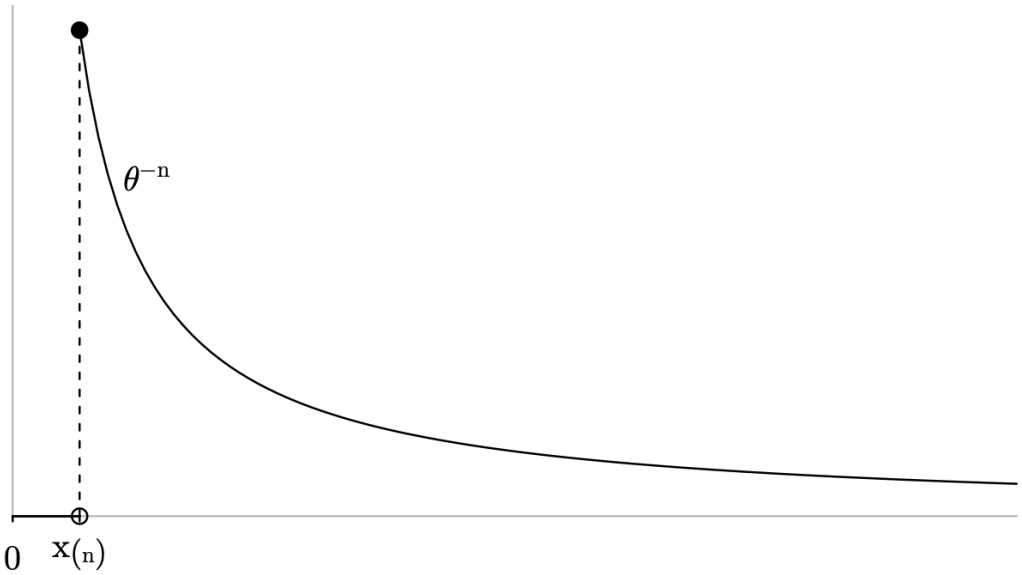


Figure 4.3: Demonstration of the MLE for Example 4.10

Example 4.11. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \beta)$, i.e., gamma with shape $\alpha > 0$ and rate $\beta > 0$. In this case there are two parameters to estimate, i.e., $\kappa = 2$. The pdf is given by

$$f(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}.$$

Then,

$$L(\alpha, \beta|\mathbf{x}) = \frac{\beta^{n\alpha}}{\Gamma(\alpha)^n} \left(\prod x_i\right)^{\alpha-1} e^{-\beta \sum x_i},$$

so

$$\ell(\alpha, \beta|\mathbf{x}) = n\alpha \log \beta - n \log \Gamma(\alpha) + (\alpha - 1) \log \left(\prod x_i\right) - \beta \sum x_i.$$

In this case we have a system of 2 equations: $\frac{d\ell}{d\alpha} = 0$ and $\frac{d\ell}{d\beta} = 0$:
$\frac{d\ell}{d\beta} = \frac{n\alpha}{\beta} - \sum x_i = 0$, and
$\frac{d\ell}{d\alpha} = n \log \beta - n\psi(\alpha) + \log\left(\prod x_i\right) = 0$, where $\psi(\alpha)$ denotes the *digamma function*, $\psi(\alpha) = \frac{d}{d\alpha} \log \Gamma(\alpha)$.

From the first equation we have $\beta = \alpha/\bar{x}$ so if α were known we could estimate β in this way. If α were unknown, we could substitute the expression for β into the second equation to get an equation in terms of α only. This becomes
$\log \alpha - \log \bar{x} - \psi(\alpha) + \sum \log x_i/n = 0$ which does not have a closed form solution. In this case the MLE for α can be obtained numerically. Once the solution is computed, say $\hat{\alpha}$, then it is plugged in the expression for β to get $\hat{\beta} = \hat{\alpha}/\bar{x}$.

4.2 Connection with decision theory

Parameter estimation can be put into a decision-theory framework, where the decision problem becomes estimating the unknown parameter. In this case, the action space $\mathcal{A} = \Theta$, i.e., the available actions are the possible values of the parameter and the action we take is the estimate of the parameter. Estimators, $T(\boldsymbol{x})$, map the data to an estimate, so an estimator is a type of decision rule.

Consider the squared-error loss, $L(\theta, a) = (\theta - a)^2$. Under this loss, the further a is from θ, the higher the loss. The risk associated with the estimator T under the squared-error loss is $R(\theta, T) = E[L(\theta, T(\boldsymbol{x}))] = E[(\theta - T(\boldsymbol{x}))^2] = \text{MSE}_\theta(T)$. This result provides an alternative interpretation of the mean squared error, as the risk of an estimator for θ under squared-error loss.

4.3 Exercises

P 1. Consider the method of moments estimator of Example 4.6. Identify potential drawbacks of this estimator.

2. Verify the formula from the mgf of the gamma distribution in Example 4.7 and use it to derive its mean and variance.

3. Explain why the MLE of Example 4.10 is biased but do not derive its bias.

4. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$, $\theta \in (0, 1)$. Derive the MLE for θ.

5 Confidence Intervals and Hypothesis Testing

Example 5.1. "Paul" was an octopus who achieved worldwide fame after consistently making correct predictions of the outcomes of the 2008 UEFA Euro and the 2010 FIFA World Cup football matches. Before a football match, Paul was offered food in two different boxes, where one box was decorated with the flag of one of the playing teams and the second box with the flag of the other team. Whichever box Paul chose to eat from, was considered his prediction for the match. Paul correctly predicted the winner in 4 out of 6 games of the 2008 Euro and in 8 out of 8 games of the World Cup, including the semifinal and final games. Overall Paul correctly predicted the winner in 12 out of 14 matches!

If Paul was choosing the winner without any prejudice, i.e., each box was chosen with probability $p = 0.5$, then the probability of 12 or more correct predictions in 14 matches is less than 0.7%, which is calculated as $\mathbb{P}(X \geq 12)$ where $X \sim \text{Bin}(14, 0.5)$. Such a rare phenomenon can lead one to think that perhaps Paul was not choosing randomly after all!

5.1 Confidence intervals

Consider a random sample $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x|\theta)$. An estimator $T = T(X_1, \ldots, X_n)$ of the parameter θ, whatever its properties, will provide only a point estimate, $\hat{\theta}$ which is likely to differ from the true value of θ. The point estimator does not provide any information about the deviation of our estimator from the true parameter value. Ideally we would like to provide a range of values which we believe to contain the true parameter value with some known probability. This range of values is called a **confidence interval** and the probability that the interval contains the parameter is called the **confidence level**.

Confidence interval

Definition 5.1. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x|\theta)$, $\theta \in \Theta$. The random interval $[L, U]$ with bounds the statistics $L = L(X_1, \ldots, X_n)$ and $U = U(X_1, \ldots, X_n)$ such that $L \leq U$ and $L, U \in \Theta$ is called a **confidence interval** for the parameter θ. If

$$\mathbb{P}(L \leq \theta \leq U) = 1 - \alpha,$$

for all $\theta \in \Theta$, the number $1 - \alpha$, $\alpha \in (0, 1)$ is called the **confidence level** of the interval.

Typical choices for the confidence level are 90%, 95%, or 99%, which correspond to α being 10%, 5%, and 1% respectively.

> **Note**
>
> Although in Definition 5.1 we define the confidence interval as a closed interval $[L, U]$, it will sometimes be more natural to quote the open interval (L, U) when the random variables L and U are continuous and $L < U$.

A useful quantity for deriving confidence intervals is defined next.

Pivot quantity

Definition 5.2. Let $\mathbf{X} = \{X_1, \ldots, X_n\}$ be a random sample for a population depending on an unknown parameter θ, and $T = T(X_1, \ldots, X_n)$ is some function of the sample. The random variable $Y = g(T, \theta)$, which is a function of T and θ, is called a **pivot quantity** if its distribution does not depend on θ.

A general procedure for constructing a confidence interval for a given confidence level $1 - \alpha$, which is applicable in many problems can be summarised in the following steps.

1. Derive a point estimator $T = T(X_1, \ldots, X_n)$ of the parameter θ and come up with a pivot quantity

   $$Y = g(T, \theta)$$

   of T and θ whose distribution does not depend of θ.

2. Using the distribution of Y, derive two quantiles, c_1 and c_2 with $c_1 \leq c_2$ such that

   $$\mathbb{P}(c_1 \leq Y \leq c_2) = 1 - \alpha,$$

   i.e., the probability that Y falls within c_1 and c_2 is $1-\alpha$, or equivalently, the probability that Y falls outside c_1 and c_2 is α, i.e.,

   $$\mathbb{P}(Y < c_1) + \mathbb{P}(Y > c_2) = \alpha.$$

   Note that the choice of c_1 and c_2 is not unique. We usually choose them so that

   $$\mathbb{P}(Y < c_1) = \mathbb{P}(Y > c_2) = \alpha/2.$$

3. Rearrange the inequality $c_1 \leq g(T, \theta) \leq c_2$ so that it has the form $L \leq \theta \leq U$, where $L = L(X_1, \ldots, X_n)$ and $U = U(X_1, \ldots, X_n)$ do not depend on θ but do depend on c_1 and c_2, and θ is only in the middle. Then,

   $$\mathbb{P}(L \leq \theta \leq U) = \mathbb{P}(c_1 \leq Y \leq c_2) = 1 - \alpha,$$

   so $[L, U]$ is a confidence interval for θ with significance level $1 - \alpha$.

Example 5.2. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exponential}(\mu)$, $\mu > 0$. As shown in Example 4.8, the MLE for μ is $\bar{X}$. To get a confidence interval, note that $\bar{X} = \sum X_i/n$ and the distribution of $W = n\bar{X}$ is Gamma$(n, 1/\mu)$, i.e., with shape n and scale μ, so $Y = n\bar{X}/\mu \sim$ Gamma$(n, 1)$. Here the parametrisation of the Gamma is the same as in Example 4.7, that is Gamma(α, β) where α is the shape parameter and β is the rate parameter.

For a given significance level $1-\alpha$, let c_1 and c_2 be the $\alpha/2$ and $1-\alpha/2$ quantiles of Gamma$(n, 1)$ respectively which can be obtained in Python using
`scipy.stats.gamma.ppf([alpha/2,1-alpha/2],a=n,scale=1)`.

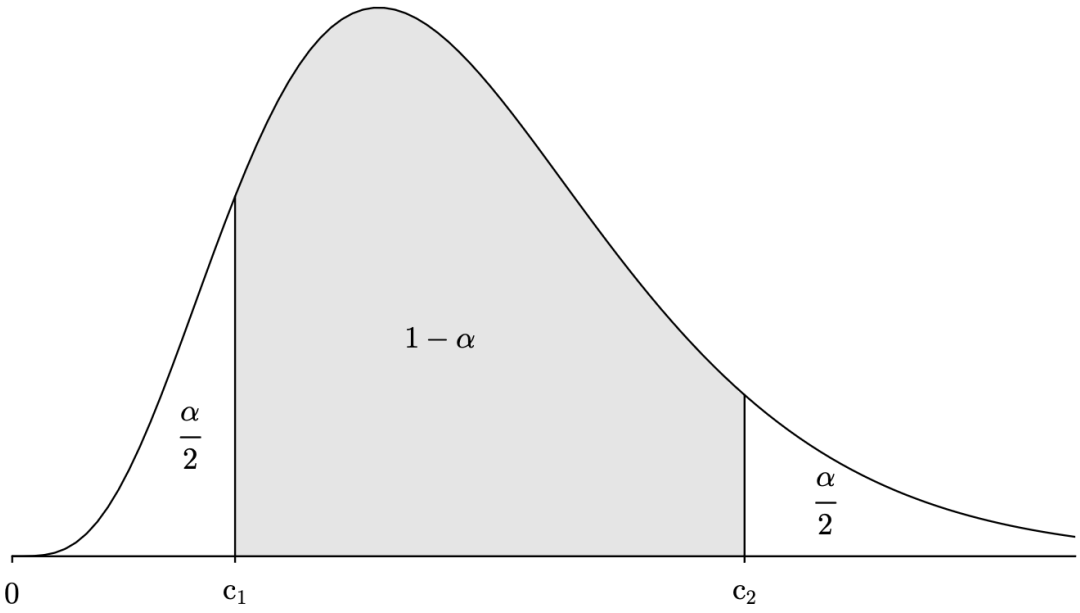


Figure 5.1: Illustration of the choice for c_1 and c_2 for Example 5.2

Then, $c_1 \leq \dfrac{n\bar{X}}{\mu} \leq c_2 \Rightarrow \dfrac{1}{c_2} \leq \dfrac{\mu}{n\bar{X}} \leq \dfrac{1}{c_1} \Rightarrow \dfrac{n\bar{X}}{c_2} \leq \mu \leq \dfrac{n\bar{X}}{c_1} \Rightarrow \left[\dfrac{n\bar{X}}{c_2}, \dfrac{n\bar{X}}{c_1}\right]$ is the confidence interval for μ.

Suppose that we observe the following sample

$$0.02, 0.11, 0.11, 0.26, 0.28, 0.44, 0.81, 0.93.$$

Then, $n = 8$, and $\bar{x} = 0.37$. For a 95% confidence interval, using Python, we find:

```
import scipy.stats
scipy.stats.gamma.ppf([0.025,0.975],a=8,scale=1)
```

```
array([ 3.45383218, 14.42267536])
```

so $c_1 = 3.45$ and $c_2 = 14.42$. We then calculate

$$L = \frac{n\bar{x}}{c_2} = \frac{8(0.37)}{14.42} = 0.21$$
$$U = \frac{n\bar{x}}{c_1} = \frac{8(0.37)}{3.45} = 0.86$$

so the 95% confidence interval is $[0.21, 0.86]$.

Example 5.3. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, 1)$, i.e., the normal distribution with unknown mean μ and known variance $\sigma^2 = 1$. We wish to derive a confidence interval for μ.

An estimator for μ is $\bar{X}$. The distribution of $\bar{X}$ is $\bar{X} \sim \mathrm{N}(\mu, \sigma^2/n)$ so

$$Y = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0, 1),$$

which is a pivot quantity. Let z_p denote the argument in the CDF of the $\mathrm{N}(0, 1)$ distribution, $\Phi(z)$ such that $\Phi(z_p) = p$ (see Figure 5.2). This can by obtained using `scipy.stats.norm.ppf` in Python, or using the standard normal distribution table. Note that, because of the symmetry of the standard normal distribution around 0, $z_p = -z_{1-p}$.

If we let $c_1 = z_{\alpha/2} = -z_{1-\alpha/2}$, $c_2 = z_{1-\alpha/2}$, then, with probability $1 - \alpha$,

$$\begin{aligned} & -z_{1-\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2} \\ \Rightarrow\ & -\bar{X} - z_{1-\alpha/2} \times \sigma/\sqrt{n} \leq -\mu \leq -\bar{X} + z_{1-\alpha/2} \times \sigma/\sqrt{n} \\ \Rightarrow\ & \bar{X} - z_{1-\alpha/2} \times \sigma/\sqrt{n} \leq \mu \leq \bar{X} + z_{1-\alpha/2} \times \sigma/\sqrt{n}. \end{aligned}$$

So,

$$\left[\bar{X} - z_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}}\right] \tag{5.1}$$

is a level $1 - \alpha$ confidence interval for μ when σ is known.

Suppose that we observe the sample

$$-1.90, -0.89, -0.87, -0.65, -0.32, -0.25, 0.90, 1.00, 1.18$$

Then, $n = 9$ and $\bar{x} = -0.2$ and recall we are assuming $\sigma = 1$. For a 95% confidence interval, we find, using the standard normal distribution table (z table) that $z_{0.975} = 1.96$ (see Figure 5.3). Then,

$$L = \bar{x} - z_{0.975}\frac{\sigma}{\sqrt{n}} = -0.2 - (1.96)\frac{1}{3} = -0.85$$
$$U = \bar{x} + z_{0.975}\frac{\sigma}{\sqrt{n}} = -0.2 + (1.96)\frac{1}{3} = 0.45.$$

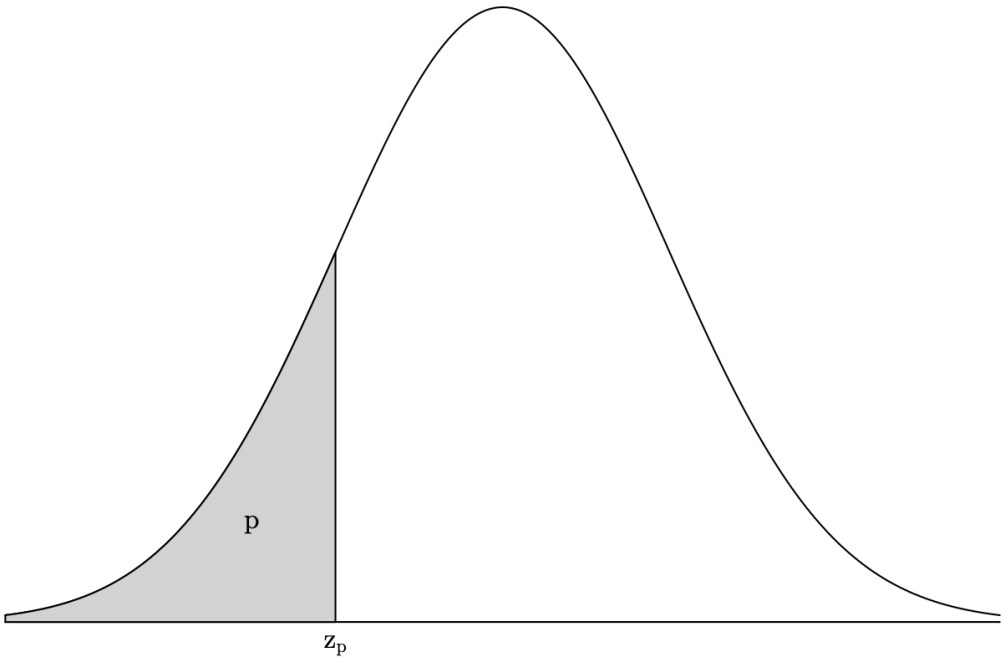
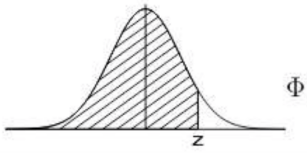


Figure 5.2: Illustration of the standard normal quantile z_p corresponding to left-tail probability p. The plotted curve corresponds to the N(0,1) pdf and for given p, z_p satisfies $p = \Phi(z_p)$ where $\Phi(z)$ is the CDF of N(0,1).

So, a 95% confidence interval for μ is $[-0.85, 0.45]$.

2 The Normal Distribution Function $\Phi(z)$



$$\Phi(z) = P(Z < z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\, dt$$

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	[illegible]	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	[illegible]	0.9808	0.9812	0.9817

Figure 5.3: Calculation of standard normal distribution quantiles. The circled number shows that $P(X < 1.96) = 0.975$ when $Z \sim N(0, 1)$ i.e., $\Phi(1.96) = 0.975$ so $z_p = 1.96$

Note

The lower and upper bounds of the confidence interval depend on the data as well as the desired significance level. The latter dependence is through the numbers c_1 and c_2. To make this dependence explicit, we can write $L(\mathbf{X}, \alpha)$ and $U(\mathbf{X}, \alpha)$ for the lower and upper bounds respectively. The width of the confidence interval is affected by the variation in the population, the sample size, and the desired confidence level.

1. The population variance, σ^2, is a measure of how different the members of the population are. If the population variance is large, then the variation within our sample will also be large, so the confidence interval will be wider.

2. If we take a large sample, i.e. if n is large, then the sample is more representative of the population, so we reduce the variability in the sample. Therefore, the confidence interval will be narrower.

3. If we decrease α, i.e., if we desire a higher confidence level for our confidence interval, then the confidence interval should become wider.

Suppose now that σ^2 is unknown. Recall that

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

can be used as its estimator and

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

(see Example 2.4). We consider the following statistic

$$Y = \frac{\bar{X} - \mu}{S/\sqrt{n}},$$

i.e., the same as before but with σ replaced by $S = \sqrt{S^2}$. To derive the distribution of Y, we use the following definition.

Student's t distribution

Definition 5.3. Let $Z \sim \mathrm{N}(0,1)$ and let $W \sim \mathcal{X}^2_\nu$ and suppose that Z and W are independent. Then the distribution of the random variable

$$Y = \frac{Z}{\sqrt{W/\nu}}$$

is called the Student's t distribution with ν degrees of freedom, written as t_ν.

i Note

The t_ν distribution has similar shape as the $\mathrm{N}(0,1)$ distribution. In particular it is symmetric around 0 and converges to $\mathrm{N}(0,1)$ as $\nu \to \infty$. This is demonstrated in Figure 5.4.

In Python this distribution is given by `scipy.stats.t`.

We can apply this definition in our problem. We know that $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$ and that $W = (n-1)S^2/\sigma^2 \sim \mathcal{X}^2_{n-1}$ and that they are independent, so

$$\begin{aligned} Y &= \frac{Z}{\sqrt{W/(n-1)}} \\ &= \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{S^2/\sigma^2}}\,, \quad \text{(note } \sigma \text{ cancels out)} \\ &= \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim \mathrm{t}_{n-1}. \end{aligned}$$

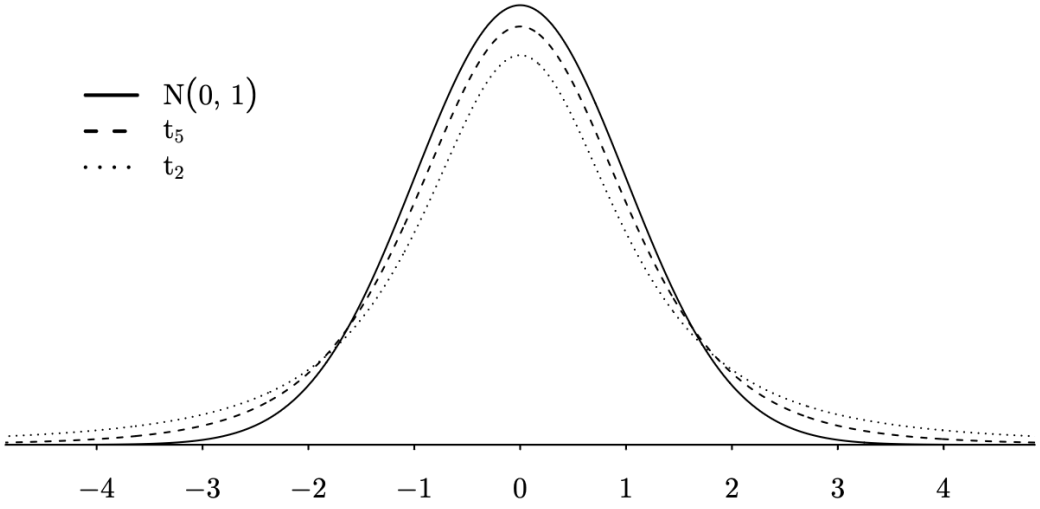


Figure 5.4: Density curves of the standard normal and t_ν distributions with degrees of freedom $\nu = 2, 5$.

Proceeding similarly with the known-variance case, we let

$$c_1 = t_{n-1;\alpha/2} = -t_{n-1;1-\alpha/2}$$

and

$$c_2 = t_{n-1;1-\alpha/2}$$

i.e., the $\alpha/2$ and $1-\alpha/2$ quantiles of t_{n-1}, then

$$\bar{X} - t_{n-1;1-\alpha/2} \times \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1;1-\alpha/2} \times \frac{S}{\sqrt{n}}$$

and therefore

$$\left[\bar{X} - t_{n-1;1-\alpha/2} \times \frac{S}{\sqrt{n}},\ \bar{X} + t_{n-1;1-\alpha/2} \times \frac{S}{\sqrt{n}}\right]$$

is a level $1-\alpha$ confidence interval for μ when σ is unknown. Compare with the confidence for μ when σ is known in Equation 5.1.

5.2 Hypothesis testing

In this section we discuss how, by using data, we can prove statements about the parameters of interest.

Consider for instance the following scenario. Train companies regularly collect passenger data on customer satisfaction. One may ask whether the frequent ticket price increases cause any drop in average customer satisfaction. The population of interest is the train passengers and the parameter of interest is the average customer satisfaction. We are interested in assessing whether the ticket price increase has an impact on the average customer satisfaction.

Statistical hypothesis

Definition 5.4. A **statistical hypothesis** is a statement about the parameter value of the population under study.

Test of significance

Definition 5.5. The **test of significance** is a rule, based on data, for deciding which hypothesis is true between two competing hypotheses:

- the *null hypothesis* (denoted by H_0), and
- the *alternative hypothesis* (denoted by H_1).

The **null hypothesis** corresponds to the common belief about the parameter in question. It is interpreted as *no change* in the value of the parameter (the *status quo*).
The **alternative hypothesis** corresponds to a new claim which we wish to prove. It is interpreted as *a change* in the value of the parameter.
The outcome of a test of significance is the decision whether to reject or not the null hypothesis.

Example 5.4. A company is selling bathroom toiletries and cosmetics. Their daily sales is a normally distributed random variable $N(\mu, \sigma^2)$ with mean $\mu = 2000$ products. They believe that their plan of giving a free sample of one type of their products when you buy another of their products will increase their average daily sales by 100.

- What are the null and alternative hypotheses?

  The claim is that with the offer the average daily sales will become 2100 (increase by 100). This statement determines the alternative hypothesis (a new claim). The null hypothesis corresponds to no change in the average daily sales, i.e. they will remain at 2000. Therefore, the two hypotheses are: H_0*:* $\mu = 2000$ *and* H_1*:* $\mu = 2100$*.*

- If the claim was that the new offer will increase the sales (without saying by how much), what would the two hypotheses be?

  In this case the claim is that the average daily sales will be some number $\mu > 2000$ *while the null hypothesis is as before. Then* H_0*:* $\mu = 2000$ *and* H_1*:* $\mu > 2000$*.*

- If the claim was that the new offer will have some impact on the sales, what would the two hypotheses be?

  In this case the claim is that the average daily sales will be some number different from 2000 *while the null hypothesis is as before. Then* H_0*:* $\mu = 2000$ *and* H_1*:* $\mu \neq 2000$*.*

- What can a statistician do to confirm the claim that the new offer improves sales?

  Suppose we want to compare H_0*:* $\mu = 2000$ *vs* H_1*:* $\mu > 2000$*. The company may consider the following experiment. Introduce the offer for a period of time, say* $n = 30$ *days, and record the number of sales on each day. Let* $x_1, \ldots, x_n$ *be the number of sales for each day, i.e. the sample, and let* $\bar{x}$ *be the sample average. If* $\bar{x}$ *turns out to be significantly larger than 2000, then there is evidence that the sales have improved.*

  There is no easy answer to what "significantly larger" means. That's a matter of personal opinion. For some people if $\bar{x} > 2010$ is enough to indicate increase in average daily sales but others may require $\bar{x} > 2100$. If we set this "critical value" for $\bar{x}$ too low, say 2010, then there is the danger of a false positive conclusion, i.e. claiming that the average sales increased while in reality they stayed the same and the fact that $\bar{x} > 2010$ was just due to variability in the sample. On the other hand, if the critical value is set to a large number, say 2100, then there is the possibility of a false negative conclusion, i.e. claiming that the average daily sales haven't increased when in reality they did but not by that much as to bring $\bar{x}$ to exceed 2100.

In order to draw a conclusion in the example above, we had to summarise the data into one number, in that case $\bar{x}$, and compare this number against a critical boundary and depending on whether $\bar{x}$ exceeded or not this critical boundary then we decide which hypothesis to accept. This brings us to the concept of the *test statistic* and its *critical value*.

Test statistic and critical value

Definition 5.6. The **test statistic** associated with a hypothesis test is the statistic, i.e. a number derived from the sample (see Definition 2.2), which is used to make a decision in a hypothesis test. The value of the test statistic is compared against some predetermined number called the **critical value** of the statistic. If the value of the test statistic exceeds the critical value then our decision is to reject the H_0.

As with any decision we make, we may reach the wrong conclusion. These are commonly referred to as "false positive" and "false negative" conclusions but in the language of statistics they are called *Type I error* and *Type II error*.

Note

Type I error Means that our decision is to reject H_0 when in reality the H_0 is true. We can also say we accept H_1 when in reality H_0 is true.

Type II error Means that our decision is to accept H_0 when in reality the H_0 is false. We can also say we accept H_0 when in reality H_1 is true.

We would like to minimise the probability of reaching the wrong decision or maximise the probability of reaching the correct decision. Therefore for every hypothesis test we need to know the probabilities $\mathbb{P}$(Type I Error) and $\mathbb{P}$(Type II Error). We define

Power of a test

Definition 5.7. For every hypothesis test, we define the **power** to be

$$\text{power} = 1 - \mathbb{P}(\text{Type II Error}) = 1 - P(\text{Accept } H_0 | H_1 \text{ true}) = P(\text{Accept } H_1 | H_1 \text{ true}).$$

The power can be interpreted as the probability of correctly rejecting H_0 or correcyly accepting H_1. So we want to have a test with high power. When the alternative hypothesis is a range of values, the power can be defined for all those values, so, in this case, it is represented by a function on θ.

Example 5.5. In the context of Example 5.4, suppose that the standard deviation is known to be $\sigma = 300$ and we want to test

$$H_0 : \mu = 2000 \quad \text{vs.} \quad H_1 : \mu > 2000$$

In a sample of $n = 30$ days we decide to use a critical value of 2070, i.e., we

$$\text{reject the } H_0 \text{ if } \bar{X} > 2070$$

- What is the probability of Type I error?

  By the definition of Type I error, the probability is $\mathbb{P}(\textit{Type I error}) = \mathbb{P}(\textit{Reject } H_0 | H_0 \textit{ is true})$.

  The statement "Reject H_0" is equivalent to $\bar{X} > 2070$ and the statement "H_0 is true" is equivalent to $\mu = 2000$. Therefore, $\mathbb{P}(\text{Type I error}) = \mathbb{P}(\bar{X} > 2070 | \mu = 2000)$.

  In order to compute this probability, we need to know the distribution of the random variable inside the parentheses, namely $\bar{X}$. This distribution is:

  $$\bar{X} \sim \text{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

  that is, the normal distribution with mean μ and variance σ^2/n. In this case $\mu = 2000$ and $\sigma^2/n = 300^2/30$. Then we have

  $$\mathbb{P}(\bar{X} > 2070 | \mu = 2000) = 1 - \mathbb{P}(\bar{X} \leq 2070 | \mu = 2000).$$

and

$$\mathbb{P}(\bar{X} \leq 2070|\mu = 2000) = \mathbb{P}\left(\frac{\bar{X} - 2000}{300/\sqrt{n}} \leq \frac{2070 - 2000}{300/\sqrt{30}}\right) = \Phi\left(\frac{2070 - 2000}{300/\sqrt{30}}\right)$$

recall Definition 1.20. Then, $z = \frac{2070-2000}{300/\sqrt{30}} = 1.28$, which corresponds to probability $\Phi(1.28) = 0.8997$. Therefore,

$$\mathbb{P}(\text{Type I error}) = 1 - 0.8997 = \mathbf{0.1003}\,.$$

- What is the probability of Type II error if the true mean is 2100?

  By the definition of Type II error, the probability is $\mathbb{P}(\textit{Type II error}) = \mathbb{P}(\textit{Accept } H_0|H_0 \textit{ is false})$.

  As before, the statement "Accept H_0" is equivalent to $\bar{X} \leq 2070$ and the statement "H_0 is false" in this case is equivalent to $\mu = 2100$. Therefore, $\mathbb{P}(\text{Type I error}) = \mathbb{P}(\bar{X} \leq 2070|\mu = 2100)$. The distribution of $\bar{X}$ in this case is normal with mean $\mu = 2100$ and variance $\sigma^2/n = 300^2/30$.

  Then, $z = \frac{2070-2100}{300/\sqrt{30}} = -0.55$, which corresponds to probability $\Phi(-0.55) = 0.2912$. Therefore,
  $\mathbb{P}(\text{Type II error}) = \mathbf{0.2912}$ and the corresponding power $= 1 - 0.2912 = \mathbf{0.7088}$. Note that if we assume a different value for μ under H_1, we will get a different probability of Type II error and therefore a different power.

- What is the power function?

  Take any μ such that $\mu > 2000$. Then, as before $z = \frac{2070-\mu}{300/\sqrt{30}}$, so, the power function is given by $\text{power}(\mu) = 1 - \Phi(\frac{2070-\mu}{300/\sqrt{30}}) = \Phi(\frac{\mu-2070}{300/\sqrt{30}})$, because of the property $\Phi(z) = 1 - \Phi(z)$ which is derived from the symmetry of the pdf of the standard normal distribution.

  This function is plotted in Figure 5.5. We observe that the further μ is from 2000, the higher the power. This can be interpreted as being more likely to reject H_0 correctly when the true mean is further from 2000.

In this example, $\bar{X}$ is the test statistic and the number 2070 is the critical value c. The probabilities of Type I and Type II errors can be illustrated in Figure 5.6. The bell curve on the left is the distribution of $\bar{X}$ under H_0 and the dark grey area is the probability of Type I error. Similarly, the bell curve on the right is the distribution of $\bar{X}$ under H_1 when the mean is 2100 and the light grey area is the probability of Type II error. Both probabilities correspond to the critical value $c = 2070$.

Now let's see what happens if the critical value of 2070 changes.

If we increase the critical value, the probability of Type I error (dark grey area) becomes smaller. This means that it becomes *less likely* to reject H_0 erroneously (a false positive).

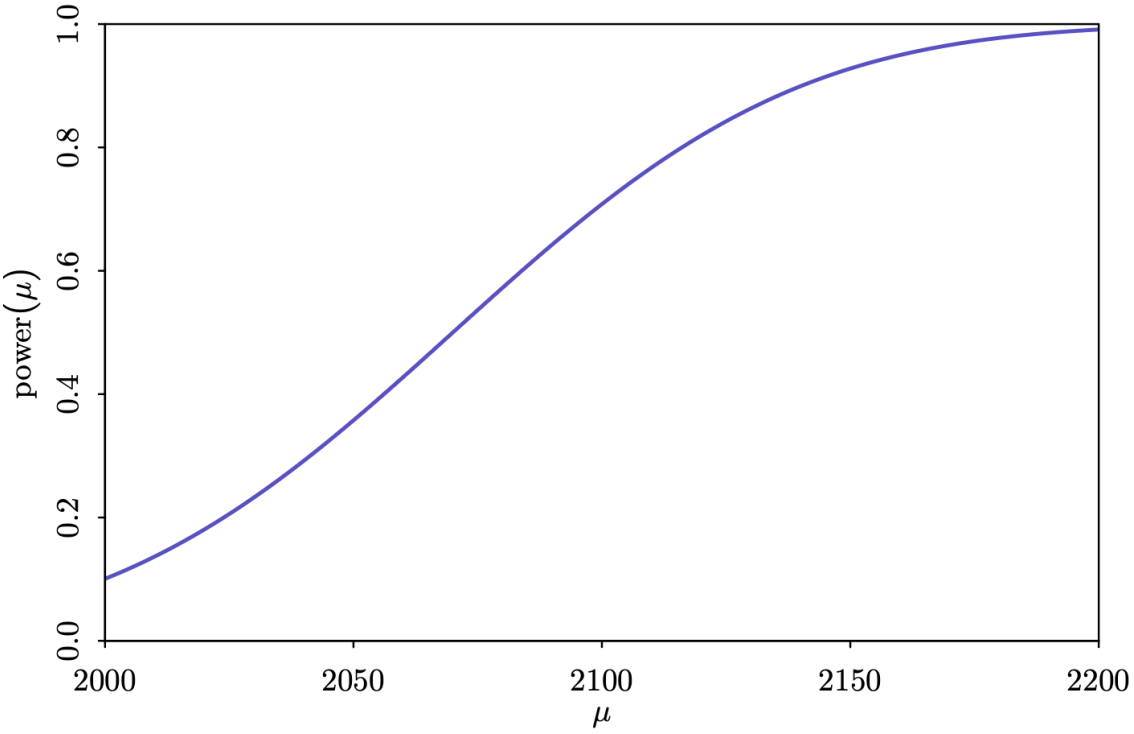


Figure 5.5: The power function for Example 5.5

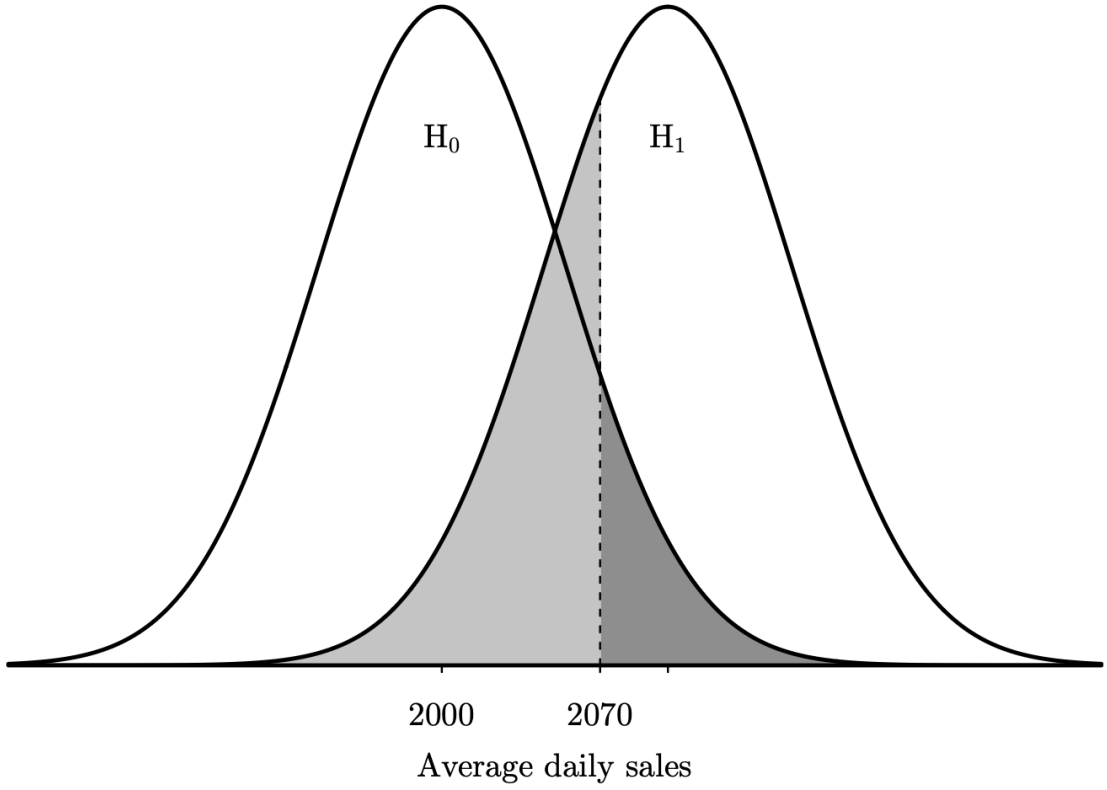


Figure 5.6: Illustration of the probabilities of Type I (dark gray area) and Type II (light gray area) errors for Example 5.5

However, we also increase the probability of Type II error (light grey area) so it becomes *more likely* to accept H_0 when we shouldn't (a false negative). Unfortunately this is a usual impediment when performing hypothesis tests. In practice, we demand the probability of Type I error to be comfortably small typically around 5% which determines the critical value. This concept gives rise to the notion of *significance level* and *p-value*.

Significance level

Definition 5.8. It is desirable to have a rule for rejecting H_0 with small probability of Type I error. In practice this probability is set to a fixed, prescribed, value denoted by α (the greek letter "alpha") called the **level of significance** and a rule with the property $\mathbb{P}(\text{Type I error}) = \alpha$ is sought.
The smaller the value of α, the more evidence needed to reject the H_0. Typical values for α are 1%, 5% and 10%.

Example 5.6. In Example 5.5, we say that the rule "reject H_0 if $\bar{X} > 2070$" has $\mathbb{P}(\text{Type I error}) \approx 10\%$. Therefore, if we want a rule with significance level $\alpha = 10\%$, we need to choose a critical value of $c = 2070$.

Suppose now that we are seeking for a rule of the form "reject H_0 if $\bar{X} > c$", where c must be chosen according to a significance level $\alpha = 5\%$. This means that

$$\begin{aligned}
0.05 &= \mathbb{P}(\text{Type I error}) \\
&= \mathbb{P}(\text{Reject } H_0 \mid H_0 \text{ true}) \\
&= \mathbb{P}(\bar{X} > c \mid \mu = 2000) \\
\Rightarrow 0.95 &= \mathbb{P}(\bar{X} \leq c \mid \mu = 2000) \\
&= \Phi\left(\frac{c - 2000}{300/\sqrt{30}}\right) \\
\Rightarrow z_{0.95} &= \frac{c - 2000}{300/\sqrt{30}} \\
\Rightarrow c &= 2000 + z_{0.95} \times 300/\sqrt{30} \\
&= 2090,
\end{aligned}$$

where we used the fact that $z_{0.95} = 1.645$ because $\Phi(1.645) = 0.95$. Therefore, using a critical value of 2090 produces a test with $\mathbb{P}(\text{Type I error}) = 0.05$, or, in other words, a level of significance $\alpha = 5\%$.

Using $c = 2090$ instead of $c = 2070$ reduces the probability of wrongly reject the null hypothesis by half. What about the probability of correctly rejecting H_0 when the true mean value is $\mu = 2100$?

As in Example 5.5, $\mathbb{P}(\text{Type II error}) = \Phi\left(\dfrac{2090-2100}{300/\sqrt{30}}\right) = \Phi(-0.1826) = 0.4276$ and power $= 1 - 0.4276 = 0.5724$. So as expected, although using the critical value $c = 2090$ compared to $c = 2070$ reduced the probability of Type I error, it increases the probability of Type II error, and consequently, reduces the power of the test.

P-value

Definition 5.9. The ***p*-value** is the smallest significance level which we can set and still be able to reject H_0 with the given data. By definition, the p-value is a probability which depends on the sample we are analysing. If p-value $< \alpha$ then the data provide enough evidence to reject H_0.

The two definitions suggest two paths one can follow to conduct a hypothesis test. In both cases, one the data and a significance level are provided. Then we could either

- Use the data to derive the test statistic and use the significance level to derive the critical value. If the value of the test statistic exceeds the critical value then we reject H_0, otherwise we accept it.
- Use the data to derive the test statistic and from there derive the corresponding p-value. If the p-value is smaller than the significance level then we reject H_0, otherwise we accept it.

These two paths are depicted below.

It shouldn't matter which of the two approaches we take to perform the hypothesis test as they are equivalent, meaning that the lead to the same conclusion regarding the rejection of the null hypothesis.

Example 5.7. Following Example 5.6, suppose we seek to reject the null hypothesis with significant level $\alpha = 5\%$. Then, according to that example, we must choose a critical value of $c = 2090$.

Suppose next that we collect $n = 30$ observations and found that the average sales were $\bar{x} = 2050$. Then, according to our hypothesis test, we do not reject the null hypothesis because $2050 \not> 2090$, so we cannot conclude that the average sales increased.

What is the smallest level of significance α that we could set, such that, the critical value c that corresponds to that level would allow us to reject the null hypothesis? Since $\alpha = 5\%$ does not allow us to reject the null hypothesis, it is clear that we should seek for a higher significance level, as we are more relaxed about wrongly reject the null hypothesis. So we expect to find p-value > 0.05. Indeed, if we choose a significance level p such that the corresponding critical value c falls exactly at $c = 2050$, then p is the p-value. This is illustrated in Figure 1.6.

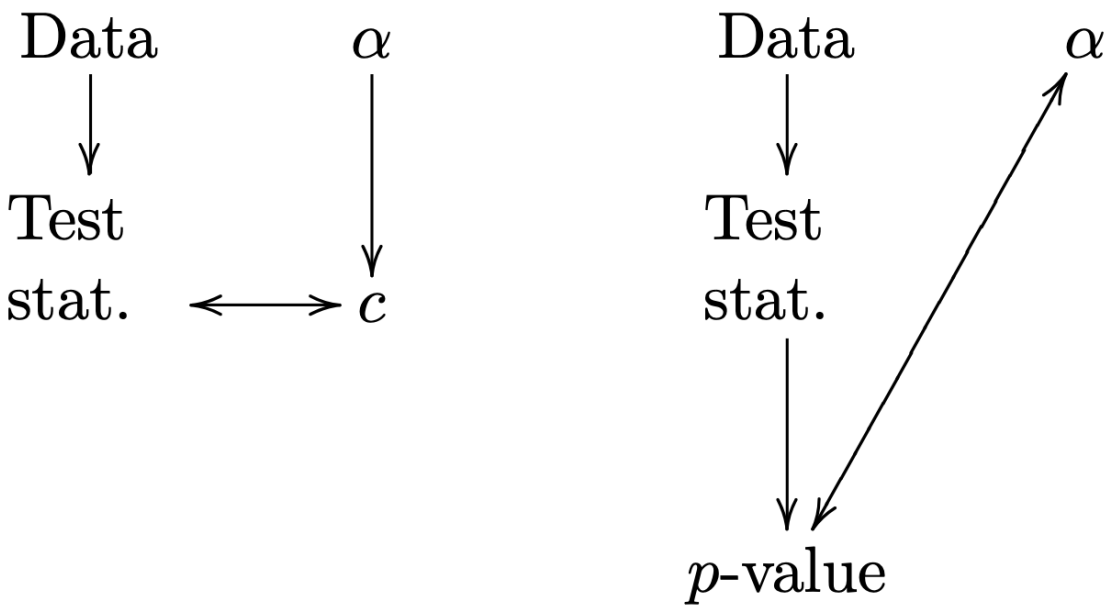


Figure 5.7

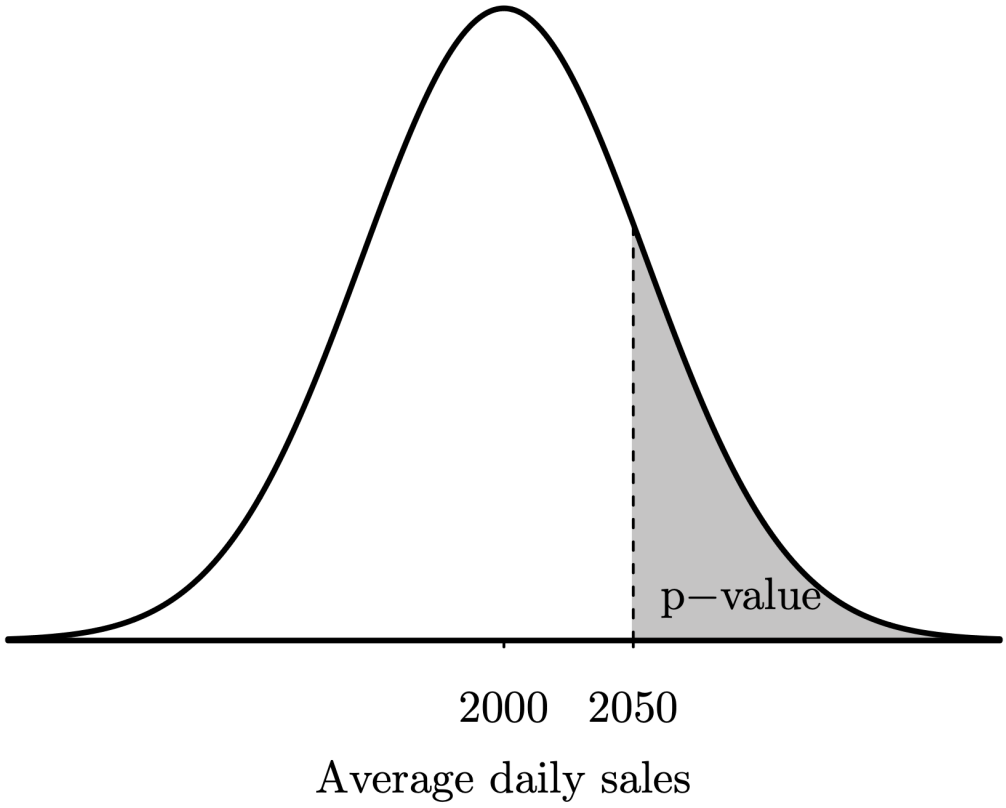


Figure 5.8: Illustration of p-value for Example 5.7

Reversing the calculations in Example 5.6, if p is the p-value, z_{1-p} is the quantile that achieves a critical value of 2050, so

$$\begin{aligned} 2050 &= 2000 + z_{1-p} \times 300/\sqrt{30} \\ \Rightarrow z_{1-p} &= \frac{2050 - 2000}{300/\sqrt{30}} \\ &= 0.9129 \\ \Rightarrow 1 - p &= \Phi(0.9129) \\ \Rightarrow p &= 1 - \Phi(0.9129) = 1 - 0.82 = 0.18. \end{aligned}$$

So p-value $= 0.18$. Accordingly, we can see that, since p-value $> 5\%$, we do not reject the null hypothesis at the 5% level. In fact, p-value $> 10\%$, so we do not reject the null hypothesis at the 10% level either.

5.3 Exercises

Confidence intervals:

1. a. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$ where both μ and σ^2 are unknown parameters. Using the fact that, the sample variance, S^2, is distributed as $(n-1)S^2/\sigma^2 \sim \mathcal{X}^2_{n-1}$, derive a level $1-\alpha$ confidence interval for σ^2. ($\mathcal{X}^2_k$ denotes the *chi-squared distribution with k degrees of freedom*, which is available in python as `scipy.stats.chi2`. It is a special case of the gamma distribution with shape $k/2$ and rate $1/2$.)

   b. Suppose that the following data were observed

   $$-1.90, -0.89, -0.87, -0.65, -0.32, -0.25, 0.90, 1.00, 1.18$$

   For these data $n = 9$, $\bar{x} = -0.2$, and $S^2 = 1.073$. Calculate a 95% confidence interval for σ^2.

2. a. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{U}(0, \theta)$, $\theta > 0$. By considering an appropriate pivot construct a level $1-\alpha$ confidence interval for θ. *Hint.* Let $W_i = X_i/\theta$.

   b. Suppose that the following data were observed

   $$0.90, 1.00, 1.18, 1.90, 2.20.$$

   Calculate a 95% confidence interval for θ. Does the confidence interval contain the maximum likelihood estimator for θ?

Hypothesis testing:

1. The author of a weight-loss diet claims that an average adult, weighting 100 Kg, who follows the proposed diet, will lose 20 Kg after 1 month. What are the null and alternative hypotheses?

2. The author of a weight-loss diet claims that an average adult, weighting 100 Kg, who follows the proposed diet, will lose weight after 1 month. What are the null and alternative hypotheses?

3. The author of a weight-loss diet claims that an average adult, weighting 100 Kg, who follows the proposed diet, will notice a change in their weight after 1 month. What are the null and alternative hypotheses?

4. The author of a weight-loss diet claims that an average adult, weighting 100 Kg, who follows the proposed diet, will lose weight after 1 month. An experiment was conducted to verify this claim. Three adults, who weighted 100 Kg, followed the diet for one month and their weights at the end of the month were recorded. The experimenters would accept the author's claim if the sample mean $\bar{X}$ of the three measured weights is less than 90. Suppose that the population standard deviation is $\sigma = 15$.

    a. The three people's weights after the end of the month were: 82, 86, and 93. What is the experimenters' conclusion?

    b. According to the central limit theorem, what is the asymptotic distribution of the sample mean of $n = 3$ measurements from a population with mean $\mu = 100$ and standard deviation $\sigma = 15$?

    c. Use the central limit theorem to calculate the probability of Type I error of the experimenters' decision rule.

    d. Use the central limit theorem to calculate the probability of Type II error of the experimenters' decision rule assuming that the average weight after one month is 85.

    e. Propose a rule of the form "accept the author's claim if $\bar{X} < c$" (in other words find c) such that the probability of Type I error is 10%.

    f. Suppose that in the sample we find that $\bar{x} = 86$. Find the p-value. What is your conclusion at significance level $\alpha = 5\%$?