Non-calculus-based Introductory Probability and Statistics

Sashwat Tanay

June 24, 2026

Contents

Chapter 1

Overview

1.1 Branches of Statistical Science

In statistical science, we analyze, study, and quantify the uncertain aspects of various phenomena of the natural world. For concreteness, consider a die which, when rolled, produces six outcomes, 1-6. As to what outcome might be produced is uncertain.

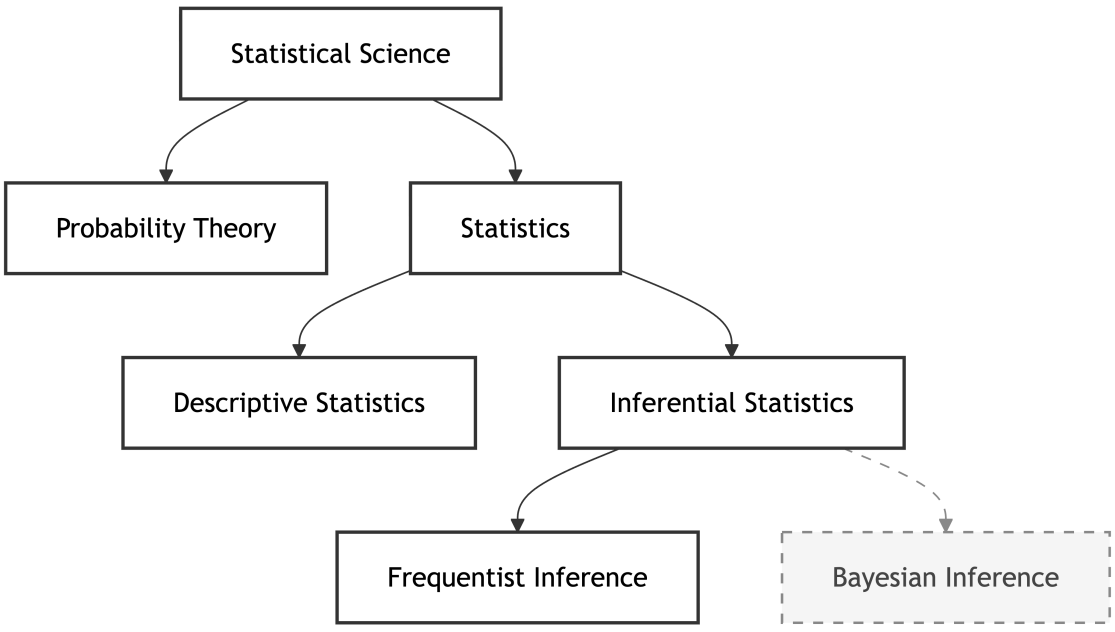

Figure 1.1: Breakdown of the field of statistical science.

Fig. 1.1 displays the lay of the land in the field of statistical science. The field branches out into two sub-fields.

1. Probability: Probability deals with predicting the outcome of an experiment given some prior or assumed knowledge of the model or the mechanism that produces the outputs of the experiment. For example, if I roll a fair die one million times, what fraction of those will result in the number "1".

2. Statistics: Statistics deals with the inverse question. Given the output of an experiment, what can we infer about the underlying model or mechanism that has produced the experimental output? For example, if I roll a die a million times, and 99% of those result

in the number "1", then is the die fair? The relation between probability and statistics is depicted pictorially in Fig. 1.2.

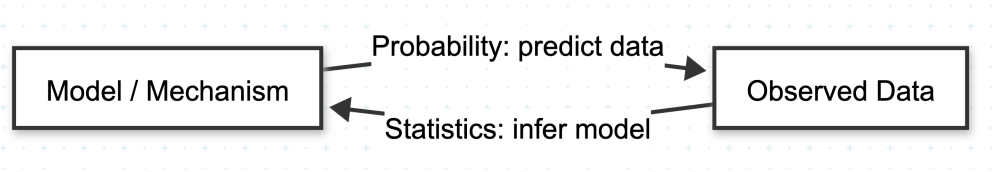

Figure 1.2: Relation between probability and statistics.

Statistics further splits into descriptive and inferential statistics. While descriptive statistics deals with summarizing the main aspects of the data at hand , the tricky inferential questions (like the one above) are dealt with within inferential statistics. Finally, within inferential statistics, there are two philosophically different branches; i.e. Frequentist and Bayesian.

In this course, the only subject that we'll not deal with is the Bayesian branch of inferential statistics, as indicated in Fig. 1.1.

Advanced Material 1.1: Frequentist vs Bayesian Inference: The two approaches are not just philosophically different, but they also find applications in practical scenarios of separate nature. While the frequentist inference is more suitable for experiments which are easily repeatable (rolling a die or trying a new vaccine), Bayesian inference is suited for experiments that are not easily repeatable (testing a new economic theory or political model). The reader is referred to Ref. [4] for Bayesian inference.

1.2 Recommended Reading

While these lecture notes cover the essential curriculum, the following texts are recommended for further study and diverse perspectives on the subject matter:

- **OpenIntro Statistics** by David M. Diez, Christopher D. Barr, and Mine Çetinkaya-Rundel [1].
- **Advanced Engineering Mathematics** by Erwin Kreyszig [2].
- **Probability & Statistics for Engineers & Scientists** by Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers, and Keying Ye [5].
- **Mathematical Methods for Physics and Engineering: A Comprehensive Guide** by K.F. Riley, M.P. Hobson, and S.J. Bence [3].

1.3 Note to the Reader

The boxes labeled as "Advanced Material" contain elements that are outside the scope of the syllabus. They are meant to motivate the students towards more advanced topics across scientific disciplines.

The various Quarto .qmd files linked throughout the lecture notes are intended to be executed on RStudio. They are meant to serve as numerical implementations of the concepts discussed in these notes.

Chapter 2

Descriptive Statistics

2.1 Data

We choose to start operationally by presenting an example data set on some college students in Tab. 2.1. Every observation in the form of a student is formally called a case and corresponds to one of the rows. The columns correspond to the features or characteristics (formally called variables) that the cases carry. The term data matrix refers to the table.

Obs.	Name	Height (cm)	Points (0–100)	Student Type	Satisfaction	Student ID
1	Alice	165.2	88	Domestic	High	345
2	Brian	172.5	74	International	Medium	1057
3	Carla	160.8	91	Domestic	High	34
4	Daniel	180.1	63	International	Low	174
5	Elena	168.4	79	Domestic	Medium	123
6	Farid	175.6	84	International	High	4

Table 2.1: Example dataset on students.

2.2 Classifying variables on the basis of the form of their values

On the basis of the form of the values variables can take, they can be divided into two broad categories (see Fig. 2.1):

1. Numerical: They appear in the form of numbers and arithmetic operations (sum, difference, average) on them are meaningful. They can either be (a) continuous (can take any value on the real number line) or (b) discrete (not continuous)

2. Categorical: Their main function is to act as labels or categories. They are either not in the form of numbers or are in the form of numbers but arithmetic operations on them are not meaningful. They further can be (a) nominal (possessing no natural order) or (b) ordinal (have a natural order).

Again, as an example, the classification of the variables in Tab. 2.1 is as follows:

- Name: nominal categorical
- Height: continuous numerical

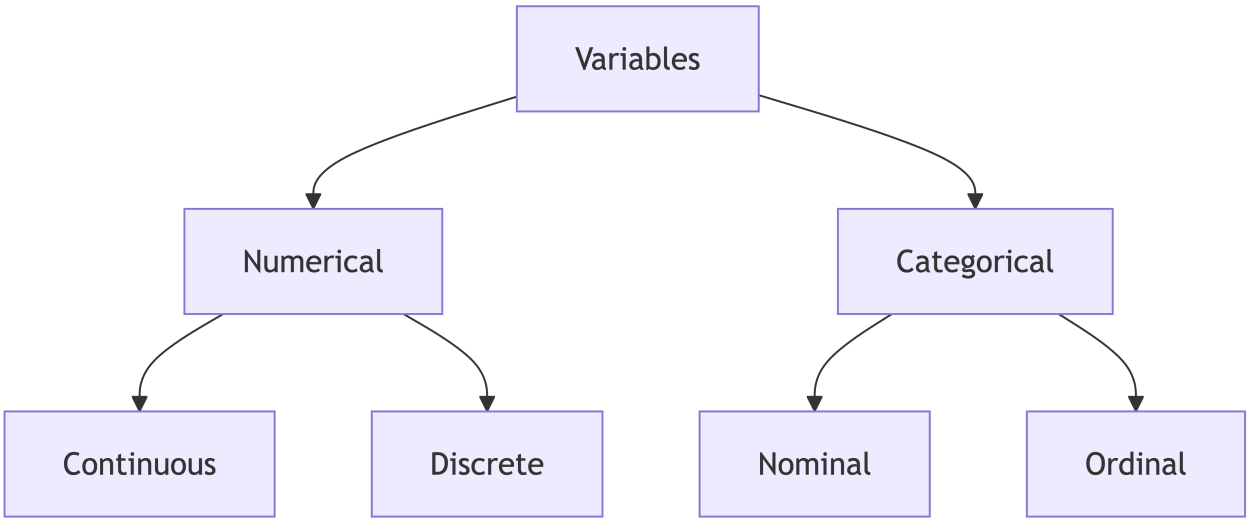

Figure 2.1: Types of Variables

- Points: discrete numerical
- Student type: nominal categorical
- Satisfaction: ordinal categorical
- Student ID: nominal categorical

Important 2.1: Variables manifesting in the form of numbers may not be numerical variables: The variable Student ID is not a numerical variable, although it manifests in the form of numbers. It does not satisfy the definitional criterion for numerical variables; it does so for categorical variables, though.

2.3 Classifying variables on the basis of their mathematical structure

The measurement scale (MS) classification [1] of variables classifies variables on the basis of the mathematical structure of the values taken by a variable; specifically which mathematical relations or operations on those values are meaningful.

It is important to note that the MS categories themselves form a hierarchy. A variable belonging to a higher category contains all the properties of the the variable from a lower category, but possesses additional structure. A variable is classified according to the highest measurement scale whose structure its values satisfy.

In this course, we have 4 MS categories in the increasing order of complexity and hierarchy:

$$\text{Nominal} < \text{Ordinal} < \text{Interval} < \text{Ratio}. \tag{2.1}$$

Let us look at them one by one.

1. Nominal: For these variables, we can decide whether their values belong to the same category or not.

[1] This classification system was introduced by American psychologist Stanley Smith Stevens (1906 – 1973).

Categorical nominal variables are MS nominal. In more abstract notation, if we define C to mean "the category of", then in the context of the Student Type variable of Tab. 2.1 (which is a categorical nominal variable) we can write valid statements like

$$C(\text{Observation 1}) = C(\text{Observation 3}) \neq C(\text{Observation 2}). \tag{2.2}$$

2. Ordinal: Ordinal variables can be classified into categories that have a natural order, but the differences between categories are not meaningful.

Categorical ordinal variables are ordinal in the context of the MS classification. In the context of the Satisfaction variable of Tab. 2.1 (which is a categorical ordinal variable) we can write statements like

$$\text{Low} < \text{Medium} < \text{High}, \tag{2.3}$$

as well as

$$C(\text{Observation 1}) = C(\text{Observation 6}) \neq C(\text{Observation 2}). \tag{2.4}$$

3. Interval: A variable of this type takes numerical values for which mathematical differences between values are quantifiable and meaningful. But the scale does not possess a true zero (representing absence of the quantity). Hence ratios are not meaningful.

Numerical variables (discrete or continuous) can satisfy this criterion, but they don't necessarily have to.

For example, consider the data matrix for three Olympic events.

Observation	Olympic Event	Year
1	Rio de Janeiro Olympics	2016
2	Tokyo Olympics	2020
3	Paris Olympics	2024

Table 2.2: Examples of Olympic events and the years they were held.

In Tab. 2.2, the variable Year belongs to the interval MS. If we want to find how much later the Paris Olympics were held as compared to the Tokyo Olympics, the result is $2024 - 2020$ years $= 4$ years (a difference). However, 2024/2020 (a ratio) does not tell us "how many times later" the Paris Olympics occurred compared to the Tokyo Olympics. This is because the origin of the Gregorian calendar does not represent the beginning of time, and is not a meaningful zero. Hence, subtractions are meaningful but not divisions.

Note that we can still write statements akin to Eqs. 2.3 and 2.4 for variables that belong to the interval MS.

Important 2.2: Zero of interval MS variables: Interval MS variables do possess a zero, but it's not physically meaningful.

4. Ratio: A variable of this type has all the properties of the interval scale but also possesses a meaningful zero representing the absence of the quantity being measured. Hence ratios are meaningful.

Numerical variables (discrete or continuous) can satisfy this criterion, but they don't necessarily have to.

Examples of a variable that belongs to the ratio MS are Height and Points of Tab. 2.1 since for these variables, differences and ratios are meaningful due to the existence of a meaningful zeros for both.

Measurement Scale	Variable Type (based on the form of values)
Nominal	Categorical (nominal)
Ordinal	Categorical (ordinal)
Interval	Numerical (continuous or discrete)
Ratio	Numerical (continuous or discrete)

Table 2.3: Correspondence between measurement scale classification and the classification based on the form of values.

Important 2.3: By definition, a variable cannot be classified as belonging to more than one MS class; recall the earlier description "highest measurement scale whose structure its values satisfy". Hence even though the Height variable of Tab. 2.1 satisfies the criterion mentioned for interval MS variables, it is not an interval MS variable since in addition, it also satisfies the criterion of the ratio MS variable, which is higher in the hierarchy 2.1.

2.4 Graphical representation of data

In this section, we introduce three basic graphical representations of data: histograms, bar plots, and scatter plots. Students should download the dataset (CSV file) and the accompanying Quarto notebook (QMD file). The QMD file contains the lecture text, the R code, and the rendered outputs (tables and plots) produced by executing the code.

2.5 Measures of central tendency

For this chapter, let $x = (x_1, x_2, \ldots, x_n)$ denote values of a certain variable arranged in nondecreasing order:

$$x_1 \leq x_2 \leq \cdots \leq x_n.$$

2.5.1 Definitions

Measures of central tendency are quantities that are invented to give us an idea of typical values in a certain data distribution. Out of numerous such quantities, we will go through the most common ones: mean, median, and mode.

Definition 2.1: Arithmetic mean $\bar{x}$ or $\mu(x)$ of a set of values $x = (x_1, x_2, ..., x_n)$ of a certain variable is defined as[a]

$$\mu(x) \equiv \bar{x} \equiv \frac{x_1 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}. \tag{2.5}$$

[a] Arithmetic mean is sometimes also called just mean or average. Also, the symbol $\equiv$ is used to introduce a definition.

Important 2.4: Definitions, axioms, and theorems: A **definition** serves to introduce a convention, a label, or a name. An example is Definition 2.1.

An **axiom** is a statement that we assume to be true. Usually, it is an encapsulation of routinely familiar natural phenomena. An example is the commutative law of addition: for real numbers $a + b = b + a$. Another one is that we get only one real number when we add two reals[a]. We will encounter more in this course.

Theorems are statements that can be deduced from axioms or other theorems.

Out of the above three kinds of statements, only the last one has a corresponding proof.

[a]These axioms are typically introduced in a course on real analysis

Definition 2.2: Case 1: n is odd. The median is the middle observation in the ordered sequence:

$$\text{med}(x) = x_{(n+1)/2}.$$

Case 2: n is even. The median is the average of the two middle observations:

$$\text{med}(x) = \frac{x_{n/2} + x_{(n/2)+1}}{2}.$$

Definition 2.3: Mode for $x = (x_1, x_2, \ldots, x_n)$ is defined as the value occurring most frequently.

Note that the mode is not unique, and that there could be more than 1 mode for a given dataset x.

2.5.2 Strengths and weaknesses of mean, median, and mode

The reason multiple measures of central tendency have been invented is that they all have their strengths and weaknesses. Below is a summary.

Measure	Strengths	Weaknesses
Mean	Uses all numerical information in the dataset.	Sensitive to extreme values (outliers). Not defined only for non-numerical variables (nominal or ordinal variables).
Median	Robust to extreme values and outliers. Defined for numerical and ordinal variables.	Not meaningful for nominal variables. Does not use all information in the dataset.
Mode	Can be defined for numerical, ordinal, and even nominal variables. Not affected by extreme values.	May not exist or may not be unique.

Table 2.4: Strengths and weaknesses of mean, median, and mode.

2.6 Measures of dispersion

2.6.1 Definitions

Measures of dispersion are invented to quantify the variability in data. We will introduce four such measures: variance, standard deviation, interquartile range, and coefficient of variation.

Definition 2.4: Variance of a set of values $x \equiv (x_1, x_2, \ldots, x_n)$ of a certain variable is defined as

$$\sigma^2(x) \equiv \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2, \tag{2.6}$$

where $\bar{x}$ is the arithmetic mean of the values.

Definition 2.5: Standard deviation of a set of values $x \equiv (x_1, x_2, \ldots, x_n)$ of a certain variable is defined as the square root of the variance:

$$\sigma(x) \equiv \sqrt{\sigma^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{2.7}$$

In short, the variance is a measure of the variability of the square of deviations around the mean, whereas the standard deviation is a measure of the variability of just the deviations.

Important 2.5: What goes wrong if we decide to invent the following simple quantity to measure the variability?

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}), \tag{2.8}$$

For a symmetrical distribution, this quantity becomes zero, as one can easily check.

Definition 2.6 (Quartiles)**:** The quartiles divide the observations into four parts.[a]

- The first quartile Q_1 is the median of the lower half of the data.
- The second quartile Q_2 is the median of the entire dataset.
- The third quartile Q_3 is the median of the upper half of the data.

If the number of observations is odd, the overall median is excluded when forming the lower and upper halves.

[a] There is no universally agreed-upon definition of quartiles. Different statistical software packages (for example the `R` programming language) may compute quartiles using slightly different rules, which can sometimes produce slightly different numerical values. The definition used here is a simple and commonly used one for introductory analysis.

Definition 2.7 (Interquartile Range)**:** The interquartile range (IQR) is a measure of variability defined as

$$\text{IQR} = Q_3 - Q_1.$$

Definition 2.8 (Coefficient of Variation)**:** The coefficient of variation (CV) is defined as

$$\text{CV} \equiv \frac{\sigma}{\mu}. \tag{2.9}$$

The CV measures the variability of a dataset relative to its mean. (unlike the standard deviation, which measures absolute dispersion).

An important feature of the CV is that it is *dimensionless*. Both the numerator and the denominator have the same units, so the units cancel. Consequently, the CV does not depend on the units (cm or feet) in which the variable is measured.

We learned in Sec. 2.3 that ratios between variable values (or mean, for that matter) are not meaningful unless the variable belongs to the ratio MS scale. This is why CV, defined as a ratio is also not meaningful unless the concerned variable belongs to the ratio MS scale.

2.6.2 Strengths and weaknesses of common measures of dispersion

Measure	Strengths	Weaknesses
Standard deviation or variance	Uses the entire data; measures absolute variability.	Sensitive to outliers. Defined only for numerical variables.
IQR	Robust to outliers.	Does not the entire dataset. Not defined for nominal variables.
CV	Dimensionless measure of relative variability. Useful across datasets with different units or different means.	Sensitive to outliers. Meaningful only for numerical variables on a ratio MS scale.

Table 2.5: Strengths and weaknesses of common measures of dispersion.

2.7 Correlation does not imply causation

2.7.1 The catastrophe

Consider the following data collected from several individuals who recently reported feeling unwell. For each person we record the average number of hours they slept during a day and their average body temperature during that same day.

Observation	Sleep Hours	Body Temperature (°C)
1	6	37.0
2	7	37.1
3	8	37.6
4	9	38.0
5	7	37.2
6	8	37.7
7	9	38.3
8	6	37.0
9	8	37.8
10	9	38.2

Table 2.6: Sleep hours and body temperature for several individuals.

The scatter plot corresponding to the data in Tab. 2.6 is shown in Fig. 2.2. The plot clearly highlights the correlation[2] between these two variables.

At first glance, the pattern in Fig. 2.2 might tempt one to conclude that increased sleep causes an increase in body temperature. In other words, one might hastily infer a causal relationship in which the number of hours slept determines the body temperature of an individual (indicated in Fig. 2.3).

[2]In this course we do not give a formal definition of correlation.

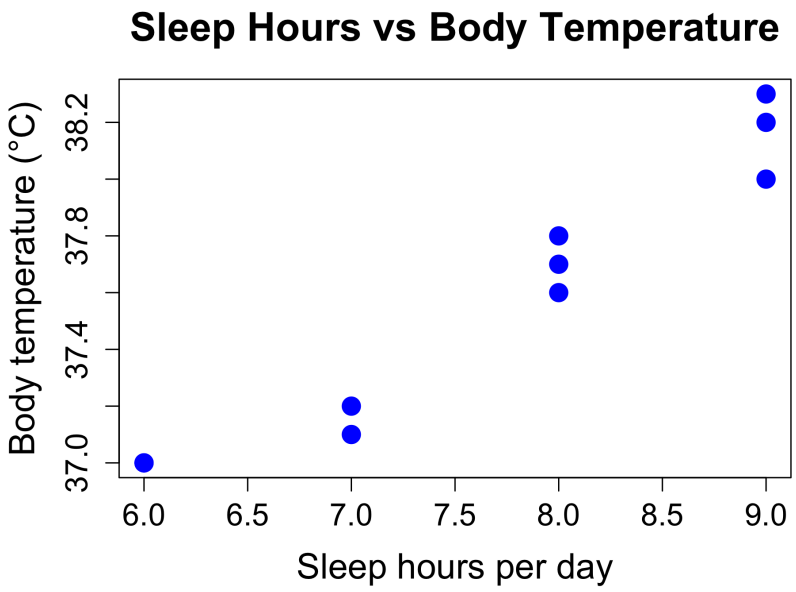

Figure 2.2: Scatter plot of sleep hours versus body temperature.

Such a conclusion, however, could lead to dangerously misguided decisions. For example, an armchair statistician physician who interprets the data in this way might attempt to reduce a patient's body temperature by restricting the amount of sleep the patient is allowed. We all know the catastrophic consequences of such a recommendation.

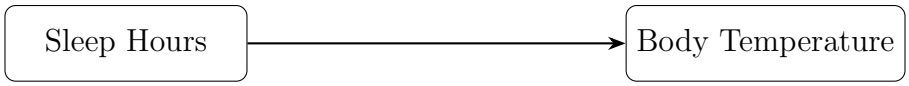

Figure 2.3: A hasty causal interpretation suggested by the observed correlation of Fig. 2.2

2.7.2 Why the catastrophe?

Suppose now that we are presented with an additional piece of information that was not included in the earlier dataset. In particular, assume that for each individual we also know the amount of virus present in their body (viral load). Once this third variable is taken into account, the dataset becomes the following.

Observation	Sleep Hours	Body Temperature (°C)	Viral Load
1	6	37.0	5
2	7	37.1	8
3	8	37.6	25
4	9	38.0	40
5	7	37.2	10
6	8	37.7	30
7	9	38.3	50
8	6	37.0	4
9	8	37.8	32
10	9	38.2	45

Table 2.7: Sleep hours, body temperature, and viral load for several individuals.

A closer inspection of Tab. 2.7 suggests that viral load may be the underlying factor influencing both sleep hours and body temperature. We explore this idea further using the scatter plot in Fig. 2.4.

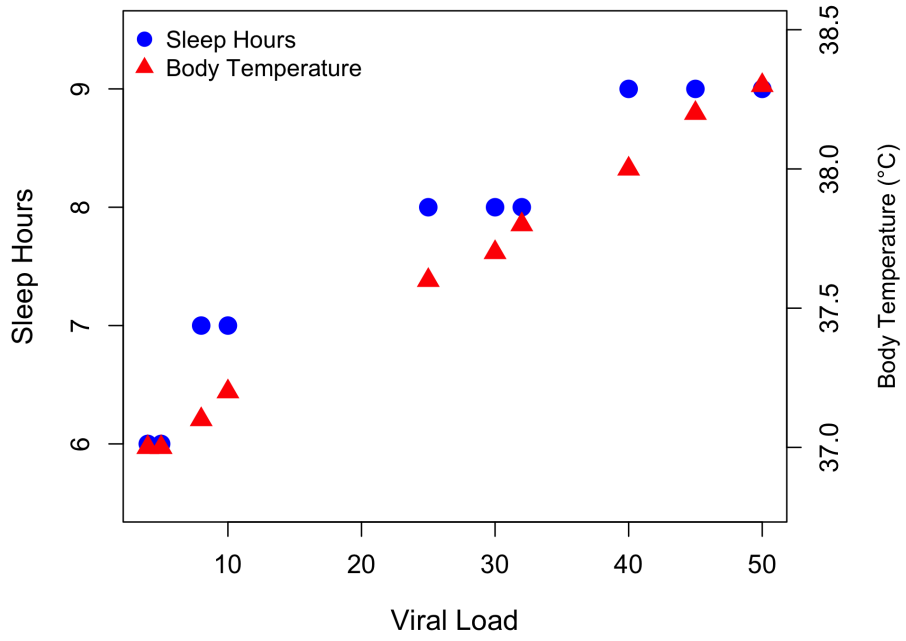

Figure 2.4: Relationship between viral load, sleep hours, and body temperature.

Fig. 2.4 suggests that the earlier observed association between sleep hours and body temperature may be explained by their common dependence on viral load; see Fig. 2.5.

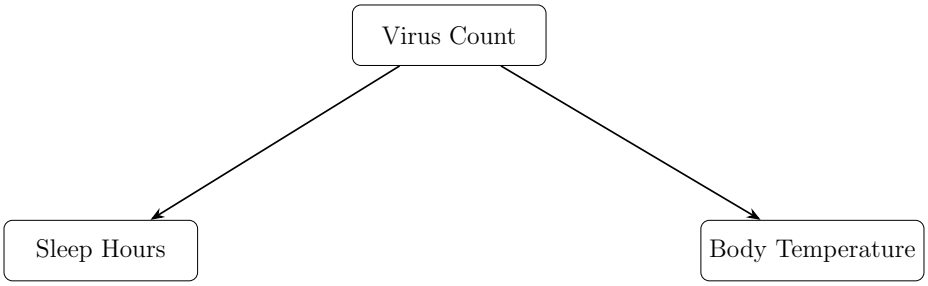

Figure 2.5: A possible causal structure in which virus count influences both sleep hours and body temperature.

If the causal structure shown in Fig. 2.5 is correct, then the appropriate medical intervention would be to treat the underlying viral infection rather than attempting to manipulate sleep hours, which would address a symptom rather than the cause. The lesson is that "Correlation does not imply causation".

Causal relationships are typically established through controlled experiments in which one variable is deliberately manipulated while other factors are held fixed or randomized. When such experiments are not feasible, specialized statistical methods from the field of causal inference are used to attempt to separate genuine causal effects from mere correlations.

2.8 A Note on Real-World Data

While the theoretical models discussed thus far assume pristine inputs, real-world data is notoriously messy—often plagued by missing values, structural inconsistencies, and human error. For a practical guide on handling these data anomalies before applying statistical models, refer to the supplementary reading: Data Cleaning Guidelines.

2.9 Exercises

1. A researcher records the concentration of a specific protein in blood samples, measured in milligrams per deciliter (mg/dL). Let this dataset be X.
   (a) What are the physical units of the arithmetic mean $\mu(X)$?
   (b) What are the physical units of the variance $\sigma^2(X)$?
   (c) What are the physical units of the coefficient of variation $\text{CV}(X)$?

2. A small startup has five employees with the following annual salaries (in thousands of dollars): $x = (45, 50, 52, 55, 950)$. The last value represents the CEO's salary.
   (a) Calculate the mean and the median for this dataset.
   (b) If a potential recruit wants to know what a "typical" employee at the company earns, which measure should you provide?
   (c) If the CEO's salary increases from 950 to 2,000, which of the two measures (mean or median) will change?

3. Suppose you are measuring the height of plants in a greenhouse in centimeters (cm).
   (a) You find the variance to be 25. Write this value with its correct units.
   (b) Calculate the standard deviation. Does the standard deviation share the same "dimension" as the original measurements?

4. Classify the following scenarios into their respective measurement scales (Nominal, Ordinal, Interval, Ratio) and state whether a "mean" is a mathematically valid operation for each:
   (a) The "Pain Scale" used in hospitals (from 0 = No Pain to 10 = Worst Possible Pain).
   (b) The weight of luggage at an airport check-in counter.
   (c) The social security numbers of a group of citizens.

Chapter 3

Probability

3.1 Set theory and Venn diagrams

Performance of an experiment on a natural system around us may lead to any one of the multiple possible outcomes. Each performance of the experiment is called a trial. First, we define a set informally as a collection of mathematical objects. For small finite sets, we may represent a set by writing out all its elements within curly braces; e.g. the set A of all even numbers on a die is $\{2, 4, 6\}$. We denote that a certain element (e.g., "2") belongs to a set A by $2 \in A$. A set that contains all the elements under consideration (all the humans on the Earth, all planets in the solar system, etc) is called the universe set.

When dealing with sets that contain a vast or infinite number of elements, listing every individual point becomes impossible. In these cases, sets are best described using a statement or the rule method. This approach defines the set by specifying the distinct properties or conditions that its elements must satisfy. The general notation is written as $\{x \mid x \text{ satisfies a specific condition}\}$. For instance, if a set A includes all cities with a population exceeding two million, the set A is written as:

$$A = \{x \mid x \text{ is a city with a population over 2 million}\}.$$

The sample space S refers to the set of all possible outcomes of the experiment under consideration; it is the universe set in the context of the outcomes of an experiment. A set B is a subset of a set A ($B \subseteq A$) if all the elements of B are also elements of A. A subset of S is called an event.

More definitions follow. The union of two sets A and B is another set $A \cup B$ which contains all the elements of A and B, but nothing more. The intersection of two sets A and B is another set $A \cap B$ which contains all the elements that are common to A and B, but nothing more. We also define an empty set $\emptyset$ which contains no element.

When two sets A and B have no common elements ($A \cap B = \emptyset$), we call them disjoint or mutually exclusive. Complement A^c (also denoted by $\bar{A}$) of a set A is a set which contains all the elements that are not present in A, and nothing more.

From the above definitions, the following theorem follows; we state it without proof.

Theorem 3.1: For a set A, $A \cap A^c = \emptyset$, and $A \cup A^c = S$.

A more intuitive way to capture the above ideas is through the use of Venn diagrams, named after the English mathematician John Venn (1834-1923). A sample Venn diagram depicting two non-disjoint sets is shown in Fig. 3.1 (to be discussed more in the class). One

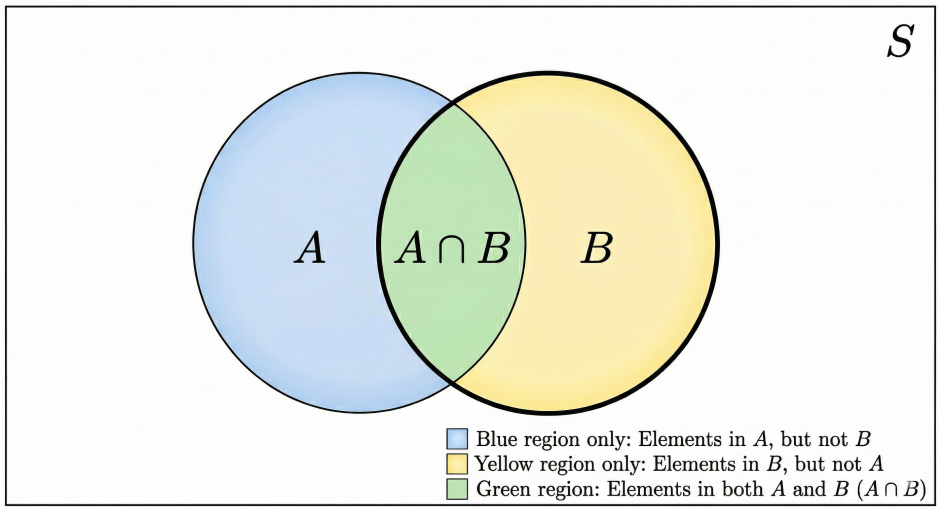

Figure 3.1: A sample Venn diagram for two non-disjoint set.

usefulness of Venn diagrams is that the veracity of certain statements can be established rather conveniently. For example, try to verify the following two de Morgan's laws with Venn diagrams.

Theorem 3.2 (De Morgan's laws)**:** For two sets A and B,

$$(A \cup B)^c = A^c \cap B^c, \tag{3.1}$$
$$(A \cap B)^c = A^c \cup B^c. \tag{3.2}$$

Another theorem that we can glean from Venn diagrams and use later is

Theorem 3.3: With $n(A)$ meant to represent the number of elements in set A, for two sets A and B,

$$n(A \cup B) = n(A) + n(B) - n(A \cap B). \tag{3.3}$$

Advanced Material 3.1: Although Theorem 3.1 can be extracted from Venn diagrams, it still does not constitute as a proof.
To prove Theorem 3.2, one has to show the equality of two sets, which can further be broken down into showing that being in one set implies being in the other, and vice versa.

3.2 Probability

3.2.1 The Axiomatic Approach to Probability

Axiom 3.1 (Kolmogorov's Axioms[a])**:** Let S be the sample space of an experiment. Probability is a real-valued function P that assigns a number to every event, satisfying the following three axioms:

1. For any event A:

$$P(A) \geq 0 \tag{3.4}$$

2.

$$P(S) = 1 \tag{3.5}$$

3. For events $A_1, A_2, \ldots$ (where $A_i \cap A_j = \emptyset$ for $i \neq j$)

$$P(A_1 \cup A_2 \cdots A_n) = \sum_{i=1}^{n} P(A_i) \tag{3.6}$$

[a]Named after the Russian mathematician Andrey Kolmogorov, who formalized probability in 1933.

Theorem 3.4: $\mathrm{P}(\emptyset) = 0$.

Proof. Since $S \cap \emptyset = \emptyset$ and $S \cup \emptyset = S$, from Eq. (3.6) we have

$$\begin{aligned} P(S) &= P(S \cup \emptyset) \\ &= P(S) + P(\emptyset). \end{aligned} \tag{3.7}$$

Substituting $P(S) = 1$ (Eq. (3.5)) yields:

$$\begin{aligned} 1 &= 1 + P(\emptyset), \\ P(\emptyset) &= 0. \end{aligned} \tag{3.8}$$

□

Theorem 3.5 (Complement Rule)**:** For any event A, $P(A^c) = 1 - P(A)$.

Proof. From Theorems 3.1, $A \cap A^c = \emptyset$ and $A \cup A^c = S$, which further imply that

$$\begin{aligned} P(S) &= P(A \cup A^c) \\ &= P(A) + P(A^c) && \text{(From Eq. (3.6))} \\ \implies 1 &= P(A) + P(A^c) && \text{(From Eq. (3.5))} \\ \implies P(A^c) &= 1 - P(A). \end{aligned} \tag{3.9}$$

□

Theorem 3.6 (Upper Bound)**:** For any event A, $P(A) \leq 1$.

Proof.

$$\begin{aligned} P(A) + P(A^c) &= 1 && \text{(From Theorem 3.5)} \\ \implies P(A) &= 1 - P(A^c) \\ P(A^c) &\geq 0 && \text{(From Eq. (3.4))} \\ \implies P(A) &\leq 1. \end{aligned} \tag{3.10}$$

□

Theorem 3.7 (Monotonicity)**:** For any events A and B such that $A \subseteq B$, $P(A) \leq P(B)$.

Proof. Since $A \subseteq B$, we can partition B as $B = A \cup (B \cap A^c)$, where $A \cap (B \cap A^c) = \emptyset$.

$$\begin{aligned} & P(B) = P(A) + P(B \cap A^c) && \text{(From Eq. (3.6))} \\ & P(B \cap A^c) \geq 0 && \text{(From Eq. (3.4))} \\ \implies \quad & P(B) \geq P(A). && \end{aligned} \tag{3.11}$$

□

Theorem 3.8 (General Addition Rule)**:** For any events A and B, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Proof. Since $A \cup B = A \cup (B \cap A^c)$ with $A \cap (B \cap A^c) = \emptyset$, and $B = (A \cap B) \cup (B \cap A^c)$ with $(A \cap B) \cap (B \cap A^c) = \emptyset$:

$$\begin{aligned} & P(A \cup B) = P(A) + P(B \cap A^c) && \text{(From Eq. (3.6))} \\ & P(B) = P(A \cap B) + P(B \cap A^c) && \text{(From Eq. (3.6))} \\ \implies & P(B \cap A^c) = P(B) - P(A \cap B) && \\ \implies & P(A \cup B) = P(A) + P(B) - P(A \cap B). && \end{aligned} \tag{3.12}$$

□

3.2.2 The Assignment Step: Specifying a Probability Distribution

Kolmogorov's axioms govern the rules of probability but leave the freedom to specify a probability distribution (the explicit mapping from the elements of the sample space to their respective probabilities).

For a coin toss with $S = \{H, T\}$, any distribution is mathematically valid provided $P(H) \geq 0$, $P(T) \geq 0$, and $P(H) + P(T) = 1$. For example, a fair coin model:

$$P(H) = 0.5, \quad P(T) = 0.5 \tag{3.13}$$

and a heavily biased coin model:

$$P(H) = 0.25, \quad P(T) = 0.75 \tag{3.14}$$

are both perfectly consistent with the axioms[1]. Since Kolmogorov's rules allow for both, we must make a specific assignment choice to fix the probabilities.

One specifically useful and popular choice for apparently symmetrical systems is embodied in the principle of equal probabilities, which leads to classical definition of probability.

Classical Probability for Symmetrical Systems

When a physical system appears reasonably symmetrical, we invoke the principle of equal probabilities, which acts as an additional axiom.

[1] Using this quarto document, one can simulate a (biased or unbiased) coin toss a large number of times.

Axiom 3.2 (Principle of Equal Probability)**:** For a finite sample space $S = \{e_1, e_2, \ldots, e_{n(S)}\}$, every outcome is equally likely. There exists a constant p such that for all $i \in \{1, 2, \ldots, n(S)\}$:

$$P(\{e_i\}) = p. \tag{3.15}$$

From this single addition to Kolmogorov's axioms, the formula for classical probability[2] follows directly.

In this course, experiments like a die roll or a draw from a lot or tossing of a coin will be assumed to follow the principle of equal probability unless contrary information is presented.

Theorem 3.9: If a system satisfies Axiom 3.2, the (classical) probability of any event A is:

$$P(A) = \frac{n(A)}{n(S)}. \tag{3.16}$$

Proof. Applying Eq. (3.6) to the mutually exclusive outcomes of S:

$$\begin{aligned} & P(S) = \sum_{i=1}^{n(S)} P(\{e_i\}) \\ \implies \quad & 1 = n(S) \cdot p \qquad \text{(From Eqs. (3.5), (3.6), and (3.15))} \\ \implies \quad & p = \frac{1}{n(S)}. \end{aligned} \tag{3.17}$$

For an event A, applying Eqs. (3.5), (3.6) and substituting Eq. (3.17) yields:

$$P(A) = \sum_{e_i \in A} P(\{e_i\}) = n(A) \cdot p = \frac{n(A)}{n(S)}. \tag{3.18}$$

□

As can be seen now, the probability distribution of Eq. (3.13) is an example of classical probability, whereas that of Eq. (3.14) is not. The classical probability model does not apply to biased systems (an unfair coin or a die).

Advanced Material 3.2: Success in Physics: The classical probability model has been remarkably successful in statistical mechanics (a branch of physics). It postulates that for an isolated physical system in equilibrium, all accessible microstates (specific microscopic configurations) are equally likely.
If Ω represents the total number of accessible microstates for a system with a fixed energy, volume, and number of particles, the probability P of finding the system in any specific microstate i is:

$$P(i) = \frac{1}{\Omega}. \tag{3.19}$$

[2]Historically known as Laplace's classical definition of probability, formulated by Pierre-Simon Laplace (1749 – 1827), a French polymath.

This single assumption provides the bridge that allows physicists to derive macroscopic thermodynamic laws from the microscopic behavior of atoms and molecules.

3.2.3 Physical Manifestation of Probability

To connect the abstract axiomatic framework of probability to observable physical phenomena, we introduce a physical postulate.

Postulate 3.1 (Cournot's Principle[a])**:** An event with a mathematical probability close to 1 will, with practical certainty, physically occur.

[a]Antoine Augustin Cournot (1801–1877) was a French mathematician, philosopher, and economist.

Theorem 3.10 (Physical Manifestation of Probability)**:** By applying Cournot's Principle (Postulate 3.1) and the Central Limit Theorem (to be formally established in Advanced Material 4.4), one can conclude that for a sufficiently large number of independent physical trials (N), the relative frequency of an event A will approximate its probability:

$$\frac{N_A}{N} \approx P(A) \quad \text{as } N \to \infty. \tag{3.20}$$

Equation (3.20) allows us to test our models against reality. If an experiment significantly violates this relation for very large N, Kolmogorov's Axioms 3.1 are not broken. Instead, the assignment step (where we assign a certain probability distribution) of Sec. 3.2.2 is flawed: the chosen probability distribution does not accurately reflect the physical system and must be revised.

In this Quarto document Cournot's principle plays out in the context of throwing a die and considering the event $\{1, 5\}$, whose probability is $1/3$.

Advanced Material 3.3: Deconstructing Theorem 3.10: Understanding how this theorem comes about involves advanced concepts, some of which are beyond the scope of these lecture notes. This discussion is intended for readers familiar with random variables, sampling statistics, and the advanced formulation of the Central Limit Theorem provided in Advanced Material 4.4 of Chapter 4.

1. The Mathematical Pillar: In the context of sampling statistics, we define an indicator random variable I_i for each trial i, where $I_i = 1$ if event A occurs and $I_i = 0$ otherwise. The observed relative frequency N_A/N is mathematically the sample mean Z of these variables. Utilizing the properties established in Advanced Material 4.4, where we discuss the Central Limit Theorem:

- (4.37a) $\implies E[Z] = P(A)$
- (4.37b) $\implies V[Z] \to 0$ as $N \to \infty$
- (4.37c) $\implies f_Z(z)$ concentrates into a narrow spike at $P(A)$

Consequently:

$$P\left(\left|\frac{N_A}{N} - P(A)\right| \leq \epsilon\right) \to 1 \text{ as } N \to \infty \tag{3.21}$$

Note that this result could also have been arrived at via the weak law of large numbers.
2. The Physical Pillar: Cournot's Principle (Postulate 3.1) bridges mathematical theory and observation by asserting that events with probability near 1 are physical certainties. Since our argument shows the probability of the condition $|N_A/N - P(A)| \leq \epsilon$ approaches 1, this relation must physically hold in practice. This results in the observable phenomenon $N_A/N \approx P(A)$ for large N.

Advanced Material 3.4: The Frequentist Approach: The Frequentist approach to probability was pioneered by Richard von Mises[a]. He essentially adopted the relationship expressed in Eq. (3.20) as the formal definition of probability itself:

$$P(A) \equiv \lim_{N\to\infty} \frac{N_A}{N}. \tag{3.22}$$

For a mathematician, this approach was problematic because the theory was grounded in the physical world rather than abstract logic. Because the limit in Eq. (3.22) involves random physical trials rather than a deterministic sequence, it is not formally well-defined in the sense of standard calculus.
To understand this undesirability, consider an analogy with the sine function (sin). In elementary geometry, the sine of an angle is defined by the properties of a physical triangle that one can draw:

$$\sin(\theta) = \frac{\text{perpendicular}}{\text{hypotenuse}}. \tag{3.23}$$

While intuitive, defining a mathematical function based on the measurements of a physical object is restrictive. A more rigorous, pure definition characterizes $\sin(x)$ as the unique solution to the following differential equation:

$$y'' + y = 0, \quad \text{with } y(0) = 0 \text{ and } y'(0) = 1. \tag{3.24}$$

Just as this differential equation provides a formal foundation for the sine function independent of physical triangles, Kolmogorov's axioms provide a formal foundation for probability independent of physical trials.

[a]Richard von Mises (1883–1953) was an Austrian-American mathematician and physicist. His brother, Ludwig von Mises (1881–1973), was a world-renowned economist.

3.3 Conditional Probability and Independence

3.3.1 Conditional Probability and its Physical Meaning

The concept of conditional probability was introduced to quantify the relative frequency of the occurrence of an event (say A) under certain restrictions (occurrence of another event B).

Definition 3.1 (Conditional probability)**:** The conditional probability $P(A \mid B)$ of an

event A given the occurrence of event B is

$$P(A \mid B) \equiv \frac{P(A \cap B)}{P(B)}. \tag{3.25}$$

Just like with basic probability, now comes the question of the physical meaning of conditional probability. The physical meaning is easy to deduce if we assume our system to be symmetrical and having a classical probability distribution. Then we have

$$\begin{aligned} P(A \mid B) &= \frac{P(A \cap B)}{P(B)} && \text{(From Def. 3.1)} \\ &= \frac{n(A \cap B)/n(S)}{n(B)/n(S)} && \text{(From Eq. (3.16))} \\ &= \frac{n(A \cap B)}{n(B)}. \end{aligned} \tag{3.26}$$

This confirms that $P(A \mid B)$ is the probability of A within a restricted universe where B has replaced S as the sample space.

What if the system is not symmetrical? Then we can't use the classical probability formula. In such scenarios, we turn to the physical manifestation of probability (Theorem 3.10). By interpreting $P(A)$ as the relative frequency of occurrence of A over N independent trials, the conditional probability $P(A \mid B)$ emerges as a ratio of two observed frequencies:

$$\begin{aligned} P(A \mid B) &= \frac{P(A \cap B)}{P(B)} && \text{(From Def. 3.1)} \\ &\approx \frac{N_{A\cap B}/N}{N_B/N} && \text{(From Eq. (3.20))} \\ &= \frac{N_{A\cap B}}{N_B}. \end{aligned} \tag{3.27}$$

This is the version of Cournot's principle (combined with the law of large numbers) applied to conditional probability. Physically, $P(A \mid B)$ represents the relative frequency of event A when we restrict ourselves to the N_B trials where event B occurs. Recall that probability $P(A)$ of an event A is approximately equal to the relative frequency N_A/N of A in a large number of trials (N) of a certain experiment.

3.3.2 Independence of events

Building on the framework of conditional probability, we can now build a language to quantify whether the occurrence of a certain event B can affect the probability of another event A.

Definition 3.2 (Independent events)**:** Two events A and B are considered independent when

$$P(A \mid B) = P(A), \tag{3.28}$$

or equivalently,

$$P(B \mid A) = P(B). \tag{3.29}$$

Otherwise, they are considered dependent.

One can easily show the equivalence of the above two definitions (Eqs. (3.28) and (3.29)) of independence of two events using the definition of conditional probability. In the context of a single roll of a die, the following two events are independent: • A = the number is greater than 4. • B = the number is odd. An example of dependent events would be textbullet A = the number is greater than 4. • B = the number is 6.

We invoke Cournot's principle to understand the physical manifestation of independent events. With N trials of an experiment

$$\begin{aligned} & P(A \mid B) = P(A) && \text{(From Eq. (3.28))} \\ \implies & \frac{N_{A\cap B}}{N_B} \approx \frac{N_A}{N}. && \text{(From Eqs. (3.27) and (3.20))} \end{aligned} \tag{3.30}$$

This shows that A being independent of B means that its overall relative frequency is the same as its relative frequency when restricted to trials where B occurred.

Advanced Material 3.5: Independence as a Spectrum: While independence is defined by the strict equality of Eq. (3.32), applied probability often treats dependence as a spectrum. We quantify this "degree of dependence" using the absolute difference:

$$\Delta = |P(A \cap B) - P(A)P(B)|$$

To formalize this association, statisticians use *covariance*. Because covariance requires numerical variables rather than event sets, we map the events to binary indicator variables:

$$I_E = \begin{cases} 1 & \text{if event } E \text{ occurs} \\ 0 & \text{if event } E \text{ does not occur} \end{cases}$$

Applying the standard definition of covariance to these indicators yields exactly our probability difference: $\Delta = |\mathrm{Cov}(I_A, I_B)|$.
Consequently, $\Delta = 0$ guarantees strict independence between A and B, while a larger Δ quantifies a stronger statistical relationship between A and B in the form of a larger magnitude covariance between the indicator variables (I_A, I_B) of A and B.

3.3.3 Multiplication Theorem

From Definition 3.1, we have the multiplication theorem.

Theorem 3.11 (Multiplication Theorem)**:**

$$P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A), \tag{3.31}$$

which specializes to the case of independent events as

$$P(A \cap B) = P(A)P(B). \tag{3.32}$$

Sometimes Eq. (3.32) is used as the definition of independence of events as it is manifestly symmetric under the switch $A \leftrightarrow B$, unlike the original Definition 3.2.

3.3.4 More than Two Events

Here we generalize the multiplication theorem and the concept of independence to more than two events.

Theorem 3.12 (General Multiplication Theorem)**:** The probability of the joint occurrence of n events is the product of their conditional probabilities. For $n = 3$:

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2). \tag{3.33}$$

Proof. By treating $(A_1 \cap A_2)$ as a single event and applying the definition of conditional probability (Def. 3.1) twice, we have:

$$\begin{aligned} P(A_1 \cap A_2 \cap A_3) &= P(A_3 \mid A_1 \cap A_2)P(A_1 \cap A_2) \\ &= P(A_3 \mid A_1 \cap A_2)\,[P(A_2 \mid A_1)P(A_1)] \\ &= P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2). \end{aligned} \tag{3.34}$$

□

The condition for the independence of two events established in Eq. (3.32) can be extended to an arbitrary number of events as follows:

Definition 3.3 (Mutual Independence)**:** A collection of n events $\{A_1, A_2, \ldots, A_n\}$ is mutually independent $\equiv$ for every possible selection of two or more events $\{A_i, A_j, \ldots, A_k\}$ from the collection:

$$P(A_i \cap A_j \cap \cdots \cap A_k) = P(A_i)P(A_j)\ldots P(A_k). \tag{3.35}$$

For three events $\{A, B, C\}$ to be mutually independent, the following four equations must hold:

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

3.3.5 The Theorems of Total Probability and Bayes

Sometimes we know the conditional probability $P(A \mid B)$, but are faced with the inverse problem of finding $P(B \mid A)$. Bayes' Theorem provides the exact mathematical machinery to flip these conditionals. To build up to it, we first need to define a way to break a sample space into non-overlapping, comprehensive pieces.

Definition 3.4 (Partition of a Sample Space)**:** A collection of events $\{B_1, B_2, \ldots, B_n\}$ is said to form a partition of the sample space S if they satisfy the following three conditions:

1. Mutually Exclusive:

$$B_i \cap B_j = \emptyset \quad \text{for all } i \neq j. \tag{3.36}$$

2. Collectively Exhaustive:

$$B_1 \cup B_2 \cup \cdots \cup B_n = S. \tag{3.37}$$

3. Non-zero Probability:

$$P(B_i) > 0 \quad \text{for all } i = 1, 2, \ldots, n. \tag{3.38}$$

Theorem 3.13 (Theorem of Total Probability)**:** If the events $(B_1, B_2, \ldots, B_n)$ constitute a partition of the sample space S, then for any event $A \subseteq S$:

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(B_i)P(A \mid B_i). \tag{3.39}$$

Proof.

$$\begin{aligned}
A &= A \cap S && \text{(Intersection with universe)} \\
&= A \cap (B_1 \cup B_2 \cup \cdots \cup B_n) && \text{(via Eq. (3.37))} \\
&= (A \cap B_1) \cup (A \cap B_2) \cup \cdots \cup (A \cap B_n) && \text{(Distributive law)} \\
P(A) &= \sum_{i=1}^{n} P(A \cap B_i) && \text{(via Axiom 3.1 \& Eq. (3.36))} \\
&= \sum_{i=1}^{n} P(B_i)P(A \mid B_i). && \text{(via Eq. (3.31))} \qquad (3.40)
\end{aligned}$$

□

Theorem 3.14 (Bayes' Theorem)**:** Let $\{B_1, B_2, \ldots, B_n\}$ be a partition of the sample space S, and let A be any event such that $P(A) > 0$. Then

$$P(B_k \mid A) = \frac{P(B_k)P(A \mid B_k)}{\sum_{i=1}^{n} P(B_i)P(A \mid B_i)} = \frac{P(B_k)P(A \mid B_k)}{P(A)}. \tag{3.41}$$

Proof.

$$\begin{aligned}
P(B_k \mid A) &= \frac{P(B_k \cap A)}{P(A)} && \text{(via Def. 3.1)} \\
&= \frac{P(B_k)P(A \mid B_k)}{P(A)} && \text{(via Eq. (3.31))} \\
&= \frac{P(B_k)P(A \mid B_k)}{\sum_{i=1}^{n} P(B_i)P(A \mid B_i)}. && \text{(via Eq. (3.39))} \qquad (3.42)
\end{aligned}$$

□

3.3.6 Independence of Experiments

Often, a single physical process is actually a sequence or combination of simpler procedures. We call the overall process a composite experiment.

Definition 3.5 (Composite Experiment)**:** A composite experiment is an experiment composed of two or more distinct sub-experiments. The final outcome of the composite experiment is determined by the combined outcomes of its individual sub-experiments.

To illustrate this, consider the following two examples of composite experiments built out of sub-experiments (denoted as $\mathcal{E}_1$ and $\mathcal{E}_2$):

- **Example 1:** You roll a standard 6-sided die (experiment $\mathcal{E}_1$). Afterward, you physically toss a coin (experiment $\mathcal{E}_2$). The final outcome is recorded as a combined result (e.g., "4 and Heads").
- **Example 2:** You roll a standard 6-sided die (experiment $\mathcal{E}_1$). If the die lands on a 1 or 2, you toss a coin (experiment $\mathcal{E}_2$) and record the combined result (e.g., "1 and Heads"). If the die lands on any other number, you do not toss the coin, and the experiment ends (e.g., just "4").

In both cases, the overall experiment is made of two sub-experiments, but the relationship between those sub-experiments is fundamentally different. To categorize this difference, we extend our concept of independence from individual events to entire experiments.

Definition 3.6 (Independence of Experiments)**:** Two experiments $\mathcal{E}_1$ and $\mathcal{E}_2$, with respective sample spaces S_1 and S_2, are considered independent if:

$$P(A_1A_2) \equiv P(A_1 \cap A_2) = P(A_1)P(A_2) \tag{3.43}$$

for all possible events $A_1 \subseteq S_1$ and $A_2 \subseteq S_2$.

Example 2 above violates Eq. (3.43). To see this, the sample space of $\mathcal{E}_1$ is $S_1 = \{1, 2, 3, 4, 5, 6\}$, and the sample space of $\mathcal{E}_2$ is $S_2 = \{H, T, \text{blank}\}$, where "blank" occurs whenever the die roll is not a 1 or 2. In this setup, the probability of getting a Heads is $P(H) = P(H \cap 1) + P(H \cap 2) = P(H \mid 1)P(1) + P(H \mid 2)P(2) = (1/2)(1/6) + (1/2)(1/6) = 1/6$. Here we used Eq. (3.39). Now, if A_1 is the event of rolling a 4 ($P(A_1) = 1/6$), then $P(A_1 \cap H) = 0$ because a die roll of 4 physically results in a "blank" for $\mathcal{E}_2$. Since $P(A_1 \cap H) \neq P(A_1)P(H)$ since $0 \neq (1/6)(1/6)$, the sub-experiments $\mathcal{E}_1$ and $\mathcal{E}_2$ are not independent.

On the other hand, Example 1 above satisfies Eq. (3.43), as one can check. Again, use has to be made of Eq. (3.39).

Theorem 3.15 (Equiprobable Outcomes in Composite Experiments)**:** If a composite experiment is formed by n sub-experiments $\mathcal{E}_1, \ldots, \mathcal{E}_n$, each with equally probable outcomes (as per Axiom 3.2), such that the $\mathcal{E}_i's$ don't affect each other, then the outcomes of the composite experiment are also equally probable.

We will not give a general proof. One can verify this result for a simple specific experiment. Consider two tosses of a fair coin. Since $P(H) = P(T) = 1/2$ for each toss, the multiplication rule of Eq. (3.31) yields a probability of $1/4$ for each of the four possible outcomes ($\{HH, HT, TH, TT\}$).

3.3.7 Sampling With and Without Replacement

Here is an example of an experiment composed of sub-experiments which exemplifies independence and dependence between events. We make two draws from an urn containing two balls: one Blue (B) and one Green (G). Let the sub-experiments $\mathcal{E}_1$ and $\mathcal{E}_2$ represent the first and second draws, where B_n and G_n denote drawing a specific color on the n-th trial ($n \in \{1, 2\}$).

Sampling Without Replacement

In this case, the ball from $\mathcal{E}_1$ is not returned. The physical setup dictates the following probability assignments:

$$\begin{aligned} P(B_1) &= 1/2, & P(G_1) &= 1/2 \\ P(B_2 \mid G_1) &= 1, & P(G_2 \mid B_1) &= 1 \\ P(B_2 \mid B_1) &= 0, & P(G_2 \mid G_1) &= 0 \end{aligned} \tag{3.44}$$

Now,

$$P(B_1 \cap B_2) = P(B_2 \mid B_1)P(B_1) = 0. \tag{3.45}$$

Now we use Def. 3.1 to find:

$$P(B_1 \mid B_2) = \frac{P(B_1 \cap B_2)}{P(B_2)} = 0. \tag{3.46}$$

Since $P(B_1 \mid B_2) = 0 \neq P(B_1) = 1/2$, it is in violation of the independence condition in Eq. (3.43) confirming that experiments $\mathcal{E}_1$ and $\mathcal{E}_2$ are dependent.

Sampling With Replacement

If the ball is returned after $\mathcal{E}_1$, the physical state is reset, leading to the following assignments:

$$\begin{aligned} P(B_1) &= 1/2, & P(G_1) &= 1/2 \\ P(B_2 \mid B_1) &= 1/2, & P(B_2 \mid G_1) &= 1/2 \\ P(G_2 \mid B_1) &= 1/2, & P(G_2 \mid G_1) &= 1/2 \end{aligned} \tag{3.47}$$

The independence is proven by first deriving the total probability of B_2:

$$\begin{aligned} P(B_2) &= P(B_2 \cap B_1) + P(B_2 \cap G_1) && \text{(From Eq. (3.39))} \\ &= P(B_2 \mid B_1)P(B_1) + P(B_2 \mid G_1)P(G_1) && \text{(From Def. 3.1)} \\ &= (1/2)(1/2) + (1/2)(1/2) = 1/2. \end{aligned} \tag{3.48}$$

Comparing the results:

$$P(B_2 \mid B_1) = 1/2 = P(B_2) \quad \text{and} \quad P(B_2 \mid G_1) = 1/2 = P(B_2). \tag{3.49}$$

Since analogous results hold for G_2, as can be similarly shown, the sub-experiments satisfy the condition in Eq. (3.43) and are independent.

Advanced Material 3.6: Two Formalisms for Composite Experiments: When modeling a composite experiment, we have two frameworks to choose from. One approach models the system sequentially, while the other imposes global structural assumptions upfront. Consider a simple experiment: two tosses of a fair coin, such that the two tosses don't affect each other (reflected in independence of the tosses).

Approach 1: The Sequential (Physical) Method This approach translates physical causality (or lack thereof) directly into conditional probabilities.

- **Axiom 1 (Initial State):** The first toss is fair. $P(H_1) = P(T_1) = 1/2$.
- **Axiom 2 (Physical Transition):** The first toss has no effect on the second toss. Therefore, the conditional probabilities are: $P(H_2 \mid H_1) = P(H_2 \mid T_1) = P(T_2 \mid H_1) = P(T_2 \mid T_1) = 1/2$.
- **The Theorem:** We use Eq. (3.39) to derive the overall probability of the second toss:

$$\begin{aligned} P(H_2) &= P(H_2 \mid H_1)P(H_1) + P(H_2 \mid T_1)P(T_1) \\ &= (1/2)(1/2) + (1/2)(1/2) = 1/2. \end{aligned}$$

  Because we have shown that $P(H_2) = P(H_2 \mid H_1)$, independence as defined in Eq. (3.28) is not assumed, but rather *proven* as a resulting theorem.

Approach 2: The Global Method This approach ignores chronological sequence and imposes an abstract mathematical definition on the entire system at once.

- **Axiom 1:** We state upfront that $P(H_1) = P(H_2) = P(T_1) = P(T_2) = 1/2$.
- **Axiom 2:** We take independence as a foundational axiom as per Eq. (3.32): $P(H_1 \cap H_2) = P(H_1)P(H_2) = 1/4$.
- **The Theorem:** From these axioms, the conditional probability can be computed trivially using Def. 3.1:

$$P(H_2 \mid H_1) = \frac{P(H_1 \cap H_2)}{P(H_1)} = \frac{1/4}{1/2} = 1/2.$$

Which approach is superior? While both are mathematically sound for simple independent systems, Approach 1 is superior and more general because it successfully handles dependent systems. Consider the following sequential dependent experiment. Roll a die. If it lands on 1 or 2 (Event A, where $P(A) = 1/3$), you toss a coin. Otherwise (A^c, where $P(A^c) = 2/3$), you do not toss the coin.
In Approach 1, we simply assign the physical transition rules: $P(H \mid A) = 1/2$ and $P(H \mid A^c) = 0$. We then easily calculate the overall probability of getting heads via Eq. (3.39):

$$P(H) = P(H \mid A)P(A) + P(H \mid A^c)P(A^c) = (1/2)(1/3) + (0)(2/3) = 1/6.$$

It works flawlessly.

However, Approach 2, completely fails here. We cannot establish upfront an "independence axiom" because the coin toss (H) does not physically exist in two-thirds of the universe of the sample space. For these reasons, Approach 1 is the primary formalism adopted in this course.

3.4 Combinatorics

3.4.1 Permutation

A permutation of certain objects refers to their arrangement in a certain order. As an example, the letters A, B, and C can be arranged in the following six permutations: ABC, ACB, BAC, BCA, CAB, and CBA.

Theorem 3.16 (Permutations With and Without Repetition)**:** Consider a set of n distinct objects. The number of ways to arrange k of these objects into a specific sequence is given by[a]:

1. Without repetition: The number of permutations is $n!/(n-k)! \equiv {}_nP_k \equiv {}^nP_k \equiv P(n,k)$.
2. With repetition: The number of permutations is n^k.

[a]The factorial of a non-negative integer n, denoted by $n!$, is defined as $n! \equiv n \times (n-1) \times \cdots \times 2 \times 1$. We also define $0! = 1$.

Proof. 1. Without repetition: We must fill k positions. We have n choices for the first position, $(n-1)$ choices for the second position, $(n-2)$ for the third position, and continuing in this pattern, $(n-k+1)$ choices for the k-th position. Multiplying these independent choices gives:

$$n \times (n-1) \times (n-2) \times \cdots \times (n-k+1)$$

Multiplying the numerator and the denominator by $(n-k)!$ yields the closed-form factorial expression:

$$\frac{n \times (n-1) \times \cdots \times (n-k+1) \times (n-k)!}{(n-k)!} = \frac{n!}{(n-k)!}$$

2. **With repetition:** We again have k positions to fill. Because objects can be selected multiple times, there are n available choices for the first position, n choices for the second position, and so on, up to n choices for the k-th position. Multiplying these choices together gives:

$$\underbrace{n \times n \times \cdots \times n}_{k \text{ times}} = n^k$$

□

3.4.2 Combinations

A combination is a selection of items from a larger set where the order of selection does not matter. For example, if you are choosing a two-person committee from a group of three students (Alice, Bob, and Charlie), the selection {Alice, Bob} is mathematically identical to {Bob, Alice}.

Theorem 3.17 (Combinations)**:** The number of different combinations of n different

things taken k at a time, without repetitions, is[a]:

$$\frac{n!}{k!(n-k)!} \equiv \binom{n}{k} \equiv {}_nC_k \equiv {}^nC_k \equiv C(n,k).$$

And the number of those combinations with repetitions is:

$$\binom{n+k-1}{k}$$

[a] The binomial coefficient notation $\binom{n}{k}$, typically read as "n choose k".

Proof. We will prove the formula for combinations only for the former case of without repetition.
From Theorem 3.16, we established that the number of ways to select and arrange k objects from n distinct objects (permutations) is $n!/(n-k)!$. However, in a combination, the internal ordering of the k chosen objects is irrelevant. Because any discrete set of k distinct objects can be ordered in $k!$ different ways, the permutation formula overcounts each distinct combination $k!$ times. To isolate the number of unique combinations, we divide the total number of permutations by this redundancy factor of $k!$:

$$\text{Combinations} = \frac{\text{Permutations}}{k!} = \frac{\frac{n!}{(n-k)!}}{k!} = \frac{n!}{k!(n-k)!}.$$

□

Important 3.1: Two Methods for Combinatorial Sampling: When solving probability problems that involve sampling, there are generally two valid approaches. To understand their connection, consider a simple problem: An urn contains 3 distinct balls, $\{A, B, C\}$. We draw 2 balls without replacement. What is the probability of selecting ball A first, followed by ball B (the specific sequence AB)?

Approach 1 (The Sequential Method): Here, our starting assumption (which we can call a micro-level axiom) is that at each individual stage of the draw, every remaining ball in the urn is equally likely to be chosen. With this, we can calculate the probability using the conditional Multiplication Theorem (Eq. 3.31):

$$\begin{aligned} P(\text{Sequence } AB) &= P(A_1 \cap B_2) \\ &= P(A_1)P(B_2 \mid A_1) \\ &= \left(\frac{1}{3}\right)\left(\frac{1}{2}\right) = \frac{1}{6}. \end{aligned}$$

Approach 2 (The Combinatorics Method): In this method, our starting assumption (a macro-level axiom) is that all possible overall permutations in the sample space are equally likely. Using the classical formula (Theorem 3.9) and permutations (Theorem 3.16), we simply count:

$$P(\text{Sequence } AB) = \frac{n(\text{favorable})}{n(S)} = \frac{1}{{}_3P_2} = \frac{1}{6}.$$

The Bridge (Deriving the Macro-Axiom): In Approach 2, we assumed that the ${}_3P_2 = 6$ possible permutations (AB, BA, AC, CA, BC, CB) were all equally likely. However, we can prove this macro-level axiom using the micro-level axiom from Approach 1, thus turning it into a theorem.
To do so, note that the joint probability of *any* specific sequence of two balls is identically $\frac{1}{3} \times \frac{1}{2} = \frac{1}{6}$. Because this local, stage-wise math is perfectly symmetric (it depends only on the number of available balls, not their specific identity), all 6 permutations are mathematically forced to be equiprobable. This establishes the macro-level assumption as a derivable theorem. This logic also naturally extends to combinations (unordered sets), ensuring they too are equiprobable.

General Principle: This specific example illustrates a universal mathematical truth: The macro-level assumption that "all permutations or combinations are equally likely" is not a blind guess. It is a direct consequence of stage-wise equiprobability. Whenever you use a combinatorics formula to find a probability, you are implicitly relying on the symmetry of the underlying sequential steps.
This concept is analogous to Theorem 3.15 (Equiprobable Outcomes in Composite Experiments). The key distinction is that Theorem 3.15 deals with strictly *independent* sub-experiments, whereas combinatorial sampling without replacement involves *dependent* sub-experiments. Despite this dependence, the symmetry of the conditional probabilities ensures the final outcomes remain uniform.

3.5 Exercises

1. Write out the complete sample space S for each of the following experiments:
   (a) A standard 6-sided die and a fair coin are tossed simultaneously.
   (b) Roll a 6-sided die. If it lands on 1 or 2, the experiment stops. If it lands on 3, 4, 5, or 6, immediately flip a coin and record the result.
   (c) Draw 3 cards from a standard deck one by one, recording only the color of the card drawn (Red or Black).
2. Let the sample space be $S = \{1, 2, 3, 4, 5, 6\}$, representing a standard die roll. Define the events A: the number is prime; B: the number is odd; and C: the number is less than 4.
   (a) Draw a Venn diagram placing all six outcomes in their correct regions.
   (b) Identify the elements in the intersection $A \cap C$.
   (c) Identify the elements in the union $B \cup C$.
   (d) Identify the elements in the complement $(A \cup B)^c$.
3. Verify the two theorems of De Morgan (discussed in the lectures) using
   (a) two example sets A and B that you invent yourself.
   (b) Venn diagrams without reference to a specific set.

4. Using Venn diagrams, graph and check the rules:

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C),$$
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C).$$

5. Use the rule method of specifying a set to describe the set S consisting of all points in the second quadrant of a Cartesian plane.

   For this problem, it is assumed that the student is aware of Cartesian system of specifying points in a place with its x and y coordinates in the form (x, y).

6. Which of the following sets are equal?

   (a) $A = \{2, 4\}$;

   (b) $B = \{x \mid x \text{ is a number on a standard die}\}$;

   (c) $C = \{x \mid x^2 - 6x + 8 = 0\}$;

   (d) $D = \{x \mid x \text{ is the number of heads when eight coins are tossed}\}$.

   For part 6c of this problem, it is assumed that the student is aware of the quadratic formula, which states that for a quadratic equation $ax^2 + bx + c = 0$, the values of x that satisfy the equation are given by

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

7. An experiment involves tossing a pair of dice, one blue and one yellow, and recording the numbers that come up. If x equals the outcome on the blue die and y the outcome on the yellow die, describe the sample space S

   (a) by listing the elements (x, y);

   (b) by using the rule method.

8. If $S = \{x \mid 0 < x < 15\}$, $M = \{x \mid 2 < x < 10\}$, and $N = \{x \mid 0 < x < 6\}$, find

   (a) $M \cup N$;

   (b) $M \cap N$;

   (c) $M^c \cap N^c$.

9. A die is such that the probability $P(\{x\})$ of rolling a number x is p if x is odd and $2p$ if it is even.

   (a) Is this a fair die?

   (b) Find p

   (c) $P(\{4, 5\})$.

10. Two fair 6-sided dice (one black and one white) are rolled simultaneously. Assume the outcomes are equally probable.

    (a) Method 1 (Enumeration): What is the probability that the sum of the two dice is exactly 8 or both dice show the same number (a "double")? Solve this by manually counting the unique favorable outcomes in the sample space.

(b) Method 2 (Addition Rule): Solve the same problem using the general addition rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Verify that the result matches your answer from part 10a.

11. Consider the same experiment involving two fair dice as described in Exercise 10. Assume the rolls are don't affect each other.

    (a) Method 1 (Enumeration): Find the probability that the black die shows an even number and the white die shows a 5 by listing all favorable outcomes in the sample space.

    (b) Method 2 (Multiplication theorem): Calculate the same probability using the multiplication theorem. Verify that it matches your result from part 11a.

12. An aircraft is equipped with two identical engines that operate independently. The aircraft crashes only if both engines fail. Let p be the probability that a chosen engine fails.

    (a) If the probability of a crash is ϵ, find p in terms of ϵ.

    (b) If $p = 10^{-3}$, find ϵ.

13. Consider an experiment where two cards are drawn sequentially with replacement from a set containing three distinct cards: one Red (R), one Blue (B), and one Green (G). Let $\mathcal{E}_1$ and $\mathcal{E}_2$ represent the first and second draws, respectively.

    (a) List the complete sample space S for this composite experiment.

    (b) Calculate the probability of each individual outcome in S.

    Note: Observe that this neat probability distribution follows from the symmetry of the individual draws and their independence; it does not need to be assumed.

    (c) Verify that your results in part (b) satisfy Kolmogorov's second axiom, $P(S) = 1$.

    (d) Without calculating the individual probabilities of every other outcome S, find the probability of *not* drawing two Green cards (G, G).

14. A carton contains twelve eggs, two of which are rotten. Two eggs are drawn at random. Find the probability that neither egg is rotten if the sampling is performed:

    (a) with replacement,

    (b) without replacement.

15. Refer back to the die-coin experiment from Exercise 1b. Assume both the die and the coin are fair. Let A represent the event that the die roll is > 2, and T be the event that the experiment ends with a "Tails".

    (a) What is the probability that the experiment ends without needing to flip the coin?

    (b) What is the probability of T?

    (c) Intuitively, if the experiment ends in a Tails, the die roll must have been > 2. Verify this intuition mathematically by calculating the conditional probability $P(A \mid T)$.

    (d) Are A and T independent?

Note: To witness the convergence of classical and frequentist probabilities for the event in part (15b), this physical experiment has been simulated on a computer. You may verify your analytical results and observe the behavior of large-scale trials by running this Quarto source document.

16. Your probability of successfully securing a job when applying to a job application is p. If you apply to n job applications, what is the probability of securing at least one job? Assume each job application process functions independently of the other ones.

17. A random 4-letter passcode is created by arranging the letters $\{A, B, C, D\}$ such that no letter is repeated. What is the probability that the passcode spells the word "BADC"? Solve in two ways, with and without using combinatorics formulae.

18. A small box contains 8 light bulbs, 3 of which are burnt out. If you select 2 bulbs at random without replacement, what is the probability that both bulbs are functional? Solve in two ways, with and without using combinatorics formulae.

19. A 3-digit lock code is generated where each digit can be any number from 0 through 9, and repetition is allowed. What is the probability that all three digits in the code are identical (e.g., 7-7-7)? Solve in two ways, with and without using combinatorics formulae.

20. A committee of 3 students is to be chosen at random from a group consisting of 5 boys and 4 girls. What is the probability that the committee consists entirely of girls? Solve in two ways, with and without using combinatorics formulae.

21. A manufacturing plant uses three machines $(\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3)$ to produce identical components. Machines $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ produce 50%, 30%, and 20% of the total output, respectively. The probability that a machine produces a defective component (D) is given by: $P(D \mid \mathcal{M}_1) = 0.02$, $P(D \mid \mathcal{M}_2) = 0.04$, and $P(D \mid \mathcal{M}_3) = 0.05$, where $\mathcal{M}_i$ also represents the event that a particular component was built with the machine $\mathcal{M}_i$.

    (a) Do the events $\{\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3\}$ form a partition of the sample space S.

    (b) Calculate the probability $P(D)$ that a randomly selected component is defective.

    (c) A component is chosen at random and found to be defective. Using Bayes' Theorem, find the probability that it was produced by machine $\mathcal{M}_3$.

Chapter 4

Special Probability Models: Discrete and Continuous

4.1 Random Variables

In Chapter 3, we defined outcomes and events as sets of qualitative descriptions (e.g., $\{H, T\}$ or $\{\text{Red}, \text{Black}\}$). However, to give an even more mathematically informative description of the system under consideration, we often need to associate these outcomes with numerical values. This is achieved through the concept of a random variable.

Definition 4.1 (Random Variable)**:** A random variable X is a real-valued function that maps every element in the sample space S to a point on the real number line $\mathbb{R}$. We denote this as $X : S \to \mathbb{R}$.

For example, consider an experiment where a fair coin is tossed twice. The sample space is $S = \{HH, HT, TH, TT\}$. If we define the random variable X as the *number of heads*, the mapping is $X(HH) = 2$, $X(HT) = 1$, $X(TH) = 1$, and $X(TT) = 0$.

A random variable is not "random"; it is a fixed function. The randomness resides entirely in the sample space S. Once an outcome $s \in S$ is realized, the value $X(s)$ is fixed.

Let us denote the distinct, countably discrete numerical values that X can take as small x_k, where $k = 1, 2, \ldots, m$. In the coin example above, the distinct values are $x_1 = 0$, $x_2 = 1$, and $x_3 = 2$.

4.1.1 The Probability Distribution Function

Definition 4.2 (Probability Distribution Function)**:** The probability distribution function (PDF) $f(x_k)$ of a discrete random variable X is defined for any distinct numerical outcome x_k of X as:

$$f(x_k) \equiv P(X = x_k) \tag{4.1}$$

4.1.2 Measures of Central Tendency and Dispersion

In Chapter 2, we defined the mean and variance for a fixed dataset x. We now define the mean and variance for an abstract random variable X. Notice that we use the exact same Greek symbols (μ and σ^2) for both concepts. We distinguish them purely by their arguments: $\mu(x)$ operates on a physical dataset, whereas $\mu(X)$ operates on a theoretical probability model.

Definition 4.3 (Expected Value or Theoretical Mean)**:** The expected value (or theoretical mean) $\mu(X)$ of a discrete random variable X which takes values $x_1, x_2, \cdots, x_m$ is defined as

$$\mu(X) \equiv E[X] = \sum_{k=1}^{m} x_k f(x_k) \tag{4.2}$$

Definition 4.4 (Theoretical Variance)**:** The theoretical variance $\sigma^2(X)$ of a discrete random variable X which takes values $x_1, x_2, \cdots, x_m$ is defined as

$$\sigma^2(X) \equiv \text{Var}(X) = \sum_{k=1}^{m} (x_k - \mu(X))^2 f(x_k) \tag{4.3}$$

4.1.3 Physical Manifestation of Mean and Variance

We now possess two parallel frameworks. In Chapter 2, we defined the sample mean $\mu(x)$ and sample variance $\sigma^2(x)$ as arithmetic operations performed on a collected physical dataset $x \equiv (x_1, x_2, \ldots, x_N)$. In contrast, this chapter defined the theoretical mean $\mu(X)$ and theoretical variance $\sigma^2(X)$ using an abstract probability model over discrete outcomes $k = 1 \ldots m$.

Given this strict distinction—physical data versus mathematical abstraction—why use identical terms and symbols for both? The justification lies in the following theorem, which guarantees that theoretical parameters physically manifest as sample statistics in the long run.

Theorem 4.1 (Physical Manifestation of Mean and Variance)**:** By applying Cournot's Principle (Postulate 3.1) and the Central Limit Theorem (formalized in Advanced Material 4.4), one can conclude that if $x \equiv (x_1, x_2, \ldots, x_N)$ is a dataset generated by N independent physical trials of a random variable X, the physical sample statistics will converge to the abstract theoretical parameters as the number of trials becomes sufficiently large:

$$\mu(x) \approx \mu(X) \quad \text{as } N \to \infty, \tag{4.4}$$

$$\sigma^2(x) \approx \sigma^2(X) \quad \text{as } N \to \infty. \tag{4.5}$$

This theorem bridges abstract theory and physical reality. As in Chapter 3, this Quarto document provides the interactive numerical verification for these claims.

Advanced Material 4.1: Deconstructing Theorem 4.1: This derivation assumes familiarity with sampling statistics and the general Central Limit Theorem (Advanced Material 4.4), which are actually beyond the scope of these notes.

1. The Mathematical Pillar: Let the sample mean of N independent trials of a random variable X be the random variable $Z = \mu(x)$. Utilizing Advanced Material 4.4:

- (4.37a) $\implies E[Z] = \mu(X)$
- (4.37b) $\implies V[Z] \to 0$ as $N \to \infty$

- (4.37c) $\implies$ $f_Z(z)$ concentrates into an infinitely narrow spike at $\mu(X)$

Consequently, the Weak Law of Large Numbers emerges directly:

$$\lim_{N\to\infty} P\left(|\mu(x) - \mu(X)| \leq \epsilon\right) = 1. \tag{4.6}$$

Similarly, the sample variance $\sigma^2(x)$ is an estimator with $E[\sigma^2(x)] \to \sigma^2(X)$ and $V[\sigma^2(x)] \to 0$ as $N \to \infty$, forcing its probability distribution to spike at $\sigma^2(X)$.
2. The Physical Pillar: Cournot's Principle (Postulate 3.1) translates mathematical probabilities of 1 into physical certainties. Thus, in physical practice, $\mu(x) \approx \mu(X)$ and $\sigma^2(x) \approx \sigma^2(X)$ for large N.

4.1.4 The Power of Random Variables: Examples

Comparing a fair system to a heavily biased system demonstrates the utility of mapping qualitative events to the real number line.

Example 1: The Fair Die vs. The Biased Die

For a standard, fair 6-sided die, let X be the number rolled. The distinct values are $x_k \in \{1, 2, 3, 4, 5, 6\}$, with PDF $f(x_k) = 1/6$.

$$\begin{aligned}\mu(X) &= 1(1/6) + 2(1/6) + 3(1/6) + 4(1/6) + 5(1/6) + 6(1/6) = 3.5\\ \sigma^2(X) &= (1-3.5)^2(1/6) + \cdots + (6-3.5)^2(1/6) \approx 2.92\end{aligned}$$

For a die tampered to roll only 5 or 6 with equal probability ($f(5) = 0.5$, $f(6) = 0.5$):

$$\begin{aligned}\mu(X) &= 5(0.5) + 6(0.5) = 5.5\\ \sigma^2(X) &= (5-5.5)^2(0.5) + (6-5.5)^2(0.5) = 0.25\end{aligned}$$

The mathematics immediately capture the physical reality: the theoretical mean shifts toward 5.5, and the variance shrinks to 0.25, reflecting tighter physical clustering.

Example 2: Colored Cards and Unlocking New Math

An urn contains cards: Red, Blue, Green, and Yellow. The draw probabilities are $P(\text{Red}) = 0.1$, $P(\text{Blue}) = 0.1$, $P(\text{Green}) = 0.1$, and $P(\text{Yellow}) = 0.7$.

Arithmetic operations are undefined for qualitative words. By defining a random variable X such that $x_1(\text{Red}) = 1$, $x_2(\text{Blue}) = 2$, $x_3(\text{Green}) = 3$, and $x_4(\text{Yellow}) = 4$, we can compute:

$$\mu(X) = 1(0.1) + 2(0.1) + 3(0.1) + 4(0.7) = 3.4$$

The result 3.4 correctly identifies the center near Yellow.

Important 4.1: The Power of Random Variables: Without mapping qualitative outcomes to $\mathbb{R}$ via random variables, computing structural measures like $\mu(X)$ or $\sigma^2(X)$ is mathematically impossible.

4.2 Continuous Distributions: Motivating via Histograms

The pedagogical approach of motivating continuous distributions via the geometric limits of histograms is deeply inspired by Ref. [1].

4.2.1 Motivating Continuous Probability

In Chapter 3, we established the connection between abstract probability and the physical world. Through Cournot's Principle (Postulate 3.1) and Theorem 3.10, we saw that the physical manifestation of probability an event is simply the relative frequency of occurrence of that event when the number of trials of the experiment is pushed to infinity.

For discrete random variables, finding this relative frequency is as simple as counting specific discrete outcomes. However, imagine an experimental setup—like a machine, a thermometer, or a sensor—that generates a continuous stream of values ranging anywhere from $-\infty$ to ∞. To extract the underlying behavior of this experimental output data, we group the data into intervals to create a histogram.

To reveal the exact nature of the machine's output, we draw a massive number of samples and progressively make the histogram bins thinner to gain more information. This transition can be explored interactively using this Quarto source document.

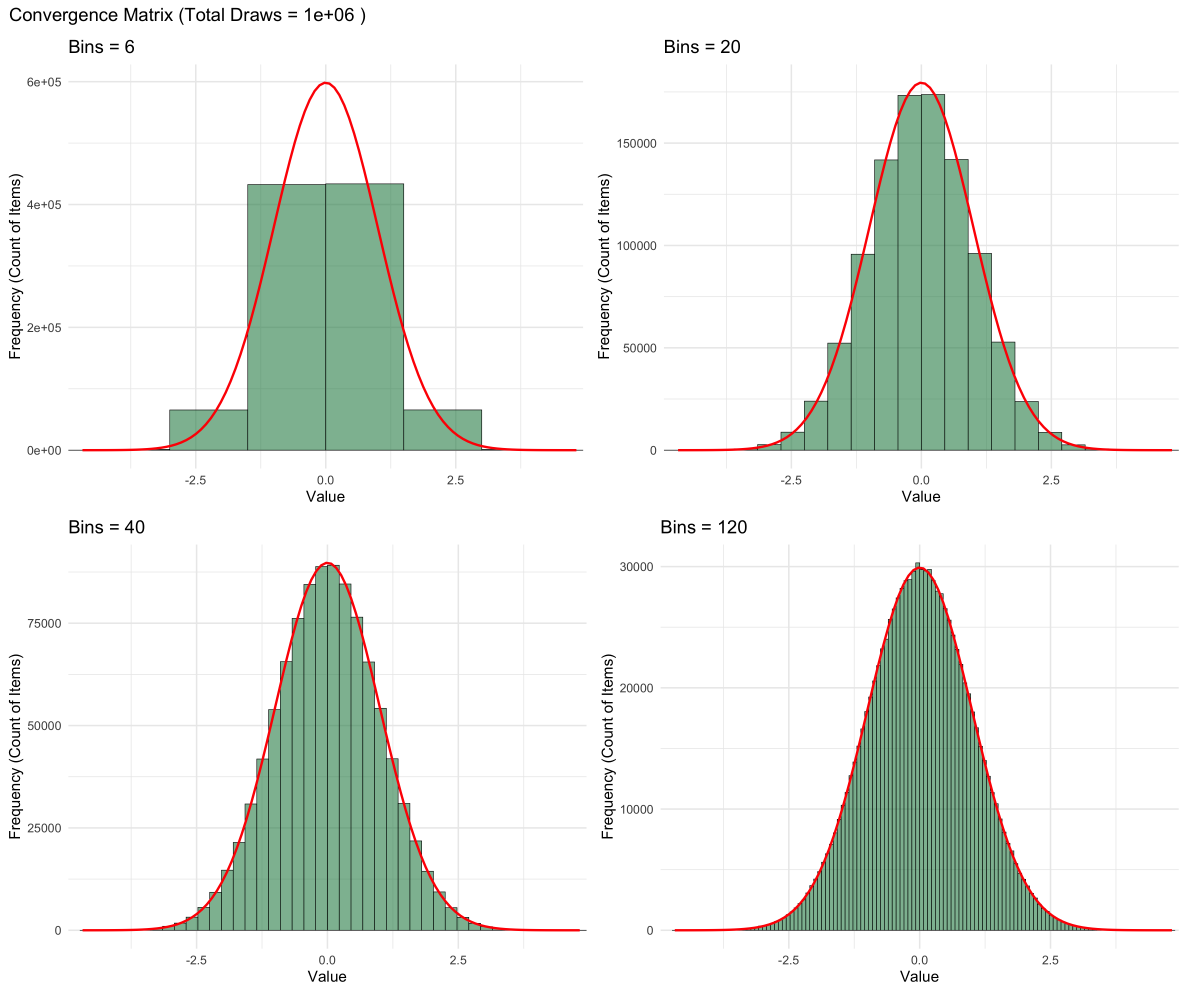

Figure 4.1: Empirical frequency histograms converging toward a smooth limiting continuous curve as the number of bins increases ($N = 1,000,000$).

As shown in the four subplots of Fig. 4.1, coarse bins yield a blocky structure. But as the bins become infinitely thin, a clear mathematical pattern emerges.

4.2.2 Extracting Features from the Histogram Limit

By observing the histogram as the bins shrink, we can tweeze out specific features that dictate how we mathematically construct a continuous probability distribution:

- **A Smooth Continuous Function:** As the bin width approaches zero, the jagged step-like perimeter of the histogram morphs into a smooth, unbroken curve. We call this boundary function the Probability Distribution Function (PDF), or $f(x)$.
- **Probability is Area:** In a histogram, the relative frequency of an interval is exactly the area of the bars within that interval. Because we established that probability *is* relative frequency, the probability of the variable falling between any two values is simply the (relative) area under the smooth curve $f(x)$ between those points[1].
- **Non-negativity:** A physical count or frequency can never be negative. Therefore, the histogram bars always extend upward, meaning our smooth curve must always sit on or above the horizontal axis.

4.2.3 Mathematical Foundation

To make these extracted features rigorous, we first introduce a notational shorthand to represent geometric area.

Definition 4.5 (Notation for Area)**:** For a continuous function $f(x)$, the total geometric area bounded between the curve and the horizontal axis, from a starting point $x = a$ to an ending point $x = b$, is denoted by the "integral" symbol:

$$\int_a^b f(x)\,dx \tag{4.7}$$

The sign of the area corresponding to a certain interval over x will be positive if the corresponding function value over that interval is positive, and vice versa.

In calculus, this symbol ($\int$) is known as the definite integral and possesses a rigorous definition. However, in this non-calculus-based text, we treat it exclusively to denote the area under a function $f(x)$.

Because the integral sign represents a continuous area, computing these boundaries manually is impossible without calculus. However, these operations can be executed numerically on a computer using the native `integrate()` function within RStudio. This process is illustrated in the following Quarto source document. With this notation established, we can formally define a continuous random variable.

Axiom 4.1 (Continuous Probability Distribution)**:** A continuous random variable X possesses a corresponding probability distribution function (PDF) $f(x)$ that satisfies the

[1] By relative area, we mean the area under $f(x)$ between the two points divided by the the total area under $f(x)$.

following conditions:

1. **Non-negativity:** $f(x) \geq 0$ for all $x \in \mathbb{R}$.
2. **Normalization:** The total area under the curve is 1.
$$\int_{-\infty}^{\infty} f(x)\,dx = 1 \tag{4.8}$$
3. **Interval Probability:** The probability that X takes a value between a and b is the area under $f(x)$ from a to b.
$$P(a \leq X \leq b) = \int_a^b f(x)\,dx \tag{4.9}$$

An immediate physical consequence of this definition is that the probability of a continuous random variable taking one exact, specific value is exactly zero (i.e., $P(X = a) = \int_a^a f(x)\,dx = 0$, because a line has no area).

Note the close resemblance between the continuous probability Axioms 4.1 and Kolmogorov's axioms for discrete events (Axioms 3.1). Upon comparing them, a mathematical dictionary emerges, which lets us practically transition from the discrete case to the continuous one. The dictionary dictates us to make the following replacement (with the limits on the sum and the integral appropriately determined)

$$\sum [\ldots] f(x_k) \quad \longrightarrow \quad \int [\ldots] f(x)\,dx \tag{4.10}$$

By applying this correspondence rule directly to our earlier discrete definitions of expected value (Definition 4.3) and variance (Definition 4.4), we immediately generate their continuous counterparts.

Definition 4.6 (Continuous Expected Value)**:** The expected value (or theoretical mean) $\mu(X)$ of a continuous random variable X is defined as:
$$\mu(X) \equiv E[X] = \int_{-\infty}^{\infty} x f(x)\,dx \tag{4.11}$$

Definition 4.7 (Continuous Theoretical Variance)**:** The theoretical variance $\sigma^2(X)$ of a continuous random variable X is defined as:
$$\sigma^2(X) \equiv \mathrm{Var}(X) = \int_{-\infty}^{\infty} (x - \mu(X))^2 f(x)\,dx \tag{4.12}$$

Just as in the discrete framework, $\mu(X)$ quantifies the central tendency or typical expected outcome of the random variable, whereas the $\sigma^2(X)$ quantifies its dispersion.

Example: The Uniform Distribution

Consider an idealized roulette wheel. A ball is spun and can land at any exact angle θ between 0 and 2π radians. Because the physical wheel is perfectly symmetrical, no angle is favored,

meaning the distribution function $f(\theta)$ must be a flat, constant line across the entire interval $[0, 2\pi]$.

To satisfy the normalization rule, the total area under this constant curve must equal 1. Geometrically, this forms a rectangle with a width of 2π. To find the constant height, we solve:

$$\begin{aligned} \text{Area} &= \text{width} \times \text{height} \\ 1 &= 2\pi \times f(\theta) \\ f(\theta) &= \frac{1}{2\pi} \end{aligned}$$

Thus, the PDF is mathematically formulated as:

$$f(\theta) = \begin{cases} 1/(2\pi) & \text{for } 0 \leq \theta \leq 2\pi \\ 0 & \text{otherwise} \end{cases} \tag{4.13}$$

The probability of the ball landing in any specific sector is simply the area of the rectangle over that specific interval.

4.3 Some Special Probability Distributions

4.3.1 The Normal Distribution

Definition 4.8 (The Normal Distribution)**:** A continuous random variable X follows a normal distribution if its probability distribution function (PDF) is governed by two parameters, μ and σ, such that:[a]

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \quad \text{for } -\infty < x < \infty. \tag{4.14}$$

We denote this mathematically as $X \sim N(\mu, \sigma)$. When $\mu = 0$ and $\sigma = 1$, it is referred to as the standard normal distribution, denoted as $N(0, 1)$.

[a] $\exp(x) \equiv e^x$ where $e \sim 2.7182$ is the irrational Euler's number.

Theorem 4.2 (Mean and Variance of the Normal Distribution)**:** For a normal random variable $X \sim N(\mu, \sigma)$, the theoretical expected value and variance are exactly its parameters:

$$\mu(X) = E[X] = \mu, \tag{4.15}$$
$$\sigma^2(X) = \text{Var}(X) = \sigma^2. \tag{4.16}$$

The proof requires calculus to evaluate the continuous integrals and is omitted here. The interested reader may refer to Ref. [3]. We can, however, verify this computationally by executing numerical integration in RStudio, as demonstrated in this Quarto source document.

For any normal distribution, the probability (area) to the left of a specific value x depends entirely on its relative distance from the mean. A formal statement requires us to first define Z-score.

Definition 4.9 (Z-Score)**:** The Z-score for a value x of the normal random variable $X \sim N(\mu, \sigma)$ is:

$$Z(x, \mu, \sigma) \equiv \frac{x - \mu}{\sigma}. \tag{4.17}$$

Theorem 4.3: For a normal random variable $X \sim N(\mu, \sigma)$, the area under the PDF of $N(\mu, \sigma)$ from $-\infty$ to x (the probability $P(X \leq x)$) depends only on the parameter $Z(x, \mu, \sigma) = \frac{x-\mu}{\sigma}$.

The reader is referred to Ref. [3] for a proof of Theorem 4.3. Because the area under the curve depends exclusively on Z, computing probabilities becomes mathematically streamlined using `pnorm()`, a built-in function in RStudio. Evaluating these areas (and the associated probabilities) gives us the widely used empirical rule:

- The probability of X falling within 1 standard deviation around the mean $P(-1 \leq Z \leq 1)$—is approximately 0.68 (or 68%).
- The probability of X falling within 2 standard deviations around the mean $P(-2 \leq Z \leq 2)$—is approximately 0.9545 (or 95.45%). 95% corresponds to approximately 1.96 standard deviations.
- The probability of X falling within 3 standard deviations around the mean $P(-3 \leq Z \leq 3)$—is approximately 0.997 (or 99.7%). 99% corresponds to approximately 2.58 standard deviations.

This is demonstrated again computationally in the same Quarto document as above.

4.3.2 The Bernoulli Distribution

Many physical experiments and observational trials possess exactly two possible qualitative outcomes. These binary systems are generically categorized as yielding either a "success" or a "failure." Common examples include tossing a coin (Heads or Tails), evaluating medical outcomes (Living or Dying), or performing quality control (Defective or Non-defective).

To analyze these systems mathematically, we associate these qualitative outcomes with numerical values using a discrete random variable X. By standard convention, a failure is mapped to the value 0, and a success is mapped to the value 1.

Definition 4.10 (The Bernoulli Distribution)**:** A discrete random variable X follows a Bernoulli distribution if it takes exactly two numerical values, 1 (success) and 0 (failure), with the probability of success governed by a single parameter p (where $0 \leq p \leq 1$). Its probability distribution function $f(x)$ is defined strictly as:

$$f(1) = P(X = 1) = p \tag{4.18}$$
$$f(0) = P(X = 0) = 1 - p \tag{4.19}$$

Theorem 4.4 (Mean and Variance of a Bernoulli Random Variable)**:** For a Bernoulli random variable X governed by a success probability p, the theoretical expected value

and variance are exactly:

$$\mu(X) = p \tag{4.20}$$
$$\sigma^2(X) = p(1-p) \tag{4.21}$$

Proof. Using Definitions 4.3 and 4.4, we sum across the two possible outcomes ($x_1 = 0$, $x_2 = 1$). The expected value evaluates to:

$$\mu(X) = \sum_{k=1}^{2} x_k f(x_k) = 0 \cdot (1-p) + 1 \cdot p = p \tag{4.22}$$

Substituting $\mu(X) = p$ into the variance formula yields:

$$\begin{aligned}\sigma^2(X) &= \sum_{k=1}^{2} (x_k - \mu(X))^2 f(x_k) \\ &= (0-p)^2(1-p) + (1-p)^2 p \\ &= p^2(1-p) + (1-p)^2 p \\ &= p(1-p)[p + (1-p)] \\ &= p(1-p)\end{aligned}$$

□

4.3.3 The Geometric Distribution

Building upon the single binary trials discussed in Section 4.3.2, imagine an experiment consisting of a sequence of independent, identical trials. Each trial yields either a "success" or a "failure," with the probability of a success remaining a constant p (where $0 < p \leq 1$).

We define a new discrete random variable X as the trial number on which the *first* success occurs. This situation is formalized by the geometric distribution.

Theorem 4.5 (The Geometric Distribution)**:** Let X be a discrete random variable representing the number of trials required to obtain the first success in a sequence of independent Bernoulli trials with success probability p. The probability distribution function $f(x)$, the mean $\mu(X)$, and the variance $\sigma^2(X)$ are given by:

$$f(x) = (1-p)^{x-1}p, \quad \text{for } x \in \{1, 2, 3, \dots\}, \tag{4.23}$$
$$\mu(X) = \frac{1}{p}, \tag{4.24}$$
$$\sigma^2(X) = \frac{1-p}{p^2}. \tag{4.25}$$

Proof. We shall provide the proofs of only the distribution and the mean. Let $q = 1 - p$.

1. The Distribution Function: For the first success to manifest on the x-th trial, the preceding $x - 1$ trials must be failures. Because the trials are independent, the probability of this sequence is the product of their individual probabilities.

$$f(x) = \underbrace{q \cdot q \cdots q}_{x-1 \text{ times}} \cdot p = q^{x-1}p. \tag{4.26}$$

2. Mean: Using the definition of the expected value for a discrete random variable, we evaluate the infinite sum:

$$\mu(X) = E[X] = \sum_{x=1}^{\infty} xf(x) = \sum_{x=1}^{\infty} xq^{x-1}p = p\sum_{x=1}^{\infty} xq^{x-1}. \tag{4.27}$$

Recognizing the summation as an arithmetico-geometric series[2], we substitute the closed-form result:

$$\mu(X) = p\left(\frac{1}{(1-q)^2}\right) = p\left(\frac{1}{p^2}\right) = \frac{1}{p}. \tag{4.28}$$

□

Recognize that the above results are totally consistent with the intuitive expectation that the mean has to be greater than one.

4.3.4 The Binomial Distribution

Consider a composite experiment consisting of a fixed number, n, of independent Bernoulli trials (Section 4.3.2). For each trial, the probability of success is a constant p, and the probability of failure is $1-p$. We define the discrete random variable X as the total number of successes observed in the n trials.

Theorem 4.6 (The Binomial Distribution)**:** Let X be the number of successes in n independent Bernoulli trials with success probability p. The probability distribution function $f(x)$, theoretical expected value $\mu(X)$, and theoretical variance $\sigma^2(X)$ are given by:

$$f(x) = \binom{n}{x} p^x(1-p)^{n-x}, \quad \text{for } x \in \{0, 1, 2, \ldots, n\}, \tag{4.29}$$

$$\mu(X) = np, \tag{4.30}$$

$$\sigma^2(X) = np(1-p). \tag{4.31}$$

Proof. We prove the functional form of the distribution and its expected value.

1. The Distribution Function: By the general multiplication theorem for independent events (Theorem 3.12), the probability of observing any single specific sequence containing exactly x successes and $n-x$ failures is $p^x(1-p)^{n-x}$. From Theorem 3.17, the number of distinct ways to arrange x successes among n trials is given by the combination $\binom{n}{x}$. Since these specific sequences represent mutually exclusive composite events, we invoke Kolmogorov's additivity axiom (Eq. (3.6) of Axiom 3.1) to obtain the total probability by summing them:

$$f(x) = \binom{n}{x} p^x(1-p)^{n-x}. \tag{4.32}$$

[2]The closed-form sum utilized here is $\sum_{x=1}^{\infty} xq^{x-1} = \frac{1}{(1-q)^2}$ for $|q| < 1$, which is obtained by taking the first derivative of the standard geometric series $\sum_{x=0}^{\infty} q^x = \frac{1}{1-q}$ with respect to q.

2. The Expected Value:

$$\begin{aligned}
\mu(X) &= \sum_{x=0}^{n} x\binom{n}{x}p^x(1-p)^{n-x} && \text{(From Def. 4.3)}\\
&= \sum_{x=1}^{n} x\binom{n}{x}p^x(1-p)^{n-x} && \text{(The } x=0 \text{ term vanishes)}\\
&= \sum_{x=1}^{n} n\binom{n-1}{x-1}p^x(1-p)^{n-x} && \left(\text{Using } x\binom{n}{x} = n\binom{n-1}{x-1}\right)\\
&= np\sum_{x=1}^{n} \binom{n-1}{x-1}p^{x-1}(1-p)^{n-x} && \text{(Rearranging terms)}\\
&= np\sum_{y=0}^{m} \binom{m}{y}p^{y}(1-p)^{m-y} && \text{(Substituting } y=x-1 \text{ and } m=n-1)\\
&= np(p+(1-p))^m && \text{(Applying the Binomial theorem)}\\
&= np. && \text{(Since } p+1-p=1) \qquad (4.33)
\end{aligned}$$

□

4.4 Central Limit Theorem

In observational and physical sciences, we rarely have the capacity to measure an entire population. Instead, we must draw a finite subset of observations, known as a sample, to infer the underlying properties of that population.

Consider a population where a certain attribute is present with a true, but often unknown, probability of success p. We draw a sample of size N, representing N independent observations or trials. If we define a discrete random variable X as the total number of observed successes within this sample, we can define the sample proportion $\hat{p}$ as the ratio of successes to the total sample size:

$$\hat{p} \equiv \frac{X}{N}. \tag{4.34}$$

Because X is a random variable, the sample proportion $\hat{p}$ is also a random variable governed by its own probability distribution. The Central Limit Theorem (CLT) dictates that as the sample size N grows, the distribution of $\hat{p}$ converges to a normal distribution, regardless of the underlying discrete nature of the individual binary trials.

Theorem 4.7 (Central Limit Theorem)**:** For N independent observations drawn from a population with a true success probability p, let $\hat{p}$ be the sample proportion. If the sample size is sufficiently large, the random variable $\hat{p}$ tends to a **normal distribution** [1]. The theoretical mean and variance of this limiting distribution are given by:

$$\mu(\hat{p}) = p, \tag{4.35}$$

$$\sigma^2(\hat{p}) = \frac{p(1-p)}{N}. \tag{4.36}$$

Advanced Material 4.2: General Formulation of the CLT: The generalized CLT applies to a sequence of independent random variables $X_1, \ldots, X_N$, where each possesses a theoretical mean $\mu(X_i)$ and theoretical variance $\sigma^2(X_i)$ [3].
For the sample mean random variable $Z = \frac{1}{N}\sum_{i=1}^{N} X_i$, its expectation $E[Z]$, variance $V[Z]$, and limiting probability density function $f_Z(z)$ are given by:

$$E[Z] = \frac{1}{N}\sum_{i=1}^{N} \mu(X_i), \tag{4.37a}$$

$$V[Z] = \frac{1}{N^2}\sum_{i=1}^{N} \sigma^2(X_i), \tag{4.37b}$$

$$\lim_{N\to\infty} f_Z(z) = \frac{1}{\sqrt{2\pi V[Z]}} \exp\left(-\frac{(z - E[Z])^2}{2V[Z]}\right). \tag{4.37c}$$

This rigorous derivation requires the algebra of sums and transformations of random variables, falling outside the scope of this course [3].

4.5 Exercises

The numerical computations and graphing components of these problems are implemented and can be interactively verified in this accompanying Quarto document.

1. Consider a coin toss where the probability of obtaining heads is p. Define a discrete random variable X that equals 0 for heads and 1 for tails.
   (a) Compute the mean and variance.
   (b) Graph the mean and variance using RStudio. Does the graph match your intuitive expectations (like zero variance for a totally biased coin)?

2. Determine whether each of the following mathematical functions can serve as a valid PDF for a continuous random variable.
   (a) $f(x) = x^2$ for $-\infty < x < \infty$.
   (b) $f(x) = x$ for $0 \leq x \leq 2$, and 0 otherwise.
   (c) $f(x) = x^3$ for $-1 \leq x \leq 1$, and 0 otherwise.

3. Consider a roulette wheel where the ball lands at an angle $\theta \in [0, 2\pi)$. The probability of landing in the interval $[0, \pi/2)$ is $1/2$, and in $[\pi/2, \pi)$ is $1/4$. The landing distribution is uniform within each of these sectors, and also uniformly distributed in the remaining sector $[\pi, 2\pi)$.
   (a) Write the complete probability distribution function $f(\theta)$.
   (b) Compute the probability that the ball lands in the interval $[\pi/4, 3\pi/2]$.

4. A continuous random variable X has a probability distribution function $f(x)$ defined on $[0, 10]$. The function increases linearly from $f(0) = 0$ to a constant value c at $x = 2$, remains constant at c from $x = 2$ to $x = 6$, and decreases linearly to $f(10) = 0$.

(a) Determine the value of c.

(b) Write the explicit piecewise form of $f(x)$.

(c) Compute $P(1 \leq X \leq 5)$.

5. Consider a continuous function $f(x)$ defined on the interval $[-1, 0]$ that forms a triangle with the horizontal axes and passes through the origin, given by:

$$f(x) = -4x \tag{4.38}$$

Assume $f(x) = 0$ outside this interval.

(a) Is $f(x)$ a valid probability distribution function?

(b) Propose a new function $g(x) = \alpha f(x)$. Determine the constant α such that $g(x)$ becomes a valid PDF.

6. Let $X_1 \sim N(50, 10)$ and $X_2 \sim N(80, 15)$ be two normally distributed random variables.

(a) Determine the value x_2 such that $P(X_2 \leq x_2) = P(X_1 \leq 65)$.

(b) Verify your result computationally using RStudio.

7. A recent college graduate is applying for jobs. Assume that each job application is an independent trial, and the probability of receiving a job offer from any single application is a constant $p = 0.05$. Let the discrete random variable X represent the number of applications submitted up to and including the first successful job offer.

(a) Identify the specific probability distribution that models X. Write its explicit probability distribution function $f(x)$.

(b) What is the chance of success in the third job application?

(c) What is the chance of first success in the third application?

(d) What is the chance of at least one success in three applications?

(e) What is the chance of at least one success in the first four applications?

(f) Calculate the minimum number of applications required such that the probability of receiving at least one job offer is greater than 50%.

(g) What is the average number of applications the graduate must submit to receive their first job offer?

8. A university is sending a sports squad of 6 players to a competition that strictly requires 4 players to field a team. Assume the health status of each player is independent, and the probability of any individual player falling sick (being unfit to play) is 0.1. Let the discrete random variable X represent the number of sick players.

(a) Identify the specific probability distribution that models X, and write its explicit probability distribution function $f(x)$.

(b) Calculate the probability that exactly zero players fall sick.

(c) Calculate the probability that exactly one player falls sick.

(d) Calculate the probability that exactly two players fall sick. Compute it using two equivalent approaches

i. Define "success" as a player falling sick.

ii. Define “success” as a player remaining fit.

(e) Calculate the overall probability that the squad will be able to successfully field a full team of at least four fit players for the competition.

(f) What is the mean number of players falling sick?

Chapter 5

Statistics

For this chapter, we will simply refer to Chapter 5 of Ref. [1].

5.1 Exercises

1. A specific genetic trait is known to exist in exactly $p = 0.15$ of a country's population, representing the true population proportion. A researcher, who does not have access to this true value, takes a random sample of $n = 500$ individuals and finds that 70 of them have the genetic trait. Assume the sample size is large enough that the Central Limit Theorem applies to the sampling distribution.

   (a) Calculate the point estimate, $\hat{p}$, for the sample proportion.

   (b) State the approximate probability distribution of $\hat{p}$, and specify its theoretical mean ($\mu_{\hat{p}}$) and true standard deviation ($\sigma_{\hat{p}}$).

   (c) Calculate the standard error, which represents the researcher's approximation of the true standard deviation using only their sample proportion.

   (d) Briefly explain why the true standard deviation from part (b) and the standard error from part (c) differ.

   (e) Calculate the Z-scores required to find the probability that the sample proportion falls between 0.14 and 0.16.

   (f) Let $\Phi(z)$ represent the cumulative area to the left of a given Z-score under the standard normal curve. Given the numerical values $\Phi(-0.63) \approx 0.2643$ and $\Phi(0.63) \approx 0.7357$, calculate the approximate probability:

   $$P(0.14 < \hat{p} < 0.16).$$

2. An academic coordinator is analyzing university records and notes that exactly 15% of all the students have failed at least one course. A student researcher randomly samples $n = 70$ students and calculates the sample proportion, $\hat{p}$, of those who failed at least one course. To study the behavior of these estimates, the researcher theoretically repeats this sampling process $1,000$ times to build a distribution of sample proportions. Assume the conditions for the Central Limit Theorem are met.

   (a) What would you expect the shape of this distribution of sample proportions to be?

   (b) Calculate both the true standard deviation of the sampling distribution and the standard error (approximation to the standard deviation) the researcher would approximate from a single sample if they found 12 students who had failed.

(c) Suppose the student researcher changes their methodology and samples $n = 140$ students per iteration instead of 70. Calculate the true standard deviation of this new distribution.

(d) Comparing your answers from part (b) and part (c), mathematically explain why increasing the sample size is a good idea.

3. A city health department is conducting a public health survey to estimate the proportion of adults in their city who have been formally diagnosed with diabetes. The true population proportion, p, is unknown. The researchers collect a random sample of $n = 1000$ adults and find that 120 of them report having a diabetes diagnosis. Assume that the central limit theorem is applicable.

(a) Calculate the point estimate, $\hat{p}$, for the proportion of adults with diabetes in the sample.

(b) Calculate the standard deviation of the sample proportion $\hat{p}$. Since p is unknown, feel free to approximate p with $\hat{p}$.

(c) Construct a 95% confidence interval. For a normal distribution, the area between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$ is 0.95.

(d) Construct a 99% confidence interval. For a normal distribution, the area between $\mu - 2.58\sigma$ and $\mu + 2.58\sigma$ is 0.99.

4. An international relations institute is conducting a survey in Country B to assess public awareness of Peru. The survey includes a question asking respondents to identify the capital of Peru, providing three options: Lima (the correct answer) and two incorrect distractors.

A baseline proportion of exactly $p = 1/3$ represents random guessing by a population that lacks knowledge or interest. A true proportion greater than $1/3$ indicates active geographic awareness, whereas a proportion less than $1/3$ suggests the presence of systematic misinformation. A researcher surveys a random sample of $n = 1200$ residents, and 480 of them select the correct capital. Assume the sample size is large enough that the Central Limit Theorem applies.

(a) Calculate the point estimate, $\hat{p}$, for the sample proportion. Using $\hat{p}$ as a substitute for the unknown true proportion p, calculate the approximate variance of the sample proportion.

(b) Construct a 95% confidence interval for the true population proportion. For a normal distribution, the area between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$ is 0.95. Does the evidence support the hypothesis of active awareness, systematic misinformation, or random guessing?

(c) Write the explicit algebraic expression for the total size of this interval. Explain how the size of the confidence interval is affected by an increase in the required confidence level (e.g., shifting from 95% to 99%, which increases the Z-score) and an increase in the sample size, n.

5. In a country-wide election between Candidate A and Candidate B, an exit poll strategist wants to determine the voting population's preference. We establish three hypotheses regarding the true population proportion p of voters who choose Candidate A:

- The population is evenly divided ($p = 0.5$).

- The population leans towards Candidate A ($p > 0.5$).
- The population leans towards Candidate B ($p < 0.5$).

The researcher surveys a random sample of $n = 900$ voters, and 414 of them state they voted for Candidate A. Assume the sample size is large enough for the Central Limit Theorem to apply.

(a) **Confidence Interval Method:** Calculate the sample proportion, $\hat{p}$, and its standard deviation.

(b) Construct a 95% confidence interval for p. For a normal distribution, the area between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$ is 0.95.

(c) Based solely on the confidence interval from part (b), which of the three hypotheses is most likely?

(d) **p-value Method:** Assume the hypothesis that the population is evenly divided is true. Calculate the standard error of the sample proportion under this assumption.

(e) Compute the Z-score corresponding to the observed sample proportion.

(f) Let $\Phi(z)$ represent the cumulative area to the left of a given Z-score under the standard normal curve. Given the numerical values $\Phi(-2.10) \approx 0.0179$, $\Phi(-2.40) \approx 0.0082$, and $\Phi(-2.70) \approx 0.0035$, calculate the p-value.

(g) Based on your computed p-value, what do you conclude? Which of the three hypotheses is most likely?

Bibliography

[1] D.M. Diez, C.D. Barr, and M. Çetinkaya-Rundel. *OpenIntro Statistics*. OpenIntro, 2019. ISBN 9781943450268. URL https://books.google.com.pe/books?id=yZLg0AEACAAJ.

[2] E. Kreyszig. *Advanced Engineering Mathematics*. John Wiley & Sons, 2010. ISBN 9780470458365. URL https://books.google.com.pe/books?id=UnN8DpXI74EC.

[3] K.F. Riley, M.P. Hobson, and S.J. Bence. *Mathematical Methods for Physics and Engineering: A Comprehensive Guide*. Cambridge University Press, 2006. ISBN 9781139450997. URL https://books.google.com.pe/books?id=Mq1nlEKhNcsC.

[4] D.S. Sivia and J. Skilling. *Data Analysis: A Bayesian Tutorial*. Oxford science publications. OUP Oxford, 2006. ISBN 9780198568315. URL https://books.google.com.pe/books?id=lYMSDAAAQBAJ.

[5] R.E. Walpole, R.H. Myers, S.L. Myers, and K.E. Ye. *Probability & Statistics for Engineers & Scientists, Global Edition*. Pearson Education, 2016. ISBN 9781292161419. URL https://books.google.com.pe/books?id=Th3aDAAAQBAJ.