Probability and Statistics Notes

Matthew Tan

April 11, 2024

Contents

1 Probability Theory

1.1 Probability Basics

Venn Diagrams

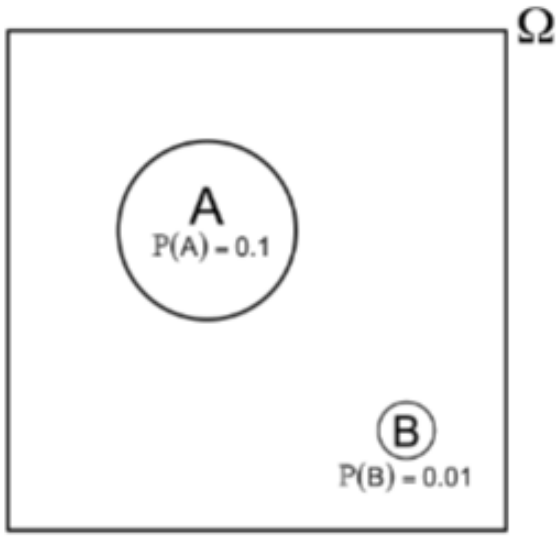


Figure 1.1: Venn Diagrams

Definition 1.1 (Sample Space): The set Ω is called the sample space. It contains all possible (primitive) outcomes we are considering.

Definition 1.2 (Event): An event is a subset of Ω, including Ω itself.

Useful results from the Venn Diagram

1. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$
   Proof. Since A either happens or it does not happen, $\mathbb{P}(A \cup A^c) = \mathbb{P}(\Omega) = 1$.
   Since A and A^c are disjoint, then $\mathbb{P}(A \cup A^c) = \mathbb{P}(A) + \mathbb{P}(A^c)$.
   Therefore, $\mathbb{P}(A) + \mathbb{P}(A^c) = 1 \implies \mathbb{P}(A^c) = 1 - \mathbb{P}(A)$
2. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
3. $\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c)$
4. $\mathbb{P}(A^c \cap B^c) = 1 - \mathbb{P}(A \cup B)$

Further rules for probability

1. Bounds on Probabilities Rule: $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$
2. If B logically entails A, then $\mathbb{P}(A) \geq \mathbb{P}(B)$

Definition 1.3 (Event Space): An event space $\mathcal{F}$ is a set of subsets of Ω which must satisfy certain properties. The event space defines the set of all describable events to which we want to assign certain properties.

Example.
Imagine the roll of a fair dice where the sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$ and the event space $\mathcal{F}$ is a set

composed of all subsets of this (there are 2^6) subsets.
For instance, $A = \{6\}$ is the event 'the result is 6' and $B = \{2, 4, 6\}$ is the event 'the result is even'. The events A, B we've described are subsets of Ω so they are in $\mathcal{F}$. An actual outcome like rolling a 2 can correspond to several events.
Consider the experiment of tossing a coin.

$$\Omega = \{H, T\}$$
$$\mathcal{F} = \{H, T, \{H, T\}, \emptyset\}$$

In English, the event space is "A head is thrown", "A tail is thrown", "either a head or tail is thrown", "neither a head or tail is thrown". The number of describable events is 2^2.

Kolmogorov Axioms of Probability

Definition 1.4 (Probability): A probability is a function $\mathbb{P}$ that satisfies, for all events in $\mathcal{F}$:

1. **Axiom 1:** $\mathbb{P}(A) \geq 0$ for all $A \in \mathcal{F}$
2. **Axiom 2:** $\mathbb{P}(\Omega) = 1$
3. **Axiom 3:** If $A_1, A_2, \cdots \in \mathcal{F}$ are mutually disjoint, then $\mathbb{P}(A_1 \cup A_2 \cup \dots) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \dots$

Revised Axioms after introducing conditional probability:

1. **Axiom 1:** $0 \leq \mathbb{P}(A|B) \leq 1$
2. **Axiom 2:** $\mathbb{P}(B|B) = 1$
3. **Axiom 3:** If $A_1, A_2, \dots$ are mutually exclusive given B then

$$\mathbb{P}(A_1 \cup A_2 \cup \dots | B) = \mathbb{P}(A_1|B) + \mathbb{P}(A_2|B) + \dots$$

Some proofs of further results using axioms:

1. **The Complement Rule:** $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$
   Proof. Since A and A^c are mutually disjoint and A either happens or it does not, $A \cup A^c = \Omega$ and $A \cap A^c = \emptyset$. Hence,

$$\begin{aligned} \mathbb{P}(A \cup A^c) = \mathbb{P}(\Omega) = 1 \quad & \text{[Axiom 2]} \\ \rightarrow \mathbb{P}(A) + \mathbb{P}(A^c) = 1 \quad & \text{[Axiom 3]} \\ \rightarrow \mathbb{P}(A^c) = 1 - \mathbb{P}(A) \quad & \blacksquare \end{aligned}$$

2. **The Probability of the Union of Two Events Rule:** $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
   Proof. The probability that A or B happens is the probability of each happening minus the probability of a joint occurrence. We know that

$$A \cup (A^c \cap B) = (A \cup B) \cap (A \cup A^c) = A \cup B$$

   So $A \cup B$ can be expressed as a union of 2 disjoint sets. By axiom 3,

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

   But $B = B \cap (A \cup A^c) = (B \cap A) \cup (B \cap A^c)$ which is also the union of 2 disjoint sets. So by axiom 3,

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

   Therefore by substituting,

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \quad \blacksquare$$

3. **The Bounds on Probabilities Rule:** $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$
Proof. From (b),
$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$
By axiom 1,
$$\mathbb{P}(A \cap B) \geq 0$$
Therefore
$$\begin{aligned} \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) &\leq \mathbb{P}(A) + \mathbb{P}(B) \\ \mathbb{P}(A \cup B) &\leq \mathbb{P}(A) + \mathbb{P}(B) \end{aligned}$$ ■

4. **The logical consequence rule:** If B logically entails A, then $\mathbb{P}(B) \leq \mathbb{P}(A)$.
Proof. AFC that if B logically entails A where $B \subseteq A$, then $\mathbb{P}(A) < \mathbb{P}(B)$. So under the circumstance that event B fully coincides with A, this implies that $\mathbb{P}(B) = \mathbb{P}(A \cup (A^c \cap B)) = \mathbb{P}(A) + \mathbb{P}(A^c \cap B)$, where $\mathbb{P}(A^c \cap B) > 0$ [Axiom 3]. So there exists an event where A does not happen yet B happens, so $B \not\subseteq A$, a contradiction.
Hence, if B logically entails A, it is not the case that $\mathbb{P}(B) > \mathbb{P}(A)$ From this, negation we conclude $\mathbb{P}(A) \geq \mathbb{P}(B)$ ■.

Alternatively, since $B \subseteq A$. Then $\mathbb{P}(B) = \mathbb{P}(A \cap B)$. So,
$$\begin{aligned} \mathbb{P}(A) &= \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c) \\ &= \mathbb{P}(B) + \mathbb{P}(A \cap B^c) \end{aligned}$$
From axiom 1, $\mathbb{P}(A \cap B^c) \geq 0$. Thus,
$$\mathbb{P}(A) = \mathbb{P}(B) + \mathbb{P}(A \cap B^c) \geq \mathbb{P}(B)$$ ■

1.2 Conditional Probability

Definition 1.5 (Conditional Probability): $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$

In general, $\mathbb{P}(A|B) \neq \mathbb{P}(B|A)$

Theorem 1.1 (Law of Total Probability): Given a partition $\{B_1, B_2\}$ of Ω then for any event A
$$\begin{aligned} \mathbb{P}(A) &= \mathbb{P}(A \cap B_1) + \mathbb{P}(A \cap B_2) \\ &= \mathbb{P}(A|B_1)\mathbb{P}(B_1) + \mathbb{P}(A|B_2)\mathbb{P}(B_2) \end{aligned}$$

1.3 Independence & Bayes' Rule

Definition 1.7 (Independence): Events A and B are mutually independent if

1. $\mathbb{P}(A|B) = \mathbb{P}(A)$
2. $\mathbb{P}(B|A) = \mathbb{P}(B)$
3. $\mathbb{P}(A)\mathbb{P}(B) = \mathbb{P}(A \cap B)$

Explanation: Since independence implies how knowing about one event tells you nothing about the other, this essentially means
$$\mathbb{P}(A|B) = \mathbb{P}(A)$$
Given the definition of conditional probability,
$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \mathbb{P}(A)$$

which implies

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$$

Theorem 1.2 (Bayes' Rule):

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}, \text{ where } \mathbb{P}(B) > 0$$

We can derive this through definitions of conditional probabilities where

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

Rearranging and substituting we have

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

The intuition is that we must compare the probability of an unlikely events with other unlikely events.
Note: For the $\mathbb{P}(B)$ we tend to use law of total probability to compute it even if it is given directly.
Bayes' Rule shows how to update probabilities given evidence. If H denotes a hypothesis and E denotes some evidence, then by Bayes' Rule:

$$\mathbb{P}(H|E) = \frac{\mathbb{P}(E|H)\mathbb{P}(H)}{\mathbb{P}(E)}$$

where

1. $\mathbb{P}(H)$ is the prior probability of H that was formed before E became available
2. $\mathbb{P}(E|H)$ is the likelihood of evidence: it is the conditional probability of seeing E if H is true
3. $\mathbb{P}(E)$ is the probability of E: the unconditional probability of witnessing the evidence
4. $\mathbb{P}(H|E)$ is called the posterior probability of H given E

Worked Example.
This is an example of the Sally Clark child. Some facts: only 30 out of 650,000 annual births in England, Scotland and Wales were known to have been murdered by their mothers; double murders is 3 out of 650,000.
The probability of a child dying from SIDS: 1 in 8500.
It is wrong to assume that two children dying are independent and therefore you cannot calculate $\mathbb{P}$(1st and second child died) by multiplying the probability together. Instead, from conditional probability

$$\mathbb{P}(C_1)\mathbb{P}(C_2|C_1) = \mathbb{P}(C_1 \cap C_2)$$

The object of interest is the probability that both children died of SIDS given that they died suddenly and unexpectedly: $\mathbb{P}(H|E)$. The hypothesis of interest is Sally Clark's two children died of SIDS and the evidence is that both children died suddenly and unexpectedly. Thus, we have

$$\mathbb{P}(H|E) = \frac{\mathbb{P}(E|H)\mathbb{P}(H)}{\mathbb{P}(H)\mathbb{P}(E|H) + [1 - \mathbb{P}(H)]\mathbb{P}(H|E^c)}$$

Pluck in the numbers to solve.

2 Data, Variable & Moments

2.1 Data Basics

Types of data

1. Categorical
   (a) Qualitative: Often binary indicators (0 = unemployed, 1 = employed)
   (b) Ordinal: the ordering of data is meaningful but difference/intervals are not, essentially ranking (0 = poor, 1 = fair, 2 = good)
2. Numerical (Quantitative). Numerical value is cardinally meaningful
   (a) Discrete: number of individuals in a household
   (b) Continuous: height, handspan, age. Many variables are treated as if they are continuous despite being discrete like wages

Definition 2.1 (Support): The set of values which a variable can take is known as its support

1. For discrete data, it is usually integers within some range ({0,1} for binary indicators)
2. for a continuous variable, the support is any real number often within an interval

We may sometimes measure several features of an object, and these are multivariate variables.

Definition 2.2 (Cross Section Data): Cross-sectional data sets have one observation per unit, observing multiple objects simultaneously.
E.g. for data on one attribute measured in N people, a cross-section dataset would be indicated as

$$\{X_1, X_2, \ldots, X_N\} = \{X_i\}_{i=1,\ldots,N}$$

and we would law them out scattered up and down on a real number line.

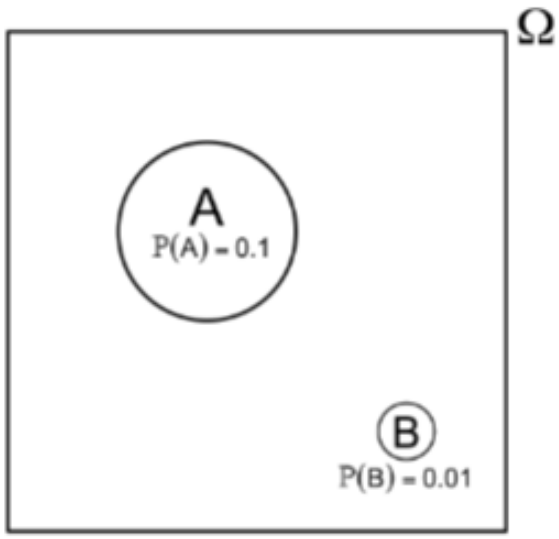


Figure 2.1: Single observation cross data

If we have more attributes, we can illustrate them in a 2D or 3D space. For example, a point $X = [5, 4]$.

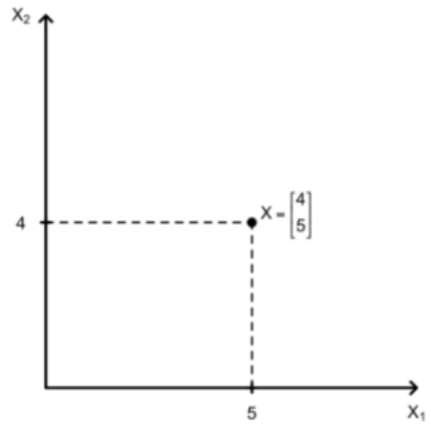


Figure 2.2: 2 observations cross data

We can easily imagine the above in more dimensions and with more datums

Definition 2.3 (Time series data): Time series data is a series of data points indexed in time order. Time series data of length T are written as

$$\{X_1, X_2, \ldots, X_T\} = \{X_t\}_{t=1,\ldots,T}$$

A time series is often thought of as having the time index as an implicit second attribute and being a sequence of bivariate observations {date, value}

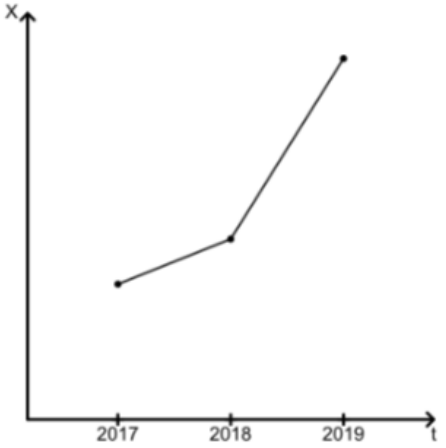


Figure 2.3: Time Series data

Actually, time series data is a single observation despite it seemingly drawn as continuous. Further, time series with 2 other observational attributes can be depicted in a 3D vector space.

Definition 2.4 (Panel Data): A Cross section of time series data. So essentially, observing multiple subjects over time.

2.2 Probability Mass Functions

Definition 2.5 (Probability of getting value x): The probability of X, which denotes some variable, yielding a particular value x, is given by

$$\mathbb{P}(X = x)$$

Definition 2.5 (Probability mass function): The probability mass function is

$$f(x) = \mathbb{P}(X = x)$$
$$f(x) \geq 0$$
$$\sum_x f(x) = 1$$

That is, for every value x in the support of X, the PMF of X gives $\mathbb{P}(X = x)$.

So the PMF for a discrete variable is a list of relative frequencies/probabilities associated with each of its possible values. A variable is discrete if it can only take a finite number of distinct values.

Consider if we have some data and relative frequency, our PMF would be

x	30	60	80	90	100
$\mathbb{P}(X = x)$	0.3	0.2	0.3	0.1	0.1

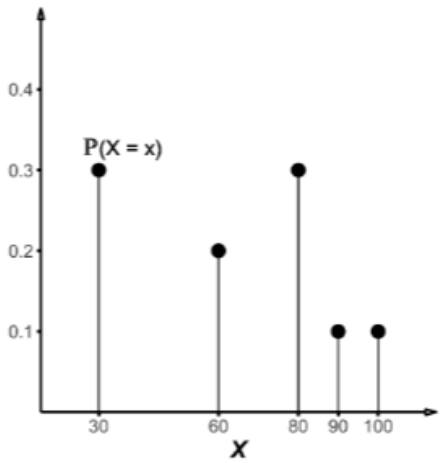


Figure 2.4 & 2.5: Table form and graphical form PMF

Bernoulli Distributions

Definition 2.6 (Bernoulli Distribution): The Binomial Distribution is when there are only 2 outcomes such that

$$f(0) = \mathbb{P}(X = 0) = p, \qquad f(1) = \mathbb{P}(X = 1) = 1 - p$$

Definition 2.7 (Generalised Bernoulli Distribution): The generalised Bernoulli distribution is when there are more than 2 possible outcomes such that, for a variable which takes the values of integers from $1, \ldots, k$, the support is $x \in \{1, \ldots, k\}$ and the PMF is

$$\mathbb{P}(X = x) = p_x$$

or

$$\mathbb{P}(X = x) = [x = 1]p_1 + [x = 2]p_2 + \cdots + [x = k]p_k$$

where $[x = y]$ denotes the 'Iverson bracket' which converts any logical proposition that is 1 if the proposition is satisfied and 0 otherwise.

We can take a variable and transform it into a new variable, like taking discrete variable X and transforming it to a new discrete variable $Y = g(X)$. For instance, suppose we independetly toss a coin a number of times and record the results

$$\{X_1, X_2, X_3, \ldots, X_N\} \qquad \{1, 0, 1, \ldots, 0\}$$

we then add up the values and divide by N to yield

$$Y = \frac{\sum_{i=1}^{N} X_i}{N}$$

This is a new discrete variable which has the support $\{0, \frac{1}{N}, \frac{2}{N}, \ldots \frac{N-1}{N}, 1\}$ which is the mean of a set of numbers. When we have a binary variable, it gives the proportion of 1s.

The multiplicative version of the sigma summation notation is

$$\Pi_{i=1}^{N} X_i$$

Definition 2.8 (Binomial Variable): A binomial variable is the number of times an outcome independently occurs (in a 2-outcome case), taking an integer value between 0 and N, such that

$$Y = \sum_{i=1}^{N} X_i$$

Since, the probability of each event, given independence, is

$$p^x (1-p)^{N-x}$$

The PMF of a binomial variable is the sum of N independent Bernoulli variables each with the same parameter p is

$$f(x) = K p^x (1-p)^{N-x}$$

where K is the scaling factor depending on N and p. The values N and p are the parameters of the distribution – they control its shape. The larger the number of trials N, the wider the range of possible outcomes.

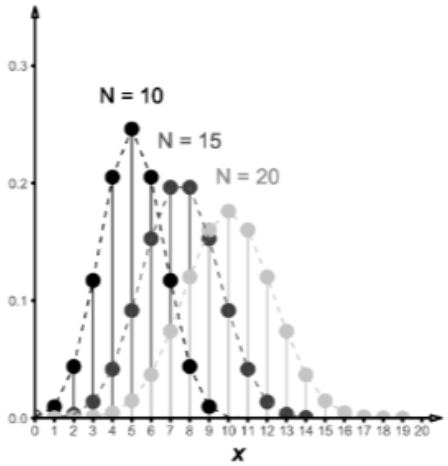


Figure 2.6: Distribution as N changes

If p is very large or small, the PMF will be skewed to the left or right respectively. The PMF is most spread out when $p = 0.5$. As $p \to 0$ or $p \to 1$, the values become increasingly concentrated

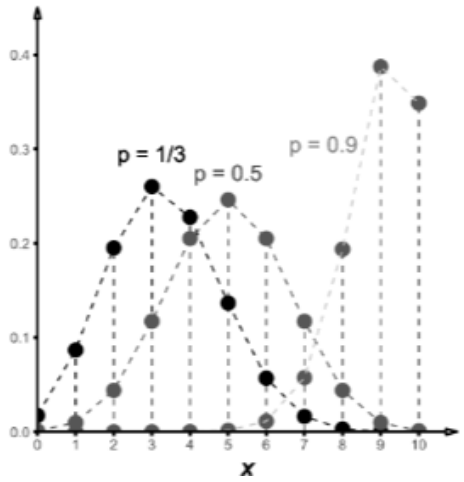


Figure 2.7: Distribution as p changes

2.3 Probability Density Functions

Definition 2.9 (Probability Density Function): Formally, the probability density function is defined as

$$f(x) = \lim_{h \to 0} \frac{\mathbb{P}(x \leq X \leq x + h)}{h}$$

$\mathbb{P}(x \leq X \leq x + h)$ means 'the probability that the variable X takes a value in the interval $[x, x + h]$. The PDF is a continuous function.
Properties of the PDF

1. Non-negativity: $f(x) \geq 0$
2. Area underneath is equal to 1:
$$\int_{-\infty}^{\infty} f(x)dx = 1$$
3. For continuous variables, the probability that X is observed to have exactly some value x is 0:
$$\mathbb{P}(X = x) = 0$$
4. $f(x)$ is not itself a probability

Properties (1) and (2) are analogous to discrete counterparts.

Types of Probability Density Functions

Characterisation 2.1 (Uniform Distribution): A random number generator acting over an interval of numbers $[a, b]$ has a continuous uniform distribution (retangular). The PDF is given by

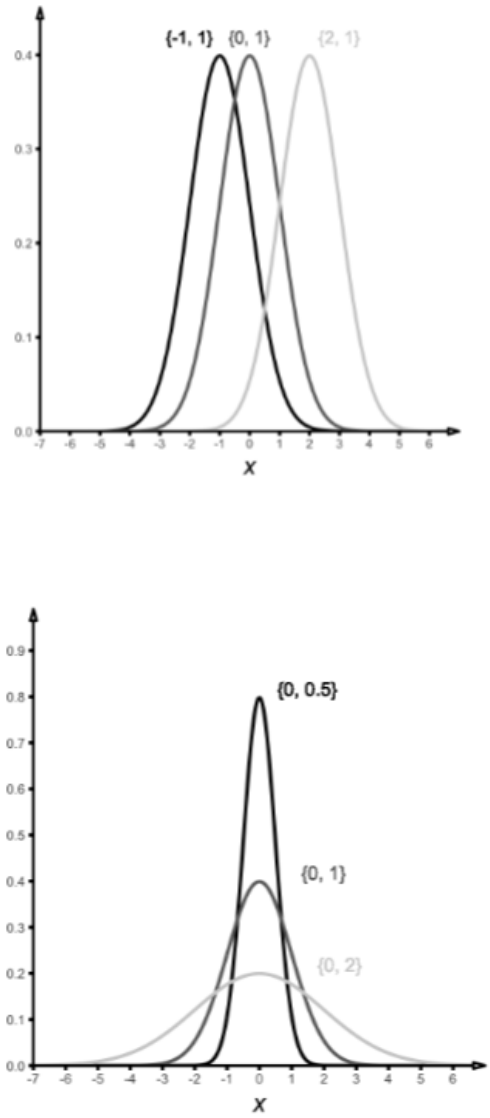
$$\begin{aligned} f(x) &= \frac{1}{b-a} \text{ if } a \leq x \leq b \\ &= 0 \text{ otherwise} \end{aligned}$$

a, b are the parameters of this distribution; they determine the upper and lower limits of the variable.

Characterisation 2.2 (Normal Distribution): The PDF is given by

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

which is just the bell curve with 2 parameters, μ and σ: μ represents the mean and thus controls the location of the distribution; σ controls the variance and thus the spread of the distribution.



Figures 2.8 & 2.9: Change of μ and σ for Normal Distributions

2.4 Cumulative Density functions

Definition 2.10 (Cumulative Distribution Functions): The CDF is defined as

$$F(x) = \mathbb{P}(X \leq x) \text{ for all } x \text{ in the support of } X$$

It is a continuous function giving the probability that the variable X is less than or equal to some value x. Unlike the PMF, the CDF is defined for every value of x; it is also defined in the gaps between outcomes in

the case of discrete variables.

Discrete variables have CDFs which are step functions.

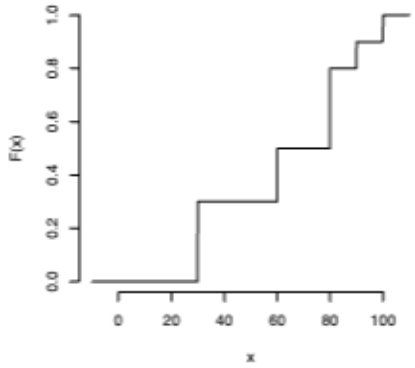


Figure 2.10: CDF for discrete variables

For continuous variables, the CDF is smooth. For a uniform distribution, it is a straight line between the supremum and infimum of the interval. For a general $U(a, b)$ variable, the CDF is

$$F(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a, b] \\ 1 & \text{for } x > b \end{cases}$$

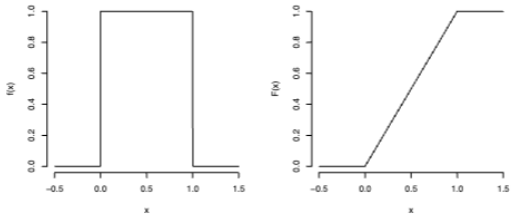


Figure 2.10: The PDF and CDF of a $U(0, 1)$

For a normal distribution, they are

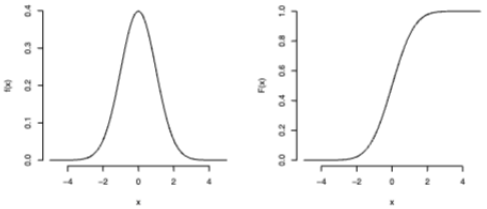


Figure 2.11: The PDF and CDF of a $N(0, 1)$

For any cumulative distribution function

1. $0 \leq F(x) \leq 1$
2. $F(x)$ is a non-decreasing function in x

3. $\mathbb{P}(X > x) = 1 - \mathbb{P}(X \leq x)$
4. $\mathbb{P}(a < x \leq b) = F(b) - F(a)$

The relationship between the CDF and PDF
The CDF is (slightly) informally the area of the PDF. Hence, when the CDF is differentiable, the PDF and CDF are connected by

$$f(x) = \frac{dF(x)}{dx}$$

or by the first fundamental theorem of calculus,

$$F(x) = \int_{-\infty}^{x} f(z)dz$$

For Uniform distribution, just trivially use area of the rectangle (of PDF) to solve for CDF.
The value of the CDF at x is the probability of getting a value less than or equal to x and it is equivalent to the area to the left of x under the PDF.
The CDF also relates directly to quantiles.

1. The median is the value of x such that $\mathbb{P}(X \leq x) = 0.5$
2. The first quartile is the value of x such that $\mathbb{P}(X \leq x) = 0.25$

Reading off Statistical Tables

Basically, the rows/columns give the values of x: the rows give the first decimal (and ones) place while the column gives the second decimal place.
When we have the value of x we scan the table for that value of x and the corresponding probability value.
When we have the value of $\mathbb{P}(X \leq x)$ we scan the table for the probability and the corresponding value of x.

Interpolation
Procedure is basically

1. Find the weights for the numbers you have all three aspects, either x values or the probabilities
2. Substitute back the weights for the numbers where you only have two aspects but not the interpolated value

Example.
Suppose we want to find the 3rd percentile of the distribution, such that $\mathbb{P}(X \leq x) = 0.75$. We realise that it is in the second table between $x = 0.67$ and $x = 0.68$.
We should linearly interpolate the gaps: suppose the table reports values for 0.7486 and 0.7517, then the value 0.75 is given by

$$0.75 = 0.7486a + 0.7517(1 - a)$$

where $a, (1 - a)$ represents the weights of each value of x. So,

$$a = \frac{0.7515 - 0.75}{0.7517 - 0.7486} = 0.5483871$$

Applying this weight to $x = 0.67$ and $x = 0.68$ gives

$$x = 0.5483871 \times 0.67 + (1 - 0.5483871) \times 0.68 = 0.6745161$$

which is close enough to the true value

2.5 Standardisation

Effects of manipulation of variable of a Normal Distribution.

1. Changing location (addition/subtraction): If $X \sim N(\mu, \sigma^2)$ and $Z = a + X$, then $Z \sim N(a + \mu, \sigma^2)$
2. Changing spread (multiplication): If $X \sim N(\mu, \sigma^2)$ and $Z = aX$, then $Z \sim N(a\mu, (a\sigma)^2)$

Turning $N(\mu, \sigma^2)$ into a Standard normal $N(0, 1)$: If $X \sim N(\mu, \sigma^2)$ and $Z = \frac{X-\mu}{\sigma}$, then $Z \sim N(0, 1)$

Suppose $X \sim N(\mu, \sigma^2)$ and further suppose that we want to know $\mathbb{P}(X \leq x)$. Since $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$, we find $z = \frac{x-\mu}{\sigma}$

2.6 Multivariate Distributions

For when we take two or more measurements.

Definition 2.11 (Distribution functions for bivariate distributions): The joint CDF for bivariate events is

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

or, through double integrals,

$$F(x, y) = \int \int f(x, y) dx dy$$

Of course it follows that

$$F(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$$

And the joint PDF is

$$f(x, , y) = \mathbb{P}(X = x, Y = y) \qquad \text{discrete}$$
$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y} \qquad \text{continuous}$$

If we know the joint distribution of two variables we can work out the distributions of each variables individually.

Definition 2.12 (Marginal PDF): The marginal PDF of x is

$$f(x) = f(x, \text{ whatever the value of } y)$$

or more formally,

$$f(x) = \int_{-\infty}^{\infty} f(x, y) dy, \qquad \text{for all } x$$

The PDF would be in terms of x, so you essentially get the value of x for all y values. Same for marginal PDF of y.

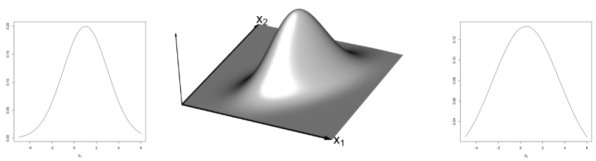


Figure 2.12: Marginal PDF for a continuous bivariate distribution

If X and Y are independent, then the joint PDF is the product of the marginals

$$f(x, y) = f(x)g(y)$$

Same for CDFs

$$F(x, y) = F(x)G(y)$$

Characterisation 2.2 (Conditional PDF): Conditional PDFs are slices through joint PDFs in one direction, holding the values of other values constant. For instance,

$$f(x_1|X_2 = 0)$$

Conditional distributions tell you things about relationships between variables especially when you take multiple values of x_2 as constant.

Useful Distinction: The marginal probability of an event is the probability distribution that describes that single event only. The conditional probability, on the other hand, is a distribution that represents the likelihood of an event to occur given a particular outcome of another event

2.7 Expected Values

Definition 2.13 (Expected Value): Expected value of X is defined as

$$\mathbb{E}[X] = \begin{cases} \sum_i x_i f(x_i) & \text{discrete} \\ \int x f(x) dx & \text{continuous} \end{cases}$$

and for the expected value of a function of a variable

$$\mathbb{E}[g(X)] = \begin{cases} \sum_i g(x_i) f(x_i) & \text{discrete} \\ \int g(x) f(x) dx & \text{continuous} \end{cases}$$

These are the probability-weighted sums of the values which variables can take. The expectation of a discrete value need not be in the support of that variable and in general expectations may not be the most probable value.

Example (Uniform Distribution). The uniform $U(a, b)$ has PDF

$$f(x) = \frac{1}{b-a} \text{ for } x \in [a, b] \text{ and 0 elsewhere}$$

The expected value is

$$\mathbb{E}[X] = \frac{a+b}{2}$$

Example (Normal Distribution). The Normal $N(\mu, \sigma^2)$ has the expected value

$$\mathbb{E}[X] = \mu$$

Example (Binary Indicator). The binary indicator has the expected value

$$\mathbb{E}[X] = 0 \times f(0) + 1 \times f(1) = f(1)$$

Example (Binomial Distribution). The Binomial Distribution has the expected value

$$\mathbb{E}[X] = np$$

where p is the probability of success while n is the number of trials.
Rules for Expected Values

1. Affine functions: $\mathbb{E}[a + bX] = a + b\mathbb{E}[X]$
2. Addition: if $h(Y)$ is another function of another variables (or even of the same variable i.e. $Y = X$), then
$$\mathbb{E}[g(X) + h(Y)] = \mathbb{E}[g(X)] + \mathbb{E}[h(Y)]$$
for example,
$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$
3. Multiplication:
$$\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)] \qquad \text{if they are independent}$$
for example,
$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$$

Theorem 2.1 (Jensen's Inequality): the expected value of a nonlinear function of a variable is not equal to the nonlinear function of the expected value

$$\mathbb{E}[f(X)] \leq f(\mathbb{E}[X]) \text{ if } f \text{ is a concave function}$$
$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X]) \text{ if } f \text{ is a convex function}$$

Examples

1. logarithmic functions are concave so $\log(\mathbb{E}[X] \geq \mathbb{E}[\log(X)]$
2. the exponential function is convex so $e^{\mathbb{E}[X]} \leq \mathbb{E}[e^X]$
3. Squaring a variable is a convex function so $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$
4. Square-roots are concave functions so $\mathbb{E}[\sqrt{X}] \leq \sqrt{\mathbb{E}[X]}$
5. Ratios of expected values in general are not equal to expected values of ratios
$$\mathbb{E}\left[\frac{X}{Y}\right] \neq \frac{\mathbb{E}[X]}{\mathbb{E}[Y]}$$

2.8 Variance

Definition 2.14 (Variance): Variance is the mean of the squared difference between the variable and its expectation.
$$Var(X) = \mathbb{E}\left([X - \mathbb{E}[X]]^2\right) = \mathbb{E}[X^2] - [\mathbb{E}(X)]^2$$
By definition, $Var(X) \geq 0$. And $Var(aX) = a^2 Var(X)$
For discrete and continuous cases, we can compute the Variance as

$$Var(X) = \begin{cases} \sum_i (x_i - \mu)^2 f(x_i) & \text{discrete} \\ \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx & \text{continuous} \end{cases}$$

Definition 2.15 (Standard Deviation): The Standard Deviation of a variable is the square root of the variance
$$\sigma = \sqrt{Var(X)}$$
Since Variances are in squared units, the standard deviation is used to give a measure of dispersion which is comparable to the mean and the variable of interest

Example. (Uniform Distribution.) The variance of the uniform $U \sim (a, b)$ is

$$Var(X) = \frac{(b-a)^2}{12}$$

Example. (Normal Distribution.) The variance of the Normal Distribution $N \sim (\mu, \sigma^2)$ is

$$Var(X) = \sigma^2$$

Example. (Binary Variables) The variance of a binary variable where p is the proportion of 1's in the population is

$$Var(X) = p(1-p)$$

Example. (Binomial Distribution). The variance of a binomial distribution with p as the proportion of 1's in the population and n as the number of trials is

$$Var(X) = np(1-p)$$

Number of standard deviations, σ, positively correlate with rareness of the event

2.9 Conditional Expectations

The notation for conditional expectations (given some X) is

$$\mathbb{E}(Y|X)$$

Uses: consider that we want to find the difference in expected salary of genders and $X = 1, 0$ is an indicator variable where 1 denotes female and 0 denotes male. So we take,

$$\mathbb{E}[Y|X=1] - \mathbb{E}[Y|X=0]$$

Definition 2.16 (Conditional Expectation): Give a conditional PDF $f(y|x)$, the conditional expectation is defined as

$$\mathbb{E}[Y|X=x] = \begin{cases} \sum_i y_i f(y_i|x) & \text{discrete} \\ \int y f(y|x) dy & \text{continuous} \end{cases}$$

Properties of Conditional Expectations

1. $\mathbb{E}[h(X)Y|X] = h(X)\mathbb{E}[Y|X]$. Conditioning on X implies treating it as known as this is akin to $\mathbb{E}[aX] = a\mathbb{E}[X]$
2. If X, Y are independent, then $\mathbb{E}[Y|X] = \mathbb{E}[Y]$. If X, Y are independent conditioning on X tells you nothing about the mean of Y
3. Law of Iterated Expectations:
$$\mathbb{E}[Y] = \mathbb{E}(\mathbb{E}[Y|X])$$

2.10 Correlation, Covariance & Independence

About the association between variables.

Definition 2.17 (Covariance): The Covariance of variables X, Y, denoted σ_{xy}, is

$$Cov(X,Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$$

or equivalently,

$$Cov(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

Proof of Equivalence: From first to second,

$$\begin{aligned} \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] &= \mathbb{E}[XY - X\mathbb{E}(Y) - Y\mathbb{E}(X) + \mathbb{E}(X)\mathbb{E}(Y)] \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \ \blacksquare \end{aligned}$$

Note: if $X = Y$, then $Cov(X,Y) = Var(X) = Var(Y)$. Covariance is an expectation of the product of derivations of X and Y from their means. When X and Y are both above or below their means, then Covariance will be positive, otherwise it will be negative. Moreover, when $Var(X) = 0$ or $Var(Y) = 0$ $\implies Cov(X,Y) = 0$.

Properties of Covariance

1. $Cov(X, X) = Var(X)$
2. $Cov(X, Y) = Cov(Y, X)$
3. $Cov(X, a) = 0$ where a is a constant
4. $Cov(aX, Y) = aCov(X, Y)$ where a is a constant
5. $Cov(X, Y + Z) = Cov(X, Y) + Cov(X, Z)$ or $Cov(X + Y, Z) = Cov(X, Z) + Cov(Y, Z)$

The last two properties are known as bilinearity.
These properties allow us to work out the variance of the sum of two variables.

$$Var(aX \pm bY) = a^2 Var(X) + b^2 Var(Y) \pm 2abCov(X, Y)$$

If X and Y are independent, then $Cov(X, Y) = 0$ since $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ but the converse, that if $Cov(X, Y) = 0$, then X and Y are independent, does not hold.

Definition 2.18 (Correlation): The Correlation between X and Y, denoted by ρ_{xy}, is

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

or equivalently,

$$Corr(X, Y) = Cov\left(\frac{X - \mathbb{E}[X]}{\sqrt{Var(X)}} \frac{Y - \mathbb{E}[Y]}{\sqrt{Var(Y)}}\right)$$

Proof left as an exercise.
i.e. correlation is just the covariance between the standardised variables.

Results of Independence of X and Y

1. $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$
2. $Corr(X, Y) = 0$
3. $Cov(X, Y) = 0$

Definition 2.19 (Mean-independence). The conditions for Mean-independence are (notion of independence which lies 'between' independence and uncorrelatedness)

$$\mathbb{E}[Y|X] = \mathbb{E}[Y] \iff Y \text{ is mean-independent of } X$$
$$\mathbb{E}[X|Y] = \mathbb{E}[X] \iff X \text{ is mean-independent of } Y$$

Moreover,

1. Independence $\implies$ Mean-independence
2. Mean-independence $\not\implies$ Independence

3 Statistics

3.1 Populations, Samples & Random Variables

Definition 3.1 (Population, Samples and Random Variables).

1. Population: a complete enumeration of some set of interest or a mathematical model which generates a set of interest

2. Sample: a subset of a population.
   If we consider a variable X and sample N values of this variable from the population, the sample is denoted $\{X_1, X_2, \ldots, X_N\}$

3. Sampling: process of selection of a subset of individuals from within a population.
   Most dominant approach is "probability sampling": simple random sampling selects the pre-determined number of respondents to be interviewed from a target population with each potential respondent having an equal chance of being selected. In theory, it produces a sample of respondents representative of the population; representative means that if we repeat the procedure many times, the features of the sample would on average, across all samples, match those of the population

4. Sampling frame: source material/device from which a sample is drawn; most straightfoward is a list of the entire population of interest with appropriate identifying information

5. Random variables: generated by random sampling.
   For instance, given a variable X which has some distribution in the population $f(X)$. When we draw our first sample X_1, since it could take any value in the support of X with probability given by $f(X)$ it is random. Thus, the first sampled value is random and has the same probability distribution as X

6. Distribution of sample values: Assuming that the population is very large or sampling is done with replacement, then X_2 is also a random variable.
   Sampling without replacement over a small population would lead to non-iid sampling

7. Independent and Identically distributed: When the sample values $\{X_1, X_2, \ldots, X_N\}$ are all drawn from the same populaton and have the same distribution then they are said to be independently and identically distributed or 'iid' for short.
   Usually, our survey population is extremely large so we do not need to sample with replacement

Therefore, for our iid random sample $\{X_1, X_2, \ldots, X_N\}$ of some variable from a population with mean μ and variance σ^2, then

$$\mathbb{E}[X_1] = \mathbb{E}[X_2] = \ldots \mathbb{E}[X_N] = \mu$$
$$Var(X_1) = Var(X_2) = \ldots Var(X_N) = \sigma^2$$

3.2 Parameters, Statistics & Estimation

Definition 3.2 (Parameters and Statistics).

1. Parameter: Numerical measure that describes a specific characteristic of a population

2. Statistic: Numerical measure that describes a specific characteristic of a sample. Formally, a statistic is a "function of a random variable".
   Statistics being computed from random variables are subject to sampling variation. Population parameters are not.

Definition 3.3 (Estimation).

1. Estimand: Parameter in the population which is to be estimated in a statistical analysis

2. Estimator: A function for calculating an estimate of a given population parameter based on randomly sampled data. An estimator is a function of a sample data which is drawn randomly. They themselves are random variables and therefore have distributions, expected values etc.

   (a) The estimator of the mean of a variable X is denoted $\bar{X}$

   (b) The estimator of the variance of a variable is often denoted by s^2 (s denotes standard deviation)

   (c) A proportion is just the mean for a binary variable but in this case the notation for the sample mean/proportion of 1's is $\hat{p}$

3. Estimate: An estimate is the numerical value of the estimator given a specific sample drawn; it is a nonrandom number (e.g. the sample mean)

Main estimators used for key parameters for a population of size M and a sample of size N

	Population Parameter	Sample Estimator
Mean	$\mu = \frac{\Sigma_{i=1}^{M} X_i}{M}$	$\bar{X} = \frac{\Sigma_{i=1}^{N} X_i}{N}$
Variance	$\sigma^2 = \frac{\Sigma_{i=1}^{M} (X_i - \mu)^2}{M}$	$s^2 = \frac{\Sigma_{i=1}^{N} (X_i - \bar{X})^2}{N-1}$
Variance (binary variables)	$\sigma^2 = p(1-p)$	$s^2 = \hat{p}(1-\hat{p})\frac{N}{N-1}$
Std. Deviation	$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{s^2}$
Covariance	$\sigma_{xy} = \frac{\Sigma_{i=1}^{M} (X_i - \mu_x)(Y_i - \mu_y)}{M}$	$s_{xy} = \frac{\Sigma_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})}{N-1}$
Covariance (binary variables)	$\sigma_{xy} = p_x p_y$	$s_{xy}^2 = \hat{p}_x \hat{p}_y$
Correlation	$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$	$r_{xy} = \frac{s_{xy}}{s_x s_y}$

Figure 3.1: Main estimators for key parameters

Characteristic of an estimator

- It is a statistic and hence subject to sampling variation
- Thus, it has a distribution (with a PMF/PDF, CDF etc) called a sampling distribution
- This sampling distribution has an expected balue and a variance

Properties of a good estimator

1. Unbiasedness: concerns the relationship between the expected value of a statistic and the parameter being estimated. If the expected value of the estimator is equal to the parameter then the statistic is an unbiased estimator.
   Definition 3.4a (Bias of Sample mean). Let $\hat{\theta}$ be an estimator of the population parameter θ. The bias of $\hat{\theta}$ is
   $$\mathbb{E}[\hat{\theta}] - \theta$$
   Therefore, $\hat{\theta}$ is an unbiased estimator of θ if $\mathbb{E}[\hat{\theta}] = \theta$.

   Example. Suppose we have a population with a mean $\mathbb{E}[X] = \mu$ and an iid random sample of size N from this population $\{X_1, X_2, \ldots, X_N\}$. The population object of interest is μ. Our estimator of this is the arithmetic mean
   $$\bar{X} = \frac{X_1 + X_2 + \cdots + X_N}{N} = \frac{\sum_{i=1}^{N} X_i}{N}$$
   Finding $\mathbb{E}[\bar{X}]$,
   $$\mathbb{E}[\bar{X}] = \mathbb{E}\left[\frac{X_1 + X_2 + \cdots + X_N}{N}\right] = \frac{\mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_N]}{N}$$
   Given that the expected value of an observation X_i drawn iid-randomly from a population with $\mathbb{E}[X] = \mu$ is also μ, this yields
   $$\mathbb{E}[\bar{X}] = \frac{\mu + \mu + \cdots + \mu}{N} = \frac{\sum_{i=1}^{N} \mu}{N} = \mu$$
   Definition 3.4b (Bias of Sample variance). Suppose we have a population which has a mean of μ and variance given by σ^2. Suppose we have an iid random sample of size N from this population $\{X_1, X_2, \ldots, X_N\}$. The population object of interest is σ^2 and the definition of population variance is
   $$\mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[(X - \mu)^2]$$
   Given that we do not know the population mean, we use the sample mean which is an unbiased estimator of μ in its place to calculate the sample mean of $(X - \bar{X})^2$, which is
   $$\hat{\sigma}^2 = \frac{\sum_i (X_1 - \hat{X})^2}{N}$$

This is the arithmetic mean of the squared deviations from the arithmetic mean. But it turns out that $\mathbb{E}[\hat{\sigma}^2] \neq \sigma^2$ and it is biased by a factor depending on the sample size N.
It is the use of the estimator $\hat{X}$ in place of the population parameter μ that is the source of the bias. The intuition is that you need the mean in order to work out the variance and once you have the mean you only have $N-1$ independent observations left (the N'th can simply be worked out from the mean and the other $N-1$ data points). So when you are averaging $(X_i - \bar{X})^2$ you are only averaging over $N-1$ true observations.

Bessel's Correction. In order to correct this biased variance estimator, you need to multiply by a factor $\frac{N}{N-1}$.

$$s^2 = \frac{N}{N-1}\sigma^2$$

which yields

$$s^2 = \frac{\sum_i (X_i - \bar{X})^2}{N-1}$$

which is the unbiased estimator of the population variance. The effect of Bessel's correction is smaller the larger the sample; intuition is that as the sample grows larger, you would expect $\bar{X}$ to be nearer μ

2. Efficiency: refers to the effect of sampling variation. The measure of variability (precision) in a statistic (estimator) is the standard error of its sampling distribution which is the square root of its variance. A small standard error indicates that the sampling distribution does not exhibit much variation and so the estimate is more precise.

Definition 3.5 (Standard Error). A measure of the variation in the sampling distribution of a statistic; it is equal to the square root of the variance of the statistic.
Derivation of Standard Error. Suppose we have a sample of iid random sample of size N from a population with mean μ and variance σ^2. The sample mean is

$$\bar{X} = \frac{\sum_i X_i}{N}$$

We will only have one number because we have only one sample. But if we repeatedly sample from our data and calculate the mean each time we know that it would be subject to sampling variation. In other words, the sample mean becomes a random variable so it will have a sampling distribution with a mean and variance. And since the expected value of the sample mean is equal to the expected value of the population variable:

$$\mathbb{E}[\bar{X}] = \mu$$

For the Variance of the Sample mean

$$\begin{aligned} Var[\bar{X}] = Var\left[\frac{\sum_i X_i}{N}\right] &= \frac{1}{N^2} Var[\sum_i X_i] \\ &= \frac{1}{N^2} Var[X_1 + X_2 + \cdots + X_N] \\ &= \frac{1}{N^2}[Var(X_1) + Var(X_2) + \cdots + Var(X_N)] \end{aligned}$$

where the last step follows from independence. We know that with iid sampling, $\mathbb{E}[X_i] = \mu$ and also that $Var[X_i] = \sigma^2$, hence

$$Var(X_1) + Var(X_2) + \cdots + Var(X_N) = \sigma^2 + \sigma^2 + \cdots + \sigma^2 = N\sigma^2$$

Hence, the variance of the sampling distribution of the mean is the variance of the variable in the population divided by the sample size,

$$Var[\bar{X}] = \frac{1}{N^2} N\sigma^2 = \frac{\sigma^2}{N}$$

The standard eror of the sampling distribution of the sample mean is just the square root of the variance

$$SE[\bar{X}] = \sqrt{\frac{\sigma^2}{N}} = \frac{\sigma}{\sqrt{N}} \quad \blacksquare$$

3. Consistency: consistency is the property that as the number of data points used increases indefinitely, the resulting sequence of estimates converges to the population parameter of interest. Estimates are still subject to sampling variation and thus move around but do so less and less as the sample size increases. Sampling distribution of the estimate becomes more concentrated near the true value of the parameter being estimated, so that the probability of the estimator being arbitrarily close to the population parameter converges to one.
One could obtain a sequence of estimates indexed by N, and consistency is a property of the behaviour of this statistic as the sample size grows; if the sequence of estimates can be mathematically shown to converge in probability to the population parameter, it is a "consistent" estimator; otherwise, it is "inconsistent". Consistent estimators are not necessarily unbiased.

   Sufficient conditions for Consistency:

   - Variance of the estimator goes to zero as the sample size grows. Thus, its variance is

     $$Var(\bar{X}) = \frac{\sigma^2}{N}$$

     and hence

     $$\text{As } N \to \infty, Var(\bar{X}) = \frac{\sigma^2}{N} \to 0$$

   - It is an unbiased estimator for μ

3.3 The Law of Large Numbers & the Central Limit Theorem

Normally, we only have a single random sample from the population of interest. Our population parameter is unknown but fixed, while our sampling statistic is known but subject to sampling variation. But it is possible to make approximate statements about the sampling distribution of the mean which are valid irrespective of the sampling variation.

The Law of Large Numbers and Central Limit Theorem are about the sequence

$$\bar{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$$

as N grows (so the behaviour of $\bar{X}_N$ as N increases).

Theorem 3.1 (Law of Large Numbers). Let $\{X_1, X_2, X_3, \ldots, X_N\}$ be an iid random sample from a population with mean and variance given by (μ, σ^2). Then

$$\mathbb{P}(|\bar{X}_N - \mu| > \epsilon) \to 0 \text{ as } N \to \infty \qquad \text{for any } \epsilon > 0$$

The probability of the a (large) absolute difference between the sample mean and the population mean gets arbitrarily small as the sample size grows; the sample mean converges to the population parameter as the sample gets bigger. **Proof given in appendix**
Some facts about LLN

- Convergence is not smooth or monotone
- The LLN states that $|\bar{X}_N - \mu|$ is ultimately small but not that every value is small; it could be that for some value of N it is big

- But the probability of such an event is extremely small in large samples

Motivation for Central Limit Theorem. We want to standardise our sample mean by subtracting the population mean and dividing by the standard error of the sampling distribution of the mean to yield

$$Z_N = \frac{\bar{X}_N - \mu}{\frac{\sigma}{\sqrt{N}}}$$

We then obtain a new sequence of random variables: $\{Z_N\}_{N=1,2,3,\ldots}$. We want to know the behaviour of this sequence.

Theorem 3.2 (Central Limit Theorem). Let $\{X_1, X_2, \ldots, X_N\}$ be an iid random sample from a population with mean and finite variance given by (μ, σ^2). Let

$$Z_N = \frac{\bar{X}_N - \mu}{\frac{\sigma}{\sqrt{N}}}$$

Then

$$\lim_{N\to\infty} Z_N \text{ is } N(0,1)$$

As the sample size grows, the sampling distribution of the standardised mean gets close to a Standard Normal.
Notice that if

$$Z_N = \frac{\bar{X}_N - \mu}{\frac{\sigma}{\sqrt{N}}} \sim N(0,1)$$

Then by rearrangement

$$\bar{X}_N = \mu + Z_N \frac{\sigma}{\sqrt{N}}$$

and then we deduce that

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

So the sampling distribution of the sample mean itself follows a normal distribution once $N \to \infty$

3.4 Intro to Statistical Inference

If we have a sample all we can calculate is an estimate $\hat{\theta}$. Since $\hat{\theta}$ is a statistic, it is subject to sampling variation and has a sampling distribution.

Standardised sample means for regular and binary variables respectively (achieved by subtracting population mean and dividing by SE)

$$Z = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{N}}} \qquad Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{N}}}$$

By the CLT, the sampling distribution of the standardised mean is approximately, $N(0,1)$ in large samples. In the case of binary/indicator variables, this approximation is good as long as p is not too close to 0 or 1.

Even when we plug in estimates of standard deviation, the standardised means are still approximately $N(0,1)$ in large samples

$$Z = \frac{\bar{X} - \mu}{\sqrt{\frac{s^2}{N}}} \qquad Z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}}$$

In this context, the population variance is a "nuisance parameter" which is any parameter which is not of immediate interest but which must be accounted for in the analysis of those parameters which are of interest.

3.5 Confidence Intervals

Our first goal in using a sample is to find a plausible range of values for the population parameter. Consider the standardised mean

$$Z = \frac{\bar{X} - \mu}{\sqrt{\frac{s^2}{N}}}$$

By the CLT we know that the sampling distribution of the standardised mean is: $Z \sim N(0, 1)$. We know that

$$\mathbb{P}[-1.96 \leq Z \leq +1.96] = 0.95$$

i.e. the probability of a $N(0, 1)$ random variable lying within 1.96 standard deviations of 0 is 0.95. Substituting Z yields

$$\mathbb{P}\left[-1.96 \leq \frac{\bar{X} - \mu}{\sqrt{\frac{s^2}{\mu}}} \leq +1.96\right] = 0.95$$

Rearranging yields

$$\mathbb{P}\left[\left(\bar{X} - 1.96\sqrt{\frac{s^2}{N}}\right) \leq \mu \leq \left(\bar{X} + 1.96\sqrt{\frac{s^2}{N}}\right)\right] = 0.95$$

The general form of a confidence interval is: sample statistics $\pm$ a number of standard deviations $\times$ standard error of the statistic.

The width of the confidence interval depends on the number of standard errors either side of zero you need to go in order to cover a fixed probability. The typical values are

- ± 1.645 SEs: Probability = 0.9
- ± 1.95 SEs: Probability = 0.95
- ± 2.58 SEs: Probability = 0.99

In general, confidence intervals (the probability of finding the population parameter in some set of values)

- Increase with the confidence level required
- Increase with the standard error of the statistic
- Increases with the variance of data
- Decreases as the sample size increases

A confidence interval is based on sample statistics so is itself a statistic: it is also subject to sampling variation

Interpreting Confidence Intervals

Suppose we have $Z \sim N(0, 1)$ and we take 20 samples each with $N = 5$ and compute the mean and the 95% confidence interval: $\bar{X} \pm 1.96 \times \sqrt{\frac{1}{N}}$. We get 20 different sample confidence intervals which look like

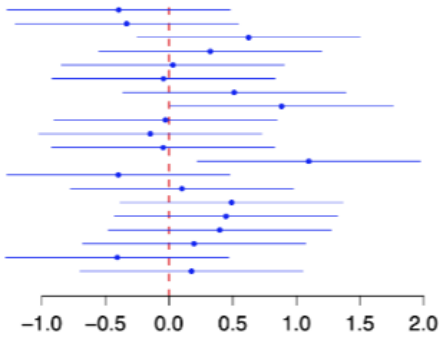


Figure 3.1: Confidence Intervals

The population parameter lies in approximately 95% of them. Your sample confidence interval corresponds to one of these. It thus has a 0.95 probability of covering the parameter of interest.
In interpreting the confidence interval, we should interpret it in terms of probabilistic behaviour of the interval, NOT the population parameter (which is fixed):
E.g. There is a 95% probability that the interval $[a, b]$ will contain the population parameter.

3.5.1 Confidence Intervals for Means

Confidence Intervals for Means for each sample type:

1. Single sample of size N from population with mean μ and variance σ^2:

$$\bar{X} \pm 1.96\sqrt{\frac{s^2}{N}}$$

2. Two independent samples of size N_x and N_y from populations with means μ_x and μ_y and variances σ_x^2 and σ_y^2:

$$(\bar{X} - \bar{Y}) \pm 1.96\sqrt{\frac{s_x^2}{N_x} + \frac{s_y^2}{N_y}}$$

3. A paired sample of size N for two variables with populations means μ_x and μ_y and variances σ_x^2 and σ_y^2:

$$(\bar{X} - \bar{Y}) \pm 1.96\sqrt{\frac{s_D^2}{N}}$$

Confidence Interval for the Difference in Two Independent Means

Derivation of Standard Error. Suppose we have two independent samples from our two populations: $\{X_1, X_2, \ldots, X_N\}$ and $\{Y_1, Y_2, \ldots, Y_N\}$.
We suppose that the X's are drawn iid from some distribution with mean μ_x and variance σ_x^2 and that the Y's are iid from another distribution with mean and variance $\{\mu_y, \sigma_y^2\}$. We would like to find a range for the difference between population means, that is

$$(\mu_x - \mu_y)$$

which is consistent with the data.
Recall for two independent variables

$$Var(X \pm Y) = Var(X) + Var(Y)$$

By using this definition, we can write the estimator for the variance of the sampling distribution of difference of two means as:

$$Var(\bar{X} - \bar{Y}) = \frac{s_x^2}{N_x} + \frac{s_y^2}{N_y}$$

And the standard error is just the square root of variance

$$SE(\bar{X} - \bar{Y}) = \sqrt{\frac{s_x^2}{N_x} + \frac{s_y^2}{N_y}}$$

Confidence Interval for the Difference in Two Dependent Means (Paired Data)

Not all random variables are independent. Sometimes the dependence occurs simply because we are measuring two features of the same, randomly sampled object. For instance if we take a random sample of couples and measure incomes of each member, there will be dependence between them because the randomisation only influence which couples appear in the sample; once we have the sample of couples the two individuals within each are tied together because if you sample one member of the couple you necessarily sample the other. **So, you have one population and one sample of pairs of random variables (i.e. bivariate data).** You also implicitly have a sample of pair-wise differences between variables.
Therefore, you need to take correlation account as this variance of the sampling distribution of difference in the means of the two variables is not just the sum of the variances of each variable.
Derivation. We can calculate the paired differences and work out the variances of differences. It turns out that

$$Var(\bar{X} - \bar{Y}) = Var(\bar{D})$$

Therefore, given $Var(\bar{D})$, you can place confidence intervals on the difference in means of two variables on the basis of a paired sample using the same general method as for a single sample. Therefore, you just need to find directly the sample variance of the differences in your sample.

Example. If you have a sample of 71 students and the data is

Student	Exam 1	Exam 2	Difference
1	57.1	60.7	-3.6
⋮	⋮	⋮	⋮
71	78.6	82.9	-4.3
Sample Means	79.6	81.4	-1.8
Sample Variances	117	151	124.45
Sample Correlation	0.54		

Figure 3.2: Paired Data and Difference

In wanting to calculate the 95% Confidence Interval on $(\mu_x - \mu_y)$, it is just like for a single mean

$$(\bar{X} - \bar{Y}) \pm 1.96\sqrt{\frac{s_D^2}{N}}$$

3.5.2 Confidence Intervals for Proportions

Confidence Intervals for Proportions are for Binary Indicators.
Confidence Intervals for Proportions for each sample type:

1. Single sample of size N from population with mean p:

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$$

2. Two independent samples of size N_x and N_y from populations with means p_x and p_y:

$$(\hat{p}_x - \hat{p}_y) \pm 1.96\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{N_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{N_y}}$$

3. A paired sample of size N, two random variables with population means p_x and p_y:

$$(\hat{p}_x - \hat{p}_y) \pm 1.96\sqrt{\frac{s_D^2}{N}}$$

Confidence Interval for the Difference in Two Independent Proportions

Derivation of Standard Error. Suppose we have two independent samples $\{X_1, X_2, \ldots, X_N\}$ and $\{Y_1, Y_2, \ldots, Y_N\}$ of two binary variables and are interested in testing the difference in proportions $p_x - p_y$. Since samples are independent, we yield the variance of the sampling distribution of the difference as

$$Var(\hat{p}_x - \hat{p}_y) = \frac{\hat{p}_x(1-\hat{p}_x)}{N_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{N_y}$$

The Standard Error is just the square root of the variance.

Confidence Interval for the Difference in Two Dependent Proportions (Paired Data)

Treatment is exactly the same as the difference in means for non-binary data because

- For paired samples, we need to recognise that we just have a single sample of differences between two random variables and place the confidence interval directly on the mean difference
- the difference measure is not binary (the difference between two binary variables is 'trinary' and take values from {1,0,-1}) and its mean is not a proportion.

So as usual, we just take calculate the sample mean and variances normally like so

Subject	Before	After	Difference (After-Before)
1	1	1	0 (No change)
2	1	0	-1 (Quit)
3	0	0	0 (No change)
	⋮	⋮	⋮
98	1	0	-1 (Quit)
99	0	1	+1 (Started)
100	0	0	0 (No change)
Sample Means	0.3	0.25	-0.05
Sample Variances	0.21	0.1875	0.4
Sample Correlation	0.66		

Figure 3.3: Paired Data and Differences for Binary Variables

3.6 Tests of Statistical Hypotheses

Hypotheses are always about population parameters and never about sample statistics. A hypothesis is a statement that some population parameter is equal to a particular value or lies in some set of values.

Some preliminary definitions

- The Hypothesis of interest is called the Null Hypothesis (denoted H_0)
- The negation of the hypothesis is called the alternative hypothesis (denoted H_1)

There are two types of errors that we can make when testing a null hypothesis about the population

1. Type 1 Errors: Rejecting the null hypothesis when it is in fact true
2. Type 2 Errors: Failing to reject the null hypothesis when it is in fact false

In summary,

	Do not reject H_0	Reject H_0
H_0 true	Correct Decision	Type 1 Error
H_0 false	Type 2 Error	Correct Decision

Figure 3.4: Error Types for Hypothesis Testing

Running a Hypothesis Test

Five step process

1. State the Hypotheses: null and alternative
2. Construct a test statistic. Usually it is

$$Z = \frac{\text{(sample statistic)} - \text{(hypothesised population parameter)}}{\text{Standard error of the sample statistic}}$$

   Intuitively, this measures how far (measured in standard errors) the sample statistic is from the hypothesised value of the population parameter
3. State the sampling distribution of the test statistic under the provisional assumption that the Null hypothesis is true. By the Central Limit Theorem (since we are talking about means and proportions), we know that in large samples standardised means are

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

4. Use the Standard Normal Distribution to control the probability of a Type 1 error (rejecting a true H_0)
5. Reject or fail to reject the Null

If our null hypothesis is true, then our test statistic has a standard normal distribution sampling distribution

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

which means that, though its observed value would vary between samples, we would expect it to still be, on average, close to 0. If we observe a Z that is far from 0, then if the null is true, we have witnessed something relatively unlikely. Or the null hypothesis might just not be true. The further Z is from 0 the more unlikely it is that the null is true.

Determining whether the null hypothesis is false

- Using our $N(0,1)$ tables, we know that a $N(0,1)$ random variable will be 1.96 or more standard deviations from 0 about 5% of the time
- So if the null hypothesis is right the probability of our observed test statistic Z being more than 1.96 or less than -1.96 is 0.05. .
- So, if we see a value of $|Z| \geq 1.96$ and thus reject the null hypothesis, we have a 5% chance of being wrong and committing a type 1 error.

Level of Significance

- Denote the probability of a Type 1 error as

$$\mathbb{P}[\text{Reject } H_0 | H_0 \text{ is true}]$$

- Define α as the maximum probability of a Type 1 error we will tolerate

- Using the sampling distribution of Z we can choose a "critical value" c, so that if H_0 is true

$$\mathbb{P}(|Z| > c) \leq \alpha$$

Example. If the maximum probability of a Type 1 error we are prepared to tolerate is 5%, then the critical value is 1.96. Hence, if we observe a sample test statistic greater in absolute value than 1.96, we reject the Null since it is probably not true and we have a 5% chance of getting it wrong.

- α is called the significance level; $1 - \alpha$ is called the confidence level

Example. Running a test for population mean

1. Set up hypotheses you want to test

$$H_0 : \mu = \mu_0 \qquad H_1 : \mu \neq \mu_o$$

2. Calculate the standardised sample mean under the assumption that the Null hypothesis is true, that the population parameter μ equals μ_0:

$$z = \frac{\bar{X} - \mu_0}{SE(\bar{X})}$$

This is your test statistic

3. Apply the CLT to state the distribution of Z under the Null: $Z \sim N(0, 1)$
4. Select a significance level α and find the associated critical value c
5. If $|Z| > c$, reject H_0; otherwise, do not reject H_0

3.6.1 Tests for Means

We consider three types of hypothesis tests regarding the means; the 5-step procedure is always the same. What differs is how you derive the test statistic (which is always [sample statistic - hypothesised population parameter]/SE of sample statistic).

Test statistic for each sample type

1. Single sample of size N from population with mean μ and variance σ^2. Our hypotheses are

$$H_0 : \mu = \mu_0; H_1 : \mu \neq \mu_0$$

Our test statistic is

$$z = \frac{\bar{X} - \mu_0}{\sqrt{\frac{s^2}{N}}}$$

2. Two independent samples of size N_x and N_y from populations with means μ_x and μ_y and variances σ_x^2 and σ_y^2. Our hypotheses are

$$H_0 : \mu_x - \mu_y = \delta_0; H_1 : \mu_x - \mu_y \neq \delta_0$$

Our test statistic is

$$z = \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{s_x^2}{N_x} + \frac{s_y^2}{N_y}}}$$

3. A paired sample of size N of two variables with population means μ_x and μ_y and variances σ_x^2 and σ_y^2. Our hypotheses are

$$H_0 : \mu_x - \mu_y = \delta_0; H_1 : \mu_x - \mu_y \neq \delta_0$$

Our test statistic is

$$z = \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{s_D^2}{N}}}$$

Derivations for independent samples and paired data are exactly the same as their counterparts in Confidence Intervals; the only difference is plugging in the respective sample statistics, hypothesised population parameter and the Standard errors.

3.6.2 Tests for Proportions

Approach closely mirrors the mean

- The first is that we work out the variance of the sampling distribution of a proportion from the proportion itself
- New: In a hypothesis test, everything is done under the null. This therefore includes the calculation of the standard error of the test statistic

Test statistic for each sample type

1. Single sample of size N from population with mean p. Our hypotheses are

   $$H_0 : p = p_0; H_1 : p \neq p_0$$

   Our test statistic is

   $$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{N}}}$$

2. Two independent samples of size N_x and N_y from populations with means p_x and p_y. Our hypotheses are

   $$H_0 : p_x - p_y = \delta_0; H_1 : p_x - p_y \neq \delta_0$$

   Our test statistic is

   $$z = \frac{(\hat{p}_x - \hat{p}_y) - (p_x - p_y)}{\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{N_x} + \frac{\hat{\pi}(1-\hat{\pi})}{N_y}}}$$

3. A paired sample of size N of two variables with population means p_x and p_y. Our hypotheses are

   $$H_0 : p_x - p_y = \delta_0; H_1 : p_x - p_y \neq \delta_0$$

   Our test statistic is

   $$z = \frac{(\hat{p}_x - \hat{p}_y) - (p_x - p_y)}{\sqrt{\frac{s_D^2}{N}}}$$

Tests regarding the Difference in Two Independent Proportions

As usual, the test statistic is

$$z = \frac{(\hat{p}_x - \hat{p}_y) - (p_x - p_y)}{SE(\hat{p}_x - \hat{p}_y)}$$

The Standard Error of the sampling distribution: under the null hypothesis, the population proportions are the same in both samples. That is $p_x = p_y = \pi$ where π is the common value.
Derivation of π. We can estimate it by applying the Law of Iterated Expectations:

$$\hat{\pi} = \left(\frac{N_x}{N_x + N_y}\right)\hat{p}_x + \left(\frac{N_y}{N_x + N_y}\right)\hat{p_y}$$

Hence

$$Var(\hat{p}_x) = \frac{\hat{\pi}(1-\hat{\pi})}{N_x} \qquad Var(\hat{p}_y) = \frac{\hat{\pi}(1-\hat{\pi})}{N_y}$$

Then, since due to independence, $Var(\hat{p}_x - \hat{p}_y) = Var(\hat{p}_x) + Var(\hat{p}_y)$, we yield

$$SE(\hat{p}_x - \hat{p}_y) = \sqrt{\frac{\hat{\pi}(1-\hat{pi})}{N_x} + \frac{\hat{\pi}(1-\hat{\pi})}{N_y}}$$

For paired data, you just have to go find and calculate the sample variance of the difference directly as per usual

3.7 Connection between Confidence Intervals and Hypothesis tests

Confidence Intervals and Hypothesis tests are closely connected. A more rigorous definition of the confidence interval is as the set of nulls consistent with H_0.

Varying the Null Hypothesis

To see what the definition means, consider varying the null. If we conducted a hypthesis test that $\mu = \mu_0$ and set the significance level $\alpha = 0.05$, then we would reject if

$$|Z| = \left| \frac{\bar{X} - \mu_0}{\sqrt{\frac{s^2}{N}}} \right| > 1.96$$

What values of μ_0 would we be just on the border of rejecting the null? To find these set

$$\left| \frac{\bar{X} - \mu_0}{\sqrt{\frac{s^2}{N}}} \right| = 1.96$$

and suppose that $\bar{X} - \mu_0 > 0$. Then,

$$\frac{\bar{X} - \mu_0}{\sqrt{\frac{s^2}{N}}} = 1.96$$

Rearranging, we would just be on the cusp of rejecting the Null if the Null were set to

$$\mu_0 = \bar{X} - 1.96\sqrt{\frac{s^2}{N}}$$

Similarly, if we supposed that $\bar{X} - \mu_0 < 0$, then

$$\frac{\bar{X} - \mu_0}{\sqrt{\frac{s^2}{N}}} = -1.96$$

and thus we would be on the cusp of rejecting the null if

$$\mu_0 = \bar{X} + 1.96\sqrt{\frac{s^2}{N}}$$

Hence, for any value of μ_0 in the range

$$\mu \in \left[\bar{X} \pm 1.96\sqrt{\frac{s^2}{N}} \right]$$

we would fail to reject the Null. This is nothing more than the 95% confidence interval. So, the 95% confidence interval is equal to the set of Null hypothesis values of the population parameter which we couldn't reject in a statistical test at $\alpha = 0.05$.

The argument is symmetric if we start from the confidence interval instead of the significance level.

Confidence intervals and tests of the hypothesis that a population parameter takes a certain value are almost the same thing

- If a value lies outside of the $(1 - \alpha)$ confidence interval, you know even before setting up the formal test, that the hypothesis that the population parameter takes that value is going to be rejected at $\alpha = 0.05$
- Symmetrically, if you reject the hypothesis that a population parameter takes a certain value at $\alpha = 0.05$, you know that that value must lie outside the $(1 - \alpha)$ confidence interval

4 Time Series

4.1 Time Series Data

Definition 4.1 (Time Series Data). Time series data are a sequence of data points recorded in chronological order, where observations are often taken at equally spaced points in time.

Cycles: these occur when a series follows an up and down pattern that is not seasonal. Cyclical variations, like business cycles, have four phases

- Peak
- Recession
- Trough/Depression
- Expansion

The periodicity of cycles is variable; the duration of each phase in the cycle also varies.

Linearity of time series: The larger the absolute value of $Corr(X, Y)$, the stronger the linear association between X and Y since correlation is the measure of association between two random variables X and Y

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

Serial correlation/autocorrelation: Some facts

- X_t denotes a variable observed at time t
- We think of X_t where $t = 1, 2, 3, \ldots$ as a sequence of random variables, any pair of which may be correlated
- Lags:
  The first lag of the variable is the value in the period before X_{t-1}; the second lag of the variable is the value of two periods before is X_{t-2}
  The h'th lag is X_{t-h}

$Corr(X_s, X_t)$ is called an autocorrelation since it is a correlation between the time series and its own past self.

Basic facts

1. Time series data is to provide a simple model of the evolution of a variable
2. Time series data is to provide a basis for prediction/forecasting
3. While cross section data are independent, time series data are not since the second observation is often dependent on the first

We often transform data to linearise data trends (by taking logarithms) like for economic growth. The standard deviation of many economic time series is approximately proportional to their level. Therefore, the standard deviation of the logarithm of such a series is approximately constant.

Definition 4.2 (Population of autocorrelation): The autocorrelated between X_t and X_{t-h} is often denoted $\rho(h)$ and its population definition is

$$\rho(h) = \frac{Cov(X_t, X_{t-h}}{\sqrt{Var(X_t)Var(X_{t-h})}}$$

Recall that

$$-1 \leq \rho(h) \leq +1$$

and by construction $\rho(0) = 1$.
When we write out a lagged variable we lose one observation. When we calculate the sample autocorrelation of any given series with a fixed sample size T, we cannot put too much confidence in the values of for large lags, since fewer pairs of current/lagged variables will be available for computing it when the lag is large. Most important attribute that time series data have is serial correlation/autocorrelation that cross-section data does not have.
The plot of the autocorrelation against the lag is called the correlogram.

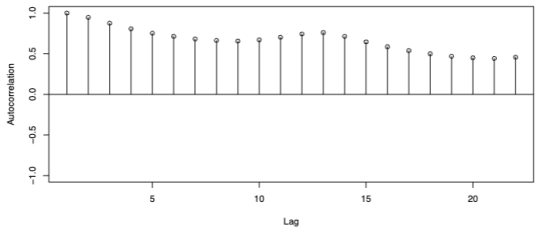


Figure 4.1: Correlogram

4.2 Basic Operations

Some Notations and Basic Operations

- X_t denotes a variable observed at time t
- Lags: The j'th lag is X_{t-j} (j is the number of periods before X_t)
- Leads: the j'th lead is X_{t+j} (j is the number of periods after X_t)
- Initial condition: Occasionally, we trace a time series back to an initial period X_0
- Differences:
  1. the first difference ΔX_t is the difference between the value in period t and $t-1$:
  $$\Delta X_t = X_t - X_{t-1}$$
  2. If your data are quarterly then the annual difference is the fourth difference of the time series:
  $$\Delta X_t = X_t - X_{t-4}$$
- Growth: The growth rate between period t and $t-1$ is the factor such that
  $$X_t = (1+g)X_{t-1}$$
  Rearranging:
  $$g = \frac{X_t}{X_{t-1}} - 1$$
- Average growth rate: The average growth rate over non-adjacent periods X_t and X_{t-h} is the factor g such that
  $$X_t = (1+g)(1+g)\dots(1+g)X_{t-h} \qquad = (1+g)^h X_{t-h}$$
  Rearranging, we yield
  $$g = \left(\frac{X_t}{X_h}\right)^{\frac{1}{h}} - 1$$
  The average growth rate between these non-adjacent periods does not equal the arithmetic mean of all the per-period growth rates in between

4.3 Deterministic & Stochastic Time Series

Characterisation 4.1 (Deterministic Time Series). A deterministic time series is one which can be expressed explicitly by an analytic expression and has no random or probabilistic aspects. Its past and future are completely specified so we can always predict its future behaviour and state how it behaved in the past.

Example. The simplest deterministic model is a linear one

$$X_t = \beta_0 X_0 + \beta_1 t$$

Or an aperiodic time series (cannot find a finite value t corresponding to a repetition period)

$$X_t = sin(\pi t) + sin(\sqrt{2}\pi t)$$

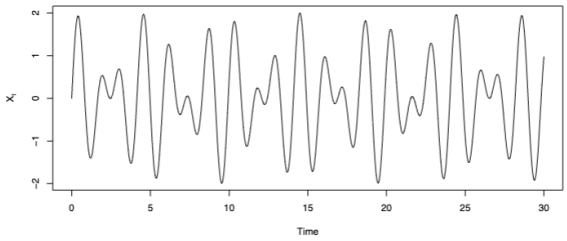


Figure 4.3: Aperiodic time series

Or a linear trend combined with an aperiodic deterministic cycle.

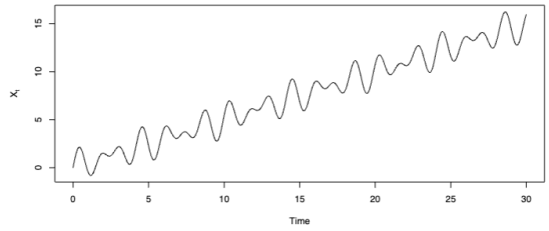


Figure 4.4: Linear + aperiodic time series

Characterisation 4.2 (Non-deterministic Time Series). A non-deterministic time series is one which cannot be described by an analytic expression. It has some as-if random aspect that prevents its behaviour from being described explicitly. A time series may be non-deterministic because

- All the information necessary to describe it explicitly is not available, although it might be in principle
- The nature of the generating process is inherently random

Since non-deterministic time series have a random aspect, they follow a probabilistic rather than deterministic laws.

Examples (Non-deterministic Time series).
Each successive observation is drawn, independently, from the distribution $N(0, 1)$.

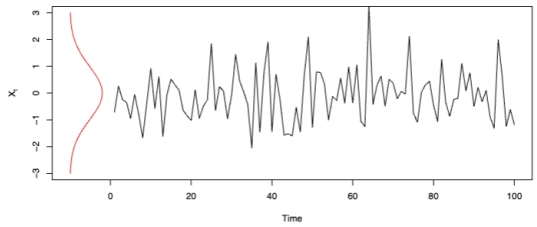


Figure 4.5: Non-deterministic time series

Stochastic process where in which the first half are drawn, independently from an $N(0,1)$ and the second half from an $N(5,1)$. Looks like a structural break.

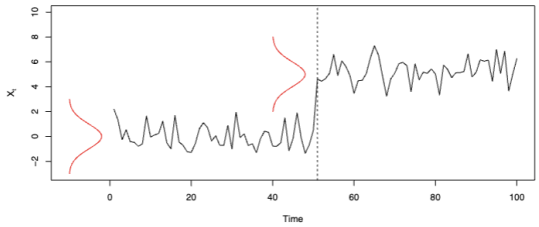


Figure 4.6: Non-deterministic time series with structural break

Stochastic process in which the data are drawn independently from an $N(\mu,1)$ distribution but the mean μ is increasing over time. Looks like a trend.

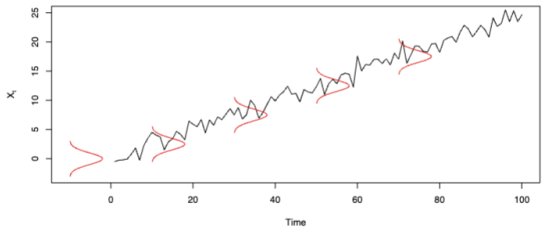


Figure 4.6: Non-deterministic time series with increasing trend

Stochastic process in which data are drawn independently from an $N(0,\sigma)$ distribution but the variance is increasing over time. Volatility is increasing.

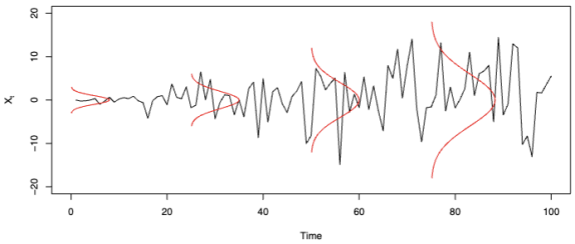


Figure 4.7: Non-deterministic time series with increasing volatility

4.4 Stationarity

Characterisation 4.3 (Stationarity). When the random process generating the stochastic component of the time series is fixed. Similar to the 'iid' assumption because the distribution is identical over time.

There is nothing statistically significant about the segment of history that you observe in the sense that the statistical properties of the process generating the data are invariant to shifts in the window of observation; the statistical properties of the time series are invariant to absolute location in time.
Relative positions may matter, but the absolute position does not. A time series is stationary if its underlying statistical structure does not evolve with time.

Definition 4.3 (Weak Stationarity). A stochastic process is weakly stationary if

- The mean, and the variance do not vary with the date, i.e.
  1. $\mathbb{E}[X_t]$ does not vary with time
  2. $Var[X_t]$ does not vary with time
- The covariance between values at different dates does not depend on the date but on the time-difference between dates, i.e.
  1. $Cov(X_t, X_{t-h})$ does not vary with time but only on h
  2. $Corr(X_t, X_{t-h})$ does not vary with time but only on h

A stationary process has the property that the mean, variance and autocorrelation structure do not change over time. For our purposes, when we refer to stationarity we mean a flat looking series, without trend, constant variance over time, a constant autocorrelation structure over time and no periodic fluctuations (seasonality).

Transforming Non-stationary time series into stationary time series

If the time series is not stationary, we can often transform it to a stationary time series with one of the following techniques.

1. **Differencing the Data:** Given the series X_t, we create a new series $Y_t = \Delta X_t = X_t - X_{t-1}$. The differenced data will contain one less point than the original data. Although you can difference the data more than once, one difference is often sufficient.
2. **Fitting Curve/Moving average:** If the data contain a trend, we can fit some type of curve to the data and then model the residuals from that fit. Since the purpose of the fit is to simply remove long term trend, a simple fit, such as a straight line, is often used. Alternatively, we can calculate a moving average.
3. **Taking logarithms/square root:** For non-constant variance, taking the logarithm or square root of the series may stabilise the variance. For negative data, you can add a suitable constant to make all the data positive before applying the transformation

4.5 Time Series Models

Basic Time Series Models

1. **White Noise Processes:**
   Definition 4.4 (White Noise Process). A time series is a white noise series if the $X_1, X_2, \ldots$ are independent and identically distributed (iid) with a mean of zero and some finite variance.
   By definition, white noise processes are serially uncorrelated. White noise processes don't have to come from normal distributions but if they do then they are called Gaussian White Noise and they have a distribution which is $N(0, \sigma^2)$. White noise processes are stationary - neither their mean nor their variance nor their degree of serial correlation change over time.
   Plots of white noise series exhibit a very erratic, jumpy, unpredictable behaviour. The autocorrelation plot shows that there is no serial correlation.

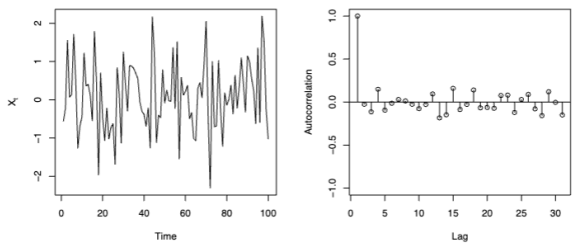


Figure 4.8: Gaussian White Noise

Since X_t are un-autocorrelated, previous values do not help us forecast future values. White noise series themselves are uninteresting from a forecasting standpoint since they are not linearly forecastable, but they are essential as they form one of the basic building blocks for more general models. In economic time series, the white noise series is often thought of as representing innovations, or shocks. Represents the aspects of the time series of interest which could not have been predicted in advance.

2. **Random Walks:**
A random walk model says that the value at time t will be equal to the last period value plus a stochastic (non-systematic) component that is a white noise (i.e. could be up or down but on average will be zero).
Characterisation 4.4 (Random Walk). A random walk is a time series where

$$X_t = X_{t-1} + e_t$$

and e_t is a white noise series. Often it is assumed that $e_t \sim N(0, \sigma^2)$. So this is a hybrid deterministic/stochastic model.

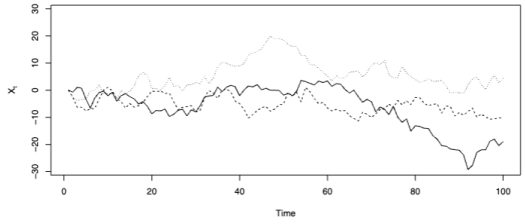


Figure 4.9: Gaussian Random Walks

Figure 4.9 shows 3 Gaussian Random walks. They are completely independent from each other, yet correlated.
A random walk can move away from its starting point in either a positive or negative direction but it will eventually come back. For instance, the mean of X_t is constant under a random walk

$$\begin{aligned}\mathbb{E}[X_t] &= \mathbb{E}[X_{t-1} + e_t] \\ &= \mathbb{E}[X_{t-1}] + \mathbb{E}[e_t] \\ &= \mathbb{E}[X_{t-1}] \qquad [\mathbb{E}[e_t] = 0]\end{aligned}$$

So, for a random walk

$$\mathbb{E}[X_t] = \mathbb{E}[X_{t-1}]$$

and thus the mean does not depend on time.
However, despite the mean being constant, a random walk is not stationary. This is because its variance

does depend on time.

$$\begin{aligned} Var(X_t) &= Var(X_{t-1} + e_t) \\ &= Var(X_{t-1}) + Var(e_t) \qquad [e_t \text{ is independent of } X_{t-1}] \\ &= Var(X_{t-1}) + \sigma^2 \qquad [e_t \text{ has variance } \sigma^2] \end{aligned}$$

To see this more clearly, we take

$$Var(X_t) = Var(X_{t-1}) + \sigma^2$$

and backdate by one period

$$Var(X_{t-1}) = Var(X_{t-2}) + \sigma^2$$

and substitute in recursively

$$\begin{aligned} Var(X_t) &= Var(X_{t-2}) + 2\sigma^2 \\ Var(X_t) &= Var(X_{t-3}) + 3\sigma^2 \\ &\vdots \\ Var(X_t) &= Var(X_0) + t\sigma^2 \end{aligned}$$

So the variance of the random walk at time t is a linear function which depends on time.

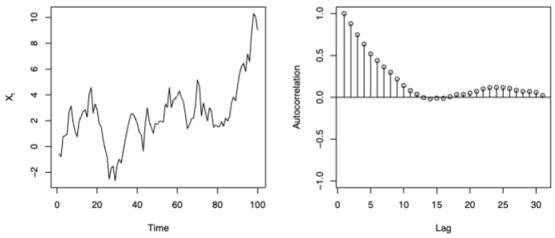


Figure 4.10: A Random Walk with its associated ACF plot

Random walks are highly persistent and have a high degree of autocorrelation out to long lags.

3. **Random Walks with Drift:**
A slight generalisation is called the random walk with drift. A random walk with drift predicts that the value at time t will be equal to the last period value plus a constant movement up or down given by θ_0, plus a random component that is white noise. It is written

$$X_t = \theta_0 + X_{t-1} + e_t$$

The resulting series is more strongly tended the further θ_0 is from zero.

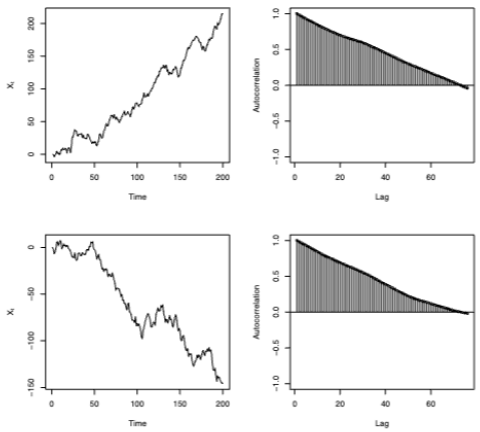


Figure 4.11: Two Gaussian random walks with drift

The first case has $\theta_0 = 1$ and the second line has $\theta_0 = -1$. In both cases, the series are strongly positively autocorrelated out to long lags.

4. **Autoregressive processes:**
An autoregressive model relates a time series variable to its past values. Autoregressive models of order p, abbreviated $AR(p)$, are commonly used in time series analyses. An $AR(p)$ model relates X_t to p lags of itself. For instance, $AR(2)$ relates X_t to X_{t-1} and X_{t-2}.
An $AR(1)$ process is

$$X_t = \theta_0 + \theta_1 X_{t-1} + e_t$$

where e_t is the white noise.
The models we have seen up to know are all special cases of an $AR(1)$

- Random Walk with Drift is an $AR(1)$ model where $\theta_1 = 1$
- Random Walk is an $AR(1)$ model where $\theta_1 = 1$ and $\theta_0 = 0$
- White Noise is an $AR(1)$ model where $\theta_1 = 0$ and $\theta_0 = 0$

4.6 Mean Reversion

Mean Reversion is a property of a stationary $AR(1)$ process.

Conditions for which $AR(1)$ is stationary

There exists parameter values such that an $AR(1)$ process is non-stationary (random walk with drift).

Condition 1: Mean is constant and invariant with time
Take $AR(1)$

$$X_t = \theta_0 + \theta_1 X_{t-1} + e_t$$

And take expectations

$$\begin{aligned}\mathbb{E}[X_t] &= \mathbb{E}[\theta_0 + \theta_1 X_{t-1} + e_t] \\ &= \theta_0 + \theta_1 \mathbb{E}[X_{t-1}] + \mathbb{E}[e_t] \\ &= \theta_0 + \theta_1 \mathbb{E}[X_{t-1}] \qquad [\mathbb{E}[e_t] = 0]\end{aligned}$$

Since $\mathbb{E}[X_t] = \mathbb{E}[X_{t-1}]$,

$$\mathbb{E}[X_t] = \theta_0 + \theta_1 \mathbb{E}[X_{t-1}]$$
$$\mathbb{E}[X_t] = \frac{\theta_0}{1-\theta_1}$$

So $\mathbb{E}[X_t]$ is independent of t and the mean is not a function of time.

Condition 2: Variance is constant and invariant with time
The variance of X_t is given by

$$\begin{aligned} Var(X_t) &= Var(\theta_0 + \theta_1 X_{t-1} + e_t) \\ &= Var(\theta_0) + Var(\theta_1 X_{t-1}) + Var(e_t) \qquad [e_t \text{ is independent of } X_t] \\ &= \theta_1^2 Var(X_{t-1}) + \sigma^2 \qquad [\theta_0 \text{ is a constant, } e_t \text{ has a variance } \sigma^2] \end{aligned}$$

If the variance is constant then $Var(X_t) = Var(X_{t-1})$, therefore

$$Var(X_t) = \frac{\sigma^2}{1-\theta_1^2}$$

Condition 3: $Corr(X_t, X_{t-h})$ was independent of time and only depends on the lag
Recall that

$$\begin{aligned} Corr(X_t, X_{t-h}) &= \frac{Corr(X_t, X_{t-h})}{\sqrt{Var(X_t)Var(X_{t-h})}} \\ &= \frac{Cov(X_t, X_{t-h})}{Var(X_t)} \qquad \text{Constant variance} \end{aligned}$$

Values of variances, covariances and correlations are not affected by the specific value of the mean. So now, we are going to assume that the data have mean 0, which happens when $\theta_0 = 0$ and $X_t = \theta_1 X_{t-1} + e_t$ (from the $AR(1)$ model). This is not necessary, but it specifies things mathematically. Thus,

$$Cov(X_t, X_{t-h}) = \mathbb{E}[X_t X_{t-h}]$$

To get the covariance, take the model for X_t and use the mean zero assumption to write $AR(1)$ as

$$X_t = \theta_1 X_{t-1} + e_t$$

Now, multiply each side of the equation by X_{t-h} to yield

$$X_t X_{t-h} = [\theta_1 X_{t-1} + e_t] X_{t-h}$$
$$X_t X_{t-h} = \theta_1 X_{t-1} X_{t-h} + e_t X_{t-h}$$

Then, we take expectations,

$$\begin{aligned} \mathbb{E}[X_t X_{t-h}] &= \mathbb{E}[\theta_1 X_{t-1} X_{t-h} + e_t X_{t-h}] \\ &= \mathbb{E}[\theta_1 X_{t-1} X_{t-h}] + \mathbb{E}[e_t X_{t-h}] \\ &= \theta_1 \mathbb{E}[X_{t-1} X_{t-h}] \qquad [\text{independence of } e_t \text{ from } X_{t-h}] \\ &= \theta_1 Cov(X_{t-1}, X_{t-h}) \end{aligned}$$

From this, for $h = 1$

$$Cov(X_t, X_{t-h}) = \theta_1 Cov(X_{t-1}, X_{t-1}) = \theta_1 Var(X_{t-1})$$

For $h = 2$

$$Cov(X_t, X_{t-2}) = \theta_1 Cov(X_{t-1}, X_{t-2})$$

Using the $h = 1$ result

$$Cov(X_t, X_{t-1}) = \theta_1 Var(X_{t-1}$$

and backdating one period

$$Cov(X_{t-1}, X_{t-2}) = \theta_1 Var(X_{t-2})$$

Substituting gives

$$Cov(X_t, X_{t-2}) = \theta_1\theta_1 Var(X_{t-2}) = \theta_1^2 Var(X_{t-2})$$

We can do so recursively to yield

$$\begin{aligned} Cov(X_t, X_{t-3}) &= \theta_1^3 Var(X_{t-3}) \\ Cov(X_t, X_{t-4}) &= \theta_1^4 Var(X_{t-4}) \\ &\vdots \\ Cov(X_t, X_{t-h}) &= \theta_1^h Var(X_{t-h}) \\ &= \theta_1^h Var(X_t) \end{aligned}$$

Since there is constant variance where $Var(X_t) = Var(X_{t-h})$ for all h. Therefore,

$$Corr(X_t, X_{t-h}) = \frac{\theta_1^h Var(X_t)}{Var(X_t)} = \theta_1^h$$

Hence, autocorrelation between periods depend on time lag between periods (denoted by h), but not the absolute location of the periods in time.

The slope of the $AR(1)$ function

$$X_t = \theta_0 + \theta_1 X_{t-1} + e_t$$

is also the one-lag autocorrelation.

- As we look at observations further apart the autocorrelation is exponentially scaled
- Because $|\theta_1| < 1$ (see above) this means that the autocorrelation is declining exponentially

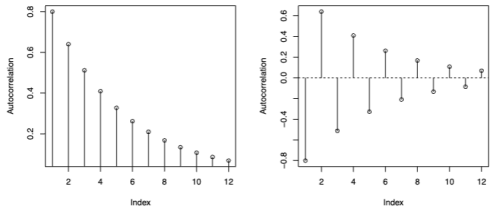


Figure 4.12: Autocorrelation for $theta_1 = 0.8$ and $theta_1 = -0.8$ respectively

Mean Reversion

Example. Consider an $AR(1)$ process linking heights in successive generations

$$X_t = \theta_0 + \theta_1 X_{t-1} + e_t$$

Assume the distribution of height in the population is the same in each generation with a mean of mu and a variance of σ^2 – a stationarity assumption. From the preceding discussion of the properties of a stationary $AR(1)$ we know that the lag coefficient $theta_1$ is equal to the intergenerational autocorrelation coeffieicnt of the height, and that the coefficient θ_0 is equal to $(1 - \theta_1)\mu$. Hence,

$$X_t = (1 - \theta_1)\mu + \theta_1 X_{t-1} + e_t$$

Taking the conditional expectation:

$$\mathbb{E}[X_t|X_{t-1}] = (1-\theta_1)\mu + \theta_1 X_{t-1}$$

Given the correlation coefficient is positive (which it is here) then this is just the weighted average of the population mean height μ and the parent's height X_{t-1}. Since the correlation is between 0 and 1 so Junior's expected height will be between their parent's and the population mean: if Mum/Dad was tall, Junior can expect to be less so – they will be expected to "regress to the mean".
This formula comes from having a stationary distribution – nothing causal about it.

Example. Consider a class of students take a 100-item true/false test on a subject.
First suppose that all students chose randomly on all questions. Then, each student's score would be a realisation of one of a set of independent and identically distributed random variables, with an expected mean of 50. Naturally, some will score above and below 50 by chance.
If one selects only the top 10% of the students and gives them a second test in which they randomly choose answers again, the expectation would be close to 50 again. So, the mean of the students would "regress" all the way back to the mean of all students who took the original test.

Now suppose the answers to the test questions were not random; then the students would be expected to score the same on the second test as they scored on the original test, and there would be no regression toward the mean.

Insights

- To the extent that a score is determined randomly, or that a score has random variation or error, as opposed to being determined by the student's academic ability or being a "true value", mean reversion will have an effect.

4.7 Spurious Correlation

Time series variables are often correlated, leading to incorrect causal inference.

Nonsense Correlation

Intuitive: stuff like divorce rate in Maine against per capita consumption of margarine are correlated but obviously there is no causal interaction between them.

Spurious correlation

Characterisation 4.5 (Spurious Correlation). Spurious correlations describes correlation which can often occur between two random walks without drift.

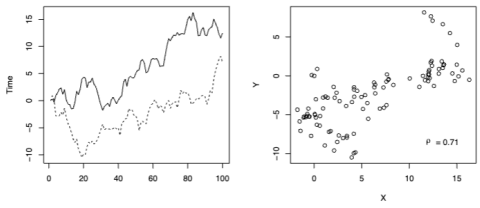


Figure 4.13: Spurious Correlation

These figures shows a simulation of two random walks without drift that are completely independent, yet have a strong positive correlation.

Some facts

- Not rare
- If you simulate and generate lots of independent pairs of time series, the absolute value of the correlation coefficient is roughly uniform over $[-1, +1]$
- So nearly half the time you would expect to see a correlation stronger than 0.5 (in either direction) between completely independent time series

5 Causal Inference

5.1 Potential Outcomes

Neyman-Rubin Causal Model

A Framework of POTENTIAL OUTCOMES.
Observed data

- An outcome variable Y
- A treatment variable D. This treatment is binary: if you get it, $D = 1$; else $D = 0$ (you do not get it)

Definition 5.1 (Outcome variable): The outcome variable Y is an observed outcome depending on whether you get the treatment. Written as

$$Y(D)$$

For each individual in the population, we postulate the existence of two potential outcomes

- $Y(O)$: the outcome under no treatment
- $Y(1)$: the outcome under treatment

Definition 5.2 (Causal Effect): The difference between the potential outcomes

$$Y(1) - Y(0)$$

is called the causal effect of the treatment. Causal effects of treatment varies across people just as potential outcomes vary across people; thus we are more interested in the population average of the distribution of causal effects.

Definition 5.3 (Average Treatment Effect): The Average Treatment Effect is

$$ATE = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$$

Definition 5.4 (Average Treatment on the Treated): The ATT is

$$ATT = \mathbb{E}[Y(1)|D = 1] + \mathbb{E}[Y(0)|D = 1]$$

We take the conditions means when $D = 1$.

However, we do not get to see the full dataset. This is because the treatment determines which of the potential outcomes actually happens and is therefore observed.

- When $D = 0$, we observe $Y(0)$
- When $D = 1$, we observe $Y(1)$

This is because, the observed outcome is linked to the potential outcomes as follows

$$Y = Y(0)(1 - D) + Y(1)(D)$$

therefore,

- When $D = 1$, $Y = Y(1)$
- When $D = 0$, $Y = Y(0)$

So we only get to observe

Person	D	Observed Outcomes Y	
1	1		$Y_1(1)$
2	0	$Y_2(0)$	
3	1		$Y_3(1)$
4	1		$Y_4(1)$
5	0	$Y_5(0)$	
6	0	$Y_6(0)$	
7	0	$Y_7(0)$	

Figure 5.1: Observable dataset

The unobservable potential outcome is known as the 'counterfactual'

5.2 Selection Bias

If we calculate the average observed outcomes for the two groups and compare them, we are comparing $\mathbb{E}[Y(0)|D = 0]$ with $\mathbb{E}[Y(1)|D = 1]$.

Theorem 5.1 (Fundamental Equation of Causal Inference): The fundamental equation of causal inference is the (Observed differences in group averages) = (average effect of treatment of the treated) + (selection bias or formally)

$$(\mathbb{E}[Y(1)|D = 1] - \mathbb{E}[Y(0)|D = 0]) = (\mathbb{E}[Y(1)|D = 1] - \mathbb{E}[Y(0)|D = 1]) + (\mathbb{E}[Y(0)|D = 1] - \mathbb{E}[Y(0)|D = 0])$$

Therefore,

Definition 5.4 (Selection Bias): Selection bias is the average difference in the no treatment outcome between the treated and un-treated groups.

$$SB = \mathbb{E}[Y(0)|D = 1] - \mathbb{E}[Y(0)|D = 0]$$

It is a comparison between what actually happened to the untreated in the group and what would have happened to the treated group if they had not been treated. Selection bias reflects the idea that bias will arise if individuals are selected for treatment on the basis of their potential outcomes.

Observed difference would be an over-estimate due to selection bias if

$$\mathbb{E}[Y(0)|D = 1] > \mathbb{E}[Y(0)|D = 0]$$

Example. Consider the fact that it is observed that Oxford Students earn \$4000 more on average than another university, where $D = 1$ is if the student has gone to Oxford. However, since Oxford selects high-ability students, it is the case that the already high ability student would do better on average than the un-treated individual who has gone to another university. Hence,

$$\mathbb{E}[Y(0)|D = 1] > \mathbb{E}[Y(0)|D = 0]$$

and \$4000 is likely an over-estimate.

Selection bias is endemic in observational microeconomic data since it is often related to rational choice; if treatment is chosen on the basis of potential outcomes (e.g. to maximise the outcome), then there is potential for selection bias. SELECTION BIAS IS COMPLETELY DIFFERENT FROM SAMPLE BIAS (we are dealing with population in causal inference).

5.3 Randomised Controlled Trials

We look at RCTs to avoid selection bias and the importance of independence of treatment assignment with respect to potential outcomes.

Randomisation of treatment makes treatment independent of potential outcomes. For example, if you split the population randomly into two groups you would expect the average income or any variable to be the same on average in the two groups (same for potential outcomes). So, mean potential outcomes are identical for the treated and untreated group. This does not mean that outcomes are independent of treatment since treatments determine outcomes – independence is required with respect to potential outcomes.

Random assignment of treatment implies that

$$\mathbb{E}[Y(0)|D] = \mathbb{E}[Y(0)]$$
$$\mathbb{E}[Y(0)|D = 1] = \mathbb{E}[Y(0)|D = 0]$$

Hence, Difference in group means = ATT since selection bias is zero as selection bias is given by

$$\mathbb{E}[Y(0)|D = 1] - \mathbb{E}[Y(0)|D = 0] = 0$$

Randomisation ensures that differences between the treatment and non-treatment groups have a causal interpretation since the Observed differences in group means = ATT

$$\mathbb{E}[Y(1)|D = 1] - \mathbb{E}[Y(0)|D = 0] = \mathbb{E}[Y(1)|D = 1] - \mathbb{E}[Y(0)|D = 1]$$

This is because Randomisation also ensures that the ATE and the ATT are the same and equal to the observed differences in means between the treatment and control groups

$$\mathbb{E}[Y(1)|D = 1] - \mathbb{E}[Y(0)|D = 0] = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$$

5.4 Internal and External Validity

While RCTs are usually considered the best possible approach to the study of causal effects, we must consider if

- A study has internal validity.
  Definition 5.5a (Internal Validity): A study is said to have internal validity if its findings for the sample are credible
- A study has external validity.
  Definition 5.5b (External Validity): A study is said to have external validity if its findings can be credibly extrapolated to the population or Real World policy of interest

Internal Validity

- Contamination:
  - People in the control group access the treatment anyway
  - Some of the untreated turn out to be treated
- Non-compliance:
  - Individuals who are offered a treatment refuse to take it
  - Some of the treated turn out to be untreated
- Hawthorne effect: A phenomenon in which participants alter their behaviour as a result of being part of an experiment or study

- Placebo effect: the placebo effect impacts final outcomes because of perceived changes (different from Hawthorne effect where the outcome changes due to imperceptible changes)

The first 2 re-introduce selection bias.

External Validity. If an RCT has external validity, it means that we can expect the distribution of outcomes that would occur in the population under the policy of interest would be the same as the distribution of outcomes realised by the experimental treatment group. Properly executed RCT's have high internal validity, but you need external validity too for credible policy evaluation. The internal validity of RCT's does not imply external validity, credible policy evaluation requires both.

Threats to external validity

- The sample used in an RCT may differ from the population of policy interest due to the small scale/local nature of the RCT, (e.g. a particular geographic area, institutional environment or demographic group).
- Establishing external validity requires enumerative and eliminative induction: doing a lot of RCTs and varying the circumstances under which the RCT takes place and seeing which aspects of the environment matter/don't matter to the outcome
- The assumption of individualistic treatment response may be valid within the design of the RCT but may not hold in the population where there is the potential for spillover effects
- Serious measurement problems when studies have short durations because we often want to learn long-term outcomes of treatments, but short studies reveal only immediate, surrogate outcomes. Even well-conducted RCT's may only reveal the distribution of surrogate outcomes in the study population, not the distribution of outcomes of real interest

5.5 Natural and Quasi-Experiments

Look at Natural and Quasi-experiments since we can't always run RCTs.

Basically, look at naturally occurring phenomenon like looking at the effects of neighbourhoods on economic outcomes by studying the effects of a tornado which passed through a community situated in Tornado Alley etc. Nature randomly (in the relevant sense of being independent of potential outcomes) allocates families to treatment/control by destroying some families' homes and leaving others' standing. Some households are then relocated to different neighbourhoods. Social scientists can then compare outcomes knowing that selection bias is not a significant factor.

Institutional arrangements can sometimes play a similar role of dividing people into the treatment and control group in a way which is independent on potential outcomes. For example, Angrist and Lavy studied the effect of class size on educational outcomes. The twelfth century rabbinic scholar Maimonides proposed a maximum class size of 40. This same maximum combined with fluctuating school enrollment (which is plausibly independent of potential outcomes) induces a quasi-experimental variation in class size in Israeli public schools today. They used Maimonides' rule of 40 to construct estimates of effects of class size on test scores.

5.6 Conditional Independence

The basic motivation is that if neither an RCT nor a natural nor quasi experiment is available, you can still learn about causal effects by studying observational data. When selection into treatment is not "as good as random" but is instead determined by observables, we can condition on those observables (i.e. hold them constant) and then look at treated and non-treated groups for fixed values of those variables. The assumption is that this removes selection on observables and all the variation that remains is independent of potential outcomes.

Conditioning

Suppose we can observe a list of things about the individual, beyond outcomes and treatment, called "covariates" like personal characteristics, information on education attainment etc.

Definition 5.5 (Covariates): Covariates are a list of observables, denoted by X.

To keep things simple, we think of X as a single variable and the data becomes

$$\{Y, D, X\}$$

The original independence requirement was that assignment to treatment was independent of potential outcomes such that $D \perp\!\!\!\perp \{Y(0), Y(1)\}$ and randomisation delivered this.

The conditional independence assumption (CIA) says assignment to treatment is independent of potential outcomes conditional on covariates.

$$D_i \perp\!\!\!\perp \{Y(0), Y(1)\}|X$$

Thus, potential outcomes don't vary with treatment holding other factors about the individual fixed.

Given CIA, one focuses on particular values of the covariates where $X = x$ and hold those fixed across the fundamental equation of causal inference

$$\begin{aligned}&\mathbb{E}[Y(1)|D = 1, X = x] - \mathbb{E}[Y(0)|D = 0, X = x] = \\ &\mathbb{E}[Y(1)|D = 1, X = x] - \mathbb{E}[Y(0)|D = 1, X = x] + \mathbb{E}[Y(0)|D = 1, X = x] - \mathbb{E}[Y(0)|D = 0, X = x]\end{aligned}$$

Everything is conditional on the covariates taking particular fixed values. Using exactly the same logic as before, what conditional independence buys us is

$$\mathbb{E}[Y(0)|D = 1, X = x] = \mathbb{E}[Y(0)|D = 0, X = x]$$

but this time we are holding the other covariates fixed. This makes the conditional selection bias term disappear and

$$(\text{Observed difference in averages}|X = x) = ATT(x) = ATE(x)$$

We can just divide the data into cells and take sample averages of outcomes for treated and untreated observations within each cell to get the $ATE(x)$ for each sub-population. Getting the overall ATE and ATT is then just a matter of averaging over cells/sub-populations (using the LIE)

$$\begin{aligned}\mathbb{E}[ATE] &= \mathop{\mathbb{E}}_{x}[ATE(x)] \\ \mathbb{E}[ATT] &= \mathop{\mathbb{E}}_{x}[ATT(x)]\end{aligned}$$

Problems with conditional independence

1. The credibility of the Conditional Independence Assumption.
   The CIA is often contestable: it rests on the assumption that you have successfully controlled for all sources of non-random selection.

2. The common support problem/curse of dimensionality.
   When we apply the CIA we need to calculate objects like

   $$\mathbb{E}[Y(1)|D = 1, X = x] \text{ and } \mathbb{E}[Y(0)|D = 0, X = x]$$

   i.e. the average outcomes for the treated and the untreated in a particular slice of the population (defined by the covariates taking certain values). You need to take both of this objects.
   Since we are working with sample data, if we need to specify the conditioning variables very finely in order to make the CIA plausible we may be left with very few or no observations in one of the group.

This is due to high-dimensionality of data rendering data sparse.
Consider a 5D unit cube with 100 observations in it and consider a cell of length 0.2. We would expect to find

$$100 \times 0.2^5 = 0.035$$

observations in a cube with side 0.2 (basically one cell). To get an expected number of 5 observations in cell would require a 5D unit cube whose side is 0.55 which is more than half the range in each dimension.
So in splitting up variables finely, we may have very few observations to work with and average over.

3. Bad controls.
Simpson's Paradox. When trends appear in dfferent groups of data but disappear when combined. Manifests in gender pay gap studies, where studies do not control for occupation and look at the pay gap between men and women across occupations. However, the reason why one may not condition on occupation is because the factor which you are conditioning on may itself be an outcome. For instance, if being a woman causes one to have differential access to different jobs, then that creates a compositional effect which invalidates conditioning on occupation.
Example. Consider

- Two types of people: {female, male}
- Two types of jobs: {Professional, Non-professional}

Let Y denote earnings and let $D \in \{0, 1\}$ denote sex

- Let $D = 0$ indicate male
- Let $D = 1$ indicate female

D will be "the treatment". We are interested in the causal effect of D on Y.
Let $W \in \{0, 1\}$ denote occupation

- Let $W = 0$ indicate a non-professional job
- Let $W = 1$ indicate a professional job

Suppose that sex has a causal effect on both earnings and also on the occupations which one can access. Then we have the usual links between potential and actual outcomes

$$Y = Y(1)D + Y(0)(1 - D)$$
$$W = W(1)D + W(0)(1 - D)$$

$Y(1)$ represents a person's potential earnings if they are female; $Y(0)$ if they are male. $W(1)$ represents a person's potential occupation if they are female; $W(0)$ if they are male. 4 possible permutations

- $W(1) = 1$ and $W(0) = 1$ indicates a person would have landed a professional job regardless of sex. When both equal 0, holds for non-professional job
- If $W(1) = 1$ and $W(0) = 0$, a person would have landed the professional job if they were a female but not if they were a man; the opposite case holds

So, we now look at what happens when you compare earnings between sexes conditional on occupation. Consider the observed differences in earnings between females and males in professional jobs:

$$\mathbb{E}[Y|W = 1, D = 1] - \mathbb{E}[Y|W = 1, D = 0]$$

In terms of potential outcomes this is

$$\mathbb{E}[Y(1)|W(1) = 1, D = 1] - \mathbb{E}[Y(0)|W(0) = 1, D = 0]$$

If we think that sex is independent of potential earnings and occupational access, then we can drop the conditioning on D and write the observed difference in average earnings between females and males in professional jobs as

$$\mathbb{E}[Y(1)|W(1)=1] - \mathbb{E}[Y(0)|W(0)=1]$$

Given the fundamental equation of causal inference, we have

$$\begin{aligned}&\mathbb{E}[Y(1)|W(1)=1] - \mathbb{E}[Y(0)|W(0)=0] = \\ &\mathbb{E}[Y(1)|W(1)=1] - \mathbb{E}[Y(0)|W(1)=1] + \mathbb{E}[Y(0)|W(1)=1] - \mathbb{E}[Y(0)|W(0)=1]\end{aligned}$$

The causal object of interest is the ATT

$$ATT = \mathbb{E}[Y(1)|W(1)=1] - \mathbb{E}[Y(0)|W(1)=1]$$

This is the difference between

- $\mathbb{E}[Y(1)|W(1)=1]$ which is the observed earnings of females in professional jobs
- $\mathbb{E}[Y(0)|W(1)=1]$ which is the counterfactual earnings which females in professional jobs would have got had they been males

The selection bias term is

$$SB = \mathbb{E}[Y(0)|W(1)=1] - \mathbb{E}[Y(0)|W(0)=1]$$

This is the difference between

- $\mathbb{E}[Y(0)|W(1)=1$ which is the counterfactual earnings which females in professional jobs would have got had they been males
- $\mathbb{E}[Y(0)|W(0)=1]$ which is the observed earnings for males in the professional job

If

$$\mathbb{E}[Y(0)|W(1)=1] \neq \mathbb{E}[Y(0)|W(0)=1]$$

then we have some form of selection bias. It is not unreasonable to argue that it is more difficult for females to attain high-paying jobs than men and thus females in those jobs would be more talented and hence their earning potential is higher. If this is true, then on average, the earnings that females who got into one of the Professions would have earned had they been male would be higher than the earnings of males in that profession.
So the selection bias would be positive.

Since we are examining the claim that once you condition on occupation, the gender pay gap disappears

$$\mathbb{E}[Y|W=1,D=1] - \mathbb{E}[Y|W=1,D=0] \approx 0$$

If this is true and $SB > 0$, then $ATT + SB \approx 0$. This is consistent with

$$ATT = \mathbb{E}[Y(1)|W(1)=1] - \mathbb{E}[Y(0)|W(1)=1] < 0$$

Hence, a causal effect on earnings which disadvantages females relative to males – a contradiction to the initial claim.
The problem is that if males and females have differential access to occupations then conditioning on occupation messes up the comparison by altering the composition of the type of males and females within a profession – females who breach the glass-ceiling are on average better than males in the same position. Apples-pears comparison problem. If male-females are paid the same, then that is consistent with a gender wage gap because they should be paid more.

6 Mathematical Appendix

6.1 Informal Proof of the Law of Iterated Expectations

Theorem (Law of Iterated Expectations). If X is a random variable whose expected value $\mathbb{E}(X)$ is defined, and Y is any random variable on the same probability space, then

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|Y))$$

Proof. Suppose a joint probability density function is well defined and the expectations are integrable, we write for the general case

$$\begin{aligned}
\mathbb{E}[X] &= \int x\mathbb{P}(X = x)dx \\
\mathbb{E}[X|Y = y] &= \int x\mathbb{P}(X = x|Y = y)dx \\
\mathbb{E}(\mathbb{E}[X|Y])) &= \int \left(\int x\mathbb{P}(X = x|Y = y)dx \right) \mathbb{P}(Y = y)dy \\
&= \int \int x\mathbb{P}(X = x, Y = y)dxdy \\
&= \int x \left(\mathbb{P}(X = x, Y = y)dy \right) dx \\
&= \int x\mathbb{P}(X = x)dx \\
&= \mathbb{E}[X]
\end{aligned}$$

6.2 Some Derivations of Covariance and Correlation