MATH2901 Higher Theory of Statistics

Noah Vinod

School of Mathematics & Statistics
Faculty of Science

1 June 2020

Contents

1. Random Variables

1.1 Definition

Consider a situation in which we want to take measurements of some variable of interest across multiple subjects. For example, we may be interested in measuring the weight loss achieved by participants in a weight loss program. Inevitably, our measurements always vary from one subject to another due to factors that are beyond our control, or beyond our knowledge. For this reason, we treat the measurements as *random variables.*

Definition 1.1.1 (Discrete Random Variable)**.** For a discrete sample space Ω, a *random variable* X is a function defined on Ω with

$$\mathbb{P}(X = x) = \sum_{s:X=x} \mathbb{P}(\{s\})$$

being the probability that X takes the value x.

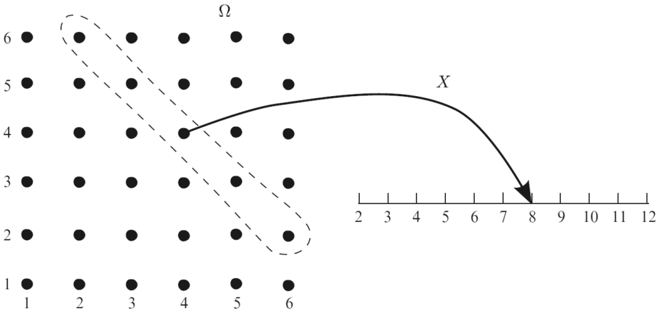
More generally, suppose we are given a sample space Ω with a collection of all subsets of Ω denoted by $\mathcal{F}$ (loosely $\mathcal{F}$ is a big bag containing all possible events) to which we can assign a probability specified by the set function $\mathbb{P}$. We call the triple $(\Omega, \mathcal{F}, \mathbb{P})$ a *probability space* and define a random variable in this space as follows:

Definition 1.1.2. A numerical *random variable* X is a function $X(\omega)$ that takes any $\omega \in \Omega$ as an input and maps it onto the real line $\mathbb{R}$ with the property that for all $x \in \mathbb{R}$

$$A(x) = \{\omega \in \Omega : X(\omega) \leq x\}$$

is an event (sometimes written as $A(x) \in \mathcal{F}$ to indicate that it is in the bag of all possible events to which we can assign a probability).

The definition is depicted pictorially below, where $\Omega = \{(1, 1), \ldots, (6, 6)\}$ is the sample space generated by two dice thrown consecutively and $X(\{(i, j)\}) = i + j$ is the random variable indicating the sum of the faces of the two dice.



Definition 1.1.3. Let A be an event. Define the *indicator function* of A as the function $I_A : \Omega \to \mathbb{R}$ such that for all $\omega \in \Omega$

$$I_A = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A. \end{cases}$$

Sometimes we may write $I\{A\}$ or $I_{\{A\}}$, depending on which happens to be the most economical, but unambiguous notation.

Example 1.1.1. Toss a fair coin 3 times.

$$\Omega = \{\text{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT}\}.$$

Let X denote the number of heads turned up. Then

$$\mathbb{P}(X=0) = \frac{1}{8}, \quad \mathbb{P}(X=1) = \frac{3}{8}, \quad \mathbb{P}(X=2) = \frac{3}{8}, \quad \mathbb{P}(X=3) = \frac{1}{8}.$$

1.2 Cumulative Distribution Function

Definition 1.2.1. The *cumulative distribution function* (cdf) of the random variable X is

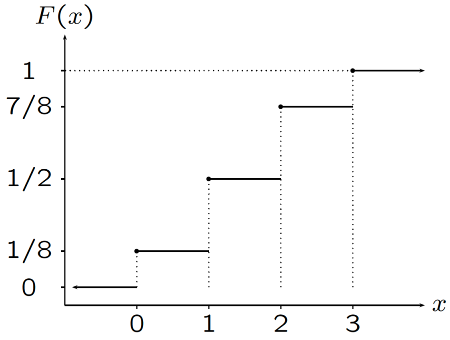
$$F_X(x) = \mathbb{P}(A(x)),$$

where $A(x) = \{\omega \in \Omega : X(\omega) \leq x\}$. Note that we can write this abbreviated as follows

$$F_X(x) = \mathbb{P}(X \leq x).$$

Note that we may write $F(x)$ if there is no ambiguity about whose cdf we refer to.

The following figure shows the cdf for Example 1.1.1.



Theorem 1.2.1 (Properties of the CDF)**.**

(i) F is bounded between 0 and 1, and such that

$$\lim_{x\to-\infty} F(x) = 0, \quad \lim_{x\to\infty} F(x) = 1.$$

(ii) F is non-decreasing: if $x < y$, then $F(x) \leq F(y)$.

(iii) For any pair of numbers $x < y$

$$\mathbb{P}(x < X \leq y) = F(y) - F(x).$$

(iv) F is right-continuous:

$$\lim_{n\to\infty} F\left(x+\frac{1}{n}\right) = F(x)$$

and

$$\mathbb{P}(X=x) = F(x) - \lim_{n\to\infty} F\left(x-\frac{1}{n}\right).$$

Proof.

(i) Define the event $A_n = \{\omega : X(\omega) \le -n\}$ and note that $A_1 \supseteq A_2 \supseteq \cdots$ is a sequence of decreasing events with limit the empty set:

$$\bigcap_{n=1}^{\infty} A_n = \varnothing.$$

It follows from the *continuity from above* property that

$$\lim_{n\to\infty} F(-n) = \lim_{n\to\infty} \mathbb{P}(X \le -n) = \lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\left(\bigcap_{n=1}^{\infty} A_n\right) = \mathbb{P}(\varnothing) = 0.$$

Similarly for the case $F(\infty) = 1$.

(ii) Since $\{X \le y\} = \{X \le x\} \cup \{x < X \le y\}$ and the events $\{X \le x\}$, $\{x < X \le y\}$ are disjoint, we have

$$\mathbb{P}(X \le y) = \mathbb{P}(X \le x) + \mathbb{P}(x < X \le y) \ge \mathbb{P}(X \le x).$$

(iii) Rearrange the last equation above to obtain the result:

$$\mathbb{P}(x < X \le y) = \mathbb{P}(X \le y) - \mathbb{P}(X \le x) = F(y) - F(x).$$

(iv) Set $B_n = \{x - 1/n < X \le x\}$ and note that $B_1 \supseteq B_2 \supseteq \cdots$ is a decreasing sequence of events with limit $B = \bigcap_{n=1}^{\infty} B_n$ equivalent to $\{X = x\}$. Hence, by the continuity from above property:

$$\lim_{n\to\infty} \mathbb{P}(B_n) = \mathbb{P}(B) = \mathbb{P}(X = x).$$

Now note that

$$\underbrace{\mathbb{P}(x - 1/n < X \le x)}_{\mathbb{P}(B_n)} = F(x) - F(x - 1/n)$$

and taking limits on both sides yields

$$\mathbb{P}(X = x) = F(x) - \lim_{n\to\infty} F(x - 1/n).$$

Now for the other case set $A_n = \{x < X \le x + 1/n\}$ and note that

$$\mathbb{P}(A_n) = \mathbb{P}(x < X \le x + 1/n) = F(x + 1/n) - F(x)$$

with $A_1 \supseteq A_2 \supseteq \cdots$ a decreasing sequence of events with limit $A_\infty \equiv \varnothing$. Hence,

$$\lim_{n\to\infty} F(x + 1/n) - F(x) = \lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(A_\infty) = 0.$$

□

1.3 Discrete Random Variables and Probability Functions

Definition 1.3.1. The random variable X is *discrete* if there are countably many values x for which $\mathbb{P}(X = x) > 0$.

Definition 1.3.2. The *probability function* of the discrete random variable X is the function f_X given by

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

Proposition 1.3.1. The probability function of a discrete random variable X has the following properties:

(i) $f_X(x) \geq 0$ for all $x \in \mathbb{R}$.

(ii) $\sum_{\text{all } x} f_X(x) = 1$.

When X is discrete, $F_X(x)$ is the sum of the probabilities for possible values of X less than or equal to x.

Example 1.3.1. A coin, with $p =$ probability of a head on a single toss, is tossed until a head turns up for the first time. Let X denote the number of tosses required. Find the probability function and the cumulative distribution function of X.

Solution. If the tosses are independent, then

$$\begin{aligned} f_X(x) &= \mathbb{P}(X = x) \\ &= \mathbb{P}(x - 1 \text{ tails then 1 head}) \\ &= (1 - p)^{x-1} \times p; \end{aligned}$$

that is, $f_X(x) = (1 - p)^{x-1}p$, $x = 1, 2, \ldots$; $0 < p < 1$.

$$\begin{aligned} F_X(x) &= \mathbb{P}(X \leq x) \\ &= \sum_{a \leq x} \mathbb{P}(X = a) \\ &= \sum_{a=1}^{x} (1 - p)^{a-1} p = 1 - (1 - p)^x, \quad x = 1, 2, \ldots. \end{aligned}$$

□

1.4 Continuous Random Variables and Density Functions

When a random variable has a continuum of possible values it is *continuous* (e.g. the lifetime of a light bulb has possible values in $[0, \infty)$).

Definition 1.4.1. The *density function* of a continuous random variable is a real-valued function f_X on $\mathbb{R}$ with the property

$$\int_A f_X(x) dx = \mathbb{P}(X \in A)$$

for any (measurable) set $A \subseteq \mathbb{R}$.

The density function (sometimes called the "probability density function") is the analogue of the probability function for continuous random variables.

Proposition 1.4.1. The density function of a continuous random variable X has the following properties:

(i) $f_X(x) \geq 0$ for all $x \in \mathbb{R}$

(ii) $\displaystyle\int_{-\infty}^{\infty} f_X(x)dx = 1$

Regardless of whether a random variable X is continuous or discrete, its cumulative distribution function is defined by

$$F_X(x) = \mathbb{P}(X \leq x).$$

Theorem 1.4.1. *The cumulative distribution function (cdf) F_X of a continuous random variable can be found from the density function f_X via*

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt.$$

Theorem 1.4.2. *The density function f_X of a continuous random variable can be found from the cumulative distribution function (cdf) F_X via*

$$f_X(x) = F'_X(x).$$

Theorem 1.4.3. *For any continuous random variable X and pair of numbers $a \leq b$*

$$\mathbb{P}(a \leq X \leq b) = \int_{a}^{b} f_X(x)dx = \text{ area under } f_X \text{ between } a \text{ and } b.$$

Continuous random variables X have the property

$$\mathbb{P}(X = a) = 0 \text{ for any } a \in \mathbb{R}.$$

It only 'makes sense' to talk about the probability of X lying in some subset of $\mathbb{R}$. A consequence of this is that, with continuous random variables, we don't have to worry about distinguishing between $<$ and $\leq$ signs. The probabilities are not affected. For example,

$$\mathbb{P}(2 < X < 3) = \mathbb{P}(2 \leq X < 3) = \mathbb{P}(2 < X \leq 3) = \mathbb{P}(2 \leq X \leq 3).$$

This is not the case for discrete random variables.

Definition 1.4.2. If F_X is strictly increasing in some interval, then F_X^{-1} is well defined and, for a specified $p \in (0, 1)$, the *p-th quantile* of F_X is x_p, where

$$F_X(x_p) = p \text{ or } x_p = F_X^{-1}(p).$$

$x_{0.5}$ is the *median* of F_X (or f_X). $x_{0.25}$ and $x_{0.75}$ are the lower and upper quartiles of F_X (or f_X).

Example 1.4.1. Let X be the random variable with cumulative distribution function $F_X(x) = 1-e^{-x}$, $x > 0$. Find the median and quartiles of X.

Solution. $F_X(x) = 1 - e^{-x}, x > 0$, then $y = F_X(x) \Rightarrow x = F_X^{-1}(y) = -\ln(1-y)$ and

$$x_p = -\ln(1-p)$$

so

$$x_{0.5} = -\ln(0.5) = \ln 2.$$

The lower quartile is $-\ln 0.75 = \ln 4/3$ and the upper quartile is $-\ln 0.25 = 2\ln 2$. □

1.5 Expectation and Moments

1.5.1 Expectation

The mean or average of the numbers $a_1, a_2, \ldots, a_n$ is

$$\frac{a_1 + \cdots + a_n}{n} = a_1 \cdot \frac{1}{n} + \cdots + a_n \cdot \frac{1}{n}.$$

Consider a random variable X with $\mathbb{P}(X = 5) = \frac{1}{5}$, $\mathbb{P}(X = 10) = \frac{4}{5}$. If we observed the values of, say, 100 random variables with the same distribution as X, we would expect to observe about 20 5's and about 80 10's so that the mean or average of the 100 numbers should be about

$$\frac{5 \times 20 + 10 \times 80}{100} = 5 \cdot \frac{1}{5} + 10 \cdot \frac{4}{5} = 9,$$

that is, the sum of the possible values of X weighted by their probabilities.

Definition 1.5.1. The *expected value* or *mean* of a discrete random variable X is

$$\mathbb{E}X = \mathbb{E}[X] = \sum_{\text{all } x} x \times \mathbb{P}(X = x) = \sum_{\text{all } x} x f_X(x),$$

where f_X is the probability function of X.

Definition 1.5.2. The *expected value* or *mean* of a continuous random variable X is

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

where f_X is the density function of X.

In both cases, $\mathbb{E}(X)$ has the interpretation of being the *long run average* of X - in the long run, as you observe an increasing number of values of X, the average of these values approach $\mathbb{E}(X)$.

Example 1.5.1. Let X be the number of females in a committee with three members. Assume that there is a 50:50 chance of each committee member being female, and that committee members are chosen independently of each other. Find $\mathbb{E}[X]$.

x	0	1	2	3
$f_X(x) = \mathbb{P}(X = x)$	1/8	3/8	3/8	1/8

$$\mathbb{E}[X] = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = \frac{3}{2}$$

Solution. The interpretation of $\frac{3}{2}$ is not that you expect X to be $\frac{3}{2}$ (it can only be 0, 1, 2 or 3) but that if you repeated the experiment, say, 100 times, then the average of the 100 numbers observed should be about $\frac{3}{2}$ $(= \frac{150}{100})$. That is, we expect to observe about 150 females in total in 100 committees. We don't expect to see exactly 1.5 females on each committee. □

Example 1.5.2. X has probability function

$$f_X(x) = \mathbb{P}(X = x) = (1-p)^{x-1} \cdot p, \quad x = 1, 2, \ldots; 0 < p < 1.$$

Find $\mathbb{E}(X)$.

Solution.

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{x=1}^{\infty} x(1-p)^{x-1} \cdot p \\
&= -p \sum_{x=1}^{\infty} \frac{d}{dp}(1-p)^x \\
&= -p \frac{d}{dp} \sum_{x=1}^{\infty} (1-p)^x \\
&= -p \frac{d}{dp} \left(\frac{1}{p} - 1 \right) \\
&= \frac{1}{p}.
\end{aligned}$$

□

Remark 1.5.1. If X is degenerate, that is, $X = c$ with probability 1 for some constant c, then X is in fact just a constant and $\mathbb{E}[X] = \sum_{\text{all } x} x\mathbb{P}(X = x) = c \cdot 1 = c$.

1.5.2 Expectation of Transformed Random Variables

Sometimes we are interested in a transformation of a random variable. For example, the circumference of a tree trunk is measured, but we want to know the cross-sectional area of the trunk. The variable of interest is $\pi(\frac{X}{2\pi})^2$. Transformations are also of interest when studying the properties of a random variable. For example, in order to understand X, it is often useful to look at the *r-th moment* of X about some constant a, defined as $\mathbb{E}[(X-a)^r]$.

Theorem 1.5.1. *The expected value of a function $g(X)$ of a random variable X is*

$$\mathbb{E}[g(X)] = \sum_{\text{all } x} g(x) f_X(x),$$

if X is discrete and

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

if X is continuous.

Proof. We prove this only in the discrete case. Set $Y = g(X)$ and $f_X(x) = \mathbb{P}(X = x)$, then

$$\begin{aligned}\mathbb{E}[g(X)] &= \mathbb{E}(Y) = \sum_y y\mathbb{P}(Y = y) \\ &= \sum_y y\mathbb{P}(g(X) = y) = \sum_y y\mathbb{P}(X \in \{x : g(x) = y\}) \\ &= \sum_y y \sum_{x:g(x)=y} \mathbb{P}(X = x) = \sum_y \sum_{x:g(x)=y} y\mathbb{P}(X = x) \\ &= \sum_y \sum_{x:g(x)=y} g(x)\mathbb{P}(X = x) = \sum_x g(x)\mathbb{P}(X = x)\end{aligned}$$

where the last line follows from the fact that if x takes on all values in its domain, then $y = g(x)$ takes on all values in its range and vice-versa. □

Note: In most situations,

$$\mathbb{E}[g(X)] \neq g(\mathbb{E}[X]).$$

1.5.3 Expectation of a Variable Under Changes of Scale

Often a change of scale is required, when studying a random variable. An example is when a change of measurement units is required (g → kg, ${}^0F \rightarrow^0 C$).

Theorem 1.5.2. *If a is a constant,*

$$\begin{aligned}\mathbb{E}[X + a] &= \sum_{\text{all } x} (x + a)\mathbb{P}(X = x) \\ &= \mathbb{E}[X] + a,\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}[aX] &= \sum_{\text{all } x} ax\mathbb{P}(X = x) \\ &= a\mathbb{E}(X).\end{aligned}$$

Similarly for X continuous. Also, $\mathbb{E}[g_1(X) + \cdots + g_n(X)] = \mathbb{E}[g_1(X)] + \cdots + \mathbb{E}[g_n(X)]$.

1.6 Standard Deviation and Variance

The standard deviation of a random variable is a measure of its *spread.* It is closely tied to the variance of a random variable, defined below:

Definition 1.6.1. If we let $\mu = \mathbb{E}(X)$, them the *variance* of X denoted by $\text{Var}(X)$ is defined as

$$\text{Var}(X) = \mathbb{E}[(X - \mu)^2]$$

(which is the second moment of X about μ).

Definition 1.6.2. The *standard deviation* of a random variable X is the square-root of its variance:

$$\text{standard deviation of } X = \sqrt{\text{Var}(X)}.$$

Proposition 1.6.1.

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - \mu^2 = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2.$$

Proof.

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}(X - \mu)^2 \\ &= \mathbb{E}(X^2 - 2\mu X + \mu^2) \\ &= \mathbb{E}(X^2) - 2\mu\mathbb{E}(X) + \mathbb{E}(\mu^2) \\ &= \mathbb{E}(X^2) - 2\mu^2 + \mu^2 \\ &= \mathbb{E}(X^2) - \mu^2.\end{aligned}$$

□

Proposition 1.6.2.

$$\begin{aligned}\mathrm{Var}(X + a) &= \mathrm{Var}(X) \\ \mathrm{Var}(aX) &= a^2\,\mathrm{Var}(X).\end{aligned}$$

1.7 Moment Generating Functions

Definition 1.7.1. The *moment generating function* (mgf) of a random variable X is

$$m_X(u) = \mathbb{E}(e^{uX}).$$

We say that the moment generating function of X exists if $m_X(u)$ is finite in some interval containing zero.

The name “moment generating function” comes from the following result concerning the r-th moment of X, that is $\mathbb{E}(X^r)$:

Theorem 1.7.1. *In general,* $\mathbb{E}(X^r) = m_X^{(r)}(0)$ *for* $r = 0, 1, 2, \ldots$.

Proof. Without going into details, the condition that the moment generating function is finite on some interval around zero guarantees the existence of all moments and that the derivative and the expectation can be interchanged and

$$\begin{aligned}m_X^{(r)}(u) &= \partial_u^{(r)}\mathbb{E}(e^{uX}) \\ &= \mathbb{E}(\partial_u^{(r)}e^{uX}) \\ &= \mathbb{E}(X^r e^{uX})\end{aligned}$$

given that the r-th derivative $m_X^{(r)}(u)$ is smooth enough around zero, the r-th moment can be computed by

$$\mathbb{E}(X^r) = \mathbb{E}(X^r e^{uX}|_{u=0}) = \mathbb{E}(X^r e^{uX})|_{u=0}.$$

□

Example 1.7.1. X has mgf

$$m_X(u) = \begin{cases}(1-u)^{-1} & u < 1, \\ \infty & u \geq 1.\end{cases}$$

Find an expression for the r-th moment of X.

Solution. Notice,

$$\begin{aligned} \mathbb{E}(X) &= m'_X(0) = (1-u)^{-2}|_{u=0} = 1, \\ \mathbb{E}(X^2) &= m_X^{(2)}(0) = 2(1-u)^{-3}|_{u=0} = 2 \qquad (\mathrm{Var}(X) = 2-1=1). \end{aligned}$$

So,

$$\begin{aligned} \mathbb{E}(X^r) &= m_X^{(r)}(0) \\ &= 1\cdot 2\cdot 3\cdots r(1-u)^{-r-1}|_{u=0} = r!. \end{aligned}$$

□

Example 1.7.2. X has probability function

$$\mathbb{P}(X=x) = e^{-\lambda}\frac{\lambda^x}{x!},\ x = 0,1,2,\dots;\lambda>0.$$

Find the mgf of X. Hence find the $\mathbb{E}(X)$ and $\mathrm{Var}(X)$.

Solution. X has mgf

$$\begin{aligned} m_X(u) = \mathbb{E}(e^{uX}) &= \sum_{x=0}^{\infty} e^{ux}\cdot\frac{e^{-\lambda}\lambda^x}{x!} \\ &= e^{-\lambda}\sum_{x=0}^{\infty}\frac{(\lambda e^u)^x}{x!} \\ &= e^{-\lambda}\cdot e^{\lambda e^u} \\ &= e^{\lambda(e^u-1)}. \end{aligned}$$

So,

$$\begin{aligned} \mathbb{E}(X) &= m'_X(0) \\ &= e^{\lambda(e^u-1)}\cdot\lambda e^u|_{u=0} \\ &= \lambda \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}(X^2) &= m_X^{(2)}(0) \\ &= \lambda(e^{\lambda(e^u-1)}\cdot e^u + e^{\lambda(e^u-1)}\cdot\lambda e^u\cdot e^u)|_{u=0} \\ &= \lambda(1+\lambda). \end{aligned}$$

Therefore, $\mathrm{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 = \lambda$. □

Example 1.7.3. X has probability function

$$f_X(x) = \mathbb{P}(X=x) = \binom{n}{x}p^x(1-p)^{n-x}, \quad x = 0,1,\dots,n; 0<p<1.$$

The Binomial Theorem states that

$$(a+b)^n = \sum_{x=0}^{n}\binom{n}{x}a^x b^{n-x}.$$

Use this result, when required, in order to:

(a) Show that $f_X(x)$ is a probability function

(b) Find the mgf of X

(c) Find $\mathbb{E}(X)$ and $\mathrm{Var}(X)$ using the mgf.

Solution.

(a)

$$\begin{aligned}\sum_{x=0}^{n} \mathbb{P}(X = x) &= \sum_{x=0}^{n} p^x (1-p)^{n-x} \\ &= (p + 1 - p)^n \\ &= 1.\end{aligned}$$

Hence $f_X(x)$ is a probability function.

(b)

$$\begin{aligned}m_X(u) &= \mathbb{E}(e^{uX}) \\ &= \sum_{x=0}^{n} e^{uX} \binom{n}{x} p^x (1-p)^{n-x} \\ &= (1 - p + pe^u)^n.\end{aligned}$$

(c)

$$\begin{aligned}\mathbb{E}(X) &= m'_X(0) \\ &= n(1 - p + pe^u)^{n-1} \cdot pe^u|_{u=0} \\ &= np.\end{aligned}$$

$$\begin{aligned}\mathbb{E}(X^2) &= m_X^{(2)}(0) \\ &= np[(1 - p + pe^u)^{n-1}u + (n-1)(1 - p + pe^u)^{n-2} \cdot pe^u \cdot e^u]|_{u=0} \\ &= np[1 + (n-1)p].\end{aligned}$$

Therefore, $\mathrm{Var}(X) = np(1-p)$.

□

Example 1.7.4. X has density

$$f_X(x) = e^{-x}, x > 0.$$

Find the moment generating function of X.

Solution.

$$\begin{aligned}m_X(u) &= \mathbb{E}(e^{uX}) \\ &= \int_0^\infty e^{ux} \cdot e^{-x} dx \\ &= \int_0^\infty e^{(u-1)x} dx \\ &= \begin{cases} (1-u)^{-1} & u < 1 \\ +\infty & u \geq 1. \end{cases}\end{aligned}$$

Then, $\mathbb{E}(X^r) = r!$, $r = 0, 1, 2, \ldots$. □

1.7.1 Properties of Moment Generating Functions

Theorem 1.7.2. *Let X and Y be two random variables all of whose moments exist. If*

$$m_X(u) = m_Y(u)$$

for all u in a neighbourhood of 0 (i.e. for all $|u| < \epsilon$ for some $\epsilon > 0$) then

$$F_X(x) = F_Y(x) \text{ for all } x \in \mathbb{R}.$$

(i.e. the mgf of a random variable is unique).

Theorem 1.7.3. *Let $\{X_n : n = 1, 2, \ldots\}$ be a sequence of random variables, each with moment generating functions $m_{X_n}(u)$. Furthermore, suppose that*

$$\lim_{n\to\infty} m_{X_n}(u) = m_X(u) \text{ for all } u \text{ in a neighbourhood of 0}$$

and $m_X(u)$ is a moment generating function of a random variable X. Then

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x) \text{ for all } x \in \mathbb{R}.$$

(i.e. convergence of mgf's implies convergence of cdf's).

1.8 Location and Scale Families of Densities

The unknown density function of a continuous random variable may belong to a family of density functions that all have similar form. Sometimes such a family has a form which identifies it as being generated by location or scale changes. The unknown location or scale constants are called *parameters*, and have a clear meaning.

Definition 1.8.1. Consider a random variable U with density function $f_U(x)$. A *location family* of densities based on the random variable U is the family of densities $f_X(x)$ where $X = U + c$ for all possible c. $f_X(x)$ is given by:

$$f_X(x) = f_U(x - c).$$

A *scale family* of densities based on the random variable U is the family of densities $f_X(x)$ where $X = cU$ for all possible c. $f_X(x)$ is given by:

$$f_X(x) = c^{-1} f_U(x/c).$$

Proof. Let $X = U + c$. Then the cdf of X is given by

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(U + c \leq x) = \mathbb{P}(U \leq x - c) = F_U(x - c).$$

Differentiating both sides we find that $f_X(x) = f_U(x - c)$.

Letting $X = cU$ we can use a similar approach to show that $f_X(x) = c^{-1} f_U(x/c)$. □

1.9 Bounding Probabilities

Theorem 1.9.1. *The inclusion-exclusion identity states that*

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_i \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \sum_{i<j<k} \mathbb{P}(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n+1}\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n).$$

Proof. Define $I\{B\}$ to be the indicator function for the event $B = \bigcup_{i=1}^n A_i$. By De Morgan's laws we have that

$$\begin{aligned}
I\{B\} &= 1 - I\{\overline{B}\} \\
&= 1 - I\left\{\overline{\bigcup_{i=1}^{n} A_i}\right\} \\
&= 1 - I\left\{\bigcap_{i=1}^{n} \overline{A}_i\right\} && \text{(by De Morgan's Law)} \\
&= 1 - \prod_{i=1}^{n} I\{\overline{A}_i\} && \text{(since } I\{A \cap B\} = I\{A\}I\{B\}\text{)} \\
&= 1 - \prod_{i=1}^{n} (1 - I\{A_i\}).
\end{aligned}$$

Finally, the inclusion-exclusion identity follows by applying the expectation operator $\mathbb{E}$ on both sides of the expansion:

$$\begin{aligned}
I\{B\} &= 1 - \prod_{i=1}^{n} (1 - I\{A_i\}) \\
&= \sum_i I\{A_i\} - \sum_{i<j} I\{A_i\}I\{A_j\} + \cdots + (-1)^{n+1} \prod_i I\{A_i\} \\
&= \sum_i I\{A_i\} - \sum_{i<j} I\{A_i \cap A_j\} + \cdots + (-1)^{n+1} I\left\{\bigcup_{i=1} A_i\right\}.
\end{aligned}$$

Without proof we now state what is commonly known as *Boole's inequalities*:

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) &\le \sum_i \mathbb{P}(A_i) \\
&\le \sum_i \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \sum_{i<j<k} \mathbb{P}(A_i \cap A_j \cap A_k) \\
&\vdots \\
\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) &\ge \sum_i \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) \\
&\ge \sum_i \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \sum_{i<j<k} \mathbb{P}(A_i, A_j, A_k) - \sum_{i<j<k<m} \mathbb{P}(A_i, A_j, A_k, A_m) \\
&\vdots
\end{aligned}$$

Thus, the inclusion-exclusion formula is a kind of 'Taylor' expansion for probabilities and can be useful for obtaining upper and lower bounds on the unknown probability. □

1.10 Chebychev's Inequality

Chebychev's Inequality is a fundamental result concerning tail probabilities of general random variables.

Theorem 1.10.1 (Chebychev's Inequality). *If X is any random variable with $\mathbb{E}(X) = \mu$, $\mathrm{Var}(X) = \sigma^2$ then*

$$\mathbb{P}(|X - \mu| > k\sigma) \leq \frac{1}{k^2}.$$

Proof. For the continuous case:

$$\begin{aligned}\sigma^2 &= \mathrm{Var}(X) \\ &= \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x)dx \\ &\geq \int_{|x-\mu|>k\sigma} (x-\mu)^2 f_X(x)dx \\ &\geq \int_{|x-\mu|>k\sigma} (k\sigma)^2 f_X(x)dx,\end{aligned}$$

since $|x-\mu| > k\sigma \Rightarrow (x-\mu)^2 f_X(x) > (k\sigma)^2 f_X(x)$. Therefore,

$$\begin{aligned}\sigma^2 &\geq k^2\sigma^2 \int_{|x-\mu|>k\sigma} f_X(x)dx \\ &= k^2\sigma^2 \mathbb{P}(|X-\mu| > k\sigma).\end{aligned}$$

Hence,

$$\mathbb{P}(|X - \mu| > k\sigma) \leq \frac{1}{k^2}.$$

□

The probability statement in Chebychev's Inequality is often stated verbally as

the probability that X is more than k standard deviations from its mean.

Note that Chebychev's Inequality makes no assumptions about the distribution of X.

Example 1.10.1. The number of items a factory produces in 1 day has mean 500 and variance 100. What is a lower bound for the probability that between 400 and 600 items will be produced tomorrow?

Solution. Let X denote the number of items produced tomorrow.

$$\mu = 500, \sigma = 10. \text{ Put } k = 10.$$

$$\mathbb{P}(|X - 500| > 10 \cdot 10) = \mathbb{P}(X < 400) + \mathbb{P}(X > 600) \leq \frac{1}{10^2}$$

$$\text{or } \mathbb{P}(400 \leq X \leq 600) \geq 1 - \frac{1}{100} = 0.99.$$

□

Theorem 1.10.2 (Markov's Inequality). *For any non-negative random variable X and $a > 0$,*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}.$$

1.11 Jensen's Inequality

Theorem 1.11.1. *If $h(x)$ is a convex function and X is a random variable, then*

$$\mathbb{E}[h(X)] \geq h(\mathbb{E}[X]).$$

Proof. Since h is convex we have that for some constants a and b:

$$h(x) \geq ax + b$$

with $h(\mu) = a\mu + b$ for $x = \mu$ (line may be tangent to h at this point). Now put $\mu = \mathbb{E}[X]$, $x = X$ and eliminate $b = h(\mu) - a\mu$ to get

$$h(X) \geq a(X - \mu) + h(\mu).$$

Hence,

$$\mathbb{E}[h(X)] \geq h(\mu) + a\mathbb{E}[X - \mu] = h(\mu).$$

□

1.12 Describing a Variable Using Data

The first two things to think about in data analysis are:

1. What is the research question? Descriptive statistics should primarily focus on providing insight into this question.

2. What are the properties of the variables of primary interest?

The most important property to think about when constructing descriptive statistics is whether each variable is *categorical* or *quantitative*.

Definition 1.12.1. A variable is *categorical* if its responses can be sorted into a finite set of (unordered) categories e.g. gender. A variable is *quantitative* if its responses are measured on some sort of scale e.g. height.

Summary of descriptive methods

Useful descriptive methods for when we wish to summarise one variable, or the association between two variables, depend on whether these variables are categorical or quantitative.

Does the research question involve:

	One variable		Two variables		
Data type:	Categorical	Quantitative	Both categorical	One of each	Both quantative
Numerics:	Table of frequencies	Mean/sd Median/quantiles	Two-way table	Mean/sd per group	Correlation
Graphs:	Bar chart	Dotplot Boxplot Histogram	Clustered bar chart	Scatterplot Boxplots Histograms etc.	Scatterplot

1.12.1 Categorical Data

Numerical Summaries of Categorical Data

The main tool for summarising categorical data is a table of frequencies (or percentages).

A *table of frequencies* consists of the counts of how many subjects fall into each level of a categorical variable.
A *two-way table* (of frequencies) counts how many subjects fall into each combination of levels from a pair of categorical variables.

Example 1.12.1. We can summarise the NSW election poll as follows:

Party	Liberal	Labour
Frequency	237	128

Example 1.12.2. Consider the question of whether there is an association between gender and whether or not a passenger on the Titanic survived. We can summarise the results from passenger records as follows:

		Outcome	
		Survived	Died
Gender	Male	142	709
	Female	308	154

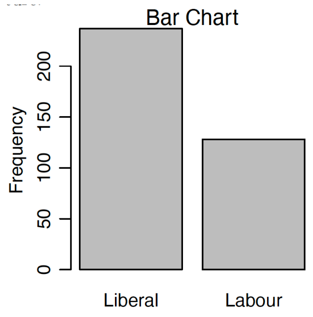
which suggests that a much higher proportion of females survived: their survival rate was 67% versus 17%.

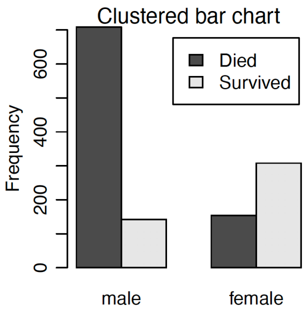
In the Titanic example, an alternative summary was the percentage survival for each gender. When one of the variables of interest has only two possible outcomes a list (or table) of percentages is useful.

If you are interested in an association between more than two categorical variables you can extend the above ideas, e.g. construct a three-way table . . .

Graphical Summaries of Categorical Data

A *bar chart* is a graph of table frequencies. A *clustered bar chart* graphs a two-way table, spacing the "bars" out as clusters to indicate the two-variable structure:





Pie charts are often used to graph categorical variables, however these are not generally recommended. It has been shown that readers of pie charts find it more difficult to understand the information that is contained in them e.g. comparing the relative size of frequencies across categories.

1.12.2 Quantitative Data

When summarising a quantitative variable, we are usually interested in three things:

- *Location* or "centre": A value around which most of the data lie
- *Spread*: How variable the values are around their centre

- *Shape*: Other information about a variable apart from location and spread. Skewness is an important example.

Numerical Summaries of Quantitative Data

The most commonly used summaries:

Definition 1.12.2. The *sample mean*

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

is a natural measure of location of a quantitative variable.

The *sample variance*

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

is a common measure of spread.

The *sample standard deviation* is defined as $s = \sqrt{s^2}$.

If we order the n values in the dataset and write them in increasing order as $\{x_{(1)}, x_{(2)}, \dots, x_{(n)}\}$. For example, $x_{(3)}$ is the third smallest observation in the dataset.

Definition 1.12.3. The *sample median* is

$$\tilde{x}_{0.5} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n+2}{2}\right)}\right) & \text{if } n \text{ is even.} \end{cases}$$

More generally, the *p-th sample quantile* of the data x is

$$\tilde{x}_p = x_{(k)} \text{ where } p = \frac{k-0.5}{n}$$

for $k \in \{1, 2, \dots, n\}$. We can estimate the sample quantile for other values of p by linear interpolation.

Example 1.12.3. The following (ordered) dataset is the number of mistakes made when ten subjects are each asked to do a repetitive task 500 times:

$$2 \quad 4 \quad 5 \quad 7 \quad 8 \quad 10 \quad 14 \quad 17 \quad 27 \quad 35.$$

Find the 5th and 15th sample percentiles of the data. Hence find the 10th percentile.

Solution. There are ten observations in the dataset, so the 5th sample percentile is

$$\tilde{x}_{(1-0.5)/10} = \tilde{x}_{0.05} = 2.$$

Similarly, the 15th sample percentile is 4. The 10th sample percentile is the average of these two. So $\tilde{x}_{0.1}$ can be estimated as

$$\tilde{x}_{0.1} = \frac{1}{2}(x_{(1)} + x_{(2)}) = \frac{1}{2}(2+4) = 3.$$

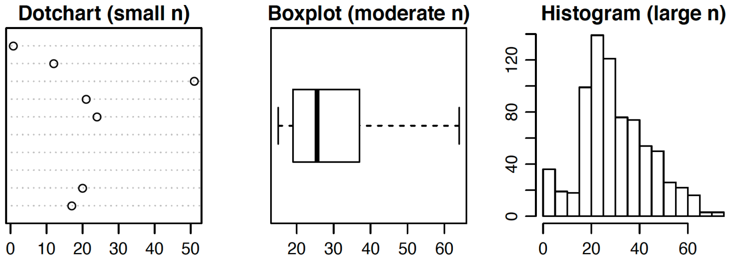
□

Apart from $\tilde{x}_{0.5}$, the two important quantiles are the *first and third quartiles*, $\tilde{x}_{0.25}$ and $\tilde{x}_{0.75}$ respectively. These terms are used to define the *interquartile range*

$$IQR = \tilde{x}_{0.75} - \tilde{x}_{0.25}$$

which is sometimes suggested as an alternative measure of spread to the sample standard deviation, because it is much less sensitive to outliers.

Graphical Summaries of Quantitative Data



A *dotchart* is a plot of each variable (x-axis) against its observation number, with data labels (if available). This is useful for small samples (e.g. $n < 20$).

A *boxplot* concisely describes location, spread and shape via the median, quartiles and extremes:

- The line in the middle of the box is the median, the measure of center
- The box is bounded by the upper and lower quartiles, so box width is a measure of spread (IQR)
- The whiskers extend till the most extreme value within one and a half interquartile ranges (1.5IQR) of the nearest quartile
- Any value farther than 1.5IQR from its nearest quartile is classified as an extreme value (or "outlier"), and labelled as a dot or open circle

Box plots are most useful for moderate-sized samples (e.g. $10 < n < 50$).

A *histogram* is a plot of the frequencies or relative frequencies of values within different intervals or *bins* that cover the range of all observed values in the sample. Note that this involves breaking the data up into smaller subsamples, and as such it will only find meaningful structure if the sample is large enough (e.g. $n > 30$) for the subsamples to contain non-trivial counts.

An issue in histogram construction is choice of number of bins. A useful rough rule is to use

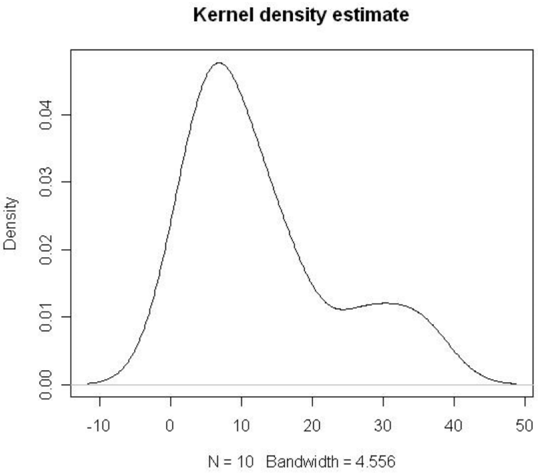
$$\text{number of bins} = \sqrt{n}.$$

A histogram is a step-wise rather than smooth function. A quantitative variable that is continuous might be better summarised by a smooth function. So an alternative estimator that often has better properties for continuous data is a *kernel data estimator*:

$$\widehat{f_h}(x) = \frac{1}{n}\sum_{i=1}^{n} w_h(x - x_i)$$

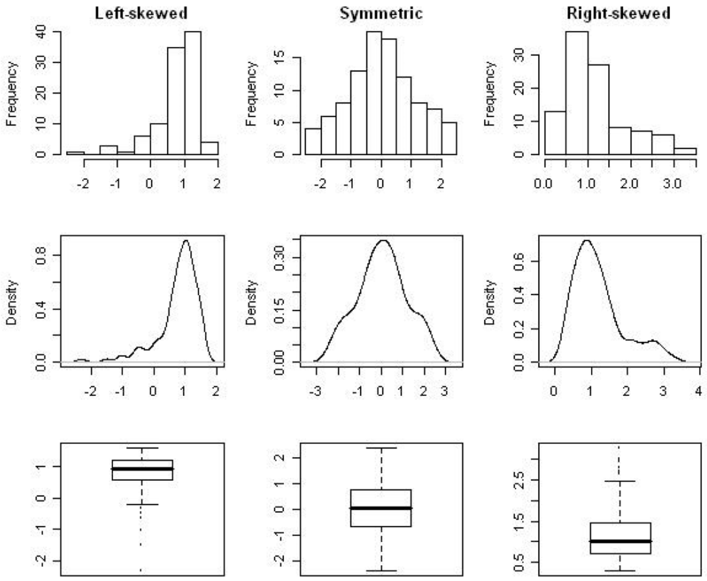
for some choice of weighting function $w_h(x)$ which includes a "bandwidth parameter" h.

Usually, $w(x)$ is chosen to be the normal density with mean 0 and standard deviation h. A lot of research has studied the issue of how to choose a bandwidth h, and most statistics packages are now able to automatically choose an estimate of h that usually performs well. The larger h is, the larger the bandwidth that is used i.e. the larger the range of observed values x_i that influence estimation of $\widehat{f}_h(x)$ at any given point x.



Shape of a Distribution

Something we can see from a graph that is hard to see from numerical summaries is the *shape* of a distribution. Shape properties, broadly, are characteristics of the distribution apart from location and spread. An example of an important shape property is *skew* - if the data tend to be asymmetric about its centre, it is skewed. We say data are "left-skewed" if the left tail is longer than the right, conversely, data are right-skewed if the right-tail is longer.



There are some numerical measures of shape, e.g. the coefficient of skewness κ_1:

$$\widehat{\kappa}_1 = \frac{1}{(n-1)s^3} \sum_{i=1}^{n} (x_i - \overline{x})^3$$

but they are rarely used - perhaps because of extreme sensitivity to outliers, and perhaps because shape properties can be easily visualised as above.

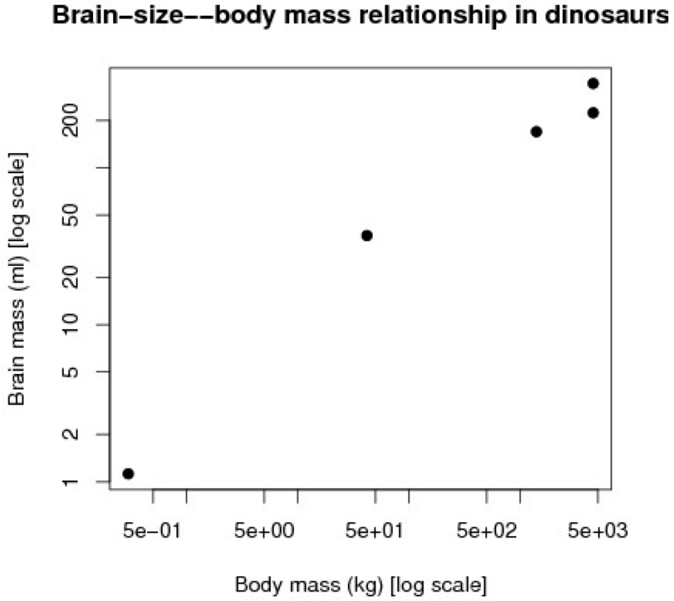
Another important thing to look for in graphs is *outliers* - unusual observations that might carry large weight in analysis. Such values need to be investigated - are they errors, are they "special cases" that offer interesting insights, how dependent are results on these outliers.

1.12.3 Summarising Associations Between Variables

Associations Between Quantitative Variables

Consider a pair of samples from two quantitative variables, $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$. We would like to understand how the x and y variables are related. An effective graphical display of the relationship between two quantitative variables is a *scatterplot* - a plot of the y_i against the x_i.

Example 1.12.4. How did brain mass change as a function of body size in dinosaurs?



An effective numerical summary of the relationship between two quantitative variables is the correlation coefficient (r):

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right)$$

where $\overline{x}$ and s_x are the sample mean and standard deviation of x, similarly for y.

Proposition 1.12.1.

(i) $|r| \leq 1$

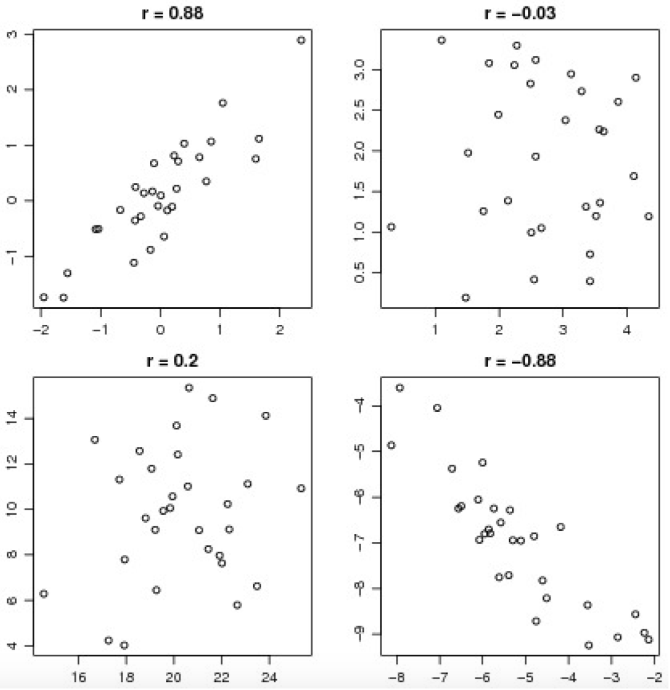
(ii) $r = -1$ if and only if $y_i = a + bx_i$ for each i, for some constants a, b such that $b < 0$

(iii) $r = 1$ if and only if $y_i = a + bx_i$ for each i, for some constants a, b such that $b > 0$

These results imply that r measures the strength and direction of associations between x and y:

- Strength of (linear) association - values closer to 1 or -1 suggest that the relationship is closer to a straight line
- Direction of association - values less than one suggest a decreasing relationship, values greater than one suggest an increasing relationship

For example,

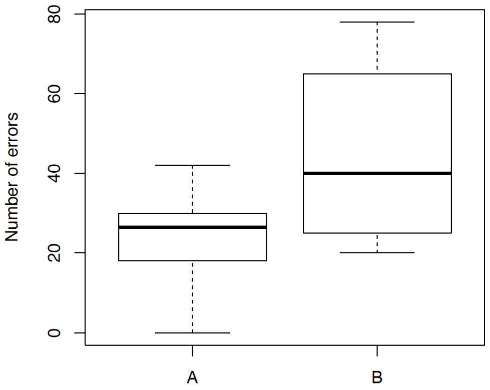


1.12.4 Associations Between Categorical and Quantitative Variables

When studying whether categorical and quantitative variables are associated, an effective strategy is to summarise the quantitative variable(s) separately for each level of the categorical variable(s).

To summarise number of errors, we might typically use mean/sd and a boxplot. For different categories, we calculate mean/sd of the number of errors for each of them and construct a boxplot for each category:

	$\bar{x}$	s
Sample A	23.4	12.3
Sample B	44.3	21.5



Note that in the above figure, the boxplots are presented on a common axis - sometimes this is referred to as *comparative boxplots* or “side-by-side boxplots”. An advantage of boxplots over histograms is that they can be quite narrow and hence readily compared across many samples by stacking them side-by-side.

2. Common Distributions

2.1 Bernoulli Distribution

Definition 2.1.1. A *Bernoulli trial* is an experiment with 2 possible outcomes. The outcomes are often labelled 'success' and 'failure'.

A simple example of a Bernoulli trial is a coin toss - 'heads' (success) and 'tails' (failure).

Definition 2.1.2. For a Bernoulli trial define the random variable

$$X = \begin{cases} 1 & \text{if the trial results in success} \\ 0 & \text{otherwise.} \end{cases}$$

Then X is said to have a *Bernoulli distribution*.

Theorem 2.1.1. *If X is a Bernoulli random variable defined according to a Bernoulli trial with success probability $0 < p < 1$ then the probability function of X is*

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

An equivalent way of writing this is $f_X(x) = p^x(1-p)^{1-x}$, $x = 0, 1$.

Definition 2.1.3. A constant like p above in a probability function or density is called a *parameter.*

2.2 Binomial Distribution

The *Binomial distribution* arises when several Bernoulli trials are repeated in succession.

Definition 2.2.1. Consider a sequence of n independent Bernoulli trials, each with success probability p. If

$$X = \text{total number of successes}$$

then X is a *Binomial* random variable with parameters n and p. A common shorthand is:

$$X \sim \text{Bin}(n, p).$$

Remark 2.2.1. The symbol "$\sim$" is commonly used in statistics for the phrase

"is distributed as" or "has distribution".

Whenever summing the number of times we observe a particular binary outcome, across n independent trials, we have a binomial distribution. For example, modelling the number of patients who survive a new type of surgery, out of 12 patients who each have 95% of surviving has distribution $\mathrm{Bin}(12, 0.95)$. In this example, we require the assumption of independence of responses across the n units in order to use the binomial distribution. This assumption is guaranteed to be satisfied if we randomly select units from some larger population.

Theorem 2.2.1. *If $X \sim \mathrm{Bin}(n, p)$ then its probability function is given by*

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, \ldots, n.$$

Proposition 2.2.1. If $X \sim \mathrm{Bin}(n, p)$ then

(i) $\mathbb{E}(X) = np$

(ii) $\mathrm{Var}(X) = np(1-p)$

(iii) $m_X(u) = (pe^u + 1 - p)^n$.

Corollary 2.2.1. *X has a Bernoulli distribution with parameter p if and only if*

$$X \sim \mathrm{Bin}(1, p).$$

2.3 Geometric Distribution

The *Geometric Distribution* arises when a Bernoulli trial is repeated until the first 'success'. In this case

$$X = \text{number of trials until first success}$$

and X is said to have a geometric distribution with parameter p, where p is the probability of success on each trial.

Theorem 2.3.1. *If X has a Geometric distribution with parameter $0 < p < 1$ then X has probability function*

$$f_X(x; p) = p(1-p)^{x-1}, \quad x = 1, 2, \ldots.$$

Proposition 2.3.1. If X has Geometric distribution with parameter p then

(i) $\mathbb{E}(X) = \frac{1}{p}$

(ii) $\mathrm{Var}(X) = \frac{1-p}{p^2}$.

Alternative definitions of the geometric distribution are possible. For example, a common definition is $X =$ number of failures before the first success. This leads to the distribution on $x = 0, 1, \ldots$ with a different mean than given above, but the variance is unchanged.

2.4 Hypergeometric Distribution

Hypergeometric random variables arise when counting the number of binary responses, when objects are sampled independently from finite populations, and the total number of "successes" in the population is known.

Suppose that a box contains N balls, m are red and $N-m$ are black, n balls are drawn at random. Let

$$X = \text{number of red balls drawn.}$$

Then X has a *Hypergeometric distribution* with parameters N, m and n. We write

$$X \sim \text{Hyp}(n, m, N).$$

Note that this can be thought of as a finite population version of the binomial distribution. Instead of assuming some constant probability p of "success" in the population, we say that there are N units in the population of which m are successes.

Theorem 2.4.1. *If X has a Hypergeometric distribution with parameters N, m and n then its probability function is given by*

$$f_X(x; N, m, n) = \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}} \quad 0 \leq x \leq \min(m, n).$$

Example 2.4.1. Lotto A machine contains 45 balls, and you select 6. Seven winning numbers are then drawn (6 main, one supplementary), and you win a major prize ($10000+) if you pick six of the winning numbers. What's the chance that you win a major prize from playing one game?

Solution. Let X be the number of winning numbers. X is hypergeometric with $N = 45$, $m = 6$, $n = 7$.

$$\begin{aligned}\mathbb{P}(X = x) &= f(x; 45, 6, 7)\\ &= \frac{\binom{6}{x}\binom{39}{7-x}}{\binom{45}{7}}.\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{P}(\text{win prize}) &= \mathbb{P}(X = 6)\\ &= \frac{\binom{6}{6}\binom{39}{1}}{\binom{45}{7}}\end{aligned}$$

which is less than 1 in a million. □

Example 2.4.2. The number of patients a town doctor sees who are in fact sick, when 800 people want to see a doctor, 500 of these are actually sick, and when the doctor only has time to see 32 of the 800 people (who are selected at random). If X is the number of sick patients, then

$$X \sim \text{Hyp}(32, 500, 800).$$

Proposition 2.4.1. If X has a Hypergeometric distribution with parameters N, m and n then

(i) $\mathbb{E}(X) = n \cdot \frac{m}{N}$

(ii) $\text{Var}(X) = n \cdot \frac{m}{N}(1 - \frac{m}{N})(\frac{N-n}{N-1})$.

It can be shown that as N gets large, a hypergeometric distribution with parameters N, m and n approaches $Y \sim \text{Bin}(n, \frac{m}{N})$. A suggestion of this can be seen in the above formulae: $\mathbb{E}(X)$ has the form of a binomial expectation with $p = \frac{m}{N}$, and $\text{Var}(X)$ only differs from the corresponding binomial variance formula by a "finite population correction factor" $\frac{N-n}{N-1}$ which tends to one as N gets large.

2.5 Poisson Distribution

Definition 2.5.1. The random variable X has a *Poisson distribution* with parameter $\lambda > 0$ if its probability function is

$$f_X(x;\lambda) = \mathbb{P}(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots.$$

A common abbreviation is

$$X \sim \text{Poisson}(\lambda)$$

.

The Poisson distribution often arises when the variable of interest is a count. For example, the number of traffic accidents in a city on any given day could be well described by a Poisson random variable. The Poisson is a standard distribution for the occurrence of rare events. Such events are often described by a Poisson process. A Poisson process is a model for the occurrence of point events in a continuum, usually a time-continuum. The occurrence or not of points in disjoint intervals is independent, with a uniform probability rate over time. If the probability rate is λ, then the number of points occurring in a time interval of length t is a random variable with a Poisson(λt) distribution.

Proposition 2.5.1. If $X \sim \text{Poisson}(\lambda)$ then

(i) $\mathbb{E}(X) = \lambda$

(ii) $\text{Var}(X) = \lambda$

(iii) $m_X(u) = e^{\lambda(e^u - 1)}$.

Example 2.5.1. Suggest a distribution that could be useful for studying X where X is the number of workplace accidents in a month (when the average number of accidents is 1.4).

Solution. X can be modelled as follows:

$$X \sim \text{Poisson}(1.4).$$

□

2.6 Exponential Distribution

The *Exponential distribution* is the simplest common distribution for describing the probability structure of *positive* random variables, such as lifetimes.

Definition 2.6.1. A random variable X is said to have an *exponential distribution* with parameter $\beta > 0$ if X has density function

$$f_X(x;\beta) = \frac{1}{\beta}e^{-x/\beta}, \quad x > 0.$$

Proposition 2.6.1. If X has an Exponential distribution with parameter β then

(i) $\mathbb{E}(X) = \beta$

(ii) $\text{Var}(X) = \beta^2$.

The exponential distribution is closely related to the Poisson distribution of the previous section. We know from previously that if a variable follows a Poisson process, counts of the number of times a particular event happens has a Poisson distribution with parameter λ. It can be shown that the *time until the next event* has exponential distribution with parameter $\beta = 1/\lambda$.

Example 2.6.1. If on average, 5 servers go offline during the day, what is the chance that no servers will go offline in the next hour? (Hint: Note that an hour is $\frac{1}{24}$ of a day).

Solution. Since an hour is $\frac{1}{24}$ of a day, on average $\lambda = \frac{5}{24}$ servers go offline per hour. If X is the time until the next server will go offline, then

$$X \sim \text{Exp}(\beta) = \text{Exp}\left(\frac{1}{\lambda}\right) = \text{Exp}\left(\frac{24}{5}\right).$$

Then

$$\mathbb{P}(X = 1 \text{ hour}) = \int_0^1 \lambda e^{-\lambda x} dx \approx 18.81\%.$$

□

An important property of the exponential distribution is *lack of memory*: if X has an exponential distribution, then

$$\mathbb{P}(X > s + t | X > s) = \mathbb{P}(X > t).$$

In other words, if the waiting time until the next event is exponential, then the waiting time until the next event is independent of the time you've already been waiting. Note, the exponential distribution is a special case of the Gamma distribution.

2.7 Uniform Distribution

The uniform distribution is the simplest common distribution for continuous random variables.

Definition 2.7.1. A continuous random variable X that can take values in the interval (a, b) with equal likelihood is said to have a *uniform distribution on* (a, b). A common shorthand is:

$$X \sim \text{Uniform}(a, b).$$

Definition 2.7.2. If $X \sim \text{Uniform}(a, b)$ then the density function of X is

$$f_X(x; a, b) = \frac{1}{b - a}, \quad a < x < b; \ a < b.$$

Note that $f_X(x; a, b)$ is simply a constant function over the interval (a, b), and zero otherwise.

Proposition 2.7.1. If $X \sim \text{Uniform}(a, b)$ then

(i) $\mathbb{E}(X) = \frac{(a+b)}{2}$

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

(iii) $m_X(u) = \frac{e^{bu} - e^{au}}{(b-a)u}$.

Note that there is also a discrete version of the uniform distribution, useful for modelling the outcome of an event that has k equally likely outcomes (such as the roll of a die). This has different formulae for its expectation and variance than the continuous case.

2.8 Special Functions Arising in Statistics

2.8.1 The Gamma Function

The *Gamma function* is essentially an extension of the factorial function (e.g. $4! = 24$) to general real numbers.

Definition 2.8.1. The *Gamma function* at $x \in \mathbb{R}$ is given by

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt.$$

Proposition 2.8.1. Some basic results for the gamma function are

(i) $\Gamma(x) = (x-1)\Gamma(x-1)$

(ii) $\Gamma(n) = (n-1)!$ for $n = 1, 2, \ldots$

(iii) $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.

Proposition 2.8.2. If m is a non-negative integer then

$$\int_0^\infty x^m e^{-x} dx = m!.$$

Example 2.8.1. Suppose $f_X(x) = \frac{1}{120} x^5 e^{-x}$ for $x > 0$. What is $\mathbb{E}(X)$?

Solution.

$$\mathbb{E}(X) = \int_{-\infty}^\infty x f_X(x) dx = \int_0^\infty \frac{1}{120} x^6 e^{-x}\, dx = \frac{6!}{120} = 6.$$

□

2.8.2 The Beta Function

Definition 2.8.2. The *Beta function* at $x, y \in \mathbb{R}$ is given by

$$B(x, y) = \int_0^1 t^{x-1} (1-t)^{y-1} dt.$$

Theorem 2.8.1. *For all* $x, y \in \mathbb{R}$,

$$B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}.$$

Example 2.8.2. Suppose $f_X(x) = 168x^2(1-x)^5$, for $0 < x < 1$. What is $\mathbb{E}(X^2)$?

Solution.

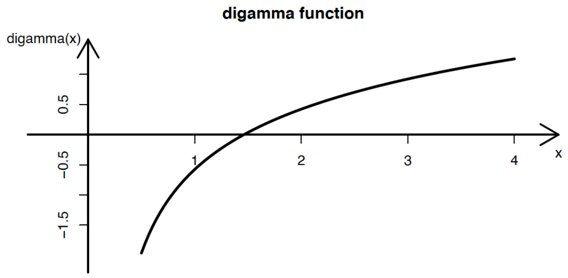
$$
\begin{aligned}
\mathbb{E}(X^2) &= \int_0^1 x^2 f_X(x)dx \\
&= \int_0^1 168x^4(1-x)^5dx \\
&= 168 \times B(5,6) \\
&= 168 \times \frac{\Gamma(5)\Gamma(6)}{\Gamma(11)} \\
&= 168 \times \frac{4!5!}{10!} \\
&= \frac{2}{15}.
\end{aligned}
$$

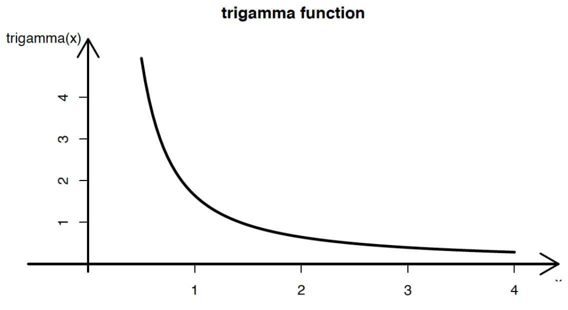
□

2.8.3 The Digamma and Trigamma Functions

Definition 2.8.3. For all $x \in \mathbb{R}$,

$$
\begin{aligned}
\text{digamma}(x) &= \frac{d}{dx}\ln(\Gamma(x)) \\
\text{trigamma}(x) &= \frac{d^2}{dx^2}\ln(\Gamma(x)).
\end{aligned}
$$





2.8.4 The Φ Function and Its Inverse

Definition 2.8.4. For all $x \in \mathbb{R}$,

$$
\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-t^2/2}dt.
$$

This function gives the cumulative distribution function of the standard normal distribution.

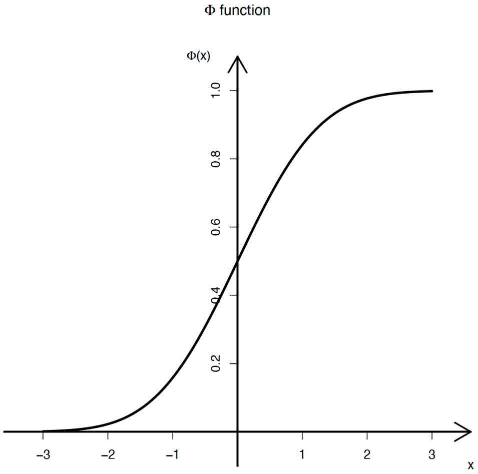
Proposition 2.8.3. The Φ function has the following properties:

(i) $\lim_{x \to -\infty} \Phi(x) = 0$

(ii) $\lim_{x \to \infty} \Phi(x) = 1$

(iii) $\Phi(0) = \frac{1}{2}$

(iv) Φ is monotonically increasing over $\mathbb{R}$.



It follows from the previous result that the inverse of Φ, $\Phi^{-1}(x)$ is well-defined for all $0 < x < 1$. Examples are

$$\Phi^{-1}\left(\frac{1}{2}\right) = 0 \text{ and } \Phi^{-1}(0.975) = 1.95996\ldots$$

2.9 Normal Distribution

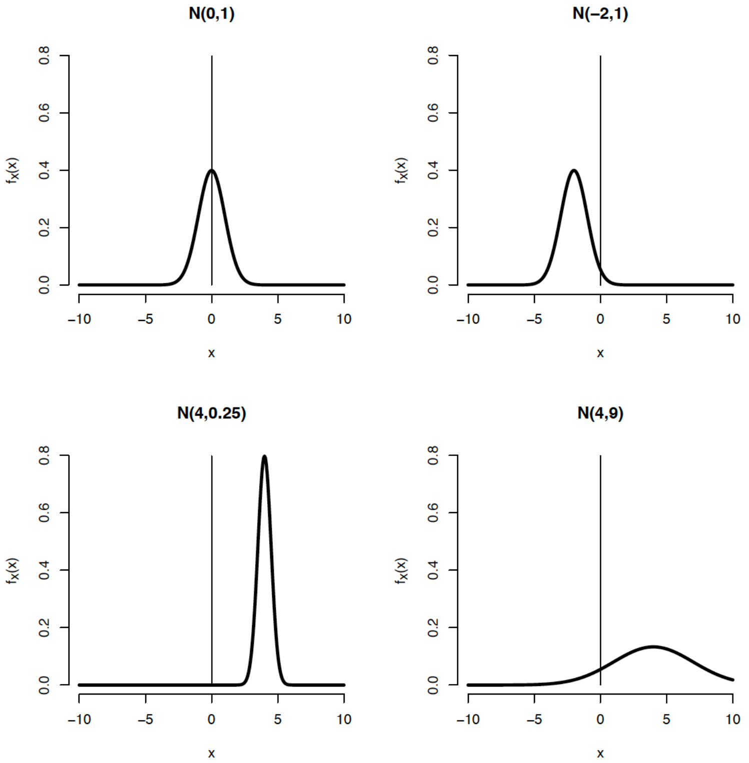
Definition 2.9.1. The random variable X is said to have a *normal distribution* with parameters μ and σ^2 (where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$) if X has density function

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

A common shorthand is

$$X \sim N(\mu, \sigma^2).$$

Normal density functions are symmetric "bell-shaped" curves symmetric about μ:


N(0,1)

N(−2,1)

N(4,0.25)

N(4,9)

The normal distribution is the most important distribution in statistical practice. One reason is that many real-life random variables are observed to have normal, or nearly normal distributions.

Proposition 2.9.1. If $X \sim N(\mu, \sigma^2)$ then

(i) $\mathbb{E}(X) = \mu$

(ii) $\text{Var}(X) = \sigma^2$

(iii) $m_X(u) = e^{\mu u + \frac{1}{2}\sigma^2 u^2}$.

The special case of $\mu = 0$ and $\sigma^2 = 1$ is known as the *standard normal distribution*. It is common to use the letter Z to denote standard normal random variables:

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

The standard normal density function is

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

2.9.1 Computing Normal Distribution Probabilities

Consider the problem:

$$\mathbb{P}(Z \leq 0.47) = \int_{-\infty}^{0.47} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

The standard normal density function does not have a closed form anti-derivative and cannot be solved in the usual way.

Corollary 2.9.1. *If $Z \sim N(0, 1)$ then*

$$\mathbb{P}(Z \leq x) = F_Z(x) = \Phi(x).$$

In other words, the Φ function is the cumulative distribution function of the $N(0, 1)$ random variable.

Probabilities concerning $Z \sim N(0,1)$ can be computed using tables for Φ. This can be used, for example, to show that:

$$\mathbb{P}(Z \leq 0.47) = \Phi(0.47) \approx 0.6808.$$

For finding a probability such as $\mathbb{P}(Z > 0.81)$, we need to work with the complement $\mathbb{P}(Z \leq 0.81)$:

$$\begin{aligned}\mathbb{P}(Z > 0.81) &= 1 - \mathbb{P}(Z \leq 0.81) \\ &= 1 - \Phi(0.81) \\ &\approx 0.2090.\end{aligned}$$

Theorem 2.9.1. *If $X \sim N(\mu, \sigma^2)$ then*

$$Z = \frac{X - \mu}{\sigma} \sim N(0,1).$$

The preceding result is an example of an operation called *standardisation.* The variable Z is said to have been *standardised.*

Example 2.9.1. Find $\mathbb{P}(X \leq 12)$ where $X \sim N(10, 9)$.

$$\begin{aligned}\mathbb{P}(X \leq 12) &= \mathbb{P}\left(\frac{X - 10}{3} \leq \frac{12 - 10}{3}\right) \\ &= \mathbb{P}(Z \leq 0.67) && \text{(where } Z \sim N(0,1)\text{)} \\ &= \Phi(0.67) \\ &\approx 0.7486.\end{aligned}$$

2.10 Gamma Distribution

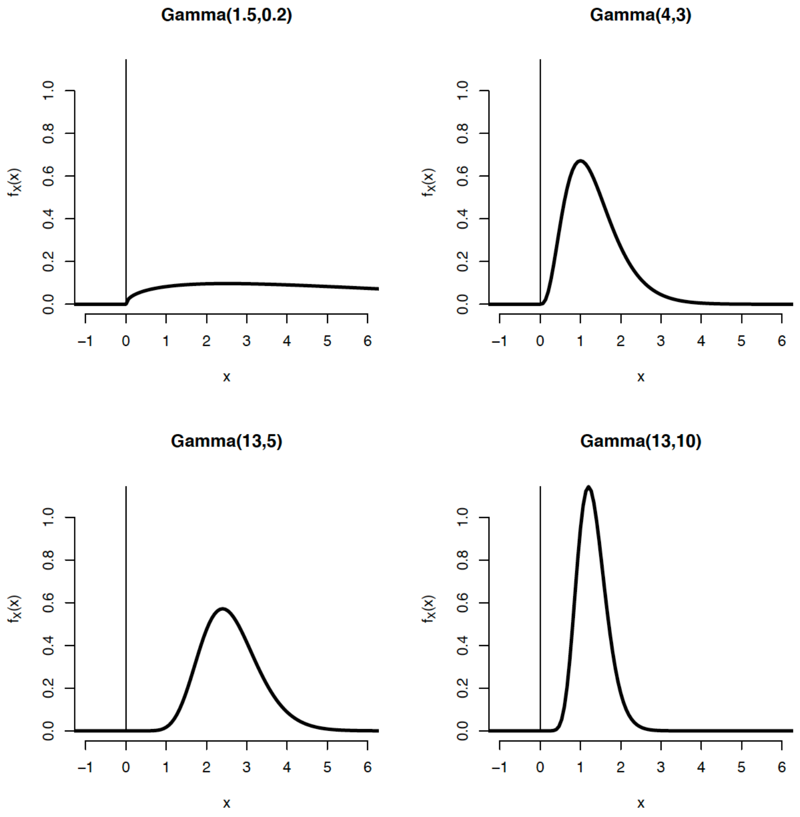
Definition 2.10.1. A random variable X is said to have a *Gamma distribution* with parameters α and β (where $\alpha, \beta > 0$) if X has density function:

$$f_X(x; \alpha, \beta) = \frac{e^{-x/\beta} x^{\alpha - 1}}{\Gamma(\alpha)\beta^\alpha}, \quad x > 0.$$

A common shorthand is:

$$X \sim \text{Gamma}(\alpha, \beta).$$

Gamma density functions are skewed curves on the positive half-line.



Gamma(1.5,0.2)

Gamma(4,3)

Gamma(13,5)

Gamma(13,10)

Proposition 2.10.1. If $X \sim \text{Gamma}(\alpha, \beta)$ then

(i) $\mathbb{E}(X) = \alpha\beta$

(ii) $\text{Var}(X) = \alpha\beta^2$

(iii) $m_X(u) = \left(\frac{1}{1-\beta u}\right)^{\alpha}$, $u < 1/\beta$.

Proposition 2.10.2. X has an exponential distribution if and only if

$$X \sim \text{Gamma}(1, \beta).$$

2.11 Beta Distribution

The *Beta distribution* generalises the Uniform$(0, 1)$ distribution, which can be thought of as a beta distribution with $a = b = 1$.

Definition 2.11.1. A random variable X is said to have a *Beta distribution* with parameters $\alpha, \beta > 0$ if its density function is

$$f_X(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}, \quad 0 < x < 1.$$

Proposition 2.11.1. If X has a Beta distribution with parameters α, β then

(i) $\mathbb{E}(X) = \frac{\alpha}{\alpha+\beta}$

(ii) $\text{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$.

2.12 Quantile-Quantile Plots of Data

Consider the situation in which we have a sample size n from some unknown random variable $\{x_1, x_2, \ldots, x_n\}$ and we want to check if these data appear to come from a random variable with cumulative distribution function $F_X(x)$. This can be achieved using a *quantile-quantile plot* (sometimes called a *Q-Q plot*).

Quantile-Quantile Plots

To check how well the sample $\{x_1, x_2, \ldots, x_n\}$ approximates the distribution with cdf $F_X(x)$, plot the n sample quantiles against the corresponding quantiles of $F_X(x)$. That is, plot the points

$$(F^{-1}(p), x_{(k)}) \text{ where } p = \frac{k-0.5}{n}, \text{ for all } k \in \{1, 2, \ldots, n\}.$$

If the data comes from the distribution $F_X(x)$, then the points will show no systematic departure from the one-to-one line.

According to the above definition, we need to know the exact cdf $F_X(x)$ to construct a quantile-quantile plot. However, for a location-scale family of distributions, a family that does not change its essential shape as its parameters change, we can construct the quantile-quantile plot using an arbitrary choice of parameters. In this case, we only need to check for systematic departures from a straight line rather than from the one-to-one line when assessing goodness-of-fit. This is the most common application of quantile-quantile plots. It allows us to see how well data approximates a whole family of location-scale distributions, without requiring any knowledge of what the values of the parameters are.

Example 2.12.1. Consider the example dataset:

$$2 \quad 4 \quad 5 \quad 7 \quad 8 \quad 10 \quad 14 \quad 17 \quad 27 \quad 35.$$

Use a quantile-quantile plot to assess how well these data approximate a normal distribution.

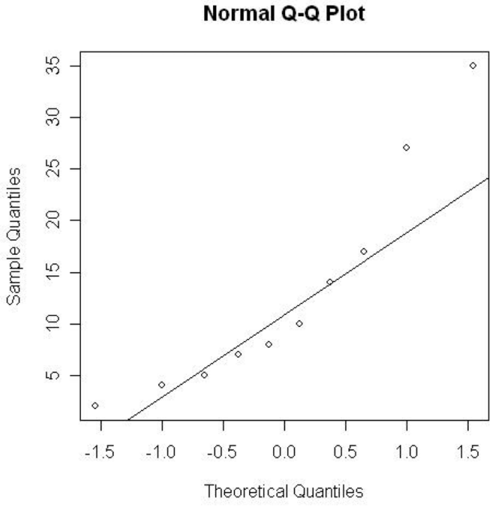
Solution. There are ten values in this dataset, so the values of p we will use to plot the data are $\frac{k-0.5}{10}$ for all $k \in \{1, 2, \ldots, 10\}$, that is, for all

$$p \in \{0.05, 0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75, 0.85, 0.95\}.$$

We want to find the quantiles corresponding to these values of p from a normal distribution, and compare those to the observed values. We have tables for the standard normal distribution, so we will use these to obtain quantiles. That is, we will plot the $x_{(k)}$ against $\Phi^{-1}(p)$ for the ten values of p displayed above. Using tables, we can show that the corresponding standard normal quantiles are

$$\{-1.64, -1.04, -0.67, -0.39, -0.13, 0.13, 0.39, 0.67, 1.04, 1.64\}$$

and so we plot these values against our ordered example dataset. This results in the following plot:



This plot does not follow a straight line - it has a systematic concave-up curve, so the data are clearly not normally distributed. In fact, because the curve is concave-up, the data are right-skewed (since the larger values in the dataset are much larger than expected for a normal distribution). □

3. Bivariate Distributions

Observations are often taken in pairs, leading to bivariate observations (X, Y) i.e. observations of two variables measured on the same subjects. For example, (height, weight) can be measured on people, as can (age, blood pressure), (gender, promotion) for employees, (sales, price) for supermarket products etc.

We are interested in exploring the nature of the relationship between two variables that have been measured on the same set of subjects.

3.1 Joint Probability Function and Density Function

Definition 3.1.1. If X and Y are discrete random variables then then *joint probability function* of X and Y is

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

the probability that $X = x$ and $Y = y$.

3.1.1 Studying Joint Probabilities

Recall that if two variables are dependent, then

$$\mathbb{P}(A \cap B) \neq \mathbb{P}(A)\mathbb{P}(B).$$

In the context of two discrete random variables X and Y,

$$\mathbb{P}(X = x, Y = y) \neq \mathbb{P}(X = x)\mathbb{P}(Y = y).$$

So when we want to calculate $\mathbb{P}(X = x, Y = y)$, or any joint probability involving both X and Y, we cannot find it using the probability functions of X and Y, which give us $\mathbb{P}(X = x)$ and $\mathbb{P}(Y = y)$. We instead need to know the joint probability function $f_{X,Y}(x, y)$.

Example 3.1.1. Suppose that X and Y have a joint probability function as tabulated below:

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

		y		
		-1	0	1
x	0	1/8	1/4	1/8
	1	1/8	1/16	1/16
	2	1/16	1/16	1/8

Find $\mathbb{P}(X = 0, Y = -1)$. Show that $\mathbb{P}(X = 0, Y = -1) \neq \mathbb{P}(X = 0)\mathbb{P}(Y = -1)$.

Solution. We have $\mathbb{P}(X = 0, Y = -1) = \frac{1}{8}$. Note that

$$\mathbb{P}(X = 0)\mathbb{P}(Y = -1) = \frac{1}{2} \times \frac{5}{16} = \frac{5}{32}.$$

This means that we wouldn't have been able to get the correct answer without looking at the *joint* probability distribution of X and Y. □

Example 3.1.2. Let X = number of successes in the first of two Bernoulli trials each with success probability p and let Y = total numbers of successes in the two trials. Then, for example,

$$f_{X,Y}(1,1) = \mathbb{P}(X=1, Y=1) = \mathbb{P}(X=1)\mathbb{P}(Y=1) = p(1-p).$$

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

		y		
		0	1	2
x	0	$(1-p)^2$	$p(1-p)$	0
	1	0	$p(1-p)$	p^2

3.1.2 Joint Density Functions

Definition 3.1.2. The *joint density function* of continuous random variables X and Y a bivariate function $f_{X,Y}$ with the property

$$\iint_A f_{X,Y}(x,y)dxdy = \mathbb{P}((X,Y) \in A)$$

any (measurable) subset A or $\mathbb{R}^2$.

For any two continuous random variables, X and Y, probabilities have the following geometrical interpretation: $f_{X,Y}$ is a surface over the plane $\mathbb{R}^2$ and probabilities over subsets $A \subseteq \mathbb{R}^2$ correspond to the *volume* under $f_{X,Y}$ over A.

Example 3.1.3. (X,Y) have joint density function

$$f(x,y) = \frac{12}{7}(x^2 + xy) \quad \text{for } x, y \in (0,1).$$

Find $\mathbb{P}(X < \frac{1}{2}, Y < \frac{2}{3})$.

Solution. We want to integrate $f_{X,Y}(x,y)$ over $(0, \frac{1}{2}) \times (0, \frac{2}{3})$:

$$\begin{aligned}
\mathbb{P}(X < 1/2, Y < 2/3) &= \int_0^{1/2} \int_0^{2/3} f_{X,Y}(x,y)dydx \\
&= \int_0^{1/2} \int_0^{2/3} \frac{12}{7}(x^2 + xy)dydx \\
&= \frac{12}{7} \int_0^{1/2} \left[x^2 y + \frac{xy^2}{2}\right]_0^{2/3} dx \\
&= \frac{12}{7} \int_0^{1/2} \frac{2}{3}x^2 + \frac{x}{2}\left(\frac{2}{3}\right)^2 dx \\
&= \frac{8}{7} \int_0^{1/2} x^2 + \frac{x}{3}dx \\
&= \frac{8}{7} \left[\frac{x^3}{3} + \frac{x^2}{6}\right]_0^{1/2} \\
&= \frac{8}{7} \left(\frac{1}{8 \cdot 3} + \frac{1}{4 \cdot 6}\right) \\
&= \frac{2}{21}.
\end{aligned}$$

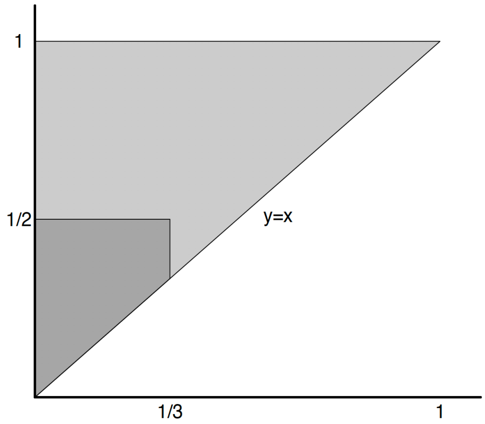
□

Example 3.1.4.

$$f_{X,Y}(x,y) = 2(x+y), \quad 0 < x < y,\ 0 < y < 1.$$

What is $\mathbb{P}(X < 1/3, Y < 1/2)$?

Solution. The entirety of the region has a triangular shape, and the area over which we need to integrate is trapezoidal, as shown below:



If we consider the shaded region as x-simple, then we have

$$\mathbb{P}(X < 1/3, Y < 1/2) = \int_0^{1/3}\int_0^y 2(x+y)dxdy + \int_{1/3}^{1/2}\int_0^{1/3} 2(x+y)dxdy = \frac{11}{108}.$$

If we consider it as y-simple, we have

$$\mathbb{P}(X < 1/3, Y < 1/2) = \int_0^{1/3}\int_x^{1/2} 2(x+y)dydx = \frac{11}{108}.$$

The latter is simpler to integrate. □

3.1.3 Other Results for $f_{X,Y}(x,y)$

Theorem 3.1.1. *If X and Y are discrete random variables then*

$$\sum_{\text{all } x}\sum_{\text{all } y} f_{X,Y}(x,y) = 1.$$

If X and Y are continuous random variables, then

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy = 1.$$

Definition 3.1.3. The *joint cdf* of X and Y is

$$\begin{aligned} F_{X,Y}(x,y) &= \mathbb{P}(X \le x, Y \le y) \\ &= \begin{cases} \displaystyle\sum_{u\le x}\sum_{v\le y} \mathbb{P}(X = u, Y = v) & (X \text{ discrete}) \\ \displaystyle\int_{-\infty}^{y}\int_{-\infty}^{x} f_{X,Y}(u,v)dudv & (X \text{ continuous}). \end{cases} \end{aligned}$$

Theorem 3.1.2. *If g is any function of X and Y,*

$$\mathbb{E}[g(X,Y)] = \begin{cases} \displaystyle\sum_{\text{all } x}\sum_{\text{all } y} g(x,y)\mathbb{P}(X=x,Y=y) & \textit{(discrete)} \\ \\ \displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f_{X,Y}(x,y)dxdy & \textit{(continuous).} \end{cases}$$

3.2 Marginal Probability/Density Functions

Theorem 3.2.1. *If X and Y are discrete, then $f_X(x)$ and $f_Y(y)$ can be calculated from $f_{X,Y}(x,y)$ as follows:*

$$f_X(x) = \sum_{\text{all } y} f_{X,Y}(x,y)$$
$$f_Y(y) = \sum_{\text{all } x} f_{X,Y}(x,y).$$

$f_X(x)$ is sometimes referred to as the marginal probability function of X.

Theorem 3.2.2. *If X and Y are continuous, then $f_X(x)$ and $f_Y(y)$ can be calculated from $f_{X,Y}(x,y)$ as follows:*

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

$f_X(x)$ is sometimes referred to as the marginal density function of X.

3.3 Conditional Probability and Density Functions

Definition 3.3.1. If X and Y are discrete, the *conditional probability function* of X given $Y=y$ is

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

Similarly,

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

Definition 3.3.2. If X and Y are continuous, the *conditional density function* of X given $Y=y$ is

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

Similarly,

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

Often we write $f_{Y|X}(y|x)$ as shorthand for $f_{Y|X}(y|X=x)$.

Let X and Y be continuous. For a given value of x, $f_{Y|X}(y|x)$ is an ordinary density function and has the usual properties such as:

Proposition 3.3.1. If X and Y are continuous then

$$\mathbb{P}(a \leq Y \leq b|X=x) = \int_a^b f_{Y|X}(y|x)dy.$$

Proposition 3.3.2. If X and Y are discrete then

$$\mathbb{P}(Y \in A|X=x) = \sum_{y\in A} f_{Y|X}(y|X=x).$$

3.4 Conditional Expected Value and Variance

Definition 3.4.1. The *conditional expected value* of X given $Y=y$ is

$$\mathbb{E}(X|Y=y) = \begin{cases} \displaystyle\sum_{\text{all } x} x\mathbb{P}(X=x|Y=y) & X \text{ discrete} \\ \displaystyle\int_{-\infty}^{\infty} xf_{X|Y}(x|y)dx & X \text{ continuous.} \end{cases}$$

Similarly,

$$\mathbb{E}(Y|X=x) = \begin{cases} \displaystyle\sum_{\text{all } y} y\mathbb{P}(Y=y|X=x) & Y \text{ discrete} \\ \displaystyle\int_{-\infty}^{\infty} yf_{Y|X}(y|x)dy & Y \text{ continuous.} \end{cases}$$

Definition 3.4.2. The *conditional variance* of X given $Y=y$ is

$$\text{Var}(X|Y=y) = \mathbb{E}(X^2|Y=y) - [\mathbb{E}(X|Y=y)]^2$$

where

$$\mathbb{E}(X^2|Y=y) = \begin{cases} \displaystyle\sum_{\text{all } x} x^2\mathbb{P}(X=x|Y=y) \\ \displaystyle\int_{-\infty}^{\infty} x^2 f_{X|Y}(x|y)dx. \end{cases}$$

Similarly for $\text{Var}(Y|X=x)$.

3.5 Independent Random Variables

Definition 3.5.1. Random variables X and Y are *independent* if and only if for all x, y

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

Corollary 3.5.1. *Random variables X and Y are independent if and only if for all x, y*

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

or

$$f_{X|Y}(x|y) = f_X(x).$$

If X and Y are independent, then the probability structure of Y is unaffected by the 'knowledge' that X takes on some value x (and vice versa).

Proposition 3.5.1. If X and Y are independent

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

Example 3.5.1.

		y			
		-1	0	1	$f_X(x)$
x	0	0.01	0.02	0.07	0.1
	1	0.04	0.13	0.33	0.5
	2	0.05	0.05	0.3	0.4
	$f_Y(y)$	0.1	0.2	0.7	1

Are X and Y independent?

Solution. X and Y are independent if

$$f_{X,Y}(x,y) = f_X(x) f_Y(y) \text{ for all } x, y;$$

that is, every entry in the body of the table equals the product of the corresponding row and column totals. X and Y are not independent if $f_{X,Y}(x,y) \neq f_X(x) f_Y(y)$ for at least one pair of values x and y. Thus X and Y are not independent in this case since, for example

$$0.04 = \mathbb{P}(X = 1, Y = -1) \neq \mathbb{P}(X = 1)\mathbb{P}(Y = -1) = (0.5)(0.1) = 0.05.$$

□

Example 3.5.2. X and Y have joint probability function

$$f_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y) = p^2(1-p)^{x+y}, \quad x = 0, 1, \ldots, \; y = 0, 1, \ldots; \; 0 < p < 1.$$

Are X and Y independent?

Solution. Now,

$$\begin{aligned} f_X(x) &= \mathbb{P}(X = x) \\ &= \sum_{\text{all } y} f_{X,Y}(x,y) \\ &= \sum_{y=0}^{\infty} p^2(1-p)^{x+y} \\ &= p(1-p)^x, \quad x = 0, 1, \ldots. \end{aligned}$$

Similarly,

$$\begin{aligned} f_Y(y) &= \sum_{x=0}^{\infty} p^2(1-p)^{x+y} \\ &= p(1-p)^y, \quad y = 0, 1, \ldots. \end{aligned}$$

Therefore, $f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)$ for all x, y and so X and Y are independent. □

Theorem 3.5.1. *If X and Y are independent,*

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

and more generally, for any functions $g(X)$ and $h(Y)$,

$$\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)].$$

Proof. For the continuous case,

$$\begin{aligned} \mathbb{E}(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_X(x) \cdot f_Y(y)dxdy \\ &= \int_{-\infty}^{\infty} xf_X(x)dx \cdot \int_{-\infty}^{\infty} yf_Y(y)dy \\ &= \mathbb{E}(X)\mathbb{E}(Y). \end{aligned}$$

□

3.6 Covariance and Correlation

3.6.1 Covariance

Definition 3.6.1. The *covariance* of X and Y is

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

where $\mu_X = \mathbb{E}(X)$ and $\mu_Y = \mathbb{E}(Y)$.

$\text{Cov}(X, Y)$ measures not only how X and Y vary about their means, but also how they vary together *linearly.* $\text{Cov}(X, Y) > 0$ if X and Y are positively associated, i.e. if X is likely to be large when Y is large and X is likely to be small when Y is small. If X and Y are negatively associated, $\text{Cov}(X, Y) < 0$.

Theorem 3.6.1.

(i) $\text{Cov}(X, X) = \text{Var}(X)$

(ii) $\text{Cov}(X, Y) = \mathbb{E}(XY) - \mu_X\mu_Y$.

Corollary 3.6.1. *If X and Y are independent then* $\text{Cov}(X, Y) = 0$.

Proposition 3.6.1. For arbitrary constants a, b

$$\text{Var}(aX + bY) = a^2 \text{Var}(X) + b^2 \text{Var}(Y) + 2ab \,\text{Cov}(X, Y).$$

Hence,

(i) $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)$

(ii) $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$, when X and Y are independent

(iii) $\mathrm{Var}(X-Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$, when X and Y are independent.

3.6.2 Correlation

Definition 3.6.2. The *correlation* between X and Y is

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

$\mathrm{Corr}(X,Y)$ measures the strength of the linear association between X and Y.

Definition 3.6.3. If $\mathrm{Corr}(X,Y) = 0$, then X and Y are said to be *uncorrelated.*

Independent random variables are uncorrelated, but uncorrelated variables are not necessarily independent; for example, if X has a distribution which is symmetric about zero and $Y = X^2$,

$$\mathbb{E}(XY) = \mathbb{E}(X^3) = 0 \text{ and } \mathbb{E}(X) = 0,$$

so $\mathrm{Cov}(X,Y) = 0$ and $\mathrm{Corr}(X,Y) = 0$, but since $Y = X^2$, X and Y are dependent.

Proposition 3.6.2.

(i) $|\,\mathrm{Corr}(X,Y)| \leq 1$

(ii) $\mathrm{Corr}(X,Y) = -1$ if and only if $\mathbb{P}(Y = a + bX) = 1$ for some constants a, b such that $b < 0$

(iii) $\mathrm{Corr}(X,Y) = 1$ if and only if $\mathbb{P}(Y = a + bX) = 1$ for some constants a, b such that $b > 0$

Proof.

1. Let $\rho = \mathrm{Corr}(X,Y)$, $\sigma_X^2 = \mathrm{Var}(X)$ and $\sigma_Y^2 = \mathrm{Var}(Y)$. Then

$$\begin{aligned} 0 &\leq \mathrm{Var}\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right) \\ &= 2 + 2\,\mathrm{Cov}\left(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right) \\ &= 2(1+\rho), \end{aligned}$$

and so $\rho \geq -1$. Also, $0 \leq \mathrm{Var}\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right) = 2(1-\rho)$, so $\rho \leq 1$.

2. If $\rho = -1$, $\mathrm{Var}\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right) = 2(1+\rho) = 0$. This means that $\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}$ is a constant, i.e. $\mathbb{P}\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y} = c\right) = 1$ for some constant. But

$$\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y} = c \Longleftrightarrow Y = \frac{-X\sigma_Y}{\sigma_X} + c\sigma_Y$$

so $\mathbb{P}(Y = a + bX) = 1$ for some constants $a = c\sigma_Y$ and $b = -\frac{\sigma_Y}{\sigma_X} < 0$.

3. Similarly, for $\rho = 1$, $\mathbb{P}(Y = a + bX) = 1$ for some constant a and $b = \frac{\sigma_Y}{\sigma_X} > 0$.

□

3.7 The Bivariate Normal Distribution

X and Y have the bivariate normal distribution if

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho)^2}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\},$$

for $-\infty < x < \infty$, $-\infty < y < \infty$; $-\infty < \mu_X < \infty$, $-\infty < \mu_Y < \infty$; $\sigma_X > 0$, $\sigma_Y > 0$, $-1 < \rho < 1$.

Theorem 3.7.1.

(i) $X \sim N(\mu_X, \sigma_X^2)$

(ii) $Y \sim N(\mu_Y, \sigma_Y^2)$.

Proposition 3.7.1.

$$\rho = \mathrm{Corr}(X, Y).$$

3.7.1 Visualisation of the Bivariate Normal Density Function

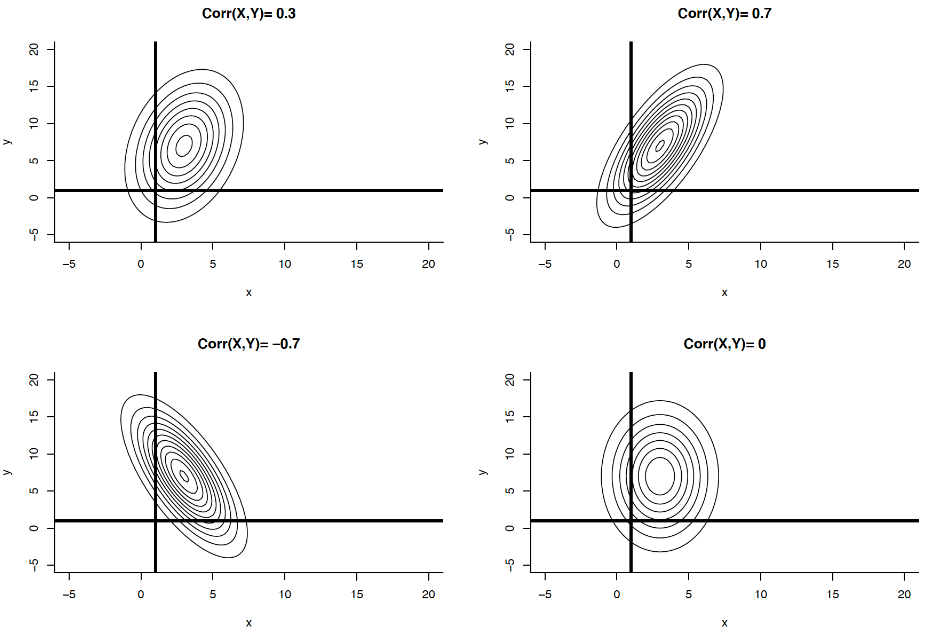
The bivariate normal density is a bivariate function $f_{X,Y}(x,y)$ with elliptical contours. The following figure provides contour plots of the bivariate normal density for

$$\mu_X = 3, \mu_Y = 7, \sigma_X = 2, \sigma_Y = 5$$

in all cases, but with

$$\rho = \mathrm{Corr}(X, Y)$$

taking four different values:



Theorem 3.7.2. *If X and Y are uncorrelated jointly normal variables, then X and Y are independent.*

3.8 Extension to n Random Variables

Definition 3.8.1. The *joint probability function* of $X_1, \ldots, X_n$ is

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n).$$

Definition 3.8.2. The *joint cdf* in both the discrete and continuous cases is

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n).$$

Definition 3.8.3. The *joint density* of $X_1, \ldots, X_n$ is

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \frac{\partial_n}{\partial x_1 \cdots \partial x_n} F_{X_1,\ldots,X_n}(x_1, \ldots, x_n).$$

Definition 3.8.4. $X_1, \ldots, X_n$ are *independent* if

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$$

or

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n).$$

4. Transformations

Definition 4.0.1. If X is a random variable, $Y = h(X)$ for some function h is a *transformation* of X.

Theorem 4.0.1. *For discrete* X*,*

$$f_Y(y) = \mathbb{P}(Y = y) = \mathbb{P}(h(X) = y) = \sum_{x:h(x)=y} \mathbb{P}(X = x).$$

Example 4.0.1.

x	-1	0	1	2
$f_X(x)$	$1/8$	$1/4$	$1/2$	$1/8$

Find $f_Y(y)$ where $Y = X^2$.

Solution.

y	0	1	4
$f_Y(y)$	$1/4$	$5/8$	$1/8$

since, for example $\mathbb{P}(Y = 0) = \mathbb{P}(X = 0) = 1/4$,

$$\mathbb{P}(Y = 1) = \mathbb{P}(X = -1) + \mathbb{P}(X = 1) = 5/8.$$

□

The density function of a transformed continuous variable is simple to determine when the transformation is *monotonic.*

Definition 4.0.2. Let h be a real-valued function defined over the set A where A is a subset of $\mathbb{R}$. Then h is a *monotonic* transformation if h is either strictly increasing or strictly decreasing over A.

Theorem 4.0.2. *For continuous* X*, if* h *is monotonic over the set* $\{x : f_X(x) > 0\}$ *then*

$$\begin{aligned} f_Y(y) &= f_X(x) \left| \frac{dx}{dy} \right| \\ &= f_X[h^{-1}(y)] \left| \frac{dx}{dy} \right| \end{aligned}$$

for y *such that* $f_X[h^{-1}(y)] > 0$.

Proof.

$$\begin{aligned}F_Y(y) &= \mathbb{P}(Y \leq y)\\ &= \mathbb{P}[h(X) \leq y]\\ &= \begin{cases}\mathbb{P}[X \leq h^{-1}(y)] = F_X[h^{-1}(y)] & \text{if } h \uparrow\\ \\ \mathbb{P}(X \geq h^{-1}(y)) = 1 - F_X[h^{-1}(y)] & \text{if } h \downarrow .\end{cases}\end{aligned}$$

Therefore,

$$f_Y(y) = \begin{cases} f_X[h^{-1}(y)]\frac{dh^{-1}(y)}{dy} = f_X(x)\frac{dx}{dy} & \text{if } h \uparrow\\ \\ -f_X[h^{-1}(y)]\frac{dh^{-1}(y)}{dy} = -f_X(x)\frac{dx}{dy} & \text{if } h \downarrow .\end{cases}$$

Now,

$$\frac{dy}{dx}\begin{cases}> 0 & \text{if } h \uparrow\\ < 0 & \text{if } h \downarrow,\end{cases}$$

and so

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|.$$

□

Example 4.0.2. $f_X(x) = 3x^2$, $0 < x < 1$. Find $f_Y(y)$ where $Y = 2X - 1$.

Solution. $x = \frac{y+1}{2}$, $\frac{dx}{dy} = \frac{1}{2}$ and

$$f_Y(y) = f_X(x)\left|\frac{1}{2}\right| = \frac{3}{8}(y+1)^2, \quad -1 < y < 1.$$

□

Example 4.0.3. Let $X \sim \text{Exp}(\beta)$, i.e. $f_X(x) = \frac{1}{\beta}e^{-x/\beta}$, $x > 0; \beta > 0$. Find $f_Y(y)$ where $Y = X/\beta$.

Solution. Let $Y = \lambda X$. Then $x = \frac{y}{\lambda}$, $\frac{dx}{dy} = \frac{1}{\lambda}$ and $f_Y(y) = f_X(x)|\frac{dx}{dy}| = \lambda e^{-y}|\frac{1}{\lambda}| = e^{-y}$, $y > 0$. □

Note that this result shows that any exponential variable can be transformed to the exponential distribution with parameter 1, by dividing by the parameter β. Hence the exponential distribution with parameter β is a *scale family* and β can be interpreted as a *scale parameter.*

Example 4.0.4. $f_X(x) = \sqrt{\frac{2}{\pi}}e^{-x^2/2}$, $x > 0$. Let $Y = \frac{X^2}{2}$. Find the density function of Y.

Solution. Then $x = \sqrt{2y}$, $\frac{dx}{dy} = (2y)^{-\frac{1}{2}}$ and

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right| = \sqrt{\frac{2}{\pi}}e^{-y}\left|(2y)^{-\frac{1}{2}}\right| = \frac{e^{-y}y^{-\frac{1}{2}}}{\sqrt{\pi}}, \quad y > 0.$$

Note, $\Gamma(\frac{1}{2}) = \int_0^\infty e^{-y}y^{-\frac{1}{2}}dy = \sqrt{\pi}$. □

4.1 Linear Transformations

The simplest monotonic transformations are linear transformations:

$$h(x) = ax + b \quad \text{for } a \neq 0.$$

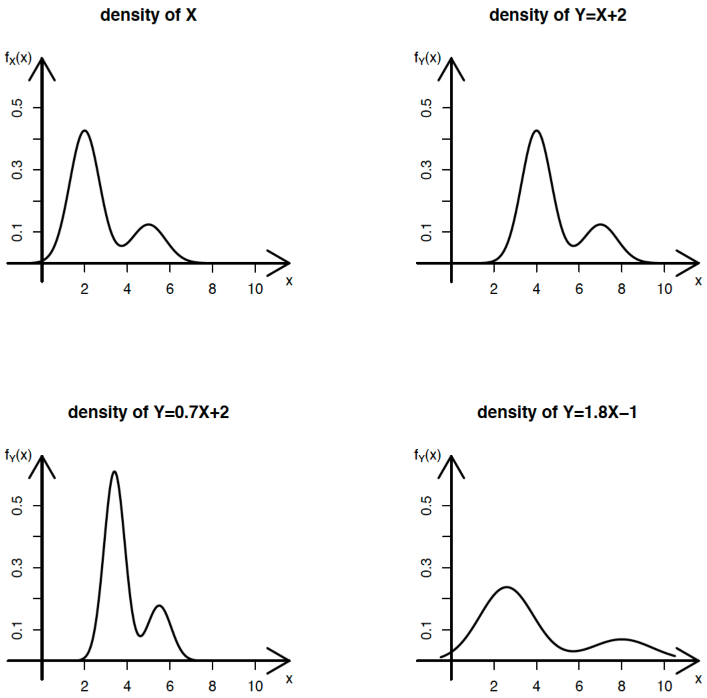
A common example of when a linear transformation is required is when applying a *change of scale* e.g. changing temperature measurements from degrees Fahrenheit to degrees Celsius, $Y = \frac{5}{9}(X - 32)$. Another example is when we are interested in calculating a summary statistic (such as the sample mean) which can be written as a linear transformation of a set of observed random variables.

Theorem 4.1.1. *For a continuous random variable X, if $Y = aX + b$ is a linear transformation of X with $a \neq 0$, then*

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$$

for all y such that $f_X\left(\frac{y-b}{a}\right) > 0$.

This result is the formula for densities of location and scale families. It follows directly from the result for general monotonic transformations. The implication is that linear transformations only change the *location* and *scale* of a density function. They do not change its shape. The following figure illustrates this:



4.2 Probability Integral Transformation

Theorem 4.2.1 (Probability Integral Transformation). *If X has density $f_X(x)$ and cdf $F_X(x)$, then $Y = F_X(X) \sim$ Uniform$(0,1)$.*

Proof. Let $Y = F_X(X)$. Then $y = F_X(x)$ and $\frac{dy}{dx} = f_X(x)$. Hence

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right| = f_X(x) \cdot \frac{1}{f_X(x)} = 1, \quad 0 < y < 1.$$

Alternatively, from first principles

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) \\ &= \mathbb{P}(F_X(x) \leq y) \\ &= \mathbb{P}[X \leq F_X^{-1}(y)] && \text{(since } F_X \uparrow\text{)} \\ &= F_X[F_X^{-1}(y)] \\ &= y. \end{aligned}$$

Therefore, $f_Y(y) = 1$, $0 < y < 1$. □

The Probability Integral Transformation allows for easy simulation of random variables from any distribution for which the inverse cdf F_X^{-1} is easily computed. A computer can be used to generate U where $U \sim \text{Uniform}(0,1)$. If we require an observation X where X has cdf F_X, then

$$U = F_X(X) \sim \text{Uniform}(0,1) \iff X = F_X^{-1}(U).$$

A Universal Random Number Generator

To generate a sample from any distribution X:

1. Use a computer to generate a random sample u from $U \sim \text{Uniform}(0,1)$.
2. Calculate your random sample from X as $x = F^{-1}(u)$.

A notable exception is when the cdf cannot be written in closed form, as is the case for the normal distribution.

Example 4.2.1. We require an observation from the distribution with cdf

$$F_X(x) = 1 - e^{-x}, \quad x > 0.$$

Explain how we can generate a random observation from this distribution.

Solution. Let $y = F_X(x)$. Then $x = F_X^{-1}(y)$ and $y = 1-e^{-x} \Rightarrow x = -\ln(1-y)$, so $F_X^{-1}(y) = -\ln(1-y)$. Thus, if $X = F_X^{-1}(U) = -\ln(1-U)$ where $U \sim \text{Uniform}(0,1)$, then X has cdf $F_X(x) = 1 - e^{-x}$, $x > 0$. □

4.3 Bivariate Transformations

If X and Y have joint density $f_{X,Y}(x,y)$ and U is a function of X and Y, we can find the density of U by calculating $F_U(u) = \mathbb{P}(U \le u)$ and differentiating.

Example 4.3.1. $f_{X,Y}(x,y) = 1$, $0 < x < 1$, $0 < y < 1$. Let $U = X + Y$. Find $f_U(u)$.

Solution.

$$\begin{aligned} F_U(y) &= \mathbb{P}(X + Y \le u) \\ &= \begin{cases} \displaystyle\int_0^u \int_0^{u-y} 1 dx dy, & 0 < u < 1 \\ \displaystyle 1 - \int_{u-1}^1 \int_{u-y}^1 1 dx dy, & 1 < u < 2 \end{cases} \\ &= \begin{cases} \dfrac{u^2}{2}, & 0 < u < 1 \\ 2u - \dfrac{u^2}{2} - 1, & 1 < u < 2. \end{cases} \end{aligned}$$

Thus

$$f_U(u) = \begin{cases} u, & 0 < u < 1 \\ 2 - u, & 1 < u < 2. \end{cases}$$

□

Theorem 4.3.1. *If U and V are functions of continuous random variables X and Y, then*

$$f_{U,V}(u,v) = f_{X,Y}(x,y) \cdot |J|$$

where

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix}$$

is a determinant called the Jacobian of the transformation. The full specification of $f_{U,V}(u,v)$ requires that the range of (u,v) values corresponding to those (x,y) for which $f_{X,Y}(x,y) > 0$ is determined.

The Jacobian has an interpretation similar to that of the factor $|\frac{dx}{dy}|$ in the univariate density transformation formula. The transformations U, V transform a small rectangle, with area $\delta x \delta y$ in the xy-plane, into a small parallelogram with area $\delta x \delta y / J$ in the u, v plane. To find $f_U(u)$ by bivariate transformation:

1. Define some bivariate transformation to (U, V)
2. Find $f_{U,V}(u,v)$
3. We want the marginal distribution of U. So now find $f_U(u) = \int_{-\infty}^{\infty} f_{U,V}(u,v)dv$.

Using a bivariate transformation to find the distribution of U is often more convenient than deriving it via the cdf. Using the cdf requires double integration, which we can avoid when using a bivariate transformation.

Example 4.3.2. X, Y are independent Uniform$(0,1)$ variables, such that $f_{X,Y}(x,y) = 1$ for $0 < x < 1$, $0 < y < 1$. Let $U = X + Y$ and $V = Y$. Use a bivariate transformation to (U, V) to find the density function of U.

Solution. First, note that $X = U - V$, $Y = V$ and so $x = u - v$, $y = v$ gives

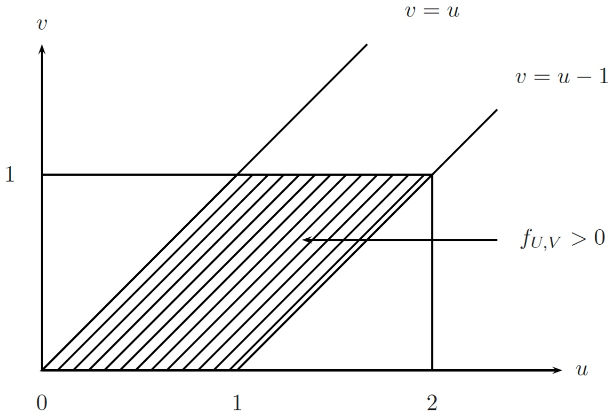
$$\frac{\partial x}{\partial u} = 1, \quad \frac{\partial x}{\partial v} = -1, \quad \frac{\partial y}{\partial u} = 0, \quad \frac{\partial y}{\partial v} = 1$$

and

$$J = \begin{vmatrix} 1 & -1 \\ 0 & 1 \end{vmatrix} = 1.$$

Now $0 < x < 1 \iff 0 < u - v < 1$ and $0 < y < 1 \iff 0 < v < 1$. Therefore

$$f_{U,V}(u,v) = f_{X,Y}(x,y)|J| = 1, \quad v < u < 1 + v, \ 0 < v < 1.$$



Hence,

$$f_U(u) = \begin{cases} \displaystyle\int_0^u 1 \cdot dv = u, & 0 < u < 1 \\ \\ \displaystyle\int_{u-1}^1 1 \cdot dv = 2 - u, & 1 < u < 2. \end{cases}$$

Note that $V = Y$, so $f_V(v) = 1$, $0 < v < 1$. □

Example 4.3.3.

$$f_{X,Y}(x,y) = 3y, \quad 0 < x < y < 1.$$

Let $U = X + Y$, $V = Y - X$. Use a bivariate transformation to (U, V) to find the density function of F.

Solution. Now,

$$X = \frac{U - V}{2}, \quad Y = \frac{U + V}{2},$$

and so $x = \frac{u-v}{2}, y = \frac{u+v}{2}$ give

$$\frac{\partial x}{\partial u} = \frac{1}{2}, \quad \frac{\partial x}{\partial v} = -\frac{1}{2}, \quad \frac{\partial y}{\partial u} = \frac{1}{2}, \quad \frac{\partial y}{\partial v} = \frac{1}{2}.$$

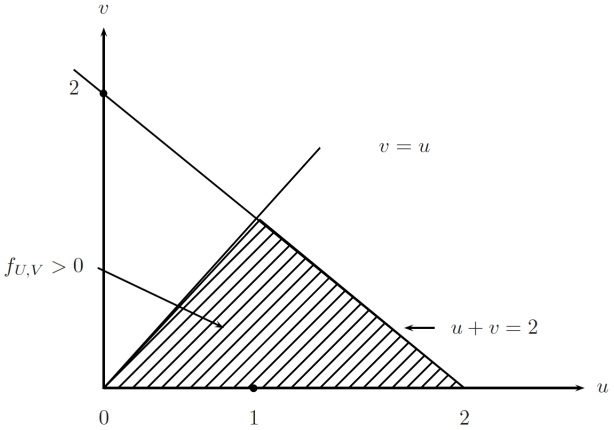
Therefore,

$$J = \begin{vmatrix} 1/2 & -1/2 \\ 1/2 & 1/2 \end{vmatrix} = \frac{1}{2}$$

and

$$0 < x < y \iff 0 < \frac{u - v}{2} < \frac{u + v}{2} \iff 0 < v < u,$$

$$y < 1 \iff \frac{u + v}{2} < 1 \iff u + v < 2.$$



Hence,

$$f_{U,V}(u,v) = \frac{3(u+v)}{2} \left|\frac{1}{2}\right| = \frac{3}{4}(u+v), \quad 0 < v < u, \ u + v < 2.$$

and

$$\text{For } 0 < u < 1, \ f_U(u) = \int_0^u \frac{3}{4}(u+v)dv = \frac{9u^2}{8}.$$

$$\text{For } 1 < u < 2, \ f_U(u) = \int_0^{2-u} \frac{3}{4}(u+v)dv = \frac{3}{2} - \frac{3u^2}{8}.$$

As an aside, note that $f_V(v) = \displaystyle\int_v^{2-v} \frac{3}{4}(u+v)du = \frac{3}{2}(1 - v^2)$, $0 < v < 1$. □

Example 4.3.4. The lifetimes X and Y of two brands of components of a system are independent with

$$f_X(x) = xe^{-x}, \ x > 0 \quad \text{and} \quad f_Y(y) = e^{-y}, \ y > 0.$$

The relative efficiency of the components, is measured as $U = \frac{Y}{X}$. Find the density function of the relative efficiency, using a bivariate transformation.

Solution.

$$f_{X,Y}(x,y) = xe^{-(x+y)}, \quad x>0, y>0.$$

Now $U = \frac{Y}{X}$. Let $V = X$. Then

$$X = V, \quad Y = UV \text{ and if } x = v, y = uv,$$

$$\frac{\partial x}{\partial u} = 0, \quad \frac{\partial x}{\partial v} = 1, \quad \frac{\partial y}{\partial u} = v, \quad \frac{\partial y}{\partial v} = u, \text{ so}$$

$$J = \begin{vmatrix} 0 & 1 \\ v & u \end{vmatrix} = -v.$$

Now $x > 0 \iff v > 0$ and $y > 0 \iff uv > 0 \Rightarrow u > 0$ since $v > 0$. Therefore,

$$f_{U,V}(u,v) = f_{X,Y}(x,y)|J| = ve^{-(v+uv)}|-v| = v^2e^{-v(1+u)}, \quad u>0, v>0.$$

and so

$$f_U(u) = \int_0^\infty v^2 e^{-v(1+u)} dv = \frac{2}{(1+u)^3}, \quad u > 0.$$

Note also, $\mathbb{E}(U) = \displaystyle\int_0^\infty u \cdot \frac{2}{(1+u)^3} du = 1.$ □

Example 4.3.5. Suppose X denotes the total time from arrival to exit from a service queue and Y denotes the time spent in the queue before being served. Suppose also that we want the density of $U = X - Y$, the amount of time spent being served when

$$f_{X,Y}(x,y) = e^{-x}, \quad 0 < y < x < \infty.$$

Now $U = X - Y$. Let $V = Y$. Find the density function of U, using a bivariate transformation.

Solution. Then

$$X = U + V, \quad Y = V \text{ and if } x = u+v, \ y = v,$$

$$\text{then } \frac{\partial x}{\partial u} = 1, \quad \frac{\partial x}{\partial v} = 0, \quad \frac{\partial y}{\partial u} = 1, \quad \frac{\partial y}{\partial v} = 1.$$

Then

$$J = \begin{vmatrix} 1 & 1 \\ 0 & 1 \end{vmatrix} = 1 \text{ and}$$

$$0 < y < x < \infty \iff 0 < v < u + v < \infty$$

$$\Rightarrow v > 0, u > 0 \text{ and } u + v < \infty \Rightarrow u < \infty, v < \infty.$$

$$f_{U,V}(u,v) = e^{-(u+v)}, u > 0, v > 0.$$

Thus the time spent being served, U, and the time spent in the queue before service, V, are independent random variables. Also,

$$f_U(u) = e^{-u}, u > 0 \text{ and } f_V(v) = e^{-v}, v > 0.$$

□

4.4 Multivariate Transformations

Theorem 4.4.1 (Change of Variable Formula)**.** *Suppose we are given the n-dimensional integral*

$$\int \cdots \int_{\mathcal{R}} f(\mathbf{x}) d\mathbf{x}$$

and the invertible transformation

$$\mathbf{z} = \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix} = \mathbf{g}(\mathbf{x}) = \begin{pmatrix} g_1(\mathbf{x}) \\ \vdots \\ g_n(\mathbf{x}) \end{pmatrix}$$

which maps the region $\mathcal{R}$ into $\mathcal{S}$. Then, if we change the variable from $\mathbf{x}$ to $\mathbf{z} = \mathbf{g}(\mathbf{x})$ we obtain

$$\int \cdots \int_{\mathcal{R}} f(\mathbf{x}) d\mathbf{x} = \int \cdots \int_{\mathcal{S}} f(\mathbf{g}^{-1}(\mathbf{z})) |\det(J_{\mathbf{g}^{-1}}(\mathbf{z}))| d\mathbf{z},$$

where $|\det(J_{\mathbf{g}^{-1}}(\mathbf{z}))|$ stands for the absolute value of the determinant of the Jacobian matrix of the transformation $\mathbf{g}^{-1}$ evaluated at $\mathbf{z}$.

The Jacobian matrix of a transformation $\mathbf{g}$ is defined as the matrix of partial derivatives

$$J_{\mathbf{g}}(\mathbf{x}) = \begin{pmatrix} \frac{\partial g_1}{\partial x_1}(\mathbf{x}) & \cdots & \frac{\partial g_1}{\partial x_n}(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial g_n}{\partial x_1}(\mathbf{x}) & \cdots & \frac{\partial g_n}{\partial x_n}(\mathbf{x}) \end{pmatrix}.$$

Note that

$$|\det(J_{\mathbf{g}^{-1}})(\mathbf{z})| = \frac{1}{|\det(J_{\mathbf{g}})(\mathbf{x})|}.$$

Theorem 4.4.2. *Suppose $\mathbf{X} \in \mathbb{R}^n$ has pdf $f_{\mathbf{X}}(\mathbf{x})$ and we transform $\mathbf{X}$:*

$$\mathbf{Z} = \mathbf{g}(\mathbf{X}),$$

where $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^n$ is invertible and continuously differentiable. Then, the pdf of $\mathbf{Z}$ is

$$f_{\mathbf{Z}}(\mathbf{z}) = f_{\mathbf{X}}(\mathbf{g}^{-1}(\mathbf{z})) |\det(J_{\mathbf{g}^{-1}})(\mathbf{z})|.$$

Proof. We can write for any sufficiently nice region $\mathcal{A} \subseteq \mathcal{S}$ that maps to $\mathcal{B} = \mathbf{g}^{-1}(\mathcal{A})$ under the transformation $\mathbf{g}$:

$$\begin{aligned} \int_{\mathcal{A}} f_{\mathbf{Z}}(\mathbf{z}) d\mathbf{z} &= \mathbb{P}(\mathbf{Z} \in \mathcal{A}) && \text{(by definition of pdf of } \mathbf{Z}\text{)} \\ &= \mathbb{P}(\mathbf{g}(\mathbf{X}) \in \mathcal{A}) \\ &= \mathbb{P}(\mathbf{X} \in g^{-1}(\mathcal{A})) \\ &= \int_{\mathbf{g}^{-1}(\mathcal{A})} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} && \text{(by definition of pdf of } \mathbf{X}\text{)} \\ &= \int_{\mathcal{A}} f_{\mathbf{X}}(g^{-1}(\mathbf{z})) |\det(J_{\mathbf{g}^{-1}})(\mathbf{z})| d\mathbf{z}. && \text{(by change of variable formula)} \end{aligned}$$

Hence it follows that for any nice region $\mathcal{A} \subseteq \mathcal{S}$ we have:

$$\int_{\mathcal{A}} f_{\mathbf{Z}}(\mathbf{z}) d\mathbf{z} = \int_{\mathcal{A}} f_{\mathbf{x}}(\mathbf{g}^{-1}(\mathbf{z})) |\det(J_{\mathbf{g}^{-1}})(\mathbf{z})| d\mathbf{z}.$$

This suggests that the integrands are equivalent in some sense, because they give us the same integral for any region $\mathcal{A}$. Without getting into technicalities, we can see why it is plausible that

$$f_{\mathbf{Z}}(\mathbf{z}) = f_{\mathbf{X}}(\mathbf{g}^{-1}(\mathbf{z})) |\det(J_{\mathbf{g}^{-1}})(\mathbf{z})|.$$

□

Theorem 4.4.3. *Suppose that* $\mathbf{X} \in \mathcal{R}$ *has pdf*

$$f_{\mathbf{X}}(\mathbf{x})$$

and we consider the pdf of $\mathbf{Z} = A\mathbf{X}$*, where* A *is an invertible* $n \times n$ *matrix. Setting* $\mathbf{g}(\mathbf{x}) = A\mathbf{x}$ *in the result above, we obtain that*

$$\mathbf{g}^{-1}(\mathbf{z}) = A^{-1}\mathbf{z}$$

and

$$J_{g^{-1}}(\mathbf{z}) = A^{-1}$$

so that

$$f_{\mathbf{Z}}(\mathbf{z}) = f_{\mathbf{X}}(A^{-1}\mathbf{z}) |\det(A^{-1})| = \frac{1}{|\det(A)|} f_{\mathbf{X}}(A^{-1}\mathbf{z}).$$

4.5 Sums of Independent Random Variables

4.5.1 Probability Function/Density Function Approach

Theorem 4.5.1 (Discrete Convolution Formula)**.** *Suppose that* X *and* Y *are independent random variables taking only non-negative integer values, and let* $Z = X + Y$*. Then*

$$f_Z(z) = \sum_{y=0}^{z} f_X(z-y) f_Y(y), \quad z = 0, 1, \ldots.$$

Proof.

$$\begin{aligned} f_Z(z) &= \mathbb{P}(X + Y = z) \\ &= \mathbb{P}(X = z, Y = 0) + \mathbb{P}(X = z-1, Y = 1) + \cdots + \mathbb{P}(X = 0, Y = z) \\ &= \mathbb{P}(X = z)\mathbb{P}(Y = 0) + \mathbb{P}(X = z-1)\mathbb{P}(Y = 1) + \cdots + \mathbb{P}(X = 0)\mathbb{P}(Y = z) \\ &= f_X(z) f_Y(0) + f_X(z-1) f_Y(1) + \cdots + f_X(0) f_Y(z) \\ &= \sum_{y=0}^{z} f_X(z-y) f_Y(y). \end{aligned}$$

□

Example 4.5.1. $X \sim \text{Poisson}(\lambda_1), Y \sim \text{Poisson}(\lambda_2)$ with

$$\mathbb{P}(X = k) = \frac{e^{-\lambda_1}\lambda_1^k}{k!}, \quad k = 0, 1, 2, \ldots$$

$$\mathbb{P}(Y = k) = \frac{e^{-\lambda_2}\lambda_2^k}{k!}, \quad k = 0, 1, 2, \ldots$$

Find the probability function of $Z = X + Y$.

Solution.

$$\begin{aligned} f_{X+Y}(z) &= \sum_{k=0}^{z} \frac{e^{-\lambda_1}\lambda_1^k}{k!} \cdot \frac{e^{-\lambda_2}\lambda_2^{z-k}}{(z-k)!} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{z!} \sum_{k=0}^{z} \binom{z}{k} \lambda_1^k \lambda_2^{z-k} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^z}{z!}, \quad z = 0, 1, 2, \ldots. \end{aligned}$$

Thus $X + Y \sim \text{Poisson}(\lambda_1 + \lambda_2)$. It follows by induction that if $X_1, X_2, \dots, X_n$ are independent with $X_i \sim \text{Poisson}(\lambda_i)$, then

$$\sum_{i=1}^{n} X_i \sim \text{Poisson}\left(\sum_{i=1}^{n} \lambda_i\right).$$

This is an important and useful property of Poisson random variables. □

Theorem 4.5.2 (Continuous Convolution Formula)**.** *Suppose X and Y are independent continuous variables with $X \sim f_X(x)$ and $Y \sim f_Y(y)$. Then $Z = X + Y$ has density*

$$f_Z(z) = \int_{\textit{all possible } y} f_X(z-y) f_Y(y) dy.$$

Proof.

$$\begin{aligned} F_Z(z) &= \mathbb{P}(Z \leq z) \\ &= \mathbb{P}(X + Y \leq z) \\ &= \int \cdots \int_{x+y \leq z} f_{X,Y}(x,y) dx dy \\ &= \int_{\text{all possible } y} \int_{-\infty}^{z-y} f_X(x) f_Y(y) dx dy \\ &= \int_{\text{all possible } y} F_X(z-y) f_y(y) dy. \end{aligned}$$

To complete the proof we differentiate w.r.t z in order to obtain the density function $f_Z(z)$:

$$f_Z(z) = \frac{d}{dz} F_Z(z) = \int_{\text{all possible } y} f_X(z-y) f_Y(y) dy.$$

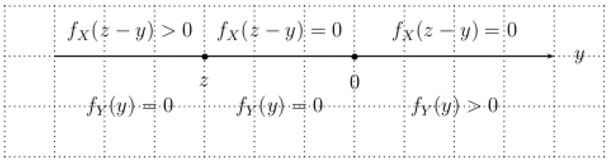
□

Example 4.5.2. X and Y are independent variables and $f_X(x) = e^{-x}$, $x > 0$, $f_Y(y) = e^{-y}$, $y > 0$. Find the density function of $Z = X + Y$.

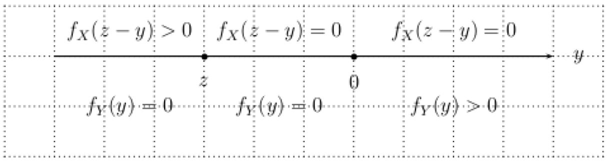
Solution.

$$Z = X + Y \sim f_Z(z) = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y) dy.$$

Now $f_X(z-y) = e^{-(z-y)}$ for $z - y > 0$ or $y < z$. If $z < 0$



Therefore if $z < 0$, $f_X(z-y)f_Y(y) = 0$ for all $y \iff f_Z(z) = 0$ for $z < 0$ as expected.



If $z < 0$,

$$f_X(z-y)f_Y(y) = \begin{cases} 0 & \text{for } y < 0 \text{ and } y > z \\ e^{-(z-y)} \cdot e^{-y} = e^{-z} & \text{for } 0 < y < z. \end{cases}$$

$$\Rightarrow f_Z(y) = \int_0^z e^{-z} dy = ze^{-z}, \quad z > 0.$$

Note that the answer is the density function of a Gamma$(2,1)$ random variable. Thus the sum of two independent exponential Exp(1) random variables is a Gamma$(2,1)$ variable. □

Example 4.5.3. Y_1 and Y_2 are independent variables and $Y_1 \sim N(0,1)$, $Y_2 \sim N(0,1)$, with

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad -\infty < x < \infty.$$

Find the distribution of $Z = Y_1 + Y_2$.

Solution. Z has density

$$f_Z(z) = \int_{-\infty}^{\infty} f_{Y_1}(z-y) f_{Y_2}(y) dy$$

where

$$f_{Y_1}(z-y) = \frac{1}{\sqrt{2\pi}} e^{-(z-y)^2/2}, \quad -\infty < z-y < \infty \quad \text{and} \quad f_{Y_2}(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}, \quad -\infty < y < \infty.$$

For any fixed $z \in (-\infty, \infty)$, when considered as a function of y,

$$f_{Y_1}(z-y) = \frac{1}{\sqrt{2\pi}} e^{-(z-y)^2/2}, \quad -\infty < y < \infty.$$

Therefore,

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(z-y)^2/2} \cdot \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-z^2/2 + zy - y^2} dy \\ &= \frac{e^{-z^2/2}}{2\pi} \int_{-\infty}^{\infty} e^{-(y^2 - zy + \frac{z^2}{4}) + \frac{z^2}{4}} dy \\ &= \frac{e^{-z^2/4}}{2\pi} \int_{-\infty}^{\infty} e^{-(y-z/2)^2} dy. \end{aligned}$$

Now for any μ and σ,

$$\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2} dy = 1 \quad \text{and} \quad \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}(y-\mu)^2} dy = \sigma\sqrt{2\pi}.$$

Put $\sigma^2 = \frac{1}{2}$ and $\mu = \frac{z}{2}$. Then $\int_{-\infty}^{\infty} e^{-(y-\frac{z}{2})^2} dy = \sqrt{\pi}$. Therefore,

$$f_Z(z) = \frac{e^{-z^2/4}}{2\pi}\sqrt{\pi} = \frac{1}{2\sqrt{\pi}} e^{-z^2/4}, \quad -\infty < z < \infty.$$

Thus $Z \sim N(0,2)$. More generally, we can show that the sum of any two normal random variables is also normal. □

Example 4.5.4. $X \sim$ Gamma$(\alpha_1, 1)$ and $Y \sim$ Gamma$(\alpha_2, 1)$. Then

$$f_X(x) = \frac{e^{-x} x^{\alpha_1 - 1}}{\Gamma(\alpha_1)}, \quad x > 0; \quad \alpha_1 \geq 1$$

$$f_Y(y) = \frac{e^{-y} y^{\alpha_2 - 1}}{\Gamma(\alpha_2)}, \quad y > 0; \quad \alpha_2 \geq 1$$

Find the distribution of $Z = X + Y$.

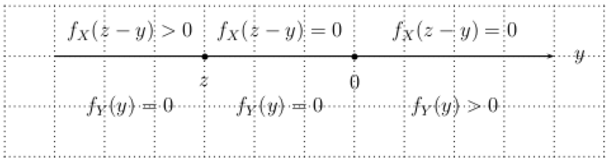
Solution. Z has density

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y) dy.$$

First we will find the limits on this integral, that is, the range of values of y for which $f_X(z-y)$ and $f_Y(y)$ are both non-zero. Both X and Y are only defined over positive values, so the range of values of y of interest to us satisfies both $z - y > 0$ and $y > 0$, i.e. $0 < y < z$.

$$f_X(z-y) = \frac{e^{-(z-y)}(z-y)^{\alpha_1 - 1}}{\Gamma(\alpha_1)} \quad \text{for } z - y > 0 \text{ or } y < z.$$

$f_Z(z) = 0$ for $z < 0$ since X and Y are non-negative random variables.
For $z > 0$



we have $f_X(z-y)f_Y(y) = 0$ for $y < 0$ or $y > z$.
For $0 < y < z$,

$$\begin{aligned} f_X(z-y)f_Y(y) &= \frac{e^{-(z-y)}(z-y)^{\alpha_1-1}}{\Gamma(\alpha_1)} \cdot \frac{e^{-y}y^{\alpha_2-1}}{\Gamma(\alpha_2)} \\ &= \frac{e^{-z}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}(z-y)^{\alpha_1-1}y^{\alpha_2-1}. \end{aligned}$$

Hence,

$$\begin{aligned} f_Z(z) &= \int_0^z \frac{e^{-z}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}(z-y)^{\alpha_1-1}y^{\alpha_2-1}dy && \left(\text{now substitute } t = \frac{y}{z}\right) \\ &= \frac{e^{-z}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^1 [z(1-t)]^{\alpha_1-1}[zt]^{\alpha_2-1}z\, dt \\ &= \frac{e^{-z}z^{\alpha_1-1+\alpha_2-1+1}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^1 (1-t)^{\alpha_1-1}t^{\alpha_2-1}dt \\ &= \frac{e^{-z}z^{\alpha_1+\alpha_2-1}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} B(\alpha_2, \alpha_1), && \text{where } B(\alpha_2,\alpha_1) = B(\alpha_1,\alpha_2) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)}, \end{aligned}$$

and so

$$f_Z(z) = \frac{e^{-z}z^{\alpha_1+\alpha_2-1}}{\Gamma(\alpha_1+\alpha_2)}, \quad z > 0.$$

That is, $Z \sim \text{Gamma}(\alpha_1 + \alpha_2, 1)$. Note, the Beta (α_1, α_2) density is $\frac{x^{\alpha_1-1}(1-x)^{\alpha_2-1}}{B(\alpha_1,\alpha_2)}$. □

Theorem 4.5.3. *If $X_1, X_2, \ldots, X_n$ are independent with $X_i \sim \text{Gamma}(\alpha_i, \beta)$, then*

$$\sum_{i=1}^n X_i \sim \text{Gamma}\left(\sum_{i=1}^n \alpha_i, \beta\right).$$

4.5.2 Moment Generating Function Approach

Theorem 4.5.4. *Suppose that X and Y are independent random variables with moment generating functions m_X and m_Y. Then*

$$m_{X+Y}(u) = m_X(u)m_Y(u).$$

Proof. If X and Y are independent, then $Z = X + Y$ has mgf

$$\begin{aligned} m_Z(u) &= \mathbb{E}(e^{u(X+Y)}) \\ &= \mathbb{E}(e^{uX} \cdot e^{uY}) \\ &= \mathbb{E}(e^{uX})\mathbb{E}(e^{uY}) \\ &= m_X(u) \cdot m_Y(u) \end{aligned}$$

so the mgf for the density

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z-y)f_Y(y)dy = m_X(u)m_Y(u).$$

An alternative proof is as follows:

$$\begin{aligned} \mathbb{E}(e^{uZ}) &= \int_{-\infty}^{\infty} e^{uz} f_Z(z)dx \\ &= \int_{-\infty}^{\infty} e^{uz} \int_{-\infty}^{\infty} f_X(z-y)f_Y(y)dydz \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{u(z-y)} f_X(z-y)dz \cdot e^{uy} f_Y(y)dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{ux} f_X(x)dx \cdot e^{uy} f_Y(y)dy, && \text{where } x = z - y \\ &= \int_{-\infty}^{\infty} m_X(u) \cdot e^{uy} f_Y(y)dy \\ &= m_X(u)\int_{-\infty}^{\infty} e^{uy} f_Y(y)dy \\ &= m_X(u)m_Y(u). \end{aligned}$$

□

Theorem 4.5.5. *If $X_1, X_2, \ldots, X_n$ are independent random variables, then $\sum_{i=1}^n X_i$ has moment generating function*

$$m_{\sum_{i=1}^n X_i}(u) = \prod_{i=1}^n m_{X_i}(u).$$

Proof.

$$\begin{aligned} m_{\sum_{i=1}^n X_i}(u) &= \mathbb{E}(e^{u\sum_{i=1}^n X_i}) \\ &= \mathbb{E}\left(\prod_{i=1}^n e^{uX_i}\right) \\ &= \prod_{i=1}^n \mathbb{E}(e^{uX_i}) \\ &= \prod_{i=1}^n m_{X_i}(u). \end{aligned}$$

□

This offers us a useful approach for deriving the distribution of the sum of independent random variables, using the one-to-one correspondence between distributions and moment generating functions. For this approach to work however we need to be able to recognise the distribution of the sum from its moment generating function.

Example 4.5.5. $X_1, X_2, \ldots, X_n$ independent Bernoulli (p) random variables. Use moment generating functions to show that $\sum_{i=1}^n X_i \sim \mathrm{Bin}(n, p)$.

Solution. Each X_i has probability function

$$f_X(x) = p^x(1-p)^{1-x}, \quad x = 0, 1;\ 0 < p < 1$$

and mgf

$$\begin{aligned} m_X(u) &= \mathbb{E}(e^{uX}) \\ &= \sum_{k=0}^{1} e^{uX}\mathbb{P}(X = x) \\ &= e^{u\cdot 0}\mathbb{P}(X = 0) + e^{u\cdot 1}\mathbb{P}(X = 1) \\ &= 1 - p + pe^u. \end{aligned}$$

Therefore, the mgf of the sum $\sum_{i=1}^n X_i$ is

$$m_{\sum_{i=1}^n X_i}(u) = \prod_{i=1}^n m_{X_i}(u) = (1 - p + pe^u)^n$$

which is the mgf of a $\mathrm{Bin}(n, p)$ random variable. □

Example 4.5.6. $X_1, X_2, \ldots, X_n$ independent random variables with $X_i \sim \mathrm{Poisson}(\lambda_i)$. Find the mgf of X_i and hence deduce the distribution of $\sum_{i=1}^n X_i$.

Solution. If $X \sim \mathrm{Poisson}(\lambda)$, then X has the mgf

$$\begin{aligned} m_X(u) &= \mathbb{E}(e^{uX}) \\ &= \sum_{x=0}^{\infty} e^{ux} \cdot \frac{e^{-\lambda}\lambda^x}{x!} \\ &= e^{-\lambda}\sum_{x=0}^{\infty} \frac{(\lambda e^u)^x}{x!} \\ &= e^{-\lambda} \cdot e^{\lambda e^u} \\ &= e^{\lambda(e^u - 1)}. \end{aligned}$$

Thus X_i above has mgf

$$m_{X_i}(u) = \mathbb{E}(e^{uX_i}) = e^{\lambda_i(e^u - 1)}$$

and $\sum_{i=1}^n X_i$ has mgf

$$m_{\sum_{i=1}^n X_i}(u) = \mathbb{E}(e^{u\sum_{i=1}^n X_i}) = \prod_{i=1}^n \mathbb{E}(e^{uX_i}) = e^{(\sum_{i=1}^n \lambda_i)(e^u - 1)}.$$

This has the form of the Poisson mgf as derived above, but now with parameter $\sum_{i=1}^n \lambda_i$. Therefore,

$$\sum_{i=1}^n X_i \sim \mathrm{Poisson}\left(\sum_{i=1}^n \lambda_i\right).$$

□

Theorem 4.5.6. *If $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ are independent then*

$$X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2).$$

Theorem 4.5.7. *If for $1 \leq i \leq n$, $X_i \sim N(\mu_i, \sigma_i^2)$ are independent then for any set of constants $a_1, \ldots, a_n$,*

$$\sum_{i=1}^{n} a_i X_i \sim N\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right).$$

4.6 Survey Design

When collecting data in a survey it is critical to try and collect data that is *representative* and *random.*

4.6.1 Representativeness

When we collect a sample from a population, typically we would like to use this sample to make inferences about some property of the population at large. However, this is only reasonable if *the sample is representative of the population.* If this is not achieved, then inferences about the population can be wrong.

Definition 4.6.1. Consider a sample $X_1, \ldots, X_n$ from a random variable X which has probability or density function $f_X(x)$. The sample is said to be representative if:

$$f_{X_i}(x) = f_X(x) \quad \text{for each } i.$$

Representativeness is typically a *more important consideration than sample size.*

4.6.2 Random Samples

Definition 4.6.2. A *random sample* of size n is a set of random variables

$$X_1, \ldots, X_n$$

with the properties

(i) the X_i's each have the same probability distribution

(ii) the X_i's are independent.

We often say that X_i are *iid* (independently and identically distributed).

Example 4.6.1. Consider sampling a variable X in a population of 10 subjects, which take the following (sorted) values:

$$2 \quad 4 \quad 5 \quad 7 \quad 8 \quad 10 \quad 14 \quad 17 \quad 27 \quad 35.$$

We sample three subjects randomly (with equal sampling probability for each subject), with replacement. Let these values be X_1, X_2 and X_3. Show that X_1, X_2 and X_3 are iid.

Solution. Now, $f_{X_1}(x) = \mathbb{P}(X_1 = x) = \frac{1}{10}$ for each $x \in \{2, 4, 5, 7, 8, 10, 14, 17, 27, 35\}$, $f_{X_2}(x) = f_{X_3}(x) = f_{X_1}(x)$ hence identically distributed. Since with replacement, $f_{X_2|X_1}(x) = f_{X_2}(x)$ (similarly for the other combinations) hence the X_i are independently distributed. □

It is more common however to sample without replacement. The most common method of obtaining such a random sample is to take a simple random sample.

Definition 4.6.3. A *simple random sample* of size n is a set of subjects sampled in such a way that all possible samples of size n are equally likely.

Strictly speaking, a simple random sample does not consist of iid random variables - they are identically distributed, but they are dependent, since knowledge that $X_i = x_i$ makes it less likely that $X_j = x_i$ because the i-th subject can only be included in the sample once. However, this dependence is very weak when the population size N is large compared to the sample size n (e.g. if $N > 100n$) and so in most instances it can be ignored.

It is important in surveys, whenever possible, to ensure sampling is random. This is important for a few reasons:

- It ensures the n values in the sample are iid, which is an important assumption of most methods of statistical inference
- Random sampling removes selection bias - the choice of who is included in the study is taken away from the experimenter, hence it is not possible for them to (intentionally or otherwise) manipulate results through choice of subjects
- Random sampling from the population of interest guarantees that the sample is representative of the population

Unfortunately, it is typically very hard to obtain a simple random sample from the population of interest, so the best we can hope for is a "good approximation".

4.6.3 Statistics Calculated from Samples

Definition 4.6.4. Let $X_1, \dots, X_n$ be a random sample. A *statistic* is any real-valued function of the random sample.

While any real-valued function can be considered a statistic, in practice we focus on particular functions which measure something of interest to us about the sample. Important examples of statistics are:

- $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, the *sample mean*
- $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$, the *sample variance*
- $\tilde{X}_{0.5}$, the *sample median.*

A key advantage of random sampling is the fact that the sample is random implies that *any statistic calculated from the sample is random.* Hence we can treat statistics as random variables and study their properties.

Theorem 4.6.1. *If $X_1, \dots, X_n$ is a random sample from a variable with mean μ and variance σ^2, then the sample mean $\overline{X}$ satisfies:*

$$\mathbb{E}(\overline{X}) = \mu \quad \textit{and} \quad \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}.$$

Theorem 4.6.2. *If $X_1, \dots, X_n$ is a random sample from a* Bernoulli(p) *variable, then the sample proportion $\widehat{p}$ satisfies:*

$$\mathbb{E}(\widehat{p}) = p \quad \textit{and} \quad \mathrm{Var}(\widehat{p}) = \frac{p(1-p)}{n}.$$

Note that while the variance results given above require the variables to be iid (hence a random sample), the expectation results only requires the observations in the sample to be identically distributed. Because of the property that $\mathbb{E}(\overline{X}) = \mu$, we say that sample means of random samples are *unbiased.* Similarly, $\widehat{p}$ is unbiased.

4.6.4 Methods of Survey Sampling

Definition 4.6.5. Consider two unbiased alternative statistics, denoted as $g(X_1, \ldots, X_n)$ and $h(Y_1, \ldots, Y_m)$. We say that $g(X_1, \ldots, X_n)$ is more *efficient* than $h(Y_1, \ldots, Y_m)$ if:

$$\mathrm{Var}[g(X_1, \ldots, X_n)] < \mathrm{Var}[h(Y_1, \ldots, Y_m)].$$

Note that the above notation implies that not only can the statistics that we are using differ (g vs h), but the observations used to calculate the statistics can also differ (X vs Y). This reflects that there are two ways to achieve efficiency - use a different statistic or by sampling differently. The most obvious way that sampling differently can increase efficiency is by increasing the sample size, but even for a fixed sample size ($n = m$) efficiency varies with sampling method.

- **Simple Random Sample**: Weaknesses are that it can be difficult to implement in practice, requiring high effort, and it can be inefficient.
- **Stratified Random Sample**: If the population can be broken into subpopulations (or "strata") which differ from each other in the variable of interest, it is more efficient to sample separately within each stratum than to sample once across the whole population e.g. Estimating average taxable income - this varies considerably with age, so a good survey design would involve sampling separately within age strata (if possible).
- **Cluster Sampling**: This is useful when subjects in the population arise in clusters, and it takes less effort to sample within clusters than across clusters. Effort-per-subject can be reduced by sampling clusters and then measuring all (or sampling many) subjects within a cluster. For example, face-to-face interviews with 100 NSW household owners - it is easier logistically to sample ten postcodes, then sample ten houses in each postcode, than to travel to a random sample 100 households spread across (potentially) 100 NSW postcodes.

Example 4.6.2. Consider estimating the average heart rate of students, μ. Males and females are known to have different heart rates, μ_M and μ_F, but the same variance σ^2. Consider estimating the mean μ using a stratified random sample, as follows:

- Take a random sample of size n of each gender.
- Calculate the sample mean of each gender $\overline{X}_M$ and $\overline{X}_F$.
- Since males and females occur with (approximately) equal frequency in the student population, we can estimate the overall mean heart rate as

$$\overline{X}_s = \frac{1}{2}(\overline{X}_M + \overline{X}_F).$$

Now,

1. Find $\mathrm{Var}(\overline{X}_s)$
2. Show that the marginal variance of heart rate across the student population (ignoring gender) is $\mathrm{Var}(X) = \sigma^2 + (\mu_M - \mu_F)^2/4$.
3. Hence show that stratified random sampling is more efficient than using a simple random sample of size $2n$, if $\mu_M \neq \mu_F$.

Solution. We are given that $\mathbb{E}(\overline{X}_M) = \mu_M$ and $\mathbb{E}(\overline{X}_F) = \mu_F$ and $\mathrm{Var}(\overline{X}_M) = \mathrm{Var}(\overline{X}_F) = \frac{1}{n}\sigma^2$. Therefore,

$$\mathrm{Var}(\overline{X}_s) = \frac{1}{4}[\sigma^2/n + \sigma^2/n] = \frac{\sigma^2}{2n}.$$

But if X is the heart rate of any arbitrary selected student (so that there is equal chance of being male and female), then we are given that $\mathbb{E}[X|M] = \mu_M$ and $\mathbb{E}[X|F] = \mu_F$ so that

$$\mathbb{E}[X] = \mathbb{E}[X|M]\mathbb{P}(M) + \mathbb{E}[X|F]\mathbb{P}(F) = \frac{\mu_M + \mu_F}{2} = \mu,$$

and similarly for the variance

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}(X-\mu)^2 \\
&= \mathbb{E}[(X-\mu)^2|M]\mathbb{P}(M) + \mathbb{E}[(X-\mu)^2|F]\mathbb{P}(F) \\
&= \frac{1}{2}\mathbb{E}(X_M-\mu)^2 + \frac{1}{2}\mathbb{E}(X_F-\mu)^2 \\
&= \frac{1}{2}\mathbb{E}(X_M-\mu_M+\mu_M-\mu)^2 + \frac{1}{2}\mathbb{E}(X_M-\mu_F+\mu_F-\mu)^2 \\
&= \sigma^2 + \frac{1}{2}\mathbb{E}(\mu_M-\mu)^2 + \frac{1}{2}\mathbb{E}(\mu_F-\mu)^2 \\
&= \sigma^2 + (\mu_M-\mu_F)^2/4.
\end{aligned}$$

Hence, if $\overline{X} = \frac{1}{2n}\sum_{i=1}^{2n} X_i$ (that is, sample mean based on $2n$ observations)

$$\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{2n} + \frac{(\mu_M-\mu_F)^2}{8n} \geq \frac{\sigma^2}{2n} = \mathrm{Var}(\overline{X}_s).$$

□

4.7 Design of Experiments

While surveying a population often provides valuable information, it is very difficult to demonstrate *causation* based on just observing an association between two variables. The reason for this is that *lurking variables* can induce an association between X and Y when there is actually no causal relationship, or when the causal relationship has a completely different nature to what we observe.

Example 4.7.1. Student survey results demonstrate that students who walk to UNSW take a lot less time than students who use public transport. Does this mean that walking to UNSW is faster than using public transport? Should we all walk to UNSW to save time?

Definition 4.7.1. An *observational study* (or survey) is a study in which we observe variables (X, Y) on subjects without manipulating them in any way. An *experiment* is a study in which subjects are manipulating in some way (changing X - the treatment variable) and we observe their response (Y - the response variable).

The purpose of an experiment is to demonstrate that changes in X cause changes in Y. Any good experiment is designed so that the only thing allowed to vary across groups is the treatment variable of interest - so if a significant effect is detected in Y, the only plausible explanation would be that it was caused by X.

4.7.1 Key Considerations in Experimental Design

Any experiment should compare, randomise and repeat:

- **Compare** to demonstrate that changes in X cause changes in Y, we need to compare across suitable designed "treatment" groups (for which we have introduced changes in the value of X). These groups need to be carefully designed so that the *only thing that differs* across groups is the treatment variable X. Double-blinding is often used for this reason, as a "placebo" or "sham treatment".

- **Randomise** the allocation of subjects to treatment groups. This ensures that any differences across groups, apart from those caused by treatment, are governed by chance (which we can then model).
- **Repeat** the *application of the treatment* to the different subjects in each treatment group. It is important that application of treatment is replicated (rather than applied once, "in bulk") in order that we can make inferences about the effect of the treatment in general.

Example 4.7.2. What error has been made in each of the following studies?

1. Consider the Mythbusters' "Is yawning contagious?" episode. The first attempt to answer this question involved sitting nine subjects together in a room for ten minutes, counting the number of yawns, and comparing results to when there was a "seed yawner" in the room with them who pretended to yawn for ten minutes. (Results were inconclusive)
2. Greg was studying how mites affect the growth of cotton plants. He applied a mite treatment to eight (randomly chosen) plants by crumpling up a mite-infested leaf and leaving it at the base of each plant. He applied a no-mite treatment by not putting any leaves at the base of each of eight "control" plants. (Surprisingly, plants in the mite treatment had faster growth)
3. The success of a course on road rules was assessed by using the RTA's practice driving test. Participants were asked to complete the test before the course, then again afterwards, and results were compared. (There was a significant improvement in scores on the test.)

Solution.

1. Replicate the application of treatment: need to separately seed yawn to (groups of) subjects, as in second Mythbusters experiment. Also randomise.
2. Compare to an appropriate control: in this case, that would be crumpled-up leaves which are not mite-infested.
3. Compare to an appropriate control: in this case, people who do the test twice without attending a road rules course. (To account for learner effects.) And randomise.

□

4.7.2 Common Experimental Designs

- **Randomised Comparative Experiment**: Define K treatment groups (each with different levels of the variable X) and randomly assign subjects to each group.
- **Randomised Blocks Design**: If there is some "blocking" variable known to be important to the response variable, break subjects into blocks according to this variable and randomise allocation of subjects to treatment groups separately within each block. This controls for the effects of the blocking variable.
- **Matched Pairs Design**: A common special case of a randomised blocks design, where the blocks come in pairs. Common examples are "before-after" experiments (a pair of measurements is taken on a subject before and after treatment application), which control for subject-to-subject variation, and twins experiments (a pair of identical twins are studied, with one assigned to each of two treatment groups), which control for genetic variation.

Example 4.7.3. Does regularly taking vitamins guard against illness? Consider two experiments on a set of $2n$ subjects:

A. Randomly assign subjects to one of two groups, each consisting of n subjects. The first group are given a vitamin supplement to take daily over the study period (three months), the second are given a placebo tablet (with no vitamins in it), to take daily. Number of illnesses are recorded over the study period.

B. All subjects are given a set of tablets (vitamins or placebo) and asked to take them daily for three months. They are then given a different set of tablets (placebo or vitamin, whichever they didn't have last time) and are asked to take these for three months. Number of illnesses are recorded and compared over the two periods.

Let the number of illnesses in the two treatment groups be Y_v and Y_p. We are interested in the mean difference in number of illnesses between takers of vitamin tablets and takers of a placebo, estimated using the sample mean difference $\overline{Y}_v - \overline{Y}_p$. Assume $\mathrm{Var}(Y_v) = \mathrm{Var}(Y_p) = \sigma^2$.

1. What type of experiment has been done in each of A and B above?
2. Find $\mathrm{Var}(\overline{Y}_v - \overline{Y}_p)$ for experiment A.
3. It is noted in analysis that there is a correlation between number of illnesses in the two study periods (because some people get sick more often than others). Find $\mathrm{Var}(\overline{Y}_v - \overline{Y}_p)$ for experiment B, assuming that the correlation in measurements (and in sample means) across the two study periods is 0.5.
4. Which experiment gives a more efficient estimate of the treatment effect?

Solution.

1. A is a randomised comparative experiment and B is a matched pairs design.
2. Here $\overline{Y}_v$ and $\overline{Y}_p$ are assumed independent (due to the randomised comparative experiment setting) and each is a sample mean of n subjects, hence $\mathrm{Var}(\overline{Y}_v) = \mathrm{Var}(Y_v)/n = \sigma^2/n$ and
$$\mathrm{Var}(\overline{Y}_v - \overline{Y}_p) = \mathrm{Var}(\overline{Y}_v) + \mathrm{Var}(\overline{Y}_p) = \frac{2\sigma^2}{n}.$$
3. Here in scenario B, the Y_p and Y_v cannot be assumed to be independent. For this reason we will include the Cov in the calculation of the variance:
$$\begin{aligned}\mathrm{Var}(Y_v - Y_p) &= \mathrm{Cov}(Y_v - Y_p, Y_v - Y_p)\\ &= \mathrm{Var}(Y_v) + \mathrm{Var}(Y_p) - 2\,\mathrm{Cov}(Y_v, Y_p)\\ &= 2\sigma^2 - 2\,\mathrm{Corr}(Y_v, Y_p)\sqrt{\mathrm{Var}(Y_p)\,\mathrm{Var}(Y_p)}\\ &= 2\sigma^2 - 2\,\mathrm{Corr}(Y_v, Y_p)\sigma^2\\ &= 2\sigma^2(1 - \mathrm{Corr}(Y_v, Y_p)).\end{aligned}$$
Therefore,
$$\mathrm{Var}(\overline{Y}_v - \overline{Y}_p) = \frac{2\sigma^2}{n}(1 - \mathrm{Corr}(Y_v, Y_p)).$$
4. Compared to study A, here we get a reduction in the variance if there is positive correlation $(1 - \mathrm{Corr}(Y_v, Y_p)) < 1$ across the study periods. Note that the variance will increase if there is a negative correlation $(1 - \mathrm{Corr}(Y_v, Y_p)) > 1$ across the study periods.

□

5. Convergence of Random Variables

5.1 Modes of Convergence

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space from Chapter 1, and let $X_1, X_2, \ldots, X$ be the random variables on this space. In other words, each $X : \Omega \to \mathcal{X}$ is a function $X(\omega)$ mapping $\omega \in \Omega$ to $\mathcal{X} \subseteq \mathbb{R}$.

5.1.1 Sure Convergence

Recall that a random variable $X : \Omega \to \mathbb{R}$ is a function. Hence we can consider the pointwise limit:

$$\lim_{n\to\infty} X_n(\omega) = X(\omega).$$

Definition 5.1.1. We say that $X_1, X_2, \ldots$ converges *surely* to X if

$$\lim_{n\to\infty} X_n(\omega) = X(\omega), \quad \text{for all } \omega \in \Omega.$$

Example 5.1.1. If $X_n(\omega) = X(\omega) + \frac{1}{n}$, then X_n converges to X surely.

5.1.2 Almost Sure Convergence

Definition 5.1.2. The sequence of numerical random variables $X_1, X_2, \ldots$ is said to converge *almost surely* to a numerical random variable X, denoted $X_n \xrightarrow{\text{a.s.}} X$, if

$$\mathbb{P}\left(\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right) = 1.$$

The last statement is equivalent to $\mathbb{P}(\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) \neq X(\omega)) = 0$. In contrast to sure convergence here $X_\infty(\omega)$ may not always be equal to $X(\omega)$ for all events $\omega \in \Omega$ but the probability of any event ω for which $X_\infty(\omega) \neq X(\omega)$ is zero. Sure convergence implies almost sure convergence, but the converse is not true.

Definition 5.1.3 (Alternative Definition)**.**

$$X_n \xrightarrow{\text{a.s.}} X$$

if and only if for every $\epsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{k\geq n} |X_k - X| > \epsilon\right) = 0.$$

Proof. Define the event

$$A_n(\epsilon) = \cup_{k\geq n}\{\omega : |X_k - X| > \epsilon\}$$

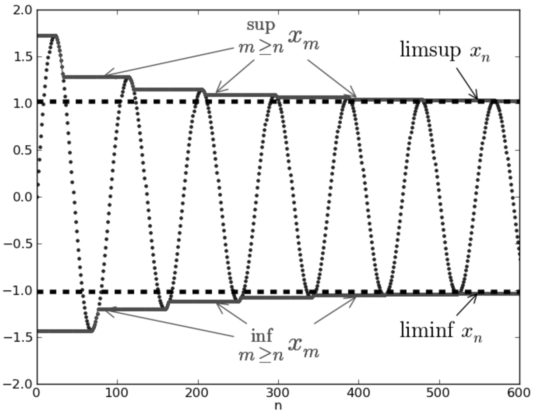
and convince yourself of the following properties:

1. $A_1(\epsilon), A_2(\epsilon), \ldots$ is a decreasing sequence of events (with increasing n we remove more and more sets from a big union) with limit
$$A(\epsilon) = \cap_{n=1}^{\infty} A_n(\epsilon) = \{\omega : |X_\infty - X| > \epsilon\}.$$

2. $A_n(\epsilon) \subseteq A_n(\epsilon')$ for $\epsilon > \epsilon'$. In other words, $A_n(\epsilon)$ is a decreasing function of ϵ.

3. $$\{\omega : \lim_{n\to\infty} X_n \neq X\} = \cup_{m=1}^{\infty} A(1/m).$$
This is because $\cup_{m=1}^{\infty} A(1/m)$ is an increasing sequence of events with limit $A(0)$ corresponding to the ω for which $X_\infty \neq X$.

4. $$A_n(\epsilon) = \{\omega : \sup_{k\geq n} |X_n - X| > \epsilon\},$$
where sup is the smallest upper bound and is same as max when the maximum of the set exists. Note that $\sup_{k\geq n} Y_k$ is a decreasing sequence as seen from the figure below.



Using these properties we can now show the following. Assume
$$\mathbb{P}\left(\omega : \lim_{n\to\infty} X_n(\omega) \neq X(\omega)\right) = 0.$$

then
$$\begin{aligned}
\lim_{n\to\infty} \mathbb{P}\left(\sup_{k\geq n} |X_k - X| > \epsilon\right) &= \lim_{n\to\infty} \mathbb{P}(A_n(\epsilon)) \\
&= \mathbb{P}\left(\lim_{n\to\infty} A_n(\epsilon)\right) \\
&= \mathbb{P}(A(\epsilon)) \\
&\leq \mathbb{P}(A(1/m), m > \epsilon^{-1}) \\
&\leq \mathbb{P}(\cap_{m=1}^{\infty} A(1/m)) \\
&= \mathbb{P}\left(\lim_{n\to\infty} X_n \neq X\right) \\
&= 0.
\end{aligned}$$

Conversely, assume that for every $\epsilon > 0$ we have
$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{k\geq n} |X_k - X| > \epsilon\right) = \mathbb{P}(A(\epsilon)) = 0,$$

then

$$\begin{aligned}\mathbb{P}\left(\lim_{n\to\infty} X_n \neq X\right) &= \mathbb{P}(\cup_{m=1}^{\infty} A(1/m)) \\ &\leq \sum_{m=1}^{\infty} \mathbb{P}(A(1/m)) \\ &= \sum_{m=1}^{\infty} 0 \\ &= 0,\end{aligned}$$

where in the last equation we used $\mathbb{P}(A(1/m)) = \mathbb{P}(A(\epsilon_m)) = 0$ for all $\epsilon_m > 0$. □

5.1.3 Convergence in Probability

Definition 5.1.4. The sequence of random variables $X_1, X_2, \ldots$ *converges in probability* to a random variable X if, for all $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0.$$

This is usually written as

$$X_n \xrightarrow{\mathbb{P}} X.$$

Example 5.1.2. $X_1, X_2, \ldots$ are independent Uniform$(0, \theta)$ variables. Let $Y_n = \max(X_1, \ldots, X_n)$ for $n = 1, 2, \ldots$. Then it can be shown that

$$F_{Y_n}(y) = \begin{cases} \left(\frac{y}{\theta}\right)^n, & 0 < y < \theta \\ 1, & y \geq \theta \end{cases}$$

and

$$f_{Y_n}(y) = \frac{ny^{n-1}}{\theta^n}, \quad 0 < y < \theta.$$

Show that $Y_n \xrightarrow{\mathbb{P}} \theta$.

Solution. For $0 < \epsilon < \theta$,

$$\begin{aligned}\mathbb{P}(|Y_n - \theta| > \epsilon) &= \mathbb{P}(Y_n < \theta - \epsilon) \\ &= \left(\frac{\theta - \epsilon}{\theta}\right)^n \\ &\to 0 \text{ as } n \to \infty.\end{aligned}$$

For $\epsilon > \theta$, $\mathbb{P}(|Y_n - \theta| > \epsilon) = 0$ for all $n \geq 1$, so

$$\lim_{n\to\infty} \mathbb{P}(|Y_n - \theta| > \epsilon) = 0.$$

Therefore, $Y_n \xrightarrow{\mathbb{P}} \theta$. □

In the last example it is easy to show that $Y_n \xrightarrow{\text{a.s.}} \theta$ as well. Consider the fact that for any n

$$\theta \geq Y_{n+1} \geq Y_n,$$

that is, Y_n is monotonically increasing, but bounded from above by θ. Hence,

$$|Y_n - \theta| = \theta - Y_n \geq \theta - \sup_{k\geq n} Y_k = \sup_{k\geq n} |Y_k - \theta|.$$

Hence, the event $\sup_{k\geq n}|Y_k-\theta|>\epsilon$ implies the event $|Y_n-\theta|>\epsilon$ and

$$\mathbb{P}(|Y_n-\theta|>\epsilon)\geq\mathbb{P}(\sup_{k\geq n}|Y_k-\theta|>\epsilon)$$

so that $\mathbb{P}(\sup_{k\geq n}|Y_k-\theta|>\epsilon)$ is squashed to zero from above as $n\to\infty$.

Example 5.1.3. For $n=1,2,\ldots$, $Y_n\sim N(\mu,\sigma_n^2)$ and suppose $\lim_{n\to\infty}\sigma_n=0$. Show that $Y_n\xrightarrow{\mathbb{P}}\mu$.

Solution. For any $\epsilon>0$,

$$\begin{aligned}\mathbb{P}(|Y_n-\mu|>\epsilon)&=\mathbb{P}(Y_n<\mu-\epsilon)+\mathbb{P}(Y_n>\mu+\epsilon)\\&=\int_{-\infty}^{-\epsilon/\sigma_n}\frac{1}{\sqrt{2\pi}}e^{-y^2/2}dy+\int_{\epsilon/\sigma_n}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-y^2/2}dy\\&\to 0\text{ as }n\to\infty.\end{aligned}$$

Thus $Y_n\xrightarrow{\mathbb{P}}\mu$. □

Theorem 5.1.1. *We have*

$$X_n\xrightarrow{\text{a.s.}}X\Rightarrow X_n\xrightarrow{\mathbb{P}}X$$

and

$$X_n\xrightarrow{\text{a.s.}}0\iff\sup_{k\geq n}|X_k|\xrightarrow{\mathbb{P}}0.$$

Proof. We know

$$\sup_{k\geq n}|X_k-X|\geq|X_n-X|\quad\Rightarrow\quad\{\sup_{k\geq n}|X_k-X|>\epsilon\}\supseteq\{|X_n-X|>\epsilon\}.$$

The last in turn implies that $\mathbb{P}(|X_n-X|>\epsilon)$ is squeezed to zero:

$$\mathbb{P}(|X_n-X|>\epsilon)\leq\mathbb{P}(\sup_{k\geq n}|X_k-X|>\epsilon)\to 0,\quad n\to\infty.$$

□

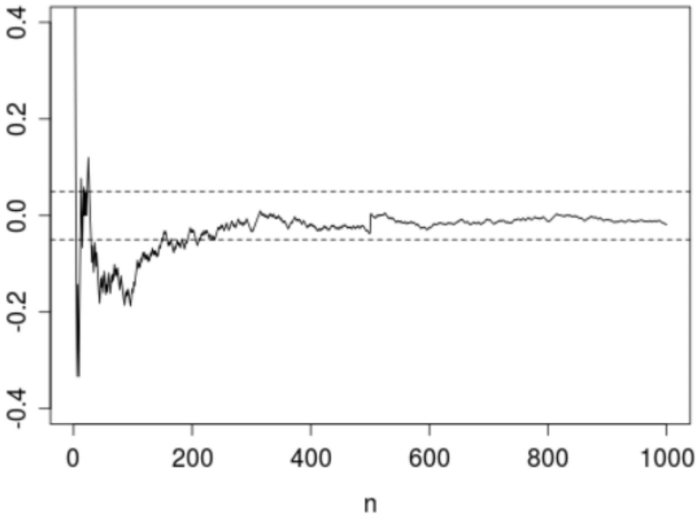
The essential difference between almost sure convergence and convergence in probability is that almost sure convergence is a property of the entire sequence $X_1,X_2,\ldots$ because the distribution of $\sup_{k\geq n}|X_k-X|$ depends on the joint pdf

$$f_{X_n,X_{n+1},\ldots}(x_n,x_{n+1},\ldots),$$

while the convergence in probability is a property of X_n only and hence of the marginal (as opposed to the joint) pdf

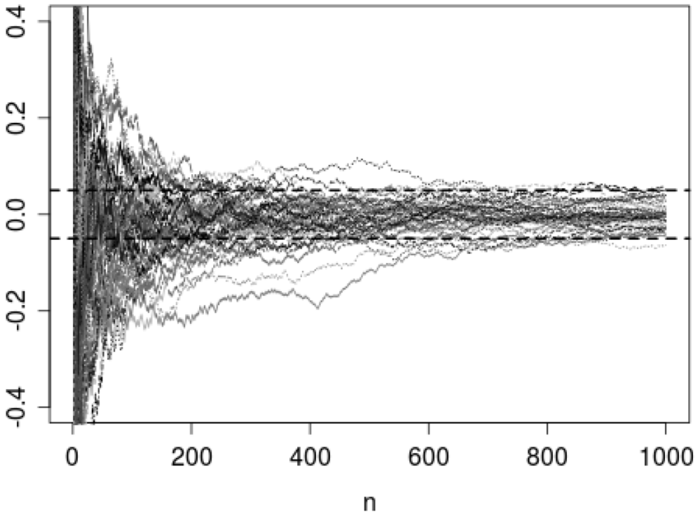
$$f_{X_n}(x_n).$$

The difference can be visualised as follows:



On the above figure, we have depicted $X_n \xrightarrow{\text{a.s.}} 0$. Here as n gets larger and larger the probability of the random path straying far away from the strip (the band $-\epsilon < X < \epsilon$) vanishes as $n \to \infty$.

In contrast, in the following figure,



we have depicted many different realisations of the random path $X_1, X_2, \ldots$. Here $X_n \xrightarrow{\mathbb{P}} 0$ means that the proportion of paths leaving the strip goes to zero as $n \to \infty$. This does not prevent a particular path from straying far away from the bands. We only want the proportion of these rogue paths to get smaller and smaller with increasing n. There is no attempt to control how far a particular path strays from the strip.

Example 5.1.4. Consider the iid sequence $X_1, X_2, \ldots$ with

$$\mathbb{P}(X_n = 1) = \frac{1}{n} = 1 - \mathbb{P}(X_n = 0).$$

Show that $X_n \xrightarrow{\mathbb{P}} 0$. Does $X_n \xrightarrow{\text{a.s.}} 0$?

Solution. The first part is easy as

$$\mathbb{P}(|X_n - 0| > \epsilon) = \mathbb{P}(X_n = 1) = \frac{1}{n} \to 0.$$

For the second part, consider the distribution of

$$Y_n = \sup_{k \geq n} |X_k - 0| = \sup_{k \geq n} X_k.$$

We have

$$
\begin{aligned}
F_{Y_n}(\epsilon) &= \mathbb{P}(\sup_{k\geq n} X_k \leq \epsilon) \\
&= \mathbb{P}(X_n \leq \epsilon, X_{n+1} \leq \epsilon, \ldots) \\
&= \mathbb{P}(X_n \leq \epsilon)\mathbb{P}(X_{n+1} \leq \epsilon)\cdots && \text{(using independence)} \\
&= \lim_{m\to\infty} \prod_{k=n}^{m} \mathbb{P}(X_k \leq \epsilon) \\
&= \lim_{m\to\infty} \prod_{k=n}^{m} \left(1 - \frac{1}{k}\right), && (\epsilon < 1) \\
&= \lim_{m\to\infty} \frac{n-1}{n} \times \frac{n}{n+1} \times \frac{n+1}{n+2} \times \cdots \times \frac{m-1}{m} \\
&= \lim_{m\to\infty} \frac{n-1}{m} \\
&= 0 && \text{(for any } n = 1, 2, \ldots).
\end{aligned}
$$

It follows that for any $0 < \epsilon < 1$ and all $n \geq 1$

$$\mathbb{P}(\sup_{k\geq n} |X_k - 0| > \epsilon) = 1.$$

Thus, it is *not* true that $X_n \xrightarrow{\text{a.s.}} 0$, that is,

$$\lim_{n\to\infty} \mathbb{P}(\sup_{k\geq n} |X_k - 0| > \epsilon) \neq 0.$$

□

5.1.4 Convergence in Distribution

Definition 5.1.5. Let $X_1, X_2, \ldots$ be a sequence of random variables. We say that X_n *converges in distribution* to X if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

for all x where F_X is continuous. A common shorthand is

$$X_n \xrightarrow{\text{d}} X.$$

We say that F_X is the *limiting distribution* of X_n.

This differs subtly (but importantly) from the idea of random variables X_i converging to the random variable X. Convergence in probability is concerned with whether the actual values of the random variables (the x_i) converge. Convergence in distribution, in contrast, is concerned with whether the distributions (the $F_{X_i}(x)$) converge.

Convergence in distribution allows us to make approximate probability statements about X_n, for large n, if we can derive the limiting distribution $F_X(x)$.

Example 5.1.5. Suppose that $\mathbb{P}(X_n = x) = \frac{1}{n}$ of $x = 1, \ldots, n$. Set $Y_n = X_n/n$. Show that $Y_n \xrightarrow{\text{d}} Y \sim U(0, 1)$.

Solution. Here we have for $0 \leq y \leq 1$

$$\begin{aligned} F_{Y_n}(y) &= \mathbb{P}(Y_n \leq y) \\ &= \mathbb{P}(X_n \leq yn) \\ &= \frac{\lfloor yn \rfloor}{n} \\ &\to y \\ &= \mathbb{P}(U \leq y) \\ &= F_U(y). \end{aligned}$$

The last convergence follows from the squeeze principle:

$$y\frac{n-1}{n} \leq \frac{\lfloor yn \rfloor}{n} \leq y\frac{n+1}{n}.$$

□

Theorem 5.1.2. *Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of random variables, each with mgf $M_{X_i}(t)$. Furthermore, suppose that*

$$\lim_{n\to\infty} M_{X_n}(t) = M_X(t).$$

If $M_X(t)$ is a mgf then there is a unique F_X (which gives a random variable X) whose moments are determined by $M_X(t)$ and for all points of continuity $F_X(x)$ we have

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x).$$

Example 5.1.6. Let $X_1, X_2, \ldots$ be independent Bernoulli random variables with success probability $1/2$, representing the outcomes of the fair coin tosses. Define new random variables $Y_1, Y_2, \ldots$ as

$$Y_n = \sum_{k=1}^{n} X_k \left(\frac{1}{2}\right)^k, \quad n = 1, 2, \ldots.$$

Show that $Y_n \xrightarrow{\mathrm{d}} Y \sim U(0,1)$.

Solution. We have

$$\begin{aligned} \mathbb{E}(e^{-uY_n}) &= \prod_{k=1}^{n} \mathbb{E}[e^{-uX_k/2^k}] \\ &= 2^{-n}\prod_{k=1}^{n}(1+e^{-u/2^k}). \end{aligned}$$

But

$$(1-e^{-u/2^n})\prod_{k=1}^{n}(1+e^{-u/2^k}) = 1-e^{-u}, \quad \text{a collapsing product,}$$

and so

$$\begin{aligned} \mathbb{E}[e^{-uY_n}] &= 2^{-n}\frac{1-e^{-u}}{1-e^{-u/2^n}} \\ &= (1-e^{-u})\frac{1/2^n}{1-e^{-u/2^n}}. \end{aligned}$$

It thus follows that

$$\lim_{n\to\infty} \mathbb{E}[e^{-uY_n}] = (1-e^{-u})\lim_{n\to\infty}\frac{1/2^n}{1-e^{-u/2^n}} = \frac{1-e^{-u}}{u}$$

using L'Hopital's rule $\lim_{n\to\infty}\frac{1/2^n}{1-e^{-u/2^n}} = \lim_{n\to\infty}\frac{\frac{d}{dn}(1/2^n)}{ue^{-u/2^n}\frac{d}{dn}(1/2)^n}$. We recognise $\frac{1-e^{-u}}{u}$ as the mgf of a random variable $U \sim U(0,1)$, evaluated at $-u$. Hence, $Y_n \xrightarrow{\mathrm{d}} Y \sim U(0,1)$. □

5.1.5 Convergence in L^p-Norm

Definition 5.1.6. The sequence of numerical random variables $X_1, X_2, \ldots$ is said to converge in *L^p-norm* to a numerical random variable X, denoted $X_n \xrightarrow{L^p} X$, if

$$\lim_{n\to\infty} \mathbb{E}|X_n - X|^p = 0.$$

Convergence in L^2-norm is often called *mean square convergence.*

Example 5.1.7. Suppose $X_1, X_2, \ldots$ are independent, each with mean μ and variance $0 < \sigma^2 < \infty$. Show that $\overline{X}_n \xrightarrow{L^2} \mu$.

Solution. We have

$$\mathbb{E}[(\overline{X} - \mu)^2] = \operatorname{Var}(\overline{X}) = \frac{\sigma^2}{n} \to 0.$$

□

The next example shows that we can have a sequence converging in mean, but not almost surely. Thus, the two types of convergences are quite distinct.

Example 5.1.8. Recall the example in which we have the iid sequence $X_1, X_2, \ldots$ with

$$\mathbb{P}(X_n = 1) = \frac{1}{n} = 1 - \mathbb{P}(X_n = 0).$$

We showed that $X_n \xrightarrow{\text{a.s.}} 0$ is not true. However, we do have $X_n \xrightarrow{L^1} 0$:

$$\mathbb{E}|X_n - 0| = 1 \times \frac{1}{n} + 0 \times (1 - 1/n) = \frac{1}{n} \to 0.$$

5.1.6 Complete Convergence

Definition 5.1.7. A sequence of random variables $X_1, X_2, \ldots$ is said to converge *completely* to X, denoted

$$X_n \xrightarrow{\text{cpl.}} X,$$

if

$$\sum_{k=1}^{\infty} \mathbb{P}(|X_k - X| > \epsilon) < \infty \quad \text{for all } \epsilon > 0.$$

Example 5.1.9. Suppose $\mathbb{P}(X_n = n^5) = 1/n^2$ and $\mathbb{P}(X_n = 0) = 1 - 1/n^2$. Then, we have

$$\begin{aligned}\sum_n \mathbb{P}(|X_n - 0| > \epsilon) &= \sum_n \mathbb{P}(X_n = n^5)\\ &= \sum_{n=1}^{\infty} \frac{1}{n^2}\\ &= \frac{\pi^2}{6} < \infty.\end{aligned}$$

Hence, by definition $X_n \xrightarrow{\text{cpl.}} 0$.

Note that if $\sum_n \mathbb{P}(|X_n - X| > \epsilon) < \infty$, then we know that $\mathbb{P}(|X_n - X| > \epsilon) \to 0$ (terms in convergent series must approach zero). Thus, complete convergence implies convergence in probability.

Theorem 5.1.3.

$$X_n \xrightarrow{\text{cpl.}} X \Rightarrow X_n \xrightarrow{\text{a.s.}} X.$$

Proof. Recall that $\mathbb{P}(\cup_k A_k) \leq \sum_k \mathbb{P}(A_k)$ is one of the defining properties of $\mathbb{P}()$. Hence, we can bound the criterion for almost sure convergence as follows:

$$\begin{aligned}
\mathbb{P}\left(\sup_{k\geq n}|X_k - X| > \epsilon\right) &= \mathbb{P}(\cup_{k\geq n}\{|X_k - X| > \epsilon\}) \\
&\leq \sum_{k\geq n}\mathbb{P}(\{|X_k - X| > \epsilon\}) && \text{(by union property of } \mathbb{P}()) \\
&\leq \underbrace{\sum_{k=1}^{\infty}\mathbb{P}(|X_k - X| > \epsilon)}_{=c<\infty \text{ from } X_n \xrightarrow{\text{cpl.}} X} - \sum_{k=1}^{n-1}\mathbb{P}(|X_k - X| > \epsilon) \\
&\leq c - \sum_{k=1}^{n-1}\mathbb{P}(|X_k - X| > \epsilon) \\
&\to c - c = 0, \text{ as } n \to \infty.
\end{aligned}$$

Hence, by definition $X_n \xrightarrow{\text{a.s.}} X$. □

5.2 Convergence Relationships

The different types of convergence form a big family with close ties to each other; the relationships produce the following diagram:

$$\begin{tikzcd}
\boxed{X_n \xrightarrow{\text{cpl.}} X} \arrow[r, Rightarrow] & \boxed{X_n \xrightarrow{\text{a.s.}} X} \arrow[d, Rightarrow] & \\
 & \boxed{X_n \xrightarrow{\mathbb{P}} X} \arrow[r, Rightarrow] & \boxed{X_n \xrightarrow{d} X} \\
\boxed{X_n \xrightarrow{L^p} X} \arrow[r, Rightarrow, "p\geq q"] & \boxed{X_n \xrightarrow{L^q} X} \arrow[u, Rightarrow] &
\end{tikzcd}$$

5.2.1 Convergence in Probability and Distribution

First note that convergence in distribution does not imply convergence in probability (unless additional assumptions are imposed).

Example 5.2.1. Suppose $X_n = 1 - X$, where $X \sim U(0, 1)$. Then,

$$\mathbb{P}(X_n \leq x) = \mathbb{P}(1 - X \leq x) = \mathbb{P}(X \geq 1 - x) = 1 - (1 - x) = x$$

so that $F_{X_n}(x)$ is the cdf of the uniform distribution for all n. Trivially, we have

$$X_n \xrightarrow{\text{d}} X \sim U(0, 1).$$

However,

$$\begin{aligned}
\mathbb{P}(|X_n - X| > \epsilon) &= \mathbb{P}(|1 - 2X| > \epsilon) \\
&= 1 - \left(\frac{\epsilon + 1}{2} - \frac{1 - \epsilon}{2}\right) \\
&= 1 - \epsilon \not\to 0.
\end{aligned}$$

Hence, $X_n \overset{\mathbb{P}}{\not\to} X$.

Before we proceed with the theorem, note that the probability of two events occurring together (in "conjunction") is always less than or equal to the probability of either one occurring alone. Formally,

$$\mathbb{P}(A \cap B) \leq \mathbb{P}(A).$$

Theorem 5.2.1.

$$X_n \xrightarrow{\mathbb{P}} X \Rightarrow X_n \xrightarrow{\mathrm{d}} X.$$

Proof. First note that

$$\begin{aligned} F_{X_n}(x) &= \mathbb{P}(X_n \leq x) \\ &= \mathbb{P}(X_n \leq x, |X_n - X| > \epsilon) + \mathbb{P}(X_n \leq x, |X_n - X| \leq \epsilon) \\ &\leq \mathbb{P}(|X_n - X| > \epsilon) + \mathbb{P}(X_n \leq x, |X - X_n| \leq \epsilon), \qquad \text{(by conjunction property)} \end{aligned}$$

and so,

$$\begin{aligned} \mathbb{P}(X_n \leq x, |X_n - X| \leq \epsilon) &\leq \mathbb{P}(X_n \leq x, X \leq X_n + \epsilon) \\ &\leq \mathbb{P}(X_n \leq x, X \leq x + \epsilon) \\ &\leq \mathbb{P}(X \leq x + \epsilon). \end{aligned}$$

Therefore,

$$F_{X_n}(x) \leq \mathbb{P}(|X_n - X| > \epsilon) + \mathbb{P}(X \leq x + \epsilon).$$

Now, in the arguments above we can switch the roles of X_n and X (there is a symmetry to deduce the analogous result:

$$F_X(x) \leq \mathbb{P}(|X - X_n| > \epsilon) + \underbrace{\mathbb{P}(X_n \leq x + \epsilon)}_{F_{X_n}(x+\epsilon)}.$$

Therefore, making the switch $x \to x - \epsilon$ gives

$$F_X(x - \epsilon) \leq \mathbb{P}(|X - X_n| > \epsilon) + F_{X_n}(x)$$

and putting it all together

$$F_X(x - \epsilon) - \mathbb{P}(|X - X_n| > \epsilon) \leq F_{X_n}(x) \leq \mathbb{P}(|X_n - X| > \epsilon) + \mathbb{P}(X \leq x + \epsilon).$$

Taking limits on both sides yields for any $\epsilon > 0$:

$$F_X(x - \epsilon) \leq \lim_{n\to\infty} F_{X_n}(x) \leq \mathbb{P}(X \leq x + \epsilon) = F_X(x + \epsilon).$$

Now if x is a point of continuity, then

$$\lim_{\epsilon\to 0} F_X(x \pm \epsilon) = F_X(x).$$

Hence, by taking the limit on both sides as $\epsilon \to 0$ we deduce by the squeeze principle that

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

at points x where $F_X(x)$ is continuous. The last agrees with the definition of convergence in probability. □

5.2.2 Convergence in L^1 Mean and Convergence in Probability

First note that convergence in probability does not imply convergence in L^1 mean (unless additional assumptions are imposed).

Example 5.2.2. Suppose $\mathbb{P}(X_n = n^5) = 1/n^2$ and $\mathbb{P}(X_n = 0) = 1 - 1/n^2$. Show that $X_n \xrightarrow{\mathbb{P}} 0$. Does X_n converge to X in L^1 mean?

Solution. We have

$$\begin{aligned}\mathbb{P}(|X_n - 0| > \epsilon) &= \mathbb{P}(X_n > \epsilon)\\ &= \mathbb{P}(X_n = n^5)\\ &= 1/n^2 \to 0.\end{aligned}$$

Hence, by definition $X_n \xrightarrow{\mathbb{P}} 0$. However,

$$\begin{aligned}\mathbb{E}|X_n - 0| &= n^5 \times \frac{1}{n^2} + 0 \times \mathbb{P}(X_n = 0)\\ &= n^3 \not\to 0.\end{aligned}$$

So the L^1 mean blows up. □

Theorem 5.2.2.

$$X_n \xrightarrow{L^1} X \Rightarrow X_n \xrightarrow{\mathbb{P}} X.$$

Proof. We have

$$\begin{aligned}\mathbb{P}(|X_n - X| > \epsilon) &\le \frac{\mathbb{E}|X_n - X|}{\epsilon} && \text{(by Chebyshev's inequality)}\\ &\le \text{constant} \times \mathbb{E}|X_n - X| \to 0.\end{aligned}$$

Hence, $\mathbb{P}(|X_n - X| > \epsilon)$ is forced to converge to 0. □

The result that $\mathbb{E}|X|^u \le (\mathbb{E}|X|^s)^{u/s}$ for $s \ge u > 0$ implies the following theorem:

Theorem 5.2.3. *For any $p \ge q \ge 1$, we have*

$$X_n \xrightarrow{L^p} X \Rightarrow X_n \xrightarrow{L^q} X.$$

Example 5.2.3. Suppose $X_1, X_2, \ldots$ are independent, each with mean μ and variance $0 < \sigma^2 < \infty$. Then, we know

$$\overline{X}_n \xrightarrow{L^2} \mu,$$

which implies

$$\overline{X}_n \xrightarrow{L^2} \mu \Rightarrow \overline{X}_n \xrightarrow{L^1} \mu \Rightarrow \overline{X}_n \xrightarrow{\mathbb{P}} \mu.$$

In general, if $p > q \ge 1$, then $X_n \xrightarrow{L^q} X$ does not imply $X_n \xrightarrow{L^p} X$,

Example 5.2.4. Assume $p > q \ge 1$, Let $\mathbb{P}(X_n = n) = \frac{1}{n^{(p+q)/2}} = 1 - \mathbb{P}(X_n = 0)$, then

$$\mathbb{E}|X_n - 0|^q = n^q/n^{(p+q)/2} = 1/n^{(p-q)/2}$$

and we will get convergence to 0 for $p > q$, but at the same time

$$\mathbb{E}|X_n - 0|^p = 1/n^{(q-p)/2} \to \infty, \quad \text{as } n \to \infty.$$

As a result, $X_n \xrightarrow{L^q} X$, but $X_n \not\xrightarrow{L^p} X$.

5.2.3 Almost Sure Convergence and in L^1 Mean

In general, these two types of convergences are quite distinct.

Example 5.2.5. Consider again the example with $\mathbb{P}(X_n = n^5) = 1/n^2$ and $\mathbb{P}(X_n = 0) = 1 - 1/n^2$. We showed that $X_n \xrightarrow{L^1} 0$ is false, because $X_n = n^5$ can take arbitrary large values as $n \to \infty$ and this forces the expectation to blow up. Show that $X_n \xrightarrow{\text{a.s.}} 0$.

Solution. This is easy, since

$$\sum_n \mathbb{P}(|X_n - 0| > \epsilon) = \sum_n \frac{1}{n^2} = \frac{\pi^2}{6} < \infty$$

implies that $X_n \xrightarrow{\text{cpl.}} 0$, which we know in turn implies $X_n \xrightarrow{\text{a.s.}} 0$. □

5.3 Converse Results on Modes of Convergence

We explore the conditions under which we can reverse the direction of the ⇒ in the diagram.

5.3.1 Convergence in Distribution to a Constant

Theorem 5.3.1. *If c is a constant, then*

$$X_n \overset{\text{d}}{\to} c \Rightarrow X_n \overset{\mathbb{P}}{\to} c.$$

Proof. We are given that

$$\lim_{n\to\infty} F_{X_n}(x) = \lim_{n\to\infty} \mathbb{P}(X_n \le x) = \begin{cases} 1 & x \ge c \\ 0 & x < c. \end{cases}$$

We now try to bound and squash to zero the criterion for convergence in probability:

$$\begin{aligned} \mathbb{P}(|X_n - c| > \epsilon) &= \mathbb{P}(X_n > c + \epsilon) + \mathbb{P}(X_n < c - \epsilon) \\ &\le 1 - F_{X_n}(c + \epsilon) + F_{X_n}(c - \epsilon) \\ &\to 1 - 1 + 0 = 0, \text{ as } n \to \infty. \end{aligned}$$

Since this is true for any small $\epsilon > 0$, we have $X_n \overset{\mathbb{P}}{\to} c$. □

5.3.2 Convergence in Probability

Theorem 5.3.2. *If $\mathbb{P}(|X_k| \le c) = 1$ for all k, then for any $p \ge 1$*

$$X_n \overset{\mathbb{P}}{\to} X \Rightarrow X_n \xrightarrow{L^p} X.$$

Proof. First note that since $X_n \overset{\text{d}}{\to} X$, we also have that X is bounded in the sense

$$\mathbb{P}(|X| \le c) = \lim_{n\to\infty} \mathbb{P}(|X_n| \le c) = 1.$$

We try to bound the criterion for convergence in mean by smuggling an indicator into the expectation:

$$\begin{aligned}
\mathbb{E}|X_n - X|^p &= \mathbb{E}[|X_n - X|^p I_{\{|X_n - X| > \epsilon\}}] + \mathbb{E}[|X_n - X|^p I_{\{|X_n - X| < \epsilon\}}] \\
&\leq \mathbb{E}[|X_n - X|^p I_{\{|X_n - X| > \epsilon\}}] + \mathbb{E}|\epsilon^p I_{\{|X_n - X| < \epsilon\}}] \\
&\leq \mathbb{E}[(|X_n| + |X|)^p I_{\{|X_n - X| > \epsilon\}}] + \epsilon^p \mathbb{P}(|X_n - X| < \epsilon) \\
&\leq (2c)^p \mathbb{E}[I_{\{|X_n - X| > \epsilon\}}] + \epsilon^p \mathbb{P}(|X_n - X| < \epsilon) \\
&\leq (2c)^p \mathbb{P}(|X_n - X| > \epsilon) + \epsilon^p \\
&\to 0 + \epsilon^p, \text{ as } n \to \infty.
\end{aligned}$$

Since this is true for any $\epsilon > 0$, no matter how small, we conclude that $\mathbb{E}|X_n - X|^p \to 0$ as $\epsilon \to 0$ and $n \to \infty$. □

Finally, of interest is when we can go from $X_n \xrightarrow{\mathbb{P}} X$ to $X_n \xrightarrow{\text{a.s.}} X$. In general, this is a difficult problem, but we can easily show that if $X_n \xrightarrow{\mathbb{P}} X$, then there is a subsequence of $X_1, X_2, X_3, \ldots$ call it $X_{k_1}, X_{k_2}, X_{k_3}, \ldots$ which converges almost surely to X.

Theorem 5.3.3. *Suppose that $X_n \xrightarrow{\mathbb{P}} X$, that is, for every $\epsilon > 0$ and $\delta > 0$, we can find a large enough n such that*

$$\mathbb{P}(|X_n - X| > \epsilon) < \delta.$$

Then

$$X_{k_n} \xrightarrow{\text{a.s.}} X, \quad n \to \infty,$$

where $k_1 < k_2 < k_3 < \ldots$ is selected such that

$$\mathbb{P}(|X_{k_j} - X| > \epsilon) < \frac{1}{j^2}.$$

Proof. Since

$$\sum_j \mathbb{P}(|X_{k_j} - X| > \epsilon) < \sum_j \frac{1}{j^2} = \frac{\pi^2}{6} < \infty,$$

then $X_{k_n} \xrightarrow{\text{cpl.}} X$, which in turn implies $X_{k_n} \xrightarrow{\text{a.s.}} X$. □

5.3.3 Almost Sure Convergence

Theorem 5.3.4. *If $X_1, X_2, \ldots$ are independent, then*

$$X_n \xrightarrow{\text{a.s.}} X \Rightarrow X_n \xrightarrow{\text{cpl.}} X.$$

Proof. First verify that the global minimum of the function $e^x - x - 1$ is 0 and hence $e^x \geq x + 1$ or equivalently

$$e^{-x} \geq 1 - x.$$

We now argue by contradiction. Assume that $X_n \xrightarrow{\text{a.s.}} X$ and that we do not have complete convergence, that is, assume

$$\sum_n \mathbb{P}(|X_n - X| > \epsilon) = \infty.$$

Then, under this assumption we show that

$$\mathbb{P}(\sup_{k \geq n} |X_n - X| > \epsilon) = \mathbb{P}(\cup_{k \geq n} \{|X_n - X| > \epsilon\}) = 1$$

instead of the expected 0, contradicting $X_n \xrightarrow{\text{a.s.}} X$. We have

$$\begin{aligned}
1 - \mathbb{P}(\cup_{k\geq n}\{|X_n - X| > \epsilon\}) &= \mathbb{P}(\cap_{k\geq n}\{|X_n - X| \leq \epsilon\}) && \text{(de Morgan's Law)} \\
&= \prod_{k\geq n} \mathbb{P}(|X_n - X| \leq \epsilon) && \text{(independence)} \\
&= \prod_{k\geq n} (1 - \mathbb{P}(|X_n - X| > \epsilon)) \\
&\leq \prod_{k\geq n} e^{-\mathbb{P}(|X_n - X|>\epsilon)}, && 1 - x \leq e^{-x} \\
&\leq e^{-\sum_{k\geq n}\mathbb{P}(|X_n - X|>\epsilon)} \\
&= e^{-\infty} \\
&= 0.
\end{aligned}$$

Hence, it is not true that $\mathbb{P}(\sup_{k\geq n}|X_n - X| > \epsilon) \to 0$ contradicting the assumption $X_n \xrightarrow{\text{a.s.}} X$. The only other logical possibility is that

$$\sum_n \mathbb{P}(|X_n - X| > \epsilon) < \infty,$$

that is, $X_n \xrightarrow{\text{cpl.}} X$. □

In summary, we have established the following diagram with reverse arrow directions:

$$\begin{tikzcd}
\boxed{X_n \xrightarrow{\text{cpl.}} X} & \boxed{X_n \xrightarrow{\text{a.s.}} X} \arrow[l, Rightarrow, "\text{indep.}"'] & \\
{} \arrow[u, Rightarrow] & \boxed{X_n \xrightarrow{\mathbb{P}} X} \arrow[u, Rightarrow, "\text{subseq.}"'] \arrow[l, Rightarrow, "\text{subseq.}"'] \arrow[d, Rightarrow, "|X|<c"] & \boxed{X_n \xrightarrow{d} X} \arrow[l, Rightarrow, "X\text{ const.}"'] \\
\boxed{X_n \xrightarrow{L^p} X} & \boxed{X_n \xrightarrow{L^q} X} \arrow[l, Rightarrow, "|X|<c"'] &
\end{tikzcd}$$

5.4 Weak Law of Large Numbers

A fundamental idea in Statistics is to use a sample to represent an unknown distribution, and to use sample characteristics to estimate the corresponding distributional characteristics. The soundness of this idea is verified by the Weak Law of Large Numbers.

Theorem 5.4.1 (Weak Law of Large Numbers)**.** *Suppose $X_1, X_2, \ldots$ are independent, each with mean μ and variance $0 < \sigma^2 < \infty$. If*

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i, \quad \textit{then } \overline{X}_n \xrightarrow{\mathbb{P}} \mu;$$

that is, for all $\epsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(|\overline{X}_n - \mu| > \epsilon) = 0$.

Theorem 5.4.2 (Slutsky's Theorem)**.** *Let $X_1, X_2, \ldots$ be a sequence of random variables that converge in distribution to X, that is,*

$$X_n \xrightarrow{\text{d}} X.$$

Let $Y_1, Y_2, \ldots$ be another sequence of random variables that converges in probability to a constant *c, that is,*

$$Y_n \xrightarrow{\mathbb{P}} c.$$

Then,

(i) $X_n + Y_n \xrightarrow{\text{d}} X + c$

(ii) $X_n Y_n \xrightarrow{\text{d}} cX$.

Example 5.4.1. Suppose that $X_1, X_2, \ldots$ converges in distribution to $X \sim N(0,1)$, i.e. $X_n \xrightarrow{\text{d}} N(0,1)$, and suppose that $nY_n \sim \text{Bin}(n, \frac{1}{2})$. What are the limiting distributions of $X_n + Y_n$ and X_nY_n?

Solution. First by the Weak Law of Large numbers $Y_n \xrightarrow{\mathbb{P}} \frac{1}{2}$. Therefore, from Slutsky's Theorem,

$$X_n + Y_n \xrightarrow{\text{d}} N(1/2, 1), \quad \text{and} \quad X_nY_n \xrightarrow{\text{d}} N(0, 1/4).$$

□

5.5 Strong Law of Large Numbers

The Weak Law corresponds to convergence in probability, while the Strong Law corresponds to almost sure convergence. Since almost sure convergence implies convergence in probability, it is in this sense that we talk about a Strong and Weak Law. The difference between the Strong and Weak Law is the same as that between almost sure convergence and convergence in probability.

Theorem 5.5.1 (Strong Law of Large Numbers)**.** *Let $X_1, X_2, \ldots$ be independent with common mean $\mathbb{E}[X] = \mu$ and variance* $\text{Var}(X) = \sigma^2 < \infty$*, then*

$$\overline{X}_n \xrightarrow{\text{a.s.}} \mu.$$

Proof. We give only a proof for the case $X_k \geq 0$ for all k. If from the sequence

$$\{\overline{X}_1, \overline{X}_2, \overline{X}_3, \ldots\}$$

we pick up the subsequence

$$\{\overline{X}_1, \overline{X}_4, \overline{X}_9, \overline{X}_{16}, \ldots\} \equiv \{\overline{X}_{j^2}\}$$

for $j = 1, 2, 3, \ldots$, then

$$\mathbb{P}(|\overline{X}_{j^2} - \mu| > \epsilon) \leq \frac{\text{Var}(\overline{X}_{j^2})}{\epsilon^2} = \frac{\sigma^2}{j^2\epsilon^2} < \infty.$$

Hence,

$$\sum_{j=1}^{\infty} \mathbb{P}(|\overline{X}_{j^2} - \mu| > \epsilon) = \frac{\pi^2}{6}\frac{\text{Var}(X)}{\epsilon^2} < \infty$$

and we can conclude that the sub-sequence of $\{\overline{X}_n\}$, namely, $\{\overline{X}_{j^2}, j = 1, 2, \ldots\}$ converges almost surely to μ. Next, it is clear that for any arbitrary n, we can always find a k so that $k^2 \leq n \leq (k+1)^2$. For example, taking $k = \lfloor\sqrt{n}\rfloor$ works. Also note that $k \to \infty$ as $n \to \infty$.
Finally, for $k^2 \leq n \leq (k+1)^2$ it can be shown (see remark below) that

$$\frac{k^2}{(k+1)^2}\overline{X}_{k^2} \leq \overline{X}_n \leq \overline{X}_{(k+1)^2}\frac{(k+1)^2}{k^2}.$$

Both $\overline{X}_{k^2}$ and $\overline{X}_{(k+1)^2}$ converge almost surely to μ as n (and hence k) go to infinity. Therefore, we conclude that $\overline{X}_n \xrightarrow{\text{a.s.}} \mu$. □

Remark 5.5.1 (Justification for Inequality)**.** Observe that for $X_k \geq 0$

$$\begin{aligned} k^2\overline{X}_{k^2} = X_1 + \cdots + X_{k^2} &\leq X_1 + \cdots + X_n && (n \geq k^2) \\ &\leq n\overline{X}_n \\ &\leq (k+1)^2\overline{X}_n && ((k+1)^2 \geq n). \end{aligned}$$

Rearranging the last gives $\frac{k^2}{(k+1)^2}\overline{X}_{k^2} \leq \overline{X}_n$. Similarly,

$$\begin{aligned} k^2\overline{X}_n &\leq n\overline{X}_n && (k^2 \leq n) \\ &= X_1 + \cdots + X_n \\ &\leq X_1 + \cdots + X_{(k+1)^2} && (n \leq (k+1)^2) \\ &= (k+1)^2\overline{X}_{(k+1)^2}. \end{aligned}$$

Hence, $\overline{X}_n \leq \overline{X}_{(k+1)^2}\frac{(k+1)^2}{k^2}$.

5.6 Central Limit Theorem

For a general sample $X_1, \ldots, X_n$ it is often of interest to make probability statements about the sample mean $\overline{X}$. The following theorem gives a way:

Theorem 5.6.1 (Central Limit Theorem)**.** *Suppose $X_1, X_2, \ldots$ are independent and identically distributed random variables with common mean $\mu = \mathbb{E}(X_i)$ and common variance $\sigma^2 = \mathrm{Var}(X_i) < \infty$. For each $n \geq 1$ let $\overline{X}_n = \frac{\sum_{i=1}^n X_i}{n}$. Then*

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\text{d}} Z$$

where $Z \sim N(0,1)$. It is common to write

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\text{d}} N(0,1).$$

Proof. Our proof will use the mgf approach, which requires that the mgf function exists. This requirement implies that all moments $\mathbb{E}[X^k] < \infty$, $k = 1, 2, \ldots$ are finite, which is much stronger than the assumption in the theorem of a finite second moment only. Nevertheless, we proceed with this proof, because it is one of the easiest.

We will denote the standardised moments

$$\kappa_k = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^k\right], \quad k = 1, 2, \ldots.$$

κ_3 is called the skewness of the distribution of X, and κ_4 is called the kurtosis of the distribution of X. The skewness parameter indicates how symmetric the distribution of X is and the kurtosis indicates how fast the tails of the density decay to zero. Without loss of generality we can assume $\mu = 0$ and $\sigma = 1$. Thus, in our case $\kappa_k = \mathbb{E}[X^k]$. Then, if

$$m_X(t) = \mathbb{E}[e^{tX}]$$

is the mgf of X, it follows from the iid assumption that

$$m_{\sqrt{n}\overline{X}_n}(t) = \left(m_X\left(\frac{t}{\sqrt{n}}\right)\right)^n.$$

Hence,

$$\zeta(t) = \ln m_{\sqrt{n}\overline{X}_n}(t) = n \ln m_X\left(\frac{t}{\sqrt{n}}\right).$$

Now, consider the Taylor expansions

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots, \quad |x| < 1$$

and

$$\begin{aligned} m_X\left(\frac{t}{\sqrt{n}}\right) &= 1 + \mathbb{E}[X]\frac{t}{\sqrt{n}} + \mathbb{E}[X^2]\frac{t^2}{2!n} + \mathbb{E}[X^3]\frac{t^3}{3!n^{3/2}} + \cdots \\ &= 1 + \frac{t^2}{2n} + \mathbb{E}[X^3]\frac{t^3}{6n^{3/2}} + \mathbb{E}[X^4]\frac{t^4}{4!n^2} + \cdots \\ &= 1 + \underbrace{\frac{t^2}{2n} + \kappa_3\frac{t^3}{6n^{3/2}} + \kappa_4\frac{t^4}{4!n^2} + \cdots,}_{x} \end{aligned}$$

where we choose t such that $|t/\sqrt{n}| < \epsilon$ with $\epsilon > 0$ small enough that

$$|x| \leq \sum_{k=2}^{\infty} |\kappa_k| \frac{\epsilon^k}{k!} < 1.$$

Now, apply both Taylor expansions to $\zeta(t)$ in a nested fashion to obtain:

$$\begin{aligned} \frac{\zeta(t)}{n} &= \left(\frac{t^2}{2n} + \kappa_3\frac{t^3}{6n^{3/2}} + \kappa_4\frac{t^4}{4!n^2} + \cdots\right) - \\ &\quad - \frac{1}{2}\left(\frac{t^2}{2n} + \kappa_3\frac{t^3}{6n^{3/2}} + \kappa_4\frac{t^4}{4!n^2} + \cdots\right)^2 + \\ &\quad + \frac{1}{3}\left(\frac{t^2}{2n} + \kappa_3\frac{t^3}{6n^{3/2}} + \kappa_4\frac{t^4}{4!n^2} + \cdots\right)^3 - \cdots \\ \text{(collect like powers of } n) &= \frac{t^2}{2n} + \frac{\kappa_3 t^3}{6n^{3/2}} + \frac{t^4}{n^2}\left(\frac{\kappa_4}{4!} - \frac{1}{8}\right) + \cdots. \end{aligned}$$

It follows that

$$\zeta(t) = \frac{t^2}{2} + \frac{\kappa_3 t^3}{6n^{1/2}} + \frac{t^4}{n}\left(\frac{\kappa_4}{4!} - \frac{1}{8}\right) + \cdots$$

or alternatively

$$m_{\sqrt{n}\overline{X}_n}(t) = \exp\left(\frac{t^2}{2} + \frac{\kappa_3 t^3}{6n^{1/2}} + \frac{t^4}{n}\left(\frac{\kappa_4}{4!} - \frac{1}{8}\right) + \cdots\right) \to \exp\left(\frac{t^2}{2}\right), \quad n \to \infty.$$

Hence, $\sqrt{n}X_n \xrightarrow{\mathrm{d}} N(0,1)$. Note that we have used the fact that

$$\left|\frac{\kappa_3 t^3}{6n^{1/2}} + \frac{t^4}{n}\left(\frac{\kappa_4}{4!} - \frac{1}{8}\right) + \cdots\right| \leq \text{const.} \sum_{k=1}^{\infty} \frac{1}{n^{k/2}} = \text{const.}\left(\frac{1}{1-n^{-1/2}} - 1\right).$$

□

Note that

$$\mathbb{E}(\overline{X}_n) = \mu, \quad \text{and} \quad \mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n}.$$

So the Central Limit Theorem states that the limiting distribution of any standardised average of independent random variables is the standard Normal or $N(0,1)$ distribution. Also note that we made no assumptions about the common distribution of the X_i. This is an important aspect of this result which makes it particularly useful in practice.
The Central Limit Theorem stated above provides the limiting distribution of

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}.$$

However, sometimes probabilities involving related quantities such as the sum $\sum_{i=1}^n X_i$ are required. Since $\sum_{i=1}^n X_i = n\overline{X}$ the Central Limit Theorem also applies to the sum of a sequence of random variables.

Theorem 5.6.2. *Suppose $X_1, X_2, \ldots$ are independent and identically distributed random variables with common mean $\mu = \mathbb{E}[X_i]$ and common variance $\sigma^2 = \mathrm{Var}(X_i) < \infty$. Then the Central Limit Theorem may also be stated in the following alternative forms:*

(i) $\sqrt{n}(\overline{X} - \mu) \xrightarrow{\mathrm{d}} N(0, \sigma^2)$.

(ii) $\frac{\sum_i X_i - n\mu}{\sigma\sqrt{n}} \xrightarrow{\mathrm{d}} N(0,1)$.

(iii) $\frac{\sum_i X_i - n\mu}{\sqrt{n}} \xrightarrow{\mathrm{d}} N(0, \sigma^2)$.

5.7 Applications of the Central Limit Theorem

The CLT tells us that the sum of many independent small random variables has an approximate normal distribution. Therefore it is plausible that any real-life random variable, formed by the combined effect of many small independent random influences, will be approximately normally distributed. Thus the CLT provides an explanation for the widespread empirical occurrence of the normal distribution.

5.7.1 Probability Calculations About a Sample Mean

Suppose we are interested in a random variable X. We take a measurement of this variable on each of n randomly selected subjects, giving us n independently and identically distributed random variables $X_1, X_2, \ldots, X_n$. $\overline{X}$ is the average of X from this sample. By the CLT, we know that the average of a sample from any random variable is approximately normally distributed. So if we know μ and σ for this random variable, we can calculate any probability we like about averages of random variables from this unknown distribution.

Example 5.7.1. It is known that Australians have an average weight of about 68kg, and the variance of this quantity is about 256. We randomly choose 10 Australians. What is an approximate distribution for the average weight of these people? What is the chance that their average weight exceeds 80kg?

Solution. From the CLT,

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{\mathrm{d}} N(0,1)$$

so we can say that

$$\overline{X} \mathrel{\dot\sim} N\left(\mu, \frac{\sigma^2}{n}\right) = N\left(68, \frac{256}{10}\right) = N(68, 25.6),$$

(where $\stackrel{\cdot}{\sim}$ means "approximately distributed"). Thus,

$$\begin{aligned}\mathbb{P}(\overline{X} > 80) &= \mathbb{P}\left(X > \frac{80-68}{\sqrt{25.6}}\right)\\ &\approx \mathbb{P}(X > 2.37)\\ &\approx 0.0089.\end{aligned}$$

□

5.7.2 Normal Approximation to the Binomial Distribution

Theorem 5.7.1 (CLT for Binomial Distribution)**.** *Suppose* $X \sim \text{Bin}(n,p)$*. Then*

$$\frac{X-np}{\sqrt{np(1-p)}} \xrightarrow{\text{d}} N(0,1).$$

Proof. Let $X_1, \ldots, X_n$ be a set of independent Bernoulli random variables with parameter p. Then

$$X = \sum_i X_i.$$

From the CLT, as it applies to sums of independent random variables,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{X-n\mu}{\sigma\sqrt{n}} \le x\right) = \mathbb{P}(Z \le x)$$

where $Z \sim N(0,1)$ and $\mu = \mathbb{E}(X_i) = p$ and $\sigma^2 = \text{Var}(X_i) = p(1-p)$. The required result follows immediately. □

The practical ramifications are that probabilities involving binomial random variables with large n can be approximated by normal probabilities. However, a slight adjustment, known as a *continuity correction*, is often used to improve the approximation:

Proposition 5.7.1 (Normal Approximation to Binomial Distribution with Continuity Correction)**.** Suppose $X \sim \text{Bin}(n,p)$. Then

$$\mathbb{P}(X \le x) \approx \mathbb{P}\left(Z \le \frac{x-np+\frac{1}{2}}{\sqrt{np(1-p)}}\right)$$

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

The continuity correction is based on the fact that a discrete random variable is being approximated by a continuous random variable.

Example 5.7.2. Adam tosses 25 pieces of toast off a roof, and 10 of them land butter side up. Is this evidence that toast lands butter side down more often than butter side up? i.e. is $\mathbb{P}(X \le 10)$ unusually small?

Solution. $X \sim \text{Bin}(25, 0.5)$. We could answer this question by calculating the exact probability, but this would be time consuming. Instead, we use the fact that

$$\frac{X/n - 1/2}{\sqrt{1/(4n)}} \xrightarrow{\text{d}} Z \sim N(0,1).$$

The normal approximation to $\mathbb{P}(X \le 10)$ gives:

$$\mathbb{P}\left(Z < \frac{10-12.5}{\sqrt{25/4}}\right) = 0.158655\ldots$$

with the correction

$$\mathbb{P}\left(Z < \frac{10 - 12.5 + 1/2}{\sqrt{25/4}}\right) = 0.211855\ldots.$$

Compare this with the exact answer, obtained from the binomial distribution:

$$\mathbb{P}(X \leq 10) = 0.212178111\ldots$$

□

A useful rule of thumb is that the normal approximation to the binomial will work well when n is large enough that both $np > 5$ and $n(1-p) > 5$.

5.7.3 Normal Approximation to the Poisson Distribution

Theorem 5.7.2. *Suppose* $X \sim \text{Poisson}(\lambda)$. *Then*

$$\lim_{\lambda\to\infty} \mathbb{P}\left(\frac{X-\lambda}{\sqrt{\lambda}} \leq x\right) = \mathbb{P}(Z \leq x)$$

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

This approximation works increasingly well as λ gets large, and it provides a reasonable approximation to most Poisson probabilities for $\lambda > 5$. Note that because X is discrete, a continuity correction will improve the accuracy of this approximation.

5.8 The Delta Method

The CLT provides a large sample approximation to the distribution of $\overline{X}_n$. But what about the other functions of a sequence $X_1, X_2, \ldots$? These random variable sequences also converge in distribution to a normal random variable.

Theorem 5.8.1 (The Delta Method). *Let* $Y_1, Y_2, \ldots$ *be a sequence of random variables such that*

$$\frac{\sqrt{n}(Y_n - \theta)}{\sigma} \xrightarrow{\text{d}} Z \sim N(0,1).$$

Suppose the function g *is differentiable in the neighbourhood of* θ *and* $g'(\theta) \neq 0$. *Then,*

$$\sqrt{n}(g(Y_n) - g(\theta)) \xrightarrow{\text{d}} N(0, \sigma^2[g'(\theta)]^2).$$

Proof. First, note that

$$\begin{aligned} Y_n - \theta &= \frac{\sigma}{\sqrt{n}} \times \frac{\sqrt{n}(Y_n - \theta)}{\sigma} \\ &= \underbrace{\frac{\sigma}{\sqrt{n}}}_{\xrightarrow{\mathbb{P}} 0} \times \underbrace{\frac{\sqrt{n}(Y_n - \theta)}{\sigma}}_{\xrightarrow{\text{d}} Z \sim N(0,1)} \xrightarrow{\text{d}} 0 \times Z \\ &= 0 \qquad\qquad \text{(by Slutsky's Theorem).} \end{aligned}$$

In other words, $Y_n - \theta \xrightarrow{\text{d}} 0$, but convergence in distribution to a constant implies convergence in probability. Hence, we have established that

$$Y_n \xrightarrow{\mathbb{P}} \theta.$$

Second, by the mean value theorem we can write

$$g'(\vartheta_n) = \frac{g(Y_n) - g(\theta)}{Y_n - \theta}$$

for some ϑ_n between θ and Y_n. Note that the mean value result is valid if $\vartheta_n \in (\theta, Y_n)$ or $\vartheta_n \in (Y_n, \theta)$ and this is why we simply say that ϑ_n is between θ and Y_n. Any such random ϑ_n satisfies

$$|\vartheta_n - \theta| \leq |Y_n - \theta|.$$

From the last inequality it follows that

$$\mathbb{P}(|\vartheta_n - \theta| > \epsilon) \leq \mathbb{P}(|Y_n - \theta| > \epsilon) \to 0$$

and hence $\vartheta_n \xrightarrow{\mathbb{P}} \theta$ and by the continuity of g' in the neighbourhood of θ

$$g'(\vartheta_n) \xrightarrow{\mathbb{P}} g'(\theta).$$

Then substituting the expression for $g'(\vartheta)$ (given by the mean value theorem), we can write

$$\begin{aligned}
\frac{\sqrt{n}(g(Y_n) - g(\theta))}{\sigma g'(\theta)} &= \frac{\sqrt{n}(Y_n - \theta)}{\sigma} \times \frac{g'(\vartheta_n)}{g'(\theta)} \\
&= \underbrace{\frac{\sqrt{n}(Y_n - \theta)}{\sigma}}_{\xrightarrow{\mathrm{d}} Z} \times \frac{\overbrace{g'(\vartheta_n)}^{\xrightarrow{\mathbb{P}} g'(\theta)}}{g'(\theta)} \\
&\xrightarrow{\mathrm{d}} Z \times 1 \sim N(0,1) \qquad\qquad \text{(by Slutsky's Theorem).}
\end{aligned}$$

□

Equivalently, the result can be stated as follows:
If

$$Y_n = \theta + \frac{1}{\sqrt{n}} \sigma Z_n + (\text{terms in } \frac{1}{n} \text{ or smaller})$$

where $Z_n \xrightarrow{\mathrm{d}} N(0,1)$ then

$$g(Y_n) = g(\theta) + \frac{1}{\sqrt{n}} \sigma g'(\theta) Z_n + (\text{terms in } \frac{1}{n} \text{ or smaller}).$$

It is often useful in statistics to use this latter notation, where one expands a statistic into a constant term and terms which vanish at different rates as n increases.

Yet another way to write the Delta method result, which is more informal but useful for practical purposes, is as follows:

$$\text{If } Y_n \mathrel{\dot\sim} N\left(\theta, \frac{\sigma^2}{n}\right) \quad \text{then} \quad g(Y_n) \mathrel{\dot\sim} N\left(g(\theta), [g'(\theta)]^2 \frac{\sigma^2}{n}\right).$$

These two expressions are only valid for finite n, but n must be *large enough* for results when $n \to \infty$ to offer reasonable approximation to the distribution of $g(Y_n)$. Hence we refer to the above as a *large sample approximation* to the distribution of $g(Y_n)$.

Example 5.8.1. Let $X_1, X_2, \ldots$ be a sequence of independently and identically distributed random variables with mean 2 and variance 7. Obtain a large sample approximation for the distribution of $(\overline{X}_n)^3$.

Solution. The CLT gives

$$\sqrt{n}(\overline{X}_n - 2) \xrightarrow{\mathrm{d}} N(0, 7).$$

Application of the Delta Method with $g(x) = x^3$ leads to $g'(x) = 3x^2$ and then

$$\sqrt{n}[(\overline{X}_n)^3 - 8] \xrightarrow{\mathrm{d}} N(0, 1008).$$

For large n the approximate distribution of $(\overline{X})^3$ is $N(8, \frac{1008}{n})$. □

6. Distributions Arising from a Normal Sample

In applications, many data sets consist of *continuous* measurements. It is common to model data such as these as a *random sample* from the *normal distribution*. Validity of the assumption is often questionable, but because of the CLT, it may be true approximately.

6.1 Samples from the Normal Distribution

Definition 6.1.1. Let $X_1, \dots, X_n$ be a random sample with common distribution $N(\mu, \sigma^2)$ for some μ and σ^2. Then $X_1, \dots, X_n$ is a *normal random sample*.

Proposition 6.1.1. If $X_1, \dots, X_n$ is a random sample from the $N(\mu, \sigma^2)$ distribution then

(i) $\sum_i X_i \sim N(n\mu, n\sigma^2)$

(ii) $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

CLT states that a sum or mean of a random sample from any distribution is approximately normal. The above result states that this result is exact if we have a normal random sample.

6.2 The Chi-Squared Distribution

Definition 6.2.1. If X has density

$$f_X(x) = \frac{e^{-x} x^{\nu/2-1}}{2^{\nu/2}\Gamma(\nu/2)}, \quad x > 0$$

then X has the χ^2 (*chi-squared*) distribution with *degrees of freedom* ν. A common shorthand is

$$X \sim \chi^2_\nu.$$

Note that the chi-squared distribution is a special case of the Gamma distribution.

Corollary 6.2.1. *If $X \sim \chi^2_\nu$ then $X \sim$* Gamma$(\nu/2, 2)$.

Proposition 6.2.1. If $X \sim \chi^2_\nu$ then

(i) $\mathbb{E}(X) = \nu$.

(ii) $\mathrm{Var}(X) = 2\nu$.

(iii) $m_X(u) = \left(\frac{1}{1-2\nu}\right)^{\nu/2}$, $u < 1/2$.

Theorem 6.2.1 (Sum of Squared iid Normals). *If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, then*

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi_n^2.$$

6.3 The t Distribution

Definition 6.3.1. If $T \sim t_\nu$ then

$$f_T(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{(\nu+1)}{2}}, \quad -\infty < t < \infty.$$

A common shorthand is

$$T \sim t_\nu.$$

Theorem 6.3.1. *If $T \sim t_\nu$ then as $\nu \to \infty$, T converges to a $N(0,1)$ random variable.*

Proof.

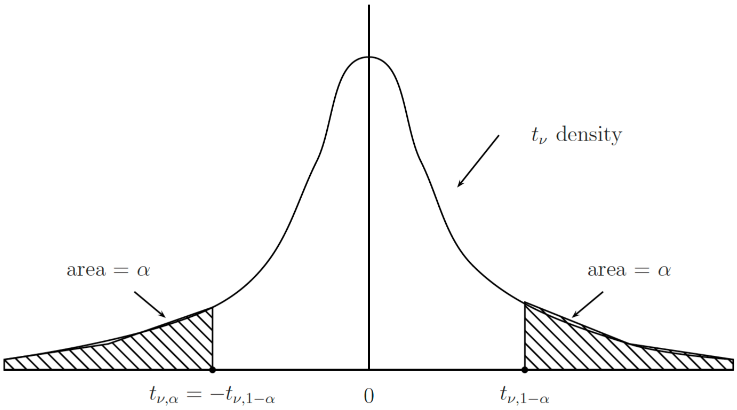
$$f_T(t) \propto \left(1 + \frac{t^2}{\nu}\right)^{-\frac{(\nu+1)}{2}} \quad \text{and} \quad \lim_{\nu\to\infty} \left(1 + \frac{t^2}{\nu}\right)^{-\nu} = e^{-t^2},$$

so

$$\lim_{\nu\to\infty} f_T(t) \propto e^{-\frac{t^2}{2}}.$$

□

The following is a plot of the t-Distribution's density function:



If $T \sim t_\nu$, then $\mathbb{P}(T \leq t_{\nu,\alpha}) = \alpha$. $t_{\nu,\alpha}$ is the α-th quantile of the t_ν distribution. $t_{\nu,1-\alpha}$ is the $(1-\alpha)$-th quantile of the t_ν distribution.

Theorem 6.3.2. *Let $X_1, \ldots, X_n$ be a random sample from the $N(\mu, \sigma^2)$ distribution. Then*

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

6.4 The F of Fisher Distribution

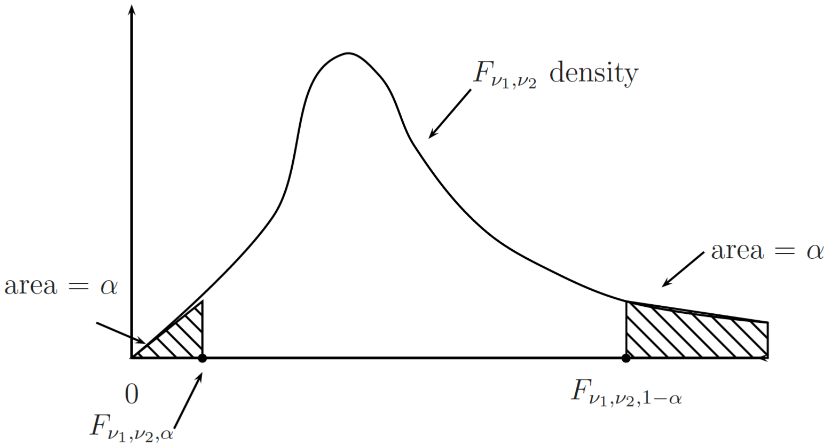
Suppose $X_1, X_2, \ldots, X_{n_X}$ are independent $N(\mu_X, \sigma_X^2)$ and $Y_1, Y_2, \ldots, Y_{n_Y}$ are independent $N(\mu_Y, \sigma_Y^2)$ and the samples are independent. When comparing the variances or drawing inferences about σ_X^2/σ_Y^2 we use S_X^2/S_Y^2 (the ratio of the sample variances) and this leads us to the F distribution.

Definition 6.4.1. Suppose $X \sim \chi^2_{\nu_1}$ and $Y \sim \chi^2_{\nu_2}$ and X and Y are independent. Then $F = \frac{X/\nu_1}{Y/\nu_2}$ has the F distribution with degrees of freedom ν_1 and ν_2. We write $F \sim \mathsf{F}_{\nu_1,\nu_2}$.

Theorem 6.4.1. *If $F \sim \mathsf{F}_{\nu_1,\nu_2}$ then F has density function*

$$f_F(u) = \frac{\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} u^{\frac{\nu_1}{2}-1}\left(1+\frac{\nu_1 u}{\nu_2}\right)^{-\frac{(\nu_1+\nu_2)}{2}}}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)}, \quad u > 0$$

where $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$.



$F_{\nu_1,\nu_2,\alpha}$ is the α-th quantile of the F_{ν_1,ν_2} distribution. $F_{\nu_1,\nu_2,1-\alpha}$ is the $(1-\alpha)$-th quantile of the F_{ν_1,ν_2} distribution.

Theorem 6.4.2 (Ratio of Sums of Squared Normal Samples)**.** *For independent samples*

$$X_1, X_2, \ldots, X_{n_X} \overset{\text{iid}}{\sim} N(\mu_X, \sigma_X^2)$$
$$Y_1, Y_2, \ldots, Y_{n_Y} \overset{\text{iid}}{\sim} N(\mu_Y, \sigma_Y^2)$$

with sample variances

$$S_X^2 = \frac{1}{n_X - 1}\sum_{i=1}^{n_X}(X_i - \overline{X})^2 = \textit{sample variance of the } X\textit{'s}$$
$$S_Y^2 = \frac{1}{n_Y - 1}\sum_{i=1}^{n_Y}(Y_i - \overline{Y})^2 = \textit{sample variance of the } Y\textit{'s}$$

we have

$$\frac{S_X^2/S_Y^2}{\sigma_X^2/\sigma_Y^2} \sim \mathsf{F}_{n_X-1, n_Y-1}.$$

7. An Introduction to Statistical Inference

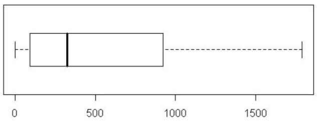
7.1 Statistical Models

Given an observed random sample $X_1, \dots, X_n$ it is common to postulate a *statistical model* for the data. This is a set of density or probability functions f_X that are consistent with the data, and facilitates answering certain questions of interest. A *parametric model* is a set of f_X's that can be parameterised by a finite number of parameters.

Example 7.1.1. A paper describes data on the life-times (hours) of 20 pressure vessels constructed of fibre/epoxy composite materials wrapped around metal liners. The data are:

```
274 28.5 1.7 20.8 871 363 1311 1661 236 828
458 290 54.9 175 1787 970 0.75 1278 776 126
```

The following figure shows a graphical representation of the data:



Given the positive and right-skewed nature of the data a plausible parametric model for the data is:

$$\{f_X(x;\beta) = \frac{1}{\beta}e^{-x/\beta}, \quad x > 0; \quad \beta > 0\}.$$

Since this family of density functions is parameterised by the single parameter $\beta > 0$, this is a parametric model. We could check how well this parametric model fits the data using a quantile-quantile plot.

Definition 7.1.1. A general parametric model with a single parameter θ is

$$\{f_X(x;\theta) : \theta \in \Theta\}.$$

The set $\Theta \subseteq \mathbb{R}$ is the set of possible values of θ and is known as the *parameter space.* If this model is assumed for a random sample $X_1, \dots, X_n$ then we write

$$X_1, \dots, X_n \sim f_X(x;\theta), \quad \theta \in \Theta.$$

Note that a model for a random sample induces a probability measure on its members. However, these probabilities depend on the value of θ. This is sometimes indicated using subscripted notation such as: $\mathbb{P}_\theta, \mathbb{E}_\theta, \mathrm{Var}_\theta$ to describe probabilities and expectations according to the model and particular values of θ, although we will minimise the use of this notation.

7.2 Estimation

Let $X_1, \ldots, X_n$ be a random sample with model

$$\{f_X(x;\theta) : \theta \in \Theta\}.$$

A fundamental problem in statistics is that of determining a single θ that is "most consistent" with the sample. This is known as the *estimation* problem, sometimes referred to as *point estimation.*

Definition 7.2.1. Suppose

$$X_1, \ldots, X_n \sim f_X(x;\theta), \quad \theta \in \Theta.$$

An *estimator* for θ, denoted by $\widehat{\theta}$, is any real-valued function of $X_1, \ldots, X_n$; i.e.

$$\widehat{\theta} = g(X_1, \ldots, X_n)$$

where the function $g : \mathbb{R}^n \to \mathbb{R}$.

Example 7.2.1. Let

$$p = \text{proportion of UNSW students who watched the Cricket World Cup final}$$

be a parameter of interest. Suppose that we survey 8 UNSW students, asking them whether they watched the Cricket World Cup final. Let $X_1, \ldots, X_8$ be such that

$$X_i = \begin{cases} 1, & \text{if } i\text{-th surveyed watched the Cricked World Cup final} \\ 0, & \text{otherwise.} \end{cases}$$

An appropriate model is

$$X_1, \ldots, X_8 \sim f_X(x;p), \quad 0 < p < 1,$$

where

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

Then the 'natural' estimator for p is

$$\widehat{p} = \frac{X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8}{8},$$

corresponding to the proportion from the survey that watched the Cricket World Cup final. However, there are many other possible estimators for p; such as

$$\widehat{p}_{\text{alt}} = \frac{X_2 + X_4 + X_6 + X_8}{4},$$

based on every second person in the survey. But even

$$\widehat{p} = \sin(X_1 e^{X_5}) + \pi \coth(X_3/(7X_8 + 12)),$$

satisfies the definition of being an estimator for p, though it is not a very good estimator.

The previous definition permits any function of the sample to be an estimator for a parameter θ. However, only certain functions have good properties.

Remark 7.2.1. The estimator $\widehat{\theta}$ is a function of the random variables $X_1, \ldots, X_n$ and is therefore a random variable itself. It has its own probability function or density function

$$f_{\widehat{\theta}}$$

that depends on θ. We use $f_{\widehat{\theta}}$ to study the properties of $\widehat{\theta}$ as an estimator of θ.

7.2.1 Bias

Definition 7.2.2. Let $\widehat{\theta}$ be an estimator of a parameter θ. The *bias* of $\widehat{\theta}$ is given by

$$\text{bias}(\widehat{\theta}) = \mathbb{E}(\widehat{\theta}) - \theta.$$

If $\text{bias}(\widehat{\theta}) = 0$ then $\widehat{\theta}$ is said to be an *unbiased* estimator of θ.

Bias is a measure of *systematic error* of an estimator - how far we expect the estimator to be from its true value θ, on average. Often, we want an estimator that has minimal bias.

Example 7.2.2. For the Cricket World Cup example

$$\widehat{p} = \frac{Y}{8}$$

where Y is the number of students who watched the Cricket World Cup final, and

$$Y \sim \text{Bin}(8, p).$$

Find $f_{\widehat{p}}(x)$, and $\text{bias}(\widehat{p})$. Compare these to the corresponding results for $\widehat{p}_{\text{alt}} = (X_2 + X_4 + X_6 + X_4)/4$.

Solution. Note that Y has probability function

$$f_Y(y) = \binom{8}{y} p^y (1-p)^{8-y}, \quad y = 0, 1, 2, \ldots, 8.$$

Because $\widehat{p} = Y/8$, $Y = 8\widehat{p}$ and using rules for transformation of a discrete random variable:

$$f_{\widehat{p}}(x) = \binom{8}{8x} p^{8x} (1-p)^{8-8x}, \quad x = 0, \frac{1}{8}, \frac{2}{8}, \ldots, 1.$$

Now to find $\text{bias}(\widehat{p}) = \mathbb{E}(\widehat{p} - p)$, we will first find $\mathbb{E}(\widehat{p})$. Since $\mathbb{E}(Y) = 8p$,

$$\mathbb{E}(\widehat{p}) = \mathbb{E}(Y/8) = \mathbb{E}(Y)/8 = 8p/8 = p$$

and so

$$\text{bias}(\widehat{p}) = \mathbb{E}(\widehat{p} - p) = \mathbb{E}(\widehat{p}) - \mathbb{E}(p) = 0.$$

This means that $\widehat{p}$ is an unbiased estimator for p. By similar arguments,

$$f_{\widehat{p}_{\text{alt}}}(x) = \binom{4}{4x} p^{4x} (1-p)^{4-4x}, \quad x = 0, \frac{1}{4}, \frac{2}{4}, \ldots, 1.$$

and $\text{bias}(\widehat{p}_{\text{alt}}) = 0$, so $\widehat{p}_{\text{alt}}$ is also unbiased. □

7.2.2 Standard Error

Definition 7.2.3. Let $\widehat{\theta}$ be an estimator of a parameter θ. The *standard error* of $\widehat{\theta}$ is the *standard deviation*:

$$\text{se}(\widehat{\theta}) = \sqrt{\text{Var}(\widehat{\theta})}.$$

To obtain the *estimated standard error* $\widehat{\text{se}}(\widehat{\theta})$, we first derive $\text{Var}(\widehat{\theta})$, the variance of $\widehat{\theta}$, and then we replace unknown parameters θ by its estimator $\widehat{\theta}$.

Like the bias, the standard error of an estimator is ideally as small as possible. However, unlike the bias the standard error can never be made zero (except in trivial cases).

Example 7.2.3. Consider, again, the Cricket World Cup example. Find $\text{se}(\widehat{p})$ and $\text{se}(\widehat{p}_{\text{alt}})$. Comment.

Solution. From properties of binomial random variables,

$$\mathrm{Var}(\widehat{p}) = \mathrm{Var}(Y/8) = (1/8)^2\,\mathrm{Var}(Y) = \frac{p(1-p)}{8}.$$

Therefore the standard error of $\widehat{p}$ is

$$\mathrm{se}(\widehat{p}) = \sqrt{\frac{p(1-p)}{8}}$$

and the estimated standard error of $\widehat{p}$ is

$$\widehat{\mathrm{se}}(\widehat{p}) = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{8}}.$$

Similarly,

$$\mathrm{se}(\widehat{p}_{\mathrm{alt}}) = \sqrt{\frac{p(1-p)}{4}} \quad \text{and} \quad \widehat{\mathrm{se}}(\widehat{p}_{\mathrm{alt}}) = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{4}}.$$

Therefore the standard error of $\widehat{p}_{\mathrm{alt}}$ will always be larger than that of $\widehat{p}$ by a factor of $\sqrt{2}$, which suggests that $\widehat{p}$ is a better estimator of p than $\widehat{p}_{\mathrm{alt}}$. □

7.2.3 Mean Squared Error

The bias and standard error of an estimator are fundamental measures of different aspects of the quality of $\widehat{\theta}$, as an estimator of θ: bias is concerned with the systematic error in $\widehat{\theta}$, while the standard error is concerned with its inherent random (or sampling) error. However, how do we choose between two estimators of θ, one with smaller bias and the other with smaller standard error?

One approach is to use a combined measure of the quality of $\widehat{\theta}$, which combines the bias and standard error.

Definition 7.2.4. The *mean squared error* of $\widehat{\theta}$ is given by

$$\mathrm{MSE}(\widehat{\theta}) = \mathbb{E}[(\widehat{\theta} - \theta)^2].$$

Theorem 7.2.1. *Let $\widehat{\theta}$ be an estimator of a parameter θ. Then*

$$\mathrm{MSE}(\widehat{\theta}) = \mathrm{bias}(\widehat{\theta})^2 + \mathrm{Var}(\widehat{\theta}),$$

and the estimated mean squared error is

$$\widehat{\mathrm{MSE}}(\widehat{\theta}) = \mathrm{bias}(\widehat{\theta})^2 + \widehat{\mathrm{se}}(\widehat{\theta})^2.$$

Proof.

$$\begin{aligned}
\mathrm{MSE}(\widehat{\theta}) &= \mathbb{E}[(\widehat{\theta} - \theta)^2] \\
&= \mathbb{E}[(\widehat{\theta} - \mathbb{E}(\widehat{\theta}) + \mathbb{E}(\widehat{\theta}) - \theta)^2] \\
&= \mathbb{E}[(\widehat{\theta} - \mathbb{E}(\widehat{\theta}))^2] + \mathbb{E}[(\mathbb{E}(\widehat{\theta}) - \theta)^2] + 2\mathbb{E}[(\widehat{\theta} - \mathbb{E}(\widehat{\theta}))(\mathbb{E}(\widehat{\theta}) - \theta)] \\
&= \mathrm{Var}(\widehat{\theta}) + \mathbb{E}[\mathrm{bias}(\widehat{\theta})^2] + 2(\mathbb{E}(\widehat{\theta}) - \theta)\mathbb{E}[(\widehat{\theta} - \mathbb{E}(\widehat{\theta}))] \\
&= \mathrm{Var}(\widehat{\theta}) + \mathrm{bias}(\widehat{\theta})^2 + 2(\mathbb{E}(\widehat{\theta}) - \theta)(\mathbb{E}(\widehat{\theta}) - \mathbb{E}(\widehat{\theta})) \\
&= \mathrm{Var}(\widehat{\theta}) + \mathrm{bias}(\widehat{\theta})^2.
\end{aligned}$$

□

Definition 7.2.5. Let $\widehat{\theta}_1$ and $\widehat{\theta}_2$ be two estimators of a parameter θ. Then $\widehat{\theta}_1$ is *better than* $\widehat{\theta}_2$ (with respect to MSE) at $\theta_0 \in \Theta$ if

$$\mathrm{MSE}_{\theta_0}(\widehat{\theta}_1) < \mathrm{MSE}_{\theta_0}(\widehat{\theta}_2).$$

If $\widehat{\theta}_1$ is better than $\widehat{\theta}_2$ for all $\theta \in \Theta$ then we say

$$\widehat{\theta}_1 \textit{ is uniformly better than } \widehat{\theta}_2.$$

Example 7.2.4. For the Cricket World Cup example, find the MSE($\widehat{p}$) and MSE($\widehat{p}_{\text{alt}}$). Is $\widehat{p}$ uniformly better than $\widehat{p}_{\text{alt}}$?

Solution. From previous results,

$$\mathrm{MSE}(\widehat{p}) = 0^2 + \frac{p(1-p)}{8} = \frac{p(1-p)}{8}$$

while

$$\mathrm{MSE}(\widehat{p}_{\text{alt}}) = \frac{p(1-p)}{4}.$$

Since

$$\mathrm{MSE}(\widehat{p}) < \mathrm{MSE}(\widehat{p}_{\text{alt}}) \quad \text{for all } 0 < p < 1$$

$\widehat{p}$ is uniformly better than $\widehat{p}_{\text{alt}}$. (This result makes intuitive sense, since $\widehat{p}$ is based on twice as many responses). □

Proposition 7.2.1. The standard error expressions for sample means and sample proportions are as follows:

$$\widehat{\mathrm{se}}(\overline{X}) = \frac{\widehat{\sigma}}{\sqrt{n}}$$

where $\widehat{\sigma}$ is the sample (estimated) standard deviation; and

$$\widehat{\mathrm{se}}(\widehat{p}) = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

7.2.4 Consistency

The consistency property of an estimator is concerned with its performance as the amount of data increases. It seems reasonable to demand that $\widehat{\theta} = \widehat{\theta}_n$ gets better as the sample size n grows.

Definition 7.2.6. The estimator $\widehat{\theta}_n$ is *consistent* for θ if

$$\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta.$$

Example 7.2.5. For the Cricket World Cup example, suppose that n people are surveyed and

$$\widehat{p}_n = \frac{Y_n}{n}$$

where

$$Y_n = \text{number of students that watched Cricket World Cup final.}$$

Prove that $\widehat{p}_n$ is consistent.

Solution. $Y_n \sim \text{Bin}(n,p)$ and

$$\mathbb{E}(\widehat{p}_n) = p.$$

Application of the Weak Law of Large numbers (to the binary X_i random variables) leads to

$$\widehat{p}_n \xrightarrow{\mathbb{P}} p.$$

Hence $\widehat{p}_n$ is consistent. □

Theorem 7.2.2. *If*

$$\lim_{n\to\infty} \text{MSE}(\widehat{\theta}_n) = 0$$

then

$$\widehat{\theta}_n \textit{ is consistent for } \theta.$$

Example 7.2.6. For the Cricket World Cup example, find $\text{MSE}(\widehat{p}_n)$. Hence show that $\widehat{p}_n$ is consistent for p.

Solution. It is easily shown that

$$\text{MSE}(\widehat{p}_n) = \frac{p(1-p)}{n}$$

so

$$\lim_{n\to\infty} \text{MSE}(\widehat{p}_n) = p(1-p) \lim_{n\to\infty} \frac{1}{n} = 0$$

and therefore $\widehat{p}_n$ is consistent for p. □

7.2.5 Asymptotic Normality

Definition 7.2.7. The estimator $\widehat{\theta}$ is *asymptotically normal* if

$$\frac{\widehat{\theta}-\theta}{\text{se}(\widehat{\theta})} \xrightarrow{\text{d}} N(0,1).$$

In particular, we know already from the CLT that a sample mean $\widehat{\mu} = \overline{X}$ and a sample proportion $\widehat{p}$ are asymptotically normal.

Example 7.2.7. In the Cricket World Cup example, let $X_1, \dots, X_n$ be defined by

$$X_i = \begin{cases} 1, & \text{if } i\text{-th surveyed student watched the final} \\ 0, & \text{otherwise.} \end{cases}$$

Assume the X_i are independent. Write $\widehat{p}_n$ in terms of the X_i and hence use convergence results from Chapter 5 to find the asymptotic distribution of $\widehat{p}_n$.

Example 7.2.8. Consider the sample mean $\overline{X}_n$ of n independent random variables with mean μ and variance σ^2. Show that $\overline{X}_n$ is consistent. Is $\overline{X}_n$ asymptotically normal?

7.2.6 Observed Values

So far we have considered the random sample $X_1, \dots, X_n$ and the properties of $\widehat{\theta}$ by treating it as a random variable. In practice, we only take one sample, and observe a single value of $\widehat{\theta}$, known as the *observed value* of $\widehat{\theta}$. This is sometimes called the *estimate* of θ (as opposed to the estimator, which is the random variable we use to obtain the estimate).

Example 7.2.9. Consider the pressure vessel example and let

$$X_i = i\text{-th lifetime before the data are observed.}$$

An unbiased estimator for β is

$$\widehat{\beta} = \overline{X} = \frac{1}{n}\sum_i X_i.$$

After the data are observed to be:

```
274 28.5 1.7 20.8 871 363 1311 1661 236 828
458 290 54.9 175 1787 970 0.75 1278 776 126
```

the observed value of $\widehat{\beta}$ becomes

$$(274 + 28.5 + \cdots + 126)/20 = 575.53.$$

Note, $\widehat{\beta}$ denotes the random variables $\overline{X}$ before the data are observed. But we will also say $\widehat{\beta} = 575.53$ for the observed value. The meaning of $\widehat{\beta}$ should be clear from the context. Moreover, in applied statistics, a common notation when reporting the observed value of an estimator (or estimate) is to add the estimated standard error in parentheses: estimate (standard error) e.g. 0.25 (0.153), where 0.25 is the observed value and 0.153 is the standard error.

7.3 Multiparameter Models

For the pressure vessel data we have previously considered the single parameter model

$$\{f_X(x;\beta) = \frac{1}{\beta}e^{-x/\beta}, \quad x > 0; \quad \beta > 0\}.$$

A *two-parameter* model is

$$\{f_X(x;\alpha,\beta) = \frac{1}{\beta^\alpha\Gamma(\alpha)}x^{\alpha-1}e^{-x/\beta}, \quad x > 0; \quad \alpha,\beta > 0\},$$

which corresponds to the X_i's having a Gamma(α,β) distribution.

Often inference is concerned with the original parameters: α and β for a Gamma(α,β) model. However, sometimes a *transformation of the parameters* is of interest. For example, in the Gamma(α,β) model a parameter of interest is often

$$\tau = \alpha\beta$$

since this corresponds to the mean of the distribution.

7.4 Confidence Intervals

An estimator $\widehat{\theta}$ of a parameter θ leads to a single number for inferring the true value of θ. For example, in the Cricket World Cup example if we survey 50 people and 16 watched the Cricket World Cup final then the estimator $\widehat{p}$ has an observed value of 0.32. However, the number 0.32 alone does not tell us much about the inherent variability in the underlying estimator. Confidence intervals aim to improve this situation with a *range* of plausible values, e.g.

$$p \text{ is likely to be in the range 0.19 to 0.45.}$$

Definition 7.4.1. Let $X_1, \dots, X_n$ be a random sample from a model that includes an unknown parameter θ. Let

$$L = L(X_1, \dots, X_n) \quad \text{and} \quad U = U(X_1, \dots, X_n)$$

be statistics (i.e. functions of the X_i's) for which

$$\mathbb{P}(L < \theta < U) \geq 1 - \alpha, \quad \text{for all } \theta \in \Theta.$$

Then (L, U) is a $1 - \alpha$*, or* $100(1 - \alpha)\%$*, confidence interval* for θ.

It is important to note that in the probability statement

$$\mathbb{P}(L < \theta < U) \geq 1 - \alpha$$

the quantity in the middle (θ) is fixed, but the limits (L and U) are random. This is the reverse situation from many probability statements e.g. $\mathbb{P}(2 \leq X \leq 7)$, for a random variable X.

Example 7.4.1. Consider the pressure vessel example

$$X_1, \dots, X_{20} \sim f_X(x; \beta)$$

where

$$f_X(x; \beta) = \frac{1}{\beta} e^{-x/\beta}, \quad x > 0; \quad \beta > 0.$$

Then it can be shown that

$$\mathbb{P}(0.52\overline{X} \leq \beta \leq 1.67\overline{X}) = 0.99 \quad \text{for all } \beta > 0.$$

Therefore

$$(0.52\overline{X}, 1.67\overline{X})$$

is a 0.99 or 99% confidence interval for β. Since the observed value for $\overline{X}$ is $\overline{x} = 575.53$, the observed value of the 99% confidence interval for β is

$$(352, 852).$$

7.5 Confidence Intervals for a Normal Random Sample

From Chapter 6, the special case where the random sample can be reasonably modelled as coming from a normal distribution:

$$X_1, \dots, X_n \sim N(\mu, \sigma^2),$$

we derived the following distribution theory results:

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1} \quad \text{and} \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

They allow for exact confidence interval statements for μ and σ.

Theorem 7.5.1. *Let $X_1, \dots, X_n$ be a random sample from the $N(\mu, \sigma^2)$ distribution. Then a $100(1 - \alpha)\%$ confidence interval for μ is*

$$\left(\overline{X} - t_{n-1,1-\alpha/2} \frac{S}{\sqrt{n}}, \overline{X} + t_{n-1,1-\alpha/2} \frac{S}{\sqrt{n}}\right).$$

Proof. By the definition of $t_{n-1,q}$,

$$\begin{aligned}
1-\alpha &= \mathbb{P}\left(-t_{n-1,1-\alpha/2} < \frac{\overline{X}-\mu}{S/\sqrt{n}} < t_{n-1,1-\alpha/2}\right) \\
&= \mathbb{P}(-t_{n-1,1-\alpha/2}S/\sqrt{n} < \overline{X}-\mu < t_{n-1,1-\alpha/2}S/\sqrt{n}) \\
&= \mathbb{P}(t_{n-1,1-\alpha/2}S/\sqrt{n} > \mu-\overline{X} > -t_{n-1,1-\alpha/2}S/\sqrt{n}) \\
&= \mathbb{P}\left(\overline{X} - \frac{S}{\sqrt{n}}t_{n-1,1-\alpha/2} < \mu < \overline{X} + \frac{S}{\sqrt{n}}t_{n-1,1-\alpha/2}\right).
\end{aligned}$$

□

The above is particularly powerful. It gives us a method of making potentially very specific statements about μ, based on a sample: we can estimate a range of values which we are arbitrarily sure will contain μ (such intervals contain μ $100(1-\alpha)\%$ of the time).

This is particularly useful since many research questions can be phrased in terms of means e.g. What is the average recovery time for patients given a new treatment?

Recall from Chapter 5 that it was noted that even when X_i are not normal, t_{n-1} is a reasonable approximation for the distribution of $\frac{\overline{X}-\mu}{S/\sqrt{n}}$, as long as n is large enough for the CLT to come into play. This means that we can use the above method to construct a confidence interval for μ even when the data is not normal.

7.6 Confidence Intervals for Two Normal Random Samples

A common situation in applied statistics is one of *comparison* between two samples e.g. "Is recovery time shorter for patients using a new treatment than for patients on the old treatment"?

Theorem 7.6.1. *Let*

$$\begin{aligned}
X_1,\ldots,X_{n_X} &\sim N(\mu_X,\sigma^2) \\
Y_1,\ldots,Y_{n_Y} &\sim N(\mu_Y,\sigma^2),
\end{aligned}$$

be two independent normal random samples; each with the same variance σ^2. Then a $100(1-\alpha)\%$ confidence interval for $\mu_X-\mu_Y$ is

$$\overline{X}-\overline{Y} \pm t_{n_X+n_Y-2,1-\alpha/2}S_p\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}$$

where

$$S_p^2 = \frac{(n_X-1)S_X^2+(n_Y-1)S_Y^2}{n_X+n_Y-2},$$

known as the pooled sample variance.

Proof. The proof is analogous to the simple sample confidence interval for μ. □

As previously, this assumption about the variables being normal is not critical when making inferences about μ, because some robustness to non-normality is inherited from the CLT.

Example 7.6.1. Peak expiratory flow is measured for 7 "normal" six-year-old children and nine-year-old children with asthma. The data are as follows:

Normal children	asthmatic children
55	53
57	39
80	56
71	54
62	49
56	53
77	45
	37
	44

We would like to know if peak flow is different between normal and asthmatic children, and if so, how different. Use a confidence interval to answer this question.

Solution. Graphical inspection of the data shows that the data do not appear to be too far from the normal distribution, for each sample. Let

$$\mu_X = \text{mean peak expiratory flow for normal children}$$
$$\mu_Y = \text{mean peak expiratory flow for asthmatic children.}$$

We will obtain a 95% confidence interval for $\mu_X - \mu_Y$ under the assumption that the variances for each population are equal. The sample sizes, sample means and sample variances are

$$n_X = 7, \qquad \overline{x} \approx 65.43, \qquad s_X^2 \approx 109.62$$
$$n_Y = 9, \qquad \overline{y} \approx 47.78, \qquad s_Y^2 \approx 47.19.$$

The pooled sample variance is

$$s_p^2 \approx \frac{6 \times 109.62 + 8 \times 47.19}{14} \approx 73.95.$$

The appropriate t-distribution quantile is $t_{14,0.975} = 2.145$. The confidence interval is then

$$(65.43 - 47.78) \pm 2.15\sqrt{73.95}\sqrt{\frac{1}{7} + \frac{1}{9}} = (8.4, 27).$$

In conclusion, we can be 95% confident that the difference in mean peak expiratory flow between the two groups is between 8.4 and 27.0 units. □

7.6.1 Confidence Intervals for a Paired Normal Random Sample

If we have a paired normal random sample from (X, Y), i.e. two normal random samples that are *dependent*, we can construct a confidence interval for the mean difference by analysing the differences $D = X - Y$ as a single normal random sample.

7.7 Confidence Intervals for Sample Proportions

Consider a binomial sample $X \sim \text{Bin}(n, p)$, where we are interested in making inferences about p, the probability of a "success". Recall that we have shown that if $X \sim \text{Bin}(n, p)$, then

$$\frac{\widehat{p} - p}{\sqrt{p(1-p)/n}} \xrightarrow{\text{d}} N(0, 1).$$

This means we can use the normal distribution to construct a confidence interval for p, based on a binomial sample.

Theorem 7.7.1. *Let $X \sim \mathrm{Bin}(n, p)$. Then an approximate $100(1-\alpha)\%$ confidence interval for p is*

$$(\widehat{p} - z_{1-\alpha/2}\widehat{\mathrm{se}}(\widehat{p}), \widehat{p} + z_{1-\alpha/2}\widehat{\mathrm{se}}(\widehat{p})),$$

where $\widehat{p} = \frac{X}{n}$ and $\widehat{\mathrm{se}}(\widehat{p}) = \sqrt{\widehat{p}(1-\widehat{p})/n}$.

Note that this is only an "approximate" confidence interval for two reasons: $\widehat{p}$ is only approximately normal, and we are using $\widehat{\mathrm{se}}(\widehat{p})$ in place of $\sqrt{p(1-p)/n}$.

8. Methods for Parameter Estimation and Inference

Throughout this chapter it is helpful to keep in mind the distinction between *estimates* and *estimators.* An estimate of a parameter θ is a function $\widehat{\theta} = \widehat{\theta}(x_1, \ldots, x_n)$ of observed values $x_1, \ldots, x_n$, whereas the corresponding estimator is the same function $\widehat{\theta}(X_1, \ldots, X_n)$ of the observable random variables $X_1, \ldots, X_n$. Thus an estimator is a random variable whose properties may be examined and considered before the observation process occurs, whereas an estimate is an actual number, the realised value of the estimator, evaluated after the observations are available.

In deriving estimation formulas it is often easier to work with estimates, but in considering theoretical properties, switching to estimators may be necessary.

For notation convenience, we usually denote the density or probability function f_X simply by f.

8.1 Method of Moments Estimation

Definition 8.1.1. Let $x_1, \ldots, x_n$ be observations from the model

$$f = f(x; \theta_1, \ldots, \theta_k)$$

containing k parameters $\theta = (\theta_1, \ldots, \theta_k)$. Form the system of k equations that equates the moments of f_X with their sample counterparts:

$$\begin{aligned}
\mathbb{E}(X) &= \frac{1}{n} \sum_i x_i \\
\mathbb{E}(X^2) &= \frac{1}{n} \sum_i x_i^2 \\
&\vdots \\
\mathbb{E}(X^k) &= \frac{1}{n} \sum_i x_i^k.
\end{aligned}$$

Then the *method of moments* estimates are the solutions of these equations in $\theta_1, \ldots, \theta_k$.

Example 8.1.1. Consider a random sample with normal model:

$$X_1, \ldots, X_n \sim N(\mu, \sigma^2).$$

Find the method of moments estimators of μ and σ^2.

Solution. The method of moments equations are:

$$\begin{aligned}
\mathbb{E}(X) &= \frac{1}{n} \sum_i x_i \\
\mathbb{E}(X^2) &= \frac{1}{n} \sum_i x_i^2.
\end{aligned}$$

But

$$\mathbb{E}(X) = \mu \quad \text{and} \quad \mathbb{E}(X^2) = \text{Var}(X) + \mathbb{E}(X)^2 = \sigma^2 + \mu^2$$

which leads to the system of equations:

$$\begin{aligned} \mu &= \overline{x} \\ \sigma^2 + \mu^2 &= \frac{1}{n}\sum_i x_i^2. \end{aligned}$$

Substitution of the first equation into the second leads to

$$\sigma = \sqrt{\frac{1}{n}\sum_i (x_i^2 - \overline{x})^2} = \sqrt{\frac{1}{n}\sum_i (x_i - \overline{x})^2}.$$

So the method of moments estimators of μ and σ are:

$$\widehat{\mu} = \overline{X} \quad \text{and} \quad \widehat{\sigma} = \sqrt{\frac{1}{n}\sum_i (X_i - \overline{X})^2}.$$

□

8.2 Consistency of Method of Moments Estimators

Assuming $\text{Var}(X^k) < \infty$, the Weak Law of Large Numbers states that

$$\frac{1}{n}\sum_i X_i^j \xrightarrow{\mathbb{P}} \mathbb{E}(X^j), \quad 1 \leq j \leq k.$$

Hence we can establish that

$$\widehat{\theta}_j \xrightarrow{\mathbb{P}} \theta_j, \quad 1 \leq j \leq k.$$

That is, the method of moments leads to consistent estimation of the model parameters.

Method of moments estimation is useful in practice because it is a simple approach that guarantees us a consistent estimator. However it is not always optimal, in the sense that it does not always provide us with an estimator that has the smallest possible standard errors and mean squared error.

8.3 Maximum Likelihood Estimation

This procedure has optimal performance for large samples, for almost any model. When this estimation method is possible, it is usually as good or better than method of moments estimation

Definition 8.3.1. Let $x_1, \ldots, x_n$ be observation from the pdf f where

$$f(x) = f(x;\theta)$$

depending on a parameter $\theta \in \Theta$. The *likelihood function* $\mathcal{L}$, a function of θ, is

$$\mathcal{L}(\theta) = f(x_1;\theta)\cdots f(x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta), \quad \theta \in \Theta,$$

and the *log-likelihood function* of θ is

$$\ell(\theta) = \ln \mathcal{L}(\theta) = \sum_i \ln f(x_i;\theta).$$

Note that the form of the likelihood function as a function of the observations $x_1, \dots, x_n$ is the same as the joint density function, but the likelihood function is regarded as a function of θ, for fixed values of $\{x_i\}$.

Example 8.3.1. Let

$$X_i = \begin{cases} 1, & \text{if } i\text{-th survey student watched the World Cup Cricket final} \\ 0, & \text{otherwise,} \end{cases}$$

for $i = 1, \dots, 8$. An appropriate probability function for $\{X_i\}$ is $f(x; p)$, where

$$\begin{aligned} f(x;p) &= \begin{cases} p, & x = 1 \\ 1-p, & x = 0 \end{cases} \\ &= p^x (1-p)^{1-x}. \end{aligned}$$

Find the likelihood and log-likelihood functions, i.e. $\mathcal{L}(p)$ and $\ell(p)$, hence find expressions for $\mathcal{L}(0.3)$ and $\ell(0.3)$.

Solution. The likelihood function for p is

$$\mathcal{L}(p) = \prod_{i=1}^{8} [p^{x_i}(1-p)^{1-x_i}] = p^{\sum_{i=1}^{8} x_i}(1-p)^{8-\sum_{i=1}^{8} x_i}$$

and the log-likelihood function for p is

$$\ell(p) = \ln \mathcal{L}(p) = \left(\sum_{i=1}^{8} x_i\right) \ln p + \left(8 - \sum_{i=1}^{8} x_i\right) \ln(1-p).$$

Taking $p = 0.3$,

$$\mathcal{L}(0.3) = (0.3)^{\sum_{i=1}^{8} x_i}(0.7)^{8-\sum_{i=1}^{8} x_i}$$

and

$$\ell(0.3) = \left(\sum_{i=1}^{8} x_i\right) \ln 0.3 + \left(8 - \sum_{i=1}^{8} x_i\right) \ln 0.7.$$

□

Definition 8.3.2. Let $x_1, \dots, x_n$ be observations from probability/density function f, where

$$f(x) = f(x; \theta)$$

containing the parameter $\theta \in \Theta$. The *maximum likelihood estimate* of θ is the choice

$$\widehat{\theta} = \theta \text{ that maximises } \mathcal{L}(\theta) \text{ over } \theta \in \Theta.$$

Example 8.3.2. In the World Cup Cricket example ($n = 8$) the likelihood function has been shown to be

$$\mathcal{L}(p) = \prod_{i=1}^{8} [p^{x_i}(1-p)^{1-x_i}] = p^{\sum_{i=1}^{8} x_i}(1-p)^{8-\sum_{i=1}^{8} x_i} = e^{(\sum_{i=1}^{8} x_i) \ln p + (8 - \sum_{i=1}^{8} x_i) \ln(1-p)}.$$

Find the maximum likelihood estimator of p by finding the value of p that maximises $\mathcal{L}(p)$.

Solution. The first derivative of $\mathcal{L}(p)$ with respect to p is then

$$\begin{aligned}\frac{d}{dp}\mathcal{L}(p) &= e^{(\sum_{i=1}^8 x_i)\ln p+(8-\sum_{i=1}^8 x_i)\ln(1-p)}\left[\left(\sum_{i=1}^8 x_i\right)/p-\left(8-\sum_{i=1}^8 x_i\right)/(1-p)\right]\\ &= \mathcal{L}(p)\left(\frac{\sum_{i=1}^8 x_i}{p}-\frac{8-\sum_{i=1}^8 x_i}{1-p}\right),\end{aligned}$$

and is zero if and only if

$$\frac{\sum_{i=1}^8 x_i}{p}-\frac{8-\sum_{i=1}^8 x_i}{1-p}=0 \iff p=\frac{\sum_{i=1}^8 x_i}{8}.$$

Further analysis (see next example) shows that this is the unique maximiser of $\mathcal{L}(p)$ over $0<p<1$ so

$$\widehat{p}=\frac{\sum_{i=1}^8 x_i}{8}=\text{proportion of people in survey that watched the World Cup Cricket final}$$

is the maximum likelihood estimate of p. □

Example 8.3.3. Suppose that the observed data are:

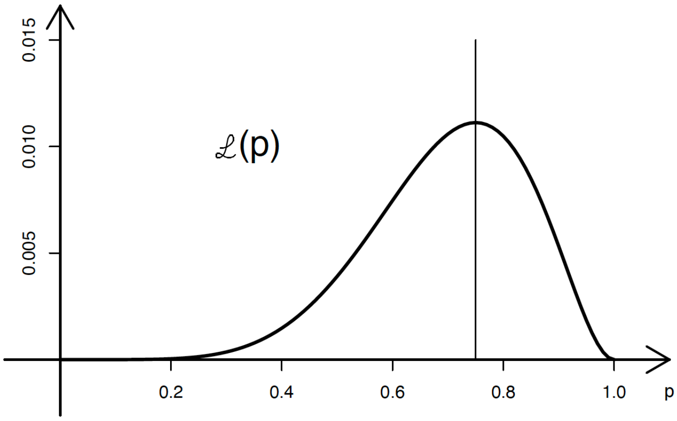
$$x_1=0, x_2=1, x_3=1, x_4=1, x_5=1, x_6=0, x_7=1, x_8=1.$$

Plot the observed likelihood function of p.

Solution. The observed value of the likelihood function is

$$\mathcal{L}(p)=p^6(1-p)^2, \quad 0<p<1.$$

The following figure is a graphical illustration of the maximum likelihood procedure in this case. Note that $\widehat{p}=6/8=0.75$ is the value of $p\in(0,1)$ that maximises $\mathcal{L}(p)$.



□

8.4 Obtaining Maximum Likelihood Estimators

8.4.1 Smooth Likelihood Functions

Consider estimation of a general parameter θ. If $\mathcal{L}(\theta)$ is smooth then calculus can be used to obtain the maximiser of $\mathcal{L}(\theta)$. However, it is usually simpler to work with the log-likelihood function $\ell(\theta)$.

Proposition 8.4.1. The point at which $\mathcal{L}(\theta)$ attains its maximum over $\theta \in \Theta$ is also that where

$$\ell(\theta) = \ln \mathcal{L}(\theta) = \sum_i \ln f(x_i; \theta)$$

attains its maximum. Therefore, the maximum likelihood estimate of θ is

$$\widehat{\theta} = \theta \text{ that maximises } \ell(\theta) \text{ over } \theta \in \Theta.$$

Example 8.4.1. Re-visit the World Cup Cricket example. Find the maximum likelihood estimator of p by finding the value of p that maximises $\ell(p)$. Then plot $\ell(p)$ when the observed data are

$$x_1 = 0, x_2 = 1, x_3 = 1, x_4 = 1, x_5 = 1, x_6 = 0, x_7 = 1, x_8 = 1.$$

Solution.

$$\ell(p) = \ln \mathcal{L}(p) = \left(\sum_{i=1}^{8} x_i\right) \ln p + \left(8 - \sum_{i=1}^{8} x_i\right) \ln(1-p).$$

The first derivative is

$$\frac{d}{dp}\ell(p) = \frac{\sum_{i=1}^{8} x_i}{p} - \frac{8 - \sum_{i=1}^{8} x_i}{1-p}$$

and is zero if and only if

$$\frac{\sum_{i=1}^{8} x_i}{p} - \frac{8 - \sum_{i=1}^{8} x_i}{1-p} = 0 \Longleftrightarrow p = \frac{\sum_{i=1}^{8} x_i}{8},$$

which is the same answer obtained previously using $\mathcal{L}(p)$. If we consider the second derivative, we will see that it is negative over all $0 < p < 1$ and so is always concave downwards. Thus the stationary point must be a maximum. □

8.4.2 Non-Smooth Likelihood Functions

Here calculus method alone cannot be used to locate the maximiser, and it is usually better to work with $\mathcal{L}(\theta)$ rather than $\ell(\theta)$.

Definition 8.4.1. Let $\mathcal{P}$ be a logical condition. Then the *indicator function* of $\mathcal{P}$, $\mathbb{I}(\mathcal{P})$ is given by

$$\mathbb{I}(\mathcal{P}) = \begin{cases} 1, & \text{if } \mathcal{P} \text{ is true} \\ 0, & \text{if } \mathcal{P} \text{ is false.} \end{cases}$$

The $\mathbb{I}$ notation allows one to write density functions in explicitly algebraic terms. For example,

$$f(x; \alpha, \beta) = \frac{e^{-x/\beta} x^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha}, \quad x > 0 \Longleftrightarrow f(x; \alpha, \beta) = \frac{e^{-x/\beta} x^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha} \mathbb{I}(x > 0).$$

Proposition 8.4.2. For any two logical conditions $\mathcal{P}$ and $\mathcal{Q}$,

$$\mathbb{I}(\mathcal{P} \cap \mathcal{Q}) = \mathbb{I}(\mathcal{P})\mathbb{I}(\mathcal{Q}).$$

Non-smooth likelihood functions arise when the range of f depends on θ.

Example 8.4.2. Suppose that the observation $x_1, \ldots, x_n$ come from pdf f, where

$$f(x; \theta) = 5(x^4/\theta^5), \quad 0 < x < \theta.$$

Use the $\mathbb{I}$ notation to find an expression for $\mathcal{L}(\theta)$, and simplify.

Solution.

$$f(x;\theta) = 5(x^4/\theta^5)\mathbb{I}(0 < x < \theta) = 5(x^4/\theta^5)\mathbb{I}(x > 0)\mathbb{I}(\theta > x).$$

The likelihood function is then

$$\begin{aligned}\mathcal{L}(\theta) &= \prod_{i=1}^{n} f(x_i;\theta)\\ &= 5(x^4/\theta^5)\mathbb{I}(x_1 > 0)\mathbb{I}(\theta > x_1)\cdots 5(x_n^4/\theta^5)\mathbb{I}(x_n > 0)\mathbb{I}(\theta > x_n)\\ &= 5^n\left(\prod_{i=1}^{n} x_i\right)^4 \left\{\prod_{i=1}^{n}\mathbb{I}(x_i > 0)\right\}\left\{\prod_{i=1}^{n}\mathbb{I}(\theta > x_i)\right\}\theta^{-5n}.\end{aligned}$$

Note that

$$\prod_{i=1}^{n}\mathbb{I}(\theta > x_i) = \mathbb{I}(\theta > x_1, \theta > x_2, \ldots, \theta > x_n) = \mathbb{I}(\theta > \max(x_1, \ldots, x_n)).$$

Also, $\prod_{i=1}^{n}\mathbb{I}(x_i > 0) = 1$ with probability 1, since $\mathbb{P}(X_i > 0) = 1$. Hence, the likelihood function is

$$\mathcal{L}(\theta) = 5^n\left(\prod_{i=1}^{n} x_i\right)^4 \theta^{-5n}\mathbb{I}(\theta > \max(x_1, \ldots, x_n))$$

or, even more digestibly,

$$\mathcal{L}(\theta) = \begin{cases} C_n\theta^{-5n}, & \theta > \max(x_1, \ldots, x_n)\\ 0, & \text{otherwise}\end{cases}$$

where $C_n = 5^n(\prod_{i=1}^{n} x_i)^4$.

Since $C_n\theta^{-5n}$ is clearly decreasing for $\theta > \max(x_1, \ldots, x_n)$ (can be verified using calculus) it is clear that $\mathcal{L}(\theta)$ attains its maximum at $\max(x_1, \ldots, x_n)$. Thus, the maximum likelihood estimator of θ is

$$\widehat{\theta} = \max(X_1, \ldots, X_n).$$

□

8.5 Properties of Maximum Likelihood Estimators

8.5.1 Consistency

Suppose

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x;\theta^*)$$

for some unknown θ^* in the interior of the set Θ, where Θ is the set of allowable values for θ. Let us have the assumptions

1. The domain of x does not depend on θ^*, that is, all pdfs $f(\cdot;\theta)$ have common support (are nonzero over the same fixed set, independent of θ^*).

2. If $\theta \neq \vartheta$, then
$$f(x;\theta) \neq f(x;\vartheta).$$
In other words, we can identify if $\vartheta = \theta$ from the equivalence of the densities $f(x;\vartheta)$ and $f(x;\theta)$.

3. The MLE
$$\widehat{\theta}_n = \operatorname*{argmax}_{\theta\in\Theta} f(\mathbf{X};\theta),$$
where $f(\mathbf{X};\theta) = \prod_{i=1}^{n} f(X_i;\theta)$, is unique and lies in the interior of Θ.

Theorem 8.5.1. *Under the assumptions above, the maximum likelihood estimator $\widehat{\theta}_n$ of θ^* is consistent; i.e.*

$$\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta^*.$$

Proof. Define the log-likelihood function based on the n data points

$$\ell_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ln f(X_i;\theta)$$

and note that by the Law of Large Numbers we have

$$\ell_n(\theta) \xrightarrow{\mathbb{P}} \ell(\theta) = \mathbb{E}_{\theta^*} \ln f(X;\theta) \quad (X \sim f(x;\theta^*)).$$

By assumption 3, we have that

$$\ell_n(\widehat{\theta}_n) > \ell_n(\theta)$$

for any $\theta \neq \widehat{\theta}_n$ with equality only when $\theta = \widehat{\theta}_n$. Similarly,

$$\begin{aligned}
\ell(\theta^*) - \ell(\theta) &= \mathbb{E}_{\theta^*} \ln f(X;\theta^*) - \ln f(X;\theta) \\
&= -\mathbb{E}_{\theta^*} \ln \frac{f(X;\theta)}{f(X;\theta^*)} \\
\text{(by Jensen's ineq. and 1)} &\geq -\ln \mathbb{E}_{\theta^*} \frac{f(X;\theta)}{f(X;\theta^*)} \\
&= -\ln 1 \\
&= 0.
\end{aligned}$$

Note that $\mathbb{E}_{\theta^*} \ln \frac{f(X;\theta)}{f(X;\theta^*)} = 0$ if and only if $f(X;\theta) = f(X;\theta^*)$. Hence, if $\theta \neq \theta^*$, then by assumption 2 we have $f(X;\theta) \neq f(X;\theta^*)$ and $-\mathbb{E}_{\theta^*} \ln \frac{f(X;\theta)}{f(X;\theta^*)} > 0$. Therefore, if $\theta \neq \theta^*$, then

$$\ell(\theta^*) > \ell(\theta)$$

with equality only if $\theta = \theta^*$. We now observe that

$$\ell_n(\theta^*) - \ell_n(\theta) = \underbrace{\ell_n(\theta^*) - \ell(\theta^*)}_{\xrightarrow{\mathbb{P}} 0} + \underbrace{\ell(\theta) - \ell_n(\theta)}_{\xrightarrow{\mathbb{P}} 0} + \underbrace{\ell(\theta^*) - \ell(\theta)}_{= c \geq 0}.$$

Hence, for $\theta \neq \theta^*$ we have $c > 0$ and

$$\ell_n(\theta^*) - \ell_n(\theta) \xrightarrow{\mathbb{P}} c > 0.$$

This can also be written as

$$\mathbb{P}(\ell_n(\theta^*) > \ell_n(\theta)) \to 1, \quad \theta \neq \theta^*.$$

Now note that by definition of MLE and its assumed uniqueness (assumption 3) we have

$$\{\omega : \widehat{\theta}_n(\omega)\} = \{\omega : \ell_n(\widehat{\theta}_n) \geq \ell_n(\theta) \text{ for all } \theta \in \Theta\}.$$

Hence, for any $\epsilon > 0$

$$\begin{aligned}
\mathbb{P}(|\widehat{\theta}_n - \theta^*| > \epsilon) &= \mathbb{P}(|\widehat{\theta}_n - \theta^*| > \epsilon, \ell_n(\widehat{\theta}_n) \geq \ell_n(\theta) \text{ for all } \theta \in \Theta) \\
\text{(by conjunction fallacy)} &\leq \mathbb{P}(|\widehat{\theta}_n - \theta^*| > \epsilon, \ell_n(\widehat{\theta}_n) \geq \ell_n(\theta^*)) \\
\text{(by conjunction fallacy)} &\leq \mathbb{P}(\ell_n(\theta) \geq \ell_n(\theta^*), \text{ for any } \theta \neq \theta^*) \\
&\leq 1 - \mathbb{P}(\ell_n(\theta^*) > \ell_n(\theta), \ \theta \neq \theta^*) \\
&\to 0.
\end{aligned}$$

Hence, by definition $\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta^*$. □

8.5.2 Equivariance

Maximum likelihood estimators are *equivariant* under functions of the parameter of interest:

Theorem 8.5.2. *Suppose $\widehat{\theta}$ is the maximum likelihood estimator of θ. Then for any function g*

$g(\widehat{\theta})$ is the maximum likelihood estimator of $g(\theta)$.

Example 8.5.1. Let $X_1, \ldots, X_n$ be random variables each with density function f, where

$$f(x;\theta) = 2\theta x e^{-\theta x^2}, \quad x > 0.$$

It has previously been shown that the maximum likelihood estimator of θ is

$$\widehat{\theta} = \frac{n}{\sum_i X_i^2}.$$

Find the maximum likelihood estimators of $\tau = 1/\theta$ and $\omega = \ln\theta$.

Solution. From the equivariance property of maximum likelihood estimation, the maximum likelihood estimator of $\tau = 1/\theta$ is

$$\widehat{\tau} = \frac{1}{\widehat{\theta}} = \frac{1}{n}\sum_i X_i^2$$

and the maximum likelihood estimator of $\omega = \ln\theta$ is

$$\widehat{\omega} = \ln\widehat{\theta} = \ln n - \ln\left(\sum_i X_i^2\right).$$

□

8.5.3 Variance and Standard Error

Let

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x;\theta)$$

be continuous random variables for some unknown $\theta \in \Theta$ and let

$$\ell_n(\theta) = \sum_{i=1}^n \ln f(X_i;\theta)$$

be the corresponding log-likelihood function (not normalised by n). In addition to the three technical assumptions we assume that

4. $f(\cdot;\theta)$ is twice differentiable in θ.
5. $\int f(x;\theta)dx$ can be twice differentiated under the integral.

Definition 8.5.1. The *Fisher score* is defined as

$$S_n(\theta) = \ell_n'(\theta)$$

and the *Fisher information* is defined as

$$I_n(\theta) = -\mathbb{E}_\theta \ell_n''(\theta),$$

where the expectation is with respect to $f(\cdot;\theta)$.

Theorem 8.5.3.

(i) $\mathbb{E}_\theta S_n(\theta) = 0$.

(ii) $I_n(\theta) = \mathbb{E}_\theta[\ell_n'(\theta)]^2 = \mathrm{Var}_\theta(S_n(\theta))$.

Proof.

(i) We have

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta}(1) \\ &= \frac{\partial}{\partial\theta}\left(\int f(x;\theta)dx\right) \\ &= \int \frac{\partial}{\partial\theta} f(x;\theta)dx. \end{aligned}$$

Hence,

$$\begin{aligned} \mathbb{E}_\theta S_n(\theta) &= n\mathbb{E}_\theta S_1(\theta), && \text{(by iid assumption)} \\ &= n\int f(x;\theta)\frac{\partial}{\partial\theta}\ln f(x;\theta)dx \\ &= n\int f(x;\theta)\frac{\frac{\partial}{\partial\theta}f(x;\theta)}{f(x;\theta)}dx \\ &= n\int \frac{\partial}{\partial\theta}f(x;\theta)dx \\ &= 0. \end{aligned}$$

(ii)

$$\begin{aligned} I_n(\theta) &= nI_1(\theta) \\ &= -n\int f(x;\theta)\frac{\partial^2}{\partial^2\theta}\ln f(x;\theta)dx \\ &= -n\int f(x;\theta)\frac{\frac{\partial^2 f}{\partial\theta^2}(x;\theta)f(x;\theta) - \frac{\partial f}{\partial\theta}(x;\theta)\frac{\partial f}{\partial\theta}(x;\theta)}{f^2(x;\theta)}dx \\ &= -n\int \frac{\partial^2 f}{\partial\theta^2}f(x;\theta)dx + n\int f(x;\theta)\frac{\frac{\partial f}{\partial\theta}(x;\theta)}{f(x;\theta)}\frac{\frac{\partial f}{\partial\theta}(x;\theta)}{f(x;\theta)}dx \\ &= 0 + n\mathbb{E}_\theta[S_1(\theta)]^2 \\ &= \mathrm{Var}_\theta(S_n(\theta)). \end{aligned}$$

□

Theorem 8.5.4. *Let $X_1, \dots, X_n$ be random variables with common density functions f depending on a parameter θ, and let $\widehat{\theta}_n$ be the maximum likelihood estimator of θ. Then as $n \to \infty$*

$$I_n(\theta)\operatorname{Var}(\widehat{\theta}_n) \xrightarrow{\mathbb{P}} 1.$$

Hence we can say that

$$\operatorname{se}(\widehat{\theta}) \approx \frac{1}{\sqrt{I_n(\widehat{\theta}_n)}}.$$

Example 8.5.2. Recall the previous example of a random sample $x_1, \dots, x_n$ from a common density function f, where

$$f(x;\theta) = 2\theta x e^{-\theta x^2}, \quad x \geq 0; \theta > 0.$$

Find the Fisher Information for θ, and hence the approximate $\operatorname{se}(\widehat{\theta})$.

Solution. From previously, the log-likelihood function, written as a random variable, is

$$\ell(\theta) = n \ln 2 + n \ln \theta + \sum_i \ln X_i - \theta \sum_i X_i^2.$$

The first and second derivatives of $\ell(\theta)$ are

$$\ell'(\theta) = n\theta^{-1} - \sum_i X_i^2 \quad \text{and} \quad \ell''(\theta) = -n\theta^{-2}.$$

Therefore, the Fisher information is

$$I_n(\theta) = -\mathbb{E}(-n\theta^{-2}) = n/\theta^2.$$

Hence the standard error of $\widehat{\theta}$ is approximately

$$\operatorname{se}(\widehat{\theta}) \approx \frac{1}{\sqrt{I_n(\widehat{\theta})}} = \frac{\widehat{\theta}}{\sqrt{n}}.$$

□

8.5.4 Asymptotic Normality

In addition to assumptions 1 through 5 we now add the following one

6. Assume the existence of a third derivative of $\ln f(x;\vartheta)$, such that

$$\left| \frac{\partial^3}{\partial \vartheta^3} \ln(f;\vartheta) \right| \leq g(x), \quad \text{for all } \vartheta \in \Theta$$

for some dominating function g with finite expectation $\mathbb{E}_\theta[g(X)] = \mu < \infty$.

Theorem 8.5.5 (Asymptotic Normality of Maximum Likelihood Estimators)**.** *Under the smoothness assumptions 1 through 6 above,*

$$\frac{\widehat{\theta} - \theta}{\sqrt{\operatorname{Var}(\widehat{\theta})}} \xrightarrow{\mathrm{d}} N(0,1) \quad \textit{and} \quad \frac{\widehat{\theta} - \theta}{\operatorname{se}(\widehat{\theta})} \xrightarrow{\mathrm{d}} N(0,1)$$

where

$$\operatorname{se}(\widehat{\theta}) = \frac{1}{\sqrt{I_n(\theta)}}.$$

Proof. By the mean value theorem we can write

$$\ell_n'(\widehat{\theta}_n) = \ell_n'(\theta) + \ell_n''(\vartheta_n)(\widehat{\theta}_n - \theta)$$

for some ϑ_n between θ and $\widehat{\theta}_n$. By the consistency of MLE we have $\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta$ and $\vartheta_n \xrightarrow{\mathbb{P}} \theta$. By the CLT, we have

$$\frac{\ell_n'(\theta) - n \times 0}{\sqrt{nI_1(\theta)}} = \frac{\ell_n'(\theta) - n\mathbb{E}_\theta S_1(\theta)}{\sqrt{nI_1(\theta)}} \xrightarrow{\mathrm{d}} Z \sim N(0,1).$$

Hence, since $\ell_n'(\widehat{\theta}_n) = 0$ (due to uniqueness and differentiability, the MLE is the critical point of the log-likelihood), we have

$$\begin{aligned}\frac{\sqrt{n}(\widehat{\theta}_n - \theta)}{I_1^{-1/2}(\theta)} &= \frac{\sqrt{n}}{I_1^{-1/2}(\theta)} \frac{-\ell_n'(\theta)}{\ell_n''(\vartheta_n)} \\ &= \frac{I_1(\theta)}{-\ell_n''(\vartheta_n)/n} \times \frac{\ell_n'(\theta)}{\sqrt{nI_1(\theta)}}.\end{aligned}$$

It seems plausible that $-\ell_n''(\vartheta_n)/n \xrightarrow{\mathbb{P}} I_1(\theta)$, because $\vartheta_n \xrightarrow{\mathbb{P}} \theta$ and $-\ell_n''(\theta)/n \xrightarrow{\mathbb{P}} I_1(\theta)$ (by the Law of Large Numbers). If we accept this, then the result follows from Slutsky's theorem:

$$\frac{\sqrt{n}(\widehat{\theta}_n - \theta)}{I_1^{-1/2}(\theta)} = \underbrace{\frac{I_1(\theta)}{-\ell_n''(\vartheta_n)/n}}_{\xrightarrow{\mathbb{P}} 1} \times \underbrace{\frac{\ell_n'(\theta)}{\sqrt{nI_1(\theta)}}}_{\xrightarrow{\mathrm{d}} Z} \xrightarrow{\mathrm{d}} Z \sim N(0,1).$$

□

This result is important because it means that maximum likelihood is not only useful for estimation, but is also a method for making inferences about parameters. Because we now know how to find the approximate distribution of any maximum likelihood estimator $\widehat{\theta}$, we can now calculate standard errors and construct confidence intervals for θ using $\widehat{\theta}$ for data from any family of distributions of known form.

Example 8.5.3. Recall the previous example:

$$X_1, \ldots, X_n \sim f, \quad \text{where} \quad f(x;\theta) = 2\theta x e^{-\theta x^2}, \; x \geq 0; \theta > 0.$$

Find the estimated standard error of $\widehat{\theta}$, and the approximate distribution of $\widehat{\theta}$.

Solution. We previously found the Fisher information to be

$$I_n(\theta) = n/\theta^2.$$

Therefore, the approximate variance of $\widehat{\theta}$ is

$$\mathrm{Var}(\widehat{\theta}) \approx \frac{\theta^2}{n}$$

and the asymptotic standard error of $\widehat{\theta}$ is

$$\widehat{\mathrm{se}}(\widehat{\theta}) = \frac{1}{\sqrt{I_n(\widehat{\theta})}} = \widehat{\theta}/\sqrt{n}.$$

Thus

$$\frac{\widehat{\theta} - \theta}{\theta/\sqrt{n}} \xrightarrow{\mathrm{d}} N(0,1).$$

This means that we can approximate the distribution of $\widehat{\theta}$ using

$$\widehat{\theta} \overset{\text{appr.}}{\sim} N(\theta, \theta^2/n)$$

□

Proposition 8.5.1. Under appropriate regularity conditions, including the existence of two derivatives of $\mathcal{L}(\theta)$, if $\tau = g(\theta)$ and $\widehat{\tau} = g(\widehat{\theta})$, where g is differentiable and $g'(\theta) \neq 0$, then

$$\frac{\widehat{\tau} - \tau}{\sqrt{\text{Var}(\widehat{\tau})}} \xrightarrow{\text{d}} N(0, 1)$$

where

$$\text{Var}(\widehat{\tau}) \approx \frac{[g'(\theta)]^2}{I_n(\theta)}.$$

Proof. The result follows directly from the delta method. □

Example 8.5.4. Recall the example:

$$X_1, \ldots, X_n \sim f, \quad \text{where} \quad f(x; \theta) = 2\theta x e^{-\theta x^2},\ x \geq 0; \theta > 0.$$

Suppose that the parameter of interest is $\omega = \ln \theta$. Use maximum likelihood estimation to find an estimator of ω and its appropriate distribution.

Solution. From previous working, the maximum likelihood estimator of ω is

$$\widehat{\omega} = \ln n - \ln \left(\sum_i X_i^2 \right)$$

and the variance of $\widehat{\omega}$ is

$$\text{Var}(\widehat{\omega}) \approx \frac{|1/\theta^2|}{n/\theta^2} = \frac{1}{n}.$$

Thus,

$$\frac{\widehat{\omega} - \omega}{1/\sqrt{n}} \xrightarrow{\text{d}} N(0, 1).$$

□

8.5.5 Asymptotic Optimality

In the case of smooth likelihood functions where asymptotic normality results can be derived it is possible to argue that, asymptotically, the maximum likelihood estimator is *optimal* or *best*.

Theorem 8.5.6 (Lower Bound on Variance). *Let*

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x; \theta), \quad \theta \in \Theta$$

and suppose that the maximum likelihood estimator $\widehat{\theta}_n$ is asymptotically normal; i.e.

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{\text{d}} N(0, [I_1(\theta)]^{-1}).$$

Let $\tilde{\theta}_n = g(X_1, \ldots, X_n)$ be any other estimator of θ with

$$\mathbb{E}_\theta g(X_1, \ldots, X_n) = \mu_n(\theta).$$

Then,

$$\text{Var}_\theta(\tilde{\theta}_n) \geq \frac{(\mu_n'(\theta))^2}{n I_1(\theta)}.$$

Proof. Assume conditions 1 through 5 and denote the joint pdf of $X_1, \ldots, X_n$ as

$$f(\mathbf{X}; \theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

and note that the log-likelihood can be written in terms of the joint pdf:

$$\ell_n'(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \ln f(X_i;\theta) = \frac{\partial}{\partial\theta} \ln f(\mathbf{X};\theta).$$

We have

$$\begin{aligned}
\mathrm{Var}_\theta(\tilde{\theta}_n) I_n(\theta) &= \mathbb{E}_\theta(\tilde{\theta}_n - \mu_n)^2 \mathbb{E}_\theta(\ell_n'(\theta))^2 \\
&\geq \left(\mathbb{E}_\theta(\tilde{\theta}_n - \mu_n)\frac{\partial}{\partial\theta}\ln f(\mathbf{X};\theta)\right)^2 \qquad \text{(using } \mathbb{E}[X^2]\mathbb{E}[Y^2] \geq (\mathbb{E}[XY])^2\text{)} \\
&\geq \left(\int f(\mathbf{x};\theta)(\tilde{\theta}_n - \mu_n)\frac{\frac{\partial}{\partial\theta} f(\mathbf{x};\theta)}{f(\mathbf{x};\theta)} d\mathbf{x}\right)^2 \\
&\geq \left(\int \frac{\partial f}{\partial\theta}(\mathbf{x};\theta)(\tilde{\theta}_n - \mu_n) d\mathbf{x}\right)^2 \\
&\geq \left(\int \frac{\partial f}{\partial\theta}(\mathbf{x};\theta)\tilde{\theta}_n d\mathbf{x} - \mu_n \int \frac{\partial f}{\partial\theta}(\mathbf{x};\theta) d\mathbf{x}\right)^2 \\
\text{(via assumption 5)} &\geq \left(\frac{d}{d\theta}\underbrace{\mathbb{E}_\theta[\tilde{\theta}_n]}_{\mu_n} - \mu_n \underbrace{\frac{d}{d\theta}\int f(\mathbf{x};\theta) d\mathbf{x}}_{\frac{d}{d\theta}(1)}\right)^2 \\
&= (\mu_n'(\theta))^2.
\end{aligned}$$

Therefore, rearranging the inequality

$$\mathrm{Var}_\theta(\tilde{\theta}_n) \geq \frac{(\mu_n'(\theta))^2}{I_n(\theta)}.$$

□

Corollary 8.5.1 (Cramer-Rao Lower Bound)**.** *If $\tilde{\theta}_n = g(X_1, \ldots, X_n)$ is an unbiased estimator of θ, then*

$$\mathrm{Var}_\theta(\tilde{\theta}_n) \geq \frac{1}{nI_1(\theta)}.$$

Since, for the MLE $\widehat{\theta}$ we have

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{\mathrm{d}} N(0, I_1^{-1}(\theta)),$$

the maximum likelihood estimator can (under some technical conditions) achieve asymptotically (as $n \to \infty$) the smallest possible variance in estimating θ.

Example 8.5.5. Recall the example:

$$X_1, \ldots, X_n \sim f, \quad \text{where} \quad f(x;\theta) = 2\theta x e^{-\theta x^2}, \quad x \geq 0; \theta > 0.$$

We have shown previously that if $\widehat{\theta}$ is the maximum likelihood estimator of θ, then

$$\widehat{\theta} = n \Big/ \sum_i X_i^2 \quad \text{and} \quad \mathrm{Var}(\widehat{\theta}) \approx \frac{1}{I_n(\theta)} = \theta^2/n.$$

It is known that

$$\mathbb{E}(X) = \frac{1}{2}\sqrt{\frac{\pi}{\theta}} \quad \text{and} \quad \mathrm{Var}(X) = \frac{1}{\theta}\left(1 - \frac{\pi}{4}\right).$$

(a) Find the method of moments estimator of $\theta, \tilde{\theta}$.

(b) Find an expression for the approximate distribution of $\overline{X}$.

(c) Hence use the delta method to find the approximate distribution of $\tilde{\theta}$.

(d) Compare the asymptotic properties of the estimators $\tilde{\theta}$ and $\widehat{\theta}$. Which is the better estimator?

Solution. Rearranging $(\mathbb{E}[X])^2 = \frac{1}{4}\frac{\pi}{\theta}$ we obtain the method of moments estimator:

$$\tilde{\theta} = \frac{\pi}{4(\overline{X})^2} = g(\overline{X}),$$

where $g(x) = \pi/(4x^2)$ with $g'(x) = -\frac{\pi}{2x^3}$ and $[g'(x)]^2 = \frac{\pi^2}{4x^6}$. We now derive its asymptotic distribution using the delta method. From the CLT

$$\sqrt{n}(\overline{X} - \sqrt{\pi/(4\theta)}) \xrightarrow{\mathrm{d}} N\left(0, \frac{1-\pi/4}{\theta}\right).$$

Hence,

$$\sqrt{n}(g(\overline{X}) - g(\sqrt{\pi/(4\theta)})) \xrightarrow{\mathrm{d}} N\left(0, \underbrace{\operatorname{Var}(X) \times \frac{4^2\theta^3}{\pi}}_{\theta^2(\frac{16}{\pi}-4)}\right).$$

Now from the MLE asymptotic theory above we know that

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{\mathrm{d}} N(0, \theta^2).$$

Thus, we compare θ^2 versus $\theta^2(\frac{16}{\pi} - 4)$. Since

$$\frac{16}{\pi} - 4 = 1.092955817\ldots > 1$$

we conclude that the MLE estimator beats the method of moments estimator, because the method of moments estimator has asymptotic variance roughly 9% higher than the variance of the MLE. □

8.6 Likelihood-Based Confidence Intervals

Because MLEs are asymptotically normal, we can construct confidence intervals easily.

Theorem 8.6.1 (Wald Confidence Intervals)**.** *Let $X_1, \ldots, X_n$ be random variables with common density function f, where*

$$f(x) = f(x;\theta), \quad \theta \in \Theta$$

and let $\widehat{\theta}$ be the MLE of θ. Under the regularity conditions for which θ is asymptotically normal

$$\left(\widehat{\theta}_n - z_{1-\alpha/2}\operatorname{se}(\widehat{\theta}), \widehat{\theta} + z_{1-\alpha/2}\operatorname{se}(\widehat{\theta})\right)$$

is an approximate $1-\alpha$ confidence interval for θ for large n, where $\operatorname{se}(\widehat{\theta}) = 1/\sqrt{I_n(\theta)}$.

Example 8.6.1. Recall the example of a sample $x_1, \ldots, x_n$ from the density function f, where

$$f(x;\theta) = 2\theta x e^{-\theta x^2}, \quad x \geq 0; \theta > 0.$$

(a) Derive a formula for a 95% confidence interval for θ.

(b) Use the following data to find an appropriate 95% confidence interval for θ, assuming that it is a random sample from a r.v. with the density function above.

```
0.366 0.568 0.300 0.115 0.204 0.128 0.277 0.391 0.328 0.451
0.412 0.190 0.207 0.147 0.116 0.326 0.256 0.524 0.217 0.485
0.265 0.375 0.267 0.360 0.250 0.258 0.583 0.413 0.481 0.468
0.406 0.336 0.305 0.321 0.268 0.361 0.632 0.283 0.258 0.466
0.276 0.232 0.133 0.316 0.468 0.496 0.573 0.523 0.256 0.491
0.127 0.054 0.440 0.228 0.249 0.754 0.430 0.111 0.459 0.233
0.257 0.640 0.147 0.273 0.112 0.389 0.126 0.356 0.273 0.296
0.433 0.253 0.234 0.514 0.177 0.221 0.534 0.509 0.510 0.269
0.262 0.625 0.183 0.541 0.705 0.078 0.847 0.149 0.031 0.453
0.299 0.226 0.069 0.211 0.195 0.381 0.317 0.467 0.289 0.593
```

Solution.

(a) Recall that the maximum likelihood estimate and corresponding standard error are:

$$\widehat{\theta} = \frac{n}{\sum_{i=1}^{n} x_i^2} \quad \text{and} \quad \widehat{\text{se}}(\widehat{\theta}) = \frac{\widehat{\theta}}{\sqrt{n}} = \frac{\sqrt{n}}{\sum_{i=1}^{n} x_i^2}.$$

For a 95% confidence interval the appropriate $N(0,1)$ quantile is

$$z_{0.975} = 1.96.$$

An approximate Wald 95% confidence interval for θ is then:

$$\left(\frac{n}{\sum_{i=1}^{n} x_i^2} - 1.96 \frac{\sqrt{n}}{\sum_{i=1}^{n} x_i^2}, \frac{n}{\sum_{i=1}^{n} x_i^2} + 1.96 \frac{\sqrt{n}}{\sum_{i=1}^{n} x_i^2} \right).$$

(b) For these data $(x_1, \ldots, x_{100})$ we have $\sum_{i=1}^{100} x_i^2 = 14.018$. So using the formula derived previously, an approximate 95% confidence interval for θ is

$$\left(\frac{100}{14.018} - 1.96 \frac{\sqrt{100}}{14.018}, \frac{100}{14.018} + 1.96 \frac{\sqrt{100}}{14.018} \right) = (5.17, 9.09).$$

□

Theorem 8.6.2. *Under the same conditions as the previous result, with* $\tau = g(\theta)$ *and* $\widehat{\tau} = g(\widehat{\theta})$,

$$\lim_{n\to\infty} \mathbb{P}(\widehat{\tau} - z_{1-\alpha/2}\,\text{se}(\widehat{\tau}) < \tau < \widehat{\tau} + z_{1-\alpha/2}\,\text{se}(\widehat{\tau})) = 1 - \alpha$$

where $\text{se}(\widehat{\tau}) = |g'(\theta)|/\sqrt{I_n(\theta)}$. *Therefore,*

$$(\widehat{\tau} - z_{1-\alpha/2}\,\text{se}(\widehat{\tau}), \widehat{\tau} + z_{1-\alpha/2}\,\text{se}(\widehat{\tau}))$$

is an approximate $1 - \alpha$ *confidence interval for* τ *for large* n.

This result is a confidence interval version of the delta method result.

8.7 Multi-parameter Maximum Likelihood Inference

In multi-parameter models such as $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ the maximum likelihood principle still applies. Instead of maximising over a single variable, the maximisation is performed simultaneously over several variables.

Example 8.7.1. Consider the model

$$X_1, \dots, X_n \sim N(\mu, \sigma^2), \quad -\infty < \mu < \infty, \sigma > 0.$$

Find the maximum likelihood estimators of μ and σ.

Solution. The log-likelihood function is

$$\ell(\mu, \sigma) = \ln \prod_{i=1}^{n} \left[\frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-\mu)^2/(2\sigma^2)} \right] = -\frac{1}{2\sigma^2} \sum_i (x_i - \mu)^2 - \frac{n}{2} \ln 2\pi - n \ln \sigma.$$

Then,

$$\frac{\partial}{\partial \mu} \ell(\mu, \sigma) = \frac{1}{\sigma^2} \sum_i (x_i - \mu) = \frac{n}{\sigma^2} (\overline{x} - \mu) = 0$$

if and only if

$$\mu = \overline{x}$$

and regardless of the value of σ. Also,

$$\frac{\partial}{\partial \sigma} \ell(\mu, \sigma) = \sigma^{-3} \sum_i (x_i - \mu)^2 - n\sigma^{-1} = 0$$

if and only if

$$\sigma = \sqrt{\frac{1}{n} \sum_i (x_i - \mu)^2}.$$

The unique stationary point of $\ell(\mu, \sigma)$ is then

$$(\widehat{\mu}, \widehat{\sigma}) = \left(\overline{x}, \sqrt{\frac{1}{n} \sum_i (x_i - \overline{x})^2} \right).$$

Analysis of the second order partial derivatives can be used to show that this is the global maximiser of $\ell(\mu, \sigma)$ over $\mu \in \mathbb{R}$ and $\sigma > 0$. Hence, the maximum likelihood estimators of μ and σ are

$$\widehat{\mu} = \overline{X} \quad \text{and} \quad \widehat{\sigma} = \sqrt{\frac{1}{n} \sum_i (X_i - \overline{X})^2}.$$

□

Definition 8.7.1. Let $\theta = (\theta_1, \dots, \theta_k)$ be the vector of parameters in a multi-parameter model. The *Fisher information matrix* is given by

$$I_n(\theta) = - \begin{bmatrix} \mathbb{E}(H_{11}) & \mathbb{E}(H_{12}) & \cdots & \mathbb{E}(H_{1k}) \\ \mathbb{E}(H_{21}) & \mathbb{E}(H_{22}) & \cdots & \mathbb{E}(H_{2k}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}(H_{k1}) & \mathbb{E}(H_{k2}) & \cdots & \mathbb{E}(H_{kk}) \end{bmatrix}$$

where

$$H_{ij} = \frac{\partial^2}{\partial \theta_i \partial \theta_j} \ell(\theta).$$

Example 8.7.2. Consider the model

$$X_1, \dots, X_n \sim N(\mu, \sigma^2), \quad -\infty < \mu < \infty, \sigma > 0.$$

Find the Fisher Information matrix for μ and σ.

Solution. It was shown previously that the first order partial derivatives are

$$\begin{aligned}\frac{\partial}{\partial\mu}\ell(\mu,\sigma) &= \frac{n}{\sigma^2}(\overline{x}-\mu)\\ \frac{\partial}{\partial\sigma}\ell(\mu,\sigma) &= \sigma^{-3}\sum_i(x_i-\mu)^2 - n\sigma^{-1}.\end{aligned}$$

The second partial derivatives are then

$$\begin{aligned}\frac{\partial^2}{\partial\mu^2}\ell(\mu,\sigma^2) &= -\frac{n}{\sigma^2}\\ \frac{\partial^2}{\partial\mu\partial\sigma}\ell(\mu,\sigma) &= -\frac{2n}{\sigma^3}(\overline{x}-\mu)\\ \frac{\partial^2}{\partial\sigma^2}\ell(\mu,\sigma) &= -3\sigma^{-4}\sum_i(x_i-\mu)^2 + n\sigma^{-2}.\end{aligned}$$

Noting that $\mathbb{E}(X_i)=\mu$ and $\mathbb{E}(X_i-\mu)^2=\sigma^2$ for each i we then get

$$\begin{aligned}\mathbb{E}\left[\frac{\partial^2}{\partial\mu^2}\ell(\mu,\sigma)\right] &= -\frac{n}{\sigma^2}\\ \mathbb{E}\left[\frac{\partial^2}{\partial\mu\partial\sigma}\ell(\mu,\sigma)\right] &= 0\\ \mathbb{E}\left[\frac{\partial^2}{\partial\sigma^2}\ell(\mu,\sigma)\right] &= -3\sigma^{-4}n\sigma^2 + n\sigma^{-2} = -2n\sigma^{-2}.\end{aligned}$$

The Fisher information matrix is then

$$I_n(\mu,\sigma^2) = -\begin{bmatrix}-n/\sigma^2 & 0\\ 0 & -2n\sigma^{-2}\end{bmatrix} = \begin{bmatrix}n/\sigma^2 & 0\\ 0 & 2n/\sigma^2\end{bmatrix}$$

□

Given the Fisher information matrix, we can find the asymptotic distribution of any function of a vector of maximum likelihood estimators $\widehat{\tau}=g(\widehat{\theta})$.

Definition 8.7.2. Let $\theta=(\theta_1,\ldots,\theta_k)$ be the vector of parameters in a multi-parameter model and $g(\theta)=g(\theta_1,\ldots,\theta_k)$ be a real-valued function. The *gradient vector* of g is given by

$$\nabla g(\theta) = \begin{bmatrix}\frac{\partial g(\theta)}{\partial g_1}\\ \vdots\\ \frac{\partial g(\theta)}{\partial\theta_k}\end{bmatrix}.$$

Theorem 8.7.1. *Let $\tau=g(\theta)$ be a real-valued function of $\theta=(\theta_1,\ldots,\theta_k)$, with maximum likelihood estimate $\widehat{\theta}$ and $\widehat{\tau}=g(\widehat{\theta})$. Under appropriate regularity conditions, including the existence of all second order partial derivatives of $\mathcal{L}(\theta)$ and first order partial derivatives of g, as $n\to\infty$.*

$$\frac{\widehat{\tau}-\tau}{\mathrm{se}(\widehat{\tau})} \xrightarrow{\mathrm{d}} N(0,1)$$

where

$$\mathrm{se}(\widehat{\tau}) = \sqrt{\nabla g(\theta)^T I_n(\theta)^{-1}\nabla g(\theta)}.$$

This is a multi-parameter extension of the delta method.

Theorem 8.7.2. *Under appropriate regularity conditions, including the existence of all second order partial derivatives of $\mathcal{L}(\theta)$, if $\tau = g(\theta)$ where each component of g is differentiable then:*

$$\lim_{n\to\infty} \mathbb{P}(\widehat{\tau} - z_{1-\alpha/2}\,\mathrm{se}(\widehat{\tau}) < \tau < \widehat{\tau} + z_{1-\alpha/2}\,\mathrm{se}(\widehat{\tau})) = 1 - \alpha$$

where

$$\mathrm{se}(\widehat{\tau}) = \sqrt{\nabla g(\theta)^T I_n(\theta)^{-1} \nabla g(\theta)}.$$

Therefore,

$$(\widehat{\tau} - z_{1-\alpha/2}\,\mathrm{se}(\widehat{\tau}_n), \widehat{\tau} + z_{1-\alpha/2}\,\mathrm{se}(\widehat{\tau}))$$

is an appropriate $1 - \alpha$ confidence interval for τ for large n.

9. Hypothesis Testing

9.1 Stating the Hypotheses

Definition 9.1.1. The *null hypothesis* labelled H_0, is a claim that a parameter of interest to us (θ) takes a particular value (θ_0). Hence H_0 has the form $\theta = \theta_0$ for some pre-specified value θ_0. The *alternative hypothesis*, labelled H_1, is a more general hypothesis about the parameter of interest to us, which we will accept to be true if the evidence against the null hypothesis is strong enough. The form of H_1 tends to be one of the following:

$$\begin{aligned} H_1 &: \theta \neq \theta_0 \\ H_1 &: \theta > \theta_0 \\ H_1 &: \theta < \theta_0. \end{aligned}$$

In a hypothesis test, we use our data to test H_0, by measuring how much evidence our data offer against H_0 in favour of H_1.

Example 9.1.1. Recall that the Mythbusters were testing whether or not toast lands butter side down more often than butter side up. In 24 trials, they found that 14 slices of bread landed butter side down. State the null and alternative hypotheses.

Solution. Here we have a sample from a binomial distribution, where the binomial parameter p is the probability that a slice of toast will land butter side down. We are interested in whether there is evidence that the parameter p is larger than 0.5.

$$H_0 : p = 0.5, \quad H_1 : p > 0.5.$$

□

9.2 The Process of Hypothesis Testing

A hypothesis test has the following steps.

1. State the null hypothesis (H_0) and the alternative hypothesis (H_1). By convention, the null hypothesis is the more specific of the two hypotheses.

2. Then we use our data to answer the question:

   "How much evidence is there against the null hypothesis?"

   A common way to achieve this is to:

   (i) Find a test statistic that measures how "far" our data are from what is expected under the null hypothesis. (You must know the approximate distribution of this test statistic assuming H_0 to be true. This is called the *null distribution*).

   (ii) Calculate a P-value, a probability that measures how much evidence there is against the null hypothesis, for the data we observed. A P-value is defined as the probability of observing a test statistic value as or more unusual than the one we observed, if the null hypothesis were true.

3. Reach a conclusion. A helpful way to think about the conclusion is to return to our original question:

“How much evidence is there against H_0?”

Example 9.2.1. The Mythbusters example can be used to illustrate the above steps of a hypothesis test.

1. Let

$$p = \mathbb{P}(\text{Toast lands butter side down}).$$

Then what we want to do is choose between the following two hypotheses:

$$H_0 : p = \frac{1}{2} \quad \text{versus} \quad H_1 : p > \frac{1}{2}.$$

2. We want to answer the question “How much evidence (if any) does our sample (14 of 24 land butter side down) give us against the claim $p = 0.5$?”

   (a) To answer this question, we will consider $\widehat{p}$, the sample proportion, and in particular we will look at the test statistic

$$Z = \frac{\widehat{p} - p}{\sqrt{p(1-p)/n}} \xrightarrow{\text{d}} N(0,1).$$

   Under the null hypothesis,

$$Z = \frac{\widehat{p} - 0.5}{\sqrt{0.5(1-0.5)/n}} \xrightarrow{\text{d}} N(0,1).$$

   This statistic measures how far our sample data are with H_0. The further $\widehat{p}$ is from 0.5, the further Z is from 0.

   (b) To find out if $\widehat{p} = \frac{14}{24}$ is unusually large, if $p = 0.5$, we can calculate

$$\mathbb{P}\left(\widehat{p} \geq \frac{14}{24}\right) \approx \mathbb{P}\left(Z \geq \frac{\frac{14}{24} - 0.5}{\sqrt{0.5(1-0.5)/n}}\right) \approx \mathbb{P}(Z > 0.82) \approx 0.2071.$$

3. So we can say that we would expect p to be at least as large as $\frac{14}{24}$ quite often (22% of the time) due to sample variation alone. Observing an event of probability 0.22 is not particularly surprising, so we conclude that we have no evidence against the claim that $p = 0.5$ - because our data are consistent with this hypothesis.

9.3 Interpreting P-Values

The most common way that a hypothesis test is conducted and reported is via the calculation of a P-value.

Definition 9.3.1. The P-value of an observed test statistic is

P-value $= \mathbb{P}$(observing a test statistic as or more “extreme” than the observed test statistic when H_0 is true).

The following are some rough guidelines on interpreting P-values. Note though that the interpretation of P-values should depend on the context to some extent, so the below should be considered as a guide only and not as strict rules.

Range of P-value	Conclusion
P-value≥ 0.1	little or no evidence against H_0
$0.01 \leq P$-value < 0.1	some, but inconclusive evidence against H_0
$0.001 \leq P$-value < 0.01	evidence against H_0
P-value < 0.001	strong evidence against H_0

It is common for people to use 0.05 as a "cut-off" between a significant finding ($P < 0.05$) and a non-significant finding ($P > 0.05$). Nevertheless, it is helpful to keep in mind that P is continuous and our interpretation of P should reflect this - a P-value of 0.049 (just less than 0.05) is hardly different from a P-value of 0.051 (just larger than 0.05), so there should be little difference in interpretation of these values. In contrast, a P-value of 0.049 offers less evidence against H_0 than a P-value of 0.0001.

9.4 Tests for Normal Samples

In the situation where

$$X_1, \ldots, X_n \sim N(\mu, \sigma^2)$$

it is possible to perform exact hypothesis tests for the parameter μ using the main result from chapter 6:

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

For testing the hypotheses

$$H_0 : \mu = \mu_0 \quad \text{versus any of} \quad \begin{cases} H_1 : \mu < \mu_0 \\ H_1 : \mu \neq \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$$

the appropriate test statistic

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

which as a t_{n-1} distribution under the null hypothesis. Using this test statistic in a hypothesis test about μ is known as a *t-test* (or a one-sample t-test).

Example 9.4.1. Before the installation of new machinery, the daily yield of fertiliser produced by a chemical plant had a mean $\mu = 880$ tonnes. Some new machinery was installed, and we would like to know if the new machinery is more efficient (i.e. if $\mu > 880$). During the first $n = 50$ days of operation of the new machinery, the yield of fertiliser was recorded. The sample mean was $\overline{x} = 888$ with a standard deviation $s = 21$. Is there evidence that the new machinery is more efficient? Use a hypothesis test to answer this question, assuming that yield is approximately normal.

Solution. Using the hypothesis testing steps given previously:

1. Our null and alternative hypotheses are:

$$H_0 : \mu = 880 \quad H_1 : \mu > 880.$$

2. We estimate μ using the sample mean $\overline{X}$. Now we want to know if our $\overline{X}$ is so far from $\mu = 880$ that is provides evidence against H_0.

   (i) We can construct a test statistic on $\overline{X}$ using

   $$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

   which is under the null hypothesis becomes

   $$T = \frac{\overline{X} - 880}{S/\sqrt{n}} \sim t_{n-1}.$$

   Larger observed values of T suggest more evidence against H_0.

   (ii) We now would like to calculate a P-value that measures how unusual it is to get a mean as large as $\overline{x} = 888$, if the true mean were $\mu = 880$:

   $$\mathbb{P}\left(T > \frac{\overline{x} - 880}{s/\sqrt{n}}\right) = \mathbb{P}\left(T > \frac{888 - 880}{21/\sqrt{50}}\right) = \mathbb{P}(T > 2.69)$$

   where $T \sim t_{49}$. From tables, $0.0025 < P\text{-value} < 0.005$.

3. This tells us that, if H_0 were true, we would be highly unlikely to observe a T statistic as large as 2.7 (hence a mean yield as high as 888 tonnes) by chance alone. We have strong evidence against the claim that $\mu = 880$. Alternatively, you could say that "the mean is significantly different from 880".

□

9.5 One-Sided and Two-Sided Tests

Definition 9.5.1. A *one-sided* hypothesis test about a parameter θ is either of the form:

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta < \theta_0$$

or

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta > \theta_0.$$

A *two-sided* hypothesis test about θ is of the form

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0.$$

A two-sided test is sometimes called a two-tailed test, because we calculate the P-value using both tails of the null distribution of the test statistic.

Example 9.5.1. A popular brand of yoghurt claims to contain 120 calories per serving. A consumer watchdog group randomly sampled 14 servings of the yoghurt and obtained the following numbers of calories per serving:

160	200	220	230	120	180	140
130	170	190	80	120	100	170

Conduct a hypothesis test to test if the manufacturer's claim is correct.

Solution. The hypotheses to be tested are:

$$H_0 : \mu = 120 \quad \text{(mean number of calories matches manufacturer's claim)}$$

versus

$$H_1 : \mu \neq 120 \quad \text{(mean number of calories differs from manufacturer's claim)}$$

where

$$\mu = \text{mean calorie value of a serving of the yoghurt.}$$

The test statistic we will use is

$$\frac{\overline{X} - 120}{S/\sqrt{14}}$$

which has a t_{13} distribution if the null hypothesis is true. The observed value is

$$\frac{\overline{x} - 120}{s/\sqrt{14}} = \frac{157.8571 - 120}{44.75206/\sqrt{14}} = 3.1651\ldots$$

Then, using the t-distribution tables,

$$\begin{aligned} P\text{-value} &= \mathbb{P}_{\mu=120}\left(\left|\frac{\overline{X} - 120}{S/\sqrt{14}}\right| > 3.1651\right) \\ &= 2\mathbb{P}(T > 3.1651), \quad T \sim t_{13} \\ &\approx 0.007 \end{aligned}$$

so there is very strong evidence against H_0. The consumer watchdog group should conclude that the manufacturer's calorie claim is not correct. □

9.6 Rejection Regions

Instead of using P-values to draw a conclusion based on data, we can use rejection regions.

Definition 9.6.1. The *rejection region* is the set of values of the test statistic for which H_0 is rejected in favour of H_1.

The term "rejection region" comes from the fact that often people speak of "rejecting H_0" (if our data provide evidence against H_0) or "retaining H_0" (if our data do not provide evidence against H_0). If our test statistic is in the rejection region we reject H_0, if the test statistic is not in the rejection region we retain H_0.

To determine a rejection region we first choose a size or *significance level* for the test, which can be defined as the P-value at which we would start rejection H_0. It should be set to a small number (typically 0.05), and by convention is usually denoted by α. Once we have determined the desired size of the test, we can then derive the rejection region.

Example 9.6.1. Recall the fertiliser example - we wanted to test

$$H_0 : \mu = 880 \quad \text{versus} \quad H_1 : \mu > 880.$$

Our test statistic is

$$T = \frac{\overline{X} - 880}{S/\sqrt{n}} \sim t_{n-1}$$

and we have a sample size of 50.

(a) Find a rejection region for a test size of 0.05

(b) Hence test H_0 versus H_1.

Solution.

(a) From tables, if $T \sim t_{49}$, then

$$\mathbb{P}(T > 1.676) \approx 0.05$$

and so our rejection region is $T > 1.676$. In other words, if our observed value of T is greater than 1.676, we will reject H_0 in favour of H_1. Alternatively, if $T < 1.676$, we retain H_0.

(b) Our observed value of T was 2.69 which is in our rejection region meaning we reject H_0 and conclude that there is evidence that $\mu > 880$.

□

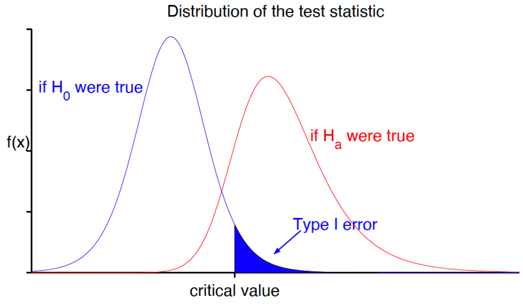
9.7 Type I and Type II Error

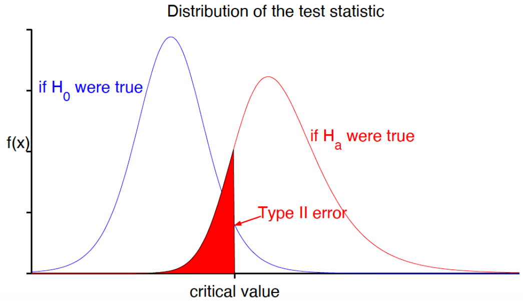
How could we choose a significance level α? This problem can be answered by considering the possible errors that can be made in reaching our decision.

Definition 9.7.1. *Type I error* corresponds to rejection of the null hypothesis when it is really true.

Type II error corresponds to acceptance of the null hypothesis when it is really false.

	reject H_0	accept H_0
H_0 true	Type I error	No error
H_0 false	No error	Type II error





Clearly we would like to avoid both Type I and II errors, however, there is always a chance that either will occur so the idea is to reduce these chances as much as possible. Unfortunately, making the probability of one type of error small has the effect of making the probability of the other large.

Definition 9.7.2. The *size* or *significance level* of a test is the probability of committing a Type I error. It is usually denoted by α. Therefore,

$$\begin{aligned}\alpha &= \text{size}\\ &= \text{significance level}\\ &= \mathbb{P}(\text{committing Type I error})\\ &= \mathbb{P}(\text{reject } H_0 \text{ when } H_0 \text{ is true}).\end{aligned}$$

A popular choice of α is $\alpha = 0.05$. This corresponds to the following situation:

"We have set up our test in such a way that if we do reject H_0 then there is only a 5% chance that we will wrongfully do so".

There is nothing special about $\alpha = 0.05$.

Example 9.7.1. If we want to test the following hypotheses about a new vaccine:

H_0 : vaccine is perfectly safe.
versus
H_1 : vaccine has harmful effects

then it is important to minimise Type II error - we want to detect any harmful side effects that are present. To assist in minimising Type II error, we might be prepared to accept a reasonably high Type I error (such as 0.1 or maybe even 0.25).

Example 9.7.2. Suppose we are testing a potential water source for toxic levels of heavy metals. If the hypotheses are

H_0 : the water has toxic levels of heavy metals
versus
H_1 : the water is OK to drink

then it is important to minimise type I error - we won't want to make the mistake of saying that toxic water is OK to drink. So we would want to choose a low Type I error, maybe 0.001 say.

9.8 Power of a Statistical Test

It is important for a particular hypothesis test procedure to have a small significance level. However, it is also important to realise that this is not the only property that a good hypothesis test should have.

Definition 9.8.1. Broadly speaking, a test procedure with good power properties, usually referred to as a "powerful test", is one that has a good chance of rejecting H_0 when it is not true.

Therefore, a test with high power is able to detect deviation from the null hypothesis with high probability.

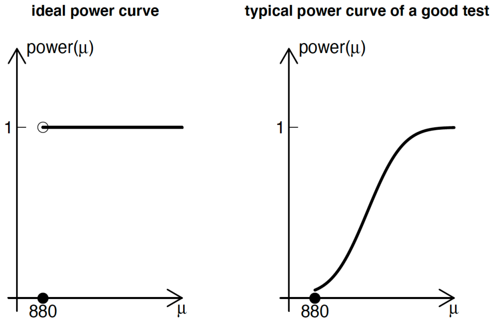
Definition 9.8.2. Consider a test about a mean μ. Then

$$\begin{aligned}\text{power}(\mu) &= \mathbb{P}(\text{reject } H_0 \text{ when the true value is } \mu)\\ &= \mathbb{P}_\mu(\text{reject } H_0).\end{aligned}$$

Therefore the power of a test is really a function, or curve, that depends on the true value of μ. It is easy to check that

$$\text{power}(\text{value of } \mu \text{ under } H_0) = \mathbb{P}(\text{Type I error}) = \alpha.$$

But for all other values of μ we want power(μ) to be as close to 1 as possible. The following graph shows the 'ideal' power curve as well as a typical actual power curve.



Proposition 9.8.1.

$$\mathbb{P}(\text{Type II error}) = 1 - \text{power}(\mu).$$

Power is difficult to calculate by hand unless the test statistic has a normal distribution - hence we will focus on one sample tests of μ, and treat σ as being known such that we can use the test statistic

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Example 9.8.1. Consider again the chemical plant example, where we are testing

$$H_0 : \mu = 880 \quad \text{against} \quad H_1 : \mu > 880$$

using the yield from 50 days of sampling.

Let us suppose that the true value of μ is $\mu = 882$ so H_0 is false (i.e. the new machine *has* increased productivity - by 2 tonnes per day). You may assume that σ is 21 in the following. What is the (approximate) power of our test when $\mu = 882$ for this test?

Solution. Using the test statistic

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

we reject the null hypothesis whenever $\mathbb{P}(Z > 1.645)$. This corresponds to a sample mean of

$$\overline{X} = \mu + 1.645 \cdot \frac{\sigma}{\sqrt{n}} \approx 884.9.$$

So our rejection region is $\overline{X} > 884.9$. If the mean is actually 882, what is the chance that our test rejects H_0?

$$\begin{aligned}\text{power}(882) &= \mathbb{P}(\text{reject } H_0 \text{ if } \mu = 882)\\ &= \mathbb{P}_{\mu=882}\left(\frac{\overline{X} - 882}{21/\sqrt{50}} > \frac{884.9 - 882}{21/\sqrt{50}}\right)\\ &= \mathbb{P}(Z > 0.9764)\\ &= 0.164.\end{aligned}$$

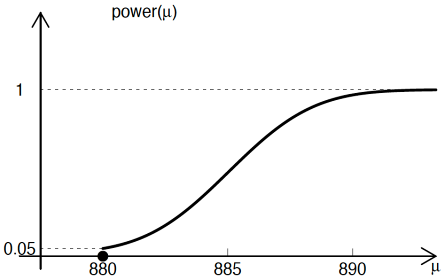
That is, if mean production really was 882 then our test has only a 16% chance of detecting this change and rejecting H_0. □

The power seems to be low in the above scenario - this is because relative to the sample variation ($\sigma = 21$), an increase in yield of 2 is not particularly large. If, on the other hand, yield had increased by 6 tonnes to 886, the power could be shown to be 0.64 in this case. And if yield had increased by 10 tonnes to 890, power would be 95%.

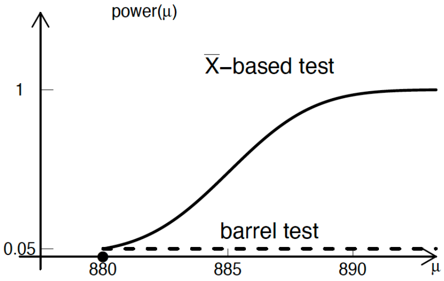
We can do such calculations for general $\mu > 880$ and arrive at a *function* of power values:

$$\begin{aligned}
\text{power}(\mu) &= \mathbb{P}_\mu(\overline{X} > 884.9) \\
&= \mathbb{P}_\mu\left(\frac{\overline{X} - \mu}{21/\sqrt{50}} > \frac{884.9 - \mu}{21/\sqrt{50}}\right) \\
&= \mathbb{P}\left(Z > \frac{884.9 - \mu}{21/\sqrt{50}}\right) \\
&= \mathbb{P}\left(Z < \frac{\mu - 884.9}{21/\sqrt{50}}\right) \\
&= \Phi\left(\frac{\mu - 884.9}{21/\sqrt{50}}\right).
\end{aligned}$$

This leads to the *power curve* shown in the following figure:



It is interesting to graphically compare the power curves from the $\overline{X}$ test and the barrel test for the "chemical plant" example:



Here we see that the $\overline{X}$ test has higher power than the Barrel test for all values of μ, so is clearly superior.

Further notes on power are:

- Power depends on n, σ and the value of μ under H_1. Power can be increased by increasing the size of the sample, n. If the alternative hypotheses are of the form

  $$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu > \mu_0$$

  then for a good test, power is larger for higher values of μ and smaller σ.

- In the planning stages of an experiment a question that is often asked is “Does the proposed test have a reasonable chance of detecting a change in population mean?”. To answer this question one needs to know how big a change in population mean (μ) that we want to be able to detect, sometimes referred to as a “least significant difference” (LSD). One would also need to know (or have reasonable estimate of) the standard deviation of data, σ. Then one can decide what value of n is necessary to have high probability of detecting an LSD. This is one of the most common applied usage of power.

- Power may also be used to compare two or more tests for a given significance level. For example, if we had two competing tests with significance level $\alpha = 0.05$ then we would want to use the test that has the higher power. The previous figure shows that the $\overline{X}$ test is clearly superior to the Barrel test.

9.9 Some More Tests

9.9.1 Two-Sample Test of Means

Consider a situation where we have two independent normal random samples:

$$X_1, \ldots, X_{n_X} \sim N(\mu_X, \sigma^2) \quad \text{and} \quad Y_1, \ldots, Y_{n_Y} \sim N(\mu_Y, \sigma^2).$$

A hypothesis test that is typically of interest is

$$H_0 : \mu_X = \mu_Y \quad \text{versus} \quad H_1 : \mu_X \neq \mu_Y.$$

An appropriate test statistic is

$$\frac{\overline{X} - \overline{Y}}{S_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \sim t_{n_X + n_Y - 2} \quad \text{under } H_0$$

where as in Section 7.6,

$$S_p^2 = \frac{1}{n_X + n_Y - 2}[(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2].$$

This statistic assumes that the two samples have equal variance - as in the two-sample confidence interval, again, in Section 7.6.

Example 9.9.1. Do MATH2801 and MATH2901 students spend the same amount at the hairdressers, on average? A class survey obtained the following results:

$$\begin{array}{lll} n_{2801} = 38 & \bar{x}_{2801} = 28.53 & s_{2801} = 32.89 \\ n_{2901} = 31 & \bar{x}_{2901} = 17.03 & s_{2901} = 19.07 \end{array}$$

Carry out the hypothesis test:

$$H_0 : \mu_{2801} = \mu_{2901} \quad \text{versus} \quad H_1 : \mu_{2801} \neq \mu_{2901}.$$

Solution. Here $s_p^2 = 605.92$ and

$$t = \frac{28.53 - 17.03}{s_p \sqrt{\frac{1}{38} + \frac{1}{31}}} \approx \frac{28.53 - 17.03}{24.61 \sqrt{\frac{1}{38} + \frac{1}{31}}} \approx 1.93.$$

Hence, the p-value is

$$p = \mathbb{P}(T > |t| \text{ or } T < -|t|) = 2\mathbb{P}(T > 1.93) \approx 0.057$$

where $T \sim t_{68}$. Therefore, we reject the H_0 hypothesis at the 5% level of significance. There is evidence against the hypothesis that MATH2801 and MATH2901 students spend the same amount on hairdressers. □

9.10 Wald Tests

So far we have only considered the special cases of normal or binomial data. How about the general situation:

$$X_1, \ldots, X_n \sim f, \quad \text{where} \quad f(x) = f(x;\theta)?$$

Theorem 9.10.1 (Wald Test)**.** *Consider the hypotheses*

$$H_0 : \theta = \theta_0 \quad \textit{versus} \quad H_1 : \theta \neq \theta_0$$

and let $\widehat{\theta}$ be an estimator of θ that is asymptotically normal:

$$\frac{\widehat{\theta} - \theta}{\mathrm{se}(\widehat{\theta})} \xrightarrow{\mathrm{d}} N(0,1).$$

The Wald test statistic is

$$W = \frac{\widehat{\theta} - \theta_0}{\widehat{\mathrm{se}}(\widehat{\theta})} \overset{\text{approx.}}{\sim} N(0,1).$$

Let w be the observed value of W. Then the approximate P-value is given by

$$P\text{-}value \approx \mathbb{P}(|Z| > |w|) = 2\Phi(-|w|)$$

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

Usually the estimator $\widehat{\theta}$ in the Wald test is the maximum likelihood estimator since, for smooth likelihood situations, this estimator satisfies the asymptotic normality requirement, and a formula for the (approximate) standard error is readily available.

Example 9.10.1. Consider the sample size $n = 100$

```
0.366 0.568 0.300 0.115 0.204 0.128 0.277 0.391 0.328 0.451
0.412 0.190 0.207 0.147 0.116 0.326 0.256 0.524 0.217 0.485
0.265 0.375 0.267 0.360 0.250 0.258 0.583 0.413 0.481 0.468
0.406 0.336 0.305 0.321 0.268 0.361 0.632 0.283 0.258 0.466
0.276 0.232 0.133 0.316 0.468 0.496 0.573 0.523 0.256 0.491
0.127 0.054 0.440 0.228 0.249 0.754 0.430 0.111 0.459 0.233
0.257 0.640 0.147 0.273 0.112 0.389 0.126 0.356 0.273 0.296
0.433 0.253 0.234 0.514 0.177 0.221 0.534 0.509 0.510 0.269
0.262 0.625 0.183 0.541 0.705 0.078 0.847 0.149 0.031 0.453
0.299 0.226 0.069 0.211 0.195 0.381 0.317 0.467 0.289 0.593
```

which may be considered to be the observed value of a random sample $X_1, \ldots, X_{100}$ with common density function:

$$f(x;\theta) = 2\theta x e^{-\theta x^2}, \quad x \geq 0; \theta > 0.$$

Use a Wald test to test the hypotheses:

$$H_0 : \theta = 6 \quad \text{versus} \quad H_1 : \theta \neq 6.$$

Solution. In Chapter 8 it was shown that the maximum likelihood estimator is

$$\widehat{\theta} = \frac{100}{\sum_{i=1}^{100} X_i^2}$$

with estimated standard error

$$\widehat{\text{se}}(\widehat{\theta}) = \frac{\widehat{\theta}}{\sqrt{100}}.$$

It may also be shown that $\widehat{\theta}$ is asymptotically normal so the Wald test applies. The Wald test statistic is

$$W = \frac{\widehat{\theta} - 6}{\widehat{\text{se}}(\widehat{\theta})} = \frac{\widehat{\theta} - 6}{\widehat{\theta}/\sqrt{100}} = \frac{\frac{100}{\sum_{i=1}^{100} X_i^2} - 6}{\frac{100}{\sum_{i=1}^{100} X_i^2}/10}.$$

Since $\sum_{i=1}^{100} x_i^2 = 14.081$ the observed value of W is

$$w = \frac{\frac{100}{14.081} - 6}{\frac{100}{14.081}/10} = 1.5514.$$

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

$$p\text{-value} = \mathbb{P}(|Z| > 1.55) = 2\Phi(-1.55) = 0.12.$$

There is little or no evidence against H_0 so we should retain the null hypothesis. □

Example 9.10.2. Do ravens intentionally fly towards gunshot sounds (to scavenge on the carcass they expect to find)? Crow White addressed this question by going to 12 locations, firing a gun, then counting raven numbers 10 minutes later. He repeated the process at 12 different locations where he didn't fire a gun. Results:

no gunshot	0	0	2	3	5	0	0	0	1	0	1	0
gunshot	2	1	4	1	0	5	0	1	0	3	5	2

Is there evidence that ravens fly towards the location of gunshots? Answer this question using an appropriate Wald test.

Solution. Here we might proceed by assuming that B_i and A_i is the number of ravens at the i-th location ($i = 1, \ldots, 12 = n$) before and after the firing of the gunshots. Note that here $A_1, \ldots, A_n$ and $B_1, \ldots, B_n$ are pairwise dependent, because the A_i depends on the B_i. We only have independence across the i-th, that is, the pairs

$$(A_1, B_1), \ldots, (A_n, B_n)$$

are independent with common mean (μ_a, μ_b). Here, due to the dependence, we cannot use a two sample student test. A crude alternative model is to consider the difference

$$X_i = A_i - B_i.$$

Then, $X_1, \ldots, X_n$ are independent and we could test

$$H_0 : \mu_a = \mu_b \quad H_1 : \mu_a > \mu_b$$

using a one sample student test:

$$T = \frac{\overline{X} - \overbrace{(\mu_a - \mu_b)}^{0 \text{ under } H_0}}{S/\sqrt{n}} \sim t_{n-1}.$$

This gives the p-value

$$2 \times \mathbb{P}\left(T > \frac{1}{2.73/\sqrt{12}}\right) \approx 2 \times 0.1153 \approx 0.23.$$

We therefore conclude that there is no evidence that ravens intentionally fly towards gunshots. □

9.11 Likelihood Ratio Tests

The Wald test is a general testing procedure for the situation where an asymptotically normal estimator is available. An even more general procedure, with good power properties, is the likelihood ratio test.

Theorem 9.11.1 (Likelihood Ratio Test (Single Parameter Case))**.** *Consider the hypotheses*

$$H_0 : \theta = \theta_0 \quad \textit{versus} \quad H_1 : \theta \neq \theta_0.$$

The likelihood ratio test statistic is

$$\Lambda(\widehat{\theta}_n) = 2 \ln \left(\frac{\mathcal{L}(\widehat{\theta}_n)}{\mathcal{L}(\theta_0)} \right) = 2[\ell_n(\widehat{\theta}_n) - \ell_n(\theta_0)].$$

Under H_0 and certain regularity conditions

$$\Lambda(\widehat{\theta}_n) \xrightarrow{\text{d}} \chi_1^2.$$

We will frequently drop the n's from the notation for convenience. Let λ be the observed value of Λ. Then the approximate P-value is given by

$$P\textit{-value} \approx \mathbb{P}_{\theta_0}(\Lambda > \lambda) = \mathbb{P}(Q > \lambda) = 2\Phi(-\sqrt{\lambda})$$

where $Q \sim \chi_1^2$.

Proof. Assume the H_0 hypothesis is true and that the data $X_1, \ldots, X_n$ are iid random variables from the 'true' $f(x; \theta_0)$. Then, recall that under some regularity conditions

$$\sqrt{nI_1(\theta_0)}(\theta_0 - \widehat{\theta}_n) \xrightarrow{\text{d}} Z \sim N(0, 1).$$

By Taylor's expansion around $\widehat{\theta}_n$ and the Mean Value Theorem we have

$$\ell_n(\theta_0) = \ell_n(\widehat{\theta}_n) + \ell_n'(\widehat{\theta}_n)(\theta_0 - \widehat{\theta}_n) + \frac{1}{2}(\theta_0 - \widehat{\theta}_n)^2 \ell_n''(\vartheta_n)$$

for some ϑ_n between $\widehat{\theta}_n$ and θ_0. Since $\ell_n'(\widehat{\theta}_n) = 0$, we obtain after rearrangement

$$\begin{aligned}
2(\ell_n(\widehat{\theta}_n) - \ell_n(\theta_0)) &= -\ell_n''(\vartheta_n)(\theta_0 - \widehat{\theta}_n)^2 \\
&= -\frac{\ell_n''(\vartheta_n)}{nI_1(\theta_0)} \left(\sqrt{nI_1(\theta_0)}(\theta_0 - \widehat{\theta}_n) \right)^2 \\
&= \underbrace{-\frac{\ell_n''(\vartheta_n)}{n}}_{\xrightarrow{\mathbb{P}} I_1(\theta_0)} I_1^{-1}(\theta) \left(\underbrace{\sqrt{nI_1(\theta_0)}(\theta_0 - \widehat{\theta}_n)}_{\xrightarrow{\text{d}} Z} \right)^2 \\
&\xrightarrow{\text{d}} 1 \times (Z^2) \sim \chi_1^2
\end{aligned}$$

where we used the result from the remark below and Slutsky's Theorem. □

Remark 9.11.1 (Relation Between Wald and Likelihood Ratio Test Statistics)**.** A Wald statistic uses the horizontal axis, for θ, to construct a test statistic - we take $\widehat{\theta}$ and compare it to θ_0, to see if $\widehat{\theta}$ is significantly far from θ_0.

In contrast, a likelihood ratio statistic uses the vertical axis, for $\ell(\theta)$, to construct a test statistic - we take the maximised log-likelihood $\ell(\widehat{\theta})$, and compare it to the log-likelihood under the null hypothesis,

$\ell(\theta_0)$, to see if $\ell(\theta_0)$ is significantly far from the maximum.

Note that if $W = (\widehat{\theta}_n - \theta_0)/\operatorname{se}(\widehat{\theta}_n)$ is the Wald statistic, then

$$\Lambda = 2(\ell_n(\widehat{\theta}_n) - \ell_n(\theta_0)) = -\frac{\ell_n''(\vartheta_n)}{n}\operatorname{se}^2(\widehat{\theta}_n)W^2 \xrightarrow{\mathbb{P}} W^2.$$

Thus, the Wald and likelihood ratio tests are asymptotically equivalent when the null hypothesis is true - and in large samples, they typically return similar test statistics hence similar conclusions. However, when the null hypothesis is not true, these tests can have quite different properties, especially in small samples.

Example 9.11.1. Consider, one last time, the sample of size $n = 100$.

0.366	0.568	0.300	0.115	0.204	0.128	0.277	0.391	0.328	0.451
0.412	0.190	0.207	0.147	0.116	0.326	0.256	0.524	0.217	0.485
0.265	0.375	0.267	0.360	0.250	0.258	0.583	0.413	0.481	0.468
0.406	0.336	0.305	0.321	0.268	0.361	0.632	0.283	0.258	0.466
0.276	0.232	0.133	0.316	0.468	0.496	0.573	0.523	0.256	0.491
0.127	0.054	0.440	0.228	0.249	0.754	0.430	0.111	0.459	0.233
0.257	0.640	0.147	0.273	0.112	0.389	0.126	0.356	0.273	0.296
0.433	0.253	0.234	0.514	0.177	0.221	0.534	0.509	0.510	0.269
0.262	0.625	0.183	0.541	0.705	0.078	0.847	0.149	0.031	0.453
0.299	0.226	0.069	0.211	0.195	0.381	0.317	0.467	0.289	0.593

which may be considered to be the observed value of a random sample $X_1, \ldots, X_{100}$ with common density function f given by

$$f(x;\theta) = 2\theta x e^{-\theta x^2}, \quad x \geq 0; \theta > 0.$$

Use a likelihood ratio test to test the hypotheses:

$$H_0 : \theta = 6 \quad \text{versus} \quad H_1 : \theta \neq 6.$$

Solution. First, note that the likelihood ratio statistic is

$$\lambda = 2\ln\left(\frac{\mathcal{L}(\widehat{\theta})}{\mathcal{L}(6)}\right) = 2[\ell(\widehat{\theta}) - \ell(6)]$$

where

$$\ell(\theta) = \sum_{i=1}^{100} \ln f(X_i;\theta) = 100\ln 2 + 100\ln\theta + \sum_{i=1}^{100}\ln X_i - \theta\sum_{i=1}^{100} X_i^2.$$

As shown before, the maximum likelihood estimator is

$$\widehat{\theta} = \frac{100}{\sum_{i=1}^{100} X_i^2}$$

so

$$\begin{aligned}\ell(\widehat{\theta}) &= 100\ln 2 + 100\ln\left(\frac{100}{\sum_{i=1}^{100} X_i^2}\right) + \sum_{i=1}^{100}\ln X_i - \left(\frac{100}{\sum_{i=1}^{100} X_i^2}\right)\sum_{i=1}^{100} X_i^2 \\ &= 100[\ln 2 + \ln 100] - 100\ln\left(\sum_{i=1}^{100} X_i^2\right) + \sum_{i=1}^{100}\ln X_i - 100.\end{aligned}$$

Also,

$$\ell(6) = 100\ln 2 + 100\ln 6 + \sum_{i=1}^{100}\ln X_i - 6\sum_{i=1}^{100} X_i^2$$

and so the likelihood ratio statistic is

$$\Lambda = 2\left[100[\ln(100/6)-1] - 100\ln\left(\sum_{i=1}^{100} X_i^2\right) + 6\sum_{i=1}^{100} X_i^2\right].$$

Since $\sum_{i=1}^{100} x_i^2 = 14.081$ the observed value of Λ is

$$\lambda = 2[100[\ln(100/6)-1] - 100\ln(14.081) + 6\times 14.081] = 2.689.$$

Then

$$\begin{aligned}
p\text{-value} &= \mathbb{P}_{\theta=6}(\Lambda > \lambda) \\
&= \mathbb{P}(Q > 2.689) && Q \sim \chi_1^2 \\
&= \mathbb{P}(Z^2 > 2.689) && Z \sim N(0,1) \\
&= 2\mathbb{P}(Z \leq -\sqrt{2.689}) \\
&= 2\Phi(-1.64) \\
&= 0.1
\end{aligned}$$

which is close to the p-value of 0.12 obtained via the Wald test in the previous section. The conclusion remains that there is little or no evidence against H_0 and that H_0 should be retained. □

9.11.1 Multiparameter Extension of the Likelihood Ratio Test

The likelihood ratio test procedure can be extended to hypothesis tests involving several parameters simultaneously.

Consider a model with parameter vector $\boldsymbol{\theta}$ and corresponding parameter space Θ. A general class of hypotheses is

$$H_0 : \boldsymbol{\theta} \in \Theta_0 \quad \text{versus} \quad H_1 : \boldsymbol{\theta} \notin \Theta_0$$

where Θ_0 is a subset of Θ. Then the likelihood ratio statistic is

$$\Lambda = 2\ln\left(\frac{\sup_{\theta\in\Theta}\mathcal{L}(\boldsymbol{\theta})}{\sup_{\theta\in\Theta_0}\mathcal{L}(\boldsymbol{\theta})}\right).$$

Under H_0 and regularity conditions on $\mathcal{L}$,

$$\Lambda \xrightarrow{\text{d}} \chi_d^2$$

where d is the dimension of Θ minus the dimension of Θ_0.

A. **Probability**

A.1 Experiment, Sample Space, Event

Probability theory is about modelling and analysing **random experiments**: experiments whose outcome cannot be determined in advance, but is nevertheless still subject to analysis.

Definition A.1.1. An *experiment* is any process leading to recorded observations.

For example, they include tossing a die, measuring the lifetime of a machine etc. Mathematically, we can model a random experiment by specifying an appropriate *probability space*, which consists of three components: a *sample space*, a set of *events* and a *probability.*

Definition A.1.2. An *outcome* is a possible result of an experiment. The set Ω of all possible outcomes is the *sample space* of an experiment. Ω is discrete if it contains a countable (finite or countably infinite) number of outcomes.

Examples of random experiments with their sample spaces are:

- Cast two dice consecutively

$$\Omega = \{(1,1),(1,2),\ldots,(1,6),(2,1),\ldots,(6,6)\}.$$

- The lifetime of a machine (in days)

$$\Omega = \mathbb{R}^+$$

Notice that for modelling purposes it is often easier to take the sample space larger than necessary. For example, the actual lifetime of a machine would certainly not span the entire positive real axis.

Definition A.1.3. An *event* is a set of outcomes (a subset of Ω). An event *occurs* if the result of the experiment is one of the outcomes in that event.

Thus, an event is a collection of some possible outcomes of the experiment. Events will be denoted by capital letters $A, B, C, \ldots$. Examples include

- The event that the sum of two dice is 10 or more,

$$A = \{(5,5),(5,6),(6,5),(6,6)\}$$

- The event that a machine lives less than 1000 days,

$$A = [0, 1000).$$

Definition A.1.4. Events are *mutually exclusive* (disjoint) if they have no outcomes in common: that is, if they cannot both occur. If A and B are mutually exclusive, we can say that $A \cap B = \varnothing$.

A.2 Advanced Material

The next thing to specify, in a model, is the collection $\mathcal{F}$, of all events "of interest". That is, the collection of all events to which we wish to assign a "probability". We can take $\mathcal{F}$ equal to the collection of *all* subsets of Ω (the power set). When Ω is *countable* this is okay, however, when Ω is uncountable, the power set of Ω is in general so large that one *cannot* assign a proper "probability" to all subsets. Thus, for an uncountable Ω we have to settle for a smaller collection $\mathcal{F}$ of events. This collection should have nice properties, such as

1. With A and B events, the set $A \cup B$ should also be an event, namely the event that A *or* B *or* both occur
2. With A and B events, the set $A \cap B$ should also be an event, namely the event that A *and* B both occur
3. With A an event, the event A^c should also be an event, namely, the event that A does *not* occur
4. The set Ω itself should be an event, namely the "certain" event. Similarly $\varnothing$ should be an event, namely the "impossible" event

The minimal assumption that we impose on $\mathcal{F}$ is that it should be an object called a σ-algebra.

Definition A.2.1. A *σ-algebra* $\mathcal{F}$ on Ω is a collection of subsets of Ω that satisfies

(i) $\Omega \in \mathcal{F}$

(ii) If $A \in \mathcal{F}$ then also $A^c \in \mathcal{F}$

(iii) If $A_1, A_2, \ldots \in \mathcal{F}$, then $\bigcup_n A_n \in \mathcal{F}$.

Events $A_1, A_2, \ldots$ are called *exhaustive* if their union is the whole sample space Ω. A sequence $A_1, A_2, \ldots$ of disjoint and exhaustive events is called a *partition* of Ω.

A.2.1 Borel σ-Algebra

The most important example of a non-trivial σ-algebra is the *Borel* σ-algebra on $\mathbb{R}$, denoted $\mathcal{B}$. This is defined as the smallest σ-algebra on $\mathbb{R}$ that contains all the intervals of the form $(-\infty, x]$, for $x \in \mathbb{R}$. We say that $\mathcal{B}$ is *generated* by the collection of intervals $(-\infty, x]$. This σ-algebra of sets is big enough to contain all important sets, and small enough to allow us to assign a natural "length measure" to all sets. This is called the *Lebesgue measure*, often denoted by Leb or m or λ.

For $\mathbb{R}^n$ we can do something similar. The smallest σ-algebra on $\mathbb{R}^n$ that contains all the "rectangles" of the form

$$(-\infty, x_1] \times \cdots \times (-\infty, x_n],$$

with $(x_1, \ldots, x_n) \in \mathbb{R}^n$ is called the Borel σ-algebra on $\mathbb{R}^n$; we write $\mathcal{B}^n$. The corresponding natural "volume" measure is again called the Lebesgue measure. For example, the Lebesgue measure of the unit disc is π.

A.2.2 Extended Real Line and Borel σ-Algebra

Instead of working with the real line, it will be convenient to work with the *extended real line* $\overline{\mathbb{R}} = [-\infty, \infty]$. The natural extension of $\mathcal{B}$ is the σ-algebra $\overline{\mathcal{B}}$ which is generated by the intervals $[-\infty, x]$. It is called the Borel σ-algebra on $\overline{\mathbb{R}}$.

Similarly, the Borel σ-algebra $\overline{\mathcal{B}}^n$ on $\overline{\mathbb{R}}^n$ (the meaning should be obvious) is defined as the σ-algebra that is generated by the rectangles of the form

$$[-\infty, x_1] \times \cdots \times [-\infty, x_n].$$

A.3 Axioms and Basic Results

Given a sample space $(\Omega, \mathcal{F})$, a probability function $\mathbb{P}$ can be defined in the following way. To every event $A \in \mathcal{F}$ we assign a number $\mathbb{P}(A)$, the *probability that* A occurs. The function $\mathbb{P}$ must satisfy the axioms:

(i) For each $A \subset \Omega$, $\mathbb{P}(A) \geq 0$

(ii) $\mathbb{P}(\Omega) = 1$

(iii) If $A_1, A_2, \ldots$ are mutually exclusive (or disjoint)

$$(A_i \cap A_j = \varnothing \text{ for all } i, j \text{ with } i \neq j)$$

then

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

Note that we will use the following interchangeable notation for the complement of event A:

$$A^c \equiv \overline{A}.$$

Proposition A.3.1.

(i) If $A_1, A_2, \ldots, A_k$ are mutually exclusive,

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

(ii) $\mathbb{P}(\varnothing) = 0$.

(iii) For any $A \subseteq \Omega$, $0 \leq \mathbb{P}(A) \leq 1$ and $P(\overline{A}) = 1 - \mathbb{P}(A)$.

(iv) If $B \subset A$, then $\mathbb{P}(B) \leq \mathbb{P}(A)$. Thus if B occurs $\Rightarrow$ A occurs then $\mathbb{P}(B) \leq \mathbb{P}(A)$.

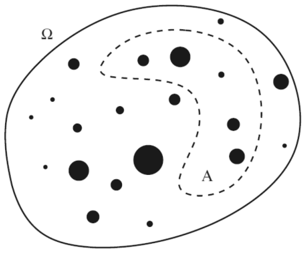
Proof. Suppose $A \subset B$. Then, we can write B as $B = A \cup (A^c \cap B)$, where A and $A^c \cap B$ are disjoint events. Hence, according to the third and first axiom

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

□

Rule for calculating probabilities: if Ω is discrete, $\mathbb{P}(A)$ = sum of probabilities for outcomes in A. This follows from axiom (iii) since, for example in the figure below (representing the computation of probabilities in a discrete sample space Ω. Each blob represents the relative weight assigned to it by the probability measure $\mathbb{P}$), if $\Omega = \{s_1, s_2, \ldots\}$ (where s_i and s_j are mutually exclusive) and $A = \{s_2, s_5, s_9, s_{11}, s_{14}\}$, then $A = \{s_2\} \cup \{s_5\} \cup \{s_9\} \cup \{s_{11}\} \cup \{s_{14}\}$, a mutually exclusive union, so

$$\mathbb{P}(A) = \mathbb{P}(\{s_2\} \cup \{s_5\} \cup \{s_9\} \cup \{s_{11}\} \cup \{s_{14}\}) = \mathbb{P}(\{s_2\}) + \mathbb{P}(\{s_5\}) + \mathbb{P}(\{s_9\}) + \mathbb{P}(\{s_{11}\}) + \mathbb{P}(\{s_{14}\}).$$



A.3.1 Monotonic Sequences of Events

Theorem A.3.1 (Continuity Property of $\mathbb{P}$)**.** *If $A_1, A_2, \ldots$ is an increasing sequence of events, i.e., $A_1 \subset A_2 \subset \cdots$, then*

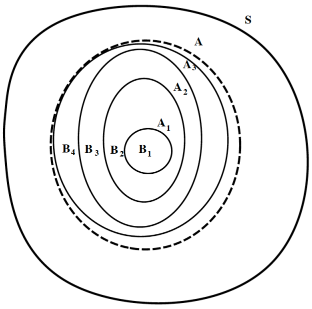
$$\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right).$$

This is a kind of continuity property. We say that $\mathbb{P}$ is continuous from below.

Proof. Suppose $A_1, A_2, \ldots$ is an increasing sequence of events. Define the event

$$A = \bigcup_n A_n = \bigcup_{n=1}^{\infty} A_n.$$

Now consider the following figure:



Note that here $S \equiv \Omega$.

Define the events $B_1, B_2, \ldots$ as

$$\begin{aligned}
B_1 &= A_1 \\
B_2 &= A_2 \cap A_1^c \\
&\vdots \\
B_n &= A_n \cap A_{n-1}^c, && n = 2, 3, \ldots
\end{aligned}$$

The "rings" $B_1, B_2, \ldots$ are disjoint and

$$\bigcup_{i=1}^{n} B_i = \bigcup_{i=1}^{n} A_i = A_n \text{ and } \bigcup_{i=1}^{\infty} B_i = A.$$

Hence, from axiom three it follows that

$$\begin{aligned}
\mathbb{P}(A) &= \mathbb{P}\left(\bigcup_{i=1}^{\infty} B_i\right) \\
&= \sum_{i=1}^{\infty} \mathbb{P}(B_i) && \text{(by axiom three)} \\
&= \lim_{n\to\infty} \sum_{i=1}^{\infty} \mathbb{P}(B_i) \\
&= \lim_{n\to\infty} \mathbb{P}\left(\bigcup_{i=1}^{n} B_i\right) \\
&= \lim_{n\to\infty} \mathbb{P}(A_n).
\end{aligned}$$

□

Theorem A.3.2 (Continuity from Above)**.** *If $A_1, A_2, \ldots$ is a decreasing sequence of events, i.e., $A_1 \supseteq A_2 \supseteq \cdots$, then*

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\left(\bigcap_{i=1}^{\infty} A_n\right).$$

A.4 Counting Rules

Counting Rule 1

If there are k experiments with n_i possible outcomes in the i-th ($i = 1, 2, \ldots, k$), then the total number of possible outcomes for the k experiments is $n_1 n_2 \ldots n_k = \prod_{i=1}^{k} n_i$.

Counting Rule 2

The number of possible permutations of r objects selected from n distinct objects is ${}^nP_r = \frac{n!}{(n-r)!}$, where $n! = n(n-1)(n-2)\cdots 3 \cdot 2 \cdot 1$ for integers $n \geq 1$ and $0! \equiv 1$.

Counting Rule 3

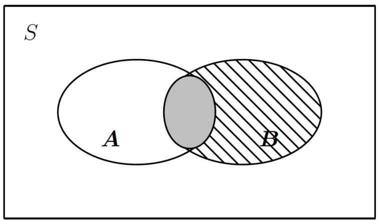
The number of ways of choosing r objects from n distinct objects is

$$\frac{n!}{r!(n-r)!} \equiv \binom{n}{r} \text{ (“}n\text{ choose }r\text{”)}, 0 \leq r \leq n.$$

A.5 Conditional Probability

Definition A.5.1. The *conditional probability* that an event A occurs, given that an event B has occurred is

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \text{ if } \mathbb{P}(B) \neq 0.$$



Given that B has occurred, the total probability for possible results of the experiment equals $\mathbb{P}(B)$, so that the probability that A occurs equals the total probability for outcomes in A (only those in $A \cap B$) divided by the total probability, $\mathbb{P}(B)$.

Lemma A.5.1. *$P(A|B) = \mathbb{P}(A) \iff \mathbb{P}(B|A) = \mathbb{P}(B)$.*

Proof. We prove "⇒". If $\mathbb{P}(A|B) = \mathbb{P}(A)$, then

$$\begin{aligned}\mathbb{P}(B|A) &= \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} \\ &= \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)} = \mathbb{P}(B).\end{aligned}$$

Interchange A and B in the above proof to prove "⇐". □

A.6 Independent Events

Events A and B are *independent* if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. For any two events A and B, $P(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$, so A and B are independent

$$\begin{aligned}&\iff \mathbb{P}(A|B) = \mathbb{P}(A) \\ &\iff \mathbb{P}(B|A) = \mathbb{P}(B) \qquad \text{(prev. lemma).}\end{aligned}$$

For a countable sequence of events $\{A_i\}$, the events are *pairwise independent* if

$$\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j) \text{ for all } i \neq j$$

and the events are *(mutually) independent* if for any collection

$$A_{i_1}, A_{i_2}, \ldots, A_{i_n},$$
$$\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_n}) = \mathbb{P}(A_{i_1}) \cdots \mathbb{P}(A_{i_n}).$$

Clearly, independence ⇒ pairwise independence, but not vice versa, as in the following example.

Example A.6.1. A coin is tossed twice. Let A be the event 'head on first toss', B the event 'head on the second toss' and C the event 'exactly one head turned up'. Note, A, B, C are pairwise independent, but

$$\mathbb{P}(A \cap B \cap C) = 0 \neq \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C) = \frac{1}{2^3},$$

so A, B, C are not independent.

Example A.6.2. A ball is drawn at random from an urn containing 4 balls numbered 1, 2, 3, 4. Let

$$\begin{aligned}A &= \{1, 2\} \qquad \text{(ball 1 or ball 2 is drawn)} \\ B &= \{1, 3\} \\ C &= \{1, 4\}.\end{aligned}$$

Show that A, B, C are pairwise independent but not independent.

Solution. Then A, B, C are pairwise independent (e.g. $\mathbb{P}(A \cap B) = \mathbb{P}(\{1\}) = \mathbb{P}(A)\mathbb{P}(B) = \frac{1}{4}$), but $\frac{1}{4} = \mathbb{P}(A \cap B \cap C) \neq \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$, so A, B, C are not independent. □

A.7 Some Probability Laws

The Multiplicative Law

For events A_1, A_2

$$\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_2 \cap A_1) = \mathbb{P}(A_2|A_1)\mathbb{P}(A_1).$$

For events A_1, A_2, A_3

$$\begin{aligned}\mathbb{P}(A_1 \cap A_2 \cap A_3) &= \mathbb{P}(A_3 \cap A_2 \cap A_1)\\ &= \mathbb{P}(A_3|A_2 \cap A_1)\mathbb{P}(A_2 \cap A_1)\\ &= \mathbb{P}(A_3|A_2 \cap A_1)\mathbb{P}(A_2|A_1)\mathbb{P}(A_1).\end{aligned}$$

The same pattern applies to higher numbers of events.

The Additive Law

For events A and B,

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

A.7.1 Corollary to the Additive Law

For events A and B, if A and B are mutually exclusive

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B).$$

The Law of Total Probability

Suppose $A_1, A_2, \ldots, A_k$ are mutually exclusive ($A_i \cap A_j = \varnothing$ for all $i \neq j$) and *exhaustive* ($\bigcup_{i=1}^k A_i = \Omega =$ sample space) events; that is, $A_1, \ldots, A_k$ form a *partition* of Ω. Then, for any event B,

$$\mathbb{P}(B) = \sum_{i=1}^k \mathbb{P}(B|A_i)\mathbb{P}(A_i).$$

Proof. Now,

$$B = \bigcup_{i=1}^k (B \cap A_i) \text{ (disjoint union since the } A_i\text{'s are disjoint)}$$

By axiom (iii) (in the finite case)

$$\begin{aligned}\mathbb{P}(B) &= \mathbb{P}\left(\bigcup_{i=1}^k (B \cap A_i)\right)\\ &= \sum_{i=1}^k \mathbb{P}(B \cap A_i)\\ &= \sum_{i=1}^k \mathbb{P}(B|A_i)\mathbb{P}(A_i)\end{aligned}$$

□

Example A.7.1. Urn I contains 3 red and 4 white balls. Urn II contains 2 red balls and 4 white. A ball is drawn from Urn I and placed unseen into Urn II. A ball is now drawn at random from Urn II. What is the probability that this second ball is red?

Solution. Let A_1 be the event '1st ball drawn red', A_2 '1st ball drawn white' and B '2nd ball drawn red'. A_1 and A_2 are mutually exclusive (they cannot both occur) and exhaustive (one of them must occur) and so

$$\begin{aligned}\mathbb{P}(B) &= \mathbb{P}(B|A_1)\mathbb{P}(A_1) + \mathbb{P}(B|A_2)\mathbb{P}(A_2)\\ &= \frac{3}{7} \times \frac{3}{7} + \frac{2}{7} \times \frac{4}{7}\\ &= \frac{17}{49}.\end{aligned}$$

□

A.8 Bayes' Formula

Bayes' formula calculates conditional probabilities when the ordering of conditioning is reversed. In the simple two event situation Bayes' Formula is:

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

Bayes' Formula

For a partition $A_1, A_2, \ldots, A_k$ and an event B,

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(B|A_j)\mathbb{P}(A_j)}{\sum_{i=1}^{k} \mathbb{P}(B|A_i)\mathbb{P}(A_i)} = \frac{\mathbb{P}(B|A_j)\mathbb{P}(A_j)}{\mathbb{P}(B)}$$

Example A.8.1. A diagnostic test for a certain disease is claimed to be 90% accurate because, if a person has the disease, the test will show a positive result with probability 0.9 while if a person does not have the disease the test will show a negative result with probability 0.9. Only 1% of the population has the disease. If a person is chosen at random from the population and tests positive for the disease, what is the probability that the person does in fact have the disease?

Solution. Let A be the event 'person has disease' and B be the event 'person tests positive'.

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

since A and A^c form a partition (they are mutually exclusive and exhaustive). Now,

$$\mathbb{P}(B|A) = 0.9 \qquad \mathbb{P}(A) = 0.01$$
$$\mathbb{P}(B|A^c) = 0.1 \qquad \mathbb{P}(A^c) = 0.99.$$

Therefore,

$$\mathbb{P}(A|B) = \frac{0.9 \times 0.01}{0.9 \times 0.01 + 0.1 \times 0.99} = \frac{1}{12};$$

that is, given that the person's test result is positive the probability that a person has the disease is $\frac{1}{12}$. □