EMTH210 Topic 7: Statistics

Topic Contents

- **Review Material (~2 lectures):** EMTH119 revision/reminder: Random variables and expectation.
- **Delta Method (~1 lecture):** Approximating expectations of functions of random variables using Taylor Series.
- **Random Vectors (~2 lectures):** Extending continuous random variables to more dimensions, using multiple integration: Joint distributions, marginal distributions, independence, correlation and a special property of normal random variables.
- **Statistics (~3 lectures):** Sample mean, sample variance, law of large numbers, central limit theorem, sample size and confidence intervals.
- **Hypothesis Testing (~2 lectures):** Making decisions with statistics: Type I and Type II errors, z-test and power.
- **Maximum Likelihood Estimation (~2 lectures):** Finding parameter values from data: Likelihood function, maximum likelihood estimator and confidence intervals.

7.1 Review Material

Probability is the mathematics for quantifying uncertainty. Three key concepts:

1. ***Experiment***: An experiment is an activity that results in distinct outcomes that cannot be predicted with certainty. A single performance of an experiment is called a ***trial***.
2. ***Sample Space***: The sample space Ω is the set of all possible outcomes of an experiment. Subsets of Ω are called events. Consider, for example, the coin toss experiment with two distinct outcomes:
$$\Omega = \{\text{Head, Tail}\}$$
3. ***Random Variable***: A random variable X is a mapping from the sample space Ω to the set of real numbers
$$X : \Omega \to \mathbb{R},$$
where each distinct outcome is assigned a unique numerical value.

In this course, we work with random variables directly and rarely mention sample spaces (but they are always there in the background).

Exercise 1 Rat

A rat is selected at random from a cage of male (M) and female (F) rats. Once selected, the gender of the rat is recorded. Define the sample space and a random variable for this experiment.

Exercise 2 Tree

A tree is randomly selected from a forest and its height is measured. Define the sample space and a random variable for this experiment.

Random variables can be either discrete or continuous:

1. ***Discrete random variable***: The set of possible outcomes is finite (as in Exercise 1) or at most countably infinite (can be put in a one-to-one correspondence with the integers).
2. ***Continuous random variable***: May take on any value in a range (as in Exercise 2) and is formally defined in Section 7.1.1.

We are usually interested in the probability of an event happening and so random variables are often described using distribution functions.

The ***distribution function*** (or ***cumulative distribution function*** (***CDF***))

$$F_X : \mathbb{R} \to [0, 1]$$

of the random variable X is

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

for any $x \in \mathbb{R}$. That is, the probability that the random variable X takes on a value less than or equal to x.

Theorem: The probability that the random variable X takes a value x in the interval $(a, b]$ is

$$\mathrm{P}(a < X \leq b) = F_X(b) - F_X(a).$$

7.1.1 Continuous Random Variables

A random variable, X, is ***continuous*** if there exists a function f_X such that

1. $f_X(x) \geq 0$ for all x;
2. $\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = 1$; and
3. for every $a \leq b$,

$$\mathrm{P}(a < X < b) = \int_a^b f_X(x)\,\mathrm{d}x.$$

The function f_X is called a ***probability density function*** (***PDF***) and the set of values for which f_X is non-zero is called the ***support*** of X.

$f_X(x)$ is not a probability! Only areas under the $f_X(x)$ curve are probabilities.

The distribution function corresponding to f_X is

$$F_X(x) = \mathrm{P}(X \leq x) = \int_{-\infty}^{x} f_X(v)\,\mathrm{d}v.$$

Properties:

1. $0 \leq F_X(x) \leq 1$;
2. $F_X(x)$ is a non-decreasing function of x; and
3. $f_X(x) = F_X'(x)$ at all points where F_X is differentiable.

Exercise 3 Continuous Random Variable

Consider the continuous random variable, X, whose probability density function is:

$$f_X(x) = \begin{cases} 4x^3 & \text{if } 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

(a) What is the support of X?

(b) Find the distribution function, $F_X(x)$.

(c) Find $\mathrm{P}\left(\frac{1}{\sqrt[4]{3}} \leq X \leq \frac{1}{\sqrt[4]{2}}\right)$.

(d) Find x such that $\mathrm{P}(X \leq x) = 0.9$.

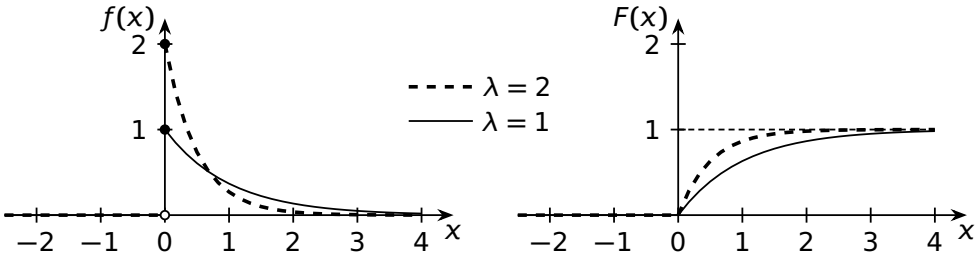
7.1.2 Three Important Continuous Random Variables

Exponential Random Variable: Given a rate parameter $\lambda > 0$, the Exp(λ) random variable X has probability density function given by

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

and distribution function given by

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



$f(x)$ and $F(x)$ of exponential random variables where $\lambda = 1$ and $\lambda = 2$.

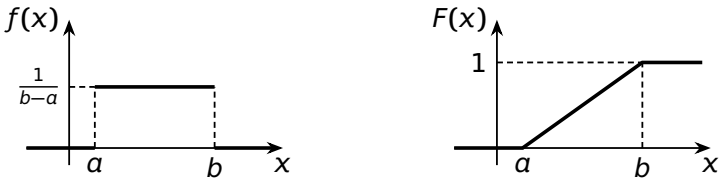
Often occurs in practice as a description of the time elapsing between unpredictable events (time between breakdowns of manufacturing equipment, time between earthquakes, arrival times of buses, time between emergency arrivals at a hospital, etc.).

Uniform Random Variable: Given two real parameters a, b with $a < b$, the Uniform(a, b) random variable X has the following PDF:

$$f(x;\, a,\, b) = \begin{cases} \dfrac{1}{b-a} & \text{if } a \leq x \leq b, \\ 0 & \text{otherwise.} \end{cases}$$

X is also said to be ***uniformly distributed*** on the interval $[a, b]$. The distribution function of X is

$$F(x;\, a,\, b) = \begin{cases} 0 & \text{if } x < a, \\ \dfrac{x-a}{b-a} & \text{if } a \leq x < b, \\ 1 & \text{if } b \leq x. \end{cases}$$



$f(x)$ and $F(x)$ of a Uniform(a, b) random variable.

Uniform random variables are natural choices for experiments in which some event is "equally likely" to happen at any time or place within some interval, e.g. angle at which a fidget spinner stops, time to wait at a subway station with trains every 10 minutes, etc.

Normal (Gaussian) Random Variable: A continuous random variable X with mean μ (*centre*) and standard deviation σ (*spread*) is called normal if its probability density function is

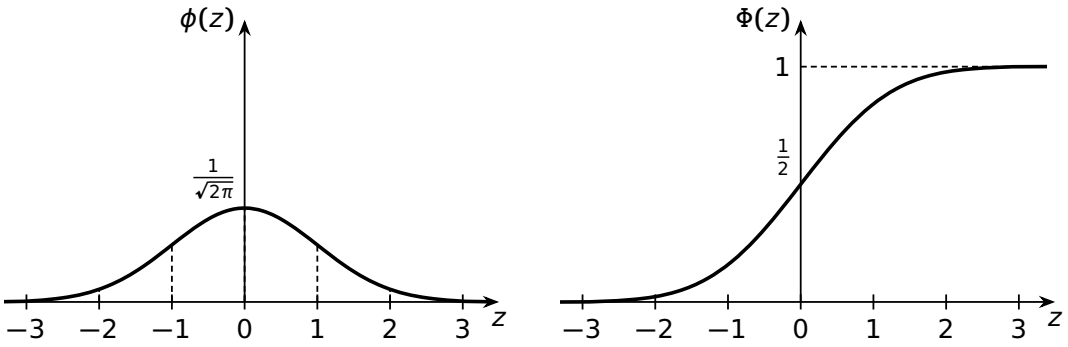
$$f_X(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right).$$

If X is Normal(μ,σ^2), then $Z = (X-\mu)/\sigma$ has a standard normal distribution ($\mu = 0$ and $\sigma^2 = 1$). Hence, if X is a normal random variable,

$$\mathrm{P}(a < X < b) = \mathrm{P}\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right).$$

We cannot integrate the normal PDF, so you will be provided with standard normal tables to compute probabilities.

The normal distribution is "bell shaped" and is the most important continuous probability distribution; we will see very useful applications of it later in this topic. It was first described by De Moivre in 1733 and subsequently by the German mathematician C. F. Gauss (1777–1855). Many random variables have a normal distribution, or they are approximately normal.



The PDF (with multiples of σ shown) and CDF for a standard normal random variable ($\mu = 0$ and $\sigma = 1$). $\mathrm{P}(-1 < Z < 1) \approx 0.68$, $\mathrm{P}(-2 < Z < 12) \approx 0.95$, and $\mathrm{P}(-3 < Z < 3) \approx 0.99$, so 3 standard deviations from the mean is quite extreme.

Notation (new)

We have a notation for statements like "X has distribution Normal$(0,1)$":

$$X \sim \text{Normal}(0,1).$$

We read the symbol "$\sim$" as "has distribution", **not** "is approximately".

For example, we could have $X \sim \text{Exp}(5)$, $Y \sim \text{Uniform}(-1,1)$, and $W \sim \text{Normal}(2,1)$.

We will leave out the parameter names (for Exponential, λ; for Uniform, a and b; and for Normal, μ and σ^2 (note the ^{2})), as they are standard and always in the same order.

7.1.3 Discrete Random Variables

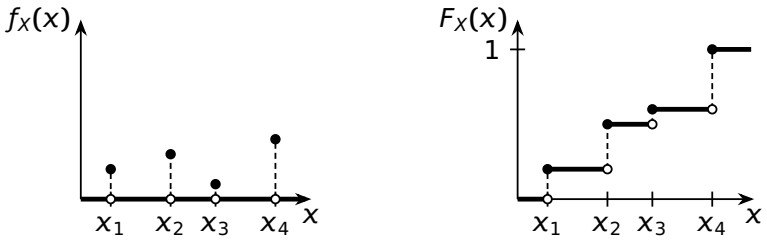
Although this course focuses primarily on continuous random variables, we should remind ourselves about discrete random variables. If X is a discrete random variable that assumes the values $x_1, x_2, x_3, \ldots$ with probabilities $p_i = \mathrm{P}(X = x_i)$, then the ***probability mass function*** (***PMF***) of X is:

$$f_X(x) = \mathrm{P}(X = x) = \begin{cases} p_i & \text{if } x = x_i \\ 0 & \text{otherwise.} \end{cases}$$

The distribution function is given by

$$F_X(x) = \sum_{x_i \le x} f_X(x_i) = \sum_{x_i \le x} p_i.$$

Here is an example:



7.1.4 Expectation

The expectation of a function $g(X)$ of a continuous random variable X is defined as

$$\mathrm{E}\big(g(X)\big) = \int_{-\infty}^{\infty} g(x) f_X(x)\,\mathrm{d}x,$$

where $f_X(x)$ is the probability density function for X. $\mathrm{E}\big(g(X)\big)$ is a *number*, not a function of X. There are two expectations of particular interest:

1. **Population Mean:** a measure of central location, usually denoted by μ, and defined as
$$\mu = \mathrm{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x,$$
($g(x) = x$). The expected value of X, $\mathrm{E}(X)$, can be thought of as the average $\frac{1}{n}\sum_{i=1}^{n} X_i$ of a large number of independent and identically distributed trials, $X_1, X_2, \ldots, X_n$.

2. **Population Variance:** a measure of variability, usually denoted by σ^2, and defined as
$$\sigma^2 = \mathrm{Var}(X) = \mathrm{E}\Big(\big(X - \mathrm{E}(X)\big)^2\Big) = \int_{-\infty}^{\infty} \big(x - \mathrm{E}(X)\big)^2 f_X(x)\,\mathrm{d}x,$$
($g(x) = \big(x - \mathrm{E}(X)\big)^2$). The square root of variance is called the ***standard deviation***, usually denoted by σ: $\sigma = \sqrt{\mathrm{Var}(X)}$.

Properties:

1. $\mathrm{E}(a) = a \quad (a \in \mathbb{R})$;
2. $\mathrm{E}\big(a\,g(X)\big) = a\,\mathrm{E}\big(g(X)\big) \quad (a \in \mathbb{R})$; and
3. $\mathrm{E}\big(g(X) + h(X)\big) = \mathrm{E}\big(g(X)\big) + \mathrm{E}\big(h(X)\big)$ ($h(X)$ is another function of X).

Using these properties, we can write

$$\begin{aligned}
\mathrm{Var}(X) &= \mathrm{E}\big((X - \mathrm{E}(X))^2\big) \\
&= \mathrm{E}\Big(X^2 - 2X\,\mathrm{E}(X) + \big(\mathrm{E}(X)\big)^2\Big) \\
&= \mathrm{E}(X^2) - 2\,\mathrm{E}(X)\big(\mathrm{E}(X)\big) + \big(\mathrm{E}(X)\big)^2 \qquad \text{(since } \mathrm{E}(X) \text{ is a constant)} \\
&= \mathrm{E}(X^2) - 2\big(\mathrm{E}(X)\big)^2 + \big(\mathrm{E}(X)\big)^2 \\
&= \mathrm{E}(X^2) - \big(\mathrm{E}(X)\big)^2.
\end{aligned}$$

We can also show that

$$\mathrm{Var}(aX + b) = a^2\,\mathrm{Var}(X).$$

To define expectations for discrete random variables, we replace the integrals with $\sum_x$ (summing over the support of X).

Exercise 4 Expectation and Variance

Let X be an exponential random variable with rate parameter $\lambda = 2$ (usually denoted Exp(2)). Find the following:

(a) The probability density function, $f_X(x; 2)$.

(b) The distribution function, $F_X(x; 2)$.

(c) The population mean, $\mathrm{E}(X)$.

(d) The population variance, Var(X).

7.2 Delta Method

In the previous sections, we computed the exact mean and variance of $g(X)$. However, $\mathrm{E}\big(g(X)\big)$ and $\mathrm{Var}\big(g(X)\big)$ may be difficult (or impossible) to compute analytically. We will use the delta method in such cases. The delta method is based on Taylor series approximations to $g(X)$ about $\mu = \mathrm{E}(X)$. The Taylor series for $g(X)$ about μ is

$$g(X) = \sum_{k=0}^{\infty} \frac{g^{(k)}(\mu)}{k!}(X-\mu)^k,$$

where $g^{(k)}$ is the kth derivative of g and $g^{(0)}(\mu) = g(\mu)$.

The first order delta method approximation to $g(X)$ is

$$g(X) \approx g(\mu) + g'(\mu)(X-\mu),$$

and the second order approximation is

$$g(X) \approx g(\mu) + g'(\mu)(X-\mu) + \frac{g''(\mu)}{2}(X-\mu)^2.$$

An approximation to $\mathrm{E}\big(g(X)\big)$ is obtained by taking the expectation of a low order Taylor approximation to $g(X)$.

The first order (one-term) approximation is $\boxed{\mathrm{E}\big(g(X)\big) \approx g(\mu)}$:

$$\begin{aligned}
\mathrm{E}\big(g(X)\big) &\approx \mathrm{E}\big(g(\mu) + g'(\mu)(X-\mu)\big) \\
&= \mathrm{E}\big(g(\mu)\big) + g'(\mu)\,\mathrm{E}(X-\mu) \\
&= g(\mu) + g'(\mu)\big(\mathrm{E}(X) - \mathrm{E}(\mu)\big) \\
&= g(\mu) + g'(\mu)(\mu-\mu) \\
&= g(\mu).
\end{aligned}$$

The second order (two-term) approximation is $\boxed{\mathrm{E}\big(g(X)\big) \approx g(\mu) + \frac{g''(\mu)}{2}\mathrm{Var}(X)}$:

$$\begin{aligned}
\mathrm{E}\big(g(X)\big) &\approx \mathrm{E}\left(g(\mu) + g'(\mu)(X-\mu) + \frac{g''(\mu)}{2}(X-\mu)^2\right) \\
&= \mathrm{E}\big(g(\mu)\big) + g'(\mu)\,\mathrm{E}(X-\mu) + \frac{g''(\mu)}{2}\,\mathrm{E}\big((X-\mu)^2\big) \\
&= g(\mu) + 0 + \frac{g''(\mu)}{2}\mathrm{Var}(X) \\
&= g(\mu) + \frac{g''(\mu)}{2}\mathrm{Var}(X).
\end{aligned}$$

The one-term approximation to $\mathrm{Var}\big(g(X)\big)$ is $\boxed{\mathrm{Var}\big(g(X)\big) \approx \big(g'(\mu)\big)^2\,\mathrm{Var}(X)}$:

$$\begin{aligned}
\mathrm{Var}\big(g(X)\big) &\approx \mathrm{Var}\big(g(\mu) + g'(\mu)(X-\mu)\big) \\
&= \big(g'(\mu)\big)^2\,\mathrm{Var}(X-\mu) \\
&= \big(g'(\mu)\big)^2\,\mathrm{Var}(X).
\end{aligned}$$

Note: These approximations are only good if X has a high probability of being close to μ (we want $\sqrt{\mathrm{Var}(X)}$ (standard deviation) to be small compared to μ).

Exercise 5 Delta Method

(a) Find a two-term approximation for $\mathrm{E}\big(g(X)\big)$ and a one-term approximation to $\mathrm{Var}\big(g(X)\big)$, where X is an $\mathrm{Exp}(\lambda)$ random variable and $g(X) = (1+X)^{-1}$.

(b) Find the mean and variance of $Y = 2X - 3$ in terms of the mean and variance of X.

(c) Suppose the random variable X has the following probability density function

$$f_X(x) = \begin{cases} 3x^2 & 0 < x \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

Find the two-term approximation to $\mathrm{E}\big(g(X)\big)$, where $g(X) = 10\ln(X)$.

7.3 Random Vectors

In the previous sections we considered the probability distribution of a random variable X, a single real value, but the real world is multidimensional. Not just the three spatial dimensions, either, as any quantity that can be measured can be another dimension with its own variation and uncertainty.

In this section we consider multivariate distributions of random vectors

$$(X_1, X_2, \ldots, X_m) \qquad (m\text{-tuples}).$$

We will focus on a pair of continuous random variables (X, Y), called a continuous bivariate random vector. For example, $X =$ thickness and $Y =$ tensile strength.

Joint Probability Density Function (***PDF***): We call the function $f_{X,Y}(x, y)$ a PDF for the random vector (X, Y) if

1. $f_{X,Y}(x, y) \geq 0$ for all (x, y);
2. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x, y)\,\mathrm{d}x\,\mathrm{d}y = 1$; and
3. $\mathrm{P}\big((X, Y) \in A\big) = \iint_A f_{X,Y}(x, y)\,\mathrm{d}x\,\mathrm{d}y$, where A is an event in the xy-plane.

The probability density function $f_{X,Y}$ defines a surface above the xy-plane, and *probabilities are volumes* under this surface.

The ***joint distribution function*** (***JDF*** or ***JCDF***) of the random vector (X, Y) is

$$\begin{aligned} F_{X,Y}(x, y) &= \mathrm{P}(X \leq x, Y \leq y) \\ &= \int_{-\infty}^{y}\int_{-\infty}^{x} f_{X,Y}(u, v)\,\mathrm{d}u\,\mathrm{d}v, \end{aligned}$$

where $f_{X,Y}$ is a PDF.

Properties:

1. $0 \leq F_{X,Y}(x, y) \leq 1$;
2. $F_{X,Y}(x, y)$ is a non-decreasing function of both x and y;
3. $F_{X,Y}(x, y) \to 1$ as $x \to \infty$ and $y \to \infty$; and
4. $F_{X,Y}(x, y) \to 0$ as $x \to -\infty$ and $y \to -\infty$.

Exercise 6

Let (X, Y) be a continuous random vector that is uniformly distributed on the unit square with the following PDF:

$$f_{X,Y}(x, y) = \begin{cases} 1 & \text{if } (x, y) \in [0, 1]^2 \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find the distribution function, $F_{X,Y}$, for any $(x, y) \in [0, 1]^2$.

(b) Find $\mathrm{P}(X \leq 1/3, Y \leq 1/2)$.

(c) Let $A = [1/4, 1/2] \times [1/3, 2/3]$. Find $\mathrm{P}\big((X, Y) \in A\big)$.

For more complicated PDFs and events (it's not hard to think of one!) the integration gets harder, but the idea is the same:

$$\mathrm{P}\big((X, Y) \in A\big) = \text{volume under the PDF over } A.$$

7.3.1 Marginal Distributions

Given a bivariate random vector (X, Y) with PDF $f_{X,Y}$, the marginal distribution of X describes the probability density of X irrespective of the value Y. We say "Y has been marginalised out".

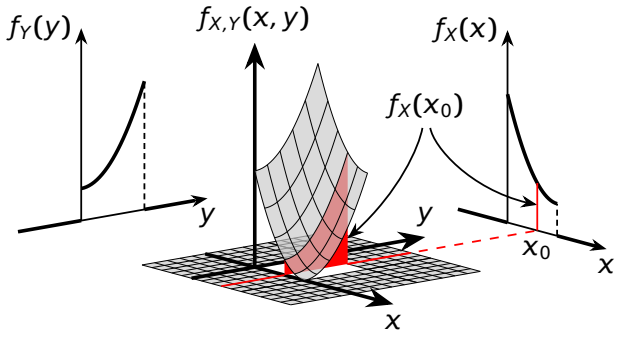
Marginal PDF: The marginal PDF for X is

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

and the marginal PDF for Y is

$$f_Y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(x, y)\,\mathrm{d}x.$$

For example:



Exercise 7 Marginals

Find the marginal PDFs $f_X(x)$ and $f_Y(y)$ from the PDF in Exercise 6 (p29).

7.3.2 Independence of Random Variables

In EMTH119 we covered independence of *events*. Here we extend this idea to random variables. Two random variables are said to be independent if and only if

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

for every $(x,y) \in \mathbb{R}^2$, or equivalently

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

for every $(x,y) \in \mathbb{R}^2$. Independence can arise in two distinct ways. We either explicitly assume that random variables are independent (think of tossing a coin twice, the coin has no memory of the first toss) or we derive independence.

Exercise 8 Independence

Let (X,Y) be the uniform random vector introduced in Exercise 6 (p29). Are X and Y independent?

Exercise 9 PDF From Marginals

If X and Y are independent random variables with PDFs defined below, what is their joint probability density function?

$$f_X(x) = \begin{cases} 2e^{-2x} & \text{if } x > 0 \\ 0 & \text{otherwise,} \end{cases} \qquad f_Y(y) = \begin{cases} 3e^{-3y} & \text{if } y > 0 \\ 0 & \text{otherwise.} \end{cases}$$

7.3.3 Expectation with Multiple Random Variables

The expectation of a function $g(X, Y)$ of a random vector (X, Y) is defined as

$$\mathrm{E}\big(g(X,Y)\big) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g(x,y) f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y.$$

The expected values of X and Y are found by setting $g(X, Y) = X$ and $g(X, Y) = Y$, respectively. If X and Y are **independent** random variables, then the following properties hold:

1. $\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y)$
2. $\mathrm{E}\big(h_1(X)\,h_2(Y)\big) = \mathrm{E}\big(h_1(X)\big)\,\mathrm{E}\big(h_2(Y)\big)$
   (h_1 is a function of X, h_2 is a function of Y)
3. $\mathrm{Var}(aX + bY + c) = a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y)$

If X and Y are **independent or dependent** random variables, then the following properties hold:

1. $\mathrm{E}\big(a\,g(X,Y)\big) = a\,\mathrm{E}\big(g(X,Y)\big)$, for $a \in \mathbb{R}$
2. $\mathrm{E}\big(g(X,Y) + h(X,Y)\big) = \mathrm{E}\big(g(X,Y)\big) + \mathrm{E}\big(h(X,Y)\big)$
3. $\mathrm{E}(aX + bY + c) = a\,\mathrm{E}(X) + b\,\mathrm{E}(Y) + c$, for $a, b, c \in \mathbb{R}$
4. $\mathrm{Var}(aX + bY + c) = a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y) + 2ab\,\mathrm{Cov}(X,Y)$, for $a, b \in \mathbb{R}$,
   where $\mathrm{Cov}(X, Y)$ is the covariance of X and Y (defined in the next section).

Exercise 10 Expectation and Variance

Let $X \sim \mathrm{Normal}(2, 4)$, $Y \sim \mathrm{Normal}(-1, 2)$, and $U \sim \mathrm{Normal}(0, 1)$ be jointly independent random variables.

(a) Find $\mathrm{E}(3X - 2Y)$.

(b) Find $\mathrm{Var}(2Y - 3U)$.

7.3.4 Covariance

The ***covariance*** of X and Y is defined as

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathrm{E}\Big(\big(X-\mathrm{E}(X)\big)\big(Y-\mathrm{E}(Y)\big)\Big)\\ &= \mathrm{E}(XY)-\mathrm{E}(X)\,\mathrm{E}(Y),\end{aligned}$$

and is a measure of the joint variability of X and Y. A positive value indicates that an increase in X mainly corresponds to an increase in Y. A negative value indicates that an increase in X mainly corresponds to a decrease in Y. However, the magnitude of covariance is difficult to interpret. A scaled version of covariance is the ***correlation coefficient***

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

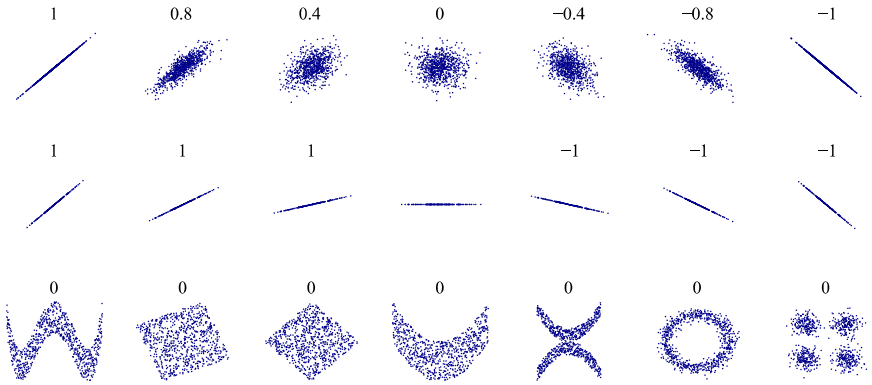
provided $\mathrm{Var}(X)$ and $\mathrm{Var}(Y)$ are nonzero. The correlation coefficient satisfies

$$-1 \le \mathrm{Corr}(X,Y) \le 1,$$

and measures the strength of a linear relationship between X and Y.

If X and Y are independent, $\mathrm{E}(XY)=\mathrm{E}(X)\,\mathrm{E}(Y)$, so $\mathrm{Cov}(X,Y)=0$ and $\mathrm{Corr}(X,Y)=0$.

$\mathrm{Corr}(X,Y)$	X and Y are:
1	perfect positive correlation: X is a linear combination of Y
-1	perfect negative correlation: X is a linear combination of Y
0	uncorrelated (in a linear manner)
> 0	positive correlation: the closer to 1, the stronger the linear relationship
< 0	negative correlation: the closer to -1, the stronger the linear relationship



Exercise 11 Linear Correlation

Find the correlation of X and $Y = a + bX$, where $a, b \in \mathbb{R}$ with $b \neq 0$.

Exercise 12 Independent Correlation

If X and Y are independent random variables, what is their covariance and correlation?

7.3.5 Multivariate Random Vectors

The bivariate random vector definitions naturally extend to the m-variate case for the random vector $(X_1, X_2, \ldots, X_m)$. For this course, we only need the definition of jointly independent random variables (for the next section and for maximum likelihood estimation). Random variables $X_1, X_2, \ldots, X_m$ are said to be jointly independent if and only if

$$f_{X_1,X_2,\ldots,X_m}(x_1, x_2, \ldots, x_m) = \prod_{i=1}^{m} f_{X_i}(x_i),$$

for every $(x_1, x_2, \ldots, x_m) \in \mathbb{R}^m$, where $f_{X_i}(x_i)$ is the marginal PDF for X_i. For example, the marginal PDF of X_1 is

$$f_{X_1}(x_1) = \int_{x_2=-\infty}^{\infty} \ldots \int_{x_m=-\infty}^{\infty} f_{X_1,X_2,\ldots,X_m}(x_1, x_2, \ldots, x_m)\, \mathrm{d}x_2 \ldots \mathrm{d}x_m.$$

We are integrating away all the other variables, so that (here) x_1 is the only variable left. The other marginals for the are found the same way: $f_{X_i}(x_i)$ is calculated by integrating $f_{X_1,X_2,\ldots,X_m}$ with respect to every variable *except* x_i.

7.3.6 Linear Combination of Normal Random Variables

If $X_1, X_2, \ldots, X_m$ are jointly independent normal random variables, i.e.

$$X_i \sim \text{Normal}\left(\mu_i, \sigma_i^2\right),$$

then

$$Y = c + \sum_{i=1}^{m} a_i X_i$$

is *also* a normal random variable:

$$Y \sim \text{Normal}\left(c + \sum_{i=1}^{m} a_i \mu_i,\ \sum_{i=1}^{m} a_i^2 \sigma_i^2\right),$$

where $c, a_1, \ldots, a_m \in \mathbb{R}$.

Exercise 13 Combining Normal Variables

Let $X \sim \text{Normal}(2, 4)$, $Y \sim \text{Normal}(-1, 2)$, and $U \sim \text{Normal}(0, 1)$ be jointly independent random variables.

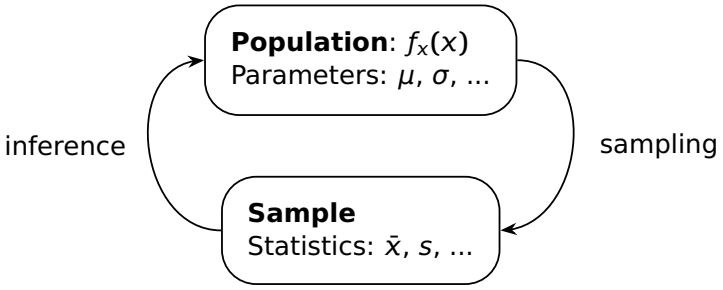
(a) Find the distribution of $K = 6 - 2U + X - Y$.

(b) Find the probability that $K = 6 - 2U + X - Y > 0$.

(c) Find $\text{Cov}(X, W)$, where $W = X - Y$.

7.4 Statistics

Most statistical approaches use a sample of observations (data) to infer the distribution that generated the data or some feature of the distribution such as its mean. A ***sample*** is a representative group from a larger population.



Statistical Approach to Learning.

Why sample at all?

1. **Cost:** too expensive to conduct a complete census.
2. **Speed:** sometimes you need a fast result.
3. **Accuracy:** precise measurements for a small sample can be better than imprecise measurements in a larger sample.
4. **Necessity:** sometimes measuring an object destroys it!

7.4.1 Simple Random Sample

If $X_1, X_2, \ldots, X_n$ are independent and each has the same marginal distribution F, we say that $X_1, X_2, \ldots, X_n$ are ***independent and identically distributed*** (***iid***) random variables, denoted

$$X_1, X_2, \ldots, X_n \sim F.$$

We also call $X_1, X_2, \ldots, X_n$ a ***simple random sample*** (***SRS***).

When we perform an experiment, we get an outcome, which we turn into a real number; a realisation of a random variable.

A realisation of an SRS gives us n observations of a random variable, denoted

$$x_1, x_2, \ldots, x_n.$$

These are measurements – *numbers*.

7.4.2 Statistics

A ***statistic*** is any function of the sample data, for example, the sample mean or the sample variance.

Collecting a sample of measurements is itself an experiment, and calculating a statistic from the sample gives us a real number; a realisation of another random variable. *Statistics are random variables.*

One important type of statistic is an ***estimator***: a rule for calculating an estimate of a given quantity based on sample data. For example, the sample mean is a ***point estimator*** (*best guess*) of the population mean (see next section). We denote a point estimator of θ by $\widehat{\theta}$, where θ is a fixed, unknown quantity called a parameter. For example, for an exponential random variable, we will be able to find $\widehat{\lambda}$, a point estimator of λ, the true (but unknown) value of the parameter.

Different samples contain different observations so each sample will produce a different value for the estimator $\widehat{\theta}$: an estimator is a random variable.

The distribution of $\widehat{\theta}$ is called the ***sampling distribution***. The standard deviation of a sampling distribution is called a ***standard error*** ,

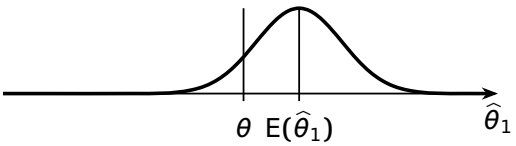
$$\text{se}(\widehat{\theta}) = \sqrt{\text{Var}(\widehat{\theta})}.$$

Often, the $\text{Var}(\widehat{\theta})$ is an unknown quantity but we can usually estimate it to give an estimated standard error denoted by $\widehat{\text{se}}$.

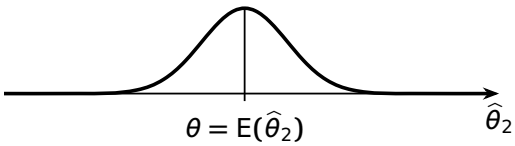
The ***bias*** of an estimator is how wrong it is, on average, defined

$$\text{Bias}(\widehat{\theta}) = \text{E}(\widehat{\theta}) - \theta,$$

and an estimator is called unbiased if $\text{E}(\widehat{\theta}) = \theta$.



a biased estimator



an unbiased estimator

Consider estimating a parameter θ using two different estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$, both with normal sampling distributions. We see that $\widehat{\theta}_1$ is a biased estimator of θ because $\text{E}(\widehat{\theta}_1) \neq \theta$. In this case, $\widehat{\theta}_1$ is more likely to overestimate θ.

7.4.3 Sample Mean

The most common statistic is the sample mean.

Let $X_1, X_2, \ldots, X_n$ be a simple random sample. Then the ***sample mean*** is

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

A *realisation* of the random variable $\bar{X}$ is

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

where $x_1, x_2, \ldots, x_n$ is an observed sample.

The sample mean is a point estimator of the population mean, μ. The expected value of $\bar{X}$ is

$$\begin{aligned}
\mathrm{E}(\bar{X}) &= \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n}\mathrm{E}(X_1 + X_2 + \ldots + X_n) \\
&= \frac{1}{n}\big(\mathrm{E}(X_1) + \mathrm{E}(X_2) + \ldots + \mathrm{E}(X_n)\big) \\
&= \frac{1}{n}(\mu + \mu + \ldots + \mu) \qquad (\mathrm{E}(X_i) = \mu \text{ for any } i \text{ (iid)}) \\
&= \frac{1}{n}(n\mu) \\
&= \mu.
\end{aligned}$$

The variance of the sample mean is

$$\begin{aligned}
\mathrm{Var}(\bar{X}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\mathrm{Var}(X_1 + X_2 + \ldots + X_n) \\
&= \frac{1}{n^2}(\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \ldots + \mathrm{Var}(X_n)) \qquad (\textbf{i}\text{id}) \\
&= \frac{1}{n^2}\big(\sigma^2 + \sigma^2 + \ldots + \sigma^2\big) \qquad (\text{i}\textbf{i}\text{d}) \\
&= \frac{1}{n^2}\big(n\sigma^2\big) \\
&= \frac{\sigma^2}{n} \qquad \left(= \frac{\mathrm{Var}(X)}{n}\right).
\end{aligned}$$

The standard error (p49) of $\bar{X}$ is $\mathrm{se} = \sqrt{\mathrm{Var}(\bar{X})} = \sigma/\sqrt{n}$.

So, the sample mean is an *unbiased point estimator* of μ and its standard error decreases as the sample size increases.

7.4.4 Sample Variance

An unbiased estimator of the population variance, σ^2, is the sample variance.

Let $X_1, X_2, \ldots, X_n$ be a simple random sample. Then the ***sample variance*** is

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

and the sample standard deviation is $S = \sqrt{S^2}$. A realisation of the random variable S^2 is

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

where $x_1, x_2, \ldots, x_n$ is an observed sample and $\bar{x}$ is the observed sample mean.

Warning: It's tempting to write $\frac{1}{\boldsymbol{n}}\sum_{i=1}^{n}(X_i - \bar{X})^2$, but this turns out to be a *biased* estimator, whose expected value is $\frac{n-1}{n}\sigma^2$, not σ^2. Correcting this by multiplying by $\frac{n}{n-1}$ gives the formula above.

7.4.5 Empirical Distribution Function

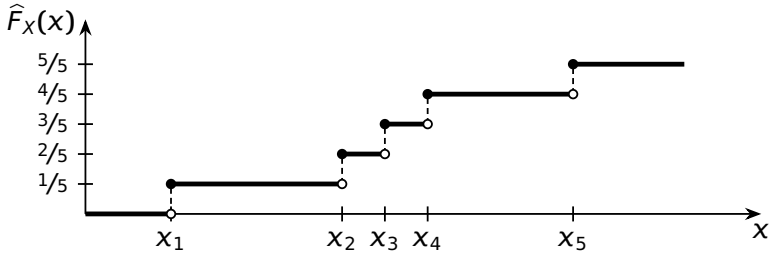
Let $X_1, X_2, \ldots, X_n \sim F$ be a simple random sample. We usually don't know the distribution function F, but we can estimate F with the ***empirical distribution function*** (***EDF***).

The EDF is a piecewise constant function that goes up a step of $1/n$ for each sample point.

That is,

$$\widehat{F}(x) = \frac{\text{number of elements in the sample} \leq x}{n}.$$

It looks like a discrete variable's distribution function, usually with all steps the same height:



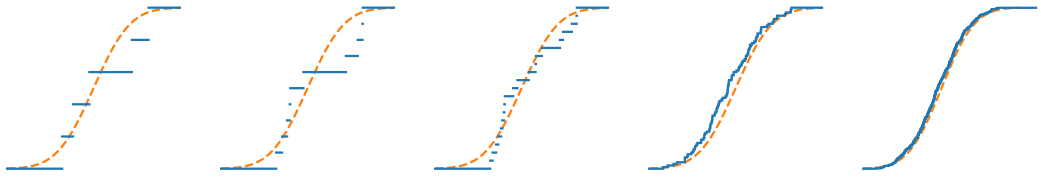
Exercise 14 Empirical Distribution Function

An SRS of $n = 6$ EMTH students was selected and the number of lectures missed by each student was recorded. (This example is discrete, but the EDF is the same for continuous random variables.) The observed sample data are:

$$0,\ 3,\ 2,\ 2,\ 1,\ 5.$$

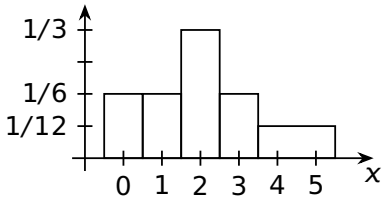
Sketch the EDF for this data.

As the sample size increases, the empirical distribution function — converges to the true distribution function - - -:



$n = 5$ $n = 10$ $n = 20$ $n = 100$ $n = 500$

A ***density histogram*** (area equal to one) can be used to approximate a PDF ($f_X(x)$). A sketch of a density histogram for Exercise 14 is:



You should use computer software to plot a density histogram and you will not be asked to sketch one by hand in this course, e.g.

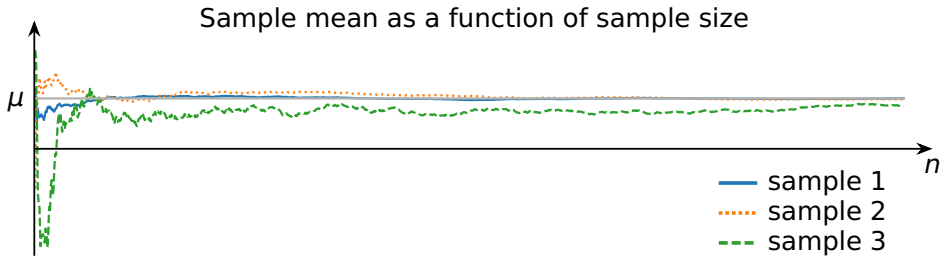
Matlab: `histogram(data, 'Normalization', 'pdf');`

Python (Matplotlib): `plt.hist(data, density=True)`

7.4.6 Law of Large Numbers

The ***law of large numbers*** (***LLN***) says that the mean of a large sample is close to the mean of the distribution. Let $X_1, X_2, \ldots, X_n$, be an SRS and let $\mu = \mathrm{E}(X_i)$ for any i (they're all the same), then the sample mean $\bar{X}$ converges in probability to μ as $n \to \infty$.

Interpretation: The distribution of $\bar{X}$ becomes more concentrated around μ as n gets large.



The trajectories of three sample means as function of sample size. Samples 1 and 2 converge to μ, whereas sample 3 requires more observations to converge. The LLN guarantees convergence, but does not give a rate of convergence.

Exercise 15 Restaurant Spending

Couples eating at a restaurant spend on average \$80.

(a) Which of the following is more likely?

(1) The average amount spent by the next 10 couples is between \$70 and \$90.
(2) The average amount spent by the next 50 couples is between \$70 and \$90.

(b) Which is more likely?

(1) The average amount spent by the next 10 couples is more than \$90.
(2) The average amount spent by the next 50 couples is more than \$90.

The next question is how large should the sample size be to get useful results? I.e. how does the sample size relate to the error in $\bar{x}$, our estimate of μ? We want to be able to say

$$\mathrm{P}\left(|\bar{X}-\mu|<\epsilon\right)=1-\alpha$$

for some high probability, $1-\alpha$. To calculate this probability, we need to know the full distribution of $\bar{X}$.

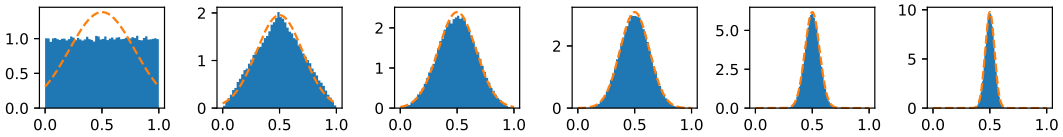
7.4.7 Central Limit Theorem

The ***central limit theorem*** (***CLT***) says that the sample mean has a distribution which is approximately normal (*nothing* is assumed about the distribution of X_i, other than a finite mean and finite variance!). Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed with finite mean μ and finite variance σ^2. Then the sample mean $\bar{X}$ converges in distribution to $\mathrm{Normal}(\mu, \sigma^2/n)$. Or equivalently, after standardisation,

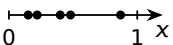
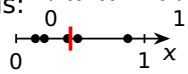
$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \rightsquigarrow \mathrm{Normal}(0,1) \text{ as } n \to \infty.$$

"$\rightsquigarrow$" means "converges in distribution". You can think of it as "goes to".

Here is an example of the distribution of sample means converging to the theoretical normal distribution as n gets larger (based on a Uniform$(0,1)$ distribution):



The process to produce this was, for each n:

1. Generate 100,000 samples of size n, like this: 
2. Find the sample mean of each sample, e.g. 
3. Plot the density histogram of these 100,000 sample means  , together with the theoretical bell curve ---.

Interpretation: Probability statements about $\bar{X}$ can be approximated using a normal distribution. Hence,

$$\begin{aligned}
\mathrm{P}(|\bar{X}-\mu|<\epsilon) &= \mathrm{P}(-\epsilon<\bar{X}-\mu<\epsilon) \\
&= \mathrm{P}\left(\frac{-\epsilon}{\sigma/\sqrt{n}}<\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}<\frac{\epsilon}{\sigma/\sqrt{n}}\right) \\
&= \mathrm{P}\left(\frac{-\epsilon}{\sigma/\sqrt{n}}<Z<\frac{\epsilon}{\sigma/\sqrt{n}}\right),
\end{aligned}$$

where Z is approximately $\mathrm{Normal}(0,1)$. We can calculate this probability using tables, and hence, calculate the probability of making an error, $|\bar{X}-\mu|$, within a specified size, ϵ.

Exercise 16 Emissions Measurement Errors

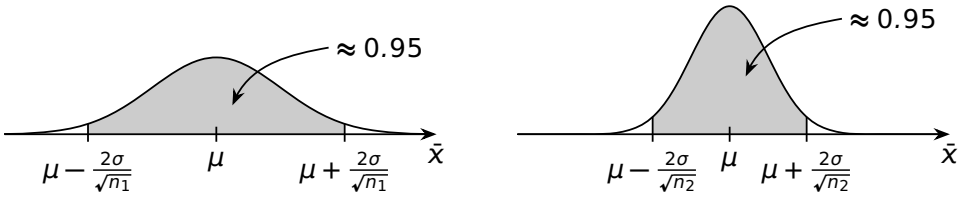
Suppose we have an SRS of 80 daily emissions measurements (in tons of sulphur oxides) from an industrial plant. The standard deviation of daily emissions is known to be 6 tons. What is the probability that an estimate of the average daily emissions will have an error less than 0.5 tons?

7.4.8 Sample Size

More data means a better estimate. We can use the CLT to calculate the required sample size to achieve an error, $|\bar{X}-\mu|$, less than ϵ with high probability. The basic process is:

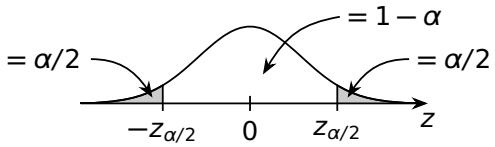
1. We choose the error threshold ϵ (which will depend on our application).
2. We choose the probability of getting an error greater than ϵ in magnitude that we are willing to tolerate, α.
3. We use the CLT to find the required sample size, n, to achieve this.

The CLT tells us that $\bar{X} \sim \text{Normal}(\mu, \sigma^2/n)$. Therefore, the variance of the sampling distribution of $\bar{X}$ decreases as the sample size increases. Hence, we can choose n large enough that the probability of achieving an error less than ϵ is high. This is illustrated here:



The sampling distribution of $\bar{X}$ for different sample sizes, where $n_1 < n_2$. The larger the sample size, the more likely that $\bar{X}$ is close to μ.

To calculate the required sample size, we use the standard normal distribution:



$z_{\alpha/2}$ denotes the z score with an area of $\alpha/2$ to its right.

We want

$$P(-\epsilon < \bar{X} - \mu < \epsilon) = 1-\alpha$$

$$P\left(\frac{-\epsilon}{\sigma/\sqrt{n}} < Z < \frac{\epsilon}{\sigma/\sqrt{n}}\right) = 1-\alpha$$

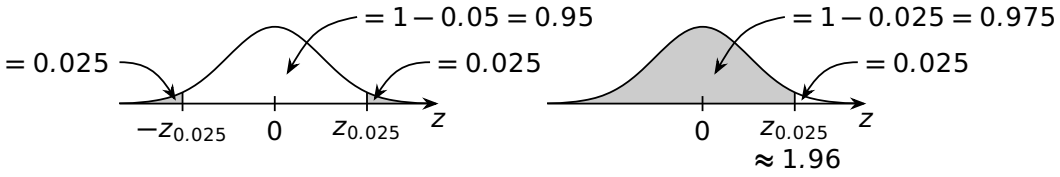
and we know (by the definition of $z_{\alpha/2}$) that

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha,$$

so we set $\dfrac{\epsilon}{\sigma/\sqrt{n}} = z_{\alpha/2}$ and rearrange to give $n = \left(\dfrac{\sigma z_{\alpha/2}}{\epsilon}\right)^2$.

This will usually need to be rounded *up* to a whole number. Common choices for α are 0.1, 0.05 and 0.01, giving $z_{\alpha/2}$ values of 1.65, 1.96, 2.58, respectively. For example, if $\alpha = 0.05$ is used, $\mathrm{P}(|\bar{X} - \mu| < \epsilon) = 0.95$.

Because the α is split in two, to get $z_{\alpha/2}$ we look up $1 - \alpha/2$ (a *probability*, not a z value) in the normal distribution table. If $\alpha = 0.05$, we need $\mathrm{P}(Z < z_{\alpha/2}) = 1 - 0.05/2 = 1 - 0.025 = 0.975$, which gives $z_{\alpha/2} = z_{0.025} \approx 1.96$, since from the table $\mathrm{P}(Z \leq 1.96) \approx 0.9750$:



Exercise 17 Emissions Measurement Errors Continued

Calculate the required sample size to make the error in Exercise 16 (p61) less than 1 ton with probability 0.95.

7.4.9 Confidence Intervals

We know how to find the sample mean, $\bar{X}$, from a sample, and we know it is a point estimator of μ, but how close to μ is it? How confident can we be with this estimate?

The CLT also gives us the $(1-\alpha)$ confidence interval, a *random interval* that contains μ with probability $1-\alpha$; μ is fixed and unknown, but we will get a different interval estimate for every sample. The sample mean $\bar{X}$ is point estimator (single value) for μ and the confidence interval is an interval estimator for μ. We know

$$\mathrm{P}\left(-z_{\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1-\alpha,$$

and rearranging we get

$$\mathrm{P}\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha.$$

Hence, the random interval

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = \left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = \left(\bar{X} - z_{\alpha/2}\mathrm{se}, \bar{X} + z_{\alpha/2}\mathrm{se}\right)$$

contains μ with probability $1-\alpha$. The observed interval is

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = \bar{x} \pm z_{\alpha/2}\mathrm{se}.$$

The quantity $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ is called the "margin of error", which you will have heard of in things like TV news polls.

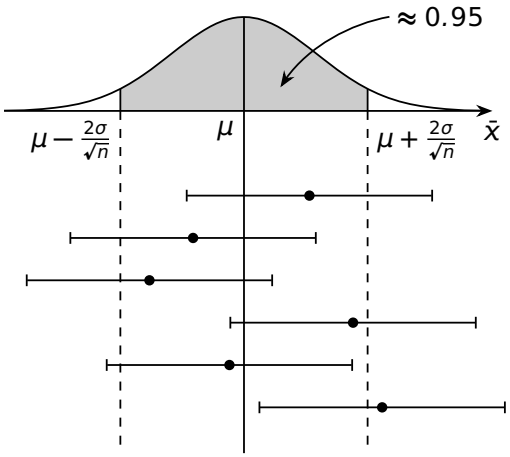
If the population standard deviation is not known, it can be replaced with the sample standard deviation, s, to give an approximate confidence interval. For 95% confidence intervals, $\alpha = 0.05$ and $z_{\alpha/2} = 1.96 \approx 2$ giving the approximate 95% confidence interval $\bar{x} \pm 2\mathrm{se}$.

Intuition: There are two possible outcomes for an observed $(1-\alpha)$ confidence interval:

1. The interval contains μ, which happens $(1-\alpha)100\%$ of the time.
2. The interval does not contain μ, which happens $\alpha 100\%$ of the time.

There is no way of knowing which outcome is true for an observed interval. Hence, a correct interpretation is: *If I repeat my experiment many times,* $(1-\alpha)100\%$ *of the* $(1-\alpha)$ *confidence intervals will contain the unknown parameter* μ*.* This is illustrated on the next slide.

A better (though more technical) interpretation of a confidence interval is: *If I perform many unrelated experiments and construct 95% confidence intervals for each parameter* $\theta_1, \theta_2, \ldots,$ 95% *of the confidence intervals will contain the true parameter value.* There is no need to force the idea of repeating the same experiment over and over.



Correct interpretation of a 95% confidence interval, $\approx \bar{x} \pm 2\sigma/\sqrt{n}$. Six 95% confidence intervals are shown, where the central dot shows the location of the observed sample mean. We see that five of the six intervals contain μ. The interval that does not contain μ has its observed sample mean more than two standard errors from μ.

Exercise 18 Grinding

The mean weight loss of 30 grinding balls after a certain length of time in mill slurry is 3.42 grams with a sample standard deviation of 0.68 grams. Find a 95% confidence interval for the population mean weight loss of such grinding balls and explain the interval in the context of the problem.

7.5 Hypothesis Testing

Suppose we want to test a manufacturer's claim that their tablet can stream HD video for at least 10 hours before the battery runs out. We select an SRS of $n = 36$ tablets and set them streaming (using the same video for each tablet) and record how long the battery lasts.

The true situation is one of two possibilities, which we call ***hypotheses***:

The **Null Hypothesis** (H_0): The average battery life is at least 10 hours

The **Alternative Hypothesis** (H_A): The average battery life is less than 10 hours

If the average battery life is much less than 10 hours, then we will conclude that the manufacturer's claim is false: we reject the null hypothesis and conclude that the evidence favours the alternative hypothesis.

This is an example of ***hypothesis testing***. Hypothesis testing is like a legal trial. We assume the accused is innocent unless the evidence *strongly* suggests that the accused is guilty. In hypothesis testing we retain the null hypothesis unless there is strong evidence to reject it.

"Strong evidence" means the probability of the observed value happening purely by chance is less than some threshold we choose, such as 5% or 1%. This threshold is called the ***level of significance*** and is denoted α.

Suppose the standard deviation of battery life for this tablet is $\sigma = 1.2$ hours, and that the observed sample mean (called the ***test statistic***) of our test is $\bar{x} = 9.65$ hours. This is lower than the claimed battery life, but may have happened by chance. What is the probability of getting a sample mean battery life this low *purely by chance*, even if the manufacturer's claim is true?

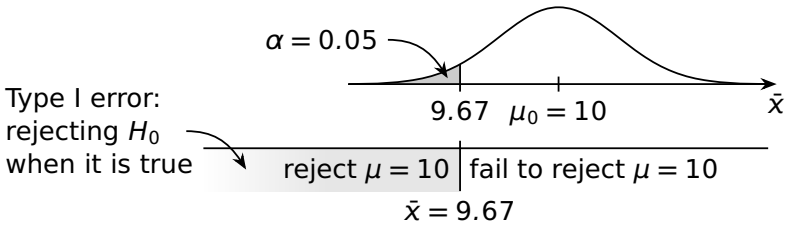
If we assume the CLT with $\mu = \mu_0 = 10$, the probability will be

$$\begin{aligned} \mathrm{P}(\bar{X} < 9.65) &= \mathrm{P}\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq \frac{9.65 - 10}{1.2/\sqrt{36}}\right) \\ &= \mathrm{P}\left(Z \leq \frac{-0.35}{0.2}\right) \\ &= \mathrm{P}(Z \leq -1.75) \\ &\approx 0.0401 \qquad \text{(tables)}. \end{aligned}$$

So if the manufacturer's claim is true, we have a 4.01% chance of getting a sample of size 36 with a sample mean of 9.65 hours or less.

This is called a ***p-value***: the probability, calculated assuming that the null hypothesis is true, that we obtain a test statistic at least as extreme as the observed value of the test statistic, just by chance. Informally, the smaller the p-value, the stronger the evidence against H_0, but a large p-value is **not** strong evidence in favour of H_0. Statistical software often provides p-values for hypothesis tests.

So "strong evidence" means a p-value less than the level of significance α. E.g. if we choose $\alpha = 0.05$, then our p-value of $0.0401 < 0.05$ constitutes strong evidence. If we instead choose $\alpha = 0.01$, we are asking for even stronger evidence.



Sampling distribution of $\bar{X}$ (assuming CLT with $\mu = \mu_0 = 10$), with a rejection region based on a level of significance of $\alpha = 0.05$: $\mathrm{P}(\bar{X} \leq 9.67) \approx 0.05$.

Type I error: If the null hypothesis is true but we reject it, we have made an error called a ***Type I error***. The probability of committing a Type I error is the level of significance we choose, α.

For the battery life example, a Type I error is concluding that the average battery life is less than 10 hours when it actually isn't, with $\alpha = 0.05$.

Here's where the 9.67 came from. Our level of significance is $\alpha = 0.05$, so there is some z_α for which $\mathrm{P}(Z < -z_\alpha) = 0.05$. From the table, we get $z_\alpha \approx 1.65$. Then we find the $\bar{x}$ value which corresponds to this z value:

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = -z_\alpha$$

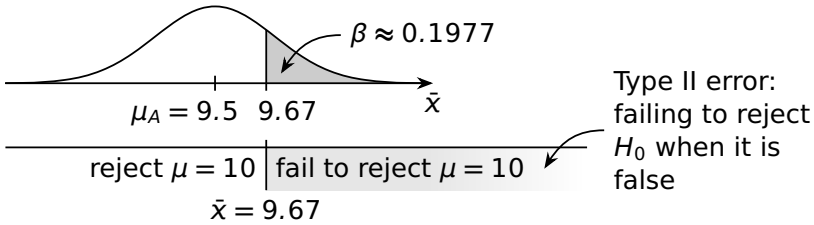
$$\frac{\bar{x} - 10}{1.2/\sqrt{36}} \approx -1.65$$

$$\bar{x} \approx 10 - 1.65\frac{1.2}{6} = 9.67$$

Suppose that the null hypothesis is *false* and the true mean battery life is actually $\mu_A = 9.5$ hours. The probability of failing to reject the null hypothesis is

$$\begin{aligned} \mathrm{P}(\bar{X} \geq 9.67) &= 1 - \mathrm{P}\left(\frac{\bar{X} - \mu_A}{\sigma/\sqrt{n}} < \frac{9.67 - 9.5}{1.2/\sqrt{36}}\right) \\ &= 1 - \mathrm{P}(Z < 0.85) \\ &\approx 0.1977 \quad \text{(tables)}. \end{aligned}$$

Hence, there is a 19.77% chance of observing a sample mean as extreme as 9.67 if $\mu_A = 9.5$.

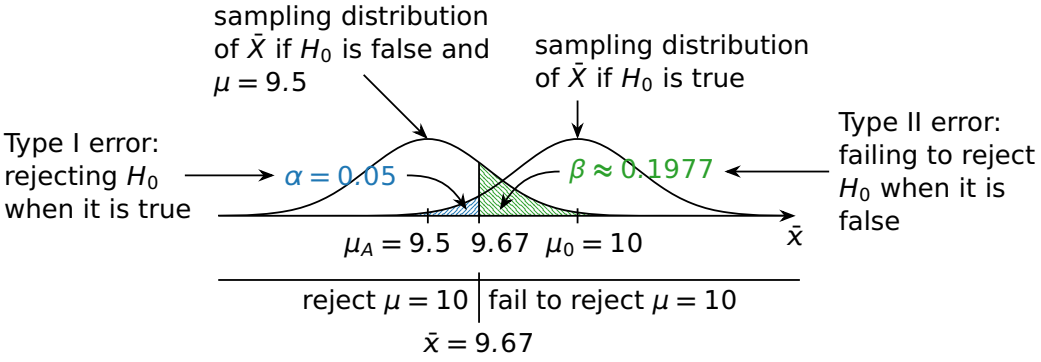


Sampling distribution of $\bar{X}$ (assuming CLT with $\mu = \mu_A = 9.5$).

Type II error: If the null hypothesis is false and it is not rejected, we have made a different error called a ***Type II error***. The probability of committing a Type II error is denoted β.

For the battery example, a Type II error is concluding that the average battery life is not less than 10 hours when it actually is, with $\beta \approx 0.1977$ if $\mu_A = 9.5$ (different μ_A values have different β values).

Summary of Outcomes:



The possible outcomes are:

	Reject H_0	Fail to Reject H_0
H_0 is true	Type I error	Correct decision
H_0 is false	Correct decision	Type II error

7.5.1 The Z-Test

In this course we are interested in tests concerning the mean. To begin, we will assume our data comes from a normal distribution (not unrealistic) with known variance σ^2 (unrealistic). In this case we know (without even needing the central limit theorem, since a linear combination of normal random variables is normal) that

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \text{Normal}(0, 1).$$

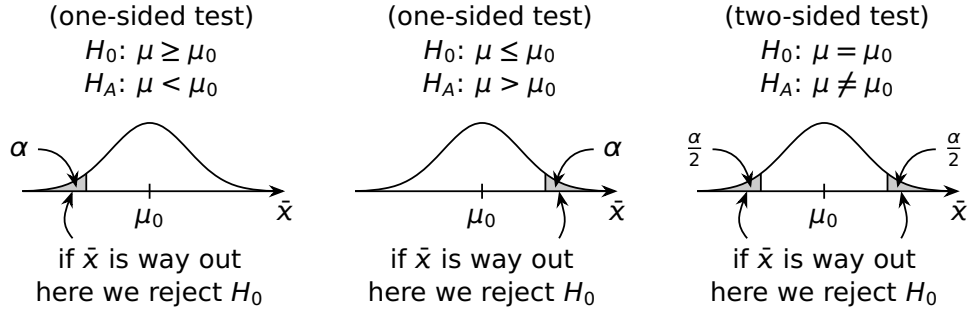
The following steps are used for the Z-test (with much more detail to follow):

1. Define a null hypothesis H_0 and an alternative hypothesis H_A.
2. Specify the probability, α, of a Type I error (H_0 is true but we reject it).
3. Define the rejection region using z_α or $z_{\alpha/2}$.
4. Compute the standardised test statistic, z^*.
5. Make the decision (whether to reject H_0).

1. **Define a null hypothesis H_0 and an alternative hypothesis H_A.**

   The ***null hypothesis***, H_0, represents the status quo, that which will be assumed to be true unless there is *strong evidence* to reject it. The ***alternative hypothesis***, H_A, contradicts the null hypothesis and will be accepted if H_0 is rejected.

   Let μ_0 be the value of the mean we are testing. There are three possible test types we are interested in:



   For one-sided tests, we test H_0: $\mu = \mu_0$, reasoning that if strong evidence exists to show that H_A is true when tested against H_0: $\mu = \mu_0$, then strong evidence exists to reject H_0: $\mu \leq \mu_0$ (or H_0: $\mu \geq \mu_0$) as well.

2. **Specify the probability, α, of a Type I error (H_0 is true & we reject it).**

   α is also called the level of significance and typical values are $\alpha = 0.05$ and $\alpha = 0.01$. α is the probability of rejecting H_0 when it's actually true.

   We are *choosing* how tolerant of this error we want to be. Avoiding Type I errors entirely is not possible, as if we make α really small, the bar for strong evidence becomes to high for a wrong H_0 to ever be rejected.

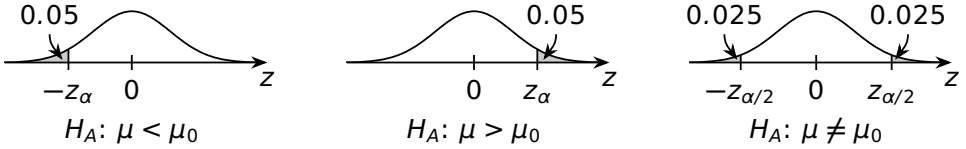
3. **Define the rejection region using z_α or $z_{\alpha/2}$.**

   Assuming that the null hypothesis is true with $\mu = \mu_0$, we have

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim \text{Normal}(0, 1).$$

   Using the sampling distribution of Z and α, we define the rejection region, the set with probability α of z values furthest from our H_0 region.

   For a one-sided test, the Type I error region is on one side, so we find z_α. For a two-sided test, the Type I error region is split into two, so we find $z_{\alpha/2}$.



The three possible rejection regions for the three possible alternative hypotheses at level of significance $\alpha = 0.05$.

4. **Compute the standardised test statistic, z^*.**

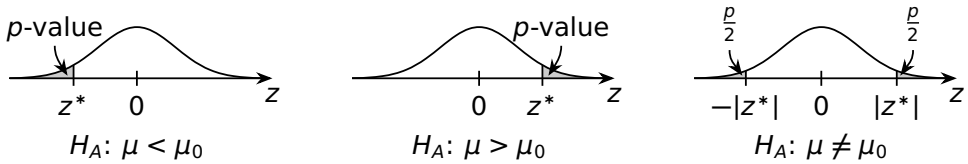
   This is the test statistic (the observed sample mean $\bar{x}$) converted to a z value:

$$z^* = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

5. **Make the decision.**

   If the test statistic is in the rejection region, we reject the null hypothesis and conclude that the alternative hypothesis is true. We know that this test will lead to this conclusion incorrectly (Type I error) $\alpha 100\%$ of the time when H_0 is true. If the test statistic is not in the rejection region, we fail to reject the null hypothesis. In this case we reserve judgement about which hypothesis is true.

   Alternatively, we can find the p-value for our z^*. If the p-value is less than α, we reject the null hypothesis and accept the alternative hypothesis.



An illustration of p-values for the possible alternative hypotheses, where z^* is the observed statistic. We reject H_0 if the p-value is less than α, which will only happen if z^* is in our rejection region.

Exercise 19 Brick Thermal Conductivity

A simple random sample of $n = 35$ cement bricks is selected from a normally distributed population with standard deviation $\sigma = 0.010$ and the thermal conductivity of each brick, x, is measured. The manufacturer claims that the average thermal conductivity is $\mu = 0.340$ and the observed sample mean is $\bar{x} = 0.343$.

(a) Using a level of significance $\alpha = 0.05$, test the manufacturer's claim.

(b) Describe a Type I error in the context of this problem.

(c) Find the p-value.

7.5.2 Large Sample Test of Hypothesis about the Population Mean

In the previous test, we knew the variance, σ^2, of the normal population we were sampling from. If σ^2 is not known, it can be replaced with its unbiased estimator, the sample variance S^2.

If the sample size is large enough (say $n > 30$, but it really depends on the population you are sampling from!), the central limit theorem will apply. Hence, the distribution of $\bar{X}$ (or Z) will be approximately normal.

Exercise 20 Truck Tires

A trucking firm is suspicious of the claim that the average lifetime of certain tires is at least 50,000 km. To test this claim, the firm selects an SRS of 40 tires and puts them on its trucks. The firm observes a sample mean lifetime of 49,083 km with a sample standard deviation of 2,301 km.

(a) Using $\alpha = 0.01$, test the claim.

(b) Describe a Type I error in the context of this problem.

(c) Find the p-value.

Exercise 21 Truck Tires Continued

Suppose the true average tire lifetime from Exercise 20 is $\mu_A = 48{,}976\,\text{km}$ (a value in the alternative hypothesis (H_0: $\mu \geq 50{,}000\,\text{km}$ is false)).

(a) Describe a Type II error in the context of Exercise 20.

(b) What is the probability of committing a Type II error in Exercise 20 if $\mu = \mu_A = 48{,}976\,\text{km}$?

7.5.3 Power

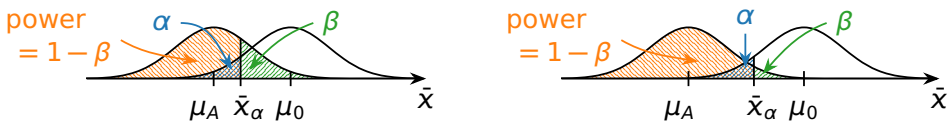
The ***power*** of a test is the probability that the test will correctly reject the null hypothesis for a particular value of μ in the alternative hypothesis. The power is equal to $(1-\beta)$ for the particular alternative considered.

Exercise 22 Truck Tires Power

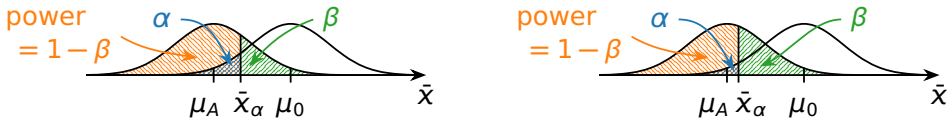
Calculate the power of the test used in Exercise 20 using the mean in the alternative hypothesis, $\mu_A = 48{,}976\,\text{km}$, from Exercise 21.

Properties of power:

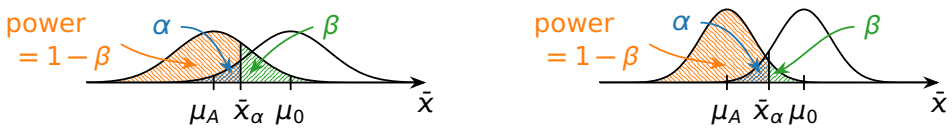
1. For fixed n and α, power increases as the distance between μ_0 and μ_A increases.



2. For fixed n and values of μ_0 and μ_A, power decreases as α is decreased.



3. For fixed α and values of μ_0 and μ_A, power increases as the sample size is increased.



7.6 Maximum Likelihood

In Section 7.4 we used point estimators to estimate features of a distribution such as the population mean, μ. In this section, we are interested in estimating a parameter value in an assumed distribution. For example, if $X_1, X_2, \ldots, X_n$ is a simple random sample from a distribution with PDF $f(x;\theta)$, **we want to estimate the parameter $\boldsymbol{\theta}$**. To estimate the parameter, we observe a simple random sample

$$x_1, x_2, \ldots, x_n,$$

and find the parameter value in the assumed distribution that makes the observed sample *most likely*. I.e., we assume the sample we got is a typical (likely) one.

We will consider estimating a *single parameter* in a PDF.

7.6.1 Likelihood Function

Let $X_1, X_2, \ldots, X_n$ be a simple random sample from a distribution that depends on an unknown parameter θ with probability density function

$$f_{X_i}(x_i;\theta).$$

Here $\theta \in \Theta$, where $\Theta \subset \mathbb{R}$ is the set of values θ can take, called the ***parameter space***. Because the X_is are independent, the joint probability density function of $X_1, X_2, \ldots, X_n$ is

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n;\theta) = \prod_{i=1}^{n} f_{X_i}(x_i;\theta).$$

However, now we have an actual sample, $x_1, x_2, \ldots, x_n$, and these are *numbers* – they are fixed and cannot change. We define another function, the ***likelihood function*** $L(\theta)$, with the same formula

$$L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i;\theta),$$

but we are now thinking of it as a function of θ (and which also depends on the **fixed** $x_1, x_2, \ldots, x_n$).

We want to find the value of θ which makes our sample as likely as possible, $\widehat{\theta}$. In other words, we want to find a *maximum of the likelihood function*.

We know how to do this: differentiate, set to 0, solve for θ, check it's a maximum. Unfortunately, that $\prod$ product makes this hard to do directly. We can turn the product into a sum by taking a log, giving us the ***log-likelihood function*** $\ell(\theta)$:

$$\ell(\theta) = \ln\big(L(\theta)\big) = \ln\left(\prod_{i=1}^{n} f_{X_i}(x_i;\theta)\right) = \sum_{i=1}^{n} \ln\big(f_{X_i}(x_i;\theta)\big).$$

It is often easier to work with the log-likelihood function: because ln is a strictly increasing function, the θ value that maximises $\ell(\theta)$ will also maximise $L(\theta)$.

Comments:

(a) Because $X_1, X_2, \ldots, X_n$ is a simple random sample, the X_i's are independent and we can find the joint PDF by multiplying all the marginals together:

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f_{X_i}(x_i; \theta).$$

(b) The PDF is seen as a function of $x_1, x_2, \ldots, x_n$ for a fixed parameter θ. In contrast, the likelihood function is seen as a function of θ for an observed sample $x_1, x_2, \ldots, x_n$. Thus, $L(\theta) : \Theta \to [0, \infty)$.

(c) The likelihood function is not a PDF: in general $\int_{-\infty}^{\infty} L(\theta)\, \mathrm{d}\theta \neq 1$.

Exercise 23 Log-likelihood for Exponential Random Variables

Find the log-likelihood function for a simple random sample of n Exp(θ) random variables.

7.6.2 Maximum Likelihood Estimation

Let the model (PDF) for the data be

$$X_i \sim f_{X_i}(x_i; \theta_A),$$

where θ_A is the actual (true) value of the parameter that we are trying to estimate. The corresponding likelihood function is

$$L(\theta) = \prod_{i=1}^{n} f_{X_i}(x_i; \theta).$$

$\widehat{\theta}$, the ***maximum likelihood estimator*** (***MLE***) for the unknown parameter θ_A, is the value of θ that maximises the likelihood function:

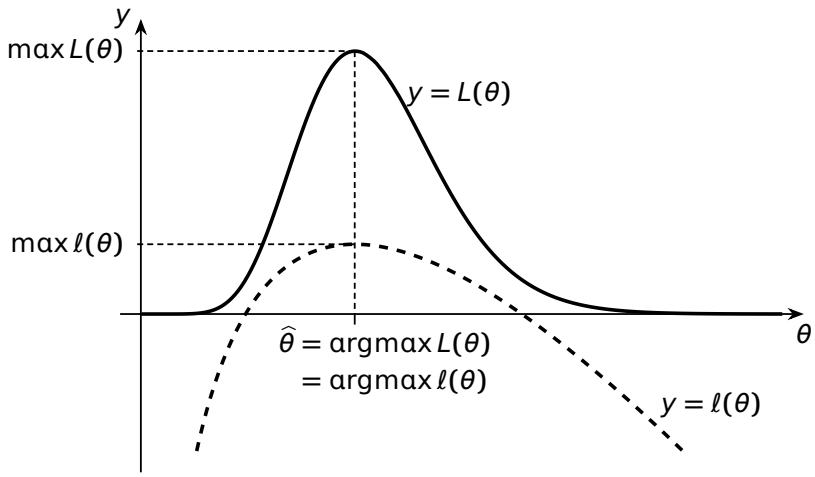
$$\widehat{\theta} = \operatorname*{argmax}_{\theta \in \Theta} L(\theta).$$

The notation argmax means $L(\theta)$ achieves its maximum value at $\widehat{\theta}$; argmax gives the *location* of the maximum, rather than the maximum value itself.

Because ln is a strictly increasing function, $\widehat{\theta}$ will maximise both $L(\theta)$ and $\ell(\theta)$, so equivalently, the MLE $\widehat{\theta}$ is the value of θ that maximises the log-likelihood function

$$\widehat{\theta} = \operatorname*{argmax}_{\theta \in \Theta} \ell(\theta).$$

It is often easier to maximise $\ell(\theta)$ rather than $L(\theta)$.



Likelihood function and the log-likelihood function for an SRS of exponential random variables. The same value of $\widehat{\theta}$ maximises both functions, but their maximum values are different.

Properties (under regularity conditions[1]):

(a) The MLE is asymptotically consistent, meaning $\widehat{\theta}$ gives the true θ_A as $n \to \infty$.

(b) For sufficiently large n,
$$\widehat{\theta} \sim \text{Normal}(\theta_A, (\widehat{\text{se}})^2),$$
where the estimated standard error is
$$\widehat{\text{se}} = \sqrt{\left(\left[-\frac{\mathrm{d}^2\ell(\theta)}{\mathrm{d}\theta^2}\right]_{\theta=\widehat{\theta}}\right)^{-1}}.$$

(c) Because $\widehat{\theta}$ is approximately normally distributed, we can calculate approximate $(1-\alpha)100\%$ confidence intervals for θ_A using
$$\widehat{\theta} \pm z_{\alpha/2}\widehat{\text{se}}.$$

Remarks:

(a) The MLE may not be unique. There is only one maximum value of the likelihood function, but there can be many values of θ that achieve the maximum value.

(b) The MLE may not exist. That is, no value of θ maximises the likelihood function.

(c) For complicated likelihood functions, numerical optimisation algorithms are used to find a maximiser.

[1] These are technical conditions that relate to the smoothness of $f(x;\theta)$. We assume these conditions hold and they are beyond the scope of this course.

7.6.3 MLE Method

Assume that we have an observed simple random sample $x_1, x_2, \ldots, x_n$ which can be modelled as a sample from the joint PDF
$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\theta_A) = \prod_{i=1}^{n} f_{X_i}(x_i;\theta_A)$$
for some unknown parameter $\theta_A \in \Theta$.

The following steps are used to find $\widehat{\theta}$, the MLE for θ_A:

1. Find an expression for the log-likelihood function
$$\ell(\theta) = \ln\left(\prod_{i=1}^{n} f_{X_i}(x_i;\theta)\right) = \sum_{i=1}^{n} \ln\left(f_{X_i}(x_i;\theta)\right).$$
2. Compute the derivative of $\ell(\theta)$ with respect to θ: $\frac{\mathrm{d}}{\mathrm{d}\theta}\ell(\theta)$.
3. Set the derivative equal to zero and solve for θ. Call this solution $\widehat{\theta}$.
4. Check that you have found a maximum of the log-likelihood function (second derivative test). If
$$\left[\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ell(\theta)\right]_{\theta=\widehat{\theta}} < 0,$$
then there is a maximum at $\theta = \widehat{\theta}$ and you have found the MLE for θ_A.
5. Provide a $(1-\alpha)100\%$ confidence interval for θ_A if required.

Exercise 24 MLE for Exponential Random Variable

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed with PDF Exp(θ_A). Find the MLE for θ_A and provide a 95% confidence interval for θ_A.

Exercise 25 Orbiter Bus Wait Time

Some UC students collected data on waiting times between buses at an Orbiter bus-stop close to campus. The observed sample mean from 132 observations was $\bar{x} = 9.08$ minutes.

(a) Assuming the waiting times follow an Exp(θ_A) distribution (reasonable assumption), find the MLE for θ_A and provide a 95% confidence interval for θ_A.

(b) What is the probability of waiting less than 15 minutes for an Orbiter bus at the bus-stop near campus?

(c) What is the expected waiting time for an Orbiter bus at the bus-stop near campus?

Exercise 26 MLE for Mean of Normal Distribution

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed with PDF Normal(θ_A, σ^2), with unknown parameter θ_A and known non-zero variance σ^2. Find the MLE for θ_A.

Exercise 27 MLE for Variance of Normal Distribution

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed with PDF Normal(μ, θ_A), with unknown parameter θ_A and known mean μ. Find the MLE for θ_A.