# Probability and Statistics

Théodore Renault

The Graduate Institute of International and Development Studies, Geneva

Events and their probabilities

Conditional probability and independence

Random variables

Bernouilli, binomial and Poisson distributions

Expectation

Two or more random variables

Random sample and limit theorems

Parameter and Interval Estimation

Hypothesis testing

# Events and their probabilities

# Events and their probabilities

Consider an experiment that can result in just one of several outcomes, which form a finite set $S$.

The probability of each outcome is a positive number indicating how likely it is to happen.

Probabilities are calibrated in such a way that the probabilities of all the outcomes sum to 1.

An **event** is a subset of $S$; the probability of an event is the sum of the probabilities of its constituent outcomes.

# Events and their probabilities

**Exercise** A coin is tossed twice. We assume that the coin is fair, in the sense that Heads (noted H) and Tails (noted T) are equally likely to appear at each toss.

(a) Give the set $S$ of this experiment.

(b) Give the probability of each outcome.

(c) Give the probability of getting exactly one Head.

**Exercise** Suppose we are interested only in the number of Heads that appear. Give the set of outcomes and their probabilities.

# Events and their probabilities

A non-empty set $S$ is called the **sample space**; its members are called outcomes and its subsets are called events.

The whole sample $S$ is sometimes referred to as the **certain event** and the empty set $\emptyset$ as the **impossible event**.

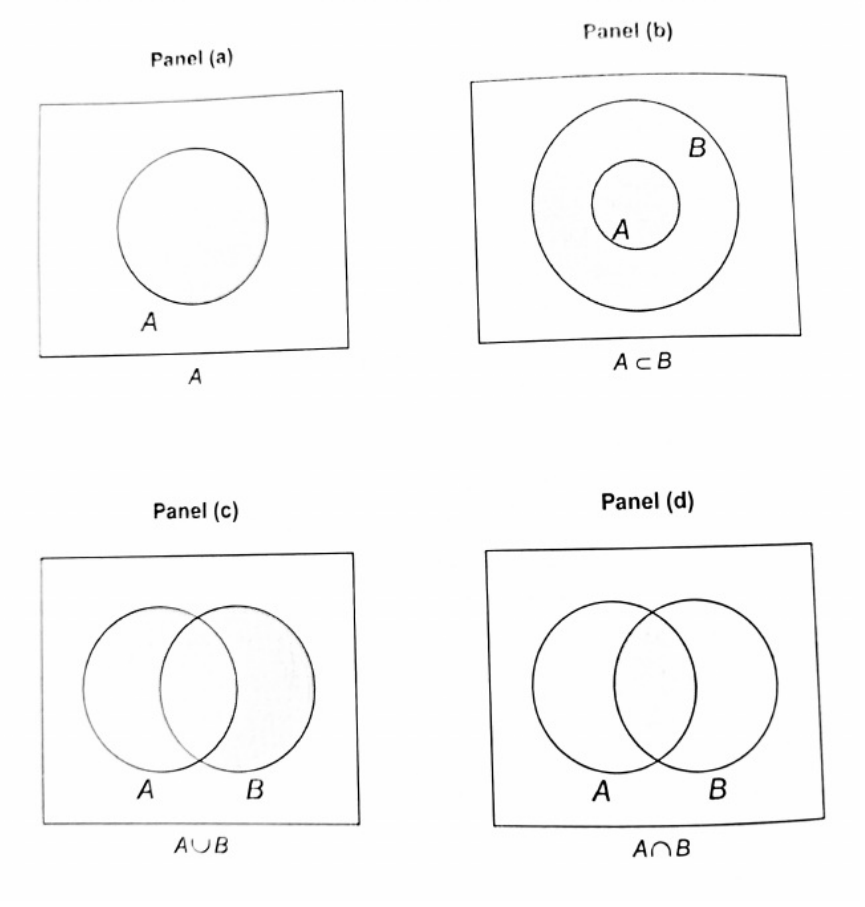
If $A$ and $B$ are events, then $A \cup B$ is the set of outcomes in at least one of $A$ and $B$, and can be interpreted as the **event that at least one of $A$ and $B$ occurs**.

$A \cap B$ is the set of outcomes that belong to both $A$ and $B$ and can be interpreted as the **event that both $A$ and $B$ occur**.

Disjoint events, in other words subsets $A$ and $B$ of $S$ with the property that $A \cap B = \emptyset$, may be interpreted as being mutually exclusive : **they cannot both occur**.

# Review of set theory

For our purpose, we make use of **Venn diagrams**.




The sample $S$ is represented by the interior of the rectangle in each plot.

# Events and their probabilities

With each event $E$ is associated a real number $P(E)$ called the probability of $S$.

The function $P$ follows the following rules :

**(P1)** Probabilities are non-negative : $P(E) \geq 0$ for every event $E$.

**(P2)** The probability of a certain event is 1 : $P(S) = 1$.

**(P3)** The probability that one of two mutually exclusive events occurs is the sum of their probabilities : if $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$.

# Additional rules

(i) $P(\emptyset) = 0$ : the probability of the impossible event is zero.

(ii) If $E$ and $F$ are events such that $E \subset F$, $E$ is a subset and $P(E) \leq P(F)$.

(iii) For any event $E$ in the sample space $S$, the complement of $E$ is the set of outcomes that are not in $E$ and is denoted by $E^c$.

$$P(E) + P(E^c) = 1$$

(iv) If the events $A_1, A_2, \ldots, A_n$ are pairwise disjoint, then

$$P(A_a \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n)$$

(v) For any events $A$ and $B$,

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

# Exercises

**Exercise** You roll a pair of dice, D1 and D2. Let $A$ be the event that D1 shows a six, B the event that D2 shows a six. Find $P(A)$, $P(B)$, $P(A \cap B)$ and $P(A \cup B)$.

**Exercise** In a certain group of men 60% are bald, 45% are fat and 30% are neither bald or fat. If a man is chosen at random from the group, what is the probability that he is both bald and fat ?

# Counting outcomes

If we have $m$ distinct objects, how many ways are there of choosing $r$ of them ?

If we think of choosing the $r$ objects one by one there are $m$ ways of choosing the first, which may be combined with any of $m-1$ ways of choosing the second and so on.

If we regard the choice of the same $r$ objects in different orders as different 'ways of choosing', the total number of ways is

$$m \times (m-1) \times \cdots \times (m-r+1) = \frac{m!}{(m-r)!}$$

# Counting outcomes

What if we are unconcerned with the order in which objects are chosen and care only about those we end up with ?

In this case, the number of possible choices is given by dividing the previous answer by $r!$, the number of ways of ordering $r$ objects.

Therefore, the number of was of choosing $r$ objects from $m$, without regard to the order in which they are chosen, is

$$\frac{m!}{r!(m-r)!}$$

This expression is usually denoted $\binom{m}{r}$.

# Conditional probability and independence

# Conditional probability and independence

Given a sample space $S$ and an event $B$ such that $P(B) > 0$, we let

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

for every event $A \subset S$. $P(A|B)$ is called the **conditional probability** if $A$ given $B$, and is interpreted as the probability assigned to $A$ when you know that $B$ has occurred.

**Exercise** A card is drawn from a standard pack. What is the probability that it is a Heart, given that is is either a Space or an Ace ?

# Bayes rule

Suppose we know $P(A|B)$ but we really want to know $P(B|A)$.

We know that

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

and we also know that

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

where $(A \cap B)$ and $(A \cap B^c)$ are two disjoint sets. Thus,

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

and

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

and

$$P(A \cap B^c) = P(A|B^c)P(B^c)$$

# Bayes rule

Therefore,

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

which is known as the **Bayes' theorem**.

**Exercise** Suppose 1 woman in every 200 has a certain disease. A test for the disease shows positive for 99% of sufferers and 5% of non-sufferers. If a randomly chosen woman is tested for the disease and shows positive, what is the probability that she has the disease ?

# Independence

Sometimes it happens that the occurrence of event $A$ does not change the probability of event $B$.

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

We say that the events $A$ and $B$ are **independent**. It follows that

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

Three events $A$, $B$, $C$ are said to be independent if

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

# Independence

**Exercise** Suppose 2 fair 6-sided dice are rolled. Let A be the event "the first roll is a 3", let B be the event "the second roll is a 6", and let C be the event "the sum of the two rolls is 6". Determine whether each of the pairs of events is independent.

# Random variables

# Random variables

A **random variable** (r.v.) is a quantity whose value depends on the result of an experiment.

More formally, a random variable is a real-valued function defined on a sample space $S$.

As with any function, the **range** of an r.v. is the set of values it can take.

If $X$ is a random variable and $U$ is a set of real numbers, the set $\{s \in S : X(s) \in U\}$ is the event that $X$ takes a value in $U$.

Given $u \in \mathbb{R}$, it is conventional to denote the event $\{s \in S : X(s) > u\}$ simply by $\{X > u\}$.

# Random variables

Let $X$ be a random variable. The **cumulative distribution function** (c.d.f) of $X$ is the function $F_X$ from $\mathbb{R}$ to $\mathbb{R}$, defined for each real number $x$ by

$$F_X(x) = P(X \leq x)$$

$F_X$ is sometimes referred to as the 'distribution' of $X$.

Since the values taken by $F_X$ are probabilities, they must all be between 0 and 1.

**Exercise** A household is chosen at random from a city in which 20% of households own no car, 55% own one and 25% own two cars. Let $X$ be the number of cars the chosen household owns.
(a) Find the values of the r.v. $X$ and their associated probabilities.
(b) Find $F_X(x)$ and sketch the function.

# Discrete Random variables

A discrete random variable is an r.v. whose range is their **finite** or **countable**.

The r.v.s of the previous exercise is discrete : its range is $\{0, 1, 2\}$.

# Continuous Random variables

Let $X$ be an r.v. whose range is contained in the closed interval $[a, b]$. Suppose there is a function $f_X$ such that $f_X(x) \geq 0$ for all $x$ and

$$F_X(x_0) = \int_a^{x_0} f_x(x)dx \quad \text{if } a \leq x_0 \leq b$$

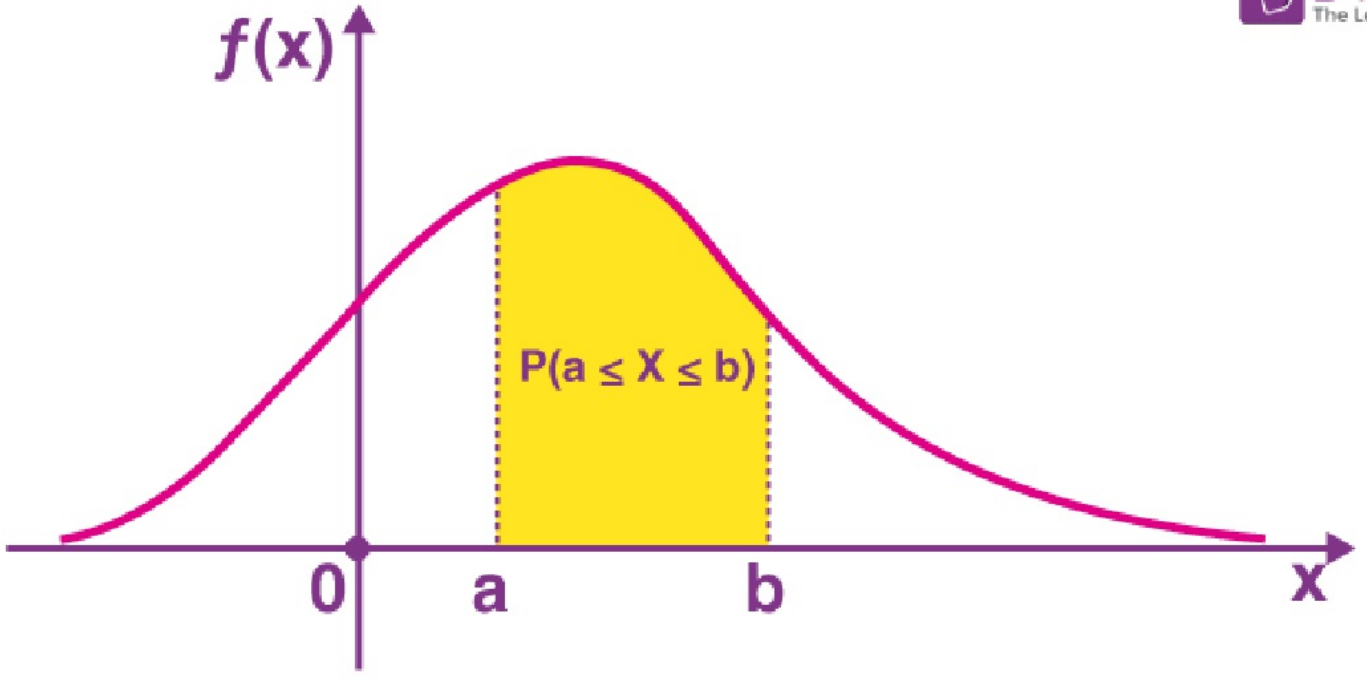
Then $X$ is said to be a continuous random variable and $f_X$ is called the density function of $X$.

It follows that

$$P(u \leq X \leq v) = \int_u^v f_X(x)dx \quad \text{if } a \leq u \leq v \leq b$$

# Continuous Random variables

Other example : $P(a \leq X \leq b) = \int_a^b f_X(x)dx$



# Continuous Random variables

If $X$ is a random variable whose c.d.f $F_X(x)$ is continuous for all $x$ and differentiable for all $x$, then $X$ is continuous and $f_X(x) = F'_X(x)$ wherever the latter is defined.

Additionally, for every continuous random variable $X$,

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

It follows that

$$\lim_{x \to -\infty} F_X(x) = 0 \text{ and } \lim_{x \to +\infty} F_X(x) = 1$$

# Continuous Random variables

**Exercise** The density function of a continuous random variable $X$ is given by

$$f_X(x) = \begin{cases} \alpha x(2-x) & \text{if } 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

Find $\alpha$ and $P(0.5 < X < 1.5)$.

# Standard normal variables

A **Gaussian random variable**, known as a **standard normal variate**, is a continuous r.v. $X$ with density function

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2)$$

The density function of a Gaussian r.v. is usually denoted by $\phi$, and the corresponding c.d.f. by $\Phi$. Thus

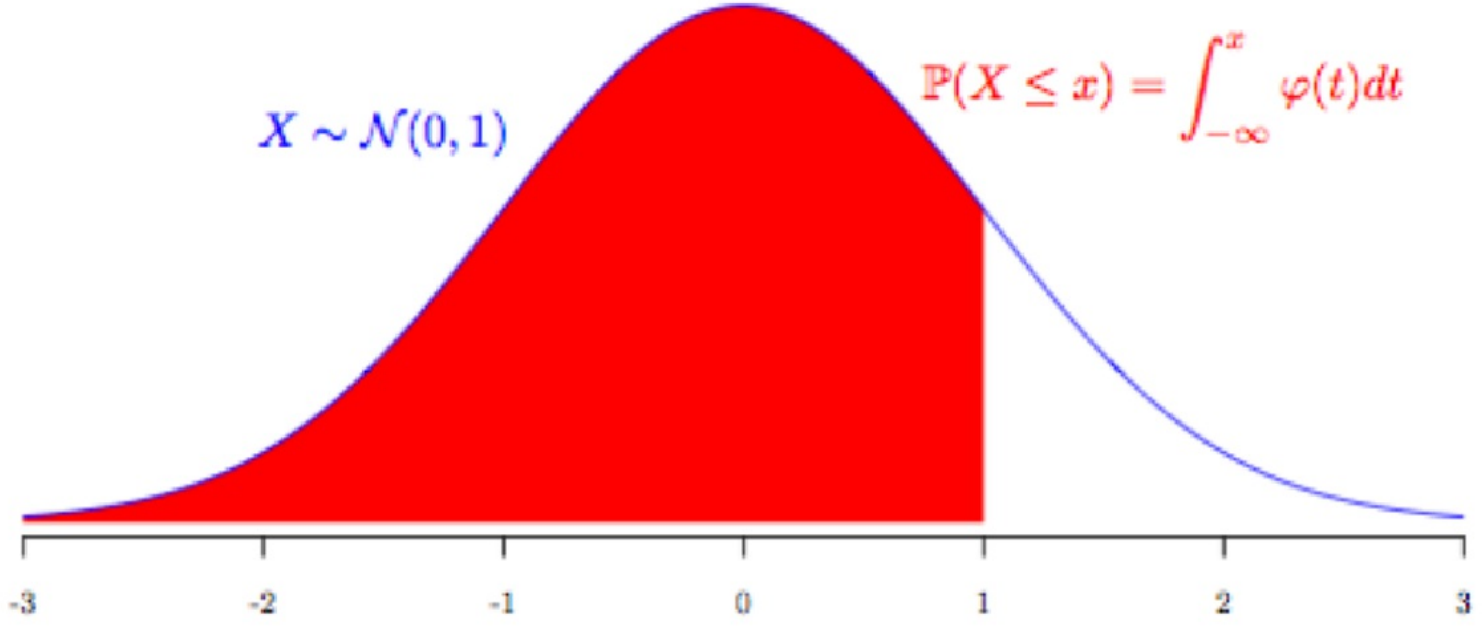
$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2) \quad \Phi(x) = \int_{-\infty}^{x} \phi(t)dt$$

and $\Phi'(x) = \phi(x)$ for all $x$.

$\phi(x)$ is symmetrical about the vertical axis :

$$\Phi(x) + \Phi(-x) = 1 \text{ for all } x$$

# Standard normal variables



# Bernouilli, binomial and Poisson distributions

# Bernouilli trials

Suppose we perform repeatedly an experiment with two possible outcomes, Success and Failure.

Each repetition is called a **trial**. The trials are said to be independent if the result of one of them does not affect any of the others.

If the trials are independent **and** the probability of Success is the same in each trial, the we have a sequence of **Bernouilli trials**.

# Binomial random variable

If we have a sequence of $n$ Bernouilli trials, each with Success-probability $p$, then the total number of Successes is called a **binomial random variable** with parameters $n$ and $p$, or a $B(n, p)$ variate, sometimes written $X \sim B(n, p)$.

If $X \sim B(n, p)$, $X$ is a discrete r.v. taking the values 0,1, ..., n

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad (k = 0, 1, \ldots n)$$

**Exercise** If a die is rolled 7 times, what is the probability that exactly 3 rolls results in a five or a six ?

# Poisson variates

Let $\lambda$ be a positive real number and let $X$ be a discrete r.v. taking the values 0, 1, 2 ...

We say that $X$ has the Poisson distribution with parameter $\lambda$ if

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} \quad (k = 0, 1, \dots)$$

The Poisson distribution provides a useful approximation to the binomial when we are dealing with rare events : if $X$ is a $B(n, p)$ variate, where $n$ is large and $p$ small, then $X$ is approximatively a Poisson variate with parameter $np$.

The approximation works very well whenever $n \geq 100$ and $np \leq 10$.

**Exercise** Suppose that 2% of people who reserve a flight fail to show up. If an airline takes 183 reservations for a flight on a plane with 180 seats, what is the probability that at least one prospective passenger is turned away ?

# Expectation

# Expected value

Let $X$ be a discrete random variable taking the values $a_1, a_2, \ldots, a_n$ with probabilities $p_1, p_2, \ldots, p_n$.

The **expected value or expectation** of $X$, denoted by $E(X)$ is defined as follows

$$E(X) = p_1 a_1 + p_2 a_2 + \cdots + p_n a_n$$

Since a random variable can take infinitely many values, we can extend the definition of expected value to such cases :

If $X$ is a discrete r.v. taking the values $a_1, a_2, \ldots$ with probabilities $p_1, p_2, \ldots$ then

$$E(X) = \sum_{i=1}^{\infty} p_i a_i$$

# Expected value

If $X$ is a continuous r.v., then

$$E(X) = \int_{-\infty}^{\infty} x f_X(x) fx$$

**Exercise** Let $X$ be uniformly distributed over the interval $[0, L]$ : then $f_X(x)$ is $1/L$ if $x$ is in that interval and 0 otherwise. Find $E(X)$.

# Expected value

If $X$ is a random variable and $g$ is a function, we can calculate $E(g(X))$.

(A) If $X$ is a discrete r.v. taking values $a_1, a_2, \ldots$ with probabilities $p_1, p_2, \ldots$, then

$$E(g(X)) = \sum_i p_i g(a_i)$$

(B) If $X$ is a continuous function random variable, then

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

**Exercise** Let $X$ be uniformly distributed over the interval $[0, L]$. Find $E(X^3)$.

# Expected value

Suppose $X$ is a random variable and $g$ and $h$ are functions. Then for any constants $a$ and $b$,

$$E(ag(X) + bh(X)) = aE(g(X)) + bE(h(X))$$

If $X$ is a continuous r.v., it follows from

$$\int_{-\infty}^{\infty} (ag(x)+bh(x))f_X(x)dx = a\int_{-\infty}^{\infty} g(x)f_X(x)dx+b\int_{-\infty}^{\infty} h(x)f_X(x)dx$$

Similarly,

$$E(aX + b) = aE(X) + b \text{ if } a \text{ and } b \text{ are constants}$$

# Variance

The **variance** of $X$, written $\text{var}(X)$, is defined by the formula

$$\text{var}(X) = E[(X - E(X))^2]$$

and measures the dispersion of a random variable.

Sometimes, we use the notation $\mu = E(X)$, then

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

The standard deviation is defined as

$$\sigma(X) = \sqrt{\text{var}(X)}$$

**Properties**

(i) $\text{var}(X) > 1$

(ii) $\text{var}(X) = E(X^2) - E(X)^2$

(iii) If $a$ and $b$ are constants, $\text{var}(aX + b) = a^2\text{var}(X)$

# The normal distribution

We say that a random variable $X$ is **normally distributed** if there exists numbers $\mu$ and $\sigma$, with $\sigma > 0$, such that $\frac{X-\mu}{\sigma}$ is a standard normal variate.

In this case

$$X = \mu + \sigma Z, \text{ where } E(Z) = 0 \text{ and } \text{var}(Z) = 1$$

It follows that $E(X) = \mu$ and $\text{var}(X) = \sigma^2$, so the standard deviation of $X$ is $\sigma$.

A normally distributed r.v. $X$ with mean $\mu$ and variance $\sigma^2$ is called an $N(\mu, \sigma^2)$ variate and is noted $X \sim N(\mu, \sigma^2)$.

# The normal distribution

**Exercise** If $X \sim N(7, 16)$, find $P(3 < X < 10)$ using the table of a standard normal distribution.

# Moment generating functions

Let $X$ be a random variable. For each positive integer $k$, $E(X^k)$ is called the $k$th **raw moment** of $X$ and $E[(X - E(X))^k]$ is called the $k$th **central moment** of $X$.

The first raw moment is the mean and the second central moment is the variance.

Moments beyond the second are often called **higher moments**.

When calculating moments of an r.v. $X$, it is often helpful to use the moment generating function of $X$, defined as follows

$$M_X(t) = E(e^{tX})$$

It can be proved that *two random variables with the same moment generating function have the same c.d.f.*

# Moment generating functions

Taking the $k$th derivative with respect to $t$

$$\frac{d^k}{dt^k} M_X(t) = E\left[\frac{d^k}{dt^k} e^{tX}\right] = E(X^k e^{tX})$$

The **raw moment theorem** states that

$$E(X^k) = \left[\frac{d^k}{dt^k} M_X(t)\right]_{t=0}$$

As a result

$$E(X) = M'(0)$$
$$E(X^2) = M''(0)$$

# Moment generating functions

**Exercise** If $X$ follows a uniform distribution on [a,b] then

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

Find the MGF of $X$. Find $E(X)$.

# Two or more random variables

# Two or more random variables

Suppose we have $n$ random variables $X_1, X_2, \ldots, X_n$ defined on the sample space, for example the height, weight, income and so on of a woman chosen at random from a population.

The probabilities associated with these r.v.s may be summarised by a function $F$ of $n$ variables known as the joint distribution function of $X_1, X_2, \ldots, X_n$ and defined as follows

$$F(x_1, x_2, \ldots, x_n) = P(X_1 \leq x_1 \text{ and } X_2 \leq x_2 \text{ and } \ldots \text{ and } X_n \leq x_n)$$

whenever $x_1, x_2, \ldots, x_n$ are real numbers.

# Two or more random variables

Let $X$ and $Y$ be random variables. Suppose for the moment that $X$ is discrete, taking value $a_1, a_2, \ldots,$ all with positive probability.

Then for $i = 1, 2, \ldots$ and any real number $y$ we can define the expression

$$H_i(y) = P(Y \leq y | X = a_i)$$

For given $i$, $H_i$ has the properties of an alternative c.d.f. for $Y$, known as the **conditional distribution** of $Y$ given $X = a_i$.

The **conditional expectation** of $Y$ given $X$, denoted by $E(Y|X)$ is defined as follows

$$E(Y|X) = E(Y|X = a_i) \text{ if } X = a_i \quad (i = 1, 2, \ldots)$$

# Two or more random variables

Suppose now that $X$ is a continuous r.v., we have a conditional distribution function similar to the discrete case.

$$E(Y|X) = E(Y|X = x) \text{ if } X = x \quad (x \in A)$$

Similarly, we have the following relation

$$E(ag(Y) + bh(Y)|X) = aE(g(Y)|X) + bE(h(Y)|X)$$

For any random variables $X$ and $Y$, the following relation holds

$$E(E(Y|X)) = E(X)$$

and is called the **law of iterated expectations**.

# Two or more random variables

**Exercise** Let each of the random variables $X$ and $Y$ take the values 0,1,2. There is an integer $k$, to be determined, such that $P(X = i \text{ and } Y = j) = (i + j)/k$ for $i = 0, 1, 2$ and $j = 0, 1, 2$.

(a) summarize the data in a table where the entries are $kP(X = i \text{ and } Y = j)$ and find $k$.

(b) find the possible of the random variable $E(Y|X)$ and their associated probabilities

(c) use this information and the law of iterated expectations to find $E(Y)$

(d) verify that the law of iterated expectations holds in this case, by showing that the same value of $E(Y)$ is obtained by direct calculation.

# Two or more random variables

For any random variables $X$ and $Y$

$$E(X+Y) = E(X) + E(Y)$$

Following the law of iterated expectations

$$E(E(X+Y|X)) = E(X) + E(E(Y|X))$$

Also

$$E(aX+bY) = aE(X) + bE(Y)$$

# Independent random variables

Let $X$ and $Y$ be independent variables r.v.s and suppose for the moment that $X$ is discrete, taking values $a_1, a_2, \ldots$ with positive probability.

Then for any $i = 1, 2, \ldots$ and any real number $y$,

$$P(Y \leq y | X = a_i) = P(Y \leq y)$$

Therefore

$$E(Y|X) = E(Y)$$

We have also

$$E(XY) = E(X)E(Y)$$

if, **but not only if**, $X$ and $Y$ are independent.

# Covariance and correlation

The **covariance** of two random variables $X$ and $Y$ is

$$\text{cov}(X, Y) = E([X - E(X)][Y - E(Y)])$$

Notice that $\text{cov}(X, Y)$ reduces to $\text{var}(X)$ if $X = Y$.

Often, we write

$$\text{cov}(X, Y) = E(XY) - E(X)E(Y)$$

In particular, $\text{cov}(X, Y) = 0$, if but not only if $X$ and $Y$ are independent.

# Covariance and correlation

**Exercise** Find the covariance of the random variables $X$ and $Y$ with joint distribution below.

| X \Y | 0 | 5 | 10 |
|---|---|---|---|
| 0 | 0.3 | 0.2 | 0 |
| 10 | 0.4 | 0 | 1 |

Are $X$ and $Y$ independent ?

# Covariance and correlation

The **coefficient of correlation** between two random variables $X$ and $Y$ is

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma(X)\sigma(Y)}$$

If $\text{corr}(X, Y) = 0$, then $X$ and $Y$ are said to be **uncorrelated**.

Finally, let $X$ and $Y$ be r.v.s

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2\text{cov}(X, Y)$$

Therefore

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$$

if, but not only if, $X$ and $Y$ are independent.

# Random variables

Some important properties of the **expected value**

- $E(a + bX) = E(a) + bE(X) = a + bE(X)$
- $E(X + Y) = E(X) + E(Y)$
- $E(X - Y) = E(X) - E(Y)$

and the **variance**

- $\text{var}(a + BX) = \text{var}(a) + \text{var}(bX) = b^2\text{var}(X)$
- $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2\text{cov}(X, Y)$
- If $X$ and $Y$ are **independent** then
  - $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$
  - $\text{var}(X - Y) = \text{var}(X) + \text{var}(Y)$

# Random sample and limit theorems

# Random sample

The random variables $X_1, X_2, \ldots, X_n$ are said to be **identically distributed** if they have the same c.d.f.

If $X_1, X_2, \ldots X_n$ are identically distributed and independent, they are said to form a **random sample**.

For example, suppose we sample at random 200 people from the population all London taxi-drivers and measure their earnings in particular week. Let $X_i$ be the earnings of the $i$th driver sampled.

Then $X_1, X_2, \ldots, X_{200}$ form a random sample; the common c.d.f. is that of the earnings of one taxi-driver chosen randomly from the same population.

# Random sample

Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample. Because $X_i$ has the same c.d.f. for all $i$, $E(X_i)$ is the same for all $i$.

We call this common expected value the **population mean** and denote by $\mu$.

Similarly, each of the r.v.s $X_1, X_2, \ldots, X_n$ has the same variance, which we call the **population variance**, denoted by $\sigma^2$.

The **sample mean** $\bar{X}$ is defined by

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \ldots X_n)$$

Notice that the sample mean $\bar{X}$ is a random variable, whereas the population mean $\mu$ is not.
In our example tax-drivers, $\mu$ is the average earnings of **all** taxi-drivers in London in the week in question, whereas $\bar{X}$ is the average earnings of the 200 drivers we happened to sample.

# Random sample

The expected value of the random variable $\bar{X}$ is

$$E(n\bar{X}) = E(X_1) + E(X_2) + \cdots + E(X_n) = n\mu$$

Since $X_1, \ldots, X_n$ are independent, the variance of $\bar{X}$ is

$$\text{var}(n\bar{X}) = \text{var}(X_1) + \cdots + \text{var}(X_n) = n\sigma^2$$

Therefore, since $E(n\bar{X}) = nE(\bar{X})$ and $\text{var}(n\bar{X}) = n^2\text{var}(\bar{X})$, it follows that

$$E(\bar{X}) = \mu; \quad \text{var}(\bar{X}) = \sigma^2/n$$

# Large random samples

Let $\bar{X}_n$ be the sample mean of a random sample from a population mean $\mu$. We are interested in how $\bar{X}_n$ when $n$ is very large.

The **law of large numbers** states that for any positive number $\delta$, however **small**, the probability that $\bar{X}_n$ differs from $\mu$ by less that $\delta$ approaches 1 as $n \to \infty$.

In other words,

$$\lim_{n\to\infty} P(\mu - \delta < \bar{X}_n < \mu + \delta) = 1 \quad \text{for all } \delta > 0$$

This law is sometimes called the **weak law of large numbers**.

# Convergence in probability

We say that a sequence of random variables $\{X_n\}$ is getting closer to another random variable $X$ as $n \to \infty$.

**Convergence in probability** Let $\{X_n\}$ be a sequence of random variables and let $X$ be a random variable defined on a sample space. We say that $X_n$ **converges in probability** to $X$ if, for all $\varepsilon > 0$

$$\lim_{n \to \infty} P[|X_n - X| \geq \varepsilon] = 0$$

This is often written $X_n \xrightarrow{P} X$.

# Convergence in probability

Suppose $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$. Then $X_n + Y_n \xrightarrow{P} X + Y$.

Suppose $X_n \xrightarrow{P} X$ and $a$ a constant. Then $aX_n \xrightarrow{P} aX$.

Suppose $X_n \xrightarrow{P} a$ and the real $g$ is continuous at $a$. Then $g(X_n) \xrightarrow{P} g(a)$.

Suppose $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$. Then $X_n Y_n \xrightarrow{P} XY$.

# Convergence in distribution

**Convergence in distribution** Let $\{X_n\}$ be a sequence of random variables and let $X$ be a random variable. Let $F_{X_n}$ and $F_X$ be, respectively, the cdfs of $X_n$ and $X$. We say that $X_n$ converges in distribution to $X$ if

$$\lim_{t \to \infty} F_{X_n} = F_X(x)$$

denoted by $X_n \xrightarrow{D} X$.

# Convergence in distribution

**Slutsky's Theorem** Let $X_n$, $X$, $A_n$ and $B_n$ be random variables and let $a$ and $b$ be constants. If $X_n \xrightarrow{D} X$, $A_n \xrightarrow{P} a$ and $B_n \xrightarrow{P} b$ then

$$A_n + B_n X_n \xrightarrow{D} a + bX$$

**Continuous Mapping Theorem** Let $g(.)$ be a continuous function then

$X_n \xrightarrow{D} X$ implies that $g(X_n) \xrightarrow{D} g(X)$

$X_n \xrightarrow{P} X$ implies that $g(X_n) \xrightarrow{P} g(X)$

If $(X_n - Y_n) \xrightarrow{P} 0$ and $Y_n \xrightarrow{D} Y$, then $g(X_n) - g(Y_n) \xrightarrow{P} 0$

# The central limit theorem

Let $\bar{X}_n$ be the sample mean of a random sample $\{X_1, \ldots, X_n\}$

For each $n$, let

$$Z_n = (\bar{X}_n - E(\bar{X}_n)) \Big/ \sqrt{\text{var}(\bar{X}_n)}$$

Then $Z_n$ has expected value 0 and variance 1, for all $n$.

The central limit theorem states that, as $n \to \infty$, the c.d.f. of $Z_n$ approaches that of a standard normal variate.

Thus, for a large $n$, the sample mean $\bar{X}_n$ is **approximatively normally distributed**.

You can prove this theorem using moment generating functions (beyond program).

# Applying the central limit theorem

Let $\{X_1, \ldots, X_n\}$ and $Z_n$ be as before, and let

$$W_n = n\bar{X}_n = X_1 + \cdots + X_n$$

Then $Z_n = (W_n - E(W_n))/\sqrt{\text{var}(W_n)}$. The theorem therefore tells us that $W_n$ is approximately normally distributed if $n$ is large.

Thus, the c.d.f. of the sum of a large number of independent, identically distributed random variables is approximately that of a normally distributed r.v. with the same mean and variance.

The approximation works well if $n \geq 50$.

# Applying the central limit theorem

**Exercise** Let $W$ be the sum of 50 independent random variables, each uniformly distributed over the interval [0,2]. Find $P(48 < W < 52)$.

# Clarifications

Suppose that $X \sim N(0,1)$, then $P(X \leq a)$ can be found in the table on Moodle.

Example : $P(X \leq 1.24) = 0.8935$.

$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a)$.

Example : $P(0.91 \leq X \leq 2.10) = P(X \leq 2.10) - P(X \leq 0.91) = 0.9821 - 0.8186 = 0.1635$.

# Clarifications

Suppose that $X \sim N(\mu, \sigma^2)$, and we want $P(a \leq X \leq b)$.

We do the following transformation :

$$P(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}) = P(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma})$$

where $Z \sim N(0,1)$ and you can find the values in the table on Moodle.

$\mu$ $\sigma^2$

Example : if $X \sim N(7,16)$ find $P(3 \leq X \leq 10)$.

$$P(\frac{3-7}{4} \leq Z \leq \frac{10-7}{4}) = P(Z \leq 0.75) - P(Z \leq -1)$$

Now we know that $P(Z \leq -1) = 1 - P(Z \leq 1)$
Then $P(-1 \leq Z \leq 0.75) = 0.7734 - (1 - 0.8413) = 0.6147$

# Parameter and Interval Estimation

# Parameter estimation

We discussed random variables, developed the notion of a probability distribution, established some fundamental results such as the CLT that give strong and useful information about the statistical properties of a sample drawn from a fixed, known distribution.

We now invert the analysis.

We start instead with data obtained by sampling a distribution or probability model with certain unknown parameters, we would like to extract information about the most reasonable values for these parameters given the observed data.

# Parameter estimation

As a starting point, we can write a general formula for the estimate in terms of the data values we observe.

Such a function is an **estimator** for our parameter of interest.

In general, there are many possible estimators for any given parameter, and it is not always clear which one we should use.

Intuitively, we might prefer to search for an estimator that tends to have the **smallest systematic error**.

# Parameter estimation

The most basic possible requirement is to ask that the estimator not have any "bias", on average, away from the expected value of the parameter.

We say that an estimator $\hat{\theta}(x_1; x_2; \ldots; x_n)$ for a set of observations $(x_1; x_2; \ldots; x_n)$ drawn by randomly sampling a random variable $X$ with probability density function $f_X(x; \theta)$ is **unbiased** if $E(\hat{\theta}) = \theta$.

$E(\hat{\theta}) = \theta + a$

bias

# Parameter estimation

**Exercise** Show that the estimate for the mean $\hat{\mu} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$ for sampling the normal distribution with mean $\mu$ and standard deviation $\theta$ is unbiased.

**Exercise** $\hat{\sigma}^2 = \frac{1}{n}(x_1^2 + x_2^2 + \cdots + x_n^2) - \left[\frac{x_1 + x_2 + \cdots + x_n}{n}\right]^2$ from sampling the normal distribution with unknown mean $\mu$ and standard deviation $\sigma$ is biased. (Recall that $\sigma^2 = E(x_i^2) - E(x_i)^2$.

# Efficiency of Estimators

As discussed, there are many different possible choices for estimators. How to choose the best one?

If $\hat{\theta}_1$ and $\hat{\theta}_2$ are two unbiased estimators for the parameter $\theta$, we we say that $\hat{\theta}_1$ is **more efficient** than $\hat{\theta}_2$ if $\text{var}(\hat{\theta}_1) < \text{var}(\hat{\theta}_2)$.

**Exercise** Given a random sample $x_1$ ; $x_2$ from the normal distribution with mean $\theta$ and standard deviation $\sigma$, we have two possible estimators for $\theta$ : $\hat{\theta}_1 = \frac{1}{2}(x_1 + x_2)$ and $\hat{\theta}_2 = \frac{1}{3}(x_1 + 2x_2)$. Show that these two estimators are unbiased. Which one if the most efficient?

$E(\hat{\theta}_1) = \theta$



$E(\hat{\theta}_1) = E\left(\frac{1}{2}(X_1 + X_2)\right)$

$= \frac{1}{2}E(X_1 + X_2)$

$= \frac{1}{2}(E(X_1) + E(X_2))$

$= \frac{1}{2} 2\theta$

$= \theta$

$E(\hat{\theta}_2) = E\left(\frac{1}{3}(X_1 + 2X_2)\right)$

$= \frac{1}{3}[E X_1 + 2 E X_2]$

$= \frac{1}{3} \cdot 3\theta = \theta$

unbiased

$\text{var}(\hat{\theta}_1) = \text{var}\left(\frac{1}{2}(X_1 + X_2)\right)$

$= \frac{1}{4} \text{var}(X_1 + X_2)$ $\text{cov}(X_1, X_2) = 0$

$= \frac{1}{4}[\text{var}X_1 + \text{var}X_2]$

$= \frac{1}{4} 2\sigma^2 = \frac{1}{2}\sigma^2$

$\text{var}(\hat{\theta}_2) = \text{var}\left(\frac{1}{3}(X_1 + 2X_2)\right)$

$= \frac{1}{9}(\text{var}X_1 + 4\,\text{var}X_2)$

$= \frac{5}{9}\sigma^2$

$\text{var}(\hat{\theta}_1) < \text{var}(\hat{\theta}_2)$

more efficient

# Interval Estimation

When estimating an unknown parameter, it is desirable to have a prediction that is as accurate as possible.

However, it is also important to be able to describe how accurate the prediction is, which is to say, how much the prediction differs from the true parameter value.

What we are seeking is to expand our discussion from **pointwise** parameter estimates, where we estimate the actual value of the parameter, to **interval** estimates, where we give an interval that we believe the parameter should lie in.

One approach to quantifying the uncertainty in our estimates is to construct a **confidence interval** : this is an interval around our estimated value in which we believe the true value should lie.

# Confidence interval

If $X$ is a random variable and $0 < \alpha < 1$ , a $100(1-\alpha)\%$ **confidence interval** for $X$ is an interval $(a; b)$ with $a < X < b$ such that $P(a < X < b) = 1 - \alpha$.

For example, a 95% confidence interval for $X$ is an interval $(a; b)$ where $X$ should land 95% of the time.

When we are constructing confidence intervals using parameter estimates, we typically will want to work with unbiased estimators that are as efficient as possible.

# Normal confidence interval

A $100(1-\alpha)\%$ confidence interval for the unknown mean $\mu$ of a normal distribution with known standard deviation $\sigma$ is given by

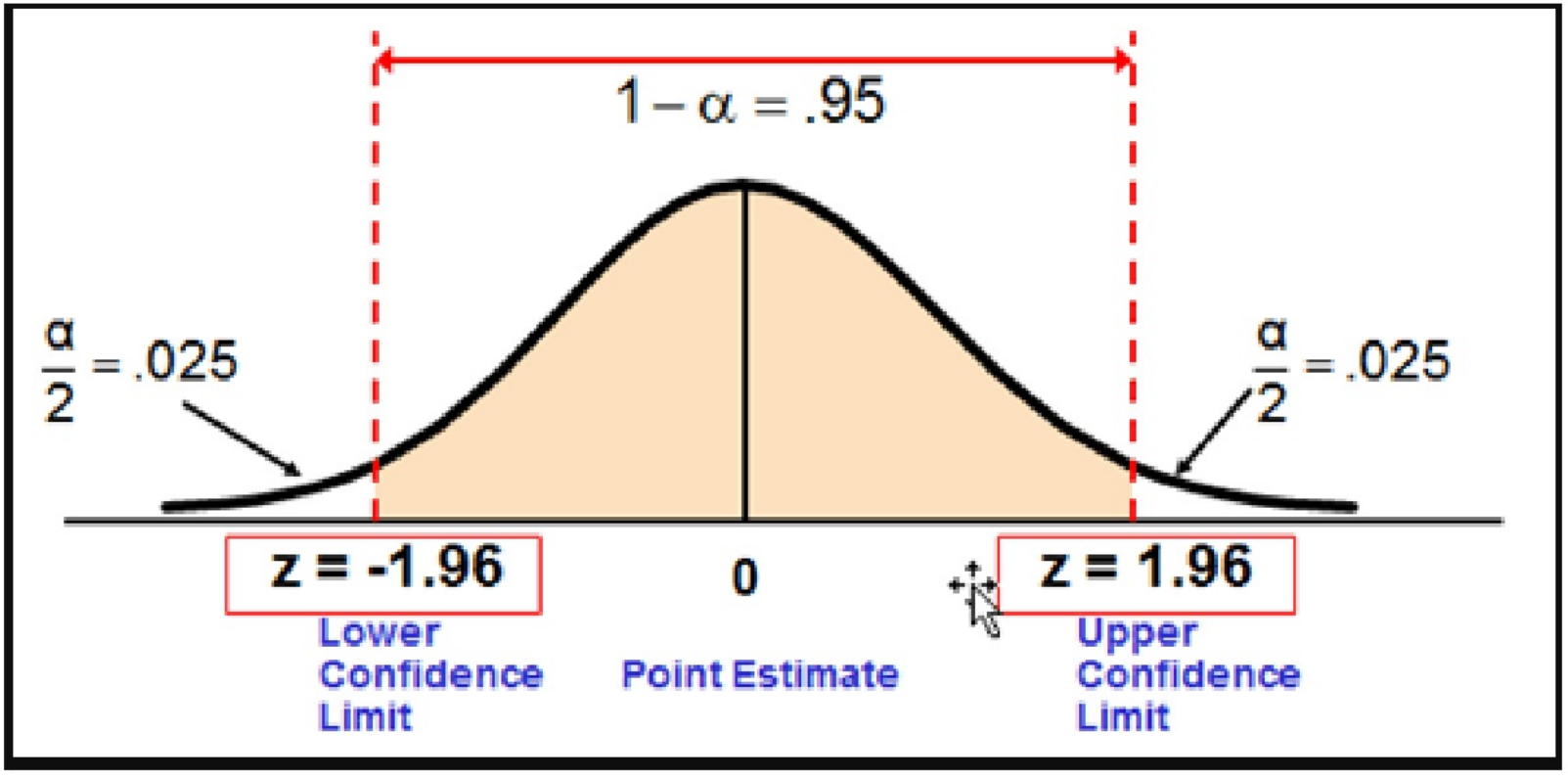
$$\hat{\mu} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = \left[\hat{\mu} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \hat{\mu} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

where $n$ sample points $x_1, \ldots, x_n$ are taken from the distribution, $\hat{\mu}$ is the sample mean and $z_{\alpha/2}$ is inverse cumulative distribution function for $N(0,1)$ satisfying $P(-z_{\alpha/2} < \frac{\hat{\mu}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}) = 1-\alpha$.

where $\frac{\hat{\mu}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$.

Some specific values $z_{\alpha/2} = 1.96$ for $\alpha = 95\%$ and $1.6449$ for $\alpha = 10\%$.

# Normal confidence interval




**Exercise** A normal distribution with unknown mean $\mu$ and standard deviation $\sigma = 1$ is sampled four times, yielding the values 1.4, 0.2, 2.9, and 1.1. Find 90%, 95% confidence intervals for $\mu$.

$X_1 = 1.4$

$x_2 = .2$

$x_3 = 2.9$

$x_4 = 1.1$

$\hat{\mu} = \frac{1}{n} \Sigma x_i$

$\hat{\mu} = \frac{5.6}{4}$

$\hat{\mu} = 1.4$

$\sigma = 1$

$n = 4$

$$\left[ -z_{\alpha/2} < \frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2} \right]$$

90%:

$$-1.64 < \frac{1.4 - \mu}{1/2} < 1.64$$

$$-.8245 < 1.4 - \mu < .8245$$

$$.57 < \mu < 2.22$$

95%:

$$-1.96 < \frac{1.4 - \mu}{1/2} < 1.96$$

$$0.42 < \mu < 2.38$$

# Hypothesis testing

# Principle of Hypothesis Testing

In the last section, we discussed methods for estimating parameters, and for constructing confidence intervals that quantify the precision of the estimate.

In many cases, parameter estimations can provide the basic framework to decide the plausibility of a particular hypothesis.

For example, to decide how plausible it is that a given coin truly is fair, we can flip the coin several times, construct an estimate for the true probability of obtaining heads and associated confidence intervals, and then decide based on the position of the confidence interval whether it is reasonable to believe the coin is fair.

However, in most of these situations, we are seeking a binary decision about a hypothesis : namely, whether or not it is justified by the available evidence.

# Null and Alternative Hypotheses

If we are making a binary decision, our first step is to explicitly identify the two possible results.

Example : "The coin is fair" versus "The coin is not fair".

We must then test a hypothesis using a statistical model. In order to do this, we must formulate the hypothesis in a way that allows us to analyze the underlying statistical distribution.

# Null and Alternative Hypotheses

The type of hypothesis we are testing in each case is a **null hypothesis**.

The other hypothesis is the **alternative hypothesis**.

Our choices are either to **reject the null hypothesis** in favor of the alternative hypothesis or to **fail to reject the null hypothesis**.

With the hypothesis tests we will study, the null hypothesis $H_0$ will be of the form "The parameter equals a specific value".

Example : "The probability of obtaining heads when flipping a coin is 1/2".

# Null and Alternative Hypotheses

The alternative hypothesis $H_a$ may then take one of several possible forms.

Two-sided : "The parameter is not equal to the given value"

One-sided : The parameter is less than the given value" or "The parameter is greater than the given value".

The two-sided alternative hypothesis is so named because it includes both possibilities listed for the one-sided hypotheses.

"The probability of obtaining heads when flipping a coin is not 1/2" is two-sided.

# Null and Alternative Hypotheses

$H_0: p = \frac{1}{2}$

$H_1: p \neq \frac{1}{2}$ two sided

**Exercise** We wish to test whether a particular coin is fair, which we do by flipping the coin 100 times and recording the proportion p of heads obtained. Give the null and alternative hypotheses for this test.

**Exercise** We wish to test whether the exam given to class A was easier than the exam given to class B, which we do by comparing the average scores $\mu_A$ and $\mu_B$ in the two classes. Give the null and alternative hypotheses for this test.

one sided

$H_0: \mu_A = \mu_B$

$H_1: \mu_A > \mu_B$

# Test Statistics, $p$-Values, and Decision Rules

Once we have properly formulated the null and alternative hypotheses, we can set up a hypothesis test to decide on the reasonableness of rejecting the null hypothesis.

Ideally, we would like to assess how likely it is to obtain the data we observed if the null hypothesis were true.

We will compute a test statistic based on the data and then assess the likelihood of obtaining this test statistic by sampling the distribution in the situation where the null hypothesis is true.

# Test Statistics, $p$-Values, and Decision Rules

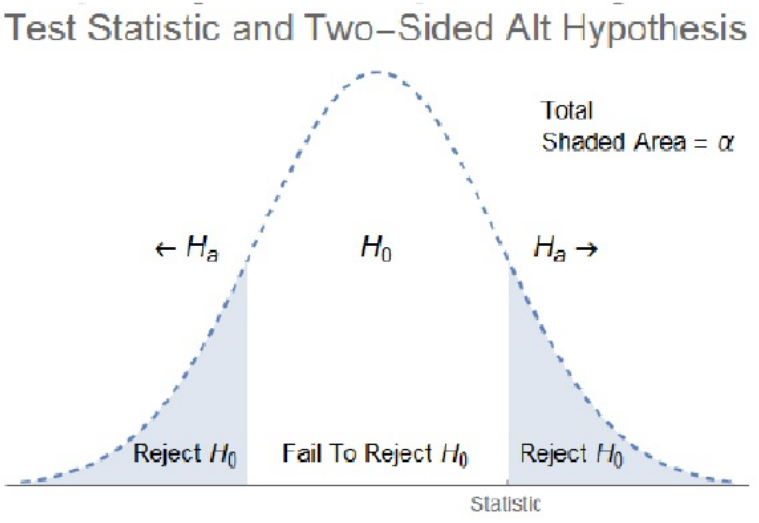
To decide whether to reject the null hypothesis, we adopt a **decision rule** of the following nature : we select a significance level $\alpha$ (often $\alpha = 0.1$, 0.05, or 0.01) and decide whether the p-value of the sample statistic satisfies $p < \alpha$ or $p \geq \alpha$.

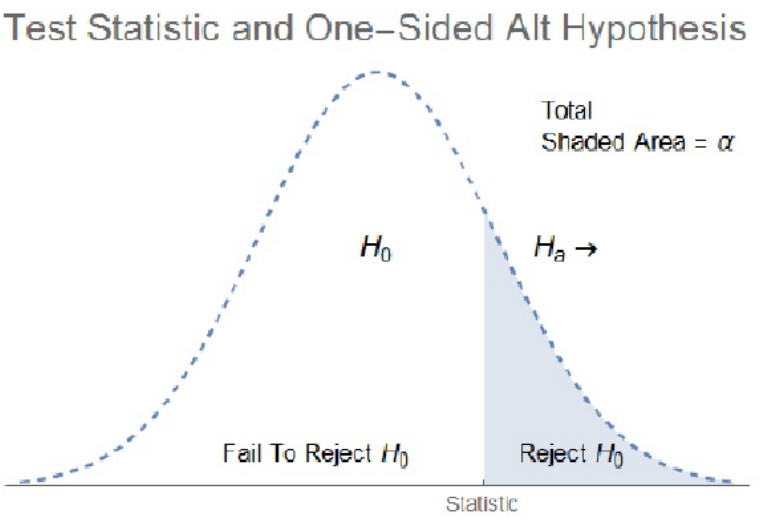
What is the **p-value** ? It's the probability of rejecting $H_0$ when true.

If $p < \alpha$, then we view the data as sufficiently unlikely to have occurred by chance : we reject the null hypothesis in favor of the alternative hypothesis and say that the evidence against the null hypothesis is **statistically significant**.

If $p \geq \alpha$, then we view as plausible that the data could have occurred by chance : we fail to reject the null hypothesis and say that the evidence against the null hypothesis is **not statistically significant**.

# Test Statistics, $p$-Values, and Decision Rules




For a two-sided alternative hypothesis, there are two regions in which we would reject the null hypothesis ("rejection regions") : one where the test statistic is too high and the other where it is too low. Together, the total area of these regions is $\alpha$.

For a one-sided alternative hypothesis, there is a single region in which we would reject the null hypothesis, corresponding to a test statistic that is sufficiently far in the direction of the alternative hypothesis. The total area of this region is $\alpha$

# $z$ Tests

We discuss the situation of testing whether a normally-distributed random variable has a particular mean.

These tests are known as **one-sample** $z$ tests after the letter $z$ traditionally used for normally-distributed quantities.

First, we must identify the appropriate null and alternative hypotheses and select a significance level $\alpha$.

We will use the test statistic $\hat{\mu}$, the sample mean, since this is the minimum-variance unbiased estimator for the population mean.

Under the assumption that $H_0$ is true, the test statistic is normally distributed with mean $\mu$ and standard deviation $\sigma$.

# $z$ Tests

If the hypotheses are $H_0 : \mu = c$ and $H_a : \mu > c$, then the $p$-value is $P(N_{\mu,\sigma} \geq z)$.

If the hypotheses are $H_0 : \mu = c$ and $H_a : \mu < c$, then the $p$-value is $P(N_{\mu,\sigma} \leq z)$.

If the hypotheses are $H_0 : \mu = c$ and $H_a : \mu \neq c$, then the $p$-value is

$$P(|N_{\mu,\sigma} - \mu| \geq |z - \mu|) = \begin{cases} 2P(N_{\mu,\sigma} \geq z) & \text{if } z \geq \mu \\ 2P(N_{\mu,\sigma} \leq z) & \text{if } z < \mu \end{cases}$$

In each case, we are simply calculating the probability that the normally-distributed random variable $N_{\mu,\sigma}$ will take a value further from the hypothesized mean $\mu$ than the observed test statistic $z$.

# $z$ Tests

$H_0: \mu = 180$

$H_a: \mu > 180$

**Exercise** The production of an assembly line is normally distributed, with mean 180 widgets and standard deviation 10 widgets for a 9-hour shift. The company wishes to test to see whether a new manufacturing technique is more productive. The new method is used for a 9-hour shift and produces a total of 197 widgets. Assuming that the standard deviation for the new method is also 10 widgets for a 9-hour shift, state the null and alternative hypotheses, identify the test statistic and its distribution, calculate the p-value, and test the claim at the 10%, 5%, and 1% levels of significance.

$$P(N_{180,10} \geq 197) = P\left(N_{0,1} \geq \frac{197 - 180}{10}\right) = P(N_{0,1} \geq 1.7)$$

$$= 4.457\%$$



$\alpha = 10\% \longrightarrow$ $p < \alpha \longrightarrow$ reject $H_0$

$\alpha = 5\% \longrightarrow$ $p < \alpha \longrightarrow$ reject $H_0$

$\alpha = 1\% \longrightarrow$ $p > \alpha \longrightarrow$ fail to reject

# $z$ Tests

**Exercise** An airline wants to measure how accurate its cross-country travel time predictions are. They believe that their predictions are accurate on average, with a standard deviation of 20 minutes. They collect data from 6 routes, whose errors in travel-time predictions are -39 minutes, +14 minutes, -21 minutes, -23 minutes, +25 minutes, and -31 minutes (positive values are flights arriving early and negative values are flights arriving late).

Test at the 10% significance level the hypothesis that the true mean error $\mu$ is 0 minutes, if (i) the airline is concerned about errors in any direction (take the average error as your test statistic), and (ii) the airline is only concerned about errors that make flights late.

$\sigma = 20$

$n = 6$

$\hat{\mu} = -12.5$

test: and

$H_0: \mu = 0$ $\quad \mu = 0$

$H_1: \mu \neq 0$ $\quad \mu < 0$

$$N_{obs} = \frac{-12.5}{20/\sqrt{6}} = -1.53$$

$P(N \leq -1.53) = 0.0643$

$p < 2$; reject $H_0$

$2P(N \leq -1.53) = 0.1286$

$p > 2$; fail to reject $H_0$

# Type I and Type II Errors

If we are testing a null hypothesis $H_0$, we commit a type I error if we reject $H_0$ when $H_0$ was actually true. We commit a type II error if we fail to reject $H_0$ when $H_0$ was actually false.

| $H_0$ / Result | Fail to Reject $H_0$ | Reject $H_0$ |
|---|---|---|
| $H_0$ is true | Correct decision | Type I Error |
| $H_0$ is false | Type II Error | Corrrect Decision |

# The $t$ Distribution and $t$ Tests

In our discussion of hypothesis testing in the previous section, we relied on $z$ tests, which require an approximately normally distributed test statistic whose standard deviation is known.

However, in most situations, it is unlikely that we would actually know the population standard deviation.

Instead, we must estimate the population standard deviation from the sample standard deviation.

We use the sample standard deviation

$$S = \sqrt{\frac{1}{n}[(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2]}$$

whose square $S^2$ is an unbiased estimator of the population variance $\sigma^2$.

# The $t$ Distribution and $t$ Tests

It may seem reasonable to say that if we use the estimated standard deviation S in place of the unknown population $\sigma$, then we should be able to use a $z$ test with the resulting approximation.

However, if we take $\frac{\bar{x}-\mu}{S/\sqrt{N}}$ as our test statistics, it turns out that this test is not normally distributed.

In fact, the distribution is a $t$ distribution.

For large $n$, the distribution looks more approximatively normal (CLT) and we are back to the $z$ tests.