UNIVERSITÀ DEGLI STUDI DI MILANO
Dipartimento di Economia, Management
e Metodi Quantitativi

DEMM

Dept. of Economics, Management and Quantitative Methods

University of Milan

Academic Year 2019-2020

Research Methods - Statistics

Fabrizio Iacone

Lecture notes, readings and module objectives

The lecture notes that have been prepared contain many definitions and basic results, so that time during the lectures can be devoted to listening and understanding, rather than copying formulae. These notes are, however, not intended to give detailed discussions with examples. Doing examples and reading more detailed discussions will be needed to gain a sound understanding of the topics to be covered. It is therefore important to use a text, as well as the lecture notes.

Acknowledgements: These notes are in part based on notes of Les Godfrey, whom I thank. All errors are mine.

Most introductory statistics book provide a useful and detailed reference to review the material of these lectures. For example,

- Miller, I. and M. Miller, 2004. John E. Freund's mathematical statistics with applications, 7th edition, Pearson Prentice-Hall.

Many introductory econometrics texts also contain useful reviews of the relevant statistical concepts. For example,

- Wooldridge, J., 2003. *Introductory econometrics*, 2nd ed., South Western College Publishing.

This course is designed to provide an introduction to topics that will be very useful in the Econometrics core module. As such, we will only give an introductory presentation of probability and probability techniques. We will not include, for example, foundations of probability (measure) theory, or applications of probability to economics or other applications such as Bayes' theory. We will only present a few of the well known distributions that are routinely used in statistics. In inferential statistics, we will focus on estimation and testing hypothesis on finite dimensional parameters in an independent random sample.

Structure

The course is in two parts: in the first one we describe the foundations of probability and distributional theory, and in the second one we apply these to statistical inference. We divided the course in ten sections:

1. Preliminary mathematics and introduction to the probability model;
2. Random variables and distributions;
3. Mathematical expectations;
4. Some useful distributions;
5. Asymptotic / large sample distribution theory;
6. Sampling distributions;
7. Point estimation;
8. Maximum likelihood estimation;
9. Interval estimation;
10. Hypothesis testing.

Section 1
Preliminary mathematics and the probability model

- Objectives: to introduce ideas concerning probability
- References: Miller and Miller, Chapter 2; Wooldridge, Appendix A (section A.1)

Preliminary mathematics - Some mathematical operators

Suppose we want to measure some characteristics, say weight, in kg, of 50 people. It is then useful to have some shorthand notation.

We will use $x_1, ..., x_{50}$, where x_i is the characteristic for the individual $i, i = 1, ..., 50$. For a generic set of n values, we write $x_1, ..., x_n$ or $\{x_1, \ldots, x_n\}$.

If two characteristics are measured simultaneously, for example weight and height, we can use $(x_1, y_1), ..., (x_n, y_n)$, with x_i and y_i being the characteristics (weight and height) for the individual $i, i = 1, ..., n$.

If one characteristic is measured for m classes of n individuals, for example gender and weight, we use $x_{11}, ..., x_{1n}, x_{21}, ..., x_{2n}, ..., x_{m1}, ..., x_{mn}$, so for example $x_{11}, ..., x_{1n}$ is the weight of the n female individuals, and $x_{21}, ..., x_{2n}$ is the weight of the n male individuals.

We are often interested in summarising the characteristics of the whole sample in one single number, or in other transformations of numbers. This may require the application of new mathematical operators (such as, the summation or the product operator).

Summation operator

For a sequence of values $X_1, X_2, ..., X_n$, the **summation operator** $\sum$ is defined by

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + \ldots + X_n$$

- n may be finite or infinite (∞);
- sometimes we do not show the limits of the summations, and use either $\sum_i X_i$ or $\sum X_i$;

- Using the summation operator, we can also define the **sample mean** (also known as **sample average**), denoted $\overline{X}$:

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

★ An example

$$X_1 = 3, X_2 = -2, X_3 = 1$$

Then,

$$\sum_{i=1}^{3} X_i = (3 + (-2) + 1) = 2$$

$$\bar{X} = \frac{1}{3}(3 + (-2) + 1) = \frac{2}{3}$$

Properties of the summation:

- if a and b are constant,

$$\sum_{i=1}^{n} a = na, \ \sum_{i=1}^{n} bX_i = b\sum_{i=1}^{n} X_i;$$

Verify this directly:

$$\sum_{i=1}^{n} a = \underbrace{(a + a + \ldots + a)}_{n \text{ times}} = na; \quad \sum_{i=1}^{n} bX_i = (bX_1 + \ldots + bX_n) = b\sum_{i=1}^{n} X_i$$

★ An example. Let $a = 2$, then

$$\sum_{i=1}^{3} a = (2 + 2 + 2) = 6.$$

★ An example. Let $b = 2$, and $X_1 = 3, X_2 = -2, X_3 = 1$, then

$$\sum_{i=1}^{3} bX_i = (2 \times 3 + 2 \times (-2) + 2 \times 1) = (6 - 4 + 2) = 4 = b\sum_{i=1}^{3} X_i$$

Properties of the summation (continued):

- for any X_i, Y_i, $(1 \leq i \leq n)$

$$\sum_{i=1}^{n}(X_i + Y_i) = \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i;$$

★ An example.

$$X_1 = 3, X_2 = -2, X_3 = 1$$

$$Y_1 = 3, Y_2 = 1, Y_3 = -2$$

Then $\sum_{i=1}^{3} X_i = (3 + (-2) + 1) = 2$ and $\sum_{i=1}^{3} Y_i = 3 + 1 - 2 = 2$ and

$$\sum_{i=1}^{3}(X_i + Y_i) = (3 + 3) + (-2 + 1) + (1 - 2) = 4 = 2 + 2 = \sum_{i=1}^{3} X_i + \sum_{i=1}^{3} Y_i$$

★ An application:

$$\sum_{i=1}^{n}(X_i - \overline{X}) = 0$$

Proof:

$$\sum_{i=1}^{n}(X_i - \overline{X}) = \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \overline{X} = \sum_{i=1}^{n} X_i - n\overline{X} = n\left(\frac{1}{n} \sum_{i=1}^{n} X_i - \overline{X} \right)$$

and $\frac{1}{n} \sum_{i=1}^{n} X_i - \overline{X} = 0$.

Further applications of the summation operator:

- **double summation**: for X_{ij} $(1 \leq i \leq n, 1 \leq j \leq m)$

$$\sum_{i=1}^{n}\sum_{j=1}^{m} X_{ij} = \sum_{i=1}^{n}(X_{i1} + X_{i2} + \ldots + X_{im}) = \sum_{j=1}^{m}(X_{1j} + X_{2j} + \ldots + X_{nj})$$

★ An example. For X_{ij} such that, for $1 \leq i \leq 2, 1 \leq j \leq 3$,

$$X_{11} = 1, X_{12} = -3, X_{13} = 3, X_{21} = -4, X_{22} = 2, X_{23} = -1.$$

Then

$$\sum_{i=1}^{2}\sum_{j=1}^{3} X_{ij} = (1 - 3 + 3 - 4 + 2 - 1) = -2$$

We can organise the data in a table

1	–3	3
–4	2	–1

It is easy to see that

fix $i = 1$ (first line of the table):

$$X_{11} = 1, X_{12} = -3, X_{13} = 3, \sum_{j=1}^{3} X_{ij} = (X_{i1} + X_{i2} + X_{i3}) = 1$$

fix $i = 2$ (second line of the table):

$$X_{21} = -4, X_{22} = 2, X_{23} = -1, \sum_{j=1}^{3} X_{ij} = (X_{i1} + X_{i2} + X_{i3}) = -3$$

Sum across all lines,

$$\sum_{i=1}^{2}\left(\sum_{j=1}^{3} X_{ij}\right) = 1 - 3 = -2$$

Or, fix $j = 1$ (first column):

$$X_{11} = 1, X_{21} = -4, \sum_{i=1}^{2} X_{ij} = (X_{1j} + X_{2j}) = -3$$

in the same way, fixing $j = 2$ or $j = 3$, respectively,

$$j = 2: X_{12} = -3, X_{22} = 2, \sum_{i=1}^{2} X_{ij} = (X_{1j} + X_{2j}) = -1$$

$$j = 3: X_{13} = 3, X_{23} = -1, \sum_{i=1}^{2} X_{ij} = (X_{1j} + X_{2j}) = 2$$

Summing all columns,

$$\sum_{j=1}^{3} \left(\sum_{i=1}^{2} X_{ij} \right) = -3 - 1 + 2 = -2$$

This corresponds to summing the columns and the rows of the table, so that

				$\sum_{j=1}^{3} X_{ij}$
	1	−3	3	1
	−4	2	−1	−3
$\sum_{i=1}^{2} X_{ij}$	−3	−1	2	$\sum_{i=1}^{2} \sum_{j=1}^{3} X_{ij} = -2$

Further operations with the Summation operator

- We can also **combine** these operations.
  For example for

$$Z_i = a + bX_i + cY_i,$$

  for a, b, c constants,

$$\sum_{i=1}^{n} Z_i = na + b\sum_{i=1}^{n} X_i + c\sum_{i=1}^{n} Y_i$$

★ An example.

For $X_1 = 3, X_2 = -2, X_3 = 1$; $Y_1 = 3$, $Y_2 = 1$, $Y_3 = -2$; $a = 4$, $b = 2$, $c = -1$:

$$\sum_{i=1}^{3} Z_i = 3 \times 4 + 2 \times 2 - 1 \times 2 = 14$$

- We can also apply the operator $\sum_{i=1}^{n}$ to **non-linear** functions, $g(X_i)$:

$$\sum_{i=1}^{n} g(X_i) = g(X_1) + \ldots + g(X_n)$$

If, for example,

$$g(X_i) = X_i^2, \text{ then } \sum_{i=1}^{n} X_i^2 = X_1^2 + \ldots + X_n^2$$

$$g(X_i) = (X_i + a)^2, \text{ then } \sum_{i=1}^{n} (X_i + a)^2 = \sum_{i=1}^{n} X_i^2 + 2a \sum_{i=1}^{n} X_i + na^2.$$

★ An example. For $X_1 = 3, X_2 = -2, X_3 = 1, \sum_{i=1}^{n} X_i^2 = 3^2 + (-2)^2 + 1^2 = 14$

$$\sum_{i=1}^{3} (X_i - 2)^2 = (3-2)^2 + (-2-2)^2 + (1-2)^2 = 18$$

$$\text{or, } \sum_{i=1}^{3} (X_i - 2)^2 = 14 + 2 \times (-2) \times 2 + 3(-2)^2 = 18$$

Product operator

For a sequence of values $X_1, X_2, ..., X_n$, the **product operator** $\prod$ is defined as

$$\prod_{i=1}^{n} X_i = X_1 \times X_2 \times \ldots \times X_n$$

- n may be finite or infinite (∞);
- sometimes either $\prod_i X_i$ or $\prod X_i$ are used;

★ An example.

For $X_1 = 3, X_2 = -2, X_3 = 1$,

$$\prod_{i=1}^{3} X_i = 3 \times (-2) \times 1 = -6$$

Relation between product and summation

- if every value of X_i is positive, then

$$\ln\left(\prod_i X_i\right) = \sum_i \ln(X_i).$$

★ An example.

For

$$X_1 = 3, X_2 = 2, X_3 = 1,$$

then

$$\prod_i X_i = 6,\ \ln\left(\prod_i X_i\right) = 1.7918\ldots$$

and

$$\ln X_1 = 1.0986\ldots,\ \ln X_2 = 0.69315\ldots,\ \ln X_3 = 0$$

$$\sum_i \ln X_i = 1.0986\ldots + 0.69315\ldots + 0 = 1.7918$$

Factorial operator

For any positive integer k, the **factorial** operator is defined so that

$$k! = k \times (k-1) \times \ldots \times 1$$

- for $k = 0$, $0! = 1$ is defined.

★ An example.

$$2! = 2 \times 1 = 2,\ 3! = 3 \times 2 \times 1 = 6,\ \ 4! = 4 \times 3 \times 2 \times 1 = 24$$

Errors to avoid:

$$\sum_{i=1}^{n} X_i^2 \neq \left(\sum_{i=1}^{n} X_i \right)^2$$

$$\sum_{i=1}^{n} X_i Y_i \neq \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i$$

$$\sum_{i=1}^{n} \frac{X_i}{Y_i} \neq \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} Y_i}$$

★ Errors to avoid, examples:

$$\sum_{i=1}^{n} X_i^2 \neq \left(\sum_{i=1}^{n} X_i \right)^2$$

Verify this setting $n = 2, X_1 = 1, X_2 = 9,$
then $\sum_{i=1}^{n} X_i^2 = (1^2 + 9^2) = 82$; $\left(\sum_{i=1}^{n} X_i\right)^2 = (1 + 9)^2 = 100.$

$$\sum_{i=1}^{n} X_i Y_i \neq \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i$$

Verify this setting $n = 2$, $(X_1 = 1, Y_1 = 2)$, $(X_2 = 2, Y_2 = 5)$,
then $\sum_{i=1}^{n} X_i Y_i = (1 \times 2 + 2 \times 5) = 12$;
$\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i = (1 + 2) \times (2 + 5) = 21.$

$$\sum_{i=1}^{n} \frac{X_i}{Y_i} \neq \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} Y_i}$$

Verify this setting $n = 2$, $(X_1 = 1, Y_1 = 2)$, $(X_2 = 2, Y_2 = 5)$, then $\sum_{i=1}^{n} \frac{X_i}{Y_i} = (\frac{1}{2} + \frac{2}{5}) = \frac{9}{10}$; $\frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} Y_i} = \frac{(1+2)}{(2+5)} = \frac{3}{7}$.

The probability model

Set theory for probability.

Random experiment
an experiment that may result in two or more different outcomes with uncertainty as to which will be observed, e.g. throwing a die.

Sample space
the set of all the potential outcomes of the experiment, usually denoted by Ω. For throwing a regular die, $\Omega = \{1,2,3,4,5,6\}$.

Event
a subset of the sample space, e.g., in the previous example,
$A = \{\textit{score less than } 4\} = \{1,2,3\}$, $B = \{\textit{score even}\} = \{2,4,6\}$.

Intersection of sets
denoted by $A \cap B$, for $A \subseteq \Omega$, $B \subseteq \Omega$. Set of elements that belong to *both A and B*. In the example of throwing a die, $A \cap B = \{2\}$.

Difference of sets

for $A \subseteq \Omega$, $B \subseteq \Omega$, indicated as $A \backslash B$. The set of elements that are in A but not in B. In the example of throwing a die, $A \backslash B = \{1, 3\}$.

Union of sets
denoted by $A \cup B$, for $A \subseteq \Omega$, $B \subseteq \Omega$. Set of elements that belong to *A or to B or to both*, e.g. In the example of throwing a die, $A \cup B = \{1, 2, 3, 4, 6\}$.

Complement set
denoted by A^c, for $A \subseteq \Omega$. Set of elements in Ω but not in A ($A^c = \Omega \backslash A$). In the previous example, $A^c = \{4, 5, 6\}$.

Empty set
a set with no elements in. Indicated as $\varnothing$.

Disjoint sets
for $A \subseteq \Omega$, $B \subseteq \Omega$. When A and B have no elements in common (so $A \cap B = \varnothing$).
In probability theory, Two events that have no outcomes in common are said to be **mutually exclusive**.
In the example of throwing a die, letting $C = \{\textit{score at least } 5\} = \{5,6\}$, $A \cap C = \varnothing$.

We can also combine these set operations, for example $A^c \cap B$,...

Some interesting formulas (De Morgan)
$(A \cap B)^c = A^c \cup B^c$, $(A \cup B)^c = A^c \cap B^c$,
$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$, $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.

Example: List the elements in these sets when A, B, C and Ω are the ones of the example of throwing a die.

Defining the probability model

Intuitively, we want probability to replicate the percentage of occurrences of an uncertain event when an experiment takes place many times.

The axioms of probability. Let Ω be a sample space composed of $n < \infty$ outcomes, and A, B are two generic events such that $A \subseteq \Omega$, $B \subseteq \Omega$. Probability, denoted $P(.)$, is a function that associates to any A, B a number in $[0,1]$ so that

$$A.1: \; P(A) \geq 0$$

$$A.2: \; P(\Omega) = 1$$

$$A.3: \; P(AUB) = P(A) + P(B) \text{ if } A \cap B = \varnothing$$

Note. This definition requires $n < \infty$: the extension to $n = \infty$ is possible but in that case the collection of all the subsets of Ω may be too large for probabilities to be assigned reasonably to all its members. Probability is then only defined for some sets A that satisfy further properties. Axiom $A.3$ is also extended to allow an infinite union of disjoint sets.

In the rest of the notes we will assume that any event A we consider is such that $P(A)$ can be assigned, unless otherwise specified.

The addition rule:

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &\leq P(A) + P(B) \end{aligned}$$

Other properties of probability:

$$P(A^c) = 1 - P(A)$$

$$P(\varnothing) = 0$$

$$\text{if } A \subseteq B, P(A) \leq P(B)$$

$$P(A) \leq 1$$

Conditional probability

Consider the example of throwing a die, with the natural probability assignment ($P(i) = 1/6$ for $i = 1, \ldots, 6$). In this case, $P(A) = P(\{1,2,3\}) = 0.5$. However, suppose the die is thrown and, although you are not told the result, you are informed that the score is even. What is the probability of a score less than 4 in this case?

For any A, B, we are interested in $P(A|B)$, the probability of "A given B" or of "A conditional upon B":

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

Multiplicative rule

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

Independent events

Two events A and B are independent if and only if

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

which implies that

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

Two mutually exclusive events A, B such that $P(A) > 0, P(B) > 0$ cannot be independent.

Example. Consider an experiment composed of two parts: we toss a coin and we throw a die. Let $A = \{\textit{die score less than } 4\}$, and $B = \{\textit{coin is head}\}$. Clearly, $P(A) = 0.5$ but also $P(A|B) = 0.5$: knowing the result of the coin toss, is not informative about the die score.

Example: bookings at the indoor court

The indoor court of the Gym can be used for three activities: basketball (BB), badminton (BM) and volleyball (VB); these are practised by two groups of people: undergraduate students (UG) and postgraduate students (PG). The percentage of bookings last year, divided per student and activity, are

	Activity		
Student	BB	BM	VB
UG	0.30	0.19	0.21
PG	0.09	0.12	0.09

For any given booking of that year,

What is the probability that a court was booked by PG to play BM?

What is the probability that a court was booked to play BM?

What is the probability that a court was booked to play BM given that the booking was made by a PG?

Are the events "the court was booked by UG" and "the court was booked to play BM" mutually exclusive?

Are the events "the court was booked by PG" and "the court was booked to play BM" independent?

★ Discussion of the example

♦ What is the probability that a court was booked by PG to play BM?

$$P(PG \cap BM) = 0.12$$

♦ What is the probability that a court was booked to play BM?

$$P(BM) = P(UG \cap BM) + P(PG \cap BM) = 0.19 + 0.12 = 0.31$$

♦ What is the probability that a court was booked to play BM given that the booking was made by a PG?

$$P(BM|PG) = \frac{P(BM \cap PG)}{P(PG)}$$

$$\begin{aligned} P(PG) &= P(PG \cap BB) + P(PG \cap BM) + P(PG \cap VB) \\ &= 0.09 + 0.12 + 0.09 = 0.3 \end{aligned}$$

$$P(BM|PG) = \frac{0.12}{0.3} = 0.4$$

♦ Are the events "the court was booked by UG" and "the court was booked to play BM" mutually exclusive?

$$P(UG \cap BM) = 0.19, \text{ so } UG \cap BM \neq \varnothing$$

♦ Are the events "the court was booked by PG" and "the court was booked to play BM" independent?

$$P(BM) = 0.31$$

$$P(BM|PG) = 0.40$$

(see above) so $P(BM|PG) \neq P(BM)$ and therefore the events are not independent.

Bayes' rule

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

Proof:

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

This is also known as Bayes' theorem.

Example: disease detection

A blood test is available, to detect the presence of a disease, which is present in 0.5% of the population. The test gives a positive outcome in 99% of the people who have the disease (it correctly detects the disease with probability 99%), and in 5% of the people who do not have the disease (it incorrectly detects the disease with probability 5%).

What is the probability of having the disease, if the outcome of the test is positive?

Let A be the event $\{\textit{the individual has the disease}\}$, and B the event $\{\textit{the outcome of the test is positive}\}$, then

$$P(A) = 0.005,\ P(B|A) = 0.99,\ P(B|A^c) = 0.05,$$

and we are interested in $P(A|B)$. We can compute $P(A') = 0.995$, so, by Bayes' rule,

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

$$\frac{0.99 \times 0.005}{0.99 \times 0.005 + 0.05 \times 0.995} \approx 0.0905$$

We can generalise Bayes' rule to sets A_j such that

$$S = \cup_{j=1}^{n} A_j, \text{ where } A_j \cap A_i = \varnothing \text{ (for } j \neq i).$$

Then,

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

$$= \frac{P(A_k \cap B)}{\sum_{j=1}^{n} P(A_j \cap B)}$$

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{j=1}^{n} P(B|A_j)P(A_j)}$$

Section 2
Random variables and probability distributions

- Objectives: to introduce random variables and probability functions, to facilitate calculating probabilities, and to model situations that are too large for a small table
- Topics: Random variables; discrete and continuous distributions; joint, marginal and conditional distributions.
- References: Miller and Miller, Chapter 3 (Sections 3.1 to 3.7); Wooldridge, Appendix B (sections B.1 and B.2)

Random variables and probability distributions

In some experiments, the outcome is a number; otherwise, it may be transformed into a number. In many cases, this may be advantageous, because the probability tables may be summarised in a generic function.

Random variable, for a finite dimensional space Ω
The random variable is a function that associates the outcomes in Ω to numbers.

Note: This definition may be generalised to comprise cases in which the dimension of Ω is not finite, but still countable, and even cases in which the dimension is not countable. When the dimension of Ω is finite or infinite but countable, then the random variable is said to be **discrete**, otherwise it is said to be **continuous**.

Notice that the random variable is neither a variable (it is a function) nor random (the association is certain).

A random variable is indicated with the use of an uppercase letter.

The value the random variable takes when the experiment is run, is called realisation and it is indicated with the use of a lowercase letter.

Example of throwing a fair die: X is "score"; the outcomes $1, 2, 3, 4, 5$ or 6 are values of x with non-zero probability.

Probability distributions for discrete random variables

Let X be a discrete random variable with realisations $\{x_1, x_2, \dots\}$, and

$$f_X(x_j) = P(X = x_j)$$

for $x_j \in \{x_1, x_2, \dots\}$, where

$$f_X(x_j) \geq 0, \sum f_X(x_j) = 1.$$

Then $\{f_X(x_1), f_X(x_2), \dots\}$ is the probability distribution, and the function $P(X = x_j)$ is the probability function.

Note. The individual $f_X(x_j)$ are probability masses, so the function $P(X = x_j)$ is sometimes known as probability mass function.

Sometimes the subscript is omitted, and $f(x_j) = P(X = x_j)$ is also used. The notation $p_j = P(X = x_j)$ is also used.

★ Example. When X is the score from throwing a fair die, $P(X = x) = 1/6$ when $1, 2, 3, 4, 5$ or 6. For any other value of x, $P(X = x) = 0$.

★ Example: Number of Tails from tossing a fair coin twice.

Let Ω be the space of outcomes of two tosses. So, let HH be the outcome "the both the first and second toss resulted in Head", HT the outcome "the first toss resulted Head, the second in Tail"... The sample space Ω is

$$\Omega : \{HH, HT, TH, TT\}$$

(notice that HT and TH are different outcomes).

Then, let X be the random variable "number of tails". If we observe the outcome HH, then X takes value 0, i.e. $x = 0$. In the same way, if observe the outcomes HT or TH, $x = 1$, ... so

$$X(HH) = 0,$$

$$X(HT) = 1,$$

$$X(TH) = 1,$$

$$X(TT) = 2.$$

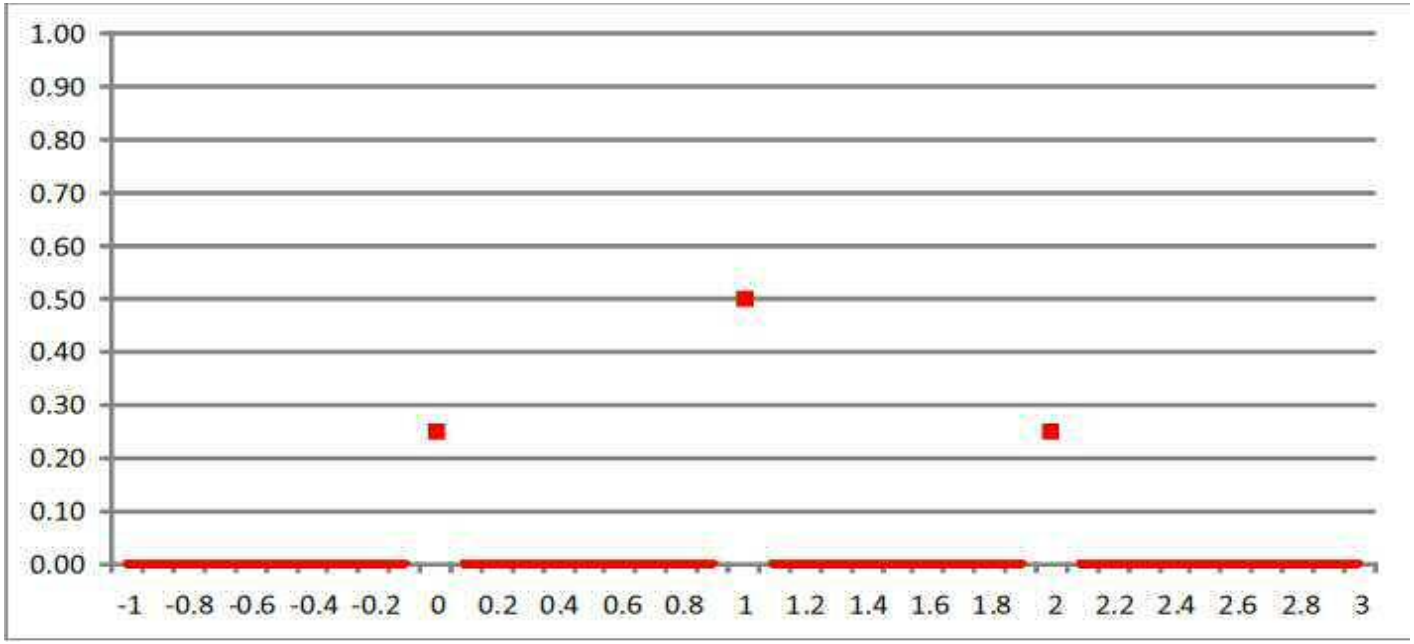
(notice that indeed X is a function of Ω with images in $\mathbb{R}$, i.e. $X : \Omega \to \mathbb{R}$, so it really is a random variable).

As each outcome has the same probability, 1/4, the probability function is

$$P(X = 0) = 1/4,\ P(X = 1) = 1/2,\ P(X = 2) = 1/4.$$

An alternative way to represent this probability function is

$$P(X = x) = \begin{cases} -1/4x^2 + 1/2x + 1/4 \text{ if } x \in \{0, 1, 2\} \\ 0 \text{ otherwise} \end{cases}$$



This function is not continuous: notice the gaps at 0, 1 and 2.

Note: this random variable is characterized by a special distribution called "binomial". For n independent trials, letting X be the random variable "number of tails" in n independent trials, the probability function is

$$P(X = x) = \frac{n!}{x!(n-x)!} 0.5^x 0.5^{n-x}.$$

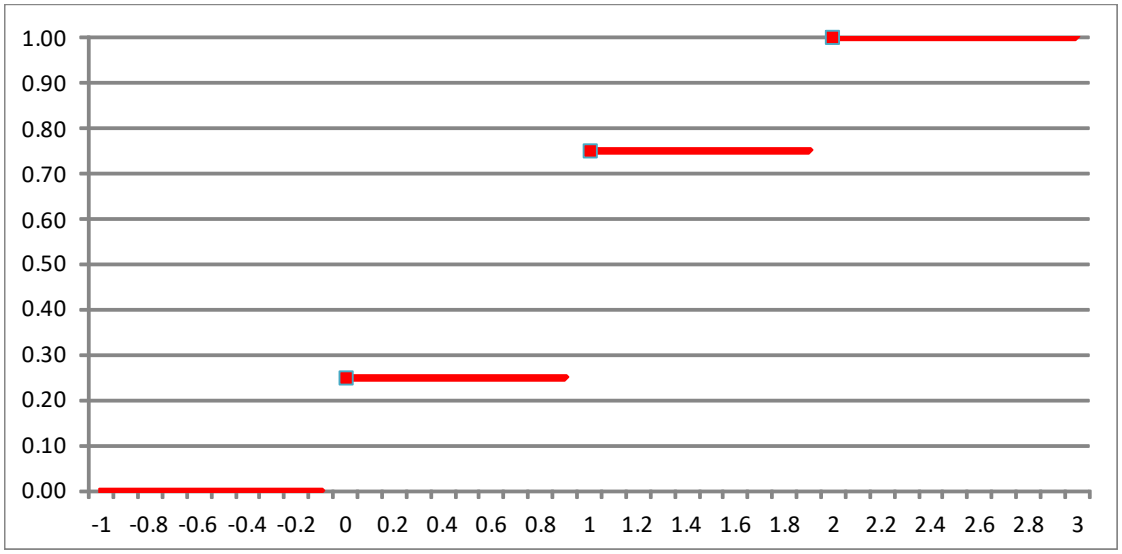
Cumulative distribution function

For any x, it is possible to define

$$F_X(x) = P(-\infty < X \leq x).$$

$F_X(.)$ is called cumulative distribution function.

★ In the example of recording the number of tails after two coin toss trials, the Cumulative distribution function is



Note: The "square" is used to indicate that when a discontinuity takes place $F_X(x)$ takes the value indicated by the square, e.g., $F_X(0) = 0.25$.

When $\boldsymbol{\Omega}$ is **not countable**, it is possible to define the probability using sets such as $(a, b]$ (for $b \geq a$). In this case, the axioms of probability imply that, for a suitable random variable X,

$$P(-\infty < X < \infty) = 1,$$

$$0 \leq P(a < X \leq b) \leq 1 \text{ for any } b \geq a.$$

Notice that this imply that

$$P(X = c) = P(c \leq X \leq c) = 0$$

and that it is still possible to define a cumulative distribution function $F_X(.)$, where

$$F_X(x) = P(-\infty < X \leq x).$$

Continuous random variables

A random variable is continuous if there is a function $f_X(.)$ such that

$$(i)\ f_X(x) \geq 0 \text{ for any } x$$

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

The function $f_X(.)$ is called **probability density function**; $F_X(x)$ is the cumulative distribution function (when it is clear that these are functions for random variable X, $f(x)$ and $F(x)$ is used instead).

For any two a, b, $a \leq b$,

$$P(a < X \leq b) = \int_{a}^{b} f_X(t)dt.$$

Example: on the beach.

Jasmine goes to the beach early in the morning. At around lunchtime, David goes to the beach as well. The beach is 4 km long, and the probability that Jasmine is in any part of the beach is continuously distributed with

$$f_X(x) = \begin{cases} 0.25 \text{ if } 0 \leq x \leq 4, \\ 0 \text{ otherwise.} \end{cases}$$

What is the probability that David will find Jasmine if he looks for her between the first and second km of the beach?

$$P(a \leq X \leq b) = \int_a^b f_X(x)dx = \int_1^2 0.25\, dx = 0.25[x]_1^2 = 0.25(2-1) = 0.25$$

★ Example/Practice. Level of charge with one battery.

The level of charge in a battery is a random variable X which may take values between 0 and 1 (so, $X = 0$ means that the battery has no charge, and $X = 1$ that it is fully charged) with density

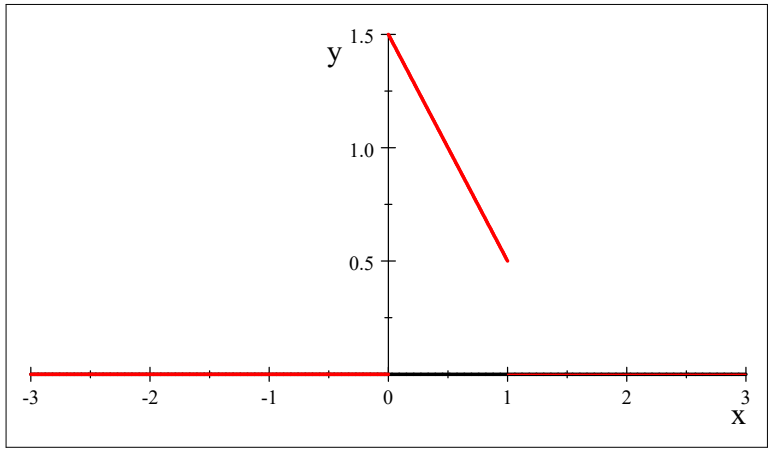
$$f(x) = \frac{3}{2} - x \text{ if } x \in [0, 1]$$

and 0 elsewhere.

Compute the probabilities $P(X \leq 0.5)$ and $P(0.1 < X \leq 0.6)$

Compute the value of x such that $P(X \leq x) = 0.5$

Before we start the discussion we plot $f(x)$



the probability is allocated between 0 and 1, and more is allocated near 0.

Therefore:

Since $X = 0.5$ is the middle of $[0,1]$, and the probability is more clustered around 0, then we should get $P(X \leq 0.5) > 1/2$ and $P(X \leq 0.5) > P(0.1 < X \leq 0.6)$.

Moreover, the value of x such that $P(X \leq x) = 0.5$ should be in the interval $(0, 1/2)$.

♦ Compute the probabilities $P(X \leq 0.5)$ and $P(0.1 < X \leq 0.6)$.

$$P(0.1 < X \leq 0.6) = \int_{0.1}^{0.6} \left(\frac{3}{2} - x\right) dx$$

$$= \left(\frac{3}{2} \times 0.6 - \frac{1}{2} \times 0.6^2\right) - \left(\frac{3}{2} \times 0.1 - \frac{1}{2} \times 0.1^2\right) = 0.575$$

$$P(X \leq 0.5) = \int_{0}^{0.5} \left(\frac{3}{2} - x\right) dx = \left[\frac{3}{2}x - \frac{1}{2}x^2\right]_0^{0.5} = 0.625$$

♦ Compute the value of x such that $P(X \leq x) = 0.5$.

$$P(X \leq x) = \int_{0}^{x} \left(\frac{3}{2} - s\right) ds = \frac{3}{2}x - \frac{1}{2}x^2$$

$$\frac{3}{2}x - \frac{1}{2}x^2 = 0.5$$

$$x = 0.38197$$

note that $\frac{3}{2}x - \frac{1}{2}x^2 = 0.5$ is a second degree equation, so it has two solutions, but the second solution, 2. 618, is discarded (because $f(x) = \frac{3}{2} - x$ only if $x \in [0,1]$)

Joint distribution of random variables

Sometimes we are interested in more than one event at the same time.

Example: stock and bond.
Let X be the return of a given stock, and Y the returns of a given bond. The stock return may either take value $x_1 = 0$ or $x_2 = 10$, and the bond return may either take value $y_1 = 0$ or $y_2 = 2$. Four outcomes are then possible, and the probability of each outcome is

		y_j	
		0	2
x_i	0	0.2	0.5
	10	0.2	0.1

so $f_{X,Y}(x_i, y_j) = P(X = x_i \cap Y = y_j)$.

Notation $p_{i,j} = P(X = x_i \cap Y = y_j)$ is sometimes also used.

Joint probability function. The list of the $f_{X,Y}(x,y)$ is the joint probability function ("joint distribution") of the two discrete random variables X and Y.

The joint probability function for discrete random variables is such that

$$(i)\ f_{X,Y}(x,y) \geq 0$$

$$(ii)\ \sum_x \sum_y f_{X,Y}(x,y) = 1$$

It is also possible to define a **joint cumulative distribution function**,

$$F_{X,Y}(x,y) = P(X \leq x, Y \leq y) = \sum_{i \leq x} \sum_{j \leq y} f_{X,Y}(x,y).$$

Marginal probability functions. The list of the $f_X(x_i) = P(X = x_i)$, $f_Y(y_j) = P(Y = y_j)$ are the "marginal probability functions" ("marginal distributions") of X and Y. They can be computed as

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

for instance, in the Stock and Bond example,

$$P(X = 0) = 0.2 + 0.5 = 0.7, P(Y = 2) = 0.5 + 0.1 = 0.6$$

i.e.

x_i	0	10
$f_X(x_i)$	0.7	0.3

and

y_j	0	2
$f_Y(y_j)$	0.4	0.6

For two **continuous** random variables X and Y, there is a **joint density** $f_{X,Y}(x,y)$ such that

$$P(a < X \leq b, c < Y \leq d) = \int_a^b \int_c^d f_{X,Y}(x,y)dydx$$

(replacing "<" with "≤" has no effect).

A function $f_{X,Y}(.)$ can serve as density if

$$(i)\ f_{X,Y}(x,y) \geq 0$$

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

It is also possible to define a **joint cumulative distribution function**,

$$F_{X,Y}(x,y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(s,t)dsdt.$$

The marginal density functions, $f_X(x)$ or $f_Y(y)$, can be obtained ("integrating out") as

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

Given the joint cumulative distribution functions, it is also possible to define the marginals.

For discrete random variables these are

$$F_X(x) = P(X \leq x) = \sum_{t \leq x} f_X(t) = \sum_{t \leq x} \sum_{y} f_{X,Y}(t,y),$$

$$F_Y(y) = P(Y \leq y) = \sum_{s \leq y} f_Y(s) = \sum_{x} \sum_{s \leq y} f_{X,Y}(x,s)$$

and, for continuous random variables,

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f_{X,Y}(t,s)dtds,$$

$$F_Y(y) = P(Y \leq y) = \int_{-\infty}^{y} f_Y(s)ds = \int_{-\infty}^{\infty} \int_{-\infty}^{y} f_{X,Y}(t,s)dtds$$

Conditional probability functions: the list of the probabilities $P(X = x|Y = y)$, $P(Y = y|X = x)$, where

$$P(X = x|Y = y) = \frac{P(X = x \cap Y = y)}{P(Y = y)}$$

$$P(Y = y|X = x) = \frac{P(X = x \cap Y = y)}{P(X = x)}$$

are the "conditional probability functions" ("conditional distribution") of X given Y and of Y given X. For instance, list the probabilities of Y when $X = 0$ in the Stock and Bond example:

$$\frac{0.2}{0.7} = 0.285\ldots \approx 0.29,\ \frac{0.5}{0.7} = 0.714\ldots \approx 0.71$$

so

$y_j\|X = 0$	0	2
$f_{Y\|X=0}(y)$	0.29	0.71

Note: this example also shows another thing: as we fixed X (to $X = 0$ in this case), $f_{Y|X=0}(y)$ is just a probability function for Y.

It is interesting to compare the conditional probability function $f_{Y|X=0}(y)$ to the marginal $f_Y(y)$,

y	0	2
$f_Y(y)$	0.4	0.6

,

y	0	2
$f_{Y\|X=0}(y)$	0.29	0.71

We can see that the information on X is important to know something about Y: for example we know that if $X = 0$ then the probability of $Y = 2$ is higher (0.71 vs 0.6) than when we know nothing about X.

We could also compute $f_{Y|X=10}(y)$, and in particular, $P(Y = 2|X = 10) = 1/3 \approx 0.33$, thus seeing that knowledge that $X = 10$ takes the probability of $Y = 2$ to 0.33.

In economics we are mostly interested in establishing conditional statements, so we are most interested in conditional probabilities and conditional probability functions.

Conditional density functions. For two continuous random variables Y and X, with joint probability density $f_{X,Y}(x,y)$ and marginals $f_X(x), f_Y(y)$, the conditional density functions $f_{X|Y}(x|y), f_{Y|X}(y|x)$ are

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}, f_Y(y) \neq 0$$

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}, f_X(x) \neq 0$$

Note: The conditional density functions are density functions, and meet all the characteristics of density functions.

★ Example/Practice. Level of charge with two batteries.

The levels of charge in two batteries are two random variables X and Y with joint density

$$f_{X,Y}(x,y) = (2-(x+y)) \text{ if } x \in [0,1] \text{ and } y \in [0,1]$$

and 0 elsewhere.

♦ Compute the probability $P(X \leq 0.5, Y \geq 0.5)$

$$P(X \leq 0.5, Y \geq 0.5) = \int_0^{0.5} \left(\int_{0.5}^{1} (2-(x+y))dy \right) dx = \int_0^{0.5} \left[2y - xy - \frac{1}{2}y^2 \right]_{0.5}^{1} dx$$

$$= \int_0^{0.5} \left(\left(2 - x - \frac{1}{2} \right) - \left(2 \times 0.5 - x \times 0.5 - \frac{1}{2} \times 0.5^2 \right) \right) dx$$

$$= \int_0^{0.5} (0.625 - 0.5x)dx = \left[0.625x - 0.5\frac{1}{2}x^2 \right]_0^{0.5}$$

$$= 0.625 \times 0.5 - 0.5\frac{1}{2}0.5^2 = 0.25$$

♦ Compute the marginal density $f_X(x)$

$$f_X(x) = \int_0^1 (2 - (x + y))dy = \left[2y - xy - \frac{1}{2}y^2 \right]_0^1 = 2 - x - \frac{1}{2} = \frac{3}{2} - x$$

♦ Compute the conditional density $f_{Y|X}(y|0.25)$

$$f_{Y|X}(y|0.25) = \frac{(2 - (\frac{1}{4} + y))}{\frac{3}{2} - \frac{1}{4}} = \frac{7}{5} - \frac{4}{5}y$$

♦ Compute the conditional probability $P(Y < 0.5|X = 0.25)$

$$F_{Y|X}(0.5|0.25) = \int_0^{0.5} \left(\frac{7}{5} - \frac{4}{5}y \right) dy = 0.6$$

Independently distributed random variables.

Two discrete random variables X and Y with joint probability function $f_{X,Y}(x,y)$ and marginals $f_X(x)$ and $f_Y(y)$ are independent if and only if

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

Notice that the reference to "all x,y" is important, because $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ may hold for some combinations of x,y, and yet fail to hold for all of them.

In the same way, when X and Y are continuous random variables, with joint probability density $f_{X,Y}(x,y)$ and marginals $f_X(x)$ and $f_Y(y)$, independence implies and is implied by

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

- When $X_1, ..., X_n$ are n **independently**, **identically distributed random variables**, we also write $X_1, ..., X_n \sim i.i.d.$.

Example. To verify that in the Stock and Bond example that the returns X and Y are not independent, consider for instance $x = 0, y = 0$. Then, $0.2 \neq 0.7 \times 0.4$.

Notice that for discrete random variables this definition is just the application of the definition of independence for two events: when two events A and B are considered, we know that they are indendependent if

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

so we have just applied this formula to the events $A = \{X = x\}$, $B = \{Y = y\}$.

However, we may change the values of x or of y: the definition of independent random variables requires that $P(A \cap B) = P(A)P(B)$ holds for all the eligible values.

Transformations of identically distributed random variables

Let X and Y be identically distributed random variables. Then, for any continuous function g, $g(X)$ and $g(Y)$ are also identically distributed random variables.

Transformations of independent random variables

Let X and Y be independently distributed random variables. Then, for any continuous function g, $g(X)$ and $g(Y)$ are also independently distributed random variables.

Section 3
Mathematical expectations

- Objectives: to introduce summary numbers to summarize the information in probability functions.
- Topics: mean, variance and higher moment; covariance and correlation; conditional moments.
- References: Miller and Miller, Chapter 4 (sections 4.1 to 4.9) and Chapter 8 (section 8.2); Wooldridge, Appendix B (sections B.3 and B.4)

It is often desirable to summarise some properties of a distribution in just a few numbers (for example, to compare random variables with different distributions).

Example. Suppose that you have a ticket of a lottery with prize value of 50£, and that 100 tickets have been sold. Calling X the value of the ticket, the probability function is

x	0	50
$f(x)$	0.99	0.01

There is also another lottery, that awards a second prize, with value 10£. Calling Y the value of the ticket, the probability function is

y	0	10	50
$f(y)$	0.98	0.01	0.01

How can we compare lotteries X and Y?

Example: In the Accident and Emergency room of a hospital, 0, 1, 2, 3, or 4 people injured arrive each hour, distributed according to the probability

x	0	1	2	3	4
$f(x)$	0.10	0.77	0.08	0.03	0.02

How can we summarize the number of injured per hour?

We will introduce the mean and the variance as two summary measures.

Mean (**expected value**)

For a discrete random variable X with $P(X = x_j) = f_X(x_j)$, the expected value $E(X)$ is

$$E(X) = \sum_j f_X(x_j)x_j$$

For a continuously distributed random variable with density $f_X(x)$, the expected value is

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

The expected value is often denoted as μ_X

Note: do not confuse the mean ($E(X)$) with the sample mean ($\bar{X}$).

In the example of the two lotteries, $E(X) = 0.5$ and $E(Y) = 0.6$.

In the example of throwing a fair die, $E(X) = 3.5$.

In the example ER,

$$E(X) = 0.10 \times 0 + 0.77 \times 1 + 0.08 \times 2 + 0.03 \times 3 + 0.02 \times 4 = 1.1.$$

In the example of the level charge of a battery,

$$E(X) = \int_0^1 x\left(\frac{3}{2} - x\right)dx = \left[\frac{3}{4}x^2 - \frac{1}{3}x^3\right]_0^1 = \frac{5}{12}$$

The expected value is a weighted average of the possible values x_j with the weights being the corresponding $f_X(x_j)$.

The expected value is not necessarily the most frequent outcome, indeed it may not be an outcome at all; it is informative about the "middle" of the distribution, but does not necessarily split the distribution in two parts; sometimes it is said that it is a measure of the central tendency, or the "centre" of gravity of the distribution, because if we repeat the experiment many times and we average the realisations, under regularity conditions that average should be very close to the expected value.

Expectations for functions of random variables.

When X is a random variable with probabilities assigned by $f(x)$, then $g(X)$ (for a given function $g : \mathbb{R} \to \mathbb{R}$) is also a random variable, with probabilities (still) assigned by $f(x)$. We can then define, for discrete random variables,

$$E(g(X)) = \sum_x f(x)g(x)$$

or, for continuous random variables and for a generic function g,

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

The **variance** $Var(X)$ is

$$Var(X) = E(X - E(X))^2$$

or, recalling $E(X) = \mu_X$,

$$Var(X) = E(X - \mu_X)^2.$$

The variance is often denoted by σ_X^2.

- For a discrete random variable X with $P(X = x_j) = f_X(x_j)$, this is

$$Var(X) = \sum_j f_X(x_j)(x_j - \mu_X)^2$$

- For a continuous random variable with density $f_X(x)$, this is

$$Var(X) = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) dx$$

- The formula is equivalent to

$$Var(X) = E(X^2) - \mu_X^2$$

★ Example: up and down

Consider the two random variables, X and Y, with probability distributions $f_X(x_j)$ and $f_Y(y_j)$ respectively, as in

x_j	-1	0	1
$f_X(x_j)$	$1/3$	$1/3$	$1/3$

,

y_j	-1	0	1
$f_Y(y_j)$	$1/4$	$1/2$	$1/4$

Here both the random variables have the same support (i.e., may take the same realisations), and in both cases the expected value is 0.

$$Var(X) = 1/3 \times (-1)^2 + 1/3 \times 0^2 + 1/3 \times (1)^2 = 2/3$$
$$Var(Y) = 1/4 \times (-1)^2 + 1/2 \times 0^2 + 1/4 \times (1)^2 = 1/2$$

The variance is a weighted average of the squared spreads $x_j - \mu_X$ with the weights being the corresponding $f_X(x_j)$.

The variance is a measure of the dispersion of the random variable around μ_X. Notice that it is scaled by squares of x_j and of μ_X, so for comparison purposes it is usually more interesting to use the **standard deviation**, σ_X, instead. This is defined as $\sigma_X = \sqrt{Var(X)}$.

Theorems for the mean and the variance

For a random variable X with $E(X) = \mu_X$, $Var(X) = \sigma_X^2$, and for constant a, b,

$$(i)\ E(a + bX) = a + b\mu_X$$

$$(ii)\ Var(a + bX) = b^2\sigma_X^2.$$

Standardisation

Let X be such that $E(X) = \mu$, $Var(X) = \sigma^2$, and let $Z \equiv -\frac{\mu}{\sigma} + (\frac{1}{\sigma})X$, so that

$$Z = \frac{X - \mu}{\sigma}$$

Then,

$$E(Z) = 0, Var(Z) = 1.$$

♦ Proof of $E(Z) = 0$

$$E(Z) = E\left(\frac{X - \mu}{\sigma}\right) = E\left(\frac{X}{\sigma}\right) - \frac{\mu}{\sigma} = \frac{1}{\sigma}E(X) - \frac{\mu}{\sigma} = \frac{1}{\sigma}\mu - \frac{\mu}{\sigma} = 0$$

using Theorem *(i)*, as μ and σ are constants.

♦ Proof of $Var(Z) = 1$

$$Var(Z) = Var\left(\frac{X - \mu}{\sigma}\right) = Var\left(\frac{X}{\sigma}\right) = \frac{1}{\sigma^2}Var(X) = \frac{1}{\sigma^2}\sigma^2 = 1$$

using Theorem *(ii)*, as μ and σ are constants.

Higher moments

The third moment about the mean is also called **skewness**, and it often denoted as μ_3

$$\mu_3 = E(X-\mu)^3.$$

When $\mu_3 = 0$, the function is symmetric; when $\mu_3 > 0$, it is "skewed to the right" (the right tail of the distribution is longer), when $\mu_3 < 0$, it is "skewed to the left" (the left tail of the distribution is longer). For comparison purposes, the normalised measure

$$\alpha_3 = \frac{\mu_3}{\sigma^3}$$

is more of interest.

In the example of the ER,

$$\mu_3 = 0.618, \sigma^2 = 0.47, \alpha_3 = 1.9180$$

The fourth moment about the mean is also called **kurtosis**, and it often denoted as μ_4

$$\mu_4 = E(X-\mu)^4.$$

The kurtosis measures the peakedness of the distribution or of the density. For comparison purposes, the normalised measure $\alpha_4 = \frac{\mu_4}{\sigma^4}$ is more of interest.

The r^{th} moment about the mean, often denoted as μ_r, is

$$\mu_r = E(X-\mu)^r.$$

Why this form for the moments? Why, for example, not taking $E|X-\mu|$ as measure of dispersion?

- *Uniqueness Theorem* Under some regularity conditions, the moments characterise the distribution univocally, i.e., if we know all the moments, then we know the distribution.

Using conditional distributions, we can also define **conditional moments**. For two random variables, X and Y,

$$\mu_{Y|X=x} = E(Y|X=x)$$
$$\sigma^2_{Y|X=x} = Var(Y|X=x) = E(Y^2|X=x) - (E(Y|X=x))^2.$$

Of course, conditional moments for X may be defined in the same way.

For discrete random variables, $E(Y|X=x) = \sum_j y_j f_{Y|x}(y_j|x)$;

For continuous random variables, $E(Y|X=x) = \int_{-\infty}^{\infty} y f_{Y|x}(y|x)dy$;

In the Stock and Bond example,

$$E(Y|X=0) = 0 \times 0.29 + 2 \times 0.71 = 1.42.$$

For comparison, notice that $E(Y) = 1.2$, and $E(Y|X=10) = 0.66\ldots$.

(In the same way, $Var(Y|X=0) \approx 0.82$ and $Var(Y|X=10) \approx 0.89$ whereas $Var(Y) = 0.96$. It is also interesting to notice here that conditioning has reduced the variance)

Covariance of random variables.

Let X and Y two random variables with joint probability function $f_{X,Y}(x_i y_j)$, and marginal probability functions $f_X(x_i)$ and $f_Y(y_j)$ respectively, and such that $E(X) = \mu_X$, $E(Y) = \mu_Y$, $Var(X) = \sigma_X^2$, $Var(Y) = \sigma_Y^2$. Then

$$Cov(X, Y) = E((X - \mu_X)(Y - \mu_Y))$$

is the covariance between X and Y (often indicated as σ_{XY}).

Theorems for the covariance

For constant a, b,

$$(i)\ Cov(a + X, b + Y) = Cov(X, Y)$$

$$(ii)\ Cov(X, Y) = E(XY) - \mu_X \mu_Y$$

$$(iii)\ Cov(aX, bY) = abCov(X, Y)$$

(iv)If X and Y are independent, then $Cov(X, Y) = 0$

Correlation of random variables

The covariance is informative of the association between two random variables. However, it is affected by the unit of measurement. The correlation is a measure of the association between two random variables that is not affected by the unit of measurement. For two random variables X, Y, the correlation ρ_{XY} is defined as

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

It can be shown that $-1 \leq \rho_{XY} \leq 1$, and that the absolute value of the ρ_{XY} coefficients are not affected by linear transformations, so if $A = \alpha_1 + \alpha_2 X$, $B = \beta_1 + \beta_2 Y$, where α_1, α_2, β_1, β_2 are constant, then $\rho_{AB} = sign(\alpha_2, \beta_2)\rho_{XY}$, where $sign(\alpha_2, \beta_2) = 1$ if $\alpha_2\beta_2 > 0$, $sign(\alpha_2, \beta_2) = -1$ if $\alpha_2\beta_2 < 0$.

In the example of Stock and Bond,

$$E(XY) = 0 \times 0 \times 0.2 + 0 \times 2 \times 0.5 + 10 \times 0 \times 0.2 + 10 \times 2 \times 0.1 = 2.$$

Since $\mu_X = 3$, $\mu_Y = 1.2$,

$$\sigma_{XY} = 2 - 3 \times 1.2 = -1.6.$$

Since $\sigma_X^2 = 21$, $\sigma_Y^2 = 0.96$,

$$\rho_{XY} = \frac{-1.6}{\sqrt{21 * 0.96}} = -0.35635.$$

Sum of random variables

Let

$$W = X + Y,$$

where $E(X) = \mu_X$, $Var(X) = \sigma_X^2$, $E(Y) = \mu_Y$, $Var(Y) = \sigma_Y^2$, $Cov(X, Y) = \sigma_{XY}$. Then,

$$E(W) = E(X + Y) = E(X) + E(Y) = \mu_X + \mu_Y,$$

and

$$\begin{aligned} Var(W) &= E(W - E(W))^2 \\ &= E(X + Y - (\mu_X + \mu_Y))^2 = E(X - \mu_X + Y - \mu_Y)^2 \\ &= E(X - \mu_X)^2 + E(Y - \mu_Y)^2 + 2E((X - \mu_X)(Y - \mu_Y)) = \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY}. \end{aligned}$$

In the Stock and Bond example, consider the portfolio $W = \frac{1}{10}X + \frac{9}{10}Y$. Then, $E(W) = \frac{1}{10} \times 3 + \frac{9}{10} \times 1.2 = 1.38$ and

$$Var(W) = \left(\frac{1}{10}\right)^2 \times 21 + \left(\frac{9}{10}\right)^2 \times 0.96 - \left(\frac{1}{10}\,\frac{9}{10}\right) \times 2 \times 1.6 = 0.6996$$

Sum of more than two random variables

The results for the mean and variance of two random variables can be extended to a generic number n of random variables. This is particularly interesting when the random variables are independently, identically distributed.

Let $X_1,...,X_n$ be n independently, identically distributed random variables with $E(X_i) = \mu_X$, $Var(X_i) = \sigma_X^2$, then

$$E\left(\sum_{i=1}^{n} X_i\right) = n\mu_X$$

$$Var\left(\sum_{i=1}^{n} X_i\right) = n\sigma_X^2$$

Further results concerning moments. Chebyshev's inequality.

Let Y be a random variable with $E(Y^2) < \infty$. Then

$$P(|Y| \geq \varepsilon) = P(Y^2 \geq \varepsilon^2) \leq \frac{E(Y^2)}{\varepsilon^2}$$

for all $\varepsilon > 0$.

Note. An interesting implication is that, letting $E(Y) = \mu$, $Var(Y) = \sigma^2$, then

$$P(|Y - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}$$

so the variance is an upper bound of the measure of the probability on the tails (defined as the support that is at least ε away from μ).

Further results concerning moments. Law of iterated expectations

For random variables X, Y

$$E(E(Y|X)) = E(Y) \text{ and } E(E(X|Y)) = E(X)).$$

Example: $E(E(Y|X))$ and $E(Y)$ in the Stock and Bond example.

Recall that the we computed $E(Y) = 1.2$, $E(Y|0) = 1.42$ and $E(Y|10) = 0.66$.

So, $E(Y|x)$ changes with the as we change $X = x$, and we know the probabilities $P(X = x)$ are $P(X = 0) = 0.7$ and $P(X = 10) = 0.3$.

$E(Y|X)$ is therefore a random variable (in X) with distribution

$E(y\|x)$	0.66	1.42
$f(x)$	0.3	0.7

We can then compute the expectation

$$E(E(Y|X)) \approx 0.66 \times 0.3 + 1.42 \times 0.7 = 1.19$$

Note: the apparent difference between $E(E(Y|X))$ and $E(Y)$ in this example is only due to rounding errors.

Note: notice that $E(Y|X)$, is a random variable, but $E(Y|x)$ is not a random variable. In $E(Y|x)$ we fixed $X = x$; it is only when we allow X to take all the values $x_1,...,x_m$, that we have the random variable $E(Y|X)$.

Further results concerning moments. Jensen's inequality

For a random variable X, and for a a convex function g,

$$E(g(X)) \geq g(E(X))$$

Example: consider the function $g(X) = X^2$, and suppose that X is distributed as

x	0	1	2	3
$f(x)$	0.25	0.25	0.25	0.25

so that

$g(x)$	0	1	4	9
$f(g(x))$	0.25	0.25	0.25	0.25

Then, $E(X) = \frac{3}{2}$ and $g(E(X)) = (\frac{3}{2})^2 = \frac{9}{4}$ while $E(g(X)) = \frac{14}{4}$.

Note:
taking $g(X) = X^2$ we can also see that, for a generic random variable X,

$$E(X^2) \geq (E(X))^2$$

Rearranging terms,

$$E(X^2) - (E(X))^2 \geq 0.$$

Since $E(X^2) - (E(X))^2 = Var(X)$, we verified that the variance is non-negative (we already knew that the variance is non-negative; this application of Jensen's inequality gives an alternative proof).

Section 4
Some special parametric distributions

- Objectives: to introduce some special parametric distributions that are routinely used in econometrics.
- Topics: Bernoulli, binomial, normal, χ^2, t and F distributions. Joint and conditional normal distributions.
- References: Miller and Miller, Chapter 5, Sections 5.1, 5.3, 5.4 and 5.7; Chapter 6, Sections 6.1 - 6.3, 6.5 - 6.8 and Chapter 8, Sections 8.4 - 8.6; Wooldridge, Appendix B (section B.5)

Bernoulli distribution

Let X be a random variable that can only take values 1 or 0, and such that

$$P(X = 1) = p, 0 < p < 1.$$

Then, X is Bernoulli distributed. For $x = 0, x = 1$, the probability function is

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

(of course, $P(X = x) = f(x;p)$ but here we have made explicit reference to the parameter p) and

$$E(X) = p, Var(X) = p(1-p).$$

The Bernoulli is used for events with two outcomes which can be classified as "success" or "failure" (e.g., tossing a coin, classifying "head" as "success").

Each Bernoulli experiment is also called a "bernoulli trial".

A random variable that can only take values 0 and 1 is also often called "indicator function", and it is often indicated as I.

Binomial distribution

The binomial distribution is appropriate if

1. the experiment can be regarded as of n independent trials
2. each trial can have one of the two mutually exclusive outcomes: success or failure
3. the probability of a successful outcome is p for each trial

Notice that the trials are identically distributed as well.

Let X be the number of successes, then

$$P(X = x) = B(x; p, n) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

and

$$E(X) = np,\ Var(X) = np(1-p).$$

Mike Walters is an estate agent. At the moment, he has a portfolio of 6 houses, and he knows that he will sell before the end of the month each one of those with probability $p = 0.6$. For each house, the sale is independent from the other ones.

Using the binomial distribution, with $p = 0.6$ and $n = 6$,

x_i	0	1	2	3	4	5	6
$f_X(x_i)$	0.004096	0.036864	0.13824	0.27648	0.31104	0.18662	0.046656

As $np = 3.6$, Mike expects to sell 3.6 houses.

Normal distribution

Let X be a continuous random variable with density (*pdf*)

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

then X is Normally distributed with mean μ and variance σ^2, and we indicate this with the notation

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

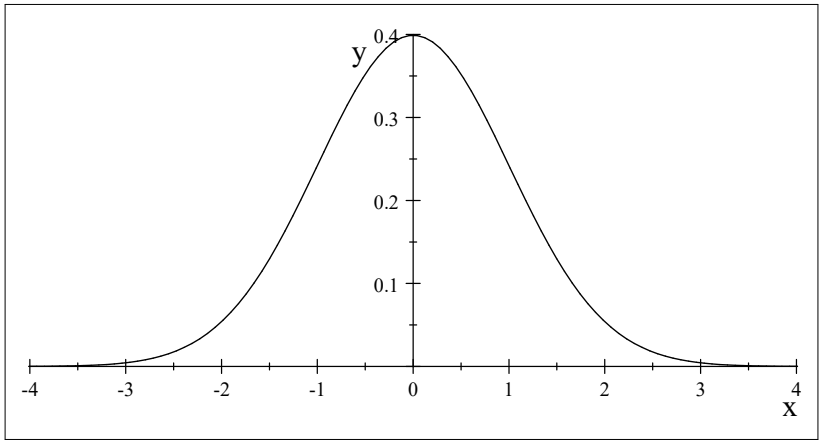
The density is

- defined for all x
- symmetric around μ
- bell shaped (shape depending on σ^2)

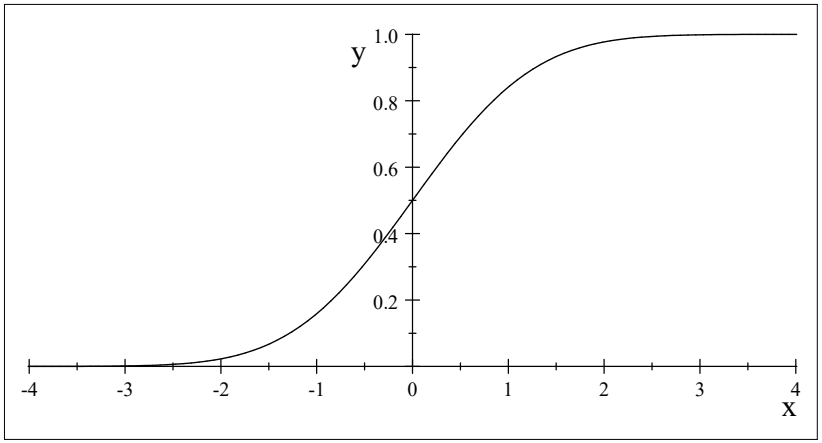
When $\mu = 0$, $\sigma^2 = 1$, the distribution is usually referred as Standard Normal, and the random variable is often indicated as Z. The *pdf* of Z is usually indicated as $\phi(z)$, the *cdf* is usually indicated as $\Phi(z)$.

Some examples: *pdf* and *cdf* of the standard normal

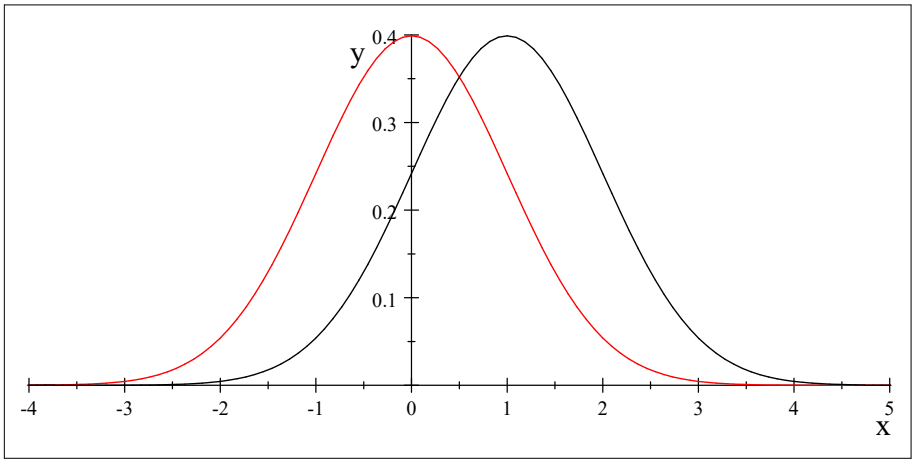
$$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$



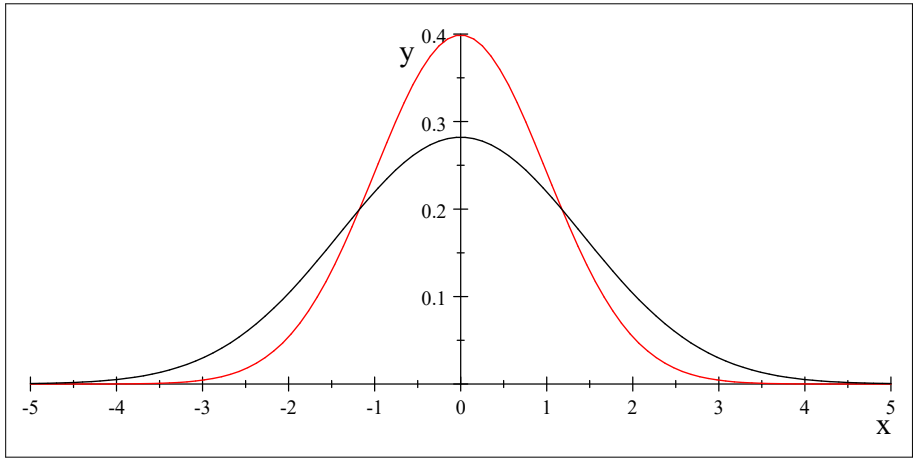
$$\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt$$



$$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \text{ and } \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-1)^2} \quad (\mu = 0 \text{ and } \mu = 1)$$



$$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \text{ and } \frac{1}{\sqrt{2\pi * 2}} e^{-\frac{1}{2*2}x^2} \quad (\sigma^2 = 1 \text{ and } \sigma^2 = 2)$$



Standardisation

When $X \sim N(\mu, \sigma^2)$, direct calculation of probabilities, such as for example in $P(X \leq c)$, is not possible, because $\int_{-\infty}^{c} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx$ does not have close form solution.

Calculation would then only be possible via numeric approximation, which of course is not practical in general.

Tables could be produced, but they would depend on μ and on σ^2: this too would be practically infeasible.

Instead, tables are only produced for $Z \sim N(0,1)$: in order to dares probability statements for X, we will transform the statement using the fact that

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

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

$$P(c \le X \le d) = P(c-\mu \le X-\mu \le d-\mu)$$

$$= P\left(\frac{c-\mu}{\sigma} \le \frac{X-\mu}{\sigma} \le \frac{d-\mu}{\sigma}\right)$$

$$= P\left(\frac{c-\mu}{\sigma} \le Z \le \frac{d-\mu}{\sigma}\right) = \Phi\left(\frac{d-\mu}{\sigma}\right) - \Phi\left(\frac{c-\mu}{\sigma}\right).$$

★ Practice. Let X have a Normal distribution with $\mu = 1$ and $\sigma^2 = 16$.

Making approximations when necessary, calculate the following probabilities: $P(X \leq 0.052)$; $P(X \leq 10.44)$; $P(X > 2.72)$; $P(X > -1.04)$; $P(-2.0 < X < 7.0)$.

Find c such that $P(X < c) = 0.25$; $P(X > c) = 0.05$.

$X \sim N(1, 16)$ and so standardization yields

$$Z = \frac{X - 1}{4} \sim N(0, 1).$$

Therefore,

$$P(X \leq 0.052) = P(Z \leq (0.052 - 1)/4) = P(Z \leq -0.237) = 0.4063.$$

$$P(X \leq 10.44) = P(Z \leq (10.44 - 1)/4) = P(Z \leq 2.36) = 0.9909.$$

$$P(X > 2.72) = P(Z > (2.72 - 1)/4) = P(Z > 0.43) = 0.3336.$$

$$P(X > -1.04) = P(Z > (-1.04 - 1)/4) = P(Z > -0.51) = 0.6950.$$

$$\begin{aligned} P(-2.0 < X < 7.0) &= P(X < 7.0) - P(X \le -2.0) \\ &= P(Z < (7.0 - 1)/4) - P(Z \le (-2.0 - 1)/4) \\ &= P(Z < 1.5) - P(Z \le -0.75) \\ &= 0.9332 - 0.2266 = 0.7066. \end{aligned}$$

$$P(X < c) = P\left(\frac{X-1}{4} < \frac{c-1}{4}\right) = P\left(Z < \frac{c-1}{4}\right)$$

from tables, $P(Z < -0.67) = 0.25$, so $\frac{c-1}{4} = -0.67$, $c = -1.68$.

$$P(X > c) = P\left(\frac{X-1}{4} > \frac{c-1}{4}\right) = P\left(Z > \frac{c-1}{4}\right)$$

from tables, $P(Z > 1.64) = 0.05$, so $\frac{c-1}{4} = 1.64$, $c = 7.56$.

★ Practice. Let Z have a Standard Normal distribution (i.e., with $\mu = 0$ and $\sigma^2 = 1$).

Calculate $P(|Z| > 1.6)$

$$P(|Z| > 1.6) = P(Z < -1.6) + P(Z > 1.6)$$
$$= 2 \times P(Z < -1.6) = 2 \times 0.0548 = 0.1096$$

Find c such that $P(|Z| > c) = 0.05$.

$$P(|Z| > c) = 2 \times P(Z < -c) \text{ so}$$
$$P(Z < -c) = 0.025, \text{ so } -c = -1.96,$$
$$c = 1.96$$

Bivariate Normal distribution

Let

$$X \sim N(\mu_1, \sigma_1^2),\ Y \sim N(\mu_2, \sigma_2^2)$$

then X and Y are jointly distributed according to a bivariate normal distribution. Letting

$$Cov(X, Y) = \sigma_{12},\ Cor(X, Y) = \rho,$$

the bivariate normal has density

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{ -\frac{\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2 \right]}{2(1-\rho^2)} \right\}$$

- If $\rho = 0$, then $f_{XY}(x,y) = f_X(x)f_Y(y)$, so X and Y are independent.
- If $\rho \neq 0$,

$$Y|x \sim N\Big(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1), \sigma_2^2(1-\rho^2)\Big)$$

- Recalling that $\rho = \frac{\sigma_{12}}{\sigma_1\sigma_2}$, the conditional mean $(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1))$ is also indicated as

$$\mu_2 + \frac{\sigma_{12}}{\sigma_1^2}(x-\mu_1);$$

- The conditional mean is linear in x;
- The conditional variance $(\sigma_2^2(1-\rho^2))$ is smaller than σ_2^2.

Sum of normally distributed random variables

We already know how to compute the expected value or the variance of sums of random variables. In some cases we also verified that some random distributions may be seen as the distribution of a sum of random variables (example, the binomial is the distribution of the sum of n independent bernoulli trials).

Moreover:

- If X_1 and X_2 are normally distributed, then $X_1 + X_2$ is also normally distributed.
- As for the moments of $X_1 + X_2$, if $X_1 \sim N(\mu_1, \sigma_1^2)$, if $X_2 \sim N(\mu_2, \sigma_2^2)$, $cov(X_1, X_2) = \sigma_{12}$, then

$$X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2 + 2\sigma_{12})$$

This is a very special result: it is not usually true that the sum of two random variables with a certain distribution is a random variable with the same distribution.

Other important distributions.

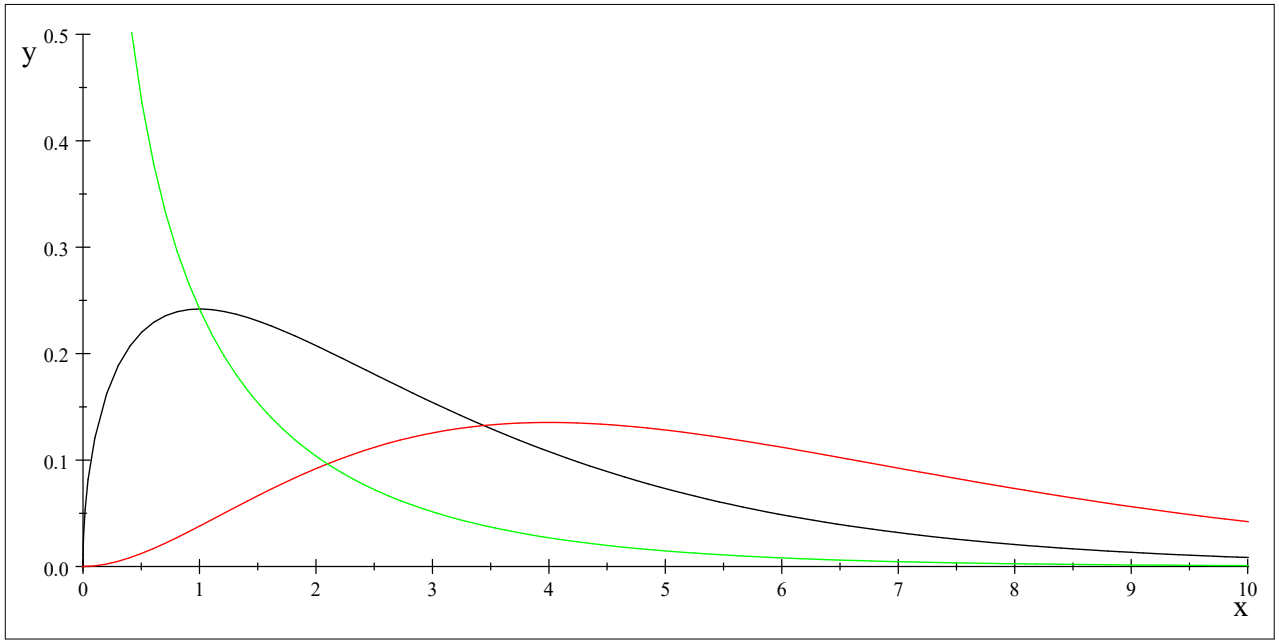
- χ^2. If $Z_1,\ldots,Z_k$ are such that

i) Z_i are identically distributed, $Z_i \sim N(0,1)$ for any $i = 1,\ldots,k$;

ii) Z_i is independent from Z_j for any $j \neq i$, for any $i,j = 1,\ldots,k$,

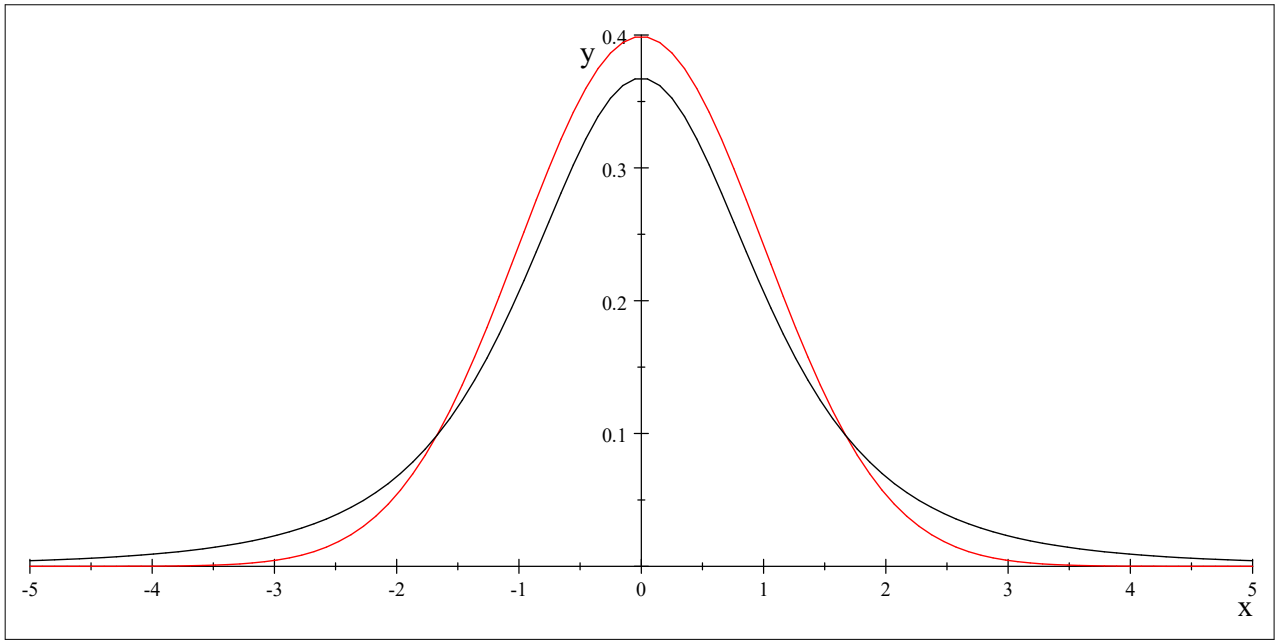
then $C_k = \sum_{i=1}^{k} Z_i^2$ is χ_k^2 (chi-squared with k degrees of freedom) distributed.

Densities of $C_1 \sim \chi_1^2$, $C_3 \sim \chi_3^2$, $C_6 \sim \chi_6^2$



- **t**. If $C_k \sim \chi^2_k$ and $Z \sim N(0,1)$, and C_k and Z are independently distributed, then $T_k = \frac{Z}{\sqrt{C_k/k}}$ is t_k (t with k degrees of freedom) distributed.

Densities of Z and T_3



i) the density f_{T_3} has thicker tails than f_Z

ii) As $k \to \infty, f_{T_k} \to f_Z$.

- **F**. If $C_k \sim \chi^2_k$ and $B_h \sim \chi^2_h$, and C_k and B_h are independently distributed, then $\frac{C_k/k}{B_h/h}$ is $F_{k,h}$ (F with k and h degrees of freedom) distributed.

★ Practice. Using tables,

χ^2: $P(C_1 > 3.84) = 0.05, P(C_2 > 5.99) = 0.05, P(C_4 > 9.49) = 0.05$

t: $P(T_3 > 2.353) = 0.05, P(T_3 < -2.353) = 0.05, P(|T_3| > 3.182) = 0.05$

t: $P(T_{10} > 1.812) = 0.05, P(T_{10} < -1.812) = 0.05, P(|T_{10}| > 2.228) = 0.05$

t: $P(T_{120} > 1.658) = 0.05, P(T_{120} < -1.658) = 0.05, P(|T_{120}| > 1.98) = 0.05$

F: $P(F_{1,10} > 4.96) = 0.05, P(F_{2,20} > 3.49) = 0.05, P(F_{4,120} > 2.45) = 0.05$

Section 5
Asymptotic / large sample properties

- Objectives: to provide an introduction to approximations that are based on consideration of the behaviour of functions of random variables as the number increases;
- Topics: Probability limits, Central limit theorems, the normal distribution as an approximation of the binomial
- References: Miller and Miller, Section 8.2; Wooldridge, Appendix C (section C.3).

Asymptotic / Large sample theory

We are interested in results for the random variables $\{h_n,\ n = 1, 2, \ldots\}$ where $h_n = h(X_1, \ldots, X_n)$.

For example, h_n could be the sample mean $\overline{X}$.

Sometimes we cannot obtain exact results for h_n with n of finite size, and we have to resort to approximations. Many approximations are derived by considering what happens as $n \to \infty$.

Probability limits.

Consider the sequence $\{h_n,\ n = 1,2,\ldots\}$ where $h_n = h(X_1,\ldots,X_n)$, and a random variable h. If

$$\lim_{n\to\infty} P(|h_n - h| \geq \varepsilon) = 0 \text{ for any } \varepsilon > 0$$

then h_n has probability limit h and we can write

$$p\lim_{n\to\infty} h_n = h$$

or

$$h_n \to_p h \text{ as } n \to \infty.$$

The condition $P(|h_n - h| \geq \varepsilon) = 0\ \forall \varepsilon > 0$ means that "h_n is close to h in a probabilistic sense". Notice that "$h_n \neq h$" (or, to be more precise, that $|h_n - h| \geq \varepsilon$) is actually possible, but the probability of it is zero.

Law of large numbers

Let $X_1,...,X_n$ be independent, identically distributed random variables, with $E(X_i) = \mu_X$, $Var(X_i) = \sigma_X^2$ ($\sigma_X^2 < \infty$), and let $\overline{X}$ be the sample mean , $\overline{X} = (1/n)\sum_{i=1}^{n} X_i$. Then

$$p\lim_{n\to\infty}\overline{X} = \mu_X.$$

Proof:

Notice that $Var(\overline{X}) = \sigma_X^2/n$. Then, by the Chebyshev inequality,

$$P(|\overline{X} - \mu| \geq \varepsilon) \leq \frac{\sigma_X^2}{n\varepsilon^2}$$

for all $\varepsilon > 0$, and notice that

$$\frac{\sigma_X^2}{n\varepsilon^2} \to 0 \text{ as } n \to \infty.$$

Discussion.

- $\bar{X}$ is "close enough" to μ_X, in a probabilistic sense, as $n \to \infty$, *regardless* of the distribution of X_i.
- The hypotheses of the Law of large numbers can be generalised, to allow for heterogeneity in the distribution, and for some types of dependence. For example, for any sequence $\{h_n\}$, and for any h, if $E(h_n - h)^2 \to 0$ as $n \to \infty$, then by the Chebyshev's inequality $p\lim_{n\to\infty} h_n = h$.

Example

Consider the experiment of tossing a fair coin, recording $X_i = 1$ if we observe head, $X_i = 0$ if tail, and let $\overline{X}$ be the sample mean. Recall that X_i is bernoulli distributed, with $E(X_i) = 1/2$, $Var(X_i) = 1/2(1 - 1/2) = 1/4$.

We can find bounds for $P(|\overline{X} - 1/2| \geq \varepsilon)$ depending on the number of tosses. For example let $\varepsilon = 0.1$, and

- suppose we did 100 tosses:

$$P(|\overline{X} - 1/2| \geq 0.1) \leq \frac{1/4}{100 \times 0.1^2} = 0.25;$$

- suppose we did 1000 tosses:

$$P(|\overline{X} - 1/2| \geq 0.1) \leq \frac{1/4}{1000 \times 0.1^2} = 0.025.$$

Notice that these bounds only depend on the Chebyshev inequality, and do not depend on the distribution of $\overline{X}$. Sharper bounds may sometimes be computed, if we know that the distribution of $\overline{X}$: however, the bounds from the Chebyshev inequality are sufficient to show $p\lim_{n\to\infty} \overline{X} = 1/2$.

Probability limits often give convergence to constants, which may be considered variables with a degenerate distribution. Convergence to non-trivial random variables (i.e., random variables having a non-degenerate distribution) is obtained by the central limit theorem.

Central limit theorem

Let $X_1, \ldots, X_n$ be independent, identically distributed random variables, with $E(X_i) = \mu_X$, $Var(X_i) = \sigma_X^2$ $(0 < \sigma_X^2 < \infty)$, and

$$Z_n = \sqrt{n}\,\frac{(\bar{X} - \mu_X)}{\sigma_X},$$

then the distribution of Z_n converges to a $N(0,1)$ as $n \to \infty$, and we indicate this as

$$Z_n \to_d N(0,1) \text{ as } n \to \infty.$$

This result may be generalised to allow for heterogeneity in the means and in the variances in the distributions, and to allow for some types of dependence.

Example. In the case of the example of the houses of Mike Walters, we may use the normal approximation of the binomial.

Letting X the number of houses sold, then X is binomial(6,0.6) distributed. Moreover, $X = \sum_{i=1}^{6} X_i$ where X_i is the sale of the individual house, which is bernoulli(0.6) distributed. So, letting $\overline{X} = \frac{1}{n}X = \frac{1}{n}\sum_{i=1}^{6} X_i$, we know that $\sqrt{n}\,\frac{(\overline{X}-\mu_X)}{\sigma_X} \rightarrow_d N(0,1)$ where in this example $\mu_X = E(X_i) = p$ (which is 0.6) and $\sigma_X^2 = Var(X_i) = p(1-p)$ (which is 0.36) so

$$\sqrt{n}\,\frac{(\overline{X}-p)}{\sqrt{p(1-p)}} \rightarrow_d N(0,1)$$

Substituting $X = n\overline{X}$, we can approximate X with

$$Y \sim N(np, np(1-p))$$

as $n \rightarrow \infty$.

Here we approximate points of the binomial by intervals of the normal: to get the best approximation it is convenient in this case (approximation of the binomial with the normal) to have the point of the binomial in the middle of the interval of the normal (except for the smallest and largest number). Thus, we could approximate the case $X = 0$ with $Y \leq 0.5$, $X = 1$ with $0.5 < Y \leq 1.5$, ... up to the approximation of $X = 6$ with with $Y > 5.5$. Setting $n = 6, p = 0.6$, then $Z = \frac{Y-np}{\sqrt{np(1-p)}} = \frac{Y-3.6}{\sqrt{1.44}} = \frac{Y-3.6}{1.2}$.

$P(Y \leq 0.5) = P(Z \leq \frac{0.5-3.6}{1.2}) = P(Z \leq -2.5833) = 0.049$

$P(0.5 < Y \leq 1.5) = P(-2.5833 < Z \leq -1.75) = 0.0352$, ...

x_j	0	1	2	3	4	5	6
$f_X(x_j)$	0.0049	0.0369	0.1382	0.2765	0.3110	0.1866	0.0467
y	$(-\infty, 0.5]$	$(0.5, 1.5]$	$(1.5, 2.5]$	$(2.5, 3.5]$	$(3.5, 4.5]$	$(4.5, 5.5]$	$(5.5, \infty)$
$f_X^*(x_j)$	0.0049	0.0352	0.1396	0.2871	0.3066	0.1700	0.0567

($f_X(x_j) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$, $f_X^*(x_j)$ is obtained by the approximation).

Section 6
Sampling distributions

- Objectives: to discuss the idea that statistics that are calculated from samples have distributions
- to examine such distributions for important combination of statistics and parent population from which the sample is selected
- Topics: Sampling distributions, some results for statistics from samples taken from normal population.
- References: Miller and Miller, Chapter 8; Wooldridge, Appendix C (Section C.1).

Sampling

Suppose we want to know the average income in town. We could go and ask each person, and compute the average, but this is not really feasible, because we cannot really ask hundreds of thousands of people (there is also another, and more important factor to take into account, and this is the fact that people in town change as well, day by day, but for the moment we neglect it).

As an alternative, consider asking the first 100 people we meet in the street, and then take the sample mean: this number is not the number we would like (the average income of the whole town) but it is fair to conjecture that it is actually informative about that number. The number we compute asking 100 people in the street may be higher, as it is if we meet many people wealthier than the average, and it may be lower, if we meet people less affluent than the average.

The action of asking 100 people is "sampling".

We now cast the problem in statistical form.

We are interested in some characteristics of a population: an observation X from this population is a random with distribution $F_X(x;\theta)$ (the population distribution), and it is assumed that we do not know θ and want to find out about it. In the example of the income in town, θ may be the mean and we want to estimate it.

Let $X_1,\ldots,X_n$ be n observations with common distribution $F_X(x;\theta)$ (e.g., when we enquire about the income with 100 individuals in town): $X_1,\ldots,X_n$ is a sample of dimension n, and the joint distribution of $X_1,\ldots,X_n$ is the **Sampling distribution** of $X_1,\ldots,X_n$.

independent random sample

Now suppose that our survey is made so that the income of each individual is distributed independently from the income of the other individuals. Then, $X_1,\ldots,X_n$ is an independent random sample.

If X_i $(1 \leq i \leq n)$ is continuously distributed,

$$P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n) = \prod_{i=1}^{n} P(X_i \leq x_i)$$

If X_i $(1 \leq i \leq n)$ is discretely distributed,

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i).$$

Sample statistics

Usually, we are interested in functions

$$h(X_1, \ldots, X_n)$$

of our sample. Functions $h(X_1, \ldots, X_n)$ of the sample are called **sample statistics**.

Because $X_1, \ldots, X_n$ are random variables, $h(X_1, \ldots, X_n)$ is a random variable too, and $h(x_1, \ldots, x_n)$ is its realisation.

For example, one such function is the **sample average**, or **sample mean**

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

If X_i, $i = 1, \ldots, n$ are

- independently and
- identically distributed variables,
- with $E(X_i) = \mu_X$, $Var(X_i) = \sigma_X^2$,

then

$$E(\overline{X}) = \mu_X$$

$$Var(\overline{X}) = \frac{1}{n}\sigma_X^2.$$

The Standard Deviation of $\overline{X}$ is often called **standard error**, and it is indicated as $SE(\overline{X})$.

Notice that the two results for $E(\overline{X})$ and $Var(\overline{X})$ do not depend on the distribution of X_i (except for the fact that X_i is iid and $E(X_i^2) < \infty$).

The distribution of $\overline{X}$ however depends on X_i.

Some results from sampling from a normal distribution.

- If X_i are *iid* and $X_i \sim N(\mu, \sigma^2)$, then

$$\bar{X} \sim N\left(\mu, \frac{1}{n}\sigma^2\right)$$

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \chi^2_{n-1}$$

$$\bar{X} \text{ and } \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \text{ are independent.}$$

So, letting

$$Z_n = \sqrt{n}\,\frac{(\bar{X} - \mu)}{\sigma},\ C_{n-1} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2,\ T_{n-1} = \frac{Z_n}{\sqrt{C_{n-1}/(n-1)}}$$

$$Z_n \sim N(0,1),\ T_{n-1} \sim t_{n-1}.$$

Notice that T_{n-1} does not depend on σ^2.

Some results from sampling from an unknown distribution.

- If X_i are *iid* with $E(X_i) = \mu$, $Var(X_i) = \sigma^2$, then

$$\bar{X} \rightarrow_d N\left(\mu, \frac{1}{n}\sigma^2\right)$$

$$\bar{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \rightarrow_p \sigma^2.$$

So, letting

$$T_n = \frac{Z_n}{\sqrt{\bar{\sigma}^2}}$$

then

$$T_n \rightarrow_d N(0,1).$$

Section 7
Point estimation

- Objectives: to calculate a single number from a sample to obtain information on an unknown quantity
- to examine the criteria that can be applied to evaluate and choose between different approaches to estimation
- to provide some important results for finite samples and asymptotic behaviour
- Topics: finite sample properties of point estimators, asymptotic properties of point estimators
- References: Miller and Miller, Ch. 11; Wooldridge, Appendix C.

Estimation

Suppose we have a random sample $X_1, \ldots, X_n$ from a population, with distribution $F_{X_1,\ldots,X_n}\left(x_1, \ldots, x_n; (\theta_1, \ldots, \theta_p)'\right)$, in which the distribution is known, but the parameters are not. The true value of the unknown parameter is assumed to be in a set $\Theta \subset \mathbb{R}^p$.

Example

If X_i for $i = 1, \ldots, n$ is independently and identically distributed as Normal(μ, σ^2), then, $p = 2$, and $\theta_1 = \mu$, $\theta_2 = \sigma^2$.

If X_i may only have two outcomes, and X_i is iid with probability of success p, so that $\sum X_i$ is binomially distributed, then $p = \theta$.

An **estimator** is a *known function* of the random variables $X_1, \ldots, X_n$, denoted as

$$\widehat{\theta}(X_1, \ldots, X_n).$$

As such, an estimator is a random variable. For given set of realisations, $x_1, \ldots, x_n$, it is possible to compute an **estimate**

$$\widehat{\theta}(x_1, \ldots, x_n).$$

Both the notations $\widehat{\theta}(X_1, \ldots, X_n)$ and $\widehat{\theta}(x_1, \ldots, x_n)$ are routinely shortened to $\widehat{\theta}_n$: the difference between the random variable and the realisation should be clear from the context.

Example

In the example of tossing a coin, assume we had n independent trials, and for each i let $X_i = 1$ for head, $X_i = 0$ for tail. Let θ be the probability of observing head: a possible estimator for θ could be

$$\widehat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Suppose we had 7 trials, and we observed 1, 0, 1, 1, 1, 1, 0, then the estimate is $\widehat{\theta}_n = 0.71429$.

Unbiasedness

An estimator is a random variable, and as such it has a distribution. If

$$E\left(\widehat{\theta}_n\right) = \theta$$

then the estimator is said to be unbiased. Conversely, $E\left(\widehat{\theta}_n\right) - \theta$ is called bias.

Unbiasedness is often regarded as a desirable properties for estimators. Intuitively, an unbiased estimator should, on average, return the parameter θ. However, this is certainly not sufficient. In the example of estimating the probability of a tail when tossing a coin n times, consider another estimator: $\widetilde{\theta}_n = X_n$. Because $E(X_n) = \theta$, this estimator too is unbiased. However, $\widetilde{\theta}_n$ discards all the information in $X_1, \ldots, X_{n-1}$. We then need another property to complement unbiasedness, in order to help us choosing between $\widetilde{\theta}_n$ and $\widehat{\theta}_n$.

Relative efficiency

Let $\widetilde{\theta}_n$ and $\widehat{\theta}_n$ be two unbiased estimators of θ: $\widehat{\theta}_n$ is efficient relative to $\widetilde{\theta}_n$ if

$$Var\left(\widehat{\theta}_n\right) \leq Var\left(\widetilde{\theta}_n\right).$$

Intuitively, from the Chebyshev inequality, an efficient estimator clusters more probability in a suitable neighbourhood of θ. Further information is conveyed by the variances if the conditions for the Central Limit Theorem hold.

In the example in which two alternative estimators are used to estimate the probability of tossing a coin on the head, $Var\left(\widehat{\theta}_n\right) = \frac{1}{n}\theta(1-\theta)$, $Var\left(\widetilde{\theta}_n\right) = \theta(1-\theta)$, so $\widehat{\theta}_n$ is more efficient.

Best / Minimum Variance Estimator

If $\widehat{\theta}_n$ is an unbiased estimator, and no other estimator has a smaller variance, then $\widehat{\theta}_n$ is the "best" (minimum variance) estimator. Often denoted as "$\widehat{\theta}_n$ is MVUE".

Linear estimator. If $\widehat{\theta}_n$ is a linear function of $X_1, \ldots, X_n$, then it is a linear estimator ($\widehat{\theta}_n = \sum_{i=1}^{n} a_i X_i$ for some known a_i).

Best Linear Estimator

If $\widehat{\theta}_n$ is an unbiased linear estimator, and no other linear estimator has a smaller variance, then $\widehat{\theta}_n$ is the "best linear" (minimum variance) estimator. Often denoted as "$\widehat{\theta}_n$ is BLUE".

Example.

Let X_i be *iid* and $X_i \sim N(\theta, \sigma^2)$, and consider $n = 2$, and $\widetilde{\theta}_n = a_1 X_1 + a_2 X_2$, so

$$E\left(\widetilde{\theta}_n\right) = E(a_1 X_1 + a_2 X_2) = a_1 E(X_1) + a_2 E(X_2) = a_1\theta + a_2\theta.$$

If $a_1 + a_2 = 1$, then $E\left(\widetilde{\theta}_n\right) = \theta$. Let $\widehat{\theta}_n = aX_1 + (1-a)X_2$ for $0 \leq a \leq 1$: $\widehat{\theta}_n$ is unbiased. Furthermore,

$$Var\left(\widehat{\theta}_n\right) = a^2\sigma^2 + (1-a)^2\sigma^2 = \sigma^2(1 - 2a + 2a^2).$$

In order to minimise the variance, compute

$$\frac{\partial(1 - 2a + 2a^2)}{\partial a} = 4a - 2$$

so from the first condition we derive

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

and $\frac{1}{2}X_1 + \frac{1}{2}X_2$ is the BLUE.

If n observations are used to estimate θ, $\frac{1}{n}\sum_{i=1}^{n} X_i$ is BLUE.

Mean Square Error

$$\begin{aligned} MSE\left(\hat{\theta}_n\right) &= E\left(\hat{\theta}_n - \theta\right)^2 \\ &= E\left(\hat{\theta}_n - E\left(\hat{\theta}_n\right)\right)^2 + \left(E\left(\hat{\theta}_n\right) - \theta\right)^2 \\ &= Var\left(\hat{\theta}_n\right) + bias^2 \end{aligned}$$

So, $MSE\left(\hat{\theta}_n\right) = Var\left(\hat{\theta}_n\right)$ for unbiased estimates.

MSE provides a criterion to compare estimates that may also be biased. This is because there may be estimates that are biased and yet cluster much information around the parameter of interest. (MSE weights bias and variance, although notice that the choice of how to weight them is arbitrary).

Example. Consider two estimators, $\tilde{\theta}_n$ and $\hat{\theta}_n$ such that

$$E\left(\tilde{\theta}_n\right) = \theta + \frac{1}{2},\ Var\left(\tilde{\theta}_n\right) = 1$$

$$E\left(\hat{\theta}_n\right) = \theta,\ Var\left(\hat{\theta}_n\right) = 2$$

then

$$MSE\left(\tilde{\theta}_n\right) = 1 + \left(\frac{1}{2}\right)^2 = 1.25$$

$$MSE\left(\hat{\theta}_n\right) = 2 + 0 = 2.$$

Asymptotic / Large Sample Properties

Sometimes we cannot obtain the properties of the estimators in finite samples, but it may be possible to describe them as n gets large.

Consistency

$\widehat{\theta}_n$ is a consistent estimator of θ if

$$p\lim_{n\to\infty}\widehat{\theta}_n = \theta.$$

Consistency and unbiasedness do not imply each other:

(counter)example.

Let X_i be *iid* and $X_i \sim N(\theta, \sigma^2)$, and consider

$$\widetilde{\theta}_n = X_n$$

$$\widehat{\theta}_n = \begin{cases} \theta \text{ with pr } 1 - \frac{1}{n} \\ n \text{ with pr } \frac{1}{n} \end{cases}$$

Then, $\widetilde{\theta}_n$ is unbiased, but it is not consistent, while $\widehat{\theta}_n$ is consistent but it is not unbiased ($\lim_{n\to\infty} E\left(\widehat{\theta}_n\right) = \lim_{n\to\infty} \theta - \frac{\theta}{n} + 1 = \theta + 1$).

However, if $\widehat{\theta}_n$ is unbiased and $Var\left(\widehat{\theta}_n\right) \to 0$ as $n \to \infty$, then $\widehat{\theta}_n$ is consistent (by the Chebyshev inequality). Also, notice that for consistency of $\widehat{\theta}_n$, it is sufficient that $MSE\left(\widehat{\theta}_n\right) \to 0$ as $n \to \infty$.

Example: Let X_i be *iid* and $X_i \sim N(\theta, \sigma^2)$ and consider $\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Since $E\left(\widehat{\theta}_n\right) = \theta$, $Var\left(\widehat{\theta}_n\right) = \frac{\sigma^2}{n}$, $MSE\left(\widehat{\theta}_n\right) \to 0$ as $n \to \infty$, so the sample mean is a consistent estimator of the population mean in this situation.

Slutzky's theorem

If $\widehat{\theta}_n$ is a consistent estimator of θ and $g(.)$ is a continuous function, then

$$g\left(\widehat{\theta}_n\right) \to_p g(\theta) \text{ as } n \to \infty.$$

Example. Let X_i be i.i.d. random variables with $E(X_i) = \mu$, $Var(X_i) = \sigma^2$ and $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$. Then, let $\widehat{\sigma} = \left(\widehat{\sigma^2}\right)^{1/2}$: by Slutzky's theorem, $p\lim_{n\to\infty}\widehat{\sigma} = \sigma$.

(NOTE: this property does not hold for the expectations. Example: let $\widehat{\theta}_n$ be such that $E\left(\widehat{\theta}_n\right) = \theta$ (and $Var\left(\widehat{\theta}_n\right) > 0$) and $p\lim_{n\to\infty}\widehat{\theta}_n = \theta$, and consider the function $g(\theta) = \theta^2$. Slutzky Theorem says that $p\lim_{n\to\infty}\widehat{\theta}_n^2 = \theta^2$; on the other hand, we know that $E\left(\widehat{\theta}_n^2\right) > \theta^2$: this follows by Jensen's inequality, or using property of the variance $Var\left(\widehat{\theta}_n\right) = E\left(\widehat{\theta}_n^2\right) - \left(E\left(\widehat{\theta}_n\right)\right)^2 > 0$).

Consistent and asymptotically normal (CAN) estimators

(definition for scalar $\widehat{\theta}_n$). If there is an estimator $\widehat{\theta}_n$ for θ which is

- consistent ($p\lim_{n\to\infty} \widehat{\theta}_n = \theta$)
- there is a scaling factor n^α and $V > 0$ such that

$$n^\alpha\left(\widehat{\theta}_n - \theta\right) \to_d N(0, V) \text{ as } n \to \infty$$

then $\widehat{\theta}_n$ is CAN.

The definition can be extended to vector estimators.

It is often the case that $\alpha = 1/2$.

Example: Let X_i be i.i.d. random variables with $E(X_i) = \mu$, $Var(X_i) = \sigma^2$ and $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Then, by the Central Limit Theorem, $\sqrt{n}\,(\overline{X} - \mu)\sigma^{-1} \to_d N(0,1)$ as $n \to \infty$. Therefore, $\overline{X}$ is a CAN estimator of θ, with scaling factor $n^{1/2}$.

Asymptotic efficiency

Consistency and limit normality are desirable properties for an estimator. However, how can we choose among two CAN estimators?

1. Compare the rate of convergence.
2. Compare the asymptotic variances.

Best Asymptotic Normal (BAN)

If $\widehat{\theta}_n$ is CAN, and there is no other CAN estimator having higher rate of convergence or smaller asymptotic variance, then $\widehat{\theta}_n$ is BAN.

Section 8
Maximum likelihood estimation

- Objectives: to provide an introduction to a very general approach to the estimation of parameters
- References Miller and Miller, Section 10.8; Wooldridge, Appendix C pp. 746-747.

Maximum likelihood estimation

The Maximum likelihood principle is based upon a reinterpretation of the joint *pdf*/probability mass function for a sample (of size n).

The $pdf\, f_{X_1,\ldots,X_n}\left(x_1,\ldots,x_n;(\theta_1,\ldots,\theta_p)'\right)$ and the probability mass function $f_{X_1,\ldots,X_n}\left(x_1,\ldots,x_n;(\theta_1,\ldots,\theta_p)'\right)$ are functions computed in each point of the support, $x_1,\ldots,x_n$, for a given set of parameters: for example, when $X_1,\ldots,X_n$ are iid normally distributed with $E(X_i)=0$, $Var(X_i)=1$, then the *pdf* is $\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x_i^2}$ for $-\infty<x_i<\infty$; when $X=\sum_i^n X_i$ is binomially distributed with parameter 1/2, the probability mass function is $\frac{n!}{x!(n-x)!}1/2^x(1-1/2)^{n-x}$ for $x=\sum_i^n x_i$, $x=0,\ldots,n$.

In maximum likelihood on the other hand these are treated as unknown functions of $(\theta_1, \ldots, \theta_p)'$, and are computed for given values of the realisations $x_1, \ldots, x_n$. For each set of parameters $(\theta_1, \ldots, \theta_p)'$, we denote the likelihood function by $L(\theta)$.

The Maximum Likelihood Estimator (MLE) is the value that maximises $L(\theta)$ or, equivalently, $l(\theta) = \ln(L(\theta))$, over a suitable parameter space Θ, i.e.

$$\widehat{\theta}_n = \arg\max_{\theta \in \Theta} L(\theta).$$

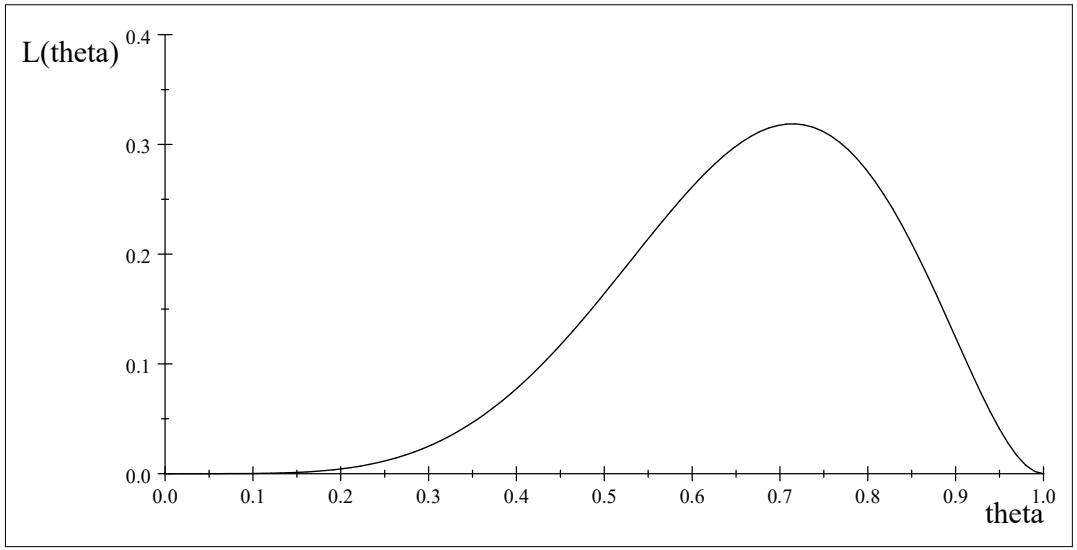
(We may drop n to shorten the notation and use $\widehat{\theta}$ instead).

In the example of estimating the probability of tossing head, when head was observed in 5 out 7 independent trials,

$$\text{if } p = 0.25,\ P(X = 5) = \frac{7!}{5!2!}0.25^5(1-0.25)^2 = 0.011$$
$$\text{if } p = 0.5,\ P(X = 5) = \frac{7!}{5!2!}0.5^5(1-0.5)^2 = 0.164$$
$$\text{if } p = 0.75,\ P(X = 5) = \frac{7!}{5!2!}0.75^5(1-0.75)^2 = 0.311$$

For the generic $p = \theta$,

$$L(\theta) = \frac{7!}{5!2!}\theta^5(1-\theta)^2, 0 \le \theta \le 1,$$



(in this example, the estimated value is $\widehat{\theta} = \frac{5}{7}$).

Example.

Let X_i be i.i.d. $N(\theta_0, 1)$, and assume that $x_1 = 2, x_2 = 3, x_3 = -1, x_4 = 0$ were observed. The likelihood function is then

$$L(\theta) = \left(\frac{1}{\sqrt{2\pi}}\right)^4 e^{-\frac{1}{2}\left[(2-\theta)^2+(3-\theta)^2+(-1-\theta)^2+(\theta)^2\right]}$$

and, taking logarithm,

$$-4/2\ln(2\pi) - \frac{1}{2}\left((2-\theta)^2 + (3-\theta)^2 + (-1-\theta)^2 + (\theta)^2\right)$$

$$= -2\ln(2\pi) + 4\theta - 2\theta^2 - 7.$$

Solving for θ,

$$\frac{\partial(-2\ln(2\pi) + 4\theta - 2\theta^2 - 7)}{\partial\theta} = 4 - 4\theta, \widehat{\theta} = 1.$$

Examples.

In the example of estimating the probability of tossing a head, when a generic x heads is observed,

$$l(\theta) = \ln \frac{n!}{x!n-x!} + x \ln\theta + (n-x)\ln(1-\theta)$$

$$\frac{\partial l(\theta)}{\partial \theta} = \frac{x}{\theta} + \frac{1}{\theta - 1}(n-x),\ \widehat{\theta} = \frac{x}{n}.$$

In the example of estimating the mean of n i.i.d. $N(\theta_0, \sigma^2)$ distributions, when a generic $x_1, \ldots, x_n$ is observed,

$$l(\theta) = -n/2 \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta)^2$$

$$\frac{\partial l(\theta)}{\partial \theta} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(-2)(x_i - \theta),\ \widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

While the likelihood function is computed using the observations, these are realisations of random variables, so changing the realisation of the sample $x_1, \ldots, x_n$, we obtain a different estimate. The functional form, however, does not change: the **estimator**, is, then, a function of the random variables, and a random variable itself; the **estimate** is the value that we compute with observations $x_1, \ldots, x_n$. For example, in both previous cases the estimator is the sample mean, $\widehat{\theta} = \frac{1}{n} \sum_{i=1}^{n} X_i$.

We only introduced maximum likelihood using an intuitive argument, but we did not show if it is consistent nor any other properties. Under weak conditions, however, it may be shown that the maximum likelihood estimator is indeed consistent, and in fact also best asymptotically normal. Moreover, if $\widehat{\theta}_n$ is a maximum likelihood estimator and $g(.)$ is a continuous function, then $g\left(\widehat{\theta}_n\right)$ is the MLE estimator of $g(\theta)$ (invariance property).

In many cases the MLE estimator is biased.

Example. Let $X_1,\ldots,X_n$ be i.i.d., $N(\mu_0,\sigma_0^2)$. The maximum likelihood estimator of σ_0^2 is obtained from

$$\begin{cases} \frac{\partial l\left((\mu,\sigma^2)'\right)}{\partial \mu} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(-2)(x_i-\mu), \\ \frac{\partial l\left((\mu,\sigma^2)'\right)}{\partial \sigma^2} = -\frac{1}{2}\frac{n}{\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2, \end{cases}$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}, \quad \widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2,$$

and notice that $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$ is a biased estimator of σ^2 (the unbiased estimator is $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$).

Other methods of estimation.

Maximum Likelihood (ML) gives one approach to estimating parameters. It is not the only approach. Two other popular approaches are Least Squares (LS) and Method of Moments (MM). These approaches have the advantage of often being computationally simpler than ML

Method of Moments

In the MM we use the sample moments to estimate the population moments: for example, for $X_i \ \ iid(\mu, \sigma^2)$, we could estimate μ and σ^2 as $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i, \widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu})^2$ (Note that these are the same estimators that we would compute if we knew that X_i is normally distributed. In general, however, the MM does not necessarily yield the same formula as ML).

Least Squares (LS)

In LS we estimate parameters minimizing an ad-hoc loss function. For example, for $X_i \ \ iid(\mu, \sigma^2)$ we estimate μ minimizing

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2.$$

This yields again $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$, but notice that again we did not assumed normality in this case. In view of the similarity between the loss function and the Gaussian likelihood, the LS is usually similar to the ML estimator if normality is assumed. However, LS gives a way to compute an estimator even if normality is not assumed.

LS is the most commonly used estimator in econometrics.

Section 9
Interval estimation

- Objectives: to discuss how interval estimates may be used to provide an idea on the strength of the information in a sample for estimating a parameter
- Topics: Confidence intervals
- References: Miller and Miller, Ch. 11; Wooldridge, Appendix C (Section C.5).

Interval estimation

May wish to give not just a point estimate but also an interval estimate that helps to reflect the precision of the estimation.

Let $C_1(X_1,\ldots,X_n)$ and $C_2(X_1,\ldots,X_n)$ be two random functions such that

- $C_2(X_1,\ldots,X_n) > C_1(X_1,\ldots,X_n)$ for any $X_1,\ldots,X_n$
- $P(C_1(X_1,\ldots,X_n) \leq \theta \leq C_2(X_1,\ldots,X_n)) = 1-\gamma$
  i.e. the probability that the random interval from C_1 to C_2 includes θ is $1-\gamma$.

then the random interval $[C_1(X_1,\ldots,X_n), C_2(X_1,\ldots,X_n)]$ is a *confidence interval estimator* of θ, for a confidence coefficient of $1-\gamma$.

It is often used $\gamma = 0.05$, so $1-\gamma = 0.95$. Sometimes confidence intervals are expressed in percentage terms, "$(1-\gamma)100$ per cent".

Because $C_1(X_1,\ldots,X_n)$ and $C_2(X_1,\ldots,X_n)$ are a function of $X_1,\ldots,X_n$, then $C_1(X_1,\ldots,X_n)$ and $C_2(X_1,\ldots,X_n)$ are random variables, so $[C_1(X_1,\ldots,X_n), C_2(X_1,\ldots,X_n)]$ is a random interval.

The true value of θ is an unknown constant, and it may, or may not, be in the interval $[C_1(X_1,\ldots,X_n), C_2(X_1,\ldots,X_n)]$.

The definition states that the probability that θ is in the confidence interval is $1-\gamma$. However, notice that we do not have the interval $[C_1(X_1,\ldots,X_n), C_2(X_1,\ldots,X_n)]$: at most, we can compute an estimate of it, by evaluating the functions $C_1(X_1,\ldots,X_n)$ and $C_2(X_1,\ldots,X_n)$ at for the sample $x_1,\ldots,x_n$. Letting

$$c_1 = C_1(x_1,\ldots,x_n),\, c_2 = C_2(x_1,\ldots,x_n)$$

be the realisation of the confidence interval in a given sample, then (c_1, c_2) is a *confidence interval estimate* (notice that this changes with the realisation of the sample).

Notice that more than one confidence interval may meet the conditions stated in the definition: for example, when estimating μ for a sample $X_1,\ldots,X_n$ of independent observations, identically distributed as a $N(\mu,1)$, then both the following intervals meet the definition ($\gamma = 0.05$):

$$C_1(X_1,\ldots,X_n) = -\infty, \qquad C_2(X_1,\ldots,X_n) = \frac{1}{\sqrt{n}}\overline{X} + 1.65$$

$$C_1(X_1,\ldots,X_n) = \frac{1}{\sqrt{n}}\overline{X} - 1.96, \; C_2(X_1,\ldots,X_n) = \frac{1}{\sqrt{n}}\overline{X} + 1.96,$$

i.e. $\left(-\infty, \frac{1}{\sqrt{n}}\overline{X} + 1.65\right]$ and $\left[\frac{1}{\sqrt{n}}\overline{X} - 1.96, \frac{1}{\sqrt{n}}\overline{X} + 1.96\right]$. How do we choose between these two (and the infinite other ones that still meet the definition in this example)? Unless more information is available, the standard criterion is to choose the interval with smaller measure of the subset of $(-\infty,+\infty)$. Since $\frac{1}{\sqrt{n}}\overline{X} + 1.65 - (-\infty) = \infty$ and $\frac{1}{\sqrt{n}}\overline{X} + 1.96 - \left(\frac{1}{\sqrt{n}}\overline{X} - 1.96\right) = 3.92$, we prefer the second one.

Example.

Interval estimation of the mean of a normal distribution.

Let $X_1, \ldots, X_n$ be a random sample from $N(\theta, \sigma^2)$, and let $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$.

- Suppose first that σ^2 is known. Then, the $(1-\gamma)100\%$ confidence interval estimator is given by $\overline{X} \pm d(\sigma/\sqrt{n})$, where d meets $P(Z > d) = \gamma/2$, for $Z \sim N(0,1)$.
- If σ^2 is unknown, then the $(1-\gamma)100\%$ confidence interval estimator is given by $\overline{X} \pm d(S/\sqrt{n})$, where d meets $P(T_{n-1} > d) = \gamma/2$, for $T_{n-1} \sim t_{n-1}$.

Let d_Z such that $P(Z > d_Z) = \gamma/2$, for $Z \sim N(0,1)$, and let d_{n-1} such that $P(T_{n-1} > d_{n-1}) = \gamma/2$, for $T_{n-1} \sim t_{n-1}$. Notice that

- the functions $d_Z(\sigma/\sqrt{n})$ and $d_{n-1}(s/\sqrt{n})$ shrink and collapse to 0 as $n \to \infty$: this reflects the increase of information that is accrued by increasing the sample size, and it is consistent with the fact that $\overline{X}$ is a Mean Squared consistent estimator of θ.
- the intervals are always random, because $\overline{X}$ is unknown. However if σ^2 is unknown, then the estimation of σ^2 includes another element of randomness, and the measure of the interval is random too. This is reflected in the fact that $d_{n-1} > d_Z$.
- if, by chance, $s = \sigma$, the confidence interval for the case in which σ is known is still smaller than the confidence interval for the case in which it is estimated. This is because the estimation of σ^2 introduces another element of randomness.
- when the sample increases, the precision of S as an estimate of σ increases (indeed, $P(|S - \sigma| > \varepsilon) \to 0$ as $n \to \infty$), so $d_{n-1} \to d_Z$.

Interval estimation of the mean of a normal distribution

Example

Suppose that the times each student spends at a canteen at the University, X_i, are independently and normally distributed, with mean μ and variance $\sigma^2 = 8^2$. A random sample of 25 had an average time $\bar{x} = 16$ minutes.

Find a 95% confidence interval for the population mean μ.

What would your answer be, if we observed $\bar{x} = 15$?

What if the sample had 64 observations instead?

What if σ^2 was unknown, and $s^2 = 8^2$ was estimated instead?

We discuss the first example.

We know that $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$, i.e.

$$\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0,1)$$

The confidence interval is obtained solving

$$P\left(\left| \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \right| \leq 1.96 \right) = 0.95$$

(verify that this matches the definition: for a 95% confidence interval, we find d from $P(Z > d) = 0.05/2$ where $Z \sim N(0,1)$, so indeed $d = 1.96$).

We can rewrite

$$P\left(\left| \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \right| \leq 1.96 \right) = 0.95$$

as

$$P\left(-1.96 \leq \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \leq 1.96 \right) = 0.95$$

$$P\left(-1.96\sqrt{\sigma^2/n} \leq \bar{X} - \mu \leq 1.96\sqrt{\sigma^2/n} \right) = 0.95$$

$$P\left(-\bar{X} - 1.96\sqrt{\sigma^2/n} \leq -\mu \leq -\bar{X} + 1.96\sqrt{\sigma^2/n} \right) = 0.95$$

$$P\left(\bar{X} - 1.96\sqrt{\sigma^2/n} \leq \mu \leq \bar{X} + 1.96\sqrt{\sigma^2/n} \right) = 0.95.$$

Replacing $\sigma = 8$, $n = 25$, the 95% confidence interval is $\left[\bar{X} - 1.96\sqrt{8^2/25}\,, \bar{X} + 1.96\sqrt{8^2/25} \right]$, and its estimate is $\left[12.864, 19.136 \right]$

★What would your answer be, if we observed $\bar{x} = 15$?

$$\left[15 - 1.96 \times \sqrt{\frac{8^2}{25}}, 15 + 1.96 \times \sqrt{\frac{8^2}{25}} \right] = [11.864, 18.136]$$

★What if the sample had 64 observations instead?

$$\left[16 - 1.96 \times \sqrt{\frac{8^2}{64}}, 16 + 1.96 \times \sqrt{\frac{8^2}{64}} \right] = [14.04, 17.96]$$

★What if σ^2 was unknown, and $s^2 = 8^2$ was estimated instead?

$$\frac{\bar{X} - \mu}{\sqrt{S^2/n}} \sim T_{n-1},$$

so when $n = 25$ we have to find d such that $P(|T_{24}| > d) = 0.05$, and therefore $d = 2.07$.

The interval is computed solving $P\left(\left| \frac{\bar{X} - \mu}{\sqrt{S^2/n}} \right| \leq 2.07 \right) = 0.95$

So

$$P\left(-2.07 \leq \frac{\overline{X}-\mu}{\sqrt{S^2/n}} \leq 2.07\right) = 0.95$$

$$P\left(-2.07 \times \sqrt{S^2/n} \leq \overline{X}-\mu \leq 2.07 \times \sqrt{S^2/n}\right) = 0.95$$

$$P\left(\overline{X}-2.07 \times \sqrt{S^2/n} \leq \mu \leq \overline{X}+2.07 \times \sqrt{S^2/n}\right) = 0.95$$

so our confidence interval is

$$\left[\overline{X}-2.07 \times \sqrt{S^2/25} \leq \mu \leq \overline{X}+2.07 \times \sqrt{S^2/25}\right]$$

Our estimate involves replacing $\overline{X}$ by $\bar{x}$, S^2 by s^2, and it is then

$$\left[16-2.07 \times \sqrt{\frac{8^2}{25}}, 16+2.07 \times \sqrt{\frac{8^2}{25}}\right] = [12.688, 19.312]$$

★ When $n = 64$, $P(|T_{63}| > d) = 0.05$ for $d = 2.00$, and our estimate is

$$\left[16-2.00 \times \sqrt{\frac{8^2}{64}}, 16+2.00 \times \sqrt{\frac{8^2}{64}}\right] = [14, 18]$$

Interval estimation of the mean of a generic distribution

In case the distribution of X_i is not known, we can only rely on asymptotic results.

Let $X_1, \ldots, X_n$ be a random sample of independent, identically distributed observations with $E(X_i) = \theta$, $Var(X_i) = \sigma^2$, and let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$.

- If σ^2 is known, then an asymptotically valid $(1-\gamma)100\%$ confidence interval estimator is given by $\overline{X} \pm d(\sigma/\sqrt{n})$, where d meets $P(Z > d) = \gamma/2$, for $Z \sim N(0,1)$.
- If σ^2 is unknown, then an asymptotically valid $(1-\gamma)100\%$ confidence interval estimator is given by $\overline{X} \pm d(S/\sqrt{n})$, where d meets $P(Z > d) = \gamma/2$, for $Z \sim N(0,1)$.

Example

The Library is interested in how many requests are made for a book in a certain reading list. The participants to the each course may be treated as a random sample of students. In the last year, 144 students were enrolled: the average request (in days) per student was $\bar{x} = 19$, and $s = 9$. Find an asymptotically valid 90% confidence interval estimate.

Discussion. Find d such that $P(|Z| > d) = 0.10$, so $d = 1.65$; proceeding as in previous examples, the approximate confidence interval is

$$\left[\bar{X} - 1.65 \times \sqrt{\frac{S^2}{144}}, \bar{X} + 1.65 \times \sqrt{\frac{S^2}{144}} \right]$$

and the estimate is

$$\left[19 - 1.65 \times \sqrt{\frac{9^2}{144}}, 19 + 1.65 \times \sqrt{\frac{9^2}{144}} \right] = [17.763, 20.238]$$

Interval estimation of the probability of success in a binomial

Let X denote the number of successes in n trials with $P(success) = \theta$. Then, an asymptotically valid $(1-\gamma)100\%$ confidence interval estimator is given by

$$\hat{\theta} \pm d\sqrt{\hat{\theta}\left(1-\hat{\theta}\right)/n} \text{ where } \hat{\theta} = \frac{X}{n}$$

d meets $P(Z > d) = \gamma/2$, for $Z \sim N(0,1)$.

★Example

A city is considering the introduction of a scheme of congestion charges to reduce pollution. A random sample of 414 citizen has been asked his opinion: 124 of these citizens are in favour. Find an asympotically valid 90% confidence interval estimate for the population proportion of voters in favour of the scheme.

Can only obtain an approximate interval. Find d such that $P(|Z| > d) = 0.10$, so $d = 1.65$, and $\widehat{\theta} = \frac{124}{414} = 0.29952$. So

$$\widehat{\theta} \pm d\sqrt{\widehat{\theta}\left(1-\widehat{\theta}\right)/n}$$

takes value

$$\frac{124}{414} \pm 1.65\sqrt{\frac{\frac{124}{414}(1-\frac{124}{414})}{414}}$$

and the estimate is therefore $[0.26237, 0.33666]$.

Section 10
Hypotheses testing

- Objectives: an introduction about the key ideas that are used when testing hypotheses about unknown coefficients, with applications.
- Topics: Statistical hypotheses null and alternative hypotheses, test statistics, critical regions and the size and power of a test; P-value.
- References: Miller and Miller, Ch. 12-13; Wooldridge, Appendix C (Section C.6).

Hypotheses testing

Suppose a probabilistic model has been specified, in which the joint distribution for a random sample for n observations depends on a set of unknown parameters $\theta_1, \ldots, \theta_p$.

We often have hypotheses concerning the value of some (possibly all) of the parameters. The hypothesis to be tested is called *null hypothesis* and it often denoted by H_0.

If H_0 is not true, then some alternatives will be true. When carrying out statistical tests we also have to formulate an *alternative hypothesis*, often denoted by H_1, against which the null is tested.

A *simple hypothesis*, null or alternative, completely specifies the joint *pdf*, by fixing the value of every parameter: for example, we may test the null $H_0 : X \sim N(0,1)$ versus a simple alternative $H_1 : X \sim N(1,1)$.

A *composite hypothesis*, null or alternative, does not completely specify the joint *pdf*, by fixing the value of every parameter: for example, we may have a simple null $H_0 : X \sim N(0,1)$ versus a composite alternative

$H_1 : X \sim N(\mu, 1)$, $-\infty < \mu < \infty$, or a composite null $H_0 : X \sim N(0, \sigma^2)$ versus a composite alternative $H_1 : X \sim N(\mu, \sigma^2)$, $-\infty < \mu < \infty$, $0 < \sigma^2 < \infty$.
In case of a composite alternative hypothesis, it is also possible that some additional information is available, and it is then taken into account in the definition of the hypothesis: suppose that H_0 specify the value of one single parameter, and it is written as $H_0 : \theta_i = \theta_i^0$, or, equivalently,
$H_0 : \theta_i - \theta_i^0 = 0$, for some i and for θ_i^0 being some constant. We can have *One Sided Alternative*, such as

$$\text{either } H_1^- : \theta_i - \theta_i^0 < 0 \text{ or } H_1^+ : \theta_i - \theta_i^0 > 0$$

or *Two Sided Alternative*,

$$H_1^{\pm} : \theta_i - \theta_i^0 \neq 0.$$

In order to determine wether not H_0 is consistent with the sample data, we must specify a test statistic and a decision rule.

A *Test Statistic*, $T(X_1, \ldots, X_n)$, is a function of the sample such that

- does not depend on any unknown parameter
- has known distribution under H_0 (and H_1), at least asymptotically.

Sometimes, we may use T to shorten the notation $T(X_1, \ldots, X_n)$, when this is clear from the context.

Given the test statistic, we can specify the *Decision rule*. Divide the sample space (all possible sets of observations) into two regions: a *rejection region* and a *non-rejection region* (the latter is also known as *acceptance region*). We then *Reject* H_0 if the realisation of the test statistic, t, falls in the rejection region, and do not reject H_0 if it falls in the non-rejection rule.

Values of t (the realisation of the test statistic T) that are taken to indicate that H_0 is not inconsistent with the data are termed as *Statistically Insignificant*, or, more appropriately, statistically insignificantly different than stated in the null hypothesis.

Notice that Reject does not mean that H_0 is false (nor that H_1 is true), and non rejection / acceptance of H_0 does not mean that H_0 is true (nor that H_1 is false).

This is obviously the case when the true model is in fact neither the one specified in H_0 nor the one specified in H_1 (for example, when $H_0 : X \sim N(0,1)$ and $H_1 : X \sim N(1,1)$, while in fact $X \sim N(0.5,1)$), but notice that it may even happen that a true hypothesis is correctly specified, and yet it is rejected by the test.

In other words, even when the hypothesis of interest is correctly specified (either under the null or the alternative) in each test two types of errors are nevertheless possible:

- a *Type I error* occurs when H_0 is rejected when H_0 is in fact true
- a *Type II error* occurs when H_0 is not rejected when H_0 is in fact false.

The probability of a Type I error is usually denoted by α and it is the *significance level* of the test (this is also known as the size of the test); the probability of a Type II error is usually denoted by β. The probability of rejecting a false H_0 is then $1 - \beta$: this is also known as the *Power* of a test. Usually, we decide on the level α, e.g. $\alpha = 0.05$ or $\alpha = 0.10$, and then choose a test with good power against the specified alternative hypothesis.

Summarising, a test is characterised by

1. the null hypothesis
2. the alternative hypothesis
3. the test statistic
4. the limit distribution of the test statistic under the null
5. the decision rule
6. the size of the test
7. the limit distribution of the test statistic under the alternative
8. the power of the test

   In practical applications, we should also specify

9. the realisation of the test statistic
10. wether the null hypothesis is rejected or not.

Example.

In order to verify if a coin is fair, we run the following test.

Toss the coin 10 times. Consider the coin fair if the number of heads is between 2 and 8 (included). Otherwise, consider the coin not fair, i.e. consider the coin not fair if the number of heads is 1 or less, or if the number of heads is 9 or more.

Let θ be $P(head)$ for each toss, X be the number of heads in 10 tosses

Then:

1. null hypothesis (... "consider the coin fair if"...)

$$H_0 : \theta = 1/2$$

2. the alternative hypothesis ("Otherwise, consider the coin not fair")

$$H_1 : \theta \neq 1/2$$

3. the test statistic

$$X$$

4. the limit distribution of the test statistic under the null

$$X \sim \frac{10!}{x!(10-x)!}\theta^x(1-\theta)^{10-x}, \theta = 1/2$$

5. the decision rule

$$Reject\ if\ x \leq 1\ or\ if\ x \geq 9$$

6. the size of the test

$$P((X=0,\theta=0.5) \cup (X=1,\theta=0.5) \cup (X=9,\theta=0.5) \cup (X=10,\theta=0.5$$

$$= P(X=0,\theta=0.5) + P(X=1,\theta=0.5) + P(X=9,\theta=0.5) + P(X=10,\theta=$$

$$= 0.0010 + 0.0100 + 0.0100 + 0.0010 = 0.0215$$

7. the limit distribution of the test statistic under the alternative

$$X \sim \frac{10!}{x!(10-x)!}\theta^x(1-\theta)^{10-x}, \theta \neq 1/2$$

8. the power of the test
We should compute $P(X \leq 1 \cup X \geq 9)$ for $\theta \neq 1/2$. There are infinite values for this, so we only compute it for a few possible parameters. For example, when $\theta = 0.6$, $P(X = 0, \theta = 0.6) = 0.0001$, $P(X = 1, \theta = 0.6) = 0.0016$, $P(X = 9, \theta = 0.6) = 0.0403$, $P(X = 10, \theta = 0.6) = 0.0060$. Overall,

θ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
size					0.021				
Power	0.736	0.376	0.149	0.048		0.048	0.149	0.376	0.736

Notice that the power increases as we move away from $\theta = 0.5$ (recall H_0): ideally we would like the power to go to 1 for each $\theta \neq 0.5$. If this happens at least as $n \to \infty$, then the test is said to be *consistent*.

Example.

Testing of a mean of a normal distribution (with known variance).

Let $X_1, \ldots, X_n$ be independent, $X_i \sim N(\mu, \sigma^2)$ $(i = 1, \ldots, n)$ for known σ^2.

We are interested in $H_0 : \{\mu = \mu_0\}$ where μ_0 is a known constant. Let

$$T = \sqrt{n}\,\frac{(\bar{X} - \mu_0)}{\sigma}$$

under H_0, $T \sim N(0, 1)$, so T is a valid test statistic. The decision rule depends on the type of alternative.

For $H_1 = H_1^+ : \{\mu > \mu_0\}$, the rejection rule is "Reject H_0 if $t > d_1$", where t is the realisation of T, and d_1 is the solution of $P(Z > d_1) = \alpha$, where α is the significance level, and Z is a $N(0,1)$ random variable.

For $H_1 = H_1^- : \{\mu < \mu_0\}$, the rejection rule is "Reject H_0 if $t < -d_1$", where t, d_1 and α are defined as above.

For $H_1 = H_1^\pm : \{\mu \neq \mu_0\}$, the rejection rule is "Reject H_0 if $|t| > d_2$", where t and α are defined as above, d_2 is the solution of $P(Z > d_2) = \alpha/2$, and Z is a $N(0,1)$ random variable.

Example.

The Railway Regulation Authority is revising the performance of the Fast Tortoises Railway Company, which is currently running the railway franchise: in the contract it was approved that the length of a regular journey is normally distributed with mean of 2 hours and standard deviation of 0.9. In the last 25 journeys, the average journey time was 2.4 hours. Should the franchise be renewed?

We first check if the problem is correctly specified, and then we address the question about renewing the franchise.

We are told to assume that $X_1, \ldots, X_n$ be independent, $X_i \sim N(\mu, \sigma^2)$ $(i = 1, \ldots, n)$ for $n = 25$, $\sigma^2 = 0.9^2$. We are interested in $H_0 : \{\mu = 2\}$.

The test statistic is

$$T = \sqrt{25}\,\frac{(\bar{X} - 2)}{0.9} \sim N(0,1),$$

under H_0. We are not told which type of alternative to take, nor the significance value. Assuming that the customers would not mind if the journey is shorter, we take the alternative

$$H_1 : \{\mu > 2\},$$

We also assume that $\alpha = 0.05$, so the critical value is 1.65. So,

1. null hypothesis

$$H_0 : \mu = 2$$

2. the alternative hypothesis

$$H_1 : \mu > 2$$

3. the test statistic

$$T = \sqrt{25}\,\frac{(\bar{X} - 2)}{0.9}$$

4. the limit distribution of the test statistic under the null

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

5. the decision rule

$$Reject\ if\ t \geq 1.65$$

6. the size of the test

$$\alpha = 0.05$$

7. the limit distribution of the test statistic under the alternative

$$T = \sqrt{25}\,\frac{(\bar{X}-2)}{0.9} = \sqrt{25}\,\frac{(\bar{X}-2-\mu+\mu)}{0.9}$$

$$= \sqrt{25}\,\frac{(\bar{X}-\mu)}{0.9} + \sqrt{25}\,\frac{(\mu-2)}{0.9}$$

$$\text{so } T = Z + \sqrt{25}\,\frac{(\mu-2)}{0.9} \text{ where } Z \sim N(0,1)$$

8. the power of the test
We should compute $P\left(Z + \sqrt{25}\,\frac{(\mu-2)}{0.9} \geq 1.65\right)$ for $\mu > 2$. There are infinite values for this, so we only compute it for a few possible parameters. For $\mu = 2.1$, $1.65 - \sqrt{25}\,\frac{(2.1-2)}{0.9} = 1.0944$, power $1 - 0.86 = 0.14$.

μ	2	2.1	2.2	2.3	2.4	2.5	2.6	2.7
size	0.05							
Power		0.14	0.30	0.51	0.72	0.87	0.95	0.99

9. the realisation of the test statistic

$$t = \sqrt{25}\,\frac{(\bar{x} - 2)}{0.9} = \sqrt{25}\,\frac{(2.4 - 2)}{0.9} = 2.22$$

10. wether the null hypothesis is rejected or not.
Since $t > 1.65$, the realisation of the test is in the rejection area so H_0 is rejected. The Fast Tortoise will lose the franchise.

Example.

Testing of a mean of a normal distribution with unknown variance.

Let $X_1, \ldots, X_n$ be independent, $X_i \sim N(\mu, \sigma^2)$ $(i = 1, \ldots, n)$.

We are interested in $H_0 : \{\mu = \mu_0\}$ where μ_0 is a known constant. We do not know σ^2, but we estimated $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$. Let

$$T = \sqrt{n} \frac{(\bar{X} - \mu_0)}{S}$$

under H_0, $T \sim t_{n-1}$, so T is a valid test statistic. The decision rule depends on the type of alternative.

For $H_1 = H_1^+ : \{\mu > \mu_0\}$, the rejection rule is "Reject H_0 if $t > d_1$", where t is the realisation of T, and d_1 is the solution of $P(T_{n-1} > d_1) = \alpha$, where α is the significance level, and T_{n-1} is a t_{n-1} distributed random variable.

For $H_1 = H_1^- : \{\mu < \mu_0\}$, the rejection rule is "Reject H_0 if $t < -d_1$", where t, d_1 and α are defined as above.

For $H_1 = H_1^\pm : \{\mu \neq \mu_0\}$, the rejection rule is "Reject H_0 if $|t| > d_2$", where t and α are defined as above, d_2 is the solution of $P(T_{n-1} > d_2) = \alpha/2$, and T_{n-1} is a t_{n-1} random variable.

★ Example

The daily withdrawals of cash at the ATM of the local bank are $N(\mu, \sigma^2)$ distributed: the bank manager conjectures that $\mu = 200$£. In a random sample of 16 days it has been observed that $\bar{x} = 240$£ and $s = 80$£.

Is the conjecture of the manager supported by the data at the 5% level?

1. null hypothesis

$$H_0 : \mu = 200$$

2. the alternative hypothesis

$$H_1 : \mu \neq 200$$

3. the test statistic

$$T = \sqrt{16}\,\frac{(\bar{X} - 200)}{S}$$

4. the limit distribution of the test statistic under the null

$$T \sim t_{15}$$

5. the decision rule

$$Reject\ if\ |t| \geq 2.13$$

6. the size of the test

$$\alpha = 0.05$$

7. the limit distribution of the test statistic under the alternative

$$\sqrt{16}\,\frac{(\bar{X}-\mu)}{S} \sim t_{15},\ \text{so}\ T = T_{15} + \sqrt{16}\,\frac{(\mu-200)}{S}\ \text{where}\ T_{15} \sim t_{15}$$

Notice that, as S is also a random variable, then T is not t_{15} distributed. In fact, its distribution is called "non-central t_{15}" (we did not see this distribution).

8. the power of the test
$T_{15} + \sqrt{16}\,\frac{(\mu - 200)}{S}$ is distributed as a "non-central t". Its distribution also depends on the unknown value σ: in the table below, we use $\sigma = 80$ (but the manager does not know σ).

μ	160	170	180	190	200	210	220	230	240
size					0.05				
Power	0.46	0.29	0.15	0.07		0.07	0.15	0.29	0.46

9. the realisation of the test statistic

$$t = \sqrt{16}\,\frac{(\bar{x} - 200)}{s} = \sqrt{16}\,\frac{(240 - 200)}{80} = 2$$

10. wether the null hypothesis is rejected or not.
As $|t| < 2.13$, the realisation is in not the rejection area so H_0 is not rejected. The conjecture of the manager is statistically supported.

Testing the mean for a general distribution.

Let $X_1, \ldots, X_n$ be independent and identically distributed, with $E(X_i) = \mu$, $Var(X_i) = \sigma^2$, $i = 1, \ldots, n$ (X_i $i.i.d.(\mu, \sigma^2)$).

We are interested in $H_0 : \{\mu = \mu_0\}$ where μ_0 is a known constant. We do not know σ^2, but we estimated $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$. Let

$$T = \sqrt{n} \frac{(\bar{X} - \mu_0)}{S}$$

even under H_0, we do not know the distribution of T. However, as $n \to \infty$, $T \to_d N(0, 1)$, so T is a valid test statistic in large samples. The decision rule depends on the type of alternative. The decision rule given for the normal distribution of X_i (with known σ^2) are applied.

Asymptotic test for population proportions.

Let X_i be the outcome of an experiment that may be successful with probability θ, and consider an independent random sample $X_1, \ldots, X_n$, and assume we are interested in the null

$$H_0 : \{\theta = \theta_0\} \ (0 < \theta_0 < 1).$$

We saw that, for given α and H_1, a valid test can be constructed using the Binomial distribution. It is, however, rather tedious to derive a critical value using that formula as the sample gets large. Fortunately, two alternative test statistics are available as the sample gets large: let

$$\widehat{\theta} = \frac{X}{n}$$

(notice that $\widehat{\theta}$ is then the MLE of θ), then, under H_0, as $n \to \infty$,

$$\widehat{T} = \frac{\sqrt{n}\left(\widehat{\theta} - \theta_0\right)}{\sqrt{\widehat{\theta}\left(1 - \widehat{\theta}\right)}} \to_d N(0,1) \text{ and } T^{(0)} = \frac{\sqrt{n}\left(\widehat{\theta} - \theta_0\right)}{\sqrt{\theta_0(1 - \theta_0)}} \to_d N(0,1) \ .$$

★Example

A firm is considering to advertise on a certain web-site. They obtained a random sample of 951 people and found that 412 looked at that site at least once a week. Test at the 5% level the hypothesis that half of all the people looked at that site at least once a week, against the alternative than only less than half of the people do so.

1. the null hypothesis

$$H_0 : \theta = 0.5$$

2. the alternative hypothesis

$$H_1 : \theta < 0.5$$

3. the test statistic

$$T = \frac{\sqrt{n}\left(\hat{\theta} - 0.5\right)}{\sqrt{0.5(1-0.5)}}$$

4. the limit distribution of the test statistic under the null

$$T \rightarrow_d N(0,1)$$

5. the decision rule

$$Reject\ if\ t < -1.65$$

6. the size of the test

$$\alpha = 0.05$$

7. the limit distribution of the test statistic under the alternative

$$\sqrt{n}\,\frac{\left(\widehat{\theta}-\theta\right)}{\sqrt{\theta(1-\theta)}} \rightarrow_d N(0,1).$$

$$\sqrt{n}\,\frac{\left(\widehat{\theta}-\theta\right)}{\sqrt{\theta(1-\theta)}} = \sqrt{n}\,\frac{\left(\widehat{\theta}-\theta\right)}{\sqrt{\theta(1-\theta)}}\,\frac{\sqrt{\theta_0(1-\theta_0)}}{\sqrt{\theta_0(1-\theta_0)}} =$$

$$\sqrt{n}\,\frac{\left(\widehat{\theta}-\theta\right)}{\sqrt{\theta_0(1-\theta_0)}}\,\frac{\sqrt{\theta_0(1-\theta_0)}}{\sqrt{\theta(1-\theta)}} = \sqrt{n}\,\frac{\left(\widehat{\theta}\pm\theta_0-\theta\right)}{\sqrt{\theta_0(1-\theta_0)}}\,\frac{\sqrt{\theta_0(1-\theta_0)}}{\sqrt{\theta(1-\theta)}}$$

$$= \sqrt{n}\,\frac{\left(\widehat{\theta}-\theta_0\right)}{\sqrt{\theta_0(1-\theta_0)}}\,\frac{\sqrt{\theta_0(1-\theta_0)}}{\sqrt{\theta(1-\theta)}} + \sqrt{n}\,\frac{(\theta_0-\theta)}{\sqrt{\theta_0(1-\theta_0)}}\,\frac{\sqrt{\theta_0(1-\theta_0)}}{\sqrt{\theta(1-\theta)}}$$

$$\text{so } \sqrt{n}\,\frac{\left(\widehat{\theta}-\theta_0\right)}{\sqrt{\theta_0(1-\theta_0)}} \rightarrow_d \frac{\sqrt{\theta(1-\theta)}}{\sqrt{\theta_0(1-\theta_0)}}Z - \sqrt{n}\,\frac{(\theta_0-\theta)}{\sqrt{\theta_0(1-\theta_0)}}$$

(Normal with mean $\sqrt{951}\,\frac{(0.5-\theta)}{\sqrt{0.5(1-0.5)}}$, variance $\frac{\theta(1-\theta)}{0.5(1-0.5)}$)

8. the power of the test

$$P(T < -1.65) \to P\left(\frac{\sqrt{\theta(1-\theta)}}{\sqrt{0.5(1-0.5)}} Z - \sqrt{951}\, \frac{(0.5-\theta)}{\sqrt{0.5(1-0.5)}} < -1.65 \right)$$

$$= P\left(Z < \frac{\sqrt{0.5(1-0.5)}}{\sqrt{\theta(1-\theta)}} \left(-1.65 + \sqrt{951}\, \frac{(0.5-\theta)}{\sqrt{0.5(1-0.5)}} \right) \right)$$

so for $\theta = 0.48$,

$$\frac{\sqrt{0.5(1-0.5)}}{\sqrt{0.48(1-0.48)}} \left(-1.65 + \sqrt{951}\, \frac{(0.5-0.48)}{\sqrt{0.5(1-0.5)}} \right) = -0.4168, P(Z < -0.41) = 0.34$$

so for $\theta = 0.45$,

$$\frac{\sqrt{0.5(1-0.5)}}{\sqrt{0.45(1-0.45)}} \left(-1.65 + \sqrt{951}\, \frac{(0.5-0.45)}{\sqrt{0.5(1-0.5)}} \right) = 1.4411, P(Z < 1.144) = 0.93$$

θ	0.45	0.48	0.5
size			0.05
Power	0.93	0.34	

9. the realisation of the test statistic

$$t = \frac{\sqrt{951}\,(\frac{412}{951} - 0.5)}{\sqrt{0.5(1-0.5)}} = -4.1183$$

10. wether the null hypothesis is rejected or not.
Since $t < -1.65$, the realisation of the test is in the rejection area so H_0 is rejected.

Note: it would have also been possible to use the statistic $\sqrt{n}\,\frac{(\hat{\theta}-\theta)}{\sqrt{\hat{\theta}(1-\hat{\theta})}}$,

because $\sqrt{n}\,\frac{(\hat{\theta}-\theta)}{\sqrt{\hat{\theta}(1-\hat{\theta})}}\Big|_{H_0} \to_d N(0,1)$ as $n \to \infty$.

Testing the difference between two means

A course of languages is taught by two instructors. They both teach the same topics and the students are randomly assigned. Instructor A has 95 students, and instructor B has 122. For each instructor, the sample averages and standard deviations of the mark in the final exam are

$\bar{x}_A$	s_A	$\bar{x}_B$	s_B
54	10	58	21

Test at the 5% the hypothesis that the means of the two distributions are the same.

★ Discussion.

Let X_{iA} be the mark of a student in Group A, and X_{iB} the mark of a student in Group B.

We assume that there is μ_A, σ_A^2 such that $E(X_{iA}) = \mu_A$, $Var(X_{iA}) = \sigma_A^2$, and μ_B, σ_B^2 such that $E(X_{iB}) = \mu_B$, $Var(X_{iB}) = \sigma_B^2$.

Let $\bar{X}_A = \frac{1}{n_A}\sum_{i=1}^{n_A} X_{iA}$, $S_A^2 = \frac{1}{n_A-1}\sum_{i=1}^{n_A}(X_{iA}-\bar{X}_A)^2$, and $\bar{X}_B = \frac{1}{n_B}\sum_{i=1}^{n_B} X_{iB}$, $S_B^2 = \frac{1}{n_B-1}\sum_{i=1}^{n_B}(X_{iB}-\bar{X}_B)^2$.

We want to test

$$H_0 : \{\mu_A = \mu_B\}.$$

By independence,

$$E(\bar{X}_A - \bar{X}_B) = \mu_A - \mu_B,\ Var(\bar{X}_A - \bar{X}_B) = \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B},$$

and it possible to verify that

$$\frac{(\bar{X}_A - \bar{X}_B) - (\mu_A - \mu_B)}{\sqrt{S_A^2/n_A + S_B^2/n_B}} \to_d N(0,1) \text{ as } n_A \to \infty,\ n_B \to \infty,$$

so our test statistic is $T = \frac{(\bar{X}_A - \bar{X}_B)}{\sqrt{S_A^2/95 + S_B^2/122}}$, and the rejection rule is "reject if the realisation of the T, t, is such that $|t| > 1.96$".

Since $t = \frac{(54-58)}{\sqrt{10^2/95 + 21^2/122}} = -1.8515$, the null hypothesis is not rejected.

Difference between two proportions

To analyse the effectiveness of a new treatment, a group of 124 of patients has been subject to the new treatment: one week later, 65 of these have been reported as healed. A second group of 256 patients has been treated in the conventional way: of these, 150 were healed after a week. Each patient was assigned to a group randomly.

Compare the two procedures using a 5% test.

★ Discussion. Let X_A be the number of patients healed with the new treatment, out of n_A patients taking the new treatment, and X_B the number of patients healed with the old treatment, out of n_A patients taking the new treatment, and let θ_A be the probability to be healed with the new treatment, and θ_B the probability to be healed with the old treatment. Then

$$H_0 : \{\theta_A = \theta_B\}, H_1 : \{\theta_A \neq \theta_B\}$$

Let

$$\widehat{\theta}_A = \frac{X_A}{n_A}, \widehat{\theta}_B = \frac{X_B}{n_B}$$

and we know that

$$E\left(\widehat{\theta}_A\right) = \theta_A, E\left(\widehat{\theta}_B\right) = \theta_B$$

$$Var\left(\widehat{\theta}_A\right) = \theta_A(1-\theta_A)/n_A, Var\left(\widehat{\theta}_B\right) = \theta_B(1-\theta_B)/n_B$$

and, using the fact that the two treatments are run independently, $Cov\left(\widehat{\theta}_A, \widehat{\theta}_B\right) = 0$, and

$$Var\left(\widehat{\theta}_A - \widehat{\theta}_B\right) = \theta_A(1-\theta_A)/n_A + \theta_B(1-\theta_B)/n_B.$$

Moreover, when $\theta_A = \theta_B$, introducing θ_0 such that $\theta_0 = \theta_A = \theta_B$,

$$Var\left(\widehat{\theta}_A - \widehat{\theta}_B, \theta_0 = \theta_A = \theta_B\right) = \theta_0(1-\theta_0)/n_A + \theta_0(1-\theta_0)/n_B$$

$$= \theta_0(1-\theta_0)(1/n_A + 1/n_B) = \theta_0(1-\theta_0)(n_A + n_B)/(n_A n_B)$$

and notice that we can estimate

$$\widehat{\theta}_0 = \frac{X_A + X_B}{n_A + n_B}$$

and

$$T = \left. \frac{\left(\widehat{\theta}_A - \widehat{\theta}_B\right)}{\sqrt{\widehat{\theta}_0\left(1-\widehat{\theta}_0\right)\frac{(n_A+n_B)}{n_A n_B}}} \right|_{H_0} \rightarrow_d N(0,1) \text{ as } n \rightarrow \infty.$$

Then, $\widehat{\theta}_A = \frac{65}{124} = 0.52419$, $\widehat{\theta}_B = \frac{150}{256} = 0.58594$ $\widehat{\theta}_0 = \frac{65+150}{124+256} = 0.56579$,

$t = \frac{\left(\frac{65}{124} - \frac{150}{256}\right)}{\sqrt{\frac{65+150}{124+256}\left(1-\frac{65+150}{124+256}\right)\frac{(124+256)}{124\times 256}}} = -1.1386$ so H_0 is not rejected.

Note: it would have also been possible to use the statistic

$$\frac{\widehat{\theta}_A - \widehat{\theta}_B}{\sqrt{\widehat{\theta}_A\left(1-\widehat{\theta}_A\right)/n_A + \widehat{\theta}_B\left(1-\widehat{\theta}_B\right)/n_B}} \rightarrow_d N(0,1)$$

as $n_A \rightarrow \infty$, $n_B \rightarrow \infty$.

P-Values

In alternative to comparing the realization of the test statistic t against the critical value, it is also possible to compute the P-value, and compare it to the size: the null hypothesis is rejected if the P-value is less than the size.

For a test statistic with realization t and limit distribution T_{H0} under the null, and one-sided positive alternative, the P-value is $P(T_{H0} > t)$; likewise, for one-sided negative alternative, the P-value is $P(T_{H0} < t)$. Finally, for a two sided alternative (with symmetric limit distribution under the null), the P-value is $P(|T_{H0}| > |t|)$.

★ Example: Fair coin (again).

Consider again the example of tossing a coin ten times and recording the number of heads to check if the coin if fair. Suppose that we set the size at 5% and we run the experiment, and we observed 7 heads. We have:

X is the number of heads: under H_0, X is binomially distributed with $\theta = 1/2$, $n = 10$, so this is the distribution we take for T_{H0}. As the alternative is two-sided, we reject H_0 if the realization is too large or too little: for 7 heads, this means using 8 or more heads or 2 or less. Using the formulas of the binomial distribution, the P-value associated to 7 heads is then 0.10938. As this exceeds 0.05 we do not reject H_0.

★ Example: The Fast Tortoise again (again).

Recall the example of the Railway Regulation Authority auditing the Fast Tortoise. We had

$$H_0 : \mu = 2, H_1 : \mu > 2$$

and test statistic

$$\sqrt{25}\,\frac{(\bar{X} - 2)}{0.9} \rightarrow_d N(0,1) \text{ under } H_0$$

The realization was $t = 2.22$; for a standard normal Z,

$$P(Z < 2.22) = 0.9868$$

so the P value is $P(Z > 2.22) = 1 - 0.9868 = 0.0132$.

Choosing the right size

So how do we choose the right size?

It is a trade-off between Type 1 and Type 2 error, so one should choose on a case by case basis. However, the practise is to set the size at 1%, 5% or 10%. Some practitioners think that rejecting at 1% we have "stronger evidence" than rejecting at 10%. We must, however, choose the size before running the test: if we set the size at 1%, we will run a much bigger risk of not rejecting a null hypothesis that we should have rejected (Type 2 error).