Statistics

MATH2089

Term 2, 2019

Some quotes

"I am not much given to regret, so I puzzled over this one a while. Should have taken much more statistics in college, I think."

Max Levchin, Paypal Co-founder, Slide Founder, 2010

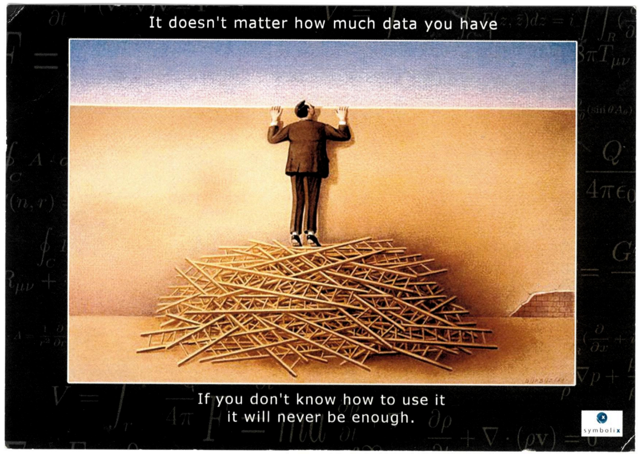
"I keep saying that the sexy job in the next 10 years will be statisticians, and I'm not kidding."

Hal Varian, Chief Economist at Google, 2009

"We live in the **Big Data** era: the world contains an unimaginably vast amount of digital information which is getting ever vaster ever more rapidly. This makes it possible to do many things that previously could not be done: spot business trends, prevent diseases, combat crime and so on.

Managed well, the data can be used to unlock new sources of economic value, provide fresh insights into science and hold governments to account. Yet, this is conditional on the existence of a statistical toolbox suitable for Big Data, the profusion and nature of which inducing commensurate statistical challenges."

The Economist, 2010


It doesn't matter how much data you have
If you don't know how to use it
it will never be enough.
symbolix

1 Introduction

What is statistics?

In order to learn about something, we must first collect observations, referred to as **data**

Definition

Statistics is the science of the (i) collection, (ii) processing, (iii) analysis, and (iv) interpretation of data

In short, statistics is learning from data and it allows us to gain new insights into the behaviour of many phenomena.

- Statistical concepts and methods are not only useful but indeed are often vital in understanding the world around us
- Further, it allows us to turn observational evidence into information for decision making, which is probably the most important aspect

What is statistics?

In engineering, this includes diversified tasks like

- calculating the average length of the downtimes of a computer
- predicting the reliability of a launch vehicle
- evaluating the effectiveness of commercial products
- studying the vibrations of airplane wings
- checking whether the level of lead in the water supply is within safety standards
- determining the strength of supports for generators at a power plant
- collecting and presenting data on the number of persons attending seminars on solar energy
- ...

What is statistics?

Statistics is a discipline that makes use of mathematics, computer science and subject matter expertise

Statistics considers the presence of randomness, uncertainty and variation, which are everywhere in real life

- If each computer had exactly the same length of downtime,
- If the level of lead was exactly identical everywhere and every time in the water supply,
- If each seminar attracted the same number of people,

  ... and if those values were known with absolute accuracy,

then a single observation would reveal all desired information, we would not need statistics :-(

Statistics

Statistics allows us to describe, understand and control the variability insofar as possible and to **take this uncertainty into account** when making judgements and decisions.

Example 1: Does cloud seeding work?

Cloud seeding is the attempt to change the amount of precipitation that falls from clouds, by dispersing substances into the air that serve as cloud condensation.

The usual intent is to increase precipitation (rain or snow).

1. A natural question may be

   "Does cloud seeding using a given substance (say, silver nitrate) really work ?"

   ⇒ **research question**

How can we answer this question?

2. First, we should observe the amount of precipitation that falls from seeded clouds, as well as from unseeded clouds

   ⇒ **experiment, collection of data**

Example 1: Does cloud seeding work?

For our experiment, we observe 52 clouds, 26 of which were chosen at random and seeded with silver nitrate.

The following rainfall (in acre-feet) are recorded:

Unseeded Clouds

```
1202.6 830.1 372.4 345.5 321.2 244.3 163.0 147.8
95.0 87.0 81.2 68.5 47.3 41.1 36.6 29.0 28.6 26.3
26.1 24.4 21.7 17.3 11.5 4.9 4.9 1.0
```

Seeded Clouds

```
2745.6 1697.8 1656.0 978.0 703.4 489.1 430.0 334.1
302.8 274.7 274.7 255.0 242.5 200.7 198.6 129.6
119.0 118.3 115.3 92.4 40.6 32.7 31.4 17.5 7.7 4.1
```

These values are our **data**

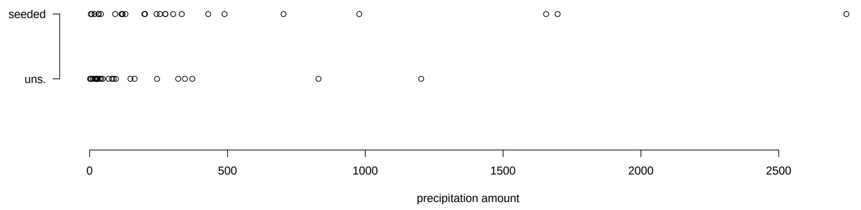
Of course, we observe variability: we could not expect each cloud (seeded or not) to give exactly the same amount of rain!

Example 1: Does cloud seeding work?

What can we do with those numbers?

3. We should present the data so that they are readily comprehensible

This includes graphical representations



as well as numerical summary measures

average prec. seeded = 441.98 average prec. unseeded = 164.58

The description and summarisation of data, is called **descriptive statistics** (Chapter 2)

Example 1: Does cloud seeding work?

At first sight, seeded clouds seem to give more precipitation than unseeded clouds.

Careful! - We must take into account the possibility of chance:

- Due to chance only, the 26 seeded clouds might be the clouds that would have given more rainfall anyway
- Due to chance only, the 26 unseeded clouds might be the clouds that would have given less rainfall anyway

⇒ Can we really conclude that the observed higher amount of rainfall for seeded clouds is due to seeding? Or is it possible that the seeding was not responsible for that but rather that the higher rainfall amount was just a **chance occurrence**?

- We have only observed 52 clouds. If we had observed 52 (or more) other clouds, would we observe different rainfall amounts?

⇒ Can we really generalise what we are seeing on a particular data set beyond that data set? How risky is it?

Example 1: Does cloud seeding work?

4. We should analyse and interpret the data bearing in mind that the observed features may be consequences of chance only

This part of statistics is called **statistical inference** (Chapters 6-12)

Inferential questions include:

- Are such generalisations reasonable or justifiable?
- How far to go with generalising from an observed data set?
- Do we need to collect more data?

Some of the most important problems in inference concern the appraisal of the risks and consequences of making wrong decisions.

- Risks are often appraised by calculating **probabilities** of some events occurring
- We will discuss **probability theory** in more details in Chapters 3-5

Example 1: Does cloud seeding work?

5. Finally, we should draw conclusions from our investigations, that is, we should answer the question

   "Does cloud seeding using silver nitrate result in more rainfall than not cloud seeding using silver nitrate?"

The statistical process

The 5 points in the above example form the **typical procedure for statistical inference**:

1. Set clearly defined goals for the investigation; formulate the research question
2. Decide what data is required/appropriate and how to collect them; collect the data
3. Display, describe and summarise the data in an efficient way; check for any unusual data features
4. Choose and apply appropriate statistical methods to extract useful information from the data
5. Interpret the information, draw conclusions and communicate the results to others

Fact

Every step in this process requires understanding statistical principles and concepts as well as knowledge and skills in statistical methods.

Example 2: Hair colour and pain tolerance

An experiment conducted at the University of Melbourne suggests that there may be a difference in pain threshold for blonds and brunettes.

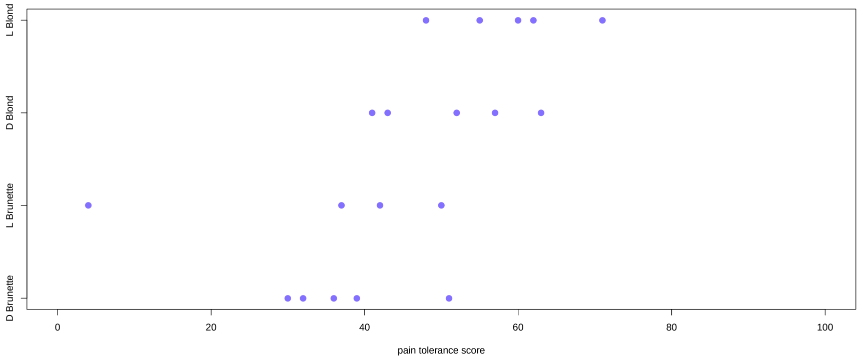
1. The research question is:

   "Is pain threshold related to hair colour?"

A group of 19 subjects was divided into light blond, dark blond, light brunette and dark brunette groups and a pain threshold score was measured for each subject

2. The data are ... (A higher score means a higher pain threshold)
3. (better seen in the next figure)

Example 2: Hair colour and pain tolerance



Example 2: Hair colour and pain tolerance

4. Pain threshold seems to increase with lighter hair colour, but is this effect real or just due to chance ? (the number of observations is quite small: we have 4 or 5 observations per hair colour)

⇒ We have to apply some inferential method to come to a conclusion

5. Depending on what we have observed, either

   It is clear from the data that pain tolerance is related to hair colour

   or

   The data do not allow us to conclude that pain tolerance is related to hair colour

Remark: In the latter case we won't say "pain tolerance is not related to hair colour". It might still be the case, but with such a small number of observations, we are not sure and it would be too risky to affirm it is.

Population

Usually, we are interested in obtaining information about a total collection of elements, which is referred to as the **population**

The elements are often called **individuals** (or units)

Given the research question, we have observed some characteristic for each individual. This characteristic, which could be quantitative or qualitative, is called a **variable**

Example 1

In Example 1 (clouds seeding), the population consists of all the clouds of the sky. An individual is a cloud and the variable of interest is the amount of rainfall.

Example 2

In Example 2 (pain tolerance), the population consists of all blonds and brunettes of the world. An individual is one of those people and the variable of interest is the pain threshold score.

Sample

- It is often physically impossible or infeasible from a practical standpoint to obtain data on the whole population
- Think also of very expensive, or very time-consuming, or destructive experiments
- In most situations, we can only observe a subset of the population, that is, we must work with only partial information

The subset of the population which is effectively observed is called the **sample**.

The **data** are the measurements that are actually collected over the sample in the course of the investigation.

Note: sometimes, we may use "sample" to designate the subset of measurements actually observed (i.e., the data)

Sample

In Example 1, the sample consists of the 52 clouds whose rainfall amounts have been recorded

(We might also consider that we have two samples: 26 seeded clouds and 26 unseeded clouds)

In Example 2, the sample consists of the 19 persons whose pain threshold scores have been measured

(We might also consider that we have 4 samples of people with different hair colour)

Fact

The distinction between the data actually acquired (the sample) and the vast collection of all potential observations (the population) is a **key to understanding statistics**

Sampling

- The process of selecting the sample is called sampling
- If the sample is to be informative about the total population, it must be representative of that population

⇒ Suppose you are interested in the average height of UNSW students, would you select the sample from the UNSW basketball team?

⇒ Suppose you are interested in the average age of UNSW students, would you select a sample made up of postgraduate students only?

- The quality of the data is paramount in a statistical study

Your results are only as good as your data !

- Sampling must be carefully done, impartially and objectively

Random sampling

In practice, the only sampling scheme that guarantees the sample to be representative of the population is **random sampling**.

⇒ The individuals of the sample are selected in a totally random fashion, without any other prior consideration

Any non-random selection of a sample often results in one which is inherently biased toward some values as opposed to others.

⇒ We must not attempt to deliberately choose the sample according to some criteria

⇒ Instead, we should just leave it up to “chance” to obtain a sample which correctly covers the underlying population

The importance of random sampling

Once a random sample is chosen, we can use statistical inference to draw conclusions about the entire population, taking the randomness into account (using probabilities).

⇒ Not possible if the sample is not random!

Information drawn in a non-random sample cannot, as a rule, be generalised to larger populations.

Fact

The statistical procedures presented in this course may not be valid when applied to non-random samples.

⇒ Never unquestioningly accept samples without knowing how the data have been generated / collected / observed

Objectives

Now you should be able to:

- identify the role that statistics can play in the engineering problem-solving process □
- discuss how variability affects the data collected and used for making engineering decisions □
- discuss how probability theory is used in engineering and science □
- discuss the importance of random sampling □

2 Descriptive Statistics

Introduction

On Slide 6, we defined statistics as **learning from data**.

However, statistical data, obtained from surveys, experiments, or any series of measurements, are often so numerous that they are virtually useless unless they are condensed.

⇒ **Data should be presented in ways that facilitate their interpretation and subsequent analysis**

The aspect of statistics which deals with organising, describing and summarising data is called **descriptive statistics**.

Essentially, descriptive statistics tools consist of

1. graphical methods and
2. numerical methods

Types of variables

There are essentially two types of variables:

1. **categorical** (or qualitative) variables: take a value that is one of several possible categories (no numerical meaning)

   Eg. gender, hair colour, field of study, status, etc.
2. **numerical** (or quantitative) variables: naturally measured as a number for which meaningful arithmetic operations make sense

   Eg. height, age, temperature, pressure, salary, etc.

Attention: sometimes categorical variables are disguised as quantitative variables.

For example, one might record gender information coded as 0 = Male, 1 = Female. It remains a categorical variable, it is not naturally measured as a number.

Types of variables

- categorical variables
  - ordinal: there is a clear ordering of the categories

    Eg. salary class (low, medium, high), opinion (disagree, neutral, agree), etc.
  - nominal: there is no intrinsic ordering to the categories

    Eg. gender, hair colour, etc.
- numerical variables
  - discrete: the variable can only take a finite (or countable) number of distinct values

    Eg. number of courses you are enrolled in, number of persons in a household, etc.
  - continuous: the variable can take any value in an entire interval on the real line (uncountable)

    Eg. height, weight, temperature, time to complete a task, etc.

Graphical representations

A picture is worth a thousand words

Graphical representations are often the most effective way to quickly obtain a feel for the essential characteristics of the data.

Fact

Any good statistical analysis of data should **always** begin with plotting the data.

- Plots often reveal useful information and opens paths of inquiry
- They might also highlight the presence of irregularities or unusual observations ("outliers")

Dotplot

- A **dotplot** is an attractive summary of numerical data when the data set is reasonably small
- Each observation is represented by a dot above the corresponding location on a horizontal measurement scale
- When a value occurs more than one time, there is a dot for each occurrence and these dots are stacked vertically

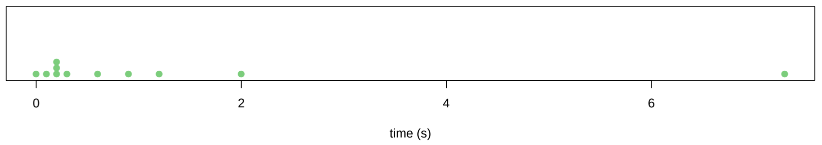
Dotplot: example

Example

In 1987, for the first time, physicists observed neutrinos from a supernova that occurred outside of our solar system. At a site in Kamiokande, Japan, the following times (in seconds) between neutrinos were recorded:

`0.107; 0.196; 0.021; 0.283; 0.179; 0.854; 0.58; 0.19; 7.3; 1.18; 2`

Draw a dotplot:



Note that the largest observation is extremely different to the others. Such an observation is called an outlier.

⇒ Could be: recording error? missed observations? real observation?

⇝ further investigation is needed

Stem-and-leaf plot

- The dotplot is useful for small samples, up to (say) about 20 observations. However, when the number of observations is moderately large, another graphical display may be more useful
- A **stem-and-leaf plot** (or just stemplot) is another effective way to organise numerical data without much effort
- Idea: separate each observation into a **stem** (all but last digit) and a **leaf** (final digit)

⇒ Example: 24 := 2|4 139 := 13|9 5 := 0|5

- Write all unique stems in vertical column with the smallest at the top, and draw a vertical line at the right of this column
- Write each leaf in the row to the right of its stem, in increasing order out from the stem

Stem-and-leaf plot: example

Example 1.5 in the textbook

Study of use of alcohol by university students. We are interested in a variable X, the percentage of undergraduate students who are binge drinkers. We observe X on 140 campuses across the US

The sample is:

```
26 57 66 66 41 46 65 35 46 38 44 29 43
14 11 68 37 27 18 46 30 32 35 59 39 32 31 39 21 58
65 50 44 29 53 27 38 52 29 58 45 34 36 56 47 22 59
46 24 51 26 39 23 55 50 42 18 48 64 44 46 66 33 61
38 35 22 57 42 42 26 47 67 37 39 58 26 41 61 51 61
56 48 53 13 28 52 36 62 31 38 42 64 51 54 33 19 25
42 37 36 55 37 56 43 28 56 49 39 57 48 52 60 17 49
61 44 18 67 36 58 47 16 33 27 29 48 45 34 57 56 48
46 49 15 52 04 41 64 37
```

(Source: "Health and Behavioural Consequences of Binge Drinking in College", J. Amer. Med. Assoc., 1994, 1672-1677)

Stem-and-leaf plot: example

⇒ Separate the tens digit ("stem") from the ones digit ("leaf")

```
0 | 4
1 | 1345678889
2 | 1223456666777889999
3 | 0112233344555666677777888899999
4 | 111222222334444556666666777888888999
5 | 00111222223345566666677777888899
6 | 01111244455666778
```

This stemplot suggests that

- a typical value is in the stem 4 row (probably in mid-40% range)
- there is a single peak, but the shape is not perfectly symmetric
- there are no observations unusually far from the bulk of the data (no outliers)

Stem-and-leaf plot

The stemplot conveys information about the following aspects of the data:

- identification of a typical value
- extent of spread about the typical values
- presence of any gaps in the data
- extent of symmetry in the distribution values
- number and location of peaks
- presence of any outlying values

Stem-and-leaf plot: variations

There are many ways in which stem-and-leaf plots can be modified to meet particular needs:

- rounding or truncating the numbers to a few digits before making a stemplot to avoid too much irrelevant detail in the stems
- splitting each stem to give greater detail in distribution
- back-to-back stemplots with common stems to compare two related distributions

Stem-and-leaf plot: splitting each stem

A more informative display can be created by repeating each stem value twice, once for the low leaves 0, 1, 2, 3, 4 and again for the high leaves 5, 6, 7, 8, 9.

For the binge-drinking data this yields (compare Slide 35)

```
0 | 4
0 |
1 | 134
1 | 5678889
2 | 12234
2 | 56666777889999
3 | 0112233344
3 | 555666677777888899999
4 | 11122222334444
4 | 5566666677788888999
5 | 001112222334
5 | 5566666777788899
6 | 011112444
6 | 55666778
```

Stem-and-leaf plot: back-to-back plots

Suppose you have two data sets, each consisting of observations on the same variable (for instance, exam scores for two different classes).

- in what ways are the two data sets similar? how do they differ?
- comparative stem-and-leaf plot, or back-to-back stemplots

The stems are common, the leaves for one data set are listed to the right and the leaves for the other to the left.

For instance, for exam scores we could observe (Class 1 | Class 2)

```
      2588 | 5 | 9
   2234578 | 6 | 01445
0225556689 | 7 | 1223567
      4479 | 8 | 01334578
           | 9 | 156688
```

⇒ The right side appears to be shifted down one row from the other side (better scores in Class 2 than in Class 1)

Frequency distribution, bar charts and histograms

A natural continuation of the stemplot is to count the number of observations (that is, the **frequency**) in each row.

(This can be done each time that proper “categories” are available)

The frequency of observations in each category then forms the **frequency distribution**.

Which categories ?

- For a categorical variable, the categories are obviously defined

⇒ count the frequency of observations in each category, mark each category on a horizontal axis and draw rectangles whose heights are the corresponding frequencies. This is called a **bar chart**

- For a discrete numerical variable, the categories are given by the distinct values taken by the variable

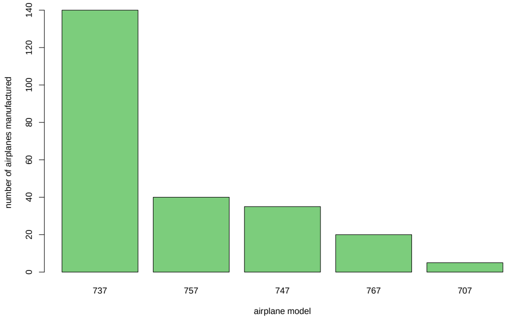
⇒ then, proceed as above. This is called a **histogram**

Bar chart: example

Example

In 1985 the Boeing Company published the figures for its production of transport aircraft. That year, they produced 5 Boeing 707's, 140 Boeing 737's, 35 Boeing 747's, 40 Boeing 757's and 20 Boeing 767's

The frequency distribution for this production is presented in the following bar chart



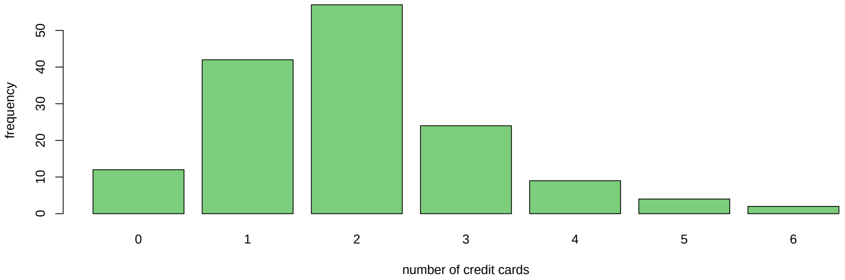
Histogram: example

Example

A sample of students from a statistics class were asked how many credit cards they carry. For 150 students, the frequency distribution is given below

Number of cards	0	1	2	3	4	5	6
Frequency	12	42	57	24	9	4	2

The frequency distribution for the number of credit cards is given in the following histogram



Histogram for a continuous numerical variable

If the variable is numerically continuous, there are no obvious categories

⇒ we have to decide on some categories, called classes, which will be intervals that do not overlap and accommodate all observations

Once the classes have been defined, we can proceed similarly to the above methods, that is:

- Determine the frequency of observations in each class, mark the class boundaries on a horizontal axis and draw rectangles whose heights are the corresponding frequencies
- However, important practical questions arise like how many classes to use and what are the limits for each class

Histogram for continuous numerical variable

Generally speaking, the number of classes depends on the total number of observations and the range of the data.

This is a trade-off between

1. choosing too few classes at a cost of losing information about actual values
2. choosing too many classes will result in the frequencies of each class to be too small for a pattern to be discernible

An empirical rule is

$$\text{number of classes} \simeq \sqrt{\text{number of observations}}$$

Note: it is common, although not essential, to choose classes of equal width

Histogram: example 1.8 in the textbook

Example

Power companies need information about customer usage to obtain accurate forecasts of demand. Here we consider the energy consumption (BTUs) during a particular period for a sample of 90 gas-heated homes

The sample is

```
10.04 13.47 13.43 9.07 11.43 12.31 4.00 9.84 10.28 8.29
6.94 10.35 12.91 10.49 9.52 12.62 11.09 6.85 15.24 18.26
11.21 11.12 10.28 8.37 7.15 9.37 9.82 9.76 8.00 10.21
6.62 12.69 13.38 7.23 6.35 5.56 5.98 6.78 7.73 9.43 9.27
8.67 15.12 11.70 5.94 11.29 7.69 10.64 12.71 9.96 13.60
16.06 7.62 2.97 11.70 13.96 8.81 12.92 12.19 16.90 9.60
9.83 8.26 8.69 6.80 9.58 8.54 7.87 9.83 10.30 8.61 7.93
13.11 7.62 10.95 13.42 6.72 10.36 12.16 10.40 5.20 10.50
8.58 14.24 14.35 8.47 7.29 12.28 11.62 7.16
```

Histogram: example

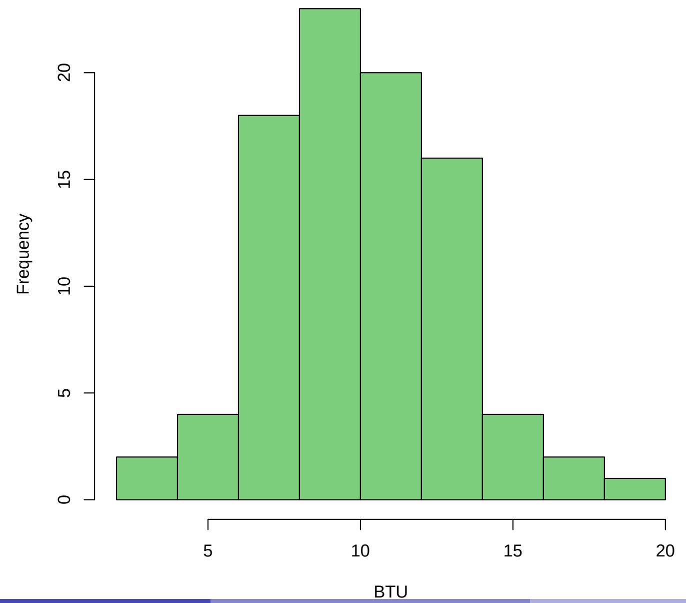
- The data set contains 90 observations, and since $\sqrt{90} \simeq 9.48$, we suspect that about nine classes will provide a satisfactory frequency distribution
- The smallest and largest data values are 2.97 and 18.26, so the classes must cover a range of at least 15.29 BTU
- As $15.29/9 \simeq 1.7$, we take the common class width equal to 2 (for simplicity), and we start at 2 (again for simplicity)

Counting the frequencies in the so-defined classes, we get

$[2,4)$	$[4,6)$	$[6,8)$	$[8,10)$	$[10,12)$	$[12,14)$	$[14,16)$	$[16,18)$	$[18,20)$
1	5	18	23	20	16	4	2	1

Note: we adopt the left-end inclusion convention, i.e. a class contains its left-end but not its right-end boundary point (interval $[a,b)$, left-closed right-open)

Histogram: example



Histogram: comments

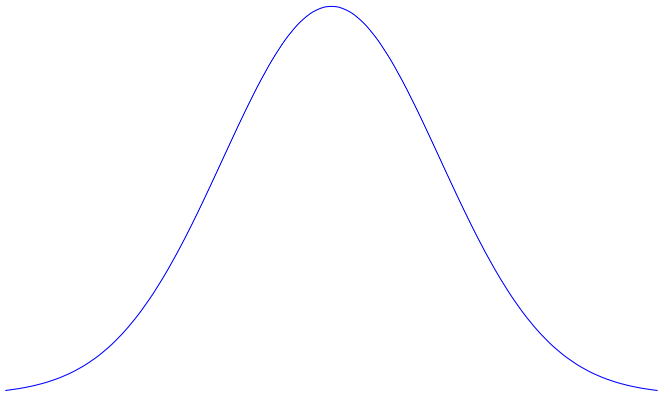
Unlike the histogram for discrete numerical variables, the histogram for continuous numerical variables consists of adjacent rectangles. This reflects the continuity of the underlying variable.

Like the stemplot, the histogram (discrete or continuous) provides a visual impression of the shape of the distribution of the observations, as well as information about the central tendency and dispersion in the data, that may not be immediately apparent from the data themselves.

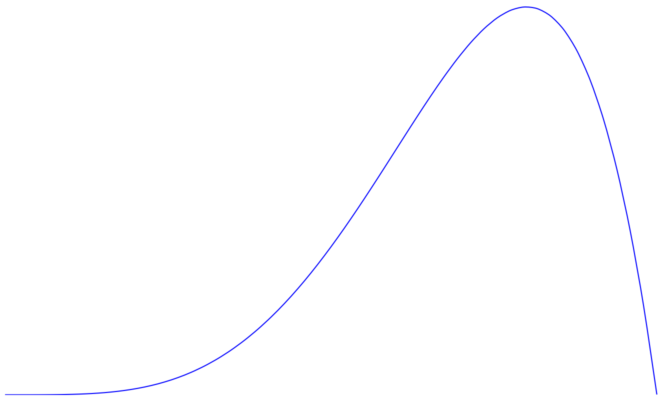
Typical words/phrases used to describe histograms:

- symmetric, or skewed to the right/left ($\Rightarrow$ with right/left tail);
- unimodal (one peak), or bimodal/multimodal;
- bell-shaped (if symmetric & unimodal);
- there are possible outliers around..., or there are no outliers;
- typical value of the data is..., the range of the data is...

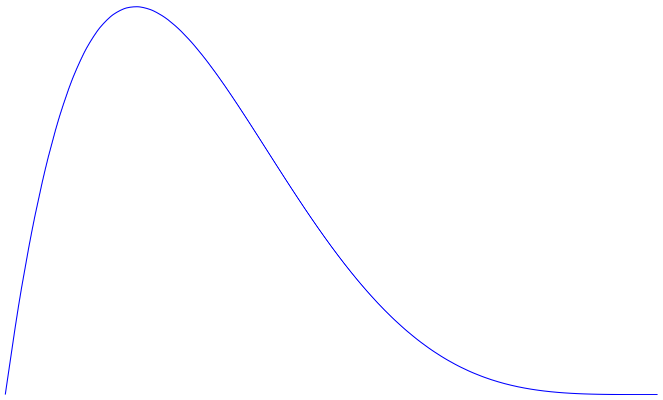
Example: Symmetric shape (‘bell-shaped’)



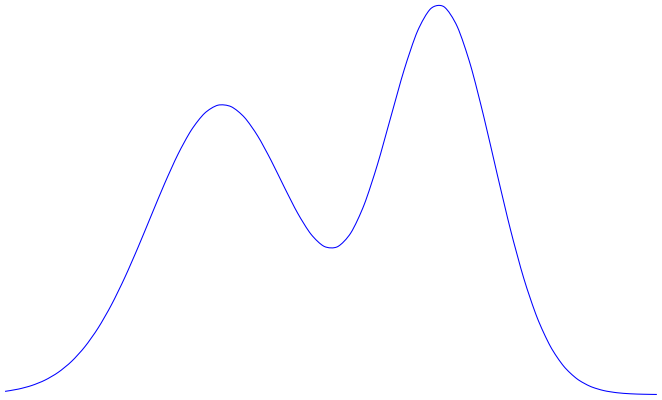
Example: Skewed to the left



Example: Skewed to the right



Example: Bimodal

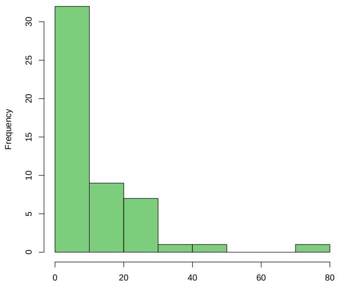


Unequal class width

Sometimes, equal-width classes may not be a sensible choice

For instance, if the data have several extreme observations or outliers, nearly all observations will fall in just a few of the classes

Consider the following histogram:



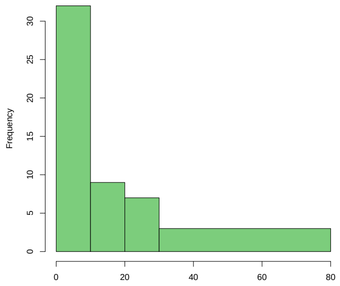
⇒ The last 5 classes together contain only 3 observations!

⇒ Might preferable to regroup the observations > 30 into a wider class

Unequal class width

However, wider classes are likely to include more observations than narrower ones!

⇒ when class widths are unequal, the frequencies will give a distorted representation of reality



⇒ the rectangular areas (not their heights) should be proportional to the frequencies : this is what a **density histogram** achieves

Density histogram

- The relative frequency of a class is the proportion of observations in that class, ie, the frequency of the class divided by the total number of observations
- We call the **density** of a class the relative frequency of the class divided by the class width
- A **density histogram** is a histogram whose rectangle heights are the densities of each class (no longer the frequencies)

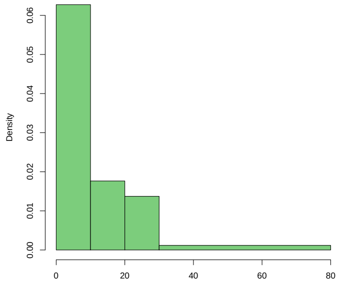
So we have:

$$\begin{aligned}\text{relative frequency} &= \text{density} \times \text{class width}\\ &= \text{rectangle height} \times \text{rectangle width}\\ &= \text{rectangle area}\end{aligned}$$

Note: the total area of the rectangles must be equal to 1, as the sum of all relative frequencies must be 1. This property is an important one, so that it is always preferable to draw a density histogram instead of a (frequency) histogram, even when the classes have equal width.

Density histogram

For the previous case, the density histogram would be



which is much more faithful to reality than histogram on Slide 54

We can check that the area of the first rectangle is $10 \times 0.063 = 0.63$, that is the first class $[0, 10)$ includes 63% of the observations

We could calculate the areas of the other rectangles the same way, and check that their sum is equal to 1

Density histogram: Example

Example

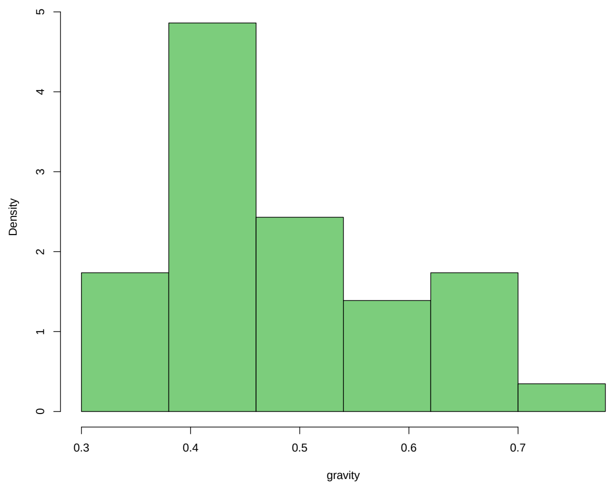
The accompanying specific gravity values for various wood types used in construction appeared in the article "Bolted Connection Design Values Based on European Yield Model" (J. of Structural Engr., 1993):

```
0.36 0.45 0.66 0.66 0.44 0.40 0.48 0.75 0.51 0.67 0.42
0.35 0.47 0.38 0.37 0.41 0.41 0.46 0.54 0.62 0.48 0.42
0.43 0.42 0.31 0.54 0.42 0.48 0.42 0.40 0.40 0.68 0.55
0.36 0.58 0.46
```

Frequency/Density table (36 observations $\rightsquigarrow$ 6 classes):

	$[0.3, 0.38)$	$[0.38, 0.46)$	$[0.46, 0.54)$	$[0.54, 0.62)$	$[0.62, 0.7)$	$[0.7, 0.78)$
Freq.	5	14	7	4	5	1
Relative Freq.	5/36	14/36	7/36	4/36	5/36	1/36
Density	$\frac{5/36}{0.08}$ $= 1.7361$	$\frac{14/36}{0.08}$ $= 4.8611$	$\frac{7/36}{0.08}$ $= 2.4306$	$\frac{4/36}{0.08}$ $= 1.3889$	$\frac{5/36}{0.08}$ $= 1.7361$	$\frac{1/36}{0.08}$ $= 0.3472$

Density histogram: Example



Descriptive measures

- Dotplots, stem-and-leaf plots, histograms and density histograms summarise a data set graphically so we can visually discern the overall pattern of variation
- It is also useful to numerically describe the data set

$\Rightarrow$ Summary measures that tell where a sample is centred (measures of centre or location), and what is the extent of spread around its centre (measures of variability)

- Usual notation:

$$x_1, x_2, \dots, x_n$$

for a sample consisting of n observations of the variable X

Note: except when indicated otherwise, we assume that X is a numerical variable.

Measure of centre: the sample mean

The most frequently used measure of centre of a sample is simply the arithmetic **mean** (or average) of the n observations.

It is usually denoted $\bar{x}$ and is given by

Sample mean

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

Note: the unit of $\bar{x}$ is the same as that of X.

Example

Metabolic rate is the rate at which someone consumes energy. Data (calories per 24 hours) from 7 men in a study of dieting are:

```
1792; 1666; 1362; 1614; 1460; 1867; 1439
```

$\Rightarrow \bar{x} = \frac{1}{7}(1792 + 1666 + 1362 + 1614 + 1460 + 1867 + 1439) = 1600$ (calories/24h)

Measure of centre: the sample median

The **median** is another descriptive measure of the centre of sample.

Sample median

The median, usually denoted m (or $\tilde{x}$), is the value which divides the data into two equal parts, half below the median and half above.

$\Rightarrow$ the sample median m is the "middlemost" value of the sample

Denote the ordered sample (smallest to largest observation)

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n-1)} \leq x_{(n)}$$

If n is odd, the median is the middle observation in the ordered data series, that is, $m = x_{\left(\frac{n+1}{2}\right)}$

If n is even, the median is *defined as* the average of the middle two observations in the ordered data series, that is

$$m = \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right)$$

Measure of centre: the sample median

Example

Metabolic rate is the rate at which someone consumes energy. Data (calories per 24 hours) from 7 men in a study of dieting are:

1792; 1666; 1362; 1614; 1460; 1867; 1439

First, order the sample

$$1362 \leq 1439 \leq 1460 \leq 1614 \leq 1666 \leq 1792 \leq 1867$$

Here, $n = 7$ (odd), so the median is the middle value $x_{(7+1)/2} = x_{(4)} = 1614$ (calories/24h).

Note: since the order of the observations matters, the median can also be defined for an ordinal categorical variable

Measure of centre: the sample median

Sometimes, it is preferable to use the median instead of the mean, as it is resistant/robust to outliers.

Example

A small company employs four young engineers, who each earn \$70,000, and the owner (also an engineer), who gets \$160,000. The latter claims that on average, the company pays \$88,000 to its engineers and, hence, is a good place to work.

The mean of the five salaries is indeed

$$\bar{x} = \frac{1}{5}(4 \times 70,000 + 160,000) = 88,000 \ \$$$

but this hardly describes the situation.

On the other hand, the median is the middle observation in

$$70,000 = 70,000 = 70,000 = 70,000 < 160,000$$

that is, \$70,000: much more representative of what a young engineer earns with the firm.

Quartiles and Percentiles

- We can also divide the sample into more than two parts
- When a sample is divided into four equal parts, the division points are called sample **quartiles**

$\Rightarrow$ The first or lower quartile q_1 is the value that has 25% of the observations below (or equal to) it and 75% of the observations above (or equal to) it

$\Rightarrow$ The third or upper quartile q_3 is the value that has 75% of the observations below (or equal to) it and 25% of the observations above (or equal to) it

$\Rightarrow$ The second quartile q_2 would split the sample into two equal halves (50% below - 50% above): that is the median ($m = q_2$)

Quartiles and Percentiles

More generally, the sample $(100 \times p)$th **percentile** (or quantile) is the value such that $100 \times p\%$ of the observations are below this value (or equal to it), and the other $100 \times (1 - p)\%$ are above this value (or equal to it).

$\Rightarrow$ q_1 is the 25th, the median is the 50th and q_3 is the 75th percentile

Quartiles and Percentiles

Practically,

$$\text{lower quartile} = \text{median of the lower half of the data}$$

$$\text{upper quartile} = \text{median of the upper half of the data}$$

Note: if n is an odd number, include the median in each half

Example (ctd.)

Metabolic rate is the rate at which someone consumes energy. Data (calories per 24 hours) from 7 men in a study of dieting are:

$$1792;\ 1666;\ 1362;\ 1614;\ 1460;\ 1867;\ 1439$$

Find the quartiles.

The median is 1614 (calories/24h). The lower half of the observations is thus

$$1362 \leq 1439 \leq 1460 \leq 1614,$$

whose median is $q_1 = \frac{1}{2}(1439 + 1460) = 1449.5$ (calories/24h).

Similarly, the third quartile $q_3 = 1729$ (calories/24h).

Five number summary

Quartiles give more detailed information about location of a data set.

Often, the three quartiles (i.e., including the median) together with the minimum and maximum observation give a good insight into the data set $\Rightarrow$ this is known as the **five number summary**

Five number summary

$$\{x_{(1)}, q_1, m, q_3, x_{(n)}\}$$

Note:

- $m = q_2$ is the 50th percentile
- q_1 is the 25th percentile, q_3 is the 75th percentile
- $x_{(1)}$ is the "0th percentile", $x_{(n)}$ is the "100th percentile"

Example: find the 5-number summary for the calories data

$$\{1362, 1449.5, 1614, 1729, 1867\}$$

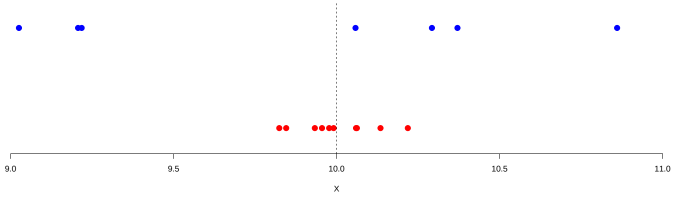
Measures of variability

One of the most important characteristics of any data set is that the observations are not all alike.

$\Rightarrow$ Mean / median describe the central location of a data set, but tell us nothing about the spread or variability of the observations

$\Rightarrow$ Different samples may have identical measures of centre, yet differ from one another in other important ways

We observe that the dispersion of a set of observations is small if the values are closely bunched about their mean (red sample), and that it is large if the values are scattered widely about their mean (blue sample) – both samples have mean 10.



Measures of variability

It would seem reasonable to measure the variability in a data set in terms of the amounts by which the values deviate from their mean

Define the deviations from the mean

$$(x_1 - \bar{x}), (x_2 - \bar{x}), \ldots, (x_n - \bar{x})$$

We might then think of using the average of those deviations as a measure of variability in the data set.

Unfortunately, this will not do, because this average is always 0:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i - \bar{x} = \bar{x} - \bar{x} = 0$$

⇒ We need to remove the signs of those deviations, so that positive and negative ones do not cancel each other out

⇒ Taking the square is a natural thing to do

Measure of variability: the sample variance

The most common measure of variability in a sample is the **sample variance**, usually denoted s^2.

The sample variance s^2 is *essentially* the average of the squared deviations from the mean $\bar{x}$.

Sample variance

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

See that the divisor for the sample variance is $n-1$, not n

⇒ s^2 is based on the n quantities $x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}$. But $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$

⇒ Thus, specifying the values of any $(n-1)$ of the deviations determines the value of the last one

The number $n-1$ is called the **number of degrees of freedom** for s^2

Measure of variability: the sample variance

It is not difficult to see that, expanding the square, we can write

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x})^2 &= \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n}\bar{x}^2 - 2\sum_{i=1}^{n} x_i\bar{x} \\ &= \sum_{i=1}^{n} x_i^2 + n\bar{x}^2 - 2\bar{x}\sum_{i=1}^{n} x_i \\ &= \sum_{i=1}^{n} x_i^2 + n\bar{x}^2 - 2n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2\end{aligned}$$

Hence,

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

This often makes the computation of the variance easier

Measure of variability: the sample standard deviation

Notice that the unit of the variance is not that of the original observations:

$$\text{unit of } s^2 = (\text{unit of } X)^2$$

⇒ difficult to interpret

Consequently, one often works with the **sample standard deviation** s.

Sample standard deviation

$$s = \sqrt{s^2} = \sqrt{\tfrac{1}{n-1}\textstyle\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

The unit of s is the same as the original unit of X

⇒ ease of interpretation in measuring spread about the mean in the original scale

The standard deviation s has a *rough* interpretation as the average distance from an observation to the sample mean.

Measure of variability: example

Example (ctd.)

Metabolic rate is the rate at which someone consumes energy. Data (calories per 24 hours) from 7 men in a study of dieting are:

`1792; 1666; 1362; 1614; 1460; 1867; 1439`

Find the variance and the standard deviation.

Here, $n = 7$ and the sample mean is $\bar{x} = 1600$ calories/24h. It follows

$$\begin{aligned} s^2 &= \frac{1}{6}\left(1792^2 + 1666^2 + \ldots + 1439^2\right) - \frac{7}{6} \times 1600^2 \\ &= 35811.87 \text{ (calories/24h)}^2 \end{aligned}$$

The standard deviation is $s = \sqrt{35811.87} = 189.24$ calories/24h

Measure of variability: iqr

The sample variance s^2 is a variability measure related to the sample mean $\bar{x}$

By constrast, the sample **Interquartile Range** (iqr) is a measure of variability related to the sample median and the quartiles

As the name suggests, iqr is given by the difference between the upper and the lower quartiles.

Sample Interquartile Range

$$\text{iqr} = q_3 - q_1$$

The interquartile range describes the amount of variation in the middle half of the observations.

It enjoys the properties of the quartiles, mainly the fact that it is less sensitive to outliers than the sample variance.

Detecting outliers from iqr

We have defined an **outlier** as an observation which is too different from the bulk of the data

⇒ How much different should an observation be to be an outlier?

An empirical rule is the following:

an outlier is an observation farther than
1.5 × iqr from the closest quartile

Besides, we say that

- an outlier is extreme if it is more than 3 × iqr from the nearest quartile
- it is a mild outlier otherwise

⇒ It is important to examine the data for possible outliers as those abnormal observations may affect most of the statistical procedures

Outliers: example

Example

Consider the energy consumption data on Slide 45. Find possible outliers.

Exercise: show that the five number summary is

$$\{2.97,\ 7.9475,\ 9.835,\ 12.045,\ 18.26\}$$

Specifically, $q_1 = 7.9475$ (BTUs) and $q_3 = 12.045$ (BTUs).

Hence, $\text{iqr} = 12.045 - 7.9475 = 4.0975$ (BTUs).

The limits for not being an outlier are thus

$$[q_1 - 1.5 \times \text{iqr}, q_3 + 1.5 \times \text{iqr}] = [1.80125,\ 18.19125]$$

⇒ only one observation (check in the data set!) falls outside that interval:

the largest value 18.26 (mild outlier)

Boxplots

A **boxplot** is a graphical display showing the five number summary and any outlier value.

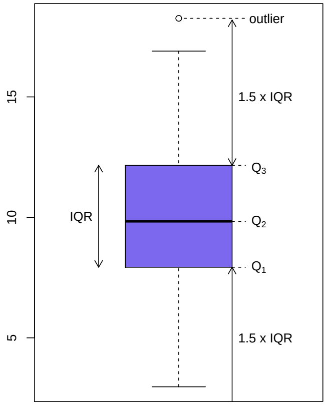
It is sometimes called **box-and-whisker** plot.

A central box spans the quartiles.

A line in the box marks the median.

Lines extend from the box out to the smallest and largest observations which are not suspected outliers.

Observations more than 1.5 × iqr outside the central box are plotted individually as outliers.



Boxplots

Example

A poll of age in years of 20 randomly chosen students led to the data:
22, 18, 20, 29, 21, 24, 21, 19, 19, 23, 19, 19, 25, 19, 19, 21, 24, 18, 21, 20
Determine the five-number summary. Draw a boxplot.

Five number summary:

$$\{ 18, 19, 20.5, 22.5, 29 \}$$

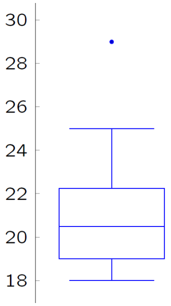
Check that:

$\text{iqr} = 22.5 - 19 = 3.5$,

$q_1 - 1.5 \times \text{iqr} = 13.75$,

$q_3 + 1.5 \times \text{iqr} = 27.75$,

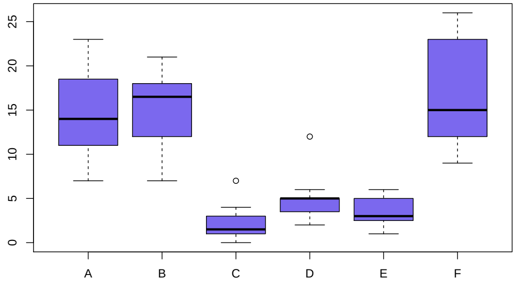
hence there is one outlier: 29



Boxplots

Boxplots are very useful graphical comparisons among data sets, because they have high visual impact and are easy to understand.

The following boxplots refer to the counts of insects in agricultural experimental units treated with six different insecticides (A to F).



At a glance you can tell that Insecticide C is the most effective (but outlier → need to investigate), that D and E are also doing well unlike A, B and F, F is especially unreliable (largest variability)

Objectives

Now you should be able to:

- understand the importance of graphical representations, construct and interpret a dotplot □
- compute an interpret the sample mean, sample variance, sample standard deviation, sample median and sample quartiles □
- construct and interpret visual data displays, including the stem-and-leaf plot, the histogram and the boxplot □
- comment and assess the overall pattern of data from visual displays □
- explain how to use the boxplots to visually compare two or more samples of data □

Recommended exercises (from the textbook):

Q5 p.20, Q17 p.23, Q3 p.69, Q15 p.77, Q35 p.86, Q37 p.86, Q39 p.86, Q53 (a,c,d) p.95, Q59 p.96, Q67 p.98, Q69 p.98 (2nd edition)

Q5 p.24, Q17 p.27, Q3 p.70, Q15 p.78, Q38 p.88, Q40 p.88, Q54 (a,c,d) p.97, Q60 p.98, Q68 p.100, Q70 p.100 (3rd edition)

3 Elements of Probability

Introduction

The previous chapter (Chapter 2) described purely descriptive methods for a given sample of data.

The subsequent chapters (Chapters 6-12) will describe **inferential methods**, that convert information from random samples into information about the whole population from which the sample has been drawn.

However, a sample only gives a partial and approximate picture of the population

⇒ drawing conclusions about the whole population, thus going beyond what we have observed, inherently involves some risk

⇒ it is important to quantify the amount of confidence or reliability in what we observe in the sample

Introduction

It is important to keep in mind the crucial role played by random sampling (Slide 23)

- Without random sampling, statistics can only provide descriptive summaries of the observed data
- With random sampling, the conclusions can be extended to the population, arguing that the randomness of the sample guarantees it to be representative of the population on average

"**Random**" is not to be interpreted as "chaotic" or "haphazard". It describes a situation in which an individual outcome is uncertain, but there is a regular distribution of outcomes in a large number of repetitions.

Probability theory is the branch of mathematics concerned with analysis of random phenomena.

⇒ Probability theory (Chapters 3-5) is a necessary link between descriptive and inferential statistics

Random experiment

Definition

A **random experiment** (or *chance experiment*) is any experiment whose exact outcome cannot be predicted with certainty.

This definition includes the ‘usual’ introduction to probability random experiments...

Experiment 1: toss a coin ; Experiment 2: roll a die ;

Experiment 3: roll two dice

... as well as typical engineering experiments...

Experiment 4: count the number of defective items produced on a given day

Experiment 5: measure the current in a copper wire

... and obviously the “random sampling” experiment

Experiment 6: select a random sample of size n from a population

Sample space

To model and analyse a random experiment, we must understand the set of all its possible outcomes

Definition

The set of all possible outcomes of a random experiment is called the **sample space** of the experiment. It is usually denoted S.

Experiment 1: $S = \{H, T\}$; Experiment 2: $S = \{1, 2, 3, 4, 5, 6\}$
Experiment 3: $S = \{(1, 1), (1, 2), (1, 3), \ldots, (6, 6)\}$
Experiment 4: $S = \{0, 1, 2, \ldots n\}$ or $S = \{0, 1, 2, \ldots\}$
Experiment 5: $S = [0, +\infty)$
Experiment 6: $S = \{$sets of n individuals out of the population$\}$

Each element of the sample space S, that is each possible outcome of the experiment, is a simple event, generically denoted ω.

From the above examples, the distinction between discrete (finite or countable) and continuous sample spaces is clear.

Events

Often we are interested in a collection of related outcomes from a random experiment, that is a subset of the sample space, which has some physical reality.

Definition

An **event** E is a subset of the sample space of a random experiment

Examples of events:

Experiment 1: $E_1 = \{H\}$ = “the coin shows up Heads”

Experiment 2: $E_2 = \{2, 4, 6\}$ = “the die shows up an even number”

Experiment 3: $E_3 = \{(1, 3), (2, 2), (3, 1)\}$ = “the sum of the dice is 4”

Experiment 4: $E_4 = \{0, 1\}$ = “there is at most one defective item”

Experiment 5: $E_5 = [1, 2]$ = “the current is between 1 and 2 A”

If the outcome of the experiment is contained in E, then we say that E has occurred.

Events

The elements of interest are the events, which are (sub)sets

$\Rightarrow$ basic concepts of set theory will be useful

Set notation

- Union $E_1 \cup E_2$ = event “either E_1 **or** E_2 occurs”
- Intersection $E_1 \cap E_2$ = event “both E_1 **and** E_2 occur”
- Complement E^c = event “E does **not** occur” $(= E')$

S is an event $\rightsquigarrow$ certain event $S^c \doteq \phi \rightsquigarrow$ impossible event

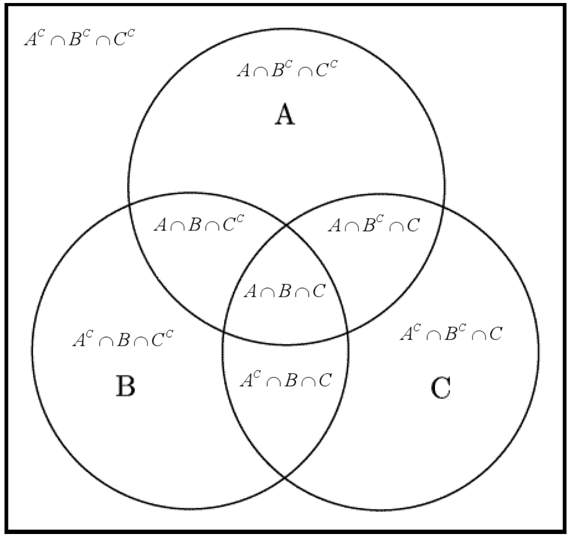
$E_1 \subseteq E_2 \Rightarrow E_1$ implies E_2

$E_1 \cap E_2 = \phi \Rightarrow$ mutually exclusive events (they cannot occur together)

De Morgan’s laws: $(E_1 \cup E_2)^c = E_1^c \cap E_2^c$
$(E_1 \cap E_2)^c = E_1^c \cup E_2^c$

These relations can be clearly illustrated by means of Venn diagrams.

Venn diagram



The axioms of probability theory

Intuitively, the probability $\mathbb{P}(E)$ of an event E is a number which should measure

how likely E is to occur

Firm mathematical footing ⇒ Kolmogorov's axioms (1933)

Kolmogorov's probability axioms

The probability measure $\mathbb{P}(\cdot)$ satisfies:

i) $0 \leq \mathbb{P}(E) \leq 1$ for any event E

ii) $\mathbb{P}(S) = 1$

iii) for any (infinite) sequence of mutually exclusive events $E_1, E_2, \ldots,$

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(E_i)$$

Useful implications of the axioms

- For any finite sequence of mutually exclusive events $E_1, E_2, \ldots, E_n$,

$$\mathbb{P}\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} \mathbb{P}(E_i)$$

- $\mathbb{P}(E^c) = 1 - \mathbb{P}(E)$
- $\mathbb{P}(\phi) = 0$
- $E_1 \subseteq E_2 \Rightarrow \mathbb{P}(E_1) \leq \mathbb{P}(E_2)$ (increasing measure)
- $\mathbb{P}(E_1 \cup E_2) = \mathbb{P}(E_1) + \mathbb{P}(E_2) - \mathbb{P}(E_1 \cap E_2)$, and by induction:
- Additive Law of Probability (or inclusion/exclusion principle)

$$\mathbb{P}\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} \mathbb{P}(E_i) - \sum_{i<j} \mathbb{P}(E_i \cap E_j) + \sum_{i<j<k} \mathbb{P}(E_i \cap E_j \cap E_k) + \ldots + (-1)^{n-1}\mathbb{P}\left(\bigcap_{i=1}^{n} E_i\right)$$

Assigning probabilities

Note:

The axioms state only the conditions an assignment of probabilities must satisfy, but they do not tell how to assign specific probabilities to events.

Experiment 1 (ctd.)

- experiment: flipping a coin ⇒ $S = \{H, T\}$
- Axioms: $\begin{cases} i) & 0 \leq \mathbb{P}(H) \leq 1, 0 \leq \mathbb{P}(T) \leq 1 \\ ii) & \mathbb{P}(S) = \mathbb{P}(H \cup T) = 1 \\ iii) & \mathbb{P}(H \cup T) = \mathbb{P}(H) + \mathbb{P}(T) \end{cases}$

⇒ The axioms only state that $\mathbb{P}(H)$ and $\mathbb{P}(T)$ are two non-negative numbers such that $\mathbb{P}(H) + \mathbb{P}(T) = 1$, nothing more!

The exact values of $\mathbb{P}(H)$ and $\mathbb{P}(T)$ depend on the coin itself (fair, biased, fake).

Assigning probabilities

To effectively assign probabilities to events, different approaches can be used, the most widely held being the **frequentist approach**.

Frequentist definition of probability

If the experiment is repeated independently over and over again (infinitely many times), the proportion of times that event E occurs is its probability $\mathbb{P}(E)$.

Let n be the number of repetitions of the experiment. Then, the probability of the event E is

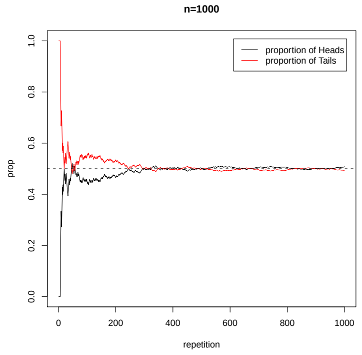
$$\mathbb{P}(E) = \lim_{n\to\infty} \frac{\text{number of times } E \text{ occurs}}{n}$$

Assigning probabilities: a simple example

Experiment 1 (ctd.)

- experiment: flipping a coin $\Rightarrow S = \{H, T\}$
- Axioms: $\begin{cases} i) & 0 \le \mathbb{P}(H) \le 1, 0 \le \mathbb{P}(T) \le 1 \\ ii) & \mathbb{P}(S) = \mathbb{P}(H \cup T) = 1 \\ iii) & \mathbb{P}(H \cup T) = \mathbb{P}(H) + \mathbb{P}(T) \end{cases}$

The coin is tossed n times, we observe the proportion of H and T:



$\Rightarrow \mathbb{P}(H) = \mathbb{P}(T) = \frac{1}{2}$

(fair coin, in this case)

Assigning probabilities

Interpretation

probability $\simeq$ proportion of occurrences of the event

- It is straightforward to check that the so-defined ‘frequentist’ probability measure satisfies the axioms
- Of course, this definition remains theoretical, as assigning probabilities would require infinitely many repetitions of the experiment
- Besides, in many situations, the experiment cannot be faithfully replicated (What is the probability that it will rain tomorrow? What is the probability of finding oil in that region?)

$\Rightarrow$ Essentially, assigning probabilities in practice relies on prior knowledge of the experimenter (belief and/or model)

- A simple model assumes that all the outcomes are equally likely, other more elaborated models define probability distributions ($\rightsquigarrow$ Chapter 5)

Assigning probabilities: equally likely outcomes

Assuming that all the outcomes of the experiment are equally likely provides an important simplification.

Suppose there are N possible outcomes $\{\omega_1, \omega_2, \dots, \omega_N\}$, equally likely to one another, $\mathbb{P}(\omega_k) = p$ for all k.

Then, Axioms 2 and 3 impose $p + p + \ldots + p = Np = 1$, that is,

$$p = \frac{1}{N}.$$

$\Rightarrow$ For an event E made up of k simple events, it follows from Axiom 3

$$\mathbb{P}(E) = \frac{k}{N} = \frac{\text{number of favourable cases}}{\text{total number of cases}}$$

$\Rightarrow$ “Classical” definition of probability

$\Rightarrow$ It is necessary to be able to effectively count the number of different ways that a given event can occur ($\Rightarrow$ combinatorics)

Basic combinatorics rules

- **Multiplication rule**: If an operation can be described as a sequence of k steps, and the number of ways of completing step i is n_i, then the total number of ways of completing the operation is

$$n_1 \times n_2 \times \ldots \times n_k$$

- **Permutations**: a permutation of the elements of a set is an ordered sequence of those elements. The number of different permutations of n elements is

$$P_n = n \times (n-1) \times (n-2) \times \ldots \times 2 \times 1 = n!$$

- **Combinations**: a combination is a subset of elements selected from a larger set. The number of combinations of size r that can be selected from a set of n elements is

$$\binom{n}{r} = C_r^n = \frac{n!}{r!(n-r)!}$$

Equally likely outcomes: example

Example

A computer system uses passwords that are 6 characters and each character is one of the 26 letters (a-z) or 10 integers (0-9). Uppercase letters are not used. Let A the event that a password begins with a vowel (either a, e, i, o or u) and let B denote the event that a password ends with an even number (either 0, 2, 4, 6 or 8). Suppose a hacker selects a password at random. What are the probabilities $\mathbb{P}(A)$, $\mathbb{P}(B)$, $\mathbb{P}(A \cap B)$ and $\mathbb{P}(A \cup B)$?

All passwords are equally likely to be selected → classical definition of probability → total number of cases $= 36^6 = 2,176,782,336$

$$\mathbb{P}(A) = \frac{5 \times 36^5}{36^6} = \frac{5}{36} = 0.1389 \qquad \mathbb{P}(B) = \frac{36^5 \times 5}{36^6} = \frac{5}{36} = 0.1389$$

$$\mathbb{P}(A \cap B) = \frac{5 \times 36^4 \times 5}{36^6} = \frac{25}{36^2} = 0.0193$$

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) = 2 \times 0.1389 - 0.0193 = 0.2585$$

Equally likely outcomes: example

Example: the birthday problem

If n people are present in a room, what is the probability that at least two of them celebrate their birthday on the same day of the year ? How large need n to be so that this probability is more than 1/2 ?

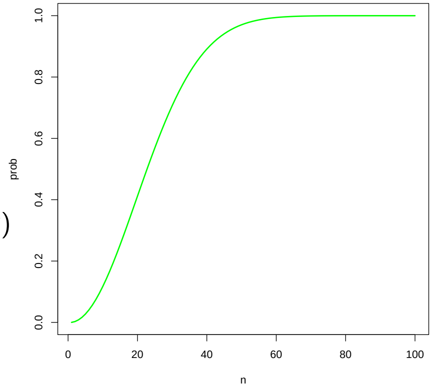
We have:

$$\mathbb{P}(\text{all birthdays are different}) = \frac{\binom{365}{n} n!}{365^n},$$

so that

$$\mathbb{P}(\text{at least two have the same birthday}) = 1 - \frac{\binom{365}{n} n!}{365^n}$$

$\Rightarrow$ Prob $> 1/2 \iff n > 23$



Conditional probabilities: definition

Sometimes probabilities need to be re-evaluated as additional information becomes available

$\Rightarrow$ this gives rise to the concept of conditional probability

Definition

The **conditional probability** of E_1, conditional on E_2, is defined as

$$\mathbb{P}(E_1|E_2) = \frac{\mathbb{P}(E_1 \cap E_2)}{\mathbb{P}(E_2)} \qquad (\text{if } \mathbb{P}(E_2) > 0)$$

= probability of E_1, given that E_2 has occurred

$\Rightarrow$ As we know that E_2 has occurred, E_2 becomes the new sample space in the place of S

$\Rightarrow$ The probability of E_1 has to be calculated within E_2 and relative to $\mathbb{P}(E_2)$

Conditional probabilities: properties

- $\mathbb{P}(E_1|E_2) =$ **probability of E_1**, (given some extra information)

  → satisfies the axioms of probability

  e.g. $\mathbb{P}(S|E_2) = 1$, or $\mathbb{P}(E_1^c|E_2) = 1 - \mathbb{P}(E_1|E_2)$
- $\mathbb{P}(E_1|S) = \mathbb{P}(E_1)$
- $\mathbb{P}(E_1|E_1) = 1, \qquad \mathbb{P}(E_1|E_2) = 1$ if $E_2 \subseteq E_1$
- $\mathbb{P}(E_1|E_2) \times \mathbb{P}(E_2) = \mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_2|E_1) \times \mathbb{P}(E_1)$

  ⇒ Bayes' first rule: if $\mathbb{P}(E_1) > 0$ and $\mathbb{P}(E_2) > 0$,

$$\mathbb{P}(E_1|E_2) = \mathbb{P}(E_2|E_1) \times \frac{\mathbb{P}(E_1)}{\mathbb{P}(E_2)}$$

- Multiplicative Law of Probability:

$$\mathbb{P}\left(\bigcap_{i=1}^{n} E_i\right) = \mathbb{P}(E_1)\times\mathbb{P}(E_2|E_1)\times\mathbb{P}(E_3|E_1\cap E_2)\times\ldots\times\mathbb{P}\left(E_n \middle| \bigcap_{i=1}^{n-1} E_i\right)$$

Example

A bin contains 5 defective, 10 partially defective and 25 acceptable transistors. Defective transistors immediately fail when put in use, while partially defective ones fail after a couple of hours of use. A transistor is chosen at random from the bin and put into use. If it does not immediately fail, what is the probability it is acceptable?

Define the following events:

- $A =$ the selected transistor is acceptable $\quad \to \mathbb{P}(A) = \frac{25}{40}$
- $PD =$ it is partially defective $\quad \to \mathbb{P}(PD) = \frac{10}{40}$
- $D =$ it is defective $\quad \to \mathbb{P}(D) = \frac{5}{40}$
- $F =$ it fails immediately $\quad \to \mathbb{P}(F) = \mathbb{P}(D) = \frac{5}{40}$

Now,

$$\mathbb{P}(A|F^c) = \mathbb{P}(F^c|A) \times \frac{\mathbb{P}(A)}{\mathbb{P}(F^c)} = 1 \times \frac{25/40}{1 - 5/40} = \frac{25}{35}$$

Example

A computer system has 3 users, each with a unique name and password. Due to a software error, the 3 passwords have been randomly permuted internally. Only the users lucky enough to have had their passwords unchanged in the permutation are able to continue using the system. What is the probability that none of the three users kept their original password?

Denote $A =$"no user kept their original password", and $E_i =$ "the ith user has the same password" ($i = 1, 2, 3$). See that

$$A^c = E_1 \cup E_2 \cup E_3,$$

for $A^c =$ at least one user has kept their original password. By the Additive Law of Probability,

$$\begin{aligned}\mathbb{P}(E_1 \cup E_2 \cup E_3) = {} & \mathbb{P}(E_1) + \mathbb{P}(E_2) + \mathbb{P}(E_3) - \mathbb{P}(E_1 \cap E_2) \\ & - \mathbb{P}(E_1 \cap E_3) - \mathbb{P}(E_2 \cap E_3) + \mathbb{P}(E_1 \cap E_2 \cap E_3).\end{aligned}$$

Clearly, for $i = 1, 2, 3$

$$\mathbb{P}(E_i) = 1/3$$

(each user gets a password at random out of 3, including their own).

From the Multiplicative Law of Probability,

$$\mathbb{P}(E_i \cap E_j) = \mathbb{P}(E_j|E_i) \times \mathbb{P}(E_i) \qquad \text{for any } i \neq j$$

Now, given E_i, that is knowing that the ith user has got their own password, there remain two passwords that the jth user may select, one of these two being their own. So

$$\mathbb{P}(E_j|E_i) = 1/2$$

and

$$\mathbb{P}(E_i \cap E_j) = 1/6.$$

Likewise, given $E_1 \cap E_2$, that is knowing that the first two users have kept their own passwords, there is only one password left, the one of the third user, and

$$\mathbb{P}(E_3|E_1 \cap E_2) = 1$$

so that (again Multiplicative Law of Probability)

$$\mathbb{P}(E_1 \cap E_2 \cap E_3) = \mathbb{P}(E_3|E_1 \cap E_2) \times \mathbb{P}(E_2|E_1) \times \mathbb{P}(E_1) = 1/6.$$

Finally,

$$\mathbb{P}(E_1 \cup E_2 \cup E_3) = 3 \times 1/3 - 3 \times 1/6 + 1/6 = 2/3$$

and

$$\mathbb{P}(A) = 1 - \mathbb{P}(E_1 \cup E_2 \cup E_3) = 1/3.$$

Independence of two events

Definition

Two events E_1 and E_2 are said to be **independent** if and only if

$$\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1) \times \mathbb{P}(E_2)$$

Note that independence implies

$$\mathbb{P}(E_1|E_2) = \mathbb{P}(E_1) \qquad \text{and} \qquad \mathbb{P}(E_2|E_1) = \mathbb{P}(E_2)$$

i.e. the probability of the occurrence of one of the event is unaffected by the occurrence or the non-occurrence of the other

→ in agreement with everyday usage of the word "independent" ("no link" between E_1 and E_2)

Caution: the 'simplified' multiplicative rule $\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1) \times \mathbb{P}(E_2)$ can only be used to assign a probability to $\mathbb{P}(E_1 \cap E_2)$ if E_1 and E_2 are independent, which can be known only from a fundamental understanding of the random experiment.

Example

We toss two fair dice, denote $E_1 =$"the sum of the dice is six", $E_2 =$"the sum of the dice is seven" and $F =$"the first die shows four". Are E_1 and F independent? Are E_2 and F independent?

Recall that $S = \{(1,1),(1,2),(1,3),\ldots,(6,5),(6,6)\}$ (there are thus 36 possible outcomes).

$E_1 = \{(1,5),(2,4),(3,3),(4,2),(5,1)\}$	$\mathbb{P}(E_1) = 5/36$
$E_2 = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$	$\mathbb{P}(E_2) = 6/36$
$F = \{(4,1),(4,2),(4,3),(4,4),(4,5),(4,6)\}$	$\mathbb{P}(F) = 6/36$
$E_1 \cap F = \{(4,2)\}$	$\mathbb{P}(E_1 \cap F) = 1/36$
$E_2 \cap F = \{(4,3)\}$,	$\mathbb{P}(E_2 \cap F) = 1/36$

Hence, $\mathbb{P}(E_1 \cap F) \neq \mathbb{P}(E_1)\mathbb{P}(F)$ and $\mathbb{P}(E_2 \cap F) = \mathbb{P}(E_2)\mathbb{P}(F)$

⇒ E_2 and F are independent, but E_1 and F are not.

Independence of more than two events

Definition

The events $E_1, E_2, \ldots, E_n$ are said to be independent iff for every subset $\{i_1, i_2, \ldots, i_r : r \leq n\}$ of $\{1, 2, \ldots, n\}$,

$$\mathbb{P}\left(\bigcap_{j=1}^{r} E_{i_j}\right) = \prod_{j=1}^{r} \mathbb{P}(E_{i_j})$$

For instance, E_1, E_2 and E_3 are independent iff

$\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1) \times \mathbb{P}(E_2)$,

$\mathbb{P}(E_1 \cap E_3) = \mathbb{P}(E_1) \times \mathbb{P}(E_3)$,

$\mathbb{P}(E_2 \cap E_3) = \mathbb{P}(E_2) \times \mathbb{P}(E_3)$ and

$\mathbb{P}(E_1 \cap E_2 \cap E_3) = \mathbb{P}(E_1) \times \mathbb{P}(E_2) \times \mathbb{P}(E_3)$

Remark

Pairwise independent events need not be jointly independent !

Example

Let a ball be drawn totally at random from an urn containing four balls numbered 1,2,3,4. Let $E = \{1,2\}$, $F = \{1,3\}$ and $G = \{1,4\}$.

Because the ball is selected at random, $\mathbb{P}(E) = \mathbb{P}(F) = \mathbb{P}(G) = 1/2$, and

$$\mathbb{P}(E \cap F) = \mathbb{P}(E \cap G) = \mathbb{P}(F \cap G) = \mathbb{P}(E \cap F \cap G) = \mathbb{P}(\{1\}) = 1/4.$$

So, $\mathbb{P}(E \cap F) = \mathbb{P}(E) \times \mathbb{P}(F)$, $\mathbb{P}(E \cap G) = \mathbb{P}(E) \times \mathbb{P}(G)$ and $\mathbb{P}(F \cap G) = \mathbb{P}(F) \times \mathbb{P}(G)$,

but $\mathbb{P}(E \cap F \cap G) \neq \mathbb{P}(E) \times \mathbb{P}(F) \times \mathbb{P}(G)$

The events E, F, G are pairwise independent, but they are not jointly independent

⇒ knowing that one event happened does not affect the probability of the others, but knowing that 2 events simultaneously happened does affect the probability of the third one

Example

Let a ball be drawn totally at random from an urn containing 8 balls numbered 1,2,3,. . . ,8. Let $E = \{1, 2, 3, 4\}$, $F = \{1, 3, 5, 7\}$ and $G = \{1, 4, 6, 8\}$.

It is clear that $\mathbb{P}(E) = \mathbb{P}(F) = \mathbb{P}(G) = 1/2$, and

$$\mathbb{P}(E \cap F \cap G) = \mathbb{P}(\{1\}) = 1/8 = \mathbb{P}(E) \times \mathbb{P}(F) \times \mathbb{P}(G),$$

but

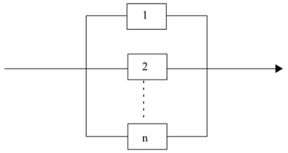
$$\mathbb{P}(F \cap G) = \mathbb{P}(\{1\}) = 1/8 \neq \mathbb{P}(F) \times \mathbb{P}(G)$$

Hence, the events E, F, G are not independent, though

$$\mathbb{P}(E \cap F \cap G) = \mathbb{P}(E) \times \mathbb{P}(F) \times \mathbb{P}(G)$$

Example

An electric system composed of n separate components is said to be a parallel system if it functions when at least one of the components functions. For such a system, if component i, **independently** of other components, functions with probability p_i, $i = 1, \ldots, n$, what is the probability the system functions?



Define the events W = the system functions and W_i = component i functions

Then, $\mathbb{P}(W^c) = \mathbb{P}(W_1^c \cap W_2^c \cap \ldots \cap W_n^c) = \prod_{i=1}^{n} \mathbb{P}(W_i^c) = \prod_{i=1}^{n} (1 - p_i)$, hence

$$\mathbb{P}(W) = 1 - \prod_{i=1}^{n} (1 - p_i)$$

Example: falsely signalling a pollution problem

Many companies must monitor the effluent that is discharged from their plants in waterways. It is the law that some substances have water-quality limits that are below some limit L. The effluent is judged to satisfy the limit if every test specimen is below L. Suppose the water does not contain the contaminant but that the variability in the chemical analysis still gives a 1% chance that a measurement on a test specimen will exceed L.
a) Find the probability that neither of two test specimens, both free of the contaminant, will fail to be in compliance

If the two samples are not taken too closely in time or space, we can treat them as independent. Denote E_i ($i = 1, 2$) the event "the sample i fails to be in compliance". It follows

$$\mathbb{P}(E_1^c \cap E_2^c) = \mathbb{P}(E_1^c) \times \mathbb{P}(E_2^c) = 0.99 \times 0.99 = 0.9801$$

Example: falsely signalling a pollution problem

Many companies must monitor the effluent that is discharged from their plants in waterways. It is the law that some substances have water-quality limits that are below the limit L. The effluent is judged to satisfy the limit if every test specimen is below L. Suppose the water does not contain the contaminant but that the variability in the chemical analysis still gives a 1% chance that a measurement on a test specimen will exceed L.
b) If one test specimen is taken each week for two years (all free of the contaminant), find the probability that none of the test specimens will fail to be in compliance, and comment.

Treating the results for different weeks as independent,

$$\mathbb{P}\left(\bigcap_{i=1}^{104} E_i^c\right) = \prod_{i=1}^{104} \mathbb{P}(E_i^c) = 0.99^{104} = 0.35$$

→ even with excellent water quality, there is almost a two-thirds chance that at least once the water quality will be declared to fail to be in compliance with the law

Bottom line: any event with positive probability will eventually occur, if we keep repeating the experiment!

Example

The supervisor of a group of 20 construction workers wants to get the opinion of 2 of them (to be selected at random) about certain new safety regulations. If 12 workers favour the new regulations and the other 8 are against them, what is the probability that both of the workers chosen by the supervisor will be against the new regulations?

Denote E_i $(i = 1, 2)$ the event "the ith selected worker is against the new regulations". We desire $\mathbb{P}(E_1 \cap E_2)$

However, E_1 and E_2 are not independent! (whether the first worker is against the regulations or not affects the proportion of workers against the regulations when the second one is selected)

So, $\mathbb{P}(E_1 \cap E_2) \neq \mathbb{P}(E_1)\mathbb{P}(E_2)$, but (by the multiplicative law of probability)

$$\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1)\mathbb{P}(E_2|E_1) = \frac{8}{20}\,\frac{7}{19} = \frac{14}{95} \simeq 0.147$$

(if E_1 has occurred, then for the second selection it remains 19 workers including 7 who are against the new regulations)

Partition

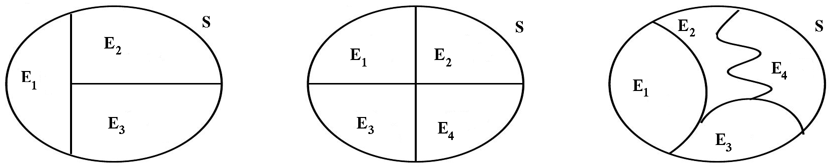
Definition

A sequence of events E_1, E_2, …, E_n such that

1. $S = \bigcup_{i=1}^n E_i$ and
2. $E_i \cap E_j = \phi$ for all $i \neq j$ (mutually exclusive),

is called a partition of S.

Some examples:

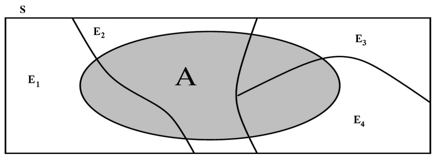


Simplest partition is $\{E, E^c\}$, for any event E

Law of Total Probability

From a partition $\{E_1, E_2, \ldots, E_n\}$, any event A can be written

$$A = (A \cap E_1) \cup (A \cap E_2) \cup \ldots \cup (A \cap E_n)$$



$$\Rightarrow \mathbb{P}(A) = \mathbb{P}(A \cap E_1) + \mathbb{P}(A \cap E_2) + \ldots + \mathbb{P}(A \cap E_n)$$

Law of Total Probability

Given a partition $\{E_1, E_2, \ldots, E_n\}$ of S such that $\mathbb{P}(E_i) > 0$ for all i, the probability of any event A can be written

$$\mathbb{P}(A) = \sum_{i=1}^n \mathbb{P}(A|E_i) \times \mathbb{P}(E_i)$$

Bayes' second rule

In particular, for any event A and any event E such that $0 < \mathbb{P}(E) < 1$, we have

$$\mathbb{P}(A) = \mathbb{P}(A|E)\mathbb{P}(E) + \mathbb{P}(A|E^c)(1 - \mathbb{P}(E))$$

Now, put the Law of Total Probability in Bayes' first rule and get

Bayes' second rule

Given a partition $\{E_1, E_2, \ldots, E_n\}$ of S such that $\mathbb{P}(E_i) > 0$ for all i, we have, for any event A such that $\mathbb{P}(A) > 0$,

$$\mathbb{P}(E_i|A) = \frac{\mathbb{P}(A|E_i)\mathbb{P}(E_i)}{\sum_{j=1}^n \mathbb{P}(A|E_j)\mathbb{P}(E_j)}$$

In particular:

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

Example

A new medical procedure has been shown to be effective in the early detection of an illness and a medical screening of the population is proposed. The probability that the test correctly identifies someone with the illness as positive is 0.99, and the probability that someone without the illness is correctly identified by the test is 0.95. The incidence of the illness in the general population is 0.0001. You take the test, and the result is positive. What is the probability that you have the disease?

Let $I =$ event that you have the illness, T = positive outcome of the screening test for illness. From the question we have,

$$\mathbb{P}(T|I) = 0.99, \quad \mathbb{P}(T^C|I^C) = 0.95, \quad \mathbb{P}(I) = 0.0001.$$

We aim to find $\mathbb{P}(I|T)$, using bayes second rule,

$$\begin{aligned}\mathbb{P}(I|T) &= \frac{\mathbb{P}(T|I)\mathbb{P}(I)}{\mathbb{P}(T|I)\mathbb{P}(I) + \mathbb{P}(T|I^C)P(I^C)} \\ &= \frac{0.99 \times 0.0001}{0.99 \times 0.0001 + 0.05 \times 0.9999} = 0.001976\end{aligned}$$

Example

Suppose a multiple choice test, with m multiple-choice alternatives for each question. A student knows the answer of a given question with probability p. If she does not know, she guesses. Given that the student correctly answered a question, what is the probability that she effectively knew the answer?

Let $C =$"she answers the question correctly" and $K =$"she knows the answer". Then, we desire $\mathbb{P}(K|C)$. We have

$$\begin{aligned}\mathbb{P}(K|C) &= \mathbb{P}(C|K) \times \frac{\mathbb{P}(K)}{\mathbb{P}(C)} \\ &= \frac{\mathbb{P}(C|K) \times \mathbb{P}(K)}{\mathbb{P}(C|K) \times \mathbb{P}(K) + \mathbb{P}(C|K^c) \times \mathbb{P}(K^c)} \\ &= \frac{1 \times p}{1 \times p + (1/m) \times (1-p)} \\ &= \frac{mp}{1 + (m-1)p}\end{aligned}$$

Objectives

Now you should be able to:

- understand and describe sample spaces and events for random experiments □
- interpret probabilities and use probabilities of outcomes to calculate probabilities of events □
- use permutations and combinations to count the number of outcomes in both an event and the sample space □
- calculate the probabilities of joint events such as unions and intersections from the probabilities of individual events □
- interpret and calculate conditional probabilities of events □
- determine whether events are independent and use independence to calculate probabilities □
- use Baye's rule(s) to calculate probabilities □

Recommended exercises

→ Q1, Q2 p.197, Q8, Q9 p.203, Q12 p.209, Q17 p.210, Q19 p.210, Q20 p.210, Q59 p.238, Q61 p.238, Q73 p.240 (2nd edition)

→ Q1, Q2 p.200, Q8, Q9 p.207, Q13 p.213, Q18 p.214, Q20 p.214, Q21 p.214, Q61 p.242, Q63 p.243, Q76 p.244 (3rd edition)

4 Random variables

Introduction

Often, we are not interested in all of the details of an experiment but only in some numerical quantities determined by the outcome.

Example 1: tossing two dice when playing a board game

$S = \{(1,1),(1,2),\dots,(6,5),(6,6)\}$

... but often only the **sum** of the points matters

→ each possible outcome ω is characterised by a real number

Example 2: buying 2 items, each either defective or acceptable

$S = \{(d,d),(d,a),(a,d),(a,a)\}$

... but we might only be interested in the **number of acceptable items** obtained in the purchase

→ again, each possible outcome ω is characterised by a real number

It is often much more natural to directly think in terms of the numerical quantity of interest, called a **random variable**.

Random variable: definition

Definition

A random variable is a real-valued function defined over the sample space:

$$X : S \to \mathbb{R}$$
$$\omega \to X(\omega)$$

Usually*, a random variable is denoted by an uppercase letter.

Define S_X the domain of variation of X, that is the set of possible values taken by X.

Example 1: tossing two dice when playing a board game

$X =$ sum of the points, $S_X = \{2,3,4,\dots,12\}$

Example 2: buying 2 electronic items

$X =$ number of acceptable items, $S_X = \{0,1,2\}$

*except in your textbook

Events defined by random variables

For any fixed real value $x \in S_X$, assertions like "$X = x$" or "$X \leq x$" correspond to a set of possible outcomes

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

→ they are events ! → meaningful to talk about their probability

Example 1 (ctd.) - If the dice are fair

$(X = 2) = \{(1,1)\}$ $\to \mathbb{P}(X = 2) = 1/36$

$(X \geq 11) = \{(5,6),(6,5),(6,6)\}$ $\to \mathbb{P}(X \geq 11) = 3/36 = 1/12$

The usual properties of probabilities apply, e.g.

- $\mathbb{P}(X \in S_X) = 1$
- $\mathbb{P}((X = x_1) \cup (X = x_2)) = \mathbb{P}(X = x_1) + \mathbb{P}(X = x_2)$ (if $x_1 \neq x_2$)
- $\mathbb{P}(X < x) = 1 - \mathbb{P}(X \geq x)$ ('$X < x$' is the complement of '$X \geq x$')

Notes

Note 1

It is important not to confuse:

- X, the name of the random variable
- $X(\omega)$, the numerical value taken by the random variable at some sample point ω
- x, a generic numerical value

Note 2

Most interesting problems can be stated, often naturally, in terms of random variables.

→ Many inessential details about the sample space can be left unspecified, and one can still solve the problem

→ Often more helpful to think of random variables simply as variables whose values are likely to lie within certain ranges of the real number line

Cumulative distribution function

A random variable is often described by its **cumulative distribution function** (cdf) (or just distribution).

Definition

The cdf of the random variable X is defined for any real number x, by

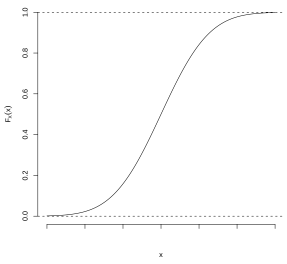
$$F(x) = \mathbb{P}(X \leq x)$$

All probability questions about X can be answered in terms of its distribution. We will denote $X \sim F$ (read 'X follows the distribution F').

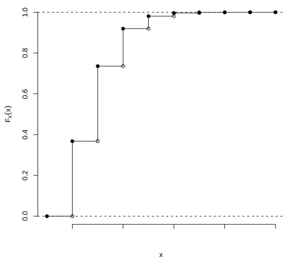
Some properties:

- For any $a \leq b$, $\mathbb{P}(a < X \leq b) = F(b) - F(a)$
- F is a nondecreasing function
- $\lim_{x \to +\infty} F(x) = F(+\infty) = 1$
- $\lim_{x \to -\infty} F(x) = F(-\infty) = 0$

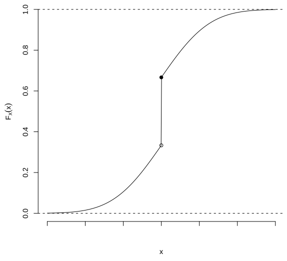
Cumulative distribution functions



Continuous distribution
→ continuous r.v.



Discrete distribution
→ discrete r.v.



Hybrid distribution
→ hybrid r.v.

Note: hybrid distributions will not be introduced in this course.

Discrete random variables

Definition

A random variable is said to be discrete if it can only assume a finite (or at most countably infinite) number of values.

Suppose that those values are $S_X = \{x_1, x_2, \ldots\}$.

Definition

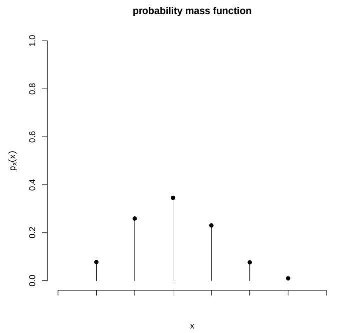
The probability mass function (pmf) of a discrete random variable X is defined for any real number x, by

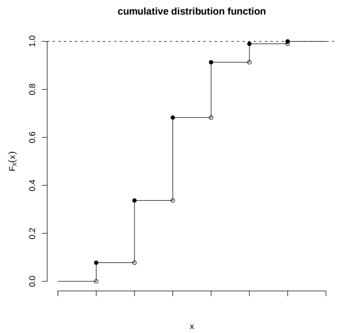
$$p(x) = \mathbb{P}(X = x)$$

→ $p_X(x) > 0$ for $x = x_1, x_2, \ldots$, and $p_X(x) = 0$ for any other value of x

Obviously:

$$\mathbb{P}(X \in S_X) = \mathbb{P}((X = x_1) \cup (X = x_2) \cup \ldots) = \sum_{x \in S_X} p(x) = 1$$





Probability mass function:

- "spikes" at $x_1, x_2, \ldots$
- height of spike at $x_i = p(x_i)$

Cumulative distribution function:

- $F(x) = \sum_{i:x_i \leq x} p(x_i)$
- step function
- jumps at $x_1, x_2, \ldots$
- magnitude of jump at $x_i = p(x_i)$

Discrete random variables: examples

Examples of discrete random variables include:

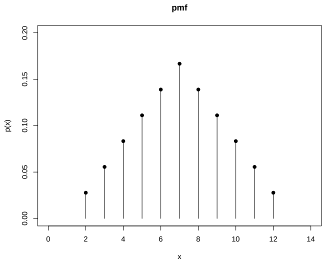
number of scratches on a surface, number of defective parts among 1000 tested, number of transmitted bits received in error, . . .

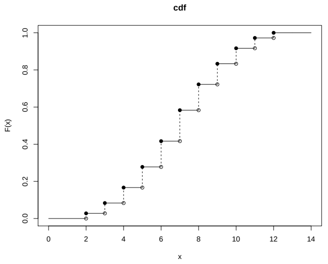
$\Rightarrow$ discrete random variables generally arise when we count things

Example: tossing 2 dice

X = sum of the points; show $p(x)$ and $F(x)$

Check that $p(x) = (6 - |7 - x|)/36$ for $x \in S_X = \{2, 3, 4, \ldots, 12\}$





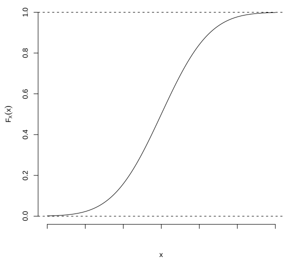
Bernoulli random variable

- Named after the Swiss scientist Jakob Bernoulli (1654-1705).
- That is the simplest random variable
- It can only assume 2 values, $S_X = \{0, 1\}$
- Its pmf is given by

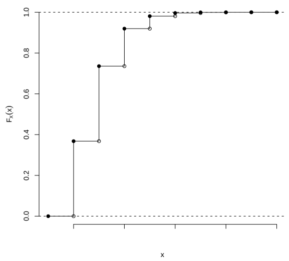
$$p(1) = \pi$$
$$p(0) = 1 - \pi$$

  for some value $\pi \in [0, 1]$
- It is often used to characterise the occurrence/non-occurrence of a given event, or the presence/absence of a given feature

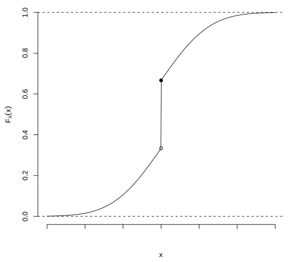
Cumulative distribution functions



Continuous distribution
→ continuous r.v.



Discrete distribution
→ discrete r.v.



Hybrid distribution
→ hybrid r.v.

Continuous random variables

As opposed to a discrete r.v., a continuous random variable X is expected to take on an uncountable number of values. S_X is therefore an uncountable set of real numbers (like an interval), and can even be $\mathbb{R}$ itself.

Definition

A random variable X is said to be continuous if there exists a nonnegative function $f(x)$ defined for all real $x \in \mathbb{R}$ such that for any set B of real numbers,

$$\mathbb{P}(X \in B) = \int_B f(x)dx$$

Consequence: $\mathbb{P}(X = x) = 0$ for any x !

$\rightarrow$ The probability mass function is useless

$\rightarrow$ The probability density function (pdf) $f(x)$ will play the central role

Continuous random variables: remark

Note 1: the fact that $\mathbb{P}(X = x) = 0$ for any x should not be disturbing

$\rightarrow$ coherent when dealing with measurements,

E.g. if we report a temperature of 74.8 degrees centigrade, owing to the limits of our ability to measure (accuracy of measuring devices), we really mean that the temperature lies “close to” 74.8, for instance between 74.75 and 74.85 degrees

Note 2: when we say that there is a zero probability that a random variable X will take on any value x, this does not mean that it is impossible that X will take on the value x!

In the continuous case, zero probability does not imply logical impossibility

$\rightarrow$ this should not be disturbing either, as we are always interested in probabilities connected with intervals and not with isolated points

Probability density function: properties

- $F(x) = \mathbb{P}(X \leq x) = \displaystyle\int_{-\infty}^{x} f(y)dy$, that is
  $$\boxed{f(x) = \frac{dF(x)}{dx} = F'(x)}$$
  (wherever F is differentiable)
- $\boxed{f(x) \geq 0}$ $\quad \forall x \in \mathbb{R} \qquad$ ($F(x)$ is nondecreasing)
- $\mathbb{P}(a \leq X \leq b) = \displaystyle\int_a^b f(x)dx = F(b) - F(a)$
- $\boxed{\displaystyle\int_{-\infty}^{+\infty} f(x)dx = 1}$
- For a small ε, $\mathbb{P}(x - \varepsilon/2 \leq X \leq x + \varepsilon/2) = \displaystyle\int_{x-\varepsilon/2}^{x+\varepsilon/2} f(y)dy \simeq \varepsilon f(x)$
  $\rightsquigarrow S_X = \{x \in \mathbb{R} : f(x) > 0\}$

Note: as $\mathbb{P}(X = x) = 0$, $\mathbb{P}(X < x) = \mathbb{P}(X \leq x)$ (for a continuous r.v.)

Continuous random variables: examples

Examples of continuous random variables include:
electrical current, length, pressure, temperature, time, voltage, weight, speed of a car, amount of alcohol in a person’s blood, efficiency of solar collector, strength of a new alloy, . . .

$\Rightarrow$ continuous r.v. generally arise when we measure things

Example

Let X denote the current measured in a thin copper wire (in mA). Assume that the pdf of X is

$$f(x) = \begin{cases} C(4x - 2x^2) & \text{if } 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

What is the value of C? Find $\mathbb{P}(X > 1.8)$

We must have $\int_{-\infty}^{+\infty} f(x)\,dx = 1$, so $C\int_0^2 (4x - 2x^2)\,dx = C \times \frac{8}{3} = 1$, that is $C = 3/8$

Then, $\mathbb{P}(X > 1.8) = \int_{1.8}^{+\infty} f(x)\,dx = 3/8 \times \int_{1.8}^{2} (4x - 2x^2)\,dx = 0.028$.

Discrete vs. Continuous random variables

	Discrete r.v.	**Continuous r.v.**
Domain of variation	$S_X = \{x_1, x_2, \dots\}$	$S_X = [\alpha, \beta] \subseteq \mathbb{R}$
Probability mass function (pmf)	$p(x) = \mathbb{P}(X = x) \geq 0$ for all $x \in \mathbb{R}$ ● $p(x) > 0$ if and only if $x \in S_X$ ● $\sum_{x \in S_X} p(x) = 1$	useless: $p(x) \equiv 0$
Probability density function (pdf)	does not exist	$f(x) = F'(x) \geq 0$ for all $x \in \mathbb{R}$ ● $f(x) > 0$ if and only if $x \in S_X$ ● $\int_{x \in S_X} f(x) = 1$

Note the similarity between the conditions for pmf and pdf.

Parameters of a distribution

Fact

Some quantities characterise a random variable more usefully (although incompletely) than the whole cumulative distribution function.

The two most important such quantities are:

- the expectation (or mean) and
- the variance

of a random variable

Often, we talk about the expectation or the variance of a distribution, understood as the expectation or the variance of a random variable having that distribution.

Expectation

The **expectation** or the **mean** of a random variable X, denoted $\mathbb{E}(X)$ or μ, is defined by

Discrete r.v.

$$\mu = \mathbb{E}(X) = \sum_{x \in S_X} x\, p(x)$$

Continuous r.v.

$$\mu = \mathbb{E}(X) = \int_{S_X} x\, f(x) dx$$

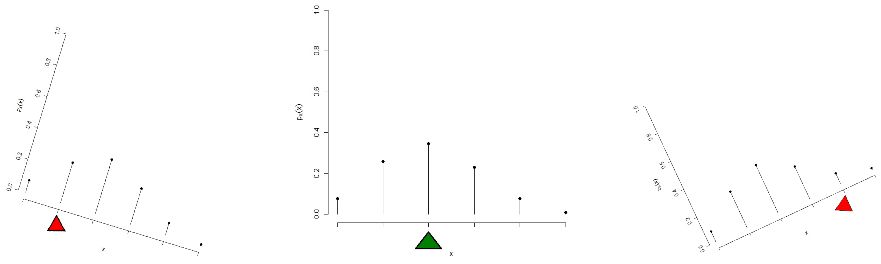
$\Rightarrow$ $\mathbb{E}(X)$ is a weighted average of the possible values of X, each value being weighted by the probability that X assumes it

Note: $\mathbb{E}(X)$ has the same units as X.

Expectation

Expectation = expected value, mean value, average value of X
= "central" value, around which X is distributed
= "centre of gravity" of the distribution

In the discrete case:



$=$

μ

$\rightarrow$ location parameter

Expectation: examples

Example 1

What is the expectation of the outcome when a fair die is rolled?

X = outcome, $S_X = \{1,2,3,4,5,6\}$ with $p(x) = 1/6$ for any $x \in S_X$

$$\begin{aligned}\mu = \mathbb{E}(X) &= 1 \times 1/6 + 2 \times 1/6 + 3 \times 1/6 + 4 \times 1/6 + 5 \times 1/6 + 6 \times 1/6 \\ &= 3.5\end{aligned}$$

→ μ need not be a possible outcome !

→ μ is not the most likely outcome (this is called the mode)

Expectation: examples

Example 2

What is the expected sum when two fair dice are rolled?

X = sum of the two dice,

$S_X = \{2,3,\ldots,12\}$ with

$p(x) = (6 - |7-x|)/36$ for any $x \in S_X$

$\rightarrow \mu = \mathbb{E}(X) = 2 \times 1/36 + 3 \times 2/36 + \ldots + 12 \times 1/36 = 7$

Example 3: Bernoulli r.v. (see Slide 131)

What is the expectation of a Bernoulli r.v.?

$\mathbb{E}(X) = 0 \times (1-\pi) + 1 \times \pi = \pi$

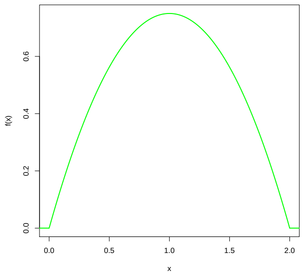
Expectation: examples

Example 4

Find the mean value of the copper current measurement X for Example on Slide 136, that is, with

$$f(x) = \begin{cases} \frac{3}{8}(4x - 2x^2) & \text{if } 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

The density is



By symmetry, it can be directly concluded that $\mu = 1$ mA

It can also be easily checked that

$$\begin{aligned}\mu = \mathbb{E}(X) &= \int_{-\infty}^{+\infty} x\, f(x)\, dx \\ &= \frac{3}{8}\int_0^2 x\,(4x - 2x^2)\, dx \\ &= 1\end{aligned}$$

Expectation of a function of a random variable

Sometimes we are not interested in the expected value of X, but in the expected value of a function of X, say $g(X)$.

There is actually no need for explicitly deriving the distribution of $g(X)$. Indeed, it can be shown

If X is a discrete r.v.

$$\mathbb{E}(g(X)) = \sum_{x \in S_X} g(x)\, p(x)$$

If X is a continuous r.v.

$$\mathbb{E}(g(X)) = \int_{S_X} g(x)\, f(x)\, dx$$

In particular, for 2 constants a and b:

Linear transformation

$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$$

With $a = 0 \rightarrow \mathbb{E}(b) = b$ (“degenerate” random variable)

Variance of a random variable

Definition

The **variance** of a random variable X, usually denoted by $\mathbb{V}\mathrm{ar}(X)$ or σ^2, is defined by

$$\sigma^2 = \mathbb{V}\mathrm{ar}(X) = \mathbb{E}\left((X-\mu)^2\right)$$

Clearly, $\boxed{\mathbb{V}\mathrm{ar}(X) \geq 0}$

If X is a discrete r.v.

$$\sigma^2 = \mathbb{V}\mathrm{ar}(X) = \sum_{x \in S_X} (x-\mu)^2 p(x)$$

If X is a continuous r.v.

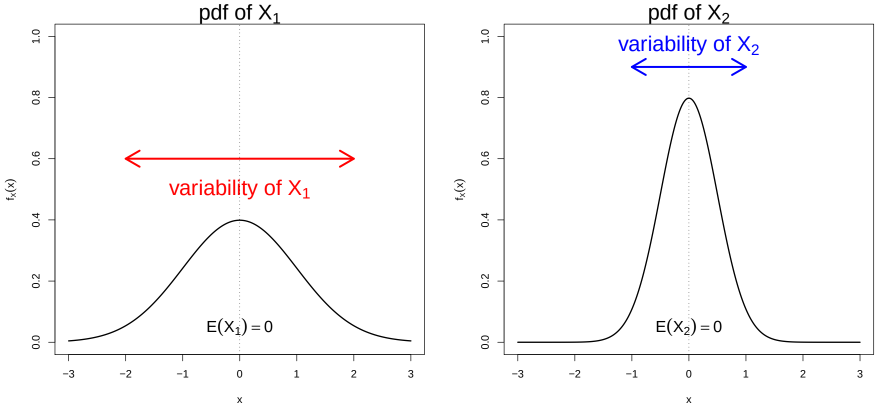
$$\sigma^2 = \mathbb{V}\mathrm{ar}(X) = \int_{S_X} (x-\mu)^2 f(x)dx$$

→ Expected square of the deviation of X from its expected value

→ The variance quantifies the dispersion of the possible values of X around the “central” value μ, that is, the variability of X

Variance: illustration

Two (continuous) random variables X_1 and X_2, with $\mathbb{E}(X_1) = \mathbb{E}(X_2)$



$\rightarrow \mathbb{V}\mathrm{ar}(X_1) > \mathbb{V}\mathrm{ar}(X_2)$

Variance: notes

Note 1

An alternative formula for $\mathbb{V}\mathrm{ar}(X)$ is the following:

$$\sigma^2 = \mathbb{V}\mathrm{ar}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \mathbb{E}(X^2) - \mu^2$$

Proof: . . . □

⇒ In practice, this is often the easiest way to compute $\mathbb{V}\mathrm{ar}(X)$, using

$$\mathbb{E}(X^2) = \sum_{x \in S_X} x^2 p(x) \quad \text{or} \quad \mathbb{E}(X^2) = \int_{S_X} x^2 f(x)dx$$

Note 2

The variance σ^2 is in ‘square units’ of X, which may make interpretation difficult.

⇒ often, we adjust for this by taking the square root of σ^2

This is called the **standard deviation** σ of X: $\sigma = \sqrt{\sigma^2} = \sqrt{\mathbb{V}\mathrm{ar}(X)}$

Variance: linear transformation

For any constants a and b, we have

Linear transformation

$$\mathbb{V}\mathrm{ar}(aX+b) = a^2\,\mathbb{V}\mathrm{ar}(X)$$

Take $a = 1$, it follows that for any b, $\boxed{\mathbb{V}\mathrm{ar}(X+b) = \mathbb{V}\mathrm{ar}(X)}$

→ variance not affected by translation

Take $a = 0$, if follows that for any b, $\boxed{\mathbb{V}\mathrm{ar}(b) = 0}$

(“degenerate” random variable)

Variance : examples

Example 1

What is the variance of the number of points shown when a fair die is rolled?

X = outcome, $S_X = \{1, 2, 3, 4, 5, 6\}$ with $p(x) = 1/6$ for any $x \in S_X$

$$\begin{aligned}\mathbb{E}(X^2) &= 1^2 \times 1/6 + 2^2 \times 1/6 + 3^2 \times 1/6 + 4^2 \times 1/6 + 5^2 \times 1/6 + 6^2 \times 1/6 \\ &= 91/6\end{aligned}$$

We know that $\mu = 3.5$ (Slide 141), so that

$$\sigma^2 = \mathbb{E}(X^2) - \mu^2 = 91/6 - 3.5^2 \simeq 2.92$$

The standard deviation is $\sigma = \sqrt{2.92} \simeq 1.71$

Example 2

What is the variance of the sum of the points when 2 fair dice are rolled?

(Exercise) Check that $\sigma^2 \simeq 5.83$, $\sigma \simeq 2.41$.

Variance: examples

Example 3

What is the variance of a Bernoulli r.v.?

$\mathbb{E}(X^2) = 0^2 \times (1 - \pi) + 1^2 \times \pi = \pi = \mathbb{E}(X)$

→ $\mathbb{V}\text{ar}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \pi - \pi^2 = \pi(1 - \pi)$

Example 4

What is the variance of the copper current measurement X for Example on Slide 136, that is, with

$$f(x) = \begin{cases} \frac{3}{8}(4x - 2x^2) & \text{if } 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

We have $\mathbb{E}(X^2) = \int_{-\infty}^{+\infty} x^2 f(x)\, dx = \frac{3}{8}\int_0^2 x^2(4x - 2x^2)\, dx = 1.2$

We know that $\mu = 1$ (Slide 24), so that $\sigma^2 = 1.2 - 1^2 = 0.2$ mA^2

→ $\sigma \simeq 0.45$ mA

Standardisation

Standardisation is a very useful linear transformation.

Suppose you have a random variable X with mean μ and variance σ^2. Then, the associated standardised random variable, often denoted Z, is given by

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

that is, $Z = \frac{1}{\sigma}X - \frac{\mu}{\sigma}$. Hence, using the linear transformations properties,

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

$$\mathbb{V}\text{ar}(Z) = \frac{1}{\sigma^2}\,\mathbb{V}\text{ar}(X) = \frac{\sigma^2}{\sigma^2} = 1$$

→ A standardised random variable has always mean 0 and variance 1.

Note 1: Z is a dimensionless variable (no unit)
Note 2: A standardised value of X is sometimes called z-score

Joint distribution function

Often, probability statements concerning two random variables, say X and Y, defined on the same sample space are of interest:

$$\omega \to (X(\omega), Y(\omega))$$

→ These two variables are most certainly related

→ They should be jointly analysed, in order to understand the degree of relationship between them

For instance, we may simultaneously measure the weight and hardness of a rock, the pressure and temperature of a gas, thickness and compressive strength of a piece of glass, etc.

Definition

The joint cumulative distribution function of X and Y is given by

$$F_{XY}(x, y) = \mathbb{P}(X \leq x, Y \leq y) \qquad \forall (x, y) \in \mathbb{R} \times \mathbb{R}$$

Note: here $(X \leq x, Y \leq y)$ means $(X \leq x) \cap (Y \leq y)$.

Joint distribution: discrete case

If X and Y are both discrete, the joint probability mass function is defined by

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

The marginal pmf of X and Y can be obtained by

$$p_X(x) = \sum_{y \in S_Y} p_{XY}(x,y) \quad \text{and} \quad p_Y(y) = \sum_{x \in S_X} p_{XY}(x,y)$$

Joint distribution: continuous case

Definition

X and Y are said to be jointly continuous if there exists a function $f_{XY}(x,y) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ such that for any sets A and B of real numbers

$$\mathbb{P}(X \in A, Y \in B) = \int_A \int_B f_{XY}(x,y)dy\,dx$$

The function $f_{XY}(x,y)$ is the joint probability density of X and Y.

The marginal densities follow from

$$\int_A f_X(x)dx = \mathbb{P}(X \in A) = \mathbb{P}(X \in A, Y \in S_Y) = \int_A \int_{S_Y} f_{XY}(x,y)dy\,dx$$

Thus,

$$f_X(x) = \int_{S_Y} f_{XY}(x,y)dy \quad \text{and} \quad f_Y(y) = \int_{S_X} f_{XY}(x,y)dx$$

Expectation of a function of two random variables

For any function $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, the expectation of $g(X,Y)$ is given by

$$\begin{aligned}\mathbb{E}(g(X,Y)) &= \sum_{x\ \in S_X} \sum_{y \in S_Y} g(x,y)p_{XY}(x,y) && \text{(discrete case)}\\ &= \int_{S_X} \int_{S_Y} g(x,y)f_{XY}(x,y)dy\,dx && \text{(continuous case)}\end{aligned}$$

Expectation of a function of two random variables

For instance, in the continuous case,

$$\begin{aligned}\mathbb{E}(aX + bY) &= \int_{S_X} \int_{S_Y} (ax + by)f_{XY}(x,y)dy\,dx\\ &= \int_{S_X} \int_{S_Y} ax\, f_{XY}(x,y)dy\,dx + \int_{S_X} \int_{S_Y} by\, f_{XY}(x,y)dy\,dx\\ &= a \int_{S_X} x \int_{S_Y} f_{XY}(x,y)dy\,dx + b \int_{S_Y} y \int_{S_X} f_{XY}(x,y)dx\,dy\\ &= a \int_{S_X} xf_X(x)dx + b \int_{S_Y} yf_Y(y)dy\\ &= a\mathbb{E}(X) + b\mathbb{E}(Y)\end{aligned}$$

Example

What is the expected sum obtained when two fair dice are rolled?

Let X be the sum and X_i the value shown on the ith die. Then, $X = X_1 + X_2$, and

$$\mathbb{E}(X) = \mathbb{E}(X_1) + \mathbb{E}(X_2) = 2 \times 3.5 = 7$$

Independent random variables

Definition

The random variables X and Y are said to be independent if, for all $(x, y) \in \mathbb{R} \times \mathbb{R}$,

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

In other words, X and Y are independent if all couples of events $(X \leq x)$ and $(Y \leq y)$ are independent.

Characterisation: For any $(x, y) \in \mathbb{R} \times \mathbb{R}$,

$$F_{XY}(x, y) = F_X(x) \times F_Y(y),$$

which reduces to

$$p_{XY}(x, y) = p_X(x) \times p_Y(y) \qquad \text{(discrete case)}$$

or

$$f_{XY}(x, y) = f_X(x) \times f_Y(y) \qquad \text{(continuous case)}$$

Independent random variables

Property

If X and Y are independent, then for any functions h and g,

$$\mathbb{E}(h(X)g(Y)) = \mathbb{E}(h(X)) \times \mathbb{E}(g(Y))$$

Proof (in the continuous case):

$$\begin{aligned}
\mathbb{E}(h(X)g(Y)) &= \iint_{S_X \times S_Y} h(x)g(y)f_{XY}(x, y)dy\,dx \\
&= \int_{S_X} \int_{S_Y} h(x)g(y)f_X(x)f_Y(y)dy\,dx \\
&= \int_{S_X} h(x)f_X(x)dx \times \int_{S_Y} g(y)f_Y(y)dy \\
&= \mathbb{E}(h(X)) \times \mathbb{E}(g(Y))
\end{aligned}$$

Covariance of two random variables

Definition

The covariance of two random variables X and Y is defined by

$$\mathbb{C}\text{ov}(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y)))$$

Properties:

- $\mathbb{C}\text{ov}(X, Y) = \mathbb{C}\text{ov}(Y, X)$
- $\mathbb{C}\text{ov}(X, X) = \mathbb{V}\text{ar}(X)$
- $\boxed{\mathbb{C}\text{ov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)}$
- $\mathbb{C}\text{ov}(aX + b, cY + d) = ac\,\mathbb{C}\text{ov}(X, Y)$
- $\mathbb{C}\text{ov}(X_1 + X_2, Y_1 + Y_2)$ $= \mathbb{C}\text{ov}(X_1, Y_1) + \mathbb{C}\text{ov}(X_1, Y_2) + \mathbb{C}\text{ov}(X_2, Y_1) + \mathbb{C}\text{ov}(X_2, Y_2)$

Note: unit of $\mathbb{C}\text{ov}(X, Y)$ = unit of X × unit of Y

Covariance: interpretation

Suppose X and Y are two Bernoulli random variables, and see that XY is then also a Bernoulli. It follows:

$$\mathbb{C}\text{ov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{P}(X = 1, Y = 1) - \mathbb{P}(X = 1)\mathbb{P}(Y = 1)$$

Then,

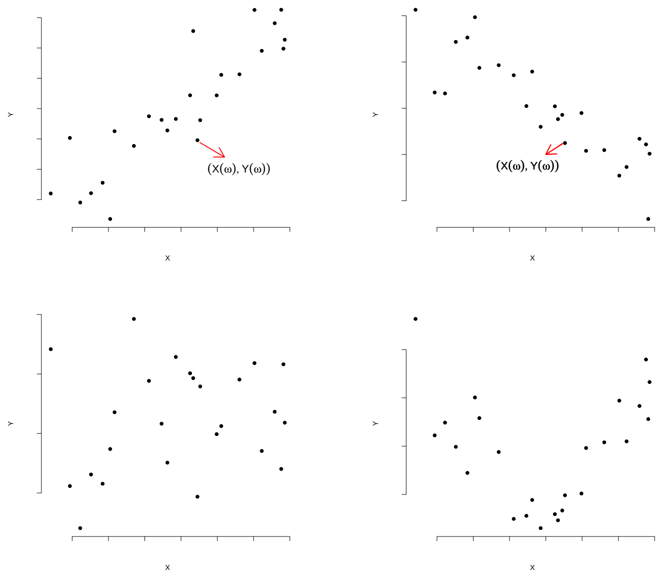
$$\begin{aligned}
\mathbb{C}\text{ov}(X, Y) > 0 &\Leftrightarrow \mathbb{P}(X = 1, Y = 1) > \mathbb{P}(X = 1)\mathbb{P}(Y = 1) \\
&\Leftrightarrow \frac{\mathbb{P}(X = 1, Y = 1)}{\mathbb{P}(X = 1)} > \mathbb{P}(Y = 1) \\
&\Leftrightarrow \mathbb{P}(Y = 1|X = 1) > \mathbb{P}(Y = 1)
\end{aligned}$$

$\rightsquigarrow$ the outcome $X = 1$ makes it more likely that $Y = 1$

$\rightsquigarrow$ *Y tends to increase when X does, and vice-versa*

This interpretation holds for any r.v. X and Y (not only for Bernoulli r.v.)

Covariance: interpretation



Covariance: interpretation

- $\mathbb{C}\mathrm{ov}(X,Y) > 0 \rightsquigarrow X$ and Y tend to increase or decrease together
- $\mathbb{C}\mathrm{ov}(X,Y) < 0 \rightsquigarrow X$ tends to increase as Y decreases and vice-versa
- $\mathbb{C}\mathrm{ov}(X,Y) = 0 \rightsquigarrow$ no linear association between X and Y (doesn't mean no association at all!)

Fact

$$X \text{ and } Y \text{ independent } \Rightarrow \mathbb{C}\mathrm{ov}(X,Y) = 0$$
$$\not\Leftarrow (X \text{ and } Y \text{ are } \textit{uncorrelated})$$

Covariance: examples

Example

Let the pmf of a r.v. X be $p_X(1) = p_X(-1) = p$ and $p_X(0) = 1-2p$ $(p \in (0,1/2))$. Define $Y = X^2$. Find $\mathbb{C}\mathrm{ov}(X,Y)$

We have $\mathbb{C}\mathrm{ov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}(X^3) - \mathbb{E}(X)\mathbb{E}(X^2)$, but as X only takes values -1, 0 and 1, $X^3 = X$. It remains

$$\mathbb{C}\mathrm{ov}(X,Y) = \mathbb{E}(X)(1-\mathbb{E}(X^2))$$

Also, $\mathbb{E}(X) = (-1)\times p + 1 \times p = 0$, so that

$$\mathbb{C}\mathrm{ov}(X,Y) = 0$$

$\rightsquigarrow X$ and Y are **uncorrelated**

However, there is a direct functional dependence between X and Y!

In particular, $\mathbb{P}(Y=0|X=0) = 1$, but $\mathbb{P}(Y=0) = 1-2p \neq 0$
$\rightsquigarrow X$ and Y are not independent!

Variance of a sum of random variables

From the properties of the covariance, it follows:

$$\begin{aligned}\mathbb{V}\mathrm{ar}\,(aX+bY) &= \mathbb{C}\mathrm{ov}\,(aX+bY, aX+bY)\\ &= \mathbb{C}\mathrm{ov}(aX,aX) + \mathbb{C}\mathrm{ov}(aX,bY)\\ &\quad + \mathbb{C}\mathrm{ov}(bY,aX) + \mathbb{C}\mathrm{ov}(bY,bY)\\ &= \mathbb{V}\mathrm{ar}(aX) + \mathbb{V}\mathrm{ar}(bY) + 2\,\mathbb{C}\mathrm{ov}(aX,bY)\\ &= a^2\,\mathbb{V}\mathrm{ar}(X) + b^2\,\mathbb{V}\mathrm{ar}(Y) + 2ab\,\mathbb{C}\mathrm{ov}(X,Y)\end{aligned}$$

Now, if X and Y are independent random variables,

$$\mathbb{V}\mathrm{ar}\,(aX+bY) = a^2\,\mathbb{V}\mathrm{ar}(X) + b^2\,\mathbb{V}\mathrm{ar}(Y)$$

For instance, if X and Y are independent,

$$\mathbb{V}\mathrm{ar}(X+Y) = \mathbb{V}\mathrm{ar}(X) + \mathbb{V}\mathrm{ar}(Y)$$

$$\mathbb{V}\mathrm{ar}(X-Y) = \mathbb{V}\mathrm{ar}(X) + \mathbb{V}\mathrm{ar}(Y)$$

Example

Example

We have two scales for measuring small weights in a laboratory. Assume the true weight of an item is 2g. Both scales give readings which have mean 2g and variance 0.05g^2. Compare using only one scale and using both scales then averaging the two measures in terms of the accuracy.

The first measure X has $\mathbb{E}(X) = 2$ and $\mathbb{V}\text{ar}(X) = 0.05$. Now, denote the second measure Y, independent of X, with $\mathbb{E}(Y) = 2$ and $\mathbb{V}\text{ar}(Y) = 0.05$. Then, take $W = \frac{X+Y}{2}$. We have

$$\mathbb{E}(W) = \frac{1}{2}\mathbb{E}(X) + \frac{1}{2}\mathbb{E}(Y) = \frac{2}{2} + \frac{2}{2} = 2 \text{ (g)}$$

and

$$\mathbb{V}\text{ar}(W) = \frac{1}{4} \times (\mathbb{V}\text{ar}(X) + \mathbb{V}\text{ar}(Y)) = \frac{1}{4} \times (0.05 + 0.05) = 0.025 \text{ (g}^2\text{)}$$

→ Averaging 2 measures reduces the variance by 2 → higher accuracy!

Correlation

The covariance of two r.v. is important as an indicator of the relationship between them.

However, it heavily depends on units of X and Y (difficult interpretation, not scale-invariant).

→ the correlation coefficient ρ is often used instead.

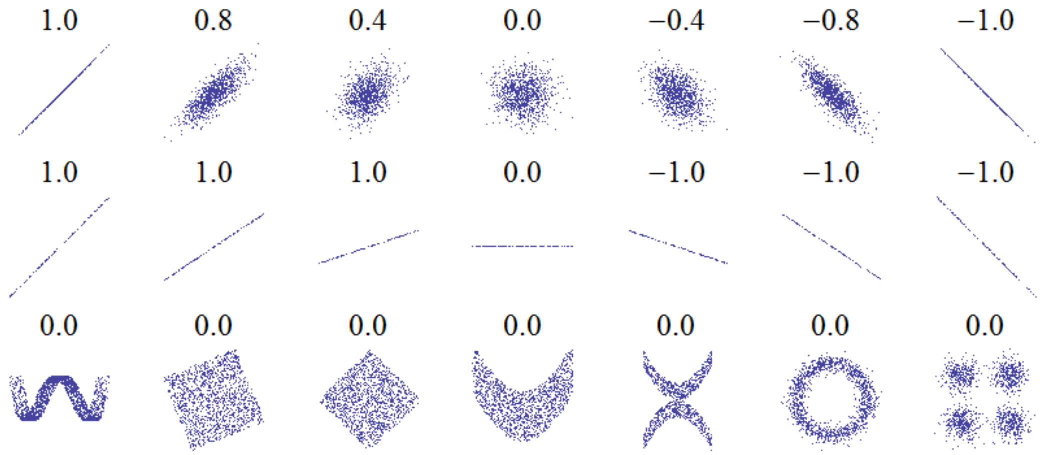
It is the covariance between the standardised versions of X and Y, or, explicitly,

$$\rho = \frac{\mathbb{C}\text{ov}(X, Y)}{\sqrt{\mathbb{V}\text{ar}(X)\,\mathbb{V}\text{ar}(Y)}}$$

Properties:

- ρ is dimensionless (no unit)
- ρ always has a value between -1 and 1.
- Positive (and negative) ρ means positive (and negative) linear relationship between X and Y
- The closer $|\rho|$ is to 1, the stronger is the linear relationship

Correlation examples



Objectives

Now you should be able to:

- understand the differences between discrete and continuous r.v. □
- for discrete r.v., determine probabilities from pmf and the reverse □
- for continuous r.v., determine probabilities from pdf and the reverse □
- for discrete r.v., determine probabilities from cdf and cdf from pmf and the reverse □
- for continuous r.v., determine probabilities from cdf and cdf from pdf and the reverse □
- calculate means and variances for both discrete and continuous random variables □
- use joint pmf and joint pdf to calculate probabilities □
- calculate and interpret covariances and correlations between two random variables □

Recommended exercises

→ Q25 p.220, Q27 p.221, Q29&30 p.221, Q69 p.57, Q40&43 p.152, Q42 p.152, Q41 p.223, Q65 p.239 (2nd edition)

→ Q27 p.225, Q29 p.225, Q31&32 p.225, Q71 p.59, Q42&45 p.157, Q44 p.157, Q43 p.227, Q67 p.243 (3rd edition)

5 Special random variables

Introduction

In practice, certain “types” of random variables come up over and over again from similar (random) experiments.

In this chapter, we will study a variety of those **special random variables**.

You can also go to

http://socr.ucla.edu/htmls/SOCR_Distributions.html

and have a look at the numerous ‘special’ distributions there.

Example

Consider the following random experiments and random variables:

- Flip a coin 10 times. Let X = number of heads obtained
- A worn machine tool produces defective parts 1% of the time . Let X = number of defective parts in the next 25 parts produced
- Each sample of air has a 10% chance of containing a particular molecule. Let X = the number of air samples that contain the molecule in the next 18 samples analysed
- Of all bits transmitted through a digital channel, 15% are received in error. Let X = the number of bits in error in the next five bits transmitted
- A multiple-choice test contains 10 questions, each with 4 choices, and you guess at each question. Let X = the number of questions answered correctly

→ Similar experiments, similar random variables

→ A general framework that include these experiments as particular cases would be very useful

The Binomial distribution

Assume:

- the outcome of a random experiment can be classified as either a "Success" or a "Failure" $(\to S = \{\text{Success}, \text{Failure}\})$
- we observe a Success with probability π
- n independent repetitions of this experiment are performed

Define X = number of Successes observed over the n repetitions. We say that X is a **binomial random variable** with parameters n and π:

$$\boxed{X \sim \text{Bin}(n, \pi)}$$

See that $S_X = \{0, 1, 2, \dots, n\}$ ($\to$ discrete r.v.) and the binomial probability mass function is given by

$$p(x) = \binom{n}{x} \pi^x (1-\pi)^{n-x}, \qquad \text{for } x \in S_X$$

where $\binom{n}{x}$ is the number of different groups of x objects that can be chosen from a set of n objects.

The Binomial distribution

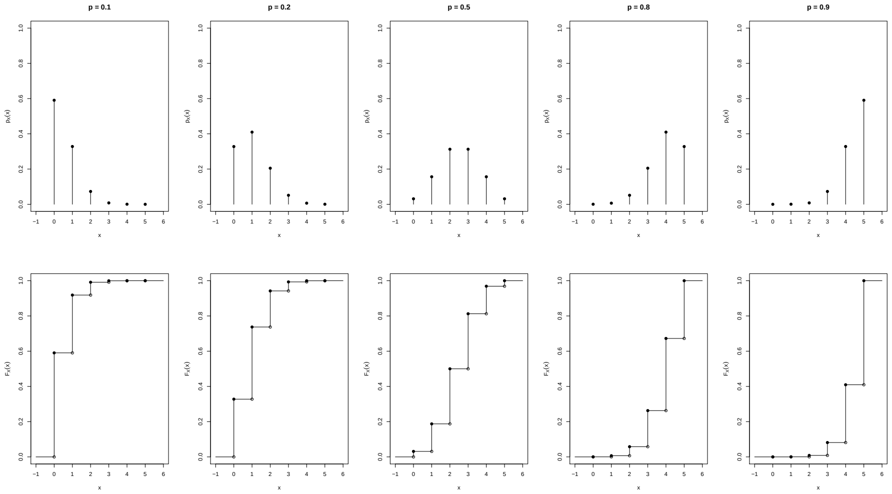
Note: the coefficients $\binom{n}{x} = n!/(x!(n-x)!)$ are called the binomial coefficients, they are the coefficients arising in Newton's famous binomial expansion.

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

These coefficients are often represented in Pascal's triangle, named after the French mathematician Blaise Pascal (1623-1662)

1
1 1
1 2 1
1 3 3 1
1 4 6 4 1
1 5 10 10 5 1
1 6 15 20 15 6 1
1 7 21 35 35 21 7 1
1 8 28 56 70 56 28 8 1
1 9 36 84 126 126 84 36 9 1
1 10 45 120 210 252 210 120 45 10 1
1 11 55 165 330 462 462 330 165 55 11 1
1 12 66 220 495 792 924 792 495 220 66 12 1
1 13 78 186 715 1287 1716 1716 1287 715 186 78 13 1

The Binomial distribution: pmf and cdf



Binomial pmf and cdf, for $n = 5$ and $\pi = \{0.1, 0.2, 0.5, 0.8, 0.9\}$

The Bernoulli distribution

Particular case: if $n = 1 \to$ the Bernoulli distribution (Slide 131)

$$\boxed{X \sim \text{Bern}(\pi)}$$

pmf:

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

Note: if $X \sim \text{Bin}(n, \pi)$, we can represent it as

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

where X_i's are n independent Bernoulli r.v. with parameters π

→ Each repetition of the experiment in the Binomial framework is called a Bernoulli trial

Binomial distribution: properties

First note that

$$\sum_{x \in S_X} p(x) = \sum_{x=0}^{n} \binom{n}{x} \pi^x (1-\pi)^{n-x} = (\pi + (1-\pi))^n = 1$$

using the binomial expansion

Second, it is easy to see that if $X_1 \sim \text{Bin}(n_1, \pi)$, $X_2 \sim \text{Bin}(n_2, \pi)$ and X_1 is independent of X_2, then

$$X_1 + X_2 \sim \text{Bin}(n_1 + n_2, \pi)$$

Binomial distribution: expectation and variance

Recall the representation $X = \sum_{i=1}^n X_i$, with $X_i \sim \text{Bern}(\pi)$

We know (Slides 142 and 150) that

$$\mathbb{E}(X_i) = \pi \qquad \text{and} \qquad \mathbb{V}\text{ar}(X_i) = \pi(1-\pi)$$

It follows $\mathbb{E}(X) = \mathbb{E}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathbb{E}(X_i) = \sum_{i=1}^n \pi = n\pi$ and

$$\mathbb{V}\text{ar}(X) = \mathbb{V}\text{ar}\left(\sum_{i=1}^n X_i\right) \overset{\text{ind.}}{=} \sum_{i=1}^n \mathbb{V}\text{ar}(X_i) = \sum_{i=1}^n \pi(1-\pi) = n\pi(1-\pi)$$

Mean and variance of the binomial distribution

If $X \sim \text{Bin}(n, \pi)$,

$$\mu = \mathbb{E}(X) = n\pi \qquad \text{and} \qquad \sigma^2 = \mathbb{V}\text{ar}(X) = n\pi(1-\pi)$$

Binomial distribution: examples

Example

It is known that disks produced by a certain company will be defective with probability 0.01 independently of each other. The company sells the disk in packages of 10 and offers a money-back guarantee if more than 1 of the disks are defective. a) In the long-run, what proportion of packages is returned? b) If someone buys three packages, what is the probability that exactly one of them will be returned?

a) Let X be the number of defective disks in a package. Then,

$$X \sim \text{Bin}(10, 0.01)$$

Hence,

$$\begin{aligned} \mathbb{P}(X > 1) &= 1 - \mathbb{P}(X = 0) - \mathbb{P}(X = 1) \\ &= 1 - \binom{10}{0} 0.01^0 0.99^{10} - \binom{10}{1} 0.01^1 0.99^9 \simeq 0.004 \end{aligned}$$

→ In the long-run, 0.4 percent of the packages will have to be returned

Binomial distribution: examples

Example

It is known that disks produced by a certain company will be defective with probability 0.01 independently of each other. The company sells the disk in packages of 10 and offers a money-back guarantee if more than 1 of the disks are defective. a) In the long-run, what proportion of packages is returned? b) If someone buys three packages, what is the probability that exactly one of them will be returned?

b) Let Y be the number of packages that the person will have to return. We have

$$Y \sim \text{Bin}(3, \pi)$$

where π is the probability that a package is returned, that is, contains more than 1 defective disk. In a), we found that $\pi = 0.004$

Thus, the probability that exactly one of the three packages will be returned is

$$\mathbb{P}(Y = 1) = \binom{3}{1} 0.004^1 0.996^2 = 0.012$$

Binomial distribution: examples

Example (Ex. 54 p.54 in the textbook)

Suppose that 10% of all bits transmitted through a digital communication channel are erroneously received and that whether any is erroneously received is independent of whether any other bit is erroneously received. Consider sending a large number of messages, each consisting of 20 bits. a) What proportion of these messages will have exactly 2 erroneously received bits? b) What proportion of these messages will have at least 5 erroneously received bits? c) What proportion of these messages will more than half the bits be erroneously received?

Let X be the number of erroneously received bits in a message of 20 bits. Clearly, we have $X \sim \text{Bin}(20, 0.1)$. Thus we have

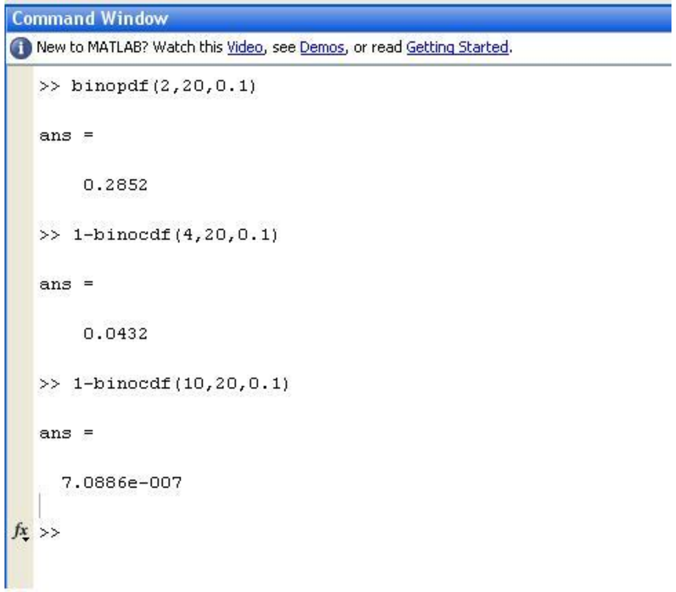
a) $\mathbb{P}(X = 2) = \binom{20}{2} 0.1^2 0.9^{18} = 0.2852$

b) $\mathbb{P}(X \geq 5) = 1 - \mathbb{P}(X = 0) - \mathbb{P}(X = 1) - \mathbb{P}(X = 2) - \mathbb{P}(X = 3) - \mathbb{P}(X = 4) = \ldots$

c) $\mathbb{P}(X > 10) = \mathbb{P}(X = 11) + \mathbb{P}(X = 12) + \ldots + \mathbb{P}(X = 20) = \ldots$

→ very tedious! → use **statistical software**

Binomial distribution: examples



Command Window
New to MATLAB? Watch this Video, see Demos, or read Getting Started.

```
>> binopdf(2,20,0.1)

ans =

    0.2852

>> 1-binocdf(4,20,0.1)

ans =

    0.0432

>> 1-binocdf(10,20,0.1)

ans =

  7.0886e-007

>>
```

The Poisson distribution

Assume you are interested in the number of occurrences of some random phenomenon in a fixed period of time.

Define $X =$ number of occurrences. We say that X is a **Poisson random variable** with parameter λ, i.e.

$$X \sim \mathcal{P}(\lambda),$$

if

$$p(x) = e^{-\lambda} \frac{\lambda^x}{x!} \qquad \text{for } x \in S_X = \{0, 1, 2, \ldots\}$$

Note: Simeon-Denis Poisson (1781-1840) was a French mathematician

Poisson distribution: how does it arise?

- Think of the time period of interest as being split up into a large number, say n, of sub-periods
- Assume that the phenomenon could occur at most one time in each of those subperiods, with some constant probability π
- If what happens within one interval is independent to others,

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

- Now, as n increases, π should decrease (the shorter the period, the less likely the occurrence of the phenomenon) → let $\pi = \lambda/n$ for some $\lambda > 0$
- Then, for any $x \in \{0, 1, \ldots, n\}$,

$$\begin{aligned}\mathbb{P}(X = x) &= \frac{n!}{x!(n-x)!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \qquad \text{(binomial pmf)} \\ &= \frac{n!}{n^x (n-x)!(1-\lambda/n)^x} \frac{\lambda^x}{x!} \left(1 - \frac{\lambda}{n}\right)^n\end{aligned}$$

Poisson distribution: how does it arise?

- Finally, as $n \to \infty$

$$\frac{n!}{n^x(n-x)!(1-\lambda/n)^x} \to 1 \quad \text{and} \quad \left(1-\frac{\lambda}{n}\right)^n \to e^{-\lambda}$$

Therefore,

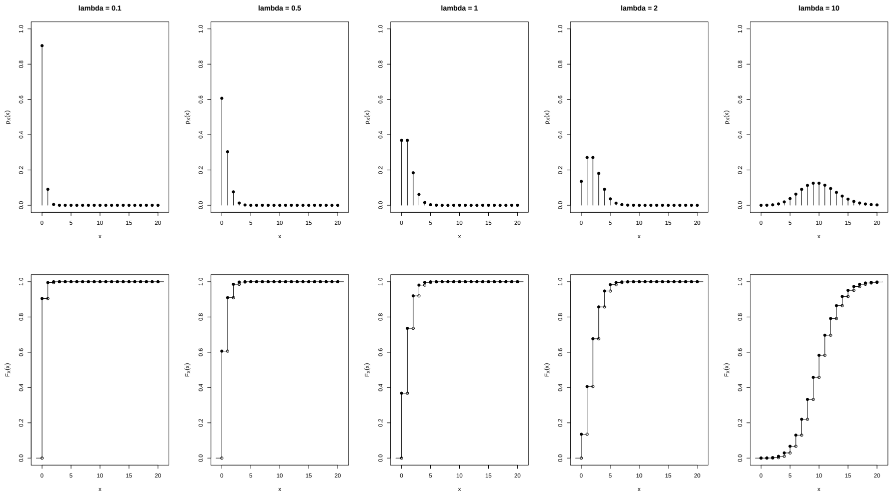
$$\mathbb{P}(X=x) = e^{-\lambda}\frac{\lambda^x}{x!} \qquad \text{for } x \in \{0,1,\ldots\}$$

which is the Poisson pmf as given on Slide 183

- The Poisson distribution is thus suitable for modelling the number of occurrences of a random phenomenon satisfying some assumptions of continuity, stationarity, independence and non-simultaneity
- λ is called the intensity of the phenomenon

Note: we defined the $\mathcal{P}(\lambda)$ distribution by partitioning a time period, however the same reasoning can be applied to any interval, area or volume

Poisson distribution: pmf and cdf



Poisson pmf and cdf, for $\lambda = \{0.1, 0.5, 1, 2, 10\}$

Poisson distribution: properties

First we have, as expected,

$$\sum_{x \in S_X} p(x) = \sum_{x=0}^{\infty} e^{-\lambda}\frac{\lambda^x}{x!} = e^{-\lambda}\sum_{x=0}^{\infty}\frac{\lambda^x}{x!} = e^{-\lambda}e^{\lambda} = 1$$

Similarly,

$$\mathbb{E}(X) = \sum_{x \in S_X} x p(x) = \sum_{x=0}^{\infty} x e^{-\lambda}\frac{\lambda^x}{x!} = \lambda \sum_{x=1}^{\infty} e^{-\lambda}\frac{\lambda^{x-1}}{(x-1)!} = \lambda$$

$$\mathbb{E}(X^2) = \sum_{x \in S_X} x^2 p(x) = \ldots = \lambda^2 + \lambda$$

$$\to \mathbb{V}\text{ar}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

Mean and variance of the Poisson distribution

If $X \sim \mathcal{P}(\lambda)$,

$$\mathbb{E}(X) = \lambda \qquad \text{and} \qquad \mathbb{V}\text{ar}(X) = \lambda$$

Poisson distribution: examples

Example

Over a 10-minute period, a counter records an average of 1.3 gamma particles per millisecond coming from a radioactive substance. To a good approximation, the distribution of the count, X, of gamma particles during the next millisecond is Poisson distributed. Determine a) λ, b) the probability of observing one or more gamma particles during the next millisecond and c) the variance of this number.

a) The mean of the Poisson distribution is λ, so we can approximate λ by the long-run average of the number of particles per millisecond, that is, $\lambda \simeq 1.3$. So we have

$$X \sim \mathcal{P}(1.3)$$

b) Thus,

$$\mathbb{P}(X \geq 1) = 1 - \mathbb{P}(X=0) = 1 - e^{-1.3}\frac{1.3^0}{0!} = 1 - e^{-1.3} = 0.727$$

c) The variance of the Poisson distribution is also equal to λ, hence

$$\mathbb{V}\text{ar}(X) = 1.3 \ \ (\text{particles}^2)$$

Poisson distribution: examples

Example (Ex. 56 p.55 in the textbook)

Suppose that the number of drivers who travel between a particular origin and destination during a designated time period has Poisson distribution with parameter $\lambda = 20$. In the long-run, in what proportion of time periods will the number of drivers a) be at most 10? b) exceed 20? c) be between 10 and 20, inclusive? Strictly between 10 and 20?

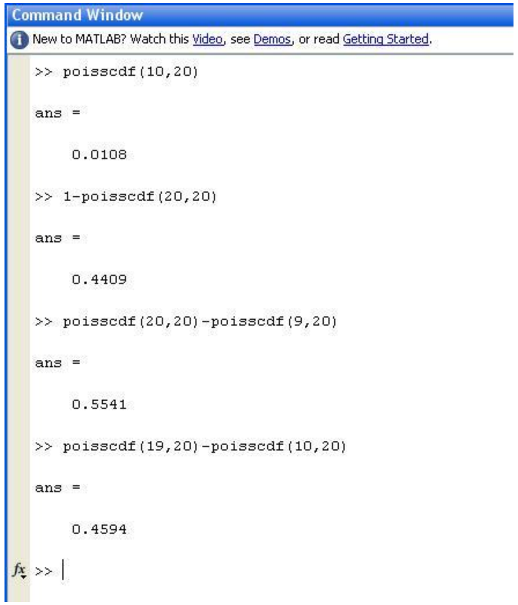
Let X be the number of drivers. It is given that $X \sim \mathcal{P}(20)$

a)

$$\begin{aligned}\mathbb{P}(X \leq 10) &= \mathbb{P}(X=0) + \mathbb{P}(X=1) + \ldots + \mathbb{P}(X=10)\\ &= e^{-20} + e^{-20} \times 20 + e^{-20}\frac{20^2}{2} + \ldots + e^{-20}\frac{20^{10}}{10!}\\ &= \ldots\end{aligned}$$

→ tedious ! → use **statistical software**

Poisson distribution: examples



New to MATLAB? Watch this Video, see Demos, or read Getting Started.

```
>> poisscdf(10,20)

ans =

    0.0108

>> 1-poisscdf(20,20)

ans =

    0.4409

>> poisscdf(20,20)-poisscdf(9,20)

ans =

    0.5541

>> poisscdf(19,20)-poisscdf(10,20)

ans =

    0.4594

fx >> |
```

Poisson approximation to the Binomial distribution

Since it was derived as a limit case of the Binomial distribution when n is 'large' and π is 'small', one can expect the Poisson distribution to be a good approximation to Bin(n, π) in that case.

As it involves only one parameter, the Poisson pmf is usually easier to handle than the corresponding Binomial

Example

It is known that 1% of the books at a certain bindery have defective bindings. Compare the probabilities that x ($x = 0, 1, 2, \ldots$) of 100 books will have defective bindings using the (exact) formula for the binomial distribution and its Poisson approximation.

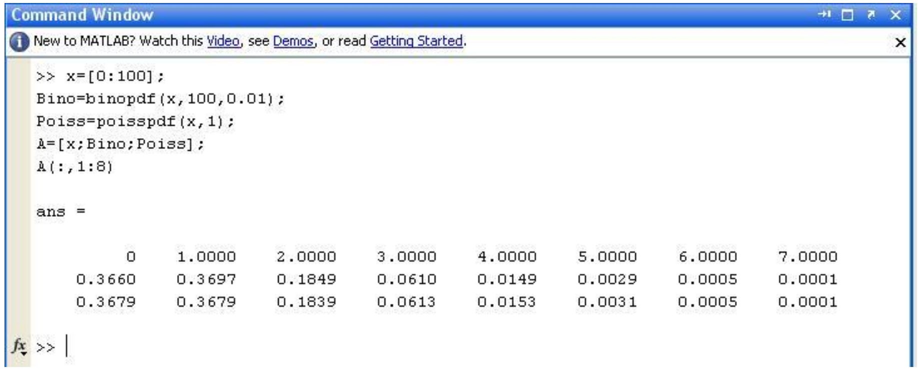
The exact Binomial pmf is $p(x) = \binom{100}{x} \times 0.01^x \times 0.99^{100-x}$, while its Poisson approximation is

$$p^*(x) = e^{-\lambda}\frac{x^\lambda}{\lambda!}$$

with $\lambda = n \times p = 100 \times 0.01 = 1$

Poisson distribution: examples

Matlab computations give:



New to MATLAB? Watch this Video, see Demos, or read Getting Started.

```
>> x=[0:100];
Bino=binopdf(x,100,0.01);
Poiss=poisspdf(x,1);
A=[x;Bino;Poiss];
A(:,1:8)

ans =

         0    1.0000    2.0000    3.0000    4.0000    5.0000    6.0000    7.0000
    0.3660    0.3697    0.1849    0.0610    0.0149    0.0029    0.0005    0.0001
    0.3679    0.3679    0.1839    0.0613    0.0153    0.0031    0.0005    0.0001

fx >> |
```

We see that the error we would make by using the Poisson approximation instead of the true distribution is only of order 10^{-3}

→ very good approximation

The Uniform distribution

There are also numerous **continuous** distributions which are of great interest. The simplest one is certainly the **uniform distribution**.

A random variable is said to be uniformly distributed over an interval $[\alpha, \beta]$, i.e.

$$\boxed{X \sim U_{[\alpha,\beta]}}$$

if its probability density function is given by

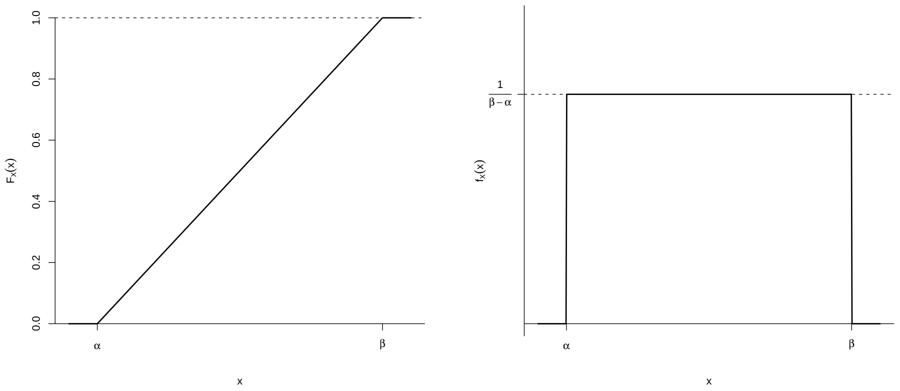
$$f(x) = \begin{cases} \frac{1}{\beta-\alpha} & \text{if } x \in [\alpha, \beta] \\ 0 & \text{otherwise} \end{cases} \qquad (\to S_X = [\alpha, \beta])$$

Constant density → X is just as likely to be "close" to any value in S_X.

By integration, it is easy to show that

$$F(x) = \begin{cases} 0 & \text{if } x < \alpha \\ \frac{x-\alpha}{\beta-\alpha} & \text{if } \alpha \leq x \leq \beta \\ 1 & \text{if } x > \beta \end{cases}$$

The Uniform distribution



cdf $F(x)$

pdf $f(x) = F'(x)$

Uniform distribution: properties

Note that the Uniform density is constant at $1/(\beta - \alpha)$ on $[\alpha, \beta]$ so as to ensure that $\int_\alpha^\beta f(x)\, dx = 1$

Now,

$$\mathbb{E}(X) = \int_\alpha^\beta x \frac{1}{\beta-\alpha} dx = \frac{1}{\beta-\alpha}\left[\frac{x^2}{2}\right]_\alpha^\beta = \frac{\beta^2-\alpha^2}{2(\beta-\alpha)} = \frac{\alpha+\beta}{2}$$

Similarly,

$$\mathbb{E}(X^2) = \int_\alpha^\beta x^2 \frac{1}{\beta-\alpha} dx = \frac{1}{\beta-\alpha}\left[\frac{x^3}{3}\right]_\alpha^\beta = \frac{\beta^3-\alpha^3}{3(\beta-\alpha)} = \frac{\beta^2+\alpha\beta+\alpha^2}{3}$$

which implies $\mathbb{V}\text{ar}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \ldots = \frac{(\beta-\alpha)^2}{12}$

Mean and variance of the Uniform distribution

If $X \sim U_{[\alpha,\beta]}$,

$$\mathbb{E}(X) = \frac{\alpha+\beta}{2} \qquad \text{and} \qquad \mathbb{V}\text{ar}(X) = \frac{(\beta-\alpha)^2}{12}$$

Uniform distribution: example

The probability that X lies in any subinterval $[a, b]$ of $[\alpha, \beta]$ is:

$$\boxed{\mathbb{P}(a < X < b) = \frac{b-a}{\beta-\alpha}} \qquad \text{(area of a rectangle)}$$

Example

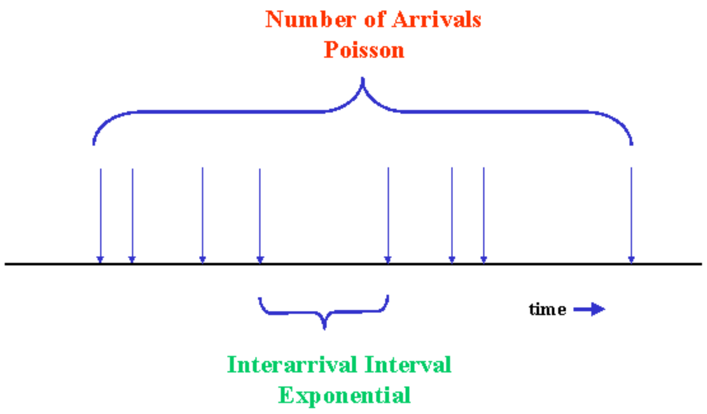
Buses arrive at a specified stop at 15-minute intervals starting at 7 A.M. That is, they arrive at 7, 7:15, 7:30, 7:45, etc. If a passenger arrives at the stop at a time uniformly distributed between 7 and 7:30, find the probability that he waits less than 5 minutes for a bus

Let X denote the time (in minutes) past 7 A.M. that the passenger arrives at the stop. We have $X \sim U_{[0,30]}$

The passenger will have to wait less than 5 min if he arrives between 7:10 and 7:15 or between 7:25 and 7:30. This happens with probability

$$\begin{aligned} \mathbb{P}((10 < X < 15) \cup (25 < X < 30)) &= \mathbb{P}(10 < X < 15) + \mathbb{P}(25 < X < 30) \\ &= \frac{5}{30} + \frac{5}{30} = \frac{1}{3} \end{aligned}$$

The Exponential distribution



Poisson and exponential distributions

The Exponential distribution

- Recall that a Poisson distributed r.v. counts the number of occurrences of a given phenomenon over a unit period of time
- The (random) amount of time before the first occurrence of that phenomenon is often of interest as well
- If $N \sim \mathcal{P}(\lambda)$ denote the number of occurrences over a unit period of time, then the number of occurrences of the phenomenon by a time x, say N_x, is $\sim \mathcal{P}(\lambda x)$ ("Poisson process")
- Denote X the amount of time before the first occurrence
- This time will exceed x ($x \geq 0$) if and only if there have been no occurrences of the phenomenon by time x, that is, $N_x = 0$

As $N_x \sim \mathcal{P}(\lambda x)$, it follows $\mathbb{P}(X > x) = \mathbb{P}(N_x = 0) = e^{-\lambda x}\frac{(\lambda x)^0}{0!} = e^{-\lambda x}$, which yields the cdf of X:

$$F(x) = \mathbb{P}(X \leq x) = 1 - e^{-\lambda x} \qquad \text{for } x \geq 0$$

This particular distribution is called the Exponential distribution.

The Exponential distribution

A random variable is said to be an **Exponential random variable** with parameter μ ($\mu > 0$), i.e.

$$X \sim \text{Exp}(\mu),$$

if its probability density function is given by

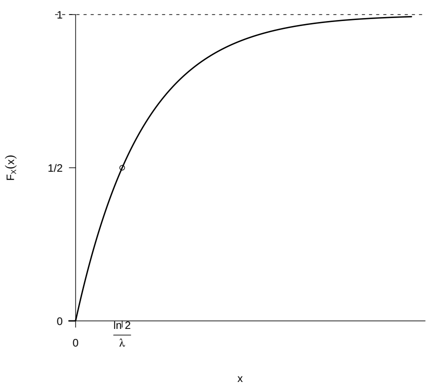
$$f(x) = \begin{cases} \frac{1}{\mu} e^{-\frac{x}{\mu}} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases} \qquad (\to S_X = \mathbb{R}^+)$$

By integration, we find

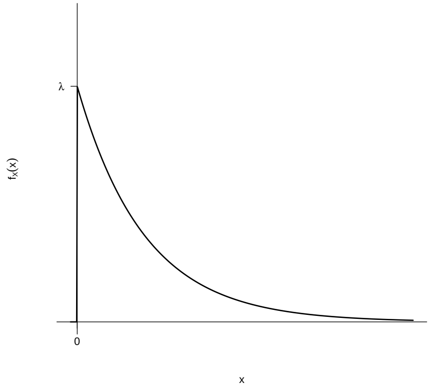
$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - e^{-\frac{x}{\mu}} & \text{if } x \geq 0 \end{cases}$$

This distribution is often useful for representing random amounts of time, like the amount of time required to complete a task, the waiting time at a counter, the amount of time until you receive a phone call, etc. Note: the parameter μ is related to λ by $\mu = 1/\lambda$.

The Exponential distribution



cdf $F(x)$



pdf $f(x) = F'(x)$

Exponential distribution: properties

We can check that (as expected)

$$\int_{-\infty}^{+\infty} f(x)\,dx = \int_0^{+\infty} \frac{1}{\mu} e^{-\frac{x}{\mu}}\,dx = \frac{1}{\mu}\left[\frac{e^{-\frac{x}{\mu}}}{(-\frac{1}{\mu})}\right]_0^{+\infty} = 1$$

Moreover,

$$\begin{aligned}\mathbb{E}(X) &= \int_0^{+\infty} x\,\frac{1}{\mu} e^{-\frac{x}{\mu}}\,dx = \left[-x\,e^{-\frac{x}{\mu}}\right]_0^{+\infty} + \int_0^{+\infty} e^{-\frac{x}{\mu}}\,dx \qquad \text{(by parts)} \\ &= 0 + \left[-\frac{e^{-\frac{x}{\mu}}}{\frac{1}{\mu}}\right]_0^{+\infty} = \mu\end{aligned}$$

Similarly, $\mathbb{E}(X^2) = \int_0^{+\infty} x^2 e^{-\frac{x}{\mu}}\,dx = \ldots = 2\mu^2$, so that $\mathbb{V}\text{ar}(X) = \mu^2$

Mean and variance of the Exponential distribution

If $X \sim \text{Exp}(\mu)$,

$$\mathbb{E}(X) = \mu \qquad \text{and} \qquad \mathbb{V}\text{ar}(X) = \mu^2$$

Exponential distribution: example

Example

Suppose that, on average, 3 trucks arrive per hour to be unloaded at a warehouse. What is the probability that the time between the arrivals of two successive trucks will be a) less than 5 minutes? b) at least 45 minutes?

Assuming the number of trucks arriving during one hour is Poisson distributed (with parameter $\lambda = 3$), then the amount of time X between two truck arrivals follows the Exp(1/3) distribution

Hence,

a) $\mathbb{P}(X \leq 1/12) = \int_0^{1/12} 3e^{-3x}\,dx = 1 - e^{-1/4} = 0.221$

b) $\mathbb{P}(X > 3/4) = \int_{3/4}^{\infty} 3e^{-3x}\,dx = e^{-9/4} = 0.105$

Other useful distributions

In the remainder of this course we will also encounter some other continuous distributions, among these are

- the Student-t (or just t) distribution, $X \sim t_\nu$;
- the Fisher-F (or just F) distribution, $X \sim \mathbf{F}_{d_1,d_2}$

We will return to them later when we will need them.

The several distributions that we have introduced so far are very useful in the application of statistics to problems of engineering and physical science.

The Normal distribution: introduction

However, the most widely used, and therefore the most important, statistical distribution is undoubtedly the

Normal distribution

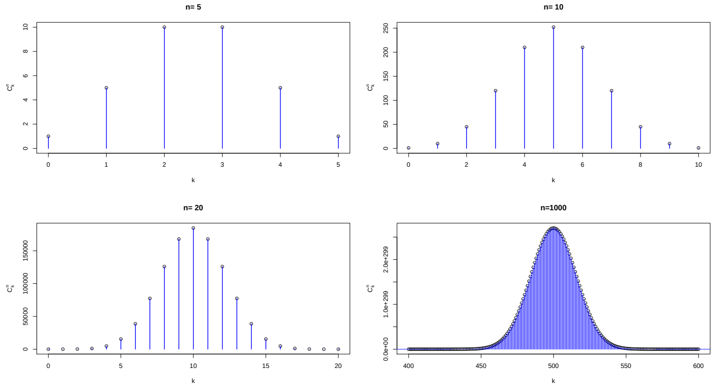
Its prevalence was first highlighted when it was observed that in many natural processes, random variation among individuals systematically conforms to a particular pattern:

- most of the observations concentrate around one single value (which is the mean)
- the number of observations smoothly decreases, symmetrically on either side, with the deviation from the mean
- it is very unlikely, yet not impossible, to find very extreme values

→ this yields the famous **bell-shaped** curve

The Normal distribution: introduction

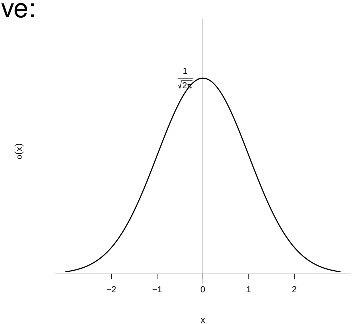
The bell-shaped curve was first spotted by the French mathematician Abraham de Moivre (1667-1754) who in his 1738 book “The Doctrine of Chances” showed that the coefficients $C_k^n = \binom{n}{k}$ in the binomial expansion of $(a+b)^n$ (see Slide 174) precisely follow the bell shape pattern when n is large



The Normal distribution: introduction

Later, Carl-Friedrich Gauss (1777-1855), a German mathematician (sometimes referred to as the *Princeps mathematicorum*, Latin for "the Prince of Mathematicians" or "the foremost of mathematicians"), was the first to write an explicit equation for the bell-shaped curve:

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



When deriving his distribution, Gauss was primarily interested in errors of measurement, whose distribution typically follows the bell-shaped curve as well. He called his curve the “normal curve of errors”, which was to become the **Normal distribution**. In honour of Gauss, the Normal distribution is also often referred to as the **Gaussian distribution**

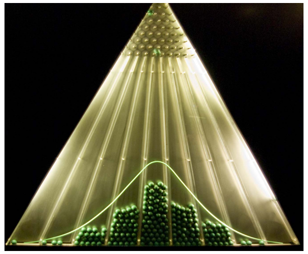
The Normal distribution: introduction

The Normal distribution is not just a convenient mathematical tool, but also occurs in natural phenomena.

For instance, in 1866 Maxwell, a Scottish physicist, determined the distribution of molecular velocity in a gas at equilibrium. As a result of unpredictable collisions with other molecules, molecular velocity in a given direction is randomly distributed, and from basic assumptions, that distribution can be shown to be the Normal distribution

The Normal distribution: introduction

Another famous example is the “bean machine”, invented by Sir Francis Galton (English scientist, 1822-1911) to demonstrate the Normal distribution. The machine consists of a vertical board with interleaved rows of pins. Balls are dropped from the top, and bounce left and right as they hit the pins. Eventually, they are collected into bins at the bottom. The height of ball columns in the bins approximately follows the bell-shaped curve.



The Normal distribution

A random variable is said to be normally distributed with parameters μ and σ ($\sigma > 0$), i.e.

$$X \sim \mathcal{N}(\mu, \sigma),$$

if its probability density function is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (\to S_X = \mathbb{R})$$

Unfortunately, no closed form exists for

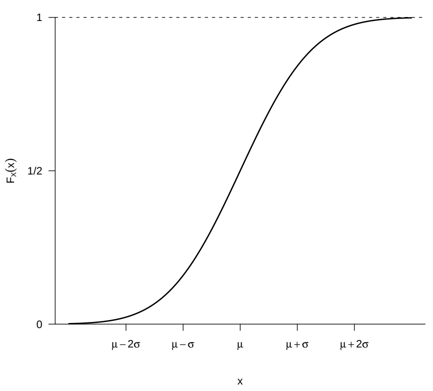
$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy$$

Important remark: Be careful! Many sources use the alternative notation

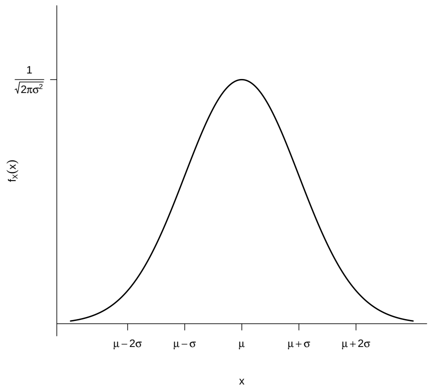
$$X \sim \mathcal{N}(\mu, \sigma^2)$$

→ in the textbook and in Matlab, the notation $\mathcal{N}(\mu, \sigma)$ is used, so we adopt it in these slides as well

The Normal distribution



cdf $F(x)$



pdf $f(x) = F'(x)$

Normal distribution: properties

It can be shown that, for any μ and σ,

$$\int_{S_X} f(x)\,dx = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = 1$$

Similarly, we can find

$$\mathbb{E}(X) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} x\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \mu$$

and

$$\mathbb{V}\text{ar}(X) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} (x-\mu)^2\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \sigma^2$$

Mean and variance of the Normal distribution

If $X \sim \mathcal{N}(\mu, \sigma)$,

$$\mathbb{E}(X) = \mu \qquad \text{and} \qquad \mathbb{V}\text{ar}(X) = \sigma^2 \qquad (\to \text{sd}(X) = \sigma)$$

The Standard Normal distribution

The **Standard Normal distribution** is the Normal distribution with $\mu = 0$ and $\sigma = 1$. This yields

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

Usually, in this situation, the specific notation

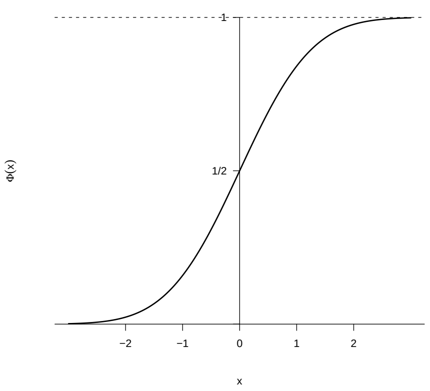
$$f(x) \doteq \phi(x) \qquad \text{and} \qquad F(x) \doteq \Phi(x)$$

is used, and a standard normal random variable is usually denoted Z.

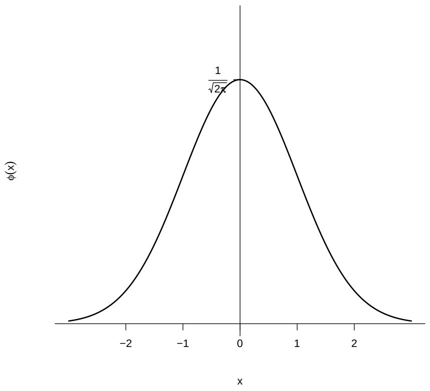
It directly follows from the preceding that

$$\mathbb{E}(Z) = 0 \qquad \text{and} \qquad \mathbb{V}\text{ar}(Z) = 1 \quad (= \text{sd}(Z))$$

The Standard Normal distribution



cdf $\Phi(x)$

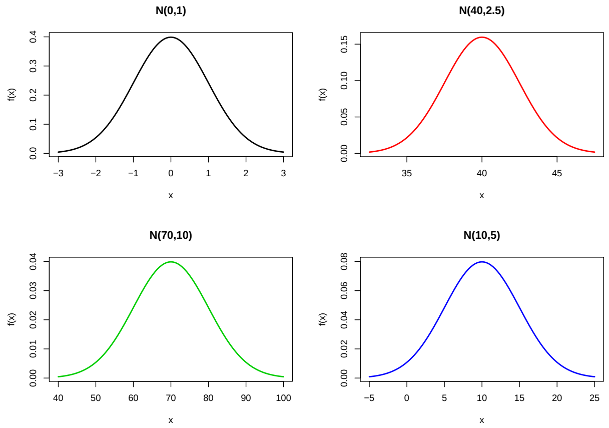


pdf $\phi(x) = \Phi'(x)$

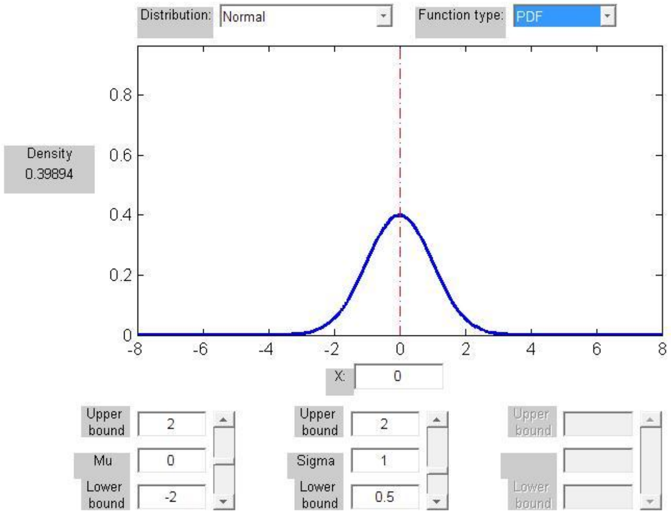
Normal distribution: properties

An important observation is that all normal probability distribution functions have the same bell shape

They only differ in where they are centred (at μ) and in their spread (quantified by σ).



In Matlab, key in `disttool` and play with the interactive probability function display tool



Distribution: Normal

Function type: PDF

Density 0.39894

X: 0

Upper bound 2

Mu 0

Lower bound -2

Upper bound 2

Sigma 1

Lower bound 0.5

Upper bound

Lower bound

Normal distribution: standardisation

It is clear from the expression and the shape of the Normal pdf that if $X \sim \mathcal{N}(\mu, \sigma)$, then $Y = aX + b$ is normally distributed with mean $\mathbb{E}(Y) = a\mu + b$ and variance $\mathbb{V}\mathrm{ar}(Y) = a^2\sigma^2$.

The following result directly follows from the foregoing:

Property: Standardisation

If $X \sim \mathcal{N}(\mu, \sigma)$, then

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

This linear transformation is called the **standardisation** of the normal random variable X, as it transforms X into a standard normal random variable Z.

Standardisation will play a paramount role in what follows.

Normal distribution: properties

This extremely important fact allows us to deduce any required information for a given Normal distribution $\mathcal{N}(\mu, \sigma)$ from the features of the 'simple' standard normal distribution.

For instance, for the standard pdf $\phi(x)$, it can be found that

$$\int_{-1}^{1} \phi(x)\,dx = \mathbb{P}(-1 < Z < 1) \simeq 0.6827$$
$$\int_{-2}^{2} \phi(x)\,dx = \mathbb{P}(-2 < Z < 2) \simeq 0.9545$$
$$\int_{-3}^{3} \phi(x)\,dx = \mathbb{P}(-3 < Z < 3) \simeq 0.9973$$

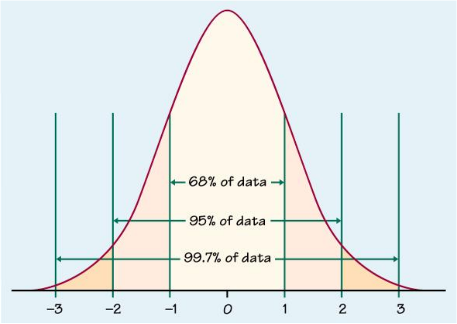
This automatically translates to the general case $X \sim \mathcal{N}(\mu, \sigma)$:

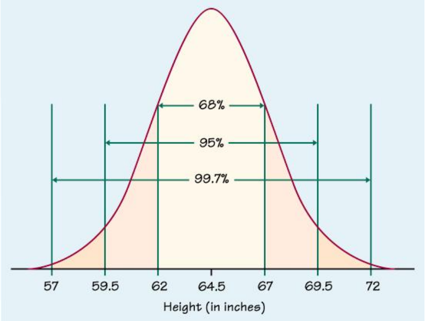
$$\mathbb{P}(\mu - \sigma < X < \mu + \sigma) \simeq 0.6827$$
$$\mathbb{P}(\mu - 2\sigma < X < \mu + 2\sigma) \simeq 0.9545$$
$$\mathbb{P}(\mu - 3\sigma < X < \mu + 3\sigma) \simeq 0.9973$$

This is known as the **68-95-99 rule** for normal distributions.

Normal distribution: properties

For instance, suppose we are told that women's heights in a given population follow a normal distribution with mean $\mu = 64.5$ inches and $\sigma = 2.5$ inches





→ we expect 68.27 % of women to be between $\mu - \sigma = 64.5 - 2.5 = 62$ inches and $\mu + \sigma = 64.5 + 2.5 = 67$ inches tall, etc.

Normal distribution: remark

- Theoretically, the domain of variation S_X of a normally distributed random variable X is $\mathbb{R} = (-\infty, +\infty)$
- However, there is a 99.7% chance to find X between $\mu - 3\sigma$ and $\mu + 3\sigma$
- It almost impossible to find X outside that interval, and virtually impossible to find it much further away from μ

→ There is in general no problem in modelling the distribution of a positive quantity with a Normal distribution, provided μ is large compared to σ

- Typical examples include weight, height, or IQ of people
- This also explains why 6σ is sometimes called the width of the normal distribution

Normal distribution: examples

Example

Suppose that $Z \sim \mathcal{N}(0, 1)$. What is $\mathbb{P}(Z \leq 1.25)$?

In principle, this should be given by

$$\mathbb{P}(Z \leq 1.25) = \Phi(1.25) = \int_{-\infty}^{1.25} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx$$

However, we know that this integral cannot be evaluated analytically.

→ we must use software (command `normcdf` in Matlab)*

This probability is the 'area under the standard normal curve to the left of z', that is

$$\mathbb{P}(Z \leq z) = \Phi(z)$$

Matlab gives: $\mathbb{P}(Z \leq 1.25) = \Phi(1.25) = 0.8944$ (> `normcdf(1.25)`)

* There also exist Tables, but we don't use them in this course

Normal distribution: examples

Any other kind of probabilities must be written in terms of $\mathbb{P}(Z \leq z)$.

Example

Suppose that $Z \sim \mathcal{N}(0,1)$. What is $\mathbb{P}(Z < 1.25)$? What is $\mathbb{P}(Z > 1.25)$? What is $\mathbb{P}(-0.38 \leq Z < 1.25)$?

As Z is a continuous random variable, $\mathbb{P}(Z < z) = \mathbb{P}(Z \leq z)$ for any z

$$\to \mathbb{P}(Z < 1.25) = \mathbb{P}(Z \leq 1.25) = \Phi(1.25) = 0.8944$$

$$\mathbb{P}(Z > 1.25) = 1 - \mathbb{P}(Z \leq 1.25) = 1 - \Phi(1.25) = 1 - 0.8944 = 0.1056$$

$$\begin{aligned}\mathbb{P}(-0.38 \leq Z < 1.25) &= \mathbb{P}(Z < 1.25) - \mathbb{P}(Z < -0.38)\\ &= \mathbb{P}(Z \leq 1.25) - \mathbb{P}(Z \leq -0.38)\\ &= \Phi(1.25) - \Phi(-0.38)\\ &= 0.8944 - 0.3520 = 0.5424\end{aligned}$$

(with Matlab: `> normcdf(1.25)-normcdf(-0.38)`)

Normal distribution: examples

Example 1.16 p.38 (textbook)

The time it takes a driver to react to the brake light on a decelerating vehicle follows a Normal distribution having parameters $\mu = 1.25$ sec and $\sigma = 0.46$ sec. In the long run, what proportion of reaction times will be between 1 and 1.75 sec? (**Hint:** $\Phi(1.09) = 0.8621$, $\Phi(-0.54) = 0.2946$.)

We have $X \sim \mathcal{N}(1.25, 0.46)$ and we desire $\mathbb{P}(1 \leq X \leq 1.75)$. We have that

$$\mathbb{P}(1 \leq X \leq 1.75) = \mathbb{P}(X \leq 1.75) - \mathbb{P}(X \leq 1)$$

The probabilities can be obtained from Matlab.

Alternatively, we know that $Z = \frac{X-1.25}{0.46} \sim \mathcal{N}(0,1)$. Then

$$\mathbb{P}(X \leq 1.75) = \mathbb{P}\left(\frac{X-1.25}{0.46} \leq \frac{1.75-1.25}{0.46}\right) = \mathbb{P}(Z \leq 1.09) = \Phi(1.09) = 0.8621$$

Similarly, $\mathbb{P}(X \leq 1) = \mathbb{P}(Z \leq -0.54) = \Phi(-0.54) = 0.2946$, so that

$$\mathbb{P}(1 \leq X \leq 1.75) = 0.8621 - 0.2946 = 0.5675$$

Normal distribution: examples

Example

The actual amount of instant coffee that a filling machine puts into "4-ounce" jars may be looked upon as a random variable having a normal distribution with $\sigma = 0.04$ ounce. If only 2% of the jars are to contain less than 4 ounces, what should be the mean fill of these jars? (**Hint:** $\Phi(-2.05) = 0.02$.)

Let X denote the actual amount of coffee put into the jar by the machine

We have $X \sim \mathcal{N}(\mu, 0.04)$, with μ such that $\mathbb{P}(X \leq 4) = 0.02$

Hence,

$$0.02 = \mathbb{P}(X \leq 4) = \mathbb{P}\left(\frac{X-\mu}{0.04} \leq \frac{4-\mu}{0.04}\right) = \mathbb{P}\left(Z \leq \frac{4-\mu}{0.04}\right)$$

According to the hint, $\mathbb{P}(Z \leq -2.05) = 0.02$

We conclude that $\frac{4-\mu}{0.04} = -2.05$, that is,

$$\mu = 4 + 0.04 \times 2.05 = 4.082 \text{ ounces}$$

Normal distribution: quantiles

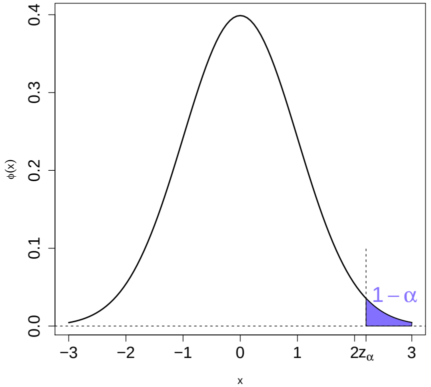
As in the previous example, we are sometimes given a probability and asked to find the corresponding value z

For instance, for any $\alpha \in (0,1)$, let z_α be such that

$$\mathbb{P}(Z > z_\alpha) = 1 - \alpha.$$

$$\text{i.e.,}\quad \mathbb{P}(Z < z_\alpha) = \alpha,$$

for $Z \sim \mathcal{N}(0,1)$

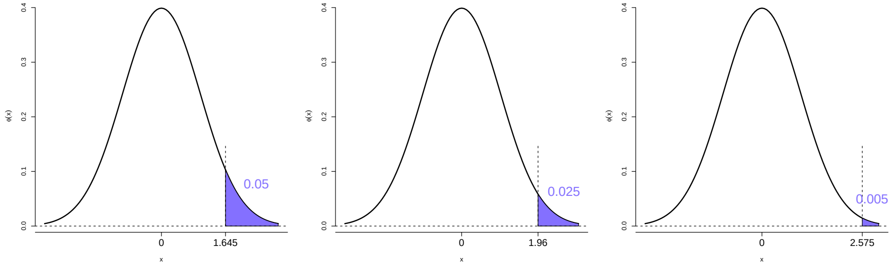


This value z_α is called the **quantile** of level α of the standard normal distribution

Normal distribution: quantiles

Some particular quantiles will be used extensively in subsequent chapters. These are the quantiles of level 0.95, 0.975 and 0.995:

$$\mathbb{P}(Z > 1.645) = 0.05, \quad \mathbb{P}(Z > 1.96) = 0.025, \quad \mathbb{P}(Z > 2.575) = 0.005$$



Note: by symmetry of the normal pdf, it is easy to see that

$$\boxed{z_{1-\alpha} = -z_\alpha}$$

→ for instance, $\mathbb{P}(Z < -1.96) = 0.025$

Some further properties of the Normal distribution

We know that if $X \sim \mathcal{N}(\mu, \sigma)$, then $aX + b$ is also normally distributed, for any real values a and b.

This generalises further: if $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$, and X_1 and X_2 are independent, then $aX_1 + bX_2$ is also normally distributed for any real values a and b.

Also, we can compute the parameters of the resulting distribution.

Property

Suppose $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$, $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$ and X_1 and X_2 are **independent**. Then, for any real values a and b,

$$aX_1 + bX_2 \sim \mathcal{N}\left(a\mu_1 + b\mu_2, \sqrt{a^2\sigma_1^2 + b^2\sigma_2^2}\right)$$

Some further properties of the Normal distribution

As a direct application of the preceding property, we have, with $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$, X_1 and X_2 independent,

$$X_1 + X_2 \sim \mathcal{N}\left(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}\right), \quad X_1 - X_2 \sim \mathcal{N}\left(\mu_1 - \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}\right)$$

Besides, the previous property can be readily extended to an arbitrary number of independent normally distributed random variables.

Example

Let X_1, X_2, X_3 represent the times necessary to perform three successive repair tasks at a certain service facility. Suppose they are independent normal random variables with expected values μ_1, μ_2 and μ_3 and variances σ_1^2, σ_2^2 and σ_3^2, respectively. What can be said about the distribution of $X_1 + X_2 + X_3$?

From the previous property, we can conclude that

$$X_1 + X_2 + X_3 \sim \mathcal{N}\left(\mu_1 + \mu_2 + \mu_3, \sqrt{\sigma_1^2 + \sigma_2^2 + \sigma_3^2}\right)$$

Some further properties of the Normal distribution

Example (ctd.)

If $\mu_1 = 40$ min, $\mu_2 = 50$ min and $\mu_3 = 60$ min, and $\sigma_1^2 = 10$ min^2, $\sigma_2^2 = 12$ min^2 and $\sigma_3^2 = 14$ min^2, what is the probability that the full task would take less than 160 min? (**Hint:** $\Phi(1.67) = 0.9525$.)

From the above, we have

$$X \doteq X_1 + X_2 + X_3 \sim \mathcal{N}\left(150, \sqrt{36} = 6\right)$$

Hence,

$$\mathbb{P}(X \leq 160) = \mathbb{P}\left(Z \leq \frac{160 - 150}{6}\right) = \mathbb{P}(Z \leq 1.67) = \Phi(1.67) = 0.9525$$

Checking if the data are normally distributed

Fact

Many of the statistical techniques presented in the coming chapters are based on an assumption that the distribution of the random variable of interest is normal.

→ In many instances, we will need to check whether a given sample has been generated by a normal random variable

How do we do that ?

Although they involve an element of subjective judgement, graphical procedures are the most helpful for detecting serious departures from normality.

Some of the visual displays we have used earlier, such as the density histogram, can provide a first insight about the form of the underlying distribution.

Density histograms to check for normality

Think of a density histogram as a piecewise constant function $h_n(x)$, where n is the number of observations in the data set

→ then, if the r.v. X having generated the data has density f on a support S_X, it can be shown that, under some regularity assumptions, for any $x \in S_X$,

$$h_n(x) \to f(x)$$

as $n \to \infty$ (and the number of classes $\to \infty$)

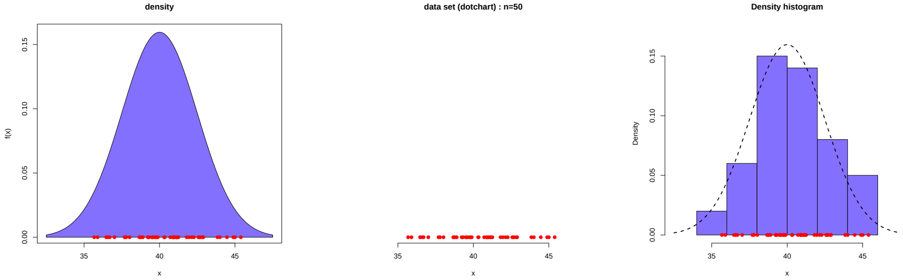
(the convergence "$h_n(x) \to f(x)$" has to be understood in a particular probabilistic sense, but details are beyond the scope of this course)

Concretely, **the larger the number of observations, the more similar the density histogram and the 'true' (unknown) density f are**

→ look at the histogram and decide whether it looks enough like the symmetric 'bell-shaped' normal curve or not

Density histograms to check for normality

Suppose we have a data set of size $n = 50$, drawn from a normal distribution

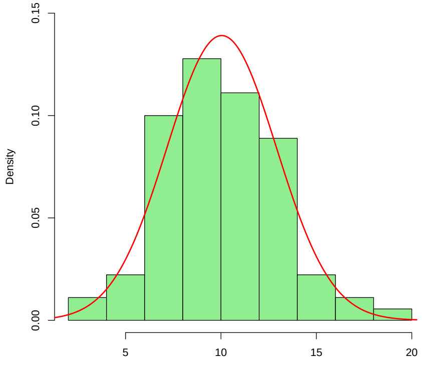


→ the density histogram 'looks like' the bell-shaped curve

Also, as both $f(x)$ and the density histogram are scaled such that the blue-ish purple areas are 1, they are easily superimposed and compared.

Density histograms to check for normality

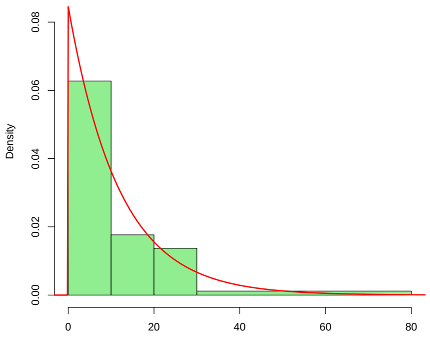
Look again at the (density) histogram on Slide 47



→ the histogram is symmetric and bell-shaped, without outliers

→ the normality assumption is reasonable

Look again at the density histogram on Slide 56



→ clear lack of symmetry (skewed to the right)

→ serious departure from normality (Exponential?)

Quantile plots

Density histograms are easy to use; however, they are usually not really reliable indicators of the distributional form unless the number of observations is very large.

→ another special graph, called a **normal quantile plot**, is more effective in detecting departure from normality

The plot essentially compares the data ordered from smallest to largest with what to expect to get for the smallest to largest in a sample if the theoretical distribution from which the data have come is normal

→ if the data were effectively selected from the normal distribution, the two sets of values should be reasonably close to one another

Note: a quantile plot is also sometimes called **qq-plot**

(→ command in Matlab: `qqplot`)

Quantile plots

Procedure for building a quantile plot:

- observations $\{x_1, x_2, \dots, x_n\}$
- ordered observations: $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$
- cumulative probabilities $\alpha_i = \frac{i-0.5}{n}$, for all $i = 1, \dots, n$
- standard normal quantiles of level α_i: for all $i = 1, \dots, n$, z_{α_i} chosen such that $\mathbb{P}(Z \leq z_{\alpha_i}) = \alpha_i$, where $Z \sim \mathcal{N}(0,1)$
- Quantile plot: plot the n pairs $(x_{(i)}, z_{\alpha_i})$

If the sample comes from the Standard Normal distribution, $x_{(i)} \simeq z_{\alpha_i}$ and the points would fall close to a 45°straight line passing by (0,0).

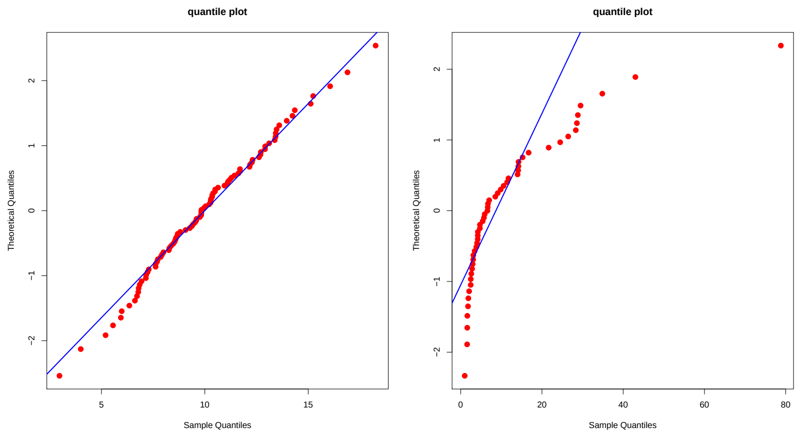
If the sample comes from some other normal distribution, the points would still fall around a straight line, as there is a linear relationship between the quantiles of $\mathcal{N}(\mu, \sigma)$ and the standard normal quantiles.

Fact

If the sample comes from some normal distribution, the points should follow (at least approximately) a straight line.

Quantile plots: examples

The figure below displays quantile plots for the previous two examples



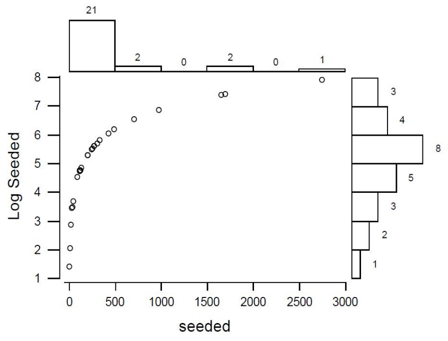
→ the normality assumption appears acceptable for the first data set, not at all for the second

Transforming observations

When the density histogram and the qq-plot indicate that the assumption of a normal distribution is invalid, transformations of the data can often improve the agreement with normality

Scientists regularly express their observations in natural logs.

Let's look again at the seeded clouds rainfall data (Slide 10)



Transforming observations

Apart from the log, other transformations may be useful:

$$\frac{-1}{x}, \quad \sqrt{x}, \quad \sqrt[4]{x}, \quad x^2, \quad x^3$$

If the observations are positive and the distribution has a long tail on the right, then concave transformations like $\log(x)$ or $\sqrt{x}$ put the large values down farther than they pull the central or small values → observations 'more symmetric'

Convex transformations work the other way

If the transformed observations are approximately normal (check with a quantile plot), it is usually advantageous to use the normality of this new scale to perform any statistical analysis.

Objectives

Now you should be able to:

- Understand the assumptions for some common discrete and continuous probability distributions □
- Select an appropriate discrete or continuous probability distribution to calculate probabilities in specific applications □
- Calculate probabilities, determine means and variances for some common discrete and continuous probability distributions □
- Calculate probabilities, determine means and variances for some common continuous probability distributions □
- Standardise normal random variables, and understand why this is useful □
- Use the cdf of a standard normal distribution to determine probabilities of interest □
- Check if a given sample may have been reasonably generated by a normal distribution □

Recommended exercises

→ Q53 p.54, Q55 p.55, Q57 p.55, Q33 p.221, Q37 p.222, Q23 p.78, Q19 p.31, Q23 p.31, Q31, Q33 p.41, Q35, Q37 p.42, Q62 p.56, Q73 p.58, Q27 p.78, Q47 p.93, Q65 p.97, Q51 (2nd edition)

→ Q54 p.56, Q56 p.57, Q58 p.57, Q35 p.226, Q39 p.227, Q24 p.79, Q19 p.34, Q23 p.35, Q31 p.44, Q33, Q35, Q37 p.45, Q64 p.58, Q75 p.60, Q28 p.79, Q66 p.100 (3rd edition)

6 Sampling distributions

Statistical Inference: Introduction

In this chapter, we introduce the last main topic for this course: **statistical inference** (that will keep us busy until the end).

Recall (Chapter 1) the general problem that is addressed:

- statistical methods are used to draw conclusions and make decisions about a **population** of interest
- however, for some reasons, we have no access to the whole population and we must do with observations on a subset of the population only. That subset is called the **sample**
- if the sample is effectively representative of the population, what we observe on the sample can be generalised to the population as a whole, at least to some extent ...
- ... taking chance factors properly into account

→ what we have learned about descriptive statistics, probability and random variables in the previous chapters, will play important roles here

Statistical Inference: Introduction

Populations are often described by the distribution of the variable of interest; e.g., we refer to a 'normal population', when the variable of interest is thought to be normally distributed

In statistical inference, we focus on drawing conclusions about one parameter of the distribution describing the population

In engineering, the parameters we are mainly interested in are

- the **mean** μ of the population
- the **variance** σ^2 (or standard deviation σ) of the population
- the **proportion** π of individuals in the population that belong to a class of interest
- the **difference in means of two sub-populations,** $\mu_1 - \mu_2$
- the **difference in two sub-population proportions,** $\pi_1 - \pi_2$

These population parameters are unknown
(otherwise, no need to make inferences about them)

→ the first part of the process is thus to estimate them

Random sampling

The importance of **random sampling** has been emphasised in Lecture 1:

- to assure that a sample is representative of the population from which it is obtained, and
- to provide a framework for the application of probability theory to problems of sampling

As we said, the assumption of random sampling is very important: if the sample is not random and is based on judgement or flawed in some other way, then statistical methods will not work properly and will lead to incorrect decisions.

Therefore, we should now properly define a random sample.

Random sampling

Before a sample of size n is selected at random from the population, the observations are modelled as random variables $X_1, X_2, \ldots, X_n$.

Definition

The set of observations $X_1, X_2, \ldots, X_n$ constitutes a random sample if

1. the X_i's are independent random variables, and
2. every X_i has the same probability distribution

This is often abbreviated to i.i.d., for 'independent and identically distributed' → it is common to talk about an i.i.d. sample

We also apply the terms 'random sample' to the set of observed values

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

of the random variables, but this should not cause any confusion.

Note: as usual, the lower case distinguishes the realisation of a random sample (the actual data) from the upper case, which represents the random variables before they are observed

Statistic, estimator and sampling distribution

A numerical measure calculated from the sample is called a **statistic**

Denote the unknown parameter of interest θ (so this can be μ, σ^2, π, or any other parameter of interest)

The only information we have to estimate that parameter θ is the information contained in the sample

$\rightarrow$ an **estimator** of θ must be a statistic, i.e. a function of the sample

$$\hat{\Theta} = h(X_1, X_2, \dots, X_n)$$

Note that an estimator is a random variable, as it is a function of random variables $\rightarrow$ it must have a distribution

That distribution is called a **sampling distribution**, it generally depends on the population distribution and the sample size

After the sample has been selected, $\hat{\Theta}$ takes on a particular value $\hat{\theta} = h(x_1, x_2, \dots, x_n)$, called the **estimate** of θ

Estimation: some remarks

Remark: the hat notation conventionally distinguishes the sample-based quantities (estimator $\hat{\Theta}$ or estimate $\hat{\theta}$) from the 'true' population parameter (θ)

Also, as usual, capital letters denote the random variables, like $\hat{\Theta}$, whereas lower-case letters are for particular numerical values, like $\hat{\theta}$

Two notable exceptions are:

- the sample mean, usually denoted $\bar{X}$; its observed value, calculated once we have observed a sample $x_1, x_2, \dots, x_n$, is denoted $\bar{x}$ (Slide 60)
- the sample standard deviation (variance), usually denoted S (S^2); its observed value, calculated once we have observed a sample $x_1, x_2, \dots, x_n$, is denoted s (s^2) (Slides 72 and 70)

An example: estimating μ in a normal population

Suppose that we have a normal population, i.e., the random variable X of interest is such that $X \sim \mathcal{N}(\mu, \sigma)$

Suppose that we are only interested in the unknown mean μ (unknown), and for simplicity suppose that σ is *known*

If we had a random sample $X_1, X_2, \dots, X_n$ from the population

– that is, $X_i \sim \mathcal{N}(\mu, \sigma)$ for all i and the X_i's are independent –

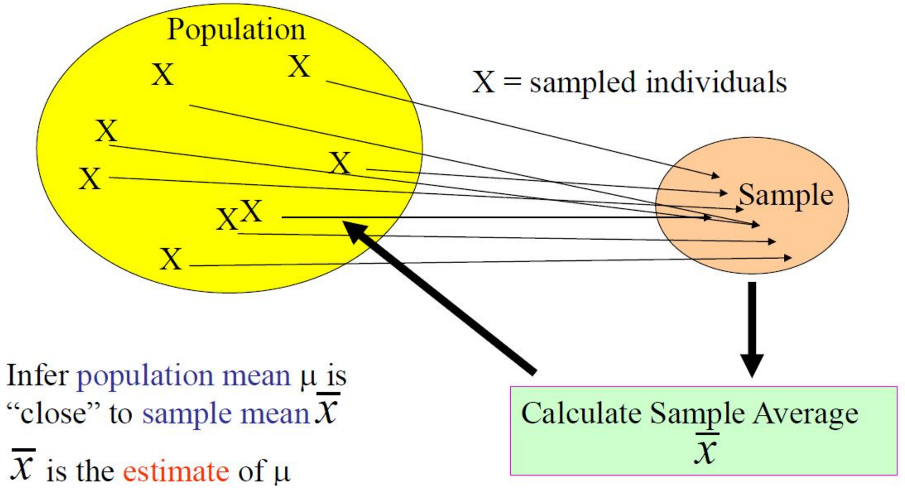
we could naturally estimate μ by the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ ($\rightsquigarrow$ **estimator**)

Now, draw such a sample and observe $x_1, x_2, \dots, x_n$

The observed sample mean is $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ ($\rightsquigarrow$ **estimate**)

As the sample is random, it should be representative of the whole population, and it is reasonable to believe that the population mean μ should be "close" to the observed sample mean $\bar{x}$

An example: estimating μ in a normal population



Infer population mean μ is "close" to sample mean $\bar{x}$

$\bar{x}$ is the estimate of μ

Calculate Sample Average $\bar{x}$

An example: estimating μ in a normal population

What does that mean, μ should be "close" to $\bar{x}$?

We know that each X_i in the sample follows the $\mathcal{N}(\mu, \sigma)$ distribution, and they are independent (i.i.d. sample).

Then, because linear combinations of independent normal r.v. remain normally distributed, we conclude that

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

has a normal distribution with expectation

$$\mathbb{E}\left(\bar{X}\right) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) \overset{\text{i.d.}}{=} \frac{1}{n}\sum_{i=1}^{n}\mu = \mu$$

and variance

$$\mathbb{V}\text{ar}\left(\bar{X}\right) = \mathbb{V}\text{ar}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \overset{\text{i.}}{=} \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\text{ar}(X_i) \overset{\text{i.d.}}{=} \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}$$

An example: estimating μ in a normal population

→ in a normal population, the sampling distribution of $\bar{X}$ is

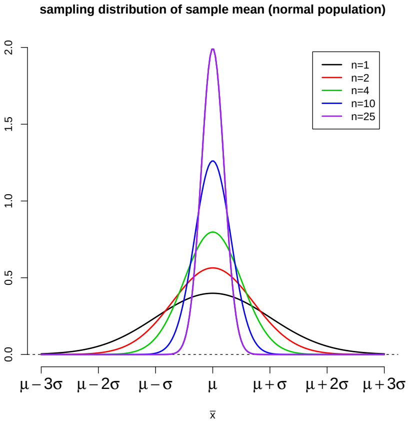
$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

$\bar{X}$ is a normal random variable, centred about the 'true' population mean μ, and with spread becoming more and more reduced as the sample size increases

→ **the larger the sample, the more accurate the estimation!**

(recall example on Slide 165)

The observed $\bar{x}$ is a value drawn from that sampling distribution



Objectives

Now you should be able to:

- Explain the general concepts of estimating the parameters of a population, in particular the difference between estimator and estimate, and the role played by the sampling distribution of an estimator □
- Illustrate those concepts with the particular case of the estimation of the mean in a normal population

Recommended exercises

→ Q53 p.237 (2nd edition)

→ Q53, Q55 p.241 (3rd edition)

7 Inferences concerning a mean

Introduction

The purpose of most **statistical inference** procedures is to generalise the information contained in an observed random sample to the population from which the sample were obtained.

This can be divided into two major areas:

- **estimation**, including point estimation and interval estimation
- **tests of hypotheses**

In this chapter we will present some theory and largely illustrate it with some methods which pertain to the estimation of means.

Point estimation

Recall that we wish to estimate an unknown parameter θ of a population. For instance, the population mean μ, from a random sample of size n, say $X_1, X_2, \ldots, X_n$.

To do so, we select an estimator, which must be a statistic (i.e. a value computable from the sample), say $\hat{\Theta} = h(X_1, X_2, \ldots, X_n)$

→ **an estimator is a random variable**, which has its mean, its variance and its probability distribution, known as the sampling distribution

For instance, to estimate the population mean μ, we suggested in the previous chapter to use the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

We derived that

$$\mathbb{E}(\bar{X}) = \mu \qquad \text{and } \mathbb{V}\text{ar}(\bar{X}) = \frac{\sigma^2}{n},$$

where σ^2 is the population variance

Specifically, if $X_i \sim \mathcal{N}(\mu, \sigma)$ for all i, we showed $\boxed{\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)}$

Properties of estimators

The choice of the sample mean to estimate the population mean seems quite natural. However, there are many other estimators that can be used to calculate an estimate. Why not:

- $\hat{\Theta}_1 = X_1$, the first observed value;
- $\hat{\Theta}_2 = (X_1 + X_n)/2$;
- $\hat{\Theta}_3 = (aX_1 + bX_n)/(a + b)$, for two constants a, b ($a + b \neq 0$)

→ criteria for selecting the 'best' estimator are needed

What do we expect from an estimator for θ ?

→ certainly that it should give estimates reasonably close to θ, the parameter it is supposed to estimate

However, this 'closeness' is not easy to comprehend: first, θ is unknown, and second, the estimator is a random variable

→ we have to properly define what "close" means in this situation

Properties of estimators: unbiasedness

The first desirable property that a good estimator should possess is that it is **unbiased**.

Definition

An estimator $\hat{\Theta}$ of θ is said to be unbiased if and only if its expectation is equal to θ, i.e.

$$\mathbb{E}(\hat{\Theta}) = \theta$$

→ an estimator is unbiased if "on the average" its values will equal the parameter it is supposed to estimate

If an estimator is not unbiased, then the difference

$$\mathbb{E}(\hat{\Theta}) - \theta$$

is called the **bias** of the estimator → systematic error

For instance, we showed that $\mathbb{E}(\bar{X}) = \mu$

→ the sample mean $\bar{X}$ is an unbiased estimator for μ

Properties of estimators: unbiasedness

The property of unbiasedness is one of the most desirable properties of an estimator, although it is sometimes outweighted by other factors.

One shortcoming is that it will generally not provide a unique estimator for a given estimation problem.

For instance, for the above defined estimators for μ,

$$\mathbb{E}(\hat{\Theta}_1) = \mathbb{E}(X_1) = \mu$$

$$\mathbb{E}(\hat{\Theta}_2) = \mathbb{E}\left(\frac{X_1 + X_n}{2}\right) = \frac{1}{2}(\mathbb{E}(X_1) + \mathbb{E}(X_n)) = \frac{1}{2}(\mu + \mu) = \mu$$

$$\mathbb{E}(\hat{\Theta}_3) = \mathbb{E}\left(\frac{aX_1 + bX_n}{a+b}\right) = \frac{1}{a+b}(a\mathbb{E}(X_1) + b\mathbb{E}(X_n)) = \frac{1}{a+b}(a\mu + b\mu) = \mu$$

→ $\hat{\Theta}_1$, $\hat{\Theta}_2$ and $\hat{\Theta}_3$ are also unbiased estimators for μ

→ we need a further criterion for deciding which of several **unbiased** estimators is best for estimating a given parameter

Properties of estimators: efficiency

That further criterion becomes evident when we compare the variances of $\bar{X}$ and $\hat{\Theta}_1$.

We have shown that $\mathbb{V}\text{ar}(\bar{X}) = \frac{\sigma^2}{n}$, while we have

$$\mathbb{V}\text{ar}(\hat{\Theta}_1) = \mathbb{V}\text{ar}(X_1) = \sigma^2$$

→ the variance of $\bar{X}$ is n times smaller than the variance of $\hat{\Theta}_1$!

→ it is far more likely that $\bar{X}$ will be closer to its mean, μ, than $\hat{\Theta}_1$ is to μ

→ the observed $\bar{x}$ will be much more accurate than x_1 as an estimate of μ!

Fact

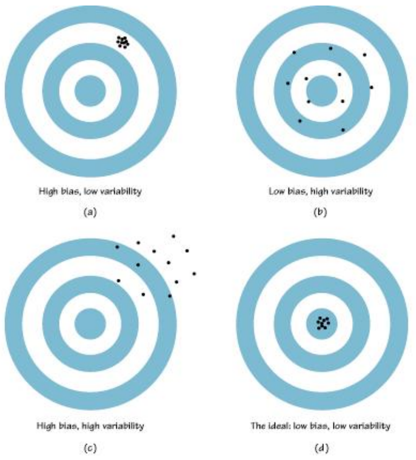
Estimators with smaller variances are more likely to produce estimates close to the true value θ

→ a logical principle of estimation is to choose the unbiased estimator that has minimum variance

Such an estimator is said to be **efficient**

Properties of estimators

A useful analogy is to think of each value taken by an estimator as a shot at a target, the target being the population parameter of interest


High bias, low variability
(a)
Low bias, high variability
(b)
High bias, high variability
(c)
The ideal: low bias, low variability
(d)

Properties of estimators: consistency

Consider the 'minimum variance' argument as the sample size increases:

We desire an estimator that is more and more likely to be close to θ as the number of observations increases.

Namely, we require that the probability that the estimator is 'close' to θ increases to one as the sample size increases.

Such estimators are called **consistent**.

An easy way to check that an unbiased estimator is consistent is to show that its variance decreases to 0 as n increases to ∞.

For instance, $\mathbb{V}\text{ar}(\bar{X}) = \frac{\sigma^2}{n} \to 0$ as $n \to \infty$, i.e. **$\bar{X}$ is consistent for μ**

On the other hand, it can be verified that

$$\mathbb{V}\text{ar}(\hat{\Theta}_1) = \sigma^2 \nrightarrow 0, \quad \mathbb{V}\text{ar}(\hat{\Theta}_2) = \frac{\sigma^2}{2} \nrightarrow 0, \quad \mathbb{V}\text{ar}(\hat{\Theta}_3) = \sigma^2 \frac{a^2+b^2}{(a+b)^2} \nrightarrow 0$$

→ none of them are consistent

Sample mean

We have seen thus far that the sample mean $\bar{X}$ is unbiased and consistent as an estimator of the population mean μ.

It can be also shown that in most practical situations where we estimate the population mean μ, the variance of no other estimator is less than the variance of the sample mean.

Fact

In most practical situations, the sample mean is a very good estimator for the population mean μ.

Note: there exist several other criteria for assessing the goodness of point estimation methods, but we shall not discuss them in this course

Here, we will always use the sample mean $\bar{X}$ when we will have to estimate the population mean μ.

Standard error of a point estimate

Although we estimate the unknown population parameter θ with an estimator $\hat{\Theta}$ that we know to have certain desirable properties (unbiasedness, consistency), the chances are slim, virtually non existent, that the estimate $\hat{\theta}$ will actually equal θ.

→ **an estimate remains an approximation of the true value!**

→ it is unappealing to report your estimate only, as there is nothing inherent in $\hat{\theta}$ that provides any information about how close it is to θ

Hence, it is usually desirable to give some idea of the precision of the estimation → the measure of precision usually employed is the standard error of the estimator.

Definition

The **standard error** of an estimator $\hat{\Theta}$ is its standard deviation sd$(\hat{\Theta})$.

Note: If the standard error involves some unknown parameters that can be estimated, substitution of those values into sd$(\hat{\Theta})$ produces an estimated standard error, denoted $\widehat{\text{sd}(\hat{\Theta})}$

Standard error of the sample mean

Suppose again that we estimate the mean μ of a population with the sample mean $\bar{X}$ calculated from a random sample of size n.

We know that $\mathbb{E}(\bar{X}) = \mu$ and $\mathbb{V}\text{ar}(\bar{X}) = \frac{\sigma^2}{n}$, so the standard error of $\bar{X}$ as an estimator of μ is

$$\text{sd}(\bar{X}) = \frac{\sigma}{\sqrt{n}},$$

However, we cannot report a numerical value for this standard error as it depends in turn on the population standard deviation σ, which is usually unknown.

→ we have a natural estimate of the population standard deviation given by the observed sample standard deviation (Slide 72)

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

→ estimate the standard error sd$(\bar{X})$ with $\widehat{\text{sd}(\bar{X})} = \frac{s}{\sqrt{n}}$

Interval estimation: introduction

When the sampling distribution of the estimator is normal, we can be 'reasonably confident' that the true value of the parameter lies within two standard errors of the estimate (recall the 68-95-99-rule, Slide 217)

→ it is often easy to determine an interval of plausible values for a parameter

→ such observations are the basis of **interval estimation**

→ instead of giving a point estimate $\hat{\theta}$, that is a single value that we know not to be equal to θ anyway, we give an interval in which we are very confident to find the true value,

and we specify what "very confident" means

Basic interval estimation: example

Example

An article in the *Journal of Heat Transfer* described a new method of measuring thermal conductivity of Armco iron. At 100°F and a power input of 550 Watts, the following measurements of thermal conductivity (in BTU/hr-ft-°F) were obtained:

41.60, 41.48, 42.34, 41.95, 41.86, 42.18, 41.72, 42.26, 41.81, 42.04

A point estimate of the 'true' thermal conductivity μ of Armco iron (at 100°F and 550 Watts) is the observed sample mean

$$\bar{x} = \frac{1}{10}(41.60 + 41.48 + \ldots + 42.04) = 41.924 \text{ BTU/hr-ft-}^\circ\text{F}$$

We know that the standard error of the sample mean as an estimator for μ is $\text{sd}(\bar{X}) = \sigma/\sqrt{n}$, and since σ is unknown, we may replace it by the sample standard deviation $s = \ldots = 0.284$ (BTU/hr-ft-°F) to obtain the estimated standard error

$$\widehat{\text{sd}(\bar{X})} = \frac{s}{\sqrt{n}} = \frac{0.284}{\sqrt{10}} = 0.0898 \text{ BTU/hr-ft-}^\circ\text{F}$$

Basic interval estimation: example

→ the standard error is about 0.2 percent of the sample mean

→ we have a relatively precise point estimate of the 'true' thermal conductivity μ under those conditions

We have that 2 times the (estimated) standard error is

$$\widehat{2\text{sd}(\bar{X})} = 2 \times 0.0898 = 0.1796$$

Hence, if we can assume that thermal conductivity is normally distributed (can we?), we are 'highly confident' that the true thermal conductivity μ is within the interval

$$[41.924 \pm 0.1796] = [41.744, 42.104]$$

'Highly confident' here means '$\sim$ 95% confident',
by the 68-95-99 rule of Normal distributions

Confidence intervals

The preceding interval is called a confidence interval.

Definition

A **confidence interval** is an interval for which we can assert with a reasonable degree of certainty (or confidence) that it will contain the true value of the population parameter under consideration.

A confidence interval is always calculated by first selecting a confidence level, which measures its degree of reliability

→ a confidence interval of level $100 \times (1-\alpha)\%$ means that we are $100 \times (1-\alpha)\%$ confident that the true value of the parameter is included into the interval (α is a real number in $[0, 1]$)

The most frequently used confidence levels are 90%, 95% and 99%

→ the higher the confidence level, the more strongly we believe that the value of the parameter being estimated lies within the interval

Confidence intervals: remarks

Remark 1: information about the precision of estimation is conveyed by the length of the interval: a short interval implies precise estimation, a wide interval, however, gives the message that there is a great deal of uncertainty concerning the parameter that we are estimating.

Note that the higher the level of the interval, the wider it must be!

Remark 2 : it is sometimes tempting to interpret a $100 \times (1-\alpha)\%$ confidence interval for θ as saying that there is a $100 \times (1-\alpha)\%$ probability that θ belongs to it

→ this is *not really true*!

Fact

The $100 \times (1-\alpha)\%$ refers to the percentage of all samples of the same size possibly drawn from the same population which would produce an interval containing the true θ.

Confidence intervals: remarks

Remark 2: (ctd.)

→ if we consider taking sample after sample from the population and use each sample separately to compute $100 \times (1-\alpha)\%$ confidence intervals, then in the long-run roughly $100 \times (1-\alpha)\%$ of these intervals will capture θ

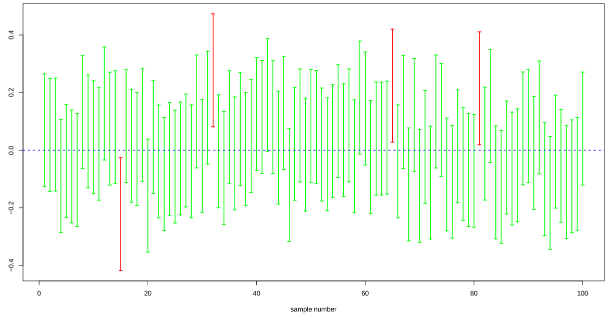
A correct probabilistic interpretation lies in the realisation that a confidence interval is a random interval, because its end-points are calculated from a random sample and are therefore random variables.

However, once the confidence interval has been computed, the true value either belongs to it or does not belong to it, and any probability statement is pointless.

That is why we use the term “confidence level” instead of “probability”

Confidence intervals: remarks

As an illustration, we successively computed 95%-confidence intervals for μ for 100 random samples of size 100 independently drawn from a $\mathcal{N}(0,1)$ population



→ 96 intervals out of 100 ($\simeq$ 95%) contain the true value $\mu = 0$

Of course in practice we do not know the true value of μ, and we cannot tell whether the interval we have computed is one of the ‘good’ 95% intervals or one of the ‘bad’ 5% intervals.

Confidence interval on the mean of a normal distribution, variance known

The basic ideas for building confidence intervals are most easily understood by first considering a simple situation:

Suppose we have a normal population with unknown mean μ and known variance σ^2.

Note that this is somewhat unrealistic, as typically both the mean and the variance are unknown.

→ we will address more general situations later

We have thus a random sample $X_1, X_2, \ldots, X_n$, such that, for all i,

$$X_i \sim \mathcal{N}(\mu, \sigma),$$

with μ unknown and σ a known constant

→ we would like a confidence interval for μ

Confidence interval on the mean of a normal distribution, variance known

In that situation, we know (Slide 250) that

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

We may standardise this normally distributed random variable:

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

Suppose we desire a confidence interval for μ of level $100 \times (1-\alpha)\%$.

From our random sample, this can be regarded as a ‘random interval’, say $[L, U]$, where L and U are statistics (i.e. computable from the sample) such that

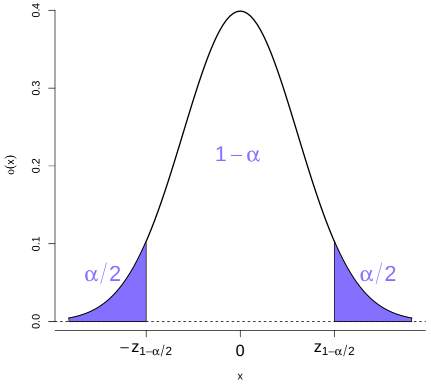
$$\mathbb{P}([L,U] \ni \mu) = \mathbb{P}(L \leq \mu \leq U) = 1-\alpha$$

Confidence interval on the mean of a normal distribution, variance known

In our situation, because $Z \sim \mathcal{N}(0,1)$, we may write

$$\mathbb{P}(-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}) = 1-\alpha$$

where $z_{1-\alpha/2}$ is the quantile of level $1-\alpha/2$ of the standard normal distribution



Hence it is the case that

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

Confidence interval on the mean of a normal distribution, variance known

Isolating μ, it follows

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

→ here are L and U, two statistics such that

$$\mathbb{P}(L \leq \mu \leq U) = 1-\alpha$$

→ L and U will yield the bounds of the confidence interval!

→if $\bar{x}$ is the sample mean of an observed random sample of size n from a normal distribution with known variance σ^2, a confidence interval of level $100 \times (1-\alpha)\%$ for μ is given by

$$\boxed{\left[\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]}$$

Confidence interval on the mean of a normal distribution, variance known

From Slide 225, $z_{0.95} = 1.645$, $z_{0.975} = 1.96$ and $z_{0.995} = 2.575$

→ a confidence interval for μ of level 90% is

$$\left[\bar{x} - 1.645 \times \tfrac{\sigma}{\sqrt{n}}, \bar{x} + 1.645 \times \tfrac{\sigma}{\sqrt{n}}\right]$$

→ a confidence interval for μ of level 95% is

$$\left[\bar{x} - 1.96 \times \tfrac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \times \tfrac{\sigma}{\sqrt{n}}\right]$$

→ a confidence interval for μ of level 99% is

$$\left[\bar{x} - 2.575 \times \tfrac{\sigma}{\sqrt{n}}, \bar{x} + 2.575 \times \tfrac{\sigma}{\sqrt{n}}\right]$$

We see that the respective lengths of these intervals are

$$3.29 \times \frac{\sigma}{\sqrt{n}},\ 3.92 \times \frac{\sigma}{\sqrt{n}} \text{ and } 5.15 \times \frac{\sigma}{\sqrt{n}}$$

Sample Size for CI on the mean

The length of a CI is a measure of the precision of the estimation
→ the precision is inversely related to the confidence level

However, it is desirable to obtain a confidence interval that is both

- short enough for decision-making purposes
- of an adequate confidence level

→ one way to reduce the length of a confidence interval with prescribed confidence level is by choosing n large enough

From the above, we know that in using $\bar{x}$ to estimate μ, the error $e = |\bar{x} - \mu|$ is less than $z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$ with $100 \times (1-\alpha)\%$ confidence

→ in other words, we can be $100 \times (1-\alpha)\%$ confident that the error will not exceed a given amount e when the sample size is

$$\boxed{n = \left(\frac{z_{1-\alpha/2}\sigma}{e}\right)^2}$$

Confidence interval on the mean of a normal distribution, variance known: example

Example

The Charpy V-notch (CVN) technique measures impact energy and is often used to determine whether or not a material experiences a ductile-to-brittle transition with decreasing temperature. Ten measurements of impact energy (in J) on specimens of steel cut at 60°C are as follows:

$$64.1, 64.7, 64.5, 64.6, 64.5, 64.3, 64.6, 64.8, 64.2, 64.3$$

Assume that impact energy is normally distributed with $\sigma = 1$ J. a) Find a 95% CI for μ, the mean impact energy for that kind of steel

An elementary computation yields $\bar{x} = 64.46$ J. With $n = 10$, $\sigma = 1$ and $\alpha = 0.05$, direct application of the previous results gives a 95% CI as follows:

$$\left[\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = \left[64.46 - 1.96 \times \frac{1}{\sqrt{10}}, 64.46 + 1.96 \times \frac{1}{\sqrt{10}}\right]$$
$$= [63.84, 65.08]$$

Confidence interval on the mean of a normal distribution, variance known: example

Example (ctd.)

b) Determine how many specimens we should test to ensure that the 95% CI on the mean impact energy μ has a length of at most 1 J

The length of the CI in part a) is 1.24 J. If we desire a higher precision, namely a confidence interval length of 1 J, then we need more than 10 observations

The bound on error estimation e is one-half of the length of the CI, thus use the expression on Slide 276 with $e = 0.5$, $\sigma = 1$ and $\alpha = 0.05$:

$$n = \left(\frac{z_{1-\alpha/2}\sigma}{e}\right)^2 = \left(\frac{1.96 \times 1}{0.5}\right)^2 = 15.37$$

→ as n must be an integer, the required sample size is 16

Confidence interval on the mean of a normal distribution, variance known: remarks

Remark 1: if the population is normal, the confidence interval

$$\left[\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \qquad (\star)$$

is valid for all sample sizes $n \geq 1$

Remark 2: this interval is not the only $100 \times (1-\alpha)\%$ confidence interval for μ. For instance, starting from $\mathbb{P}(z_{\alpha/4} \leq Z \leq z_{1-3\alpha/4}) = 1 - \alpha$ on Slide 273, another $100 \times (1-\alpha)\%$ CI could be

$$\left[\bar{x} - z_{1-3\alpha/4}\frac{\sigma}{\sqrt{n}}, \bar{x} - z_{\alpha/4}\frac{\sigma}{\sqrt{n}}\right]$$

However, interval $(\star)$ is often preferable, as it is symmetric around $\bar{x}$

Also, it can be shown that the symmetric confidence interval $(\star)$

$$\left[\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

is the shortest one among all confidence intervals of level $1 - \alpha$

Confidence interval on the mean of a normal distribution, variance known: remarks

Remark 3: in the same spirit, we have

$$\mathbb{P}(Z \leq z_{1-\alpha}) = \mathbb{P}(-z_{1-\alpha} \leq Z) = 1 - \alpha$$

Hence,

$$\left(-\infty, \bar{x} + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right]$$

and

$$\left[\bar{x} - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}, +\infty\right)$$

are also $100 \times (1-\alpha)\%$ CI for μ

These are called one-sided confidence intervals, as opposed to $(\star)$ (two-sided CI).

They are also sometimes called (upper and lower) confidence bounds.

What if the distribution is not normal?

So far, we have assumed that the population distribution is normal. In that situation, we have (Slide 250)

$$Z = \sqrt{n}\,\frac{\bar{X} - \mu}{\sigma} \sim \mathcal{N}(0,1)$$

for any sample size n.

This sampling distribution is the cornerstone when deriving confidence intervals for μ, and directly follows from $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma)$.

A natural question is now:

What if the population is not normal? (X_i not $\mathcal{N}(\mu, \sigma)$)

→ surprisingly enough, the above results still hold most of the time, *at least approximately*, due to the so-called

Central Limit Theorem

The Central Limit Theorem

The Central Limit Theorem (CLT) is certainly one of the most remarkable results in probability

Loosely speaking, it asserts that

the sum of a large number of independent random variables has a distribution that is approximately normal

It was first postulated by Abraham de Moivre in 1733 who used the bell-shaped curve to approximate the distribution of the number of heads resulting from many tosses of a fair coin (see Slide 205)

However, this received little attention until the French mathematician Pierre-Simon Laplace (1749-1827) rescued it from obscurity in his monumental work "*Théorie Analytique des Probabilités*", which was published in 1812.

But it was not before 1901 that it was defined in general terms and formally proved by the Russian mathematician Aleksandr Lyapunov (1857-1918).

The Central Limit Theorem

Central Limit Theorem

If $X_1, X_2, \ldots, X_n$ is a random sample taken from a population with mean μ and finite variance σ^2, and if $\bar{X}$ is the sample mean, then the limiting distribution of

$$\sqrt{n}\,\frac{\bar{X} - \mu}{\sigma}$$

as $n \to \infty$, is the **standard normal distribution**

Proof: no proof provided

The Central Limit Theorem

When $X_i \sim \mathcal{N}(\mu, \sigma)$, $\sqrt{n}\,\frac{\bar{X}-\mu}{\sigma} \sim \mathcal{N}(0,1)$ for all n

What the CLT states is that, when the X_i's are not normal (whatever they are!), $\sqrt{n}\,\frac{\bar{X}-\mu}{\sigma} \sim \mathcal{N}(0,1)$ when n is infinitely large

→ the standard normal distribution provides a reasonable approximation to the distribution of $\sqrt{n}\,\frac{\bar{X}-\mu}{\sigma}$ when "n is large"

This is usually denoted

$$\boxed{\sqrt{n}\,\frac{\bar{X}-\mu}{\sigma} \overset{a}{\sim} \mathcal{N}(0,1)}$$

with $\overset{a}{\sim}$ for 'approximately follows' (or 'asymptotically ($n \to \infty$) follows')

The Central Limit Theorem

The power of the CLT is that it holds true **for any population distribution**, discrete or continuous! For instance,

$$X_i \sim \text{Exp}(\mu) \quad \Longrightarrow \quad \sqrt{n}\,\frac{\bar{X}-\mu}{\mu} \overset{a}{\sim} \mathcal{N}(0,1)$$

$$X_i \sim U_{[a,b]} \quad \Longrightarrow \quad \sqrt{n}\,\frac{\bar{X}-\frac{a+b}{2}}{\frac{b-a}{\sqrt{12}}} \overset{a}{\sim} \mathcal{N}(0,1)$$

$$X_i \sim \text{Bern}(\pi) \quad \Longrightarrow \quad \sqrt{n}\,\frac{\bar{X}-\pi}{\sqrt{\pi(1-\pi)}} \overset{a}{\sim} \mathcal{N}(0,1)$$

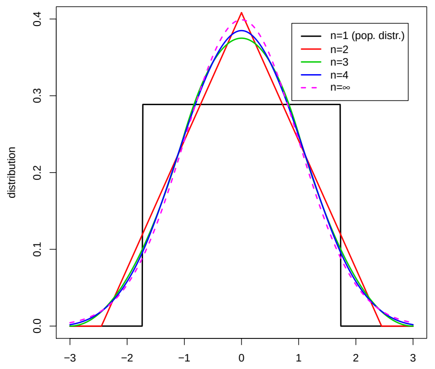
Facts:

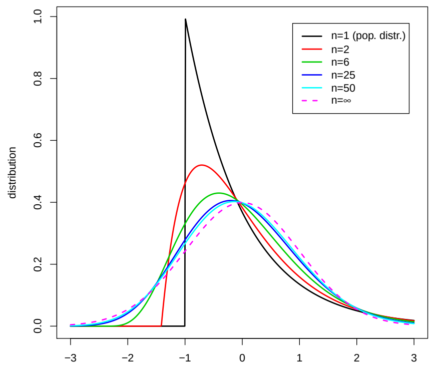
- the larger n, the better the normal approximation
- the closer the population distribution is to being normal, the more rapidly the distribution of $\sqrt{n}\,\frac{\bar{X}-\mu}{\sigma}$ approaches normality as n gets large

The Central Limit Theorem: illustration

Probability density functions for $\sqrt{n}\,\frac{\bar{X}-\mu}{\sigma}$

$\boxed{X_i \sim U_{[-\sqrt{3},\sqrt{3}]}}$ $(\mu = 0, \sigma = 1)$ $\boxed{X_i \sim \text{Exp}(1) - 1}$ $(\mu = 0, \sigma = 1)$

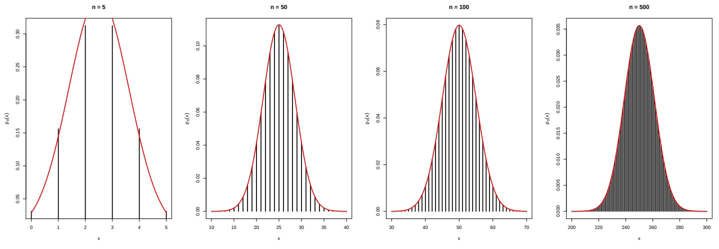




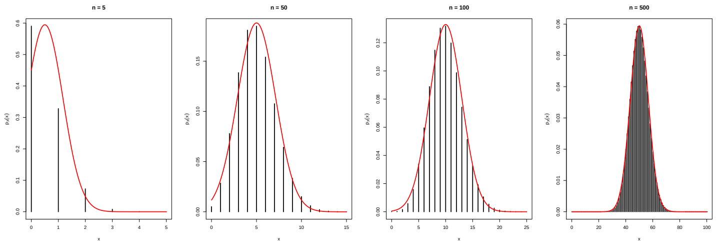
The Central Limit Theorem: illustration

Probability mass functions for $\sum_{i=1}^{n} X_i$, $X_i \sim \text{Bern}(\pi)$

$\pi = 0.5$:



$\pi = 0.1$:



The Central Limit Theorem: further illustration

Go to

http://onlinestatbook.com/stat_sim/sampling_dist/

The Central Limit Theorem: remarks

Remark 1:

The Central Limit Theorem not only provides a simple method for computing approximate probabilities for sums or averages of independent random variables.

It also helps explain why so many natural populations exhibit a bell-shaped (i.e., normal) distribution curve:

Indeed, as long as the behaviour of the variable of interest is dictated by a large number of independent contributions, it should be (at least approximately) normally distributed.

For instance, a person's height is the result of many independent factors, both genetic and environmental. Each of these factors can increase or decrease a person's height, just as each ball in Galton's board (Slide 208) can bounce to the right or the left

→ the Central Limit Theorem guarantees that the sum of these contributions has approximately a normal distribution

The Central Limit Theorem: remarks

Remark 2:

a natural question is '**how large n needs to be**' for the normal approximation to be valid

→ that depends on the population distribution!

A general rule-of-thumb is that one can be fairly confident of the normal approximation whenever the sample size n is at least 30

$$\boxed{n \geq 30}$$

Note that, in favourable cases (population distribution is not skewed and not long-tailed), the normal approximation will be satisfactory for much smaller sample sizes (like $n = 5$ in the uniform case, for instance)

If $n \geq 30$, the normal distribution will provide a good approximation to the sampling distribution of $\bar{X}$ irrespective of the shape of the population (well, except for variables with extremely right-skewed or very long-tailed distributions).

Confidence interval on the mean of an arbitrary distribution

The Central Limit Theorem allows us to use the procedures described earlier to derive confidence intervals for μ in an arbitrary population, bearing in mind that these will be **approximate confidence intervals** (whereas they were exact in a normal population)

Indeed, if n is large enough,

$$Z = \sqrt{n}\,\frac{\bar{X} - \mu}{\sigma} \overset{a}{\sim} \mathcal{N}(0,1),$$

hence

$$\mathbb{P}\left(-z_{1-\alpha/2} \leq \sqrt{n}\,\frac{\bar{X} - \mu}{\sigma} \leq z_{1-\alpha/2}\right) \simeq 1 - \alpha,$$

where $z_{1-\alpha/2}$ is the quantile of level $1 - \alpha/2$ of the standard normal distribution.

Confidence interval on the mean of an arbitrary distribution

It follows

$$\mathbb{P}\left(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \simeq 1-\alpha,$$

so that if $\bar{x}$ is the sample mean of an observed random sample of size n from any distribution with known variance σ^2, an approximate confidence interval of level $100 \times (1-\alpha)\%$ for μ is given by

$$\boxed{\left[\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]}$$

Note: because this result requires "n large enough" to be reliable, this type of interval, based on the CLT, is often called large-sample confidence interval.

One could also define large-sample one-sided confidence intervals of level $100 \times (1-\alpha)\%$: $(-\infty, \bar{x} + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}]$ and $[\bar{x} - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}, +\infty)$.

Objectives

Now you should be able to:

- Explain important properties of point estimators, including bias, variance, efficiency and consistency □
- Know how to compute and explain the precision with which a parameter is estimated □
- Understand the basics of interval estimation and explain what a confidence interval of level $100 \times (1-\alpha)\%$ for a parameter is □
- Construct exact confidence intervals on the mean of a normal distribution with known variance □
- Understand the Central Limit Theorem and explain the important role of the normal distribution in inference □
- Construct large sample confidence intervals on a mean of an arbitrary distribution □

Recommended exercises:

→ Q46+Q49, Q50 p.237, Q69 p.239, Q1, Q3 p.293, Q5, Q6 p.294, Q10, Q11 p.301 (2nd edition)

→ Q48+Q51, Q52 p.241, Q72 p.244, Q1, Q3 p.297, Q5, Q6 p.298, Q10 p.305, Q11 p.306 (3rd edition)

Confidence interval on the mean of a distribution, variance unknown

Previously we showed how to build confidence intervals for the mean μ of a distribution, assuming that the population variance σ^2 was known

→ this is probably not very realistic!

Suppose now that the population variance σ^2 is not known

→ we can no longer make practical use of the core result

$$Z = \sqrt{n}\,\frac{\bar{X}-\mu}{\sigma} \overset{(a)}{\sim} \mathcal{N}(0,1)$$

However, from the random sample $X_1, X_2, \ldots, X_n$ we have a natural estimator of the unknown σ^2: the **sample variance**

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

which will provide an estimate $s^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$ of σ^2 upon observation of a sample $x_1, x_2, \ldots, x_n$.

Confidence interval on the mean of a normal distribution, variance unknown

A natural procedure is thus to replace σ with the sample standard deviation S, and to work with the random variable

$$T = \sqrt{n}\,\frac{\bar{X}-\mu}{S}$$

In the case of a normal population, Z was just a standardised version of a normal r.v. $\bar{X}$ and was therefore $\mathcal{N}(0,1)$-distributed

However, T is now a ratio of two random variables ($\bar{X}-\mu$ and S)

→ T is not $\mathcal{N}(0,1)$-distributed !

Indeed, T cannot have exactly the same distribution as Z, as the approximation of the constant σ by a random variable S introduces some extra variability.

→ the random variable T varies more in value from sample to sample than Z (i.e. $\mathbb{V}\mathrm{ar}(T) > \mathbb{V}\mathrm{ar}(Z)$)

The Student's *t*-distribution

The first who realised that was William Gosset (1876-1937), a British chemist and mathematician who, in the early 20th century, worked at the Guinness Brewery in Dublin.

Another researcher at Guinness had previously published a paper containing trade secrets of the Guinness brewery, so that Guinness prohibited its employees from publishing any scientific papers regardless of the contained information

→ Gosset negotiated permission to publish, but without a Guinness affiliation, and using the pseudonym *Student*

He showed that, in a normal population, the exact distribution of T is the so-called t-distribution with $n-1$ degrees of freedom:

$$T \sim t_{n-1}$$

This distribution is now referred to as **Student's *t*-distribution** (which might otherwise have been Gosset's t-distribution).

The Student's *t*-distribution

A random variable, say T, is said to follow the Student's t-distribution with ν degrees of freedom, i.e.

$$\boxed{T \sim t_\nu}$$

if its probability density function is given by

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \qquad \to S_T = \mathbb{R}$$

for some integer ν.

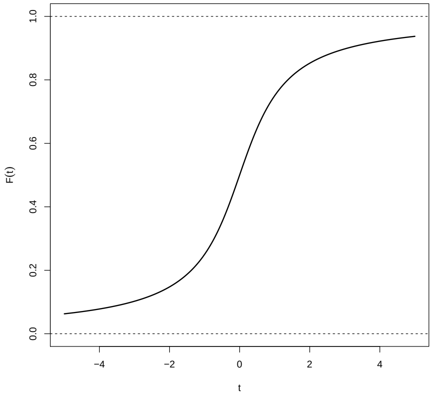
Note: the Gamma function is given by

$$\Gamma(y) = \int_0^{+\infty} x^{y-1}e^{-x}\,dx, \qquad \text{for } y > 0$$

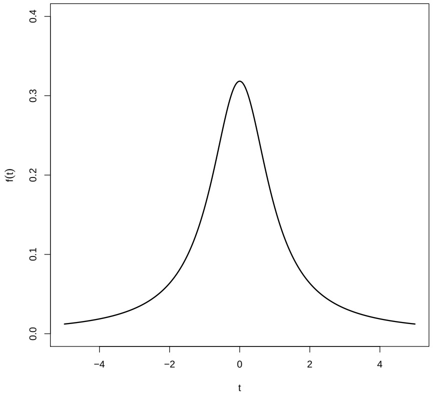
There is no simple expression for the Student's t-cdf

The Student's *t*-distribution

Student's t distribution with 1 degree of freedom

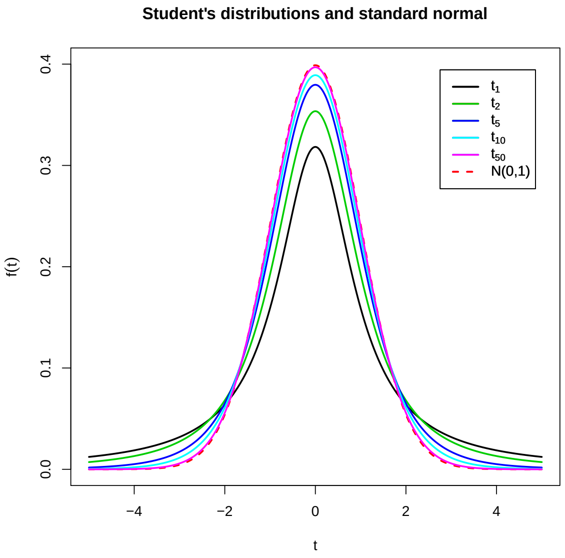


cdf $F(t)$



pdf $f(t) = F'(t)$

The Student's *t*-distribution



The Student's *t*-distribution

It can be shown that the mean and the variance of the t_ν-distribution are

$$\mathbb{E}(T) = 0 \qquad \text{and} \qquad \mathbb{V}\text{ar}(T) = \frac{\nu}{\nu - 2} \qquad (\text{for } \nu > 2)$$

The Student's *t* distribution is similar in shape to the standard normal distribution in that both densities are symmetric, unimodal and bell-shaped, and the maximum value is reached at 0.

However, the Student's *t* distribution has heavier tails than the normal

→ there is more probability to find the random variable T 'far away' from 0 than there is for Z

This is more marked for small values of ν

As the number ν of degrees of freedom increases, t_ν-distributions look more and more like the standard normal distribution

In fact, it can be shown that the Student's *t* distribution with ν degrees of freedom approaches the standard normal distribution as $\nu \to \infty$

The Student's *t*-distribution: quantiles

Similarly to what we did for the Normal distribution, we can define the quantiles of any Student's *t*-distribution:

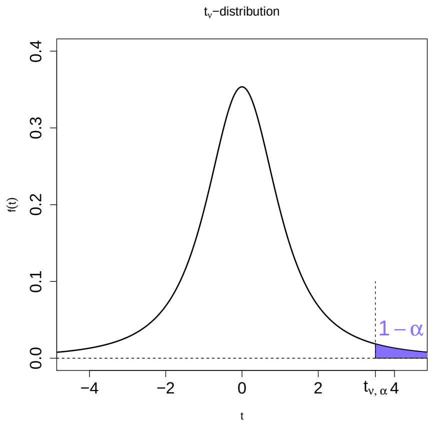
Let $t_{\nu;\alpha}$ be the value such that

$$\mathbb{P}(T > t_{\nu;\alpha}) = 1 - \alpha$$

for $T \sim t_\nu$

Like the standard normal distribution, the symmetry of any t_ν-distribution implies that

$$\boxed{t_{\nu;1-\alpha} = -t_{\nu;\alpha}}$$



Confidence interval on the mean of a normal distribution, variance unknown

So we have, for any $n \geq 2$,

$$T = \sqrt{n}\,\frac{\bar{X} - \mu}{S} \sim t_{n-1}$$

Note: the number of degrees of freedom for the *t*-distribution is the number of degrees of freedom associated with the estimated variance S^2 (recall Slide 70)

It is now easy to find a $100 \times (1 - \alpha)\%$ confidence interval for μ by proceeding essentially as we did when σ^2 was known (Slide 273)

We may write

$$\mathbb{P}\left(-t_{n-1;1-\alpha/2} \leq \sqrt{n}\,\frac{\bar{X} - \mu}{S} \leq t_{n-1;1-\alpha/2}\right) = 1 - \alpha$$

or

$$\mathbb{P}\left(\bar{X} - t_{n-1;1-\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1;1-\alpha/2}\frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

t-confidence interval on the mean of a normal distribution

→ if $\bar{x}$ and s are the sample mean and sample standard deviation of an observed random sample of size n from a normal distribution, a confidence interval of level $100 \times (1-\alpha)\%$ for μ is given by

$$\boxed{\left[\bar{x} - t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}\right]}$$

This confidence interval is sometimes called t-confidence interval, as opposed to $\left[\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$ (z-confidence interval)

Because t_{n-1} has heavier tails than $\mathcal{N}(0,1)$, $t_{n-1;1-\alpha/2} > z_{1-\alpha/2}, \forall n$

→ this reflects the extra variability introduced by the estimation of σ (less accuracy, wider interval)

Note: One can also define one-sided $100 \times (1-\alpha)\%$ t-confidence intervals $\left(-\infty, \bar{x} + t_{n-1;1-\alpha}\frac{s}{\sqrt{n}}\right]$ and $\left[\bar{x} - t_{n-1;1-\alpha}\frac{s}{\sqrt{n}}, +\infty\right)$

t-confidence interval: example

Example

An article in *Materials Engineering* describes the results of tensile adhesion test on 22 $U-700$ alloy specimens. The load at specimen failure is as follows (in megapascals):

```
7.6, 8.1, 11.7, 14.3, 14.3, 14.1, 8.3, 12.3, 15.9, 16.4,
  11.3, 12.0, 12.9, 15.0, 13.2, 14.6, 13.5, 10.4, 13.8,
                       15.6, 12.2, 11.2
```

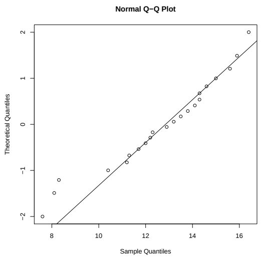
Construct a 99% confidence interval for the true average load at failure for this type of alloy. (**Matlab output:** `tinv`$(0.995, 21) = 2.831$)

Elementary computations give

$$\bar{x} = 12.67 \text{ MPa} \qquad \text{and} \qquad s = 2.47 \text{ MPa}$$

t-confidence interval: example

The quantile plot provides good support for the assumption that the population is normally distributed



Since $n = 22$, we have $n - 1 = 21$ degrees of freedom for t. According to Matlab, $t_{21;0.995} = 2.831$. The resulting CI is

$$\left[\bar{x} - t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}\right] = \left[12.67 \pm 2.831 \times \frac{2.47}{\sqrt{22}}\right]$$
$$= [11.18, 14.16]$$

→ we are 99% confident that the true average load at failure for this type of alloy lies between 11.18 MPa and 14.16 MPa

Confidence interval on the mean of an arbitrary distribution, variance unknown

What if the population is not normal ?

As in the case 'σ^2 known', we can rely on the Central Limit Theorem which asserts that, for n 'large', $Z = \sqrt{n}\frac{\bar{X}-\mu}{\sigma} \overset{a}{\sim} \mathcal{N}(0,1)$ to deduce a result like

$$T = \sqrt{n}\frac{\bar{X}-\mu}{S} \overset{a}{\sim} t_{n-1}$$

from which we could find a CI on μ for n large enough.

However, recall that, when ν is large, t_ν is very much like $\mathcal{N}(0,1)$

→ in large samples, estimating σ with S has very little effect on the distribution of T, to which the approximation by the standard normal distribution is more than enough:

$$\boxed{T \overset{a}{\sim} \mathcal{N}(0,1)}$$

Confidence interval on the mean of an arbitrary distribution

Consequently, if $\bar{x}$ and s are the sample mean and standard deviation of an observed random sample of large size n from any distribution, an approximate confidence interval of level $100 \times (1-\alpha)\%$ for μ is

$$\boxed{\left[\bar{x} - z_{1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{s}{\sqrt{n}}\right]}$$

This expression holds regardless of the population distribution, as long as n is large enough → it is called a large-sample confidence interval.

Generally, n should be at least 40 to use this result reliably (the CLT usually holds for $n \geq 30$, but a larger sample size is recommended because replacing σ by S still results in some additional variability).

As usual, corresponding one-sided confidence intervals could be defined: $(-\infty, \bar{x} + z_{1-\alpha}\frac{s}{\sqrt{n}}]$ and $[\bar{x} - z_{1-\alpha}\frac{s}{\sqrt{n}}, +\infty)$

Confidence interval on the mean: example

Example

An article in *Transactions of the American Fisheries Society* reports the results of a study to investigate the mercury contamination in largemouth bass. A sample of 53 fishes was selected from some Florida lakes, and mercury concentration in the muscle tissue was measured (in ppm):

```
1.23, 0.49, 1.08, ..., 0.16, 0.27
```

Find a confidence interval on μ, the mean mercury concentration in the muscle tissue of fish.

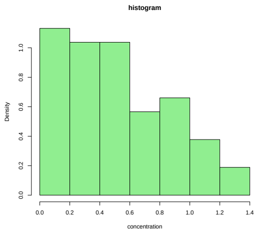
(**Matlab output:** $\texttt{norminv}(0.975) = 1.96$, $\texttt{tinv}(0.975, 52) = 2.007$)

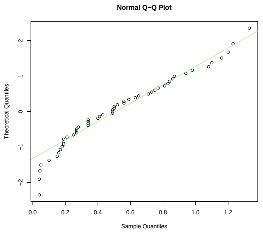
An histogram and a quantile plot for the data are displayed below

→ both plots indicate that the distribution of mercury concentration may not be normally distributed (positively skewed)

But anyway, the sample is large enough ($n = 53$) to use the Central Limit Theorem and compute an approximate confidence interval for μ.

Confidence interval on the mean: example





Elementary computations give $\bar{x} = 0.525$ ppm and $s = 0.3486$ ppm. A large sample confidence interval is given by $\left[\bar{x} - z_{1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{s}{\sqrt{n}}\right]$

With $z_{0.975} = 1.96$ and the above values, we have

$$\left[0.525 - 1.96\frac{0.3486}{\sqrt{53}}, 0.525 + 1.96\frac{0.3486}{\sqrt{53}}\right] = [0.4311, 0.6189]$$

→ we are 95% confident that the true average mercury concentration in the muscle tissue of the fishes is between 0.4311 and 0.6189 ppm

Confidence intervals on the mean: example

For large sample sizes, what if the population is not normal and you still use the t-confidence interval? In the previous example, for

$t_{52;0.975} = 2.007$ we compute

$$\left[0.525 - 2.007\frac{0.3486}{\sqrt{53}}, 0.525 + 2.007\frac{0.3486}{\sqrt{53}}\right] = [0.4289, 0.6211]$$

This is very similar to the previous result [0.4311,0.6189].

It turns out $t_{n-1;1-\alpha/2} \to z_{1-\alpha/2}$ as $n \to \infty$ and

$$\left[\bar{x} - z_{1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{s}{\sqrt{n}}\right] \subseteq \left[\bar{x} - t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}\right]$$

The t-confidence interval is similar to the large sample one and is therefore acceptable for large sample sizes (say $n \geq 40$), (however it is longer and requires more effort)

(Keep in mind that both intervals are approximated intervals, anyway)

Confidence interval on the mean: example

Example

The article "Extravisual Damage Detection? Defining the Standard Normal Tree" (*Photogrammetric Engr. and Remote Sensing, 1981: 515-522*) discusses the use of color infrared photography in identification of normal trees in Douglas fir stands. Among data reported were summary statistics for green-filter analytic optical densitometric measurements on samples of both healthy and diseased trees. For a sample of 69 healthy trees, the sample mean dye-layer density was 1.028, and the sample standard deviation was 0.163. Assume the dye-layer density follows a normal distribution. a) Calculate a 95% two-sided confidence interval for the true average dye-layer density for all such trees. (**Matlab output:** $\texttt{norminv}(0.975) = 1.96$, $\texttt{tinv}(0.975, 68) = 1.9955$)

A 95% two-sided exact t-confidence interval for the true average dye-layer density for all such tree is:

$$\left[\bar{x} - t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} - t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}\right]$$

Confidence interval on the mean: example

Example (ctd.)

a) Calculate a 95% two-sided confidence interval for the true average dye-layer density for all such trees. (**Matlab output:** $\texttt{norminv}(0.975) = 1.96$, $\texttt{tinv}(0.975, 68) = 1.9955$)

With $t_{68;0.975} = 1.9955$, $\bar{x} = 1.028$ and $s = 0.163$, the resulting CI is:

$$\left[1.028 - 1.9955\frac{0.163}{\sqrt{69}}, 1.028 + 1.9955\frac{0.163}{\sqrt{69}}\right] = [0.9888, 1.0672]$$

Since the sample size is large ($n = 69$), the Central Limit Theorem can be applied and a 95% two-sided approximate z-confidence interval for the true average dye-layer density for all such tree can be computed:

$$\left[\bar{x} - z_{1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} - z_{1-\alpha/2}\frac{s}{\sqrt{n}}\right]$$

Since $z_{0.975} = 1.96$, we have:

$$\left[1.028 - 1.96\frac{0.163}{\sqrt{69}}, 1.028 + 1.96\frac{0.163}{\sqrt{69}}\right] = [0.9895, 1.0665]$$

Confidence interval on the mean: example

Example (ctd.)

b) Suppose the investigators had made a rough guess of 0.16 for the value of σ before collecting data. What sample size would be necessary to obtain an interval width of 0.05 for a confidence level of 95%?

Using the formula on Slide 276, the required sample size should be

$$n = \left(\frac{z_{1-\alpha/2}\sigma}{e}\right)^2$$

The error e is half the interval width, $e = 0.05/2 = 0.025$. Then the sample size

$$n = \left(\frac{1.96 \times 0.16}{0.025}\right)^2 \approx 157.35$$

So, a sample size of 158 trees would be required.

Confidence intervals for the mean: summary

The several situations leading to different confidence intervals for the mean can be summarised as follows:

The first question is: **Is the population normal?** (check from a histogram and a quantile plot, for instance)

- if yes, is σ known ?
  - if yes, use an exact z-confidence interval:
    $$\left[\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$
  - if no, use an exact t-confidence interval:
    $$\left[\bar{x} - t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{n-1;1-\alpha/2}\frac{s}{\sqrt{n}}\right]$$
- if no, use an approximate large sample confidence interval:
  $$\left[\bar{x} - z_{1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{s}{\sqrt{n}}\right],$$
  (provided the sample size is large, say $n \geq 40$)

What if the sample size is small and the population is not normal ?
→ check on a case by case basis (beyond the scope of this course)

Inferences concerning proportions

Many engineering problems deal with proportions, percentages or probabilities:

we are concerned with the proportion of defectives in a lot, with the percentage of certain components which will perform satisfactorily during a stated period of time, or with the probability that a newly produced item meets some quality standards

→ qualitative information can also be included in statistical studies!

It should be clear that problems concerning proportions, percentages or probabilities are really equivalent: a percentage is merely a proportion multiplied by 100, and a probability is a proportion in a (infinitely) long series of trials.

We would like to learn about π, the **proportion of the population that has a characteristic of interest**, but as usual all we have is just a sample of size n from that population

→ inference about π → confidence interval for π

Estimation of a proportion

In this situation, the random variable to study is

$$X = \begin{cases} 1 & \text{if the individual has the characteristic of interest} \\ 0 & \text{if not} \end{cases}$$

which is Bernoulli distributed, with parameter being the value π of interest:

$$X \sim \text{Bern}(\pi)$$

The random sample $X_1, X_2, \ldots, X_n$ is a set of n independent Bern(π) random variables.

→ the number Y of individuals of the sample with the characteristic is

$$Y = \sum_{i=1}^n X_i \sim \text{Bin}(n, \pi)$$

and the **sample proportion** is

$$\widehat{P} = \frac{Y}{n}$$

Estimation of a proportion

This sample proportion $\widehat{P}$ is obviously a natural candidate for estimating the population proportion π.

From the properties of the Binomial distribution, we know that

$$\mathbb{E}(Y) = n\pi \qquad \text{and} \qquad \mathbb{V}\text{ar}(Y) = n\pi(1-\pi)$$

so that $\mathbb{E}(\widehat{P}) = \frac{1}{n}\mathbb{E}(Y) = \pi$ and $\mathbb{V}\text{ar}(\widehat{P}) = \frac{1}{n^2}\mathbb{V}\text{ar}(Y) = \frac{n\pi(1-\pi)}{n^2} = \frac{\pi(1-\pi)}{n}$

Hence, $\widehat{P}$ is an **unbiased** and **consistent estimator** for π:

$$\mathbb{E}(\widehat{P}) = \pi \qquad \text{and} \qquad \mathbb{V}\text{ar}(\widehat{P}) = \frac{\pi(1-\pi)}{n} \quad (\to 0 \text{ as } n \to \infty)$$

→ the standard error of $\widehat{P}$ is thus sd$(\widehat{P}) = \sqrt{\frac{\pi(1-\pi)}{n}}$

Upon observation of a random sample $x_1, x_2, \ldots, x_n$, in which $y = \sum_{i=1}^n x_i$ individuals have the characteristics, an estimate of π is

$$\hat{p} = \frac{y}{n}$$

Sampling distribution

We could make inference about π from $\hat{p}$ using the Binomial distribution of Y. However, it is probably easier to use the **Central Limit Theorem**. Indeed:

$$\widehat{P} = \frac{Y}{n} = \frac{1}{n}\sum_{i=1}^n X_i,$$

so that $\widehat{P}$ is actually a (particular) sample mean, for which the **CLT** guarantees that

$$\boxed{\sqrt{n}\,\frac{\widehat{P} - \pi}{\sqrt{\pi(1-\pi)}} \overset{a}{\sim} \mathcal{N}(0,1)}$$

if n is 'large' (see Slide 285)

We also know that the quality of the approximation depends on the symmetry of the initial distribution of the X_i's, here Bern(π)

→ π should not be too close to 0 or 1 → empirical rule: $n\hat{p}(1-\hat{p}) > 5$

Confidence interval for a proportion

As the sampling distribution

$$\sqrt{n}\,\frac{\widehat{P}-\pi}{\sqrt{\pi(1-\pi)}} \overset{a}{\sim} \mathcal{N}(0,1)$$

is just a particular case of $\sqrt{n}\,\frac{\bar{X}-\mu}{\sigma} \overset{a}{\sim} \mathcal{N}(0,1)$, we can use (almost) directly the large-sample confidence interval we derived for a mean

Specifically, we have that

$$\mathbb{P}\left(-z_{1-\alpha/2} \leq \sqrt{n}\,\frac{\widehat{P}-\pi}{\sqrt{\pi(1-\pi)}} \leq z_{1-\alpha/2}\right) \simeq 1-\alpha$$

or

$$\mathbb{P}\left(\widehat{P} - z_{1-\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}} \leq \pi \leq \widehat{P} + z_{1-\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}\right) \simeq 1-\alpha$$

→ a confidence interval for π takes shape

Confidence interval for a proportion

Unfortunately, the standard error of $\widehat{P}$, that is the factor $\sqrt{\frac{\pi(1-\pi)}{n}}$, contains the unknown π.

In such a situation, we may replace the unknown value by its estimate, that is, to use the estimated standard error of the estimator

$$\widehat{\text{sd}(\widehat{P})} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

in the expression of the confidence interval.

Consequently, if $\hat{p}$ is the sample proportion in an observed random sample of size n, an approximate two-sided confidence interval of level $100 \times (1-\alpha)\%$ for π is given by

$$\boxed{\left[\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]}$$

As this is based on the CLT and requires n 'large', it is a large sample confidence interval for π.

One-sided confidence intervals for a proportion

We may also find one-sided large-sample confidence intervals for the proportion π by a simple modification of the previous development

We find:

$$\boxed{\left[0, \hat{p} + z_{1-\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]}$$

and

$$\boxed{\left[\hat{p} - z_{1-\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, 1\right]}$$

Sample size

Since $\hat{p}$ is the estimate of π, we can define the error in estimating π by $\hat{p}$ as $e = |\hat{p} - \pi|$. From Slide 320, we are approximately $100 \times (1-\alpha)\%$ confident that this error is less than

$$z_{1-\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}$$

In situations where the sample size can be selected, we may choose n to be $100 \times (1-\alpha)\%$ confident that the error is less than any specified value e:

$$n = \left(\frac{z_{1-\alpha/2}}{e}\right)^2 \pi(1-\pi) \qquad \text{(compare Slide 276)}$$

→ this depends on π, for which no information is available at this point

Idea: use an upper bound which holds for any value of π

Actually, $\pi(1-\pi) \leq 1/4$, with equality for $\pi = 1/2$, thus with

$$\boxed{n = \left(\frac{z_{1-\alpha/2}}{2e}\right)^2}$$

we are at least $100 \times (1-\alpha)\%$ confident that this error is less than e and this, regardless of the value of π (this is very conservative, though).

Confidence interval for a proportion: example

Example

In a random sample of 85 car engine crankshaft bearings, 10 have a surface finish that is rougher than the specifications allow. a) Find a 95% confidence interval on the true proportion π of produced bearings that exceeds the roughness specification.

a) The estimate of π is $\hat{p} = \frac{y}{n} = \frac{10}{85} = 0.118$. Thus, the estimated standard error is

$$\widehat{\text{sd}(\hat{P})} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.118 \times (1-0.118)}{85}} = 0.035$$

and an approximated two-sided 95% confidence interval for π is

$$\left[\hat{p} \pm z_{0.975}\widehat{\text{sd}(\hat{P})}\right] = [0.118 \pm 1.96 \times 0.035] = [0.049, 0.186]$$

→ we are 95% confident that the true proportion π of produced bearings outside specifications is between 0.049 and 0.186

Confidence interval for a proportion: example

Example (ctd.)

In a random sample of 85 car engine crankshaft bearings, 10 have a surface finish that is rougher than the specifications allow. b) How large is a sample required if we want to be 95% confident that the error in estimating π is less than 0.05?

b) the previous CI has width 0.137, which is quite large for a CI for π. If we want a CI of width at most 2 × 0.05, we need

$$n = \left(\frac{z_{1-\alpha/2}}{2e}\right)^2 = \left(\frac{1.96}{2 \times 0.05}\right)^2 = 384.16$$

→ we need at least 385 observations

Note that this number would guarantee the required accuracy, regardless of the true value of π → this is why it is so high (conservative)

(Using $\hat{p} = 0.118$ as preliminary estimate of π, we would have

$$n \simeq (z_{1-\alpha/2}/e)^2\,\hat{p}(1-\hat{p}) = (1.96/0.05)^2 \times 0.118 \times 0.882 = 159.93)$$

Confidence interval for a proportion: example

Example

The article “Repeatability and Reproducibility for Pass/Fail Data" (*J. of Testing and Eval., 1997: 151-153*) reported that in $n = 48$ trials in a particular laboratory, 16 resulted in ignition of a particular type of substrate by a lighted cigarette. Let π denote the long-run proportion of all such trials that would result in ignition. Find a 95% confidence interval on the true proportion π.

The estimate of π is $\hat{p} = \frac{y}{n} = \frac{16}{48} = 0.333$. Thus, an approximated two-sided 95% confidence interval for π is

$$\begin{aligned}\left[\hat{p} \pm z_{0.975}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] &= \left[0.333 \pm 1.96\sqrt{\frac{0.333(1-0.333)}{48}}\right] \\ &= [0.333 \pm 0.133] \\ &= [0.200, 0.466]\end{aligned}$$

→ we are 95% confident that the true long-run proportion π of all such trials that would result in ignition is between 0.200 and 0.466

Objectives

Now you should be able to:

- Construct z- and t-confidence intervals on the mean of a normal distribution, advisedly using either the normal distribution or the Student’s t distribution □
- Construct large sample confidence intervals on a mean of an arbitrary distribution with unknown variance □
- Explain the difference between a confidence interval and a prediction interval □
- Construct prediction intervals for a future observation in a normal population □
- Construct confidence intervals on a population proportion □

Recommended exercises:

→ Q7, Q9, p.301, Q13, Q15 p.302, Q20 p.303, Q35 p.319, Q39 p.320, Q43(a-b) p.320, Q55 p.328, (optional) Q71, Q73 p.340, Q55 p.238 (2nd edition)

→ Q7, Q9, p.305, Q16 p.307, Q21 p.307, Q37 p.324, Q42 p.325, Q46(a-b) p.326, Q58 p.334, (optional) Q75, Q77 p.347, Q57 p.242 (3rd edition)

Hypotheses testing: Introduction

In the previous lectures we showed how a parameter of a population can be estimated from sample data, using either a point estimate or an interval of “plausible” values called a confidence interval.

However, there are many situations in which we must decide whether we believe a statement concerning a parameter is true or false, that is, we must **test a hypothesis about a parameter**.

For instance, suppose that a customer protection agency wants to test a paint manufacturer’s claim that the average drying time of his new ‘fast-drying’ paint is 20 minutes.

It instructs a member of its staff to paint each of 36 boards using a different can of the paint: the observed average drying time for this sample is 20.75 minutes

→ does that really contradict the manufacturer’s claim ?

This type of question can be answered using a statistical inference technique called **hypothesis testing**.

Statistical hypotheses

Many problems in engineering require that we decide which of two competing statements about some parameters are true

→ the statements are called hypotheses

In the previous example, we might express the hypothesis as

$$H_0 : \mu = 20$$

where μ is the ‘true’ mean drying time for this type of paint

This statement is called the **null hypothesis**, usually denoted H_0

H_0 is the default hypothesis we will assume is true unless we have enough evidence to compel us to change our minds

If so, we will favour the **alternative hypothesis**, usually denoted H_a

Typically, H_a depends on the problem and what we hope to show

In our example, it could be either $H_a : \mu \neq 20$ or $H_a : \mu > 20$

Null hypothesis

The value μ_0 of the population parameter specified in the null hypothesis

$$H_0 : \mu = \mu_0$$

is usually determined in one of three ways:

- it may result from past experience or knowledge of the process, or from previous tests or experiments

→ determine whether the parameter value has changed

- it may be determined from some theory or some model

→ check whether the theory or the model is valid

- it may result from external considerations, such as engineering specifications, or from contractual obligations

→ conformance testing

Note: in some instances, a null hypothesis of the form $H_0 : \mu \geq \mu_0$ or $H_0 : \mu \leq \mu_0$ may seem appropriate. However, the test procedure for such an H_0 is the same as $H_0 : \mu = \mu_0$

→ we always state a null hypothesis as an equality

Alternative hypothesis

The alternative hypothesis can essentially be of two types.

A **two-sided alternative** is when H_a is of the form

$$H_a : \mu \neq \mu_0$$

This is the exact denial of the null hypothesis $H_0 : \mu = \mu_0$
→ μ_0 is the only value of some interest in the problem

However, in many situations, we may wish to favour a given direction for the alternative:

$$H_a : \mu < \mu_0 \qquad \text{or} \qquad H_a : \mu > \mu_0$$

These are called **one-sided alternatives**.

Continuing with our example, the customer protection agency may only wish to highlight that the average drying time of the paint is actually longer than the advertised 20 minutes (no criticisms if this time is even shorter)

Careful! The considered alternative might change the conclusion of a hypothesis test, and should be carefully formulated!

Hypothesis testing

A procedure leading to a decision about a particular hypothesis H_0 is called a test of hypothesis.

Such procedures rely on using the information contained in a random sample from the population of interest

If this information is consistent with the hypothesis H_0, we will not reject H_0; however, any information inconsistent with H_0 cast doubt on it, and we will reject it

Fact

The truth or the falsity of a particular hypothesis can never be known with certainty, unless we can examine the entire population.

The decision we make depends on a random sample, so is a kind of 'random object'

→ a hypothesis test should be developed with the probability of reaching a wrong conclusion in mind

Hypothesis testing

To illustrate the general concepts, consider again the mean drying time problem. Imagine that we wish to test

$$H_0 : \mu = 20 \qquad \text{against} \qquad H_a : \mu \neq 20$$

We have a sample of $n = 36$ specimens and the sample mean $\bar{x}$ is observed

As the sample mean is a 'good' estimate of μ, we expect $\bar{x}$ to be reasonably close to μ

→ if $\bar{x}$ falls 'close' to 20 min, no clear contradiction with H_0, we do not reject it

→ if $\bar{x}$ is considerably 'distant' from 20 min, evidence in support of H_a, we reject H_0

The numerical value which is computed from the sample and used to decide between H_0 and H_a, here the (possibly standardised) sample mean, is called the

test statistic

Hypothesis testing

Suppose that we decide to reject H_0 if $\bar{x}$ is smaller than 19.33 or larger than 20.67 (arbitrary criterion for illustrative purposes only)

→ if $\bar{x} \in [19.33, 20.67]$, we do not reject $H_0 : \mu = 20$

The values for which we reject H_0, that is, values less than 19.33 and greater than 20.67, are called the rejection regions for the test, the limiting values (here 19.33 and 20.67) being the **critical values**.

This provides a clear-cut criterion for the decision; however, it is not infallible:

a) even if the true mean $\mu = 20$, there is a possibility that the sample mean $\bar{x}$ may be outside $[19.33, 20.67]$, due to bad luck

b) even if the true mean $\mu \neq 20$, say $\mu = 21$, there is a possibility that the sample mean $\bar{x}$ may be in $[19.33, 20.67]$

Errors

So there are essentially two possible wrong conclusions:

a) rejecting H_0 when it is true: this is defined as a **type I error**

b) failing to reject H_0 when it is false: this is defined as **type II error**

Because the decision is based on a random sample, **probabilities** can be associated with these errors.

The probability of type I error is usually denoted α

$$\mathbb{P}(\text{type I error}) = \mathbb{P}(\text{reject } H_0 \text{ when it is true}) = \alpha$$

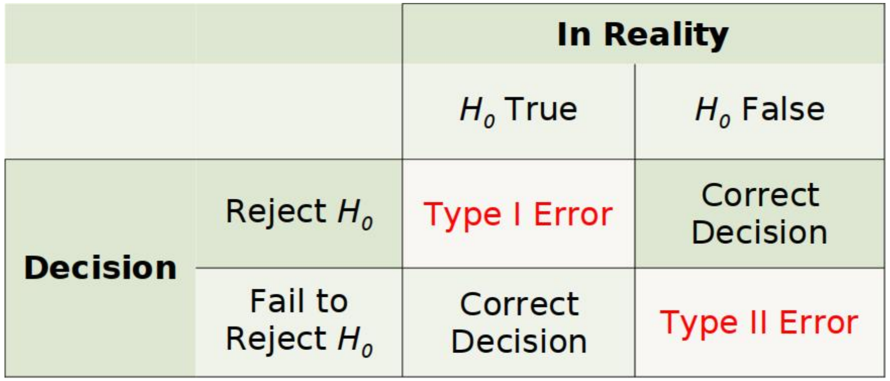
The probability of type II error is usually denoted β

$$\mathbb{P}(\text{type II error}) = \mathbb{P}(\text{fail to reject } H_0 \text{ when it is false}) = \beta$$

$1 - \beta = \mathbb{P}(\text{reject } H_0 \text{ when it is false})$ is also called the **power** of the test

Note that β actually depends on the true (unknown) value of μ.

Errors



		In Reality	
		H_0 True	H_0 False
Decision	Reject H_0	Type I Error	Correct Decision
	Fail to Reject H_0	Correct Decision	Type II Error

Quantifying Errors

Assume in our running example that it is known from past experience that the drying time is normally distributed with known standard deviation $\sigma = 2$ min.

Then we know that $Z = \sqrt{n}\frac{\bar{X}-\mu}{\sigma} \sim \mathcal{N}(0,1)$ (Slide 250) , so:

$$\begin{aligned}\mathbb{P}(\text{type I error}) &= \mathbb{P}((\bar{X} < 19.33) \cup (\bar{X} > 20.67) \text{ when } \mu = 20)\\ &= \mathbb{P}\left(Z < \sqrt{36}\,\frac{19.33-20}{2}\right) + \mathbb{P}\left(Z > \sqrt{36}\,\frac{20.67-20}{2}\right)\\ &= \mathbb{P}(Z < -2.01) + \mathbb{P}(Z > 2.01) = 0.044 = \alpha \qquad \text{(Matlab)}\end{aligned}$$

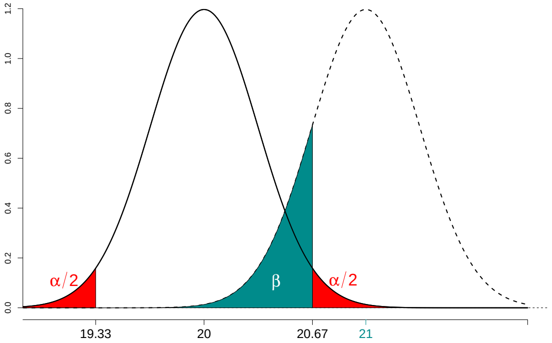
If H_0 is true, our rule has a 4.4% chance of rejecting it

Suppose now that $\mu = 21$ (so H_0 is not true!). We have:

$$\begin{aligned}\mathbb{P}(\text{type II error}) &= \mathbb{P}(\bar{X} \in [19.33, 20.67] \text{ when } \mu = 21)\\ &= \mathbb{P}\left(\sqrt{36}\,\frac{19.33-21}{2} \le Z \le \sqrt{36}\,\frac{20.67-21}{2}\right)\\ &= \mathbb{P}(-5.01 \le Z \le -0.99) = 0.16 = \beta \qquad \text{(Matlab)}\end{aligned}$$

Errors

With the decision rule: reject H_0 if $\bar{x} \notin [19.33, 20.67]$



$\rightarrow \alpha = 0.044,$
$\beta = 0.16$ if $\mu = 21$

See that β would rapidly increase as μ approached the hypothesised value μ_0

Errors

Suppose that you want to reduce the type I error probability α

→ widen the acceptance region, for instance say

$$\text{reject } H_0 \text{ if } \bar{x} \notin [19.2, 20.8]$$

(again for illustrative purpose only). Then (as on Slide 338),

$$\alpha = \ldots = 0.016 \quad (< 0.044)$$

but

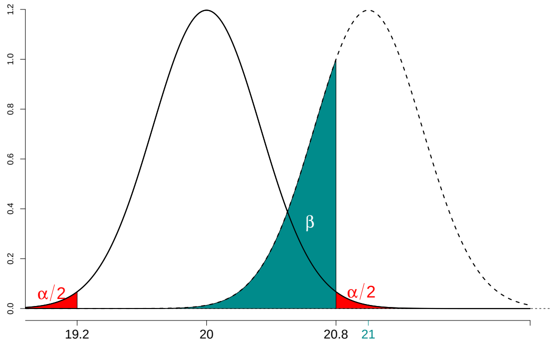
$$\beta = \ldots = 0.27 \quad (> 0.16) \qquad \text{when } \mu = 21$$

→ if α decreases, β must increase and vice-versa !

→ impossible to make both types of error as small as possible simultaneously

Errors

With the decision rule: reject H_0 if $\bar{x} \notin [19.2, 20.8]$



→ $\alpha = 0.016$,
$\beta = 0.27$ if $\mu = 21$

Errors

Usually, one decides to set α to a small predetermined level (and accept the resulting value of β).

This is because hypothesis testing was originally inspired by jury trials.

In a trial, defendants are initially **assumed innocent** (H_0). Then,

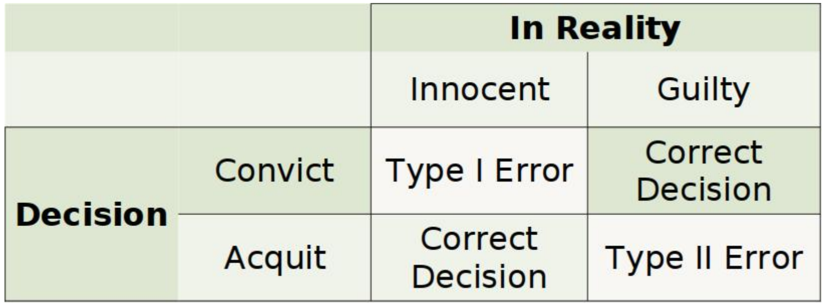
- if strong evidence is found to the contrary, then they are declared to be guilty (reject H_0)
- if there is insufficient evidence, they are declared not guilty (fail to reject H_0) → not the same as proving the defendant is innocent!

If the jury is wrong, either an innocent person is convicted (type I error) or a culprit is let free (type II error)

→ The prevailing thought is that convicting an innocent person is more a serious problem than letting a culprit free

⇝ controlling α is more important

Errors: analogy to criminal trials



		In Reality	
		Innocent	Guilty
Decision	Convict	Type I Error	Correct Decision
	Acquit	Correct Decision	Type II Error

Jury must be "convinced beyond a reasonable doubt" to convict

→ usually, the type I error probability α is set to 0.10, 0.05 or 0.01, and the decision rule is fixed accordingly

In hypothesis testing, the value of α is called the **significance level** of the test.

Significance level and decision rule

Assume for the moment that **the population is normal with known standard deviation** σ $\quad \rightarrow \bar{X} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$

At significance level α, we are after two constants ℓ and u such that

$$\alpha = \mathbb{P}(\bar{X} \notin [\ell, u] \text{ when } \mu = \mu_0) = \mathbb{P}\left(Z \notin \left[\sqrt{n}\frac{\ell - \mu_0}{\sigma}, \sqrt{n}\frac{u - \mu_0}{\sigma}\right]\right)$$

$$\rightarrow \sqrt{n}\frac{\ell - \mu_0}{\sigma} = z_{\alpha/2} = -z_{1-\alpha/2} \qquad \text{and} \qquad \sqrt{n}\frac{u - \mu_0}{\sigma} = z_{1-\alpha/2}$$

$$\rightarrow \boxed{\ell = \mu_0 - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}} \qquad \text{and} \qquad \boxed{u = \mu_0 + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}}$$

→ the decision rule is:

$$\text{reject } H_0 \text{ if } \bar{x} \notin \left[\mu_0 - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \mu_0 + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

Significance level and decision rule: example

In our example (with $\sigma = 2$), suppose we want to test $H_0 : \mu = 20$ against $H_a : \mu \neq 20$ at the 5% significance level ($\alpha = 0.05$).

Then, $\ell = 20 - 1.96 \times \frac{2}{\sqrt{36}} = 19.35$ and $u = 20 + 1.96 \times \frac{2}{\sqrt{36}} = 20.65$

→ At significance level 5%, the decision rule is

reject H_0 if $\bar{x} \notin [19.35, 20.65]$

Now, over our 36 samples, we have observed an average drying time of $\bar{x}$ =20.75 min → we reject H_0

→ we contradict the manufacturer

Now, at significance level 1% ($\alpha = 0.01$),

$$\ell = 20 - 2.575 \times \frac{2}{\sqrt{36}} = 19.14 \text{ and } u = 20 + 2.575 \times \frac{2}{\sqrt{36}} = 20.86$$

and the decision rule becomes reject H_0 if $\bar{x} \notin [19.14, 20.86]$

→ now we cannot reject H_0 from the observed sample with $\bar{x} = 20.75$

→ you do not dare contradict the manufacturer

Reject / no reject of H_0: remark

The previous situation makes it clear why we do not say “we accept H_0”:

we reject H_0 at 5% level but we don’t at 1% level. However H_0 is either true or not, regardless of the situation!

Testing at the 5% level essentially means that we tolerate being wrong in at most 5% of the cases when rejecting H_0 → here we reject H_0

Testing at the 1% level means that we tolerate being wrong in at most 1% of the cases when rejecting H_0

→ it is then too risky to reject H_0 if we need to be 99% confident in our decision

→ we do not reject H_0, the sample has not shown *enough* evidence against it. We don’t know whether H_0 is true but we cannot exclude the possibility it is (which doesn’t mean it is for sure!)

p-value

Here the evidence shown by the sample indicates that we would be wrong with a chance between 1% and 5% if we rejected H_0

→ it would be interesting to know more about how confident we can be in our decision

That is, what is the ***p*-value**.

Definition

The ***p*-value** is the smallest level of significance that would lead to rejection of H_0 with the observed sample

Concretely, the *p*-value is the probability that the test statistic will take on a value that is at least as extreme as the observed value when H_0 is true (‘extreme’ to be understood in the direction of the alternative).

It might be *roughly* interpreted as the chance of being wrong if we reject H_0

p-value

When testing $H_0 : \mu = \mu_0$ against $H_a : \mu \neq \mu_0$, the p-value will be the probability of finding the random variable $\bar{X}$ more different to μ_0 than the observed $\bar{x}$, that is,

$$\begin{aligned} p &= \mathbb{P}\left(\bar{X} \notin [\mu_0 \pm |\bar{x} - \mu_0|] \text{ when } \mu = \mu_0\right) \\ &= 1 - \mathbb{P}\left(\bar{X} \in [\mu_0 \pm |\bar{x} - \mu_0|] \text{ when } \mu = \mu_0\right) \end{aligned}$$

Define z_0 as the z-score of $\bar{x}$ if $\mu = \mu_0$, i.e.

$$z_0 \doteq \sqrt{n}\,\frac{\bar{x} - \mu_0}{\sigma} \qquad \to \text{“observed value of the test statistic”}$$

As we know that $Z = \sqrt{n}\,\frac{\bar{X} - \mu_0}{\sigma} \sim \mathcal{N}(0,1)$, we have

$$\begin{aligned} p &= 1 - \mathbb{P}\left(\sqrt{n}\,\frac{\bar{X} - \mu_0}{\sigma} \in \left[\sqrt{n}\,\frac{\mu_0 \pm |\bar{x} - \mu_0| - \mu_0}{\sigma}\right]\right) \\ &= 1 - \mathbb{P}(Z \in [-|z_0|, |z_0|]) = 2 \times (1 - \Phi(|z_0|)) \end{aligned}$$

(by symmetry of the $\mathcal{N}(0,1)$ distribution)

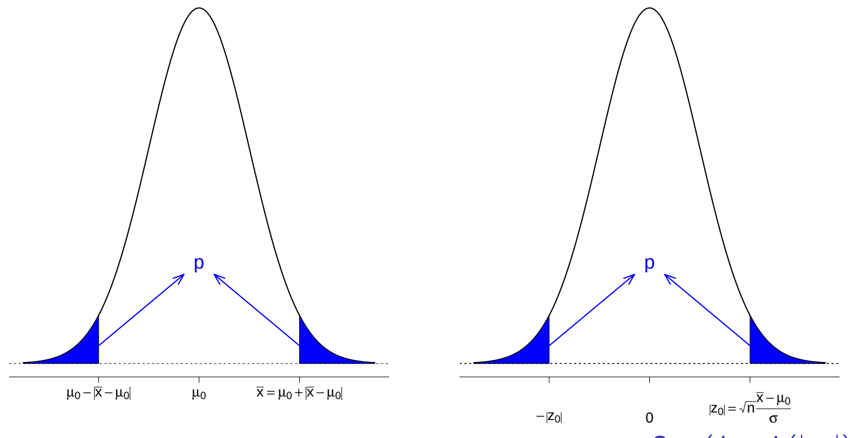
p-value

Distribution of $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$

if $H_0 : \mu = \mu_0$ is true

Standardised distribution of

$Z = \sqrt{n}\,\frac{\bar{X} - \mu_0}{\sigma}$



$\to p = 2 \times (1 - \Phi(|z_0|))$

p-value: example

In our example, we have observed a sample mean of $\bar{x} = 20.75$ min, so that the p-value is given by

$$p = 1 - \mathbb{P}(\bar{X} \in [19.25, 20.75] \text{ when } \mu = 20)$$

We have here that

$$z_0 = \sqrt{36}\,\frac{20.75 - 20}{2} = 2.25,$$

so that the p-value can easily be computed:

$$p = 1 - \mathbb{P}(-2.25 \leq Z \leq 2.25) = 2 \times (1 - \Phi(2.25)) \overset{\text{Matlab}}{=} 0.024$$

If $H_0 = 20$ is true, the probability of obtaining another random sample whose mean is at least as far from 20 as our 20.75 is 0.024.

→ if we rejected H_0, we would be wrong with a 2.4% chance

p-value and significance level

This means that $H_0 : \mu = 20$ would be rejected in favour of $H_a : \mu \neq 20$ at any level of significance greater than or equal to 0.024.

Operationally, once a p-value is computed, we typically compare it to a predefined significance level α to make a decision:

if $\boxed{p < \alpha}$, reject H_0, if $\boxed{p \geq \alpha}$, do not reject H_0

In presenting results and conclusions, it is standard practice to report the observed p-value along with the decision that is made about H_0.

This gives potential other decision makers the possibility to draw a conclusion at any specified level, not only the one you impose to them.

Here, the conclusion would be:

$p = 0.024 < \alpha = 0.05$ → reject H_0 (at significance level 5%)

or

$p = 0.024 > \alpha = 0.01$ → not reject H_0 (at significance level 1%)

Procedures in hypothesis testing

1. State the null and alternative hypotheses: H_0 and H_a
2. Determine the rejection criterion
3. Compute the appropriate test statistic and determine its distribution
4. Calculate the p-value using the test statistics computed
5. Conclusion: reject/do not reject H_0, relate back to the research question

One-sided alternatives

The whole development, yet very similar, must be slightly adapted when **one-sided alternatives** are concerned.

First, *it might occasionally be difficult to choose the appropriate formulation of the alternative*.

In our running example, suppose now that we would like to highlight that the average drying time is actually longer than the advertised 20 min. Would we test for

- $H_0 : \mu = 20$ against $H_a : \mu > 20$, hoping to reject H_0, or
- $H_0 : \mu = 20$ against $H_a : \mu < 20$, hoping not to reject H_0 ?

Recall that rejecting H_0 is a strong conclusion (we have enough evidence to do it),
unlike not rejecting (we do not have enough evidence to conclude, the decision is dictated by risk aversion, not by facts → weak conclusion)

→ always put what we want to prove in the alternative hypothesis

Here, we should test $H_0 : \mu = 20$ against $H_a : \mu > 20$

One-sided alternatives

In a two-sided test, i.e. with alternative $H_a : \mu \neq \mu_0$, an observed value $\bar{x}$ of $\bar{X}$ much smaller than μ_0 **or** much larger than μ_0 is evidence in direction of H_a.

However, if the alternative is $H_a : \mu > \mu_0$, a small value of $\bar{x}$ is not evidence against $H_0 : \mu = \mu_0$ in favour of H_a (H_0 is more likely than H_a if $\bar{X}$ takes a small value even very different to μ_0!)

→ we must only seek evidence against H_0 in the direction of H_a!

1. Thus, in testing
$$H_0 : \mu = \mu_0 \qquad \text{against } H_a : \mu > \mu_0$$
we should only reject H_0 if $\bar{X}$ is much greater than μ_0
2. Similarly, in testing
$$H_0 : \mu = \mu_0 \qquad \text{against } H_a : \mu < \mu_0$$
we should only reject H_0 if $\bar{X}$ is much smaller than μ_0

The critical region is again determined by the significance level α.

One-sided alternatives

1. With $H_a : \mu > \mu_0$, we are after a constant u such that
$$\mathbb{P}(\bar{X} > u \text{ when } \mu = \mu_0) = \alpha$$
As we know that $Z = \sqrt{n}\frac{\bar{X}-\mu}{\sigma} \sim \mathcal{N}(0,1)$, $\boxed{u = \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}}$

→ the decision rule is reject H_0 if $\bar{x} > \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$

Again with the 'observed value of the test statistic' $z_0 = \sqrt{n}\frac{\bar{x}-\mu_0}{\sigma}$, the p-value is given by
$$p = \mathbb{P}(\bar{X} > \bar{x} \text{ when } \mu = \mu_0) = \mathbb{P}\left(Z > \sqrt{n}\,\frac{\bar{x}-\mu_0}{\sigma}\right) = 1 - \Phi(z_0)$$

In our example, with $\bar{x} = 20.75$ and $H_a : \mu > 20$, we have, at significance level 5%, $u = 20 + 1.645 \times \frac{2}{\sqrt{36}} = 20.548$

→ we reject H_0. The p-value is
$$p = \mathbb{P}\left(Z > \sqrt{36}\,\frac{20.75-20}{2}\right) = \mathbb{P}(Z > 2.25) \overset{\text{Matlab}}{=} 0.012 \quad (< 0.05)$$

One-sided alternatives

2. With $H_a : \mu < \mu_0$, we are after a constant l such that

$$\mathbb{P}(\bar{X} < \ell \text{ when } \mu = \mu_0) = \alpha$$

As we know that $Z = \sqrt{n}\frac{\bar{X}-\mu}{\sigma} \sim \mathcal{N}(0,1)$, $\boxed{\ell = \mu_0 - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}}$

→ the decision rule is reject H_0 if $\bar{x} < \mu_0 - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$

The p-value is given by

$$p = \mathbb{P}(\bar{X} < \bar{x} \text{ when } \mu = \mu_0) = \mathbb{P}\left(Z < \sqrt{n}\,\frac{\bar{x}-\mu_0}{\sigma}\right) = \Phi(z_0)$$

In our example, with $\bar{x} = 20.75$ and $H_a : \mu < 20$, we have, at significance level 5%, $\ell = 20 - 1.645 \times \frac{2}{\sqrt{36}} = 19.452$

→ we do not reject H_0. The p-value is

$$p = \mathbb{P}\left(Z < \sqrt{36}\,\frac{20.75-20}{2}\right) = \mathbb{P}(Z < 2.25) \overset{\text{Matlab}}{=} 0.988 \quad (\gg 0.05)$$

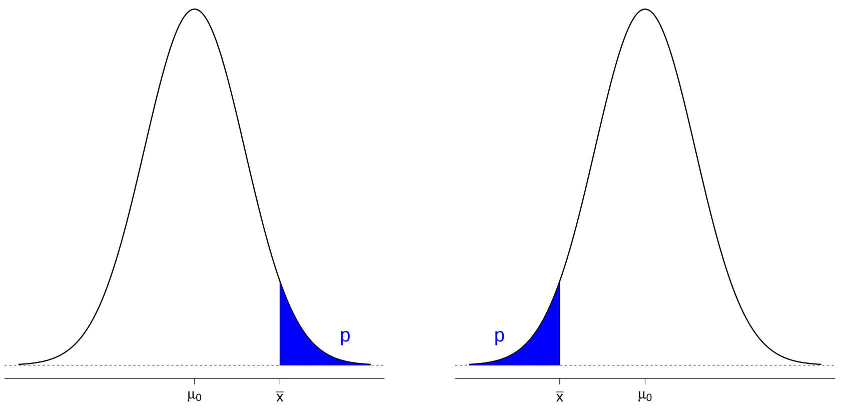
One-sided p-values

Alternative $H_a : \mu > \mu_0$

("Upper-tailed test")

Alternative $H_a : \mu < \mu_0$

("Lower-tailed test")



One-sided alternatives: remark

Remark 1: as announced earlier, the selected alternative does affect the reject/fail to reject of the same null hypothesis at the same significance level! It is therefore important to choose the alternative in a meaningful way

Remark 2: in most situations, the only meaningful way of writing H_a is in the direction which has been observed in the sample

For instance, it is obvious that an observed sample mean $\bar{x}$ greater than μ_0 can never bring evidence in favour of $H_a : \mu < \mu_0$!

In other words, we will never reject a null hypothesis which is not 'challenged' by some evidence in favour of the alternative

→ if we want the test to be relevant, the alternative must be somewhat supported by the observed evidence

The question is just to check whether that evidence is sufficiently strong to reject H_0 or not

Normal populations with unknown standard deviation

So far, we have assumed that we had a normal population with known standard deviation σ, which allowed us to directly use the procedure

$$Z = \sqrt{n}\,\frac{\bar{X}-\mu}{\sigma} \sim \mathcal{N}(0,1)$$

Suppose now that σ is unknown (still in a normal population)

Like we did when deriving confidence intervals, we can replace the unknown σ by its estimator

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

and work with

$$T = \sqrt{n}\,\frac{\bar{X}-\mu}{S}$$

However, we know (Slide 296) that the randomness of S affects the distribution of T, which is no longer normal but

$$\boxed{T \sim t_{n-1}}$$

Normal populations with unknown standard deviation

That apart, everything happens as with a known standard deviation.

Specifically, for the two-sided test $H_0 : \mu = \mu_0$ against $H_a : \mu \neq \mu_0$, the decision rule is

$$\text{reject } H_0 \text{ if } \bar{x} \notin \left[\mu_0 - t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}, \mu_0 + t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}\right],$$

with the observed sample standard deviation

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

and from the observed value of the test statistic

$$t_0 = \sqrt{n}\frac{\bar{x} - \mu_0}{s}$$

we can compute the p-value

$$p = 1 - \mathbb{P}(T \in [-|t_0|, |t_0|]) = 2 \times \mathbb{P}(T > |t_0|) \quad \text{where } T \sim t_{n-1}$$

Normal populations with unknown standard deviation

For the one-sided alternatives, the rejection criteria are

$$\text{reject } H_0 \text{ if } \bar{x} > \mu_0 + t_{n-1,1-\alpha}\frac{s}{\sqrt{n}} \quad \text{or} \quad \text{reject } H_0 \text{ if } \bar{x} < \mu_0 - t_{n-1,1-\alpha}\frac{s}{\sqrt{n}}$$

and the associated p-values are

$$p = \mathbb{P}(T > t_0) \quad \text{or} \quad p = \mathbb{P}(T < t_0)$$

It is no surprise that this test is often called the ***t*-test**, in contrast with the test based on the Normal distribution, called the ***z*-test**.

Note: as for confidence intervals, the t-distribution, with its heavier tails (compared to $\mathcal{N}$), reflects the extra variability introduced in the procedure by the estimation of σ

→ we must be more careful when making the decision, and we reject H_0 'less easily'

t-test: example

Example

The quality of a golf club is, among other things, measured by its 'coefficient of restitution', the ratio of the outgoing velocity of the ball to the incoming velocity of the club. An experiment was performed in which 15 clubs produced by a particular club maker were selected at random and their coefficient of restitution measured:

```
0.8411 0.8191 0.8182 0.8125 0.8750 0.8580 0.8532 0.8483
       0.8276 0.7983 0.8042 0.8730 0.8282 0.8359 0.8660
```

The maker claims that the mean coefficient of restitution of its clubs exceeds 0.82. From the observations we have, is there evidence (at level 0.05) to support the maker's claim? (**Hint:** You can use the following Matlab output: `tinv`$(0.95, 14) = 1.76$, `tcdf`$(2.72, 14) = 0.992$)

The observed sample mean and sample standard deviation are $\bar{x} = 0.83725$ and $s = 0.02456$. Since we would like to demonstrate that μ (the 'true' mean coefficient of restitution) exceeds 0.82, a one-sided test is appropriate:

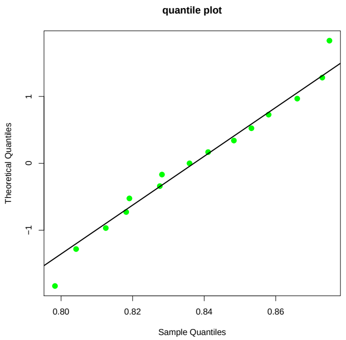
$$H_0 : \mu = 0.82 \qquad \text{against} \qquad H_a : \mu > 0.82$$

t-test: example

The quantile plot of the data supports the assumption that the coefficient of restitution is normally distributed

→ t-test

Note that $t_{14;0.95} = 1.76$ (hint)



We will reject H_0 in favour of H_a if the observed $\bar{x}$ is 'too large'. The decision rule is

$$\text{reject } H_0 \text{ if } \bar{x} > \mu_0 + t_{n-1,1-\alpha}\frac{s}{\sqrt{n}} = 0.82 + 1.76\frac{0.02456}{\sqrt{15}} = 0.8312$$

→ we reject H_0, it is clear enough that the maker is right

t-test: example

We should also compute the *p*-value. The observed value of the test statistic is

$$t_0 = \sqrt{n}\,\frac{\bar{x}-\mu_0}{s} = \sqrt{15}\,\frac{0.83725 - 0.82}{0.02456} = 2.72,$$

hence, for $T \sim t_{14}$,

$$p = \mathbb{P}(T > 2.72) = 1 - 0.992 = 0.008,$$

→ as expected, $p < 0.05$

→ we reject H_0 and we can be 99.2% confident in our decision of supporting the maker's claim

Non-normal populations

Assuming that the population is normal, we use the results

$$Z = \sqrt{n}\,\frac{\bar{X}-\mu}{\sigma} \sim \mathcal{N}(0,1) \quad \text{or} \quad T = \sqrt{n}\,\frac{\bar{X}-\mu}{S} \sim t_{n-1}$$

What if the population is not normal ?

By the Central Limit Theorem (CLT, Slide 284)

$$\sqrt{n}\,\frac{\bar{X}-\mu}{\sigma} \overset{a}{\sim} \mathcal{N}(0,1),$$

we can carry over all our *z*-test procedures to the arbitrary population case, bearing in mind that the results require *n* 'large enough' and are only approximately right.

Further, we know (Slide 307) that the estimation of σ by S does not dramatically affect that result:

$$\sqrt{n}\,\frac{\bar{X}-\mu}{S} \overset{a}{\sim} \mathcal{N}(0,1)$$

→ the above observation holds true even if σ needs to be estimated

Non-normal populations: two-sided large sample test

Hence, the test for

$$H_0 : \mu = \mu_0 \qquad \text{against} \qquad H_a : \mu \neq \mu_0$$

(two-sided test) using the decision rule

$$\text{reject } H_0 \text{ if } \bar{x} \notin \left[\mu_0 - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \mu_0 + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

(if σ is known) or

$$\text{reject } H_0 \text{ if } \bar{x} \notin \left[\mu_0 - z_{1-\alpha/2}\frac{s}{\sqrt{n}}, \mu_0 + z_{1-\alpha/2}\frac{s}{\sqrt{n}}\right]$$

(σ unknown), will have an **approximate significance level** α, provided that *n* is large enough, regardless of the population distribution

As on Slide 348, the associated **approximate *p*-value** will be given by

$$p = 2 \times (1 - \Phi(|z_0|)),$$

with $z_0 = \sqrt{n}\,\frac{\bar{x}-\mu_0}{\sigma}$ or $z_0 = \sqrt{n}\,\frac{\bar{x}-\mu_0}{s}$, the observed value of the test statistic.

Non-normal populations: one-sided large sample test

For the one-sided test for $H_0 : \mu = \mu_0$ against

$$H_a : \mu > \mu_0 \qquad \text{or} \qquad H_a : \mu < \mu_0,$$

the decision rules

$$\text{reject } H_0 \text{ if } \bar{x} > \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}} \quad \text{or} \quad \text{reject } H_0 \text{ if } \bar{x} < \mu_0 - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$$

(σ known) and

$$\text{reject } H_0 \text{ if } \bar{x} > \mu_0 + z_{1-\alpha}\frac{s}{\sqrt{n}} \quad \text{or} \quad \text{reject } H_0 \text{ if } \bar{x} < \mu_0 - z_{1-\alpha}\frac{s}{\sqrt{n}}$$

(σ unknown) will have **approximate significance level** α, provided that *n* is large enough, regardless of the population distribution

As on Slides 355-356, the associated **approximate *p*-values** are

$$p = 1 - \Phi(z_0) \quad \text{or} \quad p = \Phi(z_0)$$

These tests are called **large sample tests**, as they require *n* large ($n > 40$, say) to be (approximately) valid.

Large sample test: example

Example

A manager in charge of sales for a large corporation claims that salespeople are averaging no more than 15 sales contacts a week (he would like to increase this number). As a check on his claim, $n = 49$ salespeople are selected at random and the number of contacts made by each of them is recorded for a single randomly selected week. The mean and variance of the 49 measurements were $\bar{x} = 17$ and $s^2 = 9$, respectively. Do these figures contradict the manager's claim, at the significance level $\alpha = 0.05$? (**Hint:** You can use the Matlab outputs: $\texttt{norminv}(0.95) = 1.645$, $\texttt{normcdf}(4.67) = 1$)

We would like to show that the manager's claim is incorrect, which is formally written $H_a : \mu > 15$, where μ is the mean number of sales contacts per week. Thus, we will test

$$H_0 : \mu = 15 \qquad \text{against} \qquad H_a : \mu > 15$$

We do not need to know the distribution of the number of weekly contacts, as n is 'large enough' to ensure the (approximate) validity of the test procedure, regardless of that distribution → **large sample test**

Large sample test: example

We have a one-sided test, so that the decision rule will be (note that $z_{0.95} = 1.645$ (hint)):

$$\text{reject } H_0 \text{ if } \bar{x} > \mu_0 + z_{1-\alpha}\frac{s}{\sqrt{n}} = 15 + 1.645\frac{\sqrt{9}}{\sqrt{49}} = 15.705$$

We have observed

$$\bar{x} = 17,$$

largely beyond the critical value 15.705, so that we reject H_0

→ at the $\alpha = 0.05$ level of significance (that is, permitting a 5% chance of error), the evidence is sufficient to indicate that the manager's claim is incorrect and that the average number of contacts per week exceeds 15

We also have

$$z_0 = \sqrt{49}\,\frac{17 - 15}{\sqrt{9}} = 4.67,$$

so that the (approximate) p-value is given by

$$p = 1 - \Phi(4.67) \simeq 0$$

→ we do not take much risk when contradicting the manager

Hypothesis tests and confidence intervals

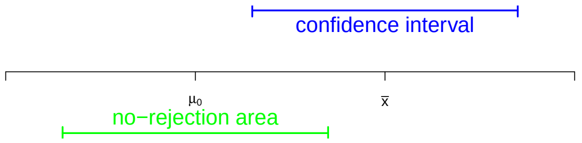
You might have noticed that the critical values for the (two-sided test) decision rule

$$\text{reject } H_0 \text{ if } \bar{x} \notin \left[\mu_0 - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \mu_0 + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

look like the limits of the (two-sided) confidence interval for μ

$$\left[\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

As the interval widths are the same, the confidence interval (centred at $\bar{x}$) cannot contain μ_0 if the 'no-rejection area' (centred at μ_0) does not contain $\bar{x}$ (and vice-versa)



Hypothesis tests and confidence intervals

Generally speaking, there is always a **close relationship between the test of hypothesis about any parameter, say θ, and the confidence interval for θ**:

If $[\ell, u]$ is a $100 \times (1 - \alpha)\%$ confidence interval for a parameter θ, then the hypothesis test for

$$H_0 : \theta = \theta_0 \qquad \text{against} \qquad H_a : \theta \neq \theta_0$$

will reject H_0 at significance level α if and only if θ_0 is not in $[\ell, u]$

→ hypothesis tests and CIs are more or less equivalent, however each provides somewhat different insights:

- CIs provide a range of likely values for θ
- tests easily display the risk levels, such as p-values, associated with a specific decision

Note: the same analogy exists between one-sided tests and one-sided confidence intervals

Objectives

Now you should be able to:

- structure engineering decision-making problems as hypothesis tests □
- understand the concepts of significance level, power, error of type I and of type II □
- test hypotheses on the mean of a normal distribution using either a z-test or a t-test □
- test hypotheses on the mean of an arbitrary distribution using the Central Limit Theorem □
- use the p-value approach for making decisions in hypothesis tests □
- explain and use the relationship between confidence intervals and hypothesis tests □

Recommended exercises:

→ Q2, Q3, Q5, Q8, Q9, Q11 p.354, Q15 p.355, Q17, Q18, Q19 p.367, Q21, Q23 p.368, Q55 p.394, Q62 p.396 (2nd edition)

→ Q2, Q3, Q5 p.361, Q8, Q9, Q11, Q15 p.362, Q18, Q19, Q20 p.374, Q22, Q24 p.375, Q65 p.406, Q73 p.408 (3rd edition)

Hypothesis tests for a proportion

When a proportion/probability π is the population parameter of interest, it can naturally be estimated from the sample by the sample proportion

$$\hat{P} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

where $X_i = 1$ if the ith individual of the sample has the characteristic, and $X_i = 0$ if not.

As a sample mean, $\hat{P}$ obeys the Central Limit Theorem and we have

$$\sqrt{n}\,\frac{\hat{P}-\pi}{\sqrt{\pi(1-\pi)}} \overset{a}{\sim} \mathcal{N}(0,1) \qquad \text{(for } n \text{ 'large')}$$

This allowed us to derive a large-sample confidence interval for π:

$$\left[\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

(Slides 317–321)

Hypothesis tests for a proportion

From

$$\sqrt{n}\,\frac{\hat{P}-\pi}{\sqrt{\pi(1-\pi)}} \overset{a}{\sim} \mathcal{N}(0,1)$$

it is also straightforward to derive testing procedures for hypotheses about the proportion π, similar to the test procedures for μ

For some value $\pi_0 \in (0,1)$ that we may have in mind, we will consider testing

$$H_0 : \pi = \pi_0 \qquad \text{against} \qquad H_a : \pi \neq \pi_0$$

Application of previous results (Slide 344) implies that the decision rule at (approximate) significance level α is

$$\text{reject } H_0 \text{ if } \hat{p} \notin \left[\pi_0 - z_{1-\alpha/2}\sqrt{\frac{\pi_0(1-\pi_0)}{n}}, \pi_0 + z_{1-\alpha/2}\sqrt{\frac{\pi_0(1-\pi_0)}{n}}\right]$$

Hypothesis tests for a proportion

Note: as $\alpha = \mathbb{P}(\text{reject } H_0 \text{ \underline{when it is true}})$, we take $\pi = \pi_0$ everywhere in the derivation of the decision rule, so that the standard error of the estimation here appears as $\sqrt{\frac{\pi_0(1-\pi_0)}{n}}$

The (approximate) p-value for this test is also calculated as in the previous chapter (Slide 348), that is,

$$p = 2 \times (1 - \Phi(|z_0|)),$$

where z_0 is the observed value of the test statistic when $\pi = \pi_0$:

$$z_0 = \sqrt{n}\,\frac{\hat{p} - \pi_0}{\sqrt{\pi_0(1-\pi_0)}}$$

This test is called a **large sample test** for a proportion.

Hypothesis tests for a proportion

For the one-sided test for $H_0 : \pi = \pi_0$ against

$$H_a : \pi > \pi_0 \qquad \text{or} \qquad H_a : \pi < \pi_0,$$

the decision rules

$$\text{reject } H_0 \text{ if } \hat{p} > \pi_0 + z_{1-\alpha}\sqrt{\frac{\pi_0(1-\pi_0)}{n}}$$

or

$$\text{reject } H_0 \text{ if } \hat{p} < \pi_0 - z_{1-\alpha}\sqrt{\frac{\pi_0(1-\pi_0)}{n}}$$

will have **approximate significance level** α.

The associated **approximate p-values** will be (Slides 355-356)

$$p = 1 - \Phi(z_0) \qquad \text{or} \qquad p = \Phi(z_0)$$

(one-sided large-sample tests for a proportion)

Hypothesis tests for a proportion: example

Example

Transceivers provide wireless communication among electronic components of consumer products. Responding to a need for a fast, low-cost test of Bluetooth-capable transceivers, engineers developed a product test at the wafer level. In one set of trials with 60 devices selected from different wafer lots, 48 devices passed. Denote π the population proportion of transceivers that would pass. Test the null hypothesis $\pi = 0.70$ against $\pi > 0.70$ at the 0.05 significance level. (**Matlab outputs:** `norminv`$(0.95) = 1.645$, `normcdf`$(1.69) = 0.9545$)

This is a one-sided hypothesis test for π:

$$H_0 : \pi = 0.70 \qquad \text{against } H_a : \pi > 0.70$$

The estimate of π is

$$\hat{p} = \frac{48}{60} = 0.80$$

Hypothesis tests for a proportion: example

The decision rule is

$$\text{reject } H_0 \text{ if } \hat{p} > 0.70 + 1.645 \times \sqrt{\frac{0.70 \times (1-0.70)}{60}} = 0.7973$$

→ reject H_0 !

(at significance level $\alpha = 0.05$, there is enough evidence to conclude that the proportion of good transceivers that would be produced is greater than 0.70)

The observed value of the test statistics is

$$z_0 = \sqrt{60}\,\frac{0.80 - 0.70}{\sqrt{0.70 \times (1-0.70)}} = 1.69$$

and the associated p-value is

$$p = 1 - \Phi(1.69) \overset{\text{hint}}{=} 1 - 0.9545 = 0.0455$$

→ at level $\alpha = 0.05$, we do reject H_0

8 Inferences concerning a variance

Inferences concerning a variance: introduction

- In the previous chapter, we saw how to make inferences about the population mean μ, and as a particular case, about a population proportion π
- Very similar methods apply to inferences about other population parameters, like the **variance** σ^2
- Variances and standard deviations are not only important in their own right, they must sometimes be estimated before inferences about other parameters can be made

Estimation of a variance

In Chapter 7, there were several instances where we estimated a population standard deviation by means of a sample standard deviation (e.g. in the derivation of the t-confidence interval for μ).

The **sample variance** of a random sample $\{X_1, X_2, \ldots, X_n\}$ with mean $\bar{X}$ is given by

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

and is obviously a natural estimator for the population variance σ^2.

We can write

$$(X_i - \bar{X})^2 = (X_i - \mu + \mu - \bar{X})^2 = (X_i - \mu)^2 + (\mu - \bar{X})^2 + 2(X_i - \mu)(\mu - \bar{X}),$$

so that

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 + n(\mu - \bar{X})^2 + 2(\mu - \bar{X})\sum_{i=1}^{n}(X_i - \mu)$$

$$= \sum_{i=1}^{n}(X_i - \mu)^2 + n(\bar{X} - \mu)^2 - 2n(\bar{X} - \mu)^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2$$

Estimation of a variance

We know that $\mathbb{V}\text{ar}(X_i) = \mathbb{E}((X_i - \mu)^2) = \sigma^2$ and $\mathbb{V}\text{ar}(\bar{X}) = \mathbb{E}((\bar{X} - \mu)^2) = \frac{\sigma^2}{n}$, hence

$$\mathbb{E}\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right) = n\sigma^2 - n\frac{\sigma^2}{n} = (n-1)\sigma^2$$

and thus

$$\mathbb{E}(S^2) = \frac{1}{n-1}\mathbb{E}\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right) = \frac{(n-1)\sigma^2}{n-1} = \sigma^2$$

→ S^2 is an **unbiased estimator** of σ^2 (and **consistent** (not shown))

Note: this makes it clear why the divisor in S^2 must be $n-1$, not n. If we divided by n, the resulting estimator would be biased!

Looking at the maths, it can be understood that we actually lose one degree of freedom because we have to estimate the unknown μ by $\bar{X}$ in the expression.

Fact: We lose one degree of freedom for each estimated parameter.

Sampling distribution in a normal population

Since S^2 cannot be negative, we should suspect that the sampling distribution of the sample variance is not normal.

Actually, in general, little can be said about this sampling distribution.

However, **when the population is normal**, the sampling distribution of S^2 can be derived and turns out to be related to the so-called

chi-square distribution

If $X_1, X_2, \ldots, X_n$ is a random sample from a normal population with mean μ and variance σ^2, then

$$\boxed{\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}}$$

→ χ^2_{n-1} denotes the chi-square distribution with $n-1$ degrees of freedom

The χ^2-distribution

A random variable, say X, is said to follow the chi-square-distribution with ν degrees of freedom, i.e.

$$\boxed{X \sim \chi^2_\nu}$$

if its probability density function is given by

$$f(x) = \frac{1}{2^{\nu/2}\Gamma\left(\frac{\nu}{2}\right)} x^{\nu/2-1} e^{-x/2} \qquad \text{for } x > 0 \qquad \to S_X = [0, +\infty)$$

for some integer ν

Note: the Gamma function is given by

$$\Gamma(y) = \int_0^{+\infty} x^{y-1} e^{-x}\, dx, \qquad \text{for } y > 0$$

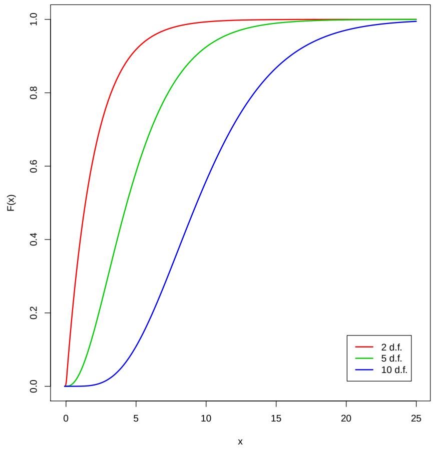
It can be shown that $\Gamma(y) = (y-1) \times \Gamma(y-1)$, so that, if y is a positive integer n,

$$\Gamma(n) = (n-1)!$$

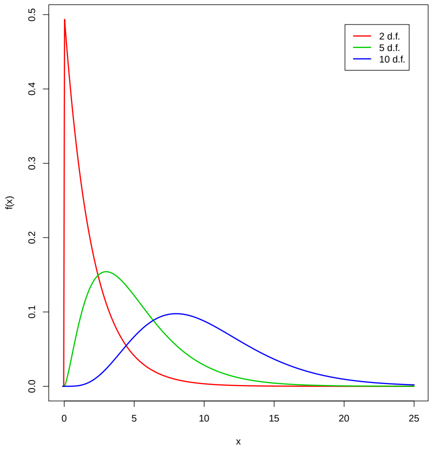
There is usually no simple expression for the χ^2-cdf.

The χ^2-distribution

Some χ^2-distributions, with $\nu = 2$, $\nu = 5$ and $\nu = 10$



cdf $F(x)$



pdf $f(x) = F'(x)$

The χ^2-distribution

It can be shown that the mean and the variance of the χ^2_ν-distribution are

$$\mathbb{E}(X) = \nu \qquad \text{and} \qquad \mathbb{V}\text{ar}(X) = 2\nu$$

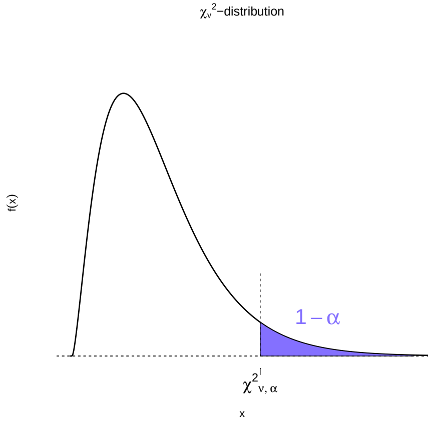
Note that a χ^2-distributed random variable is nonnegative and the distribution is skewed to the right.

However, as ν increases, the distribution becomes more and more symmetric.

In fact, it can be shown that the standardised χ^2- distribution with ν degrees of freedom approaches the standard normal distribution as $\nu \to \infty$.

The χ^2-distribution: quantiles

Similarly to what we did for other distributions, we can define the quantiles of any χ^2-distribution:



Let $\chi^2_{\nu;\alpha}$ be the value such that

$$\mathbb{P}(X > \chi^2_{\nu;\alpha}) = 1 - \alpha$$

for $X \sim \chi^2_\nu$

Careful! unlike the standard normal distribution (or the t-distribution), the χ^2-distribution is not symmetric

Confidence interval for the population variance (normal population)

As we know that $\boxed{\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}}$,

we can write $\mathbb{P}\left(\chi^2_{n-1;\alpha/2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{n-1;1-\alpha/2}\right) = 1 - \alpha$, which can be rearranged as

$$\mathbb{P}\left(\frac{(n-1)S^2}{\chi^2_{n-1;1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1;\alpha/2}}\right) = 1 - \alpha,$$

→ if s is the observed sample variance in a random sample of size n drawn from a normal population, then a two-sided $100 \times (1-\alpha)\%$ confidence interval for σ^2 is

$$\boxed{\left[\frac{(n-1)s^2}{\chi^2_{n-1;1-\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1;\alpha/2}}\right]}$$

Hypothesis test for the population variance (normal population)

Of course, the sampling distribution $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ is also the basis of test procedures for hypotheses about the population variance.

For instance, consider testing $H_0 : \sigma^2 = \sigma_0^2$ against $H_a : \sigma^2 \neq \sigma_0^2$

As S^2 is supposed to be 'close' to σ^2, we will reject H_0 whenever the observed s^2 will be too distant from σ_0^2:

at significance level α, we are after two constants ℓ and u such that

$$\alpha = \mathbb{P}(S^2 \notin [\ell, u] \text{ when } \sigma^2 = \sigma_0^2) = \mathbb{P}\left(\frac{(n-1)S^2}{\sigma_0^2} \notin \left[\frac{(n-1)\ell}{\sigma_0^2}, \frac{(n-1)u}{\sigma_0^2}\right]\right)$$

$$\rightarrow \boxed{\ell = \frac{\chi^2_{n-1;\alpha/2}\sigma_0^2}{n-1}} \quad \text{and} \quad \boxed{u = \frac{\chi^2_{n-1;1-\alpha/2}\sigma_0^2}{n-1}}$$

→ the decision rule is:

$$\text{reject } H_0 \text{ if } s^2 \notin \left[\frac{\chi^2_{n-1;\alpha/2}\sigma_0^2}{n-1}, \frac{\chi^2_{n-1;1-\alpha/2}\sigma_0^2}{n-1}\right]$$

Hypothesis test for the population variance: example

Example

The lapping process which is used to grind certain silicon wafers to the proper thickness is acceptable only if σ, the population standard deviation of the thickness of dice cut from the wafers, is at most 0.50 mm. On a given day, 15 dice cut from such wafers were observed and their thickness showed a sample standard deviation of 0.64 mm. Use the 0.05 level of significance to test the hypothesis that $\sigma = 0.50$ on that day. (**Matlab outputs:** `chi2inv`$(0.95, 14) = 23.68$, `chi2cdf`$(22.94, 14) = 0.9387$)

We are only interested in $H_a : \sigma^2 > 0.50^2 = 0.25$, thus we need a **one-sided test**. Similarly to what we did for the two-sided case, we will here reject H_0 if the observed s^2 is much larger than 0.25, with rejection criterion being

$$\text{reject } H_0 \text{ if } s^2 > \frac{\chi^2_{n-1;1-\alpha}\sigma_0^2}{n-1}$$

Here, $n = 15$, and $\chi^2_{14;0.95} = 23.68$ (hint), so the rule is

$$\text{reject } H_0 \text{ if } s^2 > \frac{23.68 \times 0.25}{14} = 0.4229$$

Hypothesis test for the population variance: example

The value we have observed is $s^2 = 0.64^2 = 0.4096$

→ do not reject H_0

Even if $s = 0.64 > 0.50$, this is not enough evidence to conclude that the lapping process was unsatisfactory on that day.

We can compute the *p*-value of the test:

The observed value of the test statistic under H_0 is

$$\frac{(n-1)s^2}{\sigma_0^2} = \frac{14 \times 0.4096}{0.25} = 22.94$$

→ The p-value $p = \mathbb{P}(X > 22.94)$ (where $X \sim \chi^2_{14}$).

According to the hint, $p = 1 - 0.9387 = 0.0613 > 0.05$.

(confirmation that we do not reject H_0 at level 5%)

9 Inferences concerning a difference of means

Inferences concerning a difference of means

Advances occur in engineering when new ideas lead to better equipment, new materials, or revision of existing production processes.

Any new procedure or device **must be compared** with the existing one and the amount of improvement assessed.

Furthermore, in many situations it is quite common to be interested in **comparing two 'populations'** in regard to a parameter of interest.

The two 'populations' may be:

- produced items using an existing and a new technique
- success rates in two groups of individuals
- health test results for patients who received a drug and for patients who received a placebo
- ...

As usual, we are not able to observe the whole populations

→ we need statistical inference methods to **make comparisons between two different populations**, having only observed two samples from them

Inferences concerning a difference of means

- For instance, suppose that the paint manufacturer of the new 'fast-drying' paint want to reduce further drying time of the paint
- Two formulations of the paint are tested: formulation 1 is the standard chemistry, while formulation 2 has a new drying ingredient that should reduce the drying time
- From experience, it is known that the standard deviation of drying time is 1.3 minutes, and this should be unaffected by the addition of the new ingredient
- Ten specimens are painted with formulation 1 and another 10 are painted with formulation 2, in random order
- The two sample average drying times are $\bar{x}_1 = 20.17$ min and $\bar{x}_2 = 18.67$ min, respectively
- What conclusions can the manufacturer draw about the effectiveness of the new ingredient ?

Hypothesis test for the difference in means

The general situation is as follows:

- Population 1 has mean μ_1 and standard deviation σ_1
- Population 2 has mean μ_2 and standard deviation σ_2

Inferences will be based on two random samples of sizes n_1 and n_2:

$$X_{11}, X_{12}, \ldots, X_{1n_1} \text{ is a sample from population 1}$$

$$X_{21}, X_{22}, \ldots, X_{2n_2} \text{ is a sample from population 2}$$

We will first assume that the samples are **independent** (i.e., observations in sample 1 are by no means linked to the observations in sample 2, they concern different individuals)

What we would like to know is whether $\mu_1 = \mu_2$ or not

→ hypothesis test

Hypothesis test for $\mu_1 = \mu_2$

We can formalise this by stating the null hypothesis as:

$$\boxed{H_0 : \mu_1 = \mu_2}$$

Then, the hypothesis test idea can be understood as it was done in the one-sample case: we observe two samples for which we compute the sample means $\bar{x}_1$ and $\bar{x}_2$.

As $\bar{x}_1$ is supposed to be a good estimate of μ_1 and $\bar{x}_2$ is supposed to be a good estimate of μ_2:

- if $\bar{x}_1 \simeq \bar{x}_2$, then H_0 is probably acceptable
- if $\bar{x}_1$ is considerably different to $\bar{x}_2$, that is evidence that H_0 is not true and we are tempted to reject it

Note that the alternative hypothesis can be

$$H_a : \mu_1 \neq \mu_2 \qquad \text{(two-sided alternative)}$$

$$\text{or} \qquad H_a : \mu_1 > \mu_2 \quad \text{or} \quad H_a : \mu_1 < \mu_2 \qquad \text{(one-sided alternatives)}$$

Hypothesis test for $\mu_1 = \mu_2$

We know (Central Limit Theorem, Slide 284) that

$$\bar{X}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1i} \overset{(a)}{\sim} \mathcal{N}\left(\mu_1, \frac{\sigma_1}{\sqrt{n_1}}\right) \quad \text{and} \quad \bar{X}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} X_{2i} \overset{(a)}{\sim} \mathcal{N}\left(\mu_2, \frac{\sigma_2}{\sqrt{n_2}}\right)$$

('$\overset{(a)}{\sim}$' means that these are exact results for any n_1, n_2 if the populations are normal, approximate results for large n_1, n_2 if they are not)

We also know (Slide 226) that if $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$ are **independent**, then $aX_1 + bX_2 \sim \mathcal{N}\left(a\mu_1 + b\mu_2, \sqrt{a^2\sigma_1^2 + b^2\sigma_2^2}\right)$

→ we deduce the sampling distribution of $\bar{X}_1 - \bar{X}_2$:

$$\boxed{\bar{X}_1 - \bar{X}_2 \overset{(a)}{\sim} \mathcal{N}\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)}$$

Now, as testing for $H_0 : \mu_1 = \mu_2$ exactly amounts to testing for $H_0 : \mu_1 - \mu_2 = 0$, the one-sample procedure we introduced in Chapter 7 can be used up to some light adaptation, with $\bar{X}_1 - \bar{X}_2$ as an estimator for $\mu_1 - \mu_2$

Hypothesis test for $\mu_1 = \mu_2$

Suppose that σ_1 and σ_2 are known, and that we have observed two samples $x_{11}, x_{12}, \ldots, x_{1n_1}$ and $x_{21}, x_{22}, \ldots, x_{2n_2}$ whose respective means are

$$\bar{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i} \quad \text{and} \quad \bar{x}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{2i}$$

For the two-sided test (with $H_a : \mu_1 - \mu_2 \neq 0$), at significance level α, the decision rule is

$$\text{reject } H_0 \text{ if } \bar{x}_1 - \bar{x}_2 \notin \left[-z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$$

(interval obviously centred at 0 by H_0)

The associated p-value is given by $p = 2 \times (1 - \Phi(|z_0|))$

where z_0 is the z-score of $\bar{x}_1 - \bar{x}_2$ if $\mu_1 - \mu_2 = 0$, i.e. $z_0 = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

Hypothesis test for $\mu_1 = \mu_2$

Similarly, for the one-sided test with alternative $H_a : \mu_1 > \mu_2$, the decision rule is

$$\text{reject } H_0 \text{ if } \bar{x}_1 - \bar{x}_2 > z_{1-\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

and the associated p-value is

$$p = 1 - \Phi(z_0),$$

while for the one-sided test with alternative $H_a : \mu_1 < \mu_2$, the decision rule is

$$\text{reject } H_0 \text{ if } \bar{x}_1 - \bar{x}_2 < -z_{1-\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

and the associated p-value is

$$p = \Phi(z_0)$$

Remark: these decision rules will lead to tests of **approximate** level α if the populations are not normal but n_1 and n_2 are large enough

Hypothesis test for $\mu_1 = \mu_2$: example

In our running example, define μ_1 the true average drying time for the formulation 1 paint, and μ_2 the true average drying time for the formulation 2 paint (with the new ingredient).

We have observed two samples of sizes $n_1 = n_2 = 10$ from both populations with known standard deviations $\sigma_1 = \sigma_2 = 1.3$, with sample means $\bar{x}_1 = 20.17$ and $\bar{x}_2 = 18.67$. Assume that both populations are normal.

At level $\alpha = 0.05$, for the alternative $H_a : \mu_1 > \mu_2$, the decision rule is

$$\text{reject } H_0 \text{ if } \bar{x}_1 - \bar{x}_2 > 1.645 \times \sqrt{\frac{1.3^2}{10} + \frac{1.3^2}{10}} = 0.956$$

→ here, we have observed $\bar{x}_1 - \bar{x}_2 = 1.5 > 0.956$, so we reject H_0

Furthermore, $z_0 = \frac{1.5}{\sqrt{\frac{1.3^2}{10} + \frac{1.3^2}{10}}} = 2.58$, so that the p-value is

$$p = 1 - \Phi(2.58) \overset{\text{Matlab}}{=} 0.0049$$

→ adding the new ingredient significantly reduces the drying time

Confidence interval for $\mu_1 - \mu_2$

As we observed, there is a strong relationship between hypothesis tests and confidence intervals

→ we can directly derive a confidence interval for $\mu_1 - \mu_2$

We note that $\bar{X}_1 - \bar{X}_2 \overset{(a)}{\sim} \mathcal{N}\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$, so

$$1 - \alpha = \mathbb{P}\left(-z_{1-\alpha/2} \leq \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq z_{1-\alpha/2}\right)$$
$$= \mathbb{P}\left(\mu_1 - \mu_2 \in \left[(\bar{X}_1 - \bar{X}_2) \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]\right)$$

→ from two observed samples, we have that a $100 \times (1 - \alpha)\%$ two-sided confidence interval for $\mu_1 - \mu_2$ is

$$\left[(\bar{x}_1 - \bar{x}_2) - z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, (\bar{x}_1 - \bar{x}_2) + z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$$

Confidence interval for $\mu_1 - \mu_2$

Similarly, $100 \times (1 - \alpha)\%$ one-sided confidence intervals for $\mu_1 - \mu_2$ are $\left(-\infty, (\bar{x}_1 - \bar{x}_2) + z_{1-\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$ and

$$\left[(\bar{x}_1 - \bar{x}_2) - z_{1-\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, +\infty\right)$$

In our running example, for instance, we have that a 95% one-sided confidence interval for $\mu_1 - \mu_2$ is

$$\left[1.5 - 1.645 \times \sqrt{\frac{1.3^2}{10} + \frac{1.3^2}{10}}, +\infty\right) = [0.544, +\infty)$$

→ we can be 95% confident that the gain in drying time is at least 0.544 minutes

A two-sided 95% confidence interval for $\mu_1 - \mu_2$ would be

$$\left[1.5 - 1.96 \times \sqrt{\frac{1.3^2}{10} + \frac{1.3^2}{10}}, 1.5 + 1.96 \times \sqrt{\frac{1.3^2}{10} + \frac{1.3^2}{10}}\right] = [0.47, 2.63]$$

Hypothesis test for $\mu_1 = \mu_2$

A generalisation of the previous procedure is to deal with the unknown variance case.

However, two different situations must be treated:

1. the standard deviations of the two distributions are unknown but equal: $\sigma_1 = \sigma_2 = \sigma$
2. the standard deviations of the two distributions are unknown but not necessarily equal: $\sigma_1 \neq \sigma_2$

These situations must be differentiated as we will need to estimate the unknown variance(s).

→ estimating one parameter σ from all the observations, or estimating two parameters σ_1 and σ_2 each from half of the observations, will lead to different results

Hypothesis test for $\mu_1 = \mu_2$ (with $\sigma_1^2 = \sigma_2^2$)

Assume for now that $\sigma_1 = \sigma_2 = \sigma$, but σ is unknown → estimate it !

Each squared deviation $(X_{1i} - \bar{X}_1)^2$ is an estimator for σ^2 in population 1, and each squared deviation $(X_{2i} - \bar{X}_2)^2$ is an estimator for σ^2 in population 2

→ we estimate σ^2 by pooling the sums of squared deviations from the respective sample means, thus we estimate σ^2 by the **pooled variance estimator**:

$$\boxed{S_p^2 = \frac{\sum_{i=1}^{n_1}(X_{1i} - \bar{X}_1)^2 + \sum_{i=1}^{n_2}(X_{2i} - \bar{X}_2)^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

where $S_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_{1i} - \bar{X}_1)^2$ and $S_2^2 = \frac{1}{n_2-1}\sum_{i=1}^{n_1}(X_{2i} - \bar{X}_2)^2$

Note: the pooled variance estimator has $n_1 + n_2 - 2$ degrees of freedom, because we have $n_1 - 1$ independent deviations from the mean in the first sample, and $n_2 - 1$ independent deviations from the mean in the second sample → altogether, $n_1 + n_2 - 2$ independent deviations to estimate σ^2

Hypothesis test for $\mu_1 = \mu_2$ (with $\sigma_1^2 = \sigma_2^2$)

In the one sample case, we had (Slide 296)

$$\boxed{\sqrt{n}\,\frac{\bar{X}-\mu}{\sigma} \sim \mathcal{N}(0,1)} + \boxed{S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2} \Rightarrow \boxed{\sqrt{n}\,\frac{\bar{X}-\mu}{S} \sim t_{n-1}}$$

Similarly, we have now

$$\boxed{\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \overset{(a)}{\sim} \mathcal{N}(0,1)} + \boxed{S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$$

$$\Rightarrow \boxed{\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \overset{(a)}{\sim} t_{n_1+n_2-2}}$$

Note: for non-normal populations, in large samples (n_1 and n_2 ‘large’), we know that $t_{n_1+n_2-2} \approx \mathcal{N}(0,1)$ and that the CLT gives approximate results anyway → we can use $\mathcal{N}(0,1)$

Hypothesis test for $\mu_1 = \mu_2$ (with $\sigma_1^2 = \sigma_2^2$)

Suppose we have observed two samples $x_{11}, x_{12}, \ldots, x_{1n_1}$ and $x_{21}, x_{22}, \ldots, x_{2n_2}$ whose respective means and standard deviations are

$$\bar{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i} \quad \text{and} \quad \bar{x}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{2i}$$

and

$$s_1 = \sqrt{\frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2} \quad \text{and} \quad s_2 = \sqrt{\frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(x_{2i} - \bar{x}_2)^2}$$

→ the observed pooled sample standard deviation is

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

which should be a good estimate of σ (if $\sigma_1 = \sigma_2$!)

Hypothesis test for $\mu_1 = \mu_2$ (with $\sigma_1^2 = \sigma_2^2$)

For the two-sided test (with $H_a : \mu_1 - \mu_2 \neq 0$), at significance level α, the decision rule is thus

reject $H_0 : \mu_1 = \mu_2$ if

$$\bar{x}_1 - \bar{x}_2 \notin \left[-t_{n_1+n_2-2;1-\alpha/2}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, t_{n_1+n_2-2;1-\alpha/2}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right]$$

The associated p-value is given by

$$p = 2 \times \mathbb{P}(T > |t_0|) \qquad \text{with } T \sim t_{n_1+n_2-2},$$

where t_0 is the observed value of the test statistic (with $\mu_1 - \mu_2 = 0$)

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

This test is known as the **two-sample t-test**.

Hypothesis test for $\mu_1 = \mu_2$ (with $\sigma_1^2 = \sigma_2^2$)

One-sided versions of this test are also available. For the alternative $H_a : \mu_1 > \mu_2$, the decision rule is

$$\text{reject } H_0 : \mu_1 = \mu_2 \text{ if } \bar{x}_1 - \bar{x}_2 > t_{n_1+n_2-2;1-\alpha}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

and the associated p-value is

$$p = 1 - \mathbb{P}(T < t_0),$$

whereas for the alternative $H_a : \mu_1 < \mu_2$, the decision rule is

$$\text{reject } H_0 \text{ if } \bar{x}_1 - \bar{x}_2 < -t_{n_1+n_2-2;1-\alpha}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

and the associated p-value is

$$p = \mathbb{P}(T < t_0)$$

Confidence intervals for $\mu_1 - \mu_2$ (with $\sigma_1^2 = \sigma_2^2$)

In the same framework, a $100 \times (1-\alpha)\%$ two-sided confidence interval for $\mu_1 - \mu_2$ is

$$\left[(\bar{x}_1 - \bar{x}_2) - t_{n_1+n_2-2;1-\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \right.$$
$$\left. (\bar{x}_1 - \bar{x}_2) + t_{n_1+n_2-2;1-\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right]$$

while two $100 \times (1-\alpha)\%$ one-sided confidence intervals for $\mu_1 - \mu_2$ are

$$\left(-\infty, (\bar{x}_1 - \bar{x}_2) + t_{n_1+n_2-2;1-\alpha} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right]$$

and

$$\left[(\bar{x}_1 - \bar{x}_2) - t_{n_1+n_2-2;1-\alpha} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, +\infty \right)$$

Confidence intervals for $\mu_1 - \mu_2$ (with $\sigma_1^2 = \sigma_2^2$)

Example

Two catalysts are being analysed to determine how they affect the mean yield of a chemical process. Catalyst 1 is currently in use, but catalyst 2 is acceptable and cheaper so that it could be adopted providing it does not change the process yield. A test is run, see data below. Is there any difference between the mean yields? Use $\alpha = 0.05$ and assume equal variances. (**Matlab outputs:** `tinv`$(0.975, 14) = 2.145$, `tcdf`$(0.35, 14) = 0.635$)

Data: Catalyst 1: $n_1 = 8$,

$$(89.19,\ 90.95,\ 90.46,\ 93.21,\ 97.19,\ 97.04,\ 91.07,\ 92.75),$$

$\to \bar{x}_1 = 92.733$, $s_1 = 2.98$

Catalyst 2: $n_2 = 8$,

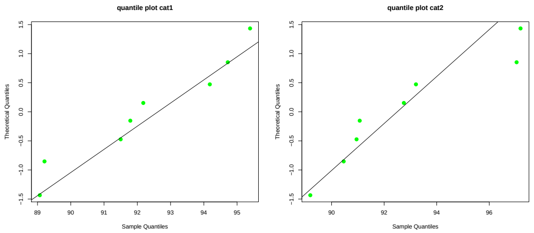
$$(91.50,\ 94.18,\ 92.18,\ 95.39,\ 91.79,\ 89.07,\ 94.72,\ 89.21),$$

$\to \bar{x}_2 = 92.255$, $s_2 = 2.39$

The parameters of interest are μ_1 and μ_2, the mean process yields using catalysts 1 and 2, respectively.

We would like to test: $H_0 : \mu_1 = \mu_2$, against $H_a : \mu_1 \neq \mu_2$

The qq-plots for the two samples do not show strong departure from normality → two-sample t-test, assuming $\sigma_1^2 = \sigma_2^2$ (for $s_1 \simeq s_2$)



With the data we have, we easily find the observed pooled standard deviation

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} = \sqrt{\frac{7 \times 2.39^2 + 7 \times 2.98^2}{8+8-2}} = 2.70$$

Now, with $n_1 + n_2 - 2 = 14$ degrees of freedom, the quantile of level 0.975 of the Student's t distribution is $t_{14;0.975} = 2.145$ (hint)

→ the rejection criterion is thus

reject $H_0 : \mu_1 = \mu_2$ if

$$\bar{x}_1 - \bar{x}_2 \notin \left[-2.145 \times 2.70 \times \sqrt{\frac{1}{8}+\frac{1}{8}}, 2.145 \times 2.70 \times \sqrt{\frac{1}{8}+\frac{1}{8}}\right]$$

$$= [-2.895, 2.895]$$

Here, $\bar{x}_1 - \bar{x}_2 = 0.478$

→ do not reject H_0!

Conclusion: at the 0.05 level of significance, we do not have enough evidence to conclude that catalyst 2 results in a mean yield that differs from the mean yield when catalyst 1 is used

→ cheaper catalyst 2 can be used without (significantly) affecting the mean process yield

The associated p-value can be found from

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{0.478}{2.70\sqrt{\frac{1}{8}+\frac{1}{8}}} = 0.35,$$

so that

$$p = 2 \times \mathbb{P}(T > 0.35) = 2 \times (1 - 0.635) = 0.73,$$

for $T \sim t_{14}$

→ do not reject H_0

A 95% confidence interval can also be derived for $\mu_1 - \mu_2$:

$$\left[0.478 - 2.145 \times 2.70 \times \sqrt{\frac{1}{8}+\frac{1}{8}}, 0.478 + 2.145 \times 2.70 \times \sqrt{\frac{1}{8}+\frac{1}{8}}\right]$$

$$= [-2.418, 3.374]$$

Of course, 0 belongs to this interval of plausible values for $\mu_1 - \mu_2$.

Hypothesis test for $\mu_1 = \mu_2$ (when $\sigma_1^2 \neq \sigma_2^2$)

In some situations, we cannot reasonably assume that the unknown variances σ_1^2 and σ_2^2 are equal

→ S_1^2 has to be used as an estimator for σ_1^2 and S_2^2 has to be used as an estimator for σ_2^2

There is no exact result available for testing $H_0 : \mu_1 = \mu_2$ in this case.

However, an approximate result can be applied:

$$\boxed{\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \overset{a}{\sim} t_\nu}$$

where the number of degrees of freedom is

$$\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$

(rounded down to the nearest integer)

Hypothesis test for $\mu_1 = \mu_2$ (when $\sigma_1^2 \neq \sigma_2^2$)

From there, the hypotheses/confidence intervals on the difference in means of two populations are tested/derived as ‘usual’, with $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ as estimated standard error, and this value of ν for the number of degrees of freedom of the t-distribution

→ this is called Welch-Satterthwaite’s approximate two-sample t-test

Remark 1: Again, if the sample sizes are ‘large’ (usually both $n_1 > 40$ and $n_2 > 40$), the test statistic has approximate standard normal distribution, and the rejection criterion and p-value can be computed by reference of the $\mathcal{N}(0,1)$-distribution (no real need for computing ν then)

Remark 2: the hypothesis of equality of variances $\sigma_1^2 = \sigma_2^2$ can be formally tested. The hypotheses would be

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{against} \quad H_a : \sigma_1^2 \neq \sigma_2^2$$

This test is beyond the scope of this course.

Hypothesis test for $\mu_1 = \mu_2$ (when $\sigma_1^2 \neq \sigma_2^2$): example

Example

The void volume within a textile fabric affects comfort, flammability, and insulation properties. Permeability of a fabric refers to the accessibility of void space to the flow of a gas or liquid. We have summary information on air permeability (in cm^3/cm^2/sec) for two different types of plain weave fabric (see below). Assuming the permeability distributions for both types of fabric are normal, calculate a 95% confidence interval for the difference between true average permeability for the cotton fabric and that for the acetate fabric. (**Hint:** You can use the following Matlab output: $\texttt{tinv}(0.975, 9) = 2.262$)

Fabric type	Sample size	Sample mean	Sample standard deviation
Cotton	10	51.71	0.79
Acetate	10	136.14	3.59

Here we have $s_1 = 0.79 \ll s_2 = 3.59$, so it would not be wise to assume $\sigma_1^2 = \sigma_2^2$!

→ Welch-Satterthwaite’s approximate two-sample t-test

Hypothesis test for $\mu_1 = \mu_2$ (when $\sigma_1^2 \neq \sigma_2^2$): example

First the right number of degrees of freedom must be determined:

$$\nu = \frac{(0.79^2/10 + 3.59^2/10)^2}{\frac{(0.79^2/10)^2}{9} + \frac{(3.59^2/10)^2}{9}} = 9.87$$

→ use $\nu = 9$ degrees of freedom

From the hint we know $t_{9;0.975} = 2.262$, so a 95% confidence interval is

$$\left[\bar{x}_1 - \bar{x}_2 \pm t_{\nu;1-\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right] = \left[51.71 - 136.14 \pm 2.262 \times \sqrt{\frac{0.79^2}{10} + \frac{3.59^2}{10}}\right]$$

$$= [-87.06, -81.80]$$

→ we can be 95% confident that the true average permeability for acetate fabric exceeds that for cotton by between 81.80 and 87.06 cm^3/cm^2/sec

Paired observations

In the application of the two-sample t-test we need to be certain the two populations (and thus the two random samples) are independent

→ this test cannot be used when we deal with “before and after” data, the ages of husbands and wives, and numerous situations where the data are naturally paired (and thus, not independent!)

Let $(X_{11}, X_{21}), (X_{12}, X_{22}), \ldots, (X_{n1}, X_{n2})$ be a random sample of n pairs of observations drawn from two subpopulations X_1 and X_2, with respective means μ_1 and μ_2.

Because X_{i1} and X_{i2} share some common information, they are certainly not independent, but they can be represented as

$$X_{i1} = W_i + Y_{i1}, \qquad X_{i2} = W_i + Y_{i2},$$

where W_i is the common random variable representing the ith pair, and Y_{i1}, Y_{i2} are the particular independent contributions of the first and second observation of the pair.

Paired observations

An easy way to get rid of the ‘dependence’ implied by W_i is just to consider the **differences**

$$D_i = X_{i1} - X_{i2} = (W_i + Y_{i1}) - (W_i + Y_{i2}) = Y_{i1} - Y_{i2}$$

→ we have just a sample of independent observations $D_1, D_2, \ldots, D_n$, one for each pair, drawn from a distribution with mean

$$\mu_D = \mu_1 - \mu_2$$

→ testing for $H_0 : \mu_1 = \mu_2$ is exactly equivalent to testing for

$$H_0 : \mu_D = 0$$

This can be accomplished by performing the usual one-sample t-test (or a large-sample test) on μ_D, from the observed sample of differences.

Note: the test will be performed on the sample of differences only
→ check if the population of differences is normal or not (the initial distributions of X_1 and X_2 do no matter)

Paired observations: example

Example

Below are the average weekly losses of worker-hours due to accidents in 10 industrial plants before and after a certain safety program was put into operation. Use a hypothesis test at significance level $\alpha = 0.05$ to check whether the safety program is effective (**Hint:** You can use the following Matlab outputs: $\texttt{tinv}(0.95, 9) = 1.833$, $\texttt{tcdf}(3.347, 9) = 0.9957$)

Data:

Plant	1	2	3	4	5	6	7	8	9	10
Before (sample 1)	47	73	46	124	33	58	83	32	26	15
After (sample 2)	36	60	44	119	35	51	77	29	26	11

Define μ_1 the true mean weekly loss before the safety program was put into operation, and μ_2 the true mean weekly loss after the safety program was put into operation. We would like to test:

$$H_0 : \mu_1 = \mu_2 \qquad \text{(no effect of the safety program)}$$

against

$$H_a : \mu_1 > \mu_2 \qquad \text{(effectiveness of the safety program)}$$

Paired observations: example

We cannot apply an independent two-sample t-test because the ‘before’ and ‘after’ weekly losses of worker-hours **in the same industrial plant** are certainly not independent.

(regardless of the safety program, some plants may be prone to frequent accidents, some others may not be)

→ pairing of the observations

→ the sample of differences is:

$$11,\ 13,\ 2,\ 5,\ -2,\ 7,\ 6,\ 3,\ 0,\ 4$$

with a sample mean $\bar{d} = 4.9$ and a sample standard deviation $s = 4.6296$

We can check that it is plausible that the sample of differences comes from a normal population (quantile plot).

→ one-sample t-test for $H_0 : \mu_D = 0$ against $H_a : \mu_D > 0$

Paired observations: example

Here $n = 10$ and $t_{9;0.95} = 1.833$ → rejection criterion:

$$\text{reject } H_0 \text{ if } \bar{d} > t_{n-1;1-\alpha}\frac{s}{\sqrt{n}} = 1.833 \times \frac{4.6296}{\sqrt{10}} = 2.684$$

→ we reject H_0

The p-value is computed from

$$t_0 = \sqrt{n}\,\frac{\bar{d}}{s} = \sqrt{10}\,\frac{4.9}{4.6296} = 3.347$$

→ $p = \mathbb{P}(T > 3.347)$ for $T \sim t_9$

From the hint, $p = 1 - 0.9957 = 0.0043$

→ clear rejection of H_0

→ Conclusion: the data shows evidence that the safety program put into operation is indeed effective

Objectives

Now you should be able to:

- test hypotheses on a population proportion □
- test hypotheses and construct confidence intervals on the variance of a normal population □
- structure comparative experiments involving two samples as hypothesis tests □
- test hypotheses and construct confidence intervals on the difference in means of two independent populations □
- test hypotheses and construct confidence intervals on the difference in means of two paired (sub)populations □

Recommended exercises:

→ Q25(a-b) p.311, Q31 p.312, Q47, Q49 p.326, Q51(b), Q53 p.327, Q75 p.341, Q25, Q27 p.368, Q29, Q31 p.369, Q35, Q37 p.370 (2nd edition)

→ Q27(a-b) p.316, Q33 p.317, Q52 p.332, Q54(b), Q56 p.333, Q79 p.348, Q27, Q29 p.376, Q32, Q34 p.377, Q38 p.378, Q40 p.379 (3rd edition)

10 Regression Analysis

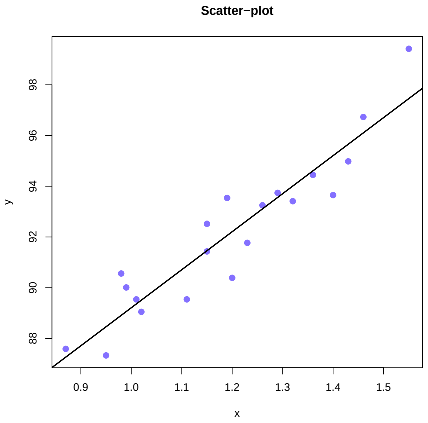
Introduction

- The main objective of many statistical investigations is to **make predictions**, preferably on the basis of mathematical equations
- For instance, an engineer may wish to predict the amount of oxide that will form on the surface of a metal baked in an oven for one hour at 200°C, or the amount of deformation of a ring subjected to a certain compressive force, or the number of miles to wear out a tire as a function of tread thickness and composition
- Usually, such predictions require that a **formula** be found which relates the dependent variable whose value we want to predict (usually it is called the **response**) to one or more other variables, usually called **predictors** (or regressors)
- The collection of statistical tools that are used to model and explore relationships between variables that are related is called **regression analysis**, and is one of the **most widely used statistical techniques**

Introduction

As an illustration, consider the following data, where y_i's are the observed purity of oxygen produced in a chemical distillation process, and x_i's are the observed corresponding percentage of hydrocarbons that are present in the main condenser of the distillation unit

i	x_i (%)	y_i (%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.54
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33



Simple linear regression model

- Inspection of the scatter-plot indicates that the points lie scattered randomly around a straight line (although no straight line will pass exactly through all the points)
- Therefore, it is reasonable to assume that the random variables X (hydrocarbon concentration) and Y (oxygen purity) are linearly related, which can be formalised by the **regression model**

$$\boxed{Y = \beta_0 + \beta_1 X + \varepsilon}$$

- The slope β_1 and the intercept β_0 are the regression coefficients
- The term ε is the random error, whose presence accounts for the fact that observed values for Y do not fall exactly on a straight line
- This model is called the **simple linear regression model**
- Sometimes a model arises from a theoretical relationship, at other times the choice of the model is based on inspection of a scatterplot

Simple linear regression model

- The random error term ε is a random variable whose properties will determine the properties of the response Y
- Assume that $\mathbb{E}(\varepsilon) = 0$ and $\mathbb{V}\text{ar}(\varepsilon) = \sigma^2$
- Suppose we fix $X = x$. At this very value of X, Y is the random variable

$$Y = \beta_0 + \beta_1 x + \varepsilon,$$

with mean $\beta_0 + \beta_1 x$ and variance $\mathbb{V}\text{ar}(\varepsilon) = \sigma^2$

→ the linear function $\beta_0 + \beta_1 x$ is thus the function giving the mean value of Y for each possible value x of X

- It is called the **regression function** (or regression line) and will be denoted

$$\mu_{Y|X=x} = \beta_0 + \beta_1 x$$

→ the slope β_1 is the change in mean of Y for one unit change in X, the intercept β_0 is the mean value of Y when $X = 0$

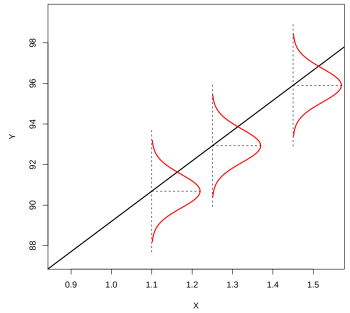
Simple linear regression model

Most of the time, the random error is supposed to be normally distributed: $\boxed{\varepsilon \sim \mathcal{N}(0, \sigma)}$ (recall Gauss, Slide 206)

It follows that, for any fixed value x for X,

$$\boxed{Y|(X = x) \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma)}$$

→ the standard deviation σ tells to which extent the observations deviate from the regression line



Note: we recognise the notation | , which means "conditionally on", as in conditional probabilities (Slide 100). Here we understand: "if we know that X takes the value x, then the distribution of Y is $\mathcal{N}(\beta_0 + \beta_1 x, \sigma)$"

Simple linear regression model

In most real-world problems, the values of the intercept β_0, the slope β_1 and the standard deviation of the error σ will not be known.

→ they are **population parameters** which must be estimated from sample data

Here the random sample consists of n pairs of observations (X_i, Y_i), assumed to be independent of one another and such that

$$Y_i | (X_i = x) \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma)$$

for all $i = 1, \dots, n$.

The straight line $\mu_{Y|X=x} = \beta_0 + \beta_1 x$ can be regarded as the population regression line, which must be estimated by a sample version

$$\hat{\mu}_{Y|X=x} = \hat{\beta}_0 + \hat{\beta}_1 x$$

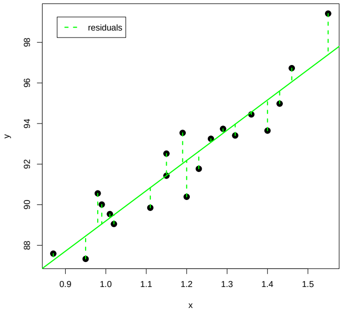
The question is how to determine the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ (and then an estimator for σ).

Least Squares Estimators

The estimates of β_0 and β_1 should result in a line that is (in some sense) a “best fit” to the data.

Gauss (again) proposed estimating the parameters β_0 and β_1 to minimise the sum of the squares of the vertical deviations between the observed responses and the fitted straight line.

These deviations are often called the **residuals** of the model, and the resulting estimators of β_0 and β_1 are the **least squares estimators**.



Least Squares Estimators

For any “candidate” straight line $Y = \hat{\beta}_0 + \hat{\beta}_1 X$, write $R(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^n (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2$

Then,

$$\frac{\partial R}{\partial \hat{\beta}_0}(\hat{\beta}_0, \hat{\beta}_1) = -2 \sum_{i=1}^n (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))$$

$$\frac{\partial R}{\partial \hat{\beta}_1}(\hat{\beta}_0, \hat{\beta}_1) = -2 \sum_{i=1}^n (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))X_i$$

→ the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ should be the solutions of the equations

$$\begin{cases} \sum_i (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)) = 0 \\ \sum_i (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))X_i = 0 \end{cases}$$

which are

$$\hat{\beta}_1 = \frac{\sum_i X_i Y_i - \frac{(\sum_i X_i)(\sum_i Y_i)}{n}}{\sum_i X_i^2 - \frac{(\sum_i X_i)^2}{n}} \quad \text{and} \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$

Least Squares Estimators

Introducing the notation

$$S_{XX} = \sum_{i=1}^n (X_i - \bar{X})^2 \qquad \left(= \sum_{i=1}^n X_i^2 - \frac{(\sum_i X_i)^2}{n} \right)$$

$$S_{XY} = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) \qquad \left(= \sum_{i=1}^n X_i Y_i - \frac{(\sum_i X_i)(\sum_i Y_i)}{n} \right)$$

we have:

Least squares estimators of β_0 and β_1

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} \quad \text{and} \quad \hat{\beta}_0 = \bar{Y} - \frac{S_{XY}}{S_{XX}} \bar{X}$$

Note: as $\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}$, the estimated straight line will always go through the point $(\bar{x}, \bar{y})$, the centre of gravity of the scatter-plot

Least Squares Estimates

Once we have observed a sample $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, we have directly the observed values

$$s_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad \text{and } s_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

and thus the estimates $\hat{b}_1$ and $\hat{b}_0$ of β_1 and β_0:

$$\hat{b}_1 = \frac{s_{xy}}{s_{xx}} \quad \text{and} \quad \hat{b}_0 = \bar{y} - \frac{s_{xy}}{s_{xx}} \bar{x}$$

The estimated or fitted regression line is therefore $\hat{b}_0 + \hat{b}_1 x$, which is an estimate of $\mu_{Y|X=x}$

Now, $\hat{b}_0 + \hat{b}_1 x$ is also the best prediction we can make of a future observation of Y when X is set to x, so it is often denoted $\hat{y}(x)$:

$$\boxed{\hat{y}(x) = \hat{b}_0 + \hat{b}_1 x}$$

Least Squares Estimation: example

Example

Fit a simple linear regression model to the data shown on Slide 428.

From the observed data, the following quantities may be computed:

$$n = 20, \qquad \sum x_i = 23.92, \qquad \sum y_i = 1,843.21$$

$$\bar{x} = 1.1960, \qquad \bar{y} = 92.1605$$

$$\sum x_i^2 = 29.2892, \qquad \sum x_i y_i = 2,214.6566$$

$$s_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = 29.2892 - \frac{23.92^2}{20} = 0.68088$$

$$s_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} = 2,214.6566 - \frac{23.92 \times 1,843.21}{20} = 10.17744$$

Least Squares Estimation: example

Therefore, the least squares estimates of the slope and the intercept are

$$\hat{b}_1 = \frac{s_{xy}}{s_{xx}} = \frac{10.17744}{0.68088} = 14.94748$$

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x} = 92.1605 - 14.94748 \times 1.196 = 74.28331$$

→ the fitted simple linear regression model is thus

$$\hat{y}(x) = 74.283 + 14.947x$$

which is the straight line shown on Slides 428 and 433

Using this model, we would predict a mean oxygen purity of 89.23% when the hydrocarbon level is $x = 1\%$.

Also, the model indicates that the mean oxygen purity would increase by 14.947% for each unit increase (1%) in hydrocarbon level.

Estimating σ^2

The variance σ^2 of the error term $\varepsilon = Y - (\beta_0 + \beta_1 X)$ is another unknown parameter

→ the residuals of the fitted model, i.e.

$$\hat{e}_i = y_i - (\hat{b}_0 + \hat{b}_1 x_i) = y_i - \hat{y}(x_i), \qquad i = 1, 2, \dots, n$$

can be regarded as a ‘sample’ drawn from the distribution of ε

→ a natural estimator for σ^2 is the sample variance of the residuals

First, it can be checked that

$$\bar{e} = \frac{1}{n} \sum_{i=1}^{n} \hat{e}_i = \bar{y} - (\hat{b}_0 + \hat{b}_1 \bar{x}) = 0$$

(by definition of the estimated coefficient $\hat{b}_0$ and $\hat{b}_1$)

Estimating σ^2

Also, recall that the number of degrees of freedom for the usual sample variance is $n-1$ because we have to estimate one parameter ($\bar{x}$ estimates the true μ)

Here we have to first estimate two parameters (β_0 and β_1)

→ the number of degrees of freedom must now be $n-2$

→ an **unbiased** estimate of σ^2 is $\boxed{s^2 = \frac{1}{n-2}\sum_{i=1}^{n} \hat{e}_i^2}$

which is the observed value taken by the estimator

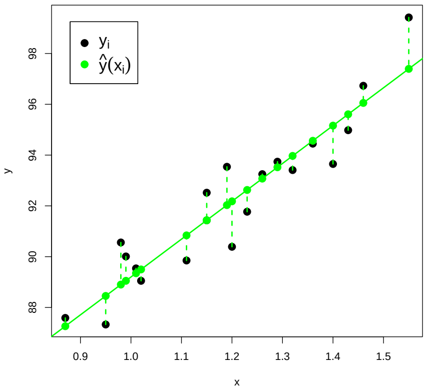
$$S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

It is clear that S is an estimator for σ.

Estimating σ^2: example

In the previous example, we fitted $\hat{y}(x) = 74.283 + 14.947x$, so that we get a series of fitted values $\hat{y}(x_i) = 74.283 + 14.947x_i$, for $i = 1, \ldots, 20$, from which the residuals can be computed: $\hat{e}_i = y_i - \hat{y}(x_i)$, for $i = 1, \ldots, 20$

i	x_i	y_i	$\hat{y}(x_i)$	$\hat{e}_i$
1	0.99	90.01	89.051	0.959
2	1.02	89.05	89.498	-0.448
3	1.15	91.43	91.435	-0.005
4	1.29	93.74	93.521	0.219
5	1.46	96.73	96.054	0.676
6	1.36	94.45	94.564	-0.114
7	0.87	87.59	87.263	0.327
8	1.23	91.77	92.627	-0.857
9	1.55	99.42	97.395	2.025
10	1.40	93.65	95.160	-1.510
11	1.19	93.54	92.031	1.509
12	1.15	92.52	91.435	1.085
13	0.98	90.56	88.902	1.658
14	1.01	89.54	89.349	0.191
15	1.11	89.85	90.839	-0.989
16	1.20	90.39	92.180	-1.790
17	1.26	93.25	93.074	0.176
18	1.32	93.41	93.968	-0.558
19	1.43	94.98	95.607	-0.627
20	0.95	87.33	88.455	-1.125



We find: $s^2 = \frac{1}{18}\sum_{i=1}^{20} \hat{e}_i^2 = 1.1824$ (%^{2}) → $s = \sqrt{1.1824} = 1.0874$ (%)

Fixed design

From now on we will assume that the value of the x_i's have been chosen before the experiment is performed, and are therefore fixed

→ this is known as a fixed design

So, only the Y_i's are random, and that substantially simplifies the coming developments, in particular the derivation of the sampling properties of the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$.

Properties of the Least Squares Estimators

We noted that $Y_i | (X_i = x_i) \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma)$

Then, because $\sum_i (x_i - \bar{x}) = 0$, we can write

$$\hat{\beta}_1 = \frac{S_{xY}}{s_{xx}} = \sum_i \frac{(x_i - \bar{x})}{s_{xx}} Y_i$$

→ which is a linear combination of the normal random variables Y_i, therefore $\hat{\beta}_1$ is normally distributed!

Its expectation is

$$\mathbb{E}(\hat{\beta}_1) = \frac{\sum_i (x_i - \bar{x})\mathbb{E}(Y_i)}{s_{xx}} = \frac{\sum_i (x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{s_{xx}} = \frac{\beta_1 \sum_i x_i (x_i - \bar{x})}{s_{xx}} = \beta_1$$

→ **unbiased** estimator of β_1

Similarly, its variance is $\mathbb{V}\text{ar}(\hat{\beta}_1) = \frac{\sum_i (x_i - \bar{x})^2\, \mathbb{V}\text{ar}(Y_i)}{s_{xx}^2} = \frac{\sigma^2 \sum_i (x_i - \bar{x})^2}{s_{xx}^2} = \frac{\sigma^2}{s_{xx}}$

Hence, the sampling distribution of $\hat{\beta}_1$ is

$$\boxed{\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma}{\sqrt{s_{xx}}}\right)}$$

Properties of the Least Squares Estimators

Now, we can write

$$\hat{\beta}_0 = \sum_{i=1}^{n} \frac{Y_i}{n} - \hat{\beta}_1 \bar{x},$$

which is again a linear combination of the Y_i's

→ the estimator $\hat{\beta}_0$ is also normally distributed! Its expectation is

$$\mathbb{E}(\hat{\beta}_0) = \sum_{i=1}^{n} \frac{\mathbb{E}(Y_i)}{n} - \mathbb{E}(\hat{\beta}_1)\bar{x} = \sum_{i=1}^{n} \frac{\beta_0 + \beta_1 x_i}{n} - \beta_1 \bar{x} = \beta_0$$

→ **unbiased** estimator of β_0

Similarly, we could find $\mathbb{V}\text{ar}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)$

Hence, the sampling distribution of $\hat{\beta}_0$ is

$$\boxed{\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}}\right)}$$

Inferences concerning β_1

An important hypothesis to consider regarding the simple linear regression model $Y = \beta_0 + \beta_1 X + \varepsilon$ is the hypothesis that $\beta_1 = 0$

→ $\beta_1 = 0$ is equivalent to stating that the response does not depend on the predictor X (as we would have $Y = \beta_0 + \varepsilon$)

We can set up a formal hypothesis test. The appropriate hypotheses are:

$$H_0 : \beta_1 = 0 \qquad \text{against} \qquad H_a : \beta_1 \neq 0$$

→ we reject H_0 when the estimate $\hat{b}_1$ is 'too different' to 0

From the sampling distribution of $\hat{\beta}_1$, we get $\sqrt{s_{xx}}\,\frac{\hat{\beta}_1 - \beta_1}{\sigma} \sim \mathcal{N}(0, 1)$

However, σ is typically unknown → replace it with its estimator S

As this estimator of σ has $n - 2$ degrees of freedom (Slide 440), we find:

$$\boxed{\sqrt{s_{xx}}\,\frac{\hat{\beta}_1 - \beta_1}{S} \sim t_{n-2}}$$

Inferences concerning β_1

From this result, all the inferential procedures that we introduced previously can be readily adapted

At significance level α, the rejection criterion for $H_0 : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$ is

$$\text{reject } H_0 \text{ if } \hat{b}_1 \notin \left[-t_{n-2,1-\alpha/2}\frac{s}{\sqrt{s_{xx}}}, t_{n-2,1-\alpha/2}\frac{s}{\sqrt{s_{xx}}}\right],$$

with the estimated standard deviation $s = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n} \hat{e}_i^2}$ (Slide 439)

and from the observed value of the test statistic under H_0 (i.e. with $\beta_1 = 0$)

$$t_0 = \sqrt{s_{xx}}\,\frac{\hat{b}_1}{s}$$

we can compute the p-value (Slide 360)

$$p = 1 - \mathbb{P}(T \in [-|t_0|, |t_0|]) = 2 \times \mathbb{P}(T > |t_0|)$$

where T is a r. v. with distribution t_{n-2}

Inferences concerning β_1

In addition to the point estimate $\hat{b}_1$ of the slope, it is also possible to obtain a confidence interval for the 'true' slope β_1.

As $\sqrt{s_{xx}}\,\frac{\hat{\beta}_1 - \beta_1}{S} \sim t_{n-2}$, we can directly write

$$\mathbb{P}\left(-t_{n-2;1-\alpha/2} \leq \sqrt{s_{xx}}\,\frac{\hat{\beta}_1 - \beta_1}{S} \leq t_{n-2;1-\alpha/2}\right) = 1 - \alpha$$

or equivalently

$$\mathbb{P}\left(\hat{\beta}_1 - t_{n-2;1-\alpha/2}\frac{S}{\sqrt{s_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{n-2;1-\alpha/2}\frac{S}{\sqrt{s_{xx}}}\right) = 1 - \alpha$$

From an observed sample for which we find s and $\hat{b}_1$, a two-sided $100 \times (1 - \alpha)$% confidence interval for the parameter β_1 is

$$\boxed{\left[\hat{b}_1 - t_{n-2;1-\alpha/2}\frac{s}{\sqrt{s_{xx}}}, \hat{b}_1 + t_{n-2;1-\alpha/2}\frac{s}{\sqrt{s_{xx}}}\right]}$$

Inferences concerning β_0

Although of less practical interest, inferences concerning the parameter β_0 can be made in the exact same way from the sampling distribution of $\hat{\beta}_0$.

We find a two-sided $100 \times (1-\alpha)\%$ confidence interval for β_0

$$\boxed{\left[\hat{b}_0 - t_{n-2;1-\alpha/2}\, s\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{s_{xx}}}, \hat{b}_0 + t_{n-2;1-\alpha/2}\, s\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{s_{xx}}}\right]}$$

as well as a rejection criterion for a hypothesis $H_0 : \beta_0 = 0$ (no intercept in the model) tested against $H_a : \beta_0 \neq 0$, at level α,

$$\text{reject } H_0 \text{ if } \hat{b}_0 \notin \left[-t_{n-2,1-\alpha/2}\, s\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{s_{xx}}}, t_{n-2,1-\alpha/2}\, s\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{s_{xx}}}\right]$$

with a p-value calculated from the observed value of the test statistic

$$t_0 = \frac{\hat{b}_0}{s\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{s_{xx}}}} \qquad \to p = 2 \times \mathbb{P}(T > |t_0|), \quad T \sim t_{n-2}$$

Inferences concerning β_1: example

Example

Test for significance of the simple linear regression model for the data shown on Slide 428 at level $\alpha = 0.01$. (**Hint:** You can use the following Matlab outputs: `tinv`$(0.995, 18) = 2.878$, `tcdf`$(11.35, 18) = 1$)

The model is $Y = \beta_0 + \beta_1 X + \varepsilon$. Testing for the significance of the model amounts to considering the hypotheses:

$$H_0 : \beta_1 = 0 \qquad \text{against} \qquad H_a : \beta_1 \neq 0$$

The estimate of β_1 is $\hat{b}_1 = 14.947$. Also, we previously found $n = 20$, $s_{xx} = 0.68088$ and $s = 1.0874$. Hence, at significance level $\alpha = 0.01$, the rejection criterion is:

$$\text{reject } H_0 \text{ if } \hat{b}_1 \notin \left[-2.878 \times \frac{1.0874}{\sqrt{0.68088}}, 2.878 \times \frac{1.0874}{\sqrt{0.68088}}\right] = [-3.793, 3.793]$$

Here, with $\hat{b}_1 = 14.947$, we clearly reject H_0

→ the 'true' slope β_1 between oxygen purity and hydrocarbon level is certainly different from 0 → **hydrocarbon level does influence oxygen purity**

Inferences concerning β_1: example

The observed value of the test statistic is

$$t_0 = \sqrt{s_{xx}}\,\frac{\hat{b}_1}{s} = \sqrt{0.68088} \times \frac{14.947}{1.0874} = 11.35$$

and the p-value is $p = 2 \times \mathbb{P}(T > 11.35) \simeq 0$ (with $T \sim t_{18}$)

We can also derive a 99% confidence interval for β_1

As $t_{18,0.995} = 2.878$, we get:

$$\left[14.947 \pm 2.878 \times \frac{1.0874}{\sqrt{0.68088}}\right] = [11.181, 18.767]$$

→ we can be 99% confident that the true value of the slope β_1 lies between 11.181 and 18.767 (so that 0 is obviously not one of the plausible values for β_1)

Simple linear regression: computer output

All statistical software programs include a least squares fit of a straight line

A typical output is as follows:

```
Regression Analysis:  Y versus X

The regression equation is Y = 74.283 + 14.947 X

 Predictor     Coef   SE Coef      T      P
 Constant    74.283     1.593  46.62  0.000
 X           14.947     1.317  11.35  0.000

S = 1.087 R-Sq = 87.74% R-Sq(adj) = 87.06%
```

The first row of the table (`Constant`) refers to the intercept (β_0), the second (`X`) to the predictor X (β_1).

The column `Coef` is for the estimates of the coefficients ($\hat{b}_0$ and $\hat{b}_1$), the column `SE Coef` is for the (estimated) standard error of these estimates ($s\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{s_{xx}}}$ and $\frac{s}{\sqrt{s_{xx}}}$), the column `T` is for the observed values t_0 of the test statistics (when testing $H_0 : \beta_0 = 0$ and $H_0 : \beta_1 = 0$), and the column `P` gives the associated p-values. Finally, `S` is the estimate s of σ.

Confidence Interval on the Mean Response

A confidence interval may be constructed on the mean response at a specified value of X, say, x.

This is thus a confidence interval for the unknown 'parameter'

$$\mu_{Y|X=x} = \beta_0 + \beta_1 x$$

We have an estimator for this parameter:

$$\hat{\mu}_{Y|X=x} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Note that, as a linear combination of normal random variables, the estimator $\hat{\mu}_{Y|X=x}$ is also normally distributed. Its expectation is:

$$\mathbb{E}(\hat{\mu}_{Y|X=x}) = \mathbb{E}(\hat{\beta}_0 + \hat{\beta}_1 x) = \mathbb{E}(\hat{\beta}_0) + \mathbb{E}(\hat{\beta}_1)x = \beta_0 + \beta_1 x = \mu_{Y|X=x}$$

→ **unbiased** estimator for $\mu_{Y|X=x}$

Confidence Interval on the Mean Response

Its variance can be found to be

$$\mathbb{V}\text{ar}(\hat{\mu}_{Y|X=x}) = \sigma^2 \left(\frac{1}{n} + \frac{(x-\bar{x})^2}{s_{xx}} \right)$$

Note 1: this is not $\mathbb{V}\text{ar}(\hat{\beta}_0) + \mathbb{V}\text{ar}(\hat{\beta}_1)x^2$, because $\hat{\beta}_0$ and $\hat{\beta}_1$ are not independent! Indeed, $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$

Note 2: because we know that the fitted straight line will always go through $(\bar{x}, \bar{Y})$, the variability in $\hat{\mu}_{Y|X=x}$ decreases as x approaches $\bar{x}$ and vice-versa → term $\frac{(x-\bar{x})^2}{s_{xx}}$

At $x = \bar{x}$, $\mathbb{V}\text{ar}(\hat{\mu}_{Y|X=x}) = \frac{\sigma^2}{n}$, which is just the variance of $\bar{Y}$!

Finally, the sampling distribution of the estimator $\hat{\mu}_{Y|X=x}$ is

$$\boxed{\hat{\mu}_{Y|X=x} \sim \mathcal{N}\left(\mu_{Y|X=x}, \sigma \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{s_{xx}}} \right)}$$

Confidence Interval on the Mean Response

If we standardise and replace the unknown σ by its estimator S, we get (as usual):

$$\frac{\hat{\mu}_{Y|X=x} - \mu_{Y|X=x}}{S\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{s_{xx}}}} \sim t_{n-2}$$

which directly leads to the following confidence interval for $\mu_{Y|X=x}$:

From an observed sample for which we find s and $\hat{y}(x)$ from the fitted model $\hat{y}(x) = \hat{b}_0 + \hat{b}_1 x$, a two-sided $100 \times (1-\alpha)\%$ confidence interval for the parameter $\mu_{Y|X=x}$, that is the mean response Y when $X = x$, is

$$\boxed{\left[\hat{y}(x) - t_{n-2;1-\alpha/2}\, s\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{s_{xx}}}, \hat{y}(x) + t_{n-2;1-\alpha/2}\, s\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{s_{xx}}} \right]}$$

Confidence Interval on the Mean Response: example

Example

Construct a 95% confidence interval on the mean oxygen purity $\mu_{Y|X=x}$ when the hydrocarbon level X is fixed to $x = 1$ (from the data shown on Slide 428).

The fitted model was $\hat{y}(x) = 74.283 + 14.947x$. We also have $n = 20$, $s = 1.0874$, $s_{xx} = 0.68088$ and $\bar{x} = 1.1960$. From Matlab, we find $t_{18;0.975} = 2.101$.

When $x = 1$, the model estimates the mean response $\mu_{Y|X=1}$ at $\hat{y}(1) = 89.23$

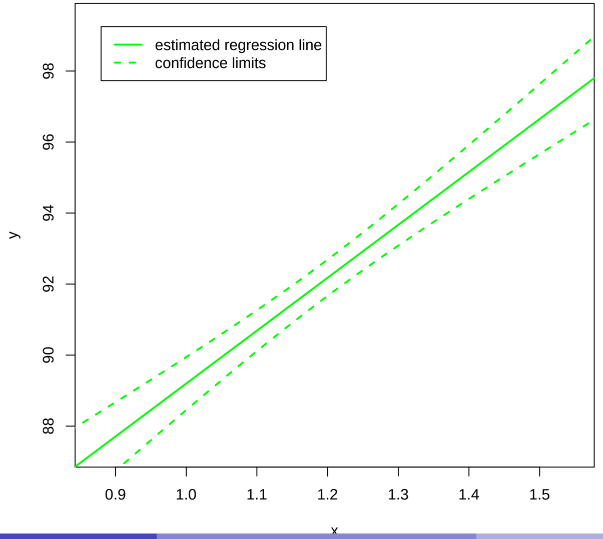
→ a 95% confidence interval for $\mu_{Y|X=1}$ is given by

$$\left[89.23 \pm 2.101 \times 1.0874 \times \sqrt{\frac{1}{20} + \frac{(1-1.1960)^2}{0.68088}} \right] = [88.48, 89.98]$$

→ when $x = 1$, we are 95% confident that the true mean oxygen purity is between 88.48 and 89.98

Confidence Interval on the Mean Response: example

By repeating these calculations for several different values for x, we can obtain confidence limits for each corresponding value of $\mu_{Y|X=x}$



Prediction of new observations

An important application of a regression model is predicting new or future observations Y corresponding to a specified level $X = x$.

$\rightarrow$ different to estimating the mean response $\mu_{Y|X=x}$ at $X = x$!

From the model, the predictor of the new value of the response Y at $X = x$, say $Y^*(x)$ is naturally given by

$$Y^*(x) = \hat{\beta}_0 + \hat{\beta}_1 x,$$

for which a predicted value is

$$\hat{y}(x) = \hat{b}_0 + \hat{b}_1 x$$

once the model has been fitted from an observed sample

$\rightarrow$ the predictor of Y at $X = x$ is the estimator of $\mu_{Y|X=x}$!

The prediction error is given by $Y|(X = x) - Y^*(x)$ and is normally distributed, as both $Y|(X = x)$ and $Y^*(x)$ are as well

Prediction of new observations

As $Y|(X = x) \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma)$ (Slide 431) and $Y^*(x) = \hat{\mu}_{Y|X=x} \sim \mathcal{N}\left(\beta_0 + \beta_1 x, \sigma\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{s_{xx}}}\right)$ (Slide 453), the expectation of the prediction error is

$$\mathbb{E}((Y|(X = x) - Y^*(x)) = \mathbb{E}(Y|X = x) - \mathbb{E}(Y^*(x)) = 0$$

$\rightarrow$ on average, the predictor will 'guess' the right value

Because the future Y is independent of the sample observations (and thus independent of $\hat{\mu}_{Y|X=x}$), the variance of the prediction error is

$$\begin{aligned}\mathbb{V}\text{ar}((Y|(X = x)) - Y^*(x)) &= \mathbb{V}\text{ar}(Y|X = x) + \mathbb{V}\text{ar}(Y^*(x)) \\ &= \sigma^2 + \sigma^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{s_{xx}}\right) = \sigma^2\left(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{s_{xx}}\right)\end{aligned}$$

and we find

$$Y|(X = x) - Y^*(x) \sim \mathcal{N}\left(0, \sigma\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{s_{xx}}}\right)$$

Prediction of new observations

Standardising and replacing the unknown σ by its estimator S, we get (as usual):

$$\frac{Y|(X = x) - Y^*(x)}{S\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{s_{xx}}}} \sim t_{n-2}$$

which directly leads to the following **prediction interval** for a new observation Y, given that $X = x$:

From an observed sample for which we find s and $\hat{y}(x)$ from the fitted model $\hat{y}(x) = \hat{b}_0 + \hat{b}_1 x$, a two-sided $100 \times (1 - \alpha)\%$ prediction interval for a new observation Y at $X = x$ is

$$\boxed{\left[\hat{y}(x) - t_{n-2;1-\alpha/2}\, s\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{s_{xx}}}, \hat{y}(x) + t_{n-2;1-\alpha/2}\, s\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{s_{xx}}}\right]}$$

Prediction of new observations: remarks

We observe:

1. a prediction interval for Y at $X = x$ will always be longer than the confidence interval for $\mu_{Y|X=x}$ because there is much more variability in one observation than in an average

   Concretely, $\mu_{Y|X=x}$ is the position of the straight line at $X = x$
   → the CI for $\mu_{Y|X=x}$ only targets that position

   However, we know that observations will not be exactly on that straight line, but ‘around’ it
   → a prediction interval for a new observation should take this extra variability into account, in addition to the uncertainty inherent in the estimation of $\mu_{Y|X=x}$

2. as n gets larger ($n \to \infty$), the width of the CI for $\mu_{Y|X=x}$ decreases to 0 (we are more and more accurate when estimating μ), but this is not the case for the prediction interval: the inherent variability in the new observation never vanishes, even when we have observed many other observations before!

Prediction of new observations: example

Example

Construct a 95% prediction interval on the oxygen purity Y when the hydrocarbon level X is fixed to $x = 1$ (from the data shown on Slide 428).

The fitted model was $\hat{y}(x) = 74.283 + 14.947x$. We also have $n = 20$, $s = 1.0874$, $s_{xx} = 0.68088$ and $\bar{x} = 1.1960$. From Matlab, we find $t_{18;0.975} = 2.101$.

When $x = 1$, the model estimates the mean response $\mu_{Y|X=1}$ to $\hat{y}(1) = 89.23$

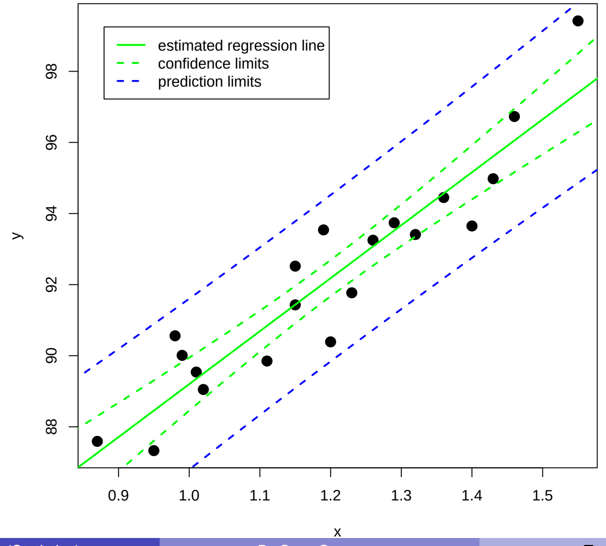
→ a 95% prediction interval for Y is given by

$$\left[89.23 \pm 2.101 \times 1.0874 \times \sqrt{1 + \frac{1}{20} + \frac{(1-1.1960)^2}{0.68088}}\right] = [86.83, 91.63]$$

→ if we fix the hydrocarbon level to $x = 1$, we can be 95% confident that the next observed value of the oxygen purity will be between 86.83 and 91.63

Prediction of new observations: example

By repeating these calculations for several different values for x, we can obtain prediction limits for each corresponding value of Y given that $X = x$



Adequacy of the regression model

In the course of fitting and analysing the simple linear regression model, we made several assumptions.

The first one is that **the model is correct**: there indeed exist coefficients β_0 and β_1, as well as a random variable ε, such that we can write $Y = \beta_0 + \beta_1 X + \varepsilon$ → **scatterplot**

The other central assumption is certainly that (Slide 432)

$$Y_i|(X_i = x_i) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma) \quad \text{for } i = 1, 2, \ldots, n,$$

which has several implications. Define the error terms

$$e_i = y_i - (\beta_0 + \beta_1 x_i), \text{ for } i = 1, \ldots, n$$

which are values drawn from the distribution of ε. We must check that:

1. the e_i’s have been drawn **independently** of one another
2. the e_i’s have the **same variance**
3. the e_i’s have been drawn from a **normal distribution**

Residual analysis

Unfortunately, we do not have access to the true e_i's (as we do not know β_0 and β_1).

However, the observed residuals of the fitted model

$$\hat{e}_i = y_i - \hat{y}(x_i) = y_i - (\hat{b}_0 + \hat{b}_1 x_i)$$

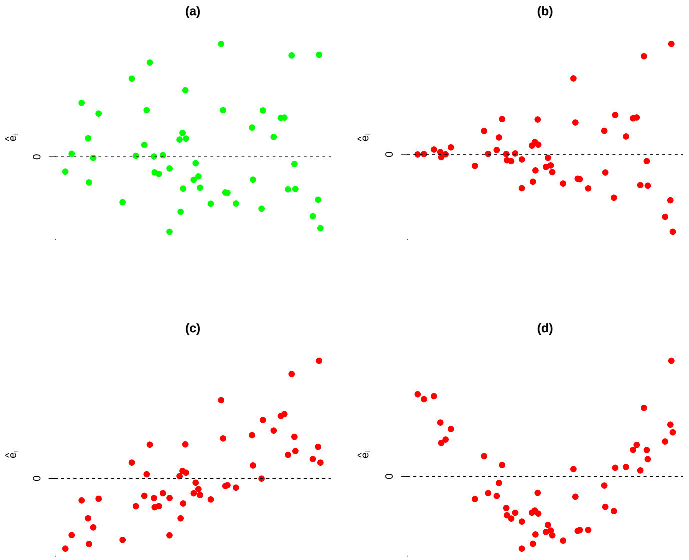
are probably good estimates of those e_i's → **residual analysis**

It is frequently helpful to plot the residuals (1) in time sequence (if known), (2) against the fitted values $\hat{y}(x_i)$, and (3) against the predictor values x_i.

Typically, these graphs will look like one of the four general patterns shown on the next slide.

As suggested by their name, the residuals are everything the model will not consider → no information should be observed in the residuals, they should look like noise.

Residual analysis

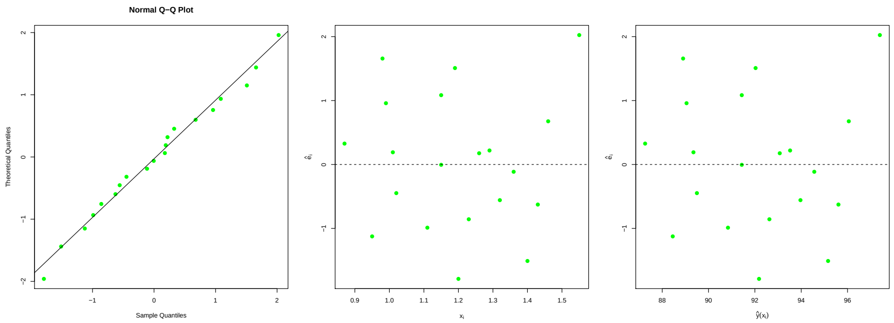


Residual analysis

- **Pattern (a) represents thus the ideal situation** (*nothing to report*, just random noise)
- In (b), the variance of the error terms e_i (and thus that of the responses Y_i) seems to be increasing with time or with magnitude of Y_i or X_i (fan shape)
- Plot (c) indicates some sort of dependence in the error terms (increasing trend)
- In (d), we get clear indication of model inadequacy: the residuals are systematically positive for extreme values and negative for medium values ⇒ the model is not complete, there is still much information in the residuals: higher-order terms (like X^2) or other predictors should be considered in the model
- Finally, a **normal probability plot (or a histogram) of residuals** is constructed so as to check the normality assumption

Residual analysis: example

From our running example (oxygen purity data), a normal quantile plot of the residuals and plots against the predicted values $\hat{y}(x_i)$ and against the hydrocarbon levels x_i for the residuals computed on Slide 441, are shown below:



→ nothing to report

→ the assumptions we made look totally valid

Variability decomposition

Similarly to the notations on Slide 435, we can define

$$s_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

→ this measures the total amount of variability in the response values, and is sometimes denoted ss_t (for '**total sum of squares**')

Now, this variability in the observed values y_i arises from two factors:

1. because the x_i values are different, all Y_i have different means. This variability is quantified by the '**regression sum of squares**':

$$ss_r = \sum_{i=1}^{n} (\hat{y}(x_i) - \bar{y})^2$$

2. each value Y_i has variance σ^2 around its mean. This variability is quantified by the '**error sum of squares**':

$$ss_e = \sum_{i=1}^{n} (y_i - \hat{y}(x_i))^2 = \sum_{i=1}^{n} \hat{e}_i^2$$

We can always write: $\boxed{ss_t = ss_r + ss_e}$

Coefficient of determination

Suppose $ss_t \simeq ss_r$ and $ss_e \simeq 0$: the variability in the responses due to the effect of the predictor is almost the total variability in the responses

→ all the dots are very close to the straight line, the predictions are very accurate: the linear regression model fits the data very well

Now suppose $ss_t \simeq ss_e$ and $ss_r \simeq 0$: almost the whole variation in the responses is due to the error terms

→ the dots are very far away from the fitted straight line, the predictions are very imprecise: the regression model is useless

→ **comparing ss_r to ss_t allows us to judge the model adequacy**

The quantity r^2, called the coefficient of determination, defined as

$$\boxed{r^2 = \frac{ss_r}{ss_t},}$$

represents the proportion of the variability in the responses that is explained by the predictor and hence taken into account in the model.

Coefficient of determination

Clearly, the coefficient of variation will have a value between 0 and 1:

- a value of r^2 near 1 indicates a good fit to the data
- a value of r^2 near 0 indicates a poor fit to the data

Fact

If the regression model is able to explain most of the variation in the response data, then it is considered to fit the data well, and is regarded as a 'good' model.

In our running example, we find in the regression output on Slide 451 a value of r^2 (`R-Sq`) is equal to 87.74%

→ almost 88% of the variation of the oxygen purity is explained by the level of hydrocarbons that was used. The remaining 12% of the variation is due to the natural variability in the oxygen purity even when the hydrocarbon level is fixed to a given level

Here r^2 is quite close to 1, which makes our model a good one.

Correlation

On Slide 166, we introduced the correlation coefficient between two random variables X and Y:

$$\rho = \frac{\mathbb{C}\text{ov}(X,Y)}{\sqrt{\mathbb{V}\text{ar}(X)\,\mathbb{V}\text{ar}(Y)}} = \frac{\mathbb{E}((X-\mathbb{E}(X))(Y-\mathbb{E}(Y))}{\sqrt{\mathbb{E}((X-\mathbb{E}(X))^2)\mathbb{E}((Y-\mathbb{E}(Y))^2)}}$$

This coefficient quantifies the strength of the linear relationship between X and Y.

→ if ρ is close to 1 or -1, there is a strong linear relationship between X and Y

→ observations in a random sample $\{(x_i, y_i), i = 1, \ldots, n\}$ drawn from the joint distribution of (X, Y) should fall close to a straight line

→ a linear regression model linking Y to X, based on that sample, should be a good model, with a value of r^2 close to 1

Correlation

We can write:

$$r^2 = \frac{ss_r}{ss_t} = \frac{ss_t - ss_e}{s_{yy}} = \frac{s_{xx}(ss_t - ss_e)}{s_{xx}s_{yy}} = \frac{s_{xy}^2}{s_{xx}s_{yy}}$$
$$= \frac{\left(\sum_i (x_i - \bar{x})(y_i - \bar{y})\right)^2}{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}$$

→ we observe that $\boxed{r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}}$

is the **sample correlation coefficient**, which can be regarded as the sample estimate of the population correlation coefficient ρ

→ except for its sign (positive or negative linear relationship), the sample correlation is the square root of the coefficient of determination (its sign is the sign of $\hat{b}_1$)

In our running example, the sample correlation coefficient is $\sqrt{0.8774} = 0.9366$ (good estimate of the 'true' correlation coefficient between hydrocarbon level and oxygen purity).

Objectives

Now you should be able to:

- Use simple linear regression for building models for engineering and scientific data □
- Understand how the method of least squares is used to estimate the regression parameters □
- Analyse residuals to determine if the regression model is an adequate fit to the data and to see if any underlying assumptions is violated □
- Test statistical hypotheses and construct confidence intervals on regression parameters □
- Use the regression model to make a prediction of a future observation and construct an appropriate prediction interval □
- Understand how the linear regression model and the correlation coefficient are related □

Recommended exercises:

→ Q7 p.104, Q13, Q15 p.114, Q21 p.126, Q1 p.499, Q5, Q8 p.500, Q13 p.507, Q17 p.508, Q19 (a-c) p.515 (2nd edition)

→ Q7 p.107, Q13 p.116, Q15 p.117, Q1 p.514, Q6 p.515, Q9 p.516, Q14 p.523, Q19 p.524, Q22 (a-d) p.531 (3rd edition)

11 ANOVA

Introduction

- In Chapter 10, we introduced testing procedures for comparing the means of two different populations, having observed two random samples drawn from those populations (two-sample z- and t-tests)
- However, in applications, it is common that we want to detect a difference in a set of **more than two populations**
- Imagine the following context: four groups of students were subjected to different teaching techniques and tested at the end of a specified period of time. Do the data shown in the table below present sufficient evidence to indicate a difference in mean achievement for the four teaching techniques?

Tech. 1	Tech. 2	Tech. 3	Tech. 4
65	75	59	94
87	69	78	89
73	83	67	80
79	81	62	88
81	72	83	
69	79	76	
	90		

Introduction: randomisation

- To answer this question, we should first note that the method of division of the students into 4 groups is of vital importance
- For instance, some basic visual inspection of the data suggests that the members of group 4 scored higher than those in the other groups. Can we conclude from this that teaching technique 4 is superior? Perhaps, students in group 4 are just better learners (regardless of the teaching technique they have been subjected to)

→ it is essential that we divide the students into 4 groups in such a way to make it very unlikely that one of the group is inherently superior to others

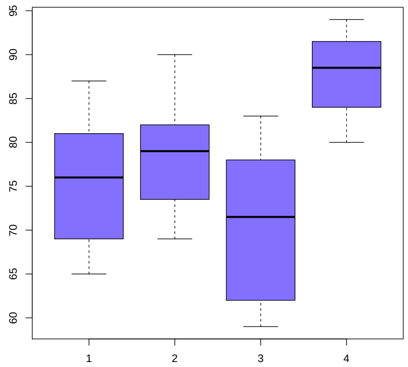
→ the only reliable method for doing this is to divide the students **in a completely random fashion**, to balance out the effect of any nuisance variable that may influence the variable of interest

- This kind of consideration is part of a very important area of statistical modelling called **experimental design**, which is not addressed in this course. Here we will always assume that the division of the individuals into the groups was indeed done "at random"

Introduction

Now, numerical summaries and a graphical display of the data are always useful:

	Tech. 1	Tech. 2	Tech. 3	Tech. 4
	65	75	59	94
	87	69	78	89
	73	83	67	80
	79	81	62	88
	81	72	83	
	69	79	76	
		90		
$\bar{x}$	75.67	78.43	70.83	87.75
s	8.17	7.11	9.58	5.80

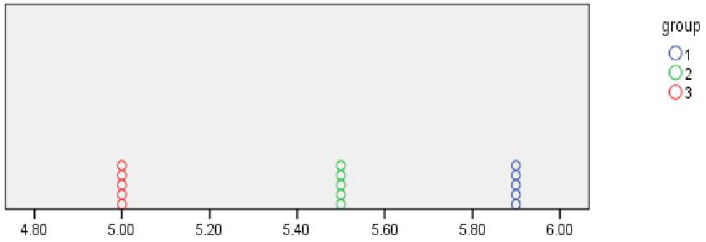


→ the boxplots show the variability of the observations within a group and the variability between the groups

→ **comparing the between-group with the within-group variability** is the key in detecting any significant difference between the groups

Between-group and within-group variability

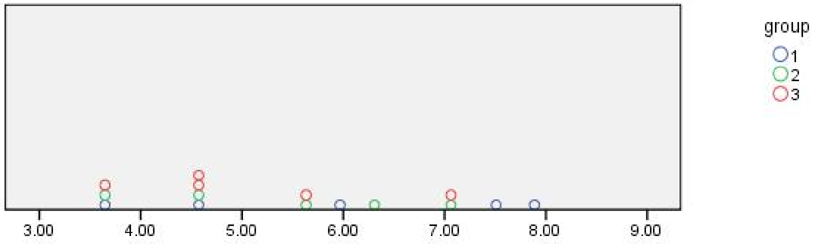
Group 1	Group 2	Group 3
5.90	5.51	5.01
5.92	5.50	5.00
5.91	5.50	4.99
5.89	5.49	4.98
5.88	5.50	5.02



Between-group variance $= 1.017$, within-group variance $= 0.00018$ (ratio $= 5545$)

Between-group and within-group variability

Group 1	Group 2	Group 3
5.90	6.31	4.52
4.42	3.54	6.93
7.51	4.73	4.48
7.89	7.20	5.55
3.78	5.72	3.52



Between-group variance = 1.017, within-group variance = 2.332 (ratio = 0.436)

Analysis of Variance

- Comparing the between-group variability and the within-group variability is the purpose of the **Analysis of Variance**

→ often shortened to the acronym **ANOVA**

- Suppose that we have k different groups (k populations, or k sub-populations of a population) that we wish to compare. Often, each group is called a treatment or treatment level (general terms that can be traced back to the early applications of this methodology in the agricultural sciences)
- The response for each of the k treatments is the random variable of interest, say X
- Denote X_{ij} the jth observation ($j = 1, \dots, n_i$) taken under treatment i

→ we have k independent samples (one sample from each of the treatments)

ANOVA samples

The k random samples are often presented as:

Treatment	1	2	...	k
	X_{11}	X_{21}	...	X_{k1}
	X_{12}	X_{22}		X_{k2}
	$\vdots$	$\vdots$		$\vdots$
	X_{1n_1}	X_{2n_2}	...	X_{kn_k}
Mean	$\bar{X}_1$	$\bar{X}_2$	...	$\bar{X}_k$
St. Dev.	S_1	S_2	...	S_k

where $\bar{X}_i$ and S_i are the sample mean and standard deviation of the ith sample. The total number of observations is

$$n = n_1 + n_2 + \ldots + n_k$$

and the **grand mean** of all the observations, usually denoted $\bar{\bar{X}}$, is

$$\bar{\bar{X}} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2 + \ldots + n_k\bar{X}_k}{n}$$

ANOVA model

The **ANOVA model** is the following:

$$\boxed{X_{ij} = \mu_i + \varepsilon_{ij}}$$

where

- μ_i is the mean response for the ith treatment ($i = 1, 2, \dots, k$)
- ε_{ij} is an individual random error component ($j = 1, 2, \dots, n_i$)

As usual for errors, we will assume that the random variables ε_{ij} are normally and independently distributed with mean 0 and variance σ^2:

$$\boxed{\varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma)} \quad \text{for all } i, j$$

Therefore, each treatment can be thought of as a normal population with mean μ_i and variance σ^2:

$$\boxed{X_{ij} \overset{\text{ind.}}{\sim} \mathcal{N}(\mu_i, \sigma)} \quad \text{for all } i, j$$

Important: the variance σ^2 is common for all treatments

ANOVA hypotheses

We are interested in detecting differences between the different treatment means μ_i, which are population parameters

→ hypothesis test!

The null hypothesis to be tested is

$$\boxed{H_0 : \mu_1 = \mu_2 = \ldots = \mu_k}$$

versus the general alternative

$$\boxed{H_a : \text{not all the means are equal}}$$

Careful! The alternative hypothesis should be that at least two of the means differ, not that they are all different !

As pointed out previously, the primary tool when testing for equality of the means is based on a comparison of the variances within the groups and between the groups.

Variability decomposition

The ANOVA partitions the total variability in the sample data, described by the **total sum of squares**

$$SS_{\text{Tot}} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{\bar{X}})^2 \qquad (\text{df} = n-1)$$

into the **treatment sum of squares** (= variability between groups)

$$SS_{\text{Tr}} = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{\bar{X}})^2 \qquad (\text{df} = k-1)$$

and the **error sum of squares** (= variability within groups)

$$SS_{\text{Er}} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2 \qquad (\text{df} = n-k)$$

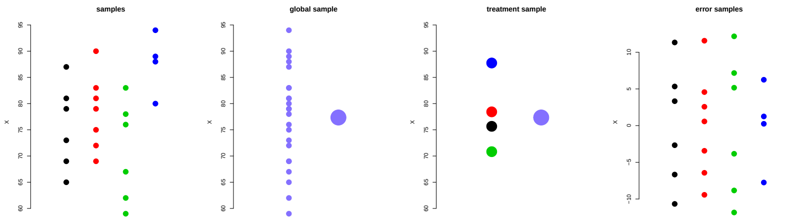
Note the number of degrees of freedom for each quantity (Slide 383)

Variability decomposition

Sum of squares identity:

one can show that $\boxed{SS_{\text{Tot}} = SS_{\text{Tr}} + SS_{\text{Er}}}$

- The total sum of squares SS_{Tot} quantifies the total amount of variation contained in the global sample
- The Treatment sum of squares SS_{Tr} quantifies the variation 'between the groups', that is the variation between the means of the groups
- The Error sum of squares SS_{Er} quantifies the variation within the groups



Mean Squared Error

In sample i, the sample variance is given by $S_i^2 = \frac{1}{n_i-1}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2$

which is an unbiased estimator for σ^2: $\mathbb{E}(S_i^2) = \sigma^2$

Since,

$$SS_{\text{Er}} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2 = \sum_{i=1}^{k}(n_i - 1)S_i^2$$

hence

$$\mathbb{E}(SS_{\text{Er}}) = \sum_{i=1}^{k}(n_i - 1)\mathbb{E}(S_i^2) = \sigma^2\sum_{i=1}^{k}(n_i - 1) = (n-k)\sigma^2$$

→ another unbiased estimator for σ^2 is the **Mean Squared Error** MS_{Er}

$$\boxed{MS_{\text{Er}} = \frac{SS_{\text{Er}}}{n-k}}$$

(generalisation of the 'pooled' sample variance, Slide 405)

→ the number of degrees of freedom for this estimator of σ^2 is $\boxed{n-k}$

Treatment mean square

Now **if H_0 is true**, that is if $\mu_1 = \mu_2 = \ldots = \mu_k = \mu$, we have $\bar{X}_i \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n_i}})$, that is $\sqrt{n_i}(\bar{X}_i - \mu) \sim \mathcal{N}(0, \sigma)$, for all $i = 1, \ldots, k$

→ $\sqrt{n_1}(\bar{X}_1 - \mu), \sqrt{n_2}(\bar{X}_2 - \mu), \ldots, \sqrt{n_k}(\bar{X}_k - \mu)$, is a random sample whose sample variance

$$\frac{1}{k-1}\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{\bar{X}})^2 = \frac{SS_{\text{Tr}}}{k-1}$$

is an unbiased estimator for σ^2

→ the **Treatment Mean Square** MS_{Tr}, defined by

$$\boxed{MS_{\text{Tr}} = \frac{SS_{\text{Tr}}}{k-1}}$$

is also an unbiased estimator for σ^2

→ the number of degrees of freedom for this estimator of σ^2 is $\boxed{k-1}$

ANOVA test

Thus we have two potential estimators of σ^2:

1. MS_{Er}, which always estimates σ^2
2. MS_{Tr}, which estimates σ^2 only when H_0 is true

Actually, if H_0 is not true, MS_{Tr} tends to exceed σ^2, as we have

$$\mathbb{E}(MS_{\text{Tr}}) = \sigma^2 + \text{ 'true' variance between the groups}$$

→ the idea of the ANOVA test now takes shape

Suppose we have observed k samples $x_{i1}, x_{i2}, \ldots, x_{in_i}$, for $i = 1, 2, \ldots, k$, from which we can find through calculations the observed values ms_{Tr} and ms_{Er}. Then:

- if $ms_{\text{Tr}} \simeq ms_{\text{Er}}$, then H_0 is probably reasonable
- if $ms_{\text{Tr}} \gg ms_{\text{Er}}$, then H_0 should be rejected

→ this will thus be a one-sided hypothesis test

We need to determine what "$ms_{\text{Tr}} \gg ms_{\text{Er}}$" means so as to obtain a hypothesis test at given significance level α.

Sampling distribution

It can be shown that, if H_0 is true, the **ratio**

$$F = \frac{MS_{\text{Tr}}}{MS_{\text{Er}}} = \frac{\frac{SS_{\text{Tr}}}{k-1}}{\frac{SS_{\text{Er}}}{n-k}}$$

follows a particular distribution known as the **Fisher's *F*-distribution** with $k-1$ and $n-k$ degrees of freedom, which is usually denoted by

$$\boxed{F \sim \mathbf{F}_{k-1,n-k}}$$

Note: Ronald A. Fisher (1890-1962) was an English statistician and biologist. Some say that he almost single-handedly created the foundation for modern statistical science. As a biologist, he is also regarded as the greatest biologist since Charles Darwin

The Fisher's *F*-distribution

A random variable, say X, is said to follow Fisher's F-distribution with d_1 and d_2 degrees of freedom, i.e.

$$\boxed{X \sim \mathbf{F}_{d_1,d_2}}$$

if its probability density function is given by

$$f(x) = \frac{\Gamma((d_1+d_2)/2)(d_1/d_2)^{d_1/2}x^{d_1/2-1}}{\Gamma(d_1/2)\Gamma(d_2/2)((d_1/d_2)x+1)^{(d_1+d_2)/2}} \qquad \text{for } x > 0$$

for some integers d_1 and d_2 → $S_X = [0, +\infty)$

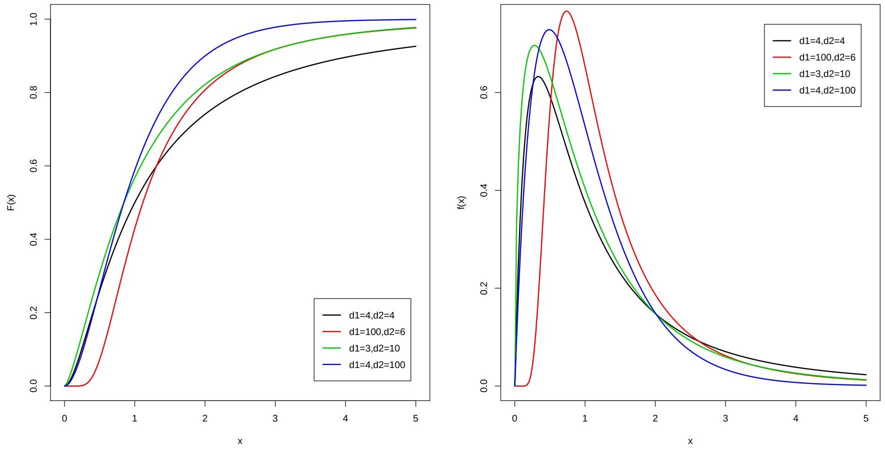
Note: the Gamma function is given by

$$\Gamma(y) = \int_0^{+\infty} x^{y-1}e^{-x}\,dx, \qquad \text{for } y > 0$$

There is usually no simple expression for the F-cdf.

The Fisher's F-distribution

Some F-distributions



cdf $F(x)$

pdf $f(x) = F'(x)$

The Fisher's F-distribution

It can be shown that the mean and the variance of the F-distribution with d_1 and d_2 degrees of freedom are

$$\mathbb{E}(X) = \frac{d_2}{d_2 - 2} \qquad \text{for } d_2 > 2$$

and

$$\mathbb{V}\text{ar}(X) = \frac{2d_2^2(d_1 + d_2 - 2)}{d_1(d_2 - 2)^2(d_2 - 4)} \qquad \text{for } d_2 > 4$$

Note that a F-distributed random variable is **nonnegative**, as expected (ratio of two positive random quantities) and the distribution is **highly skewed to the right**.

The Fisher's F-distribution: quantiles

Similarly to what we did for other distributions, we can define the quantiles of any F-distribution:

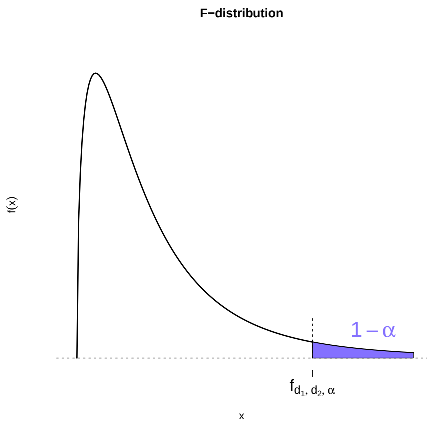
Let $f_{d_1,d_2;\alpha}$ be the value such that

$$\mathbb{P}(X > f_{d_1,d_2;\alpha}) = 1 - \alpha$$

for $X \sim \mathbf{F}_{d_1,d_2}$

The F-distribution is not symmetric, however it can be shown that

$$f_{d_1,d_2;\alpha} = \frac{1}{f_{d_2,d_1;1-\alpha}}$$



ANOVA test

The null hypothesis to test is H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$

versus the general alternative H_a: not all the means are equal

Evidence against H_0 is shown if $MS_{\text{Tr}} \gg MS_{\text{Er}}$, so we will reject H_0 whenever MS_{Tr} is much larger than MS_{Er}, i.e. $\frac{MS_{\text{Tr}}}{MS_{\text{Er}}}$ much larger than 1

→ for testing H_0 at significance level α, we need a constant c such that

$$\alpha = \mathbb{P}\left(\frac{MS_{\text{Tr}}}{MS_{\text{Er}}} > c \text{ if } H_0 \text{ is true}\right)$$

We know that, if H_0 is true, $F = \frac{MS_{\text{Tr}}}{MS_{\text{Er}}} \sim \mathbf{F}_{k-1,n-k}$

→ we have directly that $c = f_{k-1,n-k;1-\alpha}$

From observed values ms_{Tr} and ms_{Er}, the decision rule is:

$$\text{reject } H_0 \text{ if } \quad \frac{ms_{\text{Tr}}}{ms_{\text{Er}}} > f_{k-1,n-k;1-\alpha}$$

ANOVA test: p-value

The observed value of the test statistic is

$$f_0 = \frac{ms_{\text{Tr}}}{ms_{\text{Er}}}$$

and thus the p-value is given by

$$p = \mathbb{P}(X > f_0),$$

where $X \sim \mathbf{F}_{k-1,n-k}$

("the probability that the test statistic will take on a value that is at least as extreme as the observed value when H_0 is true", definition on Slide 347)

This test is also often called the ***F*-test** or **ANOVA *F*-test**.

ANOVA table

The computations for this test are usually summarised in tabular form

Source	degrees of freedom	sum of squares	mean square	F-statistic
Treatment	$df_{\text{Tr}} = k-1$	ss_{Tr}	$ms_{\text{Tr}} = \frac{ss_{\text{Tr}}}{k-1}$	$f_0 = \frac{ms_{\text{Tr}}}{ms_{\text{Er}}}$
Error	$df_{\text{Er}} = n-k$	ss_{Er}	$ms_{\text{Er}} = \frac{ss_{\text{Er}}}{n-k}$	
Total	$df_{\text{Tot}} = n-1$	ss_{Tot}		

Note 1: $df_{\text{Tot}} = df_{\text{Tr}} + df_{\text{Er}}$ and $ss_{\text{Tot}} = ss_{\text{Tr}} + ss_{\text{Er}}$

Note 2: this table is the usual computer output when an ANOVA procedure is run

ANOVA: example

Example

Consider the data shown on Slide 476. Test at significance level $\alpha = 0.05$ the null hypothesis that there is no difference in mean achievement for the four teaching techniques. (**Hint:** You can use the Matlab outputs: $\texttt{finv}(0.95, 3, 19) = 3.1274$, $\texttt{fcdf}(3.77, 3, 19) = 0.9719$)

We have $k = 4$, $n_1 = 6$, $n_2 = 7$, $n_3 = 6$ and $n_4 = 4$, with $\bar{x}_1 = 75.67$, $\bar{x}_2 = 78.43$, $\bar{x}_3 = 70.83$, $\bar{x}_4 = 87.75$ and $s_1 = 8.17$, $s_2 = 7.11$, $s_3 = 9.58$, $s_4 = 5.80$. Besides,

$$n = 6 + 6 + 7 + 4 = 23 \quad \text{and} \quad \bar{\bar{x}} = \frac{1}{n}\sum_{i=1}^{4} n_i \bar{x}_i = 77.35$$

Directly from their expressions,

$$ss_{\text{Er}} = 5 \times 8.17^2 + 6 \times 7.11^2 + 5 \times 9.58^2 + 3 \times 5.80^2 = 1196.63$$

$$ss_{\text{Tr}} = 6 \times (75.67 - 77.35)^2 + 7 \times (78.43 - 77.35)^2 + 6 \times (70.83 - 77.35)^2 + 4 \times (87.75 - 77.35)^2 = 712.59$$

ANOVA: example

From there, the ANOVA table can be easily completed:

Source	degrees of freedom	sum of squares	mean square	F-statistic
Treatments	$df_{\text{Tr}} = k-1$ $= 3$	ss_{Tr} $= 712.59$	$ms_{\text{Tr}} = \frac{ss_{\text{Tr}}}{k-1}$ $\frac{712.59}{3} = 237.53$	$f_0 = \frac{ms_{\text{Tr}}}{ms_{\text{Er}}}$ $\frac{237.53}{62.98} = 3.77$
Error	$df_{\text{Er}} = n-k$ $= 19$	ss_{Er} $= 1196.63$	$MS_{\text{Er}} = \frac{ss_{\text{Er}}}{n-k}$ $\frac{1196.63}{19} = 62.98$	
Total	$df_{\text{Tot}} = n-1$ $= 22$	ss_{Tot} $= 1909.22$		

Is $f_0 = 3.77$ 'much larger' than 1?

→ compare to the appropriate F-distribution critical value

ANOVA: example

According to the hint, $f_{3,19;0.95} = 3.1274$

→ the decision rule is:

$$\text{reject } H_0 \text{ if } \frac{MS_{\text{Tr}}}{MS_{\text{Er}}} > 3.1274$$

Here, we have observed

$$f_0 = \frac{ms_{\text{Tr}}}{ms_{\text{Er}}} = 3.77 \qquad \rightarrow \text{reject } H_0$$

We can claim that the teaching technique does have an influence on the mean achievement of the students (with less than 5% chance of being wrong)

The associated p-value is

$$p = \mathbb{P}(X > 3.77) = 1 - 0.9719 = 0.0281 \qquad \text{for } X \sim \mathbf{F}_{3,19} \text{ (hint)}$$

→ indeed, $p < \alpha = 0.05$ (reject H_0)

ANOVA: confidence intervals on treatment means

- The ANOVA F-test will tell you whether the means are all equal or not, but nothing more
- When the null hypothesis of equal means is rejected, we will usually want to know **which of the μ_i's are different from one another**
- A first step in that direction is to build confidence intervals for the different means μ_i
- From our assumptions (normal populations, random samples, equal variance σ^2 in each group), we have

$$\bar{X}_i \sim \mathcal{N}\left(\mu_i, \frac{\sigma}{\sqrt{n_i}}\right)$$

- The value of σ^2 is unknown, however we have (numerous!) estimators for it

ANOVA: confidence intervals on treatment means

For instance, the MS_{Er} is an unbiased estimator for σ^2 with $n - k$ degrees of freedom.

This one is based on all the n observations from the global sample

→ it has smaller variance (i.e. it is more accurate) than any other (like e.g. S_i), and should always be used in the ANOVA framework!

Acting 'as usual', we can conclude that

$$\sqrt{n_i}\frac{\bar{X}_i - \mu_i}{\sqrt{MS_{\text{Er}}}} \sim t_{n-k}$$

and directly write a $100 \times (1 - \alpha)\%$ two-sided confidence interval for μ_i, from the observed values $\bar{x}_i$ and ms_{Er}:

$$\left[\bar{x}_i - t_{n-k,1-\alpha/2}\sqrt{\frac{ms_{\text{Er}}}{n_i}}, \bar{x}_i + t_{n-k,1-\alpha/2}\sqrt{\frac{ms_{\text{Er}}}{n_i}}\right]$$

→ these confidence intervals for each group will tell which values μ_i's are much different from one another and which ones are 'close'

ANOVA: confidence intervals on treatment means

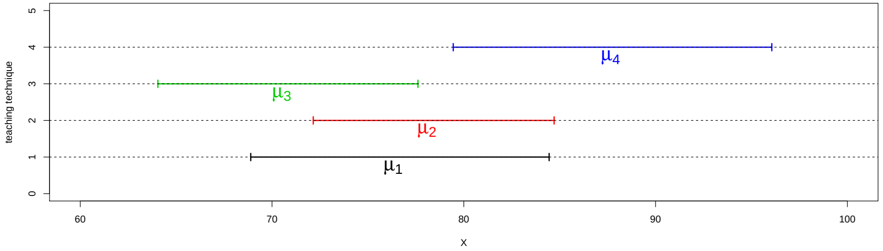
For instance, in our running example (mean achievement for teaching techniques), we would find, with $t_{19;0.975} = 2.093$ (Matlab) and $ms_{\text{Er}} = 62.98$:

$$95\% \text{ CI for } \mu_1 = [75.67 \pm 2.093 \times \sqrt{\tfrac{62.98}{6}}] = [68.89, 84.45]$$

$$95\% \text{ CI for } \mu_2 = [78.43 \pm 2.093 \times \sqrt{\tfrac{62.98}{7}}] = [72.15, 84.71]$$

$$95\% \text{ CI for } \mu_3 = [70.83 \pm 2.093 \times \sqrt{\tfrac{62.98}{6}}] = [64.05, 77.61]$$

$$95\% \text{ CI for } \mu_4 = [87.75 \pm 2.093 \times \sqrt{\tfrac{62.98}{4}}] = [79.45, 96.06]$$



→ it seems clear that $\mu_3 \neq \mu_4$ is the main reason for rejecting H_0

ANOVA: pairwise comparisons

It is also possible to build confidence intervals for the differences between two means μ_i and μ_j. From observed values $\bar{x}_i$, $\bar{x}_j$ and ms_{Er}, a $100 \times (1-\alpha)$ % confidence interval for $\mu_i - \mu_j$ is

$$\left[(\bar{x}_i - \bar{x}_j) - t_{n-k;1-\alpha/2}\sqrt{ms_{\text{Er}}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)},\right.$$
$$\left.(\bar{x}_i - \bar{x}_j) + t_{n-k;1-\alpha/2}\sqrt{ms_{\text{Er}}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}\right]$$

for any pair of groups (i,j)

Finding the value 0 in such an interval is an indication that μ_i and μ_j are not 'significantly' different. On the other hand, if the interval does not contain 0, that is evidence that $\mu_i \neq \mu_j$.

However, these confidence intervals are sometimes misleading and must be carefully analysed, in particular when related to the global null hypothesis $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$

ANOVA: pairwise comparisons

Suppose that for a pair (i,j), the $100 \times (1-\alpha)$% confidence interval for $\mu_i - \mu_j$ does not contain 0

→ at significance level α%, you would reject $H_0^{(i,j)} : \mu_i = \mu_j$

If $\mu_i \neq \mu_j$ then automatically $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ is contradicted

→ should you also reject H_0 at significance level α%? **No**

When you reject $H_0^{(i,j)} : \mu_i = \mu_j$ at significance level α%, you essentially keep a α% chance of being wrong

Successively testing $H_0^{(1,2)}$: $\mu_1 = \mu_2$, and then $H_0^{(1,3)} : \mu_1 = \mu_3$, and then . . ., and then finally $H_0^{(k-1,k)} : \mu_{k-1} = \mu_k$, that is

$$K = \binom{k}{2} = \frac{k!}{2!(k-2)!} \quad \text{pairwise comparisons,}$$

greatly increases the chance of making a wrong decision

(recall Example Slide 112)

ANOVA: pairwise comparisons

Suppose that H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ is true

If the decisions made for each of the K pairwise tests $H_0^{(i,j)} : \mu_i = \mu_j$ were *independent* (which they are not! why?), we would wrongly reject at least one null hypothesis with probability $1-(1-\alpha)^K$ (why?)

If the decisions were *perfectly dependent* (which they are not either!), we would wrongly reject at least one null hypothesis with probability α (why?)

→ if we based our decision about H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ on the pairwise comparison tests, we would wrongly reject H_0 with a probability strictly between α and $1-(1-\alpha)^K$, larger than α !

For instance, suppose $k=4$ groups, which would give $K = \binom{4}{2} = 6$ pairwise comparisons, and $\alpha = 0.05$

→ the test based on pairwise comparisons would be of effective significance level between 0.05 and $1-(1-0.05)^6 = 0.265$

Pairwise comparisons: Bonferonni adjustments

It is usually not possible to determine exactly the significance level of such a test: it all depends on the exact level of dependence between the decisions about the different pairwise comparisons.

Several procedures have been proposed to overcome this difficulty, the simplest being the Bonferonni adjustment method.

It is based on the Bonferonni inequality (see Exercise 1 Tut. Week 3):

$$\mathbb{P}(A_1 \cup A_2 \cup \ldots \cup A_K) \leq \mathbb{P}(A_1) + \mathbb{P}(A_2) + \ldots + \mathbb{P}(A_K)$$

Suppose that A_q is the event 'we wrongly reject H_0 for the qth pairwise comparison'. Then, the event $B = (A_1 \cup A_2 \cup \ldots \cup A_K)$ is the event 'we wrongly reject $H_0 : \mu_1 = \ldots = \mu_k$'

→ if we want $\mathbb{P}(B) \leq \alpha$, it is enough to take $\mathbb{P}(A_q) = \frac{\alpha}{K}$ for all q

Hence, to guarantee an overall significance level of at most α%, the pairwise comparison tests must be carried out at significance level α/K% (instead of α%), where $K = \binom{k}{2}$.

Bonferonni-adjusted t-test

To compare treatment i with treatment j, the null hypothesis is

$$H_0 : \mu_i = \mu_j$$

against the alternative

$$H_a : \mu_i \neq \mu_j$$

The observed value of the test statistic is

$$t_0 = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{ms_{\text{Er}}(\frac{1}{n_i} + \frac{1}{n_j})}}$$

The p-value is computed as

$$p = 2 \times P(T > |t_0|), \quad T \sim t_{n-k}$$

Reject H_0 if p-value is less than α/K, where K is the number of pairwise comparisons.

Pairwise comparisons: example

In our running example, we have $k = 4$ groups, and we can run $K = 6$ pairwise two-sample t-tests. We can find:

- t-test for $H_0 : \mu_1 = \mu_2$ → p-value = 0.5388
- t-test for $H_0 : \mu_1 = \mu_3$ → p-value = 0.3047
- t-test for $H_0 : \mu_1 = \mu_4$ → p-value = 0.0292
- t-test for $H_0 : \mu_2 = \mu_3$ → p-value = 0.1017
- t-test for $H_0 : \mu_2 = \mu_4$ → p-value = 0.0764
- t-test for $H_0 : \mu_3 = \mu_4$ → p-value = 0.0037

At level 5%, we reject $H_0 : \mu_1 = \mu_4$ and $H_0 : \mu_3 = \mu_4$

From this, can we reject $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ at level 5%? No

→ we must compare the above p-values to $\alpha/K = 0.05/6 = 0.0083$

The last p-value is smaller than 0.0083 → reject $H_0 : \mu_3 = \mu_4$

There is a significant difference between mean achievement for using teaching techniques 3 and 4 only

Adequacy of the ANOVA model

The ANOVA model is based on several assumptions that should be carefully checked.

The central assumption here is that the random variables $\varepsilon_{ij} = X_{ij} - \mu_i$, $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$, are (1) independent and (2) normally distributed:

$$\varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma),$$

with (3) the same variance in each group.

We do not have access to values for ε_{ij} (μ_i's are unknown!), however we can approximate these values by the observed residuals

$$\hat{e}_{ij} = x_{ij} - \bar{x}_i$$

Note that these residuals are the quantities arising in ss_{Er}.

→ as for a regression model (see Slide 464), the adequacy of the ANOVA model is established by examining the residuals

→ **residual analysis**

Residuals analysis

- The **normality** assumption can be checked by constructing a **normal quantile plot** for the residuals
- The assumption of **equal variances** in each group can be checked by plotting the residuals against the treatment level (that is, $\bar{x}_i$)

→ the spread in the residuals should not depend on any way on $\bar{x}_i$

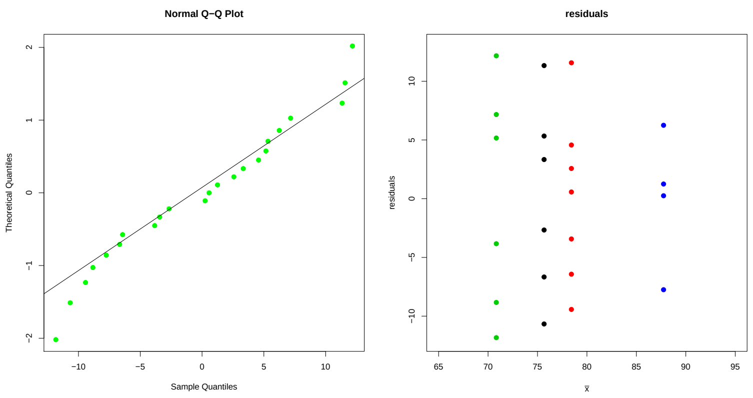
- A rule-of-thumb is that, if the ratio of the largest sample standard deviation to the smallest one is smaller than 2, the assumption of equal population variances is reasonable
- The assumption of **independence** can be checked by plotting the residuals against time, if this information is available

→ no pattern, such sequences of positive and negative residuals, should be observed

- As for the regression, the residuals are everything the model will not consider → no information should be observed in the residuals, they should look like random noise

Residual analysis: example

For our running example, a normal quantile plot and a plot against the fitted values $\bar{x}_i$ for the residuals are shown below:



→ nothing (obvious) to report

→ the assumptions we made look valid

Residual analysis: example

Example

To assess the reliability of timber structures, researchers have studied strength factors of structural lumber. Three species of Canadian softwood were analysed for bending strength (Douglas Fir, Hem-Fir and Spruce-Pine-Fir). Wood samples were selected from randomly selected sawmills. The results of the experiment are given below. Is there any significant difference in the mean bending parameters among the three types of wood? (**Hint:** You can use the Matlab outputs: $\texttt{finv}(0.95, 2, 15) = 3.68$, $\texttt{fcdf}(0.33, 2, 15) = 0.274$)

Douglas (1)	Hem (2)	Spruce (3)
370	381	440
150	401	210
372	175	230
145	185	400
374	374	386
365	390	410

→ an ANOVA was run to test the null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$, against the alternative H_a : not all the means are equal

Residual analysis: example

We computed values for the ANOVA table:

Source	degrees of freedom	sum of squares	mean square	F-statistic
Treatment	2	7544	3772	0.33
Error	15	172929	11529	
Total	17	180474		

According to the hint, we know that $f_{2,15;0.95} = 3.68$

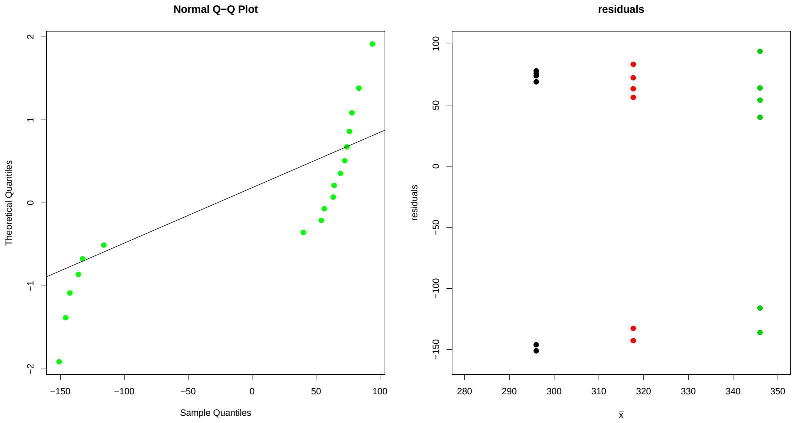
→ here we have observed $f_0 = 0.33$ → do not reject H_0 !

Associated p-value: $p = \mathbb{P}(X > 0.33) = 1 - 0.274 = 0.726$ for $X \sim \mathbf{F}_{2,15}$

→ we confidently claim that there is no significant difference in the mean bending parameters for the different wood types

Residual analysis: example

Residual analysis:

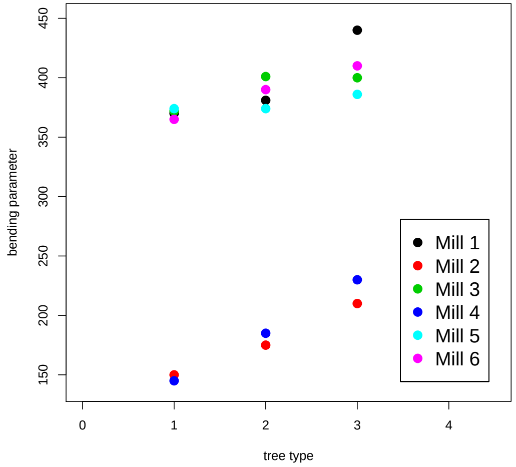


→ the assumptions are clearly not fulfilled!

→ the above conclusion is certainly not reliable!

Blocking factor

If we had plotted the data first, we would have seen:



(Bottom line: **always** plotting the data before analysing them!)

Blocking factor

It is clear that over and above the wood type, the mills where the lumber was selected is **another source of variability**, in this example even more important than the main treatment of interest (wood type).

This kind of extra source of variability is known as a **blocking factor**, as it essentially groups some observations in blocks across the initial groups $\rightarrow$ the samples are not independent! (assumption violation)

$\rightarrow$ a potential blocking factor must be taken into account!

When a blocking factor is present, the ‘initial’ Error Sum of Squares, say SS^*_{Er}, that is the whole amount of variability not due to the treatment, can in turn be partitioned into:

1. the variability due to the blocking factor, quantified by SS_{Block}
2. the ‘true’ natural variability in the observations SS_{Er}

We can write that $SS^*_{\text{Er}} = SS_{\text{Block}} + SS_{\text{Er}}$, and thus

$$SS_{\text{Tot}} = SS_{\text{Tr}} + SS_{\text{Block}} + SS_{\text{Er}}$$

Blocking factor

The ANOVA table becomes:

Source	degrees of freedom	sum of squares	mean square	F-statistic
Treatment	$k-1$	ss_{Tr}	$ms_{\text{Tr}} = \frac{ss_{\text{Tr}}}{k-1}$	$f_0 = \frac{ms_{\text{Tr}}}{ms_{\text{Er}}}$
Block	$b-1$	ss_{Block}	$ms_{\text{Block}} = \frac{ss_{\text{Block}}}{b-1}$	
Error	$n-k-b+1$	ss_{Er}	$ms_{\text{Er}} = \frac{ss_{\text{Er}}}{n-k-b+1}$	
Total	$n-1$	ss_{Tot}		

where b is the number of blocks

Note: the test statistic is again the ratio $\frac{MS_{\text{Tr}}}{MS_{\text{Er}}}$ (we have just removed the variability due to the blocking factor first), to be compared with the quantile of the $\mathbf{F}_{k-1,n-k-b+1}$ distribution

Blocking factor

In the previous example, we would have found:

Source	degrees of freedom	sum of squares	mean square	F-statistic
Treatment	2	7544	3772	15.87
Block	5	170552	34110	
Error	10	2378	238	
Total	17	180474		

From MATLAB, we can find that $f_{2,10;0.95} = 4.10$

Here, we have observed $f_0 = 15.87 \rightarrow$ clearly reject H_0!

Associated p-value: $p = \mathbb{P}(X > 15.87) = 0.0008$ for $X \sim \mathbf{F}_{2,10}$

Blocking factor: comments

- The SS_{Er} in the first ANOVA (without block) was 172,929 which contains an amount of variability 170,552 due to mills

→ about 99% of the initial SS^*_{Er} was due to mill to mill variability, and so was no *natural variability*!

- The second ANOVA (with blocking factor) adjusts for this effect
- The net effect is a substantial reduction in the 'genuine' MS_{Er}, leading to a larger F-statistic (increased from 0.33 to 15.87!)

→ with very little risk ($p \simeq 0$), we can now conclude that there is a significant difference in the mean bending for the three wood types

- An analysis of the residuals in this second ANOVA would not show anything peculiar → valid conclusion

Generally speaking, ignoring a blocking factor leads to a misleading conclusion, and it should always be carefully assessed whether a blocking factor may exist or not (plot the data!)

Objectives

Now you should be able to:

- conduct engineering experiments involving a treatment with a certain number of levels □
- understand how the ANOVA is used to analyse the data from these experiments □
- assess the ANOVA model adequacy with residual plots □
- understand the blocking principle and how it is used to isolate the effect of nuisance factors □

Recommended exercises:

→ Q3, Q6 p.406, Q9 p.407, Q10, Q11 p.412, Q13, Q15, Q17 p.413, Q19 p.414, Q22, Q23 p.415, Q35 p.428 (2nd edition)

→ Q3, Q6 p.418, Q9 p.418, Q10, Q11 p.423, Q13, Q15 p.424, Q17, Q19 p.425, Q20 p.426, Q23 p.427, Q35 p.439 (3rd edition)