MSc Mas6002 Introductory Material
Block B
Statistical Methods

1 Data types

Data come in many different forms. Typically we have a collection of records from different individuals (or components or units or ...) on one or more characteristics: data are **univariate** if there is information on just one characteristic per individual and **multivariate** otherwise.

Example 1: `Restrnt.txt`
The 1980 Wisconsin Restaurant Survey was conducted by the University of Wisconsin Small Business Development Centre, selected 19 Wisconsin counties for study. Samples were taken in each county. This is a multivariate data set since 13 characteristics are recorded for each restaurant (but some values are missing and coded as NA).

Column	Name	Count	Missing	Description
C1	ID	279	0	Identification Number of Restaurant
C2	Outlook	279	1	Business outlook from 1 = very unfavourable to 7 = very favourable
C3	Sales	279	25	Gross 1979 sales in $1000s
C4	NewCap	279	55	New capital invested in 1979 in $1000s
C5	Value	279	39	Estimated market value of the business in $1000s
C6	CostGood	279	42	Cost of goods sold as a percentage of the business
C7	Wages	279	44	Wages as a percentage of sales
C8	Ads	279	44	Advertising as a percentage of sales
C9	TypeFood	279	12	1 = fast food, 2 = supper club, 3 = other
C10	Seats	279	11	number of seats in dining area
C11	Owner	279	10	1 = sole proprietorship, 2 = partnership, 3 = corporation
C12	Ft.Empl	279	14	Number of full-time employees
C13	Pt.Empl	279	13	Number of part-time employees
C14	Size	279	16	Size of restaurant 1 = 1 to 9.5 employees, 2 = 10 to 20, 3 = over 20 (a part-time employee is 0.5)

The characteristics recorded, usually called **variables**, can be either **quantitative** or **qualitative**.

Quantitative variables are those that, of their nature, take numerical values for which arithmetic makes sense, e.g. Sales; Value; Ft.Empl. For each of these, finding a total or average value makes sense. Quantitative variables are usually either **discrete** or **continuous**. Discrete variables are often 'counts', that is the result of counting something, and continuous ones are often measurements. The possible values for a discrete variable are isolated or separated values, usually, but not necessarily, whole numbers, e.g. Seats, Pt.Empl. Continuous variables may take any value in an interval or collection of intervals.

For physical measurements (height, weight, etc) it is clear what this means, even though there will be a limit to the accuracy. In many cases the judgement that a variable is to be regarded as continuous is a practical one, based on the range and density of its possible values. Many of the variables in the data set, though recorded as whole numbers, should be regarded as continuous, e.g. Sales, NewCap, Value, CostGood.

In the above example, Owner, although recorded as 1, 2 or 3, is not quantitative; it is **qualitative**. These numerical values are just arbitrary labels, for the three kinds of owner, and could just as well have been assigned in other ways. Whenever the possibilities for a variable are really descriptions the variable is qualitative (even if numbers are used to code it). Also, any three values could have been used and arithmetic on these values does not make sense. Qualitative variables result from dividing into categories. Three examples are: sex — male/female; pain — none/mild/moderate/severe; age — young/middle-aged/old. When the categories have no intrinsic order or sequence (like male/female) they are called **nominal**. In contrast, the categories for pain and age have a natural order. Such variables with ordered categories are often called **ordinal**. Numerical values may be used to label the categories (and for ordinal variables the numerical values should respect the ordering) but this does not change the basic nature of the variable. Owner, Typefood and Outlook are qualitative; Owner and Typefood are nominal, and Outlook is also ordinal.

A (**raw**) data set may be very extensive, as in the Restaurant Survey, and so it is often very difficult to see immediately the relevant structure and variation in the data. The essential features are often obscured so that it is difficult to draw any useful conclusions from the information available. For this reason data are often presented in summary form — either through tables and diagrams or numerically.

2 Summary Tables and Diagrams

We look now at ways of extracting information from raw data to highlight the relevant structure and variation. The pattern of variation in the measurements of a variable is called its **distribution**. We shall try to assess this distribution in tabular and graphical form. First we consider forms appropriate for univariate data.

2.1 Dot plot

The simplest graphical display, so simple it is rarely used, is the Dot Plot[1]. Such plots make it easy to see the way the values are spread. However they become messy and cumbersome for larger data sets.

Example 2: `remission.txt`.
The data below are the remission times in weeks of 10 patients presenting with a certain type of carcinoma and receiving radiotherapy treatment. A **dot plot** of these is given Figure 1.

25 45 238 94 16 23 30 16 22 123

[1] In R, see the help pages for `stripchart` and `dotchart` — but frankly it is hard to imagine when this would be a sensible display for a single data set. `stripchart` does allow you to split dotplots by another variable and then this can be an alternative to box-plots (§3.4) which are discussed later.

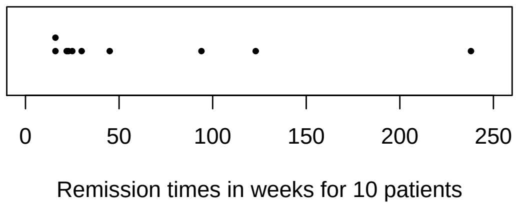

Figure 1: Simple Dot Plot

2.2 Stem-and-leaf plot

A useful alternative way to visualise the distribution of the values in larger data sets is the **stem-and-leaf plot**. Most of you will have seen these before.

Example 3: `grapeweights.txt`
The following data are the weights of 27 'one kilogram' bunches of grapes (in g).

1009	1013	996	1010	1003	1000	994
1017	988	1007	981	997	1009	1012
985	973	1063	1031	1002	1002	1020
1024	1018	1028	1025	990	1013	

Prepare a stem-and-leaf plot to illustrate this data set.

Solution
For the simplest form of a stem-and-leaf plot, split the data into groups, based on their second to last digit, to form the **stem**, with the last digit of individual values forming the **leaves**. There are various other ways of presenting the stem and the leaves but we will not look at them here. When working by hand the idea is to record the minimum detail consistent with an intelligible presentation. This display gives a simple picture of overall shape, highlights gaps in the values and picks up outliers (values far removed from the rest).

97	3
98	5 8 1
99	6 7 0 4
100	9 7 3 2 0 9 2
101	7 3 8 0 3 2
102	4 8 5 0
103	1
104	
105	
106	3

stem **leaves**

or

97	3
98	1 5 8
99	0 4 6 7
100	0 2 2 3 7 9 9
101	0 2 3 3 7 8
102	0 4 5 8
103	1
104	
105	
106	3

ordered stem-and-leaf
(not usually done by hand)

Here the display raises the question 'Is 1063 correct?' Perhaps it is a misprint for 1036. If possible the statistician should then check back with the data source.

Stem-and-leaf plots are available in R: the command is `stem()`.

2.3 Frequency Table

Again, as the number of data points increase, stem-and-leaf plots become cumbersome. We can continue splitting each leaf as above, but the retention of **all** the information is not necessary for an overall view of the pattern of variation. An alternative procedure, which condenses the data, is to **classify** it into groups.

Example 4: `Systolic-bp.txt`
The systolic blood pressures (mmHg) of 70 normal British males are measured, the men all being in the 25-45 age group.

99	148	151	120	116	143	110	110	131	110
136	123	177	117	137	163	113	120	110	105
108	120	116	133	130	138	125	123	124	127
101	123	153	118	127	132	120	147	161	121
122	168	112	186	153	120	96	155	138	123
117	121	144	117	107	115	152	146	109	133
128	118	123	106	117	121	115	130	145	136

Prepare a frequency table to summarise the data.

Solution

Class	Frequency	Relative Frequency	Relative Frequency
90–99	2	2/70	0.029
100–109	6	6/70	0.086
110–119	16	16/70	0.229
120–129	19	19/70	0.271
130–139	11	11/70	0.157
140–149	6	6/70	0.086
150–159	5	5/70	0.071
160–169	3	3/70	0.043
170–179	1	1/70	0.014
180–189	1	1/70	0.014
	70	1	1.000

The classes, defined by the class limits, must be non-overlapping so that there is no doubt as to which class an observation belongs. Since systolic blood pressure is a **continuous** variable 90–99 is interpreted to mean 89.5 to 99.5, with, it is assumed, observations exactly equal to 89.50 and 99.50 being rounded up, to 90 and 100, respectively.

If the above data had been for marks in a test out of 200, a **discrete** variable, then 90–99 would mean 90, 91,. . .,99. The relative frequency column is not essential but its inclusion, usually as a percentage or a decimal, helps when comparing frequency tables for samples of different sizes.

For large data sets it is common for only a frequency table to be published. For example, publications of the Government Statistical Service are full of frequency tables and much data can be accessed through their web site, `http://www.statistics.gov.uk/`

2.4 Bar Chart and Histogram

The bar chart and histogram are used to give a graphical representation of a frequency table for observations on discrete and continuous variables respectively.

A sample of 100 students yields the following frequency table for the variable 'number of brothers'.

Number of brothers	**Frequency**
0	30
1	34
2	20
3	12
4	3
5	0
6	1

Since this variable is discrete, taking whole number values between 0 and 6, a **bar chart** is the usual graphical representation of such a frequency table. For each observed value there is a bar or block, of constant width, with height representing frequency and each bar is separated by a gap from adjacent bars. An example is given in Figure 2.

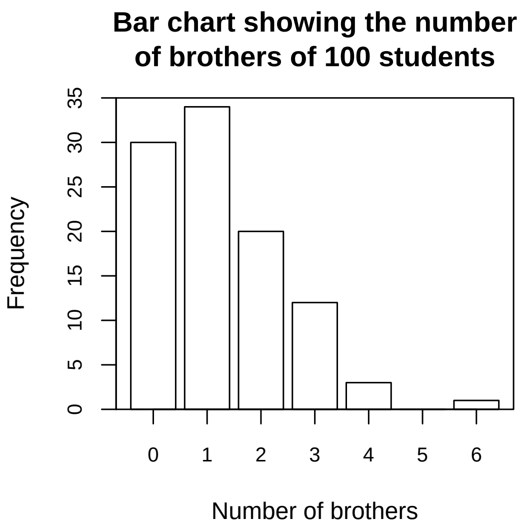

Figure 2: Example of Bar Chart

Blood pressure in `Systolic-bp.txt` is recorded as whole number values but it is a measurement and thus is a **continuous** variable. In a **histogram**, because the underlying variable is continuous, the blocks are connected. The histogram for the blood pressure data is given in Figure 3.

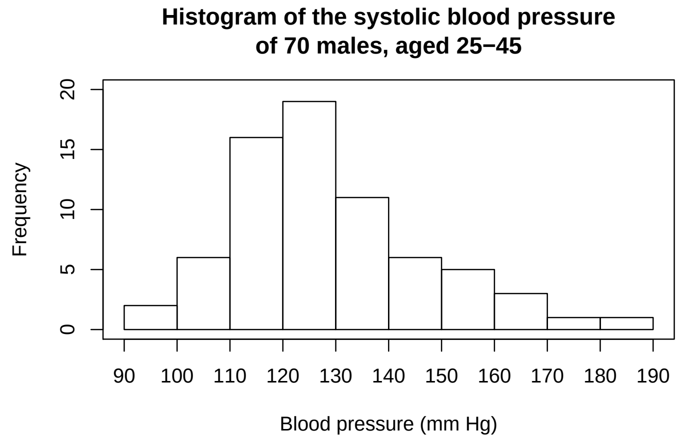

Figure 3: Example of Histogram

If the frequency table for the blood pressure data was actually the frequency table for the variable ‘marks on a test out of 200’ — a discrete variable, only taking whole number values — then it might be argued that a small gap should be left between the blocks in the above histogram to yield a bar chart. This is only the right procedure when each block corresponds to, at most, a few possible values. Here, continuity is approximately true and so a histogram representation of the data set is justifiable and produces a better display.

Since it is only the **shape** of a histogram which is of interest, it is sometimes preferable to use relative frequencies (i.e. proportions) rather than frequencies on the vertical axis. If there are n observations in the data set, then, for any class, the relative frequency is just (frequency)$/n$.

In the blood pressure example the classes are of equal width; each is of width 10mmHg. The calculations must be modified if the classes have unequal widths. (Imagine stacking counters to make the blocks — if they have to cover three times the width they can only reach one-third of the height.) For example, replace the last three classes by a single one, 160–189, which has a width 30 and a frequency 5. The correct plot is obtained by allowing for the differences in width, as suggested by the stacking counters illustration; this shows that it is really the area of the blocks that actually represents frequency, not the height. The **density** is given by adjusting relative frequencies by the class widths, so

$$\text{density} = \frac{\text{frequency}}{\text{sample size} \times \text{class width.}}$$

Using density does not change the detailed shape of the histogram if all blocks are the same width, but it will otherwise. (Of course any constant multiple of density will give the same picture, so frequency divided by any appropriate ‘width factor’ will do for constructing a single histogram.) In R, you can plot histograms using either density or frequency (check the `hist` help page for more on this).

It is possible to produce a histogram with unequal classes in R. For illustration, Figure 4 gives an example of such a histogram.

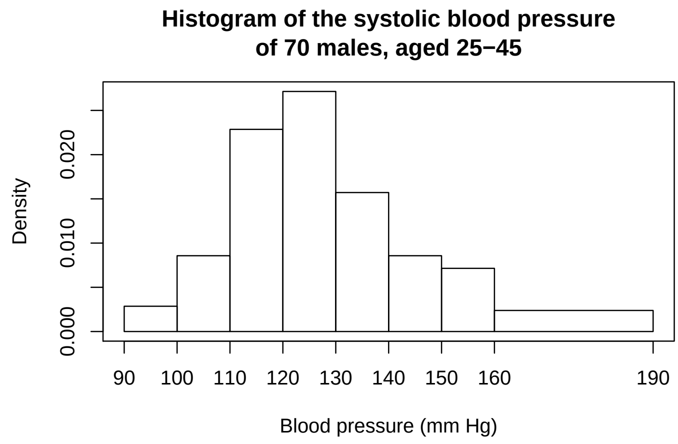

Figure 4: Example of Histogram with unequal classes

Histogram construction

Once you start to override the automatic choices in R, or want to construct such a histogram by hand, you need some ideas about what to do.

- How many classes should there be? This is a matter of judgement, and is partly arbitrary. If you select too many, you get a bumpy diagram; if you select too few, you lose a lot of information. Aim for between 5 and 15. The choice depends on the sample size: larger samples merit more classes.
- Make the classes of equal width if you can.
- Choose sensible (natural) end-points, **class-limits**, for the classes, and be clear about them – but don't worry unduly about their specification. Remember the graph is only a simple visual summary of the data. In the blood pressure example, presumably the blood pressures (on a continuous scale) are measured to the nearest mmHg, so that the intervals are 'really' $89.5 \leq x < 99.5$; $99.5 \leq x < 109.5; \ldots$ but it would be foolish to think this is important in constructing a graphical display.
- Use a block for each class with height which is either i) frequency, relative frequency, percentage or density when classes are of equal width, or ii) density or some constant multiple of this in other cases.

Notes

- For a continuous variable use a histogram, with no gaps between the boxes.
- For a discrete variable with a small number of values use a bar chart with gaps between boxes, otherwise use a histogram.
- Note that number of values is not the same as number of observations, in the 'number of brothers' examples the values are $0, 1, \ldots, 6$ but the number of observations is 100. When density is the vertical scale the total area of the blocks is 1.0. (This is of relevance in probability theory.) Usually, there is no need to include a vertical axis/scale on a histogram/bar chart. Remember that it is the shape that is important. The examples above have a vertical scale for pedagogical reasons.

Descriptive Terminology

On the basis of the shape of the histogram/bar chart the distribution of a variable might be described as positively/negatively skewed, bimodal or bell shaped. These are indicated in Figure 5. All those have a single peak, a modal class, i.e. a class with local maximum frequency. Figure 6 illustrates a shape with two peaks (bimodal).

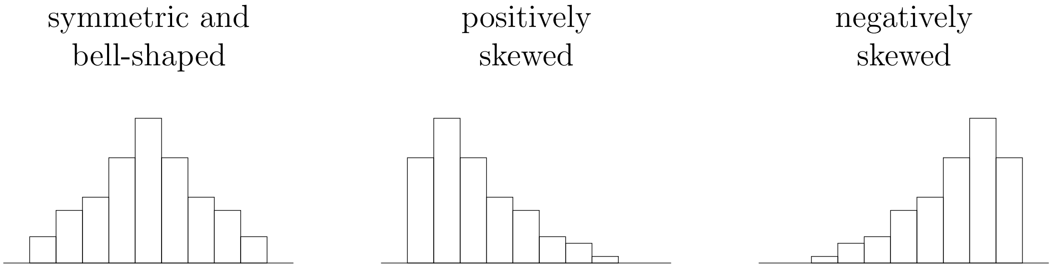

Figure 5: Histogram Shapes

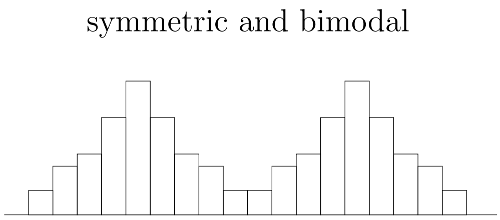

Figure 6: Bimodal Histogram

2.5 Cumulative relative frequency diagram

An alternative diagrammatic summary uses the cumulative relative frequencies. When the only data available are in grouped form, this diagram is useful for obtaining values for certain data summaries (the median and quartiles) that will be introduced later. However the real importance of the idea is theoretical, since the mathematical counterpart of cumulative relative frequencies are frequently tabulated in statistical tables and is available in R for many standard distributions.

The next table gives some data extracted from Table 10.13 of Social Trends **30**, 2000. The distances are in miles. 'Abroad' and '50 miles or over' in the original table, have been combined and treated as 50–200; obviously, for some purposes, this won't be sensible. Only the Owner-occupied data are tabulated here.

Class (distance)	**Frequency (percentage)**	**Relative Frequency**	**Cumulative Relative Frequency**
0–1	21	0.21	0.21
1–10	50	0.50	0.71
10–20	9	0.09	0.80
20–50	6	0.06	0.86
50–200	14	0.14	1.00
	100	1.00	

Thus, for example, the proportion moving less than or equal to 20 miles is 0.8.

A cumulative relative frequency diagram plots these values against the **upper end point** of the appropriate class interval. The cumulative curve is always monotonic increasing starting from 0 and rising to 1. For discrete variables with few values the display is generally presented in the form of a step-function.

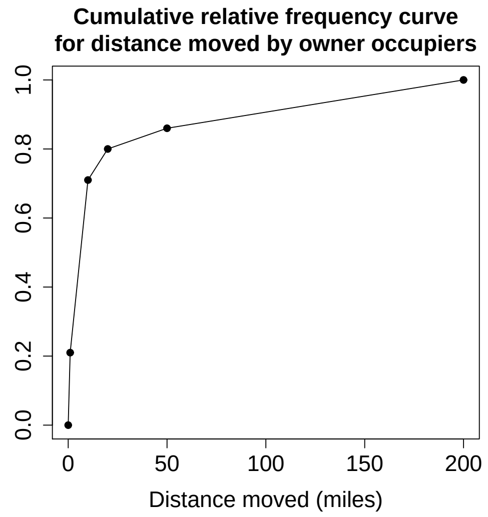

Figure 7: Example of a Cumulative Plot

Here is a histogram of the same data, which shows that they are very positively skewed.

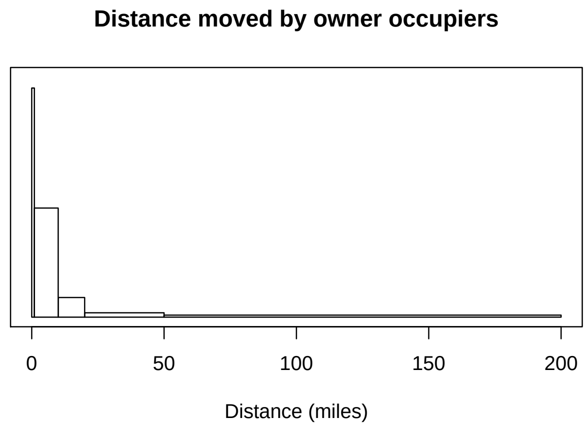

Figure 8: The Histogram for the data in Figure 7

If you have the raw data then the cumulative plot is best obtained by using all of the data. For this you have to plot each individual value against its rank divided by the sample size.

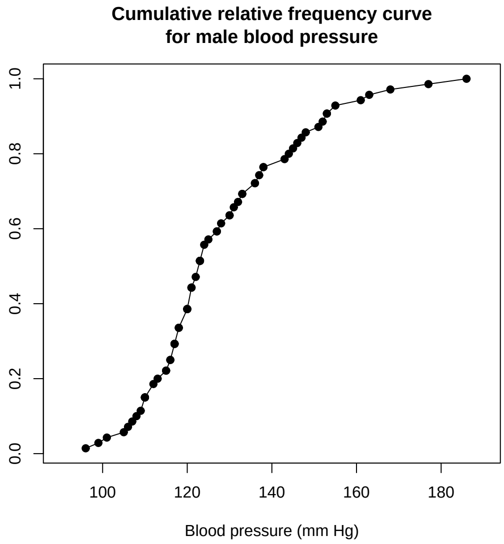

Figure 9: Example of a Cumulative Plot from individual data values

3 Numerical summaries

Although the techniques already given provide overall pictures of the variation in the data, we often require more concise (numerical) summaries — **descriptive statistics.** In this section we assume the data set represent a random sample of n observations on a variable.

3.1 Sample Mean

The first element to summarise is the general size of the numbers, to measure their **central tendency** or **location.** The most widely used measure of location is the sample mean or average.

For a random sample of n observed values, x_1, x_2, x_3, $\ldots$, x_n, the sample mean is given by $\overline{x}$ (x bar)

$$\overline{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Example 5: Means of remission, grapeweights and Systolic-bp
What are the means for data `remission.txt`, `grapeweights.txt` and `Systolic-bp.txt`?

Solution

remission times	$(25+45+\ldots+123)/10 =$	63.2 weeks
grape weights	$(1009+1017+\ldots+1020)/27 =$	1007.8 g
Systolic Blood Pressure	$(99+136+\ldots+136)/70 =$	128.0 mmHg

Notes

1. The sample mean is in the same units as each individual measurement.
2. As a summary it is usually reasonable to quote the sample mean to one significant figure more than is used for each individual measurement: in these three examples — to 1 decimal place. Giving too many decimal places at the final answer is spurious accuracy, showing poor numerical sense.
3. Notice that the sample mean does not have to take one of the values attained in the data set — or even an attainable value.
4. For anything other than small data sets, the calculation is best done with a package like R.
5. If we have grouped data we can obtain a sample mean from the frequency table, approximating the raw data value. The method assumes that within each class all the values in that class take the mid-point value. As a formula:

$$\overline{x} = \frac{\sum_j f_j y_j}{\sum_j f_j} = \sum_j (rf)_j y_j$$

where f_j is the frequency of the class j, $(rf)_j$ is the relative frequency of class j, y_j is the mid-point of class j, and summations are over the number of classes.

Example 6: Mean of frequency data
Suppose the data set `Systolic-bp.txt` were available only as a frequency table as follows.

Blood pressure	**Frequency**
90–99	2
100–109	6
110–119	16
120–129	19
130–139	11
140–149	6
150–159	5
160–169	3
170–179	1
180–189	1
	70

Calculate the sample mean blood pressure based on this table.

Solution
The sample mean, based on this table is:

$$\overline{x} = \frac{2 \times 94.5 + 6 \times 104.5 + \ldots + 1 \times 184.5}{70} = \frac{8985.0}{70} = 128.4\text{mm Hg}.$$

3.2 Sample Median

For the data set `remission.txt` the value 238 has a very large effect on the sample mean. If 238 is omitted the mean becomes 43.8 weeks (reduced from 63.2 weeks). Observations that have a large influence on the sample mean are called **outliers** or **extreme values.** A more resistant measure of location is the **sample median**; roughly, this is the value such that half the observations are above it and half the observations are below it.

The sample median is denoted by $\tilde{x}$, read as x tilde.

To find the median: arrange the observations in order (smallest to largest); count $(n+1)/2$ observations up from the bottom.

Example 7: Median of data set `remission.txt`
Data set `remission.txt` gives the remission times in weeks of 10 patients presenting with a certain type of carcinoma and receiving radiotherapy treatment as follows:

16 16 22 23 25 30 45 94 123 238

What is the median remission time?

Solution

$$\tilde{x} = \frac{25+30}{2} = 27.5 \text{ weeks}$$

Example 8: Median of data set `grapeweights.txt`
The data set `grapeweights.txt` is larger. Don't try to do the calculation by hand, but what methodology would you use to obtain its median?

Solution
In `grapeweights.txt`, $n = 27$ so the median value is obtained by sorting the data and identifying the value that lies at $(n+1)/2 = 14$. If we do this in R we get $\tilde{x} = 1009$ g.

Notes

1. If n is odd ($n = 2m+1$, say), the sample median is the $(m+1)$th ordered observation; if n is even ($n = 2m$, say), the sample median is the average of the mth and $(m+1)$th ordered observations.
2. The sample median is in the same units as each individual measurement.
3. The mean is easier to deal with mathematically and theoretically.
4. A roughly symmetrical data set has mean and median approximately equal. If the mean is much larger than the median the data have strong positive skew since the **long tail** of large values inflates the sample mean. Similarly, if the mean is much smaller there is negative skew. The relative values of the mean and median tell you something about the shape of the distribution.
5. In R, `median()` or `summary()` will return the median.

3.3 Sample Quartiles

Neither the sample mean nor the sample median tells us anything about the amount of variation or dispersion in the data. If we are using the sample median as the measure of location, then sample quartiles are often used as **resistant** measures of dispersion. The three quartiles (Q1, Q2 and Q3) divide the data in to four parts:
Q1: sample lower quartile: roughly a quarter of observations below
Q3: sample upper quartile: roughly a quarter of observations above
Q2: the sample median: in the middle of the observations

The **sample interquartile range** = Q3 − Q1. This is the range of the central 50% of observations and is often denoted by IQR. To obtain the sample quartiles and interquartile range:

1) find sample median as before;
2) find Q1 = the median of the observations below the location of the sample median;
3) find Q3 = the median of the observations above the location of the sample median;
4) evaluate Q3 − Q1.

Example 9: Quartiles for `remission.txt`
What are the quartiles for data set `remission.txt` ?

Solution
Q2 = 27.5 weeks (not one of the observed values, see Example 7)
5 observations below, therefore Q1 in 3rd position among these five:
Q1 = 22 weeks
5 observations above, therefore Q3 in 3rd position among these five:
Q3 = 94 weeks
Sample interquartile range = 94 − 22 = 72 **weeks**

Example 10: Quartiles for `grapeweights.txt`
What are the the quartiles for data set `grapeweights.txt`?

Solution
Q2 = 1009 g; the 14th observation (see Example 8)
13 observations below median, Q1 is in 7th position among these:
Q1 = 996 g
13 observations above median, Q3 is in 7th position among these:
Q3 = 1018 g
Sample interquartile range = 1018 − 996 = 22 **g**

Notes

1. The IQR, Q1, Q2 and Q3 are all unaffected by a few extreme observations, so they provide resistant measures of location and dispersion.
2. Unfortunately, quartiles are not easy to handle theoretically.
3. There are other methods for calculating sample quartiles. They can produce answers that are slightly different, but not in any important way. The rule suggested here is easy to recall and apply.

4. In R, quartiles are available via the `summary()` command.
5. A cumulative relative frequency diagram can be used to obtain quartiles graphically for grouped data. Just read back from 0.25, 0.50 and 0.75.

3.4 Box plots

The five values (minimum, Q1, Q2, Q3, maximum) provide a summary of a set of data, sometimes called the **five number summary**, which can be illustrated through a **box (-and-whisker)** plot, as can be seen Figure 10.

`grapeweights.txt`	`Systolic-bp.txt`
Max = 1063	Max = 186
Q3 = 1018	Q3 = 138
Q2 = 1009	Q2 = 123
Q1 = 996	Q1 = 116
Min = 973	Min = 96

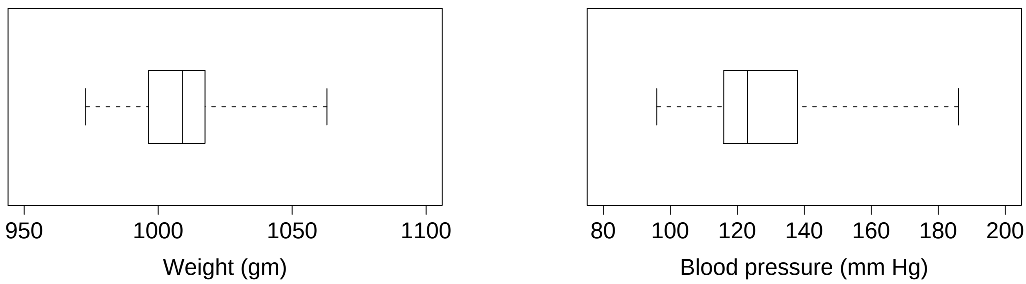

Figure 10: Examples of Boxplots

Notes

1. Box plots are useful for comparing several distributions; stem-and-leaf plots provide better displays for single data sets.
2. They are sometimes modified to identify extreme values as follows[2]:

   (a) Extend whiskers only to most extreme observation within 1.5IQR above and below Q1 and Q3.

   (b) Insert any more extreme values individually as a '*' or a line.

   For `Systolic-bp.txt`: IQR = Interquartile range = $138 - 116 = 22$
   i.e. extend whiskers at most to $116 - 33 = 83$, $138 + 33 = 171$
   i.e. extend whiskers, in fact, to 96 and to 163 with 177, 186 separate.
   The modified box plot is given in Figure 11.

[2]There are even more refined versions

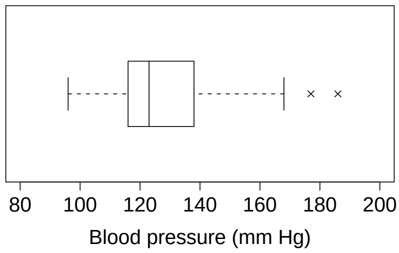

Figure 11: Example of a modified Boxplot

3. For the data `Restrnt.txt` described in Section 1, it is natural to expect that the variable Sales will vary with the variable Size. You can investigate this by drawing the boxplots. Then you would obtain Figure 12, which isn't too good. Non-negative variables that are positively skewed (as Sales is here) often produce a better spread if you take logarithms. The box plots for log(Sales) in Figure 13 are rather better.

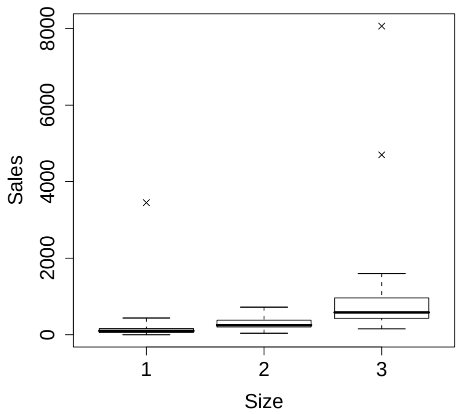

Figure 12: Boxplot: Sales by Size

4. Some published data are given in the form of five-number summaries, but for large data sets it is usual to replace the maximum and minimum by the upper and lower deciles (i.e. the values with only 10% above and 10% below). These can still be used to produce box plots, but now the whiskers will only go out to the deciles. Here is some data of that form taken from the government statistics web site: `http://www.statistics.gov.uk/`.

Table 6	
Type of Data set:	Cross-Sectional
Title:	New Earnings Survey 1999 Distribution weekly earnings
Last Updated:	29/11/99
Associated Web Links:	There are no Web links stored for this product
Time Frame:	April 1999
Geographic Coverage:	Great Britain
Universe:	Earnings distribution
Measure:	Earnings per week
Units:	£ per week

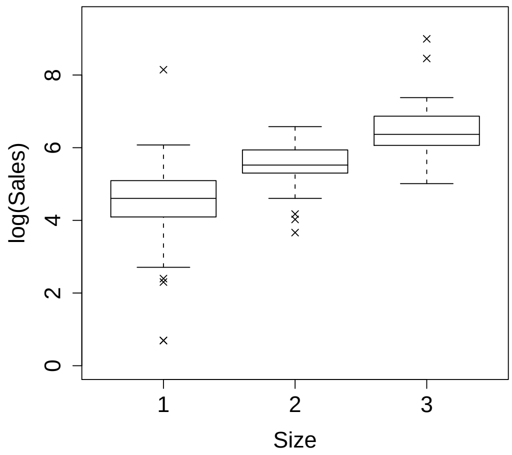

Figure 13: Boxplot: log(Sales) by Size

	Full-time All	Full-time Non-manual	Full-time Manual
Women Top 10 per cent	521	541	328
Women Top 25 per cent	398	422	261
Women Median	284	305	201
Women Bottom 25 per cent	213	230	165
Women Bottom 10 per cent	170	184	140
Men Top 10 per cent	712	863	501
Men Top 25 per cent	517	612	399
Men median	374	449	313
Men bottom 25 per cent	275	321	245
Men bottom 10 per cent	211	234	195

You might draw (by hand) box plots to compare the earnings in some of these categories.A couple are given in Figure 14. Note that both distributions are positively skewed and that the men's is higher than the women's.

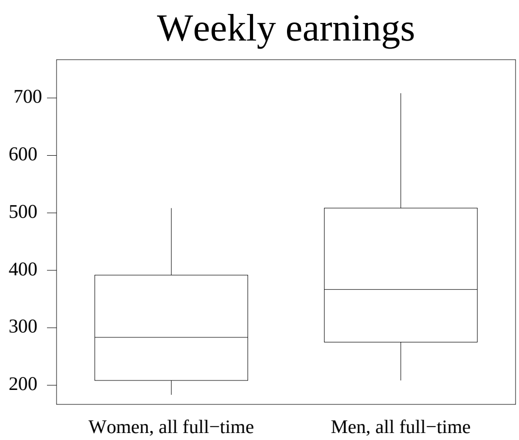

Figure 14: Boxplots Comparing Earning distributions

3.5 Sample Variance

The most commonly used measure of dispersion is the **sample variance.** This is

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

(but see Note 5 below for the formula usually used in calculation).
Clearly then, the variance is measured in the square of units of the original observations. Thus, we also define the **standard deviation**, s, as the square root of the variance, which is then in the same units as an individual observation.

Example 11: Standard deviation of task times
Given that a sample of 5 people take the following times to complete a task, what is the sample standard deviation? $n = 5$; observations 7, 8, 9, 12, 14 seconds.

Solution
The mean is 10 seconds and

$$s^2 = \tfrac{1}{4}\{(-3)^2 + (-2)^2 + (-1)^2 + 2^2 + 4^2\} = \tfrac{34}{4} = 8.5\text{sec}^2 \ ,$$

so $s = 2.92$ seconds.

Example 12: Standard deviation of `grapeweights.txt` and `Systolic-bp.txt`
What are the sample standard deviations for `grapeweights.txt` and `Systolic-bp.txt`?

Solution

`grapeweights.txt`	`Systolic-bp.txt`
$n = 27$	$n = 70$
$\overline{x} = 1007.8$ g	$\overline{x} = 128.0$ mm Hg
$s = 18.32$ g	$s = 18.5$ mm Hg

Notes

1. The rationale for $n - 1$ instead of n draws on general theory indicated later.
2. The variance effectively averages the **squares** of the deviations of individual observations about the mean.
3. The standard deviation is zero when there is no variation in the data, that is, all the values are equal; otherwise it must be strictly positive and it increases as dispersion increases.
4. The standard deviation is not resistant to extreme values. Hence, it is most appropriate as a measure of dispersion when the data show a fairly symmetric pattern of variation.
5. For actual calculations (if necessary) use

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

The term in { } is often used separately in the theory and so has its own notation. Thus

$$s_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = \sum (x - \overline{x})^2 \text{ and so } s^2 = \frac{s_{xx}}{n-1}$$

Consider again the data from Example 11 where n = 5; observations 7, 8, 9, 12, 14 seconds. These yield $\sum x = 50$, $\sum x^2 = 534$ and so

$$\begin{aligned} s_{xx} &= 534 - \frac{(50)^2}{5} = 534 - 500 = 34, \\ s^2 &= \frac{34}{4} \\ s &= 2.92 \text{ seconds} \end{aligned}$$

For large data sets, this formula is quicker than the one used in the definition. However, it is more prone to rounding error since it involves subtracting one large number from another to get an answer that is quite a small number.

6. For grouped data the variance is obtained from the frequency table, approximating the raw data value, using

$$s^2 = \frac{1}{(n-1)} \left\{ \sum_j f_j y_j^2 - \frac{(\sum f_j y_j)^2}{n} \right\}$$

(with notation as in Section 3.1). Note that this is **not** the same as:

$$\left\{ \frac{\sum_j f_j y_j^2}{n} - \left(\frac{\sum f_j y_j}{n} \right)^2 \right\}$$

although for large n the difference is minor.

For `Systolic-bp.txt`

$$\begin{aligned} \sum_j f_j y_j^2 &= 2 \times 94.5^2 + \ldots + 1 \times 184.5^2 = 1,176,948, \\ s^2 &= \frac{1}{69} \left\{ 1,176,948 - \frac{(8985)^2}{70} \right\} = 342.885, \end{aligned}$$

and so $s = 18.5$ mmHg.

7. Calculators often have a button for the standard deviation (if there are two, make sure you know which uses the $(n-1)$ divisor).

8. One use of standard deviation is to compare variability about the mean in different data sets. For example, the IQ is assessed of each student in two samples of students. For each sample the same mean IQ is found, but sample standard deviations are 3.6 and 5.8 respectively. This indicates that in the sample with $s = 3.6$ IQ is less variable, i.e. more tightly grouped around its mean.

9. You obtain the variance and the standard deviation in R using the `var()` and `sd()`.

3.6 Coefficient of Variation

For positive measurement data the **coefficient of variation** is sometimes used as a measure of spread. It is the sample standard deviation divided by the sample mean (i.e. $s/\overline{x}$). Its advantage is that it is dimensionless (it has no units): its value will be the same if we change the unit of measurement.

4 Basic inference for continuous data

Inference usually assumes that $x_1, x_2, ..., x_n$ are observed values from some probability distribution and tries to say things about that distribution.

4.1 The Normal Model

In many situations we assume that the (continuous) observations come from a $N(\mu, \sigma^2)$, and that the observations are independent (random).

- This is a reasonable approximation in many cases — by empirical verification or can be made reasonable by transformation (e.g. $x \to \log x$).
- The theory is simple and well developed.
- Inferences reduce to questions about μ and/or σ^2.

Recall the probability density function of $N(\mu, \sigma^2)$:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\} \quad x \in \mathbb{R}; \quad \mu \in \mathbb{R}, \sigma > 0.$$

The distribution function is

$$P(X \leq x) = F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

where Φ is the $N(0,1)$ distribution function tabulated in Neave 2.1–2.3 (and available in R as `pnorm(`x`)`). There are diagrams to illustrate what is tabulated associated with each of these Tables: make sure you understand them.

4.1.1 Useful distributional results

(See also the separate handout.)

- If Z is $N(0,1)$, then Z^2 is χ^2_1 — see Block A, Exercise 17.
- If $Z_1, Z_2, ..., Z_n$ are independent $N(0,1)$ r.v's, then $Z_1^2 + Z_2^2 + ... + Z_n^2$ is χ^2_n. — used in Block A, Exercise 26. Neave 3.2 gives $\chi^2_{\nu;q}$ = q-quantile of a χ^2_ν distribution, i.e. the point such that $P(X < \chi^2_{\nu;q}) = q$ if $X \sim \chi^2_\nu$. This is obtained in R by `qchisq(`q`,` ν`)`. You fill find a diagram to illustrate what $\chi^2_{\nu;q}$ at the top of the page in Neave's Table 3.2: make sure you understand it.

- If $Z \sim N(0,1)$ is independent of $W \sim \chi^2_\nu$, then

$$T = \frac{Z}{\sqrt{\dfrac{W}{\nu}}} \sim t_\nu,$$

  i.e. has a t-distribution with ν degrees of freedom.
  Neave 3.1 gives $t_{\nu;q} = q$-quantile of a t_ν distribution. Again you will find a diagram which you should ensure you understand.

- If $W_1 \sim \chi^2_{\nu_1}$ and $W_2 \sim \chi^2_{\nu_2}$ with W_1, W_2 independent, then

$$F = \frac{W_1/\nu_1}{W_2/\nu_2} \sim F_{\nu_1,\nu_2},$$

  i.e. has F-distribution with ν_1, ν_2 degrees of freedom.

  Neave Table 3.3 gives $F_{\nu_1,\nu_2;q} = q$-quantile of F_{ν_1,ν_2} distribution — of course there is another diagram.

- If $X_1, X_2, ..., X_n$ are independent $N(\mu, \sigma^2)$, then defining

$$\overline{X} = \frac{1}{n}\sum_1^n X_i \text{ so that } \overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

  and

$$S^2 = \frac{1}{n-1}\sum_1^n (X_i - \overline{X}) \text{ so that } \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

  we find that $\overline{X}$, S^2 are independent.

  Hence, from above,

$$T = \frac{\overline{X} - \mu}{\sqrt{\dfrac{S^2}{n}}} \sim t_{n-1}.$$

  In R we use the `pt` and `qt` functions to find probabilities and associated quantiles (or percentage points).

4.2 Inferences about μ and σ

For example,

- What values of μ are consistent with the data? — **confidence interval** or **point estimation**.
- Is a specified value μ_0 consistent with the data? — **hypothesis test**

4.2.1 Point estimates

The mean is estimated by

$$\hat{\mu} = \overline{X} = \frac{1}{n}\sum X_i.$$

This is unbiased (i.e. has its expectation equal to what is being estimated) since (see Block A §4.7.1)

$$E(\overline{X}) = \mu; \qquad\qquad \text{also, } \operatorname{Var}(\overline{X}) = \frac{\sigma^2}{n}.$$

The standard error of $\overline{X}$, s.e.$(\overline{X})$, is the square root of $\operatorname{Var}(\overline{X})$ and so is $\sigma/\sqrt{n}$.

The variance is estimated by

$$\hat{\sigma}^2 = S^2 = \frac{1}{n-1}\sum_1^n (X_i - \overline{X})^2.$$

This is unbiased since

$$E(S^2) = \sigma^2 \text{ as } E\left\{\frac{(n-1)S^2}{\sigma^2}\right\} = n-1.$$

Note $E(S) \neq \sigma$: i.e. S is a biased estimator of σ.

The estimated standard error[3] (e.s.e) of $\overline{X}$, sometimes written e.s.e.$(\overline{X})$ is

$$\sqrt{\frac{s^2}{n}}.$$

4.2.2 Confidence interval for μ (see also Block A §5.1)

Since

$$\frac{\overline{X} - \mu}{\sqrt{\dfrac{S^2}{n}}} \sim t_{n-1},$$

$$P\left\{-t_{n-1;1-\frac{\alpha}{2}} < \frac{\overline{X}-\mu}{\sqrt{\frac{S^2}{n}}} < t_{n-1;1-\frac{\alpha}{2}}\right\} = 1-\alpha.$$

which is equivalent to

$$P\left\{\overline{X} - t_{n-1;1-\frac{\alpha}{2}}\sqrt{\frac{S^2}{n}} < \mu < \overline{X} + t_{n-1;1-\frac{\alpha}{2}}\sqrt{\frac{S^2}{n}}\right\} = 1-\alpha.$$

[NB. This is a probability statement about $\overline{X}$, S^2 — not μ]

[3] confusingly, sometimes just called the standard error

This means that

$$\left(\overline{x} - t_{n-1;1-\frac{\alpha}{2}}\sqrt{\frac{s^2}{n}}, \overline{x} + t_{n-1;1-\frac{\alpha}{2}}\sqrt{\frac{s^2}{n}}\right)$$

is a $100(1-\alpha)$ % CI for μ.

Notes:

(a) **Interpretation**. In repeated sampling, a proportion $1-\alpha$ of such intervals will contain μ.

(b) The form is $\hat{\theta} \pm t_{\nu;1-\frac{\alpha}{2}} \times e.s.e.(\hat{\theta})$

4.2.3 Hypothesis test on μ

$H_0 : \mu = \mu_0$ v. $H_1 : \mu \neq \mu_0$.

Under H_1 one expects

$$t = \left|\frac{\overline{x} - \mu_0}{\sqrt{\frac{S^2}{n}}}\right|$$

to be large (but small under H_0).

The significance probability or p-value[4] is $P(|T| \geq t)$ for $T \sim t_{n-1}$ under H_0. This is the probability of observing something as extreme as, or more extreme than, what has been found in the actual data set. The smaller the p-value, the more evidence against H_0.

Example 1. The specification for a can of beans is that the beans should weigh 400 gm. Twenty cans provide the following contents:

404	403	391	394	402	394	401	392	394	402
401	398	392	393	405	398	395	402	406	404

Is the specification being met?

Assume $N(\mu, \sigma^2)$, $H_0 : \mu = 400$ $v.$ $H_1 : \mu \neq 400$
$\overline{x} = 398.55$, $n = 20$, $s = 4.989$, $s/\sqrt{n} = 1.116$

$$t = \left|\frac{\overline{x} - 400}{s/\sqrt{n}}\right| = 1.30 \text{ so } p = P(|T| \geq 1.30).$$

Tabulated values: $\dfrac{t_{19;0.85} \quad t_{19;0.90}}{1.066 \quad 1.328} \Longrightarrow 0.20 < p < 0.30.$ [In fact, 'exact' from R: $p = 0.21$.]

No reason to doubt H_0, ($p = 0.21$); 95% CI for $\mu : (396.21, 400.89)$.
[Note $\mu_0 = 400$ lies within 95% CI — as expected.]

[4] See also Block A §5.3

Notes

(a) We only need the p-value approximately, or within an interval

$$\text{e.g. } p \approx 0.06 \quad \text{or} \quad 0.05 < p < 0.1.$$

(b) **Conventional** interpretation:

$p > 0.10$	Data consistent with H_0
$0.05 < p < 0.10$	Perhaps weak evidence against H_0 — maybe more data needed!!
$0.01 < p < 0.05$	Some evidence against H_0
$p < 0.01$	Strong evidence against H_0
$p < 0.001$	Very strong evidence against H_0

(c) If there is evidence against H_0, it is vital to say how/why; i.e. to elaborate. In answer to a real problem (or in assessed work!) it is vital to set conclusions in context. The 'answer' is never '$p < 0.01$, reject H_0', but something like (in the context of Example 4) 'There is evidence ($p = 0.045$) to reject the hypothesis that students weigh the same, on average, before and after a semester in hall. It appears that students tend to weigh less afterwards by an average of 3.2 lbs, 95% CI $(0.09, 6.31)$ lbs.'

Be particularly careful to phrase the null hypothesis, so that any rejection does not imply a 1-sided test was performed, if not the case. For example, do **not** contract the above to say 'there was some evidence ($p = 0.045$) that students weighed less after a semester in hall'.

Full details have **not** always been given in these notes, exercises and examples.

(d) For one sided alternatives, use the one-sided version of the test statistic.

E.g. for $H_0: \ \mu = \mu_0$ v $H_1: \ \mu > \mu_0$,

$$t = \frac{\overline{x} - \mu_0}{\sqrt{\frac{s^2}{n}}}.$$

Here $p = P(T > t)$ where T is t_{n-1} under H_0.

In R the command is `t.test`.

4.2.4 Inferences for σ^2

We base CI and tests on fact that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}, \qquad \text{i.e. } \frac{\Sigma(X_i - \overline{X})^2}{\sigma^2} \sim \chi^2_{n-1}$$

Example 2. In Example 1 say that the standard deviation is supposed to be (at most) 4. Is the observed value too high for this to be credible?

$H_0 : \sigma^2 = 16$ (or $\sigma^2 \leq 16$)
$H_1 : \sigma^2 > 16$
(1-sided test more appropriate).

Under H_1 expect $(n-1)s^2/\sigma^2$ to be relatively large. Here

$$\frac{(n-1)s^2}{{\sigma_0}^2} = \frac{472.95}{16} = 29.56.$$

Thus the significance probability is $p = P(\chi^2_{19} > 29.56)$

$$\text{Tabulated values: } \frac{\chi^2_{19;0.925} \quad \chi^2_{19;0.950}}{28.46 \qquad 30.14} \Longrightarrow 0.05 < p < 0.075.$$

['Exact' $p = 1 - 0.9423 = 0.058$]. Perhaps weak evidence against H_0.

4.3 Two sample problems — separate samples

$$\text{Formulation: } \left.\begin{array}{rcl} X_1, X_2, \ldots, X_{n_1} & \sim & N(\mu_1, \sigma_1^2) \\ Y_1, Y_2, \ldots, Y_{n_2} & \sim & N(\mu_2, \sigma_2^2) \end{array}\right\} \text{all independent.}$$

Interest lies in difference $\mu_1 - \mu_2$ (for means) or ratio (for scale parameters) σ_1^2/σ_2^2, leading to CIs and tests. These will be based on the sample statistics

$$\overline{x} = \frac{1}{n_1}\sum x_i, \qquad s_1^2 = \frac{\sum(x_i - \overline{x})^2}{n_1 - 1},$$

$$\overline{y} = \frac{1}{n_2}\sum y_i, \qquad s_2^2 = \frac{\sum(y_i - \overline{y})^2}{n_2 - 1}.$$

4.3.1 Comparing variances

Base tests and CI's on fact that

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}.$$

Thus to test $H_0 : \sigma_1 = \sigma_2$ v $H_1 : \sigma_1 \neq \sigma_2$, use test statistic $f = s_1^2/s_2^2$. Values "well away from 1" are more likely under H_1.

Note: Neave's tables only give upper % points, so arrange $f > 1$ (i.e. larger s^2 in numerator). This is automatically handled in packages.

4.3.2 Comparing means

Use

$$\left.\begin{array}{l} \overline{X} \sim N\left(\mu_1, \dfrac{\sigma_1^2}{n_1}\right) \\ \overline{Y} \sim N\left(\mu_2, \dfrac{\sigma_2}{n_2}\right) \end{array}\right\} \Longrightarrow \overline{X} - \overline{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

To test

$$H_0 : \mu_1 = \mu_2, \quad v \quad H_1 : \mu_1 \neq \mu_2$$

Suppose $\sigma_1^2 = \sigma_2^2$, then

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{S^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$$

where

$$S^2 = \frac{(n-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$

is a **pooled estimator** of σ^2, which satisfies (when $\sigma_1 = \sigma_2$)

$$\frac{(n_1+n_2-2)S^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}.$$

So use test statistic

$$T = \frac{(\overline{X}-\overline{Y})}{\sqrt{S^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2} \quad \text{under } H_0.$$

Example 3. Ten cows were milked with, and ten cows without, background music, all the cows being kept under the same conditions otherwise. Over a period of a week, the following were the yields in gallons.

Cows with music:	15	18	14	12	19	13	15	15	11	17
Cows without music:	14	19	12	13	10	17	12	10	8	17

Does music influence yield?

$$\begin{array}{llll} n_1 = 10 & \overline{x} = 14.9 & s_1^2 = 6.5444 & s_1 = 2.558 \\ n_2 = 10 & \overline{y} = 13.2 & s_2^2 = 12.6222 & s_2 = 3.55 \end{array}$$

Variances equal?

$$H_0 : \sigma_1^2 = \sigma_2^2; \quad H_1 : \sigma_1^2 \neq \sigma_2^2.$$

Take $f = s_2^2/s_1^2 = 1.929$; $p-$value $= 2 \times P(F_{9,9} > 1.929)$. From Neave: $F_{9,9;0.9} = 2.44 \implies p > 0.20$ [‘exact’ $p = 2 \times 0.1710 = 0.342$]

i.e. no evidence to suggest variances are unequal.

Means equal?

Assuming $\sigma_1^2 = \sigma_2^2$: $H_0 : \mu_1 = \mu_2$; $H_1 : \mu_1 \neq \mu_2$.

Estimate of pooled variance is $s^2 = 9.5833$ (just the simple average of $s_1^2 = 6.5444$ and $s_2^2 = 12.6222$ in this case, since $n_1 = n_2 = 10$).

$$t = \frac{\overline{x}-\overline{y}}{\sqrt{s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = 1.23; \quad p\text{-value} = P(|T| > 1.23).$$

$$\begin{array}{llll} \text{Tabulated values:} & t_{18;0.85} & t_{18;0.90} & \\ & 1.067 & 1.33 & \implies p > 2 \times 0.10 = 0.20 \end{array}$$

[‘Exact’ $p = 2 \times (1 - 0.8827) = 0.23$.] Thus no evidence against equality of means ($p > 0.20$). i.e. no reason, based on this data, to conclude that music influences yield.

Note: One-sided tests may also be appropriate sometimes.

If there is any real doubt that σ_1^2 and σ_2^2 are not similar, then use

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}} \approx t_\rho$$

for an approximate test, where

$$\rho = \left[\frac{\left[\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right]^2}{\dfrac{1}{n_1 - 1}\left(\dfrac{s_1^2}{n_1}\right)^2 + \dfrac{1}{n_2 - 1}\left(\dfrac{s_2^2}{n_2}\right)^2} \right]$$

with

$$\min(n_1, n_2) - 1 \leq \rho \leq n_1 + n_2 - 2.$$

[Use of $\rho = \min(n_1, n_2) - 1$ gives a **conservative** test — i.e. the p-values are larger than the exact ones, and C.I's wider than the exact ones.]

In R use `t.test` with appropriate options. By default, R uses the approximate test allowing for $\sigma_1^2 \neq \sigma_2^2$. This is safe, and wise, since it is little different from the pooled test when the variances are roughly equal and provides protection against them being unexpectedly unequal.

4.4 Two sample problems — paired samples

Consider Example 4 below.

Here we have matched–pair data, which is clearly different from a random sample of 10 before and **another** random sample of 10 after. We can see the difference as follows.

Observation			Difference
1	X_1	Y_1	$D_1 = X_1 - Y_1$
2	X_2	Y_2	$D_2 = X_2 - Y_2$
$\vdots$			
n	X_n	Y_n	$D_n = X_n - Y_n$

Formally we could write

$$X_i = \alpha_i + \beta_1 + \epsilon_i$$
$$Y_i = \alpha_i + \beta_2 + \eta_i$$

where α_i is effect of individual i and β_1, β_2 are effects of treatments 1, 2, respectively. Then

$$\begin{aligned} D_i = X_i - Y_i &= (\beta_1 - \beta_2) + (\epsilon_i - \eta_i) \\ &= \beta_1 - \beta_2 + \zeta_i \qquad \text{where } \zeta_i \text{ is error.} \end{aligned}$$

If we assume $D_i \sim N(\beta_1 - \beta_2, \sigma^2)$, then this means that CIs + tests on $\beta_1 - \beta_2$ are as in 1-sample t-test.

Notes

(i) Suppose Var $(\alpha_i) = \sigma_\alpha^2$ and Var$(\epsilon_i) =$ Var $(\eta_i) = \tau^2$: the latter implies that $\sigma^2 = 2\tau^2$. Then

$$\begin{aligned}\text{Var}(\overline{X} - \overline{Y}) &= \text{Var}(\overline{X}) + \text{Var}(\overline{Y}) \qquad \text{if not matched pairs} \\ &= 2\left(\frac{\sigma_\alpha^2 + \frac{1}{2}\sigma^2}{n}\right).\end{aligned}$$

Whereas here

$$\text{Var}(\overline{D}) = \frac{\sigma^2}{n} \leq \frac{\sigma^2 + 2\sigma_\alpha^2}{n}$$

i.e. n differences for inferences on $\beta_1 - \beta_2$ have smaller variances.

(ii) Minor drawback is reduction in d.f. from $2(n-1)$ to $n-1$.

(iii) Basis of blocking in experimental design.

Example 4. Below are the weights (in lbs) of 10 students before and after a semester in residence at a University hall of residence.

Student:	1	2	3	4	5	6	7	8	9	10
Before x_i	140	153	156	148	167	134	190	182	178	164
After y_i	135	155	153	144	168	130	180	186	171	158
d_i	5	-2	3	4	-1	4	10	-4	7	6

$$\overline{d} = 3.2 \quad s_d = 4.3410 \quad \frac{s_d}{\sqrt{n}} = 1.373.$$

$H_0 : \beta_1 = \beta_2$ v. $H_1 : \beta_1 \neq \beta_2$

$$t = \frac{\overline{d}}{s_d/\sqrt{n}} = 2.33, \quad p = P(\mid T \mid > 2.33).$$

Tables $\Rightarrow p < 0.05$ ($p = 0.045$), i.e. some evidence to reject the hypothesis that the mean difference is zero. Note that Before > After from values of d_i [or look at CI for $\beta_1 - \beta_2$, eg. 95% CI is (0.09, 6.31)]

Note: Use of two-sample test (which is **incorrect**) yields $p = 0.70$!!

4.5 Effects of departures from assumptions

4.5.1 Inferences on μ

Non-normality. Even if the X_i's are not normal, by the Central Limit Theorem,

$$\frac{\overline{X} - \mu}{\sqrt{\dfrac{\sigma^2}{n}}} \text{ is approx } N(0,1).$$

Also $S^2 \to \sigma^2$ as $n \to \infty$ by Law of Large Numbers. Thus

$$T = \frac{\overline{X} - \mu}{\sqrt{\dfrac{S^2}{n}}} \approx N(0,1) \quad \text{for large } n$$

and, of course, then $N(0,1) \approx t_{n-1}$.

Usually OK for $n \geq 40$ if no obvious outliers.

Independence. Dependence can lead to incorrect inference — even in large samples. For example, suppose

$$\begin{aligned} \text{Corr}\,(X_i, X_{i+1}) &= \rho \qquad && \text{for } \; i = 1, ..., n-1 \\ \text{Corr}\,(X_i, X_j) &= 0 && \text{otherwise } \; (i \neq j) \end{aligned}$$

Then

$$\begin{aligned} \text{Var}(\overline{X}) &= \frac{\sigma^2}{n}\left\{1 + 2\rho\left(1 - \frac{1}{n}\right)\right\} \\ E(S^2) &= \sigma^2\left\{1 - \frac{2\rho}{n} + \;\text{ terms in }\; \frac{1}{n^2}, ...\right\}. \end{aligned}$$

Therefore for large n

$$T = \frac{\overline{X} - \mu}{\sqrt{\dfrac{S^2}{n}}} \approx N(0, 1+2\rho) \;\text{ and not }\; N(0,1).$$

So, considering $p = P(|T| > 1.96)$,

ρ	-0.3	-0.2	-0.1	0	0.1	0.2	0.3
p	0.002	0.011	0.028	0.05	0.074	0.098	0.12

Generally inferences on mean are robust to non-normality but not to dependence.

If $n_1 = n_2$ in 2-sample tests, moderate departures from $\sigma_1 = \sigma_2$ have little effect.

4.5.2 Inferences on σ^2

Non-normality.

$$\begin{aligned} E(S^2) &= \sigma^2 \\ \text{but} \quad \text{Var}\left(S^2\right) &= \frac{\sigma^4}{n-1}\left\{2 + \frac{n-1}{n}\gamma_2\right\} \end{aligned}$$

where γ_2 is the **coefficient of kurtosis** given by

$$\gamma_2 = \frac{E(X-\mu)^4}{\sigma^4} - 3 \qquad [= 0 \text{ for normal }].$$

If $\gamma_2 \neq 0$, then $(n-1)S^2/\sigma^2$ does **not** have a χ^2_{n-1} distribution. In the case of 2 samples, non-normality has so serious an effect on the F-test as to throw doubts on the wisdom of using it!

5 Basic inference for discrete data

5.1 Fundamentals

5.1.1 Discrete Data Examples

D1 Sex ratio amongst first children in India (Pakrasi-Habler, 1971) males:females 40467:32335 (actual numbers). [c.f. Sex ratio worldwide 100–110% (male/female).]

D2 Number of times a traveller was stopped by immigration officers at Fishguard (O'Dowd, 1982)

		Stopped	Not Stopped	
CND	Yes	4	2	6
Badge	No	1	5	6

D3 Blood group and social class amongst blood donors in Yorkshire (Nature, 1983)

Class	Blood Group		
	A	not A	
I-II	257	297	
III-V	866	1228	
			2648

D4 French suicides by day of week (Durkheim, 1897)

Mon	Tues	Wed	Thurs	Fri	Sat	Sun
1001	1035	982	1033	905	737	894

D5 Radioactive disintegration (Rutherford and Geiger, 1910)

No. of particles emitted in a period k	No. of 7.5 sec.periods in which this no. was observed O_k
0	57
1	203
2	383
3	525
4	532
5	408
6	273
7	139
8	45
9	27
10	10
11	4
12	0
13	1
$\geq$14	1
	2608

D6 Plums: propagation of root stock from cuttings

	Time of Planting			
	At Once		In Spring	
Condition	Long	Short	Long	Short
Alive	156	107	84	31
Dead	84	133	156	209
	240	240	240	240

D7 Number of boys in 240 American 4-child families (Rao et al, 1973)

No. of boys	0	1	2	3	4
Frequency	13	61	94	60	12

5.1.2 Some discrete distributions

Binomial (See Block A §2.4.1.) X : no. of successes in n trials, θ = probability of success

$$X \sim Bi(n, \theta) \qquad p(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

$$E(X) = n\theta \qquad \mathrm{Var}(X) = n\theta(1-\theta)$$

Data set D1, sex ratio; X = number of males

$$\text{Here } X \sim Bi(72802, \theta) \text{ and so } \hat{\theta} = \frac{40467}{72802}.$$

$$\text{The sex ratio} = \frac{\theta}{1-\theta}.$$

The question of interest is,

$$\text{is } \frac{\theta}{1-\theta} \simeq 1.1, \quad \text{say?}$$

Data set D2, CND Badge:

$$2 \text{ populations } Bi(6, \theta_1),\ Bi(6, \theta_2)$$

$$\text{Question of interest: is } \theta_1 = \theta_2?$$

Data set D6:

$$\begin{array}{lll} X_{ij} & : & \text{no. alive out of 240, } \quad i = \text{time of planting,} \quad j = \text{length} \\ X_{ij} & \sim & Bi(240, \theta_{ij}) \end{array}$$

Perhaps model θ_{ij} — e.g. $\log \theta_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$.

Data set D7, number of boys in US 4 child families:

$$\text{Number of boys } X \sim Bi(4, \theta)? \text{ Goodness of fit?}$$

Multinomial (See Block A §3.6.1) k possible classes; $P(\text{class } i) = \theta_i \quad \sum \theta_i = 1$. For n individuals, R_i = number in class $i \quad (i = 1, 2, \ldots, k)$. Then

$$\begin{aligned}
(R_1, R_2, \ldots, R_k) &\sim \text{Multi}\,(n;\ \theta_1, \ldots, \theta_k) \\
p(\mathbf{r}) &= \frac{n!}{r_1! \ldots r_n!} \theta_1^{r_1} \ldots \theta_k^{r_k} \left(\sum r_i = n\right) \\
E\,(R_i) &= n\theta_i \\
\text{Var}\,(R_i) &= n\theta_i\,(1 - \theta_i) \\
\text{Cov}\,(R_i, R_j) &= -n\theta_i\theta_j\ (i \neq j)
\end{aligned}$$

Data set D3: 1 sample, 2 binary response categories $\Rightarrow$ 4 classes; $\theta_{ij} = P(\text{individual is in row } i \text{ and col } j)$

$$(R_{11}, R_{12}, R_{21}, R_{22}) \sim \text{Multi}(2648; \theta_{11}, \theta_{12}, \theta_{21}, \theta_{22}).$$

Question of interest: Are class and blood group independent? i.e. is $\theta_{ij} = \theta_{i\cdot} \times \theta_{\cdot j}$ for all i, j.

Poisson (See Block A §2.4.2)

$$X \sim Po(\mu) \qquad\qquad p(x) = \frac{e^{-\mu}\mu^x}{x!} \quad x = 0, 1, \ldots$$

$$E(X) = \mu \qquad\qquad \text{Var}(X) = \mu.$$

Commonly occurs in conjunction with the Poisson Process where random events arise with rate λ.In this case the number of events in interval of length $t \sim Po(\lambda t)$.

Data set D4: Rate λ_i for day i; numbers $Po(\lambda_i)$. Question of interest: are λ_i equal?

Data set D5: Question of interest: would $Po(\mu)$ be a good model, i.e. 'Goodness-of-fit'.

5.1.3 Some distributional properties

- If $X \sim Bi(n, \theta)$, then $X \simeq N(n\theta, n\theta(1-\theta))$ provided $n\theta, n(1-\theta)$ not too small (say $n\theta(1-\theta) \geq 10$). Then for *integer* r

$$\begin{aligned}
P(X \leq r) &\approx P\left(\hat{X} \leq r + \frac{1}{2}\right) \text{ where } \hat{X} \sim N(n\theta, n\theta(1-\theta)) \\
&= \Phi\left(\frac{r + \frac{1}{2} - n\theta}{\sqrt{n\theta(1-\theta)}}\right)
\end{aligned}$$

  The 1/2 here is called a **continuity correction**. Similarly

$$\begin{aligned}
P(X \geq r) &\approx P(\hat{X} \geq r - \frac{1}{2}) \\
P(X < r) &\approx P(\hat{X} \leq r - \frac{1}{2}), \quad \text{etc.}
\end{aligned}$$

- If $X \sim Po(\mu)$ then $X \simeq N(\mu, \mu)$ if μ not too small, say $\mu > 5$. Again, for integer r you can use a continuity correction as for the binomial.

5.2 Inference for a binomial proportion

Suppose $X \sim Bi(n, \theta)$.

$$\begin{aligned}
\text{Point estimator} \quad \hat{\theta} &= \frac{X}{n} \\
E(\hat{\theta}) &= \theta \qquad \text{i.e. unbiased} \\
\text{Var}(\hat{\theta}) &= \frac{\theta(1-\theta)}{n}
\end{aligned}$$

so

$$\text{e.s.e.}(\hat{\theta}) = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}.$$

Confidence interval for θ Use

$$P\left(\left|\frac{X - n\theta}{\sqrt{n\theta(1-\theta)}}\right| < 1.96\right) \simeq 0.95,$$

to obtain an approximate 95% CI. There are three approaches:

(i) substitute $\hat{\theta} = x/n$ for θ in the variance to get

$$\frac{x}{n} \pm 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} \qquad [\text{i.e.} \ \ \hat{\theta} \pm 1.96\text{e.s.e.}(\hat{\theta})];$$

(ii) solve the quadratic $(x - n\theta)^2 < 1.96^2 n\theta(1-\theta)$ (for θ);

(iii) use Chart 1.2 in Neave.

Test for θ

$H_0 : \theta = \theta_0 \quad v \quad H_1 : \theta \neq \theta_0$. Test statistic: $\mid (x/n) - \theta_0 \mid$ large if H_1 true. Carry out the test in one of the following ways.

(i) Use the fact that approximately

$$\frac{\dfrac{X}{n} - \theta_0}{\sqrt{\dfrac{\theta_0(1-\theta_0)}{n}}} \sim N(0, 1)$$

under H_0 to obtain significance probability.

(ii) Use the exact test retaining $X \sim Bi(n, \theta)$. Here we find the p-value as

$$\begin{aligned}
p &= 2P(X \geq x) \ \text{ for } \ x > n\theta_0 \\
&= 2P(X \leq x) \ \text{ for } \ x < n\theta_0
\end{aligned}$$

The values can be found from tables of $Bi(n, \theta)$ distribution function, e.g. Neave 1.1

Example 5. Data set D1 for discrete data:

$$X = \#\text{males} \sim Bi(72802, \theta) \qquad \hat{\theta} = 0.5559,\ e.s.e.\ (\hat{\theta}) = 0.001842.$$

The sex ratio $= \frac{\theta}{1-\theta} \times 100$. Estimate this: $\frac{\hat{\theta}}{1-\hat{\theta}} \times 100 = 125\%$.

For a CI for the sex ratio use CI for $\hat{\theta}$ and convert using that θ to $\theta/(1-\theta)$ is a 1-1 transformation:

CI for θ	$0.5559 \pm 1.96 \times 0.001842$	$\rightarrow$	$(0.5527, 0.5595)$
transforms	via $\frac{\hat{\theta}}{1-\hat{\theta}} \times 100\%$:	$\rightarrow$	$(123, 127)$

Test

$$H_0 : \theta_0 = \frac{110}{100+110} \quad v \quad H_1 : \theta_0 \neq \frac{110}{100+110} \quad [\text{sex ratio} = 110\% \text{ worldwide, say}]$$

ie $\theta_0 = 0.5238$.

So p-value

$$\begin{aligned} p &= P\left(|Z| > \frac{0.5559 - 0.5238}{\sqrt{\dfrac{0.5238 \times 0.4762}{72802}}} \middle| Z \sim N(0,1) \right) \\ &= P(|Z| > 17.34) \\ &= 0.000.......!! \end{aligned}$$

i.e. reject H_0 — clearly sex ratio $\gg 110\%$

5.3 Comparing two binomial proportions

Suppose $X_1 \sim Bi(n_1, \theta_1)$, $X_2 \sim Bi(n_2, \theta_2)$. We wish to test $\quad H_0 : \theta_1 = \theta_2$.

5.3.1 Approximate test

The natural point estimators are $\hat{\theta}_1 = \frac{X_1}{n_1}$ and $\hat{\theta}_2 = \frac{X_2}{n_2}$.

Therefore, as in §4.3.2,

$$\begin{aligned} \hat{\theta}_1 - \hat{\theta}_2 &\sim N\left(\theta_1 - \theta_2, \frac{\theta_1(1-\theta_1)}{n_1} + \frac{\theta_2(1-\theta_2)}{n_2}\right) \\ &\sim N\left(0, \theta(1-\theta)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right) \quad \text{if} \quad \theta_1 = \theta_2 = \theta. \end{aligned}$$

Test uses

$$Z = \frac{\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{\hat{\theta}(1-\hat{\theta})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \sim N(0,1) \quad \text{under} \quad H_0$$

where $\hat{\theta} = (X_1 + X_2)/(n_1 + n_2)$ is a pooled estimator of θ since $X_1 + X_2 \sim Bi(n_1 + n_2, \theta)$ under H_0.

Similarly, for CI use

$$\hat{\theta}_1 - \hat{\theta}_2 \sim N\left(\theta_1 - \theta_2, \quad \frac{\hat{\theta}_1\left(1 - \hat{\theta}_1\right)}{n_1} + \frac{\hat{\theta}_2\left(1 - \hat{\theta}_2\right)}{n_2}\right).$$

Note that, as usual, the assumptions of H_0 are not employed here.

5.4 Goodness-of-fit tests

Suppose $(R_1, R_2, ..., R_k) \sim \text{Multi}\,(n;\ \theta_1, ..., \theta_k)$

We want to test if θ_i take specified values/form

$$\begin{array}{lcl} H_0 & : & \theta_i = \theta_{i0} \text{ for all } \ i. \\ H_1 & : & \text{some } \theta_i \text{ not as specified in } H_0 \end{array}$$

So we expect R_i to be 'close to' $e_i = n\theta_{i0}$ under H_0. Use as test statistic **Pearson's chi-square**:

$$X^2 = \sum_i \frac{(r_i - e_i)^2}{e_i} = \frac{(\text{Obs}-\text{Exp})^2}{\text{Exp}}.$$

Under H_0, $\;X^2 \simeq \chi^2_{k-1}$. Large values of X^2 are critical of H_0.

Notes:

(a) Alternative test statistic : deviance

$$D = 2\sum_i r_i \log\left(\frac{r_i}{e_i}\right) \sim \chi^2_{k-1} \text{ under } H_0$$

Tests are asymptotically equivalent.

(b) If H_0 not fully specified, then use $e_i = n\hat{\theta}_i$ under H_0.
E.g. in Example 6, $H_0 : X \sim Bi(4, \theta)$ so

$$\theta_{i+1} = \binom{4}{i}\theta^i(1-\theta)^{4-i} \text{ for } i = 0, 1, 2, 3, 4.$$

Calculate $\hat{\theta}_i$ using an efficient estimator of θ, e.g. m.l.e., then X^2 or $D \sim \chi^2_{k-1-q}$ under H_0 where q is number of parameters estimated.

(c) Provides a simple test for goodness-of-fit of any distribution.

For continuous case split range into k non-overlapping/exhaustive intervals and count number of observations in each to obtain $R_1, R_2, ..., R_k$. Find the e_i from the postulated distribution function. Power of the test increases as k increases, but can be low for continuous distributions. Other tests (e.g. Kolmogorov-Smirnov) are often better.

(d) The test is based on the asymptotic distribution of X^2. The asymptotic results are usually OK if, for $k > 4$, $\quad e_i > 1 \;\; \forall i$ and 80% of $e_i \geq 5$.

It may be acceptable to combine (usually neighbouring) classes to ensure applicability of the χ^2 approximation, but this will affect the hypotheses it is possible to test.

(e) The **Pearson residual** $= (r_i - e_i)/\sqrt{e_i} \sim$ Normal for large e_i.

If we reject H_0, look for large residuals, i.e. large contributions to X^2.

Example 6. Data set D7 for discrete data

No. of boys in 4-child family $Bi(4, \theta)$ (assuming independence)

$H_0 : \theta = \frac{1}{2} \; v \; H_1 : \theta \neq \frac{1}{2}$

So in the above notation $\Rightarrow \theta_i = \binom{4}{i}\theta^i(1-\theta)^{4-i}$ and $\theta_{i0} = \binom{4}{i}\frac{1}{2}^4$, and $k = 5$.

	R_1	R_2	R_3	R_4	R_5	
No. of boys	0	1	2	3	4	
Frequency	13	61	94	60	12	$n = 240$
Expected $e_i = n\theta_{i0}$	15	60	90	60	15	

$$X^2 = \frac{(13-15)^2}{15} + \frac{(61-60)^2}{60} + \ldots + \frac{(12-15)^2}{15} \quad = 1.06$$

Thus $p = P(\chi^2_4 > 1.06)$, and so $0.90 < p < 0.925$, i.e. no evidence to reject $\theta = \frac{1}{2}$

We might also want to test the hypothesis H_0 : no. of boys $\sim Bi(4, \theta)$ for unspecified θ .

Sample of size 240 with 13 0's, 61 1's, etc. leads to an estimate of θ of

$$\hat{\theta} = \frac{13 \times 0 + 61 \times 1 + \ldots + 12 \times 4}{240 \times 4} = 0.4969.$$

This leads to the 'expected numbers':

i	0	1	2	3	4
$e_i =$	15.38	60.74	89.99	59.26	14.63

which then give $X^2 = 1.03$ which is to be compared with a χ^2_3 (we have 3 degrees of freedom here, having estimated one parameter). The conclusion is unchanged.

5.5 χ^2-test for independence in a contingency table

In data set D3 we had a 2×2 contingency table, each individual classified by two factors into one of 4 groups. This extends to an $r \times c$ contingency table. We proceed as in §5.4

$$(R_{ij}) \sim \text{Multi}\left(n; \{\theta_{ij}\}\right); \quad \sum_i \sum_j R_{ij} = n; \quad \sum_i \sum_j \theta_{ij} = 1.$$

Under H_0: factors act independently, then

$$\theta_{ij} = \theta_{i\cdot} \times \theta_{\cdot j} \text{ for all } i,j \text{ where } \theta_{i\cdot} = \sum_j \theta_{ij}, \quad \theta_{\cdot j} = \sum_i \theta_{ij}.$$

Estimates are:

$$\hat{\theta}_{i\cdot} = \frac{r_{i\cdot}}{n}; \qquad \hat{\theta}_{\cdot j}\frac{r_{\cdot j}}{n}$$

So the expected cell counts are

$$e_{ij} = n \times \frac{r_{i\cdot}}{n} \times \frac{r_{\cdot j}}{n} = \frac{r_{i\cdot} r_{\cdot j}}{n}$$

So suitable text statistics are:

$$X^2 = \sum_i \sum_j \frac{(r_{ij} - e_{ij})^2}{e_{ij}} \text{ or } D = 2 \sum \sum r_{ij} \log \left[\frac{r_{ij}}{e_{ij}}\right].$$

Under H_0, X^2 or $D \sim \chi^2_{(r-1)(c-1)}$. Note that $(r-1)(c-1) = rc - 1 - (r-1) - (c-1)$.

We can now see this simply as an extension of goodness of fit tests, §5.4.

Example 7. Data set D3 for discrete data H_0 : blood group and class act independently (ie not associated).

		A	not A	
r_{ij} :	Class I, II	257	297	554
	III-V	866	1228	2094
		1123	1525	2648

$$e_{ij} : \begin{array}{cc} 234.95 & 319.05 \\ 888.05 & 1205.95 \end{array}$$

$$\begin{aligned} X^2 &= \sum \frac{(r_{ij} - e_{ij})^2}{e_{ij}} \\ &= 2.069 + 1.524 + 0.547 + 0.403 = 4.54 \end{aligned}$$

This is to be compared with χ^2_1. So $0.025 < p < 0.05$, and there is some evidence to suggest association — more (I, II and A) than expected, etc.

Here you might also apply the ideas in Block A §3.7. The estimate of the log odds ratio is

$$\log\left(\frac{257 \times 1228}{297 \times 866}\right) = 0.2046\ldots$$

with estimated standard error

$$\sqrt{\frac{1}{257} + \frac{1}{297} + \frac{1}{866} + \frac{1}{1228}} = 0.0960\ldots$$

so a test, based on rough normality, of whether the log odds ratio is different from zero has p-value

$$2(1 - \Phi\left(\frac{0.2046}{0.0960}\right)) \approx 0.03.$$

Thus this alternative approach gives a fairly similar p-value to the usual χ^2 test.

5.6 χ^2-test for homogeneity

The data here appear similar, but are in fact r **samples** of c categories, each multinomial.

e.g. data set D6 (condensed)

				Alive	Dead		Number alive
Sample	1	:	At once, long	156	84	240	Multi$(240; \theta_{11}, \theta_{12})$
	2	:	At once, short	107	133	240	Multi$(240; \theta_{21}, \theta_{22})$
	3	:	In Spring, long	84	156	240	Multi$(240; \theta_{31}, \theta_{32})$
	4	:	In Spring, short	31	209	240	Multi$(240; \theta_{41}, \theta_{42})$

$H_0 : \theta_{11} = \theta_{21} = \theta_{31} = \theta_{41}$ (and obviously $\theta_{12} = \ldots = \theta_{42}$ here since our Multi are Bi) i.e. **homogeneity of 4 populations**.

More generally for $r \times c$

$$H_0 : \theta_{ij} = \frac{\theta_{\cdot j}}{r} \quad \forall i, j.$$

Note

$$\theta_{i\cdot} = \sum_j \theta_{ij} \text{ for all } i \text{ and } \sum_i \sum_j \theta_{ij} = r$$

Continue as in §5.5. Estimates

$$\hat{\theta}_{ij} = \left(\widehat{\frac{\theta_{\cdot j}}{r}} \right) = \frac{r_{\cdot j}}{n}.$$

Therefore, $e_{ij} = r_{i\cdot} r_{\cdot j}/n$, etc. This the basic test procedure is the same as in §5.5. It is just the interpretation that is different.

6 Linear regression and ANOVA

6.1 Least squares

Let $Y_1, Y_2, ..., Y_n$ be r.v.s. which are approximately linearly dependent on non-random values x_i in the sense $E(Y_i) = \alpha + \beta x_i$ or $Y_i = \alpha + \beta x_i + \text{error}_i$ or $Y_i = \alpha + \beta x_i + \varepsilon_i$.

So we have parameters $\boldsymbol{\theta} = (\alpha, \beta)$ and then $E(Y_i) = \mu_i$ where

$$\mu_i = \alpha + \beta x_i.$$

We aim to get **least squares** estimators of α and β by minimising the sum of the squares of the differences between the data and the expected values

$$S = \sum_{i=1}^{n} (Y_i - \alpha - \beta x_i)^2.$$

We minimise in the usual way

$$
\begin{aligned}
\frac{\partial S}{\partial \alpha} &= -2\sum_{i=1}^{n}(Y_i - \alpha - \beta x_i) = -2n(\overline{Y} - \alpha - \beta \bar{x}) \\
\frac{\partial S}{\partial \beta} &= -2\sum_{i=1}^{n} x_i(Y_i - \alpha - \beta x_i)
\end{aligned}
$$

$\hat{\alpha}, \hat{\beta}$ satisfy

$$\frac{\partial S}{\partial \alpha} = 0 = \frac{\partial S}{\partial \beta}$$

so

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$$

where

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i(Y_i - \bar{Y})}{\sum_{i=1}^{n} x_i(x_i - \overline{x})} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

N.B. We should check second derivatives to ensure this minimises S rather than maximises it.

Given observations $y_1, y_2, \ldots, y_n$ of $Y_1, Y_2, \ldots, Y_n$, let

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2, \qquad S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}), \qquad S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2.$$

Then, from above, we have the least squares estimates

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n} x_i(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \qquad \hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}.$$

The minimised S is called the **residual sum of squares**

$$RSS = \sum \left(y_i - \hat{\alpha} - \hat{\beta}x_i\right)^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}.$$

6.2 Properties

6.2.1 Properties of $\hat{\alpha}, \hat{\beta}$

Note

$$E(\hat{\beta}) = \beta; \quad E(\hat{\alpha}) = \alpha$$

so both are unbiased.

If ϵ_i are i.i.d. with $\mathrm{Var}(\epsilon_i) = \sigma^2$, so that $\mathrm{Var}(Y_i \mid x_i) = \sigma^2$, then

$$\mathrm{Var}(\hat{\beta}) = \frac{\sigma^2}{S_{xx}}, \qquad \mathrm{Var}(\hat{\alpha}) = \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right), \qquad \mathrm{Cov}(\hat{\alpha}, \hat{\beta}) = -\frac{\overline{x}\sigma^2}{S_{xx}}.$$

6.2.2 Properties of RSS

If we write $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$, and $e_i = y_i - \hat{y}_i = i$th residual, then

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Note that $E(RSS) = (n-2)\sigma^2$, thus $RSS/(n-2)$ provides unbiased estimate of σ^2.

There is a distinction between the error term $\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$ and the residual $e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$. In vector notation we write $\boldsymbol{e} = (e_1, \ldots, e_n)^T$ so that

$$RSS = \boldsymbol{e}^T \boldsymbol{e}. \tag{1}$$

6.2.3 Normal errors

If we assume ϵ_i are i.i.d. $N(0, \sigma^2)$, then we have the distributional results

$$\begin{aligned} Y_i \mid x_i &\sim N\left(\alpha + \beta x_i, \sigma^2\right) \\ \hat{\alpha} &\sim N\left(\alpha, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\right) \\ \hat{\beta} &\sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right) \end{aligned}$$

We find that $\hat{\alpha}, \hat{\beta}$ are also the m.l. estimates (Block A, Exercise 25) but that the m.l.e. of σ^2 is $\hat{\sigma}^2 = RSS/n$ (biased).

As above, an unbiased estimator of σ^2 is

$$S^2 = \frac{RSS}{n-2}, \quad \text{and now} \quad \frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}, \quad \text{independent of } (\hat{\alpha}, \hat{\beta}).$$

(because of the assumed normality of the observations).

6.3 Tests and CI for α, β

Under the assumption that ϵ_i are i.i.d.$N(0, \sigma^2)$ we can perform tests and find CIs for the slope and intercept parameters.

6.3.1 Slope β

We use the fact that

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right) \text{ and } \frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$$

so

$$\frac{\hat{\beta} - \beta}{\sqrt{\dfrac{S^2}{S_{xx}}}} \sim t_{n-2}$$

to give $100(1-\alpha)$ % CI as

$$\hat{\beta} \pm t_{n-2;1-\frac{\alpha}{2}} \times \frac{s}{\sqrt{S_{xx}}}.$$

To test $H_0 : \beta = \beta_0 \ v \ H_1 : \beta \neq \beta_0$ (often $\beta_0 = 0$) use

$$\frac{\hat{\beta} - \beta_0}{\sqrt{\dfrac{S^2}{S_{xx}}}} \sim t_{n-2} \quad \text{under } H_0.$$

6.3.2 Intercept α

Similarly, use

$$\frac{\hat{\alpha} - \alpha_0}{\sqrt{S^2\left(\dfrac{1}{n} + \dfrac{\overline{x}^2}{S_{xx}}\right)}} \sim t_{n-2},$$

to test $H_0 : \alpha = \alpha_0 \ v \ H_1 : \alpha \neq \alpha_0$
and to give the $100(1-\alpha)\%$ CI:

$$\hat{\alpha} \pm t_{n-2;1-\frac{\alpha}{2}} \times S\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}}.$$

6.3.3 Alternative formulation of test of $\beta = 0$

Another commonly used formulation of the test of $H_0 : \beta = 0$ has neat extensions to more complex situations. We describe the test above in the new terminology as follows.

Under **full** model $y_i = \alpha + \beta x_i + \epsilon_i$ we have

$$RSS_F = S_{yy} - \frac{S_{xy}^2}{S_{xx}}; \quad \hat{\sigma}^2 = S^2 = \frac{RSS_F}{n-2}$$

Under $H_0 : \beta = 0$ we have a **reduced** model $y_i = \alpha + \epsilon_i$

By the usual process we can obtain a least squares estimate $\hat{\hat{\alpha}} = \overline{y}$ and

$$RSS_R = \sum_{i=1}^{n} (y_i - \overline{y})^2 = S_{yy}.$$

Thus $RSS_R - RSS_F = S_{xy}^2/S_{xx} = \hat{\beta}^2 S_{xx}$.

The test above of

$$H_0 : \beta = 0 \ v \ H_1 : \beta \neq 0$$

uses

$$\frac{\hat{\beta}}{\sqrt{S^2/S_{xx}}} \sim t_{n-2} \quad \text{when } H_0 \text{ is true.}$$

This test can be re-written directly using the relationship between t and F distributions as calculating $\hat{\beta}^2 S_{xx}/S^2 \sim F_{1,n-2}$ and rejecting H_0 at level α if

$$\frac{\hat{\beta}^2 S_{xx}}{S^2} > F_{1,n-2;1-\alpha}.$$

In the terminology of full and reduced models the test statistic is

$$\frac{RSS_R - RSS_F}{\frac{RSS_F}{n-2}}$$

Note:

(a) This generalizes to more complicated models (see Linear Models course).

(b) $RSS_R - RSS_F$ is known as the **regression** SS.

6.3.4 ANOVA table

The calculations are often set out in an ANOVA (Analysis of Variance) table as follows

Source of variation	Deg. Freedom	SS	Mean Sq.	F-ratio
Regression	1	$RSS_R - RSS_F$	$\frac{RSS_R - RSS_F}{1}$	$\frac{\text{Regression MS}}{\text{Residual MS}}$
Residual	$n-2$	RSS_F	$\frac{RSS_F}{n-2}$	
Total	$n-1$	RSS_R		

Notes:

(a) F-ratio is $F_{1,n-2}$ under H_0.

(b) $RSS_R = \sum_{i=1}^{n} (y_i - \overline{y})^2$ and so 'Total' is really 'corrected total' since it is centred around mean.

(c) Note we speak of a breakdown of the sum of squares, since

$$\begin{array}{ccccc} \sum_{i=1}^{n} (y_i - \overline{y})^2 & = & \frac{S_{xy}^2}{S_{xx}} & + & \sum_{i=1}^{n} e_i^2 \\ \text{Total} & & \text{Regression} & & \text{Residual} \end{array}$$

$$\text{i.e. } \sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

(d) Expected Mean Square for regression $= \sigma^2 + \beta^2 S_{xx}$
Expected Mean Square for residual $= \sigma^2$
Thus test of $\beta = 0$ compares two estimates of σ^2.

(e)

$$R^2 = \frac{\text{regression} \quad SS}{\text{total} \quad SS} \times 100\%$$

describes proportion of variation described by the regression term, i.e. measures strength of the linear relationship between x and y.

Here $R^2 =$ (sample correlation coefficient between x and y)^2.

6.4 Least squares estimators in matrix form

Matrix notation is usually used to represent linear models. Suppose we have data $\{x_i, y_i\}$ for $i = 1, \ldots, n$ and the following model is proposed.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

for $i = 1, \ldots, n$ is written in matrix form as

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

with

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Define

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}$$

to be the least squares estimator in matrix form. The aim is to find $\hat{\boldsymbol{\beta}}$ directly using matrix notation. We must first introduce some vector notation for differentiation.

6.4.1 Differentiating with respect to vectors

Let $\boldsymbol{z}$ be an $r \times 1$ column vector $(z_1, \ldots, z_r)^T$ and let $f(z_1, \ldots, z_r)$ be some function of $\boldsymbol{z}$. We define

$$\frac{\partial f(z_1, \ldots, z_r)}{\partial \boldsymbol{z}} = \begin{pmatrix} \frac{\partial f(z_1, \ldots, z_r)}{\partial z_1} \\ \vdots \\ \frac{\partial f(z_1, \ldots, z_r)}{\partial z_r} \end{pmatrix}.$$

For any $r \times 1$ column vector $\boldsymbol{a} = (a_1, \ldots, a_r)^T$ we have

$$\frac{\partial \boldsymbol{a}^T \boldsymbol{z}}{\partial \boldsymbol{z}} = \frac{\partial (a_1 z_1 + \ldots + a_r z_r)}{d\boldsymbol{z}} = (a_1, \ldots, a_r)^T = \boldsymbol{a}.$$

If M is a square $r \times r$ matrix then

$$\frac{\partial (\boldsymbol{z}^T M \boldsymbol{z})}{\partial \boldsymbol{z}} = (M + M^T)\boldsymbol{z}.$$

Proof. Let m_{ij} represent the ijth element of M. Now $(M + M^T)\boldsymbol{z}$ is a column vector with the kth element given by $\sum_{i=1}^r m_{ki}z_i + \sum_{i=1}^r m_{ik}z_i$. Hence we must show that

$$\frac{\partial(\boldsymbol{z}^T M\boldsymbol{z})}{\partial z_k} = \sum_{i=1}^r m_{ki}z_i + \sum_{i=1}^r m_{ik}z_i.$$

From the product rule

$$\begin{aligned}\frac{\partial(\boldsymbol{z}^T M\boldsymbol{z})}{\partial z_k} &= \boldsymbol{z}^T\frac{\partial(M\boldsymbol{z})}{\partial z_k} + \left(\frac{\partial \boldsymbol{z}^T}{\partial z_k}\right)M\boldsymbol{z} \\ &= (z_1,\ldots,z_r)\begin{pmatrix}\frac{\partial}{\partial z_k}\sum_{i=1}^r m_{1i}z_i \\ \vdots \\ \frac{\partial}{\partial z_k}\sum_{i=1}^r m_{ri}z_i\end{pmatrix} + (0,\ldots,0,1,0,\ldots,0)\begin{pmatrix}\sum_{i=1}^r m_{1i}z_i \\ \vdots \\ \sum_{i=1}^r m_{ri}z_i\end{pmatrix},\end{aligned}$$

(with $(0,\ldots,0,1,0,\ldots,0)$ a vector of zeros with the kth element replaced by a 1)

$$\begin{aligned}&= (z_1,\ldots,z_r)\begin{pmatrix}m_{1k} \\ \vdots \\ m_{rk}\end{pmatrix} + (0,\ldots,0,1,0,\ldots,0)\begin{pmatrix}\sum_{i=1}^r m_{1i}z_i \\ \vdots \\ \sum_{i=1}^r m_{ri}z_i\end{pmatrix} \\ &= \sum_{i=1}^r m_{ik}z_i + \sum_{i=1}^r m_{ki}z_i,\end{aligned}$$

as required. □

6.4.2 Obtaining least squares estimators

Firstly, note that

$$\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = (\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 = R(\beta_0, \beta_1).$$

Hence in vector notation, to minimise $R(\beta_0, \beta_1)$, which is the least squares procedure, we must solve the equation

$$\begin{pmatrix}\frac{\partial}{\partial \beta_0}\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} \\ \frac{\partial}{\partial \beta_1}\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}\end{pmatrix} = \begin{pmatrix}0 \\ 0\end{pmatrix}, \qquad \text{i.e. } \frac{\partial \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}}{\partial \boldsymbol{\beta}} = \mathbf{0}.$$

Now

$$\begin{aligned}\frac{\partial \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}}{\partial \boldsymbol{\beta}} &= \frac{\partial}{\partial \boldsymbol{\beta}}(\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}) \\ &= \frac{\partial}{\partial \boldsymbol{\beta}}\left(\boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{\beta}^T X^T\boldsymbol{y} - \boldsymbol{y}^T X\boldsymbol{\beta} + \boldsymbol{\beta}^T(X^T X)\boldsymbol{\beta}\right) \\ &= -X^T\boldsymbol{y} - (\boldsymbol{y}^T X)^T + \left\{(X^T X)^T + (X^T X)\right\}\boldsymbol{\beta} \\ &= -2X^T\boldsymbol{y} + 2(X^T X)\boldsymbol{\beta}.\end{aligned}$$

Thus $\hat{\boldsymbol{\beta}}$, the least squares estimator, must satisfy

$$\mathbf{0} = -2X^T\boldsymbol{y} + 2(X^T X)\hat{\boldsymbol{\beta}},$$

(sometimes referred to as the *normal equation*) which, on rearranging, gives us the result:

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1}X^T\boldsymbol{y}.$$

6.5 Extensions

Why bother with matrices given that we already have the least squares estimates of β_0 and β_1? The crucial feature of the result just derived is that it applies to **any linear model**. That is a model which expresses $E\boldsymbol{y}$ as a linear function of the parameters $\boldsymbol{\beta}$, that is any model of the form $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Now X is an $n \times p$ matrix and $\boldsymbol{\beta}$ is a vector of p unknown parameters.

6.5.1 Examples

Fitting a Polynomial We have been considering fitting a straight line model to the data. We might instead consider a quadratic relationship via the model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i.$$

How do we estimate $(\beta_0, \beta_1, \beta_2)$? The same argument of choosing $(\beta_0, \beta_1, \beta_2)$ to make the errors small still holds. However, with matrix notation we *already* have the answer. We again write the model as

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

now with

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Then, (6.4.2) gives $\hat{\boldsymbol{\beta}}$.

Grouped data Similarly we might consider what is called a *one-way* classification of the responses. Each observation is associated with a particular group. We write y_{ij} as the j-th observed response within group i. Let p be the total number of groups. We can then have $i = 1, \dots, p$. Within group i we let n_i be the total number of observations, so that we have $j = 1, \dots, n_i$. As usual, we let n denote the total number of observations, so that $n = \sum_{i=1}^{p} n_i$.

Now let μ_i denote the population mean of the dependent variable in group i. We can now write a model for the data as follows:

$$y_{ij} = \mu_i + \varepsilon_{ij},$$

for $i = 1, \dots, p$, $j = 1, \dots, n_i$ and $\varepsilon_{ij} \sim N(0, \sigma^2)$. We will call this the *one-way analysis of variance model.*

$$y_{ij} = \mu_i + \varepsilon_{ij},$$

for $i = 1, \ldots, p$, $j = 1, \ldots, n_i$ is written in matrix form as $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, with

$$\boldsymbol{y} = \begin{pmatrix} y_{1,1} \\ \vdots \\ y_{1,n_1} \\ \cdots \\ y_{2,1} \\ \vdots \\ y_{2,n_2} \\ \cdots \\ \vdots \\ \cdots \\ y_{p,1} \\ \vdots \\ y_{p,n_p} \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \ldots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & 0 & \ldots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & & \vdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \ldots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_{1,1} \\ \vdots \\ \varepsilon_{1,n_1} \\ \cdots \\ \varepsilon_{2,1} \\ \vdots \\ \varepsilon_{2,n_2} \\ \cdots \\ \vdots \\ \cdots \\ \varepsilon_{p,1} \\ \vdots \\ \varepsilon_{p,n_p} \end{pmatrix}.$$

We can immediately obtain least squares estimates of the unknown group means $\mu_1, \ldots, \mu_p$. Since

$$(X^TX)^{-1} = \begin{pmatrix} n_1 & 0 & \ldots & 0 \\ 0 & n_2 & \ldots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \ldots & n_p \end{pmatrix}^{-1}, \qquad X^T\boldsymbol{y} = \begin{pmatrix} \sum_{j=1}^{n_1} y_{1,j} \\ \vdots \\ \sum_{j=1}^{n_p} y_{p,j} \end{pmatrix}$$

we have

$$\begin{pmatrix} \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_g \end{pmatrix} = \hat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\boldsymbol{y} = \begin{pmatrix} \frac{1}{n_1}\sum_{j=1}^{n_1} y_{1,j} \\ \vdots \\ \frac{1}{n_p}\sum_{j=1}^{n_p} y_{p,j} \end{pmatrix}$$

This result is intuitive. For example, in group 1 we have n_1 observations $y_{1,1}, \ldots, y_{1,n_1}$, all with expected value μ_1: the obvious estimate for μ_1 is the sample mean $\frac{1}{n_1}\sum_{j=1}^{n_1} y_{1,j}$, as in $\hat{\boldsymbol{\beta}}$.

6.5.2 Estimating the error variance

In general the Residual Sum of Squares (RSS) is the sum of the squares of the differences between the actual observation and the estimates of their expected values. If we write $\hat{y}_i = (X\hat{\boldsymbol{\beta}})_i$, and $e_i = y_i - \hat{y}_i = i$th residual, then

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Thus, in matrix form, the residuals are

$$\boldsymbol{e} = \boldsymbol{y} - X\hat{\boldsymbol{\beta}} \quad \text{and} \quad RSS = \boldsymbol{e}^T\boldsymbol{e}.$$

The general formula for estimating σ^2 is

$$\hat{\sigma}^2 = \frac{RSS}{n-p}.$$

6.5.3 Distribution of the estimators

Now we assume $\boldsymbol{y} \sim N(X\boldsymbol{\beta}, \sigma^2 I)$, so the response variables are independent each with variance σ^2. Since $(X^TX)^{-1}X^T$ is not random

$$E(\hat{\boldsymbol{\beta}}) = E((X^TX)^{-1}X^TE\boldsymbol{y}) = (X^TX)^{-1}X^TE(\boldsymbol{y}) = (X^TX)^{-1}(X^TX)\boldsymbol{\beta} = \boldsymbol{\beta}.$$

Hence $\hat{\boldsymbol{\beta}}$ is an **unbiased** estimator of $\boldsymbol{\beta}$. Also,

$$\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol{\beta}}) &= \left\{(X^TX)^{-1}X^T\right\} \mathrm{Var}(\boldsymbol{y}) \left\{(X^TX)^{-1}X^T\right\}^T \\
&= \left\{(X^TX)^{-1}X^T\right\} \sigma^2 I_n \left\{X(X^TX)^{-1}\right\} \\
&= \sigma^2(X^TX)^{-1}(X^TX)(X^TX)^{-1} \\
&= \sigma^2(X^TX)^{-1}.
\end{aligned}$$

Since $\hat{\boldsymbol{\beta}}$ is just a linear function of $\boldsymbol{y}$ we could invoke Block A §3.6.2 to give

$$\hat{\boldsymbol{\beta}} \sim N\left\{\boldsymbol{\beta}, \sigma^2(X^TX)^{-1}\right\}.$$

We will just derive $E(\hat{\sigma}^2)$ rather than its complete distribution. First note that (because $\mathrm{tr}(AB) = \mathrm{tr}(BA)$)

$$\mathrm{tr}(X(X^TX)^{-1}X^T) = \mathrm{tr}(X^TX(X^TX)^{-1}) = \mathrm{tr}(I_p) = p$$

where I_p is the $p \times p$ identity matrix.

Now from the definition of $\boldsymbol{e}$,

$$\begin{aligned}
E(\boldsymbol{e}^T\boldsymbol{e}) &= E\left\{(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})\right\} \\
&= E\left(\boldsymbol{y}^T\boldsymbol{y} + \hat{\boldsymbol{\beta}}^TX^TX\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^TX^T\boldsymbol{y} - \boldsymbol{y}^TX\hat{\boldsymbol{\beta}}\right). \qquad (2)
\end{aligned}$$

Consider each of the terms here separately. Firstly, we have

$$E(\boldsymbol{y}^T\boldsymbol{y}) = E\left(\sum_{i=1}^n y_i^2\right) = \sum_{i=1}^n \left\{\mathrm{Var}(y_i) + E(y_i)^2\right\} = n\sigma^2 + \boldsymbol{\beta}^TX^TX\boldsymbol{\beta}.$$

Using Block A §3.2.2 equation (3) and then Block A §3.2.2 equation (2)

$$\begin{aligned}
E\left\{\hat{\boldsymbol{\beta}}^T(X^TX)\hat{\boldsymbol{\beta}}\right\} = E\left\{(X\hat{\boldsymbol{\beta}})^TX\hat{\boldsymbol{\beta}}\right\} &= \mathrm{tr}\left(\mathrm{Cov}(X\hat{\boldsymbol{\beta}})\right) + (X\boldsymbol{\beta})^TX\boldsymbol{\beta} \\
&= \mathrm{tr}\left(\sigma^2X(X^TX)^{-1}X^T\right) + \boldsymbol{\beta}^T(X^TX)\boldsymbol{\beta} \\
&= p\sigma^2 + \boldsymbol{\beta}^T(X^TX)\boldsymbol{\beta}.
\end{aligned}$$

With regard to the last two terms in (2), note that

$$\hat{\boldsymbol{\beta}}^TX^T\boldsymbol{y} = \boldsymbol{y}^TX\hat{\boldsymbol{\beta}} = \boldsymbol{y}^TX(X^TX)^{-1}X^T\boldsymbol{y} = \mathrm{tr}\left(\boldsymbol{y}^TX(X^TX)^{-1}X^T\boldsymbol{y}\right).$$

Where the last equality is because $\boldsymbol{y}^TX(X^TX)^{-1}X^T\boldsymbol{y}$ is actually a scalar. From the information on matrices in the Basic Maths handout, we have that

$$\mathrm{tr}\left(\boldsymbol{y}^TX(X^TX)^{-1}X^T\boldsymbol{y}\right) = \mathrm{tr}\left(X(X^TX)^{-1}X^T\boldsymbol{y}\boldsymbol{y}^T\right).$$

Therefore

$$
\begin{aligned}
E\left\{\boldsymbol{y}^T X(X^T X)^{-1} X^T \boldsymbol{y}\right\} &= \operatorname{tr}\left\{(X^T X)^{-1} X^T E(\boldsymbol{y}\boldsymbol{y}^T) X\right\} \\
&= \operatorname{tr}\left\{(X^T X)^{-1} X^T (\sigma^2 I_n + X\boldsymbol{\beta}\boldsymbol{\beta}^T X^T) X\right\} \\
&= \operatorname{tr}\left\{\sigma^2 (X^T X)^{-1}(X^T X) + \boldsymbol{\beta}\boldsymbol{\beta}^T (X^T X)\right\} \\
&= p\sigma^2 + \boldsymbol{\beta}^T (X^T X)\boldsymbol{\beta}.
\end{aligned}
$$

Putting these results back into (2) we get

$$
\begin{aligned}
E(\boldsymbol{e}^T\boldsymbol{e}) &= n\sigma^2 + \boldsymbol{\beta}^T(X^T X)\boldsymbol{\beta} + q\sigma^2 + \boldsymbol{\beta}^T(X^T X)\boldsymbol{\beta} - 2q\sigma^2 - 2\boldsymbol{\beta}^T(X^T X)\boldsymbol{\beta} \\
&= (n-p)\sigma^2,
\end{aligned}
$$

and so

$$
E(\hat{\sigma}^2) = E\left(\frac{\boldsymbol{e}^T\boldsymbol{e}}{n-p}\right) = \sigma^2,
$$

i.e. $\hat{\sigma}^2$ is an **unbiased** estimator of σ^2.

Finally, it is possible to prove that

$$
\frac{RSS}{\sigma^2} \sim \chi^2_{n-p} \text{ and is independent of } \hat{\boldsymbol{\beta}}.
$$

This means we can use the distribution theory on the handouts to make tests about and give CI for the elements of $\boldsymbol{\beta}$.

6.5.4 Comparing nested models

If we fit two linear models and one can be obtained from the other by setting some of the parameters to zero (the exact definition is a little bit more complicated) then the smaller is said to be *nested* in the larger. For the comparison the larger is (sometimes) called the null model and the one nested within it the reduced model.

$$
\begin{aligned}
\text{The full model} : \boldsymbol{y} &= X_f\boldsymbol{\beta}_f + \boldsymbol{\varepsilon}. \\
\text{The reduced model} : \boldsymbol{y} &= X_r\boldsymbol{\beta}_r + \boldsymbol{\varepsilon},
\end{aligned}
$$

where the dimensions of $\boldsymbol{y}, \boldsymbol{\beta}_f$ and $\boldsymbol{\beta}_r$ are $n \times 1$, $p_f \times 1$ and $p_r \times 1$ respectively.

1. Fit the full model to the data, obtain the least squares estimate

   $$\hat{\boldsymbol{\beta}}_f = (X_f^T X_f)^{-1} X_f^T \boldsymbol{y}$$

   and the corresponding residual sum of squares

   $$RSS_f = (\boldsymbol{y} - X_f\hat{\boldsymbol{\beta}}_f)^T(\boldsymbol{y} - X_f\hat{\boldsymbol{\beta}}_f)$$

2. Fit the reduced model to the data, obtain the least squares estimate

   $$\hat{\boldsymbol{\beta}}_r = (X_r^T X_r)^{-1} X_r^T \boldsymbol{y}$$

   and the corresponding residual sum of squares

   $$RSS_r = (\boldsymbol{y} - X_r\hat{\boldsymbol{\beta}}_r)^T(\boldsymbol{y} - X_r\hat{\boldsymbol{\beta}}_r)$$

3. Calculate the F-statistic defined by

$$F = \frac{(RSS_R - RSS_F)/(p_f - p_r)}{RSS_F/(n - p_f)}$$

4. For a test of size 0.05, reject the reduced model in favour of the full model if

$$F > F_{p_f - p_r, n - p_f}(0.95)$$

Note that the choice of 0.05 for the size of the test is entirely arbitrary, but is often used in practice. You should also report the p-value for the observed F, i.e. $P(F_{p_f - p_r, n - p_f} > F)$.