MAS113X Fundamentals of Statistics I

Lecture Notes

Professor S. G. Gilmour

School of Mathematical Sciences

Queen Mary, University of London

(modified version of MAS113 notes of Dr L. Pettit)

September 27, 2007

1 Introduction

1.1 What is Statistics?

Statistics is concerned with the process of finding out about real phenomena by collecting and making sense of data. Its focus is on extracting meaningful patterns from the variation which is always present in the data. An important feature is the quantification of uncertainty so that we can make firm decisions and yet know how likely we are to be right.

1.1.1 Problems and Questions

Statistical methods are applied in an enormous diversity of problems in such fields as:

- Agriculture (which varieties grow best?)
- Genetics, Biology (selecting new varieties, species)
- Economics (how are the living standards changing?)
- Market Research (comparison of advertising campaigns)
- Education (what is the best way to teach small children reading?)
- Environmental Studies (do strong electric or magnetic fields induce higher cancer rates?)

- Meteorology (is global warming a reality?)
- Medicine (which drug is best?)
- Psychology (how are shyness and loneliness related?)
- Social Science (comparison of people's reaction to different stimuli)

Questions which arise in an investigation should be posed in non-statistical terms to keep subject matter priorities first; "translating" these questions into the language of statistics usually means to answer the following:
- What should I measure?
- How should I measure it?

1.2 Ideas of Statistical modelling

In this section we are going to discuss some of the ideas of Statistical Modelling. We start with a real life problem. We think about what to measure and how to measure it. We decide how to collect some data. This may be via a survey, an experiment or carrying out an observational study. We have to *design* the method of data collection. For example by thinking carefully about questionnaire wording or in what way we decide experimental units receive different treatments or deciding which

variables to measure. We should also think of an appropriate statistical model for our data. This will often be of the form

$$\text{Observed data} = f(x, \theta) + \text{error},$$

where x are variables we have measured and θ are parameters of our model. Data often exhibit great variability. The relationship we are assuming here is NOT deterministic. That is why the "error" term is there. We usually make some assumptions about the error term but we should use our data to check if those assumptions seem justified. If not we should go back and revise our model.

Statistical model building is an iterative process. we entertain a tentative model but we are ready to revise it if necessary. Only when we are happy with our model should we stop. We can then use our model, sometimes to understand our current set of data, sometimes to help us predict what may happen in the future. We must be ready to translate what the model is telling us statistically to the client with the real life problem.

1.3 Populations and Samples

When we carry out a statistical investigation we want to find out about a population.

Definition 1 *A* population *is the collection of items under discussion. It may be finite or infinite; it may be real or hypothetical.*

Sometimes although we have a *target population* in mind the *study population* we can actually find out information about may be different.

We are interested in measuring one or more *variables* for the members of the population but to record observations for everyone would be costly. The government carries out such a *census* of the population every ten years but also carries out regular surveys based on *samples* of a few thousand.

Definition 2 *A* sample *is a subset of a population.*

The sample should be chosen to be representative of the population because we usually want to draw conclusions or *inferences* about the population based on the sample. Samples will vary and the question of whether our sample is compatible with hypotheses we may have about the population will be a large concern in this course.

We will not concern ourselves much with the mechanics of how the sample is chosen, this is a topic for the course Samples, Surveys and Simulation which some of you may be doing or may

do next year. But the following examples give you some idea of the sorts of problems:

1. A city engineer wants to estimate the average weekly water consumption for single-family dwellings in the city.

   The population is single-family dwellings in the city. The variable we want to measure is water consumption. To collect a sample if the dwellings have water meters it might be best to get lists of dwellings and annual usage directly from the water company. If not then the local authority should have lists of addresses which can be sampled from. Note we should collect data through the year as water consumption will be seasonal.

2. A political scientist wants to determine if a majority of voters favour an elected House of Lords.

   The population is voters in the UK. Electoral rolls provide a list of those eligible to vote. What we want to measure is their opinion on this issue using a neutral question. (It would be easy to bias the response by asking a leading question.) We could choose a sample using the electoral roll and then ask the question by post, on the telephone or face to face but all these methods have problems of non-

response and/or cost.

3. A medical scientist wants to estimate the average length of time until the recurrence of a certain disease.

   The population is people who are suffering from this disease or have done in the past. What we want to measure are the dates of the last bout of disease and the new bout of disease. We could take a sample of patients suffering the disease now and follow them until they have another bout. This may be too slow if the disease doesn't recur often. Alternatively we could use medical records of people who suffered the disease in one or more hospitals but records can be wrong and there may be biases introduced.

4. An electrical engineer wants to determine if the average length of life of transistors of a certain type is greater than 5000 hours.

   The population is transistors of this type. We want to record the length of time to failure by putting a sample of transistors on test and recording when they fail. Note that for such experiments where the items under test are very reliable it may be necessary to use an "accelerated" test where we subject the items to higher currents than usual.

In other parts of the course we may not emphasize the underlying population or exactly how we collect a sample but remember these questions have had to be considered.

2 Exploring Univariate Data

A good picture is worth more than a thousand words!

We examine the collected data to get a feel for they main messages and any surprising features, before attempting to answer any formal questions. This is the exploratory stage of data analysis.

2.1 Types of variables

There are two major kinds of variables:

1. Quantitative Variables (measurements and counts)
    - continuous (such as heights, weights, temperatures); their values are often real numbers; there are few repeated values;
    - discrete (counts, such as numbers of faulty parts, numbers of telephone calls etc); their values are usually integers; there may be many repeated values.
2. Qualitative Variables (factors, class variables); these variables classify objects into groups.
    - categorical (such as methods of transport to College); there is no sense of order;
    - ordinal (such as income classified as high, medium or low); there is natural order for the values of the variable.

2.2 Frequency table

A sample of each kind of variable can be summarized in a so called *frequency table* or *relative frequency table*. Such a table is often a basis for various graphical data representations. The elements of the table are values of:

- Class:
    - for a quantitative variable - an interval, part of the range of the sample, usually ordered and of equal length (class interval);

– for a qualitative variable - often the value of the variable.

- Frequency: the number of values which fall into a class.
- Relative Frequency = (frequency/sample size).
- Cumulative Frequency (Relative Frequency) at $x*$: it is the sum of the frequencies (relative frequencies) for $x \leq x*$

value	frequency	rel. freq.
1	165	34.52%
2	154	32.22%
3	112	23.43%
4	47	9.83%
Total	478	100%

Table 2.1 Frequency and relative frequency table for 'kids at school' example, variable 'sport'.

class	frequency	rel. freq.	cum. rel. freq.
$[b_0, b_1)$	n_1	$\frac{n_1}{n}$	$\frac{n_1}{n}$
$[b_1, b_2)$	n_2	$\frac{n_2}{n}$	$\frac{n_1+n_2}{n}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[b_{k-1}, b_k)$	n_k	$\frac{n_k}{n}$	1

Table 2.2 Relative frequency and cumulative relative frequency table for values of a continuous variable.

2.3 Simple plots

2.3.1 Dot Plots

The simplest type of plot we can do is to plot a batch of numbers on a scale, stacking the same values vertically one above the other.

```
Dotplot: C1

                                    .  :..
      .   ....  ..   .   . ..:::::::. :  . .    . ...:.
-----+---------+---------+---------+---------+---------+---------+-C1
  32.0       34.0       36.0       38.0       40.0       42.0
```

Dot plots display the distances between individual points, so they are good for showing features such as clusters of points, gaps and outliers. Good for a sample of small size, e.g., $n \leq 20$.

2.3.2 Stem-and-Leaf Plots

Stem-and-Leaf plots are closely related to dot plots but with the data grouped in a way that retains much, and often all, of the numerical information. They are built from the values of the data themselves. Good for a sample of medium size, e.g., $15 \leq n \leq 150$.
Each number is split into two parts:

$$\begin{array}{ccc} a & | & b \\ \uparrow & & \uparrow \\ stem & & leaf \end{array}$$

where the leaf is a single digit.
Stem shows a class, the number of leaves for a particular stem (class) shows the frequency, the span of the leaves' values indicate the class interval. You can lengthen the plot by splitting the stems or shorten the plot by rounding numbers.

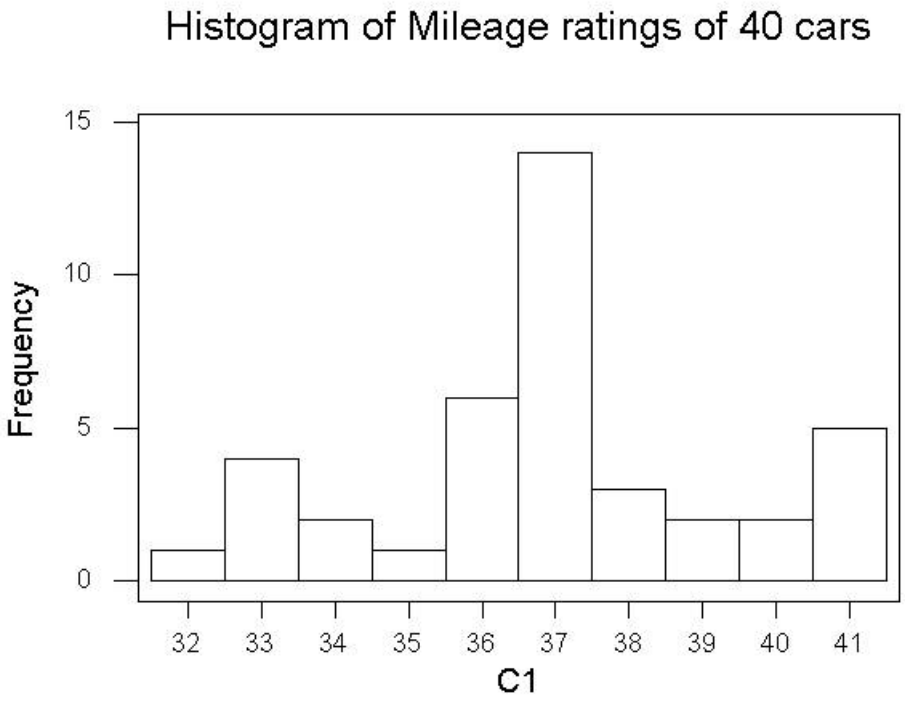
```
Stem-and-Leaf Display: C1

Stem-and-leaf of C1        N  = 40
Leaf Unit = 0.10
```

```
 1   31 8
 4   32 579
 7   33 189
 8   34 8
 9   35 5
20   36 02334567899
20   37 00112334699
 9   38 5
 8   39 09
 6   40 357
 3   41 002
```

2.3.3 Histograms

A histogram is a pictorial form of the frequency table, most useful for large data sets of a continuous variable. It is a set of boxes, whose number (number of classes), width (class interval) and height determine the shape of the histogram. The box's area represents the frequency, so that the total area of all boxes is equal to the total number of observations, resembling the property of a probability density function.



Interpreting Stem-and-Leaf Plots and Histograms

- Outliers: the observations which are well away from the main body of the data; we should look more closely at such observations to see why they are different. Are they mistakes or did something unusual happen?
- Number of peaks (modes): the mode represents the most popular value; the presence of several modes usually indicates that there are several distinct groups in the data.
- Shape of the distribution: the plot can appear to be close to symmetry, or it can show moderate or extreme skewness.
- Central values and spread: we note where the data appear

to be centered, how many modes are in the plot and where, and how spread out the data are.

- Abrupt changes: these need special attention as they may indicate some mistakes in the data or some problems coming from a wrongly executed experiment or data collection.

2.3.4 Bar Chart

A bar chart representing frequencies differs from a histogram in that the rectangles are not joined up. This visually emphasizes the discreteness of the variable; each rectangle represents a single value.
There may be various kinds of bar charts indicating other numerical measures of the sample for all sample categories.

2.3.5 Pie Chart

The pie chart displays a distribution of a variable using segments of a circle as frequencies. It is useful for presenting qualitative data sets.

2.4 Numerical Summaries of Continuous Variables

2.4.1 Locating the Centre of the Data

Two main measures of centre are:

- Mean: the average value of the sample, denoted by $\bar{x}$;

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{j=1}^{k} n_j x_j, \tag{2.1}$$

where n_j denotes the frequency of x_j. If, we have a frequency table with class intervals available only, not all observations, then x_j in the equation denotes the middle of the interval.

- Median: the middle value of the ordered data set, denoted by Med; if

$$x_{(1)} < x_{(2)} < \ldots < x_{(n)} \tag{2.2}$$

denotes the ordered data set, then

$$Med = \begin{cases} x_{(\frac{n+1}{2})} \text{ if } n \text{ is odd} \\ \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}) \text{ if } n \text{ is even.} \end{cases} \tag{2.3}$$

- Mode: the value with the highest frequency.

2.4.2 The Five-Number Summary

The five-number summary indicates the centre and the spread of the sample. It divides the ordered sample $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$ into four sections; the five numbers are the borders of the sections. The length of the sections tell us about the spread of the sample. The numbers are:

- Minimum value, $Min = x_{(1)}$;
- Lower Quartile denoted by Q_1, which 'cuts off' a quarter of the ordered data;
- Median, Med, also denoted by Q_2;
- Upper Quartile denoted by Q_3, which 'cuts off' three quarters of the ordered data;

- Maximum value, $Max = x_{(n)}$.

Quartiles are calculated in the same way as the median: Q_1 is the median of the 'lower' half of the ordered sample, Q_3 is the median of the 'upper' half of the ordered sample.

For calculation purposes

$$Q_1 = x_{(\frac{n+1}{4})}$$

and

$$Q_3 = x_{(\frac{3(n+1)}{4})}$$

.

2.5 Measuring the Spread of the data

The following two measures are simple functions of some of the 'five numbers':

- the Range

$$R = Max - Min; \tag{2.4}$$

- the Interquartile Range

$$IQR = Q_3 - Q_1. \tag{2.5}$$

Another measure of spread is the *Variance*. It is the mean of squared distances of the sample values from their average. The square root of the variance is called The *Sample Standard Deviation*, denoted by s:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{n-1}\sum_{j=1}^{k} n_j(x_j - \bar{x})^2}. \tag{2.6}$$

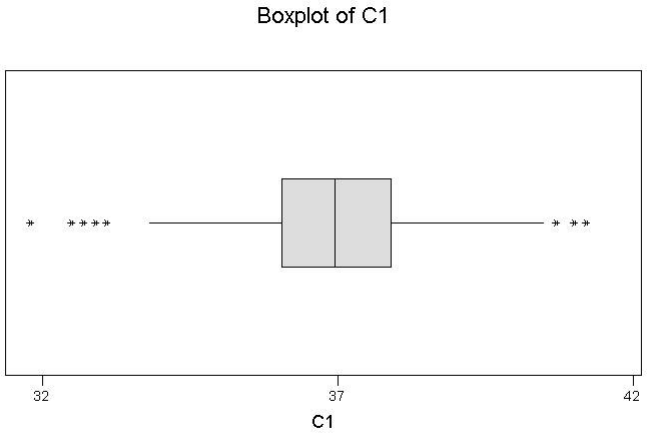
2.5.1 Pictorial Representation of The Five-Number Summary

Boxplots summarize information about the shape, dispersion, and center of your data. They can also help to spot outliers.

- The left edge of the box represents the first quartile Q_1, while the right edge represents the third quartile Q_3. Thus the box portion of the plot represents the interquartile range IQR, or the middle 50% of the observations.
- The line drawn through the box represents the median of the data.
- The lines extending from the box are called whiskers. The whiskers extend outward to indicate the lowest and highest values in the data set (excluding outliers).
- Extreme values, or outliers, are represented by dots. A value is considered an outlier if it is outside of the box (greater than Q_3 or less than Q_1) by more than 1.5 times the IQR.

The boxplot is useful to assess the symmetry of the data:

- If the data are fairly symmetric, the median line will be roughly in the middle of the IQR box and the whiskers will be similar in length.
- If the data are skewed, the median may not fall in the middle of the IQR box, and one whisker will probably be noticeably longer than the other.



2.6 Skewness

The following relations indicate skewness or symmetry:

- $Q_3 - Q_2 > Q_2 - Q_1$ and $\bar{x} > Med$ indicate positive skew;
- $Q_3 - Q_2 < Q_2 - Q_1$ and $\bar{x} < Med$ indicate negative skew;
- $Q_3 - Q_2 = Q_2 - Q_1$ and $\bar{x} = Med$ indicate symmetry.

The measure of skewness is based on the third sample moment about the mean

$$m_3 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^3.$$

This is affected by the units we measure x in. Hence, a dimensionless form is used, called the *coefficient of skewness*:

$$\text{coeff. of skew} = \frac{m_3}{s^3}.$$

A value more than or less than zero indicates skewness in the data. But a zero value does not necessarily indicate symmetry.

2.7 The Effect of Shift and Scale on Sample Measures

Denote by

$$\{x_1, x_2, \ldots, x_n\}$$

a sample from a population X.

The Effect of Shift

Let

$$y_i = x_i + a$$

for $i = 1, \ldots, n$ and for some constant a. Then

- $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(x_i + a) = \frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n} a = \bar{x} + a$;
- $Q_j(y) = Q_j(x) + a$, $j = 1, 2, 3$;
- $Q_3(y) - Q_1(y) = (Q_3(x) + a) - (Q_1(x) + a) = Q_3(x) - Q_1(x)$;
- $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\sum_{i=1}^{n}((x_i + a) - (\bar{x} + a))^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = s_x^2$.

Therefore, shifting the data shifts the measures of centre (mean, median and quartiles) but it does not affect the measures of spread (IQR and s^2). Also, you can easily show that the coefficient of skewness is not affected by a shift of the data.

The Effect of Scale

In a similar way multiplying values by a positive constant results in the measures of centre also being multiplied by that constant. The IQR and standard deviation s are multiplied by the constant and the variance s^2 is multiplied by the square of the constant. The coefficient of skewness is unaffected. Multiplying by a negative constant is similar but the IQR and s are

multiplied by the modulus of the constant and the sign of the coefficient of skewness is changed.

Check these results for yourselves as an exercise.

3 Exploratory tools for Relationships

Let

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

denote a bivariate sample from a population (X, Y). Either of X and Y can be qualitative (continuous or discrete) or quantitative (categorical or ordinal) variable. Which sort of numeric or pictorial tools we use depends on what kind of variables we examine.

3.1 Two Quantitative variables

The main graphical tools for comparing two quantitative variables are plots. Most common are:

- Scatter Plot
  A scatter plot displays points at the Cartesian coordinates: one variable provides the coordinates for the vertical y-axis, and one variable provides the coordinates for the horizontal x-axis. The scatter plot displays symbols for the data points.

- Marginal Plot
  A marginal plot is a scatter plot with graphs in the margins of the x- and/or y-axes that show the sample marginal distributions of the bivariate data. You can choose to have histograms, boxplots or dotplots as the marginal plots.

All of these give us slightly different information and each might be useful.

We can see from a plot if there seems to be some kind of relationship between the two variables. The simplest relationship is linear. The data points lying along a straight line suggest a linear relationship; the more scatter there is about the 'best fit' line, the less strong is the linear relationship.
A numerical measure of degree of linear association of two samples is the sample covariance:

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

However, it is the dimensionless *sample correlation coefficient*, usually denoted by r, which is usually used for measuring the association between x and y:

$$r = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}}.$$

The sample coefficient of correlation has the same properties as the population coefficient of correlation ρ.

See the Probability I lecture notes.

Lemma 3.1

The sample correlation coefficient r takes values between -1 and 1. $r = 1$ means positive linear correlation of the bivariate data, $r = -1$ means negative linear correlation of the bivariate data.

Proof

Assume that $x_1, \ldots, x_n$ are not all identical, otherwise the variance of x would be zero, similarly assume that $y_1, \ldots, y_n$ are not all identical.

Consider the following nonnegative function of a:

$$f(a) = \frac{1}{n-1}\sum_{i=1}^{n}((y_i - \bar{y}) - a(x_i - \bar{x}))^2 =$$
$$\frac{1}{n-1}\sum_{i=1}^{n}((y_i - \bar{y})^2 - 2a(x_i - \bar{x})(y_i - \bar{y}) + a^2(x_i - \bar{x})^2) =$$
$$s_y^2 - 2as_{xy} + a^2 s_x^2 \geq 0$$

The last term is a quadratic function of a. It is nonnegative, so

$$\Delta = (2s_{xy})^2 - 4s_x^2 s_y^2 \leq 0$$

since there must be two complex roots ($\Delta > 0$) or a repeated root ($\Delta = 0$).

This means

$$(2s_{xy})^2 \leq 4s_x^2 s_y^2$$

or

$$r^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} \leq 1.$$

Hence

$$-1 \leq r \leq 1.$$

Furthermore, if $r^2 = 1$ then $\Delta = 0$ and $a_0 = \frac{s_{xy}}{s_x^2}$ is the repeated root of the equation $f(a) = 0$. Then

$$\frac{1}{n-1}\sum_{i=1}^{n}((y_i - \bar{y}) - a(x_i - \bar{x}))^2 = \frac{1}{n-1}\sum_{i=1}^{n} z^2 = 0.$$

This can only occur if all the z terms in the summation are zero. Hence

$$y_i = \bar{y} + a(x_i - \bar{x})$$

for all $i = 1, 2, \ldots, n$. So the points (x_i, y_i) lie on a straight line with slope a. If $r = 1$, then $s_{xy} > 0$ and so the slope of the line, a, is positive. When $r = -1$, then the slope of the line is negative.
□

Note that a correlation of zero does not imply there is no relationship between the variables. It says there is no *linear* relationship. For example, a set of data which has a quadratic relationship may have zero (or near zero) correlation.

For calculation the following formula avoids rounding error. I will leave it as an exercise to show that this is equivalent to the expression given above.

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{[n\sum x_i^2 - (\sum x_i)^2][n\sum y_i^2 - (\sum y_i)^2]}}.$$

3.2 Quantitative versus Qualitative variable

To see the relationship between two such variables it is useful to plot the quantitative variable for each group of the qualitative variable against the same scale and than compare groups. Dotplots, boxplots, stem-and leaf plots, or histograms may be used for plotting depending on batch sizes. Dotplots and boxplots are good when the number of groups is large.

3.3 Two Qualitative variables

We can cross-tabulate the variables to form a two-way table of counts. Relationships can be explored by plotting proportions calculated from the table. Bar charts or pie charts may be used for each combination of the categories.

A three dimensional picture may be used as a visual representation of the two-way table.

4 One Dimensional Random Variables and Goodness of Fit Tests

In this chapter we will revise some of the material on discrete random variables and their distributions which you have seen in Probability I. We will also consider the statistical question of deciding whether a sample of data may reasonably be assumed to come from a particular discrete distribution.

First some revision:

Definition 4.1

If E is an experiment having sample space S, and X is a function that assigns a real number $X(e)$ to every outcome $e \in S$, then $X(e)$ is called a random variable (r.v.)

Definition 4.2

Let X denote a r.v. and x its particular value from the whole range of all values of X, say R_X. The probability of the event $(X \leq x)$ expressed as a function of x:

$$F_X(x) = P_X(X \leq x) \tag{4.1}$$

is called the Cumulative Distribution Function (cdf) of the r.v. X.

Properties of cumulative distribution functions

- $0 \leq F_X(x) \leq 1$, $-\infty < x < \infty$
- $\lim_{x\to\infty} F_X(x) = 1$
- $\lim_{x\to-\infty} F_X(x) = 0$
- The function is nondecreasing.
  That is if $x_1 \leq x_2$ then $F_X(x_1) \leq F_X(x_2)$.

4.1 Discrete Random Variables

Values of a discrete r.v. are elements of a countable set $\{x_1, x_2, \ldots, x_i, \ldots\}$. We associate a number $p_X(x_i) = P_X(X = x_i)$ with each outcome x_i, $i = 1, 2, \ldots$, such that:

1. $p_X(x_i) \geq 0$ for all i
2. $\sum_{i=1}^{\infty} p_X(x_i) = 1$

Note that

$$F_X(x_i) = P_X(X \leq x_i) = \sum_{x \leq x_i} p_X(x) \tag{4.2}$$

$$p_X(x_i) = F_X(x_i) - F_X(x_{i-1}) \tag{4.3}$$

The function p_X is called the *Probability Function* of the random variable X, and the collection of pairs

$$\{(x_i, p_X(x_i)),\ i = 1, 2, \ldots\} \tag{4.4}$$

is called the *Probability Distribution* of X. The distribution is usually presented in either tabular, graphical or mathematical form.

Example 4.1

$$X \sim Binomial(8, 0.4)$$

That is $n = 8$, and the probability of success p equals 0.4.

Mathematical form:

$$\{(k, P(X = k) = {}^nC_k p^k (1-p)^{n-k}\ ),\ k = 0, 1, 2, \ldots\ 8\} \tag{4.5}$$

Tabular form:

k	0	1	2	3	4	5	6	7	8
P(X=k)	0.0168	0.0896	0.2090	0.2787	0.2322	0.1239	0.0413	0.0079	0.0007
$P(X \leq k)$	0.0168	0.1064	0.3154	0.5941	0.8263	0.9502	0.9915	0.9993	1

Other important discrete distributions are:

- $Bernoulli(p)$
- $Geometric(p)$
- $Hypergeometric(n, M, N)$
- $Poisson(\lambda)$

For their properties see Probability I course lecture notes.

4.2 Goodness of fit tests for discrete random variables

4.2.1 A straightforward example

Suppose we wish to test the hypothesis that a set of data follows a binomial distribution.

For example suppose we throw three drawing pins and count the number which land pin-up. We want to test the hypothesis that a drawing pin is equally likely to land up or down. We do this 120 times and get the following data

Ups	0	1	2	3
Observed frequency	10	35	54	21

Is there any evidence to suggest that the drawing pin is not equally likely to land up or down?

Suppose it was equally likely. Then the number of ups in a single throw, assuming independent trials, would have a binomial distribution with $n = 3$ and $p = \frac{1}{2}$. So writing Y as the number of ups we would have $P[Y = 0] = \frac{1}{8}$, $P[Y = 1] = \frac{3}{8}$ $P[Y = 2] = \frac{3}{8}$ $P[Y = 3] = \frac{1}{8}$. Thus in 120 trials our expected frequencies under a binomial model would be

Ups	0	1	2	3
Expected frequency	15	45	45	15

Now our observed frequencies are not the same as our expected frequencies. But this might be due to random variation. We know a random variable doesn't always take its mean value. But how surprising is the amount of variation we have here?

We make use of a test statistic X^2 defined as follows

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

where O_i are the observed frequencies, E_i are the expected frequencies and k is the number of classes, or values that Y can take.

Now it turns out that if we find the value of X^2 for lots of samples for which our hypothesis is true it has a particular distribution called a χ^2 or

chi-squared distribution. We can calculate the value of X^2 for our sample. If this value is big, i.e. it is in the right tail of the χ^2 distribution we might regard this as evidence that our hypothesis is false. (Note if the value of X^2 was very small we might regard this as evidence that the agreement was "too good" and that some cheating had been going on.)

In our example

$$\begin{aligned} X^2 &= \frac{(10-15)^2}{15} + \frac{(35-45)^2}{45} + \frac{(54-45)^2}{45} + \frac{(21-15)^2}{15} \\ &= \frac{25}{15} + \frac{100}{45} + \frac{81}{45} + \frac{36}{15} \\ &= \frac{75+100+81+108}{45} \\ &= \frac{364}{45} \\ &= 8.08 \end{aligned}$$

Now look at Table 7, p37 in the New Cambridge statistical tables. This gives the distribution function of a χ^2 random variable. It depends on a parameter ν which is called the degrees of freedom. For our goodness of fit test the value of ν is given by $k-1$. So $\nu = 3$. For 8.0 the distribution function value is 0.9540. For 8.2 it is 0.9579. If we interpolate linearly we will get

$$0.9540 + 0.08/0.20 \times (.9579 - .9540) = .9556$$

Thus the area to the right of 8.08 is $1 - 0.9556 = 0.0444$.

This is quite a small value. It represents the probability of obtaining an X^2 value of 8.08 or more if we carry out this procedure repeatedly on samples which actually do come from a binomial distribution with $p = 0.5$ It is called the P value of the test. A P value of 0.0444 would be regarded by most statisticians as moderate evidence against the hypothesis.

An alternative approach to testing is to make a decision to accept or reject the hypothesis. This is done so that there is a fixed probability of rejecting the hypothesis when it is true. This probability is often chosen as 0.05. (Note: there is no good reason for picking this value rather than some other value.)

If we did choose 0.05 Table 8 shows us that for $\nu = 3$ the corresponding value of the χ^2 distribution is 7.815. If the value of $X^2 \leq 7.815$ we accept the

hypothesis if $X^2 > 7.815$ we reject the hypothesis. As $X^2 = 8.08$ we reject the hypothesis. To make it clear we have chosen 0.05 as our probability of rejecting the hypothesis when it is true, we say we reject the hypothesis at a 5% significance level. We call the value 7.815 the critical value.

4.2.2 Complicating factors

There are a couple of factors to complicate the goodness of fit test. Firstly if any of the expected frequencies (E_i) are less than 5 then we must group adjacent classes so that all expected frequencies are greater than 5. Secondly if we need to estimate any parameters from the data then the formula for the degrees of freedom is amended to read

$$\nu = k - p - 1$$

where k is the number of classes and p is the number of parameters estimated from the data.

We can illustrate both these ideas in the following example.

It is thought that the number of accidents per month at a junction follows a Poisson distribution. The frequency of accidents in 120 months was as follows

Accidents	0	1	2	3	4	5	6	7+
Observed frequency	41	40	22	10	6	0	1	0

To find the Poisson probabilities we need the mean μ. Since this isn't known we will have to estimate it from the data. A reasonable estimate is the sample mean of the data. This is

$$\frac{0 \times 41 + 1 \times 40 + 2 \times 22 + \cdots + 6 \times 1}{120} = 1.2$$

Now using the Poisson formula

$$P[Y = y] = \frac{e^{-\mu}\mu^y}{y!}$$

or Table 2 in New Cambridge Statistical Tables we can complete the probabilities in the following table

Accidents	Probability	E_i	O_i
0	0.3012	36.14	41
1	0.3614	43.37	40
2	0.2169	26.03	22
3	0.0867	10.40	10
4	0.0261	3.13	6
5	0.0062	0.74	0
6+	0.0015	0.18	1

Note that the probabilities have to add to one, so the last class is six or more.

Now the last three expected frequencies are all less than 5. If we group them together into a class 4+ the expected frequency will be 4.05, still less than 5. So we group the last four classes into a class 3+ with expected frequency 14.45 and observed frequency 17. We find X^2 as before.

$$\begin{aligned} X^2 &= \frac{(36.14-41)^2}{36.14} + \frac{(43.37-40)^2}{43.37} + \frac{(26.03-22)^2}{26.03} + \frac{(14.45-17)^2}{14.45} \\ &= 0.65 + 0.26 + 0.62 + 0.45 \\ &= 1.98 \end{aligned}$$

Now after our grouping there are four classes so $k = 4$ and we estimated one parameter, the mean, from the data so $p = 1$. So $\nu = 4 - 1 - 1 = 2$. Looking in Table 7 the distribution function for 1.9 is 0.6133 and for 2.0 is 0.6321. So the interpolated value for 1.98 is $0.6133 + 0.08/0.10 \times (0.6321 - 0.6133) = 0.6283$. Thus the P value is $1 - 0.6283 = 0.3717$. Such a large P value is regarded as showing no evidence against the hypothesis that the data have a Poisson distribution.

Alternatively for a significance test at the 5% level the critical value is 5.991 from table 8 and as 1.98 is smaller than this value we accept the hypothesis that the data have a Poisson distribution.

4.3 Continuous Random Variables

Values of a continuous r.v. are elements of an uncountable set, for example a real interval. The c.d.f. of a continuous r.v. is a continuous, nondecreasing,

differentiable function. An interesting difference from a discrete r.v. is that for $\delta > 0$

$$P_X(X = x) = lim_{\delta \to 0}(F_X(x + \delta) - F_X(x)) = 0$$

We define the *Density Function* of a continuous r.v. as:

$$f_X(x) = \frac{d}{dx} F_X(x) \tag{4.6}$$

Hence

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt \tag{4.7}$$

Similarly to the properties of the probability distribution of a discrete r.v. we have the following properties of the density function:

1. $f_X(x) \geq 0$ for all $x \in R_X$
2. $\int_{R_X} f_X(x)dx = 1$

The probability of an event $(X \in A)$, where A is an interval, is expressed as an integral

$$P_X(-\infty < X < a) = \int_{-\infty}^{a} f_X(x)dx = F_X(a) \tag{4.8}$$

or for a bounded interval

$$P_X(b < X < c) = \int_{b}^{c} f_X(x)dx = F_X(c) - F_X(b) \tag{4.9}$$

Example 4.2 Normal Distribution $N(\mu, \sigma^2)$
The density function is given by:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{(x-\mu)}{\sigma})^2} \tag{4.10}$$

There are two parameters which tell us about the centre and spread of the density curve: the expected value μ and the standard deviation σ.

We will return to the normal distribution in Chapter 6.

Other important continuous distributions are

- $Uniform(a, b)$
- $Exponential(\lambda)$

See Probability II course lecture notes.

Note that all distributions you have come across depend on one or more parameters, for example p, λ, μ, σ. These values are usually unknown and their estimation is one of the important problems in statistical analysis.

4.3.1 A goodness of fit test for a continuous random variable

Consider the following example.

Traffic is passing freely along a road. The time interval between successive vehicles is measured (in seconds) and recorded below.

Time interval	0-20	20-40	40-60	60-80	80-100	100-120	120+
No. of cars	54	28	12	10	4	2	0

Test whether an exponential distribution provides a good fit to these data.

We need to estimate the parameter λ of the exponential distribution. Since λ^{-1} is the mean of the distribution it seems reasonable to put $\lambda = 1/\bar{x}$. (We will discuss this further when we look at estimation). Now the data are presented as intervals so we will have to estimate the sample mean. It is common to do this by pretending that all the values in an interval are actually at the mid-point of the interval. We will do this whilst recognising that for the exponential distribution, which is skewed, it is a bit questionable.

The calculation for the sample mean is given below.

Midpoint x	Frequency f	fx
10	54	540
30	28	840
50	12	600
70	10	700
90	4	360
110	2	220
	110	3260

thus the estimated mean is $3260/110 = 29.6$. Thus we test if the data are from an exponential distribution with parameter $\lambda = 1/29.6$.

We must calculate the probabilities of lying in the intervals given this distribution.

$$\begin{aligned} P[X < 20] &= \int_0^{20} \lambda e^{-\lambda x} dx \\ &= 1 - e^{-20\lambda} \\ &= 0.4912 \end{aligned}$$

$$\begin{aligned} P[20 < X < 40] &= \int_{20}^{40} \lambda e^{-\lambda x} dx \\ &= e^{-20\lambda} - e^{-40\lambda} \\ &= 0.2499 \end{aligned}$$

Similarly

$$\begin{aligned} P[40 < X < 60] &= e^{-40\lambda} - e^{-60\lambda} = 0.1272 \\ P[60 < X < 80] &= e^{-60\lambda} - e^{-80\lambda} = 0.0647 \\ P[80 < X < 100] &= e^{-80\lambda} - e^{-100\lambda} = 0.0329 \\ P[100 < X] &= e^{-100\lambda} = 0.0341 \end{aligned}$$

Multiplying these probabilities by 110 we find the expected frequencies as given in the table below.

Time interval	0-20	20-40	40-60	60-80	80-100	100+
Observed frequency	54	28	12	10	4	2
Expected frequency	54.03	27.49	13.99	7.12	3.62	3.75

We must merge the final two classes so that the expected values are greater than 5. Thus for 80+ we have 6 observed and 7.37 expected.

We find

$$X^2 = \sum \frac{(O-E)^2}{E} = 1.71.$$

Now $\nu = 5 - 1 - 1 = 3$ since after grouping there were 5 classes and we estimated one parameter from the data. From Table 7 the P value is thus $1 - 0.3653 = 0.6347$ and there is no evidence against the hypothesis that the data follows an exponential distribution.

5 Contingency tables

As we mentioned in section 3.3 if our data take the form of two categorical variables we form a contingency table. We are often interested in whether there is some form of association or lack of independence between the two variables. Exactly what form this association takes depends on the way we collect the data.

For example consider the following:

Example 1 *227 randomly selected males were classified by eye and hair colour*

	Eye colour			
Hair colour	*Brown*	*Green/grey*	*Blue*	*Total*
Black	*10*	*24*	*8*	*42*
Brown	*16*	*41*	*26*	*83*
Fair/Red	*5*	*32*	*65*	*102*
Total	*31*	*97*	*99*	*227*

Note that in this example we selected 227 males at random and then classified them according to hair and eye colour. Apart from the grand total of 227 none of the other entries in the table were fixed. We may ask if there is an association, or lack of independence, between the two factors (at three levels). Do the proportions (or probabilities) of the three eye colours differ among the sub-populations comprising the three hair colours? Equivalently do the proportions (or probabilities) of the three hair colours differ among the three eye colours? This is a test of INDEPENDENCE.

Compare this with the following example.

Example 2 *A survey of smoking habits in a sixth form sampled 50 boys and 40 girls at random and the frequencies were noted in the following table.*

	Smoking			
	None	*Light*	*Heavy*	*Total*
Boys	*16*	*20*	*14*	*50*
Girls	*24*	*10*	*6*	*40*
Total	*40*	*30*	*20*	*90*

In this example we chose to sample 50 boys and 40 girls. Before we classified their smoking habits we knew that the row totals would be 50 and 40. We want to know if there is a difference between the sexes. We are comparing two distributions (over smoking habits) so the test is one of SIMILARITY. The hypothesis we are testing is that the population proportions of boys and girls in each smoking category are the same.

Again compare that with the following example

Example 3 *In a study of migrant birds, nestlings were ringed in four different locations A-D. One year later birds were recaptured at each location and the number of ringed birds noted. The data were*

	A	*B*	*C*	*D*	*Total*
Recovered	*30*	*75*	*24*	*31*	*160*
Not recov.	*150*	*225*	*63*	*202*	*640*
Total	*180*	*300*	*87*	*233*	*800*

In this example the column totals were fixed. We want to know if there is evidence for differences in the four recovery rates. We are comparing four proportions so the test is one of HOMOGENEITY. The hypothesis is that the proportion of recovered birds is the same for the four locations.

Now we have seen that the method of sampling is important and that this determines the hypothesis that we want to test. However it turns out that whatever the method of sampling the method we use to analyse the contingency table is the same. As with goodness of fit problems we find the expected frequencies under the hypothesis under test, calculate X^2 and compare this to an appropriate χ^2 value.

Consider the hair and eye colour example. The null hypothesis is that

$$P(\text{eye colour and hair colour}) = P(\text{eye colour}) \times P(\text{hair colour}).$$

We can estimate P(brown eyes), for example, by the number of people with brown eyes divided by the total number of people (31/227). Similarly we can estimate P(black hair) by the total number of people with black hair divided by the total number of people (42/227). So if the hypothesis of independence is true P(brown eyes and black hair) will be estimated by (31/227) $\times$

(42/227) and we would expect the number of people in our sample with brown eyes and black hair to be $227 \times (31/227) \times (42/227)$. Similarly the expected number of people in our sample with a particular combination of hair colour and eye colour if the hypothesis of independence is true will be

$$\begin{aligned} E_k &= n \times \text{Row total}/n \times \text{Column total}/n \\ &= \frac{\text{Row total} \times \text{ Column total}}{\text{overall total}(n)} \end{aligned}$$

Using this rule the table of expected frequencies is as follows:

	Eye colour			
Hair colour	Brown	Green/grey	Blue	Total
Black	5.74	17.95	18.32	42
Brown	11.33	35.47	36.20	83
Fair/Red	13.93	43.59	44.48	102
Total	31	97	99	227

We calculate X^2 as before as

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where the sum is over all the cells of the contingency table.

The number of degrees of freedom is

$$\nu = (\text{no. of rows} - 1)(\text{no. of columns} - 1).$$

Given the row and column totals we only need to know the values of ν cells in the table to determine the rest.

As before $X^2 \sim \chi^2_\nu$ under the null hypothesis of independence and a large value of X^2 gives evidence against the hypothesis.

Here $X^2 = 34.9$ on $\nu = (3-1) \times (3-1) = 4$ degrees of freedom. From table 7 we see that $P(X^2 < 25) = 0.9999$ so the P value is certainly less than 0.0001 and there is overwhelming evidence against the hypothesis that hair colour and eye colour are independent.

Consider now the smoking example. Since the row totals are fixed, under the hypothesis of similarity the row proportions or probabilities are the same

for each row. It follows that

$$\frac{E_k}{\text{Row total}} = \frac{\text{Column total}}{n}$$

or

$$E_k = \frac{\text{Row total} \times \text{ Column total}}{n}$$

Using this rule the table of expected frequencies is as follows:

	Smoking			
	None	Light	Heavy	Total
Boys	22.22	16.67	11.11	50
Girls	17.78	13.33	8.89	40
Total	40	30	20	90

For the example we find that $X^2 = 7.11$. The degrees of freedom is $(2-1)(3-1) = 2$. From table 7 we get $P(X^2 < 7.0) = 0.9698$ and $P(X^2 < 7.2) = .9727$ thus $P(X^2 < 7.11) = 0.9698 + (11/20)(.9727 - .9698) = .9714$ so the P-value is .0286. Hence there is moderate evidence against the hypothesis of similarity, moderate evidence that smoking habits differ between the boys and girls.

Finally consider the bird ringing example. Since the column totals are fixed, under the hypothesis of homogeneity the column proportions are the same for each column. It follows that

$$\frac{E_k}{\text{Column total}} = \frac{\text{Row total}}{n}$$

or

$$E_k = \frac{\text{Row total} \times \text{ Column total}}{n}$$

Using this rule the table of expected frequencies is as follows:

	A	B	C	D	Total
Recovered	36.0	60.0	17.4	46.6	160
Not recov.	144.0	240.0	69.6	186.4	640
Total	180	300	87	233	800

We find that $X^2 = 15.59$. The degrees of freedom is $(2-1)(4-1) = 3$. From table 8 we get $P(X^2 > 12.84) = 0.005$ and $P(X^2 > 16.27) = 0.001$ thus the P-value is between .001 and .005. Hence there is strong evidence against the hypothesis of homogeneity, strong evidence that the probability of recovering birds is not constant over the four locations.

As we saw with the goodness of fit test X^2 will only have a well approximated χ^2 distribution if all the $E_k > 5$. It may be possible to group rows or columns to achieve this if one of the variables is ordinal (e.g. smoking habits) but if it both are categorical any such grouping is arbitrary. In the case of contingency tables we will relax our condition to say that not more than 20% of the cells of the table should have $E_k < 5$ and none should have $E_k < 1$.

For 2×2 tables we can find a formula for the value of X^2 in terms of the entries in the table. If the table is

	Presence	Absence	Total
Group 1	a	b	$a+b$
Group 2	c	d	$c+d$
Total	$a+c$	$b+d$	n

Then

$$X^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}.$$

Example 4 *Two areas of heathland are examined; in the larger area 66 sampling units are examined and 58 of them contain a particular species of heather, while in the smaller area 22 units are examined and 12 of these contain that species. Is the species occurring at the same density over the two areas?*

The null hypothesis is that the proportion of units containing the species is the same in the two areas. The 2×2 table obtained from these data is

	Presence	Absence	Total
Area 1	58	8	66
Area 2	12	10	22
Total	70	18	88

The value of X^2 according to the formula is

$$X^2 = \frac{88(58 \times 10 - 8 \times 12)^2}{66 \times 22 \times 70 \times 18} = 11.26$$

From Table 8 we see the P-value is between .0005 and .001 so we have very strong evidence against the null hypothesis.

For 2×2 tables where any of the entries are fairly small we should apply Yates' correction. We do this by modifying the formula for X^2 to

$$X^2 = \sum \frac{(|O_i - E_i| - 0.5)^2}{E_i}.$$

Consider the heathland example. The table of expected frequencies is

	Presence	Absence	Total
Area 1	52.5	13.5	66
Area 2	17.5	4.5	22
Total	70	18	88

So using Yates' correction we find $X^2 = 9.312$ and the P-value is now between .001 and .005. This is still strong evidence against the null hypothesis but the value of X^2 has reduced considerably and in another example might have a more important effect.

6 The normal distribution, the central limit theorem and random samples

6.1 The normal distribution

We mentioned the normal (or Gaussian) distribution in Chapter 4. It has density

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{(x-\mu)}{\sigma})^2}. \tag{6.1}$$

The mean of the distribution is μ and the variance σ^2, so the standard deviation is σ.

Of special interest is the standard normal distribution which has mean 0 and variance (or standard deviation) 1. We often write the random variable with a standard normal distribution as Z. This has density

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{z^2}{2})$$

and cumulative distribution function

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-\frac{u^2}{2}) du.$$

This integral cannot be evaluated analytically but it can be solved numerically and is tabulated in The New Cambridge Statistical Tables. It is very important that you can use these tables. Note that the standard normal distribution is symmetric about zero. We use this fact often.

6.2 Central limit theorem

There are a number of results which tell us about the behaviour of sums or means of sequences of random variables. We shall state some results without proof.

Theorem 6.1 *Law of Large Numbers*

Suppose $X_1, X_2, \ldots$ are a sequence of independent and identically distributed random variables with finite means μ. Let S_n be the partial sums

$$S_n = X_1 + X_2 + \cdots + X_n.$$

Then $\frac{S_n}{n}$ *converges in distribution to* μ *as* n *tends to infinity. This means that*

$$P(\frac{S_n}{n} \leq x) \to \begin{cases} 0 & x < \mu, \\ 1 & x \geq \mu. \end{cases}$$

We shall look at some illustrations of this in Practical 6.

The central limit theorem comes in many different versions. Here we shall state, without proof, a simple version.

Theorem 6.2 *Central Limit Theorem*

If $X_1, \ldots, X_n$ *is a sequence of* n *independent random variables with*

$$\mathrm{E}[X_i] = \mu_i, \qquad \mathrm{Var}[X_i] = \sigma_i^2$$

with all means and variances finite, and $Y = \sum_{i=1}^{n} X_i$ *then*

$$Z_n = \frac{Y - \sum \mu_i}{\sqrt{\sum \sigma_i^2}} \underset{n\to\infty}{\sim} N(0,1)$$

that is Z_n *has an approximate standard normal distribution as* n *tends to infinity.*

Another way of describing this is to say that if $F_n(z)$ is the cdf of Z_n then

$$\lim_{n\to\infty} \frac{F_n(z)}{\Phi(z)} = 1$$

where $\Phi(z)$ denotes the cdf of a standard normal rv.

The following corollary is very useful.

Corollary 6.1 *If* $X_1, \ldots, X_n$ *is a sequence of independent identically distributed rvs with*

$$\mathrm{E}[X_i] = \mu \qquad \mathrm{Var}[X_i] = \sigma^2$$

then

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

This means we can approximate probabilities for $\bar{X}$ when the sample size is large whatever the distribution of X_i (so long as it has a finite mean and variance), using the normal tables. For example

$$\begin{aligned} P(\bar{X} \leq c) &= P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq \frac{c-\mu}{\sigma/\sqrt{n}}\right) \\ &= P\left(Z \leq \frac{\sqrt{n}(c-\mu)}{\sigma}\right) \\ &= \Phi\left(\frac{\sqrt{n}(c-\mu)}{\sigma}\right) \end{aligned}$$

The following theorem and corollary tell us that the approximation above is exact if the distribution of the random variables is normal to start with.

Theorem 6.3 *Let $X_1, \ldots, X_n$ be independent random variables and let $X_i \sim N(\mu_i, \sigma_i^2)$. Then the linear combination of the variables*

$$Y = a_0 + \sum_{i=1}^{n} a_i X_i$$

is also normally distributed and

$$Y \sim N(\mu, \sigma^2),$$

where $\mu = a_0 + \sum a_i \mu_i$ and $\sigma^2 = \sum a_i^2 \sigma_i^2$.

Corollary 6.2 *If $Y = \bar{X}$ and $X_i \sim N(\mu, \sigma^2)$ then $Y \sim N(\mu, \sigma^2/n)$.*

See Probability II for more details on these results.

6.3 Normal approximations to discrete random variables

Let $W_1, \ldots, W_n$ be a sequence of independent Bernoulli rvs with probability of success p. Then we know that $X = \sum W_i$ has a binomial distribution with parameters n and p and $\mathrm{E}[X] = np$ and $\mathrm{Var}[X] = npq$ where $q = 1 - p$. By the central limit theorem we have

$$Z_n = \frac{X - np}{\sqrt{npq}} \underset{n\to\infty}{\sim} N(0, 1)$$

The normal approximation to the binomial is best when p is close to 0.5 and n is large.

Note that when we approximate a discrete random variable X by a continuous one Y we know that $P(X < x + 1) = P(X \leq x)$ so we approximate both by $P(Y < x + \frac{1}{2})$. This is the so-called "continuity correction".

Another way of thinking of this is to note that we can calculate $P(X = x)$ but because of the nature of continuous distributions $P(Y = x) = 0$. Again we would approximate $P(X = x)$ by $P(x - \frac{1}{2} < Y < x + \frac{1}{2})$.

Example A fair coin is tossed 150 times. Find a suitable approximation to the probabilities of the following events

(a) more than 70 heads

(b) fewer than 82 heads

(c) more than 72 but fewer than 79 heads.

Let X be the number of heads thrown, then X has a binomial distribution with $n = 150$ and $p = 1/2$. As n is large and p moderate we may approximate X by Y a normal random variable with mean $np = 75$ and variance $np(1 - p) = 37.5$.

(a) We require $P(X > 70)$ but this is the same as $P(X \geq 71)$ so we approximate by $P(Y > 70.5)$. This is equal to

$$P(Z > (70.5 - 75)/\sqrt{37.5}) \approx P(Z > -0.735) = P(Z < 0.735) = 0.769$$

(b) We require $P(X < 82)$ but this is the same as $P(X \leq 81)$ so we approximate by $P(Y < 81.5)$. This is equal to

$$P(Z < (81.5 - 75)/\sqrt{37.5}) \approx P(Z < 1.06) = 0.855$$

(c) We require $P(72 < X < 79)$ which is the same as $P(73 \leq X \leq 78)$ and thus we approximate by $(72.5 < Y < 78.5)$. This is approximately equal to

$$P(-0.408 < Z < 0.571) = 0.716 - (1 - .658) = 0.374$$

We may similarly approximate a Poisson random variable by a normal one of the same mean and variance so long as this mean is moderately large. We again have to use the continuity correction. The justification in terms of

sums is because the sum of some Poisson random variables is also Poisson with mean equal to the sum of the means.

Example A radioactive source emits particles at random at an average rate of 36 per hour. Find an approximation to the probability that more than 40 particles are emitted in one hour.

Let X be the number of particles emitted in one hour. Then X has a Poisson distribution with mean 36 and variance 36. We can approximate X by Y which has a $N(36, 36)$ distribution. We require $P(X > 40)$. This is approximately $P(Y \geq 40.5)$ or transforming to a standard normal random variable by subtracting the mean and dividing by the standard deviation, we have

$$\begin{aligned} P(Y \geq 40.5) &= P(Z \geq \frac{40.5 - 36}{6}) \\ &= P(Z \geq 0.75) \\ &= 1 - .7734 = 0.2266 \end{aligned}$$

6.4 Random samples

In this section I shall assume we are dealing with continuous random variables. The same results apply to discrete random variables and you can translate by writing density functions as probability functions etc.

Definition 6.1 *The random variables $X_1, \ldots, X_n$ are said to be a* random sample *of size n (from the population X) if $X_1, \ldots, X_n$ are mutually independent random variables with the same density function $f(x)$, that is $X_1, \ldots, X_n$ are independent and identically distributed random variables.*

Since the X_i's are independent their joint density can be written as the product of the common marginal densities

$$f(x_1, \ldots, x_n) = f(x_1) \cdots f(x_n) = \prod_{i=1}^{n} f(x_i).$$

Any function of the random variables, e.g.

$$Y = T(X_1, \ldots, X_n)$$

is also a random variable. The distribution of such functions is of importance in statistics. The distribution of Y can usually be derived from the underlying distribution of the X_i's in the random sample. It is called the *sampling distribution* of Y

Definition 6.2 *Let $X_1, \ldots, X_n$ be a random sample of size n from a population and let $T(x_1, \ldots, x_n)$ be a real valued function. Then the random variable $Y = T(X_1, \ldots, X_n)$ is called a* statistic. *The probability distribution of a statistic Y is called the sampling distribution of Y.*

Two statistics which are often used are

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

2. The sample variance
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Lemma 6.1 *Let $X_1, \ldots, X_n$ be a random sample from a population and let $g(x)$ be a function such that $\mathrm{E}[g(X_1)]$ and $\mathrm{Var}[g(X_1)]$ exist. Then*

$$\mathrm{E}\left[\sum_{i=1}^{n} g(X_i)\right] = n\,\mathrm{E}[g(X_1)]$$

and

$$\mathrm{Var}\left[\sum_{i=1}^{n} g(X_i)\right] = n\mathrm{Var}[g(X_1)].$$

Proof

$X_1, \ldots, X_n$ are independent and identically distributed so

$$\mathrm{E}\left[\sum_{i=1}^{n} g(X_i)\right] = \sum_{i=1}^{n} \mathrm{E}[g(X_i)] = n\,\mathrm{E}[g(X_1)]$$

$$
\begin{aligned}
\mathrm{Var}\left[\sum_{i=1}^{n} g(X_i)\right] &= \mathrm{E}\left\{\left[\sum g(X_i) - \mathrm{E}[\sum g(X_i)]\right]^2\right\} \\
&= \mathrm{E}\left\{\left[\sum \{g(X_i) - \mathrm{E}[g(X_i)]\}\right]^2\right\} \\
&= \mathrm{E}\left[\sum \{g(X_i) - \mathrm{E}[g(X_i)]\}^2\right] \\
&\quad + \mathrm{E}\left[\sum_i \sum_j \sum_{i \neq j} \{g(X_i) - \mathrm{E}[g(X_i)]\}\{g(X_j) - \mathrm{E}[g(X_j)]\}\right] \\
&= \sum_i \mathrm{Var}[g(X_i)] + \sum_i \sum_j \sum_{i \neq j} \mathrm{Cov}[g(X_i), g(X_j)] \\
&= n\mathrm{Var}[g(X_1)] + \sum_i \sum_j \sum_{i \neq j} \mathrm{Cov}[g(X_i), g(X_j)]
\end{aligned}
$$

but the X_i's are independent so all the covariance terms are zero and the result follows.

Theorem 6.4 *Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and finite variance σ^2. Then*
(a) $\mathrm{E}[\bar{X}] = \mu$,
(b) $\mathrm{Var}[\bar{X}] = \sigma^2/n$,
(c) $\mathrm{E}[S^2] = \sigma^2$.

Proof For (a)

$$
\begin{aligned}
\mathrm{E}[\bar{X}] &= \mathrm{E}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[X_i] \\
&= \frac{1}{n}\sum \mu = \frac{1}{n} n\mu = \mu.
\end{aligned}
$$

For (b)

$$
\begin{aligned}
\mathrm{Var}[\bar{X}] &= \mathrm{Var}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}[X_i] \\
&= \frac{1}{n^2}\sum \sigma^2 = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}.
\end{aligned}
$$

For (c) first note that

$$
\sigma^2 = \mathrm{E}[X^2] - (\mathrm{E}[X])^2 = \mathrm{E}[X^2] - \mu^2
$$

so

$$\mathrm{E}[X^2] = \sigma^2 + \mu^2.$$

Similarly

$$\frac{\sigma^2}{n} = \mathrm{E}[\bar{X}^2] - (\mathrm{E}[\bar{X}])^2 = \mathrm{E}[\bar{X}^2] - \mu^2$$

so

$$\mathrm{E}[\bar{X}^2] = \frac{\sigma^2}{n} + \mu^2.$$

Now

$$\begin{aligned}
\mathrm{E}[S^2] &= \mathrm{E}\left[\frac{1}{n-1}\sum(X_i - \bar{X})^2\right] \\
&= \frac{1}{n-1}\mathrm{E}\left[\sum X_i^2 - n\bar{X}^2\right] \\
&= \frac{1}{n-1}\left[n\,\mathrm{E}[X_1]^2 - n\,\mathrm{E}[\bar{X}^2]\right] \\
&= \frac{1}{n-1}\left[n(\sigma^2+\mu^2) - n(\sigma^2/n + \mu^2)\right] \\
&= \frac{1}{n-1}\left[n\sigma^2 - \sigma^2\right] \\
&= \sigma^2
\end{aligned}$$

Note that this means on average if we use S^2 (with a divisor of $n-1$) we will obtain σ^2.

Definition 6.3 *A point estimator is any function $T(X_1, \ldots, X_n)$ of a random sample. We often write an estimator of the parameter θ as $\hat{\theta}$.*

An estimator of θ is a function of random variables, so is itself a random variable.

A value of the estimator for any realization $x_1, \ldots, x_n$ of the random sample, that is $T(x_1, \ldots, x_n)$ is a real number and is called an estimate.

Definition 6.4 *If $\mathrm{E}[\hat{\theta}] = \theta$ we say the estimator is unbiased. This means that the distribution of the random variable $\hat{\theta}$ should be centred about the true value θ.*

Definition 6.5 *The bias of an estimator is defined as*

$$\mathit{Bias}(\hat{\theta}) = \mathrm{E}[\hat{\theta}] - \theta.$$

Thus the sample mean is an unbiased estimator of the population mean and the sample variance is an unbiased estimator of the population variance.

6.5 Distribution of sample total

The total T of the values X measured in a sample of size n equals $n\bar{X}$. It follows that T has a distribution of the same form as $\bar{X}$, that is a normal distribution, exactly if X is normally distributed and approximately so if n is large by the central limit theorem.

The mean and variance of T are:

$$\mathrm{E}[T] = \mathrm{E}[n\bar{X}] = n\,\mathrm{E}[\bar{X}] = n\mu,$$

$$\mathrm{Var}[T] = \mathrm{Var}[n\bar{X}] = n^2\mathrm{Var}[\bar{X}] = n^2 \times \sigma^2/n = n\sigma^2.$$

Example A charter aeroplane company is asked to carry regular loads of 100 sheep. The plane available for this work has a carrying capacity of 5000kg. Records of the weights of about 1000 sheep which are typical of those that might be carried show that the distribution of sheep weight has a mean of 45kg and a standard deviation of 3kg. Can the company take the order?

Let T be the total weight of 100 sheep. Then

$$\mathrm{E}[T] = n\mu = 100 \times 45 = 4500,$$

$$\mathrm{Var}[T] = n\sigma^2 = 100 \times 9 = 900.$$

Since n is large the distribution of T will be approximately normal. So the probability that $T > 5000$ is

$$P\left(Z > \frac{5000 - 4500}{\sqrt{900}}\right) = P(Z > 16.7).$$

This probability is so small that tables do not give it. (Note $P(Z > 5) = 3 \times 10^{-7}$). the company can safely take the order.

Example An employer has to interview 20 candidates for a job. His experience has been that he may treat the length of an interview as normally distributed with mean 10 mins and standard deviation 3mins. He begins to interview at 9.00. At what time should he ask for his coffee to be brought to

him if he is to be 99% certain that he will have seen 50% of the candidates by then?

What is the probability that he will finish the interviews by 13.00 if he takes 15mins over coffee?

The length of an individual interview is $N(10, 9)$. The total length of 10 interviews is $N(100, 90)$. To determine the time t at which coffee should be brought (after 100 interviews) we require

$$\Phi\left(\frac{t-100}{\sqrt{90}}\right) = 0.99.$$

From table 5

$$\frac{t-100}{\sqrt{90}} = 2.3263.$$

Therefore $t = 100 + 2.3263 \times \sqrt{90} = 122$, so coffee should be brought at 9.00 hours + 122mins that is at 11.02.

The distribution of the total length of 20 interviews is $N(200, 180)$. The time available for interviews is 240-15 = 225 minutes. The probability of finishing by 13.00 is

$$\Phi\left(\frac{225-200}{\sqrt{180}}\right) = \Phi(1.8633) = 0.9688.$$

7 Hypothesis testing - one sample tests

7.1 Introduction

Definition 7.1 *A hypothesis is a statement about a population parameter.*

Example A hypothesis might be that the mean age of students taking MAS113X is 19 greater than years.

If it is not true what is the alternative hypothesis? Writing μ for the mean age the two hypotheses are

$$H_0 : \mu \leq 19.0 \qquad H_1 : \mu > 19.0$$

Definition 7.2 *The two complementary hypotheses in a hypothesis testing problem are called the* null hypothesis *and the* alternative hypothesis*. They are denoted by H_0 and H_1, respectively.*

If θ denotes a population parameter, the general format of the null and alternative hypotheses are

$$H_0 : \theta \in \Theta_0 \qquad H_1 : \theta \in \Theta_0^c$$

where Θ_0 is a subset of the parameter space Θ and Θ_0^c is its complement.

Thus in the age example, formally

$$\Theta = (0, \infty), \qquad \Theta_0 = (0, 19], \qquad \Theta_0^c = (19, \infty).$$

Example A null hypothesis might be that the weight of jam in a jar is 1kg. The alternative might be that the weight is not 1kg. Values less than 1kg and greater than 1kg are both covered by our alternative hypothesis.

Formally

$$\Theta = (0, \infty), \qquad \Theta_0 = \{1\}, \qquad \Theta_0^c = (0, \infty) \setminus \{1\}.$$

How do we verify a hypothesis H_0? Suppose we have collected some data. Do they support the null hypothesis or contradict it? We need a criterion, based on the collected data, which will help us answer the question.

Definition 7.3 *A hypothesis test is a rule that specifies for which sample values the decision is made to reject H_0, i.e. accept H_1, and for which sample values not to reject H_0.*

Typically, a hypothesis test is specified in terms of a test statistic $T(X_1, \ldots, X_n)$, a function of a random sample.

For example, for the age of the class a test might specify that H_0 is to be rejected if a value of $\bar{X}$ is greater than 19.5. This defines a *rejection region* or *critical region* as

$$\{(x_1, \ldots, x_n) : \bar{x} > 19.5\}$$

the set of sample points which give the value of the test statistic $T(X_1, \ldots, X_n) = \bar{X}$ bigger than 19.5.

The compliment of the rejection region is called the *acceptance region.* Here it is

$$\{(x_1, \ldots, x_n) : \bar{x} \leq 19.5\}$$

7.2 Type I and Type II errors

The decision to reject or not to reject the null hypothesis is based on a test statistic computed from values of a random sample. Hence such a decision is subject to error because of sampling variation. Two kinds of error may be made when testing hypotheses.

1. If the null hypothesis is rejected when it is true.

2. If the null hypothesis is not rejected when it is false.

	H_0 is true	H_0 is false
Not reject H_0	No error	Type II error
Reject H_0	Type I error	No error

We denote the probabilities of Type I and Type II errors by

$$\begin{aligned} \alpha &= P(\text{type I error}) = P(\text{reject } H_0 | H_0 \text{ is true}) \\ \beta &= P(\text{type II error}) = P(\text{not reject } H_0 | H_0 \text{ is false}) \end{aligned}$$

We would like to have test procedures which make both kinds of errors small. We define the *power of the test* as $1 - \beta$. Thus a test has high power if the probability of rejecting a false null hypothesis is large. In this course we shall concentrate on specifying α. The power can be used to compare two tests with the same α to see which is more powerful (better). It can also be used to decide how large a sample size should be used. Such problems will be discussed in Fundamentals of Statistics II.

The probability of a type I error, α, is often called the *significance level of the test*. It is controlled by the location of the rejection or critical region. So it can be set as small as required. Often α is taken to be 0.05 or 0.01. Of course if we choose a small α then β may be large.

7.3 Tests concerning the mean

7.3.1 Test for a mean when the variance is known

We are interested in testing the null hypothesis

$$H_0 : \mu = \mu_0.$$

We begin by assuming that the alternative hypothesis is $H_1 : \mu \neq \mu_0$.

Let $X_1, \ldots, X_n$ be a random sample from a population X, and let

$$\mathrm{E}[X] = \mu \qquad \mathrm{Var}[X] = \sigma^2$$

where μ is unknown and σ^2 is known.

There are two cases to consider. If the population X is normal then we know that

$$\bar{X} \sim N(\mu, \sigma^2/n).$$

If the population is not normal but X has finite mean and variance then for n large

$$\bar{X} \dot{\sim} N(\mu, \sigma^2/n).$$

In either case we know that

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

exactly in case 1 and approximately in case 2.

If the null hypothesis is true, i.e. $\mu = \mu_0$ then

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

Given a sample we can calculate the observed $\bar{x}$ and since we know all the other terms an observed value of z. If the null hypothesis is true z should be close to zero. Large values of $|z|$ would tend to contradict the hypothesis.

Suppose we find the value $z_{\alpha/2}$ such that

$$P(Z > z_{\alpha/2}) = \alpha/2.$$

By the symmetry of the standard normal distribution

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

If we set the rejection region as

$$\{z : z < -z_{\alpha/2} \cup z > z_{\alpha/2}\}$$

then the probability of a type I error is the probability that Z lies in the rejection region when the null hypothesis is true and this is exactly α.

The rejection region, for this alternative hypothesis, consists of the two tails of the standard normal distribution and for this reason we call it a two-tailed test.

Note this test is often called a z-test.

Example

Drills being manufactured are supposed to have a mean length of 4cm. From past experience we know the standard deviation is equal to 1cm and the lengths are normally distributed. A random sample of 10 drills had a mean of 4.5cm. Test the hypothesis that the mean is 4.0 with $\alpha = 0.05$.

We have

$$H_0 : \mu = 4.0 \quad \text{versus} \quad H_1 : \mu \neq 4.0$$

We know that

$$\bar{X} \sim N\left(\mu, \frac{1}{10}\right)$$

so if H_0 is true

$$Z = \frac{\bar{X} - 4}{\sqrt{1/10}} \sim N(0, 1).$$

The observed value of Z is

$$\frac{4.5 - 4}{\sqrt{1/10}} = 1.58$$

For a 2 sided test with $\alpha = 0.05$ the rejection region is $\{z : |z| > 1.96\}$. Since $z = 1.58$ we do not reject H_0 at the 5% level.

Suppose now our alternative is $H_1 : \mu > \mu_0$. Our null is $H_0 : \mu \leq \mu_0$ but we use the value least favourable to H_0, i.e. μ_0 as the assumed representative value from H_0.

Large values of $\bar{X}$ will give evidence against H_0, so our rejection region is given by $\{z : z > z_\alpha\}$ where $P(Z > z_\alpha) = \alpha$.

Similarly for an alternative $H_1 : \mu < \mu_0$ the region will be $\{z : z < z_{\alpha'}\}$ and by symmetry $z_{\alpha'} = -z_\alpha$.

Example An advertisement for a certain brand of cigarettes claimed that on average there is no more than 18mg of nicotine per cigarette. A test of 12 cigarettes gave a sample mean of $\bar{x} = 19.1$. Assuming $\sigma^2 = 4$ test this claim with a significance level of $\alpha = 0.05$.

We have

$$H_0 : \mu = 18.0 \quad \text{versus} \quad H_1 : \mu > 18.0$$

We know that

$$\bar{X} \sim N\left(\mu, \frac{4}{12}\right)$$

so if H_0 is true

$$Z = \frac{\bar{X} - 18}{\sqrt{1/3}} \sim N(0, 1).$$

The observed value of Z is

$$\frac{19.1 - 18.0}{\sqrt{1/3}} = 1.9053$$

For a 1 sided test with $\alpha = 0.05$ the rejection region is $\{z : z > 1.6449\}$. Since $z = 1.9053$ we can reject H_0 at the 5% level.

Note that if we had chosen $\alpha = 0.02$ then $z_\alpha = 2.0537$ we we would fail to reject H_0 at the 2% significance level

7.3.2 Test for a normal mean when variance is unknown

Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ population. We now assume the values of both μ and σ^2 are unknown.

We want to test $H_0 : \mu = \mu_0$. When σ^2 is known we have that

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

When σ^2 is unknown we estimate it by S^2 the sample variance defined by

$$S^2 = \frac{\sum(X_i - \bar{X})^2}{n - 1}.$$

The distribution of this statistic is no longer standard normal. In fact

$$T = \frac{(\bar{X} - \mu_0)\sqrt{n}}{S} \sim t_{n-1}$$

a student t distribution with $n - 1$ degrees of freedom. Note the degrees of freedom is the same as the divisor in the sample variance.

Table 9 in New Cambridge Statistical Tables gives the distribution function of T and Table 10 gives the percentage points. A t distribution has heavier tails than a normal distribution. A t with 1 degree of freedom is also called a Cauchy distribution. As the degrees of freedom tends to infinity a t distribution tends to a standard normal. We call the resulting test a t test or one sample t test.

If the alternative hypothesis is $H_1 : \mu \neq \mu_0$ then I shall write the critical region as $\{t : |t| > t_{n-1}(\alpha/2)\}$.

Example A pharmaceutical manufacturer is concerned about the impurity concentration in batches of drug and is anxious that the mean impurity doesn't exceed 2.5%. It is known that impurity concentration follows a normal distribution. A random sample of 10 batches had the following concentrations

2.1	1.9	2.4	2.3	2.6
1.5	2.8	2.6	2.7	1.8

Test at a significance level $\alpha = 0.05$ that the population mean concentration is at most 2.5.

Our null hypothesis is $H_0 : \mu = 2.5$ versus an alternative $H_1 : \mu > 2.5$.

The test statistic is

$$T = \frac{(\bar{X} - 2.5)\sqrt{n}}{S} \sim t_{n-1}$$

if H_0 is true.

Now $\bar{x} = 2.27$ and $s^2 = 0.1868$ so the observed value of T is

$$t = \frac{(2.27 - 2.5)\sqrt{10}}{.4322} = -1.683.$$

For $\alpha = 0.05$ with 9 degrees of freedom the critical region is $\{t : t > 1.833\}$ and so we fail to reject H_0 at the 5% significance level.

7.4 P values

We discussed the use of P values when we looked at goodness of fit tests. They can be useful as a hypothesis test with a fixed significance level α does not give any indication of the strength of evidence against the null hypothesis.

The P value depends on whether we have a one-sided or two-sided test.

Consider a one sided test of $H_0 : \mu = \mu_0$ versus $H_1 : \mu > \mu_0$. Assuming the population variance is unknown so that we are using a t test the P value is

$$P(T > t_{\text{obs}})$$

where t_{obs} is the observed value of t.

For $H_1 : \mu < \mu_0$ the P value is $P(T < t_{\text{obs}})$.

For a two sided test with $H_1 : \mu \neq \mu_0$ it is

$$P(|T| > |t_{\text{obs}}|) = 2P(T > |t_{\text{obs}}|).$$

In each case we can think of the P value as the probability of obtaining a value of the test statistic more extreme than we did observe assuming that H_0 is true. What is regarded as more extreme depends on the alternative hypothesis. If the P value is small that is evidence that H_0 may not be true.

It is useful to have a scale of evidence to help us interpret the size of the P value. There is no agreed scale but the following may be useful as a first indication:

P value	Interpretation
$P > 0.10$	No evidence against H_0
$0.05 < P < 0.10$	Weak evidence against H_0
$0.01 < P < 0.05$	Moderate evidence against H_0
$0.001 < P < 0.01$	Strong evidence against H_0
$P < 0.001$	Very strong or overwhelming evidence against H_0

Note that the P value is the smallest level of significance that would lead to rejection of the null hypothesis.

7.5 Test of hypothesis on the variance

Let $X_1, \ldots, X_n$ be a random sample from a population which is $N(\mu, \sigma^2)$ where μ and σ^2 are unknown. We consider now how to test a hypothesis about the population variance σ^2. We shall present the results without justification.

To test $H_0 : \sigma^2 = \sigma_0^2$ versus $H_1 : \sigma^2 \neq \sigma_0^2$ we use the test statistic

$$W = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$$

if H_0 is true.

Note the χ^2_ν distribution has mean ν.

Since the χ^2_ν distribution is defined on $(0, \infty)$ and is skewed two-sided rejection regions are more complicated than before. The rejection region is

$$\{w : w > \chi^2_{n-1}(\alpha/2) \cup w < \chi^2_{n-1}(1 - (\alpha/2))\}.$$

For example if $\alpha = 0.05$ and $n = 9$ then from Table 8 we have

$$\chi^2_8(0.025) = 17.53 \qquad \chi^2_8(0.975) = 2.180$$

and we would only reject H_0 at the 5% significance level if the observed value of W was outside the interval $[2.180, 17.53]$.

Similarly for a one-sided test, for example if $H_1 : \sigma^2 > \sigma_0^2$ then we would reject H_0 at the α significance level if

$$w > \chi^2_{n-1}(\alpha)$$

Example It is important that the variance of the percentage impurity levels of a chemical don't exceed 4.0. A random sample of 20 consignments had a sample variance of 5.62. Test the hypothesis that the population variance is at most 4.0 at a 5% level of significance and find the P value.

Our null and alternative hypotheses are

$$H_0 : \sigma^2 = 4.0 \qquad H_1 : \sigma^2 > 4.0$$

The test statistic

$$W = \frac{(n-1)S^2}{4.0} \sim \chi^2_{19}$$

if H_0 is true. The observed value of W is $w = \frac{19 \times 5.62}{4} = 26.695$. From the tables $\chi^2_{19}(0.05) = 30.14$. Since our observed value is less than this we fail to reject H_0 at the 5% significance level.

The P value is $P(W > 26.695)$. Now from Table 7 with $\nu = 19$ $P(W < 26) = 0.8698$ and $P(W < 27) = 0.8953$ so using linear interpolation

$$P(W < 26.695) \approx 0.8698 + 0.695(.8953 - .8698) = 0.8875.$$

Thus $P(W > 26.695) = 1 - 0.8875 = 0.1125$. Using our scale of evidence there is no evidence against H_0.

For a two sided alternative the P value is found by multiplying the corresponding P value by 2.

Example Take the data in the last example but suppose that we want the variance to equal 4.0. Now from the tables $\chi^2_{19}(0.025) = 32.85$ and $\chi^2_{19}(0.975) = 8.907$ and as our observed value lies between these value we fail to reject H_0. The P value is $2 \times 0.1125 = 0.225$.

7.6 Test of hypothesis on a proportion

Suppose we have a random sample of size n consisting of observations from a Bernoulli distribution with probability of success p. Then we know that $X = \sum_{i=1}^{n} X_i$ has a Binomial distribution with parameters n and p.

To test the hypothesis that $p = p_0$ we can use the test statistic

$$Z = \frac{X - np_0}{\sqrt{np_0(1-p_0)}} \sim N(0,1)$$

if H_0 is true so long as n is large by the Central Limit Theorem. The rejection region depends on the alternative hypothesis as before.

Example In a random sample of 180 voters, 75 expressed support for University top-up fees. Test the hypothesis that at least 50% of all voters support the measure. Use a significance level of $\alpha = 0.05$.

Our null and alternative hypotheses are

$$H_0 : p = 0.5 \qquad H_1 : p < 0.5$$

The test statistic

$$Z = \frac{X - np_0}{\sqrt{np_0(1-p_0)}} \sim N(0,1)$$

if H_0 is true. The observed value of Z is

$$z = \frac{75 - 180 \times 0.5}{\sqrt{180 \times 0.5 \times 0.5}} = \frac{-15}{\sqrt{45}} = -2.236$$

For a one-sided test at the 5% significance level the rejection region is $\{z : z < -1.6449\}$ and so we can reject H_0 at the 5% level.

7.7 Confidence intervals

Often we may be asked to estimate the population mean μ rather than test a hypothesis about it. Or we may have performed a test and found evidence against the null hypothesis, casting doubt on our original hypothesised value. We should give an estimate of uncertainty along with our best estimate of μ, which is $\bar{x}$, the sample mean.

One measure of uncertainty is the standard error of the mean, $\sigma/\sqrt{n}$ if σ^2 is known or $s/\sqrt{n}$ if it is not. Equivalently, but perhaps more usefully we can give a *confidence interval*. This is derived as follows:

Assume that the variance σ^2 is known. Using the result for the sampling distribution of $\bar{X}$ we know that

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

it follows that

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

cross-multiplying, we get

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

Rearranging, we have

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

We can say that the *random interval*

$$\left(\bar{X} - 1.96\sigma/\sqrt{n}, \bar{X} + 1.96\sigma/\sqrt{n}\right)$$

has a probability of 0.95 of containing or covering the value of μ, that is, 95% of all samples will give intervals (calculated according to this formula) which contain the true value of the population mean.

The corresponding interval with $\bar{X}$ replaced by $\bar{x}$ is called a 95% confidence interval for μ. Note that the 95% refers to the random variables $\bar{X} \pm 1.96\sigma/\sqrt{n}$, since μ is not a random variable. Nevertheless a confidence interval does give an interval estimate for μ which is more useful than the point estimate $\bar{x}$.

The confidence interval has another interpretation which shows its connection with hypothesis tests. We saw in practicals 7 and 8 that a number of different null hypotheses were consistent with a given sample. A 95% confidence interval contains all those values for which the P value is greater than or equal to 0.05 in a two sided test of the null hypothesis that the unknown parameter takes that value.

Again there is nothing magic in 95% and 0.05. In general we can find a $100(1-\alpha)$% confidence interval and this relates to P values greater than or equal to α. In general when σ^2 is known the $100(1-\alpha)$% confidence interval for μ is

$$\bar{x} \pm z_{\alpha/2}\sigma/\sqrt{n}.$$

Example Drills being manufactured are supposed to have a mean length of 4cm. From past experience we know the standard deviation is equal to 1cm and the lengths are normally distributed. A random sample of 10 drills had a mean of 4.5cm. Find 95% and 99% confidence intervals for the population mean.

The 95% confidence interval is given by

$$\begin{aligned}\bar{x} \pm 1.96\sigma/\sqrt{n} &= 4.5 \pm 1.96 \times \frac{1}{\sqrt{10}} \\ &= 4.5 \pm 0.62 \\ &= (3.88, 5.12).\end{aligned}$$

The 99% confidence interval is given by

$$\begin{aligned}\bar{x} \pm 2.5758\sigma/\sqrt{n} &= 4.5 \pm 2.5758 \times \frac{1}{\sqrt{10}} \\ &= 4.5 \pm 0.81 \\ &= (3.69, 5.31)\end{aligned}$$

Note that this means we would fail to reject any $\mu \in (3.69, 5.31)$ at the 1% significance level.

When σ^2 is unknown we replace it by S^2 and the corresponding percentage point of the t_{n-1} distribution. Thus a $100(1-\alpha)\%$ confidence interval for μ is

$$\bar{x} \pm t_{n-1}(\alpha/2)s/\sqrt{n}.$$

Example Drills being manufactured are supposed to have a mean length of 4cm. From past experience we know the lengths are normally distributed. A random sample of 10 drills had a mean of 4.5cm and sample variance 1.2. Find 95% and 99% confidence intervals for the population mean.

The 95% confidence interval is given by

$$\begin{aligned}\bar{x} \pm 2.262S/\sqrt{n} &= 4.5 \pm 2.262 \times \frac{\sqrt{1.2}}{\sqrt{10}} \\ &= 4.5 \pm 0.78 \\ &= (3.72, 5.28).\end{aligned}$$

The 99% confidence interval is given by

$$\begin{aligned}\bar{x} \pm 3.25S/\sqrt{n} &= 4.5 \pm 3.25 \times \frac{\sqrt{1.21}}{\sqrt{10}} \\ &= 4.5 \pm 1.13 \\ &= (3.37, 5.63)\end{aligned}$$

A confidence interval for σ^2 is

$$\left(\frac{(n-1)s^2}{\chi^2_{n-1}(\alpha/2)}, \frac{(n-1)s^2}{\chi^2_{n-1}(1-\alpha/2)}\right).$$

Example Drills being manufactured are supposed to have a mean length of 4cm. From past experience we know the lengths are normally distributed. A random sample of 10 drills had a mean of 4.5cm and sample variance 1.2. Find 95% and 99% confidence intervals for the population variance.

The 95% confidence interval is given by

$$\begin{aligned}\left(\frac{(n-1)s^2}{19.02}, \frac{(n-1)s^2}{2.700}\right) &= \left(\frac{9\times 1.2}{19.02}, \frac{9\times 1.2}{2.700}\right)\\ &= (0.568, 4.000)\,.\end{aligned}$$

The 99% confidence interval is given by

$$\begin{aligned}\left(\frac{(n-1)s^2}{23.59}, \frac{(n-1)s^2}{1.735}\right) &= \left(\frac{9\times 1.2}{23.59}, \frac{9\times 1.2}{1.735}\right)\\ &= (0.458, 6.225)\,.\end{aligned}$$

7.8 A confidence interval for a Poisson mean

Suppose we have collected data from a Poisson distribution mean λ. As long as the mean is large we can use the normal approximation to the Poisson to give a confidence interval. A suggestion is that for $\lambda \geq 10$ a 95% confidence interval will give a reasonable result but for a 99% confidence interval we should have $\lambda \geq 20$.

Suppose n Poisson observations drawn at random from a Poisson distribution of mean λ have sample mean $\bar{r}$. Since the mean and variance of the Poisson are both λ we have that

$$\frac{(\bar{r}-\lambda)}{\sqrt{\lambda/n}} \sim N(0,1).$$

An approximate 95% confidence interval could therefore be written as

$$\bar{r} \pm 1.96\sqrt{\frac{\bar{r}}{n}}$$

replacing the λ in the square root by its estimate.

However we can do better than this. The 95% limits are found by requiring that

$$\frac{|\bar{r} - \lambda|}{\sqrt{\lambda/n}} \leq 1.96$$

Let us solve this inequality. We will write 1.96 as c and recognise that we could replace 1.96 by the appropriate value for a $100(1-\alpha)\%$ confidence interval. To do this consider the corresponding equality

$$\frac{|\bar{r} - \lambda|}{\sqrt{\lambda/n}} = c$$

square to give

$$n(\bar{r} - \lambda)^2 = c^2\lambda,$$

a quadratic in λ:

$$n\lambda^2 - (2n\bar{r} + c^2)\lambda + n\bar{r}^2 = 0.$$

The two roots of this quadratic will give the upper and lower confidence limits for the true value of λ. The limits are

$$\begin{aligned} \lambda &= \frac{1}{2n}\left[2n\bar{r} + c^2 \pm \sqrt{(2n\bar{r} + c^2)^2 - 4n^2\bar{r}^2}\right] \\ &= \bar{r} + \frac{c^2}{2n} \pm \frac{c}{2n}\sqrt{c^2 + 4n\bar{r}}. \end{aligned}$$

Example In a traffic survey on a motorway, the mean number of vehicles passing per minute for 1 hour was 18. Find a 95% confidence interval for the true mean rate of vehicles passing per minute.

We apply the result with $c = 1.96$. We have $\bar{r} = 18$ and $n = 60$. So the confidence limits are

$$\begin{aligned} \lambda &= 18 + \frac{3.84}{120} \pm \frac{1.96}{120}\sqrt{3.84 + 240 \times 18} \\ &= 18.032 \pm 1.072 \\ &= (16.96, 19.10) \end{aligned}$$

The limits using $\bar{r}$ as the estimate of the variance are $\bar{r} \pm 1.96\sqrt{\bar{r}/n}$ so are $18 \pm 1.96\sqrt{0.30} = (16.93, 19.07)$. These are close to the better limits as n is large.

8 Hypothesis tests for two samples

In this chapter we consider examples with two samples. We might want to test that two means are equal or two variances.

8.1 Two independent samples - the two sample t test

Consider the situation where we have two independent samples and we want to test if they come from the same population. In particular if they have the same mean. We shall use the following notation.

We assume that the first sample $X_1, \ldots, X_{n_1}$ is of size n_1 and is normally distributed with mean μ_1 and variance σ^2. We shall denote the sample mean and variance by $\bar{X}$ and S_1^2. We assume that the second sample $Y_1, \ldots, Y_{n_2}$ is of size n_2 and is normally distributed with mean μ_2 and variance σ^2. We shall denote the sample mean and variance by $\bar{Y}$ and S_2^2. Note we are assuming that the samples come from populations with the same variance.

We want to test the null hypothesis $H_0 : \mu_1 = \mu_2$ against an alternative which is often two sided $H_1 : \mu_1 \neq \mu_2$ but which could be one sided.

Because we are assuming the population variances are the same we estimate the variance by what is called the pooled estimator of variance. This is

$$S_0^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

We showed in Coursework 6 that in these circumstances the pooled estimator of variance is an unbiased estimator of σ^2.

Note that if σ^2 were known $\bar{X} \sim N(\mu_1, \sigma^2/n_1)$ and $\bar{Y} \sim N(\mu_2, \sigma^2/n_2)$ it follows that

$$\bar{X} - \bar{Y} \sim N\left((\mu_1 - \mu_2), \sigma^2(\frac{1}{n_1} + \frac{1}{n_2})\right)$$

since the samples are assumed independent. Thus

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n_1 + 1/n_2}} \sim N(0, 1)$$

and so if σ^2 were known we could base a test of $\mu_1 = \mu_2$ on the test statistic

$$Z = \frac{\bar{X} - \bar{Y}}{\sigma\sqrt{1/n_1 + 1/n_2}}$$

which would have a $N(0,1)$ distribution if H_0 were true.

Since σ^2 is unknown we replace it by the pooled estimator S_0^2 and as in the one sample case the distribution changes from a normal to a t. The degrees of freedom are the same as the divisor in S_0^2, namely $n_1 + n_2 - 2$. thus our test statistic is

$$T = \frac{\bar{X} - \bar{Y}}{S_0\sqrt{1/n_1 + 1/n_2}}$$

which has a $t_{n_1+n_2-2}$ distribution if H_0 is true.

Example 8.1 *Two random samples were independently drawn from two populations. The first sample of size 6 had mean 49.5 and variance 280.3 and the second of size 5 had mean 64.4 and variance 310.3. Is there evidence to indicate a difference in population means?*

We are testing

$$H_0 : \mu_1 = \mu_2 \quad \textit{versus} \quad H_1 : \mu_1 \neq \mu_2$$

The test statistic is

$$T = \frac{\bar{X} - \bar{Y}}{S_0\sqrt{1/n_1 + 1/n_2}}$$

which has a $t_{n_1+n_2-2}$ distribution if H_0 is true.

The pooled estimate of variance is given by

$$s_0^2 = \frac{5 \times 280.3 + 4 \times 310.3}{6 + 5 - 2} = 293.63$$

so $s_0 = 17.14$. The observed value of T is therefore

$$t = \frac{49.5 - 64.4}{17.14\sqrt{\frac{1}{6} + \frac{1}{5}}} = -1.40.$$

We can compare this value to a t_9 distribution. The P value will be given by $2 \times P(T < -1.40) = 2 \times P(T > 1.40)$. From Table 9, $P(T < 1.4) = 0.9025$. Thus $P(T > 1.4) = .0975$ and the P value is 0.195. So there is no evidence against H_0.

Example 8.2 *The reaction times, in hundredths of a second, of two groups of subjects taking a flashing-light stimulus are given below. The first group consisted of subjects who were new to the project while the subjects in the second group had taken part in previous experiments. Test if experience has had an effect on the mean response at the 5% significance level.*

New	*2.7*	*3.0*	*3.3*	*2.9*	*3.5*	*2.7*	*3.0*	*3.1*	*2.8*	*3.0*
Experienced	*2.7*	*2.5*	*3.0*	*2.7*	*2.6*	*2.5*	*2.9*	*2.7*		

Assumptions: The response time, X, of the new subjects is $N(\mu_1, \sigma^2)$ and the response time, Y, of the experienced subjects is $N(\mu_2, \sigma^2)$.

$$H_0 : \mu_1 = \mu_2 \qquad H_1 : \mu_1 \neq \mu_2$$

The test statistic is

$$T = \frac{\bar{X} - \bar{Y}}{S_0\sqrt{1/n_1 + 1/n_2}}$$

which has a $t_{n_1+n_2-2}$ distribution if H_0 is true.

$$n_1 = 10, \bar{x} = 3.0, (n_1 - 1)s_1^2 = 0.58, s_1^2 = 0.064$$
$$n_2 = 8, \bar{y} = 2.7, (n_2 - 1)s_2^2 = 0.22, s_2^2 = 0.031$$

The pooled estimate of variance is given by

$$s_0^2 = \frac{0.58 + 0.22}{10 + 8 - 2} = 0.05$$

The observed value of T is therefore

$$t = \frac{3.0 - 2.7}{\sqrt{0.05}\sqrt{\frac{1}{10} + \frac{1}{8}}} = 2.828.$$

We can compare this value to a t_{16} distribution. The rejection region for a 5% significance test is $|t| > 2.120$ so we reject the null hypothesis at the 5% level and conclude that experience does have an effect on the mean response.

If we are asked to estimate the difference in means between two independent normal samples with the same variance we would also want the corresponding confidence interval. This is given by

$$\bar{x} - \bar{y} \pm t_{n_1+n_2-2}(\alpha/2)s_0\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

Example 8.3 *For the data in Example 8.1 above a 95% confidence interval for the difference in means for the two samples would be*

$$\begin{aligned} 49.9 - 64.4 \pm 2.262 \times 17.14 \left(\frac{1}{6} + \frac{1}{5}\right)^{1/2} &= -14.5 \pm 23.48 \\ &= (-37.98, 8.98) \end{aligned}$$

Example 8.4 *For the data in Example 8.2 a 95% confidence interval for the difference in population means for the two samples would be*

$$\begin{aligned} 3.0 - 2.7 \pm 2.120 \times \sqrt{0.05} \left(\frac{1}{10} + \frac{1}{8}\right)^{1/2} &= 0.3 \pm 0.225 \\ &= (0.075, 0.525) \end{aligned}$$

Note that 0 does not belong to the 95% confidence interval agreeing with our finding that we could reject H_0 at the 5% significance level.

For the two sample t-test we have to make the assumption that the population variances are the same. Is this reasonable? In the next section we test this assumption.

8.2 F test for comparing two variances

We suppose now that the population variances may be different. We want to test a null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ versus $H_1 : \sigma_1^2 \neq \sigma_2^2$. We need a bit of theory. We know that

$$\frac{(n_1 - 1)S_1^2}{\sigma_1^2} \sim \chi^2_{n_1 - 1} \qquad \frac{(n_2 - 1)S_2^2}{\sigma_2^2} \sim \chi^2_{n_2 - 1}$$

and they are independent. We use the following theorem.

Theorem 8.1 *If the random variables C_1 and C_2 are independent and $C_1 \sim \chi^2_{\nu_1}$ and $C_2 \sim \chi^2_{\nu_2}$ then*

$$\frac{C_1/\nu_1}{C_2/\nu_2} \sim F^{\nu_1}_{\nu_2}.$$

It follows that

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

. If we can accept H_0 we can go ahead and use our two sample t-test. If we reject it we will have to think again. We will use a 5% significance test to make this decision. A suitable test statistic is

$$F = \frac{S_1^2}{S_2^2}$$

and it can be shown that $F \sim F_{n_2-1}^{n_1-1}$ if H_0 is true where $F_{n_2-1}^{n_1-1}$ is an F distribution with $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ degrees of freedom. The upper percentage points for an F distribution are given in Table 12. To construct a rejection region for our test we will also need the lower percentage points but these can be found using the result that

$$F^{-1} = \frac{S_2^2}{S_1^2} \sim F_{n_1-1}^{n_2-1}$$

if H_0 is true. Clearly values of F much greater or smaller than 1 will tend to give evidence against H_0.

A word of warning is necessary. The validity of the F test relies heavily on the underlying populations of our samples being normally distributed. If they are not the results can be misleading. If possible we should check the normality assumption using the normal probability plots and tests found in MINITAB.

Example 8.5 *Find the rejection region in terms of F if $n_1 = 6$ and $n_2 = 11$. We want $F_{10}^5(.025)$ and $F_5^{10}(.025)$. From Table 12(c) these are*

$$F_{10}^5(.025) = 4.236 \qquad F_5^{10}(.025) = 6.619.$$

Thus our rejection region would be $F > 4.236$ and $F^{-1} > 6.619$ or $F < 0.151$. We would accept H_0 if $0.151 < F < 4.236$.

Example 8.6 *For the data in example 8.1 the observed value of F is $280.3/310.3 = 0.9033$. To carry out a test of $H_0 : \sigma_1^2 = \sigma_2^2$ versus a two-sided alternative the rejection region is found as follows.*

$$F_4^5(.025) = 9.364 \qquad F_5^4(.025) = 7.388.$$

Thus our rejection region would be $F > 9.364$ *and* $F^{-1} > 7.388$ *or* $F < 0.135$. *Thus we can certainly accept the null hypothesis.*

Example 8.7 *Random samples were independently drawn from two normal populations. The first sample of size 13 had mean 9.5 and variance 93.3 and the second of size 11 had mean 14.0 and variance 25.2. Test the hypothesis that the populations have the same variance at the 5% significance level.*

The observed value of F *is* $93.3/25.2 = 3.70$. *To carry out a test of* $H_0 : \sigma_1^2 = \sigma_2^2$ *versus a two-sided alternative the rejection region is found as follows.*

$$F_{10}^{12}(.025) = 3.621 \qquad F_{12}^{10}(.025) = 3.374.$$

Thus our rejection region would be $F > 3.621$ *and* $F^{-1} > 3.374$ *or* $F < 0.296$. *Thus we reject the null hypothesis at the 5% significance level.*

We can find a confidence interval for the ratio σ_1^2/σ_2^2.

$$P\left[F_{n_2-1}^{n_1-1}(.975) < \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} < F_{n_2-1}^{n_1-1}(.025)\right] = 0.95$$

Using the result about lower percentage points we have

$$P\left[\frac{1}{F_{n_1-1}^{n_2-1}(.025)} < \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} < F_{n_2-1}^{n_1-1}(.025)\right] = 0.95$$

Rearranging so that σ_1^2/σ_2^2 is the subject, we have the 95% random interval

$$P\left[\frac{S_1^2/S_2^2}{F_{n_2-1}^{n_1-1}(.025)} < \frac{\sigma_1^2}{\sigma_2^2} < (S_1^2/S_2^2)F_{n_1-1}^{n_2-1}(.025)\right] = 0.95$$

Thus the 95% confidence interval for σ_1^2/σ_2^2 is

$$\left(\frac{s_1^2/s_2^2}{F_{n_2-1}^{n_1-1}(.025)}, (s_1^2/s_2^2)F_{n_1-1}^{n_2-1}(.025)\right)$$

Example 8.8 *For the data in Example 8.7 the 95% confidence interval for* σ_1^2/σ_2^2 *is*

$$\left(\frac{3.70}{3.621}, 3.70 \times 3.374\right) = (1.02, 12.48)$$

If we are interested in Example 8.7 in going on to test the hypothesis that the population means are equal how should we proceed? Since we can't assume the population variances are equal we cannot use the two sample t-test. We discuss this in the next section.

8.3 An approximate test when variances are unequal

If we wanted to test equality of two means when we knew the (different) population variances we would use the test statistic

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

which would have a standard normal distribution if the null hypothesis was true. If the variances were unknown it would seem natural to use the test statistic

$$T^* = \frac{\bar{X} - \bar{Y}}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}.$$

Unfortunately the distribution of T^* is not known exactly. We can, however, approximate it by a t distribution with ν^* degrees of freedom where

$$\nu^* = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\left(\frac{s_1^4/n_1^2}{n_1-1} + \frac{s_2^4/n_2^2}{n_2-1}\right)}.$$

(I won't expect you to remember this formula!)

Note that in general ν^* is not an integer but we could interpolate in the t distribution tables. MINITAB takes the integer below ν^* and this is perhaps preferable.

Example 8.9 *We saw in Example 8.7 that we could not assume that the population variances were equal. We can test the equality of the population means using T^*.*

$$H_0 : \mu_1 = \mu_2 \quad \textit{versus} \quad H_1 : \mu_1 \neq \mu_2$$

The test statistic is

$$T^* = \frac{\bar{X} - \bar{Y}}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}.$$

which has an approximate t_{ν^*} distribution if H_0 is true where

$$\nu^* = \frac{(93.3/13 + 25.2/11)^2}{\left(\frac{93.3^2/13^2}{12} + \frac{25.2^2/11^2}{10}\right)} = 18.6.$$

The observed value of the test statistic is $t^* = -5.5/3.077 = -1.787$. *If we use a two sided test with* $\alpha = 0.05$ *the rejection region is approximately* $\{t^* : |t^*| > 2.096\}$ *so we don't reject* H_0 *at the 5% significance level.*

There is some dispute about the use of this approximate t-test. Some books recommend that it is always used, however close the sample variances, because the F test relies heavily on normality. Others argue that the approximate test has lower power than the two sample t-test and if the sample variances are close together it is better to use the two sample t-test. In this course we will adopt the latter position, checking the equality of variances by the F test and only using the approximate procedure if there is evidence against the variances being equal.

In practice unless the two sample variances are very different, in which case we will probably use the approximate test, the difference in answers between the two methods is minimal.

We can find the approximate confidence interval for the difference in means as

$$\bar{x} - \bar{y} \pm t_{\nu^*}(\alpha/2)\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}.$$

Example 8.10 *For the data in Example 8.7 the approximate 95% confidence interval for* $\mu_1 - \mu_2$ *is*

$$-5.5 \pm 2.096 \times 3.077 = -5.5 \pm 6.450 = (-11.95, 0.95).$$

8.4 Matched pairs t-test

One of the assumptions we make in the two sample t-test is that the two samples are independent. If they are not we can use another test called the matched pairs t-test. This test is appropriate if the measurements are taken of pairs of similar subjects. For example, we might have pairs of twins, pigs from the same litter, a pair of measurements on the same individual or pairs of patients who have been matched to be similar. We would expect the measurements on such similar individuals to be similar. This violates the independence assumption needed for a two sample t-test. How do we analyse

such data? We find the differences for each pair and then do a 1 sample t-test on the differences. We are assuming that the differences are normally distributed, which we should check using a normal plot in MINITAB, with an unknown mean and variance. We test the null hypothesis that this mean is zero.

Example 8.11 *Sixteen patients sampled at random were matched by age and weight. One of each pair were assigned at random to treatment A and the other to treatment B. A blood test of a certain chemical produced the following results*

A	*14.0*	*5.0*	*8.6*	*11.6*	*12.1*	*5.3*	*8.9*	*10.3*
B	*13.2*	*4.7*	*9.0*	*11.1*	*12.2*	*4.7*	*8.7*	*9.6*

Test whether there is a difference in the two treatments. Find a 90% confidence interval for the mean difference in the treatments.

The differences are $+0.8, +0.3, -0.4, +0.5, -0.1, +0.6, +0.2, +0.7$. *The mean difference is* $\bar{d} = 0.325$, *the variance of the differences is* $s_d^2 = 0.1707$ *so the standard deviation is* $s_d = 0.413$. *The null hypothesis is* $\mu_d = 0$ *versus an alternative that* $\mu_d \neq 0$. *The test statistic is*

$$T = \frac{\bar{d}\sqrt{n}}{s_d}$$

which has a t distribution with 7 degrees of freedom if H_0 *is true. The observed value of t=2.226. Comparing this with a* t_7 *distribution* $P(t_7 < 2.226) = .9692$ *so the P value is* $2(1 - 0.9692) = 0.0616$ *so there is weak evidence against the null hypothesis.*

A 90% confidence interval is of the form

$$\begin{aligned} \bar{d} \pm t_7(.05)\frac{s_d}{\sqrt{n}} &= 0.325 \pm 1.895 \times \frac{0.413}{\sqrt{8}} \\ &= 0.325 \pm 0.277 \\ &= (0.048, 0.602) \end{aligned}$$

Such matching is a simple example of a *designed experiment* with *blocking*. Here we have blocks of size 2 but in more complicated examples we might want, for example, to compare 5 animal feeds. We could do this using 5

animals from the same litter. It is important that biases are not introduced into the experiment so we allocate diets to animals at random within each litter. If we are using the same person twice in a study, once with each treatment, it is important to choose the order in which they receive the treatments randomly. With a drug treatment it may be necessary to allow time between the two treatments so that the first drug is not still affecting the subject when the second drug is taken. If the subject is a patient with a long term illness requiring continuous treatment this could be a problem. In such a clinical trial it is also important, if practically possible, that the patient receiving the treatment does not know which treatment he is receiving and the doctor assessing their improvement also does not know as again this might introduce biases. The whole subject of design of experiments is a huge one in its own right.

8.5 Test of two proportions

Suppose we have collected data in an opinion poll on whether the budget was good for the country from men and women and we want to test the hypothesis that the proportions thinking it was good are equal. Suppose we question n_1 men and n_2 women and x_1 men and x_2 women say it was good. The estimate of the proportions thinking it was good will be $\hat{p_1} = x_1/n_1$ and $\hat{p_2} = x_2/n_2$. We can estimate the difference in proportions by $\hat{p_1} - \hat{p_2}$. To test the hypothesis that the population proportions are equal $H_0 : p_1 = p_2$ we need a test statistic with known distribution if H_0 is true. If n_1 and n_2 are large then by the central limit theorem the distribution of $\hat{p_1} - \hat{p_2}$ is normal. The variance of $\hat{p_1} - \hat{p_2}$ is

$$\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}.$$

To estimate this quantity note that if H_0 is true then $p_1 = p_2 = p$ and the best estimate of p is $\hat{p} = (x_1 + x_2)/(n_1 + n_2)$. Thus our test statistic is

$$Z = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

which has a standard normal distribution if H_0 is true.

Example 8.12 *Of 1000 men asked 450 thought the budget was good for the country and of 950 women 390 thought it was good. Test the hypothesis at the 5% level that the same proportion of men and women thought it was good.*

The null hypothesis is $H_0 : p_1 = p_2$ *against* $H_1 : p_1 \neq p_2$.

$\hat{p_1} = 450/1000 = 0.45$, $\hat{p_2} = 390/950 = 0.4105$, $\hat{p} = 840/1950 = .4308$ *The test statistic* Z *given above has a standard normal distribution if* H_0 *is true. The observed value of* Z *is*

$$z = \frac{0.45 - 0.4105}{\sqrt{0.4308 \times 0.5692 \times \left(\frac{1}{1000} + \frac{1}{950}\right)}} = 1.759$$

The rejection region is $\{z : |z| > 1.96\}$ *so we don't reject* H_0.

Note that we actually know a different way to do this problem. We could write our data in the form of a 2×2 table.

	Good	Bad	Total
Men	450	550	1000
Women	390	560	950
Total	840	1110	1950

Using the formula we had for a 2×2 table the value of X^2 is

$$\frac{(450 \times 560 - 550 \times 390)^2 1950}{1000 \times 950 \times 840 \times 1110} = 3.096$$

Under the null hypothesis that the distribution of men and women are the same $X^2 \sim \chi_1^2$ and we would fail to reject at the 5% level.

Note that $\sqrt{3.096} = 1.759$. Is this a coincidence? No. It is a fact that if $X \sim N(0, 1)$ then $X^2 \sim \chi_1^2$. You can show that the observed value of z does equal the square root of the formula given before. The two P values for the test would be identical. Also note that we said for smaller sample sizes you should use a continuity correction.

The confidence interval for the difference in proportions is not quite what you would expect from the test. Because we are not assuming that $p_1 = p_2$ we estimate the variance differently. The 95% confidence interval is given by

$$\hat{p_1} - \hat{p_2} \pm 1.96 \times \sqrt{\frac{\hat{p_1}(1 - \hat{p_1})}{n_1} + \frac{\hat{p_2}(1 - \hat{p_2})}{n_2}}.$$

Example 8.13 *For the opinion poll data the 95% confidence interval is given by*

$$\begin{aligned} .45 - .4105 \pm 1.96\sqrt{\frac{(.45)(.55)}{1000} + \frac{(.4105)(.5895)}{950}} &= .0395 \pm .0439 \\ &= (-0.0044, 0.0834) \end{aligned}$$

9 Two Dimensional Random Variables

Definition 9.1
Let S be a sample space associated with an experiment E, and X_1, X_2 be functions, each assigning a real number $X_1(e), X_2(e)$ to every outcome $e \in E$. Then the pair (X_1, X_2) is called a two-dimensional random variable. The range space of the two-dimensional random variable is

$$R_{(X_1,X_2)} = \{(x_1, x_2) : x_1 \in R_{X_1},\ x_2 \in R_{X_2}\} \subset R^2.$$

Definition 9.2
The cumulative distribution function of the random variable (X_1, X_2) is

$$F_{(X_1,X_2)}(x_1, x_2) = P(X_1 \leq x_1, X_2 \leq x_2) \tag{9.1}$$

9.1 Discrete Two-Dimensional Random Variables

If possible values of (X_1, X_2) are countable, then the variable is discrete. The c.d.f. of a discrete r.v. (X_1, X_2) can be written as

$$F(x_1, x_2) = \sum_{x_{2j} \leq x_2} \sum_{x_{1i} \leq x_1} p(x_{1i}, x_{2j}) \tag{9.2}$$

where $p(x_{1i}, x_{2j})$ denotes the *joint probability function*

$$p(x_{1i}, x_{2j}) = P(X_1 = x_{1i}, X_2 = x_{2j}).$$

That is:

1. $p(x_{1i}, x_{2j}) \geq 0$, for all i, j
2. $\sum_{all\ j} \sum_{all\ i} p(x_{1i}, x_{2j}) = 1$

9.2 Continuous Two-Dimensional Random Variables

If the possible values are some uncountable set in the Euclidean plane, then the variable (X_1, X_2) is continuous, for example values might be in the range

$$R_{(X_1,X_2)} = \{(x_1, x_2) : a \leq x_1 \leq b, c \leq x_2 \leq d\}$$

for some real a, b, c, d.
The c.d.f. of a continuous r.v. (X_1, X_2) can be written as

$$F(x_1, x_2) = \int_{-\infty}^{x_2} \int_{-\infty}^{x_1} f(x, y) dx dy, \tag{9.3}$$

where $f(x_1, x_2)$ is the probability density function such that

1. $f(x_1, x_2) \geq 0$ for all $(x_1, x_2) \in R^2$
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2) dx_1 dx_2 = 1$.

Also

$$P(a \leq X_1 \leq b, c \leq X_2 \leq d) = \int_c^d \int_a^b f(x_1, x_2) dx_1 dx_2 \tag{9.4}$$