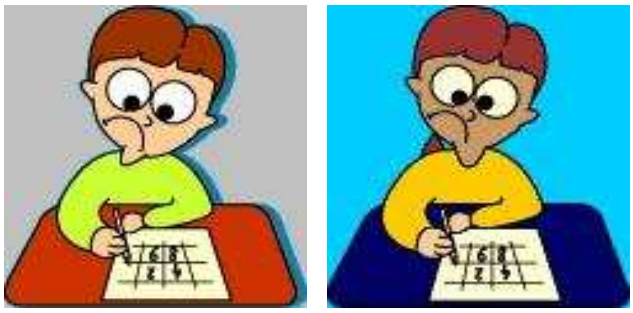
MAS1403/ACE2013

Quantitative Methods for Business Management

Statistics for Marketing and Management



Revision material for Semester 2

Dr. Lee Fawcett, 2008–2009

Exam structure
The exam for MAS1403/ACE2013 has two sections: **Section A**, which will have about five 'short' questions (and is worth a total of 50 marks), and **Section B**, which will have four 'long' questions, of which you choose to answer any two. All questions in Section B are worth 25 marks, and so the maximum mark for the exam is 100.

Open–book exam?
The exam for students registered on MAS1403 is **open–book**; however, those of you who have taken university exams before will be aware that desk space is often limited, and so you should really be selective about what you bring into the exam. You can bring in anything you like – course notes, text books, whatever – but too much stuff will probably get in your way and just cause a nuisance! And don't forget, if you rely heavily on notes and don't attempt to *learn* the techniques then you'll probably spend most of your time rummaging through reams of paper trying to find examples similar to the exam questions. So "open–book" exam doesn't mean "easy" exam!

This year, the exam for ACE2013 students – and all other students – will also be open–book!

Past papers
If you decide to use past exam papers to aid your revision, then you should note that before last year MAS1403 had the module code MAS187; similarly, the old module code for ACE2013 was AEF258. To download past papers, go to `http://www.ncl.ac.uk/exam.papers/`

Timing!
The exam is 2 hours long. Use your time wisely – there will be quite a lot of work to do in those 2 hours! I suggest spending the first hour on questions in Section A, and then moving to Section B. You'll probably find that Section B might not take a full hour to complete, so at least this time–frame cuts you some slack to go back to earlier questions you may have had problems with.

Exam/assignment/CBA weighting
The exam contributes 60% to your final mark for this module. Each of the assignments is worth 10% each (there were two of them) and the total contribution of the CBAs is 20%. So don't be frightened by the exam – in theory, this module can be passed with full marks from both assignments and the CBAs alone!

We won't try to catch you out in the exam. If you've worked steadily throughout the year, attending tutorials and doing the work from each chapter as you've gone along, you should have nothing to worry about! The questions set in the exam will be no more difficult than those set each week in the tutorials!

Remember to attend the exam for the module you are registered for – if you are registered for ACE2013 but attend the MAS1403 exam, you may be disqualified!

Semester 2 syllabus

Use this checklist to tick things off as you revise them!

- **Constructing confidence intervals**
  - when the population variance is *known*;
  - when the population variance is *unknown*.
- **Hypothesis tests for one mean**
  - when the population variance is *known*;
  - when the population variance is *unknown*.
- **Hypothesis tests for two means**
  - when both population variances are *known*;
  - when both population variances are *unknown*.
- **Goodness–of–fit tests using the χ^2 distribution**
- **Tests of independence using the χ^2 distribution**
- **Correlation and linear regression**
  - drawing/interpreting scatterplots;
  - calculating/interpreting the correlation coefficient;
  - estimating the linear regression model $Y = \alpha + \beta X + \epsilon$;
  - using a linear regression model to make predictions.
- **Time series and forecasting**
  - *describing* a time series;
  - isolating the *trend*;
  - isolating the *seasonal effects*;
  - forecasting.
- **Linear programming**
  - *formulating* linear programming problems;
  - displaying linear programming problems *graphically*;
  - *solving* linear programming problems.

1 Estimation

Recall from semester 1 that data can be summarised in two ways: **graphically** and **numerically**. When we summarise data numerically, we usually quote one measure of *location* (or average) and one measure of *spread*. The most popular measure of location is the **sample mean**, though if our data are skewed we often prefer to use the **sample median**.

The sample mean ($\bar{x}$) is a **point estimate** of the population mean (μ). We find this value by adding up our observations and dividing by how many we've got. In semester 2, we've looked at **interval estimates**, or **confidence intervals**, for the population mean.

1.1 Population variance σ^2 known

If the population variance σ^2 is known (if so, the question will say "the population variance is ..." or "the process variability is known to be ..."), then we have the following formulae for a 90%, 95% and 99% confidence interval for μ:

- **90% confidence interval for μ**

$$\bar{x} \pm 1.645 \times \sqrt{\sigma^2/n}$$

- **95% confidence interval for μ**

$$\bar{x} \pm 1.96 \times \sqrt{\sigma^2/n}$$

- **99% confidence interval for μ**

$$\bar{x} \pm 2.576 \times \sqrt{\sigma^2/n}$$

where $\bar{x}$ and n are the sample mean and sample size, and σ^2 is the population variance.

1.2 Population variance σ^2 unknown

If the population variance is unknown, then a confidence interval for μ is given by

$$\bar{x} \pm t \times \sqrt{s^2/n}$$

where t is a value obtained from t–tables (Table 1.1 in the notes) and $\bar{x}$, s^2 and n are the sample mean, sample variance (found using a calculator if it's not given) and the sample size.

1.3 Exam–style question

"Aphroditair" are an internet–based budget airline offering cheap flights to the Greek islands. From a random sample of 14 customers with Aphroditair, the mean price of flights to Kefalonia in September was £136 with a standard deviation of £25.50. Obtain a 95% confidence interval for the average price of flights to Kefalonia in September with Aphroditair.

2 Hypothesis tests for the mean

In Chapters 2 and 3 of the notes we looked at an alternative approach to statistical inference for the population mean through **hypothesis tests**. We discussed two scenarios:

1. **Hypothesis tests for one mean**, where one sample mean is compared to a target/proposed/hypothesised value for the population mean.
2. **Hypothesis tests for two means**, where two sample means are directly compared to one another.

Remember, the aim of such hypothesis tests is to use the information in our sample(s) to make conclusions about our population(s). In this section, we focus on tests for one mean.

Here, from a single population we draw a single sample, and we estimate the population mean μ with the sample mean $\bar{x}$. We'd then like to see how convincing a proposal for the population mean is, based on the information in our sample.

Case 1: Population variance σ^2 known

Steps 1 and 2 *(hypotheses)*
The null hypothesis is

$$H_0 \;:\; \mu = c.$$

If the question asks you to find out if the population mean is less than c, the alternative would be

$$H_1 \;:\; \mu < c.$$

Similarly, if the question wanted to know if the population mean was larger than c, the alternative would be

$$H_1 \;:\; \mu > c.$$

These are known as "one–tailed" tests. Otherwise, we just use the general, "two–tailed" alternative

$$H_1 \;:\; \mu \neq c.$$

Step 3 *(calculate the test statistic)*
The test statistic if the population variance is known is

$$z \;=\; \frac{|\bar{x} - \mu|}{\sqrt{\sigma^2/n}}.$$

Step 4 *(find the p–value)*
Since the population variance is known, we use tables of probabilities for the normal distribution to obtain the 10%, 5% and 1% critical values to which our test statistic can be compared (Table 2.2 in the notes); we then obtain a range for our p–value.

Step 5 *(form your conclusion)*
To get full marks, you need to

1. Use your p-value to state the strength of evidence against H_0 (see Table 2.1)
2. Say whether you're going to keep or reject H_0
3. Write a sentence in the context of the question

Case 2: Population variance σ^2 unknown

Steps 1 and 2 are as before. However, we now have:

Step 3 *(calculate the test statistic)*
The test statistic is now

$$t = \frac{|\bar{x} - \mu|}{\sqrt{s^2/n}}.$$

This is similar to before, but is now called t to remind ourselves that we compare this value to t–tables and not standard normal tables; also, the *population* variance has been replaced with its *sample* equivalent s^2.

Step 4 *(find the p–value)*
You should now use t–tables (Table 2.3) to find the 10%, 5% and 1% critical values with which to compare your test statistic. Remember, the degrees of freedom is $\nu = n - 1$.

Step 5 – the conclusion bit – is exactly the same as always!

3 Tests for two means

Here, we compare the means from two independent samples instead of comparing a single sample mean to a hypothesised value. Again, we have two situations to consider: the case when *both* population variances are known, and the case when *both* are unknown.

Case 1: Both population variances σ_1^2 and σ_2^2 known

Steps 1 and 2 *(hypotheses)*
The null hypothesis is always

$$H_0 \ : \ \mu_1 = \mu_2.$$

If the question wants to know if one population mean is bigger than the other, the alternatives might be

$$\begin{aligned} H_1 \ &: \ \mu_1 > \mu_2 \qquad \text{or} \\ H_1 \ &: \ \mu_1 < \mu_2. \end{aligned}$$

These are examples of one–tailed alternatives. Otherwise, if we are just testing for a general difference between two populations, we have

$$H_1 \ : \ \mu_1 \neq \mu_2.$$

Step 3 *(calculate the test statistic)*
Here, the test statistic is

$$z \ = \ \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}},$$

where $\bar{x}_1$, $\bar{x}_2$, n_1 and n_2 are the sample means and sample size of samples 1 and 2, and σ_1^2 and σ_2^2 are the corresponding population variances.

Step 4 *(find the p–value)*
Since both population variances are known in this case, we use standard normal tables to obtain our critical values (Table 2.2), and then locate our test statistic to obtain a range for the p–value.

Step 5 *(form your conclusion)*
Interpretation of the p–value is exactly the same as always. Remember to write a sentence in the context of the question.

Case 2: Both population variances σ_1^2 and σ_2^2 unknown

Steps 1 and 2 remain unchanged. However, we now have:

Step 3 *(calculate the test statistic)*
The first thing you should do is calculate the "pooled standard deviation", which is given by

$$s = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}.$$

The test statistic is then

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{s \times \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}},$$

where $\bar{x}_1$, $\bar{x}_2$, s_1^2 and s_2^2, are the sample means and variances of samples 1 and 2, and n_1 and n_2 are the corresponding sample sizes.

Step 4 *(find the p–value)*
We now use t–tables (Table 2.3), with degrees of freedom $\nu = n_1 + n_2 - 2$, to obtain the 10%, 5% and 1% critical values in order to find a range for our p–value.

Step 5 – the conclusion bit – exactly the same as before!

Exam–style question

"El Cheapo" are another internet–based flight company also offering cheap flights to Greece. From a random sample of 16 customers with El Cheapo, the mean price of flights to Kefalonia in September was £120 with a standard deviation of £28.30. Test the null hypothesis that there's no difference between the price of flights to Kefalonia between Aphroditair and El Cheapo. The pooled standard deviation for this problem is £27.04.

4 Goodness–of–fit tests

4.1 Quick review

In Chapter 3 we addressed the question of whether our data followed any **pattern**, or adhered to a specified **probability distribution**. The procedure for such a "goodness–of–fit" test is outlined below.

Steps 1 and 2 *(hypotheses)*
If we think our data might follow a Poisson distribution, then we have

$$\begin{aligned} H_0 &: \text{Our data follow a Poisson distribution} \qquad \text{versus} \\ H_1 &: \text{Our data do } \textit{not} \text{ follow a Poisson distribution.} \end{aligned}$$

Obviously, the Poisson distribution can be replaced with any other probability distribution which we might want to test.

Step 3 *(calculate the test statistic)*
Remember, the test statistic for a goodness–of–fit test is

$$X^2 \;=\; \sum \frac{(O-E)^2}{E},$$

where O and E represent "observed" and "expected" frequencies respectively. The observed frequencies are just the numbers we are given. The expected frequencies for each category are the numbers we would expect to see if our data follows the probability distribution we think it might (e.g. Poisson, binomial). We use the formula for the probability distribution we are testing to obtain expected *probabilities*, and then obtain our expected *frequencies* by multiplying each probability by the total sample size. Calculating the test statistic is easier if you draw up the following table:

Observed (O)	Expected (E)	$\frac{(O-E)^2}{E}$
$\vdots$	$\vdots$	$\vdots$

The test statistic is then found by adding up all the values in the last column. Remember, all expected frequencies must be ≥ 5; to achieve this, adjacent categories can be 'pooled'.

Step 4 *(find the p–value)*
Use Table 4.1 of the notes (χ^2 tables) to obtain the 10%, 5% and 1% critical values for this test which can then be used to find a range for our p–value. The degrees of freedom is given by

$$\nu \;=\; (\text{number of categories after pooling}) - (\text{number of parameters estimated}) - 1.$$

Step 5 *(form your conclusion)*
As always, we can use Table 2.1 to interpret our p–value. Remember, to reject H_0 in any hypothesis test we need a p–value of 5% or less; i.e. we need at least "moderate" evidence against H_0.

4.2 Exam–style question

The number of accidents per day in a steelworks was recorded over the period of a year; the results are shown in the table below.

Number of accidents per day	Frequency
0	175
1	125
2	46
3	14
4	5
5 or more	0

Propose a distribution that might fit these data, and test to see whether or not it is appropriate.

5 Tests of independence

It is usually fairly obvious if you are required to perform a χ^2 test of independence, since the data will be given in the form of a **contingency table**. This test is used to determine whether or not there is any association between two categorical variables.

5.1 Quick review

The steps involved in a χ^2 test for independence are shown below.

Steps 1 and 2 *(hypotheses)*
No matter what the two categorical variables are, if you are testing to see whether they are independent or not the hypotheses are

H_0 : There is no association between the two categorical variables versus
H_1 : There *is* an association between the two categorical variables.

Step 3 *(calculate the test statistic)*
The test statistic (as for the χ^2 goodness–of–fit test) is

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

where O and E represent observed and expected frequencies (respectively). We don't have to worry about any nasty probability distributions here, because we're just testing for independence. We can get directly to our expected frequencies by using the formula

$$E = \frac{\text{row total} \times \text{column total}}{\text{overall sample size}}$$

for each cell in the contingency table.

As before, once we have the expected frequencies, the test statistic can be calculated very easily by drawing up the following table:

Observed (O)	Expected (E)	$\frac{(O-E)^2}{E}$
$\vdots$	$\vdots$	$\vdots$

and then adding up the values in the final column.

Step 4 *(find the p–value)*
The 10%, 5% and 1% critical values are found in Table 4.1. This time, the degrees of freedom is given by

$$\nu = (\text{number of rows} - 1) \times (\text{number of columns} - 1).$$

Step 5 *(form your conclusion)*
This is getting boring now – exactly as before! If the test statistic is greater than the critical value, reject the null hypothesis in favour of the alternative. Otherwise, stick with H_0! If we reject the null hypothesis, that's equivalent to saying there is sufficient evidence to suggest that the two categorical variables *are* associated, or are *not* independent.

5.2 Exam–style question

The following table includes data on the number of days sick leave taken by managerial and non–managerial employees of a particular organisation. Is there an association between type of employee and number of days sick leave?

	Non–managerial	Managerial	**Total**
0–10 days	22	24	46
11–20 days	28	16	44
21 or more days	50	10	60
Total	100	50	150

6 Correlation and linear regression

Chapter 6 in the notes is concerned with *quantifying* the dependence between two continuous variables (correlation) and *modelling* this dependence if a linear relationship seems apparent (linear regression). Suppose we have two random variables (say X and Y), and our data consists of n pairs of observations on these variables, i.e.

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

The X variable is usually our **explanatory** variable and the Y variable our **response** variable. The response variable is usually the quantity which we'd like to make predictions of *given the explanatory variable*. The example in the notes looked at average monthly temperatures and ice cream sales. From an economic/business point of view, it is clear that ice cream sales should be the response (or Y) variable, and the average monthly temperature the explanatory (or X) variable, since we'd probably be interested in predicting sales of ice cream given what we know about the average temperature. We probably *wouldn't* be interested in predicting what the average temperature for, say, May will be given the sales of ice cream!

The first step in any correlation/regression analysis is usually to construct a scatterplot for our data, with the explanatory variable along the horizontal axis and the response along the vertical axis. From this, we can visualise any dependence between our variables.

- Does there seem to be any relationship between X and Y at all?
- Is there a **linear** relationship?
  - If so, is this relationship **positive** ("uphill")?
  - Is it **negative** ("downhill")?
  - Is this linear relationship **strong**?
- Does there seem to be a more complex relationship?

If you're asked to draw a scatterplot, make sure you label the axes and give your plot a title!

6.1 Correlation

The correlation coefficient, r, quantifies the amount of dependence between our two variables. This is how you should interpret it:

$r = -1$	$r = 0$	$r = 1$
'Perfect' negative dependence. All points lie on a straight line. Downhill slope.	Complete (linear) independence. Random scatter of points.	'Perfect' positive dependence. All points lie on a straight line. Uphill slope.

The formula for r is

$$r = \frac{S_{XY}}{\sqrt{S_{XX} \times S_{YY}}},$$

where

$$\begin{aligned} S_{XY} &= \left(\sum xy\right) - n\bar{x}\bar{y}, \\ S_{XX} &= \left(\sum x^2\right) - n\bar{x}^2, \qquad \text{and} \\ S_{YY} &= \left(\sum y^2\right) - n\bar{y}^2. \end{aligned}$$

Beware! A correlation coefficient close to zero does not imply no relationship at all, just no *linear* relationship!

6.2 Simple linear regression

Any straight line has the equation $Y = \alpha + \beta X$. If we can assume a straight line relationship between our variables, then we can assume the **regression equation**

$$Y = \alpha + \beta X + \epsilon,$$

where ϵ represents the "scatter" about this line. We estimate α and β using

$$\begin{aligned} \hat{\beta} &= \frac{S_{XY}}{S_{XX}} \qquad \text{and} \\ \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x}. \end{aligned}$$

Using these formulae to estimate α and β give the 'best' line of best fit through our data, i.e. the line which lies closest, on average, to all the points in the scatterplot. We estimate the scatter, ϵ as zero.

6.3 Exam–style question

A bar is trying to plan its requirements for next month. The manager believes there is a relationship between sales and the average temperature in the month. To investigate, he has collected figures for sales (y, in thousands of pounds) and average temperature (x, in ^{o}C) over the last year. His results, along with some summary statistics, are shown below.

x	18	21	20	19	13	6	4	3	5	6	11	13
y	13.8	15.3	15.1	14.8	13.1	7.8	13.3	8.2	9.6	10.2	9.8	11.5

$$\sum x = 139 \qquad \sum y = 111.7$$

$$\sum x^2 = 2107 \qquad \sum y^2 = 1772.7 \qquad \sum xy = 1814.3$$

(a) Produce a scatterplot for these data, and comment on the relationship between average temperature and sales.

(b) Calculate the sample correlation coefficient, and comment.

(c) Perform a linear regression analysis on these data, and obtain the linear regression equation.

(d) Plot the regression line on your scatterplot in part (a).

(e) If the average temperature next month is forecast as 17^{o}C, predict the level of sales the bar manager can expect.

7 Using `Minitab`

Chapter 7 in the notes was all about using `Minitab` to calculate confidence intervals, carry out hypothesis tests and perform linear regression.

8 Time Series and Forecasting

8.1 Describing time series

Before any formal statistical analysis takes place, you should examine a **time series plot** of any data and be able to describe any patterns in words. Things you should look out for are

- Trend
- Seasonality/cyclic variation
  - Simple cyclic patterns
  - Complex cyclic patterns
- Outliers?
- Stationarity

8.2 Isolating the trend

In this course, we have looked at how to estimate any **trend** in a time series through the method of **moving averages**, where we average over the cycle around an observation. These moving averages can then be plotted on the original time series plot and should illustrate changes in the underlying level of the process.

This trend can then be modelled (if it is linear) by using simple linear regression techniques (see Chapter 6). We use the regression model

$$Y = \alpha + \beta T + \epsilon,$$

where T represents "time". We estimate α and β using the same methods as in Chapter 6.

8.3 Isolating the seasonal effects

To calculate **seasonal effects** in our series, we

1. calculate *seasonal deviations* by subtracting the each moving average from its original observation;
2. find the mean of the seasonal deviations for each season (the *seasonal means*);
3. find the overall mean for *all* seasonal deviations;
4. the *seasonal effects* are then found by subtracting the overall mean from each seasonal mean.

Remember, the seasonal means should add up to one. If they don't, you need to adjust them! (see page 77 in the notes to see how to do this)

8.4 Forecasting

We can now forecast into the future by using the linear regression equation for the trend (by substituting an appropriate value for T), and then adding in the seasonal effect for the corresponding season of our forecasted value.

8.5 Exam–style question

The following data are the four–monthly sales figures for an electrical store (in thousands of pounds).

	Jan–Apr	May–Aug	Sep–Dec
2003	9	14	11
2004	10	15	13
2005	11	17	14

(a) Produce a time series plot for these data, and comment.

(b) Calculate the moving averages for these data.

(c) Use the moving averages in part (b) to estimate the linear trend

$$Y = \alpha + \beta T + \epsilon.$$

(d) Calculate the seasonal effects for each of Jan–Apr, May–Aug and Sep–Dec.

(e) Use the regression equation in part (c), and the seasonal effects in part (d), to forecast sales for May–Aug 2006.

9 Linear programming

This section in the revision guide reviews the linear programming material we covered in weeks 9 and 10 (Chapters 9 and 10 in the notes). Since it's probably still quite fresh in your head, the main points will be briefly highlighted before we look at an exam–style question.

9.1 Quick review

In an exam question on linear programming, you'll probably be asked to do three things:

1. Formulate a real–life scenario as a linear programming problem.
2. Draw a suitable diagram to enable the problem to be solved graphically.
3. Solve the problem (using the graph or by solving simultaneous equations).

To **formulate** a linear programming problem, remember to:

- State clearly the **decision variables**;
- identify your **constraints**;
- identify the **objective function**.

Remember, it might help if you draw up a table which summarises all the information given in the question.

To represent the problem **graphically**, all you have to do is plot the inequalities you identified as "constraints". Remember to shade out all the unwanted bits and clearly label the **feasible region**. You should also plot the objective line, and draw an arrow which shows the direction of this line.

To **solve** the problem, you just have to identify the point in your feasible region which optimises your objective. This is usually at the intersection of two lines on your plot. Once you have done this, the solutions can just be "read off"! You should also know how to solve these problems algebraically.

9.2 Exam–style question

A chocolate manufacturer produces two types of chocolate bar, Asteroids and Blackholes. Production of an Asteroid bar uses 10g of cocoa and 1 minute of machine time, whereas a Blackhole bar requires 5g of cocoa and 4 minutes of machine time. Altogether, 2000g of cocoa and 480 minutes of machine time are available each day. The manufacturer must make at least 50 Asteroid and 50 Blackholes each day to keep up with demand. The manufacturer makes 10p profit from each Asteroid bar and 20p profit from each Blackhole bar.

(a) Formulate the chocolate manufacturers situation as a linear programming problem.

(b) Draw a suitable diagram to enable the problem to be solved graphically, indicating the feasible region and the direction of the objective line.

(c) Use your diagram to find the company's minimum and maximum profit, $£P$.

(d) Now solve this problem algebraically to verify your solution to part (c).