NB. The publishers have updated the compansion website for Stock & Watson. The urls given here will no longer work. You can instead access the data by doing a websearch on "Stock Watson companion" and follow the links.

# Quantitative Economics Tutorial 2: Statistics

The questions in this worksheet cover probability and sampling, and introductory statistics, at the level of the lectures in the first two weeks of Trinity Term. It is not expected that students will do all of these questions; tutors will select questions from this or other sources.

*References:*

Lecture slides for lectures 1-6.

James H. Stock and Mark W. Watson, *Introduction to Econometrics*, Brief Edition, or complete Edition. Chapters Chapter 2 and 3. (Covers almost everything you need, but concisely.)

R. L. Thomas, *Using Statistics in Economics*, Prerequisites, and Chapters 1–6

T. H. Wonnacott and R. J. Wonnacott, *Introductory Statistics for Business and Economics*, Chapters 1–5 (Probability) and 6–9 (Statistics)

Sheldon Ross, *A First Course in Probability*, Chapers 1-7.

. . . and there are many other good introductions to probability and statistics.

*Questions:*

1. If the price of unleaded petrol at UK filling stations is a random variable with mean 120.8p per litre, and standard deviation 4.9p, use the Central Limit theorem to determine the probability that the average price in a random sample of 50 filling stations is below 122p.

2. Suppose that students' marks on the economics prelims paper are normally distributed with mean 61 and standard deviation 9.5.

   (Assume that the number of colleges is sufficiently large that individual observations may be considered i.i.d.)

   (a) What is the distribution of the sample mean for a random sample of size $n$?

   (b) In a random sample of 10 students, what is the probability that their average mark exceeds 63?

   (c) Suppose that you have a sample of 10 students that is selected by choosing a college at random, and then choosing 10 students at random from that college.

      i. What is the expected value of their average mark?

      ii. Explain why you cannot determine the variance of their average mark. Is it likely to be higher or lower than the variance of the sample mean in random sample of 10 students? Explain the intuition for your answer.

3. $X$ is a random variable.

   (a) If $Y = a + bX$, what is the correlation between $X$ and $Y$ if (i) $b > 0$ (ii) $b < 0$ (iii) $b = 0$ ?

(b) Now suppose that $u$ is another random variable, with mean zero; that $u$ and $X$ are independent, and that $Y = a + bX + u$. Find E($Y$) and Var($Y$) and show that:

(i) $\text{Cov}(X, Y) = b\text{Var}X$

(ii) $\text{Corr}(X, Y) = b\sqrt{\dfrac{\text{Var}X}{b^2\text{Var}X + \text{Var}u}}$

What happens to the correlation between $X$ and $Y$ as the variance of u increases?

4. If $Z_1$, $Z_2$, $\ldots Z_n$ are independent Standard Normal random variables, the random variable $W$ defined by:

$$W = Z_1^2 + Z_2^2 + \ldots + Z_n^2$$

has a $\chi^2(n)$ distribution. Using the properties of the Standard Normal distribution (or the results from question 10 on the Tutorial 1 Worksheet), find the mean and variance of the $\chi^2(n)$ distribution.

5. The 1165 Oxford PPE applicants in 2007 achieved an average score of 60.86 on the TSA test, with a standard deviation of 8.02. Construct a 95% confidence interval for the population mean score.

6. (a) Consider a random sample of size $n$ from a Bernoulli distribution with parameter $p$. If the sample mean is $\bar{X}$, show that the sample variance is given by $s^2 = \frac{n}{n-1}\bar{X}(1 - \bar{X})$. Compare the sample mean and variance with the population mean and variance.

(b) In an opinion poll of 300 voters, 140 say that they will vote for the incumbent, and 160 for the rival candidate. Estimate the proportion of votes that will be obtained by the incumbent in the election. Calculate the sample variance. Find 95% and 99% confidence intervals for the incumbent's proportion of votes in the election.

7. Download the Excel datafile CPS92_08 from the Stock and Watson Companion Website: http://wps.aw.com/aw_stock_ie_3/178/45691/11696965.cw/index.html For Average Hourly Earnings in 1992:

(a) Find the sample mean, median, maximum and minimum values, standard deviation, skewness and kurtosis. Comment on the shape of the distribution.

[Use Excel functions – for example, to find the mean of the data in cells A2 to A10 you would type into a blank cell:

$$= \text{AVERAGE(A2:A10)}$$

and similarly for functions MEDIAN, MAX, MIN, STDEV, SKEW, KURT. Note that KURT calculates *excess kurtosis*.]

(b) Estimate the population mean of average hourly earnings and find the standard error of the estimate.

(c) Find a 95% confidence interval for the population mean.

(d) Construct a new variable equal to $\ln(ahe)$ and repeat the questions above. Compare the results for the two variables.

[To construct the new variable, type:

=LN(B2)

into cell F2. Then copy cell F2, select all the cells from F3 to F7606 (for all the 1992 data), and paste the formula into them.]

8. Suppose that you are studying the effect of macroeconomic conditions on average hours worked per week. In a sample of 400 workers, their average increase in weekly working hours over the last year is -0.5 hours, with a standard deviation of 5.7 hours. Test the hypothesis that there has been no change in hours worked, against (i) the alternative that hours have changed (ii) the alternative that hours have fallen. (In each case, state the null and alternative hypotheses, and the decision rule, clearly.) How would you decide which of these tests to use?

9. (a) Show that the sample variance $s^2$ is an unbiassed estimator of the population variance.

   (b) (Quite hard) For a random sample from the normal distribution, prove the result that $(n-1)s^2/\sigma^2$ has a $\chi^2(n-1)$ distribution, for the case $n = 2$. You may use the result stated in the lecture notes, that any linear combination of normally distributed random variables has a normal distribution.

10. **Youth unemployment hits record high**
by Alan Jones, Press Association, *11 November 2009*

"The number of 18 to 24 year-olds out of work rose by 24,000 over the three months to 74600, a rate of 18%, the highest since 1992. ... Work and Pensions Secretary Yvette Cooper [PPE 1990!] said: 'The figures show more people in work and a lot more young people taking up our offer of full-time education and training, which is welcome news'. "

   (a) Assuming that the true rate of youth unemployment is indeed 18%, find the probability of observing between 160 and 200 unemployed people in a random sample of 1000 eighteen to twenty-four year olds.
   [Hint: You may use the normal approximation to the binomial distribution.]

   (b) To check the reported figures, you conduct a random survey of 1000 people between the ages of eighteen and twenty-four. To test the hypothesis that the true rate of youth unemployment is 18%, adopt the following decision rule:
   *Accept the hypothesis that the true rate of unemployment is 0.18 if the number of unemployed people in the survey is between 160 and 200 inclusive.*

      i. Find the probability of rejecting the hypothesis when it is, in fact, correct.

      ii. Sketch a graph of the normal distribution, illustrating the decision rule, and the probability that you calculated in part (a).

(c) Repeat part (b) for the tougher decision rule:

*Accept the hypothesis that the true rate of unemployment is 0.18 if the number of unemployed people in the survey is between 170 and 190 inclusive.*

(d) Suppose that you wanted to choose a symmetric interval around 180 in the decision rule so that the probability of rejecting the hypothesis when it is in fact correct is 0.05. What decision rule would you set?

11. Consider two groups A and B composed of 100 people who are suffering from a disease. A drug is given to group A. A placebo is given to group B. It is found that in group A 75 people recover whereas in group B only 65 people recover.

    At significance levels (a) 0.01, (b) 0.05 and (c) 0.1, test the hypothesis that the drug helped cure the disease. Compute the p-value in each test and show that in (a) the p-value$> 0.01$, in (b) p-value$> 0.05$ but in (c) p-value$< 0.1$.

    Briefly discuss the results and what they mean in terms of Type 1 and Type 2 errors – which kind of error is best in these circumstances?

    Now suppose that the sample sizes in each group are 300 and 225 in group A and 195 in group B recover. What do you conclude?

12. To investigate possible gender discrimination in a large firm, a sample of 100 men and 64 women with similar job descriptions are selected at random. A summary of their monthly salaries is as follows:

| | Average Salary($\bar{Y}$) | Standard Deviation | n |
|---|---|---|---|
| Men | \$3100 | \$200 | 100 |
| Women | \$2900 | \$320 | 64 |

    What do these data suggest about wage differences in the firm? Do they represent statistically significant evidence that wages of men and women are different? Do they suggest that the firm is guilty of discrimination? (From *Stock and Watson*.)

13. The table below shows the number of male and female births in the UK in 2003 and 2004 (data taken from Hendry and Nielsen, *Econometric Modeling*):

| | Boys | Girls |
|---|---|---|
| 2003 | 356578 | 338971 |
| 2004 | 367586 | 348410 |

    Is there any evidence that the probability of a girl changed between 2003 and 2004? Find the p-value of your test statistic.

14. In an opinion poll of 1000 voters in a constituency, 353 say they will vote for the Orange Party candidate, 405 for the Indigo Party candidate, and the rest will vote for other parties (or abstain). You are interested in the prospects of the Indigo Party in the election. let $p_0$ be the proportion of the electorate who support Orange, and $p_1$ be the proportion of the electorate who support Indigo. Define random variables:

    $X_i = 1$ if voter $i$ supports Orange, $X_i = 0$ otherwise, and

    $Y_i = 1$ if voter $i$ supports Indigo, $Y_i = 0$ otherwise.

(a) Find a 95% confidence interval for the proportion of the electorate who support the Indigo Party.

(b) Show that the random variables $X_i$ and $Y_i$ are not independent.

(c) Find the covariance between the sample means $\bar{X}$ and $\bar{Y}$.

(d) Hence determine the distribution of $\hat{d} = \bar{Y} - \bar{X}$.

(e) Find a 95% confidence interval for $p_1 - p_0$.

(f) Can you be confident that Indigo will beat Orange?

(g) What would have happened if you had treated the estimates of $p_0$ and $p_1$ as independent?

15. Download the Excel datafile Growth.xls from the Stock and Watson Companion Website: http://wps.aw.com/aw_stock_ie_3/178/45691/11696965.cw/index.html (and the corresponding Data Description).
Examine the relationship between the the average number of years of education of a country's adults in 1960 and: (i) its GDP in 1960 and (ii) its average growth rate between 1960 and 1995. In each case, draw a scattergram; find the sample correlation between the two variables (Excel function CORREL); and test the hypothesis that they are positively related. Comment on your results.

16. The Excel spreadsheet TSAdata.xls, on the QE Weblearn page, contains data based on (i.e. not exactly the same as the real data) the TSA test results for the PPE admissions exercise in 2008. The variable definitions are as follows:

| | |
|---|---|
| TSA: | the percentage TSA test school aggregated over the components of the test |
| Gender: | the gender of the applicant (F = Female, M = Male). |
| School Type: | the type of school the applicant was studying at at the time of application (I = Independent, S = State, O = Overseas) |
| Admit: | a dummy variable for whether or not the applicant was eventually admitted (1 = admitted, 0 = not admitted) |

(a) Draw a histogram of the TSA data and comment on whether the data seem to be normally distributed.

(b) Estimate the mean and the standard deviation of the distribution of TSA scores.

(c) Using your estimates of the mean and standard deviation draw the pdf for a fitted normal distribution and compare it to the histogram which you drew previously. Does the normal seem to have captured the density of the data adequately?

(d) Using the methods above draw the fitted normal densities on a single set of axes and comment briefly on the differences you see for

i. Females and males

ii. The three different school types

iii. Those admitted and not admitted

(e) Test the hypotheses that

i. Females perform worse than Males on the TSA test

ii. Applicants from Independent Schools perform better than applicants from State Schools

iii. There is no difference between the TSA test scores of those admitted and those not admitted

Hints on using *Excel*:

1. *Drawing Charts*
   To draw a chart in Excel 2007, first select the data that you want to display. (For example, to draw a scattergram, select the two columns that you want to relate.) Then on the *Insert* tab, choose the type of chart that you want, and it will be displayed immediately. With the chart selected, you can then use the *Layout* tab to change the appearance of the chart - add axis titles, change the axis scale, etc.

   Drawing Charts in Excel 2003 is self explanatory using the ChartWizard (the help facility is quite good should you need it). The ChartWizard is a series of dialog boxes that guide the user through the steps required to create a new chart or modify settings for an existing chart. Using ChartWizard the user specifies the worksheet range, selects a chart type and format, and indicates how the data is to be plotted. The user can add a legend, which depicts what each chart represents, a chart title and a title to each axis, and so on. Note that Chart Wizard is easier to use if you select the columns of data that are relevant for your chart first. For example, to draw a scattergram, select the two columns that you want to relate, then start the Chart Wizard.

2. *To draw a histogram*
   You need to construct a frequency table before you draw the chart:

   (a) create the bins

   (b) work out the proportion of the observations in each bin, and lay these out in the spreadsheet

   (c) draw the Chart

   Here is an example of how to do it, using 10 bins each with a width of 10 percent starting at 0 and going up to 100 (you are free to use finer bins).

   (a) In cells F2:F12 list the breakpoints 0, 10, ...,100 vertically.

   (b) Now click in G2 and enter the formula
   =FREQUENCY(A2:A1433,F2:F12)/COUNT(A2:A1433)
   and press RETURN. The number 0 will appear (because there are no applicants who scored between 0 and 10%) The FREQUENCY command is an array formula. To get the rest of the bin-proportions for the rest of the bins highlight the cells where you want the frequencies to appear (G2:G12), press the function key F2 and then press CTRL+SHIFT+ENTER. The rest of the frequencies will appear in cells G2:G12.

   (c) You can now draw the histogram using ChartWizard.

3. *Computing means and standard deviations*
   These can be done using the functions AVERAGE() and STDEV(). For example:

=AVERAGE(A2:A1433) computes the overall mean of the TSA data. To calculate averages for sub-samples you can use the AVERAGEIF() command. For example AVERAGEIF(B2:B1433,"F",A2:A1433) computes the average scores for Females.

You might also find the Filter feature useful. Filtered data displays only the rows that meet conditions that you specify and hides rows that you do not want displayed. After you filter data, you can do what you want with the subset of filtered data without rearranging or moving it. Essentially using Filter allows you to select a subset of interest and displays that subset as if it were the only data you had. You can then take the conditional mean by using AVERAGE on the filtered data. Similarly you can use STDEV to calculate the standard deviation of the filtered data and just divide by the square root of the number of filtered observations to get the standard error on the conditional mean.