# Basic Statistics Review for Economics Students

John E. Floyd
University of Toronto
May 15, 2013

This document presents a review of basic statistics for use by students who plan to study economics in graduate school or who have long-ago completed their graduate study and need a quick review of these basics. The first section deals with the nature and examination of data, proceeding partly with the aid of the spreadsheet program Gnumeric, which is a free MS-Excel clone. The second section discusses probability and examines the nature of some important probability distributions. Section three deals with hypothesis tests and section four with OLS regression analysis. Although the regression analysis will here be undertaken using Gnumeric, more convenient software for statistical analysis will discused in a subsequent document entitled *Statistical Analysis Using XLispStat, R and Gretl: A Beginning*. An appropriate background for understanding the material covered here can be obtained by reading indicated chapters of the manuscript that I prepared for my introductory statistics class, J. E. Floyd, *Statistical Analysis for Economists: A Beginning*, University of Toronto, 2010. Anyone who has difficulty understanding that material can examine one of the elementary textbooks there recommended. Those who have a good grasp of the material presented here and want a deeper review should work through a recent edition of either James H. Stock and Mark W. Watson, *Introduction to Econometrics*, published by Prentice Hall, or Jeffrey Wooldridge, *Econometrics: A Modern Approach*, published by South Western.

## 1. The Examination of Data[1]

Before undertaking statistical analysis, one should first examine carefully the data that are being analyzed. To illustrate this process, the unemployment rates for twenty-three countries in the year 2004 were obtained from *International Financial Statistics*, which is published by the International Monetay Fund. These data are presented below. They are also in the Excel spreadsheet file `datanal1.xls` and in the spreadsheet file `datanal1.gnumeric`, both produced using the Gnumeric spreadsheet program.

| | | |
|---|---|---|
| Luxembourg | 3.86819 | |
| New Zealand | 4.05 | |
| Switzerland | 4.3 | |
| Norway | 4.36667 | |
| Ireland | 4.5 | |
| Japan | 4.71667 | <— 1st. Quartile |
| United Kingdom | 4.75 | |
| Sweden | 5.51667 | |
| United States | 5.5325 | |
| Australia | 5.53333 | |
| Singapore | 5.8 | |
| Denmark | 5.85 | <— Median |
| Netherlands | 6.49167 | |
| Portugal | 6.65 | |
| Austria | 7.075 | |
| Canada | 7.19167 | |
| Italy | 8.05 | |
| Finland | 8.825 | <— 3rd. Quartile |
| Germany | 9.2 | |
| France | 9.275 | |
| Greece | 10.5 | |
| Spain | 10.8075 | |
| Belgium | 12.7917 | |

These data are cross-sectional—that is, they contain observations for a number of entities, which are countries, for a specific year. Were we to have

[1] An appropriate background for the material covered in this section can be obtained by reading the first chapter of my manuscript *Statistical Analysis for Economists: A Beginning*, noted in the introductory remarks.

obtained unemployment rates for a single country for a number of years, our data would be time series. Alternatively, we could have obtained monthly unemployment rates for all these countries for a 30 year period, in which case be would have a panel data set—360 time-series observations for each of 23 countries.

The first statistic we need to calculate is the mean, which is one of the two standard measures of central tendency. It equals

$$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N} \tag{1}$$

where $\bar{X}$ is the arithmetic mean and $X_i$ is the $i^{th}$ observation in the data above. We can obtain the mean in Gnumeric by summing the observations for the column using the command `sum(B6:B28)` in the cell `B30` and then dividing the resulting number by the number of observations using the command `B30/23` placed in the cell `B32`. Alternatively, we could have simply placed the command `average(B6:B28)` in cell `B32`. Another standard measure of central tendency is the median—the observation which half of the observations are bigger than and half the observations smaller than. To find this median observation, we copy the data to a new column and sort that column of data in ascending order by highlighting it and then using the `sort` function in the `Tools` item in the Gnumeric menu across the top of the screen. It is a sorted version of the data set that is presented above. The median observation is the $12^{th}$ one down from the top, with 11 observations bigger and 11 smaller. If there had been only 22 observations in the data set we would set the median equal to the average of the $11^{th}$ and $12^{th}$ biggest observations and would again have 11 observations smaller than and 11 observations bigger than this average.

Next, we need some measures of the variability of the data around its mean and median values. The standard measures of variability around the mean are the variance and standard deviation. The variance is defined alternatively as

$$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N} \tag{2}$$

or

$$s^2 = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N - 1} \tag{3}$$

depending upon whether we are looking at a population or a sample of that population. A population has a fixed mean, denoted here as $\mu$, so that the deviations of all its elements from that mean are independent of each other. By contrast, a sample has a mean that depends on the particular sample that has been selected. As a result, only $N-1$ deviations of the sample elements from that mean are independent—if we know $N-1$ deviations from the sample mean $\bar{X}$ we can calculate the remaining deviation. That is,

$$\sum_{i=1}^{N}(X_i - \bar{X}) = \sum_{i=1}^{N-1}(X_i - \bar{X}) + (X_N - \bar{X})$$

implies

$$(X_N - \bar{X}) = \sum_{i=1}^{N}(X_i - \bar{X}) - \sum_{i=1}^{N-1}(X_i - \bar{X})\,. \tag{4}$$

There are thus only $N-1$ degrees of freedom for variation in case of deviations from the sample mean while there are $N$ degrees of freedom in the case of deviations from the fixed population mean. Accordingly, if we treat our data as a sample from some population, whose mean we do not know, the variance can be calculated by setting up in Gnumeric a new column of numbers equal to the squares of the deviations of the respective observations from the sample mean and then summing the elements of that column and dividing by 22 (= 23 - 1). To construct the column of numbers, all we have to do is make the calculation for one element of the column and then copy the code from that cell to the remaining cells in the column—Gnumeric automatically adjusts the code numbers to match the new row element as we move down the column. Then, after calculating the variance, we simply take the square root of it to obtain the sample standard-deviation, denoted by $s$,

$$s = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N-1}}\,, \tag{5}$$

with the population standard deviation denoted by $\sigma$,

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}\,. \tag{6}$$

The problem with the variance and standard-deviation as measures of variability is that they give no information about the pattern of variability—that is, about how the elements are distributed among low and high values.

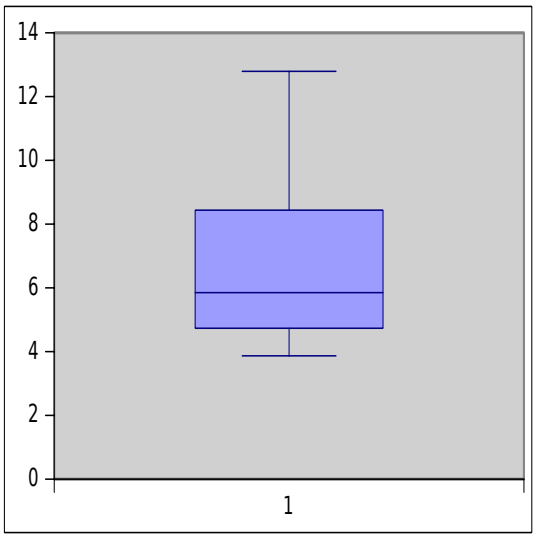
To acquire this type of information, we need to look at the range—that is the distance between the highest and lowest values—and then calculate the first and third quartiles. The first quartile is the obsevation that 25 percent of the observations are smaller than and the third quartile is the observation which 75 percent of the observations are smaller than. In the data above there are 11 observations below the median and 11 observations above it. The best pick for the 1st quartile is the $6^{th}$ observation from the bottom and the best choice of the upper quartile is the $6^{th}$ observation down from the top. As in the case of the median, sometimes the first and third quartiles have to be obtained by averaging two neighboring observations. The distance between the first and third quartiles is the inter-quartile range. We can then examine the pattern of variability by seeing where the mean and median lie relative to the first and third quartiles.

In our data set above, the mean is above the median and the third-quartile value is above the median by more than twice as much as the first-quartile value is below it. About one half of the observed unemployment rates are between 3 percent and 6 percent and about one-quarter are between 6 and 9 percent and the remaining quarter being between 9 and 13 percent. These high values drag the mean above the median and skew the distribution to the right. This can be seen from the simple but rather crude frequency distribution constructed below.

| | | | | |
|---|---|---|---|---|
| x | x | | | |
| x | x | | | |
| x | x | | | |
| x | x | x | x | |
| x | x | x | x | |
| x | x | x | x | |
| x | x | x | x | x |
| — | — | — | — | — |
| 7 | 7 | 4 | 4 | 1 |
| — | — | — | — | — |
| 3-5 | 5-7 | 7-9 | 9-11 | 11-13 |

Five equal intervals between the a minimum value of 3 and a maximum value of 13 were constructed and written in order on a piece of paper. Then for each of the 23 observations, an `x` was placed above the interval in which

the value of the observation falls. The number of observations falling in each interval is then counted to provide a frequency distribution. The left-most two categories, $3-5$ and $5-7$ are tied as indicators of the mode or modal class—that is, the category containing the largest number of occurrences. The same thing could be accomplished by drawing an appropriate histogram—unfortunately, one cannot do that easily in Gnumeric although one can obtain the mean, median, maximum, minimum, inter-quartile range, variance, standard deviation and other statistics by simply highlighting the relevant column and then clicking on the `tools` item on the menu and then choosing `Statistical Analysis` and then `Descriptive Statistics`. You can see from the pattern of `x`'s in the frequency distribution above that the distribution is skewed to the right.




This pattern can also be seen in the box-plot above that was constructed in Gnumeric. The lower and upper borders of the box in the center of the diagram give the first and third quartiles and the horizontal line in the box just above its lower border is the median. Were it shown in the box, the mean would be a horizontal line about one unit, measured on the vertical scale, above the median. The vertical whiskers above and below the box extend to the maximum and minimum values respectively. It is easy to see that there

is a high concentration of values at the lower unemployment rates, with the small number of extreme upper values pulling up the mean.

Another way of measuring the pattern of variation in data is to standardize them—that is, to measure each data-value by the number of standard-deviations it is above or below the mean. The standardized values can be expressed as

$$Z_i = \frac{X_i - \bar{X}}{s} \tag{7}$$

which, when applied to our unemployment rate data, yields the following.

| | |
|---|---|
| Luxembourg | -1.1815 |
| New Zealand | -1.1074 |
| Switzerland | -1.0055 |
| Norway | -0.9783 |
| Ireland | -0.9240 |
| Japan | -0.8357 |
| U.K. | -0.8221 |
| Sweden | -0.5096 |
| U.S. | -0.5031 |
| Australia | -0.5028 |
| Singapore | -0.3941 |
| Denmark | -0.3737 |
| Netherlands | -0.1122 |
| Portugal | -0.0477 |
| Austria | 0.1255 |
| Canada | 0.1731 |
| Italy | 0.5229 |
| Finland | 0.8388 |
| Germany | 0.9916 |
| France | 1.0222 |
| Greece | 1.5214 |
| Spain | 1.6468 |
| Belgium | 2.4554 |

Luxembourg's unemployment rate, the lowest in comparison with that in the 22 other countries, is about 1.2 standard deviations below the 22-country

mean while the that of Belgium, the highest, is nearly two-and-one half standard deviations above it. Only nine countries are above the mean, with fourteen countries below it.

It turns out that the mean of such standardized values is zero and their standard deviation and variance both equal unity. This follows from the facts that

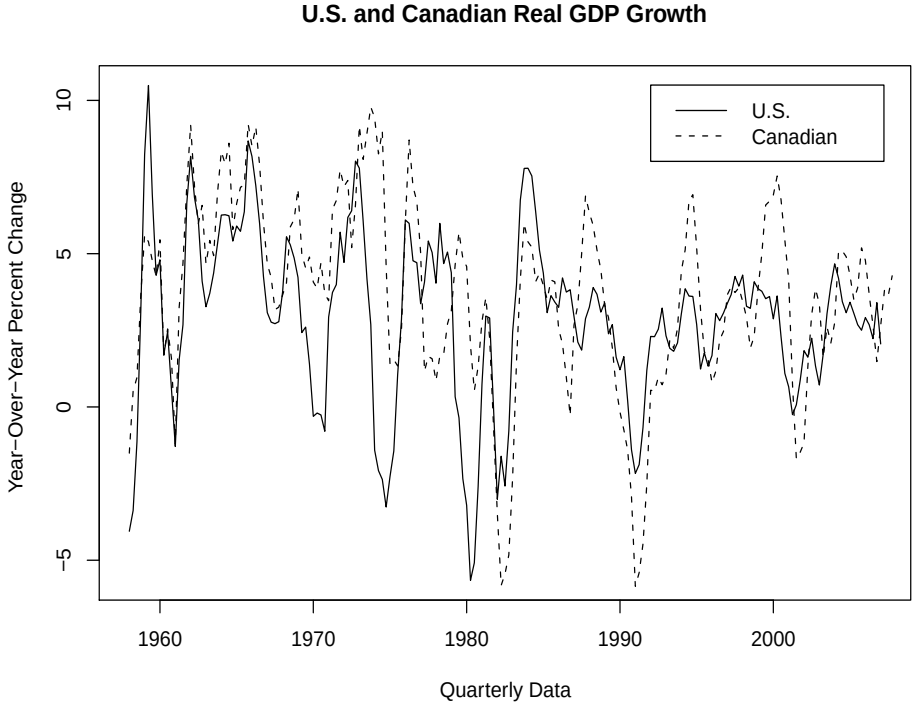
$$\bar{Z} = \frac{\sum_{i=1}^{N} Z_i}{N} = \frac{\sum_{i=1}^{N} (X_i - \bar{X})}{s\ N} = \frac{\sum_{i=1}^{N} X_i - N\bar{X}}{s\ N} = \frac{1}{s}\left(\frac{\sum_{i=1}^{N} X_i}{N} - \bar{X}\right) = 0$$

and

$$\frac{\sum_{i=1}^{N} (Z_i - \bar{Z})^2}{N-1} = \frac{\sum_{i=1}^{N} Z_i^2}{N-1} = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{s^2\,(N-1)} = \frac{s^2}{s^2} = 1\,.$$

Some additional measurement issues arise in the case of time-series data. As an example, we look at the year-over-year GDP growth rates for the United States and Canada. Gross domestic product series and consumer price indexes, with year 2000 base, were obtained for the two countries from the IMF *International Financial Statistics* and, using the Gnumeric spreadsheet program, the GDP series were divided by the corresponding CPI series and multiplied by 100 to obtain the two countries' real GDP's in year 2000 domestic dollars. Then the percentage year-over-year GDP growth rates were calculated on a quarterly basis by taking the percentage difference between each quarter's real GDP level and that of the same quarter of the previous year. These data and the corresponding calculations can be found in the MS-Excel spreadsheet file `datanal2.xls` which was created with Gnumeric and in an equivalent Gnumeric spreadsheet file `datanal2.gnumeric` which also contains the box-plot to be shown later. It turns out that Gnumeric will not read an `.xls` file in which a box-plot was created—such files have to be saved in Gnumeric format.

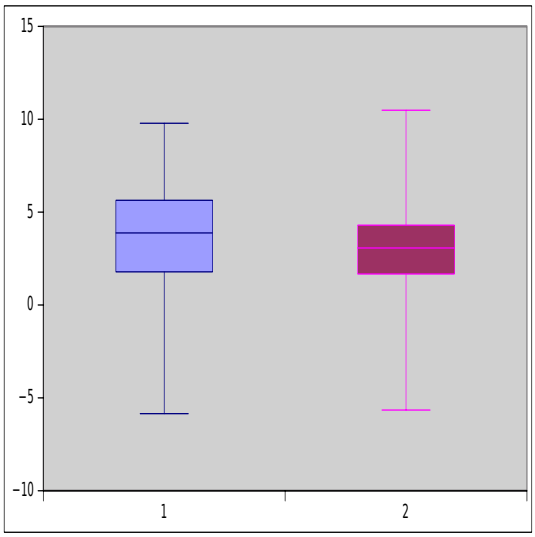
The first measurement device for time-series data is a time-series plot, shown below. The two series show little trend and seem to be somewhat related to each other, with the Canadian series showing greater variability than the U.S. series. You can also see from the figure that each of the series is somewhat serially correlated—that is, correlated with itself through time—in that high (low) values in any period tend more often than not to be followed by high (low) values in the next period.




The means and standard-deviations for each of the growth rates are calculated in the above-noted spreadsheet files—the means are 3.66 for Canada and 2.89 for the U.S. and the corresponding standard-deviations are 3.10 and 2.75. Indeed, a box-plot of the two series, shown below with the Canadian series on the left, is very instructive in helping us visualize the relative variability of the two countries' real GDP growth rates.

As you can see, the maximum and minimum values of the two series are not too different, but the inter-quartile range—that is the difference between the first-quartile and third-quartile—is much bigger for the Canadian series on the left than the U.S. series on the right, and a somewhat higher median for Canadian real GDP growth is also apparent. But the coefficients of variation, which are the standard deviations taken as percentages of the respective means, given by the formula

$$CV = \frac{100\, s}{\bar{X}}\,, \tag{8}$$




are $100 \times 3.10/3.66 = 86$ for Canada and $100 \times 2.75/2.89 = 95$ for the United States. The reason for this apparent greater variability of U.S. real GDP growth is the fact that the mean is $100 \times (3.66 - 2.89)/2.89 = 27$ percent higher in the Canadian than U.S. case while the standard deviation is only $100 \times (3.10 - 2.75)/2.75 = 13$ percent higher. As calculated in the spreadsheet file `datanal2.xls`, the median is $100 \times (3.88 - 3.07)/3.07 = 26$ percent higher in the Canadian than U.S. case.

Finally, it is worth examining the correlation between GDP growth in Canada and the United States. The covariance between two variables $X$ and $Y$ is

$$s_{xy} = s_{yx} = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{N - 1} \tag{9}$$

and the coefficient of correlation is

$$r_{xy} = r_{yx} = \frac{s_{xy}}{s_x \, s_y} \tag{10}$$

where $s_x$ and $s_y$ are the standard deviations of the two variables. As calculated in the above spreadsheet file, the coefficient of correlation of the Canadian and U.S. real GDP growth is $4.7559/(3.1019 \times 2.7499 = 0.5575$.

The correlation coefficient can also be calculated more easily in Gnumeric by clicking on `Tools`, then `Statistical Analysis` and then `Correlation`.[2]

So are we to believe that Canadian and United States year-over-year real GDP growth are correlated to a degree about half-way between zero correlation ($r_{xy} = 0$) and perfect correlation ($r_{xy} = 1$)? How likely is it that the observed correlation coefficient of 0.5575 could have arisen entirely on the basis of pure random chance? More specifically, what proportion of the observed 0.5575 coefficient could be the result of random noise in the data and what proportion is the result of a real underlying relationship between real GDP growth in the two countries? These questions can only be answered through a proper understanding of probability and hypotheses testing, the topics to which we now turn.

[2]Suppose, for example, that you want to calculate the correlation between the row elements 16 to 26 in the M column with the row elements 16 to 26 in the N column. The appropriate entry in the correlation window would be `M16:M26,N16:N26`.

**Exercises**

Before proceeding, it would be useful for you to examine carefully the Gnumeric (Excel) worksheets in which the material referred to above was produced. They are `datanal1.xls` or `datanal1.gnumeric`, and `datanal2.xls` or `datanal2.gnumeric`, with the `.gnumeric` versions including the box-plots.

Then to proceed, access the worksheet `datanalq.xls` which contain United States monthly M1 and M2 series, both seasonally adjusted and unadjusted, for the period January 1959 through May 2010. Then perform the following calculations:

1. Calculate the year-over-year percentage growth rate, monthly, for versons of the two series that are not seasonally adjusted, and the month-over-month percentage growth rate for the seasonally-adjusted series. Why should one use the non-seasonally-adjusted series for the year-over-year calculations and the seasonally-adjusted series for the month-over-month calculations?

2. Make indexed versions of the two seasonally adjusted series, starting with January 1960, on the base of 1960 = 100. Then plot these two series on the same plot. [To do the plot, highlight the two series and then click on the graph item on the tool bar and pick the item `line`. In the window that then appears, click on `insert`, then put the cursor where you want the upper-left corner of the graph to be placed and, holding down the left mouse-button, drag the cursor to where you want the bottom-right corner of the graph to be placed. Usually, it takes a lot of down-to-the-right dragging to get the cursor in an appropriate place to locate the bottom-right corner.] Which series has grown the most?

3. Calculate the means of the year-over-year and month-over-month growth rates of the two series and verify by observation that they are consistent with the results of the above plot. Then make plots of the growth rates, beginning in 1960, with both year-over-year growth rates on one plot and both month-over-month growth rates on a second plot. Which of the two series appears to have the most variable growth rate?

4. Calculate the squared deviations of the month-over-month growth rates of M1 and M2 from their means, then sum them and calculate their variances, treating the growth rates as a sample of population growth rates having measurement errors. Then obtain the standard-deviations. Are the standard-deviations consistent with your conclusion above about the relative variability of the two growth rates.

5. Calculate the product of the deviations of the month-over-month growth rates of M1 and M2 from their means and then calculate their covariance and the correlation rate between them. Why does it make sense to use the month-over-month growth rates rather than the year-over-year growth rates for the variance and covariance calculations?

6. Finally, calculate the standard deviations of the means of the two month-over-month growth rates.

After doing the above exercises, access the spreadsheet `datanala.xls` to check your answers.

## 2. Probability Theory and Probability Distributions[3]

We now carefully review the concept of probability. A classic example is the toss of a coin. If the coin is a fair one, there is a 50 percent chance of obtaining a head and a 50 percent chance of obtaining a tail on any given toss, and there is a 100 percent chance that the coin, even if not a fair one, will come up either heads or tails. Accordingly, we say that the probability of receiving a head is .5 and the probability of receiving a tail is .5 and the probability of receiving either a head or tail is $.5+.5=1$. Half of the unitary probability mass falls on head and half on tail.

Consider now an individual with expertise in stock-market analysis who assigns the following probabilities to the possibilities that the prices of each of two particular stocks, A and B, will increase by more than $\frac{1}{2}$ a percentage point (50 basis points), fall by more than $\frac{1}{2}$ a percentage point, or remain within $\frac{1}{2}$ of a percentage point of its current market value over the next year. The assigned probabilities are presented in the following two tables.

| Stock A | Probability |
|---|---|
| Increase more than 1/2 percent | .40 |
| Remain within 1/2 percent | .20 |
| Fall more than 1/2 percent | .40 |
| Do one of the above | 1.00 |

| Stock B | Probability |
|---|---|
| Increase more than 1/2 percent | .30 |
| Remain within 1/2 percent | .40 |
| Fall more than 1/2 percent | .30 |
| Do one of the above | 1.00 |

You should note that the probability of a major change in the price of stock A is greater than the probability of a major change in the price of stock B.

[3]An appropriate background for the material covered in this section can be obtained by reading chapters 2 and 3 of my manuscript, *Statistical Analysis for Economists: A Beginning*.

The individual could assign these probabilities subjectively on the basis of a general understanding of the functioning of the two firms involved and her expectations of future demand for the products they produce, or more objectively on the basis of historical data regarding the fractions of the time these stock prices rose, fell, or remained roughly the same over the previous few years.

If the probability assignment is based on an analysis of past data, she might be able to assign joint probabilities to the behaviour of the two stock prices in relation to each other as follows.

| Stock B | | Stock A | | | |
|---|---|---|---|---|---|
| | | Increase | No Change | Decrease | |
| | | $A_1$ | $A_2$ | $A_3$ | Sum |
| Increase | $B_1$ | .20 | .05 | .05 | .30 |
| No Change | $B_2$ | .15 | .10 | .15 | .40 |
| Decrease | $B_3$ | .05 | .05 | .20 | .30 |
| | Sum | .40 | .20 | .40 | 1.00 |

The nine probabilities in the center are called `joint probabilities` and the three probabilities in the right-most column and the three in the bottom row are called `marginal probabilities`, which are the probabilities that the price of the given stock will increase, remain roughly the same, or decline, regardless of what happens to the price of the other stock. The joint probabilities sum to unity, as does each set of marginal probabilities. Mathematically, letting $i$ and $j$ take the values 1, 2, or 3, the marginal probabilities in the table can be denoted as $P(A_i)$ and $P(B_j)$ and the joint probabilities as $P(A_i \cap B_j)$.[4] Notice that the marginal probabilities are the sums of the joint probabilities in the associated row or column.

If the price of stock B rises, what is the probability that the price of stock A will also rise? To calculate this we take the joint probability that both prices will rise as a fraction of the probability that the price of stock B will rise. This is a `conditional probability`— that is, the probability of a rise

[4]For example, there is a 40 percent chance that the price of stock A will increase and a 30 percent chance that the price of stock B will increase, and a 20 percent chance that they will both increase.

in the price of stock A conditional upon a rise in the price of stock B. In the above table, it equals

$$P(A_1|B_1) = P(A_1 \cap B_1)/P(B_1) = 0.20/0.30 = 0.66666666$$

which says that the price of stock A will rise two-thirds of the time when the price of stock B rises. And you can calculate from the above table that the price of stock A will remain roughly unchanged one-sixth of the time and fall one-sixth of the time, given that the price of stock B increases. By rearrangement of the above equation we can see that the joint probability of an increase in the price of stock A and the price of stock B is equal to

$$P(A_1 \cap B_1) = P(A_1|B_1)\, P(B_1)\,. \tag{1}$$

That is, the probability that the prices of both stocks will rise equals the probability that the price of stock A rises conditional upon the price of stock B rising, times the probability that the price of stock B will rise.

Consider now the probability that either the price of stock A or the price of stock B will rise. This will equal the probability that the price of stock A will rise plus the probability that the price of stock B will rise minus the probability that both stock prices will increase—that is,

$$P(A_1 \cup B_1) = P(A_1) + P(B_1) - P(A_1 \cap B_1) \tag{2}$$

which implies that $P(A_1 \cup B_1) = 0.30 + 0.40 - 0.20 = 0.50$. The joint probability has to be deducted to avoid double counting—the sum of the two probabilities $P(A_1) + P(B_1)$ gives the probability that either or both stock prices will increase. The probability that either one of the stock prices will will rise gives the probability of the `union of the two events`, denoted by the symbol $\cup$, while the symbol $\cap$ denotes the `intersection of the two events`.

Suppose, alternatively, that the joint probabilities are those given in the table below.

| Stock B | | Stock A: Increase $A_1$ | Stock A: No Change $A_2$ | Stock A: Decrease $A_3$ | Sum |
|---|---|---|---|---|---|
| Increase | $B_1$ | .12 | .06 | .12 | .30 |
| No Change | $B_2$ | .16 | .08 | .16 | .40 |
| Decrease | $B_3$ | .12 | .06 | .12 | .30 |
| | Sum | .40 | .20 | .40 | 1.00 |

Under these conditions, the probability that the price of stock A will rise given that the price of stock B rises is simply equal to the probability that the price of stock A will rise independently of what happens to the price of stock B—that is,

$$P(A_1|B_1) = P(A_1 \cap B_1)/P(B_1) = 0.12/0.30 = 0.40 = P(A_1)\,,$$

and an equivalent result holds for all the other conditional probabilities. You can see from the table that the first and third rows are the same and the elements of the second row are one-quarter larger than the corresponding elements of the other two rows. And the marginal probabilities along the bottom row are $2\frac{1}{2}$ times the joint probabilities in the second row and $3\frac{1}{3}$ times the probabilities in the first and third rows. Similarly, the joint probabilities in the first and third columns are, respectively, the same and twice as large as the joint probabilities in the second column. And the probabilities in the middle column are $\frac{1}{5}$ the size of the marginal probabilities in the rightmost column. Also, each joint probability is equal to the product of the corresponding marginal probabilities. When this condition

$$P(A_i \cap B_j) = P(A_i)\,P(B_j) \tag{3}$$

holds, the change in the price of stock A and the change in the price of stock B are `statistically independent`—that is the change in the price of one stock is in unrelated to the change in the price of the other stock. The fact that the joint probabilities are non-zero is strictly the result of the random variability of the two stocks.

Notice now that, when the two sets of events are not statistically independent, the joint probability can be obtained in two ways.

$$P(A_i \cap B_j) = P(A_i|B_j)\,P(B_j)$$

and

$$P(A_i \cap B_j) = P(B_j|A_i)\,P(A_i)\,.$$

It therefore follows that

$$P(A_i|B_j)\,P(B_j) = P(B_j|A_i)\,P(A_i)\,. \tag{4}$$

which implies that

$$P(A_i|B_j) = P(B_j|A_i)\,\frac{P(A_i)}{P(B_j)} = \frac{P(B_j|A_i)\,P(A_i)}{\sum_i [P(B_j|A_i)\,P(A_i)]}\,. \tag{5}$$

The above equation represents `Bayes Theorem`. Given an initial probability $P(A_i)$, called the `prior probability`, together with subsequent evidence regarding the probability of $B_j$ given $A_i$ from the `conditional probability` $P(B_j|A_i)$, one can upgrade that prior probability to obtain the `posterior probability` $P(A_i|B_j)$.

Consider the following (artificially constructed) example. You know that the probability of a randomly selected person in your community being a carrier of the AIDS virus is $P(A_1) = .001$ and the probability that the person is not a carrier of the virus is therefore $P(A_0) = .999$. Your prior probability of being a carrier is therefore .001 and the odds of you being a carrier are thus

$$\frac{P(A_1)}{P(A_0)} = \frac{.001}{.999} = \frac{1}{999}$$

or one to nine hundred and ninety nine. You then take an AIDS test and test positive! And you learn that empirical studies have established that carriers of the virus test positive 90 percent of the time and non-carriers test positive 1 percent of the time—that is

$$P(T_1|A_1) = .90 \qquad \text{and} \qquad P(T_1|A_0) = .01\,.$$

Your joint probability of both being a carrier and testing positive is thus

$$P(T_1 \cap A_1) = P(T_1|A_1)\,P(A_1) = .90 \times .001 = 0.0009$$

and the joint probability of testing positive and not being a carrier is

$$P(T_1 \cap A_0) = P(T_1|A_0)\,P(A_0) = .01 \times .999 = 0.0099\,.$$

Since a person testing positive must either be a carrier or not be a carrier, the marginal probability of testing positive is equal to the sum of the two joint probabilities

$$P(T_1) = P(T_1 \cap A_1) + P(T_1 \cap A_0) = .0009 + .0099 = 0.0108$$

and can be entered, along with the prior probabilities and two joint probabilities in the table below.

| An HIV Carrier? | Test Result Positive ($T_1$) | Test Result Negative ($T_0$) | Prior Probability Distribution |
|---|---|---|---|
| No ($A_0$) | 0.0099 | | 0.999 |
| Yes ($A_1$) | 0.0009 | | 0.001 |
| Total | 0.0108 | | 1.000 |

The conditional probability of you being a carrier given that you tested positive—that is your `posterior probability` is thus

$$P(A_1|T_1) = \frac{P(T_1 \cap A_0)}{P(A_1 \cap T_1) + P(A_0 \cap T_0)} = \frac{P(T_1 \cap A_0)}{P(T_1} = \frac{0.009}{0.0108} = 0.0833\,.$$

The results of the test and the evidence on its reliability has caused you to revise the probability that you are a carrier upward from a `prior probability` of 0.001 to the `posterior probability` of 0.0833 . Finally, the other two joint probabilities,

$$P(T_0 \cap A_0) = P(A_0) - P(T_1 \cap A_0)$$

and

$$P(T_0 \cap A_1) = P(A_1) - P(T_1 \cap A_1)$$

and the marginal probability $P(T_0)$ can be calculated from the numbers already in the table above to produce the following filled-out version.

| An HIV Carrier? | | Test Result: Positive ($T_1$) | Test Result: Negative ($T_0$) | Prior Probability Distribution |
|---|---|---|---|---|
| No | ($A_0$) | 0.0099 | .9891 | 0.999 |
| Yes | ($A_1$) | 0.0009 | .0001 | 0.001 |
| | Total | 0.0108 | .9892 | 1.000 |

The most common form of probability analysis in economics deals with probability distributions of random variables that can take a wide range of values, not just two as in the cases above. Suppose that we have a random variable $X$ that can take integer values in the range zero through ten. The associated probabilities of occurrence of these values form a discrete probability distribution represented by the probability function $P(X_i)$ mapped out in the bottom panel of Figure 1 below. The vertical bars measure the probabilities that $X$ will take the eleven specific values. The top panel graphs the cumulative probabilities—that is, the probability $P(\sum_0^i X_i)$ that $X$ will be equal to or less than the each of the eleven integer values. The vertical bars in that panel thus measure the cumulative probability which runs from zero to unity, and are the vertical sums of the current and all previous vertical bars in the bottom panel.

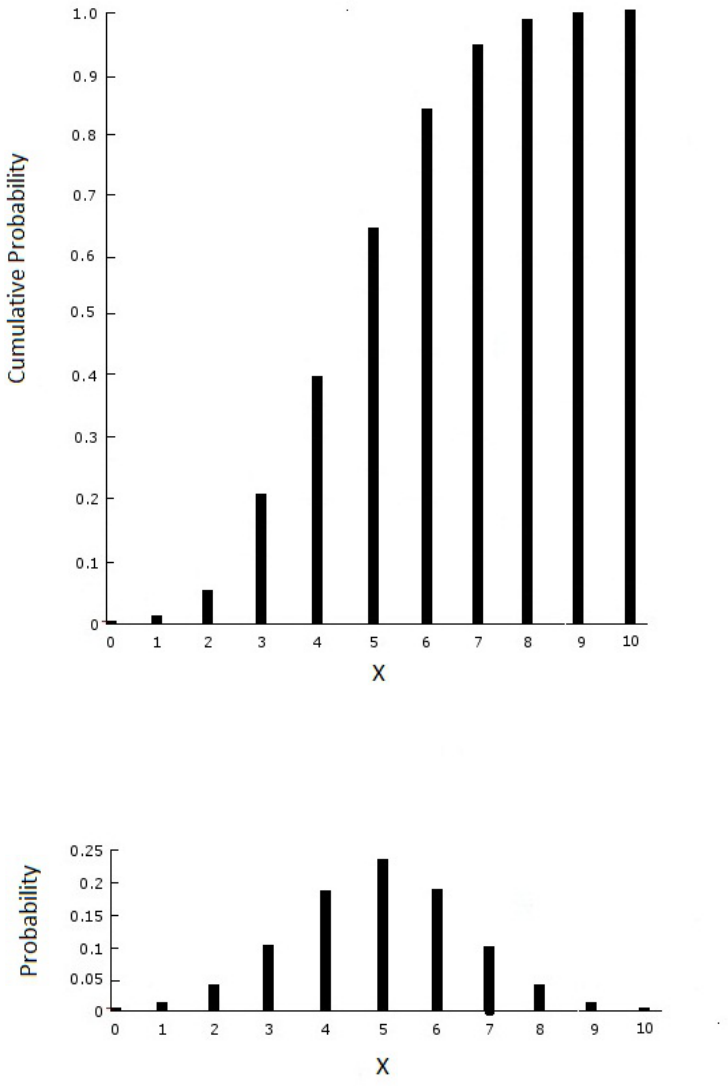



Figure 1: Probability function and cumulative probability function for a discrete variable. The vertical lines in the top panel are the sums of the current and previous probabilities in the bottom panel.

More common in economic analysis are probability distributions of the levels of continuous variables like GDP and the CPI which can take a value equal to any of a range of real numbers. The probability density function for a continuous variable, in this case the year-over-year percentage change in the market value of shares of a particular company, ranging from $-10$ to $+20$ is presented in the upper panel of Figure 2 below, with the lower panel graphing the cumulative density function. We use the term density because the distance from the horizontal axis to the curve (or density function) in the upper panel measures the probability that the variable will take a value within a range of the value on the quantity axis as the size of that range approaches zero. And the probability that the percent change in value will be between 5 and 10 percent is given by the area **A** in the top panel. Since the bottom panel gives the cumulative probability—that is, the probability that the percentage change will be equal to or below a given level—the area **A** in the top panel is represented by the distance **A** in the bottom panel. In mathematical terms, the distance of the curve from the horizontal axis in the top panel equals

$$P(v) = f(v)$$

where $P(v)$ is the probability that the percentage change will be in the immediate neighborhood of $v$ and $f(v)$ is the probability density function. The distance of the curve from the horizontal axis in the bottom panel—that is, the cumulative probability—is

$$P(v \leq v_o) \;=\; F(v_o) \;=\; \int_0^{v_o} f(v)\, dv \tag{6}$$

and the probability that the market value will be increase between 5% and 10%, given by the area **A** in the top panel and the distance **A** in the bottom panel, is

$$P(a \leq v \leq b) \;=\; F(b) - F(a) \;=\; \int_a^b f(v)\, dv\,, \tag{7}$$

where $a = 5$ and $b = 10$.

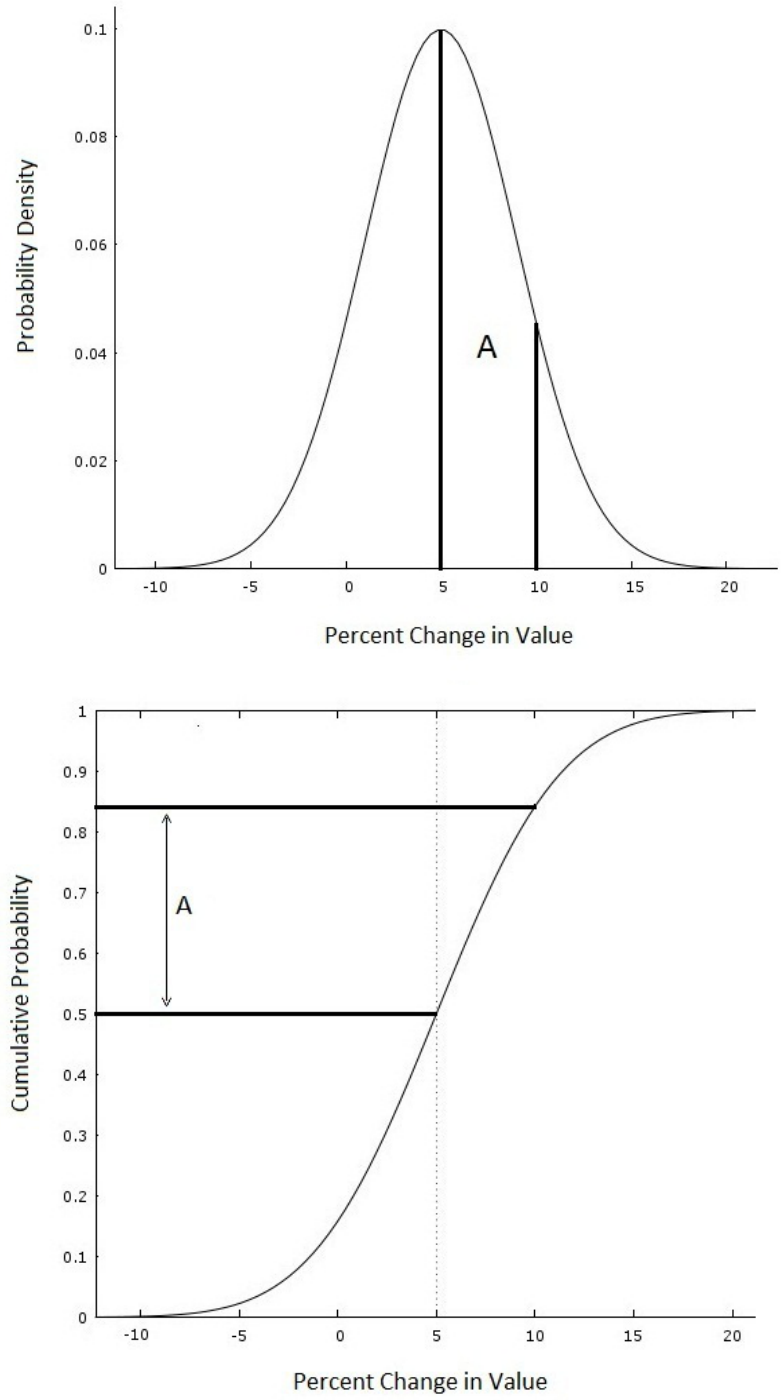



Figure 2: Probability density and cumulative probability functions for the year-over-year change in the market value of shares of a particular company, a continuous variable. The area **A** between the two vertical lines in the top panel equals the distance between the two horizontal lines in the bottom panel.

We can now look at a group of specifically defined probability distributions. Before doing so, however, it is important to understand the possible functional relationships between random variables of the sort just considered. Consider first a linear function of a random variable $X$.

$$Y = a + b\,X \tag{8}$$

A number of relationships hold. First, the expected value or mean of a linear function of a random variable is the sum of the means of its components—that is,

$$E\{Y\} = E\{a\} + E\{b\,X\} = a + b\,E\{X\}\,. \tag{9}$$

The the mean of a constant is that constant itself $(E\{a\} = a)$ and the mean or expected value of a constant times a random variable is equal to that constant times the mean of the random variable $(E\{b\,X\} = b\,E\{X\})$. The variance of a linear function of a random variable, representing either a population or a sample, is

$$\sigma^2\{Y\} = \sigma^2\{a + b\,X\} = \sigma^2\{a\} + \sigma^2\{b\,X\} = 0 + \sigma^2\{b\,X\} = b^2\sigma^2\{X\} \tag{10}$$

where the variance of a constant is obviously zero and the variance of a constant times a random variable is equal to the product of the constant squared and the variance of the random variable. This relationship

$$\sigma^2\{b\,X\} = b^2\,\sigma^2\{X\} \tag{11}$$

also implies that the standard deviation of a constant times a random variable equals

$$\sigma\{b\,X\} = |b|\,\sigma\{X\}\,. \tag{12}$$

Finally, consider the mean and variance of sums and differences of random variables.

$$\begin{aligned} E\{X + Y\} &= E\{X\} + E\{Y\} \\ E\{X - Y\} &= E\{X\} - E\{Y\} \end{aligned} \tag{13}$$

$$\begin{aligned} \sigma^2\{X + Y\} &= \sigma^2\{X\} + \sigma^2\{Y\} + 2\,\sigma\{X, Y\} \\ \sigma^2\{X - Y\} &= \sigma^2\{X\} + \sigma^2\{Y\} - 2\,\sigma\{X, Y\} \end{aligned} \tag{14}$$

The two relationships immediately above can be derived from the definitions of variance and covariance.

The first of the specific probability distributions that need to be discussed is is the binomial distribution plotted in Figure 3 below. The binomial distribution gives the probabilities of each of the $n + 1$ possible results in a sequence of $n$ random trials. An example of such a distribution would be the probabilities of observing zero through ten unfavorable results in next 10 legal situations a company is involved in. The probability $p$ that the result of a particular legal event will be unfavourable is set at 0.2 in the top panel and 0.8 in the bottom panel. A situation where $p = 0.5$ would produce a distribution quite similar to that in the bottom panel of Figure 1 above.

Mathematically, the binomial probability function, which gives the probabilities that $X$ will take values $(0, 1, 2, ....n)$, is

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x} \tag{15}$$

where $P(x) = P(X = x)$, $x = 0, 1, 2, ....n$, and $0 \leq p \leq 1$, and the `binomial coefficient` is

$$\binom{n}{x} = \frac{n!}{x!(n-1)!} \tag{16}$$

where $a! = (a)(a-1)(a-2)(a-3).....(1)$ and $0! = 1$. The probabilities can be calculated by substituting the value of $p$ and the alternative values of $x$ into the equation above. Or you can use one of the free statistical programs that will be discussed later—Gretl, R, or XLispStat—to make the calculations. The mean and variance of the binomial probability function are, respectively

$$E\{X\} = n\,p \tag{17}$$

and

$$\sigma^2\{X\} = n\,p\,(1-p)\,. \tag{18}$$

If we have two independent binomial random variables $V$ and $W$ with common probability parameter $p$ and based on $n_v$ and $n_w$ random trials, the sum $V + W$ is a binomial random variable with parameters $p$ and $n = n_v + n_w$.

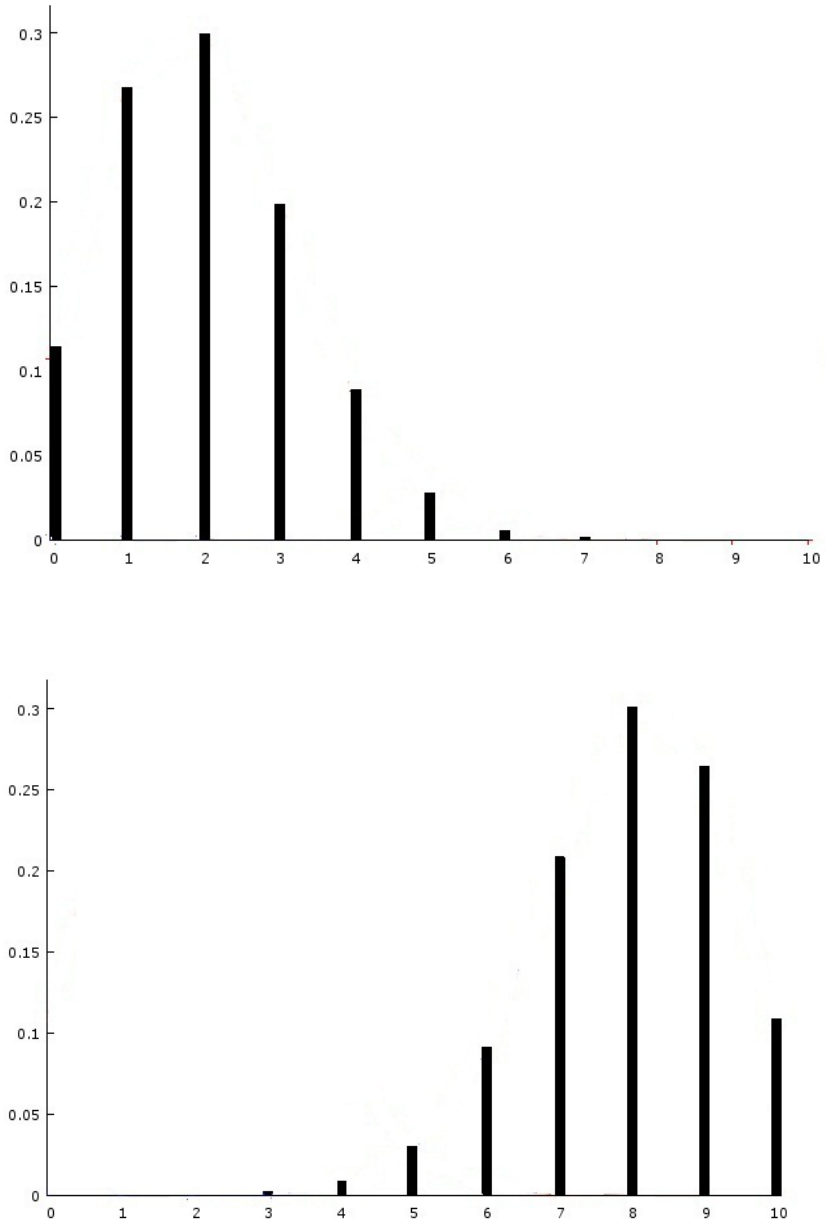



Figure 3: Binomial probability distributions with $p = 0.2$ (top panel) and $p = 0.8$ (bottom panel).

Next, consider the `poisson probability distribution`, examples of which are plotted in Figure 4 below. A poisson random variable is a discrete variable that can take any integer value from zero to infinity. The value gives the number of occurrences of the circumstance of interest during a particular period of time or within a particular spatial area. For example, a firm might be interested in the number of customer complaints occurring during a particular month. The `poisson probability function` is

$$P\{x\} = \frac{\lambda^x \, e^{-\lambda}}{x!} \tag{19}$$

where $P\{x\} = P\{X = x\}$ with $x = 0, 1, 2, 3, 4, ..., \infty$, and $0 \leq \lambda \leq \infty$ is the only parameter. The mean and variance of a poisson probability distribution are, respectively,

$$E\{X\} = \lambda \qquad \text{and} \qquad \sigma^2\{X\} = \lambda \, .$$

In the Figure below, $\lambda = 0.5$ in the top panel and $5.0$ in the bottom one—you can see that the poisson probability distribution becomes more symmetric as $\lambda$ increases.

Another distribution is the `exponential probability distribution` which is closely related to the poisson probability distribution. While the poisson probability distribution applies to the number of occurrences over an interval of time, the exponential distribution applies to the amount of time between occurrences. It is a continuous distribution because time is measured along a continuum. An exponential random variable $X$ is the time between occurrences of a random event. The probability density function is

$$f(x) = \lambda e^{-\lambda x}, \qquad (x > 0). \tag{20}$$

where $\lambda$ is, as in the case of the poisson distribution, the average number of occurrences over the period in which time-between-occurrences is being analysed. It turns out that the cumulative probability that $X \geq x$ is

$$P(X \geq x) = e^{-\lambda x}. \tag{21}$$

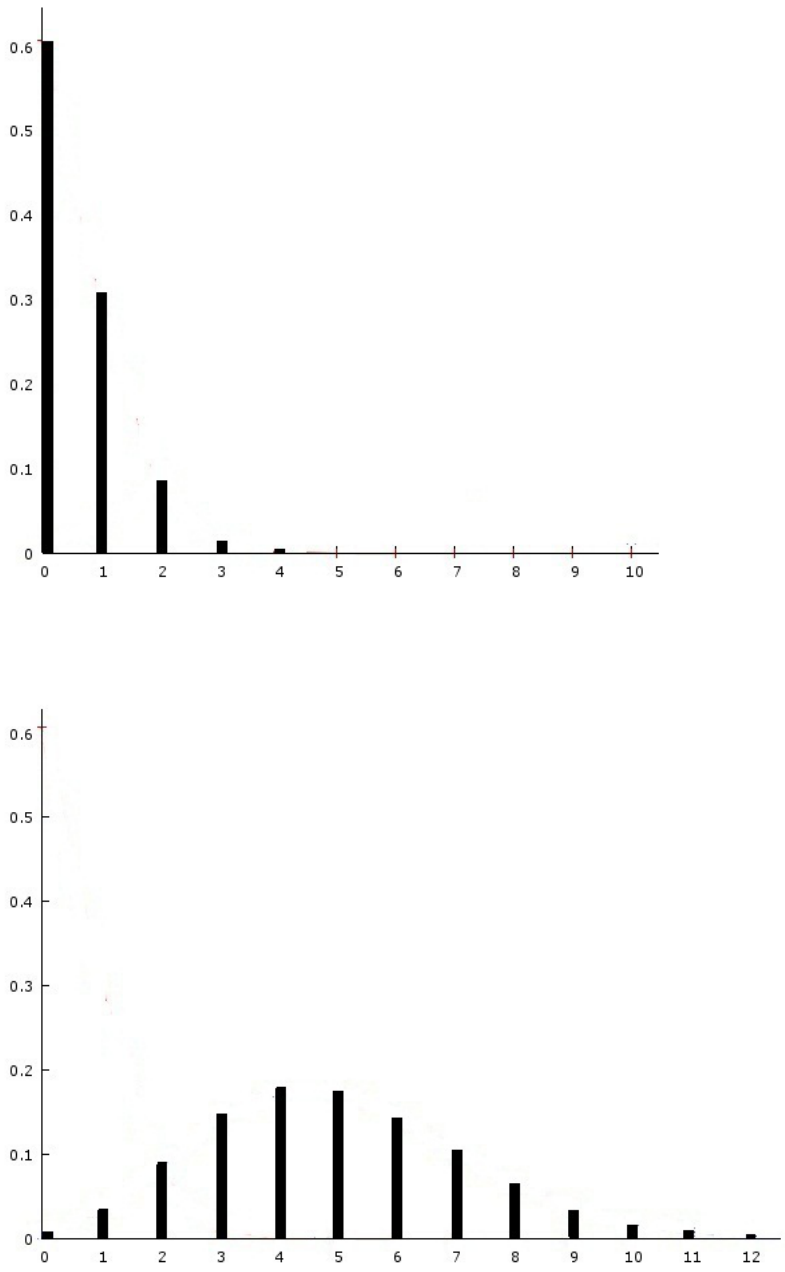



Figure 4: Poisson probability distributions with $\lambda = 0.5$ (top panel) and $\lambda = 5.0$ (bottom panel).

The mean and variance of an exponential distribution are

$$E\{X\} = \frac{1}{\lambda}$$

and

$$\sigma^2\{X\} = \frac{1}{\lambda^2}.$$

The shape of the exponential distribution is governed by the single parameter $\lambda$. As indicated in the plots of some exponential distributions in Figure 5, the exponential probability density function declines as $x$ increases from zero, with the decline being sharper the greater the value of $\lambda$. The probability density function intersects the y-axis at $\lambda$.

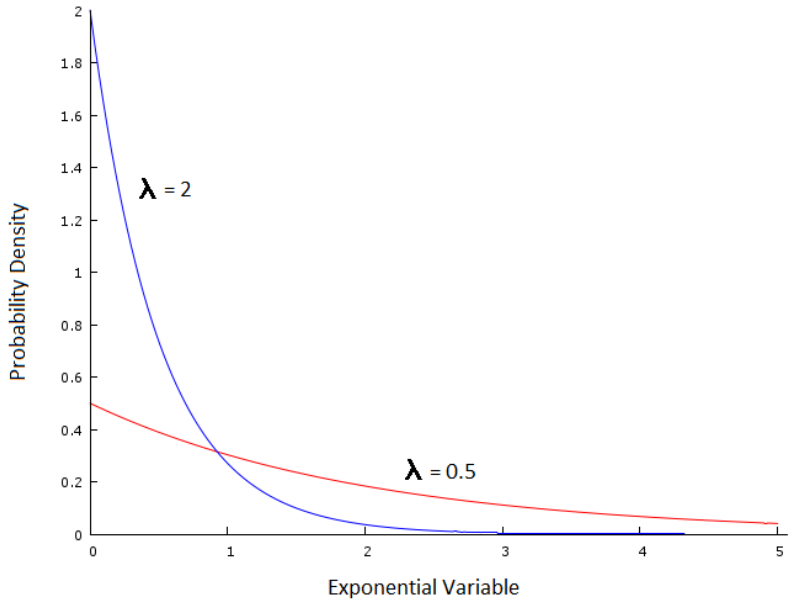



Figure 5: Two exponential probability density functions with $\lambda = 0.5$ and $\lambda = 2$ respectively.

Notice that the probability density at $X = 0$ is equal to $\lambda$ and in the case where $\lambda = 2$ the probability density exceeds unity. This does not mean that the probability that $X = 0$ is greater than unity—that would be contrary to the definition of probability. The probability of $X$ as it approaches zero is

the distance $2$, multiplied by the infinitesimal deviation of $X$ from zero—this product will clearly be less than unity.

A `uniform probability distribution` occurs when the probabilities of all occurrences in a sample space are the same. A discrete uniform random variable has a `discrete uniform probability distribution` of the sort shown in the top panel of Figure 6 below. The discrete uniform probability function is

$$P(x) \;=\; \frac{1}{s} \tag{22}$$

where $P(x) = P(X = x)$, with $x = a, a+1, a+2, \ldots, a+(s-1)$, and the parameters $a$ and $s$ are integers with $s > 0$. Parameter $a$ denotes the smallest outcome and parameter $s$ denotes the number of distinct outcomes.

The mean and variance of a discrete uniform probability distribution are, respectively,

$$E\{X\} = a + \frac{s-1}{2}$$

and

$$\sigma^2 = \frac{s^2-1}{12}.$$

The `continuous uniform or rectangular` probability distribution, an example of which is plotted in the bottom panel of the Figure below, is the continuous analog to the discrete uniform probability distribution. A continuous uniform random variable has uniform probability density over an interval. The `continuous uniform probability density function` is

$$f(x) \;=\; \frac{1}{b-a} \tag{23}$$

where the interval is $a \leq x \leq b$. Its mean and variance are

$$E\{X\} = \frac{b+a}{2}$$

and

$$\sigma^2\{X\} = \frac{(b-a)^2}{12}$$

and the `cumulative continuous uniform probability function` is

$$F(x) \;=\; P(X \leq x) \;=\; \frac{x-a}{b-a}. \tag{24}$$

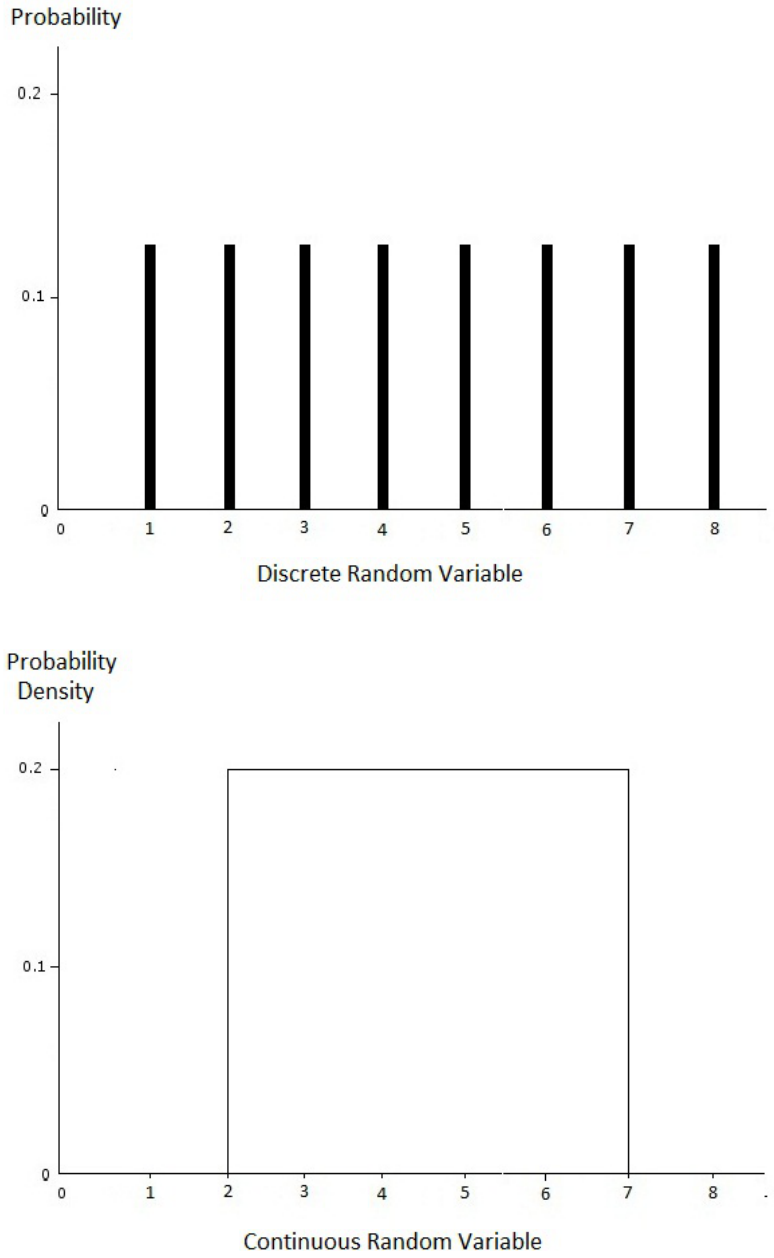



Figure 6: Uniform probability distribution for a discrete random variable (top panel) and continuous random variable (bottom panel).

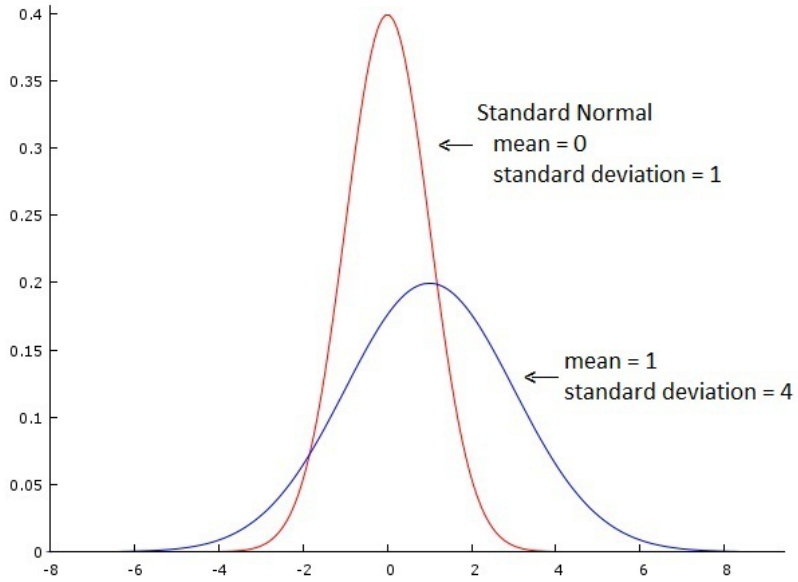



Figure 7: Normal probability distributions with mean= 0 and standard deviation = 1 (standard normal) and with mean= 1 and standard deviation = 4 .

The family of normal probability distributions, two members of which are plotted in Figure 7 above, is the most important of all. It is an excellent model for a wide variety of phenomena. The `normal random variable` is a continuous one and the `normal probability density function` is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(1/2)[(x-\mu)/\sigma]^2} \tag{25}$$

where $-\infty \leq x \leq +\infty$, $-\infty \leq \mu \leq +\infty$, $\sigma > 0$, $\pi = 3.14159$ and $e = 2.71828$. The mean and variance of a normal probability distribution are denoted respectively as

$$E\{X\} = \mu$$

and

$$\sigma^2\{X\} = \sigma^2.$$

Each parameter pair $(\mu\,,\,\sigma)$ corresponds to a different member of the family of normal probability distributions. Every normal distribution is bell shaped and symmetrical, as shown in the Figure above, with each centred at the value of $\mu$ and spread out according to the value of $\sigma$. Normal distributions are referred to using the compact notation $N(\mu\,,\,\sigma^2)\,.$

The `standardised normal distribution` is the most important member of the family of normal probability distributions—the one with $\mu = 0$ and $\sigma = 1$. The normal random variable distributed according to the standard normal distribution is called the `standard normal variable` and denoted by $Z$. It is expressed as

$$Z \;=\; \frac{X-\mu}{\sigma}\,, \tag{26}$$

so that $Z$ is measured as number of standard-deviations. A basic feature of normal distributions is that any linear function of a normal random variable is also a normal random variable. Thus

$$Z \;=\; -\frac{\mu}{\sigma} + \frac{1}{\sigma}\,X \tag{27}$$

and

$$X \;=\; \mu + \sigma\,Z \tag{28}$$

If $V$ and $W$ are two independent normal random variables with means $\mu_v$ and $\mu_w$ and variances $\sigma_v^2$ and $\sigma_w^2$ respectively, the sum $V + W$ is a normal random variable with mean $\mu = \mu_v + \mu_w$ and variance $\sigma^2 = \sigma_v^2 + \sigma_w^2$.

**Exercises**

1. Under what circumstances does $P(A_1 \cap B_1) = P(A_1)\,P(B_1)$? And under what circumstances does $P(A_1 \cup B_1) = P(A_1) + P(B_1)$?

2. Use the relationship between joint and conditional probabilities to derive Bayes Theorem.

3. On a fresh spreadsheet, construct the probability distribution of a poisson random variable with $\lambda = 2$ and then obtain the cumulative probabilies. [Start with a columm of X values ranging from 0 to 12 and then use the mathematical probability function to calculate the probabilities of observing the X's under the specified value of $\lambda$.] Then plot the probability function and the cumulative probability function on separate graphs.

4. How does one standardize a variable? On the spreadsheet above, construct and plot the standard normal probability density function and then the cumulatitive probability distribution. [In this case create a column of Z's ranging from 3.0 to -3.0 in increments of .01. Then use the probability density equation to obtain the densities for each X value.] Finally calculate the cumulative probabilities, keeping in mind that the width of each vertical slice is 0.1, and plot both the density and the cumulative probabilities on separate graphs.

To check your answers to the first three questions, simply consult the relevant parts of the main document. To check your construction and plots of the poisson and standard normal distributions, consult the spreadsheet file `probdans.xls`.

## 3. Hypotheses Tests[5]

You should already understand that a population is a finite or infinite set of elements of interest, with infinite populations normally resulting from processes. Our task is to infer information about the parameters of the probability distributions of populations by examining sample statistics obtained from samples of those populations. The two most important of these sample statistics are the sample mean

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

and sample variance

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} \tag{2}$$

from which we try to infer information about the population mean and variance.

Most important in this process is the `Central Limit Theorem` which states that when the sample size is sufficiently large the sample mean $\bar{X}$ will become approximately normally distributed with mean equal to the population mean and variance equal to the population variance divided by the sample size. And the larger the sample size, the closer the approximation of the sampling distribution of $\bar{X}$ to a normal distribution. This holds true regardless of the distribution of the population provided it has a finite standard deviation.

The true standard deviation of the sample mean is $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ but, since the population standard deviation is usually not known, we use

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

to provide an estimate of $\sigma$. The standard deviation of the sample mean is thus estimated as

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} \tag{3}$$

[5]Background for the material covered in this section can be obtained by reading chapters 4 through 7 of *Statistical Analysis for Economists: A Beginning.*

which becomes closer to $\sigma/\sqrt{n}$ as the sample size gets larger.

The sample mean is thus a `point estimator` of the population mean. When picking point estimators of a population parameter, it is important to keep in mind the desirable properties of such estimators. They should be `unbiased` in that they are no more likely to be on one side of the population parameter than on the other, `consistent` in that the larger the sample size the closer they become to the population parameter, and `relatively efficient` in the sense that they have a smaller variance than that of other unbiased estimators that could have been chosen.

To obtain an estimate of how far the sample mean is likely to deviate from the population mean—that is, how tightly it is distributed around the population mean—we use our estimate of the variance of the sample mean

$$s_{\bar{x}}^2 = \frac{s^2}{n}.$$

Given the characteristics of normal distributions, we can say that if the sample is large enough, and $\bar{X}$ is therefore normally distributed, the sample mean will lie within a distance of $\pm 2\, s_{\bar{x}}$ of $\mu$ with probability near $.95$, and within a distance of $\pm\, s_{\bar{x}}$ of the population mean with probability near $.68$. When the random sample is reasonably large we can set confidence limits $U$ and $L$ for the location of the population mean $\mu$ with approximate confidence coefficient $(1-\alpha)$, which is the probability that $U$ and $L$ will bracket the fixed population mean. These limits will be a distance from the sample mean equal to some multiple $z$ of the standard deviation of that sample mean

$$\bar{X} \pm z\, \frac{s}{\sqrt{n}}$$

where $z = z\,(1-\alpha/2)$ is the $100\,(1-\alpha/2)$ percentile of the standard normal distribution. The $100\,(1-\alpha)$ percent confidence interval for $\mu$ is

$$\bar{X} - z\, \frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + z\, \frac{s}{\sqrt{n}}\,.$$

The limits $-z\,(1-\alpha/2)$ and $z\,(1-\alpha/2)$ are given by the innermost edges of the areas beyond the vertical black lines on the left and right sides of Figure 1 below. These areas each contain a probability weight equal to $\alpha/2$. So for a 95% confidence interval each of these areas represents the probability weight $(1-.95)/2 = .05/2 = .025$ and the sum of these areas represents

the probability weight .05. The area under the probability density function between the two areas is equal to the probability weight .95. If the confidence interval actually brackets $\mu$ that confidence interval is said to be correct.

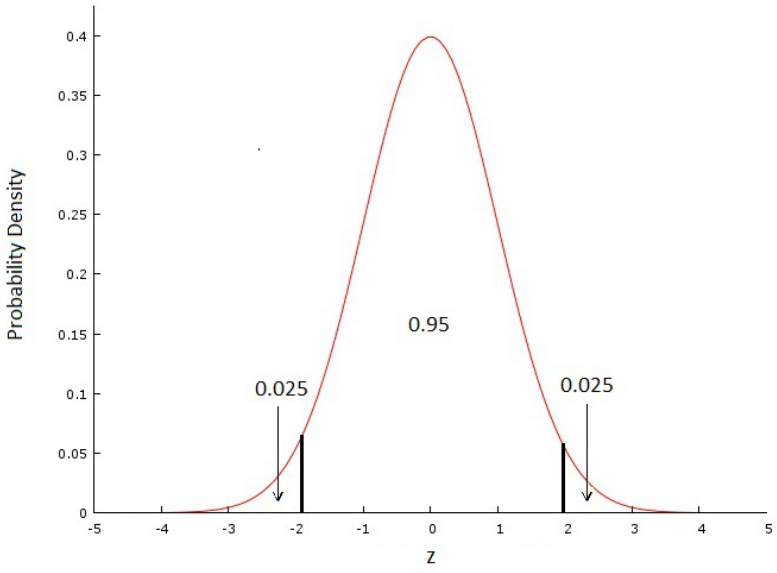



Figure 1: The areas $(1 - \alpha)$ and $\alpha/2$ for a standard normal probability distribution with $\alpha = .05$.

We have been standardized the sampling distribution of $\bar{X}$, obtaining

$$z = \frac{(\bar{X} - \mu)}{s/\sqrt{n}}$$

using $s$ as an estimator of $\sigma$ and then calculated limits for $\mu$ based on values for $z$ obtained from the standard normal probability distribution. Had we known $\sigma$, the standardized value would have been

$$z = \frac{(\bar{X} - \mu)}{\sigma/\sqrt{n}}\,.$$

It turns out that when we use the random variable $s/\sqrt{n}$ instead of the constant $\sigma/\sqrt{n}$, the random variable

$$z = \frac{(\bar{X} - \mu)}{s/\sqrt{n}}$$

is distributed according to the $t$-distribution rather than the standard normal distribution.

The $t$-distribution is symmetrical about zero like the standardized normal distribution but it is flatter, being less peaked in the middle and extending out beyond the standard normal distribution in the tails. The distribution has one parameter, $v$, equal to the degrees of freedom, which equals the sample size minus unity in the case at hand. It has mean zero and variance $v/(v-2)$ with $v > 2$. As the degrees of freedom gets larger and larger the $t$-distribution becomes a closer and closer approximation to the standard normal distribution. An example is presented in Figure 2 below, where $v$ is alternatively set at 5 and 1000. In the case where $v = 1000$, the plot is virtually indistinguishable from a plot of the standard normal distribution since the variance is $1000/98 = 1.002$.

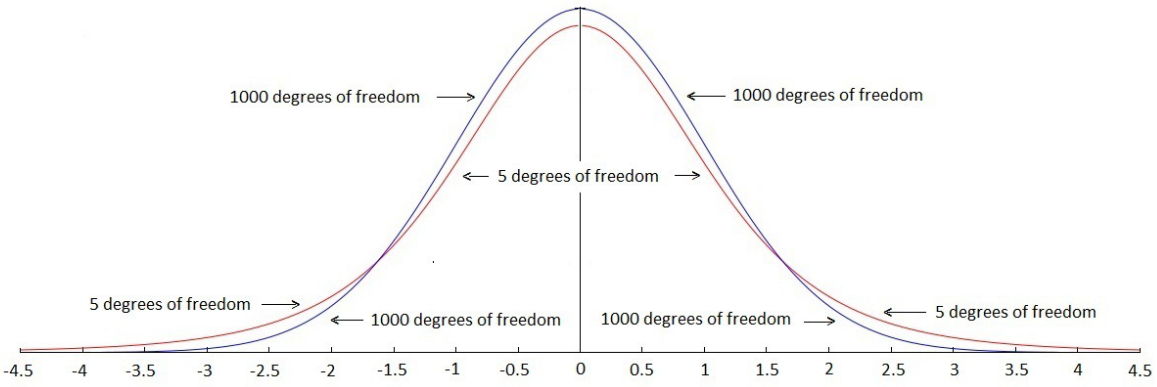



Figure 2: Two t-probability distributions with zero means and degrees of freedom equal to 5 and 1000 respectively.

Suppose our economic analysis leads to the conclusion that the population mean $\mu$ is above some level $\mu_0$, and we want to use our sample mean, which happens to be above $\mu_0$, as evidence of this. We attempt to avoid the worst mistake we could make—namely, to conclude that $\mu$ bigger $\mu_0$ when in truth it is smaller—by setting our null hypothesis as

$$H_0 : \mu \leq \mu_0$$

and then determining whether it is reasonable to reject that null-hypothesis in favour of the alternative hypothesis

$$H_1 : \mu > \mu_0$$

on the basis of our sample evidence. Had $\bar{X}$ been less than $\mu_0$, we would have no basis for rejecting the null-hypothesis. The fact that $\bar{X}$ is above $\mu_0$ casts doubt upon the null hypothesis but, since $\bar{X}$ is a random variable, the question arises as to the probability of observing a sample mean that high if the null hypothesis is really true. We impose on ourselves the decision rule that we will not reject the null hypothesis in favour of the alternative hypothesis unless the probability of observing the sample mean when the null hypothesis is in fact true is less than $0.01$. Alternatively, we could have chosen $0.025$ or $0.05$ as our decision rule. We then calculate the critical value that $\bar{X}$ would have to exceed to enable us to reject the null hypothesis and conclude that the alternative hypothesis is correct without violating our decision rule. The risk that we will incorrectly conclude that the null-hypothesis is false then it is in fact true is called the $\alpha$ risk $(= 0.01)$ and the risk that we will incorrectly conclude that the null-hypothesis is true when it is in fact false and the alternative hypothesis is true is called the $\beta$ risk. Our critical value will be

$$A = \mu_0 + z\,(1 - \alpha)\,s_{\bar{x}} = \mu_0 + z(0.99)\,\frac{s}{\sqrt{n}}$$

and we will conclude that the null hypothesis is false and $\mu > \mu_0$ if $\bar{X} > A$.

Alternatively, we can calculate the probability that $\bar{X}$ would be as large as it is or larger if $\mu = \mu_0$. This will equal one minus the cumulative probability of observing a magnitude of $z$ equal to or greater than

$$z^* = \frac{\bar{X} - \mu_0}{s_{\bar{x}}}\,.$$

If this probability, which is called the $P$-Value, is less than $\alpha$, we reject the null hypothesis. The $P$-Value is the probability of observing a sample mean as large as the one observed if $\mu$ is less than $\mu_0$. Our decision rule simply involves choosing a $P$-Value below which we reject the null hypothesis and conclude that the alternative hypotheses is correct. In all cases, of course, we use the $t$-distribution as the distribution of our sample mean—if the sample size is large enough this distribution will approximate the normal distribution and thereby be an appropriate substitute for it.

Now suppose we have independent random samples of $n_1$ and $n_2$, both reasonably large, from two populations and want to make inferences about the difference in the respective population means $\mu_2 - \mu_1$. We know that

$$E\{\bar{Y} - \bar{X}\} = E\{\bar{Y}\} - E\{\bar{X}\} = \mu_2 - \mu_1$$

and, since the samples are independent,

$$\sigma^2\{\bar{Y} - \bar{X}\} = \sigma^2\{\bar{Y}\} + \sigma^2\{\bar{X}\}.$$

We can thus use

$$s^2\{\bar{Y} - \bar{X}\} = s^2\{\bar{Y}\} + s^2\{\bar{X}\}$$

as an unbiased point estimator of $\sigma^2\{\bar{Y} - \bar{X}\}$. The standard error of the difference between the sample means is thus

$$s\{\bar{Y} - \bar{X}\} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}. \tag{4}$$

Using this standard error, we can calculate confidence limits and $P$-Values in the same way as we previously did. Again it is appropriate to use the $t$-distribution although we need large samples to obtain reasonable results.

In some cases it may be reasonable to assume that both populations are normally distributed with the same variance. In this case

$$\sigma^2\{\bar{Y} - \bar{X}\} = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2\left[\frac{1}{n_1} + \frac{1}{n_2}\right]. \tag{5}$$

To calculate confidence intervals we can then use the `pooled or combined estimator`

$$\begin{aligned} s_c^2 &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)} \\ &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \end{aligned} \tag{6}$$

as an unbiased estimator of $\sigma^2$. Then

$$s^2\{\bar{Y} - \bar{X}\} = s_c^2\left[\frac{1}{n_1} + \frac{1}{n_2}\right] \tag{7}$$

is our unbiased estimator of $\sigma^2\{\bar{Y} - \bar{X}\}$.

An interesting situation arises when we want to use samples to make inferences about a population before and after an event—for example, a shipping company might want to determine the weight loss of bananas as a result of shipment. Instead of comparing the weights of a random sample of banana bunches before shipment with a the weights of a random sample after shipment, it would be better to compare the weights of the same sample before and after shipment, analyzing these `paired differences`. This is the case because the correlation between the before and after weights is positive so that

$$\sigma^2\{\bar{Y} - \bar{X}\} = \sigma^2\{\bar{Y}\} + \sigma^2\{\bar{X}\} - 2\sigma\{\bar{Y}\bar{X}\} < \sigma^2\{\bar{Y}\} + \sigma^2\{\bar{X}\}.$$

So the best procedure is to obtain

$$D_i = Y_i - X_i\,,$$

where $Y_i$ is the weight of the $i$th bunch before shipment and $X_i$ is the weight of that same bunch after shipment, and then calculate

$$\bar{D} = \frac{\sum_{i=1}^{n} D_i}{n}$$

and

$$s_D^2 = \sum_{i=1}^{n} \frac{(D_i - \bar{D})^2}{n-1}$$

from whence

$$s_{\bar{D}} = \sqrt{\frac{s_D^2}{n}}.$$

Confidence intervals and $P$-Values are then calculated in the same fashion as we would do to make inferences about a single mean.

Suppose now that we need to make inferences about the magnitude of the variance of a population. This requires an understanding of the `chi-square ( $\chi^2$ ) distribution`. It turns out that the sum of $n$ squared standardized independent normal random variates,

$$Z_1^2 + Z_2^2 + Z_3^2 + \ldots + Z_n^2,$$

is distributed as a chi-square distribution. Accordingly,

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 = \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} \quad = \quad \chi^2(n) \tag{8}$$

where $\chi^2(n)$ is a random variable distributed according to the chi-square distribution, with parameter $n$ equal to the number of independent normal variates summed and thereby to the degrees of freedom. When we replace $\mu$ in the numerator with $\bar{X}$ we obtain the $\chi^2$ statistic

$$\sum_{i=1}^{n} Z_i^2 \;=\; \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2} \;=\; \chi^2(n-1) \tag{9}$$

where the degrees of freedom parameter is now $n-1$.

Notice now that the standard expression for $s^2$,

$$s^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1},$$

can be rewritten as

$$\sum_{i=1}^{n} (X_i - \bar{X})^2 = (n-1)\, s^2.$$

and substituted into (9), to yield

$$\frac{(n-1)\, s^2}{\sigma^2} \;=\; \chi^2(n-1). \tag{10}$$

The sampling distribution for this statistic is skewed to the right, with the skew being smaller the greater the degrees of freedom. Figure 3 shows a $\chi^2$ distribution with 12 degrees of freedom. The middle vertical line gives the mean and the other two vertical lines the critical values for $(1-\alpha) = .95$. The mean of the $\chi^2$ distribution is the number of degrees of freedom, which in the the example above equals $n-1$. The variance of the $\chi^2$ distribution is twice the number of degrees of freedom. The fractions of the probability weight below given values of $\chi^2$ for the family of chi-square distributions can be obtained from tables at the back of any standard textbook in statistics or from the free statistical software noted earlier. Rearranging (10) to put $\sigma^2$ on the right side

$$\frac{(n-1)\, s^2}{\chi^2(n-1)} = \sigma^2$$

and then choosing the lower and upper values of the $\chi^2$ statistic that delineate a $(1-\alpha)$ confidence interval—namely,

$$\chi^2(\alpha/2; n-1) \qquad \text{and} \qquad \chi^2(1-\alpha/2; n-1),$$

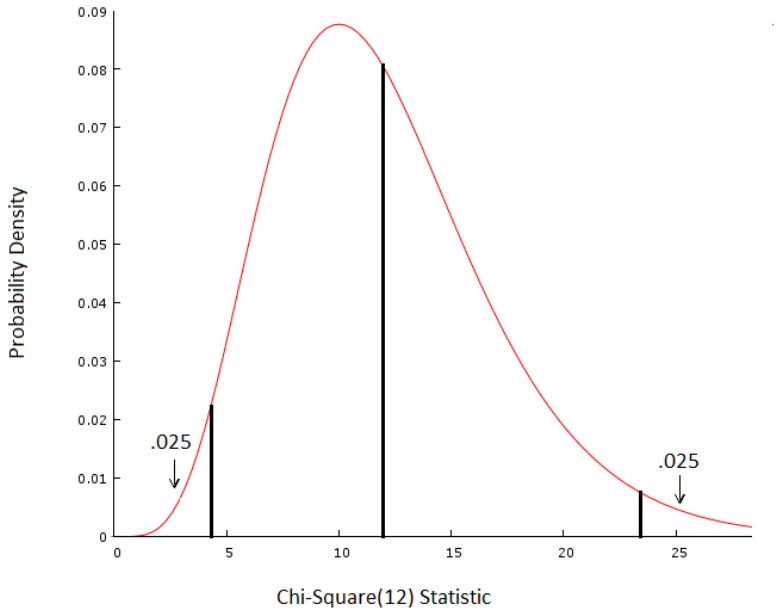



Figure 3: A chi-square distribution with 12 degrees of freedom. The middle vertical line represents the mean and probability weight equal to .025 lies beyond each of the other two vertical lines.

we obtain upper and lower critical values equal to

$$\frac{(n-1)\,s^2}{\chi^2(\alpha/2;n-1)} = U$$

and

$$\frac{(n-1)\,s^2}{\chi^2(1-\alpha/2;n-1)} = L\,.$$

Furthermore, if we want to obtain the $P$-Value for the null-hypothesis that $\sigma^2 < \sigma_0^2$ we simply calculate the cumulative probability associated with the $\chi^2$ statistic

$$\frac{(n-1)\,s^2}{\sigma_0^2}\,.$$

We are often interested in comparing the variability of two populations. To do this we need a statistic based on the two values of both $s_i$ and $n_i$ that is distributed according to an analytically tractable distribution. It turns out that the ratio of two chi-square variables, each divided by their respective degrees of freedom, is distributed according to the `F-distribution`. That is,

$$\frac{\chi^2(v_1)/v_1}{\chi^2(v_2)/v_2} \quad = \quad F(v_1, v_2) \tag{11}$$

is distributed according to the $F$-distribution with parameters $v_1$ and $v_2$, which are the degrees of freedom of the respective chi-square distributions—$v_1$ is referred to as the degrees of freedom in the numerator and $v_2$ is the degrees of freedom in the denominator. The mean and variance of the $F$-distribution are

$$E\{F(v_1, v_2)\} = \frac{v_2}{(v_2 - 2)}$$

when $v_2 > 2$, and

$$\sigma^2\{F(v_1, v_2)\} = \frac{2\, v_2^2\, (v_1 + v_2 - 2)}{v_1\, (v_2 - 2)^2\, (v_2 - 4)}$$

when $v_2 > 4$. The probability density function for an $F$-distribution with 40 degrees of freedom in the numerator and 60 degrees of freedom in the denominator is plotted in Figure 4. The mean is $60/58$, which is close to unity, and the two thick vertical lines give the critical values for $(1-\alpha) = .90$. The percentiles for this distribution can be found in the $F$-tables at the back of any textbook in statistics or calculated using one of the freely available statistical programs previously noted. The tables give only the percentiles above 50 percent. To obtain the percentiles below 50 percent we must utilize the fact that the lower tail for the $F$-value

$$\frac{\chi^2(v_1)/v_1}{\chi^2(v_2)/v_2} = F(v_1, v_2)$$

is the same as the upper tail for the $F$-value

$$\frac{\chi^2(v_2)/v_2}{\chi^2(v_1)/v_1} = F(v_2, v_1).$$

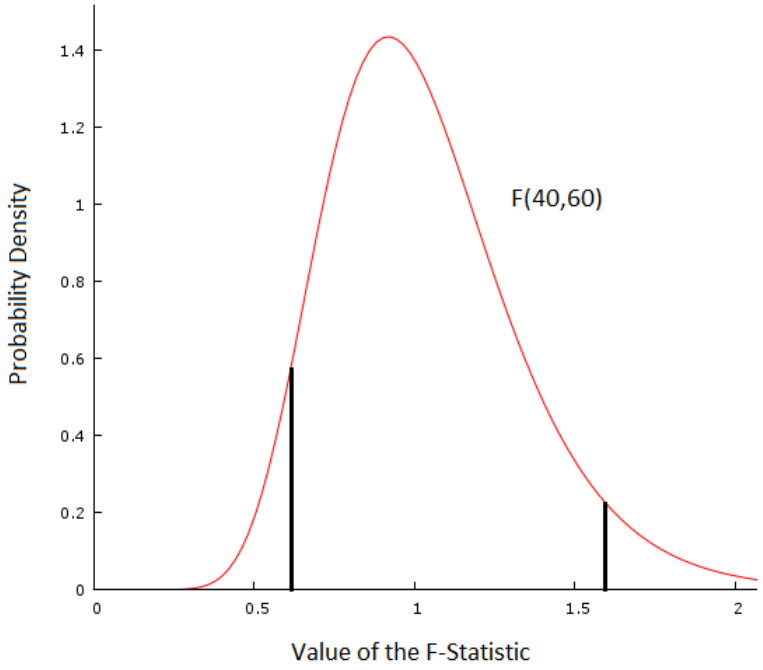



Figure 4: An F-distribution with 40 degrees of freedom in the numerator and 60 degrees of freedom in the denominator. The thick vertical lines give the extremes beyond which 5 percent of the probability mass lies, with 90 percent of the probability mass lying between these extremes.

This implies that

$$F(\alpha/2; v_1, v_2) = \frac{1}{F(1-\alpha/2; v_2, v_1)}.$$

Equation (10) can be written more generally as

$$\frac{v\, s^2}{\sigma^2} \quad = \quad \chi^2(v) \tag{12}$$

which implies that

$$\frac{s^2}{\sigma^2} = \frac{\chi^2(v)}{v}.$$

This expression can be substituted appropriately into the numerator and denominator of equation (11) to yield

$$\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \;=\; F(v_1, v_2) \;=\; F(n_1 - 1, n_2 - 1). \tag{13}$$

To establish confidence intervals in a particular case, we manipulate (13) to yield

$$\frac{\sigma_2^2}{\sigma_1^2} \;=\; F(n_1 - 1, n_2 - 1)\, \frac{s_2^2}{s_1^2}\,. \tag{14}$$

To calculate a 90 percent confidence interval we obtain the values of the statistic $F(n_1 - 1, n_2 - 1)$ at $\alpha/2 = .05$ and $1 - \alpha/2 = .95$ respectively and plug them, in turn, into the equation above. Note that this confidence interval is based on the assumption that the two populations of measurements from which the sample variances are obtained are normally distributed or approximately so.

Finally, it should be noted that in many applications involving F-statistics beyond the comparison of the variances of two samples, the degrees of freedom in the numerator will be much lower relative to the degrees of freedom in the denominator than in Figure 4. Accordingly, a better impression of the range of distributions of F-statistics is provided in Figure 5 below. The F-distribution is skewed to the right by a larger amount, the lower the degrees of freedom in the numerator. Situations with low degrees of freedom in the numerator are encountered in regression analysis, to which we now turn.

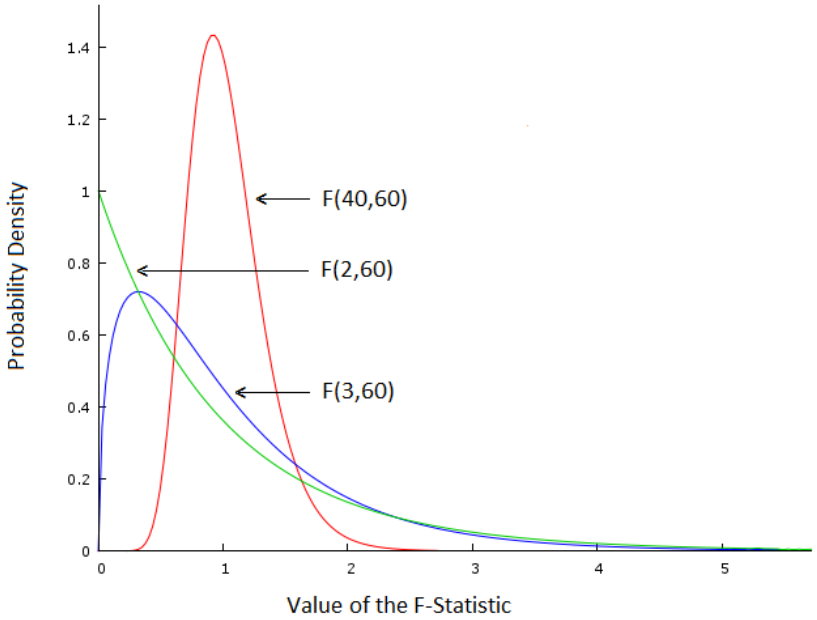



Figure 5: Three F-distributions with 60 degrees of freedom in the denominators and 2, 3 and 40 degrees of freedom in the respective numerators.

### Exercises

For these exercises, make your calculations on the answer worksheet for the exercises in the first section, `datanala.xls`.

1. Keeping in mind that the two growth rates are positively correlated, what is the $P$-Value of the null-hypothesis that the means of the month-over-month growth rates of M1 and M2 are the same. [This involves some adaptation of the technique presented above in this section.]

2. Impose an assumption (albeit incorrect) that the month-over-month growth rates of M1 and M2 are statistically independent. Given the data, what is the $P$-Value of the null-hypothesis that the variances of M1 and M2 are the same.

Answers are calculated in the spreadsheet file `hyptesta.xls` which should be consulted along with the material presented in this section above.

## 4. Ordinary Least Squares Regression Analysis[6]

We now turn to the analysis of relationships between variables, the area of statistics of most relevance to economics. For example, it might be argued that the aggregate quantity of money people choose to hold will be determined by the volume of transactions they need to make which will, in turn, be directly related to the flow of output being produced in the economy. This would suggest a possible linear relationship that can be expressed in statistical terms as

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{1}$$

where $Y_i$ is the real stock of money, which is the nominal stock measured in currency units of some base period, in the $i^{th}$ period, $X_i$ is the level of output in the $i^{th}$ period and $\beta_0$ and $\beta_1$ are parameters. In addition, an error $\epsilon_i$ has been added in each period to the deterministic relationship $Y = \alpha + \beta X$ to allow for the possibilities that $Y_i$ and $X_i$ are measured with error, that the deterministic relationship between the two variables may not be exactly linear, and that there may be additional variables affecting $Y$ that could not be measured well enough to include or are not known about at all. Often a relationship may not be linear in the actual levels of $X$ and $Y$ but may be linear in the logarithms of those variables—when the variables are in logarithms, $\beta$ represents the elasticity of $Y$ with respect to $X$ rather than the slope of a straight-line relationship.

The standard method of statistically estimating the parameters in the above relationship is to use ordinary least squares (OLS). This involves fitting a linear relationship of the form

$$Y_i = b_0 + b_1 X_i + e_i \tag{2}$$

to the data in such a way as to minimize the sum of the squared deviations of the actual level of $Y_i$ from the predicted level which we can call $\hat{Y}_i$. That is, we choose the magnitudes of $b_0$ and $b_1$ (the estimates of $\beta_0$ and $\beta_1$) that minimize

$$\sum_{i=1}^{n} e_i^2 ,$$

where $n$ is the number of observations and $e_i$ is the measured residual as compared to the true one $\epsilon_i$.

[6] An appropriate background for the material covered in this section can be obtained by reading chapters 8 and 9 of *Statistical Analysis for Economists: A Beginning*.

It turns out that under three conditions,

1. The errors are unbiased—that is, $E(\epsilon_i) = 0$ , given the level of $X_i$ (and the $i^{th}$-period levels of any other explanatory variables that might be added to the regression),

2. The variance of $\epsilon_i$ equals a constant $\sigma^2$ , again conditional on all explanatory variables,

and

3. The errors $\epsilon_i$ and $\epsilon_j$ are independent of each other, which means that $E(\epsilon_i, \epsilon_j) = 0$ for all $i$ and $j$ where $i \neq j$ ,

the OLS estimator of $\beta_1$ has the lowest variance of all possible unbiased estimators that one could use—this is the `Gauss-Markov Theorem`. There is no requirement that the errors be normally distributed, although that is often assumed. The second and third of these conditions specify that errors are `homoskedastic`. Violations of these conditions where the variance of the errors is related to one or more of the $X$ variables are called `heteroskedasticity`. And violations where the $\epsilon_i$ and $\epsilon_j$ are related to each other, which occurs often in time-series analysis, are called `serial correlation`. The major effort involved in OLS-regression analysis is dealing with violations of these conditions.

In the case where there is a single explanatory variable, estimation involves minimization of

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

which leads to the estimators

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{3}$$

and

$$b_0 = \bar{Y} - b_1 \bar{X} \,. \tag{4}$$

As an example, the worksheet `olssect.xls` contains measures of U.S. real M1 and U.S. real GDP from the first quarter of 1959 through the first quarter of 2010, in columns `C` and `K` respectively. The means are calculated by

placing the code `average(C20:C224)` in cell `C231` and, using a more indirect method, by placing the code `sum(K20:K224)` in cell `K228`, then the number of observations (= `205`) in cell `K229` and then the code `K228/K229` in cell `K231`. The resulting means for U.S. real GDP and U.S. real M1 are 7279.37 and 1049.98 respectively. Next, the squared deviation of real GDP from its mean for 1959Q1 is placed in cell `O20` by entering the code `(C20-7279.37)^2` in that cell and the code `(K20 - 1049.98)^2`, delineating the squared deviation of real M1 in 1959Q1 from its mean, is placed in cell `P20`. And the product of the deviation of real GDP from its mean and the deviation of real M1 from its mean is placed in cell `Q20` by entering the code `(C20-7279.37)*(K20-1049.98)`. The code in the adjacent cells `O20`, `P20` and `Q20` is then copied to the corresponding cells immediately below all the way down to row `224`. Then the sums of these columns are placed in `O228`, `P228` and `Q228` by simply copying the code in `K28` to these three cells. You should keep in mind that when you copy code from one cell to another, Gnumeric automatically changes that cell-numbers in that code to apply the same operation to the current row or column as was applied to the original row or column. If that automatic adjustment of row and column in the code is not what you want, then such copying cannot be undertaken without producing incorrect calculations.

Now the value of $b_1$ is obtained by entering the code `Q228/O228` in cell `O234`. [Note that copying this cell to another place on the worksheet will usually produce nonsense—for example, if you copied it to the next cell to the right the code in that cell would turn out to be `R228/P228` which would yield a value of 0.0000.] The magnitude of $b_0$ is then obtained by entering the code `K231-O234*C231` in cell `O235`. The resulting values for the two regression coefficients are $b_1 = 0.0665654$ and $b_0 = 563.97$.

The $R^2$ is the fraction of the variability of $Y$ explained by $X$. The sum of squared deviations of U.S. real M1, which represents the variability that we are trying to explain, has already been placed in cell `P228`. To obtain the unexplained variability we need to obtain the regression residuals—that is the difference between the actual and predicted levels of real M1. We put the predicted, or fitted, level of M1 for 1959Q1 in cell `R20` by entering the code `563.97 + 0.0665654*C20`. Then we place the residual for that quarter in cell `S20` by entering the code `K20-R20`, which subtracts the fitted from the actual, in that cell. The fitted and residuals for the remaining quarters are now obtained by copying the contents of these cells to the cells below down

to row 224—the row entry codes in the cells are automatically adjusted by Gnumeric. The squared residuals are then placed in column T by a coding process that should by now be obvious to you, and the sum of squared errors are then obtained by summing that column using the code sum(T20:T224) placed in cell T228. The sum of squared residuals (errors) is usually denoted as

$$SSE = \sum_{i=1}^{n} e_i^2$$

and the total sum of squares to be explained as

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 .$$

The sum of squares explained by the regression, denoted as *SSR*, is thus equal to

$$SSR = SST - SSE$$

and the fraction of the variations in U.S. real M1 explained by the regression is therefore

$$R^2 = \frac{SST - SSE}{SST} = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2 \; - \; \sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} e_i^2} . \tag{5}$$

Calculating the $R^2$ and placing it in cell O236 simply involves entering the code T229/P228 in that cell—the $R^2$ is 0.84.

Next we need to obtain the standard-errors of $b_0$ and $b_1$. This first involves obtaining the mean squared error (*MSE*), which is our estimate of $\sigma^2$.

$$\sigma^2 = MSE = \frac{SSE}{df} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} \tag{6}$$

where *df* is the degrees of freedom, which is equal to the number of observations minus the number of regressors including the constant. The mean square error in the regression we are studying is obtained by placing the code T228/(205-2) in cell T231. The standard-deviation of $b_1$ is then

$$\sigma\{b_1\} = \sqrt{\frac{MSE}{\sum_{i=1}^{n} (X_i - \bar{X})^2}} \tag{7}$$

and that of $b_0$ is

$$\sigma\{b_0\} = \sqrt{MSE\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]}. \tag{8}$$

In the equation for the standard-error of $b_0$, the term on the left in the square brackets multiplied by *MSE* measures the effect of random variations in the level of the regression line and the term on the right multiplied by *MSE* measures the effect of random variations in the slope of the regression line at the mean level of $X$. The standard-errors are calculated by entering the code `sqrt(T231/O228)` in cell `O237` and the code `sqrt(T31*((1/205)+(C231^2/O228)))` in cell `O238`. The calculated standard-errors of $b_0$ and $b_1$ are, respectively, 16.215 and .002035. The $t$-statistics—which equal the coefficients divided by their respective standard-errors—are calculated by entering `O234/O237` in cell `P234` and `O235/O238` in cell `P235` [or, in the latter case, by simply copying the code from cell `P234` to cell `P235`]. These $t$-statistics are both larger than 30, which indicates $P$-Values of virtually zero for the null hypotheses that the coefficients are zero. In the case of $b_0$, this essentially rules out the possibility that the ratio of real M1 to real GDP is constant. Another measure of the statistical significance of the regression is an $F$-test using the $F$-statistic

$$F = \frac{SST - SSE/[(n-1) - (n-2)]}{SSE/(n-2)} = \frac{SSR}{MSE} \tag{9}$$

where the degrees of freedom in the numerator $[(n-1) - (n-2)] = 1$ is the excess of the degrees of freedom used in calculating the standard-error of the dependent variable over the degrees of freedom used in calculating the standard-error of the regression. This statistic is calculated by entering the code `T229/T231` in cell `O239` and has a magnitude in excess of 1000, clearly indicating statistical significance of the regression. This calculation of the $F$-statistic was not really necessary here because when there is a single regressor other than the constant the $F$-statistic is the square of the $t$-statistic for $b_1$, as can be seen by squaring the number in cell `P234`.

Finally, it is absolutely essential to determine whether the residuals are homoskedastic. The crudest way of doing this is to plot them. In Gnumeric, this is done in the present case by high-lighting the cells `S20` through `S224`, then clicking on the `insert a chart` icon, selecting `type line`, then

insert, and then placing the cursor at the point where you want the upper left corner of the chart to be and holding down the left mouse-button while dragging the the cursor to the point where you want the bottom right corner of the chart to be. You can move the chart simply by placing the cursor on it, holding the left mouse-button and dragging the cursor. And you can change the chart's size by holding cursor at the bottom-left corner so that a double-sided arrow appears and then holding down the left mouse-button while you move the cursor. In the present example, it is obvious from looking at the chart that there is very substantial serial correlation in the residuals.

The pattern of the residuals suggests that it might be better to use the logarithms of real M1 and real GDP instead of the raw values. This is done in the Gnumeric spreadsheet using the same techniques that were used previously—nothing further would gained by a detailed discussion of the methods of doing this. The result is an elasticity of the response of real M1 to real GDP of around 0.44, with a slightly higher $R^2$ and $F$-statistic. But the plot of the residuals shows a quite similar pattern of serial correlation as occurred in the case where logarithms were not used.

One important reason for the serially-correlated residuals might be the fact that we have left out an important variable affecting the demand for money. Since the interest rate represents the cost of holding money, it is essential that it be included in the regression. To do this we use the regression-running procedure provided by Gnumeric. The exact mathematical calculations involved will be worked through in the document entitled *Statistical Analysis Using XLispStat, R and Gretl: A Beginning*. Here it is essential that the explanatory variables be next to each other in the spreadsheet. We therefore copy the three variables, logarithm of real GDP, 3-month treasury-bill interest rate and logarithm of real M1 to columns AC, AD and AE respectively. When pasting the variables in the new places we use the Paste special option and choose paste As Value on the Paste-type menu—this avoids the re-orientation of any cells referred to in codes that were in the original cell-entries. To run the regression, we click on Tools, then on Statistical Analysis and then on Regression. When the regression window appears, we click on the button that refers to X variables and enter AC20:AD142 in the window that then appears. Then we click on the Y variable button and enter AE20:AE142 in the window that then appears. Then we click on OK and a new sheet containing the regression results will then be added to the spreadsheet file. The results are in the sheet named Regression (2). I have

taken the liberty of writing in the names of the variables in the appropriate places. The $R^2$ is clearly higher as a result of adding the additional variable, but we have to be careful because adding another variable reduces the degrees of freedom—the $R^2$ could be increased to unity by adding variables until the number of variables equaled the number of observations! Accordingly, we must use an $R^2$ that takes account of this, an adjusted $R^2$ called $\bar{R}^2$ which takes the form

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{SSE}{SST} \tag{10}$$

where $n$ is the number of observations and $k$ is the number of regressors including the constant term. The addition of a variable has a positive effect on $\bar{R}^2$ in that it lowers the sum of squared errors, and a negative effect in that it lowers the ratio of $(n-1)$ over $(n-k-1)$. While $\bar{R}^2$ can thus go in either direction, it will always be less than $R^2$ which, as can be seen from equation (5), equals

$$R^2 = 1 - \frac{SSE}{SST}.$$

It turns out that the $\bar{R}^2$ in the regression with the interest rate added is higher than the $R^2$ in the earlier regression, so the fit has clearly improved. As you can see from the main spreadsheet, however, a plot of the residuals indicates that serial correlation is clearly still present.

The sheet of the spreadsheet file named `Regression (1)` shows the regression results when the actual levels rather than the logarithms of real M1 and real GDP are used and the interest rate is included. The $\bar{R}^2$ and $F$-statistic are clearly smaller in that case, indicating that the use of the logarithms of real M1 and real GDP is the better approach.

Finally, the sheets `Regression (3)` and `Regression (4)` of the spreadsheet file present the regression result when the real M2 monetary aggregate is used as the dependent variable instead of real M1. In both cases the adjusted $R^2$ statistics are higher as are the levels of the $F$-statistics used in the test for statistical significance of the regressions. While the $P$-Values for the log of real M2 shown on the two sheets are clearly larger than 0.05, it must be noted that Gnumeric makes two-tailed tests whereas we are interested only in the lower-tail. Accordingly, the single-tailed $P$-Values are one-half

the magnitudes of those shown and we can clearly reject the null-hypothesis of a positive interest elasticity of demand for real M2 at the 5% level.

A problem that has arisen throughout the foregoing regression analysis is the presence of serial correlation in the residuals. This indicates that the Gauss-Markov conditions are not being met. While the issue of how to deal with this problem are dealt with in documents dealing with the basics of econometrics, the techniques for determining whether or not the observed residuals are homoskedastic must be investigated now—clearly, situations will arise where the heteroskedasticity and serial correlation in the residuals may not be obvious from simply looking at a plot of those residuals.

A standard test for the presence of heteroskedasticity in regression residuals is the `Breusch-Pagan test`.[7] To perform this test, we regress the squared residuals of the regression we are testing on some or all of the independent variables in that regression—this involves the OLS-estimation of

$$e_i^2 = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + .... + u_i \tag{11}$$

where the $e_i$ are the residuals from the regression whose residuals we are testing, the $X_{1i}$, $X_{2i}$, etc., are some or all of the independent variables used in that regression, $\beta_0$, $\beta_1$, $\beta_2$, etc., are the coefficients and $u_i$ is the error term. From this regression, we obtain the statistic $n \times R^2$ which is distributed as chi-square with degrees of freedom, here denoted by $k$, equal to the number of independent variables in the regression excluding the constant term—that is,

$$nR^2 = \chi^2(k)\,.$$

The null hypothesis that the residuals are homogeneous is rejected in favour of the presence of heteroskedasticity if this chi-square statistic is large enough—that is, if the $P$-Value is below some critical level. The test is performed in the spreadsheet file on the regression of the log of real M1 on the 3-month treasury bill rate and the log of real GDP. The squared residual from the regression is calculated and presented in column `AH` of the spreadsheet and a regression is calculated with the dependent variable `AH20:AH224` and the independent variables `AC20:AD224`. The results are shown in the `Breusch-Pagan` sheet in the spreadsheet file. Not surprisingly, the chi-square statistic with 4 degrees

[7] T. S. Breusch and A. R. Pagan, "A Simple Test for Heteroscedasticity and Random Coefficient Variation," *Econometrica*, Vol. 47, 1979, pages 1287-1294.

of freedom is in excess of 14 and the $P$-Value for the test is less than .01, leading us to reject the null-hypothesis of homoskedasticity of the residuals.

The reason for this heteroskedasticity is most certainly the presence of serial correlation in those residuals. Traditionally, the Durbin-Watson statistic has been used to test for serial correlation in the residuals. It turns out that the test results for this statistic are difficult to establish precisely and, in addition, it ignores the possible correlations between the current residual and lags of that residual of order greater than one. The best test for serial correlation seems to be one called a Lagrange Multiplier (LM) test.[8] This test involves regressing the regression residual on a series of lags of that residual as well as the independent variables that were in the original regression—that is, fitting the equation

$$e_t = \beta_0 + \beta_1 e_{t-1} + \beta_2 e_{t-2} + .... + e_{t-p} + \gamma_1 X_{1i} + \gamma_2 X_{2i} + .... + v_i \,. \qquad (12)$$

The number of lags of the residual, denoted above by $p$, that should included in this regression is probably two or three for annual data, four or more for quarterly data and perhaps as many as twelve in the case of monthly data if the degrees of freedom are sufficient. An F-test is then performed to determine the statistical significance of the group of lags of the residual. The F-statistic is

$$F = \frac{SSE_R - SSE_U/p}{SSE_U/df} \qquad (13)$$

where $SSE_R$ is the sum of squared residuals for the restricted regression where the lagged residuals are dropped from (12), $SSE_U$ is the the sum of squared residuals of the unrestricted regression where the lagged residuals are included, and $df$ is the degrees of freedom of the unrestricted regression. The problem with this F-test is that it assumes that the residuals in question are normally distributed (since $F$ is the ratio of two Chi-square distributions divided by their degrees of freedom), which is very unlikely to be the case. Accordingly, the test is altered by multiplying the resulting F-statistic above by the number of lagged errors, $p$, to obtain a statistic that has a Chi-square distribution with degrees of freedom equal to $p$ under the assumption that the degrees of freedom in the denominator, $df$, are infinite. Thus, by simply

[8]See G. S. Maddala, *Introduction to Econometrics*, Macmillan Publishing Company, 1988, page 206, for a clear discussion of this test.

using the statistic

$$p\,F = \chi^2(p)$$

which will have a lower $P$-Value than the original F-statistic, we make it more likely that homoskedasticity will be rejected and thereby compensate for the fact that the residuals are not likely to be normally distributed by reducing any bias that would thereby result. The statistic is calculated for the two regressions using the logarithms of real M1 and real M2 as the dependent variables in the spreadsheet file `olssect.xls`. The lags of the two residuals are calculated and placed in columns `AI` through `AL` and `AN` through `AV` and the relevant $X$ variables are copied as value to neighboring columns. The restricted and unrestricted regressions are computed in the two cases, now using rows `24` through `224` because four data observations are lost when we compute the lagged residuals, and the results are placed in the sheets `Serial Corr. Test--LM1` and `Serial Corr. Test--LM2`. It is then easy to calculate the relevant $\chi^2(4)$ statistics—not surprisingly, the null-hypothesis of homoskedasticity can easily be rejected in favour of the alternative hypothesis of serially correlated residuals.

**Exercise**

The file `olssectq.xls` contains data on M1, M2, GDP, prices and government bond yields for Japan for the period 1980Q1 through 2007Q4. Run two-demand-for-money regressions for Japan, one for real M1 and one for real M2, using natural logarithms of the real M1, real M2 and real GDP variables. Plot the actual and fitted values and the residuals for the better fitting of the two regressions. Then conduct a Breusch-Pagan test for homoskedasticity and an LM-test for serial correlation in these residuals. Explain how to run these tests and how to interpret the results.

An answer to this exercise can be found in the file `olssecta.xls`.