# Presentation of Data

I've put together some "pointers" when assembling the data analysis portion of your presentation and final paper. These are not all inclusive, but are things to keep in mind when completing the empirical methodology and results section of your paper.

- As with any graph or table, it is essential to clearly label the axes so that the reader/audience knows how to read the data being presented. Don't use the abbreviations you are familiar with in your workfile. It is ok to use abbreviations if they are ones the reader can follow.
- Do not include information on the graph that you are not going to use or explain. Statistical packages like EViews often include information on graphs by default. A lot of this information you will not need to make your point.
- Don't include graphs just for the sake of having more graphs. You should spend some time thinking about which information is important in order to make your point. Some projects may not use any graphs, others will use several.
- It is rare that you will be able to take a graph or table in it's "raw" form (without formatting) from EViews or Excel and paste it into paper/presentation. In most cases this will not be satisfactory – you will need to make modifications to your graphs/tables to make them presentable. **This takes a great deal of time to do.** Don't expect to be able to simple paste your results, graphs, and tables into your paper and presentation "as is". The information is not useful to the audience and reader if it isn't presented well.
- I've included some examples of what bad versus good tables/graphs look like and why. These examples also include some information what these graphs/tables may reveal to you where applicable. I am assuming that you have a working knowledge of basic statistics from STAT 1. In your papers, you should explain conclusions you draw from your graphs/tables. This doesn't mean interpreting every single statistic, it means pointing the reader/audience to the information you want them to learn from the graph/table.

*Cross section data:*

It is standard to include some presentation of what your sample looks like. In your paper, you should include (at a minimum):

- A histogram of the dependent variable would insure that you have a sample that represents a wide range of individuals.
- A table of summary statistics for the data you are using for your analysis.

*Time series data:*

It is standard to include a presentation of how the variables you are interested in are changing over your sample period. In your paper, you should include (at a minimum):

- A time series plot of the dependent variable. Note, you should always use a line graph, not a bar graph. Bar graphs look unnecessarily cluttered for a time series plot.
- A table of summary statistics for the data you are using for your analysis if this is of interest. You may include summary statistics for different "sub-samples" of data within your larger sample to note any shifts in means or variances over time.

Notice, these rules apply to your regression tables as well.

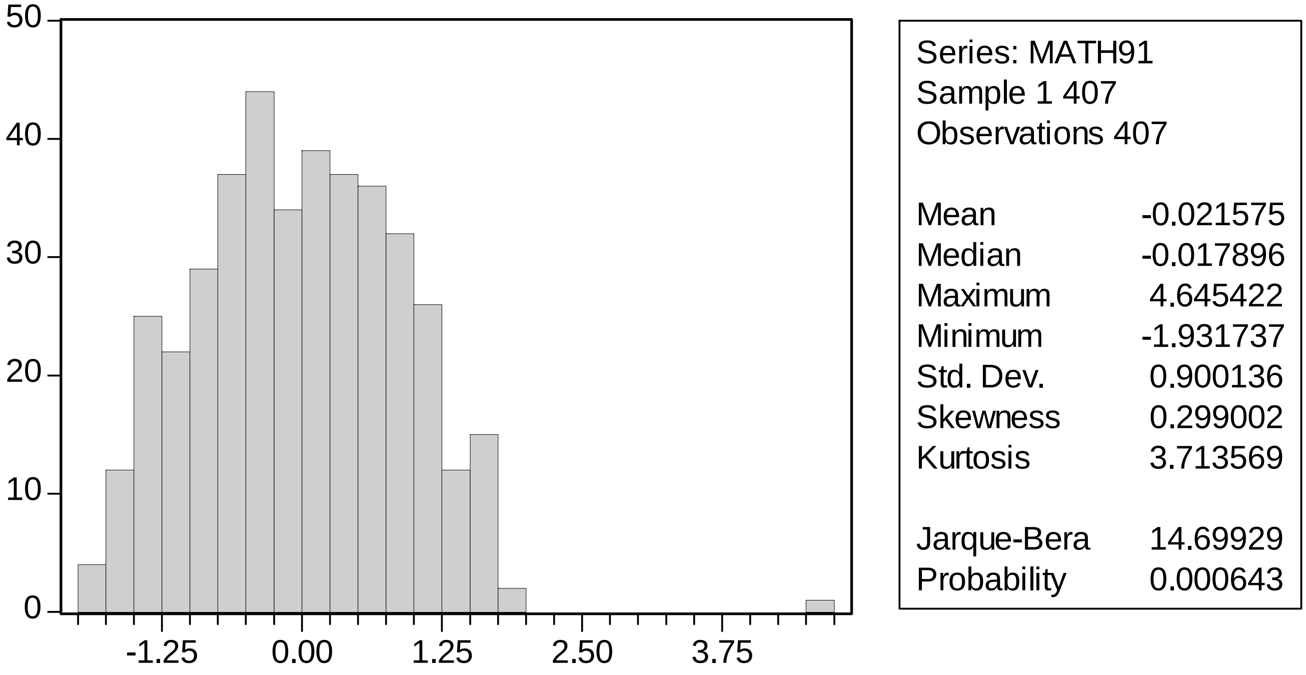
- You will probably run several regressions, many of which you won't include in your final paper. You need to be comfortable explaining and interpreting your results, graphs, and tables before you decide to how to best present them to the reader and audience.
- The regression output includes a lot of information that you don't use in interpreting your results. This information should be excluded before it is presented to the audience.
- Often, as with the tables below, you will need to "clean up" and format your regression output before including it in your paper and presentation.

Cross section data on math scores

**Graph #1**
To give the audience a sense of how the math scores are distributed, the researcher includes a histogram of the student's math scores for 1990. See the graphs on the following page.

What's wrong with the first graph (below):

- The graph at the top includes labels that the researcher may be familiar with (i.e., the names of the variables in the workfile like "MATH91), but that the reader and audience are not familiar with.
- The summary statistics on the right should not be included with the histogram – it repeats the information shown in the histogram.
- If the researcher wants to present summary statistics for the sample, then she should include a table with the summary statistics (mean, median, standard deviation, and perhaps the minimum and maximum) with all of the variables she intends to study. See Table #1 for an example of how to do this.
- This graph is missing a title – this is important to signal to the reader and audience what it is that they are supposed to be looking at.



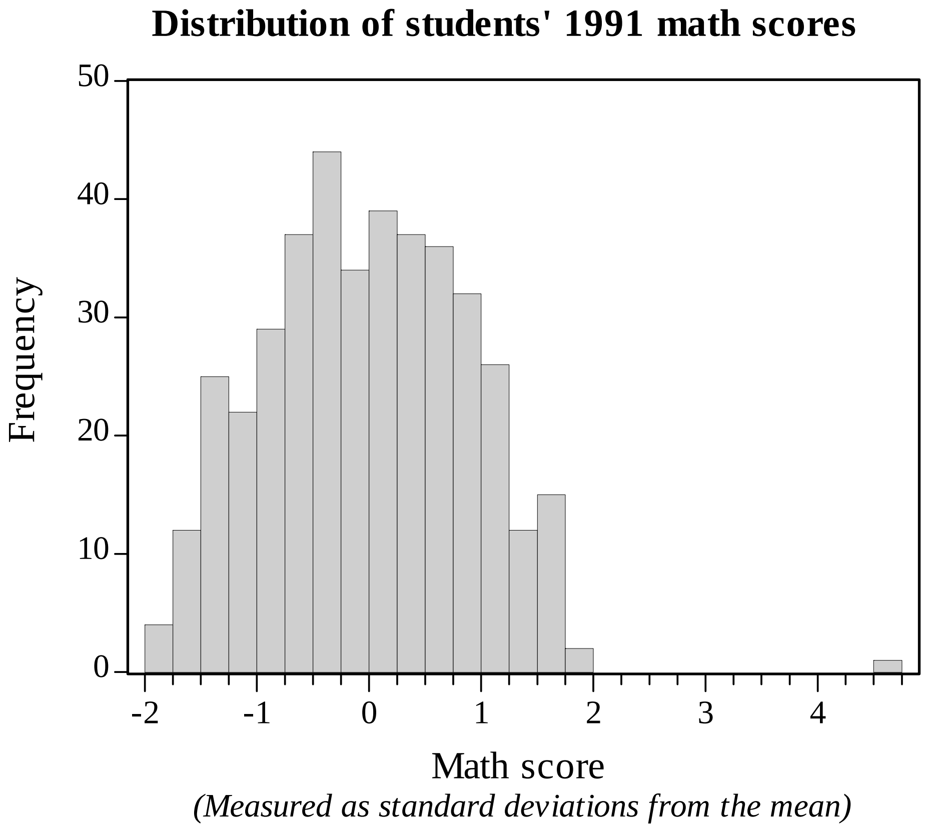
```
Series: MATH91
Sample 1 407
Observations 407

Mean          -0.021575
Median        -0.017896
Maximum        4.645422
Minimum       -1.931737
Std. Dev.      0.900136
Skewness       0.299002
Kurtosis       3.713569

Jarque-Bera    14.69929
Probability    0.000643
```

Why the second graph (following page, top) is better:

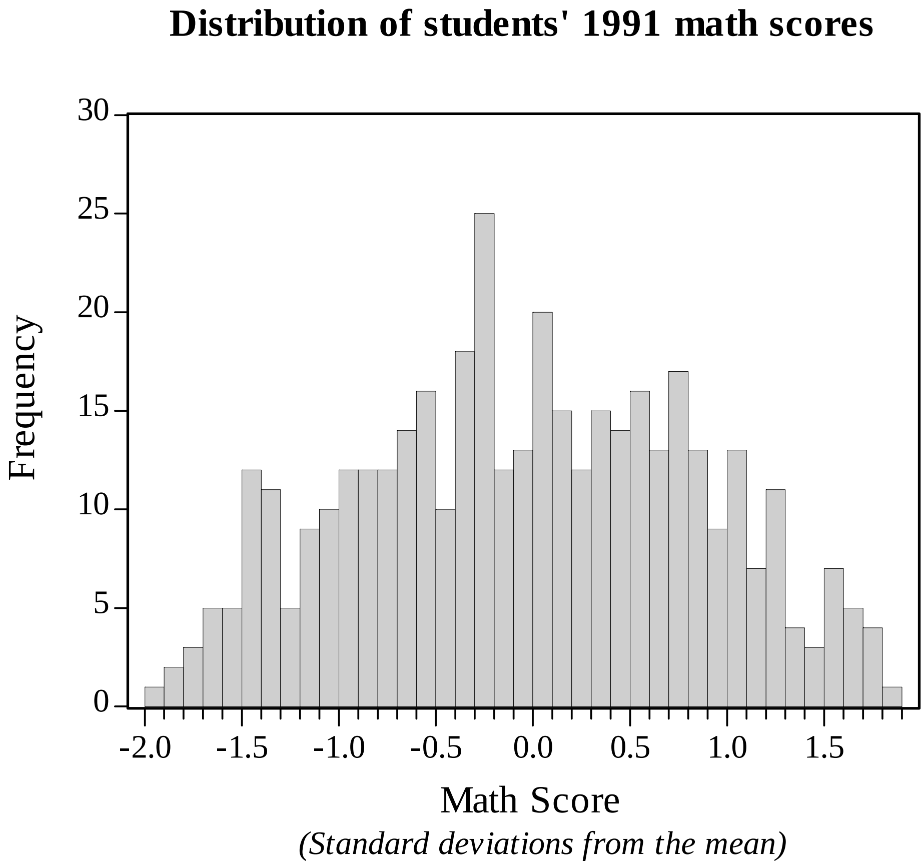
- It includes a title that clearly states what the variable is and how it is measured.
- It includes only information that is relevant for seeing the distribution of the scores (this is the information the researcher wanted to convey, see above).
- The font is the same font as the text of the paper, making it look more professional.



Notice that from the histogram, there is an "outlier" in the histogram above. One of the students in the sample had a dramatically higher score than the others (4.5 standard deviations from the mean is a dramatic difference). Based on this information, the researcher may want to exclude this student from the sample. Researchers often exclude outliers because they may skew regression results and applying these results to the general population (of students, in this case).

In EViews you can do this by using the "If" condition that appears when you click on "sample". This will allow you to limit the sample in some way (if you choose to). We did this in your earlier homework to look at differences in education and earnings among men and women.

The "adjusted" histogram is below. Notice that the distribution looks more like a normal distribution:



If you do make this type of adjustment with your data, you will explain it in the text, but you wouldn't present both histograms to make the point. The end product is what the readers and audience are interested in. It is important to document how you sorted your data, why you may exclude outliers in the text of your paper, but it isn't necessarily something you need to dwell on in your presentation.

**Table #1**

The researcher wants to present the summary statistics to the audience/reader after she has discussed how these variables are measured. She will then point out anything that is of note in her sample.

Date: 04/20/05
Time: 14:03
Sample: 1 407 IF MATH91<2.5

| | MATH91 | MATH87 | ENROLLMENT | DRUGS | SES | MALE | URBAN |
|---|---|---|---|---|---|---|---|
| Mean | -0.036853 | 0.038425 | 998.7934 | 0.155954 | -0.054176 | 0.521197 | 47.49906 |
| Median | -0.035086 | -0.131602 | 946.8889 | 0.154567 | -0.051499 | 1.000000 | 51.00000 |
| Maximum | 1.852015 | 4.644752 | 2461.695 | 0.372711 | 2.204239 | 1.000000 | 100.0000 |
| Minimum | -1.931737 | -1.465087 | 171.4286 | 0.000000 | -2.690696 | 0.000000 | 0.000000 |
| Std. Dev. | 0.872886 | 0.930617 | 528.7482 | 0.073913 | 0.879395 | 0.500175 | 45.81379 |
| Skewness | -0.002280 | 0.781764 | 0.546720 | 0.278625 | -0.136462 | -0.084864 | 0.046177 |
| Kurtosis | 2.204560 | 3.664189 | 2.577299 | 2.728794 | 2.800117 | 1.007202 | 1.128385 |
| | | | | | | | |
| Jarque-Bera | 10.57212 | 48.21636 | 22.96203 | 6.417358 | 1.912113 | 66.83420 | 58.67083 |
| Probability | 0.005062 | 0.000000 | 0.000010 | 0.040410 | 0.384406 | 0.000000 | 0.000000 |
| | | | | | | | |
| Sum | -14.77805 | 15.40844 | 400516.2 | 62.53768 | -21.72450 | 209.0000 | 19047.12 |
| Sum Sq. Dev. | 304.7717 | 346.4195 | 1.12E+08 | 2.185235 | 309.3341 | 100.0698 | 839561.2 |
| | | | | | | | |
| Observations | 401 | 401 | 401 | 401 | 401 | 401 | 401 |

What's wrong with the first table (above):

- This table has many of the same problems as the graph above. It includes labels that the researcher may be familiar with (i.e., the names of the variables in the workfile like "MATH91" and "SES"), but that the reader and audience are not familiar with.
- The table includes far too much information – the audience/reader would likely ignore it because it's difficule to look at for several reasons.
    - Most of this information the researcher won't refer to or use in her explanation of what her sample looks like. Many of the statistics above are ones that not only the reader/audience aren't familiar with, but that the researcher probably isn't familiar with either.
    - On a related note, it is usually not to your advantage to reveal when and at what time you made your table (this is included in the upper right) for your paper and presentation.
    - The data in the table above includes too many significant digits (numbers after the decimal). Most statistical packages do this by default. It is your job to decide how many significant digits are necessary to convey the point (usually a maximum of 3 and minimum of 1 or even 0).

- This table is missing a title – this is important to signal to the reader and audience what it is that they are supposed to be looking at.

**Summary statistics**

| | Mean | Std. Dev. |
|---|---:|---:|
| Math score in 1991 | -0.04 | 0.87 |
| Math score in 1987 | 0.04 | 0.93 |
| School enrollment | 998.79 | 528.75 |
| Drugs (index) | 0.16 | 0.07 |
| Socio-economic status | -0.05 | 0.88 |
| Male ( =1 if male) | 0.52 | 0.50 |
| Urban (%) | 47.50 | 45.81 |
| Number of observations | 401 | |

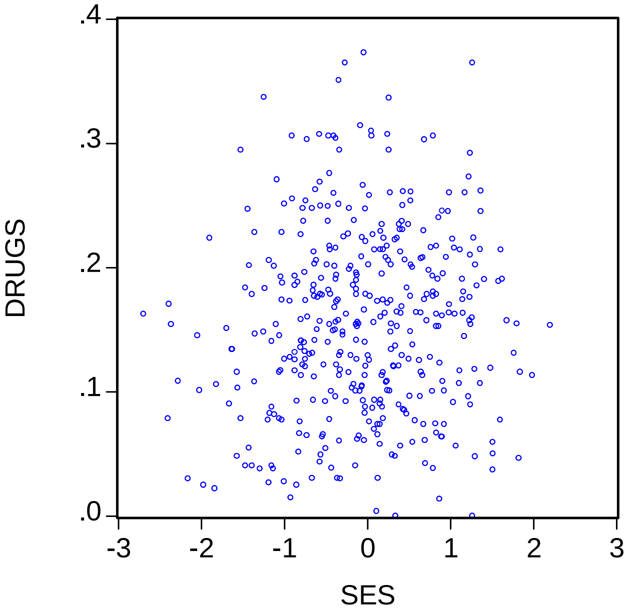
*Note: Math scores and socio-economic status are measured as standard deviations from the mean*

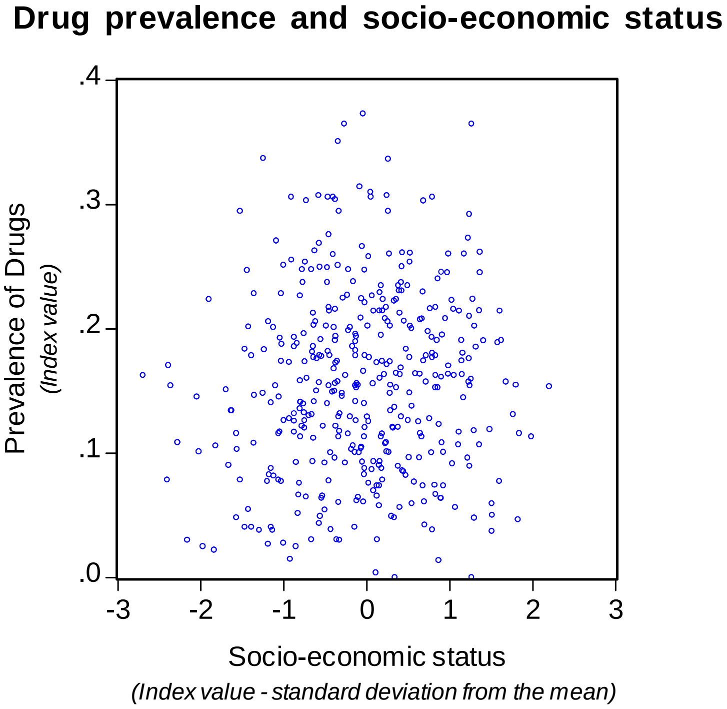
Why the second table (above) is better:

- It includes a title that clearly states what is included in the table. Also, the formatting with the lines makes it easier to look at.
- It includes a description of the data – even though the researcher has already defined each variable and how it is measured, it's useful for the audience to keep track of how these variables are measured.
- The font is the same font as the text of the paper, making it look more professional.
- The data in the table includes only two significant digits – this is all that's really needed in this case to make the point.
- Other things the researcher might have included (if she thought it necessary) on a summary statistics table would be the median, minimum and maximum, or any correlations of interest.
- You can make this type of table in Excel or in Word. I typically paste the table from EViews into Excel - it's easier to change the significant digits for all of the variables at the same time – then paste it into my research paper/presentation. Depending on how much information is in your table, it may be easier to retype the information into a table in Word.

**Graph #2**
The researcher has found from her regression that socio-economic status and math scores are positively correlated. In other words, that higher socio-economic status is associated with higher math scores. She suspects this is the case because students with lower socio-economic status tend to live and attend schools in areas with bigger drug problems.

How would we examine this? A scatterplot of socio-economic status and a measure of drug prevalence will reveal whether or not this is the case.





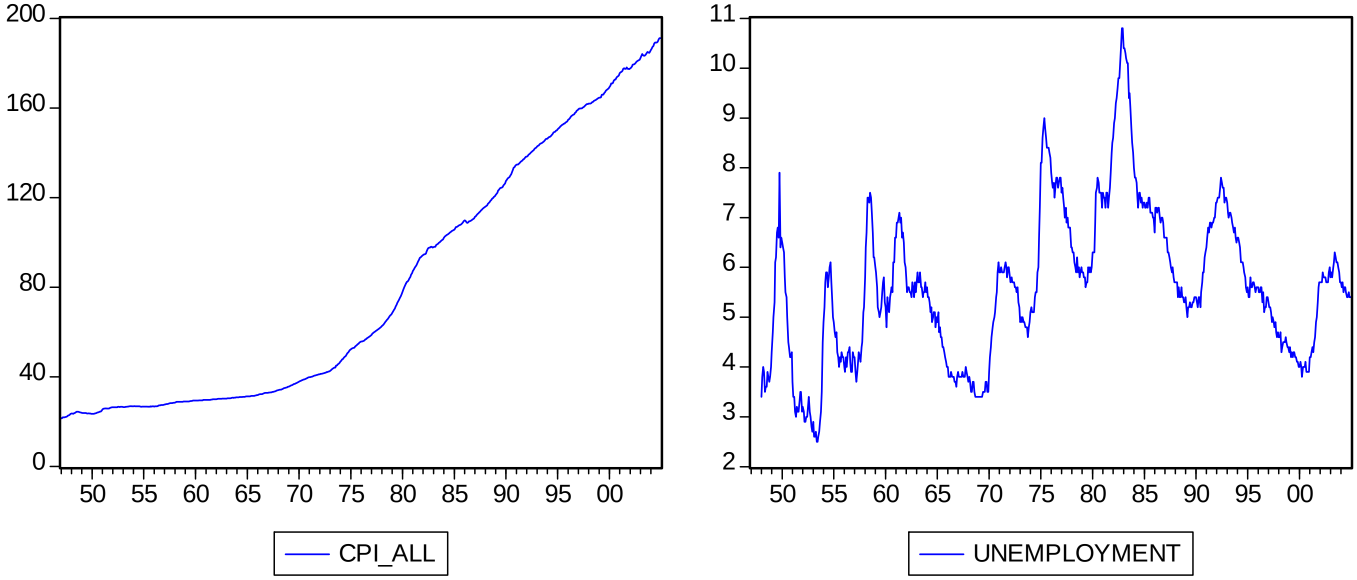
The graph on the left is not appropriately labeled. The audience can probably guess what drugs measures, but what is SES? The graph shows that the researcher's suspicion is not correct, there is no correlation between drug prevalence and socio-economic status among the students in this sample. The researcher could have included a regression line on the scatter plot is she wanted to, but it is pretty obvious that there is no relationship from the data points.

## Time series data on inflation and unemployment

The example presented in this section will be abbreviated, so please read through the cross section example, as many of those pointers will apply here as well.

**Graph #1**

Time series plot of the variables the researcher is interested in. Here, the researcher wants to show the audience how prices and unemployment fluctuate over his sample period.



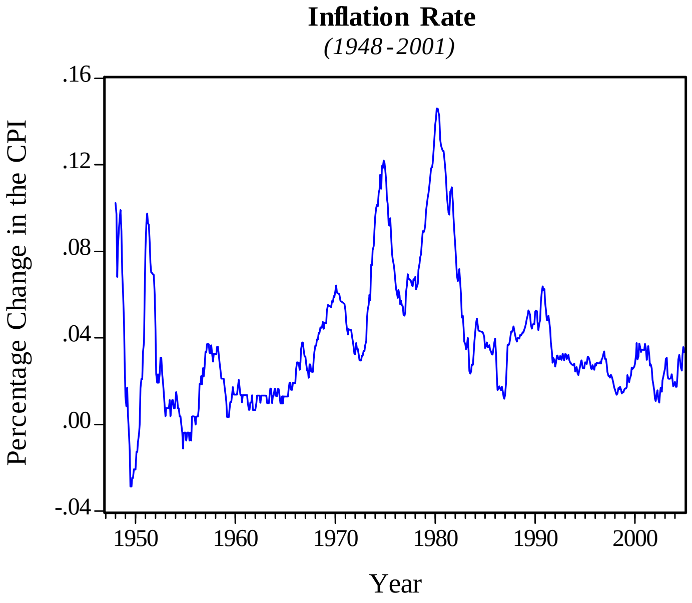
What's wrong with the first graph (above):

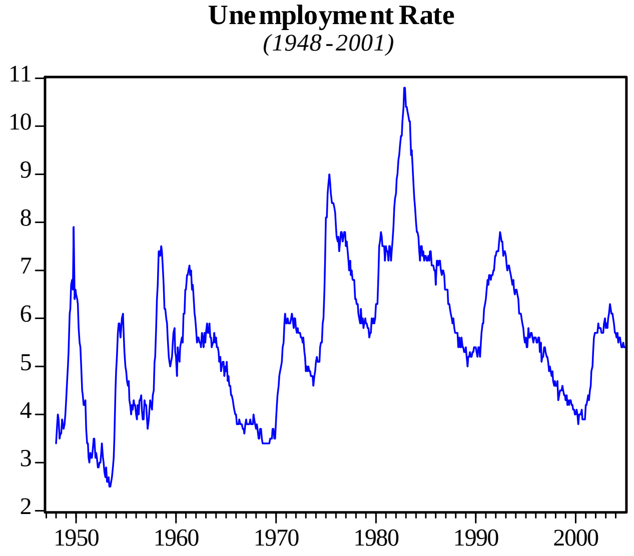
- The axes are not clearly labeled and the graphs don't have titles indicating the sample period.
- The legends that appear at the bottom are not useful because there is only one series plotted on each graph (assuming one included a title at the top of each graph).
- The years labeled at the bottom appear with two digits (55 instead of 1955). Note, this is sort of tricky to fix with the current version of EViews. I'll work with you one on one if you want to see how to fix this.
- The CPI is going up over time ("trending" upward). This isn't very interesting – we're more interested in the inflation rate (the annualized growth in the CPI).

Why the second graph is better (next page, top)

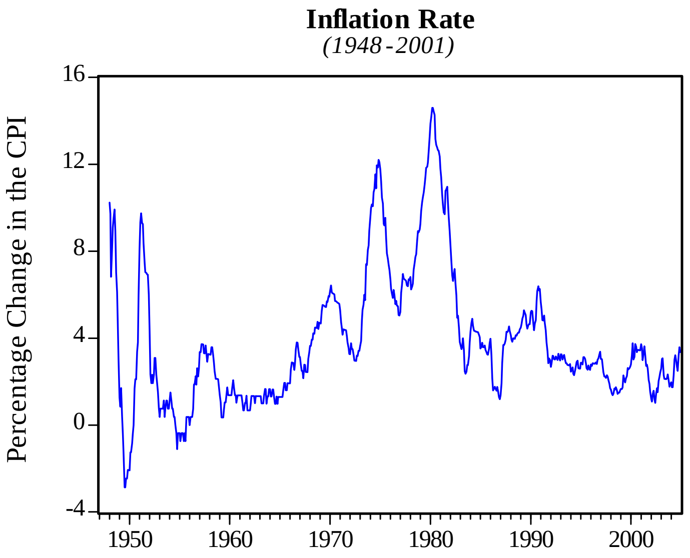
- The axes are clearly labeled, the time period indicated, and the graphs have titles indicating which data is included.
- This graph includes the inflation rate, rather than the CPI.

Note, since time series data often "trend" upward (as the economy grows), it is common to transform time series data into a growth rate (GDP growth as opposed to the dollar amount of GDP). Usually, these growth rates are annualized. If you are not transforming the time series data, it may make sense to include a "trend" in your regression to account for these effects. This can included in your regression by adding "@trend" as one of your explanatory variables. A trend simply takes the value of 1 in the first period of the sample, 2 in the second, 3 in the third, and so on.


**Inflation Rate**
*(1948-2001)*


**Unemployment Rate**
*(1948-2001)*

One problem with the graph above is, while the inflation rate and unemployment rate are measured in percentages, the CPI graph is reported in decimals, while the unemployment rate is already in percentage terms. To correct this, create another variable that multiplies the inflation rate by 100, so that the scale matches that of the unemployment rate:


**Inflation Rate**
*(1948-2001)*

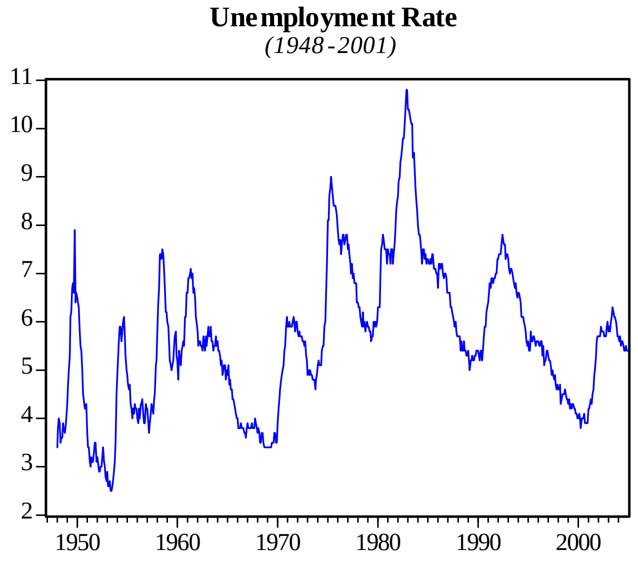

**Unemployment Rate**
*(1948-2001)*

**Table #1**

The researcher wants to present some summary statistics of the inflation and unemployment rate data. Looking ahead to what his regression results will eventually reveal, there will be a change in the relationship between these two variables after 1968. This will mean that he will need to create three tables that include summary statistics for inflation and unemployment: full sample, 1948-1968, and 1969-2001.

Date: 04/20/05
Time: 15:12
Sample: 1947M01 2004M12

| | INFLATION10 | UNEMPLOYM |
|---|---|---|
| | 0 | ENT |
| Mean | 3.862189 | 5.635965 |
| Median | 3.133663 | 5.600000 |
| Maximum | 14.59227 | 10.80000 |
| Minimum | -2.868852 | 2.500000 |
| Std. Dev. | 3.088384 | 1.528574 |
| Skewness | 1.178761 | 0.522584 |
| Kurtosis | 4.350441 | 3.379614 |

| | INFLATION10 | UNEMPLOYM |
|---|---|---|
| | 0 | ENT |
| INFLATION10<br>0 | 1.000000 | 0.215492 |
| UNEMPLOYM<br>ENT | 0.215492 | 1.000000 |

Date: 04/20/05
Time: 15:13
Sample: 1947M01 1968M12

| | INFLATION10 | UNEMPLOYM |
|---|---|---|
| | 0 | ENT |
| Mean | 2.161708 | 4.724603 |
| Median | 1.564315 | 4.550000 |
| Maximum | 10.23256 | 7.900000 |
| Minimum | -2.868852 | 2.500000 |
| Std. Dev. | 2.385814 | 1.206733 |
| Skewness | 1.324983 | 0.304221 |
| Kurtosis | 5.340331 | 2.285711 |

| | INFLATION10 | UNEMPLOYM |
|---|---|---|
| | 0 | ENT |
| INFLATION10<br>0 | 1.000000 | -0.479505 |
| UNEMPLOYM<br>ENT | -0.479505 | 1.000000 |

Date: 04/20/05
Time: 15:13
Sample: 1969M01 2001M12

| | INFLATION10 | UNEMPLOYM |
|---|---|---|
| | 0 | ENT |
| Mean | 5.097287 | 6.204040 |
| Median | 4.170337 | 5.900000 |
| Maximum | 14.59227 | 10.80000 |
| Minimum | 1.187215 | 3.400000 |
| Std. Dev. | 3.033524 | 1.501656 |
| Skewness | 1.249041 | 0.552820 |
| Kurtosis | 3.828034 | 3.204457 |

| | INFLATION10 | UNEMPLOYM |
|---|---|---|
| | 0 | ENT |
| INFLATION10<br>0 | 1.000000 | 0.212605 |
| UNEMPLOYM<br>ENT | 0.212605 | 1.000000 |

Note: in the tables above, the correlation matrix is to the right of each table for each of the three samples. Also, I cutoff some of the information that EViews includes in summary statistics by default (to save some space).

What's wrong with the first tables (above):

- This table has many of the same problems as the table in the cross section example above. It includes labels that the researcher may be familiar with (i.e., the names of the variables in the workfile like "INFLATION100"), but that the reader and audience are not familiar with.
- It includes a lot of information that is not relevant for the analysis, making it difficult to sort out what is important (presumably the mean, standard deviation, and correlation between inflation and unemployment).
- These tables are poorly organized. The information should be presented in one easy-to-read table for the reader/audience. Even though the labels are included, they are often cut off.

**Summary Statistics**

| | Inflation | Unemployment |
|---|---|---|
| *Full sample* | | |
| Mean | 3.8 | 5.6 |
| Std. Dev. | 3.1 | 1.5 |
| Correlation | | 0.22 |
| *1948-1968* | | |
| Mean | 2.2 | 4.7 |
| Std. Dev. | 2.4 | 1.2 |
| Correlation | | -0.48 |
| *1969-2001* | | |
| Mean | 5.1 | 6.2 |
| Std. Dev. | 3.0 | 1.5 |
| Correlation | | 0.21 |

Why the second table is better (above)

- The table synthesizes the information the researcher was looking for in one table. It reports the sample period, which statistics are included for which variables. From the table above, it is easy to see that average inflation and unemployment were higher, and more variable, in the later period. Also, we can see that the correlation is different across the two periods.

Note, there are several ways to format a table and to label graphs. The important thing is that the table and graphs are clear and easy to the read for your audience.