

**Homework policies:** You should give a brief and concise explanation for each question. Just writing down an answer (e.g., “SD = 3.3”) with no explanation is not sufficient.

Homework is due in the homework boxes by 11:00 AM on the due date. Late homework will not be accepted. The homework drop-off boxes are on the third floor of the Science Center, outside room 300.

Collaboration: We encourage you to discuss homework problems with other students (and with the instructor and the TAs of course), but you must write your final answer, in your own words. Solutions prepared “in committee” or by copying someone else’s paper are not acceptable.

The lowest homework grade will be dropped at the end of the semester when computing final grades.

The solutions to the odd numbered problems in IPS can be found at the back of the book.

1. Like all software, your skills using Stata will benefit from some practice, and this homework problem serves as a mini-tutorial to help you get started. The problem uses a data set discussed briefly in class, but expanded slightly to add a variable; you will need to download from the course web site. The steps required in Stata to manipulate this data set are outlined fairly specifically, and will not be repeated in such great detail in later problem sets, so you may find it useful to print and save this problem set with your notes. We will also make available a longer tutorial for getting started in Stata. **The information in problem 1 that should be handed in are indicated in bold face font.**

Some of the data come from the web site for the Statistical Abstract of the United States (<http://www.census.gov/compendia/statab/>) and gives the value of the per capita expenditure, for each state and for the District of Columbia, on the criminal justice system. Expenditures are given for the police, the judicial system, correctional facilities (essentially, prisons). The last column gives the violent crime rate (‘violent\_crime’, number of violent crimes per 10,000 members of the population aged 12 and over). Violent crime rate is typically measured using the National Crime Victimization Survey, not police records. **Explain in your pset solutions why police records are not the primary data source for this information.** More details on trends in violent crime rates can be obtained at the web site for the Bureau of Justice Statistics (<http://www.ojp.gov/bjs/glance/viort.htm>).

The data set is on the course web site in the same folder as this problem set as an Excel file (*criminal\_justice.xls*), as a plain text file with values separated by commas (*criminal\_justice.csv*), and as a Stata formatted file (*criminal\_justice.dta*). The files all contain identical information, so you need use only one version of the file.

This problem description contains more details about using Stata for a problem set because the only real barrier is in getting started, and that barrier is very low. Once you get started, you will soon learn how to navigate the Stata menus easily. Please note that all instructions, suggestions below are PC-centric. I do not own a Mac and cannot test these steps, but the steps should be similar.

The .csv and the .dta files can be read directly into Stata (instructions below). You only need to use one of these files, but you might benefit from loading each version (in separate Stata sessions), for practice. We have provided the Excel version of the file if you wish to view the file in Excel. To give you some idea of the structure of the Excel file, the first 7 rows of the file are shown below. The first row is a 'heading row' that gives variable names. The 'comma separated value file' looks a bit different, but the first row also contains the variable names. Here are the first few rows of the Excel file:

State	Police	Judicial	Corrections	violent_crime
Alabama	159	71	103	486
Alaska	412	204	273	567
Arizona	231	120	206	532
Arkansas	149	72	137	445
California	290	201	257	622
Colorado	238	86	190	334

Before you begin, you may want to open the Stata Tutorial on the course web site, in the box labeled Helpful Hints. You may find it more effective to read the file as necessary, rather than reading it all at once before you get started.

Use the following steps to load the Stata formatted file on an IBM compatible PC (the Mac version should work the same way except for differences in the operating system):

- Go to the course web site and download the file 'criminal\_justice.dta'. It will be in the Problem Sets topic box, under pset 1. You can store it anywhere, but for simplicity you might want to store it on your desktop this time.
- Start Stata (see the 'Getting Started with Stata' document on the web site under Helpful Hints for a description of how to download Stata from the FAS computing facility and use the key server).
- From the Stata 'File' menu, choose 'Open', find the file 'criminal\_justice.dta' on your disk and click 'open'.
- The file is now loaded in Stata and ready for use.

Later in the course, you will find it useful to be able to load csv files into Stata, since Excel can be used to easily produce csv files from Excel spreadsheets, and much of the data available on the internet can be saved in Excel spreadsheets. Also, we cannot always be sure that data save in .dta format on a PC can be read on a Mac. The following steps will always work.

Use the following steps to load a csv file produced in Excel into Stata:

- Open Stata. Please see the handout 'Getting started with DataDesk' for tips on how to run data desk on the web site under Helpful Hints.
- From the Stata 'File' menu, choose 'Import', choose the option 'ASCII Data Created by a Spreadsheet', click 'Browse' in the next dialog box, find the file 'criminal\_justice.csv' on your disk, and click 'open'. Stata should read the file, load it, and assign variable names.

If you have used Excel before, you may want to load the file `criminal_justice.xls` into Excel, save it in csv format, and load it into Stata using the steps above. This skill will come in handy later in the course. If you have not used Excel before, you may want to skip this step for now, since it is not needed for any of the problems below.

Now that the file is in Stata, calculate some summary statistics and make some plots. We explain how to do the problem using Stata's menus. The document 'Getting Started with Stata' provides the details on how to manipulate a similar dataset using commands directly entered into the Stata command window.

- a) Calculate means, medians and standard deviations, and interquartile ranges for the variable 'judicial' by first clicking 'Statistics' menu, choosing 'Summaries, Tables, and Test,' then 'Summaries and Descriptive Statistics', then (finally) 'Display Additional Statistics' in the dialog box that opens. Type the name of the variable 'judicial' in the box for the variable name (or click on it from the drop down menu activated with the arrow at the right side of the box), then click 'Submit'. Summary statistics for the variable will be displayed on the screen. **Copy the display and paste it into a Word document to hand in, making sure to indicate which items on the output provide the numbers asked for in this question.** You may find that it is easier to move the table of summary statistics produced by Stata into Word by using the 'copy as picture' option when you right click on the output after highlighting it.

The depth of the menus in Stata is a bit of an annoyance, but the package has many options that are used in environments like brokerage houses and medical research facilities, and the menus have been built to accommodate these options.

Note that when you select a command from a menu, Stata prints the command in its results window before the output from the command. You also have the option to save the commands and the output in a log file which you can read about in the getting started notes. Note also that once the data is loaded in Stata, you can view the data in the Data Editor available under the 'Data' menu.

- b) **Do the summary statistics indicate whether or not there are outliers in the data set for the variable 'judicial'? If so, how large (or small) do observations have to be to be labeled outliers?** Use the data editor to sort on the column 'judicial' and identify the states (including possibly the District of Columbia) with 'outlying values', then **write those states on the paper you will submit.**
- c) Plot histogram of the variable 'judicial' by selecting 'Histogram' from the 'Graphics' menu, then filling in the variable name in the dialog box. For the first plot, check the button 'frequency' in the section for the Y-axis in the lower right corner of the dialog box, and otherwise accept the defaults from Stata (do not change anything); draw a second version increasing the number of 'bins' from the default (7 in this case, I believe) to 15. **Include both plots in your homework and describe the main features of the plots (e.g., symmetry vs skewness).**

- What is visible in the second plot that is hidden in the first plot?. Put your plots in the same MS Word document as above, along with your description of the plots.** (To copy the graph to your Word or other document, use the `Edit` menu at the top of the box containing the plot.)
- d) Plot the histogram (15 bins again) with the outliers identified above eliminated. You can do this by using the if/in dialog box in the histogram plot dialog, and inserting an `if expression` of the form `judicial < x`, where x is the value that defines outliers. (This hint suggests that if there are outliers, they are on the large end of the scale...). **Include this histogram in your submitted solutions.**
  - e) Plot a boxplot for judicial with and without the outliers, using the menus and dialogs under Graphics, boxplot. Did eliminating the large values of `judicial` eliminate outliers in the reduced data set? Why or why not?
  - f) Using Stata, calculate the mean of the variable `violent\_crime`. Now recalculate that mean, but this time do the calculation each of in two groups separately: the first group should be defined by the set of states (including DC) for which the amount spent on the judicial system is lower than the median of the values for 51 states, and the second for the group defined by the set of states for which the amount of money spent on the judicial system is at least as large as the median spent by the 51 states. The mean and violent crime rates will differ between the groups. Why do you think that is the case?
  - g) If you used the csv version of the data set, you can save it in a Stata format (dta file) using the `Save as...` option under the file menu. Any changes you have made to the file will be saved, and the file can be re-opened later using the steps given earlier to open a Stata formatted data set.
2. A dentist's office decides to investigate the amount of sick leave taken by its employees. The number of days of sick leave taken by its 6 employees last year was 1, 0, 3, 5, 19, & 8.
- a. Calculate the mean and standard deviation for these data, showing each step in detail. Find the mean and then find each of the deviations ( $x_i - \bar{x}$ ) and their squares. Check that the deviations sum to zero. Calculate the variance as an average of the squared deviations (divide by n-1). Obtain s, the standard deviation, as the square root of the variance.
  - b. Prepare a modified boxplot for these data by hand. Are any of the employees outliers according to the 1.5 x IQR rule?
3. A social scientist has nearly completed a study administering a standardized IQ exam to a sample of 12-year old children with histories of asthma. In a sample of 19 children, the average score was 100 with standard deviation 10. Suppose an additional score comes in from a child not present earlier, with value 130. What will be the revised average score?

4. Each of the following lists has an average of 50. For each one, state whether the standard deviation is approximately 1, 2 or 10 without doing any arithmetic. Give reasons for your answers, even though you cannot find the new standard deviation precisely with the given information. You would need to know all of the exact measurements."

- a. 49, 51, 49, 51, 49, 51, 49, 51, 49, 51
- b. 48, 52, 48, 52, 48, 52, 48, 52, 48, 52
- c. 48, 51, 49, 52, 47, 52, 46, 51, 53, 51
- d. 54, 49, 46, 49, 51, 53, 50, 50, 49, 49
- e. 60, 36, 31, 50, 48, 50, 54, 56, 62, 53

5. You roll a standard six-sided die three times and then calculate the standard deviation (SD) for these rolls. What is the largest possible SD? What is the smallest possible SD? Explain briefly.

6. A less-than-benevolent CEO is under pressure from a union to increase the average worker's salaries. However, the CEO does not want to spend any more money for salaries. Instead she develops a plan to fire 5 employees and thereby raise the average salaries. What strategy could she use to select the 5 employees to be fired?

7. IPS 1.13

8. IPS 1.39. Use Stata for this problem. The data for the problem is on course web site in the files ips\_ex\_1.39.csv and ips\_ex\_1.39.dta.

9. IPS 1.51