HAND IN SEPARATELY: (I) PROBLEMS 1–6 (II) WRITING ASSIGNMENT

0. Install Data Desk (0 points)
Data Desk runs on both Mac and Windows. Install the Data Desk program by copying the program folder from the data-software server to your hard drive. The first time you run the program, you’ll be asked for a serial number and access code; these are given in the ReadMe file that accompanies the program. (For more information on accessing the data-software server, see http://www.swarthmore.edu/its/students/stu_software.htm. You’ll need a fileserver password; see the “Software Download” section on the web page.)

Please note: since the College has a limited number of licenses for Data Desk (enough for 20 simultaneous users), get started early to make sure you can access the program!

1. Blood Pressure (5 points)
Several years ago, I worked on a study on women with polycystic ovarian syndrome (PCOS), an endocrine disease related to diabetes (citation: Legro, Bentley-Lewis, Driscoll, Wang, and Dunaiif (2002): J. Clinical Endocrinology and Metabolism 87:5). As part of the study, 371 women with PCOS were selected. Do women with this condition tend to differ in blood pressure from the general population?

The 371 women had a mean systolic blood pressure of 121.07, with an SD of 16.59. Let μ denote the mean blood pressure of the population from which these 371 subjects were selected. (Assume they were selected via a simple random sample.)

(a) A “normal” systolic blood pressure is considered to be 120. Carry out a hypothesis test of the following hypotheses using an alpha level of .05:

\[ H_0: \mu = 120 \quad \text{vs.} \quad H_a: \mu \neq 120 \]

Be sure to label the test statistic and the p-value. Do these control subjects significantly differ from a “normal” population in systolic blood pressure?

(b) If the sample size had been \( n = 3710 \) instead of 371, how would your conclusions change? How does this illustrate the concept of statistical significance as distinct from clinical importance — that is, whether the difference in blood pressure has important consequences in a medical setting?

(c) Calculate a 95% confidence interval for \( \mu \). Does this interval contain 120? Is your confidence interval consistent with your conclusion in part (a)?

2. Alpha Levels (4 points)

(a) For some value \( k \), you test the null hypothesis \( H_0: \mu = k \) against the alternative hypothesis \( H_a: \mu \neq k \). Suppose the sample data is such that you reject \( H_0 \) using an alpha level of .01. If you had tested the same hypotheses using an alpha level of .05, would you necessarily have rejected \( H_0 \)? Explain briefly.

(b) For some value \( k \), you test the null hypothesis \( H_0: \mu = k \) against the alternative hypothesis \( H_a: \mu \neq k \). Suppose the sample data is such that you reject \( H_0 \) using an alpha level of .05. If you had tested the same hypotheses using an alpha level of .01, would you necessarily have rejected \( H_0 \)? Explain briefly.

(over)
3. P-VALUES (4 POINTS)
Suppose a researcher, Pat, tests two new drugs to see if either significantly improves cholesterol level in patients. For each test, the null hypothesis is that the drug has no effect on cholesterol levels. Pat performs a hypothesis test for the first drug and finds that the observed p-value is 0.04. Pat then performs a hypothesis test for the second drug and finds that the observed p-value is 0.001. Pat concludes that the second drug will have a larger effect on the cholesterol level of individual patients. Is this a valid way to interpret the two p-values? Explain briefly.

4. STATISTICS AND PARAMETERS (4 POINTS)
A statistic is a quantity that is calculated from a sample.
(a) Give an example of a statistic that estimates an unknown population parameter.
(b) Give an example of a statistic that does not estimate any corresponding population parameter. (There are numerous examples, several of which you should have seen in your introductory statistics class.)

5. NORMAL PROBABILITY PLOTS (6 POINTS)
(a) Use Data Desk to generate nine random samples each having 10 observations from a standard Normal distribution. (Using the command Manip ▶ Generate Random Numbers, select the Normal distribution and enter µ = 0 and sigma = 1; at the top of the window, enter 9 variables with 10 cases. Important: If you are running Data Desk on an Intel Mac, select “ULTRA” under the “Generator” menu at the bottom of the window.) Make a histogram of each random sample. (To do this, select all nine variable icons and use the command Plot ▶ Histograms. Do the samples appear to be normally distributed? (You don’t need to print anything out.)
(b) Now make a normal probability plot (NPP) of each of your nine random samples. (To do this, select all nine variable icons and use the command Plot ▶ Normal Prob Plot. Do the samples appear to be normally distributed? Did you find it easier to assess the normality of your samples using histograms or using NPPs? Explain briefly. (You don’t need to print anything out.)
(c) Repeat parts (a) and (b) using nine samples of 100 observations each. How does the appearance of the histograms and NPPs change with the increased sample size?
(d) You can think of an entire NPP as being a statistic, since it is calculated from a sample. As such, it must have a sampling distribution, like any other statistic. Make a rough sketch (by hand) of the range in which you think 95% of NPPs would fall if sampled under these conditions, for both n = 10 and n = 100. Hand in these two pictures. (Think carefully about the shape of these ranges.)

6. Q-Q PLOTS (5 POINTS)
(a) The following are quantile-quantile plots of GRE General Test Verbal scores for students intending graduate study in physical education, linguistics, genetics, and economics. Briefly describe the pattern in each of the q-q plots.

(b) Suppose we have the SAT verbal and quantitative scores for all current Swarthmore students. How do the following two plots differ: (1) a q-q plot of the verbal and quantitative score distributions, and (2) a scatterplot plotting the verbal and quantitative scores for each student? What question is answered by each plot?
Writing Assignment (hand in separately)

1. Supreme Court Justices (5 points)

2. The Good, the Bad, and the Ugly (15 points)
Find two graphics from any source: the web, magazines, newspapers, books, etc. One of these should be an example of what you think is a (generally) good graphic. The other should be an example of what you think is a (generally) bad graphic. (Clearly label which is which.) For each graphic, briefly describe its strengths and weaknesses: What’s good about this graphic? What’s bad about it?

For the bad graphic, redesign it to make it better, so that its meaning is clearer or more appropriate. For the good graphic, redesign it to make it worse, so that its meaning is less clear or less appropriate. (Don’t make the graphic incorrect; just redesign it so that the features of the data it shows are not as evident or not as clearly displayed.)

Briefly describe how your redesigned graphics achieve these goals. You may use Data Desk or any other computer program (e.g., Excel, Photoshop, etc.), or you may sketch the redesigned graphics by hand. If you do not have access to the actual data, sketch a reasonable approximation of it in your redesign.

You should hand in a copy of the two original graphics, your two redesigns, and your comments. One paragraph of comments for each example should suffice.

Try to find a graphic from a magazine, newspaper, book, or journal article, either in print or online. Don’t just google “bad graph” or “good graph”, unless you get really desperate and absolutely can’t find anything else. There are several web sites devoted to good and bad graphs, most of which I’ve seen already. I’d rather you pick your own example, rather than using an example that someone else has already chosen.