

Math 218 - Minitab

Homework #2

The goal of this assignment is to learn about statistical inference, and how a collection of independent trials contains more information than a single trial. We will also see how the power to distinguish between two alternatives grows as the sample size increases.

For this assignment, *you will have to hand in two items, a printout of your session window, and a plot.* Both these items need to be annotated as indicated below. Be sure to include as a comment in your session window (*be sure to enter all you comments after a # sign*) your name and student I.D. number. It would be helpful if you were to read the entire assignment and make sure you understand its intent and the procedures you need to follow before starting.

a) In Minitab, go to *Calc/Set Base* and issue the command to set the base of the random number generator to the last 4 digits of your social security number. This way, the results you obtain can be regenerated in the same way. In C1, using the command *calc->random data->normal*, generate 100 rows of standard normals. Then, in C2, generate 100 rows of normals with mean 0, standard deviation 1.5.

Here is some motivation. Suppose we did not know which column was which and on the basis of some data we had to make a guess. We will consider two procedures for determining how to tell which column the numbers come from. The first procedure is applied only on a small sample, a single pair of numbers. The second procedure can be applied if we are given more complete data.

Procedure A: If you are handed only a pair of numbers, and were told that one came from column 1 and one from column 2, without knowing which is which, a sensible procedure (let's call it Procedure A) would be to guess that the number with the smaller absolute value came from column 1, since both columns come from distributions symmetric about zero, but column 1 is "less spread out" than column 2, i.e. it has a smaller standard deviation. Naturally, Procedure A will guess correctly only some fraction p of the time, and we would like to know that fraction. We will

have MINITAB apply Procedure A 100 times to estimate the probability (or fraction of the time) that Procedure A gives the correct answer.

b) Using **Calc/Calculator**, fill column C3 with the ratio of the absolute value of C2 over C1 using the expression **ABS(C2/C1)**. Now code the data from C3 into column C4 using **Manip>Code>Numeric to Numeric** to be a 1 if the ratio in C3 is greater than 1, and a 0 otherwise. Notice that, row by row, C4 has a 1 if Procedure A gives the correct answer for the pair of numbers in that row, and 0 if the procedure gives the incorrect answer. Therefore, the C4 column sum, **calc->column Statistics->sum**, simply counts the number of times procedure A gives the correct answer. Actually, note that the column sum for C4 has the binomial distribution $B(100,p)$ with p the value we want to estimate. Here, $p=P(|Z| < |Y|)$ where Z and Y are independent, normal, each mean 0, but $SD(Z)=1$ and $SD(Y)=1.5$. Use the column sum to find the fraction of the time that Procedure A correctly identified which number came from which column. Enter directly into your session window your numerical estimate of p by typing it in as a comment after a # sign, for example, "My estimate of p is 0.37; that is, Procedure A worked on my data set 37% of the time."

c) Procedure B: Suppose that instead of being handed a pair of numbers, one from each of column 1 and column 2, you were handed the entire pair of columns, but without knowing which column is which. One simple way to guess would be to look at the sample standard deviation S for each column. You can compute the sample standard deviation of a column through **Calc/Column Statistics**. Compute the sample standard deviation for C1 and C2 so that these values appear in your session window (simply leave blank the box "Store result in:"). You would expect that S for C1 will be close to 1, and that S for C2 will be close to 1.5. Hence, given a pair of columns, not knowing which is C1 and which is C2, you would guess that whichever had the smaller sample standard deviation is C1. Lets call this method of guessing "Procedure B." Apply Procedure B to the data you generated in columns 1 and 2. Does Procedure B give the correct answer?

d) Use **Graph/Plot** to Plot C1 versus C2. Notice that since Minitab automatically scales the x and y axes to fit the data, the shape of the plot gives you no information about which column has which distribution. What you have to focus on, instead of the shape, are the values on the coordinate axes. You should be able to tell easily from the graph, by the coordinates, which column comes from the smaller standard deviation. Print out the plot and indicate by hand on it how you can tell from this graph which column is which.

e) In columns C3-C6 generate another 100 rows of normal mean zero, standard deviation 1 variables, and in columns C7-C10, 100 rows of normal mean zero, standard deviation 1.5 variables. Now replicate the experiment in c) 4 more times on

this new data, determining whether or not Procedure B can distinguish C3 from C7, C4 from C8, C5 from C9, and C6 from C10. Procedure B has now been applied 4 times in this part and one time in part c). Of the five times that Procedure B has been applied, what fraction of the time did Procedure B give the correct answer? Which procedure, A or B, seems to work better. Type your answers directly onto your session window as a comment of the form, for example, "Procedure B worked 4 times out of 5, that is, 80% of the time, and therefore seems to work better than Procedure A, which worked only 37% of the time."

f) Suppose that Procedure B worked 5 times out of 5. Does this indicate that Procedure B will always give the correct answer? Why or why not? Write your reasoning by hand on the printout of your session window.
