Homework #1 – Due Wednesday, Sep. 10

STAT-UB.0103 – Statistics for Business Control and Regression Models

Homework assignments will be drawn from problems from the text, as well as from other sources. Some of these will involve data files. All data files will be available on the course homepage `http://ptrckprry.com/course/ub0103/`. You are permitted to work in teams, but each student must independently write up their own solutions (no direct copying).

Assignments must be turned in at the end of class or given to the TA before his his office hours end at 2:00 pm. Electronic submissions are not accepted. Print out and turn in your Minitab plots.

Problem 1

The author Shere Hite undertook a study of women's attitudes toward love and sex by distributing 100,000 questionnaires through women's groups. Only 4.5% of the questionnaires were returned. Based on this sample of women, Hite wrote *Women and Love*, a best-selling book claiming that women are fed up with men. For example, 91% of the divorced women in the sample said that they had initiated the divorce and 70% of the married women said that they had committed adultery. Explain briefly why Hite's sampling method is nearly certain to produce a strong bias. Explicitly state what the population and the sample are. Hint: Think about the types of errors in surveys we considered in class.

. . . . . .

Problem 2

Consider this set of eight values:

250 206 235 211 261 208 174 214

Find the mean, median, and standard deviation for this set.

. . . . . .

Problem 3

Consider these values:

11 17 18 10 22 23 15 17 14 13 10 12 18 18 11 14

Find the mean, median, and mode for these.

. . . . . .

Problem 4

A set of $n = 80$ values has average 14,880.16. After all the work is completed, you discover that a value originally recorded as 12,148 should have been 11,248. If you replace the value 12,148 by 11,248, what will be the corrected average?

. . . . . .

Problem 5

Download the `97EMPLOY` dataset from the course webpage. This dataset contains five columns. Data are numbers of employees for various US airlines in 1997; these are approximately the averages of the 12 monthly employee counts.

Variable names:

AIRLINE	
FULLTIME	
PARTTIME	
TOTAL	FULLTIME + PARTTIME
FTequiv	FULLTIME + 0.5 × PARTTIME
TYPE	1 = MAJOR, 2 = NATIONAL, 3 = LARGE REGIONAL, 4 = MEDIUM REGIONAL

(a) The values of FTequiv vary substantially according to TYPE. Produce a display in which there are four side-by-side boxplots, so that the size differences among the TYPEs are displayed graphically. HINT: Use **Graph** ⇒ **Boxplot**. Then ask for **With Groups**; in the **Graph variables** box, name FTequiv (or C5) and in the **Categorical Variables** box, name TYPE (or C6).

(b) The display created in part (a) will leave you with only one clear impression. We can see more if we utilize logarithms of FTequiv to create a new variable which is $\log_{10}$(FTequiv). Use **Calc** ⇒ **Calculator** to make C7 = $\log_{10}$(FTequiv); the Minitab code for base 10 logarithms is LOGTEN. Now produce four side-by-side boxplots for the values of TYPE. Name this new variable as LogFTequiv.

You will find an unusual airline in the group TYPE = 2. What is this airline? HINT: Try **Editor** ⇒ **Brush**, then select the point on the graph. Please note that this calls for **Editor**, not **Edit**.

(c) Produce a scatter plot showing PARTTIME on the vertical axis and FULLTIME on the horizontal axis. Identify any airline which seems to have an unusual mix of part time and full time employees.

(d) Part (c) might come out differently if you plot $\log_{10}$(PARTTIME) versus $\log_{10}$(FULLTIME). Continue to use base 10 logarithms. Do you still find the same unusual airline(s)?

NOTE: A number of the airlines have no part time employees. Use the transformation LOGTEN(PARTTIME + 0.5).

. . . . . .

Problem 6

Indicate (without computation) which sample in each set has the higher standard deviation.

Set 1, Sample A: 16, 16, 16, 16, 16
Set 1, Sample B: 15, 16, 16, 16, 16

Set 2, Sample A: 20, 25, 25, 25, 30
Set 2, Sample B: 15, 25, 25, 25, 35

Set 3, Sample A: 20, 20, 30, 40, 40
Set 3, Sample B: 20, 25, 30, 35, 40

. . . . . .

Problem 7

Sincich, problem 2.66. The `HONEYCOUGH` data file (available on the course homepage) presents the data to you in two formats:

Separate columns format: Column C1 gives the scores for the 33 children in the Honey group. Column C2 gives the scores for the 35 children in the DM group. C3 gives the scores for 37 children in the the Control group.

Single column plus identifier format: Column C5 gives the scores for all $33 + 35 + 37 = 105$ children. Column C6 gives the labels corresponding to the three groups.

You can do this problem through either of the two formats provided. The single column plus identifier format generalizes to other situations and is vastly more useful.

. . . . . .

Problem 8

Larcenous Larry has a problem. The data column SALES had the following summary:

Descriptive Statistics: SALES

Variable	N	N*	Mean	StDev	Sum	Minimum	Q1	Median	Q3	Maximum
SALES	28	0	4319	2693	120925	1674	1842	3062	6542	8947

Unfortunately, he had promised that SALES would average at least 5,000. To cover this up, he maliciously edited the values to this:

Descriptive Statistics: SALES

Variable	N	N*	Mean	StDev	Sum	Minimum	Q1	Median	Q3	Maximum
SALES	28	0	5019	2693	120925	1674	1842	3062	6542	8947

When an auditor looks at the latter display, Larrys misdeeds will be quickly detected. How?

. . . . . .