Statistics 160 Fall 2008 Midterm #1 – Solutions

1. (a) The explanatory variable is *year* and the response variable is *population*.

1. (b) Since $r^2 = 0.9770393$ is the square of the correlation, we know that

$$r \in \left\{\sqrt{0.9770393}, -\sqrt{0.9770393}\right\} = \{0.988453, -0.988453\}.$$

However, the sign of r must be the same as the sign of the association which, in this case, is negative. Thus, $r = -0.988453$.

1. (c) The equation of the regression line is given by

$$\hat{y} = a + bx$$

where $b = rs_y/s_x$ and $a = \overline{y} - b\overline{x}$. Since the x-variable is year and the y-variable is population, we find

$$b = -0.988453 \times \frac{8.986712}{15.13825} = -0.5867879 \quad \text{and} \quad a = 18.29 + 0.5867879 \times 1957.5 = 1166.927.$$

That is,

$$\hat{y} = 1166.927 - 0.5867879x.$$

(Note that if you incorrectly gave the value 0.988453 for the correlation, then your resulting regression line would be $\hat{y} = -1130.347 + 0.5867879x$.)

1. (d) If $x = 200$, then the regression model predicts

$$\hat{y} = 1166.927 - 0.5867879 \times 2000 = -6.6488.$$

Obviously, it is not possible to have a negative population.

2. **(B)** There are 134 students with ages strictly less than 20 and there are 133 students with ages strictly greater than 21. Since there are 200 students aged each of 20 and 21, we see that the median must be 20. (That is, there are 334 students aged 20 or less, and 333 students aged 21 or more. The 334th student–whose age is 20–has the median age.)

3. **(C)** This is the only reasonable answer.

4. **(B)** If there were the same number of male workers as female workers, then the average salary would be \$38 000. However, since there are more male workers, and each male has a salary of \$41 000, the average must be larger than \$38 000.

5. **(D)** Consider the ages $\{20, 20, 30, 40, 50\}$. The median age is 30, but when the 50-year old leaves, the new median is 25. Consider the ages $\{20, 30, 30, 40, 50\}$. The median age is 30, but when the 50-year old leave, the median remains unchanged at 30. Thus, in this problem, the new median cannot be determined from the information given.

6. It is possible that the median change is negative while the mean change is positive if there is a data point whose value is very large. In this case, Cisco systems, whose stock went up 60 600%, is such an extreme outlier. Even though the majority of these 4567 companies experienced a negative change, the fact that a single data point, namely Cisco, saw it stock increase 60 600% was enough to outweigh the other 4566 companies that might have had negative returns. In general, the median is affected more by the number of data points and less by their actual values, as opposed to the mean which is affected more by the actual data value. (That is, the median is a resistant measure of centre.) In fact, the data in this problem suggests that the distribution of stock changes is right-skewed, and for any right-skewed distribution, the median will always be less than the mean.

7. (a) The Fidelity Technology Mutual Fund has the closer relationship to returns from the stock market as a whole.

7. (b) This information does NOT tell Joe anything about which fund has the higher returns. Correlation is only a measure of linear strength and is NOT a measure of the rate of change of returns.

8. (a) (See Exercise 9.47 on page 234.) In an observational study, we simply observe subjects who have chosen to take supplements and compare them with others who do not take supplements. In an experiment, we assign some subjects to take supplements and assign the others to take no supplements (or better yet, assign the others to take a placebo).

8. (b) *Randomized* means that the assignment to treatments is made randomly, rather than by some other method (e.g., asking for volunteers). *Controlled* means that some subjects were used as a *control group*—probably meaning that they received placebos—which gives a basis for comparison to observe the effects of the treatment.

8. (c) Subjects who choose to take supplements have other characteristics that are confounded with the effect of the supplements; one of those characteristics is that people in this group are more likely to make healthy lifestyle choices (about smoking, drinking, eating, exercise, etc.). When we randomly assign subjects to a treatment, the effect of those characteristics is erased because some of those subjects will take the supplement and some will take the placebo.

9. (a) If X denotes the yearly snowfall in Regina, then X has a normal $N(100, 10)$ distribution. Thus, the required probability is

$$\begin{aligned} P(90 \le X \le 120) &= P\left(\frac{90-100}{10} \le \frac{X-100}{10} \le \frac{120-100}{10}\right) \\ &= P(-1 \le Z \le 2) \\ &= 0.9772 - 0.1587 \\ &= 0.8185 \end{aligned}$$

using Table A.

9. (b) We must find the value of S (for snowfall) that satisfies

$$P(X \geq S) = 0.80.$$

Normalizing gives

$$P(X \geq S) = P\left(\frac{X - 100}{10} \geq \frac{S - 100}{10}\right) = P\left(Z \geq \frac{S - 100}{10}\right).$$

From Table A, we find $P(Z \geq -0.84) = 0.80$ and so S satisfies

$$\frac{S - 100}{10} = -0.84$$

Solving for S gives 91.6 cm as the required snowfall.

10. (a) **(i)** The population of interest is *all customers of this national restaurant chain.* **(ii)** The sample consists of the 140 customers of the Regina branch who were selected in the national restaurant chain's simple random sample and answered comment cards. **(iii)** The variable of interest is *quality of service.* It is *categorical* as there are 5 possibilites for it: poor, below average, average, above average, outstanding.

10. (b) If $\hat{p}$ denotes the proportion of customers who rated the quality of service as above average or outstanding, then

$$\hat{p} = \frac{67 + 19}{140} = \frac{86}{140}.$$

10. (c) The results from the given data *cannot* be extended to all other branches in the restaurant chain. This is simply because the data do not form a representative sample of the population of interest here. They *are* a simple random sample of the Regina branch's population base, but not of the restaurant chain as a whole. In order to extend results to all branches in the chain, it would be necessary to analyze the data for the simple random sample of customers from all the branches collectively. The results given can only be extended to the Regina branch; among other things, regional differences in personnel and customer expactations are likely.

11. In order to use regression, we must decide on which two variables to measure. It is clear that our response variable, or y-variable, must be *amount of money spent on books for Fall 2008 classes.* Of course, there will be variability in the typical cost of books across faculties (e.g., science and engineering students spend more on books than arts students do since arts books are typically novels or readers whereas science books are typically thick texts; fine arts and education students need to buy other materials and supplies which are not books). Furthermore, the more classes a student is taking, the more books that will be needed. Since a *full-time student* is one who is taking at least 9 credits, there will be variability in the number of credits that a full-time undergraduate student is taking. Hence, a reasonable choice for explanatory variable, or x-variable, is *number of registered credit hours.* Thus, the population of interest is *all full-time*

U of R undergraduates. A random sample of students could be obtained by randomly selecting from all students, or by stratifying by faculty. One practical difficulty is that due to privacy laws it might be very hard to obtain lists of students. Hence, a SRS might not be practical and instead it might be necessary to sample by classes (noting that information about classes is available online). An alternative arrangement might be to try and distribute questionnaires at the bookstore, but this presents the problem that not all students shop at the bookstore—many buy their books used or online.

The distribution of scores by problem number are as follows. The score is in parentheses.

- Problem 1: 12(8), 18(7), 6(6), 2(5), 2(4), 1(0)
- Problem 2: 15(2), 26(0)
- Problem 3: 41(2), 0(0)
- Problem 4: 32(2), 9(0)
- Problem 5: 21(2), 20(0)
- Problem 6: 2(6), 3(5), 26(4), 8(3), 2(0)
- Problem 7: 19(6), 12(5), 5(4), 2(2), 1(1), 2(0)
- Problem 8: 9(8), 9(7), 15(6), 5(5), 1(4), 1(2), 1(0)
- Problem 9: 1(8), 8(7), 16(6), 6(5), 7(4), 2(2), 1(0)
- Problem 10: 3(8), 4(7), 14(6), 13(5), 4(4), 1(2), 2(0)
- Problem 11: 7(8), 15(7), 13(6), 1(5), 2(1), 3(0)