Statistics 2000, Section 001, Midterm 1 (200 Points)

Friday, February 12, 2010

Your Name: _____

Question 1: z–Scores and Normal Distributions (**50 Points**)

The Graduate Record Examination (GRE) is a test taken by college students who intend to pursue a graduate degree in the United States. For around 210,000 female US citizens who took the General GRE Test in 2005–06, the mean for the quantitative ability portion of the exam was about 520 and the standard deviation was about 135 (`http://www.ets.org/Media/Tests/GRE/pdf/05-06_factors_final.%20pdf.pdf`). We can assume that the histogram follows a normal curve. **Show your work!**

-2 for each calculation error

1. (**15 Points**) The percentage of female US citizens who scored **more than 669** on the GRE test is roughly 13.57 %. (or no final result)

$z\text{-score} = \frac{669-520}{135} = 1.10$ (5)

area to the right of 1.10 = area to the left of -1.10 (5) $= 0.1357 = 13.57\%$ (5)

2. (**20 Points**) The percentage of female US citizens who scored **between 351 and 574** is about 54.98 %.

$z\text{-score}_1 = \frac{574-520}{135} = 0.40$ (4), $z\text{-score}_2 = \frac{351-520}{135} = -1.25$ (4)

area between -1.25 and 0.40 = (area to the left of 0.40) − (area to the left of -1.25)

$= 0.6554 - 0.1056 = 0.5498 = 54.98\%$

(4) (4) (4)

3. (**15 Points**) In order to be among the top 80% of all female US citizens, a student must have obtained a minimum GRE score of about 407.

to be in the top 80%, we need an area of 0.20 (20%) to the left. (5)

This is the case for $z \approx -0.84$ (5) (working with -0.85 is fine)

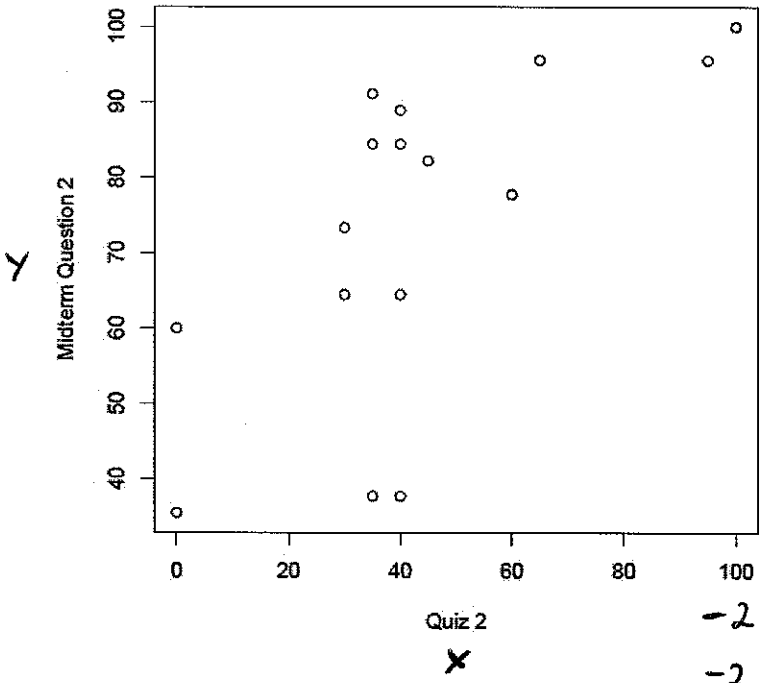
$\rightsquigarrow$ GRE score should be $-0.84 \cdot 135 + 520 = 406.6$ (≈ 407) (5)

Question 2: Regression (**50 Points**)

In a particular section of Stat 1040, students had to answer a review exercise of their textbook in Quiz 2. The result was anything but satisfactory, with the median score being an F. Detailed solutions were handed out, together with the graded quizzes. To determine whether students studied the solutions, the instructor reused the same question a few weeks later as Question 2 in Midterm 1. For a better comparability, the scores were adjusted to 100 points. In the table below, avg represents the average and SD the standard deviation.

Y Midterm Question 2 score: avg = 73 points; SD = 21 points;
X Quiz 2 score: avg = 43 points; SD = 27 points; $r = 0.65$.

The scatterplot that shows the data is displayed below and can be assumed to be football–shaped.



Y

X

-2 for each calculation error
-2 if x, y flipped

Show your work!

1. (**20 Points**) Find the regression equation for predicting the Midterm Question 2 score from the Quiz 2 score.

slope: $b_1 = r \cdot \frac{s_Y}{s_X} = 0.65 \cdot \frac{21}{27} = 0.51$ (8)

intercept: $b_0 = \bar{y} - b_1 \bar{x} = 73 - 0.51 \cdot 43 = 51.1$ (8)

regression equation: $\hat{y} = 51.1 + 0.51 \cdot x$ (4)

2. (**10 Points**) Using your regression equation, estimate the Midterm Question 2 score for a student who had a Quiz 2 score of 80 points.

$\hat{y}$ (for 80) = 51.1 + 0.51 · 80 = 91.9 ≈ 92 points (10)

3. (**10 Points**) Can we safely estimate the Midterm Question 2 score for a student who had a Quiz 2 score of 10 points? **YES** or **NO**? **Circle your answer and provide a short explanation.** (6)

$\hat{y}$ (for 10) = 51.1 + 0.51 · 10 = 56.2 ≈ 56 points

Yes: the predicted score of 56 points makes sense (it is not negative or above 100) and making a prediction for $x = 10$ is valid as this is not extrapolation (10 falls in the range of given x-values). (4)

4. (**10 Points**) As mentioned above, all scores were adjusted as if graded out of 100 points. However, the Quiz 2 scores were originally graded out of 20 points, that means, each individual Quiz 2 score was multiplied by 5 for this question. Therefore, we had an original average score of 8.6 points and an original SD of 5.4 points when grading out of 20 points. (4) (4)

As each point score initially was multiplied by 5, we now have to divide by 5. Therefore:

original avg = $\frac{\text{reported avg}}{5} = \frac{43}{5} = 8.6$ points (1)

original SD = $\frac{\text{reported SD}}{5} = \frac{27}{5} = 5.4$ points (1)

	Points	Deductions
correct & work	8	0
correct & no work	6	-2
incorrect & work	3	-5
incorrect & not work	1	-7
no answer	0	-8

based on: HW 2, Exercise 3

Question 3: Sums and Order Notation (**40 Points**)

In statistics, we usually refer to x_1 as the first observation, x_2 as the second observation, etc., and x_n as the final observation when we write down our observations in the order they were obtained (where n represents the total number of observations).

Often, we prefer to work with data that are sorted from smallest to largest, e.g., when calculating the median, we need the data to be sorted. Obviously, we can simply reorder any given list of numbers. However, we often use the notation $x_{(1)}$ to refer to the smallest observation, $x_{(2)}$ to refer to the 2nd smallest observation, etc., and $x_{(n)}$ to refer to the largest observation.

Show your work!

For $x_1 = -3, x_2 = 5, x_3 = 3, x_4 = 4, x_5 = -2, x_6 = 20$, and $n = 6$, determine the following sums (**8 Points each**):

Sorted data:

$x_{(1)} = -3$

$x_{(2)} = -2$

$x_{(3)} = 3$

$x_{(4)} = 4$

$x_{(5)} = 5$

$x_{(6)} = 20$

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{6} x_i = x_1 + x_2 + x_3 + x_4 + x_5 + x_6$$
$$= -3 + 5 + 3 + 4 + (-2) + 20$$
$$= 27$$

$$\sum_{i=2}^{n-2} x_i = \sum_{i=2}^{4} x_i = x_2 + x_3 + x_4$$
$$= 5 + 3 + 4$$
$$= 12$$

$$\sum_{i=2}^{n-2} x_{(i)} = \sum_{i=2}^{4} x_{(i)} = x_{(2)} + x_{(3)} + x_{(4)}$$
$$= -2 + 3 + 4$$
$$= 5$$

$$\sum_{i=2}^{n-1} x_1 = \sum_{i=2}^{5} x_1 = x_1 + x_1 + x_1 + x_1$$
$$= -3 + (-3) + (-3) + (-3)$$
$$= -12$$

$$\sum_{i=\frac{n}{2}}^{n^2-32} \frac{x_{n-i}}{x_{(i+1)}} = \sum_{i=3}^{4} \frac{x_{6-i}}{x_{(i+1)}} = \frac{x_{6-3}}{x_{(3+1)}} + \frac{x_{6-4}}{x_{(4+1)}}$$
$$= \frac{x_3}{x_{(4)}} + \frac{x_2}{x_{(5)}} = \frac{3}{4} + \frac{5}{5}$$
$$4 = 0.75 + 1 = 1.75$$

based on: • 5 questions directly taken from "Online Quizzes" for chapters 1 & 2 (see HW 4, Exercise 3 (EC))
• 5 questions adapted from Stat 2000, Fall 2009, Midterm 1

Question 4: Multiple Choice Questions **(60 Points)**

Mark your answer for each multiple choice question in the table below. There is only one correct answer for each question. Each correct answer is worth 4 points.

Question	(a)	(b)	(c)	(d)	Question	(a)	(b)	(c)	(d)
1	●	○	○	○	11	○	○	○	●
2	○	●	○	○	12	○	○	●	○
3	○	○	○	●	13	○	○	○	●
4	○	○	○	●	14	○	○	●	○
5	○	●	○	○	15	○	○	○	●
6	○	○	●	○					
7	○	○	●	○					
8	●	○	○	○					
9	○	○	○	●					
10	○	○	●	○					

Stat 2000, Midterm 1, Question 4 – Solutions

1. (a) The median is the average of the two observations in the middle of the ordered data, i.e., (17 + 18) / 2 = 17.5.

2. (b) The simplest way is to first compute the total of the 50 salaries and then divide by 50. The total salary of the 20 female workers is (20)($43,000) = $860,000 and the total salary of the 30 male workers is (30)($47,000) = $1,410,000. The total of all the salaries is thus $2,270,000 and the average salary is $2,270,000 / 50 = $45,400.

3. (d) The effect of adding $5,000 to each observation is to increase the first quartile by $5,000 and to increase the third quartile by $5,000. The difference, which is the interquartile range, is thus unchanged (the $5,000s cancel).

4. (d) We know that slope b1 = r * sy / sx. As sy and sx cannot be negative, b1 and r must have the same sign, i.e., be positive here. But, because sy and sx are not provided, we cannot calculate the exact value of r here.

5. (b) There are 80 white workers, 20 Asian workers and 20 in the other category. Because there are a total of 160 workers, there must be 40 blacks in the sample. Thus 40 out of the 160 workers, or 25%, are black.

6. (c) We lose all information about trends over time as we ignore the time information entirely, but we gain additional information how many observations fall into each class and thus we are able to easily determine the number of years for which the death rate was 5 or higher.

7. (c) This time plot exhibits both a trend and a seasonal variation.

8. (a) The mean is sensitive to extreme observations (outliers) which pull the mean toward them. Since the mean is much less than the median, we will suspect low outliers.

9. (d) From the graph, we are interested in the top 6 bars. Visually, these account for about 44 individuals. 44/180 is 24.4%, i.e., about 25%.

10. (c) We are dealing with one quantitative variable (change in blood pressure) and one categorical variable (type of music). (a) is impossible because we need two quantitative variables for a scatterplot. When we create a single histogram as suggested in (b), we lose the information about the type of music. (c) makes most sense as each boxplot summarizes the data for each type of music and we can use these boxplots to explore the relationship between type of music and change in blood pressure.

11. (d) We are dealing with one quantitative variable (income) and one categorical variable (gender). Correlation only can be calculated for two quantitative variables – so the answer is (d).

12. (c) price = -6677 + 175 * 200 = $28,323

13. (d) The year 2010 is far into the future (17 years beyond our last observation in 1993). Making any prediction for a value so far off from the existing data is extrapolation and the numerical result obtained wouldn't make any sense at all. Also, we cannot assume that the relationship observed from 1965 to 1995 will continue to hold into the future, it.e., that there will be improvements of the world record every few years. By the way, the last point in the scatterplot (the 29 min 31 sec 78/100 = 1771.78 sec run by Wang Junxia from China on Sept 8, 1993) still represents the current world record, i.e., no further improvement took place over the last 17 years. See: http://en.wikipedia.org/wiki/10,000_metres

14. (c) Females always mate with males that are .50 years younger, so the relationship is perfectly linear, and as females get older so will the males they mate with, so the correlation is 1.

15. (d) Goals Allowed is the explanatory variable and Winning% is the response variable. For each additional Goal Allowed, the Winning% decreases by 0.26% (as the sign in front of the slope is negative and not positive).