

Math 110
Test 1 Sample Solutions
February 12, 2010

Be sure to provide explanations for your answers as indicated. You may use your calculator and z-table.

1. (10 points each) Short Answer:

- (a) In an observational study, what is meant by *historical controls*? What issue or issues arise with the use of historical controls?

Historical controls refers to the use of data on patients with like medical histories and circumstances as the treatment group, but who did not undergo the treatment and who are not actively observed. It is difficult to find “like” populations, thus allowing for the possibility of hidden confounders, including a selection bias for sicker patients.

- (b) Explain the *regression effect* and the *regression fallacy*.

The regression effect applies in a test-retest situation. It refers to the fact that on average second test scores of individuals will be closer to the mean in standard units than first test scores. The regression fallacy is to attribute the regression effect to some cause. It is simply a property of football shaped data.

- (c) If you are given related data sets $\{x_1, x_2, \dots, x_N\}$ and $\{y_1, y_2, \dots, y_N\}$ and their means μ_x and μ_y :

- i. How do you calculate the standard deviation of the x data set? (Either explain briefly how to do it or give the formula.)

In words, it is the root-mean-square of the differences between a data point and the mean. Or, it is the square root of the average of the squares of the distances between data points and the mean. In symbols,

$$SD = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu_x)^2}$$

where μ_x is the mean.

- ii. How do you calculate the correlation coefficient of the two data sets? (Either explain briefly how to do it or give the formula.)

In words, it is the average of the products of the x and y data values in standard coordinates. In symbols,

$$r = \frac{1}{N} \sum_{i=1}^N \frac{(x_i - \mu_x)}{SD_x} \frac{(y_i - \mu_y)}{SD_y}$$

- (d) True or False:

- i. If all the numbers in a data set are multiplied by 5, then the mean and standard deviation of the data set are also multiplied by 5.

True. Replacing each value x_i by $5x_i$ in the formula for the mean will multiply the mean by 5. Similarly, the SD will be increased by a factor of 5.

- ii. If all the numbers in a data set are multiplied by 5, then the value of each data point in standard units is also multiplied by 5.

False. All the 5's in the formula will cancel:

$$\frac{5x_i - 5\mu_x}{5SD_x} = \frac{x_i - \mu_x}{SD_x},$$

which is the z -value (or standard units) for x_i .

2. (20 points) The following data on household income comes from Table HINC-01 of the Current Population Survey. It is broken down by whether the household is inside or outside a Census Bureau designated Metropolitan Statistical Area. The unit for households is thousands.

Income Range	0-\$25K	\$25-\$50K	\$50-\$75K	\$75-\$100K	Over \$100K	Row Total
Inside	22,942	23,860	17,682	11,828	21,278	97,950
Outside	6,093	5,193	3285	2,014	2,308	18,893

- (a) Construct a histogram for each row of the table. Indicate the height and area of each block. (Use \$100-\$200K for the rightmost block.)

For inside the areas and heights (in %/\$1000) of the blocks are:

Income Range	0-\$25K	\$25-\$50K	\$50-\$75K	\$75-\$100K	Over \$100K
Area	23.4%	24.3%	18.1%	12.1%	21.7%
Height	.936	.974	.724	.484	.21

For outside the areas and heights (in %/\$1000) of the blocks are:

Income Range	0-\$25K	\$25-\$50K	\$50-\$75K	\$75-\$100K	Over \$100K
Area	32.3%	27.4%	17.3%	10.6%	12.2%
Height	1.292	1.096	.692	.424	.122

- (b) Based on your histograms, what conclusions can you draw about the comparative distribution of household incomes inside and outside the Metropolitan Statistical Areas? Explain.

For outside the MSAs, household income skews to the left with higher percentages of data in the first two blocks. For inside the MSAs, household income skews to the right with a higher percentage of data in the rightmost block. On average, households inside MSAs are wealthier than those outside MSAs.

3. (20 points) In 2008, the Educational Testing Service reported that a total of 812,764 females took the SAT Critical Reading test. The mean and SD of their scores were 500 and 110.

- (a) Using the above information, estimate the number of scores between 485 and 525.

First, use z -values and the z -table to find the area of the symmetric region associated with each score:

$$z = \frac{525 - 500}{110} \approx .227 \rightarrow Area \approx 18\%$$

$$z = \frac{485 - 500}{110} \approx -.136 \rightarrow Area \approx 10\%$$

Since these areas are symmetric, we want to add half these areas to find the final area. $\frac{1}{2}18\% + \frac{1}{2}10\% = 14\%$. Then 14% of 812,764 is approximately 113,787.

- (b) The actual number of scores between 485 and 525 was 116,510. Would this lead you to believe the scores were normally distributed or not? Explain.

113,787 is close to 116,510. In fact, it is 96% of 116,510. These values are reasonably close, so it would lead us to believe the scores were normally distributed.

- (c) Using the above information, estimate the score at the 30th percentile of the scores.

The 30th percentile is a left-hand tail. It corresponds to the symmetric area of 40% and a z -value of .52. Then converting this to a score, $.52 \times 110 = 57.2$. Since the score we want is less than the mean, the answer is $500 - 57.2 \approx 443$.

4. (20 points) The attached scatter plot contains data on the point spread and actual point difference for $N = 672$ professional football games. (Data from the StatLib Datasets Archive at Carnegie-Mellon. See below for information about point spreads.)

- (a) The mean and SD for point spread are 5.3 and 3.3 respectively. The mean and SD for actual difference are 6.1 and 13.77 respectively. The correlation coefficient r is .28.

Use this information to plot the SD-line and the regression line for the data on the plot. (Be sure to label which line is which.)

- (b) Write a correct formula for the regression line.

The regression line is given by

$$y = r \cdot \frac{SD_y}{SD_x}(x - \mu_x) + \mu_y = .28 \frac{13.77}{3.3}(x - 5.3) + 6.1 = 1.17x - .09.$$

- (c) Based on the regression line what actual difference would you predict for a point spread of 7?

Evaluate $1.17 \cdot 7 - .09 \approx 8.08$.

- (d) Suppose the correlation coefficient for the data were $r=1$ (it's not, but keep going), what would the data cloud look like and what would it say about the outcomes of the games?

The data would fall in a straight line. It would say the outcome of the game can be determined exactly from the point spread. (We can't say it is the same because the SDs might be different.)

Additional Information on Point Spreads

The *point spread* is an important element in the popularity of betting on professional football. Set by Las Vegas bookmakers, the point spread on a game is a fixed number of points that is "given" to the underdog team to encourage betting on the underdog. For example, with a point spread of 7

- if one bets on the favorite, one is betting on the favorite winning by more than 7 points, or

- if one bets on the underdog, one is betting on the underdog losing by less than 7 points (or winning).

The *actual point difference* is the score of the favorite minus the score of the underdog. If this is positive, the favorite team won and, if it is negative, the underdog won.

In the scatter plot, each dot represents one football game. The x -coordinate of the dot is the bookmakers' point spread for the game and the y -coordinate is the actual point difference.

Of course, for people who bet it is worth knowing whether the bookmakers' point spread is a valid predictor of the difference in the score.