

Math 110

Test 1

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:

- In an observational study, what is meant by *historical controls*? What issue or issues arise with the use of historical controls?
- Explain the *regression effect* and the *regression fallacy*.
- 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 :
 - How do you calculate the standard deviation of the x data set? (Either explain briefly how to do it or give the formula.)
 - How do you calculate the correlation coefficient of the two data sets? (Either explain briefly how to do it or give the formula.)
- True or False:
 - 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.
 - 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.

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 |

- Construct a histogram for each row of the table. Indicate the height and area of each block. (Use \$100-\$200K for the rightmost block.)
 - Based on your histograms, what conclusions can you draw about the comparative distribution of household incomes inside and outside the Metropolitan Statistical Areas? Explain.
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.
- Using the above information, estimate the number of scores between 485 and 525.
 - 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.
 - Using the above information, estimate the score at the 30th percentile of the scores.

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.)
- 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.)
 - Write a correct formula for the regression line.
 - Based on the regression line what actual difference would you predict for a point spread of 7?
 - 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?

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.