Math 110

Test 1 Sample Solutions
September 26, 2008

Be sure to provide explanations for your answers as indicated.

1. (20 pts.) The article "Do We Really Know What Makes Us Healthy," by Gary Taubes in the New York Times Magazine of September 16, 2007, concerns the difficulties of relying on observational studies to set treatment policy. In particular, it points out four confounding factors whose effects are difficult to determine.

   (a) Define two of these four confounding factors.

      i. Healthy user bias: The people most likely to us a treatment or druc are wealthier and more educated in general, but these same people are generally healtier.

      ii. Prescriber effect: Doctors may slant their prescribing of drugs toward healthier patients. If a patient is very ill from one disease, the doctor might not prescribe additional treatments for another condition.

      iii. Compliance effect: People who strictly follow drug regimes, that is, directions for taking medication, are generally more educated and hence healthier.

      iv. Eager-patient effect: Patients who ask for drugs or therapies are more likely to get them. But these patients are also more educated and wealthier, hence healthier.

   (b) For the factors you defined in (a), explain why they might have been confounding factors in the Nurses' Health Study obse

      i. Healthy user bias: The nurses are generally well-educated (they at least finished nursing school) and well-paid, so they are generally healthier than the average person.

      ii. Prescriber effect: Since the choice of going hormone replacement therapy was prescribed by the patients' doctors, not assigned randomly, there is a bias of the users toward being healthier.

      iii. Compliance effect: Since the nurses who took the hormone replacement therapy self-reported whether or not they took the drug, there was a bias in the study toward those who comply with drug regimes.

      iv. Eager-patient effect: Sinces the nurses who opted for hormone replacement therapy did so voluntarily, they were eager patients.

2. (20 pts.)

   (a) Sketch a football shaped data cloud with $r \approx .6$.

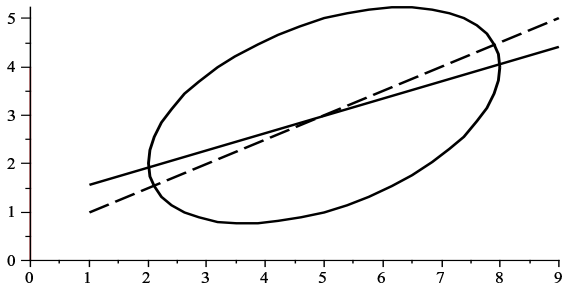

Figure 1:

(b) Sketch the SD-line on your plot. How do you know this is (roughly) the SD-line?

The SD-line is the dashed line. It passes through the tips of the football shaped cloud of data.

(c) Sketch the regression line on your plot. How do you know this is (roughly) the regression line?

The regression line is the solid line. It passes through the points near the ends of the football shaped cloud of data that lie at the extremes of the x-coordinates. Here the tangent to the curves that outlines the cloud are vertical.

3. (20 pts.) The following data on people without health insurance by family income level for 2006 is from the Current Population Survey.

Family Income	No Insurance	All People
	N= 44,815	N=293,834
Less than $25,000	18,590	70,478
$25,000 to $49,999	13,620	72,963
$50,000 to $74,999	6,445	55,258
$75,000 to $200,000	6,160	95,136

Notice N= at the top of each column is the total number of people in the four class intervals for the column. Also, the number of people is in units of 1000.

(a) For each of the four income class intervals, compute the percentage of No Insurance people in the class interval and the percentage of all people in the interval.

The percentages for the no insurace group are:

$$\frac{18596}{44815} \to 41.5\%, \ \frac{13620}{44815} \to 30.4\%, \ \frac{6445}{44815} \to 14.14\%, \ \frac{6160}{44815} \to 13.8\%,$$

The percentages for all people are:

$$\frac{70478}{292834} \to 24\%, \quad \frac{72963}{292834} \to 24.8\%, \quad \frac{55258}{292834} \to 18.8\%, \quad \frac{95136}{292834} \to 32.4\%,$$

(b) Use your percentages from (a) to construct two histograms, one for people without insurance and one for all people using the class intervals in the table. (Use the graph paper.)

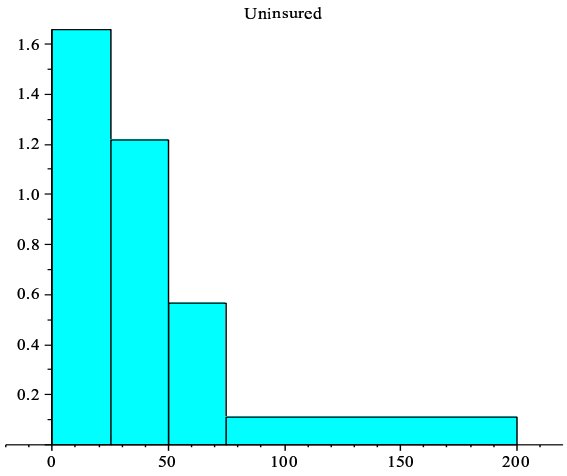

Figure 2:

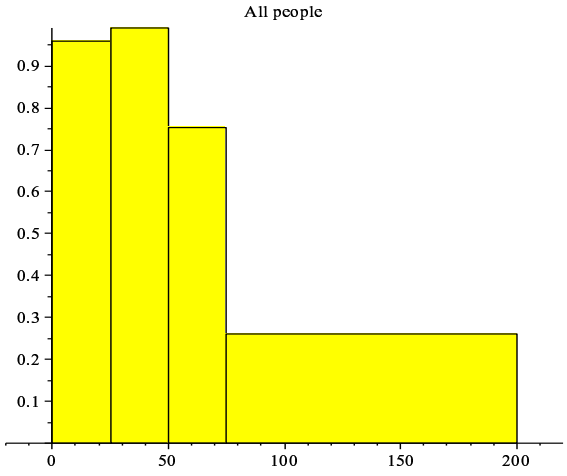

Figure 3:

(c) By comparing your histograms, what can you say about the distribution of people without health insurance by family income in comparison to the distribution of all people by family income?

The income distribution of those without health insurance is shifted to the left. The distribution of people without health insurance is significantly

higher in the lowest income category and significantly lower in the highest income category.

4. (20 pts.) The grades in a large university statistics class are normally distributed with a mean of 60 (out of a possible 100 points) and an SD of 12.

   (a) Draw a smooth histogram for the data.

   Sketch a bell shaped curve with mean located at 60 and turning or inflection points located at 48 and 72.

   (b) If 200 students took the test, estimate how many scored between 70 and 80.

   Convert 70 and 80 to standard units.

$$\frac{70-60}{12} = \frac{5}{6} \approx .83 \text{ and } \frac{80-60}{12} = \frac{5}{3} \approx 1.66.$$

   These z-values correspond to areas of approximately 59% and 90.5%. Then

$$\frac{90.5-59}{2} \approx 16.75\%.$$

   Finally, 16.75% of 200 is approximately 33 people.

5. (20 pts.) (Hypothetical) A town near Worcester administers a spelling test to all students in fourth grade and all students in fifth grade. The test consists 100 words. For one group of fourth graders, the average for the number of words correctly spelled was 50 with a standard deviation of 10. When the same group took the test again in the fifth grade, the average for the number of words correctly spelled was 60 with a standard deviation of 15. The correlation coefficient for the data is $r = .7$. Suppose in the fourth grade a student was in the 70th percentile. What would you estimate for the student's percentile in the fifth grade? (Be sure to show your calculations and draw normal approximations as needed.)

   The 70th percentile corresponds to a z-table area of 40% and a z-value of approximately .52. To convert this to a z-value on the second test, multiply by $r = .7$. This gives a z-value for the second test of .364 which corresponds to an area of approximately 29%. The second test percentile estimate is then $29/2 + 50 = 64.5$.