MATH 316 (RUSIN) TEST 1, Feb 22 2012 – POSSIBLE ANSWERS

Here are some answers that you COULD have given. Other answers can be just as correct, especially if you give a good justification of them!

1a. The latest Zogby poll shows that the percentage of registered Republicans who prefer Mitt Romney to be their presidential candidate varies from state to state. Here are the percentages in the three states voting next week:

Arizona	73%
Michigan	42%
Washington	51%

Suggest an appropriate graphical display of these data.

1b. The Zogby data actually measures candidate preferences in each precinct. (There are hundreds of precincts in each state.) Explain why it might not be appropriate to use the same type of graphical display used in 1a to present the precinct-level data. What other kind(s) of display of the precinct-level data might be useful or interesting?

ANSWER: For 1a, I would accept a bar graph with three bars (in any order). NOT a line graph, NOT a pie chart. (Why?)

For 1b the problem is that there would be many hundreds of bars to show and it's not clear what they would tell us anyway. It might make sense to give a distribution of the percentages: how many districts show support levels of 0-10%, 10-20%, etc. It might be useful to show three separate such distributions, one for each state. It might be useful to show the precincts geographically on the state map, shading each precinct according to the level of support for Romney. It all depends on what message the statistician is trying to convey!

2. The Nielsen company monitors the television-viewing habits of large numbers of Americans. In a recent week they observed the number of hours that their volunteers spent watching TV. Here are the numbers collected from the subjects who live here at UT:

13.8, 5.7, 33.8, 11.8, 27.0, 18.9, 19.3, 20.8, 25.4, 23.1, 52.0, 7.8, 10.9, 9.0, 9.0

13.8, 5.7, 33.8, 11.8, 27.0, 18.9, 19.3, 20.8, 25.4, 23.1, 52.0, 7.8, 10.9, 9.0, 9.0

(a) Which, if any, of these observations would you consider outliers, and why?

(b) The median number of hours (per week) spent watching TV is almost the same for UT students as for the general population. However, the distribution for the whole US is skewed to the right. Draw a distribution which shows this. (Be sure that each axis in your graph has an appropriate label and shows numbers that are suitable for their label.)

ANSWER: This is a numerical variable, so it makes sense to sort the data and compute quartiles: $Q_1 = 9.0, Q_2 = 18.9$ and $Q_3 = 25.4$. If we use the $1.5 \times IQR$ rule, that flags as outliers anything that's more than 24.6 away outside Q_1 and Q_3 as an outlier. Only the

52.0 qualifies under this rule. But you could make a case for one or two other very-large or very-small numbers to be considered outliers, too, based for example on the size of the gaps between it and its neighbors.

The US distribution would have to have a median of of 18.9 hours also; that means there MUST be equal area under your distribution graph on both sides of the line $x = 18.9$. To make it skewed to the right, move some of the observations further to the right (i.e. make the graph on the right be a little lower but wider). Note that your graph can NOT extend past $x = 168$! (Why?)

3. In a recent semester I gave a test and found these scores within the class:

54, 60, 61, 65, 67, 68, 70, 74, 76, 76, 77, 78, 78, 79, 80, 80, 83,
83, 83, 86, 87, 87, 88, 88, 88, 89, 90, 91, 91, 92, 93, 94, 96, 99

The students (obviously) wanted to know, "how did we do?". Give the five-number summary of these 34 data points. What would you say to the person who wanted to know how a score of 75 would compare to the rest of the class?

ANSWER: The numbers m, Q_1, Q_2, Q_3, M are easily spotted to be $54, 76, 83, 89, 99$ respectively. A score of 75 is then just below the first quartile, i.e. the student would be scoring better than only one-quarter of the class. (You could also tell the student how many standard deviations below average he or she is.)

4. A survey was done of the weights of American males in their 20s. The median weight was 160 lbs, with first and third quartiles at 150 and 180 lbs respectively. A researcher who wanted to summarize the dataset more succintly calculated instead the mean weight and the standard deviation of the weights; they are 163 lbs and 13 lbs respectively.

4a Explain why it might be reasonable for the mean weight to be higher than the median weight.

4b The second researcher chose to compute those numbers because she thought the weights would follow a normal distribution. Do the numbers given above make that seem like a reasonable assumption?

ANSWER: The mean is more than the median exactly when the heavier half of the population is further away from average than the lighter half of the population. It might be reasonable to think that there are equal numbers of 159-lb males and 161-lb males, equal numbers of 158-lb males and 162-lb males, etc; but there are no males who weigh 0 or less while there are certainly males who weigh 320 or more, so as we move up and down away from average, the heavier-than-average guys must be further from average than their lighter-than-average counterparts. (Indeed, already by the time we try to pair off the 150-lb guys with someone heavier, we've worked our way down to the first quartile; their heavier-than-average counterparts would then be at the third quartile, who weigh 180 lbs.)

These numbers are really inconsistent with having a normal distribution. For starters, the quartiles are not equidistant from the median. Also, the middle half of the population weighs between 150 and 180, but in a normal distribution we would already have 68%

of the population between 150 and 176, and so we'd expect much more than half of the population to be accounted for between 150 and 180!

5. It is generally true that the SAT math scores of all students at a university follow a roughly normal distribution. But a web search does not reveal means and standard distributions; instead, we can find the "middle half" range for each school. For example, at Auburn University, half of the students had a math SAT score between 540 and 660 (out of 800); a quarter were below 540 and a quarter were above 660.

Assuming that Auburn's distribution really is normal, compute the mean and standard deviation of their students' math SAT scores. Then, estimate what fraction of their students have a math SAT above 750.

ANSWER: A normal distribution is symmetric: the mean and median are equal, and equidistant from Q_1 and Q_3, so we would have to have a mean of 600. In that case, the given data state that one-quarter of the population has an SAT score which is 60 less than the mean.

Now use the Z-tables: to include exactly one-quarter of the population you should include all the people whose score is more than 0.675 (approx.) standard deviations below the mean. So if we compare to the previous paragraph, 60 points on the SAT represents 0.675 of a standard deviation, so the SD is about $60/0.675 = 88.9$ points.

6. There are three SAT components, so a student's total SAT score can be as high as 2400. Can we use the SAT to predict a student's likely college GPA? One group of 8 students was followed closely; their SAT totals averaged 1800 with a standard deviation of 100, and their GPAs averaged 3.00 with a standard deviation of 0.80. Here is a table showing the performances of these students:

Marsh, Stan	1860	3.48	+0.6	+0.6
Broflovski, Kyle	1850	3.80	+0.5	+1.0
Cartman, Eric	1650	3.48	−1.5	+0.6
McCormick, Kenny	1660	2.20	−1.4	−1.0
Stotch, Butters	1800	2.44	+0.0	−0.7
Black, Token	1910	3.00	+1.1	+0.0
Testaburger, Wendy	A	1.72	−0.3	−1.6
Kim, Tuong Lu	1900	3.88	+1.0	B

Here the first column shows the SAT score and the second shows the GPA; in columns 3 and 4 these scores are normalized; for example, Broflovski's SAT score is 0.5 standard deviations above the mean, and his GPA is 1.0 standard deviations above the mean.

6a. Compute the two missing entries A and B in the table.

6b. Compute the correlation coefficient r between these two variables.

6c. Does it appear that the SAT score is a good predictor of GPA ? Explain.

ANSWER: $A = 1800 + (-0.3) \times (100) = 1770$ and $B = (3.88 - 3.00)/(0.80) = 1.1$.

The correlation coefficient is computed by adding the products of the last two columns and dividing by $n-1$, in our case

$$((+0.6)\times(+0.6))+((+0.5)\times(+1.0))+((-1.5)\times(+0.6))+((-1.4)\times(-1.0))$$
$$+((+0.0)\times(-0.7))+((+1.1)\times(+0.0))+((-0.3)\times(-1.6))+((+1.0)\times(+1.1))$$

gives 2.94; dividing by 7 gives $r = .42$ That's a "modest" correlation — some would say "the SAT predicts $.42^2 = 18\%$ of the variation of GPA", which is some, but not a lot.

7. My retirement plan offers to let me buy shares in funds. The value of each share of a fund rises and falls a little bit every day. People buy shares of these funds because, generally speaking, the values of a share of a fund tend to rise over time, but there is no guarantee that the investments will grow in value. Over the course of a year, that provides approximately 250 numbers representing the value of a share each day.

The company that manages these funds reports that over the last year, Fund A ("U.S. Stocks") and Fund B ("International Bonds") had a correlation coefficient of $r = -0.25$. Separately, they compute that over the last year, Fund A and Fund C ("Canadian Stocks") had a correlation coefficient of $r = 0.95$.

7a. Explain (in terms of prices of shares of the various funds) what it means to say that first correlation coefficient is negative. Does that mean Fund B is a poor investment choice?

7b. The company did not report the correlation coefficient between Fund B and Fund C. Would you expect this coefficient to be positive or negative, or is it impossible to say? Large or small, or is it impossible to say?

ANSWER: The negative correlation coefficient means that when Fund A has gone up in value, Fund B tends to go down in value and vice versa. That does NOT mean B is a bad investment; to the contrary, it's typical to look for combinations of funds like this — a negative correlation would mean that whenever A has had a temporary drop, B is likely to rise, and vice versa. As long as the overall trend for both funds is upwards, owning a little of each fund will help smooth out the daily rises and falls.

The fact that Fund C has such a high correlation coefficient with Fund A means that Fund C tends to rise and fall almost exactly when Fund A does. So if we already know that Fund B tends to go opposite Fund A, then it follows it will go opposite to Fund C, too – the new correlation coefficient should be negative. In fact, because 0.95 is so high, the new coefficient should be very similar to the -0.25 already computed. (It is possible to specify the range of values that is possible, but we did not do that in class.)

8. A physical exam is administered to a group of women. The ages of the women have a mean of 35 and a standard deviation of 10; their resting pulse rates have a mean of 70 and a standard deviation of 5. The correlation coefficient between the ages and the pulses is $r = -0.5$.

One woman refuses to tell her age but her resting pulse is observed to be 65. What is the best estimate of her age?

ANSWER: From the correlation coefficient we can compute the slope of the line which plots age versus pulse: it would be $r \times \left(\frac{\sigma_{\text{age}}}{\sigma_{\text{pulse}}} \right) = -0.5 \times (10/5) = -1$. In other words, each unit of additional pulse above average represents one year of age below average.

In this case the woman's pulse is 5 bpm below average, so her age must be 5 years above average: we would guess that she's 40. (Of course, that correlation coefficient is not especially large; we should not be surprised if our guess is off by a few years!)

9. A random sample of adults is asked about the difficulty of their job: they were asked to describe it as "easy", "balanced", or "difficult". These same adults were also asked what was their highest level of education completed. The counts of all the combinations were

Difficulty:	Grade school	High school	College or more
Easy	31	161	81
Balanced	49	269	85
Difficult	47	112	14

Using marginal or conditional distributions as you feel appropriate, describe the connection between a person's educational level and their perception of the difficulty of their job.

ANSWER: Of the 849 people in the sample, we compute the distribution of the margin sums: 15% finished only grade school, 64% high school, and 21% college or beyond; and 32% find their jobs easy, 47% balanced, 20% difficult. Now compare those numbers to the conditional distributions: of the 15% who did not finish high school, only 24% find their job easy, while a full 37% find their job difficult; these people find work to be distinctly more challenging than the rest of the population. At the other extreme, among the 21% with a college degree, fully 45% find their job easy, and a mere 8% find their job difficult – clearly these people find their work to be noticeably different from the perceptions of the others in the sample. (As might be expected, since most of the population is in the middle-education subgroup, this subgroup's experiences are not very different from those of the population as a whole.)

Since education typically precedes work, it seems more natural to me to form the subgroups this way and then look at their work outcomes. You might be able to phrase questions for which is makes sense to break down the population along work-perception lines.

10. A drug manufacturer would like to test a new medicine which (it believes) can allow a person to study statistics better. It would like to conduct a reputable trial of this drug. Describe (in one or two paragraphs) some steps you would take to help design this experiment so that the results will be free of bias or hidden variables.

ANSWER: Your answers will surely vary, but I do hope to see that you discuss some mechanisms to choose your subjects in a random way. For example, it's not appropriate simply to choose UT students because you guys are smarter than average. You might also want to look for a stratified sample, forming pools of candidates that represent proportional

numbers of males and females, students of various ages, etc. A significant "hidden" variable is not very hidden, I think: you will probably want to measure the students' math abilities at the start, since students who have already done well in math and stats courses will likely do well again.

It is also appropriate to discuss the use of control groups, i.e. some people should be taking a placebo. And the subjects should be assigned to the drugged/placebo groups at random, preferably in a double-blind manner.

It might also be appropriate to discuss the ethical issue involved in giving some students (but not others) a drug which has the potential to give them a grade boost.

I look forward to seeing what other issues you might bring to this discussion! (How do we measure what it means to "study better"? What kind of time frame is appropriate? What about medical side effects? etc.)