MONT 106N – Identifying Patterns
Final Exam Solutions – December 17, 2009

I. The City University of New York has about 200,000 students on 21 campuses, whose ages are distributed as in the following table:

Ages	Percent
≤ 19	21.1
20 to 22	26.2
23 to 24	10.8
25 to 29	15.5
30 to 44	20.2
45 to 64	4.6
≥ 65	1.6

A) (10) On the graph paper supplied, construct a histogram for this data. To do this in a standard way, think of each year of age as corresponding to the interval starting at that year and going all the way to the next whole number of years. Thus, for instance, the age range 20 to 22 will correspond to a histogram bar with base from 20 to 23. Make the first bar extend from 15 to 20 and the last one extend from 65 to 75.

Solution: Following the directions above, the first bar in the histogram extends from 15 to 20, so its height is $21.1/5 \doteq 4.22$, the second has height $26.2/3 \doteq 8.73$, the third has height $10.8/2 = 5.4$, the fourth has height $15.5/5 = 3.1$, the fifth has height $20.2/15 \doteq 1.35$, the sixth has height $4.6/20 = .23$, and the last one has height $1.6/10 = .16$.

B) (10) Is the *average* age or the *median* age larger for the CUNY student population?

Solution: Since the distribution of ages has a long right tail, the average is larger than the median.

II. Suppose you are going to design a study to test the effectiveness of teaching elementary German by two different methods: (a) in a traditional classroom, with spoken lessons, text readings, etc.; (b) in a media lab with language tapes, music videos, cultural films, etc. There are 50 students who will be starting elementary German at your school next year and one section will be taught by each method.

A) (10) *From the point of view of the reliability of the results of the study*, would it be a good idea to have the same professor assigned to teach both sections? Explain.

Solution: Yes. This would be one way to control for possible differences in the effectiveness of different teachers.

B) (10) *From the point of view of the reliability of the results of the study*, is there a problem with letting the students select which section they register for? Explain.

Solution: Yes. If students select which section they register for, then we might expect to see some bias from confounding factors since the students who learn better in the traditional classroom might chose that option and vice versa. Note that this is one place where educational objectives might clash with the requirements of good experimental design(!)

III. The scatter plot for the first- and second-year GPA's of a large number of university students is football-shaped, with a positive r-value.

A) (10) True or False: On average, students who had GPAs above average in their first year also had GPAs above average in their second year. Explain, *briefly.*

Solution: True. Since r is positive, the regression line slopes up to the right, so an above-average GPA in the first year would lead to a predicted above-average GPA in the second year.

B) (10) True or False: A student who was at the 40th percentile of the first year GPAs is also likely to be at the 40th percentile of the second year GPAs. Explain, *briefly.*

Solution: False. From the regression effect, we would expect the second-year GPA to be somewhat higher than the 40th percentile.

IV. (Hypothetical) The grades in a large university statistics class were normally distributed with an average of 65 (out of 100) and an SD of 10.

A) (10) Draw a smoothed histogram for the data, showing the locations of the average and ave $\pm$ SD on your plot.

Solution: The graph should be a normal-shaped curve with maximum at ave $= 65$, and inflection points (changes in concavity) at $65 - 10 = 55$ and $65 + 10 = 75$.

B) (10) If 250 students took the test, estimate how many of them scored between 70 and 75.

Solution: The interval from 70 to 75 is $z = 0.5$ to $z = 1.0$ in standard units. The area under the normal curve between these z-values is

$$\frac{1}{2}(A(1.0) - A(0.5)) = \frac{1}{2}(68.26\% - 38.30\%) \doteq 15\%$$

C) (10) What exam score was at the 90th percentile of the distribution?

Solution: The 90th percentile is the score z where $A(z) = 80\%$. This happens when $z \doteq 1.28$ (with interpolation) or $z \doteq 1.30$ if you just estimate from the entries in our table. The standard unit value $z = 1.28$ corresponds to the score $65 + 1.3 \times 10 \doteq 78$.

V. All parts of this problem deal with the following data set:

x	y
1	1
8	4
10	6
10	12
14	12
17	7

A) (10) On the graph paper provided, draw the scatter plot.

Solution: Plot omitted.

B) (10) Compute the averages and SD's of the x and y data.

Solution: We compute:

$$\text{ave}_x = \frac{1+8+10+10+14+17}{6} = \frac{60}{6} = 10$$

and

$$\text{ave}_y = \frac{1+4+6+12+12+7}{6} = \frac{42}{6} = 7.$$

Then

$$SD_x = \sqrt{\frac{1}{6}((-9)^2 + (-2)^2 + 0 + 0 + 4^2 + 7^2)} = 5$$

and

$$SD_y = \sqrt{\frac{1}{6}((-6)^2 + (-3)^2 + (-1)^2 + 5^2 + 5^2 + 0)} = 4.$$

C) (10) Compute the correlation coefficient, r.

We proceed as usual. The standard unit equivalents of the xy data are

z	w
$-9/5$	$-6/4$
$-2/5$	$-3/4$
0	$-1/4$
0	$5/4$
$4/5$	$5/4$
$7/5$	0

The average of the products $z_i w_i$ is

$$r = \frac{1}{6}(54/20 + 6/20 + 0 + 0 + 20/20 + 0) = \frac{80}{120} = \frac{2}{3} \doteq .67$$

D) (10) On your scatter plot, draw in the regression line for estimating values of y, given values of x.

Solution: The equation of the regression line is $y - 7 = \frac{4}{5} \times \frac{2}{3}(x - 10)$, or

$$y = \frac{8}{15}x + \frac{5}{3}.$$

VI. (Hypothetical) At a college, a large group of incoming students took language placement tests in *both* Spanish and French. The average on the Spanish test was 70 (out of 100) with an SD of 10. The average on the French test as 75, with an SD of 12. The correlation coefficient between the Spanish and French test scores was $r = .6$. The scatter plot was football-shaped (homoscedastic).

A) (10) Estimate the average French score for the students who scored 80 on the Spanish exam.

Solution: Call x the Spanish exam score and y the French exam score. The desired estimate can be found from the equation of the regression line: $y - 75 = \frac{12 \times .6}{10}(x - 70)$. With $x = 80$ the estimate is

$$75 + \frac{12 \times .6}{10}(10) = 82.2$$

The estimated average French score for these students is about 82.

B) (10) For the students who scored 80 on the Spanish test, estimate the SD of their scores on the French exam.

Solution: Since we are given that the scatter diagram is football-shaped, this SD would be estimated by the rms error for regression, which is

$$\sqrt{1 - r^2} \times SD_y = \sqrt{1 - (.6)^2} \times 12 = .8 \times 12 = 9.6 \doteq 10$$

C) (10) Using your answers for parts A and B, among the students who scored an 80 on the Spanish test, estimate the percentage who scored above 85 on the French test.

Solution: Using the estimate of 10 for the rms error of regression from part B and the estimate of 82 for the average from part A, we want the percentage whose scores were .3 standard units or more above the average. The area from $z = -.3$ to $z = .3$ under the normal curve is $A(.3) \doteq 23.58\%$. We want the area in the upper tail which is $50 - \frac{23.58}{2} \doteq 38.21$. So about 38% scored above 85 on the French exam.

VII. Three cards are dealt (without replacement) off the top of a well-shuffled standard deck (with 52 cards – ace through king in each of the four suits: clubs, diamonds, hearts, spades).

A) (10) What is the chance that you get only kings?

Solution: The probability is

$$\frac{4}{52} \times \frac{3}{51} \times \frac{2}{50}.$$

B) (10) What is the chance that you get no kings?

Solution: There are 48 non-kings to start, then 47 non-kings left provided that the first draw is a non-king, then 46 non-kings left for the last, provided that the first two are non-kings. The probability is

$$\frac{48}{52} \times \frac{47}{51} \times \frac{46}{50}.$$

VIII. A box contains 8 red marbles and 3 green marbles.

A) (10) If you draw six marbles at random, record the color, replace the marble each time and mix up the box, what is the chance of getting 4 red and 2 green marbles?

Solution: This is given by the binomial formula:

$$\binom{6}{4}\left(\frac{8}{11}\right)^4\left(\frac{3}{11}\right)^2 = \frac{6!}{4!2!}\left(\frac{8}{11}\right)^4\left(\frac{3}{11}\right)^2.$$

(This is a chance of about 31.2%.)

B) (10) Is the chance the same if you set aside each marble as it is drawn (without replacing it in the box)? Explain why or why not.

Solution: No, the probability would not be the same. If you do not replace the marbles, then the probabilities of drawing reds and greens change according to how many of each color have been drawn.

Have a peaceful and joyous holiday season!