MONT 106N – Identifying Patterns
Solutions for Final Exam Review Questions
December 13, 2009

Note: See the Final Exam review sheet for the questions posed.

I.

A) The regression line is the line $y = mx + b$ in the plane that *minimizes* the rms error

$$\sqrt{\frac{1}{n}\left((y_1 - (mx_1 + b))^2 + \cdots + (y_n - (mx_n + b))^2\right)}.$$

(Among all lines the regression line makes the rms error the smallest.)

B) Martha is right. Assuming that the roulette game is fair, the spins of the roulette wheel are independent, and George's chance of winning with a bet on 17 is $\frac{1}{38}$ on *every spin*. The fact that he has not won for the last 50 spins does not change this.

C) Going to college does not change a man's height. What we can say is that there is a positive association between height and education level. (This could be due to confounding factors such as family income level which would tend to increase height through factors like better overall nutrition and medical care while the men were growing up, and which would also increase the chances of the men completing higher levels of education.)

D) False. The total population also grew over this period and, while it is not clear from just the data given here, the murder rate (murders/total pop.) actually *decreased* somewhat. (Total population in 1970: 203 million, in 1990: 248 million – more than a 20% increase.) *Comment:* Of course, for an exam response, I would not expect you to know the actual populations(!) The important thing to realize would be that one cannot conclude anything about the violence level of the society from just the numbers of crimes. The way to make valid comparisons is by the per capita rates.

E) The SD would probably be about 1 year. If the SD were one month, then it would be very unlikely that *any* entering first years were any age other than 18. Since there are always some people who enter college late or early compared to the average, this is not realistic. Similarly, if the SD was 5 years, then one would expect to see a significant number (maybe 34%) of entering first year students in the age range 13 to 18. Since that is also unrealistic, the best choice is "about 1 year."

F) By the binomial formula, the probability is

$$\binom{20}{7}\left(\frac{1}{2}\right)^7\left(\frac{1}{2}\right)^{13} = \frac{20!}{7!13!}\left(\frac{1}{2}\right)^{20}.$$

G) True: The education distribution has a long left tail, which means that the median is higher than the average.

H) Yes, something must be wrong. The average must be somewhere between 0 and 4, probably between 2 and 3. The rms error cannot be this large, since it is unlikely that any of the errors are as large as 3, so the rms error cannot be that large.

II.

A) This is True – a positive correlation coefficient indicates that there is a tendency for more near-sighted soldiers to also have higher intelligence (however that was measured).

B) This is False – correlation is not causation. A specific confounding factor might be the fact that the soldiers who scored higher on intelligence also did more of the reading and other close work with their eyes that leads to higher levels of near-sightedness.

III.

A) For these histograms, the best way to set up the class intervals is to take intervals of width 1 extending from $d-0.5$ to $d+0.5$ for each digit $d = 0, 1, 2, \ldots, 9$. The height of each histogram bar is then exactly the number given in the table for that digit (since the total area will then add up to 100%).

B) The most likely explanation was that in 1880, people did not keep as good records (not as many official birth certificates, school records, etc.), so they were more often unsure of their exact ages and tended to round up or down to ages that ended in 0 or 5.

C) In 1970, people were probably more aware of their exact ages due to better records, etc.

D) The even digits were more popular in 1880 (perhaps for the same sort of reason as in part B). There was little difference between the digits in 1970.

IV.

A) For the scatter plot, just plot the 6 points,

$$(x, y) = (4, 7), (5, 0), (7, 9), (8, 9), (8, 13), (10, 16).$$

The computation of the correlation coefficient is done as we discussed in class. First, we compute

$$\text{ave}_x = \frac{4+5+7+8+8+10}{6} = \frac{42}{6} = 7,$$

and

$$\text{ave}_y = \frac{7+0+9+9+13+16}{6} = \frac{54}{6} = 9.$$

Next we need the SD's:

$$SD_x = \sqrt{\frac{1}{6}((-3)^2 + (-2)^2 + 0^2 + 1^2 + 1^2 + 3^2)} = 2,$$

and

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

So the corresponding standard units values for the x and y data are

$$\begin{array}{llllll} z: -3/2 & -1 & 0 & 1/2 & 1/2 & 3/2 \\ w: -2/5 & -9/5 & 0 & 0 & 4/5 & 7/5 \end{array}$$

So the correlation coefficient is the average of the products $z_i \times w_i$:

$$r = \frac{1}{6}\left((-3/2)(-2/5) + (-1)(-9/5) + 0 + 0 + (1/2)(4/5) + (3/2)(7/5)\right) = \frac{49}{60} \doteq .8167.$$

B) The SD line passes through the point of averages and has slope SD_y/SD_x. The equation is $y - 9 = \frac{5}{2}(x-7)$, or

$$y = \frac{5}{2}x - \frac{17}{2}.$$

The regression line also passes through the point of averages, and has slope $\frac{r \cdot SD_y}{SD_x}$. The equation is $y - 9 = \frac{245}{120}(x-7)$, or

$$y = \frac{245}{120}x - \frac{127}{24}.$$

V. In the graph on page 267, (i) is the solid line (the one with least positive slope). This is the regression line for predicting values of y = verbal SAT score from x = math SAT score. (ii) is the dashed line (the one with the most positive slope). This is the regression line for predicting x from y. (iii) is the dotted line (which looks pretty close to the SD line for this data).

VI.

A) Write x for the English section score and y for the mathematics section score. Then $\text{ave}_x = 23.5$ and $SD_x = 5.6$, while $\text{ave}_y = 23.6$ and $SD_y = 5.2$. The regression line for predicting y from x has equation $y - 23.6 = \frac{(5.2)(.6)}{(5.6)}(x - 23.5)$, or approximately

$$y = (.589)x + 9.77.$$

Substituting $x = 31$, we get the predicted math section score: $y \doteq 28.0$. The SD for the scores of the the students who scored 31 would be estimated by the rms error for regression:

$$\text{rms error} = SD_y \times \sqrt{1 - r^2} = 5.2 \times \sqrt{1 - .36} = 5.2 \times .8 = 4.16.$$

B) We use the information from part A. The standard unit value corresponding to a score of 25 is $z = \frac{25-28}{4.16} \doteq -.72$. From the normal curve area table, linear interpolation gives

$$A(.72) \doteq A(.70) + (.02)\frac{A(.75) - A(.70)}{.05} = 51.61 + (.02) \times \frac{54.67 - 51.61}{.05} \doteq 52.83.$$

But this is the area under the normal curve between $-.72$ and $+.72$. We want the area in the lower tail which is $\frac{1}{2}(100 - 52.83) \doteq 23.59$. So about 23.6% of the students who scored 31 on the English portion had scores below 25 on the math exam. 23.6% of 350 is about 83 of them.

VII.

A) Out of the 8 cards in the stack, 2 are spades. Hence the chance of drawing 3 spades (when the draws are made with replacement) is

$$\frac{2}{8} \times \frac{2}{8} \times \frac{2}{8} = \left(\frac{1}{4}\right)^3 = \frac{1}{64}$$

(about 1.56%).

B) Note: This is not $1 - (1/4)^3$, since that counts draws where any one of the cards is not a spade. We want

$$\frac{6}{8} \times \frac{6}{8} \times \frac{6}{8} = \left(\frac{3}{4}\right)^3 = \frac{27}{64}$$

(about 42.2%).

C) *This* is the complementary probability to the probability from A: “at least one heart, diamond, or club” is the same as “not all spades,” so $1 - \frac{1}{64} = \frac{63}{64}$ (about 98.4%).

VIII.

A) The 70th percentile is the score such that 70% of the scores are less than or equal to that. This means that we are looking for a standard unit value z such that $A(z) +$ lower tail $= 70\%$, so $A(z) = 40\%$. From the normal table, this happens for z between .5 and .55. Interpolating to get a more exact value, $\frac{40-38.29}{z-.5} = \frac{41.77-38.29}{.05}$, so $z = .5 + \frac{40-38.29}{41.77-38.29}(.05) \doteq .525$. Then the corresponding score is $50 + (.525)(20) = 60.5$. Repeating these calculations for the 80th percentile, we get $z \doteq .85$ (a more accurate result could be obtained by interpolation here too, but the $A(.85)$ value is close to what we want, so we will just take it this time. The 80th percentile score is about $50 + (.85)(20) = 67$. The difference is about 6.5 points.

B) Reusing the 80th percentile score from part A, the first sister scored a 67. The second sister had a score of about $50 + 1.3 \times 20 \doteq 76$. so the difference in scores is about 9. (*Comment:* Note that the 90th percentile score is farther from the the 80th than the 80th is from the 70th because of the “bell-shape” of the normal curve.)

IX.

A) This can happen in two ways: either the 2 is from box (i) and the 5 is from box (ii), or vice versa. These events are mutually exclusive so the addition law gives:

$$\frac{1}{5} \times \frac{1}{6} + \frac{1}{5} \times \frac{1}{6} = \frac{1}{15}.$$

B) This can happen in several, mutually exclusive ways: $1+6, 2+5, 3+4, 4+3$, or $5+2$. So the chance is

$$\frac{5}{30} = \frac{1}{6}.$$

C) This can happen in several ways too (the first gives the draw from box (i) and the second gives the draw from box (ii): $5, 1$ $5, 2$, or $4, 1$ or $3, 1$, or $1, 6$, $2, 6$, or $1, 5$, $2, 5$, or $1, 4$, or $1, 3$. This gives a total chance of $\frac{10}{30} = \frac{1}{3}$.