MONT 106N – Identifying Patterns
Solutions for Sample Exam Questions
October 13, 2009

I.

A) What is a *confounding factor*? What are the techniques used in well-designed controlled experiments to try to minimize their effects?

Solution: A confounding factor is something that affects *both* a subject's likelihood of being in the treatment group in a study, and the outcome of the treatment. For instance, in part B below, one possible situation that would tend to explain the differences in the outcomes for the different surgeons would be if the younger doctor was consistently assigned the sicker patients, because he had more up-to-date knowledge of surgical techniques. Then the fact that more of his patients also died could also be a result. To avoid situations like this, the best-designed tests are done in a randomized, double-blind manner. The subjects are assigned to the treatment or control groups randomly, and who ends up in each group is unknown to the subjects and to the people carrying out the treatment. If appropriate, placebos are also used to minimize the possible psychological effects of patients thinking they are receiving treatment.

B) Here is a passage from Dr. Dean Edell's column in the San Francisco Chronicle of August 1, 1990:

"The more experienced a doctor is, the better. As obvious as that sounds, there are still too many people out there who never ask surgeons for a history of their work. The importance of knowing this is illustrated by this study. Peter Starek, a surgeon at the University of North Carolina, reviewed 460 heart valve replacement operations and found that only 4% of the patients of the three most senior surgeons died. But one junior surgeon lost almost 33% of his patients. Since that surgeon was technically the best in the group, says Starek, something was obviously lacking – perhaps the kind of good judgment that grows out of experience..."

1. Was Starek's study a controlled experiment or an observational study?

   Solution: This is definitely an observational study, since the results of previously performed operations are being "reviewed." There is no mention of randomization or a double-blind design.

2. Is Starek's claim justified by this evidence? Or are there possible confounding factors he is overlooking?

   Solution: There are many possible confounding factors. Note, for instance that the quote says that the young surgeon was "technically the best in the group." It is very possible that he was assigned the sickest patients with the most complicated heart valve problems, while the senior surgeons took the easier, less complicated cases.

II.

A) Find the equation of the line passing through the points $P = (2, 6)$ and $Q = (-5, 2)$ in the plane.

Solution: The slope is $m = \frac{6-2}{2-(-5)} = \frac{4}{7}$. So the equation of the line is $y - 6 = \dfrac{4}{7}(x - 2)$ in point-slope form, or $y = \dfrac{4}{7}x + \dfrac{34}{7}$ in slope-intercept form. (Either form is correct and acceptable!)

B) Find the equation of the line perpendicular to the line in part A and passing through the origin $O = (0, 0)$.

Solution: Perpendicular lines have slopes that are negative reciprocals of each other. So the line here is $y = \dfrac{-7}{4}x$.

III. The following data on people without health insurance by family income level for 2006 is from the Current Population Survey.

Family Income	No Insurance $N = 44,815$	All People $N = 293,834$
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.

Solution: The percentage of Less than \$25000 people in the No Insurance group, for instance is $\frac{18590}{44815} \times 100\% \doteq 41.5\%$. The others are as shown below:

Family Income	No Insurance $N = 44,815$	All People $N = 293,834$
Less than \$25,000	41.5%	24.0%
\$25,000 to \$49,999	30.4%	24.8%
\$50,000 to \$74,999	14.4%	18.8%
\$75,000 to \$200,000	13.7%	32.4%

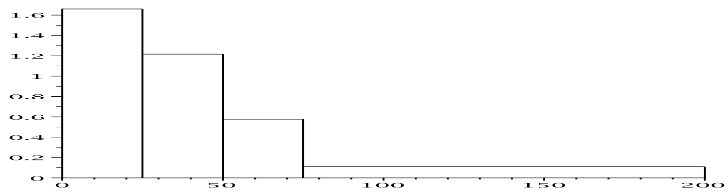

Figure 1: Histogram for III B – No Insurance

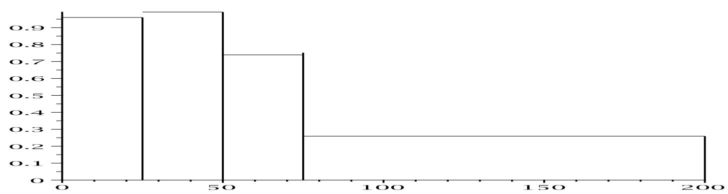

Figure 2: Histogram for III B – All People

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.)

Solution: See Figures 1 and 2.

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?

Solution: It is clear that people with no insurance tend to be concentrated in the lower income levels compared to the whole population.

D) Is the average income or the median income larger for the families with no insurance? Explain how you can tell.

Solution: The No Insurance group's income distribution has a long right tail, so the average income would be higher than the median income.

IV. 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.

Solution: See Figure 3.

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

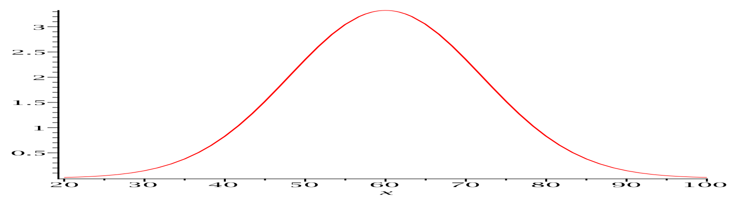

Figure 3: Histogram for IV A

Solution: In standard units, the range 70 to 80 is $z = \frac{70-60}{12} = .83$ to $z = \frac{80-60}{12} = 1.67$. The area under the normal curve between $z = .8$ and $z = 1.7$ is $\frac{1}{2}(91.09 - 57.63) = 16.73\%$ from the table. (This estimate is slightly larger than the actual area under the normal curve between $z = .83$ and $z = 1.67$ because we did not try to interpolate.) So an estimate for the number of people who scored in this range is 16.73% of 200, or about 33 students.

V. (Hypothetical) A big town near Worcester administers a spelling test to all students in fourth grade. The test consists of 100 words. The average for the number of words correctly spelled was 50 with a standard deviation of 10. Suppose a student was in the 70th percentile. What would you estimate for the number of words the student spelled correctly? (Be sure to show your calculations and draw normal approximations as needed.)

Solution: If z is the value in standard units giving the 70th percentile, the area under the normal curve to the left of z should be 70%. So the area $A(z)$ between $-z$ and z should satisfy $50 + \frac{1}{2}A(z) = 70\%$. This says $A(z) = 40\%$. Looking at the normal table, we see this happens for z between $z = .5$ and $z = .55$, probably close to $z = .525$. Then if c is the number of words correctly spelled, we want to determine the value of c that corresponds to $z = .525$ in standard units: $\frac{c-50}{10} = .525$, so $c \doteq 55$. (Note that this must be a whole number!)