Sociology 592 - Research Statistics I
Exam 1 Answer Key
September 28, 2001

Where appropriate, show your work - partial credit may be given. (On the other hand, don't waste a lot of time on excess verbiage.) Do not spend too much time on any one problem. It is legitimate (and probably essential) to refer to results that have previously been proven in class or homework, without re-proving them - for example, you wouldn't need to prove that P(-1.96 ≤ Z ≤ 1.96) = .95, since we have already shown that in class. Likewise, you are free to refer to anything that was demonstrated in the homework or handouts.

1. (5 points each, 20 points total). Indicate whether the following statements are true or false. If you think the statement is false, indicate how the statement could be corrected. For false statements, do not just say that you could substitute not equals for equals. For example, the statement P(Z ≤ 0) = .7 is false. To make it correct, don't just say P(Z ≤ 0) <> .7, instead say P(Z ≤ 0) = .5 or P(Z ≤ .525) = .7.

A. For a binomially distributed variable X, if N = 17 and p = .7, then P(13) = .1868.

TRUE. Reverse the definitions of success and failure. Then turn to Appendix E, Table II, and look up N = 17, p = .30, r = 4. Or, if you prefer, compute

$$\binom{N}{r} p^r q^{N-r} = \binom{17}{13} .7^{13} .3^4 = \frac{17!}{4!13!} .00007848 = 2380 * .00007848 = .1868$$

B. V(3X + 7) = 9 * V(X) + 49

FALSE. Adding a constant to all cases does not affect the variance. V(3X + 7) = 9 * V(X).

C. If A and B are mutually exclusive events, then P(A | B) = P(B | A).

TRUE. For mutually exclusive events, P(A | B) = P(B | A) = 0.

D. P(-.75 ≤ Z ≤ .75) = .77337272

FALSE. P(Z ≤ .75) = .77337272. Or, P(-.75 ≤ Z ≤ .75) = 2F(.75) – 1 = 0.54674544.

2. (10 points each, 30 points total) Answer <u>three</u> of the following. The answers to most of these are fairly straightforward, so do not spend a great deal of time on any one problem. NOTE: I will give up to 5 points extra credit for each additional problem you do correctly.

A. A long, hard-fought season is finally about to end. To no one's surprise, the national championship game features a rematch between the two teams that have dominated the sport all season long: The Nebraska Cornhuskers and The Notre Dame Fighting Irish (this is Women's College Soccer we are talking about, of course). The Huskers are eager to avenge their only loss of the season, a 1-0 shutout at Notre Dame on September 25. However, the Irish, fresh from a spectacular 7-0 semifinal win over hated archrival North Carolina, are determined to claim their second-ever national title.

This is the championship game so it cannot end in a tie; overtime will be played if necessary. Defense will be a key. There is a 60% chance that Notre Dame's stubborn defense will hold Nebraska to 2 goals or less. But, if Nebraska's high-powered offense does score 3 goals or more, there is a 70% chance the Huskers will win. Nebraska has a good defense too, so there is a 21% chance that the Huskers will score 2 goals or less and win the game.

What is the probability that Nebraska will win the game if it scores 2 goals or less? What is the probability that Nebraska will win the game? Finally, what is the probability that the winning team will score more goals than the losing team?

We are told:
P(2 goals or less) = .60, which implies P(3 goals or more) = .40
P(NU Winning | 3 goals or more) = .70
P(NU Winning $\cap$ Scoring 2 goals or less) = .21

Also, P(NU Winning $\cap$ scoring 3 goals or more) =
P(3 Goals or More) * P(NU Winning | 3 goals or more) = .40 * .70 = .28

So, P(NU Winning | 2 goals or less) =
P(NU Winning $\cap$ scoring 2 goals or less)/ P(2 goals or less) = .21/.60 = .35

P(NU Winning) =
P(NU Winning $\cap$ scoring 3 goals or more) + P(NU Winning $\cap$ scoring 2 goals or less) =
.28 + .21 = .49

P(Winning team scoring more goals than losing team) = 1.0 (otherwise it wouldn't be the winning team.)

B. N = 16, $\bar{x}$ = 32. Construct the 95% confidence interval when

- $\sigma = 4$
- s = 4

For σ = 4:

$$\bar{x} \pm (z_{\alpha/2} * \sigma/\sqrt{N}), \text{ i.e.,}$$
$$32 - (1.96 * 4/4) \le \mu \le 32 + (1.96 * 4/4), \text{ i.e.,}$$
$$30.04 \le \mu \le 33.96$$

For s = 4, look at Appx E, Table 3, v = 15, 2Q = .05:

$$\bar{x} \pm (t_{\alpha/2,v} * s/\sqrt{N}), i.e.$$
$$32 - (2.131 * 4/4) \le \mu \le 32 + (2.131 * 4/4), i.e.$$
$$29.869 \le \mu \le 34.131$$

C. Here are the results from a previous cohort's first exam in statistics. Compute the mean and variance of the scores. There were 9 Students in the class.

Score	Frequency
41	1
72	1
83	1
90	1
94	1
97	1
102	1
108	2

Let's expand the table as follows:

X_i	f_i	$X_i * f_i$	$(X_i - \mu)^2$	$(X_i - \mu)^2 f_i$	X_i^2	$X_i^2 f_i$
41	1	41	2240.44	2240.44	1681	1681
72	1	72	266.78	266.78	5184	5184
83	1	83	28.44	28.44	6889	6889
90	1	90	2.78	2.78	8100	8100
94	1	94	32.11	32.11	8836	8836
97	1	97	75.11	75.11	9409	9409
102	1	102	186.78	186.78	10,404	10,404
108	2	216	386.78	773.56	11,664	23,328
Σ	9	795		3,606.00		73,831

So, μ = 795/9 = 88.333. σ^2 = 3606/9 = 400.67, σ = 20.02.
Or, if you prefer, $\sigma^2 = E(X^2) - E(X)^2$ = 73,831/9 – 88.333^2 = 400.67, σ = 20.02.

D. A company wants to cut its payroll. It has decided to offer early retirement incentives to those managers whose salaries put them in the top 10% of the pay scale. If Salary ~ N(\$65000, 16000^2), how high does your salary have to be to qualify for the early retirement plan?

Turning to Appx E, Table 1, we see the critical value for Z is about 1.28 (1.29 is also ok). So,

$$x = z\sigma + \mu = 1.28 * \$16{,}000 + \$65{,}000 = \$85{,}480 \, [\$85640 \text{ if you use } 1.29]$$

E. A researcher has constructed a questionnaire that measures depression on a scale ranging from 0 to 500, where 0 = Not depressed and 500 = Extremely depressed. She is now preparing to administer this questionnaire to a cross-section of the population. The agency funding her research wants sample estimates to be very precise; thus, the true standard error of the mean must be no greater than 5. If, in the population to be studied, σ = 90, what is the minimum size her sample needs to be? How big would her sample need to be if she wanted the true standard error of the mean to be no greater than 1? Based on your results, briefly discuss how the desired level of accuracy in a study can affect data collection costs.

The True Standard Error of the Mean = $\sigma/\sqrt{N}$. Ergo, if the True SE = 5 and σ = 90,

$$\sigma/\sqrt{N} = 90/\sqrt{N} = 5, \text{Thus}$$
$$18 = \sqrt{N} \text{ and hence } N = 324$$

If the True SE = 1,

$$\sigma/\sqrt{N} = 90/\sqrt{N} = 1, \text{Thus}$$
$$90 = \sqrt{N} \text{ and hence } N = 8100$$

So, the more precise you want your estimates to be, the larger your sample will have to be and hence the greater your data collection costs will be. The researcher would have to collect a sample 25 times as large if she wanted the True SE to be 1 rather than 5. Both the researcher and the funding agency have to carefully consider whether the added precision is really necessary or not given the substantially greater costs.

3. (25 points) A Sociologist and a Biologist are both concerned about the increasing prevalence of health problems such as obesity and diabetes. They believe that these problems are related to how much one exercises as a child, and that exercise may in turn be related to the race and income of individuals. They have decided to team up and do a study of racial and economic differences in children's exercise. They have gathered data on the exercise habits and income of 1000 black and 1000 white children . Their study reveals that 70% of the black children have low incomes and the remainder have high incomes. For whites, 20% have low incomes and the rest have high incomes. Among blacks, 40% of those with low incomes regularly exercise, compared to 60% of the high-income blacks. For whites, 35% of those with low incomes regularly exercise, compared to 55% of the high-income whites.

a. Finish filling in the numbers for the following table. Remember that, as is already noted in the table, there are a total of 1000 blacks and 1000 whites. [HINT: If you don't find that more whites exercise than do blacks, you've done something wrong.]

	Black			White		
Exercise/Income	Low Income	High Income	Σ	Low Income	High Income	Σ
Exercises regularly						
Does not exercise regularly						
Σ			1000			1000

For blacks, we are told P(Low Income) = .70, implying 700 low income blacks and 300 high income blacks. Also, P(Exercise | Low Income) = .40, which means that 40% of the 700 low income blacks, or 280, exercise regularly, while the other 420 do not. Also, P(Exercise | High Income) =.60, which means that 60% of the 300 high income blacks, or 180, exercise regularly, while the other 120 do not.

For whites, we are told P(Low Income) = .20, implying 200 low income whites and 800 high income whites. Also, P(Exercise | Low Income) = .35, which means that 35% of the 200 low income whites, or 70, exercise regularly, while the other 130 do not. Also, P(Exercise | High Income) =.55, which means that 55% of the 800 high income whites, or 440, exercise regularly, while the other 360 do not. We therefore get

	Black			White		
Exercise/Income	Low Income	High Income	Σ	Low Income	High Income	Σ
Exercises regularly	280	180	460	70	440	510
Does not exercise regularly	420	120	540	130	360	490
Σ	700	300	1000	200	800	1000

b. Looking at the above table, the Biologist argued that blacks are less likely to exercise than are whites. The Sociologist, however, argued that blacks are more likely to exercise than are comparable whites. Briefly explain the evidence that supports each of their positions.

The biologist is correct, in that only 46% of all Blacks exercise regularly, compared to 51% of all whites. However, the Sociologist is also correct: 40% of all low income blacks exercise regularly, compared to only 35% of all low income whites. And, 60% of all high income blacks exercise regularly, compared to only 55% of all high income whites. That is, for blacks and whites of comparable incomes, blacks are somewhat more likely to exercise. However, blacks are much more likely to be low income, and low income people are less likely to exercise, hence blacks as a whole exercise less than whites do.

c. As these figures show, blacks are generally poorer than whites, and poor people tend to exercise less than do wealthier people. Suppose that blacks had the same income distribution as whites, i.e. 20% of blacks were low-income and the rest were high-income. Suppose further that it continued to be the case that blacks maintained their income-specific exercise rates, i.e. 40% of the poor blacks and 60% of the rich blacks exercised regularly. What percentage of blacks would then exercise regularly?

If blacks had the same income distribution as whites (20% low income, 80% high income) while maintaining their income-specific rates of exercise (40% for low income, 60% for high income), then the probability of a black exercising would be

$$P(Exercise) = P(LowInc \cap Exercise) + P(HighInc \cap Exercise) =$$
$$P(LowInc)P(Exercise \mid LowInc) + P(HighInc)P(Exercise \mid HighInc) =$$
$$(.2*.4) + (.8*.6) = .56$$

So, if blacks had the same income distribution as whites, they would actually exercise more than whites currently do. This implies that it is their lower incomes, rather than their race, that causes blacks to exercise less overall than whites do.

4. (25 points) An association has reassured a hotel that, despite recent events, no more than 15% of its members plan to cancel their reservations for an upcoming national convention. The hotel, of course, fears otherwise. A random sample of 81 association members reveals that 20 are going to cancel their reservations. Test the association's claim at the .01 level of significance. Be sure to indicate:

(a) The null and alternative hypotheses - and whether a one-tailed or two-tailed test is called for.

(b) The appropriate test statistic

(c) The critical region

(d) The computed value of the test statistic

(e) Your decision - should the null hypothesis be rejected or not be rejected? Why?

(a) The Null and Alternative Hypotheses are

H_0: p = .15 [or, E(X) = 12.15]
H_A: p > .15 [or, E(X) > 12.15]

A one-tailed alternative is appropriate. The hotel won't be upset if the cancellation rate turns out to be less than 15%, but it will be upset if it is more than that.

(b) The appropriate test statistic is

$$z = \frac{\#\ of\ cancellations \pm CC - Np_0}{\sqrt{Np_0 q_0}} = \frac{x \pm CC - (81*.15)}{\sqrt{81*.15*.85}} = \frac{x \pm CC - 12.15}{3.214}$$

For the correction for continuity, we will subtract .5 if there are more than 12.15 cancellations, we will add .5 if there are less than 12.15 cancellations.

Alternatively, we could compute

$$z = \frac{\hat{p} \pm .5/N - p_0}{\frac{\sqrt{p_0 q_0}}{\sqrt{N}}} = \frac{\hat{p} \pm .5/N - p_0}{\sqrt{\frac{p_0 q_0}{N}}} = \frac{\hat{p} \pm .5/81 - .15}{\sqrt{\frac{.15 * .85}{81}}} = \frac{\hat{p} \pm .5/81 - .15}{.039675}$$

(c) For the critical region, we will reject H_0 if $Z_c > 2.33$

(d) The computed value of the test statistic is

$$z = \frac{\#\ of\ \ cancellations \pm CC - Np_0}{\sqrt{Np_0 q_0}} = \frac{x \pm CC - 12.15}{3.214} = \frac{20 - .5 - 12.15}{3.214} = 2.29$$

Or, equivalently,

$$z = \frac{\hat{p} \pm .5/N - p_0}{\frac{\sqrt{p_0 q_0}}{\sqrt{N}}} = = \frac{\hat{p} \pm .5/81 - .15}{.039675} = \frac{.24691358 - .5/81 - .15}{.039675} = 2.29$$

(e) Decision: Do not reject the null. The computed test statistic falls just barely within the acceptance region. (I imagine the hotel is probably still going to feel a bit nervous though.)