Sociology 592 - Research Statistics I
Exam 2 Answer Key
November 7, 2003

1. (10 points each, 30 points total.) You have been asked to serve as a statistical consultant for several proposed projects. For each of the following, your employers want you to tell them:

(i) Which of the cases we have studied their problem falls under (e.g. one sample tests, case I, σ known; nonparametric tests, case II, tests of association). Briefly explain why.
(ii) the null and alternative hypotheses
(iii) whether a Z, T, chi-square, or F test is appropriate; where applicable, also tell what the degrees of freedom for the test are. You DO NOT have to give the formula for the test statistic, nor do you need to specify the acceptance region.

If values for population parameters are not specified (e.g. σ) assume they are unknown; and if two or more unknown σ's are involved, assume they are equal.

a. Notre Dame wonders what causes some graduate students to take longer to finish than others do. It suspects that field of study may have something to do with it. It will therefore draw random samples of 30 recent Ph.Ds each from the humanities, social sciences, and science and engineering. For each Ph.D., the number of years spent in graduate school will be determined.

One-way Anova. The dependent variable is continuous (Number of years to complete degree) and the independent variable is categorical (field of study). The null and alternative hypotheses are

H_0: $\mu_1 = \mu_2 = \mu_3$
H_A: The means are not all equal

Note that J = 3 and N = 90. Use an F test with DF = 2, 87.

b. A professor being considered for tenure needs proof that students are learning something from him. On the first day of his Introductory Sociology class, his 100 students will be given an exam that measures general knowledge of Sociology (like most exams, the possible scores range from 0 to 100). Students will not be told how they scored or what the correct answers were. On the last day of class, the students will once again take this same exam.

2 sample tests, case IV, matched pairs. The same people's knowledge of the material is measured both at the beginning and the end of the course. The null and alternative hypotheses are

H_0: $\mu_1 = \mu_2$
H_A: $\mu_1 < \mu_2$

The alternative is one-tailed because the professor believes (or at least hopes) that students learn something from his course. A T-test with 99 d.f. is appropriate.

c. A prosecutor has several domestic violence cases coming up. To aid her with her jury selection, she wants to know if she is correct in believing that women care more about domestic violence than men do. A sample of 60 men and 60 women will be drawn. Each person will be asked whether or not they feel domestic violence is an important problem.

2 sample tests, case 5, test of $p_1 = p_2$. Note that nonparametric tests, case II, is not appropriate, because the alternative is one-sided. The null and alternative hypotheses are

H_0: $p_W = p_M$
H_A: $p_W > p_M$

A Z statistic is appropriate.

2. (5 points each, 20 points total). For <u>each</u> of the following, indicate whether the statement is true or false. If you think the statement is false, indicate how the statement could be corrected.

NOTE: These are all pretty easy, but you could waste a great deal of time on some of them or make stupid mistakes if you don't happen to see what the easiest way to approach each problem is.

a. A researcher has a sample consisting of 522 subjects. For each subject, she has collected information on 3 variables, each of which has 3 categories. When she tests the model of conditional independence for these three variables, she gets a chi-square value of 27. If she is using the .05 level of significance, she should reject the model of conditional independence.

TRUE. The degrees of freedom are rcl – 1 – (rc – 1) – (l – 1) = 27 – 1 – 8 – 2 = 16. For a chi-square with 16 d.f., the critical value at the .05 level of significance is 26.296. Since 27 is larger than this, reject the null; at least one of her IVs is not independent of her DV.

b. The null and alternative hypothesis are

H_0: $p = .4$
H_A: $p < .4$

Data are collected from 115 cases. The researcher computes a chi-square test statistic value of -4. Using the .05 level of significance, she should reject the null.

FALSE. Since the alternative is one-tailed, she should be using a Z statistic, not chi-square. And, if she is going to use chi-square, she should at least compute it correctly; negative values for chi-square are not possible.

c. A researcher has collected data from 36 respondents on their religion (Catholic or non-Catholic), gender (male or female), and support for the Democratic Party (measured on a scale that ranges from -80 to +80). She computes

$$F_{J-1,N-JK} = \frac{SS\ Rows/(J-1)}{SS\ Error/(N-JK)} = \frac{MS\ Rows}{MS\ Error} = 347$$

If religion is her row variable, she should conclude that religion affects support for the Democratic Party but gender does not.

FALSE. Religion does indeed have a significant effect on support for the Democratic Party. But, the above test tells you nothing about the effect of gender; you have to do a separate test for that.

d. A researcher believes that a problem falls under 2 sample tests, case II, σ_1 and σ_2 are unknown and assumed equal. In reality, the problem falls under 2 sample tests, case IV, matched pairs. Fortunately, so long as N1 = N2, both approaches will yield the same computed value for the test statistic although the degrees of freedom for the two tests may differ.

FALSE. As we saw in our handouts, ignoring the fact that your samples are not independent will probably cause your test statistic to be incorrect.

Answer two of the following three questions. (25 points each; you will get up to 10 points extra credit if you answer all three correctly.)

3. In recent years, more and more Americans have turned to non-traditional, alternative forms of health care, such a chiropractic care and acupuncture. Medical researchers are interested in determining whether the use of such alternative methods is related to education. One hundred randomly selected individuals are asked whether or not they use alternative forms of healthcare. Their survey yields the following information:

ALTRNATV Does respondent use alternative health care? * EDUC Level of Education Crosstabulation

Count

		EDUC Level of Education			
		1.00 High	2.00 Medium	3.00 Low	Total
ALTRNATV Does respondent use alternative health care?	1.00 No	10	25	30	65
	2.00 Yes	15	15	5	35
Total		25	40	35	100

Using our five-step hypothesis testing procedure and the .05 level of significance, determine whether or not education is related to the use of alternative medicine. If there is a significant relationship, use the information you have been given to explain exactly what you think that relationship is, i.e. how does the use of alternative health care appear to differ by levels of education?

Step 1.

H_0: The use of alternative health care does not differ by education
H_A: The use of alternative health care does differ by education

or, equivalently,

H_0: $P(A_i \cap B_j) = P(A_i)P(B_j)$ (Model of independence)
H_A: $P(A_i \cap B_j) <> P(A_i)P(B_j)$ for some i, j

Step 2. An appropriate test statistic is

$\chi^2_v = \Sigma\Sigma (O_{ij} - E_{ij})^2/E_{ij}$, $v = rc-1-(r-1)-(c-1) = (r-1)(c-1) = 2$

Step 3. For $\alpha = .05$ and $v = 2$, accept H_0 if $\chi^2_2 \leq 5.99$

Step 4. The computed value of the test statistic is:

Educ/Altrnatv	Observed	Expected	$(O_{ij} - E_{ij})^2/E_{ij}$
High No	10	.25 * .65 * 100 = 16.25	$-6.25^2/16.25 = 2.404$
Hi Yes	15	.25 *. 35 * 100 = 8.75	$6.25^2/8.75 = 4.464$
Medium No	25	.40 * .65 * 100 = 26	$-1^2/26 = .038$
Medium Yes	15	.40 * .35 * 100 = 14	$1^2/14 = .071$
Low No	30	.35 * .65 * 100 = 22.75	$7.25^2/22.75 = 2.310$
Low Yes	5	.35 * .35 * 100 = 12.25	$-7.25^2/12.25 = 4.291$

So, $\chi^2_v = \Sigma\Sigma (O_{ij} - E_{ij})^2/E_{ij} = 13.58$.

Step 5. Reject H_0, the computed test statistic falls outside the acceptance region.

The chi-square value is highly significant, meaning that use of alternative health care varies by education level. Looking at the original table, we see that 60% of the people with high levels of education use alternative care, compared to 37.5% of those with medium levels of education and 14.3% of those with low levels of education. Therefore, it appears that, the higher the level of education, the more likely it is that a person will use alternative health care.

To confirm our calculations, here is a more complete printout from SPSS:

ALTRNATV Does respondent use alternative health care? * EDUC Level of Education Crosstabulation

			EDUC Level of Education			
			1.00 High	2.00 Medium	3.00 Low	Total
ALTRNATV Does respondent use alternative health care?	1.00 No	Count	10	25	30	65
		Expected Count	16.3	26.0	22.8	65.0
		% within EDUC Level of Education	40.0%	62.5%	85.7%	65.0%
	2.00 Yes	Count	15	15	5	35
		Expected Count	8.8	14.0	12.3	35.0
		% within EDUC Level of Education	60.0%	37.5%	14.3%	35.0%
Total		Count	25	40	35	100
		Expected Count	25.0	40.0	35.0	100.0
		% within EDUC Level of Education	100.0%	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	13.579[a]	2	.001
Likelihood Ratio	14.206	2	.001
Linear-by-Linear Association	13.442	1	.000
N of Valid Cases	100		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 8.75.

4. A school district is very concerned by its low mean score of 60 on an achievement test of its 8th graders. An outside firm contends that it can do a better job with the district's students than the current schools are. The district and the firm agree to conduct a 1 year pilot study to test the firm's claim. Thirty randomly selected eighth graders are enrolled in the firm's program. At the end of the year the firm's students get an average score of 67 on the achievement tests with a sample standard deviation of 20.

a. Using our 5-step hypothesis testing procedure and the .05 level of significance, determine whether the firms' claim that it does a better job is supported.

This falls under single sample tests, case 3, σ unknown.

Step 1:

H_0: $\mu = 60$
H_A: $\mu > 60$

The alternative is one-tailed because the firm claims it can do a better job of educating students than the schools are.

Step 2: The appropriate test statistic is

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{N}}} = \frac{\bar{x} - 60}{\frac{s}{\sqrt{30}}}$$

with d.f. = N – 1 = 29.

Step 3. Reject H_0 if $t_c \geq 1.699$

Step 4. The computed value of the test statistic is

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{N}}} = \frac{\bar{x} - 60}{\frac{s}{\sqrt{30}}} = \frac{67 - 60}{\frac{20}{\sqrt{30}}} = \frac{7}{3.651} = 1.917$$

Step 5. Reject the null. The firm does appear to do a better job of educating students than the current schools do.

b. The teachers' union opposes hiring the outside firm. To support its argument, it notes that the district's score of 60 falls within the 95% confidence interval. Is the union correct in saying that the confidence interval includes 60? If so, is its argument valid? Why or why not?

Note that the critical value of T to use is 2.045. The 95% confidence interval is

$$\bar{x} \pm (t_{\alpha/2,v} * s/\sqrt{N}), i.e.$$
$$67 - (2.045 * 20/\sqrt{30}) \leq \mu \leq 67 + (2.045 * 20/\sqrt{30}), i.e.$$
$$59.53 \leq \mu \leq 74.467$$

The teachers' union is correct in that 60 does fall within the confidence interval. Nonetheless, its argument is invalid. Confidence intervals can be used for hypothesis testing when the alternative is 2-tailed but should not be used when the alternative is 1-tailed. The firm did not claim that its students would get different scores, it said they would get better scores.

We can confirm that we got the calculations right by using Stata. The parameters for the ttesti command are N, the sample mean, the sample standard deviation, and the hypothesized mean.

```
. ttesti 30 67 20 60

One-sample t test

------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      30          67    3.651484          20    59.53188    74.46812
------------------------------------------------------------------------------
Degrees of freedom: 29

                              Ho: mean(x) = 60

     Ha: mean < 60            Ha: mean ~= 60              Ha: mean > 60
       t =   1.9170              t =   1.9170               t =   1.9170
   P < t =   0.9674          P > |t| =   0.0651           P > t =   0.0326
```

5. With his fundraising continuing to go well and support from labor unions starting to grow, Howard Dean is more confident than ever that he will be the Democratic nominee for President. Nonetheless, he wonders whether his support differs across key demographic groups. Specifically, he wants to know whether his support differs by race (coded as white, black, other) and age (coded as young, middle-aged, and elderly). For each combination of race and age, 20 voters are interviewed. Support for Dean is measured on a scale that ranges from a low of -30 to a high of 30. The survey shows that support for Dean has a mean of 5 and a variance of 125.

a) Complete the following Anova table. You do NOT need to indicate whether the F values are statistically significant or not.

Source	SS	D.F.	M. S.	F
A + B (or Main Effects)				
A (Race of voter)			100	
B (Age)	180			
AB (or 2-way interaction)				
A + B + AB (or explained)				
Error (or residual)	1710			
Total				

Note that MST = s^2 = 125, J = K = 3. Since 20 voters are selected from each of the 9 possible combinations of Race and Age, N = 180. Once we fill in the numbers we know the rest of the calculations are fairly easy:

Source	SS	D.F.	Mean Square	F
A + B (or Main Effects)	SS Main = 380	J + K - 2 = 4	SS Main / (J + K - 2) = 95	MS Main / MS Error = 9.5
A (Race of Voter)	SS Rows = 200	J - 1 = 2	SS Rows / (J - 1) = **100**	MS Rows / MS Error = 10
B (Age)	SS Columns = **180**	K - 1 = 2	SS Columns / (K - 1) = 90	MS Columns / MS Error = 9
AB (or 2-way interaction)	SS Intraction = 20285	(J - 1) * (K - 1) = 4	SS Intrction / (J -1)(K - 1) = 5071.25	MS Intrction / MS Error = 507.125
A + B + AB (or explained)	SS Cells = 20665	(J * K) - 1 = 8	SS Cells / (J * K) - 1 = 2583.125	MS Cells / MS Error = 258.3125
Error (or residual)	SS Error = **1710**	N - (J * K) = 171	SS Error / (N - J * K) = 10	
Total	SS Total = 22,375	N - 1 = 179	SS Total / (N - 1) = **125**	

b) Explain what significant interaction terms might mean. Be specific; don't just talk about interaction terms in general, rather, talk about what interactions involving the variables in this analysis might be due to.

One possibility is that the relationship between age and support for Dean differs by race (perhaps greatly so, given the F values). Among whites, the older someone is, the more likely they may be to support Dean. Among blacks, just the opposite might be true: the older someone is, the less likely they are support Dean. Older whites and younger blacks might be more inclined to support a candidate who claims to challenge the establishment the way Dean does. Older blacks might prefer a more traditional candidate, with long ties to the civil rights movement, while younger whites might prefer someone with a military background like Wesley Clark. Then again, maybe not; something else entirely could be going on. You would have to examine the data more closely to determine this.