

# POS 416z - Research Methods in Political Science I

Fall 2001.

This test booklet contains spaces for your answers. Please put everything you want me to see on these pages. Use the backs of pages if necessary to show work if you do not have room in the spaces provided.

Pages of probabilities in the normal, chi-square, and t distributions are attached at the end of this test booklet for your reference.

*I am aware of existing program, college, and university policies on academic dishonesty. My work submitted on this examination is in compliance with those policies.*

Signed \_\_\_\_\_

Student ID number \_\_\_\_\_

Name \_\_\_\_\_

*Please answer each of the following on these test pages in the blanks or spaces provided. Show any significant work (except arithmetic) for part credit. A calculator is recommended. Continue answers on the backs of these test pages if necessary. Statistical tables are attached for your use.*

1) [12 points total] In a recent poll [Christian Science Monitor, 11 Dec 2001], 921 Americans were asked if they agreed or disagreed with the following statement:

“The Bush administration is spending too much time on the war and needs to give more attention to boosting the economy.”

Forty six percent agreed; 52 percent disagreed; 2 percent were undecided.

(a) [4 points] What is the margin of error of this poll at the 95 percent confidence level?

(b) [4 points] Is it possible that the true results for the population are the reverse of these in the sample, that 52 percent of the whole population would agree while 46 percent would disagree? Explain.

(c) [4 points] How many people would we have to ask to achieve a margin of error of plus or minus 2 percent?

2) [18 points total] Multiple regression was used to study the possible effects of air pollutants on urban mortality in 59 statistical metropolitan areas in the United States. The study addressed the question of whether air pollution contributes to urban mortality (measured in deaths per 100000 per year).

Sam noticed that six of the cities in the data had *zero* concentrations of one or more of the potential pollutants NO<sub>x</sub>, or HC, or SO<sub>2</sub> in their air. He decided to remove those cities from the analysis, and to perform multiple regressions on the reduced set of 53 cities that all had at least some concentrations of each of the potential air pollutants.

Before including the potential pollution variables, Sam's favorite regression equation contained two climate variables (January temperature and Rain) and two demographic variables (% NonWhite and % White Collar), which produced an R-squared of 61.1 percent.

Part of his regression printout is shown below:

**Regression Coefficients  
Mortality vs. 4 Independents**

	Coefficient	Std. Error	Std. Coeff.	t-Value	P-Value
Intercept	993.636	54.343	993.636	18.284	<.0001
JanTemp	-2.020	.611	-.328	-3.308	.0017
Rain	1.163	.484	.216	2.404	.0197
%NonWhite	5.017	.705	.723	7.120	<.0001
%WC	-1.910	1.088	-.155	-1.755	.0850

(a) [4 points] Write the estimated regression equation for Mortality implicit in this table.

(b) [4 points] Use information in this table to find a 95 percent confidence interval for the true (population) coefficient for Rain. [Show your work (but not your arithmetic) for part credit, just in case you make a mistake somewhere.]

(c) [6 points] What can you conclude from the t-values and the P-values shown in this table?

(d) [4 points] The value of R-squared in this regression is 61.1 percent. What does that mean? And what does that tell you?

3) [14 points total] Sam then added the three pollutants (NO<sub>x</sub>, HC, and SO<sub>2</sub> concentrations) to the regression and obtained an R-squared of 73.7 percent and the following table of coefficients.

**Regression Coefficients**  
**Mortality vs. 7 Independents**  
**Row exclusion: airpolt.dat (imported)**

	Coefficient	Std. Error	Std. Coeff.	t-Value	P-Value
Intercept	917.760	46.293	917.760	19.825	<.0001
JanTemp	-1.060	.792	-.158	-1.339	.1874
Rain	1.043	.532	.210	1.959	.0564
%NonWhite	3.542	.599	.555	5.914	<.0001
%WC	-1.616	.861	-.148	-1.877	.0670
NOxPot	18.318	37.448	.151	.489	.6271
HCPot	-33.423	34.102	-.268	-.980	.3323
SO2Pot	50.176	17.919	.425	2.800	.0075

(a) [6 points] Discuss what you you learn from these facts when compared to those in (2).

(b) [4 points] Why did Sam include the climate and demographic variables in his final regression? If the focus was on pollution, why not just include the terms for the pollution concentrations?

(c) [4 points] Why did Sam do the regression in question (2)? Why not just do the regression in (3)?

4) [20 points total] The influence of race on the imposition of the death penalty has been studied a lot and repeatedly contested in the courts. The table at the right is court data showing the imposition of the death penalty in 326 murder trials, broken down by the race of the defendant and the race of the victim.

White defendant			Black defendant		
	Death Penalty			Death Penalty	
	Yes	No		Yes	No
White victim	19	132	White victim	11	52
Black victim	0	9	Black victim	6	97

(a) [4 points] Use the above data to fill in the table below to make a two-way table of Death Penalty by Defendant's Race:

		Race of defendant	
		White	Black
Death penalty	Yes		
	No		

(b) [4 points] Show that Simpson's paradox is at work here: that a higher percentage of white defendants are sentenced to death overall (second table), but for both black and white victims separately a higher percentage of black defendant's are sentenced to death (first table).

(c) [6 points] Explain how this apparent paradox occurs in this data, in language a judge could understand. Which conclusion should he believe, if any?

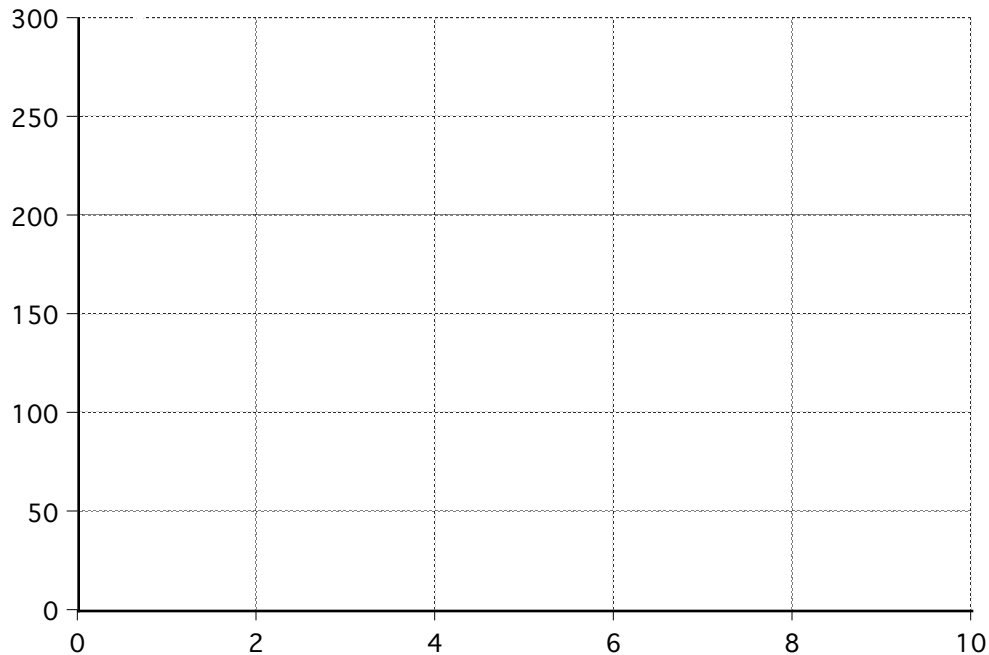
(d) [6 points] If one computes correctly the expected values that would appear in the table you constructed in (a), and then computes the chi-square statistic, one finds a value of 0.221. Use the chi-sq table attached to this exam to determine if there is a significant association in your table between the race of the defendant and the imposition of the death penalty. [Be sure to indicate how many degrees of freedom are involved here.]

(5) [24 points total] The table at the right shows data on the consumption of alcohol from wine (liters of alcohol from wine per person per year) and deaths from heart disease (deaths per 100,000 people).

(a) [4 points] Draw a stem-and-leaf plot in the space below for the Heart disease deaths per 100,000. [Pick stems and leaves carefully so you can show the distribution nicely. You might do a draft on the back of one of these pages and then put your pretty version here.]

Wine consumption and heart attacks		
<i>County</i>	<i>Alcohol from wine</i>	<i>Heart disease deaths</i>
Australia	2.5	211
Austria	3.9	167
Belgium	2.9	131
Canada	2.4	191
Denmark	2.9	220
Finland	0.8	297
France	9.1	71
Iceland	0.8	211
Ireland	0.7	300
Italy	7.9	107
Netherlands	1.8	167
New Zealand	1.9	266
Norway	0.8	227
Spain	6.5	86
Sweden	1.6	207
Switzerland	5.8	115
United Kingdom	1.3	285
United States	1.2	199
West Germany	2.7	172

(b) [4 points] Draw a scatterplot of this data, with Heart disease deaths on the y-axis and Alcohol consumption from wine on the x-axis. Label your axes. [You might do a draft on the back of one of these pages; put your final pretty version here.]



(c) [6 points] Sketch on your scatterplot in (b) the approximate regression line that would best fit this data. Determine from the plot the equation of the line you drew. [You are *not* to try to compute the regression line, or use a spiffy calculator to find it. Just draw in the approximate line, determine its slope and intercept, and then write its equation.]

(d) [4 points] Use the slope of the line you drew to compute the correlation coefficient here. You will want to know that the standard deviation of the Alcohol data is 2.5 and the standard deviation of the Heart attack data is 68.4. [Hint: the relation between the correlation  $r$  and the slope  $b$  is either

$r = b \cdot \frac{S_y}{S_x}$  or  $r = b \cdot \frac{S_x}{S_y}$ ; you'll have to figure out which. If you couldn't find the slope in (c), just pick a fake number for it and answer (d).]

(e) [6 points] Does this data convince you that consuming moderate amounts of alcohol in wine is a good way to ward off heart disease? Explain thoughtfully.