

POS 416Z - Research Methods for Political Science and Public Policy

Fall 2002.

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 Z, chi-square, t and F distributions are attached at the end of this test booklet for your reference.

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

Signed _____

Student ID number _____

1) [10 points] Mr. Bean was analyzing effects on mortality in 59 standard metropolitan statistical areas (SMSA's) and performed a multiple regression of Mortality versus January Temperature, HydroCarbon Potential, and Population per Housing Unit. He FAXed you the printout below, but unfortunately some of the numbers were unreadable. Undaunted, you realize you can compute them from the information remaining.

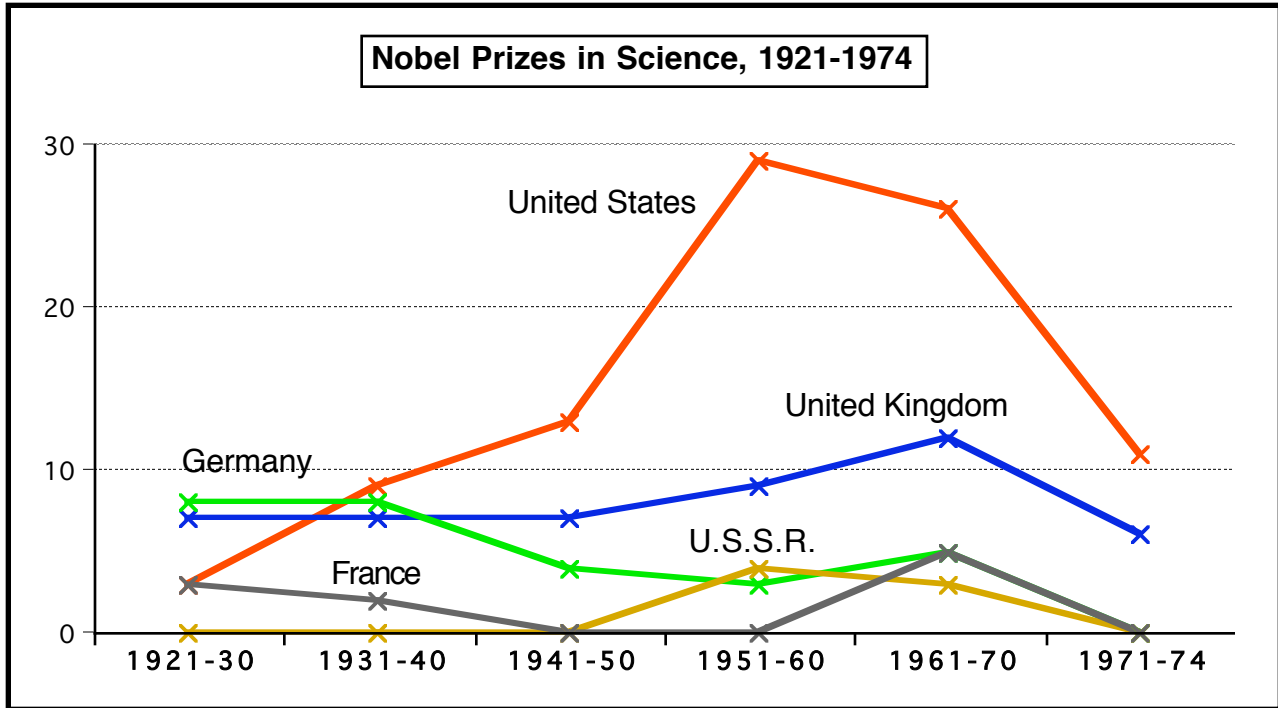
(1a) [8 points] Compute the missing information and fill in the eight blank boxes in the regression output. (Ignore the shaded regions!) [You will find t-tables and F-tables at the end of the exam.]

SUMMARY OUTPUT					
Regression Statistics					
R Square					
Observations	59				
ANOVA					
	df	SS	MS	F	Significance F
Regression	3	42095.1		4.20	
Residual	55				
Total		225992.5			
	Coefficients	Standard Error	t Stat	P-value	
Intercept	405.220	158.234			
JanTemp	0.494		0.6158	0.54	
HCPot	26.151	16.149			
pop/house	150.141	44.483			

(1b) [2 points] What is the estimated regression equation in this regression? [Don't bother interpreting significances (we'll get to that later), just write the equation. Please use words instead of X's and Y's.]

(2) [6 points] Even the best of us make graphical mistakes sometimes and mislead our audience. Below is a graph from *Science Indicators, 1974*, published by the National Science Foundation, which contains a rather glaring error that gives a misleading visual impression about the number of Nobel prizes awarded in science over the years 1921 to 1974. [Graph from Tufte, *The Visual Display of Quantitative Information*, p. 60]

(2a) [2 points] Without worrying about charting errors, what do you observe from this chart about the number of Nobel prizes in science awarded from 1921 to 1974? [Write your answer below the chart.]



(2b) [4 points] What serious mistake would you say the NSF chart maker made in setting up this chart? [Suggestion: Look at the thing(s) you observed in (2a) and find one of them that is an accident from the way the chart is set up.]

(3) [12 points] In 1974 the national speed limit was lowered to 55 in an effort to conserve gasoline after the 1973 Mideast war. Then in the mid 1980s most states raised speed limits on interstate highways to 65. Some said that the lowered speed limits saved lives.

To test the theory that the lowered speed limit had an effect on lowering highway mortality, the motor vehicle Death Rate was regressed against the Year and the average Speed Limit in force in that year, using the data at the right. The regression equation (with t-statistics in parentheses) is

$$\text{Death Rate} = 10.02 - 0.114 \text{ Year} + 0.038 \text{ Speed limit}, \quad (R^2 = 0.95)$$

$$(7.31) \quad (-13.15) \quad (2.31)$$

(3a) [4 points] Clearly, the "year" itself is not materially related to the death rate. What significant factors related to the death rate might it be substituting for? [That is to say, Why is the "year" included in this regression?]

Year	Death rate	Speed limit
1960	5.1	65
1962	5.1	65
1964	5.4	65
1966	5.5	65
1968	5.2	65
1970	4.7	65
1972	4.3	65
1974	3.5	55
1976	3.3	55
1978	3.3	55
1980	3.3	55
1982	2.8	55
1984	2.6	55
1986	2.5	65
1988	2.4	65
1990	2.2	65

(3b) [6 points] What can you conclude from this regression about the association between the speed limit and the highway death rate? [Be clear about R-squared, the coefficient for the death rate, its sign, its t-statistic, the null hypothesis about that coefficient, and what the statistics mean about reality.]

(3c) [2 points] What cautions would you suggest interpreting these results as a test of the theory that lowering speed limits saves lives? [Continue on the back if necessary.]

4) [12 points] Between November 11 and 14, 2002, the Gallop polling organization conducted telephone interviews with a sample of 1,001 adults, 18 years and older, on health care in the United States. Among the questions they asked was the following:

Are you generally satisfied or dissatisfied with the total cost of healthcare in this country?

The poll found that 22 percent said they were satisfied, 75 percent said they were dissatisfied, and 3 percent had no opinion.

(4a) [5 points] Find a 98 percent confidence interval for the proportion of U.S. adults who are generally satisfied with the total cost of healthcare in the country. [Yes, a 98% confidence interval.]

(b) [2 points] Gallup people used a 95 percent confidence interval in reporting their results. Would they have reported a larger confidence interval than yours, or a smaller one? Explain.

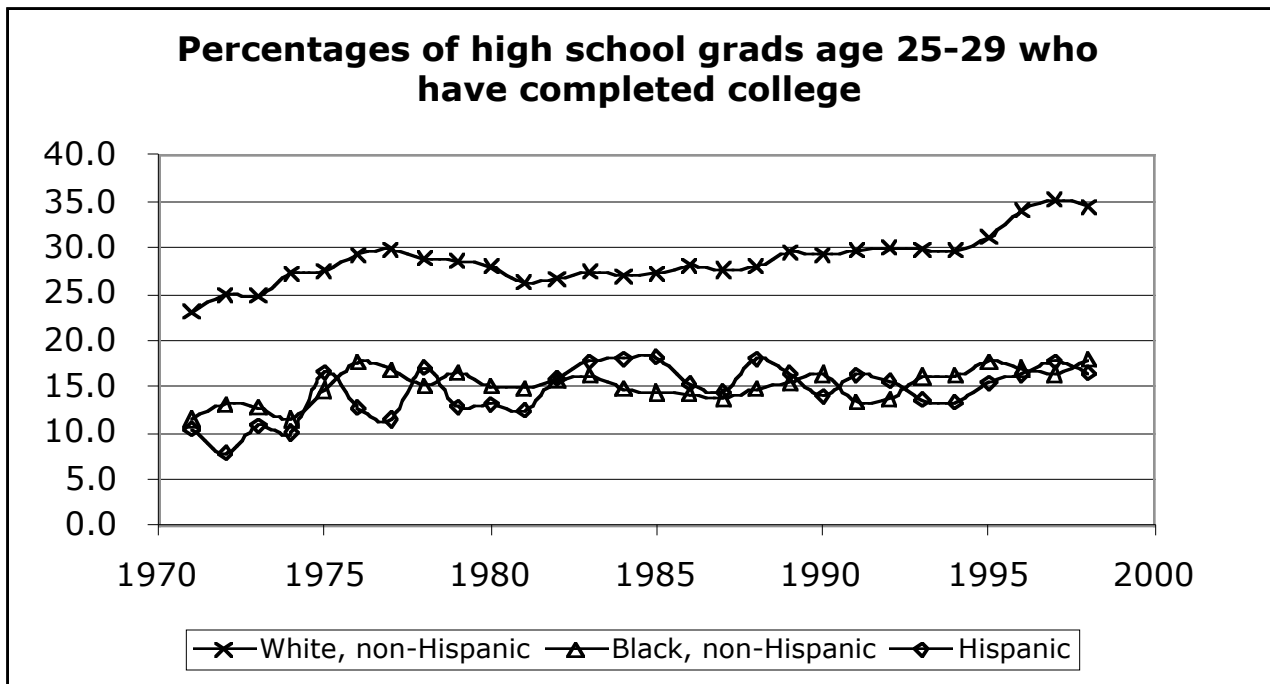
(c) [2 points] How many people would Gallup have had to poll to obtain a 95 percent confidence interval with a margin of error of plus-or-minus 1 percent?

(d) [3 points] Gallop always reports their results with the following cautionary statement:

Question wording and practical difficulties in conducting surveys can introduce error or bias into the findings of public opinion polls.

Are the errors or biases introduced by "question wording" and "practical difficulties in conducting surveys" included in the margin of error they (and we) compute? Explain your answer.

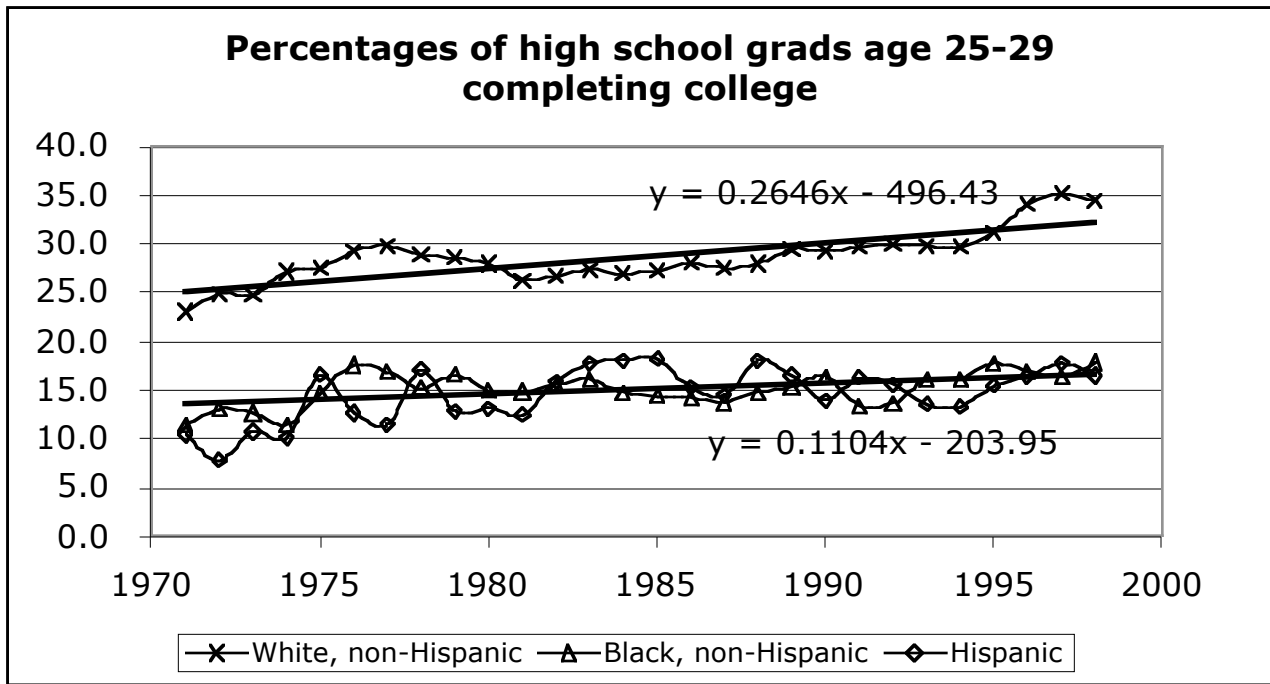
(5) [10 points] Below is a time-series graph from 1971 to 1998 showing the percentages of high school graduates in the United States between the ages of 25 and 29 who had completed college, displayed in three categories of race/ethnicity.



(5a) [2 points] What do you observe from these time-series?

(5b) [2 points, subtle] Why are the data for the Black and Hispanic populations "bumpier" than the data for the White population?

(5c) [4 points] Below, two trendlines (regressions versus time) have been added, with their equations, one for "White, non-Hispanic" and one for "Black, non-Hispanic."



Use these equations to predict for the year 2030 the percentages of White and Black high school grads who have completed college.

(5d) [2 points] Discuss briefly the public policy implications of what you see in (5a) and (5c). [The problems reflected here must be solved in your professional lifetime.]

6) [8 points] Quickies:

(a) The probability that a standard normal Z score is less than or equal to 1.50 (= $P[Z \leq 1.50]$) = ?

(b) If $n = 25$, then the probability that a t-statistic with $n-1$ degrees of freedom is greater than or equal to 2.17 (= $P[t(n-1) \geq 2.17]$) = ?

(c) If $X = 8$, $m_x = 10$, and $s_x = 2$, then the standard Z score (standard normal score) corresponding to X is ?

(e) The residuals in a regression are defined to be ...?

(f) If you were performing a chi-square analysis to determine if there is an association between Ethnic Group and Blood Type in the table at the right, how many degrees of freedom would you use?

Blood type	Ethnic Group				Total
	Hawaiians	Hawaiian-White	Hawaiian-Chinese	White	
O	1903	4469	2206	53759	62337
A	2490	4671	2368	50008	59537
B	178	606	568	16252	17604
AB	99	236	243	5001	5579
Total	4670	9982	5385	125020	145057

(g) Suppose X and Y are normally distributed with standard deviation σ . What kind of a distribution

do you think $\frac{\sum_{i=1}^m (X_i - \bar{X})^2 / m}{\sum_{i=1}^n (Y_i - \bar{Y})^2 / n}$ has?

(h) Brenda tells you that the variable W has a "bananamoid distribution with three bends," but you haven't a clue what that might be or look like (and neither does Brenda). You're pretty sure it is not normal, however. If we repeatedly draw samples of 30 observations and compute the sample means \bar{W} , what do you think the distribution of these sample means is? How do you know?

(i) How many degrees of freedom does $\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n}$ have?

7) [8 points] In 1998 Congress and the nation debated the use of sampling methods in the 2000 census. Eventually, the issue was decided by the Supreme Court, ruling out sampling because they interpreted it to contradict the phrase "actual enumeration" used in Article I, section 2 of the Constitution.

Consider the editorial excerpted below from the *Christian Science Monitor* on the subject of the controversy surrounding the use of sampling methods in the U.S. census in the year 2000.

Ignoring party positions and the Supreme Court, address this controversy from what you now know about statistics, samples, and populations. In what ways are the advocates of sampling right? In what ways might the critics be right? From a statistical viewpoint, how would you have urged your elected representative to vote on this issue to achieve the most accurate census count? Explain your position clearly, as if making a presentation or report to your elected representative, who probably needs all the statistics explained carefully. [Complete your answer below and on the back of this page.]

Down for the Count?

[Christian Science Monitor editorial, April 28, 1998]

Every census of a vast country like the United States is an estimate. Millions don't respond to the mailed census forms, and every front door can't be visited by follow-up head-counters — particularly in tightly packed urban centers.

The count came up so short in 1990 (at least 10 million) that the Census Bureau devised a plan for using sampling methods to arrive at a more accurate estimate next time around, in 2000. ... But [many] in Congress have dug their heels in — no sampling!

Why? Sampling's critics may say it's because the Constitution specifies an 'actual enumeration.' But the Constitution also says that the counting shall be done 'in such manner' as Congress directs. There's nothing barring techniques like sampling. The real issue here is political, not constitutional. ...

After 1990, the calls for improvement were loud. The sampling procedures drawn up by the Census Bureau are a far cry from 'guessing,' as some charge. The counting

process would begin with the traditional mailed census questionnaire, sent to every dwelling on a master address list for the country. In 1990, about 65 percent of households responded. Follow-up interviewers will contact a large number of those who don't respond, with an emphasis on areas with high rates of non-response. The bureau hopes this will boost the total contacted to 90 percent.

But that leaves 10 percent uncounted, and now the going gets tougher. This is where sampling would have its biggest impact. A sample of 25,000 census 'blocks' would be chosen for a second close, physical canvassing of every residence — a step that wouldn't be practical for the whole country. The results of this canvass would be compared to the earlier head count. 'Estimation factors' would emerge that could be used to correct counts in all blocks, with a close eye to corresponding demographic features like home ownership, race, and age of residents.

This spring, the bureau will conduct some dress rehearsals of this system in geographically varied parts of the country. Congress allowed for that much. But a full-scale gearing up for 2000 remains problematic. ...