

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. A table of standard normal probabilities is included for your use.

1) The table at the right shows the average monthly AFDC (Aid to Families with Dependent Children) caseload in the United States from 1974 through 1997, together with the national unemployment rate.

Year	Unemployment	AFDC Caseload
1974	6.2%	14,114
1975	8.8%	16,787
1976	10.4%	17,447
1977	8.7%	16,634
1978	7.1%	15,477
1979	6.2%	14,709
1980	6.2%	14,402
1981	6.3%	16,011
1982	6.7%	16,996
1983	6.9%	16,893
1984	5.9%	16,350
1985	5.0%	15,374
1986	4.8%	14,392
1987	3.7%	13,128
1988	3.6%	12,319
1989	4.6%	13,302
1990	4.1%	14,790
1991	6.6%	16,368
1992	7.7%	17,979
1993	6.7%	18,528
1994	6.3%	17,295
1995	5.5%	15,216
1996	4.7%	13,434
1997	4.6%	11,939

(1a) Create a stem-and-leaf plot of the unemployment data. (Do a first draft on the back of one of the test sheets; put your pretty answer below.)

(1b) Find the median, upper quartile, and lower quartile of the unemployment data.

median = upper quartile = lower quartile =

(1c) Draw a box and whiskers plot of the unemployment data (let the whiskers represent the minimum and maximum observations, rather than the 1.5*IQR criterion).

(2) A scatter plot of the AFDC caseload versus Unemployment is shown at the right, together with the regression line and regression statistics.

(2a) What are the slope and the y-intercept of the regression line?

slope =

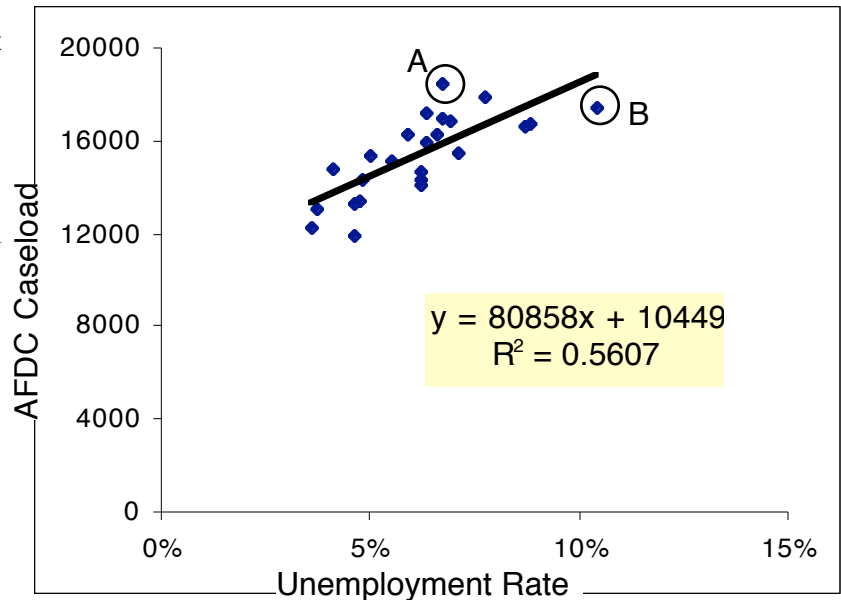
intercept =

(2b) What is the correlation between the AFDC caseload and Unemployment?

(2c) What is the meaning of R^2 here? (What does it tell you about the variance in the AFDC caseload?)

(2d) Two extreme points are circled in the scatterplot (the highest unemployment and the highest AFDC caseload). Which of these two points has the most influence on the regression line? What influence does it have?

(2e) Can you infer a causal relationship between Unemployment and the AFDC caseload from this regression? Explain (continue on the back if necessary).



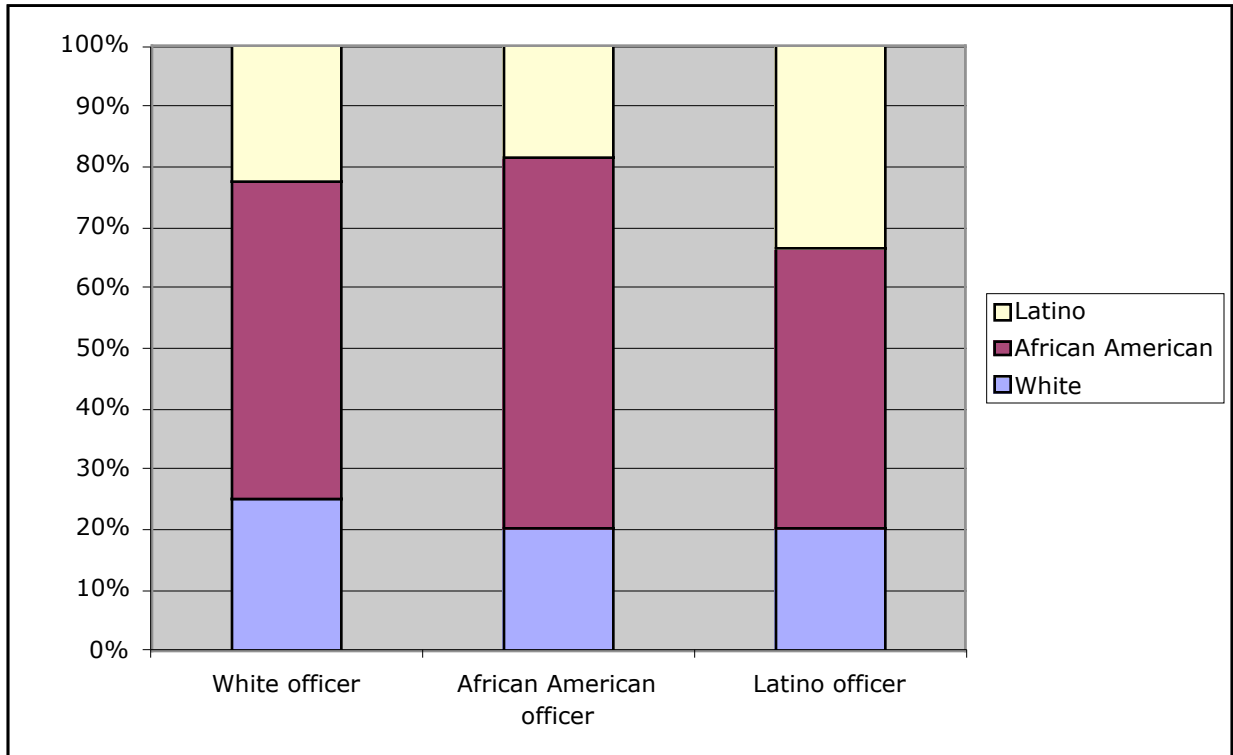
(3) In 1998, the New York City Civilian Complaint Review Board, overseeing the conduct of NYC police officers, received 1,110 complaints. The Review Board was interested in patterns in the data that might suggest racial tensions between complainants and arresting officers, so they requested the data to be broken down into racial categories, as shown below.

Complainants Race	Arresting Officer's Race		
	White officer	African American officer	Latino officer
White complainant	172	38	48
African American complainant	357	116	111
Latino complainant	153	35	80

(3a) Draw a bar chart illustrating the numbers in this data set. (Choose whatever chart you think might be interesting to look at; just make sure you use all the data.)

(3b) What can one observe from the bar chart you sketched in (a)?

(3c) Below is a stacked bar chart showing the *percentages* of complaints made by white, African-American, and Latino complainants against officers in the three racial categories. What would you point out to the Civilian Review Board that this chart shows about the interaction of race and complaints?



(4) In Cortland County, annual per capita income in 1997 averaged \$18,432, and we estimate the standard deviation of this population to be \$6,000.

(4a) Assume that the distribution of incomes is approximately normal [it's not, but let's assume in part (a) it is]. What fraction of the incomes in Cortland County would be over \$21,000? (Use the table of the standard normal distribution attached to this test.)

(4b) A random sample of 50 individuals is selected from this county for purposes of estimating tax burden as a percentage of income. What is the shape of the distribution of the means of such samples? What is that distribution?

(4c) Incomes are usually not normally distributed [the assumption in part (a) is not really true]. Incomes are usually skewed to the right. Will that have an impact on the distribution of the sample mean? Explain.

(5) In a study of smokers broken down by gender and age, it was found that the percentage of women in the study who were smokers was higher overall than the percentage of men who smoke (23.5% in the table below for women, 22.5% for men). Yet in each age category the percentage of men who smoked is greater than the percentage of women.

Age	Number of Men	Percent smokers	Number of men smokers	Number of Women	Percent smokers	Number of women smokers
18-24	52	27.8%	14	49	21.8%	11
25-34	98	29.5%	29	51	26.4%	13
35-44	102	31.5%	32	343	27.1%	93
45-64	105	27.1%	28	137	24.0%	33
65 and over	313	14.9%	47	112	11.5%	13
TOTAL	670	22.5%	151	692	23.5%	163

Explain how it can be in this data that the total percent of women who smoke exceeds that of men, while the opposite is the case for each individual age group.