

17.871

Spring 2004

Problem Set # 2: Descriptive Statistics, Correlation, and Bivariate Regressions

Handed out: February 26, 2004

Due back: March 4, 2004

Note: Unless otherwise indicated, you may use any aid to calculation (spreadsheet, calculator, STATA, abacus, fingers and toes, etc.) to answer the following questions. If the question asks you to calculate something, indicate that you knew what elements to bring together to generate the statistic that I ask you to calculate.

1. The following is the distribution of reported family incomes in the 2002 American National Election Study. Draw the histogram of the distribution on a density scale. How does it differ from the histogram that would be drawn if you entered the data in STATA and used the “hist” command to draw a graph?

Income	n
\$0-\$15K	144
\$15k-\$35k	318
\$35k-\$50k	232
\$50k-65k	210
\$65k-\$85k	231
Greater than \$85k	295

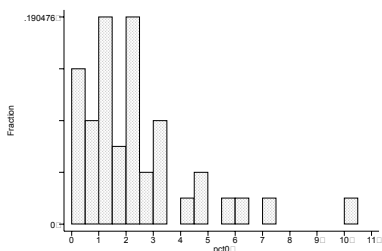
2. The following is a table of the percentage of the vote received by Roy Barnes, the Democratic candidate for Governor in Georgia in 2002, from a random sample of 15 Georgia counties, along with the total number of votes cast in the county.
 - 2.1 Which counties (if any) are beyond one standard deviation of the mean on the barnespct variable?
 - 2.2 Which counties (if any) are beyond two standard deviations of the mean on the barnespct variable?
 - 2.3. Assume for the moment that these are the only counties in Georgia. Is the average of barnespct the same as the percentage of voters who voted for Barnes in the election? Explain your answer without resorting to actually doing any calculation.

county	barnespct	turnout
Bartow	35	18,124
Bleckley	35	3,297
Clarke	57	21,702
Coffee	39	7,860
Effingham	31	11,257
Floyd	38	21,482
Haralson	34	7,187
Hart	45	6,211
Lanier	48	1,352
Morgan	38	4,819
Quitman	71	575
Sumter	55	7,523
Tift	36	7,747
Wilcox	37	1,999
Worth	38	4,969

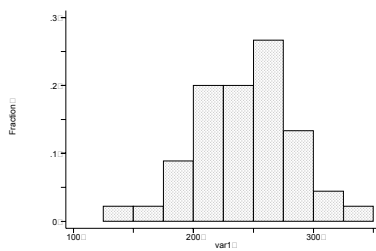
3. In Fall 2002 the mean answer to the “overall teaching quality” question on the Subject Evaluation Guide was 5.99 and the standard deviation was 0.67. Say whether each of the following scores for individual classes was unusually high or low, and why.

- (a) 6.04
- (b) 3.69
- (c) 5.96
- (d) 2.69
- (e) 6.14

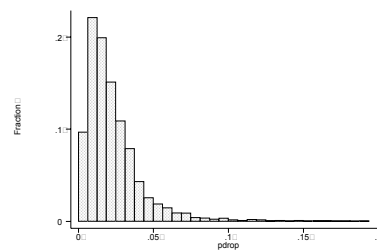
4. Estimate the mean and median of the following three distributions, from eyeballing the graphs:



(a)



(b)



(c)

5. In Fall 2002 the mean answer to the “overall teaching quality” question on the Subject Evaluation Guide for classes taught in the School of Science was 5.92, standard deviation of 0.57; the average and standard deviation for classes in the School of Humanities, Arts, and Social Sciences was 6.16 and 0.59, respectively.
- Roughly what % of School of Science classes had teaching evaluations between 4.78 and 7.0?
 - Roughly what % of SHASS classes had teaching evaluations between 5.57 and 6.75?
 - If you combined the data from the two schools, would the mean of the new distribution be more than, less than, or equal to 6.04 (the average of the two means)? Calculate the result.
 - Would the standard deviation of the new distribution be more than, less than, or equal to 0.58 (the average of the two standard deviations)? (Assume the number of observations from each school is equal.) In this case, don’t calculate the result, but reason intuitively. (The calculation is very, very tedious.)
6. Would you expect the amount of money the state of Massachusetts spends in each town to educate elementary students (on a per capita basis) to be positively or negatively correlated with performance on standardized tests? Why?
7. The graph on page 8 is a “scatterplot matrix,” showing the scatterplot of four variables (x_1 , x_2 , x_3 , and x_4) against each other. Match each scatterplot in that graph with its correlation:

- | | | |
|-------------------------|------------|--|
| (a) x_1 against x_2 | (i) .20 | |
| (b) x_1 against x_3 | (ii) .95 | |
| (c) x_1 against x_4 | (iii) -.80 | |
| (d) x_2 against x_3 | (iv) .43 | |
| (e) x_2 against x_4 | (v) .50 | |
| (f) x_3 against x_4 | (vi) .12 | |

8. Calculate the correlation between the average carbon monoxide emitted by the following cigarette brands and the amount of nicotine in the average cigarette.

Brand	Nicotine (mg)	Carbon Monoxide(mg)
Alpine	0.86	13.6
Benson & Hedges	1.06	16.6
Bull Durham	2.03	23.5
Camel Lights	0.67	10.2
Carlton	0.4	5.4
Chesterfield	1.04	15
Golden Lights	0.76	9
Kent	0.95	12.3
Kool	1.12	16.3
L & M	1.02	15.4
Lark Lights	1.01	13
Marlboro	0.9	14.4
Merit	0.57	10
Multi Filter	0.78	10.2
Newport Lights	0.74	9.5
Now	0.13	1.5
Old Gold	1.26	18.5
Pall Mall Light	1.08	12.6
Raleigh	0.96	17.5
Salem Ultra	0.42	4.9
Tareyton	1.01	15.9
TRUE	0.61	8.5
Viceroy Rich Light	0.69	10.6
Virginia Slims	1.02	13.9
Winston Lights	0.82	14.9

9. The data below is a random sample taken from responses to the Fall 2002 Subject Evaluation Guide at MIT. The variables are answers to the question about how available the instructor was (averaged across all students) and the number of students in the class.
- 9.1 What is the correlation coefficient between the overall rating of the subject and the subject's enrollment? Show your calculations.
 - 9.2 What is the correlation coefficient between the overall rating of the subject and the *log* of the subject's enrollment? Show your calculations
 - 9.3 Discuss which correlation is the better summary of the relationship between subject size and class quality.

Availability	Enrollment
5	2
5.86	14
5.27	15
6.67	6
7	6
5.96	105
3.05	20
6.62	21
5.88	17
6.53	19
6.5	6
5.4	15
4.88	25
4.90	31
6.71	14
6.71	7
5.45	147
5.43	361
4.85	103
7	6

10. Create a STATA data set of the data you hand-analyzed for Question 9.
 - 10.1 What does STATA calculate as the correlation between instructor availability and the subject's enrollment (to 3 decimal places)?
 - 10.2 What does STATA calculate as the correlation between instructor availability and the *log* of the subject's enrollment (to 3 decimal places)?
 - 10.3 Generate two graphs:
 - a. Instructor availability against enrollment

- b. Instructor availability against the log of enrollment.
- 10.4 Discuss anything about the relationship between instructor availability and enrollment that may lead you to question whether the correlation coefficient you calculated is a measure of the “true” correlation between class size and instructor availability.

For this part, turn in a log file that shows (1) the variable names in the data set, (2) the averages of the variables, and (3) the correlations among the relevant variables. Also turn in printed version of the graphs in 11.3.

11. Use the data set /mit/course_number/Examples/residual_vote_1996.dta to answer the following questions. There is a description of the variables immediately below.
- 11.1 What was the correlation coefficient between the median house value in a county and the residual vote rate of that county? Does the correlation coefficient you calculated fairly describe this relationship? (Hint: Think about reasonable transformations of the variables.)
- 11.2 Run the regression associated with the relationship between median house value and residual vote rate. If you think the variables need to be transformed to better reflect the relationship, do that. Write a sentence that describes how you would interpret this regression.

Variable name	description
stabbr	state postal abbreviation
area_name	county or similar state subdivision
median_val	median value of single family houses (uits: \$100,000)
pdrop	fraction of ballots with no vote for president (i.e., residual vote)

Graph to accompany Question 7.

