Problem Set #1
Statistics for Lawyers
March 12, 2006

1.

8.171259 3.462472 7.278197 7.40853 3.750913 6.187179 4.102017 5.361534 4.883733 6.868109

1.a Calculate the mean and median of the series above.
1.b Calculate the variance and the standard deviation of the series above.
1.c Calculate the standardized values (Z scores) for the series above.
1.d What is the variance of f(x) = 5*x +3 where x is the series above.
1.e Graph the series above and f(x) (on separate graphs) using excel. Note the difference in variance.
1.f Create a histogram of the series above using excel. Try various numbers of "bin."

2.

x1	38	54	54	73	18	30	3	60	58	68
x2	76	113	110	152	35	60	12	123	117	141

2.a Graph a scatterplot of x1 and x2 using excel. How are the two series related?
2.b Calculate the correlation coefficient between x1 and x2.
2.c Calculate the least squares regression line of x2 on x1 (preferably by hand).
2.d Graph the residuals from the regression estimated in 2.c around the regression line using excel.
2.e Calculate the r^2 for the regression calculated in 2.c (again, preferably by hand).

3. You roll 2 six-sided fair dice

3.a What's the probability that you roll 7 three times in a row?
3.b What's the probability you roll something other than 7 on each of your first three rolls?
3.c What's the probability you don't roll 7 on all of your first three rolls?
3.d You roll the dice 4 times; what's the probability you roll 7 on the first three roles but not on the fourth?
3.e Someone offers you the following game: Roll the dice once, and if you get a 7, I will pay you $100. What's the most you'd be willing to pay to play this game assuming you were risk neutral (i.e., you'd be willing to pay the expected value of the game).

Problem Set #2
Statistics for Lawyers
March 13, 2006

1. The employees of firm FSU are either managers or workers. Ten percent of the employees are managers. Also, 30% of the employees have a college degree. All managers have a college degree.

1.a Draw a Venn Diagram representing the make-up of FSU's workforce.

	College Degree	No College Degree
Managers	10%	0
Workers	20%	70%

1.b What's the probability that a randomly chosen employee is a manager if you know he had a college degree?

10/30=33.333%

1.c What's the probability that a randomly chosen worker does not have a college degree?

70/90=77.778%

2 The proportion of the overall labor force that makes the minimum wage is 10%. At the restaurant McFSU, 20% of the workforce makes the minimum wage. 30% of the workforce is female. Sixty-five percent of the workforce is men who make more than the minimum wage.

2.a You are a plaintiff's attorney specializing in employment discrimination matters. Does it appear that McFSU's compensation methods generate a disparate impact on women?

Perhaps

2.b What conditional probability are you interested in when you solve 2.a?

Probability that you make minimum wage given that you are a woman

2.c Solve the conditional probability from 2.b using Bayes's Rule.

$$\Pr\left(\text{minimum wage}|\text{female}\right)=\frac{\Pr\left(\text{female}|\text{minimum wage}\right)*\Pr\left(\text{minimum wage}\right)}{\Pr\left(\text{female}|\text{minimum wage}\right)*\Pr\left(\text{minimum wage}\right)+\Pr\left(\text{female}|>\text{minimum wage}\right)*\Pr\left(>\text{minimum wage}\right)}$$

$$=\frac{.75*.2}{.3}=50\%$$

2.d Solve the conditional probability from 2.b using a Venn Diagram.

	Min Wage	>Min Wage
Male	5	65

Female	15	15

Prob = 15/30=50%

2.e Suppose you are now the defense counsel. You find that 75% of workers with just a high school diploma make the minimum wage in the labor force at large. You also find that all of the men working at McFSU have a college degree while only half of the women working at McFSU have more than a high school diploma. Finally, you find that of the highly educated women, 75 percent of them make more than the minimum wage. What statistical argument do you make in your client's favor?

	High School	College	High School	College
	Male		Female	
Min Wage	0	5	11.25	3.75
>Min Wage	0	65	3.75	11.25

Prob(making minimum wage|woman with just high school) = 75% which is the national average

2.f If you were again the plaintiff's lawyer, what information about the labor force in general would you like to know to rebut the argument made in 2.e?

What's the national rate at which college educated people make the minimum wage since in this example

Prob(making minimum wage|woman with college degree)=25%

Further, even w/o the info you could compare it to the

Prob(making mw|man w/ college degree)=7%

Problem Set #3
Statistics for Lawyers
March 13, 2006

1. You estimate that the likelihood of Vioxx users developing a heart attack within 2 years of starting Vioxx is 12%. The known general population heart attack risk in the same period is 9%. The standard deviation of heart attack risk is 1.6%. Assume (contrary to fact) that Vioxx users are similar to the underlying population in all other relevant characteristics (e.g., distribution of age, race, comorbidities, etc.).

1.a Test the null hypothesis that Vioxx users exhibit the same heart attack risk as the general population at the 5% type-I error level.

$$t = \frac{12-9}{1.6} = 1.875 < 1.96 \therefore \text{ effect is not statistically significant}$$

1.b Argue that the appropriate test is a one-tailed test.

You don't care if Vioxx reduces heart attack risk; you only care if it increases the risk

1.c Test the null hypothesis using a one-tailed test

t is still 1.875 but now the critical value is different (if you use a 5% type I error, it is 1.645 so it would appear that Vioxx does significantly increase heart attach risk).

1.d What is the p value for the results described above?

p is about 0.030

1.e Calculate the 85% confidence interval for the population's heart attack risk in this 2 year period.

An 85% confidence interval is associated with a 15% Type I error which, in a 2 tailed test, means there will be 7.5% in each tail. This is associated with 1.44 standard deviations, therefore the confidence interval is (12-1.44*1.6, 12+1.44*1.6) or (9.7, 14.3)

1.f What does the confidence interval in 1.e denote?

If the true mean is the sample mean, then the mean from 85% of the random samples will fall within that range.

1.g What if the population of Vioxx users does not "look like" the population at large (e.g., perhaps Vioxx users are older; have more co-morbidities; etc.). How might that change the way you do your statistical testing above?

Then the appropriate hypothesized value of the heart attack risk will not be the general population mean; instead it will be the mean of the patients who would be candidates for Vioxx.

2. The average wage among software developers is $100,000 per year with a standard deviation of $8,000. You examine the pay practices of MicroFSU and find that the average pay of their female software developers is $75,000.

2.a Test the hypothesis that the average wage of female software developers at MicroFSU is not statistically different from the average wage of software developers at large.

$$t = \frac{75,000 - 100,000}{8,000} = -3.125$$

2.b Did you use a one-tailed test or a two-tailed test? Why? How might it depend on the context of the investigation?

In an employment discrimination case, women and a protected class, while men are not. Therefore, a one tailed test might be appropriate since all you care about is whether or not women are being harmed in this firm. However, if you are interested in "equality" for non-legal reasons, you may want to investigate whether or not men and women are treated the same, in which case a two tailed test will be appropriate.

2.c It turns out that wage distributions generally have mean values that are much higher than their median value. Why do you think this might be? (Note: we didn't explicitly cover this, but use your intuition to guide you here).

Wages are bound by 0 on the left (i.e., you can't pay someone a negative wage) but there is no bound on the right (i.e., LeBron James makes huge amounts of money). The extreme right values (i.e., outliers) will pull up the mean but will not affect the median.

2.d Given the fact pointed out in 2.c, how does this undermine your statistical testing in 2.a and 2.b? (Note: again, not explicitly covered in class, but use what you know about significance testing to figure it out. Hint: the testing we have done so far is based on the assumption that the variables are distributed according to the Normal Distribution. Use what we know about the Normal Distribution to guide you here).

In the Normal Distribution, mean=median. If the mean is much larger than the median, the distribution moves farther and farther away from being Normal (the distribution is skewed to the right). It turns out that if you take ln(wage) (i.e., the natural logarithm of wages), the distribution will generally be normal, so sometimes it is useful to “transform” the data before you analyze it.

Problem Set 4
Statistics for Lawyers
April 4, 2006

Download the file ps4.xls and open it in Excel. The spreadsheet contains data for all 50 states for the year 1998 on the following variables:

beerpc: per capita (ages 14+) sales of beer measured in gallons of ethanol.
winepc: per capita (ages 14+) sales of wine measured in gallons of ethanol.
liqpc: per capita (ages 14+) sales of distilled spirits measured in gallons of ethanol.
alpc: sum of beerpc, winepc, and liqpc.
unins: % of state population w/ no health insurance.
seced: % of state population w/ at least high school education.
income: per capita, inflation adjusted income (in $1,000s)
income2: income*income
unemp: unemployment rate
pctrural: % of state population living in rural areas
lfp: % of women in state who work outside the home
mormon: % of state population self-identified as belonging to LDS church
sobapt: % of state population self-identified as belonging to Southern Baptist church
catholic: % of state population self-identified as belonging to Roman Catholic church
protestant: % of state population self-identified as belonging to a mainline protestant church
per1519: % of state population ages 15-19
per2029: % of state population ages 20-29
per1529: % of state population ages 15-29
per1534: % state population ages 15-34

1) Familiarize yourself with Excel's regression function (Available under menu "Tools" and then "Data Analysis"; if you don't see data analysis, you need to add it by going to "Tools" "Add-Ins" then check "analysis toolpak" and "analysis toolpak VBA"; after you do that, you should have "data analysis" under the "tools" menu. Choose "regression" and then follow the menus) by running regressions explaining per capita beer sales on the basis of each variable separately. Now run a multiple regression using those variables you think should be important. Did the multiple regression generate coefficients of the same size and sign as the separate regressions you ran?
2) Run a regression of beerpc on unins seced income income2 unemp pctrural lfp mormon sobapt catholic protestant per1519 per2029
    a. Which coefficients are statistically significant at the 5% Type I error level (two tailed test)?
    b. Remove the age variables from the regression. Do any of the coefficients change in sign or significance?
    c. Also remove the religion variables. Do any of the coefficients change again?

d. Which of the regressions (2.a, 2.b, 2.c) do the best job fitting the data?

e. What other variables would you like to have in modeling the determinants of beer sales

3) Rerun your regressions for wine, spirits, and total alcohol.

a. What variables are important in explaining one/some of the y variables but not the others? Do you have any hypotheses for these differences?

b. Which kind of alcohol sales is most readily explained by the variables presented here?