Midterm 1 Key
Stats 11, Fall 2003

There were two versions of the test. Version A was YELLOW and Version B was BLUE. The different versions are noted below.

1. In the anonymous survey we did the first day of class, you were asked whether you would vote for the recall. Below is a table with these results, broken down by gender.

	Female	*Male*	*Total*
No recall	*18*	*26*	*44*
Yes	*6*	*11*	*17*
Total	*24*	*38*	*62*

Suppose we select a person at random from these 62 people.

This first one had a typo in the answer. The "Male" column should add to 37, not 38, and the total should be 61, not 62. Whichever total you used was fine.

a) Find the probability that the person is male ***or*** *voted for the recall.(YELLOW)*
a) Find the probability that the person is female ***or*** *voted for the recall. (BLUE)*

Both versions require that you know how to calculate probabilities of "Or"s.
P(A OR B) = P(A) + P(B) - P(A and B)
P(Male or Yes) = P(Male) + P(YES) - P(Male and Yes) = 37/61 + 17/61 - 11/61 = 43/61

P(Female or Yes) = 24/61 + 17/61 - 6/17 = 35/61

b) Are women in this survey less likely to support the recall than men? Give appropriate evidence for your answer.
To answer this you need to compare P(Yes | Female) with P(Yes | Male) (Equivalently, you could compare P(No | Female) with P(Yes | Male).
P(Yes | Female) = 6/24 = .250
P(Yes | Male) = 11/37 = .297
The answer is no. In fact, women are less likely to support the recall (at least for these 61 people.)

c) Suppose we select someone at random. Is the event "the person is male" independent of the event "the person votes for the recall?" Explain.(YELLOW)
c) Suppose we select someone at random. Is the event "the person is female" independent of the event "the person votes for the recall?" Explain.(BLUE)
This requires that you know the definition of what it means for two events to be independent: A and B are indept. means P(A | B) = P(A)

So you must check to see whether P(Male | Yes) = P(Male) or P(Yes | Male) = P(Yes), etc.
P(Male | Yes) = 11/17 = .647
P(Male) = 37/61 = .607
These probabilities are not equal and so the events are NOT independent.

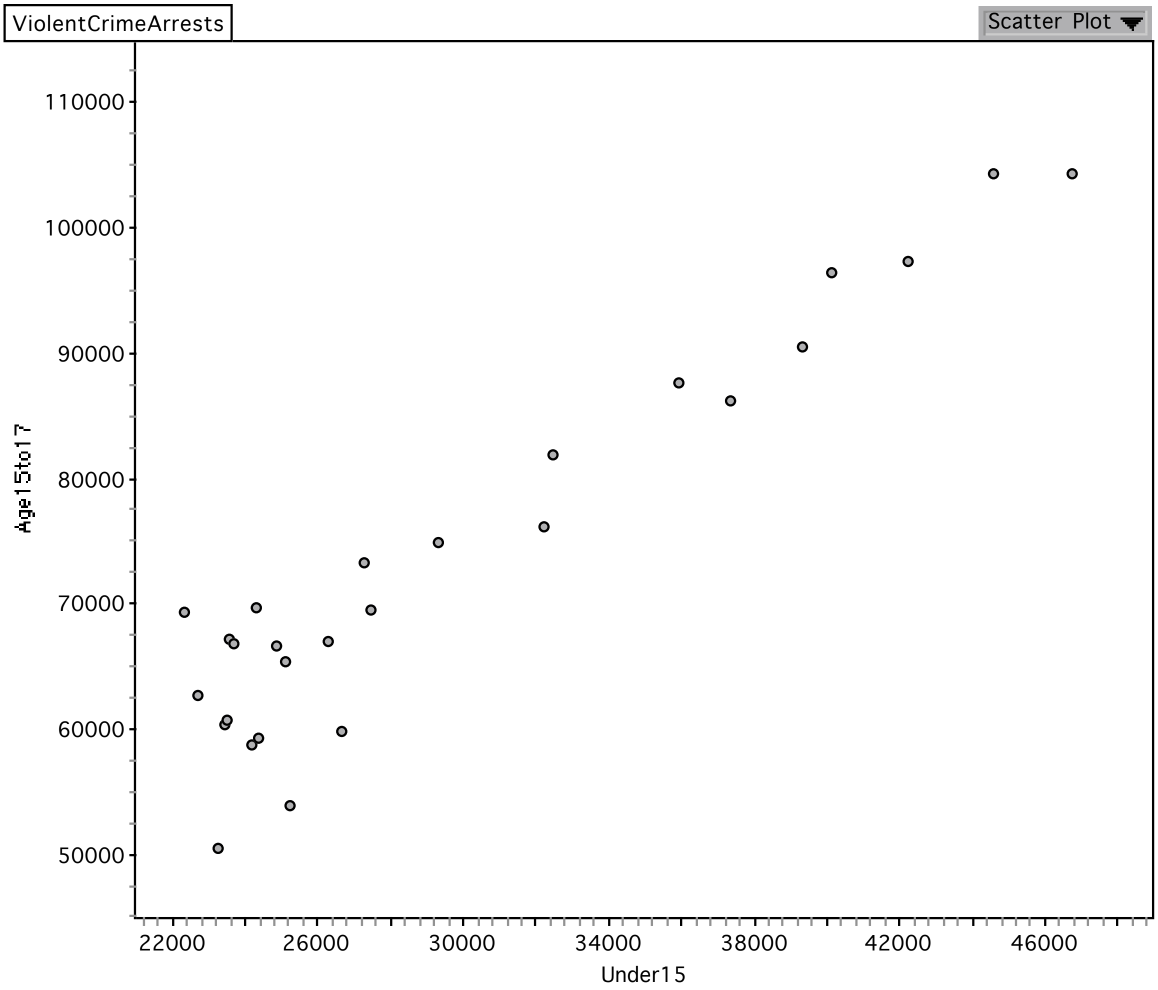
For the blue test, the calculations are similar
P(Female | Yes) = 6/17 = .353
P(Female) = 24/61 = .393

Common mistakes include:
• Using the wrong definition: P(A | B) = P(B)
• Stating the right definition but not doing any calculations
• Trying to reason logically about the situation (nice try -- but if you have the data you don't need to form an argument, you can just check it.)

Summary: This was, I felt, a rather straight-forward problem. Part (a) was the simplest and was a straight-ahead calculation problem. A surprisingly large number of people missed this. On the other hand, many more people than I expected got part (b) right, and that requires some thought; you need to know to condition on gender to compare probabilities. Part (c) is also pretty straight-forward, and exactly like an exercise I did in lecture.

To review these questions, see Section 4.5 in the book.

2) Below is a graph showing the number of violent crime arrests in the U.S. in which the offenders were aged 15 to 17 against those in which the offenders were under 15 years old. Each data point represents a year from 1971 to 1997.



a) Describe the relationship.

We looked for four things: Did you mention (1) the direction of the association (e.g. positive, negative, no direction), (2) the strength of the association (e.g. weak, strong, moderate), (3) the shape (linear? outliers?) and (4) did you provide an interpretation in the context of the data. One of the "tests" for checking for context was to see if what you wrote could be applied to absolutely any graph anywhere, or whether it could only apply to this graph. Here's an example of a good answer:

> There is a positive association between the number of children under the age of 15 arrested for crimes and the number of children between the ages of 15 and 17 arrested for crimes. The association is fairly strong and linear. Years in which many children under 15 years old were arrested also saw many children between the ages of 15 and 17 arrested.

You could stop there, or you could also go on and describe how, in years in which the number of under 15 arrests were low, there was more variability in the number of 15-17 year-olds arrested than in years in which more 15 year olds were arrested.

Some common mistakes:

- Most common was failing to describe one of the (4) things above. Some people described all three, and provided some context by saying something like "As Under15 goes up, Age15to17 tends to go up". But this doesn't tell us what "Under15" represents.
- Each data point represents a year, but the data points are not sorted by year. Many people wrote something that implied that time was passing as you moved from the left to the right. In fact, there is nothing to indicate any ordering with respect to time.
- The variables measured were a count of the number of *people* arrested, not the number of *crimes* committed. So it does not make sense to talk about what it means for someone to commit an additional crime.

Summary: I know that writing clearly and precisely is difficult, and expected some low scores on this one. But remember to read the problem carefully so that you are describing the correct data set, and are not making assumptions that are not true. Also, this question tests a bit of statistical cultural knowledge. If you've taken a statistics course, you should know that the direction, strength, and shape of an association are features to focus on.

b) Here are summary statistics for the number of offenders under the age 15 and the number between age 15-17. Write the equation of the least-squares line that fits these data.

Average number, under 15 years:	*29574*	*SD:*	*7497*
Average number, aged 15-17:	*73342*	*SD:*	*15170*

r = 0.92 (Yellow) r = 0.94 (Blue)

The two versions of the test differed only in the value of the correlation. The true value for this data is .94, but if you did the yellow exam I trust you didn't notice that you had to use a slightly lower correlation.

This was a very straight-forward calculation problem, and I confess to being disappointed that so many people missed it. It's not unusual for people to make small mistakes such as putting plus signs where there should be minus signs or calculating the slope as r*(SDx/SDy) instead of the other way around. But a great many did not know where to begin, and this is a sign either of a failure of memory or a failure to study.

The calculations here are for the yellow exam. The blue exam's calculations are the same, but with .94 used in place of .92.

slope = r*(SDy/SDx) = .92*(15170/7497) = 1.861598
intercept = avg y - b*(avg x) = 73342 - (1.861598)*(29572) = 18287.1

y-hat = 18287.1 + 1.86x
The blue equation is y-hat = 16491.8 + 1.922x

c) Interpret the slope of the regression line. Make sure your interpretation is in context of the data above.

Here's the interpretation for the blue exam. The yellow exam is quite similar:
For years in which there was one more arrest of someone under the age of 15, there were, on average, 1.922 additional arrests of children aged 15-17.

Or, if you want to use bigger numbers (which some of you did), you could say
"For years in which there were 1000 more arrests of children under the age of 15 there were, on average 1922 additional arrests of children aged 15-17.

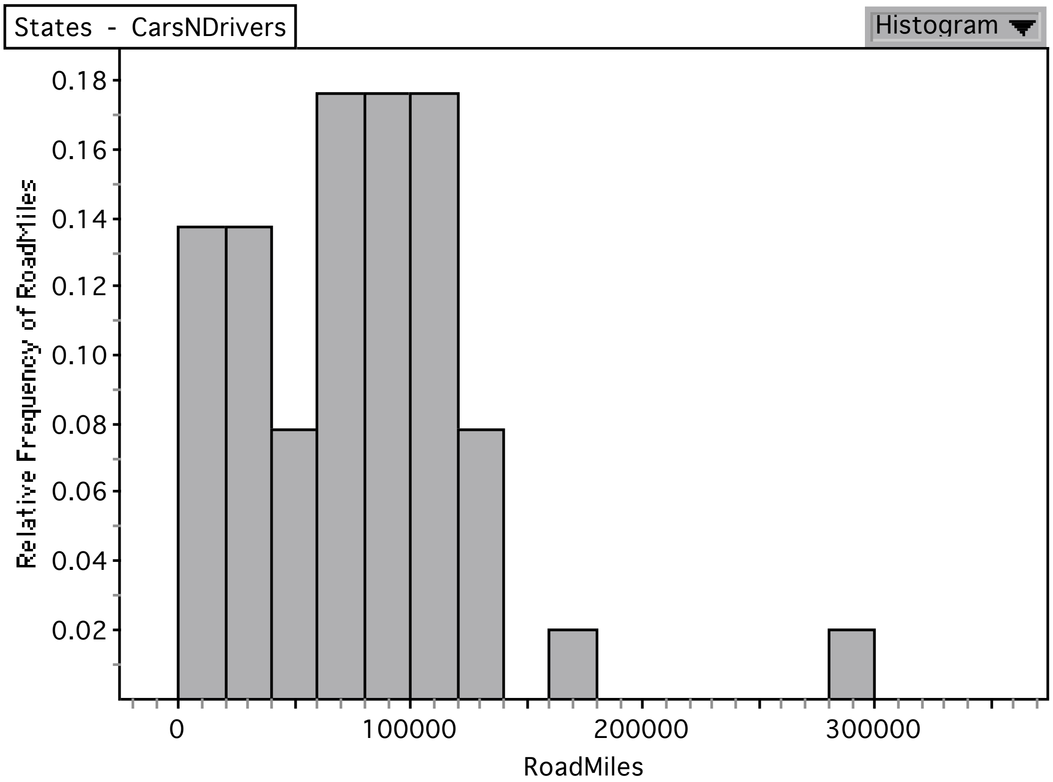
Mistakes:

- It is simply not true that *exactly* 1922 additional arrests were made of 15-17 year olds. In some years this number went up much more, in other years it actually went *down*. It is true, however, that this is how the *average* number of 15-17 year old arrests compares when you compare years in which the number of 15 year-olds differed by 1000.
- The slope tells us absolutely nothing about what happens over time. So you can't talk about what happens year to year, or from one year to the next.
- Many people rounded up and said the slope was "about 2". That's fine. But it is NOT true that this means that approximately twice as many 15-17 year olds were arrested than 15 year olds. You can see that this isn't true just by looking at the scatterplot and noting that almost always the number of 15-17 year olds is much more than twice the number of 15 year olds.
- Avoid the trap of implying that the slope tells us what will happen if we make a change to the x-variable. Some people wrote something like "If an additional arrest is made of a 15 year old, there will be an average of 1.922 more arrests made of 15-17 year olds." First, this phrasing is ambiguous. What do you mean by "an additional arrest"? Do you mean within any given year? Do you mean one more arrest this year compared to some previous year? (The latter is the correct approach.) But if so, how could we know that an additional arrest of a 15 year old will *cause* more arrests of 15-17 year olds? Think back to what we discussed in the first two days of the course; to prove that a change in x causes a change in y, you must run a controlled experiment. And it's clear that the study we're talking about here could not possibly have been a controlled experiment.

Review the handout I posted for more information on this problem:
http://web.stat.ucla.edu/~rgould/11f03/mt1examples.pdf

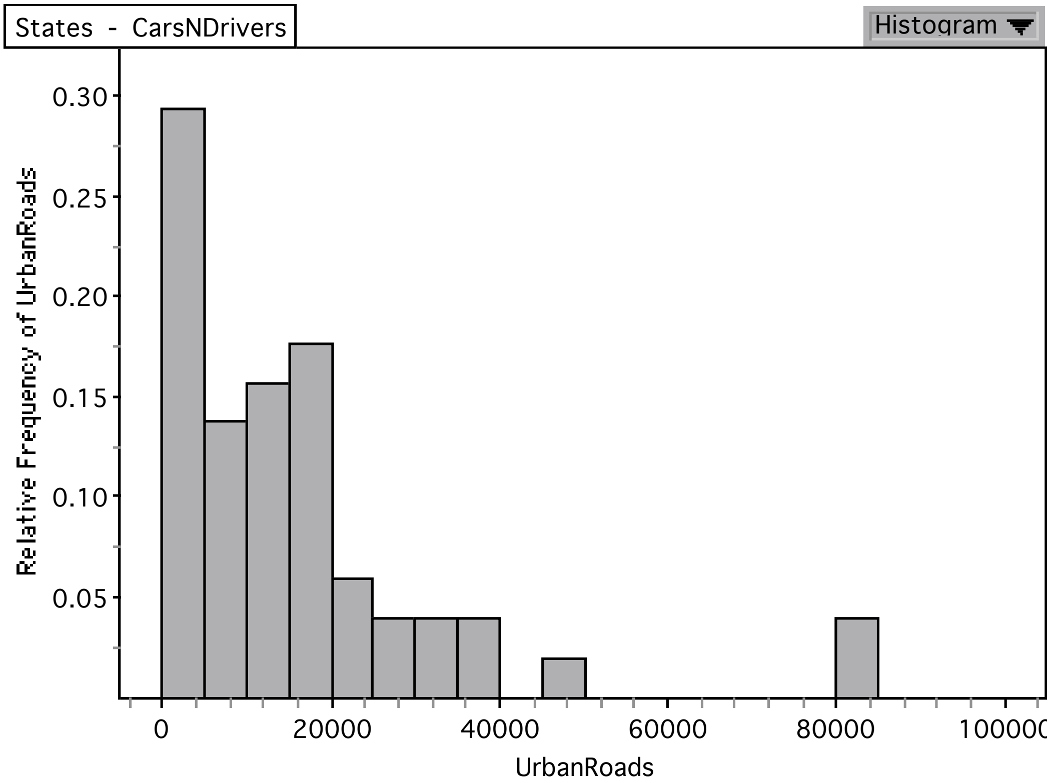
Another hint: Once you've written your interpretation, go back and look at the scatterplot and see if what you said is true. Truth matters here. For example, here's an interpretation that sounds okay, but is actually untrue: In years in which 1000 additional arrests were made of children under 15, 1922 additional arrests of 15-17 year olds were made. I challenge you to find a single instance of this being true: find two years that differ by 1000 (or 10000 or 100 or whatever) on the x-axis and see if every single corresponding value on the y-axis differs by 1922. Probably not, because the regression line tells us about *average* y values. So you will find that the average of the two groups differs by 1,922 15-17-year-old arrests.

3. The Department of Transportation keeps statistics on the number of miles of roads in each of the 50 states and the District of Columbia. Below is a histogram of these data. RoadMiles is the number of miles of road in a state.



Each bin is 20,000 miles wide, and the first bin begins at 0.

Blue had a different graph:



For the Blue exam, we measured the number of miles of urban roads, and not just overall roads. But the questions are the same for both exams.

a) Describe the distribution of RoadMiles.

We were looking for mention of three things (1) center (2) spread (3) shape. And your description should be in context of the data. Here's an example:

Yellow: A typical state had about 100,000 miles of roads (center), but some states had as little as somewhere between 0 and 20K, and some states as many as 300K. (spread). The distribution of road miles is right-skewed, in part because of outliers around 180K and 300K.

Blue: The typical state has about 15000 miles of urban roads, although some states haave as many as 80,000 miles and some as few as between 0 and 5000. (Center and spread). The distribution is right-skewed with a couple of outliers around 80000.

Most people did fine on this one.

b) Approximately what is the median number of road miles in a state, and the first and third quartiles? Describe how you got your answer. Indicate where these values are on the histogram.

I don't mean to scold, but it was clear from your answers that many of you either (a) didn't know what a median and quartiles are or (b) didn't know how to read a histogram. The median is the number such that half of the observations are below and half are above. The vertical axis in the histogram tells you the percentage of observations falling in each bin. So we know, for example, from the blue exam we know that about 30% of the states have between 0 and 5000 miles of urban roads (in other words, this first bin represents about 15 states.) The second bin has about 14% of the observations and the third has about 15%. So the first three bins together contain roughly 59%. This means that the median value must be somewhere in that third bin. And this means its some number between 10,000 and 15000 miles of urban roads.

Still on the blue exam: The first quartile is the number that has 25% of the observations below it. The first bin has 30% of the data, and so the first quartile must be in there somewhere, which means its some number between 0 and 5000 miles of roads. Similarly, the third quartile, which has 75% of the observations below it, must be in the fourth bin, between 15,000 and 20,000. (It's conceivable that you might put it in the fourth bin if you read the frequencies slightly differently.)

For the Yellow exam, the median must be somewhere in the fourth bin, since the first three bins must have 14% + 14% + 8% = 36% of the data, and the fourth bin has 17% more. So Median: between 60K and 80K miles. Q1 between 20K and 40K miles, and Q3 between 100K and 120Kmiles.

You lost points if you somewhere implied that you could know precisely what the values were (you can't because we can't see inside the bins) and if you forgot to mark these values on the histogram.

c) True or false and explain: most states have fewer miles of roads than the average state.
(Note: "most" means "more than half".)

On the whole, the class did pretty well on this one, despite the fact that I think its somewhat tricky. Either because the distribution is right-skewed, or because of the outliers, the average number of miles of road will be greater than the median number of miles of road. The median marks the 50% point: this means that more than half of the states will have number of miles of roads less than the average value. So the answer is "true".

4)

a) The 1990 U.S. Census recorded the number of housing units in each state and the District of Columbia that used a sewer tank for sewage disposal. The average number of such homes, by state, was 1499000 (one million, four hundred ninety nine thousand) and the standard deviation was 1800000 (one million eight hundred thousand). Is the distribution for the number of housing units in a state using a sewer symmetric, right-skewed, or left-skewed, or is there no way to tell for certain with the data given? Explain your answer.YELLOW

a) The 1990 U.S. Census recorded the number of housing units in each state and the District of Columbia that used a septic tank for sewage disposal. The average number of homes, by state, was 483743 and the standard deviation was 409755. Is the distribution *for the number of housing units in a state using a septic tank symmetric, right-skewed, or left-skewed, or is there no way to tell for certain with the data given? Explain your answer.BLUE*

#4 was meant to be the hard one, and that's exactly how it turned out. However, (a) is not all that tricky. Whether we're counting sewers (Yellow) or septic tanks (Blue), the number can't be less than 0. In both cases, the SD is about the same size as the average. If the distribution were symmetric, approximately 95% of the states would fall within two SDs of average. But that's impossible, since having even 1 (or so) SDs less than average puts you in negative territory. And so the distribution can't be symmetric. Its tail can't go to the left; the SD measures spread about the average, and so we know that there must be points quite a ways away from average. Since these points can't be below average (which would make them negative values -- an impossibility), they must be above the average. Hence the distribution, for both sewers and septic tanks, must be right-skewed.

b) There are three methods listed for sewage disposal: septic tank, sewer, other. Is the correlation between the number of homes in each state using sewers and the number of homes in each state using septic tanks positive, negative, or close to zero? Why?YELLOW.
b) There are three methods listed for sewage disposal: septic tank, sewer, other. Is the correlation between the number of homes in a state using septic tanks and the number of homes in a state using sewers positive, negative, or close to zero? Why?

OK. Here's the tricky one. Kudos to those who got this right, or got partial credit. There are several ways of thinking about it, but it boils down to this: you have no data, so which of the three choices is most likely? Suppose a state has many more septic tanks than average? Why might it have more than another state? Remember we're talking absolute numbers here. And the numbers are large (the average is 483,743). One good reason why a state might have a lot more septic tanks than the others is that it has a lot more houses than another state. So we'd expect heavily populated states to have lots of septic tanks. And also lots of sewers. So if a state has more than the average number of septic tanks, it is likely to have more than the average number of sewers. And if it has fewer than average septic tanks, its likely to have fewer than average sewers. Now this might not be true of all states: there might be states that have a great number of sewers (More than average), but have lower than average septic tanks. But the correlation measures the general trend -- its okay to have occaisonal exceptions.

So the correlation is most likely to be positive. (In fact, it is positive and moderately strong.)

A zero correlation is difficult to defend. It means that there's no relationship at all between these two numbers. But because we know that states vary in the number of residents, we'd expect for no other reason than that that there would be a relationship between the number of homes with septic tanks and the number of homes with sewers.

A negative correlation could happen only if there was some program through which large states tended to use one kind of system, and small states tended to use the other. This is possible, but not likely.

c) Suppose instead of looking at the number of housing units using a septic tank or sewer, we looked at the percent of housing units using septic tanks and compared to the percent of housing units using sewers. Is the correlation positive, negative, or close to zero? Why?

Now the correlation must be negative. The percent must add up to 100 within each state. So if a state has a higher than average percentage of sewers, it MUST have a lower than average percentage of septic tanks. And vice versa.