HW 2 Solutions

#A: (a) Using the data from my songlists: http://www.stat.ucla.edu/~rgould/120w05/datasets/songlist.txt, do a statistical test to determine whether the mean song lenghs is different for rock than for classical. State all assumptions, and in particular be clear about what the population is that this hypothesis applies to. Talk about how you decided to handle any outliers. Does it matter how the outliers are handled here? Check any assumptions you can. For example, if you assumed the times came from a normal distribution, can you check whether that seems reasonable? (There's no need to get fancy with this problem and do a non-parametric or bootstrap -- but you can if you want. The point is to just explore and practice with R a bit.)
(b) Compute a 95% confidence interval for the mean difference in lenghts of songs between these two groups. Again, state any assumptions you make, but not need to check them out.

Issues:
What's the population? Some people said it consists of all rock and classical songs. But if that's the case, I can guarantee you that mine is not a representative sample. (Remember that this sample came from a list of songs stored on my hard drive.) A better choice for a population (although much less interesting) is simply the songs on my hard drive.

Is it OK to remove the outlier? Well, for the t-test it is probably necessary, although a shame. It's necessary because otherwise you really can't assume normality for the rock songs. But its a shame because one should expect tracks like that in the population. In other words, although unusual, this length is not an aberration. As many of you noted, it doesn't much matter. You get the same conclusion with or without.

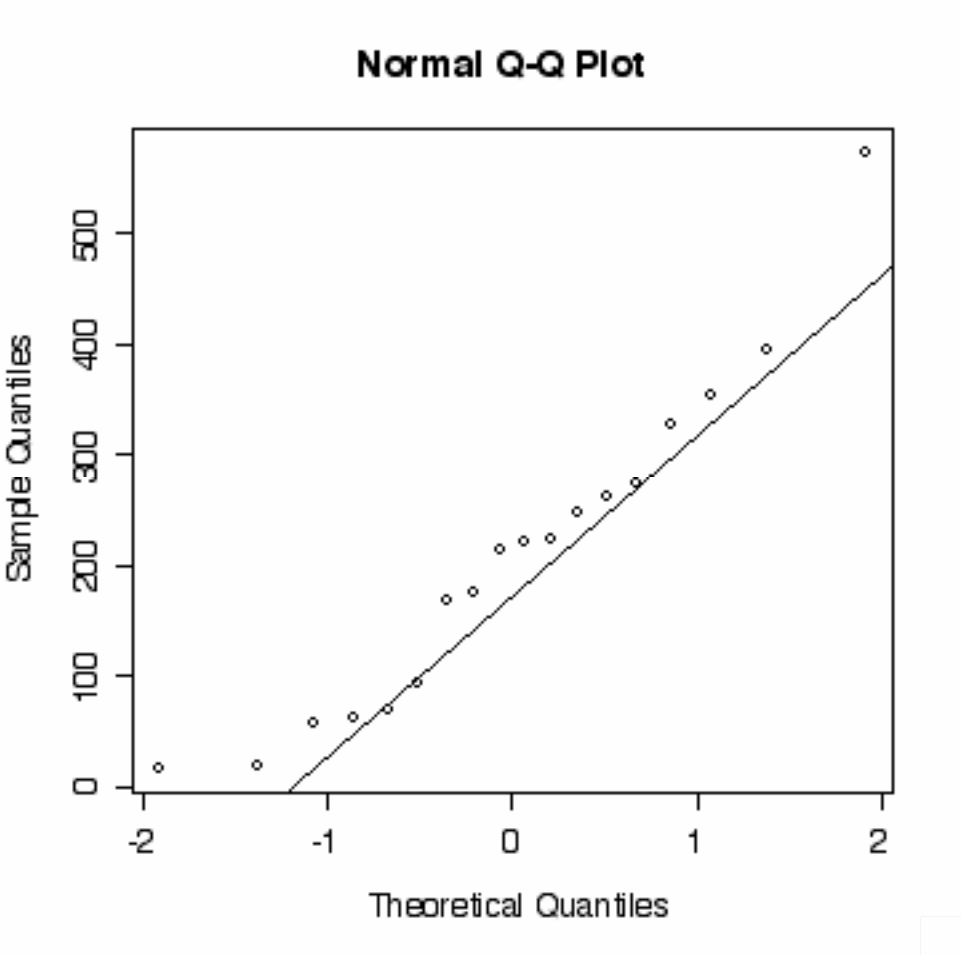
Solution:
The assumptions are that the both samples are randomly selected from their respective populations, and that not only is each observation independent of others in its population, but the two samples are independent of each other. This last idea, that the samples are independent of each other, is very important, as we'll see in #B. We also assume that the populations are normally distributed. Thus, if we looked at all of the rock songs on my harddrive, they would follow a normal distribution.

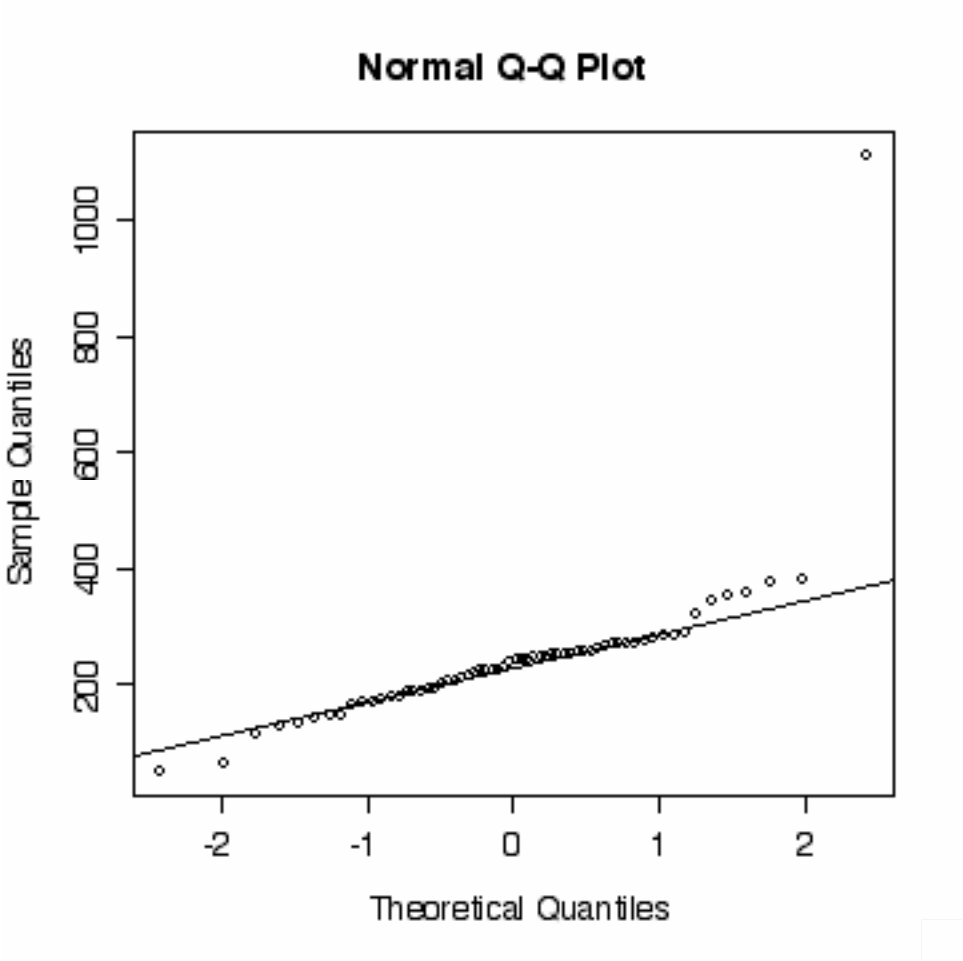
Check of Assumptions:
We don't really know how the data are collected, and so the independence assumption (within each sample) is hard to verify, as is the random sampling assumption. In reality, if these are violated there's very little we can do. (Particularly if the sample was not random.) Here, however, we'll proceed, but keep in mind that our conclusiosn could be an artifact of this poor sample. It does seem likely that the rock sample is independent of the classical sample.

We can get some insight into the normality assumption by checking to see whether the sample is close to normal. (Keep in mind that for small samples, you'll often get a non-normal appearing sample even though the population really is normal.) Histograms are a good first step (as are density curves), but qqnorm is best. These commands produce the following graphs:

```
> qqnorm(class)
```

```
> qqline(class)
> qqnorm(rock)
> qqline(rock)
```





Neither sample looks very normal. The rock sample is definitely NOT normal. We can fix the rock sample by taking out the outlier. But the classical sample is relatively small (n = 18) and non-normal. This could be a problem. (Some people said that because the sample sizes were large, you could assume normality. Typically, we want n>30 to make this assumption, and even then we could get into trouble.)

What to do? well, in truth the t-test is probably not the right test here. (If I were doing this for consulting, I'd do a permutation test -- but that's an extra-credit sort of problem.) So lets proceed.

We test the null hypothesis that the mean length of rock songs in my library is the same as the mean length of classical songs. Our alternative hypothesis is that they are not equal (which is also called a "two tailed" hypothesis.) We test at a 5% significance level. This means that we will make a decision that will result in incorrectly rejecting the null hypothesis 5% of the time. (There is a 5% probability we will say that the means are different even though they really are not.) In practice this means we reject the null hypothesis if the p-value is less than .05.

```
> t.test(rock, class)
```

```
     Welch Two Sample t-test

data:  rock and class
t = 0.8694, df = 25.359, p-value = 0.3928
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -45.50633 112.07622
sample estimates:
mean of x mean of y
 241.4516  208.1667
```

This is the test that includes the outlier in rock.

The p-value is large (certainly bigger than 5%) we means there is no evidence that the means differ.

The same conclusion is reached if we remove the outlier, although note that the p-value does increase:

```
> max(rock)
[1] 1111
> rock2 <- rock[rock<1111]
> max(rock2)
[1] 380
> t.test(rock2,class)

     Welch Two Sample t-test

data:  rock2 and class
t = 0.5351, df = 19.145, p-value = 0.5988
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -55.36992  93.43003
sample estimates:
mean of x mean of y
 227.1967  208.1667
```

I believe I started a permutation tests on one of the handouts. The permutation test proceeds as if the length for each song had a label "classical" or "rock" attached. We strip the labels off the songs, shuffle them around, and then randomly assign them to songs. Then we compare the mean lengths. In principle, we can compute the probability of getting any particular mean difference from this procedure, but in practice this isn't so

easy. An alternative is to just do a large number of runs, each time randomly assigning labels to songs, and then build up a distribution of what the mean difference would look like if "rock" and "classical" were arbitrary and had no relationship to the length of the song.

Here's a function that does this algorithm of approximating a permutation test. After writing the function, I run it with class and rock as inputs. The result is a list that contains two items. One item, named permuted.means contains the difference in average "rock" song and average "classical" song, when those labels were randomly assigned. The other item in the list, test.stat, is the difference in averages in the real sample. We want to know what the probability is of getting a difference as extreme as this, which means bigger than 33 or less than -33. In this particular simulation, this happened 383 times, and so we estimated that the probability of seeing a difference this big or bigger is about 38.3% when there really is no difference between rock and classical songs. So we again reject the null hypothesis.

```
> permute.fun
function(x,y,reps){
storage <- c()
observed.mean <- mean(x)-mean(y)
n <- length(x)
m <- length(y)
for (i in 1:reps){
     permute <- sample(c(x,y), replace=F)
     temp.rock <- permute[1:n]
     temp.class <- permute[(n+1):(n+m)]
     temp.mean <- mean(temp.rock)-mean(temp.class)
storage <- c(temp.mean, storage)}
list(permuted.means=storage, test.stat = observed.mean)}

> output <- permute.fun(rock,class,1000)
> names(output)
[1] "permuted.means" "test.stat"
> output$test.stat
[1] 33.28495
> length(output$permuted.means)
[1] 1000
> x <- output$permuted.means
> length(x[x < -33.28]) + length(x[x > 33.28])
[1] 383
> 383/1000
[1] 0.383
```

#B: Upload this dataset into a data frame: http://www.stat.ucla.edu/~rgould/datasets/bloodpressure.dat. (It's a tab-delimited file -- you'll need ot save it to your harddrive.) It consists of the systolic and diastolic blood pressure of 15 patients with moderate hypertension. Measurements were taken immediately before and two hours after taking 25 mg of th e drug captopril. The drug is expected to lower blood pressures. Does it work? Discuss. Explain.

Many of you began your analysis with graphs and descriptive statistics, and this is wonderful. But before I get into that, lets talk about what our goals for this analysis are.

The drug is intended to lower bloodpressure. We don't expect it to work the same on everyone, so we'll base our analyses on the "typical" response. But what is a response?

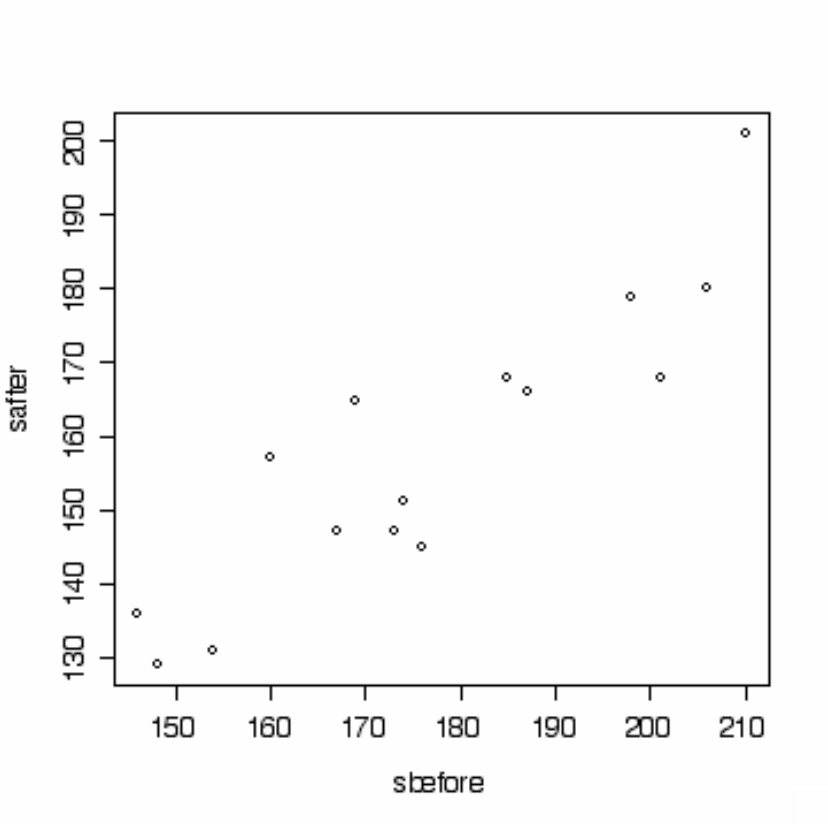
Nearly everyone decided to treat systolic separately from diastolic, and that's good. But now what? Side-by-side boxplots? If you do this, you see that, regardless of whether you look at systolic or diastolic, the before and after values have about the same amount of spread but the after group has lower medians. For systolic, the median is about 20 points lower after taking the drug.

But are these differences because of the drug, or are they just due to chance? We see quite a bit of variability in individuals' response to the drug, and so maybe the median went down this time, but could go up next time?

A hypothesis test will resolve this aspect of the problem. Our null hypothesis is the skeptical hypothesis that the mean blood pressure is unchanged. The population here consists of all people who might ever take this drug and have their blood pressure monitored before and after in this fashion.

Let's go through the assumptions. (Almost no one covered the assumptions -- but this is the most important part of a statistical test.) First, how were the data collected? Probably they are not a random sample. Random samples in medicine are rare. We are forced to "pretend" that they are. (But see alternatives at the end.) We assume that measurements are taken indpendently. We are forced to assume this, but have no idea if it's true. It could be violated, for example, if the nurse who made the measurements consistently made the same mistake, or changed his or her values based on some opinion, or if the equipment was misfunctioning somehow.

A two-sample t-test also assumes that the "before" group is independent of the "after" group. But this doesn't sound right at all. If someone has higher than average blood pressure before, we might expect him or her to have higher-than-average blood pressure after, too (although hopefully it's lower than what it was at the start of the study.) And we can see this from a scatterplot (shown for systolic only):

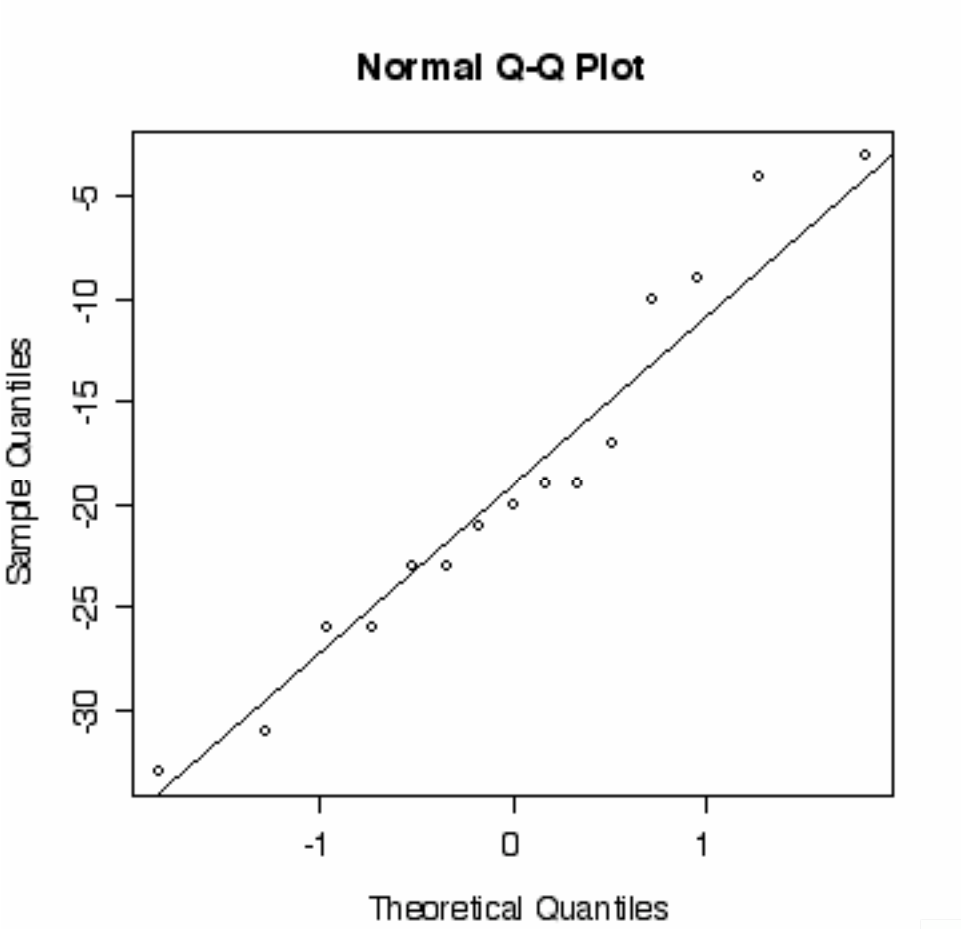


The trend looks linear, and so it makes sense to look at the correlation. Here r = .9 (about). So the assumption of independence is completely violated.

This is easy to fix, however. Instead of looking at the "before" and "after" separately, lets look at how each individual changes. The sdif variable contains each patients blood pressure after the drug minus blood pressure before --- it records the change that each patient experienced. Now there is only one variable. Our null hypothesis is that there is no change, so that the mean of this variable is 0. Our alternative is that it is negative. Again, this mean refers to the population of all patients who might take this drug.

This is called a "paired t-test" although really it is computed as an ordinary one-sample t-test. But we still have to check the assumption of normality.

```
> qqnorm(sdif)
> qqline(sdif)
```



These look pretty normal! This part of the test, at least, is straightforward.

The t-test is called using the same r function as for the two-sample:

```
> t.test(sdif)

     One Sample t-test

data:  sdif
t = -8.1228, df = 14, p-value = 1.146e-06
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -23.93258 -13.93409
sample estimates:
mean of x
-18.93333
```

We see that the mean effect was to lower the blood pressure by an average of almost 19 points. And with such a low p-value, this is extremely unlikely to be due to chance alone. We therefore reject the null hypothesis and claim that the mean change is in fact "real".

The question asked for you to conclude if the drug worked. Our t-test only tells us that the mean change is so large that it can't (reasonably) be attributed to chance alone. But this doesnt' mean the drug did it.

As many of you pointed out, this study is flawed from a lack of a control group. We really should compare how this group of patients changed compared to a group that received a placebo (or other comparison treatment). Otherwise, the large change we observed could be due to the psychological reaction of taking the drug, or being in the doctor's office, or any number of reasons.

DISCUSSION ABOUT RANDOM SAMPLE:

The entire t-test was predicated on the assumption that these patients were randomly selected from the population. This is almost certainly not true. It is not too implausible that they can be thought of as a random sample from some population, but maybe this is simply the population of patients who visit these particular doctors? Or of white middle aged males? (A criticism of medical research is that it has generalized results learned form white men to other diverse populations. In fact, a recent New Yorker article points out that a large number of medicines used on children have never been tested on children and instead the effectiveness is inferred (sometimes disasterously so) from studies done on adults). So what to do?

One example is to change our definition of the population to something less ambitious. We can say that the population consists of these same 15 patients but consists of all possible measurements we might take of their "before" and "after" blood pressure. This is a little like the population we deal with when flipping a coin. We flip a coin 10 times and estimate the probability of heads. The population consists of all possible flips of the coin. In this problem, we know that even if a patient's blood pressure is constant (which it's not), if we take the blood pressure several times, we'll get slightly different results due to measurement error. We also know that blood pressure changes constantly due to many unknown and unmeasurable reasons. Thus, we can think of a patient's blood pressure reading as a random sample selected from all possible readings for that patient.

Another approach is to do what is called a non-parametric approach. The population is the same -- just these 15 patients, but we say, "look, if the drug does nothing at all, the next measurement will either be higher or lower (or the same) as the first measurement. So lets say that there's a 50% chance it will be lower. We treat each patient's change in blood pressure as a flip of a fair coin. If we flip a fair coin 15 times, we expect 7.5 heads. In this sample, out of the 15 subjects, all 15 saw their blood pressure go down. If this were a coin, it would be like flipping it 15 times and getting 15 heads. What's the probability of that happening? Using the binomial distribution it's

```
> dbinom(15, 15, .5)
[1] 3.051758e-05
```

In other words, very unlikely. So we again conclude that the change was too large to be due to chance variation.

#C: Upload http://www.stat.ucla.edu/~rgould/120w05/datasets/anscombe.txt into a data frame. It's artificial data, so the context isn't important. (a) Without making any graphs, find the correlations between the following variables: X and Y1, X and Y2, X and Y3,

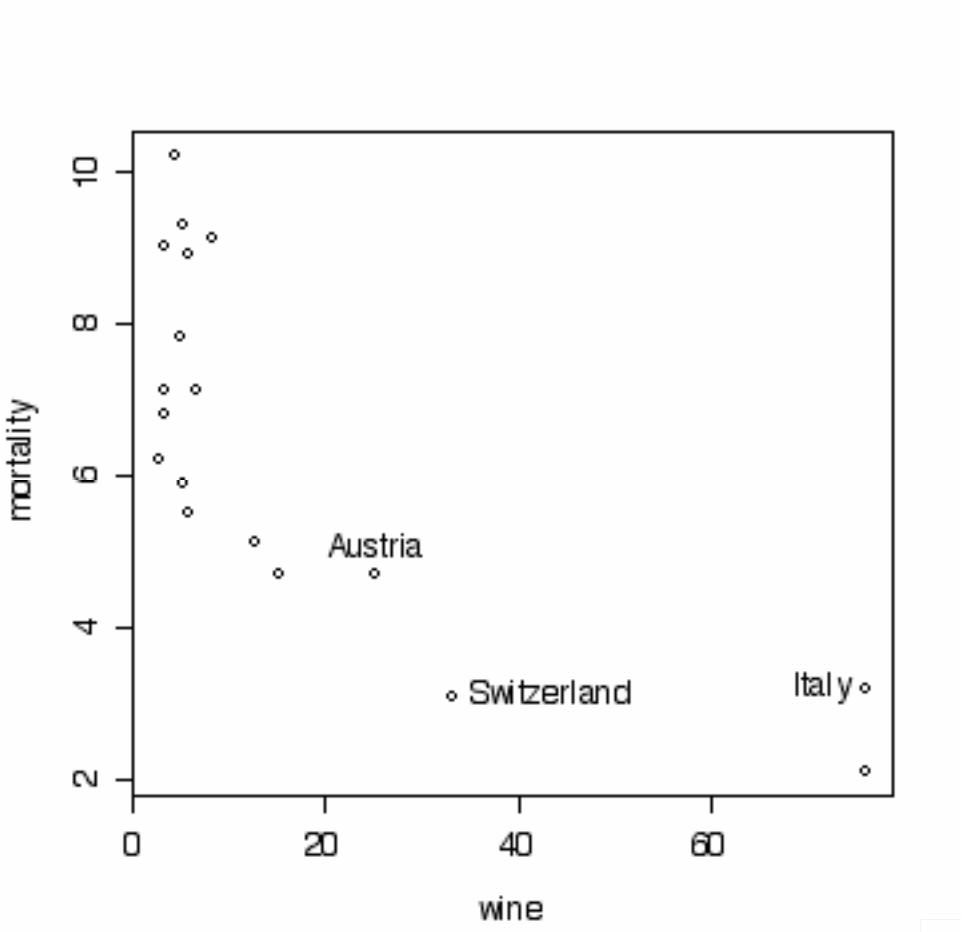
X4 and Y4. (Example, cor(X, Y1) will do it.) Comment. (b) Now make scatterplots of these pairs. (One quick way is to do "plot(dataframe)" where *dataframe* is whatever you named it. Note that this gives you too many plots (for one thing, it not only gives Y1 vs. X, but also X vs. Y1. But it also gives X vs. X4, for example.) Comment. The moral of this activity is that you need to do plots before computing correlations.

Some of you did too many plots. The problem only asked you to look at 4 plots. The others are meaningless.

The important point to note is that all 4 plots have the same correlation (at least to the first three or four decimal places.) But when you look at the plots, they do not have the same shape. In fact, only one plot is linear. The others are either quadratic or have outliers. The moral here is that if you don't look at a scatterplot, the correlation is meaningless.

BEWARE of one big potential mistake. The correlation does NOT measure linearity!! It is not correct to say that because the correlation is large, the plots must be linear. The correct thing to say is that because the plots are not linear, the correlation is not meaningful.

#D. Upload http://www.stat.ucla.edu/~rgould/120w05/datasets/wine.txt into a data frame. (Again, a tab-delimited, text file, so do
wineframe <- read.table("wine.txt", header=T, sep="\t") to read it in.) This dataset contains heart disease mortality per 1000 people, wine consumption in liters per person per year for a number of different countries. Make a scatterplot of mortality against wine consumption and comment on the shape of the relationship and what this says about wine consumption and health. Do you think the relationship is causal? Why or why not?



The relationship is non-linear, and shows a strong negative association between per-capita wine consumption and mortality. Generally, countries that drink more wine have lower (heart-releated) mortality. NOTE that because this is non-linear it is INCORRECT (or at least misleading) to talk about the correlation, as in "there is a strong negative correlation between....".

Now there seems to be some confusion about what is meant by "causal". While I'm not happy to see confusion, I'm glad we get a chance to discuss it, and I hope you feel free to bring up your ideas. Philosophers have been debating about what this means, exactly, for centuries. But here's one definition:

x causes y if, when we change x, y changes.

Many thinkers on the subject have modified this to a defiition more friendly towards statistics:

x causes y if, when we change x, there is a large probability that y will change.

And of course, we have to worry about what "large" means.

But note that however you slice it, you have to change x to conclude that changes in x cause changes in y.

These data represent a (non-random) sample of countries from the world. (Note that they represent everyone in those countries--not a sample of 1000 people as some of you thought. It's just the case that mortality is given in units of deaths per 1000 people.) This isn't really relevant to this question, though. What matters here is that this was an observational study. To decide if wine consumption improves mortality, we would have to get one (well, several) countries to increase or decrease their wine consumption, and then observe the effect on mortality. And we would have to do so in a way that avoids "confounding variables". Confounding variables are variables that affect both the predictor variable (wine consumption) and the response. For example, in these data, diet can very possibly be a confounding variable. Countries that drink lots of wine also have a certain type of diet (perhaps they eat less red meat) that also leads to a decrease in heart disease. This means that you yourself could drink as much wine as you want, but it's not going to help unless you change to a diet exactly like Italy's. There are many other possible confounders as well. The golden rule of causality (well, one of the many golden rules) is that unless you have a carefully designed, controlled study, you can't eliminate the possibility that any associations you observe (such as this one between wine and mortality) are due to unseen confounding variables.

Many of you commented that you either did or did not believe in causality because of other studies you had read or heard about. It is always good to compare what you are seeing to what you already know. However, also be sure to comment on what the strengths and limitations are of *this* data set. The question was really wondering if you could reach a causal conclusion based on these data. (Note: the jury is still out about whether alcohol consumption is healthy in moderation. There's strong associations between drinking and health, and the general trend is that moderate drinkers are healthier than abstainers who are healthier than heavy drinkers, and this seems to be true regardless of the type of alcohol. But it is difficult to completely eliminate all confounding variables, and so your doctor will probability not be advising you to drink every day anytime soon. On the other hand, there is some evidence of a biological mechanism that could explain why moderate amounts of alcohol are beneficial to some people: antioxidants appear in some alcohols, and also seem to decrease the risk of some cancers. There is also some evidence that for some people, moderate amounts of alcohol "thin" the blood and might therefore decrease the risk of heart disease.)

(Reminder: a controlled study is one in which the researchers were able to determine (hopefully by implementing some sort of randomization procedure) which subjects received which treatments. It is an observational study if they did not or could not do this.)

Things to keep in mind: describe plots in the context of the data. Don't simply say "there's a negative association", but instead say what this means: "countries that have high wine consumption tend to have low mortality." Also, always interpret the data using whatever information you can about how the data were collected.