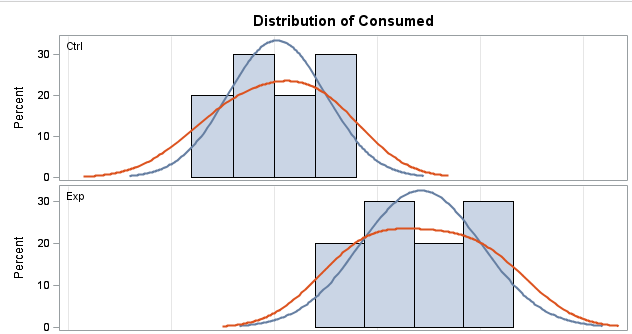
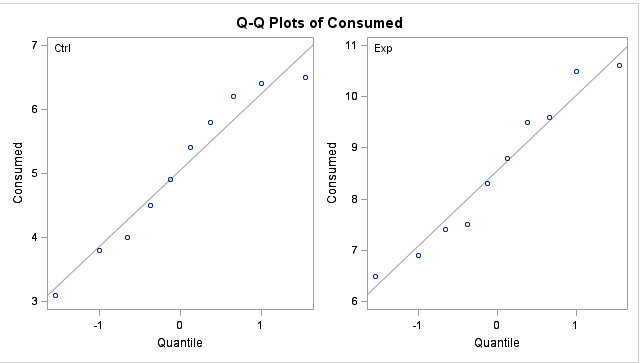
Remy Lagrois

Final Section 402

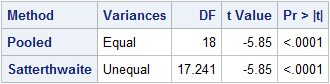
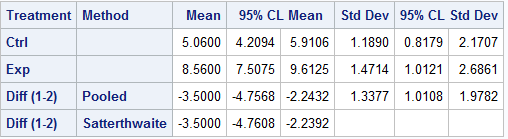
1. Using PROC POWER and the given parameters a sample size of 7 would be needed.
2. a) Since we don’t know what the alpha used for the test was we cannot compute an exact critical-t. Assuming they used an alpha between 0.01 and 0.1 then the range of possible p-values would be .0108 to .1104 (at 0.05 the p value would be .0503). In all of these cases they’d be unable to reject the null hypothesis.

b) The men were randomly assigned to each treatment group which means the different weights of the men should be pretty evenly distributed between the groups. One group won’t have mostly really heavy men while the other has mostly less heavy men, the distribution of weight within the groups should look very similar to each other.

1. The null hypothesis for this test is that the difference between the means (control – experimental) is zero. The alternative hypothesis is that the difference between means (control – experimental) is less than zero (ie experimental consumption is higher than control.



Neither group deviates far enough from normal to provide evidence they are not normally distributed (see QQ plot), this is further backed up by the histogram. The same histogram also shows similar distribution and standard deviation. The variance is confirmed to by equal by the F-test p-value of 0.5356.



The low p-value of <.0001 tells us to reject the null hypothesis that there is no difference between the means, we instead accept that the control group consumed more of the lead-acetate solution on average. The confidence interval indicates that the calcium deficient rats will consume as an average 2.243 to 4.757 (unknown units) more than rats with a normal calcium intake (95% of the time). Since the treatment was randomly assigned we can conclude the calcium deficiency is causal. However, we don’t know from what population or how the rats in the experiment were selected so we can only apply these findings to those rats which were used.

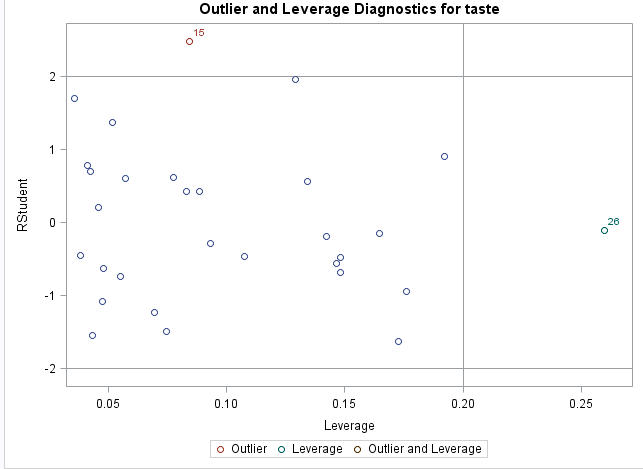
1. There are no situations where you can use your data to assess lack of model fit to the data.
2. **As long as the sample size is large enough, the F-Test for the equality of variance is robust to normality.**
3. **Adding more variables to the model will always increase the R2.**
4. An r = 0 indicates that that there is no relationship between x and y.
5. A causal relationship can be concluded if the p-value associated with the test is <0.001.
6. Type 1 error can be controlled by the researcher while type 2 error cannot.
7. **The coefficient of a categorical variable with two levels represents the difference between the group means.**
8. Paul Minton was the first chair for the Statistics Department. SMU alumni include Lamar Smith who currently serves in the House of Representatives and the award winning actress Kathy Bates.
9. a) The response variable is taste

b) taste = -26.94 + 3.80(acetic) + 5.15(H2S)

c) Yes the overall regression equation is significant, the F value of 18.81 gives a p-value of <.0001 so we reject the null hypothesis that all slopes are zero.

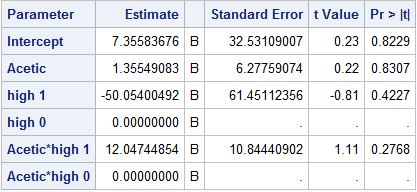
d) The slope for H2S is significant but the acetic acid slope is not. The H2S slope has a p-value of 0.0002 while the acetic slope’s is 0.4062. Therefore we can reject the null hypothesis that the slope is zero for H2S but not for acetic.

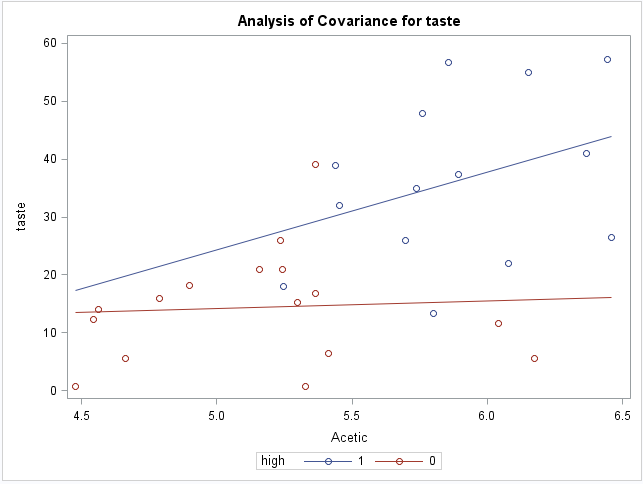
e)The slope for H2S is 5.15 with a 95% CI of 2.66-7.62. Since the slope H2S is positive the taste score will increase as the concentration of H2S increases; for each 10% increase in the concentration of H2S the taste score should increase by 0.491 points. Acetic acid’s slope of 3.80 is also positive, however it’s 95% CI ranges from -5.44 to 13.045 and the p-value indicates it isn’t significantly different from zero. However if we assume the 3.80 number is right then increasing the acetic acid concentration also increases the taste score; a 10% increase in acetic acid concentration increases the taste score by .362 points.

f)

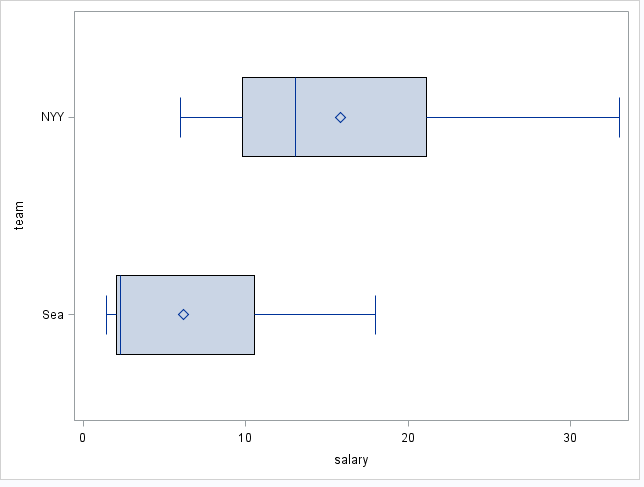
Looking at the studentized residuals plotted against leverage we can see that case 26 has high leverage which means its predictors have a value much different than the average value for the dataset. However it also has a residual of nearly zero and a low Cook’s D which indicates it doesn’t have undue influence on the model.

g) The predicted taste score would be 28.767 and the 95% PI is 17.885 to 39.649

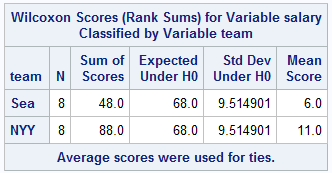
h)Yes, the acetic acid slope with high H2S is much larger than without



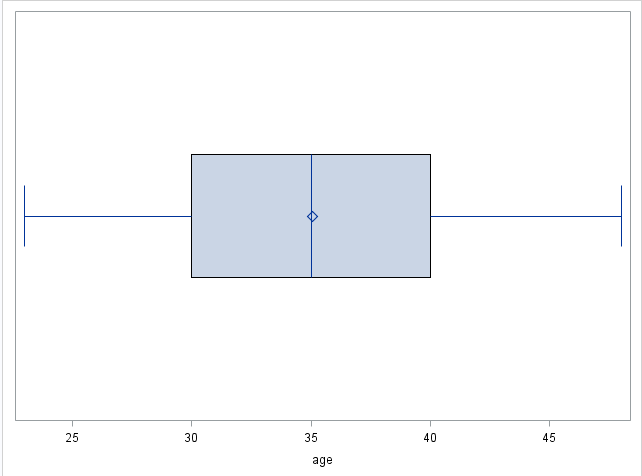
1. a) A rank of 9.5 means that the values of the 9th and 10th places of the ordered data were the same (in this case the salary for the NYY left and center fielders which is 13). When that happens, the score assigned to the ranks with equal values is the average of those ranks. So, instead of one 13 having a score of 9 and the other 10, they both get 9.5.

b)The null hypothesis is that the median salary of the two teams are the same while the alternative is that the medians are not equal. 

The salaries for the two teams have similar distributions and variance so there is no evidence of those assumptions being violated. Also since there is no salary cap in baseball the salaries are also independent.



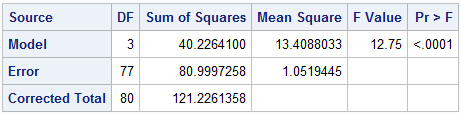
The sum of ranks for Seattle was 48 while the sum for the Yankees was 88. The z-score for this test was -2.0494 which gave a two sided p-value of 0.0404. We therefore reject the null hypothesis and conclude that the median salaries of the two teams are not equal. We can further conclude that the New York Yankees have a higher median salary.

1.  This is a box plot of the age of the salesmen. It shows that the mean and median are both around 35, the minimum is just under 25 and the max is just under 50. Half of them (ie 50%) are between the ages of 30 and 40; it appears the remaining 50% are evenly split between younger than 30 and older than 40. The ages appear to be normally distributed.

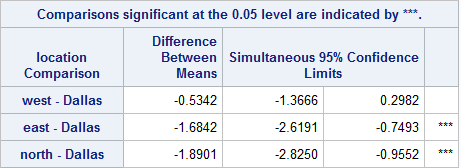
For this data a boxplot or a histogram would make sense. I chose the boxplot because you can still get an idea of the distribution while also getting more exact information as to the mean, median, min, and max (as well as 1st and 3rd quartiles). The histogram would give a better idea of the relative frequency of each age or age group but I don’t think that’s worth sacrificing the information gained from a boxplot in this case.

Bonus:

We are testing to see if the mean (average) auto interest rate in Dallas differs from the rates in the north, west, and eastern regions. In order to do this we used an alpha of 0.05, meaning there is a 95% chance the differences we see are real and not just due to chance, and looked at the confidence interval (the expected difference in rates) to determine which city/region had the higher rates.



An initial analysis of variance shows that there is a difference between at least one of these groups and making comparisons is appropriate. This is shown by the p-value of <.0001 which allows us to reject the null hypothesis that the average rate is the same for all the regions. Instead we accept the alternative that at least one of them is different from the others though this test does not tell us which one. To make the actual comparisons we ran Dunnett’s procedure on the data. This performs a t-test between each region and Dallas. The results are below:



This table shows us that there is a statistically significant difference from Dallas in the East and North. The average rate in Dallas and the West cannot be said to be different at our chosen significance level. The confidence intervals for the North and East comparisons are completely in the negative which indicates the average rate in Dallas is higher than in both of the two regions.