

Data Analysis and Statistical Methods

Statistics 651

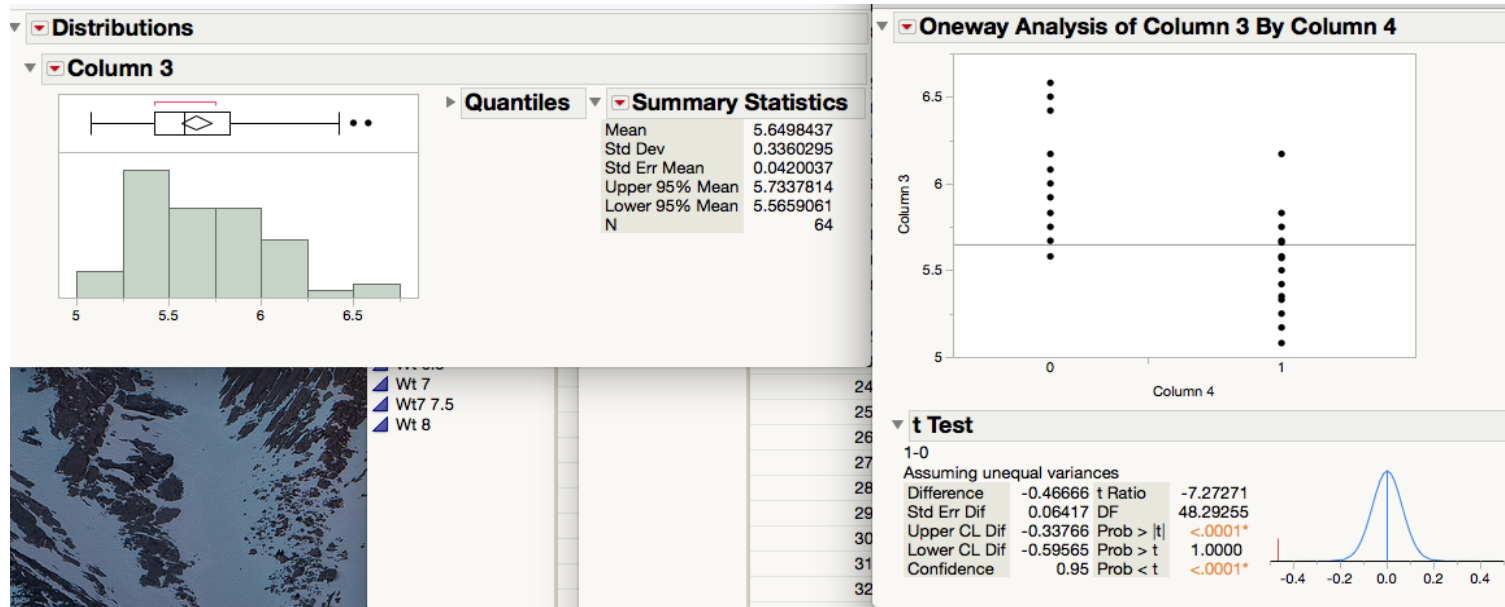
<http://www.stat.tamu.edu/~suhasini/teaching.html>

Lecture 19 (MWF) Independent two sample t-test

Suhasini Subba Rao

In general: Confidence intervals and testing

- In most software packages that you run, the estimate will come with a standard error. Below are two completely different methods. However we observe standard errors (s.e.) in both



- What we do from now onwards will be more complicated than simply analysing the sample mean.
- However, essentially, the analysis will be the same. We assume that the estimator is normally distributed. The standard error measures the variation of the estimator and can be used to locate the mean. The rule of thumb is that it is within a few standard errors of the mean.
- We can usually construct a confidence interval for the true parameter θ (θ is just a name we have given the parameter) with

$$[\text{estimate} - t_{df} \times \text{s.e.}, \text{estimate} + t_{df} \times \text{s.e.}]$$

Usually we use a t-distribution (but not always).

- If we test the hypothesis

$$H_0 : \theta = \text{number} \text{ against } H_A : \theta \neq \text{number}$$

(often number = 0), then we need to measure the distance between estimator and number with

$$t = \frac{\text{estimate} - \text{number}}{\text{s.e.}}$$

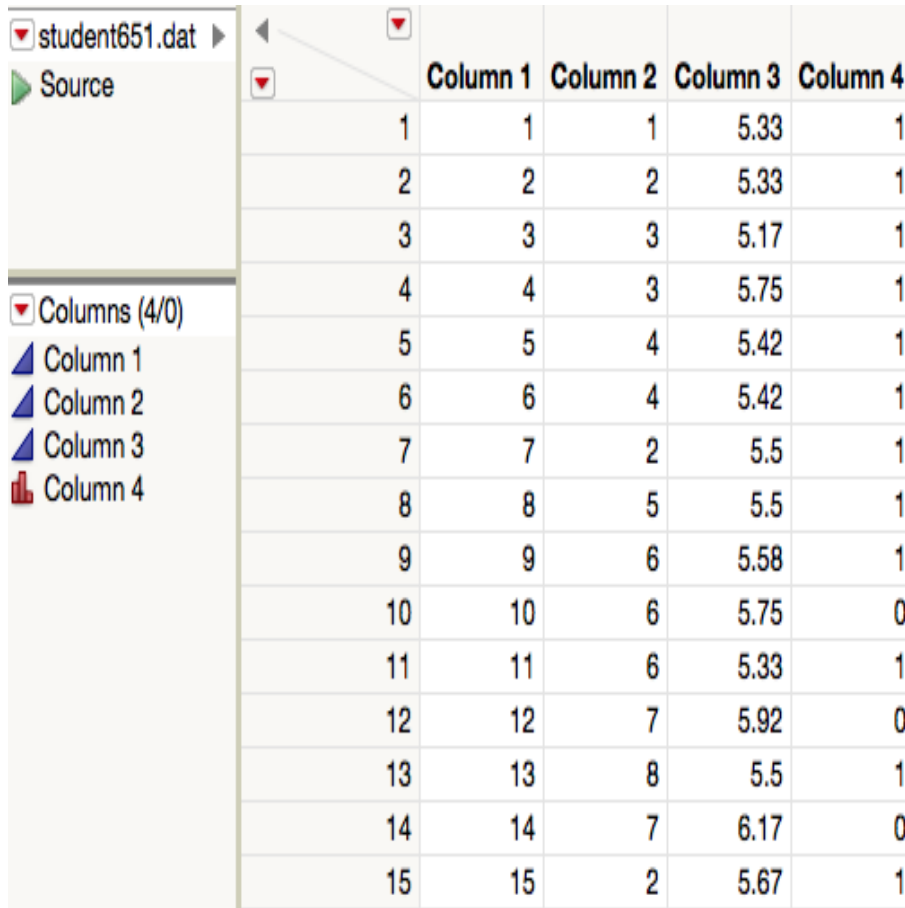
Usually the above will have a t-distribution (but not always).

Independent two sample t-test: Comparing populations

Suppose we want to compare the heights of males and females at A&M.

- Question 1: Is the mean female height less than that of the male height?
- Question 2: What is the difference in the mean male and mean female heights.
- First collect data: Then compare the sample mean of the male heights with the sample mean of female heights and use this to infer aspects on the mean height of male and female A&M students.

Snapshot of data in JMP

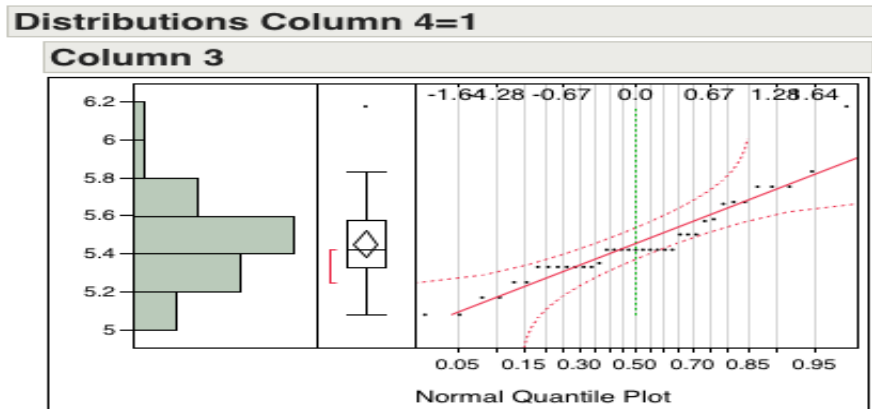
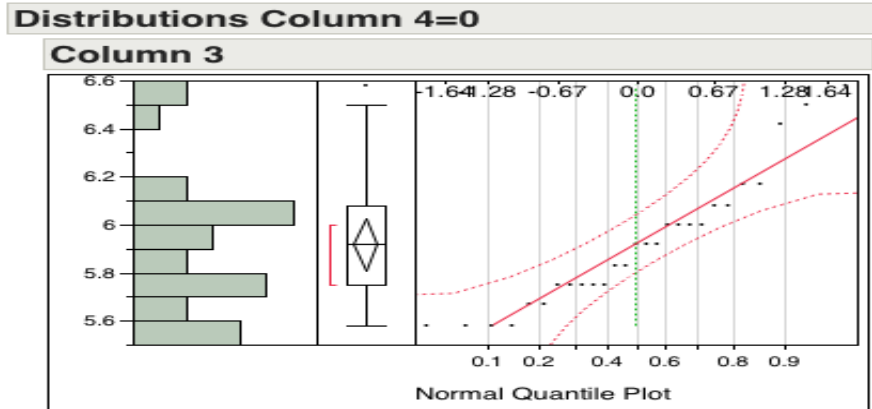


	Column 1	Column 2	Column 3	Column 4
1	1	1	5.33	1
2	2	2	5.33	1
3	3	3	5.17	1
4	4	3	5.75	1
5	5	4	5.42	1
6	6	4	5.42	1
7	7	2	5.5	1
8	8	5	5.5	1
9	9	6	5.58	1
10	10	6	5.75	0
11	11	6	5.33	1
12	12	7	5.92	0
13	13	8	5.5	1
14	14	7	6.17	0
15	15	2	5.67	1

Each row corresponds to a student. Column 3 contains the height. Column 4 their gender.

In this file 1 = Female and 0 = Male. Column 4 is a categorical/nominal variable, so make sure that JMP knows this (by changing the blue triangle to the red blob).

Plot of data collected



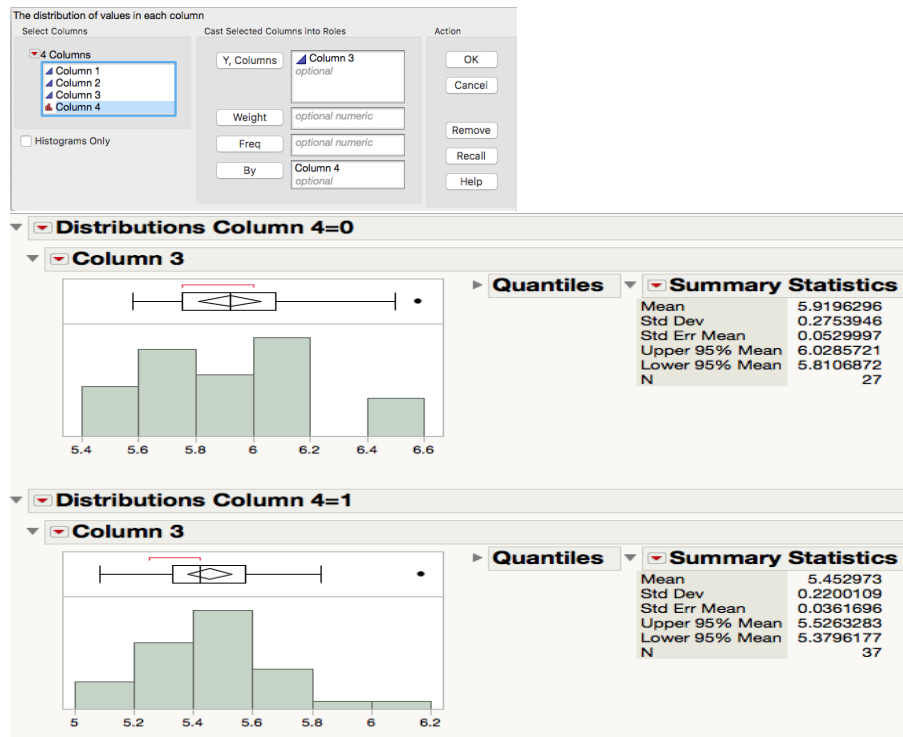
The QQplot has the horizontal lines because everyone rounded their height to the nearest inch.

Observe that each data set is not exactly normal, but we do not observe evidence of a large skew, which would slow down a CLT on the sample mean.

How these separate plots were made in JMP will become clear on the next slide.

Summary of data and instructions in JMP

- To obtain a summary of the separate male and female heights in JMP. Go to Analyze > Distribution. Then place variables in the following in the box:



Remember, By will split the data into the female and male groups.

You should then see the summary of the female and male groups.

- Let X be the height of a randomly selected female and Y the height of randomly selected male. We observe from the output that there are $n = 37$ females and $m = 27$ males in the sample.

The sample mean for females is $\bar{X} = \frac{1}{37} \sum_{i=1}^{37} X_i = 5.45$ and sample mean for males is $\bar{Y} = \frac{1}{27} \sum_{i=1}^{27} Y_i = 5.92$.

The sample standard deviations are $s_X = 0.22$ and $s_Y = 0.27$.

- Our objective is inference on the population means. Let μ_X be the female population mean height and μ_Y be the male population mean height.
- Our object of interest is the quantity $\mu_X - \mu_Y$. It will tell us how much larger or smaller females with respect to males.

Are males taller than females?

- This translates to testing $H_0 : \mu_X - \mu_Y \geq 0$ against $H_A : \mu_X - \mu_Y < 0$ or equivalently $H_0 : \mu_Y - \mu_X \leq 0$ against $H_A : \mu_Y - \mu_X > 0$
- We also want know the magnitude of the difference, this means constructing a CI for the mean difference $\mu_X - \mu_Y$.
- Clearly if $\bar{X} - \bar{Y} > 0$ we would be unable to reject the null and the p-value would be larger than 50% (recall from Lecture 15 and 16 that $\bar{X} - \bar{Y} > 0$ would lie within the non-rejection region).
- If $\bar{X} - \bar{Y} < 0$, then we should use a statistical test.
- How to make the comparison, what is the distribution of $\bar{X} - \bar{Y}$?

Objectives

- To build a confidence interval for the mean difference $\mu_X - \mu_Y$ (this will tell us where the mean difference lies).
- To test the hypothesis that $H_0 : \mu_X - \mu_Y \geq 0$ (mean female height and male height are the same or mean female height is greater than mean male height) against the alternative $H_A : \mu_X - \mu_Y < 0$.
- We can take the test further and test whether $H_0 : \mu_X - \mu_Y \geq -0.3$ against $H_A : \mu_X - \mu_Y < -0.3$.

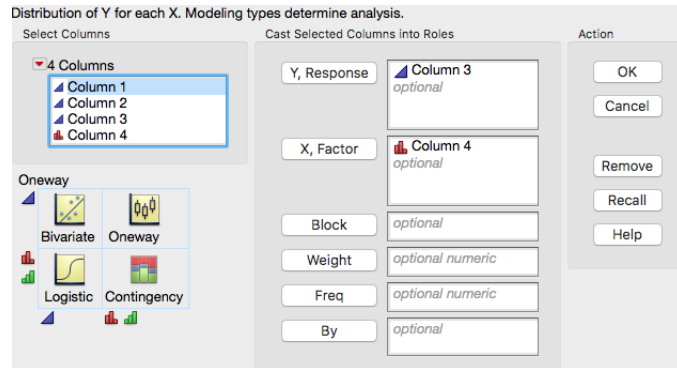
This is asking if the average males is **more** than 0.3 feet taller than the average female.

- This is called an independent two-sample t-test.

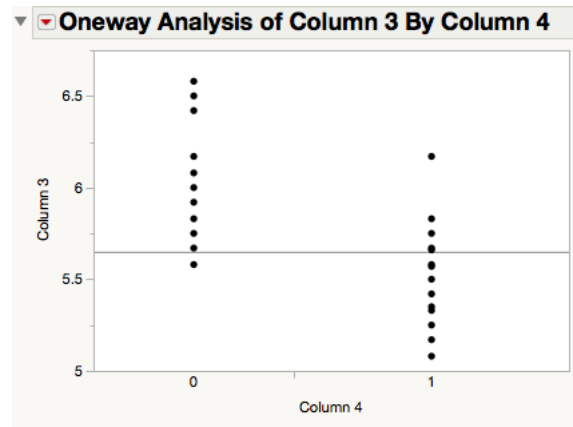
The necessary ingredients

- In order to do the test $H_0 : \mu_X - \mu_Y \geq 0$ against $H_A : \mu_X - \mu_Y < 0$ or to construct CI for $\mu_X - \mu_Y$ we need three magical ingredients:
 - (a) The difference of the sample averages: $\bar{X} - \bar{Y}$.
 - (b) The standard errors of $\bar{X} - \bar{Y}$ (this will turn out to be $\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}$, where σ_X is the standard deviation in population X and σ_Y is the standard deviation in population Y).
 - (c) \bar{X} and \bar{Y} are normally distributed. Both samples are **independent** of each other and **independent** within the sample. For example the values X_1, \dots, X_n should have no influence on Y_1, \dots, Y_m and X_1 should not have any influence on the other observations in the sample X_2, \dots, X_n .

An independent two-sample t-test in JMP

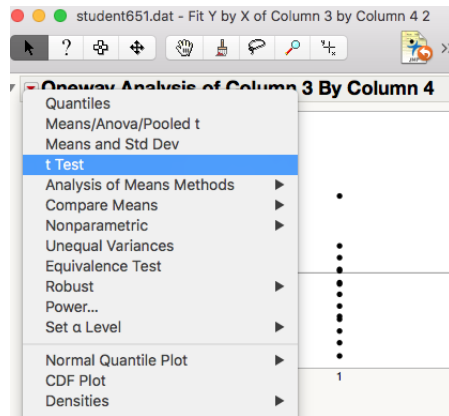


Go to Analyze > Fit Y by X. You will then see the screen shot on the right.

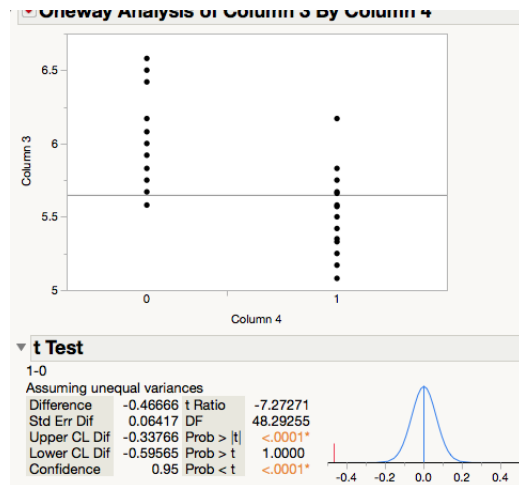


Plot of the two samples size by size. This gives you a visual guide on their differences. Remember, since there at 27 “0” heights and 37 “1” heights. So each dot can represent several students.

Lecture 19 (MWF) Independent sample t-test for testing equality of means in two populations



To do the formal t-test click on the red triangle and choose **t-Test**. You will also see Means/Anova/Pooled t. This option does the test under the assumption that both samples share the same standard deviation. We cover this later in the lecture.



Difference $\bar{X} - \bar{Y} = \text{Female} - \text{Male}$ (since is 1-0) = -0.466. Std. Err Diff = 0.064, which corresponds to the standard error of $\bar{X} - \bar{Y}$. t Ratio = $\frac{-0.466}{0.064} = -7.2$.

The t Ratio is extremely large. The p-value for the test $H_0 : \mu_X - \mu_Y \geq 0$ vs $H_A : \mu_X - \mu_Y < 0$ is $\text{Prob} < t < 0.001$ (less than 0.1%).

Concepts: Distribution of sample means $\bar{X} - \bar{Y}$

- Suppose that \bar{X} and \bar{Y} are close to normal, then the difference of the averages has the following distribution:

$$\bar{X} - \bar{Y} \sim \mathcal{N} \left(\mu_X - \mu_Y, \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} \right).$$

where σ_1 is the standard deviation of X (variability of population of X) and σ_2 is the standard deviation of Y (variability of population Y).

- Important points:
- The distribution is centered about $\mu_X - \mu_Y$, hence I am likely to draw close to $\mu_X - \mu_Y$.

- How 'close' the sample means difference is to the true means is determined by the standard error which is $\sqrt{\left(\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)}$.

The larger the sample sizes n **and** m the smaller the standard error (just like in the one sample case, where we deal with just one sample mean \bar{X} , which has standard error $\frac{\sigma^2}{n}$).

- The normality result depends (as usual) on two factors
 - (a) How close the original data sets are to normality.
 - (b) How large the two sample sizes (both need to be sufficiently large).
- If the sample size is small then the data must be close to normal. This can be checked using QQplots.

- Based on the above we make a Z -transform of the difference $\bar{X} - \bar{Y}$

$$\frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim \mathcal{N}(0, 1).$$

- In practice σ_1 and σ_2 is unknown and has to be estimated from the data.
- We estimate the two standard deviations, s_1 and s_2 from the data and plug them into

$$\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} \rightarrow \sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}$$

to give the standard error.

The distribution when the standard deviation is estimated

- We do the t-transform

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}}$$

- The distribution of the above is not known exactly, it has an approximate t-distribution with a rather complicated number of degrees of freedom.
- This is the strange DF given in the JMP output.
- JMP and most software will test $H_0 : \mu_X - \mu_Y = 0$ vs $H_A : \mu_X - \mu_Y \neq 0$. In this case replace $\mu_X - \mu_Y$ in the t-transform with 0.

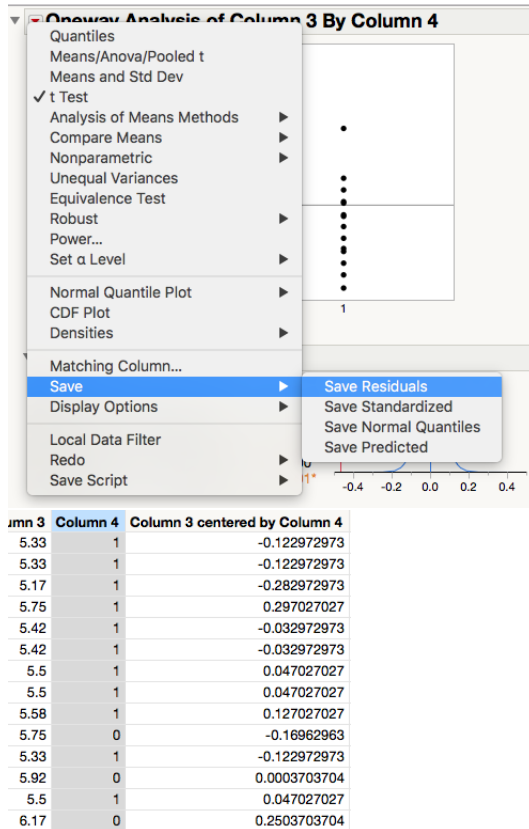
Returning to the height example: Assumptions

- Unless many of the students in the 651 class were related it is reasonable to assume that they are independent.
- To check whether the both the sample means \bar{X} (female sample average) and \bar{Y} (male sample average) are normal we return to the data.
- To check normality of the data we make a QQplot of the residuals.

What are residuals

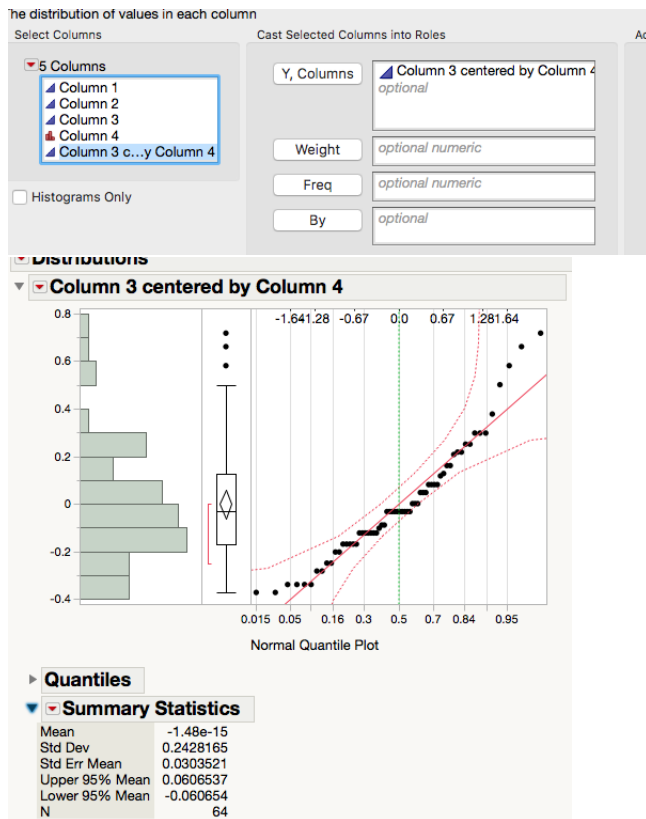
- The analysis of residuals is an important component of most data analysis.
- The residuals are the observations in each group take away the group sample mean.
- For this example it is:
 $X_i - \bar{X}$ (female heights - female average)
and $Y_i - \bar{Y}$ (male heights - male average).
- The benefit of residuals is that they can be pooled and examined altogether.

Obtaining the residuals in JMP



Go to the red triangle and choosing Save > Save Residuals. A new column is made in JMP of the residuals $X_i - \bar{X}$ and $Y_i - \bar{Y}$. Since the mean has been removed the residuals have zero sample mean (they contain no mean information).

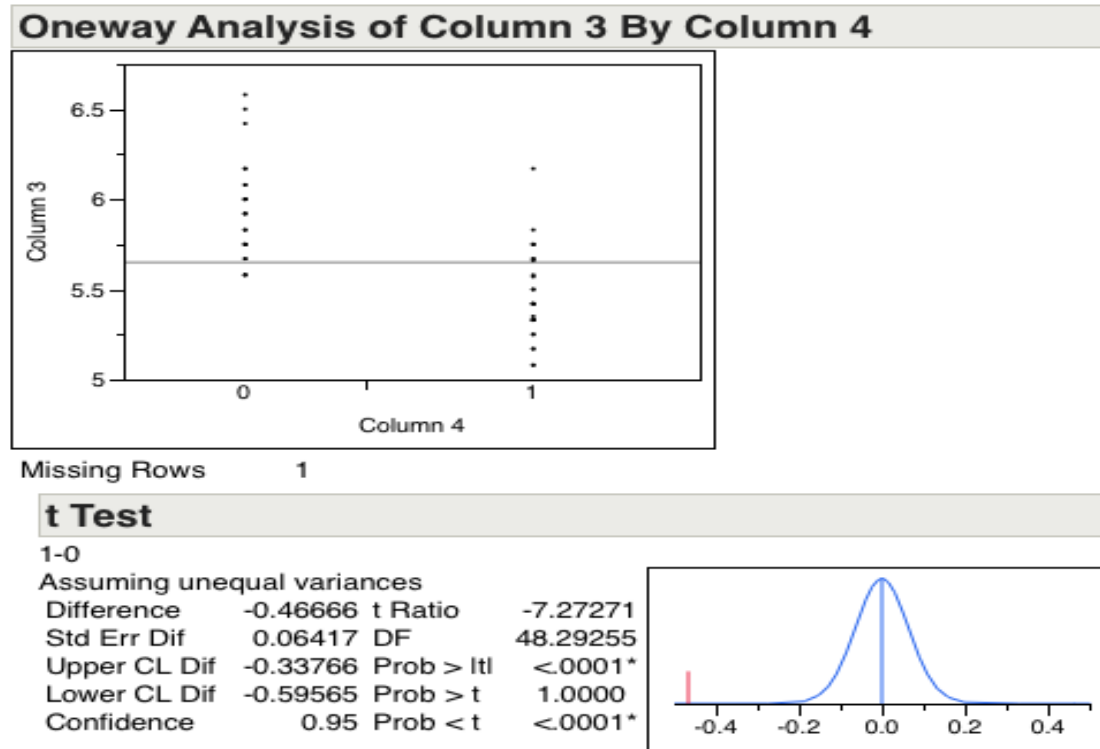
QQplot of residuals



Analyze > Distribution and select the new residual column. This will give the histogram of the residuals. To obtain the QQplot go to the red triangle and choose Quantile Normal Plot (see Lecture 10).

We observe that the residuals are not too close to normality. However, the sample sizes of 27 and 37 should make the respective sample means close to normal. Observe (as mentioned on the previous slide) that the mean is zero.

Matching the output of the t-test to the formal test



Interpreting the output

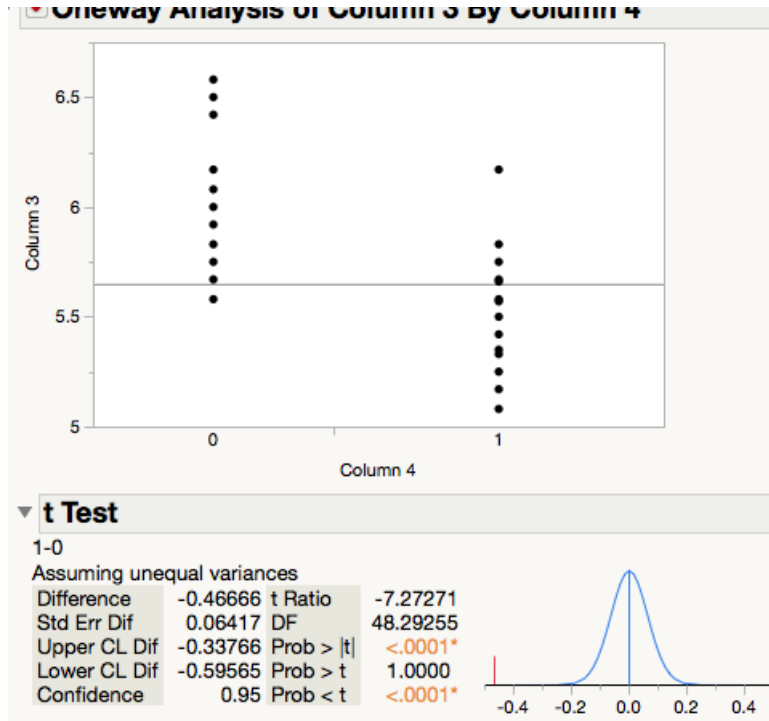
- The $df = 48.29$
- The standard error which measures the variability of the estimator of the differences is

$$\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}} = \sqrt{\frac{0.0484}{37} + \frac{0.0758}{27}} = 0.064.$$

- The difference in sample means is -0.466 .
- The t -value for the test $H_0 : \mu_X - \mu_Y \geq 0$ against $H_A : \mu_X - \mu_Y < 0$ is $= (-0.466 - 0)/0.064 = -7.27$.

- Since the alternative is pointing to the left, we need to area to the **left** of -7.27, this is $\text{Prob} < 0.0001$ (in the output). Therefore, the p-value is less than 0.0001, which means at the 1% level there is sufficient evidence to suggest that males are taller than females.
- The 95% confidence for the difference in the means is $[-0.59, -0.33]$, this means with 95% confidence we believe the mean difference in female and male heights lies between $[-0.59, -0.33]$.

Reasons for the extremely small p-value



Observe that the the difference in male and female heights is large = -0.466. Moreover the standard error is small Std. Err Dif = 0.064. The small standard error is due to the relatively large sample sizes (27 and 37).

These two ingredients lead to an extremely large t Ratio = $\frac{-0.466}{0.064} = -7.2$. This means the t -value is in the extreme tail of the t -distribution, which leads to a small p-value.

Further the CI is $[-0.59, -0.33]$ which is “far” from zero.

Using the same output for additional tests

- Suppose we believe that the mean difference is over 0.3 feet. Is there evidence in the data that this could be true:

We test $H_0 : \mu_X - \mu_Y \geq -0.3$ against $H_A : \mu_X - \mu_Y < -0.3$, we do the test at the 5% level. We calculate the t-value

$$t = \frac{-0.46 - (-0.3)}{0.064} = -2.5.$$

- To deduce the p-value we use the critical values for the t-distribution with 48.28df.

probability	0.15	0.10	0.05	0.025	0.01	0.005
t^*	1.05	1.3	1.68	2.01	2.4	2.68

- Since -2.5 lies between -2.68 and -2.4 , the area to the left of -2.5 is between 0.5% to 1% . Thus the p-value is between 0.5% - 1% . Based on this there is evidence to suggest that the mean difference between males and females is greater than 0.3 feet.

Confidence intervals at other levels

- Just as before to obtain the 99% confidence interval we use the critical value for 0.005 (remember $0.005=0.5\%$).
- We use the critical value corresponding to the t-distribution with 48.28df.

probability	0.15	0.10	0.05	0.025	0.01	0.005
t^*	1.05	1.3	1.68	2.01	2.4	2.68

- This gives the 99% confidence interval:

$$[-0.466 - 2.68 \times 0.064, -0.466 + 2.68 \times 0.064] = [-0.63, -0.29]$$

which (of course) is wider than the 95% CI, which is $[-0.6, 0.34]$.

Variants of the t-test: Pooling information

- Some improvements to the testing can procedure can made if there is reason to believe that the standard deviation in both populations is the same.
- Looking back to the comparison between male and female data we observe that the sample standard deviation of the male and female heights was 0.27 and 0.22 respectively; which are relatively close.
- If we believe that their corresponding population deviations are close or the same rather than use separate standard deviation estimates in the test, we can “pool” the information and estimate one standard deviation for both population.

Formulas behind the pooled test

- The *pooled* sample standard deviation measures the **collective spread** of the residuals ($\{X_i - \bar{X}\}$ and $\{Y_i - \bar{Y}\}$):

$$s_p = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 + \frac{1}{m-1} \sum_{j=1}^m (Y_j - \bar{Y})^2}$$

- The independent two sample t-test becomes

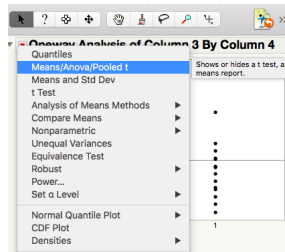
$$t = \frac{\bar{X} - \bar{Y} - (\mu_X + \mu_Y)}{\sqrt{\frac{s_p^2}{n} + \frac{s_p^2}{m}}} = \frac{\bar{X} - \bar{Y} - (\mu_X + \mu_Y)}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}}.$$

- Under the null and the assumption the sample means are normal the distribution of the t-transform t has a t-distribution with $n + m - 2$

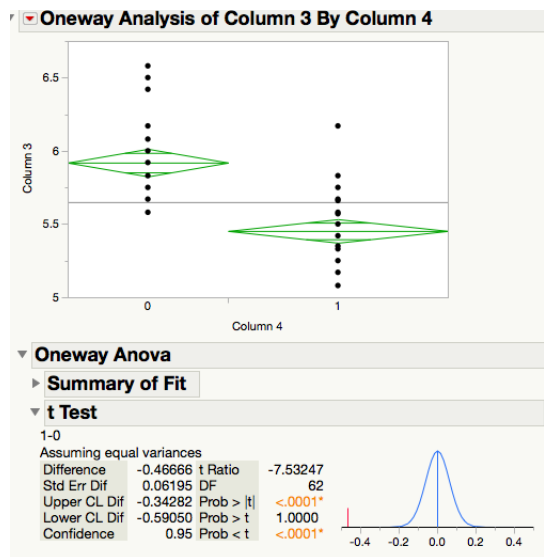
degrees of freedom. This is a little more palatable than the strange degrees of freedom we encountered above.

- If the population variances are the same, the pooled t-test gives “more correct p-values” .
- It also has more power (ability to reject the null when the alternative is true) than the regular t-test.
- One caveat is that we really have to sure that the standard deviations in both populations are the same. If they are not, the pooling method makes no sense.

The pooled t-test in JMP



To do the formal t-test click on the red triangle, do exactly what was done previously but choose **Means/Anova/Pooled t** instead of t-test.



Ignore the ANOVA stuff you see and focus on the t-test. Difference $\bar{X} - \bar{Y} = \text{Female} - \text{Male}$ (since is 1-0) = -0.466. Std. Err Diff = 0.0619, which corresponds to the standard error of $\bar{X} - \bar{Y}$.

$$t \text{ Ratio} = \frac{-0.466}{0.061} = -7.53.$$

The t Ratio is extremely large. The p-value for the test $H_0 : \mu_X - \mu_Y \geq 0$ vs $H_A : \mu_X - \mu_Y < 0$ is $\text{Prob} < t < 0.001$ (less than 0.1%). We reject the null.

Choosing the sample sizes

- The standard error is

$$\sqrt{\underbrace{\frac{\sigma_1^2}{n}}_{\text{variability of } \bar{X}} + \underbrace{\frac{\sigma_2^2}{m}}_{\text{variability of } \bar{Y}}}$$

(if the population standard deviations are known).

- As usual there are two factors which effect the size, the standard deviations σ_1 and σ_2 and the sample sizes n and m . You **cannot** control the standard deviation (variability) of the population (unless you did something drastic, such as getting rid of the extremes in the population) but, often you **can** control the sample sizes.

- Suppose that the standard deviations of both populations are the same, then the standard error is $\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}$. Suppose $n + m = 100$, how to choose the sample sizes to make this as small as possible?

Extreme examples are:

- If $n = 50$ and $m = 50$, $\sqrt{\frac{1}{50} + \frac{1}{50}} = 0.2$
- If $n = 99$ and $m = 1$, $\sqrt{\frac{1}{99} + \frac{1}{1}} = 1.01$.

- In the case that the standard deviations are the same, the standard error is smallest (and this the difference in the sample means is most reliable) when they both have sample sizes.
- Remember, in the case that the standard deviations are the **same**, having equal sample sizes leads to estimators with the smallest standard errors.

However, in the case that the standard deviations are **not the same**, equal sample sizes will not lead to the smallest standard deviation.

- The total variability/reliability on the estimator of the difference is accounted for in the standard error:

$$\sqrt{\underbrace{\frac{\sigma_1^2}{n}}_{\text{variability of } \bar{X}} + \underbrace{\frac{\sigma_2^2}{m}}_{\text{variability of } \bar{Y}}}$$

Example: Are two diets the same?

Two diets are being compared for effectiveness. 20 volunteers were randomly assigned to go on Diet 1 or Diet 2. There are 10 people in each group. After one month their weight loss (in kilos) was recorded. The data is given below.

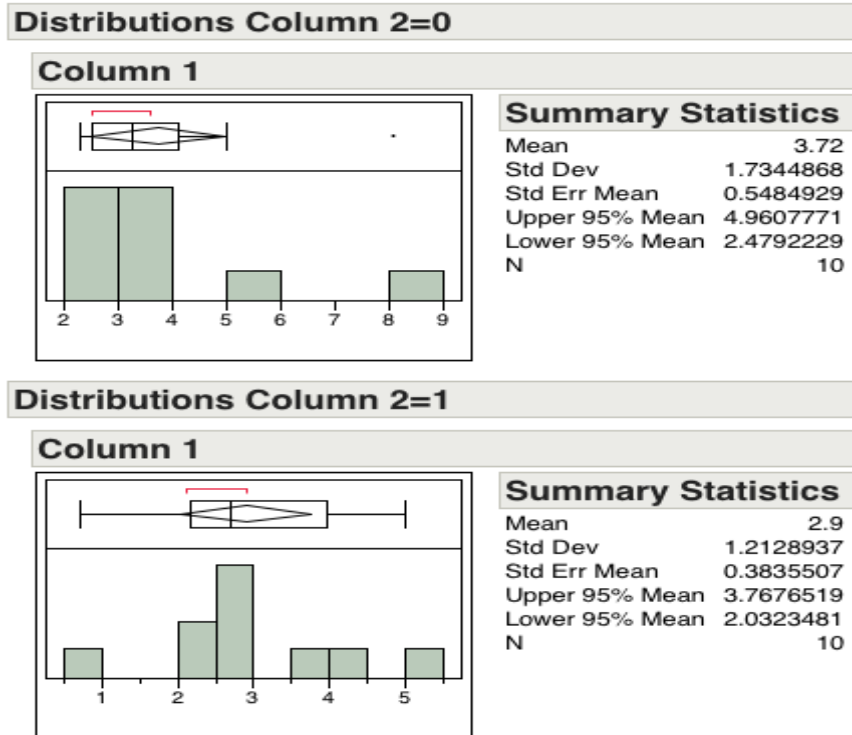
Diet I	2.9	2.7	3.9	2.7	2.1	2.6	2.2	4.2	5.0	0.7
Diet II	3.5	2.5	3.8	8.1	3.6	2.5	5.0	2.9	2.3	3

- Let μ_I be the mean weight loss of diet I and μ_{II} be the mean weight loss of diet II. Test the hypothesis that the diets are different.
- The data is

https://www.stat.tamu.edu/~suhasini/teaching651/diet_data.dat

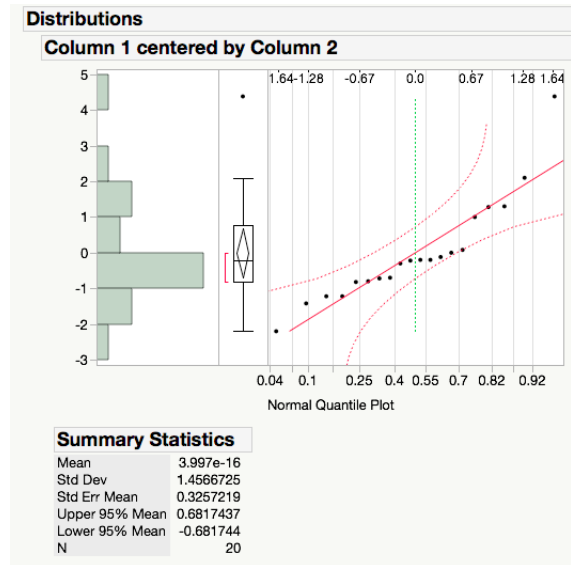
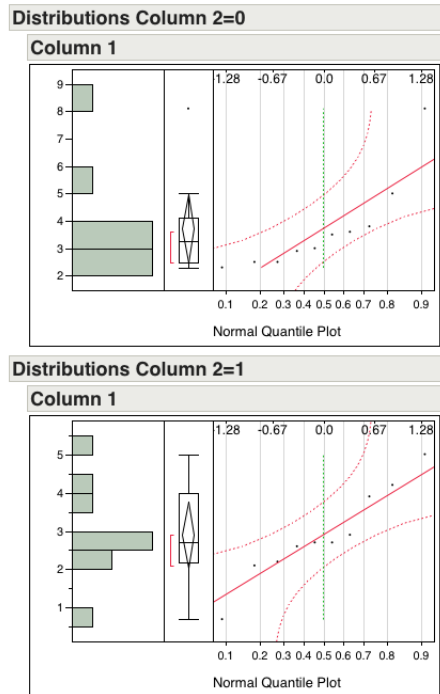
Plot of Diet data

The summary statistics and histogram is give below.



Checking the assumptions for the Diet Data

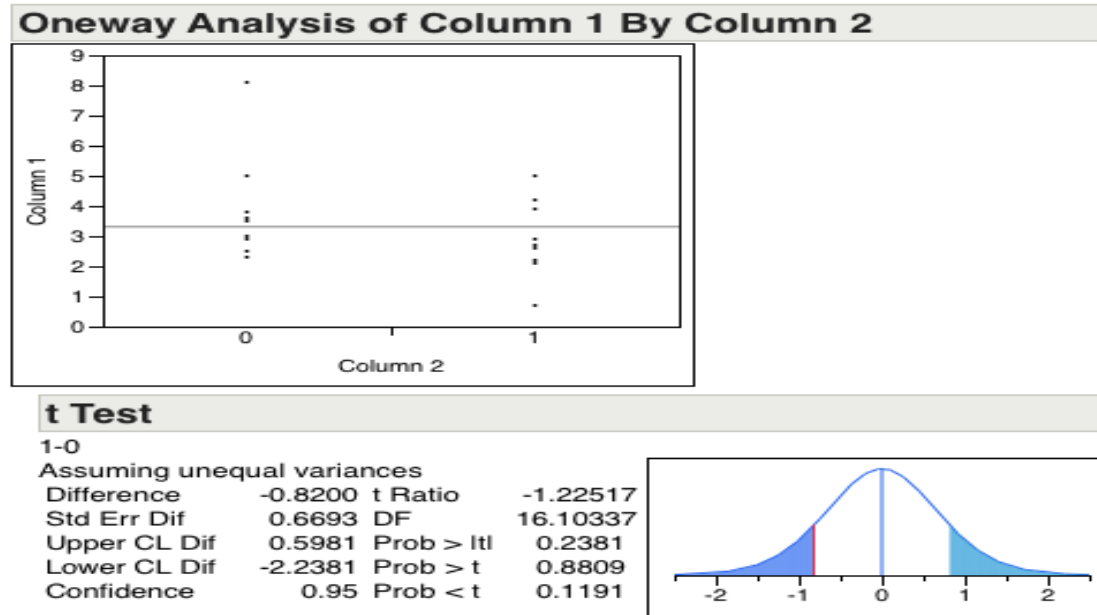
QQplot of each group and the residuals



- Since the people are not related and they were randomly allocated to one of the two diets it seems reasonable to assume that both data sets are independent.
- To check the normality assumption (for this small data set) we have made a QQplot. From the limited data that we have, it does not appear to be normal, this means that sample means (average weight loss for both diets will only be too close to normality). This will have an impact on the accuracy of the p-values that we obtain.
- However, there does not appear to be any huge outliers that may have significant impact on the outcome of the test (as we shall see later one outlier can have an dramatic impact on the test).

Lecture 19 (MWF) Independent sample t-test for testing equality of means in two populations

The JMP output



The hypothesis test and confidence interval

- Just for revision purposes the critical values for the t with 16.1df is

probability	0.15	0.10	0.05	0.025	0.01	0.005
t^*	1.07	1.33	1.74	2.12	2.58	2.91

- We test the hypothesis $H_0 : \mu_I - \mu_{II} = 0$ against $H_A : \mu_I - \mu_{II} \neq 0$, we do the test at the 5% level. We have the t-transform $t = (-0.82 - 0)/0.66 = -1.22$. This lies between -1.33 and -1.07. Therefore the area to the left of -1.22 is between 10 to 15%. Thus the p-value is between 20-30% (as seen by the JMP output which tells us it is 23%). As this is larger than 5% we cannot reject the null at the 5% level. Though we cannot accept the null, the data is consistent with there not being any difference between the two diets.

- The 95% confidence interval for the mean difference between the diets is $[-2.2, 0.59]$ pounds.
- What if we wanted to test if $H_0 : \mu_I - \mu_{II} \leq 0$ against $H_A : \mu_I - \mu_{II} > 0$ at the 5% level?
- What if we wanted to test if $H_0 : \mu_I - \mu_{II} \geq 0$ against $H_A : \mu_I - \mu_{II} < 0$ at the 5% level?

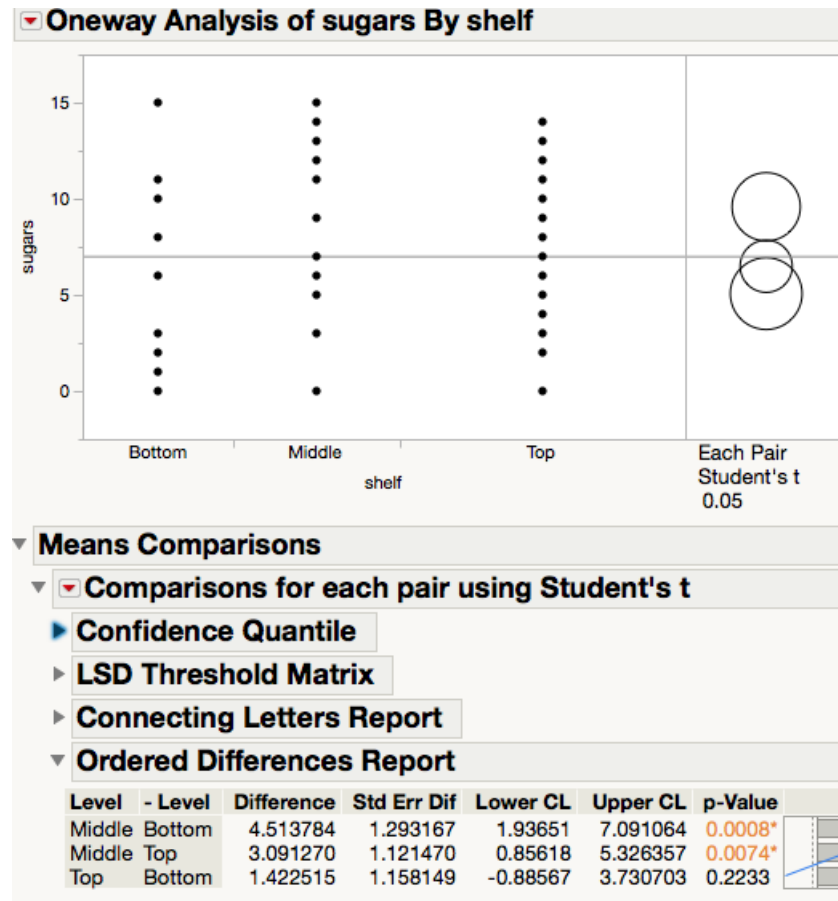
Example: Breakfast cereal and location on shelf

- We want to understand whether location of cereal packet has an influence on amount of sugar the cereal contains.
- There are usually top, middle and top shelves.
- My conjecture is that the middle shelf will contain the most sugar as this is the eye level of children which companies like to seduce with sugar. Based on this the hypotheses of interest are

$$H_0 : \mu_{Middle} - \mu_{Top} \leq 0 \text{ against } H_A : \mu_{Middle} - \mu_{Top} > 0.$$

- The data is here

https://www.stat.tamu.edu/~suhasini/teaching651/Cereal_Brands.csv

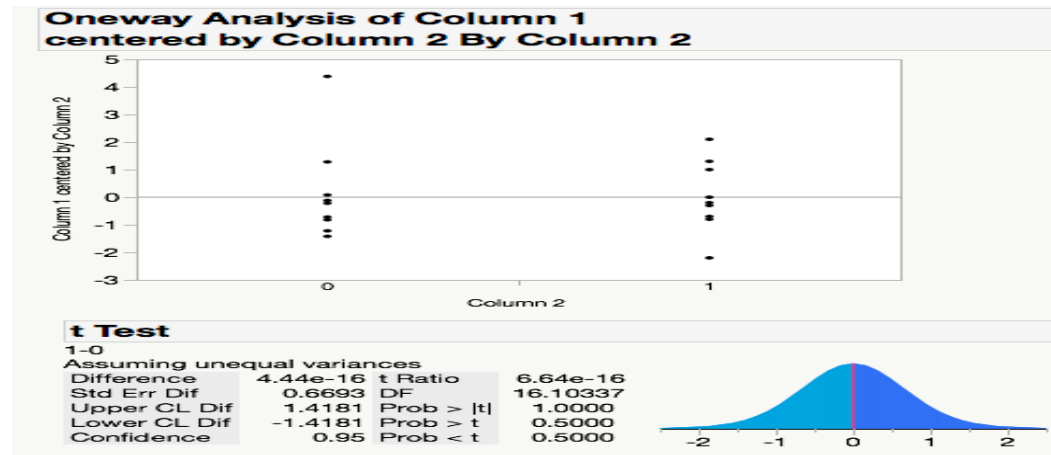


ality of means in two populations

Note The p-values that JMP give are for the two-sided tests. Therefore, you need to figure out the results for one-sided tests using this and the summary statistics.

Warning: Never apply a t-test to residuals

The residuals remove all mean information from the data. The difference in the sample means of the residuals will be zero - therefore the p-value of a 2-sided test will be 100% (and the p-value in a one-sided test will be 50%). To see this in action, here is the output of the independent sample t-test of the residuals of the diet data (considered above):



Naturally the test CANNOT detect a difference in the means, because there is no difference.