Data Analysis and Statistical Methods
Statistics 651

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

Lecture 23 (MWF) ANOVA: Testing equality of means of multiple populations

Suhasini Subba Rao

Testing for multiple groups

- Suppose we are interesting in the relationship between height of a child and their location in the world.

- We observe three samples from three populations:
  - 15 heights of 10 year old children from Country 1.
  - 15 heights of 10 year old children from Country 2.
  - 15 heights of 10 year old children from Country 3.

- Our aim is to investigate if country has an influence on the mean height of a child.

- Let $\mu_1=$ mean height of 10 year olds in Country 1. $\mu_2=$ mean height of 10 year olds in Country 2 and $\mu_3=$ mean height of 10 year olds in Country 3.

- Formally we want to test $H_0 : \mu_1 = \mu_2 = \mu_3.$ against H_A : The means are not all the same.

- How to do the test?

- One method would be to go through every combination and test individually

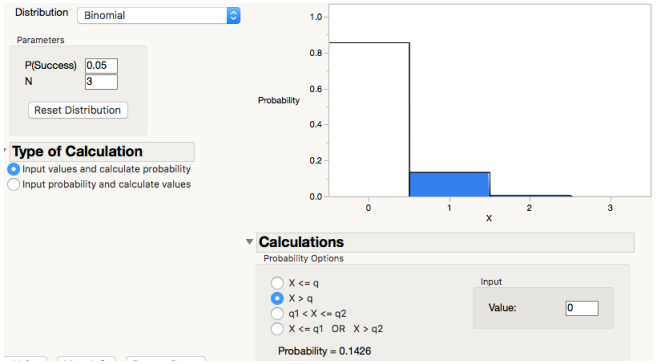
$$H_0 : \mu_1 = \mu_2 \text{ against } H_A : \mu_1 \neq \mu_2$$
$$H_0 : \mu_1 = \mu_3 \text{ against } H_A : \mu_1 \neq \mu_3$$
$$H_0 : \mu_2 = \mu_3 \text{ against } H_A : \mu_2 \neq \mu_3.$$

- There are problems associated with doing multiple tests, one of the main is false discovery or false positives. It can be shown that the probability

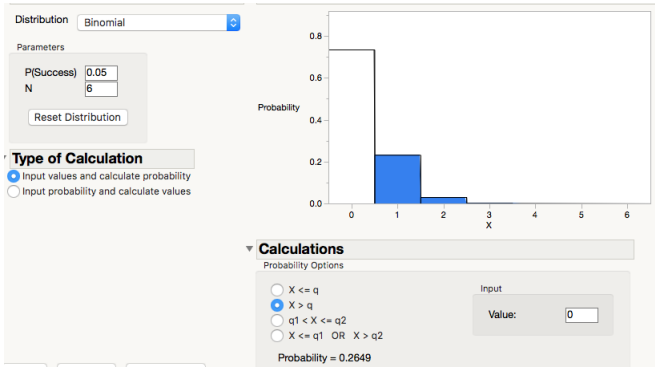
of rejecting the null for at least one of these tests (done at the 5% level) is $= (1 - (0.95)^3) \approx 14\%$ which is well above the 5% level.

- Advanced To understand why, recall that under the null that there is no difference in height. Each statistical test can be treated as Bernoulli trial, where there is with a 5% chance of a false rejection (call this a "success"). If 3 tests are done, on average there will be $0.05 \times 3 = 0.15$ rejections. In fact the number of successes follows a Binomial distribution Bin$(3, 0.05)$.

- Using this the probability of at least one false rejection is $P(S_3 \geq 1) = 0.14$; which we calculate using the Binomial calculator.



- If we now now test the equality of means between 4 groups, the number of pair-wise tests increase to 6. So probability of falsely rejecting the null follows a Bin$(6, 0.05)$. The chance of at least one false rejection increases to over 26%.



- Advanced The above is not exactly true, since even under the null the tests are not independent of each other. For example, the test for the first group against the second group depends on the test for the first group against the third (since the first group is common to both). This means the calculating for the chance of a false positive does not exactly hold.

The Bonferonni correction: controlling false positives

- If you have k groups, and you want to apply an independent two sample t-tests on all these groups there will be $k(k-1)/2$ different tests. If k is large and we do each test at the 5% level, there is a large chance the at least one test will be rejected when the **null is true**.

- To prevent this, and ensure the chance of rejecting any test less 5%, we need to do each test at the $5\%/m$-level where $m = k(k-1)/2$ is the number of tests that you are doing.

- Example: This means if there are 4 groups, then $k = 4 \times 3/2 = 6$ tests need to be done. Each at the $(5/6)\%$-significance level.

  Therefore, we can only reject the null if one of the p-values is less than $5/6\%$. This means that the p-values have to be extremely small in order reject the null.

- Using the Bonferonni correction is very conservative (can mean that is it is very difficult to reject the null when the alternative is true).

- In the case of testing equality of means we often use as ANOVA. This reduces the whole procedure to one simple test.

Example 1: M & Ms

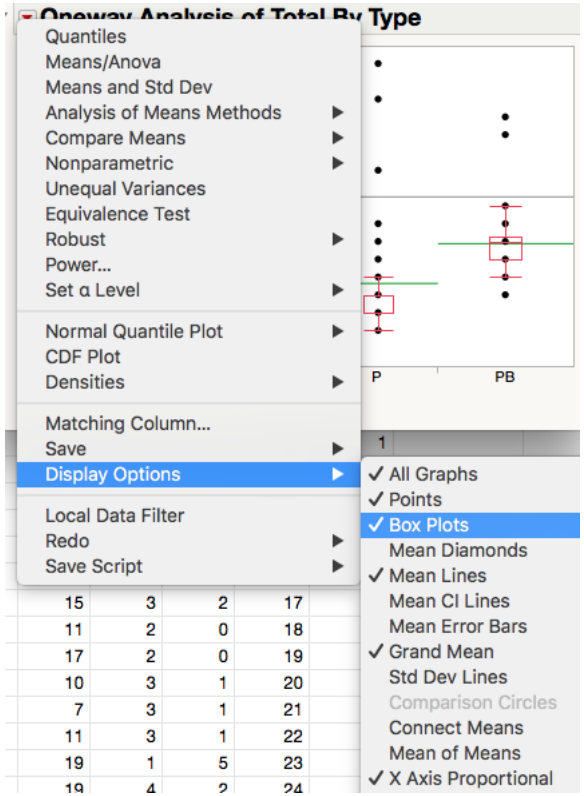
- Does the number of M&Ms in a bag depend on the type of M&M inside the bag?

- We can articulate this in statistical terms by asking whether the mean number of M&Ms vary according to the type, say Peanut Butter, Peanut or Milk Chocolate.

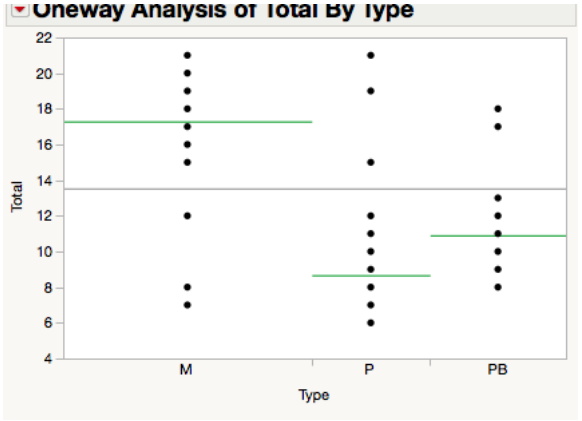
- We do not know what the means are, but we do have a sample of 170 bags. We want to use this sample to make inference about the population means. The summary statistics are:

	Milk	Peanut	Peanut Butter
sample mean	17.2	8.67	10.9
sample std deviation	2.87	3.13	1.83
sample size	84	40	46

Boxplots of M&M data: In JMP



Analyze > Fit Y by X > Place the options in the box (as in the two sample t-test test). Then in the new pop box, play with the display options. The green lines are the group sample means. the black line the global sample mean.



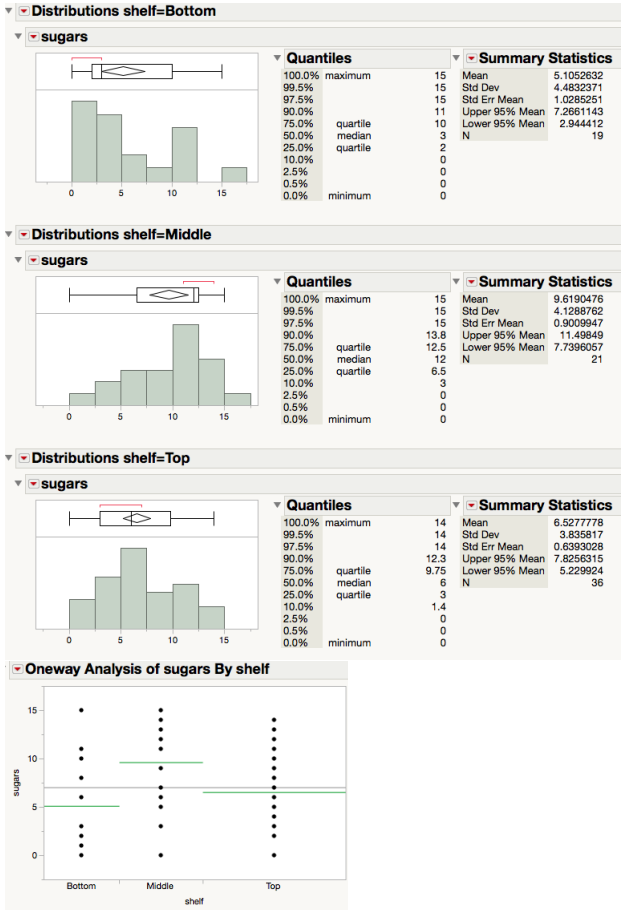
Example 2: Is there a relationship between location and sugar?

Let $\mu_B =$ denote the mean amount of cereal in the bottom shelf of a supermarket, and similar for μ_M and μ_T. We suspect that the sugar content of a cereal plays a role in its location.

Our aim is to test $H_0 : \mu_B = \mu_M = \mu_T$ against

H_A : At least one mean is different.

On the right is the summary and plots.



Example 3: Does different hormone treatments have an impact on the weight of 8 week old calves

- Calves are often put on various different hormone treatments when they are born. It is of interest to know whether the different treatments have different effects on their weights.

- Of course, we can't observe all calves and use a sample to make inference about the population.

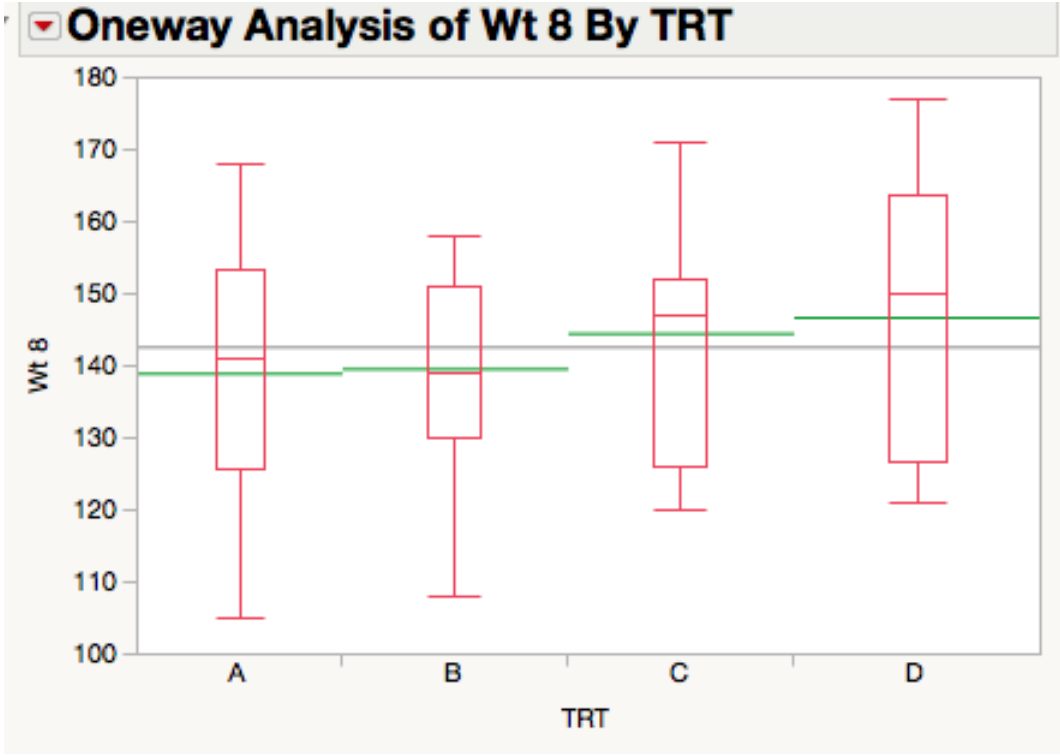
- The summary statistics are:

	A	B	C	D
sample mean	138.9	139.5	144.45	146.6
sample standard deviation	19.18	15.5	16.12	18.57
sample size	10	11	11	12

- There are differences in the sample mean, but is this statistically significant (can these difference be easily explained by sample variation)?

- We want to test $H_0 : \mu_1 = \mu_2 = \mu_3$ (the population means are the same) against H_A at least one of the means are different.

Boxplot of Calves and Treatments



It is difficult to judge. But the green lines are quite close.

Review of independent two-sample t-test

- The ANOVA is a generalisation of the independent two sample t-test (the two-sided version) to multiple populations.

- It is equivalent to the independent two sample t-test (the two-sided version) under the assumption the standard deviations in all the populations are the same (see Lecture 19, page 28).

- To understand the output, we recall how the t-test is evaluated. We test $H_0 : \mu_1 = \mu_2$ against $H_A : \mu_1 \neq \mu_2$. The statistic is

$$t = \frac{\bar{X} - \bar{Y}}{\text{standard error}} = \frac{\text{differences in sample means}}{\text{sample error of differences}}$$

  Under the null (and pooling the variance, see lecture 19) t has a t-distribution with $(n + m - 2)$df.

- We square the t-statistic

$$t^2 = F = \frac{(\bar{X} - \bar{Y})^2}{\text{standard error}^2} = \frac{\text{squared differences in sample means}}{\text{sample error of differences}^2}$$

  Under the null the distribution of t^2 has an F-distribution (we define this later) with $(1, n + m - 2)$df. The number of degrees of freedom has a double index.

- The main observation is that this is a ratio.
  - The numerator measures the distance between the sample means in each group.
  - The denominator is the squared standard error of the numerator.

- We reject the null when the standard error^2 is substantially smaller than the difference $(\bar{X} - \bar{Y})^2$.

- This corresponds to a large F-value and thus a small p-value.

- The degree of freedom for the numerator corresponds to the number of groups -1.

- The degree of freedom of the denominator corresponds to the total number of observations - number of groups.

- The ANOVA generalises this idea to the case that the number of groups is greater than two. One can do an ANOVA on two groups, the result is identical to an independent two-sample t-test using the method of pooled variance.

The principles of ANOVA: Artifical example

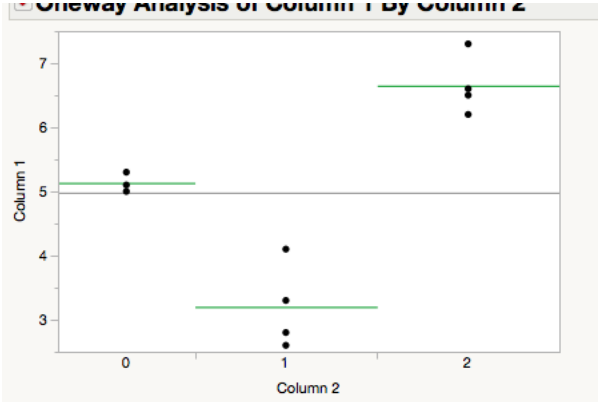
- We explain the following example very carefully. However, you will never have to do the calculations, you simply need to understand the software output.

- Consider the following artifical data example:

	Sample 1	Sample 2	Sample 3	Combined Sample
	4.1	5.1	6.6	
	3.3	5.0	6.2	
	2.6	5.3	7.3	
	2.8		6.5	
average	3.2	5.13	6.65	4.98
sample variance	0.45	0.023	0.22	0.26
sample s.d	0.67	0.15	0.46	0.50

The global mean is $\bar{X} = 4.98$

The main objective



Oneway Anova

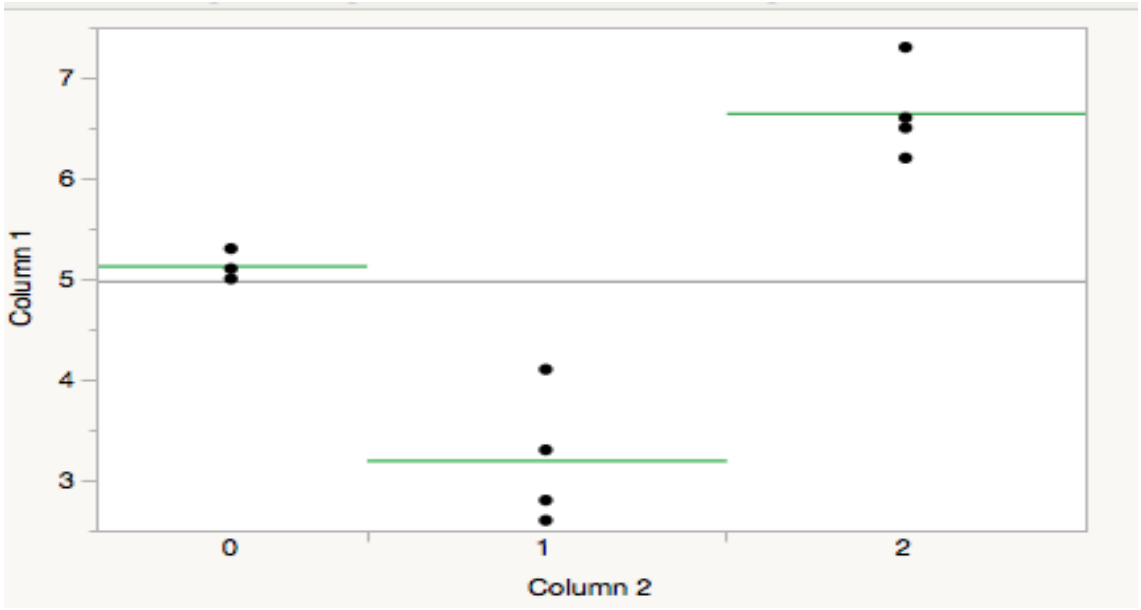
Summary of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Column 2	2	23.899697	11.9498	46.9388	<.0001*
Error	8	2.036667	0.2546		
C. Total	10	25.936364			

Our aim is to understand the above output. This output is in JMP, however, a similar output is given with other statistical software (such as R, SPSS etc).

A plot of the Artifical Data



The sample sizes are small but the data in each group appears to be quite separated. The black horizontal line is the mean of all the observations combined.

What the ANOVA does

- ANOVA works by measuring the deviation/difference between the global sample mean and the group sample means (called the SSB). As in every statistical test, it needs to be standardized by its standard error (called the SSW).

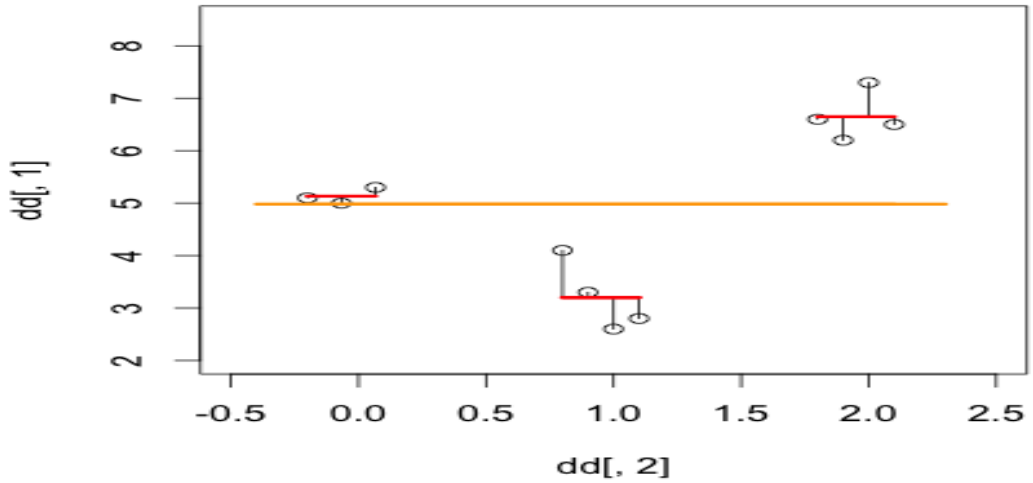
- Terminology often seen on output
  - The SST is the total sum of squares. It is the deviation squared of each observation from the global mean - basically the sample variance calculated as if the group means were is the same (multiplied by a $n-1$). Mathematically it is

$$SST = \sum_{j} \sum_{i=1} (X_{i,j} - \bar{X})^2 = SSB + SSW.$$

- The SSW is the sum of squares within each group.
  It is the sum of squares of the residuals from each observation to its group mean. Mathematically it is
  $$SSW = \sum_j \sum_{i=1} (X_{i,j} - \bar{X}_j)^2.$$
- The SSB is the sum of squares between each groups. It is the sum of the squares of the group means to global mean (multiplied by the number of observations in that group). It measures the difference between the group and global means. Mathematically it is
  $$SSB = \sum_j (\bar{X}_j - \bar{X})^2$$
  $\bar{X}_j$ denotes the group sample mean.

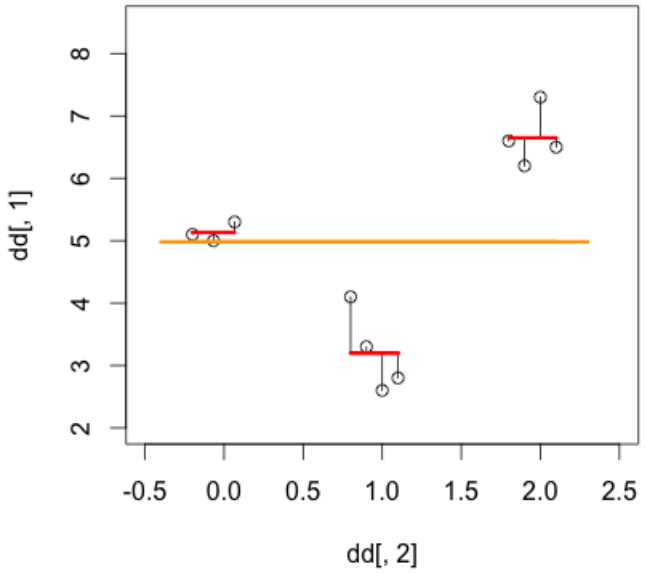
- The SSW and SSB are the main ingredients in an ANOVA.

The ANOVA plot for the Artifical example



The SSB is the difference between the red line and the yellow line. SSW is the difference between red line and the points around it. Observe that the differences 'within' each group (measured by the SSW) is a lot smaller than the differences between the groups (measure by the SSB).

Based on the above plot, we say that the within group variation is a lot lower than the between group variation.



- The average SSB (between groups) is

$$\frac{SSB}{3-1} = \frac{23.8}{2} = 11.9$$

- The average SSW (within groups) is

$$\frac{SSW}{11-3} = \frac{2.05}{8} = 0.266$$

- As in all tests we standardize the (squared) sample mean distance by its (squared) standard error. This is the F-statistic

$$F = \frac{SSB/2}{SSW/8} = \frac{11.9}{0.266}$$

(compare the t-statistic $t = (\text{estimator - mean})/\text{standard error}$).

- If the means of the populations are the same, then the F ratio will be 'small' (analogous to t being small).

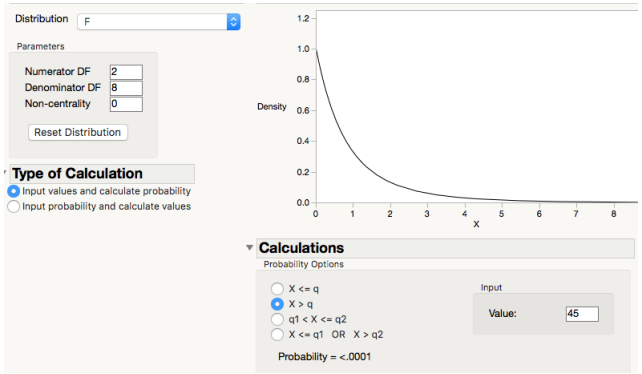
- If the population means are different, then F will be 'large'.

- The distribution of F under the null (that they share the same means) is an F-distribution with $(3-1, 11-3) = (2, 8)$-degrees of freedom.

The F-distribution and ANOVA

- Even though F is random, it will have certain characteristics if the null is true. If the null is true F is unlikely to be too large.

- The F-distribution with $(2, 8)$df looks like

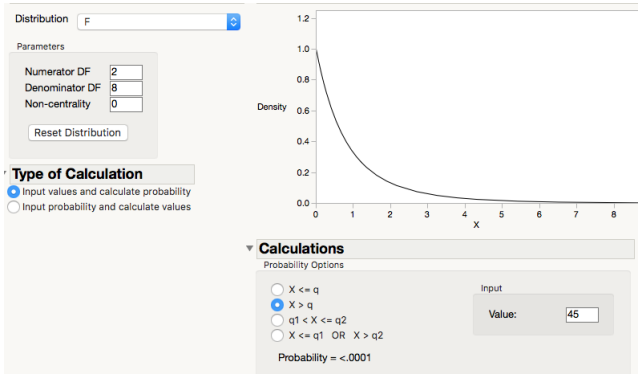


- In general, under the null the F distribution has degrees of freedom

(no. of groups -1), (no. of observations - no. groups)

- The p-value for the ANOVA is always the area to the right of F. We do not do one-sided tests with an ANOVA.

- For the artifical data set $F = 45$. The area to the right of it is



  is close to zero. The p-value for the test is close to zero. Which tells us that is extremely difficult to generate a data set such as this under the null at that the global means is the same.

- We reject the null, there is strong evidence to suggest that at least one of the population means in the group is different.

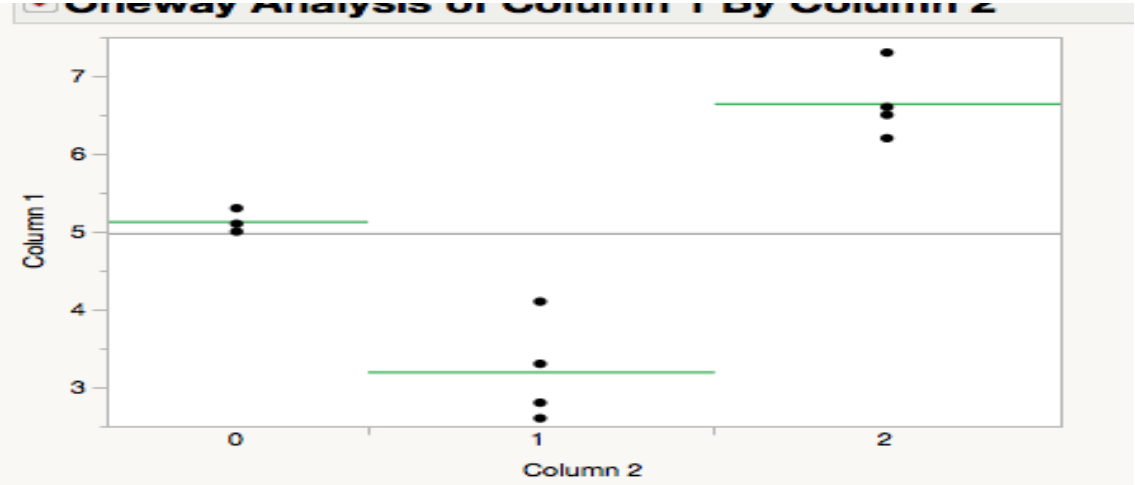
The ANOVA table: Artifical example

- When we divide the sum of squares by the degrees of freedom this is known as the mean square.

	Sum of Squares	df	Mean square	F	Sig.
Between Groups	SSB = 23.8	2	11.9	45	p-value = tiny.
Within Groups	SSW=2.05	8	0.266		
Total	25.85	10			

Sum of Squares total = SSW + SSB.

Corresponding ANOVA table in JMP



Oneway Anova

Summary of Fit

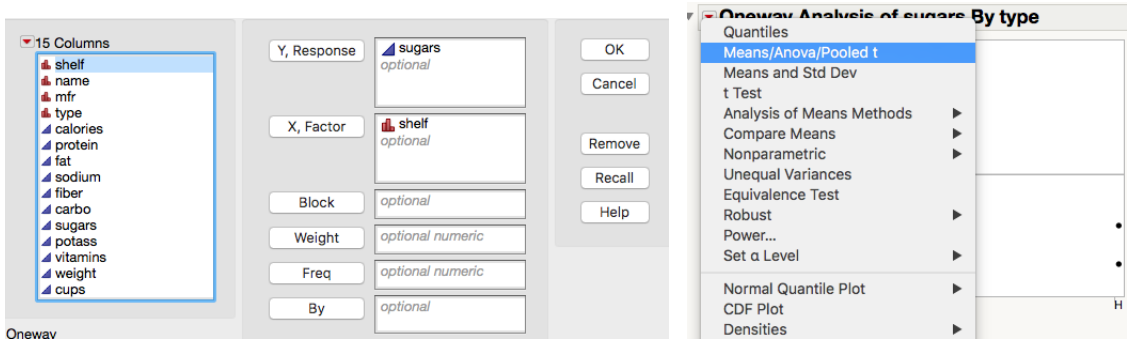
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Column 2	2	23.899697	11.9498	46.9388	<.0001*
Error	8	2.036667	0.2546		
C. Total	10	25.936364			

In an exam, I could easily cover parts of the table and ask you to fill in the rest.

Instructions for ANOVA in JMP

- Analyze > fit X by Y
- Place the box and fill in as below:



- This yields the ANOVA table on the next page.

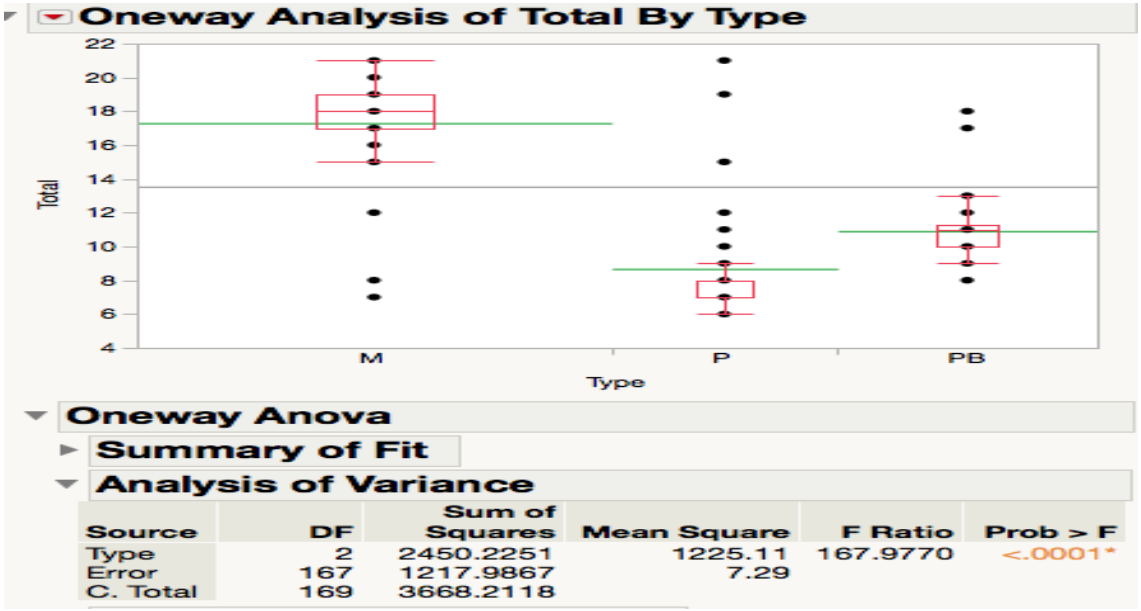
Example 1: ANOVA for M&Ms example

- The summary statistics is:

	Milk	Peanut	Peanut Butter
sample mean	17.2	8.67	10.9
sample standard deviation	2.87	3.13	1.83
sample size	84	40	46

- We want to test $H_0 : \mu_P = \mu_{PB} = \mu_M$ (the mean number of M&Ms for all types is the same) against H_A : at least one mean is different.

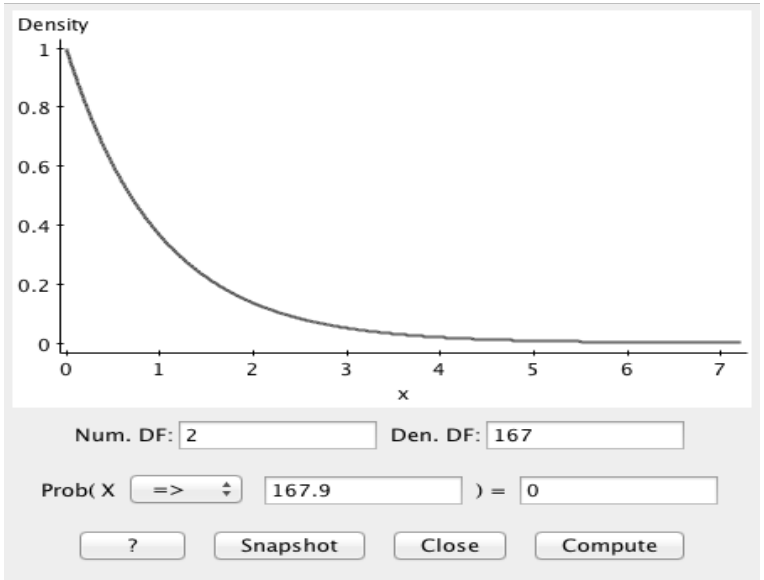
M&M ANOVA JMP output



Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Type	2	2450.2251	1225.11	167.9770	<.0001*
Error	167	1217.9867	7.29		
C. Total	169	3668.2118			

We see that the F-value is $F = 167$, which is extremely large. It is immediately clear that this is going to correspond to a very small p-value (very close to zero). Therefore, there is very strong evidence that the mean number of M&Ms in a bag varies according to type.

What the distribution looks like under the null:

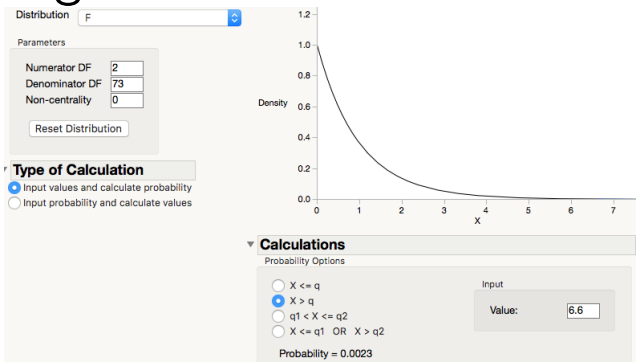


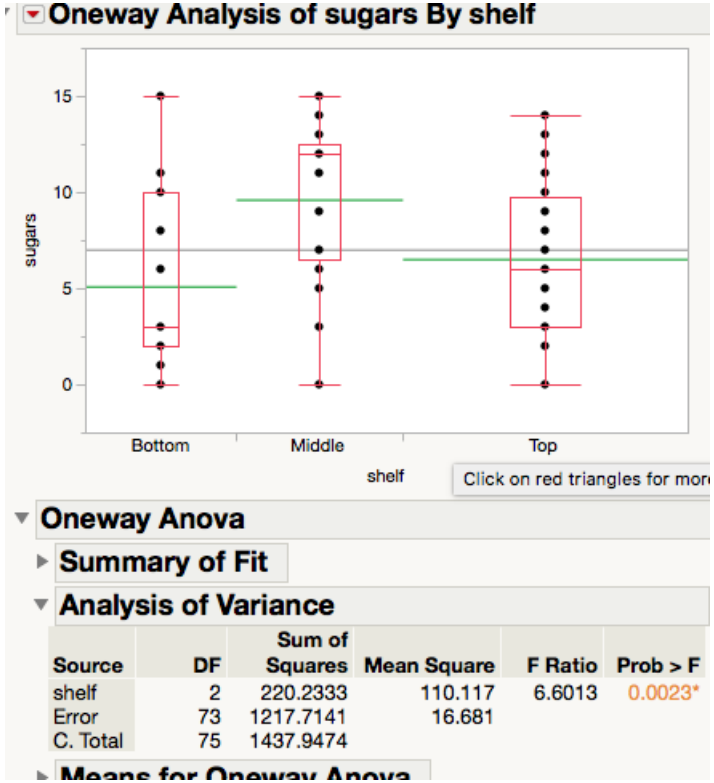
We see that $F = 167.9$ is pushed so much to the right, that the p-value will be tiny, making it the differences between the sample means extremely significant. There is substantial evidence to suggest that at least one of the means in an M&M bag is different.

Example 2: The sugar/location cereal example

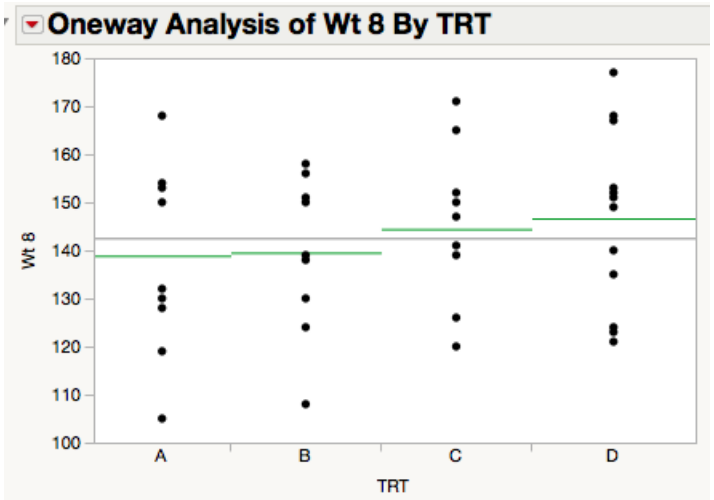
The F-value is 6.6 and the corresponding p-value is 0.23%. Since $0.23\% < 5\%$ we reject the null at the 5% level.

There is evidence to suggest there is a relationship between location and sugar levels.





Example 3: The calf example in JMP

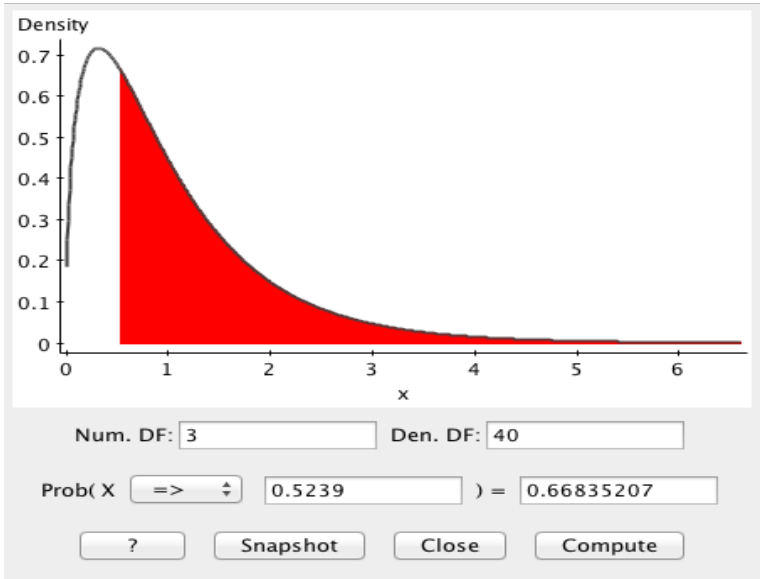


Oneway Anova

Summary of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
TRT	3	475.774	158.591	0.5239	0.6684
Error	40	12109.021	302.726		
C. Total	43	12584.795			



The ANOVA gives the $F = 0.523$, which we see corresponds to a p-value of 66%. Clearly, there is no evidence in the data, to suggest that the weights of calves vary according to treatment.

The Assumptions of an ANOVA

- To do an ANOVA we need that that the sample means are close normal. This ensures the p-value in the test is correct.

  If the sample sizes in each group are sufficient large, this does not matter since the CLT will ensure that the sample means are close to normal.

  We explain how to check this below.

- All the observations are completely independent of each other (just like the independent sample t-test).

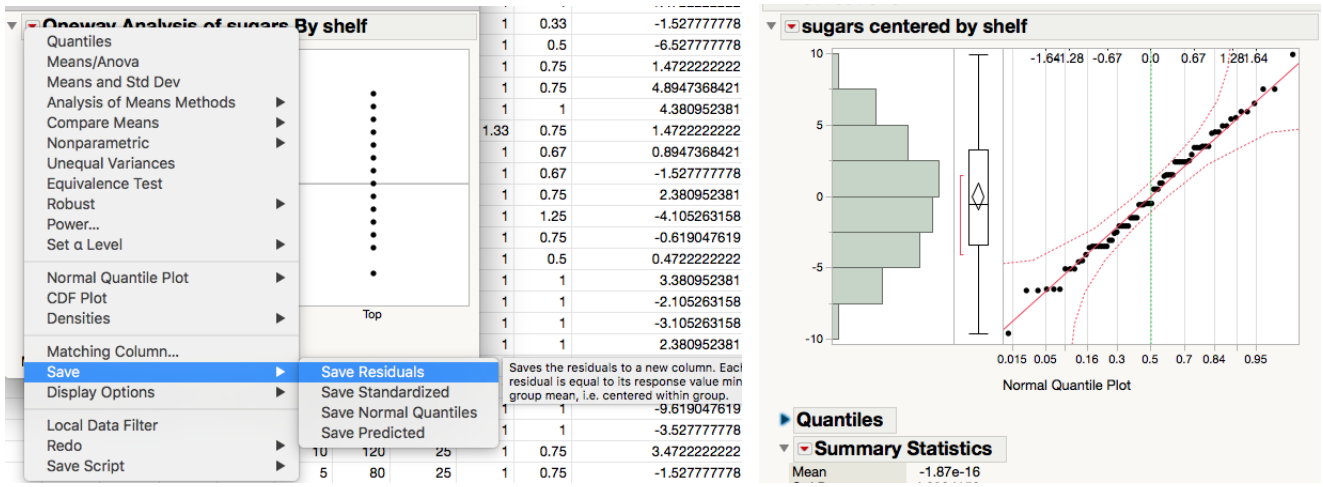
  Usually this is quite difficult to check. Typically, this can be determine done by considering how the data was collected.

- The standard deviation within each of the groups is roughly the same.

The Levene test can be used to determine if the variance between two populations is the same. It is more difficult for multiple populations. This assumptions is quite robust to differences.

Checking the Normality assumption

- To check for normality, we do exactly what was done for independent two-sample t-test. We extract the residuals and make a QQplot of the residuals (which will be a new column in the spread sheet).



- Observe that the mean of the residuals is zero (as is always the case) and that the residuals for the sugar cereal example is close to normal. Never apply an ANOVA to residuals!

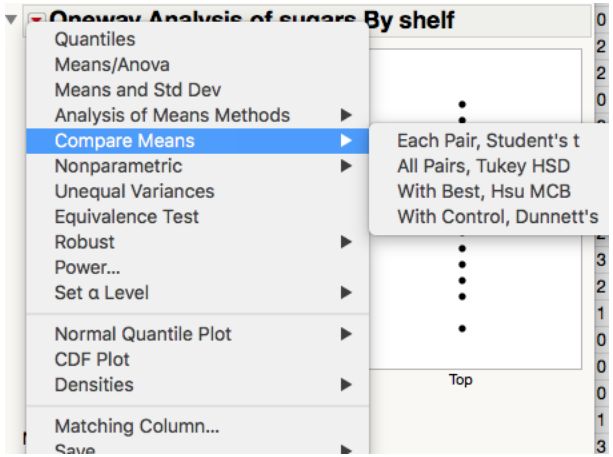
Post-Hoc tests

- If one rejects the null hypothesis in ANOVA, it is of interest to find which groups have different means.

- Post-Hoc tests apply a variation of the independent two-sample t-test to all the different pairings of the groups.

- There are several different Post-Hoc tests that one can apply, all of which control the Type I error. This means trying to globally ensure that there is close to 95% chance that no test will be falsely rejected when the null is true. Or equivalently there is only a 5% chance that one of the tests is falsely rejected when the null is true.

- JMP offers three different post-hoc tests:



Each Pair, Student's t does each test at the 5% level (and can lead to an over rejection of the null).
The three remaining test control the Type I (falsely rejective the null or a false positive) error in different ways.

Below we focus on All Pairs, Tukey HSD. Focussing on the cereal example.

Tukey's Honestly significant difference (HSD)

- Tukey's test controls the Type - I error by focussing on the distribution of the maximum and minimum averages

$$t_R = \frac{\bar{X}_{max} - \bar{X}_{\min}}{(s/\sqrt{n})}.$$

- t_R has a t-range distribution and there are tables for this (you do not have to worry about it). It is derived under the assumption that all the means are the same (the null hypothesis) and that the sample sizes for each group are the same (this does not matter so much).

- This new distribution is used to construct confidence intervals and test. The chance of a false positive for the combined set of tests is less than 5%.

- The confidence intervals given in the Tukey's test are group-wise (often called family-wise as it is over all pairwise combinations) confidence intervals. We can say with 95% confidence all the mean differences lie in these intervals, which is more informative than saying pair-wise we have 95% confidence in the intervals.

- The p-values are derived using the t-range distribution. For any p-value less than then 5% we reject the null and say there is a difference between the two groups (at the 5% level). By using the t-range distribution we are controlling the proportion of type-I errors over all the tests to 5%.

What we mean by family-wise confidence

Tukey's HSD in JMP

HSD Threshold Matrix

Abs(Dif)-HSD

	Middle	Top	Bottom
Middle	-3.0155	0.4082	1.4200
Top	0.4082	-2.3031	-1.3483
Bottom	1.4200	-1.3483	-3.1702

Positive values show pairs of means that are significantly different.

Connecting Letters Report

Level		Mean
Middle	A	9.6190476
Top	B	6.5277778
Bottom	B	5.1052632

Levels not connected by same letter are significantly different.

Ordered Differences Report

Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL	p-Value
Middle	Bottom	4.513784	1.293167	1.41995	7.607618	0.0023*
Middle	Top	3.091270	1.121470	0.40821	5.774327	0.0199*
Top	Bottom	1.422515	1.158149	-1.34830	4.193324	0.4405

The p-values correspond to the two-sided tests and are derived using the t-range distribution.

Comparing the p-values, we see that the mean amount of sugar in the Middle Shelf is significantly different to the mean amount of sugar in both the bottom or top shelves (both p-values less than 5% levels).

However, the differences in the mean amount of sugar in the bottom and top shelves are not significantly different. We conclude that the sugar content in the middle shelf differs from the sugar content in the top and bottom shelves.

Nonparametric Tests: Kruskal-Wallis

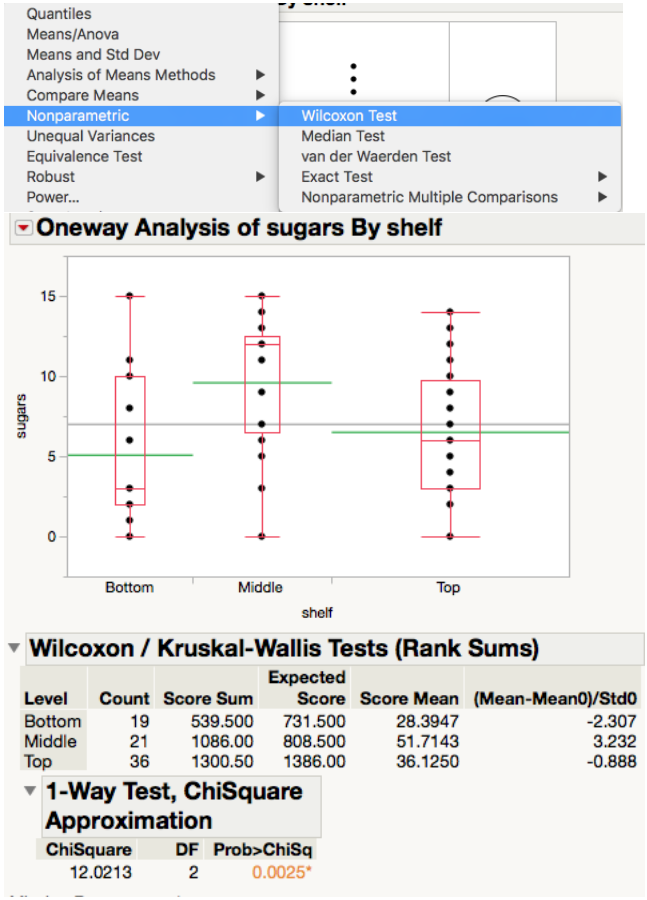
- Suppose a QQplot of the residuals shows a large departure from normality and/or there are outliers in some of the group. Then the regular ANOVA may give spurious results.

- ANOVA is quite robust to departures from normality. But when there are large outliers we may require a more robust test.

- The Kruskal-Wallis is a nonparametric procedure that one can use in the non-normal cases.

- It is a generalisation of the Wilcoxon rank sum test, that is a rank based test that uses the ANOVA machinery.

Kruskal Wallis in JMP



We need to focus on the 1-Way Test, ChiSquare Approximation.

ChiSquare=12.3 is the equivalent of the F-value (observe that it is "large"). The DF $= 2$ (since it is no. of groups - 1).

The corresponding p-value is 0.0025 (remember large t, F or Chi-value correspond to small p-value), which tells us that at there is strong evidence to suggest that at least one population mean is different to the others.

ANOVA: The Formal definition that you may see in a textbook

- We have k populations with means $\mu_1, \ldots, \mu_k$. $X_{i,j}$ denotes an observations from the ith population. In other words $X_{i,j}$ is the jth draw from the ith population:

  $X_{1,1}$ is the first draw from population one.

  $X_{1,2}$ is the second draw from population one.

  $X_{2,1}$ is the first draw from population two.

  For example, $X_{i,j}$ the height of the jth randomly selected child in country i.

We assume that $X_{i,j}$ satisfies the following model:

$$\begin{aligned} X_{i,j} &= \mu_i + \underbrace{(X_{i,j} - \mu_i)}_{\varepsilon_{i,j}} \\ &= \mu_i + \varepsilon_{i,j} = \mu + \alpha_i + \varepsilon_{i,j} \end{aligned}$$

where $\sum_k \alpha_k = 0$ and μ is the common mean.

We are testing $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ against the alternative H_A : the means are not all the same, where sample j is of size n_j. Using the notation in the above model this is the same as $H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_k$ against alternative H_A : at least one of α_js are not zero.

- From each population i we have a sample $X_{1,i}, \ldots, X_{n_i,i}$, each sample is of size n_i.

- The total sum of all observations is $N = n_1 + \ldots + n_k$.

- The data looks like

	Sample 1	Sample 2	. . .	Sample k	Global
	$X_{1,1}$ $X_{1,2}$ $\vdots$ X_{1,n_1}	$X_{2,1}$ $X_{2,2}$ $\vdots$ X_{2,n_2}	. . . .	$X_{k,1}$ $X_{k,2}$ $\vdots$ X_{k,n_k}	
mean	$\bar{X}_1$	$\bar{X}_2$	. . .	$\bar{X}_{k\cdot}$	$\bar{X}$

- The group average $\bar{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{i,j}$ and global average taken over all observations is $\bar{X} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{i,j}$.

Calculating the SSW and SSB

- We calculate the sum of squares between samples

$$SSB = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2.$$

- We calculate the sum of squares within sample

$$SSW = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{i,j} - \bar{X}_i)^2 = \sum_{i=1}^{k}(n_i - 1)s_i^2,$$

where $s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(X_{i,j} - \bar{X}_{i\cdot})^2$ which is the sample variance of sample i.

- Let $N = n_1 + \ldots + n_k$.

- We use as the test statistic

$$\frac{SSB/(k-1)}{SSW/(n-k)}.$$

- Under the null $\frac{SSB/k-1}{SSW/(N-k)} \sim F_{k-1,N-k}$.

	Sum of Squares	df	Mean square	F	Sig.
Between Groups	SSB	$k-1$	$SSB/(k-1)$	$\frac{SSB/(k-1)}{SSW/(N-k)}$	p-value.
Within Groups	SSW	$N-k$	$SSW/(N-k)$		
Total	SST	$N-1$			