## Practical and Statistical Significance

*Statistical significance* (*P*-value) indicates the extent to which the null hypothesis is contradicted by the data.

*Practical* (or *biological or whatever*) *significance* is different, and describes the practical importance of the effect in question.

A study may suggest a *statistically significant* increase in plant growth of 1% due to a treatment, but this increase may not justify the expense of the treatment. Hence, the finding is *statistically significant* but not *practically significant*.

*Statistical significance* is really only a matter of sample size. Even the slightest difference in population means will be found to be *statistically significant* given enough samples.

In contrast, even if there truly is a *practically significant* difference between population means, small sample sizes might fail to indicate the existence of a *statistically significant* difference.

Three points to consider:

1. *P*-values are sample size dependent.
2. A result with a $P = 0.08$ can be more important scientifically than one with $P = 0.001$.
3. Hypothesis tests rarely convey the full meaning of the results; they must be accompanied by confidence intervals to indicate the range of likely effects and to assess *practical significance*.

## Comparing Several Samples

### Introduction

Issues and tools for analysis of >2 independent samples are similar to comparing 2 samples. More questions are possible.

Initial question asked in this context is whether means of all samples are equal, i.e., $H_o$: $\mu_1 = \mu_2 = \mu_3 = \mu_4$.

**Analysis of Variance** (ANOVA) is an important tool for analysis of >2 samples and is a straightforward extension of the 2-sample *t*-test.

Can use ANOVA to perform each of the *t*-tests studied; i.e., an ANOVA with 1 or 2 groups is exactly the same as a *t*-test.

We will develop ANOVA as a type of **General Linear Model** and move towards a more general approach to data analysis.

## Comparing Any Two of Several Means

When subjects are divided into *distinct* experimental or observational categories, the study is a *one-way classification* problem.

A typical analysis in this context involves

1. graphical exploration
2. consider transformations
3. initial screening to evaluate differences between all groups
4. inferential techniques to address questions of interest

Besides the question of equal group means ($H_o$: $\mu_1 = \mu_2 = \mu_3 = \mu_4$), we can assess pairwise differences between means, such as:

"Does the mean of group 1 differ from the mean of group 3?" (i.e., $H_o$: $\mu_3 = \mu_1$ or $H_o$: $\mu_3 - \mu_1 = 0$).

When the number of comparisons is large, we must consider the consequences of simultaneous inferences.

**Ideal Model for Several-sample Comparisons**

An Extension of the normal model for 2-sample comparisons:

1. populations have normal distributions,
2. population standard deviations (or variances) are all equal,
3. observations *within each sample* are independent,
4. observations in any one sample are independent of those *in other samples*.

***Notation***

Population mean: $\mu$ with a subscript indicating its group (e.g., $\mu_2$)
Standard deviation (assumed "common" to all pop'ns): $\sigma$
No. treatments, populations, or groups sampled: $I$ (e.g., $I = 4$)
No. observations in the $i^{th}$ sample: $n_i$ (e.g., $n_2 = 5$)
Total no. observations from all groups: $n$ $(= n_1 + n_2 + \cdots + n_I)$

We estimate $I + 1$ parameters in the ideal model; one for each of the $I$ group means and one for the pooled standard deviation $\sigma$.

***Pooled Estimate of the Standard Deviation***

The mean for the $i$th population, $\mu_i$, is estimated with the average of the $i^{th}$ sample.

Variance ($\sigma^2$) is estimated separately for each of the $I$ samples ($s_i^2$). We pool these variance estimates to get an average weighted by their degrees of freedom ($s_p^2$):

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_I - 1)s_I^2}{(n_1 - 1) + (n_2 - 1) + \cdots + (n_I - 1)} = \frac{SS_1 + SS_2 + \cdots + SS_I}{df_1 + df_2 + \cdots + df_I}$$

If variances of all groups can be assumed equal, $\sigma^2$ is best estimated with $s_p^2$, the pooled estimate from all groups.

***t-Tests and Confidence Intervals for Mean Differences***

Use the pooled estimate of variance to calculate the standard error of the difference between groups which is used to calculate <u>*t*-statistics</u> to compare means between any 2 groups and <u>confidence intervals</u> for the difference between any 2 groups.

*Example:*

Mice-diets (Ch. 5) with 6 groups in a one-way classification.

Compare means from group 3 and group 2 ($\mu_3 - \mu_2$).

Estimate SE of $\bar{y}_3 - \bar{y}_2$:

$s_{\bar{y}_3 - \bar{y}_2} = SE(\bar{y}_3 - \bar{y}_2) = s_p\sqrt{\frac{1}{n_3} + \frac{1}{n_2}}$, where $s_p$ is the pooled estimated standard deviation from <u>all 6 groups</u>, with $(n - I)$ *df*.

Theory and computations for confidence intervals and hypothesis tests are identical the two-independent sample problem.

$$t = (\bar{y}_3 - \bar{y}_2) / \text{SE}(\bar{y}_3 - \bar{y}_2)$$

$$95\%\ \text{CI} = (\bar{y}_3 - \bar{y}_2) \pm t_{\text{df}(1 - \alpha/2)}\ \text{SE}(\bar{y}_3 - \bar{y}_2)$$

## ANOVA: Terminology and Bookkeeping

The term “variance” in ANOVA should not be misleading—these are question about means.

ANOVA approach assess differences in means by comparing the amount of variability in the data explained by different sources.

ANOVA models reflect closely the way in which data were collected (i.e., the sampling or experimental design).

Illustration: experiment assessing the effects of four different feeds on the body mass of pigs.

Randomly allocate 4-5 pigs to each treatment group and raise them on this type of feed. The resulting data look like this:

| Feed 1 | Feed 2 | Feed 3 | Feed 4 |
|---|---|---|---|
| 60.8 | 68.7 | 102.6 | 87.9 |
| 57.0 | 67.7 | 102.1 | 84.2 |
| 65.0 | 74.0 | 100.2 | 83.1 |
| 58.6 | 66.3 | 96.5 | 85.7 |
| 61.7 | 69.8 | | 90.3 |

*The following terms assume a manipulative experiment, though they usually apply to observational studies too.*

**Experimental unit** — the smallest independent unit of an experiment to which a treatment can be (randomly) assigned; here, each pig.

**Experimental design** —the way in which treatments are assigned to experimental units. The example is a Completely Randomized Design (CRD).

**Treatment** — manipulations to which experimental units are subjected; here, the treatment is feed-type. An important type of treatment is a *control*.

**Factor** — a group of related treatments examined in an experiment; this example is for a single-factor (one-way) classification (design), as feed-type is the only factor examined.

**Levels** — the number of different treatments for a particular factor; here, there are four levels of feed-type.

**Replicate** — smallest set of experimental units that receive the complete treatment set.

**Experimental error** — differences in responses from experimental units receiving the same treatment.

**Response** — variable measured to assess the effects of experimental treatments; here, the body mass of pigs studied.

For a one-way (single-factor) ANOVA, track the response for every experimental unit using two subscripts, $y_{ij}$:

- the first subscript, *i*, identifies the treatment group
- the second subscript, *j*, identifies each experimental unit (replicate) within a treatment.

For example, $y_{23}$ identifies the response for the 3rd subject in the 2nd treatment group, where $y_{23} = 74.0$.

The average for each treatment *i* is identified as $\bar{y}_i$ (or $\bar{y}_{i.}$).

The average for all observations from all treatments is the **grand mean** and is identified as $\bar{y}$ or $\bar{y}_{..}$ and is calculated as:

Sample sizes for each treatment *I* are identified as $n_i$; sample size for the entire experiment is *n*.

## Partitioning Sum of Squares

**Total Sums of Squares (SS)** estimates the total amount of variation in a data set and can be partitioned into component "sources."

We then examine how these different sources interrelate.

In the simplest case of a single sample, SS = $\Sigma(y_i - \bar{y})^2$.

- **Total SS** represents variability among all data: $\sum_{i=1}^{I}\sum_{j=1}^{n_j}(y_{ij} - \bar{y}_{..})^2$

i.e., the sum of the squared differences between every observation and the grand mean.

In a one-way classification, Total SS is partitioned two sources:

- variability due to treatments (*Treatment SS*)
- variability due to error (*Residual* or *Error SS*).

A *residual* is an observed value minus its estimated mean.

No matter how you partition them, *Total SS* for a given data set are always the same.

- **Total *df*** is the sum of all $n_i$ minus 1, or $n - 1$.

- **Treatment SS** (or *among-groups SS*) is the variability among averages from different treatments: $\sum_{i=1}^{I} n_i(\bar{y}_i - \bar{y})^2$

- **Treatment *df*** (or *among-group df*) is the number of treatment groups minus 1, or $I - 1$.

- **Residual SS** (or *error SS* or *within-group SS*) is variability among experimental units receiving the same treatment: $\sum_{i=1}^{I}\sum_{j=1}^{n_j}(y_{ij} - \bar{y}_i)^2$

- **Residual *df*** (or *error df* or *within-group df)* is: $\sum_{i=1}^{I} n_i - 1 = n - I$

*SS* and their *df* are additive:

Total *SS* = Treatment *SS* + Residual *SS*
Total *df* = Treatment *df* + Residual *df*

After calculating *Total SS* and *Treatment SS*, *Residual SS* can be obtained by subtraction:

Residual SS = Total SS – Treatment SS
Residual *df* = Total *df* – Treatment *df*

The deviation between each observation and the grand mean is the sum of:

1. the deviation of that observation from its group average
2. the deviation of that observation's group average from the grand mean:

$$(y_{ij} - \bar{y}_{..}) = (y_{ij} - \bar{y}_{I.}) + (\bar{y}_{I.} - \bar{y}_{..})$$

In the 2-group case (*t*-test), if we assumed $\sigma^2_1 = \sigma^2_2$, we estimated $\sigma^2$ with the pooled sample variance, $s_p^2$:

$$\frac{\sum_{i=1}^{2}\sum_{j=1}^{n_j}(y_{ij} - \bar{y}_i)^2}{\sum_{i=1}^{2}(n_i - 1)}$$

This is equivalent to $(SS_1 + SS_2) / (df_1 + df_2)$, which is the *Residual SS* divided by the *Residual df*.

Assume variances from all groups are equal ($\sigma^2_1 = \sigma^2_2 = \sigma^2_3 = \sigma^2_4$), and estimate $\sigma^2$ with $s_p^2$ by dividing *Residual SS* by *Residual df,* which is an estimate of error (residual) variance:

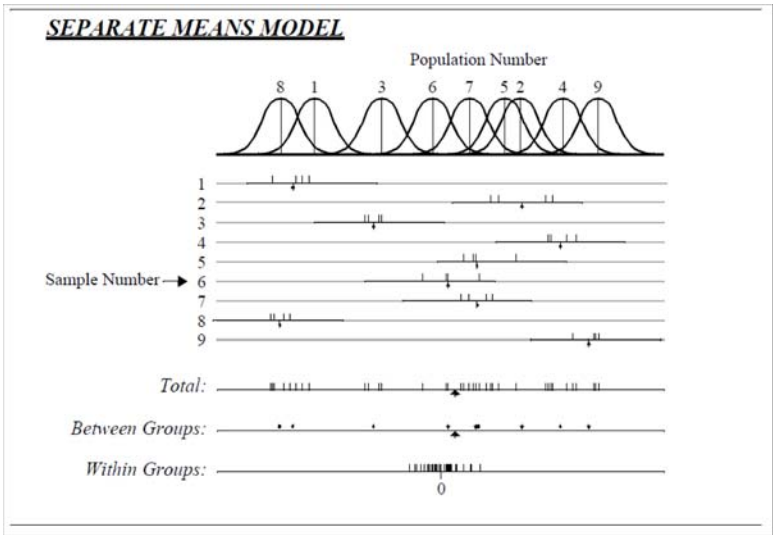
$$\textit{residual} \text{ or } \textit{error SS} = \sum_{i=1}^{I}\sum_{j=1}^{n_j}(y_{ij} - \bar{y}_i)^2$$

$$\textit{residual} \text{ or } \textit{error df} = \sum_{i=1}^{I} n_i - 1$$

The estimate of variance (*Residual SS* / *Residual df)* is called the **Residual Mean Square** (or Mean Square Error, MSE).

Dividing any *SS* by its respective *df* estimates a component of variance or its squared deviation from the mean, often called simply a **Mean Square**.

For example, to estimate variance attributable to treatment, divide *Treatment SS* by *Treatment df*, which is the **Treatment Mean Square**.



## One-way Analysis of Variance *F*-test

*Initial question:* Are there differences between *any* of the group means? Answered with ANOVA *F*-test.

Significance tests in ANOVA (*F*-tests) function by comparing ratios of different variance components (i.e., mean squares).

### *F*-Distributions

If all means are equal, the *F*-statistic has a sampling distribution of an *F*-distribution.

*F* depends on <u>two</u> parameters, the *numerator degrees of freedom* and the *denominator degrees of freedom*.

When reporting an *F*-statistic, report both numerator and denominator *df*'s. For example, $F_{2,21} = 4.54$.

For each possible pair of *df's*, there is a different *F*-distribution.

• *F* values ranging from 0.5 to 3.0 typically do not indicate strong evidence again the null hypothesis of equal means.

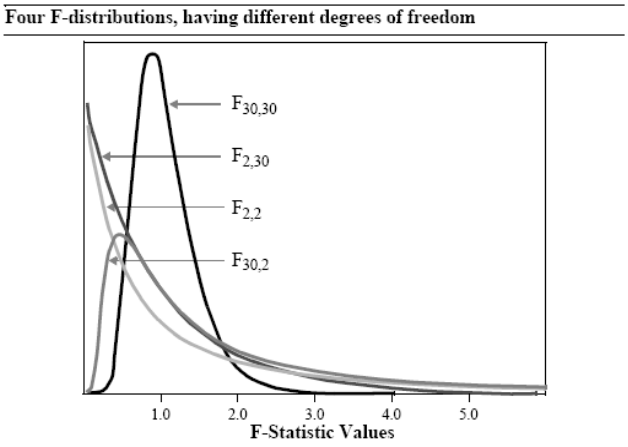
• *F* values >4.0 are strong evidence again the null.

### *F*-Tests

To generate an *F* test statistic for a treatment effect, calculate the ratio of *Treatment MS*/*Residual MS*.

For our example:

$H_o$: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
$H_a$: mean body mass of at least one treatment differs from the others.




Determine relevant averages, SS, *df,* and MS:

| | Feed 1 | Feed 2 | Feed 3 | Feed 4 | |
|---|---|---|---|---|---|
| | 60.8 | 68.7 | 102.6 | 87.9 | |
| | 57.0 | 67.7 | 102.1 | 84.2 | |
| | 65.0 | 74.0 | 100.2 | 83.1 | |
| | 58.6 | 66.3 | 96.5 | 85.7 | |
| | 61.7 | 69.8 | | 90.3 | |
| $\bar{y}_i$ | 60.62 | 69.30 | 100.35 | 86.24 | $\bar{y}_{..} = 78.01$ |
| $n_I$ | 5 | 5 | 4 | 5 | $n = 19$ |
| $Res\ SS_I$ | 37.57 | 34.26 | 22.97 | 33.55 | $Res\ SS = 128.35$ |

To calculate each *MS*, consider what each component is estimating:

- *Residual SS* estimates variation within experimental units treated alike
- *Treatment SS* estimates the variation among each treatment average from the average of all observations.

Dividing each *SS* by its *df* estimates the average squared deviation (variance) for each component.

*Residual SS* for Treatment 1 (call it *Res* $SS_1$), where $\bar{y}_1 = 60.62$:

$$\sum[(60.8 - 60.62)^2 + (57.0 - 60.62)^2 + (65.0 - 60.62)^2 + (58.6 - 60.62)^2 + (61.7 - 60.62)^2] = 37.57$$

*Res SS* = 37.57 + 34.26 + 22.97 + 33.55 = 128.35

*Total SS*: subtract every observation from the grand mean, square the result, then sum.

All relevant SS, *df*, and MS follow:

Total SS = 4354.70
Total *df* = 19 – 1 = 18

Treatment SS = 4226.35
Treatment *df* = 4 – 1 = 3
Treatment MS = Trt SS/Trt *df* = 4226.35/3 = 1408.78

Residual SS = 4354.70 – 4226.35 = 128.35
Residual *df* = *N* – *I* = 19 – 4 = 15
Residual MS = Res SS/Res *df* = 128.35/15 = 8.56

Calculate the *F*-statistic for the feeding treatment as:

$F_{3,15}$ = Trt MS/Res MS = 1408.78/8.56 = 164.64, $P < 0.0001$.

Bookkeeping is simplified by using an ANOVA table, in which calculations used in the *F*-test are organized and displayed.

Analysis of Variance

| Source (of Variation) | df | Sum of Squares | Mean Square | *F* Ratio | Prob > F |
|---|---|---|---|---|---|
| Treatment (Model) | 3 | 4226.35 | 1408.78 | 164.64 | <0.0001 |
| Error (Residual) | 15 | 128.35 | 8.56 | | |
| Total | 18 | 4354.70 | | | |

## Extra-Sum-of-Squares Principle

An alternate approach to the ANOVA *F*-test (that yields identical results) is based on the *extra-sum-of-squares principle*.

Consider the null hypothesis of equal group/treatment means.

Assume a one-way classification with *I* = 7 groups.

The initial hypothesis is: $H_o$: $\mu_1 = \mu_2 = ... = \mu_7$ and the alternative is that at least one of the means differs from the others.

***Full and Reduced Models***

ANOVA are formulated by comparing two models for the mean responses.

A *full model* is a general model that functions as a starting point; a reduced model is a special case (always simpler) of the full model obtained by imposing the restriction of the null hypothesis.

For comparing equality of all means, the *full model* includes a separate mean for each group.

The *reduced model*, obtained from the full model by supposing that the null hypothesis of equal means is true, in this case specifies a single mean for all populations:

| Group: | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Full (separate-means) model: | $\mu_1$ | $\mu_2$ | $\mu_3$ | $\mu_4$ | $\mu_5$ | $\mu_6$ | $\mu_7$ |
| Reduced (equal-means) model: | $\mu$ | $\mu$ | $\mu$ | $\mu$ | $\mu$ | $\mu$ | $\mu$ |

Specifying *full* and *reduced* models provides the framework for the *extra-sum-of-squares F*-test.

In a test of equality of group means, the *full* model is called the *"separate-means"* model and the *reduced* model is the *"equal-means"* model.

***Fitting the Models***

ANOVA functions by estimating parameters in both the *full* and *reduced* models and "asking" whether variability of responses about the estimated means is comparable in the two models.

The *estimated* means for each group different between models:

| Group: | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Full (separate-means) model: | $\bar{y}_1$ | $\bar{y}_2$ | $\bar{y}_3$ | $\bar{y}_4$ | $\bar{y}_5$ | $\bar{y}_6$ | $\bar{y}_7$ |
| Reduced (equal-means) model: | $\bar{y}$ | $\bar{y}$ | $\bar{y}$ | $\bar{y}$ | $\bar{y}$ | $\bar{y}$ | $\bar{y}$ |

where $\bar{y}$ is the average of all observations, the *grand mean*.

***Residuals***

A *residual* is an observed value minus its estimated mean.

Each observation in the data set can be linked to a group mean based on the *full* model and a different group mean based on the *reduced* model.

Therefore, if $y_{ij}$ is the response measured on the *j*th observation from the *i*th group, the residual from the *full* model is $y_{ij} - \bar{y}_i$ and the residual from the reduced model is $y_{ij} - \bar{y}_.$

If the null hypothesis is *correct*, the full and reduced models should be about equal in their ability to explain the data, so the residuals should be about the same for both models.

If the null hypothesis is *incorrect*, the residuals from the reduced model (equal-means) will be *larger* than those from the full model (separate-means).

A summary of the residuals for a given model is the *residual sums of squares* for that model, which is the sum of the squared residuals.

*Residual sums of squares* measure the variability in observations that remains unexplained by a particular model.

By determining the sum of the squared residuals for the full and reduced models separately, we can compare the differences between models to assess if the difference between models is large (indicating a poor fit) or is small enough to attribute to sampling variation.

***Extra-Sum-of-Squares F-statistic***

The *extra sum of squares* is a single number that summarizes the difference in sizes of residuals between full and reduced models:

*Extra SS* = *Residual SS* (reduced) – *Residual SS* (full)

*Extra sums of squares* measures the amount of unexplained variation in the reduced model that is explained by the full model.

To determine if there is "too much" variation left unexplained by the reduced model, compare it with the variability left unexplained by the full model using an *F*-statistic:

$$F - statistic = \frac{\text{Extra } SS \,/\, \text{Extra } df}{\hat{\sigma}^2_{full}}$$

where *Extra df* are the number of parameters representing means in the full model minus the number representing means in the reduced model, and $\hat{\sigma}^2_{full}$ is the estimate of $\sigma^2$ based on the full model.

The *F*-statistic, therefore, is the *Extra SS* per *extra df* scaled by the best estimate of variance.

Large *F*-statistics are associated with large differences in the sizes of residuals between models, which supplies the degree of evidence against the null hypothesis (here, equal means) in favor of the alternative (here, unequal means).

The test is summarized by its *P*-value which (again ☺) is the probability of finding an *F*-statistic as large or larger the observed *F* if the null hypothesis is true (i.e., means are equal).

*Feeding example*

Residual SS (reduced) = 4354.70
Residual SS (full) = 128.35

Residual *df* (reduced) = $n - 1 = 19 - 1 = 18$
Residual *df* (full) = $n - I = 19 - 4 = 15$

$\hat{\sigma}^2_{full}$ = MS Residual (full) = 8.56

Extra SS = 4354.70 – 128.35 = 4226.35
Extra *df* = 18 – 15 = 3

$F_{3, 15} = (4226.35 / 3) / 8.56 = 164.58$, $P < 0.0001$

## Fitting Full and Reduced Models with Statistical Software

Statistical packages do not automatically generate full and reduced models.

Instead, models are fit separately, and the extra SS *F* statistic calculated by hand.

Create new classification variables that reflect the structure of the full and each reduced model of interest.

The *full model* classification identifies the original levels in the study and is usually the most complex model (i.e., most parameters; often the separate-means model).

The *reduced model* classification identifies each level of interest in a simpler model (here, equal-means model).

For the feeding example, separate-means (full) and equal-means (reduced) models are:

| Group: | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Full (separate-means) model: | $\mu_1$ | $\mu_2$ | $\mu_3$ | $\mu_4$ |
| Reduced (equal-means) model: | $\mu$ | $\mu$ | $\mu$ | $\mu$ |

These *separate means* and *equal means* models are reflected by the first two classifications in the following data table:

| | Feed-type classification | | |
|---|---|---|---|
| *Y* (mass) | Full (separate means) | Reduced (equal means) | Reduced (Type 3 v. others) |
| 60.8 | 1 | 1 | 1 |
| 57 | 1 | 1 | 1 |
| 65 | 1 | 1 | 1 |
| 58.6 | 1 | 1 | 1 |
| 61.7 | 1 | 1 | 1 |
| 68.7 | 2 | 1 | 1 |
| 67.7 | 2 | 1 | 1 |
| 74 | 2 | 1 | 1 |
| 66.3 | 2 | 1 | 1 |
| 69.8 | 2 | 1 | 1 |
| 102.6 | 3 | 1 | 3 |
| 102.1 | 3 | 1 | 3 |
| 100.2 | 3 | 1 | 3 |
| 96.5 | 3 | 1 | 3 |
| 87.9 | 4 | 1 | 1 |
| 84.2 | 4 | 1 | 1 |
| 83.1 | 4 | 1 | 1 |
| 85.7 | 4 | 1 | 1 |
| 90.3 | 4 | 1 | 1 |

JMP's *Fit Y by X* platform yields the analysis for the *Separate Means* (full) classification:

**Oneway Analysis of Mass By Feed Grp**

**Summary of Fit**

| | |
|---|---|
| Rsquare | 0.970526 |
| Adj Rsquare | 0.964631 |
| Root Mean Square Error | 2.925178 |
| Mean of Response | 78.01053 |
| Observations (or Sum Wgts) | 19 |

**Analysis of Variance**

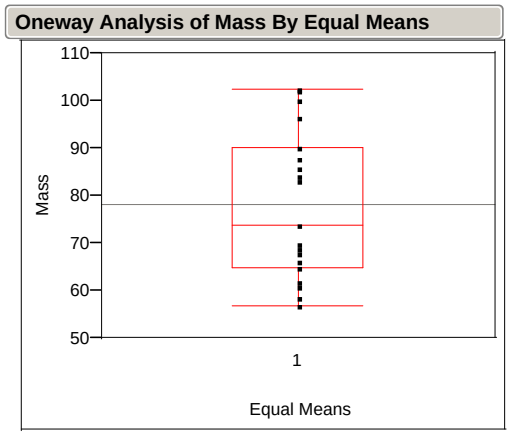
| Source | DF | Sum of Squares | Mean Square | F Ratio | Prob > F |
|---|---|---|---|---|---|
| Feed Grp | 3 | 4226.3479 | 1408.78 | 164.6415 | <.0001 |
| Error | 15 | 128.3500 | 8.56 | | |
| C. Total | 18 | 4354.6979 | | | |

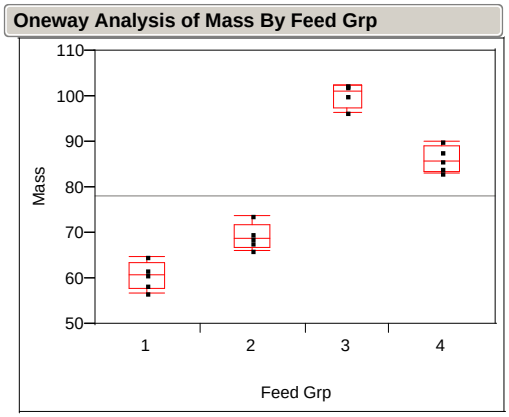
For the *Equal Means* (reduced) classification:

| | |
|---|---|
| Rsquare | 0 |
| Adj Rsquare | 0 |
| Root Mean Square Error | 15.55402 |
| Mean of Response | 78.01053 |
| Observations (or Sum Wgts) | 19 |

**Analysis of Variance**

| Source | DF | Sum of Squares | Mean Square | F Ratio | Prob > F |
|---|---|---|---|---|---|
| Equal Means | 0 | 0.0000 | . | . | . |
| Error | 18 | 4354.6979 | 241.928 | | |
| C. Total | 18 | 4354.6979 | | | |




Extra SS *F*-test to compare full and reduced models:

*F* = (Extra SS/Extra *df*) / (Residual SS (full)/Residual *df* (full)) = (Extra SS/Extra *df*) / Residual MS (full)

Extra SS = Residual SS (reduced) – Residual SS (full) = 4354.70 – 128.35 = 4226.35

Extra *df* = Residual *df* (reduced) – Residual *df* (full) = 18–15 = 3

Residual MS (full) = MSE (full) = 8.56

$F_{3,15}$ = (4226.35/3) / 8.56 = 1408.78 / 8.56 = 164.58, $p < 0.0001$

Which provides convincing evidence that the mean mass of pigs raised on at least one feed differs from those raised on other feeds.

*F* statistic identifies *df* in both the numerator and denominator.

Here, the numerator has 3 df (no. parameters in full model – no. parameters in reduced model) and the denominator has 15 *df* from error MS (or residual), taken from the full model.

In the case of a one-way classification, comparing a full model to a reduced models of *equal means* is identical to testing for a treatment effect.

## More Applications of the Extra SS F-test

Extra SS advantage becomes evident when more specific hypothesis tests are fit within a particular classification (i.e., models are *nested*).

E.g., say we're interested in whether the mean mass of pigs raised on feed 3 was the same as that pigs raised on the other 3 feeds. The appropriate set of full/reduced models is (to explain that portion of the *Treatment SS* explained by feed-type vs. others) is:

| Group: | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $H_a$: Full (others-equal) model: | $\mu_1$ | $\mu_1$ | $\mu_3$ | $\mu_1$ |
| $H_o$: Reduced (equal-means) model: | $\mu$ | $\mu$ | $\mu$ | $\mu$ |

The remaining *Treatment SS* is:

| Group: | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Full (others-equal) model: | $\mu_1$ | $\mu_2$ | $\mu_3$ | $\mu_4$ |
| Reduced (equal-means) model: | $\mu_1$ | $\mu_1$ | $\mu_3$ | $\mu_1$ |

The appropriate classification is in the last row of the data table above.

The output for the reduced model remains the same as before, and the output for this others-equal (full) model is:

**Analysis of Variance**

| Source | DF | Sum of Squares | Mean Square | F Ratio | Prob > F |
|---|---|---|---|---|---|
| Diet 1 vs 3 | 1 | 2528.5306 | 2528.53 | 23.5384 | 0.0001 |
| Error | 17 | 1826.1673 | 107.42 | | |
| C. Total | 18 | 4354.6979 | | | |

Extra SS = Res SS (reduced) – Res SS (full) = 4354.70 – 1826.17 = 2528.53

Extra *df* = Residual *df* (reduced) – Residual *df* (full) = 18 - 17 = 1

Residual MS (full) = MSE(full) = 8.56

$F_{1,15}$ = (2528.53/1) / 8.56 = 295.39, $p < 0.0001$

So there is convincing evidence that the mean mass of pigs raised on feed 3 differed from those raised on other feeds.

*Note that the F-statistic reported in JMP is not appropriate* (why?).

Several sub-models based on the same classification (nested) are possible, ranging from least specific (equal means), intermediate (others equal), and most specific (separate means). Each describe a different partitioning of the SS for treatment or group effects.

Because all of these classifications partition the same Treatment SS, they also each contribute to a reduction in the overall Residual SS (within-group SS or error SS).

We can examine how each nested model partitions the overall Treatment SS (from the separate means model) and subsequently reduces the overall Residual SS using a detailed ANOVA table:

Analysis of Variance

| Source (of Variation) | Sum of Squares | df | Mean Square | *F* Ratio | Prob > F |
|---|---|---|---|---|---|
| Among groups (Treatment) | 4226.35 | 3 | 1408.78 | 164.6 | <0.0001 |
| Feed-type 3 vs others | (2528.53) | (1) | 2528.53 | 295.4 | <0.0001 |
| Among other groups | (1697.82) | (2) | 848.91 | 99.2 | <0.0001 |
| Within groups (Residual or Error) | 128.35 | 15 | 8.56 | | |
| Total | 4354.70 | 18 | | | |

The "Among other groups" SS is found by subtraction. The appropriate *F*-stat is formed by dividing the "Among other groups" MS by the Error MS from the full model.

$F_{2,15}$ = (1697.82.82/2) / 8.56 = 99.46, $P < 0.0001$

**Robustness and Model Checking for *F*-tests**

Essentially these are simple extensions of the 2-sample case.

1. Sample Independence—Critical, both within and across groups.
2. Homogeneity of Variance—Critical.
3. Outliers—Tools not resistant to severe outliers.
4. Normality—Not critical. Extremely long tails or skewed distributions coupled with very different sample sizes, especially if sample sizes are small.

**Diagnostics Using Residual Plots**

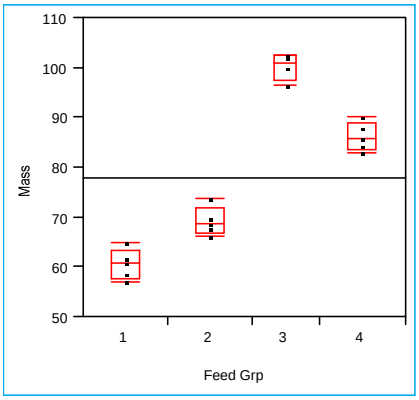
Examining side-by-side box plots is an important first step in these analyses; they help identify

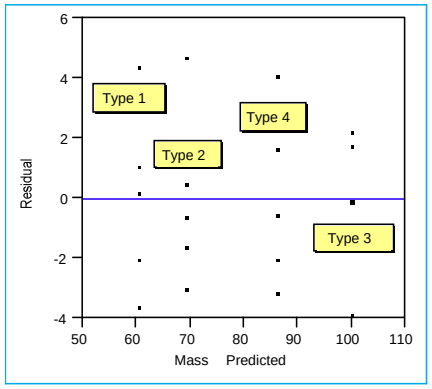
1. the centers
2. the relative spreads [and need for transformations]
3. the general shape
4. outliers.

Residual plots show the original observations with their group means subtracted out ($y_{ij} - \bar{y}_i$), thereby eliminating the visual interference of differences between means.

The features to look for are

1. an increase in spread from left to right in a *funnel-shape* pattern (suggesting need for a log transformation) or
2. serious outliers.




For the feed-type example, compare side-by-side quantile and residual plots:

## ☼ RNR 613 — Means Separation after ANOVA

In the Feeding example, we rejected the null of equal means ($F_{3,15} = 164.64$. $p < 0.0001$) and subsequently concludes mean body mass of pigs raised on different diets were not all equal.

This result suggests a series of more specific questions:

- Which treatments differed from each other?
- Which treatments were similar?

A significant *F*-statistic in a one-way classification can result from many outcomes, such as ($\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$) or ($\mu_1 \neq \mu_2 = \mu_3 = \mu_4$).

The process used to separate group means depends on whether specific comparisons were *planned* before the experiment or were *unplanned*.

**• Use multiple t-tests?**

Generate *t*-tests for all pairwise combinations of group means (for $I = 3$ groups, test $H_o$: $\mu_1 = \mu_2$, $H_o$: $\mu_2 = \mu_3$, $H_o$: $\mu_1 = \mu_3$)?

1. no. comparisons rises quickly with no. groups:
   for $I = 3$ there are 3 comparisons
   for $I = 7$ there are 21 comparisons
   for $I$ groups there are $I(I - 1)/2$ comparisons

2. With multiple comparisons, Type I error rate no longer = α; the true Type I error rate (α) is inflated: $\alpha = 1 - (1 - \alpha)^I$.

   For $I = 3$ and $\alpha = 0.05$, the true $\alpha = 1 - (1 - 0.05)^3 = 0.14$.

3. Multiple *t*-tests do not use the pooled estimate of variance from all groups — this can increase the efficiency of comparisons relative to multiple *t*-tests.

**• Planned Comparisons**

Appropriate with multiple treatment levels and a set of companions of special interest chosen ahead of time (i.e., *before any data have been collected* or *before the researcher has seen the collected data*).

If comparisons involve only a few pairs of means (say $\mu_3 - \mu_1$ and $\mu_4 - \mu_2$) then using simple *t*-tests with the pooled standard deviation ($s_p$) estimated from all groups:

$$t = \frac{(\bar{y}_1 - \bar{y}_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Questions often involve comparisons among sets of treatment means based on relationships among treatment groups.

E.g., in the Feeding study, we might compare means from pigs raised on feeds 1 and 2 (meat) to those raised on feeds 3 and 4 (soy).

We could reclassify groups so that pigs from feeds 1 and 2 are coded as “meat” and pigs from feeds 3 and 4 as “soy” and use a 2-sample *t*-test.

However, a better approach exists is to incorporate the pooled estimate of variance from all groups in the ANOVA framework.

This approach uses *Linear Combinations* of group means.

Use the parameter γ (gamma) to represent a linear combination of group means of the form: $\gamma = C_1 \mu_1 + C_2 \mu_2 + ... + C_I \mu_I$, where *I* is the number of groups and the *C*’s are coefficients chosen by the researcher to measure specific features of interest.

Estimate γ with *g*, where $g = C_1 \bar{y_1} + C_2 \bar{y_2} + ... + C_I \bar{y_I}$

*Example 1:* Compare average mass of pigs raised on feeds 1 and 2 (meat) to those on feeds 3 and 4 (soy).

Equivalent to testing: $(\mu_1 + \mu_2) = (\mu_3 + \mu_4)$ or $(\mu_1 + \mu_2) - (\mu_3 + \mu_4) = 0$.

Relevant linear combination of averages can be specified as $g = (\bar{y_1} + \bar{y_2}) - (\bar{y_3} + \bar{y_4})$,
where $C_1 = 1$, $C_2 = 1$, $C_3 = -1$, $C_4 = -1$.

Coefficients +½ +½ −½ −½ would yield identical results.

When a set of coefficients sum to zero, they are a type of linear combination called *orthogonal contrasts* or simply *contrasts*.

Compare results from these two approaches: On the left is the *t*-test using reclassified group labels, on the right is the same comparison done within ANOVA with *linear contrasts*.

Note considerable difference in standard errors between approaches and the small difference in the estimated difference between means:

**t Test**

Assuming equal variances

| | Difference | t Test | DF | Prob > \|t\| |
|---|---|---|---|---|
| Estimate | -27.551 | -8.973 | 17 | <.0001 |
| Std Error | 3.070 | | | |
| Lower 95% | -34.029 | | | |
| Upper 95% | -21.073 | | | |

UnEqual Variances

| | Difference | t Test | DF | Prob > \|t\| |
|---|---|---|---|---|
| Estimate | -27.551 | -8.792 | 13.9145 | <.0001 |
| Std Error | 3.134 | | | |
| Lower 95% | -34.140 | | | |
| Upper 95% | -20.962 | | | |

**Contrast**

**Test Detail**

| | |
|---|---|
| 1 | 0.5 |
| 2 | 0.5 |
| 3 | -0.5 |
| 4 | -0.5 |
| Estimate | -28.34 |
| Std Error | 1.3484 |
| t Ratio | -21.01 |
| Prob>\|t\| | 2e-12 |
| SS | 3778.2 |

| | |
|---|---|
| Sum of Squares | 3778.2222353 |
| Numerator DF | 1 |
| Denominator DF | 15 |
| F Ratio | 441.55304659 |
| Prob > F | 1.54011e-12 |

*Example 2*: Compare mean mass of pigs raised on feed 1 to those raised on feed 2.

Equivalent to testing $\mu_1 = \mu_2$ (or $\mu_1 - \mu_2 = 0$)
Appropriate coefficients are $C_1 = 1$, $C_2 = -1$, $C_3 = 0$, $C_4 = 0$

As above, a simple 2-sample *t*-test between groups (in isolation) versus the contrast after ANOVA:

**t Test**

Assuming equal variances

| | Difference | t Test | DF | Prob > \|t\| |
|---|---|---|---|---|
| Estimate | -8.6800 | -4.580 | 8 | 0.0018 |
| Std Error | 1.8951 | | | |
| Lower 95% | -13.0501 | | | |
| Upper 95% | -4.3099 | | | |

UnEqual Variances

| | Difference | t Test | DF | Prob > \|t\| |
|---|---|---|---|---|
| Estimate | -8.680 | -4.580 | 7.98307 | 0.0018 |
| Std Error | 1.895 | | | |
| Lower 95% | -13.052 | | | |
| Upper 95% | -4.308 | | | |

**Contrast**

**Test Detail**

| | |
|---|---|
| 1 | 1 |
| 2 | -1 |
| 3 | 0 |
| 4 | 0 |
| Estimate | -8.68 |
| Std Error | 1.85 |
| t Ratio | -4.692 |
| Prob>\|t\| | 0.0003 |
| SS | 188.36 |

| | |
|---|---|
| Sum of Squares | 188.356 |
| Numerator DF | 1 |
| Denominator DF | 15 |
| F Ratio | 22.012777561 |
| Prob > F | 0.0002893773 |

*Example 3*: Compare mean mass of pigs raised on feed 3 to those raised on feeds 1,2,4.

Equivalent to testing: $\mu_3$ = the mean of $\mu_1 + \mu_2 + \mu_4$
or: $\mu_3 = (\mu_1 + \mu_2 + \mu_4)/3$.

Appropriate coefficients are $C_3 = 1$, $C_1 = -1/3$, $C_2 = -1/3$, $C_4 = -1/3$ (note the coefficients sum to zero).

Recall that this was the same hypothesis we tested using the *extra SS* approach within the context of ANOVA. Compare that result to that of a linear contrast:

Analysis of Variance

| Source (of Variation) | Sum of Squares | df | Mean Square | *F* Ratio | Prob > F |
|---|---|---|---|---|---|
| Among groups (Treatment) | 4226.35 | 3 | 1408.78 | 164.64 | <0.0001 |
| Feed 3 vs others | 2528.53 | 1 | 2528.53 | 295.39 | <0.0001 |
| Among other groups | 1702.82 | 2 | 851.41 | 99.46 | <0.0001 |
| Within groups (Residual or Error) | 128.35 | 15 | 8.56 | | |
| Total | 4354.70 | 18 | | | |

Note that the SS, *F*-statistic, and *P* for the "Feed 3 vs others" effect in the extra SS model above is the same as that reported in the "contrast" output window.
Recall that a 2-sample *t*-test is identical to a 1-way ANOVA with 2 levels or groups.

Note that because we are interested in comparing 2 groups (feed 3 versus 1,2,4), that either a *t* or *F* is appropriate. In the contrast output we get both; these are identical as $t^2 = F$ (here $(17.19)^2 = 295.5$).

The benefit of using linear combinations within the framework of ANOVA allows us to use the pooled variance estimate from all 4 groups as denominators for test statistics.

**Contrast**

**Test Detail**

| | |
|---|---|
| 1 | -0.333 |
| 2 | -0.333 |
| 3 | 1 |
| 4 | -0.333 |
| Estimate | 28.297 |
| Std Error | 1.6461 |
| t Ratio | 17.19 |
| Prob>\|t\| | 3e-11 |
| SS | 2528.5 |

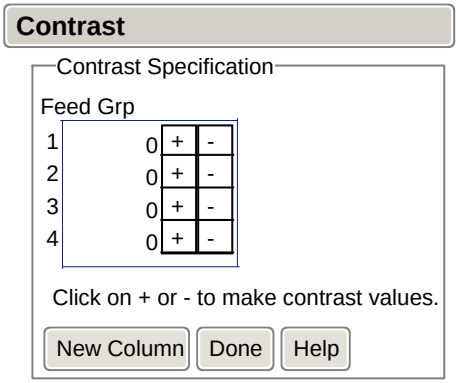
| | |
|---|---|
| Sum of Squares | 2528.5305614 |
| Numerator DF | 1 |
| Denominator DF | 15 |
| F Ratio | 295.50415599 |
| Prob > F | 2.794498e-11 |

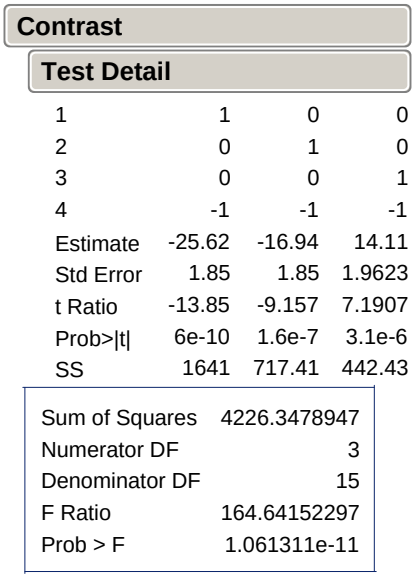
**• Using JMP to specify and solve linear combinations**

Launch *Fit Model* platform, specify the appropriate variables for the ANOVA, including the response variable *Y* (mass) and explanatory effects *X* (feed type).

Running the model generates the *F*-statistic that test the null hypothesis of equal means which we reject.

To examine the more specific questions with linear combinations, click the triangle next to the effect of interest and select *LS Means Contrast...* which yields a "Contrast Specification" dialog box where you click + or - to specify the contrast of interest. The appropriate orthogonal coefficients are generated automatically.




**• Unplanned Comparisons**

Are appropriate when there were no pre-planned comparisons of group means; hence, all possible pairs of means compared.

As the number of groups increase, the number of pairwise comparisons increases dramatically; as a result, the true α is lower than the established α-level (α is the probability of a Type I error).

E.g., with $I = 3$ groups and $\alpha = 0.05$, the true $\alpha = 0.14$.

A 95% confidence interval for an estimate captures its parameter 95% of the time. Considering several 95% CI's simultaneously is called a *family* of CI's. The frequency with which all of the intervals in the family simultaneously capture their parameters is always < 95%.

This problem increases with the number of groups and comparisons, and is known as *The Multiple Comparisons* problem for *Simultaneous Inferences*. Hence, we distinguish:

• *Individual or Pairwise confidence level:* frequency with which *a single interval* captures its parameter.

• *Overall or Familywise confidence level:* frequency with which *all intervals* capture their parameters.

With a family of $k$ CI's, each with pairwise confidence level 95%, the familywise CL can be no > 95% and no < $100(1 - 0.05k)$.

If $k = 3$, the smallest familywise CL possible is $100(1 - 0.05[3])=85\%$.

*Upshot:* This *compound uncertainty* increases the probability of making mistakes (finding differences that do not really exist which are Type I errors) when drawing more than one inference.

• If interested in only a few *planned comparisons*, the <u>pairwise</u> confidence level must be considered and controlled.

• If interested in all possible pairs of groups, the <u>familywise</u> confidence level should be considered and controlled.

Many procedures exist to combat compound uncertainty resulting from unplanned multiple comparisons;

each attempts to maintain the specified α-level for the entire "family" of pairwise comparisons.

Multiple comparisons procedures function by modifying the usual confidence interval for differences between pairs of means with a multiplier that adjusts the *t*-value used for a single interval:

Interval half-width = (Multiplier) x (Standard error)

*Standard error of the difference* is the usual pooled standard deviation times the square root of the sum of reciprocals of sample sizes.

*Interval half-width* is compared to the difference between each pair of means (e.g., $|\bar{y}_i - \bar{y}_j|$); if this difference is > the specified half-width, the difference is considered > the specified α.

Multiple comparisons procedures include *Tukey-Kramer*, *Scheffé*, *Protected LSD*, *Bonferroni*, and others*;* each attempts to controls the familywise confidence level differently.

For example, the multiplier for *Scheffé's procedure* is $\sqrt{(I-1)F_{(I-1),df\,(1-\alpha)}}$ which is based on an (1 – α) percentile *F*-distribution with *I* – 1 (between-group) numerator and *df* (within-group) denominator degrees of freedom.

The Scheffé multiplier controls the overall confidence level for the *family* of parameters consisting of *all* linear contrasts among group means.

When used for the smaller family of differences between *pairs* of group means, the overall confidence is at least 100(1 – α)% (e.g., 95%) and is generally higher (e.g., 98%).

*Using JMP for Unplanned Comparisons*

Choose the *Fit Y to X* platform, the appropriate *Y* and *X* variables, then *Compare Means.* JMP offers 4 multiple comparison tests:

• **Compare Each Pair:** LSD procedure; computes pairwise comparisons among means using Student's *t* test. This procedure does not control the familywise error rate;

• **Compare All Pairs:** Tukey or Tukey-Kramer HSD procedure; controls the familywise error rate; exact test if sample sizes are the same; conservative if sample sizes differ.

• **Compare with Best:** Hsu MCB procedure; tests whether means are less than the unknown maximum or greater than the unknown minimum.

• **Compare with Control:** Dunnett's procedure; tests whether means are different from the mean of a control group.

You can adjust the α-level used by these procedures (default is 5%) if you choose *Set Alpha Level* from the menu within the *Fit Y by X* platform.

Note that LSD and Tukey are also available in the Fit Model Platform.

**Means Comparisons**

Dif=Mean[i]-Mean[j]

| | 3 | 4 | 2 | 1 |
|---|---|---|---|---|
| 3 | 0.000 | 14.110 | 31.050 | 39.730 |
| 4 | -14.110 | 0.000 | 16.940 | 25.620 |
| 2 | -31.050 | -16.940 | 0.000 | 8.680 |
| 1 | -39.730 | -25.620 | -8.680 | 0.000 |

Alpha= 0.05

Comparisons for all pairs using Tukey-Kramer HSD

| q* | Alpha |
|---|---|
| 2.88215 | 0.05 |

Abs(Dif)-LSD

| | 3 | 4 | 2 | 1 |
|---|---|---|---|---|
| 3 | -5.961 | 8.454 | 25.394 | 34.074 |
| 4 | 8.454 | -5.332 | 11.608 | 20.288 |
| 2 | 25.394 | 11.608 | -5.332 | 3.348 |
| 1 | 34.074 | 20.288 | 3.348 | -5.332 |

Positive values show pairs of means that are significantly different.

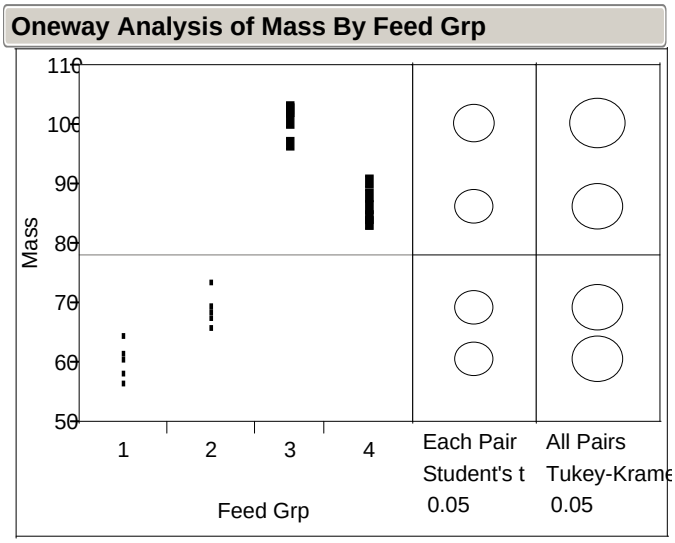
| Level | | Mean |
|---|---|---|
| 3 | A | 100.35000 |
| 4 | B | 86.24000 |
| 2 | C | 69.30000 |
| 1 | D | 60.62000 |

Levels not connected by same letter are significantly different

Additionally, in *JMP* provides *Means Comparison Circles* as a visual way to compare pairs of means.

When you click on a mean's circle, the circles for all means that are not significantly different from the selected mean at the specified α-level are highlighted. Circles for means that differ either do not intersect or intersect slightly; if circles intersect at an angle of more than 90 degrees or if they are nested, means do not differ at the specified α-level.

When you click on a mean's circle, the circles for all means that are not significantly different from the selected mean at the specified α-level are highlighted. Circles for means that differ either do not intersect or intersect slightly; if circles intersect at an angle of more than 90 degrees or if they are nested, means do not differ at the specified α-level.




**Choosing a Multiple Comparison's Procedure**

The most appropriate procedure to use depends on the specific application.

SAS offers about 15 different procedures. In general, LSD is the most liberal (i.e., has the narrowest confidence intervals) and Scheffé the most conservative (widest confidence intervals).

If the comparisons include all pairs of means, then Tukey-Kramer is recommended by *Sleuth*.