Chapter 8

ST 511 - Analysis of Variance for Comparing Means

Readings: Chapter 8 (read 8.1-8.4)

We now turn our focus back to comparing means from different populations. In chapter 6, we saw how to compare the (true) means from two (roughly) normal populations (both with equal and unequal variance).

We may however have interest in comparing the (true) means from more than two populations, say t populations. This type of question often comes up when conducting a completely randomized experiment (CRD).

In a CRD we have N total units (the book uses n_T to represent this number). We have t treatments - recall a treatment is a specific experimental condition which, in an experiment, we assign to the experimental units. This treatment may come from the levels of a single factor or the combinations of levels from several factors.

Let n_i denote the number of replicates for treatment i (i=1,...,t) (i.e. the # of units assigned to that treatment). In a balanced CRD design, we have $n_i = n$ for every treatment i. Thus, the total number of units can be given by $N = nt$ (or in the book notation $n_T = nt$).

The CRD design randomly assigns the treatments to the units (treating every unit as interchangeable).

Therefore, we can consider having t different populations that we now want to compare. For instance, we may want to know if the means are equal for each population (or the standard deviations, medians, etc.)

Example: (some description taken from Goosen, 2014)
Consider having 24 pieces of cheese. Color of the cheese is important in terms of consumer satisfaction. We have interest in how the color differs for 4 different types of corn syrup (26, 42, 55, and 62) (4 treatments). A CRD design is decided upon and we randomly assign each corn syrup type to 6 pieces of cheese (6 replicates for each treatment).

As a response, we measure the color using a 3 part CIE L*a*b* Color System.

- 'L' reflects the lightness of a sample, from black (L = 0) to white (L = 100) and runs from top to bottom.
- 'a' defines the shades from red (positive values) to green (negative values).
- 'b' defines the shades from yellow (positive values) to blue (negative values).

All three of these could be treated as responses (and analyzed together), but for our purposes we will only look at the 'L' response variable.

Again, we will focus on the means of the population. How might we make inference here?

Define

- μ_1 = mean 'L' score for **all** pieces of cheese that with corn syrup 26.
- μ_2 = mean 'L' score for **all** pieces of cheese that with corn syrup 42.
- μ_3 = mean 'L' score for **all** pieces of cheese that with corn syrup 55.
- μ_4 = mean 'L' score for **all** pieces of cheese that with corn syrup 62.

We want to test the hypotheses

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \quad vs \quad H_A : \text{at least one mean differs}$$

For two independent samples, we said if the samples came from normal populations with equal variance we could use

$$T = \frac{\bar{Y} - \bar{X}}{\sqrt{S_p^2(1/n_1 + 1/n_2)}} \quad \text{where} \quad S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

as a test statistic to make our inference. Now we have more than 2 populations, so this exact set-up won't work, but we can do something else.

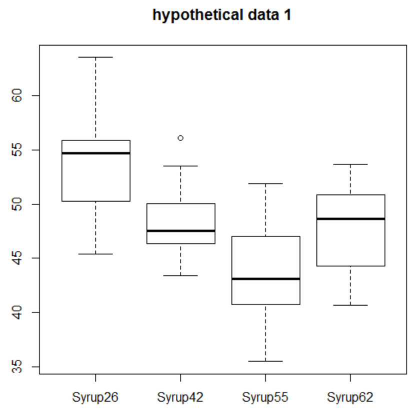
ANOVA for analyzing a CRD

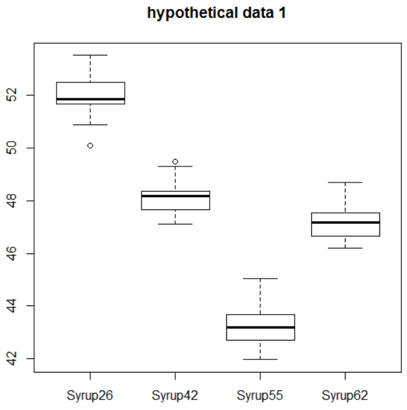
Data and labeling:

Corn Syrup	Replicate #	'L' measurement	Label
26	1	51.89	y_{11}
26	2	51.52	y_{12}
26	3	52.69	y_{13}
26	4	52.06	y_{14}
26	5	51.63	y_{15}
26	6	52.73	y_{16}
42	1	47.21	y_{21}
42	2	48.57	y_{22}
42	3	47.57	y_{23}
42	4	46.85	y_{24}
42	5	48.64	y_{25}
42	6	47.49	y_{26}
55	1	41.43	y_{31}
55	2	42.31	y_{32}
55	3	42.31	y_{33}
55	4	41.49	y_{34}
55	5	42.12	y_{35}
55	6	42.65	y_{36}
62	1	45.99	y_{41}
62	2	46.66	y_{42}
62	3	47.35	y_{43}
62	4	45.83	y_{44}
62	5	46.77	y_{45}
62	6	47.88	y_{46}

We now need two subscripts to represent which observation we are talking about. The first subscript (i) represents the treatment group, where as the second subscript (j) represent the replicate number.

Consider the following two hypothetical set of boxplots for this data. Which would give evidence that the (true) means differ?





ANOVA = Analysis of Variance

In this case, compare variation 'within' groups to variation 'between' groups.

Assumptions:

$$Y_{1j} \sim^{iid} N(\mu_1, \sigma^2)$$
$$Y_{2j} \sim^{iid} N(\mu_2, \sigma^2)$$
$$Y_{3j} \sim^{iid} N(\mu_3, \sigma^2)$$
$$Y_{4j} \sim^{iid} N(\mu_4, \sigma^2)$$

and each sample is independent of one another.

That is, the populations are independent random samples from normally distributed parent populations with equal variances. Rather than write this all out we will just say

$$Y_{ij} \sim^{iid} N(\mu_i, \sigma^2)$$

Within group variation

In two samples, to estimate the common variance σ^2 we used S_p^2. Here we use the same exact idea:

$$MS(E) = MS(W) = S_w^2$$

$$= \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2 + ... + (n_t-1)S_t^2}{n_1+n_2+...+n_t-t}$$

For a balanced design we have

$$MS(E) = MS(W) = S_w^2 = \frac{(n-1)(S_1^2 + ... + S_t^2)}{N-t}$$

$$= \frac{(n-1)(S_1^2 + ... + S_t^2)}{nt-t} = \frac{S_1^2 + ... + S_t^2}{t}$$

(i.e. just the simple average of the variances).

With the double subscript our formula for S_i^2 is given by

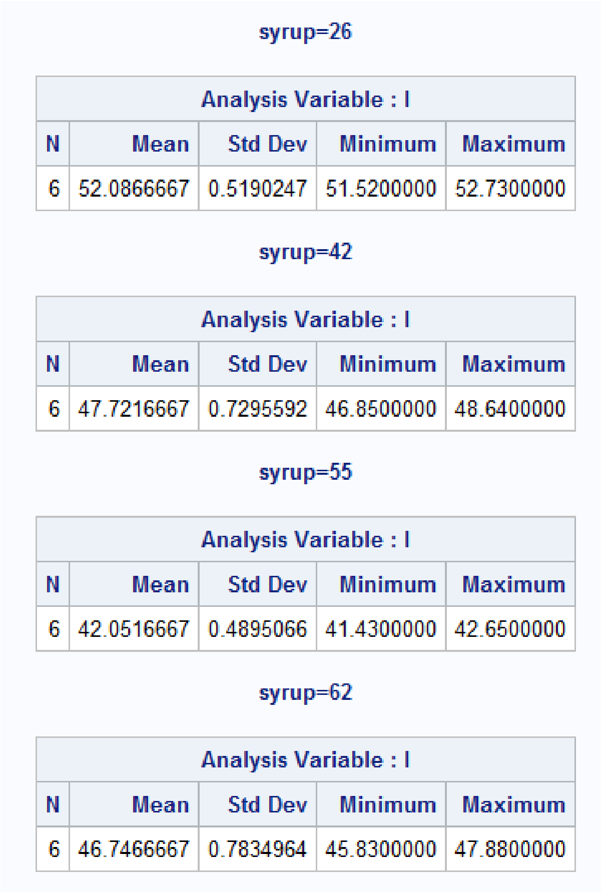
$$S_i^2 = \frac{\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_{i\bullet})^2}{n-1}$$

where the group mean $\bar{Y}_{i\bullet}$ is given by

$$\bar{Y}_{i\bullet} = \frac{\sum_{j=1}^{n} Y_{ij}}{n}$$

Thus, for a balanced design $MS(E)$ is written

$$MS(E) = \frac{\sum_{i=1}^{t}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_{i\bullet})^2}{t(n-1)}$$



syrup=26

Analysis Variable : l				
N	Mean	Std Dev	Minimum	Maximum
6	52.0866667	0.5190247	51.5200000	52.7300000

syrup=42

Analysis Variable : l				
N	Mean	Std Dev	Minimum	Maximum
6	47.7216667	0.7295592	46.8500000	48.6400000

syrup=55

Analysis Variable : l				
N	Mean	Std Dev	Minimum	Maximum
6	42.0516667	0.4895066	41.4300000	42.6500000

syrup=62

Analysis Variable : l				
N	Mean	Std Dev	Minimum	Maximum
6	46.7466667	0.7834964	45.8300000	47.8800000

Figure 8.1: summary from proc means. - proc means data=cheese; by syrup; var L; run;

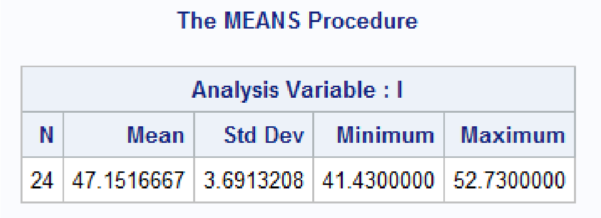
Between group variation

Variation between groups is judged by the variation between the group means:

$$MS(T) = MS(B) = S_b^2 = \frac{\sum_{i=1}^{t}\sum_{j=1}^{n}(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2}{t-1}$$

where $\bar{Y}_{\bullet\bullet}$ is the overall mean

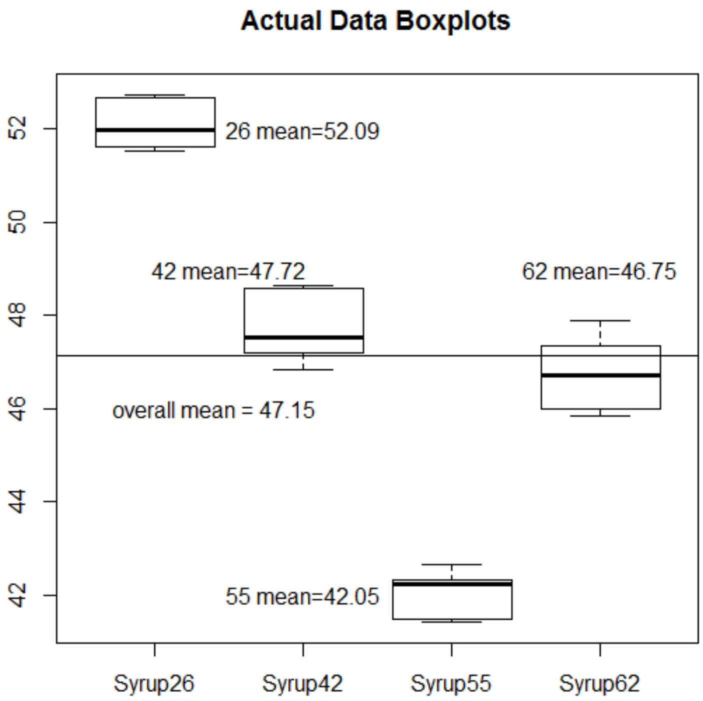
$$\bar{Y}_{\bullet\bullet} = \frac{\sum_{i=1}^{t}\sum_{j=1}^{n} Y_{ij}}{nt}$$



The MEANS Procedure

Analysis Variable : l				
N	Mean	Std Dev	Minimum	Maximum
24	47.1516667	3.6913208	41.4300000	52.7300000

Figure 8.2: summary from proc means proc means data=cheese; var L; run;



For our hypotheses:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \quad vs \quad H_A : \text{at least one mean differs}$$

Test statistic is:

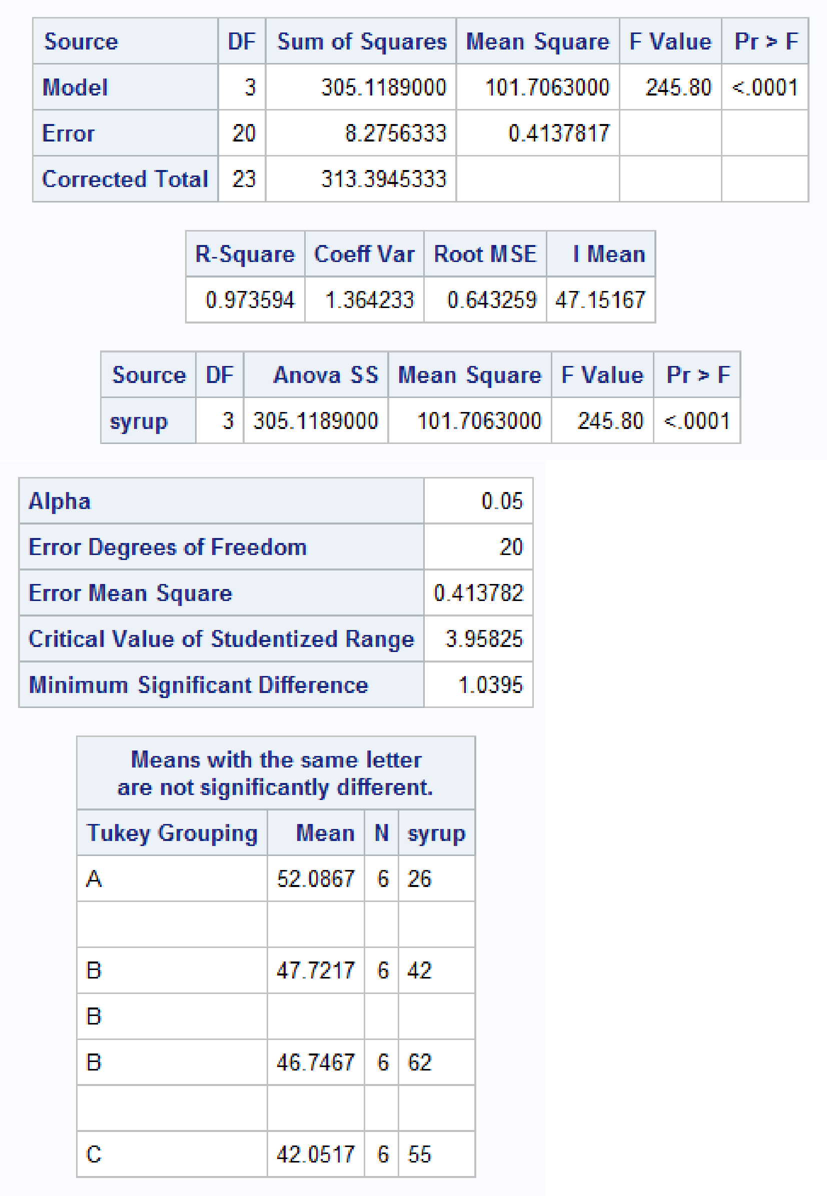
$$F = \frac{MS(T)}{MS(E)} = \frac{S_b^2}{S_w^2} \sim F_{t-1,N-t}$$

We reject H_0 for large values of F, values greater than $F_{t-l,N-t,\alpha}$.

We get a p-value by $P(F_{t-1,N-t} > F_{obs})$.

To get this analysis in SAS we can run the code:

```
proc anova data=cheese;
      class syrup;
      model L = syrup;
      means syrup/tukey;
run;
```



Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	305.1189000	101.7063000	245.80	<.0001
Error	20	8.2756333	0.4137817		
Corrected Total	23	313.3945333			

R-Square	Coeff Var	Root MSE	l Mean
0.973594	1.364233	0.643259	47.15167

Source	DF	Anova SS	Mean Square	F Value	Pr > F
syrup	3	305.1189000	101.7063000	245.80	<.0001

Alpha	0.05
Error Degrees of Freedom	20
Error Mean Square	0.413782
Critical Value of Studentized Range	3.95825
Minimum Significant Difference	1.0395

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	syrup
A	52.0867	6	26
B	47.7217	6	42
B			
B	46.7467	6	62
C	42.0517	6	55

How do we view the ANOVA ‘model’?

We made the assumption that $Y_{ij} \sim^{iid} N(\mu_{ij}, \sigma^2)$. Instead of viewing it in this form, we look at it as

$$Y_{ij} = \mu_i + E_{ij}$$

where $E_{ij} \sim^{iid} N(0, \sigma^2)$. So for $i = 1$ we have

$$Y_{1j} = \mu_1 + E_{1j}$$

which is adding the constant μ_1 to the $N(0, \sigma^2)$ distribution, giving $Y_{1j} \sim^{iid} N(\mu_1, \sigma^2)$.

Usually we then change this to a different parameterization -

$$Y_{ij} = \mu + \tau_i + E_{ij}$$

where μ is the overall (grand) mean, τ_i is the effect for the i^{th} treatment, and $E_{ij} \sim^{iid} N(0, \sigma^2)$.

So for $i = 1$ we have

$$Y_{1j} = \mu + \tau_1 + E_{1j}$$

which is adding the constant $\mu + \tau_1$ to the $N(0, \sigma^2)$ distribution, giving $Y_{1j} \sim^{iid} N(\mu + \tau_1, \sigma^2)$.

Relationship between parameterizations:

Population	Mean 1^{st} way	Mean 2^{nd} way	Variance
1	μ_1	$\mu + \tau_1$	σ^2
2	μ_2	$\mu + \tau_2$	σ^2
3	μ_3	$\mu + \tau_3$	σ^2
4	μ_4	$\mu + \tau_4$	σ^2

Our hypothesis now becomes

$$H_0 : \tau_1 = \tau_2 = ... = \tau_t \quad vs \quad H_A : \text{at least one differs}$$

We then analyze the model using an analysis of variance table (ANOVA table). **Table for balanced one-way ANOVA:**

Source	DF	SS	MS	F
Treatments	$t-1$	$SS(T)$	$MS(T) = \frac{SS(T)}{(t-1)}$	$F = \frac{MS(T)}{MS(E)}$
Error	$t(n-1)$	$SS(E)$	$MS(E) = \frac{SS(E)}{(N-t)}$	
Total	$nt-1$	$SS(TOT)$		

where

$$\begin{aligned} SS(T) &= \sum_{i=1}^{t}\sum_{j=1}^{n}(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2 = n\sum_{i=1}^{t}(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2 \\ SS(E) &= \sum_{i=1}^{t}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\bullet})^2 \\ SS(Tot) &= \sum_{i=1}^{t}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{\bullet\bullet})^2 \end{aligned}$$

Note: SS(T) is also called SS(Between) and SS(E) is also called SS(Within).

We can now see the overall idea of ANOVA.

Consider $SS(Tot)$ (sum of squares total or total sum of squares), which, if we divide by $nt-1$ we get the sample variance of the y's (over all the treatments). This is a measure of the variation in the response.

We take this $SS(Tot)$ and 'partition' it into a piece due to the different 'sources' in our model. Here the sources are the treatments and error (about the treatments). Thus

$$SS(Tot) = SS(T) + SS(E)$$

Similarly, the degrees of freedom add up.

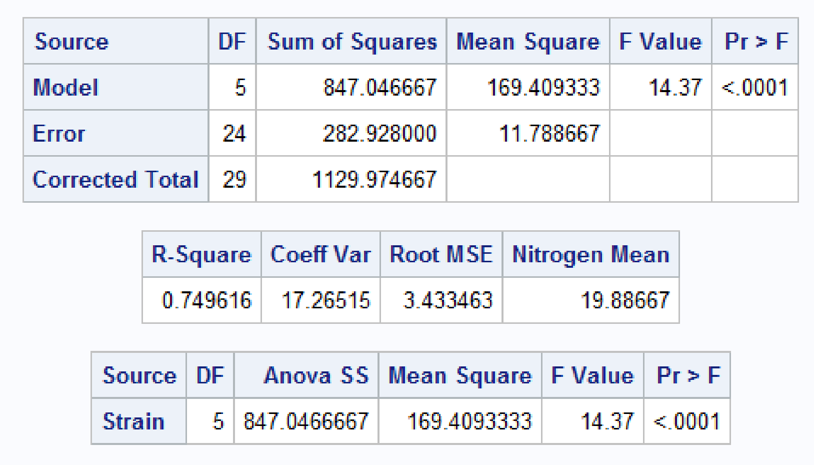
$$df_{Tot} = df_T + df_E \quad or \quad nt-1 = t(n-1) + (t-1)$$

The sum of squares represent variability from each source. When we divide by the degrees of freedom, this standardizes that measure of variation (and we call this a mean square).

Our test then becomes the ratio of the $MS(T)$ to the $MS(E)$.

$$F = \frac{MS(T)}{MS(E)}$$

Example:
The following example studies the effect of bacteria on the nitrogen content of red clover plants. The treatment factor is bacteria strain, and it has six levels. Red clover plants are inoculated with the treatments, and nitrogen content is later measured in milligrams. The data are derived from an experiment by Erdman (1946) and are analyzed in Chapters 7 and 8 of Steel and Torrie (1980). Conduct a test to determine if the means are equal at the 0.05 level. Be sure to show all 5 steps (use p-values).



Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	847.046667	169.409333	14.37	<.0001
Error	24	282.928000	11.788667		
Corrected Total	29	1129.974667			

R-Square	Coeff Var	Root MSE	Nitrogen Mean
0.749616	17.26515	3.433463	19.88667

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Strain	5	847.0466667	169.4093333	14.37	<.0001

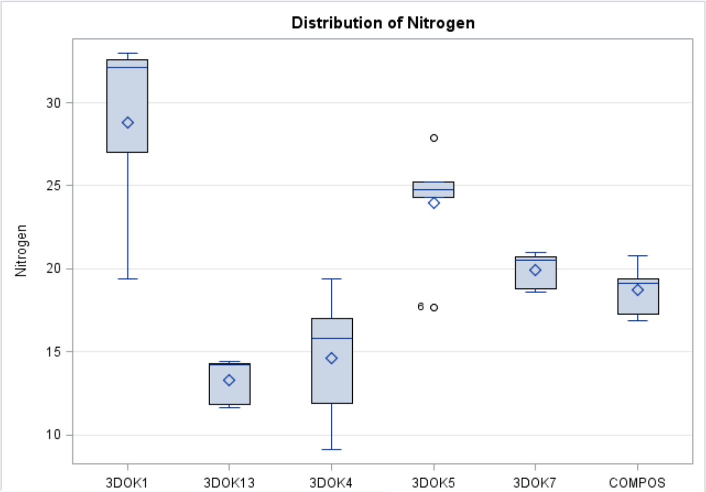
How do we check our assumptions?

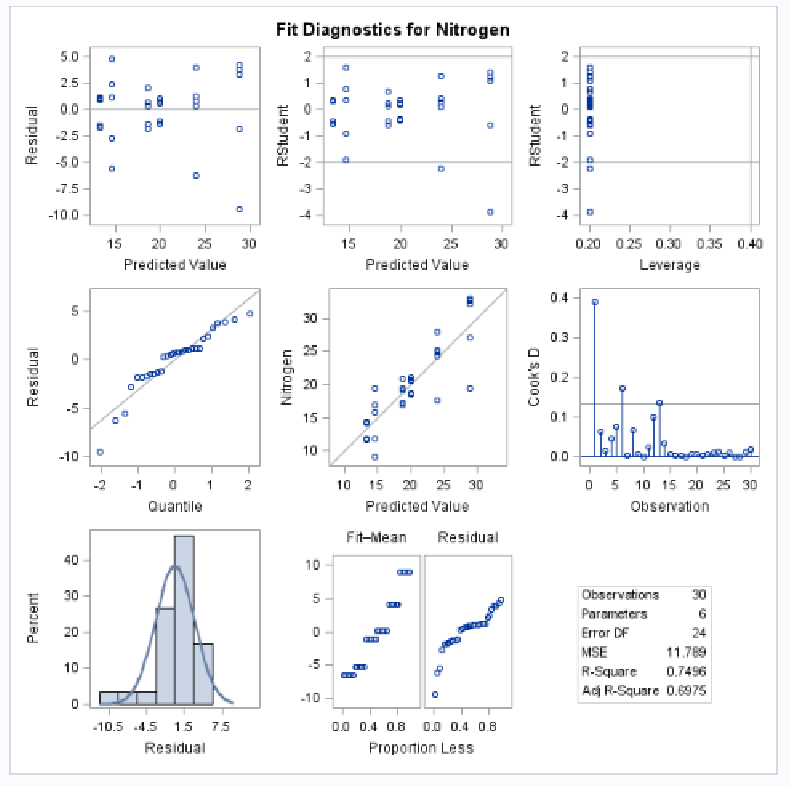
To investigate the constant variance assumption, we can look at side-by-side box plots or residual vs predicted plots. A residual is the observed value minus the predicted value. For the ij^{th} observation the residual is

$$r_{ij} = obs - pred = y_{ij} - \bar{y}_{i\bullet}$$

If the constant variance assumption is violated, one thing we might do is try a transformation of the data. This will be looked at in ST 512.

To check the normality, we look for a straight line in the qq-plot (or quantile vs residual plot) and also we hope to see a roughly normal looking histogram in the bottom left panel. (Note: the transformation idea above may also solve some nonnormality issues.)





If we reject H_0 and conclude that the treatment means differ, the next logical question to ask is which treatment means are the ones that differ.

To answer this question we usually look at all pairwise comparisons of treatment means. That is, if we reject H_0, we would look at

$$\begin{array}{cccc} \mu_1 - \mu_2 & \mu_1 - \mu_3 & \cdots & \mu_1 - \mu_t \\ & \mu_2 - \mu_3 & \cdots & \mu_{t-1} - \mu_t \end{array}$$

to see which differ.

Let's focus on $\mu_1 - \mu_2$. We get an estimator this quantity with the corresponding sample means

$$\hat{\mu}_1 - \hat{\mu}_2 = \bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}$$

The standard error of this quantity can be found by taking the square root of the variance (recall we assume our samples are independent so covariance is 0)

$$\begin{aligned} Var(\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}) &= (1^2)Var(\bar{Y}_{1\bullet}) + (-1)^2\bar{Y}_{2\bullet} + 2(1)(-1)Cov(\bar{Y}_{1\bullet}, \bar{Y}_{2\bullet}) \\ &= Var(Y_{1j})/n_1 + Var(Y_{2j})/n_2 = \sigma^2/n_1 + \sigma^2/n_2 \end{aligned}$$

For a balanced design we have

$$Var(\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}) = \sigma^2(1/n + 1/n) = 2\sigma^2/n$$

yielding a standard error of

$$SE(\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}) = \sqrt{2\sigma^2/n}$$

By the normality assumption on the data we then have a case similar to the two-sample t test with pooled variance!

$$\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet} \sim N(\mu_1 - \mu_2, \sigma^2(1/n_1 + 1/n_2)) = N(\mu_1 - \mu_2, 2\sigma^2/n)$$

We estimate σ^2 by the common pooled estimate (over all the samples, not just these two)

$$MS(E) = S_w^2 = \frac{\sum_{i=1}^{t}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_{i\bullet})^2}{N - t}$$

Thus, we can for a t-test for testing $H_0 : \mu_1 = \mu_2 \quad vs \quad H_A : \mu_1 \neq \mu_2$ using

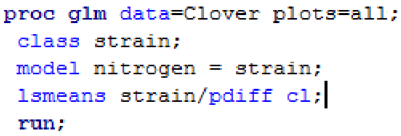
$$T = \frac{\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}}{\sqrt{MS(E)(1/n_1 + 1/n_2)}} = \frac{\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}}{\sqrt{2MS(E)/n}} \sim t_{N-t}$$

and we can form a $(1-\alpha)100\%$ CI for $\mu_1 - \mu_2$ using

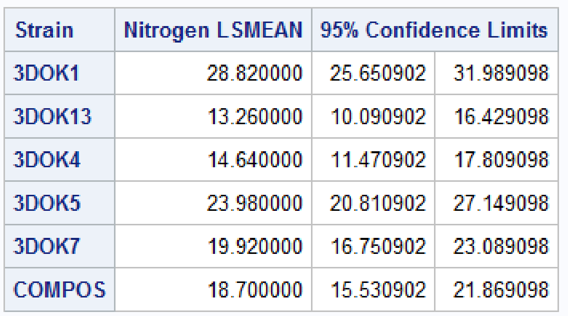
$$\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet} \pm t_{\alpha/2,N-t}\sqrt{MS(E)(1/n_1 + 1/n_2)} \quad or \quad \bar{Y}_{1\bullet} - \bar{Y}_{2\bullet} \pm t_{\alpha/2,N-t}\sqrt{2MS(E)/n}$$

and check if 0 is in the interval.

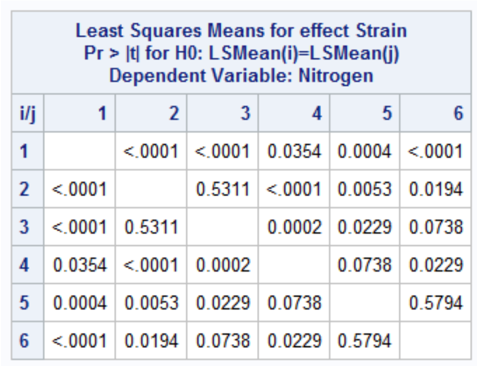
An issue arises! If we do this for all of the possible pairwise comparisons, we control the type I error rate for each test/interval, but by doing so many tests, the overall probability of making **at least one type I error** may be much more than α. This is known as the multiple comparison correction issue. This will be taken up in ST 512.



```
proc glm data=Clover plots=all;
 class strain;
 model nitrogen = strain;
 lsmeans strain/pdiff cl;
 run;
```

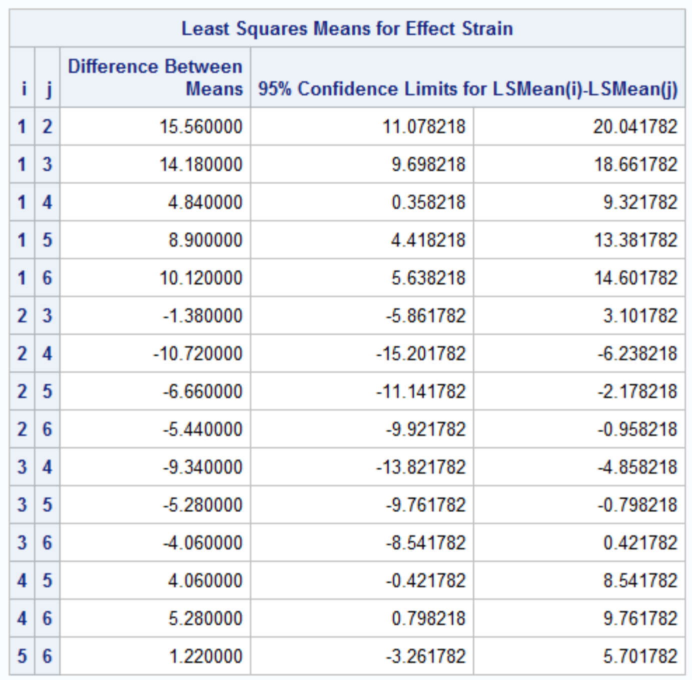


Strain	Nitrogen LSMEAN	95% Confidence Limits	
3DOK1	28.820000	25.650902	31.989098
3DOK13	13.260000	10.090902	16.429098
3DOK4	14.640000	11.470902	17.809098
3DOK5	23.980000	20.810902	27.149098
3DOK7	19.920000	16.750902	23.089098
COMPOS	18.700000	15.530902	21.869098



Least Squares Means for effect Strain
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: Nitrogen

i/j	1	2	3	4	5	6
1		<.0001	<.0001	0.0354	0.0004	<.0001
2	<.0001		0.5311	<.0001	0.0053	0.0194
3	<.0001	0.5311		0.0002	0.0229	0.0738
4	0.0354	<.0001	0.0002		0.0738	0.0229
5	0.0004	0.0053	0.0229	0.0738		0.5794
6	<.0001	0.0194	0.0738	0.0229	0.5794	



Least Squares Means for Effect Strain

i	j	Difference Between Means	95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	15.560000	11.078218	20.041782
1	3	14.180000	9.698218	18.661782
1	4	4.840000	0.358218	9.321782
1	5	8.900000	4.418218	13.381782
1	6	10.120000	5.638218	14.601782
2	3	-1.380000	-5.861782	3.101782
2	4	-10.720000	-15.201782	-6.238218
2	5	-6.660000	-11.141782	-2.178218
2	6	-5.440000	-9.921782	-0.958218
3	4	-9.340000	-13.821782	-4.858218
3	5	-5.280000	-9.761782	-0.798218
3	6	-4.060000	-8.541782	0.421782
4	5	4.060000	-0.421782	8.541782
4	6	5.280000	0.798218	9.761782
5	6	1.220000	-3.261782	5.701782

Multiple comparison correction needed!