3.0 ANOVA - ANALYSIS OF VARIANCE

(updated Spring 2005)

In our previous discussions, we considered two levels for the variable under study (i.e., engine type A versus engine type B). We will now examine a technique that can be used for examining multiple (two or more) levels for variables being studied. This technique is referred to as analysis of variance (ANOVA).

With the ANOVA technique, two estimates will be calculated:

- An estimate of the environmental error variance, σ_Y^2 :

  s_{PE}^2 (where PE stands for Pure Error).
- An estimate based on various assumptions that we will make:

  s_{test}^2

The two sample variances can be compared using the F distribution:

$$F_{calc} = \frac{S_{Test}^2}{S_{PE}^2} \sim F\nu_1, \nu_2$$

The value of F_{calc} will be inflated if the assumptions made are not satisfied.

EXAMPLE

Study the coagulation time for samples of blood drawn from 24 animals receiving 4 different diets. Does diet affect blood coagulation time? Twenty four (24) animals were randomly allocated to four diets (A,B,C, and D). The blood samples were taken and tested in random order.

The analysis that follows is exactly justified if the data are random samples from four normal populations, and is approximately justified based on the randomization dist$^{\underline{n}}$ idea presented previously.

We will assume that the four populations have equal variance, σ_y^2. Is there evidence to indicate real differences between the diets/treatments?

H_0: $\mu_A = \mu_B = \mu_C = \mu_D$.
H_A: At least one of the means differs from the others. In the following table, test order is listed in the parenthesis.

	DIET (Treatment)			
	A	B	C	D
	63 (12)	62 (20)	68 (16)	56 (23)
	67 (9)	60 (2)	66 (7)	62 (3)
	71 (15)	63 (11)	71 (1)	60 (6)
	64 (14)	59 (10)	67 (17)	61 (18)
	65 (4)		68 (13)	63 (22)
	66 (8)		68 (21)	64 (19)
				63 (5)
				59 (24)
	n_A =6	n_B = 4	n_C = 6	n_D = 8
Treat. Avg.	$\bar{Y}_A = 66$	$\bar{Y}_B = 61$	$\bar{Y}_C = 68$	$\bar{Y}_D = 61$
Grand Avg. 64				

Are the differences between the treatment averages greater than expected given the variation within treatments?

• Variation within treatments

$$s_A^2 = \frac{\sum_{j=1}^{N_A} (Y_{Aj} - \bar{Y}_A)^2}{n_a - 1} = \frac{S_A}{\nu_A} = \frac{\text{Sum of Squares within Treatment A}}{\text{Degrees of Freedom within Treatment A}}$$

$$s_A^2 = \frac{(63-66)^2 + (67-66)^2 + \ldots + (66-66)^2}{6-1} = \frac{S_A}{v_A} = \frac{40}{5} = 8$$

Similarly, $s_B^2 = \frac{S_B}{\nu_B} = \frac{10}{3} = 3.33$, $s_C^2 = \frac{S_C}{\nu_C} = \frac{14}{5} = 2.8$

and $s_D^2 = \frac{S_D}{\nu_D} = \frac{48}{7} = 6.857$.

Since s_A^2, s_B^2, s_C^2, and s_D^2 all provide estimates of the unknown variance, σ_y^2. They may be combined to provide a pooled estimate.

$$s_P^2 = \frac{\nu_A s_A^2 + \nu_B s_B^2 + \nu_C s_C^2 + \nu_D s_D^2}{\nu_A + \nu_B + \nu_C + \nu_D} = \frac{S_A + S_B + S_C + S_D}{\nu_A + \nu_B + \nu_C + \nu_D}$$

We'll refer to this sample variance as the within treatment sample variance.

$$s_w^2 = \frac{S_W}{\nu_W} = \frac{40 + 10 + 14 + 48}{5 + 3 + 5 + 7} = \frac{112}{20}$$

$$= \frac{\text{Within Treatment Sum of Squares}}{\text{Within Treatment Degree of Freedom}}$$

$s_w^2 = 5.6$. It is the within treatment mean square estimate of the true variance, σ_y^2.

One-Way ANOVA

In general,

$$\Sigma_{WithinTreatment} = \sum_{i=1}^{k} \sum_{j=1}^{n_I} (Y_{ij} - \bar{Y}_i)^2$$

$$s_w^2 = \frac{S_{WithinTreatment}}{\text{Total number of observations - Number of treatments}}$$

Variation Between Treatments

If there is no difference between treatment means, a second estimate of σ_y^2 can be obtained based on the variation of the treatment means about y.

If we calculate $\dfrac{\sum_{OverAllTreatments} (\bar{Y}_{treatment} - \bar{\bar{Y}})^2}{\#Treatments - 1}$, it can be used as a estimate of $\sigma_{\bar{y}}^2$.But we want an estimate of σ_y^2.

Since $\sigma_y^2 = n\sigma_{\bar{y}}^2$, let's scale $(\bar{y}_{treatment} - \bar{\bar{y}})^2$ by n_{treat}.

So, in general,

$$\frac{\sum_{i=1}^{k} n_i(\bar{Y}_{treatment} - \bar{\bar{Y}})^2}{k-1} = s_{BT}^2 = \frac{S_B}{\nu_B}.$$

where k is the number of treatments. It is the Between Treatment estimate of σ_y^2, or, the Between Treatment Mean Square.

For our Problem,

	Treatment			
	A	B	C	D
$\overline{Y}_i$	66	61	68	61
$\overline{\overline{Y}}$	64	64	64	64
$(\overline{Y}_i - \overline{\overline{Y}})$	2	-3	4	-3
n_i	6	4	6	8

$S_B = 6(2)^2 + 4(-3)^2 + 6(4)^2 + 8(-3)^2 = 228$, $\nu_B = 3$,

and $s_B^2 = \dfrac{S}{\nu_B} = \dfrac{228}{3} = 76.0$

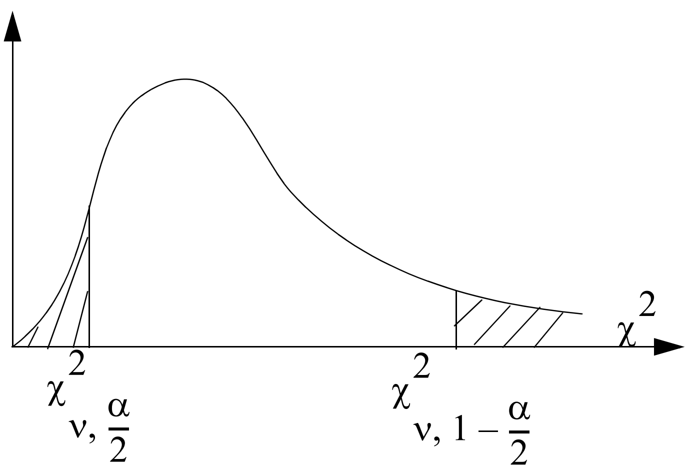
Summary:

Within Treatments:	Sum of Squares:	$S_W = 112$
	Degrees of Freedom:	$\nu_W = 20$
	Mean Square:	$s_W^2 = 5.6$
Between Treatments:	Sum of Square	$S_B = 228$
	Degrees of Freedom	$\nu_B = 3$
	Mean Square	$s_B^2 = 76.0$

Under H_0, both s_W^2 and s_B^2 are estimates of σ_y^2. If there is a difference between treatment means, the between treatment variance will be inflated. If a sample of size *n* is drawn from a normal population with variance σ^2, The quantity, $\dfrac{s^2(n-1)}{\sigma^2}$

follows the χ^2 distribution. If σ^2 is postulated, $\chi^2_{calc} = \frac{s^2(n-1)}{\sigma^2}$ can be computed and compared to the existing values. If σ^2 is not postulated, a 100(1-α)% confidence interval for σ^2 can be calculated.

$$\frac{s^2(n-1)}{\chi^2_{\nu,\,1-\frac{\alpha}{2}}} \le \sigma^2 \le \frac{s^2(n-1)}{\chi^2_{\nu,\,\frac{\alpha}{2}}}$$



• Sampling dist$^{\underline{n}}$ of Two Variances

If sample of size n_1 is drawn from normal dist$^{\underline{n}}$ with a variance of σ_1^2. and sample of size n_2 is drawn from normal dist$^{\underline{n}}$ with a variance of σ_2^2, estimates of the population variances may be calculated: s_1^2 and s_2^2.

The quantity $\frac{s_1^2}{\sigma_1^2}$ is $\frac{\chi^2_{\nu 1}}{\nu_1}$ distributed and $\frac{s_2^2}{\sigma_2^2}$ is $\frac{\chi^2_{\nu 2}}{\nu_2}$ distributed.

The ratio $\frac{\chi^2_{\nu 1}/\nu_1}{\chi^2_{\nu 2}/\nu_2}$ follows the F dist$^{\underline{n}}$ with ν_1, ν_2 degrees of freedom.

Therefore, $\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F_{\nu 1,\, \nu_2}$ or $\frac{s_1^2 \sigma_2^2}{s_2^2 \sigma_1^2} \sim F_{\nu 1,\, \nu 2}$

If $\sigma_1^2 = \sigma_2^2$, then $\frac{s_1^2}{s_2^2} \sim F_{\nu 1,\, \nu 2}$

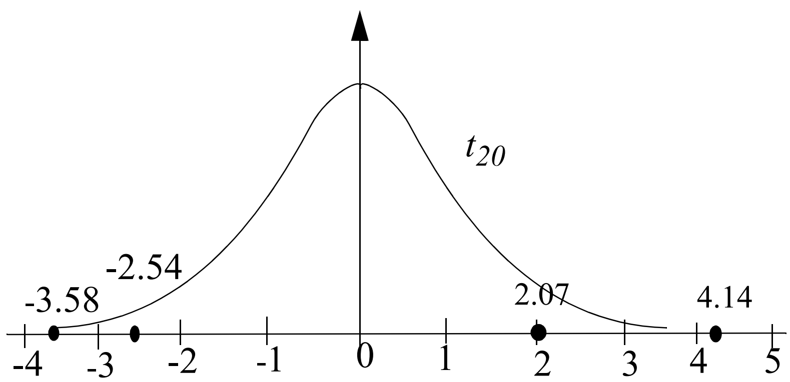
For the blood coagulation effect problem, $s_W^2 = 5.6$ (ν_w=20) and $s_B^2 = 76.0$ (ν_B=3).

Under H_0, both are estimates of σ_y^2, and $F_{calc} = s_B^2/s_W^2 = 13.57$. Note that this ratio should be distributed according to $F_{\nu_1 = 3,\, \nu_2 = 20}$ under H_0. From a F-dist$^{\underline{n}}$ table, it is found that F (3,20,.95) = 3.10 and F(3,20,.99) = 4.94. Since both numbers are less than F_{calc}, strong evidence shows that F_{calc} is not a typical value. This means that s_B^2 is inflated by differences between μ_A, μ_B, μ_C, and μ_D. Therefore, reject H_0. At least one of the means is different.

To confirm ANOVA results, we would like to display $\overline{Y}$'s on their reference dist$^{\underline{n}}$. It may be difficult, since each $\overline{Y}$ is based on different Degrees of Freedom ($\sigma_{\bar{y}}^2$ differs from each other).

$\overline{\overline{Y}} = 64$, $s_y^2 = 5.6$, s_y = 2.366. We need to calculate a t for each $\overline{Y}$.

	A	B	C	D
$\overline{Y}$	66	61	68	61
n	6	4	6	8
$s_{\overline{Y}}$	0.966	1.183	0.966	0.837
t_{calc}	2.070	-2.536	4.141	-3.584



The four calculated *t* points are plotted on a t_{20} distribution plot. From the plot, we can ask if these four t values be drawn randomly from a t_{20} dist^{n}.

Decomposition of the variance

The coagulation time for the *i*th treatment and *j*th trial within the treatment is Y_{ij} where *i* goes from 1 to k (k is the number of treatments/diets) and j goes from 1 to n_i (n_i is the number of trials for the ith treatment).
The coagulation time, Y_{ij}, may be expressed as:

$$Y_{ij} = \overline{\overline{Y}} + (\overline{Y}_i - \overline{\overline{Y}}) + (Y_{ij} - \overline{Y}_i)$$

Squaring both sides and summing over all trials and treatments, and after simplification,

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \bar{\bar{Y}}^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i} (\bar{Y}_i - \bar{\bar{Y}})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2 = N\bar{\bar{Y}}^2 + \sum_{i=1}^{k} n_i(\bar{Y}_i - \bar{\bar{Y}})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$$

$S_{TOT} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2$, is the Total Sum of Squares

$S_{AVG} = N\bar{\bar{Y}}^2$, is the Sum of Squares Due to the Mean.

$S_B = \sum_{i=1}^{k} n_i(\bar{Y}_i - \bar{\bar{Y}})^2$, is the Between Treatment Sum of Squares.

$S_W = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$, is the Within Treatment Sum of Squares.

For our blood coagulation time example, $S_{TOT} = 98644$, $S_{AVG} = 98304$, $S_B = 228$, $S_W = 112$.

We observe that in fact $S_{AVG} + S_B + S_W = S_{TOT}$

Additionally, $\nu_{TOT} = \nu_{AVG} + \nu_B + \nu_W$

(24) = (1) + (3) + (20)

What we know can be summarized in an ANOVA Table

Source of Var.	SS.	D.of F.	Mean Sqr.	F_{calc}
Average	98304	1	98304	17554.3
Between Treat.	228	3	76	13.57
Within Treat.	112	20	5.6	
Total	98644	24		

In the above table, the average mean square error (98304) is an estimate of σ_y^2 under the assumption that μ_y=0. The Between Treatment Mean Square Error (76) is an estimate of σ_y^2 under the assumption that $\mu_A = \mu_B = \mu_C = \mu_D$.

To test the hypothesis that $\mu_A = \mu_B = \mu_C = \mu_D$, compare Between/Within Treatment Mean Square

$\frac{s_B^2}{s_W^2} = 13.57 = F_{calc}$ to $F_{\nu_1 = 3, \nu_2 = 20}$. F(3,20,.95) = 3.10 and F (3,20,.99) = 4.94.

To test the hypothesis that $\mu_y = 0$, compare average and within treatment mean squares

$\frac{s_{AVG}^2}{s_W^2} = 17554.3$ to $F_{V_1 = 3, V_2 = 20}$. F (1,20,.95) = 4.35 and F (1,20,.99) = 8.10.

So, the average is significant, so as are the treatments.

Earlier, we expressed the response as:

$$Y_{ij} = \bar{\bar{Y}} + (\bar{Y}_i - \bar{\bar{Y}}) + (Y_{ij} - \bar{Y}_i)$$

where $\bar{\bar{Y}}$ is the grand mean. $(\bar{Y}_i - \bar{\bar{Y}})$ is treatment effect, and $(Y_{ij} - \bar{Y}_i)$ is experimental error.

$$y = \eta + \tau_i + \varepsilon$$

$\bar{\bar{y}}$ is an estimate of η. A model for the response is:

$$\hat{y} = \hat{\eta} + \hat{\tau}_i$$

In blood coagulation experiment, we studied one variable, diet. We will now look at two variables (1 blocking variables and 1 treatment variable or 2 blocking variables).