One-Way Analysis of Variance

- With regression, we related two quantitative, typically continuous variables.

- Often we wish to relate a quantitative response variable with a qualitative (or simply discrete) independent variable, also called a factor.

- In particular, we wish to compare the mean response value at several levels of the discrete independent variable.
  (or categorical

Example: We wish to compare the mean wage of farm laborers for 3 different races (black, white, Hispanic). Is there a difference in true mean wage among the ethnic groups?

- If there were only 2 levels, could do a: 2-sample t-test

- For 3 or more levels, must use the Analysis of Variance (ANOVA).

- The Analysis of Variance tests whether the means of t populations are equal. We test:

$H_0: \mu_1 = \mu_2 = \cdots = \mu_t$

H_a: At least one equality is not satisfied
(at least two population means differ)

• Suppose we have $t = 4$ populations. Why not test:

$H_0: \mu_1 = \mu_2$, $H_0: \mu_1 = \mu_3$, $H_0: \mu_1 = \mu_4$,
$H_0: \mu_2 = \mu_3$, $H_0: \mu_2 = \mu_4$, $H_0: \mu_3 = \mu_4$

with a series of t-tests?

• If each test has $\alpha = .05$, probability of correctly failing to reject H_0 in all 6 tests (when all nulls are true) is: $(.95)^6 = .735$

→ Actual significance level of the procedure is 0.265, not 0.05 → We will make <u>some</u> Type I error with probability 0.265 if all 4 means are truly equal.

<u>Why Analyze Variances to Compare Means?</u>

• Look at Figure 6.1, page 248.

Case I and Case II: Both have independent samples from 3 populations.

• The positions of the 3 sample means are the same in each case.

• In which case would we conclude a definite difference among <u>population means</u> μ_1, μ_2, μ_3?

Case I? <u>Yes</u>. Variance <u>between</u> <u>sample</u> <u>means</u> is <u>large</u> relative to variance <u>within</u> samples.

Case II? No. Variance <u>between</u> <u>sample</u> <u>means</u> is <u>small</u> relative to variance <u>within</u> samples.

• This comparison of variances is at the heart of ANOVA.

Assumptions for the ANOVA test:

(1) There are t independent samples taken from t populations having means $\mu_1, \mu_2, \ldots, \mu_t$.
(2) Each population has the same variance, σ^2.
(3) Each population has a normal distribution.

• The data (observed values of the response variable) are denoted: Y_{ij} , $i = 1, \ldots, t$ ← which sample

$j = 1, \ldots, n_i$ ← which observation within the sample

• Each sample has size n_i, for a total of $\sum_{i=1}^{t} n_i$ observations.

Example: Y_{47} = 7th observation in the 4th sample

Notation

The i-th level's total: $Y_{i\bullet}$ (sum over j)
The i-th level's mean: $\bar{Y}_{i\bullet} = Y_{i\bullet} / n_i$
The overall total: $Y_{\bullet\bullet}$ (sum over i and j)
The overall mean: $\bar{Y}_{\bullet\bullet} = Y_{\bullet\bullet} / \sum n_i$

Estimating the variance σ^2

- **For $i = 1, \ldots, t$, the sum of squares for each level is**

$$SS_i = \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 = \sum_{j=1}^{n_i} Y_{ij}^2 - \frac{(Y_{i\circ})^2}{n_i}$$

- **Adding all the SS_i's gives the pooled sum of squares:**

$$SS_p = \sum_i SS_i$$

- **Dividing by our degrees of freedom gives our estimate of σ^2:**

$$S_p^2 = \frac{SS_p}{(\sum n_i) - t} = \frac{\sum_i (n_i - 1) S_i^2}{\sum n_i - t}$$

- **Recall: For 2-sample t-test, pooled sample variance was:**

$$S_p^2 = \frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2}$$

(same formula as ANOVA S_p^2, just for $t=2$).

- **This is the correct estimate of σ^2 if all t populations have equal variances.**

- **We will have to check this assumption.**

Development of ANOVA F-test

- **Assume sample sizes all equal to n:**
$n_1 = n_2 = \ldots = n_t$ (= n) ← balanced data

- **Suppose H_0: $\mu_1 = \mu_2 = \ldots = \mu_t$ (= μ) is true.**

- **Then each sample mean $\overline{Y}_{i\bullet}$ has mean** μ **and variance** σ^2/n

- **Treat these group sample means as the "data" and treat the overall sample mean as the "mean" of the group means. Then an estimate of σ^2 / n is:**

$$S^2_{means} = \frac{\sum_i (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2}{t-1}$$

$\Rightarrow$ $n\,S^2_{means}$ is an estimate of σ^2.

Recall: S_p^2 was another estimate of σ^2 (independent of $n S^2_{means}$) when the populations are normal.

Consider the statistic:

$$F^* = \frac{n\,S^2_{means}}{S_p^2}$$

estimate of σ^2 if H_0 is true

estimate of σ^2

• **With normal data, the ratio of two independent estimates of a common variance has an F-distribution.**

→ **If H_0 true, we expect F* has an F-distribution.**

(This F* ratio should be "near" 1 if H_0 true)

• **If H_0 false ($\mu_1, \mu_2, \ldots, \mu_t$ not all equal), the sample means should be <u>more spread out</u>.**

→ $n s^2_{means}$ should be larger than under H_0.

→ F* ratio should be bigger than 1 if H_0 false

<u>General ANOVA Formulas (Balanced or Unbalanced)</u>

• **We want to compare the variance <u>between (among) the sample means</u> with the variance <u>within the different groups</u>.**

• **Variance between group means measured by:**

$$SSB = \sum_{i=1}^{t} \frac{Y_{i\cdot}^2}{n_i} - \frac{Y_{\cdot\cdot}^2}{\sum n_i}$$

and, after dividing by the "between groups" degrees of freedom,

$$MSB = \frac{SSB}{t-1}$$

(analogous to $n s^2_{means}$)

"between-groups mean square"

- **Variance within groups measured by:**

$$SSW = \sum_i \sum_j Y_{ij}^2 - \sum_i \frac{Y_{i\cdot}^2}{n_i}$$

also called SSE

and, after dividing by the "within groups" degrees of freedom,

$$MSW = \frac{SSW}{\sum n_i - t}$$

(analogous to S_p^2)

"within groups" mean square

- **In general, our F-ratio is:** $F^* = \dfrac{MSB}{MSW}$

- **Under H_0, F^* has an F-distribution with:**

$$df = (t-1, \sum n_i - t)$$

- **The total sum of squares for the data:**

$$TSS = \sum_i \sum_j (Y_{ij} - \bar{Y}_{\cdot\cdot})^2$$

can be partitioned into

$$TSS = SSB + SSW$$

- **The degrees of freedom are also partitioned:**

Total df = "Between groups" df + "Within groups" df

$$(\sum n_i - 1) = (t-1) + (\sum n_i - t)$$

• **This can be summarized in the ANOVA table:**

Source	df	SS	MS	F*
Between	$t-1$	SSB	MSB	MSB/MSW
Within	$\sum n_i - t$	SSW	MSW	
Total	$\sum n_i - 1$	TSS		

Example: Table 6.4 (p. 253) gives yields (in pounds/acre) for 4 different varieties of rice (4 observations for each variety)

$t = 4$ levels

$n_1 = n_2 = n_3 = n_4 = 4 \Rightarrow \sum n_i = 16$

$$\sum_i \frac{Y_{i\bullet}^2}{n_i} = \frac{3938^2}{4} + \frac{3713^2}{4} + \frac{3754^2}{4} + \frac{4466^2}{4}$$

$$= 15,832,971.25$$

$$\frac{Y_{\bullet\bullet}^2}{\sum n_i} = \frac{15871^2}{16} = 15,743,040.06$$

SSB = $15832971.25 - 15743040.06$

$= 89931.2$

$$\sum Y_{ij}^2 = 934^2 + 1041^2 + \cdots + 1140^2 + 1191^2$$
$$= 15{,}882{,}847$$

SSW = $15882847 - 15832971.25$

$= 49875.75$

$MSB = 89931.2/3 = 29977.07$, $MSW = \frac{49875.75}{12} = 4156.31$

ANOVA table for Rice Data:

Source	df	SS	MS	F*
Between	3	89931.2	29977.07	7.21
Within	12	49875.75	4156.31	
Total	15	139806.95		

$F^* = 29977.07/4156.31 = 7.21$

- **Back to original question: Do the four rice varieties have equal population mean yields or not?**

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
H_a: At least one equality is not true

Test statistic: $F^* = 7.21$

At $\alpha = 0.05$, compare to: $F_{.05,3,12} = 3.49$ from Table A.4A, p. 731

Conclusion:

If $F^* > F_\alpha$ we reject H_0. $7.21 > 3.49$, so reject H_0. We have sufficient evidence to conclude a difference among mean yields for the 4 varieties.

P-value = .005 from SAS

"Treatment Effects" Linear Model:

Our ANOVA model equation:

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad i=1,\ldots,t, \quad j=1,\ldots,n_i$$

Y_{ij} = j-th response value from i-th sample

μ_i = mean of population i

ε_{ij} = random error term

Denote the *i*-th "treatment effect" by:

$$\tau_i = \mu_i - \mu$$

"overall mean"

- **The ANOVA model can now be written as:**

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \quad \sum_i \tau_i = 0$$

- **Note that our ANOVA test of:**

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_t$$

is the same as testing:

$H_0: \tau_i = 0$ for all i

vs. $H_a: \tau_i \neq 0$ for some i

Note: For balanced data,

E(MSB) = $\sigma^2 + \frac{n}{t-1}\sum(\tau_i^2)$ **and E(MSW) =** σ^2

If H_0 is true (all $\tau_i = 0$): MSB and MSW should be approx. equal (their ratio ≈ 1)

If H_0 is false:

MSB should be somewhat greater than MSW.