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.**

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:**

● **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:**

● Suppose we have *t* = 4 populations. Why not test:

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:

→ 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?

Case II?

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

<u>Assumptions for the ANOVA test:</u>

(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:

● Each sample has size n_i, for a total of observations.

Example: Y_{47} =

<u>Notation</u>

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

Estimating the variance σ^2

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

$SS_i =$

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

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

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

● **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 $\bar{Y}_{i\bullet}$ has mean and variance**

● **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:**

Recall:

Consider the statistic:

● 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.

● If H_0 false ($\mu_1, \mu_2, \ldots, \mu_t$ not all equal), the sample means should be more spread out.

→

→

General ANOVA Formulas (Balanced or Unbalanced)

● We want to compare the variance between (among) the sample means with the variance within the different groups.

● Variance between group means measured by:

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

● Variance within groups measured by:

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

● In general, our F-ratio is:

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

● The total sum of squares for the data:

can be partitioned into

● The degrees of freedom are also partitioned:

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

Source	df	SS	MS	F

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

$$\sum_i \frac{Y_{i\bullet}^2}{n_i} =$$

$$\frac{Y_{\bullet\bullet}^2}{\sum n_i} =$$

SSB =

$$\sum Y_{ij}^2 =$$

SSW =

ANOVA table for Rice Data:

● 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:

At $\alpha = 0.05$, compare to:

Conclusion:

"Treatment Effects" Linear Model:

Our ANOVA model equation:

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

● The ANOVA model can now be written as:

● Note that our ANOVA test of:

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

is the same as testing:

Note: For balanced data,

E(MSB) = **and E(MSW) =**

If H_0 is true (all $\tau_i = 0$):

If H_0 is false: