Mathematics 231

Lecture 28

Liam O'Brien

Announcements

- Reading
  - Today | M&M 8.2 | 505-515

Today	M&M 8.2	505-515
	M&M 12.0	637
Next class	M&M 12.1	638-655

Topics

- Hypothesis testing for comparing two proportions
- Introduction to ANOVA

Comparison of Two Population Proportions

- Before, we considered the comparison of the population proportion to some null value, p_0.
- However, in studies we want to compare the proportions of "successes" in two populations p_1 and p_2.
- Example: Proportion of successes in a treatment vs. control group, among males vs. females.
- Ordinarily, we want to know if p_1 and p_2 are identical (suspecting they are not).

Comparison of Two Population Proportions

- Given SRS's from the two populations or groups, p_1 and p_2 can be estimated by their respective sample proportions.
- Question: Is the difference in sample proportions so large that it is unlikely to be due to chance alone.
- To answer this question, we consider the difference between the two sample proportions:

$$\hat{p}_1 - \hat{p}_2$$

Comparison of Two Population Proportions

When both sample sizes, n_1 and n_2, are suffciently large the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal.

$n_1\hat{p}(1-\hat{p}) \geq 10$ and $n_2\hat{p}(1-\hat{p}) \geq 10$

A test of $\mathrm{H}_0 : p_1 = p_2$ against $\mathrm{H}_A : p_1 \neq p_2$ can be based on the following statistic,

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

where $\hat{p} = \dfrac{x_1 + x_2}{n_1 + n_2}$ is the pooled estimate of p.

CI for Two Population Proportions

When both sample sizes, n_1 and n_2, are suffciently large the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal.

A confidence interval for the population proportion difference is given by:

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Example: Binge Drinking

- A survey of 17,096 college students at 4-year colleges in the U.S. was conducted in 2000. each student was asked whether or not they participated in frequent binge drinking.
- Are men and women college students equally likely to participate in the behavior?
- Men: n = 7,180; p-hat = 0.227
- Women: n = 9,916; p-hat = 0.170
- Calculate a 95% CI for the difference in means.

Example: Binge Drinking

$$n_M \hat{p}_M = 1630 > 10; \quad n_M(1-\hat{p}_M) = 5550 > 10$$

$$n_W \hat{p}_W = 1684 > 10; \quad n_W(1-\hat{p}_W) = 8232 > 10$$

A 95% CI is thus given by,

$$\hat{p}_M - \hat{p}_W \pm z^* \sqrt{\frac{\hat{p}_M(1-\hat{p}_M)}{n_M} + \frac{\hat{p}_W(1-\hat{p}_W)}{n_W}}$$

$$0.227 - 0.170 \pm 1.96(0.00622) = (0.0448, 0.0692)$$

What does this tell us about the difference between the proportions of men and women college students who frequently binge drink?

Example: Hypertension

- A major study of the effect of hypertension on risk of heart attack was performed.
- Data were collected from 3338 men with high blood pressure and 2676 men with low blood pressure.
- These men were followed over a period of time and 21 in the LBP group died of heart disease and 55 in the HBP group died of heart disease.
- Find a 95% CI for the difference between these two proportions.

Example: Hypertension

Check validity assumptions:

They all hold.

The 95% CI is given by,

$$\hat{\mathrm{p}}_{\mathrm{HBP}} - \hat{p}_{LBP} \pm z^{*} \sqrt{\frac{\hat{p}_{HBP}(1-\hat{p}_{HBP})}{n_{HBP}} + \frac{\hat{p}_{LBP}(1-\hat{p}_{LBP})}{n_{LBP}}}$$

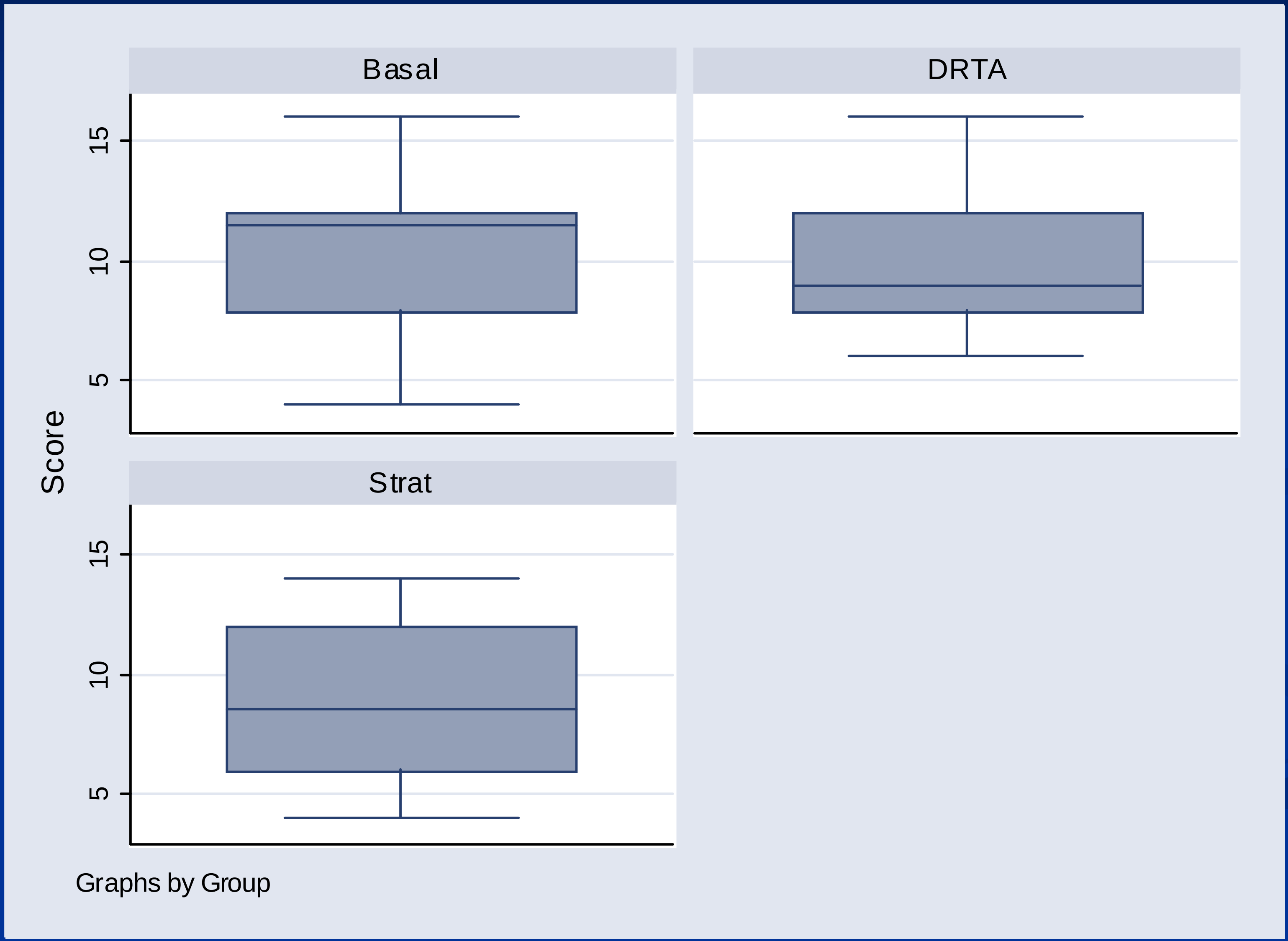
$$0.0165 - 0.0078 \pm 1.96(0.00278) = (0.00324, 0.0142)$$

What can we say about the relative risk of death from heart disease comparing the HBP and LBP groups?

Analysis of Variance (ANOVA)

- Comparison of the means of *K* independent groups.
- Populations are assumed to be normal with equal variances, $\sigma_1 = \sigma_2 = \dots \sigma_K = \sigma$.
- We obtain SRSs of size n_i from the population with mean μ_i and standard deviation σ_i, $i=1,2,\dots K$.
- We want to test the null hypothesis H_0: $\mu_1 = \mu_2 = \dots \mu_K$ against the alternative that at least one of these means differs from the others.

Example: Reading Scores



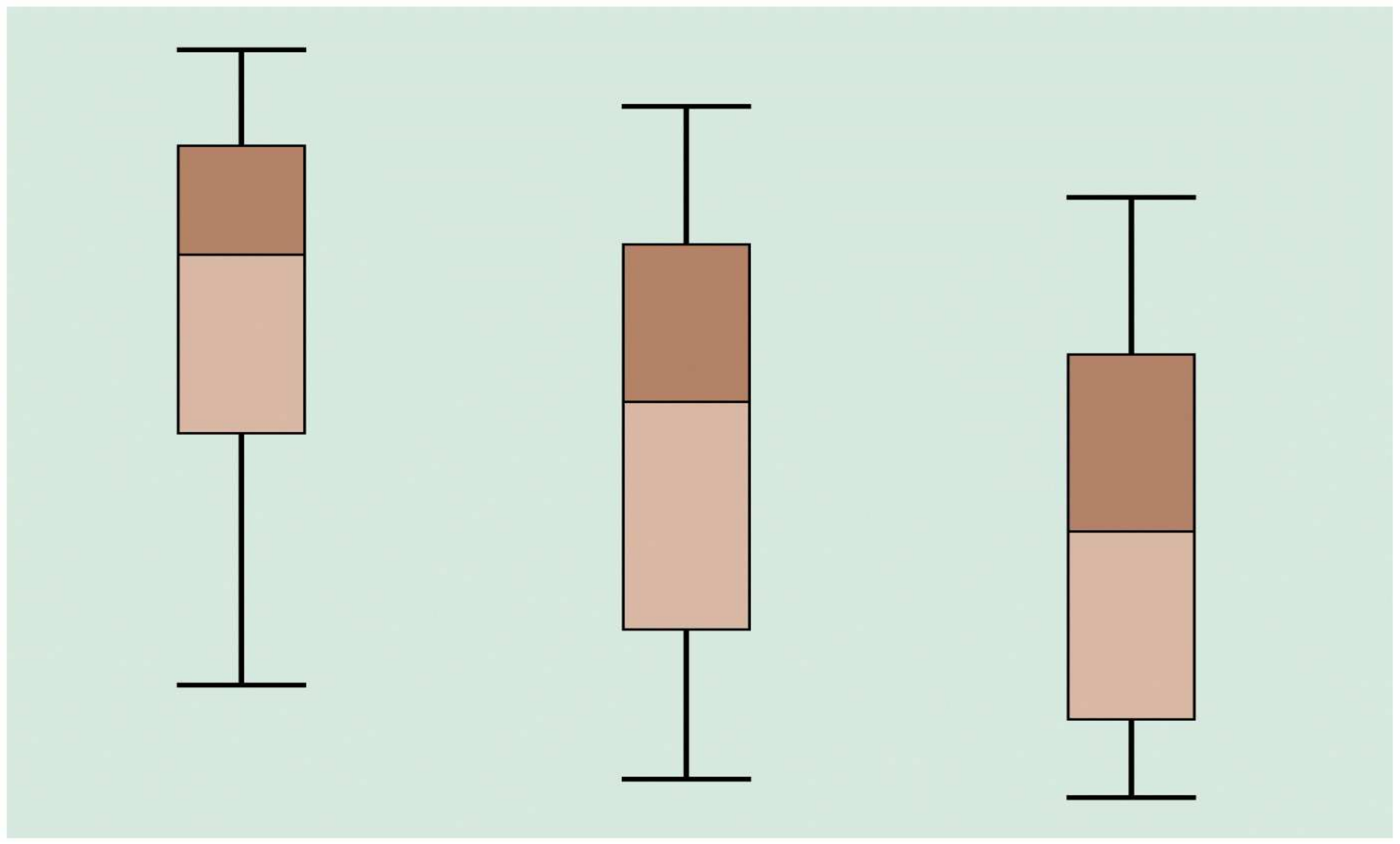
Analysis of Variance (ANOVA)

- With *K* populations there are two types of variability:

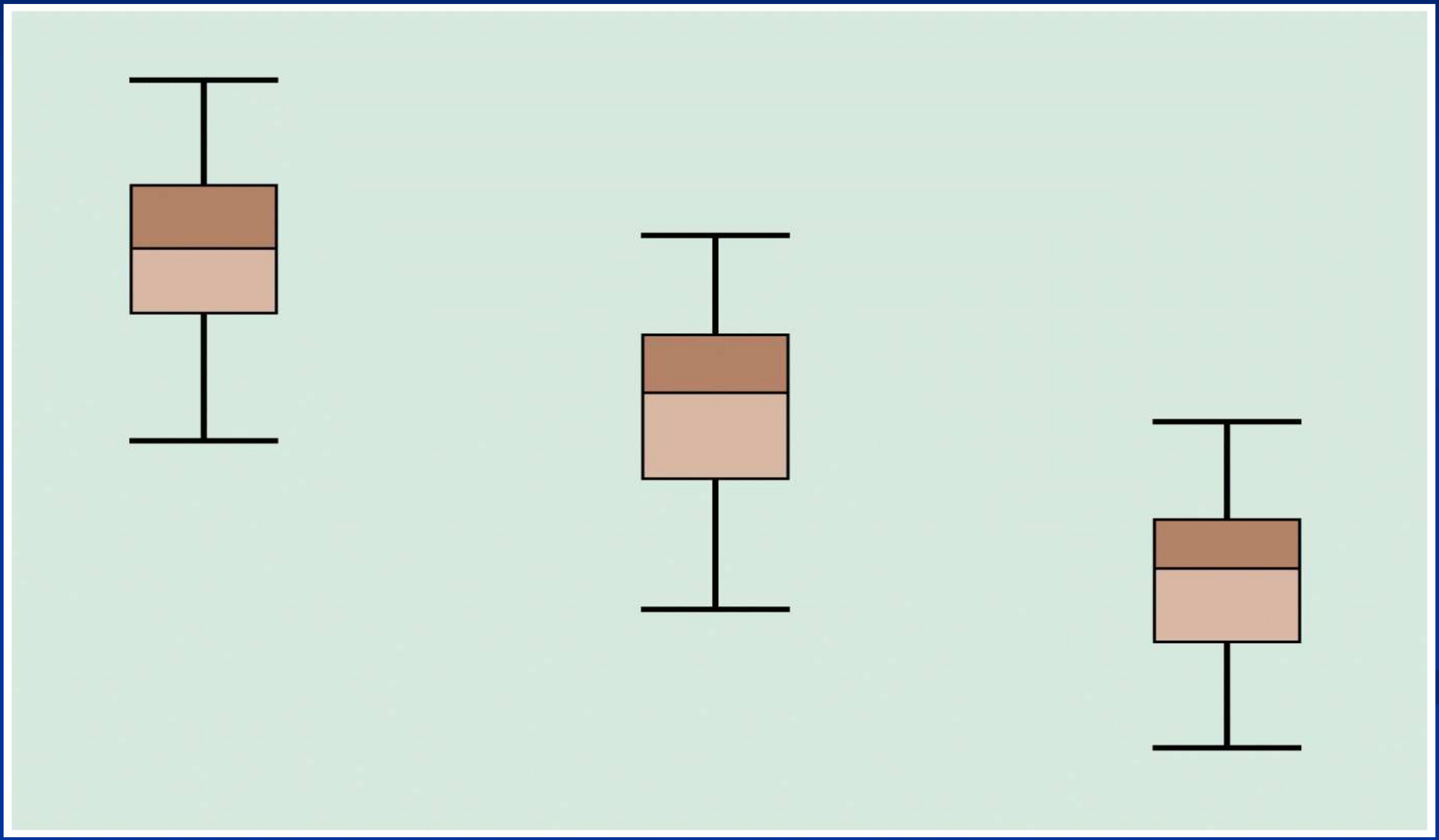
1. Variation of individual values around their group means (variation within groups).
2. Variation of group means around the overall mean (variation between groups).

- Main idea: If (i) is small relative to (ii), this implies the group (or population) means are different.
- ANOVA determines whether variability in data is mainly from variation within groups, or variation between groups.

Variation Within Versus Between Groups



Variation Within Versus Between Groups



Variance Within Groups

- From the assumption of homoscedasticity, we have that $s_1, s_2,\ldots,s_K$ all estimate σ, the common value of the sd in each of the *K* groups (or populations).
- As a result, we can combine them to obtain a better estimate of σ.
- The combined, or pooled, estimate of σ^2 is called the variance **within** groups.

Variance Within Groups

Pooled estimate of σ^2, the variance within groups,

$$s_P^2 = MSE = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_K - 1)s_K^2}{(n_1 - 1) + (n_2 - 1) + \cdots (n_K - 1)}$$

This is an extension of the pooled estimate of σ^2 used for the two-sample t-test.

The MSE refers to the within groups estimate variance

MSE is also known as the "mean square (MS) error."

Variance Between Groups

If the null hypothesis $\mathrm{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_K$ is in fact true, then it is as if we are sampling K times from the same population, with mean μ and SD σ.

From sampling distribution of sample mean, we can regard $\bar{\mathrm{x}}_1$ as an observation from a population with mean μ and SD $\sigma / \sqrt{\mathrm{n}_1}$

$\bar{x}_2$ as an observation from a population with mean μ and SD $\sigma / \sqrt{\mathrm{n}_2}$ and so on.

Variance Between Groups

So we can get a better estimate of μ using

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \cdots + n_K\bar{x}_K}{n_1 + n_2 + \cdots + n_K} = \frac{\sum_{all\, i} x_i}{n}$$

where $n = n_1 + n_2 + \cdots + n_K$

Variance Between Groups

Another estimate of σ^2 is the between groups estimate

$$s_b^2 = MSTr = \frac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + \cdots + n_K(\bar{x}_K - \bar{x})^2}{K-1}$$

This is an estimate of the variation of the group means around the overall mean; s_b^2 is also known as the "mean square (MS) between groups" or "mean square for treatments."

Note: The between groups estimate of σ^2 is valid only if $\mathrm{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_K$ is true.

Variance Between Groups

If $\text{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_K$ is true, s_w and s_b both estimate σ and should be of similar magnitude.

Therefore a test of $\text{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_K$ can be based on a comparison of the within groups and between groups estimates of the variability.

If $\text{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_K$ is not true, the between groups estimate of σ^2 will, in general, be larger than the within groups estimate of σ^2.

The F Statistic

Question: Do sample means vary around the overall mean more than the individual observations vary around the sample means?

To evaluate $\mathrm{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_K$ we use the test statistic,

$$F = \frac{MSTr}{MSE} = \frac{\text{between groups MS}}{\text{within groups MS}}.$$

The null hypothesis will be rejected if F is large.

The F Statistic

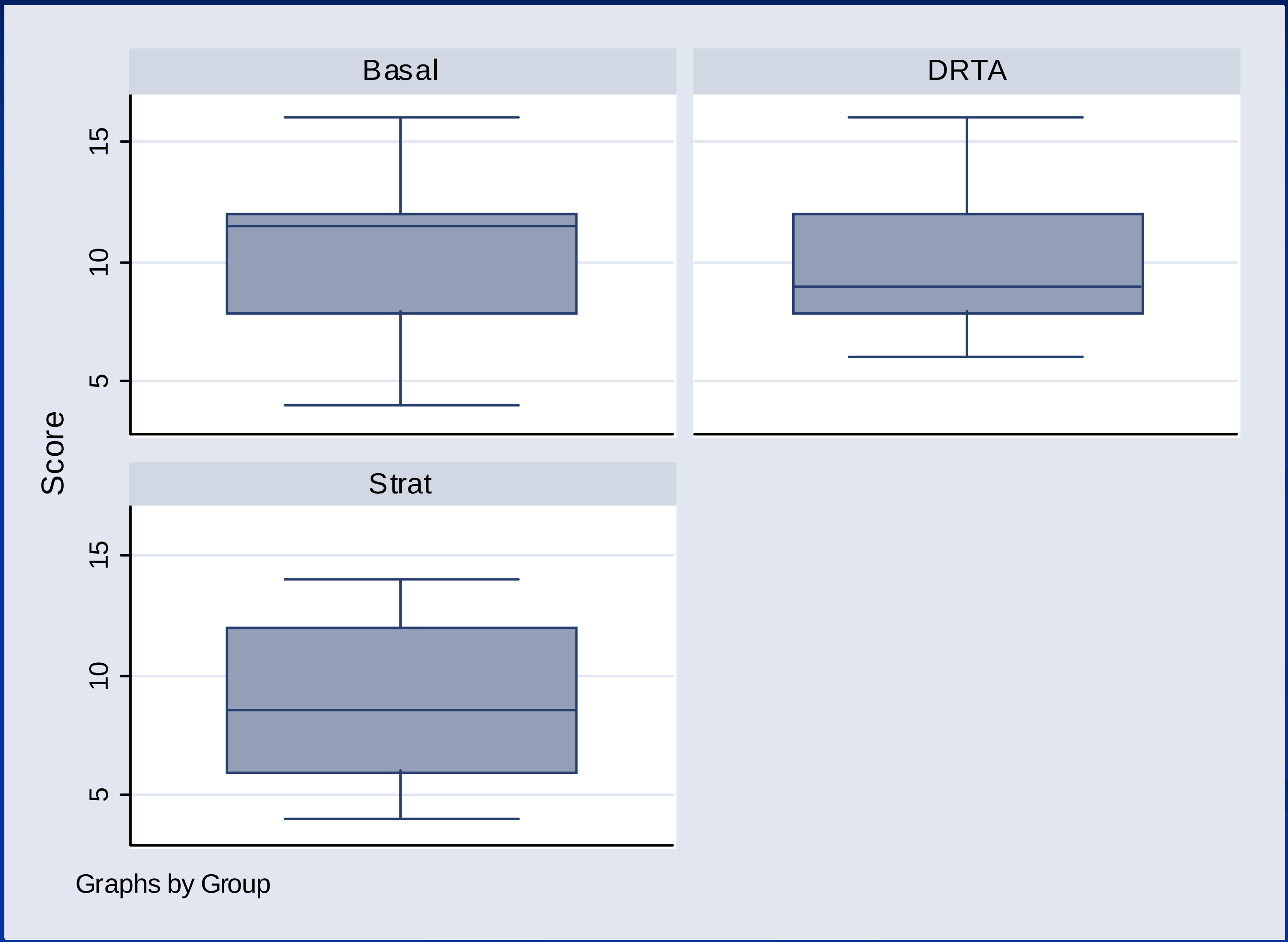
Under $H_0: \mu_1 = \mu_2 = \cdots = \mu_K$, the F statistic has an F distribution with K-1 and n-K degrees of freedom (where $n = n_1 + n_2 + \cdots + n_K$)

Note: df corresond to numerator and denominator of F.

F distribution cannot assume negative values and is skewed to the right.

Its shape depends on the degrees of freedom.

Example: Reading Scores



Example: Reading Scores

Group	Mean	Std. Dev.	Freq.
	Summary of Score		
Basal	10.5	2.9720924	22
DRTA	9.7272727	2.6935871	22
Strat	9.1363636	3.3423039	22
Total	9.7878788	3.0205203	66

Example: Reading Scores

```
. oneway score group

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      20.5757576      2   10.2878788      1.13     0.3288
 Within groups      572.454545     63   9.08658009
------------------------------------------------------------------------
    Total           593.030303     65   9.12354312

Bartlett's test for equal variances:  chi2(2) =   0.9623  Prob>chi2 =
   0.618
```

Multiple Comparisons: Bonferroni

- Suppose we wish to perform all possible pairs of comparisons among K groups.

There are $\binom{K}{2} = \dfrac{K!}{2!(K-2)!} = \dfrac{K(K-1)}{2}$ such comparison.

To protect against the overall level of α, we must perform each individual test at level,

$$\alpha^* = \frac{\alpha}{\binom{K}{2}}.$$

Multiple Comparisons: Bonferroni

- Suppose we wish to perform all possible pairs of comparisons among K groups.

  If K=3 (e.g., groups 1, 2, and 3) then there are

  $\binom{K}{2} = \frac{3!}{2!(3-2)!} = 3$ possible pairwise comparisons.

  Group 1 versus group 2

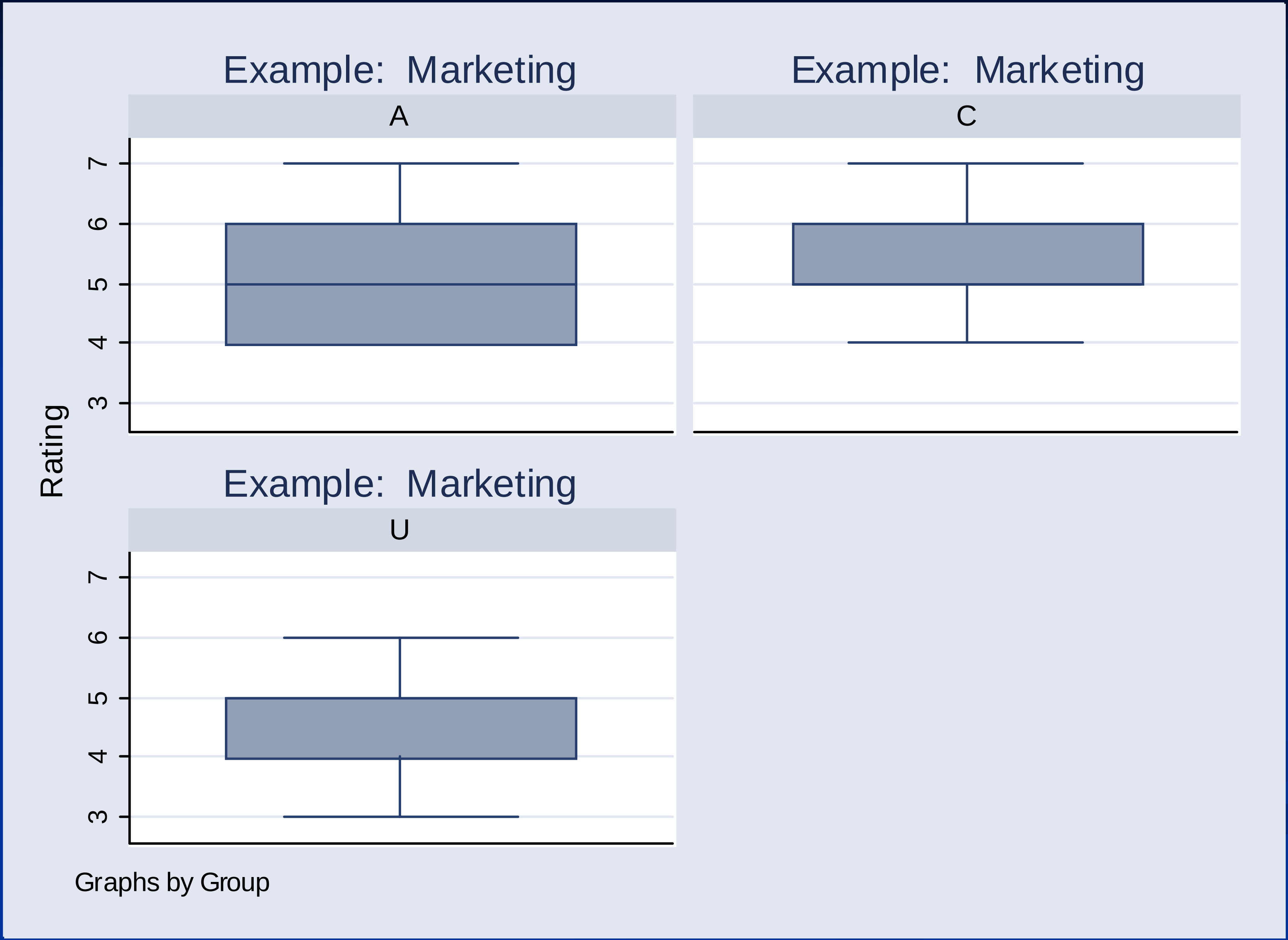
  Group 1 versus group 3

  Group 2 versis group 3

  If you want an overall 0.05 level, then do each of the three tests at the 0.05/3 = 0.0167 level.

ANOVA/Bonferroni in Stata

- To do a one-way ANOVA (1 group variable) in Stata, click on **Statistics > Linear Models and related > ANOVA/ MANOVA > one way analysis of variance.**
- Enter the "response" variable and the "group" variable.
- Then click the "Bonferroni" box under the Multiple Comparisons heading.
- Note: There are many multiple comparisons procedures, but Bonferroni is the most conservative.


Example: Marketing
A
Example: Marketing
C
Example: Marketing
U
Rating
3
4
5
6
7
Graphs by Group

Example: Marketing

```
. oneway rating group, bonferroni tabulate

            |          Summary of Rating
      Group |        Mean   Std. Dev.       Freq.
------------+------------------------------------
          A |   5.0555556    .8261596          36
          C |   5.4166667   .87423436          36
          U |   4.5090909   .69048365          55
------------+------------------------------------
      Total |   4.9212598    .8692845         127

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      18.828255      2   9.4141275     15.28     0.0000
 Within groups     76.3843434    124   .61600277
------------------------------------------------------------------------
    Total          95.2125984    126  .755655543

Bartlett's test for equal variances:  chi2(2) =   2.6669  Prob>chi2 = 0.264
```

Example: Marketing

```
Comparison of Rating by Group
                                (Bonferroni)
Row Mean-|
Col Mean |          A           C
---------+----------------------
       C |    .361111
         |      0.160
         |
       U |   -.546465   -.907576
         |      0.004      0.000
```

Bonferroni in Stata

- Stata does all possible pairwise comparisons and reports p-values that are **already** corrected for the fact you have done [K*(K-1)]/2 comparisons.
- That is, it constructs a series of two-sample t-tests using S_w as an estimate of σ, and multiplies the usual p-values by [K*(K-1)]/2.
- So you compare the corrected p-values to 0.05.