Lesson 13

ANOVA

Analysis Of Variance

Course Overview

So far in this course we've covered:

- Descriptive statistics
  - Summary statistics
  - Tables and Graphs
- Probability
  - Probability Rules
  - Probability Distributions
  - Sampling distributions and CLT
- Statistical Inference
  - Estimating parameters, confidence intervals of means and proportions
  - Hypothesis tests of means and proportions

Course Overview

Statistical Inference still to cover

- ANOVA: Lesson 13
  - Comparing means between three or more groups
- Inference about Odds Ratios and Relative Risks: Lesson 14
  - Confidence Intervals of the Odds Ratio and Relative Risk
- Simple Linear Regression: Lesson 15
  - Describing the linear relationship between two variables

ANOVA

- Many studies involve comparisons between more than two groups of subjects.
- If the outcome is continuous, ANOVA can be used to compare the means between groups.
- ANOVA is an abbreviation for the full name of the method: ANalysis Of Variance
- The test for ANOVA is the ANOVA F-test

ANOVA was developed in the 1920's by R.A. Fisher

He was described by Anders Hald as "a genius who almost single-handedly created the foundations for modern statistical science"
The ANOVA F-test is named for Fisher

Why use ANOVA instead of multiple t-tests?

- If you are comparing means between more than two groups, why not just do several two sample t-tests to compare the mean from one group with the mean from each of the other groups?

- One problem with this approach is the increasing number of tests as the number of groups increases

The problem with multiple t-tests

- The probability of making a Type I error increases as the number of tests increase.

- If the probability of a Type I error for the analysis is set at 0.05 and 10 t-tests are done, the overall probability of a Type I error for the set of tests = $1 – (0.95)^{10} = 0.40$* instead of 0.05

*Note: The formula for calculating overall Type I error on page 164 in text is incorrect.

Multiple Comparisons Problem

- Another way to describe the multiple comparisons problem is to think about the meaning of an alpha level = 0.05
- Alpha of 0.05 implies that, by chance, there will be one Type I error in every 20 tests: 1/20 = 0.05. This means that, by chance the null hypothesis will be incorrectly rejected once in every 20 tests
- As the number of tests increases, the probability of finding a ‘significant’ result by chance increases.

ANOVA: a single test for simultaneous comparison

- The advantage of using ANOVA over multiple t-tests is that the ANOVA F-test will identify if *any* of the group means are significantly different from at least one of the other group means with a *single* test.
- If the significance level is set at α, the probability of a Type I error for the ANOVA F-test = α *regardless* of the number of groups being compared.

ANOVA Hypotheses

- The Null hypothesis for ANOVA is that the means for all groups are equal:

$$H_o : \mu_1 = \mu_2 = \mu_3 = ... = \mu_k$$

- The Alternative hypothesis for ANOVA is that *at least* two of the means are not equal.
  - Rejecting the null hypothesis doesn't require that ALL means are significantly different from each other.
  - If **at least** two of the means are significantly different the null hypothesis is rejected

Analysis of Variance

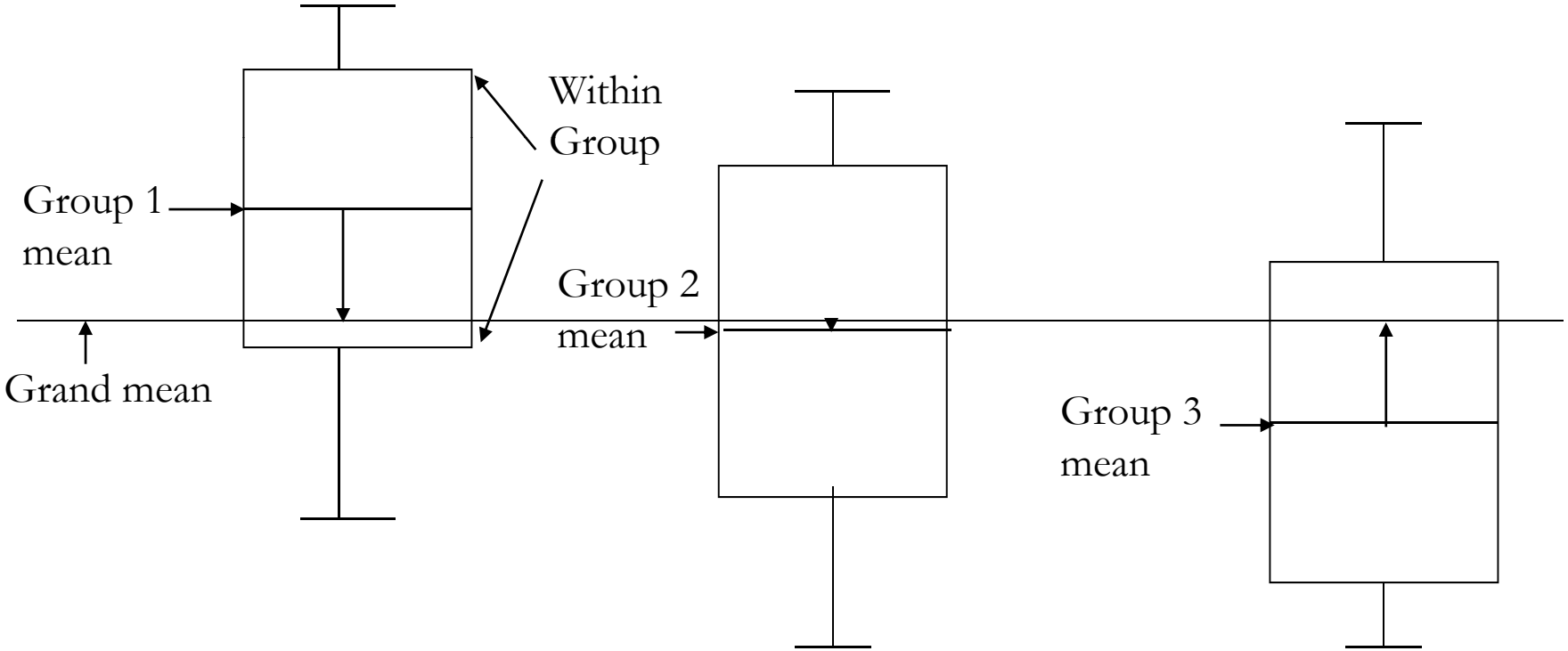
- ANOVA is used to compare *means* between three or more groups, so why is it called Analysis of *VARIANCE?*
- ANOVA is based on a comparison of two different sources of variation:
  - The average variation of observations within each group around the group mean
  - The average variation of the group means around the grand mean
    - The grand mean is the mean of all observations

Variability Among and Within Groups

- The average variation of observations within each group around the group mean is called the within group variability
- The average variation of the group means around the grand mean is called the among group variability
- The ANOVA F-test statistic is the ratio of the average variability among groups to the average variability within groups.

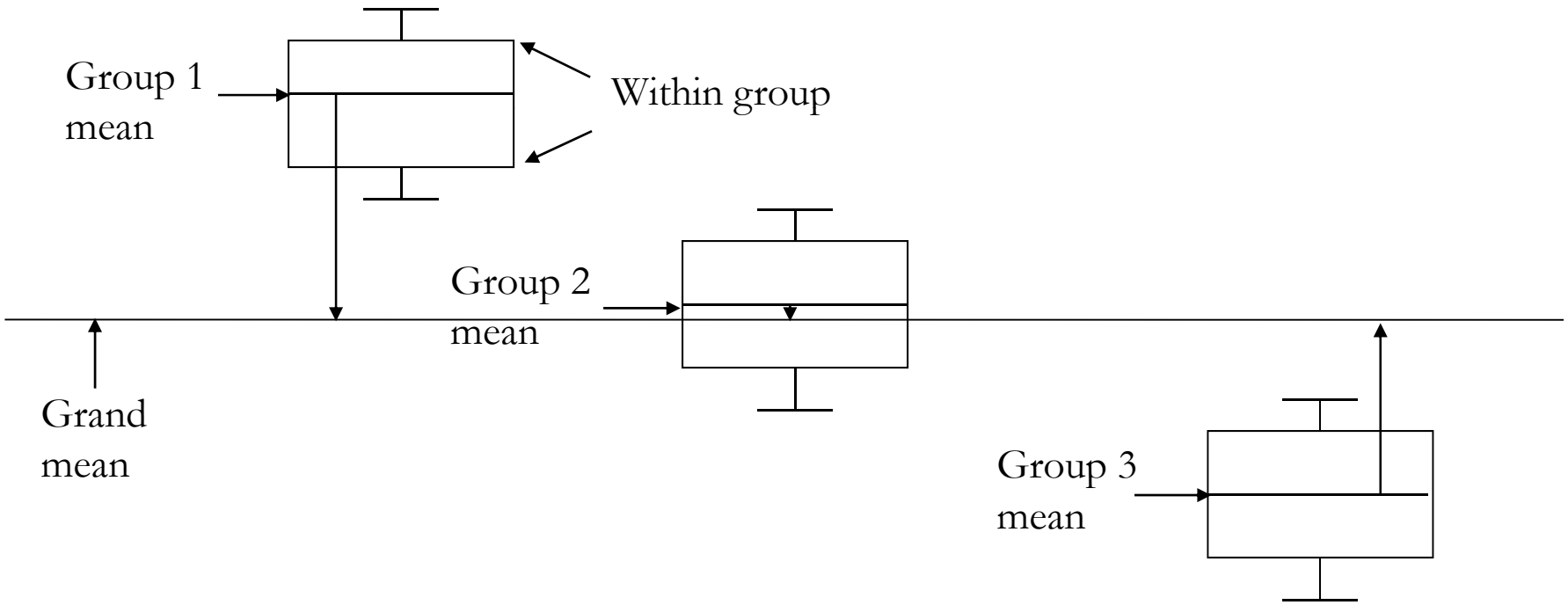
  - **F = $\frac{\text{average variability among groups}}{\text{average variability within groups}}$**

Small variability among groups compared to variability within groups: small F-statistic



Source: Introduction to the Practice of Statistics, Moore and McCabe

Large variability among groups relative to variability within groups: large F-statistic

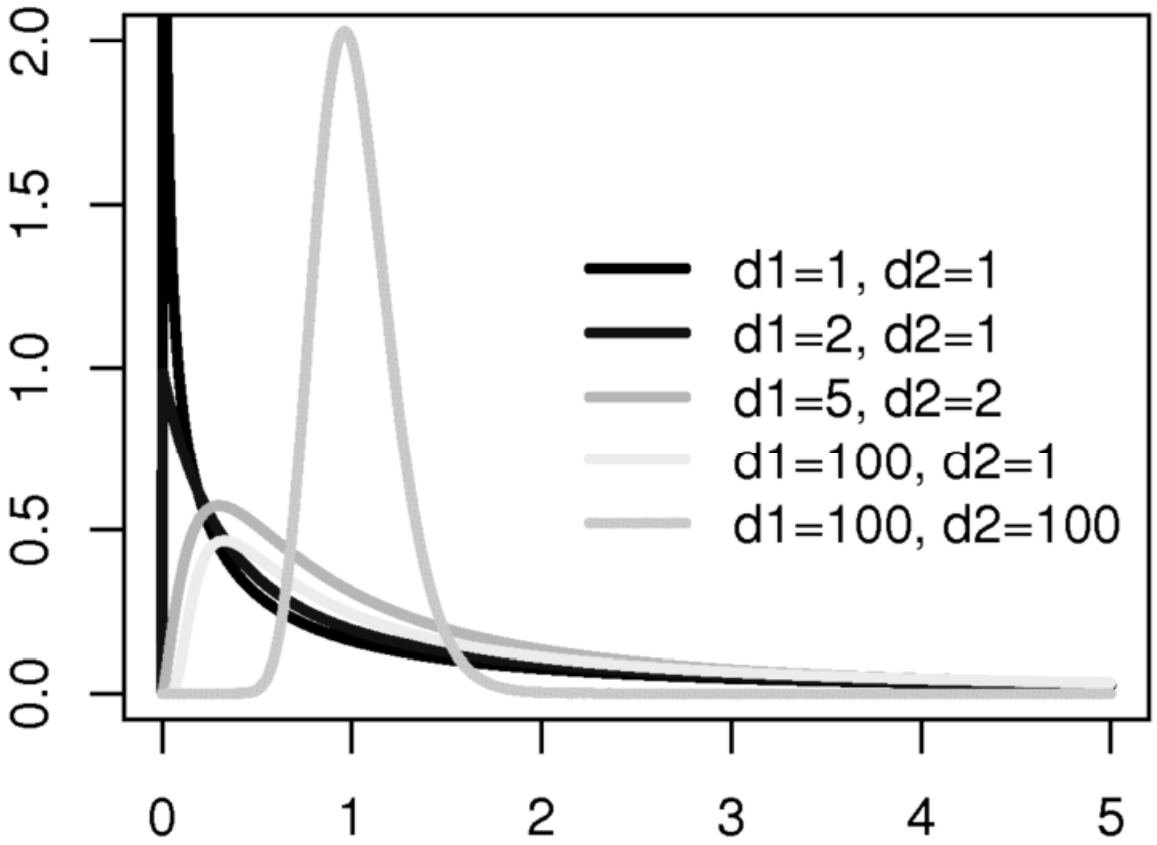


Source: Introduction to the Practice of Statistics, Moore and McCabe

ANOVA: F-statistic and F-distribution

- The F-distribution is indexed by two degrees of freedom
  - Numerator df = number of groups minus 1
  - Denominator df = total sample size minus number of groups
- The shape of the F-distribution varies depending on the degrees of freedom.

5 different F-Distributions
(Wikipedia)



The rejection region of the F-distribution is the right tail of the distribution. If the F-statistic is large enough to be in the rejection region the null hypothesis is rejected.

Partitioned difference from the grand mean

- The difference of each observation, X, from the grand mean can be partitioned into two parts:
  - The difference between the individual observation and the group mean.
  - The difference between the group mean and the grand mean

$$(X - \overline{\overline{X}}) = (X - \overline{X}_j) + (\overline{X}_j - \overline{\overline{X}})$$

Grand mean

Group Mean

Sums of Squares

- This partitioned relationship is also true for the sums of the ***squared differences*** which are called 'sums of squares'. In the formula below X_{ij} refers to an individual observation i in group j

$$\sum (X_{ij} - \overline{\overline{X}})^2 = \sum (X_{ij} - \overline{X}_j)^2 + \sum (\overline{X}_j - \overline{\overline{X}})^2$$

$$SS_T = SS_E + SS_A$$

- In words, the total sum of squares (SS_T) is equal to the Error Sum of squares (SS_E) plus the sum of squares among groups (SS_A)

Mean Squares

- MS_A is the mean square among groups= SS_A / (j-1)
  - MS_A has j-1 degrees of freedom where j= number of groups

- MS_E is the error mean square= SS_E / (N-j)
  - MS_E has N-j degrees of freedom where N= total number of observations and j= number of groups

F-statistic calculated

- **Recall that the F-statistic is the ratio of the average variability among groups divided by the average variability within groups**
  - The F-statistic = MS_A / MS_E

ANOVA Assumptions

- There are some assumptions that should be met before using ANOVA:
  - The observations are from a random sample and they are independent from each other
  - The observations are assumed to be normally distributed within each group
    - ANOVA is still appropriate if this assumption is not met but the sample size in each group is large (> 30)
  - The variances are approximately equal between groups
    - If the ratio of the largest SD / smallest SD < 2, this assumption is considered to be met.
- It is not required to have equal sample sizes in all groups.

ANOVA: the steps

- Step 1. State the Hypotheses
- Step 2. Calculate Test statistic
- Step 3. Calculate the p-value
- Step 4. Conclusion
  - If you fail to reject the null hypothesis, stop
  - If the null hypothesis is rejected indicating that at least two means are different, proceed with ***post-hoc tests*** to identify the differences

ANOVA Example

- Researchers were interested in evaluating whether different therapy methods had an effect on mobility among elderly patients.

- 18 subjects were enrolled in the study and were randomly assigned to one of three treatment groups
  - Control – no therapy
  - Trt. 1 – Physical therapy only
  - Trt. 2 – Physical therapy with support group

- After 8 weeks subjects responded to a series of questions from which a mobility score was calculated.

Data for ANOVA example

The hypothetical data below represent mobility scores for subjects in the 3 treatment groups.

A higher score indicates better mobility

Assume that mobility scores are approximately normally distributed

Control	Trt. 1	Trt. 2
35	38	47
38	43	53
42	45	42
34	52	45
28	40	46
39	46	37

ANOVA: 1. State the Hypotheses

- Step 1. State the Hypotheses
  - Null Hypothesis: $\mu_{control} = \mu_{trt\ 1} = \mu_{trt\ 2}$
  - Alternative Hypothesis: at least two of the means $(\mu_{control}, \mu_{trt\ 1}, \mu_{trt\ 2})$ are not equal

- ANOVA will identify if *at least two* of the means are significantly different

Identify the test and statistic

- The test for comparing more than 2 means is ANOVA
- Before proceeding with ANOVA check that the variances are approximately equal
- Calculate the sample SD for each group

	Control	Trt. 1	Trt. 2
SD	4.86	4.94	5.33

- If the ratio of the largest SD / smallest SD < 2 the variances are approximately equal
- 5.33 / 4.86 = 1.1 so ANOVA is appropriate
- The test statistic is the ANOVA F-statistic

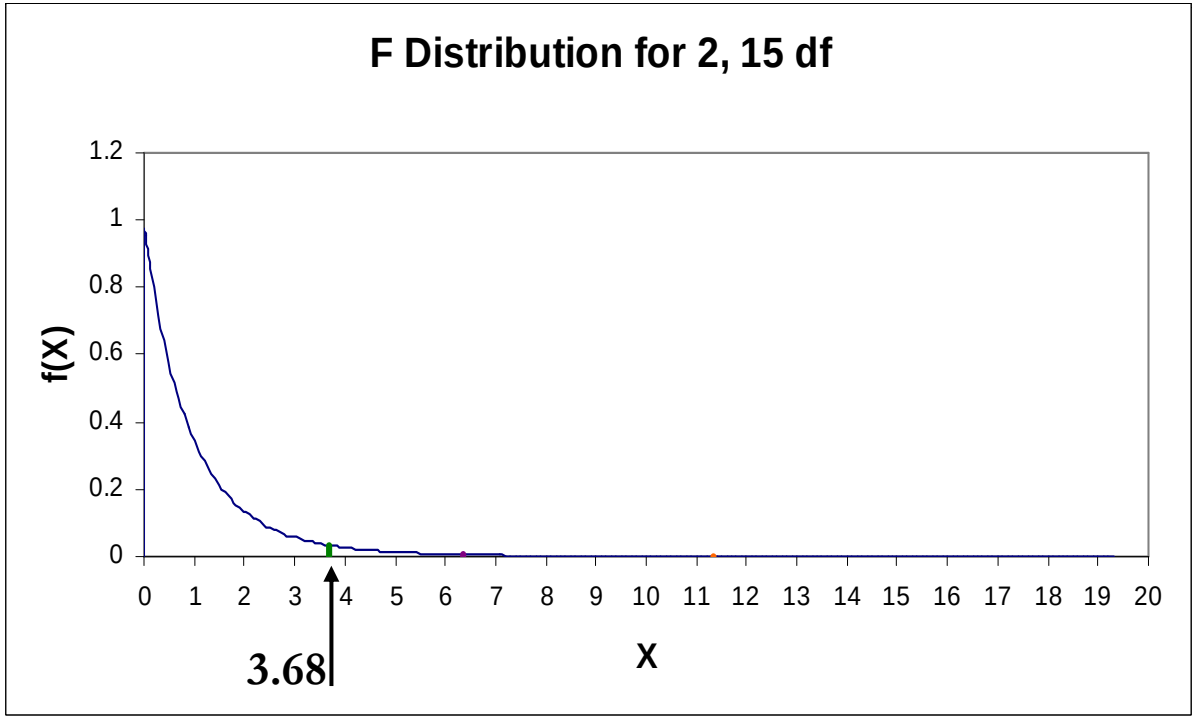
Significance Level and Critical Values

- Set the significance level – use the conventional alpha = 0.05
- Identify the critical value from the F-distribution
  - Numerator df =
  - Denominator df =

F-distribution for 2, 15 df

F-distribution for the F-statistic for 18 observations and 3 groups has 2, 15 df

The critical value for a = 0.05 in the F 2, 15 distribution is 3.68



FINV function in Excel

- Use the FINV function in Excel to find the critical value for the F-statistic
- FINV(probability, df_num, df_den)
- For the example data

  FINV(0.05, 2, 15) = 3.68
- If the F-statistic > 3.68, reject the null hypothesis
- If the F-statistic < 3.68, do not reject the null hypothesis

ANOVA step 2: Calculate the F-statistic

First calculate the grand mean, group means and group SD^2

- Grand mean = sum of all 18 observations divided by 18 = 41.7

Group	Group Mean	SD^2
Control	36	23.6
Trt 1	44	24.4
Trt 2	45	28.4

- Notice that none of the three group means are equal to the grand mean and that there are differences between the group means.
- ANOVA will identify if any of these differences are significant.

Calculate SS_A and MS_A

$$SS_A = \sum_{ij} (\overline{X}_j - \overline{\overline{X}})^2 = \sum_j n_j (\overline{X}_j - \overline{\overline{X}})^2$$

$SS_A = 6*(36 - 41.7)^2 + 6*(44 – 41.7)^2 + 6*(45 - 41.7)^2 = 292$

- MS_A
  - Divide the SS_A by the degrees of freedom (j-1)
- $MS_A = 292 / (3-1) = 292 / 2 = 146$

Calculate SS_E and MS_E

$$SS_E = \sum_{ij} (X_{ij} - \overline{X}_j)^2 = \sum_j (n_j - 1)(s_j)^2$$

- SS_E = 5*23.6 + 5*24.4 +5*28.4 =382

- MS_E = SS_E divided by the degrees of freedom (N-j)
  - MS_E = 382 / (18-3) = 382 / 15 = 25.47

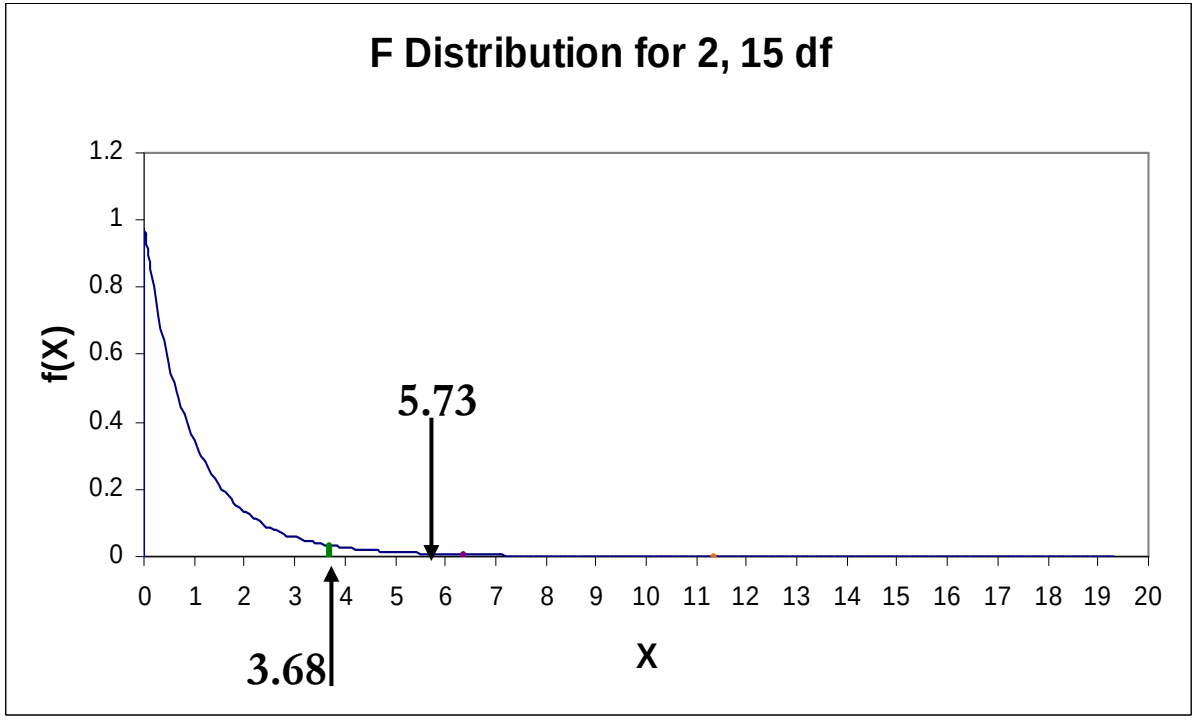
Step 3: Calculate p-value

- The ANOVA F-statistic = MS_A/MS_E
- ANOVA F-statistic =

- The p-value =

  FDIST(___, ___, ____) = 0.014

F-distribution for 2, 15 df



ANOVA Step 4: Conclusion

- The null hypothesis is _____
- Conclusion?

ANOVA in EXCEL

- Under Tools, select Data Analysis
- Select Anova: Single Factor
- Highlight all the observations
  - If you highlight the column headers check ‘labels’
- Indicate if groups are in columns or rows
- Select the alpha level for the F-test
- Select ‘Output Range’ and indicate a cell for the output table to be placed
- Select ‘OK’
- The ANOVA table and summary statistics will be placed in the indicated location.

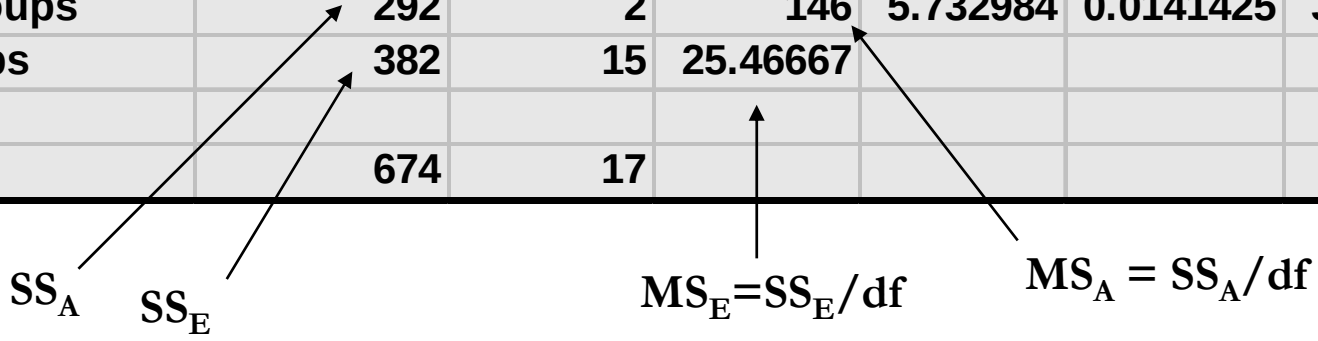
EXCEL ANOVA table

The Excel ANOVA table for the example data is on the next slide

- The ANOVA table summary provides the Count, Sum, Average and Variance for each group
- In Excel the ‘among group’ sum of squares and ‘among group’ mean square are labeled ‘*between* group’
- Excel uses the label ‘*within* group’ for the error sum of squares and error mean square
- The F-statistic, p-value and the F-critical value are provided
- The numerator and denominator degrees of freedom are also provided

ANOVA table for Example

SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
Column 1	6	216	36	23.6		
Column 2	6	264	44	24.4		
Column 3	6	270	45	28.4		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	292	2	146	5.732984	0.0141425	3.682317
Within Groups	382	15	25.46667			
Total	674	17				



ANOVA: Post-hoc tests

- A significant ANOVA F-test is evidence that not all group means are equal but it does not identify where the differences exist.
- Methods used to find group differences after the ANOVA null hypothesis has been rejected are called post-hoc tests.
  - Post-hoc is Latin for ‘after-this’
- Post-hoc comparisons should only be done when the ANOVA F-test is significant.

Adjusting the α-level for Post-hoc comparisons

- When post-hoc multiple comparisons are being done it is customary to adjust the α-level of each individual comparison so that the overall experiment significance level remains at 0.05

Bonferroni Post-hoc comparisons

- For an ANOVA with 3 groups, there are 3 combinations of t-tests.
- A conservative adjustment is to divide 0.05 by the number of comparisons.
  - For our example with 3 groups the adjusted α-level = 0.05/3 = 0.017
- A difference will only be considered significant if the p-value for the t-test statistic < 0.017.

Post-hoc Comparisons for Example

- Our example data had 3 groups so there are three two-sample post-hoc t-tests:
  - Control group compared to Treatment 1 group
  - Control group compared to Treatment 2 group
  - Treatment 1 group compared to Treatment 2 group

Post-hoc Comparisons for Example

- We know from the significant F-test (p= 0.014) that *at least* two of the groups have significantly different means.

- A two-sample t-test will be done for ***each comparison*** and the result will be considered significant if the p-value < 0.017.

Multiple Comparison Procedure

$$H_o : \mu_{group} = \mu_{another\ group}\ vs. H_A : \mu_{group} \neq \mu_{another\ group}$$

$$t_{df} = \frac{\bar{x}_{group} - \bar{x}_{anothergroup}}{\sqrt{MS_E\left(\frac{1}{n_{group}} + \frac{1}{n_{anothergroup}}\right)}}, \text{where } df = N - j$$

Post-hoc comparison: Control and Treatment 1

- Results of the two-sample t-test assuming equal variance to compare mean mobility scores between the control group and treatment 1 group:
  - Control mean score = 36
  - Treatment 1 mean score = 44
  - P-value for two-sample t-test = 0.0179
    - This is not significant at the adjusted α-level of 0.017 but could be considered marginally significant since it is close to 0.017.
- Conclusion: After adjusting for multiple comparisons, the control group mean mobility score is marginally significantly different than the mean mobility score for the group that received physical therapy ($p = 0.0179$).

Post-hoc comparison: Control and Treatment 2

- A two-sample t-test assuming equal variance is done to compare mean mobility scores between the control group and treatment 2 group:
  - Control mean score = 36
  - Treatment 2 mean score = 45
  - P-value for two-sample t-test = 0.012
    - This is a significant difference at the adjusted a-level of 0.017
- Conclusion: After adjusting for multiple comparisons, the control group mean mobility score is significantly different than the mean mobility score for the group that received physical therapy and a support group (p = 0.012).

Post-hoc comparison: Treatment 1 and Treatment 2

- A two-sample t-test assuming equal variance is done to compare mean mobility scores between the two treatment groups.
  - Treatment 1 mean score = 44
  - Treatment 2 mean score = 45
  - P-value for two-sample t-test = 0.743
    - There is not a significant difference between these two treatment groups.
- Conclusion: There is no significant difference in mean mobility score between the two treatment groups (p = 0.743).

ANOVA results summary

“ANOVA was done to evaluate differences in mean mobility score between three groups: a control group, a group that received physical therapy only and a group that received physical therapy and a support group. The significant ANOVA F-test result indicated that at least two of the mean mobility scores were significantly different.

ANOVA results summary

Post-hoc t-tests with adjusted α-level of 0.017 were done. Results of the post-hoc comparisons indicated that the treatment group with both physical therapy and a support group had a significantly different mean mobility score than the control group ($p = 0.012$); the treatment group with physical therapy only had a marginally significantly different mean mobility score than the control group ($p = 0.0179$); there was no significant difference in mean mobility score between the two treatment groups."

Post-hoc test Procedures

- There are other procedures for adjusting the α-level for multiple comparisons which you may see referenced in the Medical Literature and which are described in the text
  - Bonferroni procedure
  - Dunnett's Procedure
  - Newman-Keuls Procedure
  - Scheffe's Procedure
  - Tukey's HSD Procedure
  - Duncan Multiple range Procedure
- Many statistical software packages can accommodate these adjustments but we won't use these in this class

Other ANOVA Procedures

- More than one factor can be included for two, three or four-way ANOVA
  - Excel can provide results for a Two-way ANOVA but more complex ANOVA requires other software – PH6415

- A continuous variable can be added to the model – this is Analysis of Covariance (ANCOVA) – PH6415

- Repeated Measures ANOVA can handle replicated measurements on the same subject.

What if ANOVA is used to compare the means between two groups?

- The F-statistic for ANOVA with only two groups is equal to the t-statistic squared.
  - $F = t^2$ when both tests (ANOVA and t-test) are applied to the same two group comparison of means.

Winter Toddlers

Does it take longer to learn in the winter when babies are often bundled in clothes that restrict their movement?

At what age do babies learn to crawl? –two studies

- Data were collected from two different studies. One study in Minneapolis, MN and another completely separate study in Boulder, CO. Both studies had approximately the same number of randomly selected participants.

- Parents reported the birth month and the age at which their child was first able to creep or crawl a distance of four feet within one minute. Time was measured in weeks.

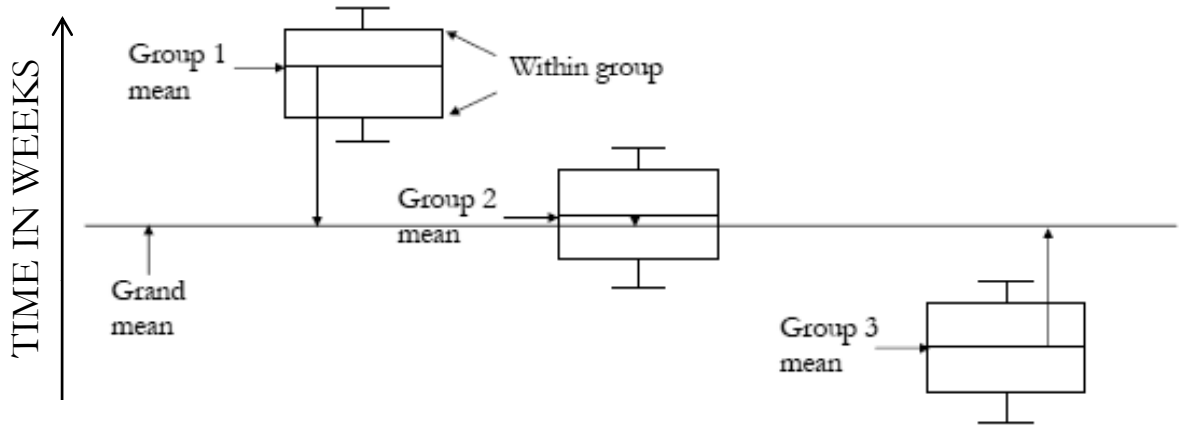
At what age do babies learn to crawl? –two studies

- The resulting data were grouped by month of birth: September (group 1) , January (group 2) and May (group 3) - all other months' data were discarded.

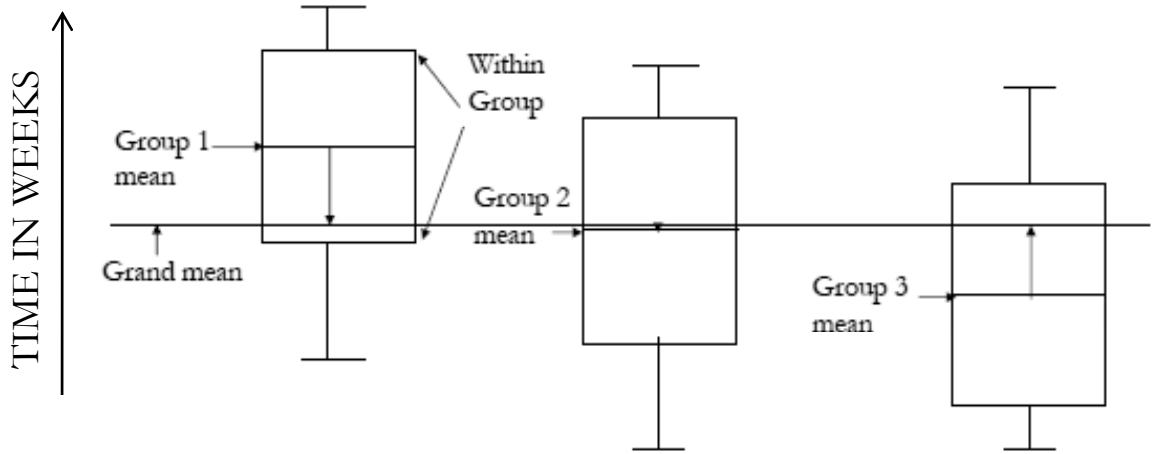
- Assume the data represent three independent simple random samples from each location, and that the populations of crawling ages have a normal distribution.

BOX-PLOTS OF THE DATA. Data is plotted on the same scale.

MINNEAPOLIS



BOULDER



ANOVA was performed separately for Denver and Minneapolis .

BASED ON THE BOX-PLOTS WHICH CITY'S DATA APPEARS TO HAVE VIOLATED THE EQUALITY OF VARIANCE ASSUMPTION?

1. Minneapolis
2. Boulder
3. Neither

BASED ON THE BOX-PLOTS, WHICH CITY'S STUDY DATA HAS A LARGER MEAN SQUARE ERROR (MSE)?

1. Minneapolis
2. Boulder
3. Neither

SUPPOSE THE ESTIMATED VARIABILITY BETWEEN MEANS (MSM) IS ABOUT THE SAME FOR BOTH LOCATIONS.

WHICH CITY'S STUDY DATA HAS A LARGER F STATISTIC?

1. Minneapolis
2. Boulder
3. Neither

ANOVA WAS PERFORMED FOR MINNEAPOLIS AND THE NULL HYPOTHESIS OF EQUALITY OF MEANS WAS REJECTED. THIS MEANS….

1. Babies born in January take significantly longer on average to crawl than babies born in both September and May.
2. Babies born in the three different months take equal amounts of time to crawl on average.
3. Babies born in the three different months do not take the same amount of time to crawl on average.
4. Babies born in the three months do not have a different average time to crawl.

Readings and Assignments

- Reading: Chapter 7 pgs. 164 – 173
  - We used the formulas for MS_A and MS_E on page 165. You don't need to know the computational formulas on page 169.
- Understand the relationships between sums of squares, mean squares and the F-statistic
- Work through the Lesson 13 practice exercise ANOVA example.