



# Analysis of Variance

## Part 5: Multiple Comparisons

STAT 705: Regression and Analysis of Variance

# How Did We Get Here?

- We have data that consist of one response variable and one predictor variable
  - The response is continuous (i.e. measured)
  - The predictor is a factor with  $t$  levels (i.e.  $t$  treatments)
- We have fit an ANOVA model to the data
  - We have verified that the assumptions are not grossly violated
  - The F test is significant  $\Rightarrow$  at least one treatment mean is different than the others
- What do we do next?

# Which Means are Different?

- Now that we know at least one mean is different, we perform additional tests to determine which means are different
- For example, we could compare the means of treatments  $i$  and  $j$  with a two-sample t test
  - $H_0: \mu_i = \mu_j$  vs.  $H_a: \mu_i \neq \mu_j$   
which can be written  $H_0: \mu_i - \mu_j = 0$  vs  $H_a: \mu_i - \mu_j \neq 0$
  - Test statistic:  $t = \frac{\bar{Y}_i - \bar{Y}_j}{s_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}$ , where  $s_p$  is the pooled std. dev.
  - Reject  $H_0$  if the absolute value of the test statistic is greater than the critical value from the t distribution with  $df = n_i + n_j - 2$

# Compare Many Pairs of Means

- If there are  $t = 3$  treatments, we would test three hypotheses
  - 1) trt 1 vs trt2  $\Rightarrow H_0: \mu_1 = \mu_2$  vs  $H_a: \mu_1 \neq \mu_2$
  - 2) trt 1 vs trt3  $\Rightarrow H_0: \mu_1 = \mu_3$  vs  $H_a: \mu_1 \neq \mu_3$
  - 3) trt 2 vs trt3  $\Rightarrow H_0: \mu_2 = \mu_3$  vs  $H_a: \mu_2 \neq \mu_3$
- If there are  $t = 4$  treatments, we would test 6 hypotheses
  - 1) trt 1 vs trt2  $\Rightarrow H_0: \mu_1 = \mu_2$  vs  $H_a: \mu_1 \neq \mu_2$
  - 2) trt 1 vs trt3  $\Rightarrow H_0: \mu_1 = \mu_3$  vs  $H_a: \mu_1 \neq \mu_3$
  - 3) trt 1 vs trt4  $\Rightarrow H_0: \mu_1 = \mu_4$  vs  $H_a: \mu_1 \neq \mu_4$
  - 4) trt 2 vs trt3  $\Rightarrow H_0: \mu_2 = \mu_3$  vs  $H_a: \mu_2 \neq \mu_3$
  - 5) trt 2 vs trt4  $\Rightarrow H_0: \mu_2 = \mu_4$  vs  $H_a: \mu_2 \neq \mu_4$
  - 6) trt 3 vs trt4  $\Rightarrow H_0: \mu_3 = \mu_4$  vs  $H_a: \mu_3 \neq \mu_4$

# Consequences of Multiple Tests

- Every time we perform a hypothesis test, the probability we incorrectly reject  $H_0$  is  $\alpha$ 
  - This is the Type I error (reject  $H_0$  when  $H_0$  is true)
  - We usually set  $\alpha = 0.05$
- If we perform 3 tests, the probability we incorrectly reject at least one null hypothesis is  $1 - (.95)^3 = .142$ 
  - We would make at least one Type I error about 14% of the time
- If we perform 6 tests, the probability we incorrectly reject at least one is  $1 - (.95)^6 = .265$ 
  - We would make at least one Type I error about 27% of the time
- We call this inflation of the Type I error rate, and this is why we do not use ordinary t tests for multiple comparisons

# Example to Illustrate Comparisons

The simulated data below were randomly selected from normally distributed populations that all have mean 100 and standard deviation 2. We should find no significant difference among the means.

Treatments	1	2	3	4	5
Responses	97	97	100	98	103
	99	101	101	100	106
	97	101	100	99	100
	100	103	99	101	100
	101	98	99	101	102
	101	103	98	99	100
	100	96	100	102	101
	102	102	98	101	100
Means	99.625	100.125	99.375	100.125	101.500
Variances	3.411	7.554	1.125	1.839	4.571

# ANOVA for Simulated Data

The ANOVA F-test shows that there are no significant differences among the means at the 5% level ( $p = 0.2354$ ).

This result is expected, since we know all of these samples came from identical populations.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	21.6000000	5.4000000	1.46	0.2354
Error	35	129.5000000	3.7000000		
Corrected Total	39	151.1000000			

When we use the ordinary t test to compare pairs of means, an interesting thing happens.

# Recall the Two-Sample t Test

Two populations	Two random samples
means $\mu_1$ and $\mu_2$	means $\bar{Y}_1$ and $\bar{Y}_2$
std. dev. $\sigma_1$ and $\sigma_2$	std. dev. $s_1$ and $s_2$

- Test  $H_0: \mu_1 = \mu_2$  vs.  $H_a: \mu_1 \neq \mu_2$











- Test statistic: 
$$t = \frac{|\bar{Y}_1 - \bar{Y}_2|}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- Critical value is from the t distribution with  $n_1 + n_2 - 1$  degrees of freedom

( The first square root in the denominator is often called the “pooled standard deviation”. )



# Pairwise Differences

Compare Treatments	p-value	Reject at $\alpha=.05$ ?	Correct Decision?
1 to 2	0.3379	No	
1 to 3	0.3724	No	
1 to 4	0.2735	No	
1 to 5	0.0408	Yes	
2 to 3	0.2417	No	
2 to 4	0.5000	No	
2 to 5	0.1414	No	
3 to 4	0.1191	No	
3 to 5	0.0123	Yes	
4 to 5	0.0734	No	

- p-values are obtained from ordinary two-sample t tests
- We reject exactly one of these ten pairwise tests
- This is simulated data
  - » We KNOW the means are all the same
  - » We should not reject any of these hypotheses
- The error rate is 2/10, or 0.20
- This is quadruple the desired error rate of 0.05

# Controlling the Type I Error Rate

- To maintain the desired Type I error rate, we must make adjustments to our criteria for rejecting  $H_0$  for the individual tests
- Many methods for adjustment have been proposed -- we consider these
  - Fisher's Least Significant Difference (LSD)
  - Tukey's Honest Significant Difference (HSD)
  - Bonferroni's Adjustment
  - Scheffe's
  - Dunnett's



# Fisher's Least Significant Difference

- Least Significant Difference is abbreviated LSD
- Fisher's LSD makes two changes to the ordinary t test
  - Degrees of freedom
    - ordinary t test:  $df = n_1 + n_2 - 2$
    - Fisher's LSD:  $df = dfE$  from overall ANOVA
    - This changes the critical value from the t distribution.
  - Pooled standard deviation
    - Ordinary t test: Combines the variances of the two groups being compared
    - Fisher's LSD: Uses root MSE from ANOVA, which uses information from all of the treatment groups

# Definition of Fisher's LSD

- Let  $t^*$  represent the critical value from the  $t$  distribution with  $df = df_E$
- Saying that two means are significantly different when  $|t| > t^*$  is mathematically the same as saying they are different when

$$|\bar{Y}_i - \bar{Y}_j| > t^* \sqrt{MSE} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- The quantity on the right side of the inequality is the smallest difference between means that is statistically significant, so it is called the least significant difference or LSD











# LSD for the Simulated Data

- For the simulated data
  - dfE = 35
  - MSE = 3.7
  - $n_i = 8$  for every group
  - critical value (for  $\alpha=0.05$ ) is  $t^* = 2.030$

$$\text{LSD} = (2.030) \sqrt{3.7 \left( \frac{1}{8} + \frac{1}{8} \right)}$$

$$= 1.952$$

- We incorrectly reject  $1/10 = 0.10$

Compare Treatments	Difference	Significant?	Correct Decision?
1 to 2	.5	No	
1 to 3	.25	No	
1 to 4	.5	No	
1 to 5	1.875	No	
2 to 3	.75	No	
2 to 4	0	No	
2 to 5	1.375	No	
3 to 4	.75	No	
3 to 5	2.125	Yes	
4 to 5	1.375	No	

# Tukey's HSD

- HSD is 'Honest Significant Difference'
- Consider the distribution of the largest sample mean minus the smallest sample mean
  - This difference follows a Studentized Range distribution
  - A table for this distribution is provided on the course website
- $HSD = (q^*) \times \sqrt{\frac{MSE}{n}}$  , where
  - where  $q^*$  is the critical value from Studentized Range
  - $n$  = sample size for each group
- Any difference of means bigger than HSD are declared to be statistically significant

# HSD for the Simulated Data

- For the simulated data
  - $df_E = 35$ ,  $MSE = 3.7$ , and  $n_i = 8$  for every groups
  - critical value (for  $\alpha=0.05$ , 5 treatments, and  $df = 35$ ) is  $q^* = 4.07$  (interpolate between 30 and 40 df)
- $HSD = 4.07 \sqrt{\frac{3.7}{8}} = 2.77$
- The largest difference in means is 2.125 (between treatments 3 and 5). It is not significant, so none of the differences are significant  $\Rightarrow$  **we would have decided correctly for every test**
- Note: If the sample sizes are not equal, we use a modified version called Tukey-Kramer

# Bonferroni's Adjustment

- Number of comparisons (k) must be known in advance
- For each test, reduce the significance level from  $\alpha$  to  $\alpha/k$
- For the simulated data
  - For each test, use 0.05/10, or 0.005
  - Reject only if the p-value is less than 0.005
  - None of the individual tests are significant.
  - **We would have decided correctly for every test**



# Scheffe's Method

- A modification of the ANOVA F test
- Critical value ( $F^*$ ) is from F distribution with numerator and denominator degrees of freedom  $t - 1$  and  $dfE$ , respectively
- Scheffe's Least Significant Difference

$$= \sqrt{(t-1)(F^*)(MSE)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

- Any difference of means bigger than this are declared to be statistically significant

# Scheffe for Simulated Data

- Scheffe's Least Significant Difference

$$= \sqrt{(5-1)(2.65)(3.7)\left(\frac{1}{8} + \frac{1}{8}\right)}$$
$$= 3.13$$

- The largest difference in means is 2.125 (between treatments 3 and 5)
- The largest difference is not significant, so none of the differences are significant
- **We would have decided correctly for every test**

# Dunnett's Method

- Is used ONLY to compare all treatments to a single reference control treatment
  - Should not be used for all possible pairwise comparisons
- Find critical value ( $d^*$ ) in Dunnett's table
  - Depends on number of treatments (excluding control) and  $df = df_E$
  - A table for this distribution is on the course website
- A difference of means is significant if it is greater than

$$(d^*) \sqrt{\frac{2 \cdot \text{MSE}}{(\# \text{replications})}}$$

- This method is not appropriate for our example data because there is not a control treatment

# Comparison of Methods

- Tukey's method
  - Preferred when we are testing all pairs of means
- Bonferroni's method
  - Can be overly strict, i.e., can fail to find important differences
- Scheffe's method
  - Primarily for a collection of tests that are suggested by data, i.e. 'data snooping'
  - The other methods should only be used for pre-planned comparisons
- Dunnett's method
  - Is used when we are comparing each treatment to a control treatment



# What You Should Know

- Why we need to make adjustments when we perform multiple hypothesis tests with the same data
- Understand the differences among the methods we presented
- Be able to choose an appropriate method

In a later lesson, we discuss more general hypotheses for comparing means (other than pairwise comparisons)