Lectures 31 & 32 – Confidence Intervals for the Difference in Means

Analysis of Treatment Effects

Suppose that, based upon an analysis of the residuals, it appears reasonable that the errors are independent $N\left(0,\boldsymbol{\sigma}^2\right)$, and that the *P*-value for the model leads us to conclude that some treatment means differ. The next step is to investigate the nature of the differences in treatment effects. From your introductory statistics course, you may be familiar with the *t*-test for the difference in two population means, μ_1 and μ_2. Although this is not the test that is usually used in the analysis of variance, it is a good place to start a discussion about analyzing treatment effects in one-way ANOVA.

A *t*-Test for the Difference in the Means of Two Treatments

Suppose that our interest lay only in considering whether the mean value of the dependent variable is different for two different treatments. If the assumption that the errors are independent $N\left(0,\boldsymbol{\sigma}^2\right)$ appears reasonable, then we can construct confidence intervals, and conduct hypothesis tests, for the difference in the means, $\mu_2-\mu_1$.

Confidence intervals and hypothesis tests are most straightforward when they involve a *single* parameter estimated by a *single* random variable! So, we should think of the difference in the means $\mu_2-\mu_1$ as a single parameter. Then the obvious choice for an estimator is the random variable $\bar{\boldsymbol{Y}}_2-\bar{\boldsymbol{Y}}_1$, where $\bar{\boldsymbol{Y}}_1$ and $\bar{\boldsymbol{Y}}_2$ are the means of independent samples drawn from the two treatments. The following notation summarizes where we are at this point.

- n_1 and n_2 are the sizes of the samples drawn from treatment one and two, respectively.
- σ_1^2 and σ_2^2, the variances for the dependent variable in each population, are assumed equal, i.e., $\sigma_1^2=\sigma_2^2=\sigma^2$, where σ^2 is the variance of the error.
- $\bar{\boldsymbol{Y}}_1=\frac{1}{n_1}\sum_{j=1}^{n_1}\boldsymbol{Y}_{1j}$ and $\bar{\boldsymbol{Y}}_2=\frac{1}{n_2}\sum_{j=1}^{n_2}\boldsymbol{Y}_{2j}$ are the sample means under the two treatments.

Now, any confidence interval for, or hypothesis test of, $\mu_2-\mu_1$ will depend upon the distribution of the estimator $\bar{\boldsymbol{Y}}_2-\bar{\boldsymbol{Y}}_1$. Proceeding step by step,

- $\bar{\boldsymbol{Y}}_1\sim N\left(\mu_1,\frac{\sigma^2}{n_1}\right)$, and $\bar{\boldsymbol{Y}}_2\sim N\left(\mu_2,\frac{\sigma^2}{n_2}\right)$, because the error is assumed to be normally distributed.
- $\bar{\boldsymbol{Y}}_2-\bar{\boldsymbol{Y}}_1\sim N\left(\mu_2-\mu_1,\frac{\sigma^2}{n_1}+\frac{\sigma^2}{n_2}\right)$, because linear combinations of normal variables are also normal, and the variance of the difference of two *independent* random variables is the *sum* of their individual variances.

Although we've assumed constant variance σ^2, we've made no assumption about the actual value of σ^2. Therefore, we must estimate σ^2 from the samples taken from the treatments. For shared variance σ^2, the most efficient way to estimate σ^2 is by through the *pooled* estimator $\boldsymbol{S}_p^2 = \dfrac{(n_1-1)\boldsymbol{S}_1^2+(n_2-1)\boldsymbol{S}_2^2}{n_1+n_2-2}$, where

$\boldsymbol{S}_1^2 = \dfrac{\sum_{j=1}^{n_1}\left(\boldsymbol{Y}_{1j}-\bar{\boldsymbol{Y}}_1\right)^2}{n_1-1}$ and $\boldsymbol{S}_2^2 = \dfrac{\sum_{j=1}^{n_2}\left(\boldsymbol{Y}_{2j}-\bar{\boldsymbol{Y}}_2\right)^2}{n_2-1}$ are the sample variances for the two samples drawn.

When the pooled estimator $\boldsymbol{S}_p$ of the variance σ^2 is used to Studentize the random variable $\bar{\boldsymbol{Y}}_2-\bar{\boldsymbol{Y}}_1 \sim N\left(\mu_2-\mu_1,\sigma^2\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)\right)$, the resulting statistic $\boxed{\dfrac{\left(\bar{\boldsymbol{Y}}_2-\bar{\boldsymbol{Y}}_1\right)-\left(\mu_2-\mu_1\right)}{\boldsymbol{S}_p\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}}$ has a ***t*** distribution with n_1+n_2-2 degrees of freedom.

Example: Five measurements of the carbon content (in ppm) of silicon wafers were recorded on successive days of production. Can we conclude, at the 5% significance level, that the mean carbon content has changed?

Day 1	2.01	2.13	2.20	2.09	2.07
Day 2	2.31	2.41	2.23	2.19	2.26

The values in the table below were computed on my calculator

Populations (Treatments)	Sample Size	Sample Mean	Sample Variance	Pooled Variance	*df*	Critical Value Used
Day 1	5	2.10	0.0050	0.0061	8	$t_{8,0.025}=2.306$
Day 2	5	2.28	0.0072			

Note: Because the sample sizes were equal in this example, the pooled variance is just the average of the two sample variances. In general, however, the pooled variance is a *weighted* average of the sample variances, where greater weight is placed on the estimate derived from the larger sample. This should seem reasonable since larger samples tend to provide more accurate estimates, and therefore should carry more weight in the *pooled* estimate.

Then a 95% confidence interval for the difference in mean carbon content for the two days is given by

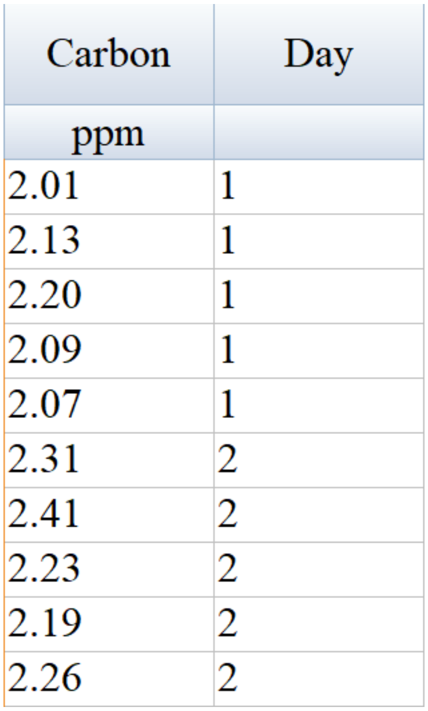
$(2.28-2.10)\pm 2.306\left(0.078\sqrt{\dfrac{1}{5}+\dfrac{1}{5}}\right)=0.18\pm 0.114$, or (0.066 ppm, 0.294 ppm).

Similarly, we can conduct a two-tailed test of the equality of the mean carbon content for the two days, with hypotheses

- ❖ $\mathbf{H}_0$: $\mu_1=\mu_2$
- ❖ $\mathbf{H}_A$: $\mu_1\neq\mu_2$

The *P*-value for this test of 0.00665 (obtained from my calculator) suggests that the mean carbon content of the wafers for the two days differs, a conclusion we could have reached based on the confidence interval for the difference in mean carbon content derived above. (Why?)

Since one-way analysis of variance is designed to detect the difference between the means of treatments, it’s natural to ask if the analysis above could have been done in ANOVA. Let’s try. The spreadsheet below contains the data for carbon content.



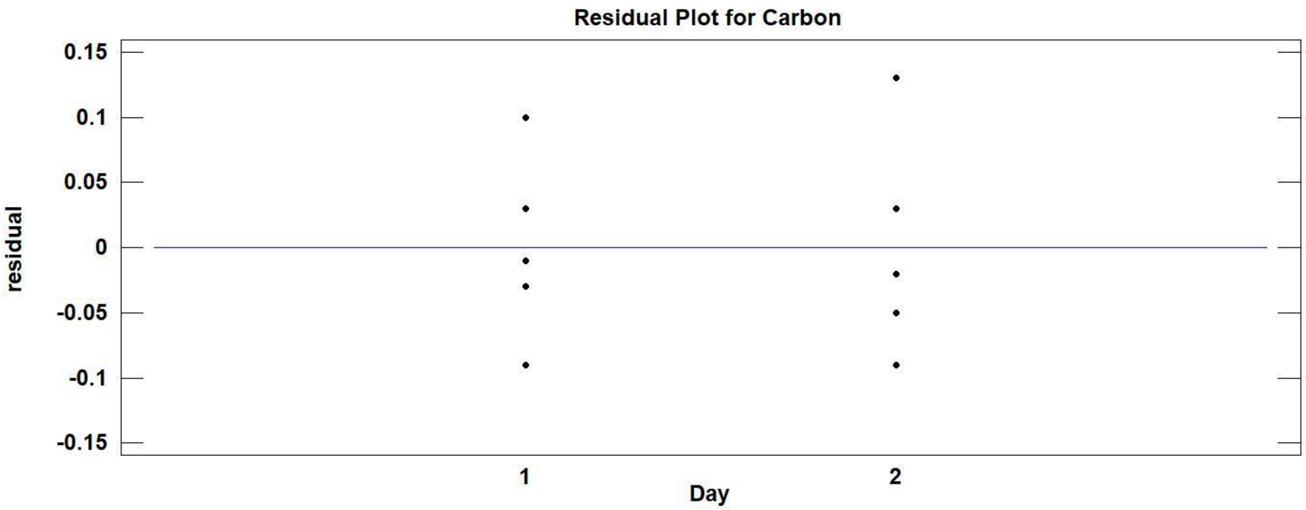
Carbon	Day
ppm	
2.01	1
2.13	1
2.20	1
2.09	1
2.07	1
2.31	2
2.41	2
2.23	2
2.19	2
2.26	2

Running a one-way analysis of variance produces the following output in Statgraphics.

ANOVA Table for Carbon by Day

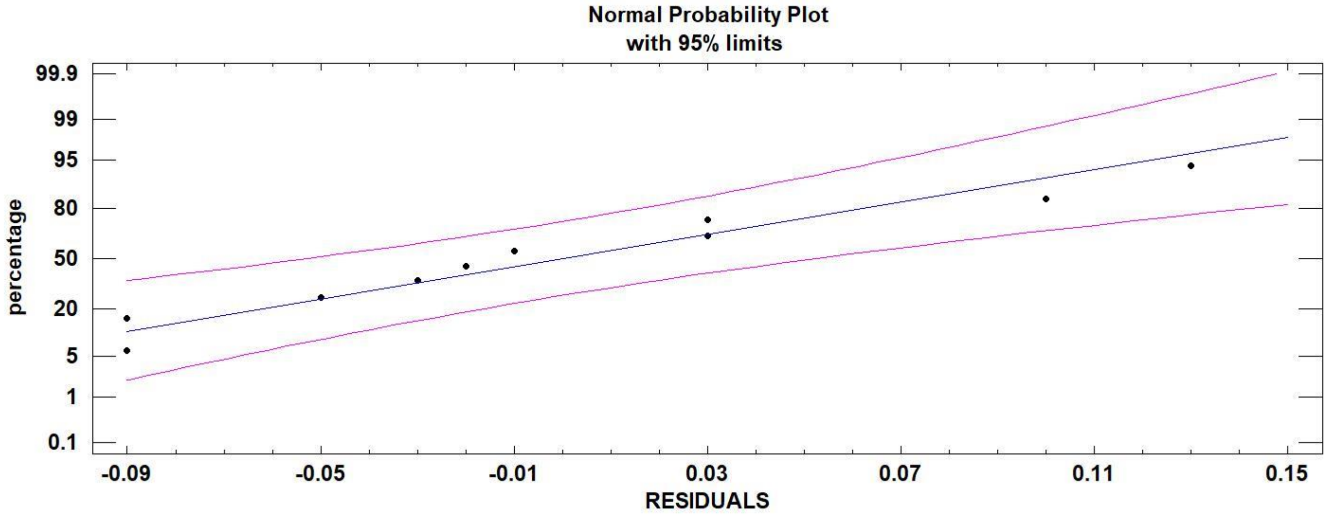
Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Between groups	0.081	1	0.081	13.28	0.0066
Within groups	0.0488	8	0.0061		
Total (Corr.)	0.1298	9			

Note that the estimated variance of 0.0061 (the Mean Square Error in the ANOVA table) is the pooled estimate s_p found earlier, and the *P*-value of 0.0066 for the *F-Ratio* in one-way ANOVA agrees with the *P*-value computed on my calculator for a *t*-test of the difference in two means when a common variance is assumed. (The *F* statistic 13.28 in the ANOVA table is the square of the *t* statistic -3.644 for the *t*-test.) Although the test of mean carbon content would usually be done as *t*-test, one advantage of preforming an analysis of variance is the wealth of graphs and tests at our disposal. For instance, the tables and graphs on the next page support the assumptions of normal error and constant variance that apply to the *t*-test as well. For the sake of brevity, they are presented without commentary, but they should be familiar to many of you by now.

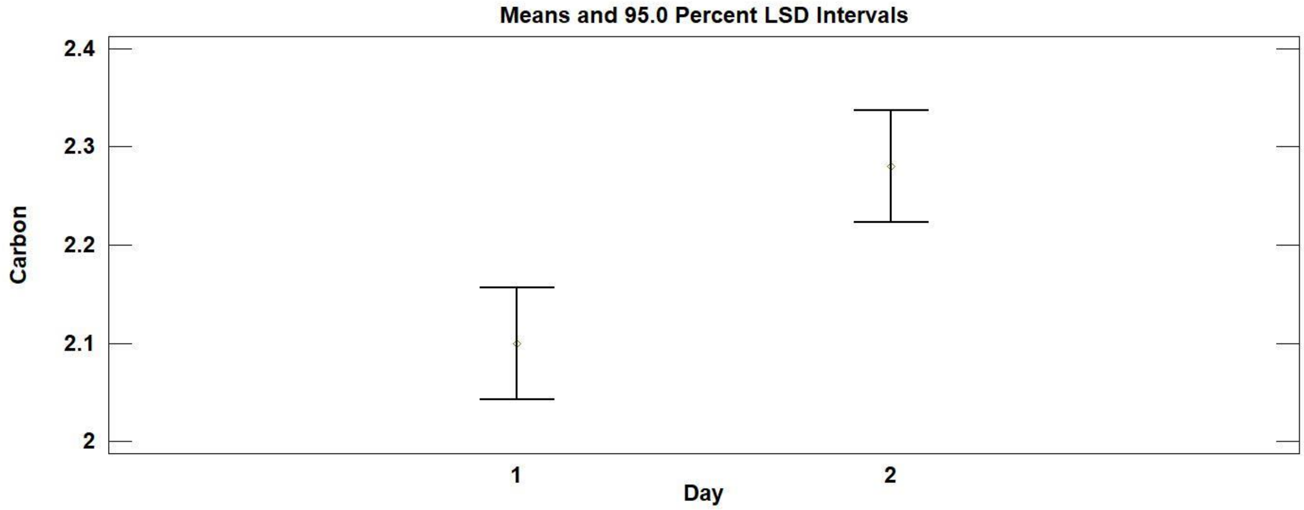


Variance Check

	Test	*P-Value*
Levene's	0.194595	0.6708



Stnd. skewness	0.732902
Stnd. kurtosis	-0.282025



Now, you're probably wondering why I've spent so much time on a test that rarely gets used in the analysis of variance. The first reason is that *t*-tests of the difference between the means of two distributions are popular in statistics, and this is a class in second semester statistics. The second reason is to argue that one-way analysis of variance can be viewed as an extension of the logic behind *t*-tests when the means of more than two treatments are of interest.

But, if $k > 2$ treatments are involved, why not simply construct $\binom{k}{2}$ $(1-\alpha)100\%$ *t*-intervals or conduct $\binom{k}{2}$ α level *t*-tests? The answer involves the idea of the family-wise error rate.

The Family-Wise Error Rate

Suppose that a factor has *k* levels, and that we've decided to conduct $r = \binom{k}{2} = \frac{k!}{2!(n-2)!} = \frac{k(k-1)}{2}$ pairwise *t*-tests of treatment means and reject the null hypothesis that all treatment means are equal if any of the pairwise tests are significant at α. What is the significance level, i.e., the probability of a **Type I** error, for the test that all treatment means are equal? (Hint: It's not α.) The problem is not merely that this could involve conducting a large number of *t*-tests, but the much less obvious problem of the family-wise error rate for such a strategy.

In the two-tailed *t*-tests being considered here, the significance level of the test, α, is the probability of rejecting the statement $\mu_{\text{treatment A}} = \mu_{\text{treatment B}}$ when the means for the two treatments are equal, i.e., the probability of committing a **Type I** error for that for that particular pairwise test of treatments A & B, i.e., $\alpha = P(\text{Concluding the means differ} \mid \text{the means are equal})$. But we are conducting not *one* such test, but $\binom{k}{2}$ such tests. The question is, what is the probability that *at least one* **Type I** error occurs for the *family* of $\binom{k}{2}$ tests. This is called the *family-wise error rate*, and, as we'll see, it can be much larger that α.

If the $\binom{k}{2}$ tests were independent (they actually aren't because the same data is used in multiple tests), then it's easy to see that *P*(at least one Type I error) = 1 - *P*(no Type I errors) = $1-(1-\alpha)^r$ where $r = \binom{k}{2}$ is the number of pairwise tests. In fact, this forms an upper bound on the family-wise rate, while a more conservative upper bound is given by $\min\{r\alpha, 1\}$. For $\alpha = 0.05$, the table below summarizes the two upper bounds on the family-wise error rate for a few choices of *k*.

Discussion: The more conservative upper bound on the family-wise **Type I** error is easily derived. If all μ_i are equal, and A_j, for $j \in \{1, \cdots, r\}$, is the event that the j^{th} pairwise *t*-test is significant at α, then $P(A_j) = \alpha$ and the probability of *at least one* pairwise Type I error among the *r* pairwise tests conducted is $P(A_1 \cup \cdots \cup A_r) \le P(A_1) + \cdots + P(A_r) = r\alpha$, so $P(A_1 \cup \cdots \cup A_r) \le min\{r\alpha, 1\}$.

Number of Levels, ***k***	***r***	**Multiplicative Upper Bound** on *P*(at least one pairwise Type I error)	**Additive Upper Bound** on *P*(at least one pairwise Type I error)
3	3	$1-(0.95)^3 =$ **0.14**	$3*0.05=$ **0.15**
4	6	$1-(0.95)^6 =$ **0.26**	$6*0.05=$ **0.30**
5	10	$1-(0.95)^{10} =$ **0.40**	$10*0.05=$ **0.50**
6	15	$1-(0.95)^{15} =$ **0.54**	$15*0.05=$ **0.75**

Upper Bounds on the Family-Wise Error Rate for Different *k*

Alternatives to All Possible Pairwise *t*-Tests:

The problem of family-wise error rates has attracted the attention of some of the biggest names in statistics, who have developed procedures for constructing simultaneous confidence intervals that can also be used to conduct pairwise tests of treatment means. Now, unlike the scenario outlined earlier, the intervals described below are *not* used to test that all treatment means are equal, but are constructed after a determination to reject equal means has been made. They can be accessed in StatGraphics by right-clicking and selecting *Pane Options*.

- Fisher's Least Significant Difference (LSD) intervals: Named after Sir R. A. Fisher, this is the method that StatGraphics defaults to (it appears in some of my solutions for no better reason than this).
- Tukey's Honest Significant Difference (HSD) intervals: Named for John Tukey, who worked for AT&T back when it had more money than God (and better service), this method was specifically designed to control the family-wise Type I error rate for *all* possible pairwise comparisons of treatment means at a fixed α. In most cases, this is the set of intervals that are preferred.
- Scheffe Intervals: Named for Henry Scheffe, who, besides deriving his intervals, wrote a classic text on the analysis of variance. This procedure is about more than just confidence intervals and pairwise comparisons. It was designed for the related problem of drawing inference on *contrasts*. We haven't discussed contrasts, but Scheffe came up with a way to conduct tests of all possible contrasts at a fixed family-wise rate α. Scheffe intervals, however, tend to be more conservative, i.e., wider, than HSD intervals because the contrasts Scheffe considered are generalizations of the pairwise comparisons considered by Tukey.
- Bonferroni Intervals: One of the original attempts at solving the problem of family-wise error rates, the eponymous Bonferroni intervals are still useful in certain situations. Generally, however, Tukey's HSD intervals are probably those most commonly employed to draw simultaneous inference in ANOVA at a fixed α.

Example

Example: (This is the first example explored in the original notes on the analysis of variance.) A city is looking to buy lightbulbs for the city’s streetlights. Seven lightbulbs from each of four brands (GE, Dot, West, and a generic) are purchased and placed in streetlights. The lifetime of each of the 28 lightbulbs is then recorded in the file *Lightbulbs*. Let's consider four different 95% confidence intervals for the difference of the means of the GE and Dot lightbulbs. (Actually, I'm only interested in the width of the intervals since they will all be centered about the same point estimate of the difference in the means.) The table below contains all of the relevant statistics for the four intervals to be created.

Brand	Sample Size	Sample Mean	Sample Variance		
GE	7	2.336	0.0460		
Dot	7	2.0	0.0213		
West	7	1.787	0.0152		
generic	7	2.1	0.0105		

Sample Means and Variances for the Four Brands of Lightbulbs

1. First, we construct a simple ***t***-interval based solely on the samples taken from the treatments GE and Dot. For $\bar{\boldsymbol{Y}}_{GE} - \bar{\boldsymbol{Y}}_{Dot}$, the pooled estimate of the variance is $s_p^2 = \frac{(7-1)s_{GE}^2 + (7-1)s_{Dot}^2}{7+7-2} = 0.03365$.

   The 95% confidence interval for $\mu_{GE} - \mu_{Dot}$ is $\left(\bar{y}_{GE} - \bar{y}_{Dot}\right) \pm t_{7+7-2,0.025}\sqrt{s_p^2\left(\frac{1}{7}+\frac{1}{7}\right)}$ =

   $\left(2.336 - 2.000\right) \pm\ t_{12,0.025}\sqrt{0.03365\left(\frac{1}{7}+\frac{1}{7}\right)} = 0.336 \pm 2.179\left(0.0981\right) = 0.336 \pm 0.214$.

2. The ***t***-interval for $\mu_{GE} - \mu_{Dot}$ using Fisher's LSD is similar to the interval above, but with the pooled estimate of the variance and the degrees of freedom derived from the *MSE* estimate of the error variance σ^2 computed from the ANOVA table. The general form for a Fisher $(1-\alpha)100\%$ LSD confidence interval for $\mu_{GE} - \mu_{Dot}$ is $\left(\bar{Y}_{GE} - \bar{Y}_{Dot}\right) \pm t_{n-k,\alpha/2}\sqrt{MSE\left(\frac{1}{7}+\frac{1}{7}\right)}$, where $n-k$ is the degrees of freedom associated the error sum of squares *SSE*.

   For the light bulb data, where *MSE* = 0.02324 and *df* = 24, the 95% Fisher *t*-interval becomes $\left(2.336 - 2.000\right) \pm t_{24,0.025}\sqrt{0.02324\left(\frac{1}{7}+\frac{1}{7}\right)} = 0.336 \pm 2.064\left(0.0815\right) = 0.336 \pm 0.168$

3. The ***t***-interval for $\mu_{GE} - \mu_{Dot}$ using Bonferroni's method is similar to the LSD interval above, but replacing α with α / r, where $r = \binom{k}{2}$ is the number of pairwise comparisons. A Bonferroni 95% family-wise ***t***-interval for $\mu_{GE} - \mu_{Dot}$ is given by

   $\left(2.336 - 2.000\right) \pm t_{24,0.025/6}\sqrt{0.02324\left(\frac{1}{7}+\frac{1}{7}\right)} = 0.336 \pm 2.875\left(0.0815\right) =\ 0.336 \pm 0.234$.

4. Tukey's HSD intervals use a critical value drawn from a **Studentized Range** distribution with $df_1 = k$ and $df_2 = n-k$ (compare this with the ***F***-test in one-way analysis of variance where $df_1 = k-1$ and $df_2 = n-k$). Tables for the Studentized Range distribution appear in statistics texts, or can be found online. The critical value for the Studentized Range is written $q_{k,n-k,\alpha}$. A Tukey 95% family-wise confidence interval for $\mu_{GE} - \mu_{Dot}$ is given by

   $\left(2.336 - 2.000\right) \pm q_{4,24,0.05}\sqrt{\frac{0.02324}{2}\left(\frac{1}{7}+\frac{1}{7}\right)} = 0.336 \pm 3.90\left(0.0576\right) = 0.336 \pm 0.225$

- The Bonferroni interval is wider than Fisher's LSD, but has the advantage of guaranteeing a fixed level of confidence for the family of all pairwise comparisons. Unless you've decided to focus on one particular comparison before gathering the samples, i.e., a-priori, the Bonferroni is better because it guarantees a family-wise error rate.

- Tukey's interval was narrower than Bonferroni's. This will be the case whenever all pairwise comparisons are considered, as when the choice of comparisons is made *after* seeing the data (post-hoc). Tukey intervals are the most widely used when multiple pairwise comparisons are envisioned.

The results of this example are summarized below, where the margin of error for each type of interval is given.

Type of Interval Used	**Margin of Error**	**Comments**
t	0.214	Does not control the family-wise error rate at 0.05
Fisher's LSD	0.168	Narrower than ***t***, but doesn't control the family-wise error rate at 0.05
Bonferroni	0.234	Controls the family-wise error rate at 0.05, but is conservative
Tukey's HSD	0.225	Controls the family-wise error rate for all pairwise comparisons at 0.05
Scheffe	*Not computed*	Controls the family-wise error rate for all possible *contrasts* at 0.05