ORIGIN ≡ 1

Multiple Pairwise Comparison Procedures in One-Way ANOVA with Fixed Effects Model

When the omnibus F-Test for ANOVA rejects the null hypothesis that all α_i's = 0, or equivalently that all μ_i are the same, then one usually wants to determine *specifically which* of the μ_i's differ from the others. In two population t-Tests, only a single comparison between population means (μ_1 with μ_2) is made. In ANOVA, greater than two populations is standard and *multiple pairwise* (i.e., $\mu_1 - \mu_2$ *and* $\mu_1 - \mu_3$ *and* $\mu_2 - \mu_3$ for three populations) or *multiple linear contrasts (i.e.,* $\mu_1+\mu_2-2\mu_3$ etc.) are often of interest. In these cases, comparisons are made simultaneously, and are therefore dependent upon the same sample data. As a result, it is standadard practice to pool the estimate of within group variance (MS_W). However, in comparing multiple hypotheses simultaneously an important complication arises. The existence of multiple dependent probabilities derived from each comparison implies that the joint probability of a *family of comparisons* together is greater than *each one separately.* Thus, if one specifies Type I error $\alpha = 0.05$ for one one test then *familywise* α_{fw} for all tests together is always greater than α (e.g., less significant - see Zar 2010 Ch. 10).

From the mathematical end of things, statisticians routinely caution experimenters about the potential pitfalls of *post hoc* (analysis following data collection, or "unplanned") "data snooping" (to borrow a term from Neter et al. 1996). By this, they mean running a large number of simultaneous tests or confidence intervals, and then proceeding to report "statistically significant" findings as if discovered outside the context of all the tests run, or worse, as the result of a strategically-chosen *a priori* (before data collection, or "planned") experimental design. The problem is that if enough simultaneous tests are run, the laws of probability predict that some tests will end up showing significance merely due to chance. This mathematically-based caution is certainly correct.

Given this, are one or more "significant" planned or unplanned results in ANOVA tests to be considered valid or not? Much depends on what exactly is meant by the foundational concept of α in often widely differing theoretical and experimental contexts. Whereas mathematicians might like to draw a bright line between *a priori* and *post hoc*, rarely in experimental practice is the distinction so clear. All experiments exist within a framework of pre-existing literature and laboratory/field data collection. In addition, data collection in research often involves cycles of refinement, and thus rarely a one-shot deal. So *of course* biologists regularly conduct complex "data snooping" in conceiving of problems, designing studies and analyzing results. They could hardly do otherwise, and engaging in a debate about what a researcher knew about data, and when he or she knew it relative to some often abstract point in time, is usually very silly.

However, in my opinion, the a *priori* vs *post hoc* distinction remains of interest from both theoretical and practical standpoints, and to be aware of the issues involved makes it possible to construct stronger scientific arguments. The distinction also points to clear limitations in statistical reasoning in the sciences to the extent that all of it must be acknowledged to be nothing more than a rough approximation. If one's data are overwhelmingly clear, then difficulties in approximation of probabilities do not really matter. However, if the data are unclear, then how one employs the approximation may influence what one might say within a test, but not necessarily what one might conclude. The take-home message remains the same - the data remain ambiguous. Biological interpretation, and experimental replication, must necessarily take precedence over abstract mathematical reasoning.

Simultaneous Inference Procedures in Practice:

Several *multiple comparison procedures* (MCP) have been developed to adjust Probabilities of Tests and associated Confidence Interval widths to accomodate familywise assessments. Some methods explicitly permit "data snooping" whereas others do not. It is important to be aware of how, and under what circumstances, each procedure is employed. As a practical matter, of course, standard statistical packages offer a full battery of possibilites and if the data permits, use of the "most conservative" (i..e, widest confidence intervals) is often considered evidence of good experimental design.

Data Structure:

k groups with not necessarily the same numbers of observations and different means.

Let index i,j indicate the ith column (treatment class) and jth row (object).

One-Way ANOVA					
	Treatment Classes:				
Objects (Replicates)	#1	#2	#3	…	#k
1					
2					
3					
…					
n	n1	n2	n3		nk
means:	X1bar	X2bar	X3bar	…	Xkbar

Model:

$X_{i,j} = \mu + \alpha_i + \varepsilon_{i,j}$ < where:

μ is the grand mean of all objects.
α_i is the mean of i = $\mu + \alpha_i$ for each class i.
$\varepsilon_{i,j}$ is the error term specific to each object i,j

Restriction:

$\sum_i n_i \cdot \alpha_i := 0$ or $\sum_i \alpha_i := 0$ or $\alpha_k := 0$ < allows estimation of k parameters.

See Rosner 2006 p. 558

Assumptions:

ε_{ij} are a random sample ~ $N(0,\sigma^2)$

Zar Example 10.3

```
#ONE-WAY ANOVA TABLE
#MULTIPLE t-TESTS
ZAR=read.table("c:/DATA/Biostatistics/ZarEX10.3R.txt")
ZAR
attach(ZAR)
Y=potassium
X=factor(variety)
options(digits=12)
#DESCRIPTIVE STATISTICS:
X1bar=mean(Y[X=="1"])
X1bar
X2bar=mean(Y[X=="2"])
X2bar
X3bar=mean(Y[X=="3"])
X3bar
n1=length(Y[X=="1"])
n1
n2=length(Y[X=="2"])
n2
n3=length(Y[X=="3"])
n3
s1=sqrt(var(Y[X=="1"]))
s1
s2=sqrt(var(Y[X=="2"]))
s2
s3=sqrt(var(Y[X=="3"]))
s3
#GENERATING ANOVA TABLE:
anova(lm(Y~X))
```

$k := 3$ < number of classes

$N := 18$ < total number of cases

$$X_{bar} := \begin{pmatrix} 26.9833333333 \\ 25.6666666667 \\ 29.55 \end{pmatrix} \quad n := \begin{pmatrix} 6 \\ 6 \\ 6 \end{pmatrix} \quad s := \begin{pmatrix} 0.679460570355 \\ 1.13078144072 \\ 1.26767503722 \end{pmatrix}$$

^ descriptive statistics derived from R

```
Analysis of Variance Table

Response: Y
          Df      Sum Sq     Mean Sq  F value     Pr(>F)
X          2 46.80333333 23.40166667 20.97341 4.5151e-05 ***
Residuals 15 16.73666667  1.11577778
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

One-Way ANOVA Table:

Values taken from R's ANOVA table:

Source:	SS	df		MS	
Between	$SS_B := 46.80333333$	$df_B := k - 1$	$df_B = 2$	$MS_B := \frac{SS_B}{df_B}$	$MS_B = 23.4017$
Within	$SS_W := 16.73666667$	$df_W := N - k$	$df_W = 15$	$MS_W := \frac{SS_W}{df_W}$	$MS_W = 1.1158$
TOTAL	$SS_T := SS_B + SS_W$				
	$SS_T = 63.54$				

Multiple t-Test / Fisher's LSD Test for Specific Treatment Pairs:

This test, offers a "solution" to the problem of distinguishing single Type I error α versus familywide α_{fw} in multiple comparisons that ignores the issue entirely! According to Sheskin 1997 *Handbook of Parametric and Nonparametric Statistical Procedures*, "Multiple t-Tests" is the term often employed for planned data, whereas "Fisher's Least Significant Difference Tests" often refers to comparisons made *post hoc*. However, the two are computationally equivalent. Among MCP's this procedure is the most powerful and most capable of detecting differences between means but, of course, also with the highest level of familywise error α_{fw}. Multiple t-Test/Fisher's LSD is often employed when the researcher feels that "data snooping" is not a major issue in the study and/or the number of multiple comparisions are relatively low. Of course, this is a judgement call. So if the data permits, use of one of the procedures below is more "conservative" and is often judged to be more prudent. Studies may report both - if it matters.

Hypotheses:

H_0: $\alpha_i = \alpha_j$ ***for specific*** **i & j** **< Means in treatment classes i & j are the same**

H_1: $\alpha_i \neq \alpha_j$ ***for specific*** **i & j** **< Two sided test**

Test Statistic:

$$t_1 := \frac{X_{bar_1} - X_{bar_2}}{\sqrt{MS_W \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad t_1 = 2.159$$

< Comparing sample 1 with 2

$$t_2 := \frac{X_{bar_1} - X_{bar_3}}{\sqrt{MS_W \cdot \left(\frac{1}{n_1} + \frac{1}{n_3}\right)}} \qquad t_2 = -4.2086$$

< Comparing sample 1 with 3

$$t_3 := \frac{X_{bar_2} - X_{bar_3}}{\sqrt{MS_W \cdot \left(\frac{1}{n_2} + \frac{1}{n_3}\right)}} \qquad t_3 = -6.3676$$

< Comparing sample 2 with 3

^ Note that pooled variance MS_W is used here, making the calculation of t statistics a little different from the 2-populations calculation

Sampling Distribution of the test Statistic t:

If Assumption hold and H_0 is true then $t \sim t_{(N-k)}$

where: k = number of classes
N = total number of observations

Critical Value of the Test:

$\alpha := 0.05$ **< Probability of Type I error must be explicitly set**

$C_1 := qt\left(\frac{\alpha}{2}, N-k\right)$ $C_1 = -2.1314$

$C_2 := qt\left(1-\frac{\alpha}{2}, N-k\right)$ $C_2 = 2.1314$

$C := |C_1|$ $C = 2.1314$

< Note: degrees of freedom = (N-k). This also differs from the 2-populations calculation

$N - k = 15$

Decision Rule:

IF |t| > C, THEN REJECT H_0 OTHERWISE ACCEPT H_0

Probability Value:

For sample 1 vs 2:

$t_1 = 2.159$

$P := 2 \cdot (1 - pt(t_1, N-k))$ $P = 0.04746$ **< if t > 0**

$P := 2 \cdot pt(t_1, N-k)$ $P = 1.95254$ **< if** $t \le 0$ **< Two sided case**

For sample 1 vs 3:

$t_2 = -4.2086$

$P := 2 \cdot (1 - pt(t_2, N-k))$ $P = 1.99924$ **< if t > 0**

$P := 2 \cdot pt(t_2, N-k)$ $P = 0.00076$ **< if** $t \le 0$ **< Two sided case**

For sample 2 vs 3:

$t_3 = -6.3676$

$P := 2 \cdot (1 - pt(t_3, N-k))$ $P = 1.999987$ **< if t > 0**

$P := 2 \cdot pt(t_3, N-k)$ $P = 1.263977 \times 10^{-5}$ **< if** $t \le 0$ **< Two sided case**

Prototype in R:

```
#PAIRWISE COMPARISON OF MEANS:
#MULTIPLE t-TEST/FISHER'S LSD:
pairwise.t.test(Y,X,p.adj="none",alternative="two.sided")
```

```
        Pairwise comparisons using t tests with pooled SD

data:  Y and X

  1       2
2 0.04746 -
3 0.00076 1.3e-05
```

Multiple t-Test / Fisher's LSD Confidence Intervals:

For sample 1 vs 2:

$$CI_1 := \left[X_{bar_1} - X_{bar_2} - C \cdot \sqrt{MS_W \cdot \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \quad X_{bar_1} - (X_{bar})_2 + C \cdot \sqrt{MS_W \cdot \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \right]$$

$$CI_1 = (0.0168 \quad 2.6165)$$

For sample 1 vs 3:

$$CI_2 := \left[X_{bar_1} - X_{bar_3} - C \cdot \sqrt{MS_W \cdot \left(\frac{1}{n_1} + \frac{1}{n_3} \right)} \quad X_{bar_1} - (X_{bar})_3 + C \cdot \sqrt{MS_W \cdot \left(\frac{1}{n_1} + \frac{1}{n_3} \right)} \right]$$

$$CI_2 = (-3.8665 \quad -1.2668)$$

For sample 2 vs 3:

$$CI_3 := \left[X_{bar_2} - X_{bar_3} - C \cdot \sqrt{MS_W \cdot \left(\frac{1}{n_2} + \frac{1}{n_3} \right)} \quad X_{bar_2} - (X_{bar})_3 + C \cdot \sqrt{MS_W \cdot \left(\frac{1}{n_2} + \frac{1}{n_3} \right)} \right]$$

$$CI_3 = (-5.1832 \quad -2.5835)$$

Bonferroni Multiple Comparisons Procedure:

If a specific and relatively small set of simultaneous tests are desired, this procedure will often give the narrowest confidence intervals among MCP's that worry about α_{fw} and is often preferred for that reason. Since the Bonferroni method requires identifying a specific set of simultaneous tests, it is not considered appropriate for *post hoc* "data snooping". The Bonferroni approach of adjusting α_{fw} in a way that's directly related to a specified number of multiple comparisons is very widespread among statistical methods.

Hypotheses:

H_0: $\alpha_i = \alpha_j$ ***for specific*** **i & j** **< Means in treatment classes i & j are the same**

H_1: $\alpha_i \neq \alpha_j$ ***for specific*** **i & j** **< Two sided test**

Test Statistic:

Same as calculated above:

$t_1 = 2.159$ **< comparing sample 1 with 2**

$t_2 = -4.2086$ **< comparing sample 1 with 3**

$t_3 = -6.3676$ **< comparing sample 2 with 3**

< Note that pooled variance MS_W is used here, making the calculation of t statistics a little different from the 2-populations calculation

Sampling Distribution of the test Statistic t:

If Assumptions hold and H_0 is true then $t \sim t_{(N-k)}$ **where: k = number of classes**
N = total number of observations

Critical Value of the Test:

$\alpha := 0.05$ **< Probability of Type I error must be explicitly set**

$g := 3$ **< Bonferroni's factor = number of planned comparisons.**
Here for three pairwise comparison of means

$C_1 := qt\left(\frac{\alpha}{2 \cdot g}, N - k\right)$ $C_1 = -2.6937$ **< Note: degrees of freedom = (N-k).**

$C_2 := qt\left(1 - \frac{\alpha}{2 \cdot g}, N - k\right)$ $C_2 = 2.6937$ **< Note also the bonferroni factor g.**

$N - k = 15$

$C_B := |C_1|$ $C_B = 2.6937$

Decision Rule:

IF $|t| > C_B$, THEN REJECT H_0 OTHERWISE ACCEPT H_0

Probability Values:

For sample 1 vs 2:

Note below the Bonferroni factor g:

$t_1 = 2.159$

$P := 2 \cdot g \cdot (1 - pt(t_1, N - k))$ $P = 0.14238$ **< if t > 0**

$P := 2 \cdot g \cdot pt(t_1, N - k)$ $P = 5.85762$ **< if** $t \le 0$ **< Two sided case**

For sample 1 vs 3:

$t_2 = -4.2086$

$P := 2 \cdot g \cdot (1 - pt(t_2, N - k))$ $P = 5.997721$ **< if t > 0**

$P := 2 \cdot g \cdot pt(t_2, N - k)$ $P = 0.002279$ **< if** $t \le 0$ **< Two sided case**

For sample 2 vs 3:

$t_3 = -6.3676$

$P := 2 \cdot g \cdot (1 - pt(t_3, N - k))$ $P = 5.999962$ **< if t > 0**

$P := 2 \cdot g \cdot pt(t_3, N - k)$ $P = 3.79193 \times 10^{-5}$ **< if** $t \le 0$ **< Two sided case**

Prototype in R:

```
#BONFERRONI COMPARISON OF MEANS:
pairwise.t.test(Y,X,p.adj="bonferroni",alternative="two.sided")
```

```
        Pairwise comparisons using t tests with pooled SD

data:  Y and X
  1      2
2 0.1424 -
3 0.0023 3.8e-05
P value adjustment method: bonferroni
```

Bonferroni Confidence Intervals:

For sample 1 vs 2:

$$CI_{B1} := \left[X_{bar_1} - X_{bar_2} - C_B \cdot \sqrt{MS_W \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \quad X_{bar_1} - (X_{bar})_2 + C_B \cdot \sqrt{MS_W \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \right]$$

$$CI_1 = (0.0168 \quad 2.6165)$$

$$CI_{B1} = (-0.3261 \quad 2.9595)$$

For sample 1 vs 3:

$$CI_{B2} := \left[X_{bar_1} - X_{bar_3} - C_B \cdot \sqrt{MS_W \cdot \left(\frac{1}{n_1} + \frac{1}{n_3}\right)} \quad X_{bar_1} - (X_{bar})_3 + C_B \cdot \sqrt{MS_W \cdot \left(\frac{1}{n_1} + \frac{1}{n_3}\right)} \right]$$

$$CI_2 = (-3.8665 \quad -1.2668)$$

$$CI_{B2} = (-4.2095 \quad -0.9239)$$

For sample 2 vs 3:

$$CI_{B3} := \left[X_{bar_2} - X_{bar_3} - C_B \cdot \sqrt{MS_W \cdot \left(\frac{1}{n_2} + \frac{1}{n_3}\right)} \quad X_{bar_2} - (X_{bar})_3 + C_B \cdot \sqrt{MS_W \cdot \left(\frac{1}{n_2} + \frac{1}{n_3}\right)} \right]$$

$$CI_3 = (-5.1832 \quad -2.5835)$$

$$CI_{B3} = (-5.5261 \quad -2.2405)$$

^ Note: Bonferroni intervals are wider than Fisher's LSD

Holm Simultaneous Testing Procedure:

This procedure is an iterative refinement of the Bonferroni approach designed to provide a simultaneous probability of α_{fw} for a specific set of tests. Holm sometimes rejects a null hypothesis that Bonferroni would not with the same data and is, thus, more powerful. However, Holm is computationally more complex and lacks direct computation of Confidence Intervals. Although Holm may be the preferred method for theoretical reasons, power consideration by itself may not necessarily be a good reason for chosing the test. As with Bonferroni, this method is unsuitable for "data snooping".

```
#HOLM COMPARISON OF MEANS:
pairwise.t.test(Y,X,p.adj="holm",alternative="two.sided")
```

```
        Pairwise comparisons using t tests with pooled SD

data:  Y and X

  1      2
2 0.0475 -
3 0.0015 3.8e-05

P value adjustment method: holm
```

Tukey Multiple Pairwise Comparisons Procedure:

This procedure is designed to provide a simultaneous probability of α_{fw} when comparing means of all possible pairs of populations within the ANOVA data structure. When sample sizes n_i differ, this procedure is also called the Tukey-Kramer Procedure. Analysis *post hoc* is permitted with this procedure as long as one is restricts "snooping" to pairwise comparisons of population means.

Tukey Test Statistic:

For sample 1 vs 2:

$$q_1 := \left| \frac{\sqrt{2} \cdot \left(X_{bar_1} - X_{bar_2} \right)}{\sqrt{MS_W \cdot \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \right| \qquad q_1 = 3.053251732725$$

For sample 1 vs 3:

$$q_2 := \left| \frac{\sqrt{2} \cdot \left(X_{bar_1} - X_{bar_3} \right)}{\sqrt{MS_W \cdot \left(\frac{1}{n_1} + \frac{1}{n_3} \right)}} \right| \qquad q_2 = 5.951908441386$$

For sample 2 vs 3:

$$q_3 := \left| \frac{\sqrt{2} \cdot \left(X_{bar_2} - X_{bar_3} \right)}{\sqrt{MS_W \cdot \left(\frac{1}{n_2} + \frac{1}{n_3} \right)}} \right| \qquad q_3 = 9.00516017411$$

Sampling Distribution of the test Statistic q:

If Assumptions hold and H_0 is true then $q \sim q_{(k,N-k)}$

where: k = number of classes
N = total number of observations

^ "studentized range distribution"

Critical Value of the Test:

$\alpha := 0.05$ **< Probability of Type I error must be explicitly set**

$C_T := \frac{1}{\sqrt{2}} \cdot qstudentizedrange(1 - \alpha, k, N - k)$ **< Studentized range distribution doesn't exist in MathCad, so, I'll calculate it using the function qtucky() in R.**

$N = 18$ $k = 3$

```
#CALCULATING C FROM THE STUDENTIZED RANGE DISTRIBUTION:
alpha=0.05
N=18
k=3
qtukey(1-alpha,k,N-k)
```

[1] 3.67337762497

$C_T := \frac{1}{\sqrt{2}} \cdot 3.67337762497$ $C_T = 2.5975$ **< critical value constructed from "studentized" range distribution.**

Decision Rule:

IF $|q| > C_T$, THEN REJECT H_0 OTHERWISE ACCEPT H_0

Probability Values:

```
#CALCULATING P FROM THE STUDENTIZED RANGE DISTRIBUTION:
#q-STATISTICS FROM MATHCAD:
q1=3.053251732725
q2=5.951908441386
q3=9.00516017411
options(digits=12)

#FOR SAMPLE 1 VS 2:          [1] 0.111388393096
P1=1-ptukey(q1,k,N-k)
P1
#FOR SAMPLE 1 VS 3:          [1] 0.00206212040461
P2=1-ptukey(q2,k,N-k)
P2
#FOR SAMPLE 2 VS 3:          [1] 3.55985514127e-05
P3=1-ptukey(q3,k,N-k)
P3
```

Tukey Confidence Interval for Multiple Pairwise Comparisons:

For sample 1 vs 2:

$$CI_{T1} := \left[X_{bar_1} - X_{bar_2} - C_T \cdot \sqrt{MS_W \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \quad X_{bar_1} - (X_{bar})_2 + C_T \cdot \sqrt{MS_W \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \right]$$

$CI_1 = (0.0168 \quad 2.6165)$

$CI_{B1} = (-0.3261 \quad 2.9595)$

$CI_{T1} = (-0.2674 \quad 2.9008)$

For sample 1 vs 3:

$$CI_{T2} := \left[X_{bar_1} - X_{bar_3} - C_T \cdot \sqrt{MS_W \cdot \left(\frac{1}{n_1} + \frac{1}{n_3}\right)} \quad X_{bar_1} - (X_{bar})_3 + C_T \cdot \sqrt{MS_W \cdot \left(\frac{1}{n_1} + \frac{1}{n_3}\right)} \right]$$

$CI_2 = (-3.8665 \quad -1.2668)$

$CI_{B2} = (-4.2095 \quad -0.9239)$

$CI_{T2} = (-4.1508 \quad -0.9826)$

For sample 2 vs 3:

$$CI_{T3} := \left[X_{bar_2} - X_{bar_3} - C_T \cdot \sqrt{MS_W \cdot \left(\frac{1}{n_2} + \frac{1}{n_3}\right)} \quad X_{bar_2} - (X_{bar})_3 + C_T \cdot \sqrt{MS_W \cdot \left(\frac{1}{n_2} + \frac{1}{n_3}\right)} \right]$$

$CI_3 = (-5.1832 \quad -2.5835)$

$CI_{B3} = (-5.5261 \quad -2.2405)$

$CI_{T3} = (-5.4674 \quad -2.2992)$

Prototype in R:

#TUKEY HSD TEST:
TukeyHSD(aov(Y~X),conf.level=0.95)

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Y ~ X)

$X
              diff             lwr            upr          p adj
2-1 -1.31666666667 -2.900752846105 0.267419512772 0.111388393032
3-1  2.56666666667  0.982580487228 4.150752846105 0.002062120403
3-2  3.88333333333  2.299247153895 5.467419512772 0.000035598551
```

^ P values and confidence intervals match, signs of values differ for CI_T