Multiple Comparisons & Statistical Power (MD4 & 5)

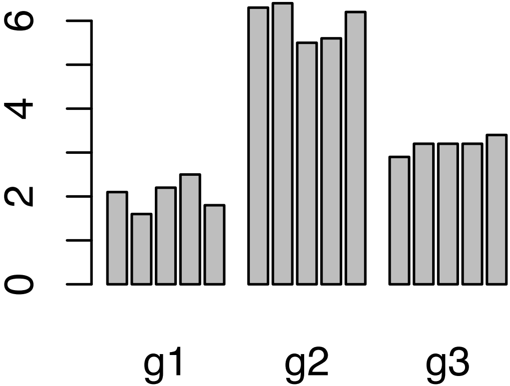
Paul Gribble

Winter, 2017

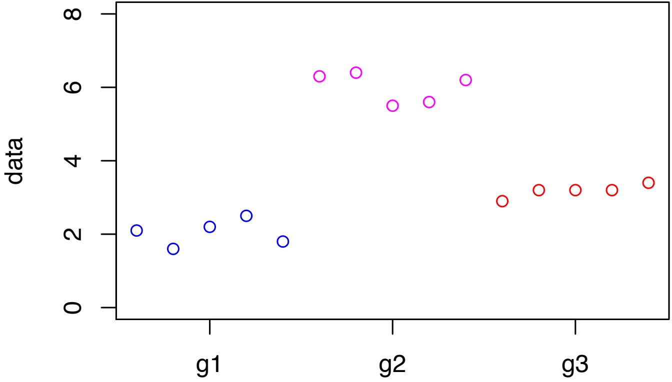
GLM & ANOVA: an example

G1	G2	G3
2.1	6.3	2.9
1.6	6.4	3.2
2.2	5.5	3.2
2.5	5.6	3.2
1.8	6.2	3.4
means		
2.0	6.0	3.2

GLM & ANOVA: an example

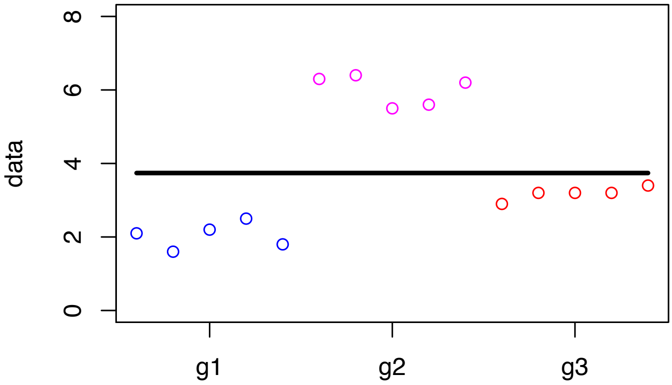


the model comparison approach: restricted model



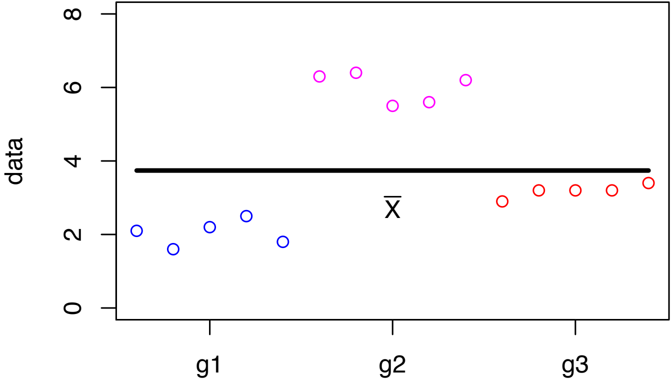
$$H_0 : Y_{ij} = \mu + \epsilon_{ij} \qquad E_r = \sum \left(Y_{ij} - \bar{X} \right)^2$$

the model comparison approach: restricted model



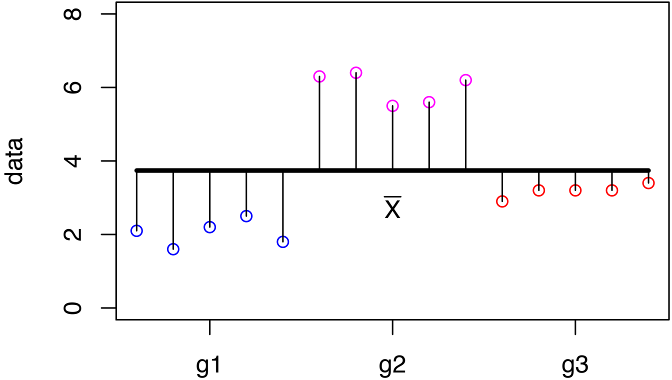
$$H_0 : Y_{ij} = \mu + \epsilon_{ij} \qquad E_r = \sum \left(Y_{ij} - \bar{X} \right)^2$$

the model comparison approach: restricted model



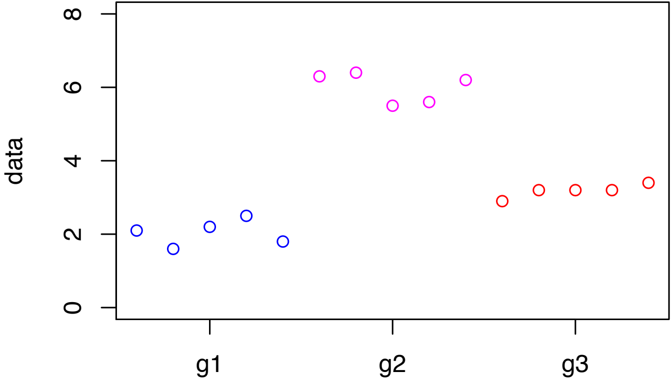
$$H_0 : Y_{ij} = \mu + \epsilon_{ij} \qquad E_r = \sum \left(Y_{ij} - \bar{X}\right)^2$$

the model comparison approach: restricted model



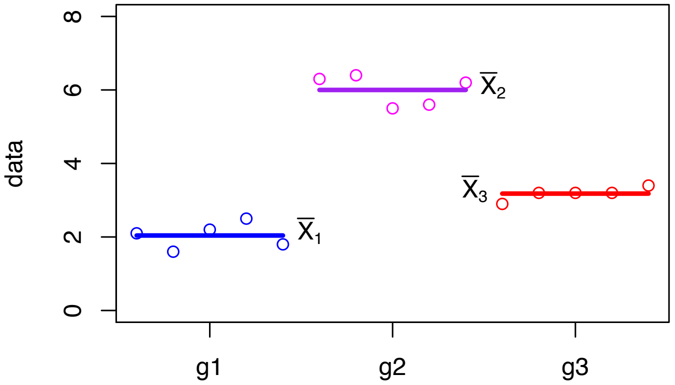
$$H_0 : Y_{ij} = \mu + \epsilon_{ij} \qquad E_r = \sum \left(Y_{ij} - \bar{X} \right)^2$$

the model comparison approach: full model



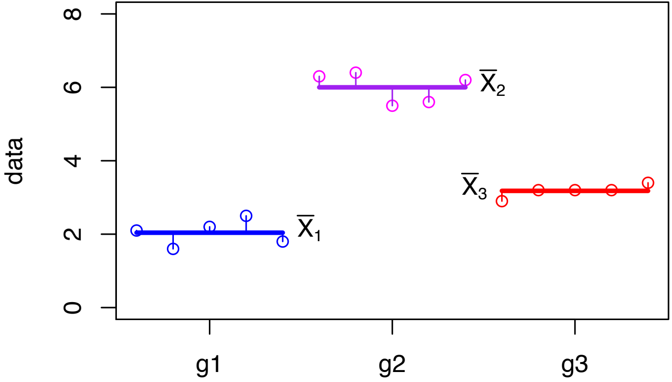
$$H_1 : Y_{ij} = \mu_j + \epsilon_{ij} \qquad E_f = \sum \left(Y_{ij} - \bar{X}_j\right)^2$$

the model comparison approach: full model



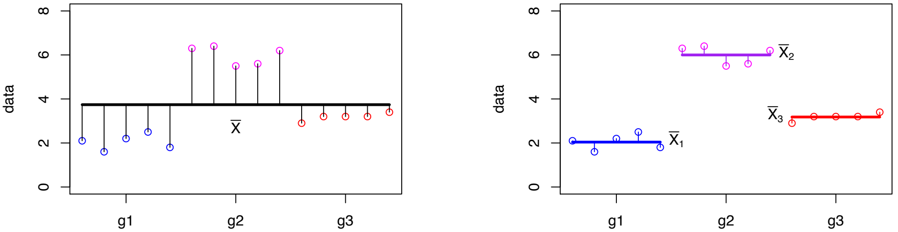
$$H_1 : Y_{ij} = \mu_j + \epsilon_{ij} \qquad E_f = \sum \left(Y_{ij} - \bar{X}_j \right)^2$$

the model comparison approach: full model



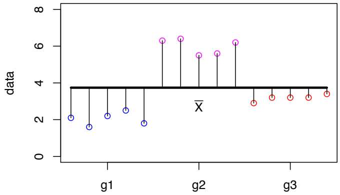
$$H_1 : Y_{ij} = \mu_j + \epsilon_{ij} \qquad E_f = \sum \left(Y_{ij} - \bar{X}_j\right)^2$$

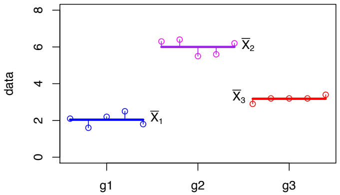
which model has smaller error?



- estimate 1 parameter
  - μ
- estimate 3 parameters
  - μ_1, μ_2, μ_3

which model has smaller error?





- Is the reduction in error you get with the full model worth the extra parameters you need to estimate in H_1?

Statistical Power

- power is the ability of a statistical test to detect real differences when they exist
- β is the probability of failing to reject the null hypothesis when it is in fact false (Type-II error)
- β is the probability of failing to reject the restricted model when the full model is a better description of the data, even with the requirement to estimate more parameters

$$\text{power} = 1 - \beta$$

- power is the probability of rejecting the null hypothesis when it is in fact false

Type-I vs Type-II error \ hypothesis testing outcomes

		Reality	
		H_0 is true	H_1 is true
Research	H_0 is true	Accurate $(1-\alpha)$	Type-II error (β)
	H_1 is true	Type-I error (α)	Accurate $(1-\beta)$

Statistical Power

- how sensitive is a given experimental design?
- how likely is our experiment to correctly identify a difference betweeen groups when there actually is one?
- what sample size is required to give an experiment adequate power?
- how many subjects do we need to include in each group sample?

Effect Size

- we need some way of assessing the expected size of the effect we are proposing to detect
- one measure is the standardized measure of effect size, f

$$
\begin{aligned}
f &= \sigma_m / \sigma_\epsilon \\
\sigma_m &= \sqrt{\frac{\sum(\mu_j - \mu)^2}{a}} = \sqrt{\frac{\sum \alpha_j^2}{a}} \\
\mu &= \left(\sum_j \mu_j \right) / a \\
\sigma_\epsilon &= \text{within-group standard deviation}
\end{aligned}
$$

Effect Size

- If you have pilot data you can compute values for f
- If not, Cohen (1977) suggests the following definitions:
  - "small" effect: $f = 0.10$
  - "medium" effect: $f = 0.25$
  - "large" effect: $f = 0.40$
- so for medium effect, standard deviation of population means across groups is $1/4$ of the within-group sd

Power Charts

- Cohen (1977) provides tables that let you read off the power for a particular combination of numerator df, desired Type-I error rate, effect size f, and subjects per group
- four factors are varying — tables require 66 pages!
  - seriously
- It's 2015, Let's use `R` instead
  - `power.t.test()`
  - `power.anova.test()`

An example

- e.g. you are planning a reaction-time study involving three groups ($a = 3$)
- pilot research & data from literature suggest population means might be 400, 450 and 500 ms with a sample within-group standard deviation of 100 ms
- suppose you want a power of 0.80 — how many subjects do you need in each sample group?

An example

```
power.anova.test(groups=3, n=NULL,
  between.var=var(c(400,450,500)),
  within.var=100**2, sig.level=0.05,
  power=0.80)

     Balanced one-way analysis of variance power calcul

         groups = 3
              n = 20.30205
    between.var = 2500
     within.var = 10000
      sig.level = 0.05
          power = 0.8

NOTE: n is number in each group
```

. . . but since we know how to program in R

- simulate! Simulate sampling from two populations
  - whose means differ by the expected amount
  - whose variances are a particular value
  - postulate a particular sample size N
- sample and do your statistical test many times (e.g. 1000) and see what proportion of times you successfully reject the null (your power)
- If power is not high enough, try a larger sample size N and repeat. Keep increasing N in simulation until you get the power you want
- computationally intensive, but allows you to test any experimental situation that you can simulate
- e.g. see `http://goo.gl/C0mI0`

Cautionary note: calculating "observed power" after rejecting the null

- you run an experiment, do stats, and end up failing to reject H_0
- two possibilities:
  1. there is in fact no difference between population means, and your experiment correctly identifies this
  2. there is a difference, but your experiment is not statistically powerful enough to detect it (for e.g. because within-group variability is high)
- can we use power calculations to see if we "had enough power" to detect the difference?
- no — not appropriate use of power analysis (although frequently taught)

Hoenig & Heisey (2001)

- doing a power analysis after an experiment that failed to reject the null, to see if "there was enough power" to detect the difference, is inappropriate
- the result of a post-hoc power analysis is completely redundant with the probability (p-value) obtained in the original analysis
- one can be obtained directly from the other
- you don't learn anything new by doing a post-hoc power analysis
- See Hoenig & Heisey (2001) for the full story

Challenges of power analyses

- you must have estimates of expected difference between means
- you must have estimates of within-group variability
- computing power for more complex experimental designs can be complicated — see Maxwell & Delaney text for examples

Testing differences between individual means

- last time we learned about one-way single-factor ANOVA
- F test of null hypothesis
  - $\mu_1 = \mu_2 = ... = \mu_n$
- called the "omnibus test"
- omnibus test doesn't tell us *which* means are different from each other
- it *does* give us permission to start looking for differences between individual means

Two kinds of multiple comparisons

planned comparisons

- in advance of looking at your results you know which groups you want to compare
- you are restricted to performing only certain comparisons
- the comparisons must be *orthogonal* to each other

post-hoc comparisons

- the results dictate which means you test (you are *chasing the biggest differences*)
- you can test as many as you like (usually)
- few (if any) restrictions on the nature of the tests you can perform
- Type-I error is controlled for by making each test more conservative

Model comparison approach

- recall the null hypothesis & restricted model:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \cdots = \mu_a \\ Y_{ij} &= \mu + \epsilon_{ij} \end{aligned}$$

- suppose we wanted to test a new hypothesis that only groups 1 and 2 are equal and the rest are different

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ Y_{i1} &= \mu^* + \epsilon_{i1} \\ Y_{i2} &= \mu^* + \epsilon_{i2} \\ Y_{ij} &= \mu_j + \epsilon_{ij}, \quad \text{for} \quad j = 3, 4, \ldots, a \end{aligned}$$

Model comparison approach

- just as before we can compare full and restricted models by computing sums of squared errors for each (see Maxwell & Delaney for details)
- just as before we end up with an F ratio:

$$
\begin{aligned}
F &= \frac{(E_R - E_F)/(df_R - df_F)}{E_F/df_F} \\
E_R - E_F &= \frac{n_1 n_2}{n_1 + n_2} \left(\bar{Y}_1 - \bar{Y}_2 \right)^2 \\
df_F &= N - a \\
df_R &= N - (a - 1) = N - a + 1 \\
df_R - df_F &= 1
\end{aligned}
$$

Model comparison approach

- after some more tedious algebra:

$$F = \frac{n_1 n_2 \left(\bar{Y}_1 - \bar{Y}_2\right)^2}{\left(n_1 + n_2\right) MS_W}$$

- or for equal group sizes n:

$$F = \frac{n \left(\bar{Y}_1 - \bar{Y}_2\right)^2}{2MS_W}$$

- MS_W is mean-square "within" term (error term) from ANOVA output
- df numerator $= 1$
- df denominator is given in ANOVA output for MS_W term

Model comparison approach

- so what we have now is an F test for a full versus restricted model
- full model is as before (different mean for each group)
- restricted model has same mean for groups 1 and 2, and different means for the rest
- restricted model is less restricted than the original restricted model with a single parameter (the grand mean)
- but still more restricted than full model

$$F = \frac{n\left(\bar{Y}_1 - \bar{Y}_2\right)^2}{2MS_W}$$

Complex comparisons

- research questions often focus on pairwise comparisons
- sometimes you may have a hypothesis that concerns a difference involving more than 2 means
- e.g. 4 groups: is group 4 different than the average of the other three?

$$H_0 : \frac{1}{3}\left(\mu_1 + \mu_2 + \mu_3\right) = \mu_4$$

- we can rewrite this as:

$$H_0 : \frac{1}{3}\mu_1 + \frac{1}{3}\mu_2 + \frac{1}{3}\mu_3 - \mu_4 = 0$$

Complex comparisons

$$H_0 : \frac{1}{3}\mu_1 + \frac{1}{3}\mu_2 + \frac{1}{3}\mu_3 - \mu_4 = 0$$

- this is just a linear combination of the 4 means so in general we can write:
$$H_0 : c_1\mu_1 + c_2\mu_2 + c_3\mu_3 + c_4\mu_4 = 0$$

- c_1 through c_4 are coefficients chosen by the experimenter to test a hypothesis of interest
- simple pairwise comparison of mean 1 vs mean 2 would be:
$$\begin{aligned} c_1 &= -1 \\ c_2 &= +1 \\ c_3 &= 0 \\ c_4 &= 0 \end{aligned}$$

Complex comparisons

an expression of the form:

$$H_0 : c_1\mu_1 + c_2\mu_2 + c_3\mu_3 + c_4\mu_4$$

is known as a "contrast" or a "complex comparison"

- linear combination of means in which *the coefficients add up to zero*
- in the general case of a groups, we can write:

$$\psi = \sum_{j=1}^{a} c_j\mu_j$$

Complex comparisons

- our expression for the F test can be simplified (see M&D) to:

$$F = \frac{\psi^2}{MS_W \sum_{j=1}^{a} \left(c_j^2 / n_j\right)}$$

where

- df denominator $= 1$
- df numerator $= N - a$

$$H_0 : \psi = \sum_{j=1}^{a} c_j \mu_j = 0$$

Complex comparisons

- some texts present contrasts not as F tests but as t-test
- when *df* $\mathrm{numerator} = 1$, t-test is just a special case of the F-test

$$
\begin{aligned}
t^2 &= F \\
t &= \sqrt{F}
\end{aligned}
$$

Testing more than one contrast

- how many contrasts can we test?
- two issues:
  1. orthogonality
  2. inflation of Type-I error
- is it permissible to perform multiple tests using an α level of 0.05?
  - better question: does it make sense to perform multiple tests and still assume that Type-I error rate remains at 0.05?
- does it matter if the contrasts were planned before the data were examined, or arrived at after looking at the data?

How many contrasts?

- if $a = 3$ there are 3 possible pairwise contrasts (`choose(3,2)`)
  - 1-2, 2-3 and 1-3
  - in addition there are an infinite of possible complex comparisons
- with an infinite $\backslash$ contrasts, some information will be redundant
- new question: how many contrasts can be tested without introducing redundancy?

Non-redundant contrasts

- are these three contrasts redundant?

$$
\begin{aligned}
\psi_1 &= \mu_1 - \mu_2 \\
\psi_2 &= \mu_1 - \mu_3 \\
\psi_3 &= \frac{1}{2}(\mu_1 + \mu_2) - \mu_3
\end{aligned}
$$

- yes, because:

$$\psi_3 = \psi_2 - \frac{1}{2}\psi_1$$

- value of ψ_3 is compelely determined if we already know ψ_1 and ψ_2

Non-redundant contrasts

- in general with a groups, there are $a-1$ contrasts without introducing redundancy
- mathematical concept for lack of redundancy is orthogonality
- two contrasts are orthogonal if:

$$\begin{aligned}
\psi_1 &= \sum c_{1j}\mu_j \\
\psi_2 &= \sum c_{2j}\mu_j \\
\sum c_{1j}c_{2j} &= 0
\end{aligned}$$

- or for unequal group sizes:

$$\sum c_{1j}c_{2j}/n_j = 0$$

Orthogonal contrasts

- e.g. what about 2 contrasts c_1 and c_2:
- $c_{11} = +1$, $c_{12} = -1$, $c_{13} = 0$
- $c_{21} = +1$, $c_{22} = 0$, $c_{23} = -1$
- orthogonality test: $\sum c_{1j} c_{2j} = 0$
  - (1)(1) + (-1)(0) + (0)(-1) = 1 + 0 + 0 = 1
  - these 2 contrasts are not orthogonal

Orthogonality

- who cares?
- primary implication: orthogonal contrasts provide non-overlapping information about how the groups differ
- formally: when two contrasts are orthogonal, then the two sample estimates ψ_1 and ψ_2 are statistically independent of one another
- each provides unique, non-overlapping information about group differences
- they are asking separate, different, distinct questions about the data

Testing multiple comparisons

- suppose you have conducted an ANOVA on 4 groups
- suppose you want to test the following 3 contrasts:

$$\begin{aligned}
\psi_1 &= \mu_1 - \mu_2 \\
\psi_2 &= \frac{1}{2}(\mu_1 + \mu_2) - \mu_3 \\
\psi_3 &= \frac{1}{3}(\mu_1 + \mu_2 + \mu_3) - \mu_4
\end{aligned}$$

- are these orthogonal?
  - ψ_1: (+1.0)(-1.0)(+0.0)(+0.0)
  - ψ_2: (+0.5)(+0.5)(-1.0)(+0.0)
  - ψ_3: (+0.3)(+0.3)(+0.3)(-1.0)

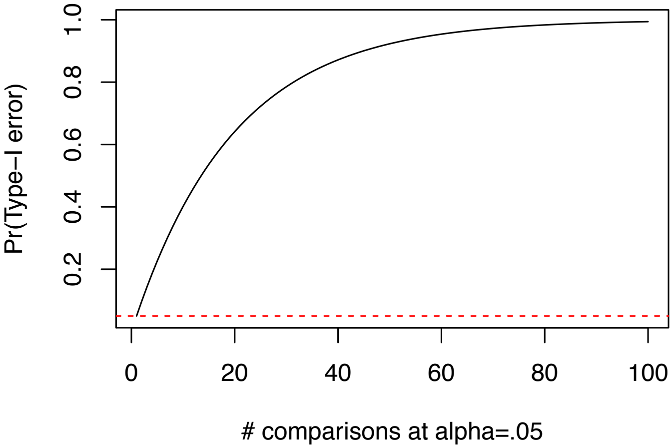
Testing multiple comparisons

- if you test each of the three contrasts at $\alpha = 0.05$, what is the true Type-I error rate?
- greater than 0.05
- we are testing three contrasts each at the 0.05 level
- at first glance you might think true error rate should be $(3)(0.05) = 0.15$
- close, but not quite right

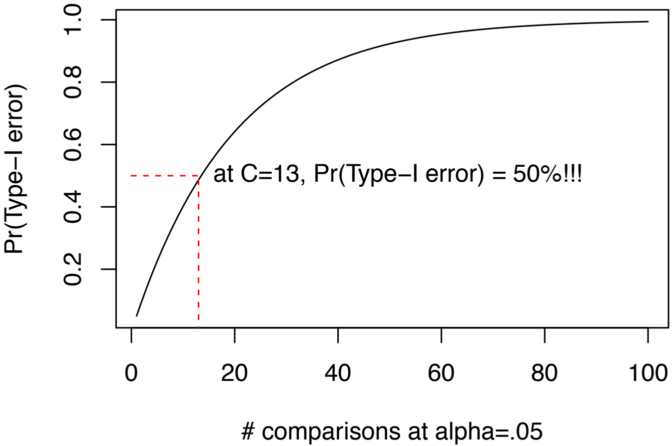
Testing multiple comparisons

- contrasts are independent events
- probabilities don't simply sum (see M&D text)
- Pr(at least one Type-I error) $=$ 1 - Pr(no Type-I errors)
- $= 1 - (1 - \alpha)^C$
- C is number of contrasts tested
- e.g. if $\alpha = 0.05$, $C = 3$, then $p = 0.143$
- if $C = 10$, $p = 0.40$ (big!)

Testing multiple comparisons



Testing multiple comparisons



Testing multiple comparisons

- is this a problem? Pr(Type-I error) $>$ 0.05 ???
- M&D text discusses some different concepts:
- error rate per contrast α_{PC}
  - probability that a particular contrast will be falsely declared significant
- experiment-wise error rate α_{EW}
  - probability that one or more contrasts will be falsely declared significant in an experiment
- family-wise error rate α_{FW}
  - has to do with multiple factor experiments (more later in the course)

Testing multiple comparisons

- In our example, $\alpha_{PC} = 0.05$
- experiment-wise error rate $\alpha_{EW} = 0.143$
- so which error rate should be controlled at the 0.05 level?
- this is an issue "about which reasonable people differ"
  - i.e. intelligent and informed people have different approaches
- M&D suggest controlling α_{EW} at the 0.05 level
- see chapter for an interesting discussion of the pros and cons of different approaches

Methods of controlling α_{EW} at 0.05

- planned vs post-hoc comparisons
- 3 methods
  - Bonferroni, Tukey, Scheffe
- M&D have a flowchart (decision tree) to help you decide which procedure to use

Planned vs Post-hoc contrasts

1. Planned Contrast
   - a contrast that an experimenter decided to test prior to any examination of the data
   - (i.e. the data do not influence your choice of which contrast(s) to test)
2. Post-Hoc Contrast
   - a contrast that an experimenter decided to test only after having looked at the data
   - i.e. a contrast "suggested by the data"
   - e.g. following large differences you observe in your dataset

Planned vs Post-hoc contrasts

- why is this distinction important?
- If the contrast(s) to be tested are suggested by the data, e.g. the largest differences are tested
- the sampling distribution of "differences between any 2 means" has a very different distribution than the "largest difference between means"
- Type-I error rate ends up being inflated if you only test the largest differences in your dataset
- M&D have a nice discussion of this in the chapter
- we will show it in R using monte-carlo simulations

Multiple Planned Comparisons

- The Bonferroni adjustment is remarkable simple
- compute the F statistic and p-value for each contrast, as usual
- then instead of comparing each p-value to α (e.g. 0.05), instead compare it to $\frac{\alpha}{C}$, where C is the total number of contrasts you will be testing
- α gets lowered in proportion to the number of contrasts
- each contrast is therefore more conservative
- OK for small values of C but overly conservative for large values of C

Multiple Planned Comparisons

- Holm-Bonferroni method : `https://en.wikipedia.org/wiki/HolmBonferroni_method`
- less conservative than straight Bonferroni
- graded adjustment with larger corrections for less significant p-values
- check online for examples
- can use the `p.adjust()` function in R

Multiple Planned Comparisons

- Keppel (and others) suggest a different approach
- you're allowed to test up to $a - 1$ orthogonal planned contrasts without any adjustment of α
- he argues that Bonferroni correction unfairly penalizes planned orthogonal contrasts
- if contrasts are planned, orthogonal and number $a - 1$ or fewer, then because the set of contrasts is not data-driven, and do not overlap, then there should be no need to adjust α level
- overall α level should be no different than that for the omnibus F test

Post Hoc Pairwise Comparisons

- Tukey's procedure allows you to perform tests of all possible pairwise comparisons in an experiment and still maintain $\alpha_{EW} = 0.05$
- the `TukeyHSD()` function in R will do this for you
- Tukey procedure makes each pairwise test more conservative
- designed to take into account the idea that data-driven tests will involve higher Type-I error rates
- there are various modifications of Tukey's procedure when sample variances are unequal or when samples sizes are unequal (see M&D)

Post Hoc Pairwise Comparions

- Scheffe method maintains α_{EW} at 0.05 when at least some of the contrasts to be tested are complex, and suggested by the data (post-hoc)
- see M&D text for a detailed description of the method
- Scheffe method is quite conservative
- see tables 5.4 & 5.5 for comparison between methods

Other Procedures

- Dunnett's procedure
  - useful when one of the groups is considered a control and is involved in all contrasts
- Fisher's LSD (least significant difference)
- Newman-Keuls
- see M&D text for details about these other methods

What should I do?

- decide which approach you think is most reasonable, given your data and your experimental design
- be ready to defend your approach to reviewers
- be ready to use a different approach if necessary
- what's the "culture" in your lab / field / journal?

R Code

- ANOVA using the `aov()` function in `R`
- computing `Fcomp` manually
- using `TukeyHSD()`
- monte-carlo simulations of multiple comparison Type-I error rates
  - planned vs pos-hoc comparisons