

Comparisons among groups within ANOVA



Problem with one-way ANOVA



- There are a couple issues regarding one-way ANOVA
- First, it doesn't tell us what we really need to know
 - We are interested in specific differences, not the rejection of the general null hypothesis as typically stated
- Second, though it can control for type I error, the tests that are conducted that do tell what we want to know (i.e. is A different from B, A from C etc.) control for type I error themselves
- So why do we do the omnibus one-way ANOVA?
 - Model testing
 - Outside of providing an estimate for variance accounted for in the DV and residual standard error, it is fairly limited if we don't go further

Multiple Comparisons



- Why multiple comparisons?
- Post hoc comparisons
- A priori comparisons
- Trend analysis
- Dealing with problems

The situation



- One-way ANOVA
- What does it tell us?
 - Means are different
 - How?
 - ✦ Don't know
- What if I want to know the specifics?
 - Multiple comparisons

Problem



- Doing multiple tests of the same type leads to increased type I error rate
- Example 4 groups:
 - 6 possible comparisons, .05 per comparison our chance of making a type I error among *any* of them is $\leq .30$
 - Yikes!

Family-wise error rate



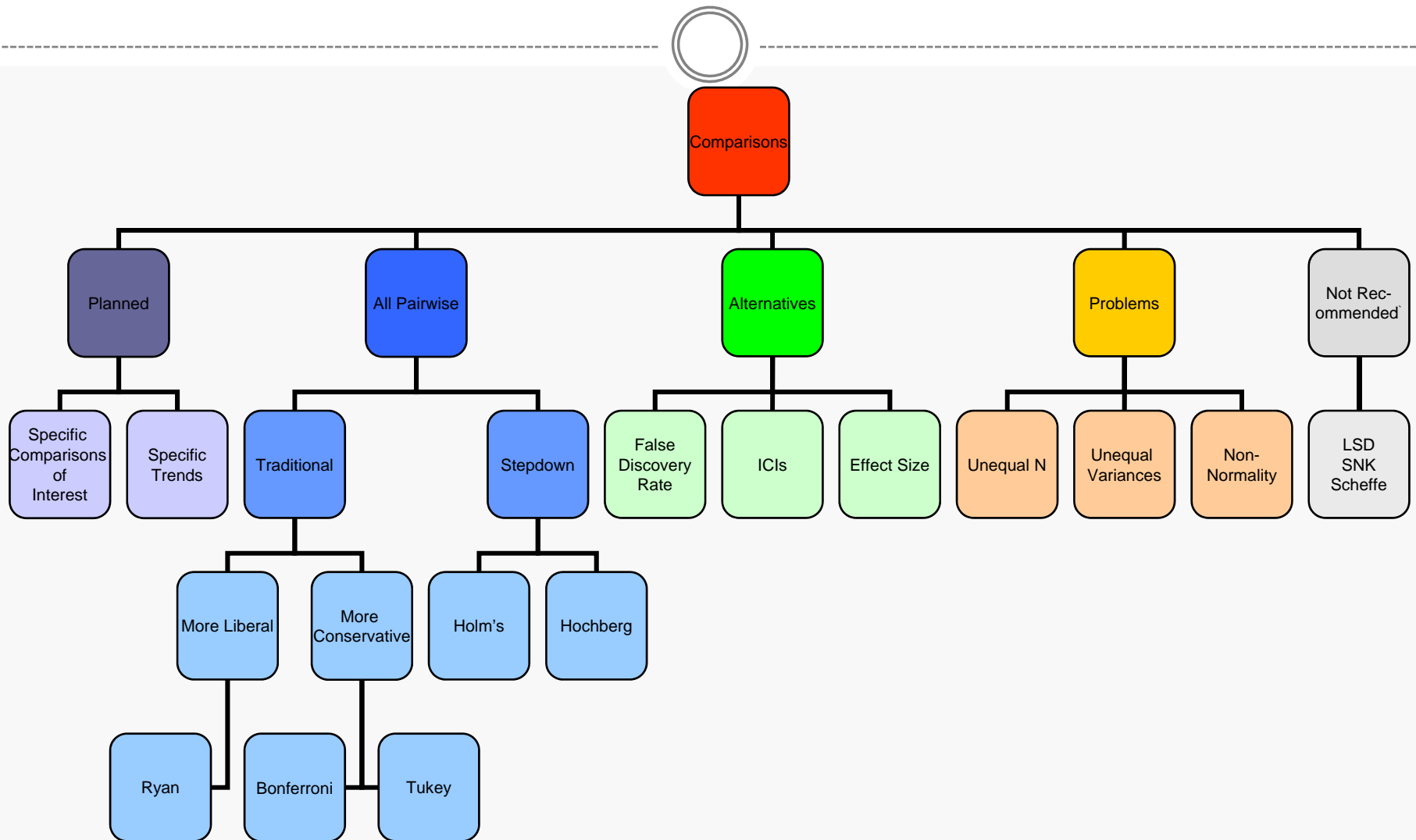
- What we're really concerning ourselves with here is familywise error rate (for the family of comparisons being made), rather than the per comparison (pairwise) error rate.
- So now what?
 - Take measures to ensure that we control FW_{α}

Some other considerations

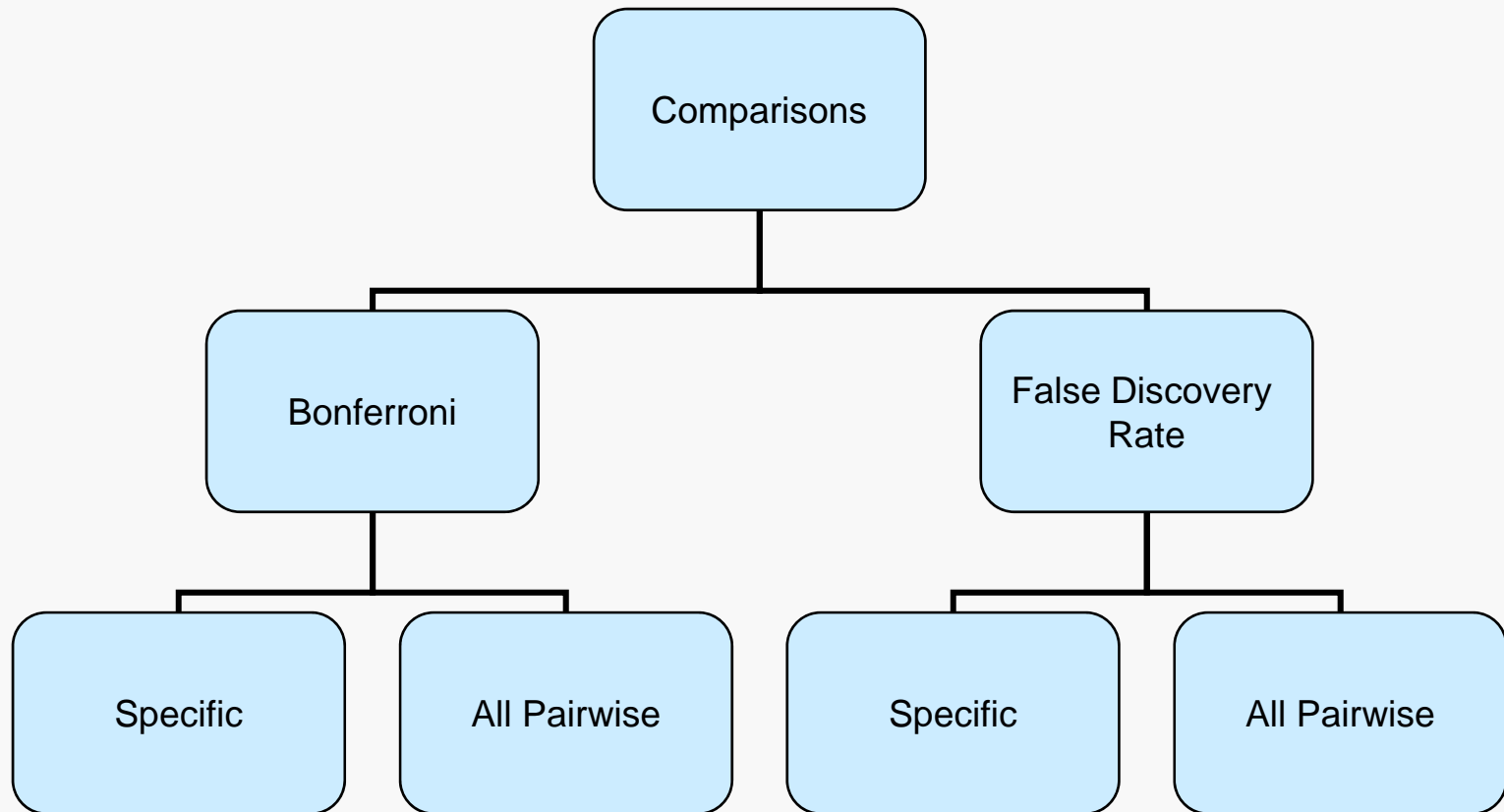


- A prior vs. Post hoc
 - Before or after the fact
- A priori
 - Do you have an expectation of the results based on theory?
 - ✦ A priori
 - ✦ Few comparisons
 - ✦ More statistically powerful than a regular one-way analysis
- Post hoc
 - Look at all comparisons of interest while maintaining type I error rate

A Basic Breakdown



A Different Breakdown



Post hoc!



- Planned comparisons are good when there are a small number of hypotheses to be tested
- Post hoc comparisons are done when one wants to check all paired comparisons for possible differences, and are typically what is seen in the literature¹
- Omnibus F test
 - Need significant F?
 - ✦ Current thinking is no
 - ✦ Most multiple comparison procedures are carried out without regard to the overall ANOVA F

Organization



- Ol' skool
 - Least significant difference ('protected' t-tests)
 - Bonferroni
- More standard fare
 - Tukey's, Student Newman-Keuls, Ryan, Scheffe etc.
- Special situations
 - HoV violation
 - Unequal group sizes
- Stepdown procedures
 - Holm's
 - Hochberg
- Newer approaches
 - FDR
 - ICI
 - Effect size

Least significant difference/Multiple t



- Do not use, it does not control for the familywise error rate for more than three comparisons¹
 - And if you only had 3 comparisons one could question the need for a correction
- However the basic approach here can be seen elsewhere so we'll use it as a starting point
- It is essentially multiple t-tests with no correction, however LSD requires a significant overall F to start
 - The only one that does that we will discuss
- Rather than pooled or individual variances use MS_{error} and t_{cv} at $df_{w/in}$

• So:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{MS_{error}}{n} + \frac{MS_{error}}{n}}}$$
$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{2MS_{error}}{n}}}$$

Bonferroni and Sidak test



- The ‘Bonferroni’ is actually a general approach regarding corrective measures for familywise error rate maintenance that can be applied to many situations
 - We will compare these type of procedures to controlling for the *false discovery rate* later
- The Bonferroni *t* adjustment specifically simply reduces alpha for the comparisons of interest based on the number of comparisons being made
 - Use $\alpha^* = \alpha/c$ where *c* is the number of comparisons
 - Technically we could adjust in such a fashion that some comparisons are more liberal than others, but this is the default approach in most statistical packages
- The Sidak is a modified version
 - ✦ Same story except our $\alpha^* = 1 - (1 - \alpha)^{1/c}$
 - ✦ Example 3 comparisons
 - Bonferroni $\alpha^* = .05/3 = .0167$
 - Sidak $\alpha^* = 1 - (1 - .05)^{1/3} = .0170$
 - In other words, the Sidak correction is not quite as strict (slightly more powerful)

Bonferroni



- While the traditional Bonferroni adjustment is widely used (and makes for an easy approach to eyeball comparisons yourself), it generally is too conservative in its standard from
- It is not recommended that you use it if there are a great many comparisons, as your pairwise comparisons would be using very low alpha levels
 - E.g. 7 groups: each comparison would be tested at $\alpha = .002$

Tukey's studentized range statistic



$$q_r = \frac{\bar{X}_L - \bar{X}_S}{\sqrt{\frac{MS_{error}}{n}}}$$

- This can be used (in lieu of the standard F test) to test the overall hypothesis that there is a significant difference among the means by using the largest and smallest means among the groups
- It is tested against the q critical value for however many groups are involved
 - The r above refers to the number of groups, and is used to obtain the q critical value for a given n
- Depending on how the means are distributed, it may or may not lead to the same conclusion as the F test
- Many post hoc procedures will use this q approach in order to determine significant differences

Tukey's HSD



- Tukey's HSD is probably the most common post hoc utilized for comparing individual groups you'll see in psych research
- It compares all mean differences against the largest q_{cv}
 - Conducted as though means were the maximum number of steps apart
 - ✦ E.g. if 6 means the largest and smallest would be 6 steps apart
- Thus Familywise type I error rate is controlled, but compared to many available procedures this is at the cost of a rise in type II error (i.e. loss in power)

Newman-Keuls



- Uses a different q depending on how far apart the means of the groups are in terms of their ordered series.
- In this way q_{cv} will change depending on how close the means are to one another (in terms of their ordering from large to small)
 - Closer values will need a smaller difference to be significantly different
- Problem: turns out that NK test does not control for FW type I error rate any better than the LSD test
 - Inflates for more than three groups
- Do not use if controlling FW error rate is a concern

Ryan Procedure



- Happy medium

$$\alpha_r = \frac{\alpha}{k / r}$$

- Uses the Newman-Keuls method but changes alpha to reflect the number of means involved (k) *and* how far apart those in the comparison are (r)
 - As you can see, at max number of steps apart $k=r$ and we will be testing at α , while closer means are tested at more stringent alpha levels
- Others came after to slightly modify it to ensure α_{FW} rate is maintained
 - Hence REGWQ in stat packages¹
- α controlled, power retained \rightarrow happy post hoc analysis

Comparison of procedures



- Example SPSS output

VAR00001

VAR00002	N	Subset		
		1	2	3
Student-Newman-Keuls ^{a,f}	8	2.0000		
	8		4.0000	
	8			6.0000
Sig.		1.000	1.000	1.000
Tukey HSD ^{a,b}	8	2.0000		
	8	4.0000	4.0000	
	8		6.0000	
Sig.		.063	.063	
Ryan-Einot-Gabriel-Wel sch Range ^b	8	2.0000		
	8		4.0000	
	8			6.0000
Sig.		1.000	1.000	1.000

Means for groups in homogeneous subsets are displayed.

Based on Type III Sum of Squares

The error term is Mean Square(Error) = 2.762.

a. Uses Harmonic Mean Sample Size = 8.000.

b. Alpha = .05.

Unequal n and Heterogeneity of Variance



- The output there mentions the harmonic mean
- If no HoV problem and fairly equal n, can use the harmonic mean of the sample sizes to calculate means and proceed as usual
 - k is number of groups, n the number in a particular group

$$\bar{n}_h = \frac{k}{\sum \frac{1}{n_i}}$$

- Stat packages already do this for you when using e.g. ‘Type III’ sums of squares (SPSS default)
- Also, if one compares this to the arithmetic mean for unequal samples, one can see that it is less than if the group means were equal to the largest sample (obvious) but also less than the arithmetic mean, further demonstrating that with unequal samples sizes we are losing power
- For 3 groups with ns equal to 10 15 20, $n_h = \sim 14$

Tests for specific situations

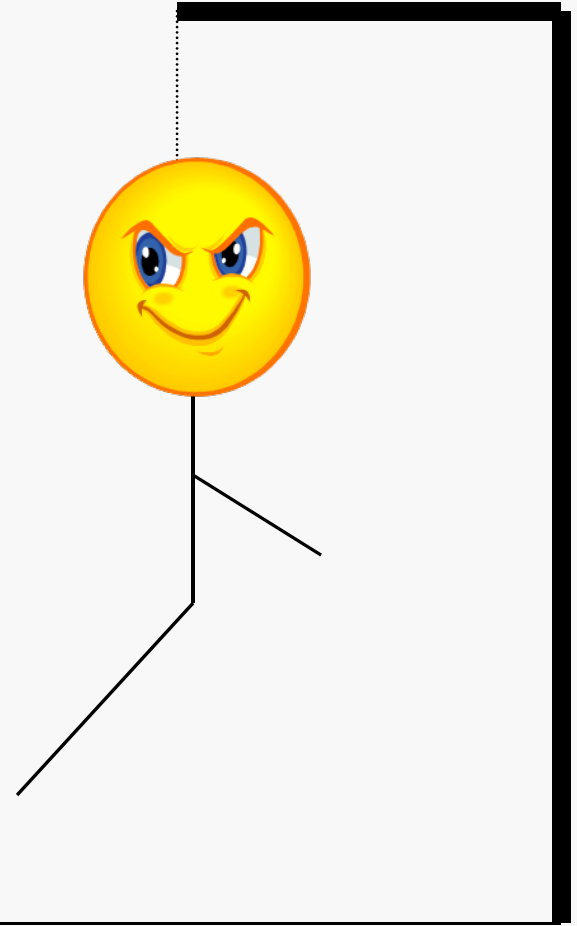


- For heteroscedasticity
 - Dunnett's T3
 - ✦ Think of as a Welch t-test with adjusted critical value
 - Games-Howell
 - ✦ Similar to Dunnett's
 - ✦ Works for unequal N
 - ✦ Creates a confidence interval for the difference, if doesn't include zero then sig diff
 - ✦ Performs better with larger groups than Dunnett's¹
- Nonnormality can cause problems with these however

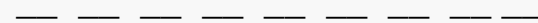
Non-normality



- Can we save him?
- One may start with nonparametric techniques for analysis, and then adjust p-values accordingly via a Bonferroni or FDR approach



B T S T P



Others



- Scheffe
 - Uses the F distribution rather than the studentized range statistic, with $F(k-1, df_{\text{error}})$ rather than $(1, df_{\text{error}})$
 - It allows for testing of *any* type of linear contrast¹ in a post hoc fashion
 - Much more conservative than most, suggested alpha = .10 if used
 - Not to be used to do all pairwise comparisons or a priori comparisons
- Dunnett
 - Use when wanting to compare a control against several treatments

Multiple comparisons



- Most modern methods control for type I FW error rate (the probability of at least 1 incorrect rejection of H_0) such that rejection of omnibus F not needed
- However if the F test is applied and rejected, alpha might in reality actually be *lower* than .05 (meaning raise in type II i.e. reduced power)
- Stepdown procedures

Holm and Larzelere & Mulaik



- Holm's:
- Change α depending on the number of hypotheses remaining to be tested.
- First calculate t s and associated observed p -values for all comparisons and arrange in increasing magnitude (disregard sign of the t -statistic)
- Test the largest mean difference at $\alpha^* = \alpha/c$,
- If significant test the next largest at $\alpha/(c-1)$ and so forth until you get to a nonsig result
- At that point where you do not reject a comparison, stop the procedure and do not reject any remaining either

- Logic: if one H_0 is rejected it leaves only $c-1$ null hypotheses left for possible incorrect rejection (type I error) to correct for
- Controls alpha but is more powerful than other approaches
- Can be used for testing any specific contrasts

- L&M
- Provided same method but concerning correlation coefficients
- As much as people love to flag correlation matrices for significance, let me know if you ever see this correction ever applied

Hochberg



- Order the observed p-values $P_{[1]}, P_{[2]} \dots P_{[k]}$ smallest to largest¹
- Test largest p at α , if you don't reject move to next one and test the next p-value at $\alpha/(k-1)$
- If rejected, reject all those that follow also
- In other words:
 - Reject if $P_{[k]} \leq \alpha/k$
- Essentially backward Holm's method
 - Stop when we reject rather than stop when we don't
 - Turns out to be more powerful, but assumes independence of groups (unlike Holm's)

False Discovery Rate



- More recent efforts have supplied corrections that are more statistically powerful and would be more appropriate in some situations e.g. when the variables of interest are dependent
- To compare, the Bonferroni family of tests seeks to control the chance of even a single false discovery among all tests performed
- The False Discovery Rate (FDR) method on the other hand controls the proportion of errors among those tests whose null hypothesis were rejected.

$$\frac{\textit{number of false rejections}}{\textit{total number of rejections}}$$

- Another way to think about it is- why control for alpha for a test in which you aren't going to reject the H_0 ?

False Discovery Rate



- Benjamini & Hochberg¹ defined the FDR as the expected proportion of errors *among the rejected hypotheses*
 - Proportion of falsely declared pairwise tests among all pairwise tests declared significant
- FDR is a family of procedures much like the Bonferroni although conceptually distinct in what it tries to control for

False Discovery Rate



- In terms of alpha, starting with the largest observed p-value (which will have no adjustment)

$$P_k \leq \alpha^* \quad \text{reject}$$

$$\alpha^* = \frac{(C - k + 1)\alpha}{C}$$

c = number of comparisons

k = which comparison we are on in terms of order

- In terms of the specific p-value

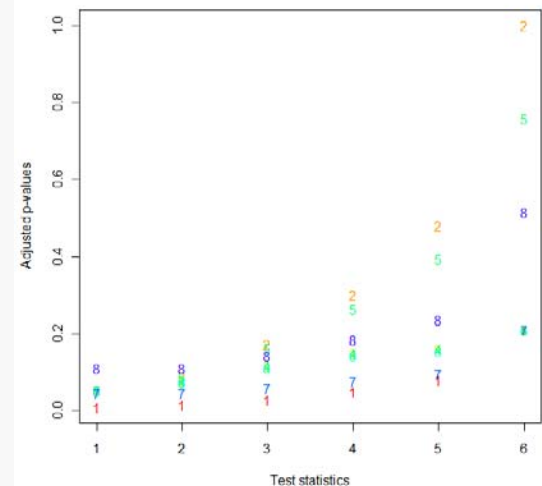
$$P_{adj} = p\left(\frac{\alpha}{\alpha^*}\right)$$

R library multtest



- Example for a four group setting
- <http://www.unt.edu/benchmarks/archives/2002/april02/rss.htm>
- `pairwise.t.test(DV, groupvar, p.adjust.method = "fdr")`
- The output below uses the `multtest`¹ for comparison of several methods

1	2	3	4	5	6	7	8
rawp	Bonf	Holm	Hochberg	SidakSS	SidakSD	BH	BY
0.009	0.054	0.054	0.054	0.053	0.053	0.045	0.110
0.015	0.090	0.075	0.075	0.087	0.073	0.045	0.110
0.029	0.174	0.116	0.116	0.162	0.111	0.058	0.142
0.050	0.300	0.150	0.150	0.265	0.143	0.075	0.184
0.080	0.480	0.160	0.160	0.394	0.153	0.096	0.235
0.210	1.000	0.210	0.210	0.757	0.210	0.210	0.514



False Discovery Rate



- It has been shown that the FDR performs comparably to other methods with few comparisons, and better (in terms of power, they're all ok w/ type I error) with increasing number of comparisons
- An issue that one must remind themselves in employing the FDR regards the emphasis on p-values
- Knowing what we know about p-values, sample size and practical significance, we should be cautious in interpretation of such results, as the p-value is not an indicator of practical import
- However, the power gained by utilizing such a procedure may provide enough impetus to warrant its usage at least for determining statistical significance

Another option



- Inferential Confidence Intervals (see 5700 notes)
- One could perform post hoc approaches to control for type I error
 - E.g. simple Bonferroni-style correction to our initial critical value
 - E reduction term depends on the pair of groups involved
 - More comparisons will result in larger t_{cv} to be reduced
- Alternatively, one could calculate an average E over all the pairwise combinations, then go back and retest with that E
 - Advantage: creates easy comparison across intervals
 - Disadvantage: power will be gained in cases where E goes from larger to smaller (original to average), and lost in the converse situation

Yet another alternative



- **Something to think about: Where is type I error rate in the assessment of practical effect?**
 - All the standard approaches have statistical significance as the sole focus
- **Calculate interval estimates for the effect size of each comparison**
- **As with other procedures you could correct the interval estimate**
 - E.g. start with .95 CI but widen depending on number of intervals calculated for all the group comparisons in a Bonferroni fashion
- **Technically however, the effect sizes, if done appropriately, are evaluated without regard to the design/complexity of the study**
 - i.e. The effect size for A vs. D in an ANOVA four group setting would already reflect the same as that if they were the only groups under consideration
- **This may in fact provide the best solution, because the emphasis is shifted from statistical significance to practical importance**

Which to use?



- Some are better than others in terms of power, control of a familywise error rate, data behavior
- Try alternatives, but if one is suited specifically for your situation use it
- Some suggestions
 - Assumptions met: Tukey's or REWQ of the traditional options, FDR for more power
 - Unequal n: Gabriel's or Hochberg (latter if large differences)
 - Unequal variances: Games-Howell
- However one must remember that the entire emphasis of post hoc procedures is on the some of the biggest problems NHST has to offer
 - Dichotomous thinking
 - Ignoring effect size
 - Ignoring otherwise noticeable trends
 - Overemphasis on type I at the expense of type II
 - ✦ False negatives may be as important or even much more so depending on the research question

A priori Analysis (contrast, planned comparison)



- The point of these type of analyses is that you had some particular comparison in mind before even collecting data.
- Why wouldn't one do a priori all the time?
 - Though we have some idea, it might not be all that strong theoretically
 - Might miss out on other interesting comparisons

Still doing t-tests¹



- For any comparison of means:

$$\psi = \sum a_j \bar{X}_j$$

$$t' = \frac{\psi}{\sqrt{\frac{(\sum a_j^2) MS_{error}}{n}}}$$

Linear contrasts



- Testing multiple groups against another group
- Linear combination
 - A weighted sum of group means
 - Sum of the weights (a) should equal zero

$$\psi = a_1\bar{X}_1 + a_2\bar{X}_2 \dots \quad a_k\bar{X}_k = \sum a_j\bar{X}_j$$

Example



- From the t.v. show data (ANOVA notes):
 - 1) 18-25 group Mean = 6 SD = 2.2
 - 2) 25-45 group Mean = 4 SD = 1.7
 - 3) 45+ group Mean = 2 SD = .76
- Say we want to test whether the youngest group is significantly different from the others:
 - $\Psi = 2(6) + (-1)(4) + (-1)(2) = 6$
 - Note: we can choose anything for our weights as long as they add to zero and reflect the difference we want to test
 - ✦ However, as Howell notes, having the weights sum to two will help us in effect size estimation (more on that later)
- $$SS_{\text{contrast}} = \frac{n\psi^2}{\sum a_j^2}$$
 - Equals MS_{contrast} as df 1 for comparison of 2 groups

Example cont'd.



- $SS_{\text{contrast}} = (8 \cdot 6^2) / 6 = 48$
- $df = 1$
 - SS_{contrast} will always equal MS_{contrast}
- $F = 48 / MS_{\text{error}} = 48 / 2.76 = 17.39$
- Compare to $F_{\text{cv}}(1, 21)$, if you think you need to.

- Note SPSS gives a t-statistic which in this case would be 4.17 ($4.17^2 = 17.39$)

Choice of coefficients



- Use whole numbers to make things easier
 - Though again we will qualify this for effect size estimates
- Use the smallest numbers possible
- Those with positive weights will be compared to those with negative weights
- Groups not in the comparison get a zero
- In orthogonal contrasts, groups singled out in one contrast should not be used in subsequent contrasts

Orthogonal Contrasts



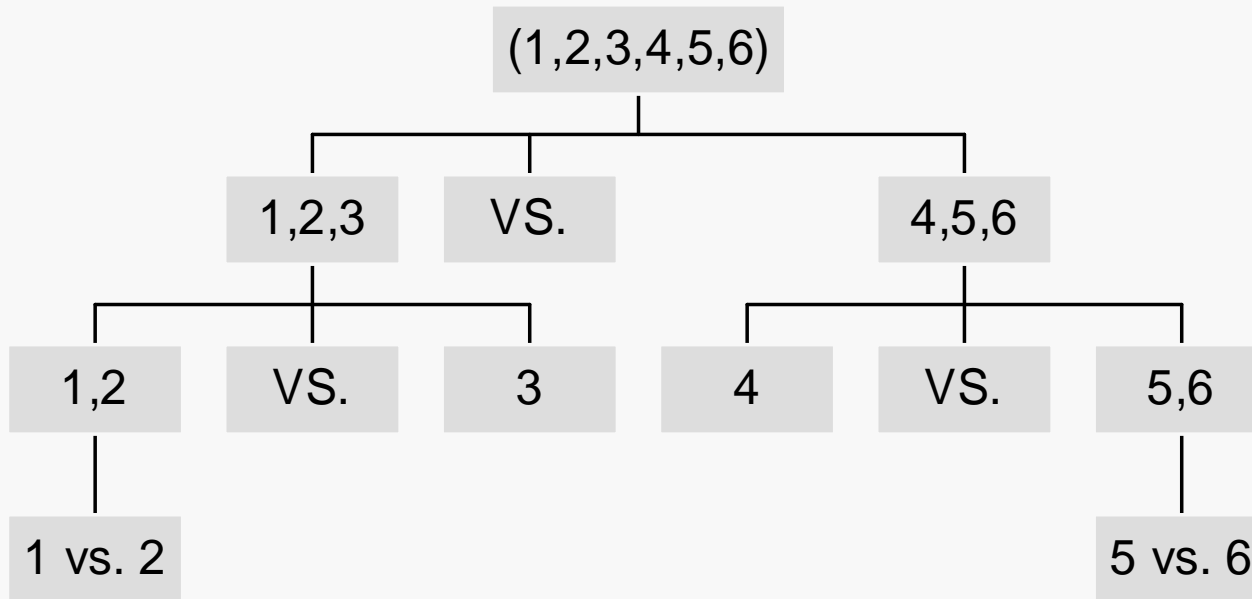
- Contrasts can be said to be independent of one another or not, and when they are they are called orthogonal
- Example: 4 groups
 - If a contrast is conducted for 1 vs. 2, it wouldn't tell you anything (is independent of) the contrast comparing 3 vs. 4
 - A complete set of orthogonal contrasts will have their total SS equal to SS_{treat}

Requirements

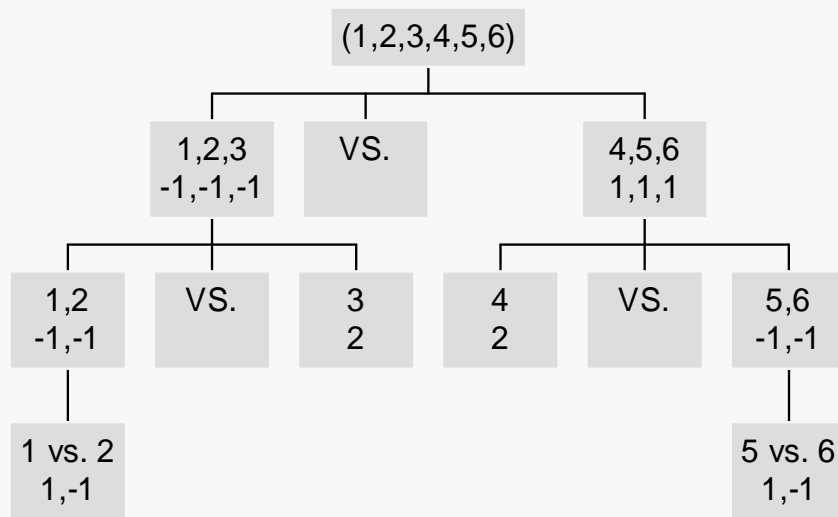


- Sum of weights (coefficients) for individual contrasts must equal zero
- Sum of the products of the weights for any two contrasts sum to zero
- The number of comparisons must equal the df for treatments

Example



Weights



- **-1 -1 -1 1 1 1**
- **-1 -1 2 0 0 0**
- **0 0 0 2 -1 -1**
- **1 -1 0 0 0 0**
- **0 0 0 0 1 -1**

Orthogonal contrasts



- Note that other contrasts could have been conducted and given an orthogonal set
- Theory should drive which contrasts you conduct
- Orthogonal is not required
 - Just note that the contrasts would not be independent
 - ✦ We couldn't add them up to get SS_{treat}

Contrast Types



- Stats packages offer some specific types of contrasts that might be suitable to your needs
- Deviation
 - Compares the mean of one level to the mean of all levels (grand mean); reference category not included.
- Simple
 - Compares each mean to some reference mean (either the first or last category e.g. a control group)
- Difference (reverse Helmert)
 - Compares each level (except the first) to the mean of the previous levels

Contrast Types

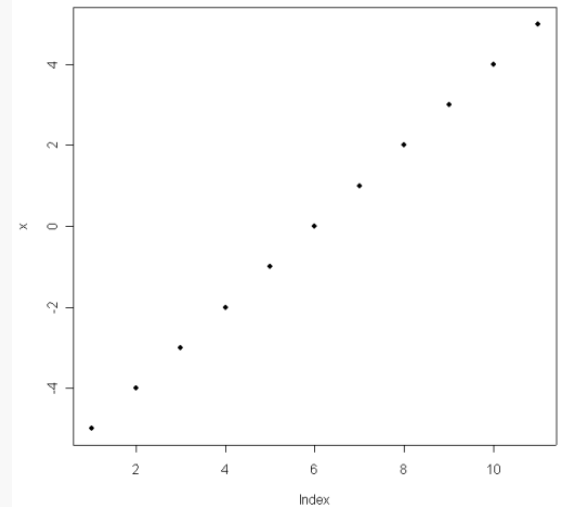


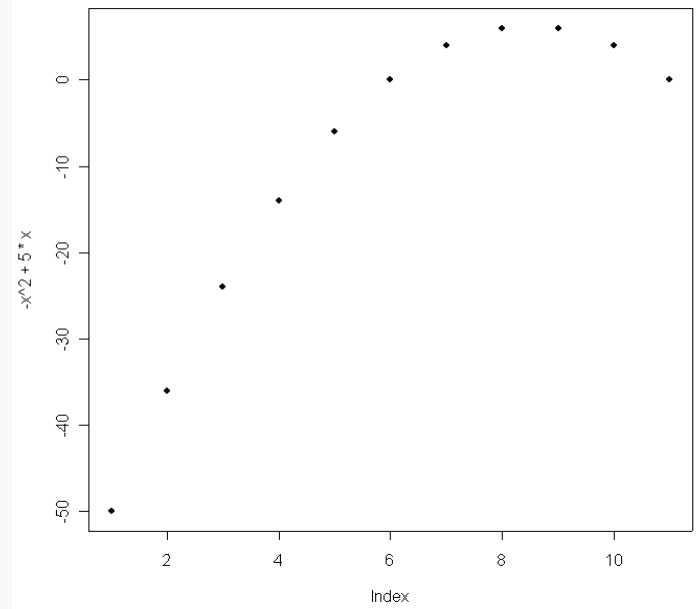
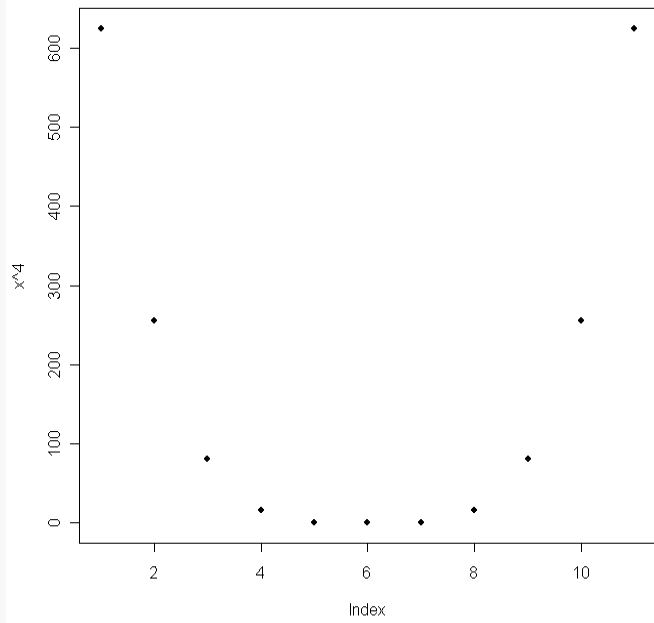
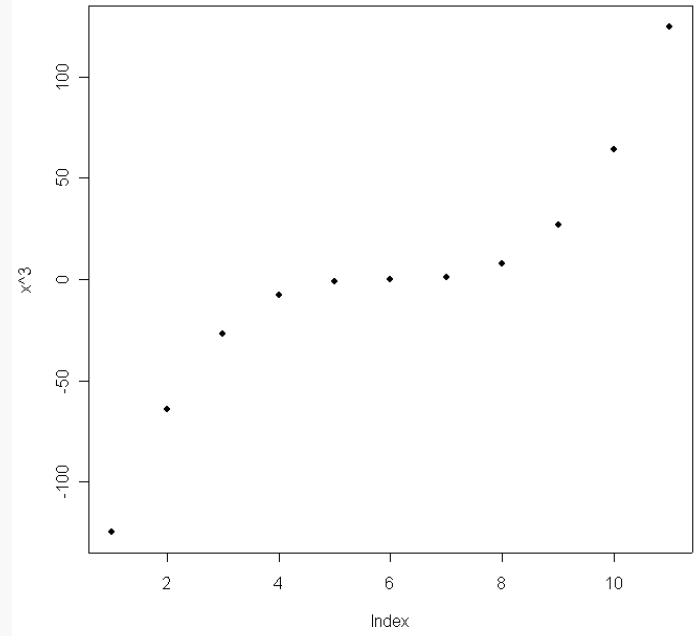
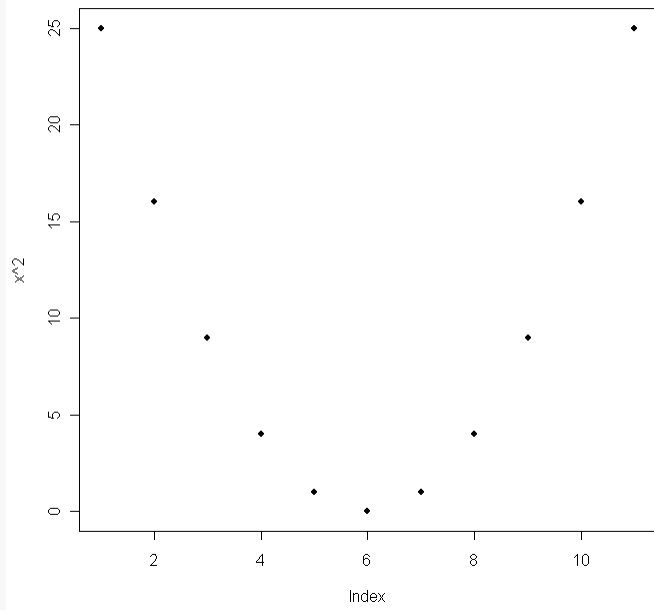
- Helmert
 - Compares mean of level 1 with all later, level 2 with the mean of all later, level 3 etc.
- Repeated
 - Compares level 1 to level 2, level 2 to level 3, 3 to 4 and so on
- Polynomial
 - Tests for trends (e.g. linear) across levels
- Note that many of these would most likely be more useful in a repeated measures design

Trend Analysis



- The last contrast mentioned (polynomial) regards trend analysis.
- Not so much interested in mean differences but an overall pattern
- When used?
 - Best used for categorical data that represents an underlying continuum
- Example linear







- Strategy the same as before, just the weights used will be different
- Example coefficients (weights):
 - Linear: $-2 -1 0 1 2$
 - Quadratic: $-2 1 2 1 -2$
 - Cubic: $-1 2 0 -2 1$

Summary for multiple comparisons



- Let theory guide which comparisons you look at
 - Perform a priori contrasts whenever possible
- Test only comparisons truly of interest
- Use more recent methods for post hocs for more statistical power