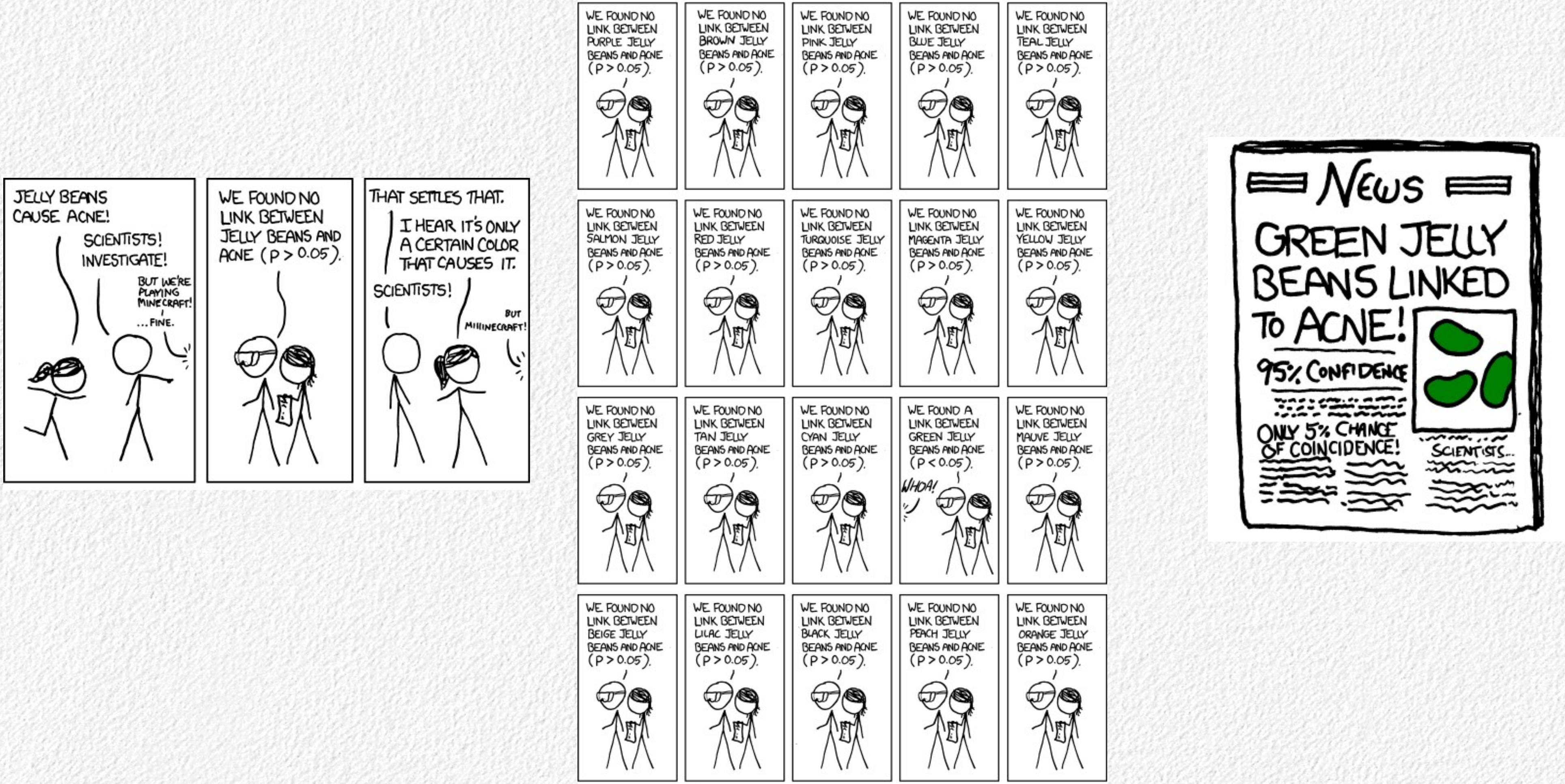
Post-hocs and contrasts


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Example: cost of selection in fruit flies

- Experiment to test for a fitness cost of selection
- Three different treatments
  - Not selected (NS)
  - Selected for resistance to pesticide (RS)
  - Selected for susceptibility to pesticide (SS)
- Measure reproductive output for each (fecundity)
- Question: is fecundity different between treatments?
  - If any kind of selection is costly, which will be different?
  - If the direction of selection matters, which will be different?

ANOVA doesn’t tell us what we want to know

- ANOVA gives an omnibus test of differences among the levels, but not specific about which means differ
- We need to know which means are different to interpret the results
- This we get from post-hoc procedures, which are done after a significant ANOVA
- Why do we put up with this?

The multiple testing problem

- Our α-level is an error rate (usually use a nominal $\alpha = 0.05$)
  - A single hypothesis test has a 5% chance of a false positive if the null is true
- With two tests we only avoid Type I error if we don’t get one on the first test AND don’t get one on the second, or $(1-0.05)(1-0.05)$
- Probability that at least one of two tests is a false positive is

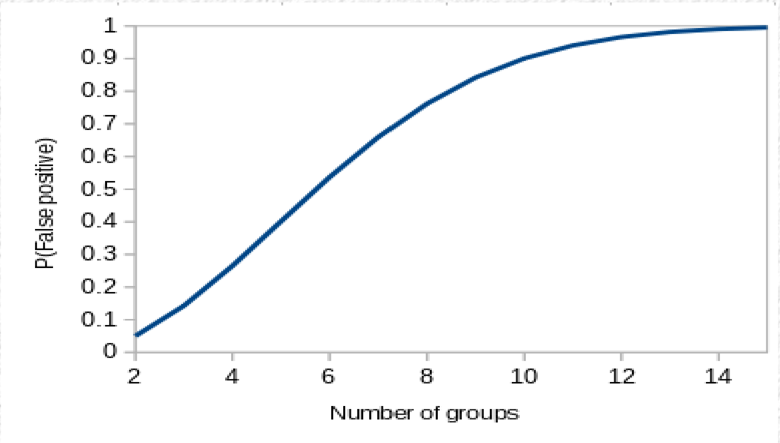
  $$1 - (1-0.05)(1-0.05) = 1 - (1-0.05)^2 = 0.0975$$

- Probability that at least one of k tests is a false positive is

  $$1 - (1 - \alpha)^k = 1 - (1 - 0.05)^k$$

Groups	Comparisons	P(False positive)
2	1	0.05
3	3	0.14
4	6	0.26
5	10	0.40
6	15	0.54
7	21	0.66
8	28	0.76
9	36	0.84
10	45	0.90

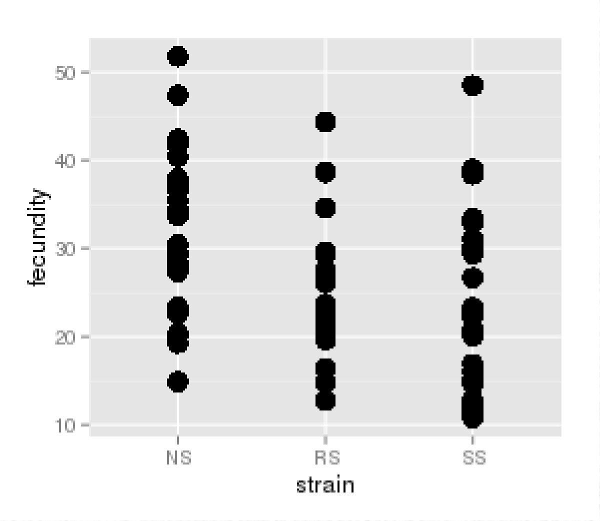
More groups, more comparisons, bigger problem

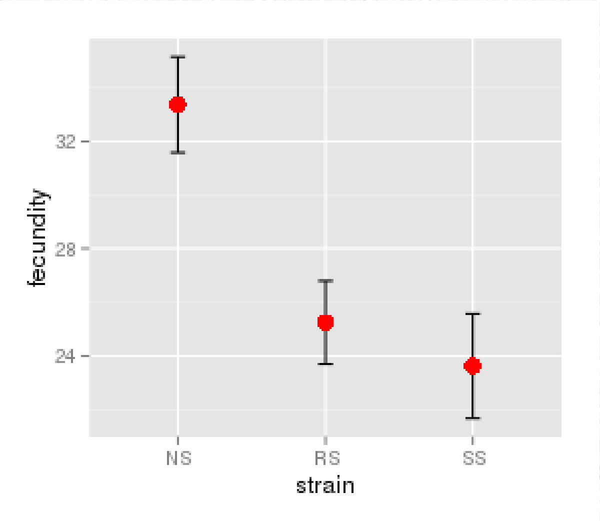


Essential to address this problem when comparisons are not independent

ANOVA’s two-step is designed to control Type I error rate

- The first step establishes that there is evidence for differences between at least two of the groups
- If (and only if) this omnibus test is significant we move on to the post-hocs
- The post-hocs adjust the amount of difference required to be significant to maintain a 5% **family-wise** error rate





The data

```
Response: fecundity
          Df Sum Sq Mean Sq F value    Pr(>F)
strain     2 1362.2  681.11  8.6657 0.0004244
Residuals 72 5659.0   78.60
```

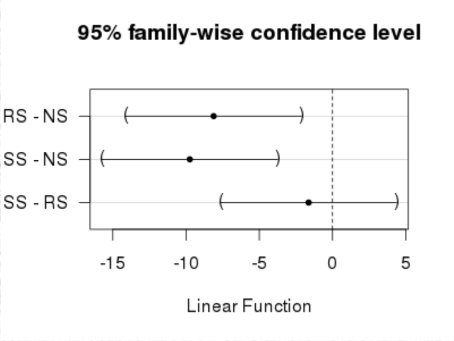
We have three comparisons to make: NS vs. RS, NS vs. SS, RS vs. SS

Making comparisons between all possible pairs of means is usually done with Tukey post-hocs

Tukey-Kramer HSD

- Uses the “studentized range” distribution instead of t for critical values, p-values
  - Studentized range is flatter than t - takes a greater difference between means to be significant than t
  - Gets flatter still as the number of comparisons increases – amount of difference required gets bigger the more groups are compared
  - For example, Tukey's requires 2.39 se between means, t-test requires 2.01 se for a pair of means to be significantly different for the fly data
- Can be used with unequal sample sizes between groups
- Since the amount of difference needed is adjusted we still consider $p < 0.05$ to be significant for each comparison

Tukey's comparisons



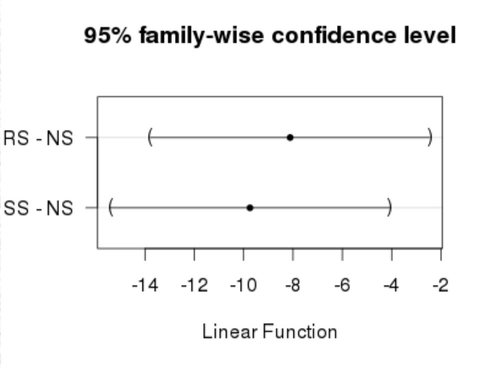
Implemented by using the t-distribution, but with lower d.f.

```
          Simultaneous Tests for General Linear Hypotheses
Multiple Comparisons of Means: Tukey Contrasts
Fit: aov(formula = fecundity ~ strain, data = fruitfly.df)
Linear Hypotheses:
              Estimate Std. Error t value Pr(>|t|)
RS - NS == 0   -8.116      2.508  -3.237 0.005105 **
SS - NS == 0   -9.744      2.508  -3.886 0.000662 ***
SS - RS == 0   -1.628      2.508  -0.649 0.793406
```

Post-hocs for fewer than all possible comparisons

- More tests → bigger adjustment to avoid Type I error → more differences missed → higher Type II error (and lower power)
- If you only actually care about a subset of the possible comparisons, better to only test the ones you care about
- For example:
  - Dunnett's method compares each mean to a single comparison group (usually the control)
  - Scheffe's method can compare any combinations of group means (e.g. RS and SS vs NS)

Dunnett's method
Compare each group against control



```
         Simultaneous Tests for General Linear Hypotheses
Multiple Comparisons of Means: Dunnett Contrasts
Fit: aov(formula = fecundity ~ strain, data = fruitfly.df)
Linear Hypotheses:
             Estimate Std. Error t value Pr(>|t|)
RS - NS == 0   -8.116      2.508  -3.237 0.003543 **
SS - NS == 0   -9.744      2.508  -3.886 0.000441 ***
```

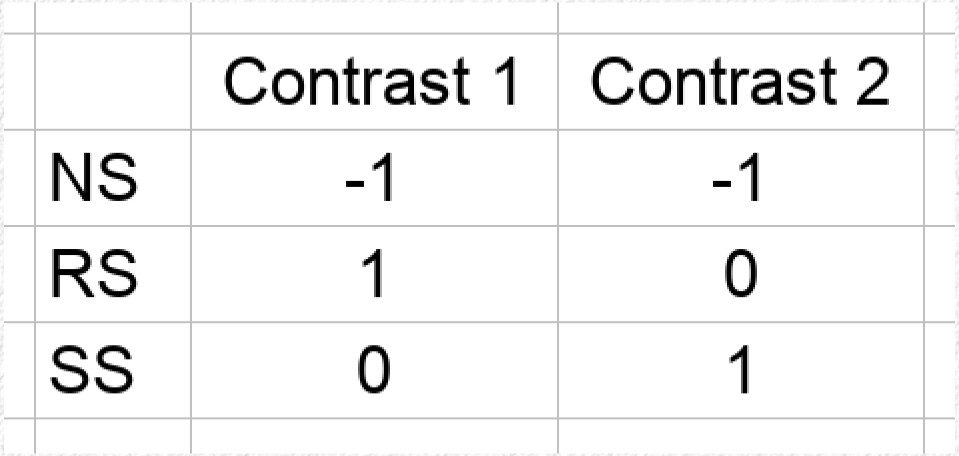
Omitting the SS – RS comparison makes the p-values smaller

Orthogonal contrast

- We need to use post-hocs following ANOVA because the comparisons are not independent
  - One group that by chance is unusually large or small can result in more than one false positive
- But, we don’t adjust our alpha level for analysis of completely different data sets
  - We don’t worry about a career-wise Type I error rate
  - p-values for different experiments aren’t adjusted
- If we can make comparisons within a data set that are independent then we don’t need to adjust alpha
- How do we use orthogonal contrasts?

The contrast matrix

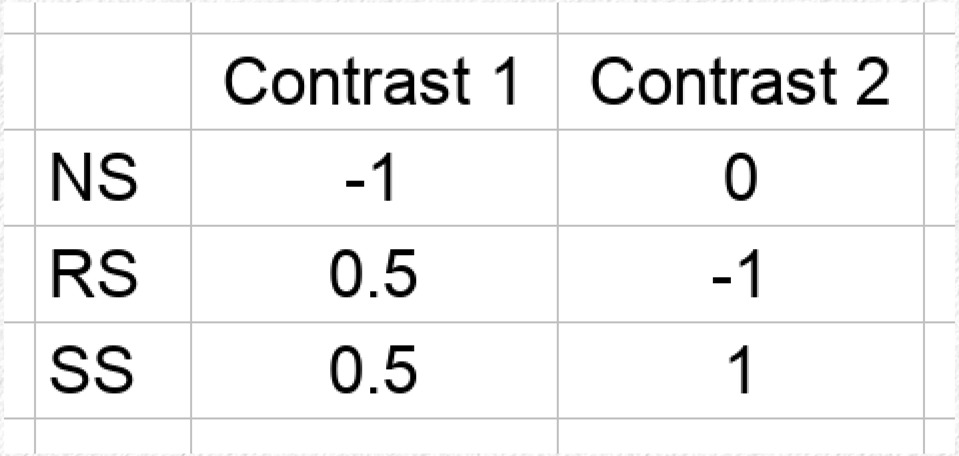
- Numbers in the matrix are **weights** – define the comparisons made
  - 0 indicates the mean isn't included in the comparison
  - Negative weights are compared with positive
- Contrast 1 compares RS to NS
- Contrast 2 compares SS to NS
- To be orthogonal, these weights have to:
  - Sum to zero for each contrast (down the columns)
  - Sums of products of any two contrasts has to be zero (multiply across columns, sum products)
- This set defines the Dunnett's comparisons – are they orthgonal?



	Contrast 1	Contrast 2
NS	-1	-1
RS	1	0
SS	0	1

Other possibility...

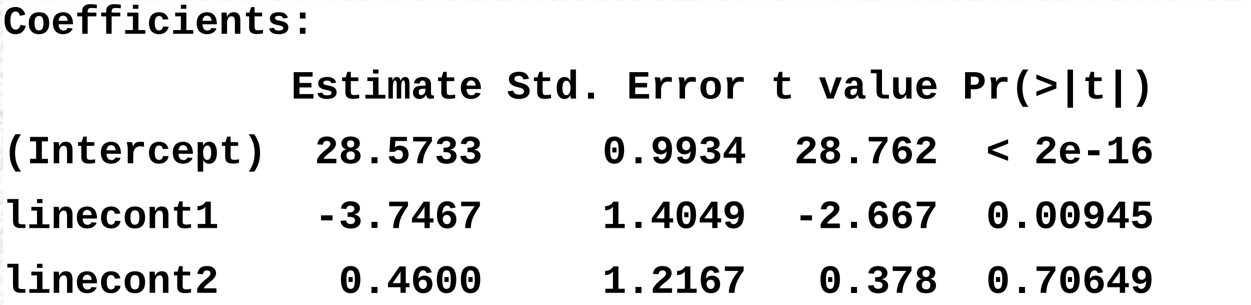
- Contrast 1 compares control (NS) with mean of the two selected lines
- Contrast 2 compares selected lines against each other



	Contrast 1	Contrast 2
NS	-1	0
RS	0.5	-1
SS	0.5	1

- Are Contrast 1 and Contrast 2 orthogonal?
- This is an example of Helmert coding – each level against the mean of the following levels

Results



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  28.5733     0.9934  28.762  < 2e-16
linecont1    -3.7467     1.4049  -2.667  0.00945
linecont2     0.4600     1.2167   0.378  0.70649
```

- Intercept is the grand mean
- First coeff. Is the difference between NS and mean of remaining two lines
- Second is the difference between second and third selected lines
- Independent – can interpret the p-values at the 0.05 level without fear of increasing Type I error
- What isn’t being compared?
- Is this what we want to know?

Advantages of orthogonal contrasts

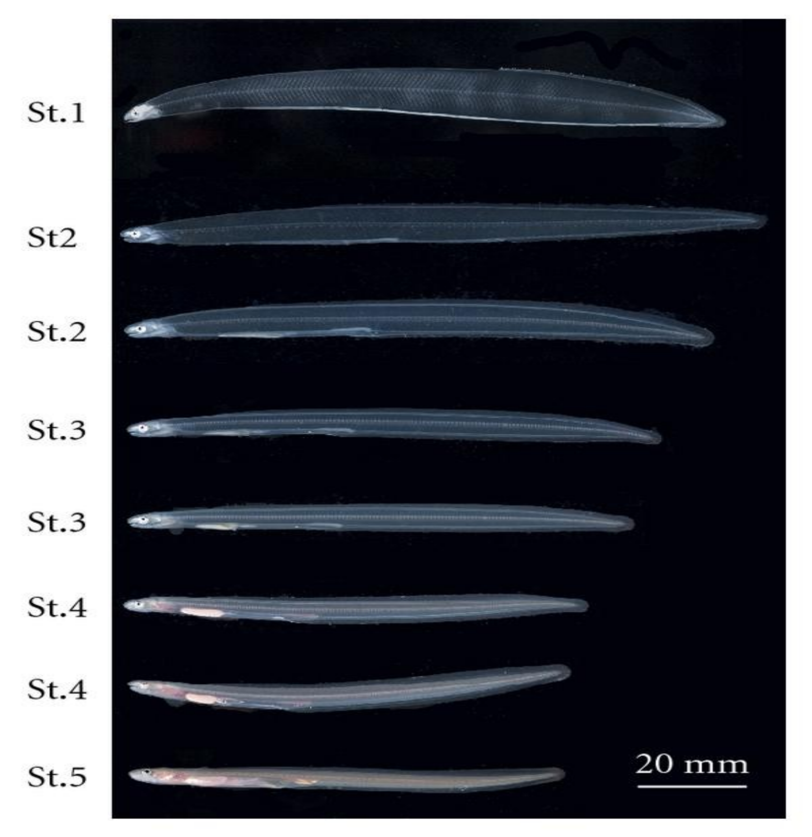
- The most statistically powerful method for comparing means (no need to adjust for multiple comparisons)
- Can be interpreted even if the omnibus ANOVA is not significant
- Can test hypotheses about combinations of groups (two selected vs. single non-selected group)

Disadvantages of orthogonal contrasts

- Sample size has to be equal between groups = balanced design
- Not all comparisons can be made orthogonal
  - For k groups there are at most k-1 independent contrasts
  - Some sets of comparisons aren't orthogonal, even if there are only k-1 of them
- If the question you want to ask can't be answered with orthogonal contrasts, better off with post-hocs

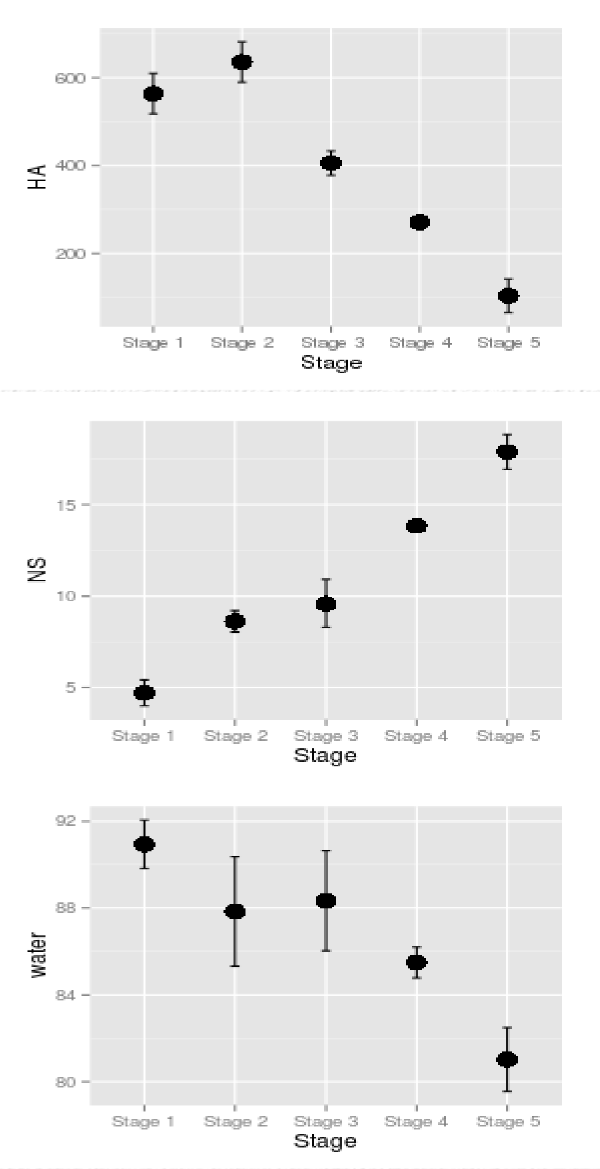
Example with ordinal levels: physiological changes during development

- Japanese Conger eel
- Five stages of metamorphosis from larvae to adult identified – stage is ordinal
- Four animals at each stage selected for measurement of several variables, including percent body water, hyaluronan (HA), and neutral sugar (NS)



Contrasts with ordinal categories

- Developmental stage is a categorical variable with natural ordering (it is an **ordinal** categorical variable)
- The questions we ask should account for this ordering
- For example, we could use contrasts that compare each level to the means of later levels



Sequential contrasts

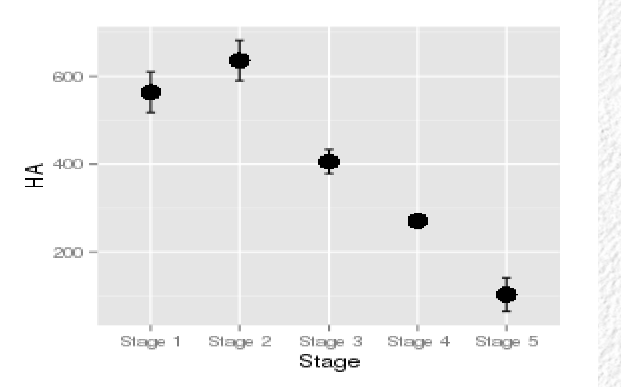
- This set of contrasts compares each level to the mean of the levels that follow
- Each level is different from the means of subsequent levels for HA

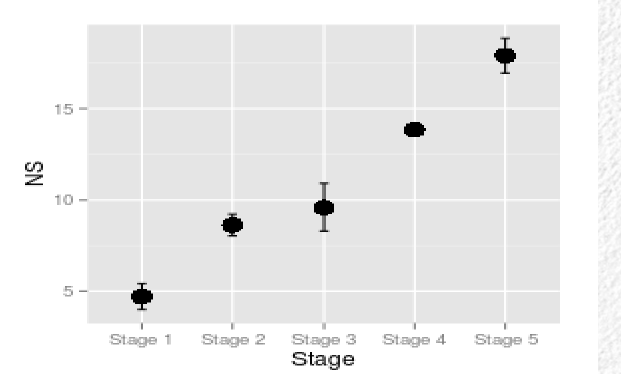
	Contrast 1	Contrast 2	Contrast 3	Contrast 4
Stage 1	4	0	0	0
Stage 2	-1	3	0	0
Stage 3	-1	-1	2	0
Stage 4	-1	-1	-1	1
Stage 5	-1	-1	-1	-1

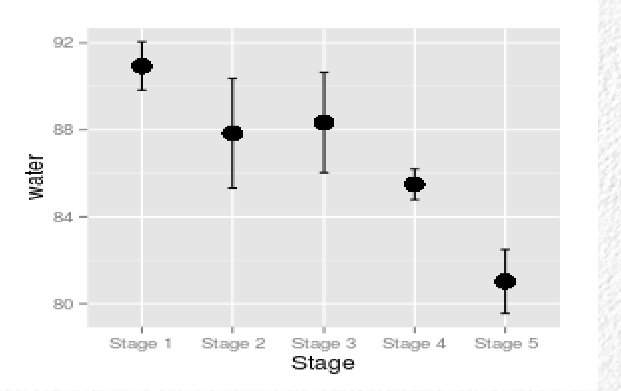
```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         395.938     16.261  24.349 1.79e-13 ***
Stage.seqContrast1  -41.966      8.131  -5.162 0.000116 ***
Stage.seqContrast2  -93.896     10.497  -8.945 2.12e-07 ***
Stage.seqContrast3  -72.813     14.844  -4.905 0.000190 ***
Stage.seqContrast4  -83.705     25.711  -3.256 0.005322 **
```

Pattern of change across the levels

- Instead of focusing on statistically significant differences in means, we could ask about the pattern of change
- Like a linear regression, but using the ordinal levels instead of a numeric predictor
- Done with **orthogonal polynomial contrasts**

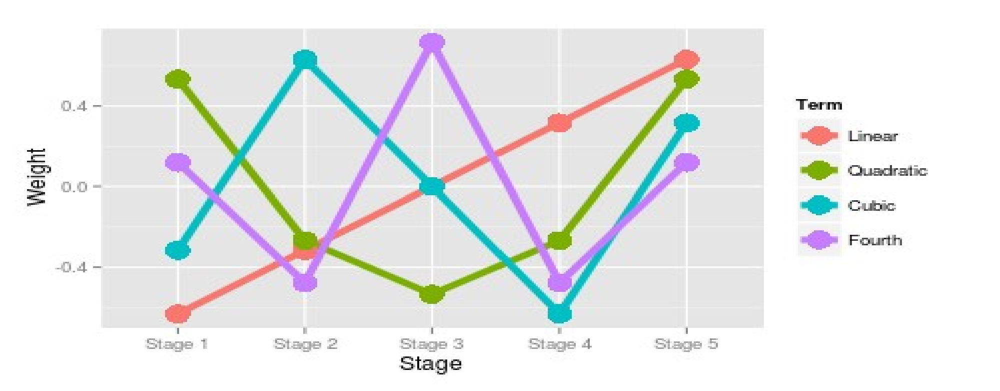




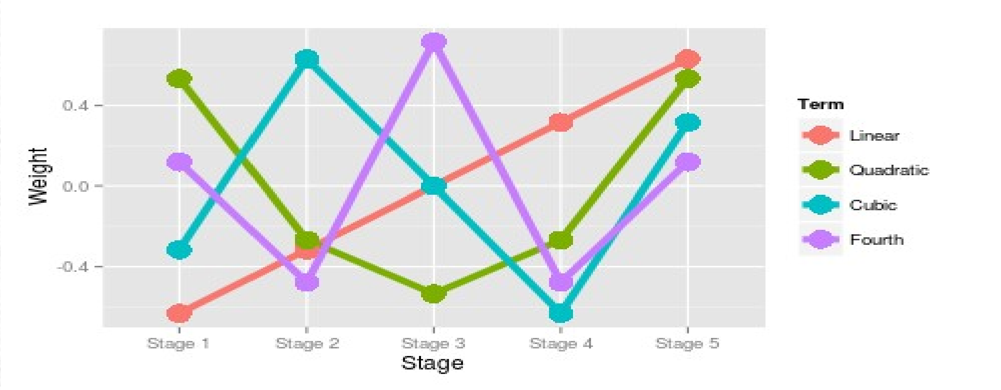


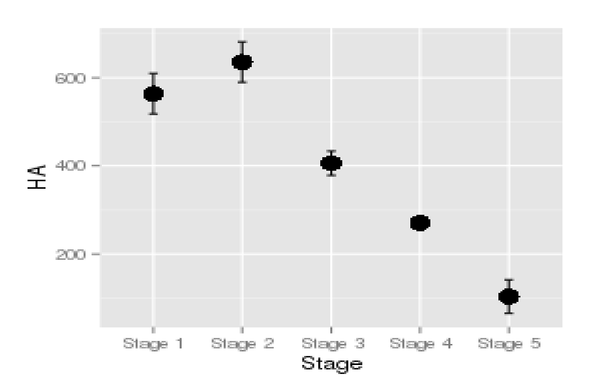
Orthogonal polynomial contrasts

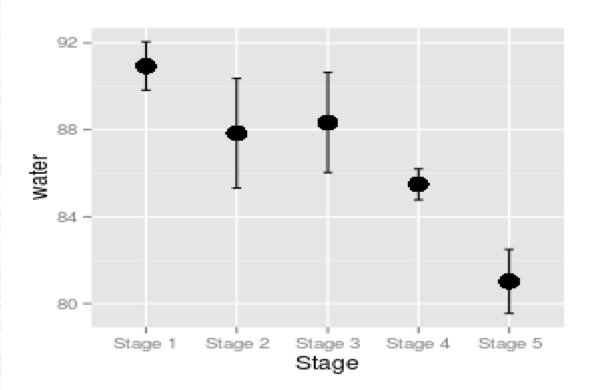
- Assign numeric weights to each developmental stage
  - Weights describe a pattern of change
  - With five stages can have up to a fourth degree polynomial
- The numbers used as weights are arbitrary, but have to meet orthogonality criteria (sum to 0, sums of products are 0)

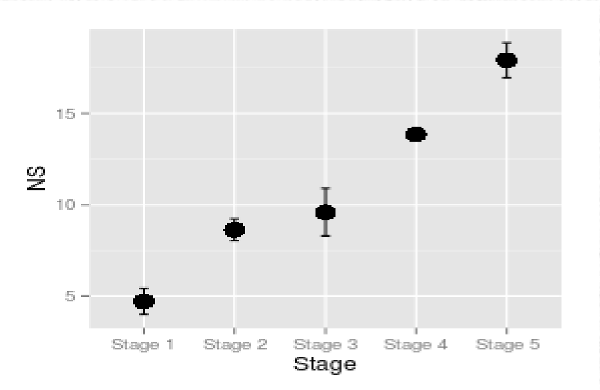


	Linear	Quadratic	Cubic	4th degree
Stage 1	-0.63	0.53	-0.32	0.12
Stage 2	-0.32	-0.27	0.63	-0.48
Stage 3	0	-0.53	0	0.72
Stage 4	0.32	-0.27	-0.63	-0.48
Stage 5	0.63	0.53	0.32	0.12







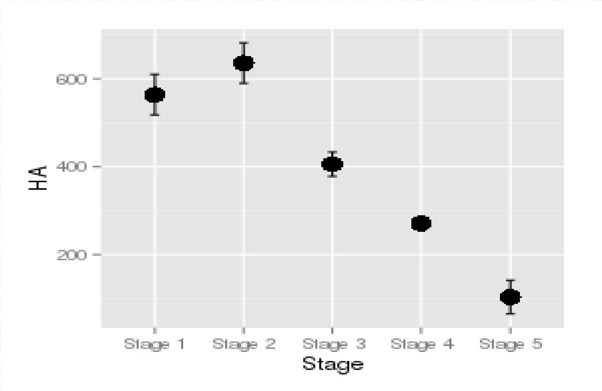


Which variable shows a linear trend?

Which shows a quadratic trend?

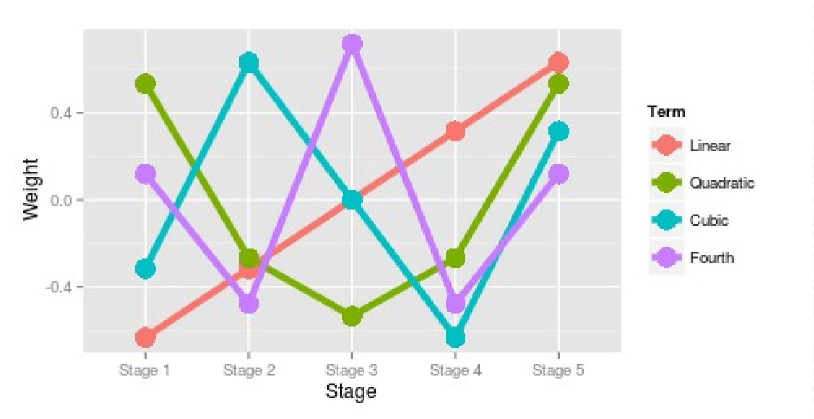
Results: HA

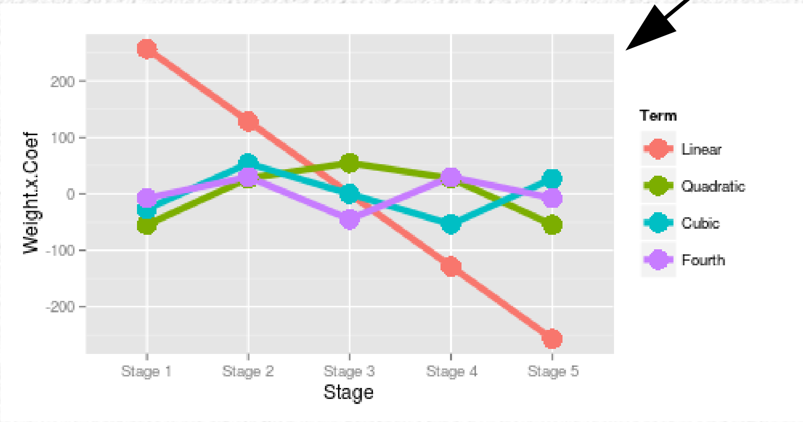
```
Call:
lm(formula = HA ~ Stage, data = eels)
Residuals:
    Min      1Q  Median      3Q     Max
-116.94  -43.12   12.68   30.44  122.39
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   395.94      16.26  24.349 1.79e-13 ***
Stage.L      -406.41      36.36 -11.177 1.13e-08 ***
Stage.Q      -102.45      36.36  -2.817   0.0130 *
Stage.C        85.11      36.36   2.341   0.0335 *
Stage^4       -62.75      36.36  -1.726   0.1049
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 72.72 on 15 degrees of freedom
Multiple R-squared:  0.904,     Adjusted R-squared:  0.8785
F-statistic: 35.33 on 4 and 15 DF,  p-value: 1.805e-07
```



Relating the weights to the effects in the data set

	Orthogonal polynomial weights					*x*	*Coefficients*
	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5		
Linear	-0.63	-0.32	0.00	0.32	0.63	x	-406.41
Quadratic	0.53	-0.27	-0.53	-0.27	0.53	x	-102.45
Cubic	-0.32	0.63	0.00	-0.63	0.32	x	85.11
4th degree	0.12	-0.48	0.72	-0.48	0.12	x	-62.75





Predicting mean HA

Ex: Stage 1 HA, linear trend

Start with intercept, add scaled weights for the linear trend(coefficients multiplied by weights)

	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5
Linear	-0.63	-0.32	0.00	0.32	0.63
Quadratic	0.53	-0.27	-0.53	-0.27	0.53
Cubic	-0.32	0.63	0.00	-0.63	0.32
4th degree	0.12	-0.48	0.72	-0.48	0.12

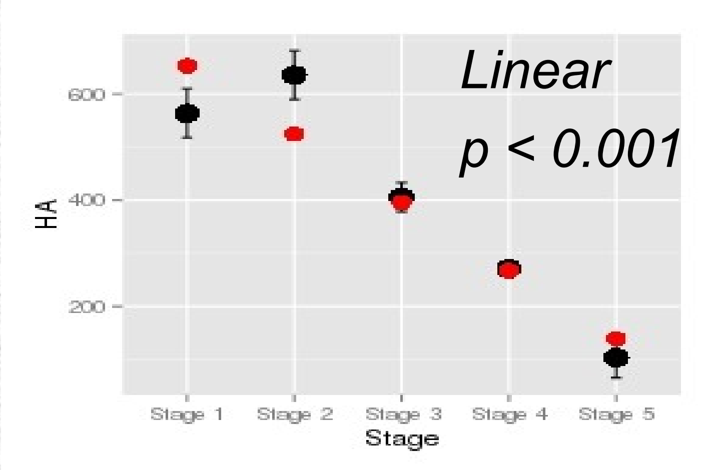
Stage 1: 395.94 – 406.41 (-0.63) = 651.97

Stage 2: 395.94 – 406.41 (-0.32) = 525.99

Stage 3: 395.94 – 406.41 (0) = 395.94

Stage 4: 395.94 – 406.41 (0.32) = 265.89

Stage 5: 395.94 – 406.41 (0.63) = 139.90



HA = 395.94 - 406.41 (linear)

Adding the quadratic trend

Start with linear trend, add the quadratic scaled weights (quadratic coefficient multiplied by the weights):

	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5
Linear	-0.63	-0.32	0.00	0.32	0.63
Quadratic	0.53	-0.27	-0.53	-0.27	0.53
Cubic	-0.32	0.63	0.00	-0.63	0.32
4th degree	0.12	-0.48	0.72	-0.48	0.12

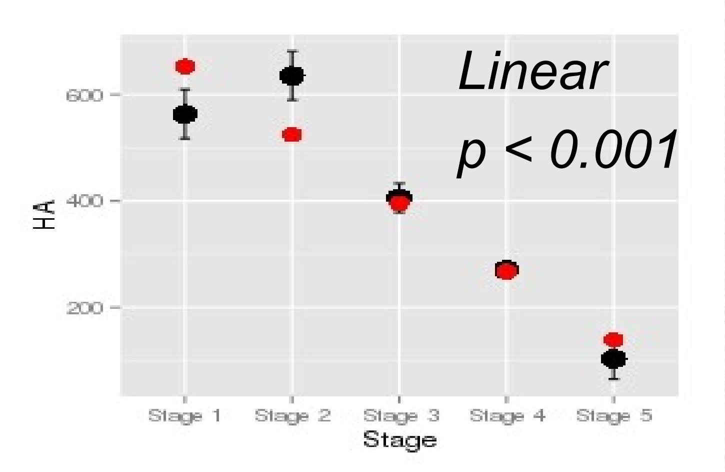
Stage 1: Intercept + Linear – 102.45 (0.53) = 597.67

Stage 2: Intercept + Linear – 102.45 (-0.27) = 553.67

Stage 3: Intercept + Linear – 102.45 (-0.53) = 450.24

Stage 4: Intercept + Linear – 102.45 (-0.27) = 293.55

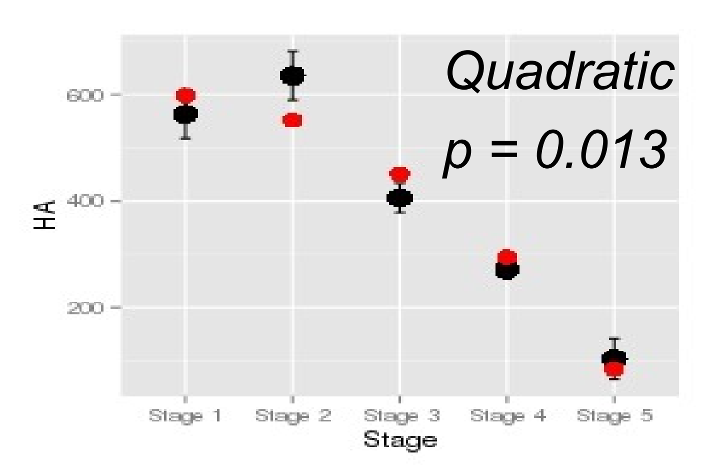
Stage 5: Intercept + Linear – 102.45 (0.53) = 85.60



Linear

$p < 0.001$

HA = 395.94 - 406.41 (linear)



Quadratic

$p = 0.013$

HA = 395.94 - 406.41 (linear)- 102.45 (quadratic)

Cubic trend

Start with the quadratic, and add the cubic scaled weights

	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5
Linear	-0.63	-0.32	0.00	0.32	0.63
Quadratic	0.53	-0.27	-0.53	-0.27	0.53
Cubic	-0.32	0.63	0.00	-0.63	0.32
4th degree	0.12	-0.48	0.72	-0.48	0.12

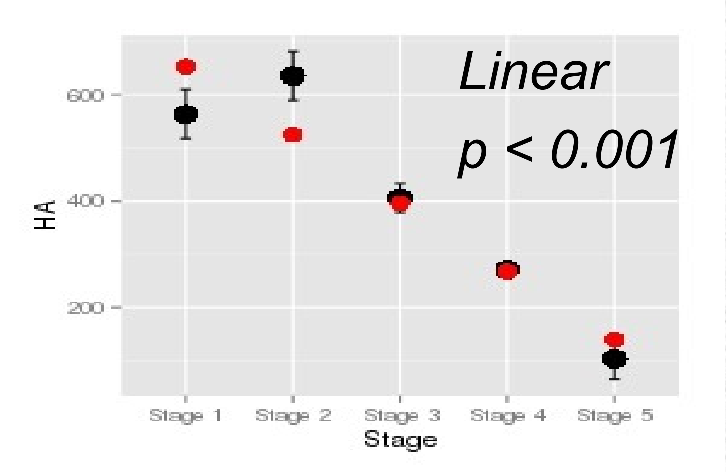
Stage 1: Intercept + Linear + Quadratic + 85.11 (-0.32) = 570.44

Stage 2: Intercept + Linear + Quadratic + 85.11 (0.63) = 607.27

Stage 3: Intercept + Linear + Quadratic + 85.11 (0) = 450.24

Stage 4: Intercept + Linear + Quadratic + 85.11 (-0.63) = 239.93

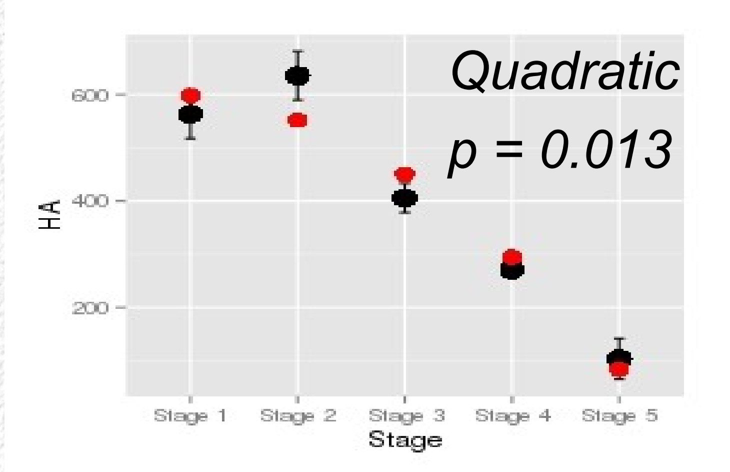
Stage 5: Intercept + Linear + Quadratic + 85.11 (0.32) = 112.84



Linear

$p < 0.001$

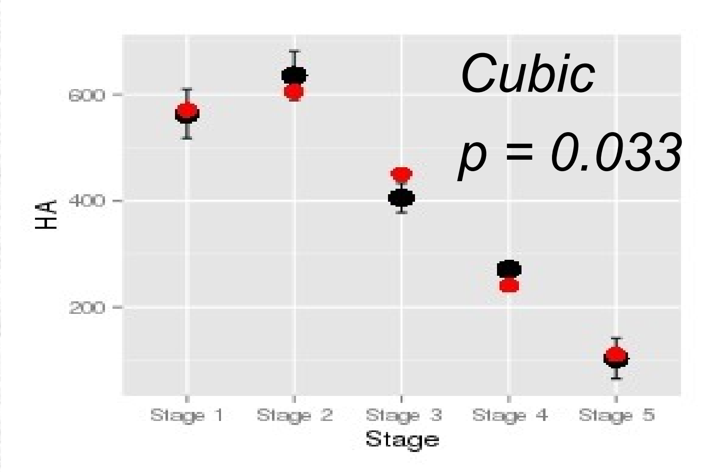
HA = 395.94 - 406.41 (linear)



Quadratic

$p = 0.013$

HA = 395.94 - 406.41 (linear)- 102.45 (quadratic)



Cubic

$p = 0.033$

HA = 395.94 - 406.41 (linear) - 102.45 (quadratic)
+ 85.11 (cubic)

4th degree trend

Add the 4th degree scaled weights to the cubic trend

	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5
Linear	-0.63	-0.32	0.00	0.32	0.63
Quadratic	0.53	-0.27	-0.53	-0.27	0.53
Cubic	-0.32	0.63	0.00	-0.63	0.32
4th degree	0.12	-0.48	0.72	-0.48	0.12

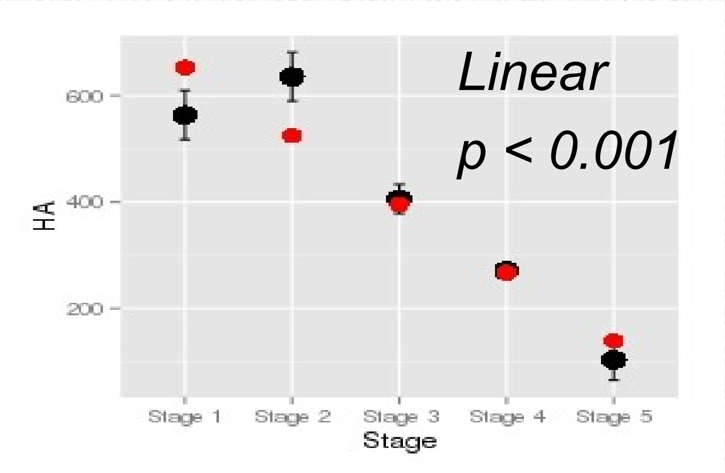
Stage 1: Intercept + Linear + Quadratic + Cubic - 62.75 (0.12) = 562.91

Stage 2: Intercept + Linear + Quadratic + Cubic - 62.75 (-0.48) = 637.39

Stage 3: Intercept + Linear + Quadratic + Cubic - 62.75 (0.72) = 405.06

Stage 4: Intercept + Linear + Quadratic + Cubic - 62.75 (-0.48) = 270.05

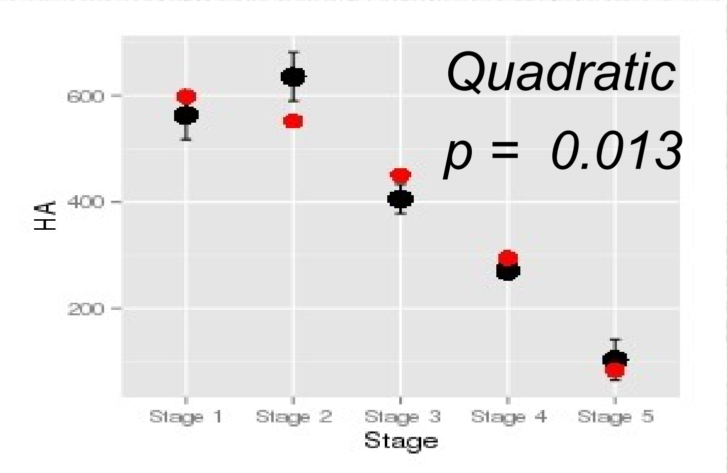
Stage 5: Intercept + Linear + Quadratic + Cubic - 62.75 (0.12) = 105.31



Linear

$p < 0.001$

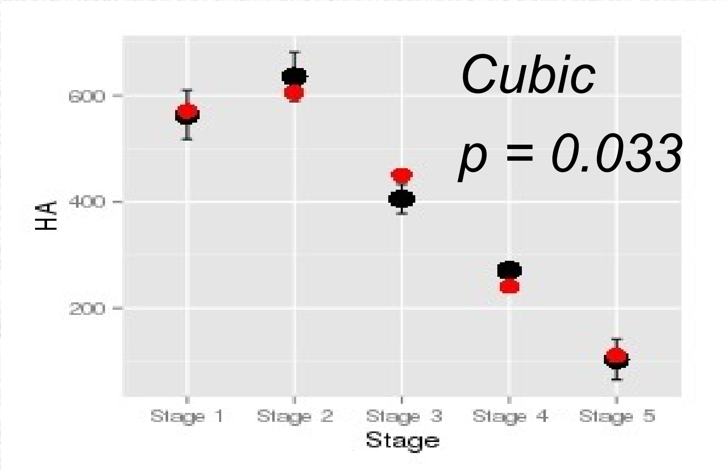
HA = 395.94 - 406.41 (linear)



Quadratic

$p = 0.013$

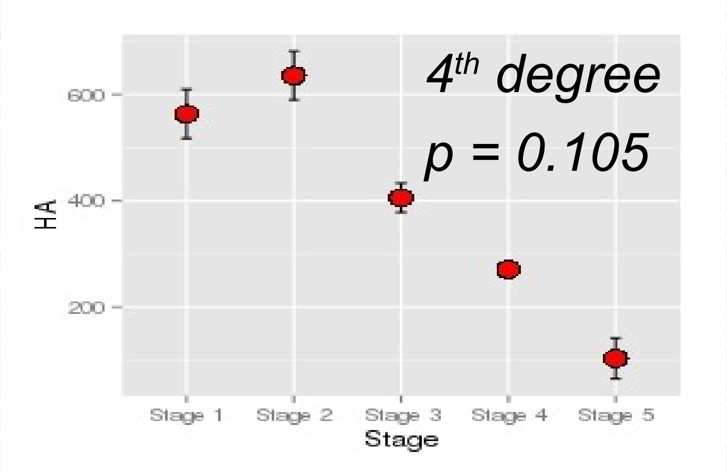
HA = 395.94 - 406.41 (linear)- 102.45 (quadratic)



Cubic

$p = 0.033$

HA = 395.94 - 406.41 (linear) - 102.45 (quadratic)
+ 85.11 (cubic)



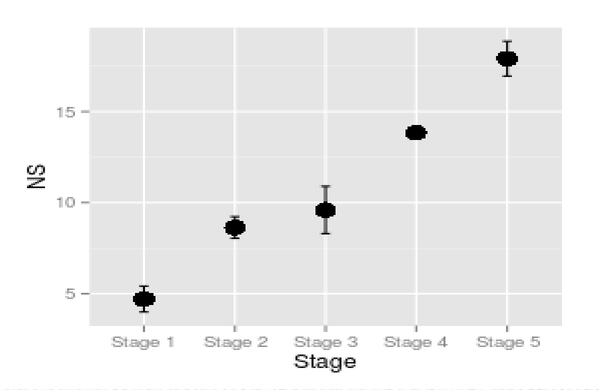
4th degree

$p = 0.105$

HA = 395.94 - 406.41 (linear) - 102.45 (quadratic)
+ 85.11 (cubic) - 62.75 (4th degree)

Results: NS

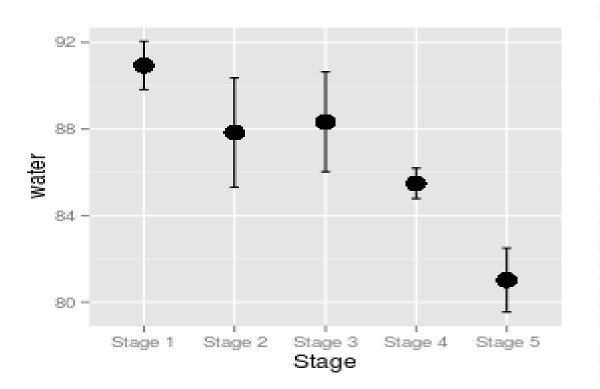
```
Call:
lm(formula = NS ~ Stage, data = eels)
Residuals:
    Min      1Q  Median      3Q     Max
-3.2575 -1.0269  0.2263  0.7137  2.4450
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.9370     0.3763  29.067 1.33e-14 ***
Stage.L       9.9928     0.8413  11.877 4.98e-09 ***
Stage.Q       0.9501     0.8413   1.129    0.277
Stage.C       0.8815     0.8413   1.048    0.311
Stage^4      -1.1597     0.8413  -1.378    0.188
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.683 on 15 degrees of freedom
Multiple R-squared:  0.9064,    Adjusted R-squared:  0.8815
F-statistic: 36.33 on 4 and 15 DF,  p-value: 1.496e-07
```



Which trend is significant?

Results: Water

```
Call:
lm(formula = water ~ Stage, data = eels)
Residuals:
    Min      1Q  Median      3Q     Max 
-4.9475 -1.9519 -0.8325  1.8881  6.6650 
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  86.7245     0.7913 109.600  < 2e-16 ***
Stage.L      -7.0131     1.7694  -3.964  0.00125 ** 
Stage.Q      -1.6243     1.7694  -0.918  0.37314    
Stage.C      -1.6507     1.7694  -0.933  0.36562    
Stage^4       1.0351     1.7694   0.585  0.56725    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 3.539 on 15 degrees of freedom
Multiple R-squared:  0.5422,    Adjusted R-squared:  0.4201 
F-statistic: 4.442 on 4 and 15 DF,  p-value: 0.01745
```



Which trend is significant?

Protecting experiment-wise error rates outside of ANOVA

- There are other situations that generate more than one non-independent p-value
  - Multiple predictor variables
  - Multiple response variables from the same subjects
- Post-hoc procedures only cover comparisons among levels of a single categorical predictor, don’t work with these
- Need to either:
  - Adjust α
  - Use model selection methods (more later)

Adjustments to α

- To achieve an experiment-wise α = 0.05 with 3 p-values, test each p-value at:

- Dunn-Šidák method $\alpha' = 1 - \sqrt[k]{(1-\alpha)} = 1 - \sqrt[3]{1-0.05} = 0.0169$

- Bonferroni $\alpha' = \alpha / k \quad = 0.05/3 = 0.0167$

- Advantage: these can be applied to any procedure (e.g. which of the eel ANOVA’s would still be significant at 0.0167?)

Eel analysis – three responses, same stages

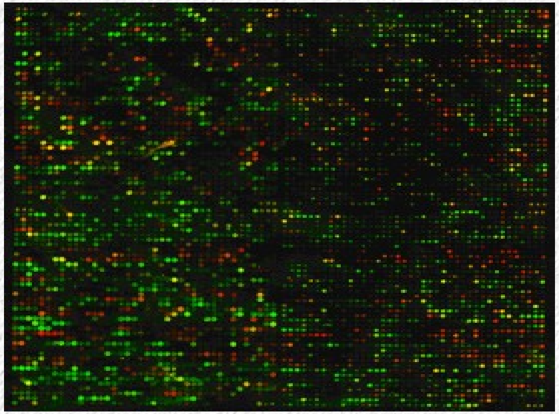
Variable	p	Un-adjusted	Bonferroni: 0.0167	Dunn-Šidák: 0.0169
HA	1.8e-7	Signif.	Signif.	Signif.
Water	0.0174	Signif.	NS	NS
NS	1.5e-7	Signif.	Signif.	Signif.

Bonferroni is always a little lower than Dunn-Šidák, and is thus more "conservative" = fewer significant differences will be found

The false discovery rate problem

- Exploratory data analysis is becoming more common
  - Data mining
  - Automated data collection on thousands of variables at once
- The number of “false discoveries” (Type I errors, false positives) may be huge
  - Expect 5% of the p-values to give us Type I errors if the null is true
  - With 1,000 p-values that's 50 false discoveries expected
- False discoveries waste time, money

Example: microarray analysis



- Microarrays express many, many genes (20,000 is not atypical)
- Expression measured by intensity of fluorescence on a chip
- Wish to separate those that are differentially expressed from those that are not
  - Any that are differentially expressed will be studied further
- Initially, differential expression was based on “fold change” (i.e. 2 fold increase, 3 fold increase, etc.)
- Fold change is an arbitrary criteria, not grounded in probability – better methods needed

The usual approaches: rock and a hard place

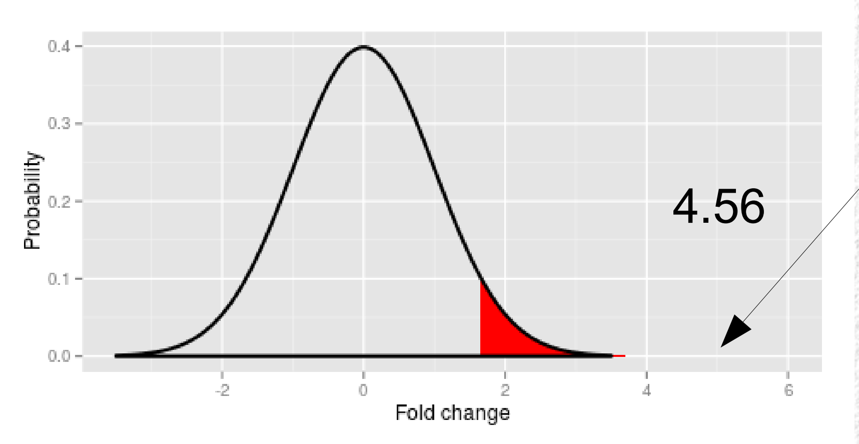
Rock: only very large differences will be significant

Bonferroni

$$\frac{0.05}{20{,}000} = 0.0000025$$

Dunn-Šidák method

$$1 - \sqrt[20000]{1 - 0.05} = 0.00000256$$



Hard place: with no adjustment expect huge number of false positives

$$20{,}000 \times 0.05 = 1{,}000$$

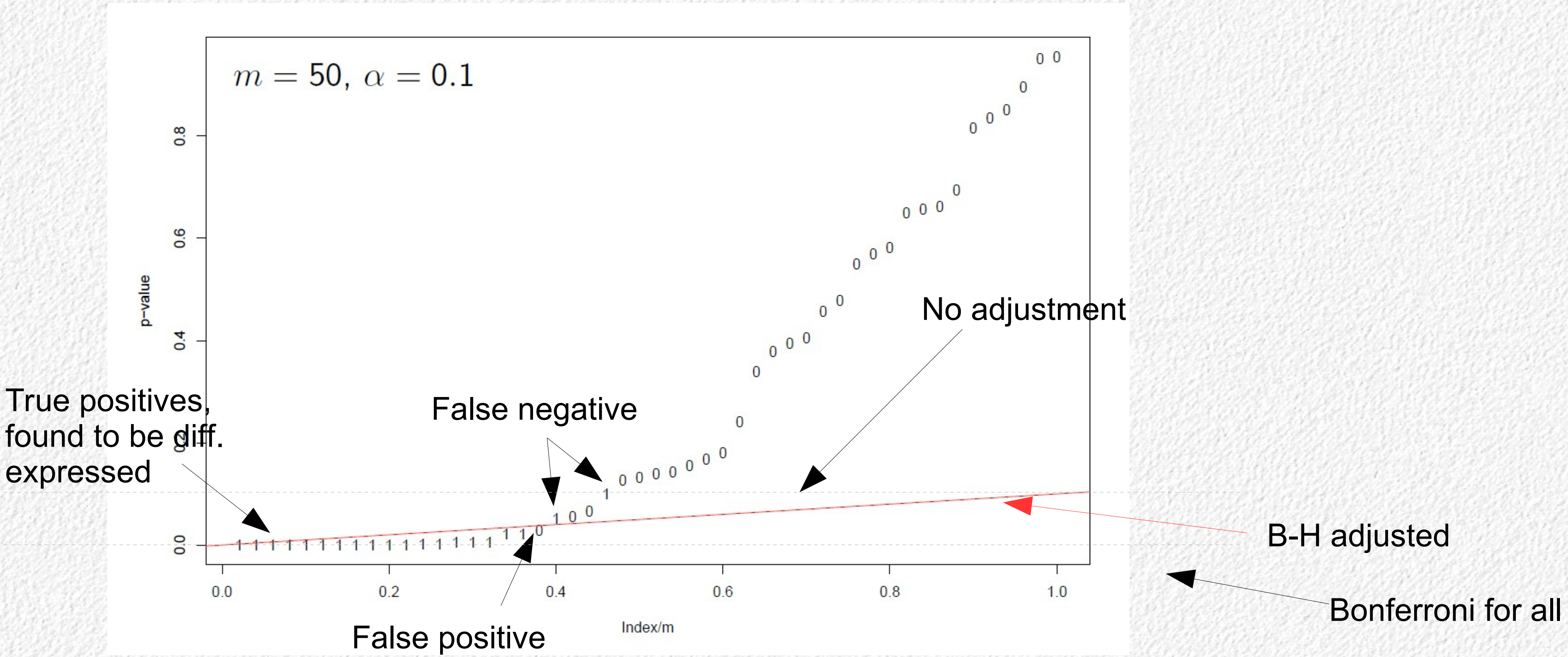
Benjamini and Hochberg's solution

- Calculate p-values for each gene (t-tests, ANOVA)
- Sort them from smallest to largest
- Test the smallest at the most stringent level
- Test successive p-values at increasingly less stringent level
- Specifically, for m tests, ordered from lowest to highest p-value, from k = 1 to m, and test at:

$$P_k \leq \frac{k}{m} \alpha$$

- E.g. test smallest p-value at α/m (Bonferroni), second at 2α/m, third at 3α/m, final (biggest) at mα/m = α.

A BH FDR graph for simulated data



True positives, found to be diff. expressed

False negative

No adjustment

B-H adjusted

Bonferroni for all

False positive

BH balances between excessive missed positives and excessive false positives, gives the lowest combination of false positives and false negatives