Class 15: Analysis of Variance II

Alison Presmanes Hill

all models are Wrong but some are Useful

George E. P. Box (1979). *Robustness in the strategy of scientific model building.* In Robustness in Statistics, pages 201-236. Academic Press.

http://www.dtic.mil/cgi-bin/GetTRDoc?Location=U2&doc=GetTRDoc.pdf&AD=ADA070213

there is no need to ask the question

Is the model true?

If the "truth" is to be the "whole truth" then the answer must be

No.

The only question of interest is :

Is the model illuminating and useful?

ANOVA: when and why

- When:
  - We want to compare means we can use a *t*-test. This test has limitations:
    - You can compare only 2 means: often we would like to compare means from 3 or more groups.
    - It can be used only with one Predictor/Independent Variable.
- ANOVA
  - Compares several means.
  - Can be used when you have manipulated more than one Independent Variables.
  - It is an extension of regression (the General Linear Model)

PLOS ONE

RESEARCH ARTICLE

Being Sticker Rich: Numerical Context Influences Children's Sharing Behavior

Tasha Posid[1]*, Allyse Fazio[2], Sara Cordes[2]

1 Department of Psychology, The Ohio State University, Columbus, Ohio, United States of America,
2 Department of Psychology, Boston College, Chestnut Hill, Massachusetts, United States of America

* posid.1@osu.edu

Being sticker rich

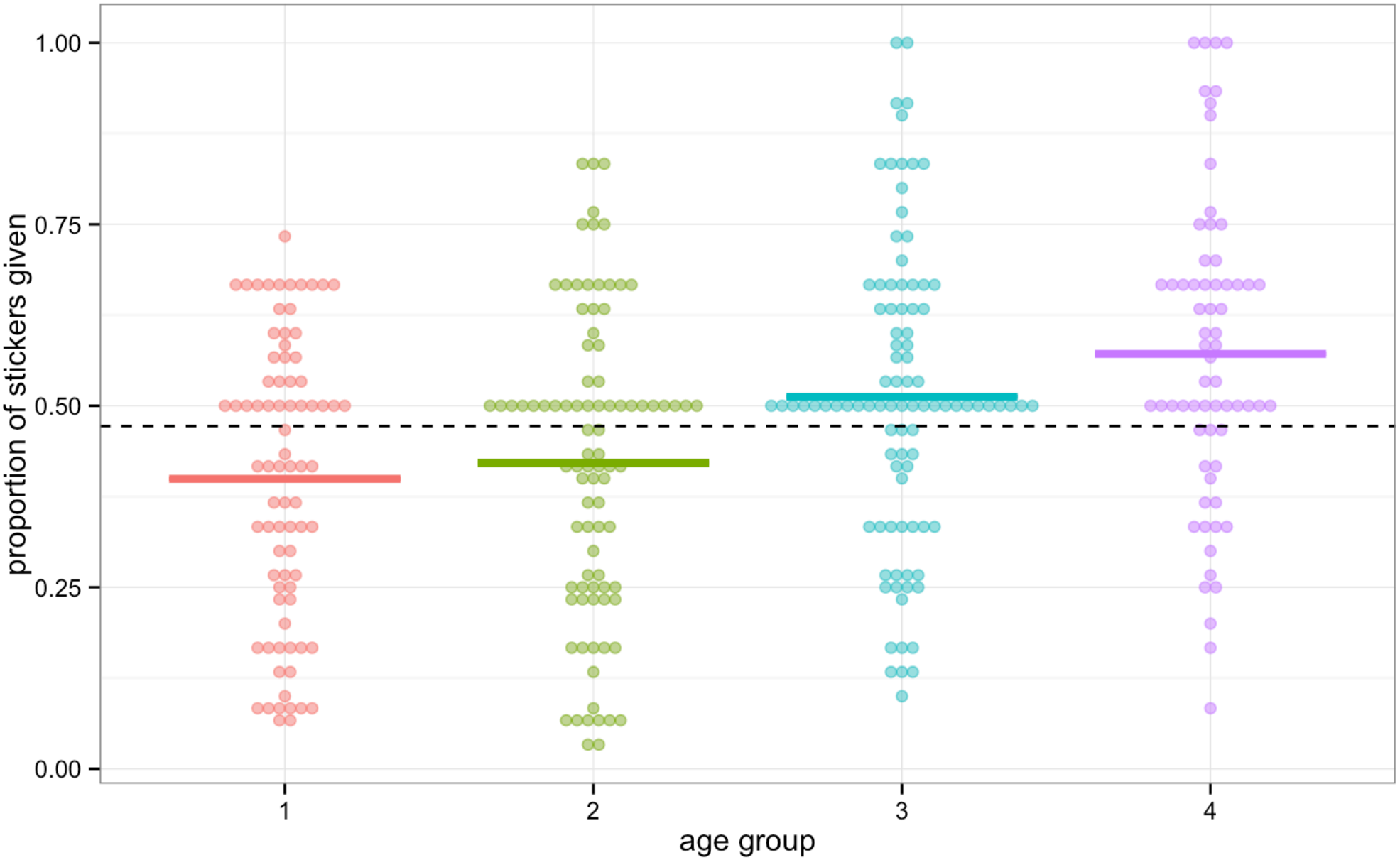
- Children (ages 3–11) received a small (12, “sticker poor”) or large (30, “sticker rich”) number of stickers, and were then given the opportunity to share their windfall with either one or multiple anonymous recipients (Dictator Game).
- Do the number of available resources and/or the number of potential recipients alter the likelihood of a child donating and/or the amount they donate?

Approach

- "Givers" only analyzed
- *"A univariate [read: one response/outcome variable] ANOVA was conducted investigating the impact of the between-subjects factors [read: all levels of factors are measured from independent samples] of age (4: 3–4 years, 5–6 years, 7–8 years, 9–11 years), number of resources (2: 12 or 30 stickers), number of recipients (2: 1 or 2 anonymous recipients), and gender (2: female, male) on the proportion of resources shared."*

Focus on one factor first…

- One-way ANOVA
  - Really only needed if > 2 levels for our one factor *(what if = 2 levels?)*
- Age group (4 levels)
  - 3–4 years,
  - 5–6 years,
  - 7–8 years,
  - 9–11 years


1.00
0.75
0.50
0.25
0.00
proportion of stickers given
1
2
3
4
age group

Let's start with garden variety linear regression

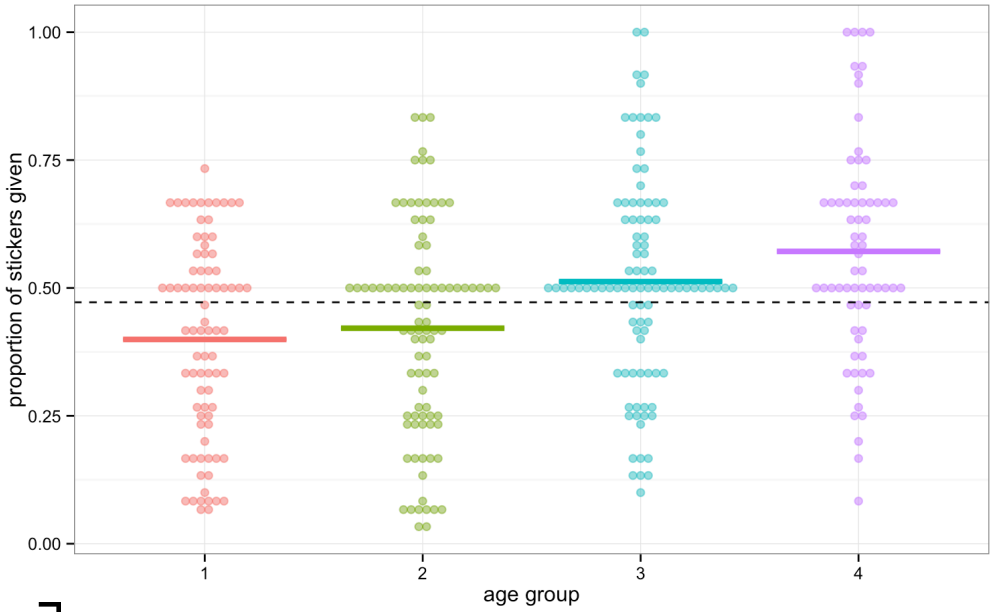
```
sticker_lm <- lm(prop_given ~ age_group, data = givers)
summary(sticker_lm)

Call:
lm(formula = prop_given ~ age_group, data = givers)

Residuals:
     Min       1Q   Median       3Q      Max
-0.49179 -0.15559 -0.00463  0.12047  0.48566

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.39938    0.02285  17.479  < 2e-16 ***
age_group2   0.02192    0.03140   0.698 0.485639
age_group3   0.11496    0.03116   3.689 0.000264 ***
age_group4   0.17575    0.03413   5.150 4.53e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2044 on 324 degrees of freedom
Multiple R-squared:  0.1004,        Adjusted R-squared:  0.09212
F-statistic: 12.06 on 3 and 324 DF,  p-value: 1.67e-07
```

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

```
means <- givers %>%
+   group_by(age_group) %>%
+   summarise(cell_means = mean(prop_given)) %>%
+   mutate(tx_effects = cell_means - cell_means[1])
means
```

```
Source: local data frame [4 x 3]
```

	age_group (fctr)	cell_means (dbl)	tx_effects (dbl)
1	1	0.3993750	0.0000000
2	2	0.4212963	0.0219213
3	3	0.5143369	0.1149619
4	4	0.5751282	0.1757532



$$\begin{bmatrix} Y_{11} \\ Y_{21} \\ \vdots \\ Y_{n_3 3} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \theta \\ \tau_2 \\ \tau_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{22} \\ \vdots \\ \varepsilon_{n_3 3} \end{bmatrix}$$

$$Y_{ij} = \theta + \tau_j + \varepsilon_{ij}, \text{ where } \tau_1 = 0$$

```
means
Source: local data frame [4 x 3]

  age_group cell_means tx_effects
     (fctr)      (dbl)      (dbl)
1         1  0.3993750  0.0000000
2         2  0.4212963  0.0219213
3         3  0.5143369  0.1149619
4         4  0.5751282  0.1757532
```

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{\alpha} = (\theta, \tau_2, \tau_3, \tau_4)$$

```
tidy(sticker_lm)
          term   estimate  std.error  statistic      p.value
1  (Intercept) 0.3993750 0.02284908 17.4788218 1.179975e-48
2   age_group2 0.0219213 0.03140306  0.6980625 4.856388e-01
3   age_group3 0.1149619 0.03116379  3.6889579 2.640087e-04
4   age_group4 0.1757532 0.03412684  5.1499998 4.531246e-07
```

$$\begin{bmatrix} Y_{11} \\ Y_{21} \\ \vdots \\ Y_{n_3 3} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \theta \\ \tau_2 \\ \tau_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{22} \\ \vdots \\ \varepsilon_{n_3 3} \end{bmatrix}$$

$$Y_{ij} = \theta + \tau_j + \varepsilon_{ij}, \text{ where } \tau_1 = 0$$

Omnibus ANOVA in R

```
sticker_lm <- lm(prop_given ~ age_group, data = givers)
anova(sticker_lm)
Analysis of Variance Table
Response: prop_given
            Df  Sum Sq Mean Sq F value       Pr(>F)
age_group    3  1.5111 0.50370   12.06 0.000000167 ***
Residuals  324 13.5323 0.04177
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$$\begin{bmatrix} Y_{11} \\ Y_{21} \\ \vdots \\ Y_{n_3 3} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{22} \\ \vdots \\ \varepsilon_{n_3 3} \end{bmatrix}$$

$$Y_{ij} = \mu_j + \varepsilon_{ij}$$

aov versus anova in R

- **`?aov`**:
  "aov is designed for balanced designs, and the results can be hard to interpret without balance: beware that missing values in the response(s) will likely lose the balance."
- Use **`anova(lm())`** if you don't have a balanced design
- But really, only use **`anova(lm())`** if you:
  - Have an unbalanced design (but not horribly so- generally bad idea)
  - Are certain you wish to assume equal variances across groups
    - Foreshadowing: Behrens-Fisher problem revisited
  - Only have one predictor (i.e., one-way ANOVA)
    - Foreshadowing: types of sums of squares

What are the ANOVA assumptions?

- The *k* samples are randomly selected from the k populations of interest
- Each of the *k* populations have a normal distribution
- All *k* populations have the same variance
- Why does this matter?
  - Our observed *F*-statistic is based on the assumption that when H_0 is true, our *F*-statistic will have an *F*-distribution

ANOVA assumptions

- Parallel those for linear regression, except for the assumption of linearity
- We assume:
  - Which group an observation falls into is fixed, not random
  - Residuals are all independent
  - Residuals are normally distributed
  - Residuals have constant variance

Checking ANOVA assumptions

A priori assumptions check

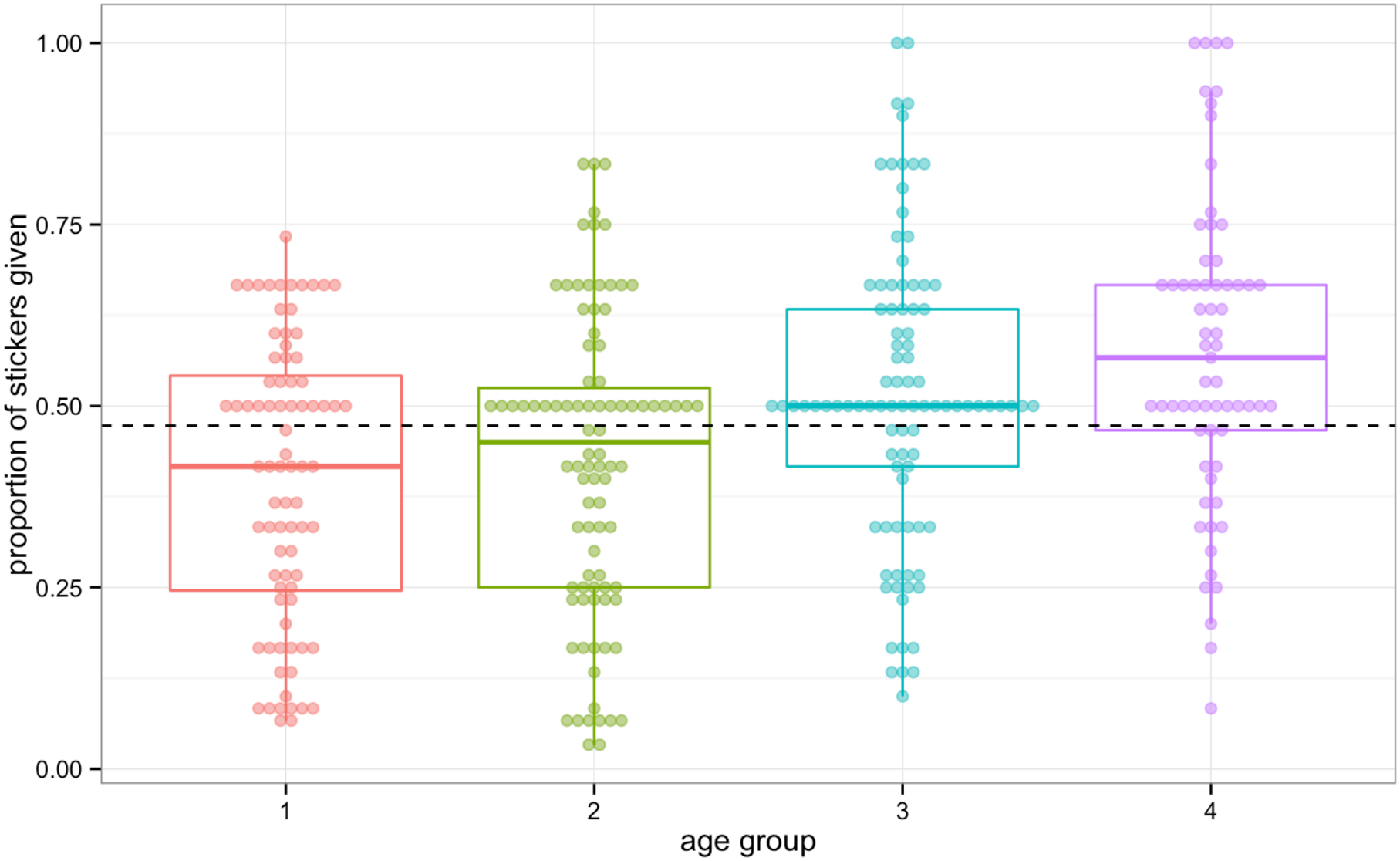
- Normality
  - Look at Q-Q plot by group
  - Look at skewness/kurtosis ratios by group
- Homogeneity of variance
  - Look at boxplots by group
  - Look at actual standard deviations by group
  - Run levene test

A posteriori assumptions check

- Normality of residuals
  - Look at scatterplot of residuals by group (what will be the mean?)
  - Look at Q-Q plot of residuals by group
  - Look at skewness/kurtosis ratios by group
- Homogeneity of residuals variance
  - Look at boxplots of residuals by group
  - Look at actual standard deviations by group
  - Run levene test

- The plot of each sample's values against its mean (or its sample ID) will consist of vertical "stacks" of data points, one stack for each unique sample mean value. If the assumptions for the samples' population distributions are correct, the stacks should be about the same length. Outliers may appear as anomalous points in the graph.
- A fan pattern like the profile of a megaphone, with a noticeable flare either to the right or to the left (one or more of the "stacks" of data points is much longer than the others), suggests that the variance in the values increases in the direction the fan pattern widens (usually as the sample mean increases), and this in turn suggests that a transformation may be needed.

Side-by-side boxplots of the samples can also reveal lack of homogeneity of variances if some boxplots are much longer than others, and reveal suspected outliers.



Unequal variances, one-way

```
oneway.test(prop_given  ~ age_group,  data = givers)

        One-way analysis of means (not assuming equal variances)

data:  prop_given and age_group
F = 11.413, num df = 3.00, denom df = 174.57, p-value = 7.167e-07
```

B. L. Welch (1951), On the comparison of several mean values: an alternative approach. *Biometrika*, **38**, 330–336.

Casella & Berger (2001)

11.2.2 The Classic ANOVA Hypothesis

The classic ANOVA test is a test of the null hypothesis

$$H_0: \quad \theta_1 = \theta_2 = \cdots = \theta_k,$$

a hypothesis that, in many cases, is silly, uninteresting, and not true. An experimenter would not usually believe that the different treatments have *exactly* the same mean.

Why do contrasts?

- The *F*-ratio tells us only that the experiment was successful
  - i.e. group means were different
- It does not tell us specifically which group means differ from which.
- We need additional contrasts to find out where the group differences lie.

How?

- Multiple *t*-tests
  - We saw earlier that this is a bad idea
- Orthogonal Contrasts/Comparisons
  - Hypothesis driven
  - Planned a priori
- *Post Hoc* Tests
  - Not Planned (no hypothesis)
  - Compare all pairs of means
- Trend Analysis

Planned Contrasts

- Basic Idea:
  - The variability explained by the Model (Model sums of squares) is due to participants being assigned to different groups.
  - This variability can be broken down further to test specific hypotheses about which groups might differ.
  - We break down the variance according to hypotheses made *a priori* (before the experiment).

Contrasts

- Let $\boldsymbol{t = (t_1, \ldots, t_k)}$ be a set of variables, either parameters or statistics, and let $\boldsymbol{a = (a_1, \ldots, a_k)}$ be known as constants. The function:

$$\sum_{i=1}^{k} a_i t_i$$

- If $\boldsymbol{a}$ linear combination of $\boldsymbol{t_i s}$; if

$$\sum_{i=1}^{k} a_i = 0$$

- Then it is called a **contrast**.

For example:

- Let's say we have means $\theta_1, \ldots, \theta_k$ and constants a = (1, -1, 0, …, 0), then:

$$\sum_{i=1}^{k} a_i \theta_i = \theta_1 - \theta_2$$

- Is a contrast that compares θ_1 to θ_2

$$H_0 = \sum_{i=1}^{k} a_i \theta_i = 0 \text{ for all } (a_1, \ldots, a_k) \text{ such that } \sum_{i=1}^{k} a_i = 0$$

$$H_1 = \sum_{i=1}^{k} a_i \theta_i \neq 0 \text{ for all } (a_1, \ldots, a_k) \text{ such that } \sum_{i=1}^{k} a_i = 0$$

The ANOVA null is really just a statement about contrasts, and allows us to think and operate in a univariate space...

What is a contrast matrix?

- In R, we parameterize our contrasts (just linear combinations of the cells means) in a matrix
- The rows of this matrix correspond to the levels of the factor, and there is one column for each predictor included in the model
- The coding scheme defined by the contrasts attribute in R is the *inverse* of the matrix of contrast weights.

```
dummy <- contrasts(givers$age_group)
dummy
  2 3 4
1 0 0 0
2 1 0 0
3 0 1 0
4 0 0 1
```

Coefs

Group means

```
solve(cbind(1, dummy))
  1 2 3 4
  1 0 0 0
2 -1 1 0 0
3 -1 0 1 0
4 -1 0 0 1
```

Contrast matrices in R

```
contrasts(givers$age_group)
  2 3 4
1 0 0 0
2 1 0 0
3 0 1 0
4 0 0 1
contr.treatment(4)
  2 3 4
1 0 0 0
2 1 0 0
3 0 1 0
4 0 0 1
```

The default contrast setting in R here is called "dummy coding" in that each level is compared to the "reference level", so the intercept is the cell mean of the reference group (here, age group = 1)

Group means

Coefs

```
solve(cbind(1, dummy))
  1 2 3 4
  1 0 0 0
2 -1 1 0 0
3 -1 0 1 0
4 -1 0 0 1
```

Linear regression with dummy coding

```
contrasts(givers$age_group) <- contr.treatment(4)
sticker_dummy <- lm(prop_given ~ age_group, data = givers)
summary(sticker_dummy)

<snip, snip>

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept)  0.39938    0.02285  17.479 < 0.0000000000000002 ***
age_group2   0.02192    0.03140   0.698             0.485639
age_group3   0.11496    0.03116   3.689             0.000264 ***
age_group4   0.17575    0.03413   5.150          0.000000453 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2044 on 324 degrees of freedom
Multiple R-squared:  0.1004,  Adjusted R-squared:  0.09212
F-statistic: 12.06 on 3 and 324 DF,  p-value: 0.000000167
```

```
solve(cbind(1, dummy))
  1 2 3 4
  1 0 0 0
2 -1 1 0 0
3 -1 0 1 0
4 -1 0 0 1
```

The default contrast setting for "lm"

Contrast matrices in R

```
solve(cbind(1, simple_c))
     1    2    3    4
1 0.25 0.25 0.25 0.25
2 -1.00 1.00 0.00 0.00
3 -1.00 0.00 1.00 0.00
4 -1.00 0.00 0.00 1.00
```

```
c <- contr.treatment(4)
simple_mat <- matrix(rep(1/4, 12), ncol = 3)
simple_mat
     [,1] [,2] [,3]
[1,] 0.25 0.25 0.25
[2,] 0.25 0.25 0.25
[3,] 0.25 0.25 0.25
[4,] 0.25 0.25 0.25
simple_c <- c - simple_mat
simple_c
      2     3     4
1 -0.25 -0.25 -0.25
2  0.75 -0.25 -0.25
3 -0.25  0.75 -0.25
4 -0.25 -0.25  0.75
contrasts(givers$age_group) <- simple_c
```

Let's do "simple coding" instead, which still compares each level to the reference level, intercept being the grand mean across all 4 groups

Linear regression with simple coding

```
contrasts(givers$age_group) <- simple_c
sticker_simple <- lm(prop_given ~ age_group, data = givers)
summary(sticker_simple)
```

```
solve(cbind(1, simple_c))
      1    2    3    4
  0.25 0.25 0.25 0.25
2 -1.00 1.00 0.00 0.00
3 -1.00 0.00 1.00 0.00
4 -1.00 0.00 0.00 1.00
```

```
<snip, snip>

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.47753    0.01140  41.905  < 2e-16 ***
age_group2   0.02192    0.03140   0.698 0.485639
age_group3   0.11496    0.03116   3.689 0.000264 ***
age_group4   0.17575    0.03413   5.150 4.53e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2044 on 324 degrees of freedom
Multiple R-squared:  0.1004,      Adjusted R-squared:  0.09212
F-statistic: 12.06 on 3 and 324 DF,  p-value: 1.67e-07
```

Linear regression with deviation coding

```
contr.sum(4)
  [,1] [,2] [,3]
1    1    0    0
2    0    1    0
3    0    0    1
4   -1   -1   -1
```

```
solve(cbind(1, contr.sum(4)))
          1     2     3     4
[1,]  0.25  0.25  0.25  0.25
[2,]  0.75 -0.25 -0.25 -0.25
[3,] -0.25  0.75 -0.25 -0.25
[4,] -0.25 -0.25  0.75 -0.25
```

```
contrasts(givers$age_group) <- contr.sum(4)
sticker_dev <- lm(prop_given ~ age_group, data = givers)
summary(sticker_dev)

<snip, snip>

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.47753    0.01140  41.905  < 2e-16 ***
age_group1  -0.07816    0.01977  -3.953 9.47e-05 ***
age_group2  -0.05624    0.01902  -2.956  0.00334 **
age_group3   0.03680    0.01883   1.955  0.05145 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2044 on 324 degrees of freedom
Multiple R-squared:  0.1004,  Adjusted R-squared:  0.09212
F-statistic: 12.06 on 3 and 324 DF,  p-value: 1.67e-07
```

Let's do "deviation coding" instead, which compares each level to the grand mean

```
contr.sum(4)
  [,1] [,2] [,3]
1    1    0    0
2    0    1    0
3    0    0    1
4   -1   -1   -1
```

$$\beta_0 = (\mu_1 + \mu_2 + \mu_3 + \mu_4)/4$$
$$\beta_1 = \mu_1 - (\mu_1 + \mu_2 + \mu_3 + \mu_4)/4$$
$$\beta_2 = \mu_2 - (\mu_1 + \mu_2 + \mu_3 + \mu_4)/4$$
$$\beta_3 = \mu_3 - (\mu_1 + \mu_2 + \mu_3 + \mu_4)/4$$

$$\mu_4 - (\mu_1 + +\mu_2 + \mu_3 + \mu_4)/4 = -(\beta_1 + \beta_2 + \beta_3)$$

```
solve(cbind(1, contr.sum(4)))
     1     2     3     4
[1,]  0.25  0.25  0.25  0.25
[2,]  0.75 -0.25 -0.25 -0.25
[3,] -0.25  0.75 -0.25 -0.25
[4,] -0.25 -0.25  0.75 -0.25
```

```
solve(cbind(1, contr.sum(4)))
     1     2     3     4
[1,] 0.25  0.25  0.25  0.25
[2,] 0.75 -0.25 -0.25 -0.25
[3,] -0.25  0.75 -0.25 -0.25
[4,] -0.25 -0.25  0.75 -0.25
```

$$\beta_0 = (\mu_1 + \mu_2 + \mu_3 + \mu_4)/4$$
$$\beta_1 = \mu_1 - (\mu_1 + \mu_2 + \mu_3 + \mu_4)/4$$
$$\beta_2 = \mu_2 - (\mu_1 + \mu_2 + \mu_3 + \mu_4)/4$$
$$\beta_3 = \mu_3 - (\mu_1 + \mu_2 + \mu_3 + \mu_4)/4$$

$$\beta_0 = .25\mu_1 + .25\mu_2 + .25\mu_3 + .25\mu_4$$
$$\beta_1 = .75\mu_1 - .25\mu_2 - .25\mu_3 - .25\mu_4$$
$$\beta_2 = -.25\mu_1 + .75\mu_2 - .25\mu_3 - .25\mu_4$$
$$\beta_3 = -.25\mu_1 - .25\mu_2 + .75\mu_3 - .25\mu_4$$

These all say the same thing. This is typically the set of contrasts you wish to test with ANOVA

Why am I boring all of you with this?

- Once you get beyond two group comparisons, you need to know a bit about how factors are utilized in linear models and what the resulting parameter estimates mean. One day you may even want to exert control on this.
- A popular R package for performing linear modelling for thousands of, e.g., genes at once, while borrowing strength across the genes, is called limma. And, unlike lm(), limma does NOT make the design matrix for you. limma does not use the same formula interface as lm().

Multiple comparisons

- At this point, you may be thinking, why do an ANOVA at all? Why not just do a whole heap of t-tests?
- Proliferation of Type I error → family-wise error rates
- Post-hoc inference

Multiple comparison procedures

Family type	Simultaneous methods		Sequential methods	
	Equal variances	**Unequal variances**	**Equal variances**	**Unequal variances**
Planned	Dunn-Bonferroni	Dunn-Bonferroni using Welch's t'	Hochberg	Hochberg using Welch's t'
All pairwise	Tukey HSD (equal n) or Tukey-Kramer (unequal n)	Games-Howell or Dunnett T3	Fisher-Hayter	
Experimental vs control	Dunn-Bonferroni (unequal n)	Dunn-Bonferroni using Welch's t'		
Post-hoc	Sheffé	Sheffé using Welch's t'		

The “Big 3” MCPs

- Planned contrasts
  - Bonferroni (orthogonal contrasts)
  - Holm (non-orthogonal, overlapping contrasts)
  - Hochberg (non-orthogonal, overlapping contrasts)
- Sheffé
- Tukey HSD

Planned contrasts

- Sometimes we are less interested in the overall *F* test than in testing highly specific linear combination hypotheses

- Planned orthogonal pair-wise contrasts
  - In some cases, these hypotheses "break down" the information to a set of means into non-overlapping components that are statistically uncorrelated
  - For example, contrasting groups A & B is orthogonal to contrasting groups C & D (they are linearly independent)
  - When contrasts are all orthogonal, can use ordinary t-tests with Bonferroni adjusted *p*-values (*k*-1 possible orthogonal contrasts)

- Planned pair-wise contrasts (overlapping- not orthogonal)
  - Contrasting groups A & B is NOT orthogonal to comparing B & C
  - When this happens, we gain power by accounting for that overlap (Bonferroni will be overly conservative, increased Type II errors)

Bonferroni

- If all k tests are done at α level, then the FWE must be less than or equal to $k\alpha$
- Thus, we use a new critical p-value for each contrast equal to α/k to keep FEW less than or equal to α
- Pros: easy/straightforward
- Cons: raises possibility of Type II errors, may be overly conservative
- Several procedures developed early on to improve on Bonferroni for non-orthogonal contrasts

Holm: Method of Closure

- Assume that the *k* tests are independent; that is, the falsity of one hypothesis is not precluded by the falsity of another

- Rank order the p-values for your k tests from lowest to highest

- Look at the lowest p-value first; if it is < α/k, reject null for that test and move on to the next. If not, stop.

Holm: Method of Closure

- For each successive test, reduce divisor by 1 when previous was significant
- For example, for second test, compare the second smallest p-value to α/k-1
- Continue until you fail to reject (p > adjusted α)
- This method has been shown to have more statistical power than Bonferroni

Hochberg: False Discovery Rate

- Rank order the p-values for your k tests from lowest to highest
- Look at the lowest p-value first; if it is < $1^{*}\alpha/k$, reject null for that test and move on to the next. If not, stop.
- For each successive test, add 1 to the numerator multiplier when previous was significant (leave denominator at *k*)
- For example, for second test, compare the second smallest p-value to $(2^{*}\alpha)/k$

Scheffé

- Allows you to perform *any* contrast hypothesis test after having viewed the data
- Provides FEW protection AND protection against post-hoc cherry-picking of data
- High protection → high cost
- The larger the family of tests, the more protection needed, lower statistical power

Scheffé

- Compute the generalized t-statistic exactly the way we have computed it in class
- Then compare to a *new* t-critical value, t'

$$t' = \sqrt{(k-1)F_{\alpha,k-1,N-k}}$$

- *k* = total number of independent groups before *contrast* hypothesis
- What does the above formula assume?
- Note: I use t' here instead of F' because you use the t' to calculate confidence intervals in place of t

Tukey HSD

- You are interested in performing all possible pairwise contrasts between pairs of groups of means

- By stating this family of contrasts a priori, we get full protection at a lower cost

Tukey HSD

- Works from two-sample t-test assuming equal variances

$$t_{obs} = \frac{\bar{Y}_{\bullet 1} - \bar{Y}_{\bullet 2}}{\sqrt{s^2_{pooled}\left[\left(1/n_1\right)+\left(1/n_2\right)\right]}}$$

- Standard t-test: compute a t-statistic and compare to a critical value from the t-distribution (standard t-test)

- Tukey suggested a new statistic, q, to compare to a new critical value from the studentized range distribution

Tukey HSD

- The HSD is the minimum difference between each possible pair of sample means that would be larger than one would expect by chance

- q is the same for all pair-wise contrasts: it is your new critical value to beat

- What does this imply about the 95% confidence interval for q?

Tukey HSD procedure

- Find q, the critical value for the studentized range statistic (qtukey()), for α where $q_{k\ ,N-k}$
  - Note that these degrees of freedom are *not* the same as the overall *F* df in the ANOVA

$$HSD = q\sqrt{\frac{MSE}{n}}$$

- Where MSE = mean square error or mean square within-groups for between-groups ANOVA
- What does the above formula assume?

Planned contrasts

- Similar to t-tests with one important distinction
  - For σhat^2, you use the data from all the groups (e.g., MS_{wg} or MSE) , regardless of whether a group is involved (i.e., has a nonzero weight) in the linear combination contrast
  - For standard t-test, what is our σhat^2 assuming equal variances?

$$s^2_{pooled} = \frac{SS_{wg1} + SS_{wg2}}{n_1 + n_2 - 2}$$

$$MS_{wg} = \frac{SS_{wg}}{k(n-1)}$$

False positive rate

	Call based on observed data		
True state of the world	Fail to reject H_0	Reject H_0	
H_0	True negative $1 - \alpha$	False positive Type I error α	# true H_0's
H_1	False negative Type II error β	True positive $1 - \beta$	# true H_1's
		# rejected H_0's	# total tests

Family-wise error rate: P(≥ 1 false positive)

	Call based on observed data		
True state of the world	Fail to reject H_0	Reject H_0	
H_0	True negative $1 - \alpha$	False positive Type I error α	# true H_0's
H_1	False negative Type II error β	True positive $1 - \beta$	# true H_1's
		# rejected H_0's	# total tests

Power: probability of true positive

	Call based on observed data		
True state of the world	Fail to reject H_0	Reject H_0	
H_0	True negative $1 - \alpha$	False positive Type I error α	# true H_0's
H_1	False negative Type II error β	True positive $1 - \beta$	# true H_1's
		# rejected H_0's	# total tests

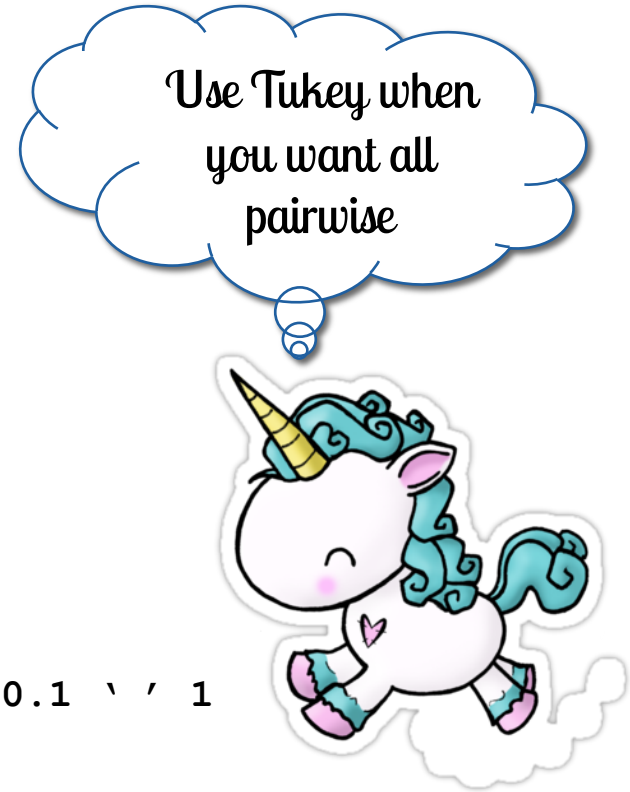
False discovery rate: E(# false pos / # discoveries)

	Your decision based on observed data		
True state of the world	Fail to reject H_0	Reject H_0	
H_0	True negative $1 - \alpha$	False positive Type I error α	# true H_0's
H_1	False negative Type II error β	True positive $1 - \beta$	# true H_1's
		# rejected H_0's	# total tests

Contrasts when variances are equal

```
# library(multcomp)
sticker_mcp <- glht(sticker_lm, mcp(age_group = "Tukey"))
confint(sticker_mcp)
summary(sticker_mcp, test = univariate()) #unadjusted p values
summary(sticker_mcp, test = adjusted("bonferroni")) #p value adjustment
```

```
          Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Linear Hypotheses:
            Estimate Std. Error t value   Pr(>|t|)
2 - 1 == 0  0.02192    0.03140   0.698    1.00000
3 - 1 == 0  0.11496    0.03116   3.689    0.00158 **
4 - 1 == 0  0.17575    0.03413   5.150 0.00000272 ***
3 - 2 == 0  0.09304    0.03022   3.079    0.01353 *
4 - 2 == 0  0.15383    0.03327   4.624 0.00003263 ***
4 - 3 == 0  0.06079    0.03304   1.840    0.40017
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- bonferroni method)
```



Contrasts when variances are unequal

```
with(givers, pairwise.t.test(prop_given, age_group, pool.sd = FALSE, paired = FALSE, p.adjust = "bonferroni"))
            Pairwise comparisons using t tests with non-pooled SD
data:  prop_given and age_group

  1        2        3
2 1.0000   -        -
3 0.0013   0.0152   -
4 0.000006 0.000090 0.4482

P value adjustment method: bonferroni
# library(DTK) for Dunnett T3 (aka Dunnett-Tukey-Kramer)
sticker_dtk <- DTK.test(x = givers$prop_given, f = givers$age_group, a = 0.05)
sticker_dtk
[[1]]
[1] 0.05
[[2]]
          Diff    Lower CI  Upper CI
2-1 0.02192130 -0.05883434 0.1026769
3-1 0.11496192  0.03541347 0.1945104
4-1 0.17575321  0.08602450 0.2654819
3-2 0.09304062  0.01343253 0.1726487
4-2 0.15383191  0.06405540 0.2436084
4-3 0.06079129 -0.02829979 0.1498824
DTK.plot(sticker_dtk)Cannot use TukeyHSD
```