STAT 113
Comparing Multiple Means

Colin Reimer Dawson

Oberlin College

December 5, 2017

Outline

Comparing Multiple Means

A Randomization Test

The F-statistic

Inferences After ANOVA

Exercise and Changes in Brain Size

Researchers in China recently investigated whether different kinds of exercise/activity might help to prevent brain shrinkage or perhaps even lead to an increase in brain size (Mortimer et al., 2012). The researchers randomly assigned elderly adult volunteers into four activity groups: tai chi, walking, social interaction, and no intervention. Each participant had an MRI to determine brain size before the study began and again at its end. The researchers measured the percentage increase or decrease in brain size during that time.

Variables and Hypotheses

1. Here, the response variable (change in brain size) is quantitative, and the explanatory variable (activity group) is categorical.
2. A natural set of parameters to focus on is the typical response in each group. For example, focus on the four group population means of the change in brain size variable.
4. For activity and change in brain size to be associated, that would mean that the group distributions are not identical. In particular, we would expect the *means* to differ:

$$H_0 : \mu_{\text{TaiChi}} = \mu_{\text{Walking}} = \mu_{\text{Social}} = \mu_{\text{Nothing}}$$
$$H_1 : \text{At least one } \mu \text{ differs from at least one other}$$

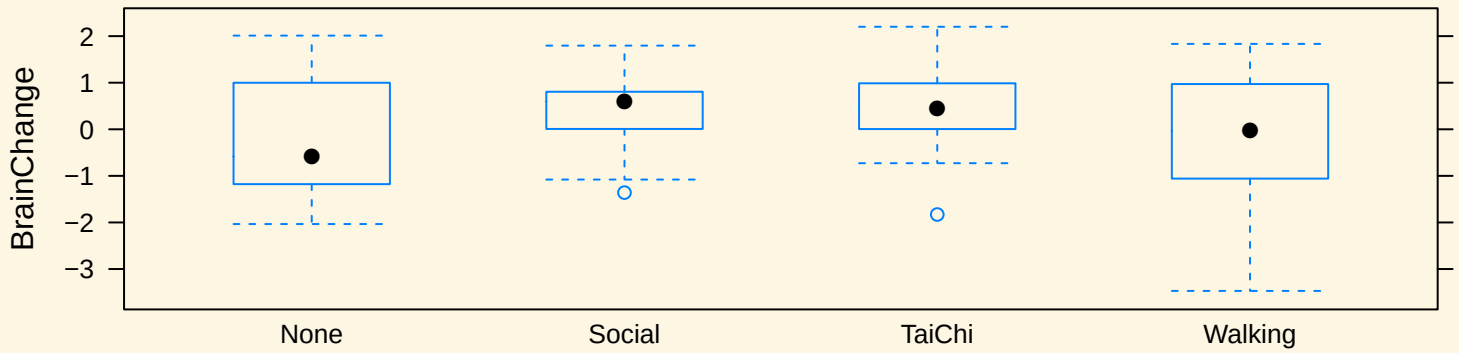
The Data

```
Brain <- read.file("http://colinreimerdawson.com/data/brain_size.txt")
sample(Brain) %>% head()
```

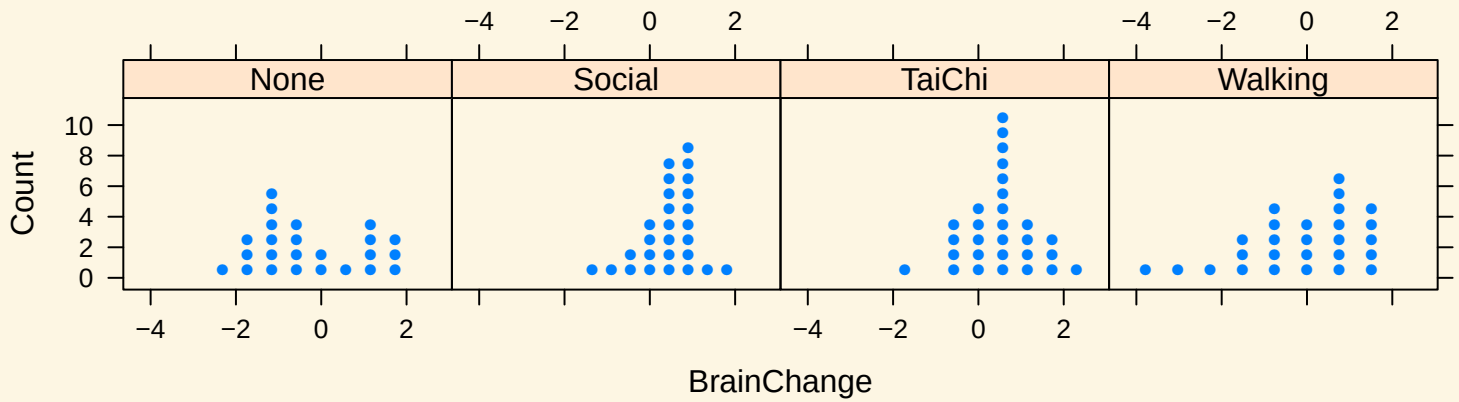
```
      Treatment BrainChange orig.id
50   Walking       1.492      50
48   Walking       1.145      48
59    Social       0.276      59
84      None      -1.347      84
74    Social       0.596      74
29    TaiChi       2.201      29
```

How does it look?

```
bwplot(BrainChange ~ Treatment, data = Brain)
```




```
dotPlot(~BrainChange | Treatment, data = Brain)
```



Descriptive Stats

```
favstats(BrainChange ~ Treatment, data = Brain)
```

	Treatment	min	Q1	median	Q3	max	mean	sd	n	missing
1	None	-2.034	-1.16875	-0.585	0.9725	2.011	-0.2401250	1.2584309	24	0
2	Social	-1.359	0.00750	0.596	0.8060	1.796	0.4056296	0.6968969	27	0
3	TaiChi	-1.829	0.00500	0.449	0.9870	2.201	0.4710690	0.8557466	29	0
4	Walking	-3.470	-1.05850	-0.026	0.9710	1.833	-0.1503333	1.3868388	27	0

A Randomization Test

- We are testing for an association. We can randomize by randomly pairing responses and group assignments. Randomly re-group the data.
- But how to we measure deviation from expectations under H_0?

Possible Test Statistics

- Take $\bar{x}_{\text{largest}} - \bar{x}_{\text{smallest}}$
- Take average of all pairwise absolute differences:

$$\frac{|\bar{x}_2 - \bar{x}_1| + |\bar{x}_3 - \bar{x}_1| + |\bar{x}_4 - \bar{x}_1| + |\bar{x}_3 - \bar{x}_2| + |\bar{x}_4 - \bar{x}_2| + |\bar{x}_4 - \bar{x}_3|}{6}$$

- Take standard deviation of sample means:

$$\sqrt{\frac{\sum_{g=1}^{G}(\bar{x}_g - \bar{\bar{x}})^2}{G-1}}$$

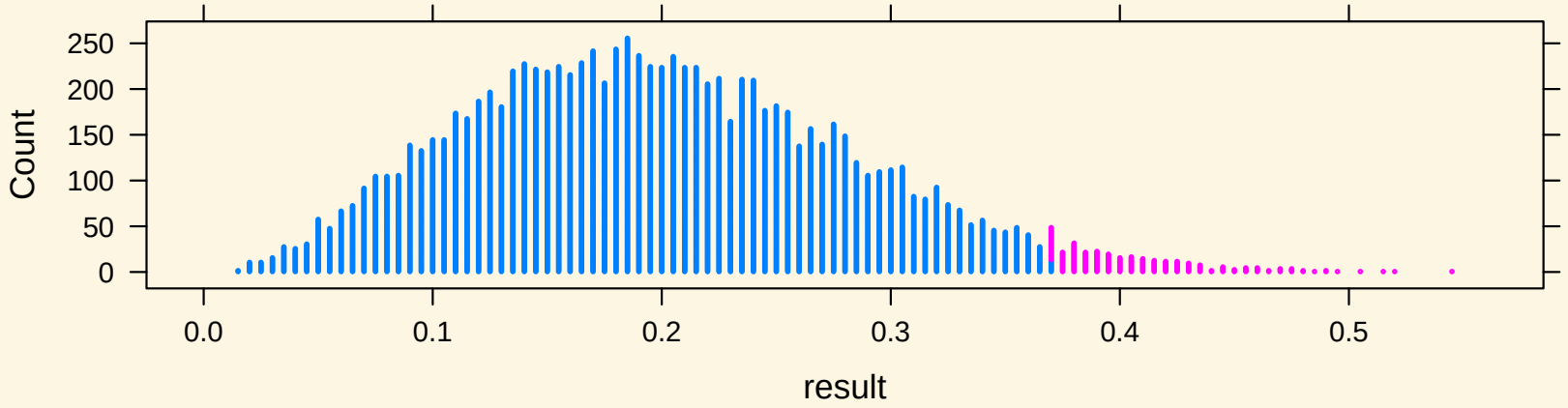
where $\bar{x}_g$ = mean of group g, $\bar{\bar{x}}$ = mean of means, and G = number of groups

Possible Randomization Test: Std. Dev. of Means

```
## Construct the randomization distribution
Random.sd.of.means <- do(10000) *
    mean(BrainChange ~ shuffle(Treatment), data = Brain) %>% sd()
## Compute the observed variance of means
obs.sd.of.means <-
    mean(BrainChange ~ Treatment, data = Brain) %>% sd()
```

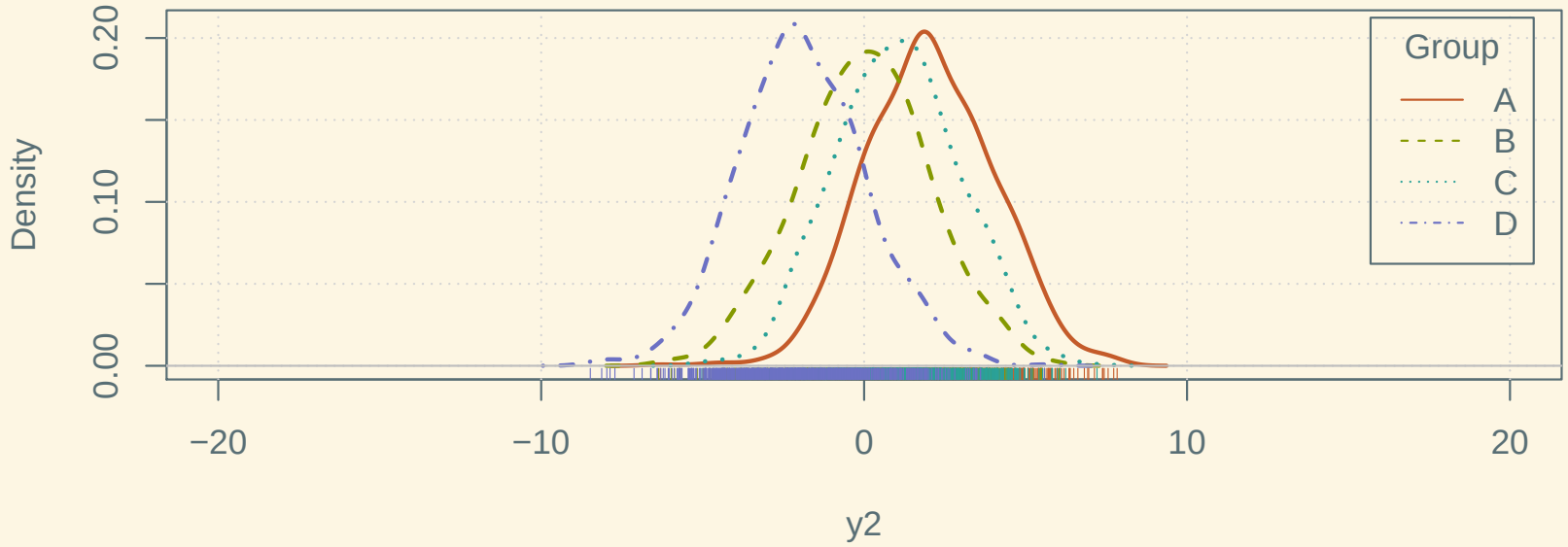
Possible Randomization Test: Std. Dev. of Means

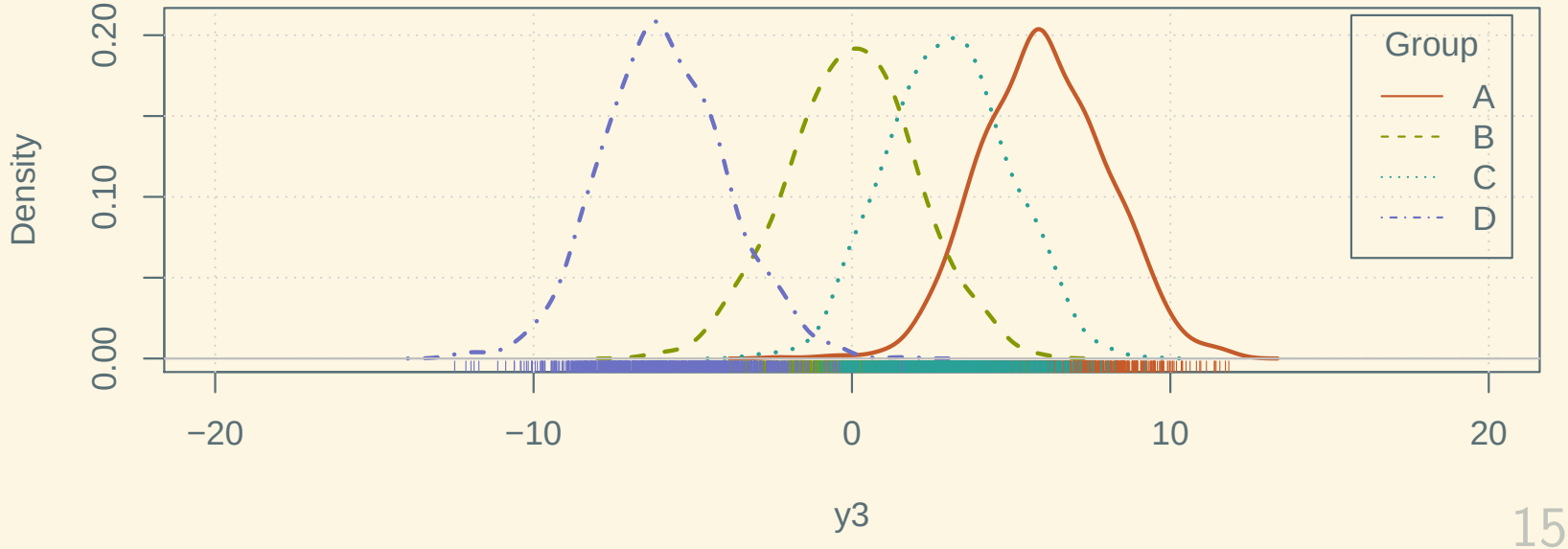
```
dotPlot(~result, data = Random.sd.of.means, width = 0.005, cex = 5,
        groups = (result >= obs.sd.of.means))
```



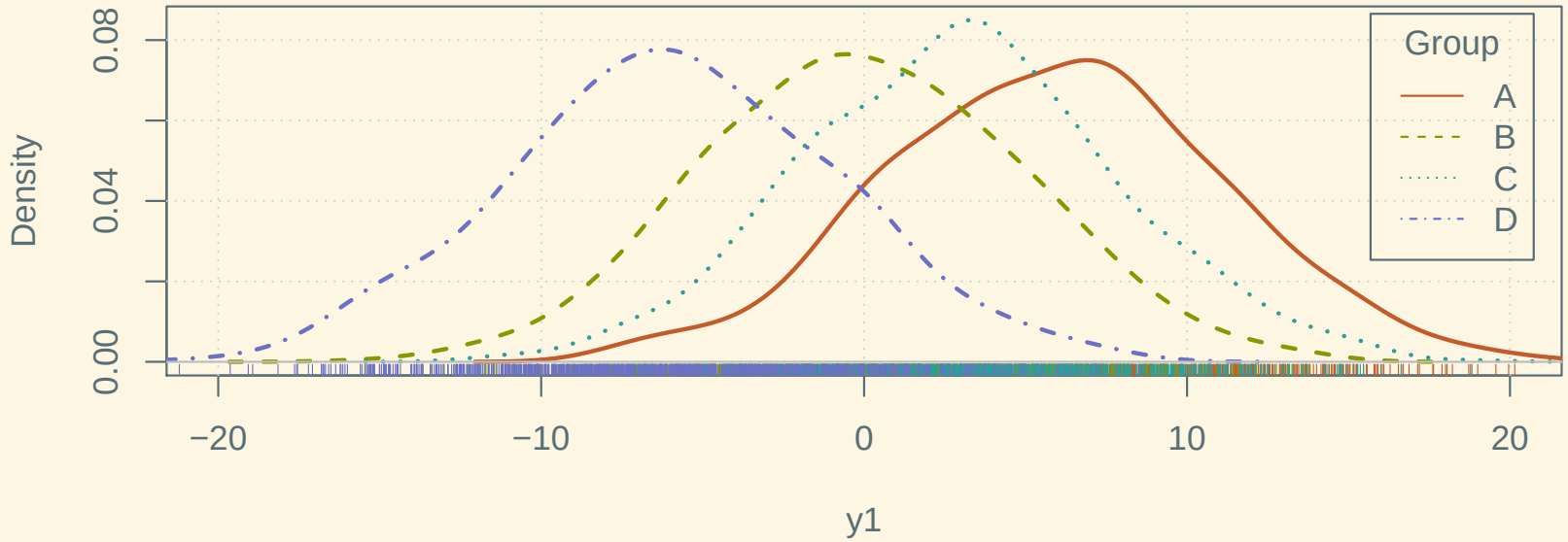

```
### P-value
prop(~(result >= obs.sd.of.means), data = Random.sd.of.means)

   TRUE
 0.0299
```

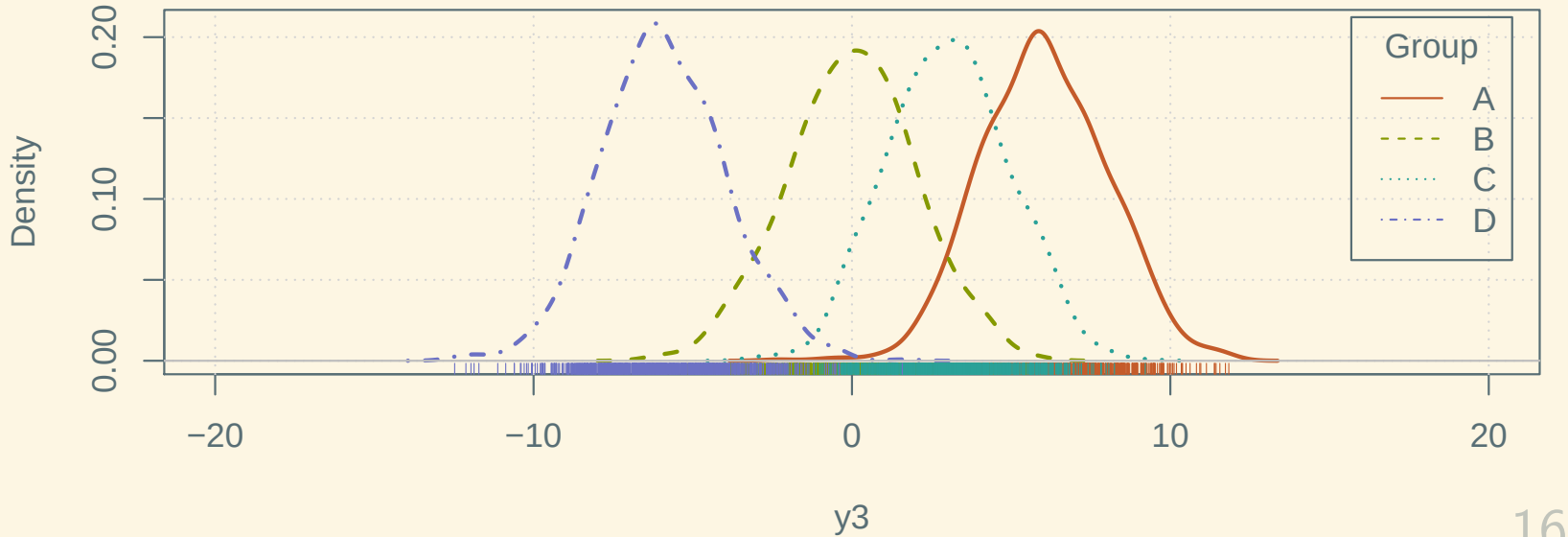
Which set of groups seem more distinct?





Which set of groups seem more distinct?





Within Groups Vs. Between Groups Variability

- Not only the differences of the sample means, but also the variation within groups seems to matter.
- Intuitively, if response values tend to differ more *between* groups than they do *within* groups, that points to a larger deviation from H_0.
- Idea: Compare variance (**between groups variance**) to variance within groups (**within groups variance**)

The Analysis of Variance (ANOVA)

- Conceptually, a standardized measure of dispersion of means is $(\sigma^2_{between}/\sigma^2_{within})$, the ratio of the between-groups variability to the within-groups variability.
- The F-statistic is based on this ratio:

$$F = \frac{\sum_{g=1}^{G} n_g(\bar{y}_g - \bar{\bar{y}})^2/(G-1)}{\sum_{g=1}^{G}\sum_{i=1}^{n_g}(y_{g,i} - \bar{y}_g)^2/(N-G)}$$

  where g indexes groups (of G), i indexes observations within groups, and

n_g and N	are the sample sizes in group g and overall
$\bar{y}_g$ and $\bar{\bar{y}}$	are the means in group g and overall
$y_{g,i}$	is the ith response in group g

- When H_0 is true, this has an F-**distribution**, with $G-1$ "between groups" df and $N-G$ "within groups" df, for $N-1$ df total.

The Analysis of Variance (ANOVA) Table

Component	Symbol	Computation
Sum of Squares (SS) "Between"	SS_{Between}	$\sum_{g=1}^{G} n_g(\bar{y}_g - \bar{\bar{y}})^2$
Sum of Squares (SS) "Within"	SS_{Within}	$\sum_{g=1}^{G}\sum_{i=1}^{n_g}(y_{g,i} - \bar{y}_g)^2$
Sum of Squares (SS) "Total"	SS_{Total}	$\sum_{g,i}(y_{g,i} - \bar{\bar{y}})^2$
Degrees of Freedom (df) "Between"	df_{Between}	$G - 1$
Degrees of Freedom (df) "Within"	df_{Within}	$N - G$
Degrees of Freedom (df) "Total"	df_{Total}	$N - 1$
Mean Square (MS) "Between"	MS_{Between}	$SS_{\text{Between}}/df_{\text{Between}}$
Mean Square (MS) "Withinn"	MS_{Within}	$SS_{\text{Within}}/df_{\text{Within}}$
F-statistic	F	$MS_{\text{Between}}/MS_{\text{Within}}$

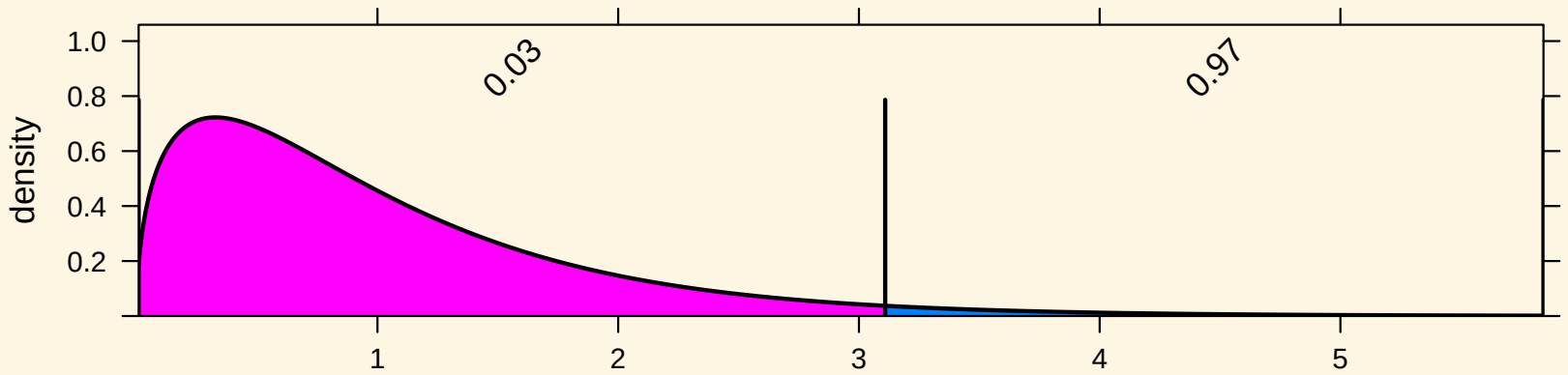
You won't need to compute the SS pieces by hand; just have a sense of what they're doing

The ANOVA Table: Exercise and Brain Size Change

```
aov(BrainChange ~ Treatment, data = Brain) %>% summary()
```

```
                Df Sum Sq Mean Sq F value Pr(>F)
    Treatment    3  10.83   3.609   3.109 0.0297 *
    Residuals  103 119.56   1.161
    ---
    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
xpf(3.1091, df1 = 3, df2 = 103, lower.tail = FALSE)
```




```
    [1] 0.02966933
```

Conditions for (Analytic) F-test

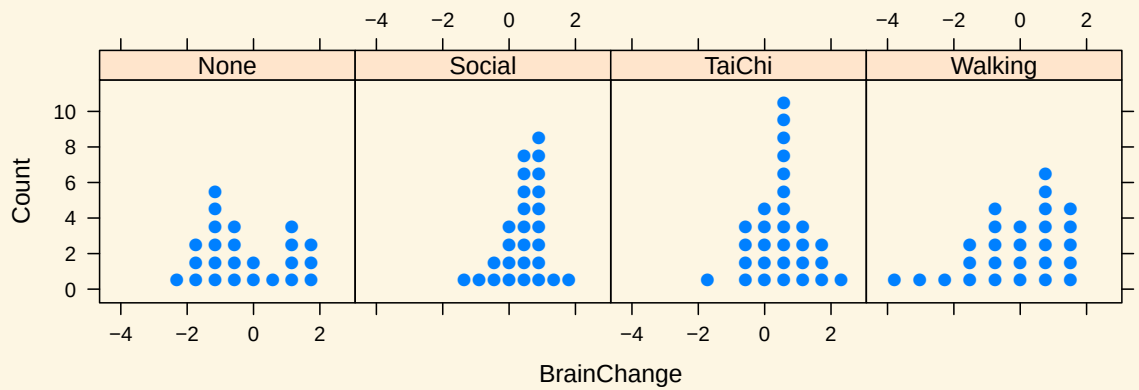
1. Within groups, Normally distributed responses
2. Similar standard deviations in each group

In practice, reasonably symmetric distributions with reasonably similar standard deviations works OK (largest / smallest ≤ 2)

```
sd(BrainChange ~ Treatment, data = Brain)
```

```
      None    Social    TaiChi   Walking
 1.2584309 0.6968969 0.8557466 1.3868388
```

```
dotPlot(~BrainChange | Treatment, data = Brain)
```



Example: Sandwich Ants

Adapted from Lock Ex. 8.22

Some students did an experiment asking how different sandwich fillings might affect the mean number of ants attracted to pieces of a sandwich. The students running this experiment also varied the type of bread for the sandwiches, randomizing between four types: Multigrain, Rye, Wholemeal, and White. The ant counts in 6 trials and summary statistics for each type of bread and the 24 trials as a whole are given below.

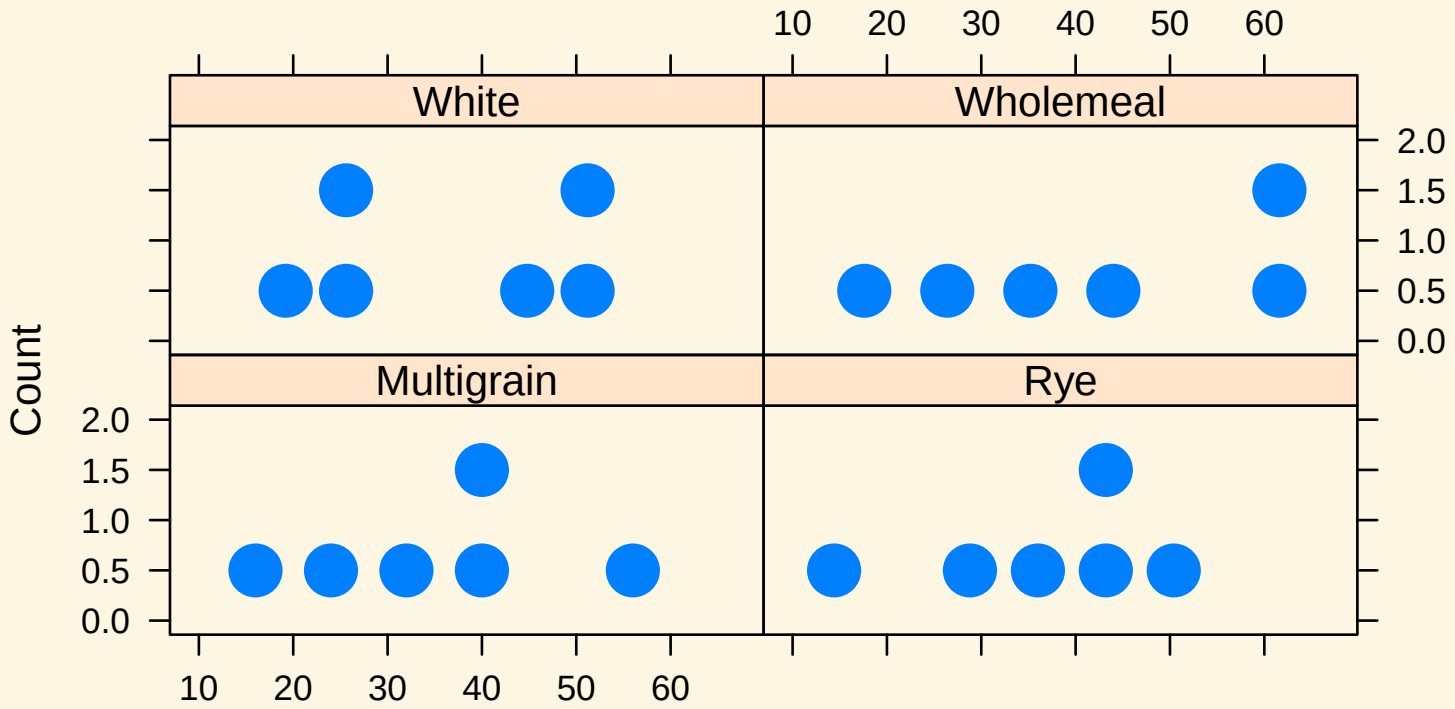
Bread	Ants						Mean	SD
Multi	42	22	36	38	19	59	36.00	14.52
Rye	18	43	44	31	36	54	37.67	12.40
Whole	29	59	34	21	47	65	35.83	13.86
White	42	25	49	25	21	53	42.50	17.41
					Total		38.00	13.95

Example: Sandwich Ants

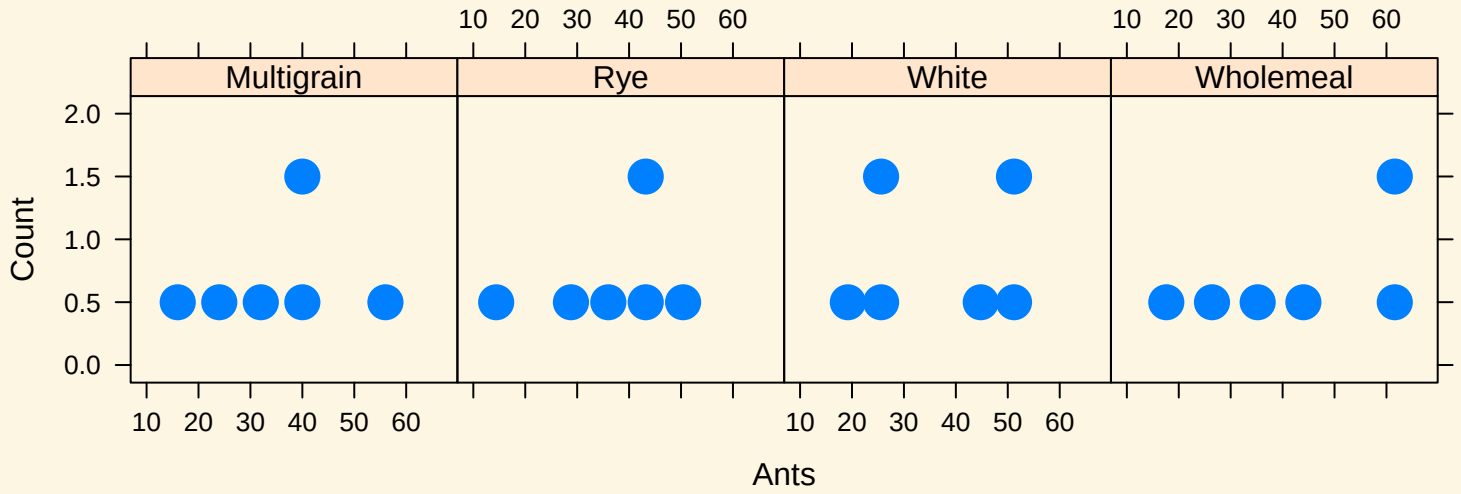
$$H_0 : \mu_{Multi} = \mu_{Rye} = \mu_{Whole} = \mu_{White}$$

$$H_1 : \text{not } H_0$$

```
library("Lock5Data"); data("SandwichAnts")
dotPlot(~Ants | Bread, data = SandwichAnts, cex = 0.75)
```



Are the Conditions for Analytic Inference Satisfied?



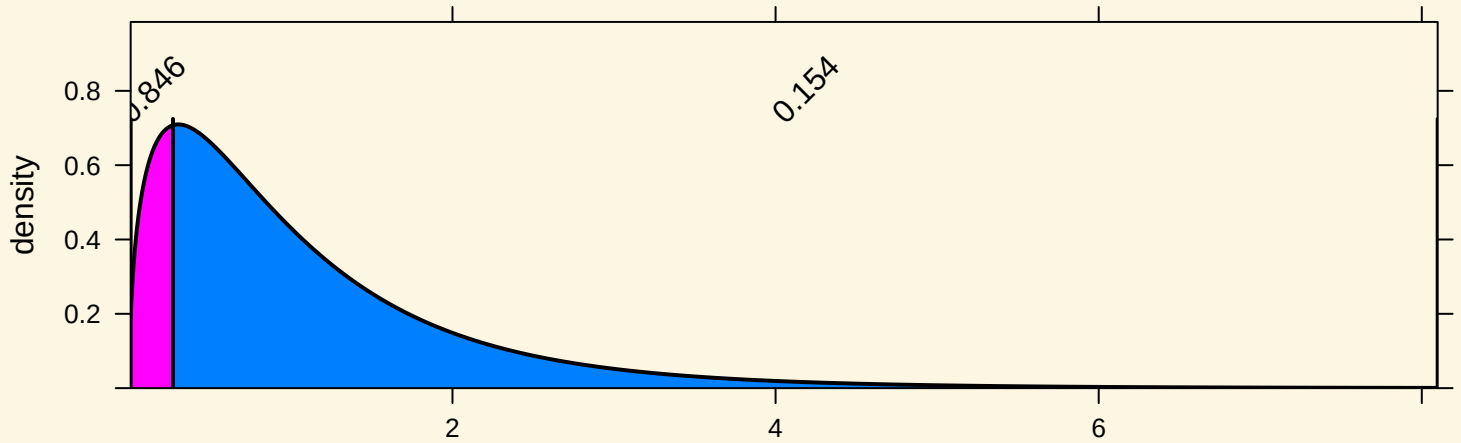
Bread	Ants						Mean	SD
Multi	42	22	36	38	19	59	36.00	14.52
Rye	18	43	44	31	36	54	37.67	12.40
Whole	29	59	34	21	47	65	35.83	13.86
White	42	25	49	25	21	53	42.50	17.41
					Total		38.00	13.95

Example: Sandwich Ants

```
aov(Ants ~ Bread, data = SandwichAnts) %>% summary()
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
Bread        3    174   58.11    0.27  0.846
Residuals   20   4300  214.98
```

```
xpf(0.27, df1 = 3, df2 = 20, lower.tail = FALSE)
```




```
[1] 0.8462571
```

Example: Stereotype Threat and Student Athletes

The term "stereotype threat" refers to a phenomenon whereby reminders of particular components of an individual's identity (race, gender, ethnicity) can result in the individual conforming to stereotypes about that group. For example, women perform worse on a math test after being reminded of their gender (Spencer et al., 1999). Some researchers (Steele, 1997) believe this is due to anxiety about the possibility of confirming negative stereotypes. Yopyk and Prentice (2005) administered a math test to student-athletes after either (A) reminding them of their athlete status, (B) reminding them of their student status, or (C) not reminding them of either component of their identity. The test scores had the following mean and standard deviations.

	Athlete Prime	No Prime	Student Prime
n	12	13	12
$\bar{x}$	66.97	82.46	86.17
s	5.60	4.99	4.58

Example: Stereotype Threat and Student Athletes

	Athlete Prime	No Prime	Student Prime
n	12	13	12
$\bar{x}$	66.97	82.46	86.17
s	5.60	4.99	4.58

Pairs: Fill in the ANOVA table (the hard part is done). What is your conclusion?

Source	df	SS	MS	F	P-value
Prime		2504.38			
Residuals		874.5			

Solutions

	Athlete Prime	No Prime	Student Prime
n	12	13	12
$\bar{x}$	66.97	82.46	86.17
s	5.60	4.99	4.58

Source	df	SS	MS	F	P-value
Prime	2	2504.38	1252.19	48.68	1.05e-10
Residuals	34	874.5	25.72		

Conclusion: Some group mean is different from some other group mean.

Inferences After ANOVA

- If we find evidence that the means are not equal, we will want to ask which ones differ.
- Why not do this from the start?
- Doing F-test first keeps our overall chance of Type I Error at 5% (or whatever α is), provided we *stop if it's not significant*.

Inference After ANOVA

Following a Significant F-test...

1. CIs for individual means (estimate each μ_g)
2. CIs for pairwise differences in means (estimate $\mu_A - \mu_B$ for each A, B pair)
3. t-tests for pairwise differences (test whether $\mu_A - \mu_B = 0$ for each pair)

In general...

Do these as we normally would, but use the "pooled within groups variance", estimated by MS_{Within}, in place of s_A, s_B, etc.

Examining Individual Groups

- Since we assume σ^2_{Within} is the same across groups, the standard error of any individual mean is:

$$SE_{\bar{y}_A} = \sqrt{\frac{\sigma^2_{\text{Within}}}{n_A}}$$

- Estimate with

$$\widehat{SE}_{\bar{y}_A} = \sqrt{\frac{MS_{\text{Within}}}{n_A}}$$

Confidence Interval for a Single Group Mean

$$\bar{y}_A \pm t^*_{df_{\text{Within}}} \cdot \sqrt{\frac{MS_{\text{Within}}}{n_A}}$$

Pairwise Comparisons

CI and Test Statistic for Pairwise Difference

A CI or test of a difference is based on a t-test with degrees of freedom df_{Within} (the number of pieces of information available to estimate σ^2_{Within})

- CI: $\bar{y}_A - \bar{y}_B \pm t^*_{df_{\text{Within}}} \cdot \sqrt{MS_{\text{Within}}(\frac{1}{n_A} + \frac{1}{n_B})}$
- Test statistic: $t_{obs} = \frac{\bar{y}_A - \bar{y}_B - 0}{\sqrt{MS_{\text{Within}}(\frac{1}{n_A} + \frac{1}{n_B})}}$
  - t_{obs} has a t-distribution with $df_{\text{Within}} = N - G$ degrees of freedom.

The ANOVA Table: Stereotype Threat

Source	df	SS	MS	F	P-value
Prime	2	2504.38	1252.19	48.68	1.05e-10
Residuals	34	874.5	25.72		

Let's compute confidence intervals for the means and do tests for differences of pairs.