STAT 113
Comparing Multiple Means

Colin Reimer Dawson

Oberlin College

January 14, 2022

Outline

Comparing Multiple Means

A Randomization Test

An Analytic Approach

Effect Size

Statistics by Variable Type

Explanatory	Response	Statistic
None	Binary	Single Proportion
None	Quantitative	Single Mean
None	Categorical	χ^2 (Goodness of Fit)
Binary	Binary	Difference of Proportions
Binary	Quantitative	Difference of Means
Quantitative	Quantitative	Correlation or Slope
Categorical	Categorical	χ^2 (Association)
Categorical	Quantitative	??

Outline

Comparing Multiple Means

A Randomization Test

An Analytic Approach

Effect Size

Exercise and Changes in Brain Size

- Researchers in China 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 one of 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

Variables: The **response variable** (% change in brain size) is quantitative, and the **explanatory variable** (activity group) is categorical (w/ 4 levels).

Parameters: Some natural parameters to focus on are the **typical responses in each group** (e.g., **mean % increase** in brain size).

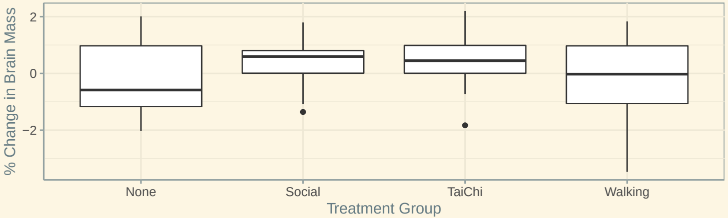
Hypotheses: If there is no association between group and response, **the population means by group** (of the % increase variable) **would have to be equal**:

$$H_0 : \mu_{\text{TaiChi}} = \mu_{\text{Walking}} = \mu_{\text{Social}} = \mu_{\text{Nothing}}$$

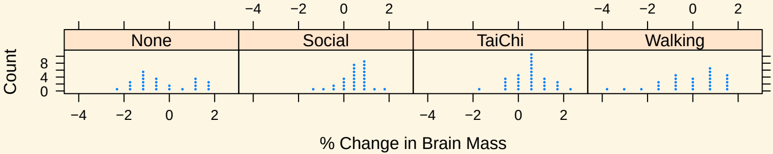
H_1 : At least one μ differs from at least one other

The Data

```
gf_boxplot(BrainChange ~ Treatment, data = Brain,
          xlab = "Treatment Group", ylab = "% Change in Brain Mass")
```




```
dotPlot(~BrainChange | Treatment, data = Brain, layout = c(4,1),
      xlab = "% Change in Brain Mass")
```



Descriptive Stats

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

    Treatment  min      Q1 median   Q3 max  mean   sd  n missing
1        None -2.0 -1.1687 -0.585 0.97 2.0 -0.24 1.26 24       0
2      Social -1.4  0.0075  0.596 0.81 1.8  0.41 0.70 27       0
3      TaiChi -1.8  0.0050  0.449 0.99 2.2  0.47 0.86 29       0
4     Walking -3.5 -1.0585 -0.026 0.97 1.8 -0.15 1.39 27       0
```

- The **Social Interaction** and **Tai Chi** groups showed an **increase in average brain mass** from start to end.
- The **Control** and **Walking** groups saw a **decrease**.
- But can this reasonably be attributable to chance?

Outline

Comparing Multiple Means

A Randomization Test

An Analytic Approach

Effect Size

A Randomization Test

- We use the same basic randomization procedure whenever our null hypothesis is that **two variables are not associated**.
- Randomize by **randomly pairing** responses and group assignments
- In other words, **randomly re-group** the data.

Possible Test Statistics

- How to measure **how far the data is from what the "skeptic" expects to see on average** (i.e., if H_0 is accurate)?
- Some possibilities:
  - Range of means: $\bar{x}_{\text{largest}} - \bar{x}_{\text{smallest}}$
  - Average pairwise absolute difference:

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

  - Standard deviation of sample means

Possible Test Statistic: Std. Dev. of Means

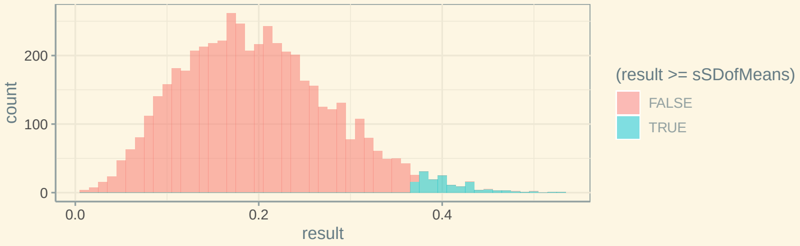
```
## Compute the observed SD of means
sSDofMeans <-
    ## calculate the four means
    mean(BrainChange ~ Treatment, data = Brain) %>%
    ## take the sd() of the set of four group means
    sd()
sSDofMeans

     [1] 0.37
```

Possible Test Statistic: Std. Dev. of Means

```r
set.seed(42)
## Construct the randomization distribution
RandomizationDistribution <- do(5000) *
    mean(BrainChange ~ shuffle(Treatment), data = Brain) %>%
    sd()
```

Possible Randomization Test: Std. Dev. of Means

```
gf_histogram(~result, data = RandomizationDistribution,
        binwidth = 0.01, fill = ~(result >= sSDofMeans))
```




```
### P-value
prop(~(result >= sSDofMeans), data = RandomizationDistribution)

     prop_TRUE
          0.03
```

Outline

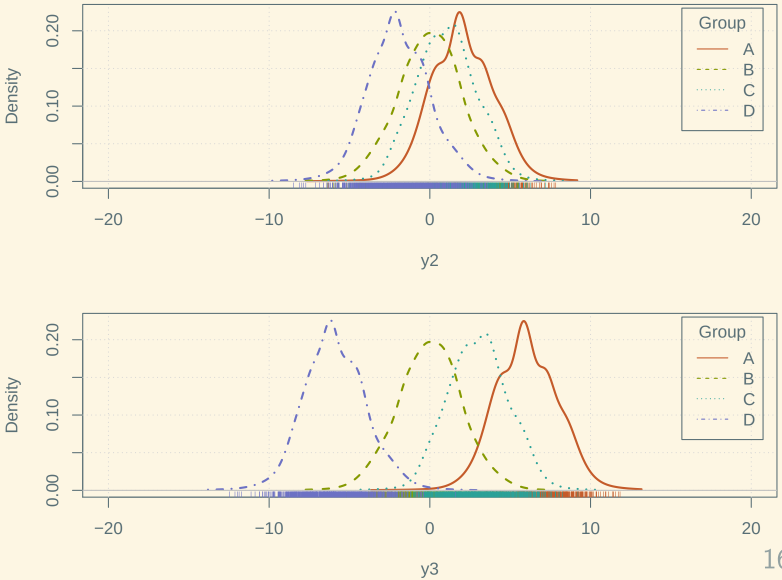
Comparing Multiple Means

A Randomization Test

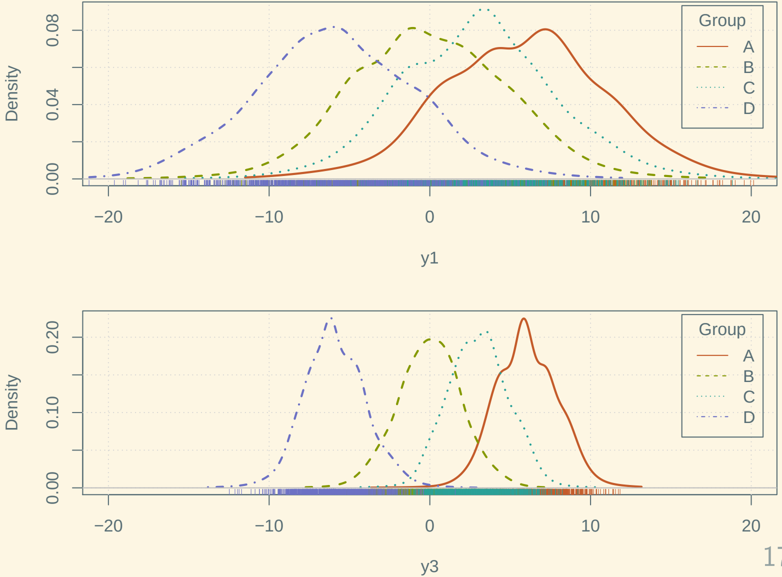
An Analytic Approach

Effect Size

Which set of groups seem more distinct?



Which set of groups seem more distinct?



Within Groups Vs. Between Groups Variability

- Not only the differences **among the sample means**, but also the variation **within groups** seems to matter.

Within Groups Vs. Between Groups Variability

- Not only the differences **among the sample means**, but also the variation **within groups** seems to matter.
- The more the response values differ *between* groups **relative to the natural within group variation**, the less likely that is to happen by chance.

Within Groups Vs. Between Groups Variability

- Not only the differences **among the sample means**, but also the variation **within groups** seems to matter.
- The more the response values differ *between* groups **relative to the natural within group variation**, the less likely that is to happen by chance.
- Idea: Compare variation **between groups** to variation **within groups**

The F statistic and the Analysis of Variance (ANOVA)

- To test for differences among **means**, we analyze different aspects of **variability**: between groups vs within groups.

The F statistic and the Analysis of Variance (ANOVA)

- To test for differences among **means**, we analyze different aspects of **variability**: between groups vs within groups.
- This is called the **Analysis of Variance** (ANOVA)

The F statistic and the Analysis of Variance (ANOVA)

- To test for differences among **means**, we analyze different aspects of **variability**: between groups vs within groups.
- This is called the **Analysis of Variance** (ANOVA)
- A **standardized** measure of variability among means is $(\sigma^2_{\text{between}}/\sigma^2_{\text{within}})$, the **ratio** of the **between-means variance** to the **within-group variance**.

The F statistic and the Analysis of Variance (ANOVA)

- To test for differences among **means**, we analyze different aspects of **variability**: between groups vs within groups.
- This is called the **Analysis of Variance** (ANOVA)
- A **standardized** measure of variability among means is $(\sigma^2_{\text{between}}/\sigma^2_{\text{within}})$, the **ratio** of the **between-means variance** to the **within-group variance**.
- The F-**statistic** (named for Ronald Fisher; remember him?) is (sort of) an estimate of this ratio.

The F statistic and the Analysis of Variance (ANOVA)

- To test for differences among **means**, we analyze different aspects of **variability**: between groups vs within groups.
- This is called the **Analysis of Variance** (ANOVA)
- A **standardized** measure of variability among means is $(\sigma^2_{\text{between}}/\sigma^2_{\text{within}})$, the **ratio** of the **between-means variance** to the **within-group variance**.
- The F-**statistic** (named for Ronald Fisher; remember him?) is (sort of) an estimate of this ratio.
- If there is no association at all, all groups have the **same distribution**, so there's only one σ^2_{within}

The F statistic and the Analysis of Variance (ANOVA)

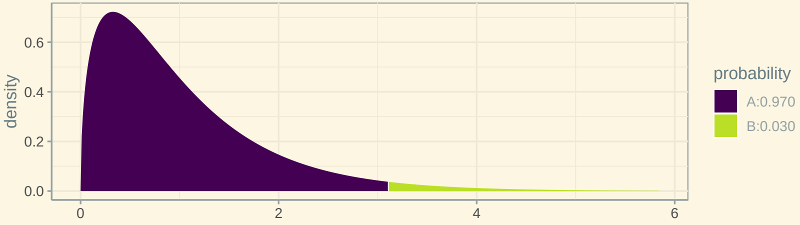
- To test for differences among **means**, we analyze different aspects of **variability**: between groups vs within groups.
- This is called the **Analysis of Variance** (ANOVA)
- A **standardized** measure of variability among means is $(\sigma^2_{\text{between}}/\sigma^2_{\text{within}})$, the **ratio** of the **between-means variance** to the **within-group variance**.
- The F**-statistic** (named for Ronald Fisher; remember him?) is (sort of) an estimate of this ratio.
- If there is no association at all, all groups have the **same distribution**, so there's only one σ^2_{within}
- Aside: Groups could have equal means but different variability; but this test isn't set up to look for that.

Properties of the F statistic

- Like χ^2, the F statistic cannot be negative, and **larger values constitute bigger discrepancies** from H_0.
- Thus (also like χ^2) all tests are **"right-tailed"**, despite H_1 being **non-directional**

Analytic Inference: Exercise and Brain Size Change

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

                Df Sum Sq Mean Sq F value Pr(>F)
    Treatment     3   10.8    3.61    3.11   0.03 *
    Residuals   103  119.6    1.16
    ---
    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

pdist("f", q = 3.11, df1 = 3, df2 = 103, lower.tail = FALSE)
```




```
    [1] 0.03
```

Exercise and Brain Change: Conclusion

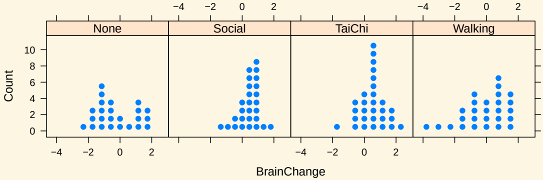
There is statistically significant evidence ($F = 3.11$, $p = 0.03$) that at least some of the treatments in this study have an impact on the decline in brain mass for the population of older adults.

Conditions for (Analytic) F-test

In theory, assumes:

1. **Normally distributed responses** *within* groups
2. **Same population standard deviation** for *each* group

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

       None  Social  TaiChi Walking
       1.26    0.70    0.86    1.39

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

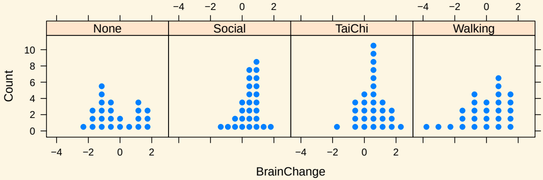


Conditions for (Analytic) F-test

In practice, look for:

1. **Reasonably symmetric** within-group distributions
2. A **ratio of 2 or less** between the largest and smallest standard deviation

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

       None  Social  TaiChi Walking
       1.26    0.70    0.86    1.39

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



Sandwich Ants: Adapted from Lock Ex. 8.22

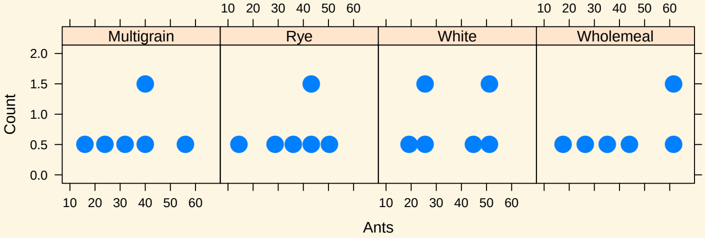
A group of intro stats students did an experiment asking how different types of sandwich bread affect the mean number of ants attracted to pieces of a sandwich.
The students placed sandwiches with either Multigrain, Rye, Wholemeal, or White bread on the ground in randomized order, and counted how many ants crawled on each sandwich.
The ant counts for 6 sandwiches of each type are given below.

Bread	Ants						Mean ($\bar{x}$)	SD (s)
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
					Overall		38.00	13.95

Sandwich Ants: Hypotheses and Plots

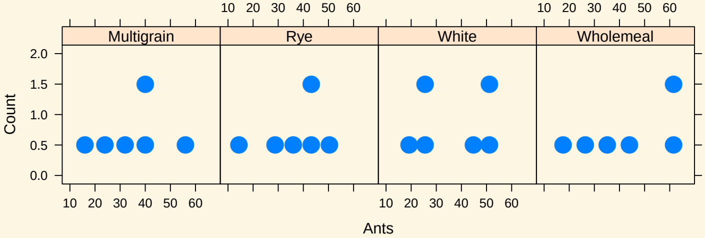
$$H_0 : \mu_{\text{Multi}} = \mu_{\text{Rye}} = \mu_{\text{Whole}} = \mu_{\text{White}}$$
$$H_1 : \text{not } H_0$$

```
library(Lock5Data); data(SandwichAnts)
dotPlot(~Ants | Bread, data = SandwichAnts, cex = 0.4)
```



Sandwich Ants: Conditions

Checking Symmetry Within Groups:

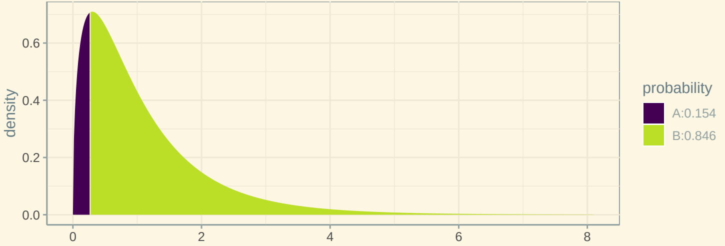


Checking Standard Deviations Within Groups:

```
sd(Ants ~ Bread, data = SandwichAnts)

    Multigrain        Rye      White  Wholemeal
          14.5       12.4       13.9       17.4
```

Sandwich Ants: Test Statistic and P-value

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

               Df Sum Sq Mean Sq F value Pr(>F)
    Bread       3    174    58.1    0.27   0.85
    Residuals  20   4300   215.0

pdist("f", q = 0.27, df1 = 3, df2 = 20, lower.tail = FALSE)
```




```
    [1] 0.846
```

Conclusion: Sandwich Ants

There was not statistically significant evidence that ants prefer any type of bread over any other type of bread ($F = 0.27$, $p = 0.85$).

Outline

Comparing Multiple Means

A Randomization Test

An Analytic Approach

Effect Size

Effect Size for ANOVA

- The F-statistic and P-value are measures of how surprising the pattern of means would be if all differences were due to chance.

Effect Size for ANOVA

- The F-statistic and P-value are measures of how surprising the pattern of means would be if all differences were due to chance.
- But as always, with enough data, any difference is distinguishable from chance variation.

Effect Size for ANOVA

- The F-statistic and P-value are measures of how surprising the pattern of means would be if all differences were due to chance.
- But as always, with enough data, any difference is distinguishable from chance variation.
- We can quantify the magnitude of the differences on a standardized scale with

$$R^2 = 1 - \frac{SS_{Within}}{SS_{Between} + SS_{Within}}$$

Effect Size for ANOVA

- The F-statistic and P-value are measures of how surprising the pattern of means would be if all differences were due to chance.
- But as always, with enough data, any difference is distinguishable from chance variation.
- We can quantify the magnitude of the differences on a standardized scale with

$$R^2 = 1 - \frac{SS_{Within}}{SS_{Between} + SS_{Within}}$$

- Same concept as in regression: what proportion of total variability is predictable if we know the groups?

Effect Size: Brain Change

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

                 Df Sum Sq Mean Sq F value Pr(>F)
     Treatment     3   10.8    3.61    3.11   0.03 *
     Residuals   103  119.6    1.16
     ---
     Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

lm(BrainChange ~ Treatment, data = Brain) %>% rsquared()

     [1] 0.083
```

There is significant evidence that differences among treatment groups are not due to chance ($F = 3.11$, $P = 0.03$). However, only 8.3% of the variability across individuals in changes in brain size during the study period is attributable to differences in treatments ($R^2 = 0.083$).