Inference for Many Means

Nate Wells

Math 141, 4/23/21

Outline

In this lecture, we will...

- Construct a statistic to measure the differences in mean among several groups
- Discuss the theoretical and simulation-based distribution of the F statistic
- Use ANOVA to test for a difference in means among several groups

Section 1

Differences Among Several Populations

Comparing a Quantatiative and Categorical Variable

- Suppose we want to compare the values of a common quantitative variable across several different populations.

Comparing a Quantatiative and Categorical Variable

- Suppose we want to compare the values of a common quantitative variable across several different populations.
    - That is, we wish to determine whether a quantitative and categorical variable are independent

Comparing a Quantatiative and Categorical Variable

- Suppose we want to compare the values of a common quantitative variable across several different populations.
    - That is, we wish to determine whether a quantitative and categorical variable are independent
- Previously, we. . .

Comparing a Quantatiative and Categorical Variable

- Suppose we want to compare the values of a common quantitative variable across several different populations.
  - That is, we wish to determine whether a quantitative and categorical variable are independent
- Previously, we. . .
  - Used the chi-square test to determine whether two categorical variables were independent

Comparing a Quantatiative and Categorical Variable

- Suppose we want to compare the values of a common quantitative variable across several different populations.
  - That is, we wish to determine whether a quantitative and categorical variable are independent
- Previously, we. . .
  - Used the chi-square test to determine whether two categorical variables were independent
  - Determine whether a quantitative variable and categorical variable (with only 2 levels) were independent

Comparing a Quantatiative and Categorical Variable

- Suppose we want to compare the values of a common quantitative variable across several different populations.
  - That is, we wish to determine whether a quantitative and categorical variable are independent
- Previously, we. . .
  - Used the chi-square test to determine whether two categorical variables were independent
  - Determine whether a quantitative variable and categorical variable (with only 2 levels) were independent
- The Analysis of Variance (ANOVA) test will allow us to assess whether the mean values of a quantitative variable differ across the levels of a categorical variable.

There's No Accounting For Taste

Research Question: Certainly, individual tastes in movie genres vary. But in aggregate, do audience ratings of movies depend on genre? To answer, we assess the Rotten Tomatoes audience rating for 132 films from 2011 spread across 7 genress.

There's No Accounting For Taste

Research Question: Certainly, individual tastes in movie genres vary. But in aggregate, do audience ratings of movies depend on genre? To answer, we assess the Rotten Tomatoes audience rating for 132 films from 2011 spread across 7 genress.

Movie	AudienceScore	Genre
Insidious	65	Horror
Paranormal Activity 3	58	Horror
Bad Teacher	38	Comedy
Bridesmaids	77	Comedy
Midnight in Paris	84	Romance
The Help	91	Drama

There's No Accounting For Taste

Research Question: Certainly, individual tastes in movie genres vary. But in aggregate, do audience ratings of movies depend on genre? To answer, we assess the Rotten Tomatoes audience rating for 132 films from 2011 spread across 7 genress.

Movie	AudienceScore	Genre
Insidious	65	Horror
Paranormal Activity 3	58	Horror
Bad Teacher	38	Comedy
Bridesmaids	77	Comedy
Midnight in Paris	84	Romance
The Help	91	Drama

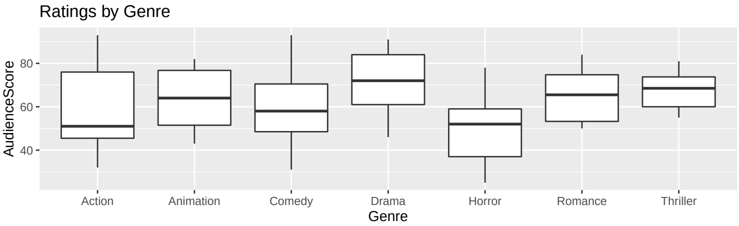
- **Observational unit**: a single film
- **Sample**: 132 films from 2011
- **Population**: All films (maybe from last 20 years?)
- **Variables**: Audience Rating and Genre
- **Parameters**: Average audience rating for each genre, $\mu_1, \ldots, \mu_7$.
- **Null Hypothesis**: $H_0 : \mu_1 = \mu_2 = \cdots = \mu_7$
- **Alternative Hypothesis**: At least one μ is not equal to the others

Data Exploration

Do ratings differ by genre?

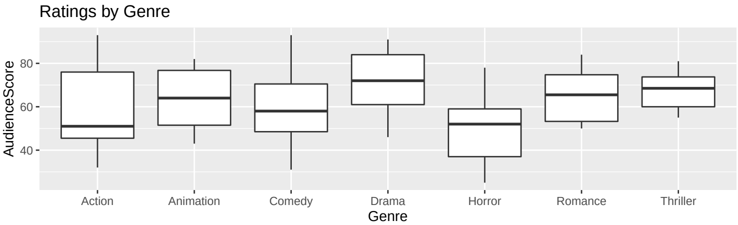
Data Exploration

Do ratings differ by genre?



Data Exploration

Do ratings differ by genre?




```
movies %>% group_by(Genre) %>%
  summarize(number = n(), avg_rating = mean(AudienceScore), st_dev = sd(AudienceScore))
```

```
## # A tibble: 7 x 4
##   Genre     number avg_rating st_dev
##   <fct>      <int>      <dbl>  <dbl>
## 1 Action        32       58.6  18.4
## 2 Animation     12       64.1  13.9
## 3 Comedy        27       59.1  15.7
## 4 Drama         21       72.1  14.5
## 5 Horror        17       48.6  15.9
## 6 Romance       10       64.8  12.9
## 7 Thriller      12       67.7   9.01
```

```
##       Movie number avg_rating st_dev
## 1 All Films    131         61     17
```

Independence?

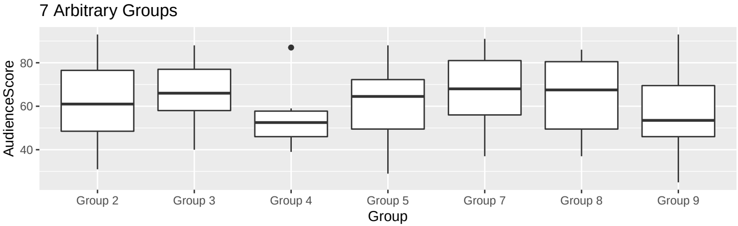
- We saw a clear visual difference in mean scores for different genres.

Independence?

- We saw a clear visual difference in mean scores for different genres.
    - But maybe we would see a similar difference just by separating into 7 arbitrary groups

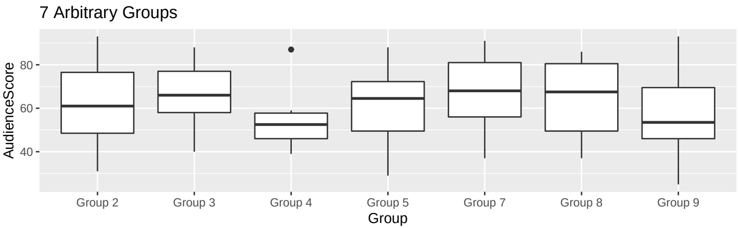
Independence?

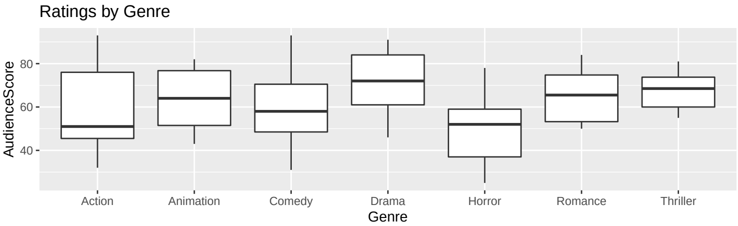
- We saw a clear visual difference in mean scores for different genres.
  - But maybe we would see a similar difference just by separating into 7 arbitrary groups



Independence?

- We saw a clear visual difference in mean scores for different genres.
  - But maybe we would see a similar difference just by separating into 7 arbitrary groups



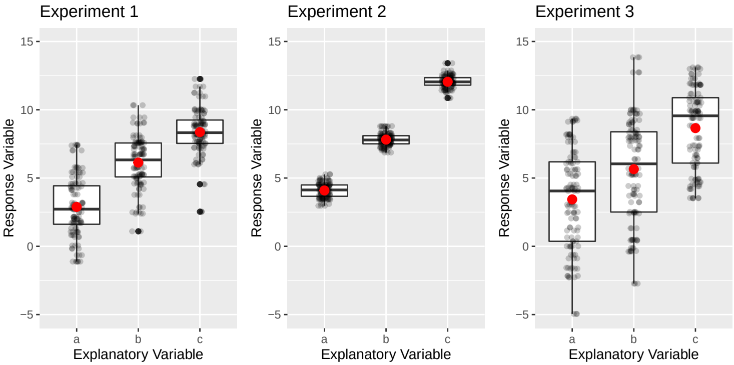


There's No Accounting for Taste. . . But There is Accounting for Variance

Which of the following experiments gives *strongest* evidence of a difference in population means? Which gives the *weakest* evidence?

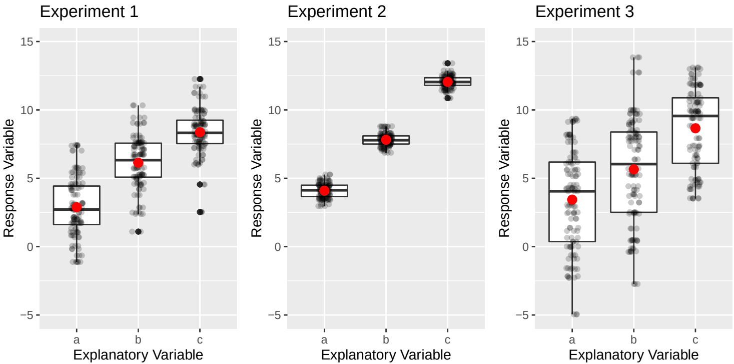
There's No Accounting for Taste. . . But There is Accounting for Variance

Which of the following experiments gives *strongest* evidence of a difference in population means? Which gives the *weakest* evidence?



There's No Accounting for Taste. . . But There is Accounting for Variance

Which of the following experiments gives *strongest* evidence of a difference in population means? Which gives the *weakest* evidence?



- **Strongest**: Experiment 2
- **Weakest**: Experiment 3

Constructing a Statistic

- To assess whether a collection of population means are equal, we treat the sample means as a data set.

Constructing a Statistic

- To assess whether a collection of population means are equal, we treat the sample means as a data set.
  - If the population means are equal, the sample means should be close.
  - Otherwise, the sample means should be spread out.

Constructing a Statistic

- To assess whether a collection of population means are equal, we treat the sample means as a data set.
  - If the population means are equal, the sample means should be close.
  - Otherwise, the sample means should be spread out.
- But how spread out do sample means need to be to give good evidence that population means are different?

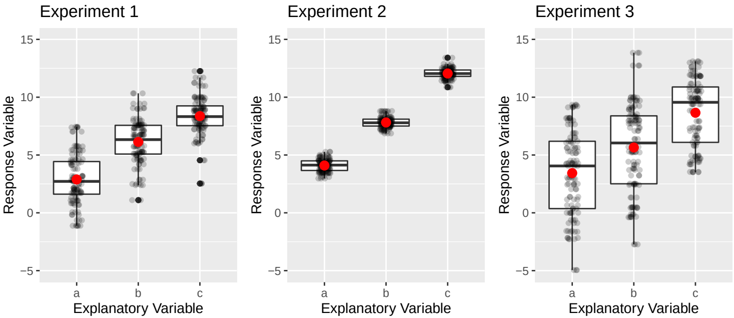
Constructing a Statistic

- To assess whether a collection of population means are equal, we treat the sample means as a data set.
  - If the population means are equal, the sample means should be close.
  - Otherwise, the sample means should be spread out.
- But how spread out do sample means need to be to give good evidence that population means are different?
  - If the populations have large standard deviation, we would expect the sample means to exhibit greater spread (even if the population means are equal)

Constructing a Statistic

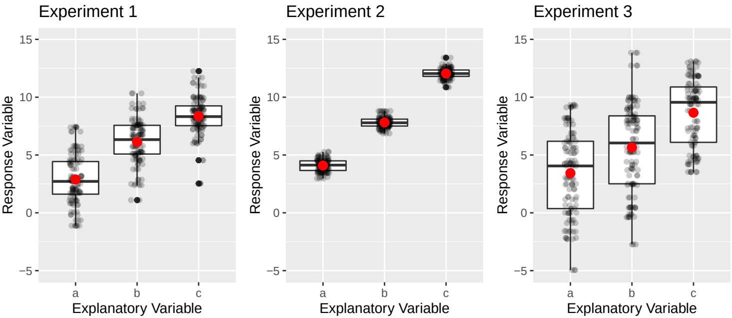
- To assess whether a collection of population means are equal, we treat the sample means as a data set.
  - If the population means are equal, the sample means should be close.
  - Otherwise, the sample means should be spread out.
- But how spread out do sample means need to be to give good evidence that population means are different?
  - If the populations have large standard deviation, we would expect the sample means to exhibit greater spread (even if the population means are equal)
- Is the variation observed among sample means greater than what can be explained by variability in observations within each group alone?

Partitioning Variability



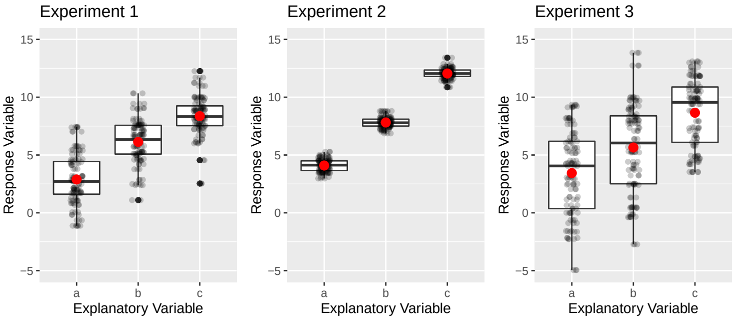
- The **Total Variability** among all observations is the sum of **Variability Between Groups** and **Variability Within Groups**

Partitioning Variability



- The **Total Variability** among all observations is the sum of **Variability Between Groups** and **Variability Within Groups**
- Variability Between Groups: How much do means vary?
  - Compare red dots

Partitioning Variability



- The **Total Variability** among all observations is the sum of **Variability Between Groups** and **Variability Within Groups**
- Variability Between Groups: How much do means vary?
  - Compare red dots
- Variability Within Groups: How much do observations in groups vary from mean?
  - Within each group, compare black dots to red dot

Partitioning Variability

- $\text{TotalVariability} = \text{ Variability Between Groups } + \text{ Variability Within Groups}$

Partitioning Variability

- $\text{TotalVariability} = \text{ Variability Between Groups } + \text{ Variability Within Groups}$
- Variability Between Groups: How much do means vary?

$$\begin{aligned}\text{Variability Between Groups } &= \sum n_i(\bar{x}_i - \bar{x})^2 \\ &= \text{Sum of Squares Group} \\ &= \text{SSG}\end{aligned}$$

Partitioning Variability

- $\text{TotalVariability} = \text{ Variability Between Groups } + \text{ Variability Within Groups}$
- Variability Between Groups: How much do means vary?

$$\begin{aligned} \text{Variability Between Groups } &= \sum n_i(\bar{x}_i - \bar{x})^2 \\ &= \text{Sum of Squares Group} \\ &= \text{SSG} \end{aligned}$$

- Variability Within Groups: How much do observations in groups vary from mean?

$$\begin{aligned} \text{Variability Within Groups } &= \sum (x_i - \bar{x}_i)^2 \\ &= \text{Sum of Squares Error} \\ &= \text{SSE} \end{aligned}$$

Partitioning Variability

- $\text{TotalVariability} = \text{ Variability Between Groups } + \text{ Variability Within Groups}$
- Variability Between Groups: How much do means vary?

$$\begin{aligned}\text{Variability Between Groups } &= \sum n_i(\bar{x}_i - \bar{x})^2\\ &= \text{Sum of Squares Group}\\ &= \text{SSG}\end{aligned}$$

- Variability Within Groups: How much do observations in groups vary from mean?

$$\begin{aligned}\text{Variability Within Groups } &= \sum (x_i - \bar{x}_i)^2\\ &= \text{Sum of Squares Error}\\ &= \text{SSE}\end{aligned}$$

- Total Variability: How much do observations vary from overall mean?

$$\begin{aligned}\text{Variability Within Groups } &= \sum (x - \bar{x})^2\\ &= \text{Sum of Squares Total}\\ &= \text{SST}\end{aligned}$$

Mean Squares

- Suppose we have a single population of 120 people which we divide randomly into...
  a. 3 groups
  b. 20 groups

Mean Squares

- Suppose we have a single population of 120 people which we divide randomly into...
  a. 3 groups
  b. 20 groups
- All else equal, which of these divisions do we expect to have higher SSG?

$$\text{SSG} = \text{Variability Between Groups} = \sum n_i(\bar{x}_i - \bar{x})^2$$

Mean Squares

- Suppose we have a single population of 120 people which we divide randomly into...
  a. 3 groups
  b. 20 groups
- All else equal, which of these divisions do we expect to have higher SSG?

$$\text{SSG} = \text{Variability Between Groups} = \sum n_i(\bar{x}_i - \bar{x})^2$$

- We standardize sum of squares to compare SSG to SSE

$$\text{Mean Variability Between Groups} = \frac{\text{SSG}}{K-1} = \text{MSG}$$

$$\text{Mean Variability Within Groups} = \frac{\text{SSE}}{n-k} = \text{MSE}$$

Mean Squares

- Suppose we have a single population of 120 people which we divide randomly into...
  a. 3 groups
  b. 20 groups
- All else equal, which of these divisions do we expect to have higher SSG?

$$\text{SSG} = \text{Variability Between Groups} = \sum n_i(\bar{x}_i - \bar{x})^2$$

- We standardize sum of squares to compare SSG to SSE

$$\text{Mean Variability Between Groups} = \frac{\text{SSG}}{K-1} = \text{MSG}$$
$$\text{Mean Variability Within Groups} = \frac{\text{SSE}}{n-k} = \text{MSE}$$

- Our goal is to use MSG and MSE to build a test statistic which measures when variability between groups is much greater than variability within groups

The F Statistic

$$\text{Mean Variability Between Groups} = \frac{\text{SSG}}{K-1} = \text{MSG}$$

$$\text{Mean Variability Within Groups} = \frac{\text{SSE}}{n-k} = \text{MSE}$$

The F Statistic

$$\text{Mean Variability Between Groups } = \frac{\text{SSG}}{K-1} = \text{MSG}$$
$$\text{Mean Variability Within Groups } = \frac{\text{SSE}}{n-k} = \text{MSE}$$

- The F statistic is

$$F = \frac{\text{MSG}}{\text{MSE}} = \frac{\frac{1}{K-1}\sum n_i(\bar{x}_i - \bar{x})^2}{\frac{1}{n-K}\sum(x_i - \bar{x}_i)^2}$$

The F Statistic

$$\text{Mean Variability Between Groups} = \frac{\text{SSG}}{K-1} = \text{MSG}$$
$$\text{Mean Variability Within Groups} = \frac{\text{SSE}}{n-k} = \text{MSE}$$

- The F statistic is

$$F = \frac{\text{MSG}}{\text{MSE}} = \frac{\frac{1}{K-1}\sum n_i(\bar{x}_i - \bar{x})^2}{\frac{1}{n-K}\sum(x_i - \bar{x}_i)^2}$$

- If all observations come from the same population, what is a typical value for F?

The F Statistic

$$\text{Mean Variability Between Groups } = \frac{\text{SSG}}{K-1} = \text{MSG}$$
$$\text{Mean Variability Within Groups } = \frac{\text{SSE}}{n-k} = \text{MSE}$$

- The F statistic is

$$F = \frac{\text{MSG}}{\text{MSE}} = \frac{\frac{1}{K-1}\sum n_i(\bar{x}_i - \bar{x})^2}{\frac{1}{n-K}\sum(x_i - \bar{x}_i)^2}$$

- If all observations come from the same population, what is a typical value for F?

$$F \approx 1$$

The F Statistic

$$\text{Mean Variability Between Groups } = \frac{\text{SSG}}{K-1} = \text{ MSG}$$
$$\text{Mean Variability Within Groups } = \frac{\text{SSE}}{n-k} = \text{ MSE}$$

- The F statistic is

$$F = \frac{\text{MSG}}{\text{MSE}} = \frac{\frac{1}{K-1}\sum n_i(\bar{x}_i - \bar{x})^2}{\frac{1}{n-K}\sum (x_i - \bar{x}_i)^2}$$

- If all observations come from the same population, what is a typical value for F?

$$F \approx 1$$

- If mean of groups are not equal, what values of F are typical?

The F Statistic

$$\text{Mean Variability Between Groups} = \frac{\text{SSG}}{K-1} = \text{MSG}$$
$$\text{Mean Variability Within Groups} = \frac{\text{SSE}}{n-k} = \text{MSE}$$

- The F statistic is

$$F = \frac{\text{MSG}}{\text{MSE}} = \frac{\frac{1}{K-1}\sum n_i(\bar{x}_i - \bar{x})^2}{\frac{1}{n-K}\sum(x_i - \bar{x}_i)^2}$$

- If all observations come from the same population, what is a typical value for F?

$$F \approx 1$$

- If mean of groups are not equal, what values of F are typical?

$$F > 1$$

F is for Films

- We *could* use the previous formulas to calculate the F statistic by hand. . .

F is for Films

- We *could* use the previous formulas to calculate the F statistic by hand. . .
  - But let's use technology!

F is for Films

- We *could* use the previous formulas to calculate the F statistic by hand...
  - But let's use technology!

```
movies_F<- movies %>%
  specify(AudienceScore ~ Genre) %>%
  calculate(stat = "F")
movies_F
```

```
## # A tibble: 1 x 1
##    stat
##   <dbl>
## 1  4.34
```

- Is this a large value of F?

Section 2

The Distribution of the F statistic

The setup for Hypothesis Tests

- Hypotheses
  - **Null Hypothesis**: $H_0 : \mu_1 = \mu_2 = \cdots = \mu_7$
  - **Alternative Hypothesis**: At least one μ is not equal to the others

The setup for Hypothesis Tests

- Hypotheses
  - **Null Hypothesis**: $H_0 : \mu_1 = \mu_2 = \cdots = \mu_7$
  - **Alternative Hypothesis**: At least one μ is not equal to the others
- The F statistic

$$F = \frac{\text{MSG}}{\text{MSE}} = \frac{\text{Variability Between Groups}}{\text{Variability Within Groups}}$$

The setup for Hypothesis Tests

- Hypotheses
  - **Null Hypothesis**: $H_0 : \mu_1 = \mu_2 = \cdots = \mu_7$
  - **Alternative Hypothesis**: At least one μ is not equal to the others
- The F statistic

$$F = \frac{\text{MSG}}{\text{MSE}} = \frac{\text{Variability Between Groups}}{\text{Variability Within Groups}}$$

- Extreme values of the F statistic should give evidence that the null is not true

The setup for Hypothesis Tests

- Hypotheses
  - **Null Hypothesis**: $H_0 : \mu_1 = \mu_2 = \cdots = \mu_7$
  - **Alternative Hypothesis**: At least one μ is not equal to the others
- The F statistic

$$F = \frac{\text{MSG}}{\text{MSE}} = \frac{\text{Variability Between Groups}}{\text{Variability Within Groups}}$$

- Extreme values of the F statistic should give evidence that the null is not true
  - How do we know which values of F are extreme?

The setup for Hypothesis Tests

- Hypotheses
  - **Null Hypothesis**: $H_0 : \mu_1 = \mu_2 = \cdots = \mu_7$
  - **Alternative Hypothesis**: At least one μ is not equal to the others
- The F statistic

$$F = \frac{\text{MSG}}{\text{MSE}} = \frac{\text{Variability Between Groups}}{\text{Variability Within Groups}}$$

- Extreme values of the F statistic should give evidence that the null is not true
  - How do we know which values of F are extreme?
- We can find the distribution F under the null hypothesis by...

The setup for Hypothesis Tests

- Hypotheses
  - **Null Hypothesis**: $H_0 : \mu_1 = \mu_2 = \cdots = \mu_7$
  - **Alternative Hypothesis**: At least one μ is not equal to the others
- The F statistic

$$F = \frac{\text{MSG}}{\text{MSE}} = \frac{\text{Variability Between Groups}}{\text{Variability Within Groups}}$$

- Extreme values of the F statistic should give evidence that the null is not true
  - How do we know which values of F are extreme?
- We can find the distribution F under the null hypothesis by...
  - Randomization
  - Theoretical Approximation.

Randomization and Permutation

- If we assume that the quantitative and categorical variable are independent, then all samples are actually drawn from the same population

Randomization and Permutation

- If we assume that the quantitative and categorical variable are independent, then all samples are actually drawn from the same population
- We can imitate drawing new samples from this population by `permuting` the group labels among observations

Randomization and Permutation

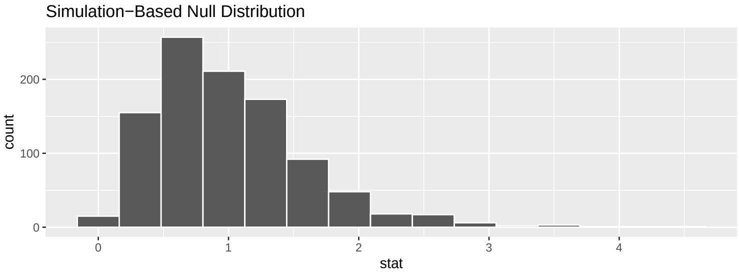
- If we assume that the quantitative and categorical variable are independent, then all samples are actually drawn from the same population
- We can imitate drawing new samples from this population by `permuting` the group labels among observations
    - i.e. we assume that the genre label on a movie is superfluous and shuffle those labels around, while preserving Audience Rating.

Randomization and Permutation

- If we assume that the quantitative and categorical variable are independent, then all samples are actually drawn from the same population
- We can imitate drawing new samples from this population by `permuting` the group labels among observations
  - i.e. we assume that the genre label on a movie is superfluous and shuffle those labels around, while preserving Audience Rating.
- This way, we can study how the size of the F statistic changes just due to random sampling

Randomization and Permutation II

```
null_dist<-movies %>%
  specify(AudienceScore ~ Genre) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute" ) %>%
  calculate(stat = "F")
null_dist %>% visualize()
```



- Most F statistics are at most 3
  - i.e. Assuming independence, Variance BETWEEN groups is at most 3 times variance WITHIN groups

Randomization and Permutation III

How does the observed F statistic compare?

Randomization and Permutation III

How does the observed F statistic compare?

```
movies_F<-movies %>% specify(AudienceScore ~ Genre) %>%  calculate(stat = "F")
movies_F
```

```
##   stat
## 1  4.3
```

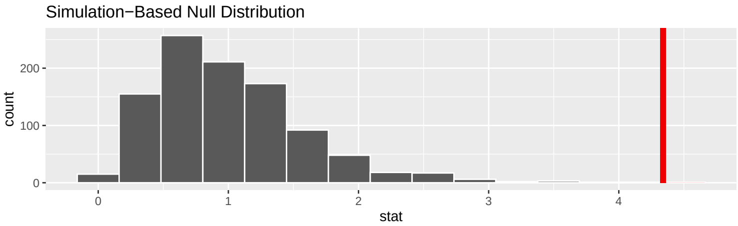
Randomization and Permutation III

How does the observed F statistic compare?

```
movies_F<-movies %>% specify(AudienceScore ~ Genre) %>%  calculate(stat = "F")
movies_F
```

```
##    stat
## 1  4.3
```

```
null_dist %>% visualize()+shade_p_value(obs_stat = movies_F, direction = "right")
```



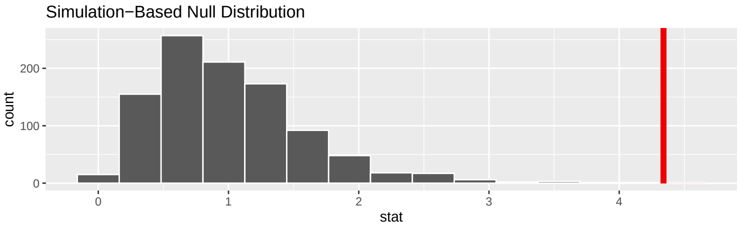
Randomization and Permutation III

How does the observed F statistic compare?

```
movies_F<-movies %>% specify(AudienceScore ~ Genre) %>%  calculate(stat = "F")
movies_F
```

```
##   stat
## 1  4.3
```

```
null_dist %>% visualize()+shade_p_value(obs_stat = movies_F, direction = "right")
```




```
null_dist %>% get_p_value(obs_stat = movies_F, direction = "right")
```

```
##   p_value
## 1   0.001
```

Theoretical Approximation

Like other statistics, the F statistic also has a theoretical distribution

Theoretical Approximation

Like other statistics, the F statistic also has a theoretical distribution

- Suppose we have a total of n observations among k groups and that the following conditions hold:

Theoretical Approximation

Like other statistics, the F statistic also has a theoretical distribution

- Suppose we have a total of n observations among k groups and that the following conditions hold:
    1. Observations are independent
    2. Within each group, values are approximately Normal
    3. Standard deviation is relatively constant between groups

Theoretical Approximation

Like other statistics, the F statistic also has a theoretical distribution

- Suppose we have a total of n observations among k groups and that the following conditions hold:
    1. Observations are independent
    2. Within each group, values are approximately Normal
    3. Standard deviation is relatively constant between groups
- Then the distribution for the F statistic under the null hypothesis is well approximated by the F-distribution with parameters $df_1 = k - 1$ and $df_2 = n - k$.

Theoretical Approximation

Like other statistics, the F statistic also has a theoretical distribution

- Suppose we have a total of n observations among k groups and that the following conditions hold:
    1. Observations are independent
    2. Within each group, values are approximately Normal
    3. Standard deviation is relatively constant between groups
- Then the distribution for the F statistic under the null hypothesis is well approximated by the F-distribution with parameters $df_1 = k - 1$ and $df_2 = n - k$.
    - The p-value is the area in the right tail.

Theoretical Approximation

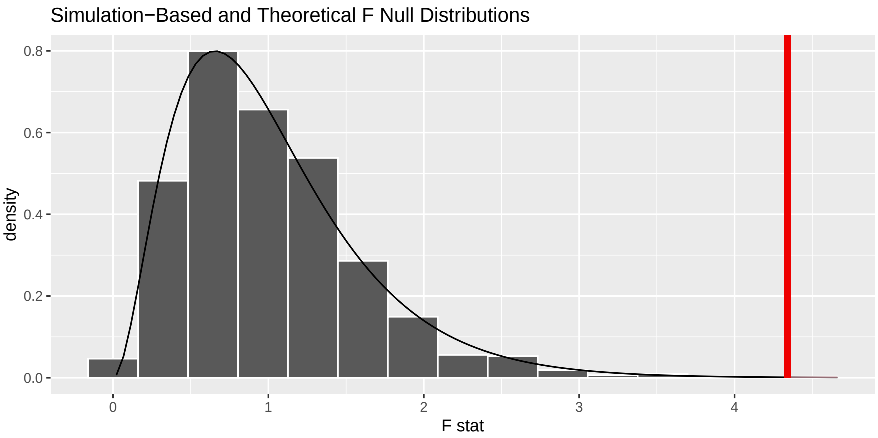
Like other statistics, the F statistic also has a theoretical distribution

- Suppose we have a total of n observations among k groups and that the following conditions hold:
  1. Observations are independent
  2. Within each group, values are approximately Normal
  3. Standard deviation is relatively constant between groups
- Then the distribution for the F statistic under the null hypothesis is well approximated by the F-distribution with parameters $df_1 = k - 1$ and $df_2 = n - k$.
  - The p-value is the area in the right tail.

```
p_value<- pf(q = 4.340672, df1 = 6, df = 125, lower.tail = FALSE)
p_value
```

```
## [1] 0.00052
```

Theory-based and Simulation-based Distributions



There is No Accounting for Taste . . . Even on Average

- The observed F statistic had P-value less than $\alpha = 0.001$

There is No Accounting for Taste ... Even on Average

- The observed F statistic had P-value less than $\alpha = 0.001$
  - This gives extremely good evidence against the Null hypothesis.

There is No Accounting for Taste . . . Even on Average

- The observed F statistic had P-value less than $\alpha = 0.001$
    - This gives extremely good evidence against the Null hypothesis.
    - We conclude that Audience Rating does depend on genre.