Notes and Exercises Chapter 22 – Inference for Comparing Many Means
Math 217 **Name** _____
Saidi **Date** _____

22 Inference for comparing many means

Sometimes we want to compare means across many groups. We might initially think to do pairwise comparisons. For example, if there were three groups, we might be tempted to compare the first mean with the second, then with the third, and then finally compare the second and third means for a total of three comparisons. However, this strategy can be treacherous. If we have many groups and do many comparisons, it is likely that we will eventually find a difference just by chance, even if there is no difference in the populations. Instead, we should apply a holistic test to check whether there is evidence that at least one pair groups are in fact different, and this is where **ANOVA** saves the day.

In this section, we will learn a new method called **analysis of variance (ANOVA)** and a new test statistic called an F-statistic (which we will introduce in our discussion of mathematical models). ANOVA uses a single hypothesis test to check whether the means across many groups are equal:

- Ho: The mean outcome is the same across all groups. In statistical notation, $\mu_1=\mu_2 = \mu_3 = \cdots = \mu_i$, where μ_i represents the mean of the outcome for observations in category i.
- Ha: At least one mean is different.

Generally we must check three conditions on the data before performing ANOVA:

- the observations are independent within and between groups,
- the responses within each group are nearly normal, and
- the variability across the groups is about equal.

When these three conditions are met, we may perform an ANOVA to determine whether the data provide convincing evidence against the null hypothesis that all the μ_i are equal.

Example 1 **Sweet Corn**

Researchers want to study the question of whether organic methods can be successfully used to control harmful insects. They compared the weights of ears of corn under 5 conditions.

Treatment 1: Nematodes
Treatment 2: Parasitic wasps
Treatment 3: Nematodes and parasitic wasps
Treatment 4: Bacteria
Treatment 5: Control

I = 5

Result: Ears of corn were randomly sampled from each plot and weighed.

Look below at the summary and distribution information. ***Notice there is variation both between the groups and within each individual treatment group.***

Table 11.1.1 Weights (ounces) of ears of sweet corn

	Treatment				
	1	2	3	4	5
	16.5	11.0	8.5	16.0	13.0
	15.0	15.0	13.0	14.5	10.5
	11.5	9.0	12.0	15.0	11.0
	12.0	9.0	10.0	9.0	10.0
	12.5	11.5	12.5	10.5	14.0
	9.0	11.0	8.5	14.0	12.0
	16.0	9.0	9.5	12.5	11.0
	6.5	10.0	7.0	9.0	9.5
	8.0	9.0	10.5	9.0	18.5
	14.5	8.0	10.5	9.0	17.0
	7.0	8.0	13.0	6.5	10.0
	10.5	5.0	9.0	8.5	11.0
Mean	11.58	9.63	10.33	11.13	12.29
SD	3.47	2.42	1.96	3.12	2.87
n	12	12	12	12	12

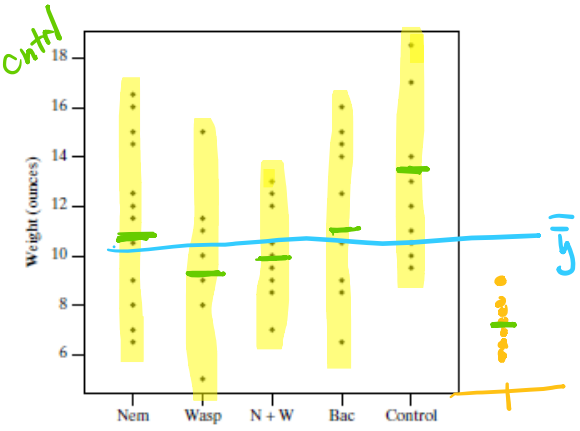


Figure 11.1.1 Weights of ears of corn receiving five different treatments

We extend the idea of a **randomization test** from Chapter 20 to compare all ***I*** groups.

Question:

If each of the results in the table were placed in one of the 5 groups at random, how likely would it be that one of the five would be as large as, say, 12.29 (the mean weight of the control group)? Look at the largest difference between (control – treatment 2 = 2.66)

In a simulation of 10,000 trials to answer this question, 1626 had a difference as large or more than 2.66. Just as we did this type of simulation in chapter 7 with the randomization test, we can see that this proportion of differences of 1626/10000 gives a p-value of 0.163. This means the observed differences could have been due to chance.

22.2.1 Observed data

The method of analysis of variance in this context focuses on answering one question: is the variability in the sample means so large that it seems unlikely to be from chance alone? This question is different from earlier testing procedures since we will *simultaneously* consider many groups, and evaluate whether their sample means differ more than we would expect from natural variation. We call this variability the **mean square between groups (MSG)**, and it has an associated degrees of freedom, $df_{groups} = k - 1$ when there are k groups. The MSG can be thought of as a scaled variance formula for means. If the null hypothesis is true, any variation in the sample means is due to chance and shouldn't be too large. Details of MSG calculations are provided in the footnote.

The mean square between the groups is, on its own, quite useless in a hypothesis test. We need a benchmark value for how much variability should be expected among the sample means if the null hypothesis is true. To this end, we compute a pooled variance estimate, often abbreviated as the **mean square error (**MSE), which has an associated degrees of freedom value $df_{error} = n - k$. It is helpful to think of MSE as a measure of the variability within the groups.

When the null hypothesis is true, any differences among the sample means are only due to chance, and the MSG and MSE should be about equal. As a test statistic for ANOVA, we examine the F-statistic, which is a ratio:

$$F = \frac{MSG}{MSE}$$

The **MSG represents a measure of the between-group variability,** and **MSE measures the variability within** each of the groups.

Conditions:

- independent observations, both within and across groups
- large samples and no extreme outliers

Example 2:

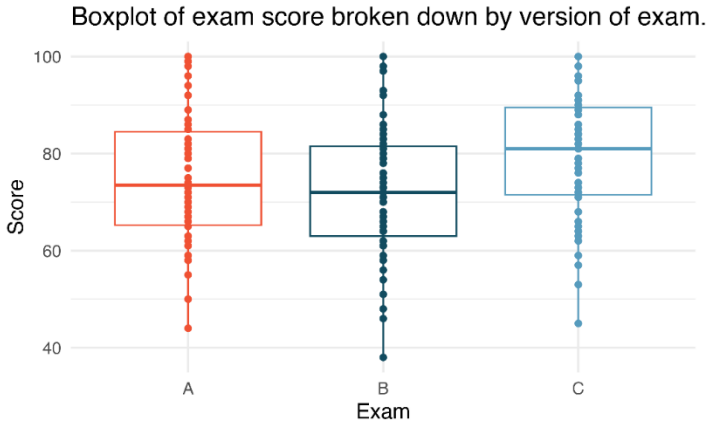
Suppose that a teacher had had such an extremely large class that three different exams were given: A, B, and C. The table and figure below provide a summary of the data including exam C. Again, we would like to investigate whether the difficulty of the exams is the same across the three exams, so the test is

- Ho: μ_A=$\mu_B = \mu_C$ The inherent average difficulty is the same across the three exams.
- At least one of the exams is inherently more (or less) difficult than the others.

The classdata data can be found in the **openintro** R package.

Table: Summary statistics of scores for each exam version.

Exam	n	Mean	SD	Min	Max
A	58	75.1	13.9	44	100
B	55	72.0	13.8	38	100
C	51	78.9	13.1	45	100



The figure below shows the process of randomizing the three different exams to the observed exam scores. If the null hypothesis is true, then the score on each exam should represent the true student ability on that material. It shouldn't matter whether they were given exam A or exam B or exam C. By reallocating which student got which exam, we are able to understand how the difference in average exam scores changes due only to natural variability. There is only one iteration of the randomization process in Figure 22.4, leading to three different randomized sample means (computed assuming the null hypothesis is true).

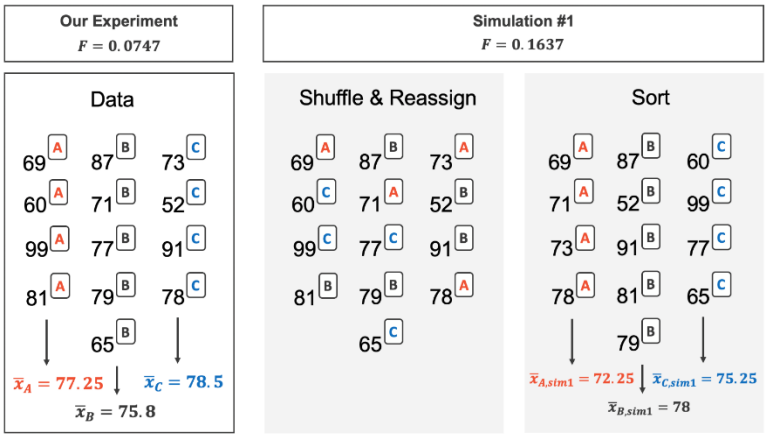


Figure: The version of the test (A or B or C) is randomly allocated to the test scores, under the null assumption that the tests are equally difficult.

In the two-sample case, the null hypothesis was investigated using the difference in the sample means. However, as noted above, with three groups (three different exams), the comparison of the three sample means gets slightly more complicated. We have already derived the F-statistic which is exactly the way to compare the averages across three or

more groups! Recall, the F statistic is a ratio of how the groups differ (MSG) as compared to how the observations within a group vary (MSE).

The figure below shows the values of the simulated F statistics over 1,000 random simulations. We see that, just by chance, the F statistic can be as large as 7.

22.2.3 Observed statistic vs. null statistic

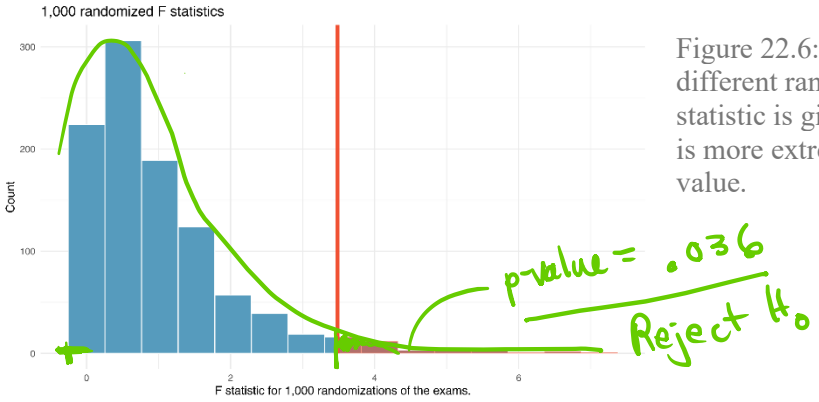


Figure 22.6: Histogram of F statistics calculated from 1000 different randomizations of the exam type. The observed F statistic is given as a red vertical line 3.48. The area to the right is more extreme than the observed value and represents the p-value.

Using statistical software, we can calculate that 3.6% of the randomized F test statistics were at or above the observed test statistic of F=3.48. That is, the p-value of the test is 0.036. Assuming that we had set the level of significance to be α=0.05, the p-value is smaller than the level of significance which would lead us to reject the null hypothesis. We claim that the difficulty level (i.e., the true average score, μ) is different for at least one of the exams.

While it is tempting to say that exam C is harder than the other two (given the inability to differentiate between exam A and exam B in Section 20.1), we must be very careful about conclusions made using different techniques on the same data.

When the null hypothesis is true, random variability that exists in nature produces data with p-values less than 0.05. How often does that happen? 5% of the time. That is to say, if you use 20 different models applied to the same data where there is no signal (i.e., the null hypothesis is true), you are reasonably likely to get a p-value less than 0.05 in one of the tests you run.

The details surrounding the ideas of this problem, called a **multiple comparisons test** or **multiple comparisons problem**, are outside the scope of this textbook, but should be something that you keep in the back of your head. To best mitigate any extra type I errors, we suggest that you set up your hypotheses and testing protocol before running any analyses.

Notation

"y-i-j"

y_{ij} = observation j in group i

Thus, the first observation in the first group is y_{11}, the second observation in the first group is y_{12}, the third observation in the second group is y_{23}, and so on.

We will also use the following notation:

I = number of groups
n_i = number of observations in group i
$\bar{y}_i$ = mean for group i
s_i = standard deviation for group i

The total number of observations is

$$n_\cdot = \sum_{i=1}^{I} n_i$$

"n-dot"

Finally, the **grand mean**—the mean of all the observations—is

$$\bar{\bar{y}} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i} y_{ij}}{n_\cdot}$$

Equivalently we can express $\bar{\bar{y}}$ as a weighted average of the group means

$$\bar{\bar{y}} = \frac{\sum_{i=1}^{I} n_i \bar{y}_i}{\sum_{i} n_i} = \frac{\sum_{i=1}^{I} n_i \bar{y}_i}{n_\cdot}$$

$$S_{pooled} = \sqrt{MS(within)} = \sqrt{\frac{SSwithin}{df\ within}}$$, where **df within** = $n_\cdot - I$

$$MSbetween = \frac{SSbetween}{df\ between}$$, where **df between** = $I - 1$

** Note **

Here, you will see the following notation and definitions we will NOT calculate by hand:

- Grand mean $\bar{\bar{y}}$
- SS(within) – "sum of squares **within groups"** (this is the numerator of the pooled variance)
- MS(within) – "mean square **within groups"** Also known as **MSE,** or Mean Square Error (this also means the pooled variance) The **MSE** can ***quantify the training error*** MSW = MSE
- SS(between) – "sum of squares **between groups"**
- MS(between) – "mean square **between groups**"
- SS(total) – total sum of squares

The Fundamental Relationship of ANOVA

$$y_{ij} - \bar{\bar{y}} = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{\bar{y}})$$

grand "within" difference "between" difference

This expresses the deviation of an observation from the grand mean as the sum of two parts: the "within group" deviation $(y_{ij} - \bar{y}_i)$ and the "between group" deviation $(\bar{y}_i - \bar{\bar{y}})$.

This means that: **SS(total) = SS(within) + SS(between)**

ANOVA Table

ANOVA Quantities with Formulas			
Source	**df**	**SS(Sum of squares)**	**MS(Mean Square)**
Between groups	$I - 1$	$\sum_{n=1}^{I} n_i(\bar{y}_i - \bar{\bar{y}})^2$	SS/df SS Between / df between
Within groups	$n_\bullet - I$	$\sum_{i=1}^{I} (y_{ij} - \bar{y})^2$	SS/df SS within / df within = MSW = MSE
Total	$n_\bullet - 1$	$\sum_{i=1}^{I}\sum_{j=1}^{n_i} (y_{ij} - \bar{\bar{y}})^2$	

22.3.3 Reading an ANOVA table from software

The calculations required to perform an ANOVA by hand are tedious and prone to human error. For these reasons, it is common to use statistical software to calculate the F-statistic and p-value.

An ANOVA can be summarized in a table very similar to that of a regression summary, which we saw in Chapters 7 and 8. The table below shows an ANOVA summary to test whether the mean of on-base percentage varies by player positions in the MLB. Many of these values should look familiar; in particular, the F-statistic and p-value can be retrieved from the last two columns.

term	df	sumsq	meansq	statistic	p.value
position	2	0.0161	0.0080	5.08	0.0066
Residuals	426	0.6740	0.0016		

One statistic not presented on the ANOVA table that might be of interest is the percentage of the variability is in the position explained by the factor levels of the categorical variable. We can find this as the ratio of the sum of squares for position divided by the total sum of squares.

$$SST = \frac{0.0161}{(0.0161+0.6740)} = 0.02332995$$

Percentage of explained variability = 2.33 %

In this case, 2.33% of the variability in position is explained by factor levels of the categorical variable. This is the same as the R^2 value we would obtain if instead performed a linear regression.

Chapter 22 interactive notes

R Saidi

Load the libraries and data

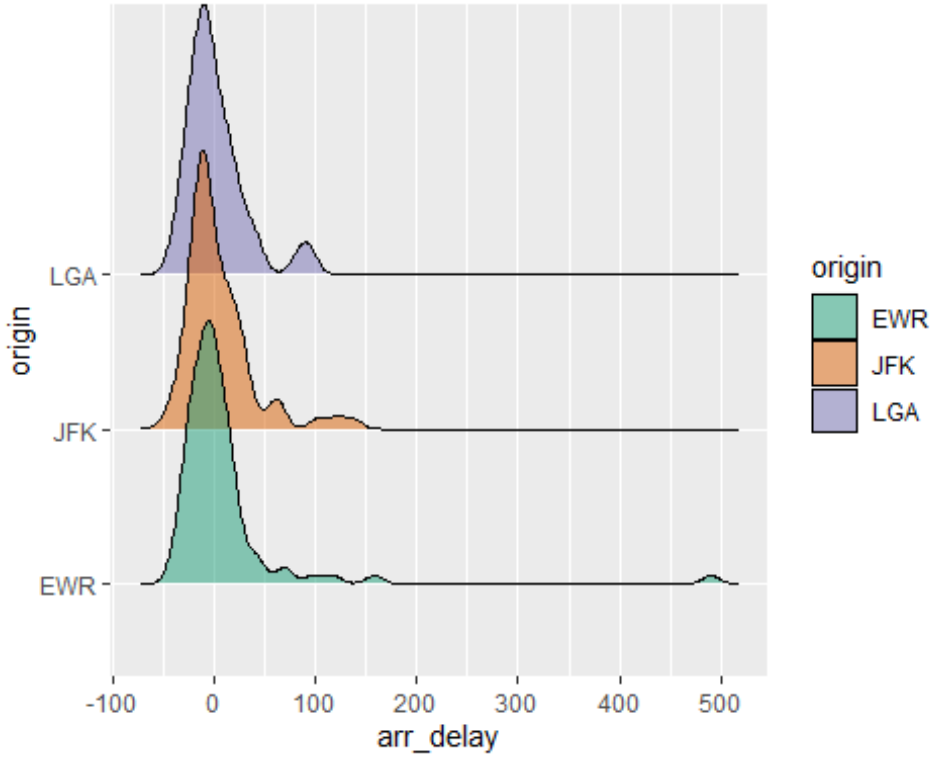
Load this data from the class

```
library(tidyverse)
library(tidymodels)
library(openintro)
data("nycflights")
head(nycflights)
```

```
# A tibble: 6 × 16
   year month   day dep_time dep_delay arr_time arr_delay carrier tailnum flight
  <int> <int> <int>    <int>     <dbl>    <int>     <dbl> <chr>   <chr>    <int>
1  2013     6    30      940        15     1216        -4 VX      N626VA     407
2  2013     5     7     1657        -3     2104        10 DL      N3760C     329
3  2013    12     8      859        -1     1238        11 DL      N712TW     422
4  2013     5    14     1841        -4     2122       -34 DL      N914DL    2391
5  2013     7    21     1102        -3     1230        -8 9E      N823AY    3652
6  2013     1     1     1817        -3     2008         3 AA      N3AXAA     353
#
```

```
summarytable
```

```
# A tibble: 35 × 6
# Groups:   origin [3]
   origin carrier mean_delay sd_delay median_delay count
   <chr>  <chr>        <dbl>    <dbl>        <dbl> <int>
 1 EWR    9E           3.11      39.7         -9     121
 2 EWR    AA          -2.55      41.6        -13     350
 3 EWR    AS         -11.3       41.9        -19.5    66
 4 EWR    B6           9.35      44.2         -5     625
 5 EWR    DL          10.7       61.6         -3     445
 6 EWR    EV          17.4       49.8          1    4170
 7 EWR    MQ          14.8       58.1         -1.5   210
 8 EWR    OO          -5         NA           -5       1
 9 EWR    UA           3.91      40.0         -6    4559
10 EWR    US           0.444     32.4         -7     444
#
```



Again, the distributions appear very similar.

```
# Run an analysis of variance on score vs. rank
aov_delay_origin <- aov(arr_delay ~ origin, data = samp)
#summary(aov_delay_origin)
# Tidy the model
tidy(aov_delay_origin)
```

```
# A tibble: 2 × 6
  term         df   sumsq meansq statistic p.value
  <chr>     <dbl>   <dbl>  <dbl>     <dbl>   <dbl>
1 origin        2   1702.   851.     0.385   0.681
2 Residuals   197 435435.  2210.    NA      NA
```

```
# run the ANOVA on the original dataset and you get the opposite result
# Run an analysis of variance on score vs. rank
aov_delay_nyc <- aov(arr_delay ~ origin, data = nycflights)
#summary(aov_delay_origin)
# Tidy the model
tidy(aov_delay_nyc)
```

```
# A tibble: 2 × 6
  term         df     sumsq meansq statistic   p.value
  <chr>     <dbl>     <dbl>  <dbl>     <dbl>     <dbl>
1 origin        2    91477. 45739.      22.9  1.11e-10
2 Residuals 32732 65276924.  1994.      NA   NA
```

Interpret the results

For the original data, the p-value of the ANOVA test is very small. Therefore we reject the null. There is very strong evidence that at least one mean arrival delay from one of the 3 airport origins is different.

BUT from a random sample, the p-value is very large, meaning we fail to reject the null. This suggests that simply having a large sample size will coerce an effect, when there really is no effect.

Simulating samples under the null hypothesis

First calculate the observed statistic

In the case of an ANOVA, the statistic we are interested in is the F-statistic. While we can plot this statistic on an F-distribution, it really is just another statistic that we can calculate (like the mean or median). We like this statistic because it allows for us to summarize how different multiple means are from each other, relative to how variable the observations are within each group.

We can calculate the F-statistic using the tools from the infer package we are familiar with. The only part that is new is the stat that we calculate. Here, we use the "F" statistic.

```
obs_stat <- nycflights |>
  specify(arr_delay ~ origin) |>
  calculate(stat = "F")
obs_stat
```

```
Response: arr_delay (numeric)
Explanatory: origin (factor)
# A tibble: 1 × 1
   stat
  <dbl>
1  22.9
```

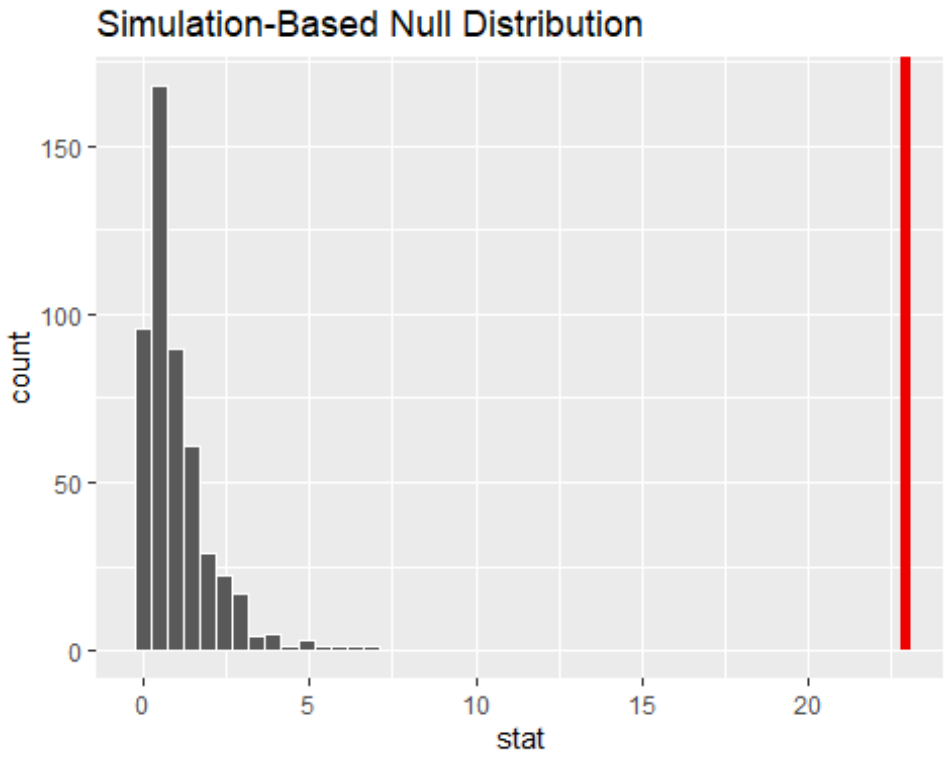
The next step is to simulate what we would expect for arrival delays to look like, if the null hypothesis was true. This is similar to the method for a difference in means, except now we have three groups: jfk, lga, and ewr. The underlying process, however, looks the same:

Step 1: Write the values of arr_delay on 32735 index cards (one card per person). Step 2: Shuffle the cards and randomly split them into three new piles, of the same size as the original groups. Step 3: Calculate and record the test statistic: F-statistic Step 4: Repeat steps (1) and (2) 1000 to generate the sampling distribution of the difference in means under the null hypothesis. Step 5: Calculate p-value as the percentage of simulations where the test statistic is at least as extreme as the observed F-statistic

```
null_distr <- nycflights |>
 specify(arr_delay ~ origin) |>
  hypothesize(null = "independence") |>
  generate(reps = 500, type = "permute") |>
  calculate(stat = "F")

null_distr |>
  visualise() +
  shade_p_value(obs_stat = obs_stat, direction = "greater")
```

```
Warning in min(diff(unique_loc)): no non-missing arguments to min; returning
Inf
```



We can see that the observed F-statistic is very far above the F distribution. therefore we reject the null. There is very strong evidence that the mean arrival delay is different at at least one airport.

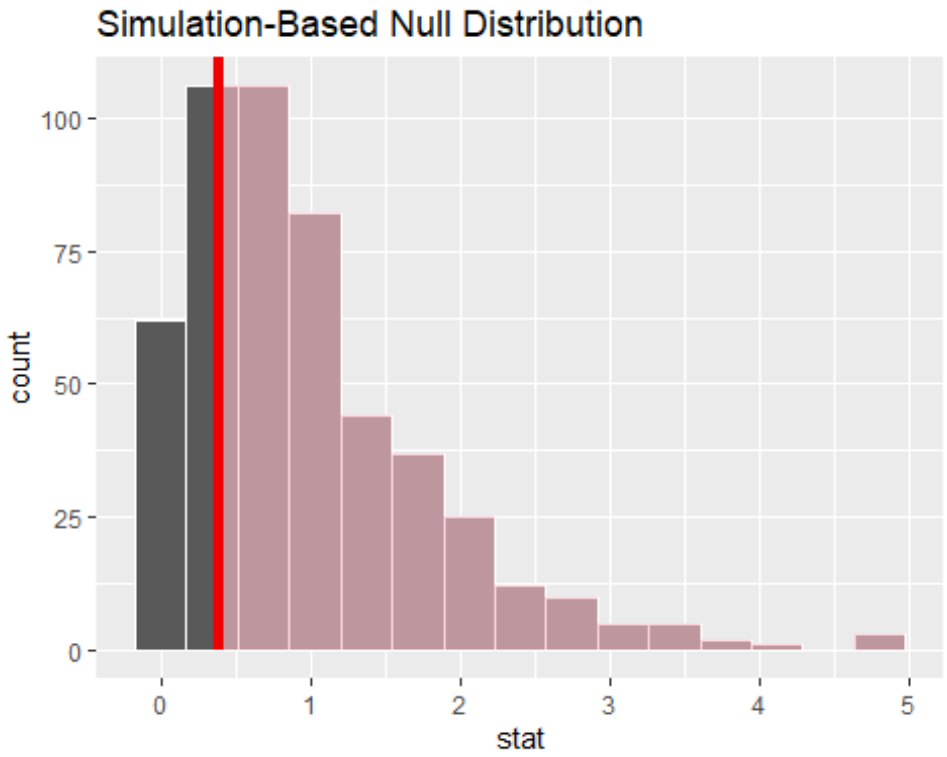
Try all of that again, but use the small sample

```
obs_stat_samp <- samp |>
  specify(arr_delay ~ origin) |>
  calculate(stat = "F")
obs_stat_samp
```

```
Response: arr_delay (numeric)
Explanatory: origin (factor)
# A tibble: 1 × 1
   stat
  <dbl>
1 0.385
```

```
null_distr_samp <- samp |>
 specify(arr_delay ~ origin) |>
  hypothesize(null = "independence") |>
  generate(reps = 500, type = "permute") |>
  calculate(stat = "F")

null_distr_samp |>
  visualise() +
  shade_p_value(obs_stat = obs_stat_samp, direction = "greater")
```



Again, we get the opposite effect. This means sampling is incredibly important!!!!

Post hoc testing

If you reject the null, now you can determine by pair-wise testing which group's mean is different using a "family" error rate, and then distribute that level to each of the tests we are performing. The "family" error rate specifies an overall Type I error rate we are willing to have for all of tests you wish to perform. We will use alpha = 0.05

There are many pairwise tests. We will use the TukeyHSD test on our original model, aov_delay_origin.

```
TukeyHSD(aov_delay_nyc)

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = arr_delay ~ origin, data = nycflights)

$origin
               diff       lwr       upr     p adj
JFK-EWR -3.3429714 -4.734340 -1.951603 0.0000001
LGA-EWR -3.6197575 -5.040598 -2.198917 0.0000000
LGA-JFK -0.2767861 -1.723655  1.170083 0.8951358
```

How to read this output

The p-adjusted values for comparing mean delays for JFK and EWR are significant The p-adjusted values for comparing mean delays for LGA and EWR are significant The p-adjusted values for comparing mean delays for LGA and JFK are not significant.

Non-parametric test Kruskall wallis Test

```
ktest_nyc <- kruskal.test(arr_delay ~ origin, data = nycflights)
ktest_nyc
```

```
    Kruskal-Wallis rank sum test

data:  arr_delay by origin
Kruskal-Wallis chi-squared = 58.591, df = 2, p-value = 1.893e-13
```

```
# and on the sample
ktest_samp <- kruskal.test(arr_delay ~ origin, data = samp)
ktest_samp
```

```
    Kruskal-Wallis rank sum test

data:  arr_delay by origin
Kruskal-Wallis chi-squared = 0.20372, df = 2, p-value = 0.9032
```

Homework Chapter 22

1. Review section 22.4 (the chapter review)
2. Suggested problems from textbook section 22.5 exercises: 4-7, 9, 14
3. Suggested tutorials:

   8 - Comparing many means