

# Bootstrap Applications Solutions

Lt Col Ken Horton

Lt Col Kris Pruitt

Professor Bradley Warner

25 October, 2020

## Exercises

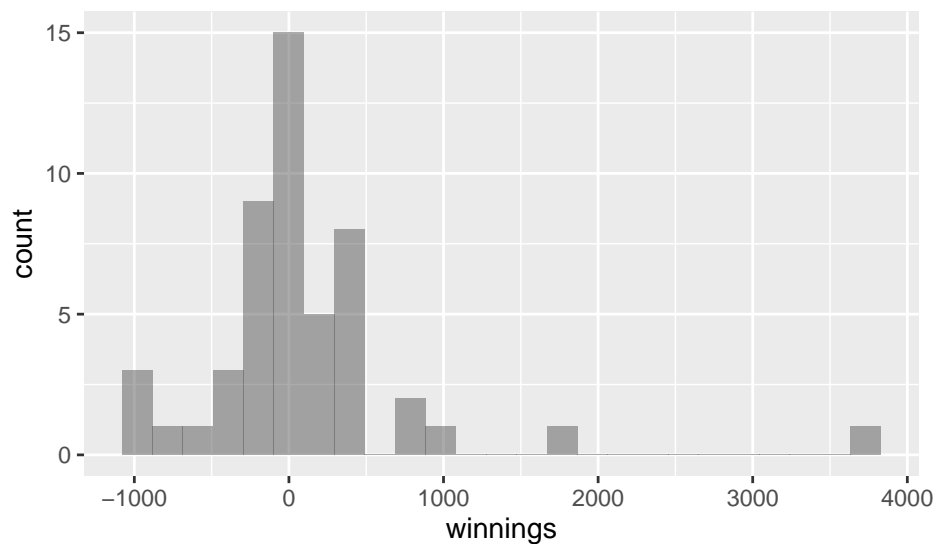
### 1. Poker

An aspiring poker player recorded her winnings and losses over 50 evenings of play, the data is in the `openintro` package in the object `poker`. The poker player would like to better understand the volatility in her long term play.

- a. Load the data and plot a histogram.

```
poker<-read_csv("data/poker.csv")
```

```
poker %>%  
  gf_histogram(~winnings)
```



- b. Find the summary statistics.

```
favstats(~winnings,data=poker)
```

```
##   min   Q1 median  Q3   max  mean      sd  n missing  
## -1000 -187    11 289 3712 90.08 703.6835 50      0
```

- c. *Mean absolute deviation* or *MAD* is a more intuitive measure of spread than variance. It directly measures the average distance from the mean. It is found by the formula:

$$mad = \sum_{i=1}^n \frac{|x_i - \bar{x}|}{n}$$

Write a function and find the *MAD* of the data.

```
mad<-function(x){  
  xbar<-mean(x)  
  sum(abs(x-xbar))/length(x)  
}
```

```
obs<-mad(poker$winnings)  
obs
```

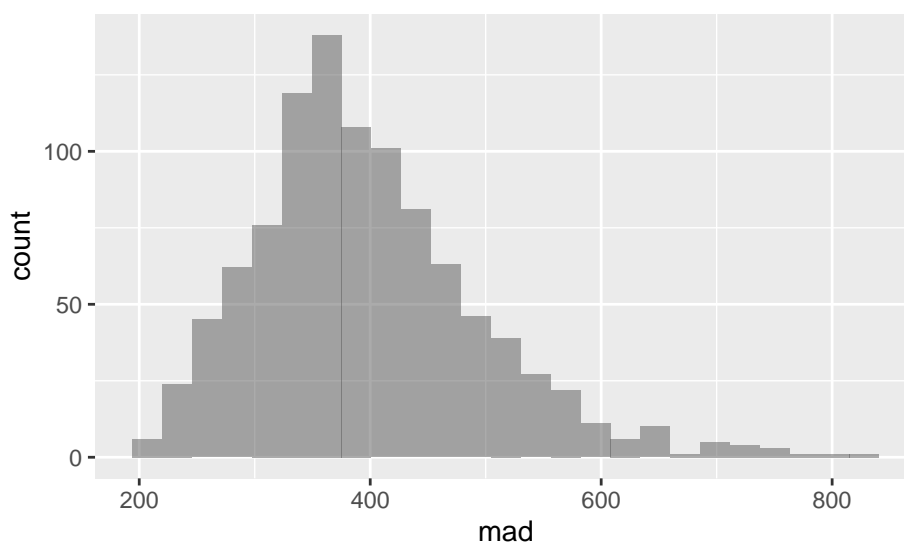
```
## [1] 394.1792
```

- d. Find the bootstrap distribution of the *MAD* using 1000 replicates.

```
set.seed(1122)  
results<-do(1000)*mad(resample(poker$winnings))
```

- e. Plot a histogram of the bootstrap distribution.

```
results %>%  
  gf_histogram(~mad)
```



- f. Report a 95% confidence interval on the MAD.

```
cdata(~mad,data=results)
```

```
##          lower    upper central.p  
## 2.5% 243.9448 636.0925      0.95
```

g. ADVANCED: Do you think sample MAD is an unbiased estimator of population MAD? Why or why not?

We don't know without doing some math. We do know that the sample standard deviation is biased and part of that is because we have to use the sample mean in its calculation. We are doing the same thing here, so our estimate might also be biased for the same reason.

## 2. Bootstrap hypothesis testing

Bootstrap hypothesis testing is relatively undeveloped, and is generally not as accurate as permutation testing. Therefore in general avoid it. But for our problem in the notes, it may work. We will sample in a way that is consistent with the null hypothesis, then calculate a P-value as a tail probability like we do in permutation tests. This example does not generalize well to other applications like relative risk, correlation, regression, or categorical data.

a. Using the HELPrct data set, store the observed value of the difference of means for male and female.

I am going to just select the two columns I need.

```
HELP_sub <- HELPrct %>%  
  select(age,sex)
```

```
obs <- diffmean(age~sex,data=HELP_sub)  
obs
```

```
## diffmean  
## -0.7841284
```

b. The null hypothesis requires the means of each group to be equal. Pick one group to adjust, either male or female. First zero the mean of the selected group by subtracting the sample mean of this group from data points only in this group. Then add the sample mean of the other group to each data point in the selected group. Store in a new object called HELP\_null.

This is tricky, we doing some data wrangling here.

```
means<-mean(age~sex,data=HELP_sub)  
means
```

```
## female    male  
## 36.25234 35.46821
```

```
means['female']
```

```
## female  
## 36.25234
```

Let's get all the female and adjust the mean.

```
H_female <- HELP_sub %>%
  filter(sex=="female") %>%
  mutate(age=age-means['female']+means['male'])
```

```
mean(~age,data=H_female)
```

```
## [1] 35.46821
```

Combine back into one data set.

```
HELP_sub_new<-HELP_sub %>%
  filter(sex=="male") %>%
  rbind(H_female)
```

c. Run `favstats()` to check that the means are equal.

```
favstats(age~sex,data=HELP_sub_new)
```

```
##      sex      min      Q1  median      Q3      max      mean      sd      n
## 1 female 20.21587 30.21587 34.21587 39.71587 57.21587 35.46821 7.584858 107
## 2  male 19.00000 30.00000 35.00000 40.00000 60.00000 35.46821 7.750110 346
##   missing
## 1         0
## 2         0
```

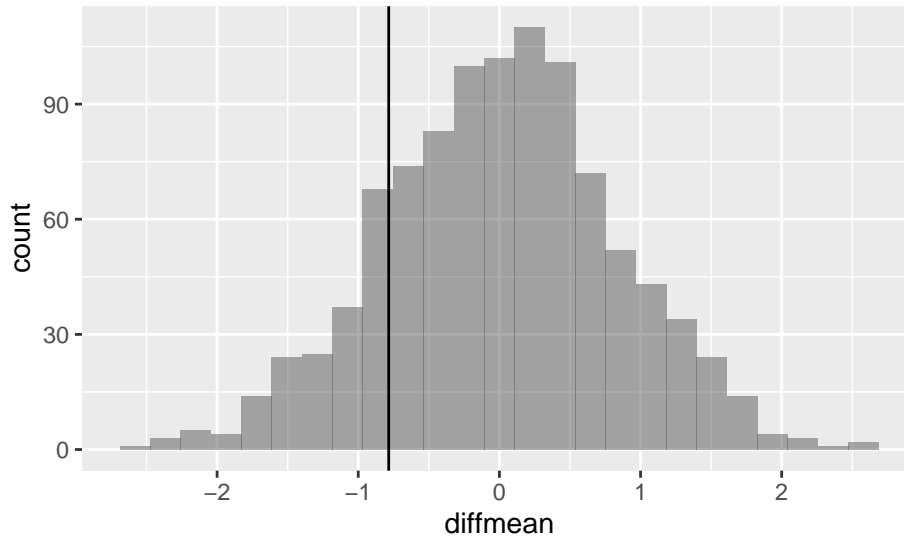
d. On this new adjusted data set, generate a bootstrap distribution of the difference in sample means.

```
set.seed(1159)
results<-do(1000)*diffmean(age~sex,data=resample(HELP_sub_new))
```

e. Plot the bootstrap distribution and a line at the observed difference in sample means.

```
results %>%
  gf_histogram(~diffmean) %>%
  gf_vline(xintercept=obs)
```

```
## Warning: geom_vline(): Ignoring 'mapping' because 'xintercept' was provided.
```



f. Find a p-value.

```
2*prop1(~(diffmean<=obs),data=results)
```

```
## prop_TRUE
## 0.3476523
```

g. How does the p-value compare with those in the notes.

This is a similar p-value.

### 3. Paired data

Are textbooks actually cheaper online? Here we compare the price of textbooks at the University of California, Los Angeles' (UCLA's) bookstore and prices at Amazon.com. Seventy-three UCLA courses were randomly sampled in Spring 2010, representing less than 10% of all UCLA courses. When a class had multiple books, only the most expensive text was considered.

The data is in the file `textbooks.csv` under the data folder.

```
textbooks<-read_csv("data/textbooks.csv")
```

```
## Parsed with column specification:
## cols(
##   dept_abbr = col_character(),
##   course = col_character(),
##   isbn = col_character(),
##   ucla_new = col_double(),
##   amaz_new = col_double(),
##   more = col_character(),
##   diff = col_double()
## )
```

```
head(textbooks)
```

```
## # A tibble: 6 x 7
##   dept_abbr course isbn          ucla_new amaz_new more   diff
##   <chr>      <chr> <chr>          <dbl>     <dbl> <chr> <dbl>
## 1 Am Ind    C170  978-0803272620    27.7     28.0 Y   -0.28
## 2 Anthro    9     978-0030119194    40.6     31.1 Y    9.45
## 3 Anthro   135T  978-0300080643    31.7     32   Y   -0.32
## 4 Anthro   191HB  978-0226206813    16      11.5 Y    4.48
## 5 Art His   M102K  978-0892365999    19.0     14.2 Y    4.74
## 6 Art His   118E   978-0394723693    15.0     10.2 Y    4.78
```

Each textbook has two corresponding prices in the data set: one for the UCLA bookstore and one for Amazon. Therefore, each textbook price from the UCLA bookstore has a natural correspondence with a textbook price from Amazon. When two sets of observations have this special correspondence, they are said to be **paired**.

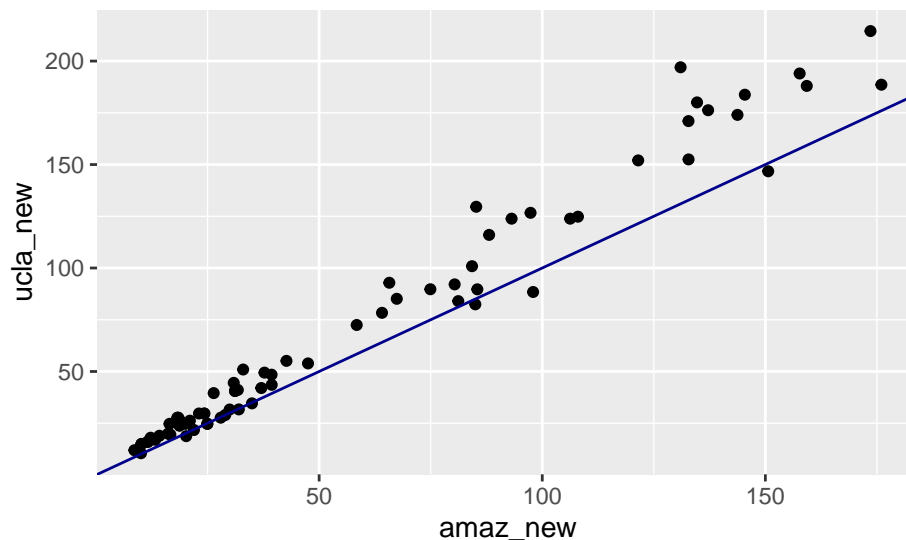
To analyze paired data, it is often useful to look at the difference in outcomes of each pair of observations. In `textbooks`, we look at the difference in prices, which is represented as the `diff` variable. It is important that we always subtract using a consistent order; here Amazon prices are always subtracted from UCLA prices.

a. Is this data tidy? Explain.

Yes, because each row is a textbook and each column is a variable.

b. Make a scatterplot of the UCLA price versus the Amazon price. Add a 45 degree line to the plot.

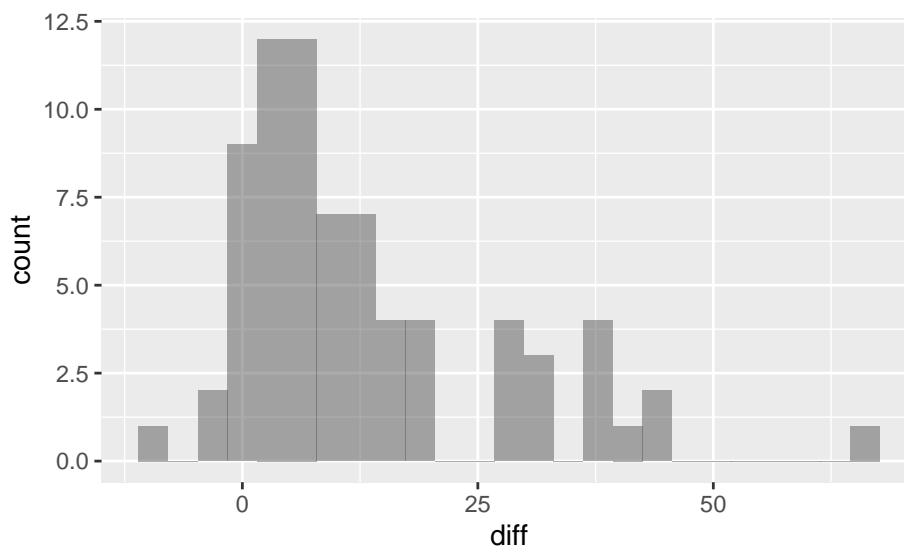
```
textbooks %>%
  gf_point(ucla_new~amaz_new) %>%
  gf_abline(slope=1,intercept = 0,color="darkblue")
```



It appears the books at the UCLA bookstore are more expensive. One way to test this is with a regression model; we will learn about in the next block.

c. Make a histogram of the differences in price.

```
textbooks %>%
  gf_histogram(~diff)
```



The distribution is skewed.

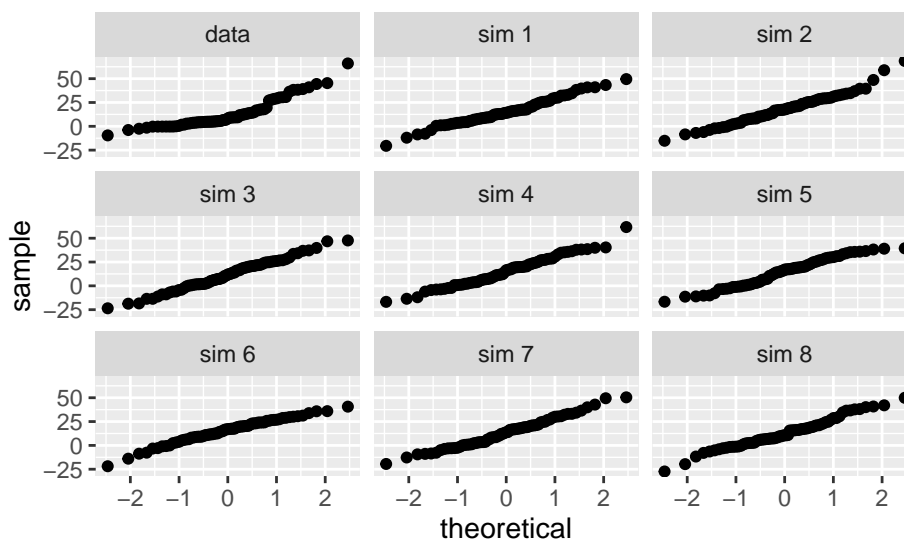
The hypotheses are:

$H_0: \mu_{diff} = 0$ . There is no difference in the average textbook price.

$H_A: \mu_{diff} \neq 0$ . There is a difference in average prices.

d. To use a  $t$  distribution, the variable `diff` has to be independent and normally distributed. Since the 73 books represent less than 10% of the population, the assumption that the random sample is independent is reasonable. Check normality using `qqnormsim()` from the `openintro` package. It generates 8 qq plots of simulated normal data that you can use to judge the `diff` variable.

```
qqnormsim(diff, textbooks)
```



The normality assumption is suspect but we have a large sample so it should be acceptable to use the  $t$ .

- e. Run a  $t$  test on the `diff` variable. Report the p-value and conclusion.

```
t_test(~diff,textbooks)
```

```
##
## One Sample t-test
##
## data: diff
## t = 7.6488, df = 72, p-value = 6.928e-11
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  9.435636 16.087652
## sample estimates:
## mean of x
## 12.76164
```

We did not have to use the `paired` option since we already took the difference. Here is an example of using the `paired` option.

```
t_test(textbooks$ucla_new,textbooks$amaz_new,paired=TRUE)
```

```
##
## Paired t-test
##
## data: textbooks$ucla_new and textbooks$amaz_new
## t = 7.6488, df = 72, p-value = 6.928e-11
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  9.435636 16.087652
## sample estimates:
## mean of the differences
##                12.76164
```

The p-value is so small that we don't believe the average price of the books from the UCLA bookstore and Amazon are the same.

- f. Create a bootstrap distribution and generate a 95% confidence interval on the mean of the differences, the `diff` column.

```
textbooks %>%
  summarise(obs_diff=mean(diff))
```

```
## # A tibble: 1 x 1
##   obs_diff
##   <dbl>
## 1      12.8
```

We need to just pull the difference.



```
obs_stat<- textbooks %>%
  summarise(obs_diff=mean(diff)) %>%
  pull(obs_diff)
```

```
obs_stat
```

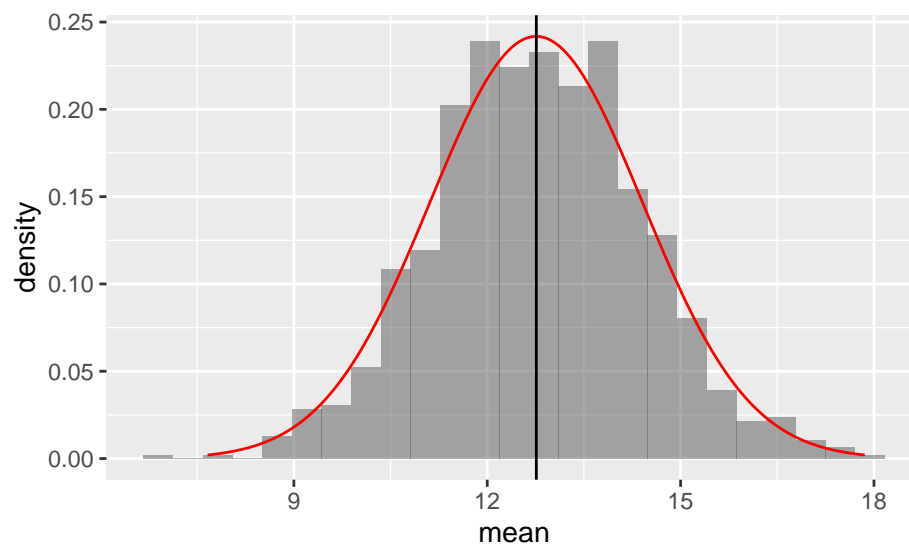
```
## [1] 12.76164
```

Next a bootstrap distribution.

```
set.seed(843)
results<-do(1000)*mean(~diff,data=resample(textbooks))
```

```
results %>%
  gf_dhistogram(~mean) %>%
  gf_dist("norm",mean=12.76,sd=14/sqrt(72),color="red") %>%
  gf_vline(xintercept = obs_stat)
```

```
## Warning: geom_vline(): Ignoring 'mapping' because 'xintercept' was provided.
```



```
cdata(~mean,data=results)
```

```
##      lower    upper central.p
## 2.5% 9.583829 16.05705      0.95
```

Not a bad solution for this problem.

- g. If there is really no differences between book sources, the variable `more` is a binomial and under the null the probability of success is  $\pi = 0.5$ . Run a hypothesis test using the variable `more`.

```
inspect(textbooks)
```

```
##
## categorical variables:
##      name      class levels  n missing
## 1 dept_abbr character    41 73      0
## 2   course character    66 73      0
## 3    isbn character    73 73      0
## 4    more character     2 73      0
##
##                                distribution
## 1 Mgmt (8.2%), Pol Sci (6.8%) ...
## 2 10 (4.1%), 101 (2.7%), 180 (2.7%) ...
## 3 978-0030119194 (1.4%) ...
## 4 Y (61.6%), N (38.4%)
##
## quantitative variables:
##      name  class  min   Q1 median   Q3  max   mean   sd  n
## ...1 ucla_new numeric 10.50 24.70 43.56 116.00 214.5 72.22192 59.65913 73
## ...2 amaz_new numeric  8.60 20.21 34.95  88.09 176.0 59.46027 48.99557 73
## ...3   diff numeric -9.53  3.80  8.23  17.59  66.0 12.76164 14.25530 73
##      missing
## ...1      0
## ...2      0
## ...3      0
```

We have 45 books that were more expensive out of the total of 73.

```
prop_test(45,73,p=0.5)
```

```
##
## 1-sample proportions test with continuity correction
##
## data: 45 out of 73
## X-squared = 3.5068, df = 1, p-value = 0.06112
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.4948968 0.7256421
## sample estimates:
##      p
## 0.6164384
```

Notice that this test failed to reject the null hypothesis. In the paired test, the evidence was so strong but in the binomial model it is not. There is a loss of information making a discrete variable out of a continuous one.

h. Could you use a permutation test on this example? Explain.

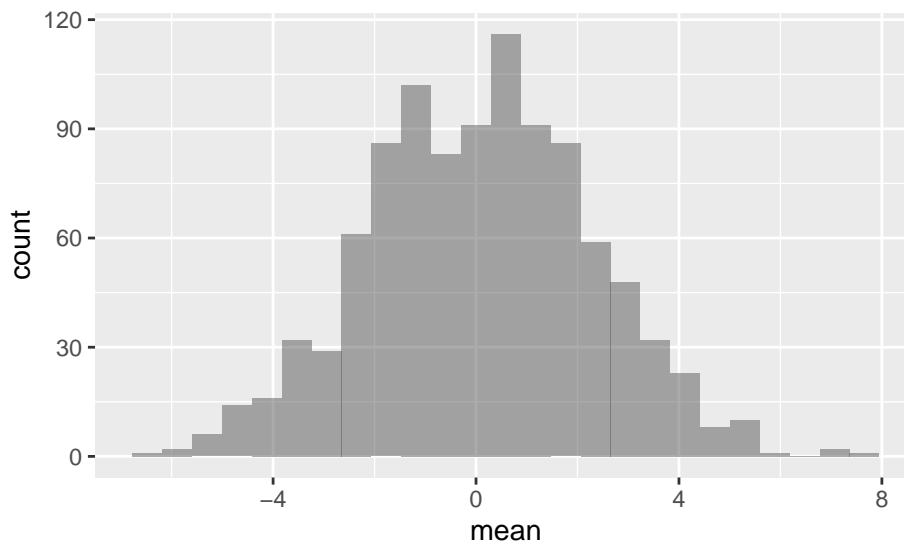
Yes, but you have to be careful because you want to keep the pairing so you can't just shuffle the names. You have to shuffle the names within the paired values. This means to simply randomly switch the names within a row. This is easier to do by just multiplying the diff column by a random choice of -1 and 1.

```
sample(c(-1,1),size=73,replace = TRUE)
```

```
## [1]  1 -1 -1  1 -1 -1 -1  1 -1 -1  1 -1 -1  1 -1  1  1 -1 -1 -1 -1  1 -1 -1
## [26] -1 -1  1  1  1  1  1 -1  1  1  1  1  1 -1 -1  1  1 -1  1  1  1  1 -1 -1
## [51] -1 -1 -1  1 -1  1 -1 -1  1  1 -1  1  1 -1 -1 -1  1  1 -1  1 -1  1  1
```

```
set.seed(406)
results <- do(1000)*mean((~diff*sample(c(-1,1),size=73,replace = TRUE)),data=textbooks)
```

```
results %>%
  gf_histogram(~mean)
```



```
prop1((~mean>=obs_stat),data=results)
```

```
## prop_TRUE
## 0.000999001
```

None of the permuted values is at or greater than the observed value.

## File Creation Information

- File creation date: 2020-10-25
- Windows version: Windows 10 x64 (build 18362)
- R version 3.6.3 (2020-02-29)
- mosaic package version: 1.7.0
- tidyverse package version: 1.3.0
- openintro package version: 2.0.0