Hypothesis Testing

El Mex

Contents


```
library(tidyverse)
library(infer)
library(moderndive)
library(nycflights13)
library(ggplot2movies)
```

1 Hypothesis Testing

1.1 Promotions activity

1.1.1 Does gender affect promotions at a bank?

Say you are working at a bank in the 1970s and you are submitting your résumé to apply for a promotion. Will your gender affect your chances of getting promoted? To answer this question, we'll focus on data from a study published in the *Journal of Applied Psychology* in 1974. This data is also used in the OpenIntro series of statistics textbooks.

The `promotions` data frame on `moderndive` contains the data on the 48 applicants. Let's explore it by looking at six randomly selected rows:

```
glimpse(promotions)
```

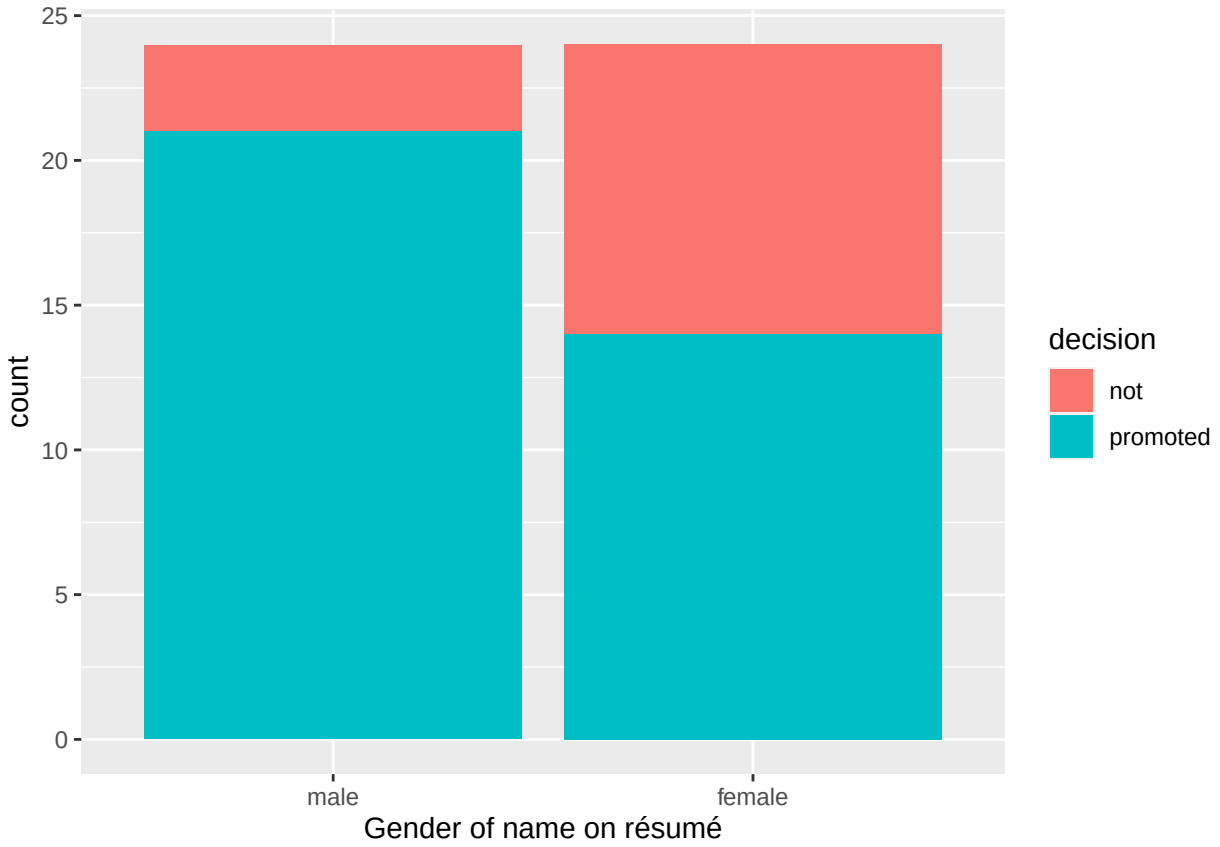
```
## Rows: 48
## Columns: 3
## $ id       <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,...
## $ decision <fct> promoted, promoted, promoted, promoted, promoted, promoted...
## $ gender   <fct> male, male, male, male, male, male, male, male, male, male...
```

```
promotions %>%
  sample_n(size = 6) %>%
  arrange(id)
```

```
## # A tibble: 6 x 3
##      id decision gender
##   <int> <fct>    <fct>
## 1     8 promoted male
## 2    19 promoted male
## 3    21 promoted male
## 4    22 promoted female
## 5    34 promoted female
## 6    42 not      female
```

Perform an exploratory data analysis of the relationship between the two categorical variables `decision` and `gender`. You can use a stacked barplot.

```
ggplot(promotions, aes(x = gender, fill = decision)) +
  geom_bar() +
  labs(x = "Gender of name on résumé")
```



It seems résumés with female names were much less likely to be accepted for promotion. Let's quantify these promotion rates by computing the proportion of résumés accepted for promotion for each group.

```
# Use of the tally() function here which is a shortcut for summarize(n = n()) to get counts
promotions %>%
  group_by(gender, decision) %>%
  tally()
```

```
## # A tibble: 4 x 3
## # Groups:   gender [2]
##   gender decision     n
##   <fct>  <fct>    <int>
## 1 male   not          3
## 2 male   promoted    21
## 3 female not         10
## 4 female promoted    14
```

Of the 24 male résumés, 21 were promoted ($21/24 = 0.875 = 87.55\%$). On the other hand, of the 24 female résumés, 14 were selected for promotion ($14/24 = 0.583 = 58.3\%$).

Does this provide conclusive evidence that there is gender discrimination in promotions at banks? Could a difference in promotion rates of 29.2% still occur by chance, even in a hypothetical world where no gender-based discrimination existed? In other words, what is the role of sampling variation in this hypothesized world?

1.1.2 Shuffling once

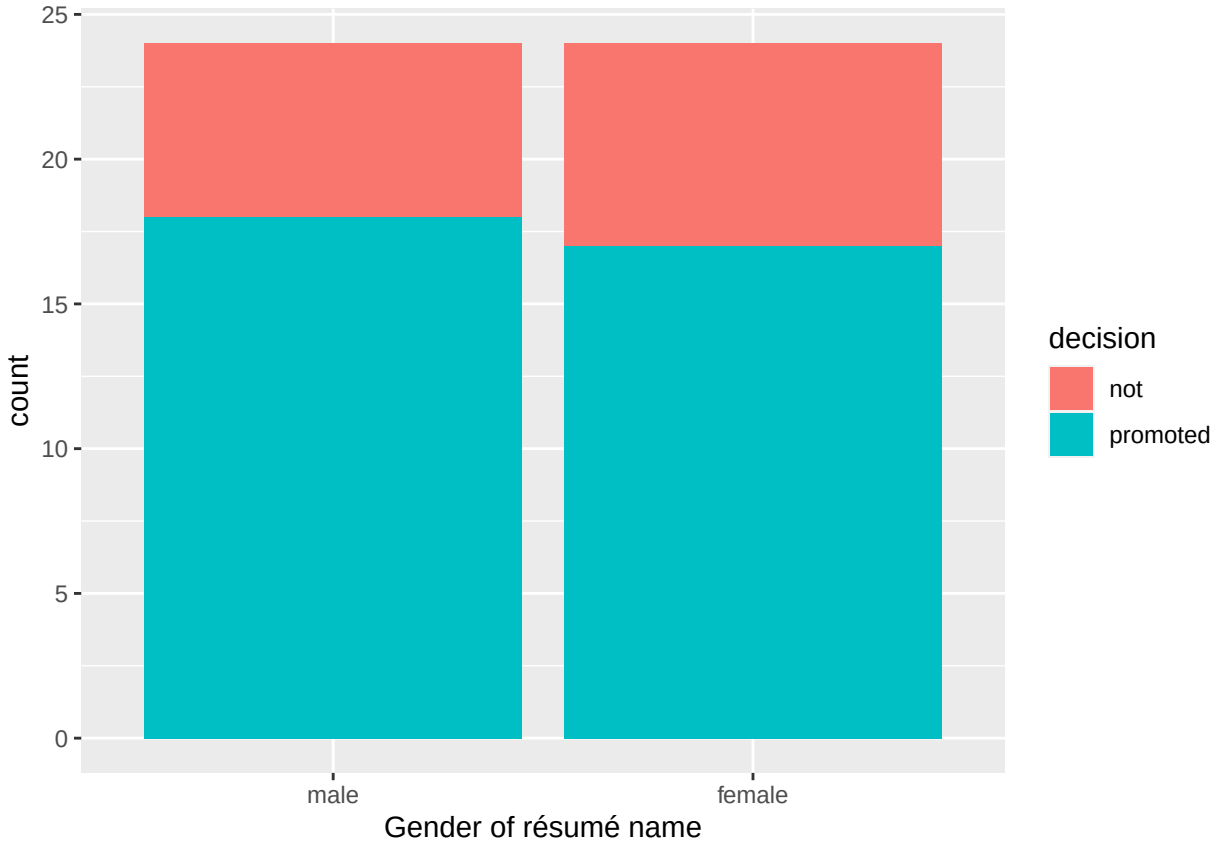
Imagine a hypothetical universe where no gender discrimination in promotions existed. If these gender labels were irrelevant, then we could randomly reassign them by "shuffling" them to no consequence!

We've saved one such shuffling in the `promotions_shuffled` data frame. Explore it and create the previous barplot.

```
glimpse(promotions_shuffled)
```

```
## Rows: 48
## Columns: 3
## $ id       <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,...
## $ decision <fct> promoted, promoted, promoted, promoted, promoted, promoted...
## $ gender   <fct> female, female, male, female, male, male, male, female, ma...
```

```
ggplot(promotions_shuffled, aes(x = gender, fill = decision)) +
  geom_bar() +
  labs(x = "Gender of résumé name")
```


25
20
15
10
5
0
count
male
female
Gender of résumé name
decision
not
promoted

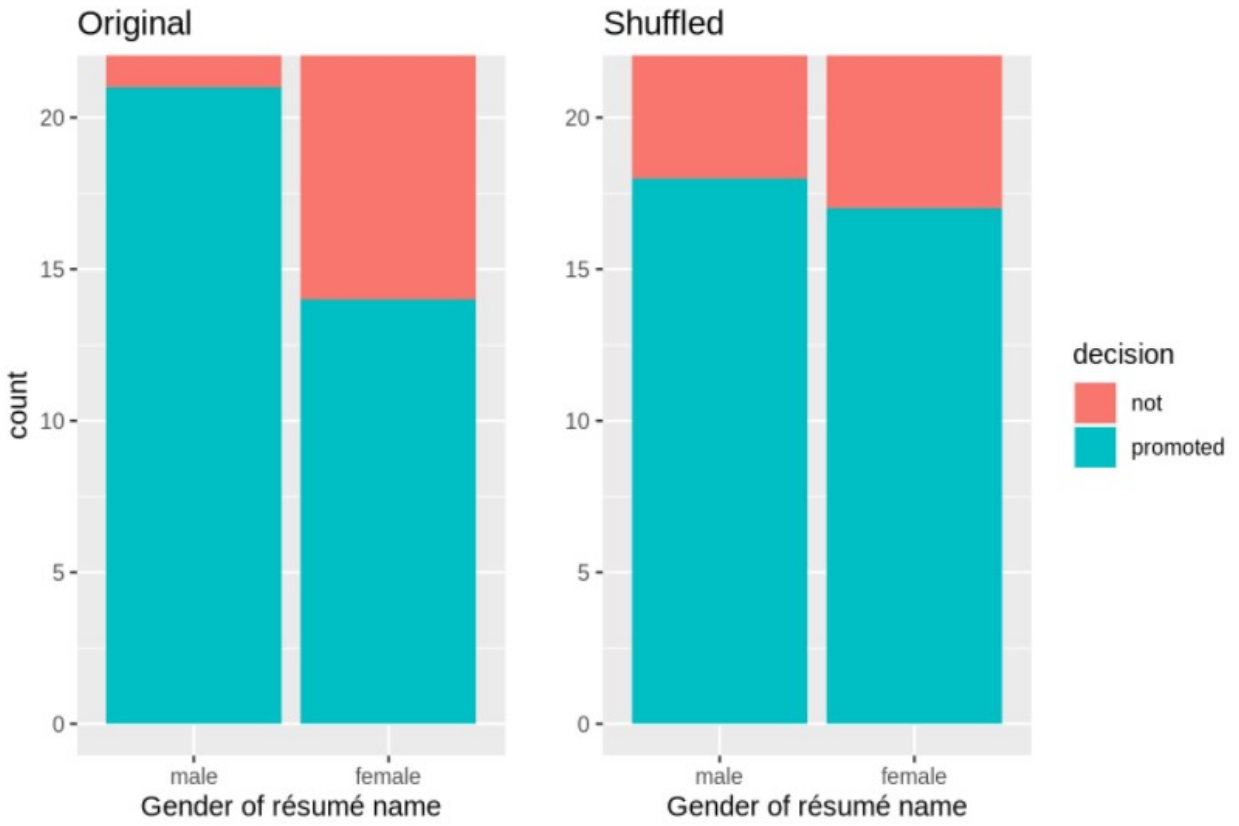


FIGURE 9.4: Barplots of relationship of promotion with gender (left) and shuffled gender (right).

Compared to the original data in the left barplot, the new "shuffled" data in the right barplot has promotion rates that are much more similar.

Let's also compute the proportion of résumés accepted for promotion for each group:

```
promotions_shuffled %>%
  group_by(gender, decision) %>%
  tally() # Same as summarize(n = n())
```

```
## # A tibble: 4 x 3
## # Groups:   gender [2]
##   gender decision     n
##   <fct>  <fct>    <int>
## 1 male   not          6
## 2 male   promoted    18
## 3 female not          7
## 4 female promoted    17
```

In this hypothetical universe of no discrimination, $18/24 = 0.75 = 75\%$ of males were promoted. By contrast, $17/24 = 0.78 = 70.8\%$ of females were promoted. It appears that male names were selected for promotion at a rate that was 0.75 - 0.708 = 0.042 = 4.2% different than females.

This difference in rates is not the same as the difference in rates of $0.292 = 29.2\%$ we originally observed. This is once again due to sampling variation.

1.1.3 Understanding hypothesis tests

1. A hypothesis is a statement about the value of an unknown population parameter.
2. A hypothesis test consists of a test between two competing hypotheses: (1) a null hypothesis H_0 vs (2) an alternative hypothesis H_1.
    - If as in the example we think men are promoted at a higher rate, we call such alternative hypotheses *one-sided alternatives*. If someone else however does not share such suspicions and only wants to investigate that there is a difference, whether higher or lower, they would set what is known as a *two-sided alternative*.
    - H_0 : men and women are promoted at the same rate **VERSUS** H_1 : men are promoted at a haigher rate than women.
    - Mathematical notation: $H_0 : p_m - p_f = 0$ vs $H_1 : p_m - p_f > 0$, if two-sided alternative: $p_m - p_f \neq 0$.
3. A **test statistic** is a point estimate/sample statistic formula used for hypothesis testing. Note that a sample statistic is merely a summary statistic based on a sample of observations. Here, the samples would be the $n_m = 24$ resumes with males and $n_f = 24$ resumes of female, so the point estimate is the difference in sample proportions $\hat{p}_m - \hat{p}_f$.
4. The observed **test statistic** is the value of the test statistic that we observed in real life. In our data, it was the observed difference of $\hat{p}_m - \hat{p}f = 0.875 - 0.583 = 0.292 = 29.2$ in favor of male resumes.
5. The **null distribution** is the sampling distribution of the test statistic assuming the null hypothesis H_0 is true.
    - Assuming the null hypothesis H_0, also stated as "Under H_0", how does the test statistic vary due to sampling variation? In our case, how will the difference in sample proportions $\hat{p}_m - \hat{p}_f$ vary due to sampling under H_0?.
6. The p-value is the probability of obtaining a test statistic just as extreme or more extreme than the observed test statistic assuming the null hypothesis H_0 is true.
7. In many hypothesis testing procedures, it is commonly recommended to set the **significance level** of the test beforehand (0.1, 0.01, and 0.05).

1.2 Conducting hypothesis tests

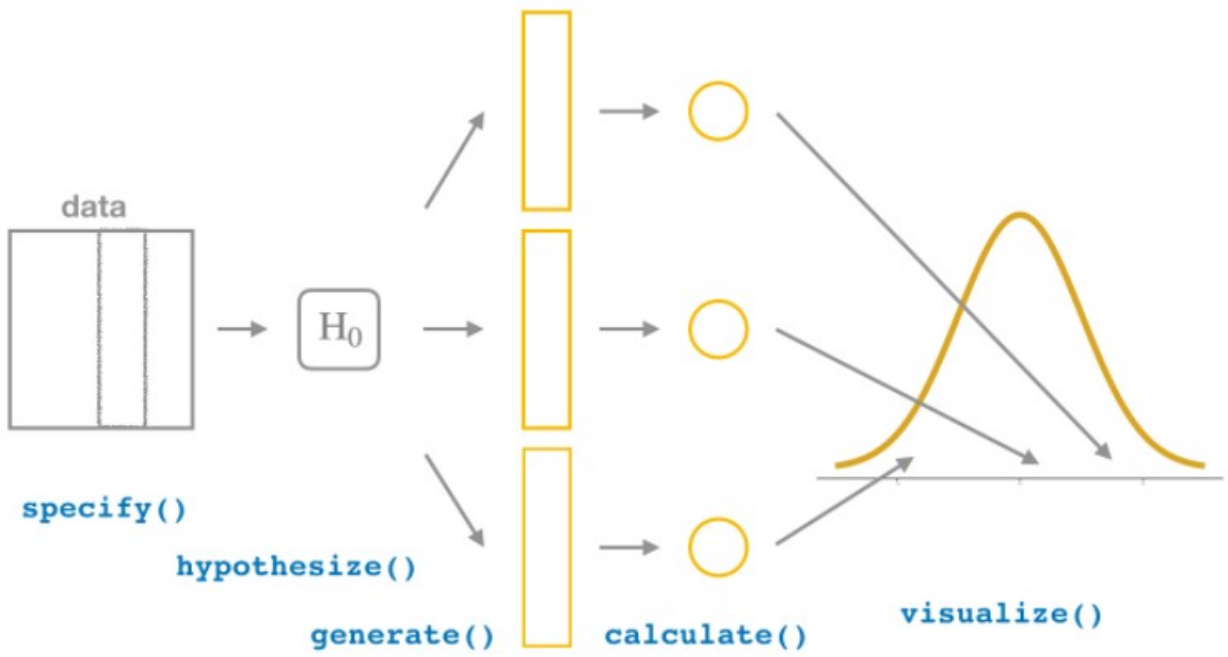


FIGURE 9.9: Hypothesis testing with the infer package.

We'll use a pre-specified significance level $\alpha = 0.05$ for this hypothesis test.

1.2.1 `infer` package workflow

1. `specify` variables

Recall that we use the `specify()` verb to specify the response variable and, if needed, any explanatory variables for our study. In this case, since we are interested in any potential effects of gender on promotion decisions, we set `decision` as the response variable and `gender` as the explanatory variable. We do so using `formula = response ~ explanatory` where `response` is the name of the response variable in the data frame and `explanatory` is the name of the explanatory variable. So in our case it is `decision ~ gender`.

Furthermore, since we are interested in the proportion of résumés `"promoted"`, and not the proportion of résumés `not` promoted, we set the argument `success` to `"promoted"`.

```
promotions %>%
  specify(formula = decision ~ gender, success = "promoted")
```

```
Response: decision (factor)
Explanatory: gender (factor)
# A tibble: 48 x 2
   decision gender
   <fct>    <fct>
 1 promoted male
 2 promoted male
 3 promoted male
 4 promoted male
 5 promoted male
 6 promoted male
 7 promoted male
 8 promoted male
 9 promoted male
```

Notice how the `promotions` data itself doesn't change, but the `Response: decision (factor)` and `Explanatory: gender (factor)` meta-data do.

2. `hypothesize` the null

In order to conduct hypothesis tests using the `infer` workflow, we need a new step not present for confidence intervals: `hypothesize()`. Recall from Section 9.2 that our hypothesis test was

$$H_0 : p_m - p_f = 0$$
$$\text{vs. } H_A : p_m - p_f > 0$$

In other words, the null hypothesis H_0 corresponding to our "hypothesized universe" stated that there was no difference in gender-based discrimination rates. We set this null hypothesis H_0 in our `infer` workflow using the `null` argument of the `hypothesize()` function to either:

- `"point"` for hypotheses involving a single sample or
- `"independence"` for hypotheses involving two samples.

In our case, since we have two samples (the résumés with "male" and "female" names), we set `null = "independence"`.

```r
promotions %>%
  specify(formula = decision ~ gender, success = "promoted") %>%
  hypothesize(null = "independence")
```

```
Response: decision (factor)
Explanatory: gender (factor)
Null Hypothesis: independence
# A tibble: 48 x 2
   decision gender
   <fct>    <fct>
 1 promoted male
 2 promoted male
 3 promoted male
 4 promoted male
 5 promoted male
 6 promoted male
 7 promoted male
 8 promoted male
 9 promoted male
10 promoted male
# … with 38 more rows
```

Again, the data has not changed yet. This will occur at the upcoming `generate()` step.

Where do the terms `"point"` and `"independence"` come from? These are two technical statistical terms. The term "point" relates from the fact that for a single group of observations, you will test the value of a single point. Going back to the pennies example from Chapter 8, say we wanted to test if the mean year of all US pennies was equal to 1993 or not. We would be testing the value of a "point" μ, the mean year of *all* US pennies, as follows

$$H_0 : \mu = 1993$$
$$\text{vs } H_A : \mu \neq 1993$$

The term "independence" relates to the fact that for two groups of observations, you are testing whether or not the response variable is *independent* of the explanatory variable that assigns the groups. In our case, we are testing whether the `decision` response variable is "independent" of the explanatory variable `gender` that assigns each résumé to either of the two groups.

3. `generate` replicates

After we `hypothesize()` the null hypothesis, we `generate()` replicates of "shuffled" datasets assuming the null hypothesis is true. We do this by repeating the shuffling exercise you performed in Section 9.1 several times. Instead of merely doing it 16 times as our groups of friends did, let's use the computer to repeat this 1000 times by setting `reps = 1000` in the `generate()` function. However, unlike for confidence intervals where we generated replicates using `type = "bootstrap"` resampling with replacement, we'll now perform shuffles/permutations by setting `type = "permute"`. Recall that shuffles/permutations are a kind of resampling, but unlike the bootstrap method, they involve resampling *without* replacement.

```
promotions_generate <- promotions %>%
  specify(formula = decision ~ gender, success = "promoted") %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute")
nrow(promotions_generate)
```

```
[1] 48000
```

Observe that the resulting data frame has 48,000 rows. This is because we performed shuffles/permutations for each of the 48 rows 1000 times and $48,000 = 1000 \cdot 48$. If you explore the `promotions_generate` data frame with `View()`, you'll notice that the variable `replicate` indicates which resample each row belongs to. So it has the value `1` 48 times, the value `2` 48 times, all the way through to the value `1000` 48 times.

4. `calculate` summary statistics

Now that we have generated 1000 replicates of "shuffles" assuming the null hypothesis is true, let's `calculate()` the appropriate summary statistic for each of our 1000 shuffles. From Section 9.2, point estimates related to hypothesis testing have a specific name: *test statistics*. Since the unknown population parameter of interest is the difference in population proportions $p_m - p_f$, the test statistic here is the difference in sample proportions $\hat{p}_m - \hat{p}_f$.

For each of our 1000 shuffles, we can calculate this test statistic by setting `stat = "diff in props"`. Furthermore, since we are interested in $\hat{p}_m - \hat{p}_f$ we set `order = c("male", "female")`. As we stated earlier, the order of the subtraction does not matter, so long as you stay consistent throughout your analysis and tailor your interpretations accordingly.

Let's save the result in a data frame called `null_distribution`:

```r
null_distribution <- promotions %>%
  specify(formula = decision ~ gender, success = "promoted") %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in props", order = c("male", "female"))

head(null_distribution, 10)
```

```
## # A tibble: 10 x 2
##    replicate    stat
##        <int>   <dbl>
##  1         1  0.125
##  2         2 -0.0417
##  3         3  0.0417
##  4         4  0.125
##  5         5 -0.0417
##  6         6  0.0417
##  7         7 -0.125
##  8         8  0.125
##  9         9  0.0417
## 10        10  0.208
```

Observe that we have 1000 values of stat, each representing one instance of $\hat{p}_m - \hat{p}_f$ in a hypothesized world of no gender discrimination. What was the observed difference in promotion rates? In other words, what was the observed test statistic $\hat{p}_m - \hat{p}_f$?

```r
# Let's save this in obs_diff_prop:
obs_diff_prop <- promotions %>%
  specify(decision ~ gender, success = "promoted") %>%
  calculate(stat = "diff in props", order = c("male", "female"))

  head(obs_diff_prop, 10)
```

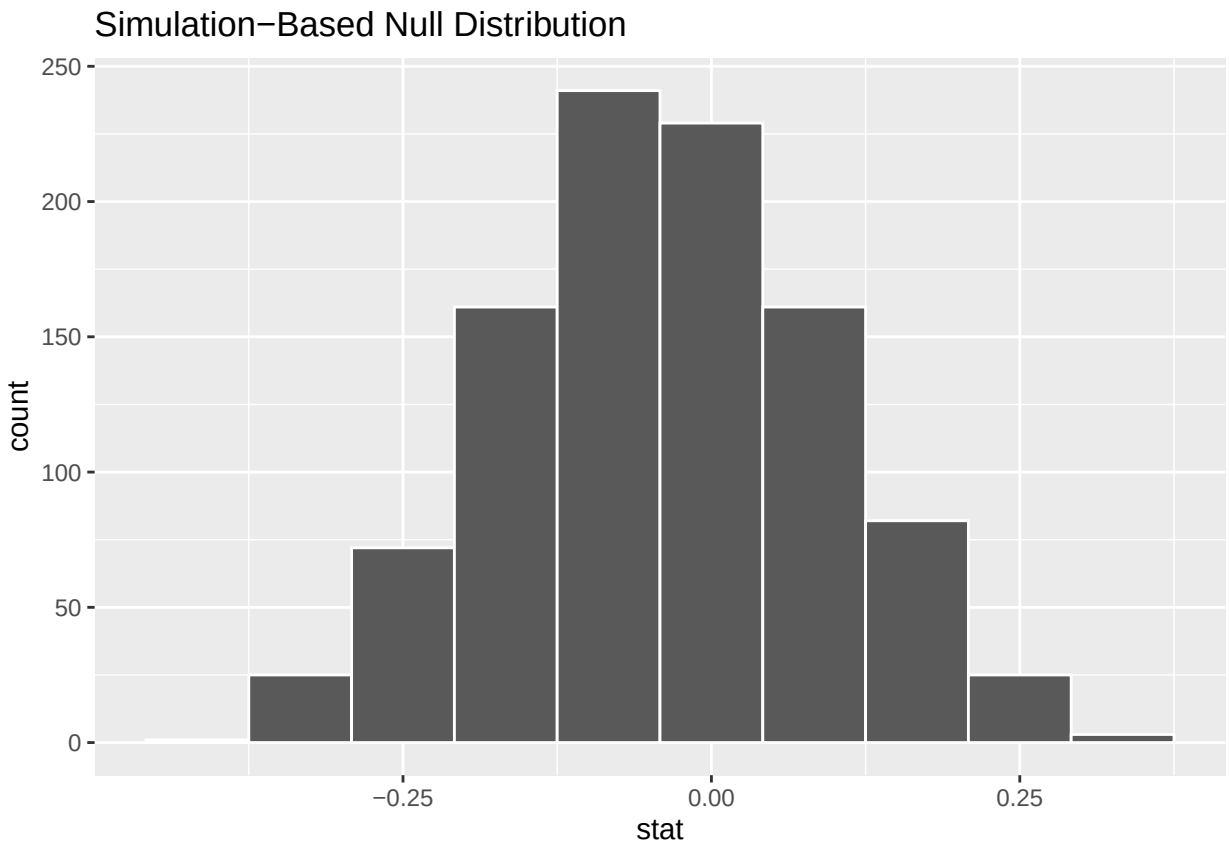
```
## # A tibble: 1 x 1
##    stat
##   <dbl>
## 1 0.292
```

5. `visualize` the p-value

The final step is to measure how surprised we are by a promotion difference of 29.2% in a hypothesized universe of no gender discrimination. If the observed difference of 0.292 is highly unlikely, then we would be inclined to reject the validity of our hypothesized universe.

We start by visualizing the *null distribution* of our 1000 values of $\hat{p}_m - \hat{p}_f$ using `visualize()` in Figure 9.10. Recall that these are values of the difference in promotion rates assuming H_0 is true. This corresponds to being in our hypothesized universe of no gender discrimination.

```
visualize(null_distribution, bins = 10)
```

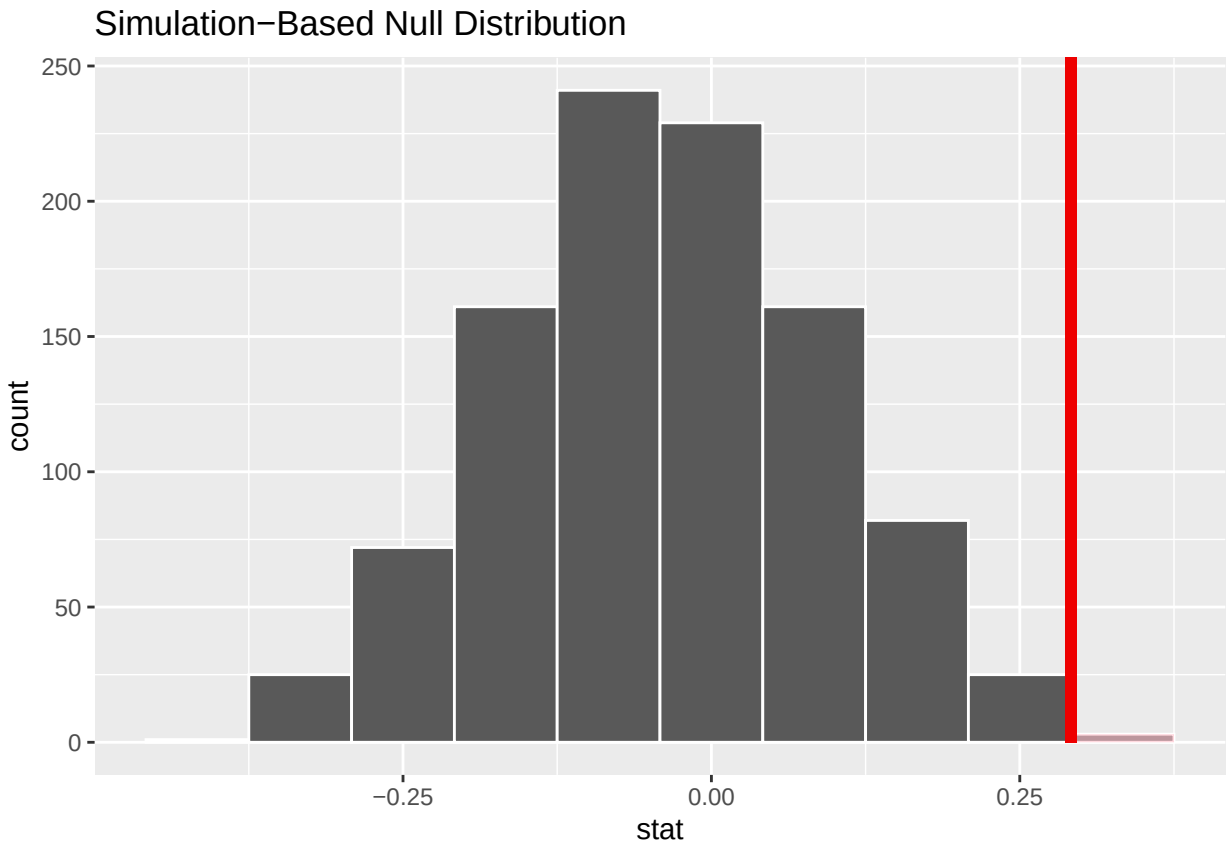


Let's now add what happened in real life to Figure 9.10, the observed difference in promotion rates of 0.875 - 0.583 = 0.292 = 29.2%. However, instead of merely adding a vertical line using `geom_vline()`, let's use the `shade_p_value()` function with `obs_stat` set to the observed test statistic value we saved in `obs_diff_prop`.

Furthermore, we'll set the `direction = "right"` reflecting our alternative hypothesis $H_A : p_m - p_f > 0$. Recall our alternative hypothesis H_A is that $p_m - p_f > 0$, stating that there is a difference in promotion rates in favor of résumés with male names. "More extreme" here corresponds to differences that are "bigger" or "more positive" or "more to the right." Hence we set the `direction` argument of `shade_p_value()` to be `"right"`.

On the other hand, had our alternative hypothesis H_A been the other possible one-sided alternative $p_m - p_f < 0$, suggesting discrimination in favor of résumés with female names, we would've set `direction = "left"`. Had our alternative hypothesis H_A been two-sided $p_m - p_f \neq 0$, suggesting discrimination in either direction, we would've set `direction = "both"`.

```
visualize(null_distribution, bins = 10) +
  shade_p_value(obs_stat = obs_diff_prop, direction = "right")
```



The solid dark line marks 0.292 = 29.2%. However, what does the shaded-region correspond to? This is the *p*-value:

> A *p*-value is the probability of obtaining a test statistic just as or more extreme than the observed test statistic assuming the null hypothesis H_0 is true.

So judging by the shaded region in Figure 9.11, it seems we would somewhat rarely observe differences in promotion rates of 0.292 = 29.2% or more in a hypothesized universe of no gender discrimination. In other words, the *p*-value is somewhat small. Hence, we would be inclined to reject this hypothesized universe, or using statistical language we would "reject H_0".

What fraction of the null distribution is shaded? In other words, what is the exact value of the *p*-value?

```
# use get_p_value()
null_distribution %>%
  get_p_value(obs_stat = obs_diff_prop, direction = "right")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.028
```

Keeping the definition of a *p*-value in mind, the probability of observing a difference in promotion rates as large as 0.292 = 29.2% due to sampling variation alone in the null distribution

is 0.027 = 2.7%. Since this *p*-value is smaller than our pre-specified significance level = 0.05, we reject the null hypothesis $H_0 : p_m - p_f = 0$. In other words, this *p*-value is sufficiently small to reject our hypothesized universe of no gender discrimination. We instead have enough evidence to change our mind in favor of gender discrimination being a likely culprit here.

1.2.2 Comparison with confidence intervals

To create the bootstrap distribution needed to construct a 95% confidence interval to our problem, we only need to make two changes. First, we remove the `hypothesize()` step. Second, we switch the `type` of resampling in the `generate()` step to be `"bootstrap"` instead of `"permute"`.

```
bootstrap_distribution <- promotions %>%
  specify(formula = decision ~ gender, success = "promoted") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "diff in props", order = c("male", "female"))
```

Now let's compute the percentile-based confidence intervals.

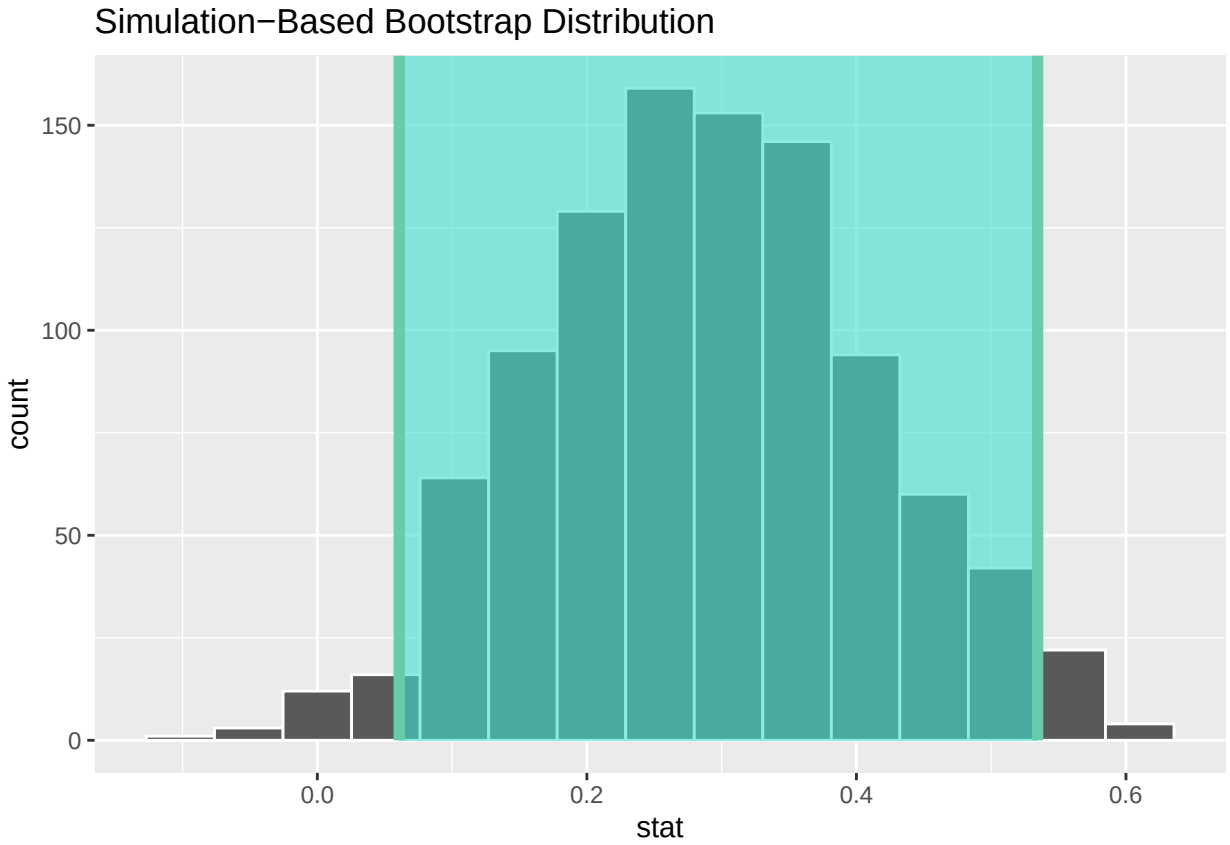
```
percentile_ci <- bootstrap_distribution %>%
  get_confidence_interval(level = 0.95, type = "percentile")

percentile_ci
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1   0.0608    0.534
```

We are 95% "confident" that the true difference in population proportions $p_m - p_f$ is between (0.034, 0.524).

```
# visualize it
visualize(bootstrap_distribution) +
  shade_confidence_interval(endpoints = percentile_ci)
```



Notice a key value that is not included in the 95% confidence interval for $p_m - p_f$: the value 0. In other words, a difference of 0 is not included in our net, suggesting that p_m and p_f are truly different! Furthermore, observe how the entirety of the 95% confidence interval for $p_m - p_f$ lies above 0, suggesting that this difference is in favor of men.

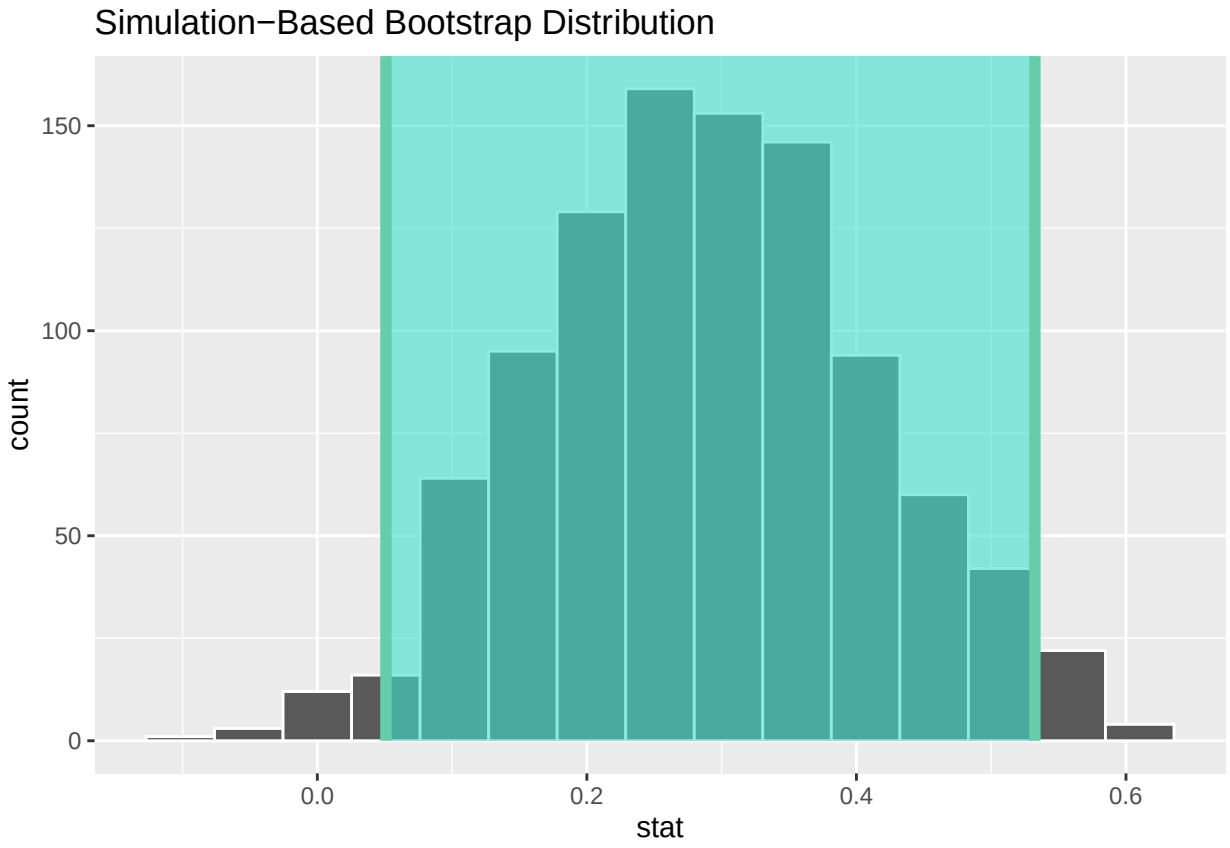
Since the bootstrap distribution appears to be roughly normally shaped, we can also use the standard error method as we did in Section 8.4. In this case, we must specify the `point_estimate` argument as the observed difference in promotion rates 0.292 = 29.2% saved in `obs_diff_prop`. This value acts as the center of the confidence interval.

```
se_ci <- bootstrap_distribution %>%
  get_confidence_interval(level = 0.95, type = "se",
                          point_estimate = obs_diff_prop)

se_ci
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1   0.0509    0.532
```

```r
# visualize it
visualize(bootstrap_distribution) +
  shade_confidence_interval(endpoints = se_ci)
```



1.2.3 "There is only one test"

Let's recap the steps:

1. `specify()` the variables of interest in your data frame.
2. `hypothesize()` the null hypothesis H_0. In other words, set a "model for the universe" assuming H_0 is true.
3. `generate()` shuffles assuming H_0 is true. In other words, simulate data assuming H_0 is true.
4. `calculate()` the test statistic of interest, both for the observed data and your simulated data.
5. `visualize()` the resulting null distribution and compute the p-value by comparing the null distribution to the observed test statistic.

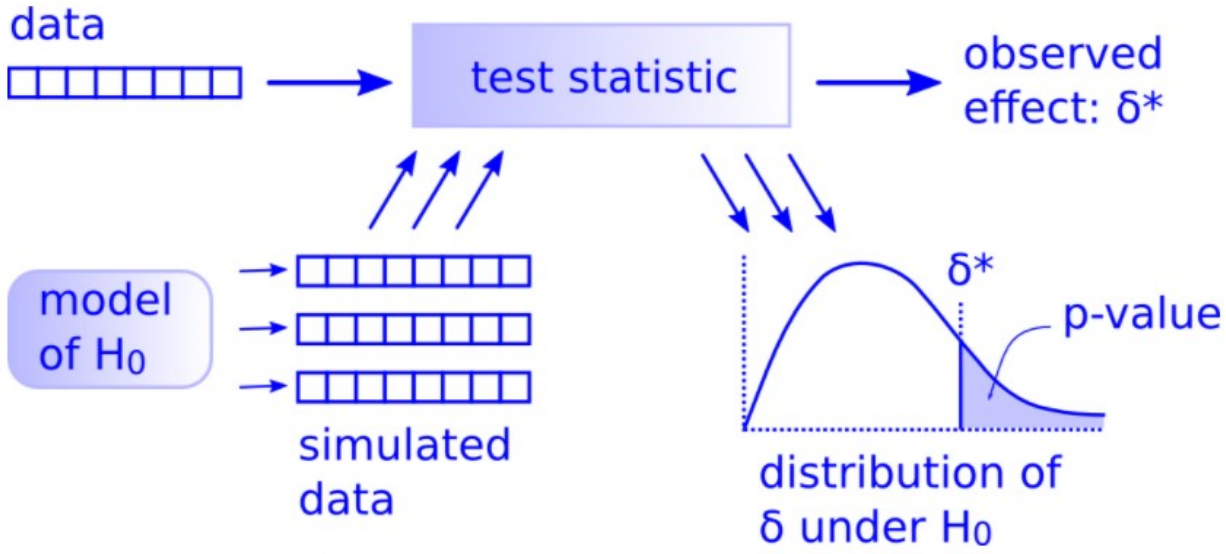


FIGURE 9.14: Allen Downey's hypothesis testing framework.