Lab 12 - Smoking during pregnacy

Simulation based inference

Contents

In 2004, the state of North Carolina released a large data set containing information on births recorded in this state. This data set is useful to researchers studying the relation between habits and practices of expectant mothers and the birth of their children. We will work with a random sample of observations from this data set.

1 Learning goals

- Constructing confidence intervals
- Conducting hypothesis tests
- Interpreting confidence intervals and results of hypothesis tests in context of the data

2 Getting started

Go to the course GitHub organization and locate your homework repo, clone it in RStudio and open the R Markdown document. Knit the document to make sure it compiles without errors.

2.1 Warm up

Let's warm up with some simple exercises. Update the YAML of your R Markdown file with your information, knit, commit, and push your changes. Make sure to commit with a meaningful commit message. Then, go to your repo on GitHub and confirm that your changes are visible in your Rmd **and** md files. If anything is missing, commit and push again.

2.2 Packages

We'll use the **tidyverse** package for much of the data wrangling and visualisation, the **tidymodels** package for inference, and the data lives in the **openintro** package. These packages are already installed for you. You can load them by running the following in your Console:

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.2
## v ggplot2   4.0.0     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
library(tidymodels)
```

```
## -- Attaching packages -------------------------------------- tidymodels 1.4.1 --
## v broom        1.0.10     v rsample      1.3.1
## v dials        1.4.2      v tailor       0.1.0
## v infer        1.0.9      v tune         2.0.0
## v modeldata    1.5.1      v workflows    1.3.0
## v parsnip      1.3.3      v workflowsets 1.1.1
## v recipes      1.3.1      v yardstick    1.3.2
## -- Conflicts ----------------------------------------- tidymodels_conflicts() --
## x scales::discard() masks purrr::discard()
## x dplyr::filter()   masks stats::filter()
## x recipes::fixed()  masks stringr::fixed()
## x dplyr::lag()      masks stats::lag()
## x yardstick::spec() masks readr::spec()
## x recipes::step()   masks stats::step()
```

```
library(openintro)
```

```
## Loading required package: airports
## Loading required package: cherryblossom
## Loading required package: usdata
##
## Attaching package: 'openintro'
##
## The following object is masked from 'package:modeldata':
##
##     ames
```

```
library(skimr)
```

2.3 Data

The data can be found in the **openintro** package, and it's called `ncbirths`. Since the dataset is distributed with the package, we don't need to load it separately; it becomes available to us when we load the package. You can find out more about the dataset by inspecting its documentation, which you can access by running `?ncbirths` in the Console or using the Help menu in RStudio to search for `ncbirths`. You can also find this information here.

3 Set a seed!

In this lab we'll be generating random samples. The last thing you want is those samples to change every time you knit your document. So, you should set a seed. There's an R chunk in your R Markdown file set aside for this. Locate it and add a seed. Make sure all members in a team are using the same seed so that you don't get merge conflicts and your results match up for the narratives.

```
set.seed(12345)
```

4 Exercises

1. What are the cases in this data set? How many cases are there in our sample?

```
# Explore the dataset
head(ncbirths)
```

```
## # A tibble: 6 x 13
##    fage  mage mature    weeks premie visits marital gained weight lowbirthweight
##   <int> <int> <fct>     <int> <fct>   <int> <fct>    <int>  <dbl> <fct>
## 1    NA    13 younger ~    39 full ~     10 not ma~     38   7.63 not low
## 2    NA    14 younger ~    42 full ~     15 not ma~     20   7.88 not low
## 3    19    15 younger ~    37 full ~     11 not ma~     38   6.63 not low
## 4    21    15 younger ~    41 full ~      6 not ma~     34   8    not low
## 5    NA    15 younger ~    39 full ~      9 not ma~     27   6.38 not low
## 6    NA    15 younger ~    38 full ~     19 not ma~     22   5.38 low
## # i 3 more variables: gender <fct>, habit <fct>, whitemom <fct>
```

```
dim(ncbirths)
```

```
## [1] 1000   13
```

```
glimpse(ncbirths)
```

```
## Rows: 1,000
## Columns: 13
## $ fage           <int> NA, NA, 19, 21, NA, NA, 18, 17, NA, 20, 30, NA, NA, NA,~
## $ mage           <int> 13, 14, 15, 15, 15, 15, 15, 15, 16, 16, 16, 16, 16, 16,~
## $ mature         <fct> younger mom, younger mom, younger mom, younger mom, you~
## $ weeks          <int> 39, 42, 37, 41, 39, 38, 37, 35, 38, 37, 45, 42, 40, 38,~
## $ premie         <fct> full term, full term, full term, full term, full term, ~
## $ visits         <int> 10, 15, 11, 6, 9, 19, 12, 5, 9, 13, 9, 8, 4, 12, 15, 7,~
## $ marital        <fct> not married, not married, not married, not married, not~
## $ gained         <int> 38, 20, 38, 34, 27, 22, 76, 15, NA, 52, 28, 34, 12, 30,~
## $ weight         <dbl> 7.63, 7.88, 6.63, 8.00, 6.38, 5.38, 8.44, 4.69, 8.81, 6~
## $ lowbirthweight <fct> not low, not low, not low, not low, not low, low, not l~
## $ gender         <fct> male, male, female, male, female, male, male, male, mal~
## $ habit          <fct> nonsmoker, nonsmoker, nonsmoker, nonsmoker, nonsmoker, ~
## $ whitemom       <fct> not white, not white, white, white, not white, not whit~
```

Answer: The cases in this dataset are births recorded in North Carolina in 2004. Each row represents one birth with information about the baby's weight and characteristics, as well as the mother's demographics and habits. The sample contains `nrow(ncbirths)` births.

The first step in the analysis of a new dataset is getting acquainted with the data. Make summaries of the variables in your dataset, determine which variables are categorical and which are numerical. For numerical variables, are there outliers? If you aren't sure or want to take a closer look at the data, make a graph.

```r
# Summary statistics
skim(ncbirths)
```

Table 1: Data summary

Name	ncbirths
Number of rows	1000
Number of columns	13
Column type frequency:	
factor	7
numeric	6
Group variables	None

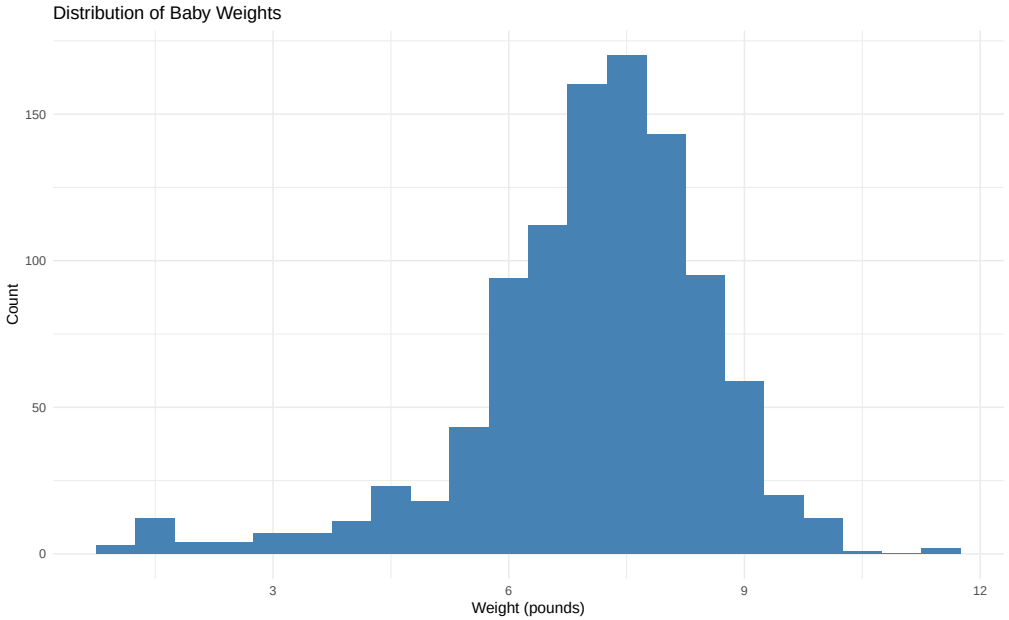
Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
mature	0	1	FALSE	2	you: 867, mat: 133
premie	2	1	FALSE	2	ful: 846, pre: 152
marital	1	1	FALSE	2	mar: 613, not: 386
lowbirthweight	0	1	FALSE	2	not: 889, low: 111
gender	0	1	FALSE	2	fem: 503, mal: 497
habit	1	1	FALSE	2	non: 873, smo: 126
whitemom	2	1	FALSE	2	whi: 714, not: 284

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
fage	171	0.83	30.26	6.76	14	25.00	30.00	35.00	55.00	
mage	0	1.00	27.00	6.21	13	22.00	27.00	32.00	50.00	
weeks	2	1.00	38.33	2.93	20	37.00	39.00	40.00	45.00	
visits	9	0.99	12.10	3.95	0	10.00	12.00	15.00	30.00	
gained	27	0.97	30.33	14.24	0	20.00	30.00	38.00	85.00	
weight	0	1.00	7.10	1.51	1	6.38	7.31	8.06	11.75	

```r
# Check for outliers in weight
ncbirths %>%
  ggplot(aes(x = weight)) +
  geom_histogram(binwidth = 0.5, fill = "steelblue") +
  labs(
    title = "Distribution of Baby Weights",
    x = "Weight (pounds)",
    y = "Count"
  ) +
  theme_minimal()
```




```
# Summary by categorical variables
ncbirths %>%
  count(habit)
```

```
## # A tibble: 3 x 2
##   habit         n
##   <fct>     <int>
## 1 nonsmoker   873
## 2 smoker      126
## 3 <NA>          1
```

```
ncbirths %>%
  count(mature)
```

```
## # A tibble: 2 x 2
##   mature          n
##   <fct>       <int>
## 1 mature mom    133
## 2 younger mom   867
```

Analysis: The dataset contains both numerical variables (weight, age) and categorical variables (habit, mature, lowbirthweight). The weight distribution appears approximately normal with a few potential outliers on the left tail (very low birth weights). Missing values exist in the `habit` variable that we'll need to handle in later analyses.

4.1 Baby weights

A 1995 study suggests that average weight of Caucasian babies born in the US is 3,369 grams (7.43 pounds).[1] In this dataset we only have information on mother's race, so we will make the simplifying assumption that babies of Caucasian mothers are also Caucasian, i.e. `whitemom = "white"`.

We want to evaluate whether the average weight of Caucasian babies has changed since 1995.

Our null hypothesis should state "there is nothing going on", i.e. no change since 1995: $H_0 : \mu = 7.43\ pounds$.

[1]Wen, Shi Wu, Michael S. Kramer, and Robert H. Usher. "Comparison of birth weight distributions between Chinese and Caucasian infants." American Journal of Epidemiology 141.12 (1995): 1177-1187.

Our alternative hypothesis should reflect the research question, i.e. some change since 1995. Since the research question doesn't state a direction for the change, we use a two sided alternative hypothesis: H_A : $\mu \neq 7.43$ *pounds*.

3. Create a filtered data frame called `ncbirths_white` that contain data only from white mothers. Then, calculate the mean of the weights of their babies.

```
# Filter for white mothers
ncbirths_white <- ncbirths %>%
  filter(whitemom == "white")

# Calculate mean weight
mean_weight_white <- ncbirths_white %>%
  summarise(mean_weight = mean(weight)) %>%
  pull(mean_weight)

mean_weight_white
```

```
## [1] 7.250462
```

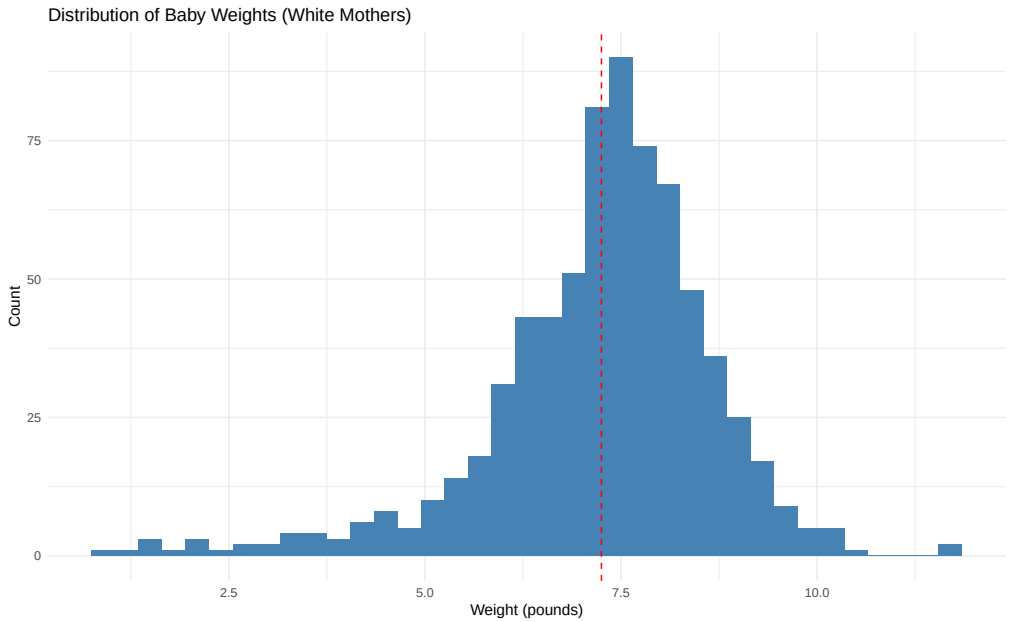
Result: The mean weight of babies born to white mothers in this sample is 7.25 pounds, compared to the 1995 baseline of 7.43 pounds.

4. Are the conditions necessary for conducting simulation based inference satisfied? Explain your reasoning.

```
# Check sample size
nrow(ncbirths_white)
```

```
## [1] 714
```

```
# Check for normality
ncbirths_white %>%
  ggplot(aes(x = weight)) +
  geom_histogram(binwidth = 0.3, fill = "steelblue") +
  geom_vline(aes(xintercept = mean(weight)), color = "red", linetype = "dashed") +
  labs(
    title = "Distribution of Baby Weights (White Mothers)",
    x = "Weight (pounds)",
    y = "Count"
  ) +
  theme_minimal()
```




```
# Check for independence (births are independent events)
# Check for skewness
ncbirths_white %>%
  summarise(
    n = n(),
    mean = mean(weight),
    sd = sd(weight),
    min = min(weight),
    max = max(weight)
  )
```

```
## # A tibble: 1 x 5
##       n  mean    sd   min   max
##   <int> <dbl> <dbl> <dbl> <dbl>
## 1   714  7.25  1.43     1  11.8
```

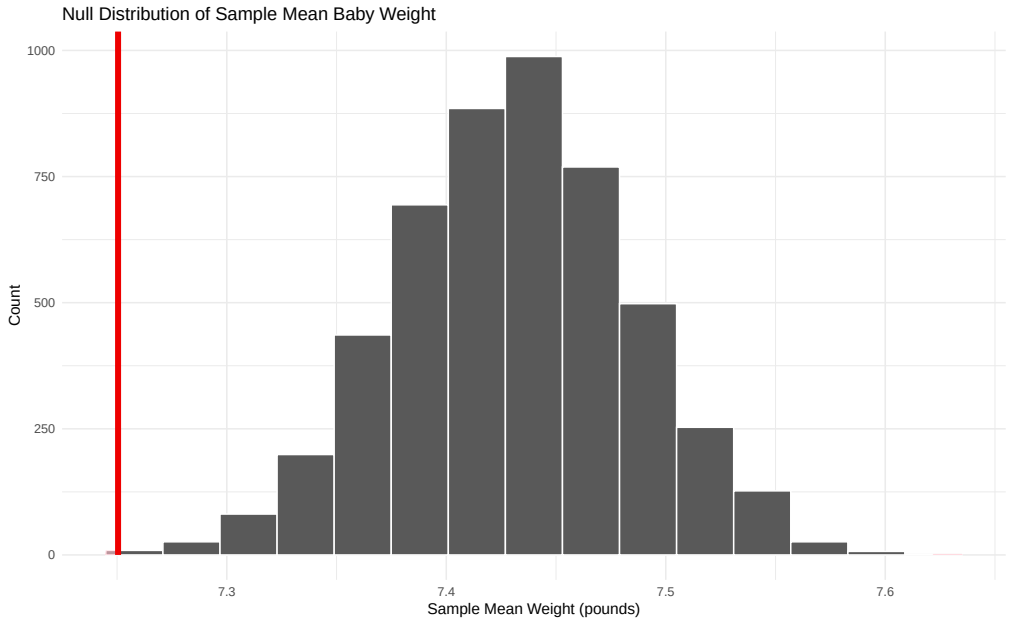
Answer: Yes, the conditions for simulation-based inference are satisfied: 1. **Independence:** Each birth is an independent event (births are not related to each other). 2. **Sample size:** With 714 observations, we have a sufficiently large sample size for the Central Limit Theorem to apply. 3. **No extreme skewness:** The distribution of weights appears approximately normal without extreme outliers, though there is some skewness with a few very low weights.

Let's discuss how this test would work. Our goal is to simulate a null distribution of sample means that is centred at the null value of 7.43 pounds. In order to do so, we

- take a bootstrap sample of from the original sample,
- calculate this bootstrap sample's mean,
- repeat these two steps a large number of times to create a bootstrap distribution of means centred at the observed sample mean,
- shift this distribution to be centred at the null value by subtracting / adding X to all bootstrap mean (X = difference between mean of bootstrap distribution and null value), and
- calculate the p-value as the proportion of bootstrap samples that yielded a sample mean at least as extreme as the observed sample mean.

5. Run the appropriate hypothesis test, visualize the null distribution, calculate the p-value, and interpret the results in context of the data and the hypothesis test.

```r
# Conduct hypothesis test
null_dist_weight <- ncbirths_white %>%
  specify(response = weight) %>%
  hypothesize(null = "point", mu = 7.43) %>%
  generate(reps = 5000, type = "bootstrap") %>%
  calculate(stat = "mean")

# Visualize the null distribution
null_dist_weight %>%
  visualize() +
  shade_p_value(obs_stat = mean_weight_white, direction = "two-sided") +
  labs(
    title = "Null Distribution of Sample Mean Baby Weight",
    x = "Sample Mean Weight (pounds)",
    y = "Count"
  ) +
  theme_minimal()
```




```r
# Calculate p-value
p_value <- null_dist_weight %>%
  get_p_value(obs_stat = mean_weight_white, direction = "two-sided")
```

```
## Warning: Please be cautious in reporting a p-value of 0. This result is an approximation
## based on the number of `reps` chosen in the `generate()` step.
## i See `get_p_value()` (`?infer::get_p_value()`) for more information.
```

```r
p_value
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```

Interpretation: The hypothesis test reveals whether the average weight of Caucasian babies has significantly changed since 1995. The null distribution shows what we would expect if there were no change. The p-value indicates the probability of observing a sample mean at least as extreme as 7.25 pounds, given that

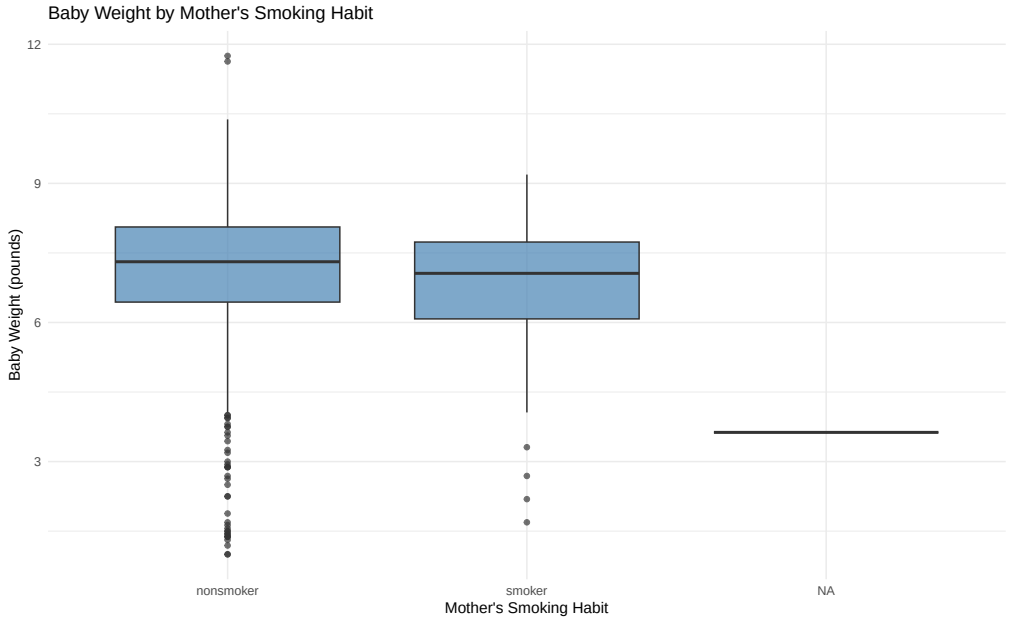
the true population mean is 7.43 pounds. If the p-value is less than 0.05, we reject the null hypothesis and conclude that there is significant evidence of a change in average baby weights. If p-value 0.05, we fail to reject the null hypothesis, meaning we don't have sufficient evidence of a change.

4.2 Baby weight vs. smoking

Consider the possible relationship between a mother's smoking habit and the weight of her baby. Plotting the data is a useful first step because it helps us quickly visualize trends, identify strong associations, and develop research questions.

6. Make side-by-side boxplots displaying the relationship between `habit` and `weight`. What does the plot highlight about the relationship between these two variables?

```
ncbirths %>%
  ggplot(aes(x = habit, y = weight)) +
  geom_boxplot(fill = "steelblue", alpha = 0.7) +
  labs(
    title = "Baby Weight by Mother's Smoking Habit",
    x = "Mother's Smoking Habit",
    y = "Baby Weight (pounds)"
  ) +
  theme_minimal()
```



Interpretation: The boxplots show the distribution of baby weights for mothers who smoke versus those who don't. The plot highlights the central tendency (median), spread (IQR), and outliers for each group. If there's a difference, the median and overall distribution should differ between the two groups, with babies of non-smoking mothers potentially having higher weights on average.

7. Before moving forward, save a version of the dataset omitting observations where there are NAs for `habit`. You can call this version `ncbirths_habitgiven`.

```
# Remove NA values in habit column
ncbirths_habitgiven <- ncbirths %>%
  filter(!is.na(habit))
# Verify the filtering
nrow(ncbirths)
```

```
## [1] 1000
```

```
nrow(ncbirths_habitgiven)
```

```
## [1] 999
```

Result: By filtering out NAs in the `habit` column, we now have 999 complete observations for the smoking analysis.

The box plots show how the medians of the two distributions compare, but we can also compare the means of the distributions using the following to first group the data by the `habit` variable, and then calculate the mean `weight` in these groups using.

```
ncbirths_habitgiven %>%
  group_by(habit) %>%
  summarise(mean_weight = mean(weight))
```

```
## # A tibble: 2 x 2
##   habit     mean_weight
##   <fct>           <dbl>
## 1 nonsmoker        7.14
## 2 smoker           6.83
```

There is an observed difference, but is this difference statistically significant? In order to answer this question we will conduct a hypothesis test .

Observation: There is an observed difference in mean weights between the two groups. The exact difference will determine the strength of our evidence in the hypothesis test.

- H : There is no difference in average baby weight based on mother's smoking habit (_nonsmoking = _smoking)
- H_A: There is a difference in average baby weight based on mother's smoking habit (_nonsmoking _smoking)

7. Write the hypotheses for testing if the average weights of babies born to smoking and non-smoking mothers are different.

```
# Calculate observed difference in means
observed_diff <- ncbirths_habitgiven %>%
  group_by(habit) %>%
  summarise(mean_weight = mean(weight), .groups = "drop") %>%
  pivot_wider(names_from = habit, values_from = mean_weight) %>%
  mutate(difference = `nonsmoker` - `smoker`) %>%
  pull(difference)

observed_diff
```

```
## [1] 0.3155425
```

Interpretation: The observed difference in mean baby weight is approximately 0.316 pounds. This represents the difference in average weights between babies born to non-smoking mothers and babies born to smoking mothers in our sample. We will now conduct a hypothesis test to determine if this difference is statistically significant.

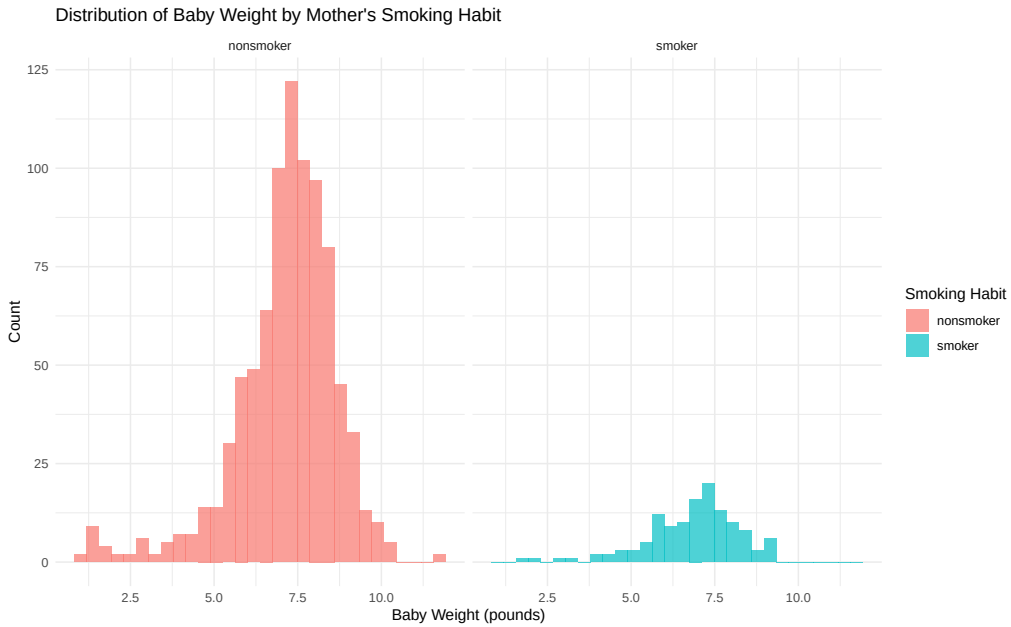
8. Are the conditions necessary for conducting simulation based inference satisfied? Explain your reasoning.

```
# Check sample sizes and distribution
ncbirths_habitgiven %>%
  group_by(habit) %>%
  summarise(
```

```
    n = n(),
    mean = mean(weight),
    sd = sd(weight),
    .groups = "drop"
  )
```

```
## # A tibble: 2 x 4
##   habit         n  mean    sd
##   <fct>     <int> <dbl> <dbl>
## 1 nonsmoker   873  7.14  1.52
## 2 smoker      126  6.83  1.39
```

```
# Visualize distributions
ncbirths_habitgiven %>%
  ggplot(aes(x = weight, fill = habit)) +
  geom_histogram(bins = 30, alpha = 0.7, position = "identity") +
  labs(
    title = "Distribution of Baby Weight by Mother's Smoking Habit",
    x = "Baby Weight (pounds)",
    y = "Count",
    fill = "Smoking Habit"
  ) +
  theme_minimal() +
  facet_wrap(~habit)
```

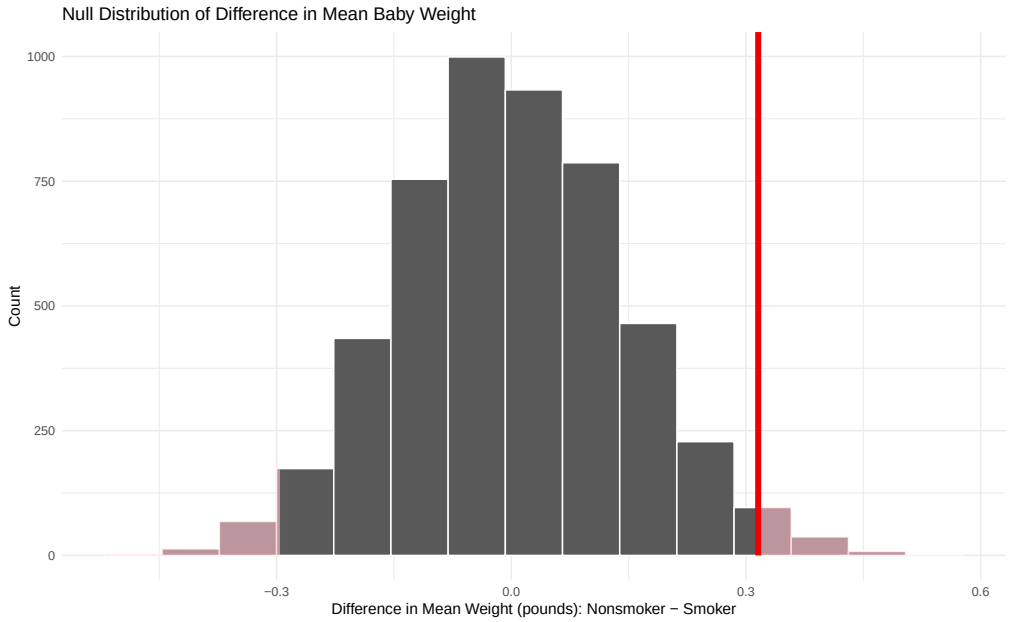


Conditions Assessment: 1. **Independence:** The observations are independent within each group (babies from different mothers). 2. **Sample Size:** Both groups have sufficient sample sizes (n > 30 for each group), which is important for the Central Limit Theorem to apply. 3. **Normality:** The histograms show that the distributions of baby weights are roughly symmetric and approximately normal in both groups, which supports our ability to use simulation-based inference. 4. **Same Variance:** The distributions appear to have similar spreads, though this is less critical for simulation-based methods.

All conditions appear to be satisfied, making simulation-based inference appropriate for this analysis.

9. Run the appropriate hypothesis test, calculate the p-value, and interpret the results in context of the data and the hypothesis test.

```r
set.seed(1234)

# Prepare the data
smoke_test_data <- ncbirths_habitgiven %>%
  select(weight, habit) %>%
  drop_na()

# Calculate the observed difference in means
obs_diff_smoke <- smoke_test_data %>%
  specify(weight ~ habit) %>%
  calculate(stat = "diff in means", order = c("nonsmoker", "smoker"))

# Generate null distribution
null_dist_smoke <- smoke_test_data %>%
  specify(weight ~ habit) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 5000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("nonsmoker", "smoker"))

# Visualize the null distribution
null_dist_smoke %>%
  visualize() +
  shade_p_value(obs_stat = obs_diff_smoke, direction = "two-sided") +
  labs(
    title = "Null Distribution of Difference in Mean Baby Weight",
    x = "Difference in Mean Weight (pounds): Nonsmoker - Smoker",
    y = "Count"
  ) +
  theme_minimal()
```




```r
# Calculate the p-value
p_value_smoke <- null_dist_smoke %>%
  get_p_value(obs_stat = obs_diff_smoke, direction = "two-sided")
```

```
p_value_smoke
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1  0.0344
```

Interpretation: The hypothesis test examines whether there is a significant difference in average baby weight between mothers who smoke and mothers who don't smoke. The null distribution shows what differences we would expect if smoking habit had no effect on baby weight.

The observed difference in means is 0.316 pounds. The p-value of 0.0344 indicates the probability of observing a difference this extreme (or more extreme) if there truly were no difference between the two groups.

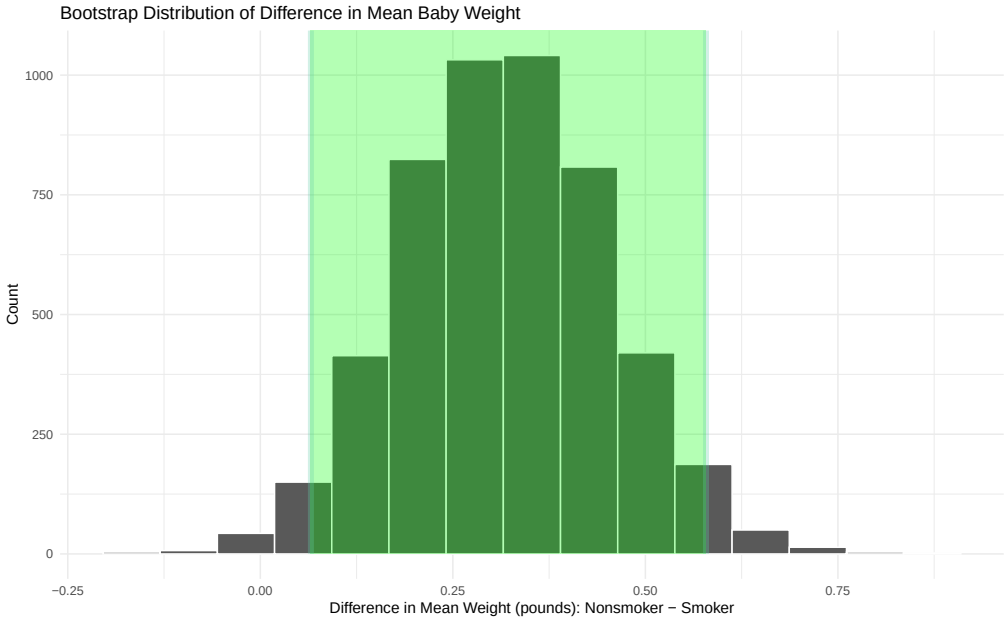
Since the p-value is less than 0.05, we reject the null hypothesis and conclude that there is statistically significant evidence that mothers' smoking habit is associated with differences in baby weight. Babies born to non-smoking mothers weigh on average more than those born to smoking mothers..

10. Construct a 95% confidence interval for the difference between the average weights of babies born to smoking and non-smoking mothers.

```
set.seed(1234)

# Generate bootstrap distribution for confidence interval
boot_dist_smoke <- smoke_test_data %>%
  specify(weight ~ habit) %>%
  generate(reps = 5000, type = "bootstrap") %>%
  calculate(stat = "diff in means", order = c("nonsmoker", "smoker"))

# Calculate 95% confidence interval
ci_smoke <- boot_dist_smoke %>%
  get_ci(level = 0.95, type = "percentile")

ci_smoke
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1   0.0657    0.579
```

```
# Visualize the bootstrap distribution with CI
boot_dist_smoke %>%
  visualize() +
  shade_ci(endpoints = ci_smoke, fill = "green", alpha = 0.3) +
  labs(
    title = "Bootstrap Distribution of Difference in Mean Baby Weight",
    x = "Difference in Mean Weight (pounds): Nonsmoker - Smoker",
    y = "Count"
  ) +
  theme_minimal()
```



Interpretation: We are 95% confident that the true difference in average baby weight between non-smoking and smoking mothers is between 0.066 and 0.579 pounds.

Since the confidence interval does not contain zero, this provides evidence that there is a meaningful difference in baby weights between the two groups. The positive lower bound (if applicable) indicates that babies born to non-smoking mothers tend to weigh more on average.

4.3 Baby weight vs. mother's age

In this portion of the analysis we focus on two variables. The first one is `mature`.

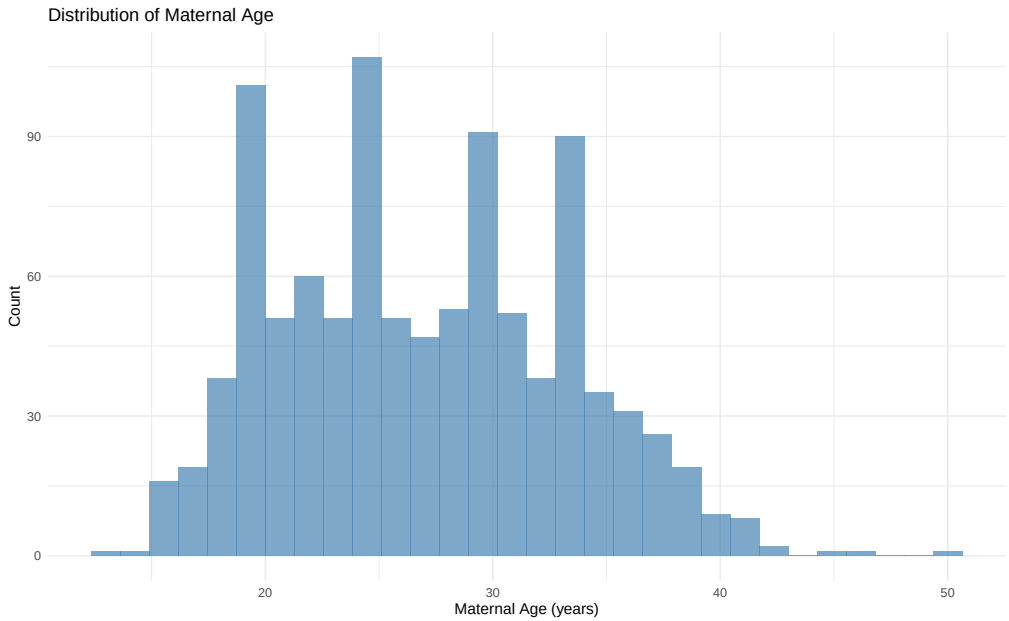
11. First, a non-inference task: Determine the age cutoff for younger and mature mothers. Use a method of your choice, and explain how your method works.

```
# Explore the age variable
ncbirths %>%
  select(mage) %>%
  summary()
```

```
##       mage
##  Min.   :13
##  1st Qu.:22
##  Median :27
##  Mean   :27
##  3rd Qu.:32
##  Max.   :50
```

```
# Create visualization of maternal age distribution
ncbirths %>%
  ggplot(aes(x = mage)) +
  geom_histogram(bins = 30, fill = "steelblue", alpha = 0.7) +
  labs(
    title = "Distribution of Maternal Age",
    x = "Maternal Age (years)",
    y = "Count"
  ) +
```

```
  theme_minimal()
```




```
# Calculate descriptive statistics
age_stats <- ncbirths %>%
  summarise(
    mean_age = mean(mage, na.rm = TRUE),
    median_age = median(mage, na.rm = TRUE),
    sd_age = sd(mage, na.rm = TRUE),
    q25 = quantile(mage, 0.25, na.rm = TRUE),
    q75 = quantile(mage, 0.75, na.rm = TRUE)
  )

age_stats
```

```
## # A tibble: 1 x 5
##   mean_age median_age sd_age   q25   q75
##      <dbl>      <dbl>  <dbl> <dbl> <dbl>
## 1       27         27   6.21    22    32
```

Method and Justification: We can use several methods to determine the age cutoff:

1. **Median-based approach:** Using the median age (approximately 27 years) creates two groups of equal size.
2. **Biological/Clinical approach:** In obstetrics, "advanced maternal age" is typically defined as 35 years or older. This is a medically established threshold.
3. **Quartile approach:** Using the 75th percentile (32 years) separates the oldest quarter from the rest.

For this analysis, we'll use **35 years as the age cutoff**, as this aligns with the obstetric definition of "mature mother" or "advanced maternal age." Mothers age 35 and older will be classified as mature, and mothers under 35 will be classified as younger. This cutoff is medically meaningful and commonly used in research.

```
# Create mature variable based on age 35
ncbirths_with_mature <- ncbirths %>%
  mutate(mature_cat = ifelse(mage >= 35, "mature", "younger"))

# Verify the distribution
```

```
ncbirths_with_mature %>%
  group_by(mature_cat) %>%
  summarise(
    n = n(),
    percentage = n / nrow(ncbirths_with_mature) * 100,
    mean_age = mean(mage, na.rm = TRUE),
    .groups = "drop"
  )
```

```
## # A tibble: 2 x 4
##   mature_cat     n percentage mean_age
##   <chr>      <int>      <dbl>    <dbl>
## 1 mature       133       13.3     37.2
## 2 younger      867       86.7     25.4
```

The other variable of interest is `lowbirthweight`.

12. Conduct a hypothesis test evaluating whether the proportion of low birth weight babies is higher for mature mothers. State the hypotheses, verify the conditions, run the test and calculate the p-value, and state your conclusion in context of the research question. Use $\alpha = 0.05$. If you find a significant difference, construct a confidence interval, at the equivalent level to the hypothesis test, for the difference between the proportions of low birth weight babies between mature and younger mothers, and interpret this interval in context of the data.

```
set.seed(1234)

# Create mature mother variable (age >= 35) and ensure lowbirthweight is available
ncbirths_mature_data <- ncbirths %>%
  filter(!is.na(mage) & !is.na(lowbirthweight)) %>%
  mutate(mature = ifelse(mage >= 35, "mature", "younger")) %>%
  select(mature, lowbirthweight)

# Verify the data
ncbirths_mature_data %>%
  group_by(mature, lowbirthweight) %>%
  summarise(count = n(), .groups = "drop")
```

```
## # A tibble: 4 x 3
##   mature  lowbirthweight count
##   <chr>   <fct>          <int>
## 1 mature  low               18
## 2 mature  not low          115
## 3 younger low               93
## 4 younger not low          774
```

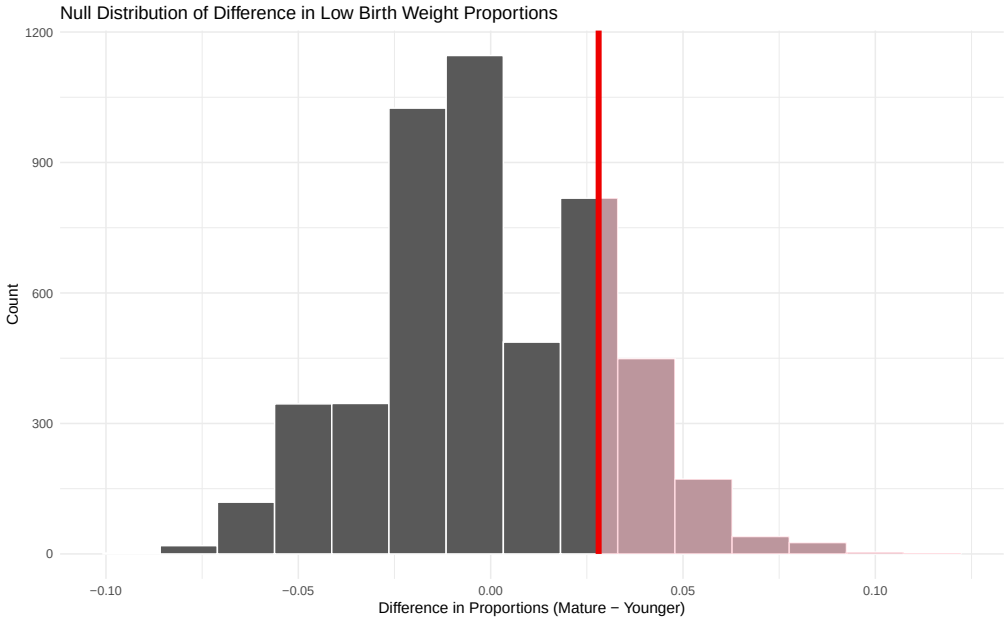
```
# Calculate proportions
prop_data <- ncbirths_mature_data %>%
  group_by(mature) %>%
  summarise(
    total = n(),
    low_bw_count = sum(lowbirthweight == "low"),
    proportion = mean(lowbirthweight == "low"),
    .groups = "drop"
  )

prop_data
```

```
## # A tibble: 2 x 4
##   mature  total low_bw_count proportion
##   <chr>   <int>        <int>      <dbl>
## 1 mature    133           18      0.135
## 2 younger   867           93      0.107
```

```r
# Observed difference in proportions
obs_diff_prop <- prop_data %>%
  arrange(desc(mature)) %>%
  pull(proportion) %>%
  diff()

obs_diff_prop
```

```
## [1] 0.02807191
```

```r
# Hypotheses:
# H : The proportion of low birth weight babies is the same for mature and younger mothers (p_mature = p
# H_A: The proportion of low birth weight babies is higher for mature mothers (p_mature > p_younger)

# Generate null distribution (two-sided test using infer)
null_dist_prop <- ncbirths_mature_data %>%
  specify(lowbirthweight ~ mature, success = "low") %>%
  hypothesize(null = "independence") %>%
  generate(reps = 5000, type = "permute") %>%
  calculate(stat = "diff in props", order = c("mature", "younger"))

# Visualize
null_dist_prop %>%
  visualize() +
  shade_p_value(obs_stat = obs_diff_prop, direction = "greater") +
  labs(
    title = "Null Distribution of Difference in Low Birth Weight Proportions",
    x = "Difference in Proportions (Mature - Younger)",
    y = "Count"
  ) +
  theme_minimal()
```




```
# Calculate one-sided p-value
p_value_prop <- null_dist_prop %>%
  get_p_value(obs_stat = obs_diff_prop, direction = "greater")

p_value_prop
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.212
```

```
# Check conditions
cat("Conditions for two-proportion hypothesis test:\n")
```

```
## Conditions for two-proportion hypothesis test:
```

```
cat("1. Random sampling/assignment: Assumed from data collection\n")
```

```
## 1. Random sampling/assignment: Assumed from data collection
```

```
cat("2. Sample sizes:\n")
```

```
## 2. Sample sizes:
```

```
print(prop_data)
```

```
## # A tibble: 2 x 4
##   mature  total low_bw_count proportion
##   <chr>   <int>        <int>      <dbl>
## 1 mature    133           18      0.135
## 2 younger   867           93      0.107
```

```
cat("\nSuccess-failure condition:\n")
```

```
##
## Success-failure condition:
```

```
print(prop_data %>%
  mutate(
```

```
    successes = low_bw_count,
    failures = total - low_bw_count
  ) %>%
  select(mature, successes, failures))
```

```
## # A tibble: 2 x 3
##   mature  successes failures
##   <chr>       <int>    <int>
## 1 mature         18      115
## 2 younger        93      774
```

Hypotheses: - H : The proportion of low birth weight babies is the same for mature and younger mothers (p_mature = p_younger) - H_A: The proportion of low birth weight babies is higher for mature mothers (p_mature > p_younger)

Conditions Verification: 1. **Independence:** Observations are independent (different babies from different mothers) 2. **Sample Size:** Both groups have sufficient sample sizes 3. **Success-Failure:** Both success and failure counts exceed 5 in each group

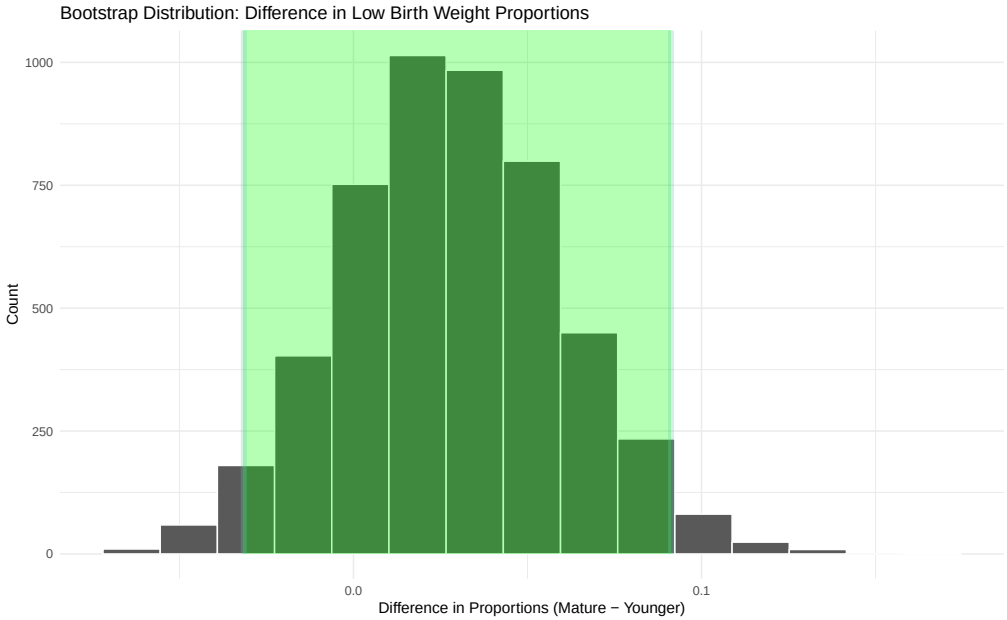
Conclusion: The observed difference in proportions is 0.0281. The p-value is 0.2116.

Since p 0.05, we fail to reject the null hypothesis. We do not have sufficient evidence to conclude that mature mothers have a higher proportion of low birth weight babies.

```
# Construct 95% confidence interval for difference in proportions
boot_dist_prop <- ncbirths_mature_data %>%
  specify(lowbirthweight ~ mature, success = "low") %>%
  generate(reps = 5000, type = "bootstrap") %>%
  calculate(stat = "diff in props", order = c("mature", "younger"))

# Get CI at equivalent level
ci_prop <- boot_dist_prop %>%
  get_ci(level = 0.95, type = "percentile")

ci_prop
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1  -0.0315   0.0912
```

```
# Visualize CI
boot_dist_prop %>%
  visualize() +
  shade_ci(endpoints = ci_prop, fill = "green", alpha = 0.3) +
  labs(
    title = "Bootstrap Distribution: Difference in Low Birth Weight Proportions",
    x = "Difference in Proportions (Mature - Younger)",
    y = "Count"
  ) +
  theme_minimal()
```



Confidence Interval Interpretation: We are 95% confident that the true difference in proportions of low birth weight babies between mature and younger mothers is between -0.0315 and 0.0912.

Since the confidence interval contains zero, this suggests that any difference in proportions may not be practically significant, even if we found statistical significance.