Final Report Group 35

August 14, 2023

1 Final Report: Which city in Georgia has the most expensive real estate: Atlanta or Savannah?

1.0.1 Introduction and Background Information

In STAT 201, we have learned the process of statistical inference. Statistical inference is a method involving the estimation of a parameter of a population based on the use of random sampling. The sample can be used to infer about the population given that it is random, representative, generalizable, and unbiased.

For this project, we will be using a Real Estate Georgia dataset with real estate information in Georgia for the first half of 2021 to perform statistical inference and compare two cities: Atlanta and Savannah. The dataset consists of variables such as updates on the listing, the price, the city, and the home type. The cost of living in Georgia, USA is an estimated 11% below the national average, with its housing cost being 23% below the national average. In spite of these state percentages, Atlanta, the state's most populated city, has a cost of living that is 14% higher than the state average and 2% higher than the national average. In comparison, Savannah—another well-populated city in Georgia—has a cost of living that is 1% higher than the state average and 10% lower than the national average. According to Georgia Realtors, the average real estate sales price in Georgia increased by 17.4% from the previous year which brings up the mean price to be $352,069. With this background information in mind, we want to determine which of the two cities has the higher average (mean) real estate value, and which city has the higher median real estate value.

1.0.2 Purpose

The purpose of our project is to infer the mean and median price within the time period of the first six months of 2021 for the population of all real estate in Atlanta and Savannah by using the sample data of listings in each city. Mean was chosen, as the primary goal of the project was to compare the average real estate value between the two cities. The median was then selected to account for the few very expensive listings for sale that may skew the mean. The result of this inference could be taken into consideration when choosing which city to buy real estate.

1.0.3 Methods and Results

[1]:
```
# loading packages

library(tidyverse)
library(broom)
```

```
library(repr)
library(digest)
library(infer)
library(gridExtra)
```

```
  Attaching packages                                   tidyverse
1.3.2
 ggplot2 3.3.6       purrr   0.3.4
 tibble  3.1.8       dplyr   1.0.10
 tidyr   1.2.1       stringr 1.4.1
 readr   2.1.2       forcats 0.5.2
  Conflicts
tidyverse_conflicts()
 dplyr::filter() masks stats::filter()
 dplyr::lag()    masks stats::lag()

Attaching package: ‘gridExtra’


The following object is masked from ‘package:dplyr’:

    combine
```

[2]:
```
# setting the seed
set.seed(2021)

# reading in the dataset
url <- "https://raw.githubusercontent.com/shauna06/STAT-201-Project-Rep/main/
 ↪RealEstate_Georgia.csv"
real_estate_georgia <- read_csv(url)
head(real_estate_georgia)
```

```
New names:
• `` -> `…1`
Rows: 13804 Columns: 39
  Column specification

Delimiter: ","
chr  (12): id, country, event, city, state, streetAddress,
description, curr…
dbl  (26): …1, stateId, countyId, cityId, is_bankOwned,
is_forAuction, tim…
date  (1): datePostedString

 Use `spec()` to retrieve the full column specification for this
```

```
data.
 Specify the column types or set `show_col_types = FALSE` to quiet
this message.
```

A tibble: 6 × 39

...1 <dbl>	id <chr>	stateId <dbl>	countyId <dbl>	cityId <dbl>	country <chr>	datePostedString <date>	is_bar <dbl>
0	31503-110785431	16	17	55064	USA	2021-07-12	0
1	31503-76611082	16	18	55064	USA	2021-07-12	0
2	31503-93126153	16	19	55064	USA	2021-07-10	0
3	31503-110785598	16	20	55064	USA	2021-07-09	0
4	31503-2101070583	16	21	55064	USA	2021-07-06	0
5	31503-227421330	16	22	55064	USA	2021-07-05	0

Table 1 (above): The first six rows from the Georgia real estate dataset.

The dataset contains many variables; however, we will focus on only the city and price variables. We are also only interested in the cities of Atlanta and Savannah so we will filter for those cities only. Some listings have the price at 0; we will consider these to be missing values and filter them out of the dataset.

[3]:
```
# filtering for Atlanta and Savannah and selecting variables of interest, then
 ↪filtering out listings with price of 0

real_estate_filtered <- real_estate_georgia |>
    filter(city == "Atlanta" | city == "Savannah") |>
    select(city, price) |>
    filter(price != 0)

head(real_estate_filtered)
```

A tibble: 6 × 2

city <chr>	price <dbl>
Atlanta	299900
Atlanta	350000
Atlanta	99000
Atlanta	536068
Atlanta	799990
Atlanta	399900

Table 2 (above): The first six rows of the price data for Georgia and Savannah.

We are interested in the distribution of each individual city, so we will calculate the mean, median, and standard deviation of each city:

[4]:
```
# calculating the sample mean, median, and standard deviation for real estate
 ↪prices in Atlanta and in Savannah

summary <- real_estate_filtered |>
    group_by(city) |>
    summarise(sample_mean = mean(price),
```

```
            sample_median = median(price),
            sample_std = sd(price))
summary
```

A tibble: 2 × 4

city	sample_mean	sample_median	sample_std
<chr>	<dbl>	<dbl>	<dbl>
Atlanta	616754.7	449714	699213.8
Savannah	358771.2	225000	483254.2

Table 3 (above): The mean, median, and standard deviations of the Georgia and Savannah samples.

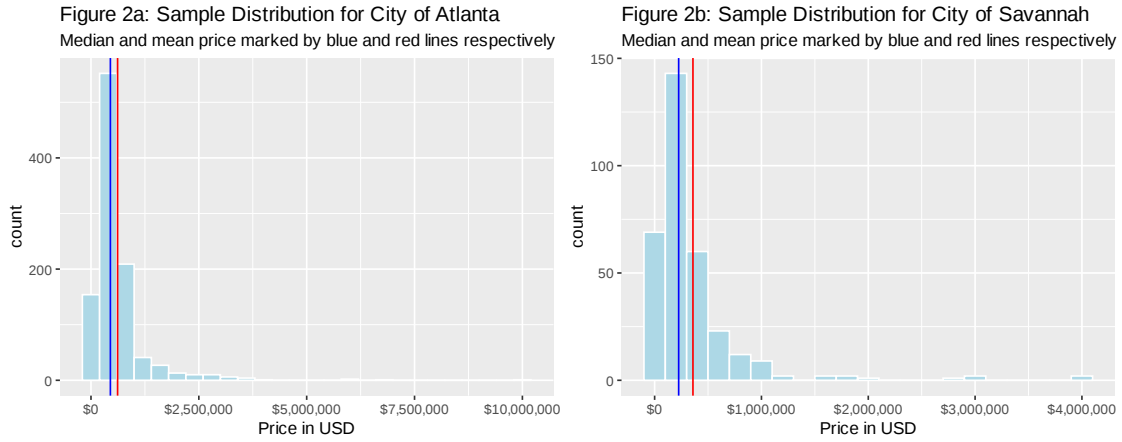
We notice that the mean, median, and standard deviation of the Atlanta sample are all greater than those of the Savannah sample.

Let us also plot the price distributions of each of the two cities. The median price of each city is marked with a blue line and the mean price is marked with a red line.

[5]:
```
# plotting sample mean and median lines on distributions

# setting plot width and height
options(repr.plot.width = 10, repr.plot.height = 4)

atlanta_sample_dist_plot <- real_estate_georgia |>
    filter(city == "Atlanta") |>
    select(city, price) |>
    ggplot(aes(x = price)) +
    geom_histogram(binwidth = 400000, colour = "white", fill = "light blue") +
    geom_vline(xintercept = summary$sample_mean[1], colour = "red") +
    geom_vline(xintercept = summary$sample_median[1], colour = "blue") +
    xlab("Price in USD") +
    ggtitle("Figure 2a: Sample Distribution for City of Atlanta") +
    labs(subtitle = "Median and mean price marked by blue and red lines␣
 ↪respectively") +
    scale_x_continuous(labels = scales::dollar_format())


savannah_sample_dist_plot <- real_estate_georgia |>
    filter(city == "Savannah") |>
    select(city, price) |>
    ggplot(aes(x = price)) +
    geom_histogram(binwidth = 200000, colour = "white", fill = "light blue") +
    geom_vline(xintercept = summary$sample_mean[2], colour = "red") +
    geom_vline(xintercept = summary$sample_median[2], colour = "blue") +
    xlab("Price in USD") +
    ggtitle("Figure 2b: Sample Distribution for City of Savannah") +
    labs(subtitle = "Median and mean price marked by blue and red lines␣
 ↪respectively") +
    scale_x_continuous(labels = scales::dollar_format())
```

```
plots <- grid.arrange(atlanta_sample_dist_plot, savannah_sample_dist_plot, ncol␣
  ↪= 2)
plots
```

```
TableGrob (1 x 2) "arrange": 2 grobs
  z     cells    name           grob
1 1 (1-1,1-1) arrange gtable[layout]
2 2 (1-1,2-2) arrange gtable[layout]
```



We observe that the real estate prices in both Atlanta and Savannah are right-skewed (mean is greater than median for both cities).

1.0.4 Do Atlanta and Savannah have different mean housing prices?

To answer this question, we will do a two-sample t-test for difference in means, with a significance level of 5%. Our null hypothesis (H_0) is that the mean real estate price in Atlanta is equal to the mean real estate price in Savannah ($\mu_1 = \mu_2$). The alternative hypothesis (H_1) is that the mean real estate price in the two cities is not equal ($\mu_1 \neq \mu_2$).

A two-sample t-test is appropriate because: * the population standard deviation is unknown * the sample size is large enough (>30 for both cities), so we can apply the Central Limit Theorem for normality of the sampling distribution

Afterward, we will also construct a confidence interval with bootstrapping and compare the results with the hypothesis test.

Two-Sample T-Test To use the t.test function we are first going to separate each city into their own groups using the filter function and then pulling the price for each group. We would then use Atlanta as x and Savannah as y. We are also using "two.sided" as the alternative since we have an alternative hypothesis of $\mu_1 \neq \mu_2$.

[17]:
```
# setting the seed
set.seed(2021)

# doing a two-sample t-test for difference in means

two_sample_t_test <- tidy(t.test(x = real_estate_filtered |> filter(city ==␣
 ↪"Atlanta") |> pull(price),
                                 y = real_estate_filtered |> filter(city ==␣
 ↪"Savannah") |> pull(price),
                                 alternative = "two.sided"))

two_sample_t_test
```

A tibble: 1 × 10

estimate	estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
257983.5	616754.7	358771.2	7.443283	2.612355e-13	776.082	189945.2	326021.8

Table 4 (above): The results of the two sample t-test for the difference in means between Georgia and Savannah real estate prices.

We obtain a test statistic of 7.443283. For a two-sided test, this corresponds to a p-value of approximately 2.61×10^{-13}, which is below our significance level of 0.05. Therefore, we reject the null hypothesis that the mean real estate price in Atlanta is equal to the mean real estate price in Savannah. We have sufficient evidence that the mean real estate price in the two cities is different.

Constructing 95% Confidence Interval To construct the confidence interval we are going to use the infer package. First we are going to specify the targeted variables, in this case comparison of price depending on the city. After selecting the variables we are going to generate a bootstrap of 2000 samples and calculate the difference of means between the two variables.

[18]:
```
# setting the seed
set.seed(2021)

# constructing bootstrap confidence interval for difference in mean

# finding bootstrap distribution for difference in means
bootstrap_diff_in_means <- real_estate_filtered |>
    specify(formula = price ~ city) |>
    generate(type = "bootstrap", reps = 2000) |>
    calculate(stat = "diff in means", order = c("Atlanta", "Savannah"))

head(bootstrap_diff_in_means)
```

A infer: 6 × 2

replicate <int>	stat <dbl>
1	265115.4
2	257891.0
3	285799.7
4	255840.9
5	254967.3
6	324673.8

Table 5 (above): The first six rows of the bootstrap sampling distribution for the difference in means for Georgia and Savannah real estate prices.

To plot the bootstrap distribution and confidence interval we are going to use the ggplot() function and the get_ci() function. After obtaining the upper and lower confidence bound of the interval we are going to shade the area in between in the plot.

[19]:
```
# making a plot of the bootstrap distribution
bootstrap_diff_in_means_plot <- bootstrap_diff_in_means |>
    ggplot(aes(x = stat)) +
    geom_histogram(bins = 30) +
    ggtitle("Figure 3: Bootstrap Distribution of Difference in Mean (Atlanta␣
 ↪minus Savannah)",
            subtitle = "95% confidence interval shaded in blue") +
    scale_x_continuous(labels = scales::dollar_format()) +
    xlab("Difference in Mean Price (US Dollars)") +
    theme(text = element_text(size = 14))

# finding a 95% confidence interval based on the bootstrap distribution
bootstrap_diff_in_means_ci <- bootstrap_diff_in_means |>
    get_ci(level = 0.95, type = "percentile")

# shading the region in the plot containing the 95% confidence interval
bootstrap_diff_in_means_ci_plot <- bootstrap_diff_in_means_plot +
    annotate("rect", xmin = bootstrap_diff_in_means_ci[[1]], xmax =␣
 ↪bootstrap_diff_in_means_ci[[2]], ymin = 0, ymax = Inf,
             fill = "skyblue",
             alpha = 0.3)

bootstrap_diff_in_means_ci
```

A tibble: 1 × 2

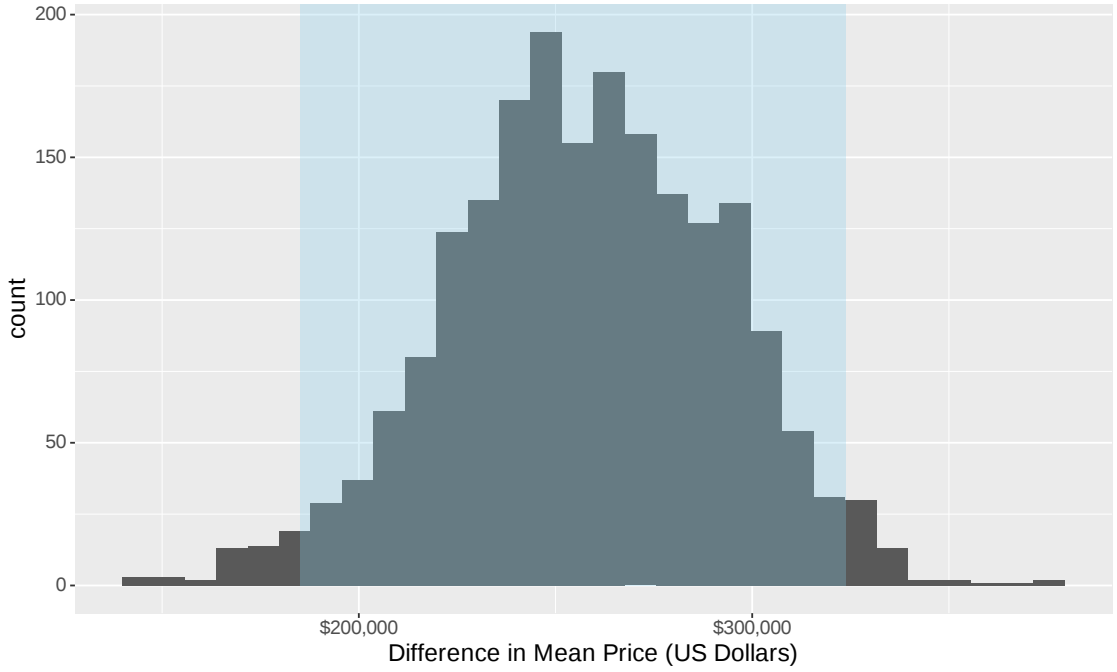
lower_ci <dbl>	upper_ci <dbl>
185232	323789.4

Table 6 (above): The 95% confidence interval constructed using the bootstrap sampling distribution from Table 5.

[20]:
```
# setting plot width and height
options(repr.plot.width = 9, repr.plot.height = 6)
```

```
bootstrap_diff_in_means_ci_plot
```

Figure 3: Bootstrap Distribution of Difference in Mean (Atlanta minus Savannah)
95% confidence interval shaded in blue



From our confidence interval, we can say that we are 95% confident that the interval from 185232 to 323789.4 US dollars captures the true difference in the mean price between Atlanta and Savannah. Since 0 does not lie in our confidence interval, this suggests that the mean prices in the two cities are different. This is consistent with the two-sample t-test that we conducted above.

1.0.5 Do Atlanta and Savannah have different median real estate prices?

Earlier, we found that Atlanta and Savannah had different mean real estate prices. However, sometimes the mean is not the best measurement when the data is skewed. We found that for both cities, the sample had a price distribution that was skewed to the right. Here, we will do a permutation test to see if there is a statistically significant difference in the median real estate prices between the two cities. Afterward, we will construct a 95% confidence interval using the bootstrap distribution for the difference in medians.

Permutation Test We will first observe the difference in median real estate prices within our sample.

[21]:
```
# doing a permutation test for difference in medians

# Finding the difference in medians in our sample

observed_median_price_diff <- real_estate_filtered |>
```

```
    group_by(city) |>
    summarize(median = median(price)) |>
    pivot_wider(names_from = city, values_from = median) |>
    transmute(difference = Atlanta - Savannah)

observed_median_price_diff
```

A tibble: 1 × 1

difference
<dbl>
224714

Table 7 (above): Observed difference in median price between the Atlanta and Savannah sample.

The observed difference in the median prices (Atlanta minus Savannah) is $224714. Let's see how likely we are to observe such a difference with the null model.

We are now going to test for independence with 2000 permutations with difference in medians for its statistic.

[22]:
```
# setting the seed
set.seed(2021)

# generating a null model

null_model_real_estate <- real_estate_filtered |>
    specify(formula = price ~ city) |>
    hypothesize(null = "independence") |>
    generate(reps = 2000, type = "permute") |>
    calculate(stat = "diff in medians", order = c("Atlanta", "Savannah"))


head(null_model_real_estate)
```

A infer: 6 × 2

replicate	stat
<int>	<dbl>
1	-43801.5
2	-9450.0
3	-9550.0
4	24501.0
5	-49550.0
6	550.0

Table 8 (above): First six rows of the null model for the difference in median of real estate prices in Atlanta and Savannah.

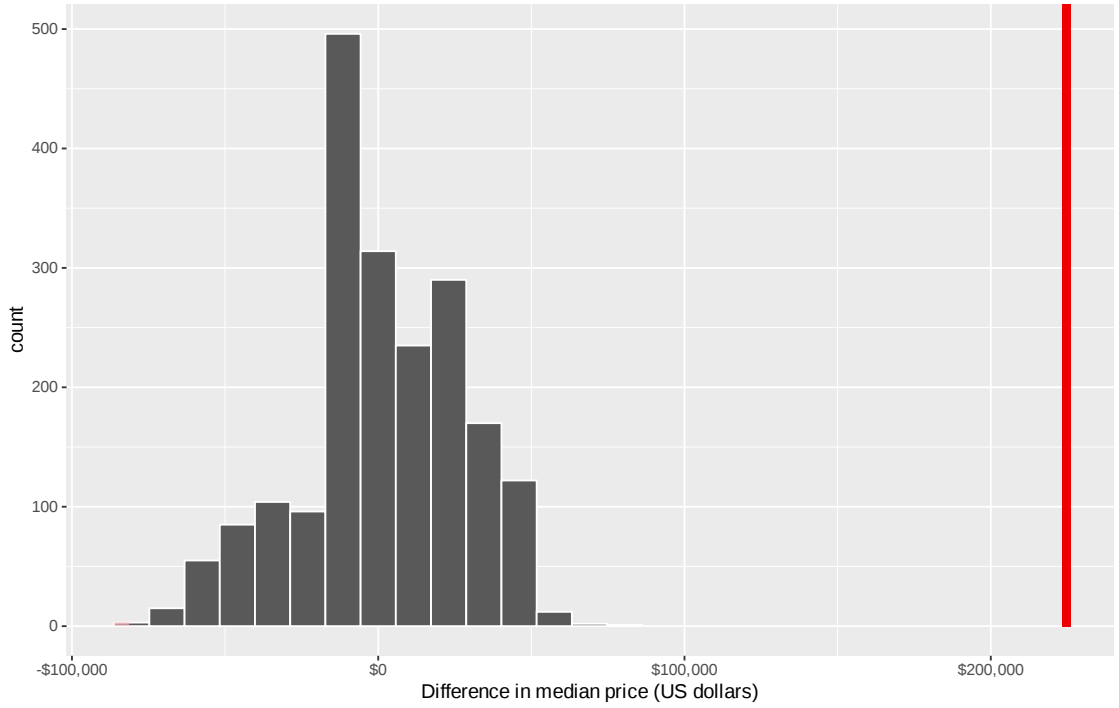
Now we are going to plot the histogram of the values generated while also shading its p-value using the shade_p_value function and obtaining the numbers from the get_p_value function.

[23]:
```
# plotting the result of the hypothesis test, and obtaining the p-value

median_diff_plot <- null_model_real_estate |>
```

```
    visualize() +
    shade_p_value(obs_stat = observed_median_price_diff, direction = "both") +
    xlab("Difference in median price (US dollars)") +
    scale_x_continuous(labels = scales::dollar_format()) +
    theme(text = element_text(size = 11.5)) +
    ggtitle("Figure 4: Simulation-Based Null Distribution for Difference in␣
  ↪Medians (Atlanta minus Savannah)")

p_value_2 <- null_model_real_estate |>
    get_p_value(obs_stat = observed_median_price_diff, direction = "both")


median_diff_plot
```

```
Warning message:
"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. See
`?get_p_value()` for more information."
```

Figure 4: Simulation-Based Null Distribution for Difference in Medians (Atlanta minus Savannah)



[24]:
```
p_value_2
```

A tibble: 1×1

p_value
<dbl>
0

Table 9 (above): Observed p-value of the permutation test.

Even though our p-value above is 0, it is not recommended to report such a p-value as it suggests that making a Type I error is impossible. Since we have 2000 repetitions in the null model, we find that it is best to report the p-value as being less than $\frac{1}{2000}$; in other words, $p < 0.0005$. This is below our significance level of 0.05, so we reject the null hypothesis that the median real estate price is the same in the two cities of Atlanta and Savannah. We have sufficient evidence that the median real estate price in the two cities is different.

Constructing 95% confidence interval We are also going to construct a confidence interval for difference in medians using bootstrapping.

To construct the confidence interval for medians, we are going to use the same method when generating the confidence interval for means. The difference is that in the calculate() method, instead of using "diff in means" we are going to use "diff in medians." As for getting and shading the confidence interval, the same process was done.

[25]:
```
# setting the seed
set.seed(2021)

# constructing bootstrap confidence interval for difference in median

# finding bootstrap distribution for difference in medians
bootstrap_diff_in_medians <- real_estate_filtered |>
    specify(formula = price ~ city) |>
    generate(type = "bootstrap", reps = 2000) |>
    calculate(stat = "diff in medians", order = c("Atlanta", "Savannah"))

# making a plot of the bootstrap distribution
bootstrap_diff_in_medians_plot <- bootstrap_diff_in_medians |>
    ggplot(aes(x = stat)) +
    geom_histogram(bins = 30) +
    ggtitle("Figure 5: Bootstrap Distribution of Difference in Median (Atlanta␣
 ↪minus Savannah)",
            subtitle = "95% confidence interval shaded in blue") +
    xlab("Difference in Median Price (US Dollars)") +
    scale_x_continuous(labels = scales::dollar_format()) +
    theme(text = element_text(size = 14))

# finding a 95% confidence interval based on the bootstrap distribution
bootstrap_diff_in_medians_ci <- bootstrap_diff_in_medians |>
    get_ci(level = 0.95, type = "percentile")

# shading the region in the plot containing the 95% confidence interval
bootstrap_diff_in_medians_ci_plot <- bootstrap_diff_in_medians_plot +
    annotate("rect", xmin = bootstrap_diff_in_medians_ci[[1]], xmax =␣
 ↪bootstrap_diff_in_medians_ci[[2]], ymin = 0, ymax = Inf,
             fill = "skyblue",
```

```
            alpha = 0.3)


head(bootstrap_diff_in_medians)
```

A infer: 6 × 2

replicate <int>	stat <dbl>
1	214000
2	238100
3	254900
4	224100
5	210979
6	245000

Table 10 (above): The first six rows of the bootstrap sampling distribution for the difference in medians for Georgia and Savannah real estate prices.

[26]:
```
bootstrap_diff_in_medians_ci
```

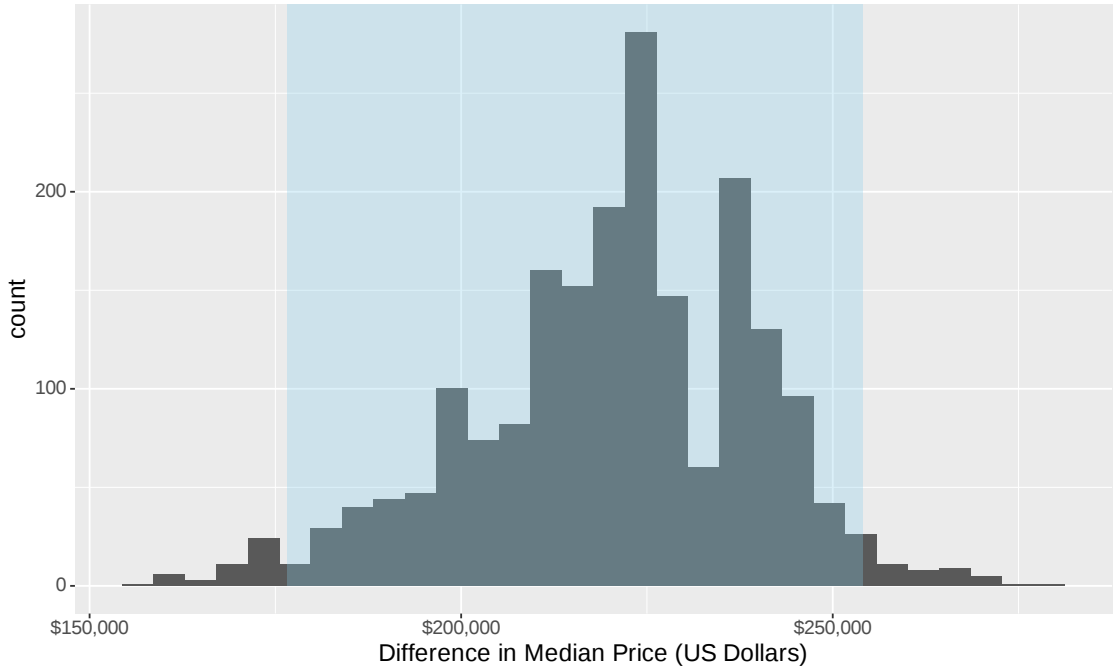
A tibble: 1 × 2

lower_ci <dbl>	upper_ci <dbl>
176598.8	254000

Table 11 (above): The 95% confidence interval constructed using the bootstrap sampling distribution from Table 10.

[27]:
```
bootstrap_diff_in_medians_ci_plot
```

Figure 5: Bootstrap Distribution of Difference in Median (Atlanta minus Savannah)
95% confidence interval shaded in blue



From our confidence interval, we can say that we are 95% confident that the interval from 176598.8 to 254000 US dollars captures the true difference in the median price between Atlanta and Savannah. Since 0 does not lie in our confidence interval, this suggests that the median prices in the two cities are different. This is consistent with our permutation test above.

1.0.6 Discussion

Statistical inference and hypothesis testing via the bootstrap method and the asymptotic method were performed throughout this project to address whether there was a difference in the mean and median real estate prices in Atlanta and Savannah, Georgia. After performing a two sample t-test with a 5% significance level for the difference in means, we obtained a p-value less than our significance level and subsequently rejected our null hypothesis. This meant that there was enough evidence to support a difference in mean real estate price between Atlanta and Savannah. After performing a permutation hypothesis test with a 5% significance level for the difference in medians, we, again, obtained a p-value less than 5% and rejected the null hypothesis, signifying that there was enough evidence to support a difference between the median real estate price between Atlanta and Savannah. As well, for both the difference in mean and median price, our 95% confidence intervals constructed using bootstrapping did not include zero, suggesting that the mean and median prices were different. Overall, the findings in our report show that Atlanta is the more expensive city to buy real estate, which may impact which city consumers decide to live in.

While standard deviation as a parameter was originally included in the proposal to determine the city with the most varied housing price, the hypothesis testing was unable to generate a difference in standard deviations. As well, we had not learned in class about any hypothesis tests for the difference in standard deviation between two populations. Because of this, the parameter was omitted from the final project.

Based on the background research done for this project, we expected there to be a difference in means and medians between the two cities, given that Atlanta is 14% more expensive than the state average to Savannah's 1%. The findings align well with our expectations.

Future questions regarding real estate in Georgia could potentially look at multiple (more than 3) counties or multiple groups of counties in the state and compare real estate variance using the ANOVA method. Additional research could also be done to further investigate finding the difference in standard deviations between two independent populations, so that this parameter may also be analyzed.

1.0.7 References

Cost of Living in Georgia 2023 | RentCafe. (n.d.). Retrieved July 30, 2023, from Www.rentcafe.com. https://www.rentcafe.com/cost-of-living-calculator/us/ga/

FOR RESIDENTIAL REAL ESTATE ACTIVITY IN THE STATE OF GEORGIA Annual Report on the Georgia Housing Market. (n.d.). Retrieved July 30, 2023, from https://garealtor.com/wp-content/uploads/GA-Annual-Housing-Report-2021.pdf

Real Estate Georgia. (n.d.). Kaggle. Retrieved July 30, 2023, from https://www.kaggle.com/datasets/yellowj4acket/real-estate-georgia