Homework 7

Interval Estimation

Jaiden Atterbury

Instructions

Please answer the following questions in the order in which they are posed. Add a few empty lines below each and write your answers there. **Focus on answering in complete sentences and show work whether we ask for it or not**. You will also need scratch paper/pen to work out the answers before typing it.

For help with formatting documents in RMarkdown, please consult R Markdown: The Definitive Guide. Another option is to search using Google.

Exercises

1. (Measurement error) Recall the pH-meter from Homework 6 which was known to give readings that were systematically higher or lower by a quantity δ_0. In order to estimate δ_0, six measurements $X_1, X_2, \ldots, X_6$ were made from a solution with pH **known** to be 4.84. In your previous homework, you were asked to come up with an estimator for δ_0. Let's call it $\hat{\delta}_0^{mom}$.

   Now, suppose four measurements - Y_1, Y_2, Y_3, Y_4 - are made from a solution with an unknown pH-level μ_0 resulting in 4.33, 4.22, 4.23, 4.37. As in the previous homework, the measurement error model is that Y_1, Y_2, Y_3, Y_4 is drawn independently from a distribution with mean $\mu_0 + \delta_0$ and variance σ_0^2.

   Consider the estimator
   $$\hat{\mu}_0 = \bar{Y} - \hat{\delta}_0^{mom}$$
   for μ_0.

a. Show that $\hat{\mu}_0$ is an unbiased estimator of μ_0.

As found in Homework 6, $\hat{\delta}_0^{mom} = \bar{X} - 4.84$ where $\bar{X} = \frac{X_1+X_2+X_3+X_4+X_5+X_6}{6}$. We now consider the estimator $\hat{\mu}_0 = \bar{Y} - \hat{\delta}_0^{mom}$ for μ_0. Where $\bar{Y} = \frac{Y_1+Y_2+Y_3+Y_4}{4}$, $E[Y] = \mu_0 + \delta_0$, and $Var[Y] = \sigma_0^2$. In order to see if our estimator is unbiased for μ_0, we must show that $E[\hat{\mu_0}] = \mu_0$.

$$
\begin{aligned}
E[\hat{\mu}_0] &= E[\bar{Y} - \hat{\delta}_0^{mom}] \\
&= E[\bar{Y}] - E[\hat{\delta}_0^{mom}] \quad \text{(Linearity of expectation)} \\
&= E[\frac{1}{4}\sum_{i=1}^{4} Y_i] - \delta_0 \quad \text{(Homework 6.1.b)} \\
&= \frac{1}{4}\sum_{i=1}^{4} E[Y_i] - \delta_0 \quad \text{(Linearity of expectation)} \\
&= \frac{1}{4} \cdot 4 \cdot (\mu_0 + \delta_0) - \delta_0 \quad \text{(Since } E[Y] = \mu_0 + \delta_0\text{)} \\
&\therefore \hat{\mu}_0 \text{ is an unbiased estimator of } \mu_0
\end{aligned}
$$

b. Give an expression for the standard error of $\hat{\mu}_0$. That is, find $\sqrt{Var\,(\hat{\mu}_0)}$. Show your work. (State any assumptions you need to make)

In order to give and expression for the standard error of $\hat{\mu}_0$, we will need to find $Var[\hat{\mu}_0]$. Furthermore, based on the problem descriptions from Homework 6 Problem 1 and this problem, we are given that $Var[X] = \sigma_0^2$ and $Var[Y] = \sigma_0^2$. Thus, we will assume that the variability of the X and Y measurements are the same, in other words $Var[X] = Var[Y]$. For completeness, we will also assume that the X and Y measurements are independent of each other.

$$
\begin{aligned}
Var[\hat{\mu}_0] &= Var[\bar{Y} - \hat{\delta}_0^{mom}] \\
&= Var[\bar{Y}] + Var[\hat{\delta}_0^{mom}] \quad \text{(Non-linearity of variance)} \\
&= Var[\bar{Y}] + \frac{\sigma_0^2}{6} \quad \text{(Homework 6.1.c)} \\
&= Var[\frac{1}{4}\sum_{i=1}^{4} Y_i] + \frac{\sigma_0^2}{6} \\
&= \frac{1}{16} Var[\sum_{i=1}^{4} Y_i] + \frac{\sigma_0^2}{6} \quad \text{(Non-linearity of variance)} \\
&= \frac{1}{16}\sum_{i=1}^{4} Var[Y_i] + \frac{\sigma_0^2}{6} \quad \text{(Since the } Y_i\text{'s are independent)} \\
&= \frac{1}{16} \cdot 4 \cdot \sigma_0^2 + \frac{\sigma_0^2}{6} \quad \text{(Since } Var[Y] = \sigma_0^2\text{)} \\
&= \frac{\sigma_0^2}{4} + \frac{\sigma_0^2}{6} \\
&= \frac{5}{12}\sigma_0^2 \quad \text{(Assuming } Var[X] = Var[Y]\text{)} \\
&\therefore \sqrt{Var\hat{\mu}_0} = \sqrt{\frac{5}{12}\sigma_0^2} = \sqrt{\frac{5}{12}}\sigma_0
\end{aligned}
$$

c. The variability in the pH measurements - σ_0 - is the same for both the X measurements and also the Y measurements. This makes sense since the variability in the readings is related to the meter, not the specific solution it is being used on.

A natural estimate for σ_0 is a pooled standard deviation s_p calculated from both samples. The formula for s_p is below:

$$
s_p^2 = \frac{\sum_{i=1}^{6}(x_i - \bar{x})^2 + \sum_{j=1}^{4}(y_i - \bar{y})^2}{6 + 4 - 2}
$$

Calculate s_p, the pooled estimate of σ_0.

```
# Six measurements for solution with pH = 4.84 from homework 6
x <- c(4.71, 4.63, 4.69, 4.76, 4.58, 4.83)

# Four measurements for solution with unknown pH from this homework
y <- c(4.33, 4.22, 4.23, 4.37)

# Find the means of the measurements
mean_x <- mean(x)
mean_y <- mean(y)

# Find the sum of squared differences
ssd_x <- sum((x - mean_x)^2)
ssd_y <- sum((y - mean_y)^2)

# Use the above formula to find the pooled variance
pool_var <- (ssd_x + ssd_y) / (6 + 4 - 2)

# Take the square root of the pooled variance to get the pooled estimate of the
# standard deviation.
pool_sd <- sqrt(pool_var)
```

Therefore, based on the above calculations, s_p, the pooled estimate of σ_0 is 0.0840201.

d. Calculate the estimated standard error of $\hat{\mu}_0$. Show your steps.

Based on part b, we know that the standard error of $\hat{\mu}_0$ is $\sqrt{\frac{5}{12}}\sigma_0$. Furthermore, in part c, we calculated the pooled estimate of σ_0, s_p, and found that this estimate was equal to 0.0840201. Putting this all together we can see that the estimated standard error of $\hat{\mu}_0$ is $\sqrt{\frac{5}{12}} \cdot s_p$. Which we can calculate using R.

```
est_se <- sqrt(5 / 12) * pool_sd
```

Therefore our estimated standard error of $\hat{\mu}_0$ is 0.0542347.

2. (Force) A type of metal bar breaks when a force of size X is applied, where X has PDF

$$f(x) = 2\alpha_0 x\, e^{-\alpha_0\, x^2} \qquad x > 0$$

where $\alpha_0 > 0$ is an unknown parameter. We observe a breaking force of 40. Find a 95% confidence interval for α_0.

Hint: We are looking for a random interval $[L, U]$ which contains α_0 with probability 95%. Construct the interval by "inverting" the probability statement

$$P\left(q_{0.025} \leq X \leq q_{0.975}\right) = 0.95$$

where $q_{0.025}$ and $q_{0.975}$ are the 2.5th and 97.5th percentiles of the distribution of X.

Setup:
Since we are looking for an interval $[L, U]$ which contains α_0 with probability 95%, we can do this by "inverting" the probability statement $P\left(q_{0.025} \leq X \leq q_{0.975}\right) = 0.95$. In order to invert this probability statement, we need to find expressions for $q_{0.025}$ and $q_{0.975}$ in terms of α_0. We will start by finding an expression for $q_{0.025}$ in terms of α_0, then move to finding an expression for $q_{0.975}$.

Finding an expression for $\mathbf{q_{0.025}}$:

$$\begin{aligned}
\int_0^{q_{0.025}} f(x)dx &= \int_0^{q_{0.025}} 2\alpha_0 x e^{-\alpha_0 x^2} dx \\
&= \int_0^{-\alpha_0(q_{0.025})^2} \frac{2\alpha_0 x}{-2\alpha_0 x} e^u du \quad (\text{Let } u = -\alpha_0 x^2 \implies du = -2\alpha_0 x) \\
&= -\int_0^{-\alpha_0(q_{0.025})^2} e^u du \\
&= [-e^u]_0^{-\alpha_0(q_{0.025})^2} \\
&= 1 - e^{-\alpha_0(q_{0.025})^2}
\end{aligned}$$

Notice however, that $q_{0.025}$ represents the 0.025th quantile of the PDF of X. Thus $\int_0^{q_{0.025}} f(x)dx = 0.025$ by the definition of a quantile. We will use this fact to find an expression for $q_{0.025}$ in terms of α_0.

$$\begin{aligned}
1 - e^{-\alpha_0(q_{0.025})^2} &= 0.025 \\
e^{-\alpha_0(q_{0.025})^2} &= 0.975 \\
-\alpha_0(q_{0.025})^2 &= \ln(0.975) \\
(q_{0.025})^2 &= \frac{-\ln(0.975)}{\alpha_0} \\
q_{0.025} &= \sqrt{\frac{-\ln(0.975)}{\alpha_0}}
\end{aligned}$$

Finding an expression for $\mathbf{q_{0.975}}$:

$$\begin{aligned}
\int_0^{q_{0.975}} f(x)dx &= \int_0^{q_{0.975}} 2\alpha_0 x e^{-\alpha_0 x^2} dx \\
&= \int_0^{-\alpha_0(q_{0.975})^2} \frac{2\alpha_0 x}{-2\alpha_0 x} e^u du \quad (\text{Let } u = -\alpha_0 x^2 \implies du = -2\alpha_0 x) \\
&= -\int_0^{-\alpha_0(q_{0.975})^2} e^u du \\
&= [-e^u]_0^{-\alpha_0(q_{0.975})^2} \\
&= 1 - e^{-\alpha_0(q_{0.975})^2}
\end{aligned}$$

Notice however, that $q_{0.975}$ represents the 0.975th quantile of the PDF of X. Thus $\int_0^{q_{0.975}} f(x)dx = 0.975$ by the definition of a quantile. We will use this fact to find an expression for $q_{0.975}$ in terms of α_0.

$$\begin{aligned}
1 - e^{-\alpha_0(q_{0.975})^2} &= 0.975 \\
e^{-\alpha_0(q_{0.975})^2} &= 0.025 \\
-\alpha_0(q_{0.9755})^2 &= \ln(0.025) \\
(q_{0.975})^2 &= \frac{-\ln(0.025)}{\alpha_0} \\
q_{0.975} &= \sqrt{\frac{-\ln(0.025)}{\alpha_0}}
\end{aligned}$$

Putting it all together:

$$\begin{aligned}
0.95 &= P(q_{0.025} \le X \le q_{0.975}) \\
&= P(\sqrt{\frac{-\ln(0.975)}{\alpha_0}} \le X \le \sqrt{\frac{-\ln(0.025)}{\alpha_0}}) \\
&= P(\frac{-\ln(0.975)}{\alpha_0} \le X^2 \le \frac{-\ln(0.025)}{\alpha_0}) \\
&= P(-\ln(0.975) \le X^2\alpha_0 \le -\ln(0.025)) \\
&= P(\frac{-\ln(0.975)}{X^2} \le \alpha_0 \le \frac{-\ln(0.025)}{X^2})
\end{aligned}$$

Thus a 95% confidence interval for α_0 is $[\frac{-\ln(0.975)}{X^2}, \frac{-\ln(0.025)}{X^2}]$. In our case, since $x_{obs} = 40$, the interval becomes $[\frac{-\ln(0.975)}{40^2}, \frac{-\ln(0.025)}{40^2}]$ which is approximately $[0.000016, 0.002306]$.

3. (CLT) A sample of 83 observations for an integer-valued random variable Y is shown below:

value	0	1	2	3	4	5
frequency	13	18	23	15	6	8

Use the Central Limit Theorem to find a 90% confidence interval for $\pi_0 = P(Y \ge 2)$. Show your work, develop your answer. We are grading on style.

Hint: You actually have 83 independent Bernoulli random variables - $X_1, X_2, \ldots, X_{83}$ - where each X_i is one if $Y \ge 2$ and zero otherwise. Therefore you can think of $X_1, X_2, \ldots, X_{83} \overset{i.i.d.}{\sim} Binom(1, \pi_0)$ and you wish to construct a confidence interval for the mean of the distribution - π_0 - using the CLT.

Setup:
In this problem, we are given 83 observations for an integer-valued random variable Y and told to find a 90% confidence interval for $\pi_0 = P(Y \ge 2)$. Furthermore, since an observed value of Y can either be greater than or equal to 2 or not, it turns out that we actually have 83 independent Bernoulli random variables - $X_1, X_2, \ldots, X_{83}$ - where each X_i is one if $Y \ge 2$ and zero otherwise. Thus, we can think of $X_1, X_2, \ldots, X_{83} \overset{i.i.d.}{\sim} Binom(1, \pi_0)$ and we want to construct a 90% confidence interval for π_0 using the Central Limit Theorem.

Notice, that π_0 is actually the mean of the distribution of the X_i's. Furthermore, since we are looking to find a 90% confidence interval for π_0, we are going to need to find an estimator for π_0. However, since π_0 is the mean of the distribution, it follows that $\bar{X}$ is the optimal estimator for this value. Thus, we will build our confidence interval around $\bar{X}$. From the sample of 83 observations for an integer-valued random variable Y, we see that 52 are greater than or equal to 2. Therefore the amount of successes in our 83 Bernoulli random variables is 52. In our case it follows that $\hat{\pi}_0 = \bar{x} = \frac{52}{83}$.

Since $\bar{X}$ includes the sum of 83 independent Bernoulli random variables, and 83 is large (in particular $83 > 30$), it follows that the Central Limit Theorem applies. This theorem tells us that $\bar{X} \approx Norm(\mu_0, \frac{\sigma_0}{\sqrt{n}})$. Furthermore, slide 7 of the "summary of confidence intervals" lecture notes tells us that if $X_i \sim Binom(1, \pi_0)$, then $\bar{X} \approx Norm(\pi_0, \sqrt{\frac{\pi_0(1-\pi_0)}{n}})$.

However, as expected, we don't have the true value of σ_0. Thus, we can either 1.) find the standard deviation of our sample values, or 2.) we could use our estimate of π_0 to estimate σ_0. In our case it is easier and more practical to use our estimate of π_0 to estimate σ_0, this turns out to be $s = \sqrt{\frac{\hat{\pi}_0(1-\hat{\pi}_0)}{n}} = \sqrt{\frac{\frac{52}{83}(1-\frac{52}{83})}{83}} = \sqrt{\frac{1612}{571787}}$.

Putting it all together:
Putting this all together, we can see that the 90% confidence interval for π_0 is of the form $[\bar{x} - 1.65 \cdot s, \bar{x} + 1.65 \cdot s]$.

Plugging in our numbers we get $[\frac{52}{83} - 1.65\sqrt{\frac{1612}{571787}}, \frac{52}{83} + 1.65\sqrt{\frac{1612}{571787}}]$, to avoid rounding error we will calculate this in R.

```r
lower_pi <- 52/83 - qnorm(0.95) * sqrt(1612/571787)
upper_pi <- 52/83 + qnorm(0.95) * sqrt(1612/571787)
```

Therefore, our 90% confidence interval for π_0 is [0.5391702, 0.7138419]

4. (Airbnb) Read sections 18.3 and 19.2 in the Notes where I constructed a confidence interval for the mean (daily) price of 2 bedroom apartment rentals in Seattle. In this section you will repeat this calculation for a different subset of rentals: houses with 3 or more bedrooms where the entire home is for rent. The variables you will be filtering on and their values are shown below:
   - property_type: Houses
   - room_type: Entire home/apt
   - bedrooms: 3 or more

a. In this part, you will construct a large sample 95% confidence interval for the mean price of all such house rentals in Seattle. Be sure to

New filtered data frame:

```r
airbnb <- read_csv("listings.csv")

airbnb_subset <- airbnb %>%
  filter(property_type == "House",
         room_type == "Entire home/apt",
         bedrooms >= 3) %>%
  mutate(price = parse_number(price)) %>%
  select(price)
```

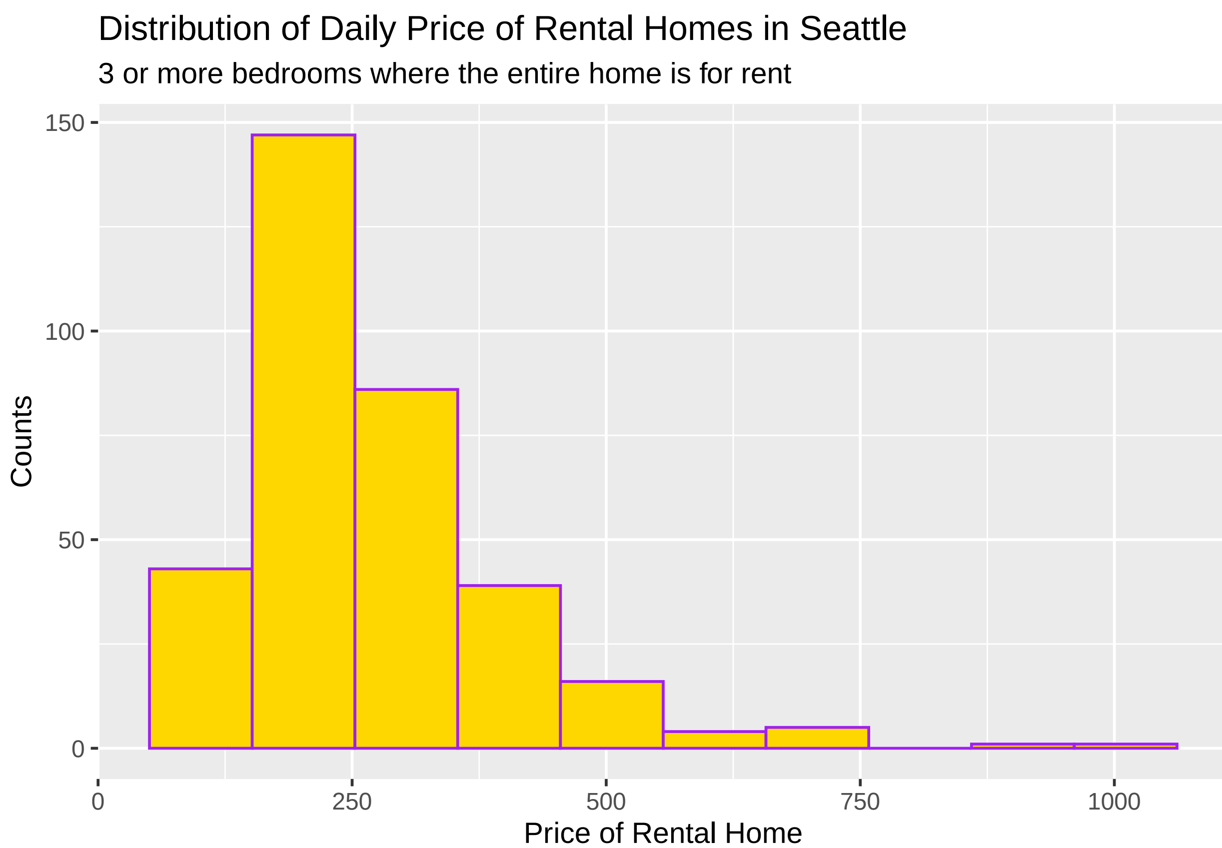
- display the first five rows of the filtered data frame

```r
head(airbnb_subset, 5)
```

```
## # A tibble: 5 x 1
##   price
##   <dbl>
## 1   975
## 2   450
## 3   461
## 4   700
## 5   450
```

- make a histogram of `price` and

```r
# Number of bins = ceiling of log_2(342) + 1

ggplot(data = airbnb_subset,
       mapping = aes(x = price)) +
  geom_histogram(bins = 10,
                 color = "purple",
                 fill = "gold") +
  labs(title = "Distribution of Daily Price of Rental Homes in Seattle",
       subtitle = "3 or more bedrooms where the entire home is for rent",
       x = "Price of Rental Home",
       y = "Counts")
```



- calculate and report a large sample 95% confidence interval for the mean daily price. (See section 18.3 from pages 206-208 for example code.)

```
airbnb_subset %>%
  summarise(xbar = mean(price),
            s = sd(price),
            n = n(),
            se = s / sqrt(n),
            lower = xbar - qnorm(0.975) * se,
            upper = xbar + qnorm(0.975) * se )
```

```
## # A tibble: 1 x 6
##    xbar     s     n    se lower upper
##   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
## 1  277.  131.   342  7.08  263.  291.
```

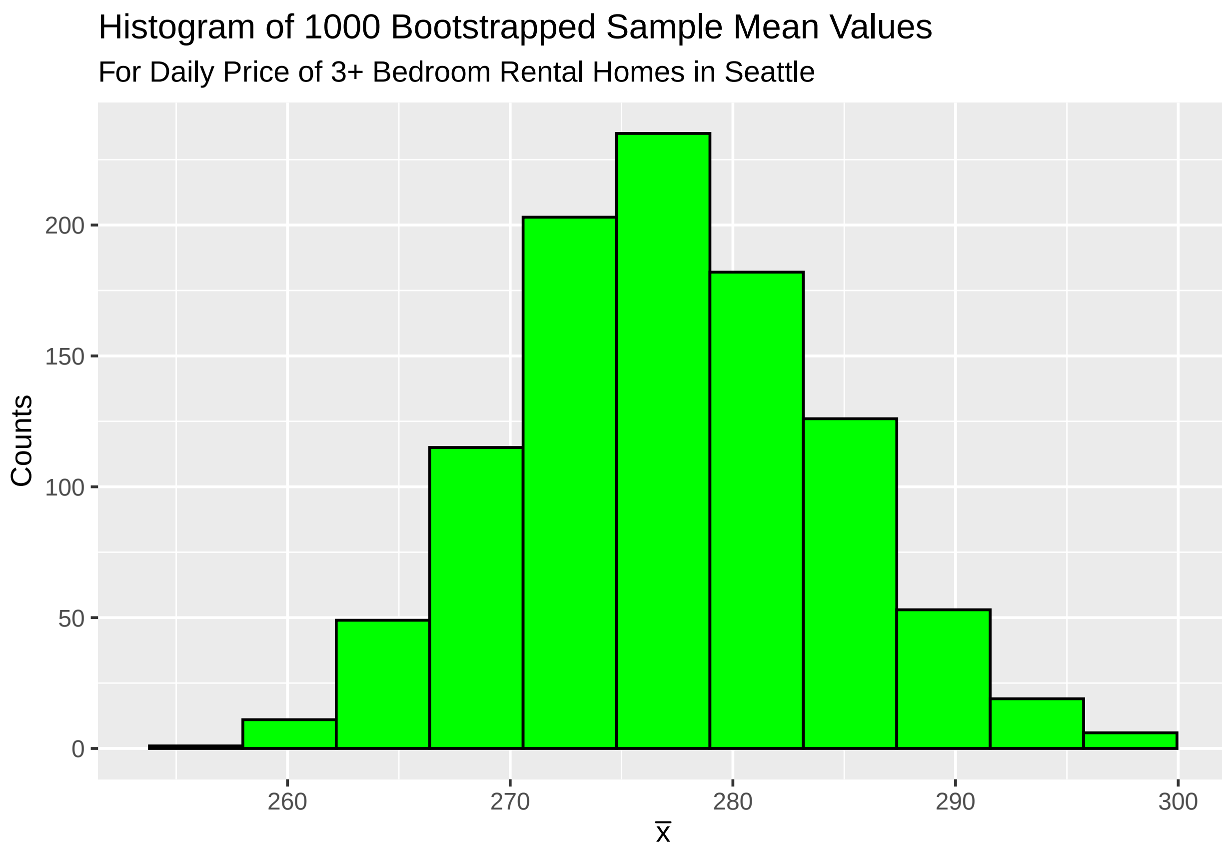
As calculated above, a large sample 95% confidence interval for the mean daily price of houses with 3 or more bedrooms where the entire home is for rent in Seattle is $[263.3743, 291.1345]$.

b. In this part, you will construct a (non-parametric) bootstrap confidence interval for the mean price of houses with 3 or more bedrooms where the entire home is for rent. Be sure to

   - display the bootstrap sampling distribution of the sample mean

```
set.seed(14141)
B = 1000

# Create the data frame of mean values
boot_df <- tibble(
```

```r
  xbar_star = replicate(n = B,
                        expr = mean(sample(x = airbnb_subset$price,
                        size = nrow(airbnb_subset),
                        replace=TRUE)))

# Number of bins = ceiling of log_2(1000) + 1
ggplot(data=boot_df,
       mapping = aes(x=xbar_star)) +
  geom_histogram(bins = 11,
                 color = "black",
                 fill = "green") +
  labs(title="Histogram of 1000 Bootstrapped Sample Mean Values",
       subtitle="For Daily Price of 3+ Bedroom Rental Homes in Seattle",
       x = expression(bar("x")),
       y = "Counts")
```

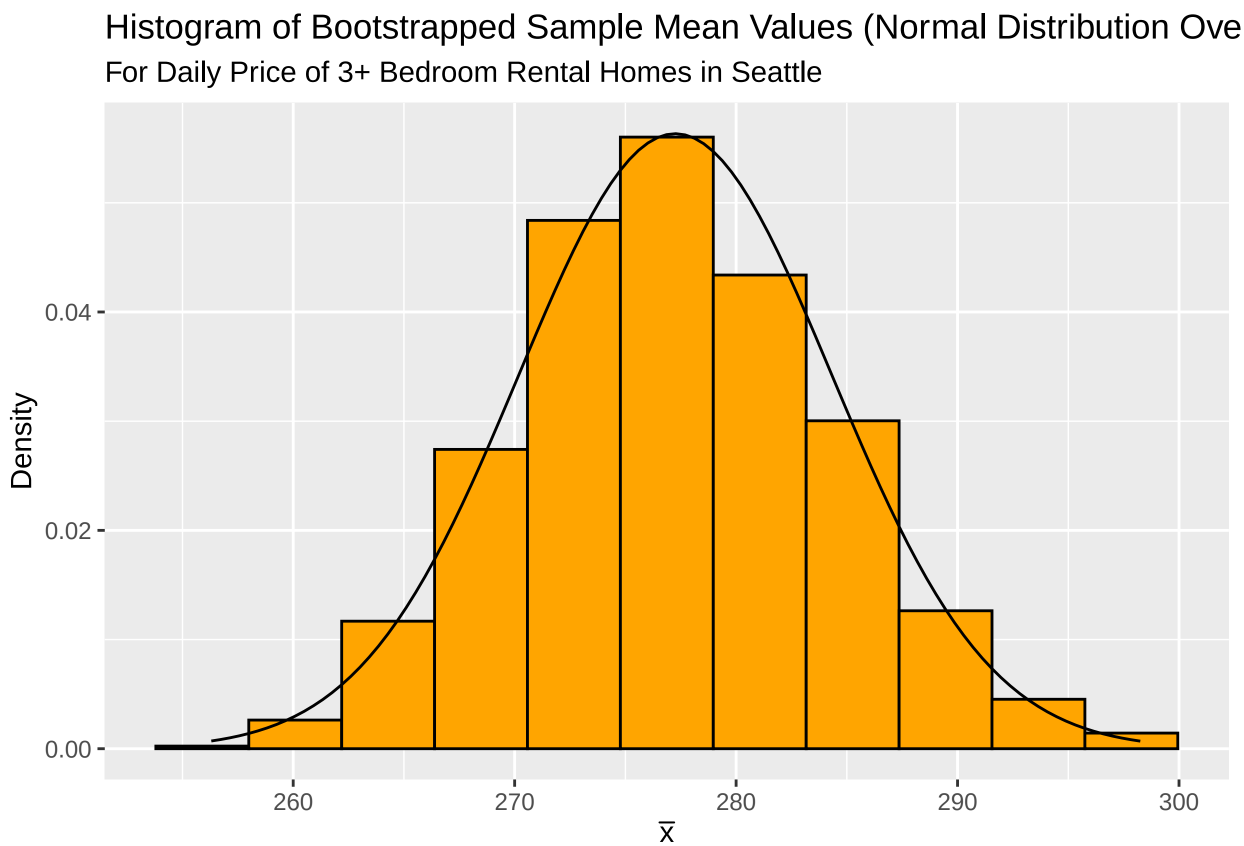


- compare the bootstrap sampling distribution with the normal distribution

To compare the bootstrap sampling distribution with the normal distribution, we will overlay a normal distribution onto the above histogram, as well as make and analyze a QQplot.

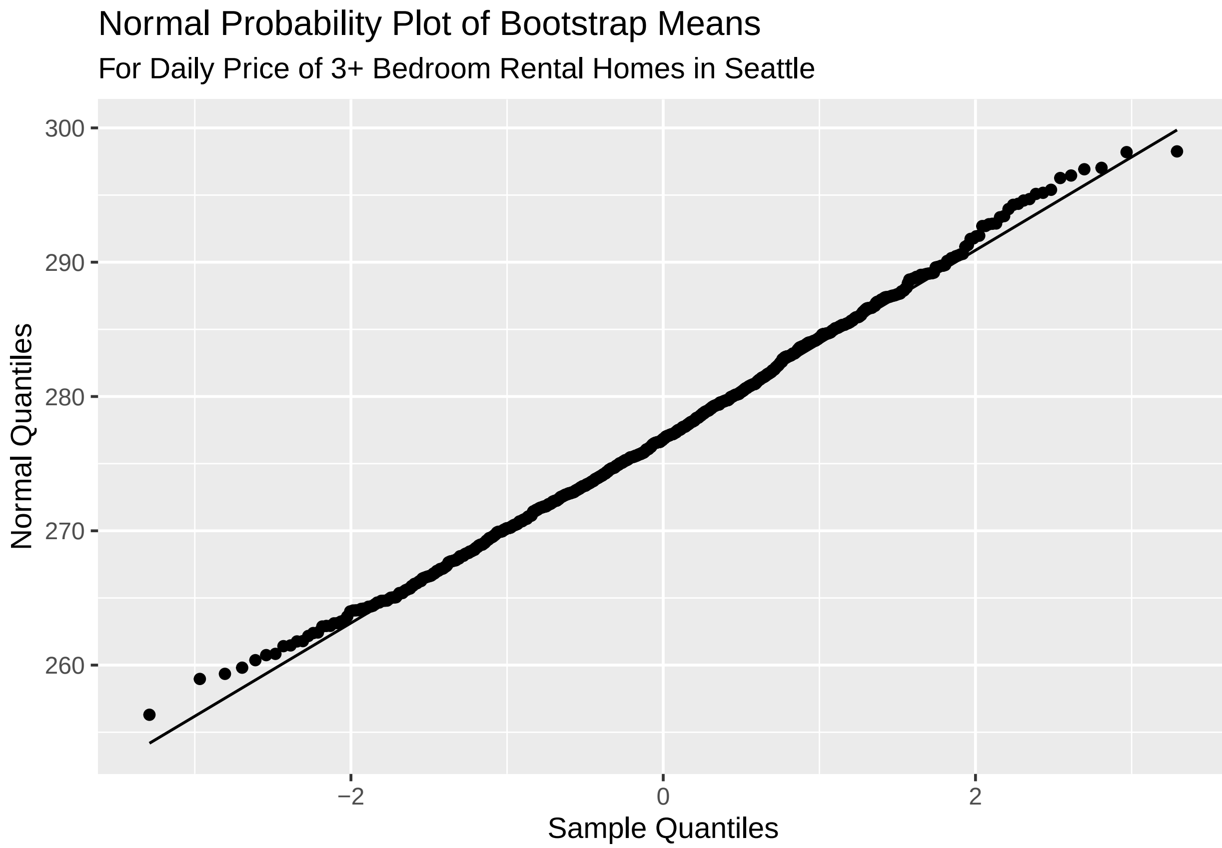
```r
n <- nrow(airbnb_subset)
norm_mean <- mean(airbnb_subset$price)
norm_sd <- sd(airbnb_subset$price) / sqrt(n)

# Number of bins = ceiling of log_2(1000) + 1
ggplot(data=boot_df) +
```

```
  geom_histogram(mapping = aes(x=xbar_star, y=after_stat(density)),
                 bins = 11,
                 color = "black",
                 fill = "orange") +
  geom_function(fun = dnorm, args = list(mean = norm_mean, sd = norm_sd)) +
  labs(title="Histogram of Bootstrapped Sample Mean Values (Normal Distribution Overlaid)",
       subtitle="For Daily Price of 3+ Bedroom Rental Homes in Seattle",
       x = expression(bar("x")),
       y = "Density")
```



As can be seen from the above histogram with the normal distribution overlaid, the bootstrapped sampling distribution of $\bar{X}$ seems to be approximately normal. We will now analyze the corresponding qqplot to further emphasize this relatioship.

```
ggplot(data=boot_df,
       mapping = aes(sample = xbar_star))+
  stat_qq(distribution = qnorm)+
  stat_qq_line(distribution = qnorm)+
  labs(title = "Normal Probability Plot of Bootstrap Means",
       subtitle="For Daily Price of 3+ Bedroom Rental Homes in Seattle",
       x = "Sample Quantiles",
       y = "Normal Quantiles")
```



As we can see, despite a little bit of discrepancy at the tails, due to the linear relationship between the sample quantiles and the normal quantiles the sampling distribution of the estimator seems to follow a normal distribution. This normality of $\bar{X}$ is because the sample size of $n = 342$ is large enough for the Central Limit Theorem to apply.

- calculate and report the standard bootstrap confidence limits (See section 19.2 on pages 219 - 222 for example code)

Since our bootstraped sampling distribution is approximately normal, we will use the standard bootstrap method to construct our confidence interval.

```
lower <- mean(airbnb_subset$price) -  qnorm(0.975) * sd(boot_df$xbar_star)
upper <- mean(airbnb_subset$price) +  qnorm(0.975) * sd(boot_df$xbar_star)
```

As calculated above, the standard bootstrap confidence limits are [263.4878915, 291.0208804]. Due to the approximate normality of our bootstrapped estimator, the resulting confidence interval is very close to the confidence interval calculated above using theory.