Making Inferences on Two Proportions and Inferences on One Mean

Blake Pappas

10/26/2021

Example: Voters

Suppose that we polled a random sample of voters. We found that there were 359 voters who voted by mail in the 2020 election, 192 of whom intend to vote by mail in 2022. Further, there were 161 voters who did not vote by mail in 2020. Of these, 38 intend to vote by mail in 2022.

We will use this (fictional) data to infer whether 2020 voting method are predictive of their intentions to vote by mail in 2022.

Recap: Descriptive Summaries of Two Binary Variables

First, a little setup. The code below will create a table called mail_voting that holds the counts. A little bit of arithmetic will find the number who do not intend to vote by mail in each category.

The `rbind` function creates a matrix of the counts. The grouping variable will be the "row variable" and the outcome variable will be the "column" variable. The row and column names are then set to keep track of the values.

```
voter_table <- as.table(rbind(c(192, 359 - 192), c(38, 161 - 38)))
rownames(voter_table) <- c("mail2020", "not_mail2020")
colnames(voter_table) <- c("mail2022", "not_mail2022")
voter_table
```

```
##              mail2022 not_mail2022
## mail2020          192          167
## not_mail2020       38          123
```

I'll look for evidence of association by comparing the row proportions. The table below shows that among those who voted by mail in 2020, 53.5% intend to vote by mail in 2022. This is higher than the proportion among those who didn't vote by mail in 2020, 23.6%.

```
prop.table(voter_table, margin = 1)
```

```
##               mail2022 not_mail2022
## mail2020     0.5348189    0.4651811
## not_mail2020 0.2360248    0.7639752
```

Inference on Two Proportions in R

Let's define π_1 to be the proportion of 2020 mail-in voters who plan to vote by mail in 2022. Then, π_2 is the proportion of 2020 non-mail voters who plan to vote by mail in 2022. We will use inferential methods to investigate whether the difference $\pi_1 - \pi_2$ is significantly greater than zero, which indicates that 2020 voting preference is associated with greater likelihood of mail voting in 2022.

When using `prop.test()` function for inference on two proportions, the input can either be a `table` or vectors of successes and sample sizes.

Input is a Table

If the input object is a `table`, make sure that the grouping variable is the row variable and the outcome variable is the column variable.

The code below makes a 90% confidence interval for $\pi_1 - \pi_2$:

```
prop.test(voter_table, conf.level = 0.9)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  voter_table
## X-squared = 39.027, df = 1, p-value = 4.18e-10
## alternative hypothesis: two.sided
## 90 percent confidence interval:
##  0.2242593 0.3733289
## sample estimates:
##    prop 1    prop 2
## 0.5348189 0.2360248
```

We are 90% confident that $\pi_1 - \pi_2$ is between 0.2242593 and 0.3733289. Since all of the values in this interval are greater than zero, it seems that there is evidence that 2020 mail-in voters are more likely to vote by mail in 2022 than 2020 non-mail voters.

Input is a Vector of Counts

We can also give `prop.test` a vector of successes and a vector of sample sizes. The code below will test the hypotheses

$$H_0 : \pi_1 - \pi_2 = 0; H_A : \pi_1 - \pi_2 > 0$$

with an α of 0.05.

```
prop.test(x = c(192, 38), n = c(359, 161), alternative = 'greater')
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(192, 38) out of c(359, 161)
## X-squared = 39.027, df = 1, p-value = 2.09e-10
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.2242593 1.0000000
```

```
## sample estimates:
##    prop 1    prop 2
## 0.5348189 0.2360248
```

The p-value for the test is $2.09 \times 10^{-}10$. The squared test statistic (z_0^2) is 39.027. The absolute value of the test statistic is $\sqrt{39.027}$, or 6.247159. The sample proportions are about 6 standard errors apart.

Code to print $|z_0|$ and the p-value:

```
test.result <- prop.test(x = c(192, 38), n = c(359, 161), alternative = 'greater')
abs.z0 <- sqrt(test.result$statistic)
p.value <- test.result$p.value

# Print Values in Console
abs.z0
```

```
## X-squared
##  6.247164
```

```
p.value
```

```
## [1] 2.089861e-10
```

Remember that `prop.test` automatically applies a continuity correction, and so its results will not exactly match the large-sample methods seen in lecture. Just for fun, let's calculate the z0 test statistic using a user-defined formula.

```
n1 <- 359
n2 <- 161
pihat1 <- 192 / n1
pihat2 <- 38 / n2
pihat_pool <- (192 + 38) / (n1 + n2)
z0 <- (pihat1 - pihat2) / sqrt(pihat_pool * (1 - pihat_pool) / n1 + pihat_pool * (1 - pihat_pool) / n2)
z0
```

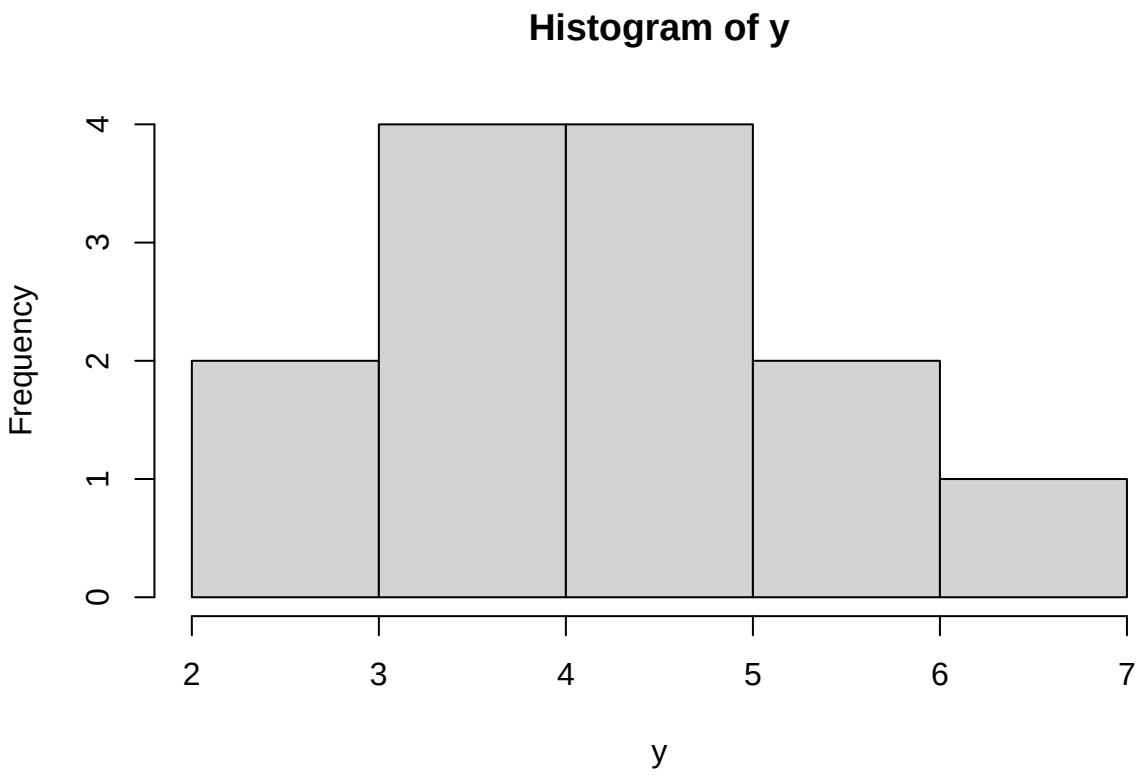
```
## [1] 6.342653
```

If you set "correct = FALSE" in the `prop.test` call, R will not use a continuity correction. Under this setting, the test statistic will be equal to the absolute value of z0.

Inference on One Mean in R

Inference on one mean is very simple in R. First, we'll generate some artificial data upon which to perform the t-test and look at the data distribution.

```
# Generate 13 Samples from a N(4, 1^2) Distribution
set.seed(109202)
n <- 13
y <- rnorm(n, 4, 1)
hist(y)
```




```
summary(y)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.569   3.647   4.152   4.220   4.769   6.381
```

The histogram above has a symmetric shape that looks approximately bell-shaped. The observations range from about 2.6 to 6.4.

t Confidence Interval and t Test in R

The `t.test` function can be use to perform the one-sample interval and test for μ.
The first input value is a vector of numeric data. By default, it will calculate a 95% confidence interval and test the hypotheses

$$H_0 : \mu = \mu_0; \quad H_A : \mu \neq \mu_0.$$

The following code makes a 95% confidence interval for the population mean. (Note that the data was generated from a N(4, 1^2) distribution, so the true μ value is 4.)

```
t.test(y, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  y
```

```
## t = 13.588, df = 12, p-value = 1.197e-08
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  3.543655 4.897085
## sample estimates:
## mean of x
##   4.22037
```

Save the results in an object if you want to extract certain pieces. Here is the confidence interval for μ:

```
results <- t.test(y, conf.level = 0.95)
results$conf.int
```

```
## [1] 3.543655 4.897085
## attr(,"conf.level")
## [1] 0.95
```

To test

$$H_0 : \mu = 5; \quad H_A : \mu \neq 5,$$

set the `mu` option to 5 in the function.

```
results2 <- t.test(y, conf.level = 0.95, mu = 5, alternative = 'two.sided')
results2$statistic
```

```
##         t
## -2.510169
```

```
results2$p.value
```

```
## [1] 0.02739933
```

The test statistic is $t_0 = -2.510169$, meaning that the observed y-bar is about 2.5 standard errors below 5. The p-value is 0.0274. For an alpha of 0.05, the null hypothesis is rejected and we conclude that there is evidence that $\mu \neq 5$.

Use the option "alternative" to change the direction of the alternative.

```
# Test H0: mu=5; HA: mu <5
results3 <- t.test(y, mu = 5, alternative = "less")
results3
```

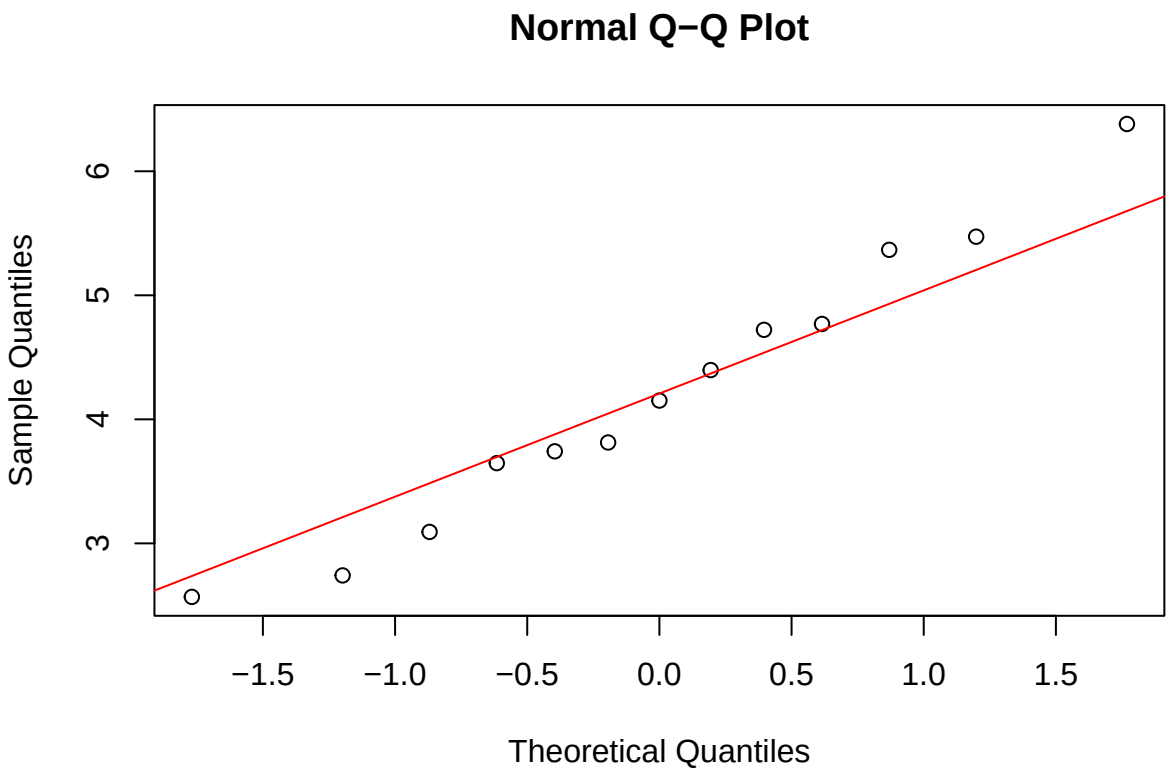
```
##
##  One Sample t-test
##
## data:  y
## t = -2.5102, df = 12, p-value = 0.0137
## alternative hypothesis: true mean is less than 5
## 95 percent confidence interval:
##      -Inf 4.773928
## sample estimates:
## mean of x
##   4.22037
```

Unlike those from `prop.test`, the outputs of `t.test` will exactly match the formulas.

Checking Assumptions

Remember that the t-test assumes that the data come from an approximately normal distribution. Let's circle back with a normal quantile plot to see if this is reasonable for the data.

```
qqnorm(y)
qqline(y, col = 'red')
```



The pattern in this plot is close to a line and so normality seems to be a good assumption. (Of course, this data was generated from a normal distribution, so it's definitely going to be normal. But with real data, you never know what you might see!)

Exercises

Example: Mushrooms

Use the data in `mushrooms.csv` for the following exercises.

a. Create a contingency table (two-way table) summarizing the variables `edible` ("e" for edible and "p" for poisonous) and `bruises` ("t" for true, the species bruises and "f" for false, the species does not bruise.) What proportion of species that bruise are edible?

```r
mushrooms <- read.csv("mushrooms.csv")

cont_tbl <- as.table(rbind(c(219, 425 - 219), c(175, 425 - 175)))
rownames(cont_tbl) <- c("edible", "poisonous")
colnames(cont_tbl) <- c("bruised", "not bruised")
cont_tbl
```

```
##           bruised not bruised
## edible        219         206
## poisonous     175         250
```

```r
prop.table(cont_tbl, margin = 1)
```

```
##             bruised not bruised
## edible    0.5152941   0.4847059
## poisonous 0.4117647   0.5882353
```

Answer: Approximately 51.53% of mushroom species that bruise are edible.

b. Create an interpret a 90% confidence interval for the difference between the proportion of mushrooms that bruise that are edible (group 1) and the proportion of mushrooms that do not bruise that are edible.

```r
prop.test(cont_tbl, conf.level = 0.9)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  cont_tbl
## X-squared = 8.7477, df = 1, p-value = 0.0031
## alternative hypothesis: two.sided
## 90 percent confidence interval:
##  0.04521272 0.16184610
## sample estimates:
##    prop 1    prop 2
## 0.5152941 0.4117647
```

Answer: We are 90% confident that $\pi_1 - \pi_2$ is between 0.04521272 and 0.16184610. Since all values in this interval are greater than zero, we can conclude that there is evidence to suggest that the proportion of mushrooms that bruise that are edible is greater than the proportion of mushrooms that do not bruise that are edible.

c. Now consider `gill.size` as a grouping variable. Gill.size is equal to "b" for mushrooms with broad gills and "n" for mushrooms with narrow gills. Conduct a hypothesis test, using $\alpha = 0.05$, to assess whether the data provide strong evidence that that mushrooms with narrow gills are less likely to bruise than those with broad gills. State the test statistic, p-value, and conclusion of your test.

```r
cont_tbl2 <- as.table(rbind(c(137, 425 - 137), c(175, 425 - 175)))
rownames(cont_tbl2) <- c("narrow", "broad")
colnames(cont_tbl2) <- c("bruised", "not bruised")
cont_tbl2
```

```
##        bruised not bruised
## narrow     137         288
## broad      175         250
```

```
prop.table(cont_tbl2, margin = 1)
```

```
##          bruised not bruised
## narrow 0.3223529   0.6776471
## broad  0.4117647   0.5882353
```

```
results <- prop.test(cont_tbl2, conf.level = 0.95)
results
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  cont_tbl2
## X-squared = 6.9324, df = 1, p-value = 0.008465
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.15629179 -0.02253174
## sample estimates:
##    prop 1    prop 2
## 0.3223529 0.4117647
```

```
test_statistic <- sqrt(results$statistic)
test_statistic
```

```
## X-squared
##  2.632951
```

```
p.val <- results$p.value
p.val
```

```
## [1] 0.00846466
```

Answer: See above for the hypothesis test assessing whether the data provide strong evidence that that mushrooms with narrow gills are less likely to bruise than those with broad gills. It has a test statistic of 2.632951 and a p-value of 0.00846466. Based on the results, we reject the null hypothesis. It seems there is strong evidence that mushrooms with narrow gills are not less likely to bruise than those with broad gills.

Example: Promotion

A political research group asked a random sample of 200 homeowners and asked if they plan to vote for Candidate A. 36% of the volunteers planned to vote for A. They then asked a random sample of 200 non-homeowners, and they found that 43% planned to vote for Candidate A.

a. Think about the data collection in this scenario, and how it differs from other similar scenarios. Convince yourself that using the inference on two proportions procedures is reasonable in this case.

b. Is there a significant difference in levels of support for candidate A across the two groups? Perform a large sample test. Use $\alpha = 0.05$. Report the test statistic, p-value, and conclusion.

```
x1 <- 72
x2 <- 86
n1 <- 200
n2 <- 200

results1 <- prop.test(x = c(x1, x2), n = c(n1, n2), conf.level = 0.95, alternative = 'two.sided')
results1
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(x1, x2) out of c(n1, n2)
## X-squared = 1.768, df = 1, p-value = 0.1836
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.17056709  0.03056709
## sample estimates:
## prop 1 prop 2
##   0.36   0.43
```

```
test_stat <- sqrt(results1$statistic)
test_stat
```

```
## X-squared
##  1.329649
```

```
p.vals <- results1$p.value
p.vals
```

```
## [1] 0.1836338
```

Answer: See above for the hypothesis test assessing the significance in difference between the two groups. The test statistic is 1.329649 and the p-value is 0.1836338. Based on the results, we fail to reject the null hypothesis. There is strong evidence that suggests there is not a significant difference in levels of support for candidate A across the two groups.

c. Find a large sample 95% confidence interval for the the difference in support for Candidate A across the groups. Does the interval contain 0? Is this consistent with your finding in part (b)?

```
tbl <- as.table(rbind(c(72, 200 - 72), c(86, 200 - 86)))
rownames(tbl) <- c("homeowners", "non-homeowners")
colnames(tbl) <- c("support", "no support")
tbl
```

```
##                support no support
## homeowners          72        128
## non-homeowners      86        114
```

```r
prop.table(tbl, margin = 1)
```

```
##                support no support
## homeowners        0.36       0.64
## non-homeowners    0.43       0.57
```

```r
prop.test(tbl, conf.level = 0.95)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  tbl
## X-squared = 1.768, df = 1, p-value = 0.1836
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.17056709  0.03056709
## sample estimates:
## prop 1 prop 2
##   0.36   0.43
```

Answer: Looking at the confidence interval above, it indeed does contain zero. This is also consistent with my finding in part (b).

Example: Prices

Use the Airbnb data from the file `Airbnb_NOLA.csv`. These represent a simple random sample of active Airbnb listings in New Orleans in August 2018.

a. Find a 99% confidence interval for the average price of Airbnb listing in New Orleans in August 2018.

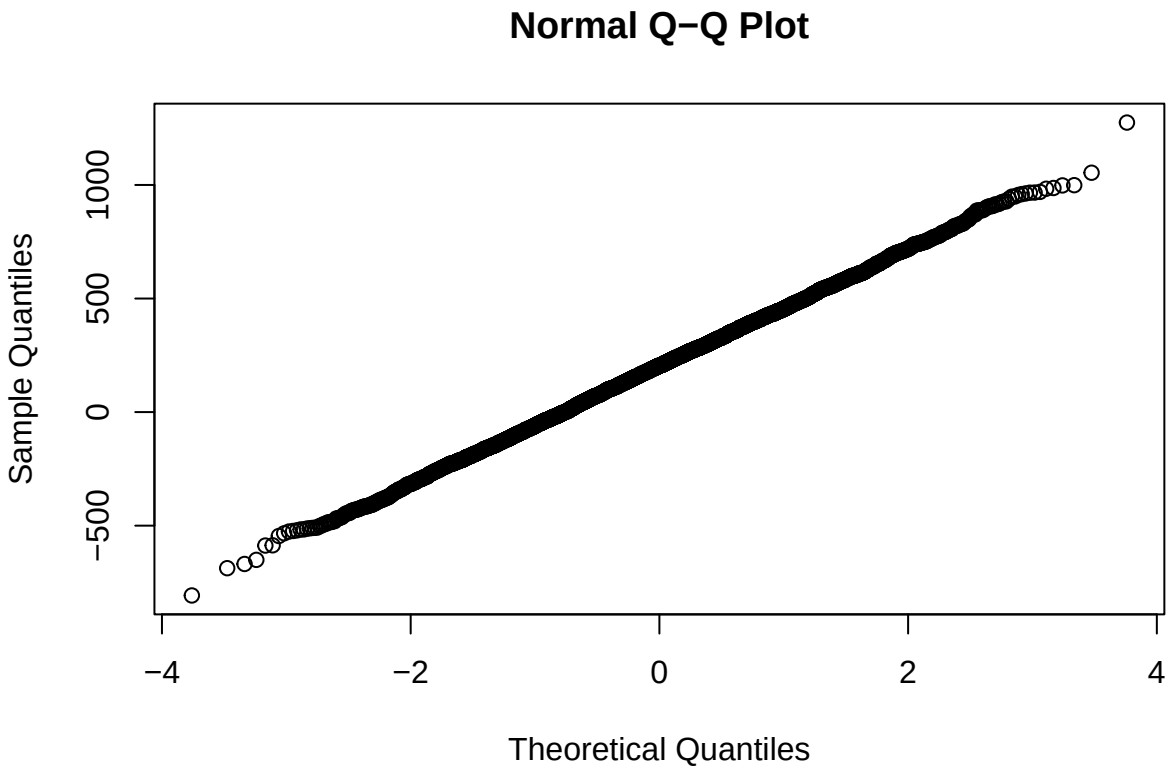
```r
NOLA <- read.csv("Airbnb_Listings_NOLA.csv")

n <- 5878
avg <- mean(NOLA$Price)
st.dev <- sd(NOLA$Price)

y <- rnorm(n, avg, st.dev)

t.test(y, conf.level = 0.99)
```

```
##
##  One Sample t-test
##
## data:  y
## t = 59.683, df = 5877, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
##  192.4616 209.8294
## sample estimates:
## mean of x
##  201.1455
```

Answer: The confidence interval for the average price of Airbnb listings in New Orleans in August 2018 is (188.1952, 205.2516).

b. Make a normal quantile plot of the data. Does the sample appear to be approximately normal? Based on this answer and the sample size, are you concerned with the validity of your results?

```
qqnorm(y)
```



Answer: See above for the normal quantile plot of the Airbnb data. This sample does appear to be approximately normal. Based on the shape of the plot, as well as the sample size of 5,878, I am not concerned with the validity of the results.

c. Take the natural log of price. Use this transformed data to find a 95% confidence interval for the log of the average Airbnb price.

```
ln_price <- log(NOLA$Price)
mean <- mean(ln_price)
standard_dev <- sd(ln_price)
n <- 5878

y1 <- rnorm(n, mean, standard_dev)

t.test(y1, conf.level = 0.95)
```

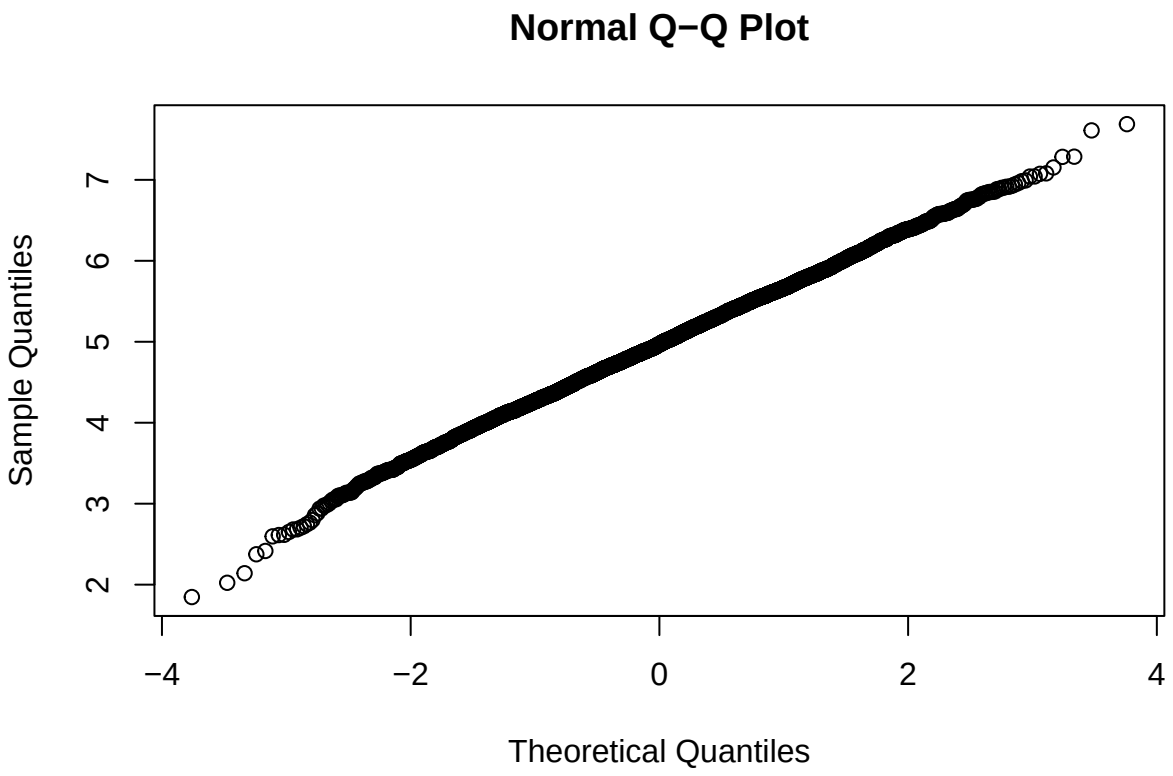
```
##
```

```
##  One Sample t-test
##
## data:  y1
## t = 538.19, df = 5877, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  4.953525 4.989744
## sample estimates:
## mean of x
##  4.971635
```

Answer: The confidence interval for the log of the average Airbnb price is (4.953525, 4.989744).

d. Make a normal quantile plot of the natural log of price. Does the sample appear to be approximately normal?

```
qqnorm(y1)
```



Answer: See above for the normal quantile plot of the Airbnb data. This sample does appear to be approximately normal.