USP 634 Lab 5- Hypothesis Testing

Jamaal Green

April 26, 2016

```
#set your working directory
cali <- read.csv("californiatod.csv", header = TRUE)
```

Learning Objectives

1. Finish Simulation work
2. Determining normality using QQ-Plot
3. Run t-tests

I. Simulation (continued)

Remeber we simulated a single dice roll:

sample(1:6, 1)

Check whether a single roll results in a 4:

sample(1:6, 1) == 4

Roll two die and check if we get a 4 and a 6:

setequal(c(sample(1:6,1), sample(1:6,1)), c(4,6))

Repeat the simulation 10000 times:

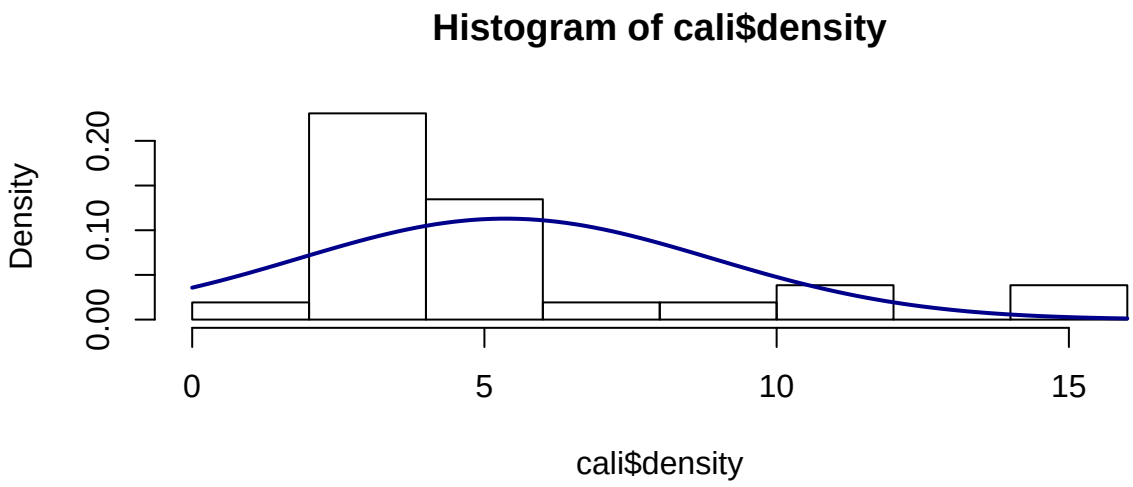
replicate(10000,setequal(c(sample(1:6,1), sample(1:6,1)), c(4,6)))

II. QQ Plot

Recall that superimposed a normal curve on top of a histogram to get a rough idea of whether a variable is normally distributed.

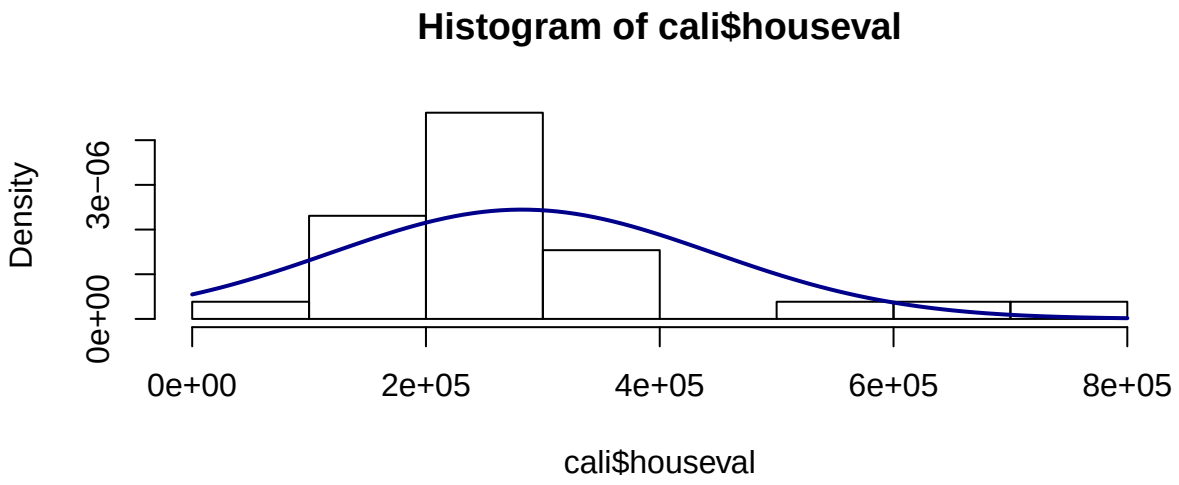
First, we will create a histogram for density and superimpose a curve on top of it.

```
hist(cali$density, freq= FALSE)
curve(dnorm(x, mean=mean(cali$density), sd = sd(cali$density)), lwd=2, col='darkblue', add = TRUE)
```



Now for the housing value variable:

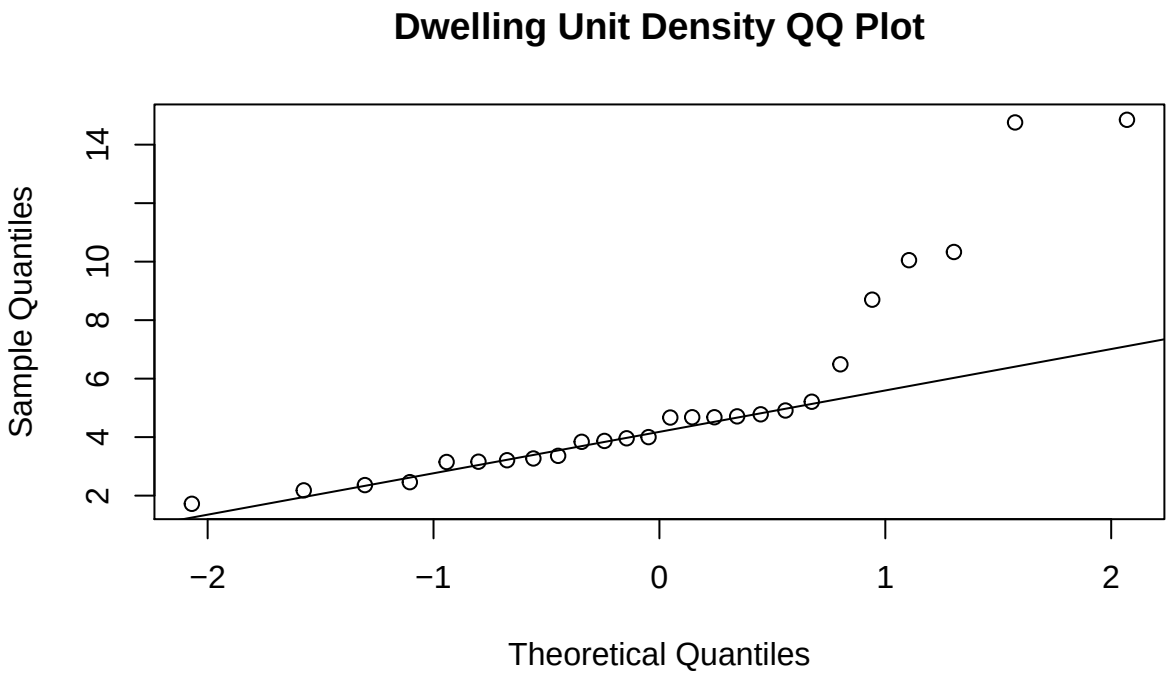
```
hist(cali$houseval, freq= FALSE)
curve(dnorm(x, mean=mean(cali$houseval), sd = sd(cali$houseval)), lwd=2, col='darkblue', add = TRUE)
```



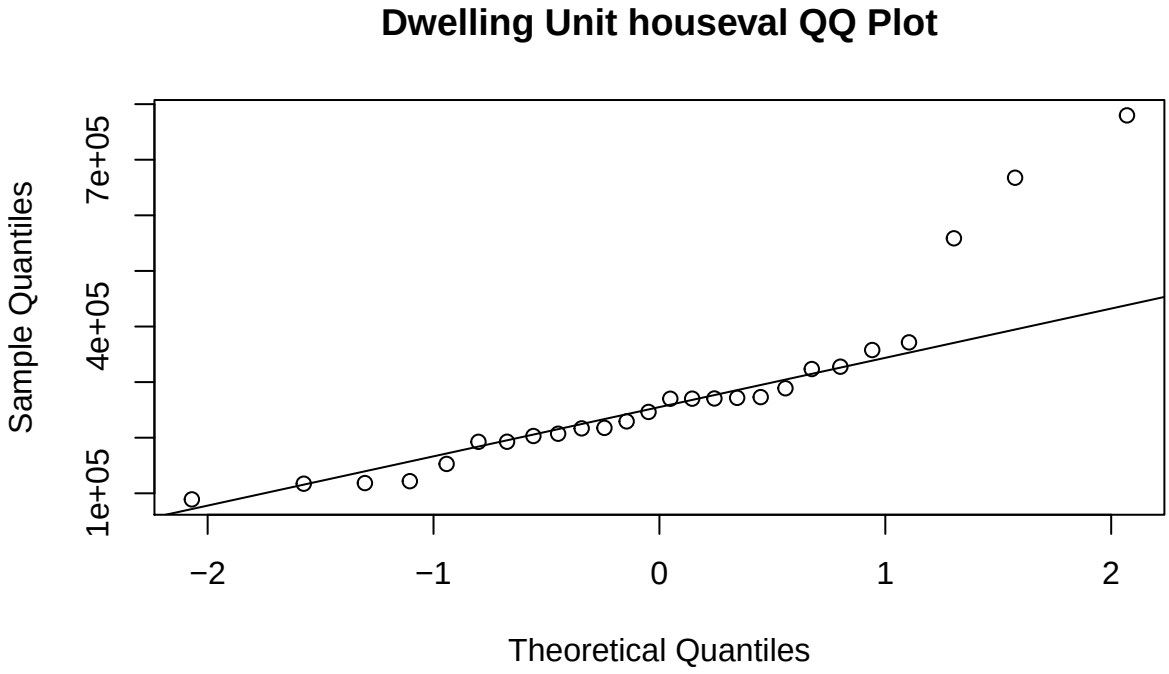
There are better ways to do this. One wya to determine if a variable is normally distributed is to use a QQ plot. You can use the **qqnorm()** function to create a Quantile-Quantile plot evaluating the fit of a sample data to a normal distribution. More generally, the **qqplot()** function creates a Quantile-Quantile plot for any theoretical distribution, **qqline()** draws a diagonal line for the fit of a theoretical distribution.

Let's create a qqplot for the density and house value variables:

```
qqnorm(cali$density, main =  "Dwelling Unit Density QQ Plot")
qqline(cali$density)
```




```
qqnorm(cali$houseval,main =  "Dwelling Unit houseval QQ Plot")
qqline(cali$houseval)
```



Which of these two variables are more likely to be normally distributed?

III. T-Tests

Conduct single sample hypothesis test by hand

We will not compute a **single sample hypothesis test** with a known mean by hand. We shall test whether the mean housing value in the TOD dataset is statistically significantly different from the mean of all housing

on the market (assuming mean = 250000).

The formular R uses for computing the test statistic is:

$$\frac{\bar{x} - \mu}{\frac{s}{\sqrt{n-1}}}$$

1. Find the mean and standard deviation of the variable houseval:

```
xbar <- mean(cali$houseval)
mu <- 250000
s <- sd(cali$houseval)
n <- nrow(cali)
```

2. Now use R as a calculator to compute the t-statistic by hand. Enter the correct numerical values to compute this statement:

```
(sample_mean - population_mean)/(sample_sd/(square_root(n-1)))
```

Replace each term with the correct corresponding numerical value. Note that to designate a square root you can use **sqrt()**.

```
tstar <- (xbar - mu)/(s/sqrt(n-1))
```

3. You can get the p-value with

```
p <- 2*pt(-abs(tstar), df = 26-1)
p
```

Question: What is the null hypothesis for this test? Should you reject the null hypothesis based on the p-value?

Compute the single sample t-test in R.

```
t.test(cali$houseval, alternative = 'two.sided', mu = 250000, conf.level = .95)
```

```
##
##  One Sample t-test
##
## data:  cali$houseval
## t = 1.0053, df = 25, p-value = 0.3244
## alternative hypothesis: true mean is not equal to 250000
## 95 percent confidence interval:
##  216280.3 348029.1
## sample estimates:
## mean of x
##  282154.7
```

Compare the results to our hand calculations.

Compute a group-mean t-test

The formula R uses for computing the test statistics is:

$$\frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

1. Find the mean and standard deviation for the variable houseval

```
xbar1 <- with(cali, mean(houseval[railtype == "Light rail"]))

xbar2 <- with(cali, mean(houseval[railtype == "Heavy or commuter rail"]))
```

```
s1 <- with(cali, sd(houseval[railtype == "Light rail"]))

s2 <- with(cali, sd(houseval[railtype == "Heavy or commuter rail"]))

n1 <- with(cali, length(houseval[railtype == "Light rail"]))

n2 <- with(cali, length(houseval[railtype == "Heavy or commuter rail"]))
```

2. Now use R as a calculator to compute the t-statistic by hand. Enter the correct numerical values to complete the formula:

$$\frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Replace each term with the correct corresponding numerical value. Note that to designate a square root use the **sqrt()** function.

```
tstar <- (xbar1 - xbar2)/ sqrt(s1^2/n1 + s2^2/n2)
```

3. You can get the p-value for the two-sided test with

```
p <- 2*pt(-abs(tstar)), df =  pmin(n1,n2)-1

p
```

Theoretically the formula is not exactly appropriate for our small sample size, but for now it shall suffice. You can use this same equation for calculating the p-value in Assignment 2.

Compute the single sample t-test in R

The general syntax for the t-test in R is

t.test(var1 ~ var2, data)

In this var1 is a continuous variable and var2 is a categorical variable.

Execute the test by placing the proper variable names into the generic command (in this case *houseval* and *railtype*)

```
t.test(houseval ~ railtype, data = cali)
```

```
##
##  Welch Two Sample t-test
##
## data:  houseval by railtype
## t = 3.4087, df = 23.927, p-value = 0.002314
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   65647.09 267225.93
## sample estimates:
## mean in group Heavy or commuter rail             mean in group Light rail
##                              339767.3                             173330.8
```

Can we reject the null hypothesis given these results?

- The group-means t-test (like many statistical tests you will learn) measures the probability that the null hypothesis can be rejected. For a two-tailed test, the typical null hypothesis is that the population means are the same for two different groups. Rejecting the null is the same as concluding that the population means are different.

- The null hypothesis is $H_0 : diff = 0$, or in other words, there is not a difference in the mean density by railtype in the population from which the sample is drawn. And the alternative hypothesis is "the true difference in means is not equal to 0"
- R accepts arguments for alternative hypotheses and confidence levels. The alternative argument can be 'two-sided' for a two-tailed hypothesis (mean1 = mean2) as well as two one-tailed hypotheses ('greater' for mean1 > mean2, and 'less' for mean1 < mean2)

Twosided versus one-sided

By default, the **t.test()** function is a two-tailed test, i.e. the alternative hypothesis is that the difference in means is not equal to zero. Put another way, the alternative hypothesis is that there is a difference in average density between railtypes. We can switch to one-sided tests by changing the 'alternative' argument, an example follows

```
#Default t.test()
t.test(houseval ~ railtype, data = cali, alternative = 'two.sided', conf.level = .95)
```

```
##
##  Welch Two Sample t-test
##
## data:  houseval by railtype
## t = 3.4087, df = 23.927, p-value = 0.002314
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   65647.09 267225.93
## sample estimates:
## mean in group Heavy or commuter rail             mean in group Light rail
##                             339767.3                             173330.8
```

```
#one sided t.test()

t.test(houseval ~ railtype, data = cali, alternative = 'greater', conf.level = .95)
```

```
##
##  Welch Two Sample t-test
##
## data:  houseval by railtype
## t = 3.4087, df = 23.927, p-value = 0.001157
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  82889.75      Inf
## sample estimates:
## mean in group Heavy or commuter rail             mean in group Light rail
##                             339767.3                             173330.8
```

In order to get the p-value for one-sided t distribution

pt(-abs(tstar), df = length(cali) - 1)

The output for the two-tailed test has a p-value of .001 and thus we would reject the null hypothesis that the difference between our group means in 0.

Paired t-test

Note that R also allows you to compute a **two-sample t-test**, appropriate in situations where you have "matched pairs" data available in separate variables. For example, in the Krizek paper we read this week, he tests whether the means of variables differ significantly pre and post-move (treatment). You can do a paired t-test in R!

The general form of the syntax is:

t.test(var1, var2, data = Dataset, alternative = 'two.sided', conf.level = .95, paired = TRUE)

This is the type of t-test you want to conduct for Question 4 in Assigment 2.