Normal Distribution Tutorial

2022-11-16

Learning Goals

- Visualize properties of the normal distribution.
- Understand the Central Limit Theorem.
- Calculate sampling properties of sample means.
- Decide whether a data set likely comes from a normal distribution

Tutorial

This lab mainly focuses on some exercises to better understand the nature of the normal distribution. We will also learn a couple of tools that help us decide whether a particular data set is likely to have come from population with an approximately normal distribution.

Many statistical tests assume that the variable being analyzed has a normal distribution. Fortunately, many of these tests are fairly robust to this assumption—that is, they work reasonably well even when this assumption is not quite true, especially when sample size is large. Therefore it is often sufficient to be able to assess whether the data come from a distribution whose shape is even approximately normal (the bell curve).

Draw a histogram

A good way to start is to simply visualize the frequency distribution of the variable in the data set by drawing a histogram. Let's use the age of passengers on the Titanic for our example.

```
#read in titanic
titanicData <- read.csv("input_files/titanic.csv", stringsAsFactors = TRUE)
#have a look at the data
head(titanicData)
```

```
##   passenger_class                                      name     age    embarked
## 1             1st                 Allen,MissElisabethWalton 29.0000 Southampton
## 2             1st                  Allison,MissHelenLoraine  2.0000 Southampton
## 3             1st           Allison,MrHudsonJoshuaCreighton 30.0000 Southampton
## 4             1st Allison,MrsHudsonJ.C.(BessieWaldoDaniels) 25.0000 Southampton
## 5             1st                Allison,MasterHudsonTrevor  0.9167 Southampton
## 6             1st                          Anderson,MrHarry 47.0000 Southampton
##              home_destination    sex survive
## 1                  StLouis,MO female     yes
## 2 Montreal,PQ/Chesterville,ON female      no
## 3 Montreal,PQ/Chesterville,ON   male      no
## 4 Montreal,PQ/Chesterville,ON female      no
## 5 Montreal,PQ/Chesterville,ON   male     yes
## 6                  NewYork,NY   male     yes
```

```
str(titanicData)
```

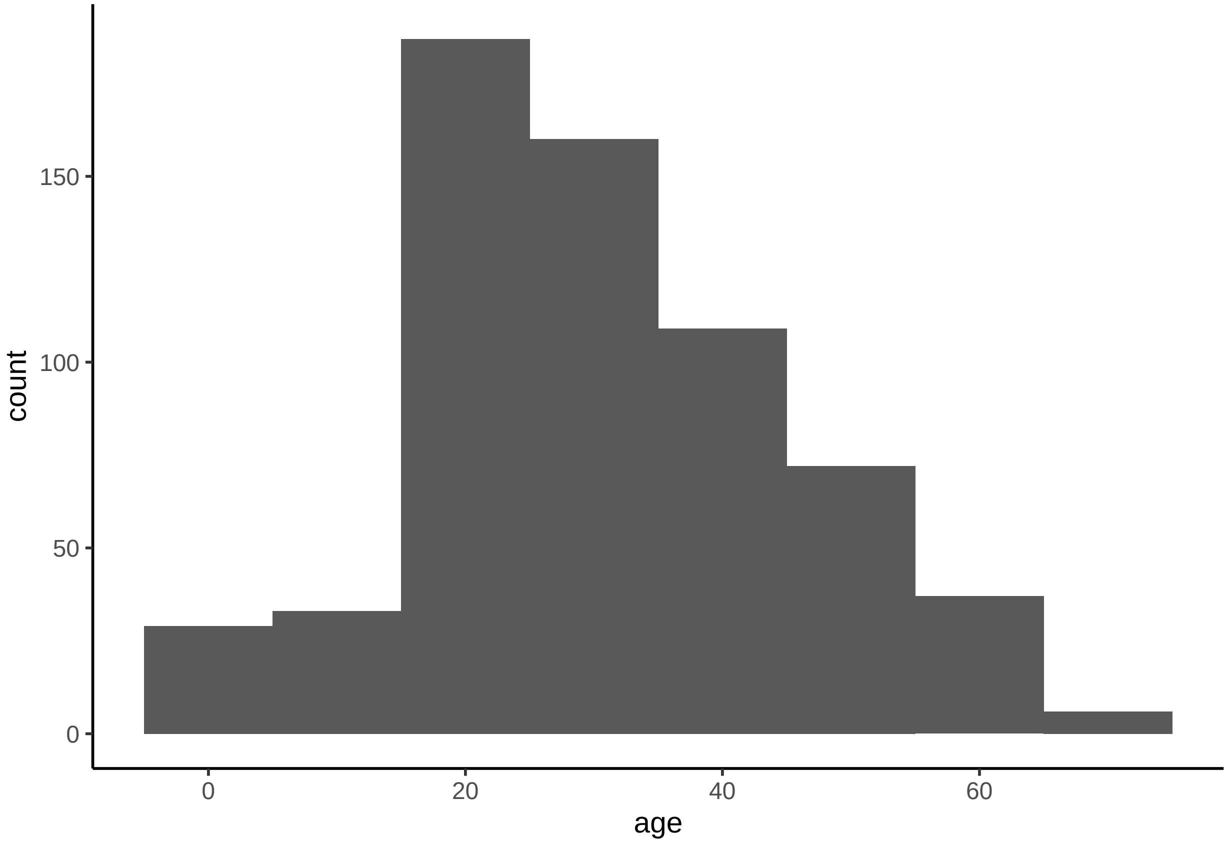
```
## 'data.frame':    1313 obs. of  7 variables:
```

```
##  $ passenger_class : Factor w/ 3 levels "1st","2nd","3rd": 1 1 1 1 1 1 1 1 1 1 ...
##  $ name            : Factor w/ 1310 levels "Abbing,MrAnthony",..: 22 25 26 27 24 31 45 46 50 54 ...
##  $ age             : num  29 2 30 25 0.917 ...
##  $ embarked        : Factor w/ 4 levels "","Cherbourg",..: 4 4 4 4 4 4 4 4 4 2 ...
##  $ home_destination: Factor w/ 372 levels "","?Havana,Cuba",..: 314 234 234 234 234 241 165 26 24 23
##  $ sex             : Factor w/ 2 levels "female","male": 1 1 2 1 2 2 1 2 1 2 ...
##  $ survive         : Factor w/ 2 levels "no","yes": 2 1 1 1 2 2 2 1 2 1 ...
```

Remember we can use ggplot() to draw histograms.

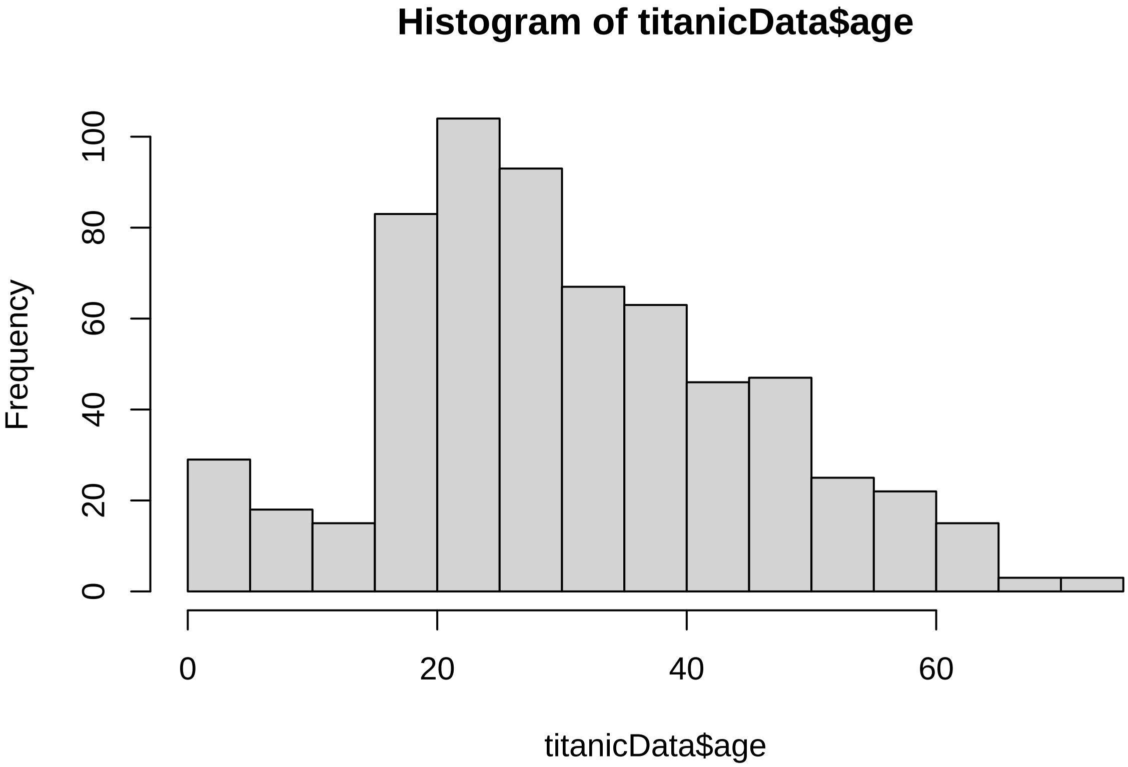
```
#use ggplot to make a histogram of age
ggplot(titanicData, aes(x = age)) +
  geom_histogram(binwidth = 10) +
  theme_classic()
```

```
## Warning: Removed 680 rows containing non-finite values (`stat_bin()`).
```



If we are just drawing a histogram for ourselves to better understand the data, it is even easier to just use a function from base R, hist(). Give hist() a vector of data as input, and it will print a histogram in the plots window.

```
hist(titanicData$age)
```



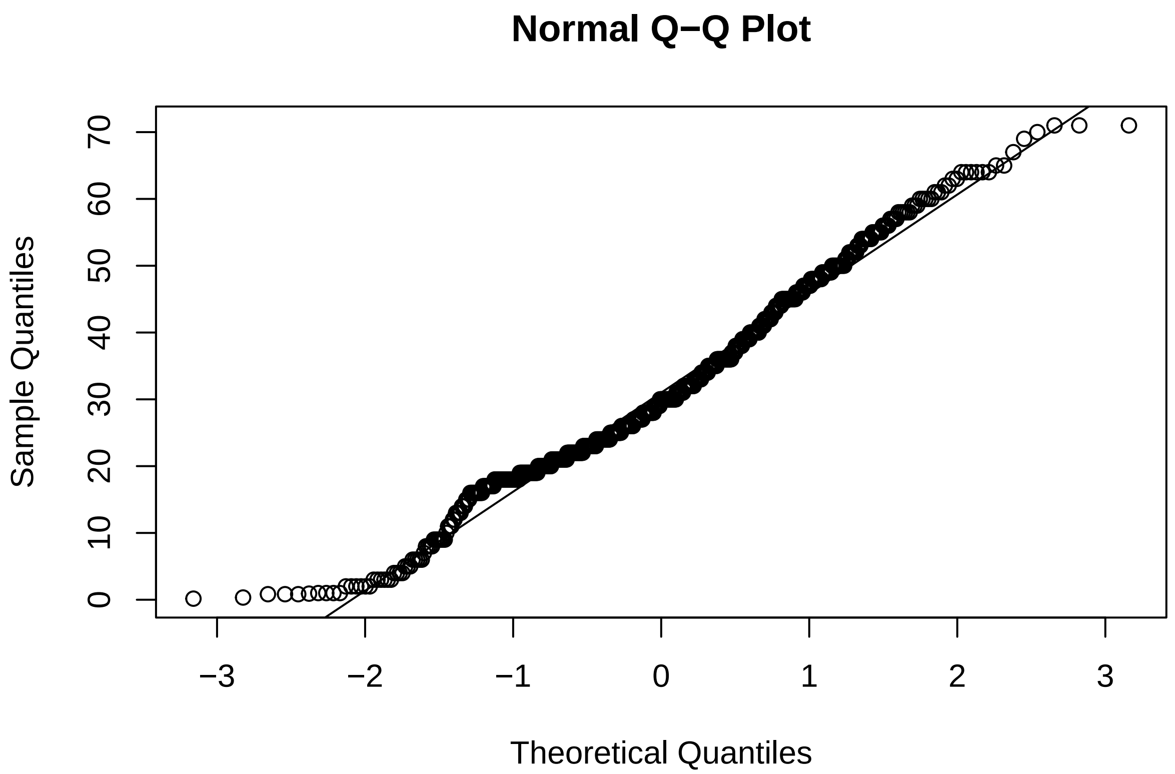
Looking at this histogram, we see that the frequency distribution of the variable is not exactly normal; it is slightly asymmetric and there seems to be a second mode near 0. On the other hand, like the normal distribution, the frequency distribution has a large mode near the center of the distribution, frequencies mainly fall off to either side, and there are no outliers. This is close enough to normal that most methods would work fine.

QQ plot

Another graphical technique that can help us visualize whether a variable is approximately normal is called a quantile plot (or a QQ plot). The QQ plot shows the data on the vertical axis ranked in order from smallest to largest ("sample quantiles" in the figure below). On the horizontal axis, it shows the expected value of an individual with the same quantile if the distribution were normal ("theoretical quantiles" in the same figure). The QQ plot should follow more or less along a straight line if the data come from a normal distribution (with some tolerance for sampling variation).

QQ plots can be made in R using a function called qqnorm(). Simply give the vector of data as input and it will draw a QQ plot for you. (qqline() will draw a line through that Q-Q plot to make the linear relationship easier to see.)

```
qqnorm(titanicData$age)
qqline(titanicData$age)
```



This is what the resulting graph looks like for the Titanic age data. The dots do not land along a perfectly straight line. In particular the graph curves at the upper and lower end. However, this distribution definitely would be close enough to normal to use most standard methods, such as the t-test.

It is difficult to interpret QQ plots without experience. One of the goals of today's exercises will be to develop some visual experience about what these graphs look like when the data is truly normal. To do that, we will take advantage of a function built into R to generate random numbers drawn from a normal distribution. This function is called rnorm().

rnorm()

The function rnorm() will return a vector of numbers, all drawn randomly from a normal distribution. It takes three arguments:

n: how many random numbers to generate (the length of the output vector)

mean: the mean of the normal distribution to sample from

sd: the standard deviation of the normal distribution

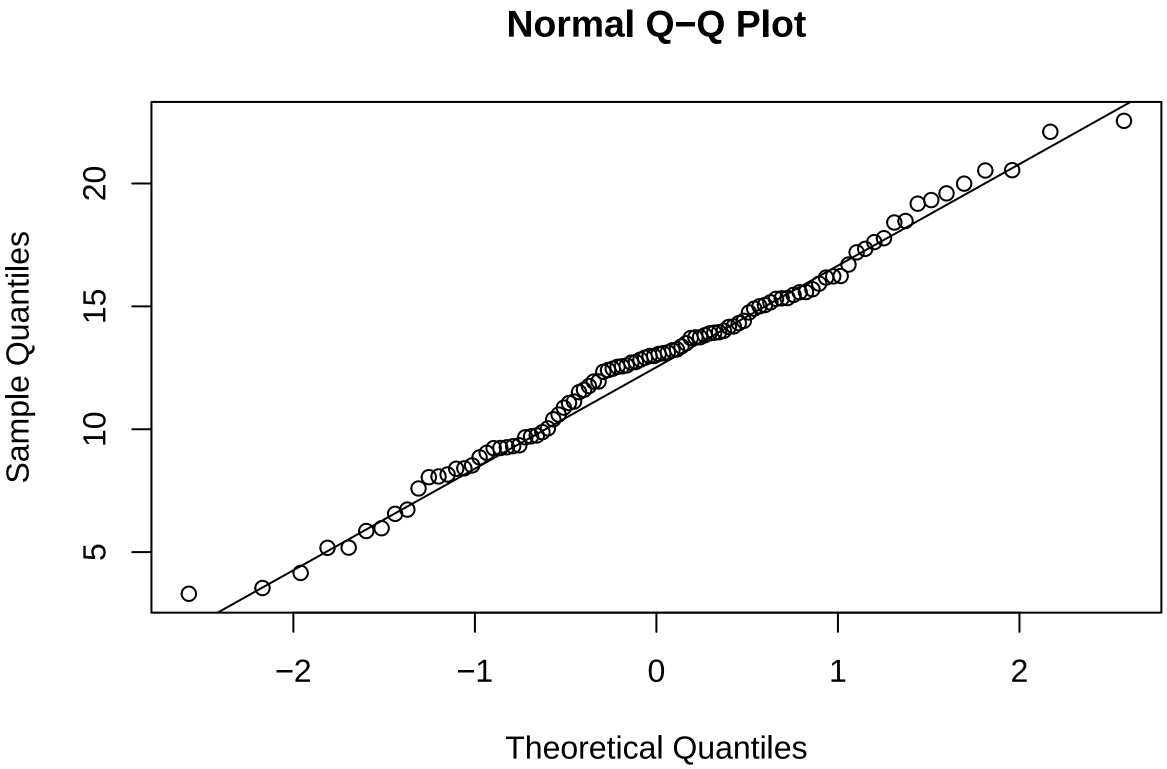
For example, the following command will give a vector of 20 random numbers drawn from a normal distribution with mean 13 and standard deviation 4:

```
rnorm(n = 20, mean = 13, sd = 4)
```

```
##  [1]  9.591368 10.756576 14.122650 10.421640 11.473464 13.409227 15.726861
##  [8] 11.256788 17.167426 17.108852 11.714879 10.696451 13.316005 15.579084
## [15] 14.136816 13.412374 13.881502 18.608371 11.115633 13.871874
```

Let's look at a QQ plot generated from 100 numbers randomly drawn from a normal distribution:

```
normal_vector <- rnorm(n = 100, mean = 13, sd = 4)
qqnorm(normal_vector)
qqline(normal_vector)
```



These points fall mainly along a straight line, but there is some wobble around that line even though these points were in fact randomly sampled from a known normal distribution. With a QQ plot, we are looking for an overall pattern that is approximately a straight line, but **we do not expect a perfect line**.

In the exercises, we'll simulate several samples from a normal distribution to try to build intuition about the kinds of results you might get.

When data are not normally distributed, the dots in the quantile plot will not follow a straight line, even approximately. For example, here is a histogram and a QQ plot for the population size of various counties, from the data in countries.csv. These data are very skewed to the right, and do not follow a normal distribution at all.

Transformations

When data are not normally distributed, we can try to use a simple mathematical transformation on each data point to create a list of numbers that still convey the information about the original question but that may be better matched to the assumptions of our statistical tests. Chapter 13 of Whitlock and Schluter has more info on this, but for now let's learn how to do one of the most common data transformations, the log-transformation.

With a transformation, we apply the same mathematical function to each value of a given numerical variable for individual in the data set. With a log-transformation, we take the logarithm of each individual's value for a numerical variable.

- We can only use the log-transformation if all values are greater than zero.
- Also, it will only improve the fit of the normal distribution to the data in cases when the **frequency distribution of the data is right-skewed.**

To take the log transformation for a variable in R is very simple. We simply use the function `log()`, and apply it to the vector of the numerical variable in question. For example, to calculate the log of age for all passengers on the Titanic, we use the command:

```
#eval=F so we don't print three pages of values in the pdf, but you should run it
log(titanicData$age)
```

This will return a vector of values, each of which is the log of age of a passenger.

Activity

We return to the applet we used in tutorial 3, located at http://www.zoology.ubc.ca/~whitlock/Kingfisher/SamplingNormal.htm to investigate three points made in the text.

- Point 1: The distribution of sample means is normal, if the variable itself has a normal distribution. First, hit "COMPLETE SAMPLE OF 10" and "CALCULATE MEAN" a few times, to remind yourself of what this applet does. (It takes a sample from the normal distribution shown when you click "SHOW POPULATION". The top panel shows a histogram of that sample, and the bottom panel shows the distribution of sample means from all the previous samples.)

Next, hit the "MEANS FOR MANY SAMPLES" button. This button makes a large number of separate samples at one go, all of the same sample size, to save you from making the samples one by one. Notice that the sample mean differs from sample to sample. The sample mean produced by random sampling from a probability distribution is itself a random variable.

Look at the distribution of sample means. Does it seem to have a normal distribution? Click the checkbox by "SHOW SAMPLING DISTRIBUTION" off to the right, which will draw the curve for a normal distribution with the same mean and variance as this distribution of sample means.

- Point 2: The standard deviation of the distribution of sample means is reduced with larger sample sizes. The standard deviation of the population is controlled by the right slider marked with σ. The sample size is set by left slider. (The default when it opens is set to n=10.)

For n=10, have the applet calculate a large number of sample means as you did in the previous exercise. If each sample size is 10 and the standard deviation is 30 (as in the default), what do you predict the standard deviation of the sample means to be? (Use the equation you have learned in class to make this calculation.)

Change the sample size to n=100, and recalculate many sample means. Calculate the predicted standard deviation of all the sample means. Should the sampling distribution of sample means be wider or narrower with this larger sample size than in the previous case with a smaller sample size? What do you observe in the simulations?

- Point 3: The distribution of sample means is approximately normal no matter what the distribution of the variable, as long as the sample size is large enough. (The Central Limit Theorem) Load another web page: http://www.zoology.ubc.ca/~whitlock/Kingfisher/CLT.htm

This will simulate a very skewed distribution of data, showing the number of cups of coffee drunk per week for a population of university students. Click on "COFFEE" to see the distribution of the variable among individual students. Describe the ways that this looks different from a normal distribution.

Set $n = 2$ for the sample size, and simulate many sample means. Does the distribution of sample means look normal? Is it closer to normal in its shape than the distribution of individuals in the population?

Now set $n = 25$ and simulate many sample means. How does the distribution of sample means look now? It should look much more like a normal distribution, because of the Central Limit Theorem.