Central Limit Theorem

MATH 3512, BCIT

Matrix Methods and Statistics for Geomatics

November 19, 2018

Sampling Distributions

So far we have had random variables X which gave us the numerical results of a random process. Now let's consider a special kind of random variable Y. Let's say you roll a die three times and

$$\left.\begin{array}{rcl} X & = & 3 \\ X & = & 4 \\ X & = & 1 \end{array}\right\} Y = \frac{3+4+1}{3} = \frac{8}{3}$$

Let's do it again, and we get a different random outcome

$$\left.\begin{array}{rcl} X & = & 5 \\ X & = & 5 \\ X & = & 2 \end{array}\right\} Y = \frac{5+5+2}{3} = \frac{12}{3}$$

Sampling Distributions

The distribution of Y is called a sampling distribution. It can be based on

1. the mean of a sample
2. the variance of a sample
3. the proportion of a sample

There is therefore a sampling distribution of the mean, a sampling distribution of the variance, and a sampling distribution of the proportion of a population. If I don't want to be specific about which of the three I am talking about, I will call it the sampling distribution of a statistic.

Sampling Distributions

Exercise 1: Take a deck of cards and label the cards $1 - 52$. Draw five cards with replacement. What are the values for the sampling distributions (proportion here refers to proportion of even-numbered cards)?

sample #	first	second	third	fourth	fifth	sixth
first card	7	40	18	28	46	40
second card	7	17	41	46	33	20
third card	35	1	41	19	30	48
fourth card	8	17	16	10	45	18
fifth card	36	18	7	3	22	25
mean						
variance						
proportion						

Sampling Distributions

Exercise 2: Take a deck of cards and label the cards $1 - 52$. Draw five cards with replacement. What are the values for the sampling distributions (proportion here refers to proportion of even-numbered cards)?

sample #	first	second	third	fourth	fifth	sixth
first card	7	40	18	28	46	40
second card	7	17	41	46	33	20
third card	35	1	41	19	30	48
fourth card	8	17	16	10	45	18
fifth card	36	18	7	3	22	25
mean	18.6	18.6	24.6	21.2	35.2	30.2
variance	238.3	193.3	241.3	280.7	104.7	173.2
proportion	0.4	0.4	0.4	0.6	0.6	0.8

Sampling Distributions

sample #	first	second	third	fourth	fifth	sixth
first card	7	40	18	28	46	40
second card	7	17	41	46	33	20
third card	35	1	41	19	30	48
fourth card	8	17	16	10	45	18
fifth card	36	18	7	3	22	25
mean	18.6	18.6	24.6	21.2	35.2	30.2
variance	238.3	193.3	241.3	280.7	104.7	173.2
proportion	0.4	0.4	0.4	0.6	0.6	0.8

The random variable X has a uniform distribution because the probability of any card to be drawn is $1/52$. The mean of this probability distribution is 26.5, the variance is 225.25, and the mean proportion of even cards is 0.5. How good are the estimates based on the samples? Are they biased?

Unbiased Estimators

mean The sample mean $\bar{x}$ is an **unbiased** estimator of the population mean μ because $\text{mean}(\bar{x}) \longrightarrow \mu$.

variance The sample variance s^2 is an **unbiased** estimator of the population variance σ^2 because $\text{mean}(s^2) \longrightarrow \sigma^2$.

proportion The sample proportion $\hat{p}$ is an **unbiased** estimator of the population proportion p because $\text{mean}(\hat{p}) \longrightarrow p$.

st dev The sample standard deviation s is a **biased** estimator of the population standard deviation σ because it is NOT true that $\text{mean}(s) \longrightarrow \sigma^2$.

The mean of the sample statistic is taken from all possible (m choose n) samples, where m is the population size and n is the sample size.

Sampling Distributions

Example 1: Sampling in R. First, we roll a die $m = 200$ times and plot the results. You can see the distribution, which should be approximately uniform.

```
x1<-round((runif(200,0,1)*6)+0.5)
barplot(table(x1))
```

Sampling Distributions

Next, we pick 1000 samples of these 200 die rolls (how many are there in total?) and plot their mean. For example, if the sample is "$4, 2, 5$," then the mean is $11/3$. The samples have a sample size of $n = 3$.

```
sampling<-function(x) {
t<-c()
for (i in 1:1000) {
t<-append(t,mean(sample(x,3)))
}
return(t)
}
x2<-sampling(x1)
hist(x2,breaks=seq(1,6,1),col="blue")
```

Central Limit Theorem

Given

1. The original population has mean μ and standard deviation σ.
2. Simple random samples of the same size n are selected from the population.

Then

Case 1 Original population is normally distributed.

Case 2 Original population is not normally distributed.

Central Limit Theorem

In **Case 1**, for any sample size n, the distribution of $\bar{x}$ is a normal distribution with these parameters.

1. The mean of all values of $\bar{x}$ is

$$\mu_{\bar{x}} = \mu \tag{1}$$

2. The standard deviation of all values of $\bar{x}$ is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{2}$$

3. The z-score conversion of $\bar{x}$ works according to the following formula,

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{3}$$

Central Limit Theorem

In **Case 1**, for any sample size n, the distribution of $\bar{x}$ is a normal distribution with these parameters.

1. The mean of all values of $\bar{x}$ is

$$\mu_{\bar{x}} = \mu \tag{1}$$

2. The standard deviation of all values of $\bar{x}$ is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{2}$$

3. The z-score conversion of $\bar{x}$ works according to the following formula,

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{3}$$

Central Limit Theorem

In **Case 1**, for any sample size n, the distribution of $\bar{x}$ is a normal distribution with these parameters.

1. The mean of all values of $\bar{x}$ is

$$\mu_{\bar{x}} = \mu \tag{1}$$

2. The standard deviation of all values of $\bar{x}$ is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{2}$$

3. The z-score conversion of $\bar{x}$ works according to the following formula,

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{3}$$

Central Limit Theorem

In **Case 2**, for sample sizes $n > 30$, the distribution of $\bar{x}$ is approximately a normal distribution with these parameters.

1. The mean of all values of $\bar{x}$ is

$$\mu_{\bar{x}} = \mu \tag{4}$$

2. The standard deviation of all values of $\bar{x}$ is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{5}$$

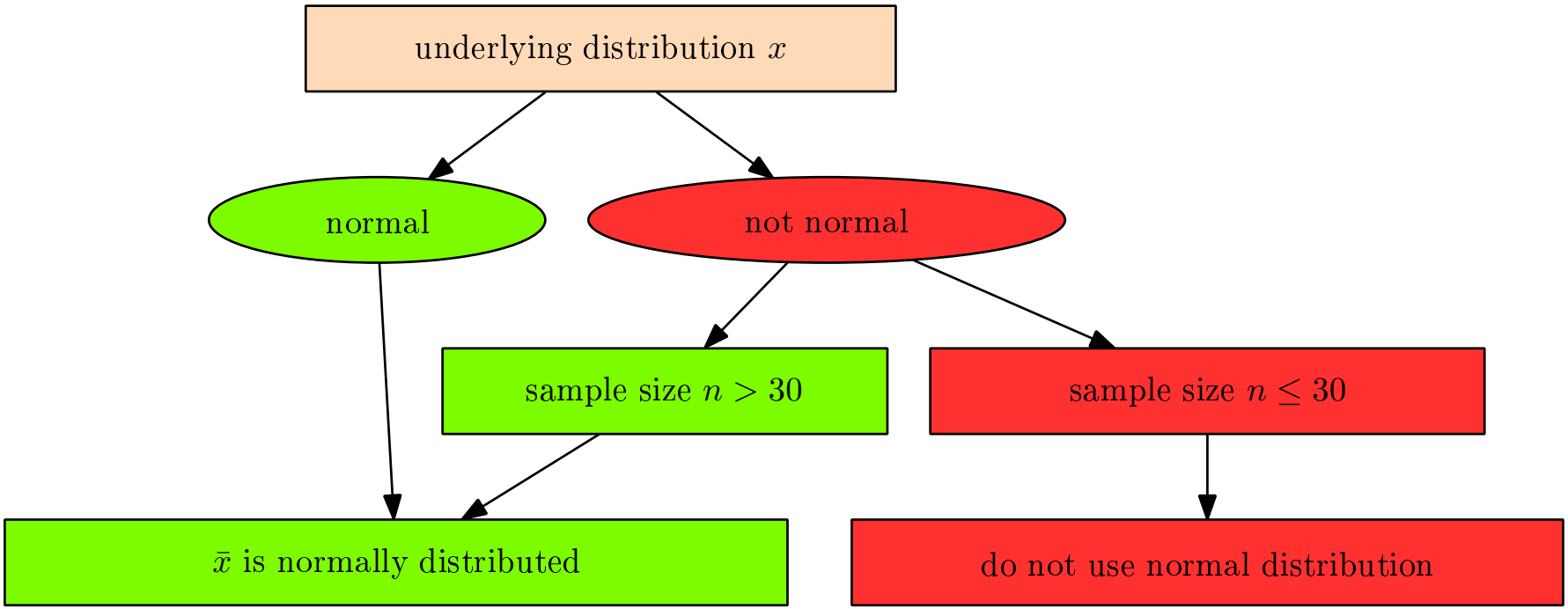
3. The z-score conversion of $\bar{x}$ works according to the following formula,

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{6}$$

Central Limit Theorem

In **Case 2**, for sample sizes $n \leq 30$, the distribution of $\bar{x}$ cannot be approximated well by a normal distribution and the methods of this section do not apply.

Central Limit Theorem Flow Chart



Central Limit Theorem Class Exercise

Look at your hardcopy of `agesheights.pdf`. Pick a line number. Consider the first n measurements of height in your line. What is their mean? What is their standard deviation? Remember the following formulas:

$$\text{mean of a sample } = \bar{x} = \frac{\sum x}{n} \tag{7}$$

$$\text{standard deviation of a sample } = s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} \tag{8}$$

Now take all the sample means and sample standard deviations in class and consider how they are distributed. Then do the same for ages. (The measurements of height were randomly generated from a normal distribution with mean $\mu = 69.5$ inches and standard deviation $\sigma = 2.4$ inches. The ages were randomly generated from the age distribution in Canada in 2017 published on the Statistics Canada website.)

Correction for a Finite Population

The Central Limit Theorem assumes that the population is infinite. We can achieve this by sampling with replacement. In practice, however, we usually sample without replacement. In this case, all is fine unless $n > 0.05m$, where m is the size of the population. If n is more than five percent of m, we introduce a correction factor so that

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{m-n}{m-1}} \tag{9}$$

Note that the standard deviation of a sampling distribution (rather than a population or a sample) is sometimes called the standard error. This standard error can be with respect to different parameters, but for our purposes it will always be about the mean of the sampling distribution.

Central Limit Theorem

1. Check Requirements. When working with the mean from a sample, verify that the normal distribution can be used by confirming that the original population has a normal distribution or $n > 30$.
2. Individual Value or Mean from a Sample? Determine whether you are using a normal distribution with a single value x or the mean $\bar{x}$ from a sample of n values.

Central Limit Theorem

Individual Values When working with an individual value from a normally distributed population, use the methods talked about earlier with

$$z = \frac{x - \mu}{\sigma} \tag{10}$$

Mean from a Sample of Values When working with a mean for some sample of n values, be sure to use the value of $\sigma/\sqrt{n}$ for the standard deviation of the sample mean, so use

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{11}$$

Exercises

Exercise 3: When designing elevators, an obviously important consideration is the weight capacity. An Ohio college student died when he tried to escape from a dormitory elevator that was overloaded with 24 passengers. The elevator was rated for a capacity of 16 passengers with a total weight of 2500 lb. The table below shows values of recent adult weight parameters. For the following, we assume a worst-case scenario in which all of the passengers are males (which could easily happen in a dormitory setting). If an elevator is loaded to a capacity of 2500 lb with 16 males, the mean weight of a passenger is 156.25 lb.

	Males	Females
μ	182.9 lb	165.0 lb
σ	40.8 lb	45.6 lb
Distribution	normal	normal

Exercises

Exercise 4: When designing elevators, an obviously important consideration is the weight capacity. An Ohio college student died when he tried to escape from a dormitory elevator that was overloaded with 24 passengers. The elevator was rated for a capacity of 16 passengers with a total weight of 2500 lb. The table below shows values of recent adult weight parameters. For the following, we assume a worst-case scenario in which all of the passengers are males (which could easily happen in a dormitory setting). If an elevator is loaded to a capacity of 2500 lb with 16 males, the mean weight of a passenger is 156.25 lb.

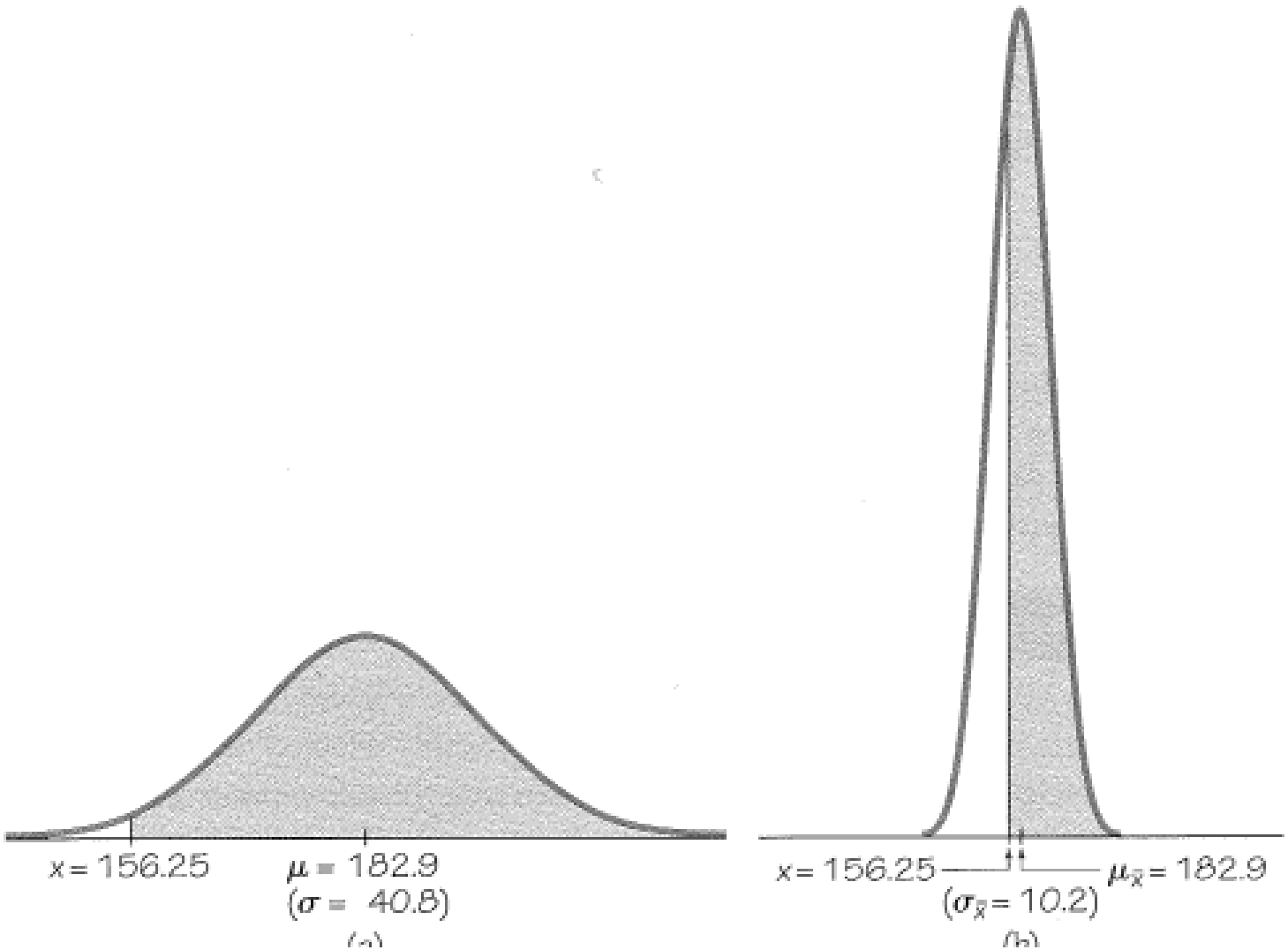
1. Find the probability that one randomly selected adult male has a weight greater than 156.25 lb.
2. Find the probability that a sample of 16 randomly selected adult males has a mean weight greater that 156.25 lb (so that the total weight exceeds the maximum capacity of 2500 lb).

Exercises

Exercise 5: When designing elevators, an obviously important consideration is the weight capacity. An Ohio college student died when he tried to escape from a dormitory elevator that was overloaded with 24 passengers. The elevator was rated for a capacity of 16 passengers with a total weight of 2500 lb. The table below shows values of recent adult weight parameters. For the following, we assume a worst-case scenario in which all of the passengers are males (which could easily happen in a dormitory setting). If an elevator is loaded to a capacity of 2500 lb with 16 males, the mean weight of a passenger is 156.25 lb.

1. Find the probability that one randomly selected adult male has a weight greater than 156.25 lb.
2. Find the probability that a sample of 16 randomly selected adult males has a mean weight greater that 156.25 lb (so that the total weight exceeds the maximum capacity of 2500 lb).

Exercises



Exercises

Exercise 6: You need to obtain new desks for an incoming class of 25 kindergarten students who are all 5 years of age. An important characteristic of the desks is that the it must accommodate the sitting heights of those students. (The sitting height is the height of a seated student from the bottom of the feet to the top of the knee.) The table below lists the parameters for sitting heights of 5-year-old children

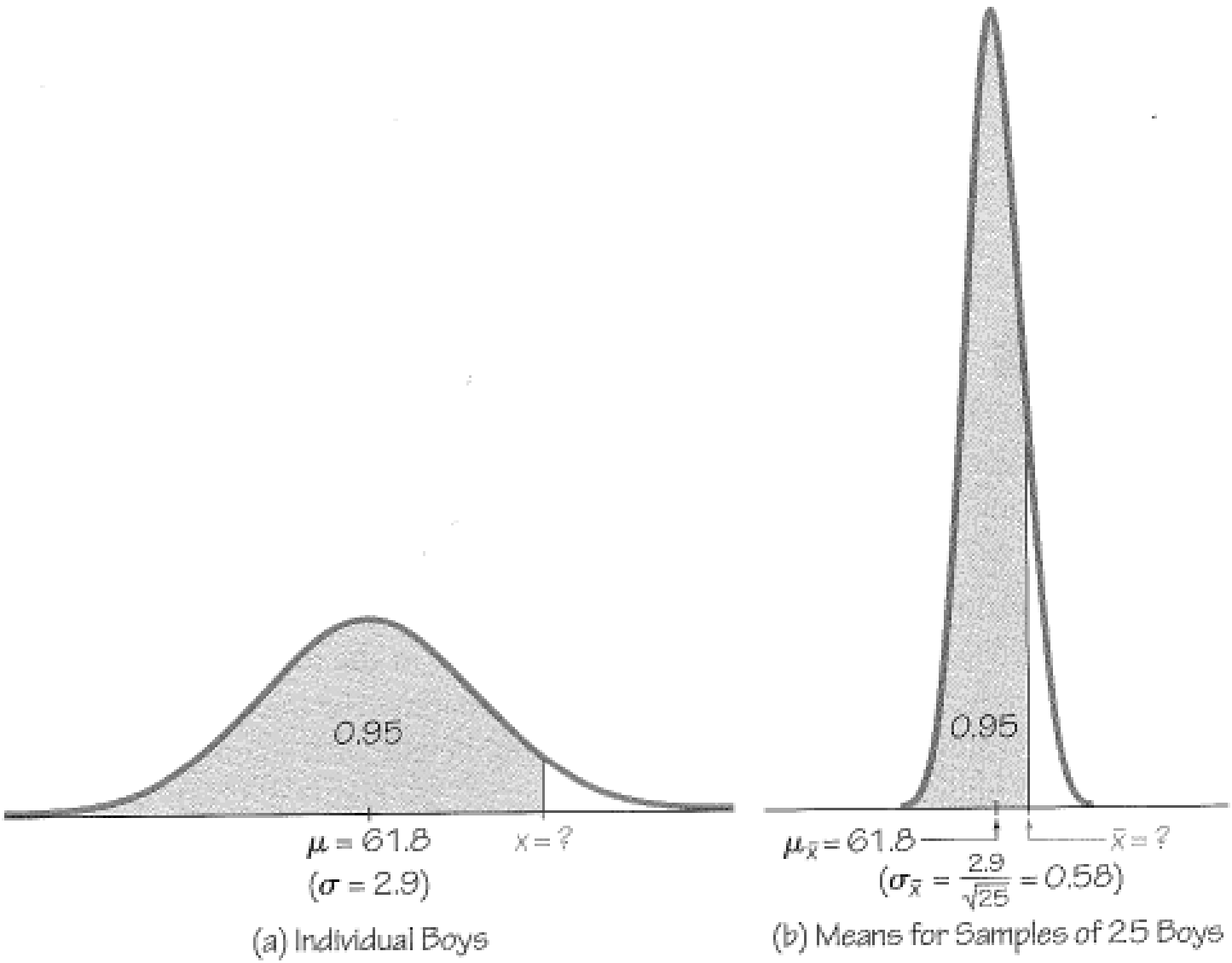
	Boys	Girls
μ	61.8 cm	61.2 cm
σ	2.9 cm	3.1 cm
Distribution	normal	normal

Exercises

Exercise 7: You need to obtain new desks for an incoming class of 25 kindergarten students who are all 5 years of age. An important characteristic of the desks is that the it must accommodate the sitting heights of those students. (The sitting height is the height of a seated student from the bottom of the feet to the top of the knee.) The table below lists the parameters for sitting heights of 5-year-old children

1. What sitting height will accommodate 95% of the boys?
2. What sitting height is greater than 95% of the means of sitting heights from random samples of 25 boys?
3. Based on the preceding results, what single value should be be the minimum sitting height accommodated by the desks? Why are the sitting heights of girls not included in the calculations?

Exercises


0.95
$\mu = 61.8$
$(\sigma = 2.9)$
$x = ?$
(a) Individual Boys
0.95
$\mu_{\bar{x}} = 61.8$
$(\sigma_{\bar{x}} = \frac{2.9}{\sqrt{25}} = 0.58)$
$\bar{x} = ?$
(b) Means for Samples of 25 Boys

Exercise 8: Cans of regular Coke are labeled to indicate that they contain 12 oz. The corresponding sample statistics are $n = 36$ and $\bar{x} = 12.19$ oz. Assuming that the Coke cans are filled so that $\mu = 12$ oz (as labeled) and the population standard deviation is $\sigma = 0.11$ oz (based on the sample results), find the probability that a sample of 36 cans will have a mean of 12.19 oz or greater. Do these results suggest that the Coke cans are filled with an amount greater than 12.00 oz?

Exercises

Exercise 9: Women have head circumferences that are normally distributed with a mean of 22.65in and a standard deviation of 0.80in.

1. If the hats by Leko company produce women's hats so that they fit head circumferences between 21.00in and 25.00in, what percentage of women can fit into these hats?
2. If the company wants to produce hats to fit all women except for those with the smallest 2.5% and the largest 2.5% head circumferences, what head circumferences should be accommodated?
3. If 64 women are randomly selected, what is the probability that their mean head circumference is between 22.00in and 23.00in? If this probability is high, does it suggest that an order for 64 hats will very likely fit each of 64 randomly selected women? Why or why not?

Exercises

Exercise 10: According to the web site www.torchmate.com, "manhole covers must be a minimum of 22in in diameter, but can be as much as 60in in diameter." Assume that a manhole is constructed to have a circular opening with a diameter of 22 in. Men have shoulder breadths that are normally distributed with a mean of 18.2in and a standard deviation of 1.0in (based on data from the National Health and Nutrition Examination Survey).

1. What percentage of men will fit into the manhole?
2. Assume that the Connecticut Light and Power company employs 36 men who work in manholes. If 36 men are randomly selected, what is the probability that their mean shoulder breadth is less than 18.5in? Does this result suggest that money can be saved by making smaller manholes with a diameter of 18.5in? Why or why not?

Exercises

Exercise 11: Passengers died when a water taxi sank in Baltimore's Inner Harbor. Men are typically heavier than women and children, so when loading a water taxi, assume a worst-case scenario in which all passengers are men. Assume that weights of men are normally distributed with a mean of 182.9lb and a standard deviation of 40.8lb. The water taxi that sank had a stated capacity of 25 passengers, and the boat was rated for a load limit of 3500 lb.

Exercises

1. Given that the water taxi that sank was rated for a load limit of 3500 lb, what is the mean weight of the passengers if the boat is filled to the stated capacity of 25 passengers?
2. If the water taxi is filled with 25 randomly selected men, what is the probability that their mean weight exceeds the value?
3. After the water taxi sank, the weight assumptions were revised so that the new capacity became 20 passengers. If the water taxi is filled with 20 randomly selected men, what is the probability that their mean weight exceeds 175lb, which is the maximum mean weight that does not cause the total load to exceed 3500lb?
4. Is the new capacity of 20 passengers safe?

Exercises

Exercise 12: Loading M&M Packages M&M plain candies have a mean weight of 0.8565g and a standard deviation of 0.0518g. The M&M candies used in a data set came from a package containing 465 candies, and the package label stated that the net weight is 396.9g. (If every package has 465 candies, the mean weight of the candies must exceed $396.9/465 = 0.8535g$ for the net contents to weigh at least 396.9g.)

1. If 1 M&M plain candy is randomly selected, find the probability that it weighs more than 0.8535g.
2. If 465 M&M plain candies are randomly selected, find the probability that their mean weight is at least 0.8535g.
3. Given these results, does it seem that the Mars Company is providing M&M consumers with the amount claimed on the label?

Exercises

Exercise 13: A ski gondola in Vail, Colorado, carries skiers to the top of a mountain. It bears a plaque stating that the maximum capacity is 12 people or 2004lb. That capacity will be exceeded if 12 people have weights with a mean greater than $2004/12 = 167$lb. Because men tend to weigh more than women, a worst-case scenario involves 12 passengers who are all men. Assume that weights of men are normally distributed with a mean of 182.9 lb and a standard deviation of 40.8lb.

1. Find the probability that if an individual man is randomly selected, his weight will be greater than 167lb.
2. Find the probability that twelve randomly selected men will have a mean weight that is greater than 167lb (so that their total weight is greater than the gondola maximum capacity of 2004lb).
3. Does the gondola appear to have the correct weight limit? Why or why not?

Instructor Note

Introduce students to Kolmogorov-Smirnov test. See `http://www.real-statistics.com/tests-normality-and-symmetry` and wikipedia. In R Statistics, try `ks.test((rnorm(55,124,4.3)-124)/4.3,pnorm)`.

Assessing Normality

IQ is not a number stenciled onto a person's forehead. Psychologists develop tests that are supposed to measure intelligence (and whether there is anything to be measured is a matter of controversy). One necessary feature of a useful IQ test is that its results are normally distributed with a mean of 100 and a standard deviation of 16 (or 15). You are the manager of a company that creates IQ tests. A psychologist introduces a test to you which, applied to a random sample of appropriate test subjects, has the following results (next slide). The mean is 101.005, the standard deviation is 16.20518. Later in the course we will learn that these statistics are consistent with the requirement that the test has a mean of 100 and a standard deviation of 16. The question remains: is the data normally distributed?

Assessing Normality

104	127	91	91	87	113	112	100	97	87
78	99	87	88	103	95	102	104	114	87
63	93	106	110	82	97	102	100	107	91
105	109	103	115	95	95	128	92	120	107
121	81	101	102	105	119	82	105	73	115
116	95	85	119	113	108	160	128	101	69
87	104	78	85	93	95	128	85	83	82
124	67	106	126	106	103	105	98	76	128
104	122	105	90	110	86	82	87	100	108
83	118	102	109	80	78	112	89	113	92
107	79	111	111	102	84	101	82	93	87
108	113	131	108	87	89	83	92	117	95
85	104	114	113	78	120	102	114	74	97
103	88	90	124	92	120	81	91	139	115
142	99	119	87	109	73	94	95	91	101
101	122	89	107	118	108	97	109	123	125
107	89	117	105	122	92	91	44	106	74
133	125	95	111	128	74	97	112	79	107
127	104	98	109	99	101	104	121	99	119
94	119	84	87	94	92	93	86	104	113

Assessing Normality

To assess normality informally, we follow a three-step procedure. Once we have learned hypothesis testing, we will be able to assess normality more formally.

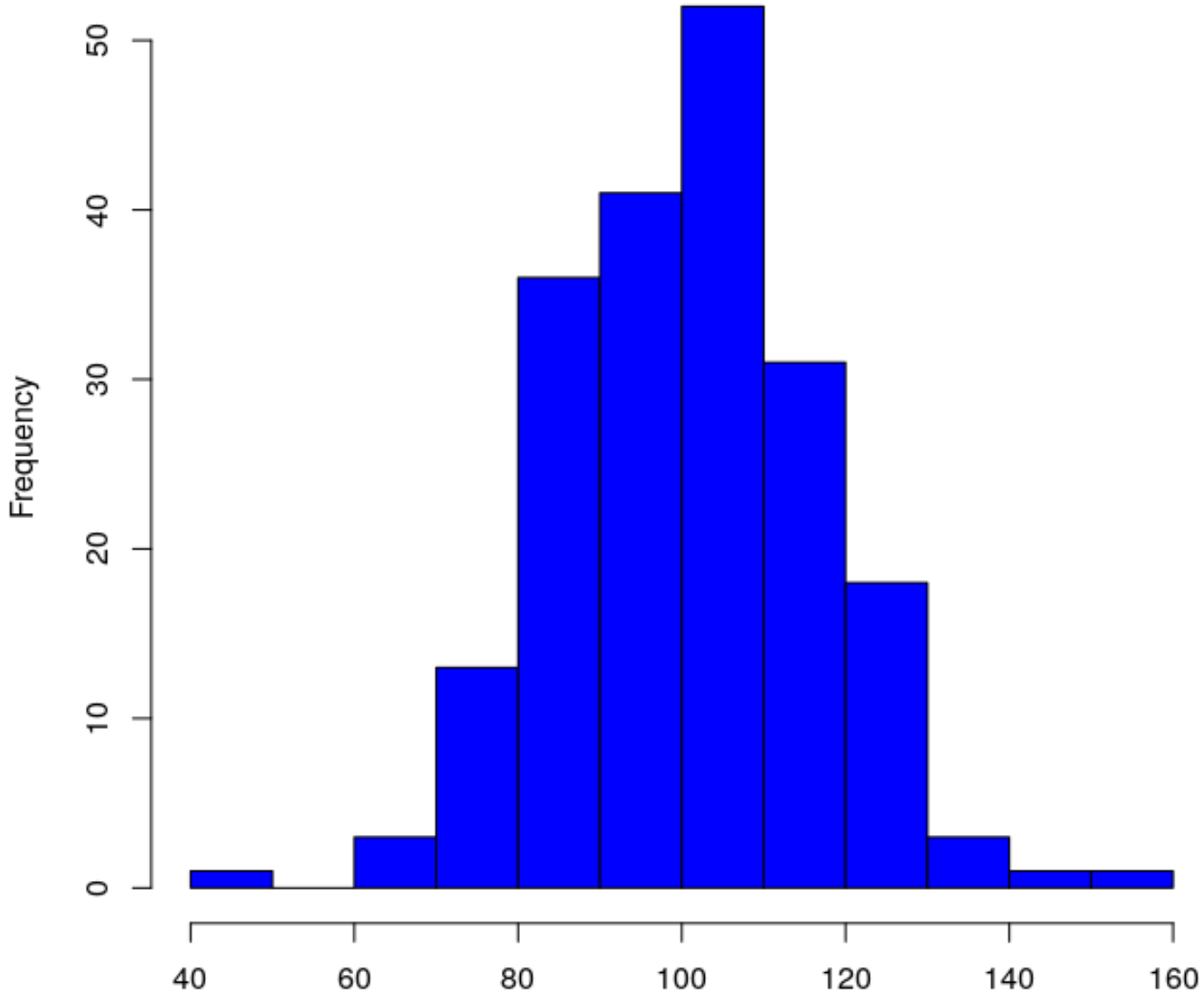
Histogram Construct a histogram. If the histogram departs dramatically from a bell shape, conclude that the data does not have a normal distribution.

Outliers Use a boxplot to identify outliers. If there is more than one outlier present, conclude that the data does not have a normal distribution.

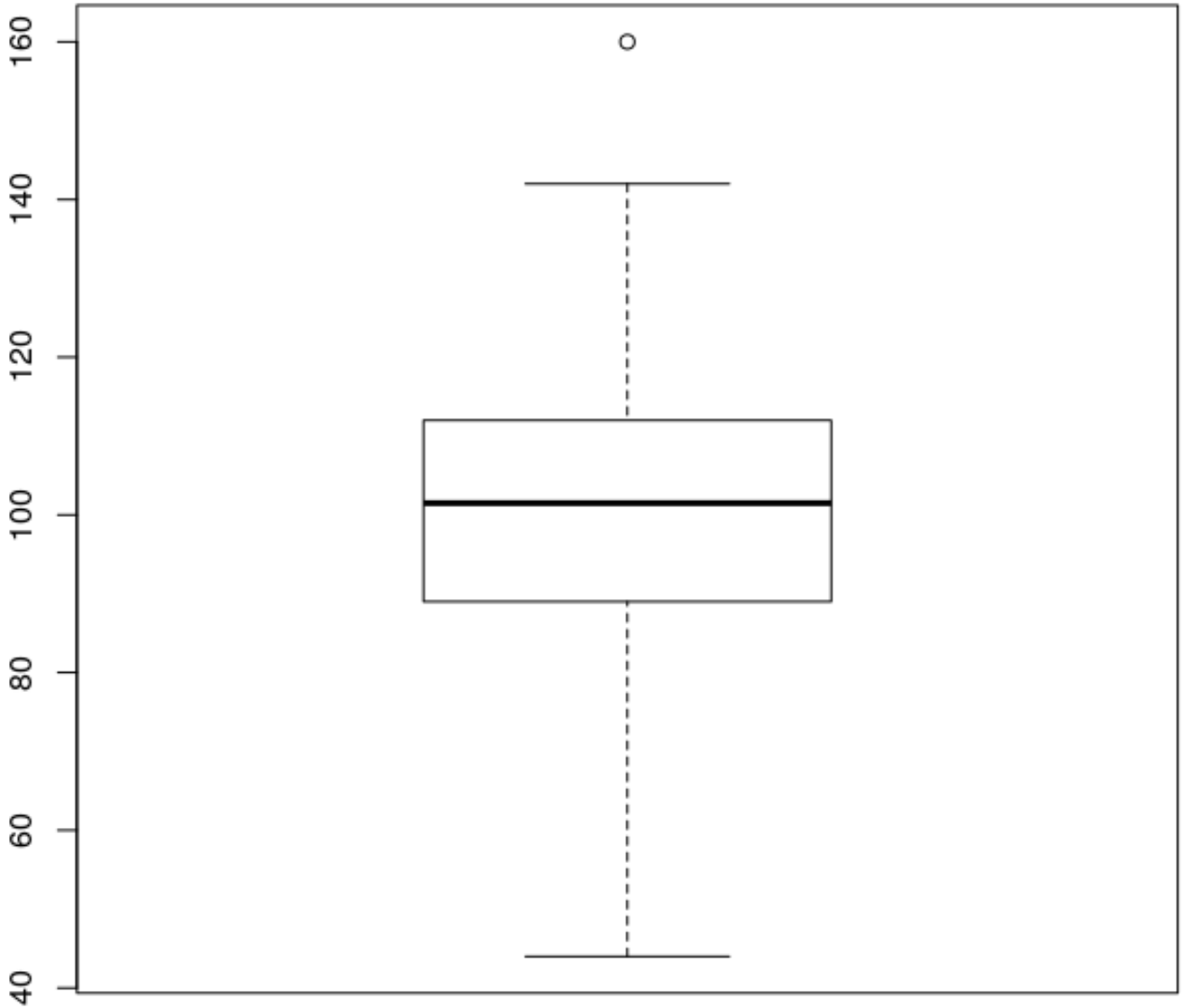
NQP If the data passes the first two tests, use technology to generate a normal quantile plot. The population is probably not normally distributed if either one of the following two conditions apply:

1. The points do not lie reasonably close to a straight line.
2. The points show some systematic pattern that is not a straight-line pattern.

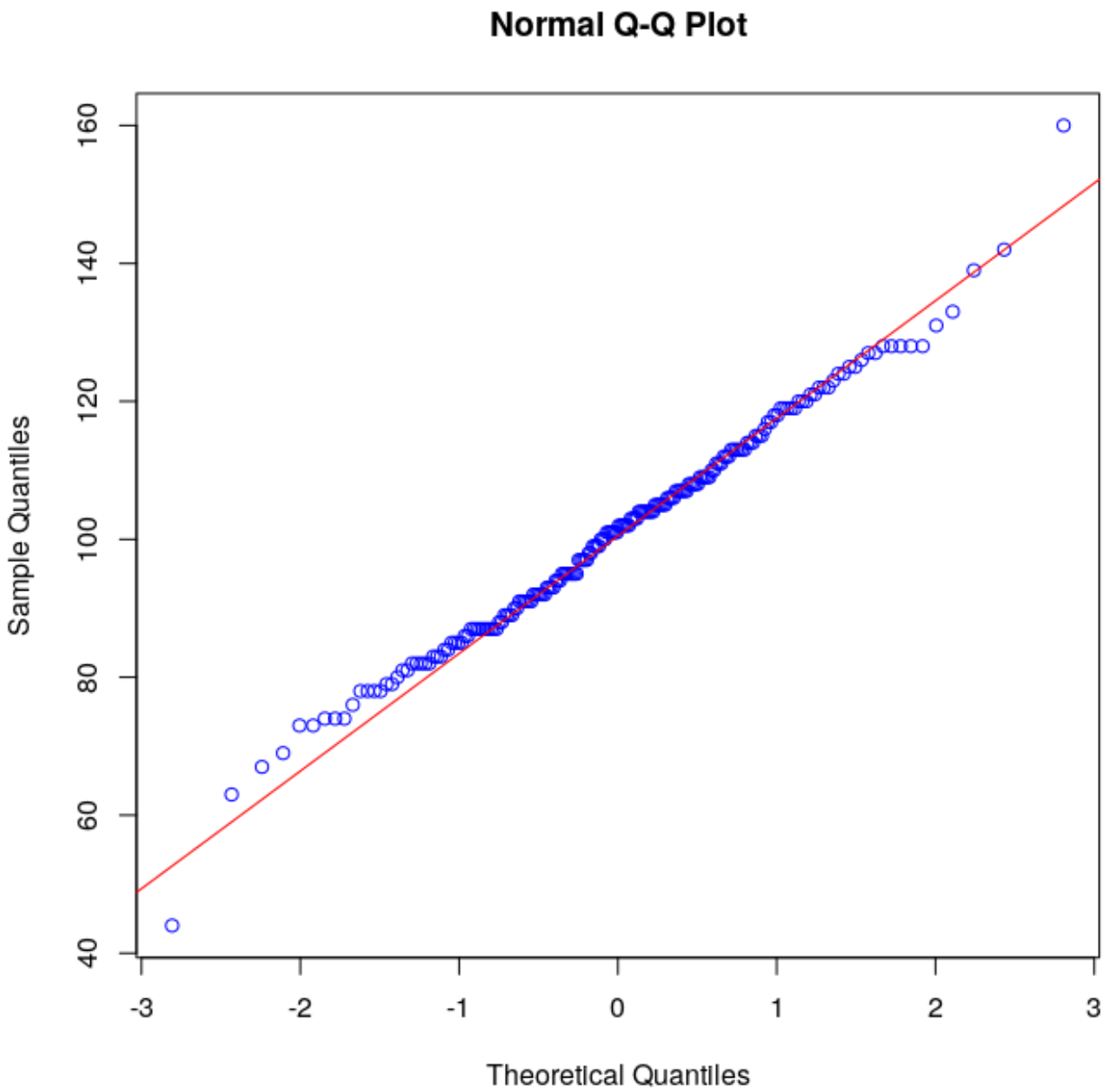
Histogram



Boxplot



Normal Quantile Plot



R Code

The following R Statistics code will generate normal data and the graphs associated with each of the three steps.

```
x<-round(rnorm(200,100,16),0)
hist(x,col="blue")
boxplot(x,range=2)
qqnorm(x,col="blue")
qqline(x,col="red")
```

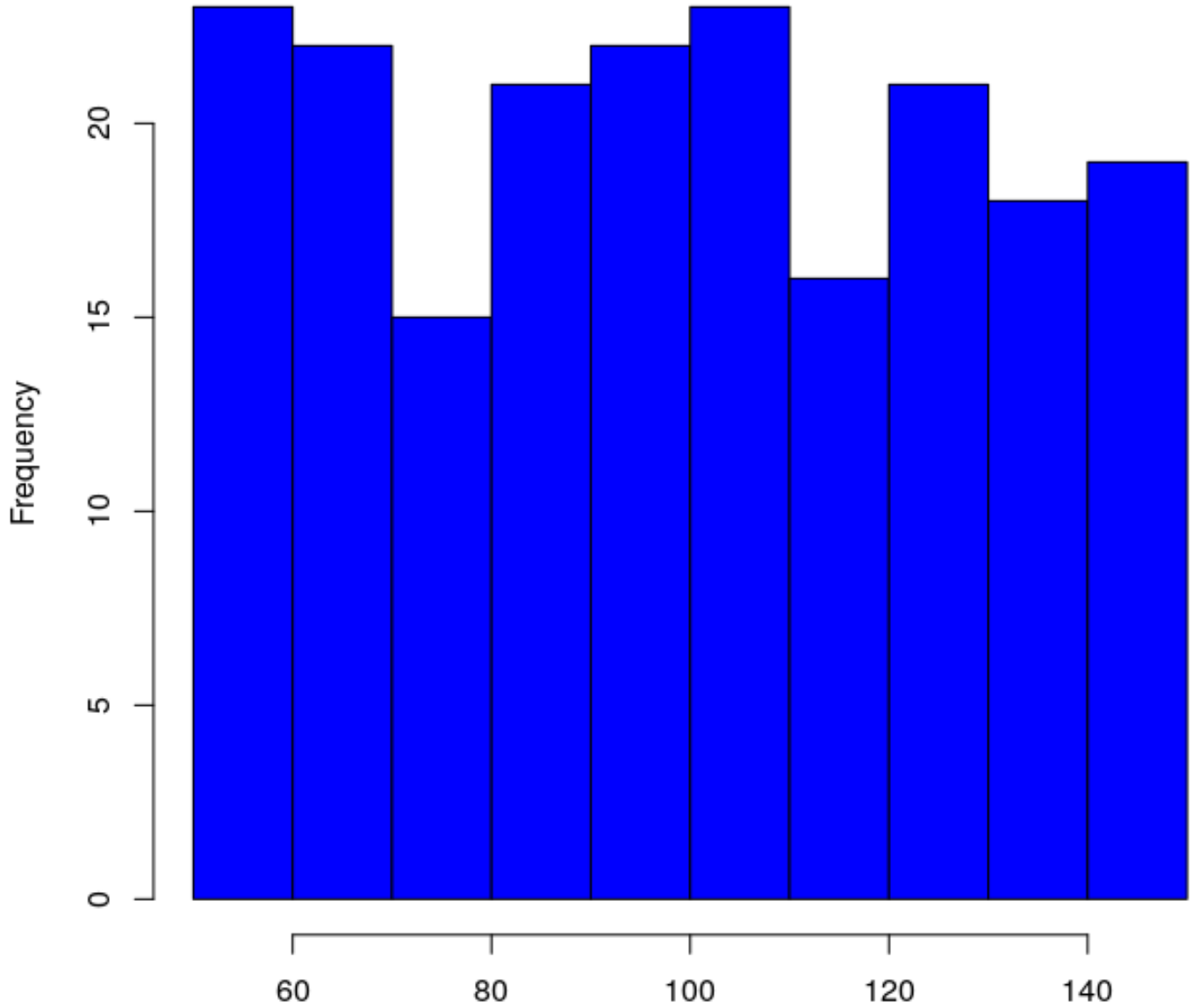
Non-Normal Distributions

Try the following R Statistics code for some non-normal distributions.

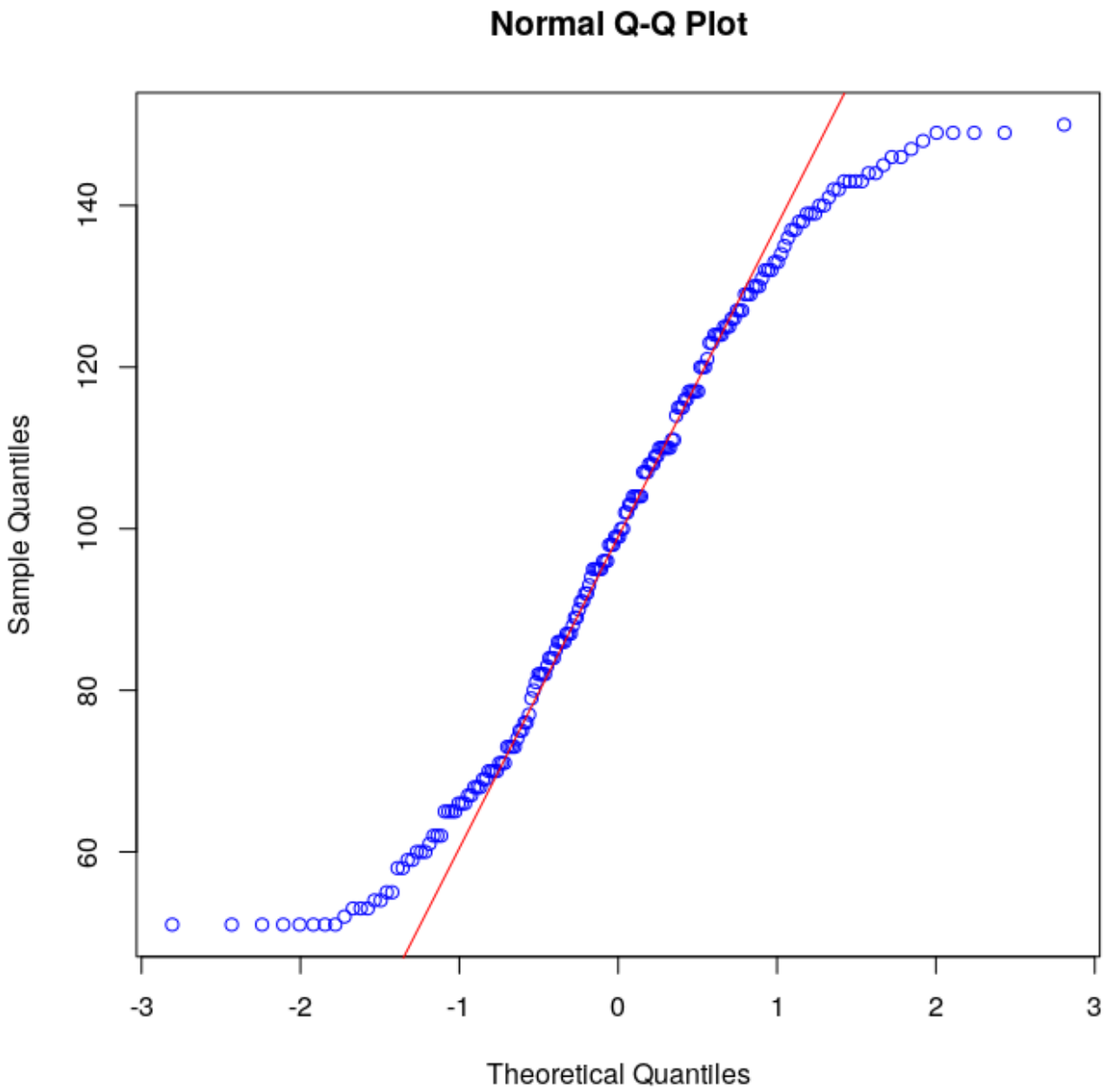
```
y<-round(runif(200,50,150),0)
w<-rnorm(200,0,20)
z<-100-(((abs(w)/w)*50)-w)
v<-rnorm(200,500,40)
v<-append(v,rnorm(200,300,40))
```

y is a uniform distribution, z is a U-shaped distribution (w helps to generate z), and v is a two-peaked distribution.

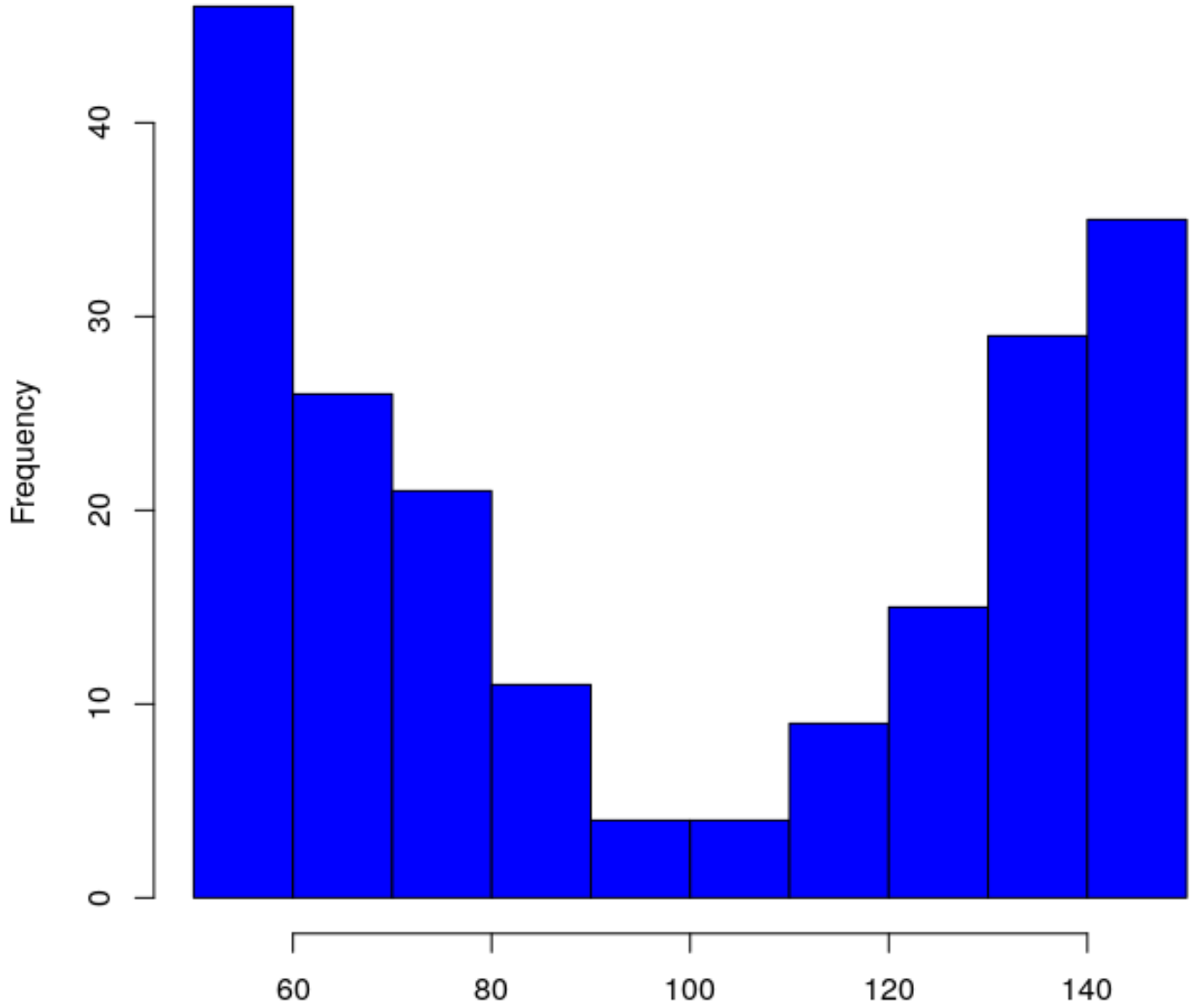
Histogram of Uniform Distribution



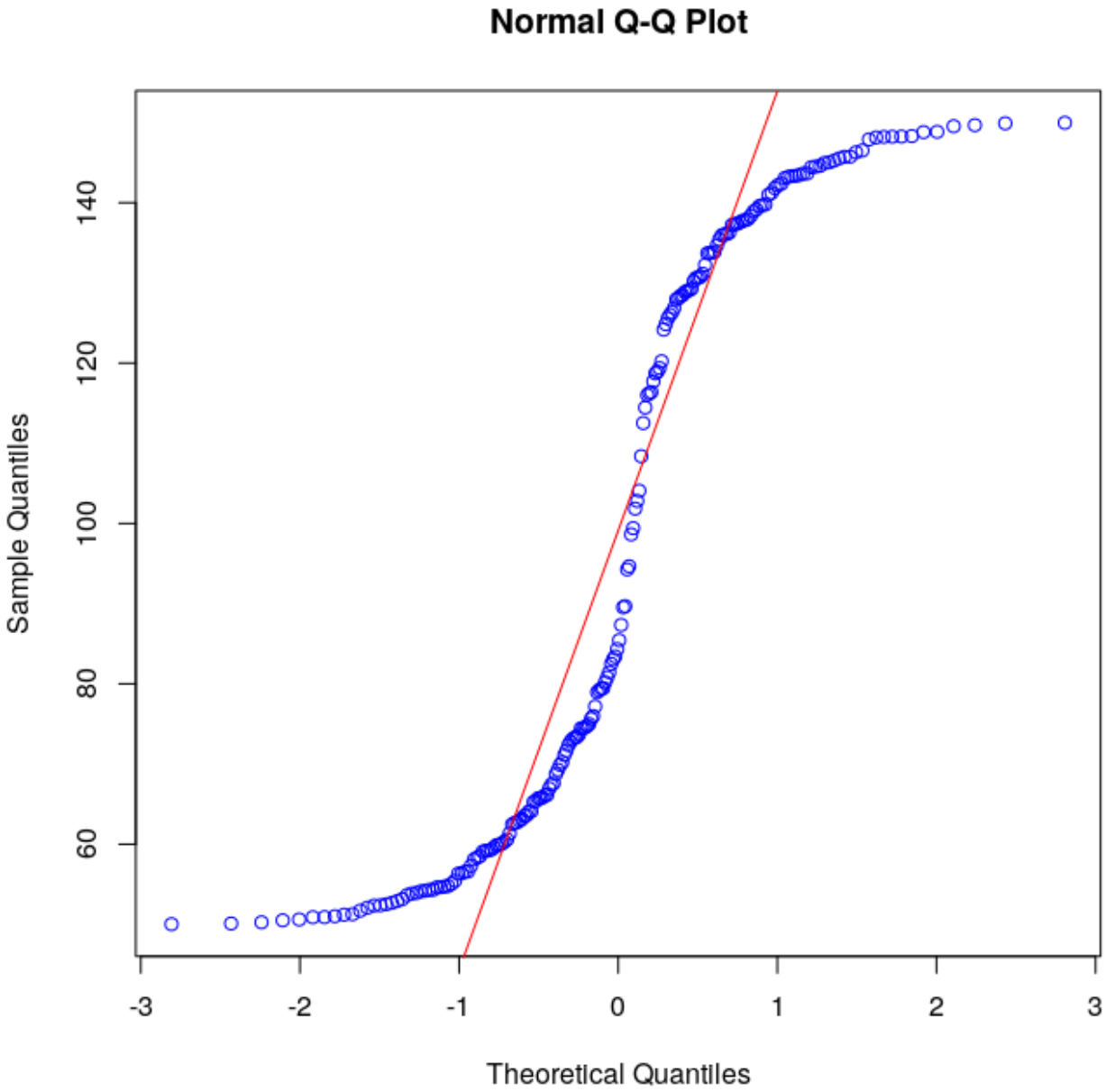
Normal Quantile Plot of Uniform Distribution



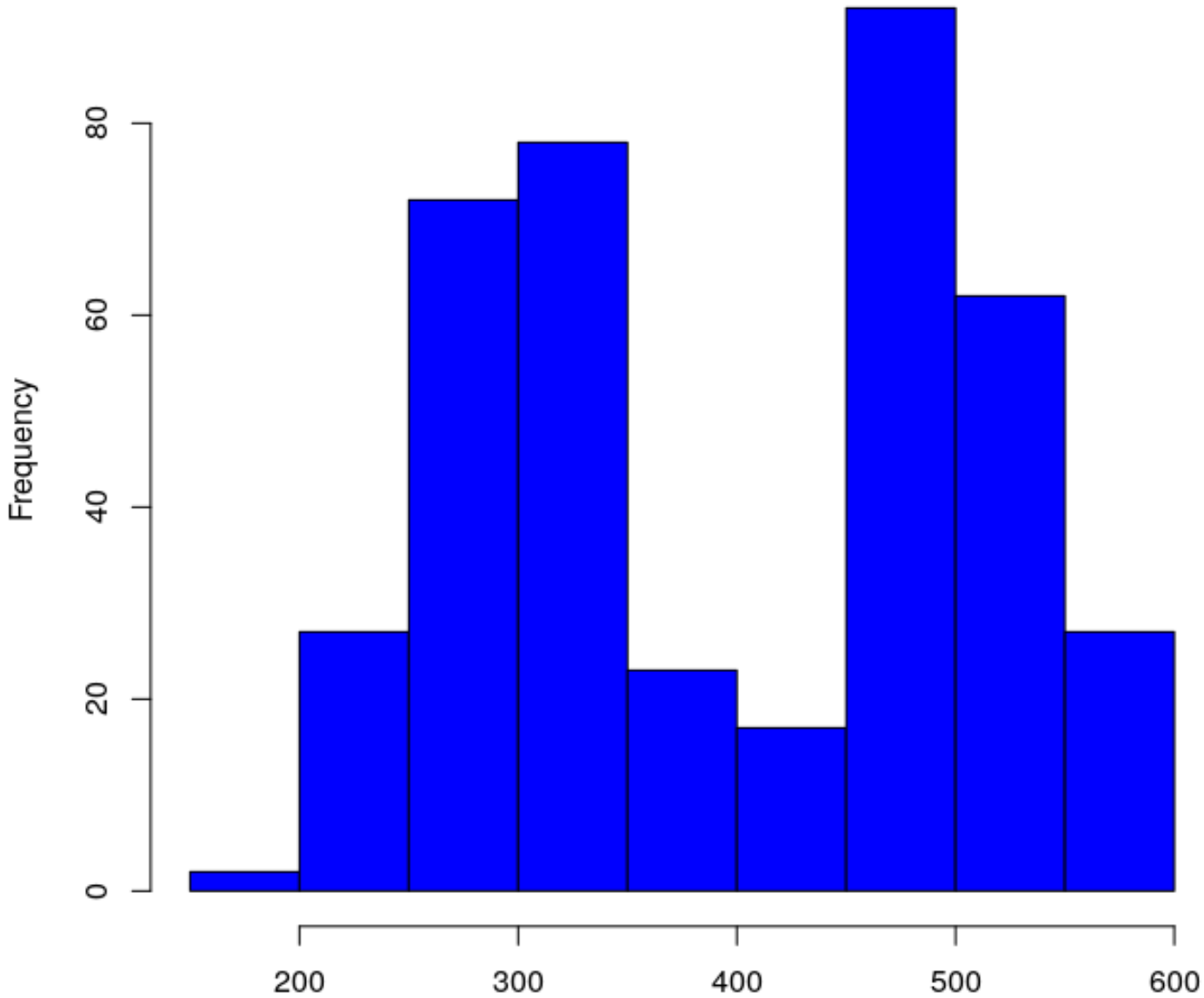
Histogram of U-Shaped Distribution



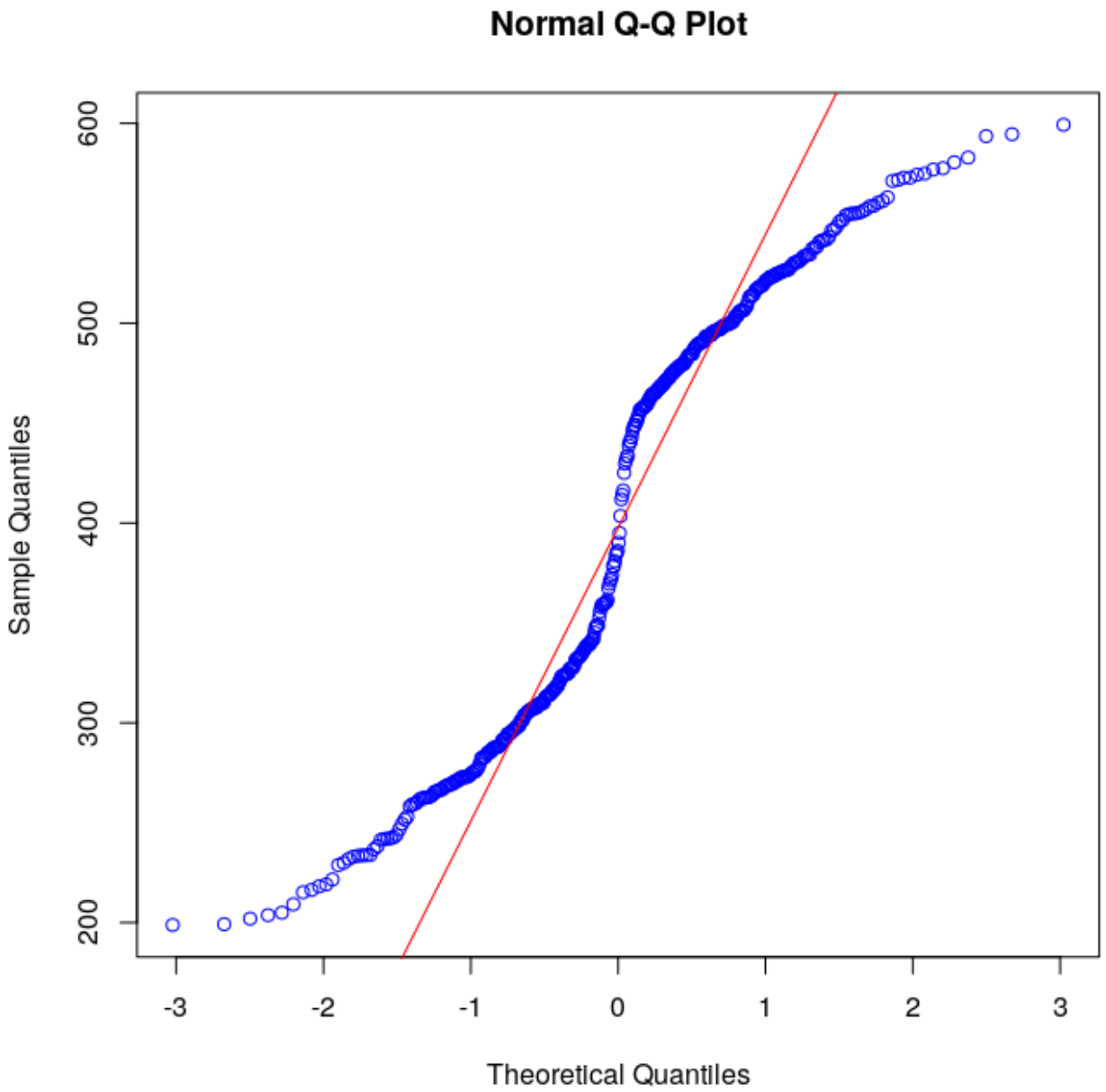
Normal Quantile Plot of U-Shaped Distribution



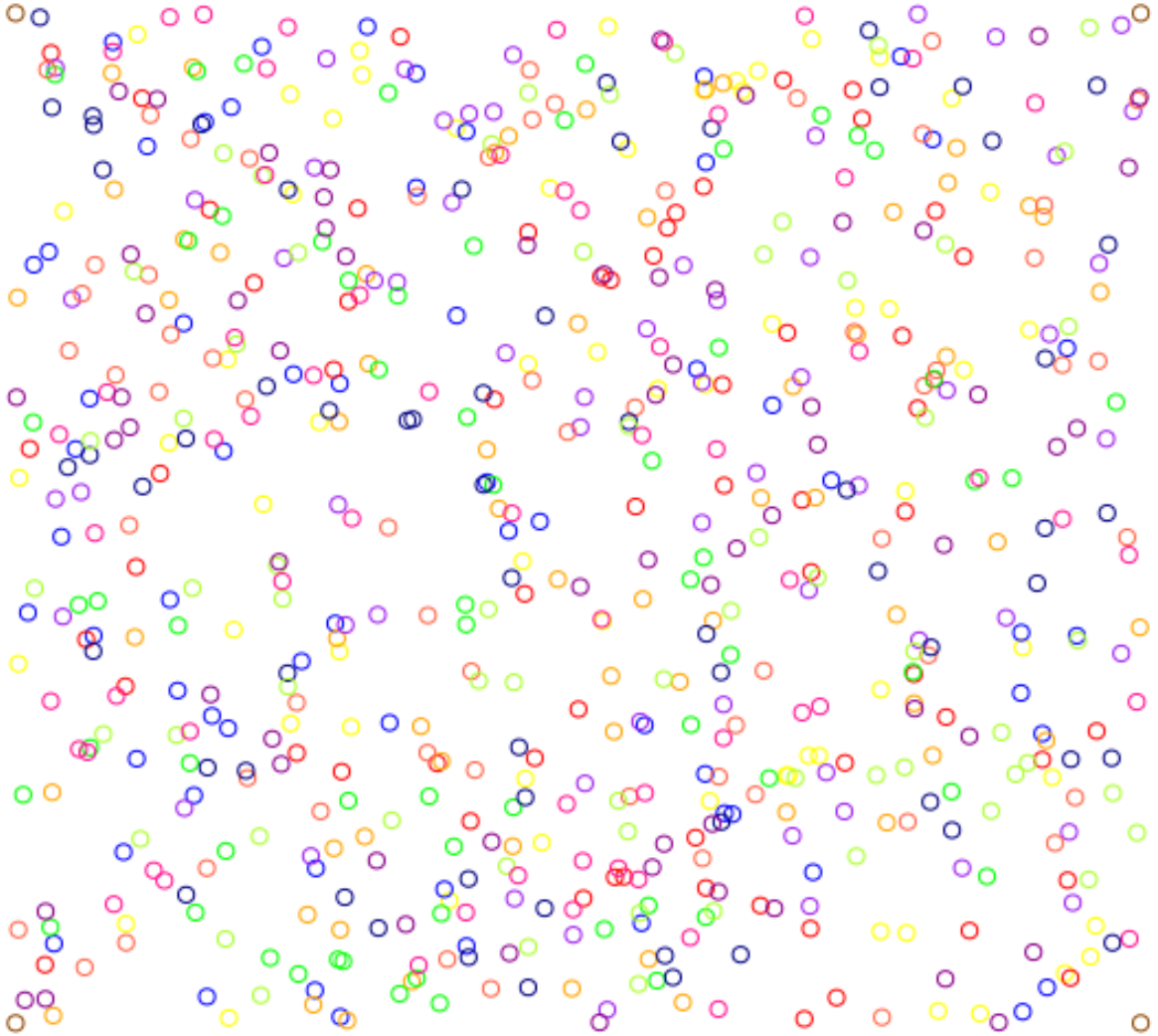
Histogram of Two-Peaked Distribution



Normal Quantile Plot of Two-Peaked Distribution



Guess the Number of Jelly Beans



Exercises

Exercise 14: Follow the three-step procedure to find out if the following data sets are normally distributed. You can find the data on D2L.

1. Ages of Oscar-Winning Actresses `OSCRF.TXT`
2. Body Temperatures `BTEMP.TXT`
3. White Blood Cell Counts for Males `MWHT.TXT`
4. Flight Departure Delays `DPDLY.TXT`

End of Lesson

Next Lesson: Confidence Intervals