

# Chapter 2: Descriptive Statistics and Displays

DSCC 462  
Computational Introduction to Statistics

Anson Kahng  
Fall 2022

# Plan for Today

# Plan for Today

- Visualize datasets using graphs

# Plan for Today

- Visualize datasets using graphs
- Describe important aspects of data using measures of **center** and **spread**

# Plan for Today

- Visualize datasets using graphs —
- Describe important aspects of data using measures of **center** and **spread**
- Transform data to allow for statistical tests

# Describing Data

# Describing Data

- Once we have data collected, we need to distill it down into summaries that provide essential information

# Describing Data

- Once we have data collected, we need to distill it down into summaries that provide essential information
- **Goal:** Describe important attributes of a dataset in a concise manner

# Describing Data

- Once we have data collected, we need to distill it down into summaries that provide essential information
- **Goal:** Describe important attributes of a dataset in a concise manner
- When we summarize, some information is lost, but a great deal of communicative power can be gained

# Describing Data

- Once we have data collected, we need to distill it down into summaries that provide essential information
- **Goal:** Describe important attributes of a dataset in a concise manner
- When we summarize, some information is lost, but a great deal of communicative power can be gained
- Often, we want to use both numerical and graphical summaries

# Describing Data

- Once we have data collected, we need to distill it down into summaries that provide essential information
- **Goal:** Describe important attributes of a dataset in a concise manner
- When we summarize, some information is lost, but a great deal of communicative power can be gained
- Often, we want to use both numerical and graphical summaries
- The type of summary statistics you use will depend on the type of data

# Describing Data

# Describing Data

- Generally, we want to know what the distribution of our variables is

# Describing Data

- Generally, we want to know what the distribution of our variables is
- **Distribution:** the values a variable can take on and how often it takes each value

# Describing Data

- Generally, we want to know what the distribution of our variables is
- **Distribution:** the values a variable can take on and how often it takes each value
- Make plots and construct tables to see what the distribution looks like

# Describing Categorical Distributions

# Describing Categorical Distributions

- **Absolute frequency table:** A table whose values correspond to how often we observe each of the categories of a variable

# Describing Categorical Distributions

- **Absolute frequency table:** A table whose values correspond to how often we observe each of the categories of a variable
- Categories and numeric count in each category

# Describing Categorical Distributions

- **Absolute frequency table:** A table whose values correspond to how often we observe each of the categories of a variable
- Categories and numeric count in each category

Cause of Death	Number of Deaths
Cancer	12
Heart Attack	30
Stroke	10
Car Accident	53
Other	37

# Describing Categorical Distributions

# Describing Categorical Distributions

- **Relative frequency table:** A table whose values are the percentage of the total number of observations that appear in each category

# Describing Categorical Distributions

- **Relative frequency table:** A table whose values are the percentage of the total number of observations that appear in each category
- Categories and proportion (relative frequency) in each category

# Describing Categorical Distributions

- **Relative frequency table:** A table whose values are the percentage of the total number of observations that appear in each category
- Categories and proportion (relative frequency) in each category

Cause of Death	Number of Deaths	Relative Frequency (%)
Cancer	12	8.45
Heart Attack	30	21.13
Stroke	10	7.04
Car Accident	53	37.32
Other	37	26.06

# Describing Categorical Distributions: Graphically

# Describing Categorical Distributions: Graphically

- Graphically: We typically use a bar chart to illustrate the information in a frequency (or relative frequency) table

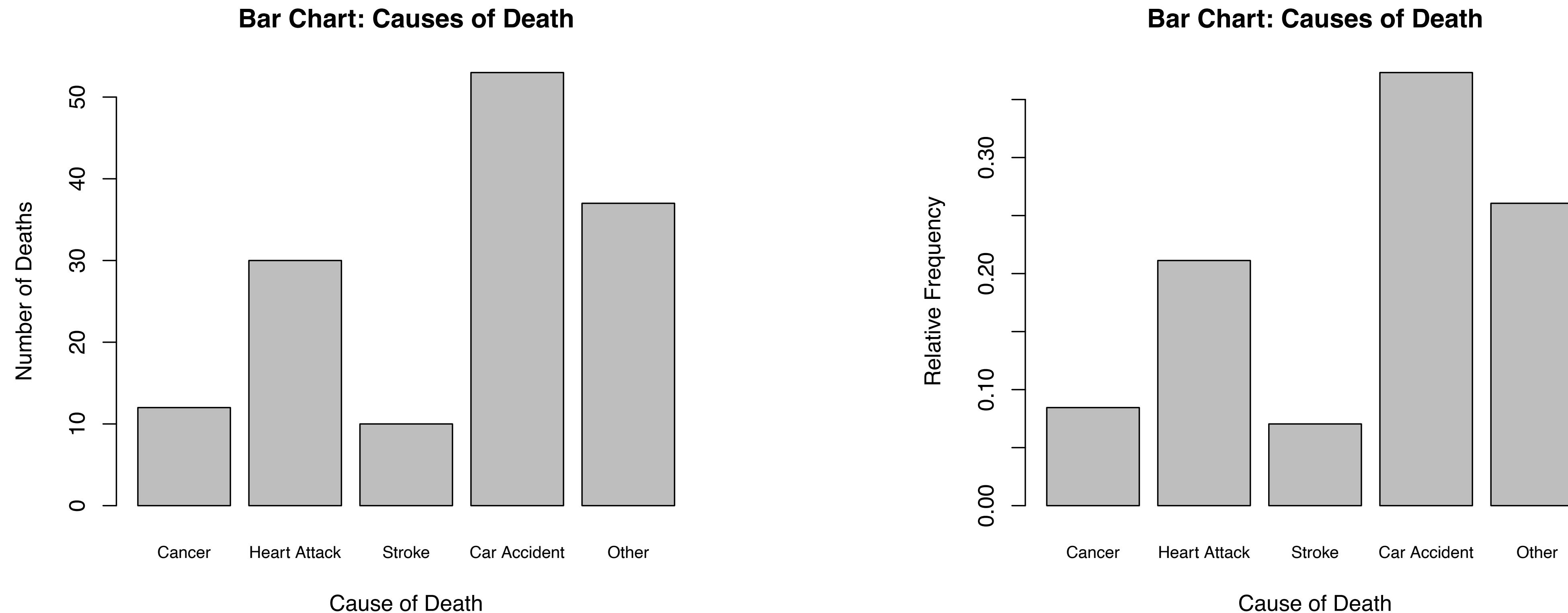
# Describing Categorical Distributions: Graphically

- Graphically: We typically use a bar chart to illustrate the information in a frequency (or relative frequency) table
- **Bar chart:** a graphical display with categories listed on the horizontal axis and vertical bars drawn for each category, where the height of the bar represents the frequency (or relative frequency) of observations within that category

# Describing Categorical Distributions: Graphically

- Graphically: We typically use a bar chart to illustrate the information in a frequency (or relative frequency) table
- **Bar chart:** a graphical display with categories listed on the horizontal axis and vertical bars drawn for each category, where the height of the bar represents the frequency (or relative frequency) of observations within that category
- Bars do not touch and are all of the same width

# Bar Charts



# Describing Continuous Distributions

# Describing Continuous Distributions

- When data are continuous, we will want to describe distributions in a different way

# Describing Continuous Distributions

- When data are continuous, we will want to describe distributions in a different way
- Typically, we are interested in how much of the data falls within a given range

# Describing Continuous Distributions

- When data are continuous, we will want to describe distributions in a different way
- Typically, we are interested in how much of the data falls within a given range
  - Generally, ranges should be of equal length

# Histogram

# Histogram

- **Histogram**: a graphical display of a frequency distribution for discrete or continuous data where the horizontal axis displays intervals for which the data is binned over, and the vertical axis represents the frequency (or relative frequency)

# Histogram

- **Histogram**: a graphical display of a frequency distribution for discrete or continuous data where the horizontal axis displays intervals for which the data is binned over, and the vertical axis represents the frequency (or relative frequency)
- Must determine *binwidth*, or the length of the intervals used for the horizontal axis

# Histogram

- **Histogram**: a graphical display of a frequency distribution for discrete or continuous data where the horizontal axis displays intervals for which the data is binned over, and the vertical axis represents the frequency (or relative frequency)
- Must determine *binwidth*, or the length of the intervals used for the horizontal axis
  - No best way for determining binwidth, though many possibilities exist (often, we first determine how many bins we want, and then evenly divide the range into that many bins)

# Histogram

- **Histogram**: a graphical display of a frequency distribution for discrete or continuous data where the horizontal axis displays intervals for which the data is binned over, and the vertical axis represents the frequency (or relative frequency)
- Must determine *binwidth*, or the length of the intervals used for the horizontal axis
  - No best way for determining binwidth, though many possibilities exist (often, we first determine how many bins we want, and then evenly divide the range into that many bins)
- By default, R uses Sturges' formula: number of bins  $k = \lceil \log_2(n) \rceil + 1$

# Histogram

# Histogram

- Since the horizontal axis represents a continuous number line, there should *not* be breaks between bars of the histogram

# Histogram

- Since the horizontal axis represents a continuous number line, there should *not* be breaks between bars of the histogram
- The only time that there should be space between histogram bars is when 0 entries are observed in a given interval

# Histogram Workflow

# Histogram Workflow

1. Determine the minimum and maximum of the dataset

# Histogram Workflow

1. Determine the minimum and maximum of the dataset
2. Calculate binwidth

# Histogram Workflow

1. Determine the minimum and maximum of the dataset
2. Calculate binwidth
3. Define a series of class intervals that are equal in size and adjacent to each other

# Histogram Workflow

1. Determine the minimum and maximum of the dataset
2. Calculate binwidth
3. Define a series of class intervals that are equal in size and adjacent to each other
  - The first class should contain the minimum; the last class should contain the maximum

# Histogram Workflow

1. Determine the minimum and maximum of the dataset
2. Calculate binwidth
3. Define a series of class intervals that are equal in size and adjacent to each other
  - The first class should contain the minimum; the last class should contain the maximum
  - If an observation is exactly on the boundary, put it in the lower bin

# Histogram Workflow

1. Determine the minimum and maximum of the dataset
2. Calculate binwidth
3. Define a series of class intervals that are equal in size and adjacent to each other
  - The first class should contain the minimum; the last class should contain the maximum
  - If an observation is exactly on the boundary, put it in the lower bin
4. Determine how many observations fall within each bin

# Histogram Workflow

1. Determine the minimum and maximum of the dataset
2. Calculate binwidth
3. Define a series of class intervals that are equal in size and adjacent to each other
  - The first class should contain the minimum; the last class should contain the maximum
  - If an observation is exactly on the boundary, put it in the lower bin
4. Determine how many observations fall within each bin
5. Indicate position of the class interval on the horizontal axis

# Histogram Workflow

1. Determine the minimum and maximum of the dataset
2. Calculate binwidth
3. Define a series of class intervals that are equal in size and adjacent to each other
  - The first class should contain the minimum; the last class should contain the maximum
  - If an observation is exactly on the boundary, put it in the lower bin
4. Determine how many observations fall within each bin
5. Indicate position of the class interval on the horizontal axis
6. At each class interval, draw a vertical bar equal in height to the number of observations in that class

# Histogram Example

165.5	158.1	154.7	178.9	180.9	194.5	174.1	160.5
202.0	259.8	125.5	160.0	160.9	179.0	195.0	149.3
194.8	148.0	193.9	162.2	159.1	206.3	178.7	164.5
204.4	179.5	126.9	196.6	231.1	202.8	174.0	181.9
186.4	198.0	150.8	172.2	187.1	152.8	169.6	206.6
177.3	148.0	196.7	179.7	180.6	152.6	185.3	158.7
211.3	185.1	177.3	184.6	165.7	170.1	178.4	184.2

Table 1: Weights (lbs) of 56 patients

# Histogram Example

# Histogram Example

- Minimum: 125.5

# Histogram Example

- Minimum: 125.5
- Maximum: 259.8

# Histogram Example

- Minimum: 125.5
- Maximum: 259.8
- How many bins?  $\lceil \log_2 56 \rceil + 1 = 7$  bins

# Histogram Example

- Minimum: 125.5
- Maximum: 259.8
- How many bins?  $\lceil \log_2 56 \rceil + 1 = 7$  bins
- Range: 120 - 260, so each bin has length 20

# Histogram Example

- Minimum: 125.5
- Maximum: 259.8
- How many bins?  $\lceil \log_2 56 \rceil + 1 = 7$  bins
- Range: 120 - 260, so each bin has length 20
- Frequency table

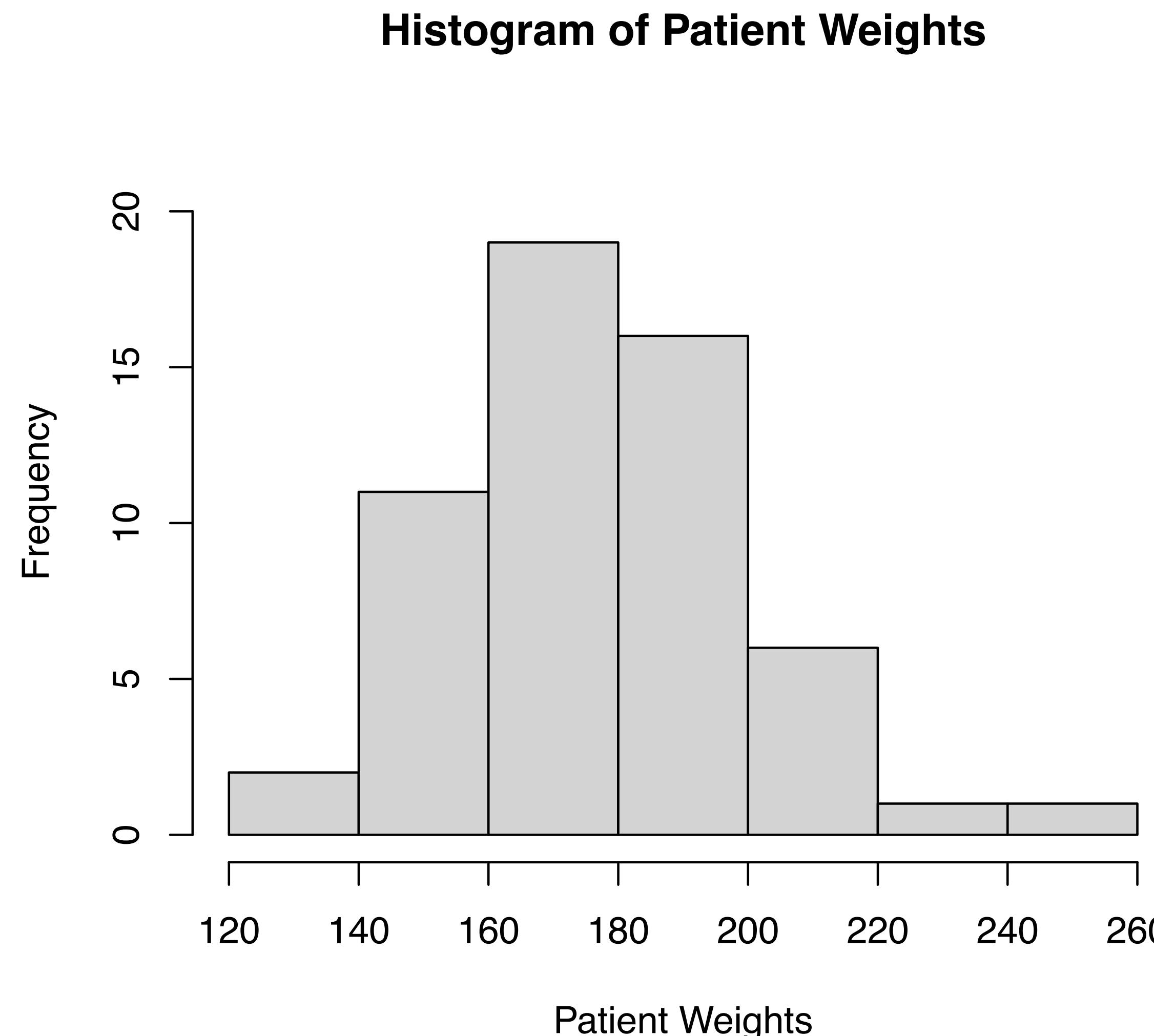
# Histogram Example

- Minimum: 125.5
- Maximum: 259.8
- How many bins?  $\lceil \log_2 56 \rceil + 1 = 7$  bins
- Range: 120 - 260, so each bin has length 20
- Frequency table

Weight	Frequency
(120, 140]	2
(140, 160]	10
(160, 180]	20
(180, 200]	16
(200, 220]	6
(220, 240]	1
(240, 260]	1

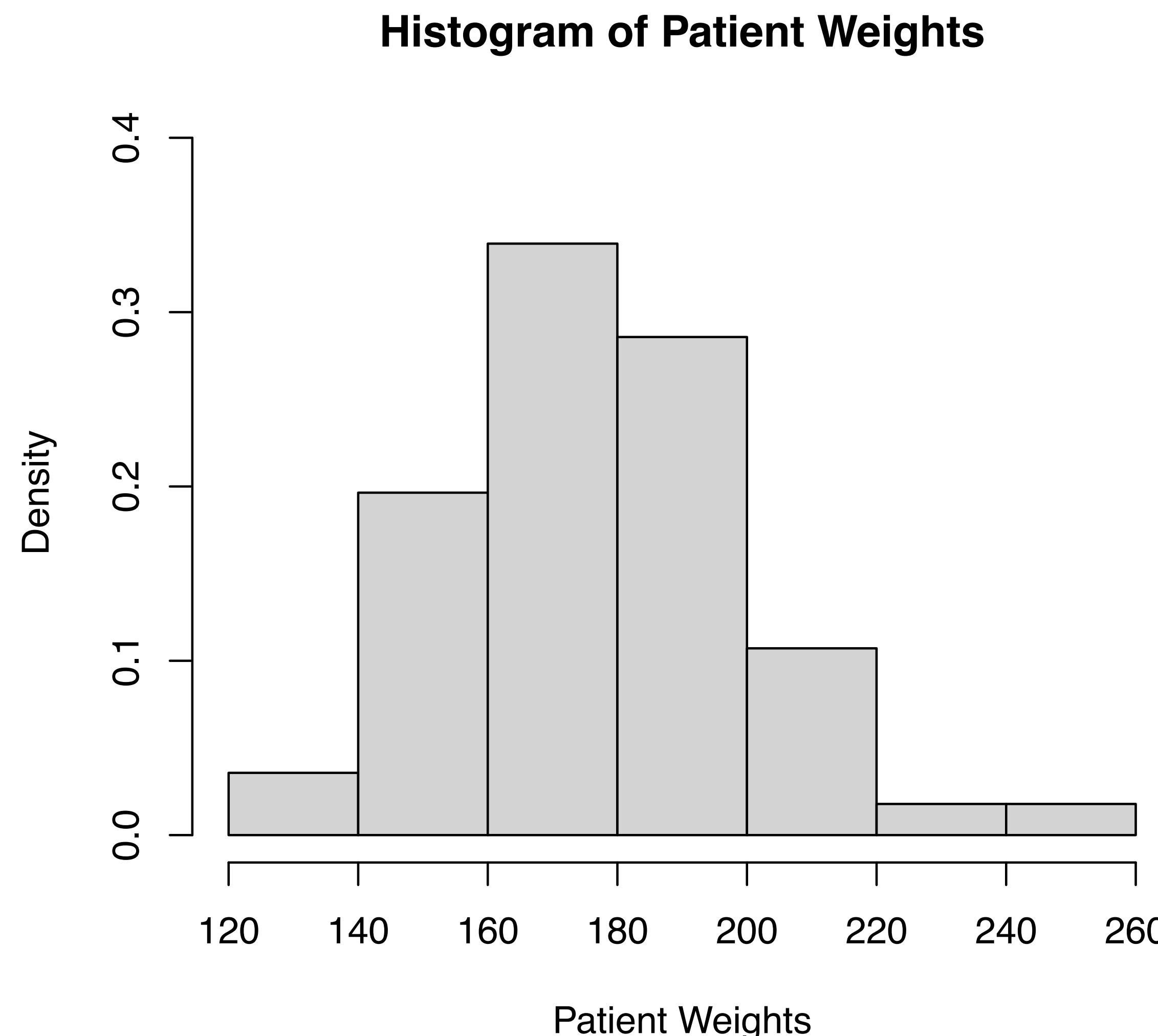
# Histogram Example: Count

Weight	Frequency
(120, 140]	2
(140, 160]	10
(160, 180]	20
(180, 200]	16
(200, 220]	6
(220, 240]	1
(240, 260]	1



# Histogram Example: Density

Weight	Rel. Frequency
(120, 140]	0.04
(140, 160]	0.20
(160, 180]	0.34
(180, 200]	0.29
(200, 220]	0.11
(220, 240]	0.02
(240, 260]	0.02



# Histogram Properties

# Histogram Properties

- Typically, we want to examine a histogram's:

# Histogram Properties

- Typically, we want to examine a histogram's:
  - Center

# Histogram Properties

- Typically, we want to examine a histogram's:
  - Center
  - Shape

# Histogram Properties

- Typically, we want to examine a histogram's:
  - Center
  - Shape
  - Spread

# Histogram Properties

- Typically, we want to examine a histogram's:
  - Center
  - Shape
  - Spread
- Be aware of deviations from the typical pattern / extreme points

# Histogram Properties

- Typically, we want to examine a histogram's:
  - Center
  - Shape
  - Spread
- Be aware of deviations from the typical pattern / extreme points
- **Outliers:** Data that are not typical of the rest of the values in the dataset

# Center, Shape, and Spread

# Center, Shape, and Spread

- Does the distribution appear to have one center, or are there multiple peaks?

# Center, Shape, and Spread

- Does the distribution appear to have one center, or are there multiple peaks?
- Is the distribution symmetric?

# Center, Shape, and Spread

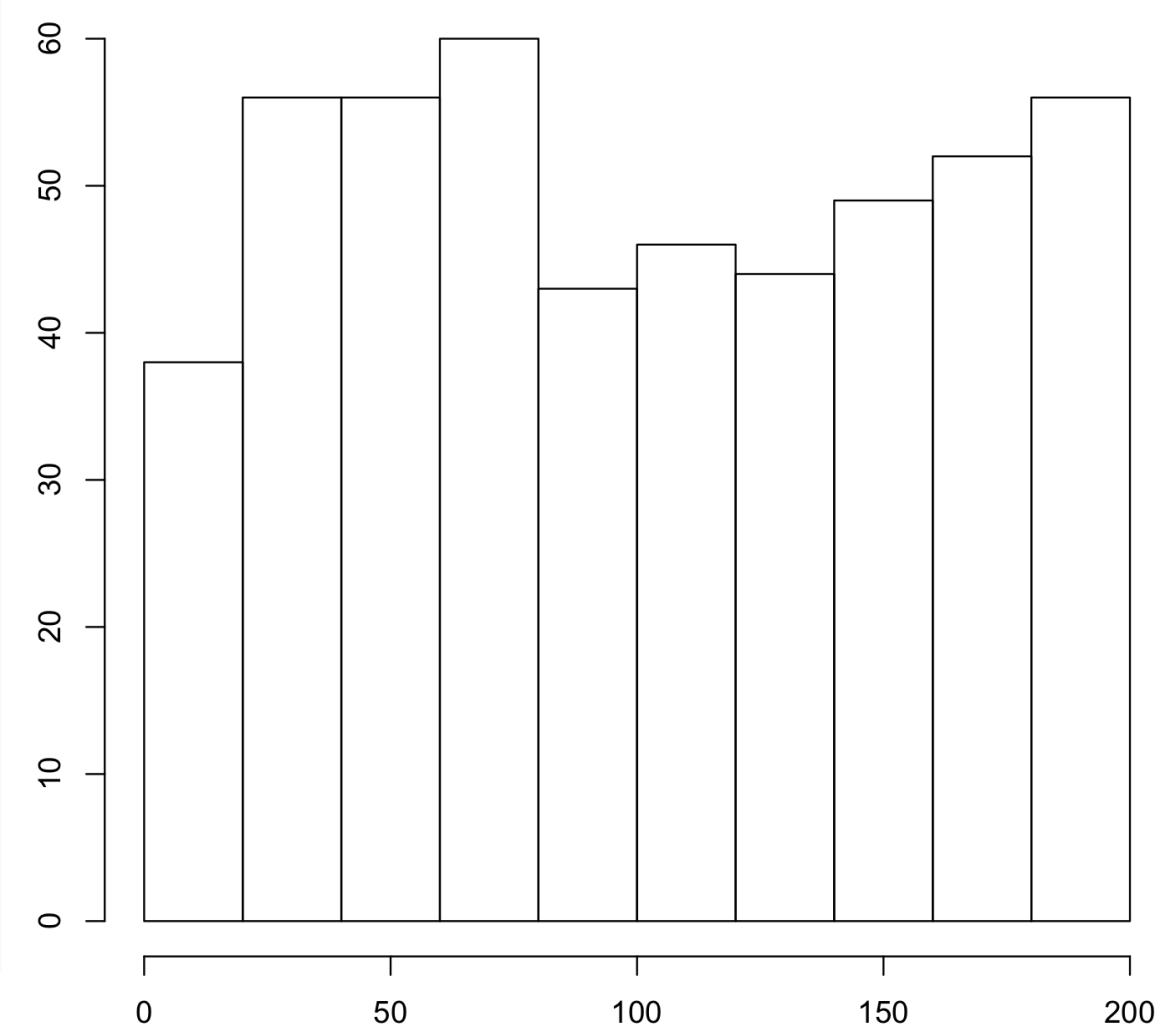
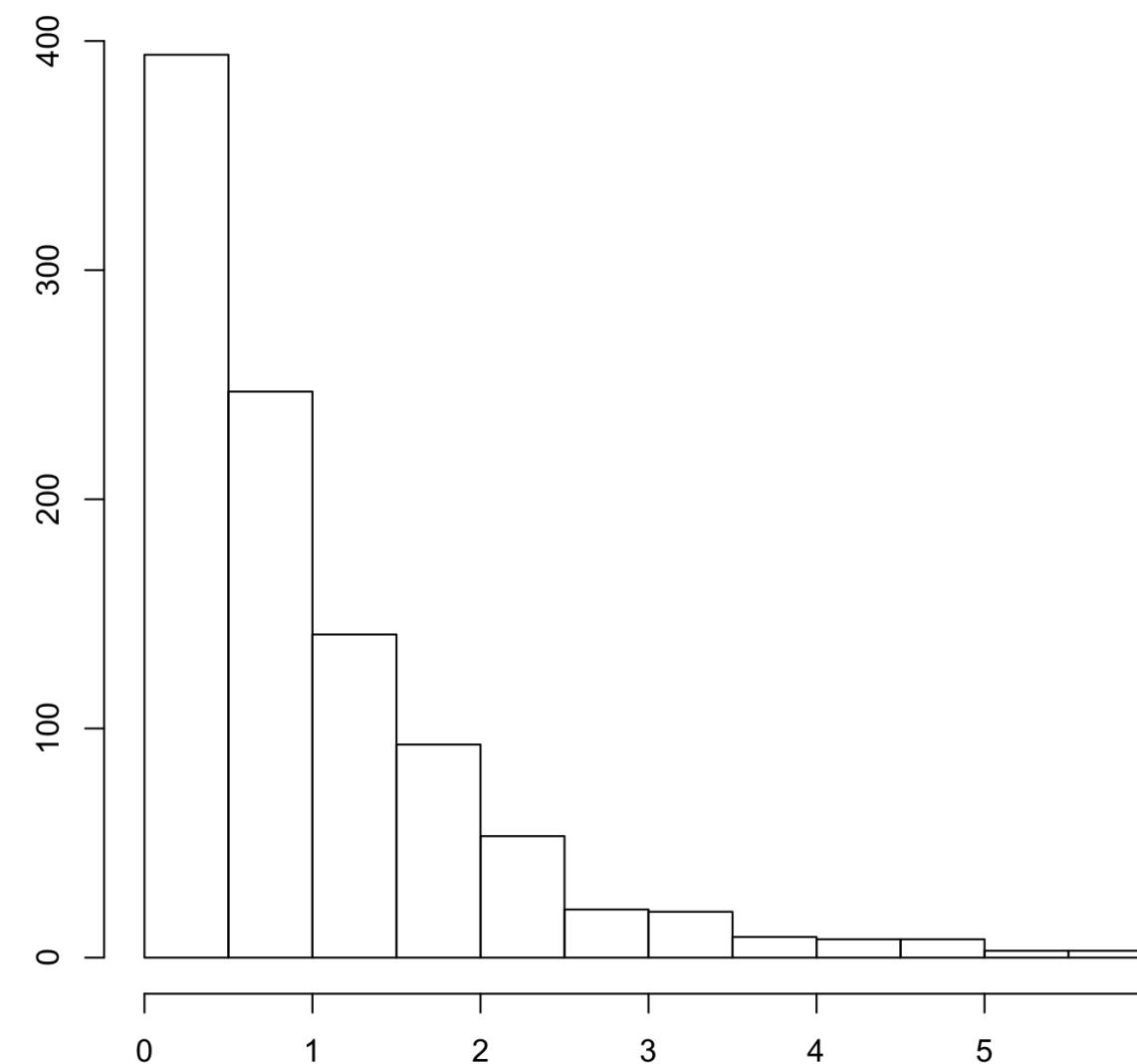
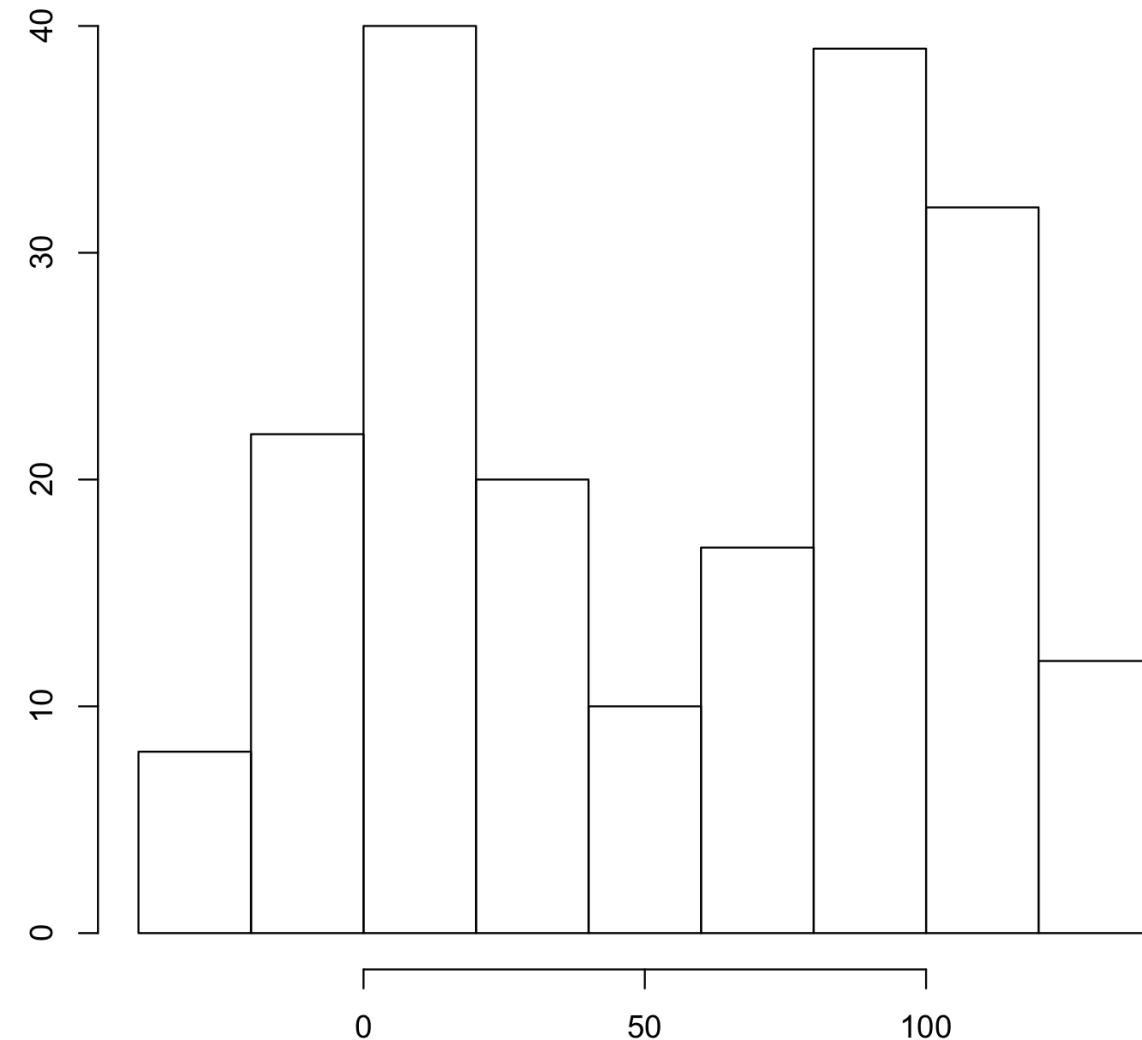
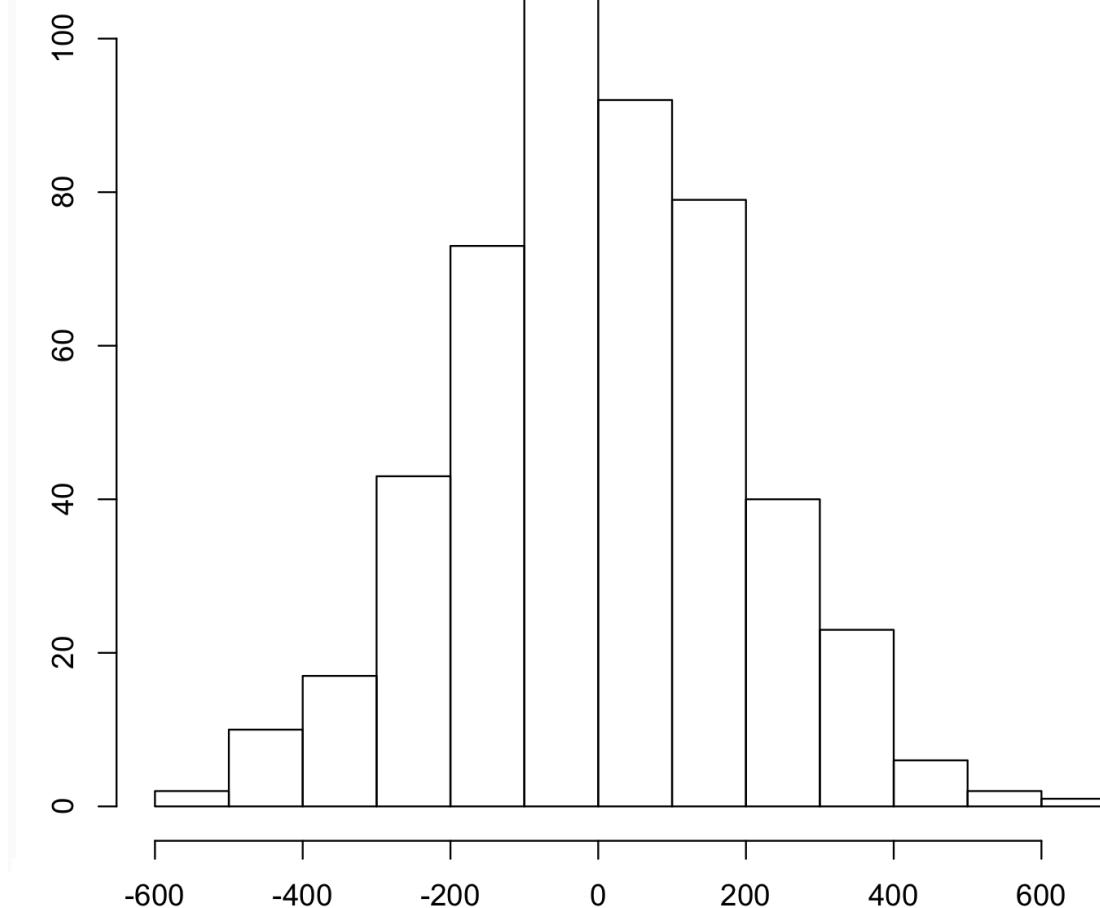
- Does the distribution appear to have one center, or are there multiple peaks?
- Is the distribution symmetric?
- Are the data all close together, or are they largely spread over the domain?

# Center, Shape, and Spread

- Does the distribution appear to have one center, or are there multiple peaks?
- Is the distribution symmetric?
- Are the data all close together, or are they largely spread over the domain?
- Do any data points seem to deviate from typical patterns?

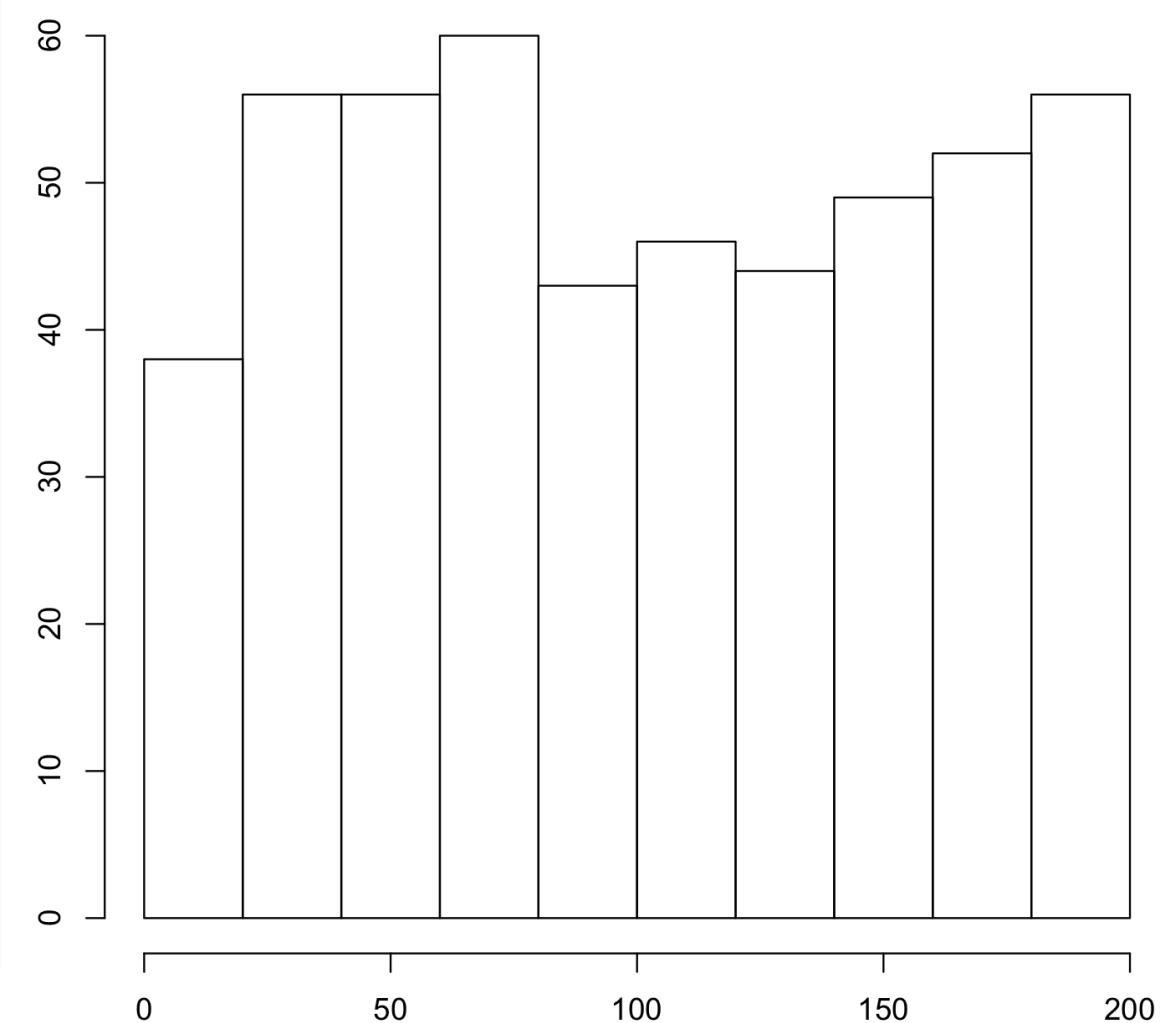
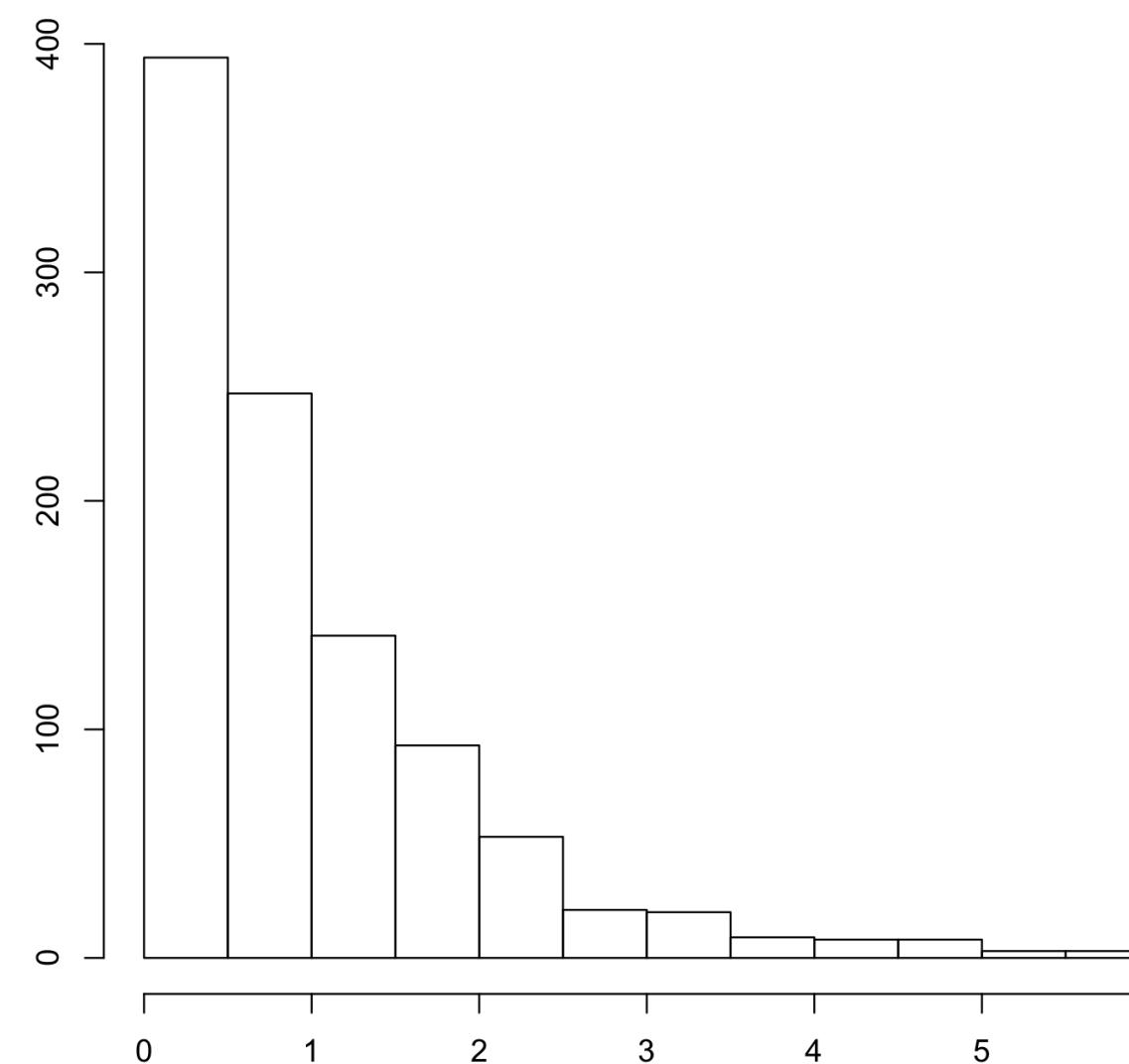
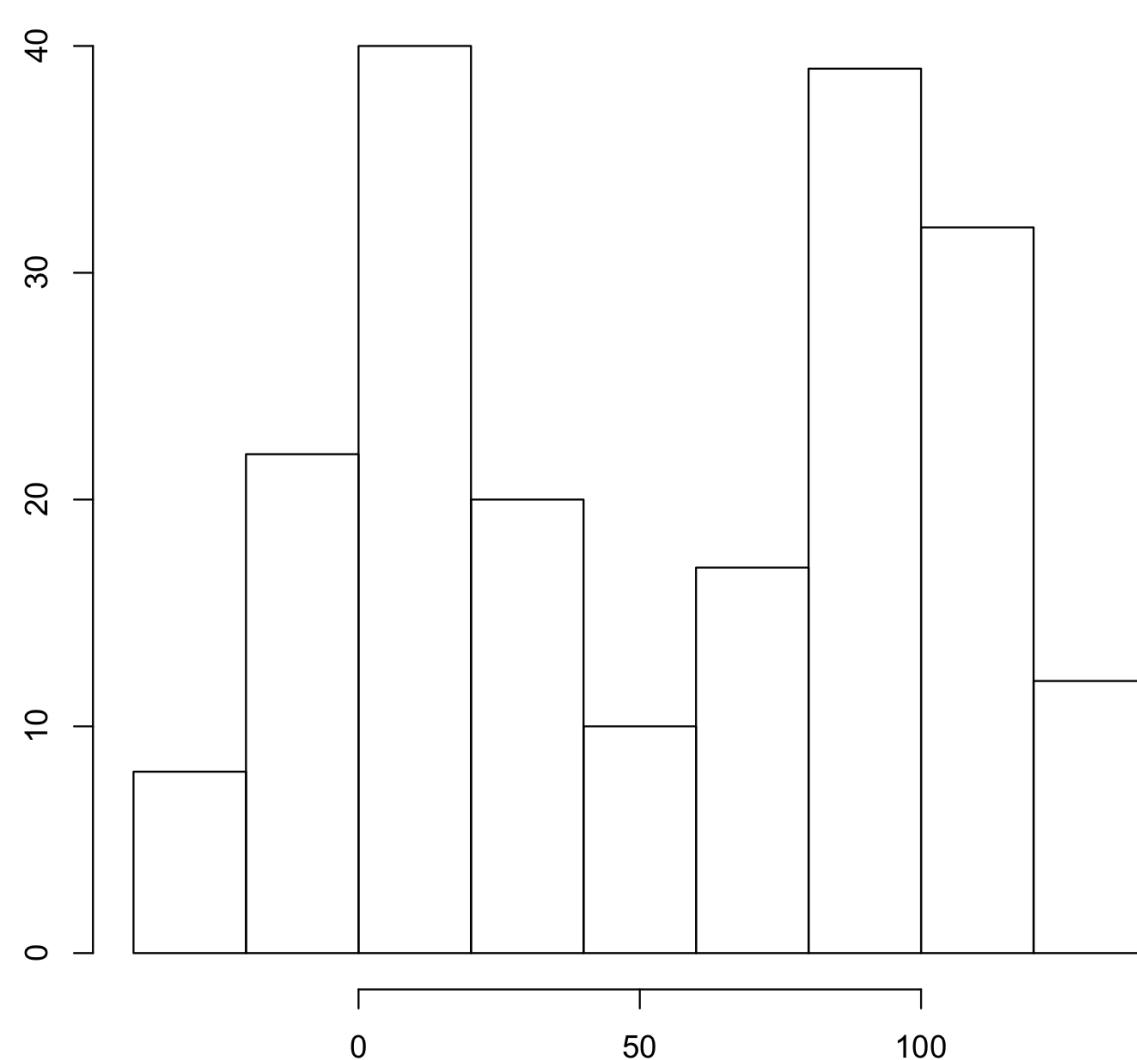
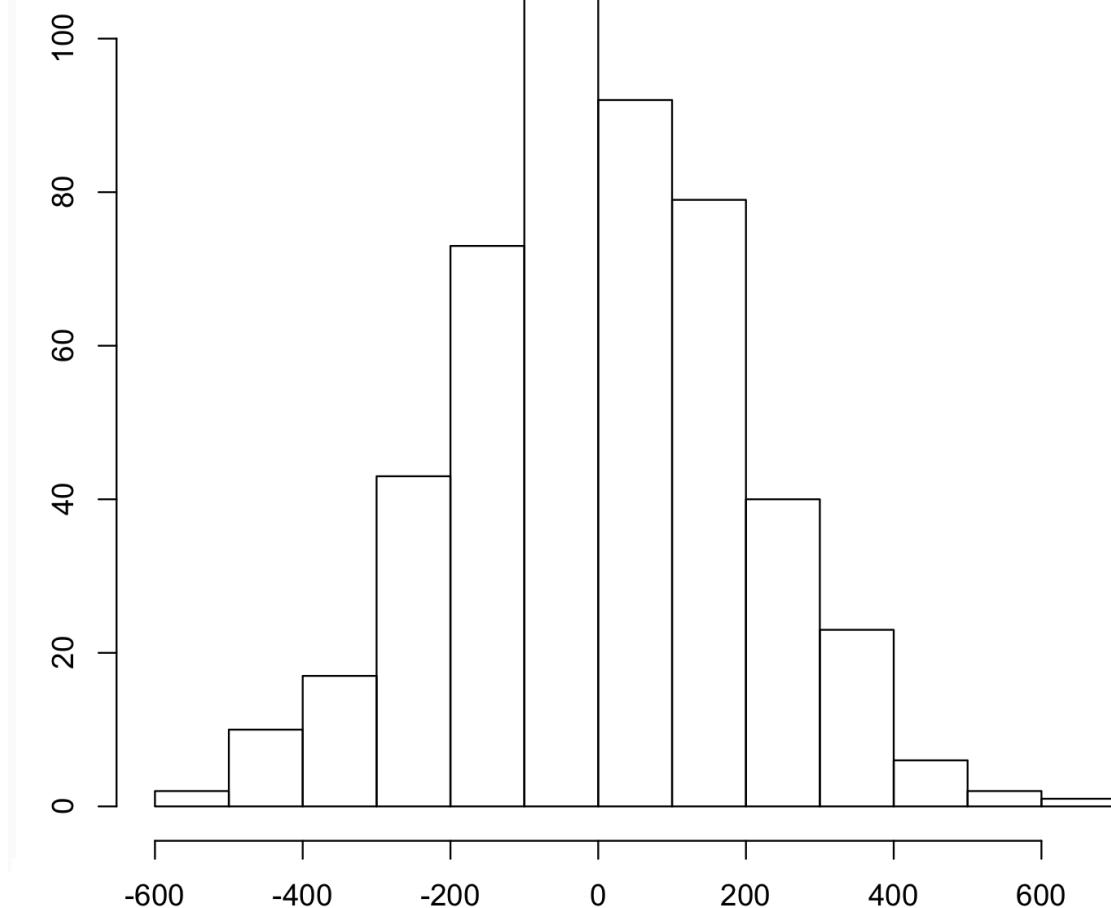
# Center, Shape, and Spread: Examples

A



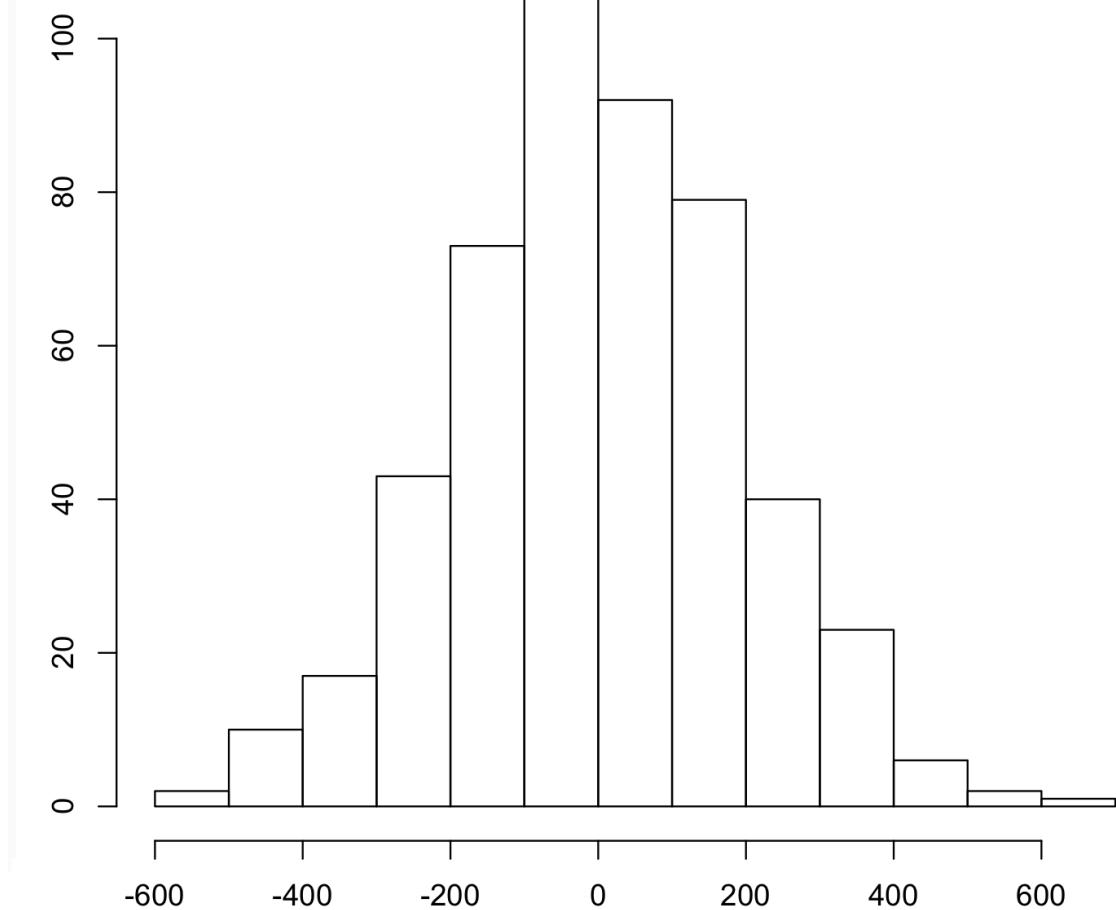
# Center, Shape, and Spread: Examples

B

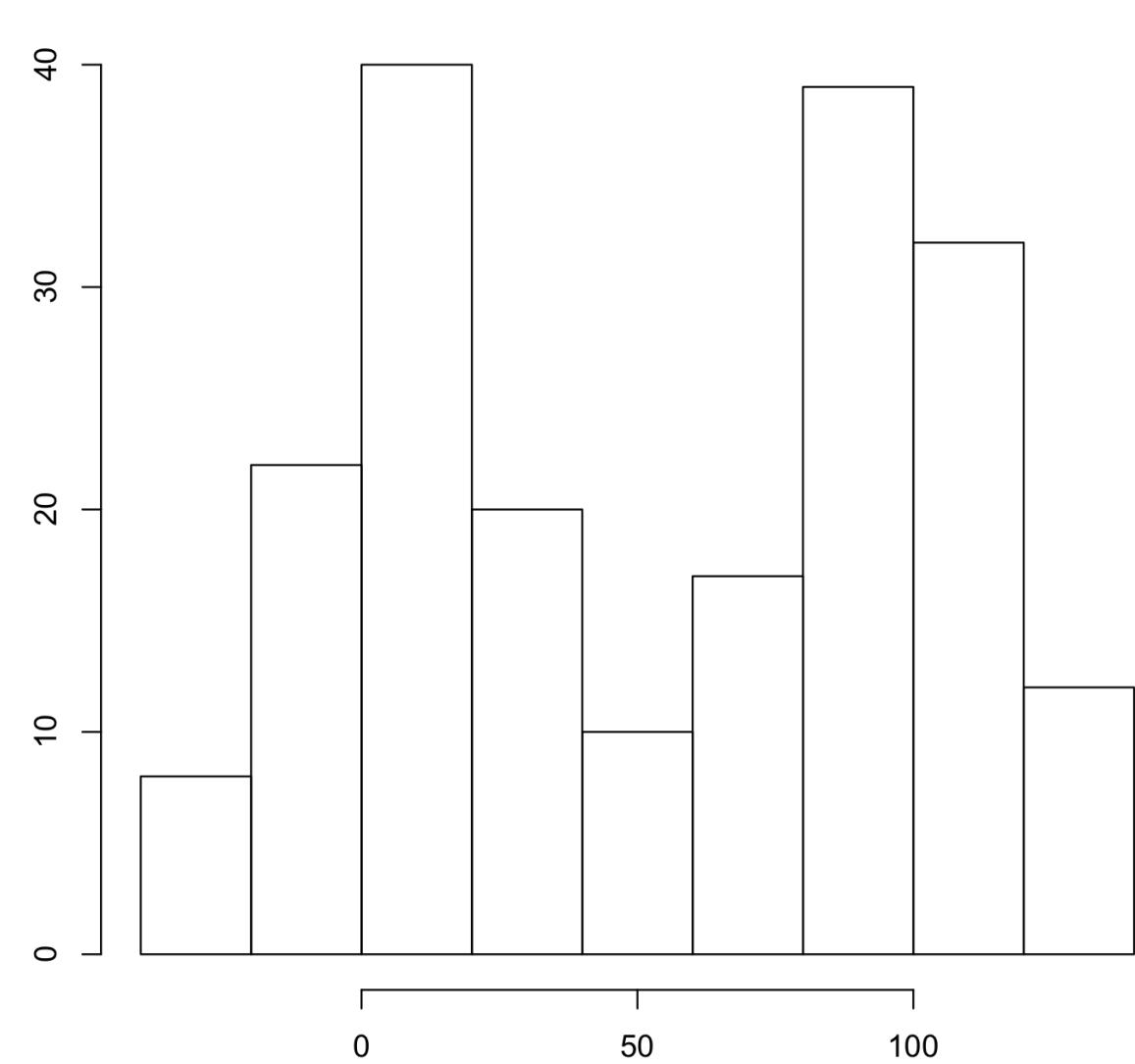


One center  
Symmetric

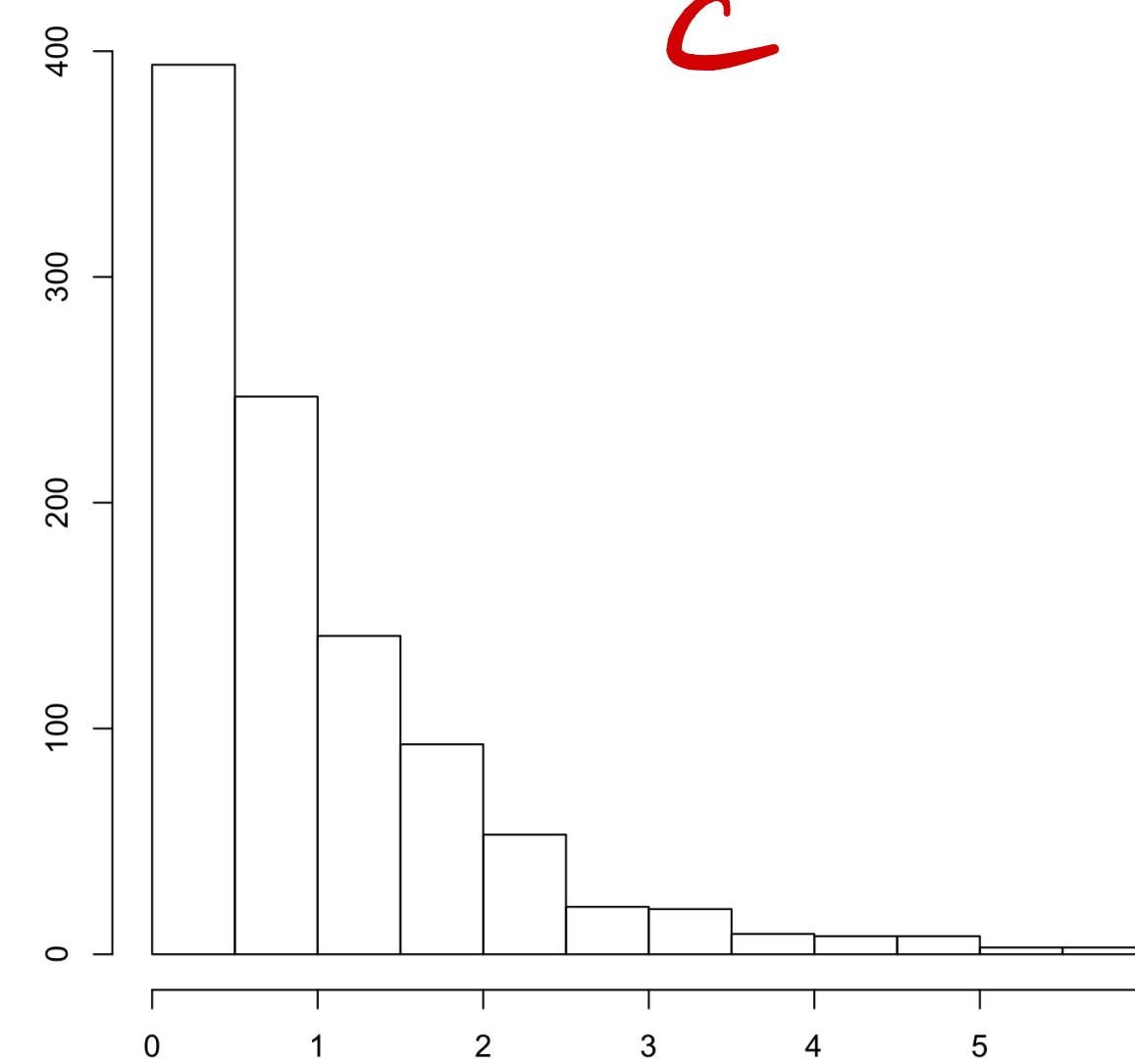
# Center, Shape, and Spread: Examples



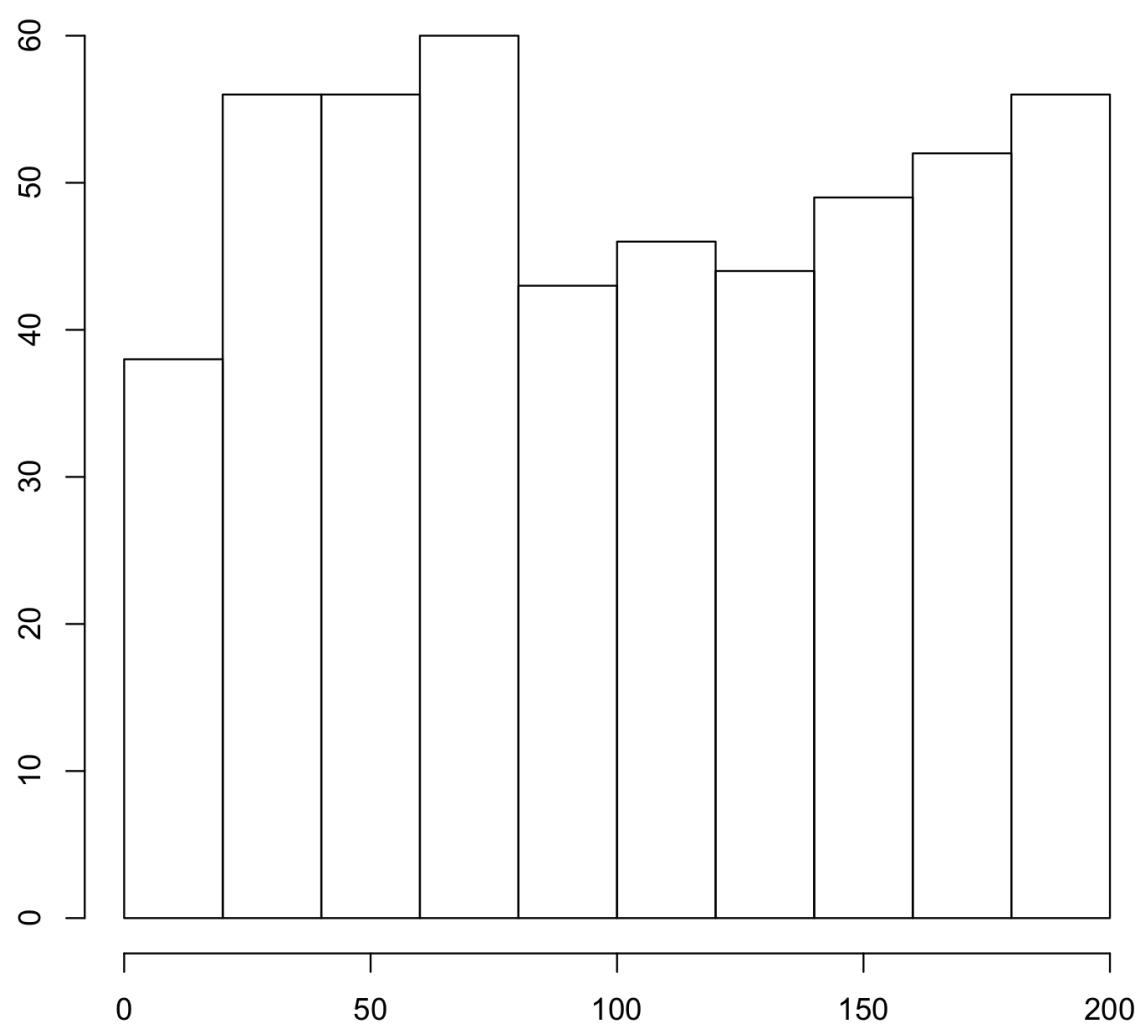
One center  
Symmetric



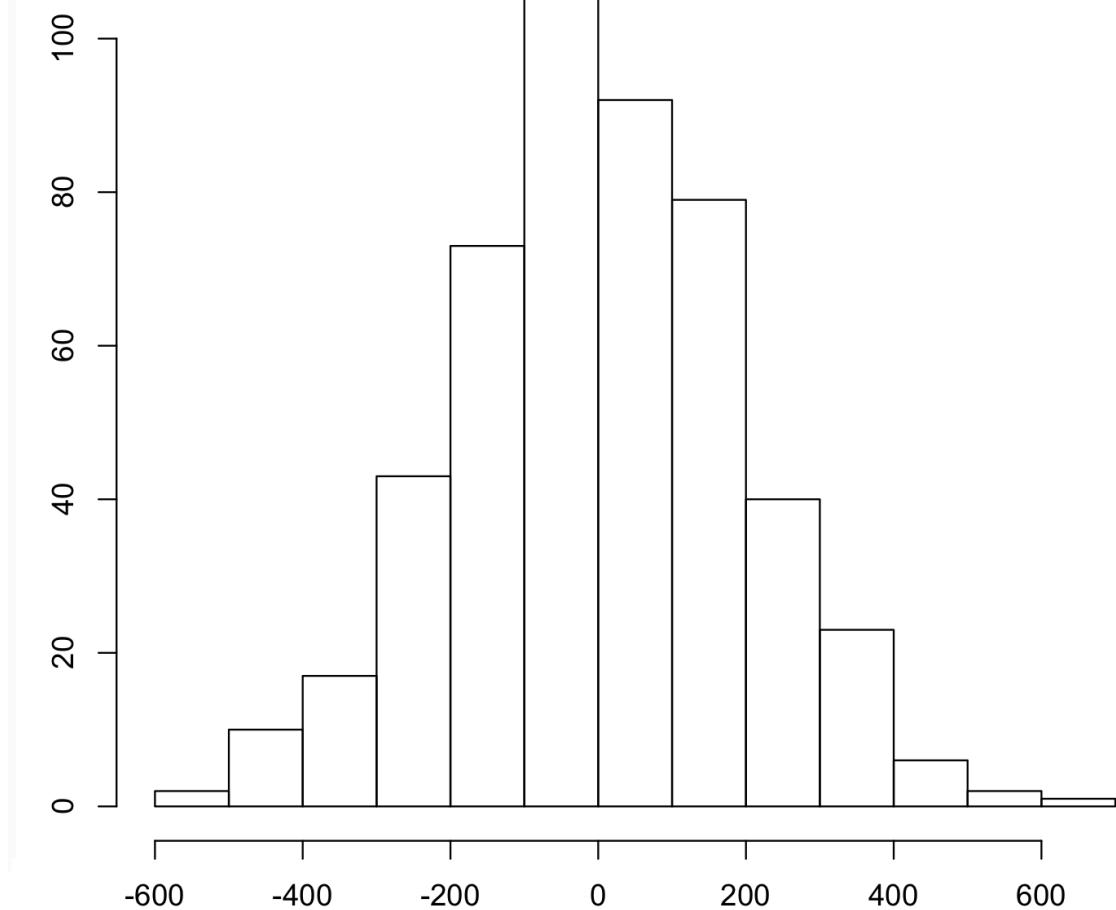
Two centers  
Symmetric



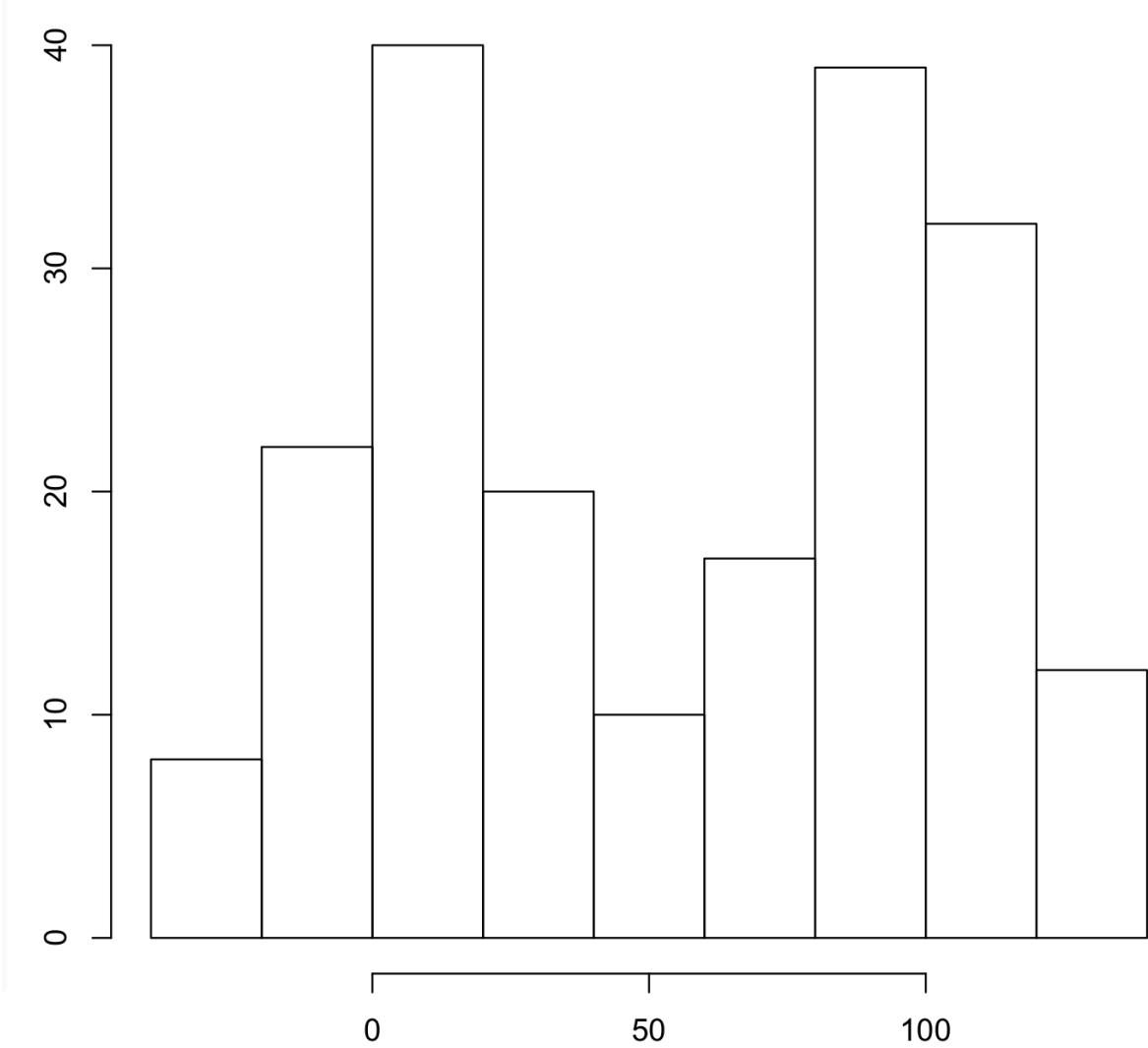
C



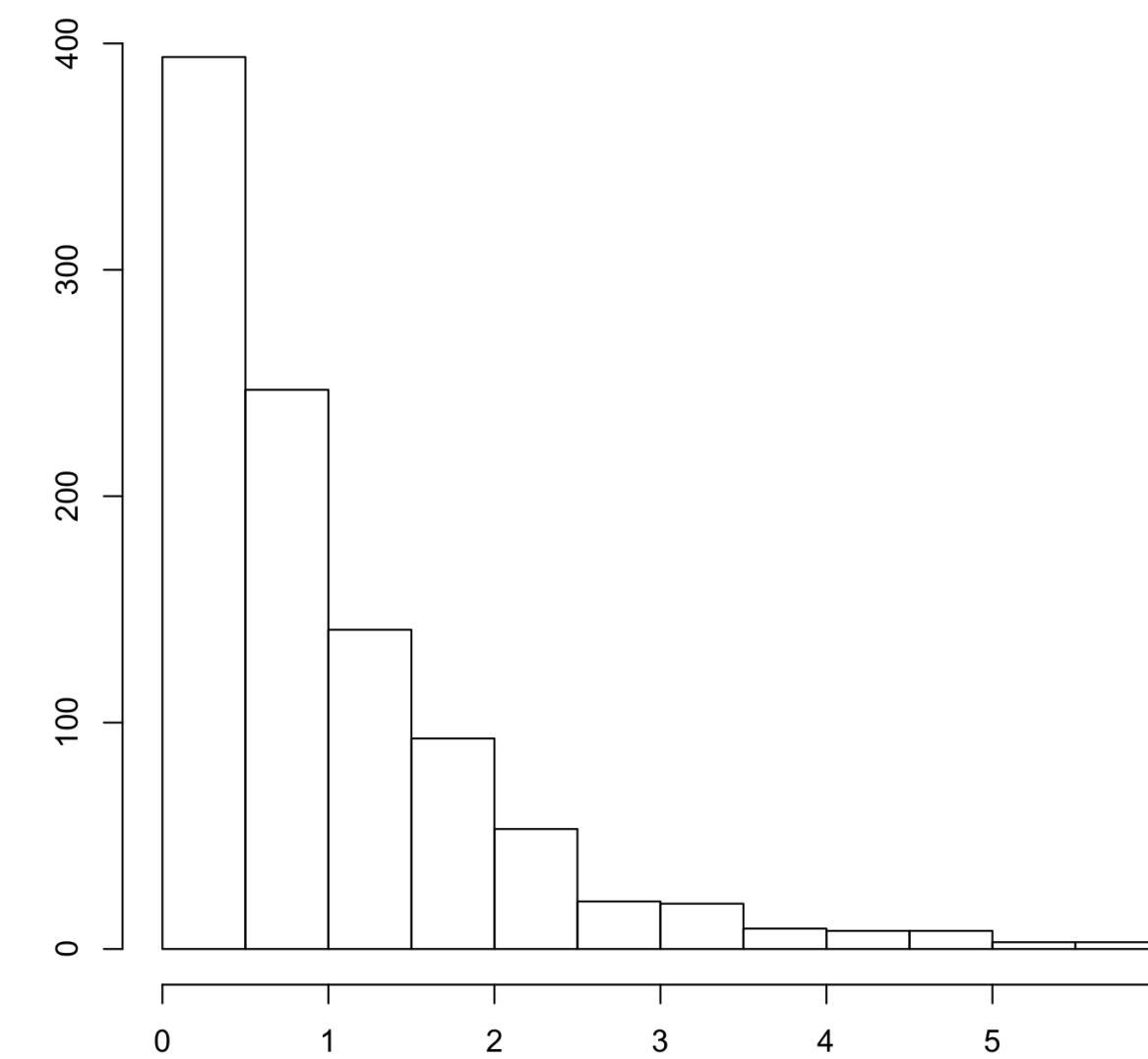
# Center, Shape, and Spread: Examples



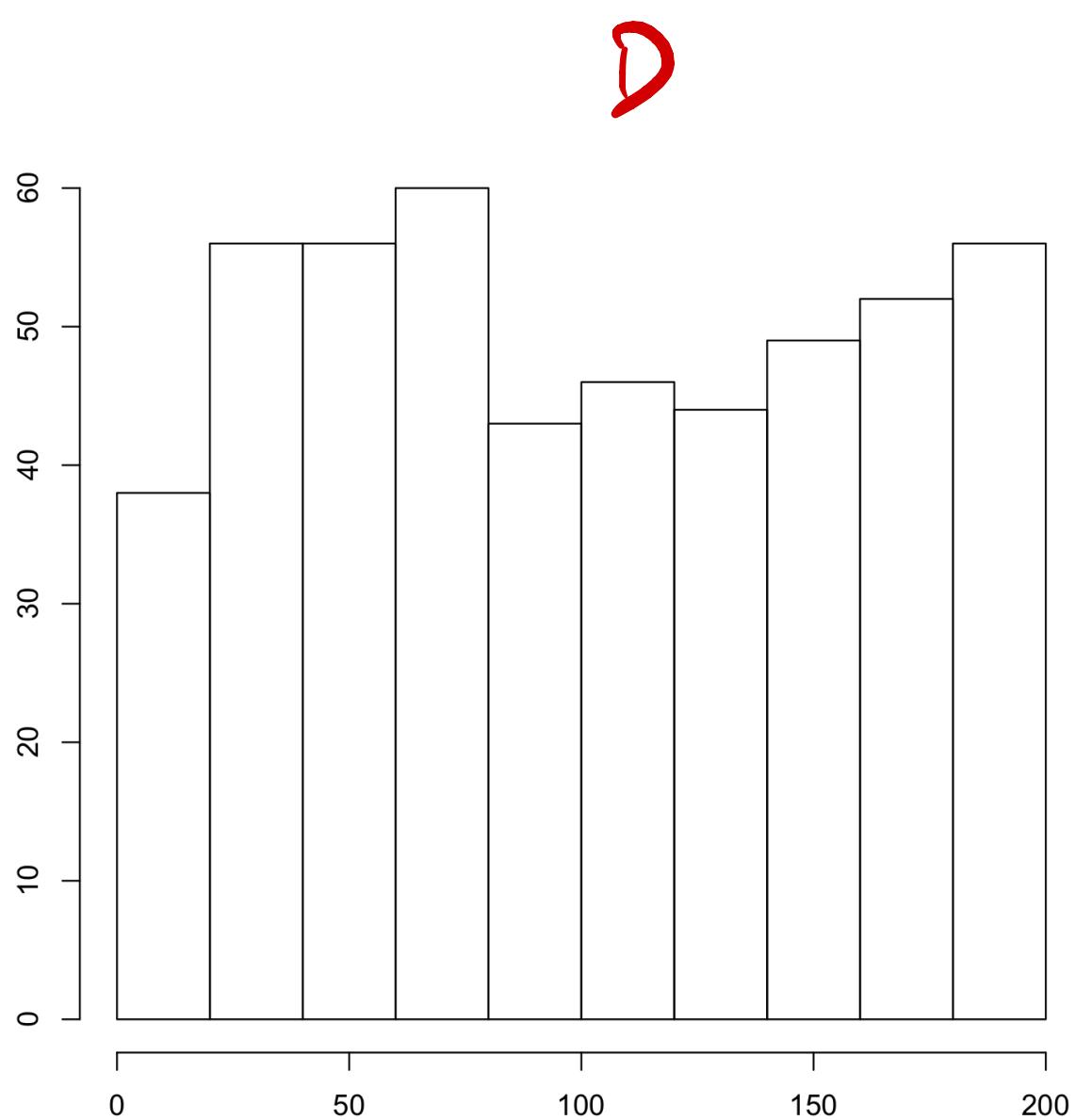
One center  
Symmetric



Two centers  
Symmetric

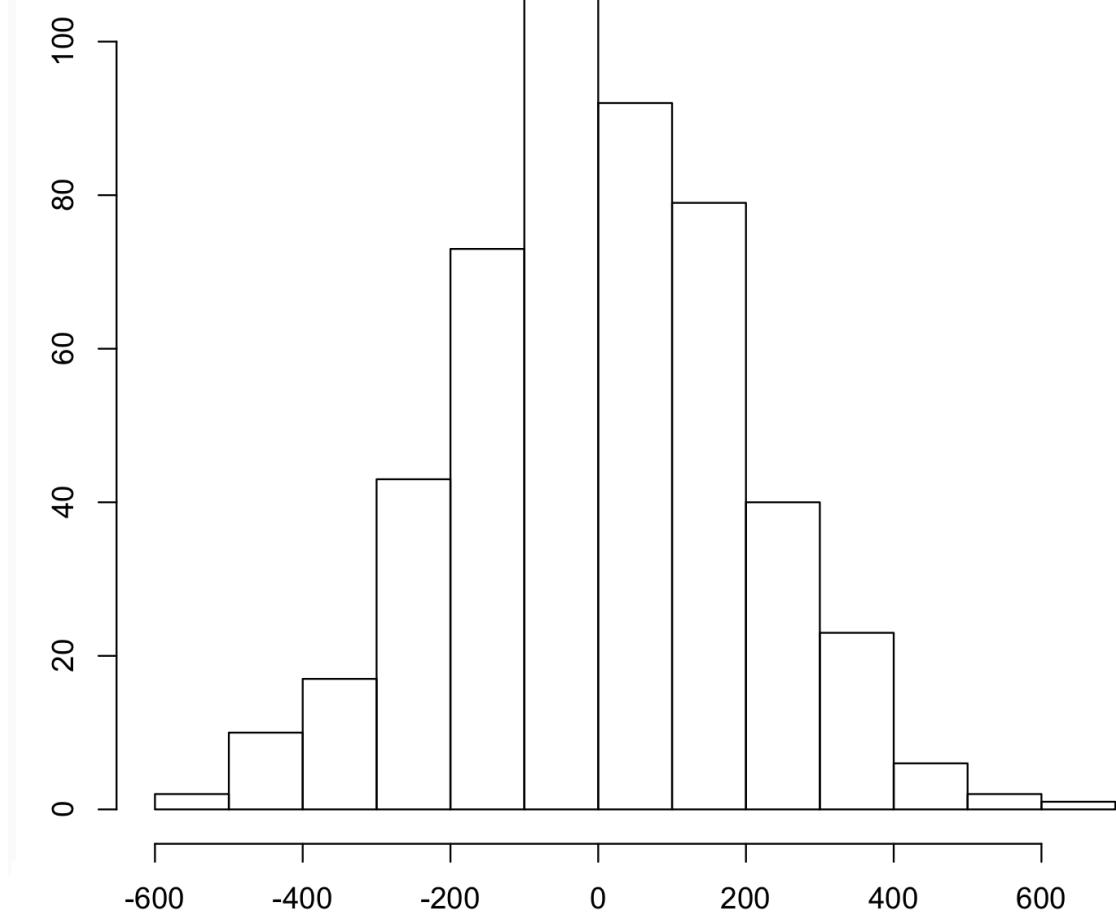


One peak  
Asymmetric

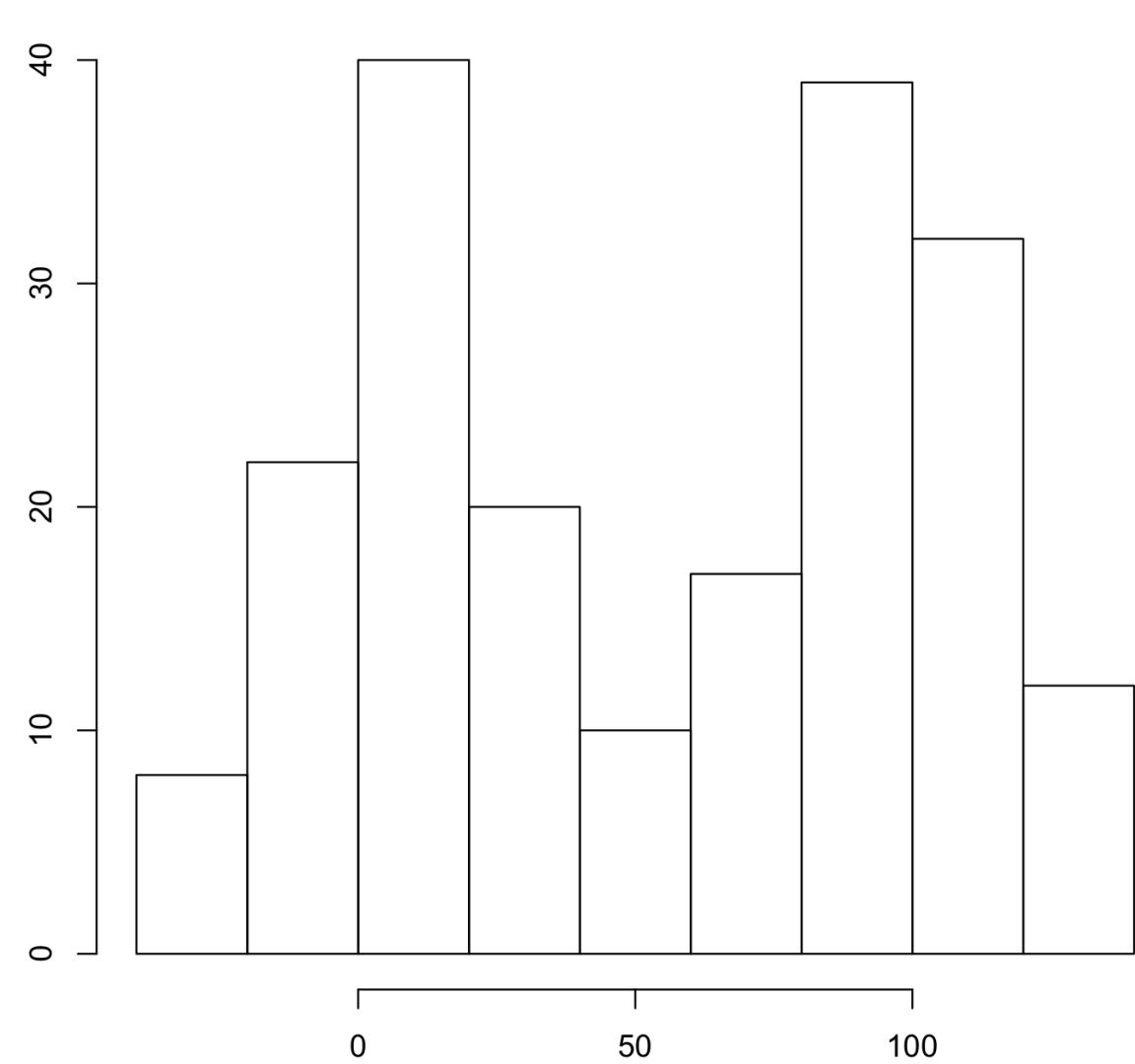


D

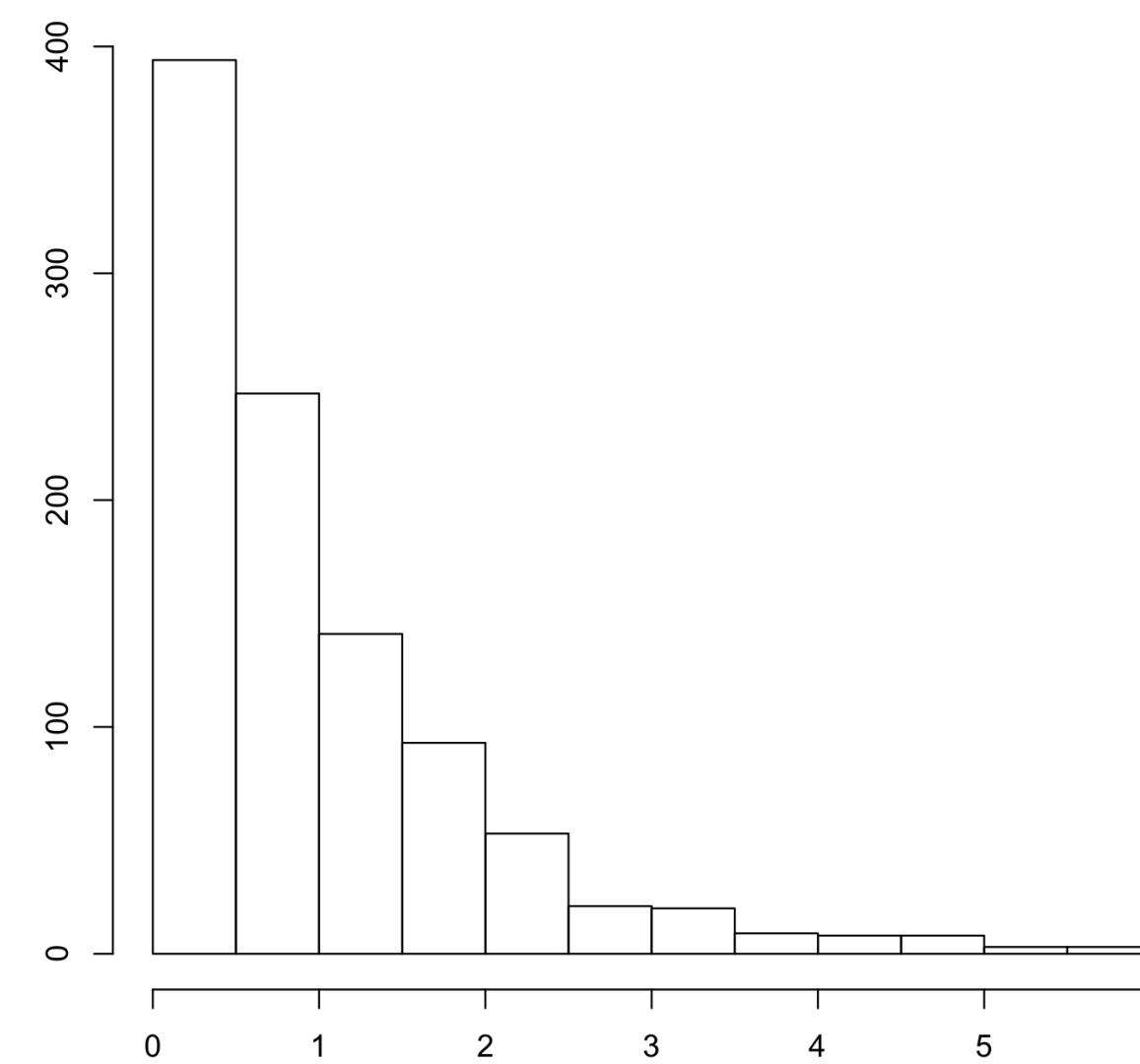
# Center, Shape, and Spread: Examples



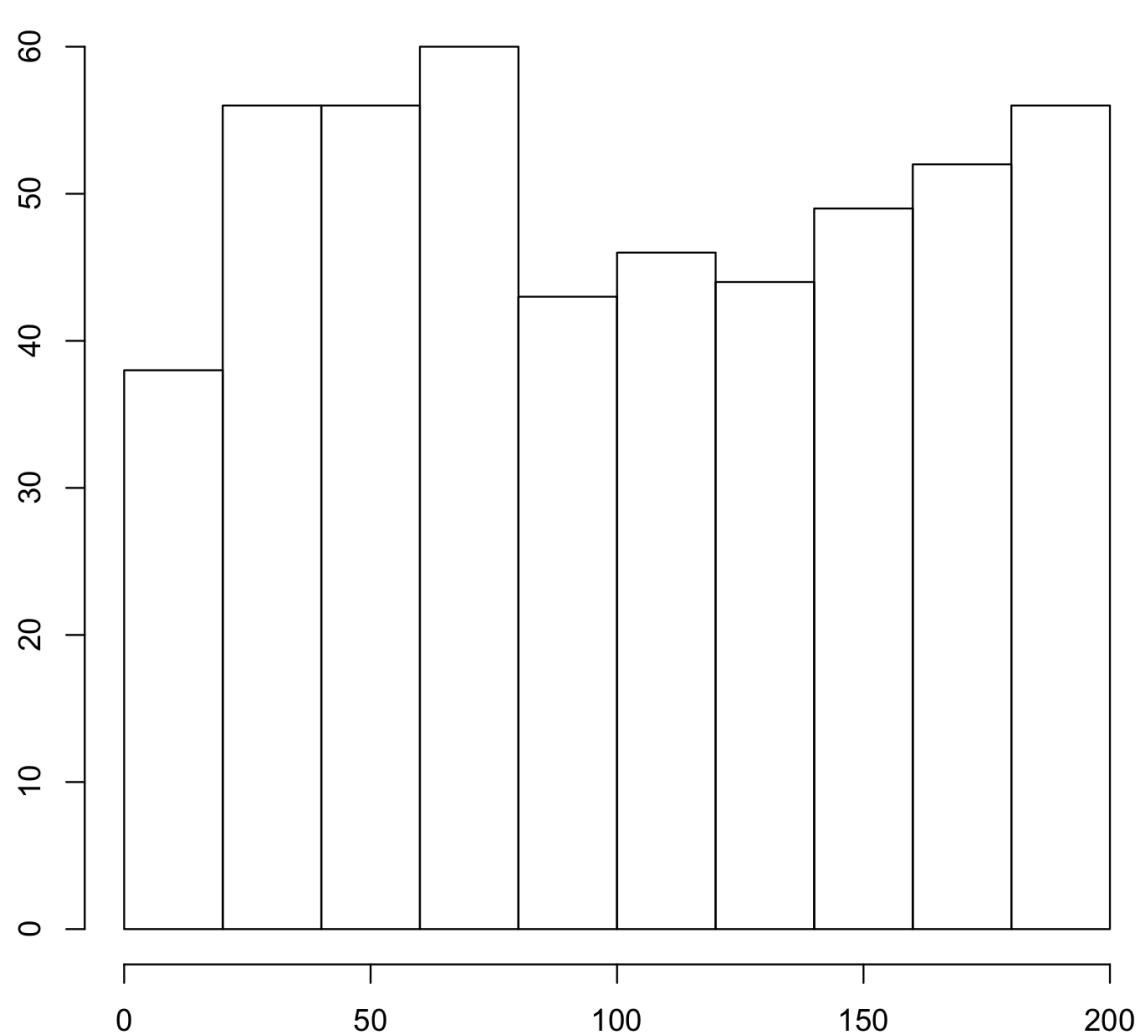
One center  
Symmetric



Two centers  
Symmetric



One peak  
Asymmetric



Large spread  
Uniform

# Measure of Center: Mean

# Measure of Center: Mean

- Mean ( $\bar{x}$ ): The sum of all observations divided by the total number of observations ( $n$ ):

# Measure of Center: Mean

- Mean ( $\bar{x}$ ): The sum of all observations divided by the total number of observations ( $n$ ):

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1 + x_2 + \dots + x_n}{n}$$

# Measure of Center: Mean

- Mean ( $\bar{x}$ ): The sum of all observations divided by the total number of observations ( $n$ ):

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1 + x_2 + \dots + x_n}{n}$$

- The mean is sensitive to outliers

# Measure of Center: Mean

- Mean ( $\bar{x}$ ): The sum of all observations divided by the total number of observations ( $n$ ):

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1 + x_2 + \dots + x_n}{n}$$

- The mean is sensitive to outliers
- Example: Find the mean of the following five heights (in): 66, 46, 68, 71, 72

$\sim 64$

# Measure of Center: Mean

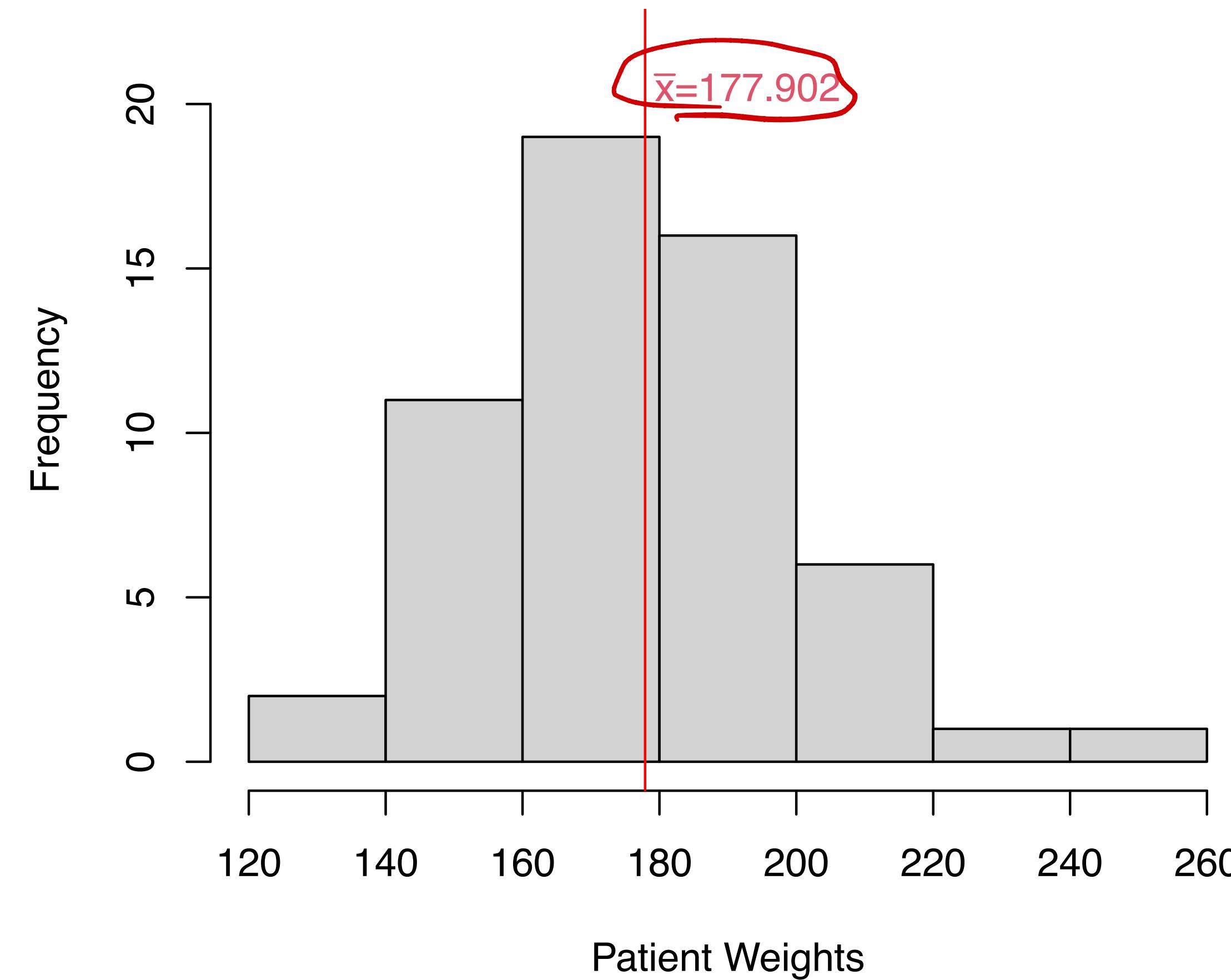
- Mean ( $\bar{x}$ ): The sum of all observations divided by the total number of observations ( $n$ ):

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1 + x_2 + \dots + x_n}{n}$$

- The mean is sensitive to outliers
- Example: Find the mean of the following five heights (in): 66, 46, 68, 71, 72
  - R code: `mean(c(66, 46, 68, 71, 72))`

# Mean for Patient Weights

Histogram of Patient Weights



```
hist(weights, xlab="Patient Weights", main="Histogram of Patient Weights", ylim=c(0,22))
abline(v=mean(weights), col="red")
text(mean(weights)+15,20.5,substitute(paste(bar(x),"=",m),list(m=round(mean(weights),3))),col=2)
```

# Order Statistics

# Order Statistics

- Consider data  $X_1, X_2, \dots, X_n$

# Order Statistics

- Consider data  $X_1, X_2, \dots, X_n$
- We can order these data from smallest to largest

# Order Statistics

- Consider data  $X_1, X_2, \dots, X_n$
- We can order these data from smallest to largest
- Let  $\underline{X_{(k)}}$  be the  $k^{th}$  largest observation in the dataset

# Order Statistics

- Consider data  $X_1, X_2, \dots, X_n$
- We can order these data from smallest to largest
- Let  $X_{(k)}$  be the  $k^{th}$  largest observation in the dataset
- The order statistics are then  $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$

$$\begin{array}{ccc} 5 & 7 & 2 \\ x_{(3)} = 2 \end{array}$$

# Measure of Center: Median

# Measure of Center: Median

- **Median:** The middle value of the ordered data, denoted  $\tilde{X}$ , also called the 50<sup>th</sup> percentile or Q2 (second quartile)

# Measure of Center: Median

- **Median:** The middle value of the ordered data, denoted  $\tilde{X}$ , also called the 50<sup>th</sup> percentile or Q2 (second quartile)
- Rank the data and find the observation for which half the data are greater than or equal to it and half the data are less than or equal to it

# Measure of Center: Median

- **Median:** The middle value of the ordered data, denoted  $\tilde{X}$ , also called the 50<sup>th</sup> percentile or Q2 (second quartile)
- Rank the data and find the observation for which half the data are greater than or equal to it and half the data are less than or equal to it

- As an equation:  $\tilde{x} = \frac{x_{(\lfloor \frac{n+1}{2} \rfloor)} + x_{(\lceil \frac{n+1}{2} \rceil)}}{2}$

n odd :  $x_{(\frac{n+1}{2})}$

n even :  $\frac{x_{(\lfloor \frac{n+1}{2} \rfloor)} + x_{(\lceil \frac{n+1}{2} \rceil)}}{2}$

# Measure of Center: Median

- **Median:** The middle value of the ordered data, denoted  $\tilde{X}$ , also called the 50<sup>th</sup> percentile or Q2 (second quartile)
- Rank the data and find the observation for which half the data are greater than or equal to it and half the data are less than or equal to it
- As an equation: 
$$\tilde{x} = \frac{x_{(\lfloor \frac{n+1}{2} \rfloor)} + x_{(\lceil \frac{n+1}{2} \rceil)}}{2}$$
- The median is not as sensitive to outliers (it is more robust than the mean)

# Measure of Center: Median

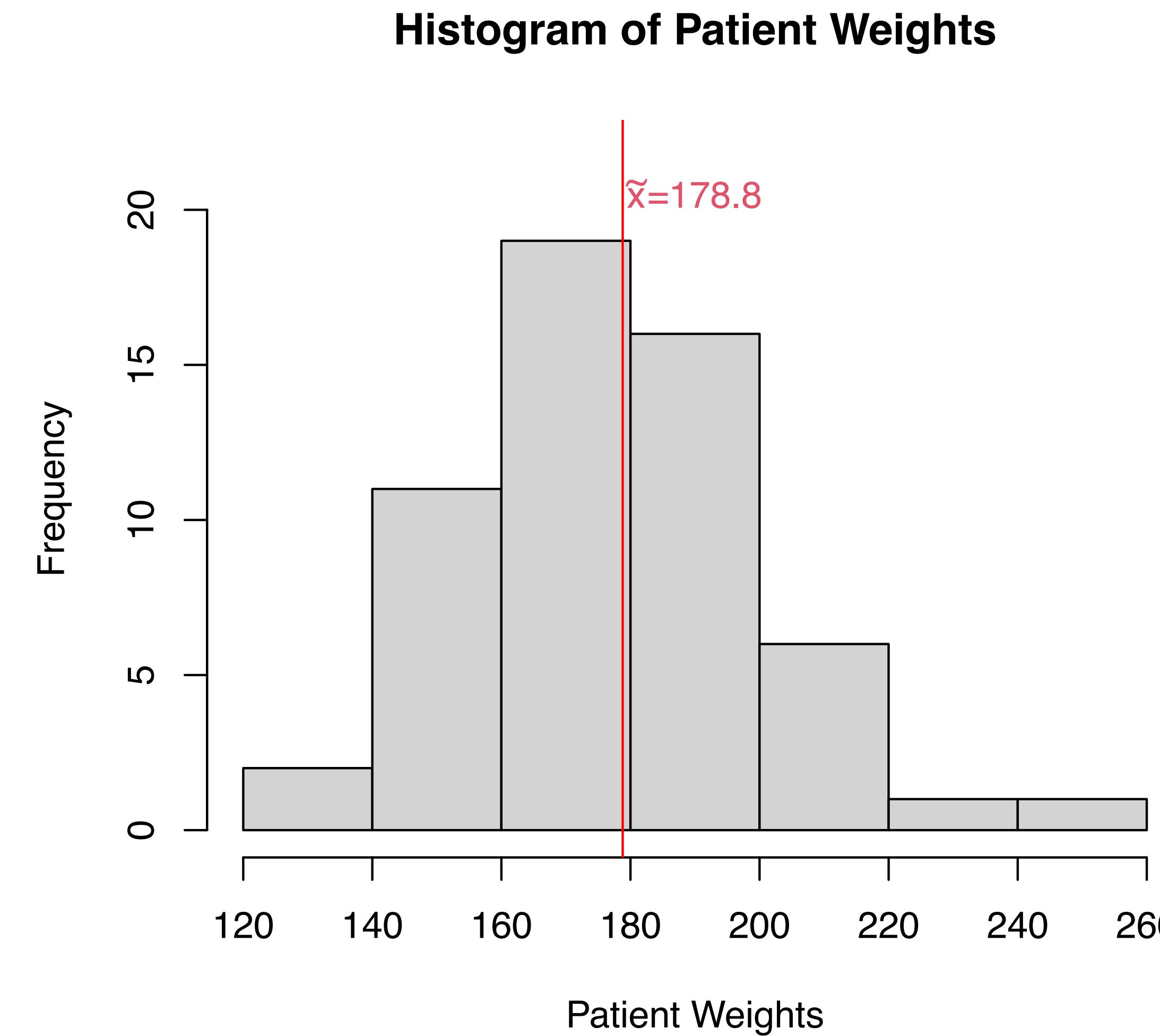
- **Median:** The middle value of the ordered data, denoted  $\tilde{X}$ , also called the 50<sup>th</sup> percentile or Q2 (second quartile)
- Rank the data and find the observation for which half the data are greater than or equal to it and half the data are less than or equal to it
- As an equation: 
$$\tilde{x} = \frac{x_{(\lfloor \frac{n+1}{2} \rfloor)} + x_{(\lceil \frac{n+1}{2} \rceil)}}{2}$$
- The median is not as sensitive to outliers (it is more robust than the mean)
- Example: Find the median of the following five heights (in): 66, 46, 68, 71, 72

46, 66, 68, 71, 72

# Measure of Center: Median

- **Median:** The middle value of the ordered data, denoted  $\tilde{X}$ , also called the 50<sup>th</sup> percentile or Q2 (second quartile)
- Rank the data and find the observation for which half the data are greater than or equal to it and half the data are less than or equal to it
- As an equation: 
$$\tilde{x} = \frac{x_{(\lfloor \frac{n+1}{2} \rfloor)} + x_{(\lceil \frac{n+1}{2} \rceil)}}{2}$$
- The median is not as sensitive to outliers (it is more robust than the mean)
- Example: Find the median of the following five heights (in): 66, 46, 68, 71, 72
  - R code: `median(c(66, 46, 68, 71, 72))` c(1, 2, 3, 4)

# Median for Patient Weights



```
hist(weights, xlab="Patient Weights", main="Histogram of Patient Weights", ylim=c(0,22))
abline(v=median(weights), col="red")
text(median(weights)+11,20.5, substitute(paste(tilde(x),"=",m)), list(m=round(median(weights),3))), col=2)
```

# Measure of Center: Trimmed Mean

# Measure of Center: Trimmed Mean

- **$K\%$  trimmed mean:** Sample mean of the data remaining after removing the highest  $K\%$  and lowest  $K\%$  of observations

# Measure of Center: Trimmed Mean

- **K% trimmed mean:** Sample mean of the data remaining after removing the highest K% and lowest K% of observations
- Example: Find the 10% trimmed mean of the sample 5, 15, 18, ~~2~~, 17, 10, ~~23~~, 20, 17, 16

$$\frac{\sum x_i}{8}$$

# Measure of Center: Trimmed Mean

- **K% trimmed mean:** Sample mean of the data remaining after removing the highest K% and lowest K% of observations
- Example: Find the 10% trimmed mean of the sample 5, 15, 18, 2, 17, 10, 23, 20, 17, 16
  - $\bar{x}_{10\%} =$

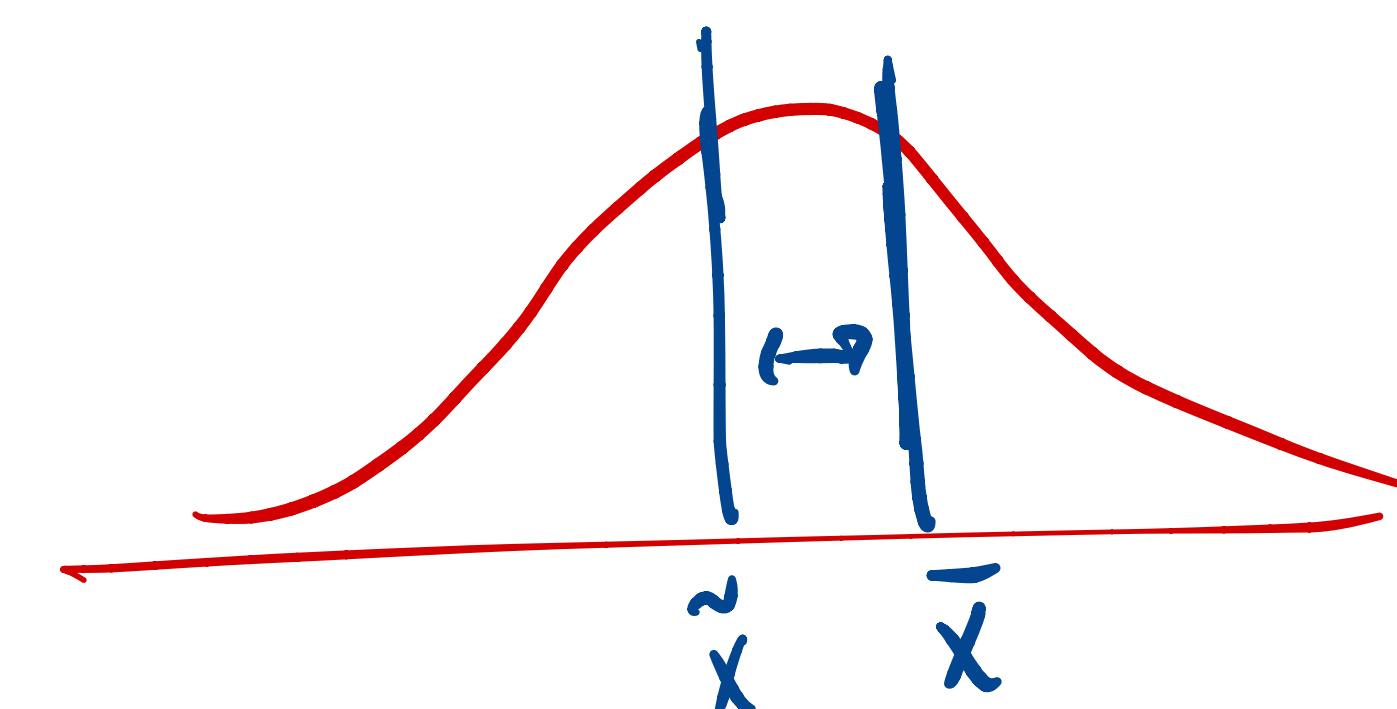
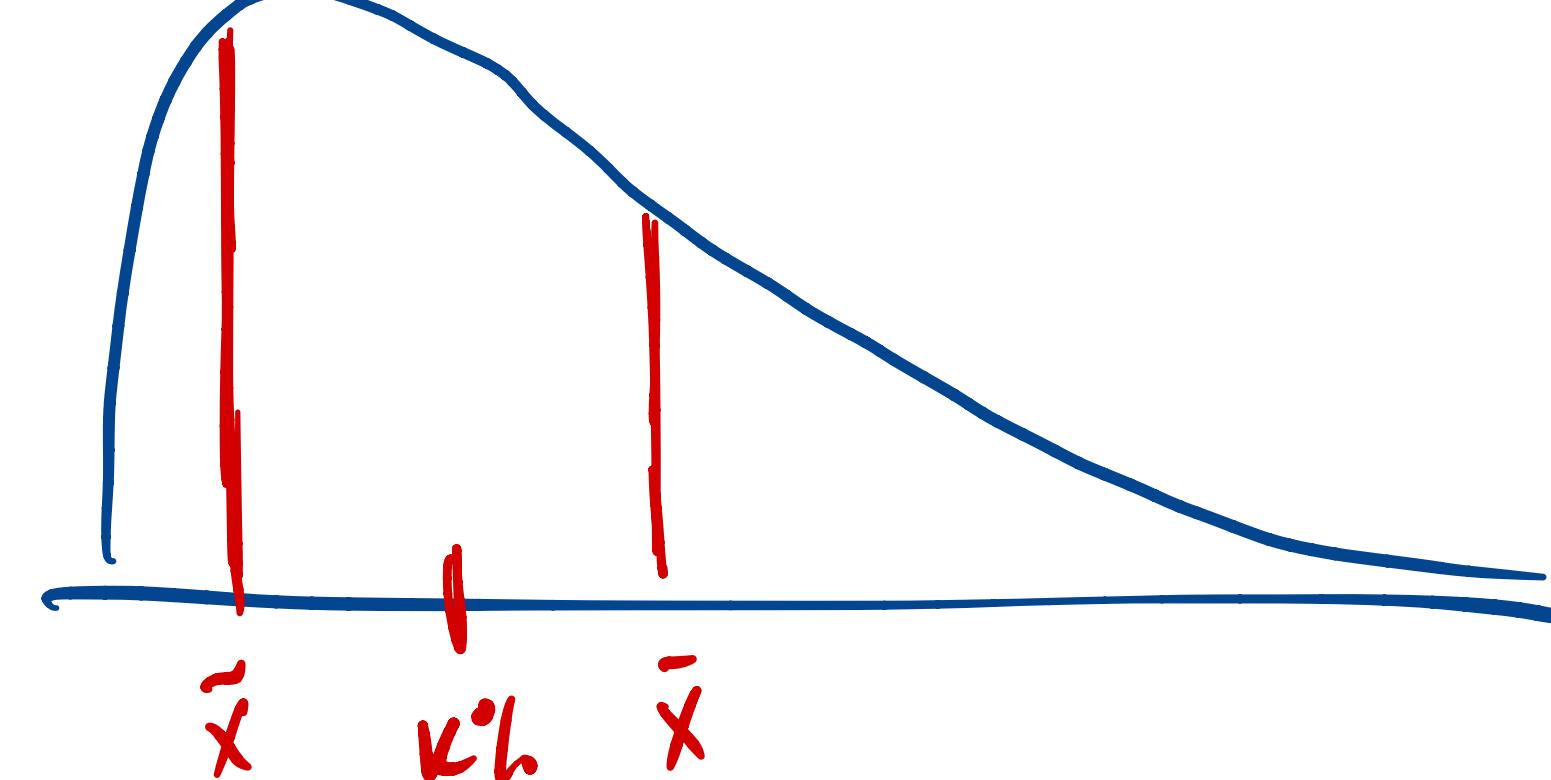
# Measure of Center: Trimmed Mean

- **K% trimmed mean:** Sample mean of the data remaining after removing the highest K% and lowest K% of observations
- Example: Find the 10% trimmed mean of the sample 5, 15, 18, 2, 17, 10, 23, 20, 17, 16
  - $\bar{x}_{10\%} =$
  - R code:

```
x <- c(5,15,18,2,17,10,23,20,17,16)
mean(x, trim=0.1)
```

# Measure of Center: Trimmed Mean

# Measure of Center: Trimmed Mean



- $K = 0\%$  is the sample mean;  $K = 50\%$  is the median

# Measure of Center: Trimmed Mean

- $K = 0\%$  is the sample mean;  $K = 50\%$  is the median
- The trimmed mean is a compromise between the mean and median

# Measure of Center: Trimmed Mean

- $K = 0\%$  is the sample mean;  $K = 50\%$  is the median
- The trimmed mean is a compromise between the mean and median
- Trimmed mean is relatively stable and not too sensitive to outliers

# Measure of Center: Mode

# Measure of Center: Mode

- **Mode:** The observation that occurs most frequently

# Measure of Center: Mode

- **Mode:** The observation that occurs most frequently
- If all values occur the same number of times, then there is no unique mode

# Measure of Center: Mode

- **Mode:** The observation that occurs most frequently
- If all values occur the same number of times, then there is no unique mode
- Example:

# Measure of Center: Mode

- **Mode:** The observation that occurs most frequently
- If all values occur the same number of times, then there is no unique mode
- Example:
  - Find the mode: 3.1, 3.2, 4.5, 5.1, 5.9, 6.0

# Measure of Center: Mode

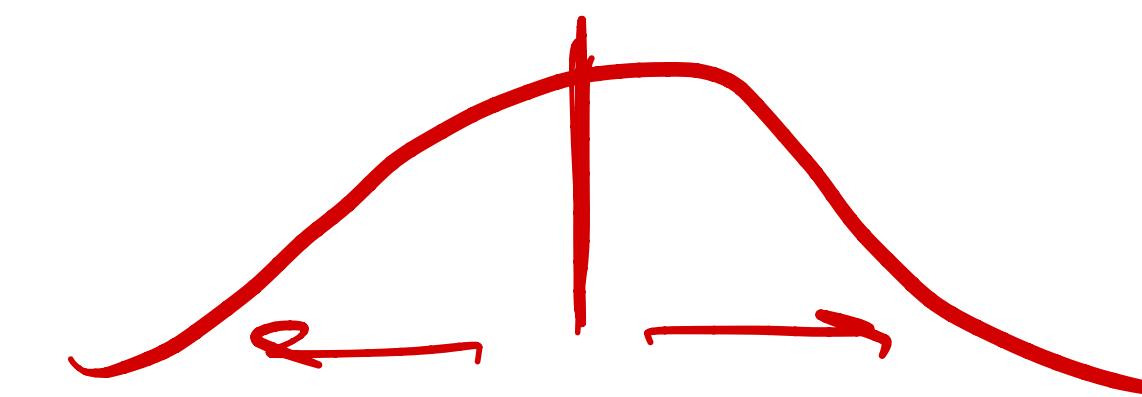
- **Mode:** The observation that occurs most frequently
- If all values occur the same number of times, then there is no unique mode
- Example:
  - Find the mode: 3.1, 3.2, 4.5, 5.1, 5.9, 6.0
  - Find the mode: 7.1, 7.8, 7.8, 9.1, 9.3, 9.4, 9.4, 9.4

# Measures of Center

# Measures of Center

- The “best” measure of central tendency (mean, trimmed mean, median, mode) generally depends on the distribution of values

# Measures of Center



- The “best” measure of central tendency (**mean, trimmed mean, median, mode**) generally depends on the distribution of values
- If a distribution is symmetric and unimodal, then the mean, median, and mode should all approximately be the same

# Measures of Center

- The “best” measure of central tendency (mean, trimmed mean, median, mode) generally depends on the distribution of values
- If a distribution is symmetric and unimodal, then the mean, median, and mode should all approximately be the same
- **Unimodal:** A histogram or density plot only has one peak

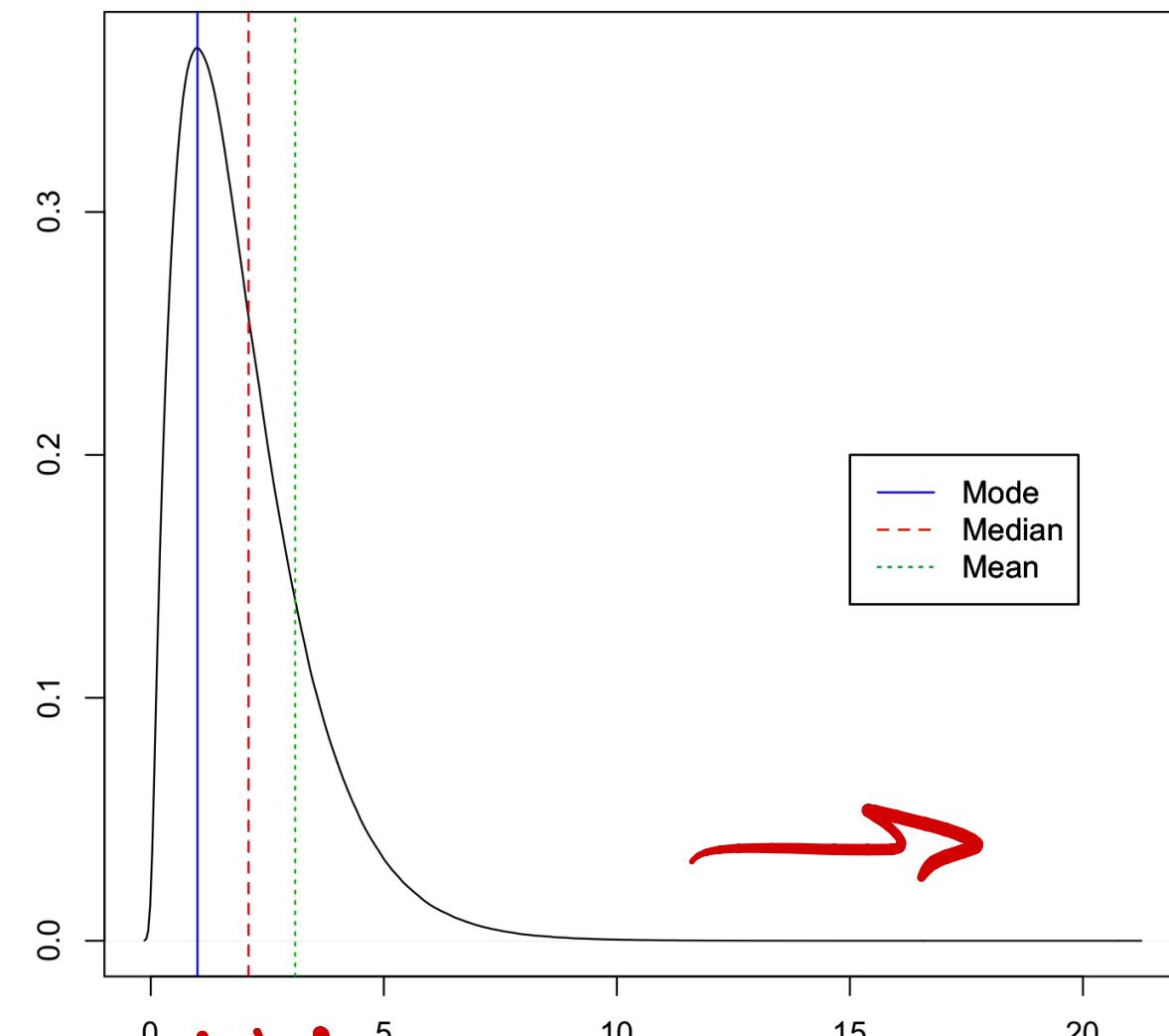
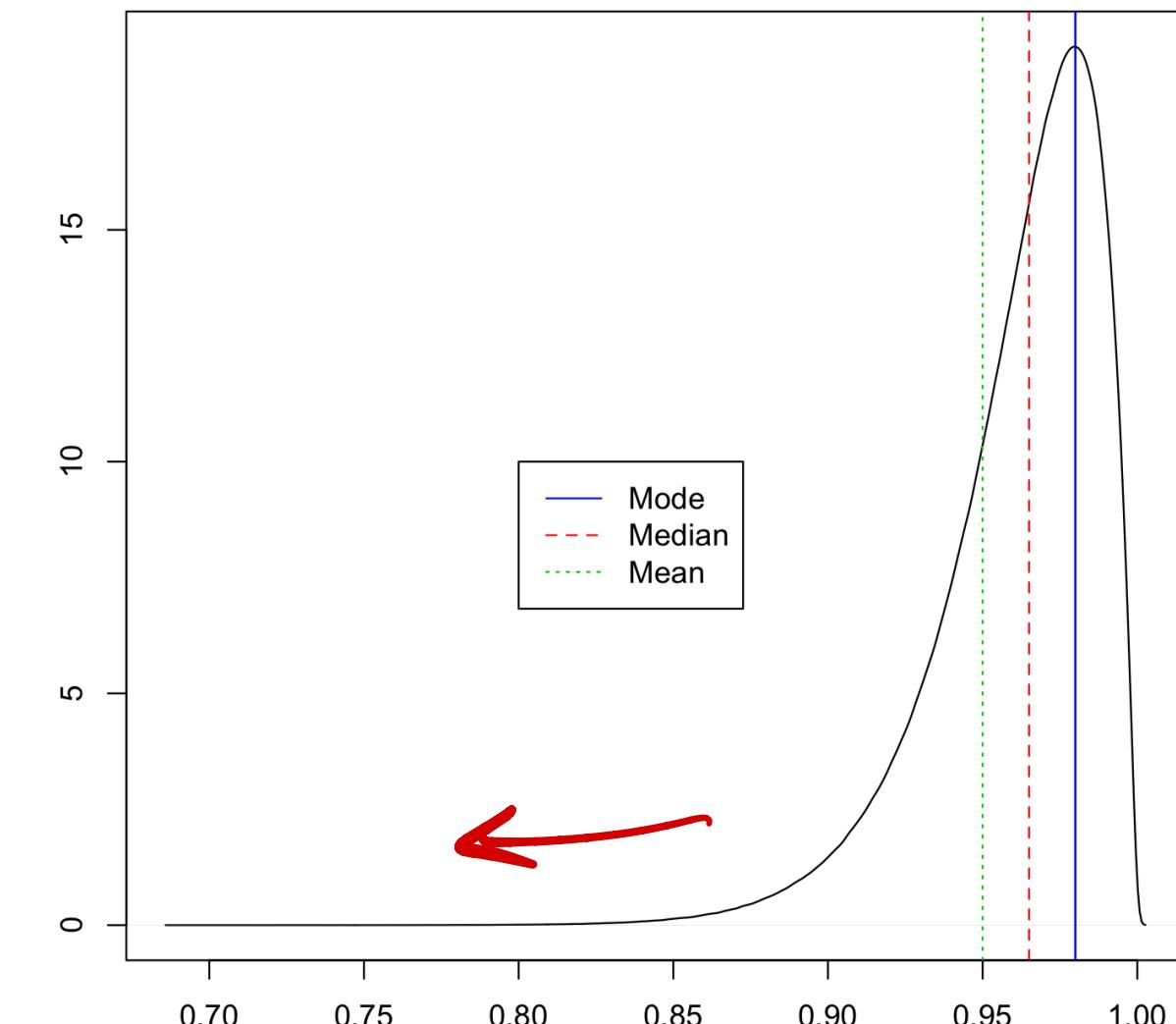
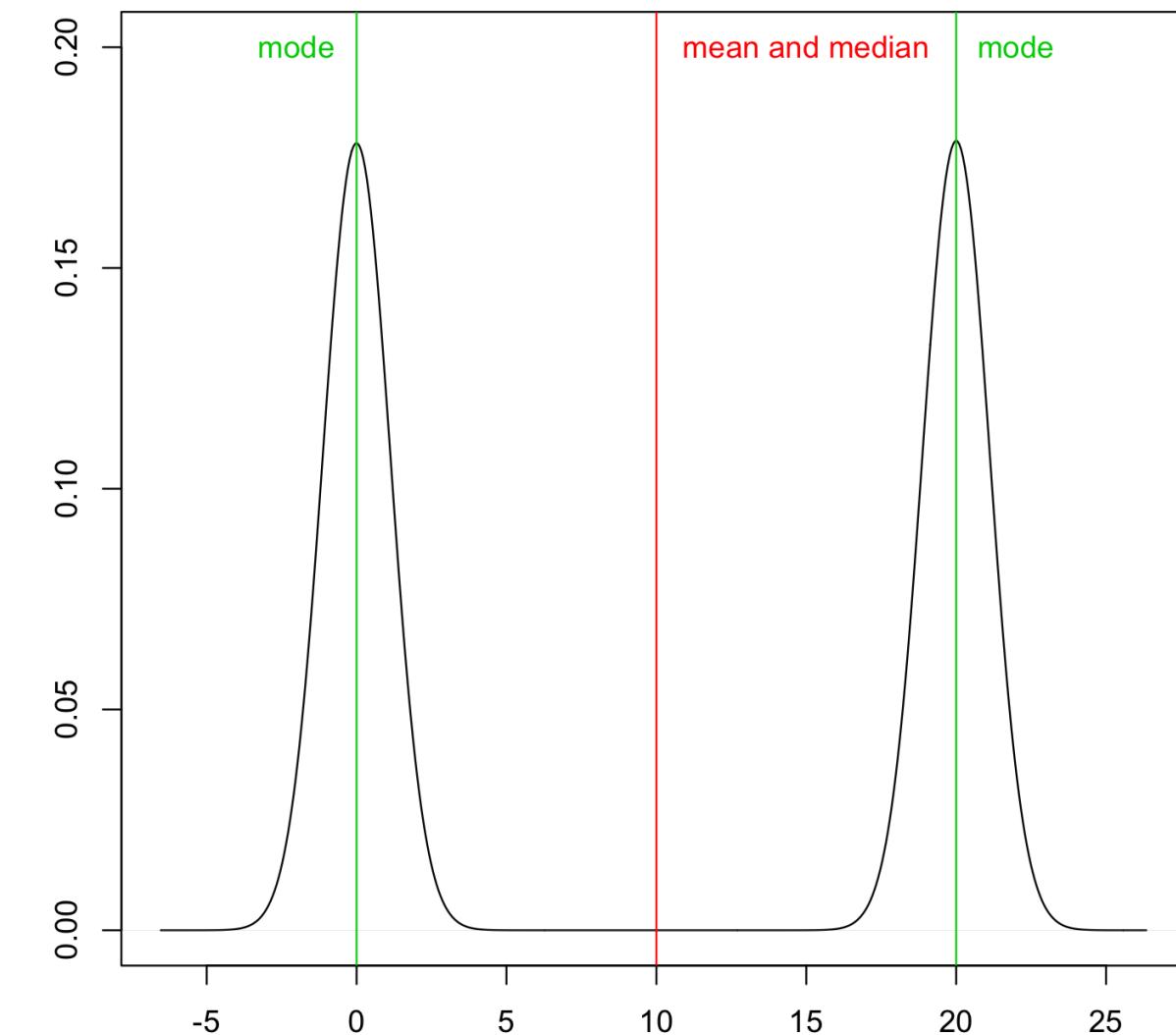
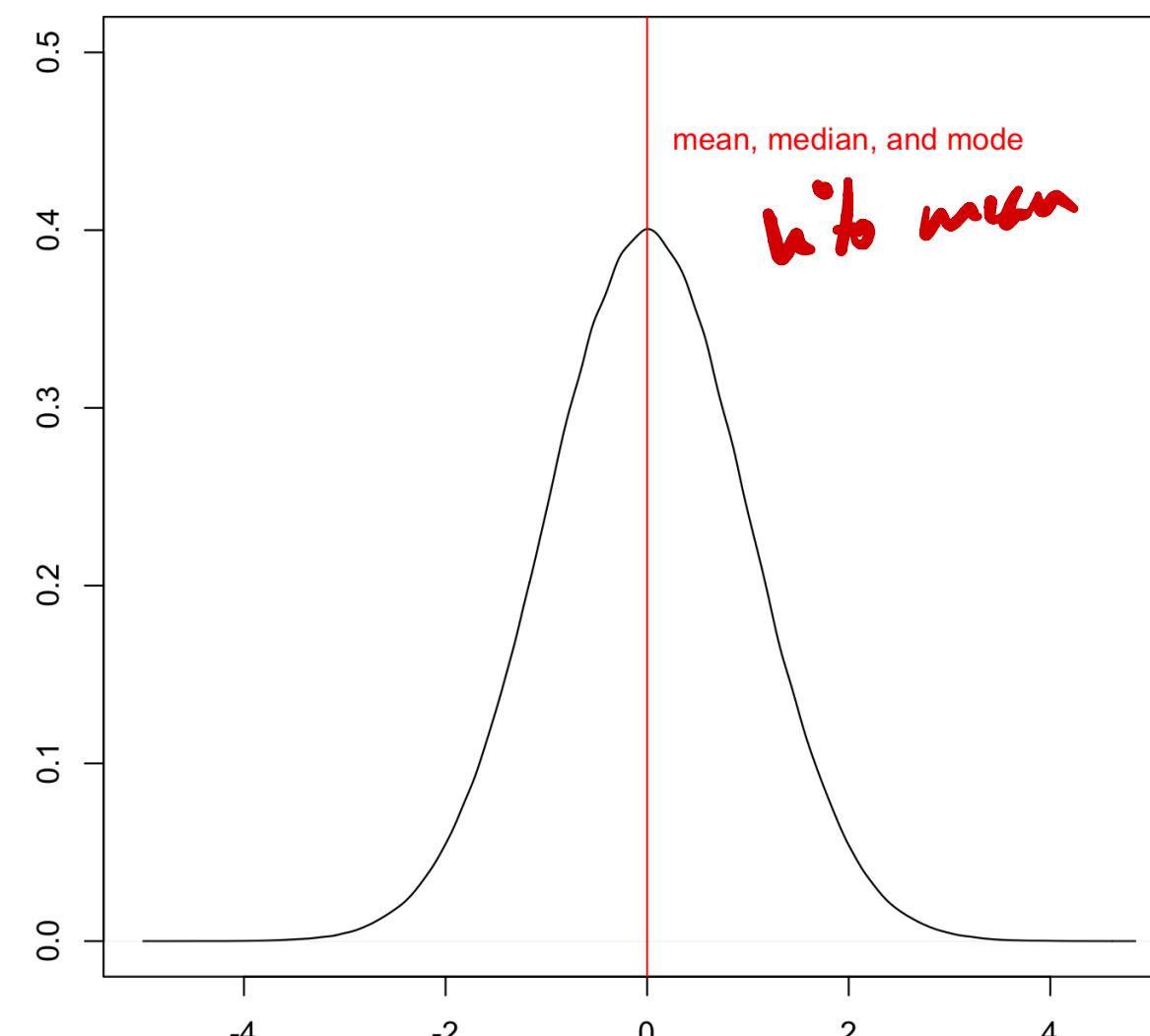
# Measures of Center

- The “best” measure of central tendency (mean, trimmed mean, median, mode) generally depends on the distribution of values
- If a distribution is symmetric and unimodal, then the mean, median, and mode should all approximately be the same
- **Unimodal:** A histogram or density plot only has one peak
- **Bimodal:** A histogram or density plot has two peaks

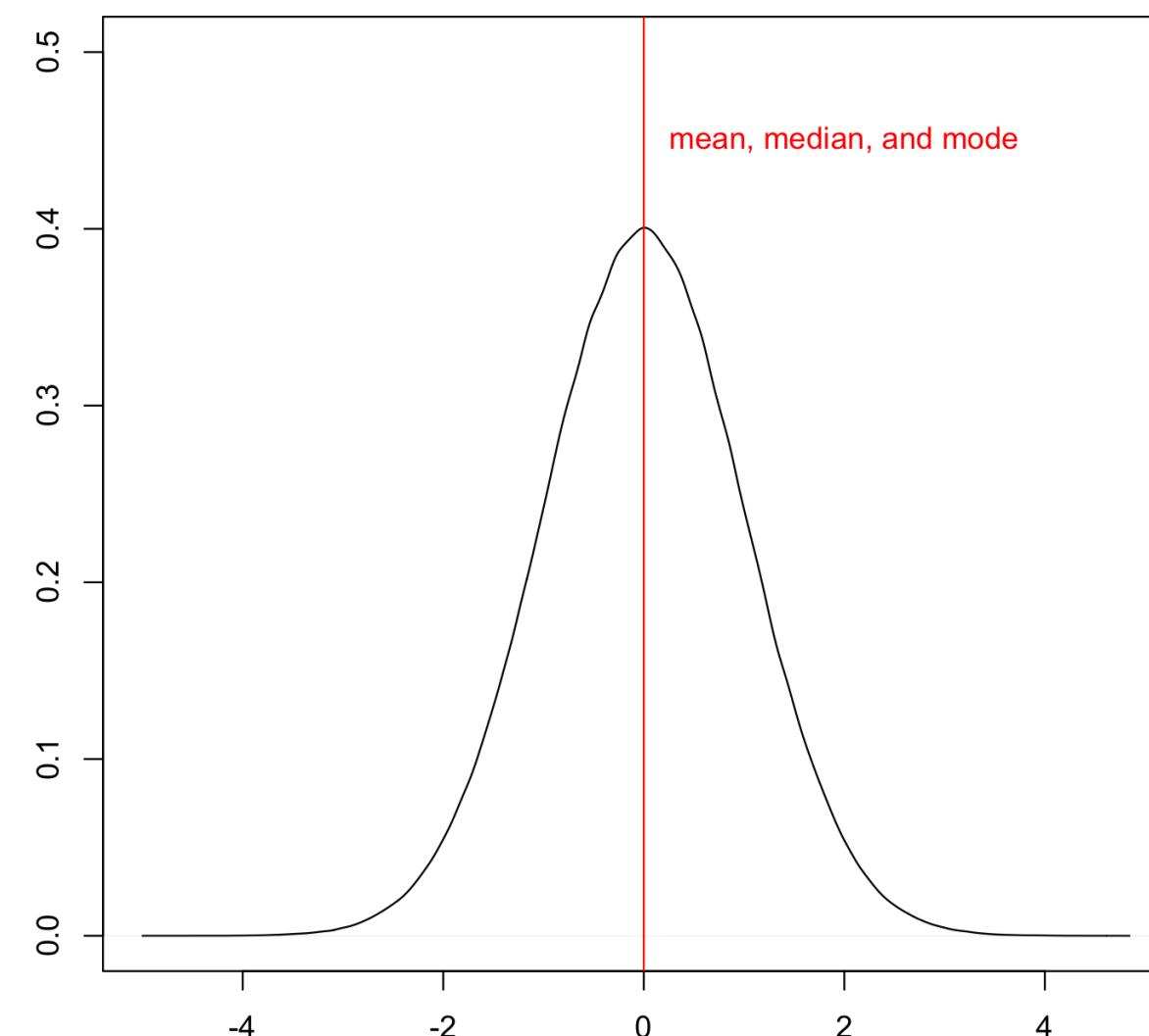
# Measures of Center

- The “best” measure of central tendency (mean, trimmed mean, median, mode) generally depends on the distribution of values
- If a distribution is symmetric and unimodal, then the mean, median, and mode should all approximately be the same
- **Unimodal:** A histogram or density plot only has one peak
- **Bimodal:** A histogram or density plot has two peaks
- **Multimodal:** A histogram or density plot has multiple peaks

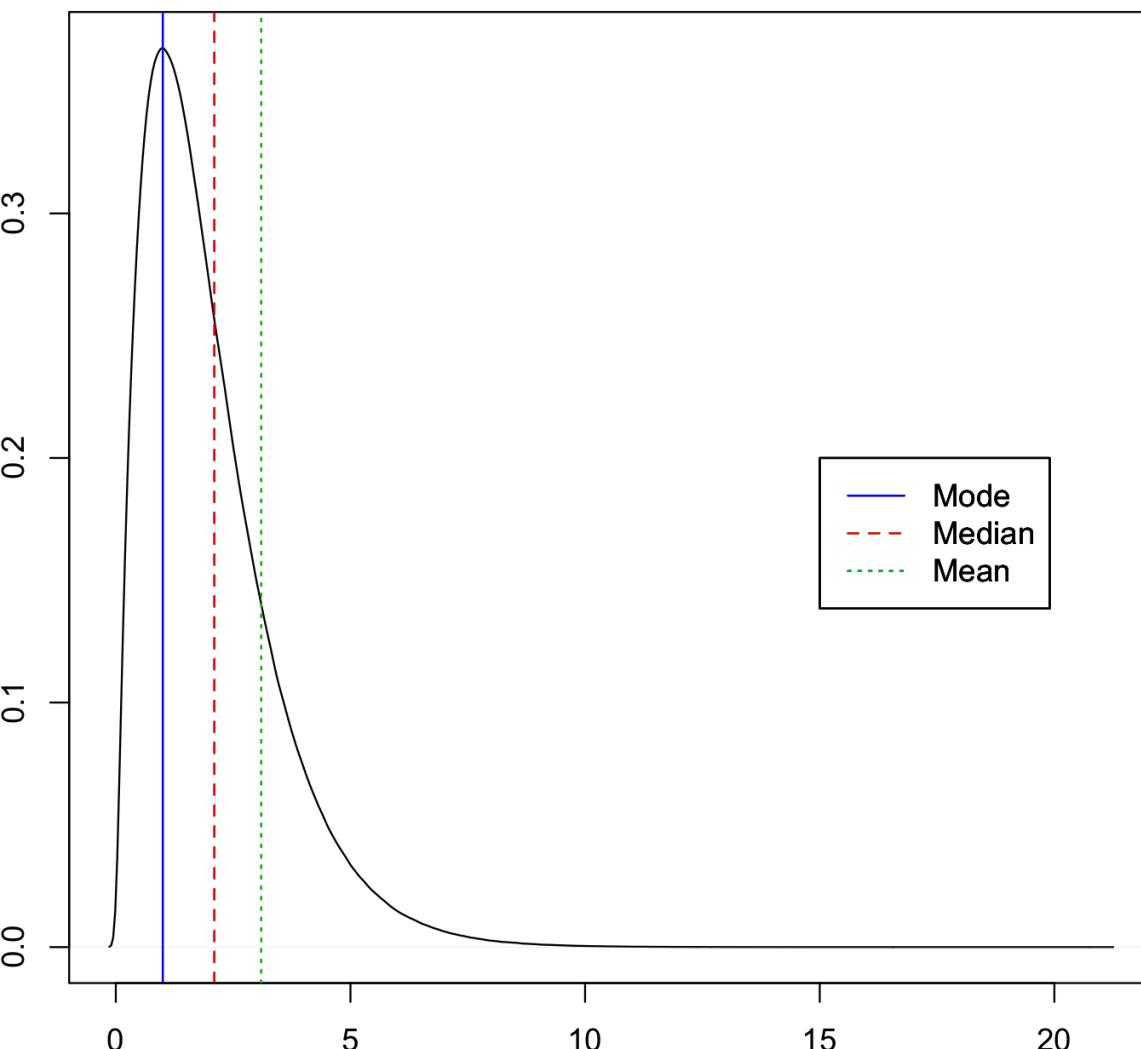
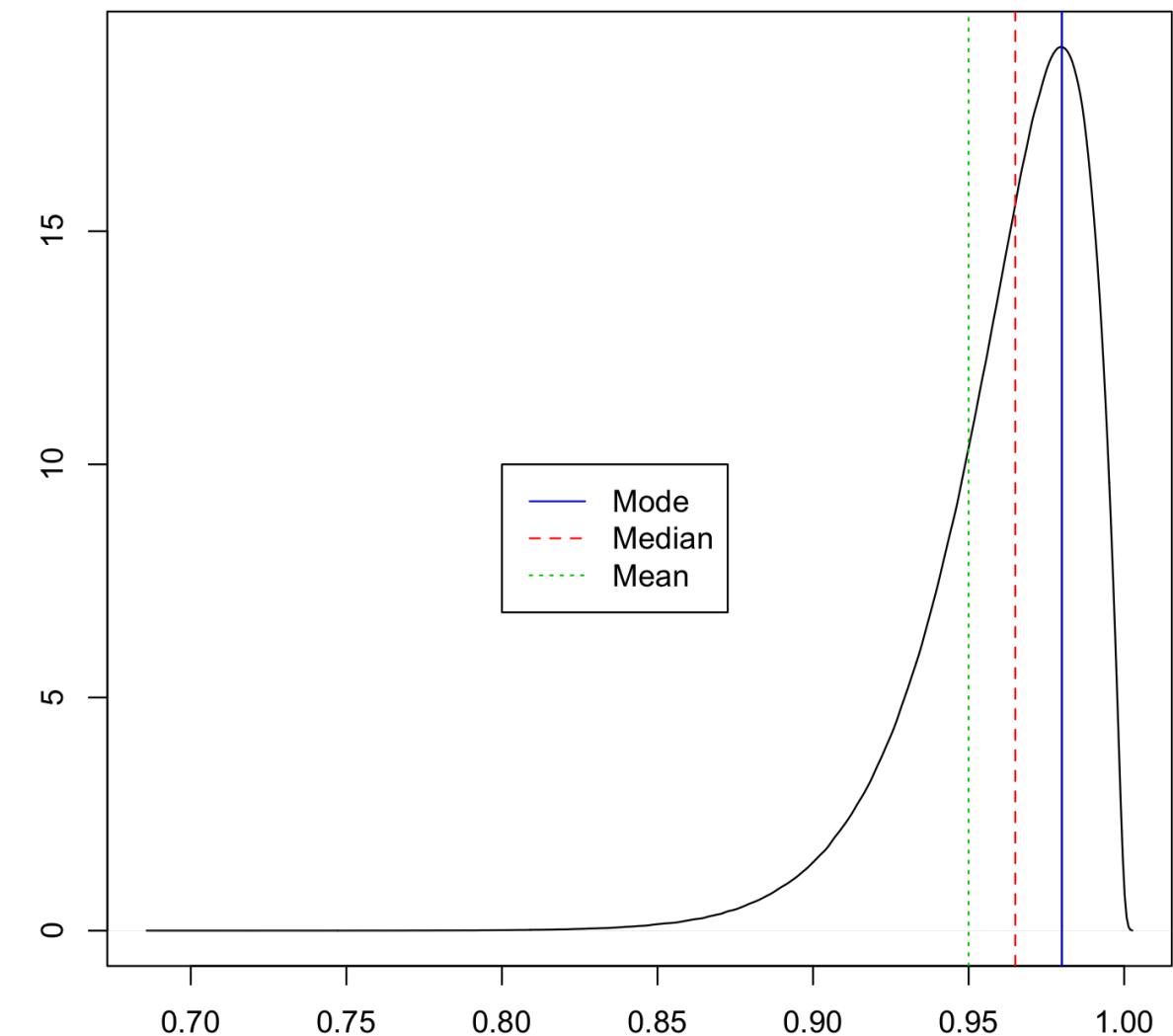
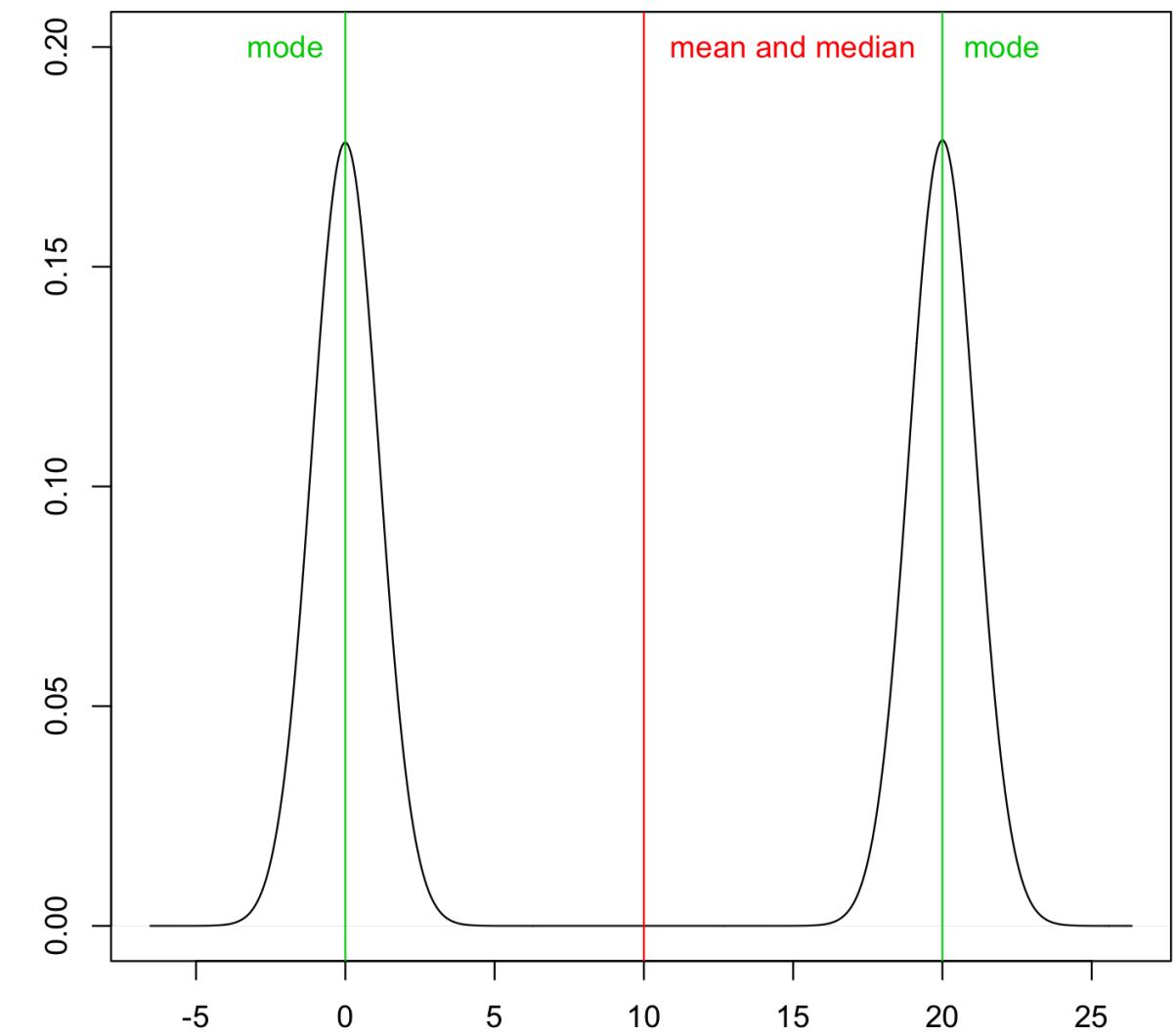
# Measures of Center: Comparison



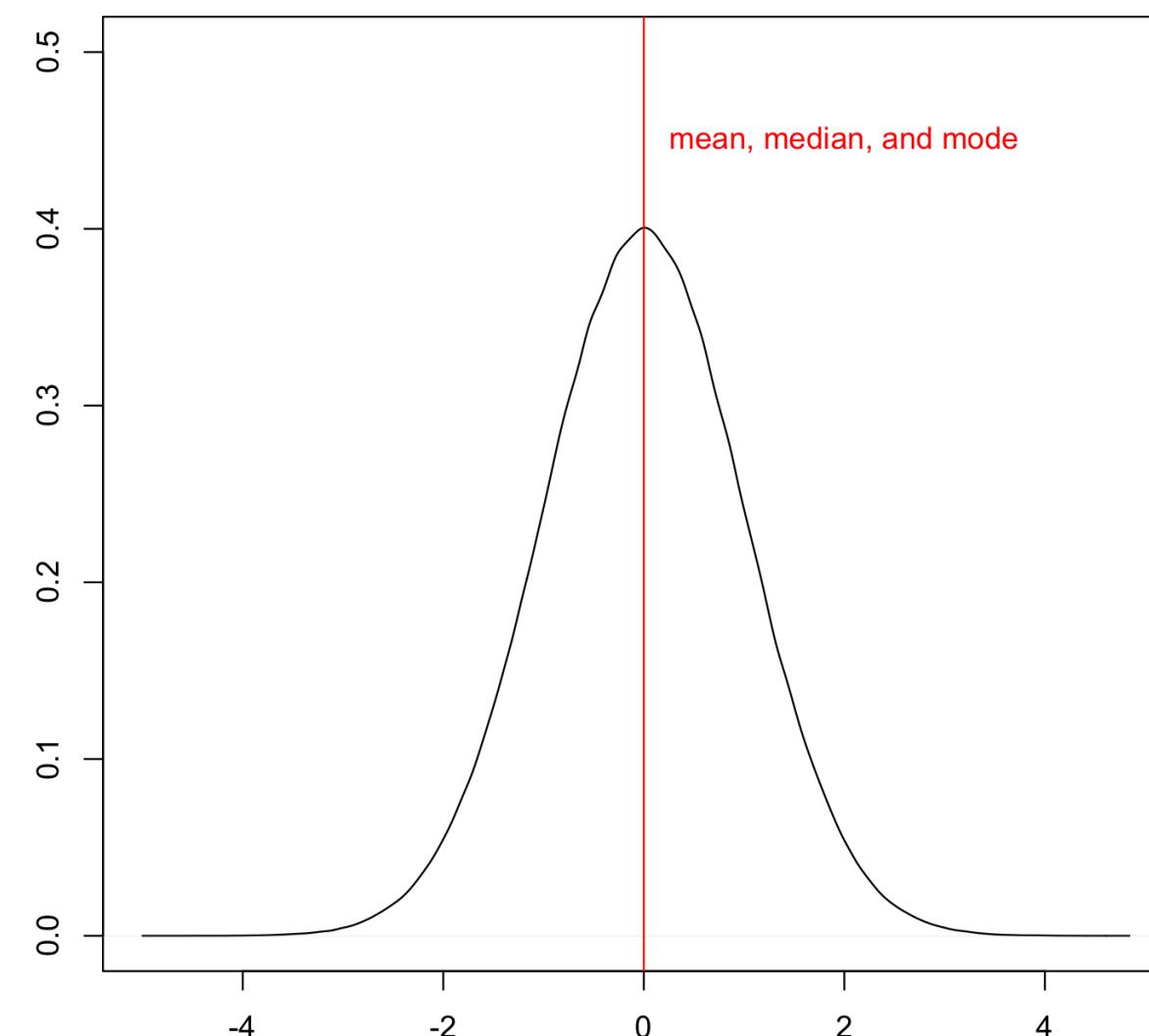
# Measures of Center: Comparison



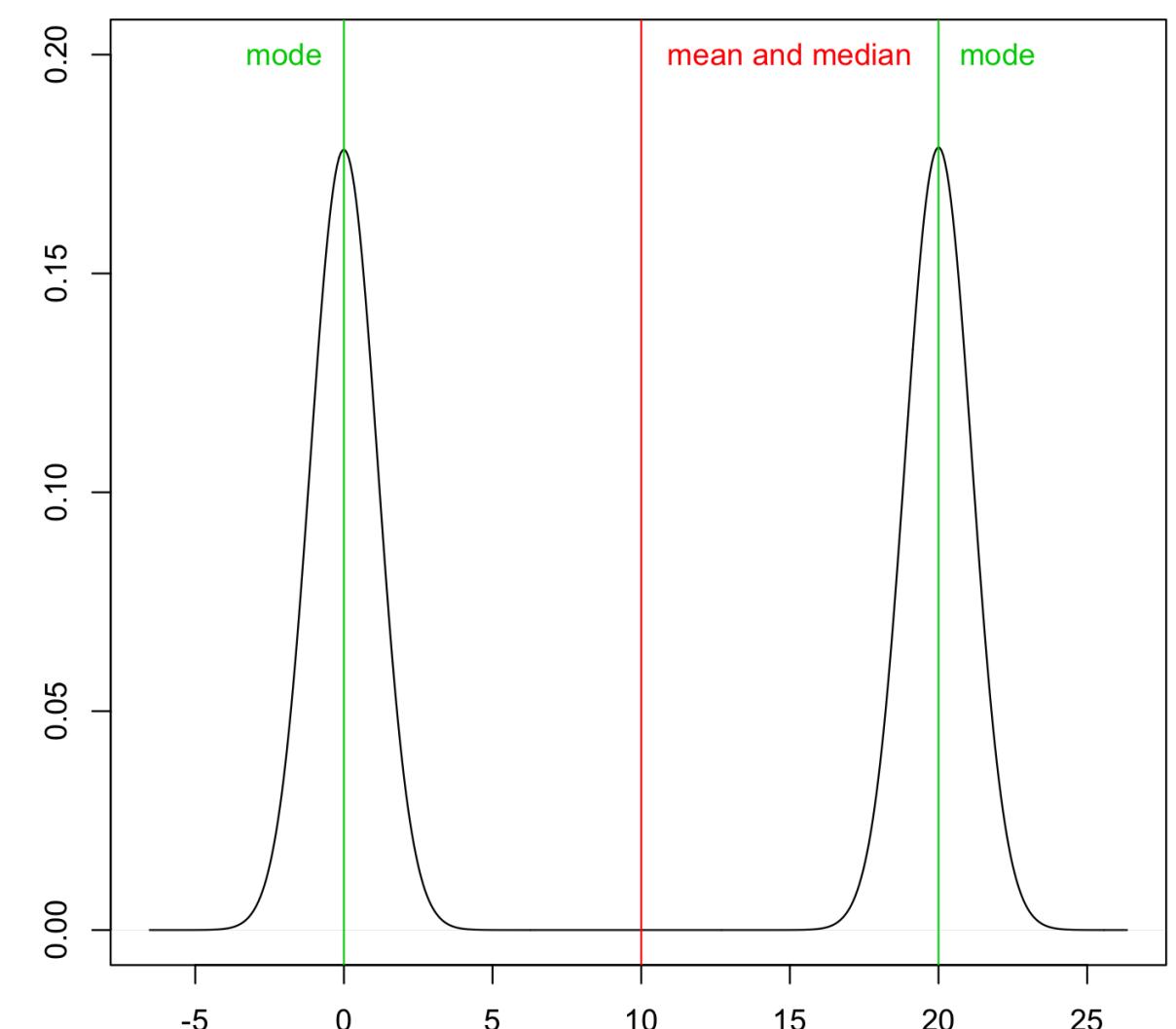
Symmetric  
Unimodal



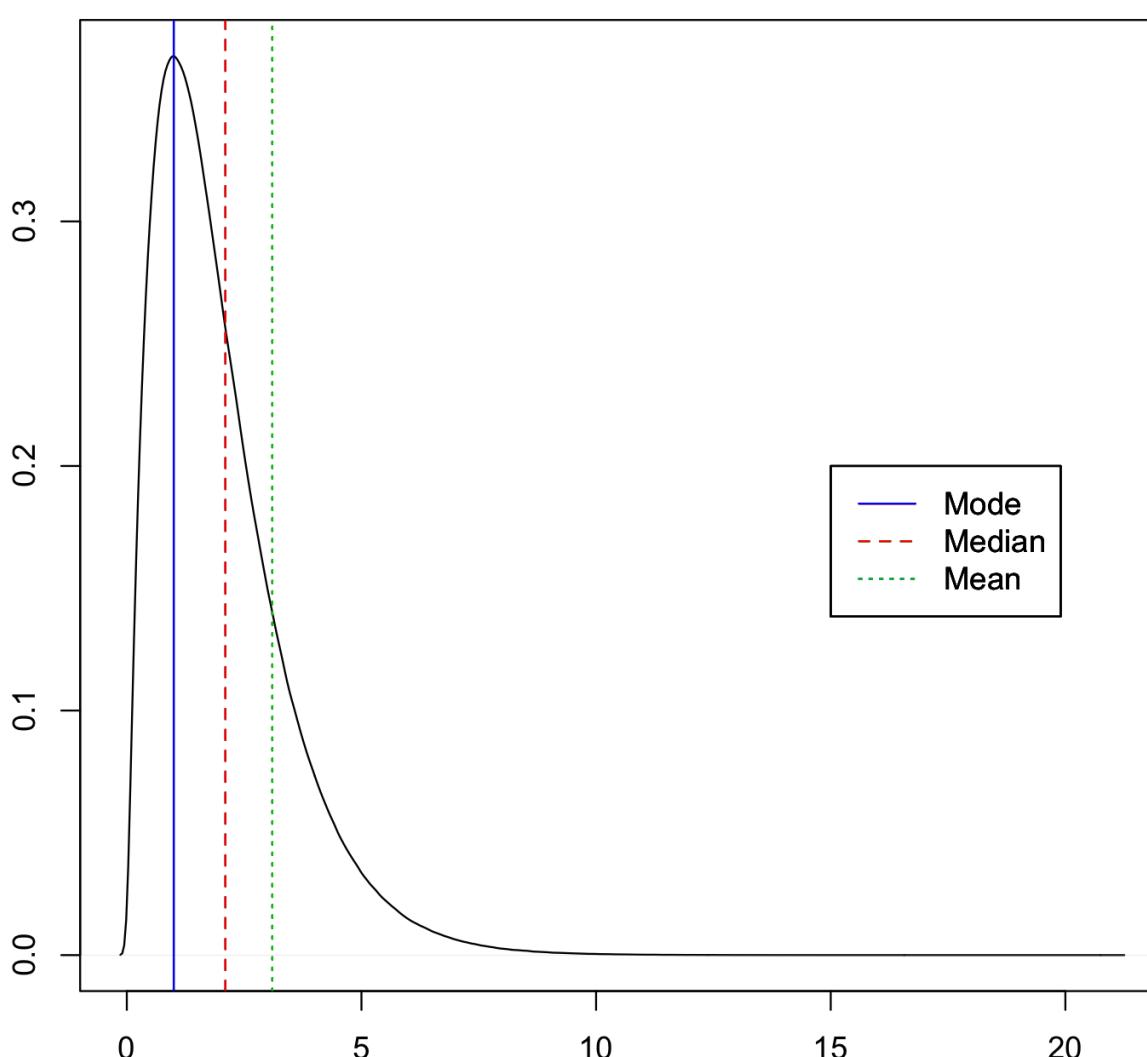
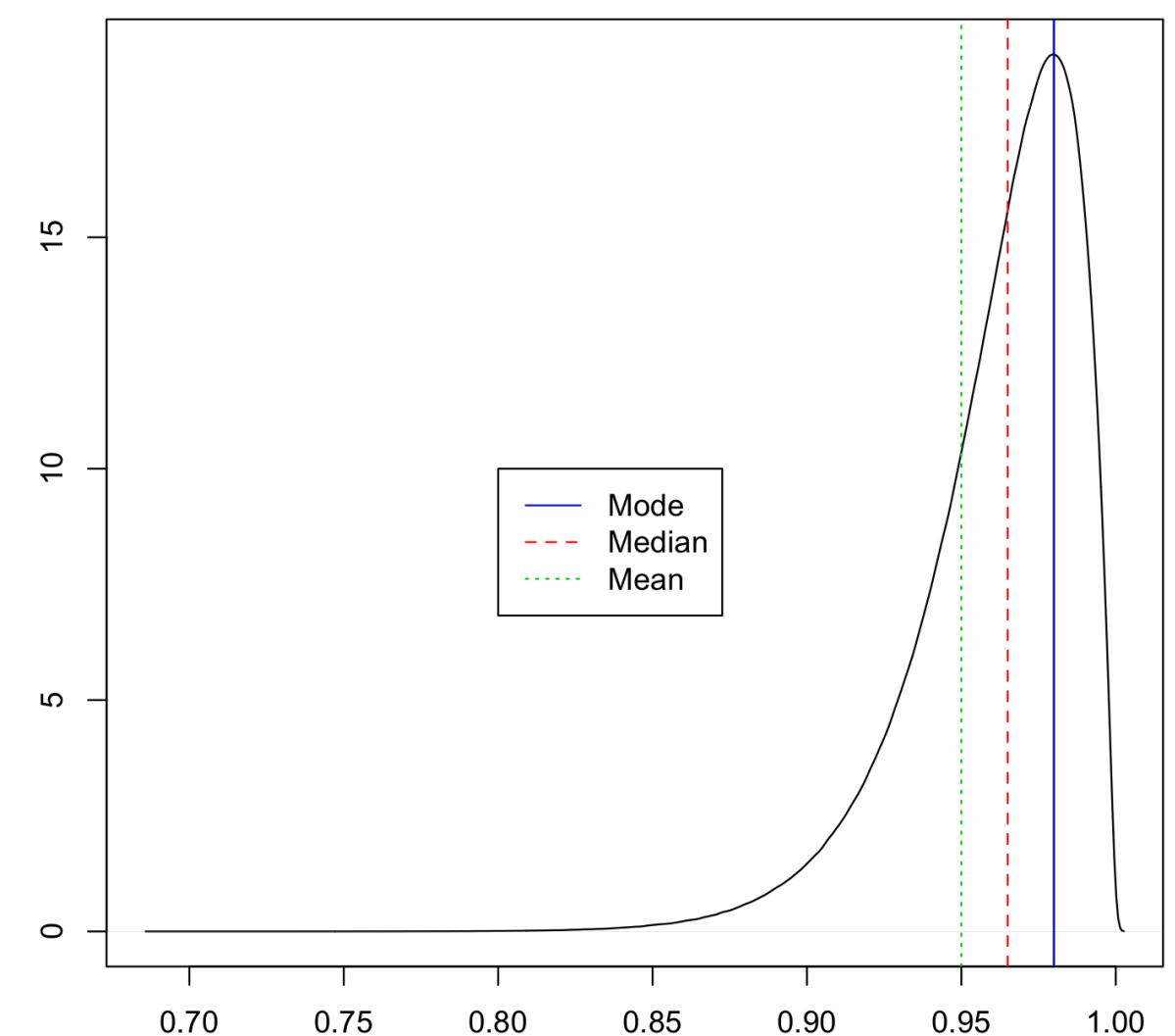
# Measures of Center: Comparison



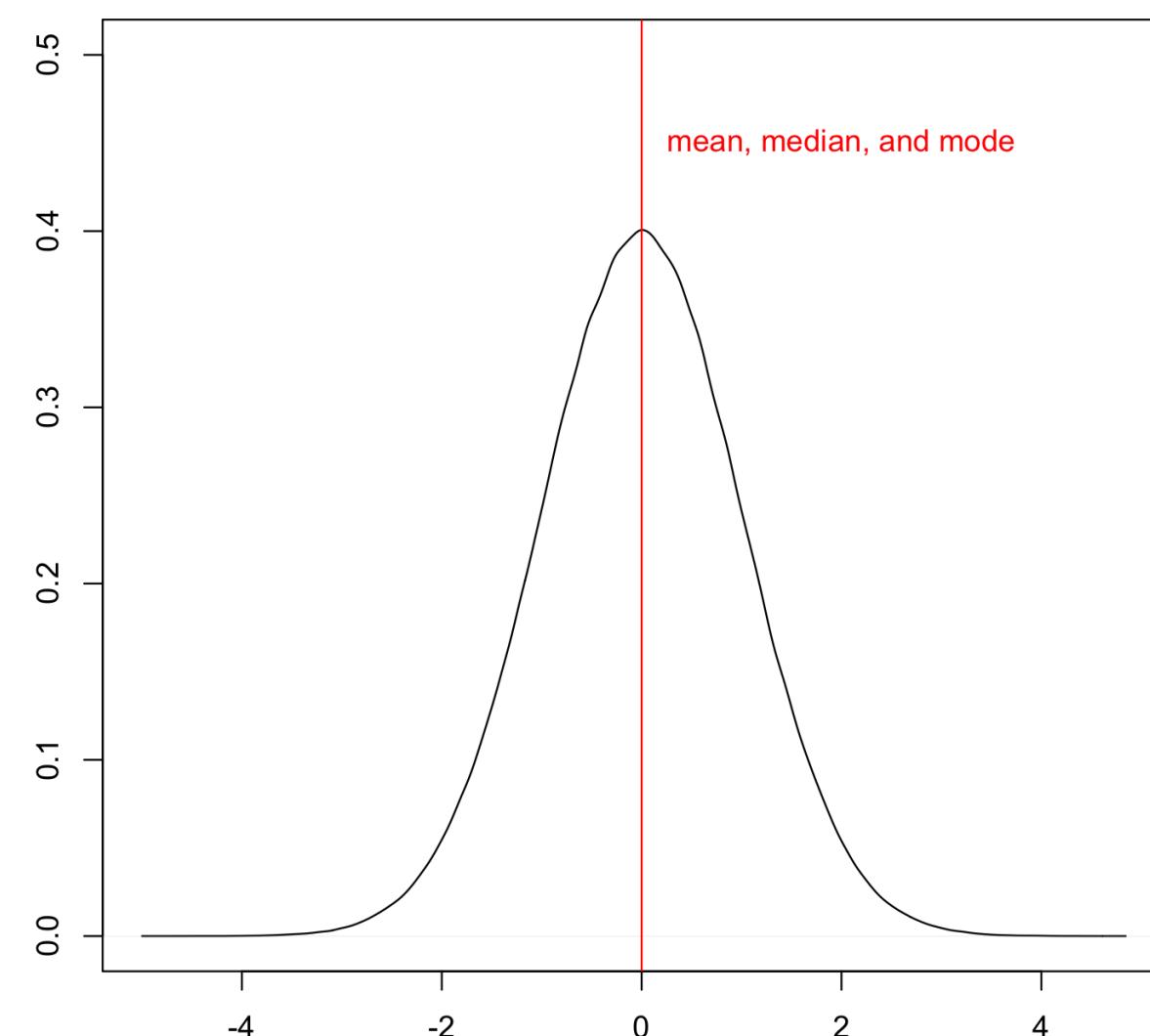
Symmetric  
Unimodal



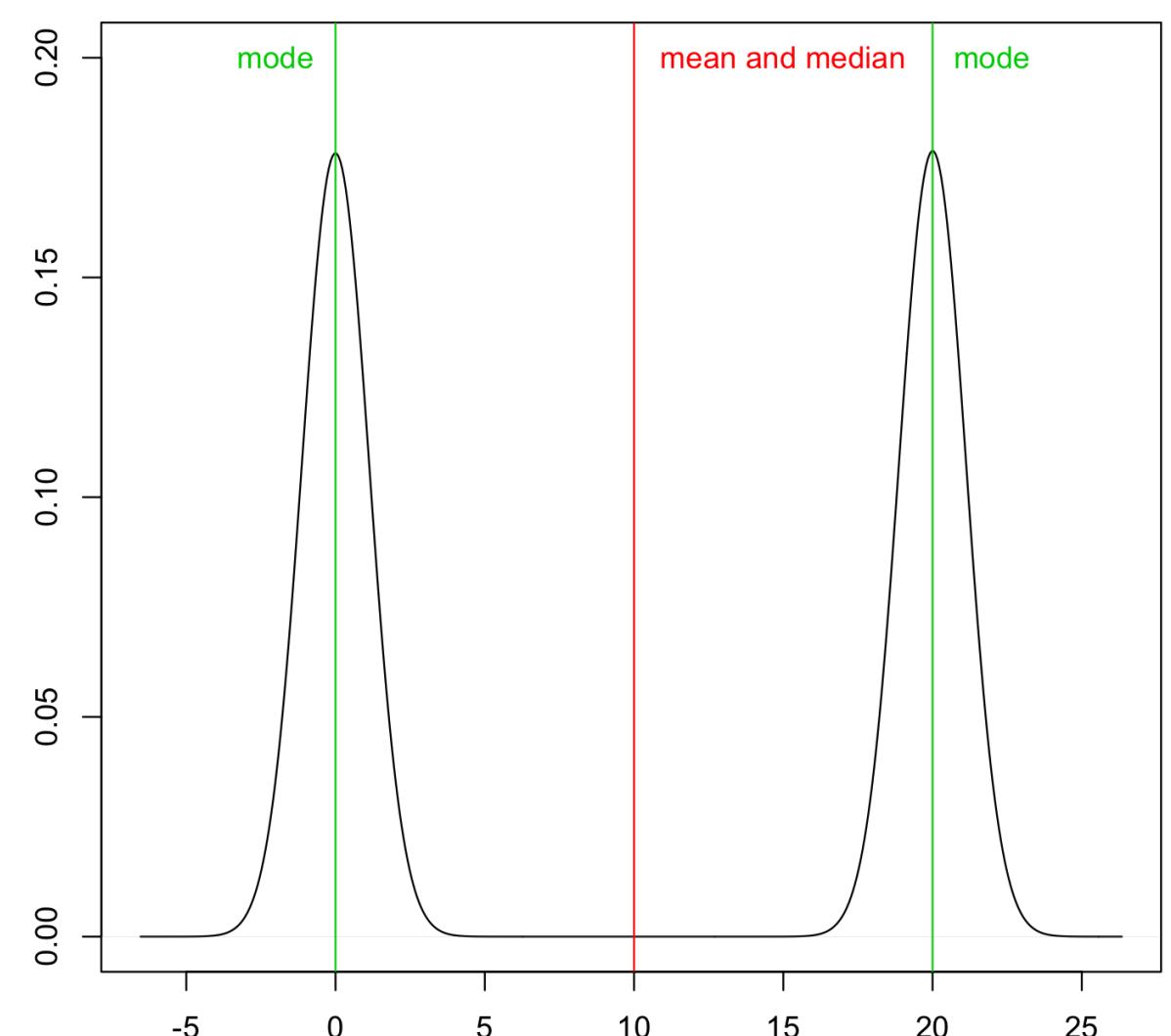
Symmetric  
Bimodal



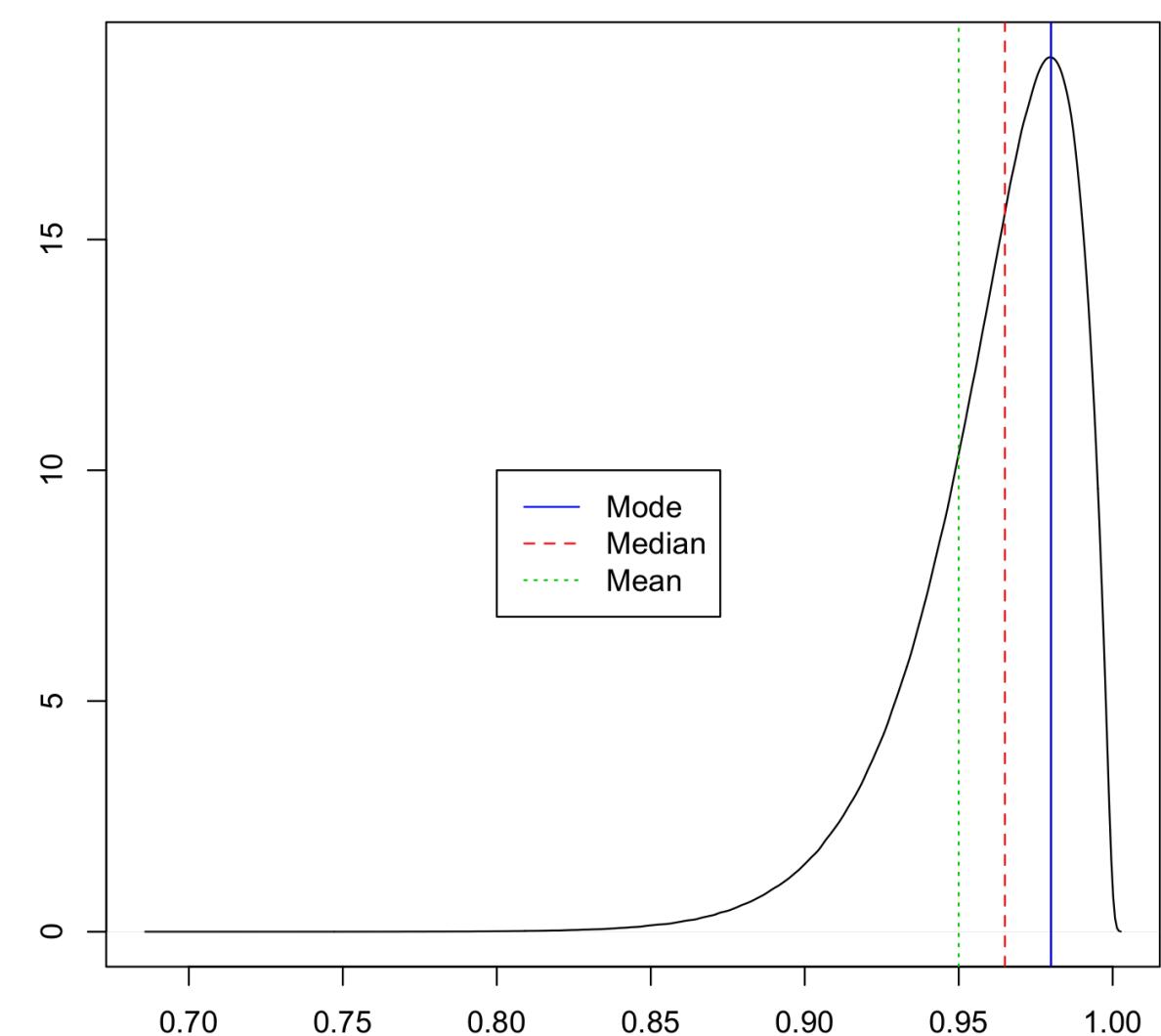
# Measures of Center: Comparison



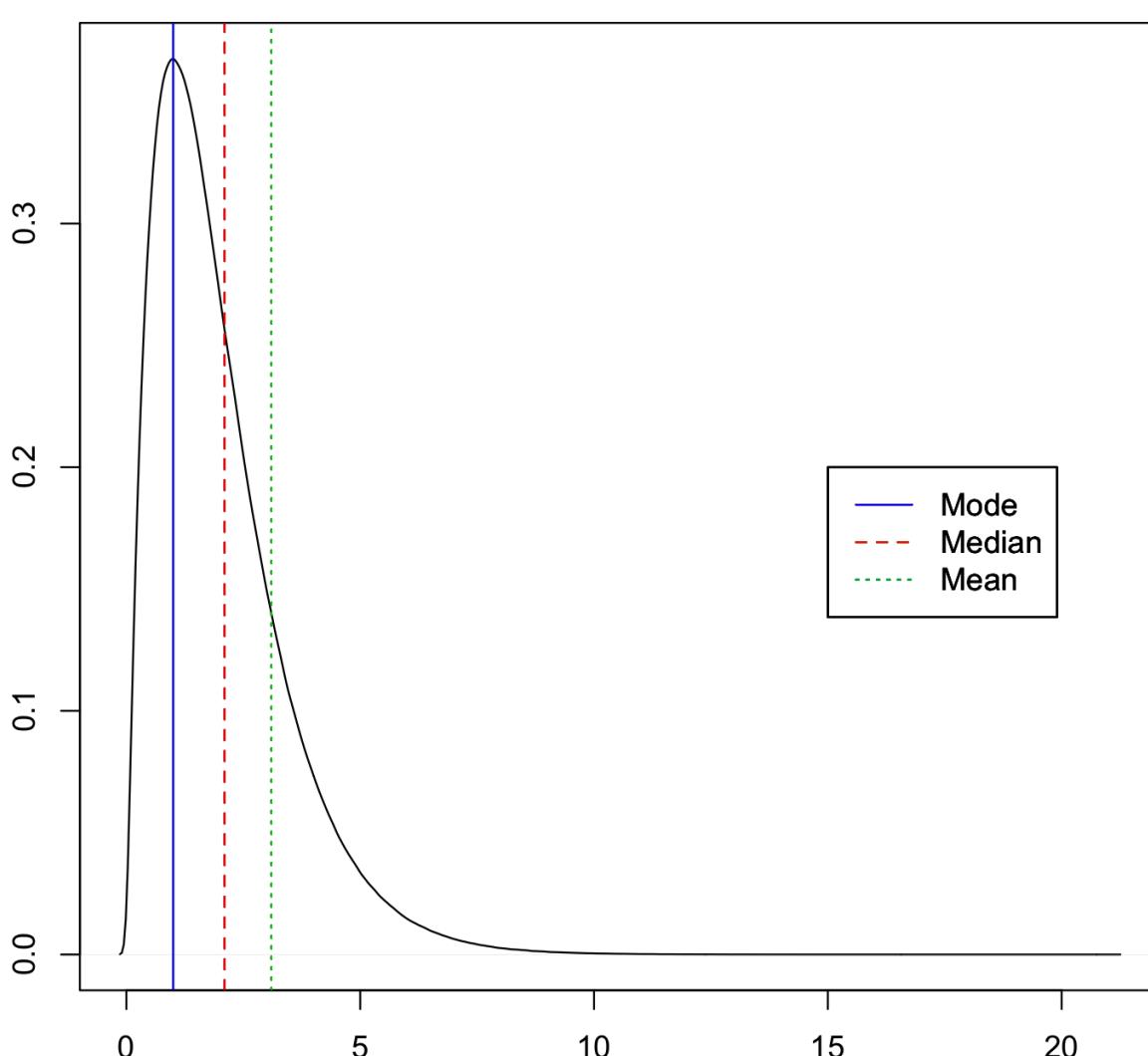
Symmetric  
Unimodal



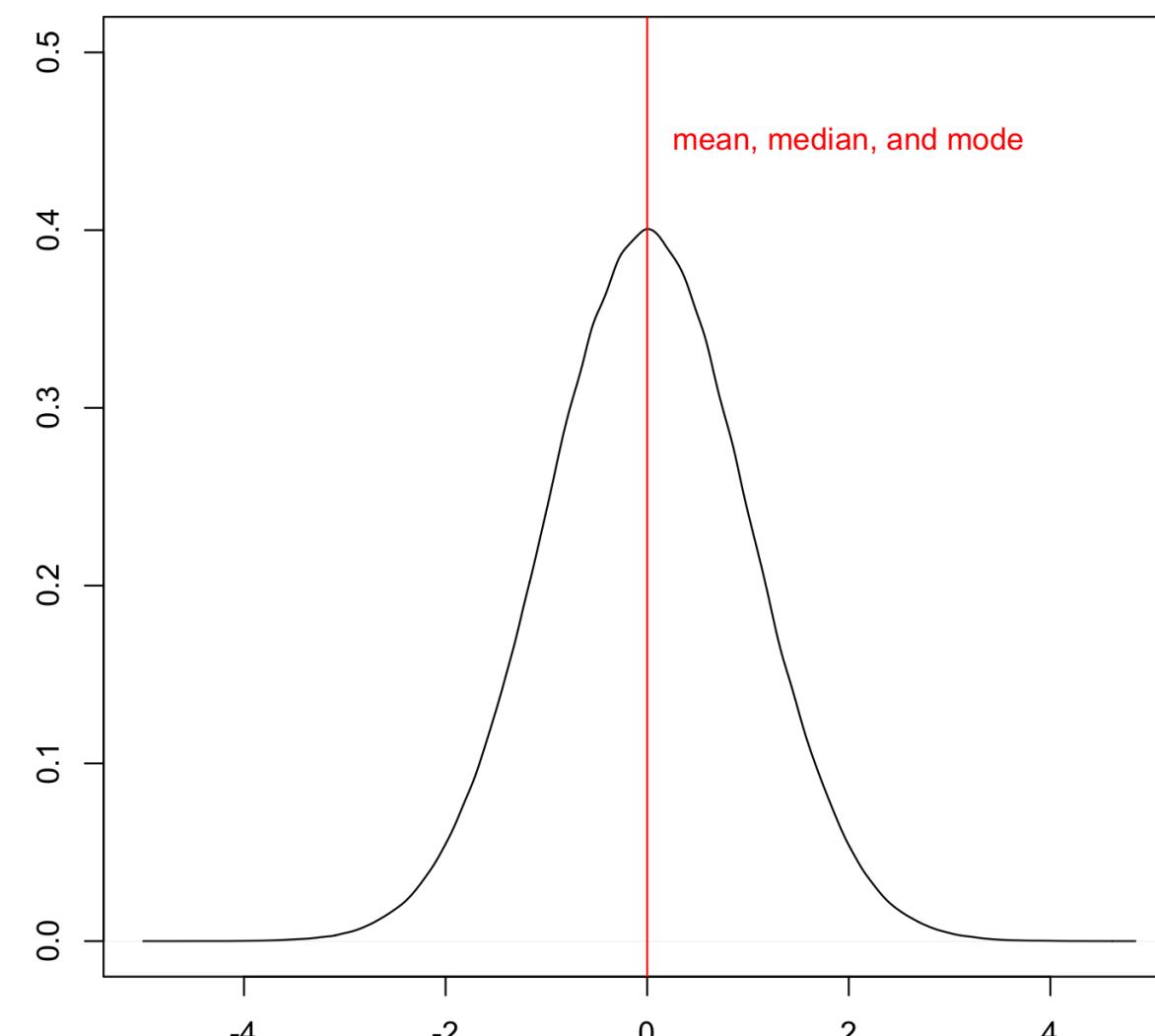
Symmetric  
Bimodal



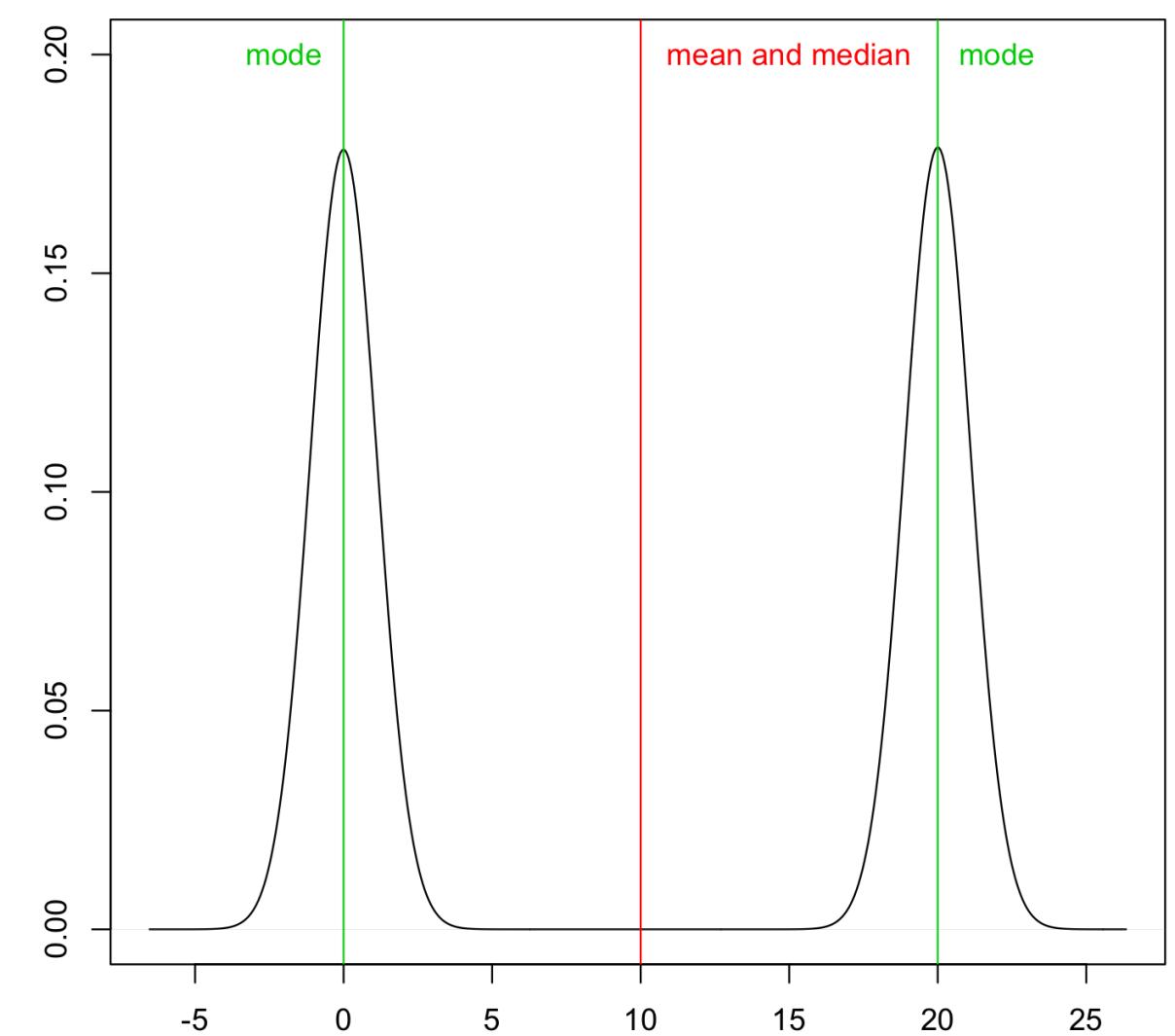
Left-skewed  
Unimodal



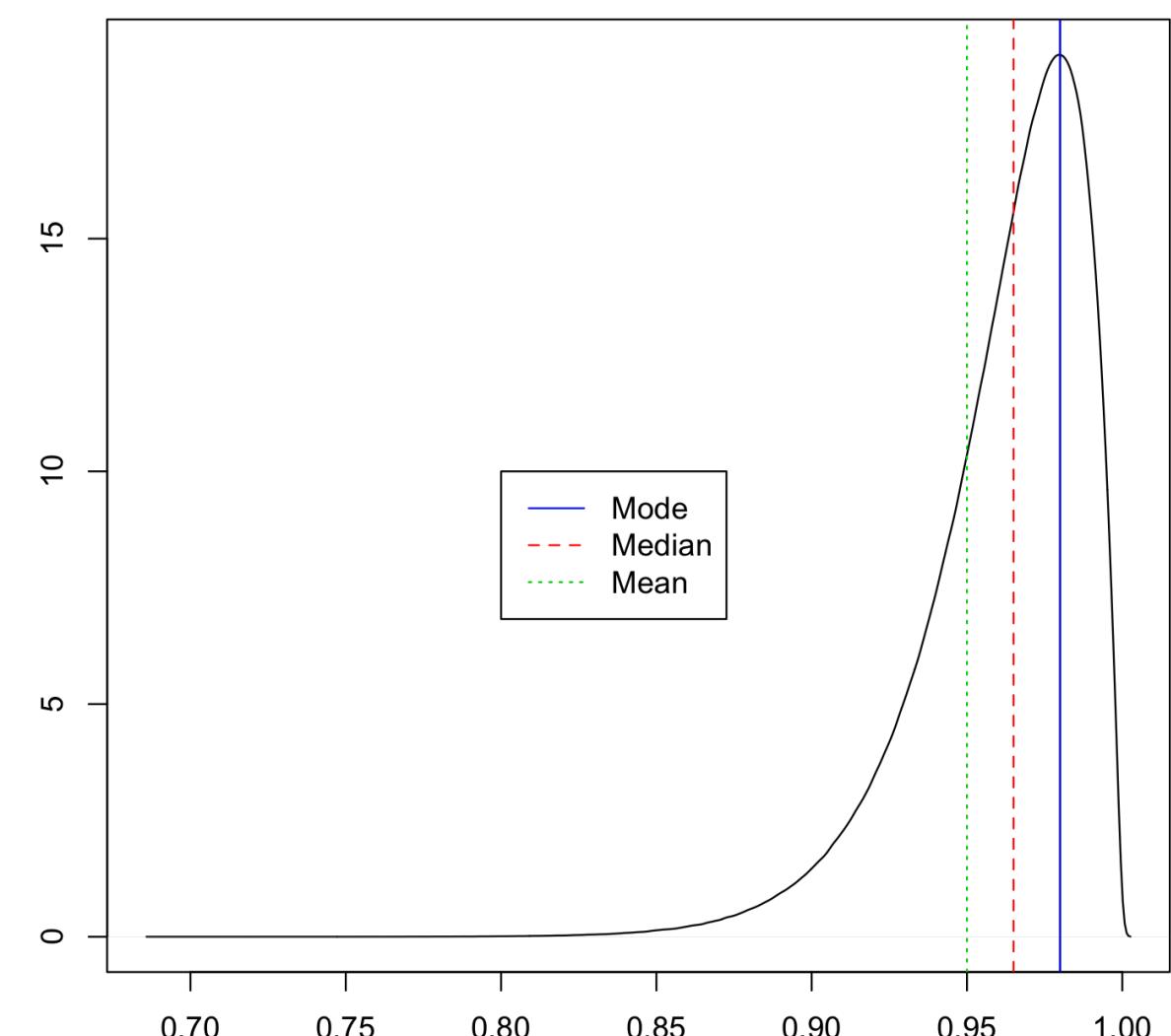
# Measures of Center: Comparison



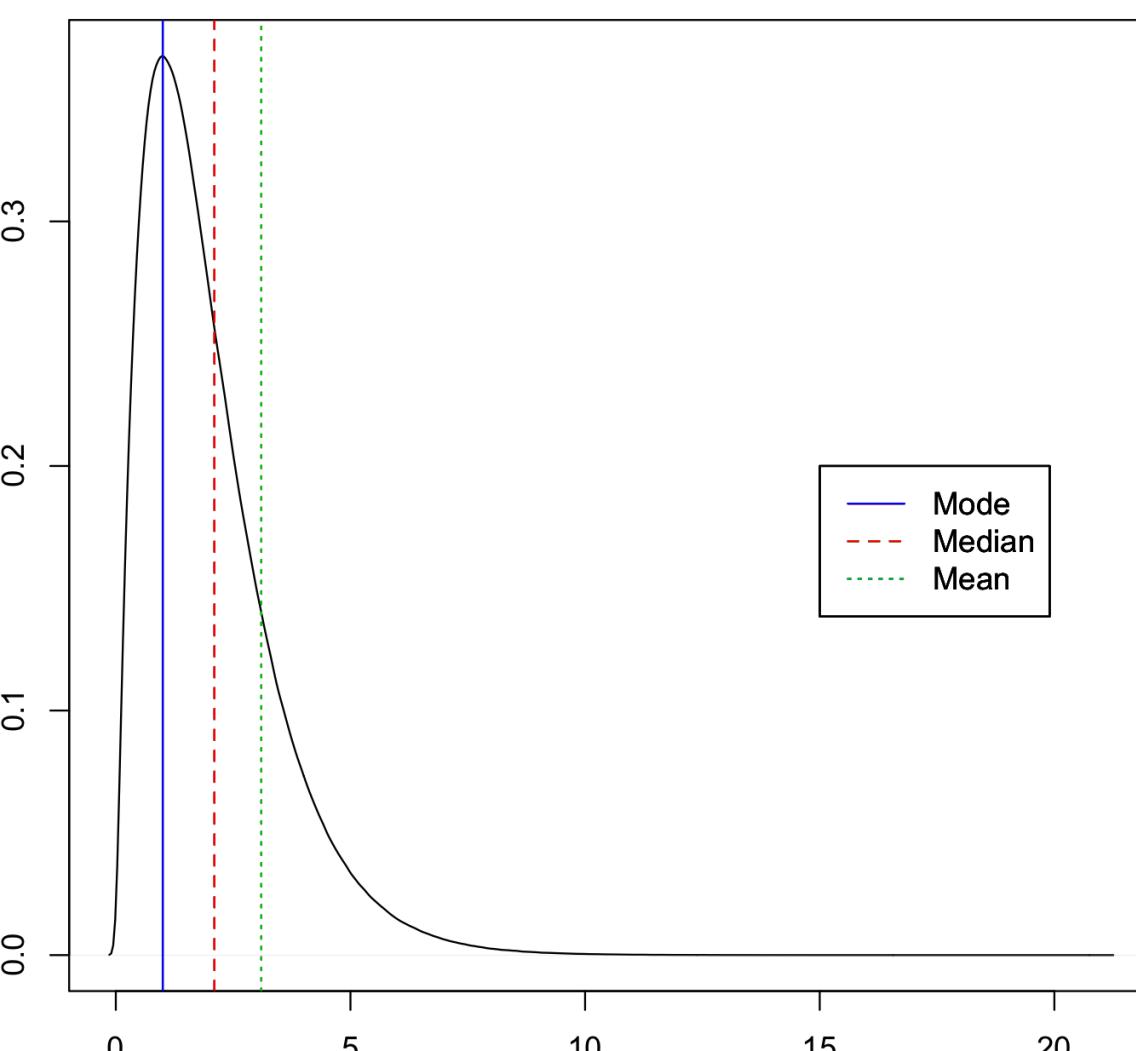
Symmetric  
Unimodal



Symmetric  
Bimodal

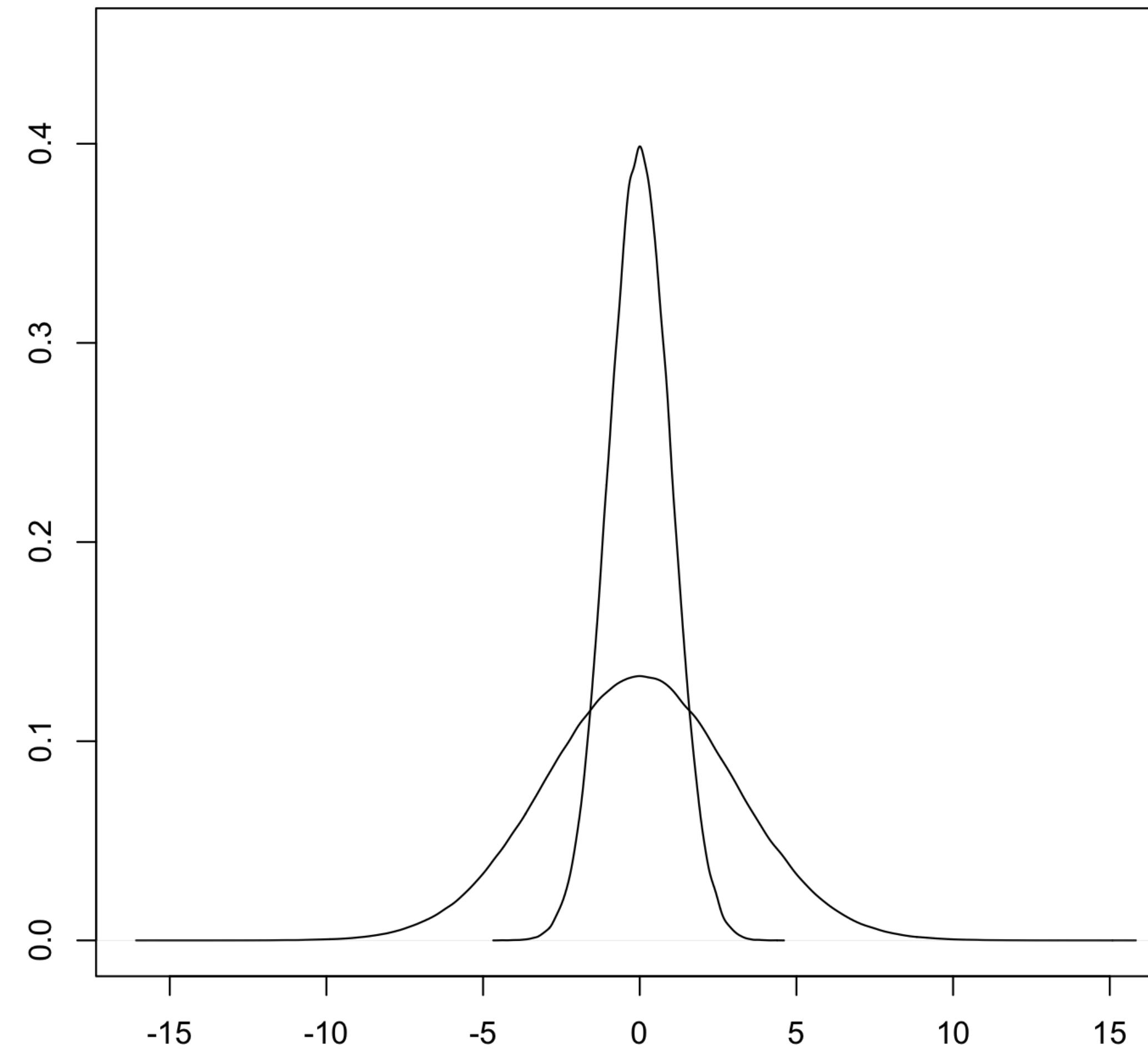


Left-skewed  
Unimodal



Right-skewed  
Unimodal

# Measures of Center: Not the Whole Picture



# Measures of Dispersion

# Measures of Dispersion

- Measures of dispersion give us information regarding the *spread* of the data (i.e., how variable it is)

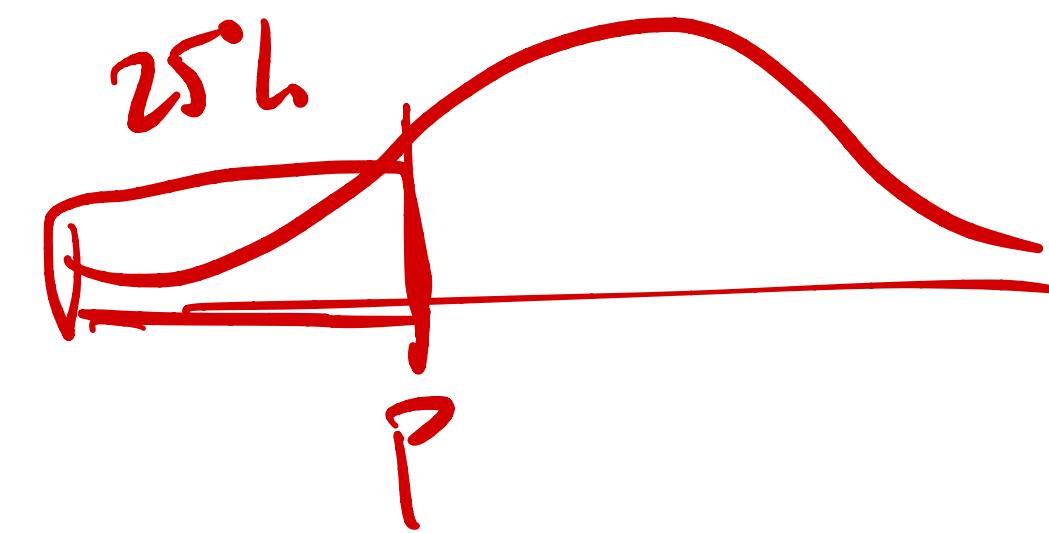
# Measures of Dispersion

- Measures of dispersion give us information regarding the *spread* of the data (i.e., how variable it is)
- While knowing where the center of a distribution is may be important, it is also essential to know how disperse the data is around that center to better understand how the variable acts

# Quantiles (and Percentiles and Quartiles)

# Quantiles (and Percentiles and Quartiles)

- The  $p$  sample quantile is the value below which a proportion  $p$  of the data are located (and above which a proportion  $1 - p$  of the data are located)



# Quantiles (and Percentiles and Quartiles)

- The  $p$  sample quantile is the value below which a proportion  $p$  of the data are located (and above which a proportion  $1 - p$  of the data are located)
- The  $k^{th}$  percentile is the  $(k/100)$  sample quantile

# Quantiles (and Percentiles and Quartiles)

- The  $p$  sample quantile is the value below which a proportion  $p$  of the data are located (and above which a proportion  $1 - p$  of the data are located)
- The  $k^{\text{th}}$  percentile is the  $(k/100)$  sample quantile
  - E.g., if your birth weight is at the 95<sup>th</sup> percentile, then you weighed more than 0.95 of all newborn babies

# Quantiles (and Percentiles and Quartiles)

- The  $p$  sample quantile is the value below which a proportion  $p$  of the data are located (and above which a proportion  $1 - p$  of the data are located)
- The  $k^{\text{th}}$  percentile is the  $(k/100)$  sample quantile
  - E.g., if your birth weight is at the 95<sup>th</sup> percentile, then you weighed more than 0.95 of all newborn babies
- Quartiles correspond to 0.25, 0.50, and 0.75 sample quantiles (split the data into four equal parts)

# Quantiles (and Percentiles and Quartiles)

- The  $p$  sample quantile is the value below which a proportion  $p$  of the data are located (and above which a proportion  $1 - p$  of the data are located)
- The  $k^{\text{th}}$  percentile is the  $(k/100)$  sample quantile
  - E.g., if your birth weight is at the 95<sup>th</sup> percentile, then you weighed more than 0.95 of all newborn babies
- Quartiles correspond to 0.25, 0.50, and 0.75 sample quantiles (split the data into four equal parts)
- Sample quantiles are constructed based on the order statistics

# Quantiles (and Percentiles and Quartiles)

- The  $p$  sample quantile is the value below which a proportion  $p$  of the data are located (and above which a proportion  $1 - p$  of the data are located)
- The  $k^{\text{th}}$  percentile is the  $(k/100)$  sample quantile
  - E.g., if your birth weight is at the 95<sup>th</sup> percentile, then you weighed more than 0.95 of all newborn babies
- Quartiles correspond to 0.25, 0.50, and 0.75 sample quantiles (split the data into four equal parts)
- Sample quantiles are constructed based on the order statistics
- In R: `quantile()`

# Quantiles (and Percentiles and Quartiles)

- The  $p$  sample quantile is the value below which a proportion  $p$  of the data are located (and above which a proportion  $1 - p$  of the data are located)
- The  $k^{\text{th}}$  percentile is the  $(k/100)$  sample quantile
  - E.g., if your birth weight is at the 95<sup>th</sup> percentile, then you weighed more than 0.95 of all newborn babies
- Quartiles correspond to 0.25, 0.50, and 0.75 sample quantiles (split the data into four equal parts)
- Sample quantiles are constructed based on the order statistics
- In R: `quantile()`
  - 9 methods for calculating sample quantiles (types 1-3 are for discrete distributions; types 4-9 are for continuous distributions).

$$Q1 = 25^{\text{th}} \text{ percentile} = 0.25 \text{ quantile}$$

# Quantiles (and Percentiles and Quartiles)

- The  $p$  sample quantile is the value below which a proportion  $p$  of the data are located (and above which a proportion  $1 - p$  of the data are located)
- The  $k^{\text{th}}$  percentile is the  $(k/100)$  sample quantile
  - E.g., if your birth weight is at the  $95^{\text{th}}$  percentile, then you weighed more than 0.95 of all newborn babies
- Quartiles correspond to 0.25, 0.50, and 0.75 sample quantiles (split the data into four equal parts)
- Sample quantiles are constructed based on the order statistics
- In R: `quantile()`
  - 9 methods for calculating sample quantiles (types 1-3 are for discrete distributions; types 4-9 are for continuous distributions).
  - Default: `type=7`

# Interquartile Range (IQR)

# Interquartile Range (IQR)

$$Q_3 - Q_1$$

- **Interquartile Range (IQR):** the difference between the 75<sup>th</sup> percentile and the 25<sup>th</sup> percentile

# Interquartile Range (IQR)

- **Interquartile Range (IQR)**: the difference between the 75<sup>th</sup> percentile and the 25<sup>th</sup> percentile
- $IQR = Q3 - Q1$

# Interquartile Range (IQR)

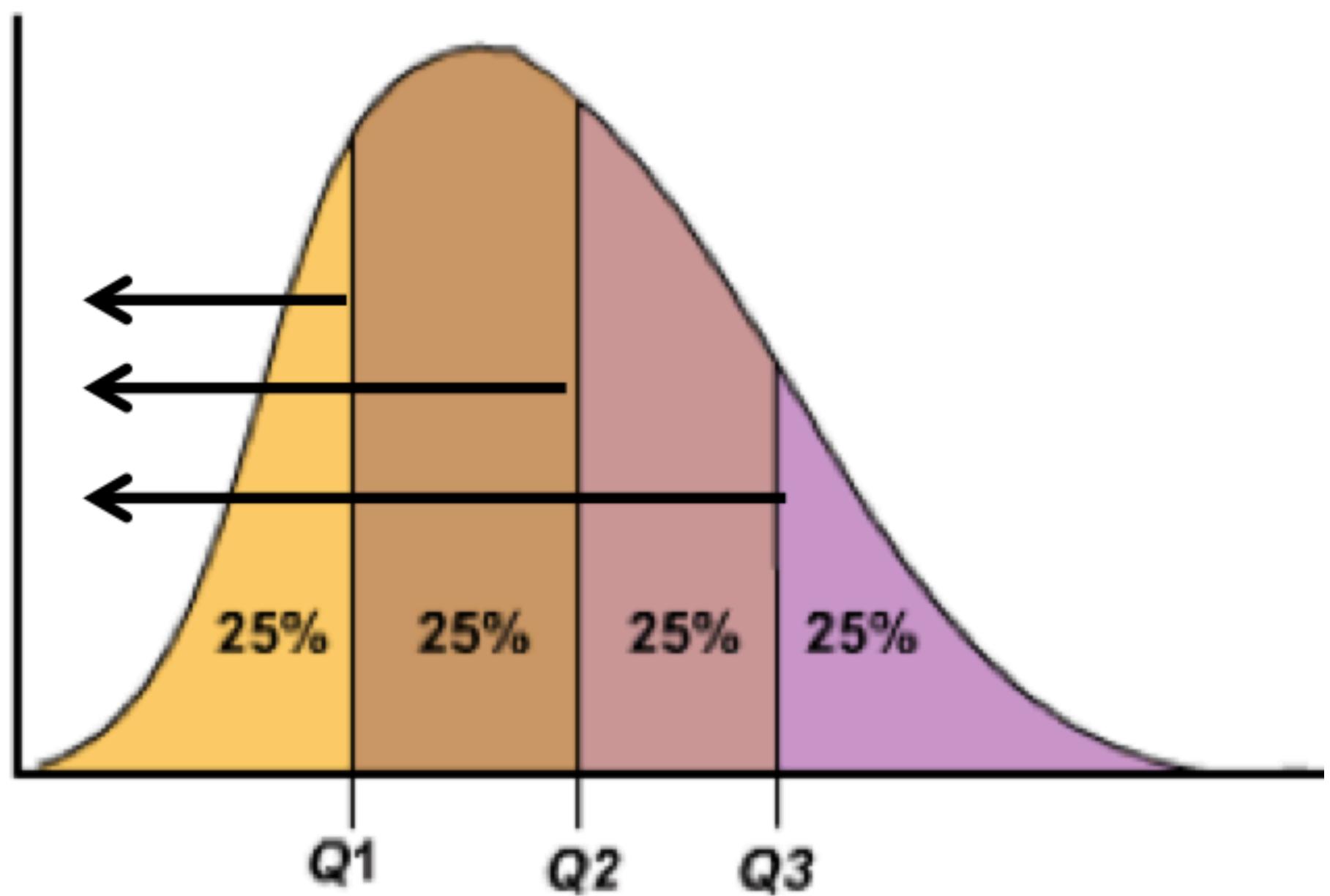
- **Interquartile Range (IQR)**: the difference between the 75<sup>th</sup> percentile and the 25<sup>th</sup> percentile
- $IQR = Q3 - Q1$
- Middle 50% of the observations in a given dataset

# Interquartile Range (IQR)

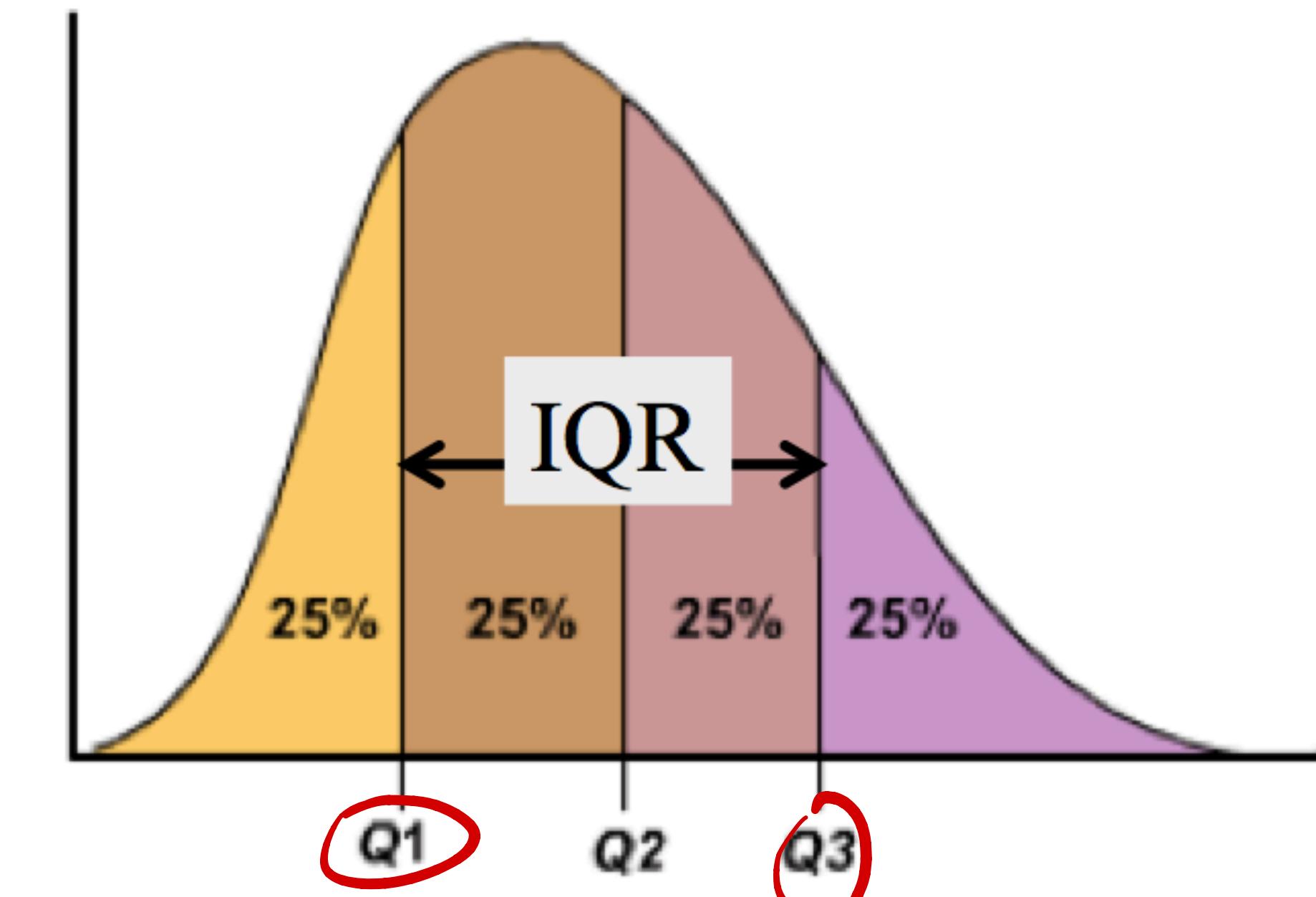
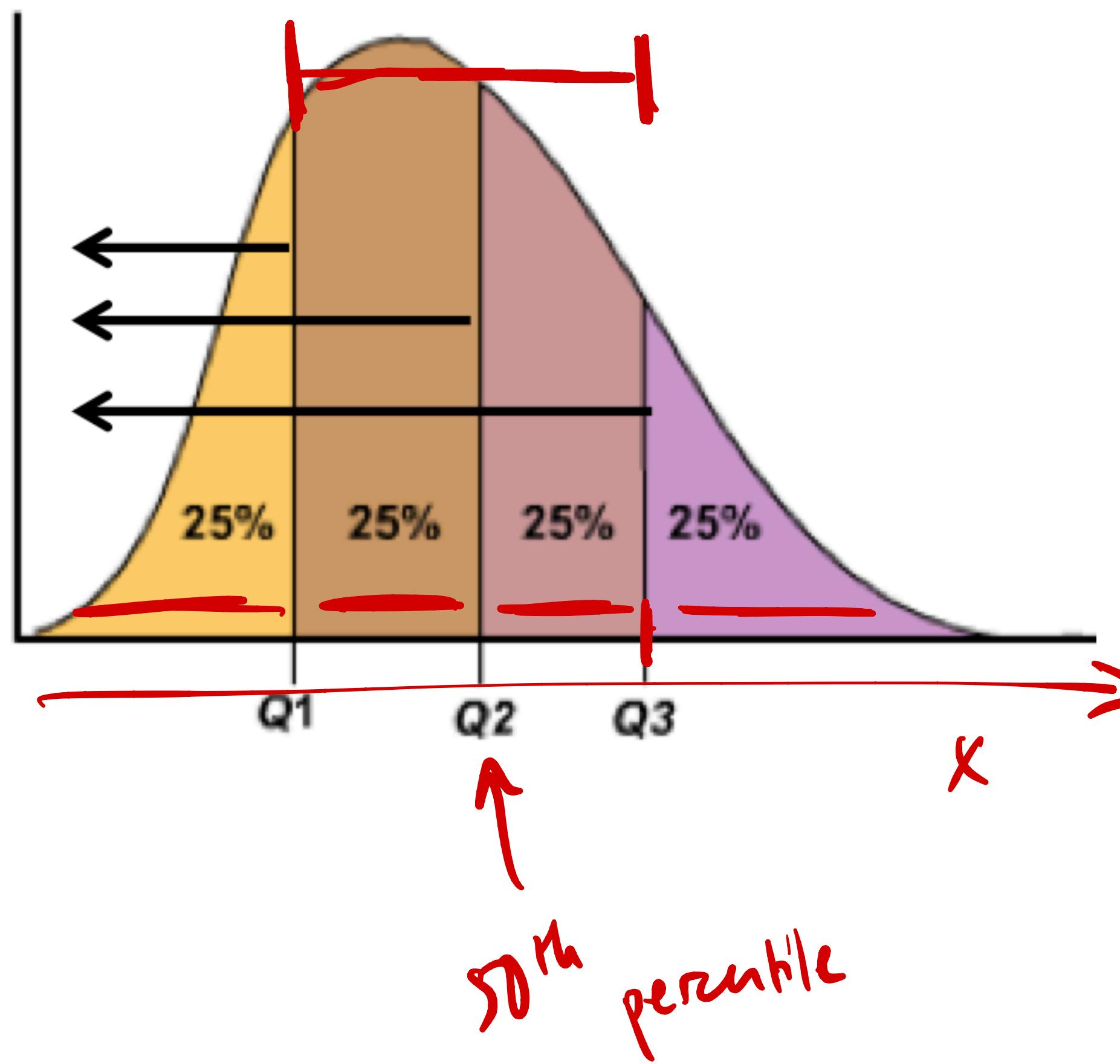
- **Interquartile Range (IQR)**: the difference between the 75<sup>th</sup> percentile and the 25<sup>th</sup> percentile
- $IQR = Q3 - Q1$
- Middle 50% of the observations in a given dataset
- R code: (Option 1) `IQR(data)`  
OR (Option 2) `quantile(data, 0.75) - quantile(data, 0.25)`

# Interquartile Range (IQR)

# Interquartile Range (IQR)



# Interquartile Range (IQR)



# Interquartile Range (IQR): Step by Step

# Interquartile Range (IQR): Step by Step

- Find the median:  $\tilde{x} = \frac{x_{(\lfloor \frac{n+1}{2} \rfloor)} + x_{(\lceil \frac{n+1}{2} \rceil)}}{2}$

# Interquartile Range (IQR): Step by Step

- Find the median:  $\tilde{x} = \frac{x_{(\lfloor \frac{n+1}{2} \rfloor)} + x_{(\lceil \frac{n+1}{2} \rceil)}}{2}$
- Lower half of data: Everything less than or equal to median ( $\leq \tilde{x}$ )

# Interquartile Range (IQR): Step by Step

- Find the median:  $\tilde{x} = \frac{x_{(\lfloor \frac{n+1}{2} \rfloor)} + x_{(\lceil \frac{n+1}{2} \rceil)}}{2}$
- Lower half of data: Everything less than or equal to median ( $\leq \tilde{x}$ )
- Upper half of data: Everything greater than or equal to median ( $\geq \tilde{x}$ )

# Interquartile Range (IQR): Step by Step

- Find the median:  $\tilde{x} = \frac{x_{(\lfloor \frac{n+1}{2} \rfloor)} + x_{(\lceil \frac{n+1}{2} \rceil)}}{2}$
- Lower half of data: Everything less than or equal to median ( $\leq \tilde{x}$ )
- Upper half of data: Everything greater than or equal to median ( $\geq \tilde{x}$ )
- Q1 (25<sup>th</sup> percentile): Median of the lower half of data

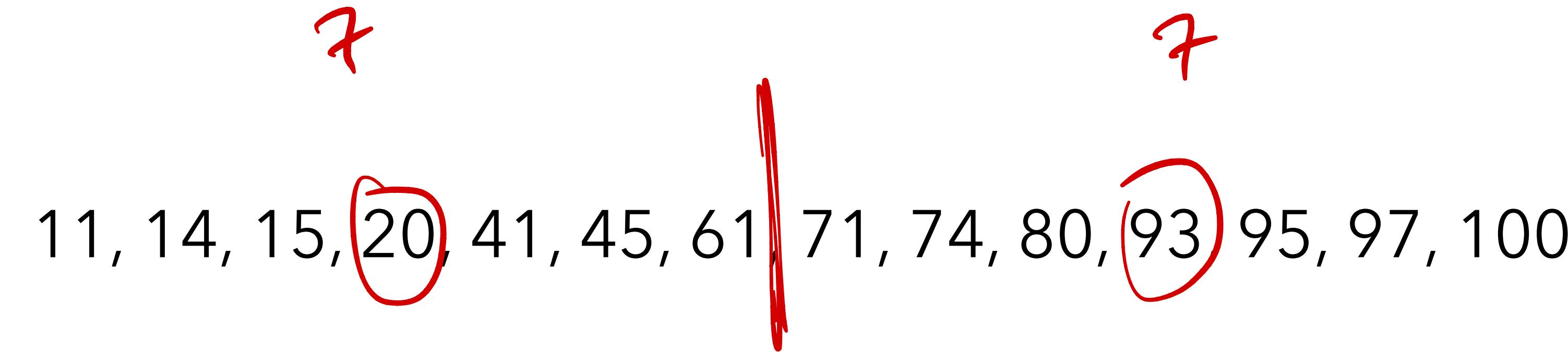
# Interquartile Range (IQR): Step by Step

- Find the median:  $\tilde{x} = \frac{x_{(\lfloor \frac{n+1}{2} \rfloor)} + x_{(\lceil \frac{n+1}{2} \rceil)}}{2}$
- Lower half of data: Everything less than or equal to median ( $\leq \tilde{x}$ )
- Upper half of data: Everything greater than or equal to median ( $\geq \tilde{x}$ )
- Q1 (25<sup>th</sup> percentile): Median of the lower half of data
- Q3 (75<sup>th</sup> percentile): Median of the upper half of data

# Interquartile Range (IQR): Step by Step

- Find the median:  $\tilde{x} = \frac{x_{(\lfloor \frac{n+1}{2} \rfloor)} + x_{(\lceil \frac{n+1}{2} \rceil)}}{2}$
- Lower half of data: Everything less than or equal to median ( $\leq \tilde{x}$ )
- Upper half of data: Everything greater than or equal to median ( $\geq \tilde{x}$ )
- Q1 (25<sup>th</sup> percentile): Median of the lower half of data
- Q3 (75<sup>th</sup> percentile): Median of the upper half of data
- IQR: Q3 - Q1

# Interquartile Range (IQR): Example, $n = 14$



# Boxplot

# Boxplot

- Boxplots differ from histograms in that they only require one axis and display only summary statistics based on *quartiles*

# Boxplot

- Boxplots differ from histograms in that they only require one axis and display only summary statistics based on *quartiles*
- **Quartile**: percentiles that break the data into four equal groups

# Boxplot

- Boxplots differ from histograms in that they only require one axis and display only summary statistics based on *quartiles*
- **Quartile**: percentiles that break the data into four equal groups
  - 1<sup>st</sup> quartile (Q1): 25<sup>th</sup> percentile

# Boxplot

- Boxplots differ from histograms in that they only require one axis and display only summary statistics based on *quartiles*
- **Quartile**: percentiles that break the data into four equal groups
  - 1<sup>st</sup> quartile (Q1): 25<sup>th</sup> percentile
  - 2<sup>nd</sup> quartile (Q2): 50<sup>th</sup> percentile (median)

# Boxplot

- Boxplots differ from histograms in that they only require one axis and display only summary statistics based on *quartiles*
- **Quartile**: percentiles that break the data into four equal groups
  - 1<sup>st</sup> quartile (Q1): 25<sup>th</sup> percentile
  - 2<sup>nd</sup> quartile (Q2): 50<sup>th</sup> percentile (median)
  - 3<sup>rd</sup> quartile (Q3): 75<sup>th</sup> percentile

# Boxplot

# Boxplot

- Five number summary: *minimum, Q1, Q2, Q3, and maximum*

# Boxplot

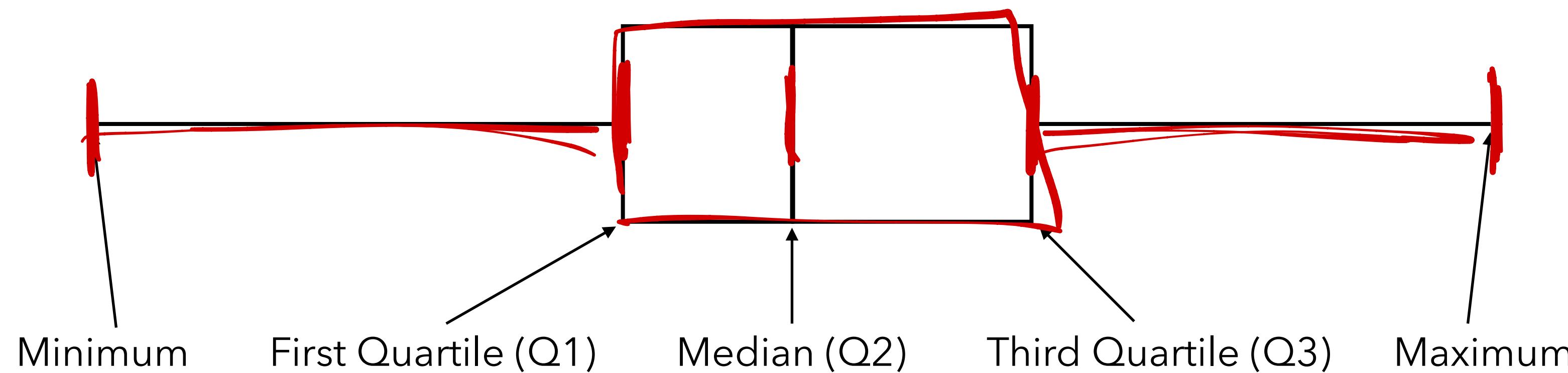
- Five number summary: *minimum, Q1, Q2, Q3, and maximum*
- A skeletal boxplot uses only the five number summary

# Boxplot

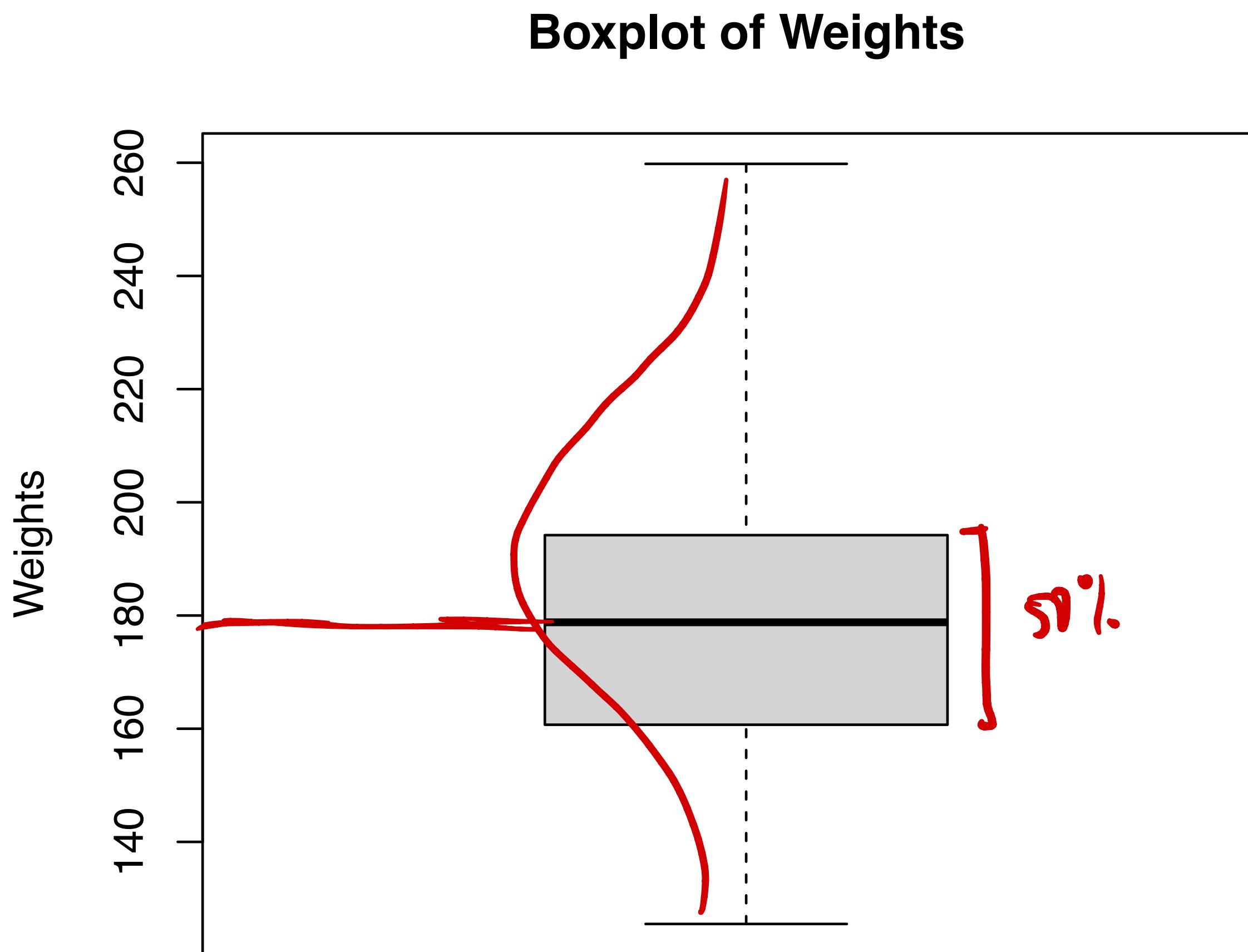
- Five number summary: *minimum, Q1, Q2, Q3, and maximum*
- A skeletal boxplot uses only the five number summary
- In R: fivenum(data)

# Boxplot

- Five number summary: *minimum*,  $Q_1$ ,  $Q_2$ ,  $Q_3$ , and *maximum*
- A skeletal boxplot uses only the five number summary
- In R: `fivenum(data)`



# Boxplot: Example



```
boxplot(weights, range=0, main="Boxplot of Weights", ylab="Weights")
```

# Modified Boxplot

# Modified Boxplot

- A **modified boxplot** uses the quartiles, but it also distinguishes outliers from the rest of the data

# Modified Boxplot

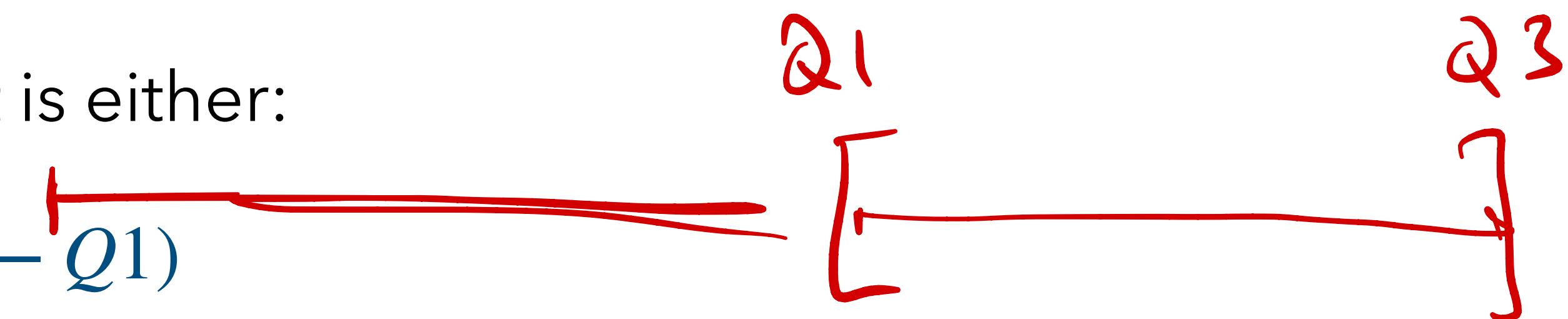
- A **modified boxplot** uses the quartiles, but it also distinguishes outliers from the rest of the data
- Modified boxplots are the standard boxplots that are, generally, the most beneficial to use

# Modified Boxplot

- A **modified boxplot** uses the quartiles, but it also distinguishes outliers from the rest of the data
- Modified boxplots are the standard boxplots that are, generally, the most beneficial to use
- An outlier is a data point that is either:

# Modified Boxplot

- A **modified boxplot** uses the quartiles, but it also distinguishes outliers from the rest of the data
- Modified boxplots are the standard boxplots that are, generally, the most beneficial to use
- An outlier is a data point that is either:
  - Less than:  $Q1 - 1.5 \times (Q3 - Q1)$



# Modified Boxplot

- A **modified boxplot** uses the quartiles, but it also distinguishes outliers from the rest of the data
- Modified boxplots are the standard boxplots that are, generally, the most beneficial to use
- An outlier is a data point that is either:
  - Less than:  $Q1 - 1.5 \times (Q3 - Q1)$
  - Greater than:  $Q3 + 1.5 \times (Q3 - Q1)$

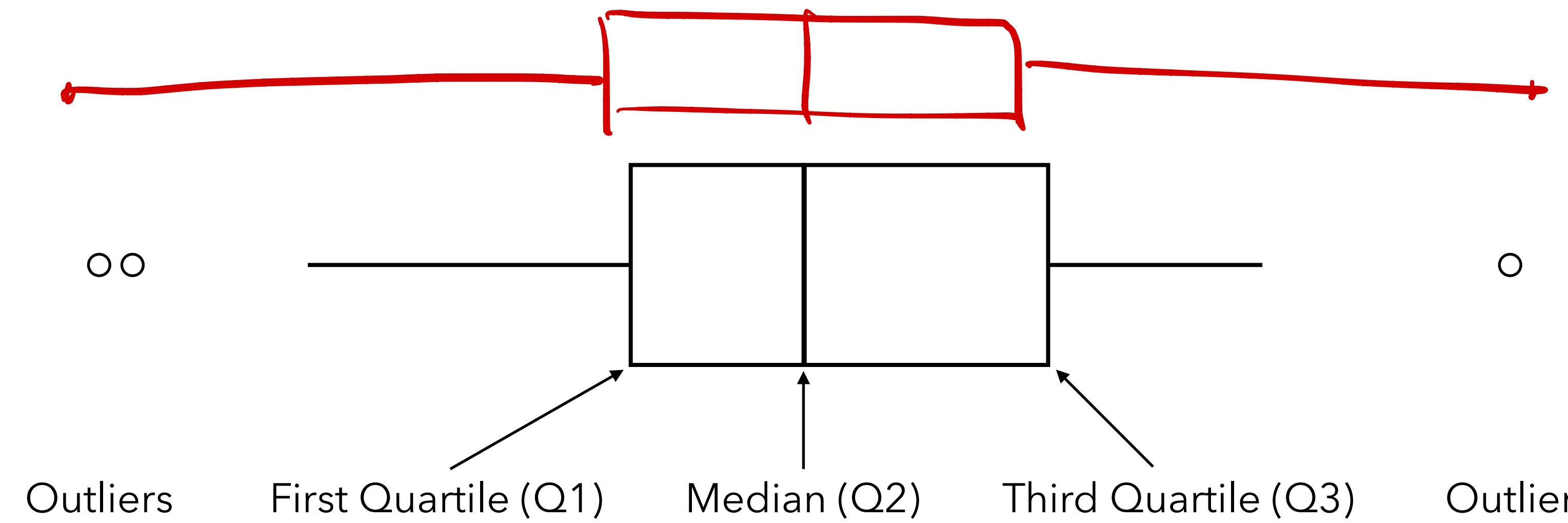
# Modified Boxplot

- A **modified boxplot** uses the quartiles, but it also distinguishes outliers from the rest of the data
- Modified boxplots are the standard boxplots that are, generally, the most beneficial to use
- An outlier is a data point that is either:
  - Less than:  $Q1 - 1.5 \times (Q3 - Q1)$
  - Greater than:  $Q3 + 1.5 \times (Q3 - Q1)$
  - Standard span:  $1.5 \times (Q3 - Q1) = 1.5 \times IQR$

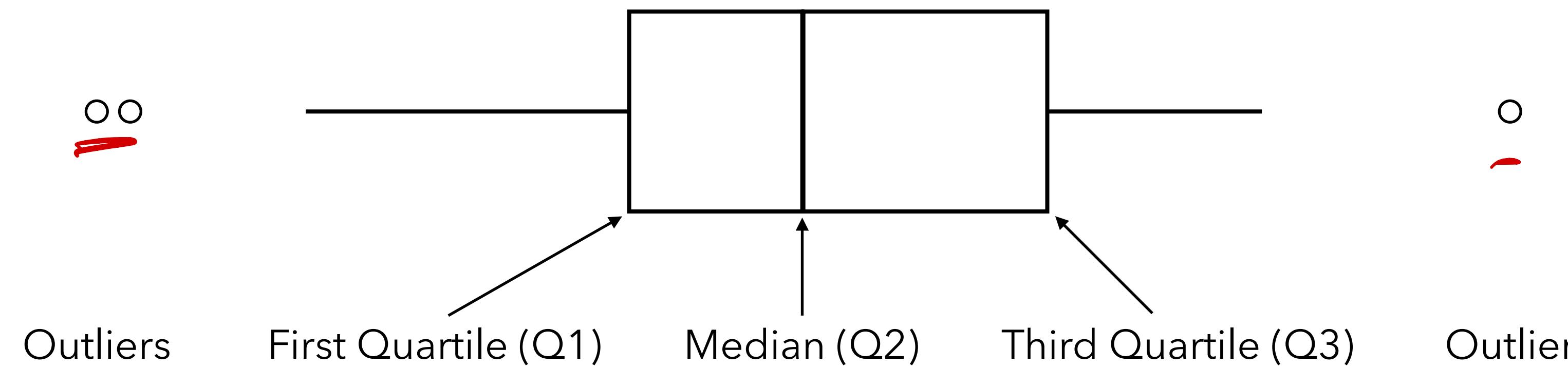
# Modified Boxplot

- A **modified boxplot** uses the quartiles, but it also distinguishes outliers from the rest of the data
- Modified boxplots are the standard boxplots that are, generally, the most beneficial to use
- An outlier is a data point that is either:
  - Less than:  $Q1 - 1.5 \times (Q3 - Q1)$
  - Greater than:  $Q3 + 1.5 \times (Q3 - Q1)$
- Standard span:  $1.5 \times (Q3 - Q1) = 1.5 \times IQR$
- Now, whiskers extend only to the highest / lowest non-outlier points, and the outliers are distinguished as separate points

# Modified Boxplot

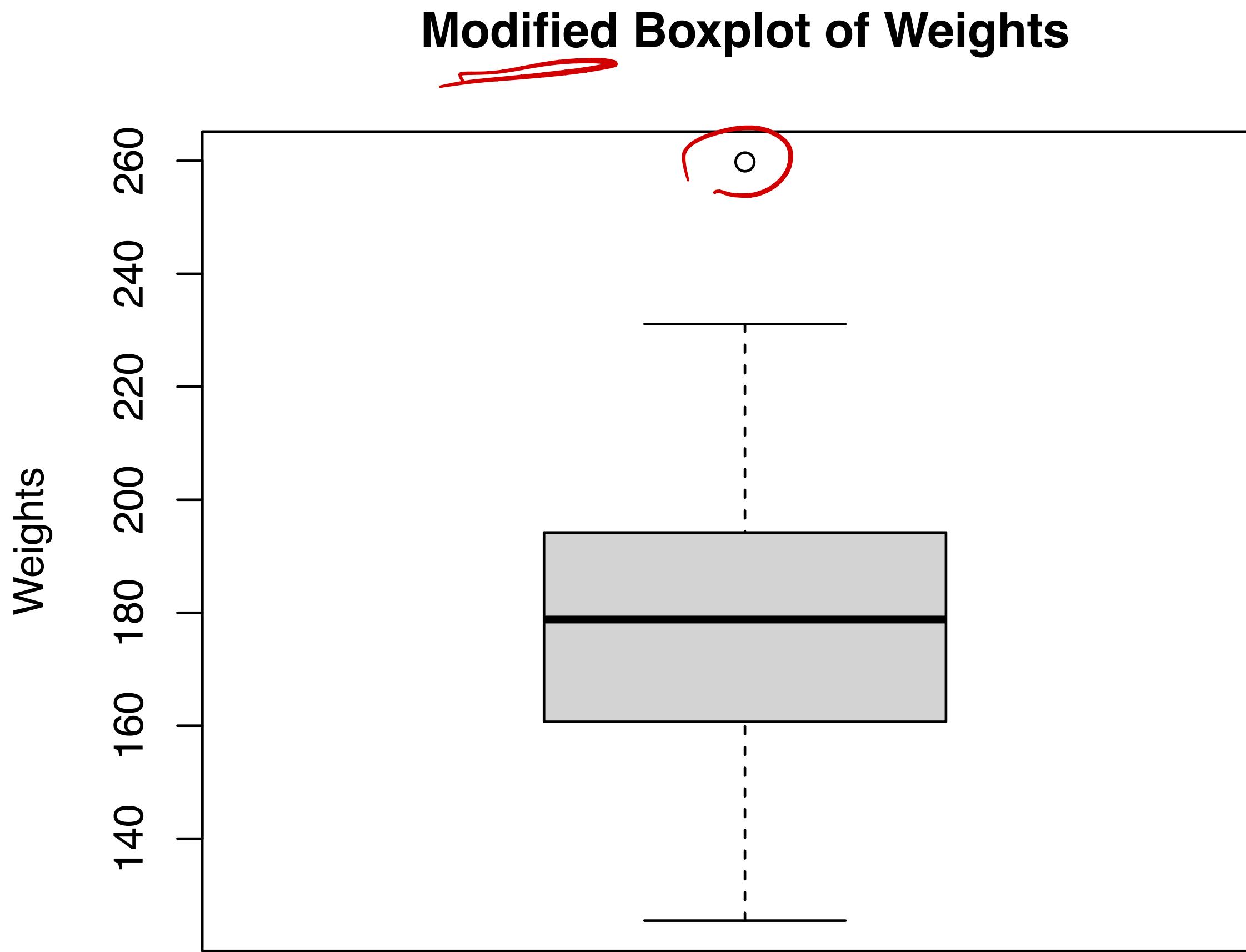


# Modified Boxplot



- Note that whiskers extend to the most extreme observation that isn't an outlier, and there can be multiple outliers beyond the whiskers

# Modified Boxplot: Example



```
boxplot(weights, main="Modified Boxplot of Weights", ylab="Weights")
```

# Modified Boxplot: Step by Step

# Modified Boxplot: Step by Step

- Construct an axis covering all observations

# Modified Boxplot: Step by Step

- Construct an axis covering all observations
- Draw a box with lines at Q1, Q2, and Q3

# Modified Boxplot: Step by Step

- Construct an axis covering all observations
- Draw a box with lines at  $Q_1$ ,  $Q_2$ , and  $Q_3$
- Locate the *upper fence*:  $Q_3 + 1.5 \times (Q_3 - Q_1)$  and *lower fence*:  
 $Q_1 - 1.5 \times (Q_3 - Q_1)$

# Modified Boxplot: Step by Step

- Construct an axis covering all observations
- Draw a box with lines at  $Q_1$ ,  $Q_2$ , and  $Q_3$
- Locate the *upper fence*:  $Q_3 + 1.5 \times (Q_3 - Q_1)$  and *lower fence*:  
 $Q_1 - 1.5 \times (Q_3 - Q_1)$
- Indicate the upper whisker, given by the largest observation that is less than the upper fence

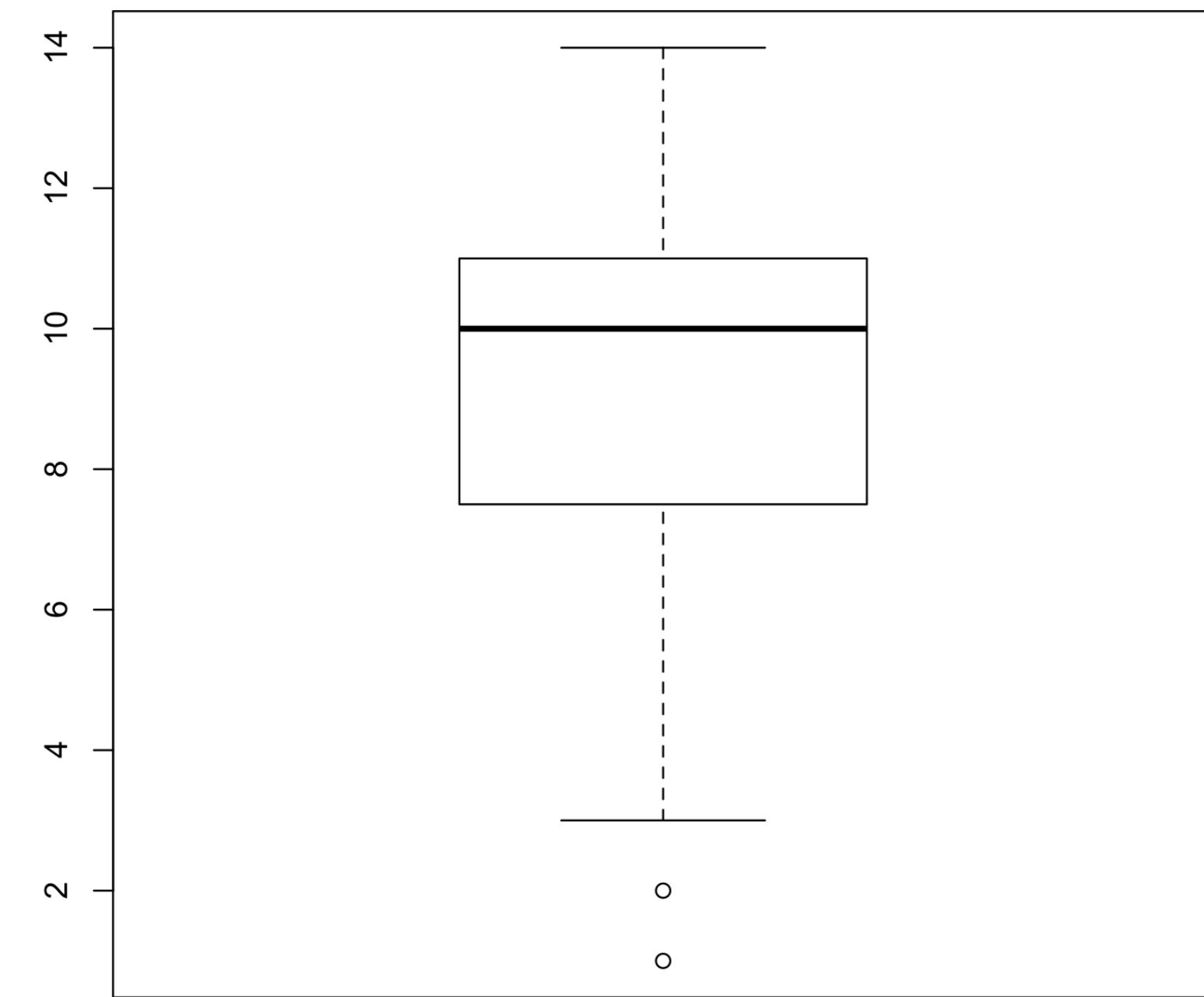
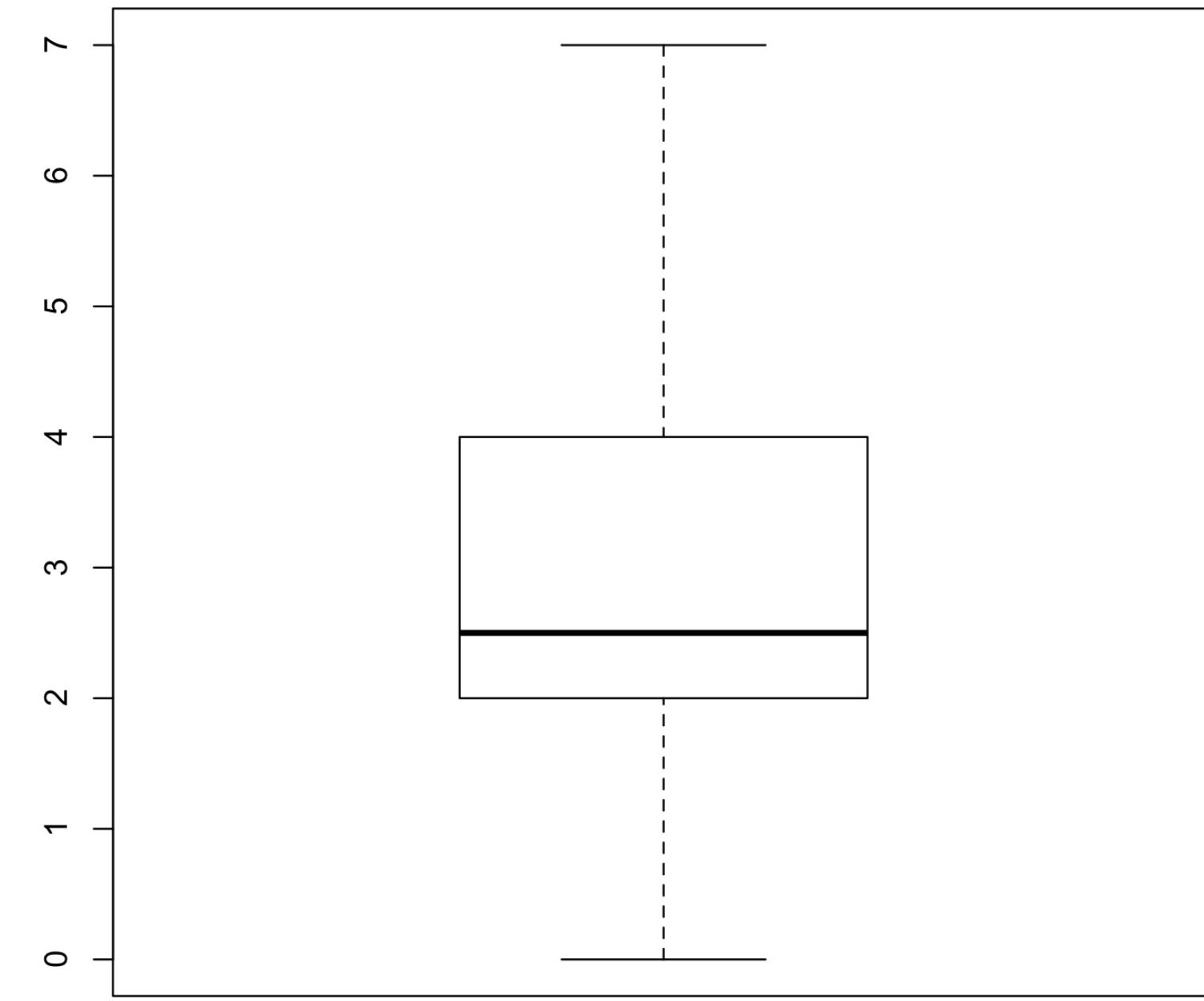
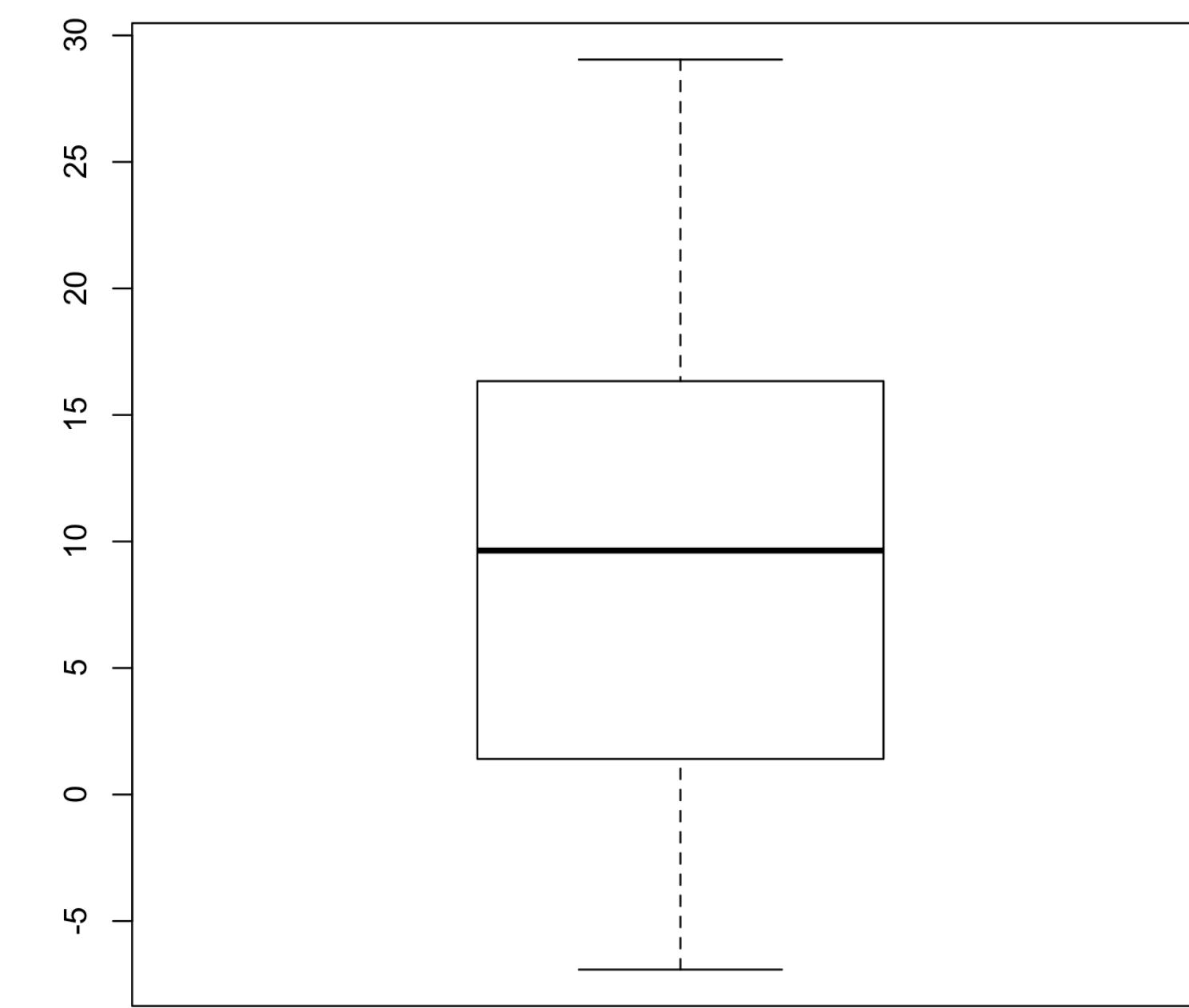
# Modified Boxplot: Step by Step

- Construct an axis covering all observations
- Draw a box with lines at  $Q_1$ ,  $Q_2$ , and  $Q_3$
- Locate the *upper fence*:  $Q_3 + 1.5 \times (Q_3 - Q_1)$  and *lower fence*:  
 $Q_1 - 1.5 \times (Q_3 - Q_1)$
- Indicate the upper whisker, given by the largest observation that is less than the upper fence
- Indicate the lower whisker, given by the smallest observation that is greater than the lower fence

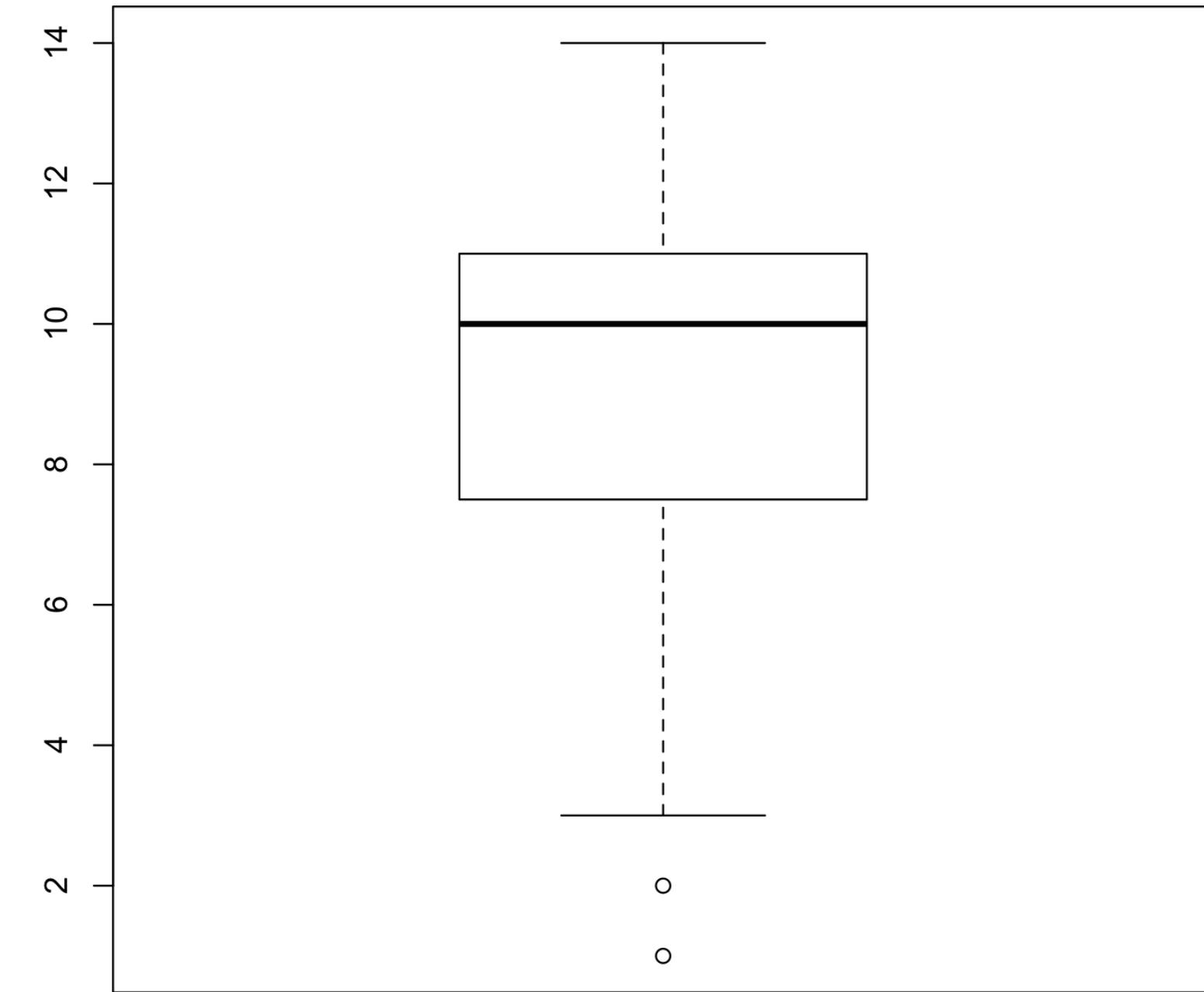
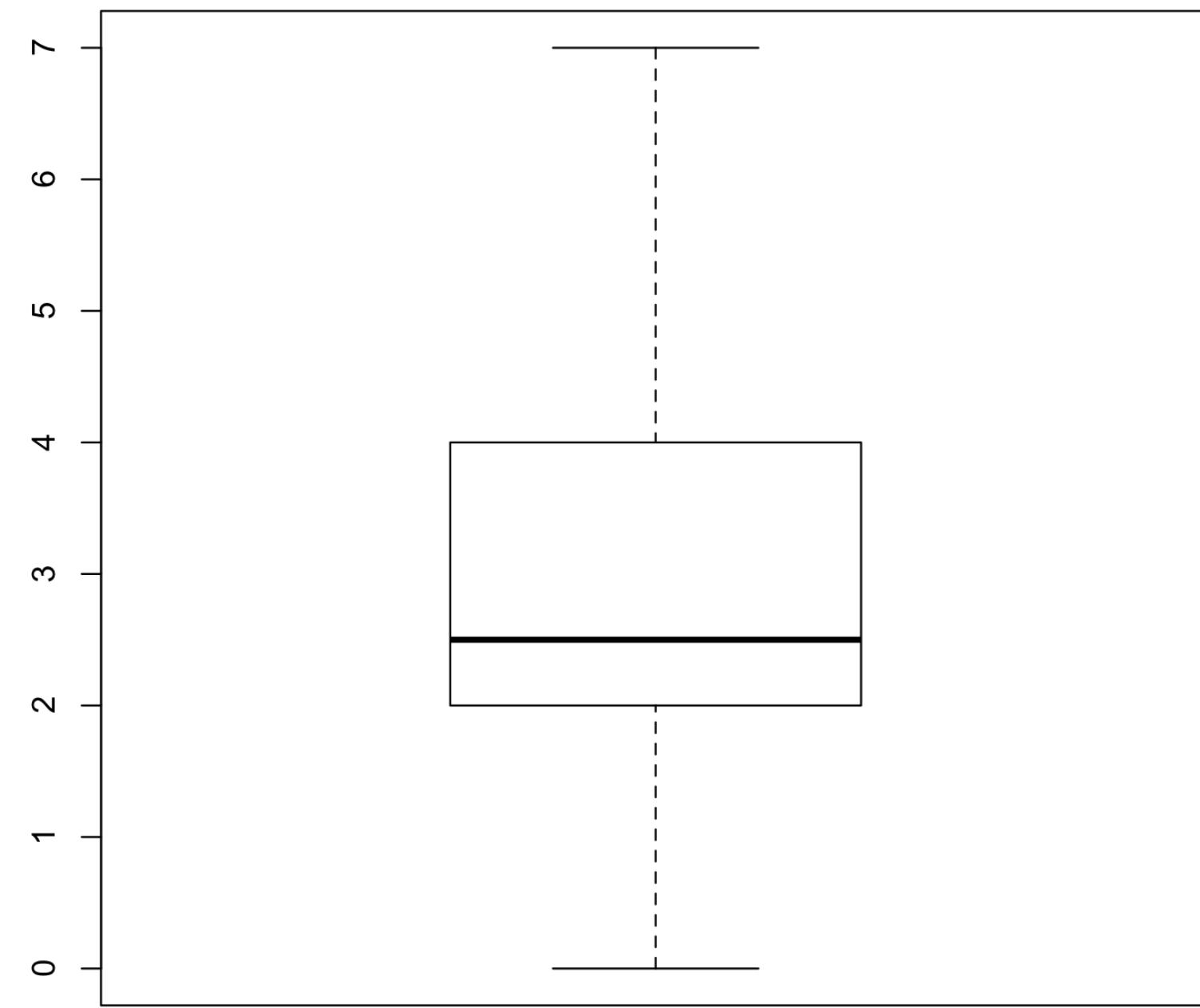
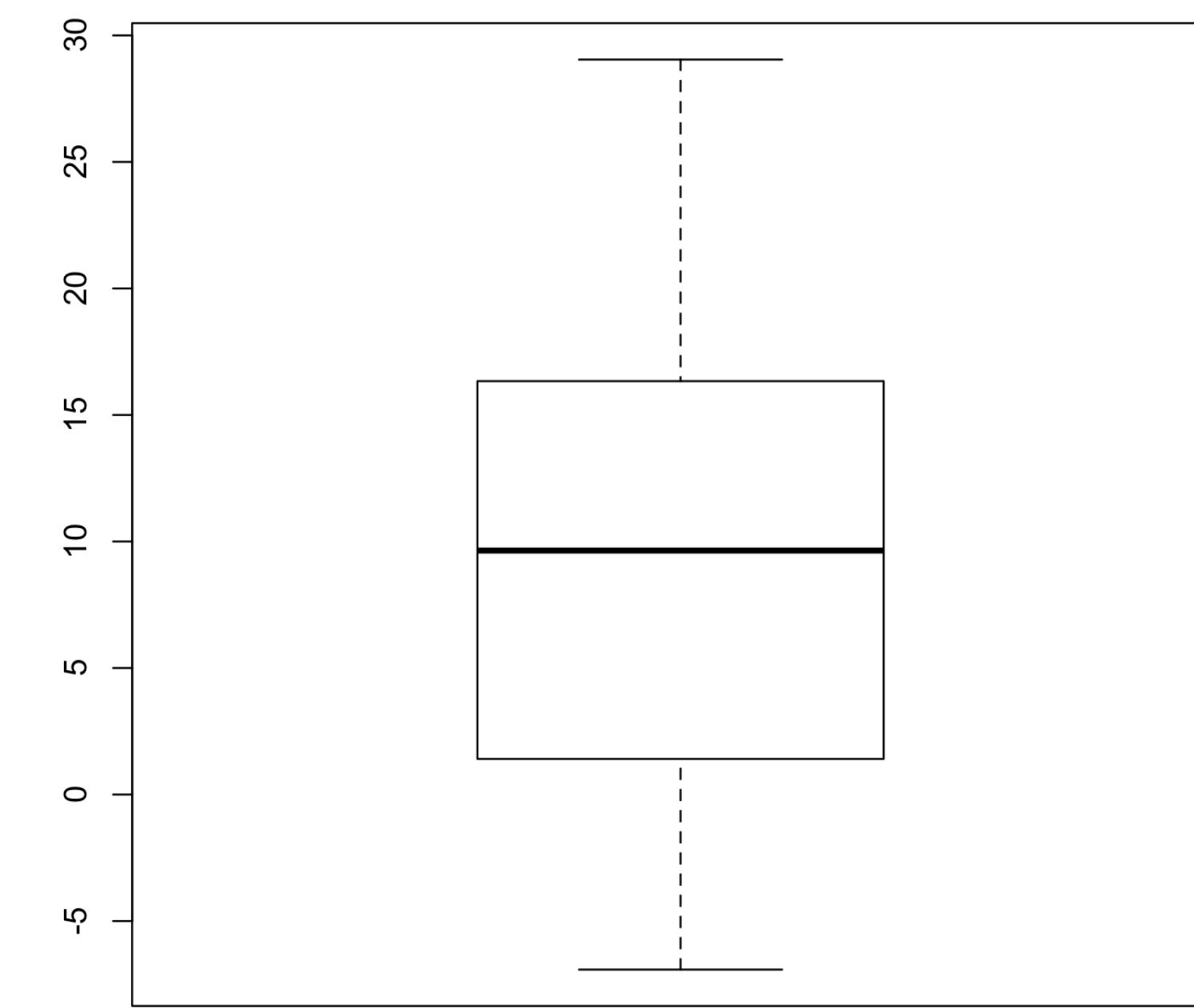
# Modified Boxplot: Step by Step

- Construct an axis covering all observations
- Draw a box with lines at  $Q_1$ ,  $Q_2$ , and  $Q_3$
- Locate the *upper fence*:  $Q_3 + 1.5 \times (Q_3 - Q_1)$  and *lower fence*:  
 $Q_1 - 1.5 \times (Q_3 - Q_1)$
- Indicate the upper whisker, given by the largest observation that is less than the upper fence
- Indicate the lower whisker, given by the smallest observation that is greater than the lower fence
- Mark outliers as individual points beyond the fences

# (Modified) Boxplot Examples

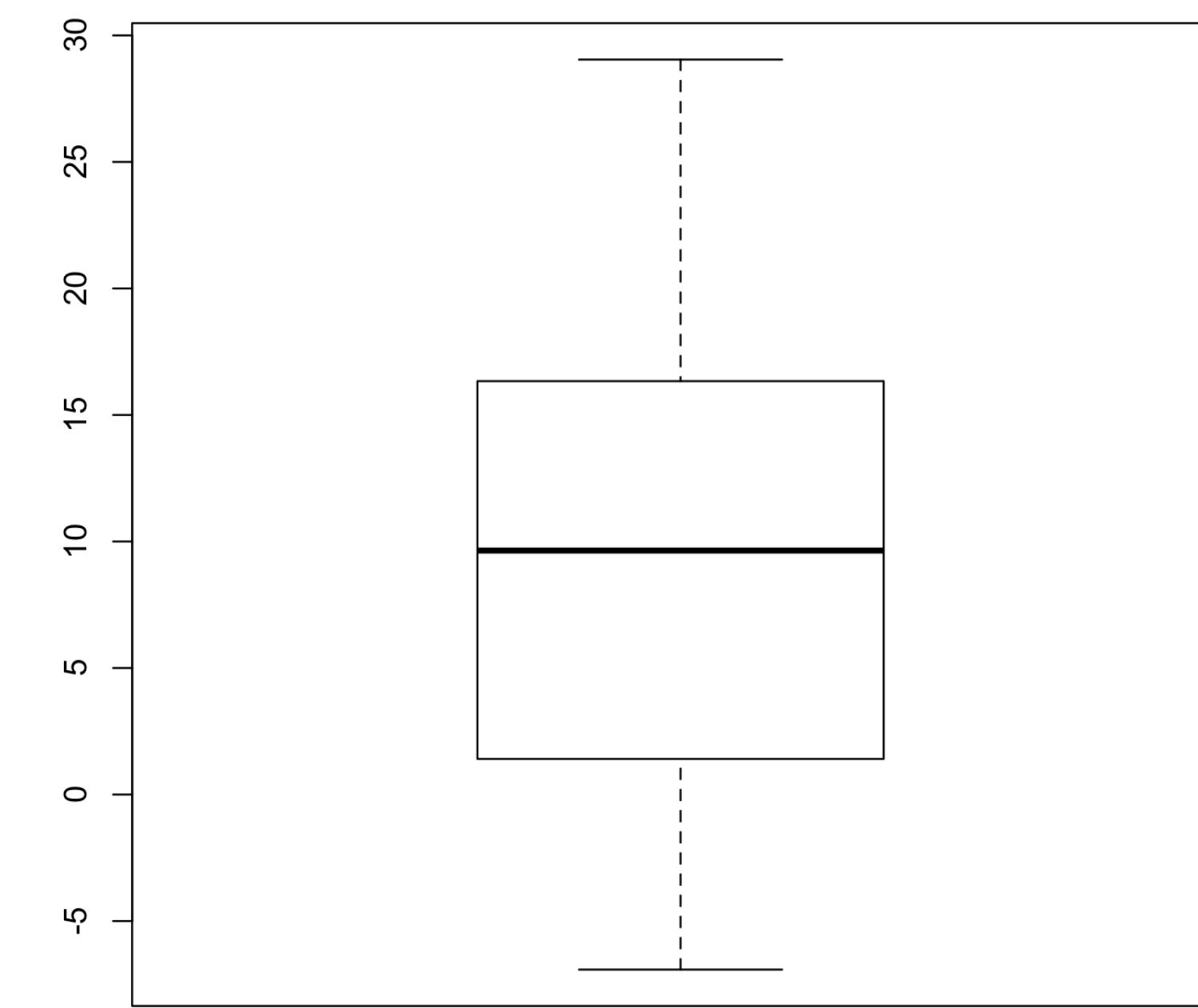


# (Modified) Boxplot Examples

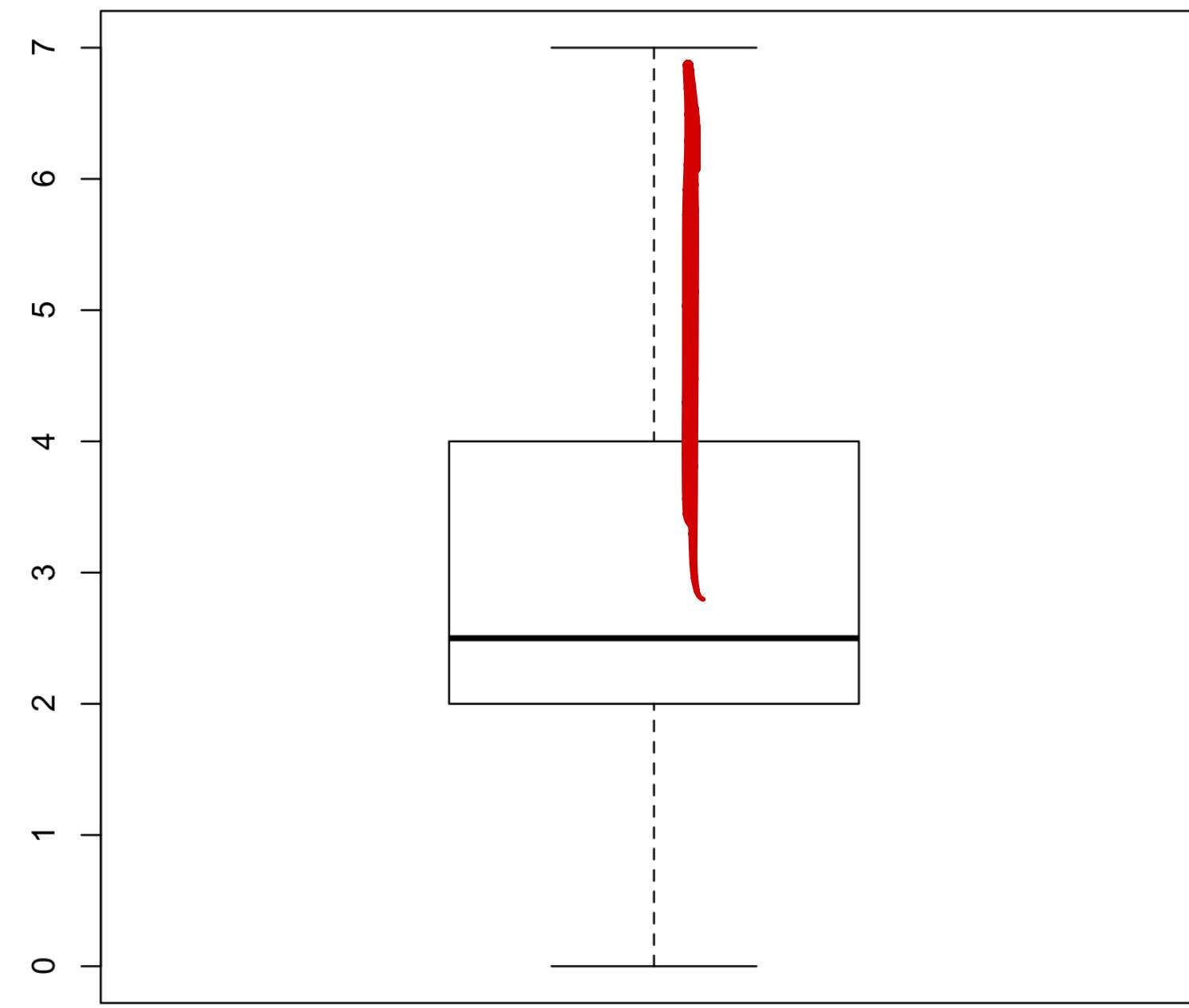


Symmetric

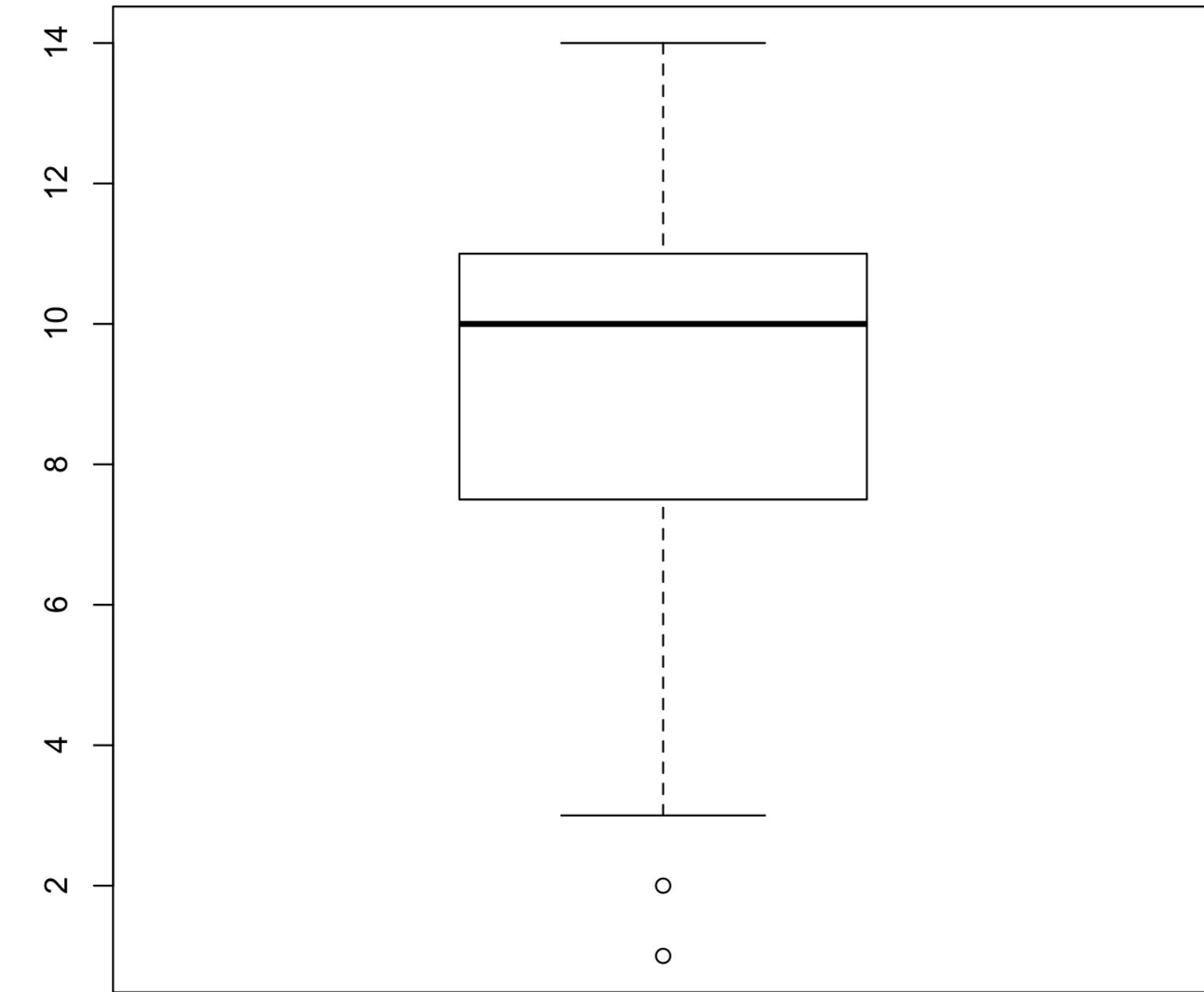
# (Modified) Boxplot Examples



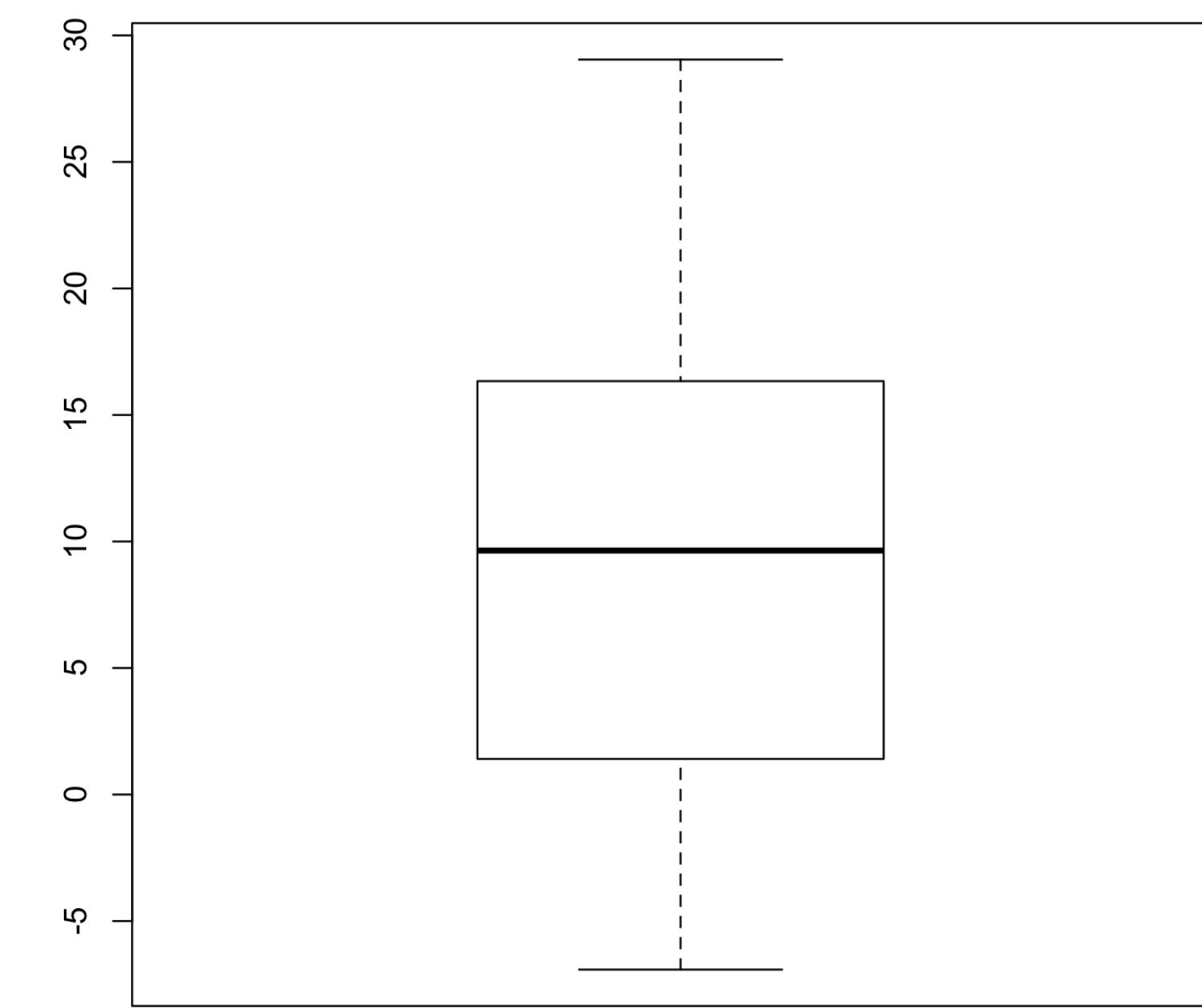
Symmetric



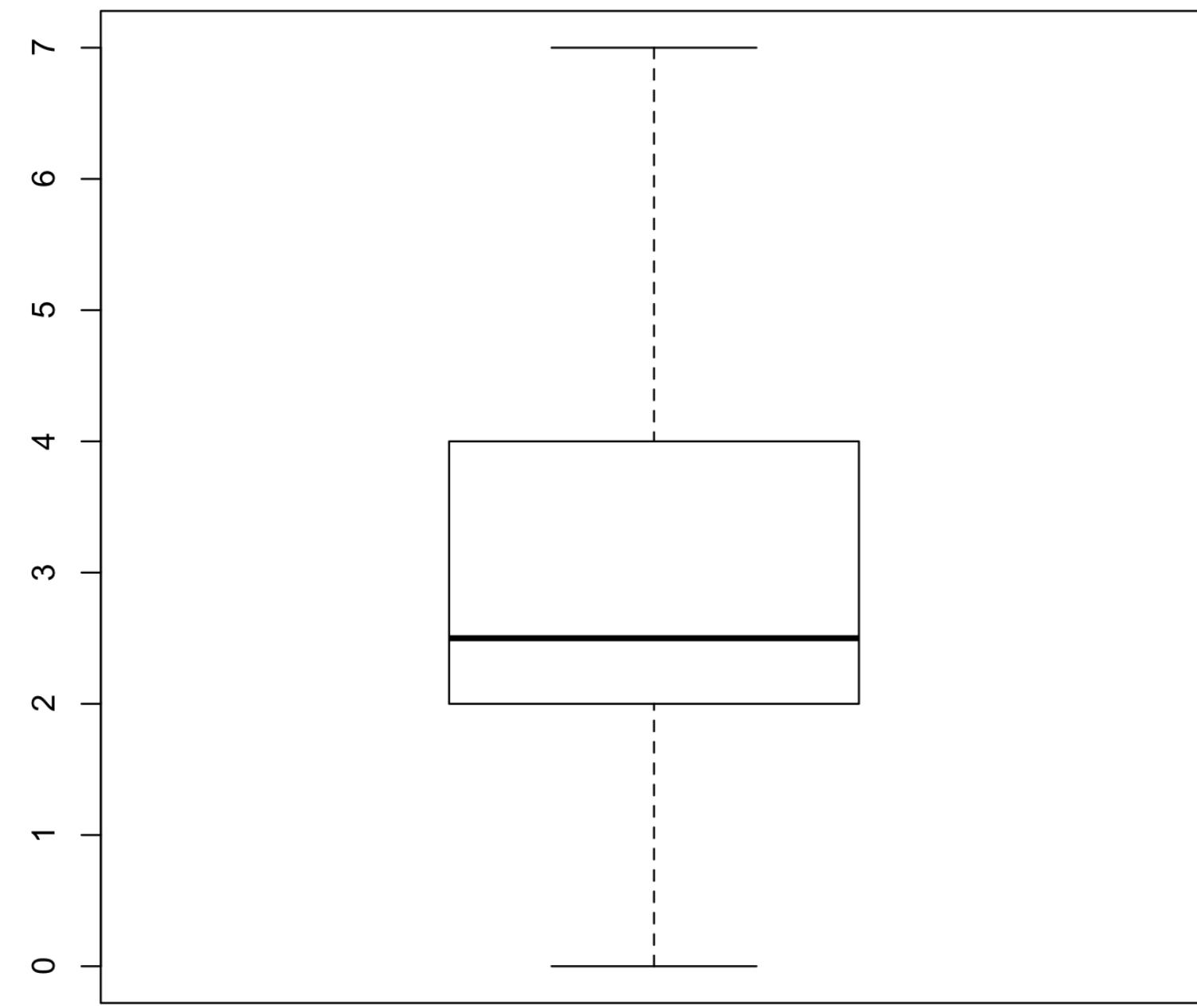
Asymmetric  
(Right Skew)



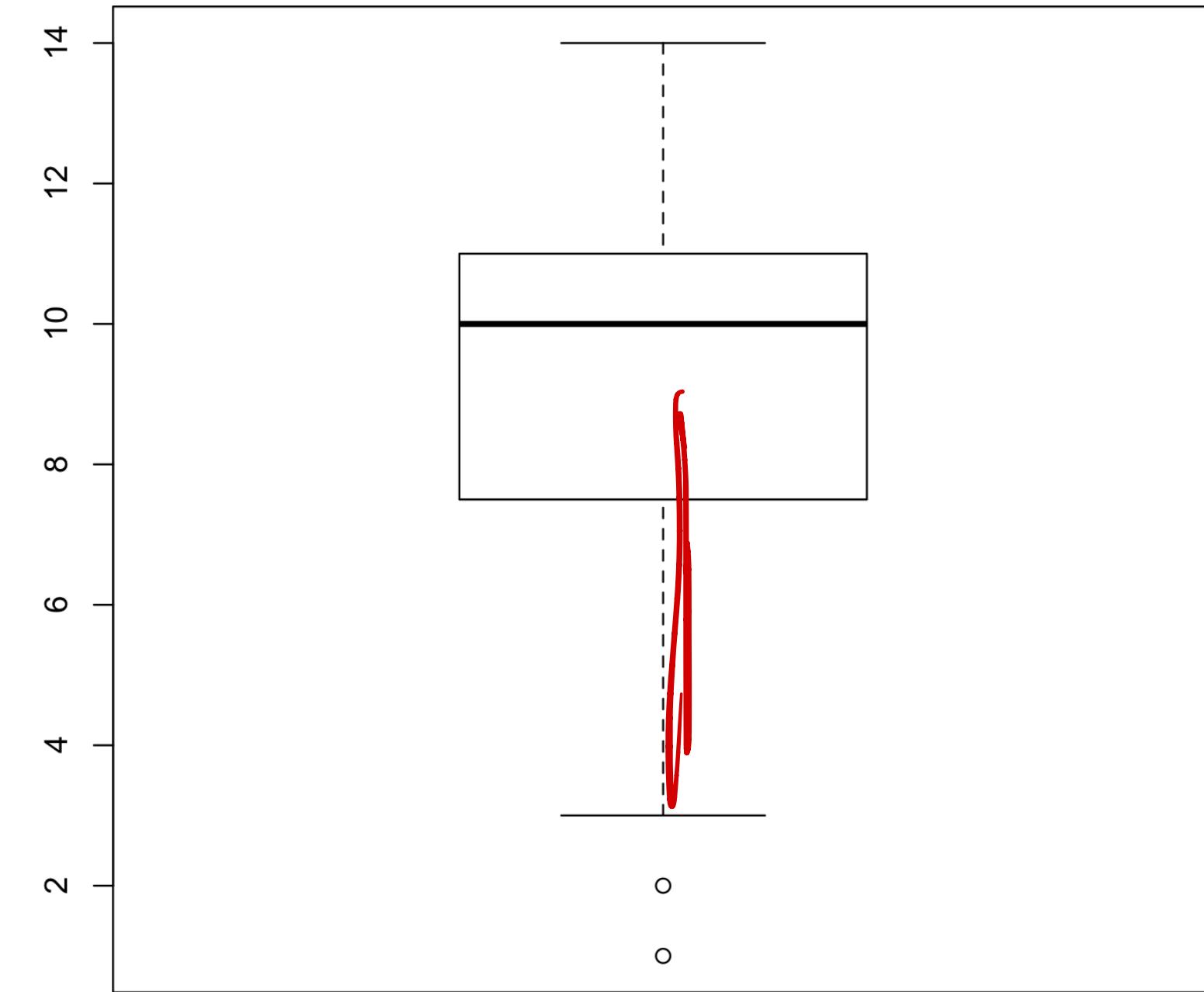
# (Modified) Boxplot Examples



Symmetric



Asymmetric  
(Right Skew)



Asymmetric  
(Left Skew)

# Measures of Dispersion: Variance

# Measures of Dispersion: Variance

- **Variance:** The average squared deviation of observations from the mean

# Measures of Dispersion: Variance

- **Variance:** The average squared deviation of observations from the mean
- Measured in squared units (e.g., if measuring weights, units are pounds<sup>2</sup>)

# Measures of Dispersion: Variance

- **Variance:** The average squared deviation of observations from the mean
- Measured in squared units (e.g., if measuring weights, units are pounds<sup>2</sup>)
- We calculate the *sample variance* of our observations as follows:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

# Measures of Dispersion: Variance

- **Variance:** The average squared deviation of observations from the mean
- Measured in squared units (e.g., if measuring weights, units are pounds<sup>2</sup>)
- We calculate the *sample variance* of our observations as follows:

$$s^2 = \frac{1}{n - 1} \sum_{i=1}^n (x_i - \bar{x})^2$$

- R code: var (data)

# Measures of Dispersion: Standard Deviation

# Measures of Dispersion: Standard Deviation

- **Standard deviation (SD):** The positive square root of the variance

# Measures of Dispersion: Standard Deviation

- **Standard deviation (SD):** The positive square root of the variance
- Size of the “typical” deviation from the mean, measured in the same units as observations

# Measures of Dispersion: Standard Deviation

- **Standard deviation (SD):** The positive square root of the variance
- Size of the “typical” deviation from the mean, measured in the same units as observations
- We calculate the *sample standard deviation* of our observations as follows:

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

# Measures of Dispersion: Standard Deviation

- **Standard deviation (SD):** The positive square root of the variance
- Size of the “typical” deviation from the mean, measured in the same units as observations
- We calculate the *sample standard deviation* of our observations as follows:

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

- R code: `sd(data)`

# Variance and SD vs. IQR

# Variance and SD vs. IQR

- In general, variance and standard deviation are more sensitive to outliers than the interquartile range

# Variance and SD vs. IQR

- In general, variance and standard deviation are more sensitive to outliers than the interquartile range
- Typically, use mean and variance (and standard deviation) when the data is symmetric; when the data is skewed, use median and IQR

# Variance and SD vs. IQR

- In general, variance and standard deviation are more sensitive to outliers than the interquartile range
- Typically, use mean and variance (and standard deviation) when the data is symmetric; when the data is skewed, use median and IQR
- The magnitude of the variance or SD depends on the problem

# Variance and SD vs. IQR

- In general, variance and standard deviation are more sensitive to outliers than the interquartile range
- Typically, use mean and variance (and standard deviation) when the data is symmetric; when the data is skewed, use median and IQR
- The magnitude of the variance or SD depends on the problem
  - What is large for one group may be small for another

# Variance and SD vs. IQR

- In general, variance and standard deviation are more sensitive to outliers than the interquartile range
- Typically, use mean and variance (and standard deviation) when the data is symmetric; when the data is skewed, use median and IQR
- The magnitude of the variance or SD depends on the problem
  - What is large for one group may be small for another
  - Remember that SD is on the same scale as the observations

# Updating Sample Mean and Variance

# Updating Sample Mean and Variance

- When a new data point is added to a set of  $n$  data points, we don't have to completely recalculate the mean and variance

# Updating Sample Mean and Variance

- When a new data point is added to a set of  $n$  data points, we don't have to completely recalculate the mean and variance
- Let  $\bar{x}_n$  and  $s_n^2$  be the sample mean and variance, respectively, of our  $n$  original data points

# Updating Sample Mean and Variance

- When a new data point is added to a set of  $n$  data points, we don't have to completely recalculate the mean and variance
- Let  $\bar{x}_n$  and  $s_n^2$  be the sample mean and variance, respectively, of our  $n$  original data points
- Updating for the  $(n + 1)^{st}$  data point, we get:

# Updating Sample Mean and Variance

- When a new data point is added to a set of  $n$  data points, we don't have to completely recalculate the mean and variance
- Let  $\bar{x}_n$  and  $s_n^2$  be the sample mean and variance, respectively, of our  $n$  original data points
- Updating for the  $(n + 1)^{st}$  data point, we get:

$$\bar{x}_{n+1} = \frac{n}{n+1}\bar{x}_n + \frac{1}{n+1}x_{n+1}$$

# Updating Sample Mean and Variance

- When a new data point is added to a set of  $n$  data points, we don't have to completely recalculate the mean and variance
- Let  $\bar{x}_n$  and  $s_n^2$  be the sample mean and variance, respectively, of our  $n$  original data points
- Updating for the  $(n + 1)^{st}$  data point, we get:

$$\bar{x}_{n+1} = \frac{n}{n+1}\bar{x}_n + \frac{1}{n+1}x_{n+1}$$

$$s_{n+1}^2 = \frac{n-1}{n}s_n^2 + \frac{1}{n+1}(x_{n+1} - \bar{x})^2$$

# Coefficient of Variation

# Coefficient of Variation

- The **coefficient of variation** is defined as the ratio of the standard deviation to the mean:  $CV = s/\bar{x}$

# Coefficient of Variation

- The **coefficient of variation** is defined as the ratio of the standard deviation to the mean:  $CV = s/\bar{x}$
- Describes dispersion in a way that does not depend on units

# Coefficient of Variation

- The **coefficient of variation** is defined as the ratio of the standard deviation to the mean:  $CV = s/\bar{x}$
- Describes dispersion in a way that does not depend on units
- Beneficial for comparing dispersion across different scenarios

# Coefficient of Variation

- The **coefficient of variation** is defined as the ratio of the standard deviation to the mean:  $CV = s/\bar{x}$
- Describes dispersion in a way that does not depend on units
- Beneficial for comparing dispersion across different scenarios
- The higher the CV, the higher the dispersion

# Coefficient of Variation

- The **coefficient of variation** is defined as the ratio of the standard deviation to the mean:  $CV = s/\bar{x}$
- Describes dispersion in a way that does not depend on units
- Beneficial for comparing dispersion across different scenarios
- The higher the CV, the higher the dispersion
- Only applicable for data on a ratio scale (meaningful zero)

# Skewness

# Skewness

- **Skewness:** a measure of asymmetry of a distribution

# Skewness

- **Skewness:** a measure of asymmetry of a distribution

$$\text{skew} = \frac{\cancel{m_3}}{s^3} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left[ \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \right]^{3/2}}$$

# Skewness

- **Skewness:** a measure of asymmetry of a distribution

$$\text{skew} = \frac{m_3}{s^3} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left[ \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \right]^{3/2}}$$

- Left (negative) skew: The left tail extends farther out than the right tail

# Skewness

- **Skewness:** a measure of asymmetry of a distribution

$$\text{skew} = \frac{m_3}{s^3} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left[ \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \right]^{3/2}}$$

- Left (negative) skew: The left tail extends farther out than the right tail
- Right (positive skew): The right tail extends farther out than the left tail

# Skewness

- **Skewness:** a measure of asymmetry of a distribution

$$\text{skew} = \frac{m_3}{s^3} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left[ \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \right]^{3/2}}$$

- Left (negative) skew: The left tail extends farther out than the right tail
- Right (positive skew): The right tail extends farther out than the left tail
- Symmetric distributions have skew 0

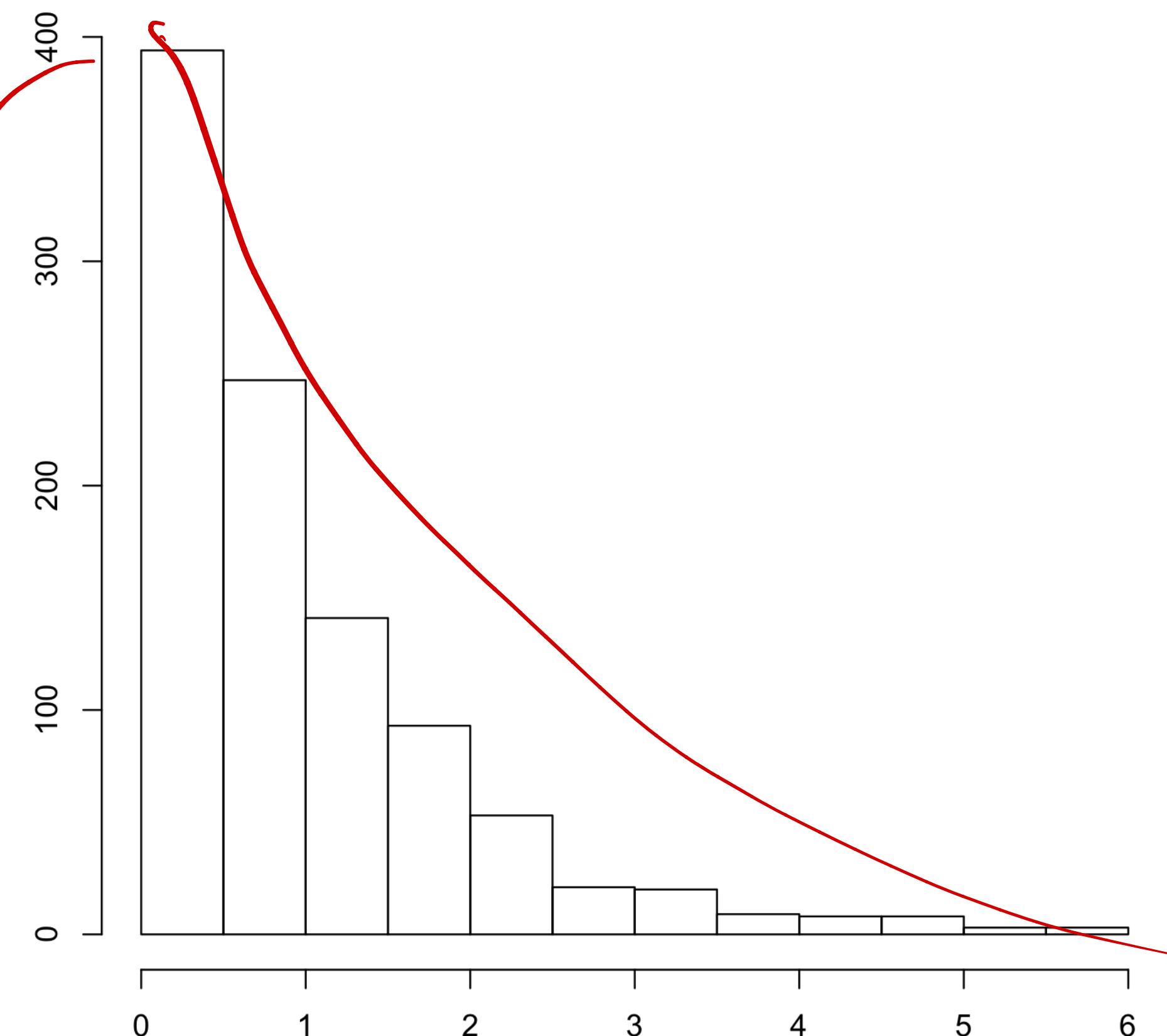
# Skewness

- **Skewness:** a measure of asymmetry of a distribution

$$\text{skew} = \frac{m_3}{s^3} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left[ \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \right]^{3/2}}$$

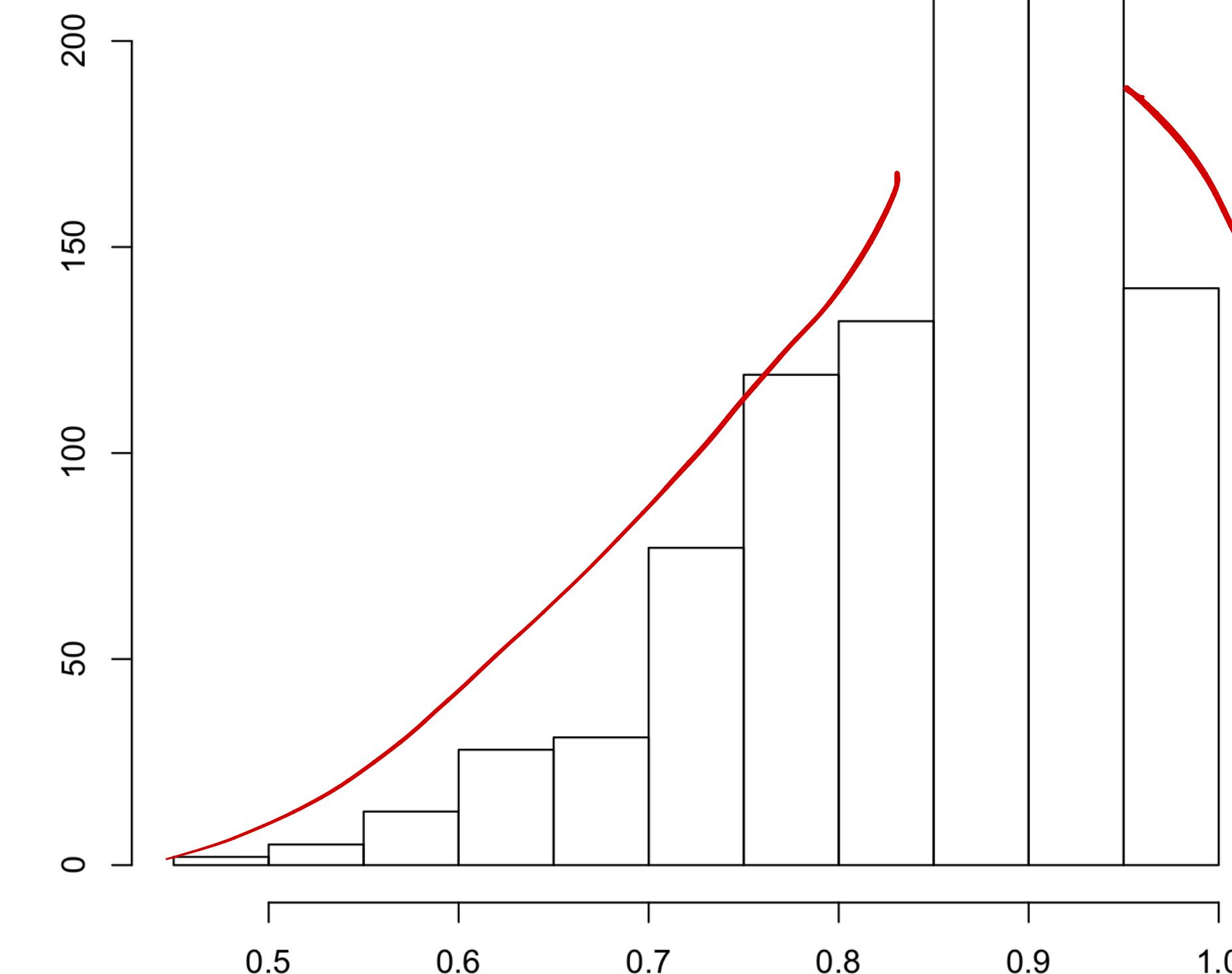
- Left (negative) skew: The left tail extends farther out than the right tail
- Right (positive skew): The right tail extends farther out than the left tail
- Symmetric distributions have skew 0
- R code: `library(moments); skewness(data)`

# Skewness: Examples



Right Skew (1.890)

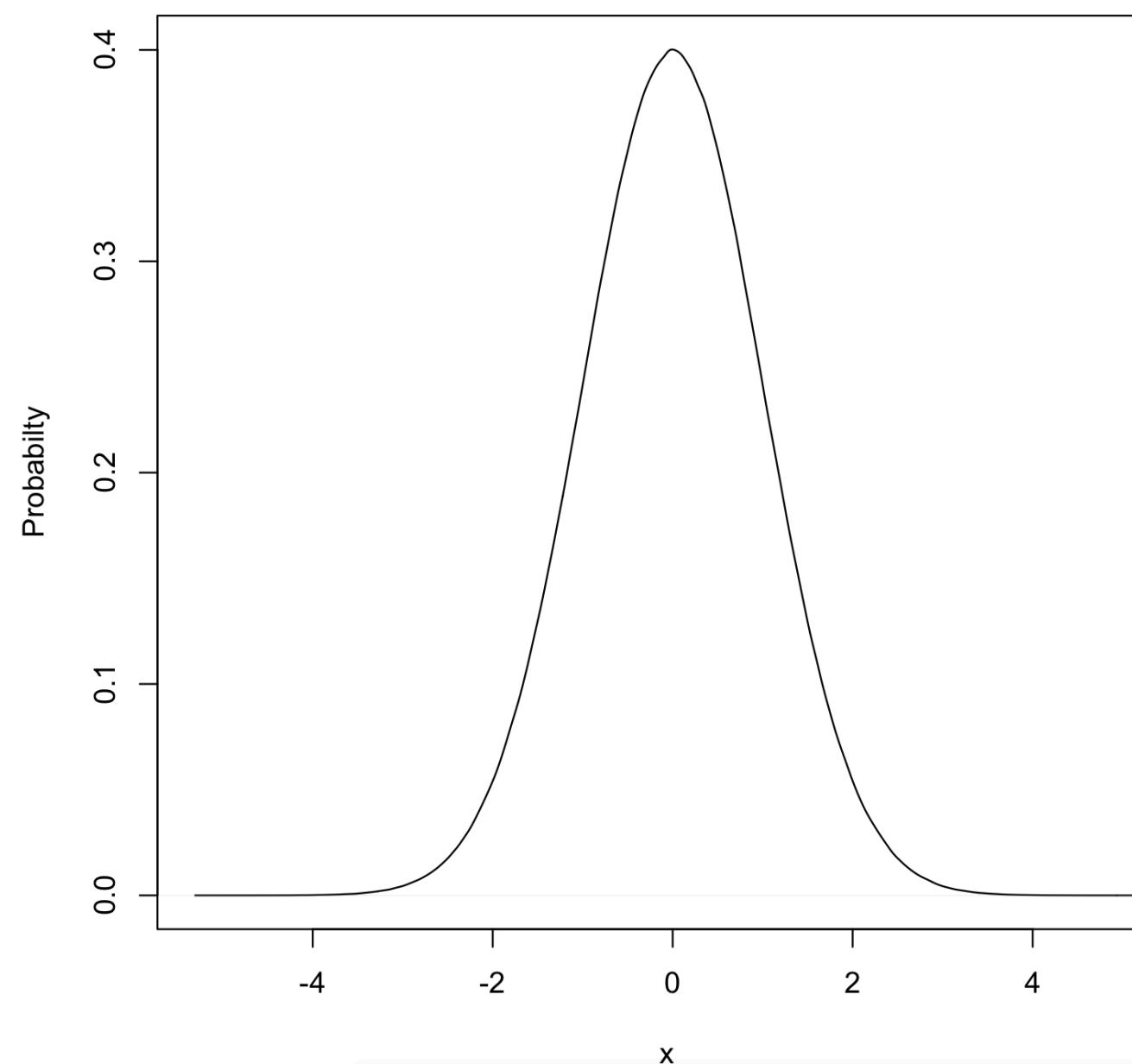
1



Left Skew (-0.963)

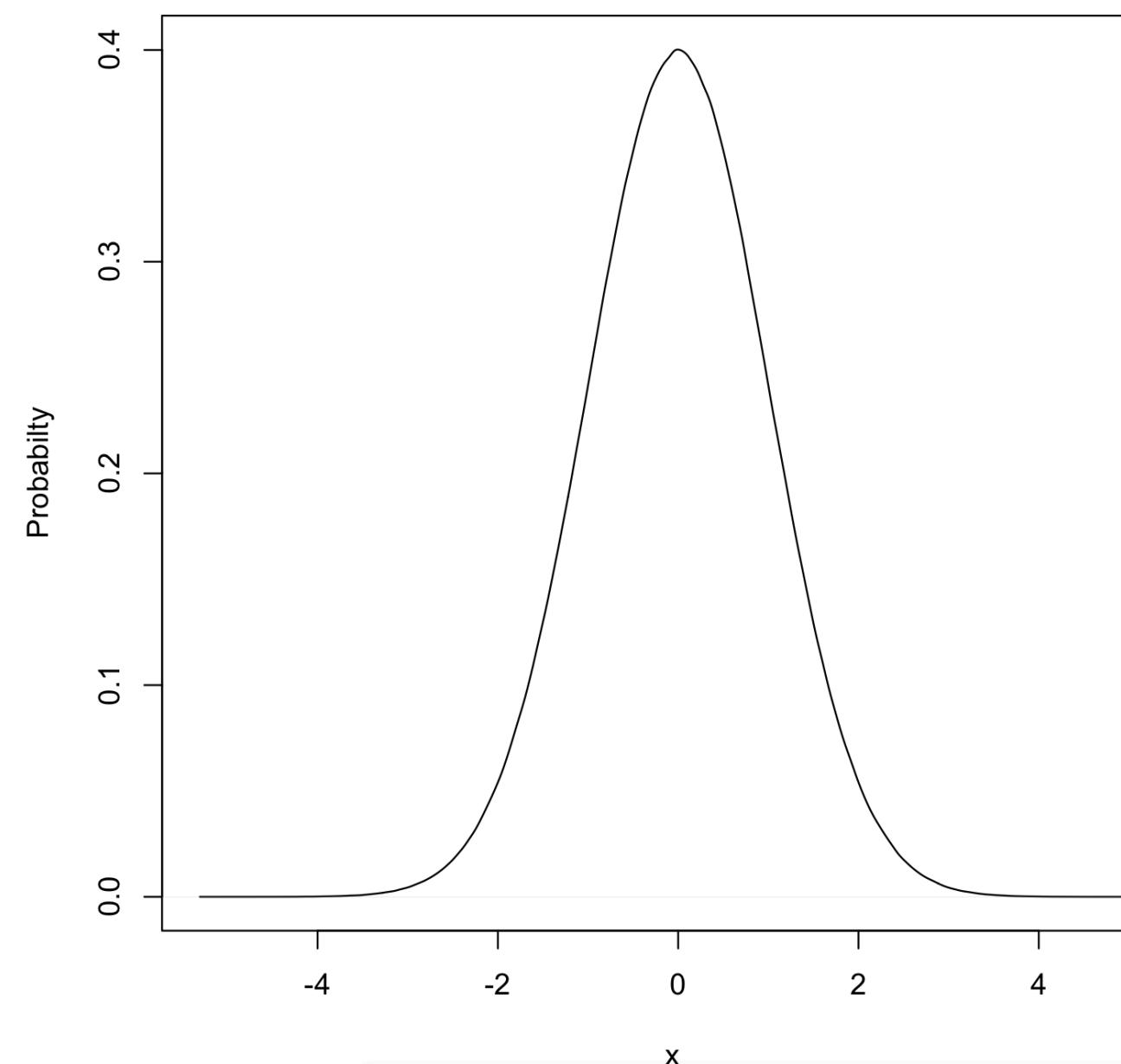
1

# Describing Distributions: Normality



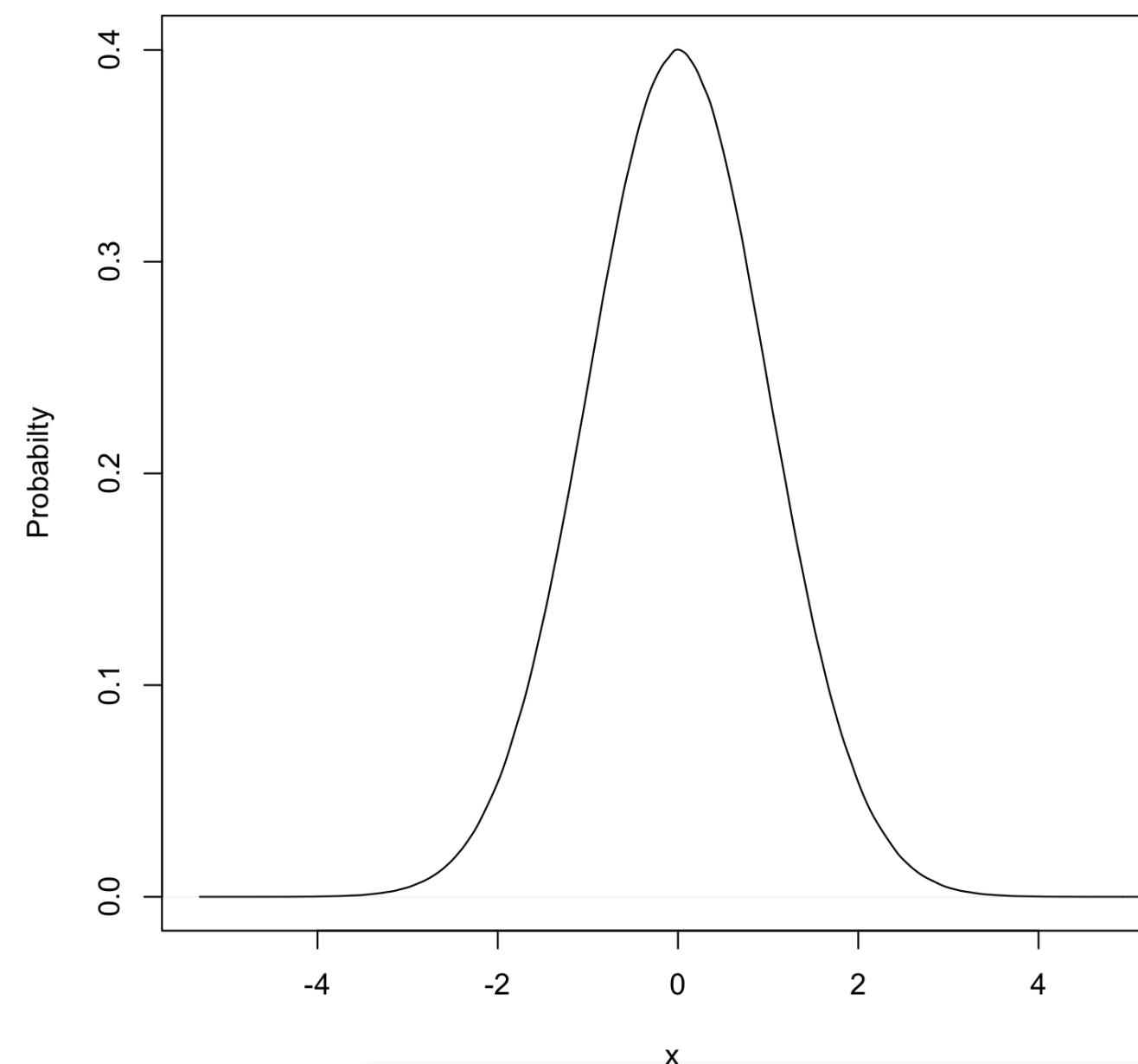
# Describing Distributions: Normality

- In most statistical applications, *normality* of data is essential



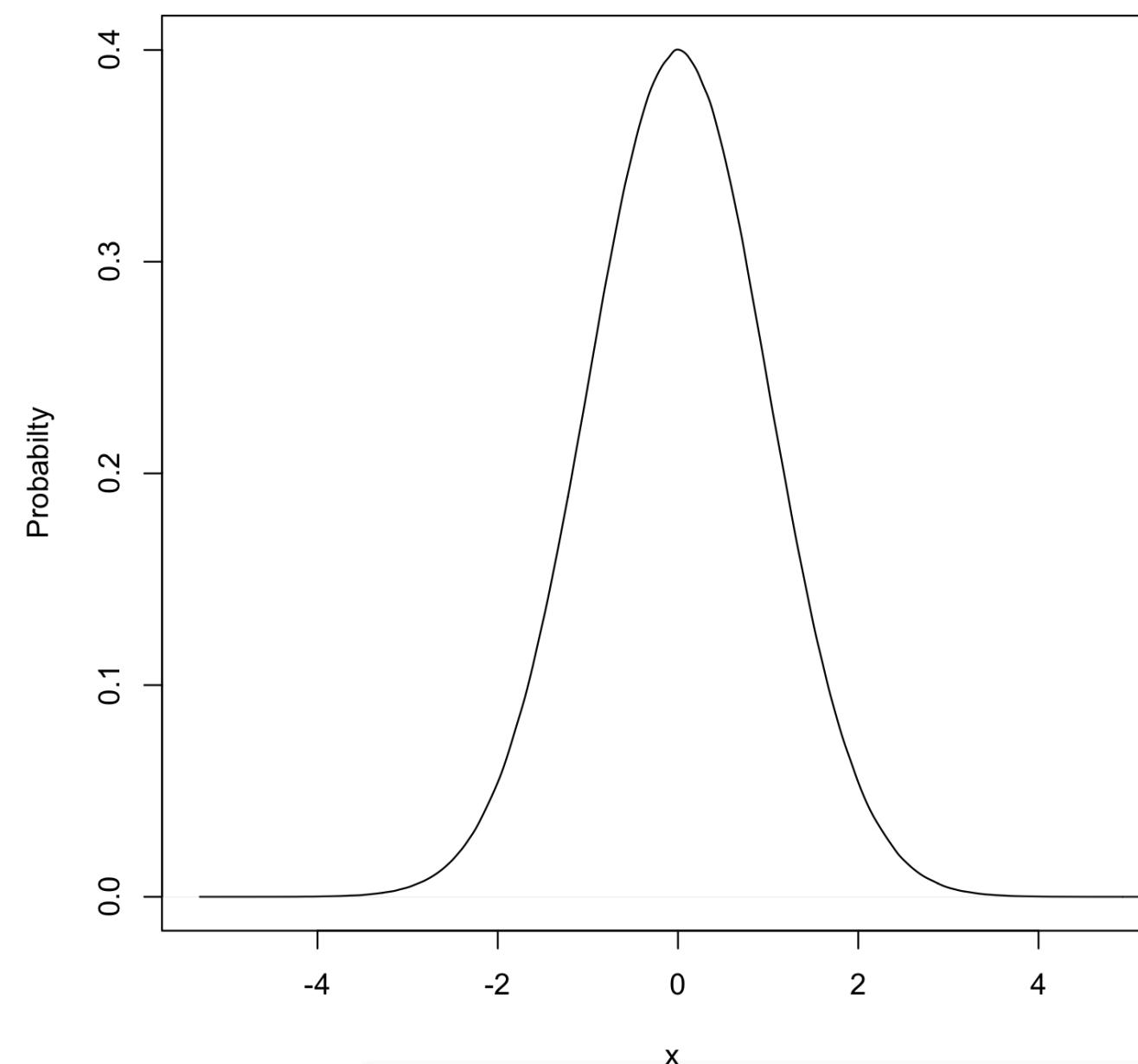
# Describing Distributions: Normality

- In most statistical applications, *normality* of data is essential
- In particular, in order to apply many common statistical procedures, data should follow a *normal (Gaussian) distribution*



# Describing Distributions: Normality

- In most statistical applications, *normality* of data is essential
- In particular, in order to apply many common statistical procedures, data should follow a *normal (Gaussian) distribution*
- To see if the data is approximately normal, we can use the *Empirical Rule*



# Empirical Rule: z-scores

# Empirical Rule: z-scores

- To discuss the Empirical Rule, we must first introduce the concept of z-scores

# Empirical Rule: z-scores

- To discuss the Empirical Rule, we must first introduce the concept of z-scores
- A **z-score** tells us how many standard deviations an observation is from its mean:

# Empirical Rule: z-scores

- To discuss the Empirical Rule, we must first introduce the concept of z-scores
- A **z-score** tells us how many standard deviations an observation is from its mean:

$$z = \frac{x - \bar{x}}{s}$$

! !

# Example: z-scores

# Example: z-scores

- The mean on Exam 1 is 86, and the standard deviation is 4

# Example: z-scores

- The mean on Exam 1 is 86, and the standard deviation is 4
- Student A scores 90 on the exam. What is their z-score?

$$z = \frac{90 - 86}{4} = 1$$

# Example: z-scores

- The mean on Exam 1 is 86, and the standard deviation is 4
- Student A scores 90 on the exam. What is their z-score?

$$\bullet \quad z = \frac{90 - 86}{4} = 1$$

# Example: z-scores

- The mean on Exam 1 is 86, and the standard deviation is 4
- Student A scores 90 on the exam. What is their z-score?

$$\bullet \quad z = \frac{90 - 86}{4} = 1$$

- Student B scores 78 on the exam. What is their z-score?

# Example: z-scores

- The mean on Exam 1 is 86, and the standard deviation is 4
- Student A scores 90 on the exam. What is their z-score?

$$\bullet \quad z = \frac{90 - 86}{4} = 1$$

- Student B scores 78 on the exam. What is their z-score?

$$\bullet \quad z = \frac{78 - 86}{2} = -2$$

# Example: z-scores

# Example: z-scores

- A student takes two tests. The mean on the first test is 86 with a standard deviation of 4. The mean on the second test is 400 with a standard deviation of 15. The student scored 91 on the first test and 425 on the second. Which test did the student score better on relative to the other test takers?

# Empirical Rule

# Empirical Rule

- If a distribution is symmetric, unimodal, and bell-shaped (i.e., normally distributed), then the following hold:

# Empirical Rule

- If a distribution is symmetric, unimodal, and bell-shaped (i.e., normally distributed), then the following hold:
- Approximately 68% of observations fall within one SD of the mean:  
 $\bar{x} \pm s$ , or  $(\bar{x} - s, \bar{x} + s)$

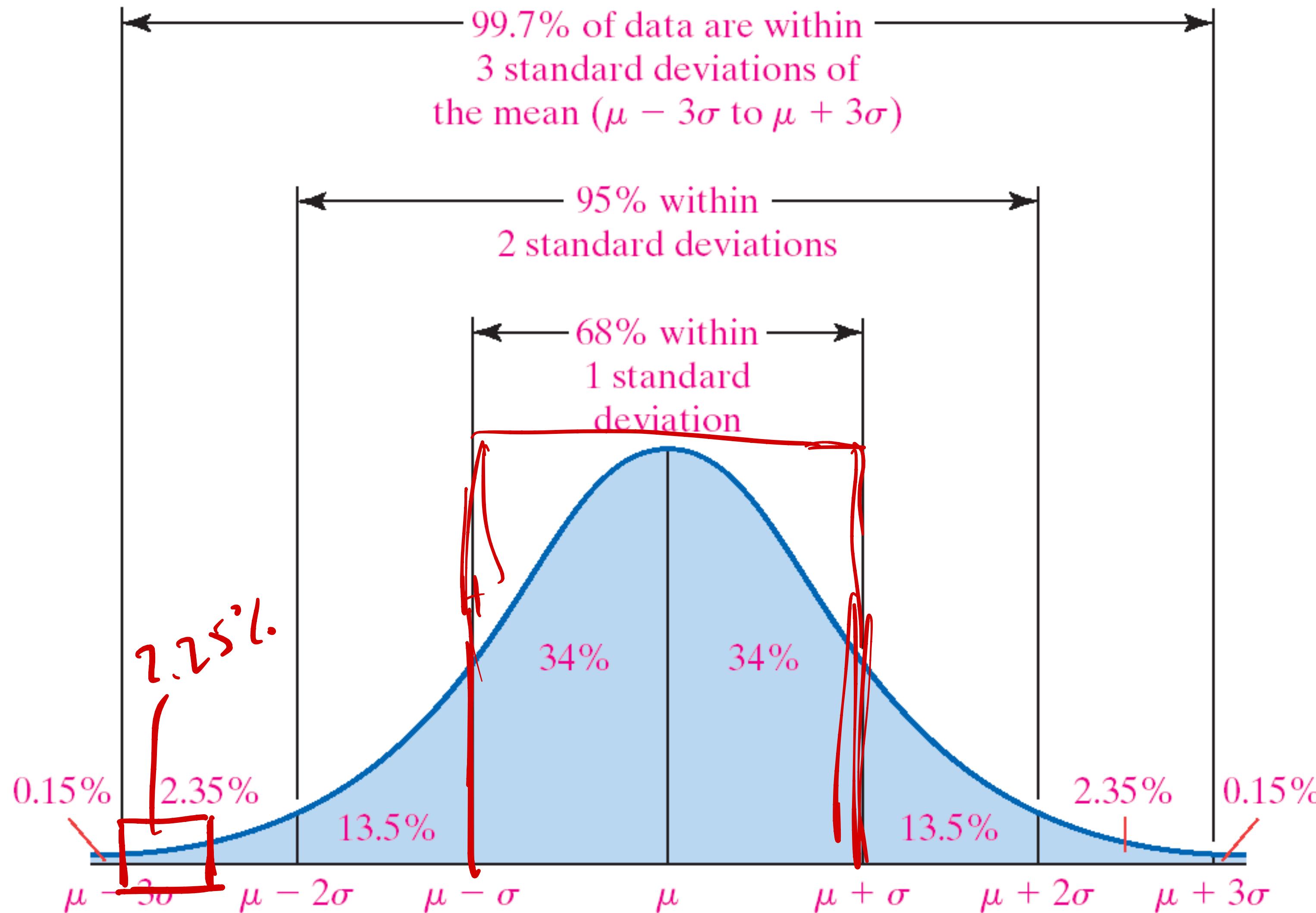
# Empirical Rule

- If a distribution is symmetric, unimodal, and bell-shaped (i.e., normally distributed), then the following hold:
- Approximately 68% of observations fall within one SD of the mean:  
 $\bar{x} \pm s$ , or  $(\bar{x} - s, \bar{x} + s)$
- Approximately 95% of observations fall within two SDs of the mean:  
 $\bar{x} \pm 2s$ , or  $(\bar{x} - 2s, \bar{x} + 2s)$

# Empirical Rule

- If a distribution is symmetric, unimodal, and bell-shaped (i.e., normally distributed), then the following hold:
  - Approximately 68% of observations fall within one SD of the mean:  
 $\bar{x} \pm s$ , or  $(\bar{x} - s, \bar{x} + s)$
  - Approximately 95% of observations fall within two SDs of the mean:  
 $\bar{x} \pm 2s$ , or  $(\bar{x} - 2s, \bar{x} + 2s)$
  - Approximately 99.7% of observations fall within three SDs of the mean:  
 $\bar{x} \pm 3s$ , or  $(\bar{x} - 3s, \bar{x} + 3s)$

# Empirical Rule



# Empirical Rule: Heart Rate Data

# Empirical Rule: Heart Rate Data

- Suppose heart rates have an approximately normal distribution with  $\bar{x} = 71.2$  and  $s = 9.15$

# Empirical Rule: Heart Rate Data

- Suppose heart rates have an approximately normal distribution with  $\bar{x} = 71.2$  and  $s = 9.15$
- Approximately 68% of observations fall in the interval  $71.2 \pm 9.15 = (62.05, 80.35)$

# Empirical Rule: Heart Rate Data

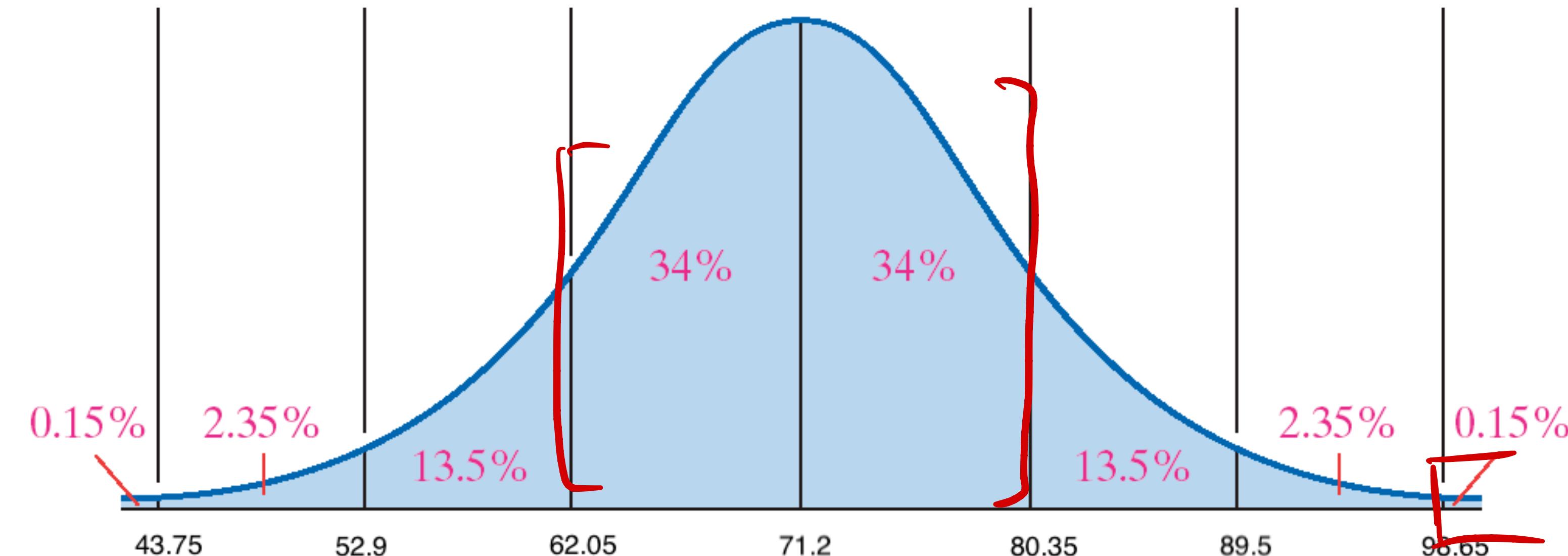
- Suppose heart rates have an approximately normal distribution with  $\bar{x} = 71.2$  and  $s = 9.15$
- Approximately 68% of observations fall in the interval  $71.2 \pm 9.15 = (62.05, 80.35)$
- Approximately 95% of observations fall in the interval  $71.2 \pm 2 \times 9.15 = (52.9, 89.5)$

# Empirical Rule: Heart Rate Data

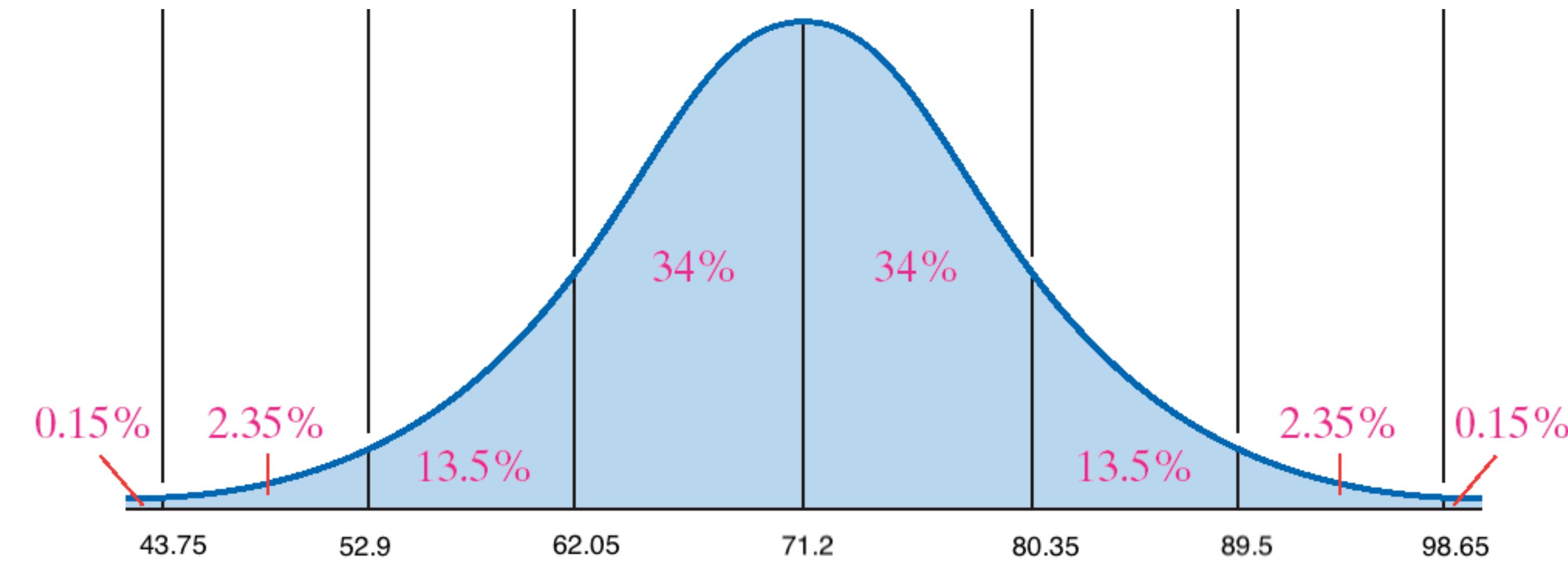
- Suppose heart rates have an approximately normal distribution with  $\bar{x} = 71.2$  and  $s = 9.15$
- Approximately 68% of observations fall in the interval  $71.2 \pm 9.15 = (62.05, 80.35)$
- Approximately 95% of observations fall in the interval  $71.2 \pm 2 \times 9.15 = (52.9, 89.5)$
- Approximately 99.7% of observations fall in the interval  $71.2 \pm 3 \times 9.15 = (43.75, 98.65)$

# Empirical Rule: Heart Rate Data

- Suppose heart rates have an approximately normal distribution with  $\bar{x} = 71.2$  and  $s = 9.15$
- Approximately 68% of observations fall in the interval  $71.2 \pm 9.15 = (62.05, 80.35)$
- Approximately 95% of observations fall in the interval  $71.2 \pm 2 \times 9.15 = (52.9, 89.5)$
- Approximately 99.7% of observations fall in the interval  $71.2 \pm 3 \times 9.15 = (43.75, 98.65)$

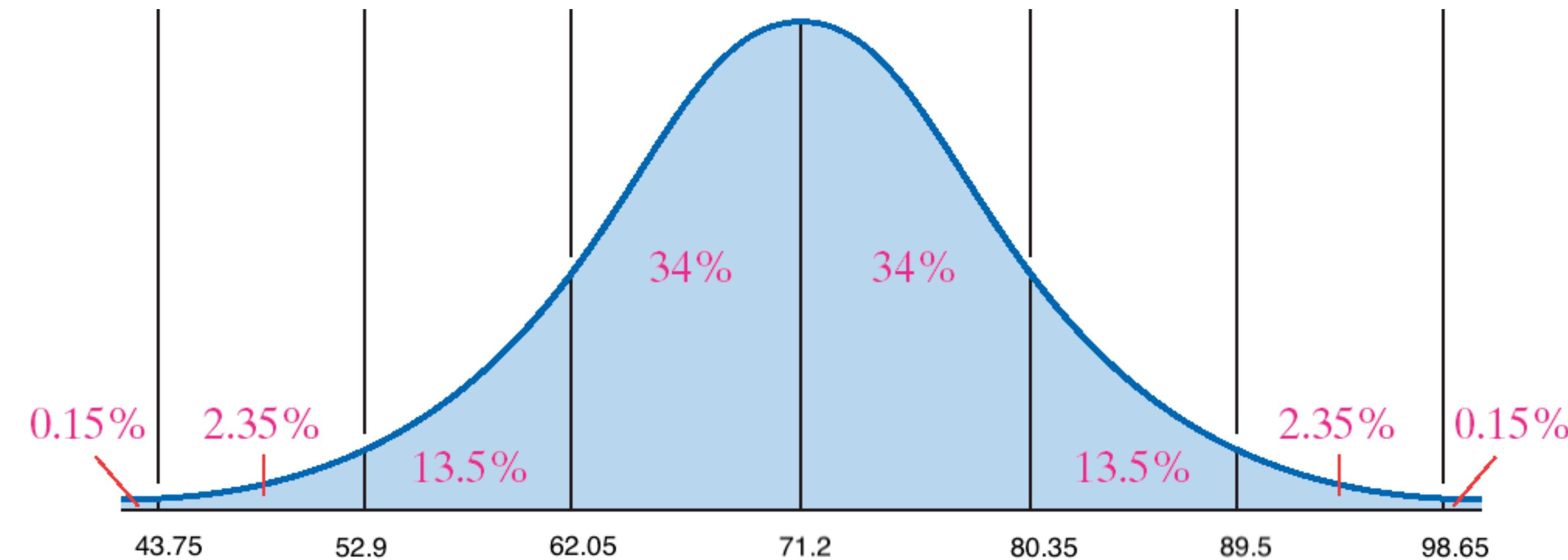


# Empirical Rule: Heart Rate Data



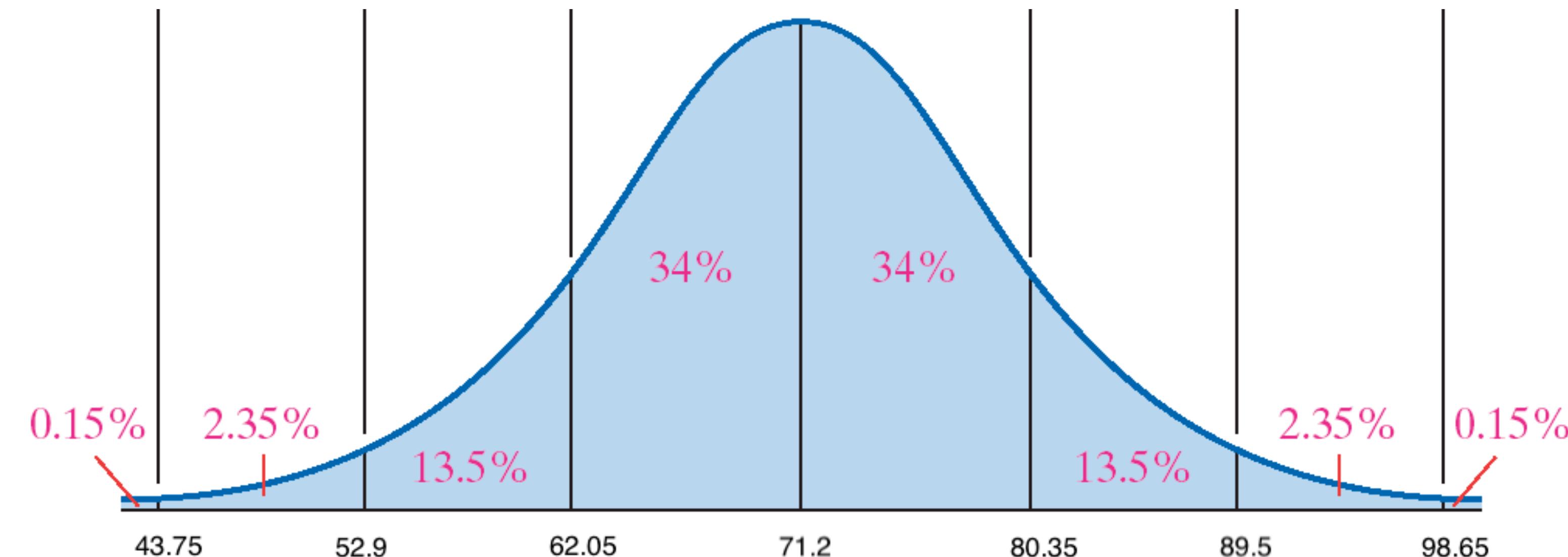
# Empirical Rule: Heart Rate Data

- How fast does a heart rate have to be in order to be at the 97.5<sup>th</sup> percentile?



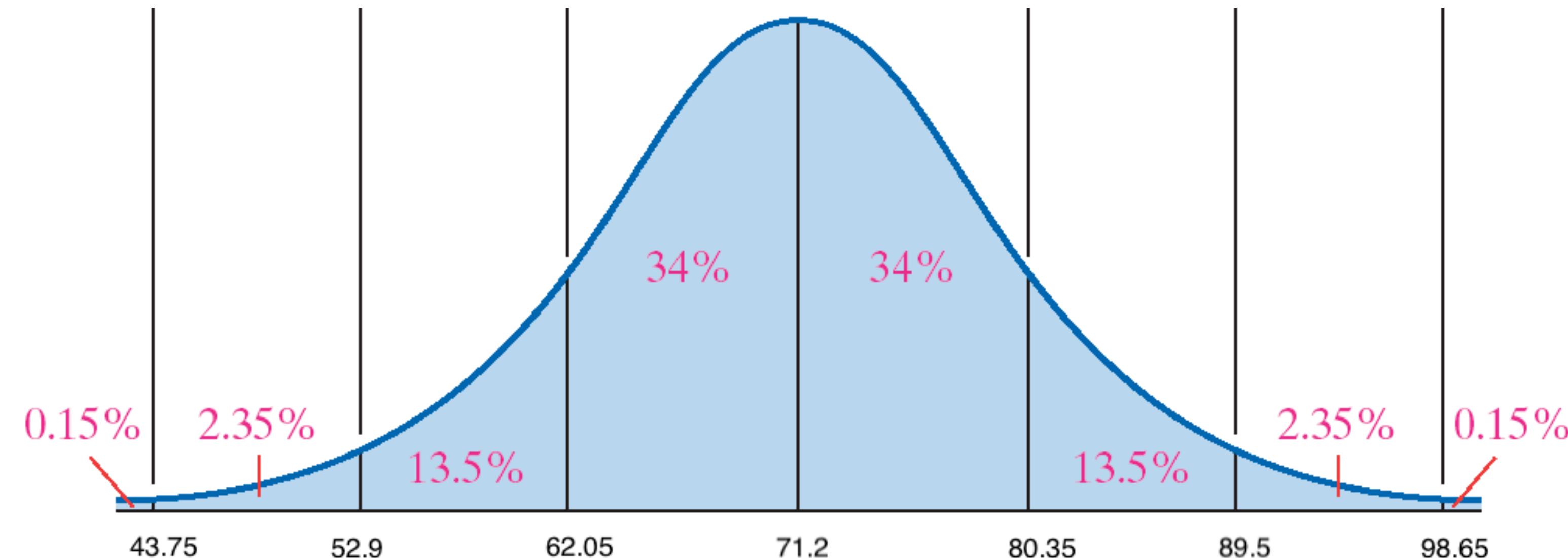
# Empirical Rule: Heart Rate Data

- How fast does a heart rate have to be in order to be at the 97.5<sup>th</sup> percentile?
- Approximately what percentage of heart rates are greater than 80.35?



# Empirical Rule: Heart Rate Data

- How fast does a heart rate have to be in order to be at the 97.5<sup>th</sup> percentile?
- Approximately what percentage of heart rates are greater than 80.35?
- Approximately what percentage of heart rates are between 52.9 and 62.05?



# Chebyshev's Inequality

# Chebyshev's Inequality

- The empirical rule works for symmetric, unimodal, and bell-shaped data

# Chebyshev's Inequality

- The empirical rule works for symmetric, unimodal, and bell-shaped data
- If the data is *not* symmetric and unimodal, we may instead use *Chebyshev's Inequality* to summarize the distribution (holds for *any distribution*)

# Chebyshev's Inequality

- The empirical rule works for symmetric, unimodal, and bell-shaped data
- If the data is *not* symmetric and unimodal, we may instead use *Chebyshev's Inequality* to summarize the distribution (holds for *any distribution*)
- **Chebyshev's Inequality** (informal): At least  $1 - \left(\frac{1}{k}\right)^2$  of observations in a set of data lie within  $k$  standard deviations of the mean

# Chebyshev's Inequality

- The empirical rule works for symmetric, unimodal, and bell-shaped data
- If the data is *not* symmetric and unimodal, we may instead use *Chebyshev's Inequality* to summarize the distribution (holds for *any distribution*)
- **Chebyshev's Inequality** (informal): At least  $1 - \left(\frac{1}{k}\right)^2$  of observations in a set of data lie within  $k$  standard deviations of the mean
- Intuitively: “No more than a certain fraction of values be too far from the mean”

# Chebyshev's Inequality

- The empirical rule works for symmetric, unimodal, and bell-shaped data
- If the data is *not* symmetric and unimodal, we may instead use *Chebyshev's Inequality* to summarize the distribution (holds for *any distribution*)
- **Chebyshev's Inequality** (informal): At least  $1 - \left(\frac{1}{k}\right)^2$  of observations in a set of data lie within  $k$  standard deviations of the mean
- Intuitively: “No more than a certain fraction of values be too far from the mean”
- Weaker than the Empirical Rule because it makes fewer assumptions

# Quantile Plots

# Quantile Plots

- We may extend the intuition of the empirical rule to a more sophisticated method for examining whether data is in fact normally distributed

# Quantile Plots

- We may extend the intuition of the empirical rule to a more sophisticated method for examining whether data is in fact normally distributed
- Suppose we have data  $x_1, x_2, \dots, x_n$  and we want to see if they follow a normal distribution

# Quantile Plots

- We may extend the intuition of the empirical rule to a more sophisticated method for examining whether data is in fact normally distributed
- Suppose we have data  $x_1, x_2, \dots, x_n$  and we want to see if they follow a normal distribution
- We can compare the sample quantiles based on the observed data to the theoretical quantiles of the normal distribution



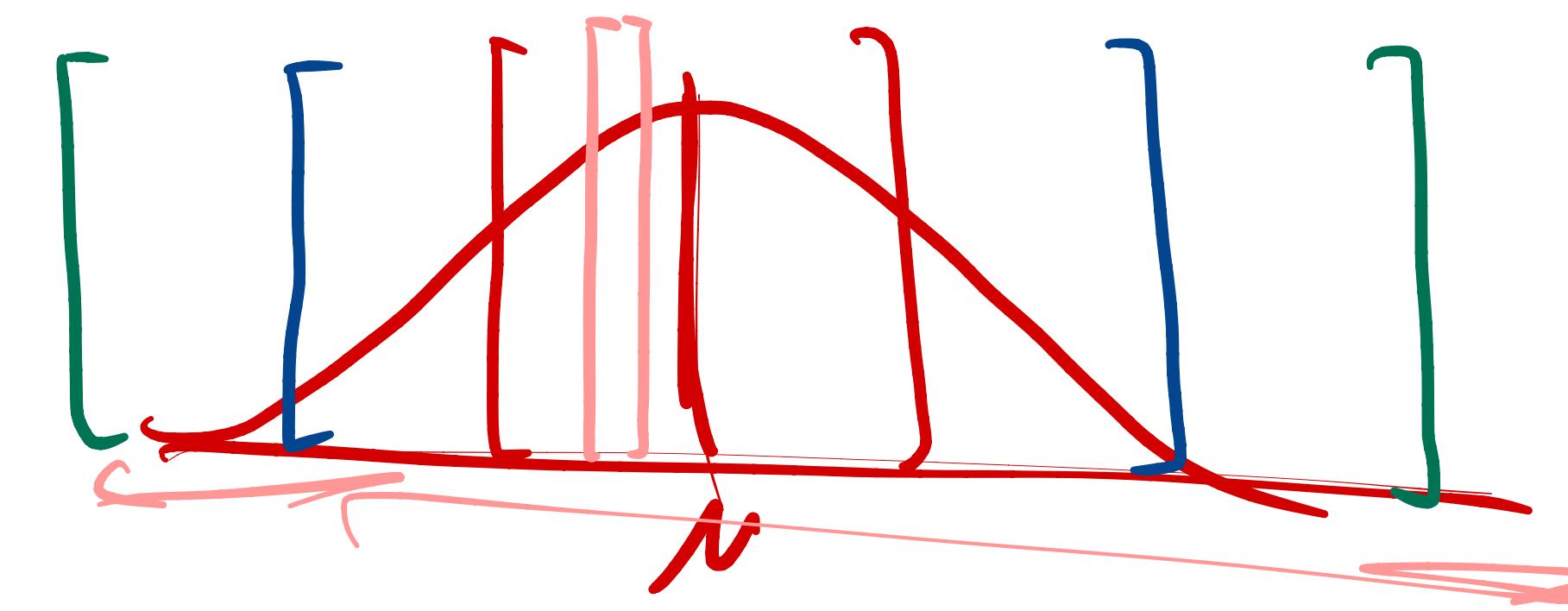
# Quantile Plots

- We may extend the intuition of the empirical rule to a more sophisticated method for examining whether data is in fact normally distributed
- Suppose we have data  $x_1, x_2, \dots, x_n$  and we want to see if they follow a normal distribution
- We can compare the sample quantiles based on the observed data to the theoretical quantiles of the normal distribution
- A *quantile plot* plots the sample quantiles (y-axis) against the theoretical quantiles (x-axis)

# Quantile Plots

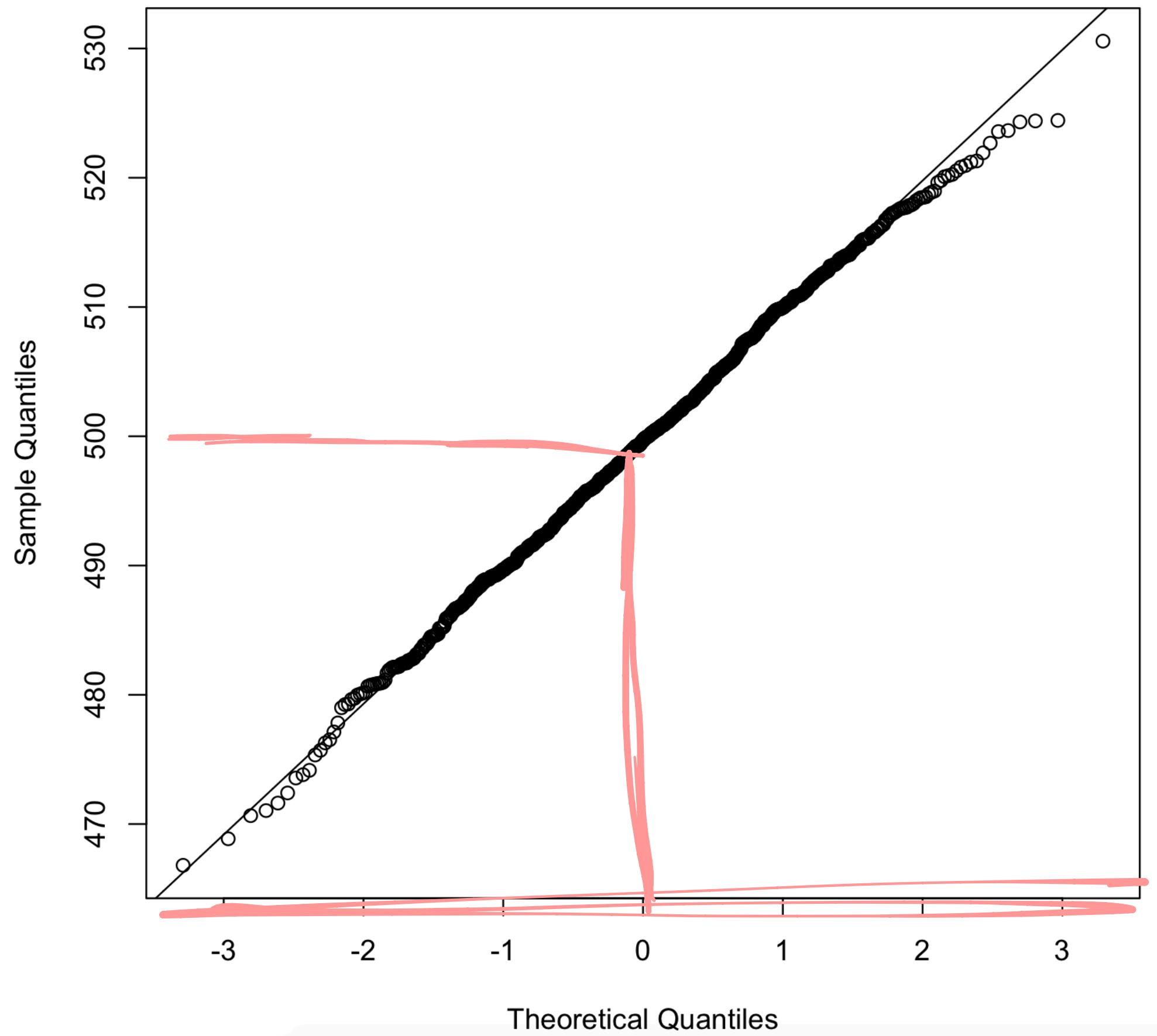
- We may extend the intuition of the empirical rule to a more sophisticated method for examining whether data is in fact normally distributed
- Suppose we have data  $x_1, x_2, \dots, x_n$  and we want to see if they follow a normal distribution
- We can compare the sample quantiles based on the observed data to the theoretical quantiles of the normal distribution
- A *quantile plot* plots the sample quantiles (y-axis) against the theoretical quantiles (x-axis)
- If the points lie approximately on the line  $y = x$ , then the data is approximately normally distributed

# Quantile Plots

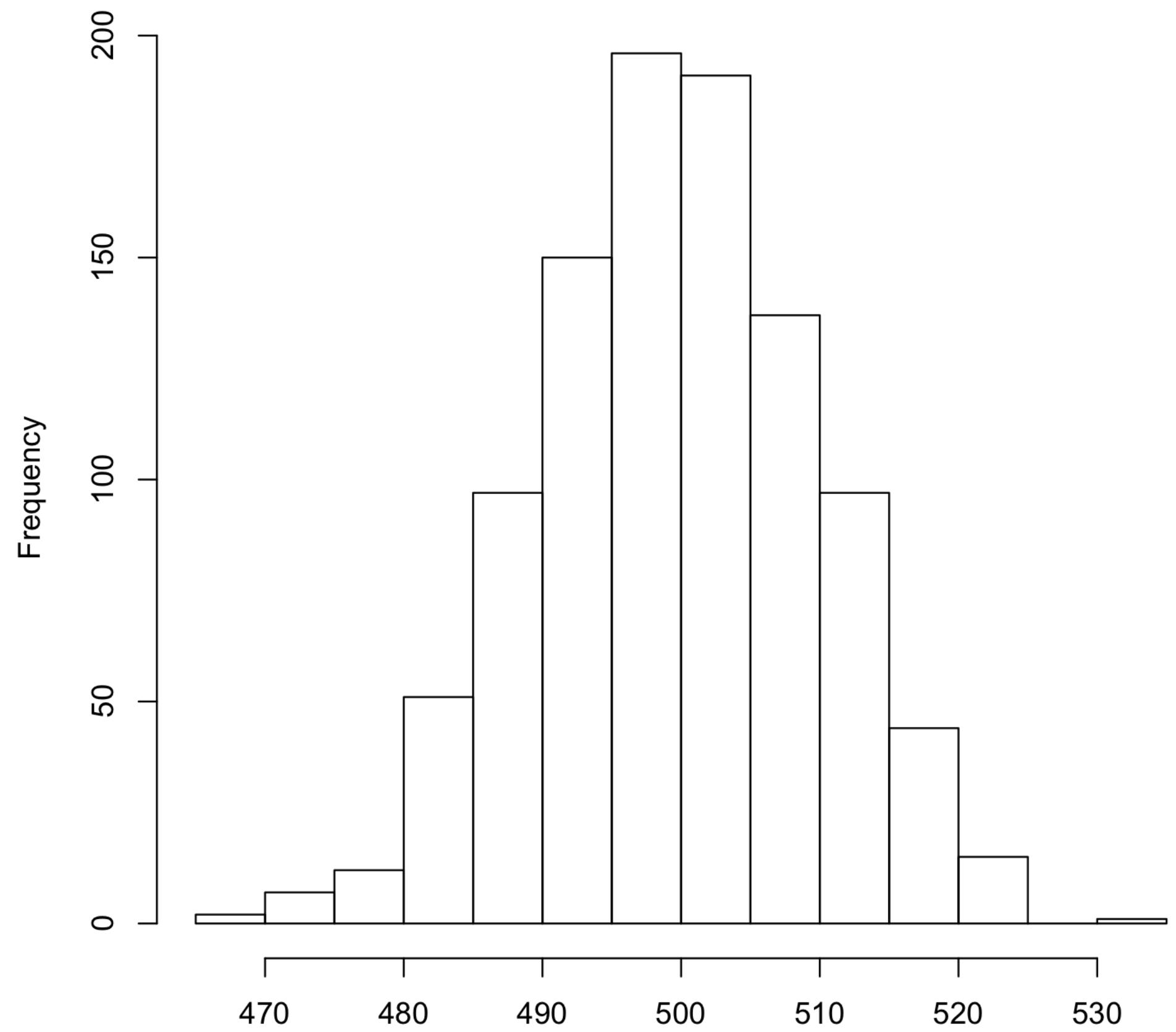
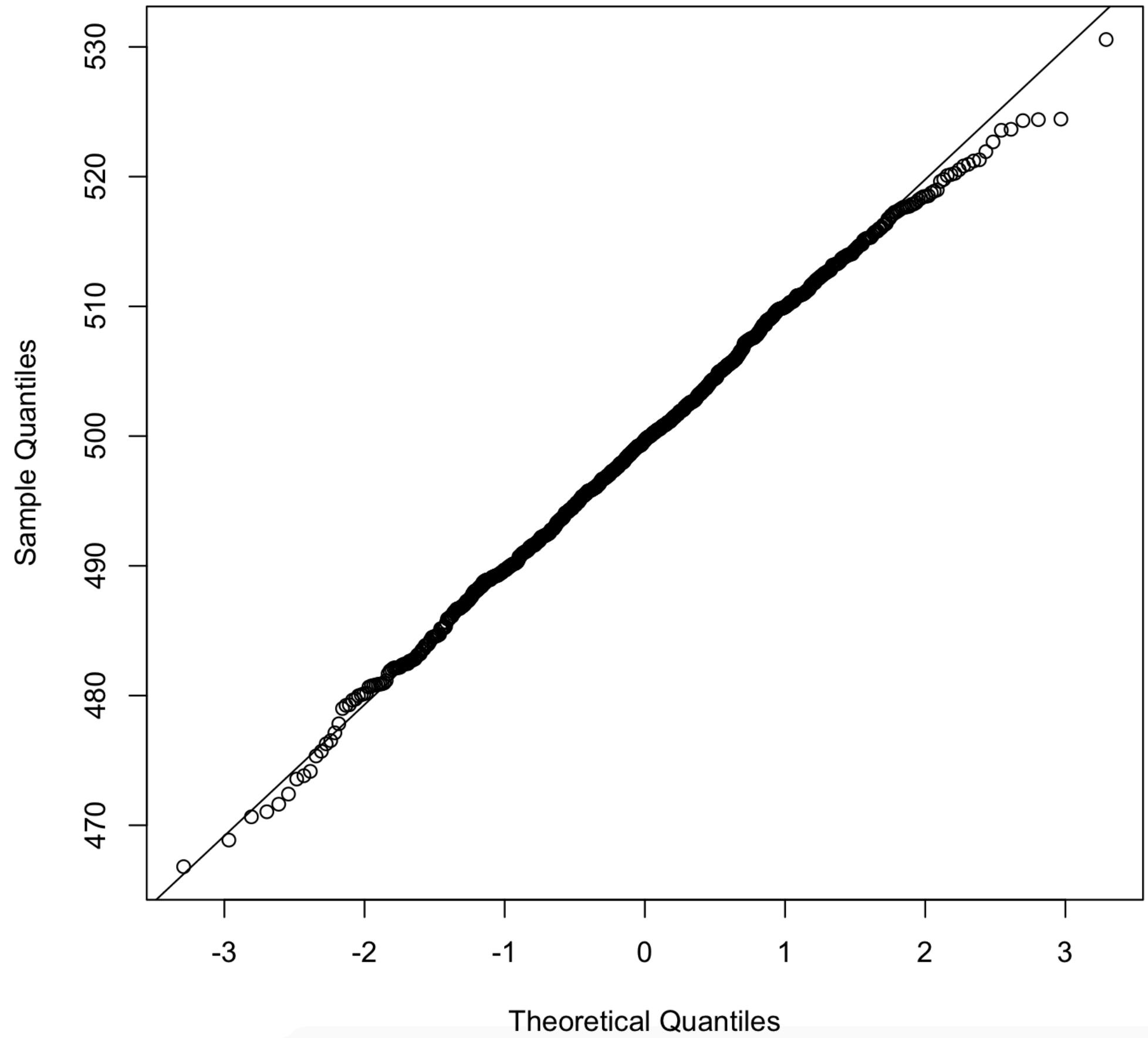


- We may extend the intuition of the empirical rule to a more sophisticated method for examining whether data is in fact normally distributed
- Suppose we have data  $x_1, x_2, \dots, x_n$  and we want to see if they follow a normal distribution
- We can compare the sample quantiles based on the observed data to the theoretical quantiles of the normal distribution
- A *quantile plot* plots the sample quantiles (y-axis) against the theoretical quantiles (x-axis)
- If the points lie approximately on the line  $y = x$ , then the data is approximately normally distributed
- R code: `qqnorm(data); qqline(data)`

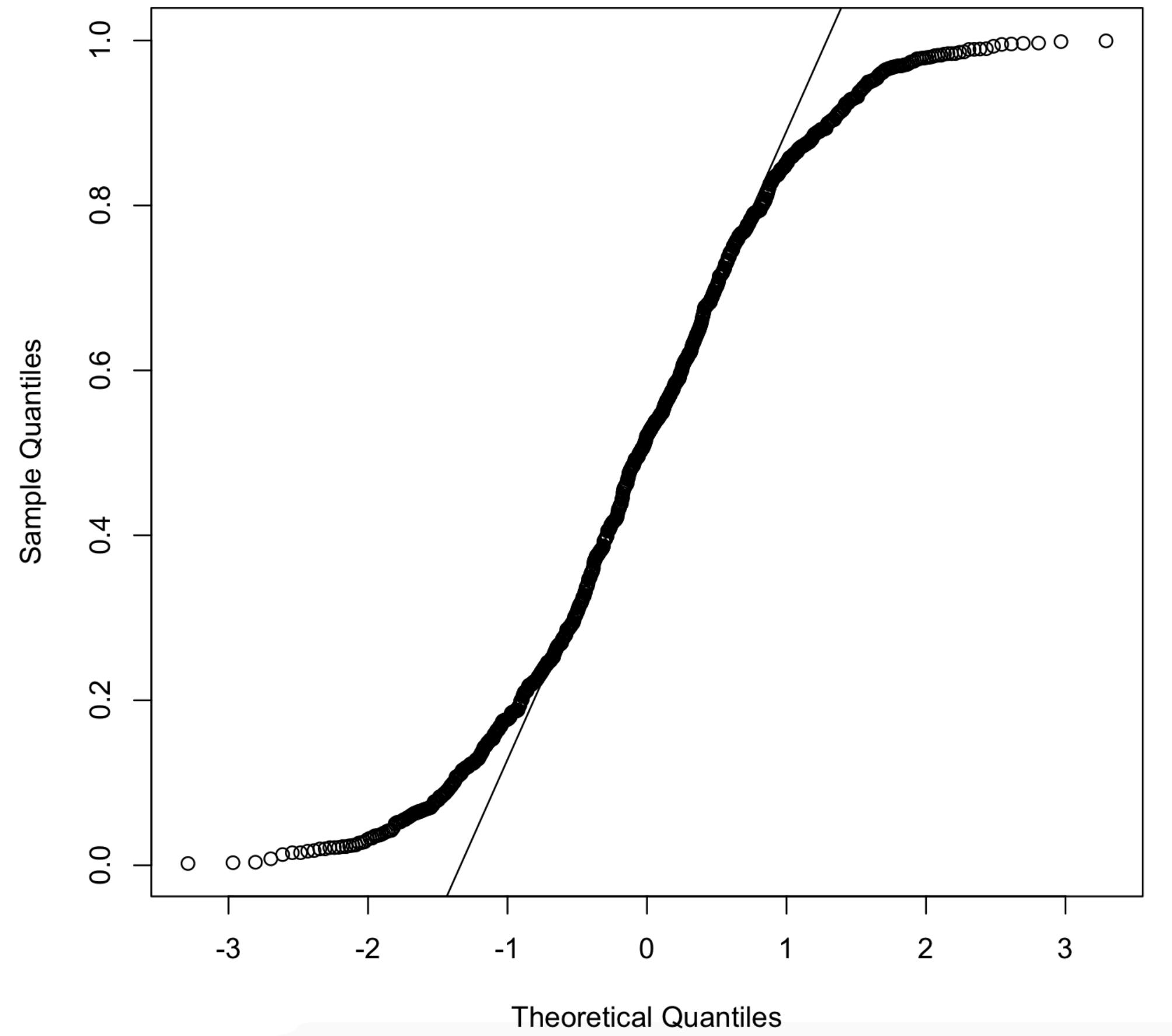
# Quantile Plots



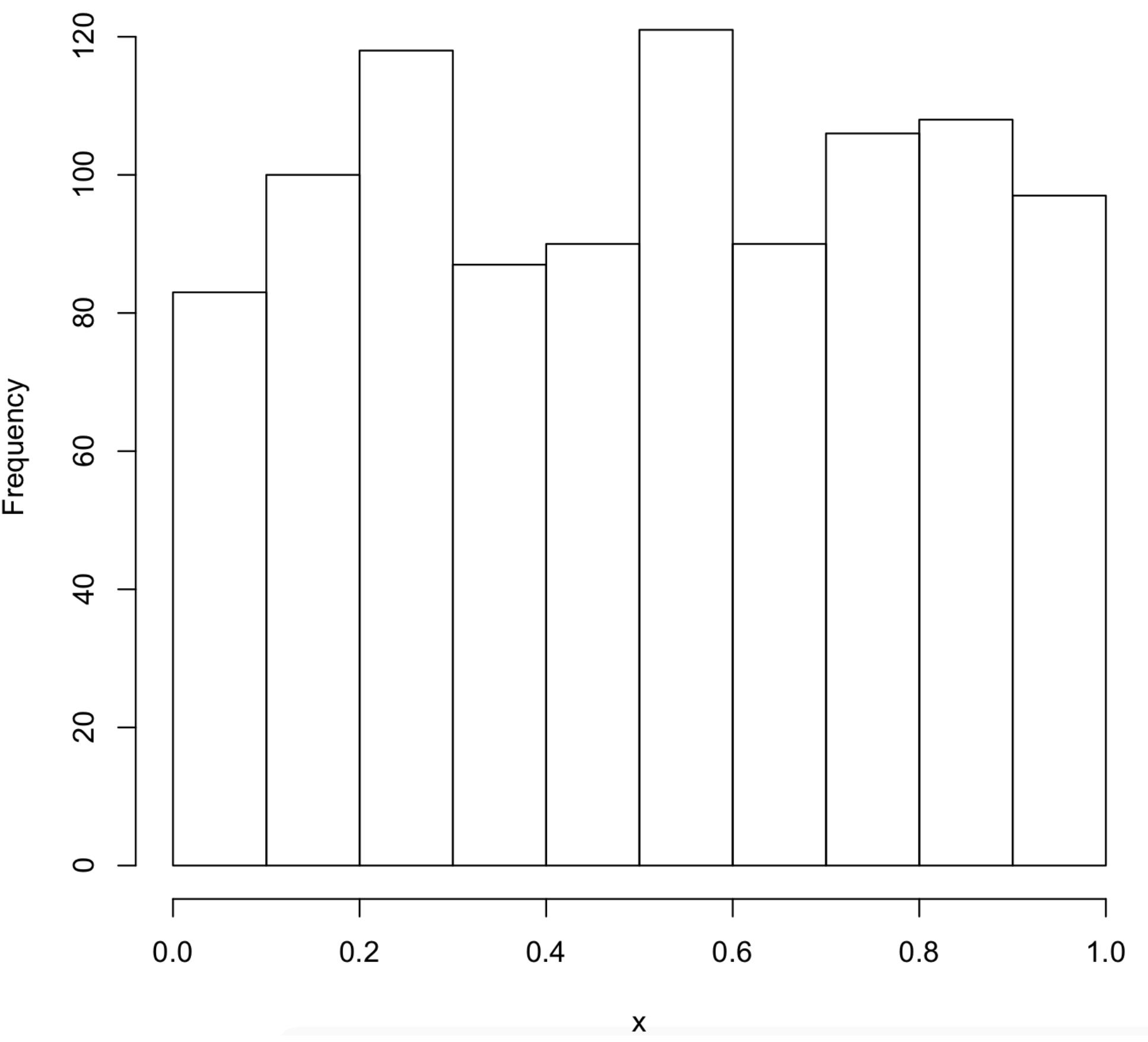
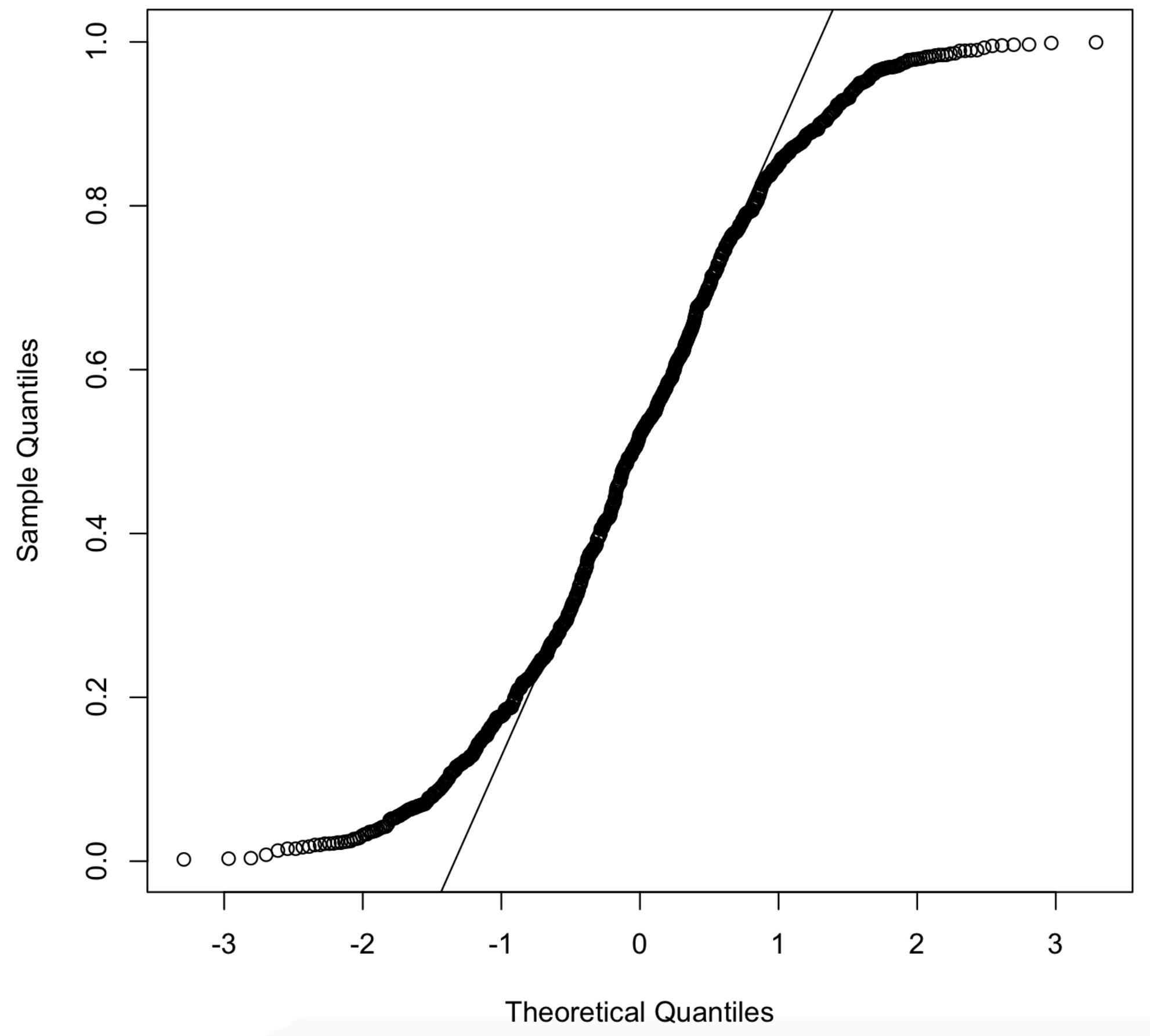
# Quantile Plots



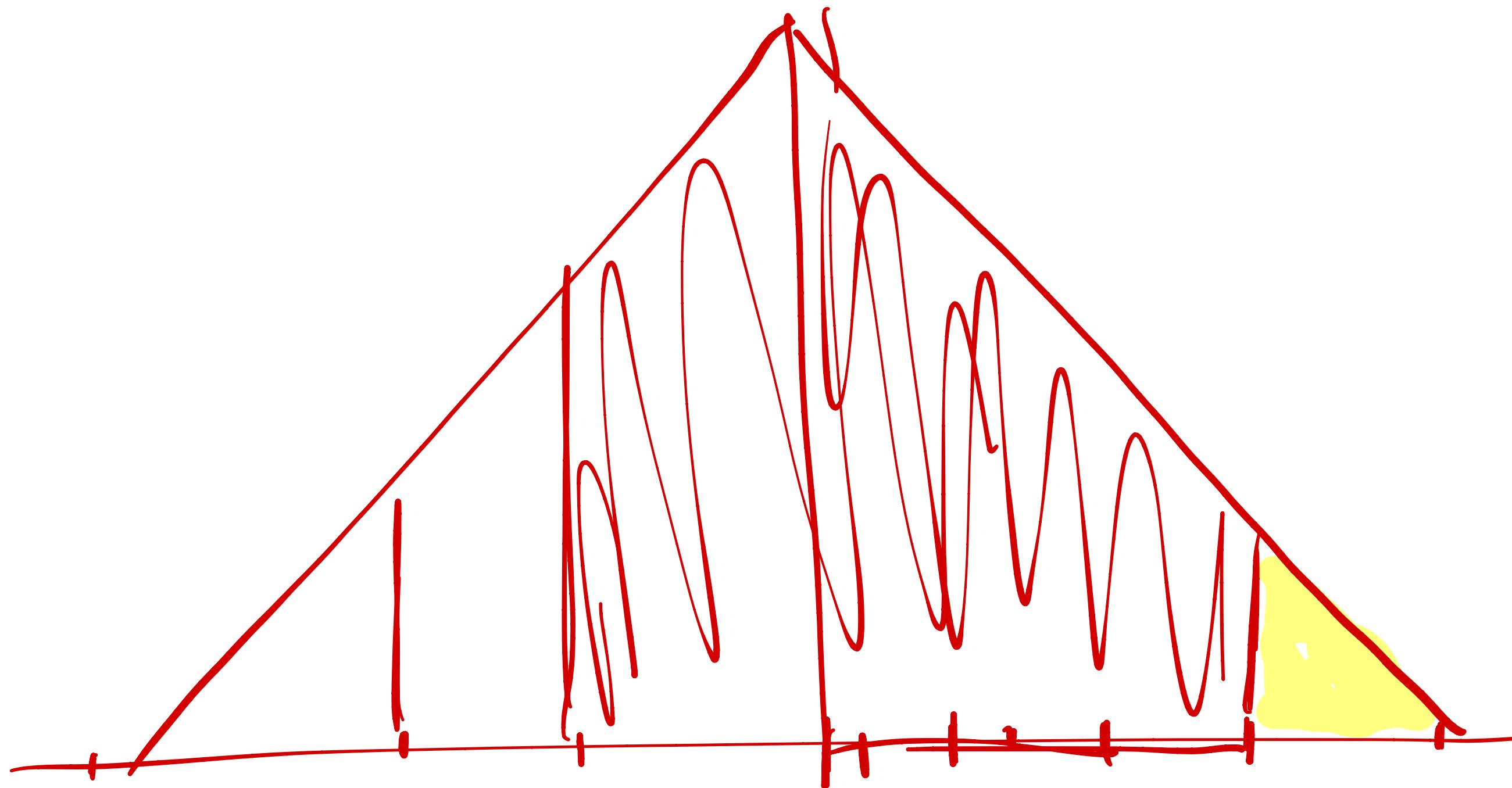
# Quantile Plots



# Quantile Plots



# Transformations



# Transformations

- In some cases, we may need to transform our data to change the measurement units or to make an analytical method simpler, more accurate, or more effective

# Transformations

- In some cases, we may need to transform our data to change the measurement units or to make an analytical method simpler, more accurate, or more effective
- Often, when doing a transformation, we attempt to take skewed data and transform it onto a scale in which the data is then approximately normally distributed

# Linear Transformations

# Linear Transformations

- *Linear transformations* are typically used to change units

# Linear Transformations

- *Linear transformations* are typically used to change units
- A linear transformation is of the form  $Y = aX + b$

# Linear Transformations

- *Linear transformations* are typically used to change units
- A linear transformation is of the form  $Y = aX + b$
- $X$  is the original data,  $Y$  is the transformed data, and  $a$  and  $b$  are the transforming constants

# Linear Transformations

- *Linear transformations* are typically used to change units
- A linear transformation is of the form  $Y = aX + b$
- $X$  is the original data,  $Y$  is the transformed data, and  $a$  and  $b$  are the transforming constants
- Example: When converting Celsius to Fahrenheit, we have  $F = 1.8C + 32$  with  $a = 1.8$  and  $b = 32$

# Linear Transformations

# Linear Transformations

- When doing linear transformations, many summary statistics do not need to be explicitly recalculated

# Linear Transformations

- When doing linear transformations, many summary statistics do not need to be explicitly recalculated
- Instead, we can transform each summary statistic to represent the transformed data

# Linear Transformations

- When doing linear transformations, many summary statistics do not need to be explicitly recalculated
- Instead, we can transform each summary statistic to represent the transformed data

Statistic	Original	Transformed
Mean	$\bar{x}$	$a\bar{x} + b$
Median	$\tilde{x}$	$a\tilde{x} + b$
Trimmed Mean	$\bar{x}_{K\%}$	$a\bar{x}_{K\%} + b$
Variance	$s^2$	$a^2 s^2$
Standard Deviation	$s$	$ a s$
Interquartile Range	IQR	$ a IQR$
Lower Quartile	$Q_1$	$aQ_1 + b$ if $a > 0$ $aQ_3 + b$ if $a < 0$
Upper Quartile	$Q_2$	$aQ_3 + b$ if $a > 0$ $aQ_1 + b$ if $a < 0$
Quantile	$x_{K\%}$	$ax_{K\%} + b$ if $a > 0$ $ax_{(100-K)\%} + b$ if $a < 0$

$y = ax + b$

# Transformations

# Transformations

- Linear transformations do little to change the overall shape of data

# Transformations

- Linear transformations do little to change the overall shape of data
- Other transformations can be used to reduce skewness and thus affect overall shape

# Transformations

- Linear transformations do little to change the overall shape of data
- Other transformations can be used to reduce skewness and thus affect overall shape
- Often, we want to transform data so they are approximately normally distributed in order to apply standard statistical procedures

# Transformations

- Linear transformations do little to change the overall shape of data
- Other transformations can be used to reduce skewness and thus affect overall shape
- Often, we want to transform data so they are approximately normally distributed in order to apply standard statistical procedures
- Many types of data are not naturally symmetric (e.g., income, age, survival time)

# Transformations

- Linear transformations do little to change the overall shape of data
- Other transformations can be used to reduce skewness and thus affect overall shape
- Often, we want to transform data so they are approximately normally distributed in order to apply standard statistical procedures
- Many types of data are not naturally symmetric (e.g., income, age, survival time)
- Transform data to meet assumptions needed for analyses

# Transformations

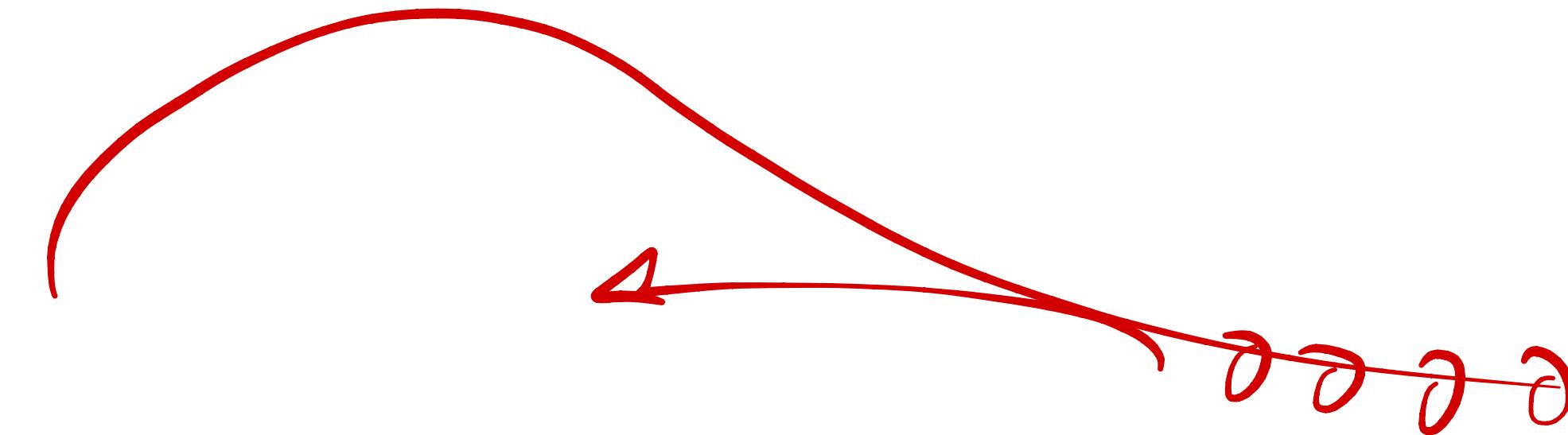
# Transformations

- Find an appropriate function  $f(x)$  that will transform original data  $x_1, x_2, \dots, x_n$  into  $y_1, y_2, \dots, y_n$  through the transformation  $y_i = f(x_i)$

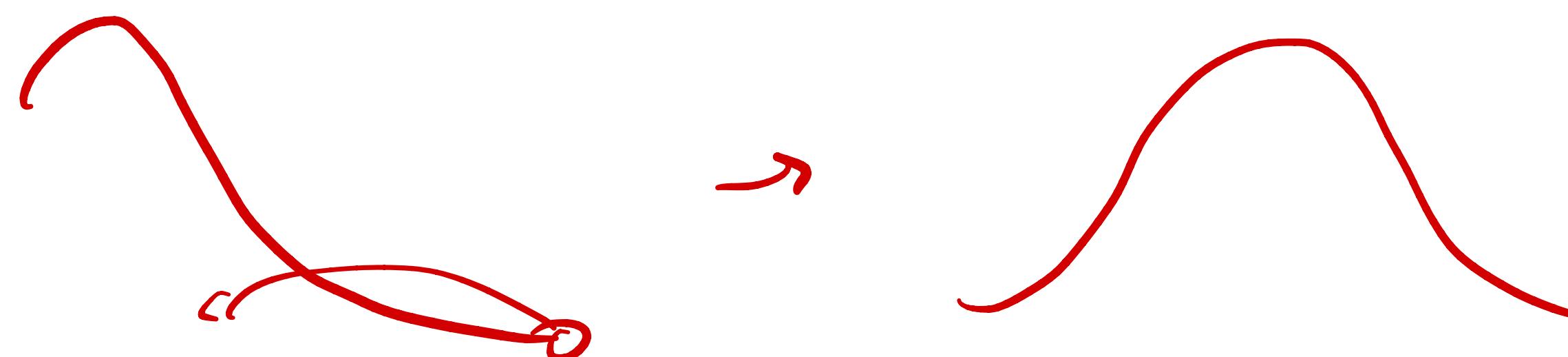
# Transformations

- Find an appropriate function  $f(x)$  that will transform original data  $x_1, x_2, \dots, x_n$  into  $y_1, y_2, \dots, y_n$  through the transformation  $y_i = f(x_i)$
- Often,  $f(x)$  must be an increasing function so that if  $x_i > x_j$ , then  $y_i > y_j$

# Transformations



- Find an appropriate function  $f(x)$  that will transform original data  $x_1, x_2, \dots, x_n$  into  $y_1, y_2, \dots, y_n$  through the transformation  $y_i = f(x_i)$
- Often,  $f(x)$  must be an increasing function so that if  $x_i > x_j$ , then  $y_i > y_j$
- For right skewed data, use a function that tends to reduce larger values in proportion to smaller ones (i.e., an increasing function whose slope is decreasing)



# Transformations

- Find an appropriate function  $f(x)$  that will transform original data  $x_1, x_2, \dots, x_n$  into  $y_1, y_2, \dots, y_n$  through the transformation  $y_i = f(x_i)$
- Often,  $f(x)$  must be an increasing function so that if  $x_i > x_j$ , then  $y_i > y_j$
- For right skewed data, use a function that tends to reduce larger values in proportion to smaller ones (i.e., an increasing function whose slope is decreasing)
  - $f(x) = \log(x)$

# Transformations

- Find an appropriate function  $f(x)$  that will transform original data  $x_1, x_2, \dots, x_n$  into  $y_1, y_2, \dots, y_n$  through the transformation  $y_i = f(x_i)$
- Often,  $f(x)$  must be an increasing function so that if  $x_i > x_j$ , then  $y_i > y_j$
- For right skewed data, use a function that tends to reduce larger values in proportion to smaller ones (i.e., an increasing function whose slope is decreasing)
  - $f(x) = \log(x)$
  - $f(x) = \sqrt{x}$

# Transformations

# Transformations

- General Social Survey (GSS) collects data annually on a sample of Americans regarding demographic and lifestyle characteristics

# Transformations

- General Social Survey (GSS) collects data annually on a sample of Americans regarding demographic and lifestyle characteristics
- Consider the 2006 survey, which collected data on the age at which respondents were first married

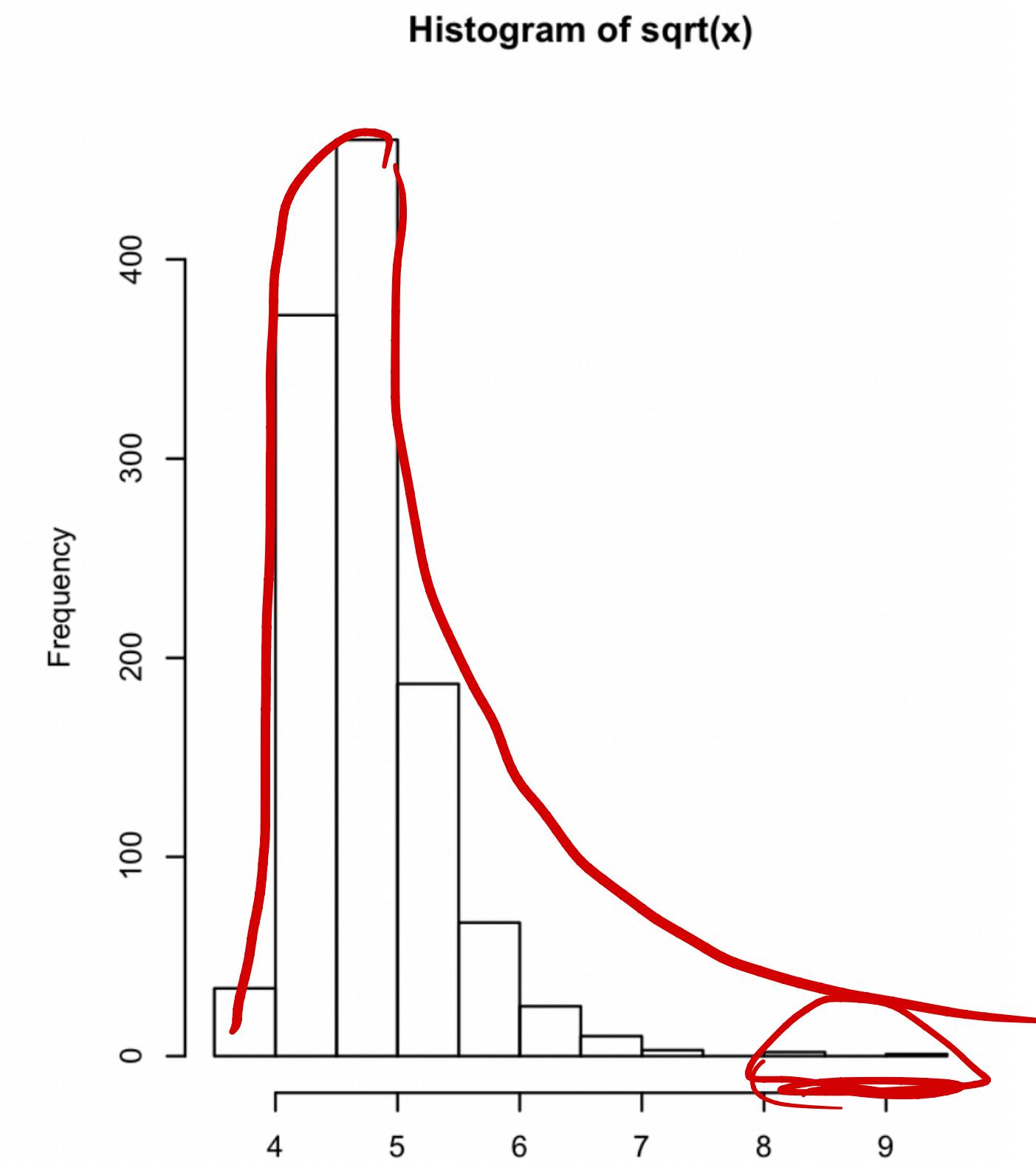
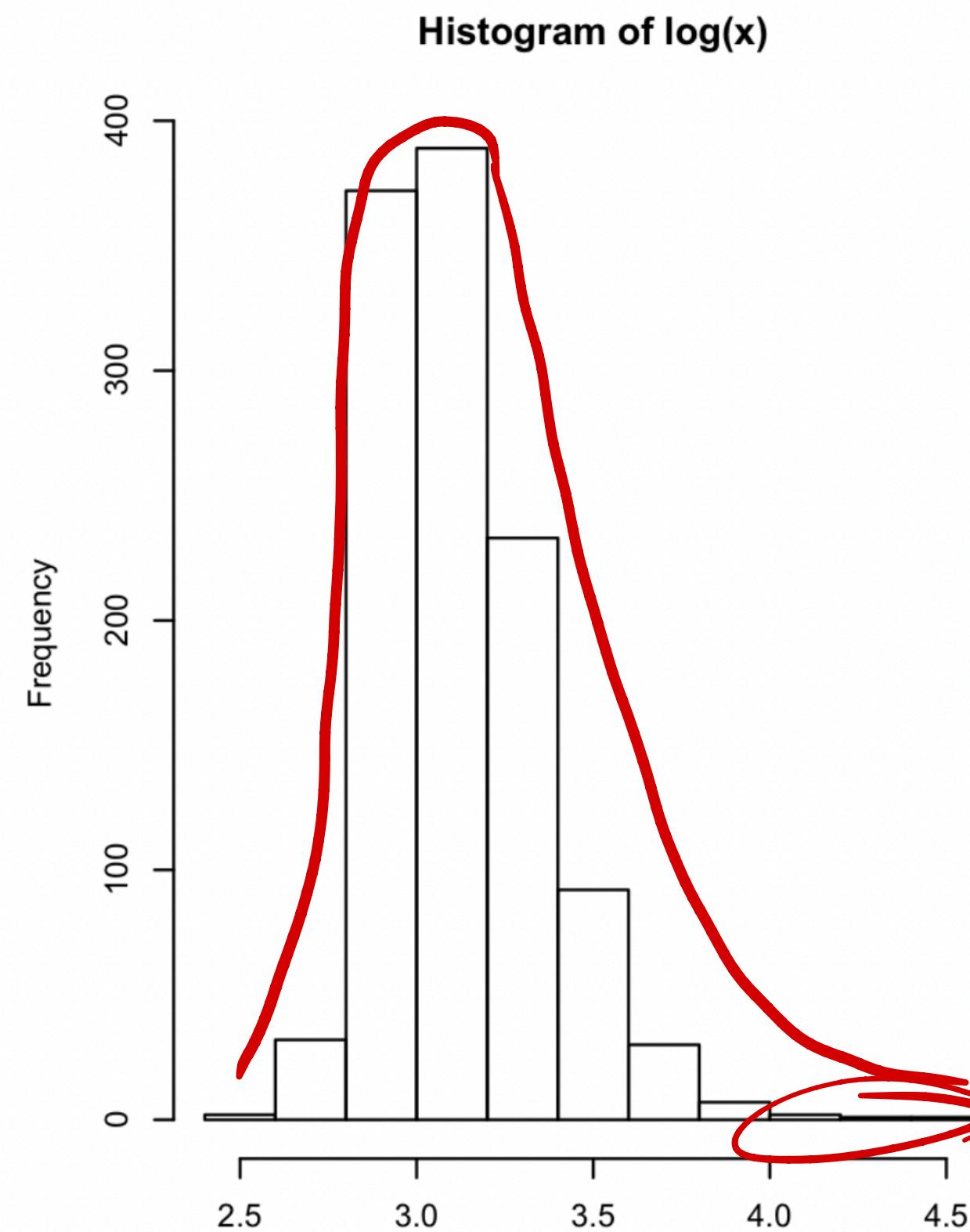
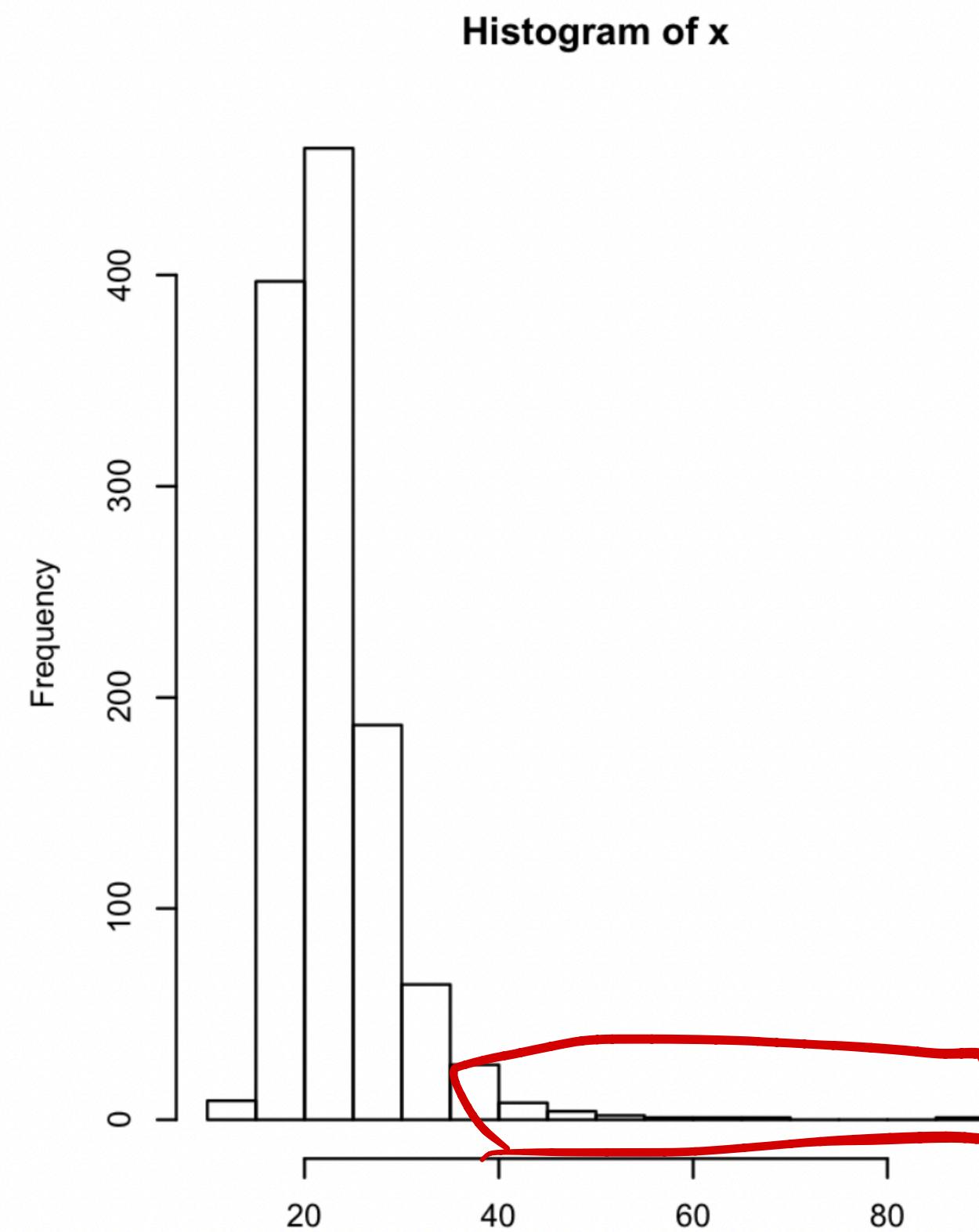
# Transformations

- General Social Survey (GSS) collects data annually on a sample of Americans regarding demographic and lifestyle characteristics
- Consider the 2006 survey, which collected data on the age at which respondents were first married
- Right skew to data

# Transformations

- General Social Survey (GSS) collects data annually on a sample of Americans regarding demographic and lifestyle characteristics
- Consider the 2006 survey, which collected data on the age at which respondents were first married
- Right skew to data
- Apply  $f(x) = \log(x)$  and  $f(x) = \sqrt{x}$  transformations to see which fits better

# Transformations



# Transformations

# Transformations

- The  $f(x) = \log(x)$  transformation seems to reduce skewness more, although the distribution is not perfectly symmetric

# Transformations

- The  $f(x) = \log(x)$  transformation seems to reduce skewness more, although the distribution is not perfectly symmetric
- Check empirical rule:

# Transformations

- The  $f(x) = \log(x)$  transformation seems to reduce skewness more, although the distribution is not perfectly symmetric
- Check empirical rule:

	k	$\bar{x} - ks$	$\bar{x} + ks$	Theoretical % in Range	Actual % in Range
x	1	17.20	29.45	68	81.6
	2	11.08	35.58	95	96.2
	3	4.95	41.71	99.7	98.4
$\log(x)$	1	2.90	3.34	68	69.6
	2	2.68	3.57	95	95.8
	3	2.46	3.79	99.7	99.1
$\sqrt{x}$	1	4.22	5.37	68	79.7
	2	3.65	5.94	95	96.0
	3	3.08	6.51	99.7	98.6

# Box-Cox Power Transformations

# Box-Cox Power Transformations

- Data is often log-transformed for convenience and interpretability

# Box-Cox Power Transformations

- Data is often log-transformed for convenience and interpretability
- A more flexible approach is the *Box-Cox* power transformation, parameterized by a value  $\lambda$

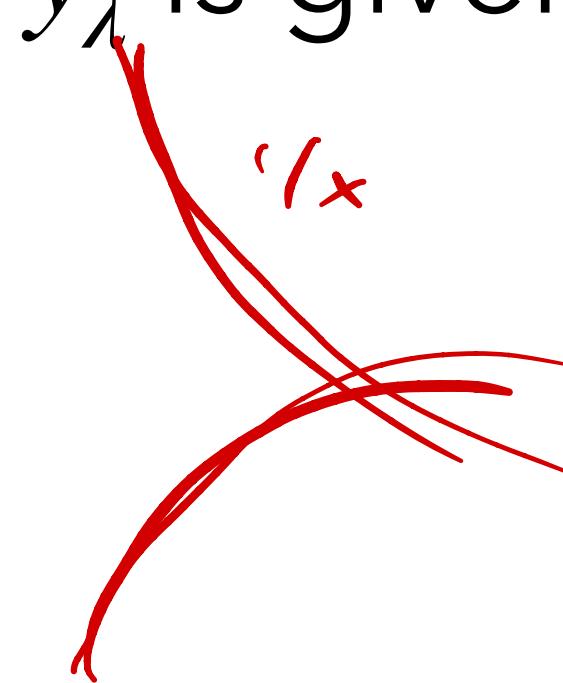
# Box-Cox Power Transformations

- Data is often log-transformed for convenience and interpretability
- A more flexible approach is the *Box-Cox* power transformation, parameterized by a value  $\lambda$
- For a positive  $x$ , the transformed value  $y_\lambda$  is given as follows:

# Box-Cox Power Transformations

- Data is often log-transformed for convenience and interpretability
- A more flexible approach is the *Box-Cox* power transformation, parameterized by a value  $\lambda$
- For a positive  $x$ , the transformed value  $y_\lambda$  is given as follows:

$$y_\lambda = \begin{cases} \frac{x^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log(x), & \lambda = 0 \end{cases}$$



# Box-Cox Power Transformations

# Box-Cox Power Transformations

- R has a built-in function to estimate the optimal parameter  $\hat{\lambda}$  given data  $x$

# Box-Cox Power Transformations

- R has a built-in function to estimate the optimal parameter  $\hat{\lambda}$  given data  $x$
- R code:

# Box-Cox Power Transformations

- R has a built-in function to estimate the optimal parameter  $\hat{\lambda}$  given data  $x$
- R code:

```
library(MASS)
```

# Box-Cox Power Transformations

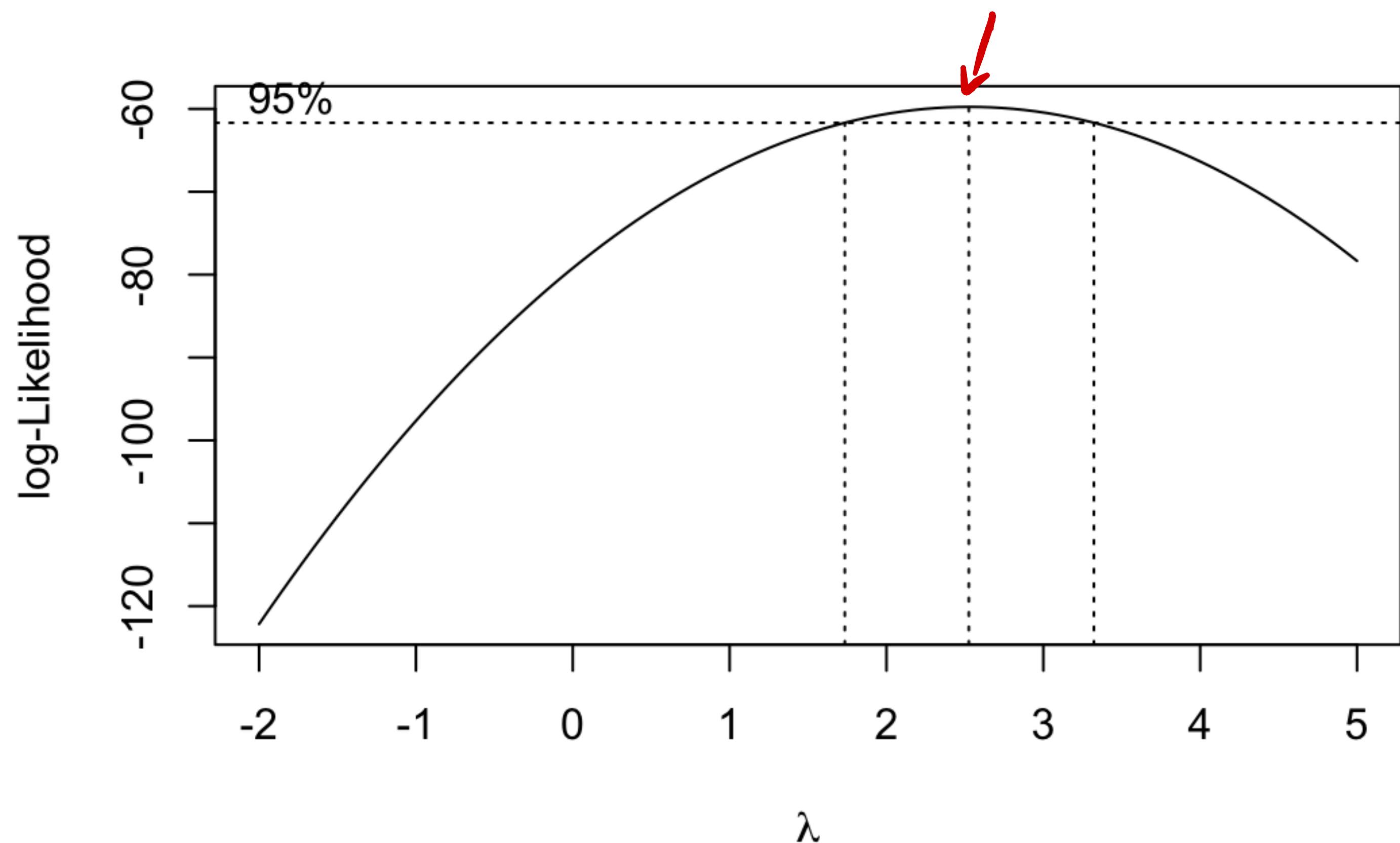
- R has a built-in function to estimate the optimal parameter  $\hat{\lambda}$  given data  $x$
- R code:

```
library(MASS)  
  
bc1 <- boxcox(x~1)
```

# Box-Cox Power Transformations

- R has a built-in function to estimate the optimal parameter  $\hat{\lambda}$  given data  $x$
- R code:

```
library(MASS)  
  
bc1 <- boxcox(x~1)
```



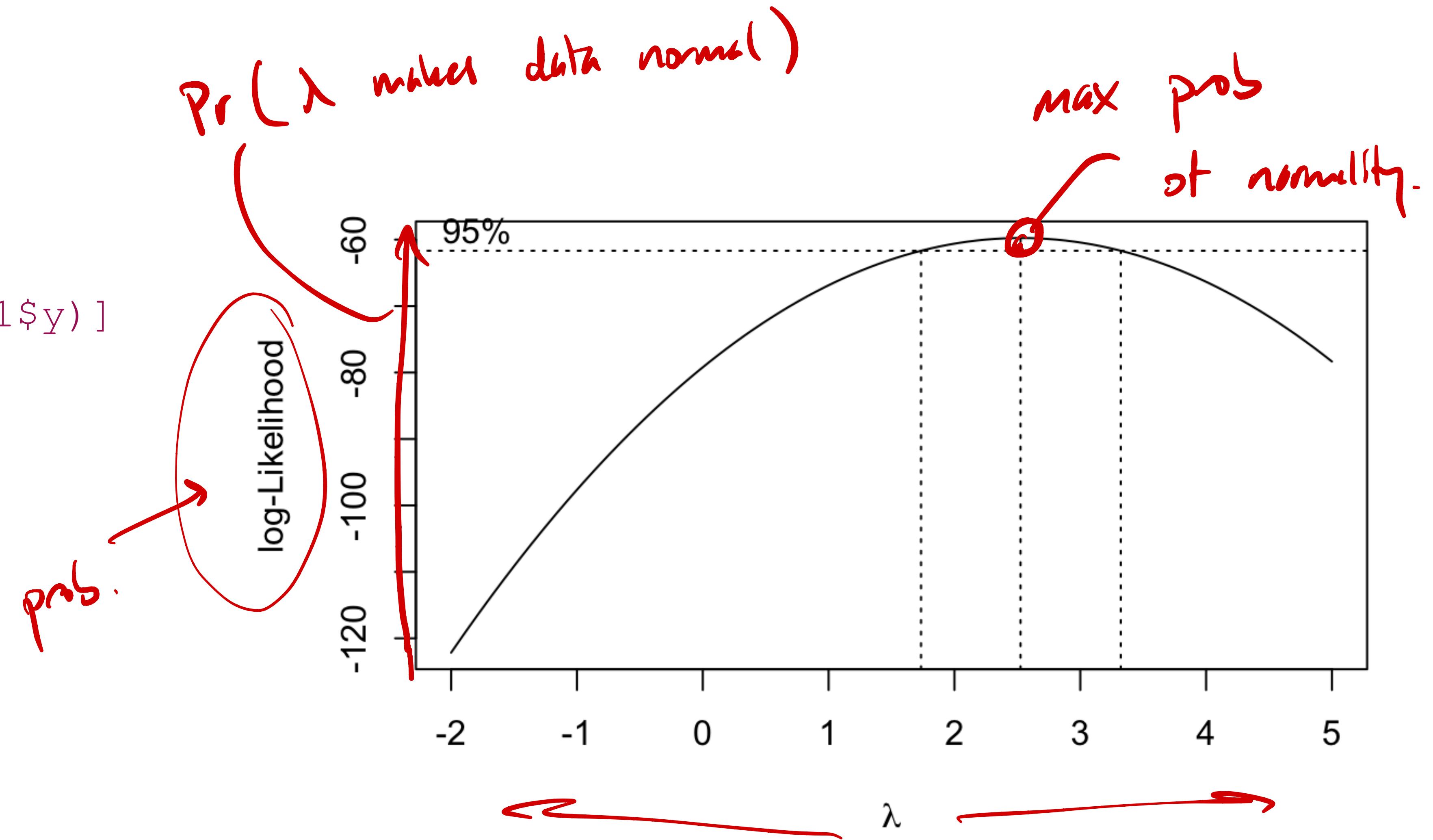
# Box-Cox Power Transformations

- R has a built-in function to estimate the optimal parameter  $\hat{\lambda}$  given data  $x$
- R code:

```
library(MASS)
```

```
bc1 <- boxcox(x~1)
```

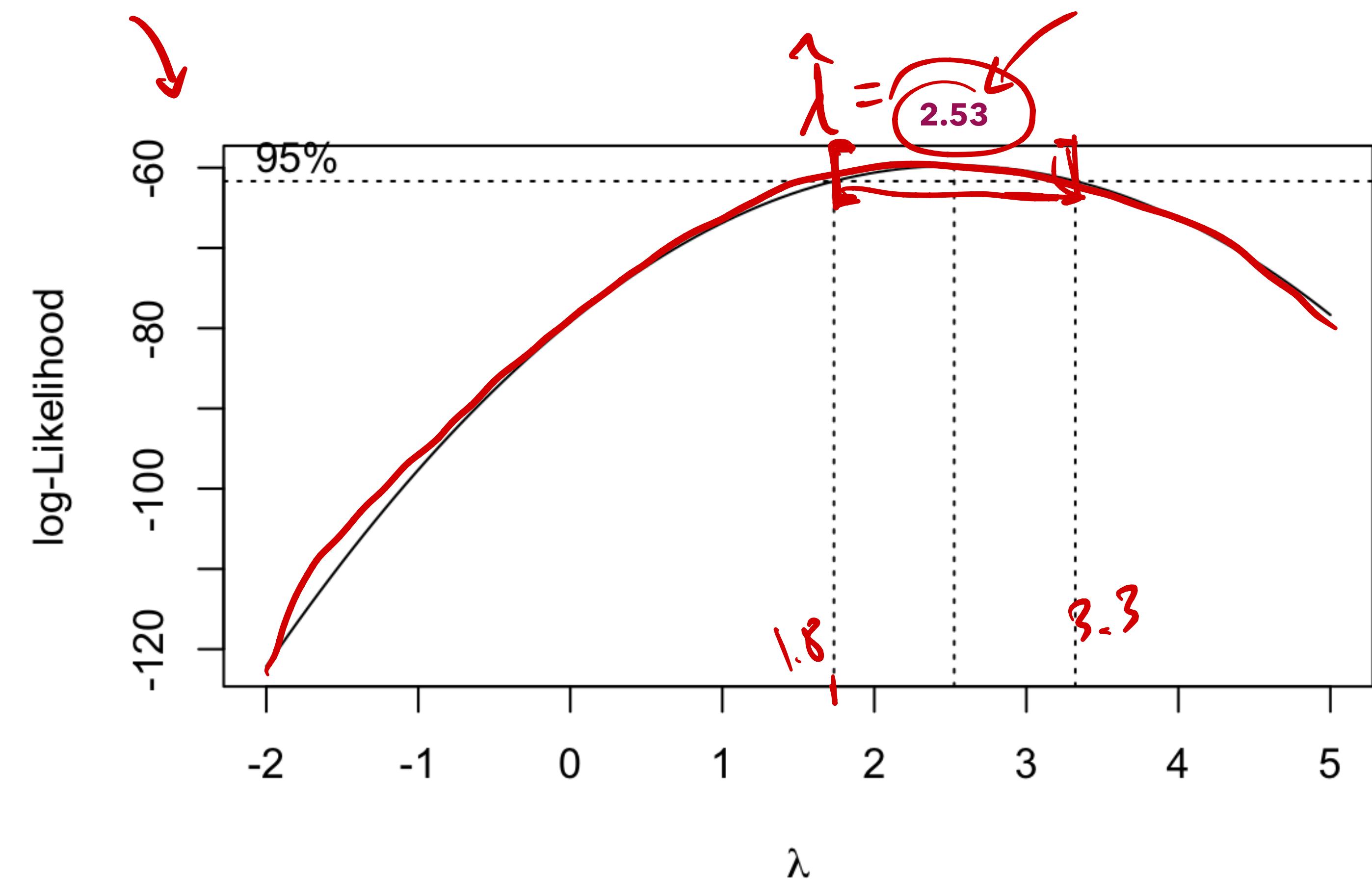
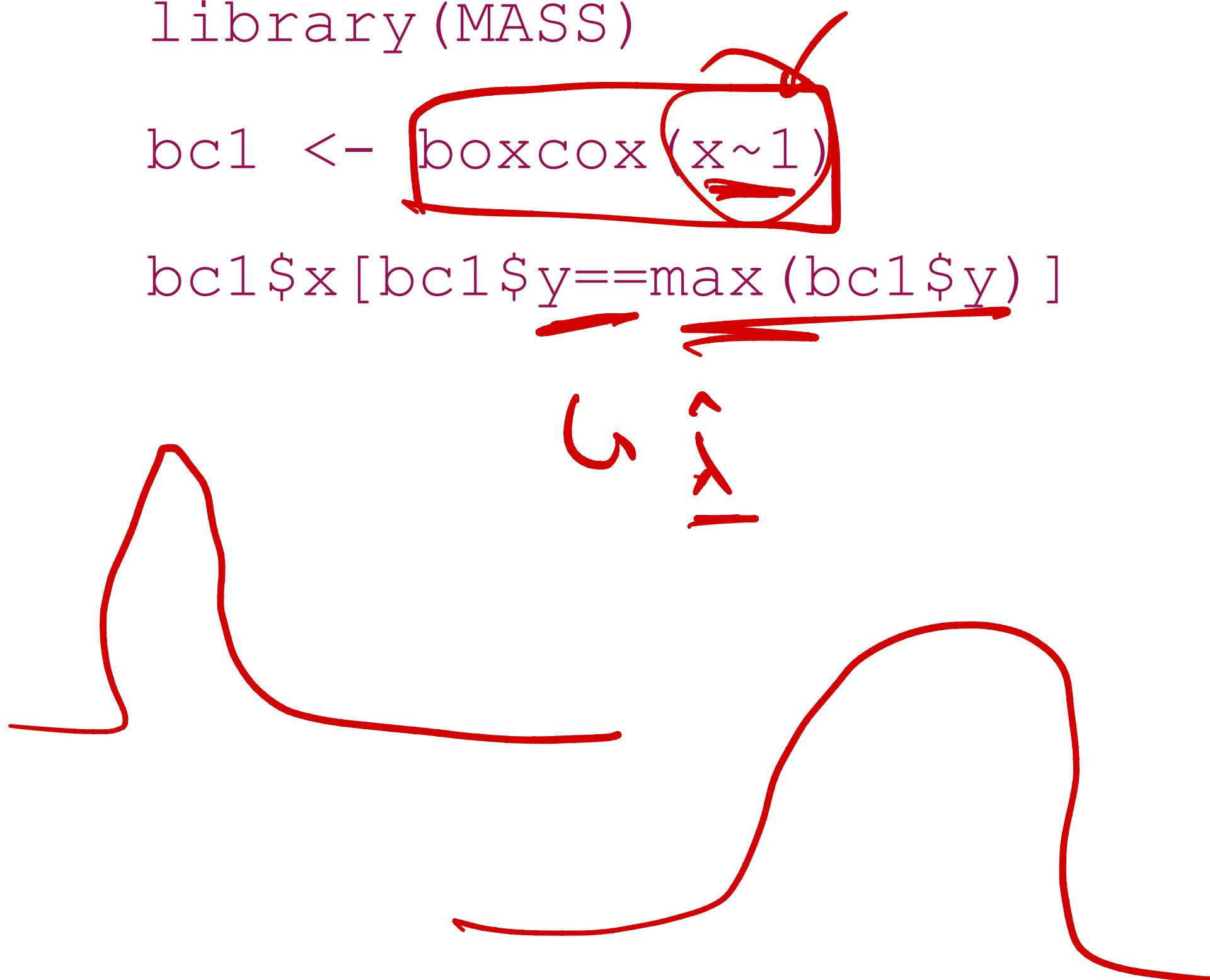
```
bc1$x[bc1$y==max(bc1$y) ]
```



# Box-Cox Power Transformations

- R has a built-in function to estimate the optimal parameter  $\hat{\lambda}$  given data  $x$
- R code:

```
library(MASS)  
bc1 <- boxcox(x~1)  
bc1$x[bc1$y==max(bc1$y)]
```



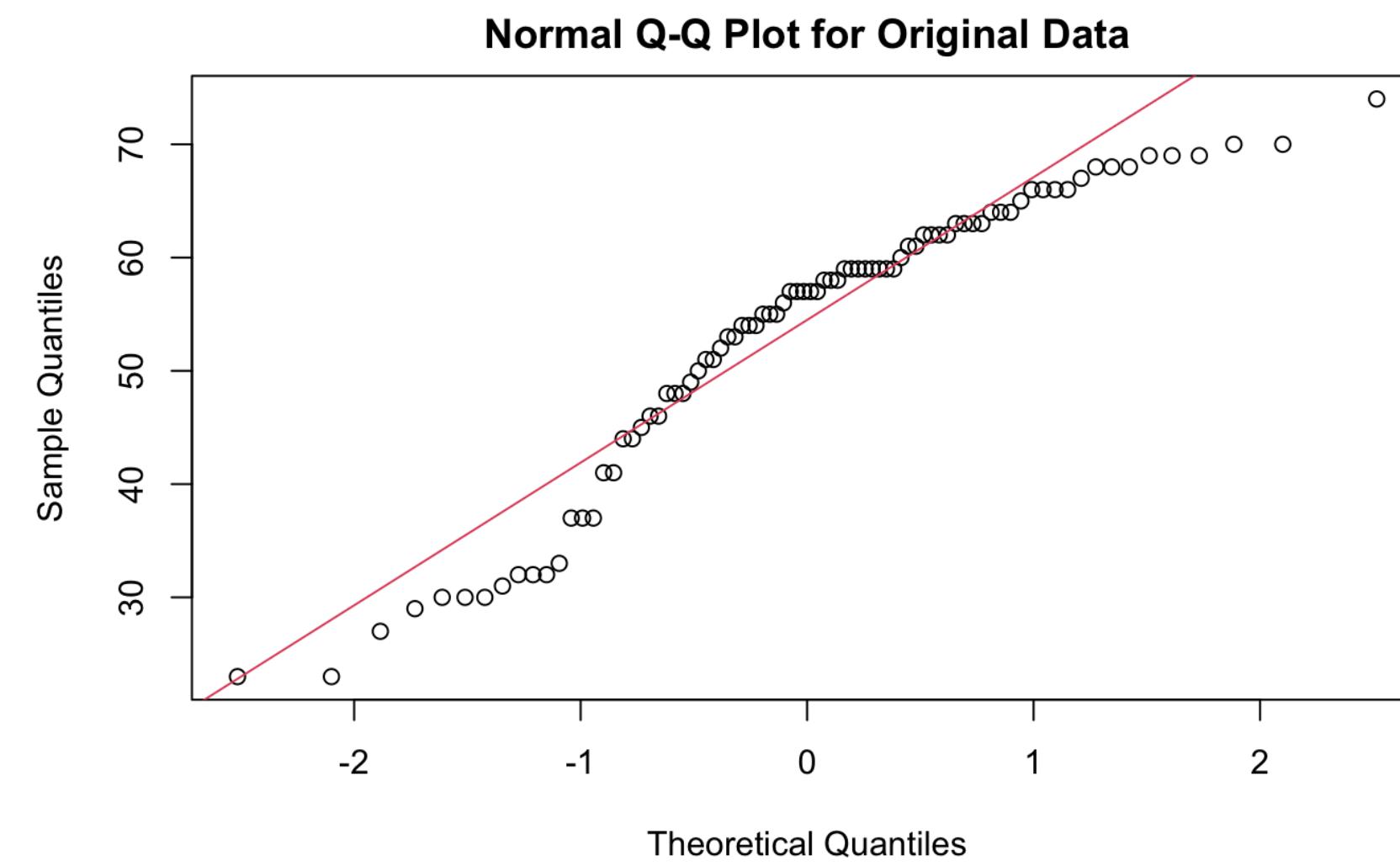
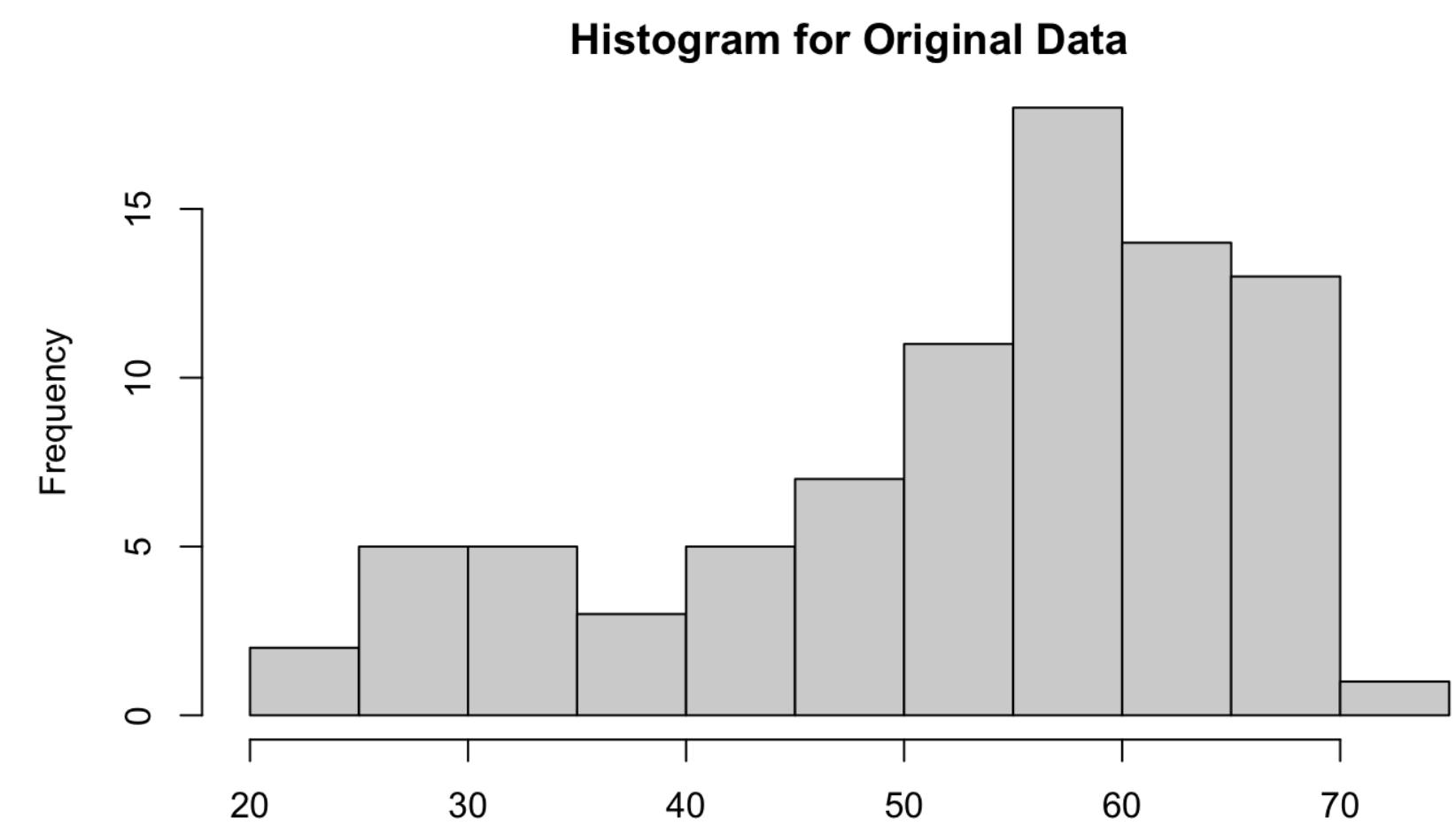
# Box-Cox Power Transformations

# Box-Cox Power Transformations

- Let's examine a histogram and Q-Q plot of the Box-Cox transformed data:

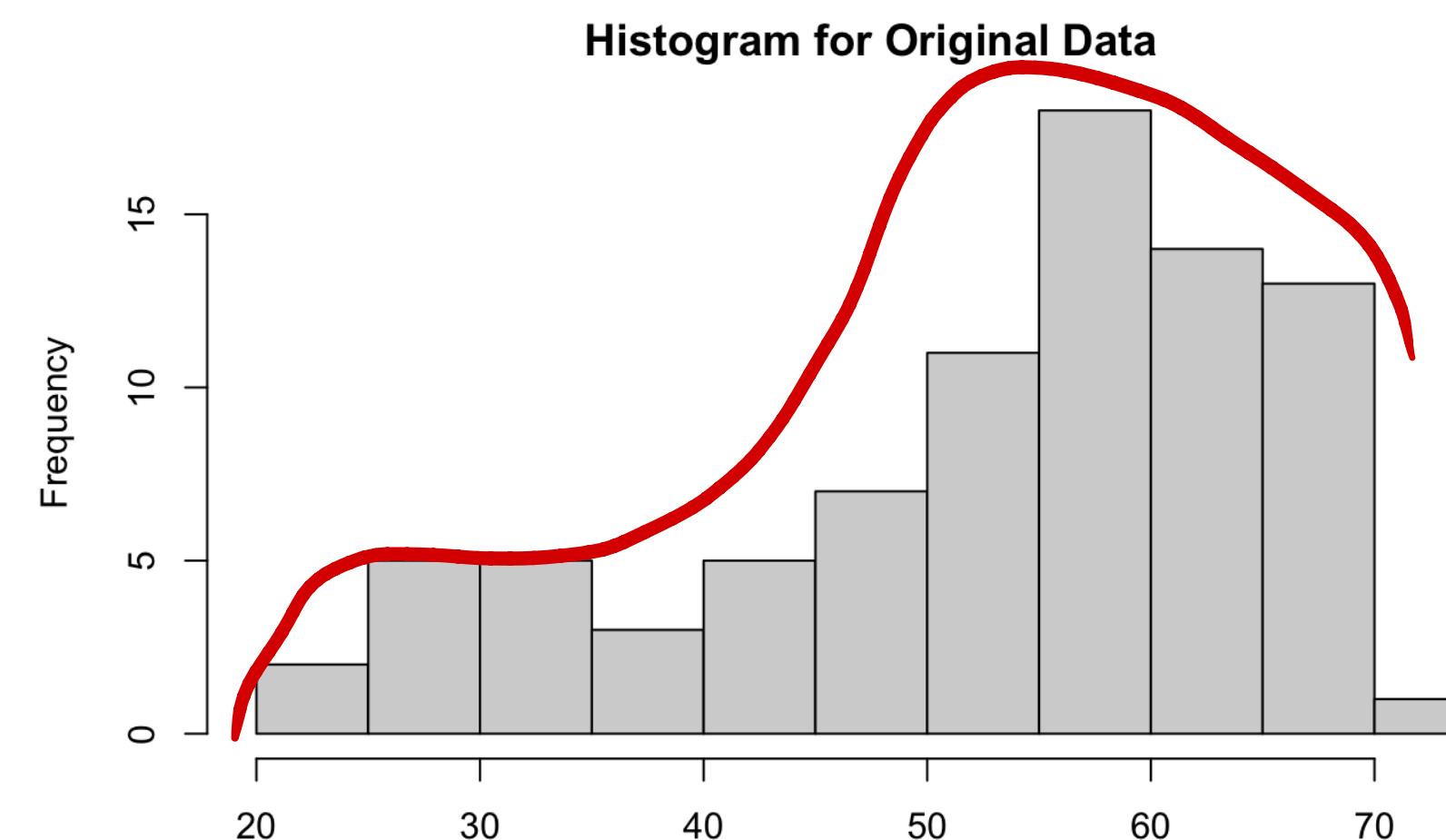
# Box-Cox Power Transformations

- Let's examine a histogram and Q-Q plot of the Box-Cox transformed data:



# Box-Cox Power Transformations

- Let's examine a histogram and Q-Q plot of the Box-Cox transformed data:



$\hat{\lambda}$

