Chapter 1. Looking at Data – Distributions

1.1 Data

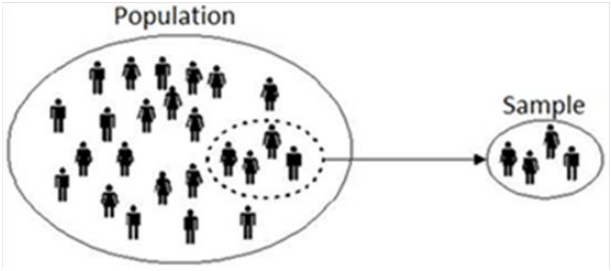
Statistics: science of learning/reasoning from data (data =)

Population versus Sample

- **Population**: the entire group to be studied
  - Population Size:
  - Population **Parameter**: a numerical summary of a _____
- **Sample**: a subset of the population we actually select for collecting data
  - Sample Size:
  - Sample **Statistic**: a numerical summary of a _____

< Ex > Ben's Project: Ben is about to graduate and will work full-time. He's been exercisesing and wants to know how many hours, on average, full-time workers exercise per week.

- target population:
- population parameter he wants to know:
- information he needs to collect from each person:



He randomly selects _____ and finds _____ hours per week

- sample size:
- sample statistics he obtained:

He reports that _____ exercise an average of _____ hours per week.

Descriptive Statistics: Organizing and Summarizing data

Inferential Statistics: Taking the result from a sample and extending it to the population

< Ex > Project: Does the Media Affect You?

Main Research Question: Do media and advertisements affect how college students feel about their physical appearance?

Some Hypotheses:

- Media and advertisements negatively affect student's happiness with their physical appearance.
- Female students are more negatively affected by media and advertisements in terms of body image than men are.

Data Collection: Survey

1. What is your gender? Males or Female

4. How many magazines do you read in a month? _____

5. What type of magazines do you read? Circle all that apply.

Heath/Fitness Sport Fashion Adult
Celebrity News Hobby Other

6. How many hours do you watch TV per week? _____

14. How happy are you with your physical appearance on a scale from 1 to 7 (1: very unhappy, 7: very happy)? _____

Identify the variable type.

(a) What is your gender?

(b) How many magazines do you read in a month?

(c) What type of magazines do you read?

(d) How many hours do you watch TV per week?

(e) How happy are you with your physical appearance on a scale from 1 to 7 (1: very unhappy, 7: very happy)?

Variables: the information we want to learn about the individuals
the characteristics of the individuals

Types of Variables: $\begin{cases} \text{Qualtative or Categorical} \\ \text{Quantitative or Numerical} \end{cases} \rightarrow \begin{cases} \text{Discrete} \\ \text{Continuous} \end{cases}$

- **Qualitative/Categorical** V**ariable:** Allow for classification of individuals based on some attribute or characteristic
- **Quantitative/Numerical Variable:** Provide numerical measures of individuals. The values can be added or subtracted and provide meaningful results
  - **Discrete Variable:** has either a finite number of possible values or a countable number of possible values. (counting such as 1, 2, 3, and so on)
  - **Continuous Variable:** has an infinite number of possible values that are not countable

Level of Measurement of a Variable

Rather than classify a variable as qualitative or quantitative, we can assign a level of measurement to the variable

Nominal Level Ordinal Level Interval Level Ratio Level

Nominal Level vs. Ordinal Level

- A variable is at the **nominal level of measurement** if the values of the variable are names or categorizes. It does not allow for the values of the variables to be arranged in a ranked order or specific order
- A variable is at the **ordinal level of measurement** if it has the properties of the nominal level of measurement, however It allows for the values of the variables to be arranged in a ranked order or specific order

1.2 Displaying Distributions with Graphs

We want to explore data or summarize data to describe their main features.

Summarize Qualitative Data: { Frequency or Relative Frquency Table (Distribution); Bar Plot (Bar Graph, Bar Chart); Pie Chart }

Frequency Tables and Relative Frequency Tables

- **Frequency Table**: lists each category of data together with the frequency for each category
- **Relative Frequency Table:** lists each category of data together with the relative frequency

Note: The Relative Frequency is the proportion (or percent) of observations within a category

< Ex > A physical therapist wants to determine types of rehabilitation required by her patients. She obtains a sample of 30 of her patients and records the body part requiring rehabilitation.

(a) Construct a frequency table.

Back	Back	Hand
Wrist	Back	Groin
Elbow	Back	Back
Back	Shoulder	Shoulder
Hip	Knee	Hip
Neck	Knee	Knee
Shoulder	Shoulder	Back
Back	Back	Back
Knee	Knee	Back
Hand	Back	Wrist

Frequency Table

Body Part	Frequency (Count)
Back	
Wrist	
Elbow	
Hip	
Shoulder	
Knee	
Hand	
Groin	
Neck	

```
> table(BodyPart)
BodyPart
    Back    Elbow    Groin     Hand      Hip     Knee     Neck
      12        1        1        2        2        5        1
Shoulder    Wrist
       4        2
```

(b) Construct a relative frequency table.

Relative Frequency Table

Body Part	Relative Frequency
Back	
Wrist	0.0667
Elbow	0.0333
Hip	0.0667
Shoulder	0.1333
Knee	0.1667
Hand	0.0667
Groin	0.0333
Neck	0.0333

Relative Frequency Table (in Percent)

Body Part	Percent (%)
Back	
Wrist	6.67
Elbow	3.33
Hip	6.67
Shoulder	13.33
Knee	16.67
Hand	6.67
Groin	3.33
Neck	3.33

```
> table(BodyPart) / 30
BodyPart
      Back      Elbow      Groin       Hand        Hip       Knee
0.40000000 0.03333333 0.03333333 0.06666667 0.06666667 0.16666667
      Neck   Shoulder      Wrist
0.03333333 0.13333333 0.06666667
```

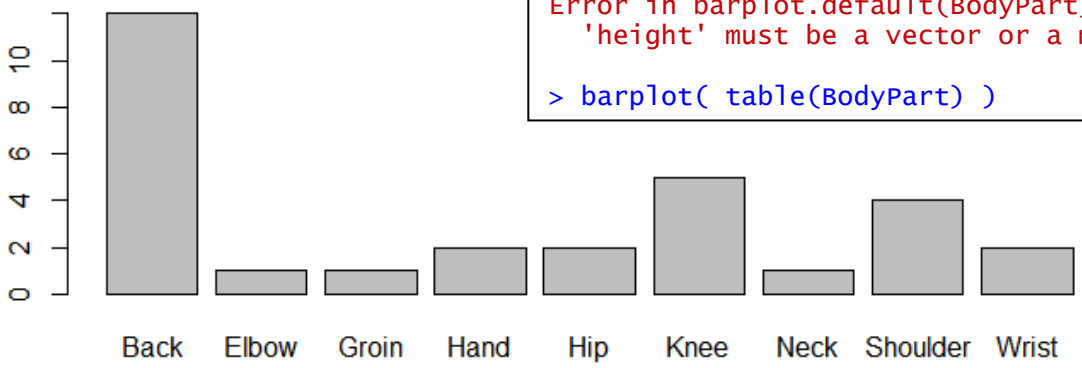
<u>Bar Plot</u>: Label each category of data on either the horizontal (or vertical) axis and the frequency or relative frequency of the category on the other axis.

Note: The _____ of each rectangle represents the category's frequency or relative frequency

< Ex > (continue) A physical therapist wants to determine types of rehabilitation required by her patients. She records the body part requiring rehabilitation from 30 patients. Draw a bar plot.

```
> barplot(BodyPart)
Error in barplot.default(BodyPart) :
  'height' must be a vector or a matrix

> barplot( table(BodyPart) )
```



< Ex > Titanic Data with 1,309 passengers (A portion of the data is shown below)

Class	Status	Sex	Age
First	Survived	male	80
First	Survived	female	76
Third	Died	male	74
First	Died	male	71

```
> table(Status, Class)
          Class
Status     First Second Third
  Died       123    158   528
  Survived   200    119   181
```

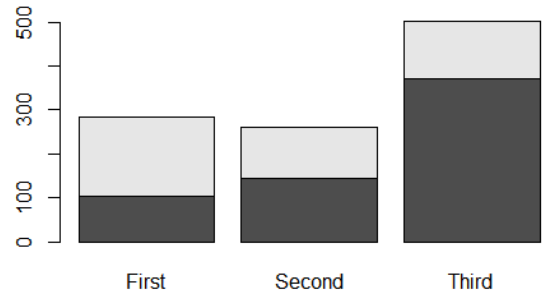
The following summarizes the frequencies for each category of Class & Status.

Survival Status	**Passenger Ticket Class**			
	First	**Second**	**Third**	
Died	123	158	528	
Survived	200	119	181	

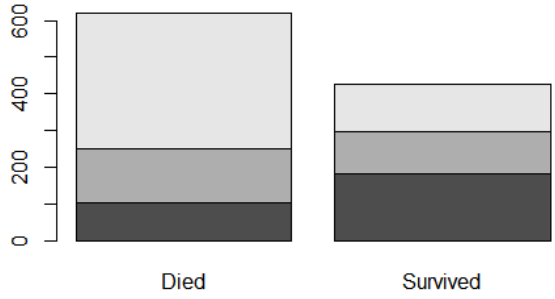
(a) We want to **compare the death rates** for three passenger ticket classes (First, Second, Third). We can use a **Side-by-Side Bar plot.** Draw a side-by-side bar plot.

Stacked Bar Plots (default plot)

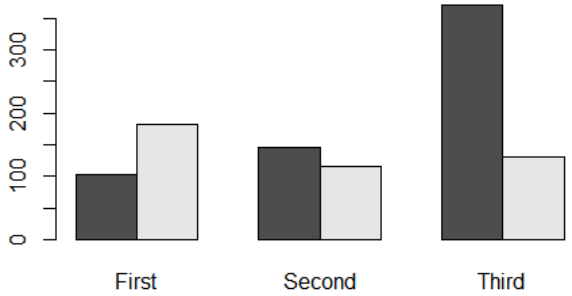
```
> barplot(table(Status,Class))
```

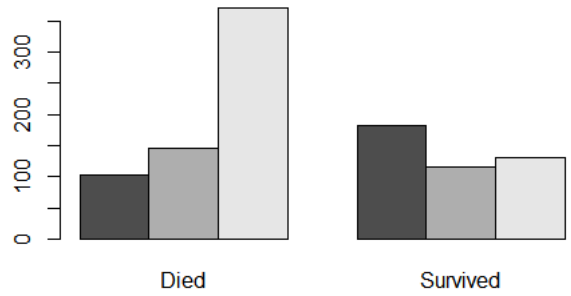



```
> barplot(table(Class,Status))
```



Side-by-Side Bar Plots






```
> barplot(table(Status,Class),
beside=T)
```

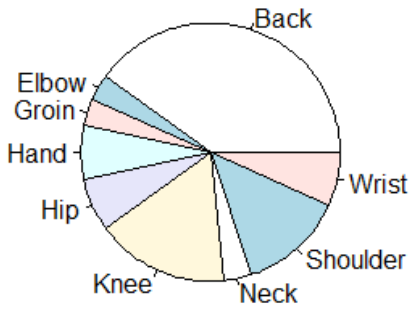
```
> barplot(table(Class,Status),
beside=T)
```

(b) Make some general conclusions based on the graph in (a).

Pie Chart: A pie chart is a circle divided into sectors (i.e. pie slides). Each sector represents a category of data.

Note: The _____ of each sector (i.e. pie slide) is proportional to the frequency of the category.

< Ex > (continue) A physical therapist wants to determine types of rehabilitation required by her patients. She records the body part requiring rehabilitation from 30 patients. Draw a pie chart.




```
> pie(BodyPart)
Error in pie(BodyPart) : 'x' values
must be positive.

> pie( table(BodyPart) )
```

Graphs Using Quantitative Variables

We want to explore data or summarize data to describe their main features.

Summarize Quantitative Data: { Frequency or Relative Frquency Table; Histogram; Stem and Leaf Plot and Dot Plot; Boxplot }

< Ex > Data: 20 Test Scores. Draw a dot plot.

77 78 79 80 80 81 82 83 83 83
84 85 86 87 87 88 89 92 93 98

Dot Plot

Place each observation horizontally and place a dot above the observation each time it is observed

Histogram

A histogram is constructed by drawing rectangles for each class of data. The width of each rectangle is the same and the rectangles tough each other.

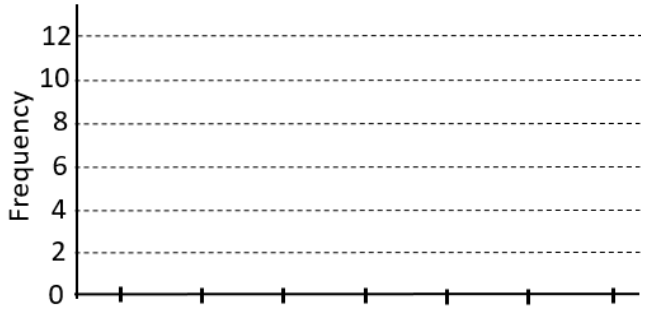
< Ex > Data: weights (in pounds) from a team of 18 rowers (they are arranged in order)

109.0	109.5	174.0	178.5	183.0	183.0	184.5	184.5	185.0
186.0	186.0	188.5	194.5	195.5	200.0	202.5	203.5	214.0

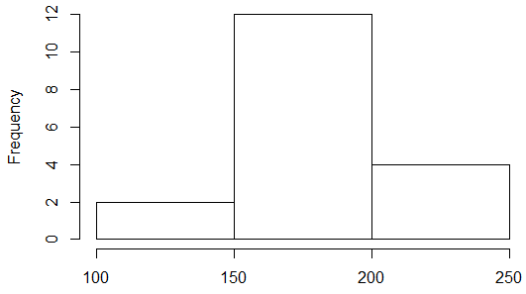
(a) Create a frequency table.

Class (Weight)	Frequency

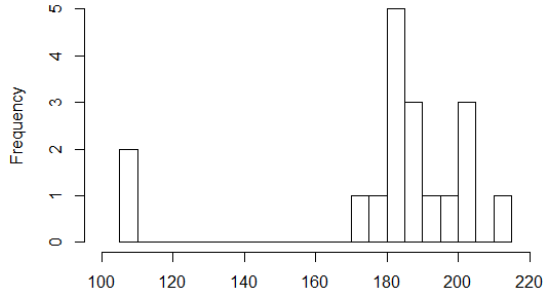
(b) Draw a histogram.



< Side Note > Histograms when the number of classes are too few or too many.

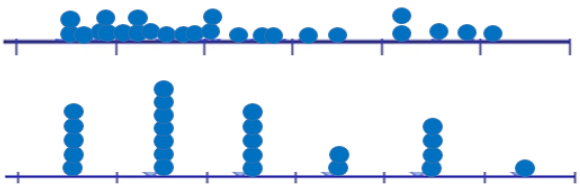


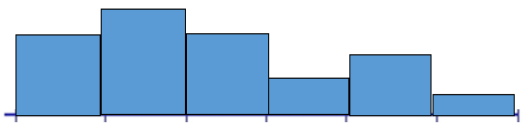

```
> hist(weight, nclass=3)
```




```
> hist(weight, nclass=20)
```

Histogram





Stem-and-Leaf Plot (Stem Plot): Separate each number into two parts, a **stem** and a **leaf**.

- Use the digits to the left of the rightmost digit to form the **stem**
- Each rightmost digit forms a **leaf**

< Ex > Data: 20 Test Scores Draw a stem plot.

77 78 79 80 80 81 82 83 83 83
84 85 86 87 87 88 89 92 93 98

Construction of a Stem Plot

1. Write the stems in a vertical column in increasing order. Draw a vertical line to the right of the stems

2. Write each leaf corresponding to the stems to the right of the vertical line. Within each stem, rearrange the leaves in ascending order.

3. Include the legend to indicate what the values represent

<< Attention >> If the data values are as shown below, what will be different in the stem plot?

7.7 7.8 7.9 8.0 8.0 8.1 8.2 8.3 8.3 8.3
84. 8.5 8.6 8.7 8.7 8.8 8.9 9.2 9.3 9.8

Split Stems: The data appear rather bunched, we can use split stems.

27	17	11	24	36
13	29	22	18	17
23	30	12	46	17
32	48	11	18	23
18	32	26	24	38
24	15	13	31	22
18	21	27	20	16
15	37	19	19	29

1	1 1 2 3 3 5 5 6 7 7 7 8 8 8 8 9 9
2	0 1 2 2 3 3 4 4 4 6 7 7 9 9
3	0 1 2 2 6 7 8
4	6 8

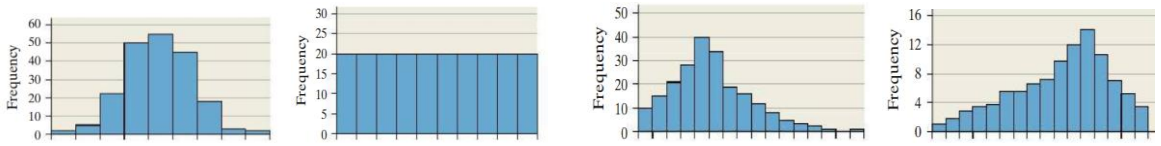
Legend: 1|1 represents 11

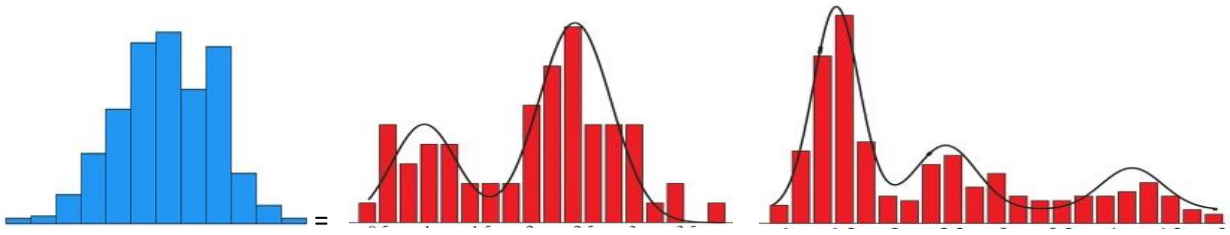
1	1 1 2 3 3
1	5 5 6 7 7 7 8 8 8 8 9 9
2	0 1 2 2 3 3 4 4 4
2	6 7 7 9 9
3	0 1 2 2
3	6 7 8
4	
4	6 8

Legend: 1|1 represents 11

Identify the Shape of a Distribution (Shape of a Plot)

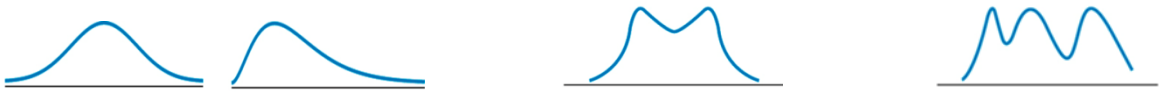
Symmetric **Skewed**





Note: Points at which distributional shapes peak are called **modes** of distribution

Unimodal Shape Bimodal Shape Multimodal Shape



< Ex > Which distributional shape (symmetric, right-skewed, or left-skewed) is expected?

(a) scores from a hard test **(b)** ages for heart attack patients

1.3 Displaying Distributions with Numbers

Measuring Central Tendency (= typical value): Mean (Arithmetic Mean), Median, Mode

Sample Mean = Average: $\bar{x}$

Data Values: $x_1,\ x_2,\ \cdots,\ x_n$

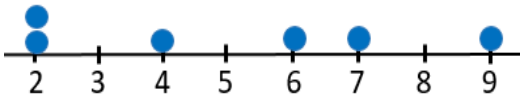
$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x_i}{n}$$

Population Mean: μ

Values in the Population: $x_1,\ x_2,\ \cdots,\ x_N$

$$\mu = \frac{x_1 + x_2 + \cdots + x_n}{N} = \frac{\sum x_i}{N}$$

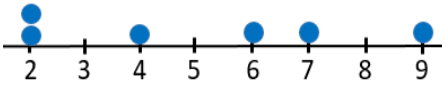
< Ex > Weekly Study Hour for 6 students: { 2, 6, 7, 2, 4, 9 } Calculate the sample mean.



Sample Median: M

The median is he value that lies in the middle when arranged un ascending order

< Ex > Weekly Study Hour for 6 students: { 2, 6, 7, 2, 4, 9 } Calculate the sample median.

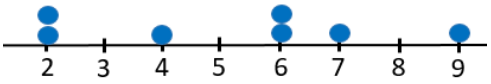


Finding the Median of a Data Set

1. Arrange the data in ascending order

2. Determine the value in the middle of the data set

< Ex > Find the median of the data { 2, 6, 7, 2, 4, 9 , 6}



< Ex > Amy wants to know how much time she spends on her cell phone. She goes to her phone's website and records the call lengths for a random sample of 12 calls.

1	7	4	1
2	4	3	48
3	5	3	6

Average Call Length: 7.3 minutes

Median Call Length: 3.5 minutes

(a) Which measure of central tendency better describes the length of a typical call length?

(b) There is one phone call that took far longer than the others (it is called an _____)
It was a mistake, and it actually was a 5-minute call. Recalculate the mean and median.

Mean:

Median:

(c) Compare the values in (a) and (b). What do we see? Which measure (mean or median) is **resistant** to extreme values?

Note: A numerical summary of data is said to be **resistant** if extreme values (very large or small) relative to the data do not affect its value substantially.

< Ex > Data: weights (in pounds) from a team of 18 rowers (they are arranged in order)

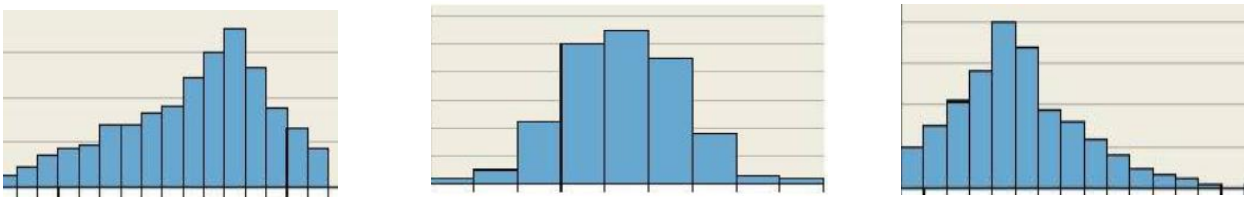
109.0	109.5	174.0	178.5	183.0	183.0	184.5	184.5	185.0
186.0	186.0	188.5	194.5	195.5	200.0	202.5	203.5	214.0
Mean Weight: 181.1 pounds Median Weight: 185.5 pounds								

(a) Extreme Values in the data:

(b) What will happen to the values of mean and median if we removed extreme values?

<u>Mean versus Median and the Shape of Distribtion</u>

Skewed- Symmetric Skewed-



Note: We generally use _____. May prefer _____ when we expect extreme values

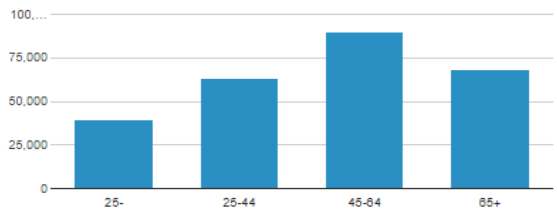
< Ex > Household incomes in OC: https://www.point2homes.com/US/Neighborhood/CA/Orange-County-Demographics.html

_____ Household Income: $101,547 _____ Household Income: $76,312

Household Income and Average Income in Orange County

top ▲

Median Income Under 25	$39,663
Median Income 25-44	$63,234
Median Income 45-64	$90,067
Median Income Over 65	$68,195

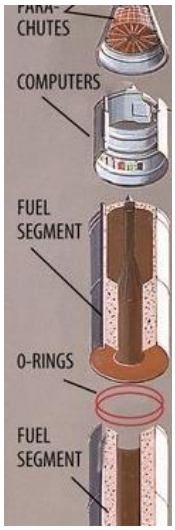


The **Mode** of a variable is the most frequently occurring value in the data

< Ex > The number of O-ring failures on the shuttle Columbia for its 17 flights.

0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 2, 3

Find the mode.

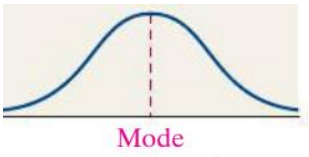


< Ex > Find the mode.

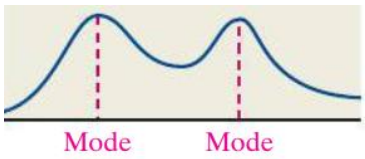
(a) Test Scores: 82, 77, 90, 71

(b) Test Scores: 82, 77, 90, 71, 71, 68, 92, 82

The data is unimodal



The data is bimodal



< Ex > The data represent the location of injuries that required rehabilitation by a physical therapist from a sample of 30 patients. Variable Type = _____

Hip	Back	Back	Back	Hand	Neck	Knee	Knee	Knee	Hand
Knee	Shoulder	Wrist	Back	Groin	Shoulder	Shoulder	Back	Knee	Back
Hip	Shoulder	Elbow	Back	Back	Back	Back	Back	Back	Wrist

Find the mode, which represents the location of injury that occurs the most frequently.

Measuring Variability/Spread: Range, Variance, Standard Deviation, IQR

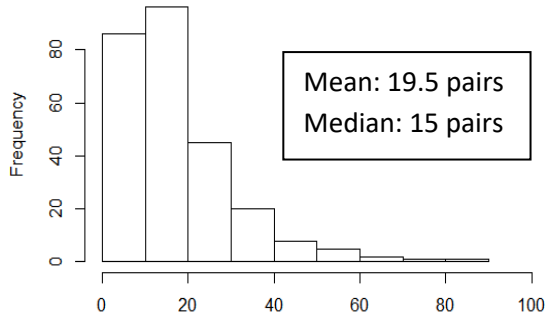
Note: Dispersion is the degree to which the data are spread out.

< Ex > Data Set with 455 students (A portion of the data is shown below)

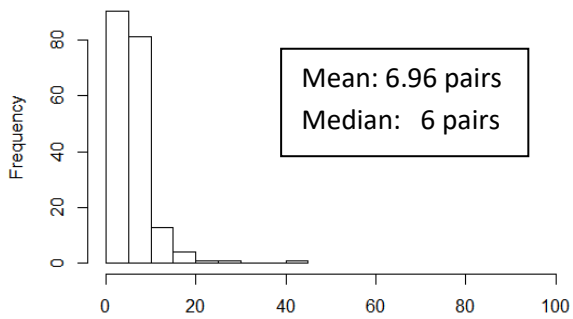
GENDER	AGE	WORKHR	TATTOO	SHOES	EXERCISE	GPA	CLASS	CARAGE	MARRIED
FEMALE	18	20	NO	23	5	2.9	FRESHMAI	1	NO
FEMALE	18	16	NO	12	3	3.9	FRESHMAI	2	NO

Variable: the number of pairs of shoes a person owns

264 Female Students



191 Male Students

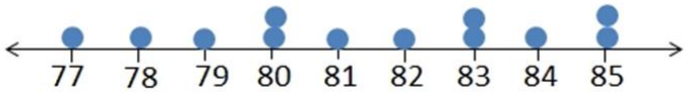


Range R: The **Range** is the difference between the largest and the smallest data value

R = largest data value – smallest data value (spread of the entire data)

< Ex > Data: 12 Test Scores { 77 78 79 80 80 81 82 83 83 83 84 85 }

Compute the range.



Sample Variance s^2

Data Values: $x_1,\ x_2,\ \cdots,\ x_n$

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{\text{sum of squared deviations from the mean}}{\text{one less than the sample size}}$$

Note: Its unit is square of the unit of the data values

Population Variance σ^2

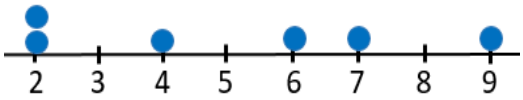
$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$$

< Ex > Weekly Study Hour for 6 students: { 2, 2, 4, 6, 7, 9 }

(a) Calculate the sample variance.

Note: Sample Mean: $\bar{x} = \frac{2+2+4+6+7+9}{6} =$ hours

Data Value (x_i)	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
2		
2		
4		
6		
7		
9		



(b) Calculate the sample standard deviation.

Sample Standard Deviation s

$s = \sqrt{s^2}$ i.e. $s = \sqrt{\text{sample variance}}$

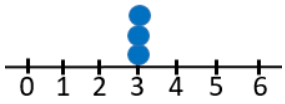
Population Standard Deviation σ

$\sigma = \sqrt{\sigma^2}$ i.e. $\sigma = \sqrt{\text{population variance}}$

Note: The unit of standard deviation is the same as the unit of the data values

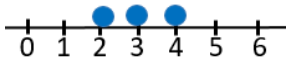
< Ex > Calculate the sample variance and sample standard deviation for each data set.

(a) Data: {3, 3, 3}



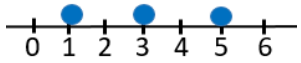
Data	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
2		
3		
4		

(b) Data: {2, 3, 4}



Data	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
2		
3		
4		

(c) Data: {1, 3, 5}



Data	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
1		
3		
5		

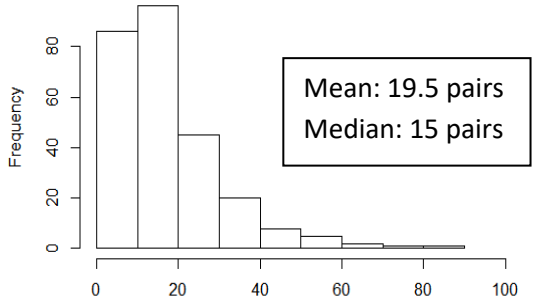
Interpretations of Standard Deviation

The standard deviation measures the spread (= variability) of the distribution. We can interpret the standard deviation as the **typical deviation from the mean**.

Note: The larger the standard deviation (variance), the _____ spread the distribution has.

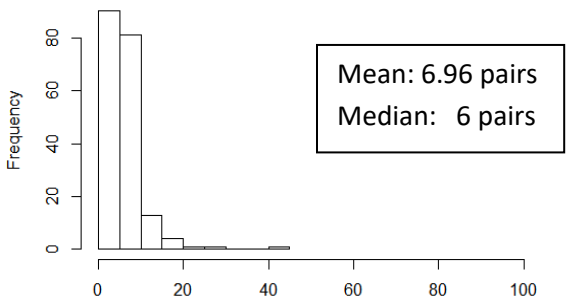
< Ex > Data with 455 students: Variable: the number of pairs of shoes a person owns

264 Female Students



Sample Standard Deviation: _____ pairs

191 Male Students



Sample Standard Deviation: _____ pairs

< Ex > Data Set with 455 students (A portion of the data is shown below)

GENDER	AGE	WORKHR	TATTOO	SHOES	EXERCISE	GPA	CLASS	CARAGE	MARRIED
FEMALE	18	20	NO	23	5	2.9	FRESHMAI	1	NO
FEMALE	18	16	NO	12	3	3.9	FRESHMAI	2	NO

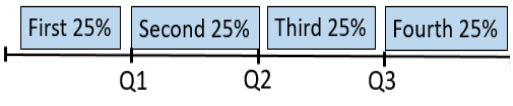
```
> mean(GPA)
[1] 3.051692
> median(GPA)
[1] 3
> var(GPA)
[1] 0.3050269
> sd(GPA)
[1] 0.5522924
```

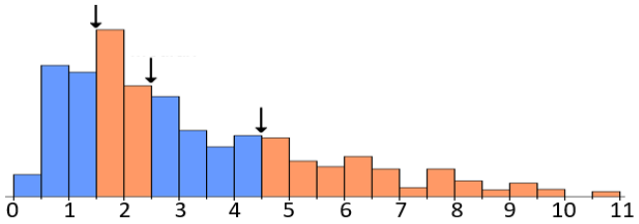
```
> mean(SHOES)
[1] 14.22198
> median(SHOES)
[1] 10
> var(SHOES)
[1] 156.3845
> sd(SHOES)
[1] 12.50538
```

```
> tapply(GPA, GENDER, mean)
  FEMALE     MALE
3.121591 2.955079
```

```
> tapply(SHOES, GENDER, sd)
   FEMALE      MALE
13.659613  4.897725
```

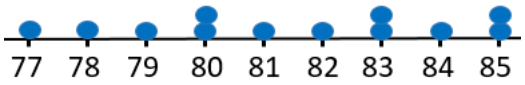
Quartiles divide data into four equal parts $\begin{cases} Q_L: \text{first quartile} \\ Q_2: \text{second quartile} \\ Q_3: \text{third quartile} \end{cases}$





< Ex > Data: 12 Test Scores { 77 78 79 80 80 81 82 83 83 83 84 85 }

What are the quartiles?

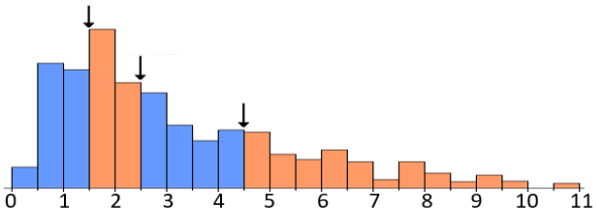



```
> summary(score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  77.00   79.75   81.50   81.25   83.00   85.00
```

```
> range(score)
[1] 77 85
> min(score)
[1] 77
> max(score)
[1] 85
```

The **Interquartile Range (IQR)** is the range of the middle 50% of the data

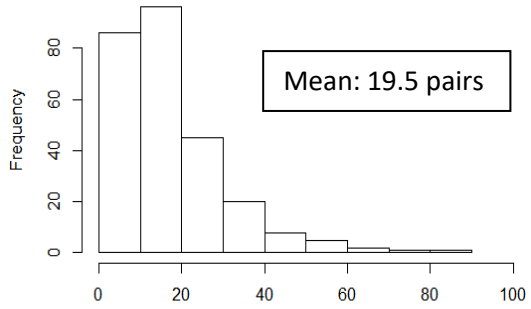
$$\text{IQR} = Q_3 - Q_1$$



< Ex > Data with 455 students: Variable: the number of pairs of shoes a person owns

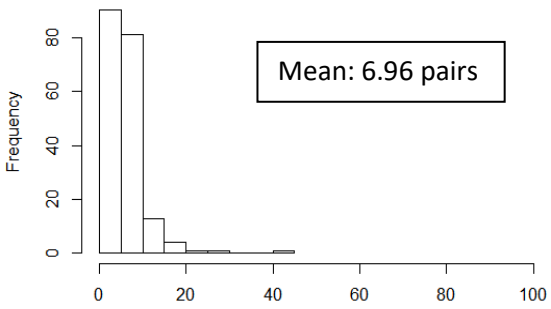
Calculate the Interquartile Range (IQR) for each gender.

264 Female Students



Min.	Q1	Q2	Q3	Max.
3	10	15	25	90

191 Male Students



Min.	Q1	Q2	Q3	Max.
1	4	6	8	45

Note: The more spread a data set has, the _____ the interquartile range will be.

< Ex > We've learned measures of spread (or variability); range, variance, standard deviation, IQR. Which one is resistant to extreme values?

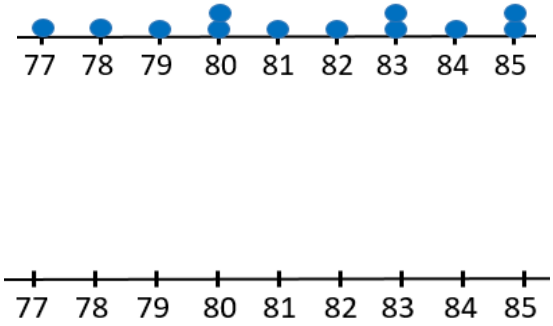
Five Number Summary

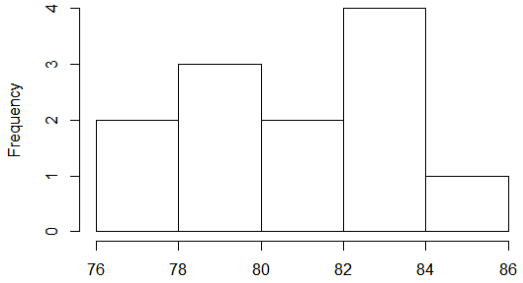
Five Numbers: Min., Q_1, Median (or Q_2) , Q_3, Max.

Boxplot (or **Box-and-Whisker Plot)**: graphical display of the five-number summary

< Ex > Data: 12 Test Scores. Draw a boxplot.

Min.	Q1	Med.	Q3	Max.
77.00	79.75	81.50	83.00	85.00



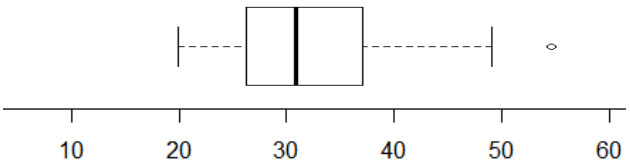


Checking for Outliers : Using the 1.5×IQR Rule

If a data value is less than the lower fence or greater than the upper fence, it is considered an outlier.

- Lower Fence: more than 1.5×IQR below Q_1 $(Q_1 - 1.5 \times IQR)$
- Upper Fence: more than 1.5×IQR above Q_3 $(Q_3 + 1.5 \times IQR)$

< Ex > The data show the finishing times (in minutes) for the men in the 60- to 64-year-old age group in a 5 kilometer race. The boxplot and the five-number summary are given below.



19.95	23.25	23.32	25.55	25.83	26.28	42.47
28.58	28.72	30.18	30.35	30.95	32.13	49.17
33.23	33.53	36.68	37.05	37.43	41.42	54.63

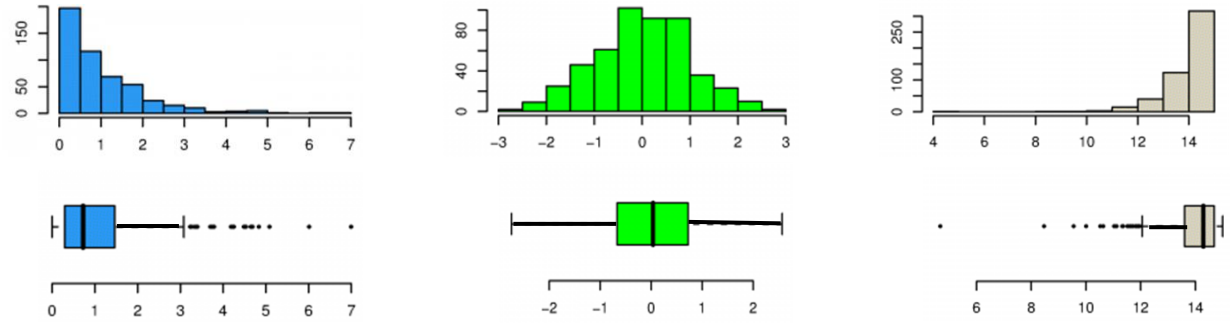
Min.	Q1	Med.	Q3	Max.
19.95	26.06	30.95	37.24	54.63

(a) Find the IQR.

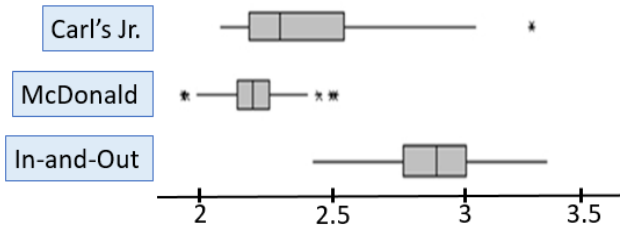
(b) Find the Lower Fence and the Upper Fence.

(c) Are there any outliers in the data set?

Histograms and Boxplots



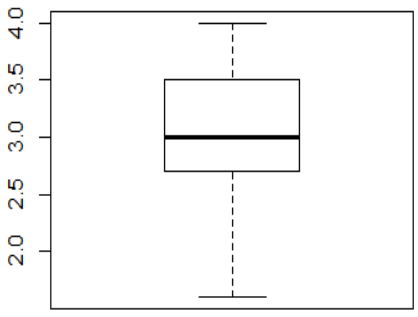
< Ex > Variable: Service Time (in minutes) Compare Distributions by Using boxplots.



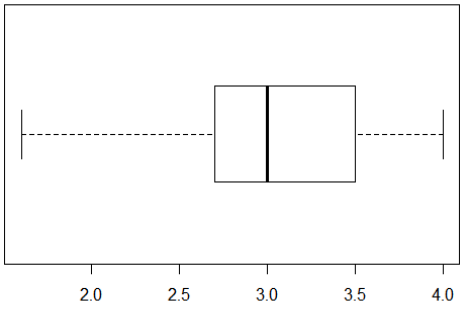
< Ex > Data with 455 students: Variable: GPA

GENDER	AGE	WORKHR	TATTOO	SHOES	EXERCISE	GPA	CLASS	CARAGE	MARRIED
FEMALE	18	20	NO	23	5	2.9	FRESHMAI	1	NO
FEMALE	18	16	NO	12	3	3.9	FRESHMAI	2	NO

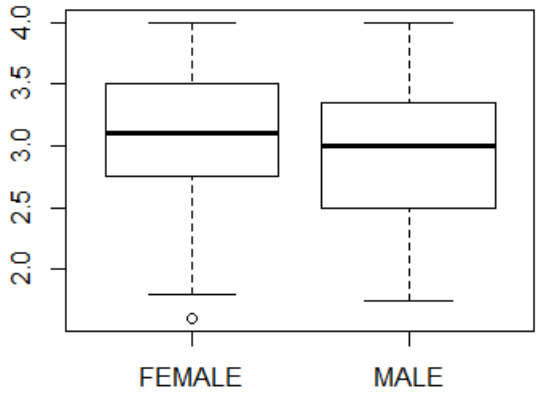
```
> boxplot(GPA)
```



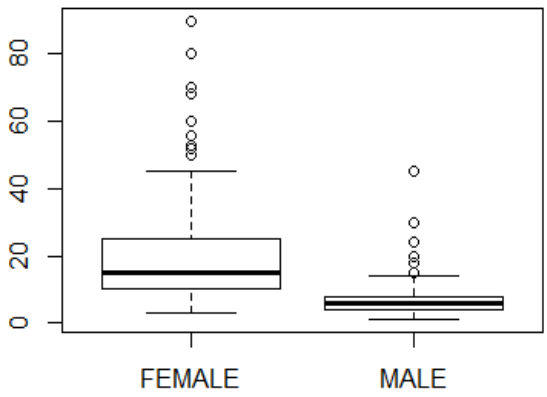

```
> boxplot(GPA, horizontal=T)
```



```
> boxplot(GPA~GENDER)
```




```
> boxplot(SHOES~GENDER)
```



Changing the Unit of Measurement: $y = ax + b$

A **linear transformation** changes the original variable x into the new variable y

< Ex > Linear Transformation: $y = ax + b$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{\sum(ax_i + b)}{n}$$

Original Data (x_i)	Transformed Data (y_i)
x_1	$y_1 = ax_1 + b$
x_2	$y_2 = ax_2 + b$
⋮	⋮
x_n	$y_n = ax_n + b$

Mean: $\bar{x}$	Mean: $\bar{y}$
Variance: s_x^2	Variance: s_y^2
SD: s_x	SD: s_y

$$s_y^2 = \frac{\sum(y_i - \bar{y})^2}{n-1} = \frac{\sum[(ax_i + b) - (a\bar{x} + b)]^2}{n-1}$$

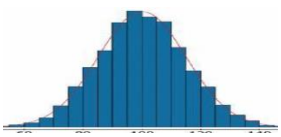
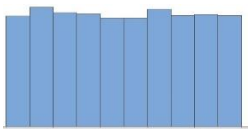
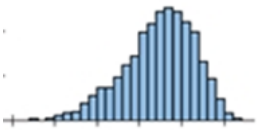
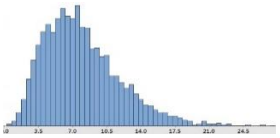
1.4 Density Curves and Normal Distributions

Density Functions or Density Curves or Distribution Functions or Probability Distributions

Probability Density Function $f(x)$

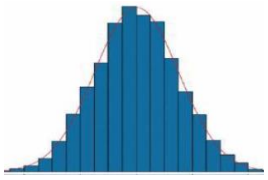
A **probability density function (pdf)** is an equation used to compute probabilities of continuous random variables. It must satisfy the following two properties:

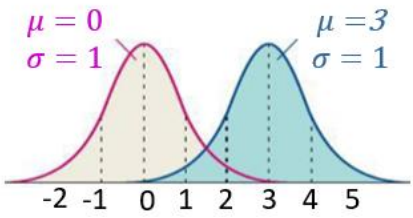
1. The total area under the graph of the equation over all possible values of the random variable must equal 1 i..e. $\int f(x)\,dx = 1$
2. The height of the graph of the equation must be greater than or equal to 0 for all possible values of the random variable i.e. $f(x) \geq 0$

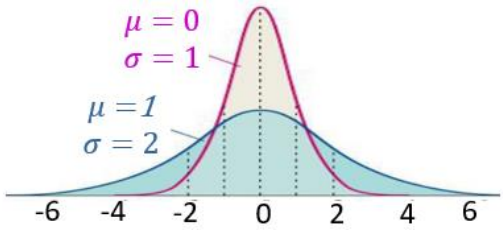





Normal Probability Distribution

A random variable is normally distributed, or has a normal distribution if its relative frequency histogram, has the shape of a normal curve

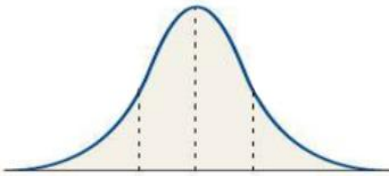






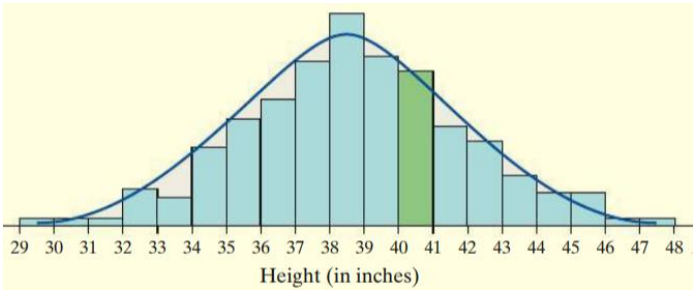
Properties of the Normal Density Curve

1. The normal curve is symmetric about its _____
2. The area under the normal curve to the right of mean equals the area under the curve to the left of mean
3. The total area under the normal curve is _____



< Ex > We are interested in the heights of a pediatrician's three-year-old female patients. The raw data indicate that the mean height of the patients is $\mu = 38.72$ inches with standard deviation $\sigma = 3.17$ inches.

The relative frequency distribution and the normal curve are shown below.



The area of the rectangle for heights between 40 and 50 inches

The area under the normal curve for heights between 40 and 50 inches

Area under a Normal Curve

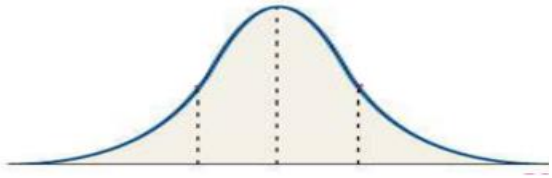
Suppose that a random variable X is normally distributed with mean μ and standard deviation σ. The area under the normal curve for any interval of values of the random variable X represents either

- The proportion of the population with the characteristic described by the interval of values or
- The probability that a randomly selected individual from the population will have the characteristic described by the interval of values

< Ex > The serum total cholesterol for males 20-29 years old is approximately normally distributed with mean $\mu = 180$ and standard deviation $\sigma = 36$.

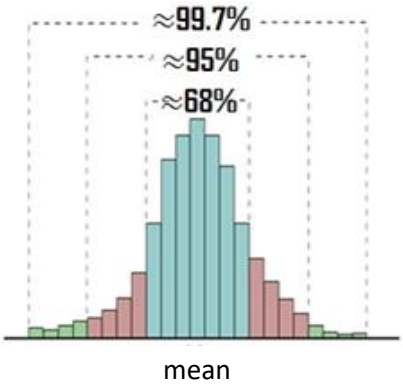
(a) The normal curve is shown. Label the mean.

(b) An individual with total cholesterol greater than 200 is considered to have high cholesterol. Shade the region under the normal curve.



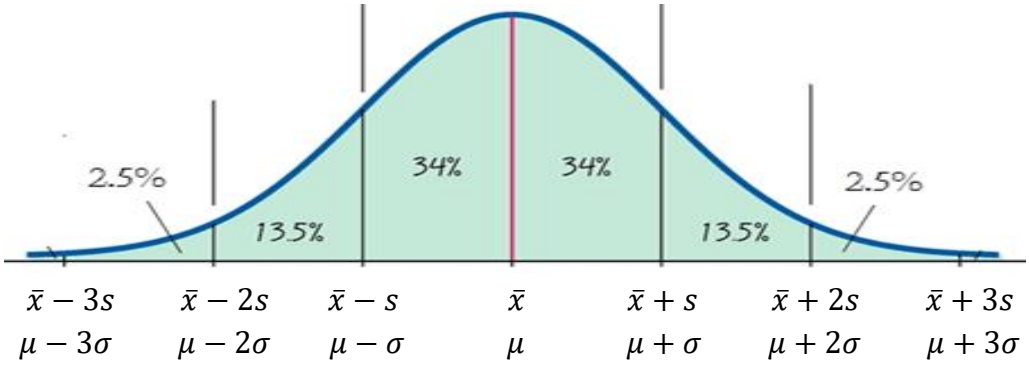
(c) Suppose that the area under the normal curve to the right of $x = 200$ is 0.2892. Interpret it.

The Empirical Rule

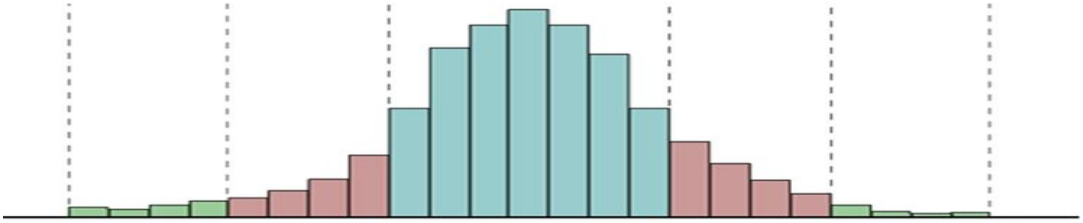


If data have a **distribution that is bell-shaped**, the **Empirical Rule** can be used to determine the percentage of data that will be within k **standard deviations of the mean**.

- About **68%** of values lie within 1 sd from the mean
- About **95%** of values lie within 2 sd from the mean
- About **99.7%** values lie within 3 sd from the mean



< Ex > The life spans of lizards have a bell-shaped distribution. The average lizard lives 3.5 years with standard deviation 0.5 years. Use the Empirical Rule to answer the questions.



(a) What is the percentage of lizards that live longer than 4.5 years?

(b) What is the percentage of lizards that live between 2.5 and 4.0 years?

< Ex > Which one lives longer?

The life span of domesticated cats has a bell-shaped distribution with mean 15.7 years and standard deviation 1.6 years. One cat lives to be 13 years.

The life span of a particular species of turtles has a bell-shaped distribution with mean 180 years and standard deviation 40 years. A turtle lives to be 112 years

Which one, a cat or a turtle, lives longer relative to their own species? → Will answer it later

The **z-score** represents the distance a data value is from the mean in terms of the number of standard deviations.

Population z-Score

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

Sample z=Score

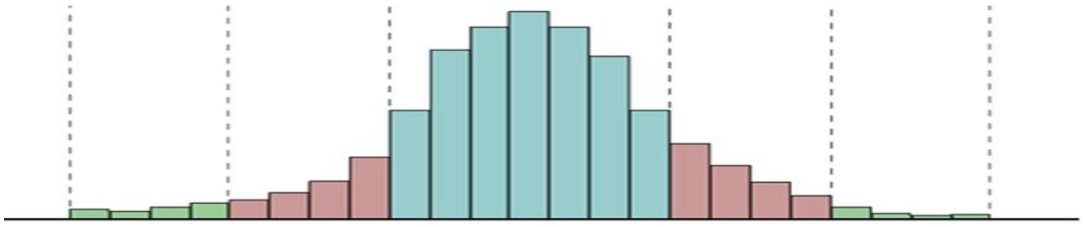
$$z = \frac{x - \bar{x}}{s}$$

That is, the s-score is

$$z = \frac{\text{value} - \text{mean}}{\text{stadanrd deviaion}}$$

< Ex > The life spans of lizards have a bell-shaped distribution. The average lizard lives 3.5 years with standard deviation 0.5 years. Find the z-score of the following

(a) 4.0 years **(b)** 2.7 years **(c)** 4.8 years **(d)** 1.3 years



years 3.5

z-score

(e) If a lizard lives shorter than 1.3 years, is it typical/normal or out of the norm? Explain.

< Ex > (continue) **Which one lives longer?**

The life span of domesticated cats has a bell-shaped distribution with mean 15.7 years and standard deviation 1.6 years. One cat lives to be 13 years. The life span of a particular species of turtles has a bell-shaped distribution with mean 180 years and standard deviation 40 years. A turtle lives to be 112 years

Which one, a cat or a turtle, lives longer relative to their own species?

Probability Using a Normal Distribution

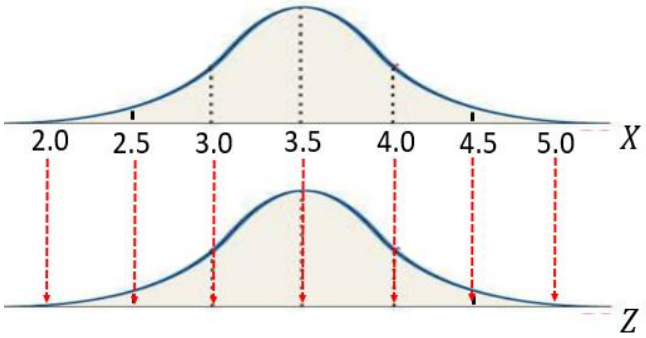
We use z-scores to help find the area under a normal curve by hand.

Standardizing a Normal Random Variable

Suppose that the random variable X has mean μ and standard deviation σ.

Then, the random variable $\boldsymbol{Z = \frac{X - \mu}{\sigma}}$ is normally distributed with mean $\boldsymbol{\mu = 0}$ and standard deviation $\boldsymbol{\sigma = 1}$. The random variable Z is said to have the **standard normal distribution**.

< Ex > Let $X =$ the life spans (in years) of lizards, then X is normally distributed with mean 3.5 years and standard deviation 0.5 years.

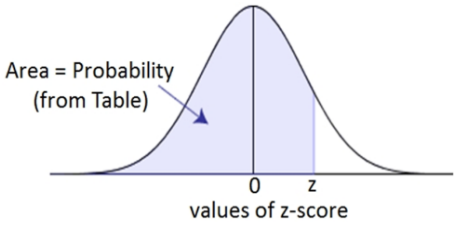


The area under the normal curve for lifespans (x) between 2.5 and 4 years

The area under the normal curve for values (z) between _____ and _____

<u>Table A</u>: Standard Normal Distribution Table

It gives the area under $N(0,1)$ to the left of a z-score



< Ex > Find the area under the $N(0,1)$ distribution.

(a) left of -0.41

$P(Z \leq -0.41)$

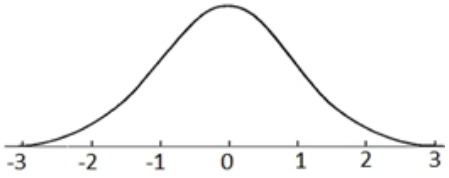
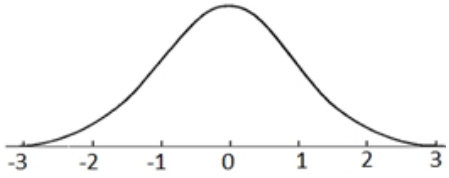


TABLE A Standard Normal probabilities (continued)

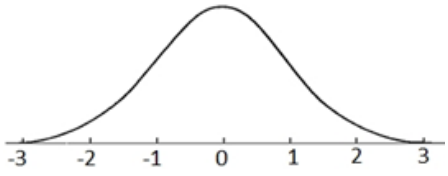
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121

(b) $P(Z > 1.12)$

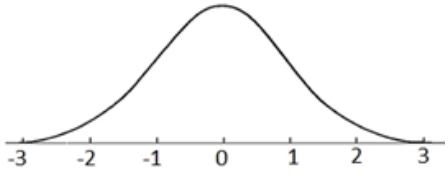


z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830

(c) $P(-0.41 < Z < 1.12)$

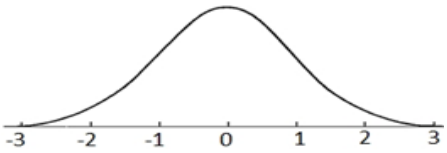


(d) $P(Z \leq -6.55)$



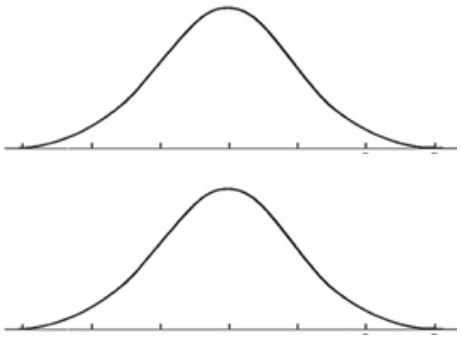
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002

(e) $P(Z > 8.02)$

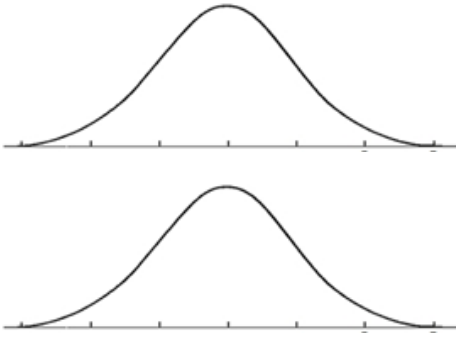


< Ex > The heights of three-year-old girls are normally distributed with mean 38.72 inches and standard deviation 3.17 inches.

(a) Find the probability that a randomly selected three-year-old girl is between 35 and 40 inches tall.

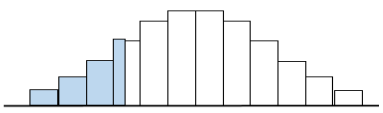


(b) Find the probability that a randomly selected three-year-old girl is shorter than 35 inches tall.



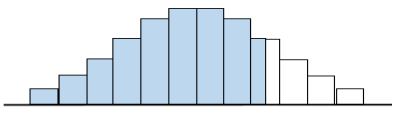
The ***k*th percentile** of a data set is a value such that k percent of the observations are less than or equal to the value.

30th Percentile



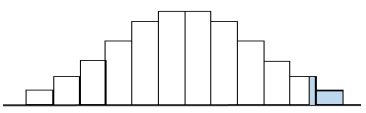
A

75th Percentile



B

Top 5%



Finding the Value of a Normal Random Variable

< Ex > Find the z score from a $N(0,1)$ distribution.

(a) Find the value z such that the area to its left is 0.90

That is, find z such that $P(Z < z) = 0.9$

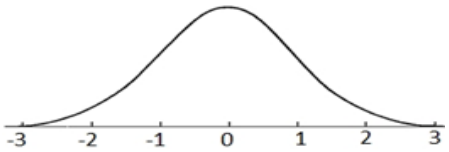
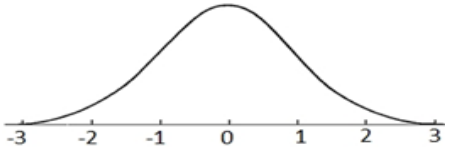


TABLE A Standard Normal probabilities (continued)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015

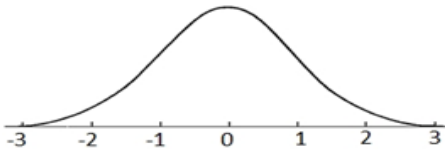
(b) Find z such that the area to its right is 0.03

That is, find z such that $P(Z > z) = 0.03$



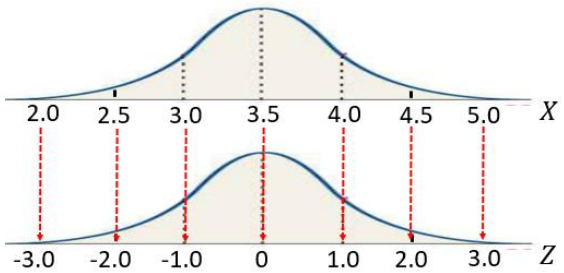
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706

(b) Find z such that $P(Z < z) = 0.05$



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455

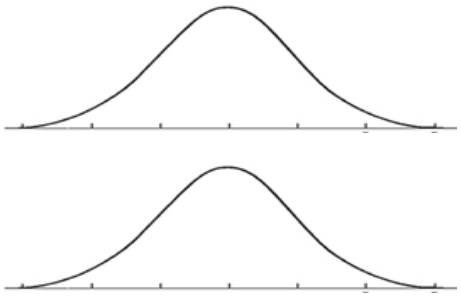
< Ex > Let $X =$ life spans (in years) of lizards. X is normally distributed with mean 3.5 and sd 0.5. Suppose the z-score of a lizard's life span is -1.28, what is the life span of this lizard?



Converting z-scores to Actual Measurement Scales

< Ex > The heights of three-year-old girls are normally distributed with mean 38.72 inches and standard deviation 3.17 inches.

Find the height of a three-year-old girl at the 20th percentile.



< Ex > The time required for Speedy Lube to complete an oil change service on an automobile approximately follows a normal distribution with a mean of 17 minutes and a standard deviation of 2.5 minutes.

Speedy Lube guarantees customers that the service will take no longer than 20 minutes. If it does take longer, the customer will receive the service for half price. What percent of customers receive the service for half price?