# Probability & Statistics for IS

## Chapter 1: Descriptive Statistics


**Lecturers**

Dr. Ghandi Manasra
Dr. Monjed H. Samuh

Palestine Polytechnic University
(ghandi@ppu.edu, monjedsamuh@ppu.edu)


Term 191

# Topics to be Covered

- Descriptive Statistics (Tables, Graphs, Measures)
- Introduction to R.
- T Test Procedures (One-sample t-test, Two-sample t-test).
- Univariate and Multivariate ANOVA.
- Correlation and Regression.
- Nonparametric Procedures.

# Table of Contents

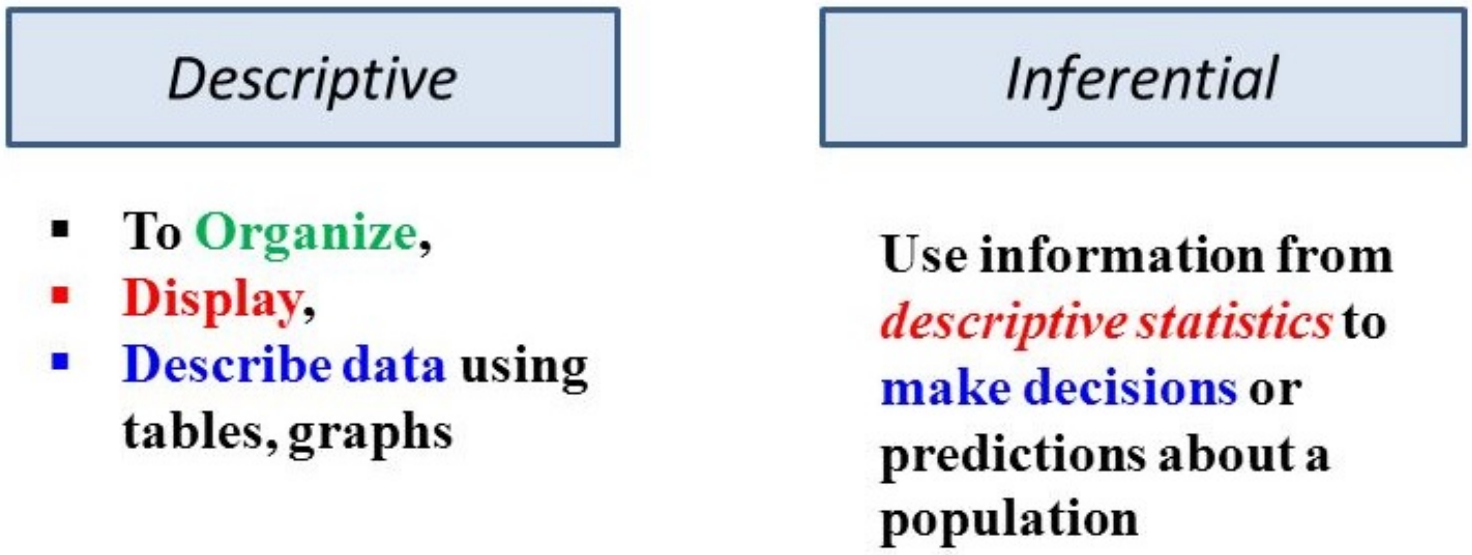
# Learning Objectives

After studying this chapter, the student will:

- Define statistics generally.
- Differentiate between the two branches of statistics.
- Identify types of data.
- Identify data level of measurement.
- Differentiate between population and sample data.
- Identify the four basic sampling techniques.
- Understand how data can be appropriately organized and displayed (Tables and Graphs).
- Understand how to reduce data sets into a few useful, descriptive measures (measures of central tendency, measures of dispersion, measures of position).

# What is statistics?

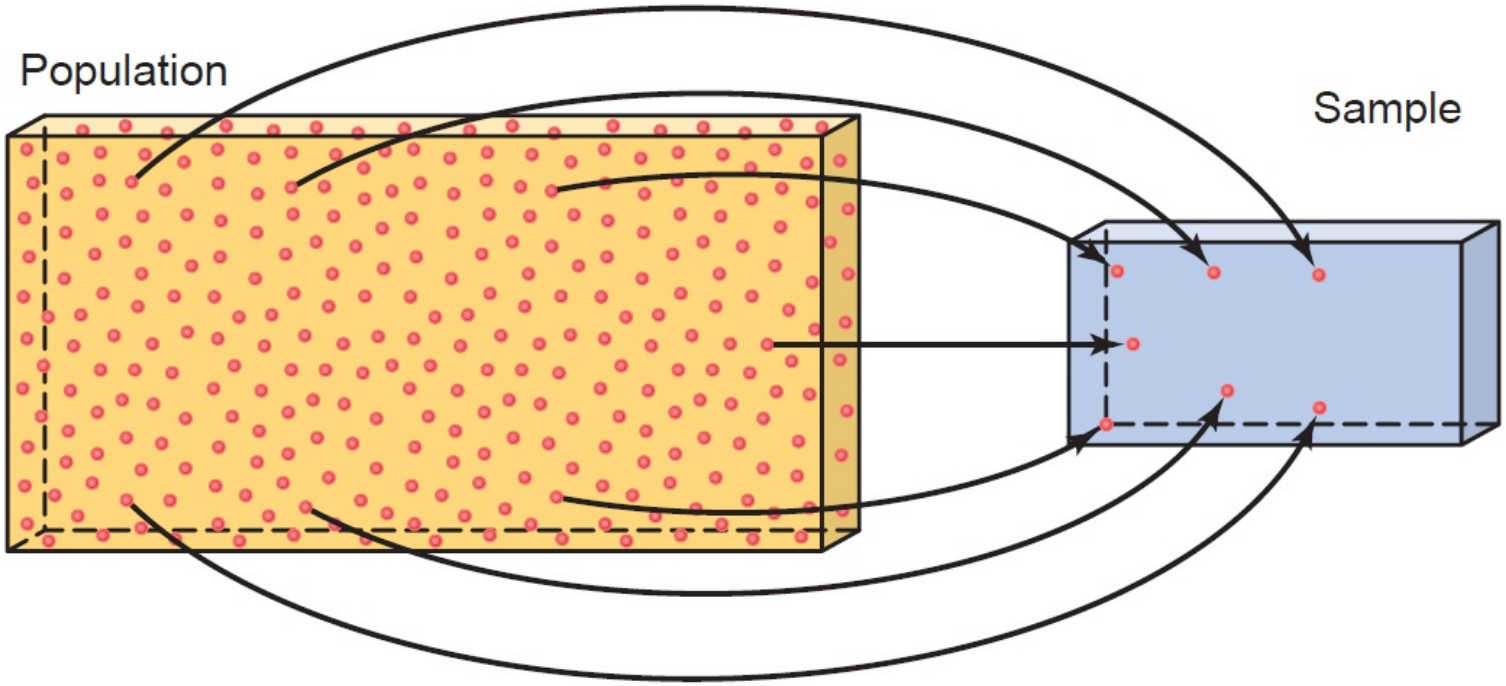
- The study of **statistics** explores the collection, organization, analysis, and interpretation of numerical data.
- The concepts of statistics may be applied to a number of fields that include business, psychology, and agriculture.
- Types of Statistics



# Population and Sample

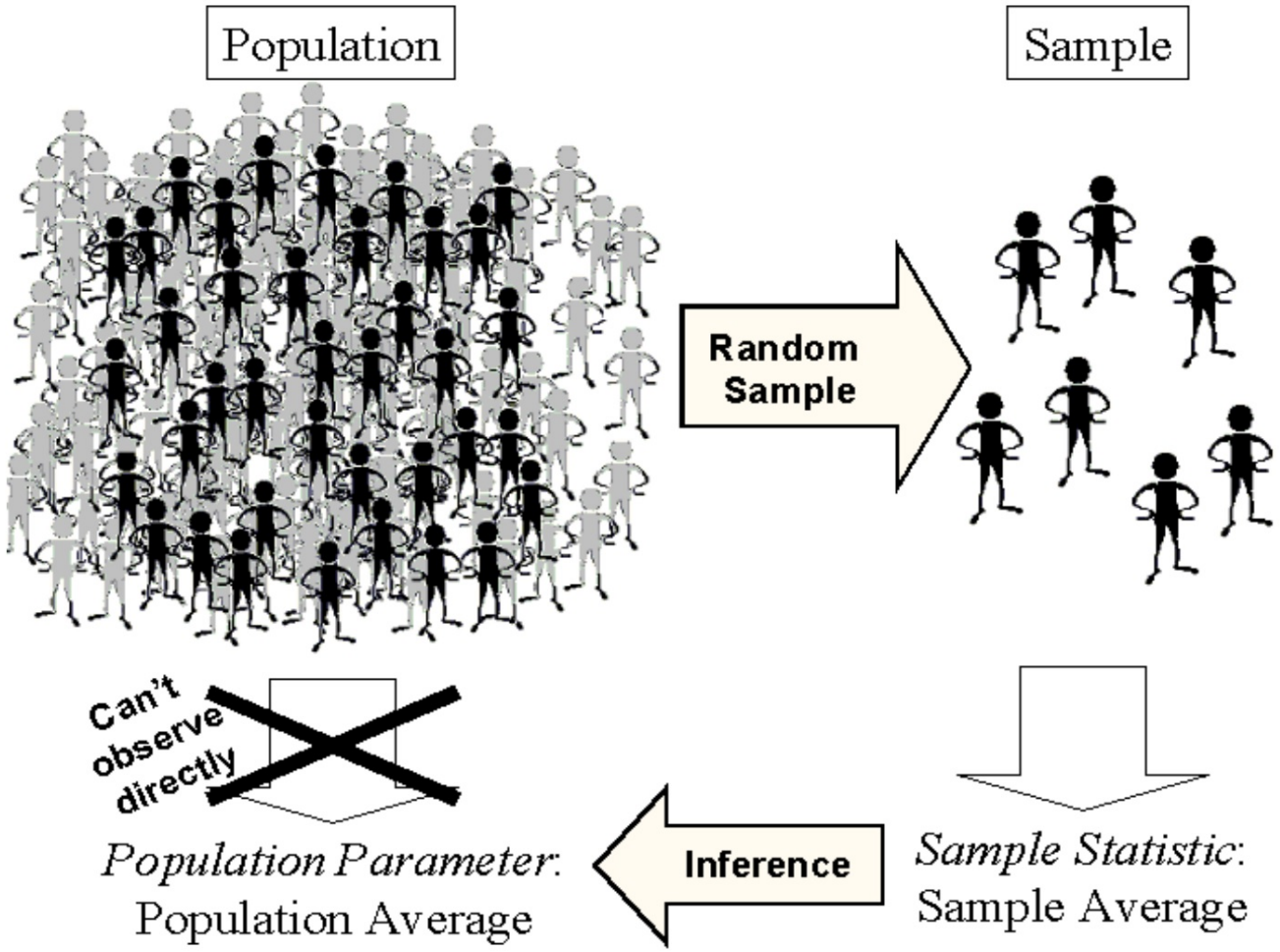
- A **population** is a complete set of elements (persons or objects) for which we have interest.

- A **sample** is a subset of the population.



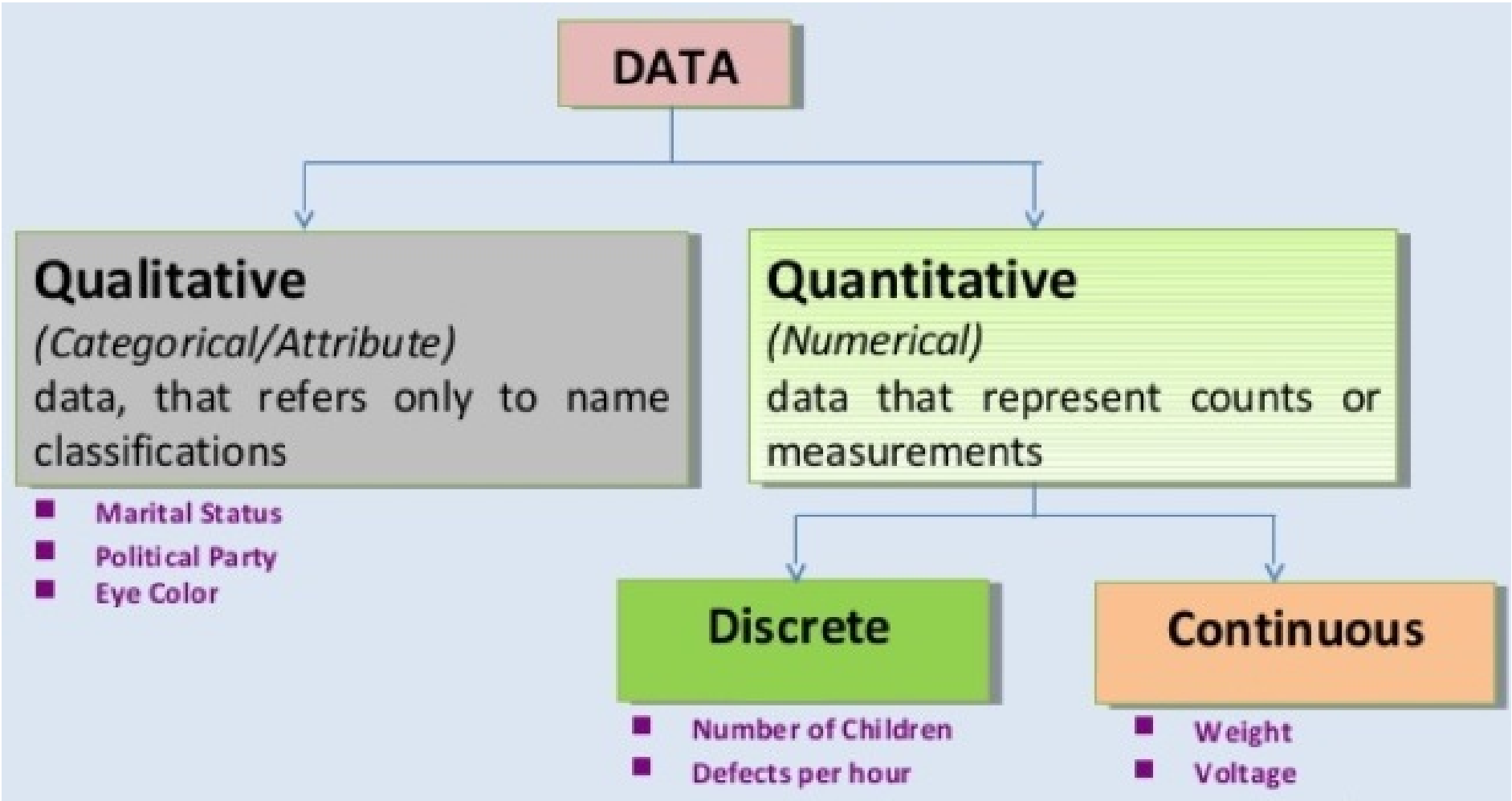
# Statistic and Parameter

- A **statistic** is a descriptive measure computed from the data of the sample.
- A **parameter** is a descriptive measure computed from the data of the population.



# Data

- The raw material of statistics is **data**.
- Data are the **quantities** (numbers) or **qualities** (attributes) measured or observed that are to be collected and/or analyzed.
- The word *data* is plural, *datum* is singular.
- A collection of data is often called a **data set** (singular).
- Sources of data: Data are obtained from
  1. Routinely kept records (Hospital and medical records).
  2. Surveys (Questionnaires).
  3. Experiments.
  4. Reports (Published records)


DATA
Qualitative
(Categorical/Attribute)
data, that refers only to name classifications
Marital Status
Political Party
Eye Color
Quantitative
(Numerical)
data that represent counts or measurements
Discrete
Number of Children
Defects per hour
Continuous
Weight
Voltage

# Data Measurement Scales

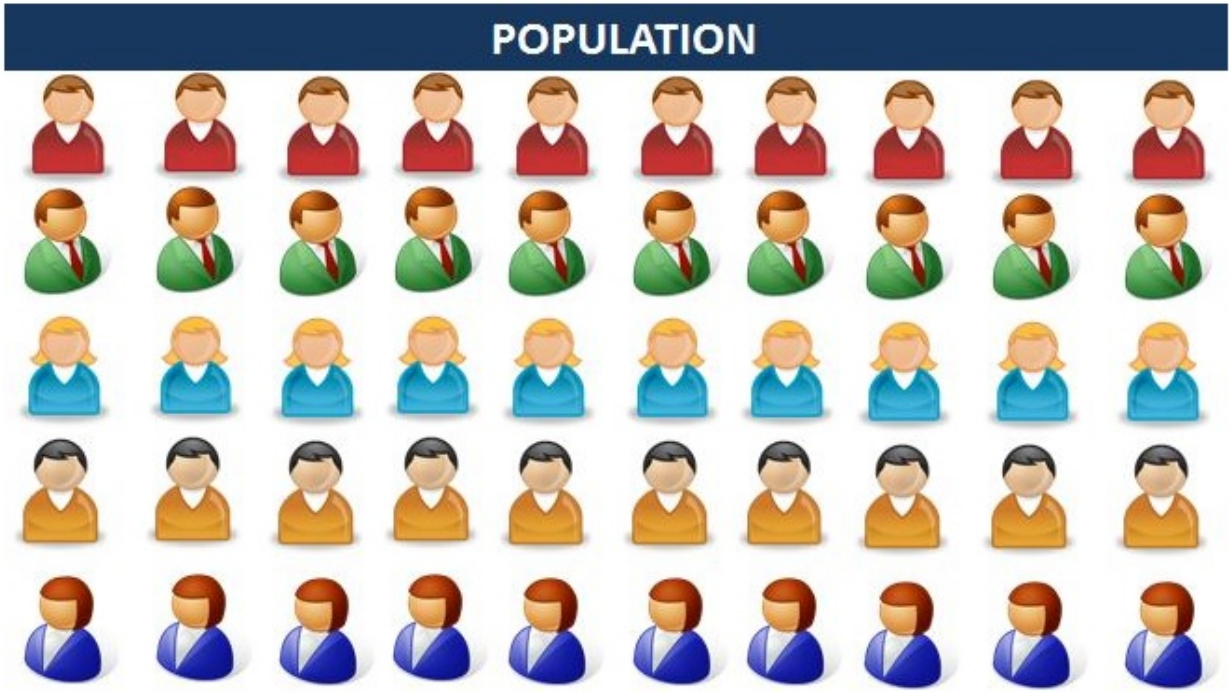
- Data can be classified as being on one of four scales:

| | < > | − | cx + ÷ | |
|---|---|---|---|---|
| *Scale* | *Order* | *Distance* | *True Zero* | *Examples* |
| **Nominal** | **no** | **no** | **no** | Color, Gender, Ethnicity, Country |
| **Ordinal** | **yes** | **no** | **no** | Rating scales, Rank orders |
| **Interval** | **yes** | **yes** | **no** | Time of day, Year, IQ, Likert scales |
| **Ratio** | **yes** | **yes** | **yes** | Age, Height, Weight, Rates |

# Sampling Techniques

- Two advantages of sampling:
  1. the cost is lower, and
  2. data collection is faster than measuring the entire population.
- Because the sample will be used to draw conclusions about the entire population, it should be a **representative sample**, that is, it should reflect as closely as possible the relevant characteristics of the population under consideration

# Representative Sample






# Sampling Techniques

**Sampling Methods (Probability Sampling Methods):**

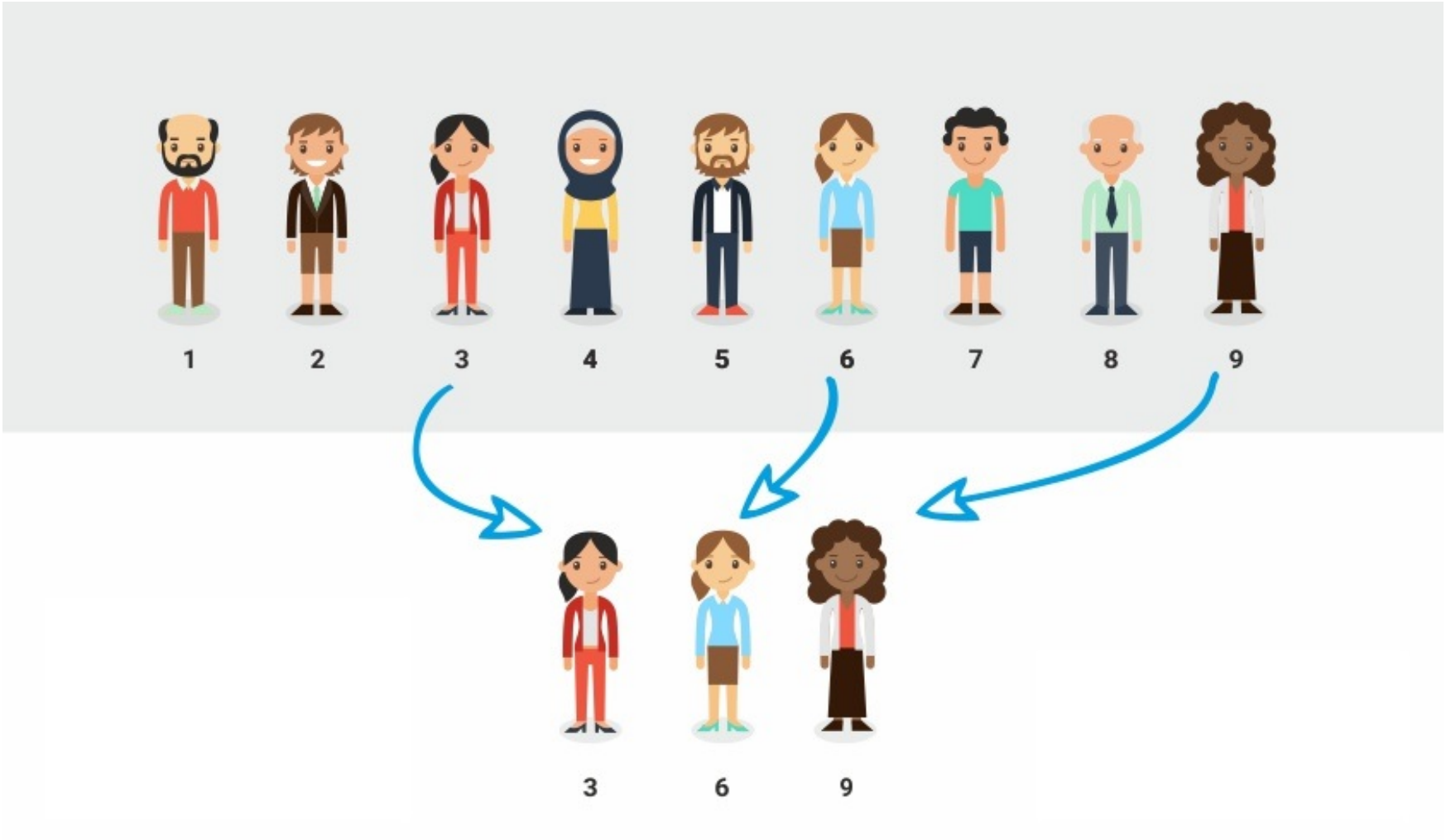
1. Simple Random Sampling
2. Systematic Sampling
3. Stratified Sampling
4. Cluster Sampling

# Simple Random Sampling

- If a sample of size $n$ is drawn from a population of size $N$ in such a way that every possible sample of size $n$ has the same chance of being selected, the sample is called a **simple random sample**.

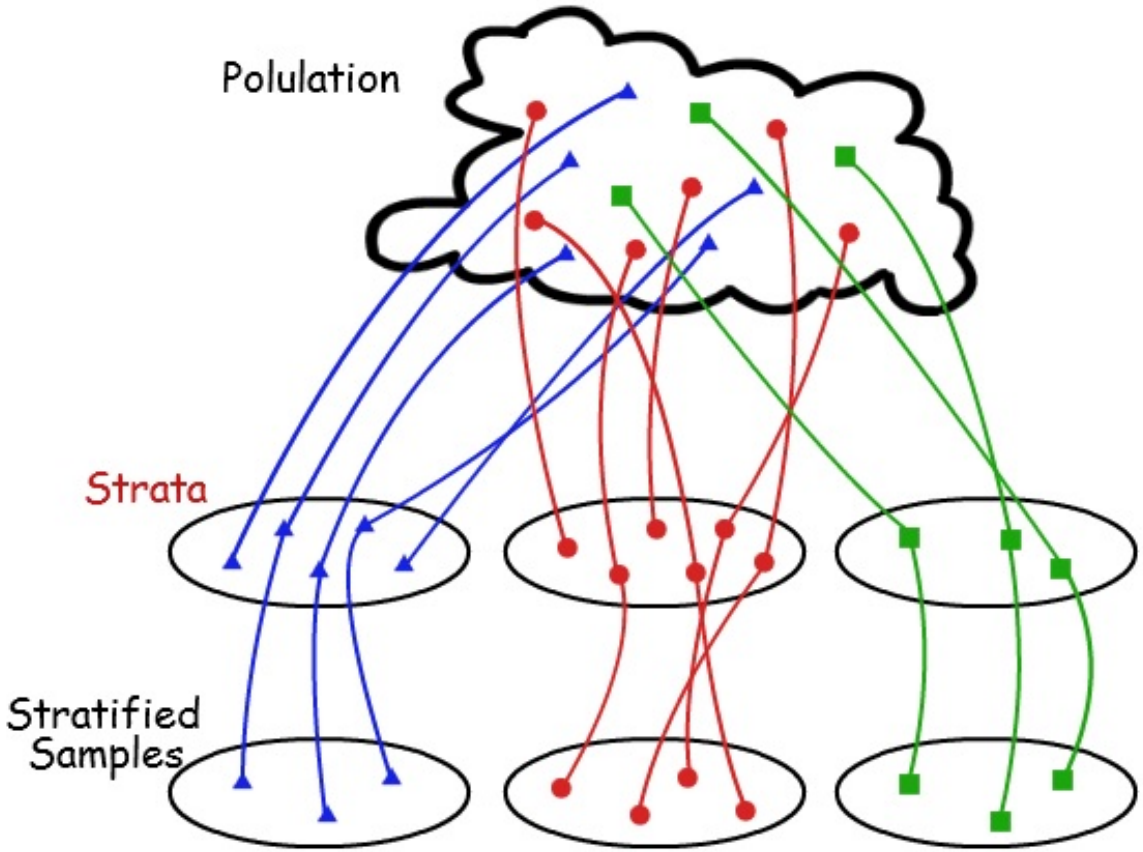
# Systematic Sampling

- Systematic samples are obtained by numbering each value in the population and then selecting the $k^{th}$ value.



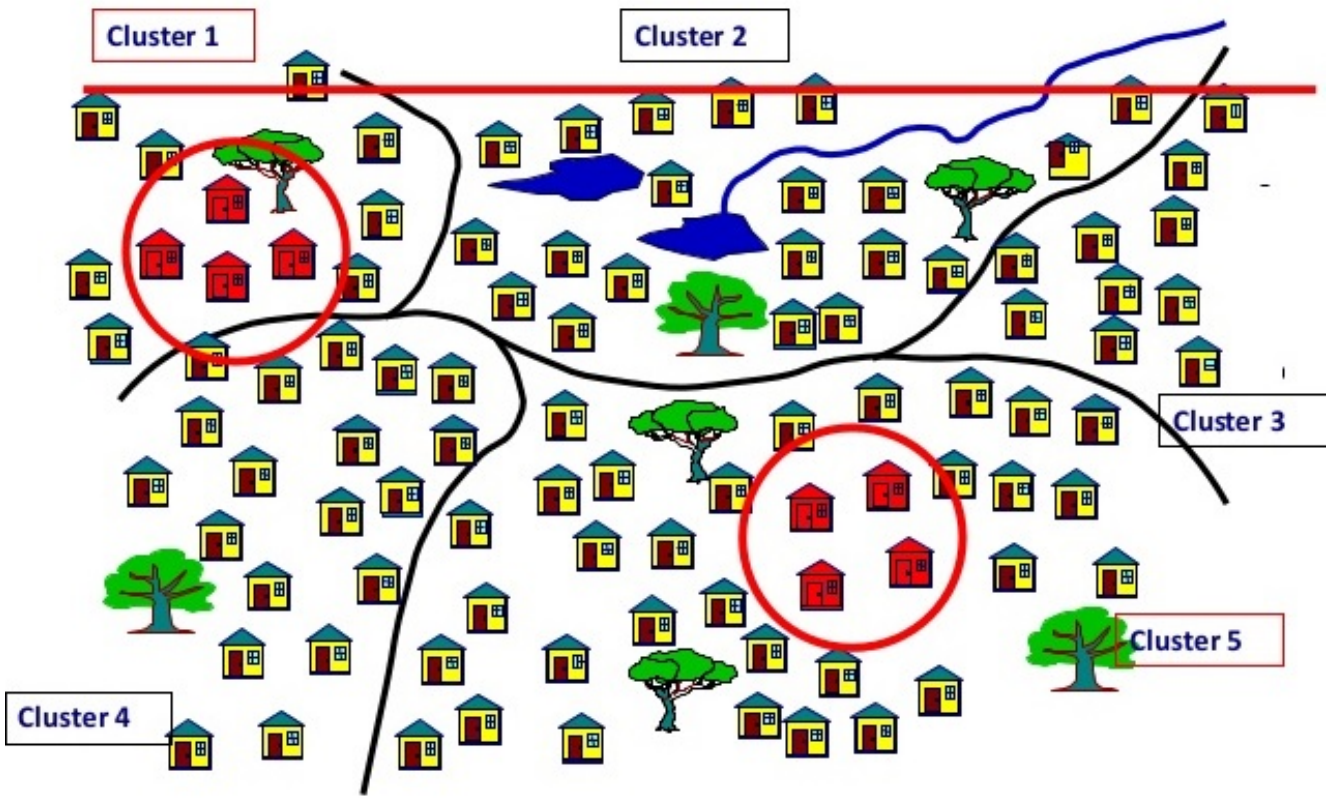
# Stratified Sampling

- A stratified sample has members from each segment of a population.
- This ensures that each segment from the population is represented.

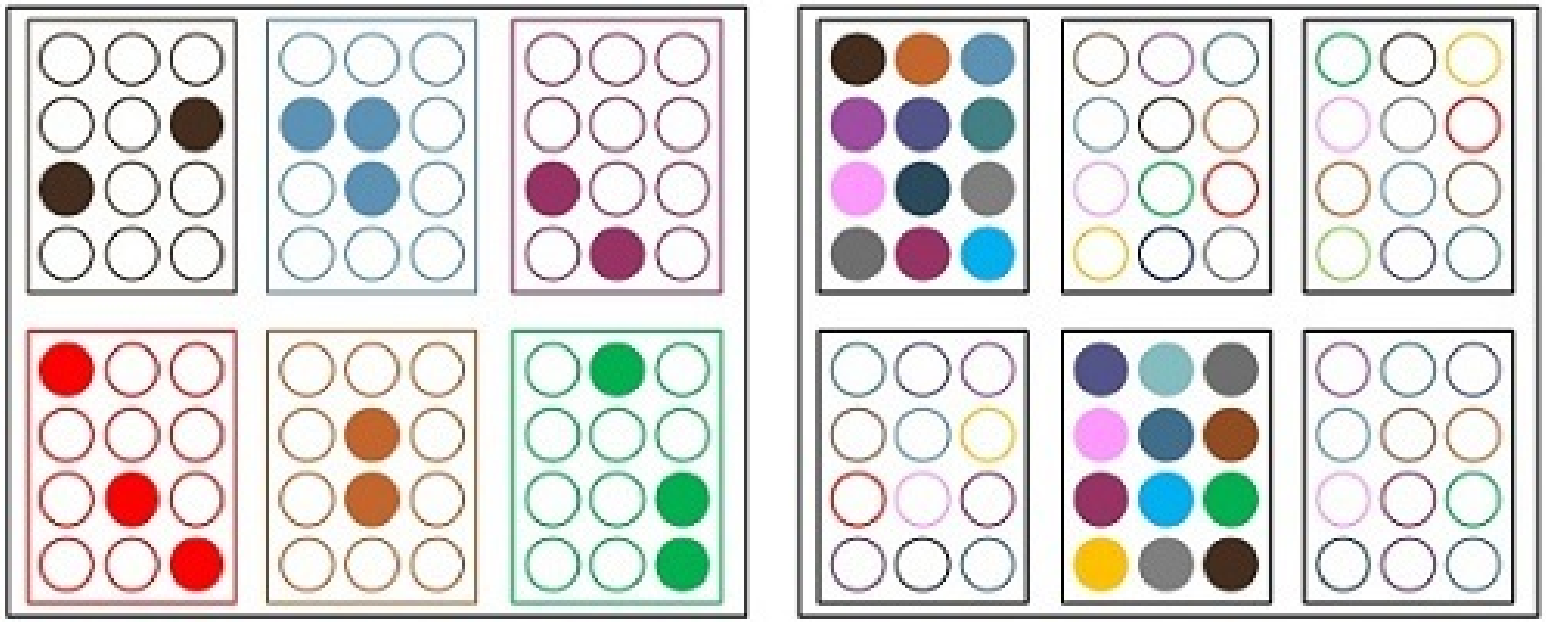


# Cluster Sampling

- Cluster samples are selected by dividing the population into groups and then taking samples of the groups.



# Difference between Cluster Sampling and Stratified Sampling



Stratified Sampling Vs Cluster Sampling

# Introduction to R: Basics

### basic arithmetic operations

```
> 2+3
[1] 5
> 2-3
[1] -1
> 2*3
[1] 6
> 2/3
[1] 0.6666667
> 2^3
[1] 8
```

### precedence of operators

```
> 4^2-3*2
[1] 10
> (4^2) - (3*2)
[1] 10
>
> -2--3
[1] 1
> -2 - -3
[1] 1
>
> 1 - 6 + 4
[1] -1
>
> 2^-3
[1] 0.125
```

### functions, arguments to functions

```
> log(100)
[1] 4.60517
> log(100, base=10)
[1] 2
> log(100, b=10)
[1] 2
```

### obtaining help

```
> help(log)
starting httpd help server ...
> ?log
```

# Introduction to R: Basics

**vectors**

```
> c(1,2,3,4)  # combine
[1] 1 2 3 4
>
> 1:4     # sequence operator
[1] 1 2 3 4
> 4:1
[1] 4 3 2 1
> -1:2    # note precedence
[1] -1  0  1  2
> seq(1,4)
[1] 1 2 3 4
> seq(2, 8, by=2)
[1] 2 4 6 8
> seq(0, 1, by=.1)
 [1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
> seq(0, 1, length=11)
 [1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
```

**vectorized arithmetic**

```
> c(1,2,3,4)/2
[1] 0.5 1.0 1.5 2.0
> c(1,2,3,4)/c(4,3,2,1)
[1] 0.2500000 0.6666667 1.5000000 4.0000000
> log(c(0.1,1,10,100), 10)
[1] -1  0  1  2
>
> c(1,2,3,4) + c(4,3)
[1] 5 5 7 7
> c(1,2,3,4) + c(4,3,2)
[1] 5 5 5 8
Warning message:
In c(1, 2, 3, 4) + c(4, 3, 2) :
```

# Introduction to R: Basics

**variables**

```
> x <- c(1,2,3,4)
> x
[1] 1 2 3 4
> x/2
[1] 0.5 1.0 1.5 2.0
> y <- sqrt(x)
> y
[1] 1.000000 1.414214 1.732051 2.000000
> x <- rnorm(100)
> x
  [1] -0.24060344 -0.29593451  0.15322553 -1.502
  [7] -1.70157666 -1.46867857 -1.71481489 -0.158
 [13]  0.39550288  1.18905999 -1.04930267  0.413
 [19]  0.82744597 -0.73162973  1.46864575 -0.124
 [25]  0.91688706 -0.30194873  2.39263916 -0.208
 [31] -0.58526232 -1.59496278  0.61702932  0.209
 [37] -0.09961610  0.24964521 -0.40260034  0.437
 [43] -3.29702414  0.95761763  0.06590933 -0.395
 [49]  0.37428035  2.24894202  1.53268485  0.762
 [55] -1.07393592  0.08195079  1.05490233  3.518
 [61]  0.19373456 -1.07652350 -0.80980120  0.624
 [67]  0.43990064  1.36759602  1.57151297  1.329
 [73]  1.37787508 -0.69587561  0.99337252  0.382
 [79] -0.50997428 -0.38551756  0.11701363 -0.292
 [85]  0.26968593  1.88209974  0.98094460 -0.731
 [91] -1.41831337  0.23258858 -0.93869222  0.672
 [97]  1.20794268  0.65164709  0.55659825  0.875
> summary(x)  # a "generic" function
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-3.2970 -0.4923  0.2210  0.1260  0.7851  3.5187
```

**basic indexing**

```
> x[21]
[1] 1.468646
> x[11:20]
 [1] -0.0911527  0.7941572  0.3955029  1.1890600
> x[-(11:100)]    # careful here!
 [1] -0.2406034 -0.2959345  0.1532255 -1.5028443
>
> z <- x[1:10]
> z
 [1] -0.2406034 -0.2959345  0.1532255 -1.5028443
> z < -0.5
 [1] FALSE FALSE FALSE  TRUE  TRUE FALSE  TRUE
> z > 0.5
 [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE F
> z < -0.5 | z > 0.5   # | is vectored "or", & i
 [1] FALSE FALSE FALSE  TRUE  TRUE FALSE  TRUE
> abs(z) > 0.5
 [1] FALSE FALSE FALSE  TRUE  TRUE FALSE  TRUE
> z[abs(z) > 0.5] # indexing by a logical vector
[1] -1.502844 -0.646307 -1.701577 -1.468679 -1.7
```

**user-defined functions**

```
> mean(x)
[1] 0.1260197
> sum(x)/length(x)
[1] 0.1260197
>
> my.mean <- function(x) sum(x)/length(x)
> my.mean(x)
[1] 0.1260197
> my.mean(1:100)
[1] 50.5
```

# Introduction to R: Basics

**Vectors, Matrices, and Arrays**

```
> x5 <- array(x1, dim=c(2, 2, 2, 2)) # 2*2*2*2 array
> x1 <- 1:16 # Vector
> x2 <- matrix(x1, nrow=4, ncol=4) # 4*4 matrix
> x3 <- matrix(x1, nrow=4, ncol=4, byrow=TRUE) # 4*4 matrix by rows
> x4 <- array(x1, dim=c(4, 2, 2)) # 4*2*2 array
> x5 <- array(x1, dim=c(2, 2, 2, 2)) # 2*2*2*2 array
```

```
> x2
     [,1] [,2] [,3] [,4]
[1,]    1    5    9   13
[2,]    2    6   10   14
[3,]    3    7   11   15
[4,]    4    8   12   16
```

```
> x3
     [,1] [,2] [,3] [,4]
[1,]    1    2    3    4
[2,]    5    6    7    8
[3,]    9   10   11   12
[4,]   13   14   15   16
```

```
> x4
, , 1

     [,1] [,2]
[1,]    1    5
[2,]    2    6
[3,]    3    7
[4,]    4    8

, , 2

     [,1] [,2]
[1,]    9   13
[2,]   10   14
[3,]   11   15
[4,]   12   16
```

```
> x5
, , 1, 1

     [,1] [,2]
[1,]    1    3
[2,]    2    4

, , 2, 1

     [,1] [,2]
[1,]    5    7
[2,]    6    8

, , 1, 2

     [,1] [,2]
[1,]    9   11
[2,]   10   12

, , 2, 2

     [,1] [,2]
[1,]   13   15
[2,]   14   16
```

# Introduction to R: Basics

**Creating Data Frames**

```
> Height <- c(156, 140, 172)
> Weight <- c(60, 55, 75)
> rbind(Height, Weight)
       [,1] [,2] [,3]
Height  156  140  172
Weight   60   55   75
> cbind(Height, Weight)
     Height Weight
[1,]    156     60
[2,]    140     55
[3,]    172     75
> mydata <- data.frame(Height, Weight)
> mydata
  Height Weight
1    156     60
2    140     55
3    172     75
> dim(mydata)
[1] 3 2
```

**Sampling Data**

```
> sample(1:10)
 [1]  2  9  4  1  5 10  6  8  7  3
> sample(c("a", "b", "c", "d", "e"))
[1] "e" "d" "a" "b" "c"
> sample(1:100, size=5)
[1] 44 63 12  9 97
> sample(0:1, size=10, replace=T)
 [1] 1 1 1 1 1 0 0 0 0 1
>
> rnorm(5)
[1]  0.2032140 -0.2836550  0.4045802  1.37
> rnorm(5, mean=10, sd=2)
[1]  6.307223 11.346789 11.097564  9.74761


> round(rnorm(5),2)
[1] -0.32 -1.27  0.64  0.64  1.20
> round(rnorm(5, mean=10, sd=2),3)
[1] 11.068  9.763  9.702 11.178 10.045
```

# Raw Data

- The ages of 34 patients who suffered stress strokes were as follows:

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 29 | 30 | 36 | 41 | 45 | 50 | 57 | 61 | 28 | 50 | 36 | 58 |
| 60 | 38 | 36 | 47 | 40 | 32 | 58 | 46 | 61 | 40 | 55 | 32 |
| 61 | 56 | 45 | 46 | 62 | 36 | 38 | 40 | 50 | 27 | | |

- The blood types of a random sample of 50 patients were as follows:

| | | | | |
|---|---|---|---|---|
| O | O | A | A | O |
| B | O | B | A | O |
| AB | B | A | B | AB |
| O | O | A | A | O |
| AB | O | A | B | A |
| O | A | A | O | A |
| O | A | O | AB | A |
| O | B | A | A | O |
| O | O | O | A | O |
| O | A | O | A | O |

**Data in raw form are usually not easy to use for decision making.**

# Ranked Data

- It is just a sorted list of data.

```
> x <- c(29,30,36,41,45,50,57,61,28,50,36,58,60,38,36,47,40,32,
         58,46,61,40,55,32,61,56,45,46,62,36,38,40,50,27)
>
> sort(x)
 [1] 27 28 29 30 32 32 36 36 36 36 38 38 40 40 40 41 45 45 46 46 47 50 50 50 55
[26] 56 57 58 58 60 61 61 61 62
```

1. Shows range (min to max).
2. Provides some signals about variability within the range.
3. May help identify outliers (unusual observations).

- If the data set is large, the ordered array is less useful.

# Frequency Distributions: Qualitative Data

- One type of table that is commonly used to evaluate data is known as a **frequency distribution**.
- For nominal and ordinal data, a frequency distribution consists of a set of classes or categories along with the numerical counts that correspond to each one.
- Blood types of a random sample of 50 patients

| Blood type | Frequency |
|---|---|
| A | 18 |
| B | 6 |
| AB | 4 |
| O | 22 |
| SUM | 50 |

# Frequency Distributions: Qualitative Data

```
> BT <- c("O","O","A","A","O","B","O","B","A","O","AB","B","A","B","AB","O","O","A","A","O",
          "AB","O","A","B","A","O","A","A","O","A","O","A","O","AB","A","O","B","A","A","O",
          "O","O","O","A","O","O","A","O","A","O")
> length(BT)
[1] 50
>
> sort(BT)
 [1] "A"  "A"  "A"  "A"  "A"  "A"  "A"  "A"  "A"  "A"  "A"  "A"  "A"  "A"  "A"
[16] "A"  "A"  "A"  "AB" "AB" "AB" "AB" "B"  "B"  "B"  "B"  "B"  "B"  "O"  "O"
[31] "O"  "O"  "O"  "O"  "O"  "O"  "O"  "O"  "O"  "O"  "O"  "O"  "O"  "O"  "O"
[46] "O"  "O"  "O"  "O"  "O"
>
> table(BT)
BT
 A AB  B  O
18  4  6 22
> as.data.frame(table(BT))
  BT Freq
1  A   18
2 AB    4
3  B    6
4  O   22
>
>
> tab <- as.data.frame(table(BT))
> names(tab)[1] <- "BloodType"
> tab
  BloodType Freq
1         A   18
2        AB    4
3         B    6
4         O   22
```

# Frequency Distributions: Quantitative Data

- To display discrete or continuous data in the form of a frequency distribution, we must break down the range of values of the observations into a series of distinct, nonoverlapping intervals.
- If there are too many intervals, the summary is not much of an improvement over the raw data.
- If there are too few, a great deal of information is lost.
- Although it is not necessary to do so, intervals are often constructed so that they all have equal widths; this facilitates comparisons among the classes.
- Once the upper and lower limits for each interval have been selected, the number of observations whose values fall within each pair of limits is counted, and the results are arranged as a table.

# Frequency Distributions: Quantitative Data

- Ages of 34 patients who suffered stress strokes

```
> x <- c(29,30,36,41,45,50,57,61,28,50,36,58,60,38,36,47,40,32,
+ 58,46,61,40,55,32,61,56,45,46,62,36,38,40,50,27)
>
> range(x)
[1] 27 62
>
> breaks <- seq(25,65, by=5)
> breaks
[1] 25 30 35 40 45 50 55 60 65
>
> cuts <- cut(x, breaks)
> cuts
 [1] (25,30] (25,30] (35,40] (40,45] (40,45] (45,50] (55,60] (60,65] (25,30]
[10] (45,50] (35,40] (55,60] (55,60] (35,40] (35,40] (45,50] (35,40] (30,35]
[19] (55,60] (45,50] (60,65] (35,40] (50,55] (30,35] (60,65] (55,60] (40,45]
[28] (45,50] (60,65] (35,40] (35,40] (35,40] (45,50] (25,30]
Levels: (25,30] (30,35] (35,40] (40,45] (45,50] (50,55] (55,60] (60,65]
>
> FreqTab <- table(cuts)
> FreqTab
cuts
(25,30] (30,35] (35,40] (40,45] (45,50] (50,55] (55,60] (60,65]
      4       2       9       3       6       1       5       4
> cbind(FreqTab)
        FreqTab
(25,30]       4
(30,35]       2
(35,40]       9
(40,45]       3
(45,50]       6
(50,55]       1
(55,60]       5
(60,65]       4
```

# Frequency Distributions: Other Examples

***TABLE 2.4***

Cases of Kaposi's sarcoma for the first 2560 AIDS patients reported to the Centers for Disease Control in Atlanta, Georgia

| Kaposi's Sarcoma | Number of Individuals |
|---|---|
| Yes | 246 |
| No | 2314 |

***TABLE 2.5***

Cigarette consumption per person aged 18 or older, United States, 1900–1990

| Year | Number of Cigarettes |
|---|---|
| 1900 | 54 |
| 1910 | 151 |
| 1920 | 665 |
| 1930 | 1485 |
| 1940 | 1976 |
| 1950 | 3522 |
| 1960 | 4171 |
| 1970 | 3985 |
| 1980 | 3851 |
| 1990 | 2828 |

***TABLE 2.6***

Absolute frequencies of serum cholesterol levels for 1067 U.S. males, aged 25 to 34 years, 1976–1980

| Cholesterol Level (mg/100 ml) | Number of Men |
|---|---|
| 80–119 | 13 |
| 120–159 | 150 |
| 160–199 | 442 |
| 200–239 | 299 |
| 240–279 | 115 |
| 280–319 | 34 |
| 320–359 | 9 |
| 360–399 | 5 |
| Total | 1067 |

# Relative Frequency Distributions

- It is sometimes useful to know the proportion of values that fall into a given interval in a frequency distribution rather than the absolute number.
- The **relative frequency** for an interval is the **proportion** of the total number of observations that appears in that interval.
- The relative frequency is computed by dividing the number of values within an interval by the total number of values in the table.
- The proportion can be left as it is, or it can be multiplied by 100% to obtain the percentage of values in the interval.
- The relative frequencies for all intervals in a table sum to 100%.

| Cholesterol Level (mg/100 ml) | Number of Men | Relative Frequency (%) |
|---|---|---|
| 80–119 | 13 | 1.2 |
| 120–159 | 150 | 14.1 |
| 160–199 | 442 | 41.4 |
| 200–239 | 299 | 28.0 |
| 240–279 | 115 | 10.8 |
| 280–319 | 34 | 3.2 |
| 320–359 | 9 | 0.8 |
| 360–399 | 5 | 0.5 |
| Total | 1067 | 100.0 |

# Relative Frequency Distributions

- Relative frequencies are useful for comparing sets of data that contain unequal numbers of observations.

| Cholesterol Level (mg/100 ml) | Ages 25–34 | | Ages 55–64 | |
|---|---|---|---|---|
| | Number of Men | Relative Frequency (%) | Number of Men | Relative Frequency (%) |
| 80–119 | 13 | 1.2 | 5 | 0.4 |
| 120–159 | 150 | 14.1 | 48 | 3.9 |
| 160–199 | 442 | 41.4 | 265 | 21.6 |
| 200–239 | 299 | 28.0 | 458 | 37.3 |
| 240–279 | 115 | 10.8 | 281 | 22.9 |
| 280–319 | 34 | 3.2 | 128 | 10.4 |
| 320–359 | 9 | 0.8 | 35 | 2.9 |
| 360–399 | 5 | 0.5 | 7 | 0.6 |
| Total | 1067 | 100.0 | 1227 | 100.0 |

- Because there are more men in the older age group, it is inappropriate to compare the columns of absolute frequencies for the two sets of males.
- Comparing the relative frequencies is meaningful, however.
- We can see that in general, the older men have higher serum cholesterol levels than the younger men; the younger men have a greater proportion of observations in each of the intervals below 200 mg/100 ml, whereas the older men have a greater proportion in each class above this value.

# Cumulative Relative Frequency Distributions

- The **cumulative relative frequency** for an interval is the percentage of the total number of observations that have a value less than or equal to the upper limit of the interval.
- The cumulative relative frequency is calculated by summing the relative frequencies for the specified interval and all previous ones.

| Cholesterol Level (mg/100 ml) | Ages 25–34 | | Ages 55–64 | |
|---|---|---|---|---|
| | Relative Frequency (%) | Cumulative Relative Frequency (%) | Relative Frequency (%) | Cumulative Relative Frequency (%) |
| 80–119 | 1.2 | 1.2 | 0.4 | 0.4 |
| 120–159 | 14.1 | 15.3 | 3.9 | 4.3 |
| 160–199 | 41.4 | 56.7 | 21.6 | 25.9 |
| 200–239 | 28.0 | 84.7 | 37.3 | 63.2 |
| 240–279 | 10.8 | 95.5 | 22.9 | 86.1 |
| 280–319 | 3.2 | 98.7 | 10.4 | 96.5 |
| 320–359 | 0.8 | 99.5 | 2.9 | 99.4 |
| 360–399 | 0.5 | 100.0 | 0.6 | 100.0 |

- For example, 56.7% of the 25- to 34-year-olds have a serum cholesterol level less than or equal to 199 mg/100 ml, whereas only 25.9% of the 55- to 64- year-olds fall into this category.

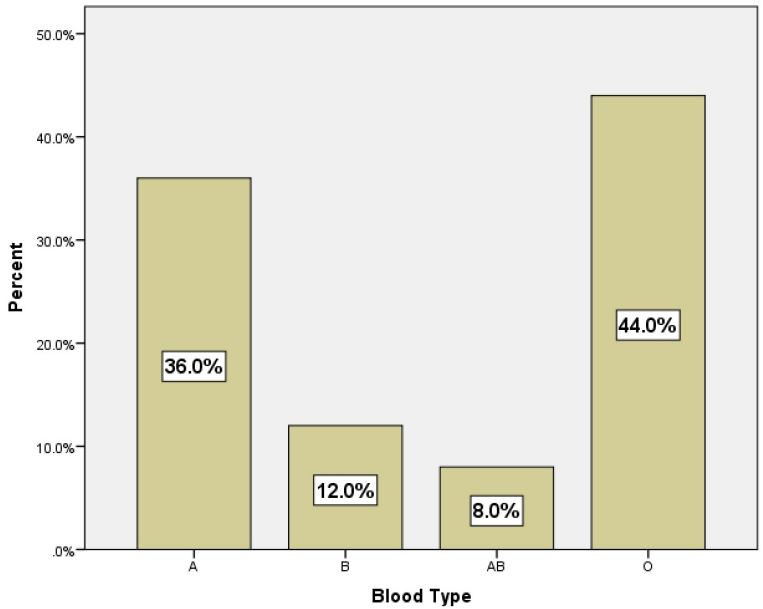
# Cumulative Relative Frequency Distributions

- Ages of 34 patients who suffered stress strokes

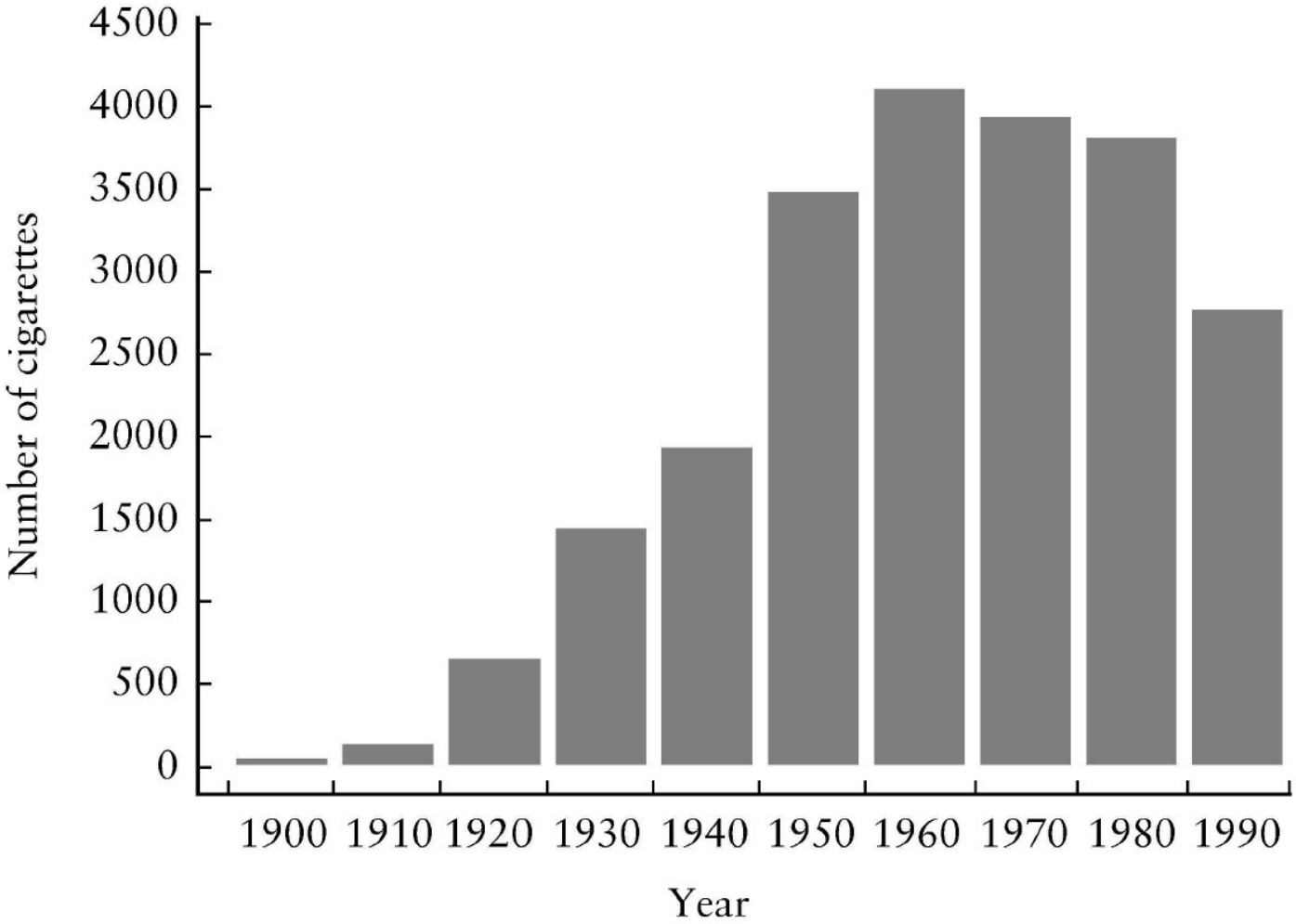
```
> RelFreq <- prop.table(FreqTab)
> RelFreq
cuts
   (25,30]    (30,35]    (35,40]    (40,45]    (45,50]    (50,55]    (55,60]    (60,65]
0.11764706 0.05882353 0.26470588 0.08823529 0.17647059 0.02941176 0.14705882 0.11764706
>
> CumFreq <- cumsum(FreqTab)
> CumFreq
(25,30] (30,35] (35,40] (40,45] (45,50] (50,55] (55,60] (60,65]
      4       6      15      18      24      25      30      34
>
> CumRelFreq <- cumsum(RelFreq)
> CumRelFreq
  (25,30]   (30,35]   (35,40]   (40,45]   (45,50]   (50,55]   (55,60]   (60,65]
0.1176471 0.1764706 0.4411765 0.5294118 0.7058824 0.7352941 0.8823529 1.0000000
>
> cbind(FreqTab, RelFreq, CumFreq, CumRelFreq)
        FreqTab    RelFreq CumFreq CumRelFreq
(25,30]       4 0.11764706       4  0.1176471
(30,35]       2 0.05882353       6  0.1764706
(35,40]       9 0.26470588      15  0.4411765
(40,45]       3 0.08823529      18  0.5294118
(45,50]       6 0.17647059      24  0.7058824
(50,55]       1 0.02941176      25  0.7352941
(55,60]       5 0.14705882      30  0.8823529
(60,65]       4 0.11764706      34  1.0000000
>
```

# Bar Charts

- **Bar charts** are a popular type of graph used to display a frequency distribution for **nominal or ordinal data**.
- In a bar chart, the various categories into which the observations fall are presented along a horizontal axis.
- A vertical bar is drawn above each category such that the height of the bar represents either the frequency or the relative frequency of observations within that class.
- The bars should be of **equal width** and **separated from one another** so as not to imply continuity.



# Bar Charts: Other Example

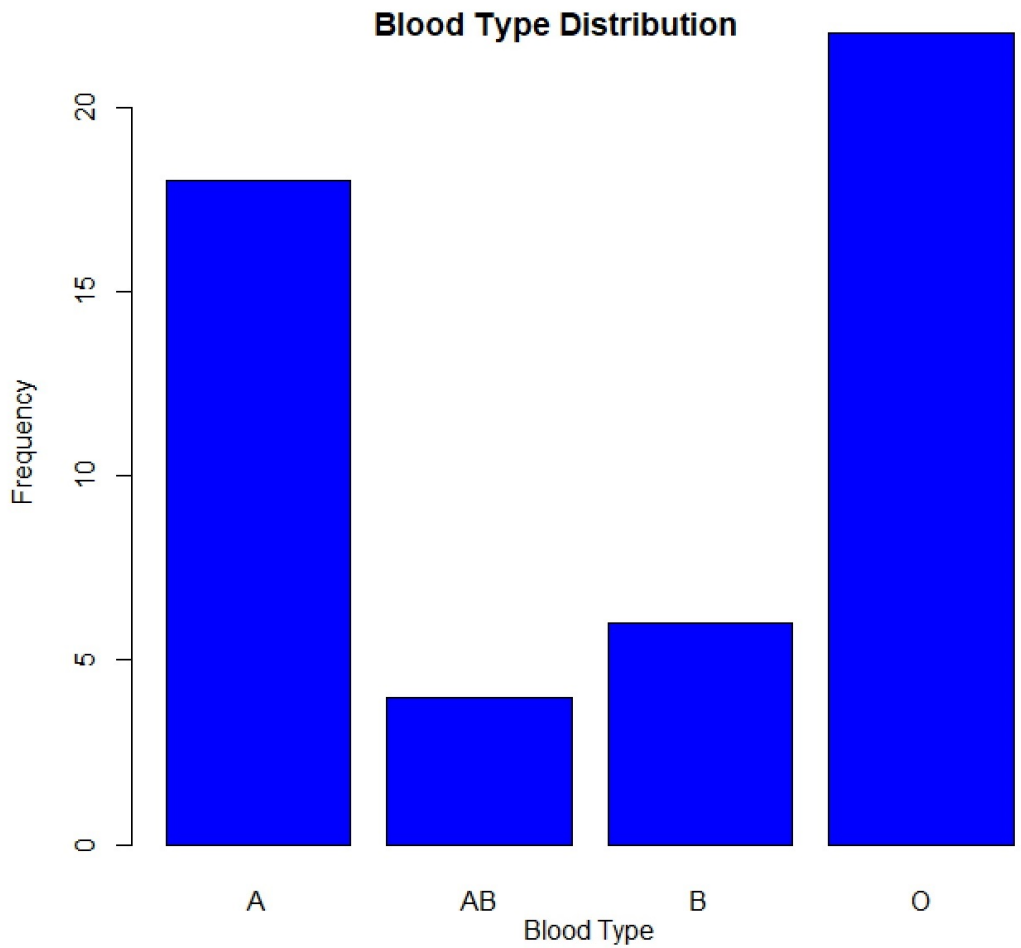



Cigarette consumption per person 18 years of age or older, 1900–1990

# Bar Charts: Blood Type Example

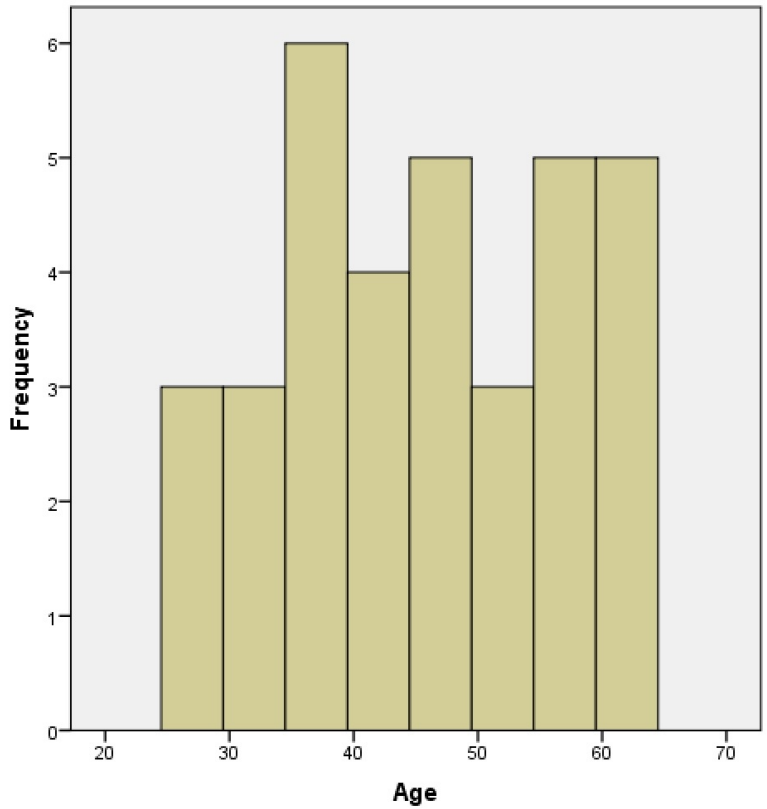
- Blood type of 50 patients.

```
> ## Bar Chart for Blood Type
> barplot(table(BT),xlab="Blood Type",ylab="Frequency",main="Blood Type Distribution", col=4)
```

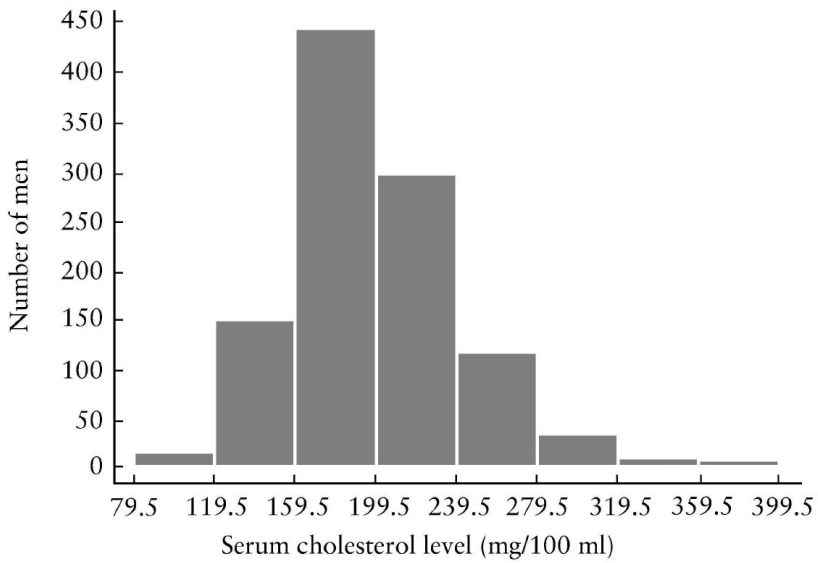


# Histograms

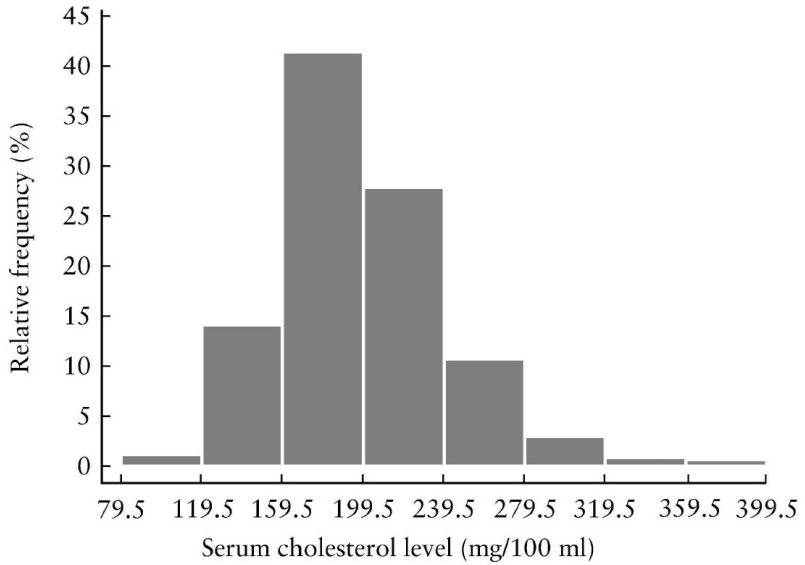
- **Histogram** depicts a frequency distribution for discrete or continuous data.
- The horizontal axis displays the true limits of the various intervals.
- The vertical axis of a histogram depicts either the frequency or the relative frequency of observations within each interval.



# Histograms: Other Examples




Absolute frequencies of serum cholesterol levels for 1067 U.S. males, aged 25 to 34 years, 1976–1980

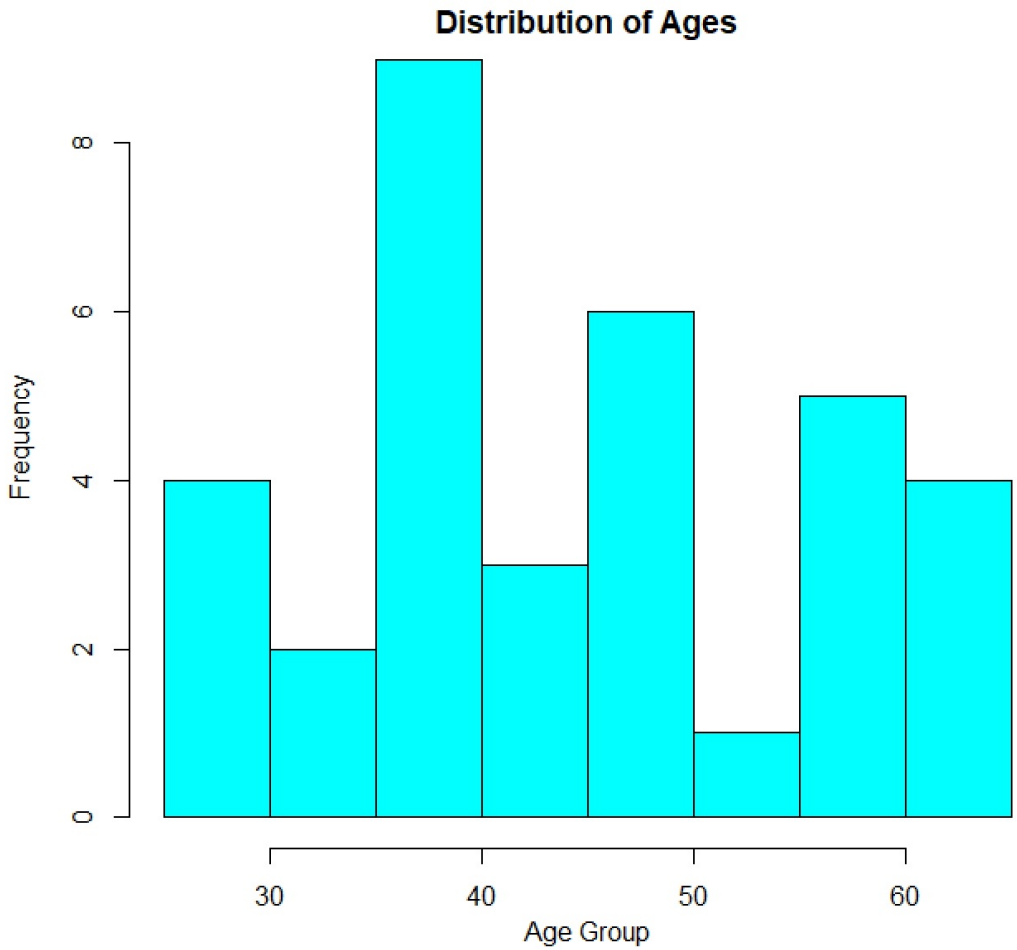



Relative frequencies of serum cholesterol levels for 1067 U.S. males, aged 25 to 34 years, 1976–1980

# Histograms: Ages of 34 patients

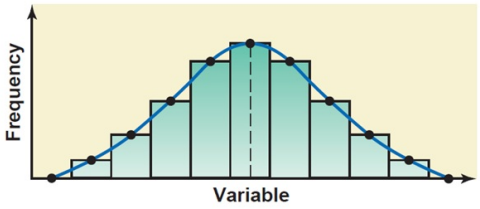
- Ages of 34 patients.

```
> hist(x, breaks=breaks,xlab="Age Group",ylab="Frequency",main="Distribution of Ages", col=5)
```

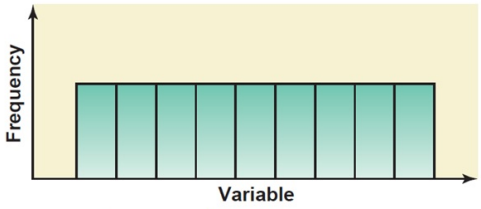


# Histograms: Shapes

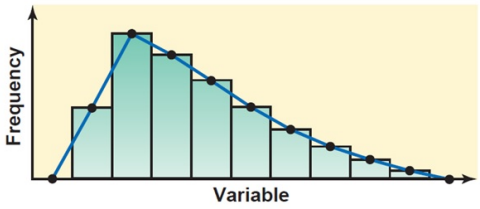
A histogram can assume any one of a large number of shapes. The most common of these shapes are: (1) **Symmetric**. (2) **Skewed**. (3) **Uniform** or **rectangular**.



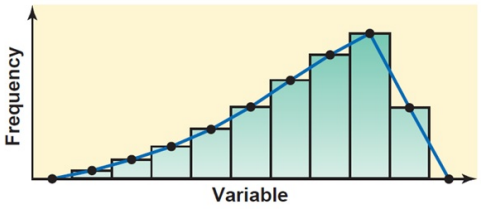

Symmetric histograms.



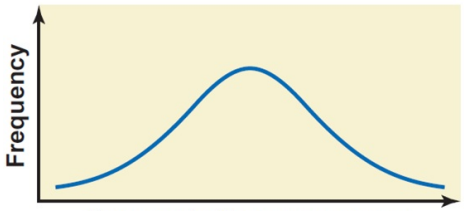

A histogram with uniform distribution.



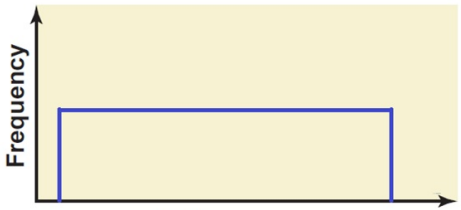

**skewed-to-the-right histogram**



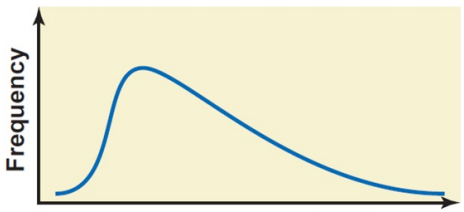

**skewed-to-the-left histogram**



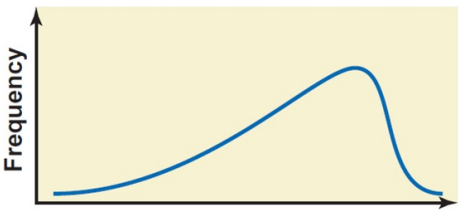

Symmetric frequency curves.




Uniform



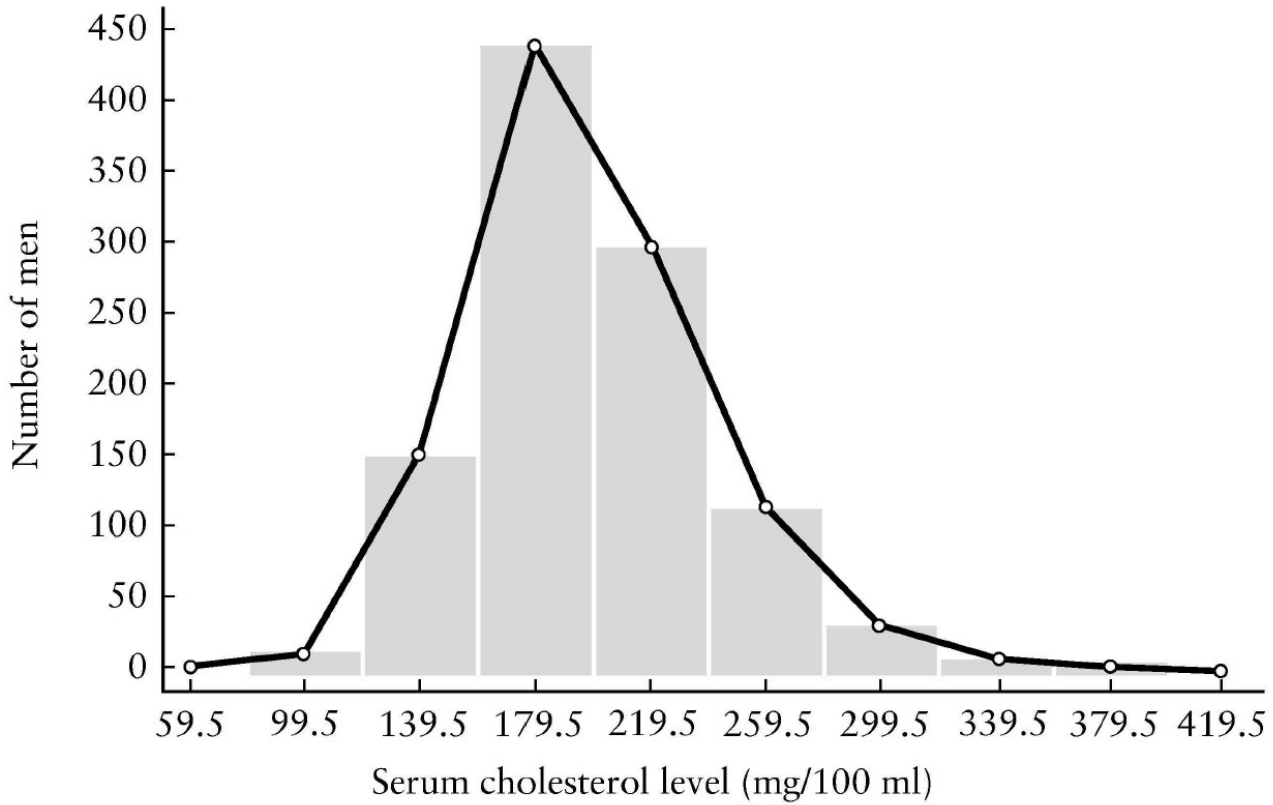

Frequency curve skewed to the right.




Frequency curve skewed to the left.

# Frequency Polygons

- A frequency polygon uses the same two axes as a histogram.
- It is constructed by placing a point at the center of each interval such that the height of the point is equal to the frequency or relative frequency associated with that interval.
- Points are also placed on the horizontal axis at the midpoints of the intervals immediately preceding and immediately following the intervals that contain observations. The points are then connected by straight lines.



# Frequency Polygons

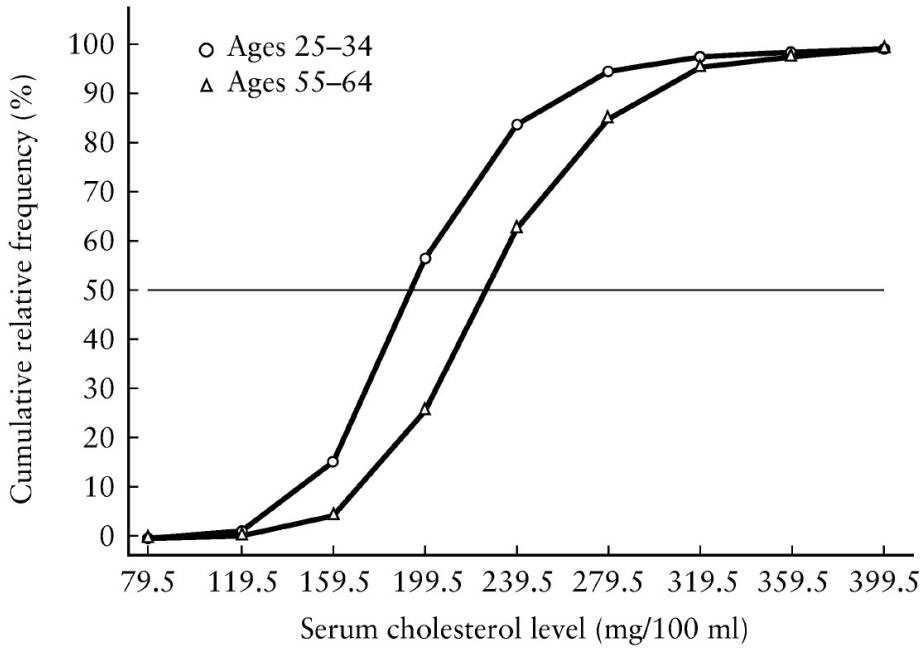
- Because they can easily be superimposed, frequency polygons are superior to histograms for comparing two or more sets of data.


Relative frequencies of serum cholesterol levels for 2294 U.S. males, 1976–1980

- Since the older men tend to have higher serum cholesterol levels, their polygon lies to the right of the polygon for the younger men.

# Cummulative Frequency Polygons

- A point is placed at the true upper limit of each interval; the height of the point represents the cumulative relative frequency associated with that interval.
- Like frequency polygons, cumulative frequency polygons can be used to compare sets of data.



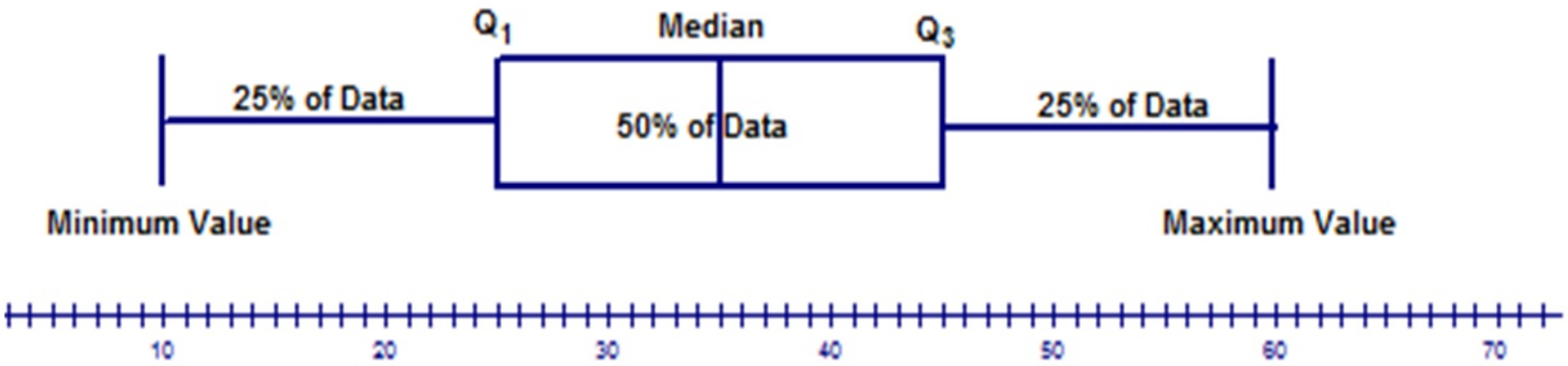

- By noting that the cumulative frequency polygon for 55- to 64-year-old males lies to the right of the polygon for 25- to 34-year-old males for each value of serum cholesterol level, we can see that the distribution for older men is **stochastically larger** than the distribution for younger men.

# Cummulative Frequency Polygons

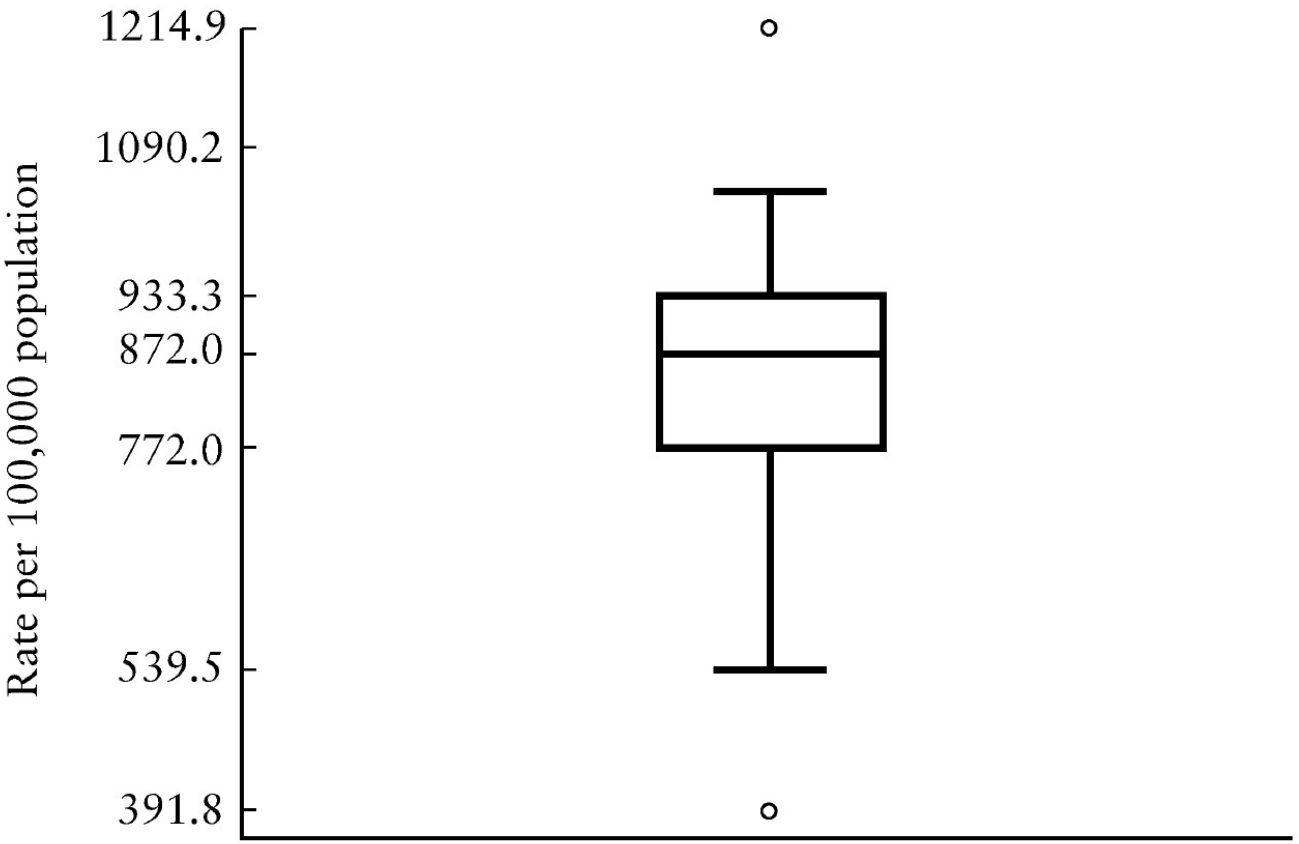
- Cumulative frequency polygons can also be used to obtain the **percentiles** of a set of data. Roughly, the 95th percentile is the value that is greater than or equal to 95% of the observations and less than or equal to the remaining 5%.
- the $50^{th}$ percentile of the serum cholesterol levels for the group of 25- to 34-year-olds-the value that is greater than or equal to half of the observations and less than or equal to the other half-is approximately 193 mg/100 ml; the $50^{th}$ percentile for the 55- to 64-year-olds is about 226 mg/100 ml.
- **Percentiles** are useful for describing the **shape of a distribution**.
- For example, if the 40th and 60th percentiles of a set of data lie an equal distance away from the midpoint, and the same is true of the 30th and 70th percentiles, the 20th and 80th, and all other pairs of percentiles that sum to 100, the data are **symmetric**; that is, the distribution of values has the same shape on each side of the 50th percentile.
- Alternatively, if there are a number of outlying observations on one side of the midpoint only, the data are said to be **skewed**.
- If these observations are smaller than the rest of the values, the data are **skewed to the left**; if they are larger than the other measurements, the data are **skewed to the right**.

# Box Plots

- Examination of a box-and-whisker plot for a set of data reveals information regarding the amount of spread, location of concentration, and symmetry of the data.



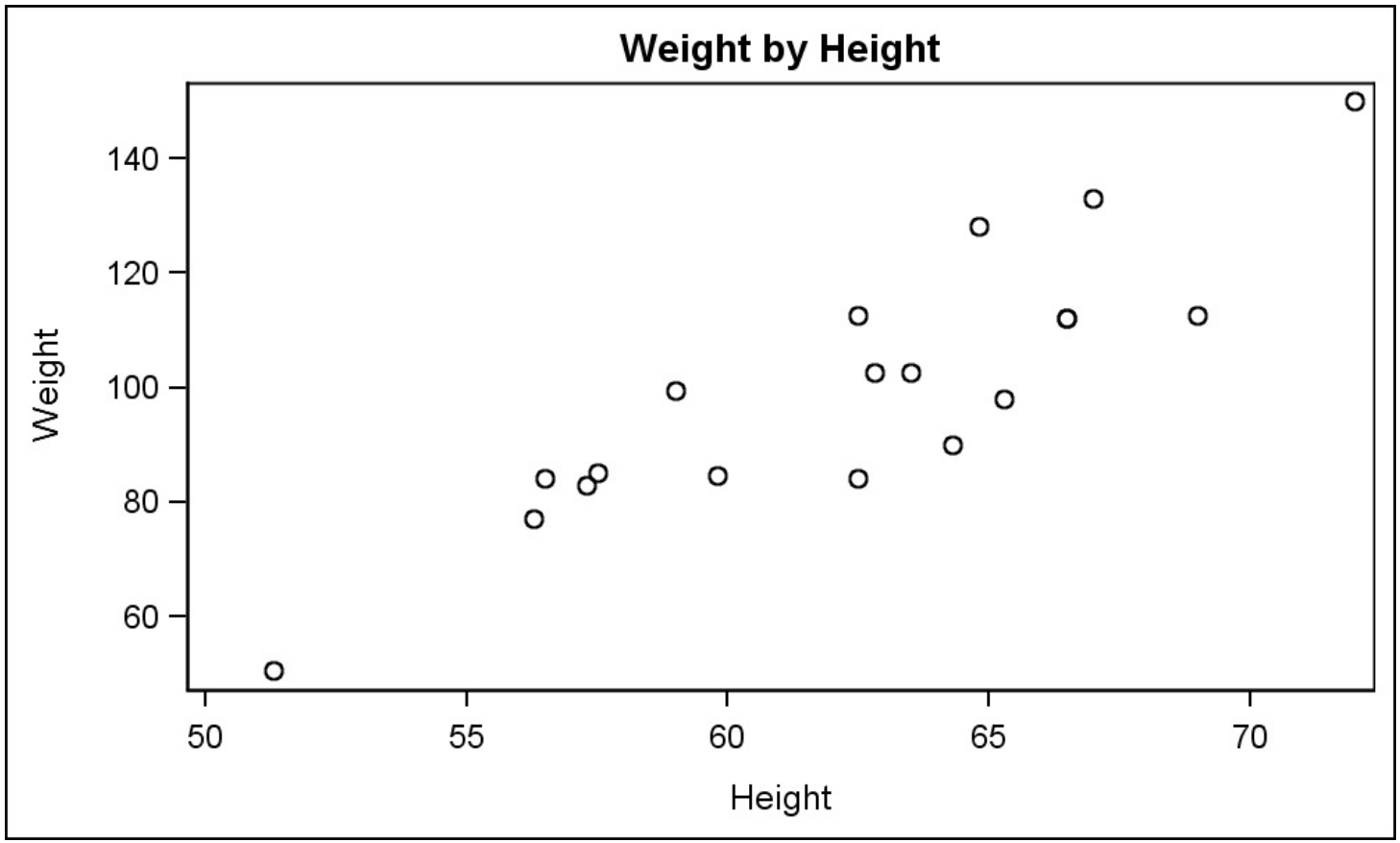
# Box Plots




Box plot: Crude death rates for the United States, 1992

# Scatter Plots

- A scatter plot is used to depict the relationship between two different continuous measurements.



# Measures of Central Tendency

- A **measure of central tendency** is a value that represents a typical, or central, entry of a data set.
- The three most commonly used measures of central tendency are the:
  1. Arithmetic mean (Average),
  2. Median,
  3. Mode.

# Measures of Central Tendency: Mean

- The **mean** of a data set is the sum of the data entries divided by the number of entries.
- It is the most popular and best understood measure of central tendency for quantitative data (discrete and continuous).
- To find the mean of a data set, use one of the following formulae:

(1) **Population Mean**: $\mu = \frac{1}{N}\sum_{i=1}^{N} X_i$, where

  - $\mu$ is the population mean ($\mu$ is a Greek letter, read as mu)
  - $\sum_{i=1}^{N}$ is the summation notation
  - $X_i$ is the value of element $i$ in the population
  - $N$ is the population size

- Note that $\mu$ is a descriptive measure of the entire population, so we call it a **parameter**.

# Measures of Central Tendency: Mean

(2) **Sample Mean**: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, where

- $\bar{X}$ is the sample mean (read as $X$ bar)
- $X_i$ is the value of element $i$ in the sample
- $n$ is the sample size

- Note that $\bar{X}$ is a descriptive measure of a sample, so we call it a **statistic**.
- We will often use the sample mean $\bar{X}$ to estimate the population mean $\mu$.

# Measures of Central Tendency: Mean

***Example***: The following cholesterol levels of 10 people were measured in mg/dl:

$$260, 150, 165, 201, 212, 243, 219, 227, 210, 240.$$

Calculate the mean.

```
> Y <- c(260,   150,   165,   201,   212,   243,   219,   227,   210,   240)
> mean(Y)
[1] 212.7

> summary(Y)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  150.0   203.2   215.5   212.7   236.8   260.0
```

# Measures of Central Tendency: Properties of the mean

- Most common measure of central tendency.
- The mean is unique: A data set has only one mean (generally not part of the data set).
- All the values are included in computing the mean (and hence is a good representative of the data).
- The mean is affected accordingly if the data values are given mathematical treatment by any constant item.
- The sum of the deviations of each value from the mean will always be zero. Expressed symbolically:

$$\sum_{i=1}^{N}(X_i - \mu) = 0 \quad \text{OR} \quad \sum_{i=1}^{n}(X_i - \bar{X}) = 0.$$

- $\mu$ is fixed but usually unknown.
- $\bar{X}$ is random but known.

# Measures of Central Tendency: Properties of the mean

- The mean is not an appropriate measure for ordinal or nominal variables.
  - For these types of data, the numbers are merely labels, so that even if we choose to represent the blood types A, B, AB and O by the numbers 1, 2, 3, and 4, an average blood type of 1.8 is meaningless.
  - One exception to this rule applies when we have dichotomous data and the two possible outcomes are represented by the values 0 and 1. In this situation, the mean of the observations is equal to the proportion of 1s in the data set.
  - ***Example***: Listed in the following table are the relevant dichotomous data; the value 1 represents a male, and 0 designates a female. If we compute the mean of these observations, we find that $\bar{X} = 0.615$.

| Subject | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Gender | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 |

Table: Indicators of gender for 13 adolescents suffering from asthma

  - Therefore, 61.5% of the study subjects are males.

# Measures of Central Tendency: Properties of the mean

- The mean is affected by outliers (unusually large or small data values).

***Example***: The following cholesterol levels of 10 people were measured in mg/dl:

$$931, 150, 165, 201, 212, 243, 219, 227, 210, 240.$$

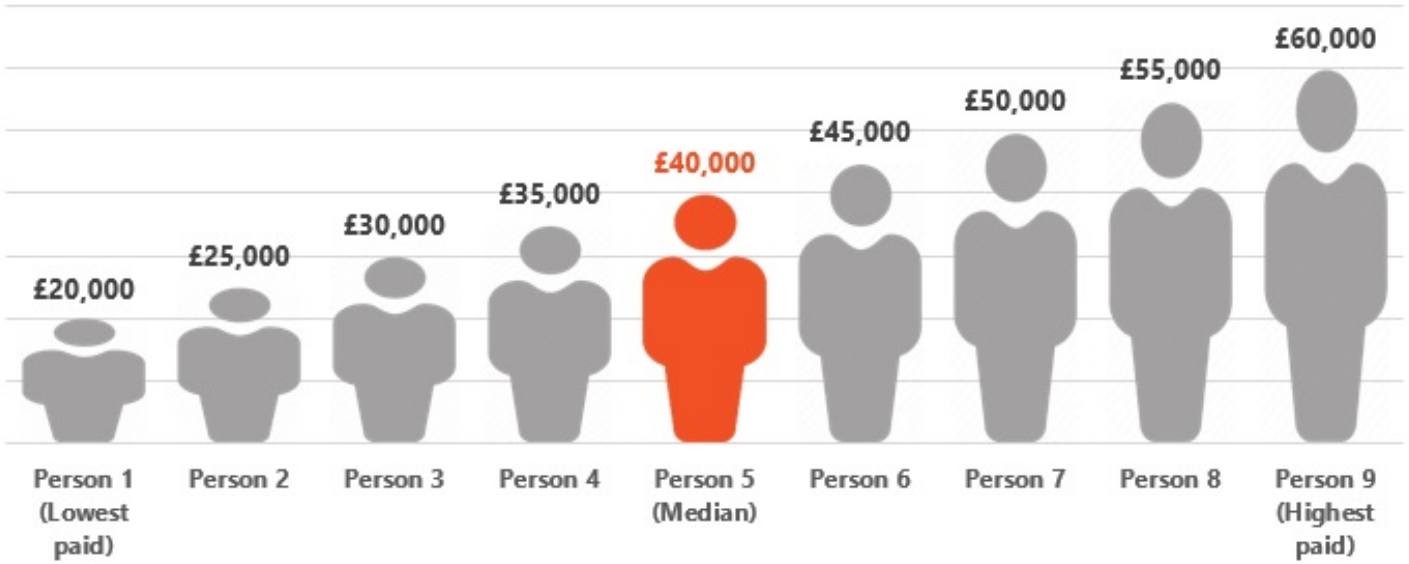
Calculate the mean.

```
> Y <- c(260,   150,   165,   201,   212,   243,   219,   227,   210,   240)
> mean(Y)
[1] 212.7

> #Outlier
> Y <- c(931,   150,   165,   201,   212,   243,   219,   227,   210,   240)
> mean(Y)
[1] 279.8
```

# Measures of Central Tendency: Median

- The **median** is defined as the $50^{th}$ percentile of a set of measurements; if a list of observations is ranked from smallest to largest, half the values are greater than or equal to the median, whereas the other half are less than or equal to it.




- If the data set has an odd number of entries, the median is the **middle data entry**.

- If the data set has an even number of entries, the median is the **mean of the two middle data entries**.

# Measures of Central Tendency: Median

***Example***: The following cholesterol levels of 10 people were measured in mg/dl:

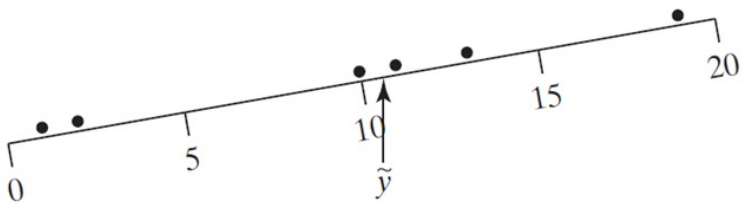
$$260, 150, 165, 201, 212, 243, 219, 227, 210, 240.$$

Calculate the median.

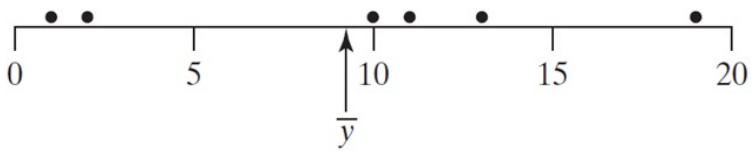
***Example***: In the previous example, add the value 120 and recalculate the median.

# Measures of Central Tendency: Properties of the median

- The median can be used as a summary measure for ordinal observations as well as for discrete and continuous data.
- There is only one median for a set of data (may be part of the data set).
- It is useful as a descriptive measure for skewed distributions.
- Unlike the mean, the median is said to be **robust**; that is, it is much less sensitive to **unusual data points**.
- Mean versus Median:




The median divides the data into two equal pieces




The mean is the "point of balance" of the data

# Measures of Central Tendency: Mode

- The **mode** of a set of values is the observation that occurs **most frequently**.


- If no entry is repeated, the data set has no mode.
- If two entries occur with the same most frequency, each entry is a mode and the data set is called **bimodal**.

# Measures of Central Tendency: Mode

***Example***: Six strains of bacteria were tested to see how long they could remain alive outside their normal environment. The time, in minutes, $\{2, 3, 5, 7, 8, 10\}$. Find the mode.

***Example***:

1. The data set: $\{1, 2, 3, 4, 6, 8, 9\}$ has **no mode**.
2. The data set: $\{1, 2, 3, 3, 4, 5\}$ has one mode (**Unimodal**).
3. The data set: $\{1, 1, 2, 3, 4, 4, 5\}$ has two modes (**Bimodal**).
4. The data set: $\{1, 1, 2, 3, 3, 4, 5, 5\}$ has three modes (**Trimodal**).

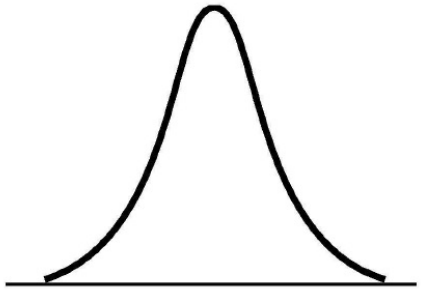
# Measures of Central Tendency: Properties of the mode

- The mode may not exist.
- The mode may not be unique.
- The mode may or may not equal the mean and median.
- The mode is not affected by extreme values.
- The mode always corresponds to one of the actual observations (unlike the mean and median).
- The mode can be used as a summary measure for all types of data (qualitative or quantitative).

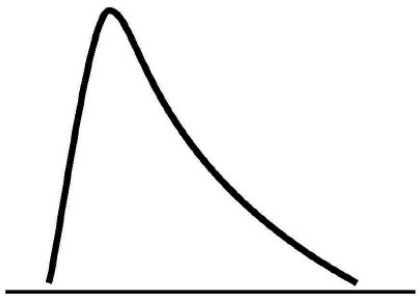
# When to use Mean, Median, and Mode

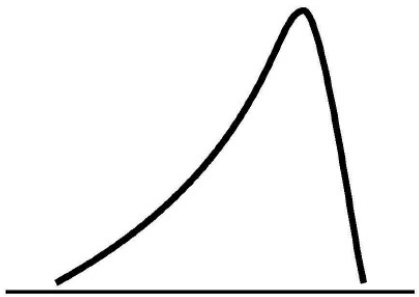
The best measure of central tendency for a given set of data often depends on the way in which the values are distributed.

- If they are **symmetric** and **unimodal**-meaning (one peak), then the mean, the median, and the mode should all be roughly the same.



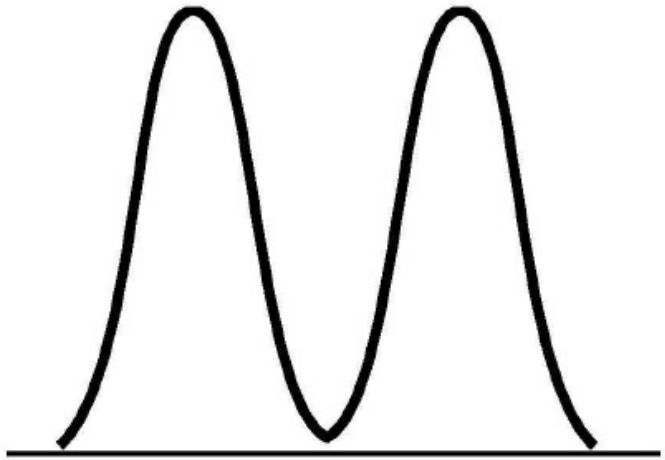
- When the data are **not symmetric** (**skewed**), the median is often the best measure of central tendency. Because the mean is sensitive to extreme observations.





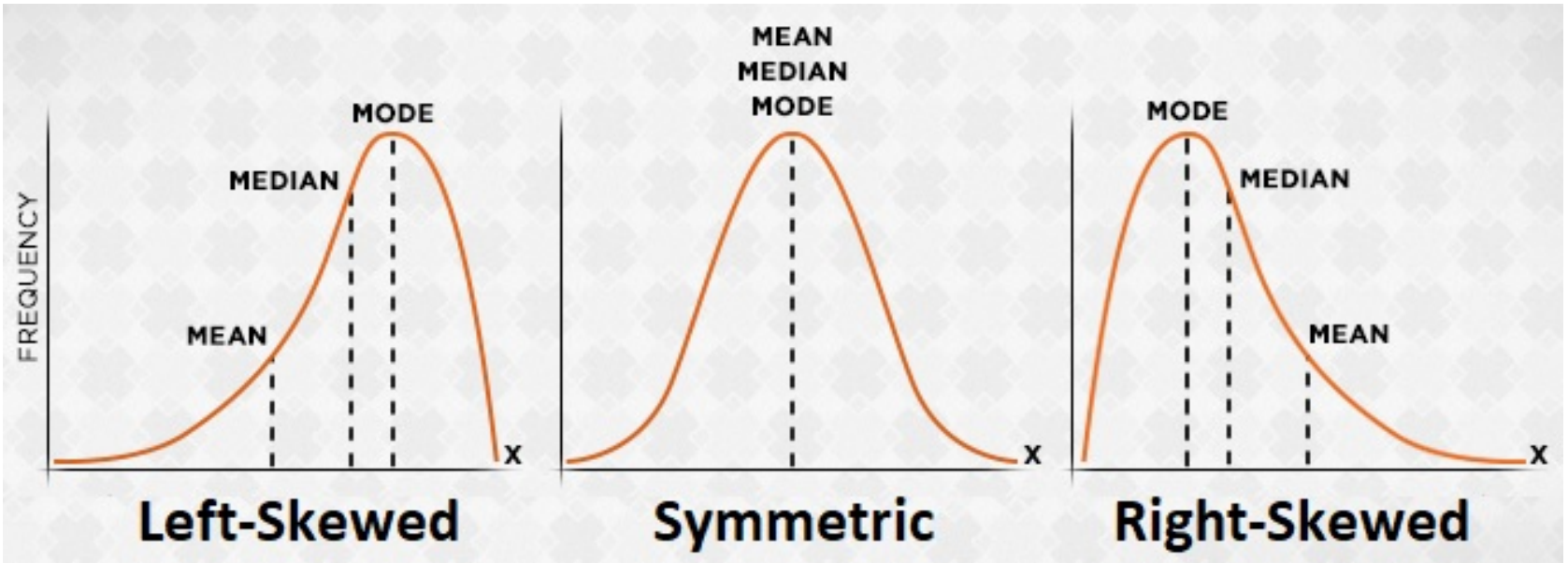
# When to use Mean, Median, and Mode

- If the distribution of values is **symmetric** but **bimodal** (two peaks), then the mean and the median should again be approximately the same.
  - Note, however, that this common value could lie between the two peaks, and hence be a measurement that is extremely unlikely to occur.
  - A bimodal distribution often indicates that the population from which the values are taken actually consists of two distinct subgroups that differ in the characteristic being measured; in this situation, it might be better to report two modes rather than the mean or the median, or to treat the two subgroups separately.



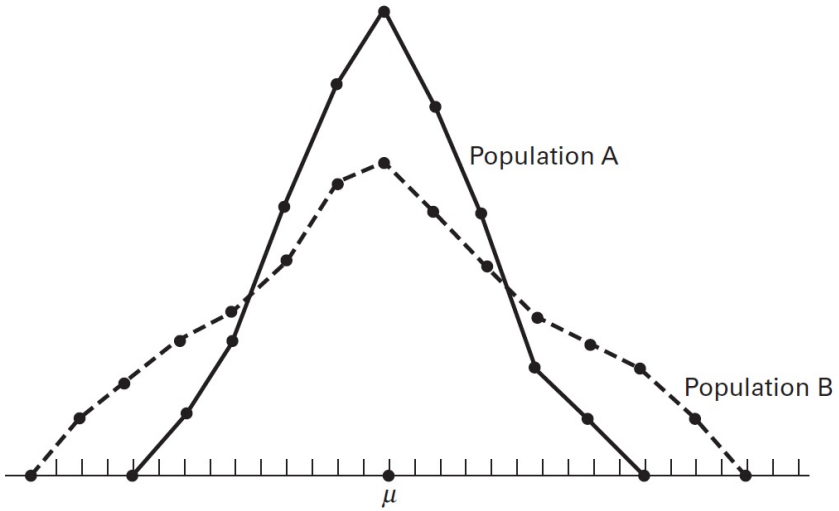
# Relationship between mean, median, and mode

1. Symmetric: Mean = Median = Mode.
2. Left-Skewed: Mean < Median < Mode.
3. Right-Skewed: Mean > Median > Mode.



# Measures of Dispersion

- A measure of dispersion conveys information regarding the amount of variability present in a set of data.
- If all the values are the same, there is no dispersion; if they are not all the same, dispersion is present in the data.
- The amount of dispersion may be small when the values, though different, are close together.




Two frequency distributions with identical means, medians, and modes but different amounts of dispersion.

- Population B, which is more variable than population A, is more spread out (has more dispersion).

# Measures of Dispersion: Range

- The **range** of a group of measurements is defined as the difference between the largest observation and the smallest.
- $Range = Max - Min$.
- It is not a good measure of dispersion to use for a data set that contains outliers (It can be misleading).
- It is not a very satisfactory measure of dispersion. Its calculation is based on two values only: the largest and the smallest. All other values in a data set are ignored when calculating the range.

***Example***: In $\{4, 6, 9, 3, 7\}$ the minimum value is 3, and the maximum is 9. So the range is $9 - 3 = 6$.

# Measures of Dispersion: Variance

- The **variance** quantifies the amount of variability, or spread, around the mean of the measurements.
- It is nonnegative and is zero only if all observations are the same.
- To find the variance of a data set, use one of the following formulae.

(1) **Population Variance**: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$, where

  - $\sigma^2$ is the population variance ($\sigma$ is a Greek letter, read as sigma)
  - $\mu$ is the population mean
  - $X_i$ is the value of element $i$ in the population
  - $N$ is the population size

- Note that $\sigma$ or $\sigma^2$ is a descriptive measure of the entire population, so we call it a **parameter**.

# Measures of Dispersion: Variance

(2) **Sample Variance**: The formula for the sample variance is

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

where

- $S^2$ is the sample variance
- $\bar{X}$ is the sample mean
- $X_i$ is the value of element $i$ in the sample
- $n$ is the sample size

- Note that $S^2$ is a descriptive measure of a sample, so we call it a **statistic**.
- We will often use the sample variance $S^2$ to estimate the population variance $\sigma^2$.

# Measures of Dispersion: Variance

- Short-Cut Formula for the Sample Variance

$$S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right].$$

OR

$$S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}\right].$$

- **Degrees of freedom**: In computing the variance there are $n-1$ degrees of freedom because if $n-1$ values are known, the $n^{th}$ one is determined automatically. This is because all of the values of $(X_i - \bar{X})$ must add to zero.

# Measures of Dispersion: Standard Deviation

The **variance** represents squared units and, therefore, is not an appropriate measure of dispersion when we wish to express this concept in terms of the original units. To obtain a measure of dispersion in original units, we merely take the positive square root of the variance. The result is called the **standard deviation**.

(1) The population standard deviation is $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2}$.

(2) The sample standard deviation is $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$.

- It is the most commonly used measure of dispersion.
- It shows variation about the mean.
- It has the same units as the original data.

# Measures of Dispersion: Example

***Example***: The following cholesterol levels of 10 people were measured in mg/dl:

$$260, 150, 165, 201, 212, 243, 219, 227, 210, 240.$$

1. Calculate the range.
2. Calculate the variance and the standard deviation.

```
> Y <- c(260,   150,   165,   201,   212,   243,   219,   227,   210,   240)
> range(Y)
[1] 150 260

> var(Y)
[1] 1166.233
> sd(Y)
[1] 34.15016
```

# Some Summary Descriptive Measures

| Some summary descriptive functions | |
|---|---|
| length | Number of elements in a vector |
| sum | Sum of the values in a vector |
| min, max, range | Minimum, maximum, and range (min, max) of a vector |
| mean, median | Mean and median of the values in a vector |
| sd, var | Standard deviation and variance |
| cov, cor | Covariance and Pearson correlation |

```
> x <- rnorm(100)
> y <- rnorm(100)
> var(x)
[1] 1.180011
> var(y)
[1] 0.928549
> sd(x)
[1] 1.086283
> sqrt(var(x))
[1] 1.086283
> cov(x,y)
[1] 0.1062208
> cor(x,y)
[1] 0.1014762
```

```
> cov(cbind(x, y)) # Covariance matrix
          x         y
x 1.1800106 0.1062208
y 0.1062208 0.9285490
> cor(cbind(x, y)) # Correlation matrix
          x         y
x 1.0000000 0.1014762
y 0.1014762 1.0000000
```

# Grouped Data

Here we discuss how to calculate the mean and variance for **grouped data** (data arranged in a frequency distribution).

(1) Mean for grouped data:

$$\bar{X} = \frac{\sum_{i=1}^{k} m_i f_i}{\sum_{i=1}^{k} f_i},$$

where

- $k$ is the number of intervals in the table (frequency distribution),
- $m_i$ is the midpoint of the $i^{th}$ interval,
- $f_i$ is the frequency associated with the $i^{th}$ interval,
- $\sum_{i=1}^{k} f_i = n$.

(2) Variance for grouped data:

$$S^2 = \frac{\sum_{i=1}^{k} m_i^2 f_i - \frac{\left(\sum_{i=1}^{k} m_i f_i\right)^2}{\sum_{i=1}^{k} f_i}}{\sum_{i=1}^{k} f_i - 1}$$

# Grouped Data: Example

***Example***: Displayed below is a frequency distribution containing a summary of the resting systolic blood pressures for a sample of 35 patients with ischemic heart disease, or suppression of blood flow to the heart.

| **Blood Pressure (mm Hg)** | **Number of Patients** |
|---|---|
| 115–124 | 4 |
| 125–134 | 5 |
| 135–144 | 5 |
| 145–154 | 7 |
| 155–164 | 5 |
| 165–174 | 4 |
| 175–184 | 5 |
| Total | 35 |

Compute the grouped mean and the grouped standard deviation of the data.

# Grouped Data: Example

***Example***: All 57 residents in a nursing home were surveyed to see how many times a day they eat meals.

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 3 | 3 |
| 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 3 | 3 | 3 | 3 | 3 | 3 | 3 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| 4 | 4 | 4 | 4 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | | | |

What is the mean for the number of meals eaten per day?

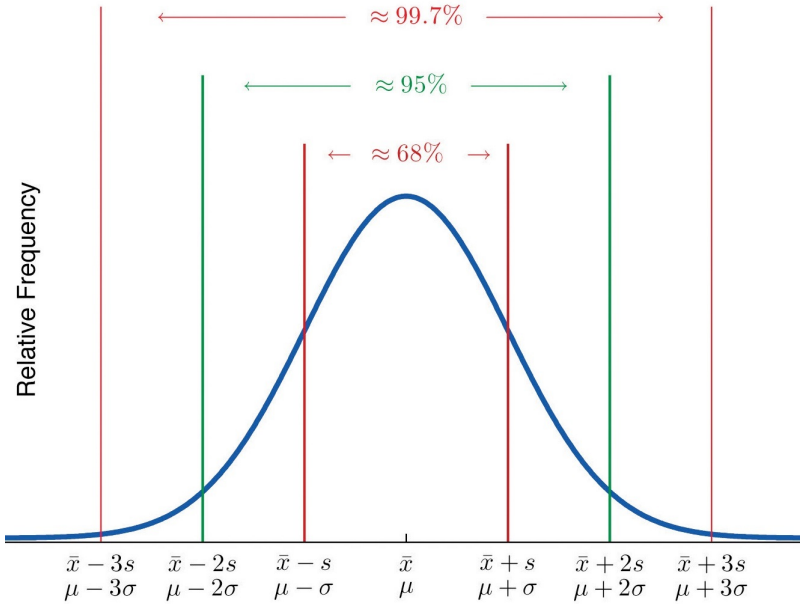
# Interpreting Standard Deviation

- When interpreting the standard deviation, remember that it is a measure of the typical amount an entry deviates from the mean.
- The more the entries are spread out, the greater the standard deviation.

(1) **Empirical Rule**:

For data with a (symmetric) bell-shaped distribution, the standard deviation has the following characteristics:

1. About 68% of the data lie within one standard deviation of the mean.
2. About 95% of the data lie within two standard deviations of the mean.
3. About 99.7% of the data lie within three standard deviations of the mean.



# Empirical Rule: Example

***Example***: Suppose the pulse rates of 200 college men are **bell-shaped** with a mean of 72 and standard deviation of 6.

- About 68% of the men have pulse rates in the interval ————————-.
- About 95% of the men have pulse rates in the interval ————————-.
- About 99.7% of the men have pulse rates in the interval ————————-.

# Empirical Rule: Example

***Example***: Heights of 18-year-old males have a bell-shaped distribution with mean 69.6 inches and standard deviation 1.4 inches.

- About what proportion of all such men are between 68.2 and 71 inches tall?

- What interval centered on the mean should contain about 95% of all such men?

# Empirical Rule: Example

***Example***: BMIs of 15-Year-old Boys.

- At age 15, suppose that BMI follows an approximately bell-shaped distribution with mean 19.83 and standard deviation 3.43.
- Then we would expect approximately 68% of 15-year-old boys to have BMIs falling in the interval (16.4, 23.26).
- Suppose that a particular 15-year-old boy, Ahmad, has a BMI equal to 25. Then, Ahmad's BMI is more extreme than two thirds of boys his age.
- We would expect 95% of 15-year-old boys to have BMIs falling in the interval (12.97, 26.69) and nearly all to fall in the interval (9.54, 30.12).
- In fact, BMI is probably not quite bell-shaped for 15-year-olds. It may be for 5-year-olds, but by age 15 there are many obese children who probably skew the distribution to the right (lots of large values in the right tail).
- Therefore, the empirical rule may be somewhat inaccurate for this variable.

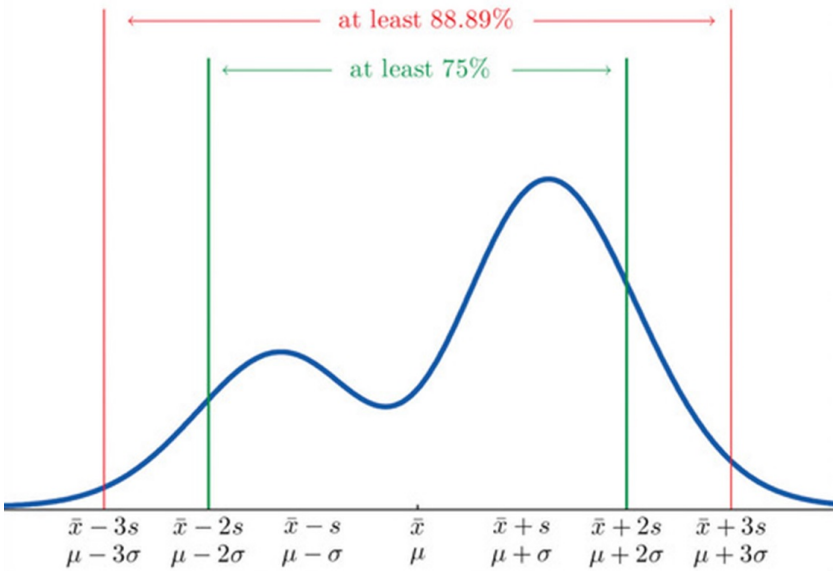
# Interpreting Standard Deviation

## (2) **Chebyshev's Rule**:

It applies to any data set, regardless of the shape of the distribution.

1. No useful information is provided on the fraction of measurements that fall within one standard deviation of the mean.
2. At least 75% of the data will fall within two standard deviations of the mean.
3. At least 89% of the data will fall within three standard deviations of the mean.

Generally, for any $k > 1$, at least $1 - \frac{1}{k^2}$ of the data will fall within $k$ standard deviations of the mean.



# Chebyshev's Rule: Example

***Example***: A sample of size 50 has mean 28 and standard deviation 3. Without knowing anything else about the sample, what can be said about the number of observations that lie in the interval (22, 34)? What can be said about the number of observations that lie outside that interval?

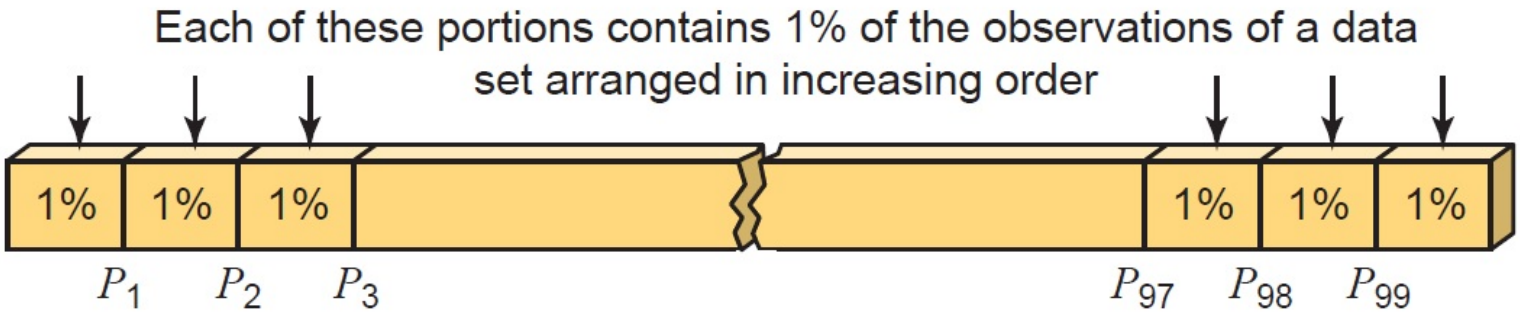
# Measures of Position

Statisticians often talk about the position of a value, relative to other values in a set of data. The most common measures of position are percentiles, quartiles, and standard scores ($z$-scores).

## (1) **Percentiles**:

- The $k^{th}$ percentile $P$ of a sample of $n$ observations (denoted by $P_k$) is that value of the variable such that $k$ percent or less of the observations are less than $P_k$ and $(100 - k)$ percent or less of the observations are greater than $P_k$.

- For example, the $25^{th}$ percentile is that value of a variable ($P_{25}$) such that 25% of the observations are less than that value and 75% of the observations are greater.



# Measures of Position

**Calculating the $k^{th}$-percentile ($P_k$):**

1. Order the data from smallest to largest.
2. Determine the location of the percentile (index):

$$index = \frac{n \times k}{100}.$$

   (a) If index is not an integer, round it up to the next integer and find the corresponding ordered value.

   (b) If index is an integer, say $m$, calculate the average of the $m^{th}$ and $(m+1)^{th}$ ordered values.

# Percentiles: Example

***Example***: Sixteen people sign up for a weight-loss class and the amount of weight lost at the end of the two-month period (in pounds) is as follows:

5 10 8 7 25 12 5 14 11 10 21 9 8 11 18 10.

1. Calculate the approximate value of the $90^{th}$ percentile.
2. Calculate the approximate value of the $50^{th}$ percentile.

```
> W <- c(5, 10, 8, 7, 25, 12, 5, 14, 11, 10, 21, 9, 8, 11, 18, 10)

> quantile(W)
  0%  25%  50%  75% 100%
 5.0  8.0 10.0 12.5 25.0

> quantile(W, c(.10, .20, .30, .70, .80, 0.9))
 10%  20%  30%  70%  80%  90%
 6.0  8.0  8.5 11.5 14.0 19.5
```

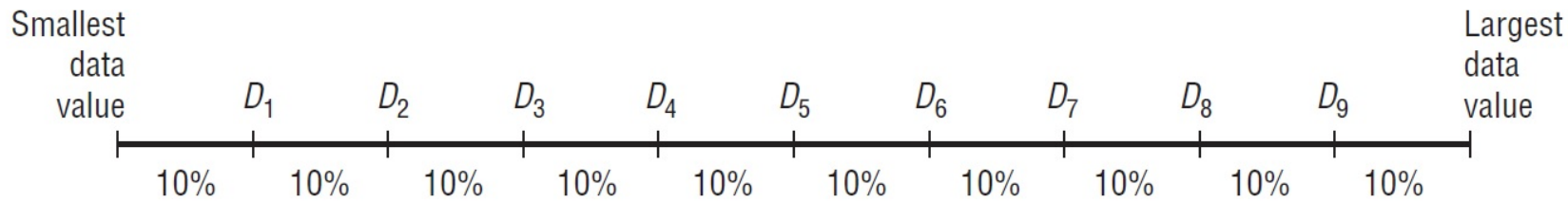
# Measures of Position

## (2) Deciles:

- The deciles are special cases of percentiles;

$$D_1 = P_{10}, \quad D_2 = P_{20}, \quad \cdots\cdots\cdots, D_9 = P_{90}.$$

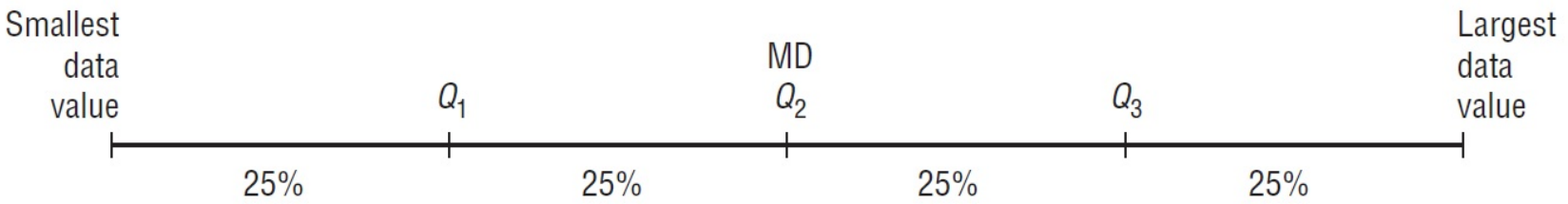
- Thus, deciles divide the data set into 10 equal parts.




## (3) Quartiles:

- Quartiles divide the data set into four equal parts, separated by $Q_1$, $Q_2$, and $Q_3$.

- Note that $Q_1$ is the first quartile (same as the $25^{th}$ percentile); $Q_2$ is the second quartile (same as the $50^{th}$ percentile, or the median, or $D_5$); and $Q_3$ is the third quartile (corresponds to the $75^{th}$ percentile).



# Measures of Position

**Calculating the quartiles:**

1. Arrange the data in order from lowest to highest.
2. Find the median of the data values. This is the value for $Q_2$.
3. Find the median of the data values that fall below $Q_2$. This is the value for $Q_1$.
4. Find the median of the data values that fall above $Q_2$. This is the value for $Q_3$.

# Measures of Position: Example

***Example***: Prior to the start of class, resting heart rates were recorded for all females who had signed up for a low-impact aerobics program. The rates for the fifteen females, ages 20 to 25, were

73 77 80 83 73 82 75 73 77 84 76 81 75 79 70.

1. Find the values of $Q_1$, $Q_2$, and $Q_3$.

2. Calculate the approximate value of $D_1$.

# Percentile Rank

- A **percentile rank** of a given score $X$ indicates the proportion or percentage of data values that fall below $X$.

- Finding percentile rank of a value:

$$\text{Percentile rank of } X = \frac{\text{Number of observations below } X + 0.5}{n} \times 100\%.$$

***Example***: Sixteen people sign up for a weight-loss class and the amount of weight lost at the end of the two-month period (in pounds) is as follows:

5 10 8 7 25 12 5 14 11 10 21 9 8 11 18 10.

Find the percentile rank of 8.

# Five Number Summary

- The combination of the five numbers ($Min$, $Q_1$, $MD$, $Q_3$, $Max$) is called the **five number summary**.
- It provides a quick numerical description of both the center (location) and spread of a distribution.
- Each of the values represents a measure of position in the dataset.
- The $Min$ and $Max$ providing the boundaries and the quartiles and median providing information about the $25^{th}$, $50^{th}$, and $75^{th}$ percentiles.

# Measures of Position

## (4) Standard score ($z$-score):

- Sometimes we want to compare data which come from different samples or populations. This is sometimes done using standard units or scores.
- The standard score, or $z$-score, represents the distance between a given measurement $X$ and the mean, expressed in standard deviations.
- In other words, the number of standard deviations that a particular measurement deviates from the mean is called $z$-score.
- The sample $z$-score for a measurement $X$ is

$$z = \frac{X - \bar{X}}{S}$$

- The population $z$-score for a measurement $X$ is

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

# Measures of Position

**Properties of the standard score:**

- A $z$-score can be negative, positive, or zero.
- If $z$ is negative, the corresponding $X$-value is below the mean.
- If $z$ is positive, the corresponding $X$-value is above the mean.
- If $z = 0$, the corresponding $X$-value is equal to the mean.
- Given values for $\mu$ (or $\bar{X}$) and $\sigma$(or $S$), we can go from the "$x$ scale" to the "$z$ scale", and vice versa. Algebraically, we can solve for $X$ and get $X = \sigma z + \mu$ or $X = Sz + \bar{X}$.
- This is also the procedure that is used to get from degrees Celsius ($^\circ$C) to degrees Fahrenheit ($^\circ$F). The relationship is

$$^\circ C = \frac{^\circ F - 32}{1.8}.$$

Similarly,

$$^\circ F = 1.8 \times ^\circ C + 32.$$

# Measures of Position

***Example***: Suppose that blood sugar levels are normally distributed with a mean of 100 mg/dl and a standard deviation of 10 mg/dl. Ahmad has a blood sugar level of 85 mg/dl. Calculate and interpret Ahmad's $z$-score.

***Example***: Women's heights have a mean of 63.6 inches and a standard deviation of 2.5 inches. The $z$-score corresponding to a woman with a height is 2.56. How tall is the woman?

# Other Measures of Dispersion

## (1) **Interquartile Range**:

- The interquartile range (IQR) is the difference between the third and first quartiles. That is,

$$IQR = Q_3 - Q_1.$$

- The *IQR* is essentially the range of the middle 50% of the data.

- It is a better measure of dispersion than range because it leaves out the extreme values (outliers).

# Other Measures of Dispersion

## (2) **Coefficient of Variation**:

- In some cases the variance of a set of observations changes with its mean.
- For example: suppose we are measuring the weights of children of various ages.
  - 5-year-old children (relatively light, on average)
  - 15-year-old children (much heaver, on average)
- Clearly, there's much more variability in the weights of 15-year-olds.
- A valid question to ask is "Do 15-year-old children's weights have more variability relative to their average?"
- The coefficient of variation ($CV$) allows such comparisons to be made.
- Population $CV = \frac{\sigma}{\mu} \times 100\%$.
- Sample $CV = \frac{S}{\bar{X}} \times 100\%$.

# Coefficient of Variation: Example

***Example***: The standard deviations and means for the weights (in kg) of 5- and 15-year-old male children are obtained as follow:

| Age | $s$ | $\bar{x}$ | CV |
|---|---|---|---|
| 5 | 2.74 | 18.39 | |
| 15 | 12.05 | 56.28 | |

- Thus, 15-year-olds' weights are **more variable** relative to their average weight than 5-year-olds.
- Note that the CV is a unitless quantity (unit-free measure).

# Identifying Outliers

- An **outlier** is a value that is very small or very large relative to the majority of the values in a data set.

- Steps for identifying outliers:

  1. Arrange data in order.

  2. Calculate the first quartile ($Q_1$), third quartile ($Q_3$), and the interquartile range ($IQR$).

  3. Compute $Q_1 - 1.5 \times IQR$ and $Q_3 + 1.5 \times IQR$.

  4. Anything outside the interval ($Q_1 - 1.5 \times IQR$, $Q_3 + 1.5 \times IQR$) is an outlier.

# Identifying Outliers: Example

***Example***: Sixteen people sign up for a weight-loss class and the amount of weight lost at the end of the two-month period (in pounds) is as follows:

5 10 8 7 25 12 5 14 11 10 21 9 8 11 18 10.

Check this data set for outliers.