Summary Statistics

What Is Statistics?

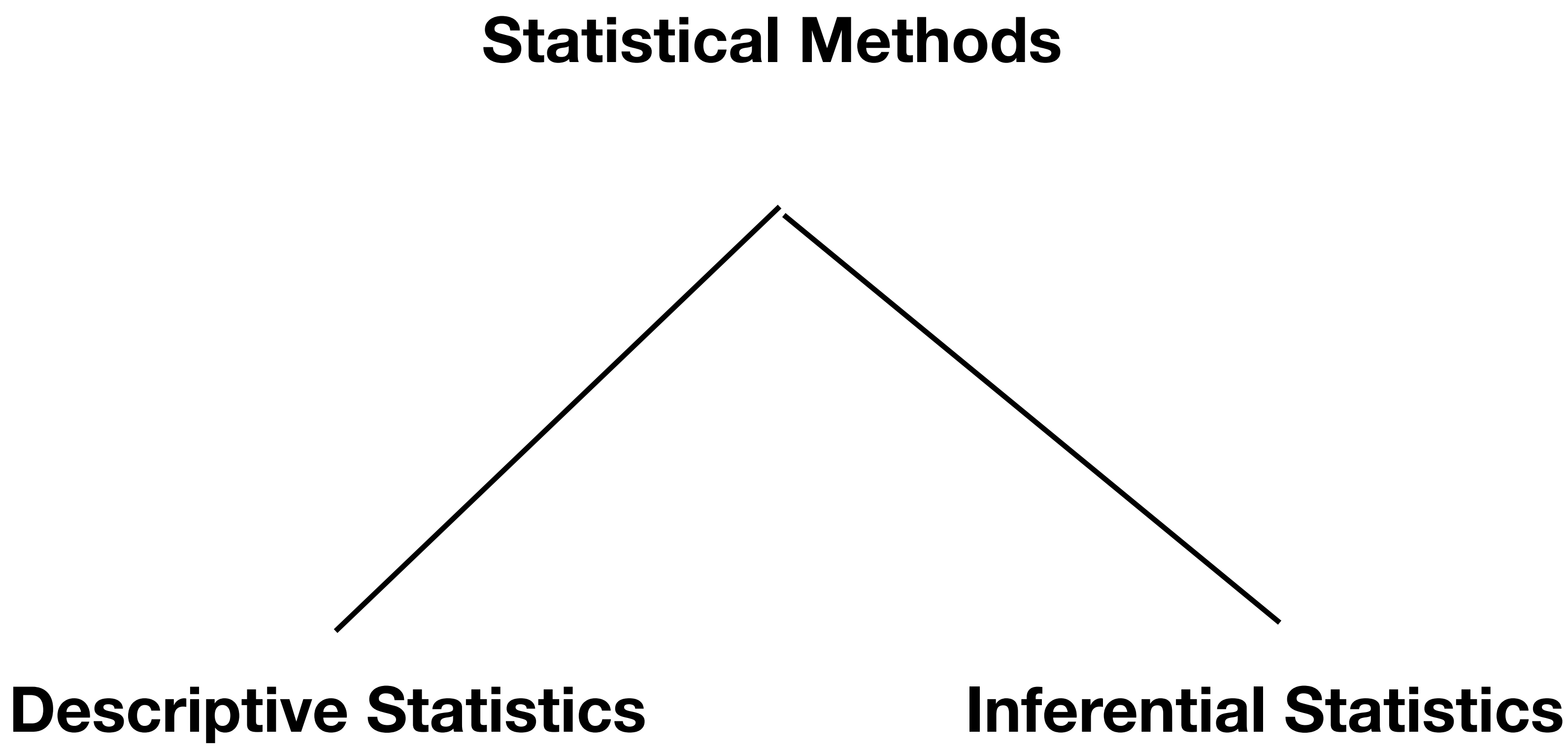
1. **Collecting Data**
   e.g., Survey
2. **Presenting Data**
   e.g., Charts & Tables
3. **Characterizing Data**
   e.g., Average

What Is Statistics?

Statistics is the science of data. It involves collecting, classifying, summarizing, organizing, analyzing, and interpreting numerical information.

Types of Statistical Applications in Business

- Economics
  - Forecasting
  - Demographics
- Sports
  - Individual & Team Performance
- Engineering
  - Construction
  - Materials
- Business
  - Consumer Preferences
  - Financial Trends


Statistical Methods
Descriptive Statistics
Inferential Statistics

Introduction

- **Descriptive Statistics** vs. **Inferential Statistics**
- Descriptive Statistics - Data summarization
- Inferential Statistics - Use of sample data to make inferences about a population parameter.

Introduction

- **Population:** the collection of objects upon which measurements **could** be taken.
- **Sample:** a subset of the population.
- **Variable** is the measurable characteristic of an entity.

Types of Data

- Quantitative or Qualitative?
  - Quantitative: presented as numbers permitting arithmetic
    - Interest rate
    - Temperature
  - Qualitative (categorical): everything else
    - Country of birth
    - Supplier

Types of Data

Quantitative

ID	Age
1	17
2	29
3	54
4	33

Qualitative

ID	Country
1	1
2	2
3	1
4	3

1 : China ,2 : US ,3 : Japan

Types of Data

- Univariate or Multivariate?
  - Univariate: one fact for each object in a dataset (“one column in a spreadsheet”)

    It means One variable
  - Multivariate: two or more facts for each object in a dataset (“many columns in a spreadsheet”)

Types of Data

Univariate

ID	Age
1	17
2	29
3	54
4	33

Multivariate

ID	Country	Age	...
1	1	17	
2	2	29	
3	1	54	
4	3	33	

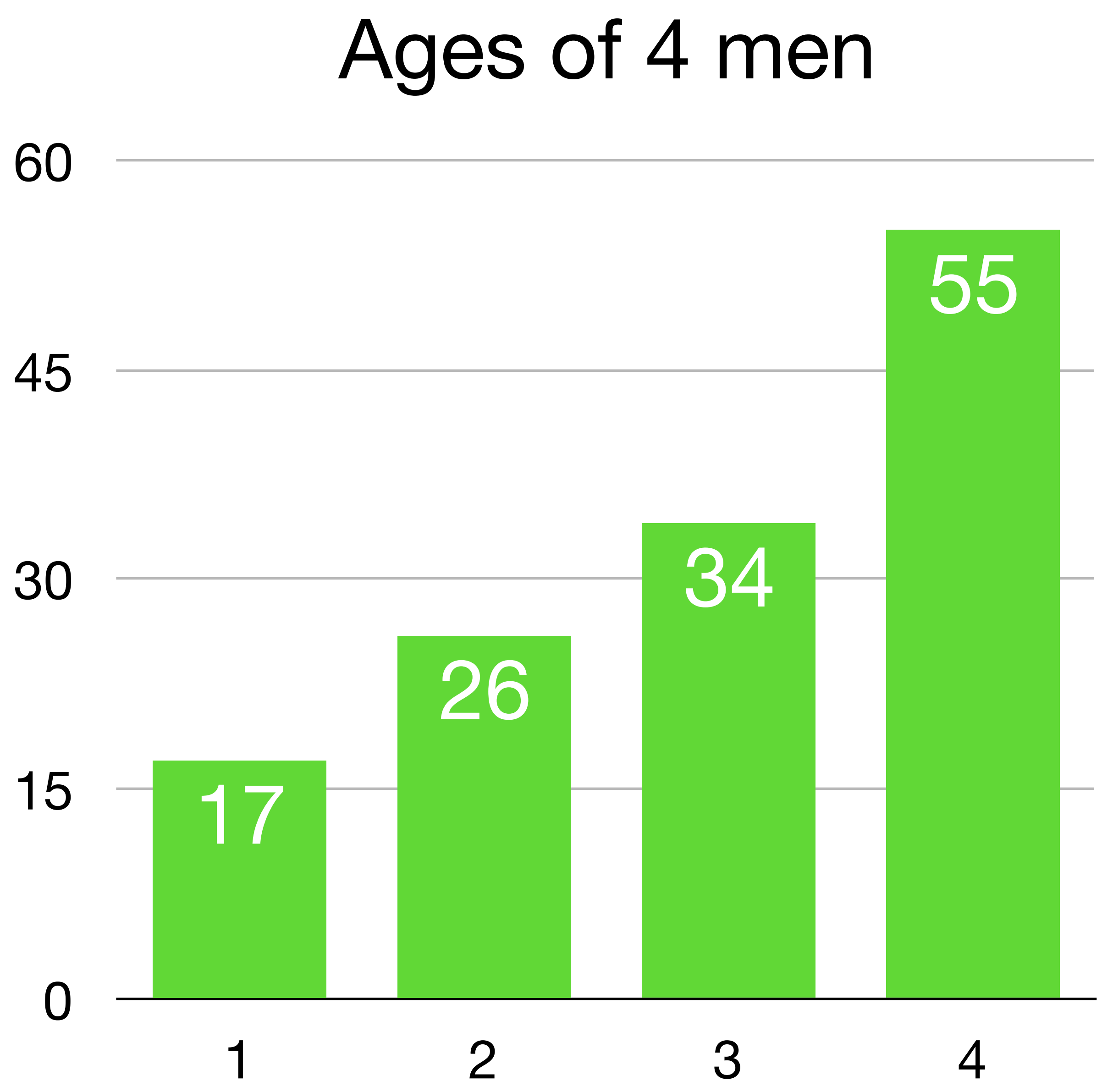
1 : China ,2 : US ,3 : Japan

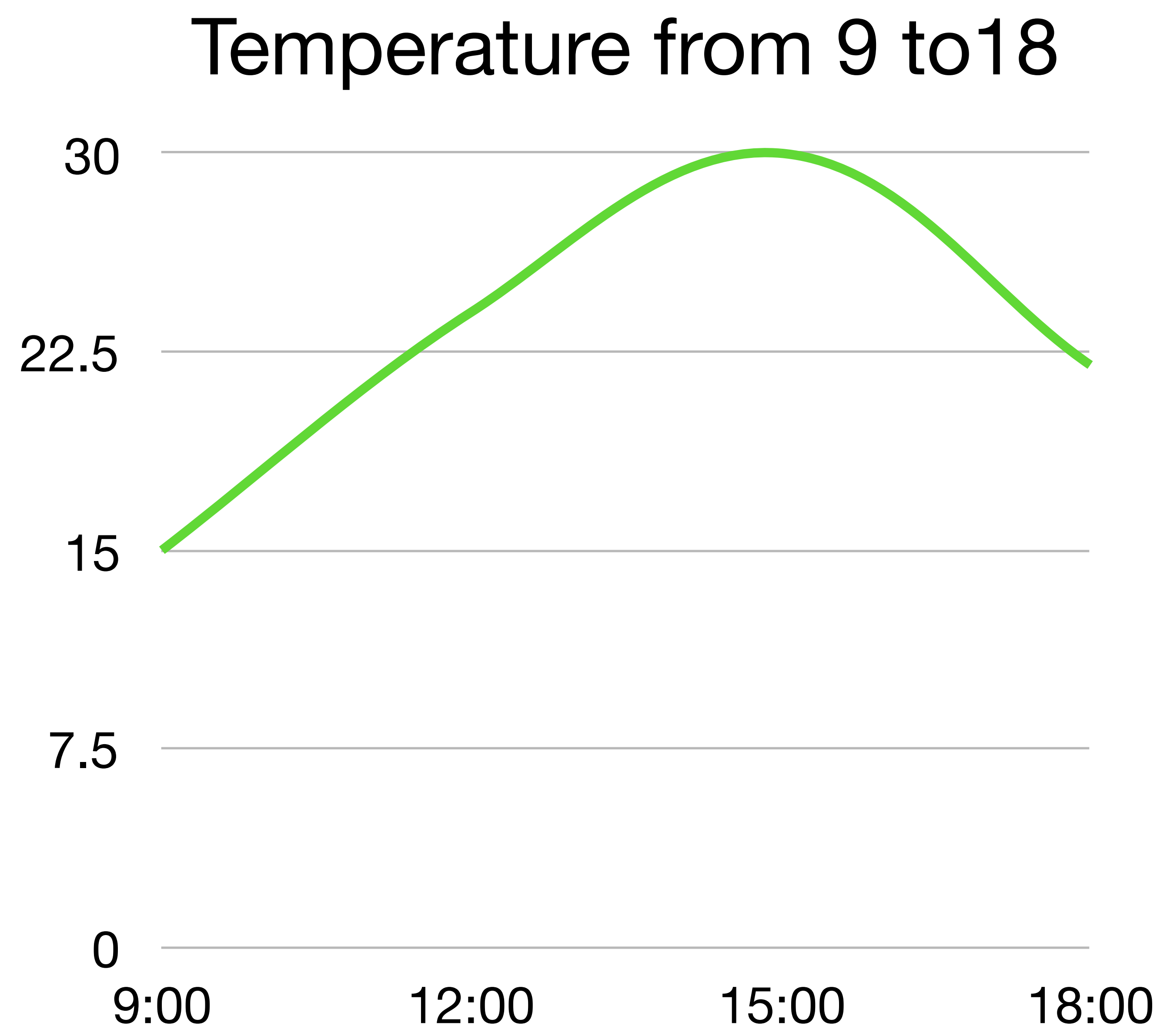
Types of Data

- Discrete or Continuous?
  - Discrete: counted
    - Cars sold
    - Number of children
  - Continuous: measured (always allow “in-between” values)
    - Gallons of oil sold
    - Temperature
  - What about age? Money?

- Let us turn to next slide

Types of Data





(Temperature is a continuous variable because it could be measured to any degree of precision desired)

The Distribution of Values of a Variable (Graphical Procedures)

Frequency Distribution

What is a Frequency Distribution?

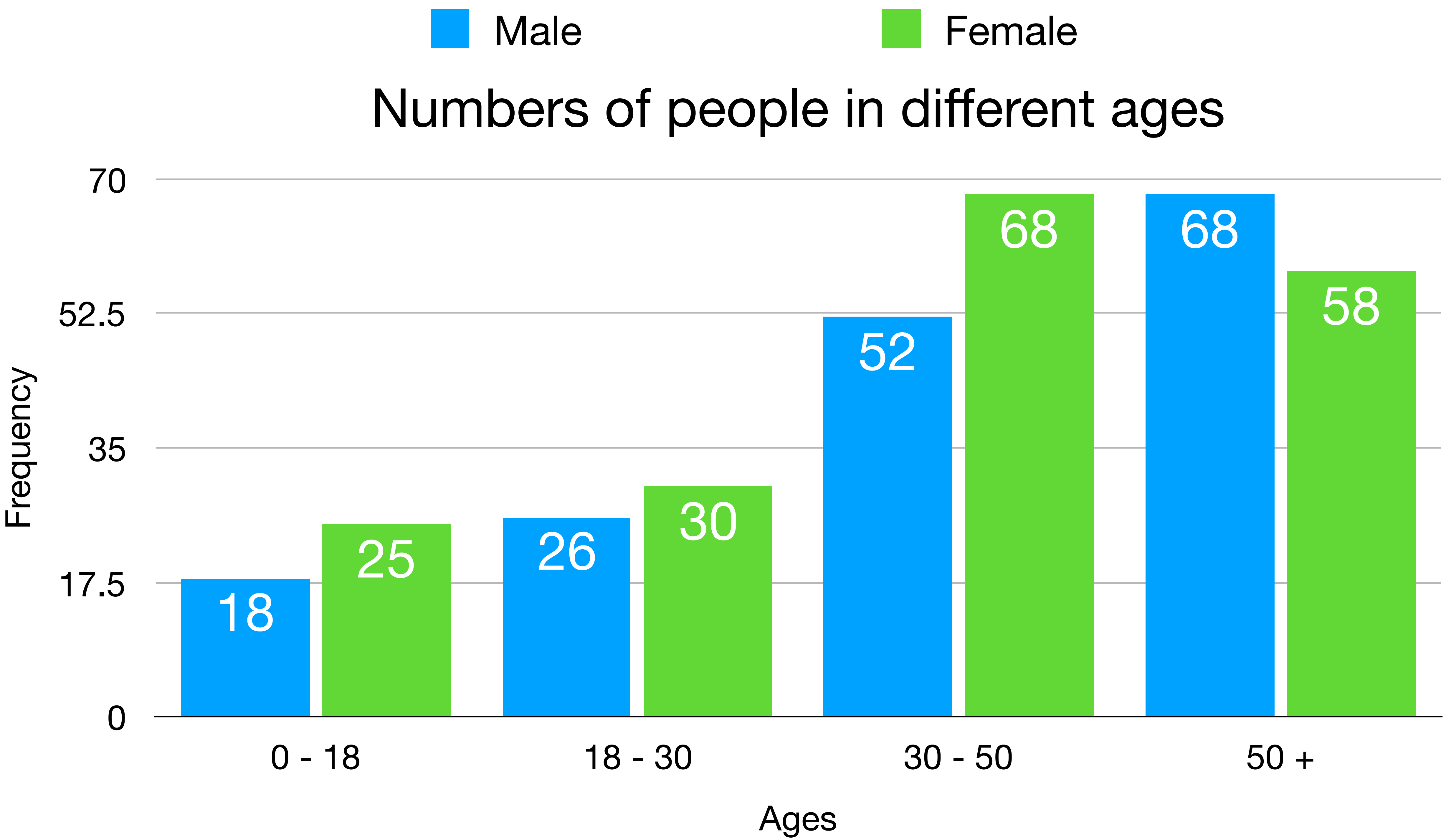
- A frequency distribution is a list or a table …
- containing the values of a variable (or a set of ranges within which the data fall) ...
- and the corresponding frequencies with which each value occurs (or frequencies with which data fall within each range)

Why Use Frequency Distributions

- A frequency distribution is a way to summarize data
- The distribution condenses the raw data into a more useful form...
- and allows for a quick visual interpretation of the data

Frequency Distribution: Discrete Data

- Discrete data: possible values are countable



Frequency Distribution: Continuous Data

Example: A manufacturer of insulation randomly selects 20 winter days and records the daily high temperature

24, 35, 17, 21, 24, 37, 26, 46, 58, 30,
32, 13, 12, 38, 41, 43, 44, 27, 53, 27

(Temperature is a continuous variable because it could be measured to any degree of precision desired)

Grouping Data by Classes

Sort raw data in ascending order:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44,…46, 53, 58

- Find range: 58 - 12 = 46

- Select number of classes: 5 (usually between 5 and 20, we can use 2^k >= n where k is number of classes and n is the number of data values or use k= 1+ 3.3 log (n))

- Compute class width: =

$$\frac{\textbf{Largest value – Smallest value}}{\textbf{Number of Classes}}$$

(46/5 then round off to 10)

- Determine class boundaries:10, 20, 30, 40, 50

- Count observations & assign to classes

Frequency Distribution Example

Data in ordered array:
12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

Frequency Distribution		
Class	Frequency	Relative Frequency
10 - 20	3	0.15
20 - 30	6	0.30
30 -40	5	0.25
40 - 50	4	0.20
50 - 60	2	0.10
Total	20	1.00

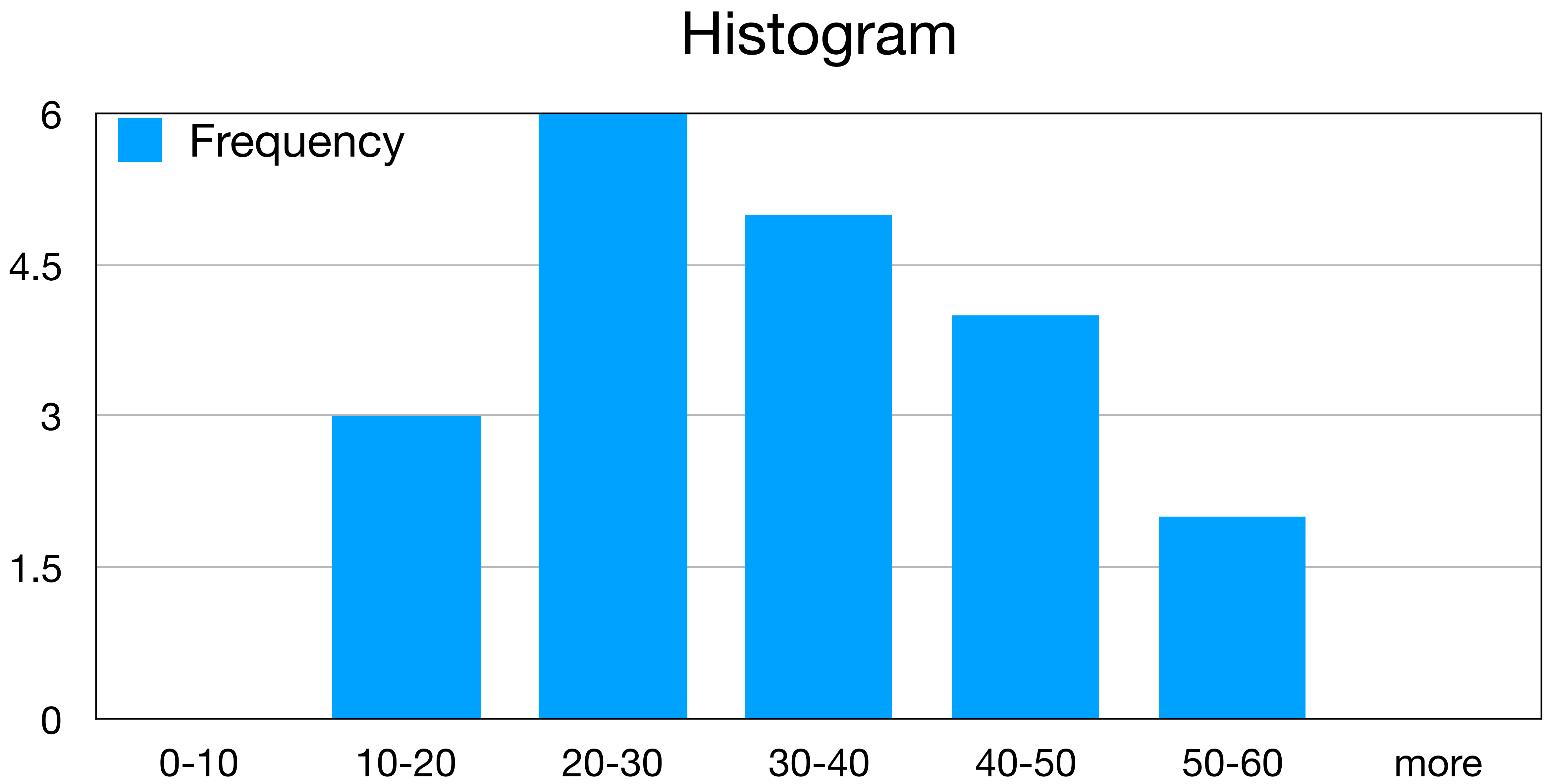
Histograms

- The classes or intervals are shown on the horizontal axis
- frequency is measured on the vertical axis
- Bars of the appropriate heights can be used to represent the number of observations within each class
- Such a graph is called a histogram

Histogram Example

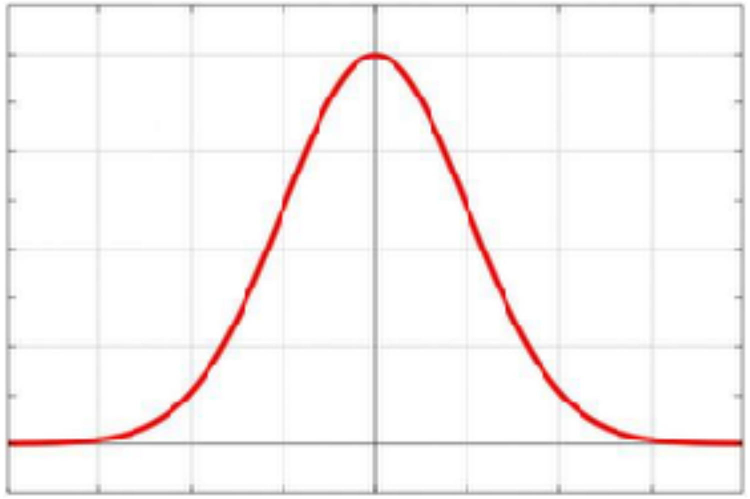
Data in ordered array:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

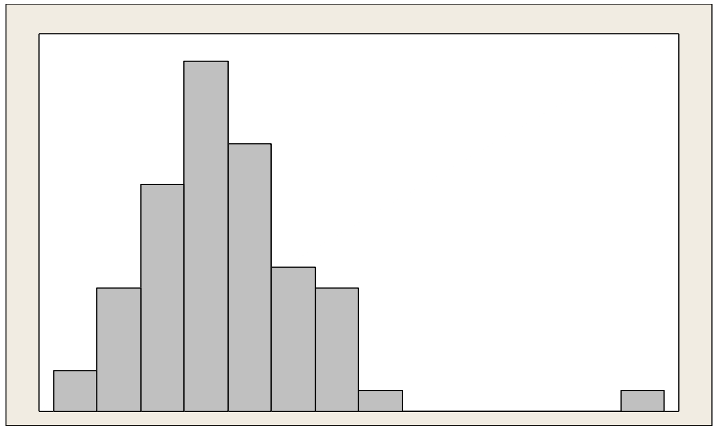


The Histogram

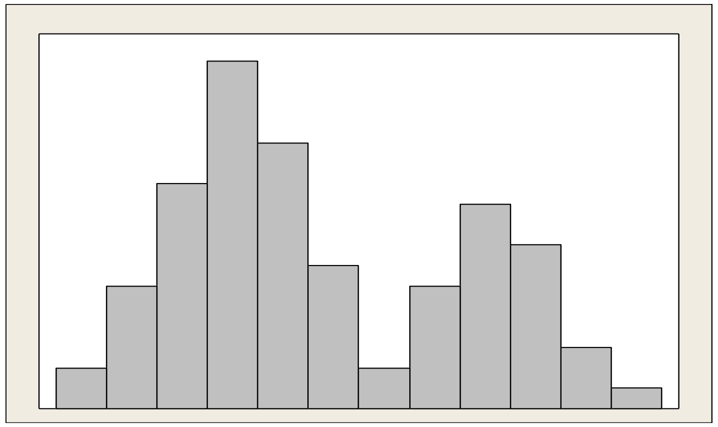
- With a histogram, can detect
- Skewness:
  (here: skewed to the right)



- Outliers:
example: electricity consumption in each level.
And the outlier may are big factories



- Bimodal distribution:



Bar and Pie Charts

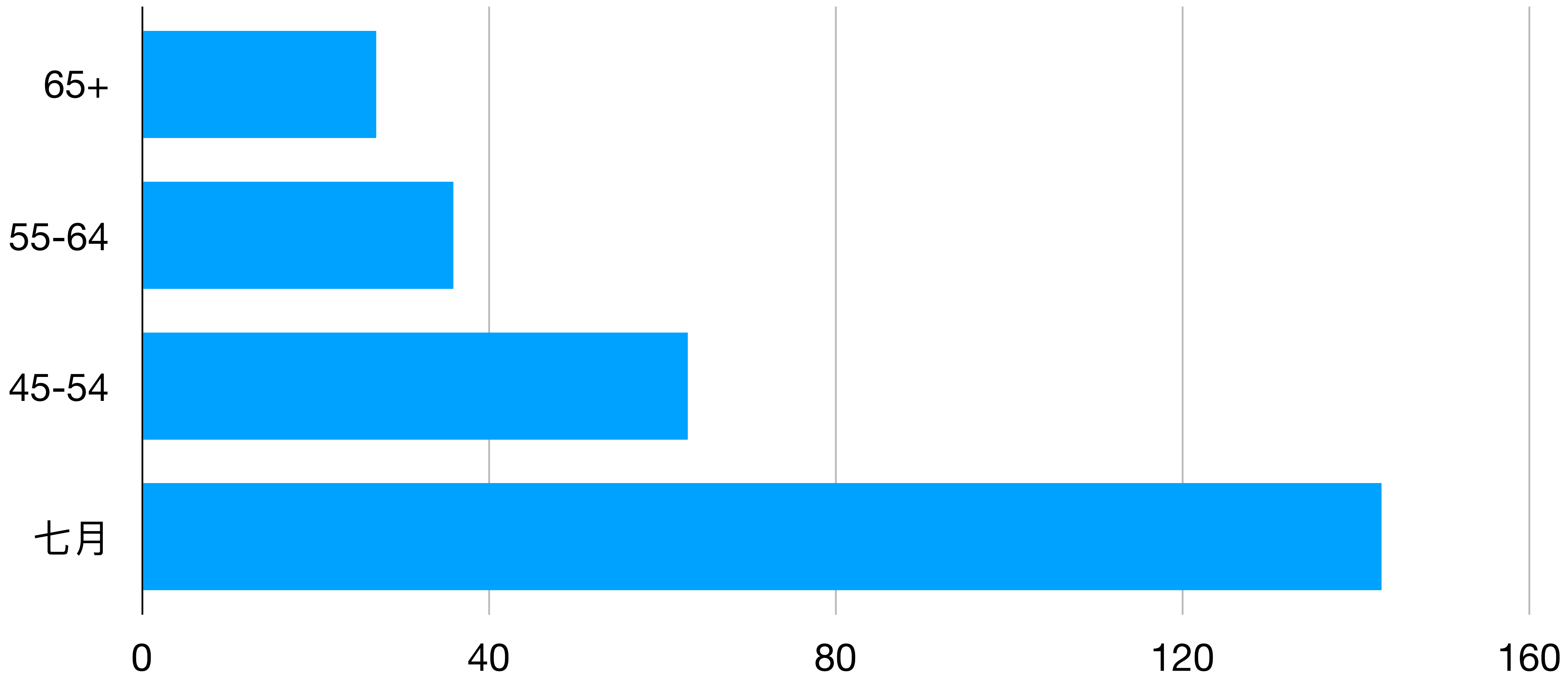
- Bar charts and Pie charts are often used for qualitative (category) data
- Height of bar or size of pie slice shows the frequency or percentage for each category

Bar chart Example:

Use JALC Survey
to display a Bar chart
For the Responder's age

Age	Frequency
Under 18	1
18 - 24	17
25-34	114
35-44	130
45-54	135
55-64	105
65+	40

Example

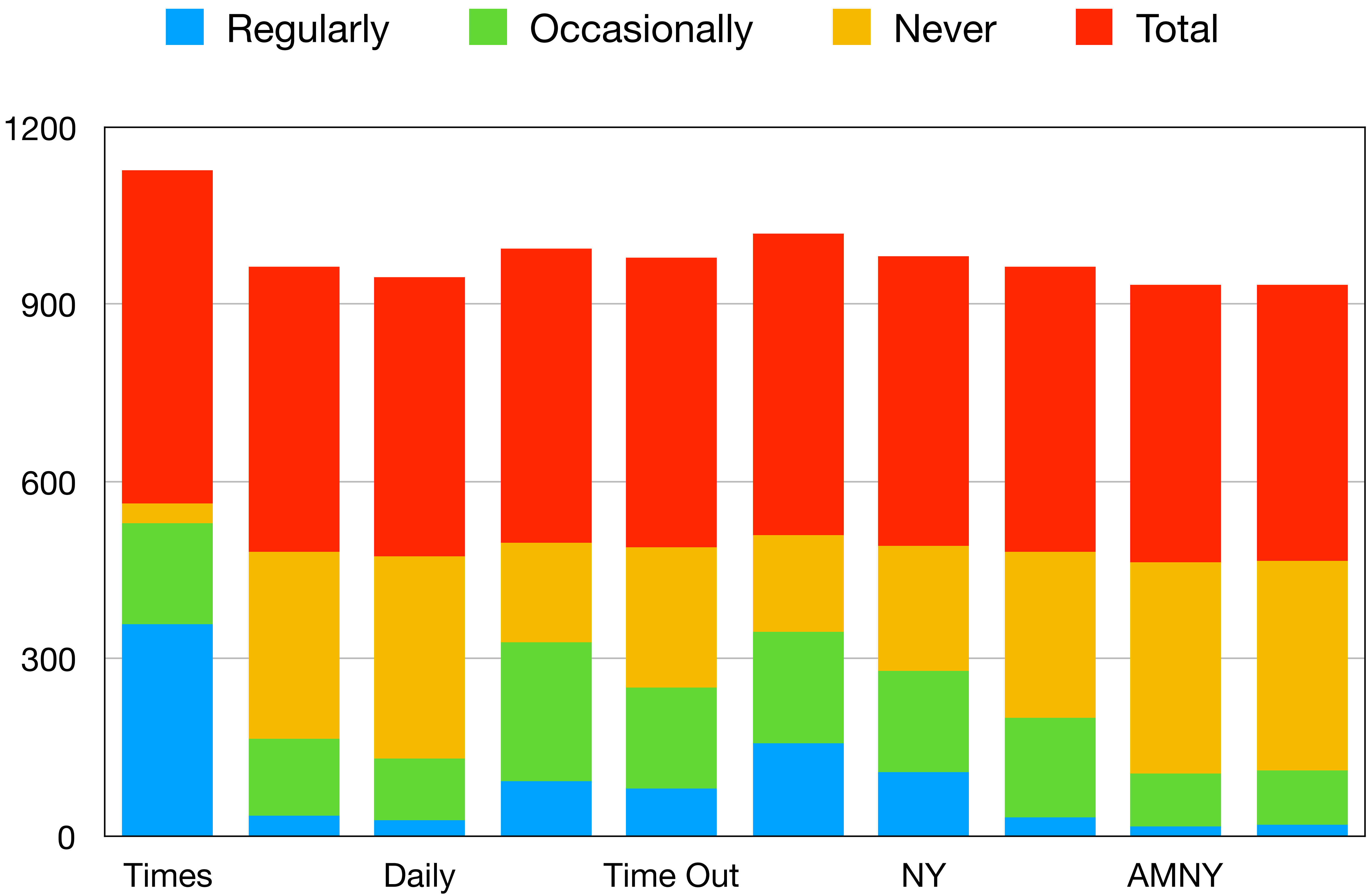


The Stacked Bar Chart

- Example : JALC Survey
- New York area Publications read by respondent

Publication	Regularly	Occasionally	Never	Total
NY Times	359	170	34	563
NY Post	34	130	317	481
Daily News	28	104	340	472
Wall St.	93	235	169	497
Time Out NY	80	172	237	489
New Yorker	157	188	165	510
NY Magazine	108	171	212	491
Village Voice	31	169	282	482
AM NY	16	91	355	471
Metro	19	92	355	466

The Stacked Bar Chart

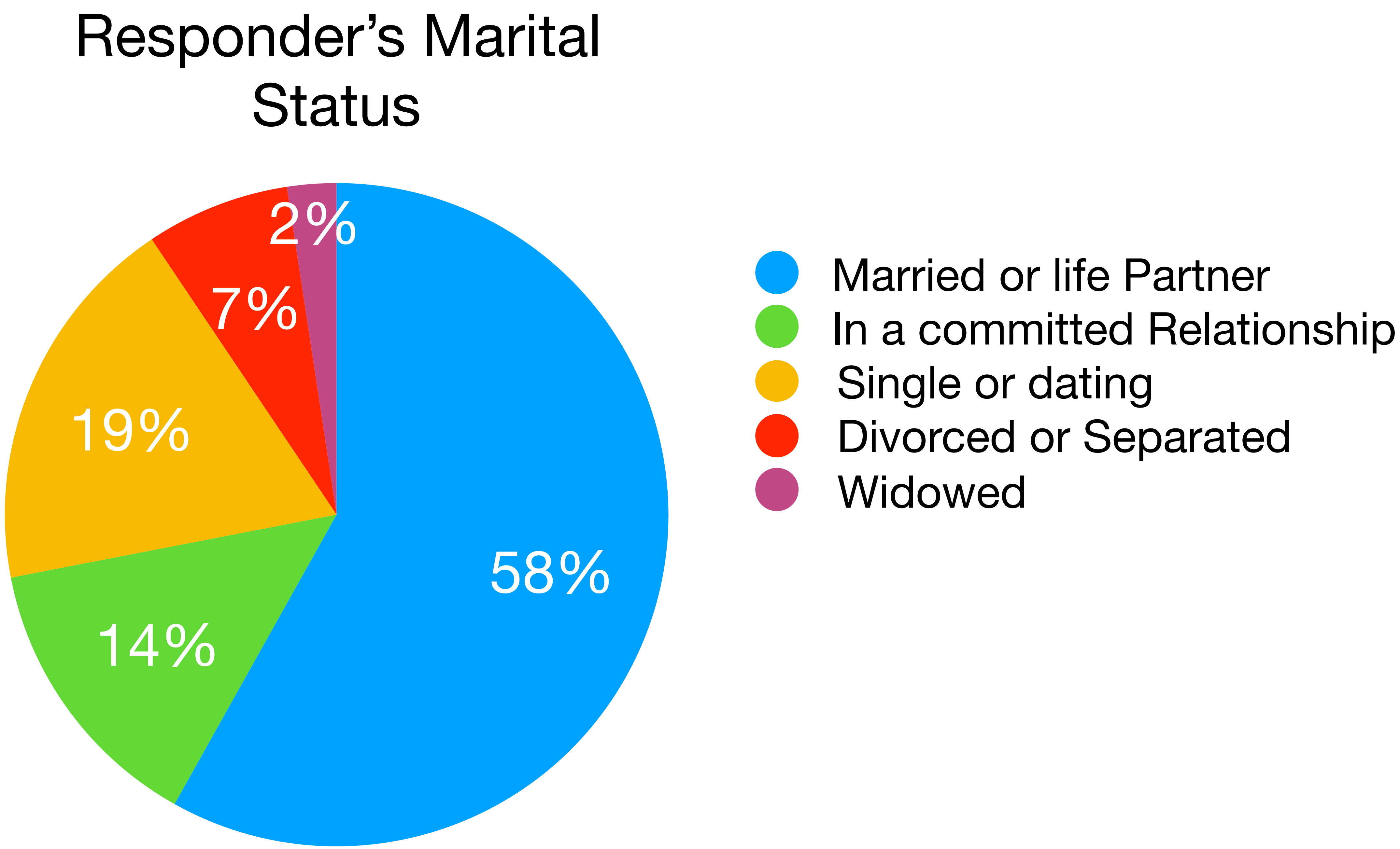


Example

Pie chart Example: Use JALC Survey to display a Pie chart For the Responder's Marital Status.

Marital Status	Response
Married or life Partner	315
In a committed Relationship	75
Single or dating	101
Divorced or Separated	38
Widowed	13
Total	542

Pie Chart



Responder's Marital Status
2%
7%
19%
14%
58%
Married or life Partner
In a committed Relationship
Single or dating
Divorced or Separated
Widowed

Pareto Diagram Example

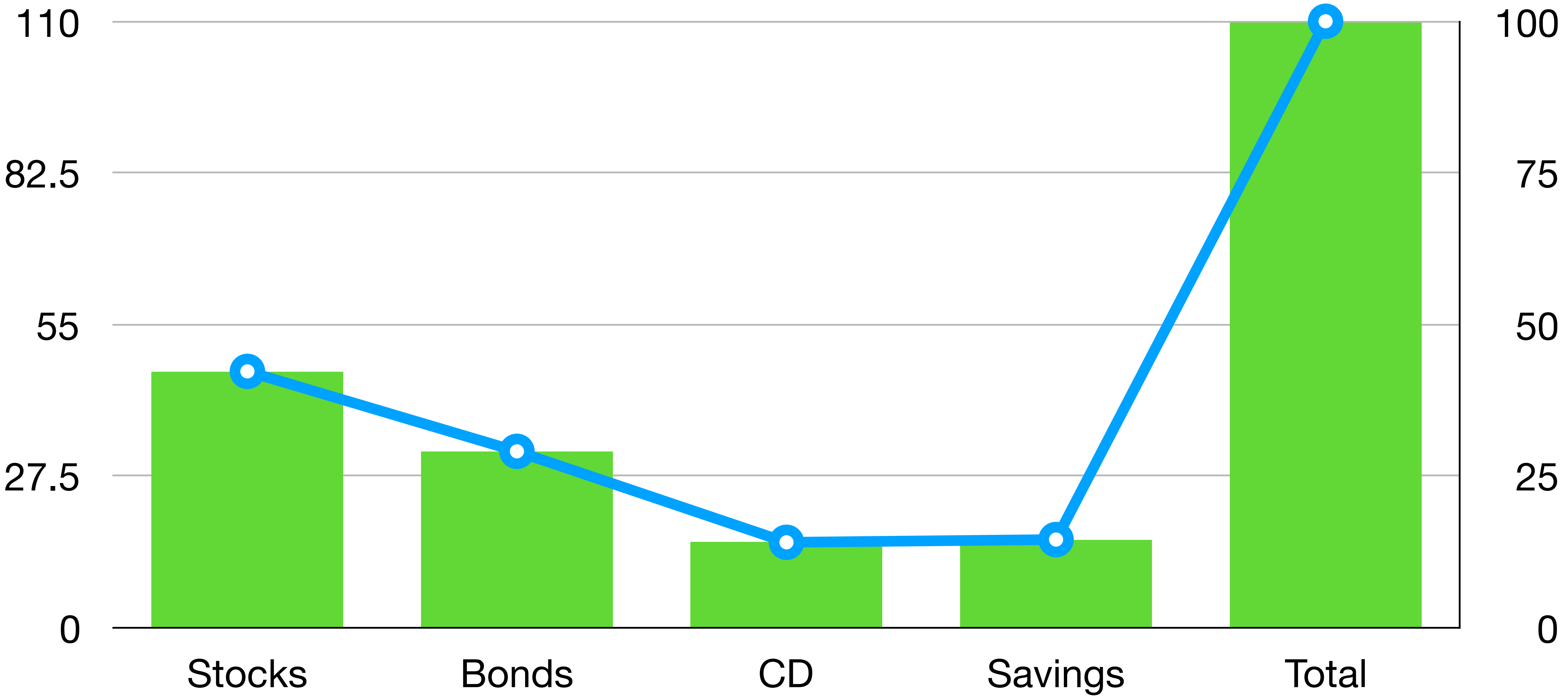
- Similar to a bar chart, but columns are sorted from tallest to shortest, with *cumulative* count

Example:

Current Investment Portfolio

InvestmentType	Amount	Percentage
Stocks	46.5	42.27
Bonds	32.0	29.09
CD	15.5	14.09
Savings	16.0	14.55
Total	110	100

Pareto Diagram Example



The Stem and Leaf Diagram

- Each value has a *stem* and a *leaf* :

28.23 turns into 28 | 2

stem **leaf**

- Leaf is always a single digit, not necessarily the first after the decimal point
- Low-order digits (here the “3”) may be dropped (no rounding, please)

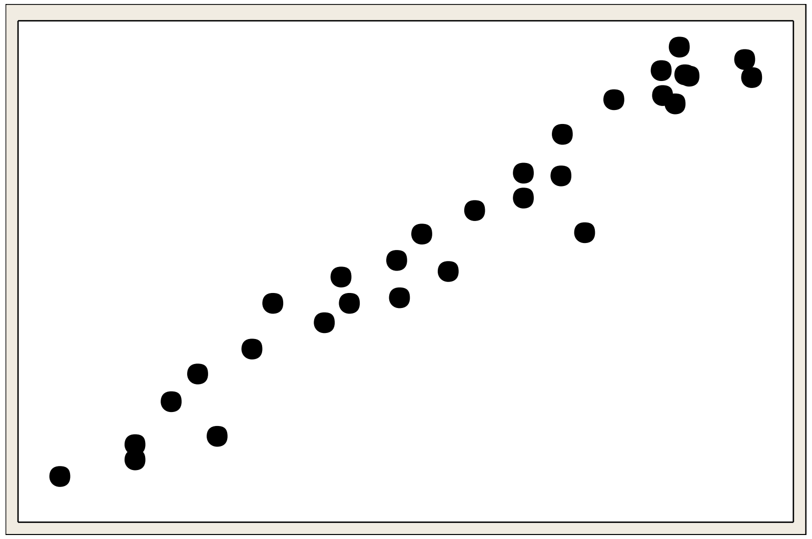
The Stem and Leaf Diagram

Example: Stem and leaf for A the daily high temperature:

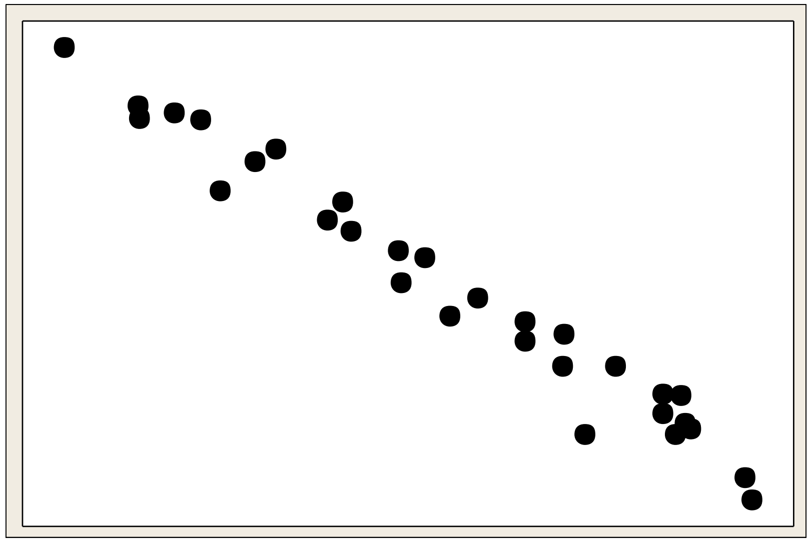
Stem	Leaves
1	2 3 7
2	1 4 4 6 7 7
3	0 2 5 7 8
4	1 3 4 6
5	3 8

The Scatter Plot

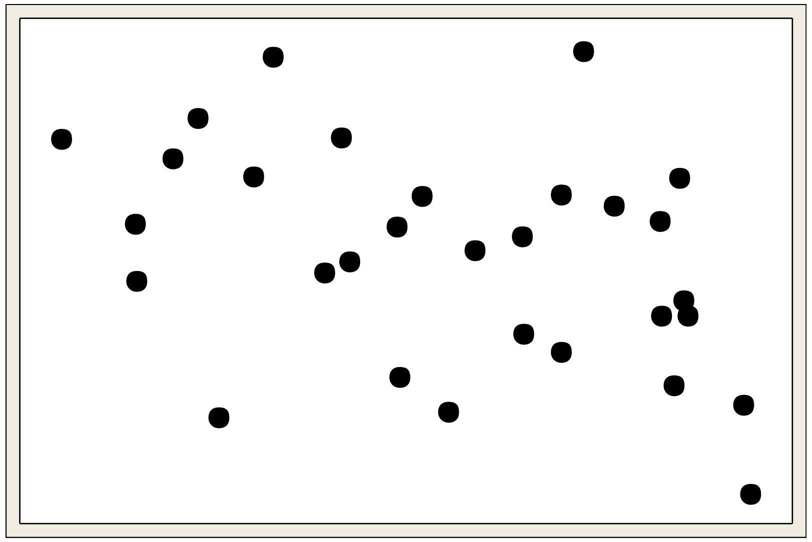
- Positive relation:



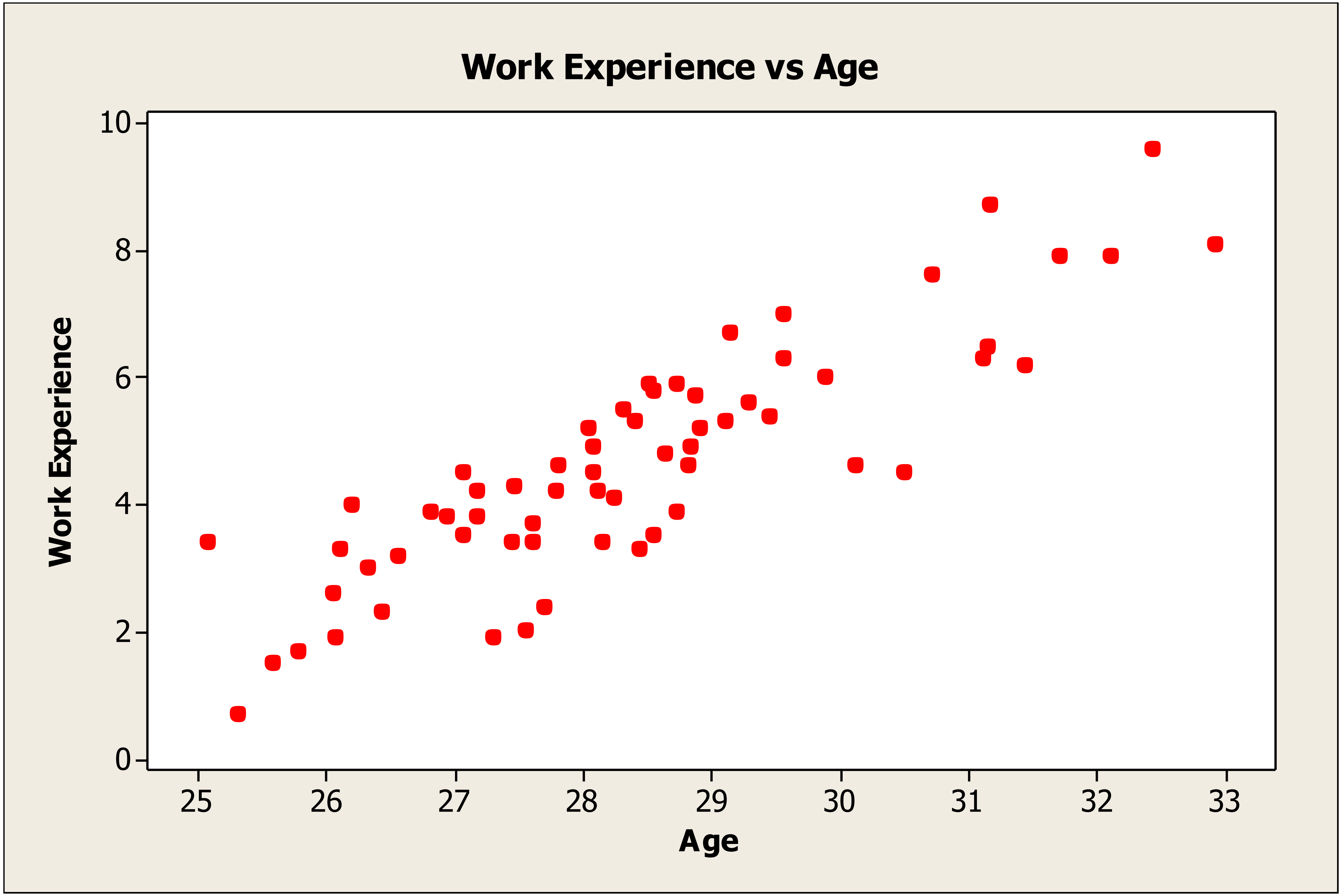
- Negative relation:



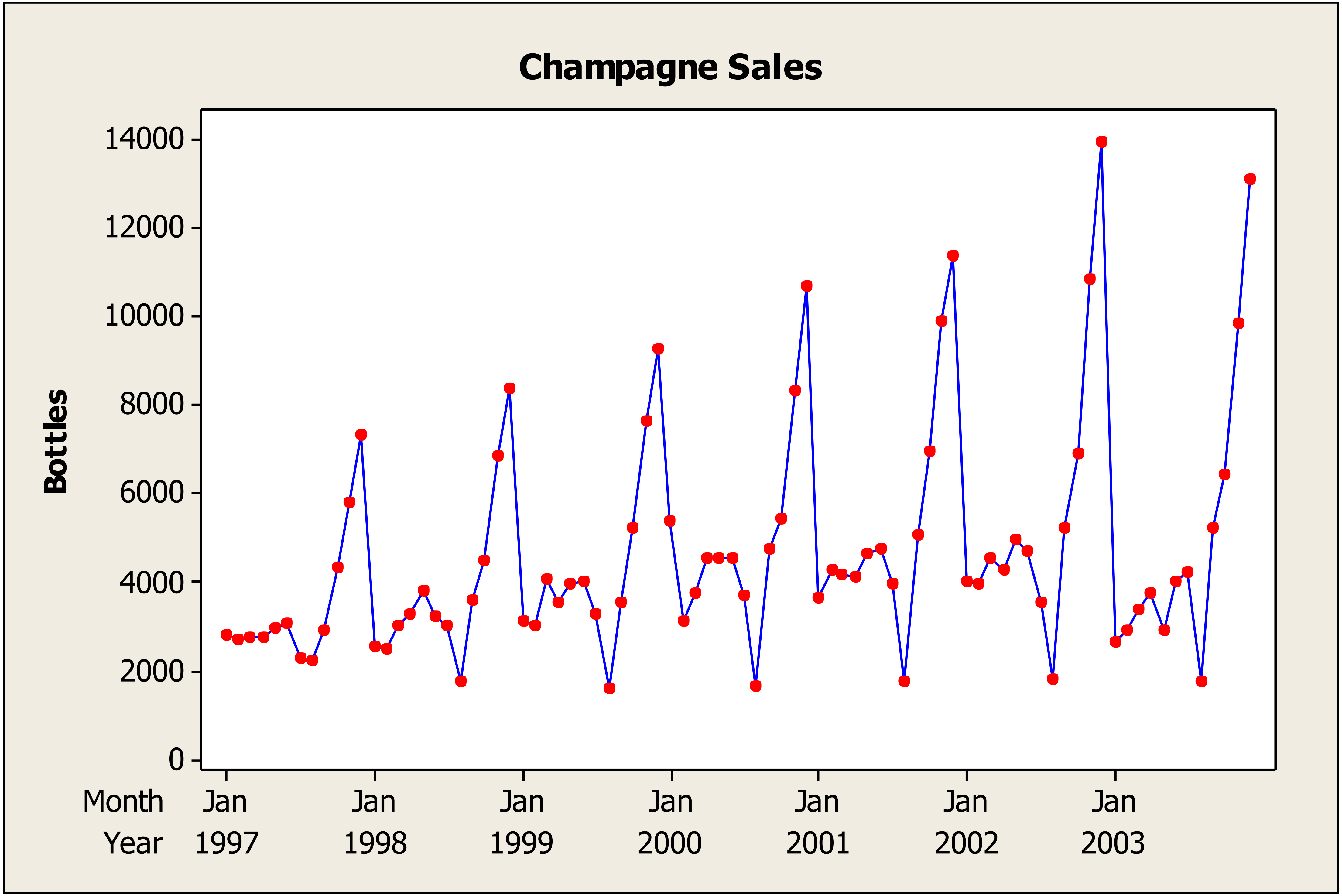
- No relation:



The Scatter Plot



Time Series



Champagne Sales
Bottles
14000
12000
10000
8000
6000
4000
2000
0
Month
Year
Jan
1997
Jan
1998
Jan
1999
Jan
2000
Jan
2001
Jan
2002
Jan
2003

Numerical Methods for Summarizing Data

Measures of Central Tendency or Location

- **Mean**
- **Median**
- **Trimmed Mean**

The Mean

- Have n = 8 numbers:

12	6	13	6	19	8	4	0
Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8

Mean = (12+6+13+6+19+8+4+0)/8 = 8.5

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \qquad \mu = \frac{\sum_{i=1}^{N} y_i}{N}$$

By convention, ÿ and n are used for samples, and μ , N are used for whole populations.

The Median

- Median = the middle value in a *sorted* dataset

0, 4, 6, 6, 8, 12, 13, 19

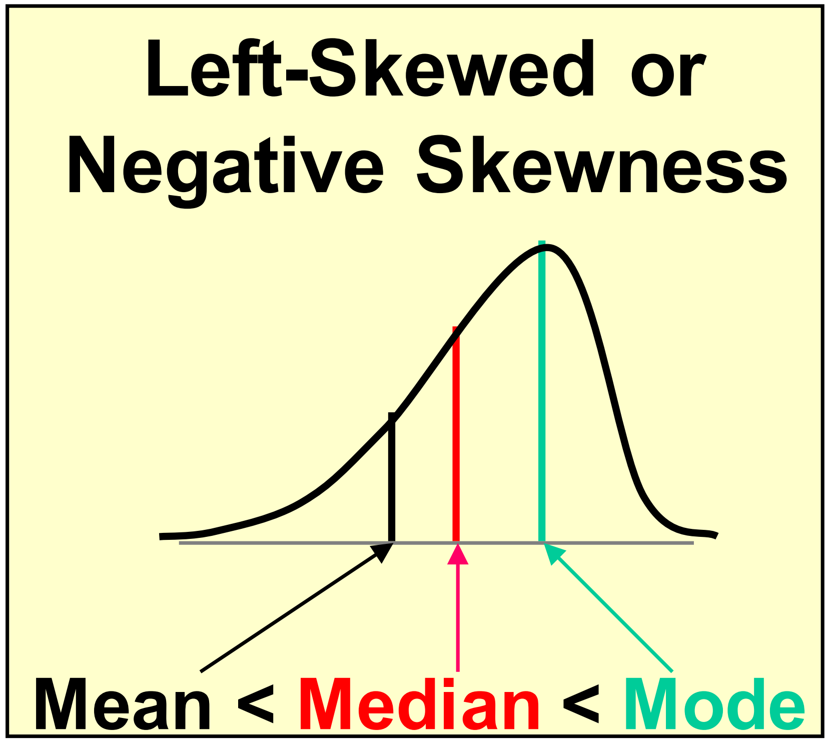
- Note: 6 is listed *twice*
- When n is odd, take obvious middle value.
- When n is even, take average of two middle values.
- In our case: Median = (6+8)/2 = 7

Mean vs. Median

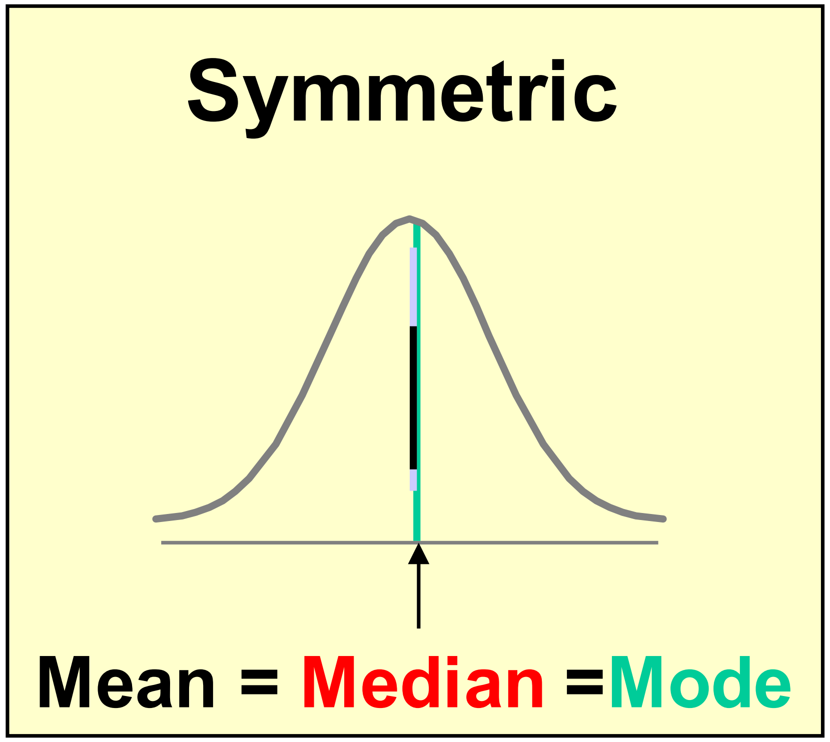
- Mean is
  - Sensitive to outliers (very big or very small values)
  - Useful when interested in long-term average outcomes, and have large dataset
- Median is
  - Useful when ranking is important (GMAT score)
  - Important in demographics
- Other “typical value” measures
  - Mode = the most common value
  - Trimmed mean (ignore upper and lower x% of data)

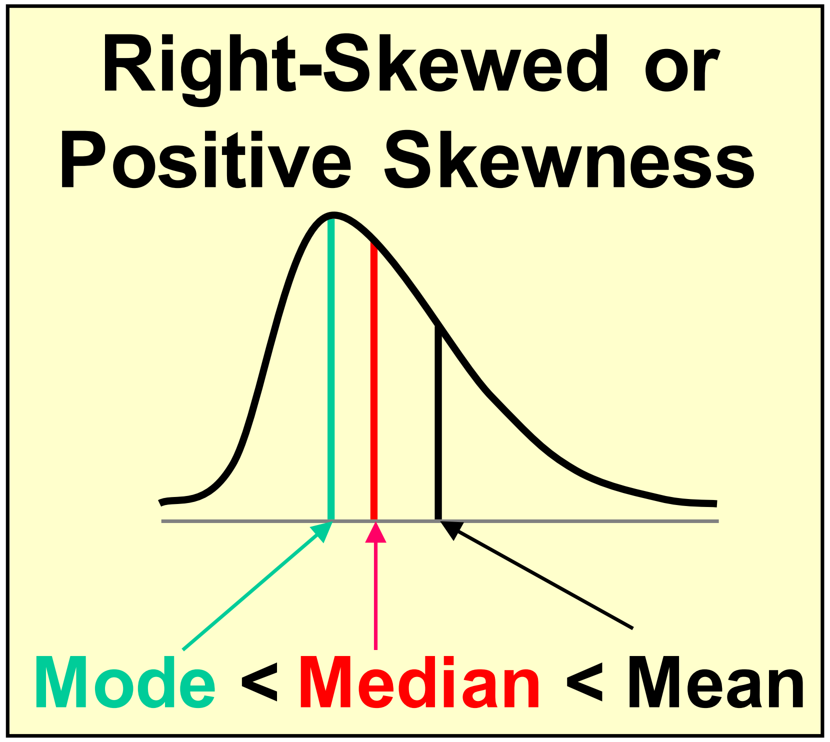
Shape of a Distribution

- Describes how data is distributed
- Symmetric or skewed



(Longer tail extends to left)





(Longer tail extends to right)

Trimmed Mean

- **Trimmed Mean**
- Trim off the largest 5%, for example, and the smallest 5% of the observations and then calculate the sample mean of the remaining 90% of the data values.
- Purpose: Minimize the effect of unusual observation

Measuring Variability

- Four features
- 1.Variance
- 2.Standard Deviation
- 3.Range
- 4.Quartiles

Measuring Variability

- **Sample Variance** (denoted by s^2)

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2$$

Measures squared distance between each observation and the sample mean

- An easier-to-use formula:

$$s^2 = \frac{\sum_{i=1}^{n} y_i^{\,2} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}{n-1}$$

Measuring Variability

Degrees of freedom

In statistics, the number of **degrees of freedom** is the number of values in the final calculation of a statistic that are free to vary.

The number of independent ways by which a dynamic system can move, without violating any constraint imposed on it, is called *number of degrees of freedom*. In other words, the number of degrees of freedom can be defined as the minimum number of independent coordinates that can specify the position of the system completely.

Measuring Variability

- The divisor (n – 1) is the “degrees of freedom”.

Example: Let n=3. Suppose we have three data values.

$$y_1 = 7, \quad y_2 = 3, \quad y_3 = 2 \quad \Rightarrow \quad \bar{y} = 4$$

The building blocks of s^2 are **deviations**:

$$y_1 - \bar{y} = 3, \quad y_2 - \bar{y} = -1, \quad y_3 - \bar{y} = -2$$

- There is **one** constraint on these deviations: S(y$_i$ - y^) = 0.
- For this example, the degrees of freedom = 3-1 = 2.
  - You have freedom to specify any 2 of the 3 deviations.
  - Once you specify any two of the deviations, the third deviation has to be a value so that all deviations add to 0.
- In general, the degrees of freedom = n -1.

Measuring Variability

- **Sample Standard Deviation**

$$s = \sqrt{s^2}$$

To find *s*, don't forget to compute square root

Variance and standard deviation are *always* positive

Measuring Variability

- **Sample Range** (denoted by R)

  R = Largest observation - Smallest observation
- Minimum = smallest value in a dataset
- Maximum = largest value in a dataset
- Don’t have much inferential power
- Always look at them, though, to detect errors
- “The range is very sensitive to outliers …”
- “... as the sample size increases, the range tends to increase …”

Measuring Variability

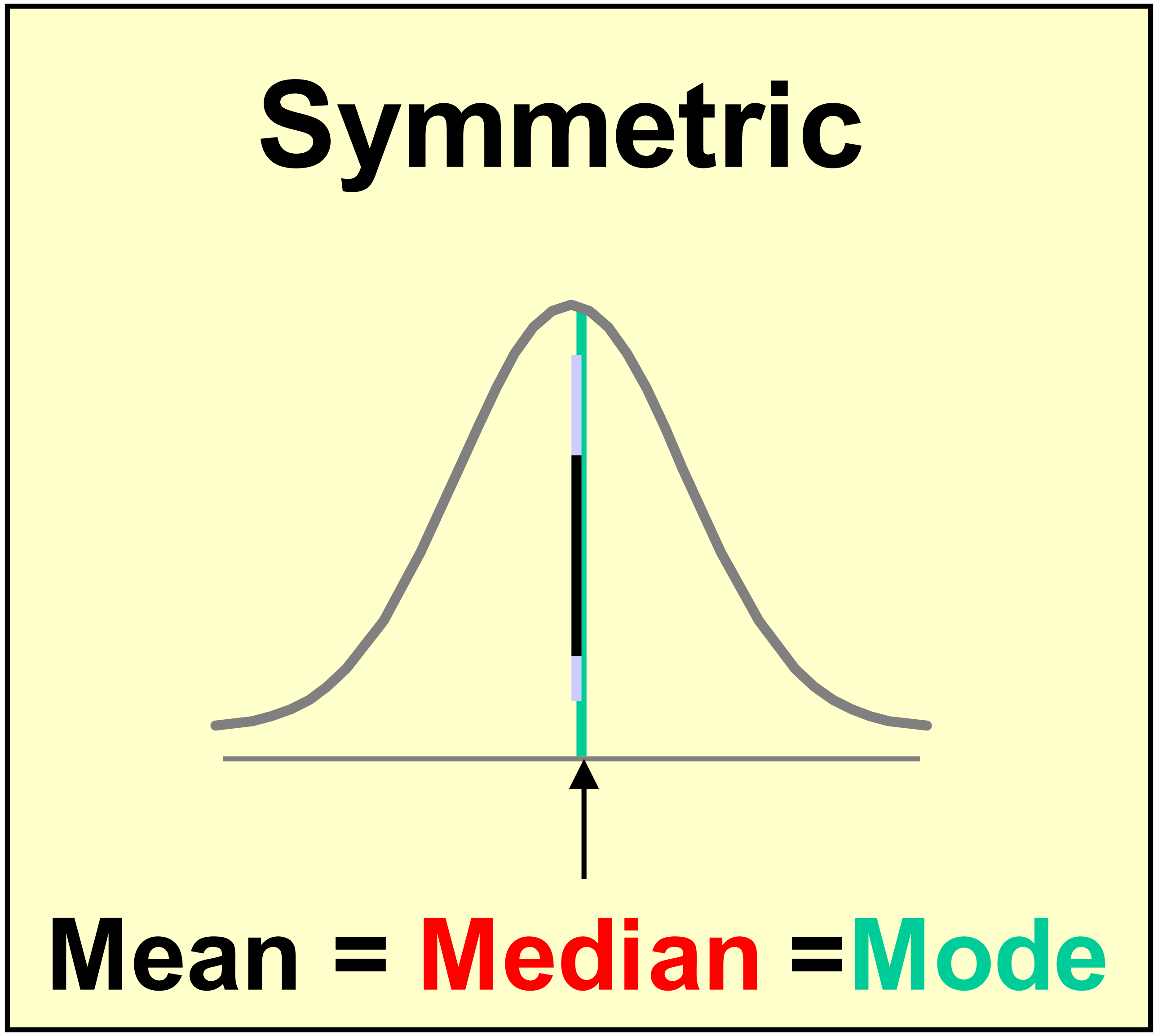
- Linking the histogram with the sample mean and sample standard deviation.
- **Empirical Rule:**

  For a set of measurements having a **mound-shaped histogram**, the interval

$$\bar{y} \pm 1s \textit{ contains approximately } 68\% \textit{ of the measurements;}$$
$$\bar{y} \pm 2s \textit{ contains approximately } 95\% \textit{ of the measurements;}$$
$$\bar{y} \pm 3s \textit{ contains approximately all of the measurements.}$$

- The approximation may be poor if the data are severely skewed or bimodal, or contain outliers.

Example



Measuring Variability

- **Quartiles**
  - Separate data into 4 sections
  - First Quartile = $\mathbf{Q_1} = \mathbf{y_{(k)}}$, where $k = (n + 1) / 4$
    also called 25th percentile
  - Third Quartile = $\mathbf{Q_3} = \mathbf{y_{(3k)}}$
    also called 75th percentile
- How are the quartiles used to measure variability?
- Median is second quartile

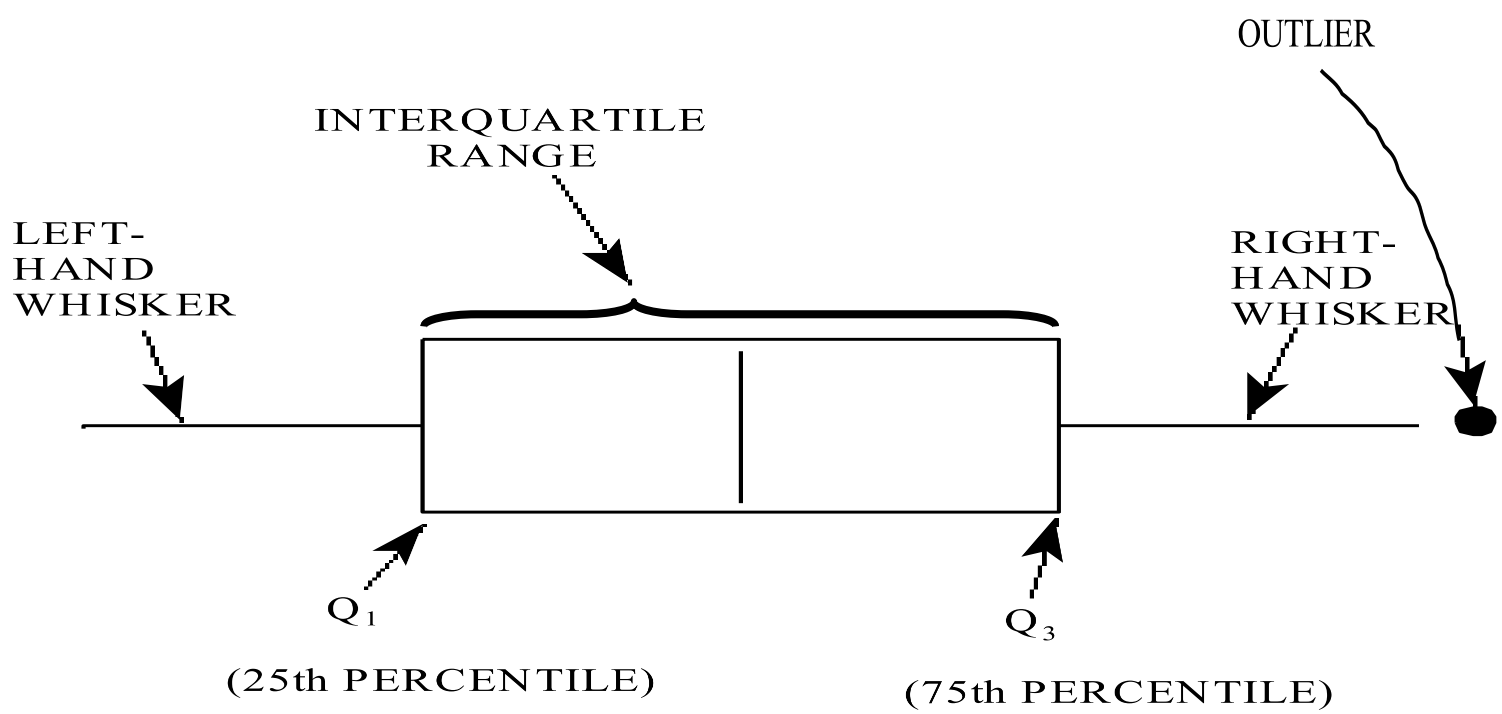
Measuring Variability

- Inter-Quartile Range (IQR) = Q3 – Q1
- Inner fences:
  - Lower inner fence = Q1 – 1.5 IQR
  - Upper inner fence = Q3 + 1.5 IQR
- Outer fences:
  - Lower outer fence = Q1 – 3.0 IQR
  - Upper outer fence = Q3 + 3.0 IQR
  - Data outside inner fence = outlier
  - Data outside outer fence = serious outlier

The “1.5” and “3.0” were decided by John Tukey. In reading box plots, it is not critical to know these.

Measuring Variability

- The **boxplot** uses 5 numbers to represent the data distribution.



Example

A manufacturer of insulation randomly selects 20 winter days and records the daily high temperature

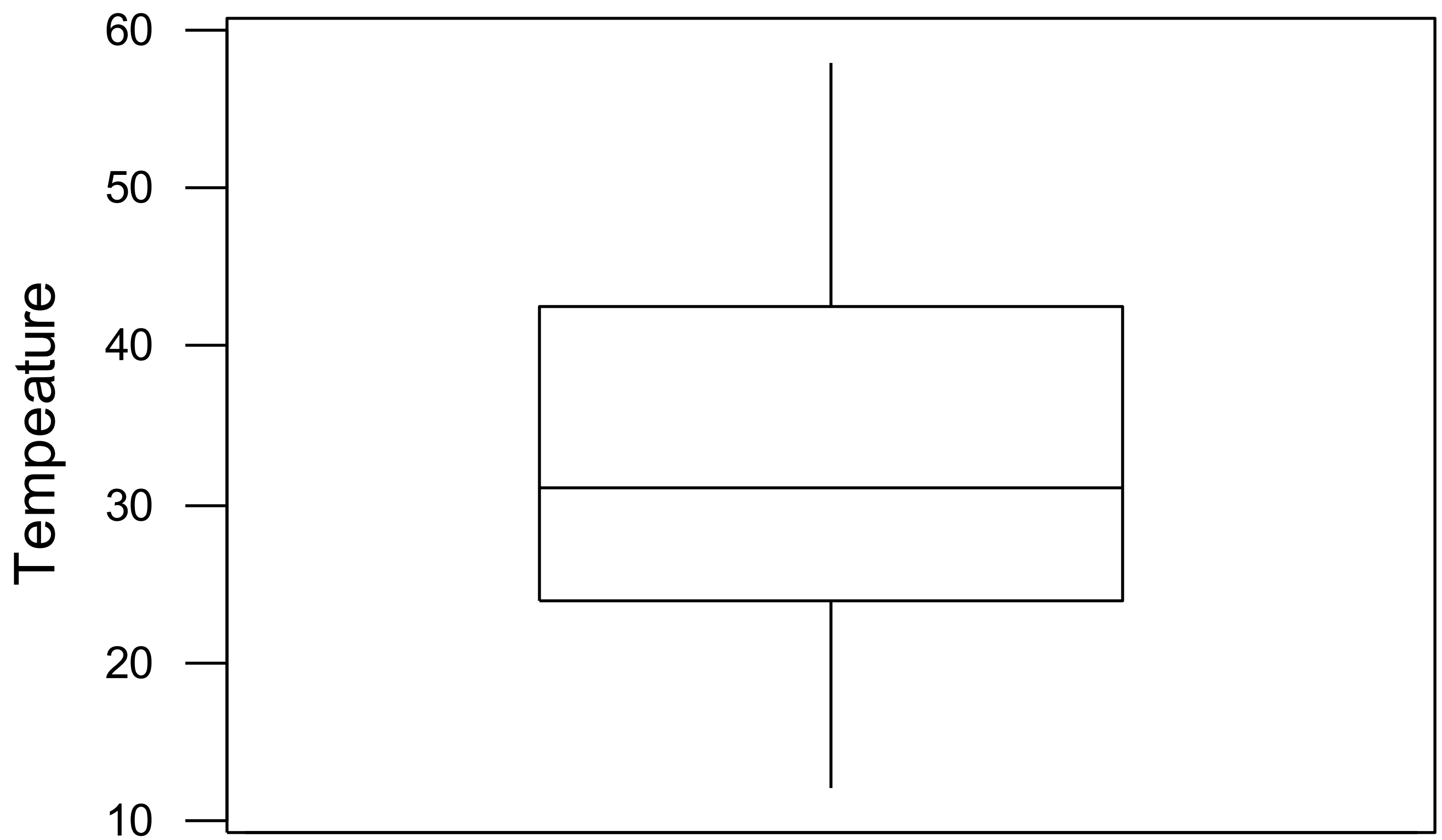
24, 35, 17, 21, 24, 37, 26, 46, 58, 30, 32, 13, 12, 38, 41, 43, 44, 27, 53, 27

Descriptive Statistics: Temperature

Variable	N	Mean	Median	TrMean	StDev	SE Mean
Temperature	20	32.40	31.00	32.11	12.67	2.83

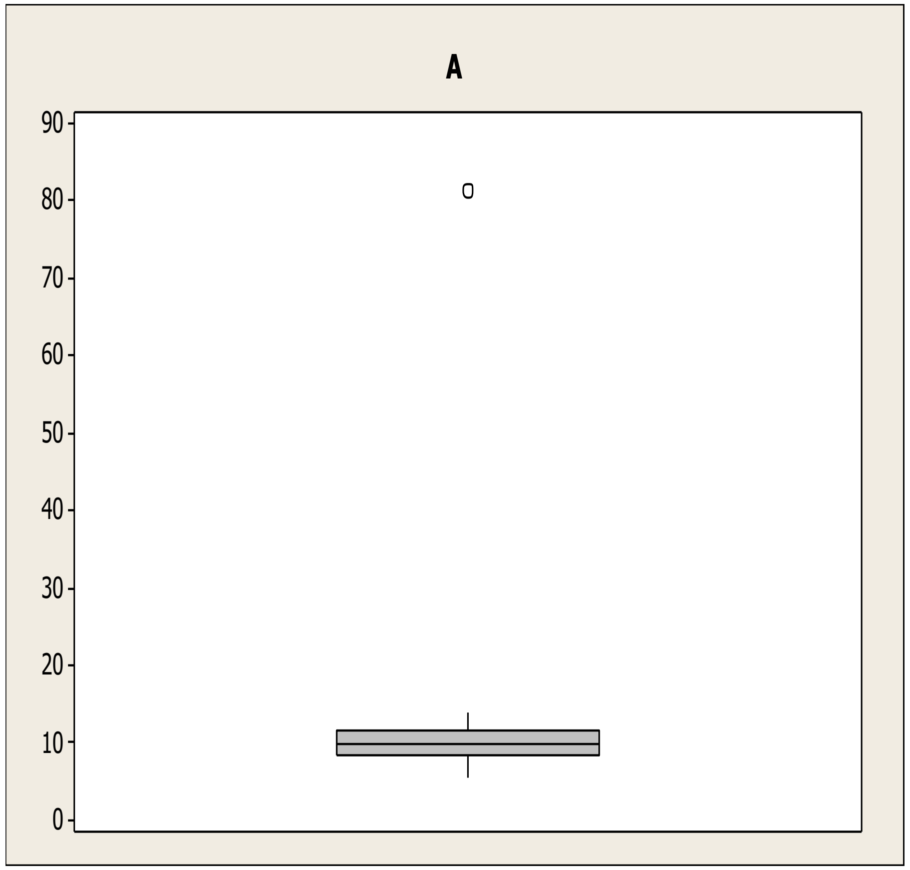
Variable	Minimum	Maximum	Q1	Q3
Temperature	12.00	58.00	24.00	42.50

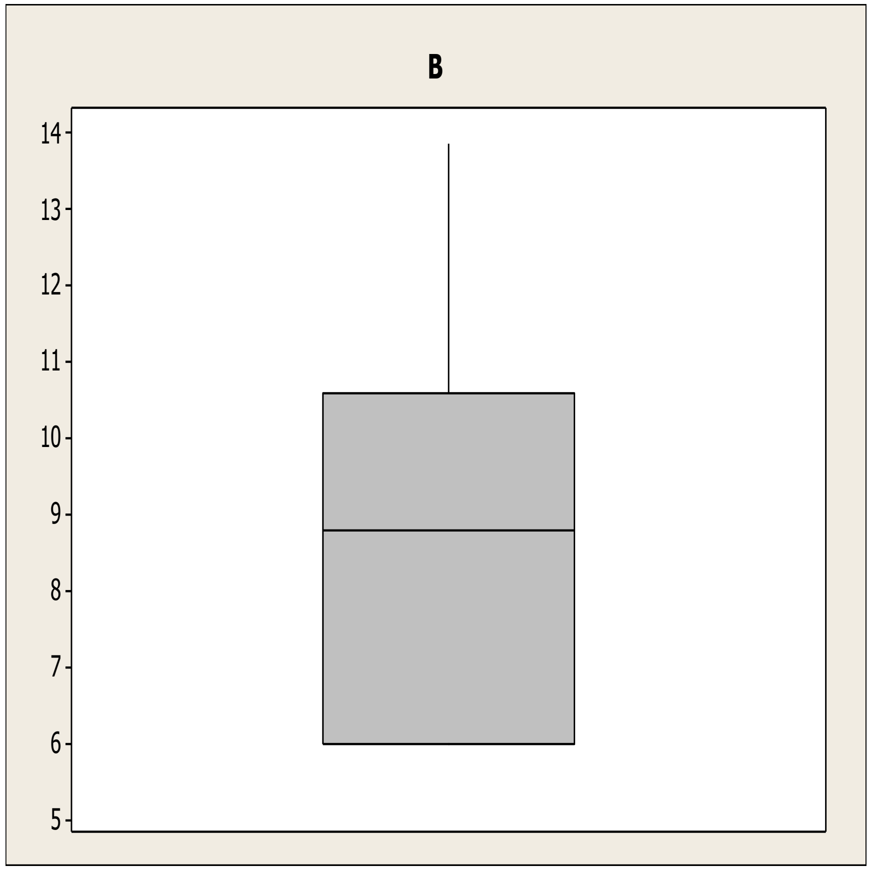
Example



The Box Plot

- Excellent for comparing datasets
- Box plots “gone bad”:





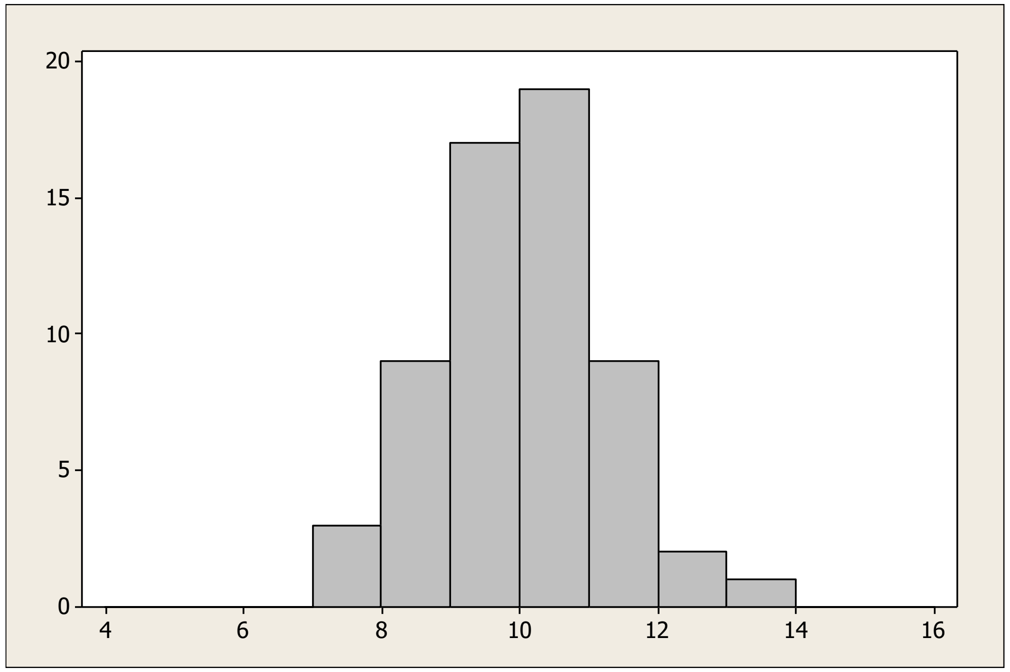
Measuring Variability

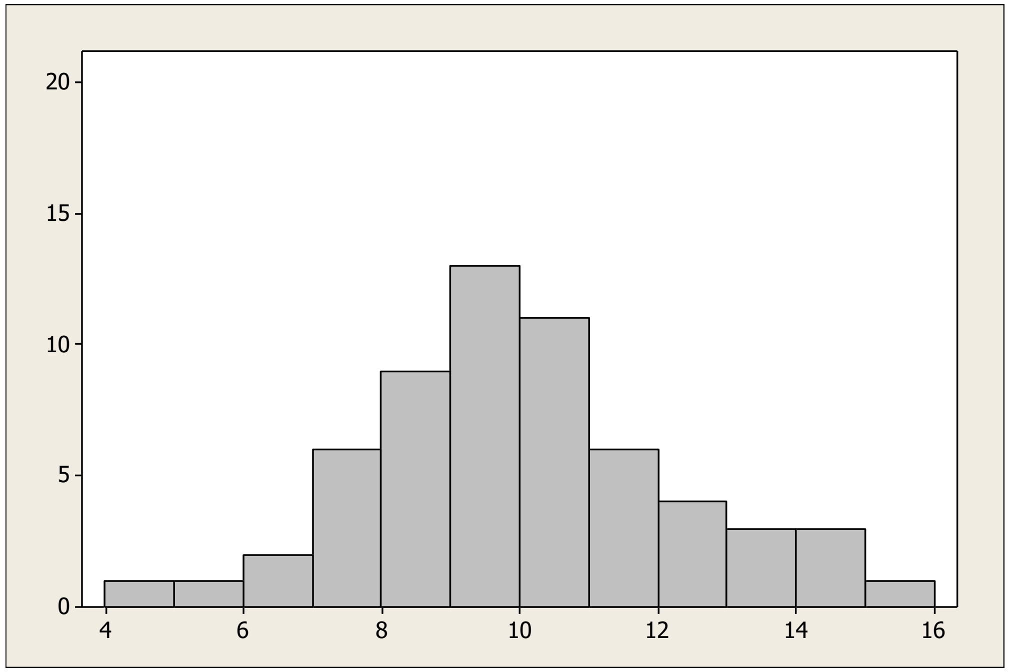
- IQR = Q3 – Q1
- Range = max – min
- The most important measure of variability:
  s = standard deviation

- Empirical rule:
  - About 2/3 of the data is between – s and + s
  - About 95% of the data is between – 2s and + 2s
- s^2 = variance

Variability

- Two datasets – same mean, but different *variability*





- Is variability “good” or “bad”?

Probability

Important Terms

- Probability – the chance that an uncertain event will occur (always between 0 and 1)
- Experiment – a process of obtaining outcomes for uncertain events
- Elementary Event – the most basic outcome possible from a simple experiment
- Sample Space – the collection of all possible elementary outcomes

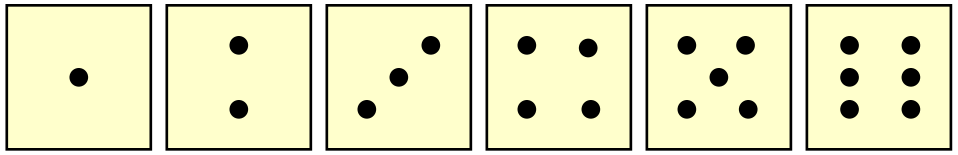
Events

- Elementary event – An outcome from a sample space with one characteristic
  - Example: A red card from a deck of cards
- Event – May involve two or more outcomes simultaneously
  - Example: An ace that is also red from a deck of cards

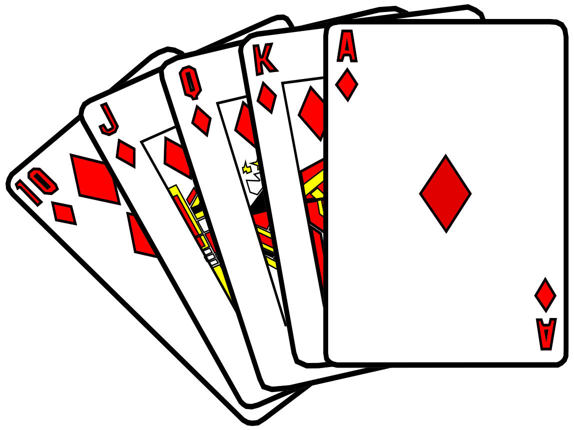
Sample Space

The sample space is the collection of all possible outcomes

e.g. All 6 faces of a die:

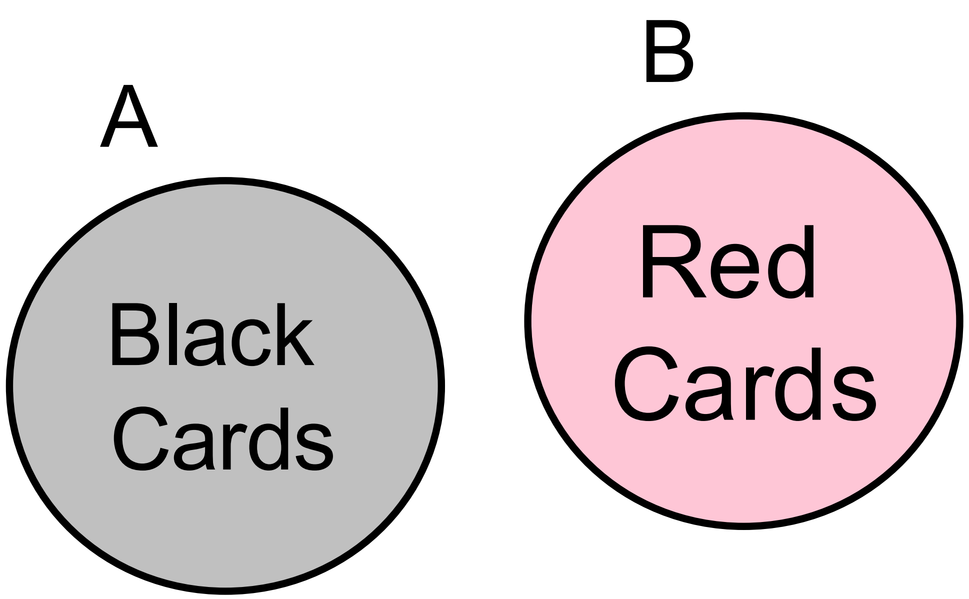


e.g. All 52 cards of a bridge deck:



Events

- **Mutually Exclusive Events:**
- If A occurs, then B cannot occur
- A and B have no common elements



A card cannot be Black and Red at the same time.

Events

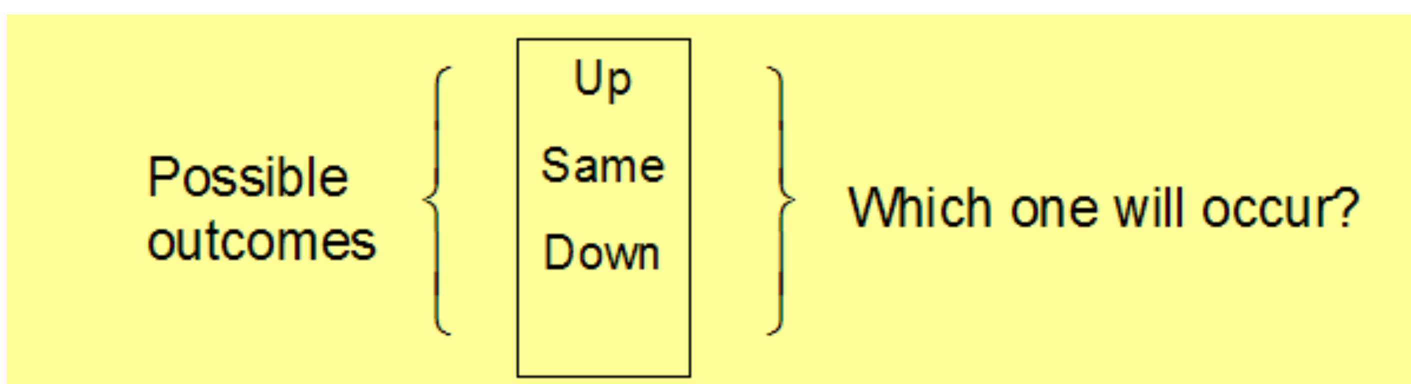
- Inevitable events：
- In every experiment the event will happen
- Impossible events
- In every experiment the event will never happen

Basic Principles of Probability

- Synonyms for probability : chance , likelihood.

- Only consider random experiments
- Characteristics of random experiments
  - Can specify all possible outcomes.
  - Cannot predict a specific outcome with certainty.

Example: Today's closing price of a security relative to yesterday.



- Probability measures the **uncertainty** of the outcomes.

Basic Principles of Probability

- Three interpretations of probability

  1. The **classical interpretation**

  N possible mutually exclusive equally likely outcomes; the probability of an event E equals the ratio of the number of outcomes (N_E) pertaining to E to the total number of outcomes (N):

  $$\mathbf{P(E) = N_E / N}$$

<u>Example:</u> Roll a fair die one time and observe the up face

P(rolling a 6) = 1/6

Roll a pair of fair dice and observe the up faces

P(both up faces are 6’s) = 1/36 = (1/6) (1/6)

Basic Principles of Probability

2. The **relative frequency interpretation**

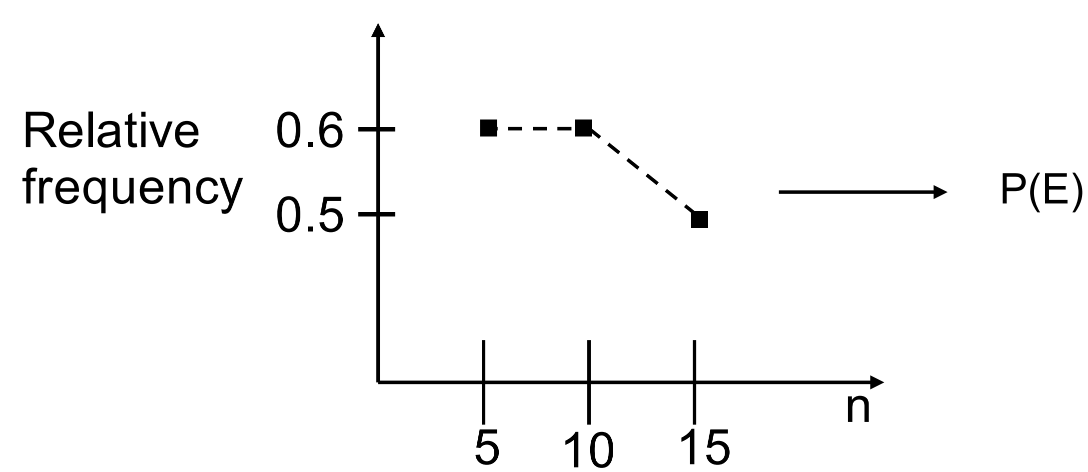
If an experiment has been repeated n times under identical conditions, and n_e of these trials have resulted in event E, then the **relative frequency** of event E is n_e/n. Assume the experiment can be repeated *indefinitely* under identical conditions. Under these assumptions, the long-run relative frequency with which event E occurs is the probability of event E:

$$\mathbf{P(E) = \text{limit } (n_e/n) \text{ as } n \longrightarrow \infty}$$

$$\approx \mathbf{n_e/n,} \text{ for n sufficiently large.}$$

<u>*Example:*</u> Flipping a coin to find the P (head).

Number of heads
<u>In each of 5 tosses</u>



Basic Principles of Probability

3. The **subjective interpretation**

The subjective probability of an event E is reached when you are **indifferent** between wagering on this event or on the drawing of a red bead from an urn in which a fraction n_e/n of the beads are red.

An opinion or judgment by a decision maker about the likelihood of an event.

Example: What is the probability that the Ravens will win the Super Bowl in 2009? Suppose you say .2.

Rules of Probability

Rules for Possible Values and Sum

Individual Values

$0 \leq P(A) \leq 1$

For any event A

Sum of All Values

$$\sum_{i=1}^{k} P(e_i) = 1$$

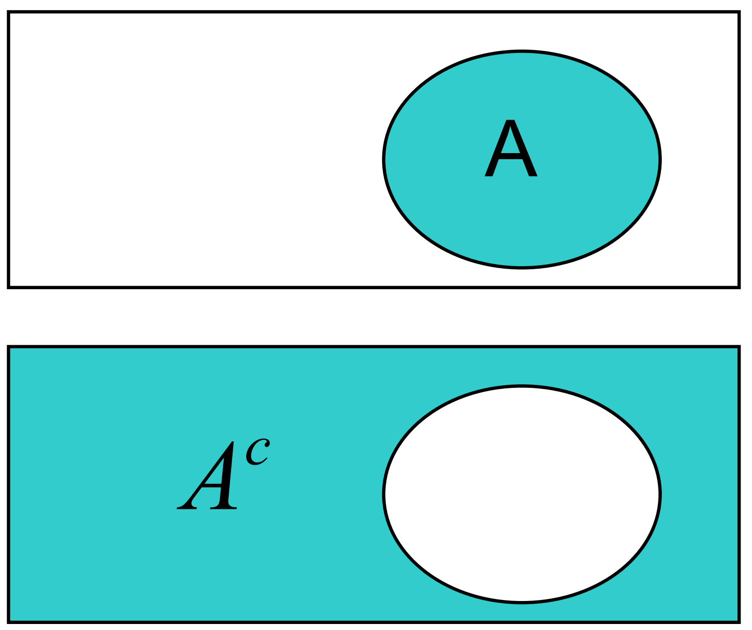
where:

k = Number of elementary events in the sample space

e_i = i^{th} elementary event

Complement Rule

- The complement of an event **A** is the collection of all possible elementary events **not** contained in event **A**. The complement of event **A** is represented by   .
- Complement Rule:



$$P(A^c) = 1 - P(A)$$

Or, $P(A) + P(A^c) = 1$

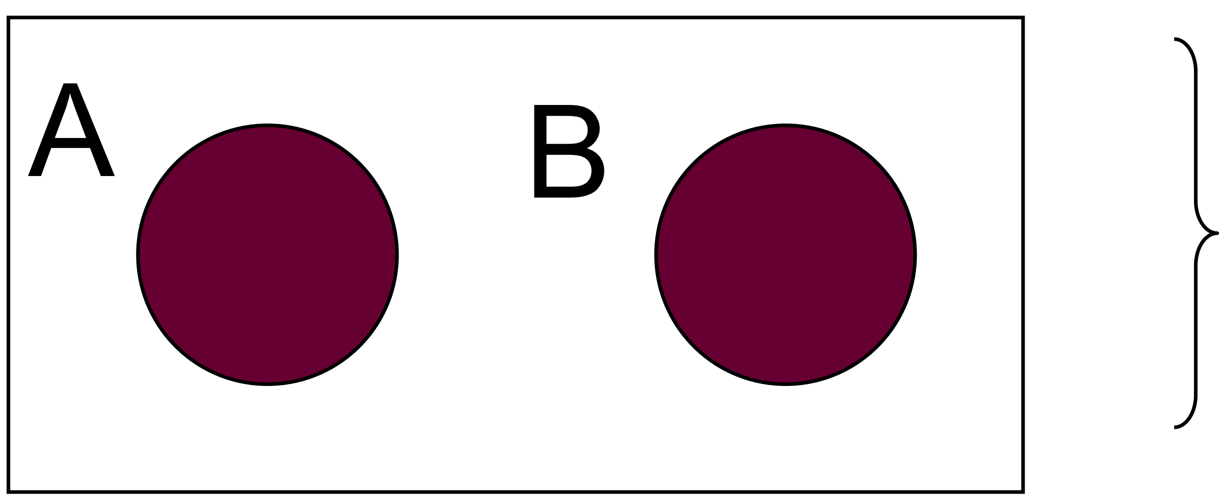
Basic Principles of Probability

- Let A and B denote any events in a random experiment.
- **P(A) ^{3} 0, P(B) ^{3} 0**
- **Addition Law for Mutually Exclusive Events**

  If A and B are mutually exclusive events

  (A and B = empty set), then

  P(A or B) = P(A) + P(B)

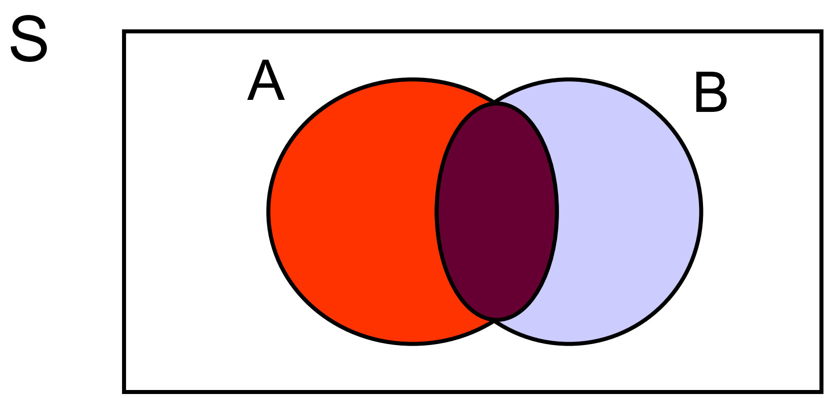


Venn diagram

Basic Principles of Probability

- General Addition Law

P(A or B) = P(A) + P(B) – P(A and B)



(A and B) has been counted twice
==>Subtract once

Conditional Probability

Conditional probability of event A occurring given that event B has occurred, denoted by P(A|B), is:

P(A|B) = P(A and B) / P(B),

provided P(B) > 0.

Addition Rule Example

P(Red or Ace) = P(Red) +P(Ace) - P(Red and Ace)

$= 26/52 + 4/52 - \boxed{2/52} = 28/52$

Type	Color		Total
	Red	Black	
Ace	2	2	4
Non-Ace	24	24	48
Total	26	26	52

Don't count the two red aces twice!

Conditional Probability Example

- Of the cars on a used car lot, 70% have air conditioning (AC) and 40% have a CD player (CD). 20% of the cars have both.
-

What is the probability that a car has a CD player, given that it has AC ?

i.e., we want to find $P(CD \mid AC)$

Example

- Given AC, we only consider the top row (70% of the cars). Of these, 20% have a CD player. 20% of 70% is about 28.57%.

$$P(\mathrm{CD} \mid \mathrm{AC}) = \frac{P(\mathrm{CD\ and\ AC})}{P(\mathrm{AC})} = \frac{0.2}{0.7} = 0.2857$$

	CD	Not CD	Total
AC	0.2	0.5	0.7
No AC	0.2	0.1	0.3
Total	0.4	0.6	1

Statistical Independence

- Multiplication Rule

From the definition of conditional probability,

P(A and B) = P(B) P(A|B)

Obviously,

P(A and B) = P(A) P(B|A)

<u>Example:</u> Randomly pick (without replacement) 2 cards from a standard deck. Find probability of 2 hearts.

A = {1st card is a heart}, B = {2nd card is a heart}

P (A and B) = P(A) P(B|A) = (13/52) (12/51)

- The multiplication rule is useful in Probability Trees.

Probability Trees

- Probability trees

  In a **probability tree**, the probability for a specific path is found by using the multiplication rule.

Exercise:

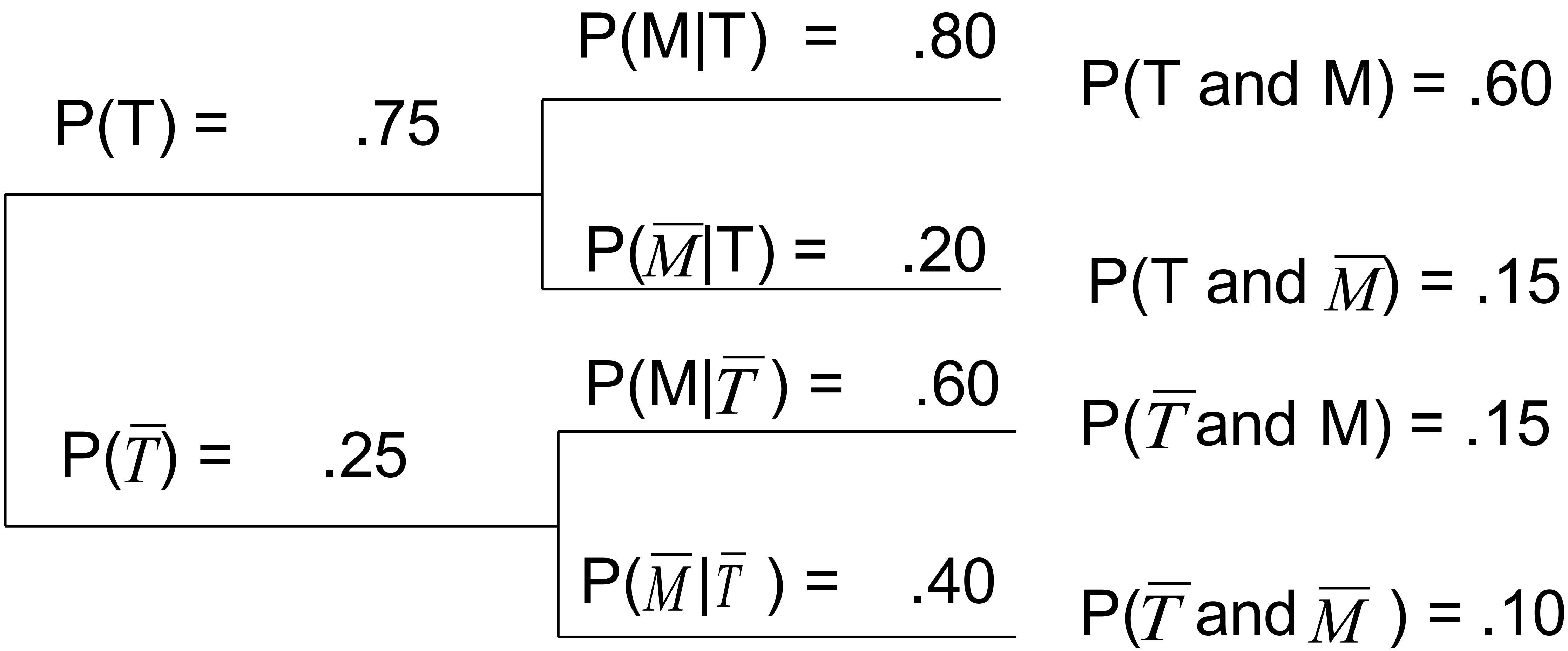
A purchasing dept. finds that 75% of its special orders are received on time. Of those orders that are on time, 80% meet specifications completely; of those orders that are late, 60% meet them.

Find the probability that an order is on time and meets specifications.

T = {Order is on time} M = {Meets specifications}

P(T) = .75 P(M|T) = .80 P $(M|\bar{T})$ = .60

Probability Trees



Statistical Independence

- Multiplication Law for Independent Events
- If A and B are **independent**, then
  P(A and B) = P(A) P(B)

- Reason: From Multiplication Rule,
  P(A and B) = P(A|B) P(B)
  From independence,
  P(A|B) = P(A)
  P(A and B) = P(A) P(B)

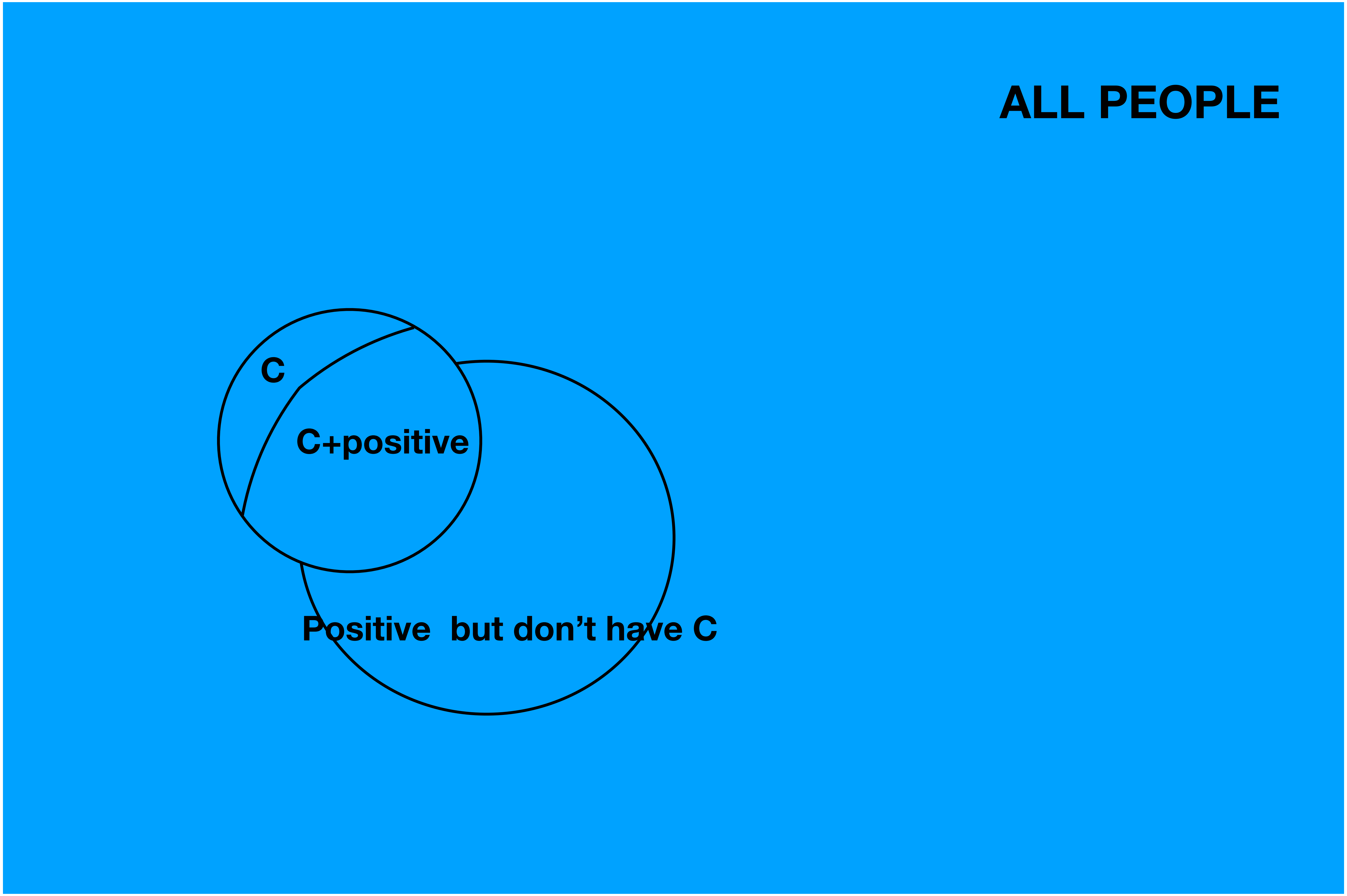
Bayes' Rule

- Need a way to relate P(A | B) to P(B | A)
- Bayes' rule:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes' Theorem Example

- P(C) = 0.01 （there's a specific cancer that occurs in one percent of the population.）
- Test :
- 90% it is positive if you have C (sensitive)
- 90% it is negative if you don't have C (specificity)
- Question : Test = Positive
- What is the probability of having cancer


ALL PEOPLE
C
C+positive
Positive but don't have C

Bayes' Rule

- Prior probability + Test evidence -> Posterior probability

Bayes' Theorem Example

- A drilling company has estimated a 40% chance of striking oil for their new well.
- A detailed test has been scheduled for more information. Historically, 60% of successful wells have had detailed tests, and 20% of unsuccessful wells have had detailed tests.
- Given that this well has been scheduled for a detailed test, what is the probability that the well will be successful?

Example

- Let S = successful well and U = unsuccessful well
- P(S) = .4 , P(U) = .6   (prior probabilities)
- Define the detailed test event as D
- Conditional probabilities:

  P(D|S) = .6   P(D|U) = .2

$$P(S|D) = \frac{P(D|S)P(S)}{P(D)}$$

Example

- P(D)=?
- P(D and S)= P(D/S)P(S)=(0.6)(0.4)=0.24
- P(D and U)= P(D/U)P(U)=(0.2)(0.6)=0.12
- P(D)=0.36

$$P(S|D) = \frac{P(D|S)P(S)}{P(D)} = \frac{0.24}{0.36} = 0.67$$