Pattern Recognition

Basics of Probability and Statistics

Content

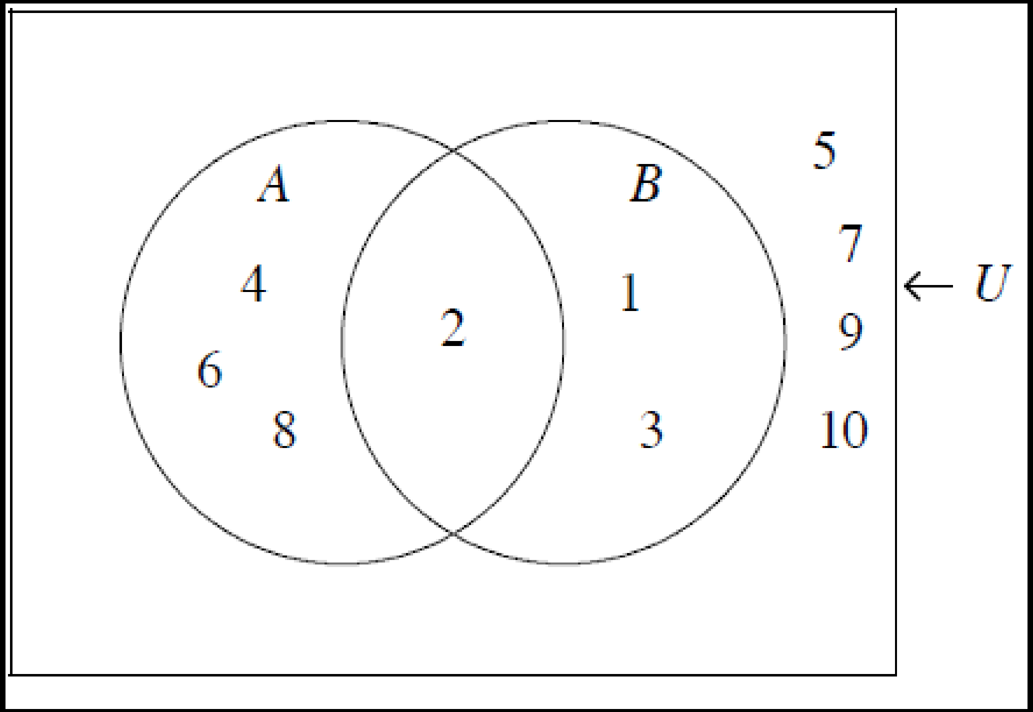
Probabilities: **Basic concepts in probability by Sue Gordon**
Set Notation
Finite Equiprobable Spaces
Complementary Events
Mutually Exclusive Events
Conditional Probability
Independence
Statistics: **Elementary Statistics by Ron Larson**
Basic Definitions
Measures of Central Tendency
Measures of Variations

Set Notation

- A **set** is a collection of objects.
- The elements of a set are specified using {}.
  - $A = \{2, 4\ , 6, 8\}$
  - $A = \{even\ numbers\ less\ than\ 9\}$
- Given two sets A and B:
  - The **union** of A and B ($A \ \cup \ B$) is the set of elements which belong to A or to B (or both).
  - The **intersection** of A and B ($A \ \cap \ B$) is the set of elements which belong to both A and B.
  - The **complement** of A ($\bar{A}$), is the set of all elements which do not belong to A.

Set Notation

- The **empty** set, written ∅ or {}, means the set with no elements in it.
- A set C is a **subset** of A if all the elements in C are also in A.
- For example, let
  - U = {all positive numbers ≤ 10}
  - A = {2, 4, 6, 8}
  - B = {1, 2, 3}
  - C = {6, 8}



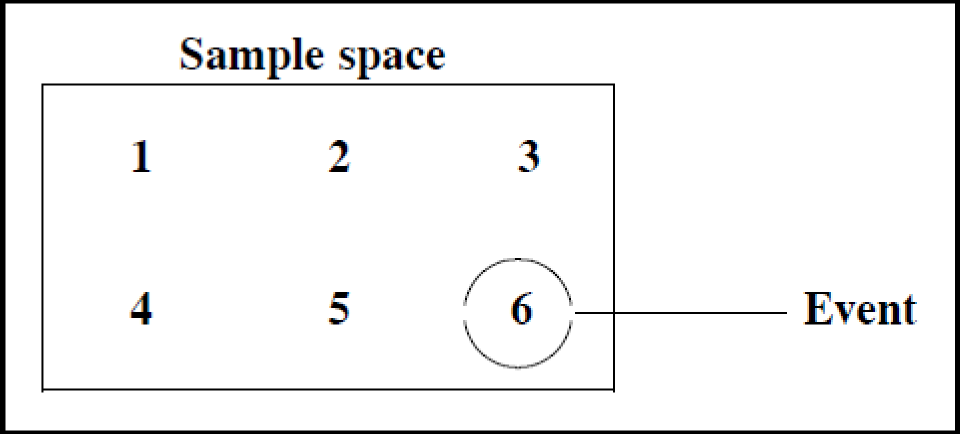
Content

Finite Equiprobable Spaces

- **Finite Equiprobable Spaces** refer to cases where there are a finite number of equally likely outcomes.
  - If a coin is tossed 100 times, it's 50% be a head and 50% be a tail.
    - Probability of heads is 1/2
    - Probability of tails is 1/2
  - If a die is rolled 600 times, it's probable that each value (1,2,3,4,5,6) happen 100 times.
    - Probability of each value is 1/6

Finite Equiprobable Spaces

- **Sample Space** is the set of all possible outcomes of an experiment.
- **Event** is a subset of the sample space.
- Example: rolling a die
    - The sample space is the set {1,2,3,4,5,6}
    - The event of getting the value ‘6’ is the subset {6}



Finite Equiprobable Spaces

The probability of an event A occurring is

$$P(A) = \frac{number\ of\ elements\ in\ A}{total\ number\ of\ elements\ in\ the\ sample\ space}$$

Finite Equiprobable Spaces

- **Example**

A library has 20 programming books, 30 medical books, and 10 engineering books. What is the probability of choosing a programming book and an engineering book?

- **Solution**

P(programming book) = 20/60 = 1/3

P(engineering book) = 10/60 = 1/6

Finite Equiprobable Spaces

- **Example**

Two coins are tossed. Let A be the event 'two heads are obtained', and, B be the event 'one head and one tail is obtained'. Find P(A), P(B).

- **Solution**

The sample space= $\{HH, HT, TH, TT\}$. $A = \{HH\}$. $B = \{HT, TH\}$.

Since there are 4 outcomes in the sample space.

$$P(A) = 1/4$$
$$P(B) = 2/4 = 1/2$$

Content

Probabilities: **Basic concepts in probability by Sue Gordon**
Set Notation
Finite Equiprobable Spaces
Complementary Events
Mutually Exclusive Events
Conditional Probability
Independence
Statistics: **Elementary Statistics by Ron Larson**
Basic Definitions
Measures of Central Tendency
Measures of Variations

Complementary Events

- If an event is a **certainty**, then its probability is one.
- **Example**

If a normal die is rolled, what is the probability that the number showing is less than 7?

- **Solution**

Sample space = {1,2,3,4,5,6}

Event = {1,2,3,4,5,6}

Hence the probability (number is less than 7) = 6/6 = 1.

Complementary Events

- If an event is **impossible**, then its probability is zero.
- **Example**

Find the probability of throwing an 8 on a normal die.

- **Solution**

Sample space = {1,2,3,4,5,6}

Event = {}, i.e. the empty set.

Hence the probability of throwing an 8 is 0/6 = 0.

Complementary Events

- Two events are **complementary** if they cannot occur at the same time and they make up the whole sample space.
- **Example**

When a coin is tossed, the sample space is {H, T} and the events H = ‘obtain a head’ and T = ‘obtain a tail’ are complementary.

If we calculate the probabilities we find that

$P(H) = 1/2$, $P(T) = 1/2$ and $P(H) + P(T) = 1$.

Complementary Events

- **Example**

A die is rolled. Let A be the event ‘a number less than 3 is obtained’ and let B be the event ‘a number of 3 or more is obtained’.

Then $P(A) = 2/6$, and $P(B) = 4/6$.

So that $P(A) + P(B) = 1$.

Complementary Events

- If two events are **complementary**, then **their probabilities add up to 1.**
- **Example**

A marble is drawn at random from a bag containing 3 red, 2 blue, 5 green and 1 yellow marble. What is the probability that it is not green?

- Solution

the probability that the marble is green: $P(G) = 5/11$.

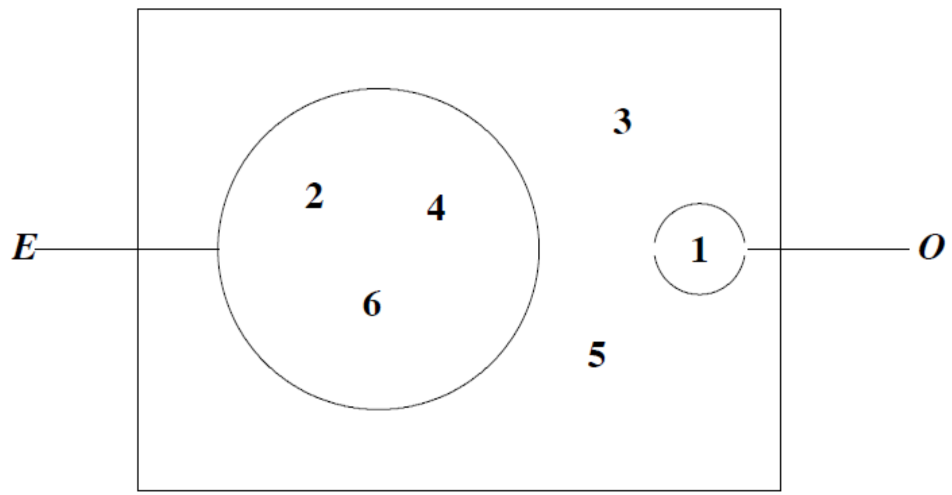
the probability that it is not green, $P(\bar{G}) = 1 - 5/11 = 6/11$.

Content

Probabilities: **Basic concepts in probability by Sue Gordon**
Set Notation
Finite Equiprobable Spaces
Complementary Events
Mutually Exclusive Events
Conditional Probability
Independence
Statistics: **Elementary Statistics by Ron Larson**
Basic Definitions
Measures of Central Tendency
Measures of Variations

Mutually Exclusive Events

- Two events are **incompatible**, **disjoint** or **mutually exclusive** when two events **cannot** occur at the same time.
    - o we can never have a head and a tail side of a coin face up at the same time.
- **Example**: suppose a die is tossed. Then the events E = ‘obtaining an even number’ and O = ‘obtaining a one’ are mutually exclusive.



On our diagram the events do not intersect.

Mutually Exclusive Events

- Exercise: **What is the flaw in the following argument?**

'Seventy percent of first year science students study mathematics. Thirty percent of first year science students study chemistry. If a first-year science student is selected at random, the probability that the student is taking maths is $\frac{70}{100}$, the probability that the student is taking chemistry is $\frac{30}{100}$, hence the probability that the student is taking maths or chemistry is

$\frac{70}{100} + \frac{30}{100} = 1$ (i.e., a certainty).'

Mutually Exclusive Events

- Solution:

The two events are not mutually exclusive; therefore, we cannot add the probabilities.

That is, to count all students doing maths and/or chemistry, we need to count all the maths students, all the chemistry students, and subtract from this the number of students who were counted twice because they were in both classes.

$$P(M \cup C) = P(M) + P(C) - P(M \cap C)$$

Mutually Exclusive Events

- To Summarise:

For any two events A and B, then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

If A and B are mutually exclusive, then

$$P(A \cap B) = 0,$$

A and B cannot happen together,

so that $P(A \cup B) = P(A) + P(B)$.

Content

Probabilities: **Basic concepts in probability by Sue Gordon**
Set Notation
Finite Equiprobable Spaces
Complementary Events
Mutually Exclusive Events
Conditional Probability
Independence
Statistics: **Elementary Statistics by Ron Larson**
Basic Definitions
Measures of Central Tendency
Measures of Variations

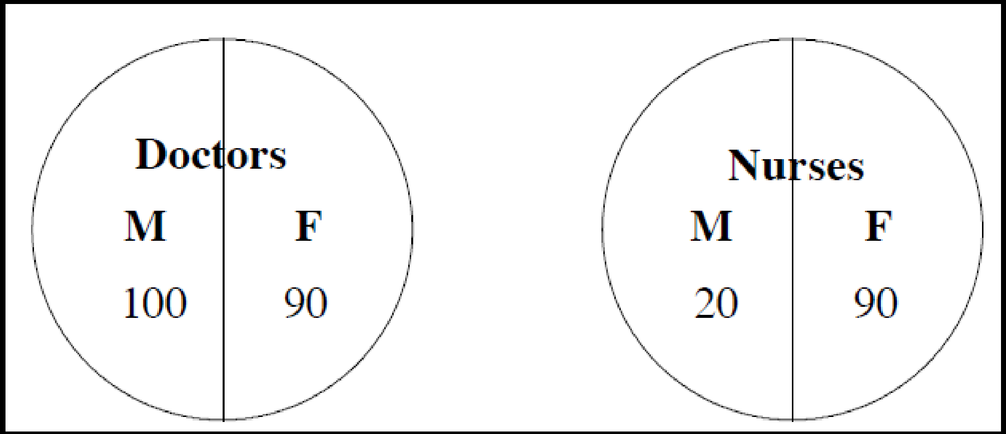
Conditional Probability

- A lecture on a topic of public health is held and 300 people attend. They are classified in the following way:
- If one person is selected at random, find the following probabilities:
  - P(a doctor is chosen);
  - P(a female is chosen);
  - P(a nurse is chosen);
  - P(a male is chosen);
  - P(a female nurse is chosen);
  - P(a male doctor is chosen).

Gender	Doctors	Nurses	Total
Female	90	90	180
Male	100	20	120
Total	190	110	300

Conditional Probability

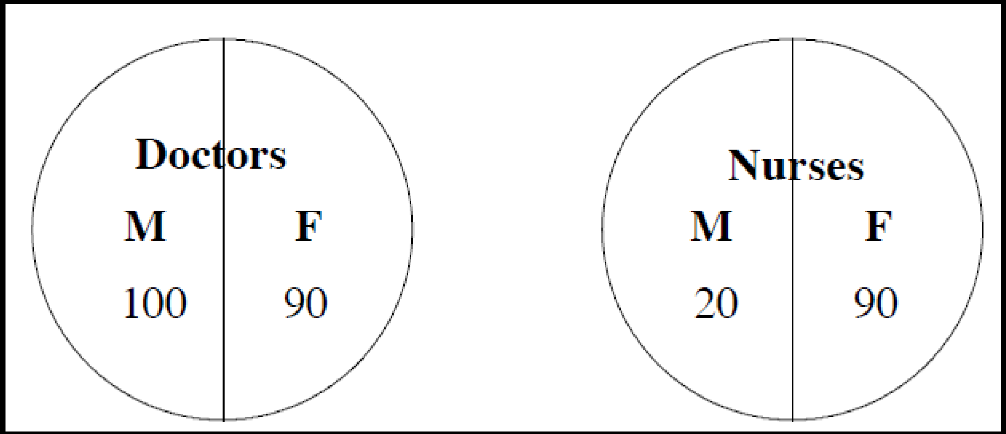
- A lecture on a topic of public health is held and 300 people attend. They are classified in the following way:
- If one person is selected at random, find the following probabilities:
  - P(a doctor is chosen); 190/300
  - P(a female is chosen); 180/300
  - P(a nurse is chosen); 110/300
  - P(a male is chosen); 120/300
  - P(a female nurse is chosen); 90/300
  - P(a male doctor is chosen); 100/300



Conditional Probability

- Now suppose you are given the information that a female is chosen and you wish to find the probability that she is a nurse.
  - P(nurse | female): "The probability that a chosen is a nurse, given that she is female"
- $P(nurse|female) = \frac{P(nurse \cap female)}{P(female)} =$

$$\frac{90}{300} / \frac{180}{300} = \frac{90}{180} = \frac{1}{2}$$



Conditional Probability

- **Definition:** The conditional probability of A given B is defined by

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ provided that } P(B) \neq 0.$$

$$P(A \cap B) = P(A|B)\,.\,P(B) = P(B\,|A)\,.\,P(A),$$

Note that: $P(A|\,B)$ is not the same as $P(B|A)$.

Conditional Probability

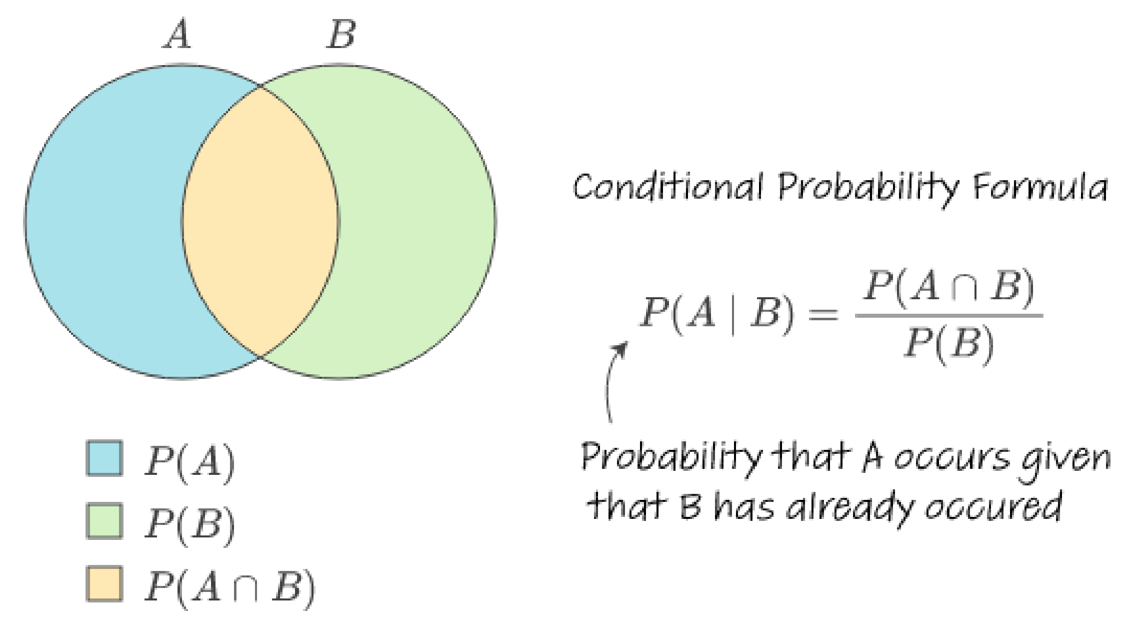
- **Exercise:** find
  - P(female | nurse),
  - P(doctor | male),
  - P(male | doctor).

Conditional Probability

- **Exercise:** find
  - P(female | nurse), $= \frac{90}{110} = \frac{9}{11}$, since there are 110 nurses and of these 90 are female.
  - P(doctor | male), $= \frac{100}{120} = \frac{5}{6}$, since there are 120 males of whom 100 are doctors.
  - P(male | doctor). $= \frac{100}{190} = \frac{10}{19}$, since there are 190 doctors and of these 100 are male.

Conditional Probability

- **To summarize:**



https://stats.stackexchange.com/questions/587109/why-is-the-denominator-in-a-conditional-probability-the-probability-of-the-condi

  - To calculate P(A|B), choose the whole set of B, then from the set of B, choose A.
  - To calculate P(B|A), choose the whole set of A, then from the set of A, choose B.

Content

Probabilities: **Basic concepts in probability by Sue Gordon**
Set Notation
Finite Equiprobable Spaces
Complementary Events
Mutually Exclusive Events
Conditional Probability
Independence
Statistics: **Elementary Statistics by Ron Larson**
Basic Definitions
Measures of Central Tendency
Measures of Variations

Independence

- **Definition:**

Two events A and B are said to be **independent** if and only if $P(A|B) = P(A)$, that is, when the conditional probability of A given B is the same as the probability of A.

  - The occurrence of A **does not depend on** the occurrence of B

- Example: tossing a coin, the probability to get heads does not depend on the probability of getting a tails.

Independence

- When two events are independent, the chance that both will happen is found by multiplying their individual chances.

  A and B are independent events if and only if $P(A \cap B) = P(A).P(B)$.

- **Example**
  What is the probability of obtaining '6' and '6' on two successive rolls of a die?

- **Solution**
  P(obtaining 6 on a roll of a die) = 1/6.
  The two rolls are independent
  So P(6 and 6) = 1/6 . 1/6 = 1/36.

Independence

- **Example:** A box contains three white cards and three black cards numbered:

White			Black		
1	2	2	1	1	2

One card is picked out of the box at random. If A is the event 'the card is black' and B is the event 'the card is marked 2', are A and B independent?

- **Solution**

$P(A) = 1/2.$ $\qquad P(B) = 1/2.$

$P(A \cap B) = P(card\ is\ black\ and\ marked\ 2) = 1/6.$

Now $1/6 \neq 1/2 . 1/2,$ so A and B are not independent.

Independence

- **Exercise**

A couple has two children. Let A be the event 'they have one boy and one girl' and B the event 'they have at most one boy'. Are A and B independent?

Independence

- **Exercise**

A couple has two children. Let A be the event 'they have one boy and one girl' and B the event 'they have at most one boy'. Are A and B independent?

- **Solution**

Sample space = {GG, BG, GB, BB}. Note that a 'girl followed by a boy' is not the same event as 'a boy followed by a girl'.

A = {BG, GB}, B = {GG, BG, GB}, A ∩ B = {BG,GB},

P(A ∩ B) = 2/4 = 1/2 P(A).P (B) = 2/4 . 3/4 = 3/8 .

Since $P(A \cap B) \neq P(A).P(B)$, A and B are **not independent**.

Content

Probabilities: Basic concepts in probability by Sue Gordon
Set Notation
Finite Equiprobable Spaces
Complementary Events
Mutually Exclusive Events
Conditional Probability
Independence
Statistics: Elementary Statistics by Ron Larson
Basic Definitions
Measures of Central Tendency
Measures of Variations

Basic Definitions

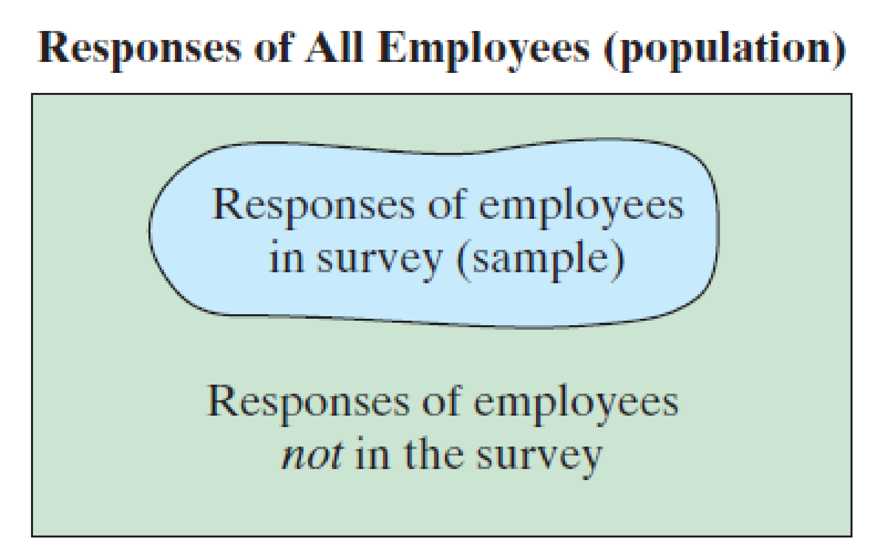
- **Data**
  consist of information coming from observations, counts, measurements, or responses.

- **Statistics**
  is the science of collecting, organizing, analyzing, and interpreting data in order to make decisions.

Basic Definitions

Two types of data sets

- **population**
  is the collection of all outcomes, responses, measurements, or counts that are of interest.
- **sample**
  is a subset, or part, of a population.

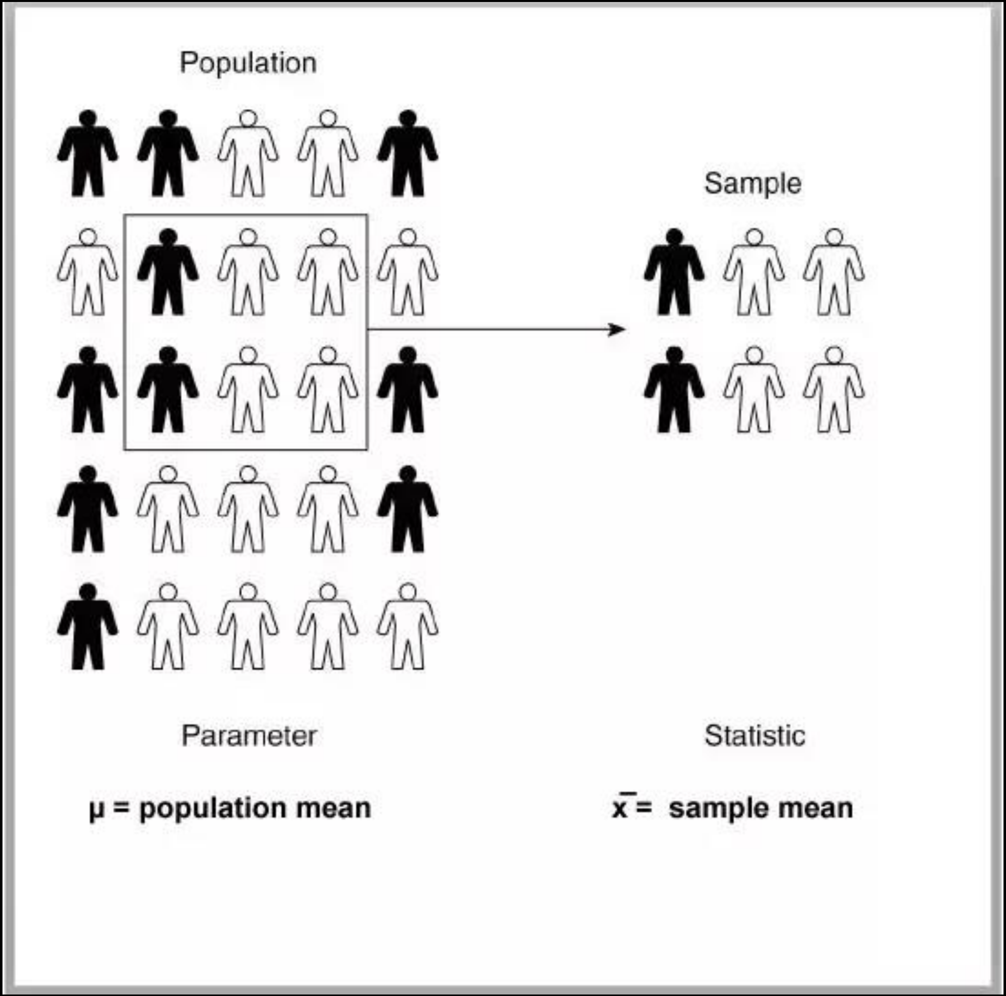


Basic Definitions

- **parameter**
  is a numerical description of a population characteristic.
- **statistic**
  is a numerical description of a sample characteristic.

  - Note that a sample statistic can differ from sample to sample, whereas a population parameter is constant for a population.

Basic Definitions

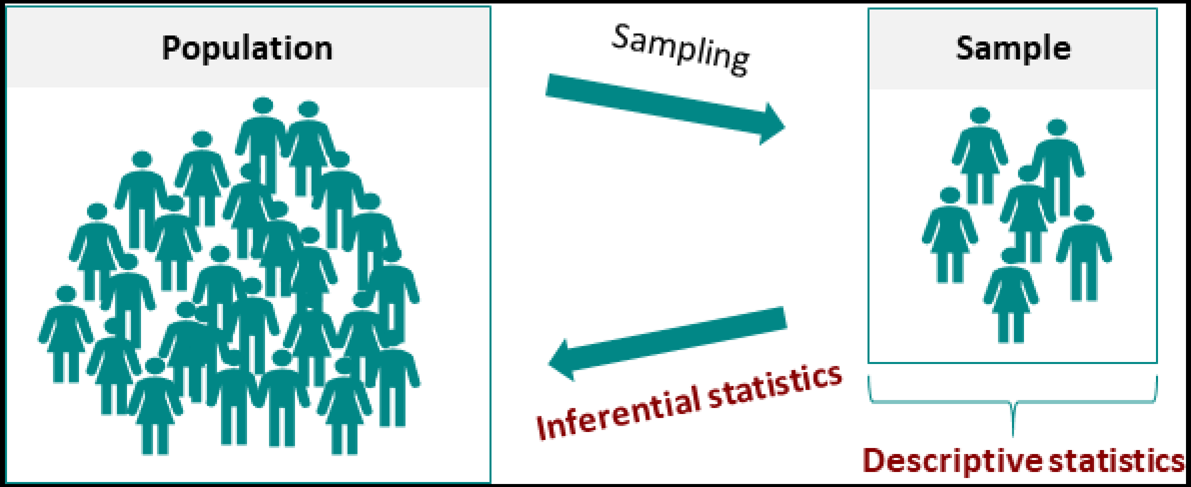


https://www.quora.com/What-are-some-examples-of-population-parameters-and-sample-statistics

Basic Definitions

Two branches of statistics

- **Descriptive statistics**
  involves the organization, summarization, and display of data.
- **Inferential statistics**
  involves using a sample to draw conclusions about a population.
  - A basic tool in the study of inferential statistics is probability.



https://datatab.net/tutorial/descriptive-inferential-statistics

Basic Definitions

Types of data

- **Qualitative data**
  consist of attributes, labels, or nonnumerical entries.
  - Gender, color, martial status.
- **Quantitative data**
  consist of numbers that are measurements or counts.
  - Age, height, weight, price

Basic Definitions

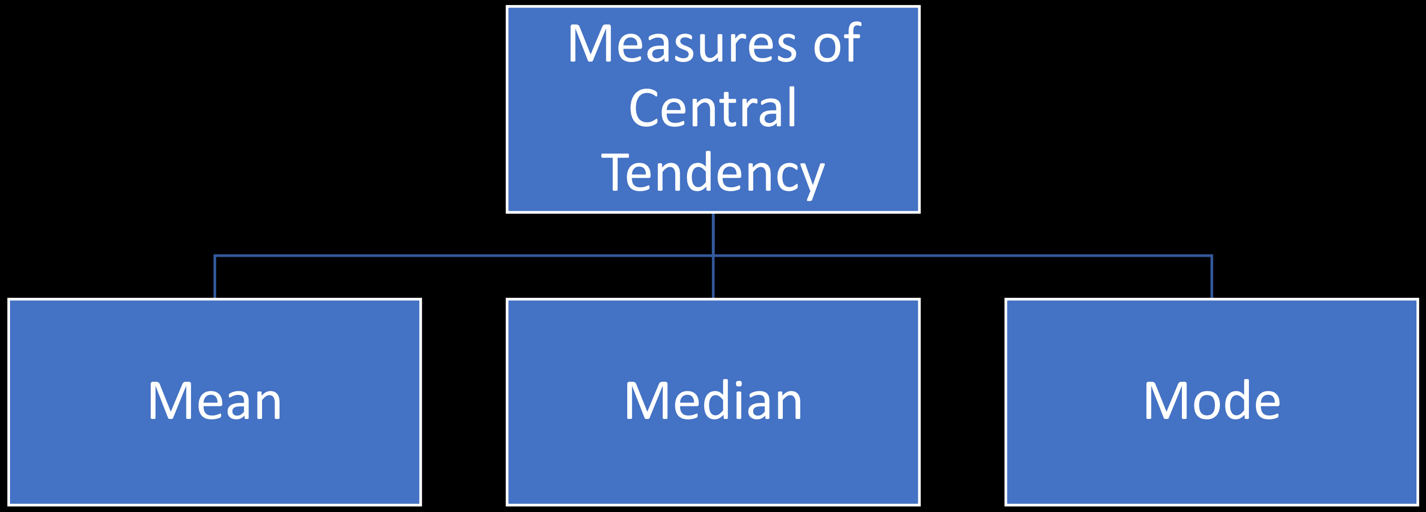
Task: What nominal and ordinal data?

Content

Probabilities: **Basic concepts in probability by Sue Gordon**
Set Notation
Finite Equiprobable Spaces
Complementary Events
Mutually Exclusive Events
Conditional Probability
Independence
Statistics: **Elementary Statistics by Ron Larson**
Basic Definitions
Measures of Central Tendency
Measures of Variations

Measures of Central Tendency

- A measure of central tendency is a value that represents a typical, or central, entry of a data set.



Measures of Central Tendency

- The **mean** of a data set is the sum of the data entries divided by the number of entries.

$$\mu = \frac{\sum(X)}{N}$$

- **Example**: The weights for a sample of adults before starting a weight-loss study are listed. What is the mean weight of the adults?

274 235 223 268 290 285 235

- **Solution**:

$$\frac{274 + 235 + 223 + 268 + 290 + 285 + 235}{7} = 258.6$$

Measures of Central Tendency

- The **median** of a data set is the value that lies in the middle of the data when the data set is ordered.
- **Example**: Find the median of the weights

  274 235 223 268 290 285 235

- **Solution**: first order the data.

  223 235 235 268 274 285 290

The median is the middle value, 268

Measures of Central Tendency

- The **mode** of a data set is the entry that occurs with the greatest frequency.
  - A data set can have one mode, more than one mode, or no mode.
  - When no entry is repeated, the data set has no mode.
  - When two entries occur with the same greatest frequency, each entry is a mode and the data set is called **bimodal**.
- **Example**: Find the mode

274 235 223 268 290 285 235

- **Solution**: first order the data.

223 235 235 268 274 285 290

the mode is 235.

Measures of Central Tendency

- **Exercise**: Find the mean, median, and mode of the ages.

Ages in a class						
20	20	20	20	20	20	21
21	21	21	22	22	22	23
23	23	23	24	24	65	

Measures of Central Tendency

- **Exercise**: Find the mean, median, and mode of the ages.

Ages in a class						
20	20	20	20	20	20	21
21	21	21	22	22	22	23
23	23	23	24	24	65	

- **Solution**:

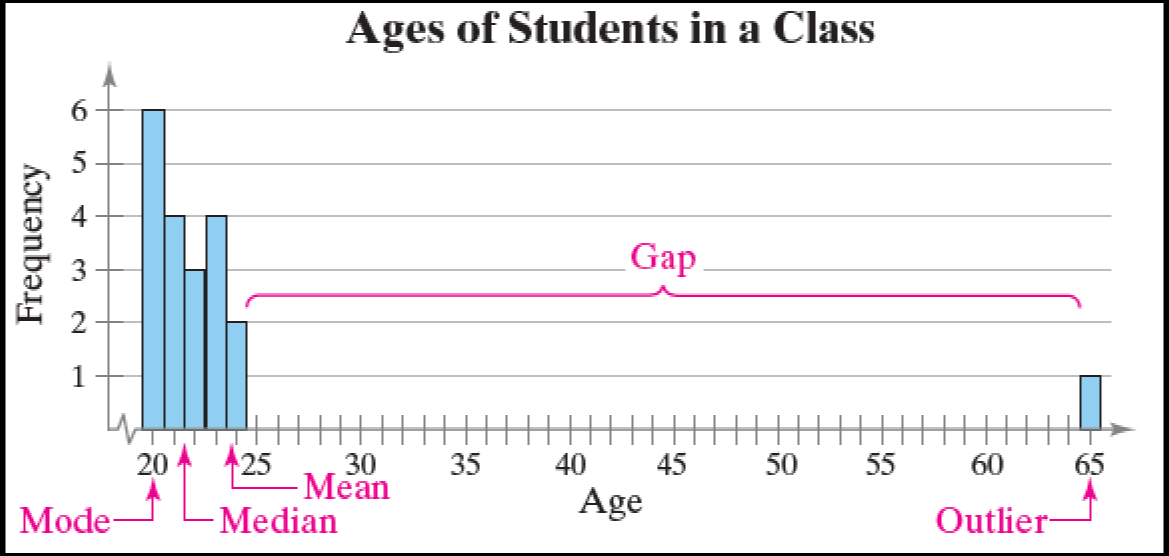
Mean = $\frac{475}{20} = 23.8$ years

Median = $\frac{21+22}{2} = 21.5$ years

Mode: The entry occurring with the greatest frequency is 20 years.

Measures of Central Tendency

- An **outlier** is a data entry that is far removed from the other entries in the data set.



- While some outliers are valid data, other outliers may occur due to data-recording errors.

Measures of Central Tendency

- Task: What is weighted mean? Compute the weighted mean of the following data:

Your grades from last semester are in the table. The grading system assigns points as follows: A = 4, B = 3, C = 2, D = 1, F = 0. Determine your grade point average (weighted mean).

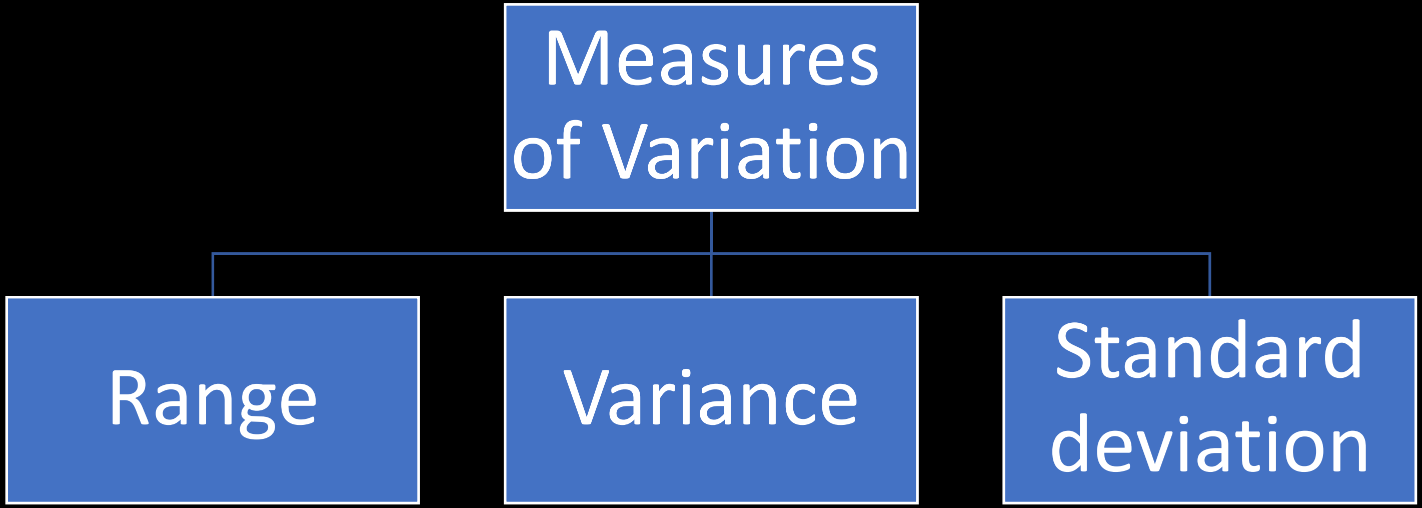
Final Grade	Credit Hours
C	3
C	4
D	1
A	3
C	2
B	3

Content

Probabilities: **Basic concepts in probability by Sue Gordon**
Set Notation
Finite Equiprobable Spaces
Complementary Events
Mutually Exclusive Events
Conditional Probability
Independence
Statistics: **Elementary Statistics by Ron Larson**
Basic Definitions
Measures of Central Tendency
Measures of Variations

Measures of Variation

- Ways to measure the variation (or spread) of a data set.



Measures of Variation

- The **range** of a data set is the difference between the maximum and minimum data entries in the set.

Range = (Maximum data entry) - (Minimum data entry)

- **Example**: Find the range of the starting salaries for Corporation A.

Starting Salaries for Corporation A (in thousands of dollars)

Salary	41	38	39	45	47	41	44	41	37	42

- **Solution**:

Ordering the data helps to find the least and greatest salaries.

37 38 39 41 41 41 42 44 45 47

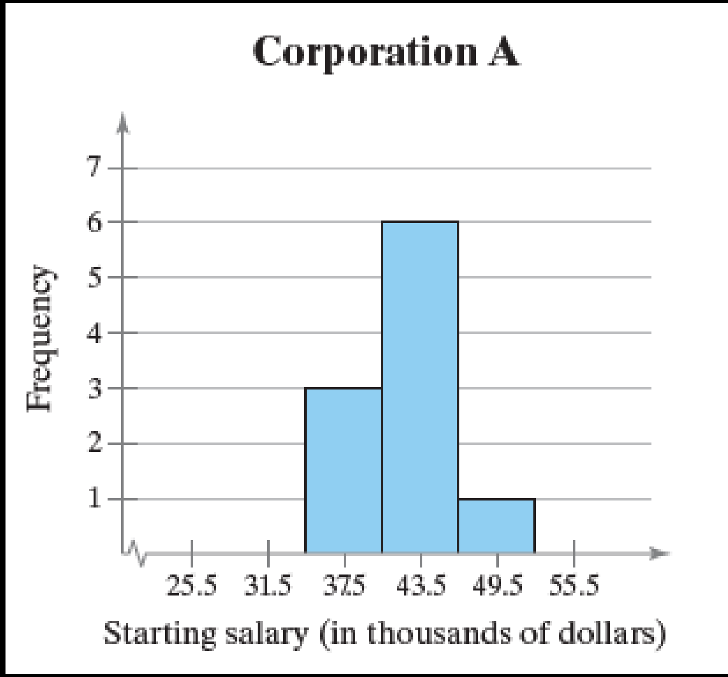
Minimum

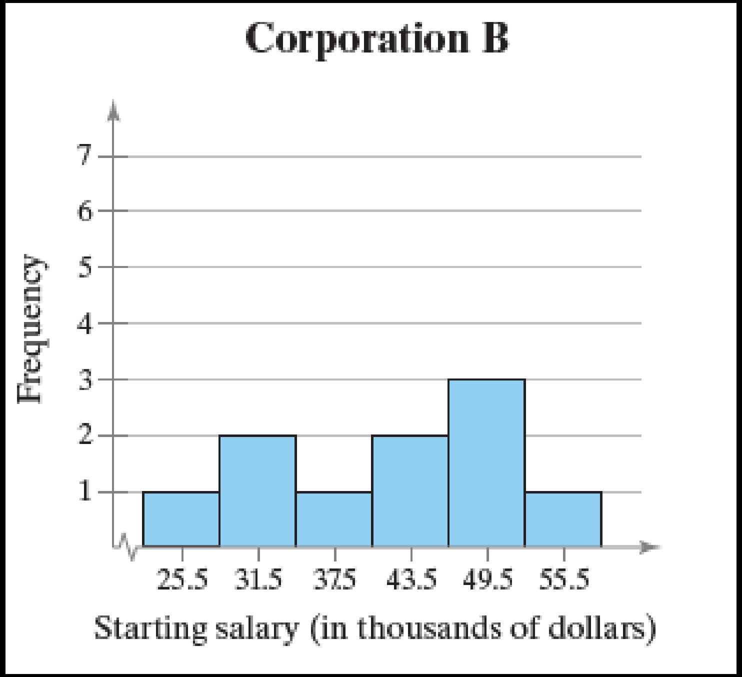
Maximum

$$\begin{aligned}\text{Range} &= (\text{Maximum salary}) - (\text{Minimum salary})\\ &= 47 - 37\\ &= 10\end{aligned}$$

Measures of Variation

- Differences in range in a dataset





Range = 10

Range = 35

Measures of Variation

- The **variance** (or **deviation)** of an entry x in a population data set is the difference between the entry and the mean μ of the data set.

$$Deviation\ of\ x = x - \mu$$

- **Example**: The mean starting salary is

$$\mu = 415/10 = 41.5$$

  - the sum of the deviations is 0.

Salary (in 1000s of dollars) x	Deviation (in 1000s of dollars) $x - \mu$
41	−0.5
38	−3.5
39	−2.5
45	3.5
47	5.5
41	−0.5
44	2.5
41	−0.5
37	−4.5
42	0.5
$\Sigma x = 415$	$\Sigma(x - \mu) = 0$

The sum of the deviations is 0.

Measures of Variation

- **Population Variance**
  is the average of the squares of the deviations.

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

- **Standard deviation**
  is the square root of the population variance.

$$stdv = \sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum(x-\mu)^2}{N}}$$

Measures of Variation

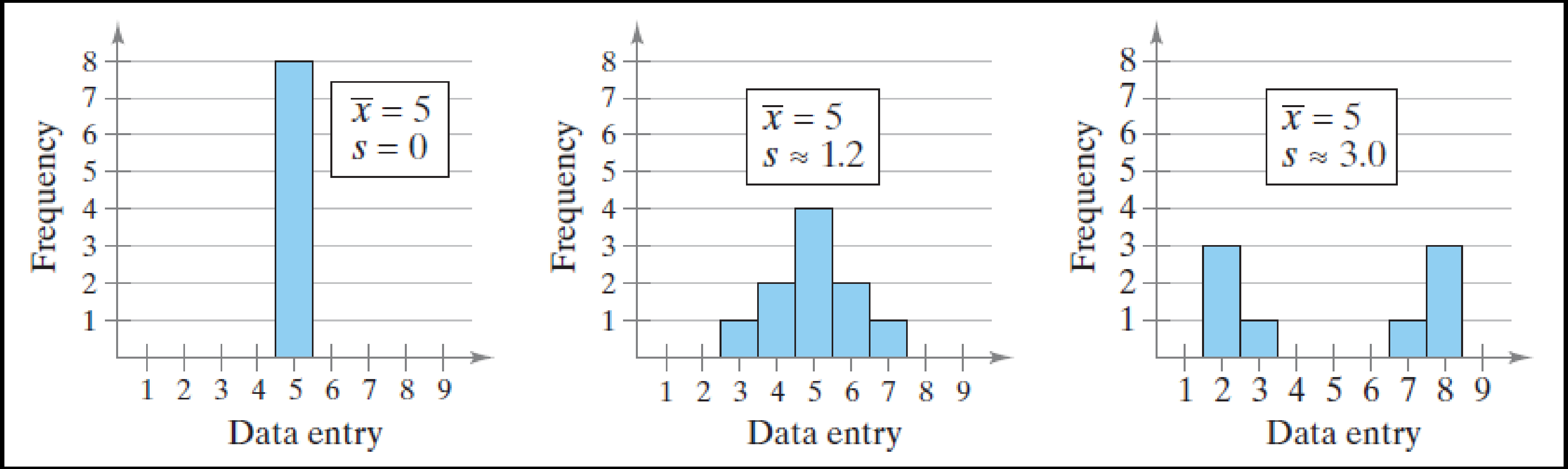
- **Example:** Find the population variance and standard deviation of the starting salaries for Corporation A.

Salary x	Deviation $x - \mu$	Squares $(x - \mu)^2$
41	−0.5	0.25
38	−3.5	12.25
39	−2.5	6.25
45	3.5	12.25
47	5.5	30.25
41	−0.5	0.25
44	2.5	6.25
41	−0.5	0.25
37	−4.5	20.25
42	0.5	0.25
$\Sigma x = 415$		$SS_x = 88.5$

- **Solution:** we have 10 entries and $\sum(x) = 415$
  - $\mu = \frac{415}{10} = 41.5$
  - $\sigma^2 = \frac{88.5}{10} = 8.9$
  - $\sigma = \sqrt{8.9} \approx 3$

Measures of Variation

- Interpreting Standard Deviation:
  - 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.



Measures of Variation

- For data sets with distributions that are symmetric and bell-shaped, the standard deviation has these characteristics.
  - About 68% of the data lie within one standard deviation of the mean.
  - About 95% of the data lie within two standard deviations of the mean.
  - About 99.7% of the data lie within three standard deviations of the mean.