Statistical Methods for Data Science (DSCI 602)

Dr. Aihua Li

Ball State University

Spring, 2021

Outline

- Data science
- Application of probability and statistics
- **Introduction to Probability**
  - ➢Event and its types
  - ➢Algebra of events
  - ➢Conditional probability
  - ➢Rules of probability
- **Introduction to Statistics**
  - ➢Descriptive Statistics
  - ➢Inferential Statistics

Data Science

- Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data.[1][2] Data science is related to data mining, machine learning and big data. (from wiki)
- Data science is a "concept to unify statistics, data analysis and their related methods" in order to "understand and analyze actual phenomena" with data.

Abraham Lincoln:

Give me six hours to chop down a tree and I will spend the first four sharpening the axe.

Probability is the sharpened axe for statistics.

1 Application of probability and statistics

- Ever heard about a weather forecast at the end of a news bulletin on TV or read about the weather conditions of your city/country for the next few days in any newspaper? They specifically use the term “probability.”

1 Application of probability and statistics

- The field of probability and statistics (which, for convenience, I will refer to simply as statistics“ below) impacts many aspects of our daily lives, business, medicine, the law, government and soon.
- **Statistics include numerical facts and figures**. For instance:
  - The largest earthquake measured 9.2 on the Richter scale.
  - By the year 2020, there will be 15 people aged 65 and over for every new baby born.

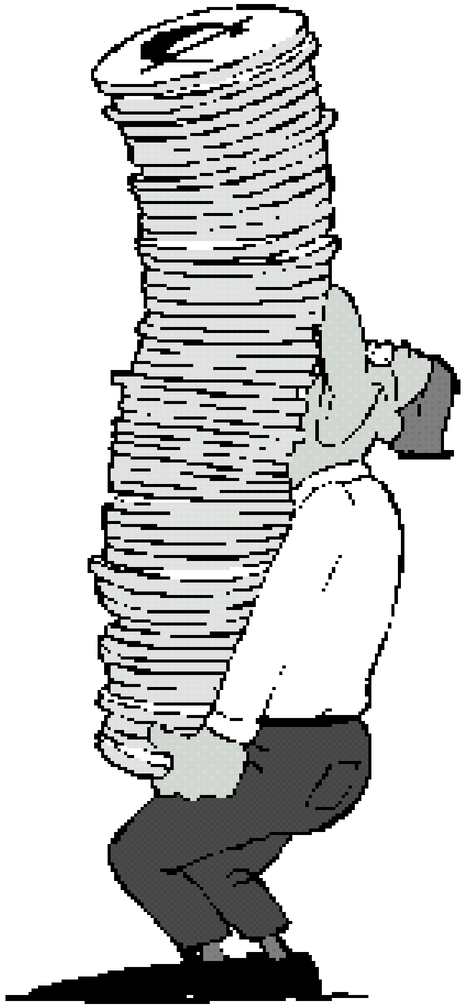
- **But, statistics are *not only facts and figures***; they are something more than that. In the broadest sense, “statistics” refers to a range of techniques and procedures for analyzing, interpreting, displaying, and making decisions based on data.
- The study of statistics involves math and relies upon calculations of numbers. But it also relies heavily on how the numbers are chosen and how the statistics are interpreted.

Symmetrical Outcomes

Frequentist Approach

- If we flipped millions of coins, we'd expect half of them to come up heads.

- So, we say that the probability of getting a head is one half.



Weather Forecasts

- Consider the forecast for August 1 in Seattle.
- If it rained 62,000 of the last 100,000 days, you might say the probability of rain is 62%.

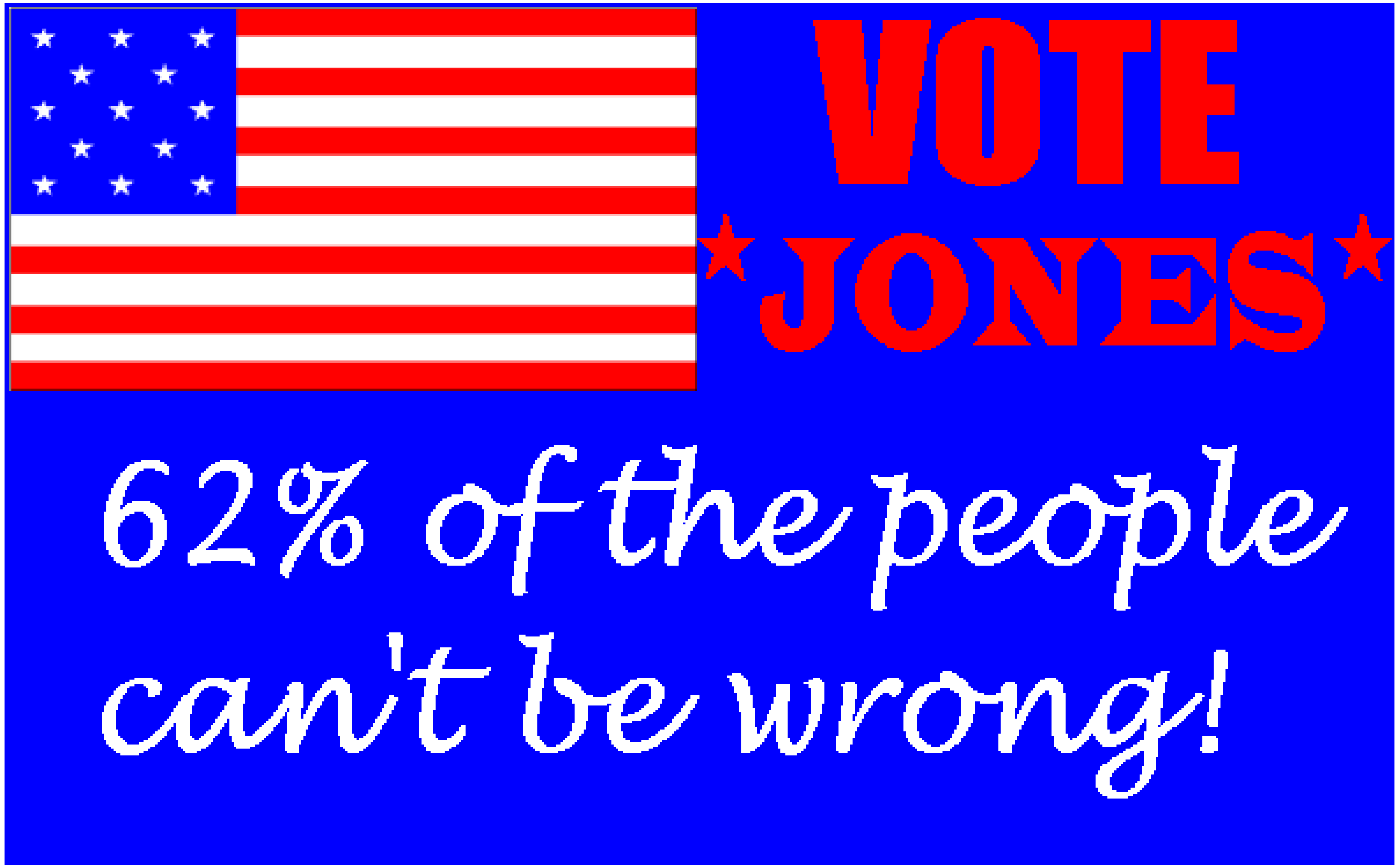
Complications

- It rains less often of August 1.

- Can take humidity, wind direction, and the weather on July 31 into account.

- Sample of exactly similar prior cases is zero.

Subjective Probability



Different Probabilities

- You say there is a 1/6 probability of getting a six when rolling a die.
- Your friend says it is 2/3.
- You are still right, even if a 6 is rolled.

More Weather

A “10% chance of rain”
means that it will rain on 10% of the
days on which rain is forecast
with this probability.

2 What is Probability?

Terms related to Probability:

- Random Experiment: A random experiment is the one in which all the possible results are known in advance but none of them can be predicted with certainty.
- Outcome: The result of a random experiment is called an outcome.
- Sample Space: The set of all the possible outcomes of a random experiment is called a Sample Space.
- Event: Any subset of **the sample space** is called an Event.

Probability

- The probability of any event is defined as the chance of occurrence of the events to the total possible outcomes.

Formula of Probability

The most basic and general formula to calculate probability is

$$\text{probability} = \frac{\text{Number of favorable outcomes}}{\text{Number of possible equally-likely outcomes}}$$

$$\text{probability} = \frac{\text{Number of favorable outcomes}}{\text{Number of possible equally-likely outcomes}}$$

- Consider rolling a six-sided dice. The sample space has six possible outcomes.

The **sample space** has six possible outcomes, *U* = {side facing up is 1, side facing up is 2, …, side facing up is 6}. ***P*[*U*] = 1.**

Define **event *A*** = {side facing up is number 3}, then ***P*[*A*] = 1/6** or 1 out of 6 possible and equally likely outcomes.

3 Event and its types

Based on random experiments, events are of following types.

- Impossible Events
- Sure Events

- Simple Event
- Compound Event
- Complimentary Event (A’ or A^c or $\bar{A}$)

Event A or B (union, $A \cup B$)

Event A and B (intersection, $A \cap B$)

Event A but not B ($A \cap B'$)

Exhaustive Events (Sample Space)

Favorable Events

Mutually Exclusive Events

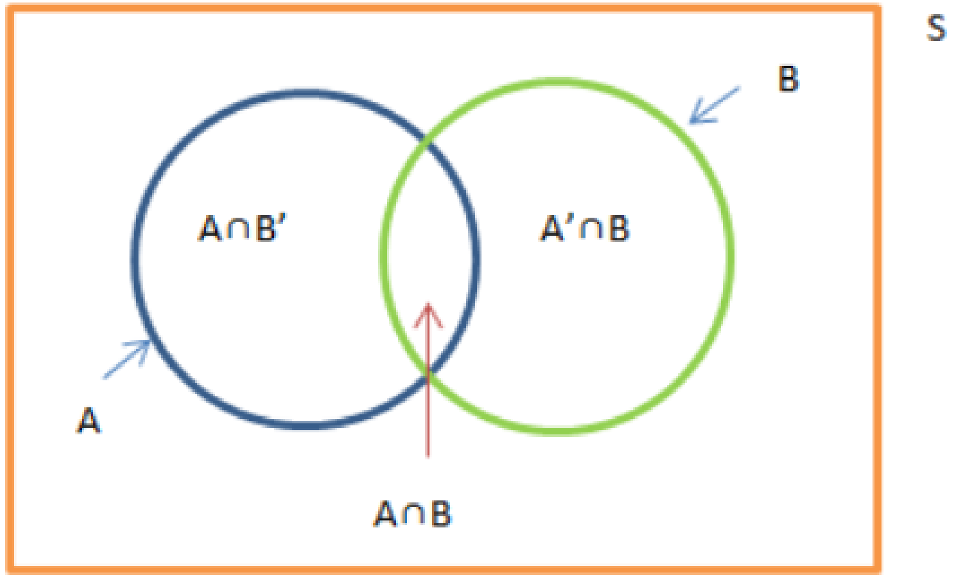
Equally Likely Events

Independent Events

Dependent Events:

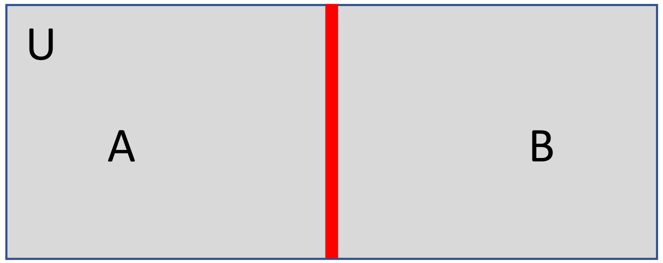
Events Bases on Venn Diagram

A Venn diagram showing various algebra of events.



Single Event

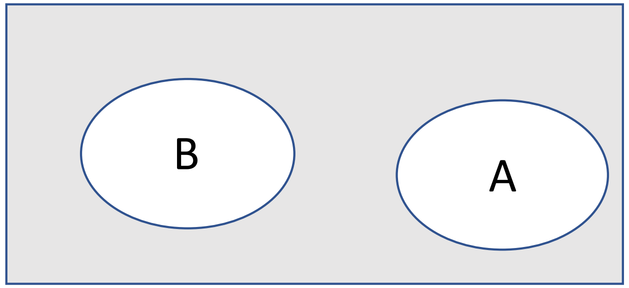
- Coin: two sides
- All possible outcomes: head (A) and tail (B)
- Event: tossing a fair coin: P(A)=1/2, P(B)=1/2, P(A)+P(B)=1



- Dice: six sides
- All possible outcomes:1,2,3,4,5,6
- Event: tossing a dice: P(A1)+ P(A2)+ P(A3)+ P(A4)+ P(A5)+ P(A6)=1

4 Algebra Of Events

Disjoint

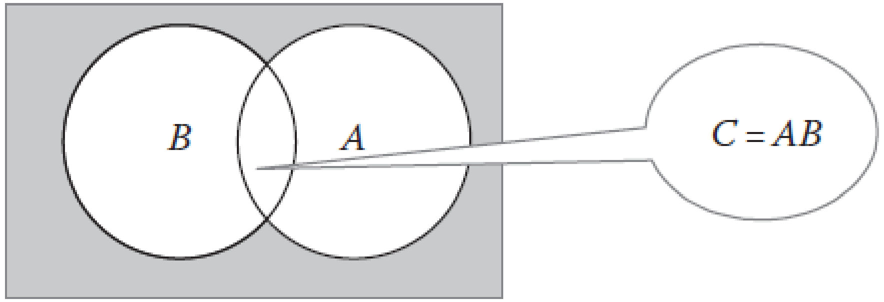


- **Disjoint**: Having no members in common;
- **Joint:** Having common outcomes;

Unions and Intersections

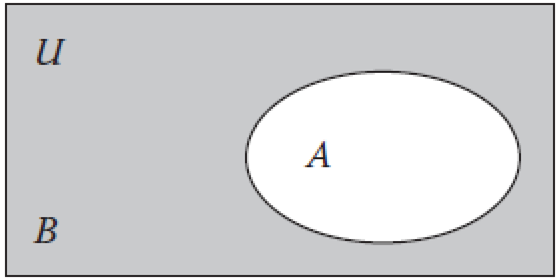
- The word "and" in mathematics means the same thing in mathematics as the **intersection**, which uses the following symbol: ∩
- The word "or" means the same thing in mathematics as the **union**, which uses the following symbol: ∪.

Joint



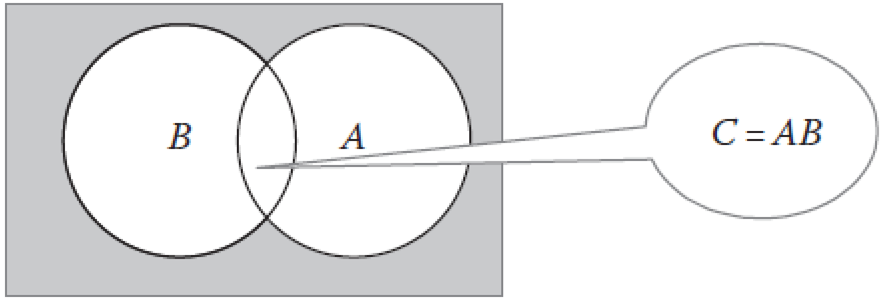
4 Algebra Of Events

1: A and B are **mutually exclusive**



P(A)+P(B)=1 P(A or B)=P(A ∪ B)=P(A)+P(B)

2: A and B are inclusive events

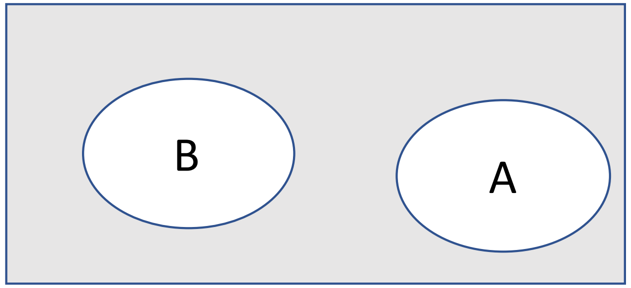


P(C)= P(A and B)=P(A ∩ B)>0

Event C is the Union of A and B : Inclusive events are events that can happen at the same time. To find the probability of an inclusive event we first add the probabilities of the individual events and then subtract the probability of the two events happening at the same time.

P(A or B)=P(A ∪ B)=P(A)+P(B)−P(A and B)

Union (or)



P(C)= P(A and B)=P(A ∩ B)=0

Intersection (and)

Examples:

What is the probability of drawing a black card or a ten in a deck of cards?

There are 4 tens in a deck of cards P(10) = 4/52

There are 26 black cards P(black) = 26/52

There are 2 black tens P(black and 10) = 2/52

P(black or ten)=4/52+26/52−2/52=30/52−2/52=28/52=7/13

Examples:

$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) - P(A \text{ and } B)$

What is the probability of drawing a black card or a ten in a deck of cards?

There are 4 tens in a deck of cards P(10) = 4/52

There are 26 black cards P(black) = 26/52

There are 2 black tens P(black and 10) = 2/52

$P(\text{black or ten}) = 4/52 + 26/52 - 2/52 = 30/52 - 2/52 = 28/52 = 7/13$

Two Events

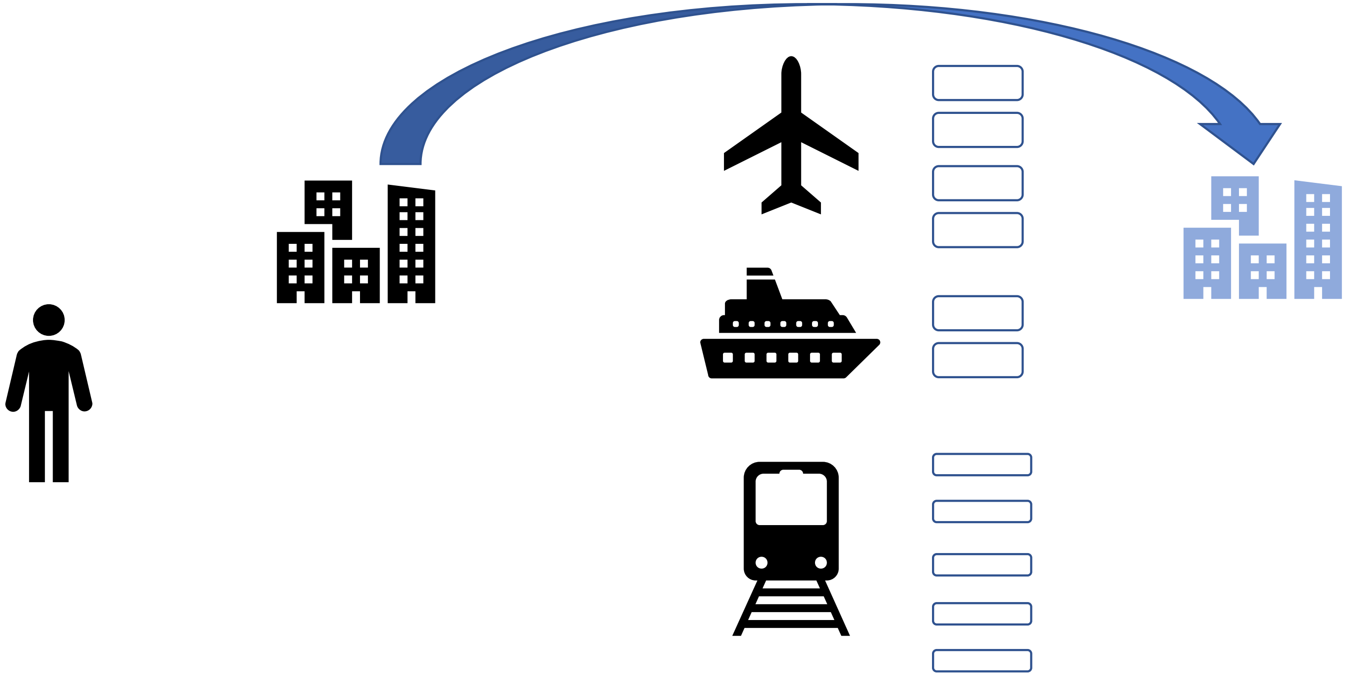
Independent events: Two events are independent when the outcome of the first event does not influence the outcome of the second event.

$$P(A \text{ and } B)=P(A)*P(B)$$

Dependent events: Two events are dependent when the outcome of the first event influences the outcome of the second event.

$$P(A \text{ and } B)=P(A)*P(B \text{ after } A)$$

- **Multiplication Theorem**
- : e.g., different steps to solve a problem, each step in different same space
- **Addition Theorem (Algebra)**
- : e.g., different ways to solve a problem, in the same sample space, total is 1.

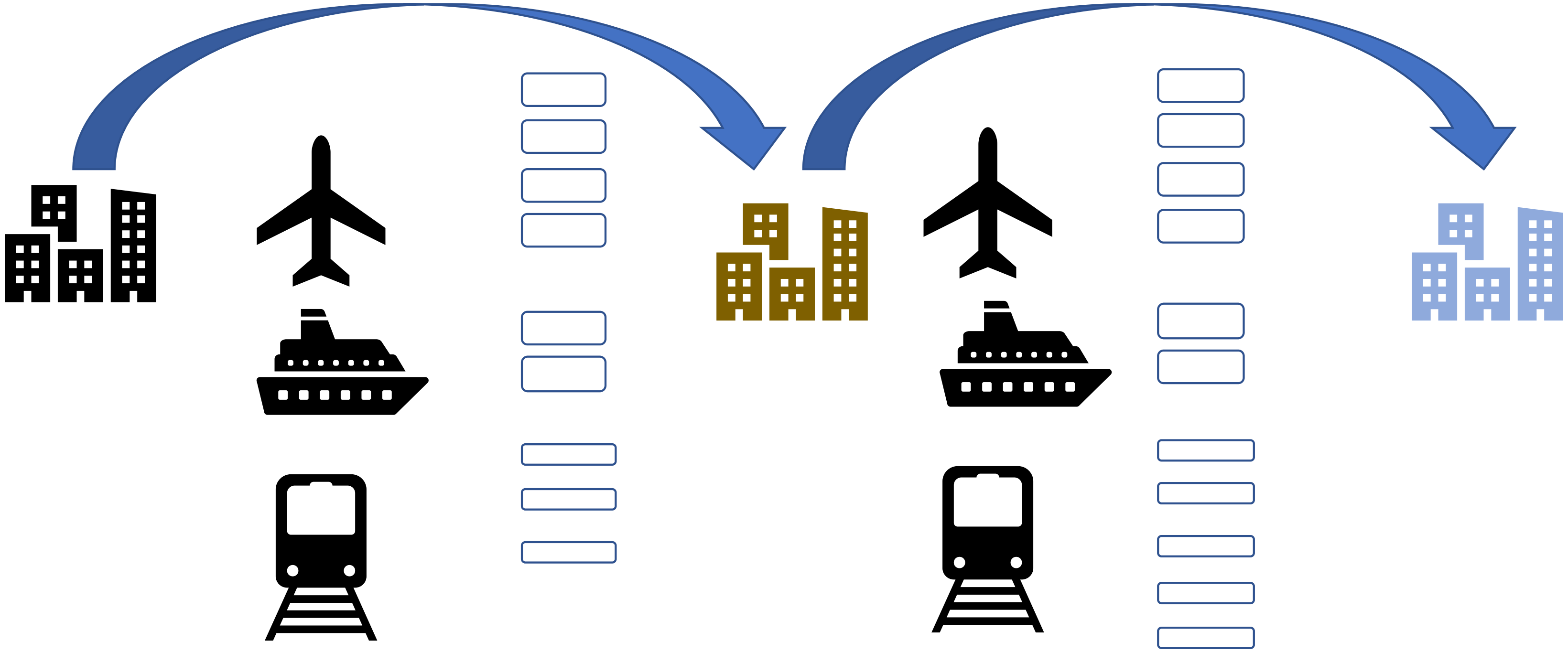


4+2+5=11

P(Flight)=4/11

P(Ship)=2/11

P(Train)=5/11



9*11=99

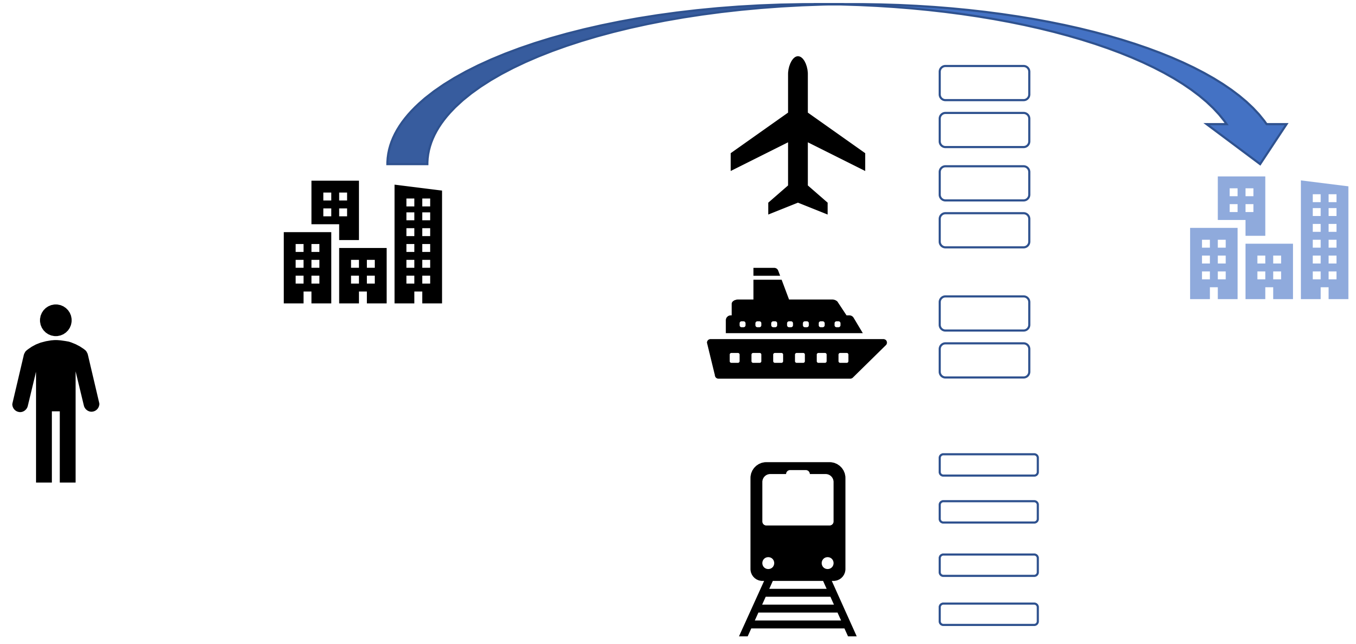
P(Flight and Flight)=16/99

P(Flight then Ship)=8/99

P(Flight then Train)=20/99

P(Ship then Flight)=8/99

...

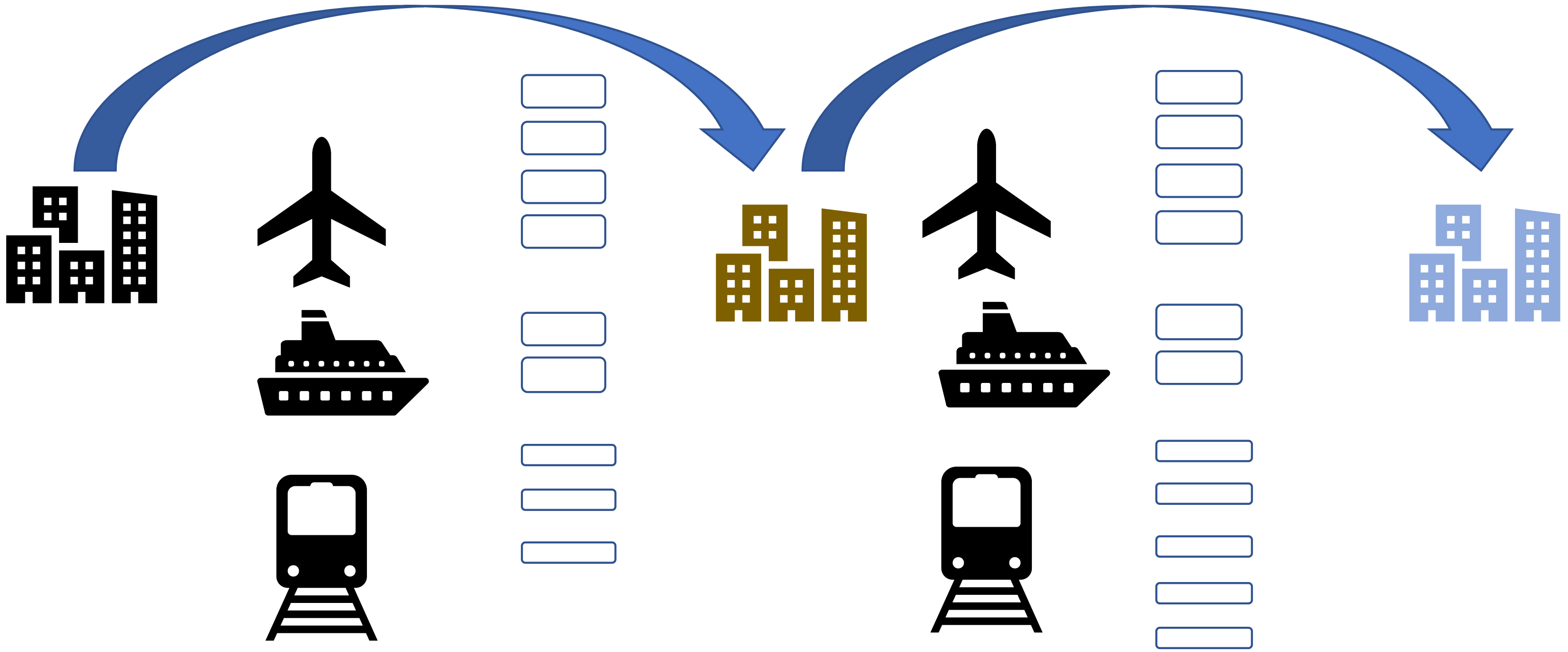


4+2+5=11

P(Flight)=4/11

P(Ship)=2/11

P(Train)=5/11



9*11=99

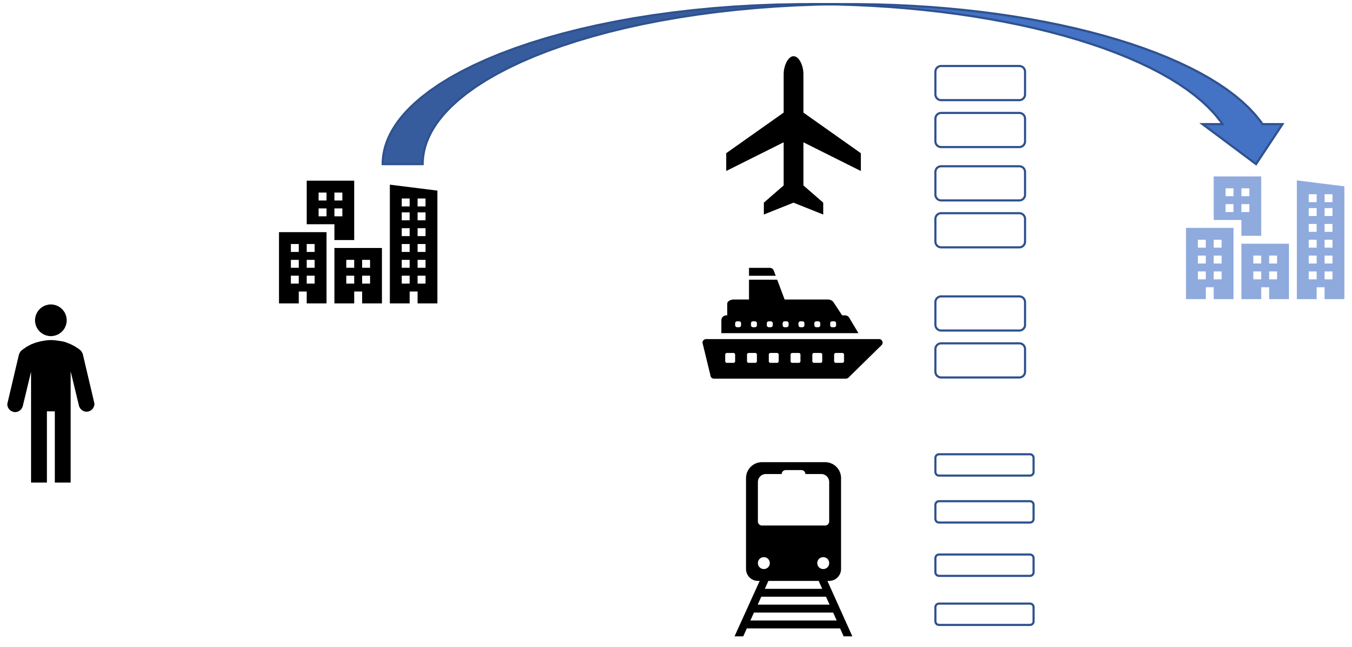
P(Flight and Flight)=16/99

P(Flight then Ship)=8/99

P(Flight then Train)=20/99

P(Ship then Flight)=8/99

...

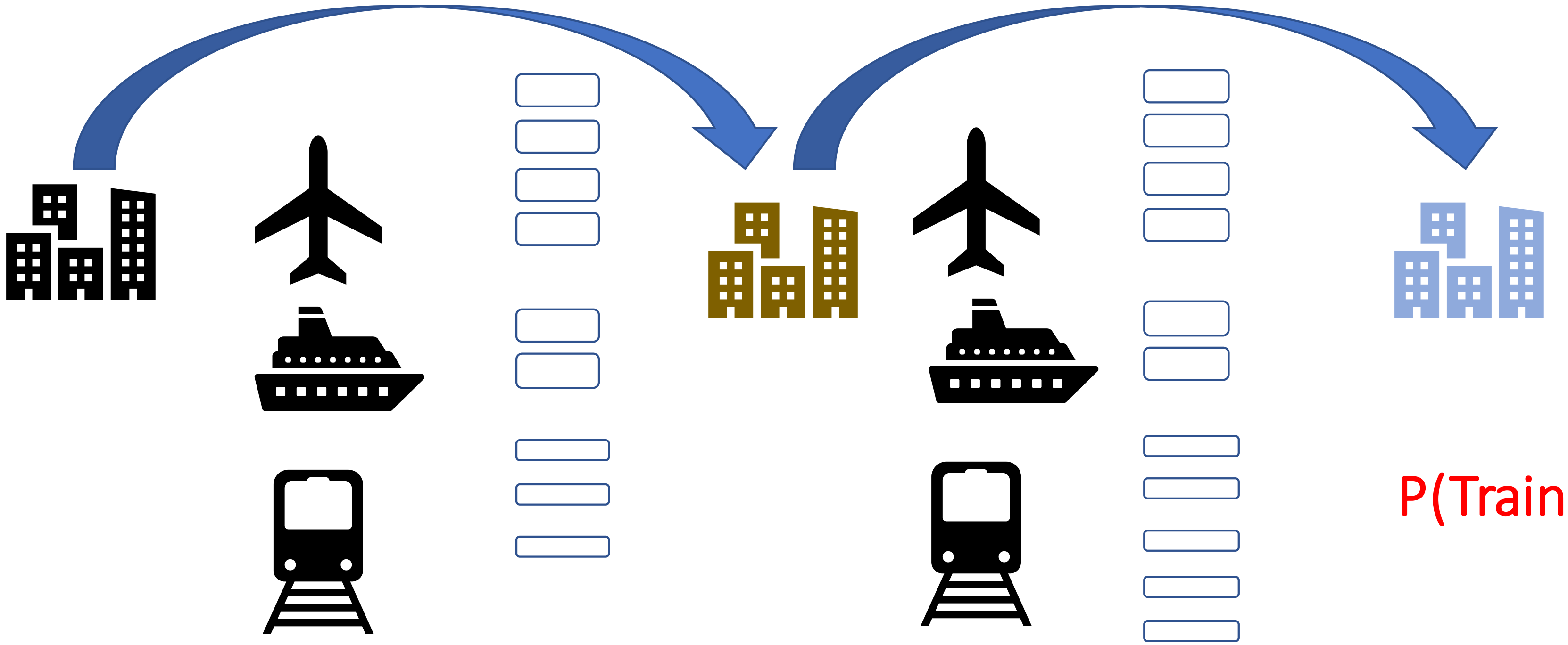


4+2+5=11

P(Flight)=4/11

P(Ship)=2/11

P(Train)=5/11



9*11=99

P(Flight and Flight)=16/99

P(Flight then Ship)=8/99

P(Flight then Train)=20/99

P(Ship then Flight)=8/99

...

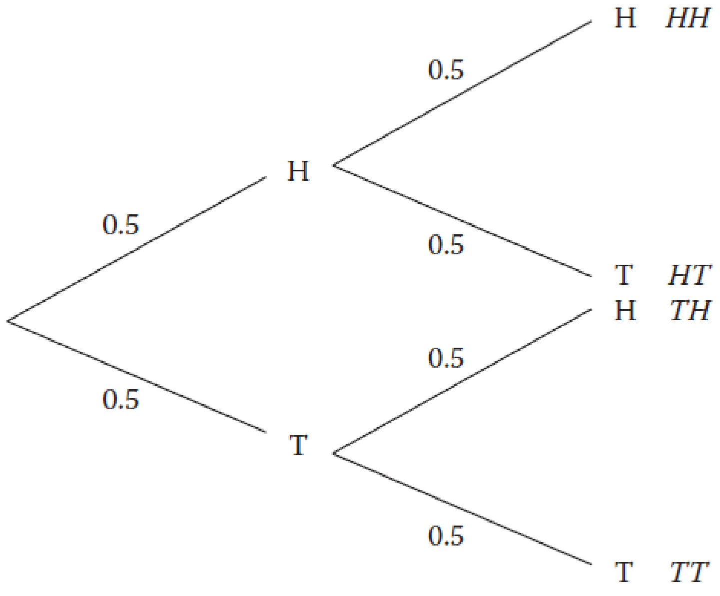
P(Train then Ship)=(3/9)*(2/11)= 6/9

Two or more independent Events

Toss two coin:

Head: H
Tail: T
P(HH)=0.5*0.5=0.25
P(HT)=0.5*0.5=0.25
P(TH)=0.5*0.5=0.25
P(TT)=0.5*0.5=0.25
P(HH)+P(HT)+P(TH)+P(TT)=1

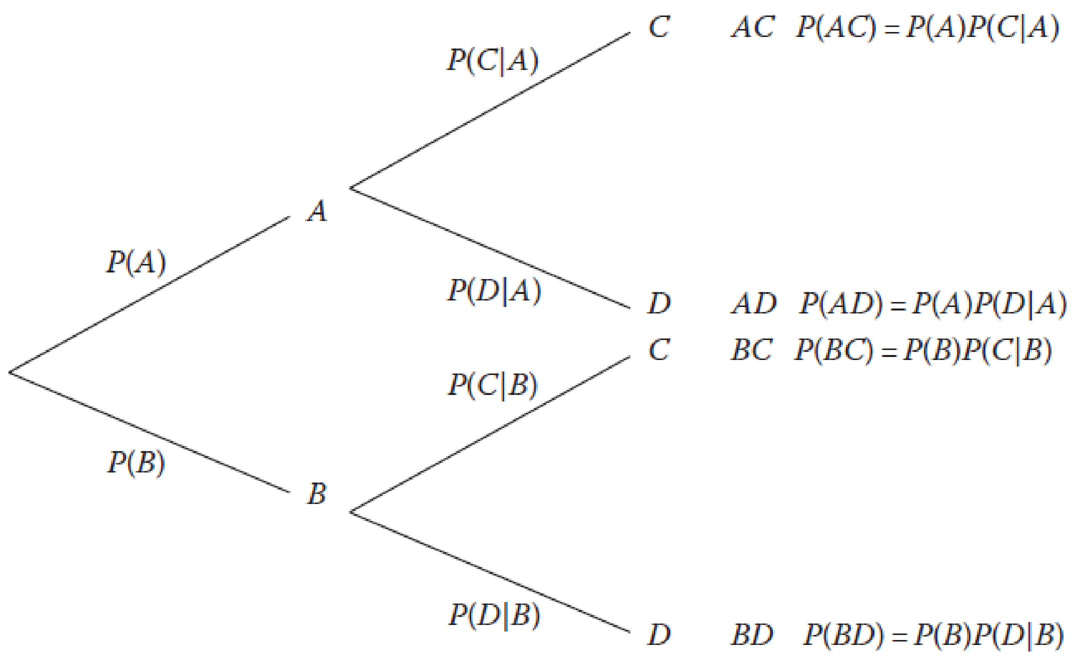
If one has three dice what is the probability of getting three 4s?
The probability of getting a 4 on one die is 1/6
The probability of getting 3 4s is:
P(4and4and4)=1/6*1/6*1/6=1/216

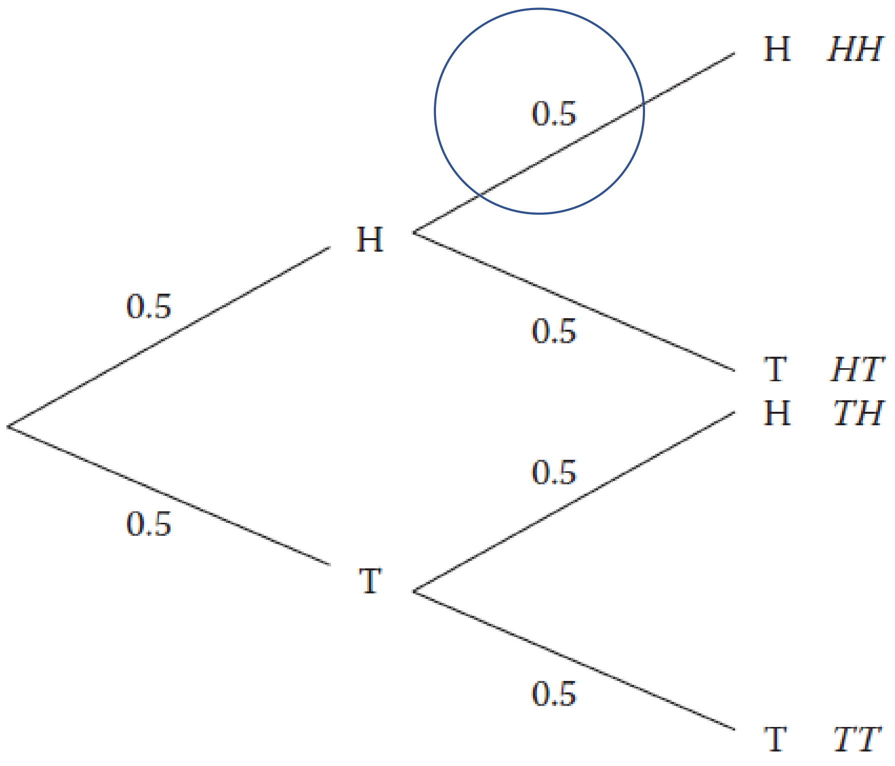


Coin1	Coin2		
Head	Head	P(HH)	All possible outcomes (U)
Head	Tail	P(HT)	
Tail	Head	P(TH)	
Tail	Tail	P(TT)	

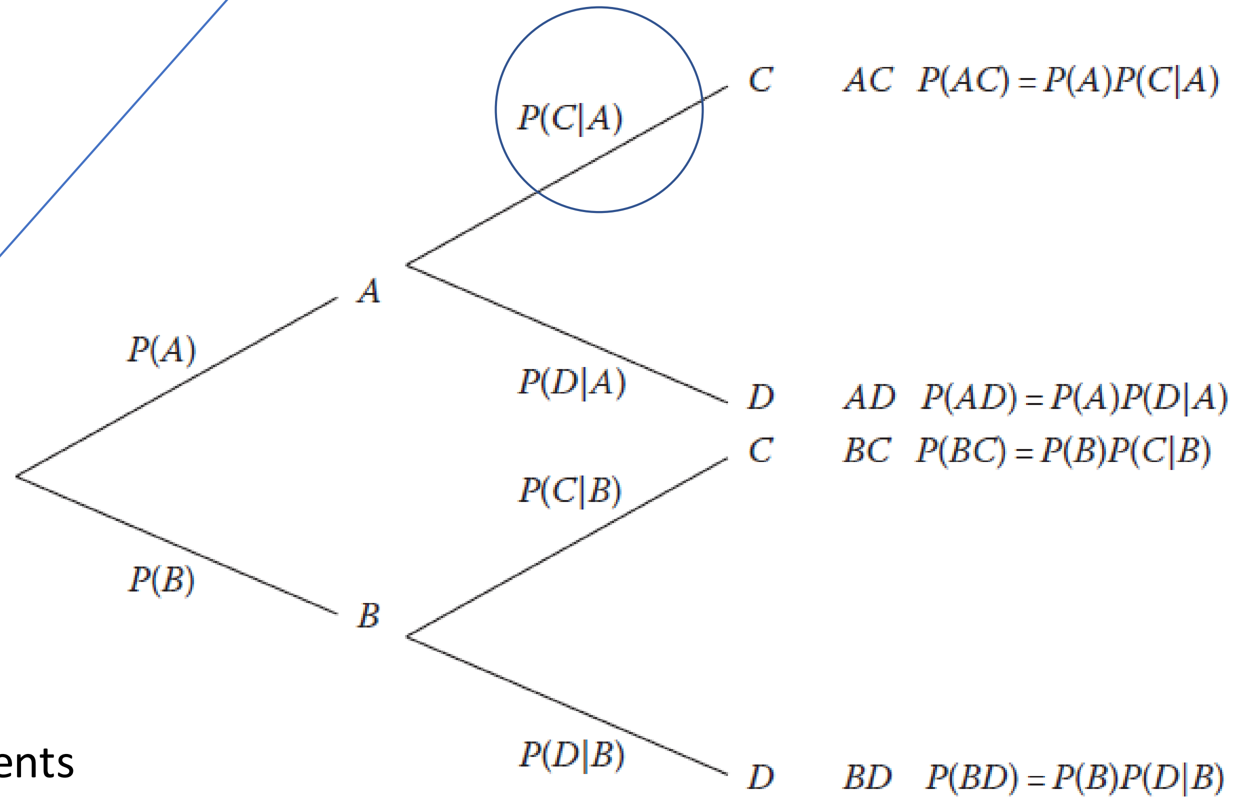
Two or more dependent Events

- What is the probability for you to choose two red cards in **a deck of cards**?
- P(red)=26/52=1/2
- P(red)=25/51 (probability)
- P(2red)=1/2*25/51=25/102





Independent events



dependent events

Example

Draw three red heart As from three decks of cards (one card from one deck, totally we will draw three times), each deck of card has 52 cards, then

What the probability of drawing three red heart As?

If we mix the three decks of cards together, we draw three times and get three red heart As, then

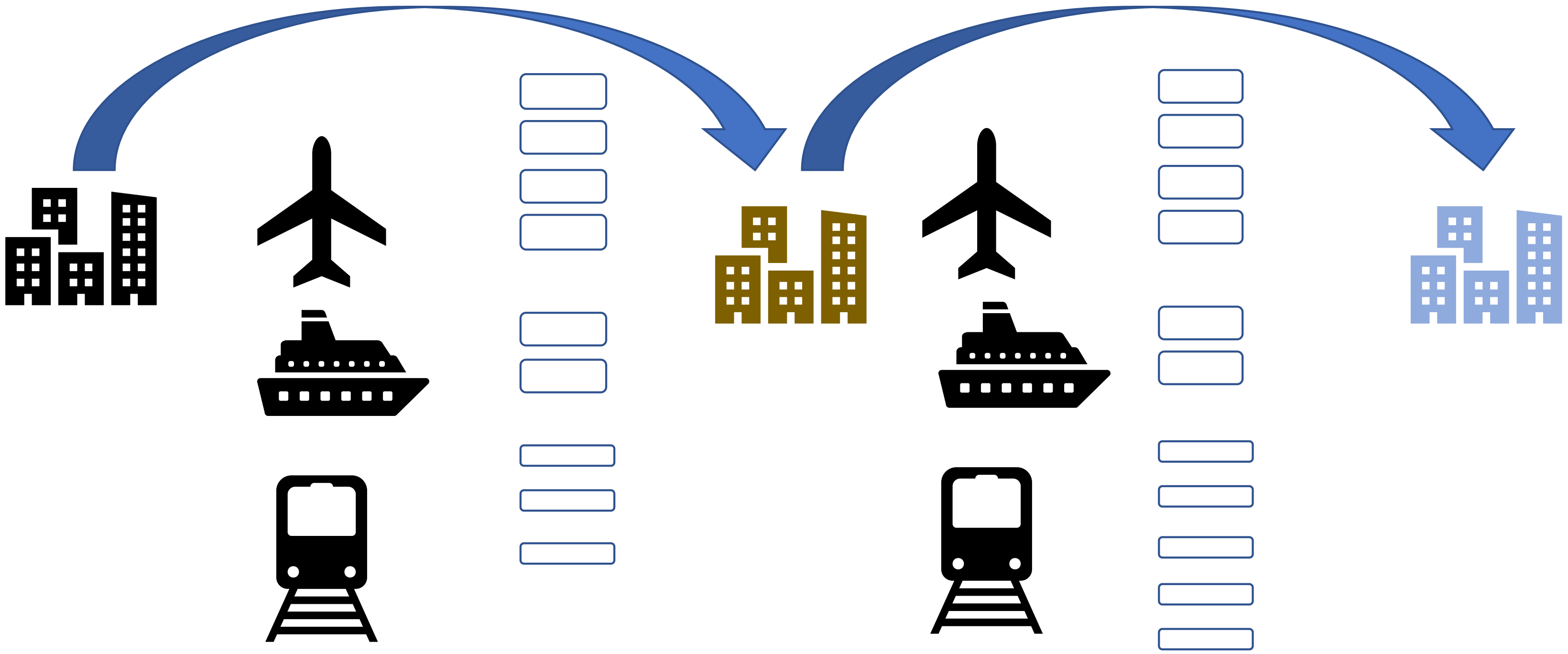
What the probability of drawing three red heart As?

If one has three decks of cards, what is the probability of getting three red heart As?
The probability of getting a red heart A in a deck of cards, is 1/52
The probability of **getting 3 red heart cards** is:
P(**red heart A** and **red heart A** and **red heart A**)=1/52*1/52*1/52=1/(52*52*52)=1/140608

If the three decks of cards are mixed together,
The probability of **getting 3 red heart cards** is:
P(**red heart A** and **red heart A** and **red heart A**)=3/(52*3)*2/(52*3-1)*1/(52*3-2)=1/620620

If the three decks of cards are mixed together, then The probability of **getting a red heart cards:**
P(**red heart A)** =3/(52*3)=1/52

Permutations and combinations



All ways: 9 * 11 = 99

Permutation and Combination

- The concepts of and differences between permutations and combinations can be illustrated by examination of all the different ways in which a pair of objects can be selected from five distinguishable objects—such as the letters A, B, C, D, and E. If both the letters selected and the order of selection are considered, then the following 20 outcomes are possible:

- List of the 20 potential combinations of the letters A, B, C, D, and E.

AB	BA	AC	CA	AD
DA	AE	EA	BC	CB
BD	DB	BE	EB	CD
DC	CE	EC	DE	ED

- For combinations, *k* objects are selected from a set of *n* objects to produce subsets without ordering. Contrasting the previous permutation example with the corresponding combination, the AB and BA subsets are no longer distinct selections; by eliminating such cases there remain only 10 different possible subsets—AB, AC, AD, AE, BC, BD, BE, CD, CE, and DE.

Permutation and Combination

"Order does/does not matter" and "Repeats are/are not allowed".

- When the order **doesn't** matter, it is a **Combination**.
- When the order **does** matter it is a **Permutation**.

Permutations

$$\mathbf{n^r}$$

where ***n*** is the number of things to choose from,
and we choose ***r*** of them,
repetition is allowed,
and order matters.

- There are basically two types of permutation:

  - **Repetition is Allowed**: such as the lock above. It could be "333".
  - **No Repetition**: for example the first three people in a running race. You can't be first *and* second.

$$\frac{n!}{(n-r)!}$$

where ***n*** is the number of things to choose from,
and we choose ***r*** of them,
no repetitions,
order matters.

The **factorial function** (symbol: **!**) just means to multiply a series of descending natural numbers. Examples:

- $4! = 4 \times 3 \times 2 \times 1 = 24$
- $7! = 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 5{,}040$
- $1! = 1$

$$P(n,r) = {}^nP_r = {}_nP_r = \frac{n!}{(n-r)!}$$

Combinations

$$\binom{r+n-1}{r} = \frac{(r+n-1)!}{r!(n-1)!}$$

where ***n*** is the number of things to choose from,
and we choose ***r*** of them
repetition allowed,
order doesn't matter.

- There are also two types of combinations (remember the order does **not** matter now):
- **Repetition is Allowed**: such as coins in your pocket (5,5,5,10,10)
- **No Repetition**: such as lottery numbers (2,14,15,27,30,33)

$$\frac{n!}{r!(n-r)!} = \binom{n}{r} = \binom{n}{n-r}$$

where ***n*** is the number of things to choose from,
and we choose ***r*** of them,
no repetition,
order doesn't matter.

$$C(n,r) = {}^{n}C_{r} = {}_{n}C_{r} = \binom{n}{r} = \frac{n!}{r!(n-r)!}$$

Just remember the formula:

$$\frac{n!}{r!(n-r)!}$$

It is often called "n choose r" (such as "16 choose 3")
And is also known as the **Binomial Coefficient**.

$$C(n,r) = {}^{n}C_{r} = {}_{n}C_{r} = \binom{n}{r} = \frac{n!}{r!(n-r)!}$$

$$P(n,r) = {}^{n}P_{r} = {}_{n}P_{r} = \frac{n!}{(n-r)!}$$

Just remember the formula:

$n!$: n *factorial*

5 Conditional Probability

- Consider **an experiment performed in two steps**. At the first step, an event can occur with probability $P[A]$ and then at the second step the occurrence of an event B is **dependent** on whether A occurred.
- We use a vertical bar to denote the occurrence of one event conditioned on another

P(A and B)=P(A)*P(B after A)

=P(B)*P(A after B)

$$\Pr[B \text{ occurs given } A] = \Pr[B|A] = P[B|A]$$

The key relation here is that A and B must have non-null intersection $P[AB] \neq 0$ and that the ratio of the probability of this intersection to the probability of A will determine P[B|A]; this is to say,

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

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, where\ P(B) \neq 0$$

Rules of probability

- **Rule of Subtraction**.

The probability that event A will occur is equal to 1 minus the probability that event A will <u>not</u> occur.

$P(A) = 1 - P(A')$

- **Rule of Addition**

The rule of addition applies to the following situation. We have two events, and we want to know the probability that either event occurs.

- **Rule of Multiplication**

The rule of multiplication applies to the situation when we want to know the probability of the intersection of two events; that is, we want to know the probability that two events (Event A and Event B) both occur.

What are Statistics?

- Statistics are *not only facts and figures*; they are something more than that. In the broadest sense, “statistics” refers to a range of techniques and procedures for analyzing, interpreting, displaying, and making decisions based on data.

Importance of Statistics

- It may be the single most important subject matter you study.

- You will continue to inundated with statistics throughout your life.

- Consider some examples...

Examples of Statistics

- People tend to be more persuasive when they look others directly in the eye and speak loudly and quickly.
- Women make 75 cents to every dollar a man makes at the same job.
- Eating egg whites can increase life span.
- 79.48% of all statistics are made up on the spot (just kidding!).

Some of These Statistics are Not True.

- But they are universal.
- They seem to add credibility.
- It is important to become a consumer of research and of statistics.

First step: Question statistics

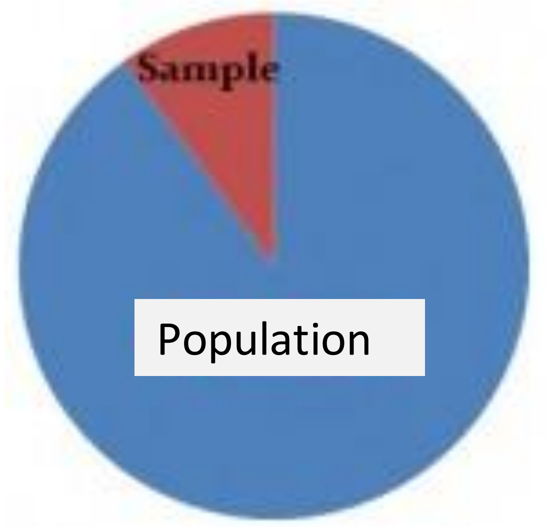
"There are three kinds of lies -- lies, damned lies, and statistics."

- Benjamin Disraeli,
Former British Prime Minister

Populations and Samples

- The study of statistics revolves around the study of data sets
- Two important types of data sets - **populations** and **samples**. Along the way, we'll introduce simple random sampling, the main method used in this tutorial to select samples.

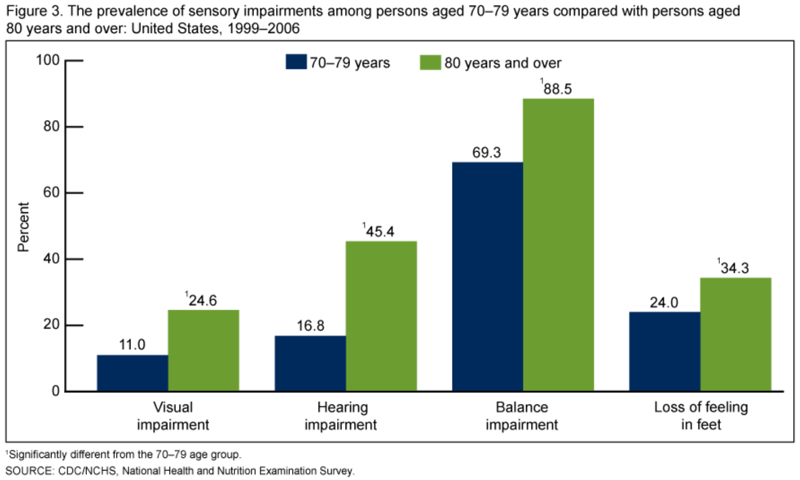
Statistics	----	Probability
Sample		Event
Population		Outcomes Space



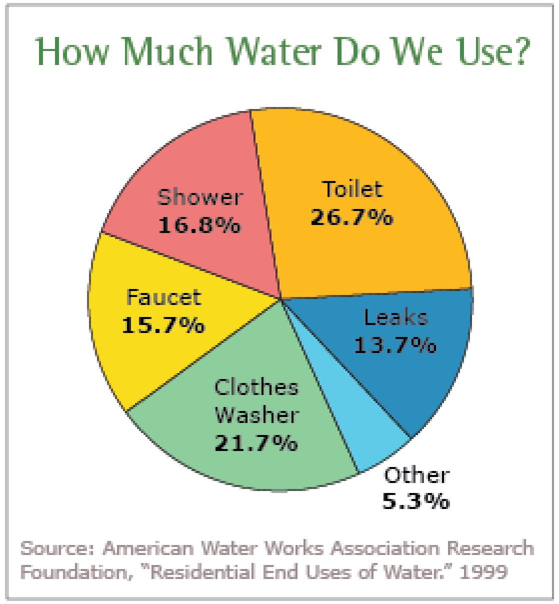
Descriptive Statistics

- By descriptive we mean that we are not attempting to make inferences or predictions but mainly to characterize the data, typically a **sample** of the theoretical population defined by the distribution.
- In simpler words, we want to tell what the data look like. Here, we use the term **statistic** to refer to a calculation on the sample, such as the sample mean and variance that we will discuss in **Exploratory Data Analysis** refers to performing descriptive statistics, and using a collection of visual and numerical tools such as quantile–quantile plots, boxplots, and autocorrelation

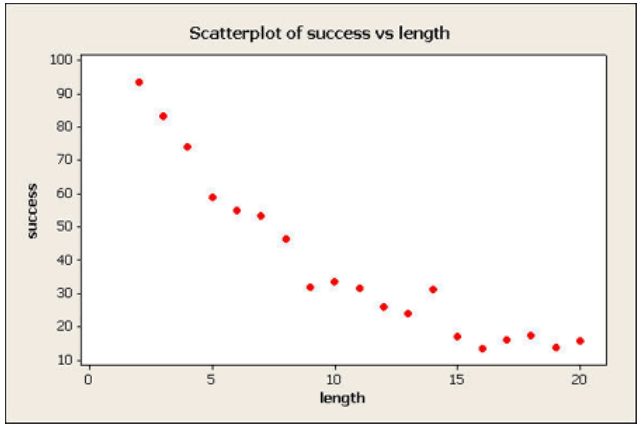
Common graph idioms



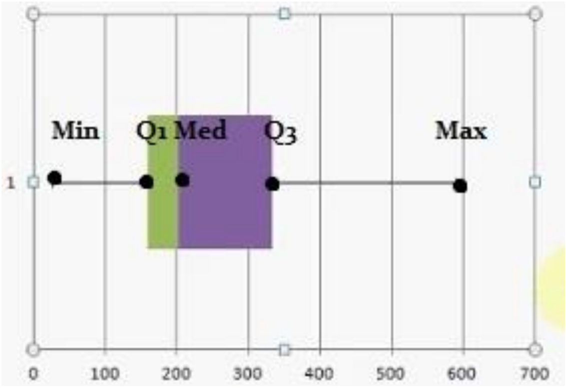
Grouped bar graph. Image: CDC.



Pie chart showing water consumption. Image courtesy of EPA



Scatter plot Image: Penn State



Box and whiskers graph

Inferential Statistics

- Once we require answers to specific questions about the samples, we enter the realm of **inferential statistics**. The following are examples of questions we typically pose. Is a sample drawn from a **normal** distribution? Is it drawn from the same distribution as this other sample? Is there a trend?
- The question is often posed as a **hypothesis** that can be tested to be falsified. We must learn two important classes of methods: **parametric** (e.g., t and F tests) and **nonparametric** (e.g., Wilcoxon, Spearman). The first type is applied when the *assumed distribution complies with certain conditions,* and thus more conclusive, *whereas the second type is less restrictive in assumptions but less conclusive.*

Sampling

- Such sampling requires every member of the population to have an equal chance of being selected into the sample. In addition, the selection of one member must be independent of the selection of every other member. That is, picking one member from the population must not increase or decrease the probability of picking any other member (relative to the others).

- A substitute teacher wants to know how students in the class did on their last test. The teacher asks the 10 students sitting in the front row to state their latest test score. He concludes from their report that the class did extremely well.

What is the sample? What is the population? Can you identify any problems with choosing the sample in the way that the teacher did?

Sampling Techniques

- **Random Assignment**

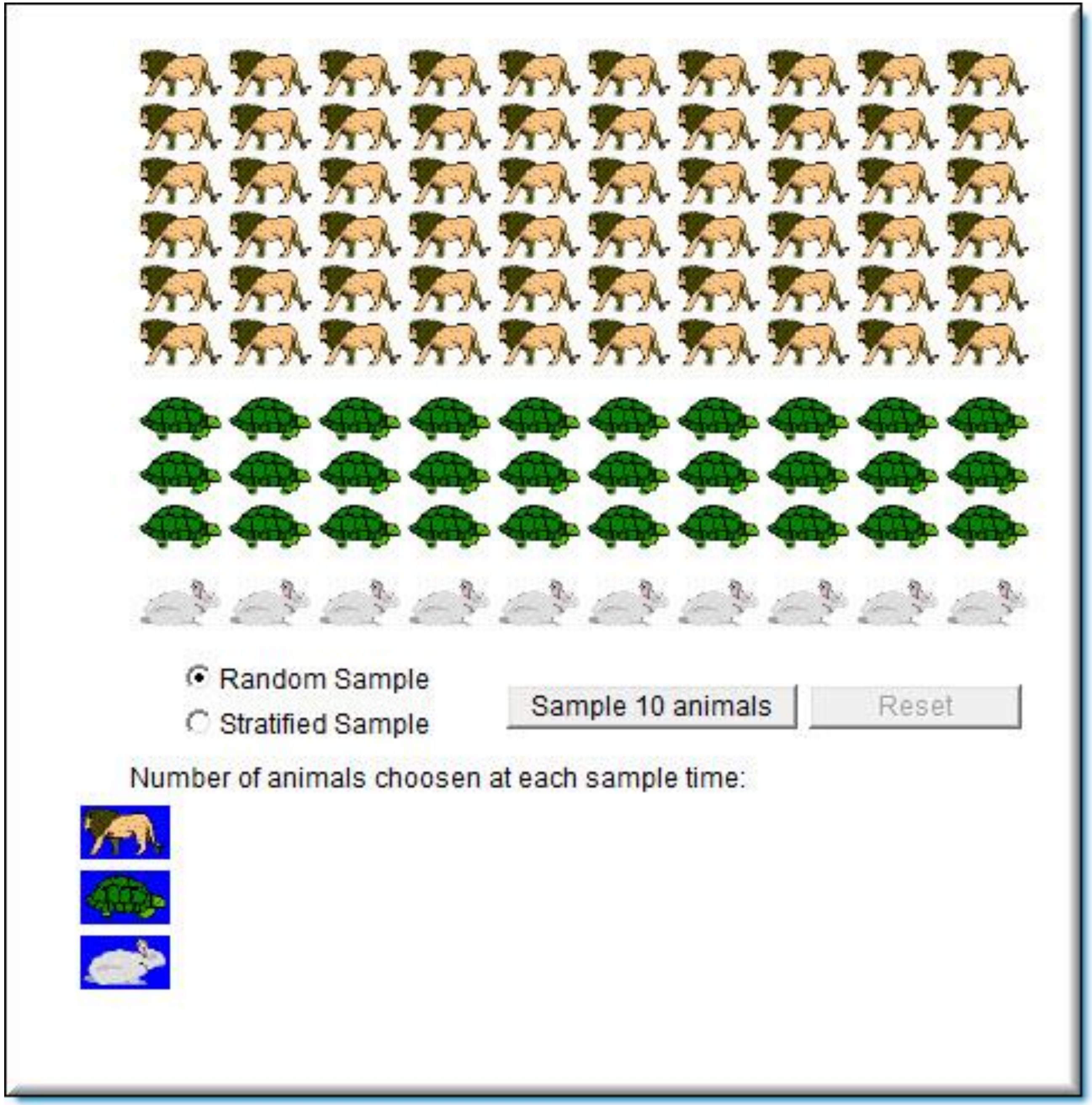
  This random division of the sample into different groups is called random assignment

- **Stratified Sampling**

  This method can be used if the population has a number of distinct "strata" or groups. In stratified sampling, you first identify members of your sample who belong to each group. Then you randomly sample from each of those subgroups in such a way that the sizes of the subgroups in the sample are proportional to their sizes in the population.

Sample Size Matters

- Random samples, especially if the sample size is small, are not necessarily representative of the entire population.
- Only a large sample size makes it likely that our sample is close to representative of the population.
- Inferential statistics take into account the sample size when generalizing results from samples to populations.


Random Sample
Stratified Sample
Sample 10 animals
Reset
Number of animals choosen at each sample time:

Math Review

Rounding. In this text, values are usually rounded to two decimal places (to the nearest hundredth). To do so, if the first digit of the numbers to be dropped is less than 5, simply drop them. If the first digit of the numbers to be dropped is 5 or greater, round up. One exception is if a calculation results in a whole number. In this case, it is optional to add zeros in the decimal places.

9.34782 rounds to 9.35

123.39421 drops to 123.39

74.99603 rounds to 75 or 75.00

Proportions and Percentages

Proportions and Percentages. A proportion is a part of a whole number that can be expressed as a fraction or as a decimal. For example, in a class of 40 students, 6 earned "A's." The proportion of the class that received A's can be expressed as a fraction (6/40), or as a decimal (.15).

To change a fraction to a decimal, simply divide the numerator by the denominator:

$$6/40 = 6 \div 40 = .15$$

Proportions and Percentages (Cont'd.)

To change a proportion (decimal) to a percentage, simply multiply by 100 (or move the decimal two places to the right) and place a percent sign (%) after the answer:

.1823 X 100 = 18.23%

To change a percentage to a proportion (decimal), remove the percent sign and divide by 100 (or move the decimal two places to the left):

15% = 15 ÷ 100 = .15
23.68% = 23.68 ÷ 100 = .2368

Signed Numbers

The following rules will help you to determine how to add, subtract, multiply, and divide signed numbers:

Addition. When adding values that include *both positive and negative* numbers: a) add all of the positive numbers, b) add all of the negative numbers, and c) determine the difference between the two sums, using the sign of the larger number for the result.

$$(-6) + (-8) + (10) + (-7) + (5) + (-1) = 15 + (-22) = -7$$

When adding only negative values, add as if they were positive and then attach a negative sign in the result.

$$(-3) + (-17) + (-42) + (-18) + (-5) + (-21) = -106$$

Signed Numbers (Cont'd.)

Subtraction. Change the sign of the value to be subtracted, and add.

$(-14) - (-5) = -14 + 5 = -9$
$20 - (-4) = 20 + 4 = 24$
$52 - (+73) = 52 + (-73) = -21$

Signed Numbers (Cont'd.)

Multiplication. Multiplication will be indicated by either a times sign (X) or by two values, each in parentheses, next to each other with no sign between them. If two values with the *same sign* are to be multiplied together, the result will be positive.

$-12 \times (-3) = 36$

$(-7)(-7) = 49$

If two values with the *opposite sign* are to be multiplied together, the result will be negative.

$$(-7)(7) = -49$$

Signed Numbers (Cont'd.)

Division. If the numerator and the denominator both have the *same sign*, the result will be positive.

$$(-15) \div (-3) = 5$$

If they have *opposite signs*, the result will be negative.

$$6 \div (-3) = -2$$

Order of Operations

Remember the mnemonic "*<u>P</u>lease <u>E</u>xcuse <u>M</u>y <u>D</u>ear <u>A</u>unt <u>S</u>ally*" from childhood as the order in which mathematical operations should be completed. It is still relevant today and translates as *<u>P</u>arentheses, <u>E</u>xponents, <u>M</u>ultiplication, <u>D</u>ivision, <u>A</u>ddition, <u>S</u>ubtraction*. Some rules follow:

- Compute all values that are in parentheses first. Brackets and parentheses may be used interchangeably, but parentheses may also be nested inside of brackets. If this kind of nesting occurs, compute values in the innermost parentheses first.

$$
\begin{aligned}
&7 \text{ X } [4 - (3 \text{ X } 6)] \\
&= 7 \text{ X } [4 - 18] \\
&= 7 \text{ X } [-14] \\
&= -98
\end{aligned}
\qquad\qquad
\begin{aligned}
&-3 + [(2 \text{ X } 4) - (7 \text{ X } 7)] + 8 \\
&= -3 + [8 - 49] + 8 \\
&= -3 + [-41] + 8 \\
&= -36
\end{aligned}
$$

Order of Operations (Cont'd.)

- Exponents are next. The only exponent we will be using is 2. In other words, some of our values will need to be squared. To square a number means to multiply that number by itself. "Six squared" means "six times six" and is written as 6^2.

$$
\begin{aligned}
&4 + (9 \times 3) + (6 - 4^2) \\
&= 4 + 27 + (6 - 16) \\
&= 4 + 27 + (-10) \\
&= +21
\end{aligned}
$$

Order of Operations (Cont'd.)

- After parentheses and exponents have been taken care of, multiplication and division (from left to right) are next. Addition and subtraction (from left to right) are performed last.

$$(-7 + 4^2) \text{ X } 2 - [(3 \text{ X } 6) \div 2]$$
$$= (-7 + 16) \text{ X } 2 - [18 \div 2]$$
$$= 9 \text{ X } 2 - 9$$
$$= +9$$

Order of Operations (Cont'd.)

Order is important because the same numbers with parentheses and brackets in different places would result in a different amount as shown below.

$$
\begin{aligned}
& -7 + [(4^2 \text{ X } 2) - (3 \text{ X } 6)] \div 2 \\
&= -7 + [(16 \text{ X } 2) - 18] \div 2 \\
&= -7 + [32 - 18] \div 2 \\
&= -7 + [14] \div 2 \\
&= -7+7 \\
&= 0
\end{aligned}
$$

Summation Operator

The Greek letter *sigma* (Σ) is used as a symbol for the summation operator. This is a frequently used notation in statistics that tells you to add the value of whatever variable(s) follows to the right of the symbol. Variables are often represented as X or Y. Remember to keep the order of operations in mind when using the summation operator. Suppose you have the following X and Y scores:

X	Y	X^2	XY	(X -3)	Y + 2)	$(Y + 2)^2$
6	4	36	24	3	6	36
5	7	25	35	2	9	81
9	2	81	18	6	4	16
8	1	64	8	5	3	9
28	14	206	85	16	22	142

Summation Operator (Cont'd.)

$\Sigma X = 28$	Simply add all of the X values.
$(\Sigma X)^2 = 784$	Remember, parentheses first. Sum the X values. Then, square the sum of the X values.
$\Sigma X^2 = 206$	Here, exponents come first. Square each X value first, then find the sum of the X^2 column.
$\Sigma XY = 85$	Multiplication comes before addition. Thus, multiply each X value by each Y value (XY), then add the column for the XY values.
$\Sigma(X - 3) = 16$	Subtract 3 from each X value, then add the $(X - 3)$ column.
$\Sigma(Y + 2)^2 = 142$	Parentheses first, then exponents, then addition. Thus, add 2 to each Y value, then square the $(Y + 2)$ values, and finally sum the $(Y + 2)^2$ column.