Chapter 3

Revision to Probability and Statistics

Introduction

Probability and Statistics

- Statistics is the mathematical science behind the problem “what can I know about a population if I’m unable to reach every member?”

Probability and Statistics

- If we could measure the height of every resident of Australia, then we could make a statement about the average height of Australians at the time we took our measurement.
- This is where random sampling comes in.

Probability and Statistics

- If we take a reasonably sized random sample of Australians and measure their heights, we can form a statistical inference about the population of Australia.
- Probability helps us know how sure we are of our conclusions!

Data

What is Data?

- Data = the collected observations we have about something.
- Data can be continuous: *"What is the stock price?"*
- or categorical: *"What car has the best repair history?"*

Why Data Matters

- Helps us understand things as they are:

"What relationships if any exist between two events?"

"Do people who eat an apple a day enjoy fewer doctor's visits than those who don't?"

Why Data Matters

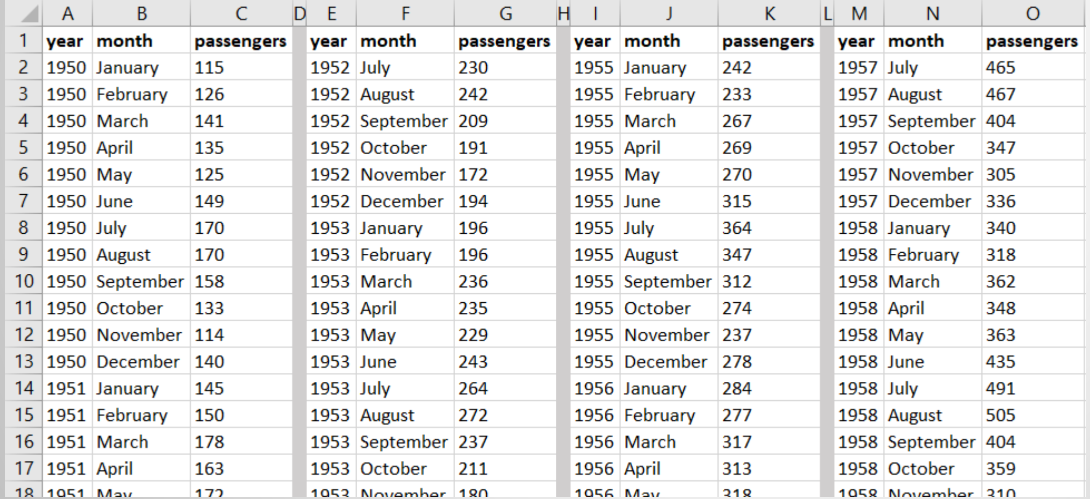
- Helps us predict future behavior to guide business decisions:

"Based on a user's click history which ad is more likely to bring them to our site?"

Visualizing Data

- Compare a table:

Flights



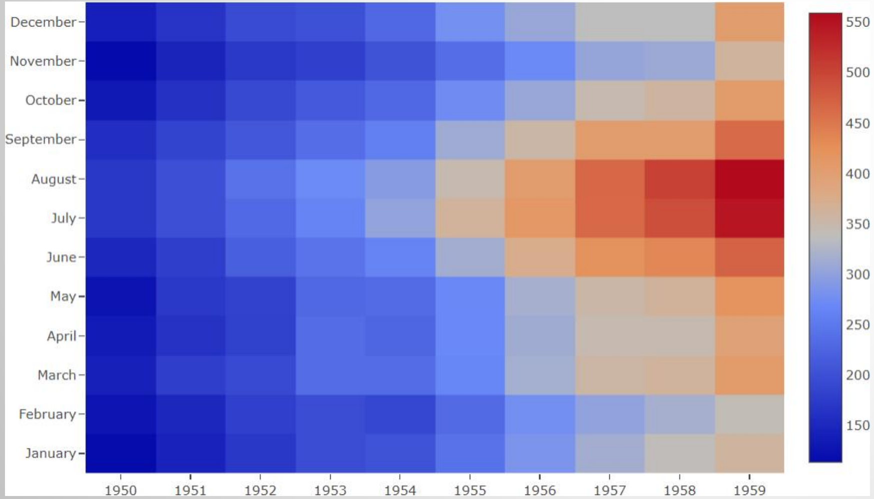
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	**year**	**month**	**passengers**		**year**	**month**	**passengers**		**year**	**month**	**passengers**		**year**	**month**	**passengers**
2	1950	January	115		1952	July	230		1955	January	242		1957	July	465
3	1950	February	126		1952	August	242		1955	February	233		1957	August	467
4	1950	March	141		1952	September	209		1955	March	267		1957	September	404
5	1950	April	135		1952	October	191		1955	April	269		1957	October	347
6	1950	May	125		1952	November	172		1955	May	270		1957	November	305
7	1950	June	149		1952	December	194		1955	June	315		1957	December	336
8	1950	July	170		1953	January	196		1955	July	364		1958	January	340
9	1950	August	170		1953	February	196		1955	August	347		1958	February	318
10	1950	September	158		1953	March	236		1955	September	312		1958	March	362
11	1950	October	133		1953	April	235		1955	October	274		1958	April	348
12	1950	November	114		1953	May	229		1955	November	237		1958	May	363
13	1950	December	140		1953	June	243		1955	December	278		1958	June	435
14	1951	January	145		1953	July	264		1956	January	284		1958	July	491
15	1951	February	150		1953	August	272		1956	February	277		1958	August	505
16	1951	March	178		1953	September	237		1956	March	317		1958	September	404
17	1951	April	163		1953	October	211		1956	April	313		1958	October	359
18	1951	May	172		1953	November	180		1956	May	318		1958	November	310

Not much can be gained by reading it.

Visualizing Data

- to a graph:

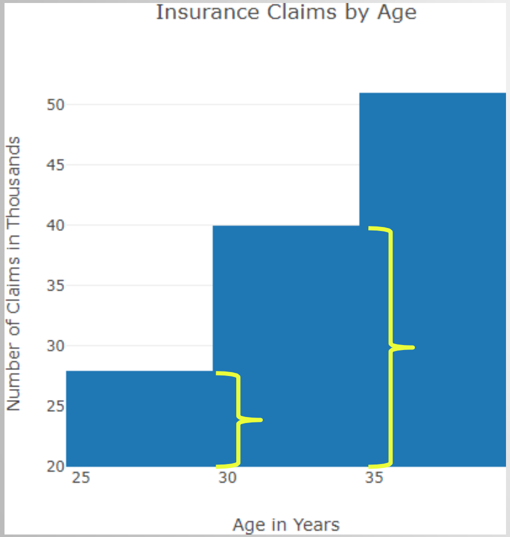
Flights

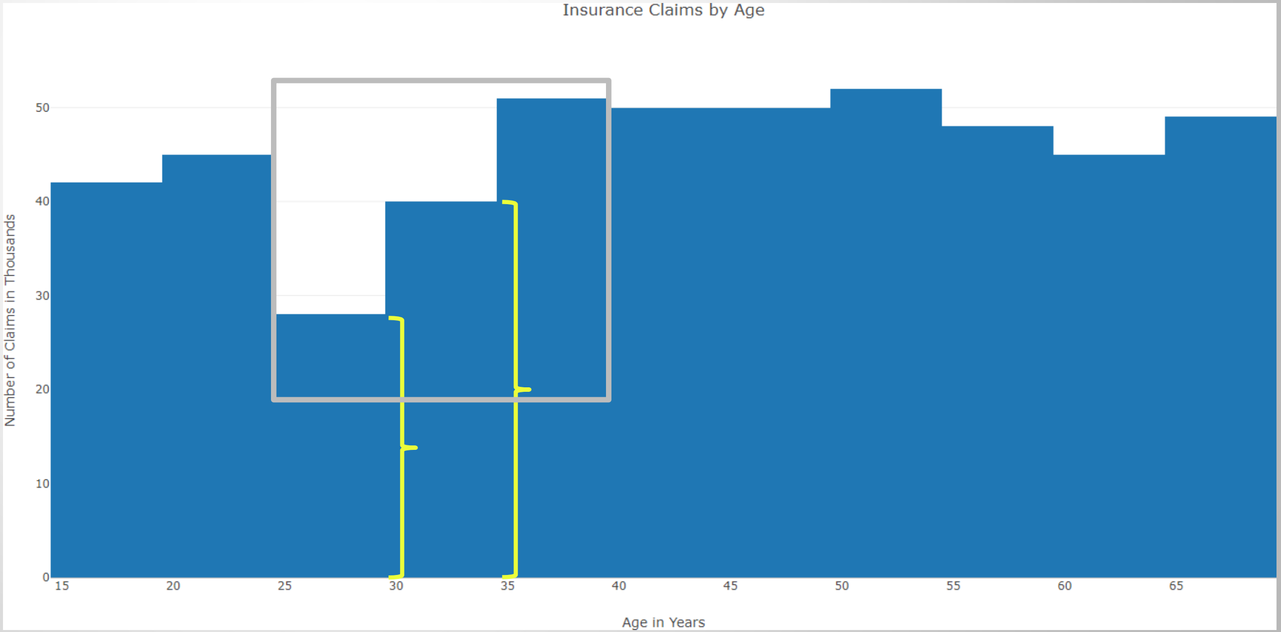


The graph uncovers two distinct trends - an increase in passengers flying over the years and a greater number of passengers flying in the summer months.

Analyze Visualizations Critically!

- Graphs can be misleading:





Measuring Data

Levels of Measurement

Nominal

- Predetermined categories
- Can’t be sorted

Animal classification (*mammal fish reptile*)

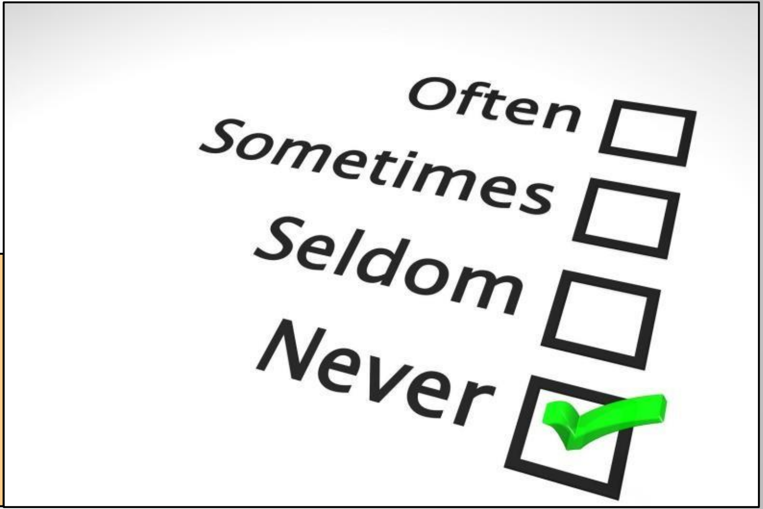
Political party (*republican democrat independent*)

Levels of Measurement

Ordinal

- Can be sorted
- Lacks scale

Survey responses

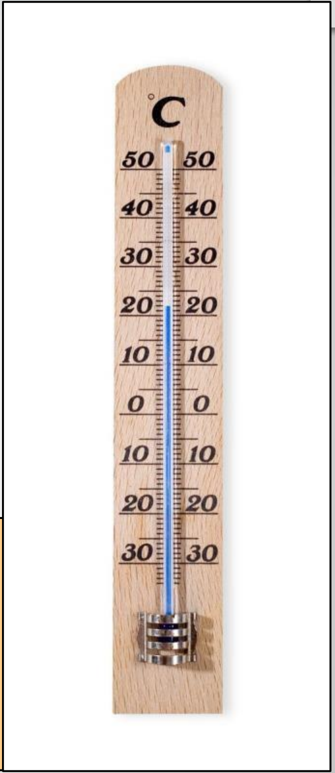


Levels of Measurement

Interval

- Provides scale
- Lacks a “zero” point

Temperature



Levels of Measurement

Ratio

- Values have a true zero point

Age, weight, salary

Population vs. Sample

- Population = every member of a group
- Sample = a subset of members that time and resources allow you to measure

Mathematical Symbols & Syntax

Symbol/Expression	Spoken as	Description
x^2	x squared	x raised to the second power $x^2 = x \times x$
x_i	x-sub-i	a subscripted variable (the subscript acts as a label)
$x!$	x factorial	$4! = 4 \times 3 \times 2 \times 1$
xЂ	x bar	symbol for the sample mean
μ	“mew”	symbol for the population mean (Greek lowercase letter mu)
Σ	sigma	syntax for writing sums (Greek capital letter sigma)

Exponents

$$\boldsymbol{x^5} = x \times x \times x \times x \times x$$

1 2 3 4 5

EXAMPLE: $3^4 = 3 \times 3 \times 3 \times 3 = 81$

Exponents – special cases

$$x^{-3} = \frac{1}{x \times x \times x}$$

EXAMPLE: $2^{-3} = \frac{1}{2\times2\times2} = \frac{1}{8} = 0.125$

$$x^{\left(\frac{1}{n}\right)} = \sqrt[n]{x}$$

EXAMPLE: $8^{\left(\frac{1}{3}\right)} = \sqrt[3]{8} = 2$

Factorials

$$\boldsymbol{x}! = x \times (x-1) \times (x-2) \times \cdots \times 1$$

EXAMPLE: $6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$

EXAMPLE: $\frac{5!}{3!} = \frac{5\times4\times\cancel{3\times2\times1}}{\cancel{3\times2\times1}} = 5 \times 4 = 20$

Series Sums

$$\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{x_i} = x_1 + x_2 + x_3 + \cdots + x_n$$

EXAMPLE: $x = \{5,3,2,8\}$

$$n = \#\ elements\ in\ x = 4$$

$$\sum_{\boldsymbol{i=1}}^{\boldsymbol{4}} \boldsymbol{x_i} = 5 + 3 + 2 + 8 = 18$$

Equation Example

- Formula for calculating a sample mean:

$$\bar{x} = \sum_{i=0}^{n} \frac{x_i}{n}$$

- Read out loud:

“$\boldsymbol{x}$ bar (the symbol for the sample mean) is equal to the sum (indicated by the Greek letter sigma) of all the $\boldsymbol{x}$-sub-$\boldsymbol{i}$ values in the series as $\boldsymbol{i}$ goes from 1to the number $\boldsymbol{n}$ items in the series divided by $\boldsymbol{n}$."

Equation Example

$$\bar{x} = \sum_{i=0}^{n} \frac{x_i}{n}$$

1. Start with a series of values:

{7, 8, 9, 10}

2. Assign placeholders to each item

{7, 8, 9, 10}

1 2 3 4 n=4

3. These become x_1 x_2 etc.

$$x_1 = 7 \quad x_2 = 8 \quad x_3 = 9 \quad x_4 = 10$$

Equation Example

4. Plug these into the equation:

$$\bar{x} = \sum_{i=0}^{n} \frac{x_i}{n} = \frac{x_1 + x_2 + x_3 + x_4 \ldots + x_n}{n}$$

$$= \frac{7 + 8 + 9 + 10}{4} \quad = \frac{34}{4} = 8.5$$

Measures of Central Tendency

Measurements of Data

- “What was the average return?”

  Measures of Central Tendency

- “How far from the average did individual values stray?”

  Measures of Dispersion

Measures of Central Tendency (mean, median, mode)

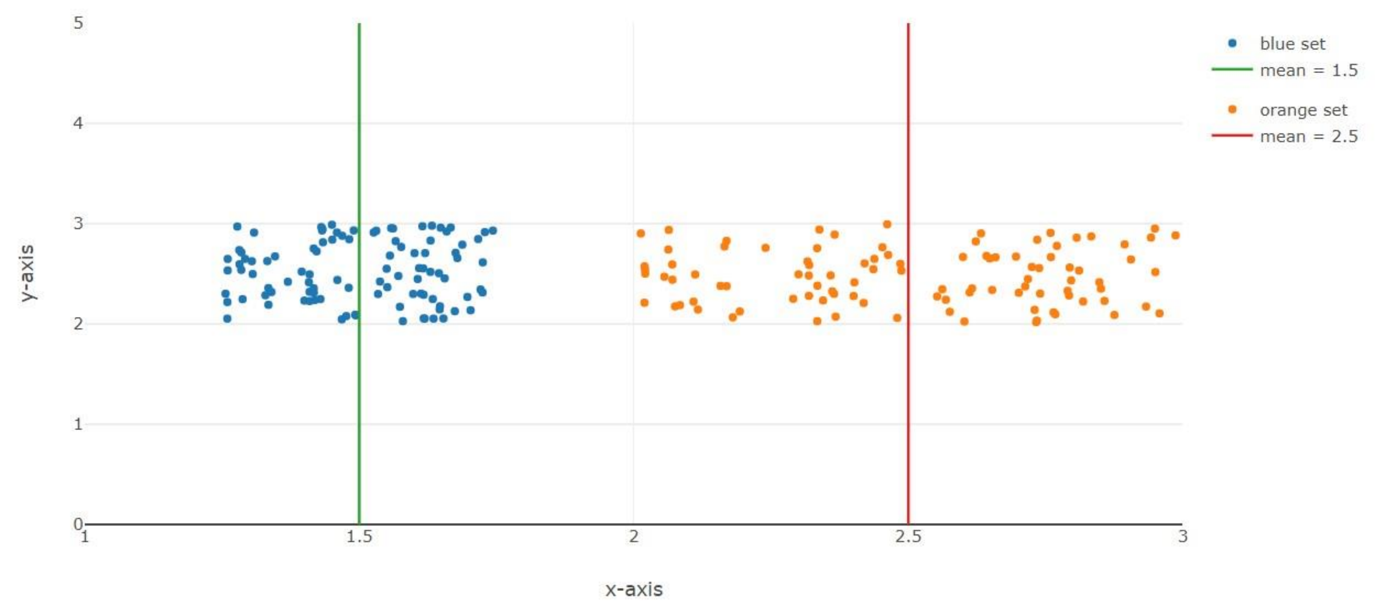
- Describe the “location” of the data
- Fail to describe the “shape” of the data

mean = “calculated average”

median = “middle value”

mode = “most occurring value”

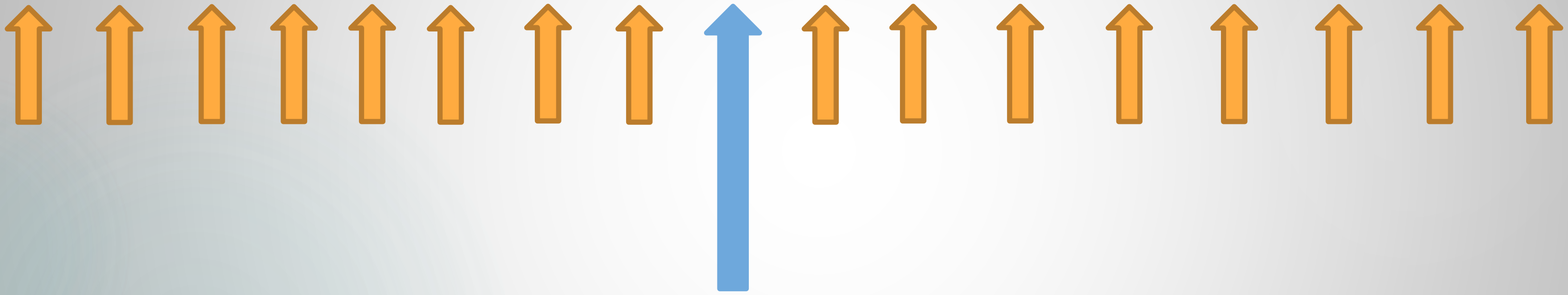
Measures of Central Tendency



- Shows “location” but not “how spread out”

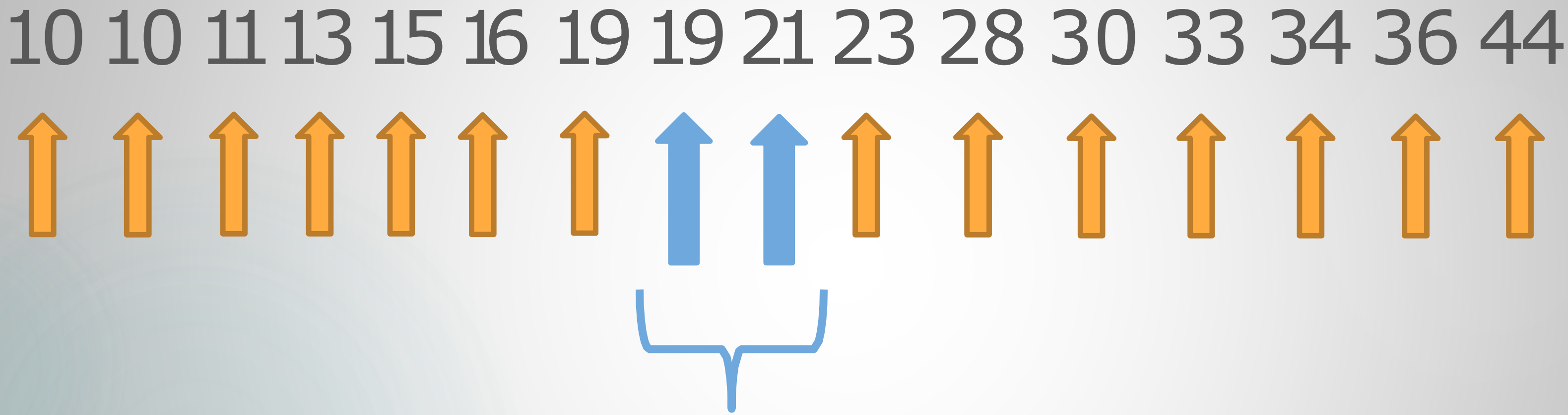
Median – *odd number of values*

9 10 10 11 13 15 16 19 19 21 23 28 30 33 34 36 44



$= 19$

Median - *even number of values*



10 10 11 13 15 16 19 19 21 23 28 30 33 34 36 44

$$\frac{19+21}{2} = 20$$

Mean vs. Median

- The mean can be influenced by *outliers*.
- The mean of {2,3,2,3,2,12} is 4
- The median is 2.5
- The median is much closer to most of the values in the series!

Mode

10 10 11 13 15 16 16 16 21 23 28 30 33 34 36 44

$= 16$

Measures of Dispersion

Measures of Dispersion (range, variance, standard deviation)

9 10 11 13 15 16 19 19 21 23 28 30 33 34 36 39

- In this sample the mean is 22.25
- How do we describe how "spread out" the sample is?

Range

9 10 11 13 15 16 19 19 21 23 28 30 33 34 36 39

$$Range = max - min$$

$$= 39 \ - 9$$

$$= 30$$

Variance

- Calculated as the sum of square distances from each point to the mean
- There’s a difference between the SAMPLE variance and the POPULATION variance
- subject to Bessel's correction $(\boldsymbol{n-1})$

Variance

SAMPLE VARIANCE: $$s^2 = \frac{\Sigma(x-\bar{\mathrm{x}})^2}{n-1}$$

POPULATION VARIANCE: $$\sigma^2 = \frac{\Sigma(X-\mu)^2}{N}$$

Sample Variance

$$s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$$

4 7 9 8 11

$$\bar{x} = \frac{4+7+9+8+11}{5} = \frac{39}{5} = 7.8$$

sample mean

$$s^2 = \frac{(4-7.8)^2+(7-7.8)^2+(9-7.8)^2+(8-7.8)^2+(11-7.8)^2}{5-1}$$

$$= 6.7$$

sample variance

Standard Deviation

- square root of the variance
- benefit: same units as the sample
- meaningful to talk about *“values that lie within one standard deviation of the mean”*

Sample Standard Deviation

$$s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$$

Sample:

4 7 9 8 11

xпђ $= \frac{4+7+9+8+11}{5} = \frac{39}{5} = 7.8$

sample mean

$$s = \sqrt{\frac{(4-7.8)^2+(7-7.8)^2+(9-7.8)^2+(8-7.8)^2+(11-7.8)^2}{5-1}}$$

$$= \sqrt{6.7} = 2.59$$

sample standard deviation

Population
Standard Deviation

$$\sigma = \sqrt{\frac{\Sigma(X-\mu)^2}{N}}$$

Population:

4	7	9	8	11

$$\mu = \frac{4+7+9+8+11}{5} = \frac{39}{5} = \boxed{7.8}$$

population mean

$$\sigma = \sqrt{\frac{(4-7.8)^2+(7-7.8)^2+(9-7.8)^2+(8-7.8)^2+(11-7.8)^2}{5}}$$

$$= \sqrt{5.36} = 2.32$$

population standard deviation

Probability

What is Probability?

- Probability is a value between 0 and 1 that a certain event will occur
- For example, the probability that a fair coin will come up heads is 0.5
- Mathematically we write:

$$P(E_{heads}) = 0.5$$

What is Probability?

- In the above “heads” example, the act of flipping a coin is called a trial.
- Over very many trials, a fair coin should come up “heads” half of the time.

Trials Have No Memory!

- If a fair coin comes up tails 5 times in a row, the chance it will come up heads is *still* 0.5
- You can’t think of a series of independent events as needing to “catch up” to the expected probability.
- Each trial is independent of all others

Experiments and Sample Space

- Each trial of flipping a coin can be called an experiment
- Each mutually exclusive outcome is called a simple event
- The sample space is the sum of every possible simple event

Experiments and Sample Space

- Consider rolling a six-sided die
- One roll is an experiment
- The simple events are:

$E_1=1 \quad E_2=2 \quad E_3=3$

$E_4=4 \quad E_5=5 \quad E_6=6$

- Therefore, the sample space is:

$S = \{E_1 , E_2 , E_3 , E_4, E_5, E_6\}$

Experiments and Sample Space

- The probability that a fair die will roll a six:

  The simple event is:

  $E_6 = 6$ (one event)

  Total sample space:

  $S = \{E_1, E_2, E_3, E_4, E_5, E_6\}$ (six possible outcomes)

  The probability:

  $P(\text{Roll Six}) = 1/6$

Probability Exercise

- A company made a total of 50 trumpet valves
- It is determined that one of the valves was defective
- If three valves go into one trumpet, what is the probability that a trumpet has a defective valve?

Probability Exercise

1. Calculate the probability of having a defective valve:

$$P\left(E_{defective\ valve}\right) = \frac{1}{50} = 0.02$$

Probability Exercise

2. Calculate the probability of having a defective trumpet:

$$P\left(E_{defective\ trumpet}\right) = 3 \times P\left(E_{defective\ valve}\right)$$

$$= 3 \times 0.02 = \mathbf{0.06}$$

Permutations

Permutations

- A permutation of a set of objects is an arrangement of the objects in a certain order.
- The possible permutations of letters a, b and c is:

abc acb bac bca cab cba

Permutations

- For simple examples like abc, we calculate the number of possible permutations as $n!$ *("n factorial")*
- abc = 3 items
- $n! = 3! = 3 \times 2 \times 1 = 6 \text{ permutations}$

Permutations

- You can also take a subset of items in a permutation
- The number of permutations of a set of n objects taken r at a time is given by the following formula:

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

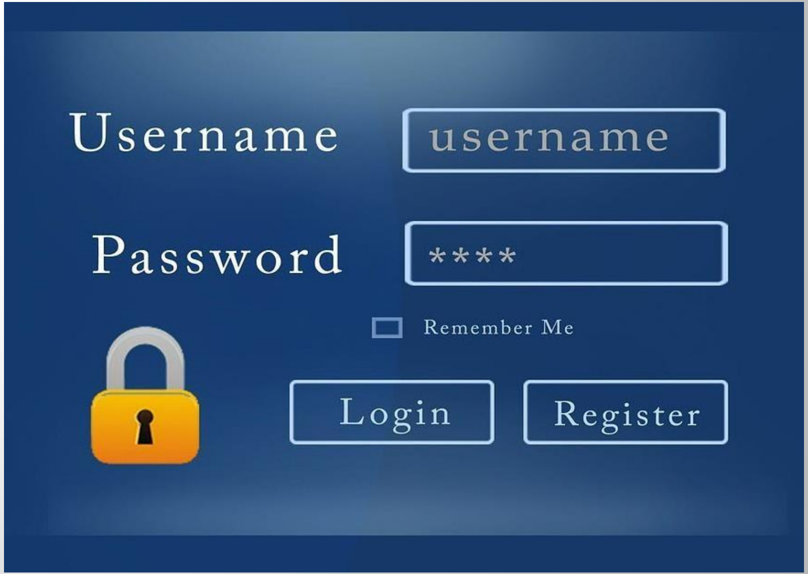
Permutations Example #1

A website requires a 4 character password

Characters can either be lowercase letters or the digits 0-9.

You may not repeat a letter or number.

How many different passwords can there be?



Permutations Solution #1

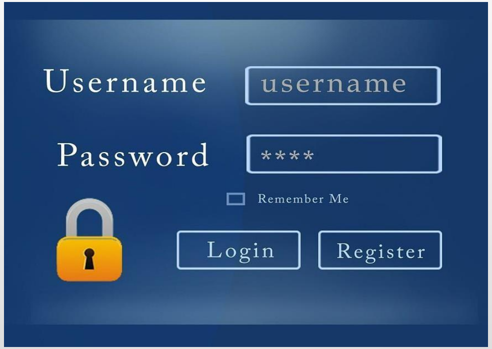
- Recognize that n, or the number of objects is 26 letters $+ 10$ numbers $= 36$
- r, or the number of objects taken at one time is 4
- Plug those numbers into the formula:

$$_{36}P_4 = \frac{36!}{(36-4)!}$$

Permutations Solution #1

$$ {}_{36}P_4 = \frac{36!}{(36-4)!} = \frac{36 \times 35 \times 34 \times 33 \times \cancel{32 \times 31 \ldots}}{\cancel{32 \times 31 \ldots}} $$

$$ = 36 \times 35 \times 34 \times 33 = \mathbf{1{,}413{,}720} \text{ permutations} $$



Permutations Allowing Repetition

- The number of arrangements of n objects taken r at a time, *with repetition* is given by

$$n^r$$

Permutations Example #2

How many 4 digit license plates can you make using the numbers 0 to 9 while allowing repetition?

Permutations Solution #2

Recognize there are 10 objects taken 4 at a time. Plug that into the formula:

$$n^r = 10^4 = 10{,}000 \text{ permutations}$$

Permutations Formulas

- Total Permutations of a set n

$$n!$$

- Permutations taken r at a time given set n (no repetition)

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

- Permutations taken r at a time given set n (with repetition)

$$n^{r}$$

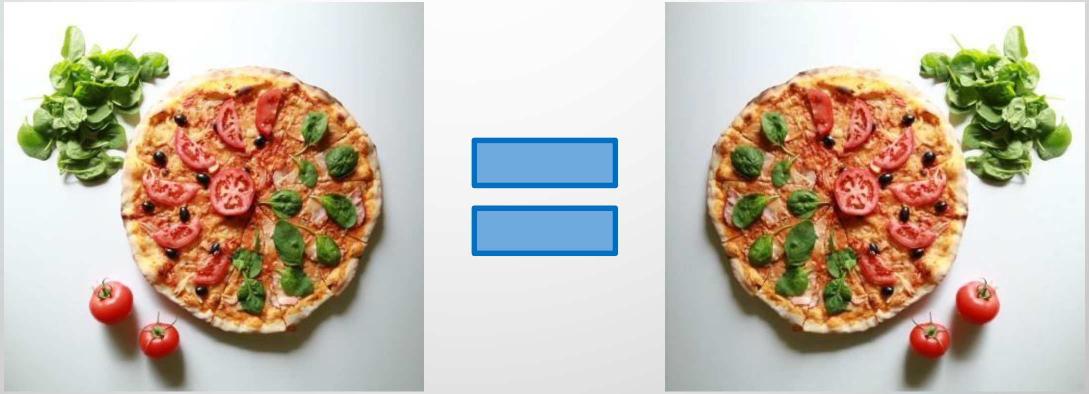
Combinations

Combinations

- *Unordered* arrangements of objects are called combinations.
- A group of people selected for a team are the same group, no matter the order.

Combinations

- *Unordered* arrangements of objects are called combinations.
- A pizza that is half tomato, half spinach is the same as one half spinach, half tomato.



Combinations

- The number of combinations of a set of n objects taken r at a time is given by:

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

Combinations vs. Permutations

How many 3-letter combinations can be made from the letters ABCDE?

1. Permutations:

$${}_5P_3 = \frac{5!}{(5-3)!} = 5 \times 4 \times 3 = \mathbf{60}$$

ABC	ACB	BAC	BCA	CAB	CBA
ABD	ADB	BAD	BDA	DAB	DBA
ABE	AEB	BAE	BEA	EAB	EBA
ACD	ADC	CAD	CDA	DAC	DCA
ACE	AEC	CAE	CEA	EAC	ECA
ADE	AED	DAE	DEA	EAD	EDA
BCD	BDC	CBD	CDB	DBC	DCB
BCE	BEC	CBE	CEB	EBC	ECB
BDE	BED	DBE	DEB	EBD	EDB
CDE	CED	DCE	DEC	ECD	EDC

Combinations vs. Permutations

How many 3-letter combinations can be made from the letters ABCDE?

2. Realize each row contains the same letters

ABC	ACB	BAC	BCA	CAB	CBA
ABD	ADB	BAD	BDA	DAB	DBA
ABE	AEB	BAE	BEA	EAB	EBA
ACD	ADC	CAD	CDA	DAC	DCA
ACE	AEC	CAE	CEA	EAC	ECA
ADE	AED	DAE	DEA	EAD	EDA
BCD	BDC	CBD	CDB	DBC	DCB
BCE	BEC	CBE	CEB	EBC	ECB
BDE	BED	DBE	DEB	EBD	EDB
CDE	CED	DCE	DEC	ECD	EDC

Combinations vs. Permutations

How many 3-letter combinations can be made from the letters ABCDE?

3. Combinations:

$${}_nC_r = \frac{n!}{r!\,(n-r)!} = \frac{5!}{3!\cdot 2!}$$

$$= \frac{5\times 4\times 3}{3\times 2} = \mathbf{10}$$

ABC	ACB	BAC	BCA	CAB	CBA
ABD	ADB	BAD	BDA	DAB	DBA
ABE	AEB	BAE	BEA	EAB	EBA
ACD	ADC	CAD	CDA	DAC	DCA
ACE	AEC	CAE	CEA	EAC	ECA
ADE	AED	DAE	DEA	EAD	EDA
BCD	BDC	CBD	CDB	DBC	DCB
BCE	BEC	CBE	CEB	EBC	ECB
BDE	BED	DBE	DEB	EBD	EDB
CDE	CED	DCE	DEC	ECD	EDC

Combinations Example #1

- For a study, 4 people are chosen at random from a group of 10 people.
- How many ways can this be done?

Combinations Solution #1

Since you're going to have the same group of people no matter the order they’re chosen, you can set up the problem as a combination:

$${}_{n}C_{r} = \frac{n!}{r!\,(n-r)!} = \frac{10!}{4!\,(10-4)!} = 210$$

Combinations Example #1a

For a pizza, 4 ingredients are chosen from a total of 10 ingredients.

How many different combinations of pizza can we have?
In this situation we're only allowed to use each ingredient once.

Combinations Solution #1a

Same as before, there will be 210 different types of pizza you can make:

$$ {}_{n}C_{r} = \frac{n!}{r!\,(n-r)!} = \frac{10!}{4!\,(10-4)!} = 210 $$

Combinations Solution #1a

But what if we're allowed to repeat ingredients? (Use pepperoni 3 times and then add tomato once)

Combinations with Repetition

- The number of combinations taken r at a time from a set n and allowing for repetition:

$$ {}_{n+r-1}C_r = \frac{(n+r-1)!}{r!\,(n-1)!} $$

Combinations Example #2

- For a pizza, 4ingredients are chosen at random from a possible of 10 ingredients.

- How many different pizza topping combinations are there,
- allowing repetition?

Combinations Solution #2

4 ingredients selected from 10 possible ingredients, allowing for repetition is:

$${}_{n+r-1}C_r = \frac{(n+r-1)!}{r!\,(n-1)!} = \frac{13!}{4!\,(9)!} = 715$$

Combinations with/without repetition

How many 3-letter combinations can be made from the letters ABCDE?

without repetition:

$$ {}_{n}C_{r} = \frac{n!}{r!\,(n-r)!} = \frac{5!}{3!\cdot 2!} = \mathbf{10} $$

with repetition:

$$ {}_{n+r-1}C_{r} = \frac{(n+r-1)!}{r!\,(n-1)!} = \frac{7!}{3!\,(4)!} = \mathbf{35} $$

ABC	ABD	ABE	ACD	ACE
ADE	BCD	BCE	BDE	CDE

ABC	ABD	ABE	ACD	ACE
ADE	BCD	BCE	BDE	CDE
AAA	AAB	AAC	AAD	AAE
BBA	BBB	BBC	BBD	BBE
CCA	CCB	CCC	CCD	CCE
DDA	DDB	DDC	DDD	DDE
EEA	EEB	EEC	EED	EEE

Permutations & Combinations in Excel

Order matters?	Repetition?	Formula	**In Excel**
Yes (permutation)	No	$_nP_r = \frac{n!}{(n-r)!}$	=PERMUT(n,r)
No (combination)	No	$_nC_r = \frac{n!}{r!\,(n-r)!}$	=COMBIN(n,r)
Yes (permutation)	Yes	n^r	=PERMUTATIONA(n,r)
No (combination)	Yes	$_{n+r-1}C_r = \frac{(n+r-1)!}{r!\,(n-1)!}$	=COMBINA(n,r)

Intersections, Unions & Complements

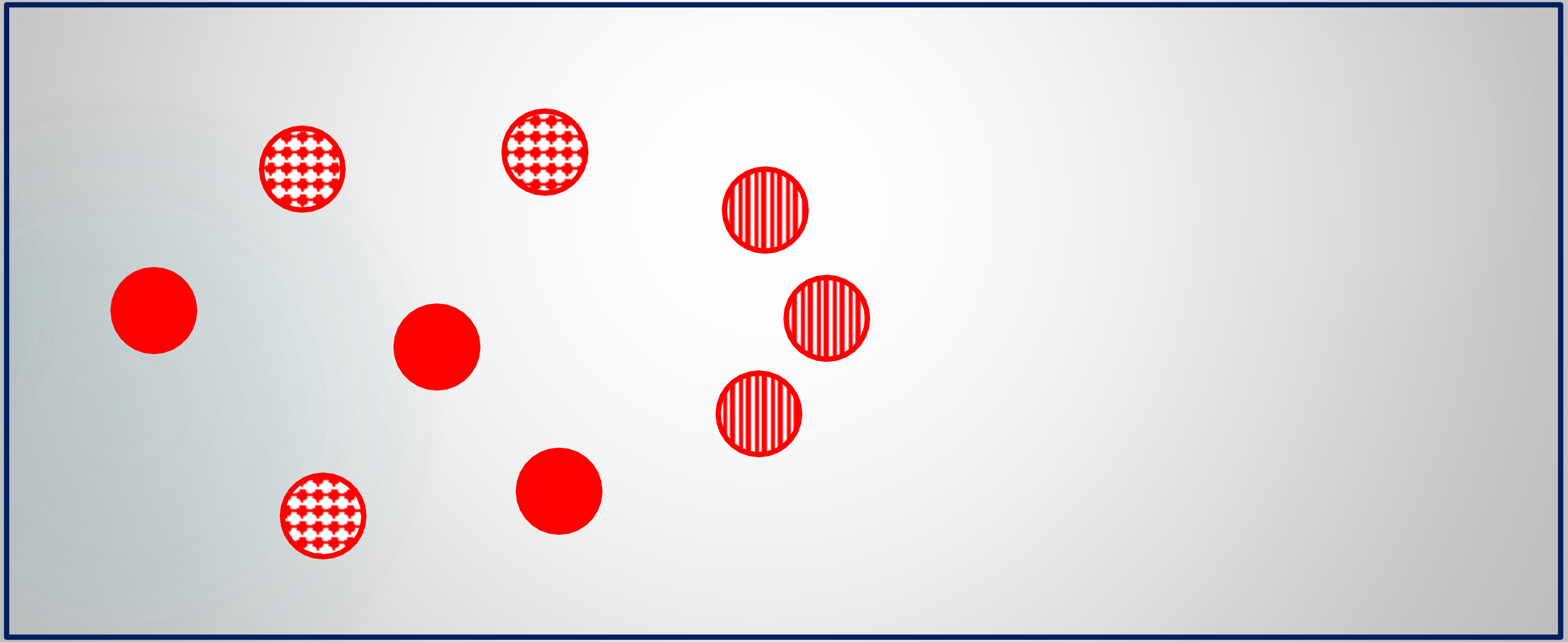
Intersections

- In probability, an intersection describes the sample space where two events *both* occur.

- Consider a box of patterned, colored balls

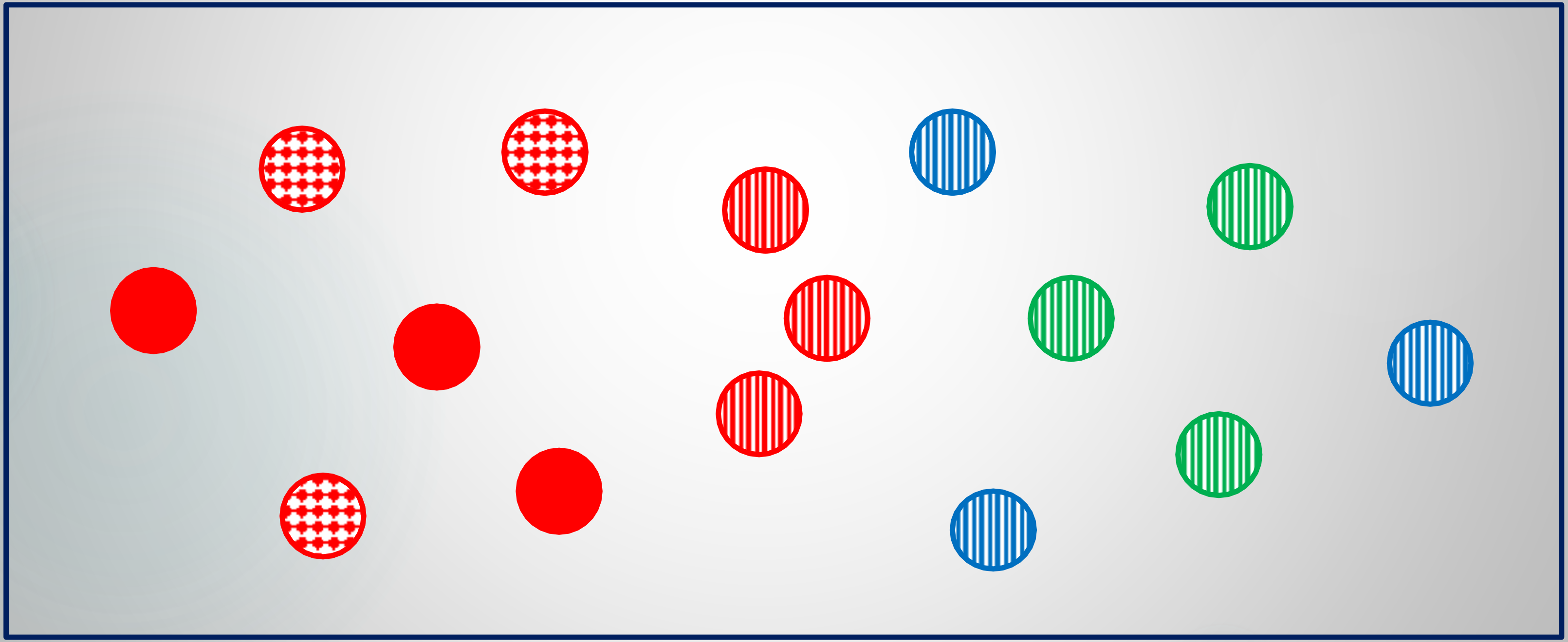
Intersections

- 9 of the balls are red:



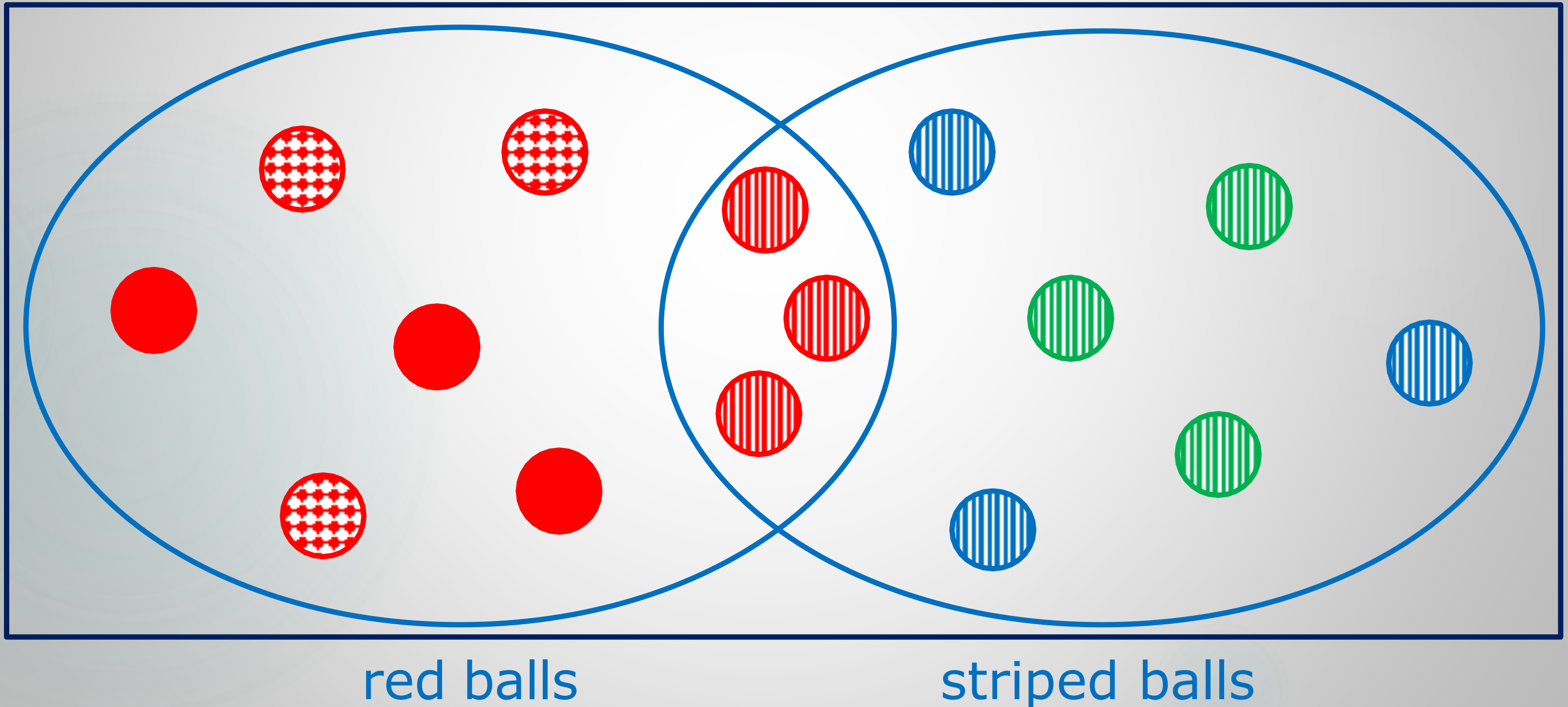
Intersections

- 9 of the balls are striped:



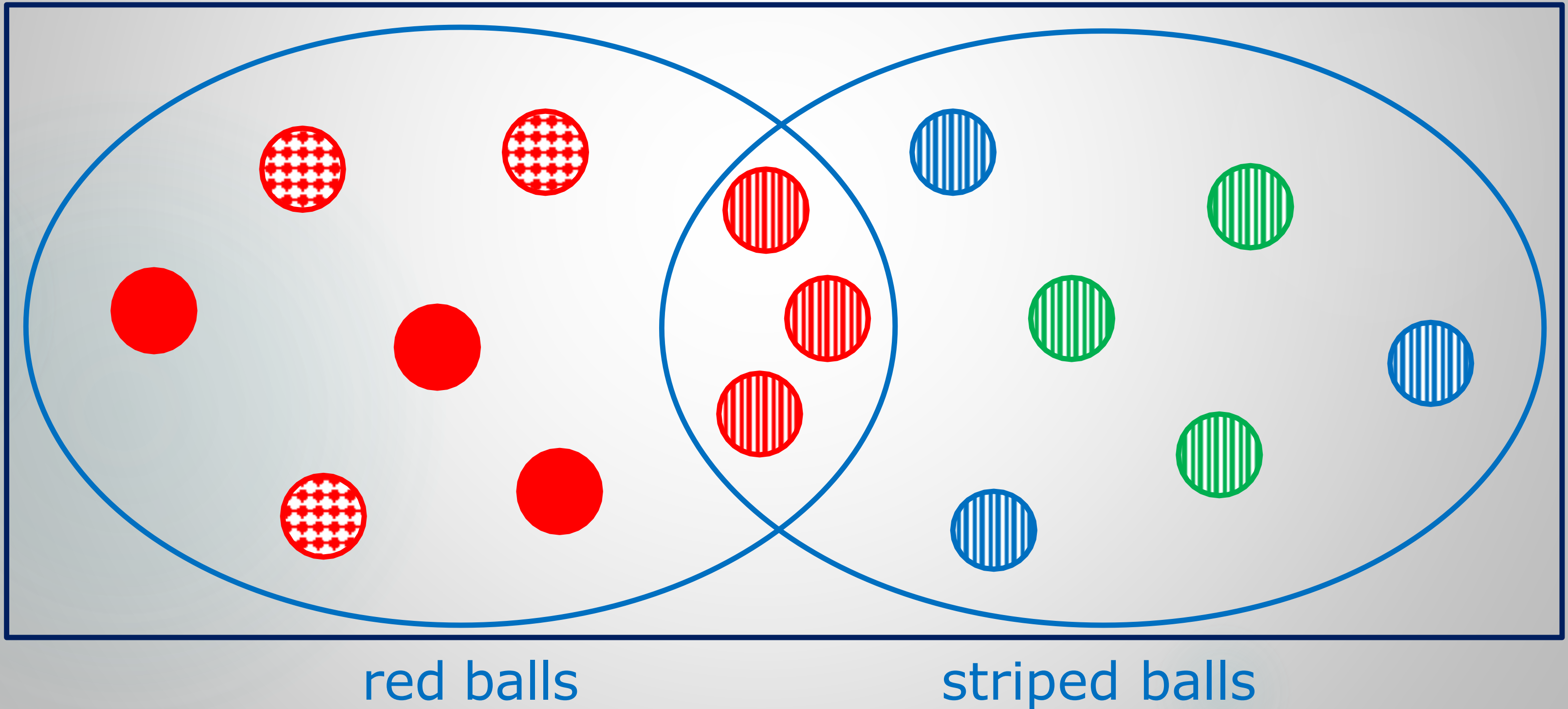
Intersections

- 3 of the balls are both red and striped:

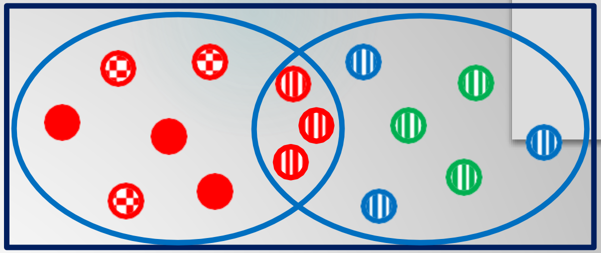


Intersections

- What are the odds of a red, striped ball?



Intersections



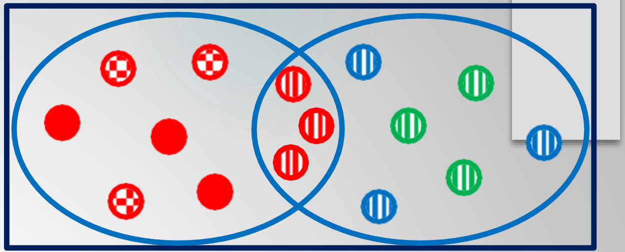
- If we assign A as the event of red balls, and B as the event of striped balls, the intersection of A *and* B is given as:

$$A \cap B$$

- Note that order doesn't matter:

$$A \cap B = B \cap A$$

Intersections



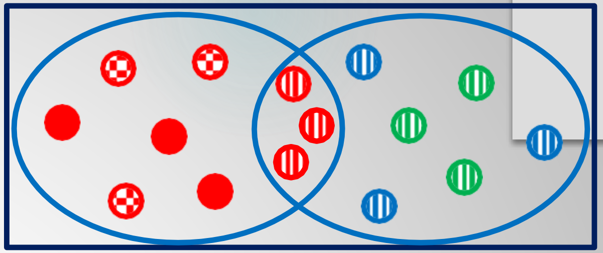
- The probability of A *and* B is given as

$$P(A \cap B)$$

- In this case:

$$P(A \cap B) = \frac{3}{15} = \mathbf{0.2}$$

Unions



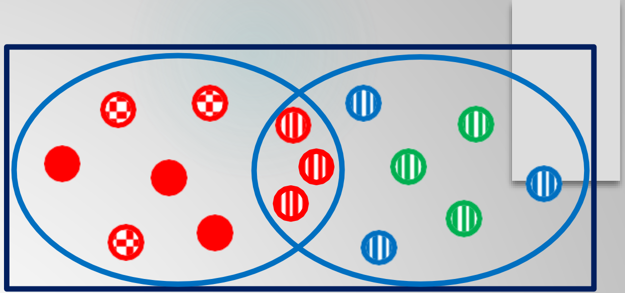
- The union of two events considers if A *or* B occurs, and is given as:

$$A \cup B$$

- Note again, order doesn't matter:

$$A \cup B = B \cup A$$

Unions



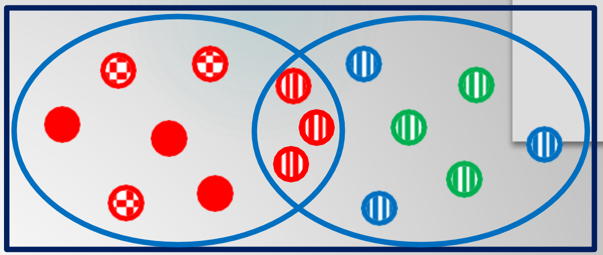
- The probability of A *or* B is given as:

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

- In this case:

$$P(A \cup B) = \frac{9}{15} + \frac{9}{15} - \frac{3}{15} = \frac{15}{15} = \mathbf{1.0}$$

Complements



- The complement of an event considers everything outside of the event, given by:

$$A^{\complement}$$

- The probability of *not* A is:

$$P(A^{\complement}) = 1 - P(A) = \frac{15}{15} - \frac{9}{15} = \frac{6}{15} = \mathbf{0.4}$$

Independent & Dependent Events

Independent Events

- An independent series of events occur when the outcome of one event has no effect on the outcome of another.
- An example is flipping a fair coin twice
- The chance of getting heads on the second toss is independent of the result of the first toss.

Independent Events

- The probability of seeing two heads with two flips of a fair coin is:

$$P(H_1H_2) = P(H_1) \times P(H_2)$$

$$= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

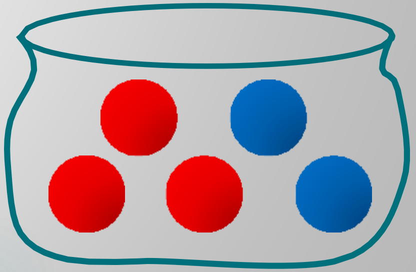
1st Toss	2nd Toss
H	H
H	T
T	H
T	T

Dependent Events

- A dependent event occurs when the outcome of a first event does affect the probability of a second event.
- A common example is to draw colored marbles from a bag *without replacement*.

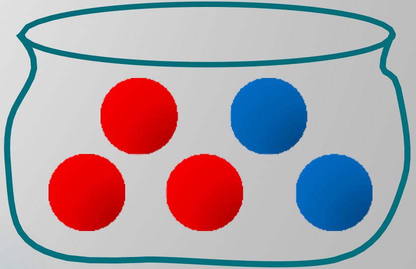
Dependent Events

- Imagine a bag contains 2 blue marbles and 3 red marbles.
- If you take two marbles out of the bag, what is the probability that they are both red?



Dependent Events

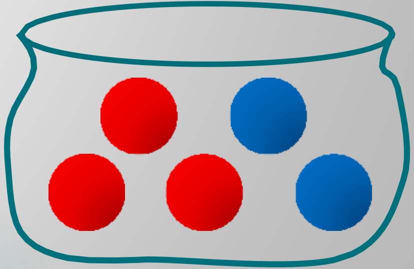
- Here the color of the first marble affects the probability of drawing a 2nd red marble.



Dependent Events

- The probability of drawing a first red marble is easy:

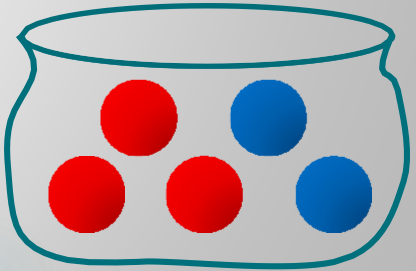
$$P(R_1) = \frac{3}{5}$$



Dependent Events

- The probability of drawing a second red marble *given that* the first marble was red is written as:

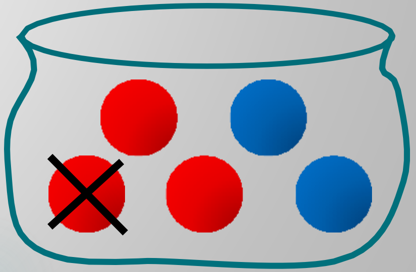
$$P(R_2|R_1)$$



Dependent Events

- After removing a red marble from the sample set this becomes:

$$P(R_2|R_1) = \frac{2}{4}$$



Dependent Events

- So the probability of two red marbles is:

$$P(R_1 \cap R_2) = P(R_1) \cdot P(R_2|R_1)$$

$$= \frac{3}{5} \times \frac{2}{4} = \frac{6}{20} = \mathbf{0.3}$$