L09: Introduction to Probability

Statistics is everywhere: the "chance of rain"

Berkeley, CA 10 Day Weather

7:54 am PST Print

DAY	DESCRIPTION	HIGH / LOW	PRECIP	WIND	HUMIDITY
TODAY FEB 13	Partly Cloudy	56°/45°	10%	SW 8 mph	78%
FRI FEB 14	Mostly Sunny	62°/45°	10%	WSW 5 mph	75%
SAT FEB 15	Partly Cloudy	64°/48°	10%	SW 7 mph	71%
SUN FEB 16	Partly Cloudy	62°/47°	10%	W 11 mph	71%
MON FEB 17	Sunny	66°/45°	0%	N 16 mph	36%
TUE FEB 18	Sunny	64°/44°	0%	N 11 mph	34%

Different definitions

According to the U.S. National Weather Service (NWS), PoP is the probability of excedance that more than 0.01 inches (0.25 mm) of precipitation will fall in a single spot, averaged over the forecast area. This can be expressed mathematically as

$$PoP = C \times A$$

where C is the confidence that any form of precipitation (e.g., snow or rain) will occur somewhere in the forecast area and A is the percent of the area that will receive measurable precipitation, if it occurs at all. For instance, if there is a 100% probability of rain covering one half of a city, and a 0% probability of rain on the other half of the city, the POP for the city would be 50%. A 50% chance of a rainstorm covering the entire city would also lead to a PoP of 50%.

The PoP measure is meaningless unless it is associated with a period of time. NWS forecasts commonly use PoP defined over 12-hour periods (PoP12), A "daytime" PoP12 means from 6 am to 6 pm.

Different definitions

Environment Canada reports a chance of precipitation (COP) that is defined as "The chance that measurable precipitation (0.2 mm of rain or 0.2 cm of snow) will fall on any random point of the forecast region during the forecast period." The values are rounded to 10% increments.

Different definitions

If there is a 20% chance that at least .3mm of rain will fall on 50% of a city, which forecast (US or Canadian) will give a higher percent probability of rain?

How might this change your behavior (bringing an umbrella or not)?

Learning objectives for today

- Why does probability matter?
- Terminology for probability: sample space (discrete vs. continuous), event, discrete probability model, continuous probability model
- The frequentist definition of probability and how sample size affects our estimation
- Some basic rules for probability
- Venn diagrams

Why is probability important:

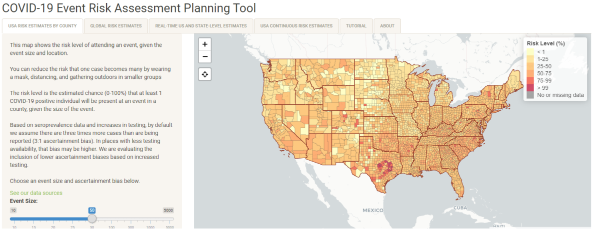
Although we don't always explicitly discuss probability, we think about probability all the time:

- Probability of rain today
- Probability of getting a job of a certain income range with a certain college degree
- Probability of a question showing up on the next exam

Statistics can be misleading, knowing the basic rules of probability and understanding how they are generated can help you to interpret statistics clearly, think critically about information and draw relevant conclusions.

Statistics is everywhere: COVID risk

An interesting project to check out is the COVID-19 Event Risk Assessment Planning Tool



Why is probability important:

- Statistics can be used to make and defend policy decisions - sometimes these are misleading (can be intentionally) You should know how to interpret those statistics for yourself
- Determining probability can be difficult and not intuitive. Our gut instinct about the probability of an event may be way off and lead to poor decision making in policy, medical or personal settings.
- Using predictive models to calculate probabilities sounds like an objective process, however these models can encode bias and discrimination.

Example. 1: Misleading statistics

- Kirstjen Nielsen, former Homeland Security Secretary stated, about folks at the US-Mexico border:
  "Again, let's just pause to think about this statistic: 314 percent increase in adults showing up with kids that are not a family unit... Those are traffickers, those are smugglers, that is MS-13, those are criminals, those are abusers."
- Nielsen was speaking about a relative increase in the probability of the event of "adults with kids who are not their own at the US-Mexico border". The relative increase is very large (314%). However, how often did the event happen in the first place?
- In a Washington Post analysis[1], the increase was from 0.19% in 2017 to 0.61% in 2018. Thus the actual chance of the event happening is very small and increased by 0.61%-0.19% = 0.42 percentage points.

Example. 1: Misleading statistics

- Takeaway: looking at the increase in absolute percentage points provides a different interpretation than the increase on the relative scale.

Reference:
https://www.washingtonpost.com/news/politics/wp/2018/06/18/how-to-mislead-with-statistics-dhs-secretary-nielsen-edition/?noredirect=on&utm_term=.9193534ee80c

Example 2: Calculating probabilities in medical settings can be difficult and not intuitive

- Suppose that there is test for a specific type of cancer that has a 90% chance of a positive screening test result for cancer if the individual truly has cancer and a 90% chance of testing negative for cancer when the individual does not have it.
- 1% of patients in the population have the cancer being tested for.
- What is the chance that a patient has cancer given that they test positive?

a) Between 0% - 24.9%
b) Between 25.0% - 49.9%
c) Between 50.0% - 74.9%
d) Between 75.0% - 100%

Example 2: Calculating probabilities in medical settings can be difficult and not intuitive

Many people choose 4), but the true answer is 1)! Why do we get this so wrong?

Video link (2 mins): click here

we will delve more into screening probabilities as we move forward

What is probability?

Fundamental components

Probability can be thought of in three fundamental components:

- A random experiment
- All possible outcomes of that experiment
- An event or events of interest

Sample space

We refer to the entire set of possible outcomes as the Sample Space

All possible outcomes of a sample space (S) together have a probability of 1.
$P(S) = 1$

Sample space

- Discrete sample space
  - e.g., Marital status: S = {Single, married, divorced, widowed}
  - careful, discrete spaces remind us of nominal, ordinal, or discrete variables
  - anything that is countable with "gaps" between the events
  - the notation is important: S = {elements in the space}
- Continuous sample space
  - e.g., The interval [0, 1]: S = {all numbers between 0 and 1}
  - continuous sample spaces remind us of continuous variables only
  - the events are not countable (i.e., we cannot list the numbers between 0 and 1, there are infinite)

Defining probability

The probability or occurrence of an event A often called the probability of A and denoted as P(A) is the ratio of the number of outcomes where event A occurs to the total number of possible outcomes.

For a coin what is the probability of heads?

- Probability model: Description of random phenomena. Consists of sample space S and a way of assigning probabilities to events

Frequentist probability

Frequentist definition

- Probability corresponds to frequency over many repetitions
- For example, the probability of a coin landing heads on a single toss is 0.5: if we toss a coin numerous times, we would expect one-half of the tosses to land heads, the more times we toss the coin the closer we expect the fraction of tosses to come to exactly half.

From B&M: How common is the common cold?

- Suppose that there are 100,000 people in your community. You work for your community's public health office and want to estimate the number of people who had a common cold. You are not able to sample everyone but suppose you could randomly call people in your community and ask them "Did you have a cold yesterday?" and then calculate the proportion of the sample who had a cold.

From B&M: How common is the common cold?

- Here are the dimensions, a data frame for the whole population, and the mean of the variable `had_cold_yesterday`:

```
dim(cold_data)
```

```
## [1] 100000      2
```

```
cold_data %>% summarize(population_mean = mean(had_cold_yesterday))
```

```
##   population_mean
## 1         0.11214
```

- Note that the mean of a 0/1 variable is called a proportion. This is because the mean is the number of individuals with a cold (coded as `had_cold_yesterday = 1`) divided by the total number of individuals.

From B&M page 216: How common is the common cold?

How big should your sample size be?

- We want to sample enough people such that the proportion of those with colds in the sample is close to the proportion of those with colds in the population
- Let's take samples of size 5, 100, and so on using dplyr's `sample_n` function:

```
sample_5 <- dplyr::sample_n(tbl = cold_data, size = 5)
sample_100 <- dplyr::sample_n(tbl = cold_data, size = 100)
sample_1000 <- dplyr::sample_n(tbl = cold_data, size = 1000)
sample_10000 <- dplyr::sample_n(tbl = cold_data, size = 10000)
sample_100000 <- dplyr::sample_n(tbl = cold_data, size = 100000)
```

Now estimate the proportion of those with a cold in the random samples

```
##   sample_mean_n5
## 1              0
```

```
##   sample_mean_n100
## 1             0.08
```

```
##   sample_mean_n1k
## 1           0.112
```

```
##   sample_mean_n10k
## 1           0.1091
```

```
##   sample_mean_n100k
## 1           0.11214
```

Estimate the proportion of those with a cold in the random samples

- What do you notice about the proportion estimates?
- Do they approach the true estimate as the sample size increases?

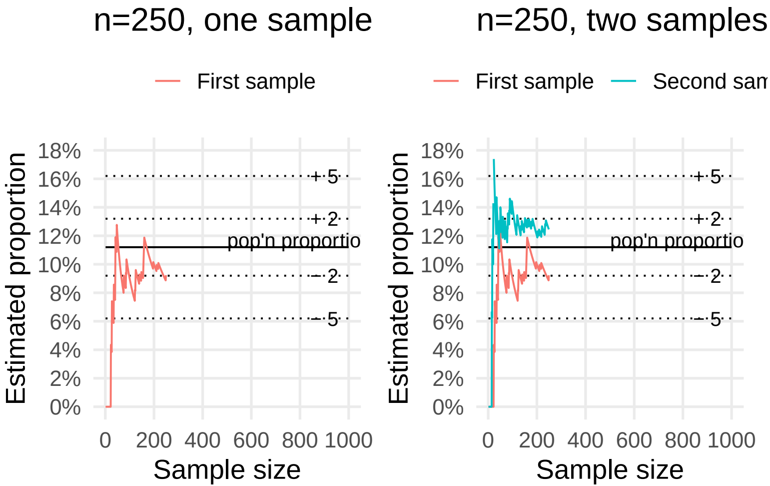
How many people should we sample?

- So we know we should sample more than five people, but feasibly can't sample everyone.
- We need to sample *enough* people to reasonably estimate the true chance of having a cold. But how many is enough?

Look at the sample's estimated proportion as a function of sample size

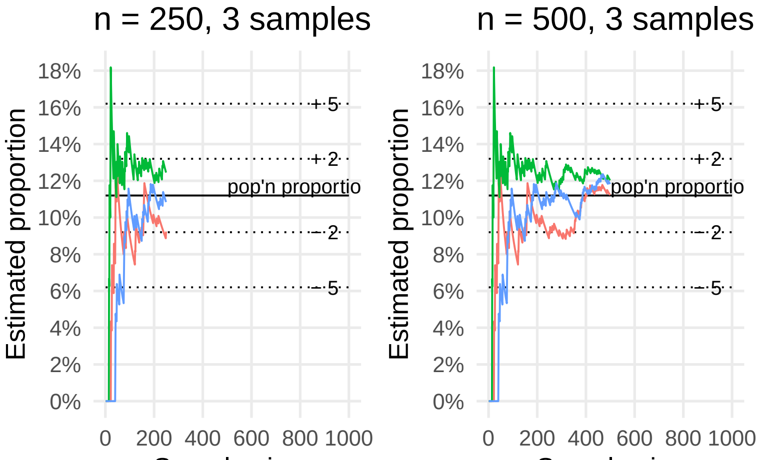
- To help us decide I first sampled one person and took the mean of that sample.
- Then I added another person and took the mean of that sample of size 2 (n=2). ... and so on, until I had 5000 people.
- The plot on the next slide shows the estimated proportion vs. the sample size.

Estimated proportion vs. sample size for n = 250

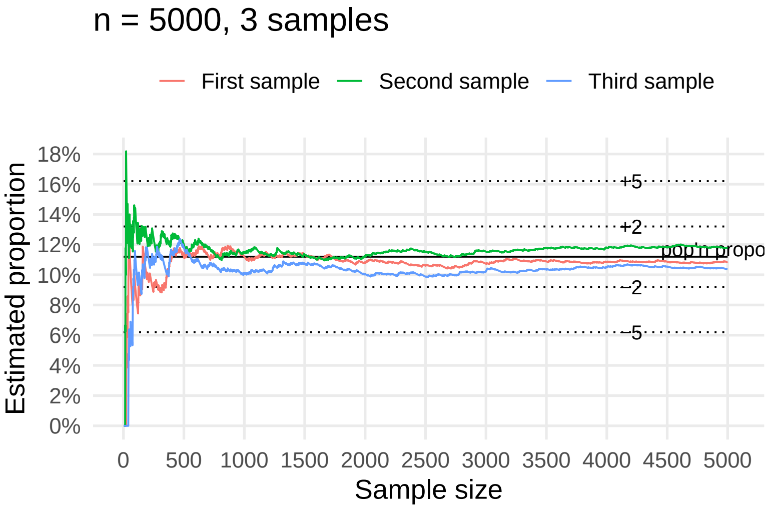


Estimated proportion vs. sample size for n = 500

- Increase the sample size how the estimate becomes closer to the true value
- Add in a third sample to compare how different samples perform in the short vs. the long run



Estimated proportion vs. sample size for n = 5000



Summary of the example on estimating the proportion with a cold

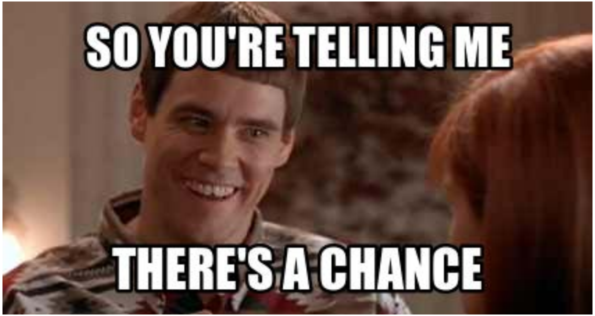
- As the sample size increases, the estimated proportion becomes closer to the true proportion
- Random samples of the same size will provide different estimates of the true proportion, but will be closer to each other (and the true value) if they are "large enough"

Some vocabulary, definitions, and rules

Rule of Range

Probabilities are numbers between 0 and 1.

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



What is the probability of a certain event? Of an impossible event?

Complement

The total number of outcomes in a random experiment (the sample space) can always divide into two mutually exclusive groups:

outcomes where A occurs P(A)

outcomes where A does not occur $P(\bar{A})$

These are called complementary events These must cover the entire set of possibilities which sums to 1

Thus $P(A) + P(\bar{A}) = 1$

Complement

The probability of the complement is 1 minus the probability of the event occurring.

- P(A does not occur) = 1 - P(A)
- Shorthand: $P(\bar{A}) = 1 - P(A)$ or $P(A^c) = 1 - P(A)$ or $P(A') = 1 - P(A)$

Composite Events

We stared by defining probability in terms of one event, but we can expand this to think about the probability of more than one event. For example, let A and B be two separate events. A composite event would then be the event which describes the outcomes of both A and B.

The composite event where both A and B occur is also referred to as the intersect of A and B.

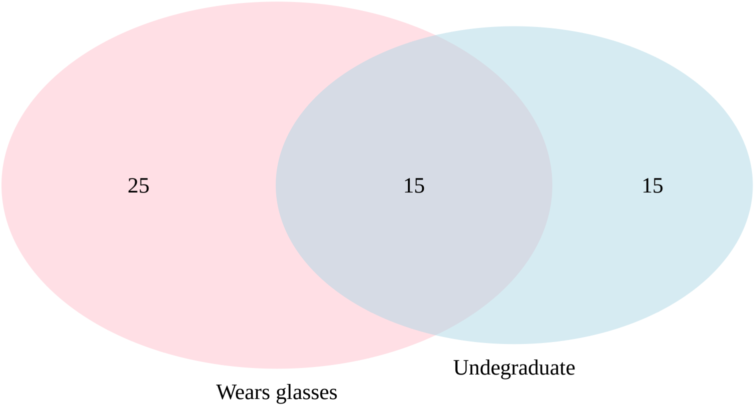
This is sometimes written as P(AB) or $P(A \cap B)$

The composite event where either A or B occurs is also referred to as the union of A and B.

This is sometimes written as P(A or B) or $P(A \cup B)$

Composite Events

Imagine we have 100 students in a classroom.



Composite Events

There are 40 students who wear glasses, and 30 who are undergraduates.

Based on the Venn diagram, what is P(Glasses)?

What is the complement $P(\overline{Glasses})$

What is the union of these two events $P(Glasses \cup Undergraduate)$

What is the intersect of these two events $P(Glasses \cap Undergraduate)$

Composite Events

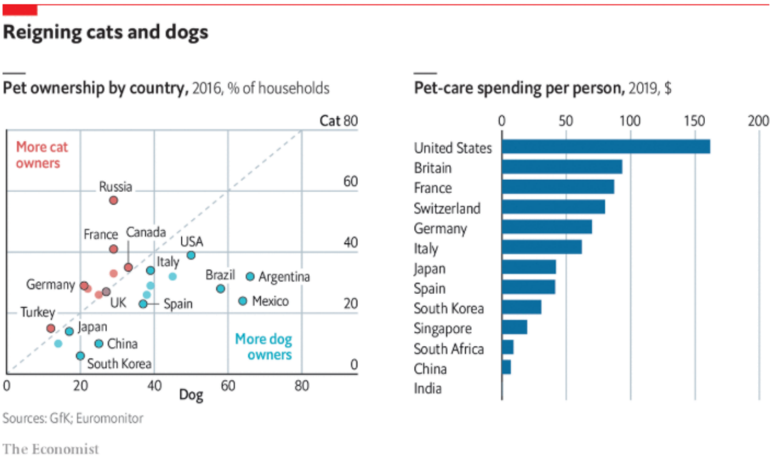
If we toss a coin twice, what are the possible composite events in our sample space?

What is the probability of tossing a combination of one Heads and one Tails?

What is the complement of that?

Probability of pet ownership

What is the probability space here?



from the economisit article here

Disjoint events

If two events have a joint probability of 0 (i.e., no overlap in their event spaces $P(A \cap B) = 0$) then they are disjoint and the probability of either event occurring is the summation of their individual probabilities.

$P(AorB) = P(A) + P(B)$, if A and B are disjoint events.

Disjoint events are also described as mutually exclusive meaning that it is not possible to have both events

Discrete probability models

Discrete probability model

- A probability model with a sample space made up of a list of individual outcomes is called discrete
- To assign probabilities in a discrete model, list the probabilities of all the individual outcomes. These probabilities must be numbers between 0 and 1 and must sum to 1. The probability of any event is the sum of the probabilities of the outcomes making up the event.

Discrete probability model example

For example, we could survey a sample of people and ask them their marital status. Based on this survey we can calculate the portion of each event in the sample space:

Single	Married	Divorced	Widowed
47%	30%	18%	5%

This is a discrete probability model shown in a table. How else could you display these data?

Continuous probability model

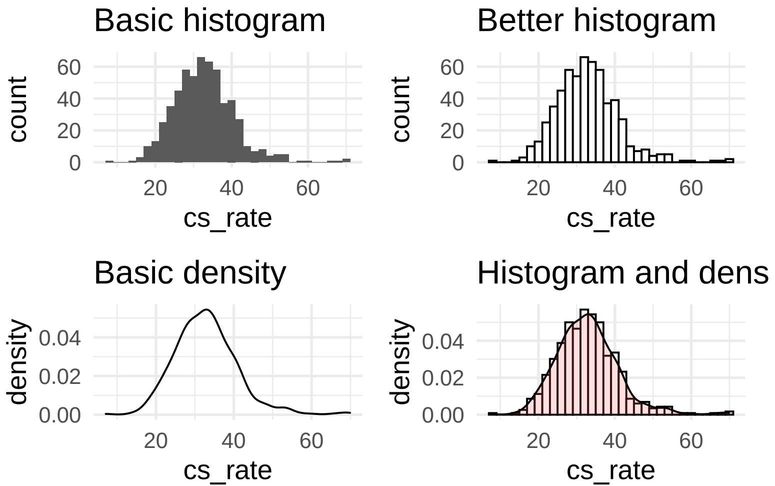
- A continuous probability model assigns probabilities as areas under a density curve. The area under the curve and between a range of specified values on the horizontal axis is the probability of an outcome in that range.
- What is a density curve?

Density curves

- Density curves are also known as probability density functions
- You can think of density curves as smoothed histograms.

Density curves using `geom_density()`

- Recall the data on cesarean delivery rates across hospitals in the US. We can use these data to also make a density plot (also called density curve):



Density curves

- From this plot, we can see that the density curve approximates the shape of the histogram very well.
- Remember because there are infinitely many cesarean delivery rates that could be observed between 0 and 1 it is impossible to assign a finite probability to any specific number.
- If we did, we could do this infinitely and their summed probability would surpass 100%.
- Instead, the density curve is used to determine the probability of an observed event within a specific range.

Density curves

You could use the density curve to calculate:

- $P(CS < 0.20)$
- $P(0.20 < CS < 0.40)$
- $P(CS < 0.2 \text{ or } CS > 0.4) = P(CS < 0.2) + P(CS > 0.4)$ because this events are independent
- $P(CS > 0.4) = 1 - P(CS \leq 0.4)$
- The calculations can be interpreted as either:
  - the proportion of hospitals with cesarean delivery rates in the specified range
  - the probability that a randomly chosen hospital will have cesarean delivery rate in the specified range.

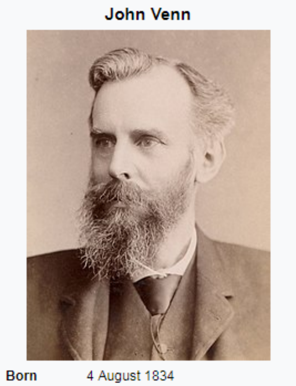
Random variables

- A random variable is a variable whose value is a numerical outcome of a random phenomenon
- Random variables are represented by capital letters, most popularly X.
- A lower case letter represents a particular value for the random variable has been taken. For example $P(X = x)$ asks, what is the probability that random variable X takes the value x?
- For continuous random variables, we ask $P(X < x)$ for example, because $P(X = x) = 0$ for continuous random variables.
  - Why is this the case?

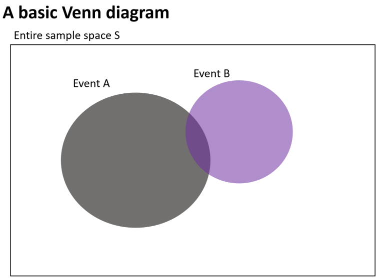
Venn diagrams

Venn diagrams

Popularized in a paper in 1880



Venn diagrams

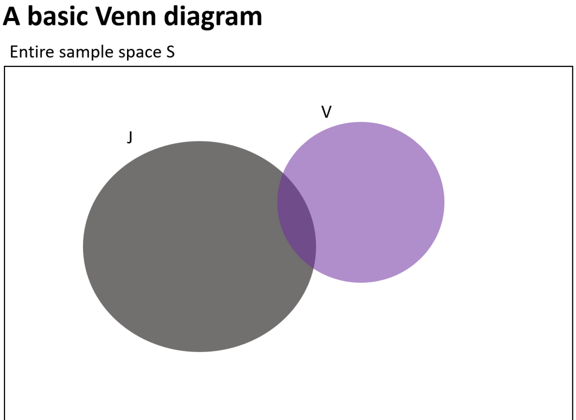


- Venn diagrams generally consist of at least two events
- When we draw a rectangle around the diagram this is the sample space. All the probabilities inside must sum to 1.
- When you make a Venn diagram label the events using capital letters to represent the random variables

Venn diagrams

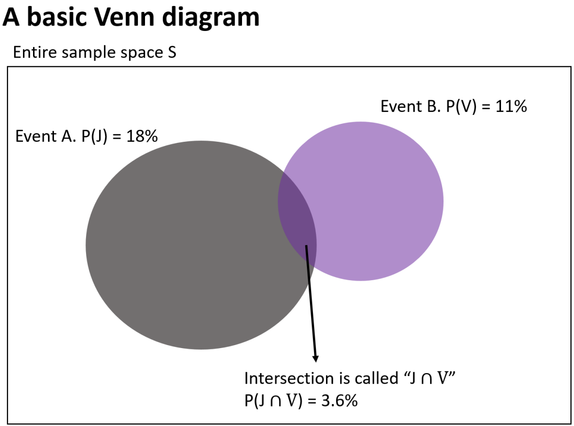
Suppose you had access to survey data about vaping and JUUL-related advertisements.

You could define the following random variables: J is the event "seen ad for JUUL" V is the event "vaped in the last 30 days"



Venn diagrams

Now let's say we do a survey - we can add the percents to the diagram:

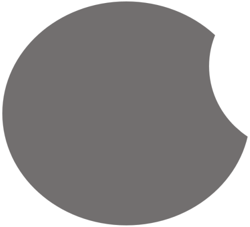


what percent of the individuals saw an ad for JUUL and do not vape?

Venn diagram

What percent of individuals saw an ad for JUUL and do not vape?

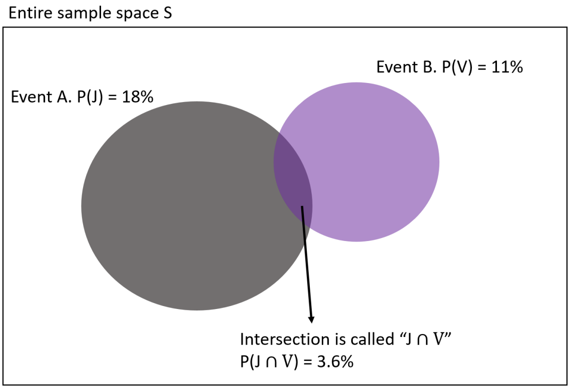
This percent is represented by this area:



How would we write this as a probability statement? What percentage is this?

Venn diagram

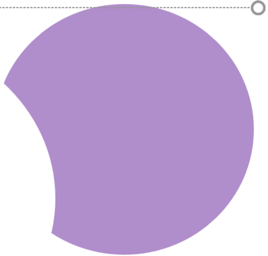
A basic Venn diagram



What percent of individuals vape but have not seen an ad for JUUL?

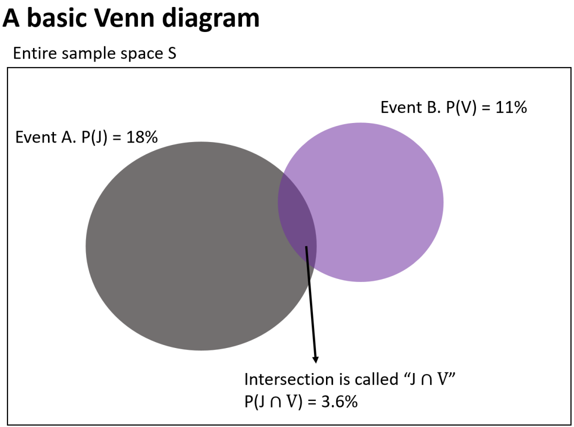
Venn diagram

What percent of individuals vape but have not seen an ad for JUUL?



How would we write this as a probability statment? What percentage is this?

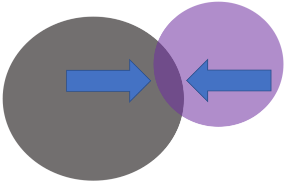
Venn diagram



What percent of individuals vape AND have seen an ad for JUUL? What is this called?

Venn diagram

What percent of individuals have seen an ad for JUUL and vape?

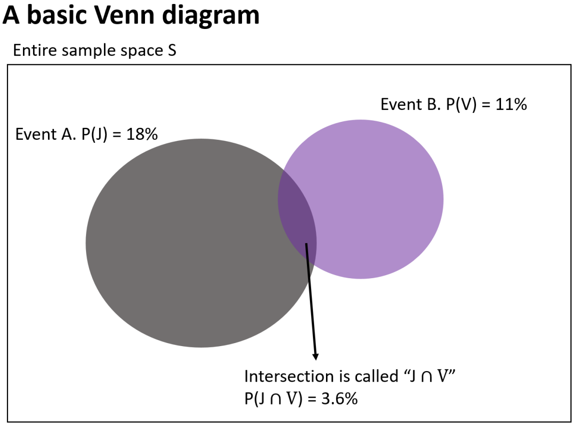


The "and" represents the intersection, the area where both events have occurred:

$P(J \cap V) = 3.6\%$

3.6% of individuals have seen an ad for JUUL and have vaped in past 30 days

Venn diagram



What percent of individuals vape OR seen an ad for JUUL? What is this called?

Venn diagram

The union is all events that have either of the outcomes - all the shaded parts of this diagram are part of the union.

P(J or V)
$P(J \cup V)$
$= P(J) + P(V) - P(J \cap V)$ this is the general rule for addition
$= 0.18 + 0.11 – 0.036$
$= 0.256 = 25.6\%$

R recap:

The only new code in this lecture you need to remember is

```
geom_density()
```

Which is the function we use to generate a density plot within the ggplot package

Comic Relief

From xkcd.com