MSDS 660 Week 2: Probability Theory and Data Exploration in R

REGIS UNIVERSITY

Recap

- Scales of Measurement
- Statistical Process
- Sampling

Scales of Measurement

Qualitative data –
Non- numeric

- Nominal data, Categories
  - No order
    - Customer ID, zip code, etc.
- Ordinal Data
  - Ordered but no scale
    - Likert scale

Quantitative data – Numeric or Boolean

- Interval data
  - The interval between observations is expressed in terms of a fixed unit of measure.
    - Temperature in Celsius or Fahrenheit, SAT and ACT scores
- Ratio data
  - There is a true zero on a ratio scale and equal intervals between points.
    - Car speed, monetary values, temperature in Kelvin, etc.

Week 2 Content

- Discuss the Measures of Central Tendency and Measures of Variability
- Understand Probability Theory and how it relates to statistics
- Compute Events and their probabilities
- Discuss Random Variables and compute expected value and variance

Measures of Central Tendency

Descriptive Statistics

- help us organize and summarize the data so it's easier to understand
- measures of central tendency tell us about **location** on the normal distribution
    - **Mean**: the average of values in any given data set
    - **Median**: the middle score in a distribution when you order your values.
    - **Mode**: the score that occurs with the greatest frequency in any given data set

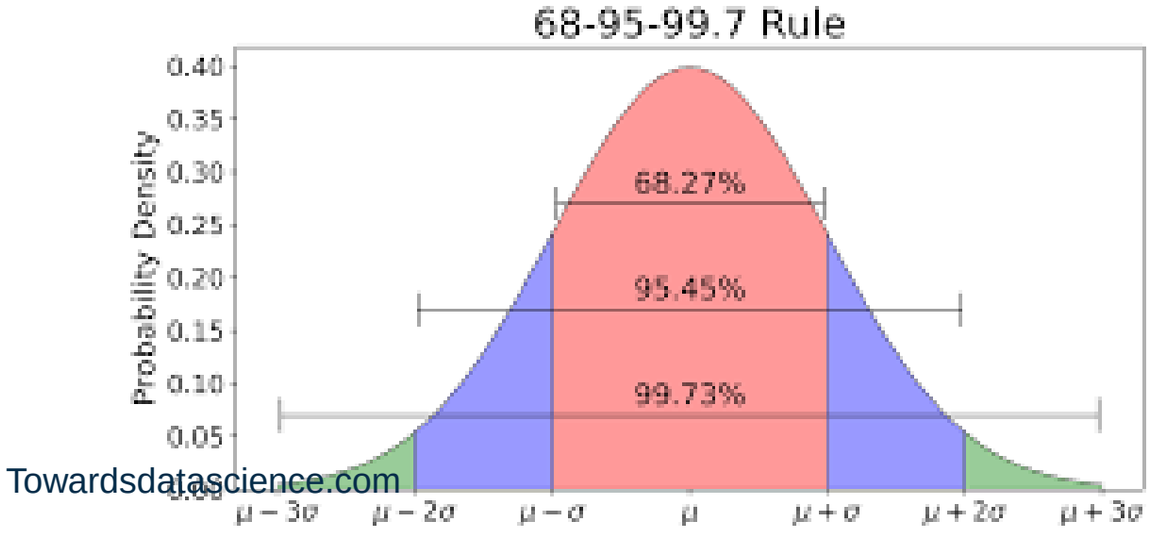
Examples?

Inferential Statistics

- allow us to draw conclusions about a population based on a sample taken from that population.
- p-values, correlation, regression, hypotheses testing

Measures of Central Tendency: Mean

- Sample means ($\overline{X}$) are used to estimate population means, symbolized by mu (μ).
- Mean is a good measure for a Gaussian or t-distribution
- In a perfectly normal curve, the bell-shaped curve, mean, median, and mode are all in the exact same place.



- Why would I ever use a median if I can use the mean?

Measures of Variability

Characterize the spread or variability of data.

- **Range**
  - the difference between the highest and lowest scores in a distribution.
- **Variance**
  - the average squared distance from the mean.
- **Standard deviation**
  - How far the data are distributed from the mean. It’s the square root of the variance.
- **Interquartile range (IQR)**
  - measures the variability of points near the center, where (50%) of the data are located.

Probability

Probability of an outcome is the proportion of times the outcome would occur if we observed the random process an infinite number of times.

Probability values are always assigned on a scale from 0 to 1.

A probability near zero indicates an event is quite unlikely to occur.

A probability near one indicates an event is almost certain to occur.

Basic Relationships of Probability

Basic probability relationships are useful to compute the probability of an event without knowledge of all the sample point probabilities.

1. Addition Rule
2. Complement of an event
3. Multiplication Rule for Independent Events
4. Union of two events
5. Intersection of two events
6. Mutually exclusive events

General Addition Rule

The general addition rule provides a way to compute the probability of event *A*, or *B*, or both *A* and *B* occurring.

The law is written as:

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

Where the P (*A* and *B*) is the probability that both events occur.

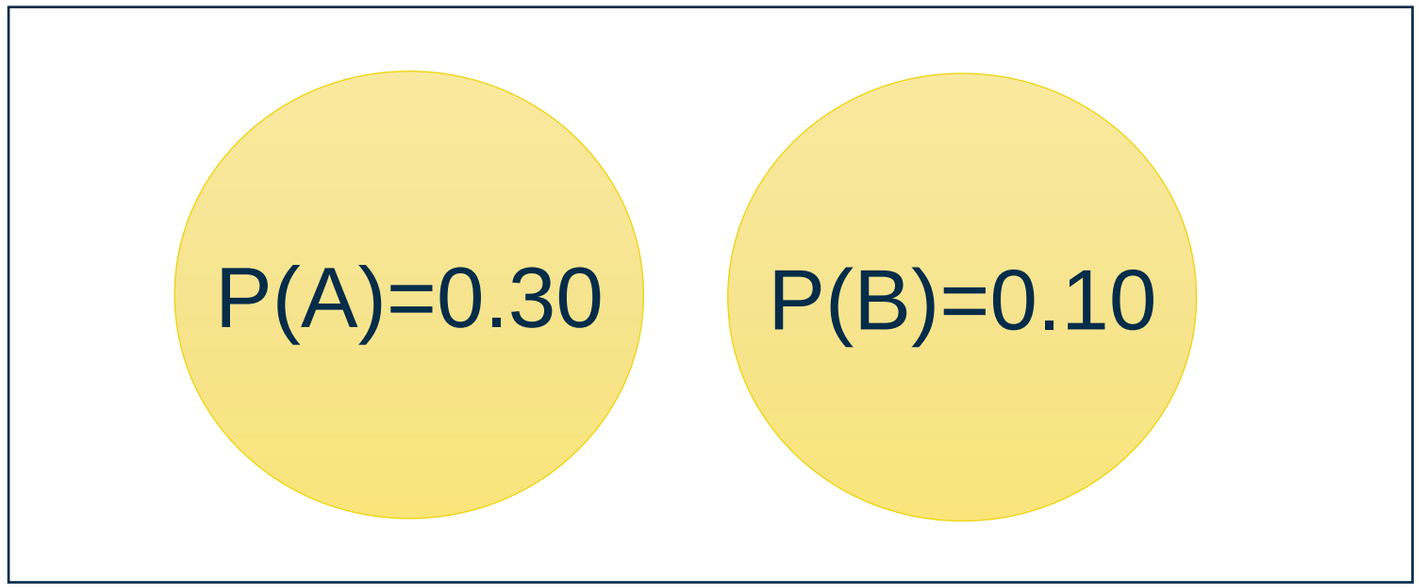
Complement of an Event

The complement of event *A* is defined to be the event consisting of all sample points that are not in *A.*

The complement of *A* is denoted by A^c.

$$P(A) + P(A^c) = 1, \qquad \text{i.e} \qquad P(A) = 1 - P(A^c)$$

Q. What is $P(A^c)$?



1. .70
2. .60
3. .40
4. 1.00

Multiplication Rule for Independent Events

If the probability of event *A* is not changed by the existence of event *B*, we would say that events *A* and *B* are <u>independent</u>.

Example: Rolling 2 six-sided dice.

If *A* and *B* represent events from two different and independent processes, the probability that **both** *A* and *B* occur can be calculated as the product of their separate probabilities:

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

Question

About 9% of people are left-handed. Suppose 2 people are selected at random from the U.S. population. Because the sample size of 2 is very small relative to the population, it is reasonable to assume that these two people are independent.

What is the probability that both are left-handed?

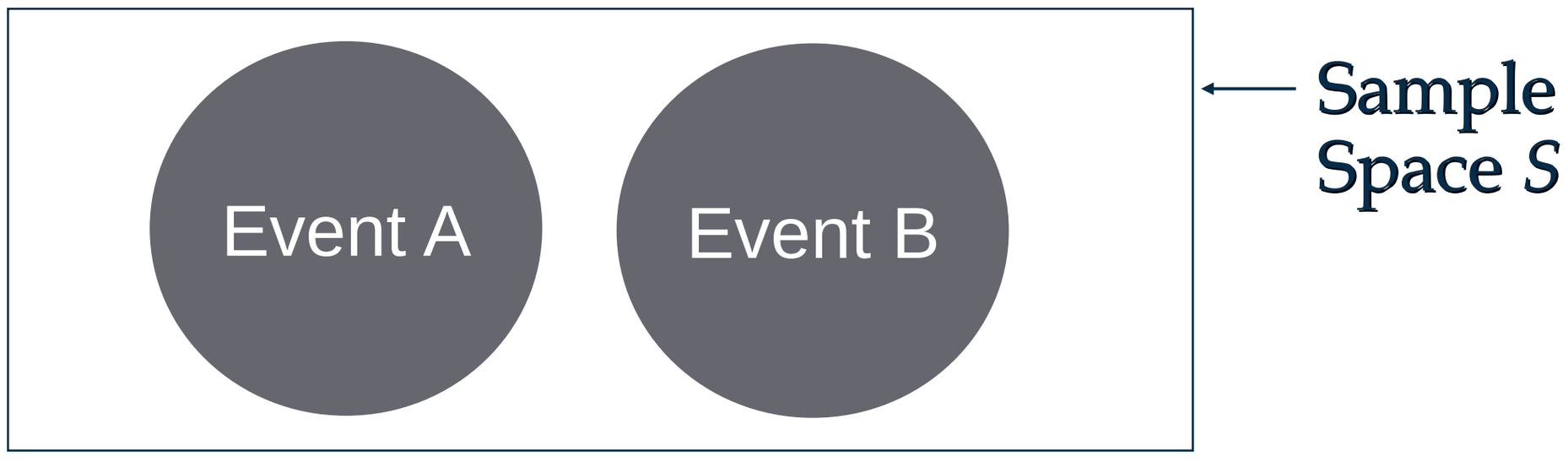
1. .82
2. .0.0081
3. 0.09

OpenIntro, Guided Practice 3.22, pg. 90

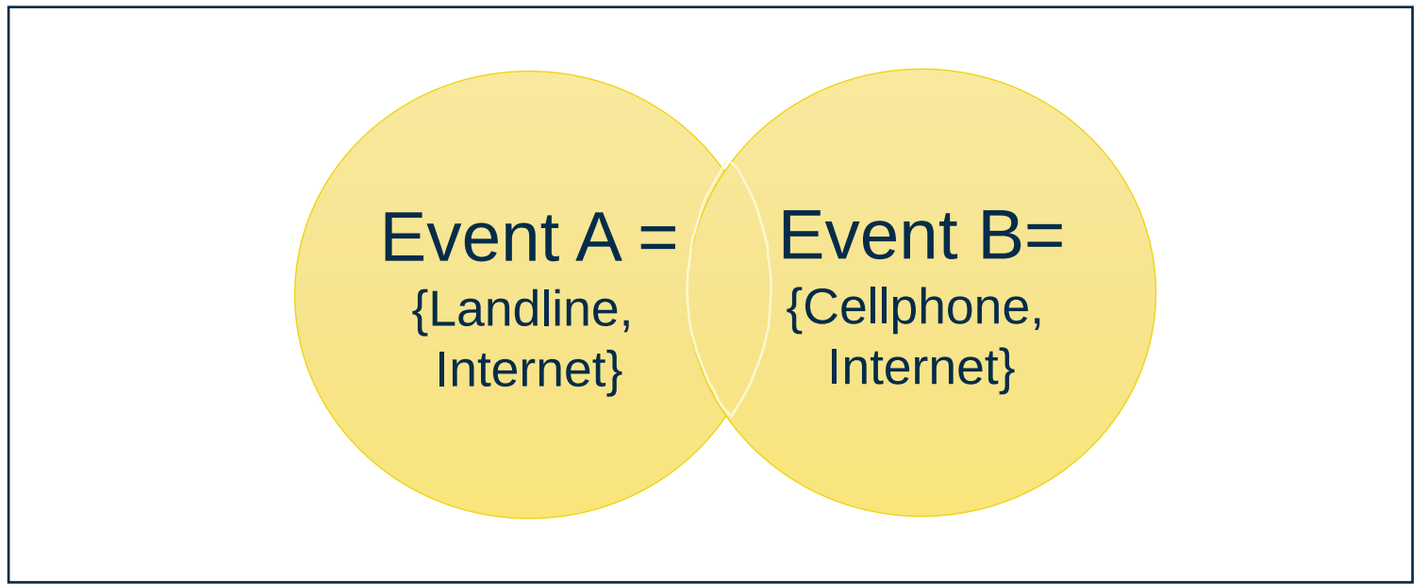
Union of Two Events

The union of events *A* and *B* is the event containing all sample points that are in *A* or *B* or both.

The union of events *A* and *B* is denoted by $A \cup B$.



Q. What is AUB?

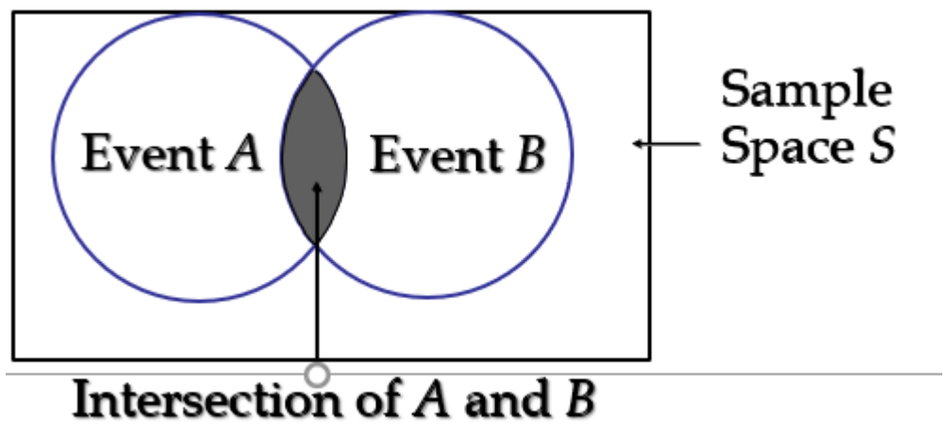


1. {Internet}
2. {Landline, Internet, Cellphone}
3. {Landline, Cellphone}

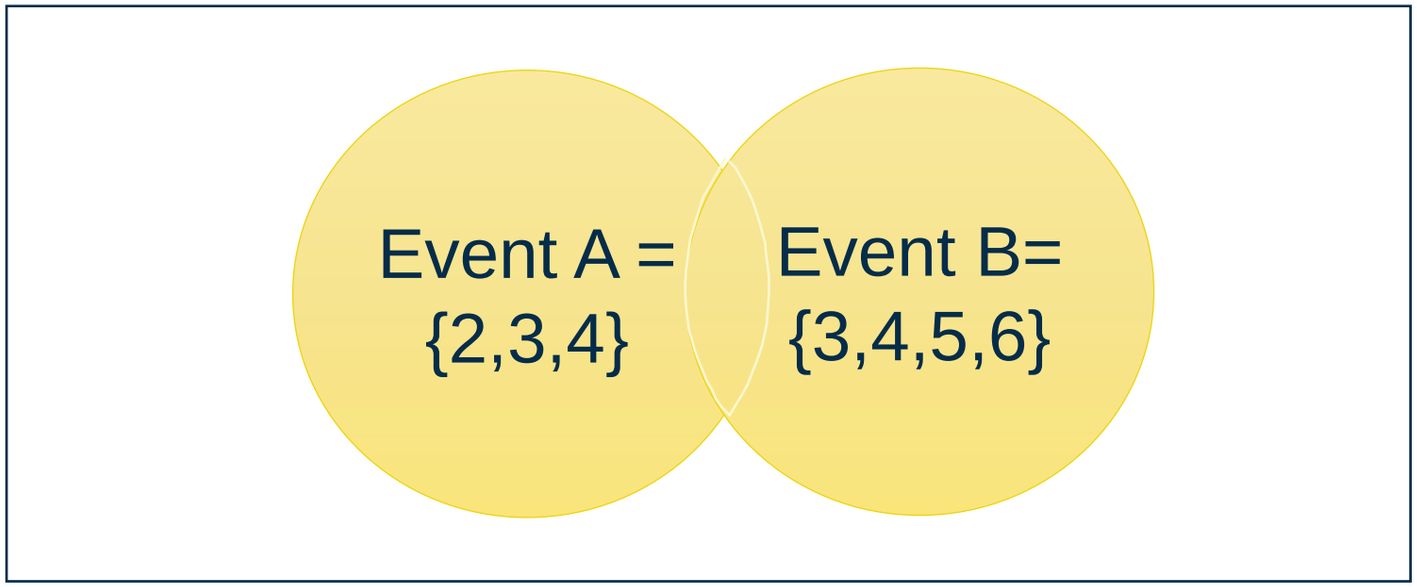
Intersection of Two Events

The intersection of events *A* and *B* is the set of all sample points that are in both *A* and *B at the same time*.

The intersection of events *A* and *B* is denoted by $A \cap B$.



Q. What is A ∩ B?

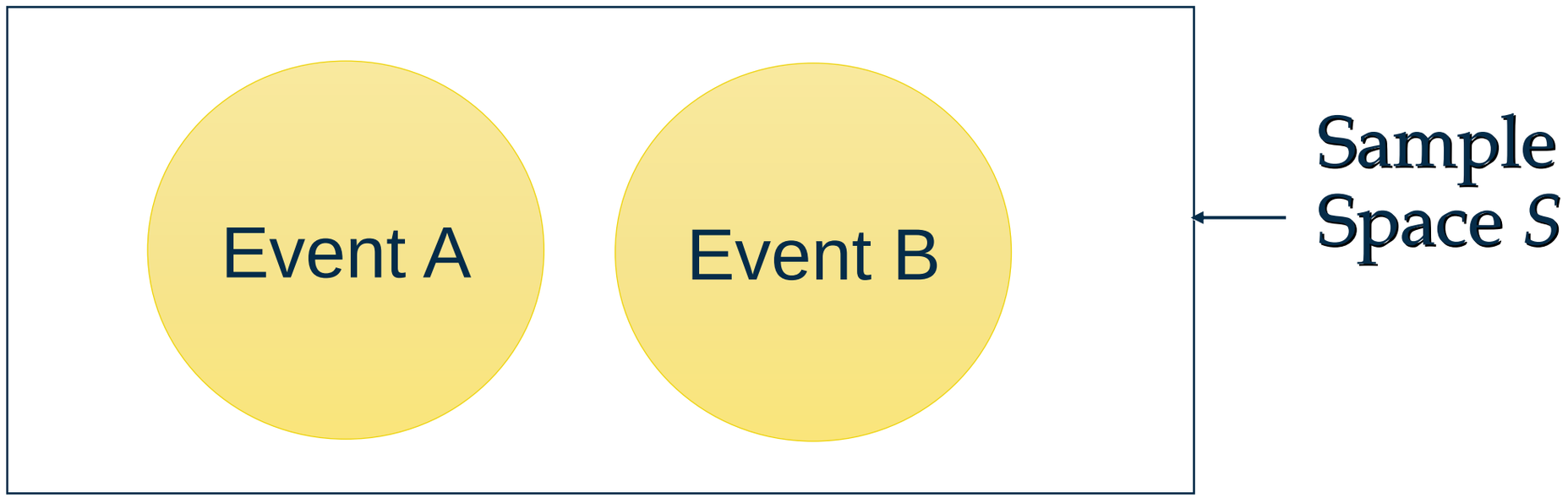


1. {2,3,4,5,6}
2. {3,4,5,6}
3. {3,4}

Mutually Exclusive Events

Two events are said to be mutually exclusive if the events have no sample points in common.

Two events are mutually exclusive if, when one event occurs, the other cannot occur.



Mutually Exclusive Events

If events *A* and *B* are mutually exclusive,

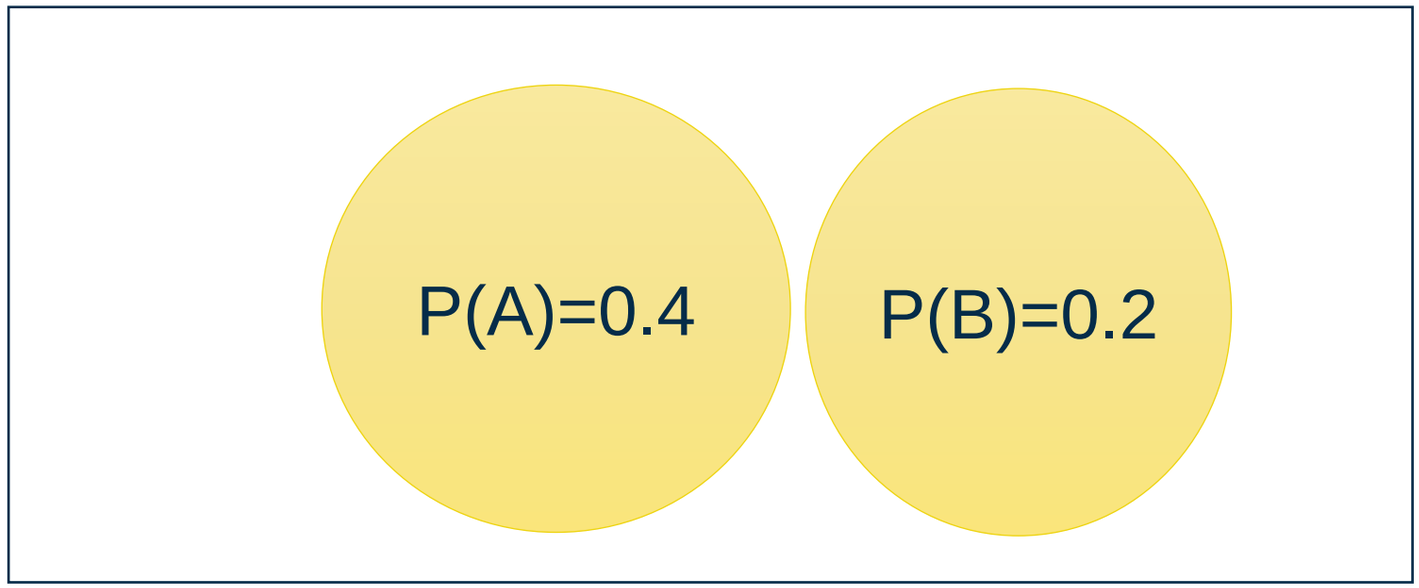
$$P(A \cap B) = 0.$$

The addition law for mutually exclusive events is:

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

There is no need to include "$- P(A \cap B)$"

Q. What is P(AUB)?



1. .4 + .2
2. .4
3. .2
4. .4 + .2 - .1

Conditional Probability

The probability of an event given that another event has occurred is called a conditional probability.

The conditional probability of *A* given *B* is denoted by $P(A|B)$.

A conditional probability is computed as follows :

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

Conditional Probability

Contingency table summarizing the photo_classify data set

		truth		
		fashion	not	Total
mach_learn	pred_fashion	197	22	219
	pred_not	112	1491	1603
	Total	309	1513	1822

This data set a sample of 1822 photos from a photo sharing website. A Machine Learning (ML) classifier serves as a test for these data. Each photo receives 2 classifications based on whether ML classified a photo is about fashion or not. The photos have also been classified by humans which constitutes the true nature of the photos.

OpenIntro: Figure 3.11. pg. 95

Conditional Probability

P (pred_fashion | true_fashion) or if a photo is actually about fashion, what is the chance the ML Classifier correctly identified the photo as being about fashion?

The ML algorithm correctly classified 197 of the 309 photos that were actually about fashion.

So, P (pred_fashion | true_fashion) = 197/309=0.64

OR you can do it using $$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$$

P (pred_fashion | true_fashion) = $\frac{(197/1822) = \ 197}{(309/1822) \quad 309} = .64$

Question

If the ML classifier suggests a photo is **not** about fashion, what is the probability that it was incorrect and the photo is about fashion?

1. 112/309
2. ✓ 112/1603
3. 197/219

Discrete Probability Distributions

- The probability distribution for a random variable describes how probabilities are distributed over the values of the random variable.
- We can describe a discrete probability distribution with a table, graph, or formula.
- The probability distribution is defined by a probability function, denoted by $f(x)$, which provides the probability for each value of the random variable.
- Requirements: $1 \geq f(x) \geq 0$ and $\Sigma f(x) = 1$
- A random variable is a numerical description of the outcome of an experiment.

Example

Using past data on TV sales, tabular representation of the probability distribution for TV sales was developed. We can calculate our probability function which will give the probability for each value of the random variable. The number of TV units sold to a customer is a random variable, and we represent it by *X*.

Units Sold	Number of Days
0	80
1	50
2	40
3	10
4	20
	200

x	$f(x)$
0	.40
1	.25
2	.20
3	.05
4	.10
	1.00

80/200

What is the value of *f(x)* associated with *x* = 1?

Expected Value and Variance of a Random Variable *X*

The expected value, or mean, of a random variable is a measure of its central location:

$$E(x) = \mu = \Sigma x f(x)$$

$E(x)$ is the most widely used measure to evaluate the expected return of an asset or a project or the cost of a part.

The variance summarizes the variability in the values of a random variable:

$$\text{Var}(x) = \sigma^2 = \Sigma(x - \mu)^2 f(x)$$

The standard deviation, σ, is defined as the positive square root of the variance.

Calculation of Expected value and Variance of *X*

Number of Units Sold	Number of of Days	x	$f(x)$	$xf(x)$	$x - \mu$	$(x - \mu)^2$	$(x - \mu)^2 f(x)$
0	80	0	.40	.00	-1.2	1.44	.576
1	50	1	.25	.25	-0.2	0.04	.010
2	40	2	.20	.40	0.8	0.64	.128
3	10	3	.05	.15	1.8	3.24	.162
4	20	4	.10	.40	2.8	7.84	.784
	200		$E(x) =$	1.20	$Var(x) = \sigma^2 =$		1.660

Demo

- Download Lab_week2.rmd and follow along

Discussion Activity

- Work as a group. Continue working to look for any trends in bird/wildlife data set (1995- March 2022). You can pick a different time frame and parameters if you are interested using the faa's query tool (https://wildlife.faa.gov/search).
- Create a bar plot showing a relationship between year and number of birds struck
- Create a plot with ggplot() showing the extent of damage or aircraft by bird collisions
- Use a subset of data where the incident occurred after 2015 and compute the means and SDs of distance by the number of birds struck. The function tapply() is useful for this.
- Create a final plot of bird collisions by engine type and height (feet above ground level). You'll need to revalue engine type and do some class conversions to create a good-looking plot. Be sure to include labels and title.
- Post to the discussion thread:

1. Any problems/concerns with data.
2. Plots that you think may be relevant to support your assertions.
3. A one paragraph summary of your analysis process and interpretation of results.

Assignment

Use the *modified* marketing data set that you had manipulated from last week **or** a data set of your choice. You may find it helpful to use the same dataset each week to apply concepts we are learning and building on your knowledge base.

- Explore the relationships between variables that might be correlated.
- A boxplot and histogram plot(s) (others?) of the data.
- A brief explanation ***1 paragraph !*** of your analysis process and your interpretation of the data. (What is the data, what did you do with the data, what do the results mean?)
- Your Rmd file and knitted pdf file