Probability & Statistics: A Brief Overview

PHS SummR Camp
Unnati Mehta
Fall 2020

Things to keep in mind:

- You don't have to remember everything we talk about number for number, word for word
- This is going to be an interactive session
- A lot of these concepts serve as the foundation for what we'll be covering later in the course, so this is a warm-up and a reference
- We are all here to support you however we can!
- This is a challenging time, and we are learning what works and doesn't work right along with you, so please don't hesitate to give feedback

Our goals today:

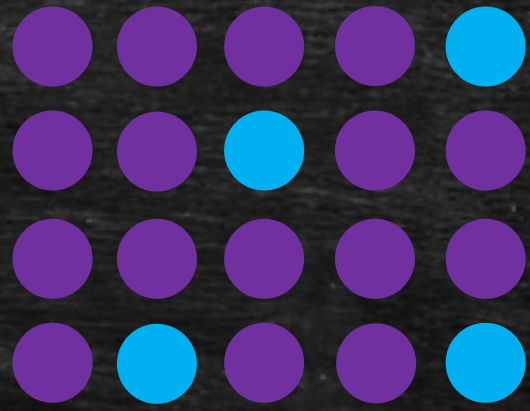
- What is probability?
- Set notation
- Random variables
- Probability distributions
- Introduction to statistics
- Why and do we use statistical concepts?
- Putting it all in context!

What is probability?

http://www.PollEv.com/unnatimehta240

What is probability?

- It is the relative likelihood that a given outcome or event will occur
- Ranges from **0** to **1**
- What is the probability that we would randomly select a blue dot?
- How can we use probabilities?
  - Start with a **population**
  - Understand the probability **distribution**
  - Observe the **parameter** of choice



4/20 = 1/5 = 0.2

1 – 0.2 = 0.8

Set Notation

- In minion world, there are two hairstyles (long and short) and there are three possible colors (green, red, and blue)
- The sample space Ω consists of all possible outcomes of a particular trial
- We can also subset our sample space Ω to include only certain elements

G ⊂ Ω → this means that **G** is a "subset" of Ω

(this is our sample space!)

(this is our subset!)

Probability and Sets

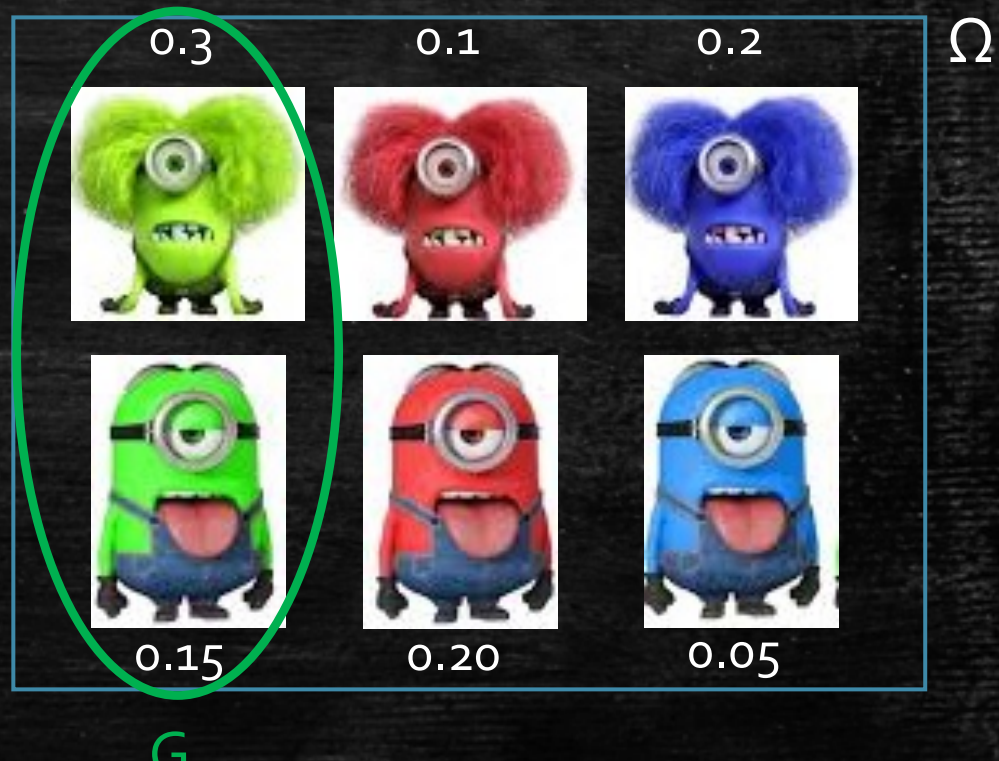
The function P(*) denotes the probability of an event * occurring, where event * is a subset of the sample space (Ω). The output is a real number between 0 and 1.

We define P() so that the probability of seeing each type of minion is shown in the diagram.

$P(\Omega) = 1$

P() =

P(G) =



Probability and Sets

Operations

$A \cup B$

- The union of A and B
- A set of elements contained in A or B

$A \cap B$

- The intersection of A and B
- A set of elements contained in A and B

Statements

$A \subset B$

- A is a subset of B (in this case, it's not)

A and B are mutually exclusive

- There is no overlap in elements between A and B

Probability and Sets

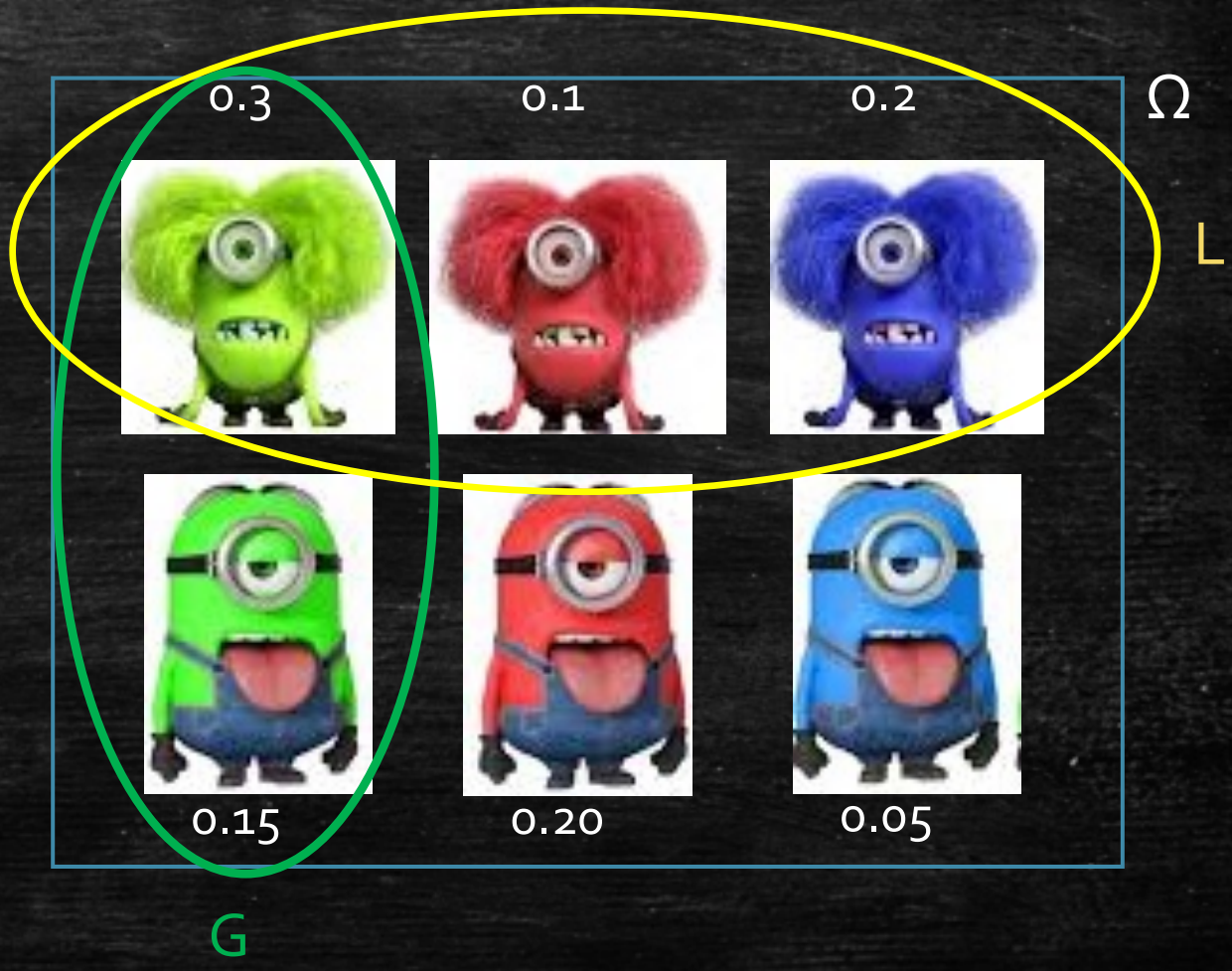
Quick tip:
∪ → "or"
∩ → "and"

$P(G) =$

$P(L) =$

$P(G \cup L) =$

$P(G \cap L) =$

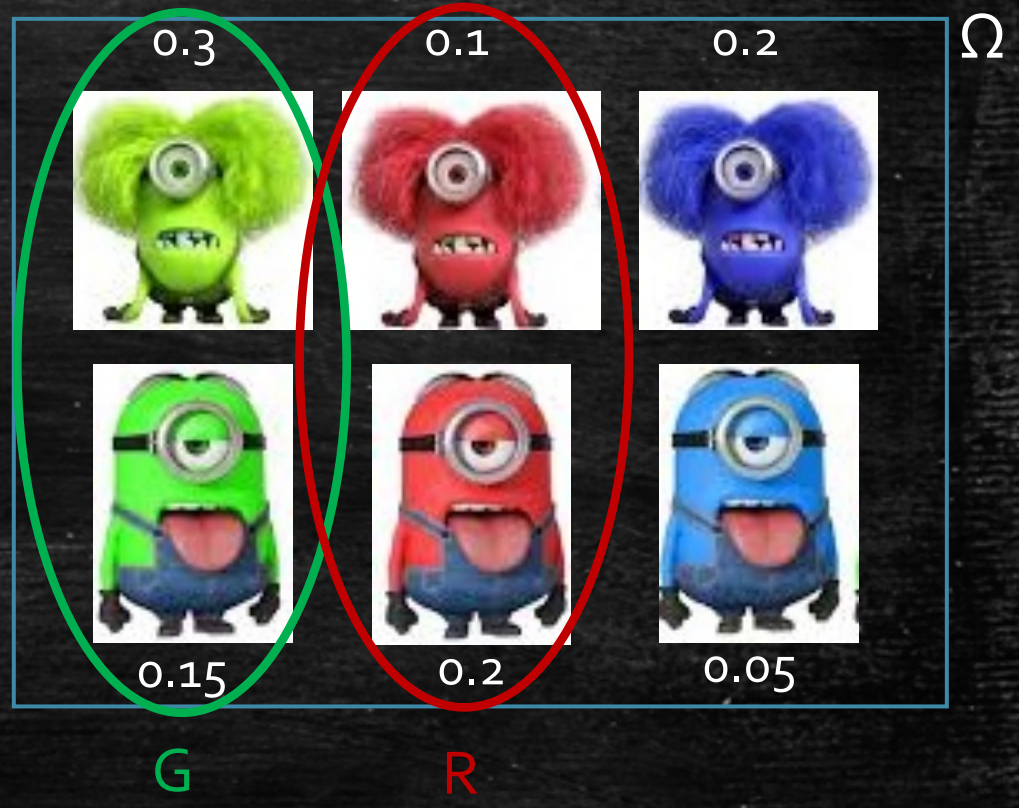


Mutual Exclusivity

- G and R are mutually exclusive if no element in G is also in R, and no element in R is also an element of G, i.e.

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

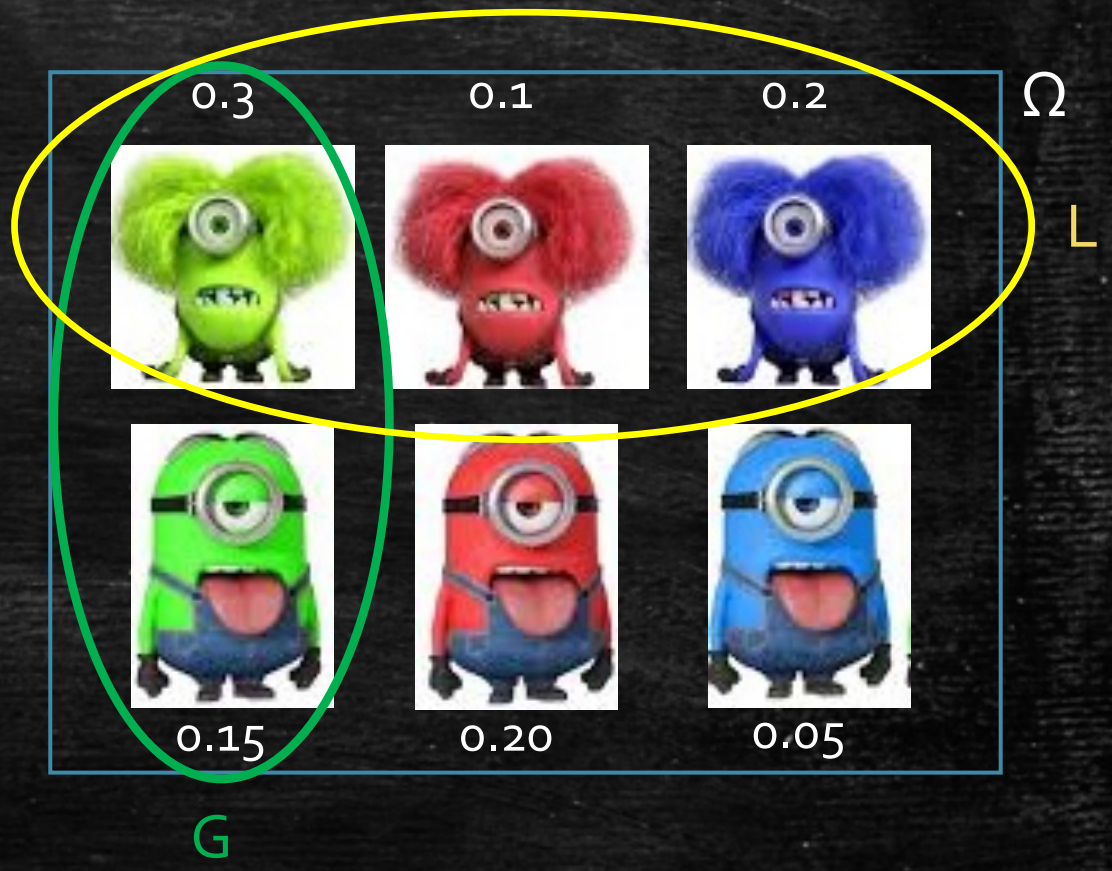
- In this case, if G ⊂ Ω and R ⊂ Ω, and G and R are mutually exclusive, then:
  - P(Ω) = 1
  - P(G) ≥ 0, P(R) ≥ 0
  - P(G ∪ R) = P(G) + P(R) =



Conditional Probability

- We can also assess probability within subsets of the sample space.
- Conditional probability is defined as the probability that an event will occur *given* that another event has already occurred.

$$P(G|L) = \frac{P(G \cap L)}{P(L)} =$$



Decomposition

We can use a trick to calculate the probability of event Z

$$P(Z)$$

$$= P(G \cap Z) + P(R \cap Z) + P(B \cap Z)$$

$$=$$

Decomposition

Let Ω be the sample space. Say $A_1, A_2 \ldots A_n$ are mutually exclusive, and $A_1 \cup A_2 \cup \ldots \cup A_n = \Omega$

Then:

$$P(B) = \sum_{i=1}^{n} P(A_i \cap B)$$

Now, let's say we want to calculate $P(A_j|B)$…

Bayes' Theorem

Recall the definition of conditional probability:

$$P(A_j|B) = \frac{P(A_j \cap B)}{P(B)}$$

We can rewrite the numerator of this equation as:

$$P(A_j \cap B) = P(B|A_j)P(A_j)$$

Recall the definition of decomposition:

$$P(B) = \sum_{i=1}^{n} P(A_i \cap B) \rightarrow P(A_i \cap B) = P(B|A_i)P(A)$$

(this goes back to rules of conditional probability)

So we get:

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}$$

Bayes' Theorem

Let Ω be the sample space. Say $A_1, A_2 \ldots A_n$ are mutually exclusive; and $A_1 \cup A_2 \cup \ldots \cup A_n = \Omega$; and $P(B) > 0$

Then:

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}$$

Can anyone think of a context in which Bayes' Theorem is commonly used/can be used?

Random Variables

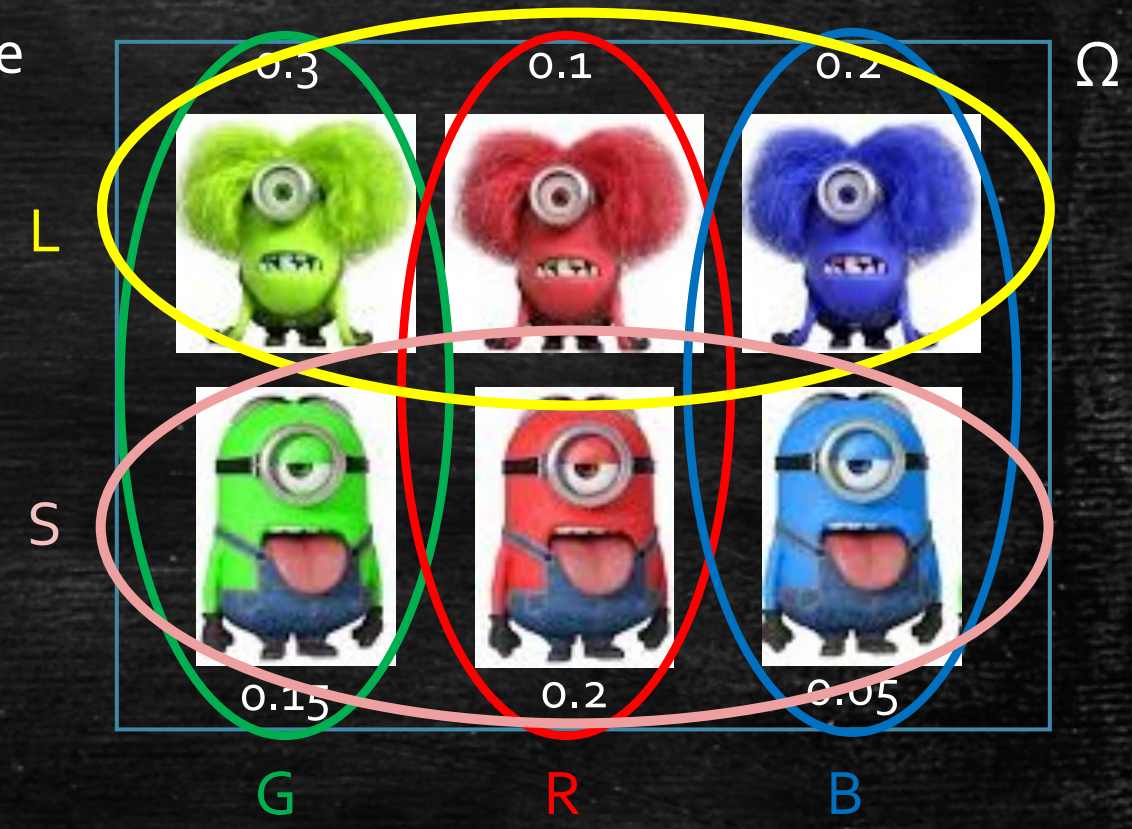
Random Variables

- A random variable can be written as a function that delineates subsets of Ω on the real line, $\mathbb{R}$
- Let's define X as color of minion:

$$X = \begin{cases} 1 \; if \; G \\ 2 \; if \; R \\ 3 \; if \; B \end{cases}$$

- Let's define Y as hair type of minion:

$$Y = \begin{cases} 1 \; if \; L \\ 2 \; if \; S \end{cases}$$



Random Variables

- **Discrete** random variables can take on separate values
  - Refers to binary, categorical, or count variables
  - Example → Colors ($X \in \{1,2,3\}$)
  - Example → Smoking ($X \in \{0,1\}$)

- **Continuous** random variables can take on any value in an interval
  - Example → Test scores ($X \in (0,100)$)
  - Example → Annual income ($X \in (0, \mathbb{R})$)

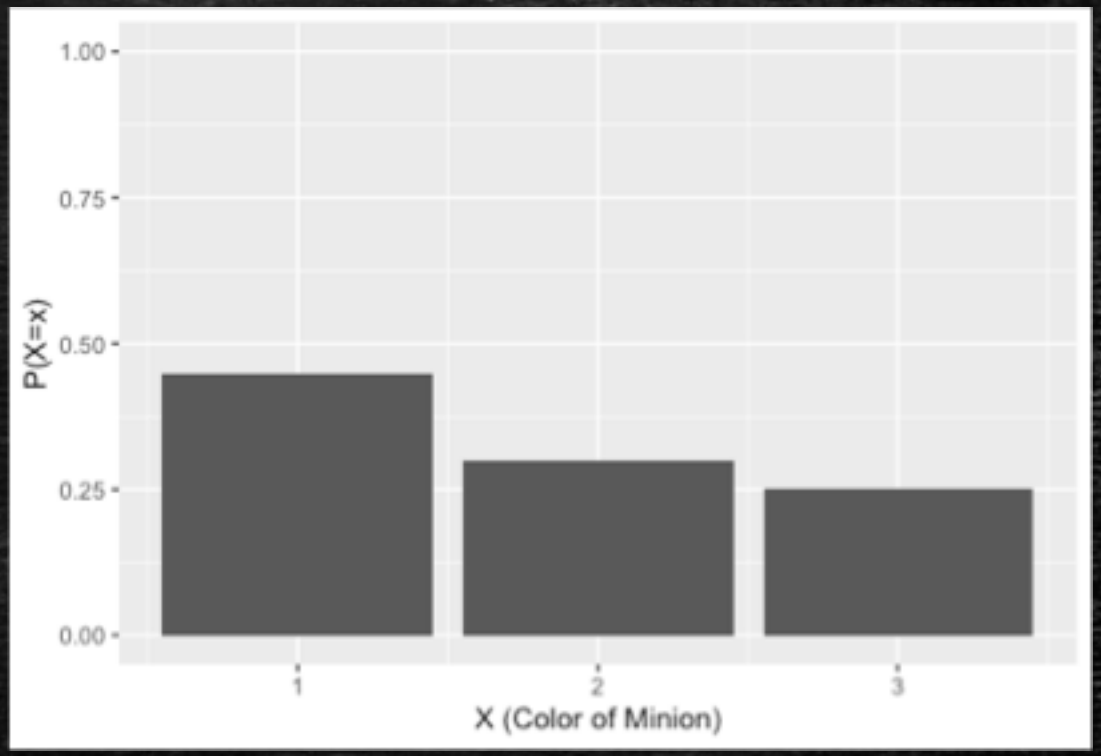
- Random variables help to simplify probability statements
  - P(G) can be written as P(X=1)
  - P(R) can be written as P(X=2)
  - P(L) can be written as P(Y=1)
  - $P(G \cap L)$ can be written as P(X=1, Y=1)

Probability Mass Function (PMF): Discrete RVs

- A PMF is a quantitative rule that assigns probability to a value of a discrete random variable
- Probability mass functions plot the possible values of a random variable against the probability of observing those values
- We've defined X as color of minion:

$$f_X(x) = P(X=x) = \begin{cases} 0.45 \rightarrow x = 1 \\ 0.30 \rightarrow x = 2 \\ 0.25 \rightarrow x = 3 \\ 0 \rightarrow otherwise \end{cases}$$

This subscript reminds us that this function is associated with random variable X. It is often suppressed.



Cumulative Distribution Function (CDF)

A CDF is a quantitative rule that assigns a probability to a set of values for a random variable and is written as:

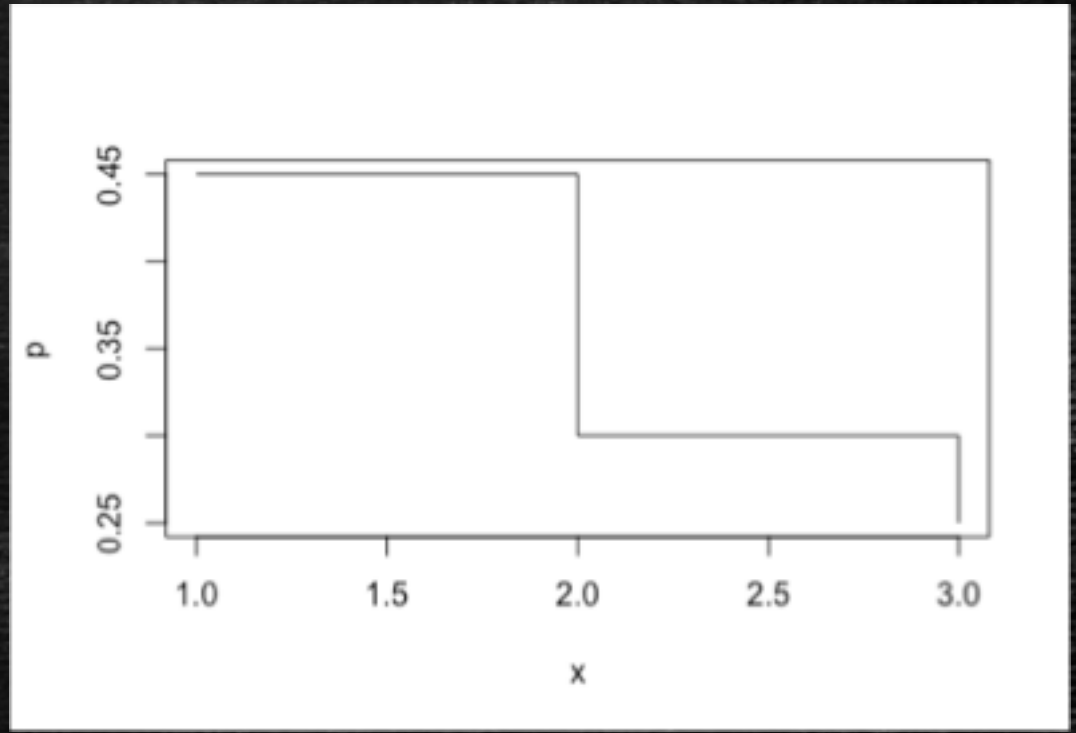
$$F_X(x) = P(X \leq x)$$

…or the probability that *X* is less than or equal to some specific *x*.

Cumulative Distribution Function

X = color of minions, and $F_X(x)$ is the CDF for X:

$$F_X(x) = P(X \leq x) = \begin{cases} o\ if\ x < 1 \\ 0.45\ if\ x = 1 \\ 0.75\ if\ x = 2 \\ 1.00\ if\ x \geq 3 \end{cases}$$



Notice, that if we accumulate the probabilities observed in the PMF, we get the CDF!

Discrete Random Variables: Expectation and Variance

The expected value and variance for random variable X are defined as:

$$\mu_X = E_X[X] = \sum_x (x f_X(x))$$

$$\sigma_X^2 = Var_X[X] = \sum_x (x - E_X[X])^2 * f_X x$$

Discrete Random Variables:
Expectation and Variance

X = color of minions, and $f_X(x)$ is the PMF for X:

$$f_X(x) = P(X = x) = \begin{cases} 0.45 \rightarrow x = 1 \\ 0.30 \rightarrow x = 2 \\ 0.25 \rightarrow x = 3 \\ 0 \rightarrow otherwise \end{cases}$$

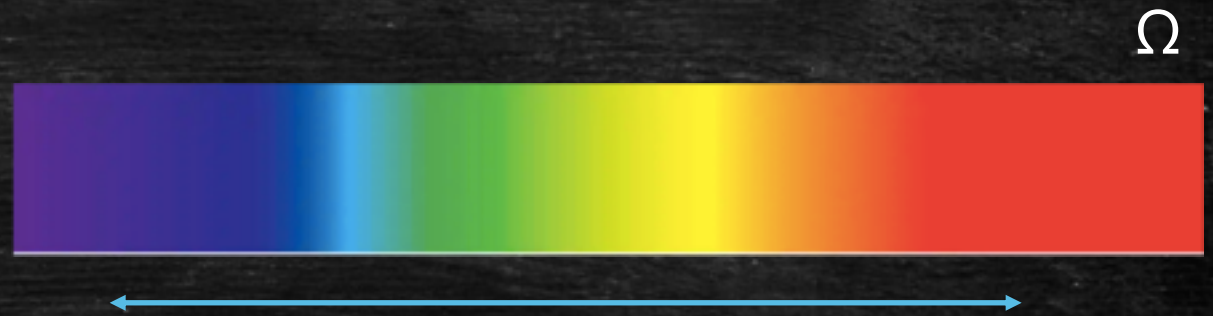
$$E_X[X] = 0.45 * 1 + 0.30 * 2 + 0.25 * 3 = 1.8$$

$$Var_X[X] = (1 - 1.8)^2 * 0.45 + (2 - 1.8)^2 * 0.30 + (1 - 1.8)^2 * 0.25 = 0.46$$

Continuous Random Variables

- Now let's say that in minion world, minions could have any hair length or be of any color

- Our sample space Ω would then consist of an infinite number of possible minions



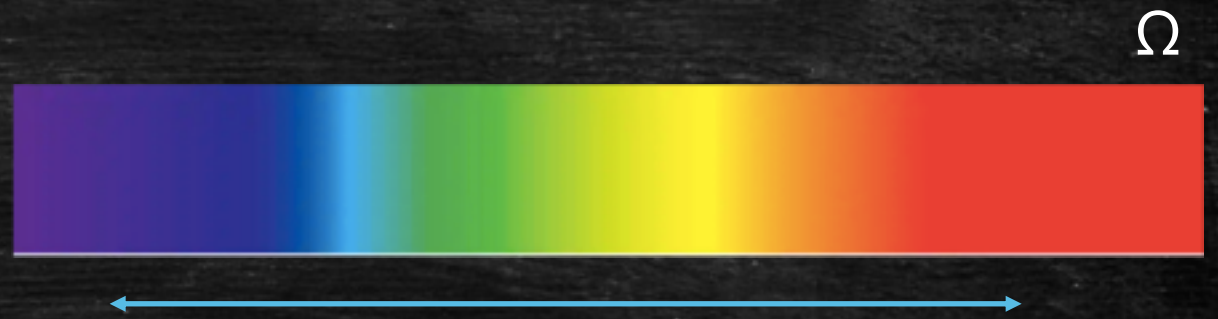
Continuous Random Variables

- Let's define X as color of minion

$$X = wavelength\ in\ nm$$

- Let's define Y as hair length of minion

$$Y = length\ in\ mm$$



Probability Density Function (PDF)

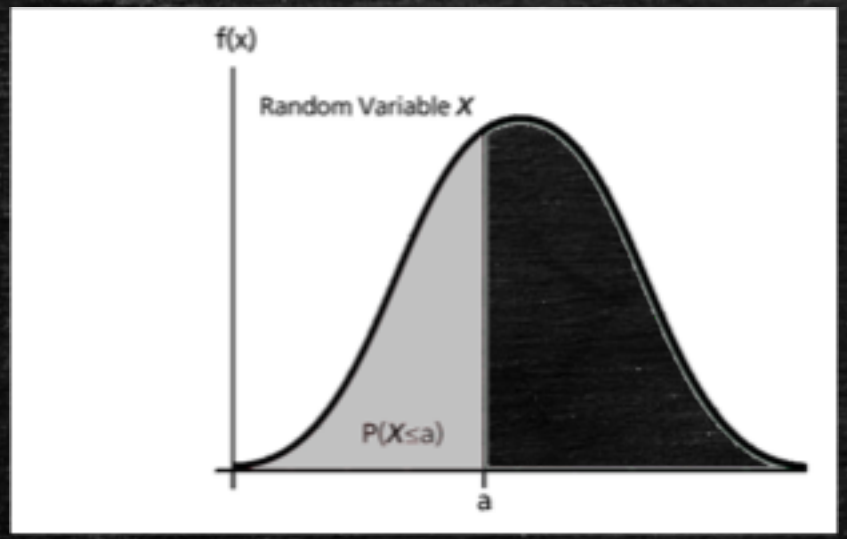
- The PDF for random variable X, $f_X(x)$, has the following properties:
  - $f_X(x) > 0$ for all values of x
  - The area under the PDF over all values of random variable X equals 1
  - If we integrate the PDF, we get the CDF, as it assigns a probability to a set of values from a random variable

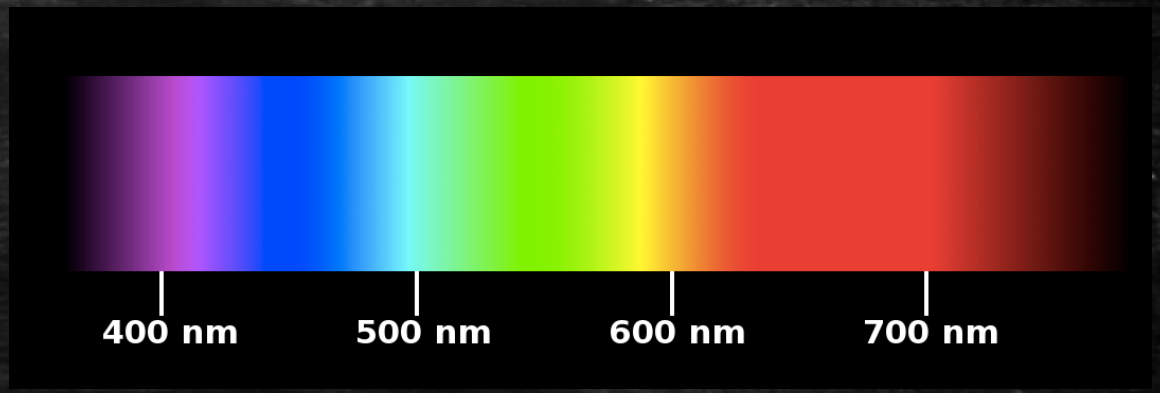
- The definition for the CDF is essentially the same for continuous and discrete random variables → $f_X(x) = P(X \leq x)$

$$CDF = F_X(x) = \int PDF = \int_{x_{min}}^{x_{max}} f_x(x)$$

PDF and CDF

$$F_X(x) = P(X \leq x)$$





$$F_X(550) = P(X \leq 550)$$

$$= P(\quad)$$

Continuous Random Variables:
Expectation and Variance

The expected value and variance for random variable X are defined as:

$$\mu_X = E_X[X] = \int_{x_{min}}^{x_{max}} x f_X(x) dx$$

$$\sigma_X^2 = Var_X[X] = \int_{x_{min}}^{x_{max}} (x - E_X[X])^2 f_X(x) dx$$

Joint Probability Mass Functions

- Joint probability functions can be used to assign probabilities to values obtained by a vector of discrete random variables
- Below, we've defined a joint PMF for 2 random variables, color X and hair type Y

$$f(x, y) = P(X = x, Y = y)$$

X = color of minion
Y = hair type of minion

Joint Probability Mass Function

- Remember:

$$X = \begin{cases} 1 \; if \; G \\ 2 \; if \; R \\ 3 \; if \; B \end{cases}$$

- And:

$$Y = \begin{cases} 1 \; if \; L \\ 2 \; if \; S \end{cases}$$

$f(x,y) = P(X = 1, Y = 1) =$

X = color of minion
Y = hair type of minion

Independence

If X and Y are independent random variables, their joint PMF $f_{XY}(x, y)$ can be written as the product of their marginal PMFs $f_X(x)$, and $f_Y(y)$:

$$f_{XY}(x, y) = f_X(x) * f_Y(y)$$

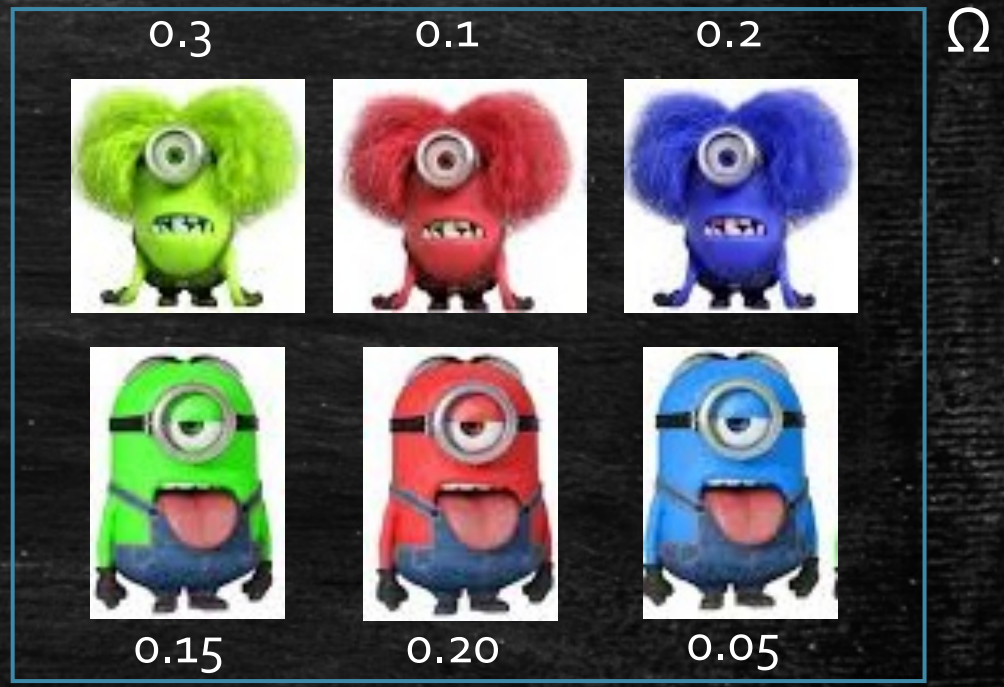
In other words:

$$P(X = x, Y = y) = P(X = x) * P(Y = y)$$

Independence

$$f_X(x) = P(X = x) = \begin{cases} 0.45 \rightarrow x = 1 \\ 0.30 \rightarrow x = 2 \\ 0.25 \rightarrow x = 3 \\ 0 \rightarrow otherwise \end{cases}$$

$$f_Y(Y) = P(Y = y) = \begin{cases} 0.60 \rightarrow y = 1 \\ 0.40 \rightarrow y = 2 \\ 0 \rightarrow otherwise \end{cases}$$



$$f_{XY}(1,1) = f_X(1) * f_Y(1) = 0.45 * 0.60 = 0.27 \sim 0.30$$

X = color of minion
Y = hair type of minion

Common Probability Distributions

Probability Distributions

- We make the assumption that the process by which our data were "generated" can be described by a particular **probability distribution**, whose parameters, θ, are ultimately unknown

Discrete

- Bernoulli
- Binomial
- Poisson
- Geometric
- Hypergeometric
- Negative Binomial

Continuous

- Normal ("Gaussian")
- Exponential
- Gamma
- Uniform
- Beta

Probability Distributions

- A **probability distribution** is a function that gives us the probability for every possible value of a random variable
- When we make distributional assumptions, we often use the following notation:

$$X \sim Distribution(\theta_1, \ldots, \theta_n)$$

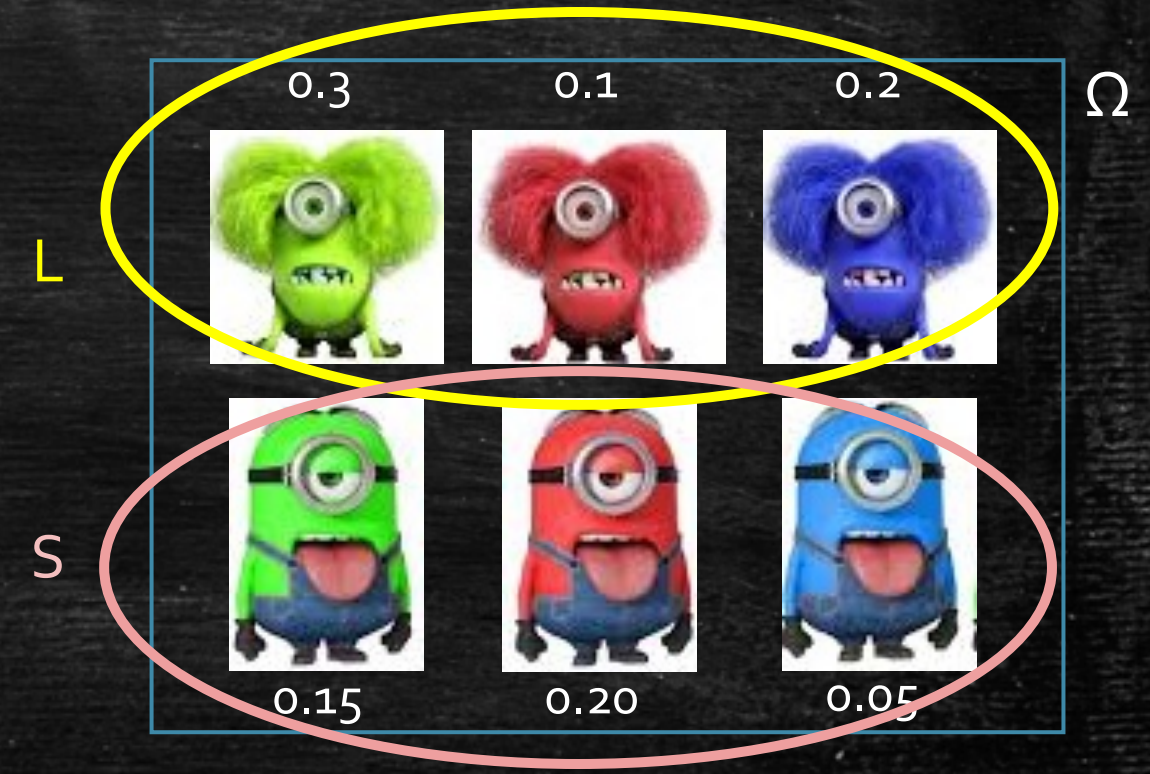
e.g. $X \sim N(\mu, \sigma^2)$

- Specifying a distributional assumption inherently implies that we expect our data to have certain properties**,** e.g.
  - mean
  - variance
  - kurtosis
  - etc.

A Bernoulli Trial

We can think of observing hair length as a binary variable, *X*: 0) short, 1) long
If we were to randomly choose one minion, what would be your best guess? It may be useful to characterize our intuition mathematically:

$$P(X = x) = \begin{cases} 0.40 & \text{for } x = 0 \\ 0.60 & \text{for } x = 1 \end{cases}$$



A Bernoulli Trial

If we assume:

1. Each observation, or "Bernoulli trial," results in one of two possible outcomes (minion with short hair, minion with long hair)
2. The probability of success remains constant across trials (picking X = 1 this trial does not affect what you will pick next trial)
3. Each trial is independent

Then, we can model the probability of possible outcomes in our sample with a binomial distribution

Binomial Distribution

Random Variable → Y = the number of successes in *n* independent Bernoulli trials

Discrete/Continuous? → Discrete

Parameters: *n* – number of trials; *p* – probability of successes

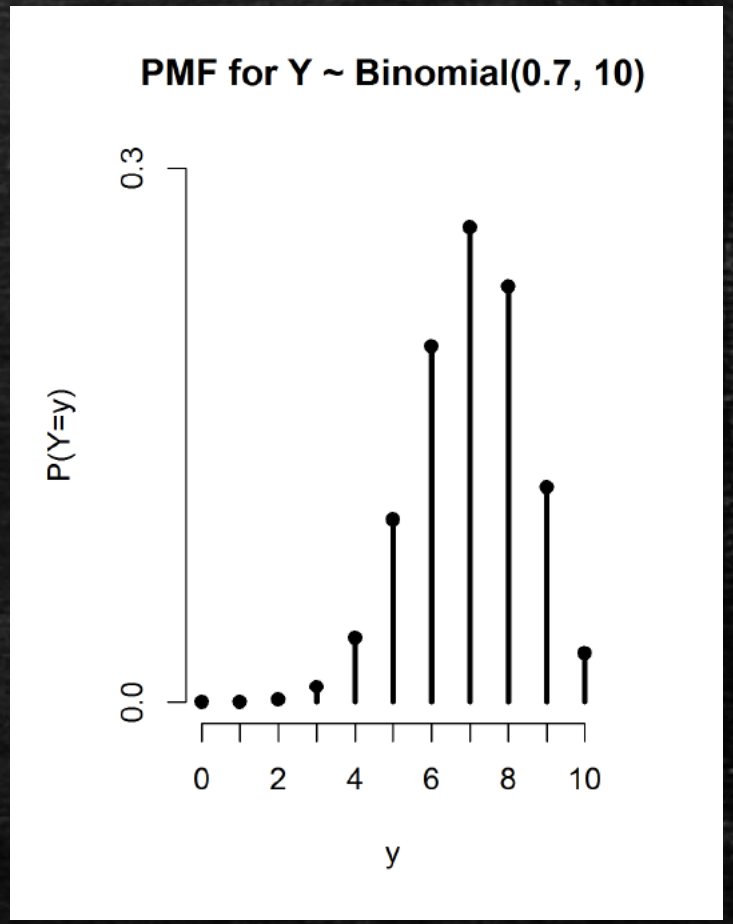
$$\text{PMF} = P(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}$$

Properties

Possible values: $Y \in 0,1,\ldots, n$

Mean: $\mathrm{E}[Y] = np$

Variance: $\mathrm{Var}[Y] = np(1-p)$



Normal Distribution

Random Variable → Y = value of normally distributed variable

Discrete/Continuous? → Continuous

Parameters :μ = mean, σ^2 = variance

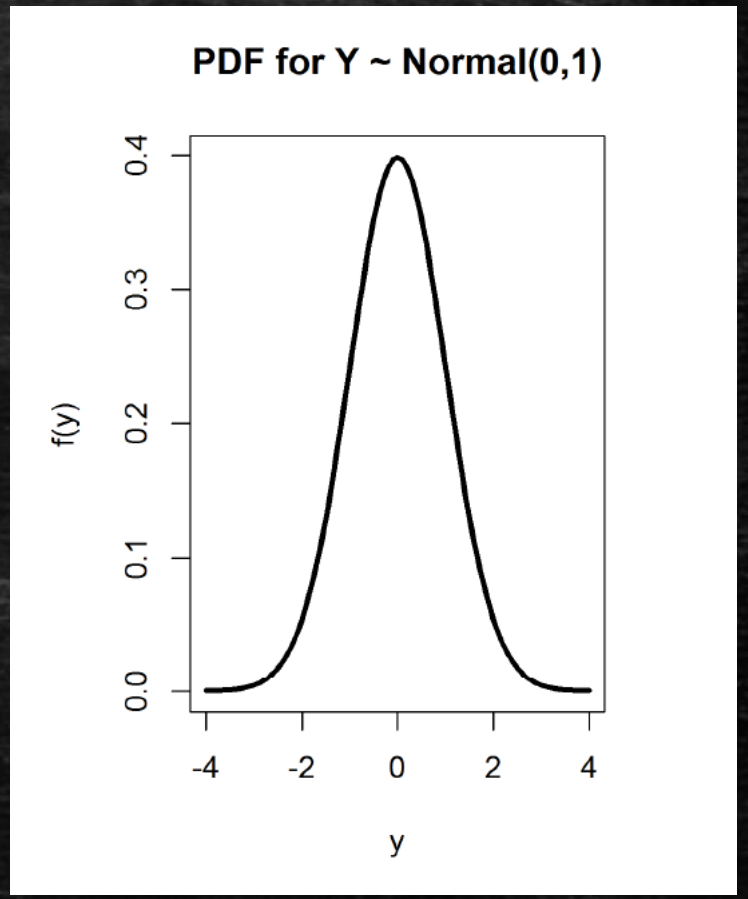
$$PDF = f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

Properties

Possible values: $Y \in (-\infty, \infty)$

Mean: $\mathrm{E}[Y] = \mu$

Variance: $\mathrm{Var}[Y] = \sigma^2$



What have we done so far?

- Probability
  - Stated and applied the rules (axioms) of probability
  - Covered set notation
  - Defined conditional probability, decomposition, and Bayes' Theorem
- Random Variables
  - Defined random variables
  - Characterized PMF/PDF and CDF
  - Understood probability distributions

Remember, Google is your friend! You do not have to memorize the long PMFs and PDFs!

Focus on remembering the properties, and most importantly, the *applications*, of what we have learn.

Let’s take a break!

Please be back in 5 minutes ☺

Introduction to Statistics

Why do we use statistics?

- Probability distributions are not real world observations; rather, they are mathematical constructs that can help us estimate unknown parameters, θ, with our data
- Statistics is the science of connecting these distributions with data
- We'll get into this later, but the terminology of statistics can be a little tricky.
  - Properties of *distributions* differ from properties of *data*
  - e.g. distribution/sampling mean vs. sample mean

Properties of Data: Central Tendency

- The central tendency of a variable reflects its typical, average, or expected value
- For a continuous variable with values of x ranging from [1…n], the **sample mean** is:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- If we were to take more and more numbers into our sample (i.e. increase the size of our sample size *n*), then $\bar{x}$ → μ (the sample mean will converge to the true mean). This is called the **Law of Large Numbers**.
- Other measures of central tendency include median, mode

Properties of Data: Variability

- A data's dispersion, or variability, characterizes how data departs from the center and from each other
- **Sample variance**
  - Biased: $\mathrm{Var}(x) = s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
  - Unbiased: $\mathrm{Var}(x) = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- **Sample standard deviation** is on the same scale as the random variable.
  - $\mathrm{SD} = \sqrt{\mathrm{Var}(x)}$
- Others: range, interquartile range (IQR)

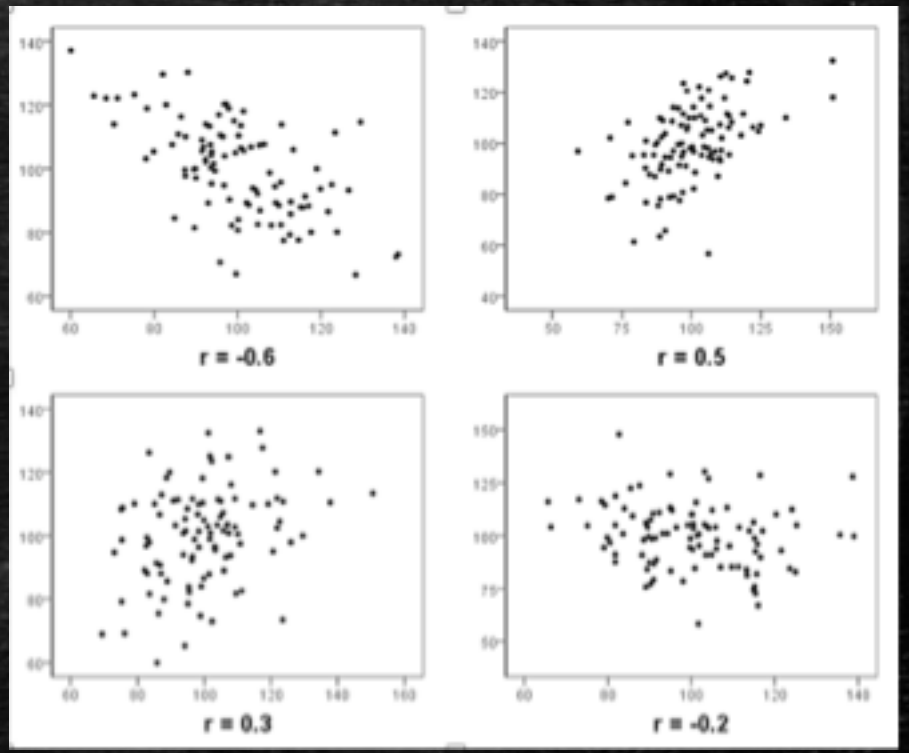
Properties of Data: Covariance

We may also want to describe the joint variability of *X* and *Y*.

Covariance:

$$\widehat{\text{cov}}(X, Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

We can think of this as a measure of the extent to which extreme values of each variable tend to occur together. If they tend in similar directions, the covariance is positive; if they tend in opposite directions, it's negative.



Properties of Data: Correlation

Although covariance can tell us the direction of the relationship, it can be hard to compare magnitudes.

If we standardize the covariance by dividing by the product of the standard deviations, we get the **correlation**, which we often refer to with ρ.

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

The correlation is constrained to be between **-1 and 1**.

Estimation

Estimand: the parameter of interest whose true value we would like to know (e.g. θ, μ)

Estimator: the method of estimating the estimand (for example, if we would like to know the value of estimand μ, we can take a sample and use the sample mean (e.g. $\bar{x}$) as an estimator of μ)

Estimate: the numerical estimate of the estimand that results from the use of a particular estimator

Estimation

Let us say that you want to know μ, the true mean height in the population. This is your _____ .

You take a random sample of 50 people in the population and measure their heights. You then take the mean, $\bar{x}$, of the heights of these 50 people in your sample. $\bar{x}$ is your _____.

Let's say the sample mean, $\bar{x}$, is 64 inches (I made this up). This is your _____.

SO we made the journey from:

Estimation: Expectation

We can think about our estimands as **expectations**, $\mathrm{E}[X]$, of their probability distributions.

For example, the sample mean, $\bar{x}$, is an estimator of the **expected value** of a distribution:

$$\mathrm{E}[X] = \mu$$

The sample variation is the expectation of the squared distance from the mean.

$$\mathrm{Var}(X) = \mathrm{E}[(\mathrm{X} - \mathrm{E}[\mathrm{X}])^2]$$

Similarly, the covariance can be defined as:

$$\mathrm{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$

What makes for a "good" estimator?

A good estimator has a sampling distribution that usually gives us something as close to the estimand as possible.

The distribution of a statistic across infinitely many data sets is called a **sampling distribution**.

Good estimators are often:

- Unbiased: the mean of the estimator's sampling distribution is the target parameter
- Efficient: the estimator has a low variance – since we'll only get to observe one data set, we'd like the estimator to return values close to its mean
  - "more efficient" = "has smaller standard error"
  - Standard error (SE) = $\sqrt{variance}$ of an estimator's sampling distribution

Sample Mean vs. Sampling Distribution

Tool developed by Rice University

Central Limit Theorem

- Many estimators involve using the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, calculated on independent and identically distributed (i.i.d.) data

- Through the central limit theorem (CLT), we can learn what happens to the sampling distribution of $\bar{x}$ asymptotically (i.e., as the sample size *n* grows large)

- If $x_1,\ldots,x_n$ have common mean μ and finite variance σ^2, we can conclude that:

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \text{ as } n \to \infty$$

- Likewise, our **standard error** (the standard deviation of our sampling distribution) is: $\mathrm{SE} = \sqrt{\mathrm{Var}(X)/n}$.

For large *n*, the sampling distribution of the sample mean looks a lot like a normal distribution

Statistical Approaches to Estimation

- The formulas we have discussed are conventional approaches to parameter estimation
- However, how do other measures (e.g. median) fit into this mold?
- What happens when we want to make more complex distributional assumptions? Multivariate assumptions?

- There are actually many ways to go about estimation, e.g.
  - One Least Square (OLS)
  - Maximum Likelihood Estimation (MLE)
  - Method of Moments
  - etc.

True Distribution ≠ Distributional Assumptions

Most of the time, the true distribution that generates our sample data is unknown. When we make distributional assumptions, we are just applying mathematical constructs based on what we know about our data.

In order to build confidence around our estimates, we need validity and hypothesis testing.

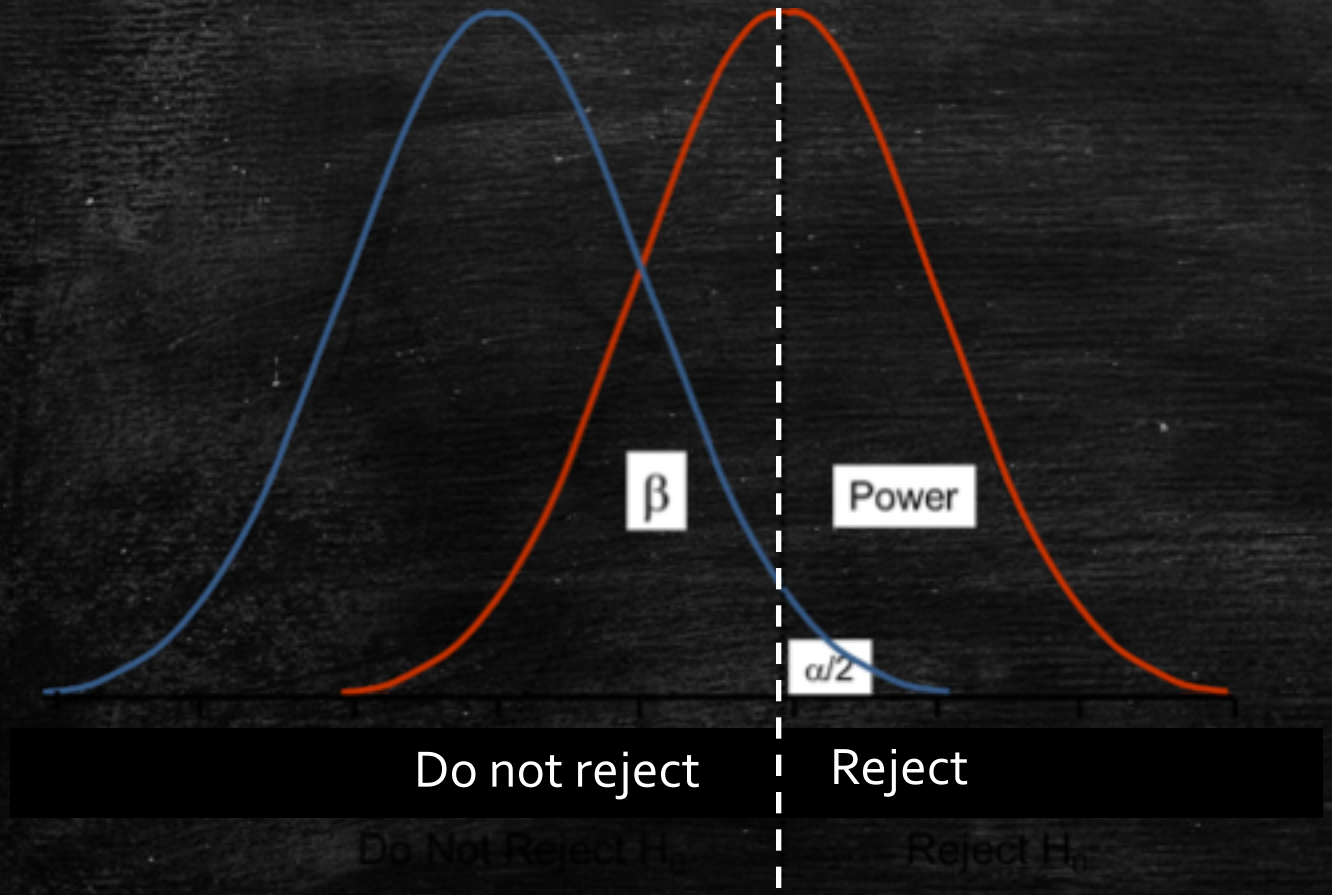
Null Hypothesis Significance Testing (NHST) framework

- Evaluates the compatibility of observed data with some "null" hypothesis H_o, which is the default assumption for the model generating the data
- Two possible options:
  - Reject the null
  - Fail to reject the null (we **never** "accept" the null)
- Usually the null hypothesis is selected so that rejecting the null identifies something of scientific importance
- Distinguishing between statistically and practically significant results is part of our job as researchers!

Null Hypothesis Significance Testing (NHST) framework

- H_0: the null hypothesis
- H_A: the alternative hypothesis
- T: The test statistic calculated from the data
- $Null\ distribution$: the sampling distribution for T if H_0 were true
- $Rejection\ region$: the set of values t for which T = t would lead us to reject the null
- $Acceptance\ or\ non-rejection\ region$: the set of values t for which T=t would lead us to fail to reject the null

Null Hypothesis Significance Testing (NHST) framework



Blue = null

Red = alternate

Type I Error, α: we reject H_0 when H_0 is true

Type II Error, β: we fail to reject H_0 when H_0 is false

Power, 1- β: we reject H_0 when H_0 is false

Null Hypothesis Significance Testing (NHST) framework

Our **significance level**, or the probability of rejection when the null is true, is conventionally set at $\alpha = 0.05$.
Due to the CLT, we can obtain a **test statistic** that is normally distributed:

$$\frac{\bar{x} - E(X)}{\sqrt{\frac{\widehat{\mathrm{Var}(X)}}{n}}} \sim N(0,1) \text{ as } n \to \infty$$

Using the **normal pdf,** we obtain its corresponding **p-value**, or the probability *under the null* of observing a test statistic as extreme or more extreme than the observed test statistic value.

Confidence Intervals

We can also construct 95% **confidence intervals** around our statistic by:

$$\bar{x} \pm 1.96 \sqrt{\frac{\widehat{\mathrm{Var}(X)}}{n}}$$

This is essentially the inverse of what we did to calculate our test statistic.
If we repeatedly take large samples from the population and construct a confidence interval around the sample proportion, they should contain the true μ 95% of the time.

Don’t worry, we got this!

- Today was A LOT, so first, give yourself a pat on the back
- PHS2000A and 2000B will teach you:
  - The relationship between study design and statistics
  - Thoughtful selection of estimands that reflect causal parameters
  - Modern methods for parameter estimation
- All of these concepts that we discussed will be properly contextualized as this course progresses, so even if you don’t understand some of the topics we covered today, that is **okay**
- The goal of this review is to give you an idea of the kinds of concepts that we’ll be drawing on throughout the semester

Acknowledgements

- Thank you so much to Abrania Marrero, Keletso Makofane, Louisa Smith, and Leah Comment for working on this material before me and providing the foundation/inspiration for it
- Thank you to Kat, Matt, Jack, Jarvis, and Michael for their comments and guidance
- All my peers in the PHS program for being an awesome group of people to learn with, grow with, and take on PHS2000A/B with!