

LECTURE 03

Descriptive Statistics and Intuition for Probability

September 18, 2023

PBHLTH 198, Fall 2023 @ UC Berkeley

Andrew O'Connor

Class Outline

- Recap of lecture 2
- Types of Variables
- Descriptive Statistics
- Probability Preview
- Pandas

Recap: Managing Projects

"System"

"System" Files



"System" version
of Python: 3.9.18



"System" version
of Pandas: 2.0.1



Anaconda

Virtual Environment 1
(Used for Project 1)

Python version 3.9.18
pandas version 1.4.4
seaborn version 0.12.0

Virtual Environment 2
(Used for Project 2)

Python version 3.10.12
pandas version 2.0.1
scikit-learn version 1.1

[...]

Recap: Managing Projects

"System"

"System" Files



"System" version
of Python: 3.9.18

"System" version
of Pandas: 2.0.1



Depending on the requirements of new
projects, there could be an error

Notice how different projects have different needs
Not necessary to install every library and clutter your
"system", install libraries as necessary per project

Anaconda

Virtual Environment 1
(Used for Project 1)

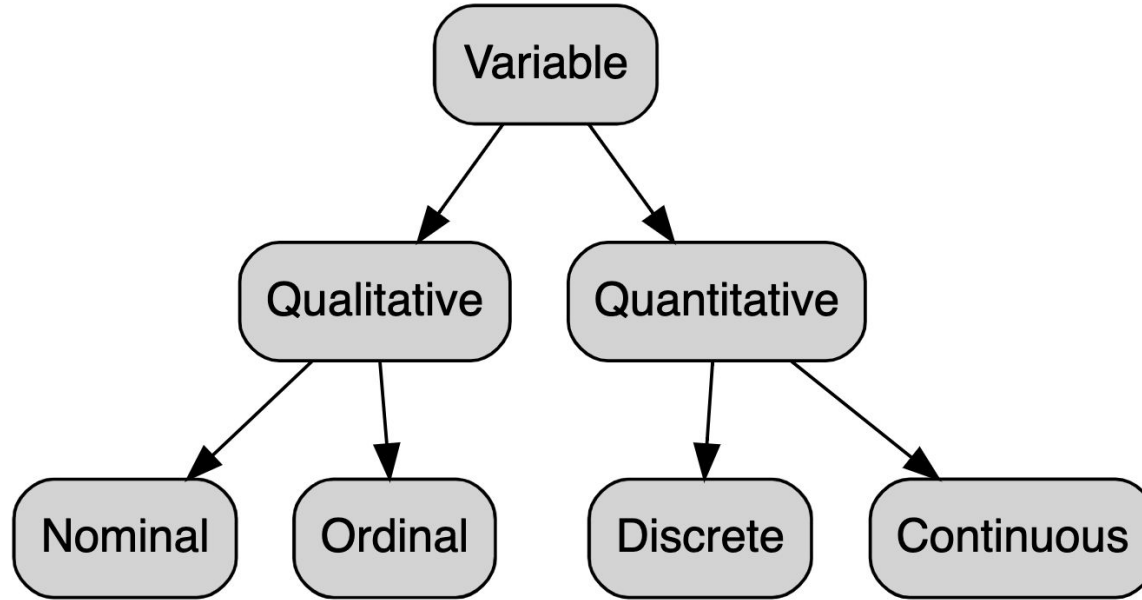
Python version 3.9.18
pandas version 1.4.4
seaborn version 0.12.0

Virtual Environment 2
(Used for Project 2)

Python version 3.10.12
pandas version 2.0.1
scikit-learn version 1.1

[...]

Types of Variables

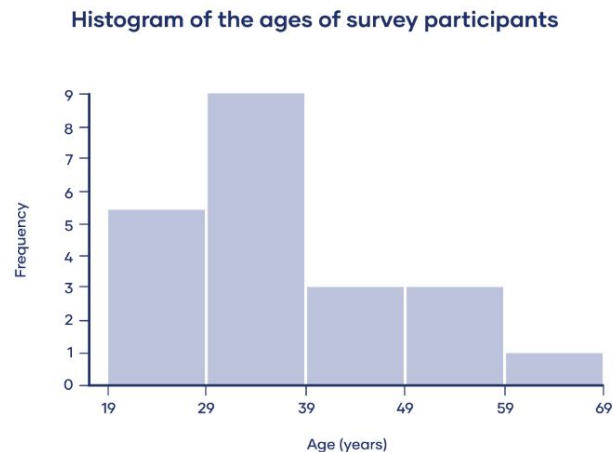
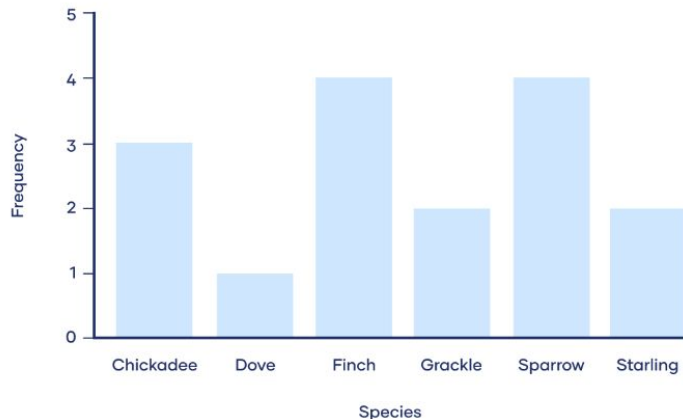


Statistics

- There are 2 main "branches" of statistics:
 - **Descriptive statistics:** A set of techniques used to summarize, organize, and present data in a meaningful way; Objective is to describe the main features of a dataset, such as its
 - **Central tendency** (mean, median, mode)
 - **Spread** (range, variance, standard deviation)
 - **Shape** (skewness, kurtosis)
 - **Inferential Statistics:** Statistics that involves making predictions, inferences, or generalizations about a larger population based on a sample of data. These techniques use probability theory to estimate parameters and test hypotheses; Helps researchers draw conclusions about the population characteristics, relationships, or effects by analyzing sample data

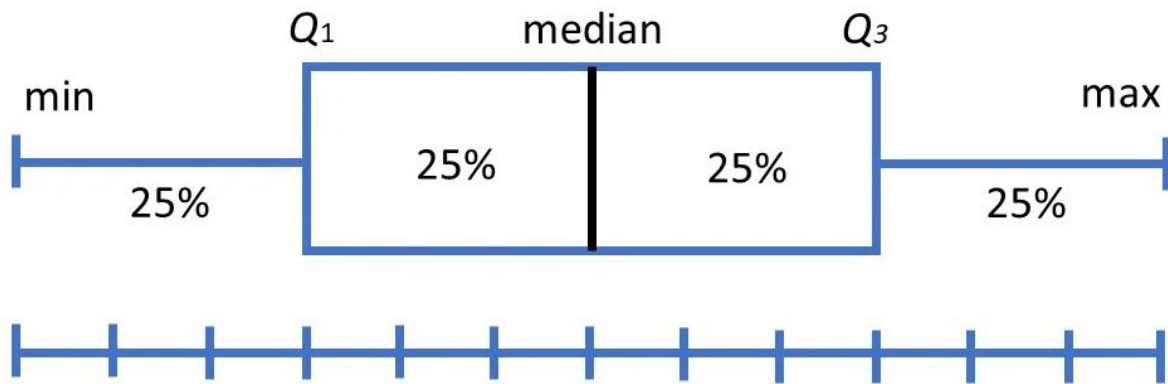
Frequency Distributions

- **Frequency distributions** describe the number of observations for each possible value of a variable
- The frequencies of each possible value can be visually depicted by **histograms**

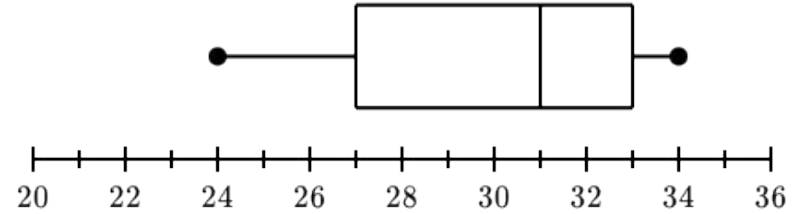


Box Plots

- **Boxplots** are a graphical tool used in to display the distribution, central tendency, and variability of a dataset
- It has 5 main features: the "**Five Number Summary**"



Practice: boxplots

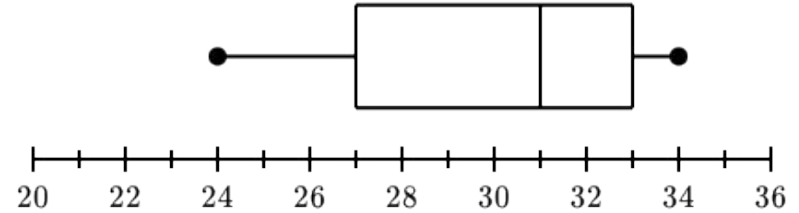


From the figure above, determine the following:

1. Minimum
2. Maximum
3. 25th Percentile
4. 50th Percentile
5. 75th Percentile

Is there any skew in the data?

Practice: boxplots



From the figure above, determine the following:

1. Minimum: **24**
2. Maximum: **34**
3. 25th Percentile: **27**
4. 50th Percentile: **31**
5. 75th Percentile: **33**

Is there any skew in the data?

Yes, depicted by the large gap between Q1 and Q3

Probability

Probability

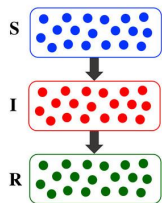
- Exactly what probabilities are has been the subject of contentious debate
 - Some think that probabilities are long-run frequencies that only apply to events that can happen over and over again under identical conditions
 - Others think that probabilities quantify an individual's subjective degree of uncertainty about events of any kind and can vary across individuals
 - Still others don't fall rigidly into either of those groups
- Regardless of any specific interpretation of probability, the fundamentals still hold

[Video]

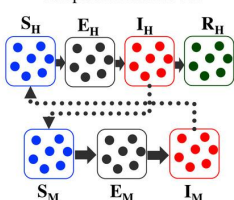


Application: Epidemic Modeling

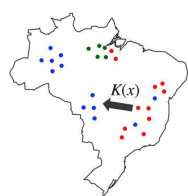
A. Compartmental model



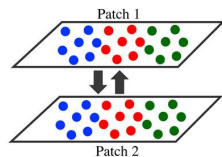
B. Vector-borne compartmental model



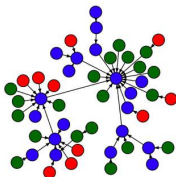
C. Spatial model



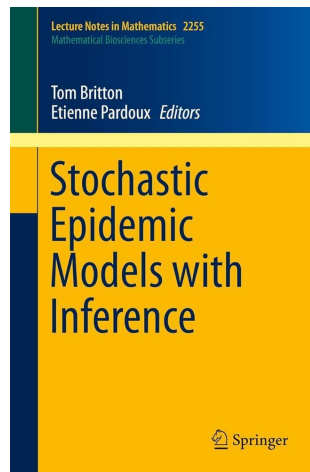
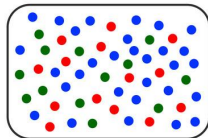
D. Metapopulation model



E. Network model



F. Individual-based model



Requires solid understanding of:

- Probability Theory
- Linear Algebra
- Stochastic Processes
- Programming
- Math
- Everything!

Definitions

- **Outcome space (Ω):** The set of all possible outcomes of a random process
 - EX: Spots you may see on N number of rolls of a die: {1, 2, 3, 4, 5, 6}
- **Event (ω):** A subset of the outcome space
 - EX: The specific event where you draw an Ace on the first card in a standard deck
- **Experiment/Trial:** Any activity that involves chance
 - EX: Rolling a die N times and recording the number of spots you see each time
- **Random Variables:** Numerical mapping of a random process to the number line; The values of random variables are unknown; A function that assigns values to each of an experiment's outcomes
 - EX: Counting the number of trials it takes to flip a heads on a fair coin

Example: Rolling a Fair Die

- Suppose we roll a fair six-sided die a fixed N number of times.
- How can we model the number of times we roll a 6 in N number of rolls?

Here's how we can define the following

Experiment: Each "trial" (in this case, roll of a die), we have an equally likely chance of rolling any of the 6 faces on a die

Outcome Space: $\{0, 1, 2, 3, \dots, N\}$; In our most extreme cases, we can never roll a 6 or we can roll a 6 every time (N)

Random Variable: Let X be a random variable denoting the number of times we roll a 6

Example: Drawing from a Standard Deck

- Suppose we draw cards from a standard deck of cards without replacement
- How can we model the number of cards it takes to draw the first ace?

Here's how we can define the following

Experiment: Each "trial" (in this case, drawing a card), we have an equally likely chance of drawing any of the 52 cards

Outcome Space: $\{1, 2, 3, \dots, 49\}$; In our most extreme cases, we must draw a card so the minimum number of cards we draw is 1. At the other extreme, we draw every card before we finally get to the remaining 4 Aces.

Random Variable: Let X be a random variable denoting the number of cards until we draw an Ace

Practice: Definitions

Suppose you are a researcher studying the occurrence of disease X in a population. You sample N individuals from an eligible study population and record the total number of people that have previously been diagnosed with lung cancer.

Answer the following:

1. What is the outcome space?
2. Define a relevant random variable.
3. What can you consider your "trial" to be?

1. $\{0, 1, 2, 3, 4, \dots, N\}$
2. let X be the number of people who have previously been diagnosed with lung cancer
3. Each trial would be every individual we sample, since each person either has or has not been previously diagnosed

Practice: Definitions

Suppose you are a researcher studying the occurrence of disease X in a population. You sample N individuals from an eligible study population and record the total number of people that have previously been diagnosed with lung cancer.

Answer the following:

1. What is the outcome space?
2. Define a relevant random variable.
3. What can you consider your "trial" to be?

Probability

- The probability an event occurs is equal to the probability of the event occurring divided by the set of all possible outcomes

$$P(A) = \frac{n}{N} = \frac{\text{\textit{\#outcomes in A}}}{\text{\textit{\#outcomes in Sample Space}}}$$

$P(A)$	Event A	The probability of event A happening.
$P(A')$	Complement	The probability of event A not happening.
$P(A \cup B)$	Union	The probability of event A or B happening.
$P(A \cap B)$	Intersection	The probability of event A and B happening.

Standard deck of cards: From now on we will assume the following unless otherwise specified.

- Cards being dealt from a deck means that the cards are drawn at random without replacement.
- A *hand* of cards is a subset dealt from the deck. The order in which the cards appear in the hand doesn't matter.
- A standard deck consists of 52 cards.
 - There are 13 cards in each of 4 *suits*: clubs, spades, hearts, and diamonds. Clubs and spades are black. Hearts and diamonds are red.
 - Each suit has 13 different *ranks*: Ace, 2, 3, 4, 5, 6, 7, 8, 9, 10, Jack, Queen, King.

Practice: Probability

Answer the following:

1. What is the probability of drawing a diamond in a deck of cards?
2. What is the probability of drawing a club OR a spade?
3. If the deck has no face cards (J, Q, K), what happens to the probability of drawing a 4?

Standard deck of cards: From now on we will assume the following unless otherwise specified.

- Cards being dealt from a deck means that the cards are drawn at random without replacement.
- A *hand* of cards is a subset dealt from the deck. The order in which the cards appear in the hand doesn't matter.
- A standard deck consists of 52 cards.
 - There are 13 cards in each of 4 *suits*: clubs, spades, hearts, and diamonds. Clubs and spades are black. Hearts and diamonds are red.
 - Each suit has 13 different *ranks*: Ace, 2, 3, 4, 5, 6, 7, 8, 9, 10, Jack, Queen, King.

Practice: Probability

Answer the following:

1. What is the probability of drawing a diamond in a deck of cards?

$$13 / 52 = 1/4$$

2. What is the probability of drawing a club OR a spade?

$$(13 + 13) / 52 = 26 / 52$$

3. If the deck has no face cards (J, Q, K), what happens to the probability of drawing a 4 or 5?

$$(4 + 4) / (52 - 12) = 8 / 40$$

Increases; The outcome space decreases

Conditional Probability

- With conditional probability, we consider the probability of an event given that another event occurred
- Consider the previous slide: we divide the event occurring divided by the sample space
 - What does it mean for the sample space to decrease? What does that imply on probability?
 - If we decrease the sample space, we remove possible outcomes (given that a certain outcome occurred)
- Leads us to independent vs. dependent events

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

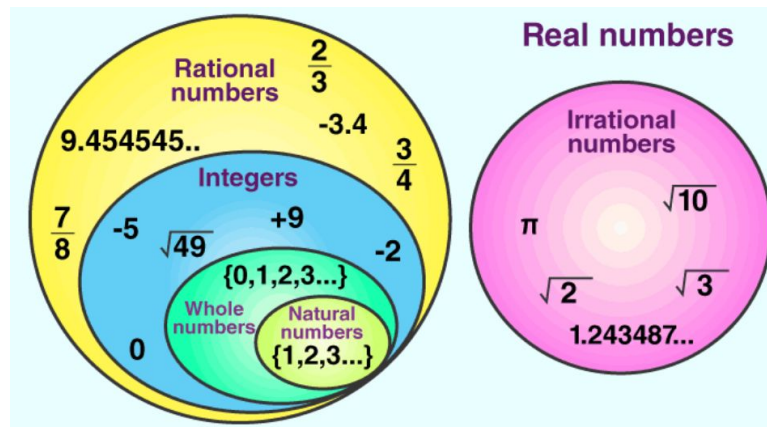
Probability of
A and B

Probability of
A given B

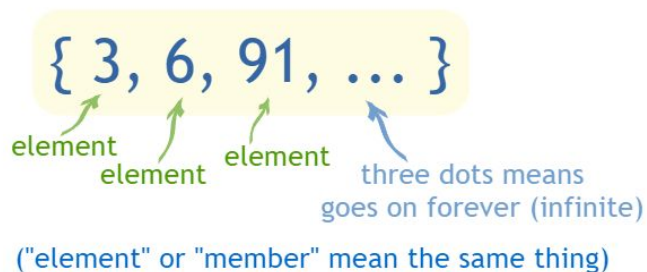
Probability of B

Set Theory

- **Set notation** is fundamental in probability theory because it provides a concise way to describe events and outcomes.
- Understanding sets is crucial for working with probabilities
 - Independence/Dependence
 - Intersections/Unions



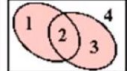
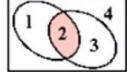
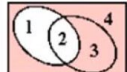
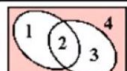
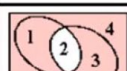
Set Notation



Symbol	Meaning	Example
$\{ \}$	<u>Set</u> : a collection of elements	$\{1, 2, 3, 4\}$
$A \cup B$	<u>Union</u> : in A or B (or both)	$C \cup D = \{1, 2, 3, 4, 5\}$
$A \cap B$	<u>Intersection</u> : in both A and B	$C \cap D = \{3, 4\}$
$A \subseteq B$	Subset: every element of A is in B.	$\{3, 4, 5\} \subseteq D$
$A \subset B$	Proper Subset: every element of A is in B, but B has more elements.	$\{3, 5\} \subset D$
$A \not\subset B$	Not a Subset: A is not a subset of B	$\{1, 6\} \not\subset C$
$A \supseteq B$	Superset: A has same elements as B, or more	$\{1, 2, 3\} \supseteq \{1, 2, 3\}$
$A \supset B$	Proper Superset: A has B's elements and more	$\{1, 2, 3, 4\} \supset \{1, 2, 3\}$
$A \not\supset B$	Not a Superset: A is not a superset of B	$\{1, 2, 6\} \not\supset \{1, 9\}$
A^c	<u>Complement</u> : elements not in A	$D^c = \{1, 2, 6, 7\}$ When $\bigcup = \{1, 2, 3, 4, 5, 6, 7\}$
$A - B$	<u>Difference</u> : in A but not in B	$\{1, 2, 3, 4\} - \{3, 4\} = \{1, 2\}$

Set Operations

- Union
- Intersection
- Difference
- Complement

Set Notation	Pronunciation	Meaning	Venn Diagram	Answer
$A \cup B$	"A union B"	Everything in both sets		$\{1, 2, 3\}$
$A \cap B$	"A intersect B"	Only what is in common with both sets		$\{2\}$
\bar{A} or A'	"A complement"	Everything NOT in set A		$\{3, 4\}$
$(A \cup B)'$	"not A union B"	Everything NOT in set A and set B		$\{4\}$
$(A \cap B)'$	"not A intersect B"	Everything NOT in common between set A and set B		$\{1, 3, 4\}$

Operation	Notation	Meaning
Intersection	$A \cap B$	all elements which are in both A and B
Union	$A \cup B$	all elements which are in either A or B (or both)
Difference	$A - B$	all elements which are in A but not in B
Complement	\bar{A} (or A^C)	all elements which are not in A

Practice: Set Notation

Let $A = \{1, 2, 3, 4\}$ and let $B = \{3, 4, 5, 6\}$

Write out the following using set notation and their respective symbols

1. The **intersection** of A and B
2. The **union** of A and B
3. The **difference** of A and B
4. The **complement** of A

solutions

$$A \cap B = \{3, 4\}$$

$$A \cup B = \{1, 2, 3, 4, 5, 6\}$$

$$A - B = \{1, 2\}$$

$$A^C = \{\text{all real numbers except } 1, 2, 3 \text{ and } 4\}$$

Practice: Set Notation

Let $A = \{1, 2, 3, 4\}$ and let $B = \{3, 4, 5, 6\}$

Write out the following using set notation and their respective symbols

1. The **intersection** of A and B
2. The **union** of A and B
3. The **difference** of A and B
4. The **complement** of A

Axioms of Probability

- Probability is a numerical function on events that satisfies three conditions (**axioms**) that are reasonable based on our natural sense of what probabilities should be

Let Ω be a sample space. The axioms of probability are:

Axiom 1 $P(A) \geq 0$, for all events $A \subseteq \Omega$.

Axiom 2 $P(\Omega) = 1$.

Axiom 3 If events A and B satisfy $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$.

1. The probability of any event cannot be negative
2. The probability of an outcome for a trial is 1 (it must have happened)
3. If two events are mutually exclusive, the probability of them both happening is their sum

Example: Rolling a Fair Die

- Suppose we roll a fair six-sided die a fixed N number of times.
- How can we model the number of times we roll a 6 in N number of rolls?

Here's how we can define the following

Experiment: Each "trial" (in this case, roll of a die), we have an equally likely chance of rolling any of the 6 faces on a die

Outcome Space: $\{0, 1, 2, 3, \dots, N\}$; In our most extreme cases, we can never roll a 6 or we can roll a 6 every time (N)

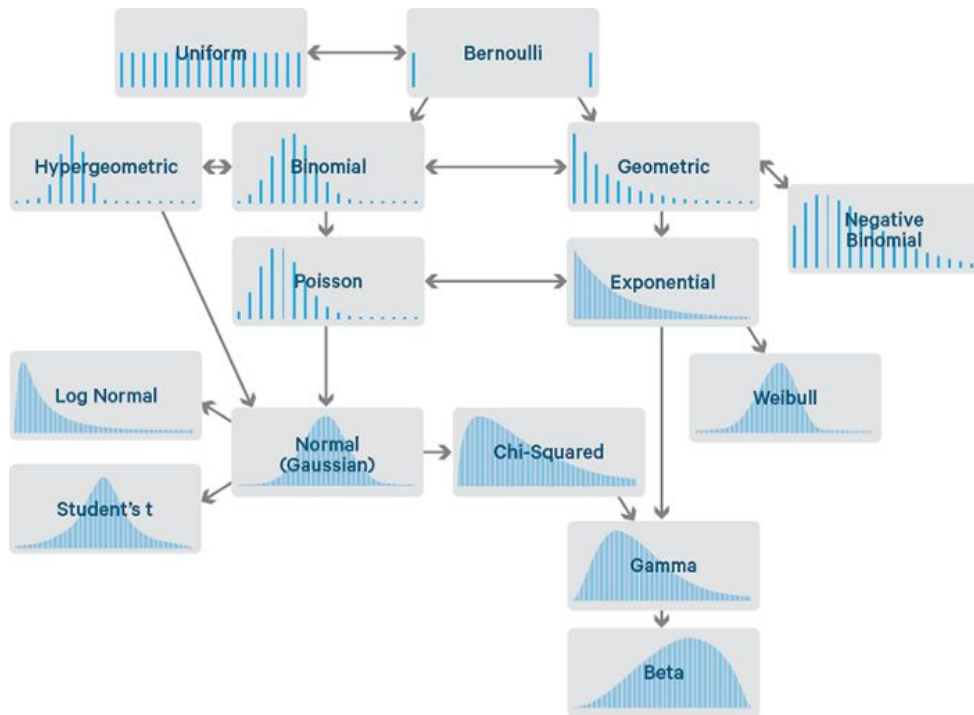
Random Variable: Let X be a random variable denoting the number of times we roll a 6

- Each roll of a die, we have $\frac{1}{6}$ probability of rolling a 6 (our "success")
- Therefore, $X \sim \text{Binomial}(N, p)$

Distributions

- Random variables are associated with **probability distributions**, and these distributions describe how the probabilities are assigned to various values or outcomes of the random variables
- In other words, probability distributions provide a way to understand the likelihood of different outcomes occurring based on the nature of the question we're asking
 - EX: Determine if a single trial is a "success" (**Bernoulli**)
 - EX: Count the number of "successes" in N fixed number of trials (**Binomial**)
 - EX: Count the number of multiple types of "success" in N fixed trials (**Multinomial**)
 - EX: Count the number of trials until our first "success" (**Geometric**)
 - EX: Count the number of "good elements" in a population of N with G good elements sampled without replacement with size n (**Hypergeometric**)
- The named distributions are famous and are defined by **parameters**

Distributions



X	X Counts	$p(x)$	Values of X	$E(x)$	$V(x)$
Discrete uniform	Outcomes that are equally likely (finite)	$\frac{1}{b-a+1}$	$a \leq x \leq b$	$\frac{b+a}{2}$	$\frac{(b-a+2)(b-a)}{12}$
Binomial	Number of successes in n fixed trials	$\binom{n}{x} p^x (1-p)^{n-x}$	$x = 0, 1, \dots, n$	np	$np(1-p)$
Poisson	Number of arrivals in a fixed time period	$\frac{e^{-\lambda} \lambda^x}{x!}$	$x = 0, 1, 2, \dots$	λ	λ
Geometric	Number of trials up through 1st success	$(1-p)^{x-1} p$	$x = 1, 2, 3, \dots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Negative Binomial	Number of trials up through k th success	$\binom{x-1}{k-1} (1-p)^{x-k} p^k$	$x = k, k+1, \dots$	$\frac{k}{p}$	$\frac{k(1-p)}{p^2}$
Hyper-geometric	Number of marked individuals in sample taken without replacement	$\frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}}$	$\max(0, M+n-N) \leq x \leq \min(M, n)$	$n \frac{M}{N}$	$\frac{nM(N-M)(N-n)}{N^2(N-1)}$

Properties of Distributions

- Distributions have multiple properties that define them
 - Expected Value (Mean)
 - Variance
 - Standard Deviation
 - Covariance
- Each of these properties vary depending on the distribution

Why care about distributions?

- **Distributions** are useful in helping us model real world situations and for inferring population parameters
- Applications
 - Estimate population parameters, such as means, variances, or rates, from sample data
 - In clinical trials and healthcare policy research, probability distributions are used to evaluate treatment effectiveness, estimate risks and benefits, and inform healthcare decisions
 - In epidemiology and public health, researchers assess the risk of disease occurrence or exposure to environmental factors; Distributions help quantify these risks

Applications of Distributions

Poisson probability function can help optimize trauma operating room availability

[David R. Sinclair MD](#) 

[Canadian Journal of Anesthesia/Journal canadien d'anesthésie](#) 58, 342–343 (2011) | [Cite this article](#)

3426 Accesses | 1 Citations | 1 Altmetric | [Metrics](#)

To the Editor,

The Poisson probability distribution is a statistical function that describes the occurrence of random events. The probability of the event must be the same for any two intervals of equal length, and the occurrence or non-occurrence of the event in any interval must be independent of the occurrence or non-occurrence in any other interval.¹ The Poisson probability

The arrival of patients at a trauma centre is a random event that follows a Poisson probability distribution. The likelihood of patients requiring immediate surgery is similar in any two time periods of equal length, and the arrival or non-arrival of the patient is independent of the arrival or non-arrival in any other time period. The Poisson probability function can be used to calculate the probability of a patient arriving while all operating rooms are in use. Consider a Level 1 medical centre that has two trauma operating rooms. Equate the number of operating rooms to the number of patients arriving during the average operative time period, as this time interval represents the time period in which each operating room would be occupied if both rooms were being utilized simultaneously. Comparing the number of patients arriving in the average operative time period according to the day of the week or the time of the year would account for weekday and seasonal variability. The probability that a third patient requiring emergency surgery could arrive at the trauma centre while the two operating rooms are occupied is 0.18 (Table C). Three patients from different accidents could also arrive at one time. The probability that three or more patients from independent accidents could arrive in a three-hour period is 0.323 (Table D). There is a 33% probability of needing additional operating room availability. It is conceivable that a single event causing multiple victims could occur, such as a major motor vehicle collision. The arrival of each patient at the trauma centre in any time interval would not be independent of the arrival of any other patient in the same time interval. Tracking of actual arrival rates for a specified period and analysis of correlation would help to address this arrival pattern.

Takeaways

- Descriptive Statistics
 - Mostly interested in describing out data via the 5 number summary and looking at the distribution of the values that were collected; Boxplots and Histograms are useful
- Probability
 - Probability theory is the entire foundation for modern Biostatistics. It is very mathematical and utilizes set notation to describe events and outcomes
 - Numerical functions can be applied on processes that involve chance, to give us random variables
 - Random variables have distributions, in which some are already well defined and proved
 - Distributions of random variables have special properties that can be rigorously proved (in advanced probability classes)