FUNDAMENTALS OF (BIO)STATISTICS PART ONE:
INTRODUCTION TO BIOSTATISTICS

Emma K. T. Benn, DrPH, MPH (*she/her)*
Associate Professor
Founding Director, Center for Scientific Diversity
Director of Data Science Training and Enrichment, Graduate School of Biomedical Sciences
Center for Biostatistics & Department of Population Health Science and Policy
emma.benn@mountsinai.org
Twitter: @EKTBenn

***Note that these slides have been adapted from the NIGMS-funded Applied Statistics in Biological Systems (ASIBS) Short Course (Co-PIs: Emilia Bagiella & Emma Benn) – www.asibs-statistics.org/online-lectures .**

Mount Sinai

Online Texts to Learn More About the Fundamentals of (Bio)Statistics

- **OpenIntro Statistics by David Diez, Mine Centinkaya-Rundel, Christopher Barr, and OpenIntro** - https://leanpub.com/openintro-statistics

- **Introductory Statistics for the Life and Biomedical Sciences by Julie Vu, Dave Harrington, and OpenIntro**- https://leanpub.com/biostat

**Minimum contribution for the above texts is $0 so you can access them for free.*

PART 1.1

What is biostatistics?

Goal of statistics relative to the research path

Internal and external validity

Reliability

What is Biostatistics?

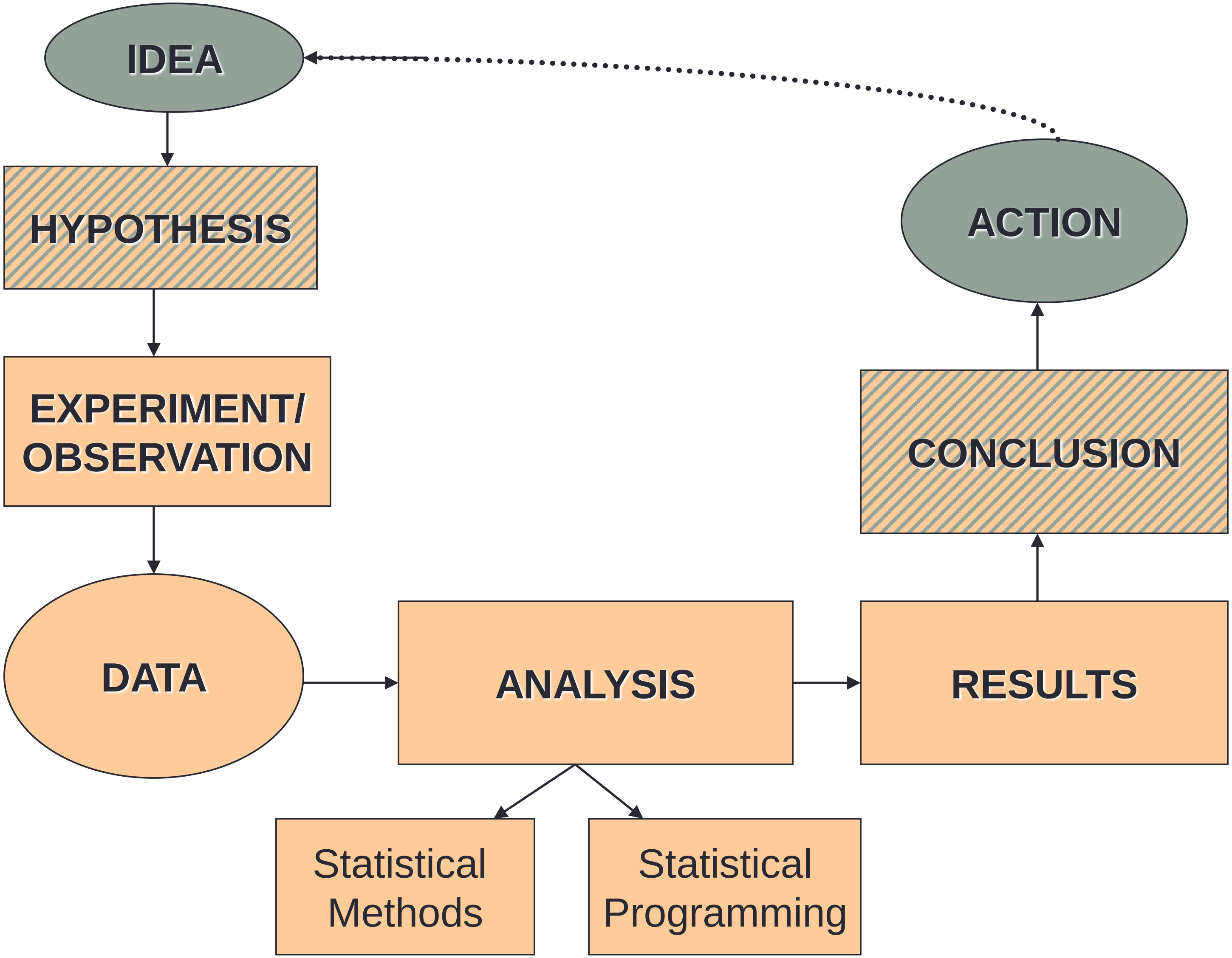
“**Biostatistics is the discipline concerned with how we ought to make decisions when analyzing biomedical data.** It is the evolving discipline concerned with formulating explicit rules to compensate both for the fallibility of human intuition in general and for biases in study design in particular.” (Berger VW, Matthews JR. What does biostatistics mean to us. Mens Sana Monogr. 2006;4(1):89–103.)

“The branch of **statistics** that deals with **data relating to living organisms”** (Wikipedia)

….what is statistics?

“**Statistics** is the science of learning from **data**, and of measuring, controlling, and communicating **uncertainty**…” (Amstat.org)

RESEARCH PATH



IDEA

HYPOTHESIS

EXPERIMENT/ OBSERVATION

DATA

ANALYSIS

Statistical Methods

Statistical Programming

RESULTS

CONCLUSION

ACTION

Fundamental principle

No statistical analysis, no matter how cleverly conducted, can rescue a poorly designed or conducted study.

Data analysis **starts with the design of the study** and the collection of the data.

Validity

- Internal validity: degree to which experiment identifies and measures the actual *causal relationship* in question.

- External validity: degree to which the experiment produces results that can be generalized to the entire population of interest

Internal Validity

- Internal validity refers to how well an experiment is done, especially whether it avoids confounding and bias.

- The less chance for confounding and bias in a study, the higher its internal validity.

External Validity

- External validity refers to how well data and theories from one setting apply to another setting.

- This question is usually asked about laboratory research: Does it apply in the everyday "real" world outside the lab?

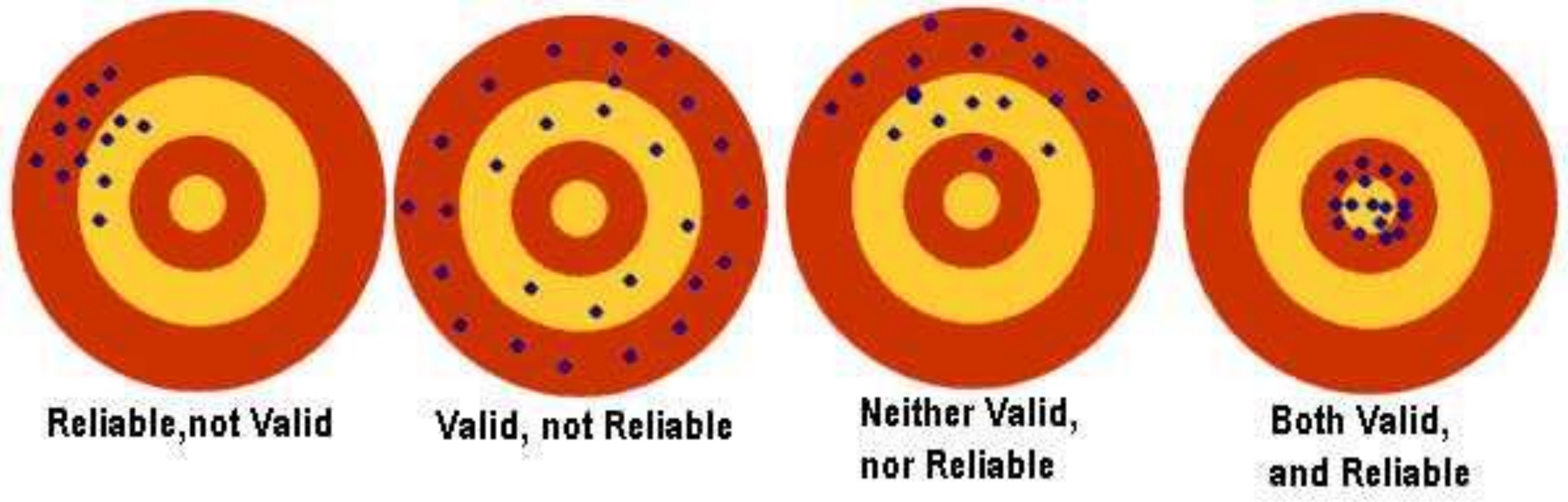
Reliability

- It refers to the **consistency, stability and dependability** of a measure or a result

- Examples
  - Imaging results
  - Disease staging or classification
  - Assay results

Reliability and Validity

Note: reliability is different from validity. A measure may be valid but not reliable and vice versa



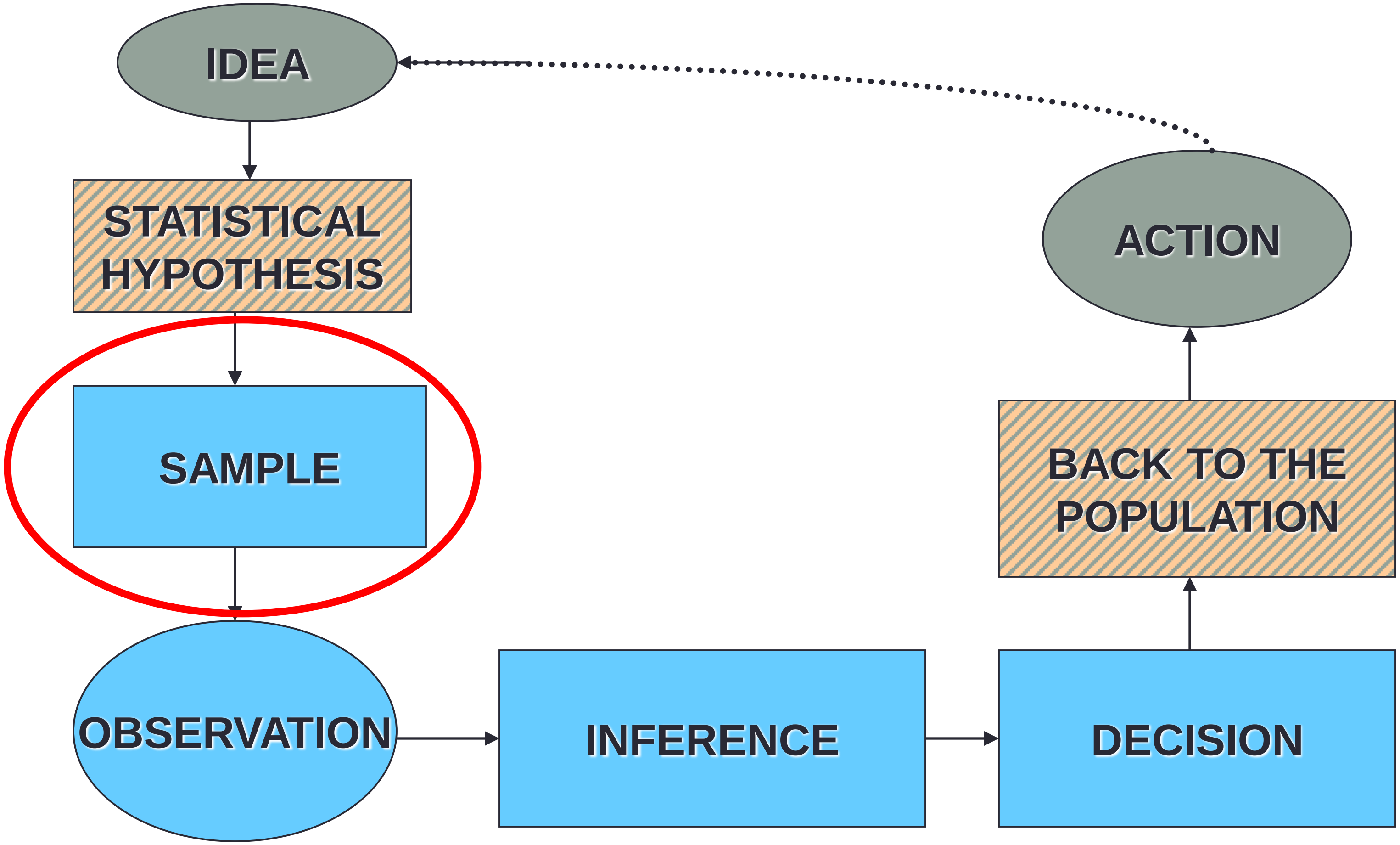
PART 1.2

A first look at the data
Population and sample
Parameters and statistics
Data types
Descriptive Statistics

**Unused*

**I highly recommend reading Chapter 1 (Sections 1.2-1.3) in the freely available “Introductory Statistics for the Life and Biomedical Sciences” by Julie Vu, David Harrington, and OpenIntro: https://leanpub.com/biostat .*

RESEARCH PATH



IDEA

STATISTICAL HYPOTHESIS

SAMPLE

OBSERVATION

INFERENCE

DECISION

BACK TO THE POPULATION

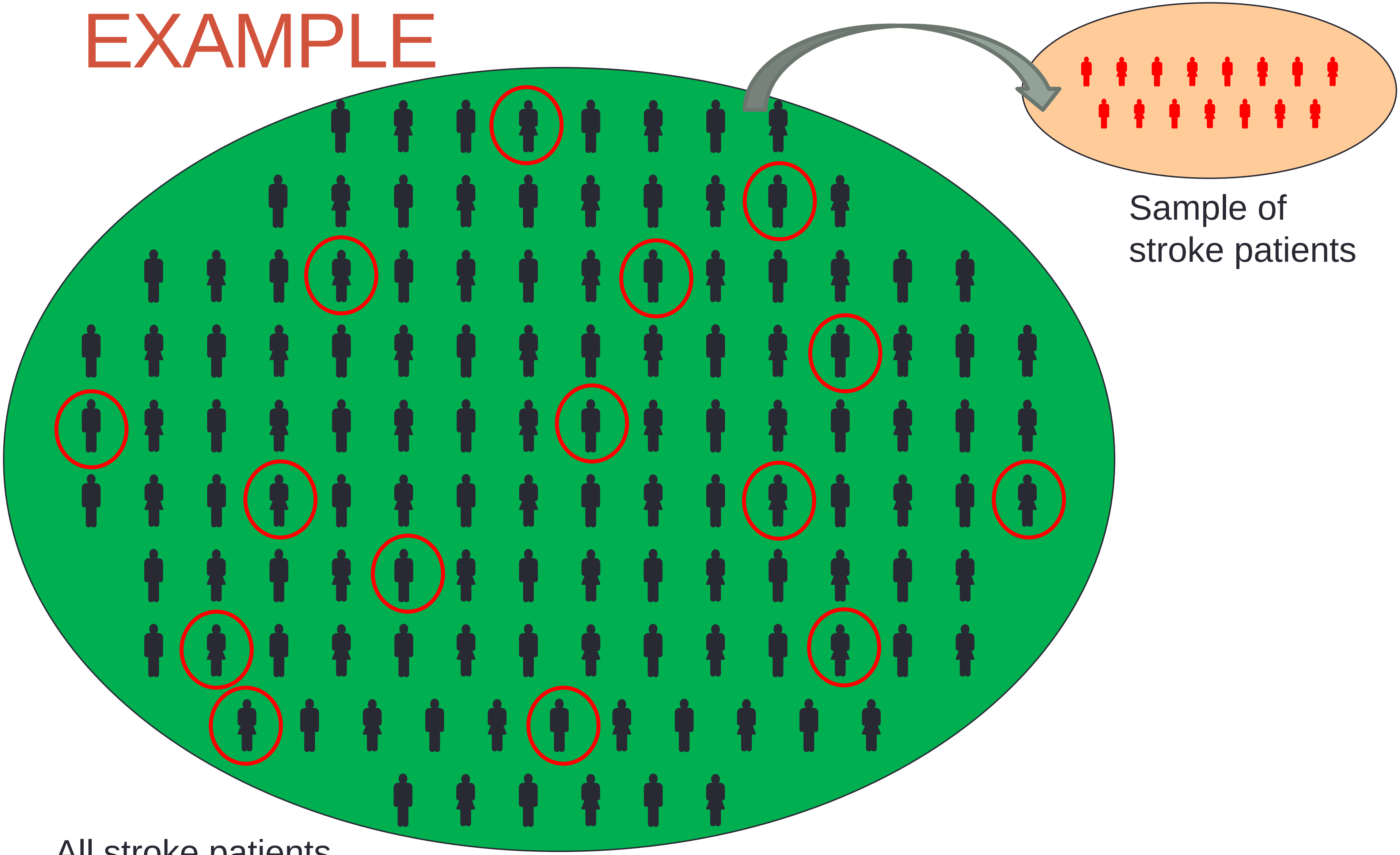
ACTION

Population and Sample

- A **population** is the collection of all individuals (units) who are the target of a specific research question.

- A **sample** is a subgroup of individuals **drawn from the population** of interest.

EXAMPLE



Sample of stroke patients

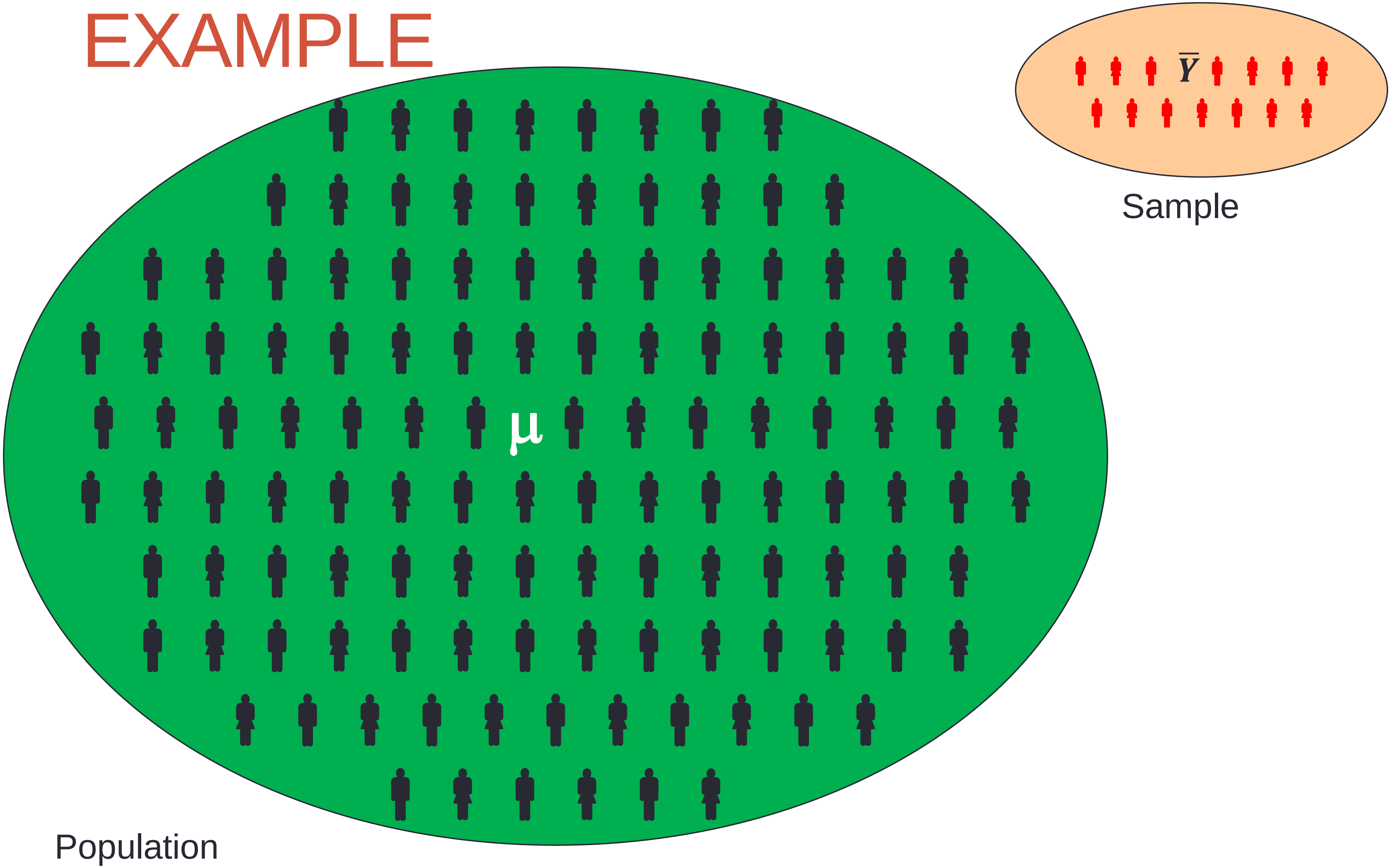
All stroke patients

Parameters and Statistics

- A **parameter** is a numerical characteristic of a *population* (e.g. the average SBP of all stroke patients).
  - We indicate parameters with Greek letters (μ, σ, θ).

- A **statistic** is a numerical characteristic of a *sample* (e.g. the average SBP of a sample from all stroke patients).
  - We indicate statistics with Latin letters ($\bar{Y}$, s).

EXAMPLE



Introduction to statistical data analysis

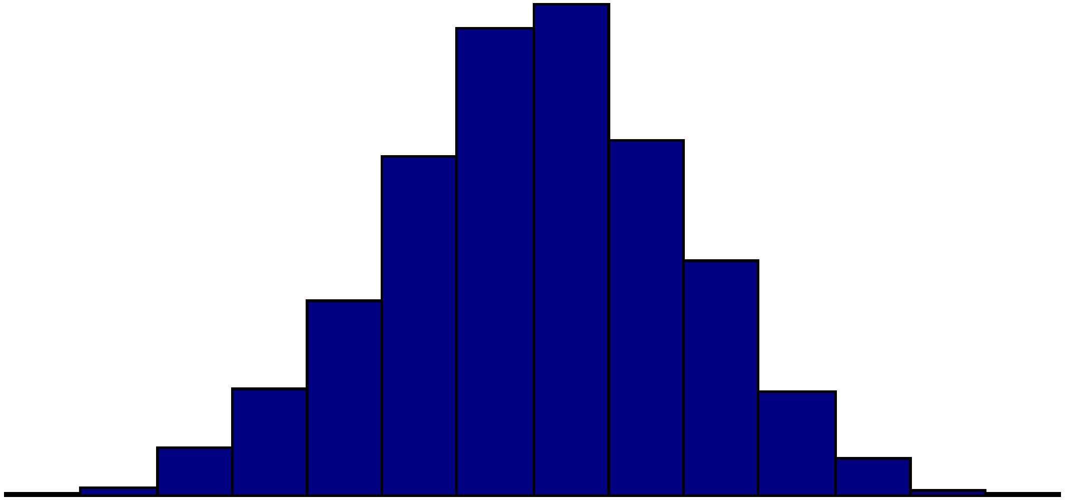
First, know your data!

Before you use complicated statistical tools to analyze your data and find your results you have to become familiar with the data.

Introduction to statistical data analysis

Descriptive statistics:

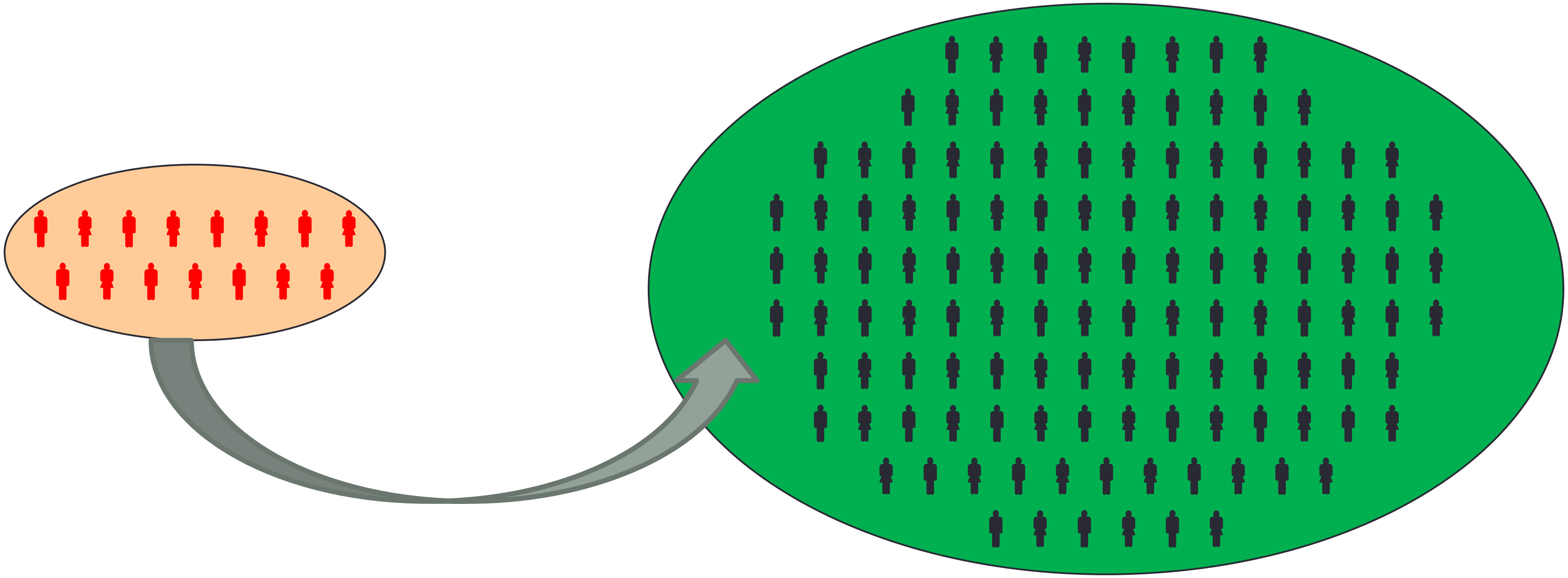
A technique for summarizing and presenting data



Introduction to statistical data analysis

Inferential Statistics:

A technique for reliably generalizing from a sample to the general population



Data Types

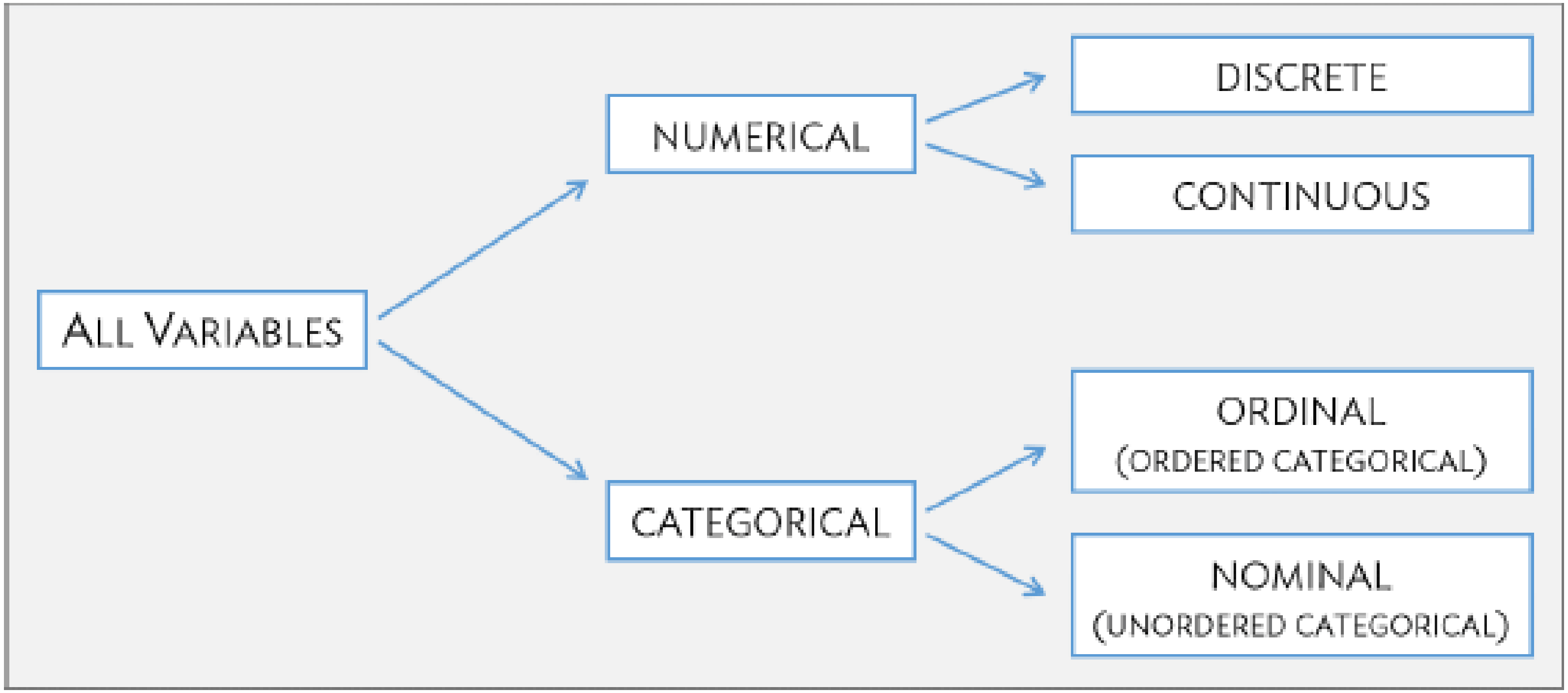


Figure 1.8 from *"Introductory Statistics for the Life and Biomedical Sciences" by Julie Vu, David Harrington, and OpenIntro: https://leanpub.com/biostat*

Categorical (Qualitative) Data Types

- **Nominal scale:**
  - The lowest measurement scale.
  - Un-ordinable.
  - Cannot perform operations (e.g. gender, race, religion, eye color, possible genotypes at a particular locus)
- **Ordinal scale:**
  - Values are ordered on a rank scale.
  - Cannot perform operations (e.g. academic rank, education, stage or severity of disease)

Numerical (Quantitative) Data Types

- **Discrete scale:**
  - Values obtained by **counting** (e.g., number of hospital visits, number of children per family, number of participants in the CREiGS Short Course, number of COVID-19 cases in a particular region)
- **Continuous scale:**
  - Values obtained by measurement. Values can be ordered.
  - Can *theoretically* take on infinite number of values.
  - Operations can be performed (e.g. age, weight, height, heart rate, BMI, glucose level, temperature)

Mayya, S. S., Monteiro, A. D., & Ganapathy, S. (2017). Types of biological variables. *Journal of thoracic disease*, *9*(6), 1730. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5506151/

Tools for data analysis: Descriptive statistics

Descriptive statistics allow us to reduce the space of the study data to a manageable dimension or reorganize it in a way that allows us to identify clear patterns.

Tools for data analysis: Descriptive statistics

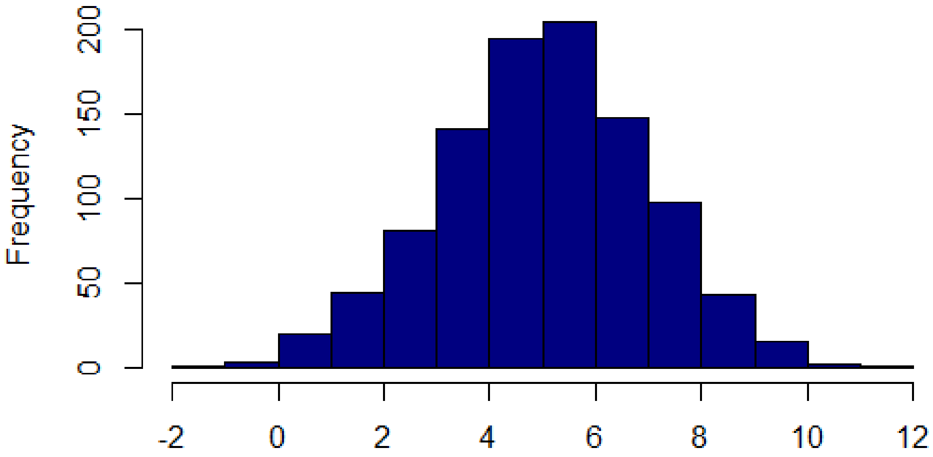
All statistical techniques that allow us to describe the sample

1. Graphs (bar graphs, histograms, charts)
2. Frequency tables
3. Measures of Central tendency
4. Measures of Dispersion
5. Measures of Shape

Frequency Distribution

- The pattern of variation of a variable is called its *distribution*, which can be described both mathematically and graphically.

- The distribution records all possible numerical values of a variable and how often each value occurs (its frequency).



Continuous Measures - Notation

- Let’s denote a variable of interest (e.g. age or SBP) with Y

- Let Y_i denote the value of the variable Y in the i^{th} individual in a population of N individuals or in a sample of n individuals.

Example

SBP in a sample of 10 patients

Y	SBP
Y_1	135
Y_2	122
Y_3	122
Y_4	117
Y_5	135
Y_6	146
Y_7	105
Y_8	135
Y_9	117
Y_{10}	109

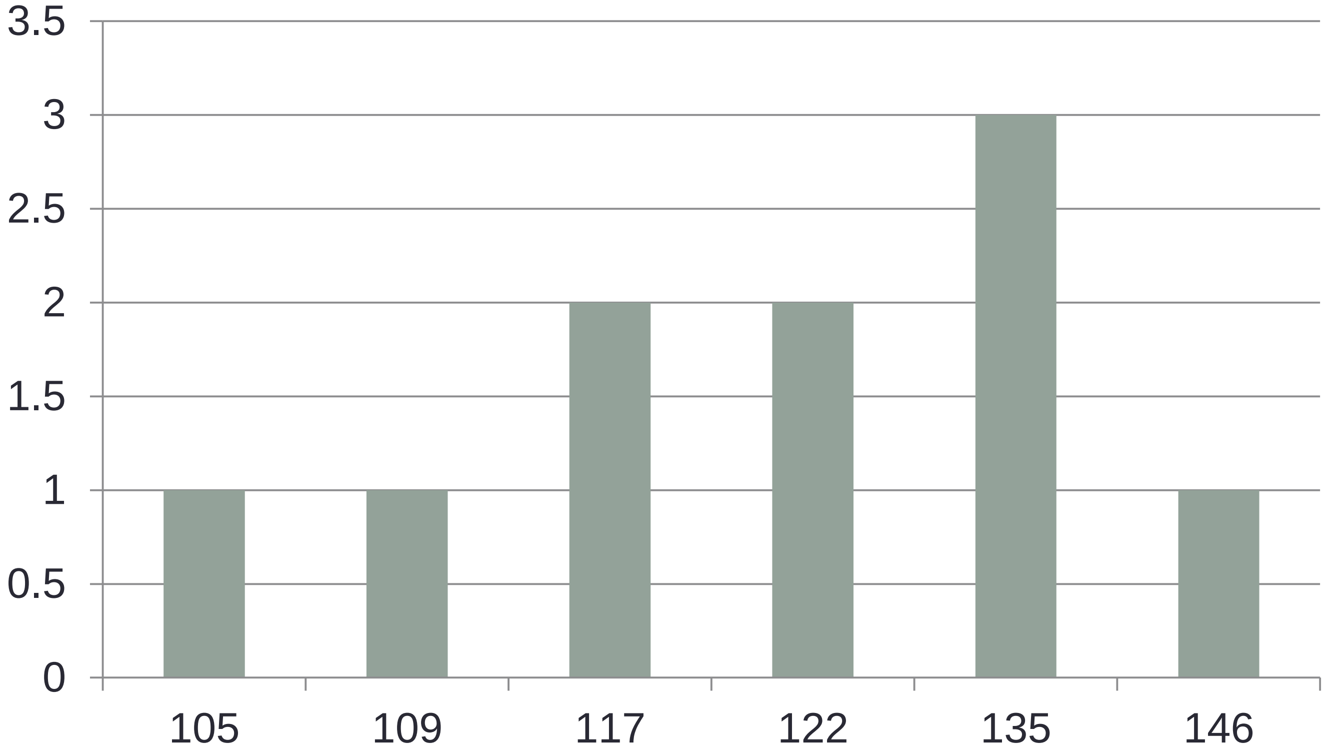
FREQUENCY DISTRIBUTION

SBP in a sample of 10 patients

SBP	Frequency	Relative Frequency	Percent	Cumulative Percent
105	1	0.1	10.0	10.0
109	1	0.1	10.0	20.0
117	2	0.2	20.0	40.0
122	2	0.2	20.0	60.0
135	3	0.3	30.0	90.0
146	1	0.1	10.0	100.0
total	10	1.00	100.0	

FREQUENCY DISTRIBUTION

SBP in a sample of 10 patients



Central Tendency

- Central tendency is a statistical measure that identifies a single score as representative for the entire distribution.

- Measures of central tendency are
  - Mode
  - Median
  - Mean

Mode

- Is the value in the distribution that occurs more frequently.
- If all the values are different there is no mode.
- A distribution may have more than one mode (e.g., bimodal distribution).
- A simple way to find the mode is to plot the frequency distribution and look for the tallest “bump.”

Mode

SBP in a sample of 10 patients

SBP	Frequency	Percent
105	1	10.0
109	1	10.0
117	2	20.0
122	2	20.0
135	3	30.0
146	1	10.0

Median

The value that divides the frequency distribution **of the ordered values** exactly in half

$$M_Y = \begin{cases} \dfrac{Y_{\left(\frac{n}{2}\right)} + Y_{\left(\frac{n}{2}+1\right)}}{2} & \text{if } n \text{ is even} \\ Y_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd} \end{cases}$$

Properties:

For a given set of data there is only one median
The median is easily understood and easy to compute
The median is NOT affected by extreme values

Median

SBP in a sample of 10 patients

Y	SBP
$Y_{(1)}$	**105**
$Y_{(2)}$	**109**
$Y_{(3)}$	**117**
$Y_{(4)}$	**117**
$Y_{(5)}$	**122**
$Y_{(6)}$	**122**
$Y_{(7)}$	**135**
$Y_{(8)}$	**135**
$Y_{(9)}$	**135**
$Y_{(10)}$	**146**

$$M_Y = \frac{Y_{\left(\frac{n}{2}\right)} + Y_{\left(\frac{n}{2}+1\right)}}{2} = \frac{Y_{(5)} + Y_{(6)}}{2} = \frac{122+122}{2} = 122$$

Arithmetic Mean

Is the “center of gravity” of the distribution

$$\mu = \frac{1}{N}\sum_{i=1}^{N} Y_i \qquad \bar{Y} = \frac{1}{\mathrm{n}}\sum_{i=1}^{n} Y_i$$

Properties:

1. For a given set of data there is only one arithmetic mean
2. The arithmetic mean is easily understood and easy to compute
3. Each and every value in a sample contributes to the mean

Mean

SBP in a sample of 10 patients

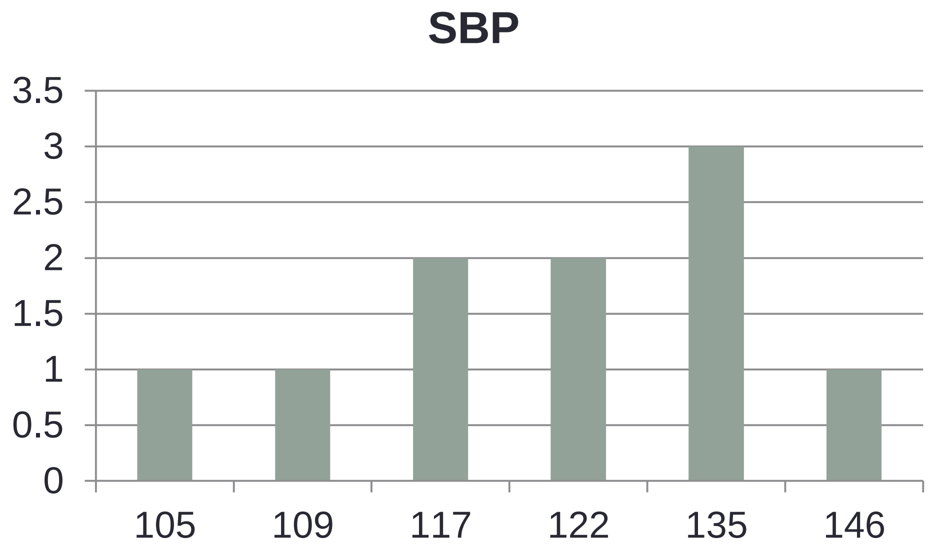
Y	SBP
Y_1	**135**
Y_2	**122**
Y_3	**122**
Y_4	**117**
Y_5	**135**
Y_6	**146**
Y_7	**105**
Y_8	**135**
Y_9	**117**
Y_{10}	**109**

$$\bar{Y} = \frac{1}{\mathrm{n}} \sum_{i=1}^{n} Y_i = \frac{1243}{10} = 124.3$$

SUMMARY

For our sample of 10 patients:

- Mode = 135
- Median = 122
- Mean = 124.3



Dispersion

- Variability provides a quantitative measure of the degree to which scores in a distribution are spread around or clustered together.

- One of the goals of statistical analysis is to understand variability.

Variance and Standard Deviation

Variance

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \mu)^2$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Standard Deviation

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \mu)^2}$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Other measures of dispersion

Inter-quartile range (IQR)

$$I_R = Y_{75} - Y_{25}$$

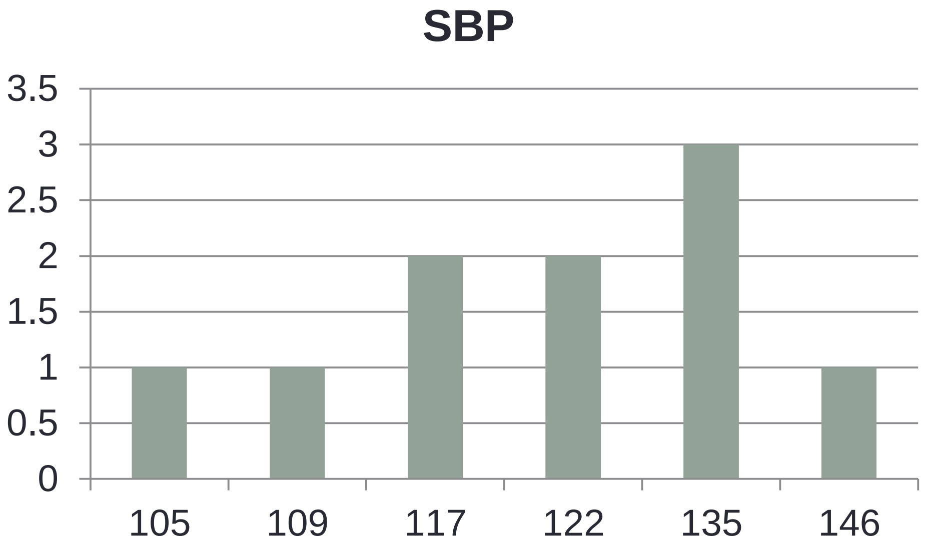
Range

$$R = Y_{\max} - Y_{\min}$$

SUMMARY

For our sample of 10 patients:

- Mode = 135
- Median = 122
- Mean = 124.3
- Variance = 170.9
- St. Dev. = 13.07
- IQR = 135-117 = 18
- Range = 146-105 = 41



Measures of shape

Skewness

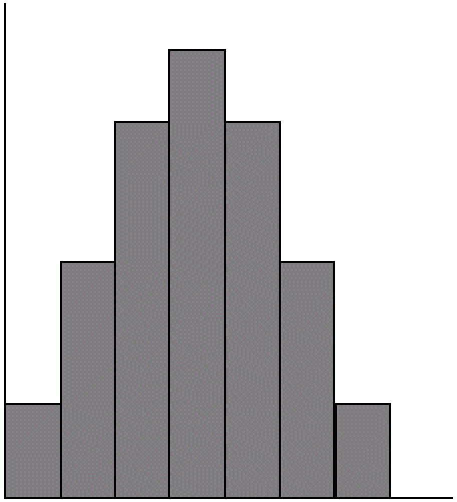
- A measure of symmetry.
- Can be positive or negative

Kurtosis

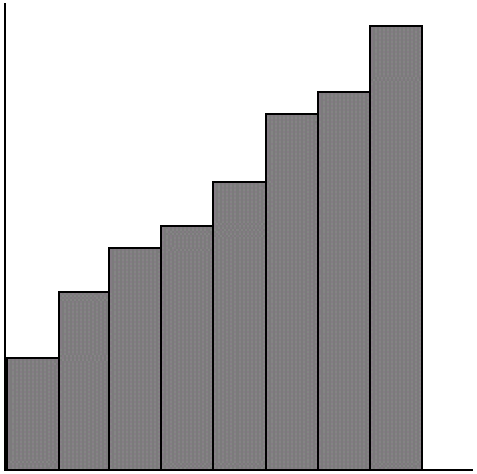
- A measure of the heaviness of the tails of a distribution

Both measures are independent of the location and scale parameters

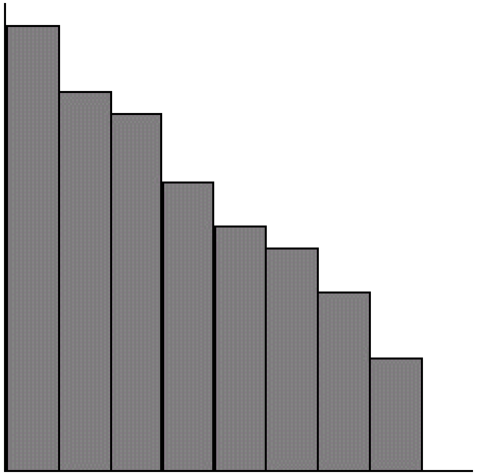
Shape of a distribution



Symmetric Distribution Histogram



Negatively Skewed Histogram



Positively Skewed Histogram

END OF PART ONE

FUNDAMENTALS OF (BIO)STATISTICS PART TWO:
PROBABILITY & PROBABILITY DISTRIBUTIONS

Emma K. T. Benn, DrPH, MPH (*she/her)*
Associate Professor
Founding Director, Center for Scientific Diversity
Director of Data Science Training and Enrichment, Graduate School of Biomedical Sciences
Center for Biostatistics & Department of Population Health Science and Policy
emma.benn@mountsinai.org
Twitter: @EKTBenn

***Note that these slides have been adapted from the NIGMS-funded Applied Statistics in Biological Systems (ASIBS) Short Course (Co-PIs: Emilia Bagiella & Emma Benn) – www.asibs-statistics.org/online-lectures .**

Mount Sinai

PART 2.1

Introduction to Probability

I highly recommend that you read Chapter 2 of *"Introductory Statistics for the Life and Biomedical Sciences" by Julie Vu, David Harrington, and OpenIntro:* *https://leanpub.com/biostat*

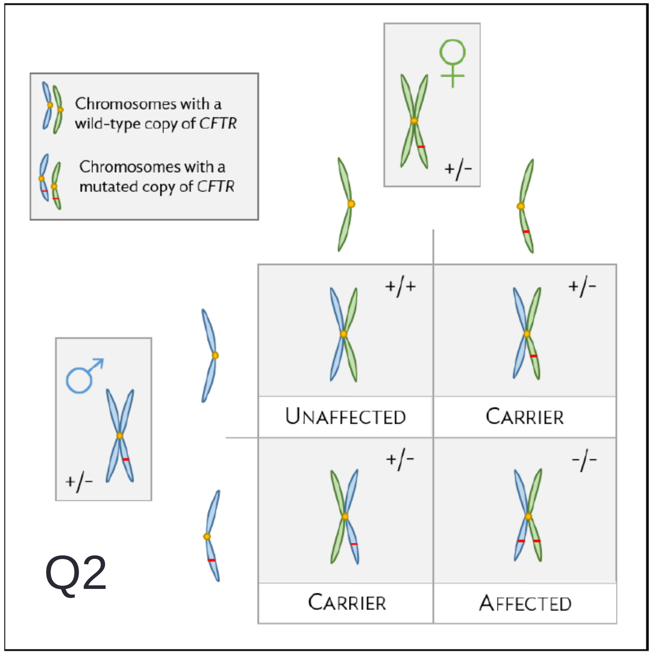
Probability Basics

- Q1: If we were to toss a coin, what’s the probability of getting heads?

- Q2: If each member of a heterosexual couple have one wild-type copy and one mutated copy of CFTR (i.e., they are both Cystic Fibrosis carriers), what is the probability that a child of this couple will be affected by Cystic Fibrosis?

Definitions

- Sample Space – **set of all possible outcomes**
  - Q1: Heads, Tails
- Event – **set of outcomes of interest**
  - Q1: Getting Heads
  - Q2: Child is Affected

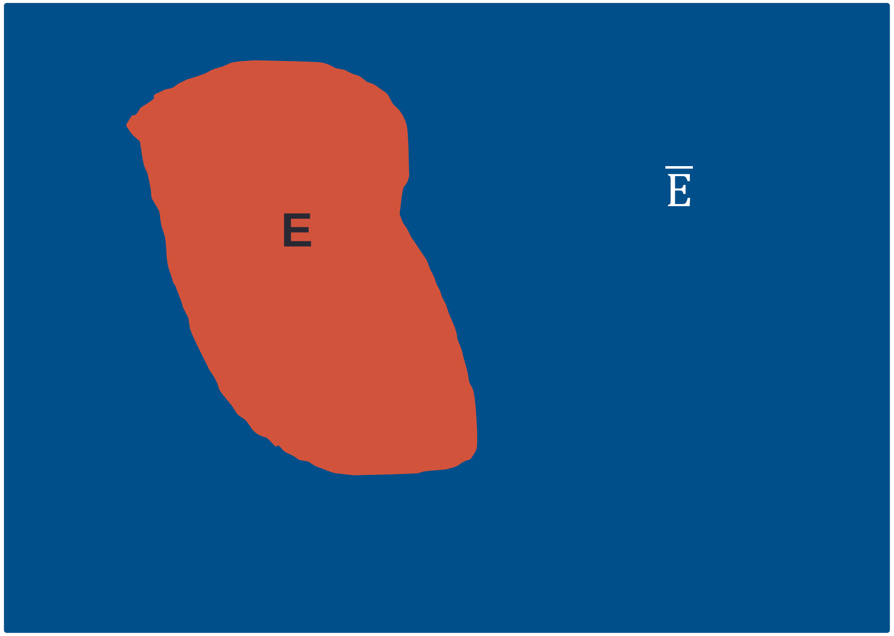

Chromosomes with a wild-type copy of *CFTR*
Chromosomes with a mutated copy of *CFTR*
♀
+/–
♂
+/–
+/+
+/–
UNAFFECTED
CARRIER
+/–
–/–
CARRIER
AFFECTED
Q2

- Probability of an Event – **Relative frequency of the outcome(s) of interest** over an **indefinitely large (or infinite)** number of trials.
  - **Empirical probabilities rely on finite set of data**
    - **Imagine tossing a coin 10,000 times and getting 4950 Heads + 5050 Tails**
    - **Pr(Getting Heads) =** $\frac{\text{no. of Heads}}{\text{total no. of outcomes}} = \frac{\text{4950 Heads}}{\text{4950 Heads} + \text{5050 Tails}} =$ **0.495**
- General Probability Rule
  - The probability of event E, denoted by Pr(E) **always satisfies 0 ≤ Pr(E) ≤ 1**.

Figure 2.1: Pattern of CF inheritance for a child of two unaffected carriers from "Introductory Statistics for the Life and Biomedical Sciences" by Julie Vu, David Harrington, and OpenIntro: *https://leanpub.com/biostat*

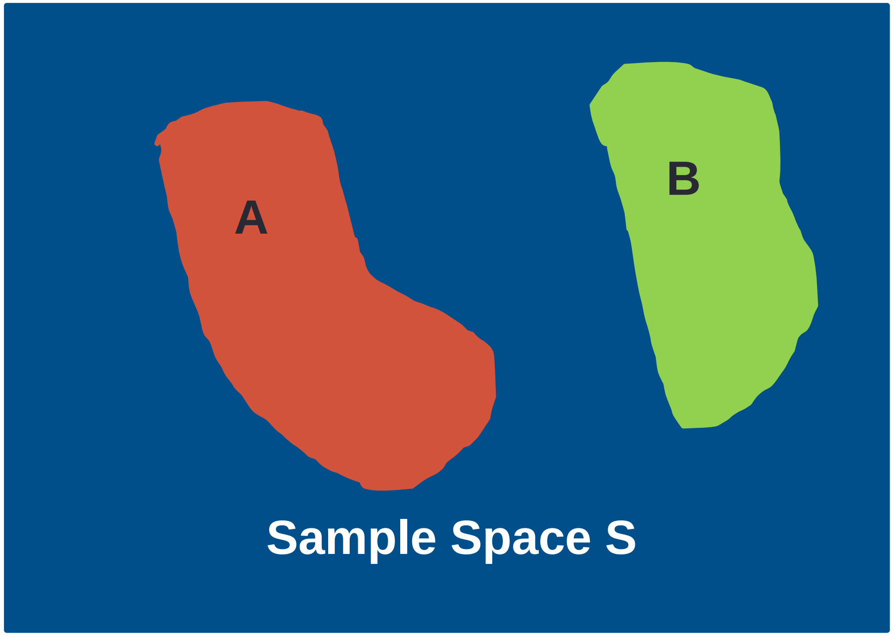
The Complement

- We define the **complement of an event E** as all of the sample space that does not include event E. Denoted by:
  - E’
  - E^c
  - $\overline{E}$



- In probabilistic terms:
  - $\Pr(\overline{E}) = 1 – \Pr(E)$
  - $\Pr(E) = 1 – \Pr(\overline{E})$

Visualization 1



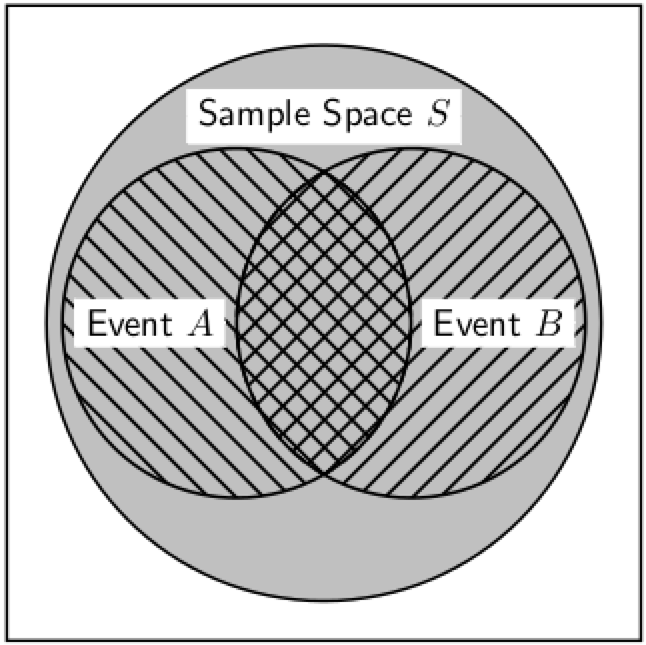
In this diagram, events A and B are **mutually exclusive**. Based on the diagram above, how would you define mutual exclusivity.

Mutual Exclusive Events

- If two events cannot happen at the same time, they are mutually exclusive.
  - If you flip one coin, you cannot get Heads and Tails at the same time.

- Probability Rules for Mutual Exclusivity
  - Pr(A or B) = Pr(A ∪ B) = **Pr(A) + Pr(B)**
  - Pr(A and B) = Pr(A ∩ B) = **0**

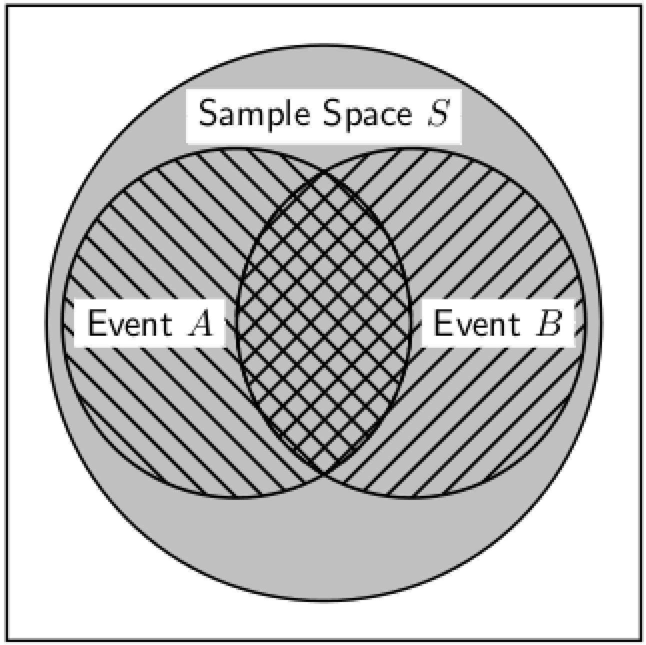
Challenge



Based on this diagram, would you consider A and B
to mutually exclusive events?

Challenge cont'd

- Based on your decision as to whether A and B are mutually exclusive events, how would you calculate the **Pr**(A ∪ B) ?



Independence and Probability

- If A does not depend on B, then events are **independent.**
  - Hypertensive status of Parent 1 **independent** of hypertensive status of Parent 2 only if…

- Multiplication Law for Independence
  - **$\Pr(A \cap B) = \Pr(A) \times \Pr(B)$**
  - When no. events (k) > 2 then general law
    - **$\Pr(A_1 \cap A_2 \cap \ldots \cap A_k) = \Pr(A_1) \times \Pr(A_2) \times \ldots \times \Pr(A_k)$**

- Addition Law for Independence
  - **$\Pr(A \cup B) = \Pr(A) + \Pr(B) \times \Pr(\bar{A})$**
  - Pr(A or B) = Pr(A occurs) + Pr(B occurs and A does not occur)

- **Law does not apply for Dependent events**
  - **$\Pr(A \cap B) \neq \Pr(A) \times \Pr(B)$**

Independence and Probability

- Example revisited:
  - Hypertension screening program in 2-parent households
  - Event Assignment
    - A = Parent 1 DBP ≥ 95
    - B = Parent 2 DBP ≥ 95
  - Suppose the following:
    - $Pr(A) = 0.10$
    - $Pr(B) = 0.20$

  - **Challenge: What do we mean when we ask for the $Pr(A \cap B)$?**

  - **Challenge: Assume that $Pr(A \cap B) = 0.05$. Are the two events independent?**

Conditional Probability

- Suppose we are conducting breast cancer screening among older women.
- Let A = {has breast cancer} and B = {mammogram+}
- We are interested in **the probability that a woman has breast cancer <u>given</u> she has a positive mammogram.**
- When **we first condition on the occurrence of B and then assess the probability of A,** we are computing a conditional probability.
- Conditional Probability denoted as **Pr(A|B).**

Conditional Probability

- Mathematically:$\Pr(A|B) = \dfrac{\Pr(A \cap B)}{\Pr(B)}$

- If A and B are **independent events**, $\Pr(A|B) = \Pr(A)$
  - How would you interpret this law for the non-biostatistician?

- Challenge: What is the **complement of Pr(A|B)?**
  - Is that the same as Pr(B|A)?

Conditional Probability

- We can derive Pr(B|A) from the Pr(A|B).

- Three step process:
  - First: Define Pr(B|A) using our knowledge about conditional probabilities.

  - Second: We can obtain the Pr(B∩A) from our definition of Pr(A|B)?
    - Note that Pr(B∩A) = Pr(A∩B).

  - Third: We define Pr(A) with respect to two intersections.
    - Pr(A∩B) and Pr(A∩ $\overline{\mathrm{B}}$)
    - Known as the **Total Probability Rule**.

Bayes' Theorem

- Once we put our three steps together, we will have derived Bayes' Theorem.

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

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

$$\Pr(B|A) = \frac{\Pr(A|B) \text{ x } \Pr(B)}{\Pr(A \cap B) + \Pr(A \cap \bar{B})}$$

Total Probability Rule

$$\mathbf{\Pr(B|A) = \frac{\Pr(A|B) \text{ x } \Pr(B)}{\Pr(A|B) \text{x} \Pr(B) + \Pr(A|\bar{B}) \text{ x} \Pr(\bar{B})}}$$

Probability Challenge

- If the sensitivity of a screening test is 70%, the specificity of the test is 95%, and the prevalence of the disease of interest in the population is 10%, what is the positive predictive value of the test?

PART 2.2

Probability Distributions

I highly recommend that you read Chapter 3 of *"Introductory Statistics for the Life and Biomedical Sciences" by Julie Vu, David Harrington, and OpenIntro:* *https://leanpub.com/biostat*

Probability Distributions

- For statistical inference, we often make the following assumption:
  - Data are **a random sample selected from a population** and that the **distribution of the population has a known theoretical form.**

  - We call this distribution a **frequency or probability distribution.**

  - A **probability distribution** is used to calculate the theoretical **probability of different values occurring in the population**.

Types of probability distributions

- Many types of probability distributions and which ones we use **depend on the type of data** with which we are working.

- **Discrete distributions** can model discrete data (e.g. number of heart attacks)

- **Continuous distributions** can model continuous data (e.g. serum cholesterol levels)

Properties of Discrete Probability Distributions

- The probability that a random variable X can take a specific value x is Pr(X = x) sometimes denoted as p(x).

- p(x) is non-negative.

- The **sum** of p(x) over all possible values of x is 1, where:

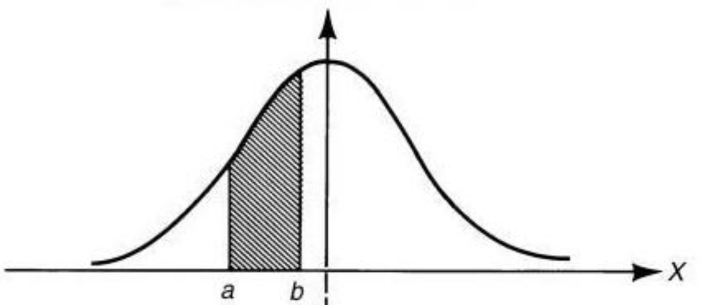
$$\sum_{x=0}^{x=\infty} p(x) = 1\,.$$

- Recall: 0 ≤ p(x) ≤ 1.

Properties of continuous probability distributions

- The probability that x is between two points a and b is

$$p(a \le x \le b) = \int_a^b f(x)dx$$



- It is non-negative

- The **integral** of the probability function is 1, that is,

$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

Consider this scenario

- We have 3 unrelated patients have been prescribed, by their neurologist, to take an experimental drug X. Any individual taking drug X has a 50% chance of having a seizure.

- 1 out of the 3 patients has a seizure.

- How can we determine the probability of our outcome in this scenario?

Introduction to the **Binomial distribution**
(a common distribution for binary data)

- A Bernoulli random variable, X, takes on only two values: 0 and 1, with probabilities 1-p and p, respectively.
  - A patient either has a seizure or does not have a seizure.
    - **Pr(Seizure) = p** and **Pr($\overline{\text{Seizure}}$) = Pr(No Seizure) = 1 - p**

- Usually Bernoulli random variables are used to indicate a dichotomous outcome for health research (e.g. **success or failure** of an intervention, staying alive or dying, disease remission or relapse, etc.)

The Binomial distribution

- The sum of Bernoulli random variables gives rise to the Binomial distribution
  - *n* independent trials (*n* patients prescribed to drug X).
  - Each trial has a dichotomous outcome (e.g.“seizure” or “no seizure”)
  - The probability of “success”, *p*, remains the same for each trial.

- Express getting *x* “successes” in *n* trials mathematically as:

$$\Pr(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

- Where $\binom{n}{x}$ is the no. of ways you can have *x* successes and *(n – x)* failures

Back to our prior binomial scenario

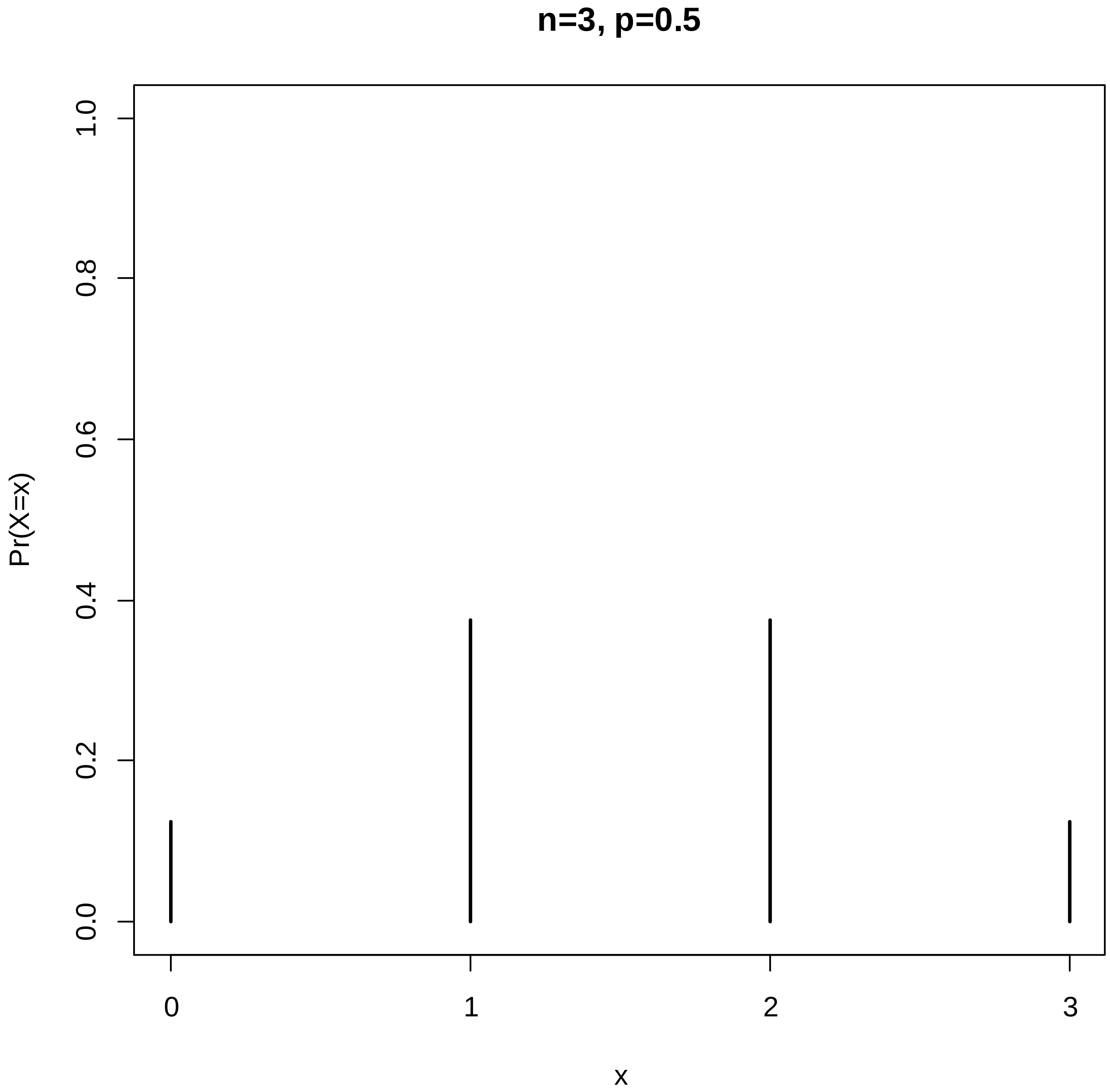
- Example for determing $\binom{n}{x}$
  - 3 unrelated patients prescribed to drug X (n=3)
    - Define Seizure as a success (1) and No Seizure as a failure (0)
  - How many ways can we be successful once (x = 1) in the three trials?
    - 1 0 0
    - 0 1 0
    - 0 0 1
  - Calculated mathematically as:

$$\frac{n!}{x!(n-x)!} = \frac{3 \text{ x } 2 \text{ x } 1}{1 \text{ x } (2 \text{ x } 1)} = 3$$

Binomial scenario cont'd

- Now lets find the probability of 1 out of 3 patients having a seizure

$$\Pr(X = 1) = \binom{3}{1} 0.5^{1} (1 - 0.5)^{3-1} = 0.375$$


n=3, p=0.5
Pr(X=x)
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
x

The Binomial distribution

Properties

- Has two parameters π and N
- Has mean $\mu = N\pi$
- Has variance $\sigma^2 = N\pi(1-\pi)$

Consider this scenario…

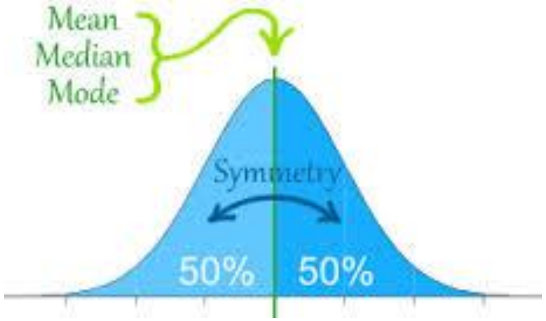
- Prior research has posited that the amount of time in a day that children spend in the upright position is distributed normally and that the average time spent in an upright position is about 5.4 ± 1.3 hours, on average.

- How can we use this information to determine the probability that a randomly selected child spends less than 5 hours in the upright position?

Introduction to the **Normal Distribution** (a common continuous distribution)

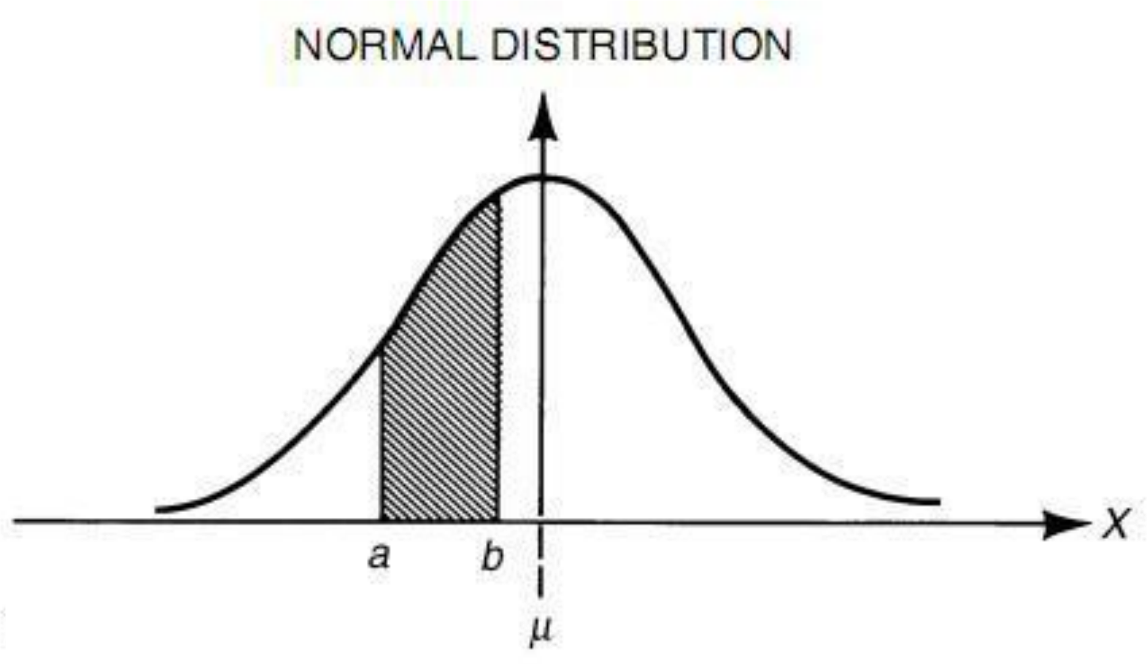
Properties:

1. Bell shaped
2. Symmetric around the mean μ
3. Mean, median and mode coincide
4. Total area under the curve=1
5. Completely determined by parameters **μ and σ**



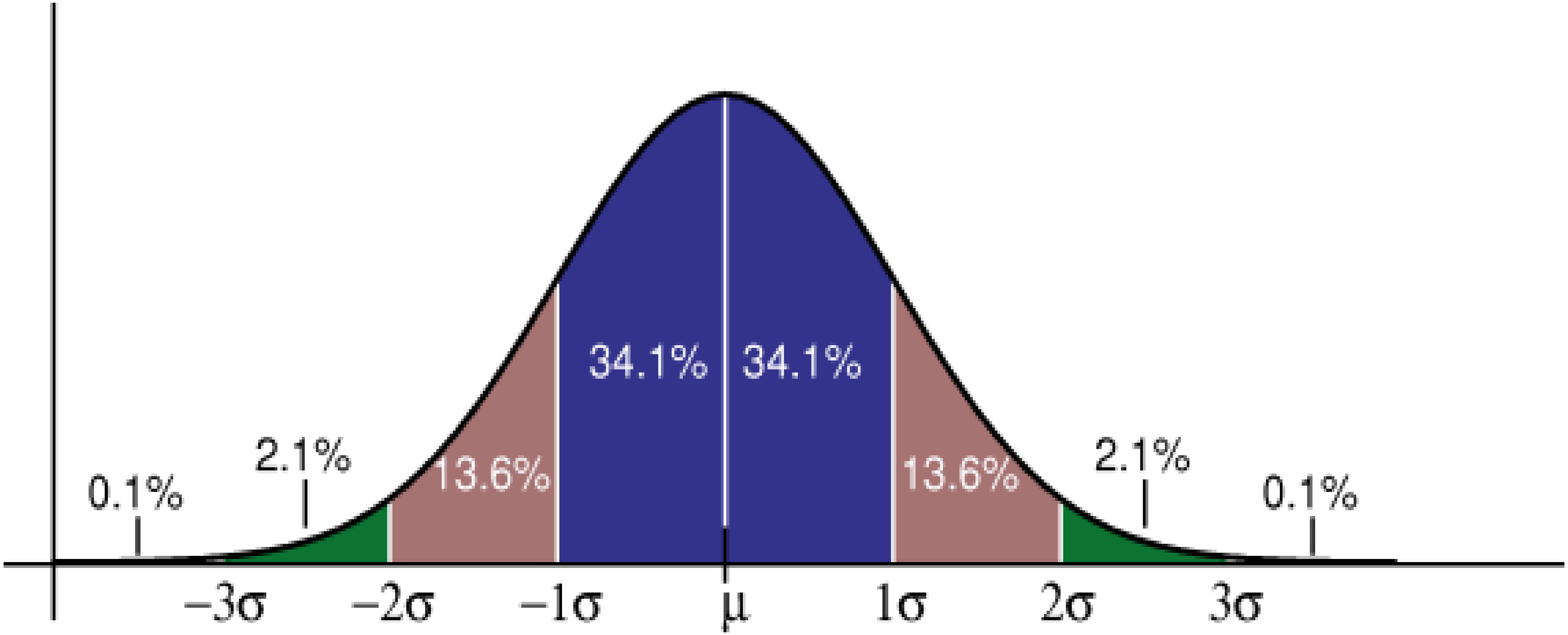
A continuous random variable Y has a Normal Distribution function N(μ, σ) if

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



$$p(a \le x \le b) = \int_{a}^{b} f(x)dx$$

Normal Distribution



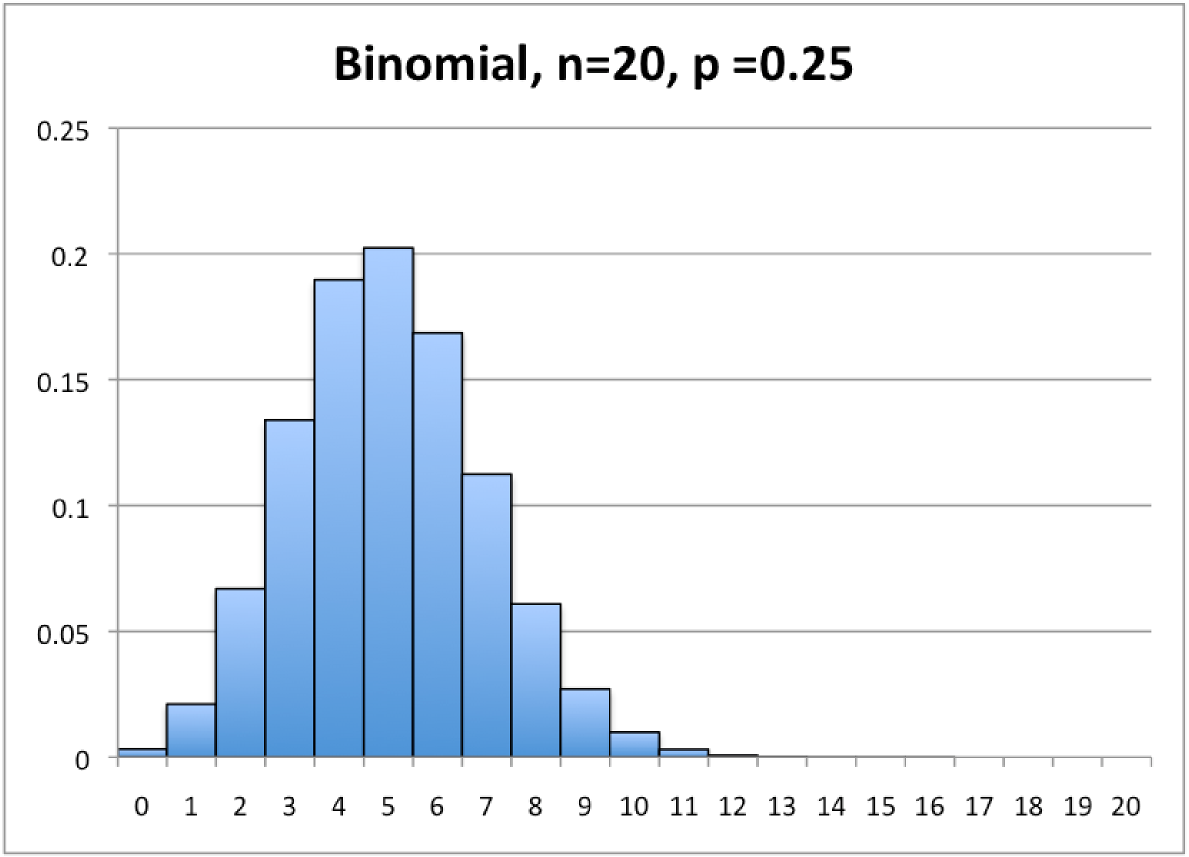
68.26%

95.44%

99.74%

Link between binomial and normal distributions

As *n* (number of trials) increases the Binomial distribution converges to the **Normal Distribution.**



Standard Normal Distribution

- The Normal Distribution function is quite common but very complicated.

- To ease the complexity, we can **transform any N($\boldsymbol{\mu}, \boldsymbol{\sigma}$) into the Standard Normal Distribution**.

- The Standard Normal Distribution is a special Normal distribution with $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\sigma} = \mathbf{1}$.

- There are known probabilities associated with any value under the **N(0,1)**.

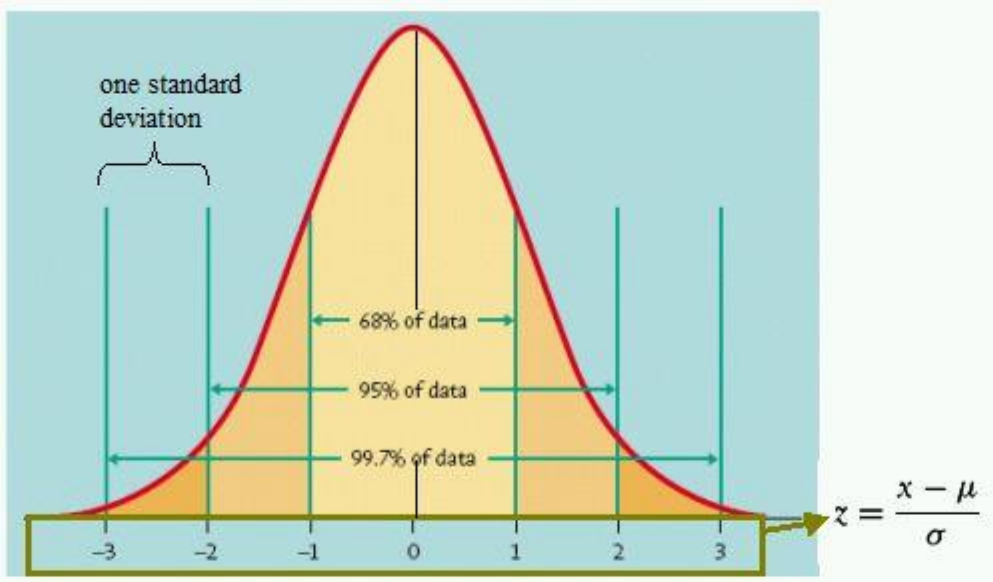
Standard Normal Distribution

If $Y \sim N(\mu,\sigma^2)$ then to standardize Y we do the following,

$$Z = \frac{(Y-\mu)}{\sigma} \sim N(0,1)$$

1. Has mean $\mu = 0$ and variance $\sigma^2 = 1$
2. It is tabulated
3. $P(-a \leq Z \leq a) = P(Z \leq a)-P(Z \leq -a)$
4. By symmetry: $P(Z \geq a)=P(Z \leq -a)$
5. By sum to 1 property: $P(Z \geq a)=1- P(Z < a)$

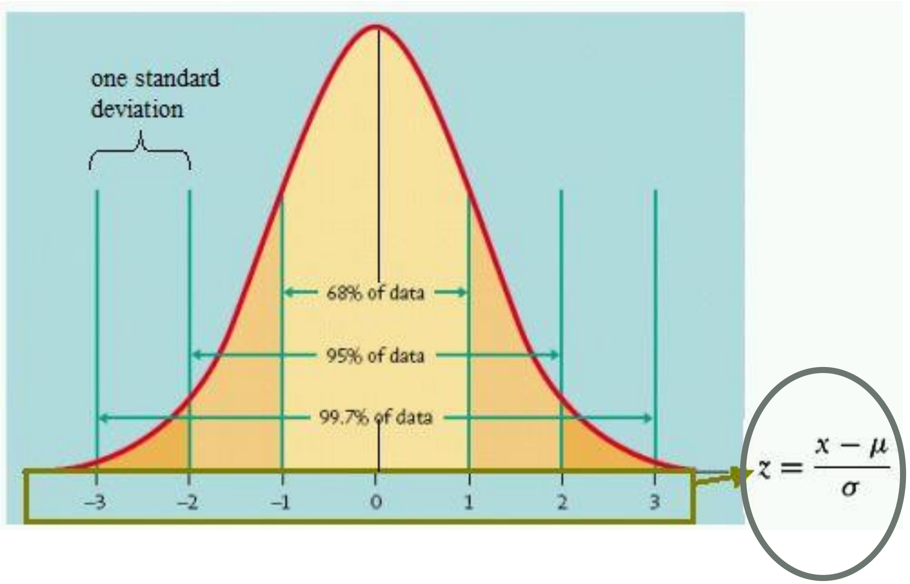
Standard Normal Distribution



Back to our prior scenario

The amount of time in a day children spend in the upright position (X) is distributed normally with mean=5.4 hours and standard deviation = 1.3 hours.

1. What is the probability that a randomly selected child spends less than 5 hours in the upright position



- **$X \sim N(5.4, (1.3)^2)$**

- **thus, $Z = \frac{5-5.4}{1.3} \sim N(0,1)$**

- **$P(X<5) = P\left(Z < \frac{5-5.4}{1.3}\right) = P(Z<-0.31) = 0.3783$**

**Note that the probabilities for most common distributions including the standard normal distribution can be obtained using most statistical softwares (R, Python, SAS, etc.)*

END OF PART TWO

FUNDAMENTALS OF (BIO)STATISTICS PART THREE:
STATISTICAL INFERENCE FOR CONTINUOUS DATA

Emma K. T. Benn, DrPH, MPH (*she/her)*
Associate Professor
Founding Director, Center for Scientific Diversity
Director of Data Science Training and Enrichment, Graduate School of Biomedical Sciences
Center for Biostatistics & Department of Population Health Science and Policy
emma.benn@mountsinai.org
Twitter: @EKTBenn

***Note that these slides have been adapted from the NIGMS-funded Applied Statistics in Biological Systems (ASIBS) Short Course (Co-PIs: Emilia Bagiella & Emma Benn) – www.asibs-statistics.org/online-lectures .**

Mount
Sinai

PART 3.1

Point estimates and confidence intervals

I highly recommend that you read Chapter 5 of *“Introductory Statistics for the Life and Biomedical Sciences” by Julie Vu, David Harrington, and OpenIntro:* [*https://leanpub.com/biostat*](https://leanpub.com/biostat)

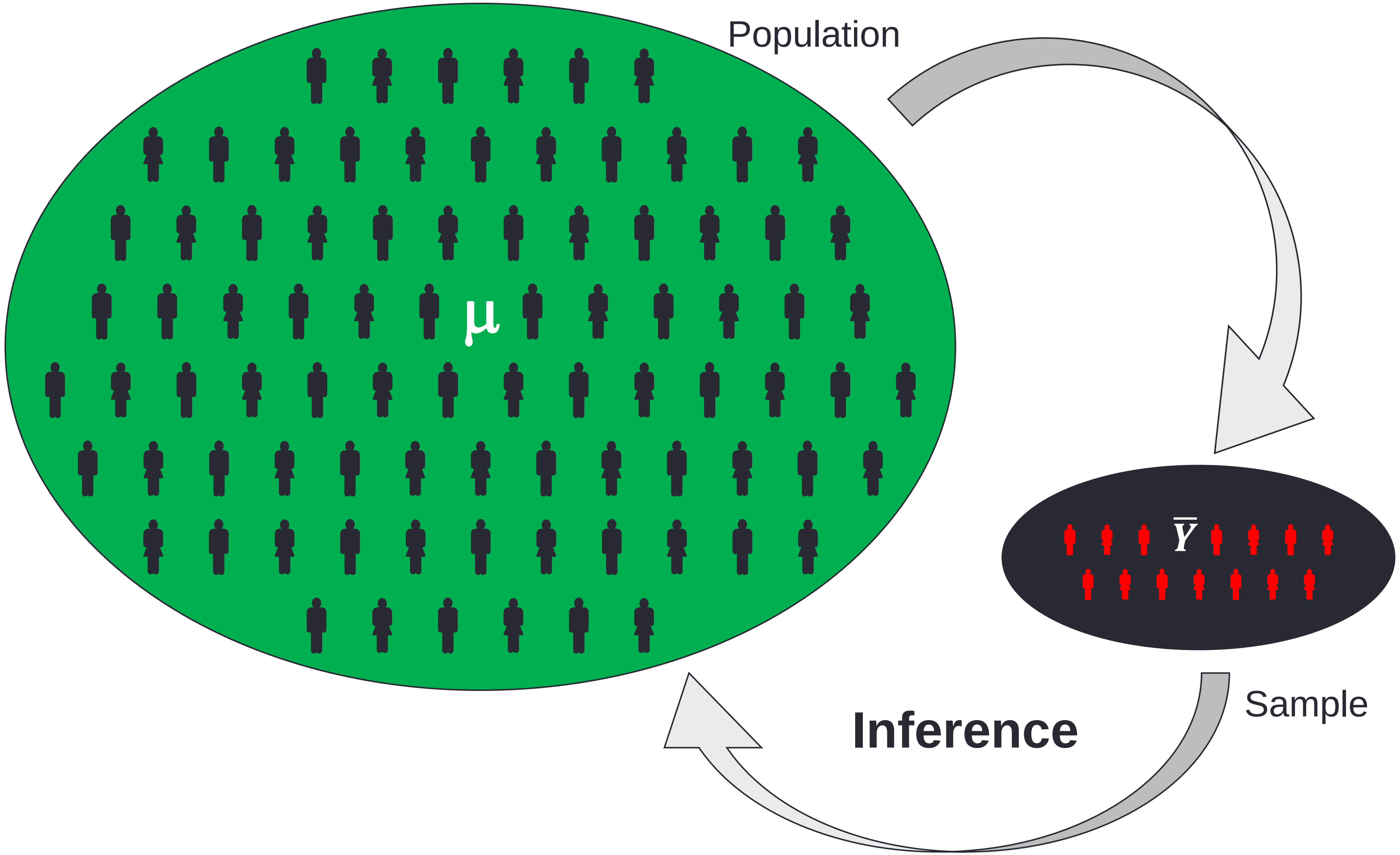
What is Statistical Inference?

Statistical inference allows to draw conclusions based on observed data.

A generalization made about a larger group or population from the observation of a sample of that population.

Note that this is inevitably an imprecise process…

Inference from the sample



Statistical inference

- Two Approaches:
  - **Estimation** (point and interval estimation)
  - **Hypothesis Testing**

Estimation: Underlying Idea

We are interested in the value of a specific parameter in the population (e.g. mean cholesterol level in hypertensive patients), but we can only observe a small portion of the whole population (the sample)

Goal

To give the most **precise estimate** of the parameter of interest by looking at the sample

Point Estimates

- One way to make inference about a population parameter is to use the sample point estimate of that parameter
- **Point estimates**: summary statistics from the sample that are used to estimate the parameter of interest.

$$\boxed{\begin{array}{c} \boxed{\bar{x} \rightarrow \mu} \\ s \rightarrow \sigma \\ p \rightarrow \pi \end{array}}$$

Example

We are interested in the proportion of time in a day children spend in the upright position.

In a random sample of 10 children selected from the larger population we find that the mean time children in the sample spend in the upright position is = 3 hours.

We could at this point conclude that on average children spend 3 hours a day in the upright position.

The truth is…

We do not know what the mean number of hours children spend in the upright position is. Therefore, we try to "guess" it by looking at a sample from this population and making inference from this sample.

During this process we aim at:

1. Giving the most precise estimate of that mean
2. Making the smaller error possible in stating our estimate is the true mean

Can we do better than that?

Recall

- The distribution of a statistic calculated from a sample drawn a random from the population of interest is called the **statistic's sampling distribution**

- The sampling distribution of the sample mean is (by virtue of the Central Limit Theorem) distributed as Normal with mean equal to the population mean μ and variance equal to σ^2/n

- Therefore, we can make inference about the mean of the population using the distribution of the sample mean

Why is this important?

Because we want to answer questions like:

If we draw a random sample of *n* units from the population of interest, what is the probability that the mean of this sample will be between two reasonable values?

Given an unknown parameter of interest in the population, with what precision can we estimate it from a random sample of size *n*?

If we draw a random sample of size *n*, what is the probability that that sample comes from a population with mean μ?

Interval estimates

- A point estimate consist of a single value that is used to estimate the parameter of interest (say, the mean)
- An interval estimate gives us a *range of plausible values* of the population parameter.
- The confidence level describes the uncertainty associated with the *sampling method*.

Interval Estimation

- Add more information to our point estimate.
- Interval estimates give us a **range of plausible values** for the population parameter.

Interval estimates

- Consist of two numerical values defining an interval that, with a specific degree of confidence, we feel includes the parameter of interest
- A 100(1-α)% confidence interval for a parameter θ is a random interval, based on the data, such that

$$P(L \leq \theta \leq U) = 1- \alpha$$

Degree of confidence

- The **degree of confidence** is determined by us through a parameter called **α**

- A 100(1-α)% confidence interval for a parameter θ is a random interval, based on the data, such that:

$$\mathbf{Pr(L \le \theta \le U) = 1- \alpha ,}$$

Where: L = lower limit**, U** = upper limit

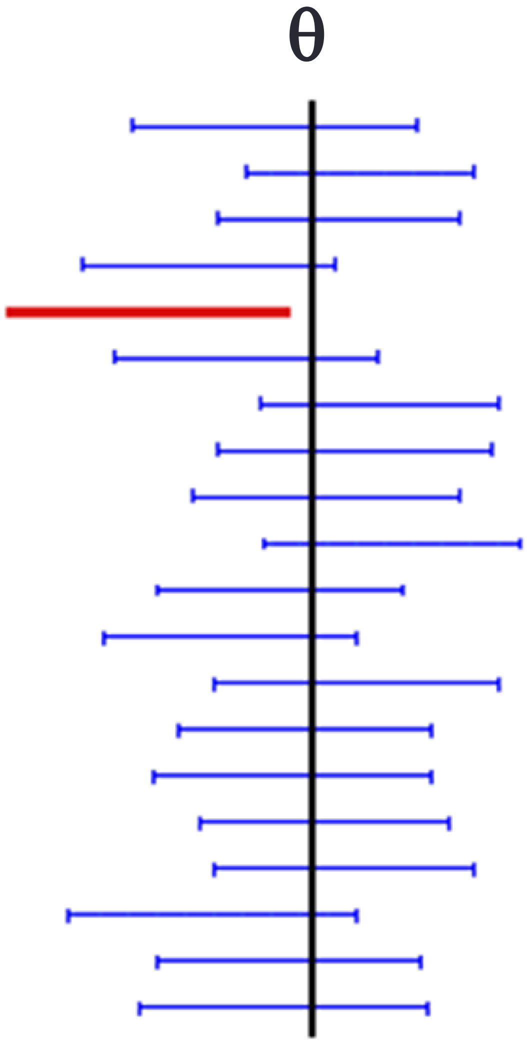
α is our preset **confidence level**.

Degree of confidence

$\Pr(L \leq \theta \leq U) = 1-\alpha$

If we keep drawing samples of the same size, calculate the sample statistic and build a confidence interval, **1- α% will contain the true parameter, α% of them will not.**

For α=0.05 or 5%, 95% of the intervals will contain the true parameter, 5% of them will not.



Form of a confidence interval

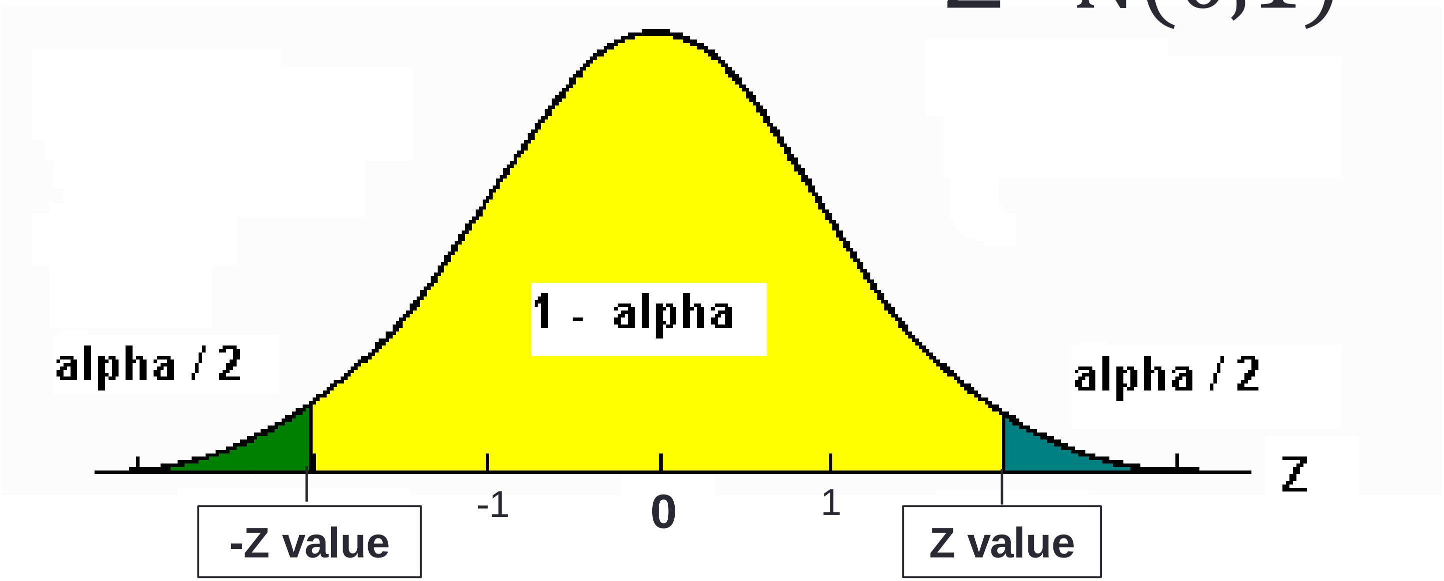
$$\begin{pmatrix}\text{Point}\\ \text{Estimate}\end{pmatrix} \pm \begin{pmatrix}\mathit{Confidence}\\ \mathit{Level}\end{pmatrix}\begin{pmatrix}\mathit{s.e.of}\\ \mathit{estimate}\end{pmatrix}$$

Note:

1. The point estimate is a **statistic** calculated from the sample (e.g. the sample mean)
2. The confidence level depends on the sampling distribution of the statistic

Using the Normal Distribution

$Z \sim N(0,1)$



Confidence level (1- α)	Z value
90%	1.65
95%	1.96
99%	2.58
99.9%	3.291

Confidence interval for the sample mean

Recall:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

A (100)(1-α)% Confidence interval for the sample mean has the form

$$\bar{X} \pm z_{(1-\alpha/2)} s.e._{\bar{X}} = \bar{X} \pm z_{(1-\alpha/2)} \frac{\sigma}{\sqrt{n}}$$

Interpretation

- Probabilistic: In repeated sampling from a normally distributed population 100(1-α)% of the intervals of the form $\bar{X} \pm z_{(1-\alpha/2)} \frac{\sigma}{\sqrt{n}}$ will in the long run contain the true value of μ
- Practical: We are 100(1-α)% confident that the computed interval $\bar{X} \pm z_{(1-\alpha/2)} \frac{\sigma}{\sqrt{n}}$ contains the population mean μ

Wrong Interpretation!

There is a 100(1-α)% chance that the population parameter falls between the limits of the confidence interval. This is incorrect. The population parameter, is a constant, not a random variable. Therefore we cannot make probabilistic statements about it.

Note

- The confidence interval is symmetric around the sample mean, not around the population mean.
- In fact, the confidence interval may not contain the population mean.

What if we do not know σ^2?

- When the standard deviation of the population σ is not known (most times) we can use its estimate from the sample **s**
- If n is large then we can use the normal distribution to determine the confidence level
- The formula for a 100(1-a)% CI for the population mean then becomes:

$$\overline{X} \pm z_{(1-\alpha/2)} \frac{s}{\sqrt{n}}$$

What if we do not know σ²?

- When n is small (<30)

$$\frac{\bar{X} - \mu}{s / \sqrt{n}} \sim t_{n-1}$$

- We use the Student’s t-distribution to determine the confidence level
- The formula for a 100(1-α)% CI for the population mean then becomes:

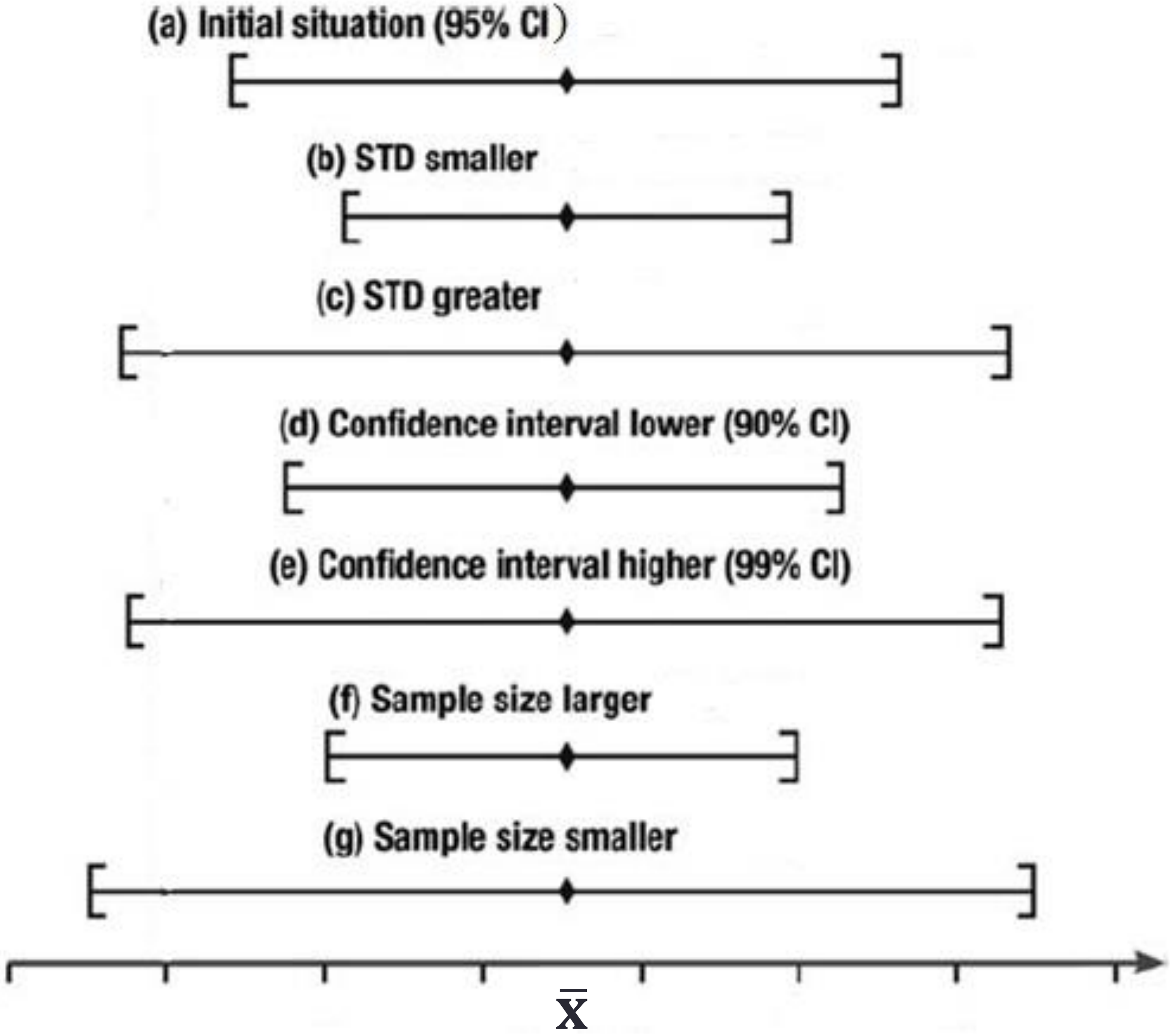
$$\bar{X} \pm t_{(1-\alpha/2, n-1)} \frac{s}{\sqrt{n}}$$

Notes

- We are not restricted to calculating 95% confidence intervals. We can calculate 90% confidence intervals, 99% confidence intervals, etc.

- The higher the desired confidence level the wider will be the confidence interval

- The smaller the sample size, the wider the confidence interval


(a) Initial situation (95% CI)
(b) STD smaller
(c) STD greater
(d) Confidence interval lower (90% CI)
(e) Confidence interval higher (99% CI)
(f) Sample size larger
(g) Sample size smaller
$\bar{X}$

PART 3.2

Fundamentals of hypothesis testing

I highly recommend that you read Chapter 5 of *“Introductory Statistics for the Life and Biomedical Sciences” by Julie Vu, David Harrington, and OpenIntro:* *https://leanpub.com/biostat*

Example

- There is reason to believe that the average cholesterol level in children whose father died of heart disease is **different than that of the general children population, which is claimed to be 160 mg/dL**.

Data collection

- A **sample mean of 177 mg/dL** is observed in a **random sample of 25 children** whose fathers died of heart disease.

Question:

- Do children whose father died of heart disease have the same total cholesterol level as children in the general population?

Hypothesis testing approach

- Instead of using the sample to estimate the unknown population parameter (with some precision) we first make a (smart) guess of what the value of the parameter may be and then we use the sample to determine how good our guess is.
- We call the null hypothesis (H_0) the statement that we want to test or the value of the parameter under our guess

Hypothesis Testing

- Does the information we obtained through our sample support the claim/hypothesis about the parameter?
  - What is the probability of observing a sample mean of 177 mg/dL, if the sample indeed comes from a population with mean μ = 160 mg/dL?

- Hypothesis testing is a method for **testing a claim or hypothesis about a parameter** in a population, by analyzing data from a random sample.

"... the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only to give the facts a chance of disproving the null hypothesis."

R. A. Fisher*

**Note that while I acknowledge Fisher’s contributions to the field of statistics, I vehemently oppose his support of racism and eugenics.*

Elements of hypothesis testing

- Random sample: a subset of units form a population such that each unit has the same probability of being selected
- Null hypothesis H_0: A hypothesis about a population (or parameter) of interest, generally chosen to represent the "status quo" (the default, no change, no difference)

Elements of hypothesis testing

- Alternative hypothesis H_A: Hypothesis specifying something different than the null
- Test statistic: A decision rule for choosing between H_0 and H_A
- Error probabilities: The probability of making the right (or wrong) decision about the parameter of interest

General form of set of hypothesis

Let θ be the (unknown) parameter of interest (e.g. mean, variance, risk ratio, odds ratio, effect size etc.). We set our hypothesis as

$H_0 : \theta = \theta_0$		$H_0 : \theta \geq \theta_0$		$H_0 : \theta \leq \theta_0$
$H_A : \theta \neq \theta_0$	or	$H_A : \theta < \theta_0$	or	$H_A : \theta > \theta_0$

The null and alternative hypothesis are ALWAYS complementary and mutually exclusive.

The **outcome** of a statistical test can be either

- “Reject H_0 in favor of H_1”

or

- “Fail to reject H_0”.

NOTE: If we fail to reject the null hypothesis, it does not mean that the null hypothesis is true or correct. It just means that, based on the data, we do not have sufficient evidence to support the claim stated in the alternative hypothesis.

Different outcomes of hypothesis testing

- We **fail to reject H_0** when **H_0 is really true**.
- We **fail to reject H_0** when **H_A is really true**.
- We **reject H_0** when **H_0 is really true**.
- We **reject H_0** when **H_A is really true**.

Error probabilities

α = Type I error: the probability of rejecting the null hypothesis when it is actually true

β = Type II error: the probability of accepting the null hypothesis when it is false

$1-\beta$= Power: the probability of correctly rejecting the null hypothesis

α is also called the significance level of the test

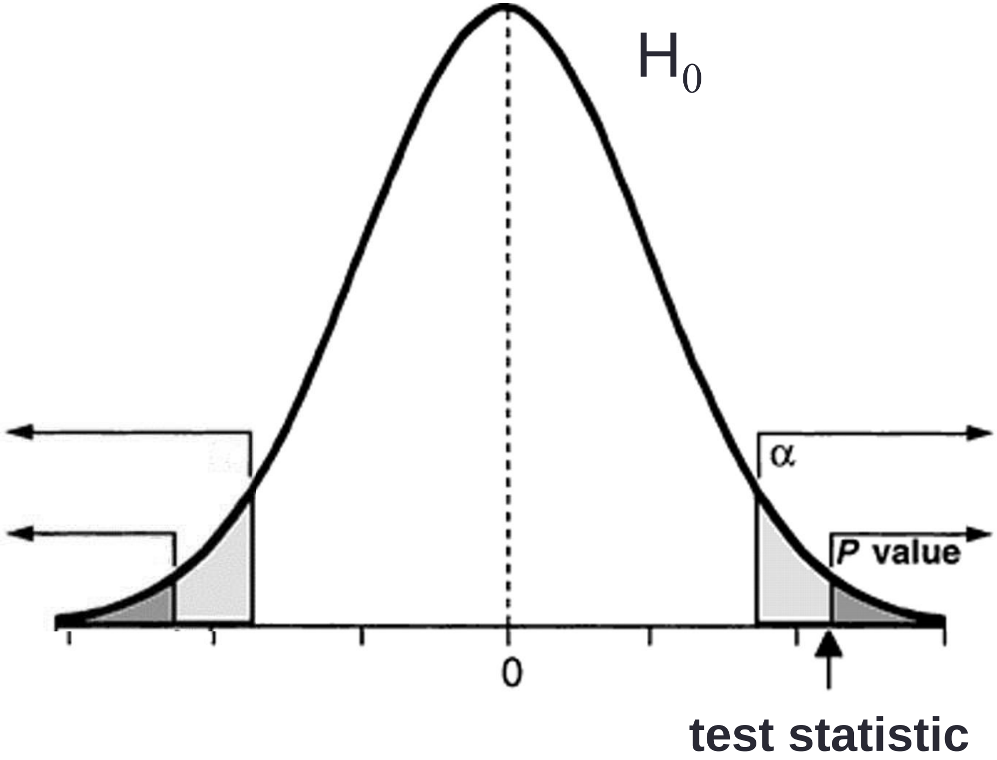
	H_0 is true	H_1 is true
Accept Null Hypothesis	Right decision	Wrong decision Type II Error
Reject Null Hypothesis	**Wrong decision Type I Error**	Right decision

Decision Rule

- **Logic:** reject the null hypothesis if the sample data are not consistent with the null hypothesis.
  - Our sample data is not consistent with null hypothesis if our test statistic has a very low chance of occurring if the null hypothesis is true

The P-value

- The **p-value** is the conditional probability of obtaining a value of the test statistic as extreme or more extreme than the one observed**, under the null hypothesis.**



Remarks

- We cannot control both α and β.
- A common strategy is to fix α at an acceptable level and then choose a test procedure that maximizes the power.
- For fixed α the power of the test can be increased by increasing the sample size
- For fixed sample size reducing α will reduce the power of the test

Example

- Infant blood pressure (BP) was collected during the first week of life in the newborn nursery. One question is **whether the level of consciousness of the infant affect BP**?

- We have the following result for the SBP of the first week for **two different group of infants**:

Level of consciousness	n	Mean	Sd
Quiet sleep	64	81.9	9.8
Awake and quiet	175	86.1	10.3

Assumptions

- Two sample test of means
  - The samples consist of **independent observations** drawn at random from the respective populations of interest.
  - The **two samples are independent from each other**.
  - The samples are drawn from **normally** distributed populations.
  - The two populations **have equal variance**.

Test Statistic (Equal Variance)

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)_{H_0}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1 + n_2 - 2)}$$

- The quantity **Sp is called the pooled standard** error and is a weighted average of the estimated variances of the two populations (assumed to be equal).

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

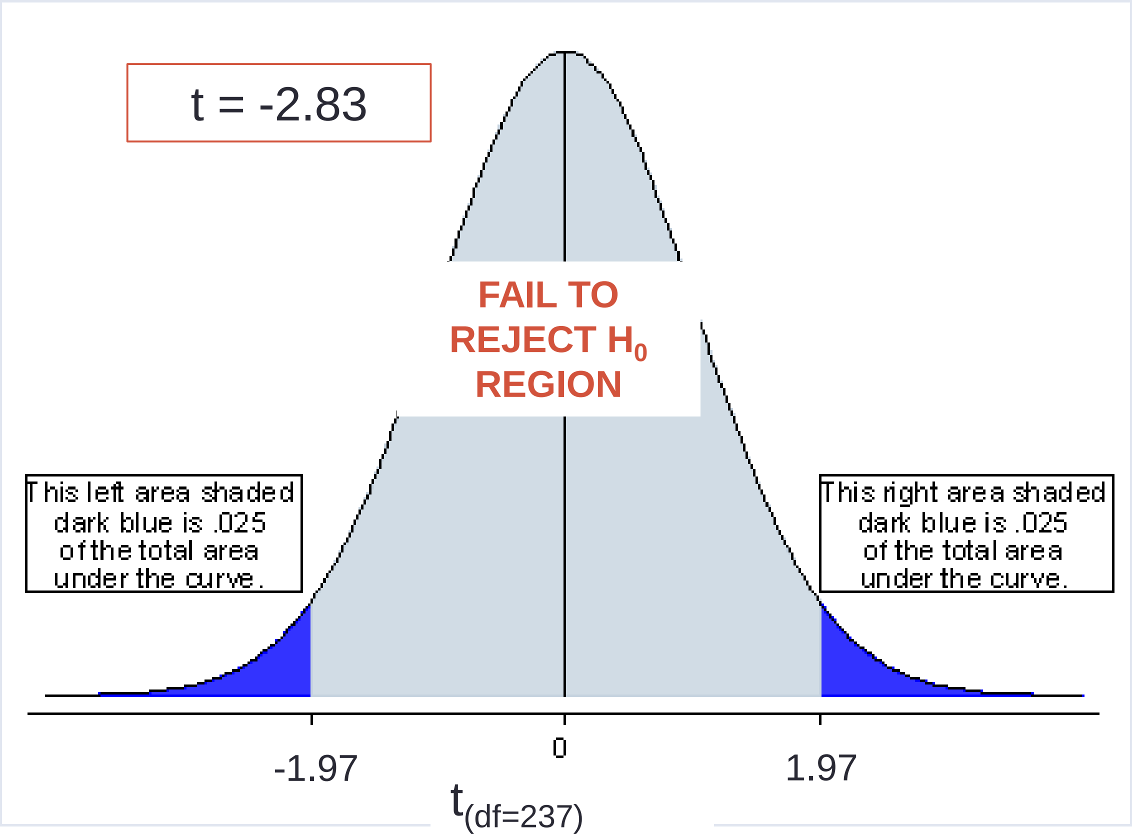
Test Statistic for infant BP example

$$s_p = \sqrt{\frac{(64-1)9.8^2 + (175-1)10.3^2}{64+175-2}} = 10.17$$

$$t = \frac{(81.9-86.1)-(0)_{H_0}}{10.17\sqrt{\frac{1}{64}+\frac{1}{175}}} = -2.83 \sim t_{(239-2=237)}$$

Quiet Sleep Group (n = 64, $\bar{x}$ = 81.9, s = 9.8)
Awake and Quiet Group (n=175, $\bar{x}$ = 86.1, s = 10.3)

Decision rule



Decision and Conclusion

- Since the test statistic exceeds the critical value for α=0.05 we reject the null hypothesis that the BP of infants with different level of consciousness is equal.

- Our sample does not support the hypothesis that blood pressure is the same in infants with different level of consciousness

Equal Variance Assumption

- The equal variance assumption allowed us to calculate the pooled standard deviation

- In the previous example, we assumed that the population variances were equal given the similarity between the sample standard deviations/variances of each group.

- However we usually **need to test whether the two groups meet the equal variance** assumption.
  - There is a specific test (the F test) that tells you whether we are meeting this assumption.
  - If you fail to meet the equal variance assumption, you will need to conduct a two-sample test of means assuming unequal variance.

Test for Equal Variance

- A rigorous way to make sure the equal variance assumption is met is to test the hypothesis

$$H_0 : \sigma^2{}_1 = \sigma^2{}_2 \quad \text{vs.} \quad H_A : \sigma^2{}_1 \neq \sigma^2{}_2$$

- If the F test for variance is **not significant** the null hypothesis of equality of variances cannot be rejected
  - We then proceed and calculate the **pooled variance**

- If the F test for variance is **significant** we reject the null hypothesis of equality of variance
  - We cannot assume that the two populations have equal variance
  - We cannot calculate the pooled variance

Two Sample Tests: Unequal Variance

- When we cannot assume equal variance between two populations we must use a different test statistic that takes into account both sample variances

$$t' = \frac{(\overline{X}_1 - \overline{X}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- This test statistic does **NOT** follow the t-Distribution

Paired Comparisons

- What if the groups are not independent?
  - Matched pairs
  - Before/After experiments
  - Repeated tests
- Need to take into account the correlation between samples

Example

IDEA: A high energy snack may increase brain performance.

HYPOTHESIS: Students perform differently if they eat a high energy snack.

EXPERIMENT: 10 students were asked to solve a set of math problems on an empty stomach. Two days later the same students were asked to solve another set of math problems after eating a high energy snack. Time to complete the test was recorded both times.

- RESULTS:
- With energy snack:
  - Scores = 72, 82, 93, 65, 76, 89, 81, 58, 95, 91
- Without energy snack:
  - Scores = 75, 79, 84, 71, 82, 91, 85, 68, 90, 92

Different approach from two sample test

Snack	No Snack
72	75
82	79
93	84
65	71
76	82
89	91
81	85
58	68
95	90
91	92
$\bar{x}_1$	$\bar{x}_2$

Difference with two sample test

Snack	No Snack
72	75
82	79
93	84
65	71
76	82
89	91
81	85
58	68
95	90
91	92
$\bar{x}_1$	$\bar{x}_2$

Snack	No Snack	Score Difference
72	75	-3
82	79	3
93	84	9
65	71	-6
76	82	-6
89	91	-2
81	85	-4
58	68	-10
95	90	5
91	92	-1
		$\bar{D}$

H_A: $\mu_1 \neq \mu_2$ **or** $\mu_1 - \mu_2 \neq 0$

$H_A : \mu_d \neq 0$

Example: Paired Comparisons

- Interested in the **individual differences**

$$H_0 : \mu_d = 0$$

$$H_A : \mu_d \neq 0$$

Where μ_d represents the mean of all differences $Y_i^{(1)}$-$Y_i^{(2)}$.

Test statistic

$$t = \frac{\overline{D} - \mu_d}{s_{\overline{D}} / \sqrt{n}} \sim t_{n-1}$$

- Note 1: *n* here is the **number of pairs**, not the total number of subjects
- Note 2: This is in practice a **one sample test** on the differences $Y_i^{(1)}$-$Y_i^{(2)}$.

Test statistic

$$t = \frac{\overline{D} - \mu_d}{s_{\overline{D}} / \sqrt{n}} = \frac{-1.5 - 0}{5.72 \;\; / \sqrt{10}} = -0.83 \sim t_{10-1}$$

P-value = 0.2142

Conclusion?

Does getting a high energy snack make students perform differently? *We have insufficient evidence to suggest that students perform differently.*

END OF PART THREE

FUNDAMENTALS OF (BIO)STATISTICS PART FOUR:
INTRODUCTION TO THE ANALYSIS OF CATEGORICAL DATA

Emma K. T. Benn, DrPH, MPH (*she/her)*
Associate Professor
Founding Director, Center for Scientific Diversity
Director of Data Science Training and Enrichment, Graduate School of Biomedical Sciences
Center for Biostatistics & Department of Population Health Science and Policy
emma.benn@mountsinai.org
Twitter: @EKTBenn

***Note that these slides have been adapted from the NIGMS-funded Applied Statistics in Biological Systems (ASIBS) Short Course (Co-PIs: Emilia Bagiella & Emma Benn) – www.asibs-statistics.org/online-lectures .**

Mount
Sinai

PART 4.1

Introduction to the analysis of categorical data

Measures of association

I highly recommend that you read Chapter 8 of *"Introductory Statistics for the Life and Biomedical Sciences" by Julie Vu, David Harrington, and OpenIntro:* *https://leanpub.com/biostat*

Analysis of Categorical Response Data

- Analysis of categorical data is primarily focused on analyzing **how categorical response variables (i.e., outcomes/dependent variables) are influenced by predictor variables**.
  - Categorical restriction only pertains to response variable.
  - While we do have categorical response variables that have >2 levels, this lecture will be focused on binary/dichotomous response variables.

Describing bivariate relationships between two binary categorical variables

- Motivating Example:
  - The Physician's Health Study was a 5-year randomized study of whether regular aspirin intake reduces mortality from cardiovascular disease. Every other day, physicians took either one aspirin or placebo. Of the 11,034 physicians taking placebo, 189 suffered a heart attack, as compared to 104 of the 11,037 physicians taking aspirin.

NEJM 318: 262-264, 1988.

Contingency Tables

- When we work with a single categorical variable, we can easily count the number of observations in each category and compute sample proportions.
- However, when we have ≥2 categorical variables, we need an effective way to display all possible combinations of outcomes along with their corresponding frequencies and probabilities.
- A table that cross-classifies two or more categorical variables to show all possible combinations of outcomes is a contingency table.

Contingency Tables: Two-way Contingency Table

The Physician’s Health Study was a 5-year randomized study of whether regular aspirin intake reduces mortality from cardiovascular disease. Every other day, physicians took either one aspirin or placebo. Of the 11,034 physicians taking placebo, 189 suffered a heart attack, as compared to 104 of the 11,037 physicians taking aspirin. *NEJM 318: 262-264, 1988.*

Two-way Contingency Table

	MI (Y=1)	**No MI (Y=0)**	
Aspirin (X=1)	104	10,933	11,037
Placebo (X=0)	189	10,845	11,034
	293	21,778	22,071

Contingency Tables: Two-way Contingency Table

	MI (Y=1)	No MI (Y=0)	
Aspirin (X=1)	104	10,933	11,037
Placebo (X=0)	189	10,845	11,034
	293	21,778	22,071

General Layout of the Contingency Table

Let **I** be the **no. of categories** for the **row variable**.
Let **J** be the **no. of categories** for the **column variable**.

-Let $\mathbf{n_{i.}}$ be the **marginal frequency** of people with the **i^{th} outcome**.

-Let $\mathbf{n_{.j}}$ be the **marginal frequency** of people with the **j^{th} outcome**.

-Let $\mathbf{n_{ij}}$ be the **joint frequency** of people with the **i^{th} and j^{th} outcomes**.

-Let $\mathbf{n_{..}}$ be the **total sample size**.

Contingency Tables: Two-way Contingency Table

	MI (Y=1)	No MI (Y=0)	
Aspirin (X=1)	104	10,933	11,037
Placebo (X=0)	189	10,845	11,034
	293	21,778	22,071

	MI (Y=1)	No MI (Y=0)	
Aspirin (X=1)	n_{11}	n_{12}	$n_{1.}$
Placebo (X=0)	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	$n_{..}$

General Layout of the Contingency Table

Let **I** be the **no. of categories** for the **row variable**.
Let **J** be the **no. of categories** for the **column variable**.

-Let $\mathbf{n_{i.}}$ be the **marginal frequency** of people with the **ith outcome**.

-Let $\mathbf{n_{.j}}$ be the **marginal frequency** of people with the **jth outcome**.

-Let $\mathbf{n_{ij}}$ be the **joint frequency** of people with the **ith and jth outcomes**.

-Let $\mathbf{n_{..}}$ be the **total sample size**.

Contingency Tables: Two-way Contingency Table

	MI (Y=1)	No MI (Y=0)	
Aspirin (X=1)	104	10,933	11,037
Placebo (X=0)	189	10,845	11,034
	293	21,778	22,071

	MI (Y=1)	No MI (Y=0)	
Aspirin (X=1)	n_{11}	n_{12}	$n_{1.}$
Placebo (X=0)	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	$n_{..}$

1. **What is the probability of MI?**

2. **What is the probability of having an MI and taking placebo?**

3. **What is the probability that aspirin-takers will have an MI?**

4. **What is the probability that placebo-takers will have an MI?**

Contingency Tables: Two-way Contingency Table

	MI (Y=1)	No MI (Y=0)	
Aspirin (X=1)	104	10,933	11,037
Placebo (X=0)	189	10,845	11,034
	293	21,778	22,071

	MI (Y=1)	No MI (Y=0)	
Aspirin (X=1)	n_{11}	n_{12}	$n_{1.}$
Placebo (X=0)	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	$n_{..}$

1. **What is the probability of MI?**

$$\Pr(Y=1) = \frac{n_{.1}}{n_{..}} = \frac{293}{22{,}071} = 0.013$$

Contingency Tables: Two-way Contingency Table

	MI (Y=1)	No MI (Y=0)	
Aspirin (X=1)	104	10,933	11,037
Placebo (X=0)	189	10,845	11,034
	293	21,778	22,071

	MI (Y=1)	No MI (Y=0)	
Aspirin (X=1)	n_{11}	n_{12}	$n_{1.}$
Placebo (X=0)	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	$n_{..}$

1. What is the probability of MI?

$$\Pr(Y=1) = \frac{n_{.1}}{n_{..}} = \frac{293}{22{,}071} = 0.013$$

2. What is the probability of having an MI and taking placebo?

$$\Pr(Y=1 \cap X=0) = \frac{n_{21}}{n_{..}} = \frac{189}{22{,}071} = 0.0086$$

Contingency Tables: Two-way Contingency Table

	MI (Y=1)	No MI (Y=0)	
Aspirin (X=1)	104	10,933	11,037
Placebo (X=0)	189	10,845	11,034
	293	21,778	22,071

	MI (Y=1)	No MI (Y=0)	
Aspirin (X=1)	n_{11}	n_{12}	$n_{1.}$
Placebo (X=0)	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	$n_{..}$

1. What is the probability of MI?

$$\Pr(Y=1) = \frac{n_{.1}}{n_{..}} = \frac{293}{22{,}071} = 0.013$$

2. What is the probability of having an MI and taking placebo?

$$\Pr(Y=1 \cap X=0) = \frac{n_{21}}{n_{..}} = \frac{189}{22{,}071} = 0.0086$$

3. What is the probability that aspirin-takers will have an MI?

$$\Pr(Y=1|X=1) = \frac{n_{11}}{n_{1.}} = \frac{104}{11{,}037} = 0.0094$$

Contingency Tables: Two-way Contingency Table

	MI (Y=1)	No MI (Y=0)	
Aspirin (X=1)	104	10,933	11,037
Placebo (X=0)	189	10,845	11,034
	293	21,778	22,071

	MI (Y=1)	No MI (Y=0)	
Aspirin (X=1)	n_{11}	n_{12}	$n_{1.}$
Placebo (X=0)	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	$n_{..}$

1. What is the probability of MI?

$$\Pr(Y=1) = \frac{n_{.1}}{n_{..}} = \frac{293}{22{,}071} = 0.013$$

2. What is the probability of having an MI and taking placebo?

$$\Pr(Y=1 \cap X=0) = \frac{n_{21}}{n_{..}} = \frac{189}{22{,}071} = 0.0086$$

3. What is the probability that aspirin-takers will have an MI?

$$\Pr(Y=1|X=1) = \frac{n_{11}}{n_{1.}} = \frac{104}{11{,}037} = 0.0094$$

4. What is the probability that placebo-takers will have an MI?

$$\Pr(Y=1|X=0) = \frac{n_{21}}{n_{2.}} = \frac{189}{11{,}034} = 0.017$$

Measures of Association

- While the studies generally are interested in examining associations between predictors and outcomes of interest, we must have a rigorous method by which we can assess the magnitude of these associations.

- However, in some cases, there are restrictions imposed on the estimable measures of association by the type of study design.

Measures of Association: Risk Difference

- Risk Difference (RD) is a measure of the absolute effect of the exposure, or the excess risk of disease attributable to the exposure.

- In the population:

$$RD = P(D|E) - P(D|E')$$

- Using sample probabilities will yield an unbiased estimate, $\widehat{RD}$.

- -1 ≤ RD ≤ 1

- Estimable in Cross-sectional and Prospective Studies.

Measures of Association: Risk Difference

- **RD = 0: NO ASSOCIATION**
  - Incidence of disease is the same for exposed and unexposed.
  - Removing the exposure will have no impact on the disease.

- **RD < 0: NEGATIVE ASSOCIATION**
  - The exposure reduces the incidence of disease (PROTECTIVE EFFECT).
  - If those who are not exposed become exposed, their incidence of disease will decrease by |RD| .

- **RD > 0: POSITIVE ASSOCIATION**
  - The exposure increases the incidence of disease.
  - If those who are not exposed become exposed, their incidence of disease will increase by RD.

- **Number Needed to Treat (NNT) = 1 / |RD|**
  - How many subjects need to be exposed/treated to prevent 1 sick patient?

Measures of Association: Risk Difference – standard error

Cross-sectional study examining relationship between osteoporosis (Osteo.) and severe periodontal disease (SPD1).

	SPD1 (Y=1)	No SPD1 (Y=0)	
Osteo. (X=1)	42	56	98
No Osteo. (X=0)	108	174	282
	150	230	380

Statistic	Value	95% Confidence Limits	
Risk Difference	0.046	-0.068	0.160

Measures of Association: Risk Ratio (Relative Risk)

- Risk ratio (RR) is a measure of the risk of disease in the exposed relative to that of the unexposed.

- In the population:

$$RR = \frac{P(D|E)}{P(D|E')}$$

- Sample probabilities will yield an unbiased estimate, $\widehat{RR}$.

- RR ≥ 0

- Estimable in Cross-sectional and Prospective Studies.

Measures of Association: Risk Ratio (Relative Risk)

- If RR = 1 ⇒ No association
  - The risk of disease is the same for exposed and unexposed.
- If RR < 1 ⇒ Negative association
  - The exposure decreases the risk of disease.
  - The exposure has a protective effect.
  - **The risk of disease in the exposed is (1-RR)% lower than among the unexposed.**
- If RR > 1 ⇒ Positive association
  - Exposure increases the risk of disease.

Measures of Association: Risk Ratio

Prospective study examining relationship between oral contraceptives (OC) and myocardial infarction (MI).

	MI (Y=1)	No MI (Y=0)	
OC (X=1)	13	4987	5,000
No OC (X=0)	7	9993	10,000
	20	14,980	15,000

Statistic	Value	95% Confidence Limits	
Relative Risk	3.7143	1.4829	9.3036

Measures of Association: Odds Ratio

- **Odds ratio (OR) is a measure of the odds of:**
  - exposure in the diseased relative to that of the non-diseased [exposure OR], or
  - disease in the exposed relative to that of the unexposed [disease OR].

- **In the population, the exposure OR is equivalent to the disease OR:**

$$OR = \frac{P(E|D)/[1 - P(E|D)]}{P(E|D')/[1 - P(E|D')]} = \frac{P(D|E)/[1 - P(D|E)]}{P(D|E')/[1 - P(D|E')]}$$

Measures of Association: Odds Ratio

- Sample probabilities will yield an unbiased estimate, $\widehat{OR}$.
- OR ≥ 0
- Estimable in Cross-sectional, Prospective, and Retrospective Studies.
- OR is a good approximation of the RR when the disease is rare.
  - Challenge: Why is the above statement true?

Measures of Association: Odds Ratio

- **If OR=1 ⇒ No Association**

- **If OR<1 ⇒ Negative Association**
  - Exposure is protective against disease.
  - The odds of disease in the exposed are (1-OR)% lower than among the unexposed.

- **If OR>1 ⇒ Positive Association**

Measures of Association: Odds Ratio

Retrospective study examining the relationship between age group at first birth and breast cancer (BC) among mothers.

	BC (Y=1)	No BC (Y=0)	
Age ≥ 30 (X=1)	683	1,498	2,181
Age < 30 (X=0)	2,537	8,747	11,284
	3,220	10,245	13,465

Statistic	Value	95% Confidence Limits	
Odds Ratio	1.5720	1.4214	1.7385

PART 4.2

Hypothesis testing for categorical data

I highly recommend that you read Chapter 8 of *“Introductory Statistics for the Life and Biomedical Sciences” by Julie Vu, David Harrington, and OpenIntro:* *https://leanpub.com/biostat*

Measures of Association to Tests of Association

- We have discussed parameters that can help us measure the magnitude of the association between two categorical variables, along with the uncertainty surrounding our estimates of those parameters.

- We will now discuss how to formally test our hypotheses.
  - H0: No association between E and D ($RD=0$ or $RR=1$ or $OR=1$; depends on study design).
  - H1: Association exists between E and D ($RD\neq0$ or $RR\neq1$ or $OR\neq1$; depends on study design).

Bivariate Tests of Association for Categorical Data

- But what would we expect if there were no association between exposure and disease?

- To answer this question, you must recall your basic probability rules.

Bivariate Tests of Association for Categorical Data

- If Exposure and Disease are independent then the following must be true when each variable consists only of 2 categories.

$$P(E \cap D) = P(E) \times P(D)$$

$$P(E \cap D') = P(E) \times P(D')$$

$$P(E' \cap D) = P(E') \times P(D)$$

$$P(E' \cap D') = P(E') \times P(D')$$

Bivariate Tests of Association for Categorical Data: Chi-Squared Test for Independence/Homogeneity

- The Chi-Squared Test allows us to test for an association between two categorical variables.

- However, with any hypothesis test, there are several assumptions:
  - Independent observations (observations must not be correlated)
  - Cell counts expected under the null hypothesis ($\hat{\mu}_{ij}$) must be greater than 5

Bivariate Tests of Association for Categorical Data: Chi-Squared Test for Independence/Homogeneity

How do we compute the cell counts expected under the null hypothesis (aka expected cell counts) μ_{ij} ?

	D	D’	
E	μ_{11}	μ_{12}	$n_{1.}$
E’	μ_{21}	μ_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	$n_{..}$

$$\pi_{ij} = \pi_i * \pi_j$$

$$\frac{n_{ij}}{n_{..}} = \frac{n_{i.}}{n_{..}} * \frac{n_{.j}}{n_{..}}$$

$$\hat{\mu}_{ij} = n_{..} * \frac{n_{i.}}{n_{..}} * \frac{n_{.j}}{n_{..}}$$

$$\boldsymbol{\hat{\mu}_{ij} = \frac{n_{i.} n_{.j}}{n_{..}}}$$

Bivariate Tests of Association for Categorical Data: Chi-Squared Test for Independence/Homogeneity

- As in any hypothesis test, our test statistic for the chi-squared test must allow for the examination of how far our results (what we’ve observed) deviate from what would be expected under the null hypothesis.

- Our test statistic (i.e. the Pearson chi-squared statistic) can be computed as follows:

$$\chi^2 = \sum \frac{\left(n_{ij} - \mu_{ij}\right)^2}{\mu_{ij}} \sim under\ H_0$$

**Note that while I acknowledge the contributions of Pearson to the field of statistics, I strongly oppose his support of the eugenics movement.*

Bivariate Tests of Association for Categorical Data: Chi-Squared Test for Independence/Homogeneity

Properties of the chi-squared distribution:

- Ranges from 0 to ∞ (concentrated over non-negative values).
- As with the student’s t distribution, it is defined by its degrees of freedom (df), with:
  - $df = (I-1)(J-1)$
  - $\mu = df$
  - $\sigma = \sqrt{2df}$
- As degrees of freedom increase, the distribution becomes more bell-shaped (normal).

Bivariate Tests of Association for Categorical Data: Chi-Squared Test for Independence/Homogeneity

Properties of the chi-squared distribution:

- Ranges from 0 to ∞ (concentrated over non-negative values).

- As with the student’s t distribution, it is defined by its degrees of freedom (df), with:
  - $\mu = df$
  - σ = $\sqrt{2df}$

- As degrees of freedom increase, the distribution becomes more bell-shaped (normal).

- Additionally…
  - The test statistic has a chi-squared distribution for large n
  - The chi-squared approximation improves as $\hat{\mu}_{ij}$ increases, with $\hat{\mu}_{ij}$≥5 being sufficient for this approximation.

Bivariate Tests of Association for Categorical Data: Chi-Squared Test for Independence/Homogeneity

- Conducting the **chi-squared test for homogeneity** for earlier mentioned retrospective study.
  - Hypotheses:
    - **H_0: Breast cancer and breast cancer-free populations are homogeneous with respect to age at first birth (OR=1).**
    - **H_1: "   " are not homogeneous "   ".**
  - Set $\alpha = 0.05$ and check your assumptions.

OBSERVED

	BC (Y=1)	No BC (Y=0)	
Age ≥ 30 (X=1)	683	1,498	2,181
Age < 30 (X=0)	2,537	8,747	11,284
	3,220	10,245	13,465

EXPECTED UNDER H0, $\hat{\mu}_{ij}$

	BC (Y=1)	No BC (Y=0)	
Age ≥ 30 (X=1)	**521.561**	**1659.439**	2,181
Age < 30 (X=0)	**2698.439**	**8585.561**	11,284
	3,220	10,245	13,465

Bivariate Tests of Association for Categorical Data: Chi-Squared Test for Independence/Homogeneity

- Conducting the **chi-squared test for homogeneity** for Motivating Example 3
  - Hypotheses:
    - **H_0: Breast cancer and breast cancer-free populations are homogeneous with respect to age at first birth (OR=1).**
    - **H_1: “ “ are not homogeneous “ “.**
  - Set $\alpha = 0.05$ and check your assumptions.

OBSERVED

	BC (Y=1)	No BC (Y=0)	
Age ≥ 30 (X=1)	683	1,498	2,181
Age < 30 (X=0)	2,537	8,747	11,284
	3,220	10,245	13,465

EXPECTED UNDER H0, $\hat{\mu}_{ij}$

	BC (Y=1)	No BC (Y=0)	
Age ≥ 30 (X=1)	**521.561**	**1659.439**	2,181
Age < 30 (X=0)	**2698.439**	**8585.561**	11,284
	3,220	10,245	13,465

Statistic	DF	Value	Prob
Chi-Square	1	78.3698	<.0001

Bivariate Tests of Association for Categorical Data: Chi-Squared Test for Independence/Homogeneity

- Some thoughts about the chi-squared test:
  - The test statistic is highly affected by sample size.
  - **The test statistic is not a measure of the strength of the association,** but of **the strength of the evidence against the null hypothesis of no association.**
  - No normality assumption for the underlying distribution from which the data come.
  - While we have focused on 2x2 tables, easily extends to mxn tables.

Bivariate Tests of Association for Categorical Data: Chi-Squared Test for Independence/Homogeneity

- While the chi-squared test is commonly used to test for an association between a categorical predictor and outcome of interest, however it is not an exact test procedure.

- When we do not meet the underlying assumption about our expected cell counts for a 2x2 table, we can instead use the **Fisher's Exact Test**.

- The Fisher's Exact test gives the exact levels of significance for any 2x2 table.

**Note that while I acknowledge the contributions of Fisher to the field of statistics, I strongly oppose his support of the eugenics movement.*

Bivariate Tests of Association for Categorical Data: Fisher’s Exact Test

- Motivating Example for the Fisher’s Exact Test:
  - Researchers are investigating the relationship between high salt intake and death from CVD. Given limited resources, researchers evaluated death records and classified cause of death as either CVD related or unrelated. They subsequently asked close relatives about the salt intake (low or high) of the deceased.

Bivariate Tests of Association for Categorical Data

OBSERVED

	CVD (Y=1)	No CVD (Y=0)	
High Salt (X=1)	2	23	25
Low Salt (X=0)	5	30	35
	7	53	60

EXPECTED UNDER NULL HYPOTHESIS

	CVD (Y=1)	No CVD (Y=0)	
High Salt (X=1)	2.917	22.083	25
Low Salt (X=0)	4.083	30.917	35
	7	53	60

Bivariate Tests of Association for Categorical Data

OBSERVED

	CVD (Y=1)	No CVD (Y=0)	
High Salt (X=1)	2	23	25
Low Salt (X=0)	5	30	35
	7	53	60

EXPECTED UNDER NULL HYPOTHESIS

	CVD (Y=1)	No CVD (Y=0)	
High Salt (X=1)	**2.917**	22.083	25
Low Salt (X=0)	**4.083**	30.917	35
	7	53	60

Since $\hat{\mu}_{11} < 5$ and $\hat{\mu}_{21} < 5$, we must conduct the Fisher's Exact test for 2x2 tables.

Bivariate Tests of Association for Categorical Data: Fisher's Exact Test – the Hypergeometric Distribution

- To compute our p-value (the probability of our observed table or something more extreme under H0), we must be familiar with the hypergeometric distribution.

- Consider all possible tables with fixed margins $(n_{1.}, n_{2.}, n_{.1}, n_{.2})$ and that the rows and columns can be re-arranged such that $n_{1.} \leq n_{2.}$ and $n_{.1} \leq n_{.2}$.

- For each 2x2 table, once we have n_{11}, then all other cell counts can be computed given fixed margins.

Bivariate Tests of Association for Categorical Data: Fisher’s Exact Test – the Hypergeometric Distribution

- Thus our p-value can be computed by summing the probabilities of our observed table and the tables that are more extreme than what we have observed.

$$\Pr(X = n_{11}) = \frac{n_{1.}!\, n_{2.}!\, n_{.1}!\, n_{.2}!}{n_{..}!\, n_{11}!\, (n_{1.} - n_{11})!\, (n_{.1} - n_{11})!\, (n_{.2} - n_{1.} + n_{11})!},$$

where $n_{11} = 0, \ldots, \min(n_{.1}, n_{1.})$

	CVD (Y=1)	No CVD (Y=0)	
High Salt (X=1)	2	23	25
Low Salt (X=0)	5	30	35
	7	53	60

For our data, $n_{11} = 0, \ldots, \min(7, 25)$, thus $n_{11} = 0, \ldots, 7$.

Bivariate Tests of Association for Categorical Data: Fisher's Exact Test – the Hypergeometric Distribution

- Thus our p-value can be computed by summing the probabilities of our observed table and the tables that are more extreme than what we have observed.

$$\mathbf{Pr(X = 0) = 0.017}$$
$$\mathbf{Pr(X = 1) = 0.105}$$
$$\mathbf{Pr(X = 2) = 0.252}$$
$$\Pr(X = 3) = 0.312$$
$$\mathbf{Pr(X = 4) = 0.214}$$
$$\mathbf{Pr(X = 5) = 0.082}$$
$$\mathbf{Pr(X = 6) = 0.016}$$
$$\mathbf{Pr(X = 7) = 0.001}$$

	CVD (Y=1)	**No CVD (Y=0)**	
High Salt (X=1)	2	23	25
Low Salt (X=0)	5	30	35
	7	53	60

Bivariate Tests of Association for Categorical Data: Fisher's Exact Test – the Hypergeometric Distribution

- Thus our p-value can be computed by summing the probabilities of our observed table and the tables that are more extreme than what we have observed.

$$\mathbf{Pr}(X = 0) = \mathbf{0.017}$$
$$\mathbf{Pr}(X = 1) = \mathbf{0.105}$$
$$\mathbf{Pr}(X = 2) = \mathbf{0.252}$$
$$\Pr(X = 3) = 0.312$$
$$\mathbf{Pr}(X = 4) = \mathbf{0.214}$$
$$\mathbf{Pr}(X = 5) = \mathbf{0.082}$$
$$\mathbf{Pr}(X = 6) = \mathbf{0.016}$$
$$\mathbf{Pr}(X = 7) = \mathbf{0.001}$$

	CVD (Y=1)	**No CVD (Y=0)**	
High Salt (X=1)	2	23	25
Low Salt (X=0)	5	30	35
	7	53	60

Exact p-value = Pr(X = 0) + Pr(X =1)+Pr(X=2)+Pr(X=4)+Pr(X=5)+Pr(X=6)+Pr(X=7)

= **0.687**

END OF PART FOUR