Probability Theory and Introductory Statistics

- ALY6010, Section 81229
- Week 1: May 24, 2022
- Professor Dan Koloski (Professor K)
- d.koloski@northeastern.edu
- Roux Institute at Northeastern University

Agenda: 5/24/2022

Time Slot	
6:00-6:30 (30min)	Intro (will be a recap in weeks 2+)
6:30-6:55 (25min)	Section 1: Discrete Probability Distributions
6:55-7:00 (5min)	Break
7:00-7:40 (40min)	Section 2: Continuous Probability Distributions
7:40-7:45 (5min)	Break
7:45-8:35 (40min)	Section 3: Assignments Overview and Some Helpful new R Techniques
8:35-8:40 (5min)	Wrap-Up

Some Ground Rules

- We are here to support each other
  - Respect each other
  - Please use proper Zoom and Discussion Board etiquette
- You are in charge of your own learning
- Academy Integrity matters
  - Your work must be your work
  - Cite all sources in APA style
  - TurnItIn will be used for all assignments
- If you need help, please ask for it!

Professor K Introduction *(Pronouns: he/him/his)*

ORACLE

EMPIRIX

The Kennedy Center

PORT FRINGE

MENSK

Blue Fuse Jazz

stages

TA: Oscar Castro Carrillo *(he, him, his)*

- Masters in Analytics – AMI concentration
- Expected to complete program in Dec 2022
- Data Analyst at Arkatechture
- Castrocarrillo.o@northeastern.edu or send me a message via Slack

Student Introductions

- Name
- Where you live
- Profession (if currently working)
- Favorite summer activity?

Office Hours & Program Meetups

- Oscar Castro Carrillo (TA)
  - Office Hours Saturday 10AM
  - TA Zoom Meeting https://northeastern.zoom.us/j/98903544418
  - castrocarrillo.o@northeastern.edu
- Professor K (d.koloski@northeastern.edu)
  - Drop-in Thursday 12PM-1PM
  - By appt: Email me for other times
- CPS Program Meetups
  - Quarterly
  - Open to all Roux ALY students
  - Meet your fellow students and Roux faculty (maybe guests)
  - Informal discussion around topics of interest – participation optional

Lecture Notes Poster

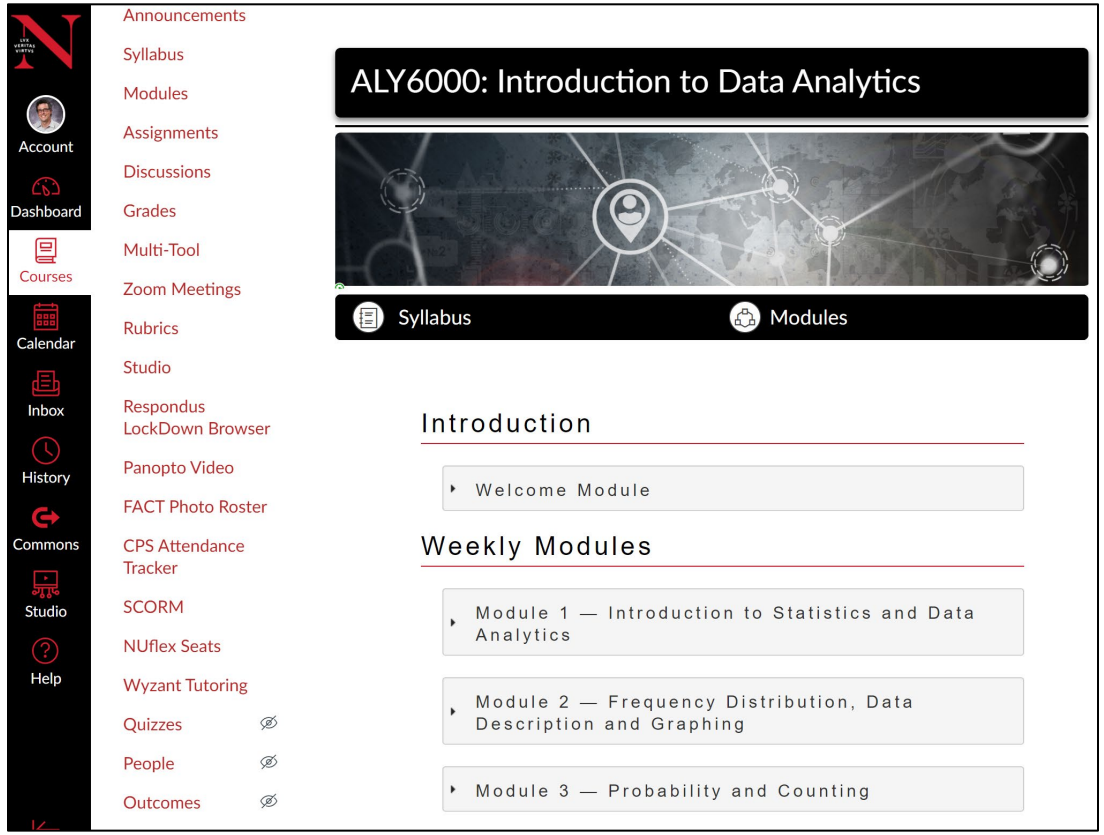
Student Name:

WEEK:	Topic:	
BEFORE WHAT YOU *ALREADY* KNOW	**BEFORE** WHAT YOU *WANT* TO LEARN	**AFTER** WHAT YOU *DID* LEARN
What you already know about this topic	What you want to learn about this topic	Highlights of what you learned from the Lecture / Class Activity
	Potential questions for Lecture	

Where You Started: ALY6000

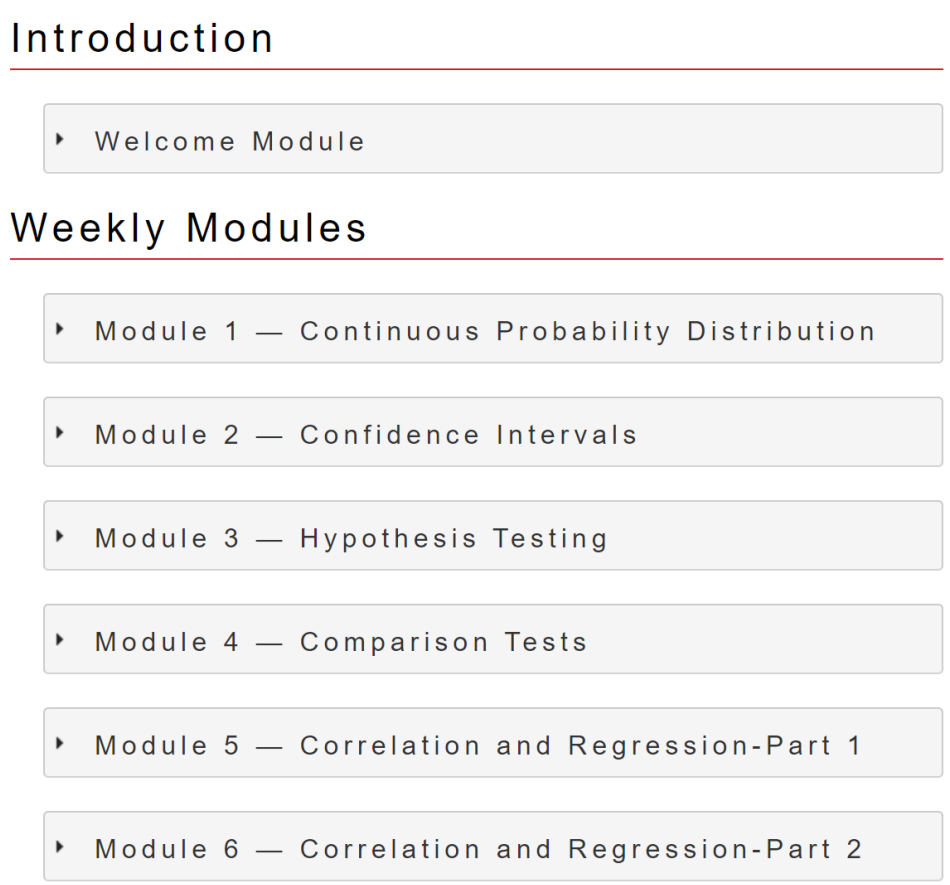
- Learning focus areas
  - Understanding the math
  - Basic Programming and R
  - Communicating your findings

- 6 weekly modules
  - Weeks 1-3 emphasize math and using R
  - Weeks 4-6 emphasize industry, communication and final project prep
  - Ungraded assignments: Blumen Problem Sets, R Practice, In-module Quizzes
  - Graded Assignments: Discussion Boards, Module Projects



ALY 6010 Course and Canvas Overview

- CLO1: Develop **strategic and operational questions based on the data** and the need of the organization
- CLO2: Use **data analysis techniques (hypothesis testing, correlation, t-testing, variance, and single variable linear regression)** to answer research questions
- CLO3: **Use "R"** to perform computations and a wide range of statistical techniques
- CLO4: **Interpret the results** from data analysis techniques
- CLO5: **Explain the implications** of the results from data analysis for the purpose of answering essential business questions



ALY 6010 Grading

Graduate Programs Final Grading Scale

	Title	Description	Grade (Pts or %)
1	Discussions	6 total	10%
2	R Practice assignment	6 total	20%
3	Quizzes	6 total	10%
4	Final Project and Milestone ssignments	2 milestone assignments and final project	60%
	Total		**100%**

Week 1 Objectives and Task List

Learning Objectives

By the end of this module, you should be able to:

- Distinguish between descriptive statistics and inferential statistics
- Calculate expected value and variance
- Use computational tools (such as R) to calculate probability distribution
- Use R to create a descriptive statistics table by sub-group

Task List

- View lessons in Canvas
- Read Elementary Statistics, Chapters 5 & 6
- Review R in Action, Chapters 1-5
- Complete primary Discussion post by Thursday
- Complete Practice Problem Set (not submitted)
- Complete R Practice assignment
- Take quiz
- Review and start Final Project

Section 1

Discrete Probability Descriptions

Basic Statistics Vocabulary: Reminders

- Statistical theory
  - How to THINK about data problems
- Data structure
  - How to ACTUALLY SOLVE the problems
- Distribution
  - Behavior of statistical measures, can be discrete or continuous
- Sample distribution
  - Values and frequency of those values
- Descriptive Statistics
- Inferential Statistics
- Predictive Statistics

Descriptive Stats for Numerical Data

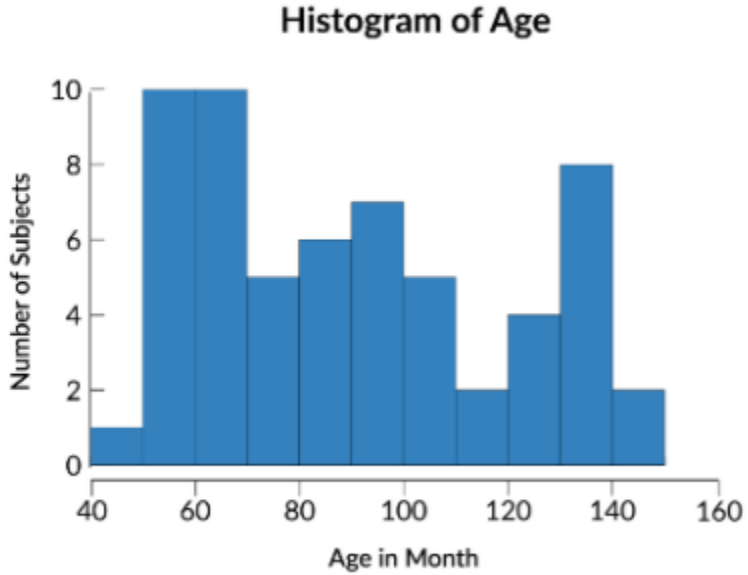
Measures of central tendency

Measures of dispersion

Measures of shape

Measures of location

Mean	90.42
Standard Error	3.90
Median	84
Mode	60
Standard Deviation	30.23
Sample Variance	913.84
Kurtosis	-1.18
Skewness	0.39
Range	95
Minimum	48
Maximum	143
Sum	5425
Count	60



Source: Northeastern University, *ALY6000: Introduction to Analytics*. Course Materials.

Probability Distributions

- Vocabulary
  - Random Variable
    - Discrete
    - Continuous

- A *probability distribution* for a random variable is a table, a graph, or a formula that assigns probabilities to each value the random variable takes

- Example: Toss a coin twice and let X be the number of tails observed. Construct a probability distribution for X.
  - Sample Space: S = {HH, HT, TH, TT}
    - X = 0 when {HH} is result (1/4)
    - X = 1 when {HT or TH} is result (2/4)
    - X = 2 when {TT} is result (1/4)

# of observations of X	Probably of outcome P(X)
0	1/4, or 0.25
1	2/4 or 0.5
2	1/4, or 0.25

Discrete Distributions

- A random variable *X* is a numerical outcome of a random experiment
- The distribution of the random variable, *X*, is the collection of possible outcomes along with their probabilities, matching the outcomes of *X* to the probability of *X*, *P(X)*.
- Random variable X draws from the sample space of all possible probabilities
- P(X) is between 0 and 1 for any given event
  - Σ P(X) = 1
- Expected value E(X) is same as the weighted average:
  - E(X) = Σ (X*P(X))
- Variance can be calculated by reworking the standard Variance formula to be
  - Var = E(X^2) – E(X)^2

Displaying Discrete Distributions

Frequency Table

Species	Frequency Distribution	Relative Frequency Distribution	Cumulative Relative Frequency Distribution
Largemouth Bass	228	0.33727811	0.3372781
Bluegill	220	0.32544379	0.6627219
Bluntnose Minnow	103	0.15236686	0.8150888
Yellow Perch	38	0.05621302	0.8713018
Black Crappie	36	0.05325444	0.9245562
Iowa Darter	32	0.04733728	0.9718935
Pumpkinseed	13	0.01923077	0.9911243
Tadpole Madtom	6	0.00887574	1.0000000

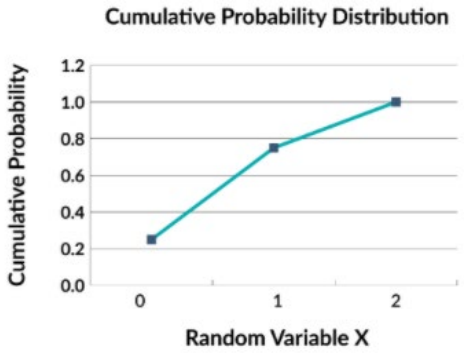
Discrete Probability Distribution Table

Species	Probability Distribution: f(X) or P(X)	Cumulative Density Function: F(X) or CDF(X)
Largemouth Bass	0.33727811	0.3372781
Bluegill	0.32544379	0.6627219
Bluntnose Minnow	0.15236686	0.8150888
Yellow Perch	0.05621302	0.8713018
Black Crappie	0.05325444	0.9245562
Iowa Darter	0.04733728	0.9718935
Pumpkinseed	0.01923077	0.9911243
Tadpole Madtom	0.00887574	1.0000000

More on Discrete Probability Distributions

- Cumulative Probability Distribution
  - For any value x of a random variable X, the cumulative probability is given by $P(X \leq x)$

x	$P(X \leq x)$
0	0.25
1	0.75
2	1



- Mean/Expected Value
  - $\mu = E(x) = E\ x\ p(X) = \sum x\ P(x)$
- Variance of Discrete Prob. Dist.
  - $\sigma^2 = \sum(x - \mu)^2\ P(x)$
- Standard Deviation of Discrete Prob. Distribution
  - $\sigma = \sqrt{\sum(x - \mu)\ 2\ P(x)}$

Binomial Probability Distributions (1 of 2)

- The experiment consists of `n` identical and independent trials.
- There are only two possible outcomes in each trial; one of which is called the Success (S), and the other is called the Failure (F).
- The probability of success
  - denoted by `P(S) = p`, stays the same from trial to trial.
  - The probability of failure is denoted by `q`. Note that `p + q = 1`, or `q = 1 – p`.
- The binomial random variable X is the number of successes in `n` trials

Binomial Probability of a Success result after n number of trials

- *n* is the sample size (or the number of trials)
- *p* and *q* are respectively the probabilities of success and failure
- *n!* ("n factorial") is computed as follows:
  - *n! = n(n - 1)(n - 2) ... (3)(2)(1)(0)*, where 0! = 1

- $P(X = x) = \frac{n!}{X!(n-x!)} p^x q^{n-x}$

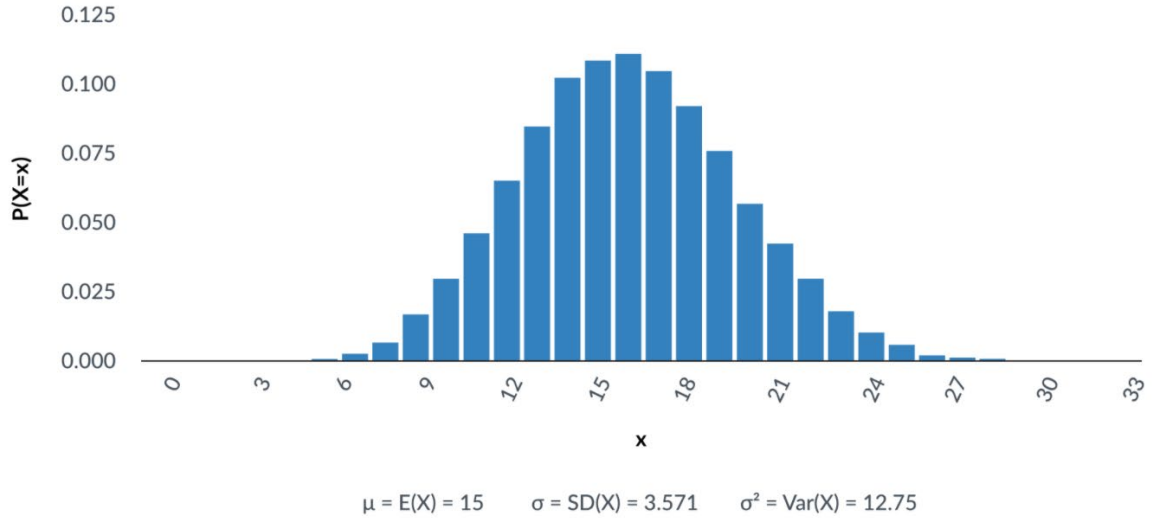
Binomial Probability Distributions (2 of 2)

- The binomial random variable X is the number of successes in `n` trials

- Expected Value:
  - **`E(X) = np`**

- Population Variance
  - **`Var(X) = np(1-p)= npq`**

- Example Histogram of Binomial Probability



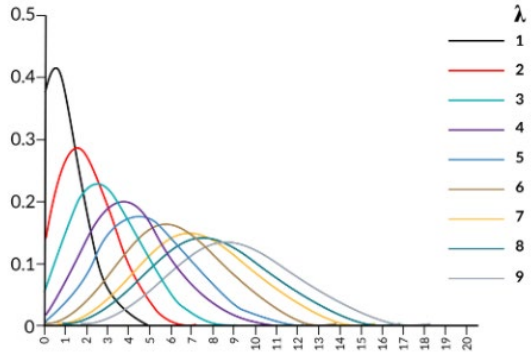
Poisson Probability Distributions

- Calculates the **probability that an event will occur a given number of times in a given interval**
- Requires the following conditions:
  - Each event is independent of others
  - The rate of occurrence is constant
  - Two events cannot occur at the same instant
  - The probability of an event occurring in an interval is proportional to the length of that interval

- Calculating Poisson Probabilities
  - `k`: number of times event occurs in interval, integer
  - λ (Greek “lambda”): average number of event occurrences in interval
  - `e`: Euler's number, a mathematical constant approximately equal to 2.71828

$$P(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

$$E[X] = \lambda$$



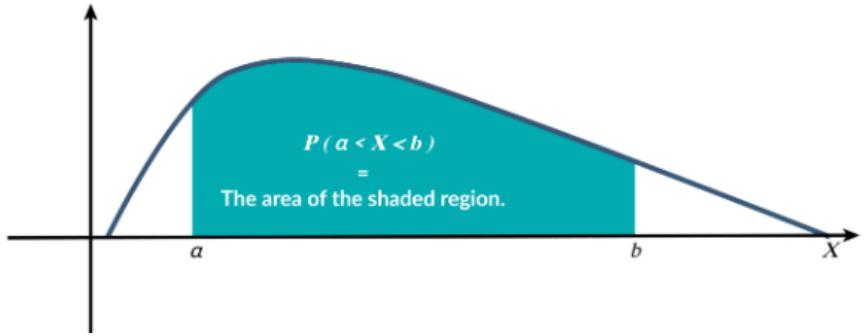
[Break]

Section 2

Continuous Probability Distributions

Continuous Probability Distributions

- Takes an infinite number of possible values in a certain range
  - Talking about P of an single point a is irrelevant ($P(x=a)=0$)
  - $P(a < X < b)$ = The area under the curve between the points a and b



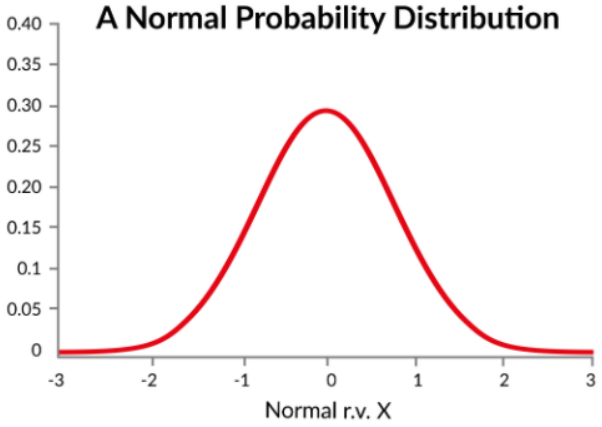
- With continuous random variables, getting the P of the endpoints doesn't matter, since $P(X=a)=0$

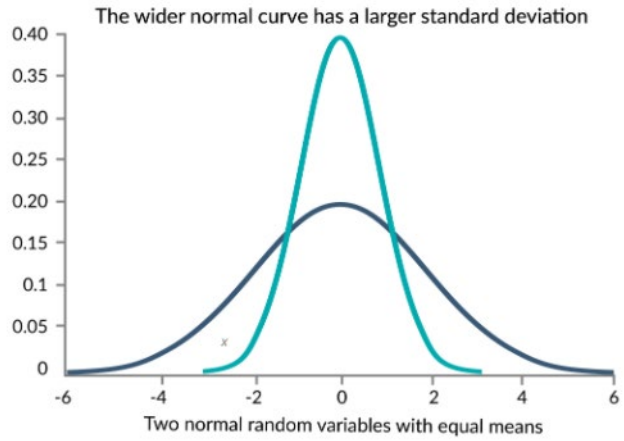
- Therefore

$$P(a < X < b) =$$
$$P(a \leq X < b) =$$
$$P(a < X \leq b) =$$
$$P(a \leq X \leq b)$$

Normal Probability Distribution

- **“The Bell Curve”**





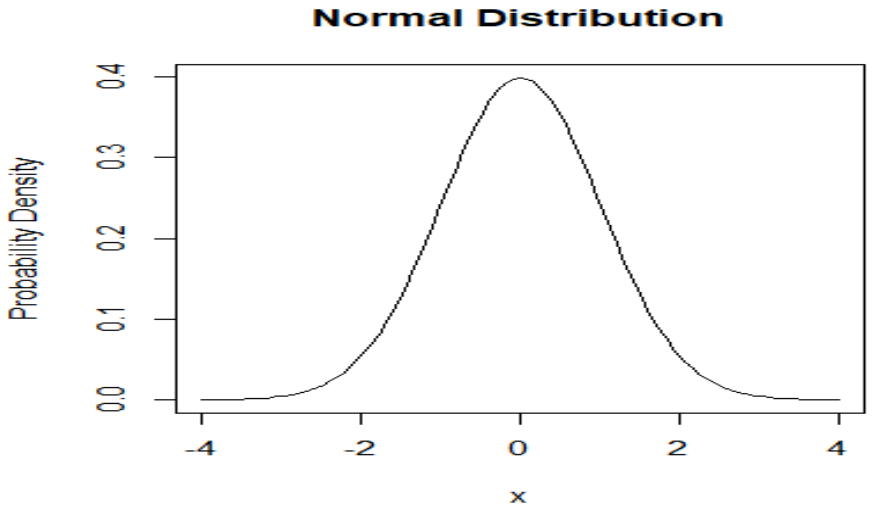
Normal Random Variable

- Any normal random variable *X* with a mean μ and a standard deviation σ can be converted into a standard normal random variable *Z* by the following formula:
  - $Z = \frac{x-\mu}{\sigma}$
- This is helpful because you can use a standard normal distribution table to calculate the area under the curve (P(a<x<b)

Checking for Normality

- Much of inferential statistics is based on the presumption of a normally-distributed set of data ... so check your assumptions!!!!

- Some ways to check for normality:
  - Histogram in R: `hist(x)`
    - Does it look bell-shaped?



  - Pearson Coefficient (PC) of Skewness
    - $\boldsymbol{PC} = \frac{\boldsymbol{3(\bar{X} - median)}}{\boldsymbol{s}}$, where s = sample standard deviation
    - IF PC <= -1 OR PC >= 1, THEN data is significantly skewed

Z Score and T-Statistic

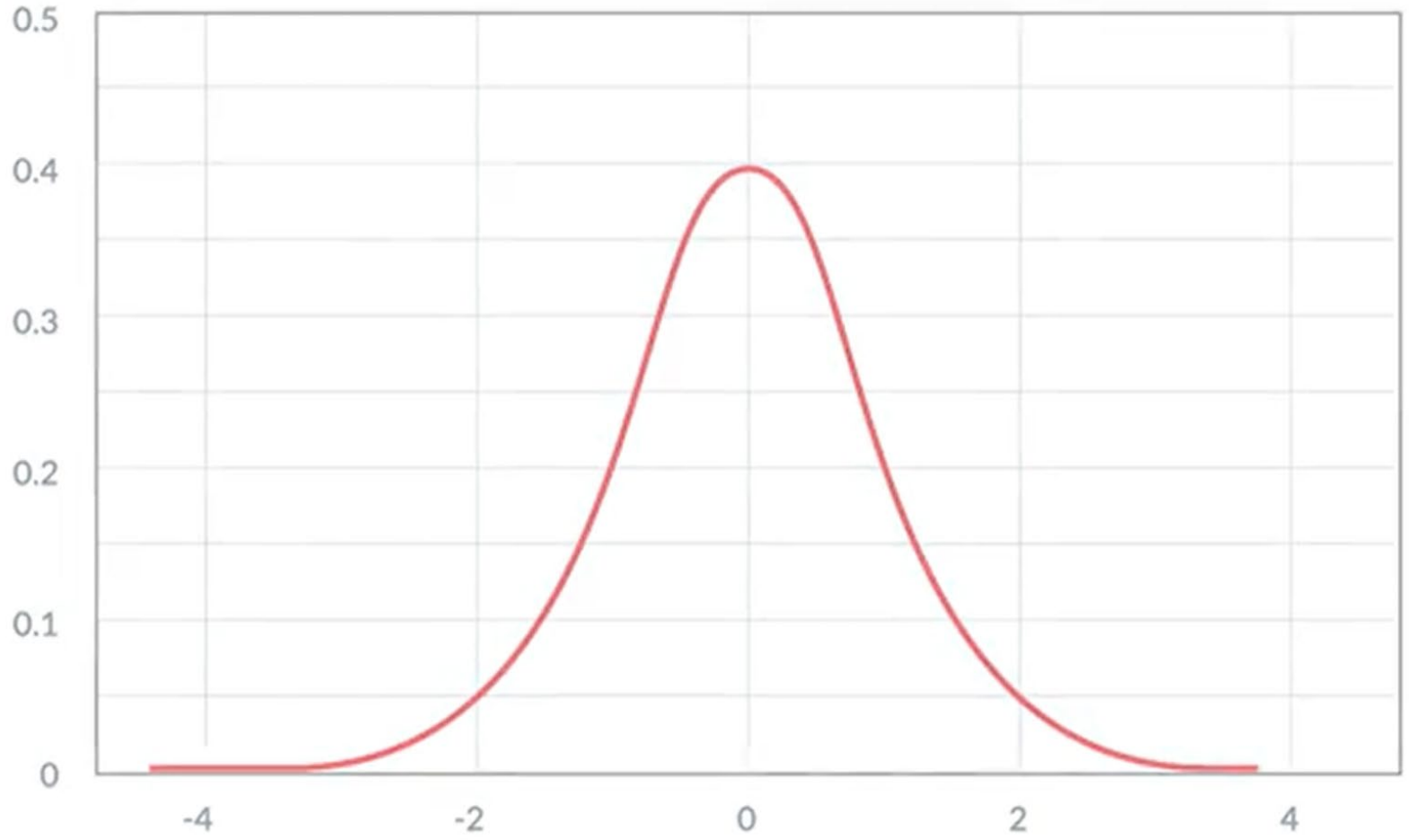
Z Score (Standard Normal Dist.)

- Used when sample size **n** >= 30
- Used when you know population mean and population standard deviation
  - $Z = \frac{x-\mu}{\sigma}$
- Z helps estimate probability (P-value, or area under the curve) of value of X being a certain distance from the mean
- Z(0) = Z(μ) = 0.5

T-Statistic (T-Distribution)

- Used when sample size **n** < 30
- T-Distribution is wider and flatter than Standard Norm. Dist.
  - Less precise estimation as a result
- Used when you know population mean and population standard deviation
- T-Statistic helps estimate probability (P-value, or area under the curve) of value of X being a certain distance from the mean

Z Score Graphically



Solving Std. Norm. Dist. Problems for P(X)

STEP 1:
State the question & data you have

- Ex. question: “Find P(x>50)” or “Find P(x=25)” or “Find P(2<x<10)”
- Ex. data: x, μ, and σ (or $\bar{x}$ or s)

STEP 2: DRAW the curve & area

- Help you figure out which Zs to use

STEP 3: Calculate relevant Z score(s)

- May be just a single Z or two Zs depending on question (= , < , >) and whether X is above or below the mean

STEP 4: Find the relevant area (P)

- Different combinations depending on the question (draw the curve!)
  - Ex: Z, 1-P(Z), P(Z2)-P(Z1), 1-P(Z2)-P(Z1)), or P(|Z|)-0.5

Helpful Functions in Excel and R

IN EXCEL

`=NORM.DIST(q,mean,stdev,TRUE/FALSE)`

Z-Score -> P-Value

- **q:** The z-score
- **mean:** The mean of the normal distribution. Default is 0.
- **stdev:** The standard deviation of the normal distribution. Default is 1.
- **Cumulative:** If TRUE, the cumulative probability to the left of **q** in the normal distribution is returned. If FALSE, the output of the prob. density function is produced.

`=NORM.INV(p,mean,stdev)`
P-Value -> Z-Score

- **p:** The p-value . Rest of variables are same as `norm.dist`

IN R

`pnorm(q, mean = 0, sd = 1, lower.tail = TRUE)`
Z-Score -> P-Value

- **q:** The z-score
- **mean:** The mean of the normal distribution. Default is 0.
- **sd:** The standard deviation of the normal distribution. Default is 1.
- **lower.tail:** If TRUE, the probability to the left of **q** in the normal distribution is returned. If FALSE, the probability to the right is returned. Default is TRUE.

`qnorm(p, mean = 0, sd = 1, lower.tail = TRUE)`
P-Value -> Z-Score

- **p:** The p-value . Rest of variables are same as `pnorm`

Bluman Examples

Inferring Sample Stats from Population Stats

- We can **use population statistics and Z-score or T-statistic to make inferences about hypothetical sample statistics**

- Example
  - Given information about a population mean $\boldsymbol{\mu}$ and population standard deviation $\boldsymbol{\sigma}$ or population variance $\boldsymbol{\sigma^2}$ …
  - What is the probability that a sample of size $\boldsymbol{n}$ from that population will have a certain sample mean $\boldsymbol{\bar{x}}$ or sample standard deviation $\boldsymbol{s}$ or sample variance $\boldsymbol{s^2}$?

- Very useful for “smell-testing” data from sampling techniques

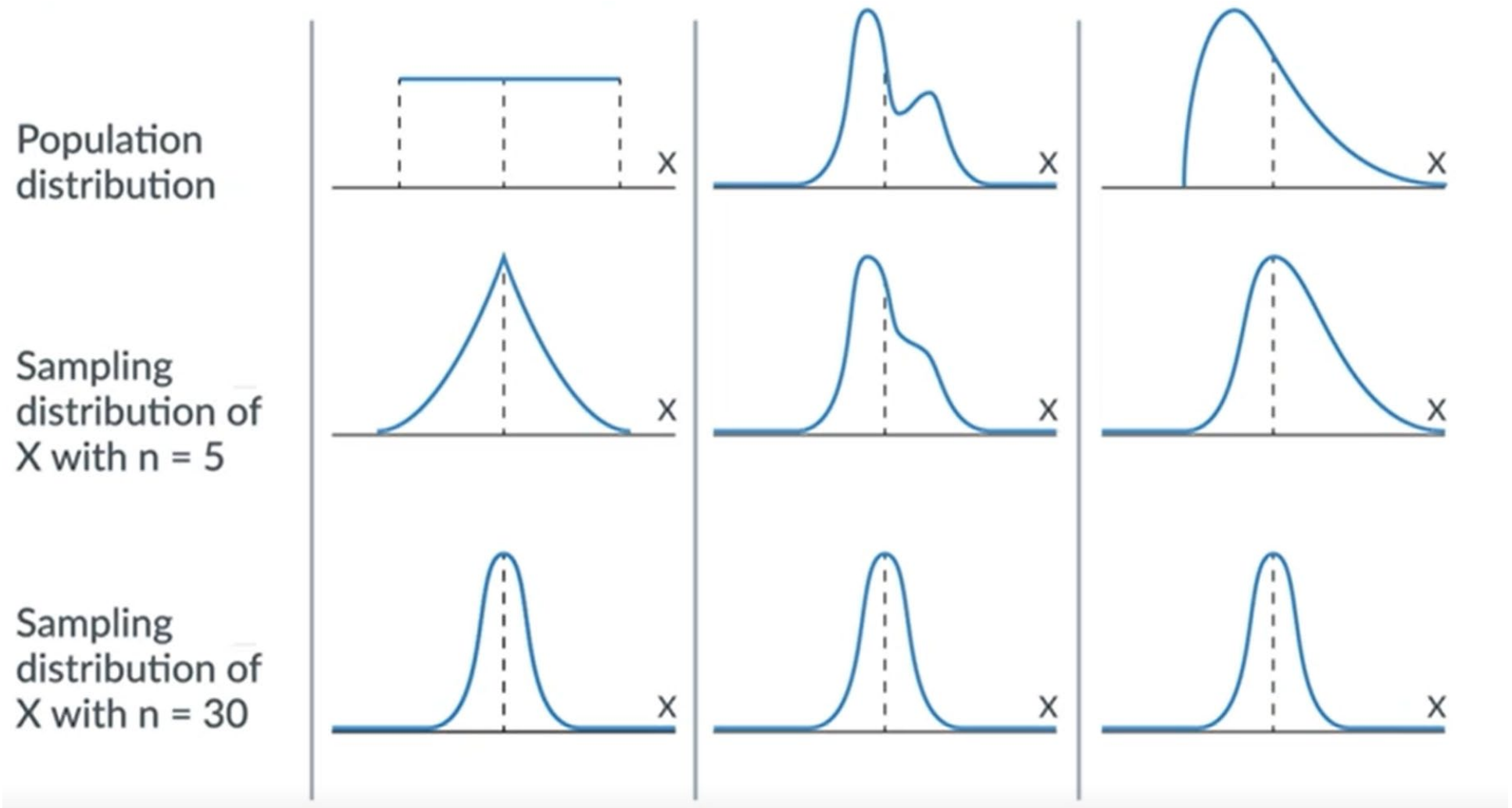
Wait. Say that again. What can we infer?

- Central Limit Theorem
  - A distribution of Sample Means ($\bar{x}$) approaches normal distribution when n >=30 and approaches t-distribution when n <30, regardless of the actual distribution of X
  - Distribution of Sample Variance $\mathbf{s^2}$ approaches chi-square distribution when n >=30, regardless of distribution of X
  - N <= 30 is a rule of thumb for using t-distribution vs normal distribution

- Law of Large Numbers
  - ($\bar{x}$) approaches Expected Value of X, EV(X) = μ Population Mean as n approaches infinity

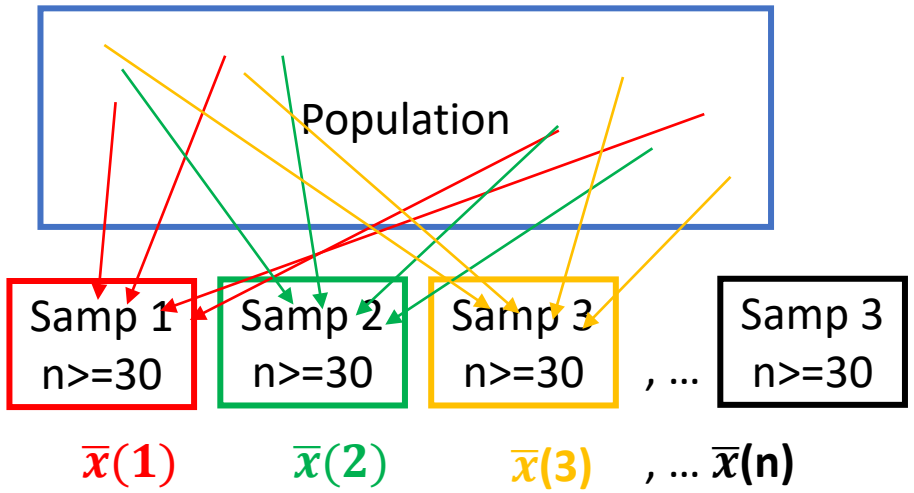
- THEREFORE
  - If we know something about the population statistics, we can use the standard normal distribution to make inferences about hypothetical sample statistics

Central Limit Theorem Graphically



Distribution of Sample Means

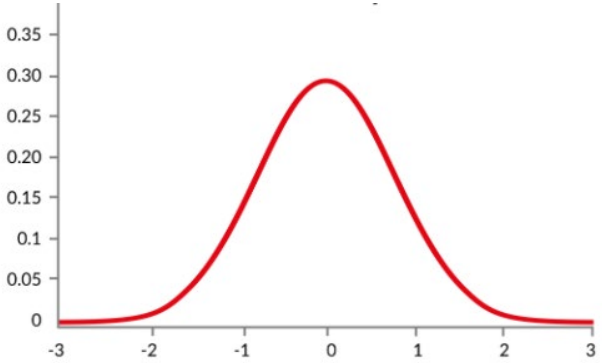
- From a given population ….
- Take a series of random samples where n>=30 and find the sample means
- Distribution of those sample means should ALSO be approximately normally distributed



Sample Means
$\bar{x}(1)$
$\bar{x}(2)$
$\bar{x}(3)$

$$Mean\ of\ Means(\bar{x})$$
$$\mu(\bar{x}) = \frac{\Sigma(\bar{x})}{n(\bar{x})}$$

$$Standard\ Error\ of\ Means$$
$$SE(\bar{x}) = \sigma_{\bar{x}} = \frac{\sigma}{n}$$



Z-score Equations for Sample Statistics

Z-Score for Distribution of Sample Means $\bar{x}$

- $z = \frac{\bar{x}-\mu}{\sigma_{\bar{x}}}$
- $\sigma_{\bar{x}}$ =

  "(population) standard deviation for the sampling distribution"
  = "standard error of the means"

Z-Score for Distribution of Sample Means $\bar{x}$

- $z = \frac{\bar{x}-\mu}{\sigma_{\bar{x}}}$
- $\sigma_{\bar{x}} = \frac{\sigma}{\mathrm{SQRT(n)}}$, therefore
- $\mathrm{z} = \frac{\bar{x}-\mu}{\frac{\sigma}{\mathrm{SQRT(n)}}}$

Solving Std. Norm. Dist. Problems for P($\bar{x}$)

STEP 1:
State the question & data you have

- Ex. question: "Given certain population statistics, what is the probability that in a random sample of 100 adults, the average would be X.

- Ex. data: x, μ, σ and n (or $\bar{x}$ or s)

STEP 2: DRAW the curve & area

- Help you figure out which Zs to use

STEP 3: Calculate relevant Z score(s)

- May be just a single Z or two Zs depending on question (= , < , >) and whether **x** is above or below the mean

STEP 4: Find the relevant area (P)

- Different combinations depending on the question (draw the curve!)
  - Ex: Z, 1-P(Z), P(Z2)-P(Z1), 1-P(Z2)-P(Z1)), or P(|Z|)-0.5

Bluman Examples

Normal Approximation of Binomial Probability

- Binomial Reminders
  - The experiment consists of `n` identical and independent trials with outcomes of Success (S) and Failure (F).
- The probability of success
  - denoted by `P(S) = p`, stays the same from trial to trial.
  - The probability of failure is denoted by `q`. Note that `p + q = 1`, or `q = 1 – p`.
- The binomial random variable X is the number of successes in `n` trials

Ensuring Sample Size is Big Enough to Use Normal Approx.

- IFF **n*p AND n*q are both >=5**

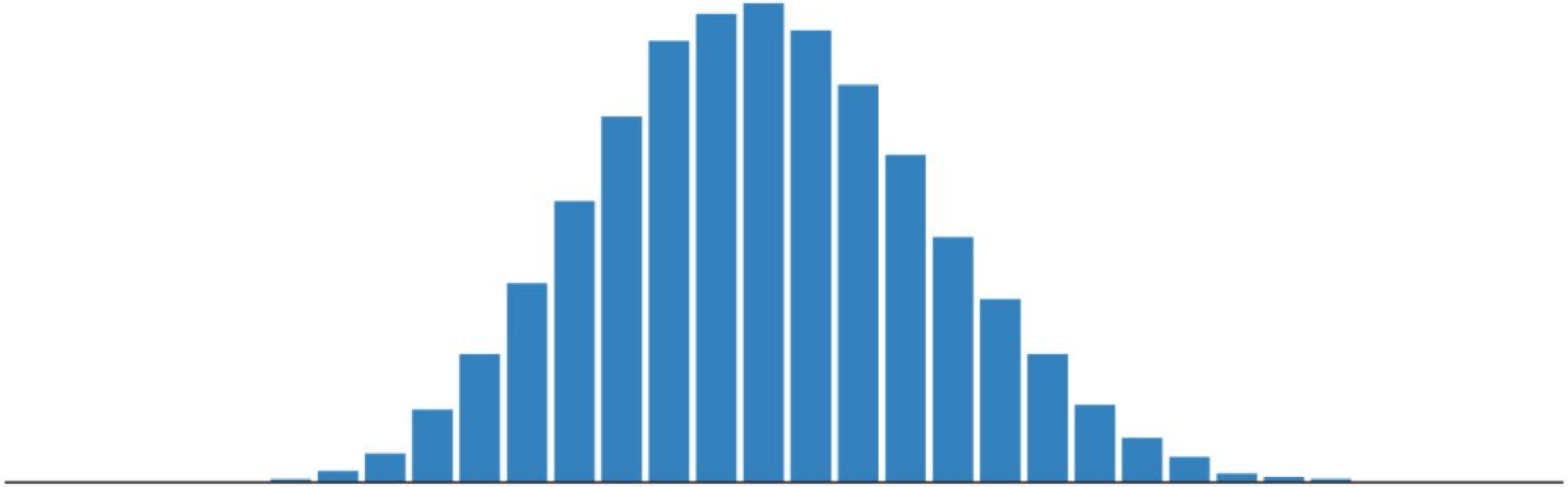
Since Binomial Distribution is discrete and X values are integers, need to apply a **continuity correction** to make it an area under the curve rather than a point

- For P(X=a), use P(a-0.5 < X < a+0.5)
- For P(X>a), use P(>a + 0.5)
- For P(X>=a), use P(>a - 0.5)
- For P(X<a), use P(<a - 0.5)
- For P(X<=a), use P(<a + 0.5)

Equations to find the mean and Stdev for use with Z

- $\mu = n*p$
- $\sigma = SQRT(n*p*q)$

Graphical Example



Solving Std. Norm. Dist. Problems for P($\bar{x}$)

STEP 1:
State the question & data you have

- Ex. question: "Given certain population statistics about an allergy (Yes/No), what is the probability that in a random sample of 100 adults, X or more would be allergic.

- Ex. data: $\mathbf{x}$, $\mathbf{n}$, $\mathbf{p}$, $\mathbf{q}$, and (calculate μ, σ)

STEP 2: Make sure you can use norm. approx.; add the continuity correction

- Are np and nq both >= 5?
- Add 0.5 to one or both sides of X, depending on the question

STEP 3: DRAW the curve & area

- Help you figure out which Zs to use

STEP 4: Calculate relevant Z score(s)

- May be just a single Z or two Zs depending on question (= , < , >) and whether $\mathbf{x}$ is above or below the mean

STEP 5: Find the relevant area (P)

- Different combinations depending on the question (draw the curve!)
  - Ex: Z, 1-P(Z), P(Z2)-P(Z1), 1-P(Z2)-P(Z1)), or P(|Z|)-0.5

Bluman Examples

[Break]

Section 3

Assignments Overview and Some Helpful new R Techniques

Discussion Boards - 10% of final grade

Weekly - PRIMARY response due **Thursday** by 11:59 PM EST
2 SECONDARY responses due **Saturday** by 11:59 PM EST

Purpose - Expand our conversation and share your insights, views, experiences

- It is essential that you post on time so that your colleagues can respond!

R Practice Assignments - 25% of final grade

Weekly - due **Sunday** by 11:59 PM EST

*except for Module 6 - due **Friday** 7/1 by 11:59 PM EST*

Purpose - Practical experience applying course material in R and interpreting results

We will use the LungCap.csv dataset referenced in the Canvas Module.

Assignment Strategy

R code for the R Practices

- Trial runs for Final Project milestones
- Place for me to give you feedback & check in

Report for the R Practices

- 35% - Summary & Report Format - you can **present analyses in a (semi-)formal report** (aim for 1-2 paragraphs max!)
- 35% - Data Analysis & Interpretation - you **understand what the statistics behind the data mean**
- 30% - Data Visualizations - you can **create the visuals (or stats tests) in R**

Strategies

- Time box R work
- Outline Report early
- Identify Needed analyses

- Troubleshoot together
- Email with questions
- Attend office hours

Final Project & Milestone Assignments - 65% of grade

Bi-weekly - (all due by 11:59 PM EST Sunday)

- **Data selection decision** - Due this **Sunday 5/29**
- Milestone 1 - Exploratory Data Analysis - Due **Sunday 6/5**
- Milestone 2 - Basic Hypothesis Testing - Due **Sunday 6/19**
- Final Project - Multi-variable Relationships & Report - Due **Friday 7/1**

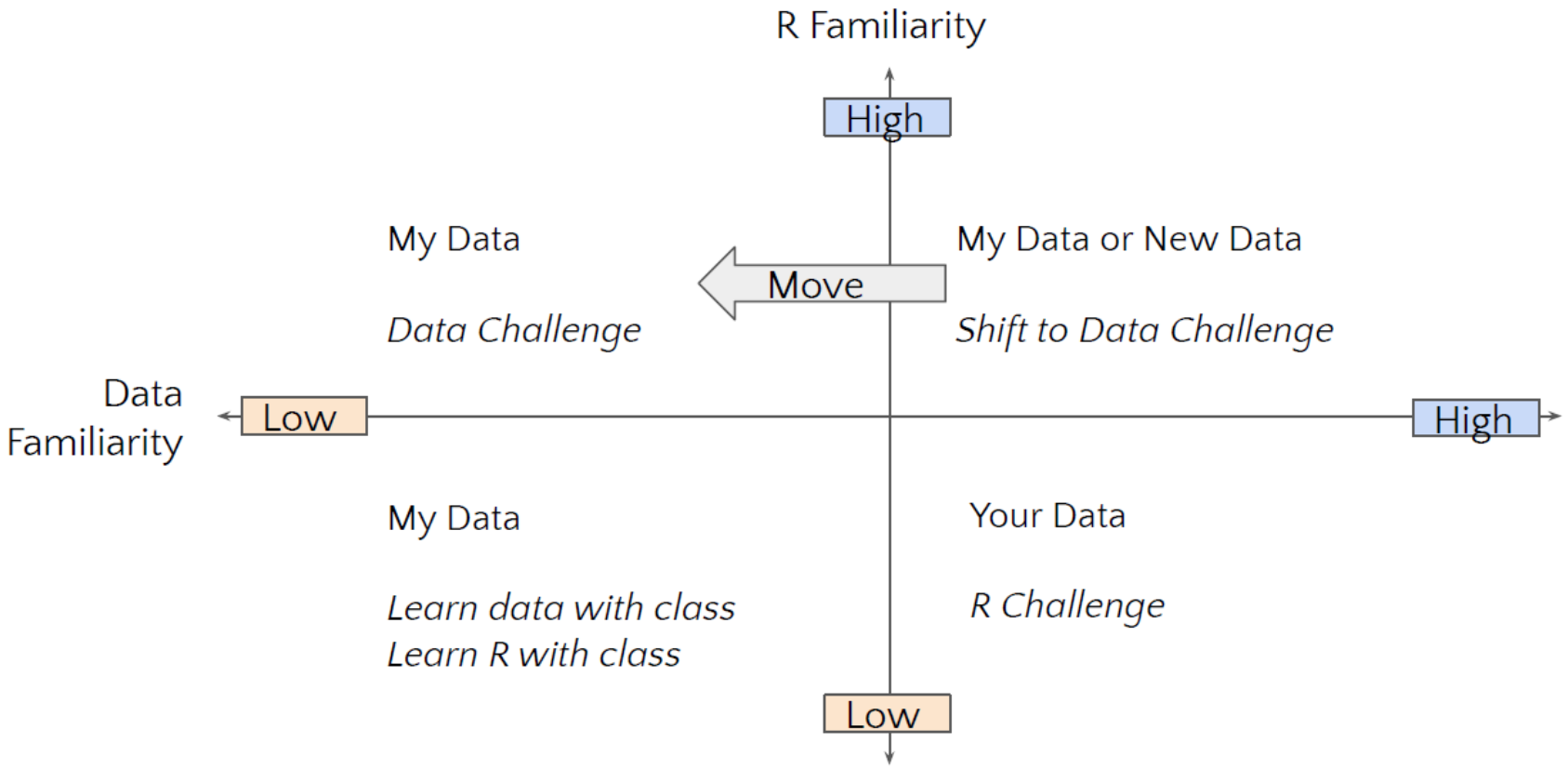
Purpose - End-to-end analysis including data cleanup, exploration, hypothesis testing, presentation to stakeholders

Data source -

1. Choose your own dataset
   - Submit dataset to Dan for approval by this Saturday 5/29
   - At least 6000 rows, 3+ categorical variables, 2+ continuous or discrete variables
2. Use provided data - Study of Women's Health Across the Nation (SWAN): Baseline Dataset, [United States], 1996-1997 (ICPSR 28762)
   - Do not download the original. I have cleaned this data and posted information about it HERE.

Which Data Set To Use? (h/t Prof. Joy-El Talbot)

To choose your own data or not?



In-Class Exercise

- Cool R techniques for this week's assignment

Wrap Up

This Week's Learning Objectives and Task List

Week 1 Objectives and Task List

Learning Objectives

By the end of this module, you should be able to:

- Distinguish between descriptive statistics and inferential statistics
- Calculate expected value and variance
- Use computational tools (such as R) to calculate probability distribution
- Use R to create a descriptive statistics table by sub-group

Task List

- View lessons in Canvas
- Read Elementary Statistics, Chapters 5 & 6
- Review R in Action, Chapters 1-5
- Complete primary Discussion post by Thursday
- Complete Practice Problem Set (not submitted)
- Complete R Practice assignment
- Take quiz
- Review and start Final Project

Thank you! See you next week!

Appendix 1: Review Slides from ALY6000

Central Tendency: Mean, Median and Mode

- *Mean*: Arithmetic average of a set of n numbers
  - Calc: Sum the *n* numbers and divide by *n*
  - $\bar{X}$ *("X bar"):* mean of sample data
  - μ (Greek "mu"): mean of population

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

- *Median*: point at which half of the members of a data set are abolve, and half are below
  - Calc: sort the *n* numbers, and find the middle value

- *Mode*: most frequently-occurring value in the data set
  - Sometimes there is no mode if multiple values occur at same max frequency

Dispersion: Range and IQR

- *Range:* the difference between the maximum and minimum values in the data set.
- A larger range usually (but not always) indicates a larger spread or deviation in the values of the data set.
- Expressed as (min,max), though the value is actually max-min

- *Interquartile range (IQR):* the range of the middle most 50% of the observations.
  - Interquartile range=Q3-Q1
  - =75th percentile - 25th percentile
  - =P75 - P25
- *Outliers* are identified as values beyond 1.5*IQR (inter-quartile range).
  - The lowest value not considered an outlier = Q1 – 1.5*IQR.
  - The largest value, not considered an outlier, is Q3 + 1.5*IQR.

Dispersion: Variance

Variance is mostly a means to an end (Standard Deviation) – except when doing advanced probabilities

- Calculating the *Population Variance* of a Population of N members
  1. Find the population mean μ
  2. Calculate the difference between EACH data point and the population mean $(x_i - \mu)$
  3. Square each difference $(x_i - \mu)^2$
  4. Sum the squares of each difference $\sum_{i=1}^{i=N} (x_i - \mu)^2$
  5. Divide the sum of squares by **N**

$$1/n * \sum_{i=1}^{iN} (x_i - \mu)^2$$

- Denoted as σ^2 ("sigma squared")

- Calculating the *Sample Variance* of a Sample of n members
  1. Find the sample mean $\overline{X}$
  2. Calculate the difference between EACH data point and the sample mean $(x_i - X)$
  3. Square each difference $(x_i - \overline{X})^2$
  4. Sum the squares of each difference $\sum_{i=1}^{i=n} (x_i - X)^2$
  5. Divide the sum of squares by **n-1**

$$1/n-1 * \sum_{i=1}^{i=n} (x_i - \overline{X})^2$$

- Denoted as s^2 ("s squared")

Dispersion: Standard Deviation

- *Standard deviation*: measures deviation from the mean of the data set
- **The most widely used measure of dispersion**
- Calc: square root of the Variance of the data set
  - *To calculate the STDEV, first we need to calculate the Variance!*

- Population Standard Deviation
  - Denoted as σ (Greek "sigma")
  - $\sigma = \sqrt{\sigma^2}$
- Sample Standard Deviation
  - Denoted as "s"
  - $s = \sqrt{s^2}$

Appendix 2: Data Cleansing (h/t Berkeley Almand-Hunter)

Handling Missing Values (20 min)

3 types of missing data

- Missing Completely at Random (MCAR)
  - Assumed for many imputation methods
  - Missing data is unrelated to any other observed or unobserved variable
- Missing at Random (MAR)
  - Missing data is related to other observed variables, but is unrelated to its own unobserved value
  - Example: animals with lower body weights are more likely to have missing values for dream sleep because they are harder to study, but the missingness is unrelated to time spent dreaming
- Not Missing at Random (NMAR)
  - Missing values are neither MCAR or MAR
  - Example: animals that spend less time dreaming are more likely to have a missing dream value

3 Steps for Handling Missing Data

1. Identify the missing data
    a. What % of the data is missing? In which variables?
    b. VIM Package is very helpful for this
        i. Get Started Guide
        ii. Datacamp article
2. Examine the causes of the missing data and correlation with other variables
    a. Does missing data seem to be MCAR, MAR or NMAR?
        i. Correlation between missing variables
    b. Look at VIM plots
3. Delete, ignore, or impute missing data

Options for Handling missing Data

1. Drop observations (rows) with missing values
    - Quick and easy
    - Can only be used if data is MCAR and the amount of missing data is small (< 2% of sample)
2. Rely on algorithm to deal with missing values
    - Some algorithms require you to handle missing data first
    - An algorithm's method for imputation might not make sense for your dataset
3. Impute missing values
    - Once you impute, missing values are handled as if they were observed, which introduces new uncertainty!
    - You have lots of options if your data is MCAR or MAR.
    - From *R in Action,* "if your data are NMAR, you can turn to specialized methods, collect new data, or go into an easier and more rewarding profession."

Source: Hastie, T., Friedman, J., & Tisbshirani, R. (2017). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.

Options 1 & 2

1. Remove Missing Data
    - Listwise deletion: Delete all rows where with any missing values
    - Pairwise deletion: Delete only the rows that have missing values in the columns used for the analysis
    - To remove observations (rows) with missing data, use drop_na()
2. Leave missing value handling to the algorithm
    - some basic R functions default to returning NA if there are any NAs in your dataset
        - sum(), mean(), median(), max(), min()
        - If you specify na.rm = TRUE, the function will ignore NAs and calculate a result based on the remaining data (make sure this is a reasonable thing to do).
    - ggplot will create a plot, but give a warning message (clean your data first!)
    - caret package (short for Classification And REgression Training) will not run with missing values
        - preProcess() has lots of imputation options

Option 3: Impute Missing Data

- Calculate using other columns if possible
  - example: if you have a radius column and are missing area of a circle, you can calculate area
- Impute to the mean or median
  - Calculate using values for all observations that are not missing in the column
  - Pros
    - easy and preserves data
  - Cons
    - doesn’t work for categorical data
    - doesn’t factor for correlations between features (e.g. temperature and day of year)
    - Can change variance, especially if there’s a lot of missing data
    - Produces biased results for data that isn’t MCAR
  - Can improve this for data that is MAR by imputing to mean/median of a group of another variable
  - Examples:
    - find the mean temperature for each day of year
    - find the mean temperature overall

Option 3: Impute Missing Data

- Most frequent values imputation
  - Replace missing values in column with most frequent value
  - Pros
    - Works with categorical data
  - Cons
    - doesn’t account for correlations between features
    - Can introduce bias in the data
- Constant value imputation
  - Impute missing value with specified constant
  - Works well for categorical features (example: replace survey response values with “not answered”)
- Previous/Next Value Imputation
  - Good for time series or ordered data
  - Works for numeric or categorical data

Option 3: Impute Missing Data

- K-nearest neighbors
  - the K closest neighbors are found in the training set and the value for the predictor is imputed using these values (e.g. using the mean)
  - Available in VIM and caret
- Linear regression, ML methods
  - Predict missing values based on other values
- Multiple Imputation Methods
  - Predict missing values multiple times to create multiple datasets and results are combined
  - See MICE package in R
- Option: Add missing data feature (column)
  - Add a boolean column to indicate missing data in your model
  - Still need to use other imputation methods

Caveats

- Missing values will not always be “NA” in your dataset
  - Sometimes people use -99, ‘NaN’ etc. Make sure you convert them to the format R recognizes:
    - mutate(col_name = na_if(col_name, -99))

Resources

- *R in Action,* Chapter 18
- *Elements of Statistical Learning*, Section 9.6 (free pdf)
- Medium Article on handling missing values
- KD Nuggets