MSDS 660: Statistical Methods and Experimental Design

Week 3

Normal Probability Distribution and Analyzing survey responses

kpolson@regis.edu

Recap Week 2

Basic Laws:

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

Compliment of an Event: $$P(A) + P(A^c) = 1, \quad \text{i.e} \quad P(A) = 1 - P(A^c)$$

Multiplication Rule for Independent Events: $$P(A \text{ and } B) = P(A) * P(B)$$

Addition Law for Mutually Exclusive Events: $$P(A \cup B) = P(A) + P(B)$$

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

Topics for Week 3

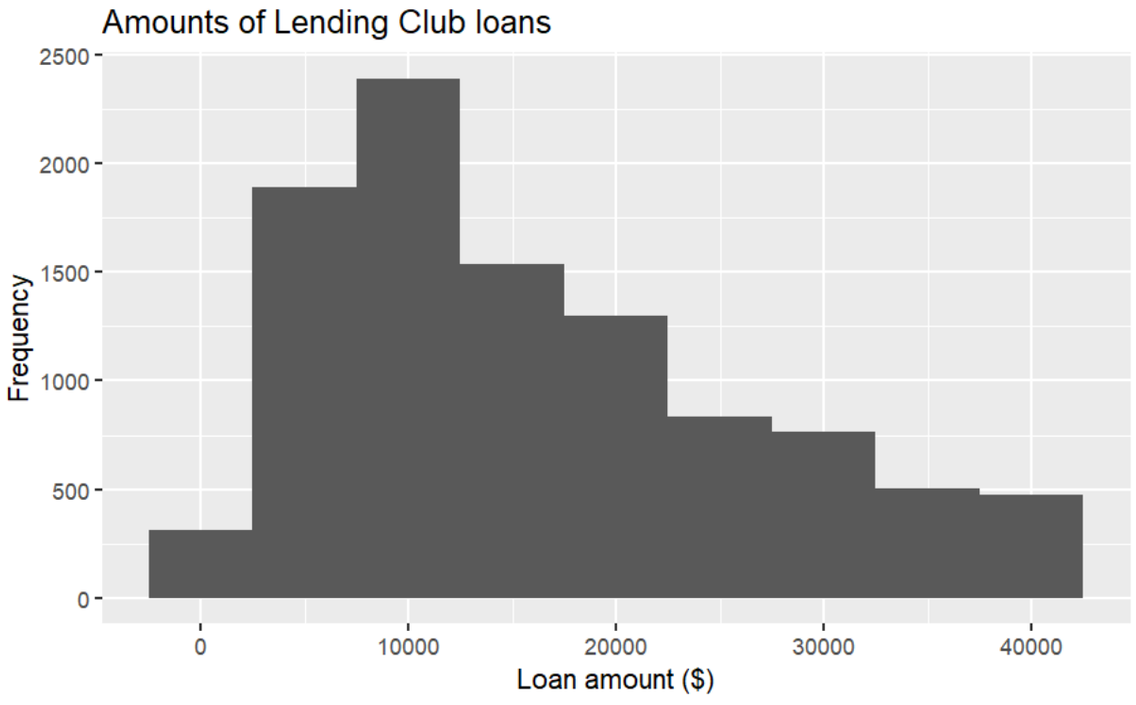
- Continuous probability distributions
  - Properties of a Normal Probability Distribution
- Point estimates and sampling variability

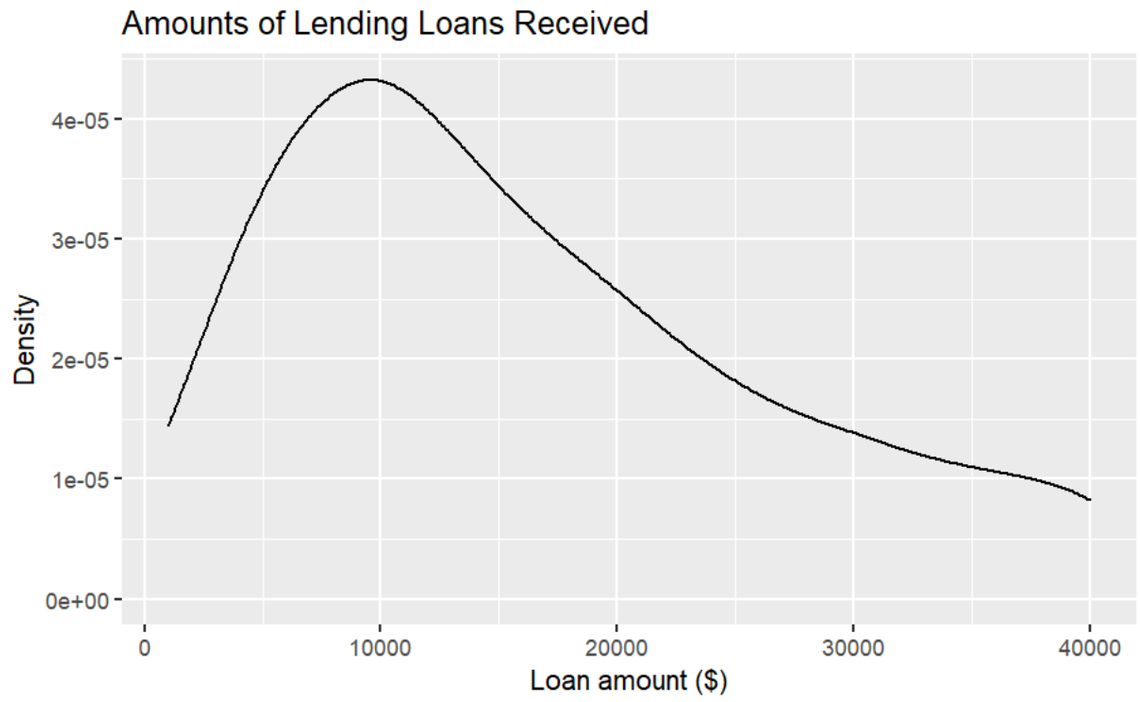
Recap: Discrete Probability Distribution

- **The probability distribution** for a random variable describes how probabilities are distributed over the values of the random variable.
- The probability distribution is defined by a probability function, denoted by ***f*(*x*),** which provides the probability for each value of the random variable.
- A **random variable** is a numerical description of the outcome of an experiment.

Continuous probability distribution

- Continuous probability distributions are used in parametric statistical analyses.
- **Parametric statistics** refers to a branch of statistics which assumes that a sample data comes from a population that can be modeled by a probability distribution with a fixed set of parameters.
- The smooth curve on the right- hand side of the slide represents a probability density function (aka density or distribution).





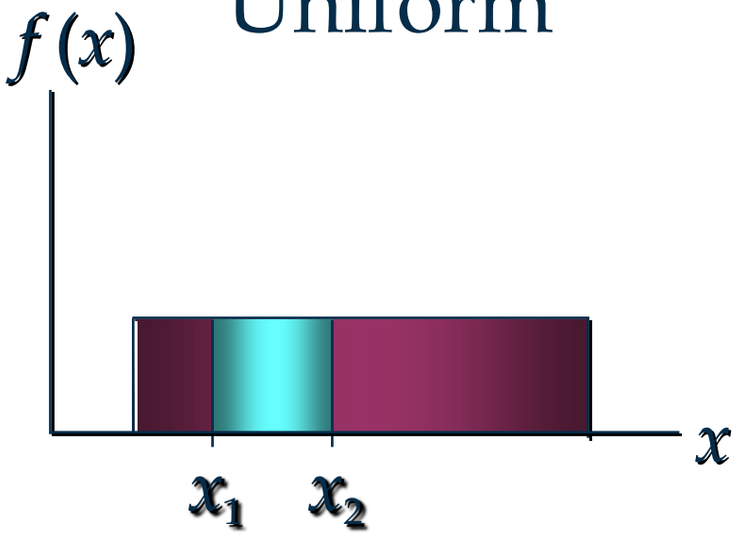
Probability

	Discrete Random Variable	Continuous Random Variable
What kind of?	A RV can take on only certain values along an interval, with the possible values having gaps between them Eg.) Number of customers, number of successes, etc.	A RV can take on a value at any point along an interval. (usually with several decimal places) Eg.) Temperature outside, height, weight, amount of gas, etc.
How to Obtain certain Probabilities	By Probability Function, f(x) Eg.) f(x=1)=0.23, f(x>=1)=f(1)+f(2)+(3)…, f(x<=3)=f(0)+f(1)+f(2)+f(3)	By Probability Density Function (PDF), f(x) **All the time, f(x=a certain value)=0.** We have to compute the probability that x will be within a specified interval of values. Probability= Area under the PDF with a certain interval.
Specific Types	Binomial, Poisson, Negative Binomial, etc.	Uniform, **Normal**, Standard Normal, Exponential, Chi-square, F, Gamma, etc.

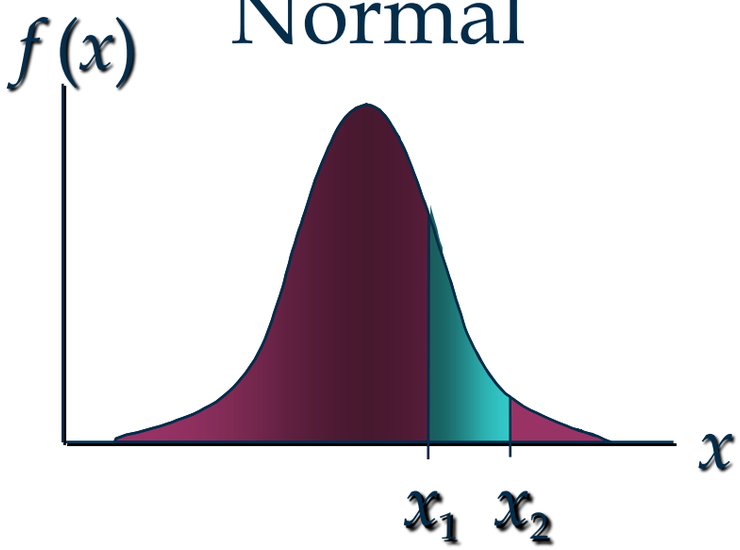
Continuous Distributions

The probability of the random variable assuming a value within some given interval from x_1 to x_2 is defined to be the **Area** under the graph of the Probability Density Function between x_1 and x_2.

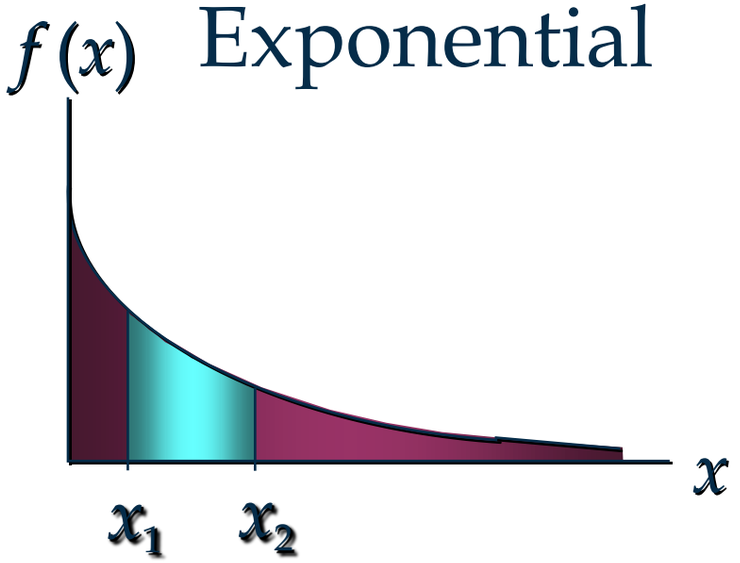
Uniform



Normal

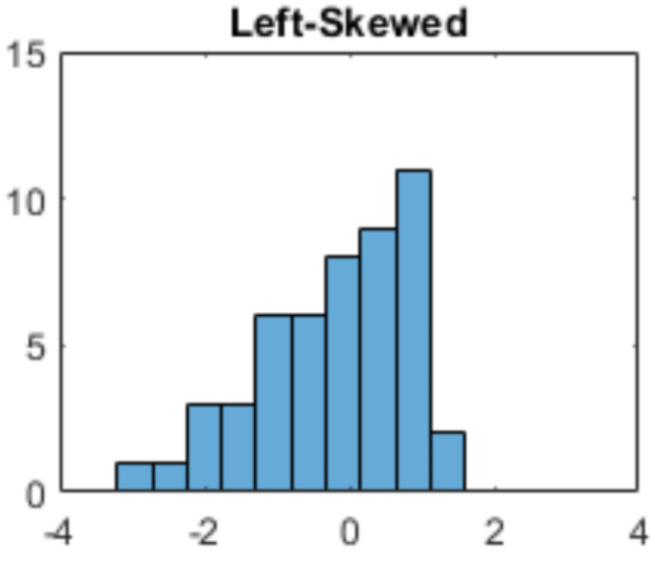


Exponential

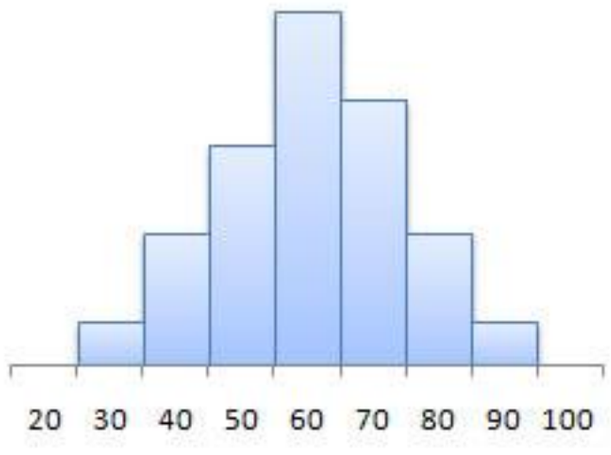


Shape of Numerical Distributions

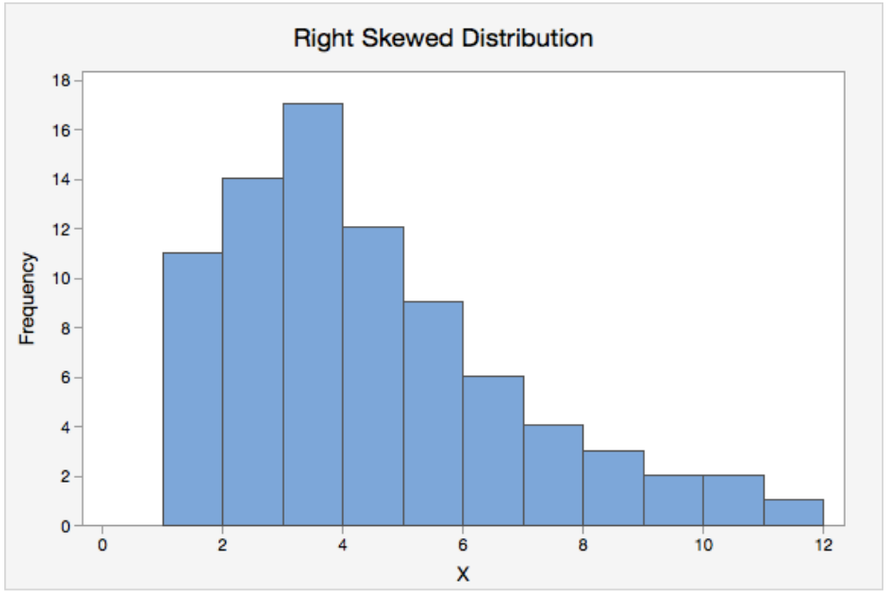
- **Skewness**: Graph shape represents characteristics of a distribution
  - **Right-skewed**: If the distribution has a long right tail
  - **Left-skewed**: If the distribution has a long left tail
  - **Symmetric**: Distribution has roughly equal trailing off in both directions
- Depending on the distribution of the variables, they may need to be transformed to be able to apply parametric statistics



This Photo by Unknown Author is licensed under CC BY-SA-NC



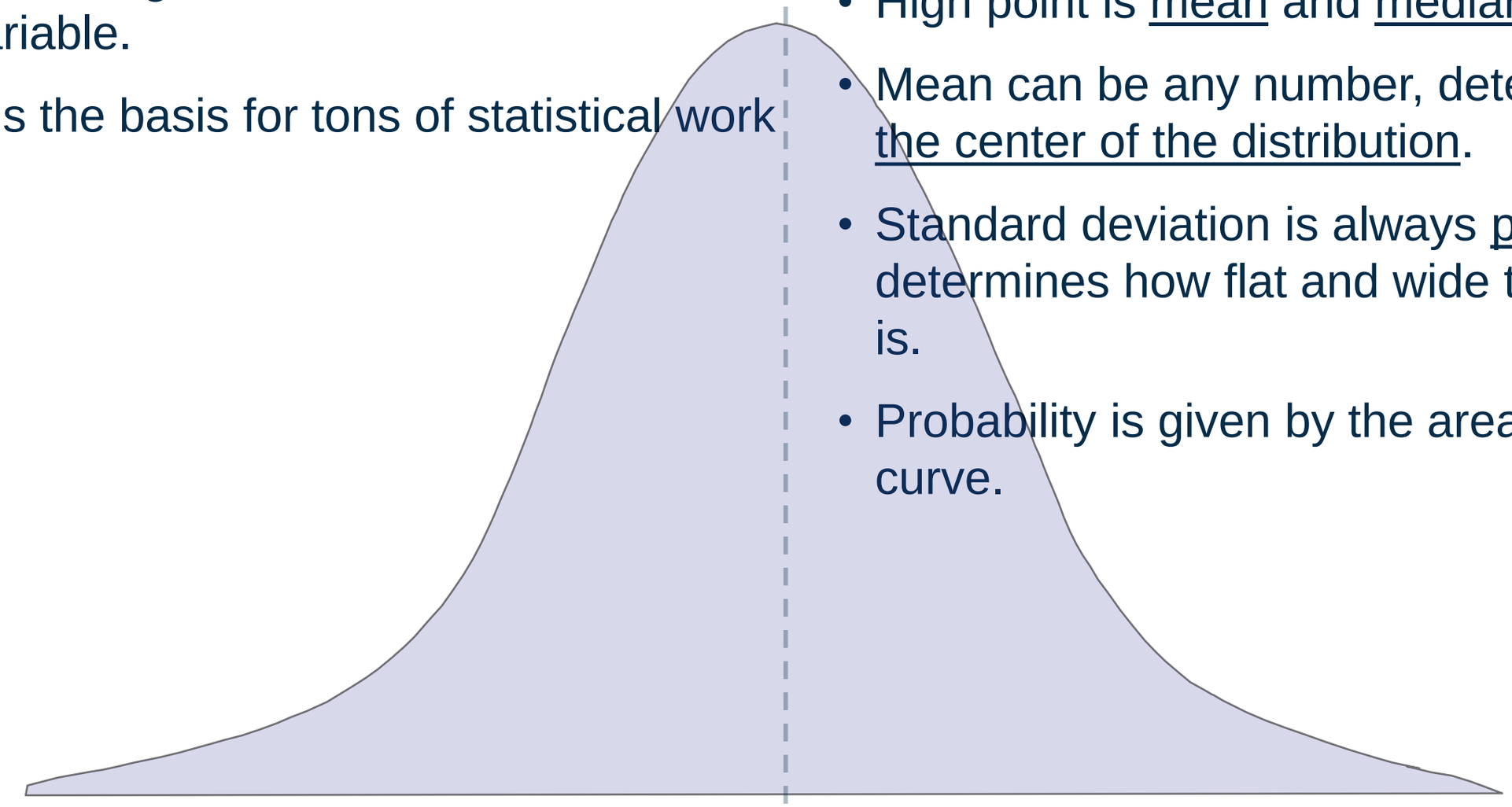
This Photo by Unknown Author is licensed under CC BY-SA-NC



This Photo by Unknown Author is licensed under CC BY-NC

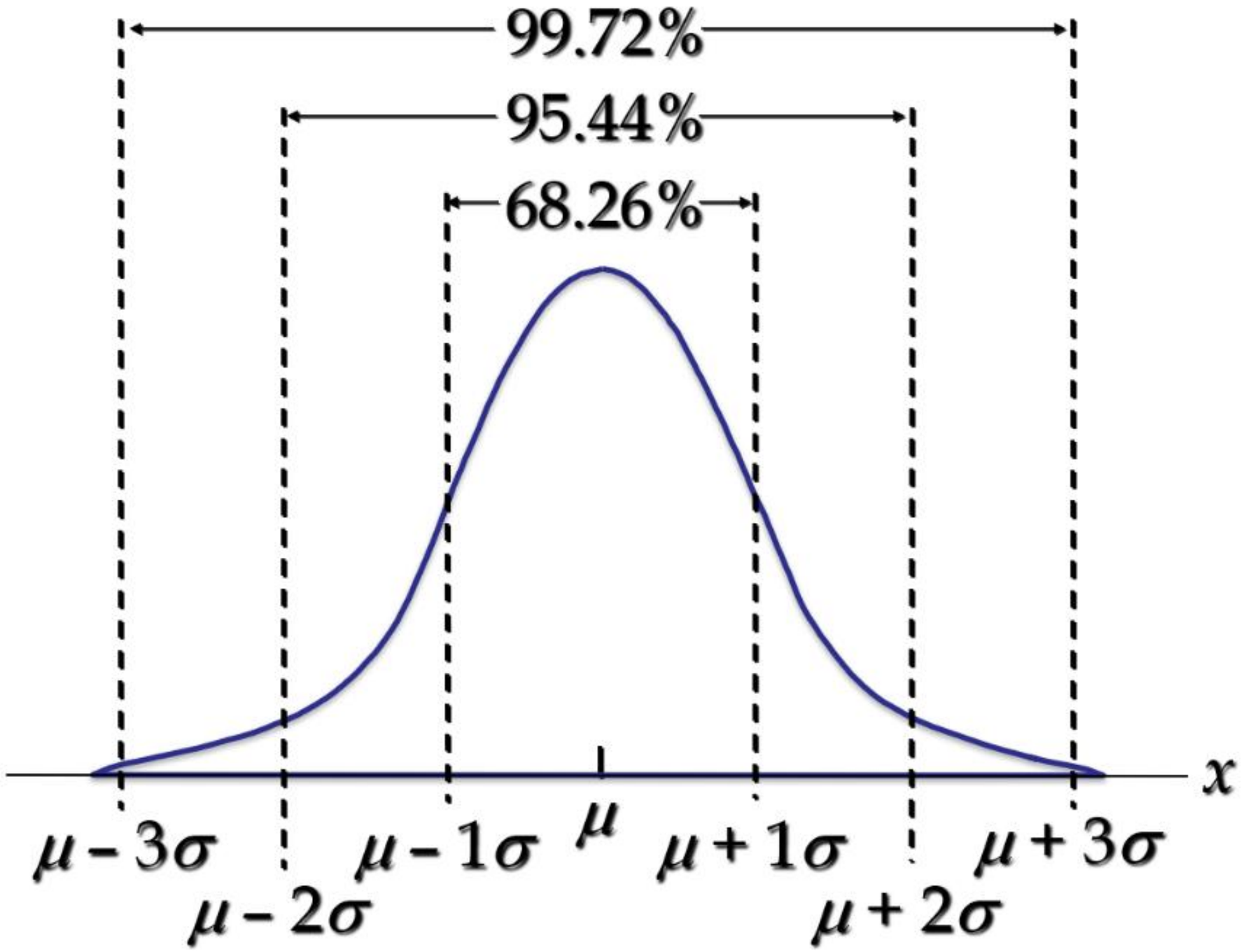
Normal Probability Distribution

- The **normal probability distribution** is the most important distribution for describing a continuous random variable.
- It is the basis for tons of statistical work



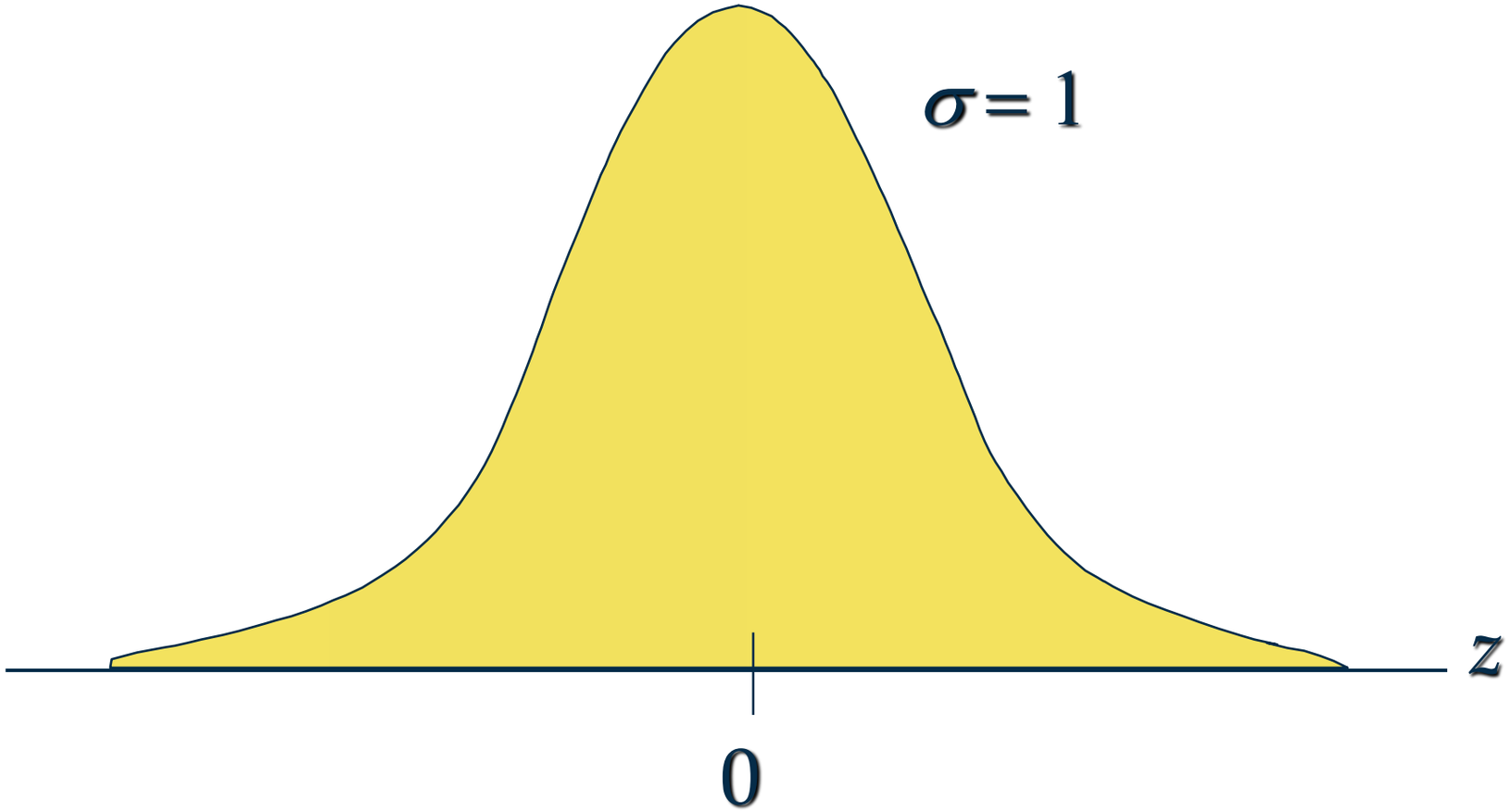
- Mean (μ) and standard deviation (σ) determine shape and location of curve.
- High point is mean and median.
- Mean can be any number, determines the center of the distribution.
- Standard deviation is always positive, determines how flat and wide the curve is.
- Probability is given by the area under the curve.

Normal Probability Distribution



Standard Normal Distribution

- Special case of normal distribution with a mean (μ) of 0 and a standard deviation (σ) of 1.
- Basis of charts that are found in most stats books.
- pg. 410-411 of Open Intro textbook



Z -score

- A Z-score of an observation is the number of standard deviations (SDs) that observation falls above or below the mean.
- If the observation (value of X) is 1 SD above the mean, its Z-score = 1
- If the observation (value of X) is 1.5 SDs below the mean, its Z-score = - 1.5
- If X is an observation from a distribution N (μ, σ):

the Z-score is defined as: $$z = \frac{\bar{x} - \mu}{\sigma}$$

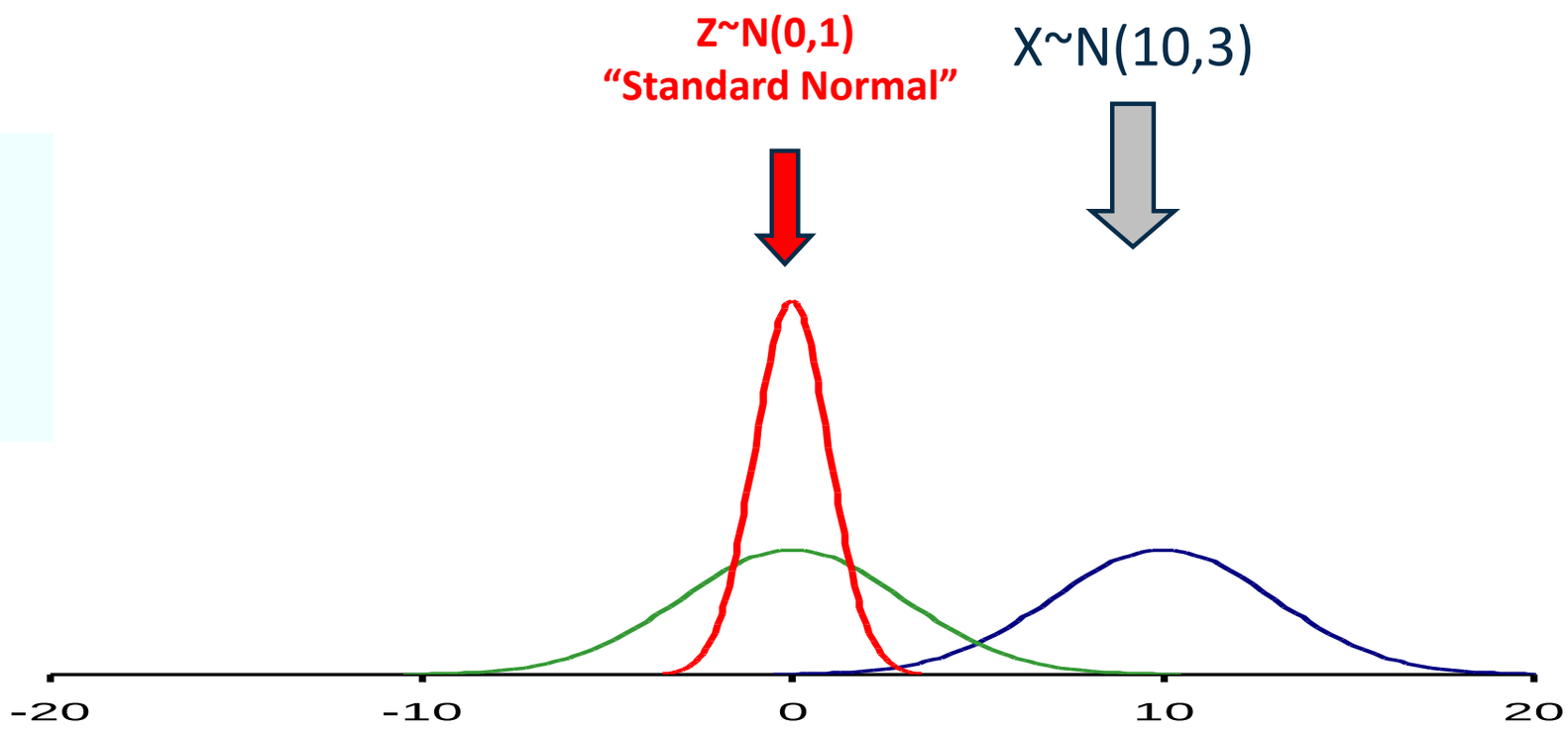
Standardization

Let X be a random variable representing <u>monthly</u> salary ($K) of an MSDS graduate. Suppose X~N(μ=10, σ=3).

$$Z = \frac{X - 10}{3}$$

is a new normal random variable scaled so that μ = 0 and σ = 1.

N(0,3) ⟹



Ways to compute a z-score and p-value

1. Use R

pnorm() function takes a z-score and returns the lower tail area under the curve.

qnorm() function take a p-value or the percentage of observations that lie towards the left of the X value that it corresponds to within the cumulative distribution function.

For our example of monthly MSDS salaries where X =15:

pnorm(1.66) =.9515428

qnorm(.9515428) = 1.66

According to this calculation, the area below 15 represents a proportion of .95 of MSDS graduates who had z-scores below 1.66.

2. Use probability table in Appendix C1.of Openintro pg. 410-411 or online

Z- score Exercise

This exercise is from Open Intro's Chapter 4, Q: 4.39, pg. 166

Auto Insurance premiums. Suppose a newspaper article states that the distribution of auto insurance premiums for residents of California is approximately normal with a mean of $1650. The article also states that 25% of California residents pay more than $1800.

a. What is the Z-score that corresponds to the top 25% (or the 75th percentile) of the standard normal distribution?
b. What is the mean insurance cost? What is the cutoff for the 75th percentile?
c. Identify the SD of insurance premiums in California.

Hint: you can find the z-score using the qnorm() function in R

Point estimates

So far, we have focused on exploring sample statistics, but we may also be interested in generalizing our findings to a much bigger population! For example, Pew Research conducts polls in order to understand public opinion on topics of interest. The aim of this work is to use the responses from a poll and estimate opinions of the broader U.S. population.

A Pew poll in 2022 asked: Do you think widespread use of driverless passenger vehicles would be a good idea for society?

46% of respondents said it would be a good idea.

The 46% is a point estimate of the approval rating we may see if we collected responses from the entire population.

The entire-population response proportion is a parameter of interest. We use the sample proportion to estimate our parameter.

Point estimates in Statistical Inference

- Sample Statistics becomes **point estimators** in statistical inferences.

- Point estimator $\bar{x}$ (sample mean) to estimate μ (population mean)
- Point estimator $\bar{p}$ (sample proportion) to estimate ρ (population proportion)
- Point estimator s (sample standard deviation) to estimate σ (population standard deviation)
  - sample standard deviation is the standard error of the mean

Central Limit Theorem

Definition:

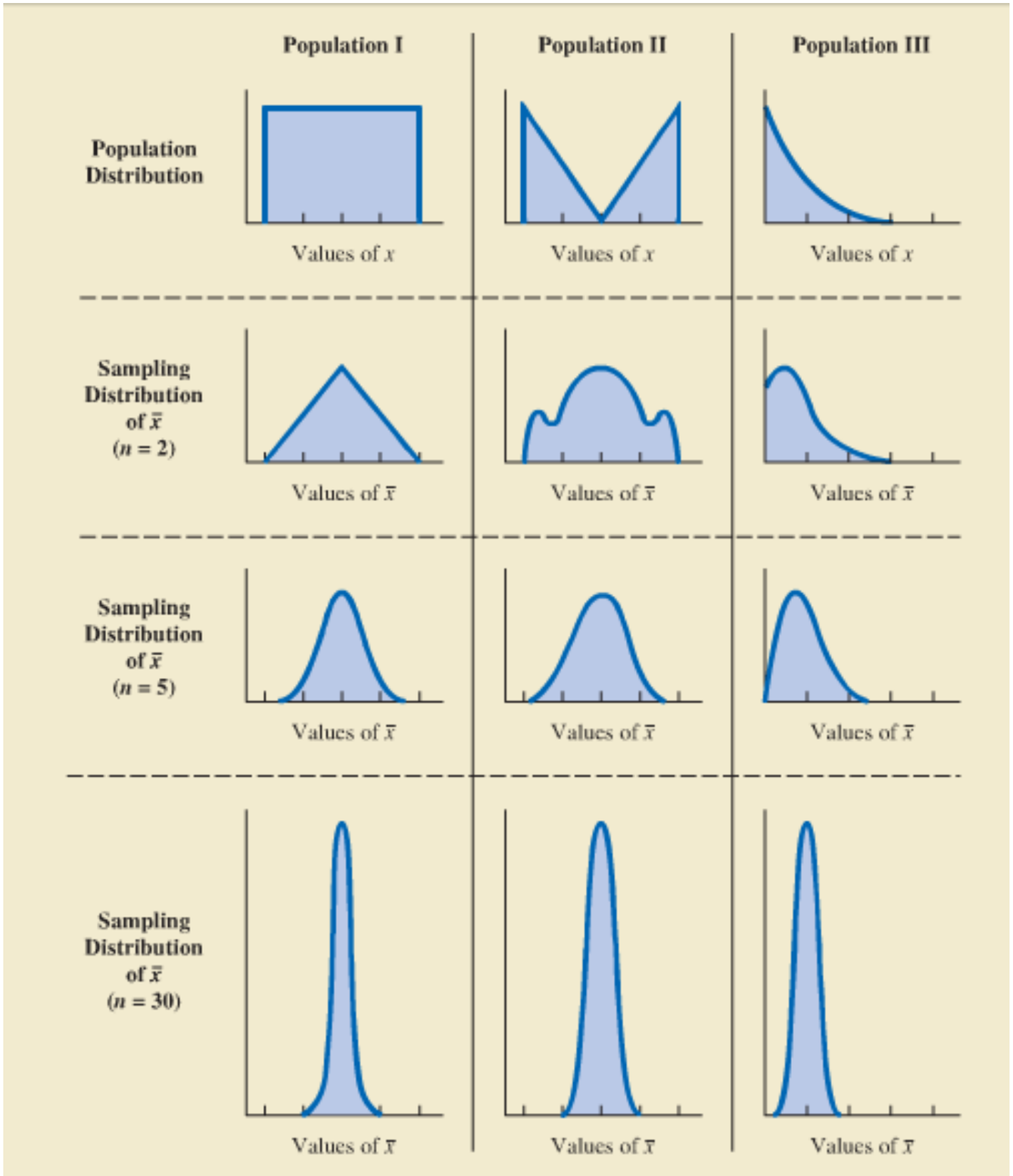
The observations are independent, and the sample size is sufficiently large, the sample proportion will follow a normal distribution with the following mean and standard error:

$$\mu_{\hat{p}} = p \qquad SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

In order for the Central Limit Theorem to hold, the sample size is typically considered sufficiently large when $np \geq 10$ and $n(1-p) \geq 10$, which is called the **success-failure condition**.

In selecting random samples of size *n* from a population, the sampling distribution of the sample mean $\bar{x}$ can be approximated by **a normal distribution** as the sample size becomes **large**.

Central Limit Theorem



As sample size becomes **30 or more**, the sampling distribution becomes a **normal** distribution

Hypothetical Sampling Example

Four individuals are available to be surveyed about their views on driverless cars (population), but you only survey two of them (sample), drawn randomly.

- Population is {1, 2, 3, 4}.
- 6 Possible samples: {1, 2}, {1, 3}, {1, 4}, {2, 3}, {2, 4}, {3, 4}.

Their responses are in a range of 1-5.

Survey Results

Participant	Survey Question Score
1	5
2	3
3	2
4	5

What is the average of responses in the population?

Mean of the population: $\mu = \frac{5+3+2+5}{4} = 3.75.$

Survey Results

Sample	Survey Question Score	Mean of the Sample ($\overline{x}$)
(1,2)	(5, 3)	4
(1,3)	(5, 2)	3.5
(1,4)	(5, 5)	5
(2,3)	(3, 2)	2.5
(2,4)	(3, 5)	4
(3,4)	(2, 5)	3.5

- None of the samples result in a mean of 3.75 as shown.
- The mean of the sample takes on a distribution of values.

Demo and Discussion

- Follow along with Lab_week3.Rmd
- Follow along with Discussion_Wk3.Rmd
  - Continue using this R file to look at the GAP21Q28 of the survey. What can you generalize to the U.S. population by looking at demographics of your choosing. Post to the discussion thread that includes:
  - Choose at least 2 demographics to compare their response on this variable
  - Your plot(s) - try to label axis and title the plots.
  - R file

Assignment

Using the Pew Research Center's data, (any wave) find something that seems interesting and pick two or three questions that might be correlated.

In the assignments folder post:

- Include the weight variable in your response estimates.
- Plot any visualizations as appropriate.
- A brief explanation ***1 paragraph !*** of your analysis process and your interpretation of the data. (What is the data, what did you do with the data, what do the results mean?)
- Your Rmd file and knitted pdf file