# Stat 201: Introduction to Statistics

Standard 23 – Sampling Distributions for Sample Means

# Recall Definitions from Ch 2

- **Statistic**: numerical summary of a sample
  - Mean($\bar{x}$), proportion($\hat{p}$),median, mode, standard deviation($s$), variance($s^2$), Q1, Q3, IQR, etc.
  - We use US alphabet letters to denote these
- **Parameter**: numerical summary of a population
  - Mean($\mu_x$), proportion($\rho$), median, mode, standard deviation($\sigma$), variance($\sigma^2$), Q1, Q3, IQR, etc.
  - We usually don’t know these values
  - We use Greek letters to denote these

# Sampling Distributions

- Intro: https://www.youtube.com/watch?v=DmZJ1blQOns

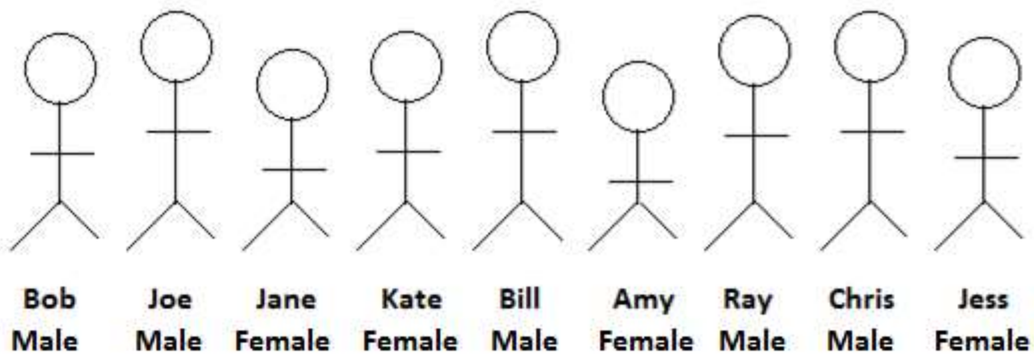
- A **sampling distribution** is the **probability distribution** that specifies probabilities for the possible values of the mean or proportion.
  - Proportions – consider the Binomial from Chapter 6
  - Means – consider the standard normal from Chapter 6
- A **sampling distribution** is a special case of a probability distribution where the outcome of an experiment that we are interested in is a sample statistic such as a **sample proportion($\hat{p}$)** or **sample mean ($\bar{x}$)**
  - It's the same as what we were doing before, but now instead of singular observations we're looking at groups

# Sampling Distributions

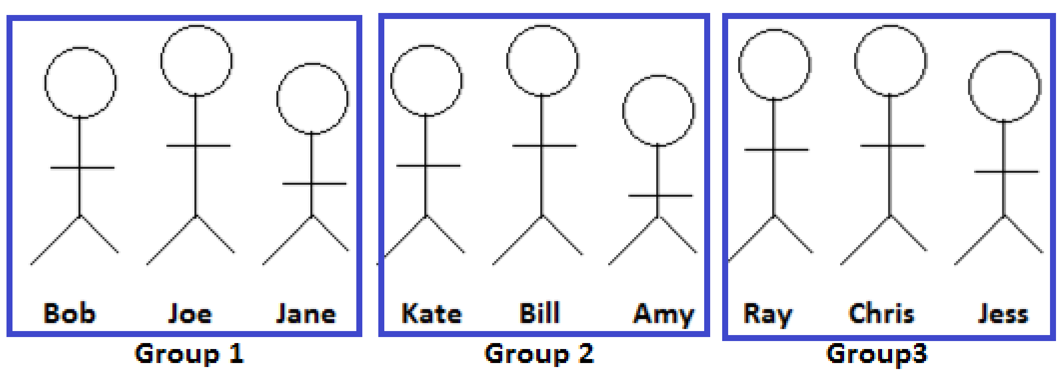
- This is confusing.
  - Remember, before we talked about events and random variables in n trials
  - Now, we're talking about m groups of n trials which yield m sample means or m sample proportions
    - $\overline{x_i} = \frac{\sum x}{n}\ for\ i = 1, 2, \ldots, m$
    - $\widehat{p_i} = \frac{x}{n}\ for\ i = 1, 2, \ldots, m$

# Sampling Distributions

- Variable: Gender of Students
  - Before, we measured individuals:



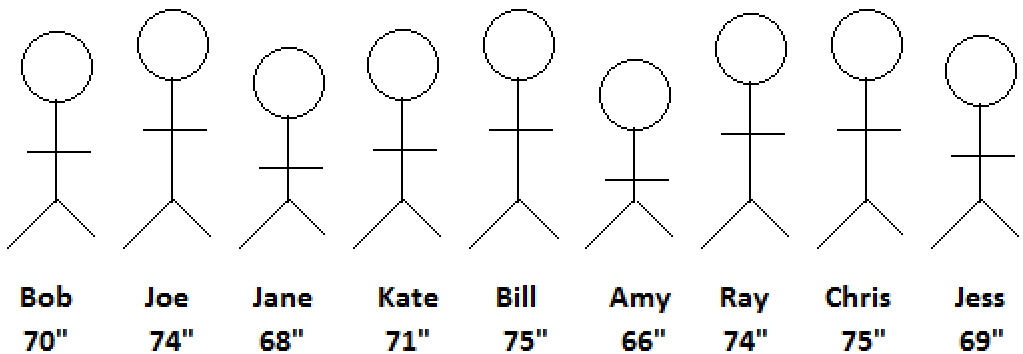

  - Now, we have one measurement across groups:



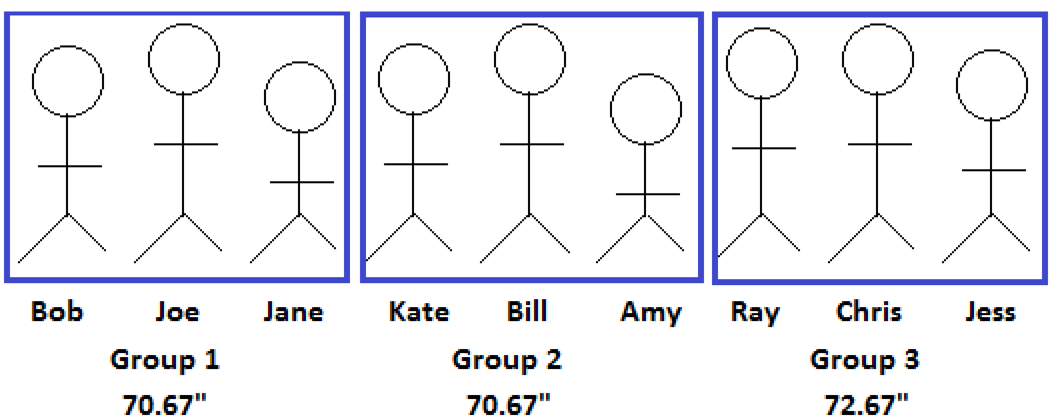

$\hat{p}$ = 2/3 = .6667 $\hat{p}$ = 1/3 = .3333 $\hat{p}$ = 2/3 = .6666

# Sampling Distributions

- Variable: Heights of Americans
  - Before, we measured individuals:

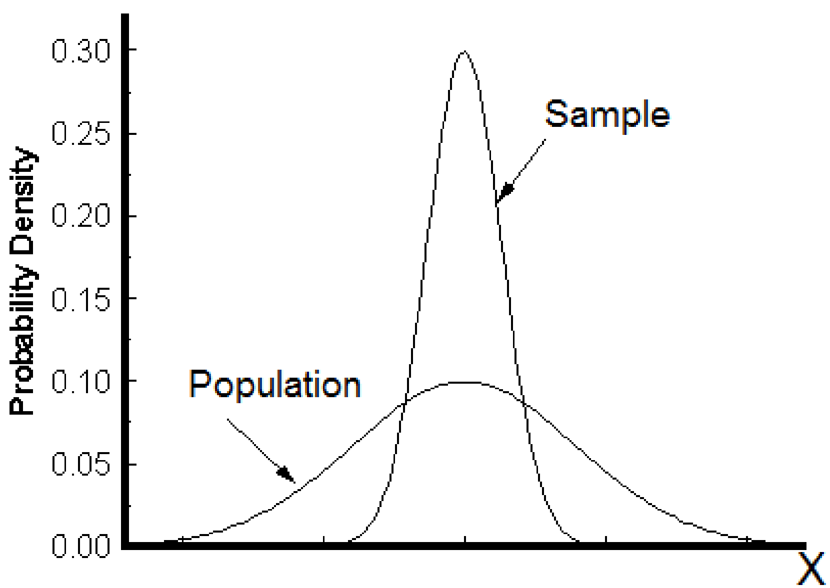



  - Now, we have one measurement across groups:



# Sampling Distribution - Graphs

- Sample vs. Population: the sampling distribution is narrower than the population because grouping the data reduces the variation; pay attention to the standard error equations



# Sampling Distributions: Means

- This first sampling distribution we'll talk about is the **sampling distribution for the sample mean**.

- The idea is that there is some **true population mean out there, $\mu$,** but it might not be feasible to know it
  - We may not have enough time or money to poll the population
  - It may be infeasible to get a population measure

# Sampling Distributions: Means

- Instead, we look at **sample mean, $\bar{x}$,** the mean of quantitative observations

- We've looked at this before in the **descriptive statistics** but now we're going to talk about **all possible sample means from repeated random samples from our population**

# Sampling Distributions: Means

- **Before we had quantitative observations:** $x_1, x_2, x_3, \ldots, x_n$

  - We would summarize all x's with one **sample mean, one $\boldsymbol{\bar{x}}$**

  - $\bar{x} = \dfrac{\text{the sum of x's}}{\text{the total sample size}} = \dfrac{\sum x}{\text{n}}$

    $=$ the mean of the observations in our sample

# Sampling Distributions: Means

- **Now we have m groups of n subjects with categorical observations:** $\{x_{1,1}, x_{1,2}, x_{1,3}, \ldots, x_{1,n}\}, \{x_{2,1}, x_{2,2}, x_{2,3}, \ldots, x_{2,n}\}, \ldots, \{x_{m,1}, x_{m,2}, x_{m,3}, \ldots, x_{m,n}\}$
- Now, we find summary statistics for each group

  $\overline{x_1}, \overline{x_2}, \overline{x_3}, \overline{x_4}, \ldots, \overline{x_m}$

  – We have m sample means, one $\bar{x}$ for each group

  - $\overline{x_1} = \frac{\text{the sum of x}'\text{s } from\ group\ 1}{\text{the total sample size of } group\ 1} = \frac{\sum x}{\text{n}}$
  - $\overline{x_2} = \frac{\text{the sum of x}'\text{s } from\ group\ 2}{\text{the total sample size of group 2}} = \frac{\sum x}{\text{n}} \ldots$
  - $\overline{x_m} = \frac{\text{the sum of x}'\text{s } from\ group\ m}{\text{the total sample size } of\ group\ m} = \frac{\sum x}{\text{n}}$

# Sampling Distributions: Means

- You could think of each group as a barrel and we're only interested in the mean of each barrel; we are no longer interested in the individual responses
- The example below shows how we could summarize 40 observations, into four representative sample means

| $\overline{x_1}$ | $\overline{x_2}$ | $\overline{x_3}$ | $\overline{x_4}$ |
|---|---|---|---|
| $x_1, x_2, \dots, x_{10}$ | $x_{11}, x_{12}, \dots, x_{20}$ | $x_{21}, x_{22}, \dots, x_{30}$ | $x_{31}, x_{32}, \dots, x_{40}$ |

# Sampling Distribution – Mean and SD

- The mean of the sampling distribution for a sample mean will always equal the population mean: $\boldsymbol{\mu_{\overline{x}} = \mu_x}$
  - This is the mean of all possible sample means, but we note that some $\boldsymbol{\overline{x}}$ will be lower and some will be higher

# Sampling Distribution – Mean and SD

- **Think about it this way:**
  - **Q:** If the population mean of time Americans spend on social media is 100 minutes with a standard deviation of 25 minutes what would you expect the average time a sample of 35 Americans spent on social media?
  - **A:** 100 minutes is our best guess.

- Later, we'll do this the other way around and we will call $\bar{x}$ the **point estimate for $\boldsymbol{\mu_x}$** since it's our best guess for the population mean if we don't know it

# Sampling Distribution – Mean and SD

- The standard error, the standard deviation of all possible sample means, is:

$$\boldsymbol{\sigma_{\overline{x}} = \frac{\sigma_x}{\sqrt{n}}}$$

$$= \boldsymbol{St.Dev}(\overline{x_1}, \overline{x_2}, \overline{x_3}, \overline{x_4}, \ldots, \overline{x_m})$$

# Sampling Distribution – Mean and SD

- **Think about it this way:**
  - **Q:** If our best guess for $\boldsymbol{\mu}$ is $\bar{x}$ we need a **measure of reliability** for our estimate
  - **A:** We'll talk more about this later, but our standard error calculator is a big part of this
- Later, in the case we don't know $\mu_x$ or $\sigma_x$ we're estimating it with our **point estimate** $\boldsymbol{\bar{x}}$
  - Recall: $\boldsymbol{\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}}$
  - Consider: $\boldsymbol{\frac{s_x}{\sqrt{n}}}$ **[Note: we estimate $\boldsymbol{\sigma_x = s_x}$]**

# Sampling Distribution – Mean and SD

- The mean of the sampling distribution for a sample mean will always equal the population mean: $\boldsymbol{\mu_{\overline{x}} = \mu_x}$
  - This is the mean of all possible sample means, but we note that some $\boldsymbol{\overline{x}}$ will be lower and some will be higher
- The standard error, the standard deviation of all possible sample means, is:

$$\boldsymbol{\sigma_{\overline{x}} = \frac{\sigma_x}{\sqrt{n}}}$$

# Sampling Distribution – Mean and SD

- $\boldsymbol{\mu_{\bar{x}} = mean\ of\ all\ sample\ means = \mu_x}$
  - Even though we know the mean is the population mean, we note that some $\bar{x}$ will be lower and some will be higher
- $\boldsymbol{\sigma_{\bar{x}} = the\ std.dev.of\ all\ sample\ means = \frac{\sigma_x}{\sqrt{n}}}$
- Aside:
    - What if we increase n?
      - The standard deviation shrinks
    - What if we decrease n?
      - The standard deviation grows

# Sampling Distribution:

- Now that we know the mean and standard error of the sample means we can calculate z-scores to find some probabilities associated with sample means just like we did before.

$$\boldsymbol{\mu_{\bar{x}} = \mu_x}$$

$$\boldsymbol{\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}}$$

$$z = \frac{observation - mean}{st.dev} = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu_x}{\frac{\sigma_x}{\sqrt{n}}}$$

# Sampling Distribution:

$$P(\bar{x} > c) = 1 - P\left(z < \frac{c - \mu_{\bar{x}}}{\sigma_{\bar{x}}}\right) = 1 - P\left(z < \frac{c - \mu_x}{\frac{\sigma_x}{\sqrt{n}}}\right)$$

$$P(\bar{x} < c) = P\left(z < \frac{c - \mu_{\bar{x}}}{\sigma_{\bar{x}}}\right) = P\left(z < \frac{c - \mu_x}{\frac{\sigma_x}{\sqrt{n}}}\right)$$

$$P(c_1 < \bar{x} < c_2) = P\left(z < \frac{c_2 - \mu_{\bar{x}}}{\sigma_{\bar{x}}}\right) - P\left(z < \frac{c_1 - \mu_{\bar{x}}}{\sigma_{\bar{x}}}\right)$$

$$= P\left(z < \frac{c_2 - \mu_x}{\frac{\sigma_x}{\sqrt{n}}}\right) - P\left(z < \frac{c_1 - \mu_x}{\frac{\sigma_x}{\sqrt{n}}}\right)$$

# Sampling Distribution Example 1

- The students that live in University of South Carolina Dormitories throw away an average of 600,000 beer cans per month with a standard deviation of 100,000 cans.
- Find the mean and standard error of the **sampling distribution for the sample mean** with n = 48 months.
- Let's find the sampling distribution!

# Sampling Distribution Example 1

- Let's find the sampling distribution mean:
- $\mu_{\bar{x}} = the\ mean\ of\ all\ possible\ sample\ means = \mu_x = the\ population\ mean$ = 600,000
  - Some $\bar{\boldsymbol{x}}$ will be lower and some will be higher but **the mean of all sample means of n=4 months will be 600,000**

# Sampling Distribution Example 1

- Let's find the sampling distribution st. deviation:
- $\sigma_{\bar{x}} = standard\ error$
  $= the\ standard\ deviation\ of\ all\ possible\ sample\ means\ of\ n = 4\ months$

$$= \frac{\sigma_x}{\sqrt{n}} = \frac{100{,}000}{\sqrt{48}} = 14433.7567$$

andard error, **the standard deviation of all possible**
**ll possible sample means of n=48 months**, is

# Sampling Distribution Example 1

- <u>Let’s find the sampling distribution:</u>

- $\mu_{\bar{x}} = \mu_x$ = 600,000

- $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{100{,}000}{\sqrt{48}} = 14433.7567$

# Sampling Distribution Example 1

- What is the probability that the sample mean number of beer cans thrown away per month for the University of South Carolina Dormitories for a random sample of four months is less than 550,000?
- $\mathrm{P}(\bar{x} < 550{,}000) = P\left(Z < \frac{550{,}000 - 600{,}000}{\frac{100{,}000}{\sqrt{48}}}\right) =$

$\mathrm{P}\left(\mathrm{Z} < \frac{550{,}000 - 600{,}000}{14433.7567}\right) = \mathrm{P}(\mathrm{Z} <$-3.46)=.0003

# Sampling Distribution Example 1

- What is the probability that the sample mean number of beer cans thrown away per month for the University of South Carolina Dormitories for a random sample of four months is less than 550,000?
- $\mathrm{P}(\bar{x} < 550{,}000) = .0003$
- This is an **very unusual** occurrence, we only see less than 550,000 cans thrown away .03% of the time
- **Note:** this assumes the number of beer cans thrown away follows the normal distribution – you'll see why soon.

# Sampling Distributions – Example 2

- Say, we know that **the average American spends 100 minutes on social media per day with a standard deviation of 25 minutes.**
- **What is the sampling distribution of the sample mean** of time Americans spend on social media for n=35?
  - Note, we aren't interested in the individuals but the group of thirty five
  - Here, X=the proportion of the ten Americans in each group

# Sampling Distributions – Example 2

- Say, we know that **the average American spends 100 minutes on social media per day with a standard deviation of 25 minutes.**
- **What is the sampling distribution of the sample mean** of time Americans spend on social media for n=35?

  - n = sample size = **sample size of thirty five** = 35
  - $\mu_x$ = population mean = 100
  - $\sigma_x$ = population standard deviation = 25

# Sampling Distributions – Example 2

- Let's find the sampling distribution mean:

- **The mean of all sample means of n=35**
  = $\mu_{\bar{x}} = \mu_x = 100$

  – Some $\overline{\boldsymbol{x}}$ will be lower and some will be higher but **the mean of all sample means of n=35 will be 100**

# Sampling Distributions – Example 2

- Let's find the sampling distribution st. error:

- **The st. deviation of all sample means of n=35**

  = Standard Error

$$= \boldsymbol{\sigma_{\bar{x}}} = \frac{\boldsymbol{\sigma_x}}{\sqrt{\boldsymbol{n}}} = \frac{25}{\sqrt{35}} = 4.2258$$

# Sampling Distributions – Example 2

- Let's find the sampling distribution :

$$\mu_{\bar{x}} = \mu_x = 100$$

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{25}{\sqrt{35}} = 4.2258$$

# Sampling Distributions – Example 2

- The probability that a sample of n=35 spend **more than two hours** on social media on average:

$$P(\bar{x} > 120) = P\left(z > \frac{120 - 100}{4.2258}\right) = P(Z > 4.73)$$

$$= 1 - P(Z < 4.73) \approx 1 - 1$$

$$= 0$$

# Sampling Distributions – Example 2

- The probability that a sample of n=35 spend **less than one hour** on social media on average:

$$P(\bar{x} > 60) = P\left(z < \frac{60 - 100}{4.2258}\right) = P(Z < -9.47)$$

$$= P(Z < -9.47)$$

$$\approx 0$$

# Sampling Distributions – Example

- The probability that a sample of n=35 spend **between 1 and 1.5 hours** on social media on average:

$$P(60 < \bar{x} < 90) = P\left(\frac{90-100}{4.2258} < z < \frac{60-100}{4.2258}\right)$$
$$= P(Z < -2.37) - P(Z < -9.47)$$
$$\approx .0089 - 0$$
$$= 0$$

# Sampling Distributions – Example

- **<u>Note:</u>** we had to assume normality of $\bar{x}$ to use the Z-score transformation to solve the previous probabilities

- We are able to make that assumption – unlocking all of the nice methodologies of the Normal distribution – by utilizing the central limit theorem

# Central Limit Theorem: Means

- For random sampling with a **large sample size n, the sampling distribution of the sample mean** is approximately a normal distribution
  – For us, 30 is close enough to infinity

- Introduction:
  – https://www.youtube.com/watch?v=Pujol1yC1_A

# Central Limit Theorem: Means

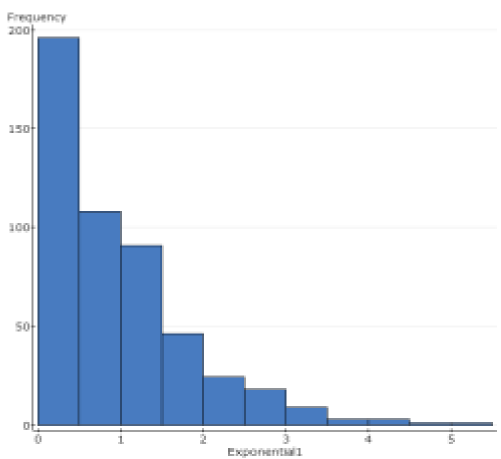
1) For any population the sampling distribution of $\bar{x}$ is bell shaped when the sample size n is large, when n is thirty or more
2) The sampling distribution of $\bar{x}$ is bell-shaped when the population distribution is distribution is bell-shaped, regardless of sample size
3) We do not know the shape of the sampling distribution of $\bar{x}$ if the sample size is small and the population distribution isn’t bell-shaped

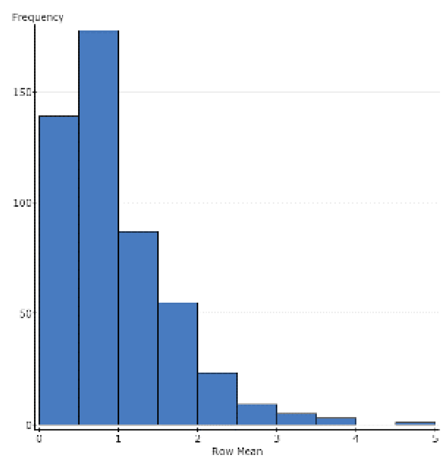
# Central Limit Theorem

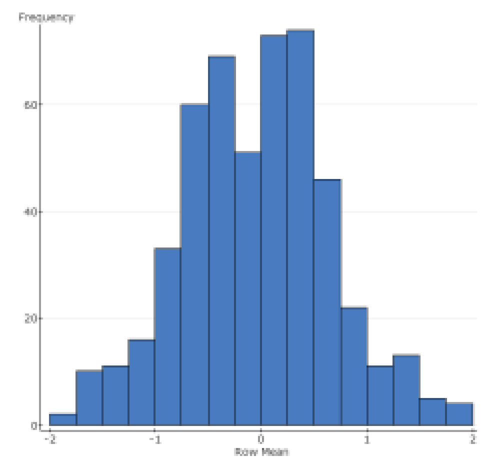
For any population the sampling distribution of $\bar{x}$ is bell shaped when the sample size n is large, when n is thirty or more

**Note:** for small sample size we can't say this.

Population $\qquad$ $\bar{x}$ when n=2 $\qquad$ $\bar{x}$ when n=30



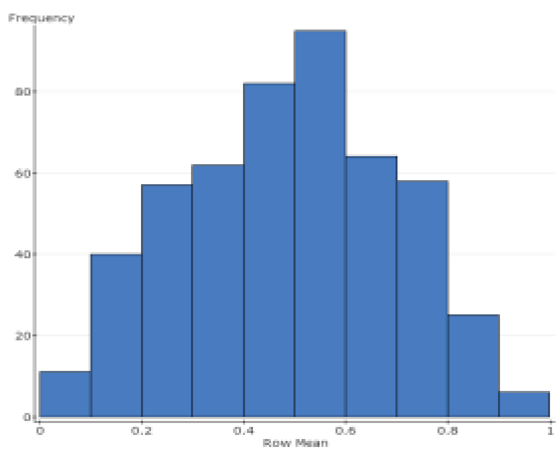



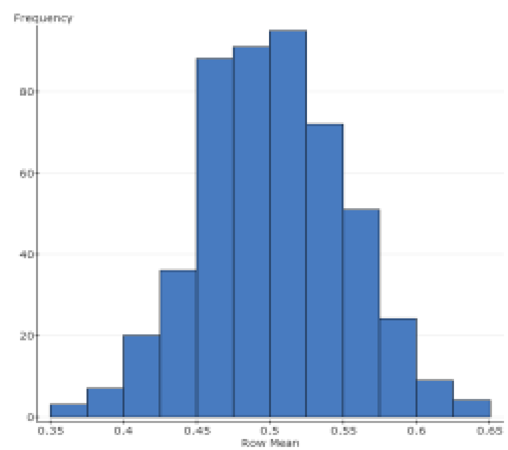
# Central Limit Theorem

The sampling distribution of $x_{bar}$ is bell-shaped when the population distribution is distribution is bell-shaped, regardless of sample size

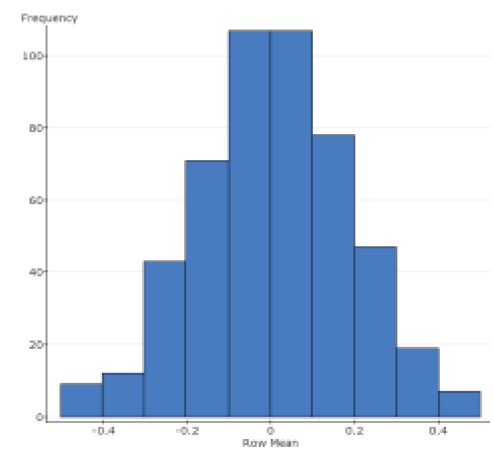
Population




$\bar{x}$ when n=2




$\bar{x}$ when n=30



# Sampling Distribution for the Sample Mean Summary

| Shape, Center and Spread of Population | Shape of sample | Center of sample | Spread of sample |
|---|---|---|---|
| Population is normal with mean μ and standard deviation $\sigma$. | Regardless of the sample size n, the shape of the distribution of the sample mean is normal | $\mu_{\bar{x}} = \mu$ | $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$ |
| Population is not normal with mean μ and standard deviation $\sigma$. | As the sample size n increases, the distribution of the sample mean becomes approximately normal | $\mu_{\bar{x}} = \mu$ | $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$ |

# Sampling Distribution:

$$P(\bar{x} > c) = 1 - P\left(z < \frac{c - \mu_{\bar{x}}}{\sigma_{\bar{x}}}\right) = 1 - P\left(z < \frac{c - \mu_x}{\frac{\sigma_x}{\sqrt{n}}}\right)$$

$$P(\bar{x} < c) = P\left(z < \frac{c - \mu_{\bar{x}}}{\sigma_{\bar{x}}}\right) = P\left(z < \frac{c - \mu_x}{\frac{\sigma_x}{\sqrt{n}}}\right)$$

$$P(c_1 < \bar{x} < c_2) = P\left(z < \frac{c_2 - \mu_{\bar{x}}}{\sigma_{\bar{x}}}\right) - P\left(z < \frac{c_1 - \mu_{\bar{x}}}{\sigma_{\bar{x}}}\right)$$

$$= P\left(z < \frac{c_2 - \mu_x}{\frac{\sigma_x}{\sqrt{n}}}\right) - P\left(z < \frac{c_1 - \mu_x}{\frac{\sigma_x}{\sqrt{n}}}\right)$$