# Sampling and Sampling Distributions

# Statistics

- **Descriptive statistics**
  - Collecting, presenting, and describing data
- **Inferential statistics**
  - Drawing conclusions and/or making decisions concerning a population based only on sample data

# Populations and Samples

- A **Population** is the set of all items or individuals of interest

  - Examples: All likely voters in the next election
    All parts produced today
    All sales receipts for November

- A **Sample** is a subset of the population

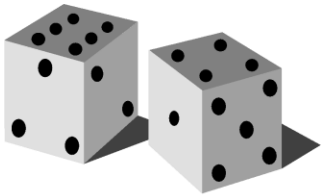
  - Examples: 1000 voters selected at random for interview
    A few parts selected for destructive testing
    Random receipts selected for audit

# Why Sample?

- Less time consuming than a census
- Less costly to administer than a census
- It is possible to obtain statistical results of a sufficiently high precision based on samples.

# Simple Random Samples

- Every object in the population has an equal chance of being selected
- Objects are selected independently
- Samples can be obtained from a table of random numbers or computer random number generators



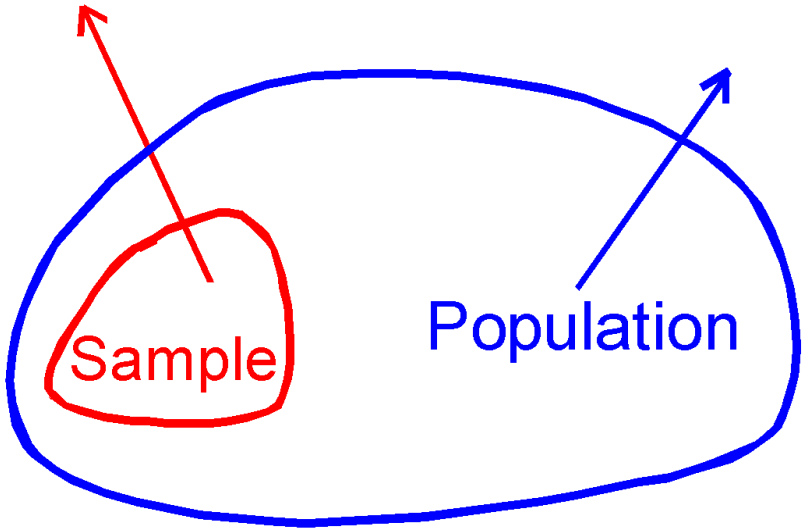
- A simple random sample is the ideal against which other sample methods are compared

# Inferential Statistics

- Making statements about a population by examining sample results

Sample statistics → Population parameters

(known) **Inference** (unknown, but can be estimated from sample evidence)



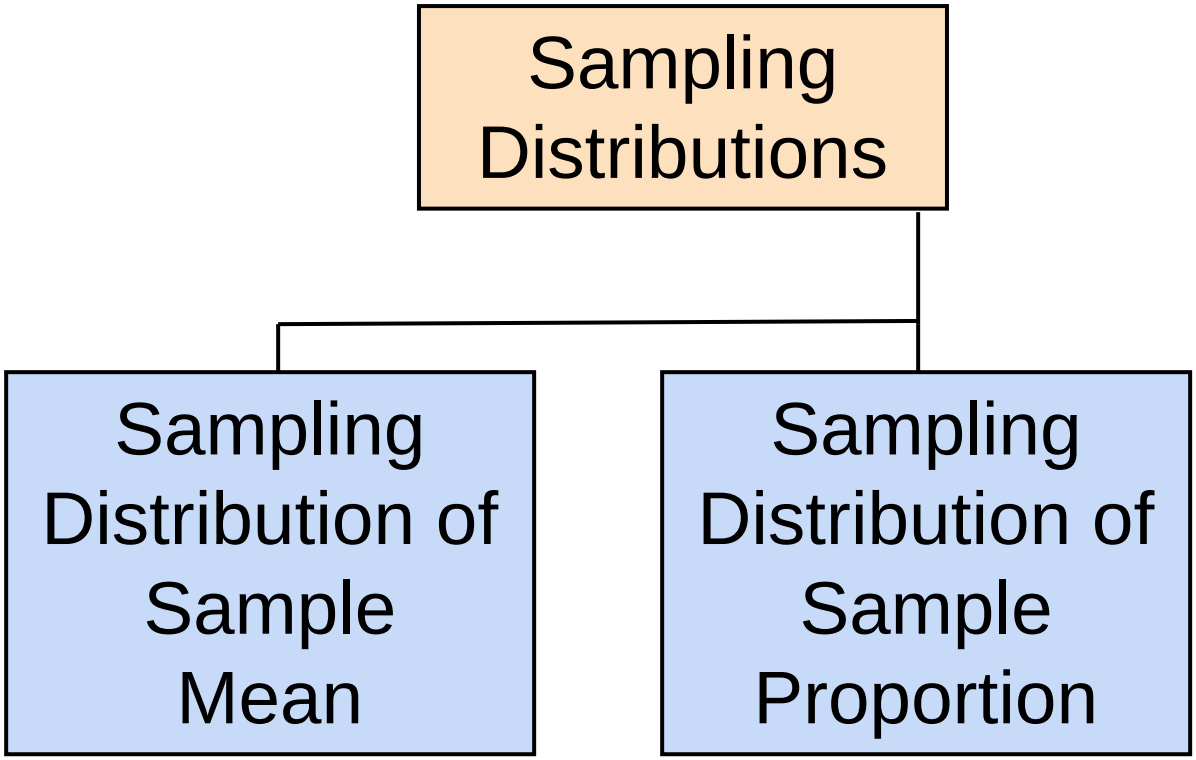
# Inferential Statistics

**Drawing conclusions and/or making decisions concerning a population based on sample results.**

- **Estimation**
  - e.g., Estimate the population mean weight using the sample mean weight
- **Hypothesis Testing**
  - e.g., Use sample evidence to test the claim that the population mean weight is 120 pounds

# Sampling Distributions

- A sampling distribution is a distribution of all of the possible values of a statistic for a given size sample selected from a population


Sampling Distributions
Sampling Distribution of Sample Mean
Sampling Distribution of Sample Proportion

# Developing a Sampling Distribution

- **Assume there is a population …**
- Population size N=4
- Random variable, X, is age of individuals
- Values of X:

  18, 20, 22, 24 (years)

# Developing a Sampling Distribution

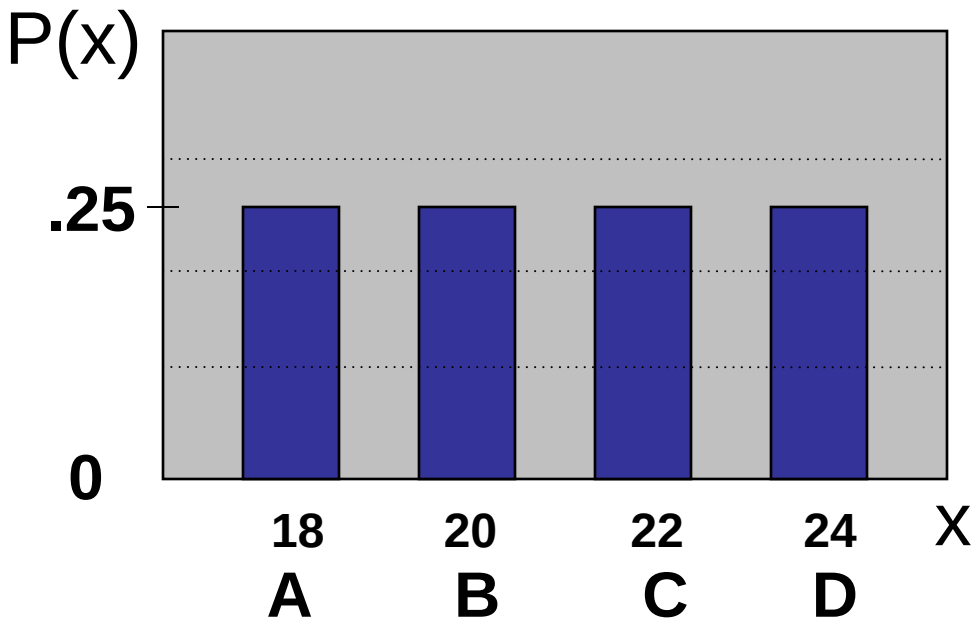
*(continued)*

Summary Measures for the Population Distribution:

$$\mu = \frac{\sum X_i}{N}$$

$$= \frac{18 + 20 + 22 + 24}{4} = 21$$

$$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}} = 2.236$$




Uniform Distribution

# Developing a Sampling Distribution

*(continued)*

Now consider all possible samples of size n = 2

| 1st Obs | 2nd Observation 18 | 20 | 22 | 24 |
|---|---|---|---|---|
| 18 | 18,18 | 18,20 | 18,22 | 18,24 |
| 20 | 20,18 | 20,20 | 20,22 | 20,24 |
| 22 | 22,18 | 22,20 | 22,22 | 22,24 |
| 24 | 24,18 | 24,20 | 24,22 | 24,24 |

16 possible samples (sampling with replacement)

16 Sample Means

| 1st Obs | 2nd Observation 18 | 20 | 22 | 24 |
|---|---|---|---|---|
| 18 | 18 | 19 | 20 | 21 |
| 20 | 19 | 20 | 21 | 22 |
| 22 | 20 | 21 | 22 | 23 |
| 24 | 21 | 22 | 23 | 24 |

# Developing a Sampling Distribution

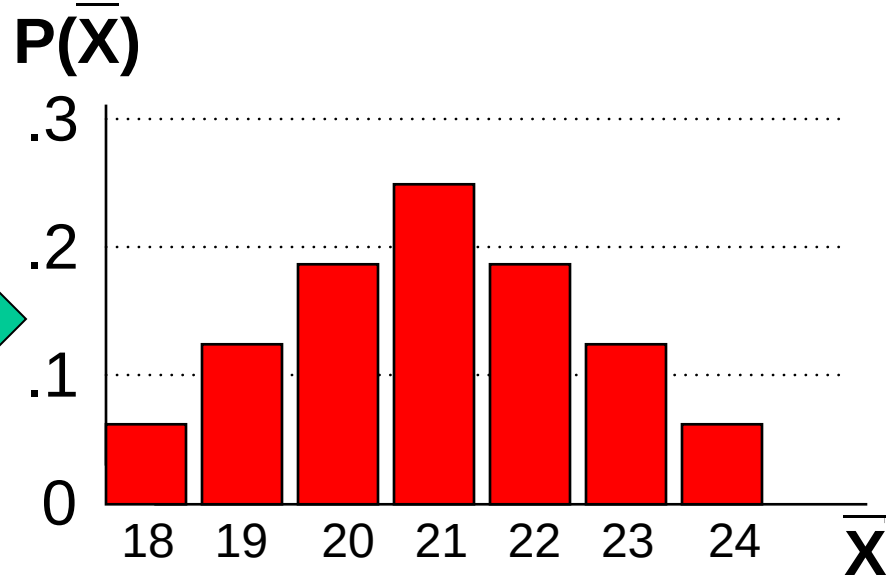
*(continued)*

## Sampling Distribution of All Sample Means

16 Sample Means

| 1st Obs | 2nd Observation 18 | 20 | 22 | 24 |
|---|---|---|---|---|
| **18** | 18 | 19 | 20 | 21 |
| **20** | 19 | 20 | 21 | 22 |
| **22** | 20 | 21 | 22 | 23 |
| **24** | 21 | 22 | 23 | 24 |

Sample Means Distribution




(no longer uniform)

# Developing a Sampling Distribution

*(continued)*

Summary Measures of this Sampling Distribution:

$$E(\bar{X}) = \frac{\sum \bar{X}_i}{N} = \frac{18 + 19 + 21 + \square + 24}{16} = 21 = \mu$$

$$\sigma_{\bar{X}} = \sqrt{\frac{\sum (\bar{X}_i - \mu)^2}{N}}$$

$$= \sqrt{\frac{(18-21)^2 + (19-21)^2 + \square + (24-21)^2}{16}} = 1.58$$

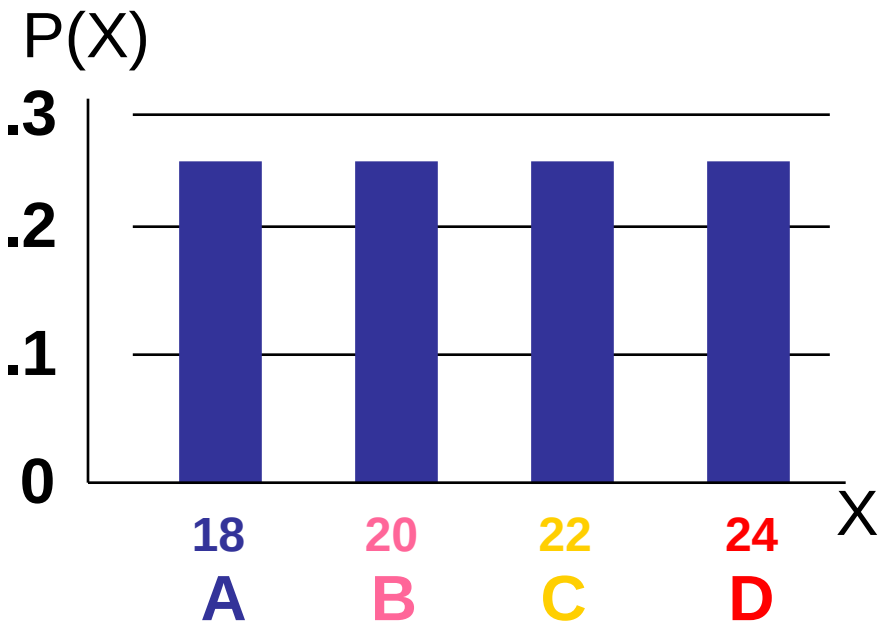
# Comparing the Population with its Sampling Distribution

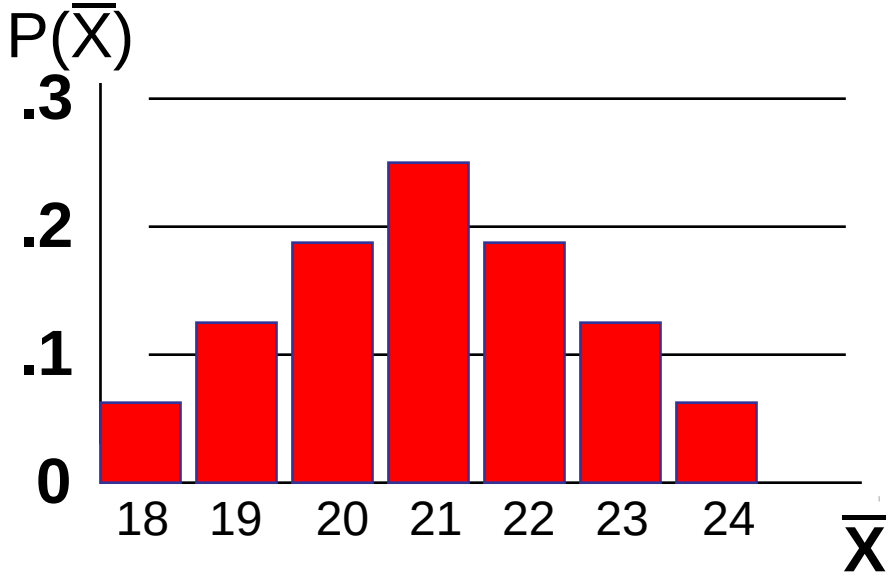
**Population**
**N = 4**

$\mu = 21 \quad \sigma = 2.236$

**Sample Means Distribution**
**n = 2**

$\mu_{\bar{X}} = 21 \quad \sigma_{\bar{X}} = 1.58$



# Expected Value of Sample Mean

- Let $X_1, X_2, \ldots X_n$ represent a random sample from a population

- The sample mean value of these observations is defined as

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

# Standard Error of the Mean

- Different samples of the same size from the same population will yield different sample means
- A measure of the variability in the mean from sample to sample is given by the Standard Error of the Mean:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

- Note that the standard error of the mean decreases as the sample size increases

# If the Population is Normal

- If a population is normal with mean μ and standard deviation σ, the sampling distribution of $\bar{X}$ is also normally distributed with

$$\mu_{\bar{X}} = \mu \quad \text{and} \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

# Z-value for Sampling Distribution of the Mean

- Z-value for the sampling distribution of $\overline{X}$:

$$Z = \frac{(\overline{X} - \mu)}{\sigma_{\overline{x}}} = \frac{(\overline{X} - \mu)}{\frac{\sigma}{\sqrt{n}}}$$

where: $\overline{X}$ = sample mean

$\mu$ = population mean

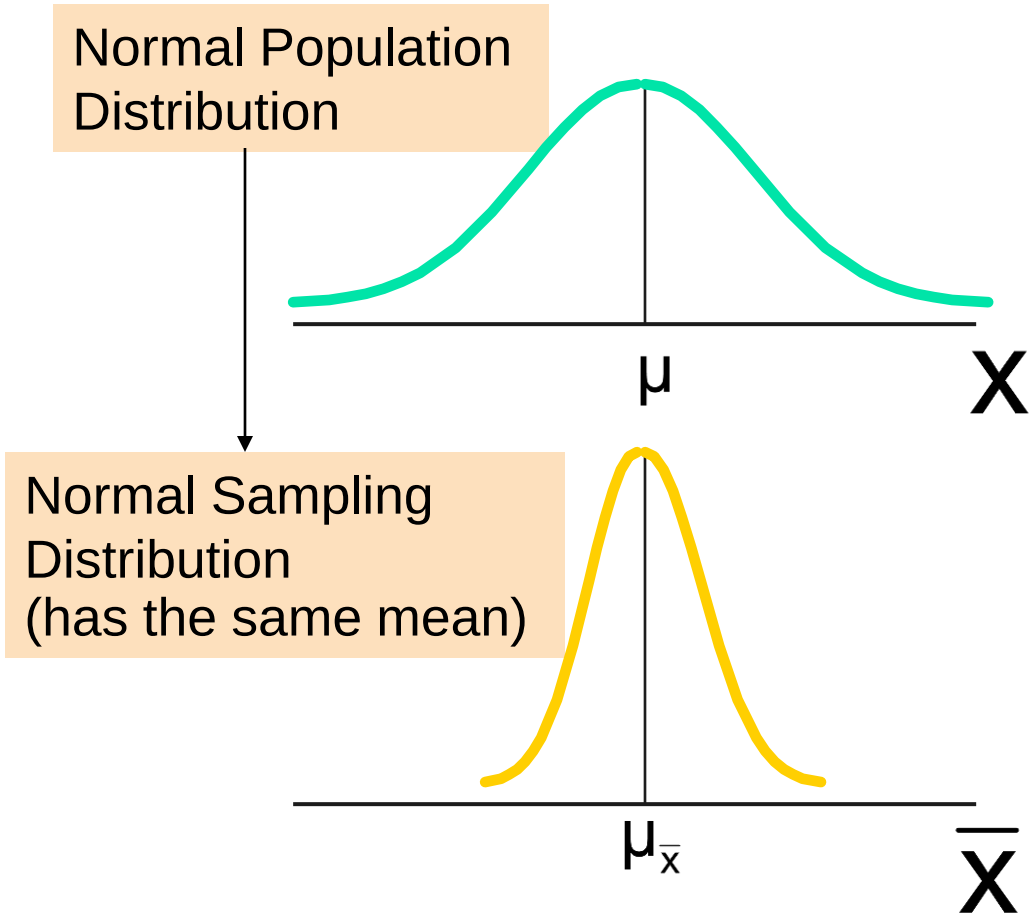
$\sigma$ = population standard deviation

$n$ = sample size

# Sampling Distribution Properties

$$\mu_{\bar{x}} = \mu$$

(i.e. $\bar{X}$ is unbiased)



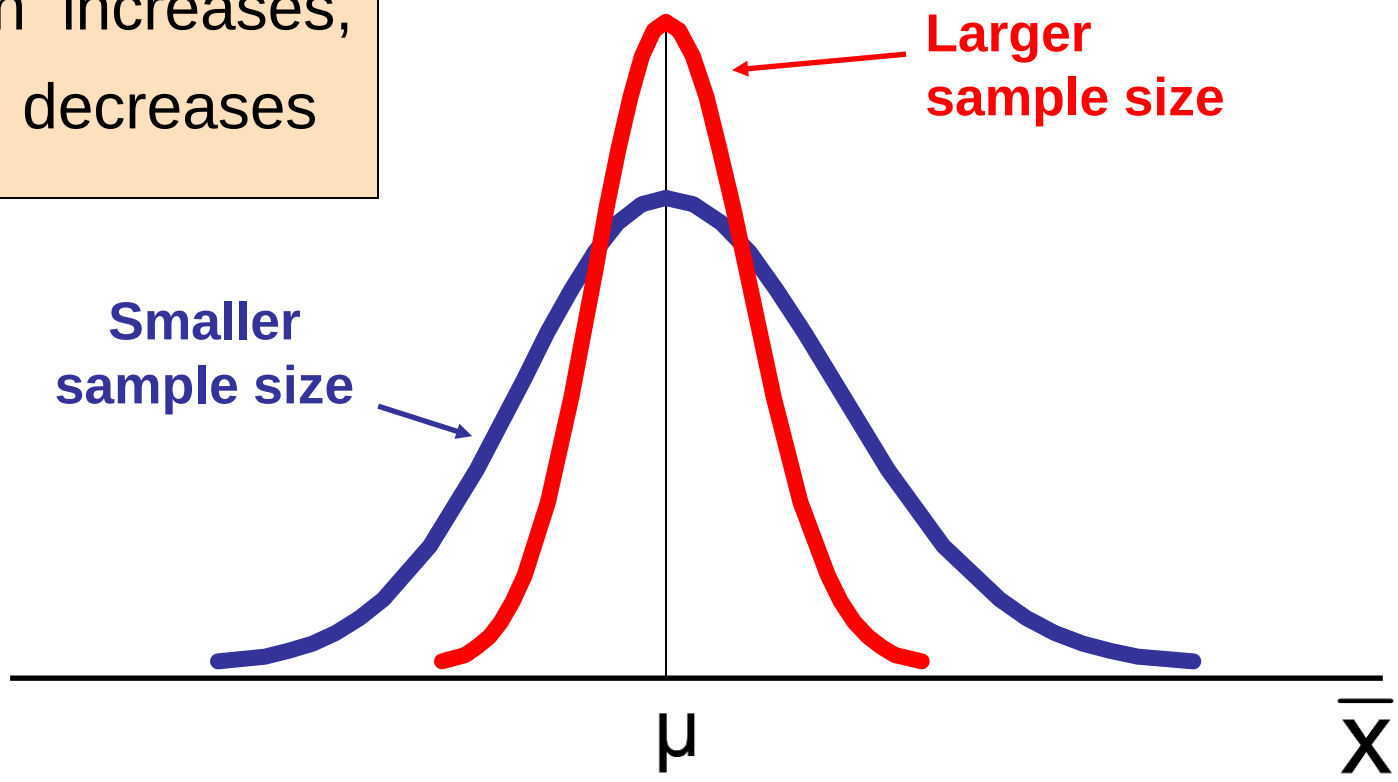
# Sampling Distribution Properties

*(continued)*

- For sampling with replacement:

As n increases,

$\sigma_{\bar{x}}$ decreases



# If the Population is **not** Normal

- We can apply the Central Limit Theorem:
  - Even if the population is not normal,
  - …sample means from the population will be approximately normal as long as the sample size is large enough.

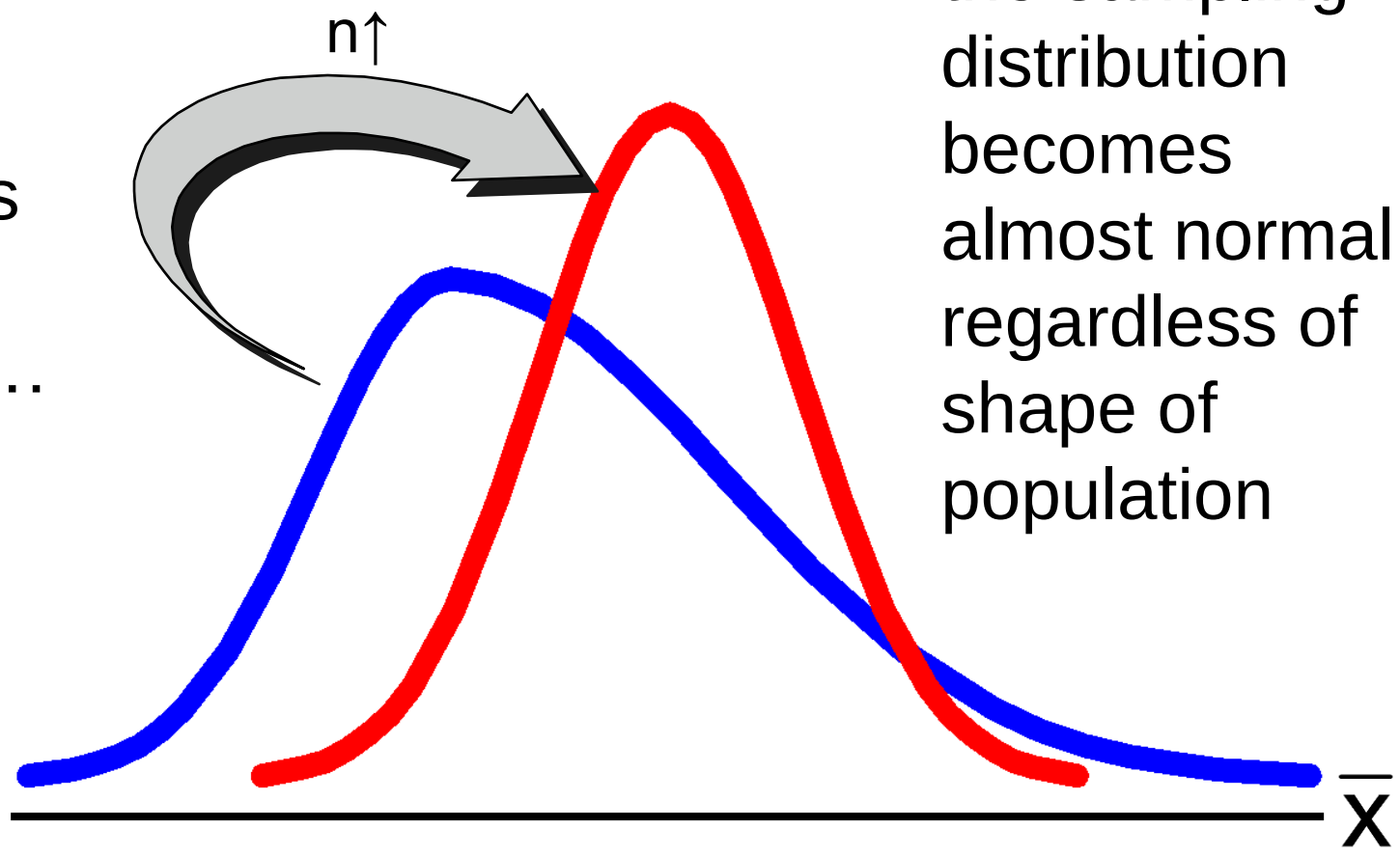
---

Properties of the sampling distribution:

$$\mu_{\bar{x}} = \mu \quad \text{and} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

# Central Limit Theorem

As the sample size gets large enough…




the sampling distribution becomes almost normal regardless of shape of population

# If the Population is **not** Normal

*(continued)*

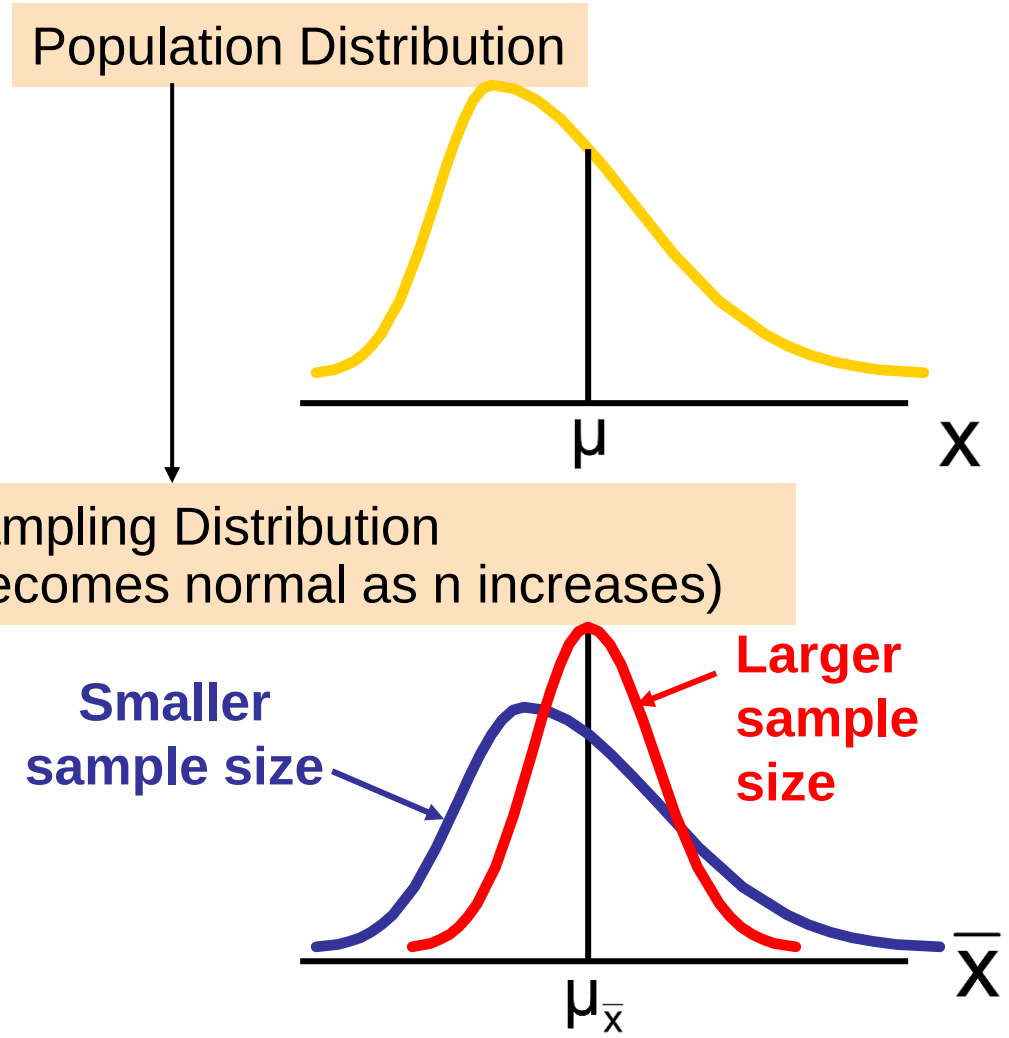
Sampling distribution properties:

Central Tendency

$$\mu_{\bar{x}} = \mu$$

Variation

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$



# How Large is Large Enough?

- For most distributions, $n > 25$ will give a sampling distribution that is nearly normal
- For normal population distributions, the sampling distribution of the mean is always normally distributed

# Example

- Suppose a population has mean $\mu = 8$ and standard deviation $\sigma = 3$. Suppose a random sample of size $n = 36$ is selected.

- What is the probability that the sample mean is between 7.8 and 8.2?

# Sampling Distributions of Sample Proportions

P = the proportion of the population having some characteristic

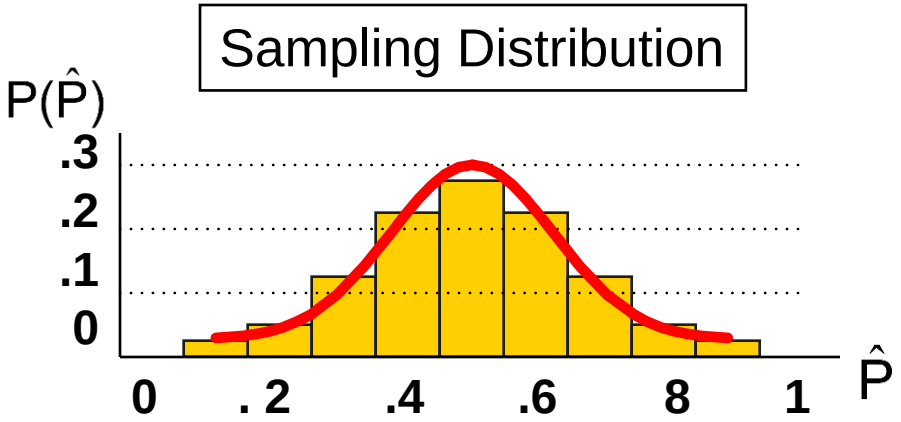
- Sample proportion ($\hat{P}$) provides an estimate of P:

$$\hat{P} = \frac{X}{n} = \frac{\text{number of items in the sample having the characteristic of interest}}{\text{sample size}}$$

- $0 \le \hat{P} \le 1$
- $\hat{P}$ has a binomial distribution, but can be approximated by a normal distribution when $nP(1 - P) > 9$

# Sampling Distribution of $\hat{P}$

- Normal approximation:




Properties:

$$E(\hat{P}) = p \quad \text{and} \quad \sigma^2_{\hat{P}} = Var\left(\frac{X}{n}\right) = \frac{P(1-P)}{n}$$

(where P = population proportion)

# Z-Value for Proportions

Standardize $\hat{P}$ to a Z value with the formula:

$$Z = \frac{\hat{P} - P}{\sigma_{\hat{P}}} = \frac{\hat{P} - P}{\sqrt{\frac{P(1-P)}{n}}}$$

# Example

- If the true proportion of voters who support Proposition A is P = .4, what is the probability that a sample of size 200 yields a sample proportion between .40 and .45?

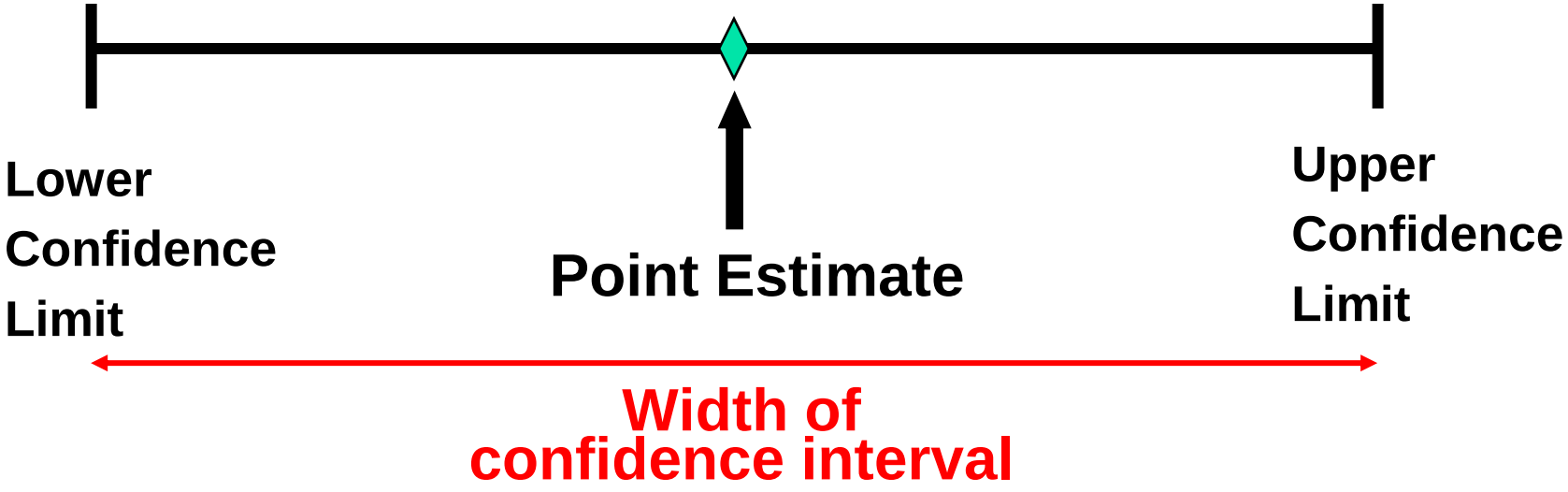
# Point estimate and confidence interval estimate

# Definitions

- An estimator of a population parameter is
  - a random variable that depends on sample information . . .
  - whose value provides an approximation to this unknown parameter
- A specific value of that random variable is called an estimate

# Point and Interval Estimates

- A point estimate is a single number,
- a confidence interval provides additional information about variability



# Point Estimates

| We can estimate a Population Parameter … | | with a Sample Statistic (a Point Estimate) |
|---|---|---|
| Mean | $\mu$ | $\bar{x}$ |
| Proportion | P | $\hat{p}$ |

# Confidence Intervals

- How much uncertainty is associated with a point estimate of a population parameter?
- An interval estimate provides more information about a population characteristic than does a point estimate
- Such interval estimates are called confidence intervals

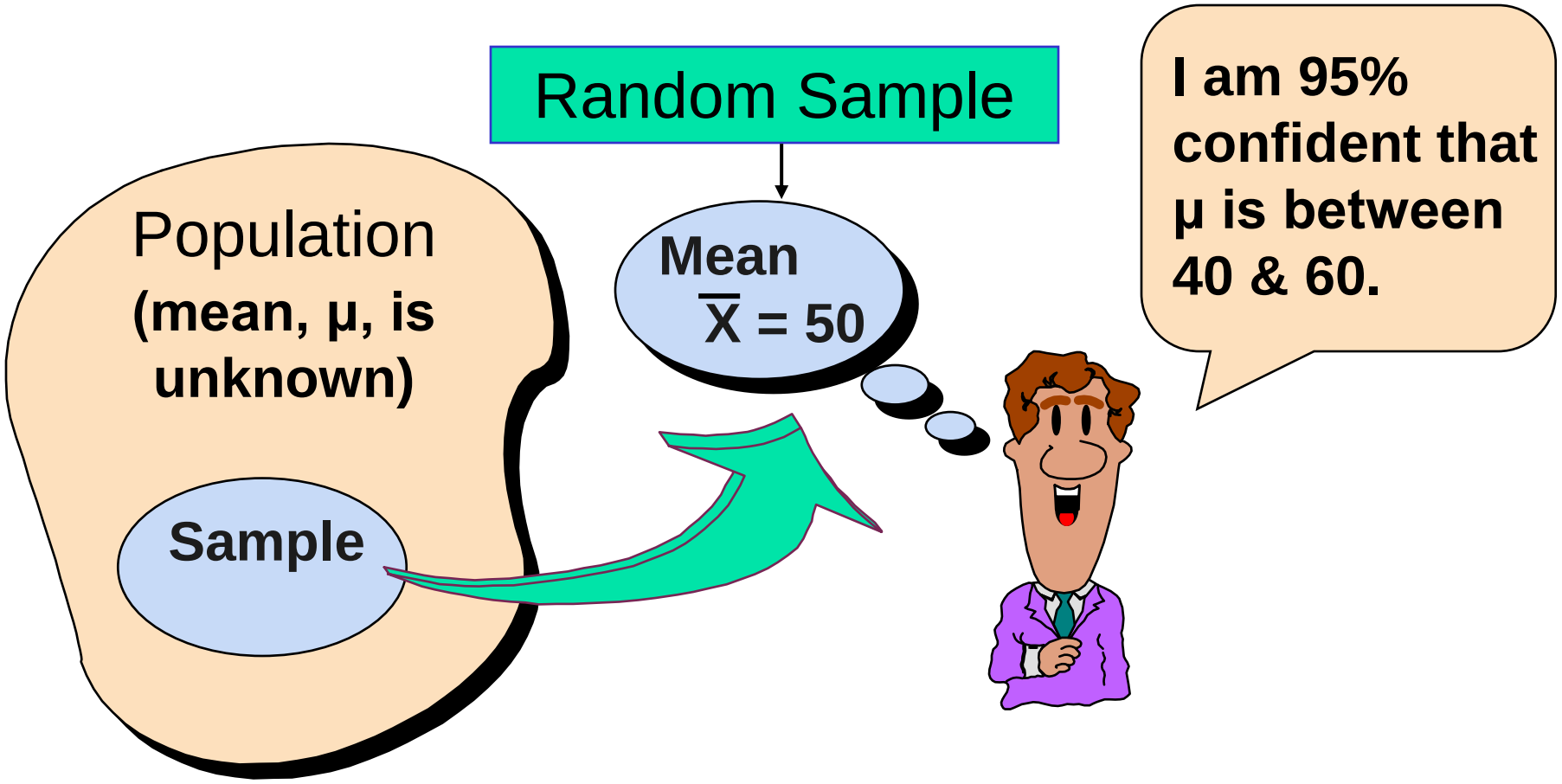
# Confidence Interval Estimate

- An interval gives a range of values:
  - Takes into consideration variation in sample statistics from sample to sample
  - Based on observation from 1 sample
  - Gives information about closeness to unknown population parameters
  - Stated in terms of level of confidence
    - Can never be 100% confident

# Confidence Interval and Confidence Level

- If P(a < **θ** < b) = 1 - **α** then the interval from a to b is called a 100(1 - **α**)% confidence interval of **θ**.
- The quantity (1 - **α**) is called the confidence level of the interval (**α** between 0 and 1)
  - In repeated samples of the population, the true value of the parameter **θ** would be contained in 100(1 - **α**)% of intervals calculated this way.
  - The confidence interval calculated in this manner is written as a < **θ** < b with 100(1 - **α**)% confidence

# Estimation Process



# Confidence Level, (1-α)

*(continued)*

- Suppose confidence level = 95%
- Also written (1 - **α**) = 0.95
- A relative frequency interpretation:
  - From repeated samples, 95% of all the confidence intervals that can be constructed will contain the unknown true parameter
- A specific interval either will contain or will not contain the true parameter
  - No probability involved in a specific interval

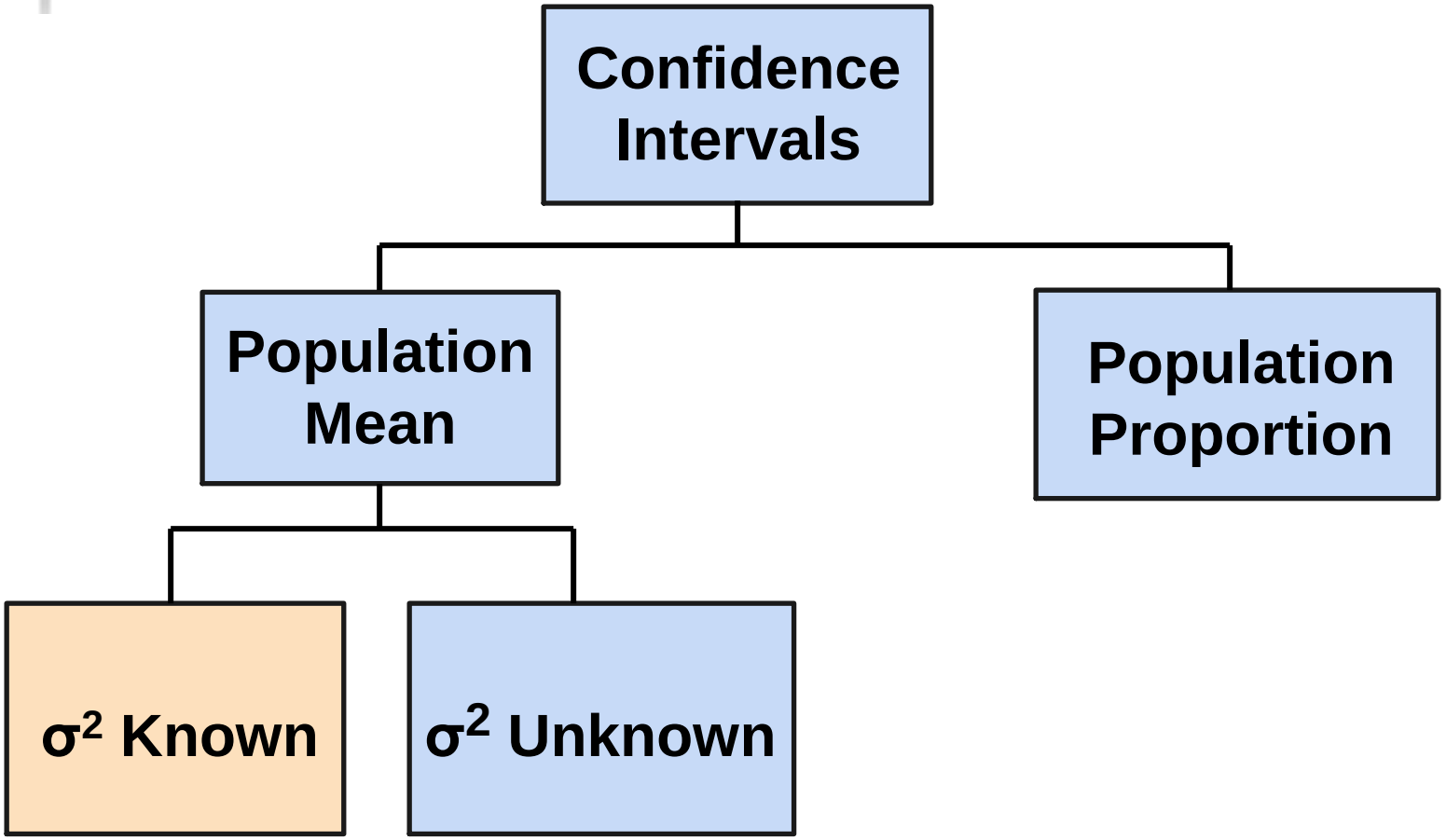
# General Formula

- The general formula for all confidence intervals is:

**Point Estimate ± (Reliability Factor)(Standard Error)**

- The value of the reliability factor depends on the desired level of confidence

# Confidence Intervals



# Confidence Interval for μ ($\sigma^2$ Known)

$$Z = \frac{(\bar{X} - \mu)}{\sigma_{\bar{x}}} = \frac{(\bar{X} - \mu)}{\frac{\sigma}{\sqrt{n}}}$$

- Assumptions
  - Population variance $\sigma^2$ is known
  - Population is normally distributed
  - If population is not normal, use large sample
- Confidence interval estimate:

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

(where $z_{\alpha/2}$ is the normal distribution value for a probability of $\alpha/2$ in each tail)

# Margin of Error

- The confidence interval,

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

- Can also be written as $\bar{x} \pm ME$ where ME is called the margin of error

$$ME = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

- The interval width, w, is equal to twice the margin of error

# Reducing the Margin of Error

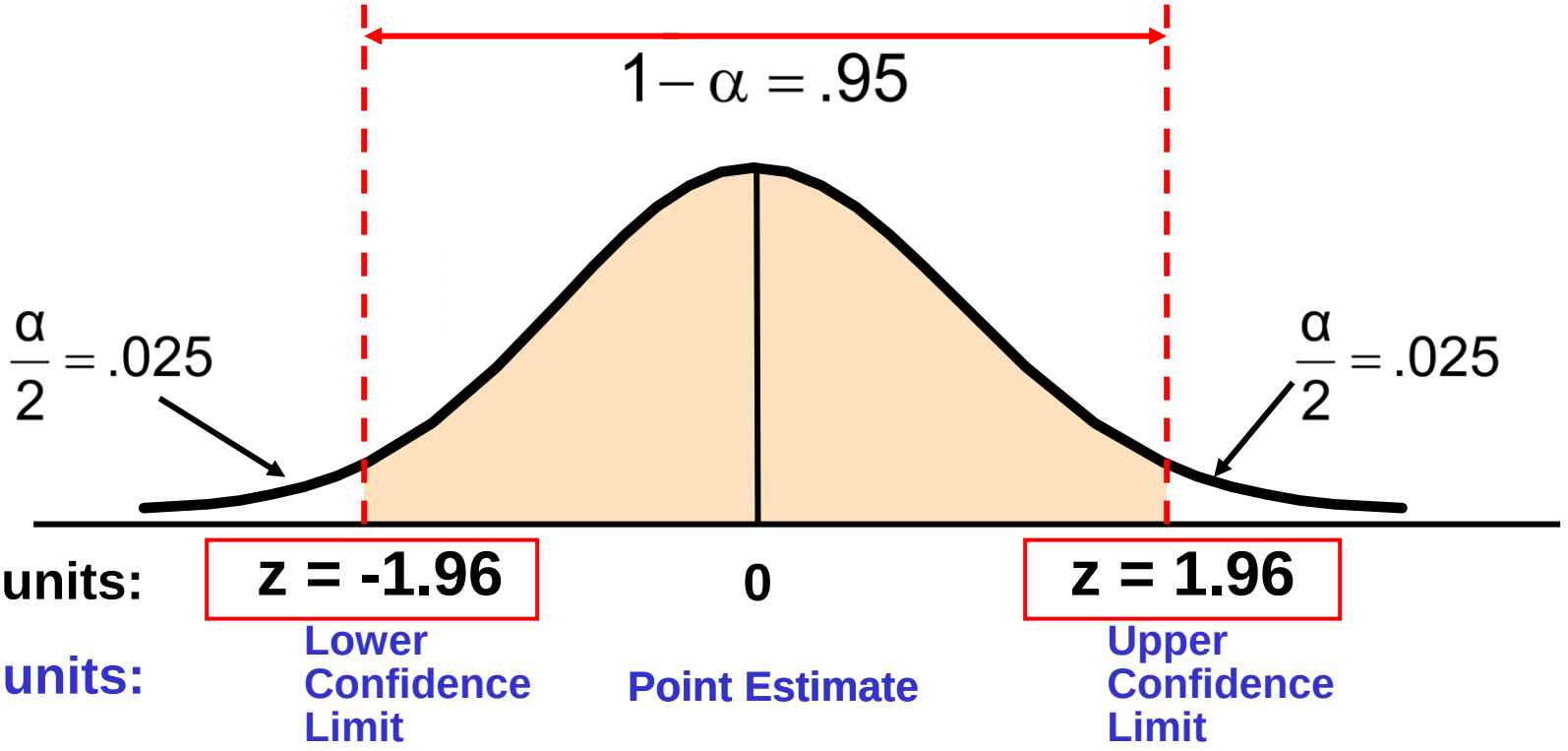
$$ME = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

The margin of error can be reduced if

- the population standard deviation can be reduced ($\sigma\downarrow$)
- The sample size is increased ($n\uparrow$)
- The confidence level is decreased, $(1 - \alpha)\downarrow$

# Finding the Reliability Factor, $z_{\alpha/2}$

- Consider a 95% confidence interval:



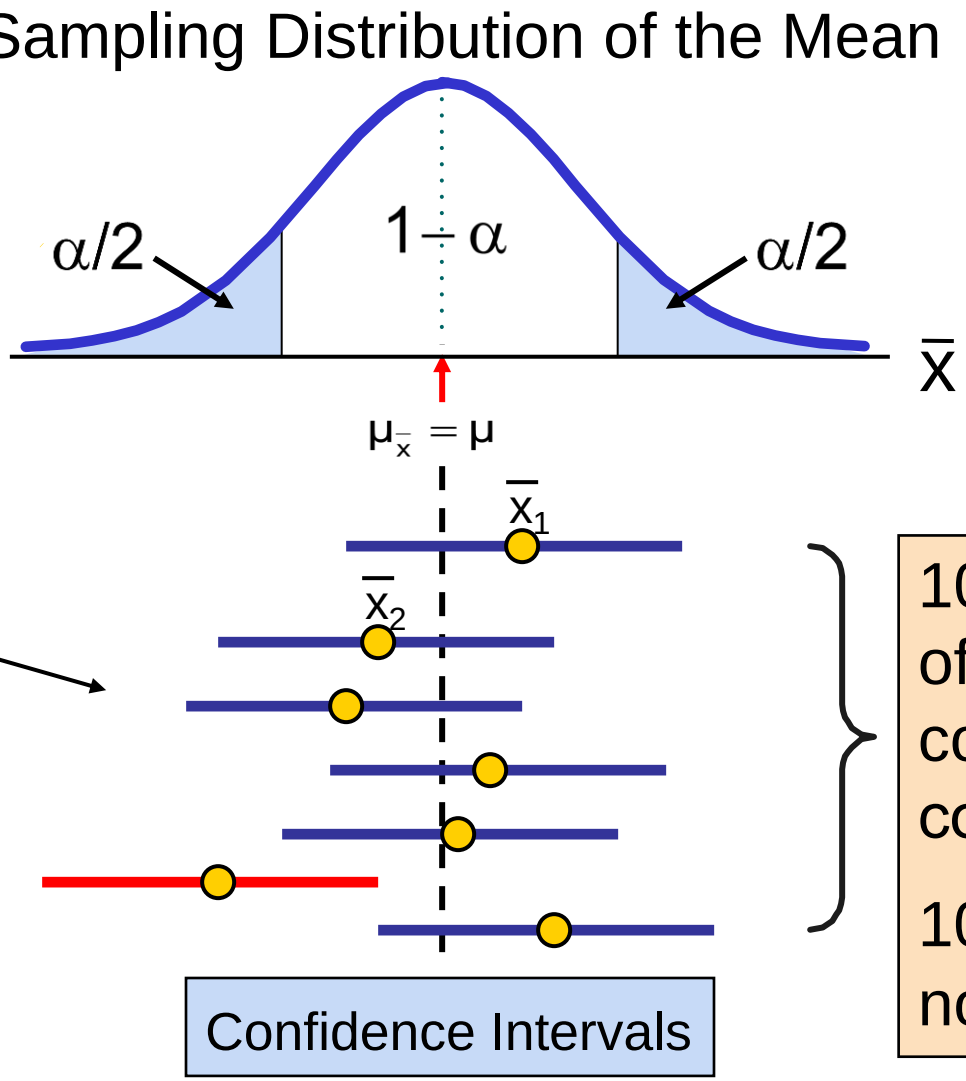

- Find $z_{.025} = \pm 1.96$ from the standard normal distribution table

# Common Levels of Confidence

- Commonly used confidence levels are 90%, 95%, and 99%

| *Confidence Level* | *Confidence Coefficient,* $1-\alpha$ | $Z_{\alpha/2}$ *value* |
|---|---|---|
| 80% | .80 | 1.28 |
| 90% | .90 | 1.645 |
| 95% | .95 | 1.96 |
| 98% | .98 | 2.33 |
| 99% | .99 | 2.58 |
| 99.8% | .998 | 3.08 |
| 99.9% | .999 | 3.27 |

# Intervals and Level of Confidence




Intervals extend from

$$\bar{x} - z\frac{\sigma}{\sqrt{n}}$$

to

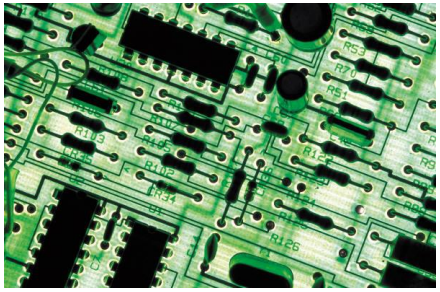
$$\bar{x} + z\frac{\sigma}{\sqrt{n}}$$

100(1-α)% of intervals constructed contain μ;

100(α)% do not.

# Example

- A sample of 11 circuits from a large normal population has a mean resistance of 2.20 ohms. We know from past testing that the population standard deviation is 0.35 ohms.
- Determine a 95% confidence interval for the true mean resistance of the population.



# Example

*(continued)*

- A sample of 11 circuits from a large normal population has a mean resistance of 2.20 ohms. We know from past testing that the population standard deviation is .35 ohms.

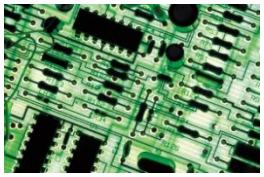
- Solution:

$$\bar{x} \pm z\frac{\sigma}{\sqrt{n}}$$

$$= 2.20 \pm 1.96\,(.35/\sqrt{11})$$

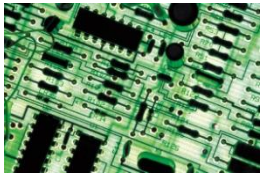
$$= 2.20 \pm .2068$$

$$1.9932 < \mu < 2.4068$$

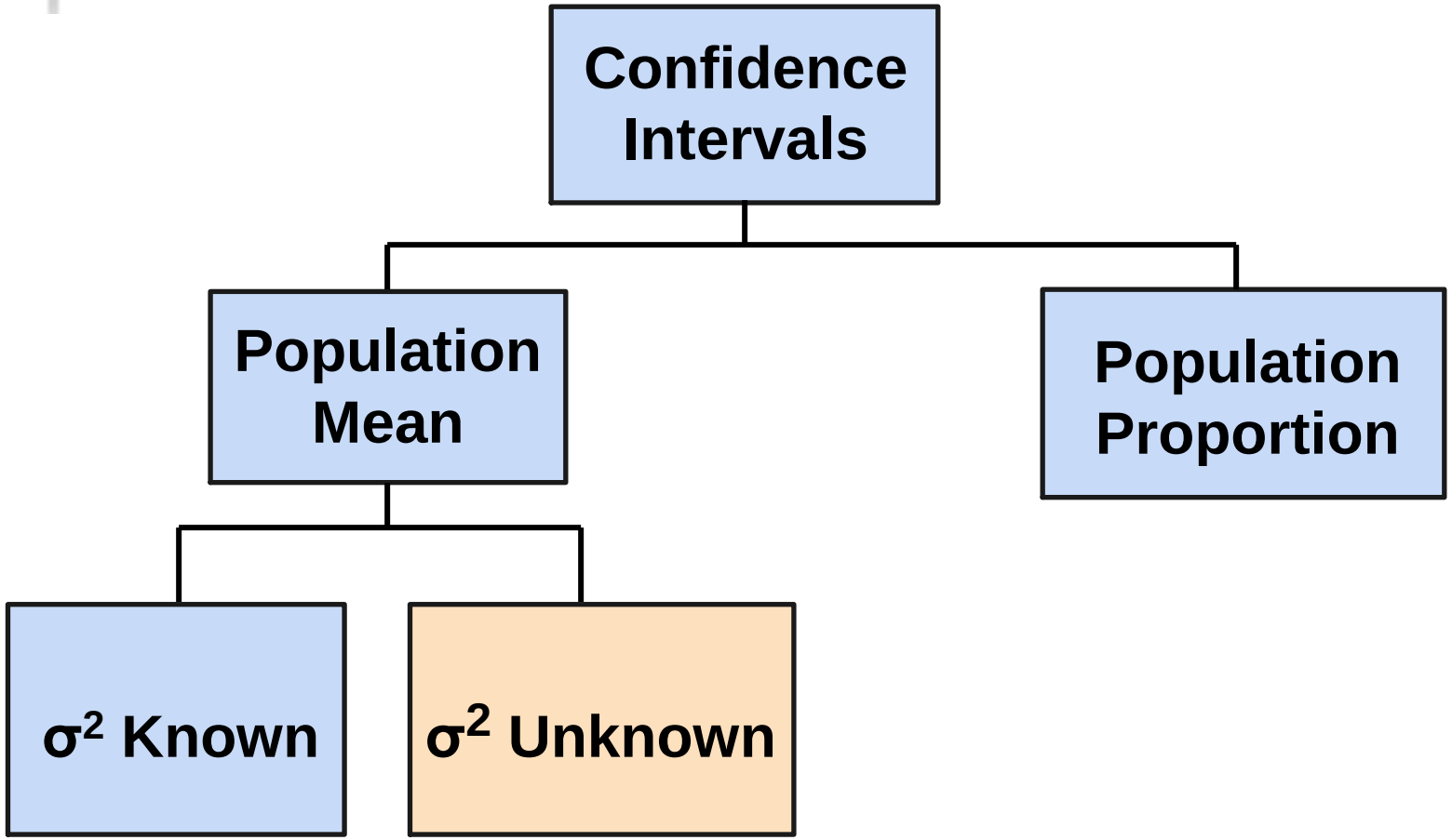


# Interpretation

- We are 95% confident that the true mean resistance is between 1.9932 and 2.4068 ohms
- Although the true mean may or may not be in this interval, 95% of intervals formed in this manner will contain the true mean



# Confidence Intervals



# Student's t Distribution

- Consider a random sample of n observations
  - with mean $\bar{x}$ and standard deviation s
  - from a normally distributed population with mean μ

- Then the variable

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

follows the Student's t distribution with (n - 1) degrees of freedom

# Confidence Interval for μ ($\sigma^2$ Unknown)

- If the population standard deviation σ is unknown, we can substitute the sample standard deviation, s
- This introduces extra uncertainty, since s is variable from sample to sample
- So we use the t distribution instead of the normal distribution

# Confidence Interval for μ (σ Unknown)

*(continued)*

- Assumptions
  - Population standard deviation is unknown
  - Population is normally distributed
  - If population is not normal, use large sample
- Use Student's t Distribution
- Confidence Interval Estimate:

$$\bar{x} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$$

where $t_{n-1,\alpha/2}$ is the critical value of the t distribution with n-1 d.f. and an area of α/2 in each tail:

$$P(t_{n-1} > t_{n-1,\alpha/2}) = \alpha/2$$

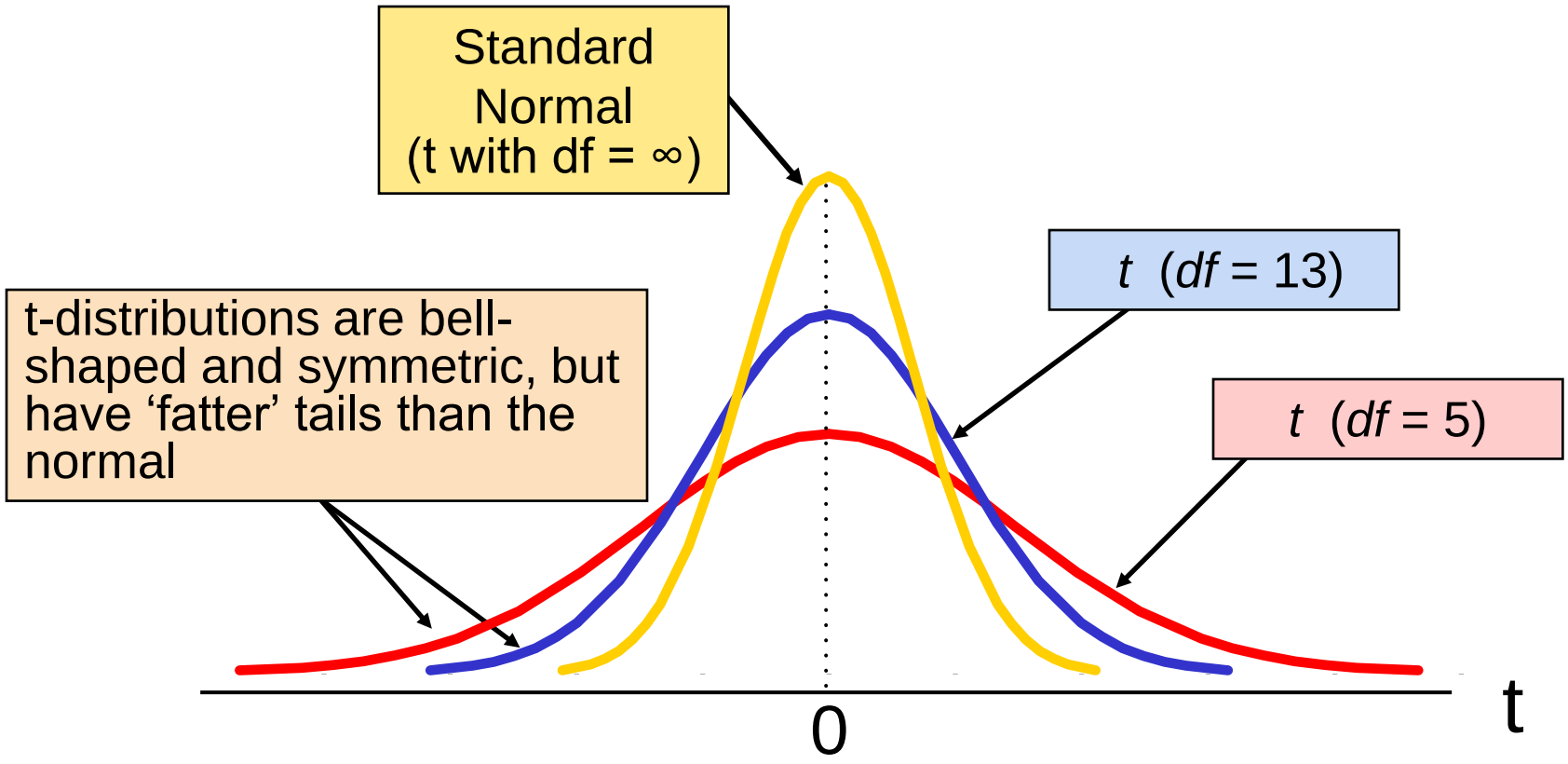
# Student's t Distribution

- The t is a family of distributions
- The t value depends on degrees of freedom (d.f.)
  - Number of observations that are free to vary after sample mean has been calculated

$$d.f. = n - 1$$

# Student's t Distribution

Note: t → Z as n increases

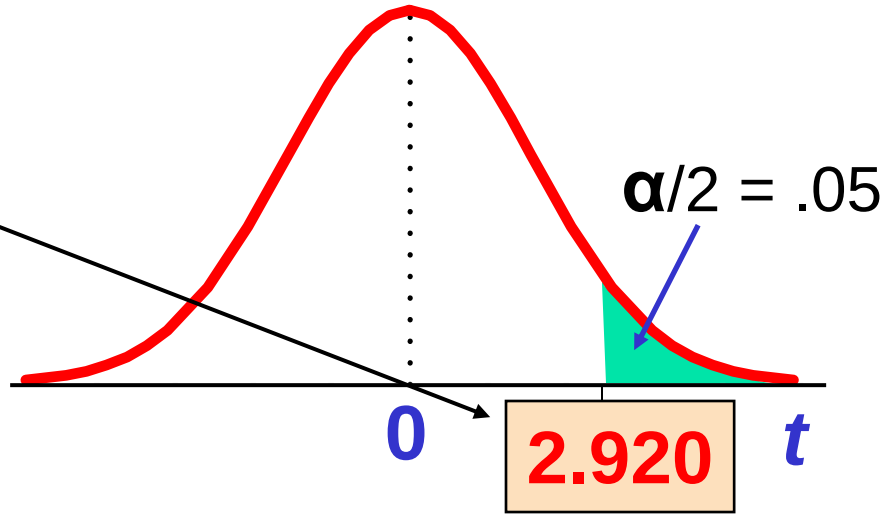


# Student's t Table

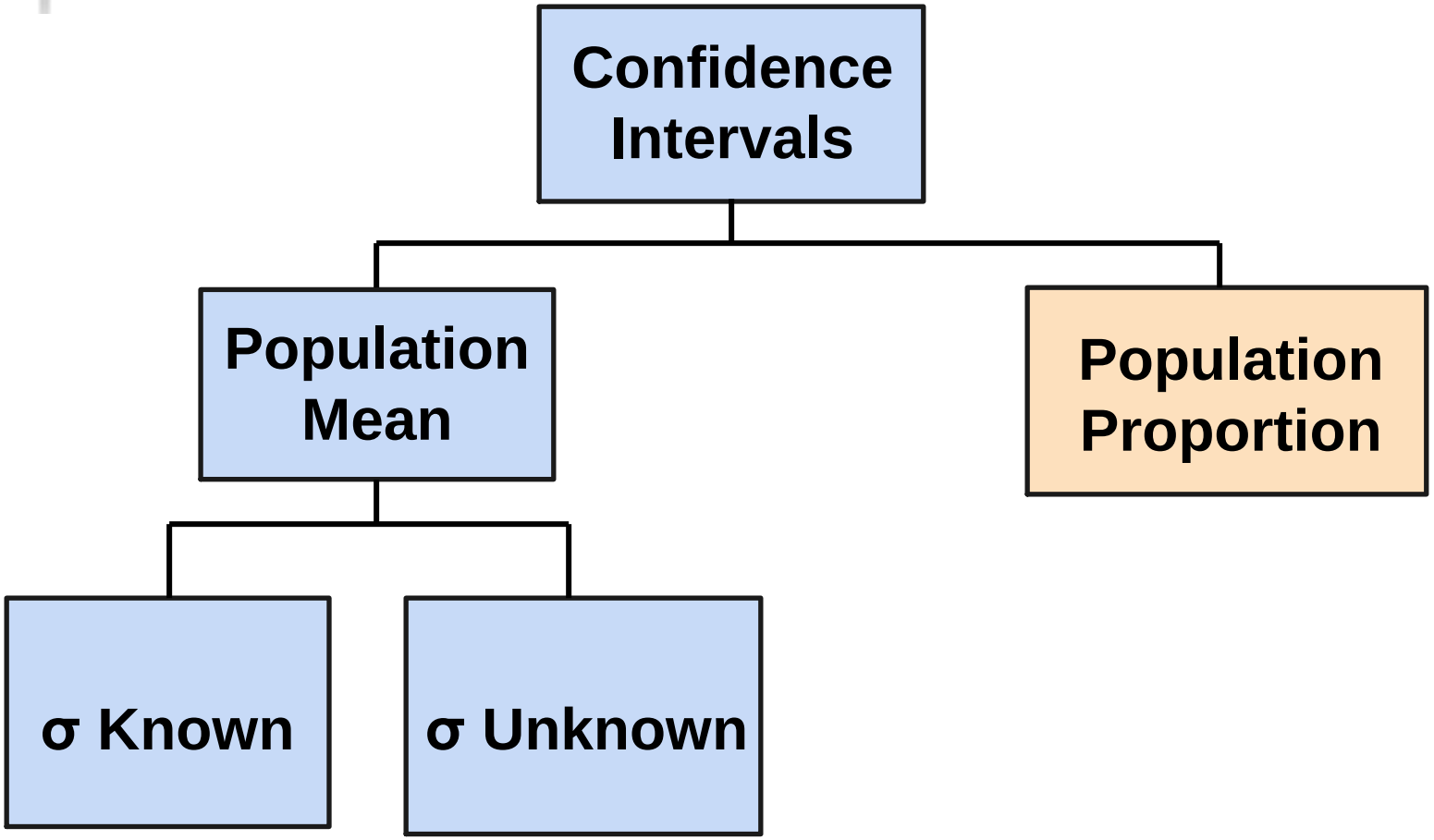
| | Upper Tail Area | | |
|---|---|---|---|
| **df** | .10 | **.05** | .025 |
| 1 | 3.078 | 6.314 | 12.706 |
| **2** | 1.886 | **2.920** | 4.303 |
| 3 | 1.638 | 2.353 | 3.182 |

The body of the table contains t values, not probabilities

Let: n = 3
df = *n* - 1 = 2
**α** = .10
**α**/2 =.05



# Confidence Intervals



# Confidence Intervals for the Population Proportion, p

- An interval estimate for the population proportion ( P ) can be calculated by adding an allowance for uncertainty to the sample proportion ( $\hat{p}$ )

# Confidence Intervals for the Population Proportion, p

*(continued)*

- Recall that the distribution of the sample proportion is approximately normal if the sample size is large, with standard deviation

$$\sigma_P = \sqrt{\frac{P(1-P)}{n}}$$

- We will estimate this with sample data:

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

# Confidence Interval Endpoints

- Upper and lower confidence limits for the population proportion are calculated with the formula

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < P < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- where
  - $z_{\alpha/2}$ is the standard normal value for the level of confidence desired
  - $\hat{p}$ is the sample proportion
  - n is the sample size

# Example

- A random sample of 100 people shows that 25 are left-handed.
- Form a 95% confidence interval for the true proportion of left-handers