Eser

350

EGEN350

Chapter 2

Probability

(Skip Sections 2.1 and 2.4 - 2.9)

Probability Definition and Rules

<u>Probability</u>: the numerical likelihood, or chance, that an outcome (<u>Ex</u>: x_1, x_2, etc.) will occur

i = index
1, 2, 3, ...

<u>Rules</u>:

never always

1. $0 \leq P(x_i) \leq 1$ where x_i is one possible outcome

2. Total probability of all possible outcomes = 1

3. <u>Additive</u>: $P(x_1 \ or \ x_2) = P(x_1) + P(x_2)$

4. <u>Complement</u>: $P(x_1') = 1 - P(x_1)$ b/c $\cancel{P(x_1)} + P(x_1') = 1$

"opposite of"

$\cancel{-P(x_1)}$ $-P(x_1)$

$P(x_1') = 1 - P(x_1)$

Example: Digital Channel

There is a chance that a bit transmitted through a digital transmission channel is received in error. Let X equal the number of bits in error in the next four bits transmitted. The possible values for X are {0, $\underbrace{1, 2, 3, 4}_{X \geq 1}$}. Given the following:

complement of $X \geq 1$

X	0	1	2	3	4
$P(X = x)$	0.6561	0.2916	0.0486	0.0036	0.0001

- Do all x satisfy: $\underbrace{0 \leq P(x_i) \leq 1}_{\text{yes!}}$ and $\underbrace{\text{Total probability} = 1}_{\text{Yes}}$?

- $P(X = 0 \; or \; X = 1) =$ $P(X=0) + P(X=1)$
  $0.6561 + 0.2916 = \boxed{0.9477}$

- $P(\underbrace{X \; is \; at \; least \; 1}_{X \geq 1}) =$ $P(X=1) + \cdots + P(X=4)$
  $= 1 - P(X=0) = 1 - 0.6561 = \boxed{0.3439}$

> iClicker Question:

Consider the variable X = number of courses a randomly selected university student is registered for.

X	1	2	3	4	5	6	7
$P(X = x)$	0.01	0.03	0.13	0.25	0.39	0.17	0.02

$$P(X = 3\ or\ X = 4\ or\ X = 5) =$$ add $= 0.77$

> iClicker Question:

Consider the variable X = number of courses a randomly selected university student is registered for.

$X \le 6$

X	1	2	3	4	5	6	7
$P(X = x)$	0.01	0.03	0.13	0.25	0.39	0.17	0.02

$P(X \text{ is at most } 6) = P(X \le 6)$

$= 1 - P(X = 7)$

$= 1 - .02$

$= \boxed{.98}$

$P(X \le 5) = 1 - [P(X = 6) + P(X = 7)]$

Counting Techniques

1. Multiplication Rule

2. Combination

3. Permutation

Multiplication Rule

Defn: A particular order in which events/things follow each other.

- If an operation can be described as a sequence of k steps and…
  - The number of ways of completing step 1 is n_1 and
  - The number of ways of completing step 2 is n_2 and
  - The number of ways of completing step 3 is n_3

  and so on…, then the

Total number of ways of completing the operation is:

$$n_1 \times n_2 \times n_3 \times \cdots \times n_k$$

Think: Tree Diagram

Examples: Passwords and Contractors

- An 8-character password consisting of all lower-case letters.

  How many passwords are possible?

$$\underset{1^{st}}{\underline{26}} \times \underset{2^{nd}}{\underline{26}} \times \underset{3^{rd}}{\underline{26}} \times \underset{4^{th}}{\underline{26}} \times \underset{5^{th}}{\underline{26}} \times \underset{6^{th}}{\underline{26}} \times \underset{7^{th}}{\underline{26}} \times \underset{8^{th}}{\underline{26}} = \boxed{26^8}$$

use blanks for each step

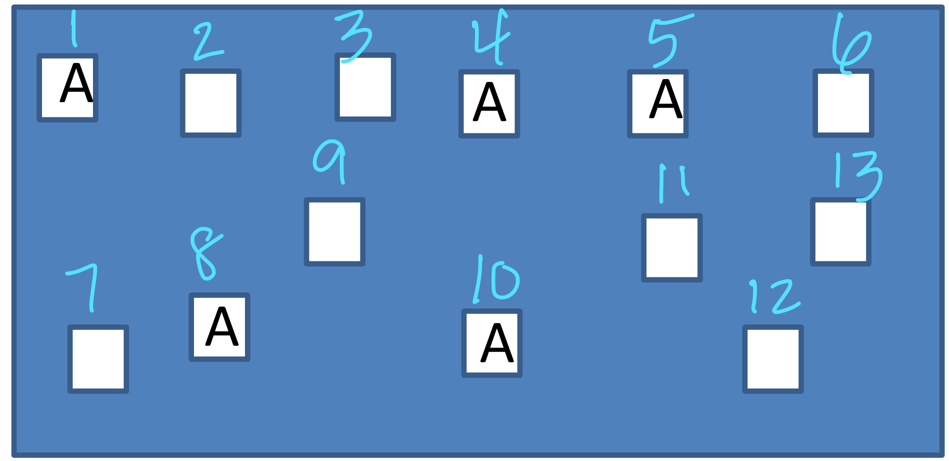
- A homeowner doing renovations has to choose from 5 drywallers, 3 plumbers, and 2 electrical contractors.

  How many ways can the contractors be chosen for the job?

$$\underset{\text{drywall}}{\underline{5}} \times \underset{\text{plumb.}}{\underline{3}} \times \underset{\text{elec.}}{\underline{2}} = \boxed{30}$$

> iClicker Question:

A printed circuit board has 13 different locations for a component to be placed. Five (5) IDENTICAL components (call them A, A, A, A, A) are placed on the board in the locations shown. →



Would reordering the components (A, A, A, A, A) within the same positions make a different circuit board?

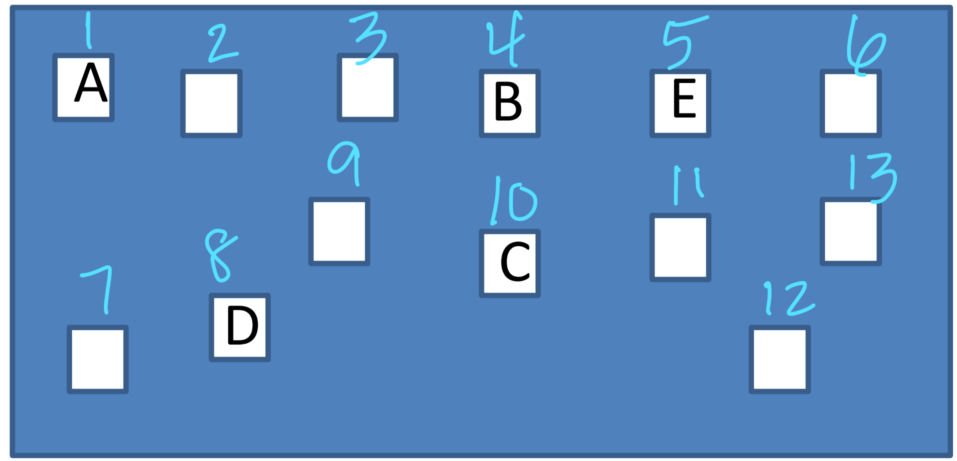
A. Yes

B. No

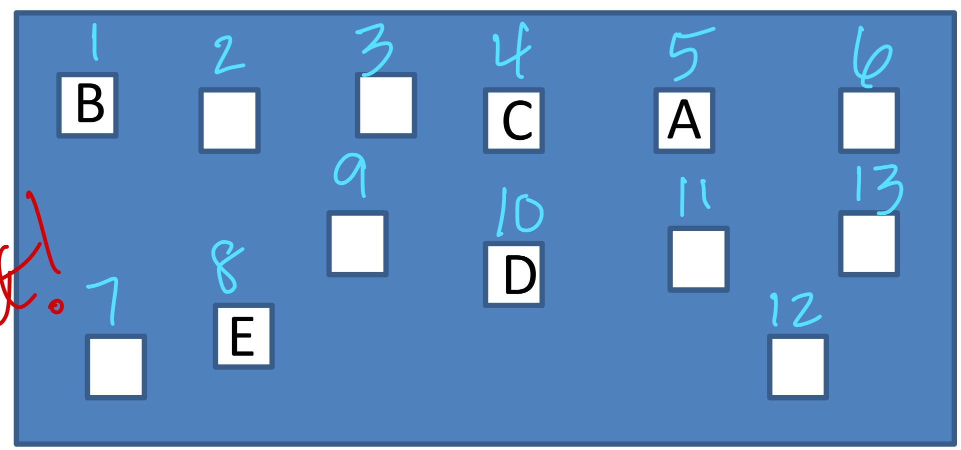
Order does not matter when components are identical!

> iClicker Question:

What if we had 5 DIFFERENT components (call them A, B, C, D, E) are placed on the board? One possible arrangement is shown here. →



Would reordering the components (A, B, C, D, E) make a different circuit board? (See second board.)



A. Yes

B. No

Order matters when components are different!

Combination

The number of combinations, subsets of size r that can be selected from a set of size n, where $r \leq n$, is:

said "n choose r"

$$C_r^n = \binom{n}{r} = \frac{n!}{r!(n-r)!}$$

Divides out re-orderings!

Example: In the circuit board problem using 5 IDENTICAL components, we are choosing a subset of $r = 5$ out of $n = 13$ choices. This is a **combination**. (*Order does not matter*!)

Calculator math > Prob nCr

$$C_5^{13} = \binom{13}{5} = \frac{13!}{5!(13-5)!} = 1287$$

= 13 nCr 5

RECALL:

Factorial notation: n! = n*(n-1)*(n-2)*(n-3)*….*1

Permutation

The number of permutations of subsets of r elements selected from a set of size n different elements, where $r \leq n$, is:

to permute means to re-order

$$P_r^n = \frac{n!}{(n-r)!}$$

Example: In the circuit board problem using 5 DIFFERENT components, this is a **permutation** of size $r = 5$ chosen from $n = 13$ choices. (*Order matters!*)

Mult Rule:

$\underline{13} \times \underline{12} \times \underline{11} \times \underline{10} \times \underline{9}$
A B C D E
is the same as P_5^{13}

$$P_5^{13} = \frac{13!}{(13-5)!} = 154{,}440$$

$= 13 \, nPr \, 5$

$r!$ = # of ways to re-order r things

mult. by $5! = 120$

Note: $P_r^n > C_r^n$ In our example: 154,440 > 1,287

Example: Fiber composites

(a) How many ways can a Materials Scientist select 3 fiber types to add to a 3-fiber type-composite from 25 fiber types?

$$C_3^{25} = \binom{25}{3} = \frac{25!}{3!(25-3)!} = 2300$$

(b) Now suppose that each fiber type has a different role: primary reinforcing phase, secondary phase, and tertiary phase. How many ways can 3 different fibers be selected?

$$P_3^{25} = \frac{25!}{(25-3)!} = \frac{25 \cdot 24 \cdot 23 \cdot \cancel{22 \cdot \ldots \cdot 1}}{\cancel{22 \cdot \ldots \cdot 1}} = 13,800$$

reduce by 1 each step

or

Mult. Rule : $\underset{\text{prim.}}{\underline{25}} \times \underset{\text{sec.}}{\underline{24}} \times \underset{\text{tert.}}{\underline{23}} = 13,800$

Does Order Matter?
(Are there distinct roles / positions / items?)

Yes

A letter can be used more than once

With Replacement

Mult. Rule

Ex: Passwords

Same fiber cannot hold multiple positions

Without Replacement

Permutation P^{25}_{3}

Ex: Select Prim/Sec/Tertiary phase fibers from 25 fibers

Mult. Rule

$\underline{25} \times \underline{24} \times \underline{23}$

Prim. Sec. Tert.

No

Without Replacement

Combination

Ex: Select 3 fiber types from 25

> iClicker Question:

In a study of tensile strength, there are 4 technicians who are trained to operate 3 machines. How many different ways can a technician and a machine be chosen?

What counting technique would you use to answer the question above?

A. Multiplication Rule
B. Combination
C. Permutation

> iClicker Question:

How many different ways can one make a first, second, third, and fourth choice among 12 firms leasing construction equipment?

What counting technique would you use to answer the question above?

A. Multiplication Rule
B. Combination
C. Permutation

> iClicker Question:

A calibration study needs to be conducted to see if the readings on 15 test machines are giving similar results. In how many ways can 3 of the 15 be selected for the initial investigation?

What counting technique would you use to answer the question above?

A. Multiplication Rule
B. Combination
C. Permutation

Summary

- Probability: the numerical likelihood, or chance, that an outcome will occur

- Rules of Probability:
  1. $0 \leq \mathrm{P}(x_i) \leq 1$ where x_i is one possible outcome
  2. Total probability of all possible outcomes = 1
  3. Additive: P(x_1 or x_2) = P(x_1) + P(x_2)
  4. Complement: $P(x_1') = 1 - P(x_1)$

- Counting Techniques:
  1. Multiplication Rule
  2. Combinations (order does not matter)
  3. Permutations (order matters)

EGEN350

Chapter 3

Overview of
Discrete Random Variables

(Skip Sections 3.3 - 3.4 and 3.6 - 3.8)

Terminology

- Data

  - Observations used to answer research questions

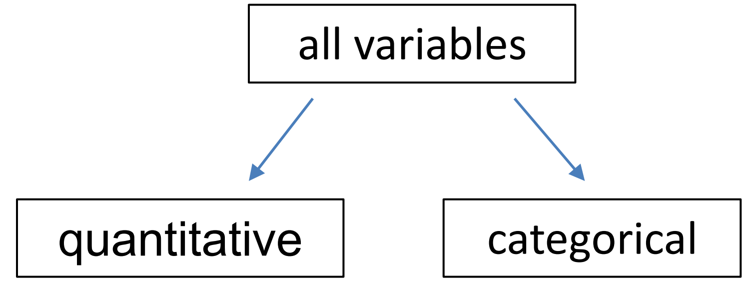
- Observational unit or Case or individual

  - The people or things we collect data on

- Variable or Random Variable

  - What is measured on each case

Types of Variables



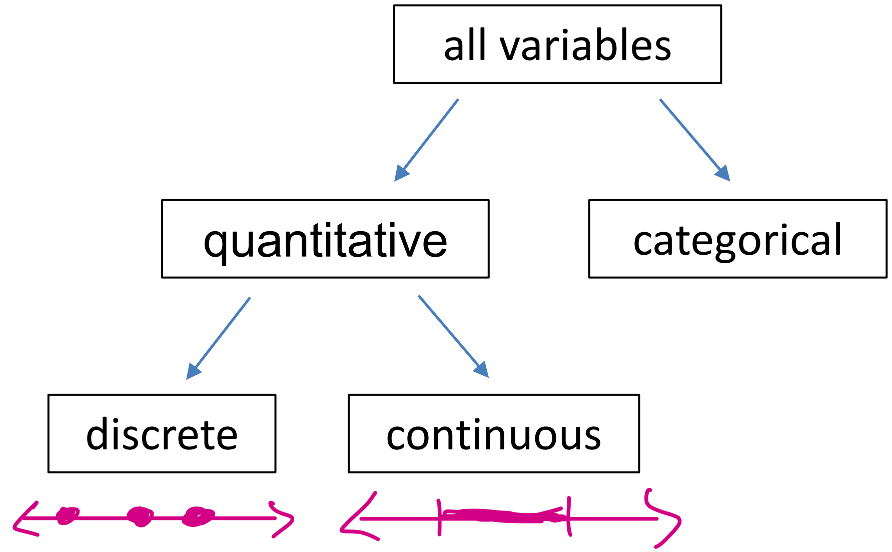
- Quantitative Variable

  where the focus of this class will be

  - Takes on numbers where arithmetic operations (such as finding an average) make sense
  - Examples: time, weight, strength

- Categorical Variable
  - Puts cases into groups (or categories)
  - Examples: adhesive type, eye color, dog breed

Types of Quantitative Variables



- Discrete
  - Can only be specific values, with jumps between
  - Examples: number of pets, number of courses

- Continuous
  - Value can be any number in a range (interval)
  - Examples: time, weight, strength

X versus x ?

What is the difference between X and x?

Ex:

X: a variable

X: age

$x = 46$

x: a possible value of the variable

Discrete Random Variables have a Probability Mass Function (pmf)

For a discrete random variable X with possible values $x_1, x_2, \ldots x_n$, a **probability mass function** is a function such that:

(1) $f(x_i) = P(X = x_i)$ (The value $f(x_i)$ is the probability of x_i occurring.)

(2) $f(x_i) \geq 0$ (Probability is greater than or equal to 0.)

(3) $\sum_{i=1}^{n} f(x_i) = 1$ (Total probability must equal 1.)

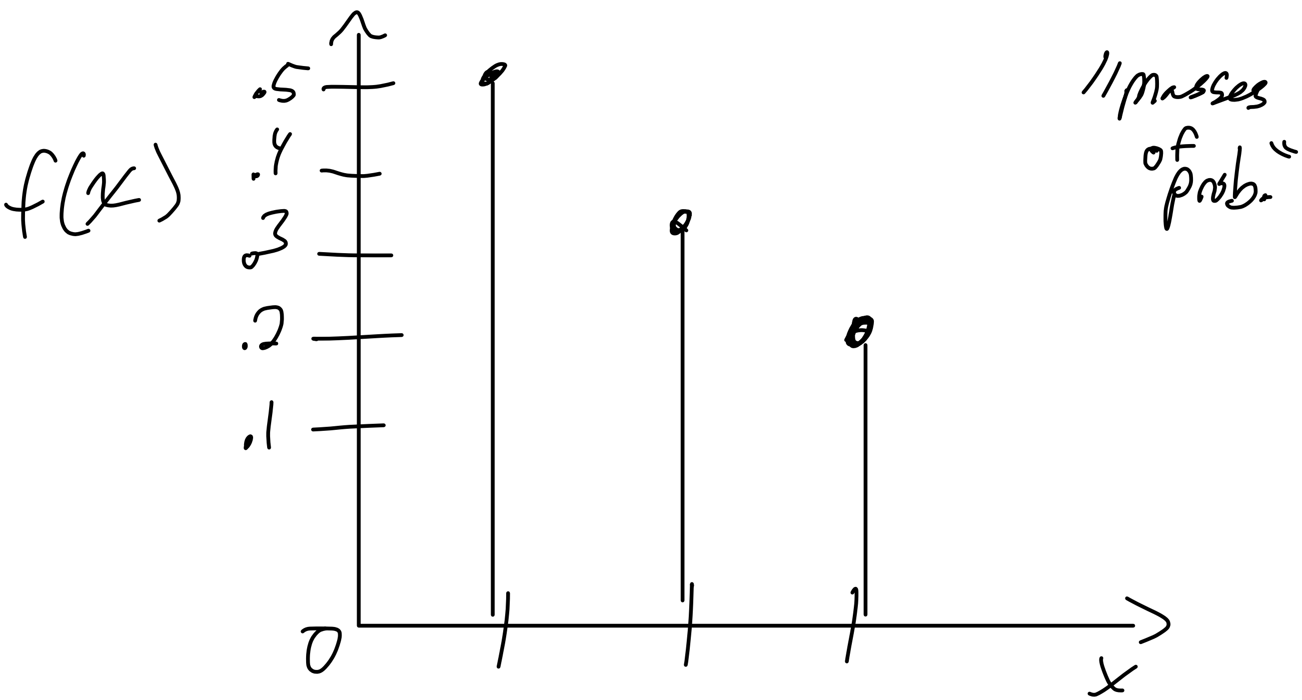
Example: Graph of pmf

Lots of components are ready to be shipped by a supplier. The number of defective components in each lot is:

Number of defectives	0	1	2
Probability	0.5	0.3	0.2

x's

$f(x) = P(X = x)$



Example: Find Probability using pmf

Lots of components are ready to be shipped by a supplier. The number of defective components in each lot is:

Number of defectives	0	1	2
Probability	0.5	0.3	0.2

$P(X \le 1) =$ $P(X=0) + P(X=1)$

$.5 + .3 = .8$

$P(X < 1) =$ $P(X=0) = .5$

* Pay attention to the equal sign for discrete probabilities!

Cumulative Distribution Function (CDF)

The **cumulative distribution function** of a discrete random variable X, denoted as $F(x)$, satisfy the following:

accumulates prob. up to xi

(1) $F(x_i) = P(X \leq x_i) = \sum_{x \leq x_i} f(x_i)$

(2) $0 \leq F(x) \leq 1$

(3) If $\mathrm{x} \leq y$, then $F(x) \leq F(y)$. (Nondecreasing function)

Important:

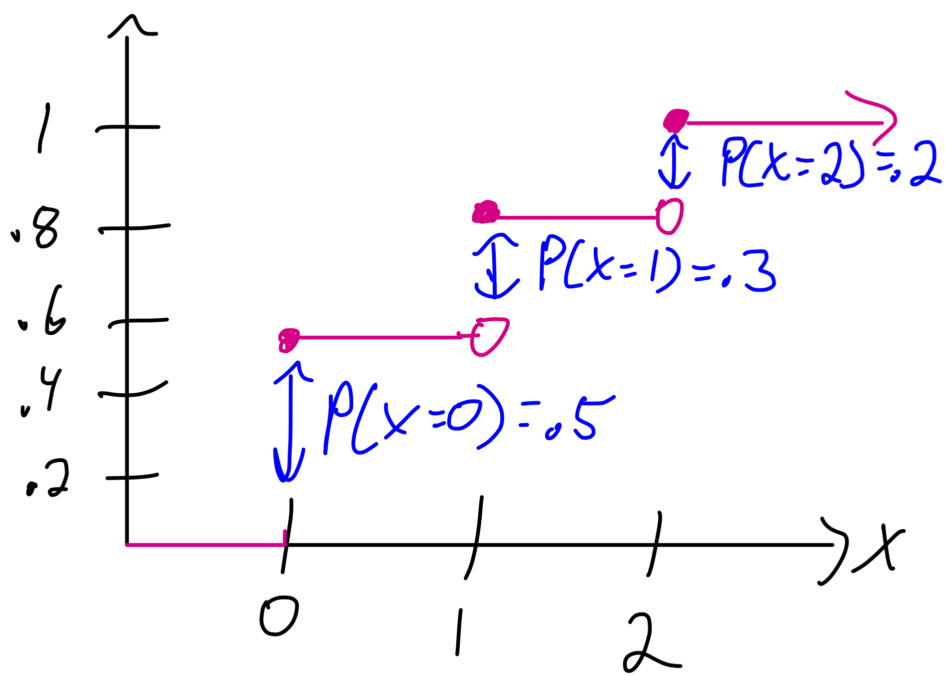
- The CDF is capital $F(x)$. The pmf is lower-case $f(x)$.
- CDF is the sum of all probs *up to* and *including* prob at x, $P(X \leq x)$.

accumulate prob. from left to right

Example: Graph of CDF

Lots of components are ready to be shipped by a supplier. The number of defective components in each lot is:

Number of defectives	0	1	2
Probability	0.5	0.3	0.2



Example: Find Probability using CDF

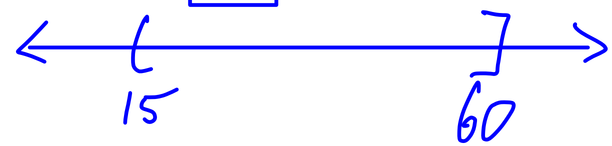
$$F(x) = \begin{cases} 0 & x < 0 \\ 0.4 & 0 \le x < 20 \\ 0.9 & 20 \le x < 100 \\ 1 & x \ge 100 \end{cases}$$

$F(x) = P(X \le x)$

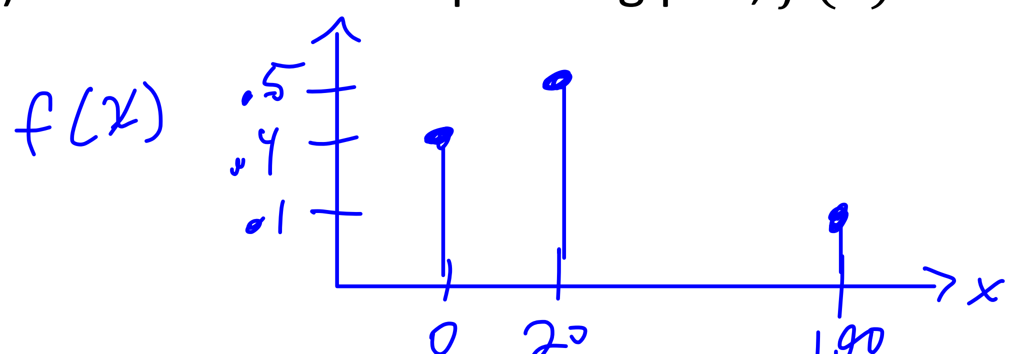
a) Find $P(X \le 15) = F(15) = .4$

b) Find $P(X > 40) = 1 - P(X \le 40) = 1 - .9 = \boxed{.1}$

c) Find $P(15 < X \le 60) = F(60) - F(15) = .9 - .4 = \boxed{.5}$



d) Sketch the corresponding pmf, $f(x)$:



Additional Discrete Distributions

- Binomial (covered here in Ch3 Notes)
- Discrete Uniform
- Geometric
- Negative Binomial
- Hypergeometric
- Poisson

Extra Credit

Binomial Distribution, $\underset{\text{Variable}}{X}\overset{\text{"is distributed as"}}{\sim}BIN(n, p)$

X = number of successes in n trials, where:

- n = number of independent $\boxed{\text{trials}}$
- Each trial has 2 possible outcomes (success/failure).
- p = probability of $\boxed{\text{success}}$

Probability formula:

$$P(\underset{\text{variable}}{X} = \underset{\text{value}}{x}) = \underbrace{\binom{n}{x}}_{\text{\# of ways to choose } x \text{ successes out of } n \text{ trials}} \underbrace{(p)^x}_{x \text{ successes}} \underbrace{(1-p)^{n-x}}_{n-x \text{ failures}}$$

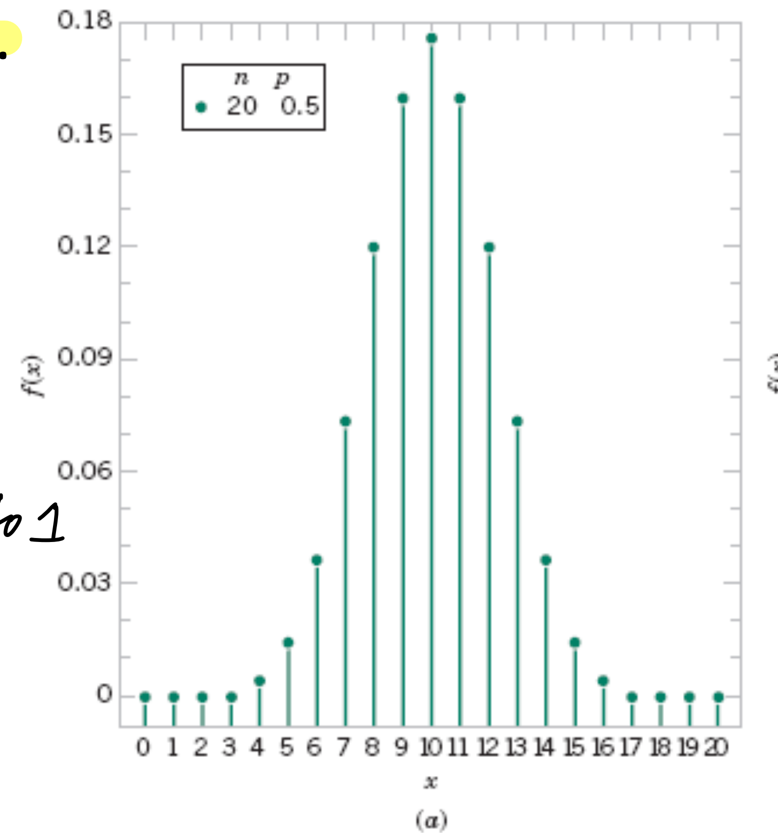
Note:

- The "SUCCESS" is the **outcome of interest**, which need not be a 'good' outcome. For example, a success could be a transmission error, defect, etc.
- When trials are "independent", you can multiply probabilities. For "x successes" p should be multiplied by itself x times. Hence, $(p)^x$.

Example: Binomial Distribution

Toss a fair coin 20 times.

- The number of trials (tosses) is $n = 20$.
- The number of successes (X) is the number of heads observed.
- The probability of success is $p = 0.5$.



What is the probability of observing 12 heads out of 20 tosses?

exponents add to n

$$P(X = 12) = \binom{20}{12}(.5)^{12}(.5)^{8} = \boxed{.12}$$

combinations

bases add to 1

12 heads

8 tails

Example (cont.):

Possible # of heads:
0, 1, 2, 3, ..., 20

Toss a fair coin 20 times.

- The number of trials (tosses) is $n = 20$.
- The number of successes (X) is the number of heads observed.
- The probability of success is $p = 0.5$.

What is the probability of observing at least 1 head out of 20 tosses?

$P(X \geq 1) = 1 - P(X=0)$

$= 1 - \left[\binom{20}{0}(.5)^0(.5)^{20}\right] = .9999$

> iClicker Question:

A lot contains ten thousand components, 10% of which are defective. Seven (7) components are sampled from the lot. Let X represent the number of defective components in the sample. What is the "success"?

A. The component is defective.

B. The component is non-defective.

*success is what the variable is counting

> iClicker Question:

A lot contains ten thousand components, 10% of which are defective. Seven (7) components are sampled from the lot. Let X represent the number of defective components in the sample. What is the distribution of X?

A. $X \sim BIN(10000, 0.1)$

B. $X \sim BIN(7, 0.1)$

C. $X \sim BIN(10000, 0.9)$

D. $X \sim BIN(7, 0.9)$

$n = 7$

$p = .1$

> iClicker Question:

A lot contains ten thousand components, of which 10% are defective. Seven (7) components are sampled from the lot. Let X represent the number of defective components in the sample. What is the probability of obtaining 1 defective in the sample of 7 components?

A. $P(X = 1) = \binom{7}{1}(0.1)^1(0.7)^6$

B. $P(X = 1) = \binom{7}{1}(0.1)^1(0.9)^6$

C. $P(X = 1) = \binom{10000}{1}(0.1)^1(0.9)^6$

> iClicker Question:

A lot contains ten thousand components, of which 10% are defective. Seven (7) components are sampled from the lot. Let X represent the number of defective components in the sample. What is the probability of obtaining "at least 1" defective in the sample of 7 components?

A. $P(X \geq 1) = \binom{7}{1}(0.1)^1(0.9)^6$

B. $P(X \geq 1) = 1 - P(X = 1) = 1 - \binom{7}{1}(0.1)^1(0.9)^6$

C. $\underbrace{P(X \geq 1) = 1 - P(X = 0)}_{\text{complement rule}} = 1 - \binom{7}{0}(0.1)^0(0.9)^7$

Summary

- Data are observations used to answer research questions.
- Observational unit or Case are the people or things we collect data on.
- Variable is what is measured on each individual/unit/case.
- There are two types of variables:

  (1) Quantitative - takes on numbers where arithmetic operations (such as finding an average) make sense, and

  (2) Categorical - puts cases into groups (or categories)
- There are two types of quantitative variables:

  (1) Discrete - can only be specific values, with jumps between

  (2) Continuous - value can be any number in a range (interval)

Summary

- What is the difference between X and x?

  X: a variable

  x: a possible value of the variable

- For a discrete random variable X with possible values $x_1, x_2, \dots x_n$, a <u>probability mass function</u>, denoted as $f(x)$, is a function such that:

  (1) $f(x_i) = P(X = x_i)$

  (2) $f(x_i) \geq 0$

  (3) $\sum_{i=1}^{n} f(x_i) = 1$

- The <u>cumulative distribution function</u> of a discrete random variable X, denoted as $F(x)$, satisfy the following:

  (1) $F(x) = P(X \leq x_i) = \sum_{x \leq x_i} f(x_i)$

  (2) $0 \leq F(x) \leq 1$

  (3) If $\mathrm{x} \leq y$, then $F(x) \leq F(y)$.

Summary

- Binomial Distribution

  X = number of successes in n trials, where:
  - n = number of independent trials
  - Each trial has 2 possible outcomes (success/failure).
  - p = probability of success
  - Distribution Notation: $X \sim BIN(n, p)$
  - Probability Formula: $P(X = x) = \binom{n}{x}(p)^x(1-p)^{n-x}$

- Other discrete distributions:
  - Discrete Uniform
  - Geometric
  - Negative Binomial
  - Hypergeometric
  - Poisson

Ch3

pmf

$f(x) = P(X = x)$

Ex:

x	0	1	2
$f(x)$	0.5	0.3	0.2

CDF

$F(x) = P(X \le x)$

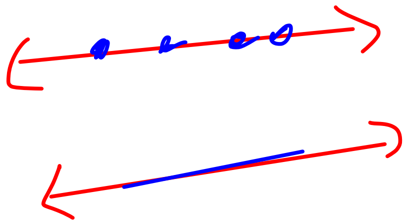
$$F(x) = \begin{cases} 0 & , \quad x < 0 \\ .5 & , \quad 0 \le x < 1 \\ .8 & , \quad 1 \le x < 2 \\ 1 & , \quad x \ge 2 \end{cases}$$

Find prob using pmf:

- $P(X \le 1) = 0.8$
- $P(X > 2) = 0$
- $P(1 < X \le 2) = 0.2$

Find prob using CDF:

- $P(X \le 1) = F(1)$
- $P(X > 2) = 1 - P(X \le 2)$
  $= 1 - F(2)$
- $P(1 < X \le 2) =$



discrete

Continuous

EGEN350

Chapter 4
Continuous Random Variables (RVs)
(Skip Sections 4.6 – 4.11)

> iClicker Question:

You will often see a probability statement such as:

$$P(X \ \leq x)$$

What is the difference between X and x?

A. X is a variable;
x is a possible value of the variable.

B. x is a variable;
X is a possible value of the variable.

Continuous Random Variables have a Probability Density Function (pdf)

For a continuous random variable X with a range of possible values, a **probability density function** is a function such that:

(1) $f(x) \geq 0$ (Height of the curve is on or above the x axis.)

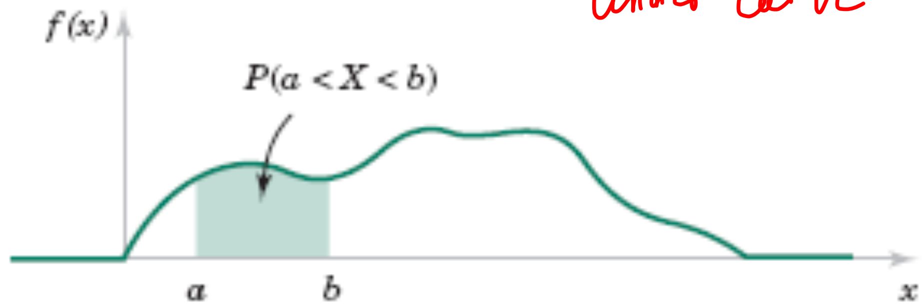
(2) $\int_{-\infty}^{\infty} f(x)dx = 1$ (Total area under the curve is 1.)

How to Find Probability using a pdf

- Probability is found by integrating under the curve over a range of values (a, b).

Prob = area under curve

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



$= F(b) - F(a) \leftarrow$ fundamental thm of calc.

- Recall from calculus: $P(X = a) = \int_a^a f(x)dx = \underline{\ 0\ }$

- Since the area of a line is 0, then: $P(X \le a) = P(X < a)$

No prob. at single value

Note

Example: Find Probability using pdf

X is a continuous RV whose pdf is: $\mathrm{f}(x) = \begin{Bmatrix} 0 & x < 0 \\ 2e^{-2x} & x \geq 0 \end{Bmatrix}$

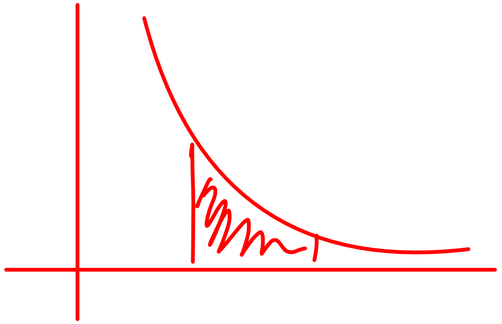
What is $P(1 < X < 3)$?

$= \int_1^3 2e^{-2x}\,dx$

$= \left.\frac{2e^{-2x}}{-2}\right|_1^3$

$= \left[-e^{-2(3)}\right] - \left[-e^{-2(1)}\right]$

$= -e^{-6} + e^{-2} = .1329$



> iClicker Question:

X is a continuous RV whose pdf is : $\mathrm{f}(x) = \begin{Bmatrix} 0 & x < 0 \\ 2e^{-2x} & x \geq 0 \end{Bmatrix}$

What is $P(X > 3)$? (Use 4 decimal places.)

$$\int_3^\infty 2e^{-2x}\,dx = -e^{-2x}\Big|_3^\infty = 0$$

Cumulative Distribution Function (CDF)

adds up all prop $\leq x$

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)\, du$$

<u>Note</u>: Since x is a limit of integration, change the variable of integration to u.

$$\int f(x)\, dx = F(x)$$

(Find integral)

$f(x)$

(pdf)

$F(x)$

(CDF)

$$\frac{d}{dx} F(x) = f(x)$$

(Find derivative)

Example: Find the CDF

X is a continuous RV whose pdf is: $\mathrm{f}(x) = \begin{cases} 0 & x < 0 \\ 2e^{-2x} & x \geq 0 \end{cases}$

Find the CDF.

$F(x) = \int_{-\infty}^{x} f(u)\,du = \underbrace{\int_{-\infty}^{0} 0\,du}_{\text{can ignore}} \to 0 + \int_{0}^{x} 2e^{-2u}\,du$

Interesting Piece

$= \frac{2e^{-2u}}{-2}\Big|_0^x$

$\left[-e^{-2u}\right]\Big| -e^{-2x}+1$

$\boxed{1-e^{-2x}}$

FULL ANSWER

$$\boxed{F(x) = \begin{cases} 0 & x < 0 \\ 1-e^{-2x} & x \geq 0 \end{cases}}$$

Example: Find Probability using CDF

$$F(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-2x} & x \geq 0 \end{cases}$$

Found by: $F(x) = \int_{-\infty}^{x} 2e^{-2u}du$

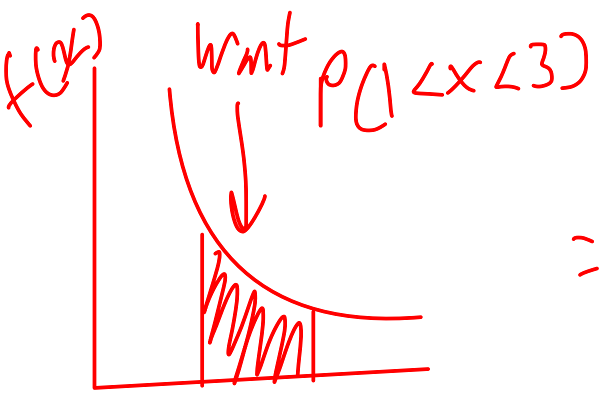
What is $P(1 < X < 3)$? $= P(X<3) - P(X<1)$

$P(X \le x) = F(x)$

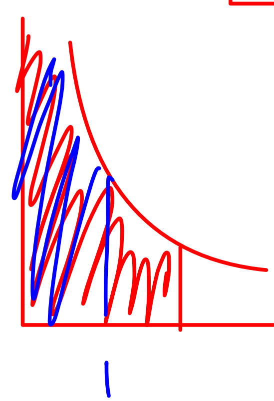
$= F(3) - F(1)$

$= \left[1-e^{-2(3)}\right] - \left[1-e^{-2(1)}\right]$

$= \boxed{.1329}$



=

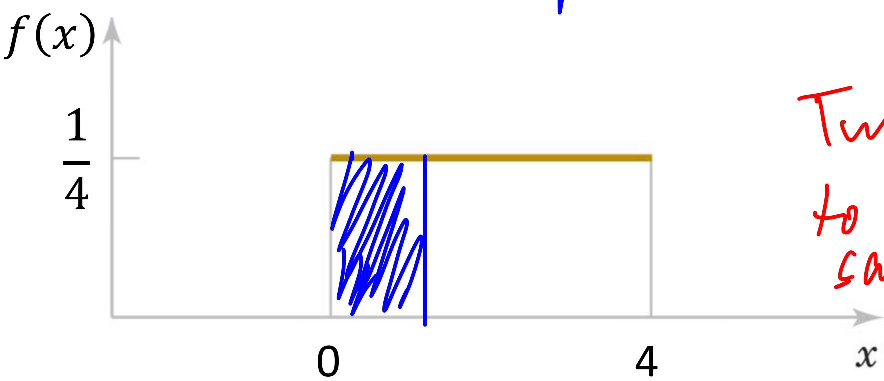


* integrate PDF to find prob.

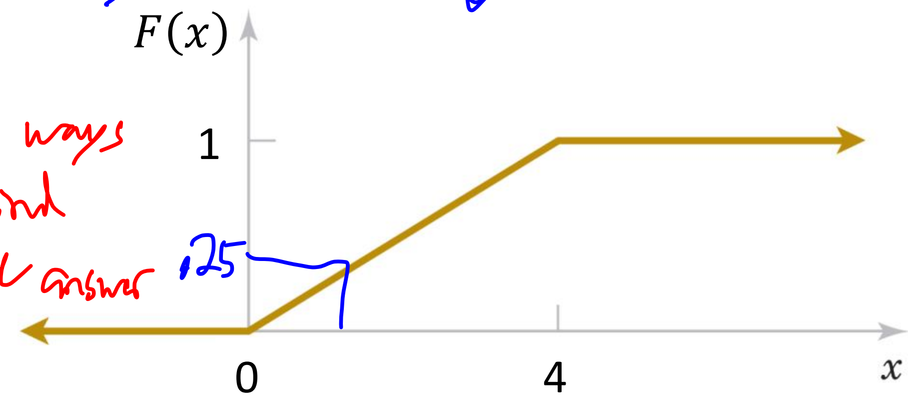
Relating pdf and CDF

evaluate CDF to find prob.

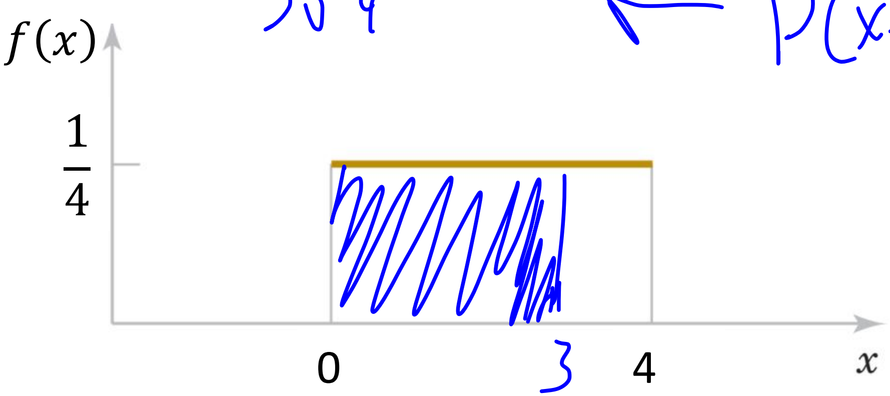
$\int_0^1 \frac{1}{4}\,dx = .25$ $\quad P(x \le 1)$ $\quad F(1) = .25$

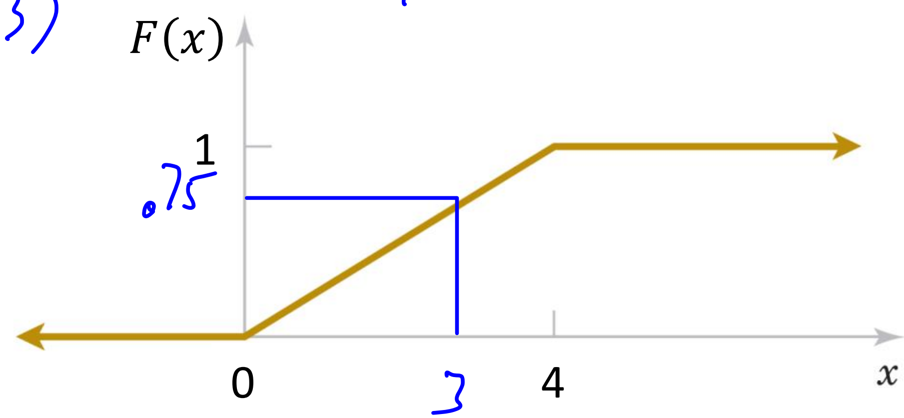


Two ways to find same answer



$\int_0^3 \frac{1}{4}\,dx = .75$ $\quad P(x \le 3)$ $\quad F(3) = .75$





> iClicker Question:

The CDF of X is: $F(x) = \begin{Bmatrix} 0 & x < 0 \\ 1 - e^{-2x} & x \geq 0 \end{Bmatrix}$

Using the CDF, find $P(X > 3)$?

A. $F(3) = 1 - e^{-2(3)}$

B. $F(3) - F(1) = (1 - e^{-2(3)}) - (1 - e^{-2(1)})$

C. $1 - F(3) = 1 - (1 - e^{-2(3)})$

> iClicker Question:

X is a cont. RV with pdf:

$$f(x) = \begin{Bmatrix} \dfrac{20000}{(x+100)^3} & X > 0 \\ 0 & elsewhere \end{Bmatrix}$$

How many pieces are in the CDF (piece-wise function)?

A. 1
B. 2
C. 3

> iClicker Question:

X is a cont. RV with pdf: $\mathrm{f}(x) = \begin{Bmatrix} 2x & 0 \leq X \leq 1 \\ 0 & elsewhere \end{Bmatrix}$

How many pieces are in the CDF (piece-wise function)?

A. 1
B. 2
C. 3

Example: Find Percentiles* using the CDF

$$F(x) = \begin{cases} 0 & x < 0 \\ x^2 & 0 \leq x \leq 1 \\ 1 & x \geq 1 \end{cases}$$

Find the median (50th percentile) of X:

$F(x) = P(X \leq x) \leftarrow$ RECALL

$F(x) = x^2 = .5$

$x = \sqrt{.5}$

[$-\sqrt{.5}$ is impossible]

* Percentiles: $100p^{\text{th}}$ percentile is a value that has $100p\%$ of the observations *below* its value.

> iClicker Question:

$$F(x) = \begin{cases} 0 & x < -1 \\ \frac{x^3+1}{9} & -1 \le x < 2 \\ 1 & x \ge 2 \end{cases}$$

What x value is the 90th percentile? (Use 2 decimal places.)

$$\boxed{F(x) = \frac{x^3+1}{9} = .9}$$

$$x^3+1 = 8.1$$

$$x^3 = 7.1$$

$$\boxed{x = \sqrt[3]{7.1}}$$

Where you've been; Where you're going

- Now you:
  - Know how to find a probability using a pdf
  - Know how to find a CDF
  - Know how to find a probability using a CDF
  - Understand how pdf and CDF relate

- Up next, the Most Common Continuous Distribution:

Distribution	Shorthand Notation
Normal	$X \sim N(\mu, \sigma^2)$

Normal Distribution

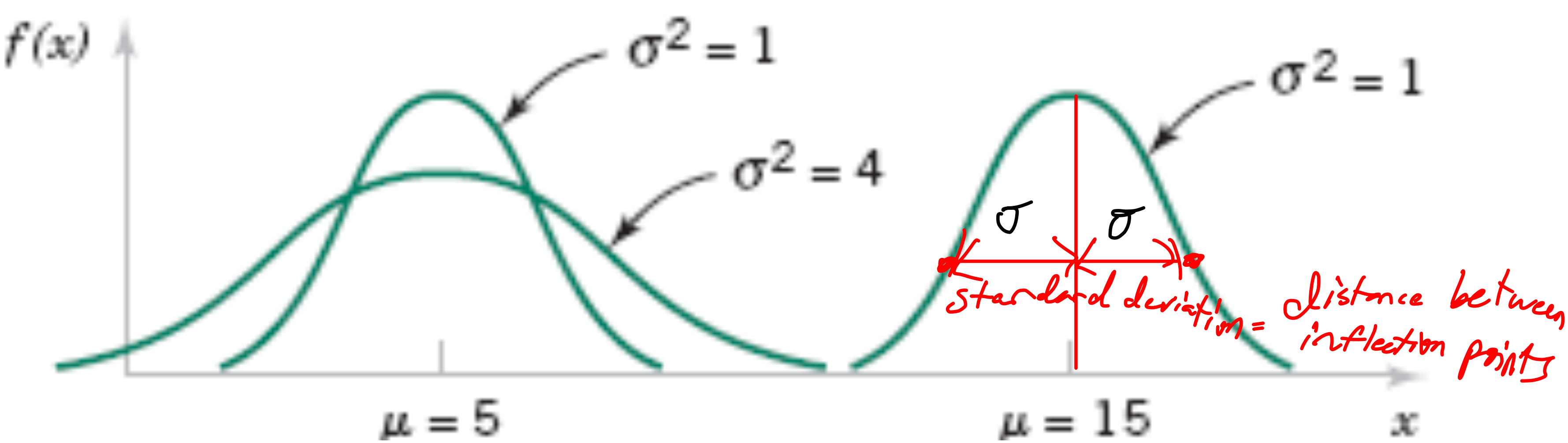
(AKA, Gaussian dist or Bell-shaped curve)

Mean Variance

$$X \sim N(\mu, \sigma^2)$$

- Mean – center of the distribution
- Variance – spread of the distribution
- Standard Deviation (SD) – spread of the dist

where $\sigma = SD = \sqrt{Variance}$



NOTE: The inflection points (where the concavity changes) are 1 SD away from the mean.

Standard Normal Distribution (Z)

- Note, there is *no closed form* answer to the integral of the Normal pdf.

- Thus, to find a probability under a normal pdf, we must:
  1) 'Standardize' the distribution (formula is on next 2 slides) &
  2) Find the probability in a table (or use a computer/calculator).

- Since there are infinitely many normal distributions (i.e. infinitely many combinations for mean/variance), we will 'standardize' all normal distributions to one normal distribution, called the "Standard Normal Distribution".

$$\boxed{Z \sim N(0, 1)}$$

> iClicker Question:

The Standardize Formula is: $Z = \frac{X-\mu}{\sigma}$

Which term in the Z formula is being used as the "ruler" to allow for standardized comparisons across applications?

The difference (x-μ
is measured in what?

A. Z

B. X

C. μ

D. σ

How to Standardize X to Z

If X is a normal random variable with mean μ and variance σ^2, the random variable Z is a normal random variable with mean 0 and variance 1.

$X \sim N(\mu, \sigma^2)$ $\longrightarrow$

$$\boxed{Z = \frac{X - \mu}{\sigma}}$$

$\longrightarrow Z \sim N(0, 1)$

<u>NOTE</u>: X is the original variable with the original units.

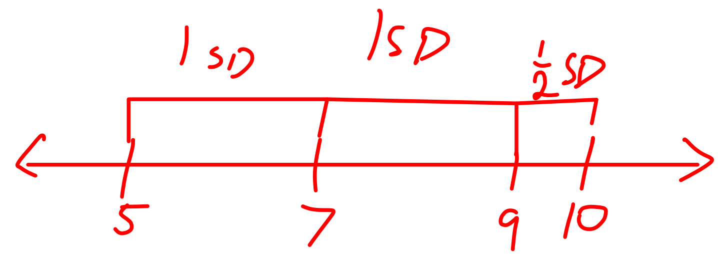
Z is a standardized variable with <u>no</u> units.

Z Interpretation

Z = the number of standard deviations X is above/below μ

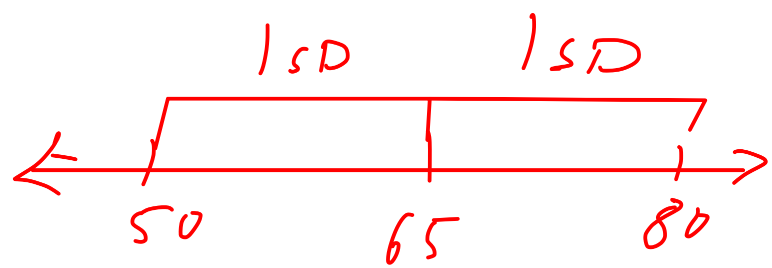
$Z = \frac{x-\mu}{\sigma}$

<u>Example</u>: $z = \frac{10-5}{2} = +2.5$



Interpretation: $x = 10$ is 2.5 SDs above $\mu = 5$

<u>Example</u>: $z = \frac{50-80}{15} = -2$



Interpretation: $x = 50$ is 2 SDs below $\mu = 80$

> iClicker Question:

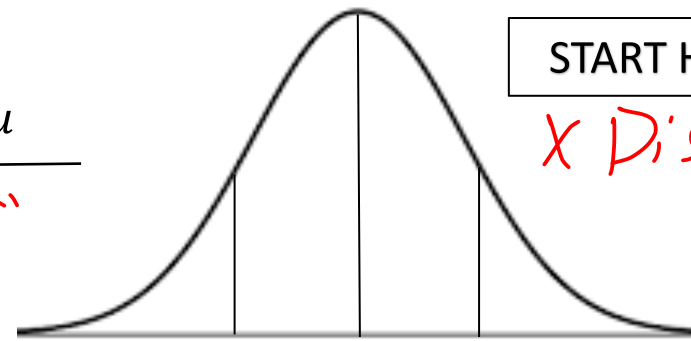
What is the interpretation of $\mathrm{z} = \frac{x-\mu}{\sigma} = -1.85$?

A. z is 1.85 units below x
B. z is 1.85 units above x
C. x is 1.85 standard deviations below μ
D. x is 1.85 standard deviations above μ

How Does Standardizing Work?

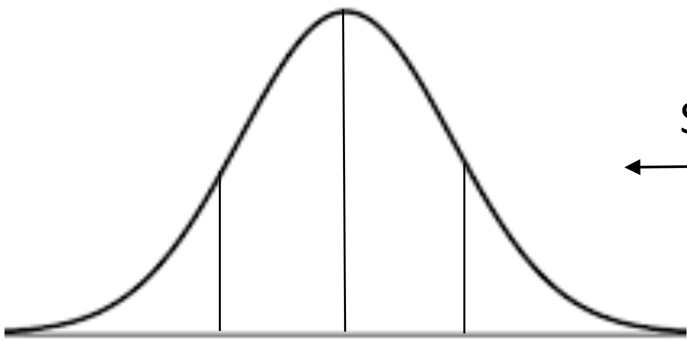
START HERE

X Dist



Subtract μ

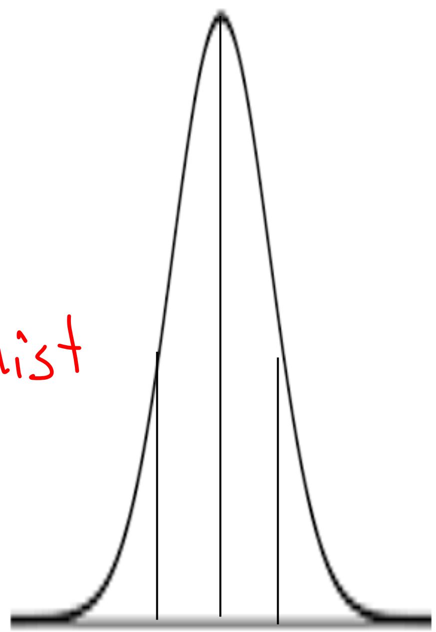
"shift"



Divide by σ

Rescale

Z dist



$$Z = \frac{X - \mu}{\sigma}$$

<u>IMPORTANT</u>:

The prob. under X curve

is the same as

the prob. under Z curve!

Z-table

$$\Phi(z) = P(Z \le z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}\,du$$

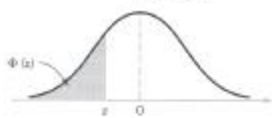
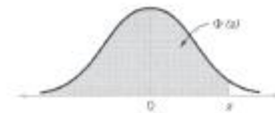


TABLE III Cumulative Standard Normal Distribution

z	−0.09	−0.08	−0.07	−0.06	−0.05	−0.04	−0.03	−0.02	−0.01	−0.00
−3.9	0.000033	0.000034	0.000036	0.000037	0.000039	0.000041	0.000042	0.000044	0.000046	0.000048
−3.8	0.000050	0.000052	0.000054	0.000057	0.000059	0.000062	0.000064	0.000067	0.000069	0.000072
−3.7	0.000075	0.000078	0.000082	0.000085	0.000088	0.000092	0.000096	0.000100	0.000104	0.000108
−3.6	0.000112	0.000117	0.000121	0.000126	0.000131	0.000136	0.000142	0.000147	0.000153	0.000159
−3.5	0.000165	0.000172	0.000179	0.000185	0.000193	0.000200	0.000208	0.000216	0.000224	0.000233
−3.4	0.000242	0.000251	0.000260	0.000270	0.000280	0.000291	0.000302	0.000313	0.000325	0.000337
−3.3	0.000350	0.000362	0.000376	0.000390	0.000404	0.000419	0.000434	0.000450	0.000467	0.000483
−3.2	0.000501	0.000519	0.000538	0.000557	0.000577	0.000598	0.000619	0.000641	0.000664	0.000687
−3.1	0.000711	0.000736	0.000762	0.000789	0.000816	0.000845	0.000874	0.000904	0.000935	0.000968
−3.0	0.001001	0.001035	0.001070	0.001107	0.001144	0.001183	0.001223	0.001264	0.001306	0.001350
−2.9	0.001395	0.001441	0.001489	0.001538	0.001589	0.001641	0.001695	0.001750	0.001807	0.001866
−2.8	0.001926	0.001988	0.002052	0.002118	0.002186	0.002256	0.002327	0.002401	0.002477	0.002555
−2.7	0.002635	0.002718	0.002803	0.002890	0.002980	0.003072	0.003167	0.003264	0.003364	0.003467
−2.6	0.003573	0.003681	0.003793	0.003907	0.004025	0.004145	0.004269	0.004396	0.004527	0.004661
−2.5	0.004799	0.004940	0.005085	0.005234	0.005386	0.005543	0.005703	0.005868	0.006037	0.006210
−2.4	0.006387	0.006569	0.006756	0.006947	0.007143	0.007344	0.007549	0.007760	0.007976	0.008198
−2.3	0.008424	0.008656	0.008894	0.009137	0.009387	0.009642	0.009903	0.010170	0.010444	0.010724
−2.2	0.011011	0.011304	0.011604	0.011911	0.012224	0.012545	0.012874	0.013209	0.013553	0.013903
−2.1	0.014262	0.014629	0.015003	0.015386	0.015778	0.016177	0.016586	0.017003	0.017429	0.017864
−2.0	0.018309	0.018763	0.019226	0.019699	0.020182	0.020675	0.021178	0.021692	0.022216	0.022750
−1.9	0.023295	0.023852	0.024419	0.024998	0.025588	0.026190	0.026803	0.027429	0.028067	0.028717
−1.8	0.029379	0.030054	0.030742	0.031443	0.032157	0.032884	0.033625	0.034379	0.035148	0.035930
−1.7	0.036727	0.037538	0.038364	0.039204	0.040059	0.040929	0.041815	0.042716	0.043633	0.044565
−1.6	0.045514	0.046479	0.047460	0.048457	0.049471	0.050503	0.051551	0.052616	0.053699	0.054799
−1.5	0.055917	0.057053	0.058208	0.059380	0.060571	0.061780	0.063008	0.064256	0.065522	0.066807
−1.4	0.068112	0.069437	0.070781	0.072145	0.073529	0.074934	0.076359	0.077804	0.079270	0.080757
−1.3	0.082264	0.083793	0.085343	0.086915	0.088508	0.090123	0.091759	0.093418	0.095098	0.096801
−1.2	0.098525	0.100273	0.102042	0.103835	0.105650	0.107488	0.109349	0.111233	0.113140	0.115070
−1.1	0.117023	0.119000	0.121001	0.123024	0.125072	0.127143	0.129238	0.131357	0.133500	0.135666
−1.0	0.137857	0.140071	0.142310	0.144572	0.146859	0.149170	0.151505	0.153864	0.156248	0.158655
−0.9	0.161087	0.163543	0.166023	0.168528	0.171056	0.173609	0.176185	0.178786	0.181411	0.184060
−0.8	0.186733	0.189430	0.192150	0.194894	0.197662	0.200454	0.203269	0.206108	0.208970	0.211855
−0.7	0.214764	0.217695	0.220650	0.223627	0.226627	0.229650	0.232695	0.235762	0.238852	0.241964
−0.6	0.245097	0.248252	0.251429	0.254627	0.257846	0.261086	0.264347	0.267629	0.270931	0.274253
−0.5	0.277595	0.280957	0.284339	0.287740	0.291160	0.294599	0.298056	0.301532	0.305026	0.308538
−0.4	0.312067	0.315614	0.319178	0.322758	0.326355	0.329969	0.333598	0.337243	0.340903	0.344578
−0.3	0.348268	0.351973	0.355691	0.359424	0.363169	0.366928	0.370700	0.374484	0.378281	0.382089
−0.2	0.385908	0.389739	0.393580	0.397432	0.401294	0.405165	0.409046	0.412936	0.416834	0.420740
−0.1	0.424655	0.428576	0.432505	0.436441	0.440382	0.444330	0.448283	0.452242	0.456205	0.460172
0.0	0.464144	0.468119	0.472097	0.476078	0.480061	0.484047	0.488033	0.492022	0.496011	0.500000

Left margin
z-value to 1st
decimal place

$$\Phi(z) = P(Z \le z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}\,du$$



Top margin
2nd decimal
of Z

TABLE III Cumulative Standard Normal Distribution (*continued*)

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500000	0.503989	0.507978	0.511967	0.515953	0.519939	0.532922	0.527903	0.531881	0.535856
0.1	0.539828	0.543795	0.547758	0.551717	0.555760	0.559618	0.563559	0.567495	0.571424	0.575345
0.2	0.579260	0.583166	0.587064	0.590954	0.594835	0.598706	0.602568	0.606420	0.610261	0.614092
0.3	0.617911	0.621719	0.625516	0.629300	0.633072	0.636831	0.640576	0.644309	0.648027	0.651732
0.4	0.655422	0.659097	0.662757	0.666402	0.670031	0.673645	0.677242	0.680822	0.684386	0.687933
0.5	0.691462	0.694974	0.698468	0.701944	0.705401	0.708840	0.712260	0.715661	0.719043	0.722405
0.6	0.725747	0.729069	0.732371	0.735653	0.738914	0.742154	0.745373	0.748571	0.751748	0.754903
0.7	0.758036	0.761148	0.764238	0.767305	0.770350	0.773373	0.776373	0.779350	0.782305	0.785236
0.8	0.788145	0.791030	0.793892	0.796731	0.799546	0.802338	0.805106	0.807850	0.810570	0.813267
0.9	0.815940	0.818589	0.821214	0.823815	0.826391	0.828944	0.831472	0.833977	0.836457	0.838913
1.0	0.841345	0.843752	0.846136	0.848495	0.850830	0.853141	0.855428	0.857690	0.859929	0.862143
1.1	0.864334	0.866500	0.868643	0.870762	0.872857	0.874928	0.876976	0.878999	0.881000	0.882977
1.2	0.884930	0.886860	0.888767	0.890651	0.892512	0.894350	0.896165	0.897958	0.899727	0.901475
1.3	0.903199	0.904902	0.906582	0.908241	0.909877	0.911492	0.913085	0.914657	0.916207	0.917736
1.4	0.919243	0.920730	0.922196	0.923641	0.925066	0.926471	0.927855	0.929219	0.930563	0.931888
1.5	0.933193	0.934478	0.935744	0.936992	0.938220	0.939429	0.940620	0.941792	0.942947	0.944083
1.6	0.945201	0.946301	0.947384	0.948449	0.949497	0.950529	0.951543	0.952540	0.953521	0.954486
1.7	0.955435	0.956367	0.957284	0.958185	0.959071	0.959941	0.960796	0.961636	0.962462	0.963273
1.8	0.964070	0.964852	0.965621	0.966375	0.967116	0.967843	0.968557	0.969258	0.969946	0.970621
1.9	0.971283	0.971933	0.972571	0.973197	0.973810	0.974412	0.975002	0.975581	0.976148	0.976705
2.0	0.977250	0.977784	0.978308	0.978822	0.979325	0.979818	0.980301	0.980774	0.981237	0.981691
2.1	0.982136	0.982571	0.982997	0.983414	0.983823	0.984222	0.984614	0.984997	0.985371	0.985738
2.2	0.986097	0.986447	0.986791	0.987126	0.987455	0.987776	0.988089	0.988396	0.988696	0.988989
2.3	0.989276	0.989556	0.989830	0.990097	0.990358	0.990613	0.990863	0.991106	0.991344	0.991576
2.4	0.991802	0.992024	0.992240	0.992451	0.992656	0.992857	0.993053	0.993244	0.993431	0.993613
2.5	0.993790	0.993963	0.994132	0.994297	0.994457	0.994614	0.994766	0.994915	0.995060	0.995201
2.6	0.995339	0.995473	0.995604	0.995731	0.995855	0.995975	0.996093	0.996207	0.996319	0.996427
2.7	0.996533	0.996636	0.996736	0.996833	0.996928	0.997020	0.997110	0.997197	0.997282	0.997365
2.8	0.997445	0.997523	0.997599	0.997673	0.997744	0.997814	0.997882	0.997948	0.998012	0.998074
2.9	0.998134	0.998193	0.998250	0.998305	0.998359	0.998411	0.998462	0.998511	0.998559	0.998605
3.0	0.998650	0.998694	0.998736	0.998777	0.998817	0.998856	0.998893	0.998930	0.998965	0.998999
3.1	0.999032	0.999065	0.999096	0.999126	0.999155	0.999184	0.999211	0.999238	0.999264	0.999289
3.2	0.999313	0.999336	0.999359	0.999381	0.999402	0.999423	0.999443	0.999462	0.999481	0.999499
3.3	0.999517	0.999533	0.999550	0.999566	0.999581	0.999596	0.999610	0.999624	0.999638	0.999650
3.4	0.999663	0.999675	0.999687	0.999698	0.999709	0.999720	0.999730	0.999740	0.999749	0.999758
3.5	0.999767	0.999776	0.999784	0.999792	0.999800	0.999807	0.999815	0.999821	0.999828	0.999835
3.6	0.999841	0.999847	0.999853	0.999858	0.999864	0.999869	0.999874	0.999879	0.999883	0.999888
3.7	0.999892	0.999896	0.999900	0.999904	0.999908	0.999912	0.999915	0.999918	0.999922	0.999925
3.8	0.999928	0.999931	0.999933	0.999936	0.999938	0.999941	0.999943	0.999946	0.999948	0.999950
3.9	0.999952	0.999954	0.999956	0.999958	0.999959	0.999961	0.999963	0.999964	0.999966	0.999967

Interior: $P(Z \le z)$

How to Read the Z-table?

- Find the Cumulative Standard Normal Table (Z-table) in Appendix
- Find the z value to the first-decimal place on the left margin
- Find the second-decimal place of z on the top margin
- Where the row and column intersect gives the prob: $P(Z \leq z)$

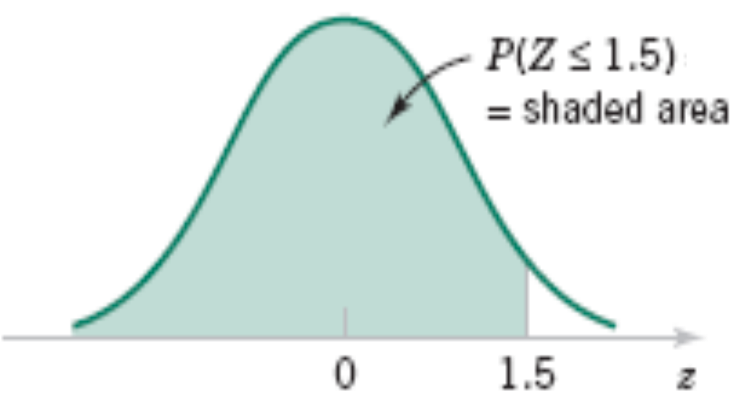
Examples:

$P(Z \leq 1.50) =$

$P(Z > 1.50) = 1 - P(Z \leq 1.5)$

$1 - .933193$

$\boxed{.066807}$



Cumulative Standard Normal Table

z	0.00	0.01	0.02	0.03
0	0.50000	0.50399	0.50398	0.51197
⋮		⋮		
1.5	0.93319	0.93448	0.93574	0.93699

Types of Normal Probability Questions

- Forward Probability (Given x, find probability.)

  Examples:

  - "Less Than" Prob, such as $P(X < 10) = ?$
  - "Greater Than" Prob, such as $P(X > 10) = ?$
  - "Interval" Prob, such as $P(5 < X < 10) = ?$

- Backward Probability (Given probability, find x.)

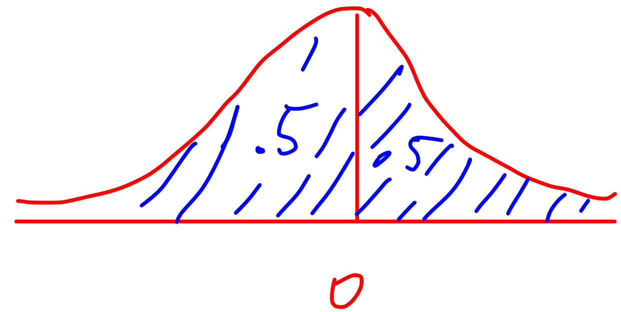
  Examples:

  - $P(X < x) = 0.05$, find $x = ?$
  - $P(X > x) = 0.05$, find $x = ?$

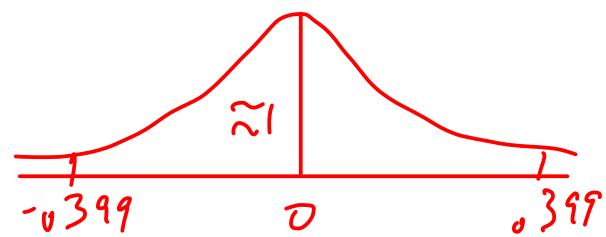
Note: Use Closest/Nearest Value in the Z table

(No need to interpolate!)

Facts about Z-table



- $P(Z < 0) \; = \; P(Z > 0) \; = \; 0.5$
- $P(-\infty < Z < \infty) = 1$
- $P(-3.99 < Z < 3.99) \approx 1$
- $P(Z < -3.99) = P(Z > 3.99) \approx 0$
- $P(Z > -3.99) = P(Z < 3.99) \approx 1$



- <u>Note</u>, the "less than" probabilities are given in the table.
  - ➢ Find $P(Z \; < \; -0.75) =$ .226627
    Where "-0.7" row and "-0.05" column intersect
  - ➢ Find $P(Z \; < 1.13) =$
    Where "1.1" row and "0.03" column intersect

Facts about Z-table: "Greater Than" Prob.

To find the "greater than" probability, you must either:

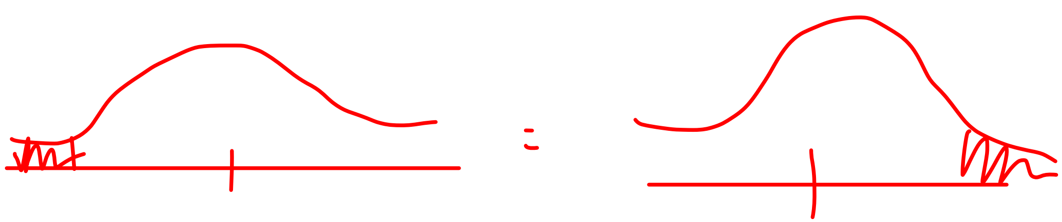
- *Use complement "1 −"*

  $P(Z > 1.13) = 1 - P(Z \leq 1.13) = 1 - .870762 = \boxed{.129238}$

OR

- *Use symmetry about 0*

  $P(Z > 1.13) = P(Z < -1.13) = .129238$



Example: “Greater Than” Prob.

Assume that the current measurement in a strip of wire (X) follows a normal distribution a mean of 10 mA and a variance of 4 mA². What is the probability that a measurement exceeds 13 mA?

$X \sim N(10, 4)$ $\sigma^2 = 4$

$\sigma = 2$

Steps: (1) Standardize to Z.

(2) Read the Z-table to obtain the probability.

$$P(X > 13) = P\left(\frac{X-10}{2} > \frac{13-10}{2}\right)$$

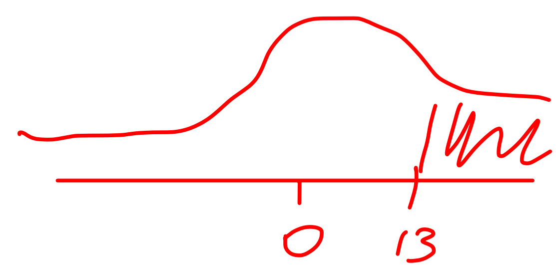
$$= P(Z > 1.5)$$

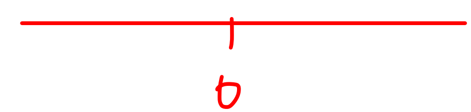
$$= 1 - P(Z \leq 1.5)$$

$$1 - .933193 = \boxed{.066807}$$

or

$$= P(Z < -1.50) = \boxed{.066807}$$



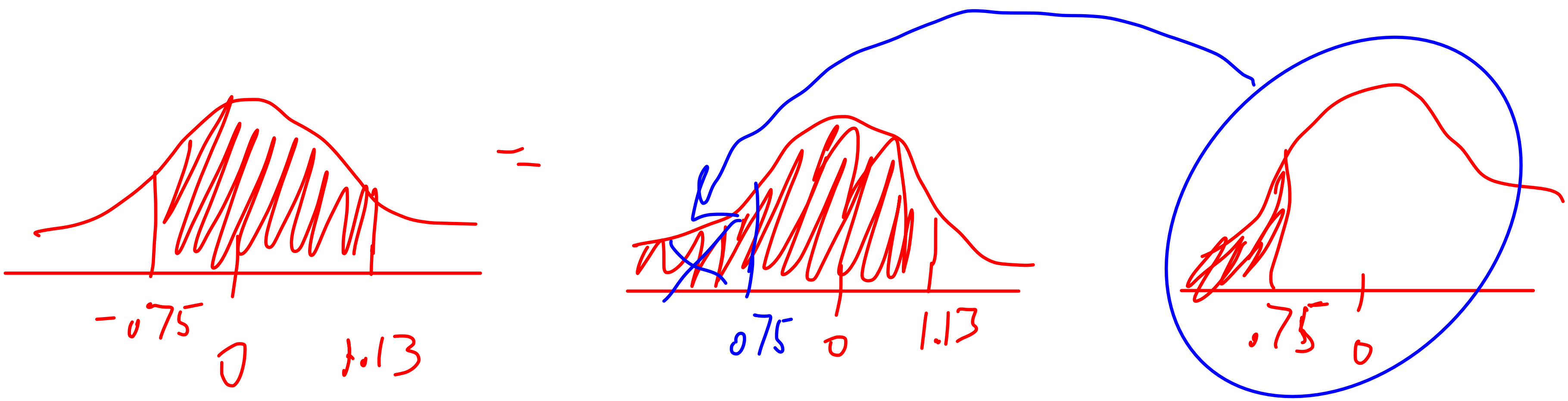


Facts about Z-table: “Interval” Prob.

To find the “interval” probability:

- Find “less than” probability of larger z, then subtract off the “less than” prob of smaller z

$$P(-0.75 < Z < 1.13) = P(Z < 1.13) - P(Z < -0.75) =$$



Example: “Interval” Prob.

What is the probability that a current measurement is between 9 and 11 mA? Recall: $X \sim N(10,4)$

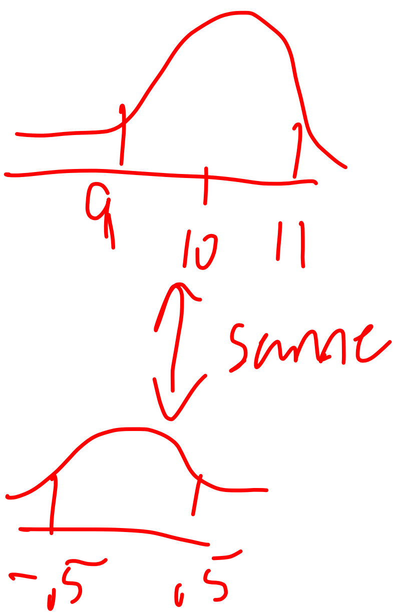
$$P(9 < X < 11) = P\left(\frac{9-10}{2} < Z < \frac{11-10}{2}\right)$$

$$= P(-.5 < Z < .5)$$

$$= P(Z < .5) - P(Z < -.5)$$

$$= .691462 - .308538$$

$$= \boxed{.382924}$$



Types of Normal Probability Questions

- Forward Probability (Given x, find probability.)

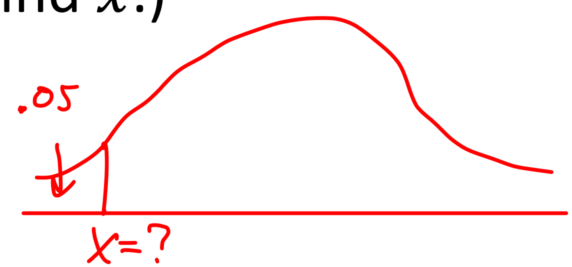
  Examples:

  - "Less Than" Prob, such as $P(X < 10) = ?$
  - "Greater Than" Prob, such as $P(X > 10) = ?$
  - "Interval" Prob, such as $P(5 < X < 10) = ?$

- Backward Probability (Given probability, find x.)

  Examples:

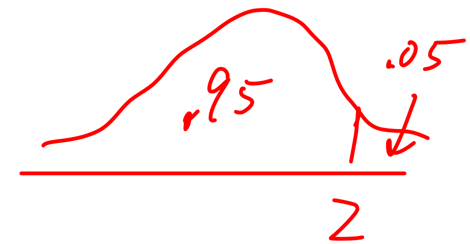
  - $P(X < x) = 0.05$, find $x = ?$
  - $P(X > x) = 0.05$, find $x = ?$



Note: Use Closest/Nearest Value in the Z table
(No need to interpolate!)

Facts about Z-table: Backward Prob.

Given the probability $P(Z > z) = 0.05$, find the z:



- The area under the Z curve to the right of z is 0.05.

$$P(Z > z) = 0.05$$

- Thus, the area under the Z curve to the left of z is 0.95.

$$P(Z < z) = 0.95$$

- Find z: 1.64 / 1.65 / 1.645

* Look up .95 on interior, Check row/column headings to find Z

Example: "Backward" Prob.

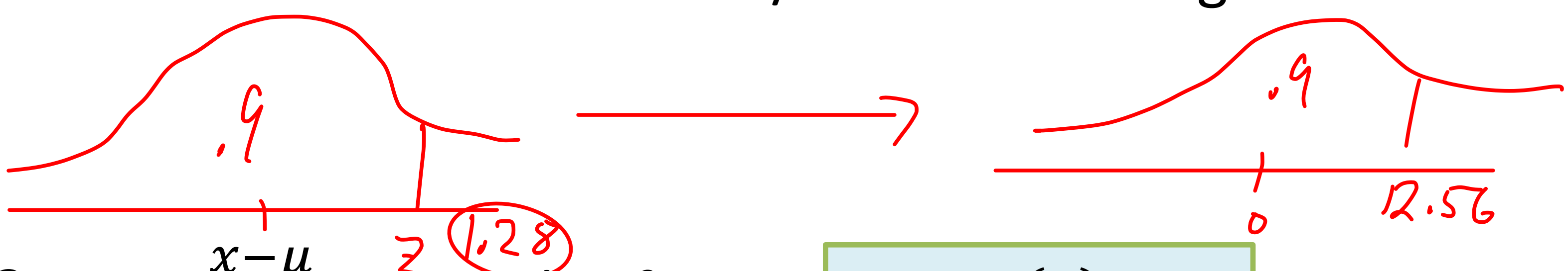
<u>Current Example</u>: $X \sim N(10,4)$

1. Convert the prob. involving X to a prob. involving Z

$P(X < x) = .9$

$P(Z < z) = .9$

2. Find the prob. on the interior of the Z-table, then find the z value in the row/column headings.



3. $z = \dfrac{x-\mu}{\sigma}$, now solve for x: $\boxed{x = \sigma(z) + \mu}$

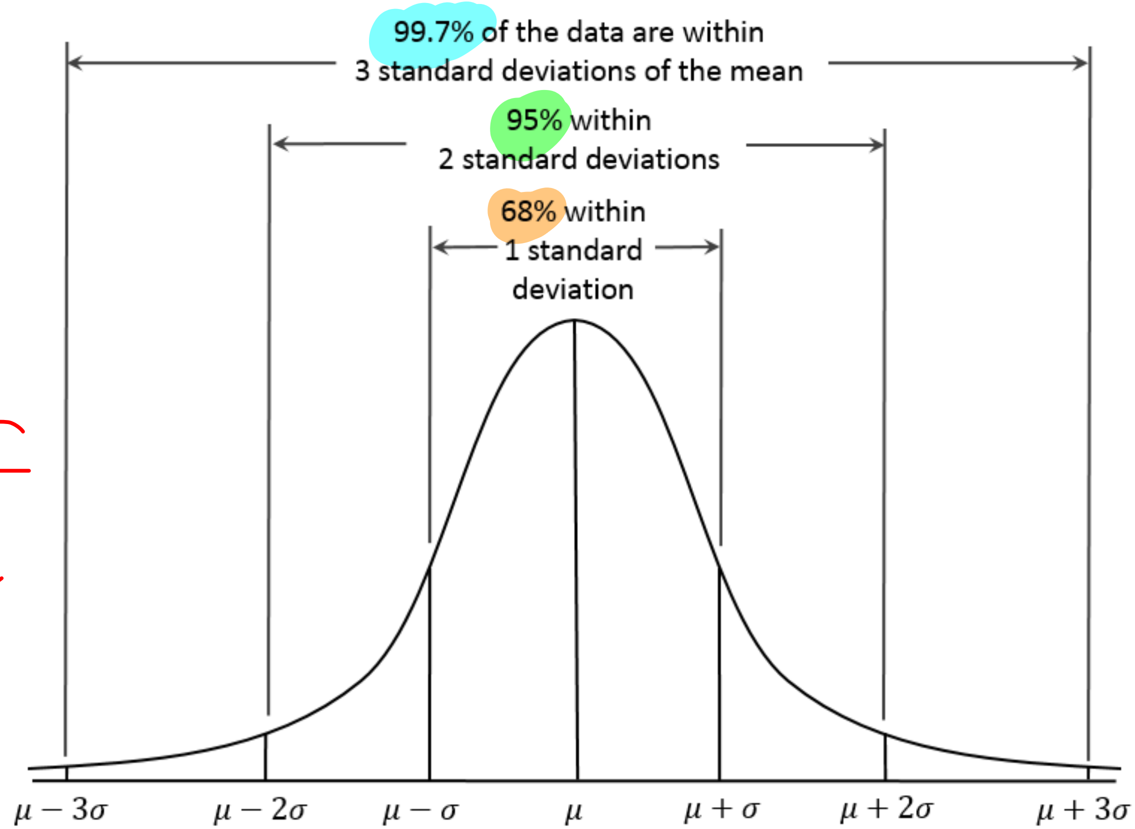
$1.28 = \dfrac{x-10}{2}$

$x = 2(1.28) + 10 = 12.56$

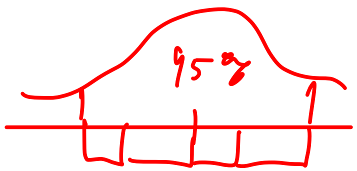
> iClicker Question:

What x value marks out the highest 1% of current measurements? (Hint: What % is less than the value x?)
Recall: $X \sim N(10,4)$
Use 2 decimal places.

Normal X dist: 68 – 95 – 99.7% Rule



Shape



Rough estimate

$\sigma = \frac{max - min}{4}$

- 68% of the pop values lie within 1 σ of μ.
- 95% of the pop values lie within 2 σ of μ.
- 99.7% of the pop values lie within 3 σ of μ.

Example (Using 68-95-99.7% Rule):

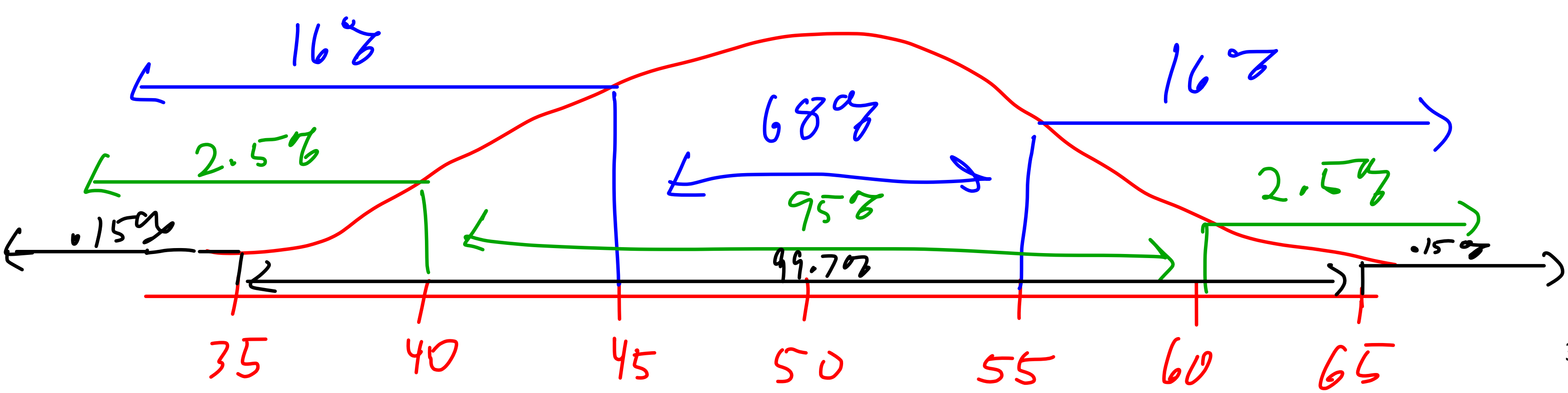
Lifetimes of batteries in a certain application are normally distributed with mean 50 hours and standard deviation 5 hours.

a) $P(40 < X < 60)$ $= .95$

b) $P(X > 45)$ $= .84$

c) $P(X < x) = 0.16$, find $x = 45$

d) $P(X > x) = 0.0015$, find $x = 65$



Summary

- A random variable (RV) has a probability distribution. A probability distribution includes: (1) the possible values of X and (2) their corresponding probabilities.

- A discrete RV has a probability mass function (pmf). A continuous RV has a probability density function (pdf).

- For continuous RVs:
  - A valid pdf, $f(x)$, must satisfy: (1) $f(x) \geq 0$ and (2) $\int_{-\infty}^{\infty} f(x)dx = 1$.
  - There is 0 probability at a single value, $x = a$: $P(X = a) = 0$
  - Probability is found by integrating the pdf: $P(a \leq X \leq b) = \int_{a}^{b} f(x)dx$
  - Integrate the pdf to obtain the CDF. Differentiate the CDF to obtain the pdf.
  - The cumulative density function (CDF) is: $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)\, du$
  - Cumulative probability is found by evaluating the CDF: $P(X \leq a) = F(a)$

- There are many continuous distributions. Examples are Normal, Uniform, and Exponential.

Summary

- For Normal RVs:
  - ➢ The normal distribution is a symmetric bell-shaped curve, described by its mean, μ, and variance, σ^2. The mean is the center of the distribution. The variance and standard deviation (SD=$\sqrt{Var}$) are measures of spread of the distribution. The inflection points of the distribution are 1 SD (σ) away from μ.
  - ➢ The distribution notation is: $X \sim N(\mu, \sigma^2)$
  - ➢ The normal pdf is not integrable, therefore a table (or calculator/computer) must be used to find probabilities.
  - ➢ To find normal probabilities, first standardize the Normal RV (X) to the Standard Normal RV (Z) using: $Z = \frac{X-\mu}{\sigma}$, then find prob in the Z-table.
  - ➢ Know how to find the probability, given z. Also know how to find z, given prob.
  - ➢ The Standard Normal RV Z is a symmetric, bell-shaped curve with mean $\mu = 0$ and variance $\sigma^2 = 1$. Distribution Notation: $\mathrm{Z} \sim N(0,1)$
  - ➢ The value of z tells us "the number of standard deviations x is above(below) μ"
  - ➢ 68%, 95%, and 99.7% of the values of a normal distribution lie within 1, 2, and 3 σ's of μ, respectively.

EGEN350

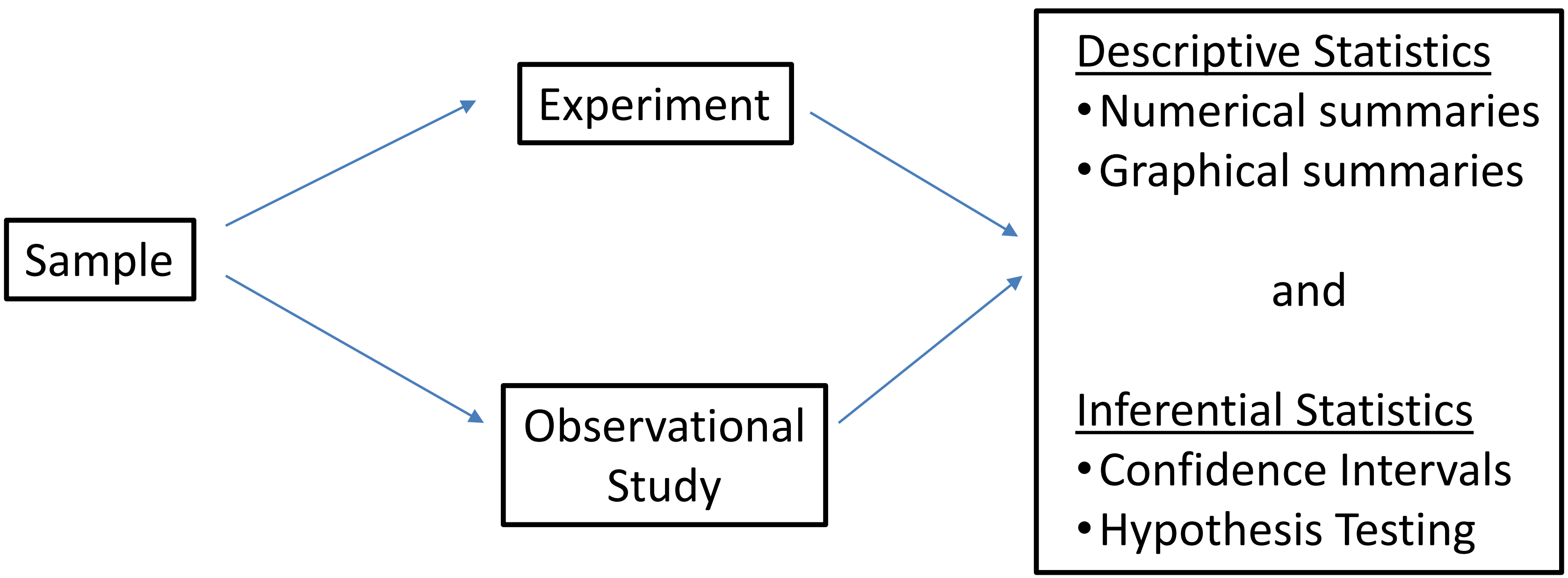
Chapter 1

Collecting Data

(Skip Sections 1.3 and 1.4)

Steps of Collecting & Analyzing Data

1. SELECT SAMPLE **2. COLLECT DATA** **3. ANALYZE DATA**



Sampling Designs

Examples:

1. Simple Random Sample (SRS)
2. Stratified Random Sample
3. Cluster Sample
4. Systematic Sample
5. Convenience Sample – BAD!

Good

Simple Random Sample (SRS)

- Each possible sample of size n has an equal chance of being selected.

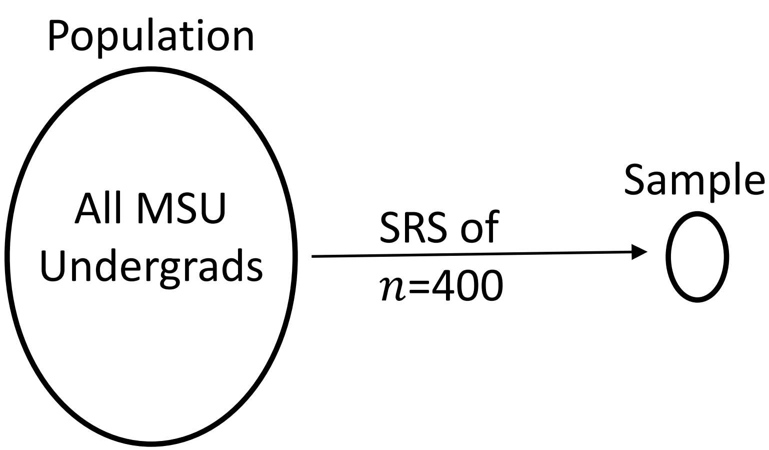
N = population size
n = sample size

Steps:

1. Create a numbered list of all N individuals in the pop.
2. Use a computer random number generator to randomly select n individuals from the list.

$\binom{N}{n} = \binom{17000}{400}$

Example:



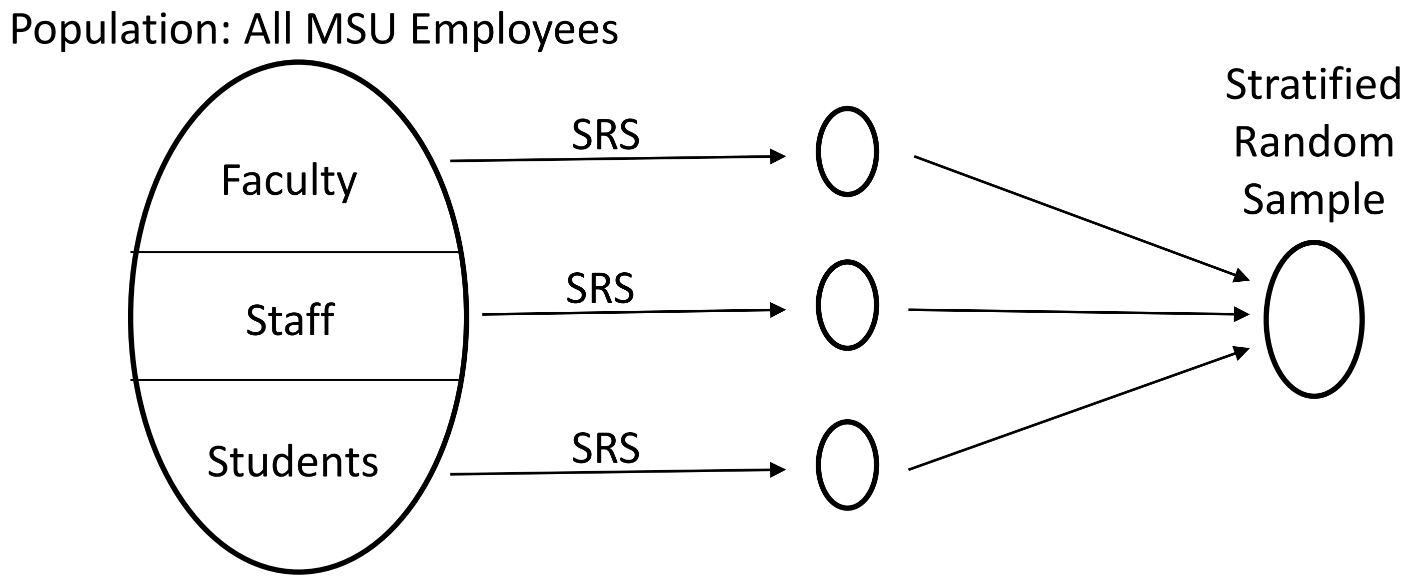
# of samples of 400 from population of 17000

Stratified Random Sample

Steps:

1. Divide the population into non-overlapping homogeneous groups, called strata. (Homogenous indicates the individuals within a stratum are similar with respect to the response.)
2. A SRS is drawn from each stratum.
3. Combine the SRSs to form the stratified random sample.

Example: Response = Annual Salary



Cluster Random Sampling

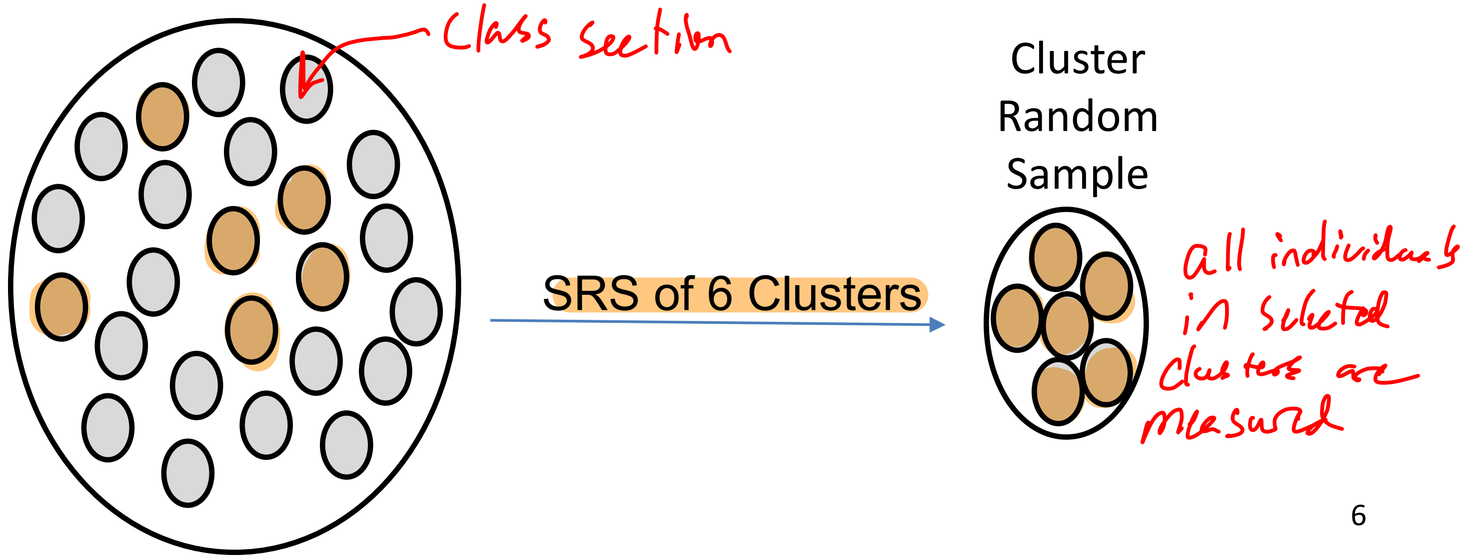
Steps:

1. The population naturally consists of non-overlapping heterogenous groups, called clusters. (Heterogenous indicates the cluster represents and reflects the variability present in the population.)
2. A SRS of clusters is selected.
3. All individuals in the selected clusters form the cluster sample.

Example:

Population:

All MSU STAT216 Students



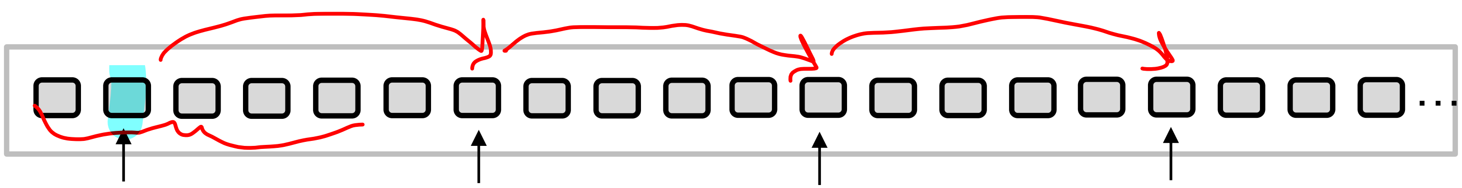
Systematic Random Sample

<u>Steps</u>:

1. Randomly select an individual between 1 and k from the numbered pop. list, then select every k^{th} individual thereafter.
2. Be certain that the variable of interest is not related to the order of the list. Or if the variable is related to the order of the list, be certain it is not related in a cyclic manner.

<u>Example</u>: Suppose k=5. Randomly select one of first 5 items, then continue to select every 5^{th} item thereafter…

Production line:



Suppose item 2 is randomly selected

Then item 7 is selected

Then item 12 is selected

Then item 17 is selected

Convenience Sample

- An "easily available" group is the sample.

➢ This is a BAD sampling plan if the individuals in your convenience sample systematically differ from the rest of the population and therefore does not represent the entire population.

Example: You are interested in all MSU student and employee opinions about updating gym equipment. You stand at the gym entrance and ask the next 50 people their opinion. This is a Poor Sample, because gym-goers may have systematically different opinions than non-gym-goers. This sample does not represent the population!

> iClicker Question:

To obtain an estimate of the amount of sleep residents of Bozeman are getting, John split the residents of Bozeman into multiple age groups (0-4, 5-10, 11-15, etc), and took a simple random sample of individuals from each age group. What is the name of the sampling method described here?

A. Simple Random Sample
B. Stratified Random Sample
C. Cluster Random Sample
D. Systematic Random Sample
E. Convenience Sample

> iClicker Question:

To obtain an estimate of the amount of sleep residents of Bozeman are getting, Joan split the residents of Bozeman by neighborhood, took a simple random sample of 10 neighborhoods, and asked all residents within each selected neighborhood. What is the name of the sampling method described here?

A. Simple Random Sample
B. Stratified Random Sample
C. Cluster Random Sample
D. Systematic Random Sample
E. Convenience Sample

Experiment vs. Observational Study

- Experiment – an experiment deliberately imposes a treatment on individuals in order to record their responses.
- Observational study – a study which observes individuals and measures variables, but does not attempt to influence the response.

To Recognize the Difference, Ask:

"Was there a treatment imposed on the individuals?"

- In an experiment, the researcher determines which individuals receive which treatment.
- In an observational study, the individuals have already self-chosen their groups.

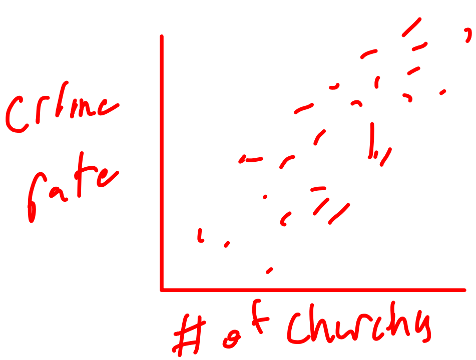
Experimental Design

(i.e. Method of assigning treatments to individuals)

a steel rod (may be different coatings applied)

DEFINITIONS:

- Unit: an individual in an experiment (subject=humans)
- Response: variable of interest ← how strong is the rod?
- Factor: a categorical explanatory variable
- Treatment: combination of factors' levels (categories)
- Extraneous Variable: a variable that is not of primary interest, but it affects the response.



Extraneous variable: population size

Example: Are cell phones harmful?

Cell phones have become the must-have gadget, but how safe are they? Cell phones emit electromagnetic radiation. There is concern that the heavy use of cell phones may increase a person's risk of getting cancer. Is there an association between cell phone use and occurrence of cancer?

- <u>Study 1</u>: A German study compared 118 patients with a rare form of eye cancer to 475 healthy patients who did not have the eye cancer. The patients' cell phone use was measured using a questionnaire. The eye cancer patients used cell phones more often, on average, than the healthy patients.

- <u>Study 2</u>: An Australian study used 200 mice. One hundred (100) mice were exposed for two half-hour periods a day to the same kind of microwaves with roughly the same power as the kind transmitted from a cell phone. The other 100 mice were not exposed. After 18 months, the brain tumor rate for the mice exposed to cell phone radiation was twice as high as the brain tumor rate for the unexposed mice.

Which study of more convincing?

> iClicker Question:

Regarding the German study, can we conclude the cell phone use *caused* the eye cancer? Why?

A. Yes, because this study was conducted in people, and the data show that people with eye cancer use their cell phones more often.
B. No, because other extraneous variables are not accounted for in this study. For example, what if people that use their cell phone more are generally less healthy?

> iClicker Question:

Regarding the Australian study, can we conclude the microwaves <u>*caused*</u> the brain tumors? Why?

A. Yes, because this experiment eliminates the effect of extraneous variables, <u>on average</u>. Treatments are randomly assigned to units, thus the groups receiving the treatments are similar (i.e. balanced) with respect to extraneous variables such as healthiness.

B. Yes, because many mice (n = 200) were studied.

C. No, because some mice are less healthy than others. What if some of the less healthy ones were assigned to the microwave group?

D. No, because we don't exactly understand the mechanism why microwaves cause brain tumors.

Principles of Experimental Design

- <u>Replication</u> – treatments are applied to *<u>many</u>* units.
  - ➢ The experiment is replicated on many units to reduce the role of random variation due to uncontrolled extraneous variables.
  - ➢ An experiment can survey how variable the effect of interest is. Greater replication increases the signal-to-noise ratio.

- <u>Random Assignment</u> of treatments to the units.
  - ➢ Treatments are randomly assigned to units so that, <u>on average</u>, the values of extraneous variables for units receiving each treatment will be identical.

- <u>Control Group</u>
  - ➢ A control group is a treatment group that receives no active treatment. The units in a control group could receive a placebo (sham) treatment. A placebo treatment is a treatment that has no active ingredients but otherwise resembles the real treatment(s).

> iClicker Question: RhinoDeck

Wood-fiber composites are common building materials (Ex: RhinoDeck). However, wood fibers may lose strength when exposed to moisture or when frozen. A research team designed a study where sections of wood-fiber composite were subjected to three different humidities (24%, 45%, 99%), and were either frozen or not. Maximum strength was then determined for each sample using flexural testing.

What is the response variable?

A. Maximum Strength
B. Humidity (24%, 45%, 99%)
C. Frozen (Yes, No)
D. Composite materials

> iClicker Question: RhinoDeck

Wood-fiber composites are common building materials (Ex: RhinoDeck). However, wood fibers may lose strength when exposed to moisture or when frozen. A research team designed a study where sections of wood-fiber composite were subjected to three different humidities (24%, 45%, 99%), and were either frozen or not. Maximum strength was then determined for each sample using flexural testing. This is an experiment.

What are the factors (and their factor levels)?

A. Maximum Strength
B. Humidity (24%, 45%, 99%)
C. Frozen (Yes, No)
D. B. & C.

> iClicker Question: RhinoDeck

Wood-fiber composites are common building materials (Ex: RhinoDeck). However, wood fibers may lose strength when exposed to moisture or when frozen. A research team designed a study where sections of wood-fiber composite were subjected to three different humidities (24%, 45%, 99%), and were either frozen or not. Maximum strength was then determined for each sample using flexural testing. This is an experiment.

How many treatments are there?

Two Elements of Scope of Inference: Generalization vs. Causation

(1) Generalization (Inference) to the population

- This is valid when the "sample is randomly selected" from the population. The sample results can be 'generalized to the population' or 'inferred to be true in the population'.

(2) Causation (or Causal Relationship between the response and the explanatory variables)

- This is valid when the "treatments are randomly assigned" to the units in an experiment. When random assignment is used, the effect of potential extraneous variables will be balanced (i.e. have the same effect) on all treatment groups. Thus, the observed difference between the treatment groups must be <u>caused</u> by the treatment(s).

Scope of Inference

	Type of Study	
Type of Sample	Randomized Experiment	Observational Study
Random sample	**Causal relationship, and can generalize results to population.**	**Cannot conclude causal relationship, but can generalize results to population.**
NOT a random sample	**Causal relationship, but cannot generalize results to a population.**	**Cannot conclude causal relationship, and cannot generalize results to a population.**

Inferences to population can be made

Can only generalize to those similar to the sample due to sample not being representative of population

Can draw cause-and-effect conclusions

Can only discuss association due to potential extraneous variables

> iClicker Question:

What is the scope of inference for the cell phone study on Germans?

A. Infer Association to Those Similar to Sample
B. Infer Causation to Those Similar to Sample
C. Infer Association to Population
D. Infer Causation to Population

> iClicker Question:

What is the scope of inference for the cell phone study on mice?

A. Infer Association to Those Similar to Sample
B. Infer Causation to Those Similar to Sample
C. Infer Association to the Population
D. Infer Causation to the Population

> iClicker Question:

What do you need for the inference to apply to the population (i.e. generalize to the pop.)?

A. Random-selected sample
B. Randomly-assigned treatments in experiment

Summary

- Steps: 1) Select sample, 2) Collect data, 3) Analyze data
- Good sampling designs are Simple Random Sampling, Stratified Random Sampling, Cluster Sampling, and Systematic Sampling. A bad sampling design is Convenience Sampling. Know what each means.
- An observational study is a study which observes individuals and measures variables, but does not attempt to influence the response.
- An experiment deliberately imposes a treatment on individuals in order to record their responses.
- A unit is an individual in an experiment. The response is the variable of interest. A factor is a categorical explanatory variable. A treatment is a combination of factors' levels (categories). An extraneous variable is a variable that is not of primary interest, but it affects the response.
- Random Selection is a method of selecting individuals from a population. Random selection of the sample allows for Inference to the Population.
- Random Assignment is a method of assigning treatments to units. Random assignment of treatments to units allows for Causation to be stated.

EGEN350

Chapter 6

Numerical & Graphical Summaries

(Skip Section 6.2 and 6.6)

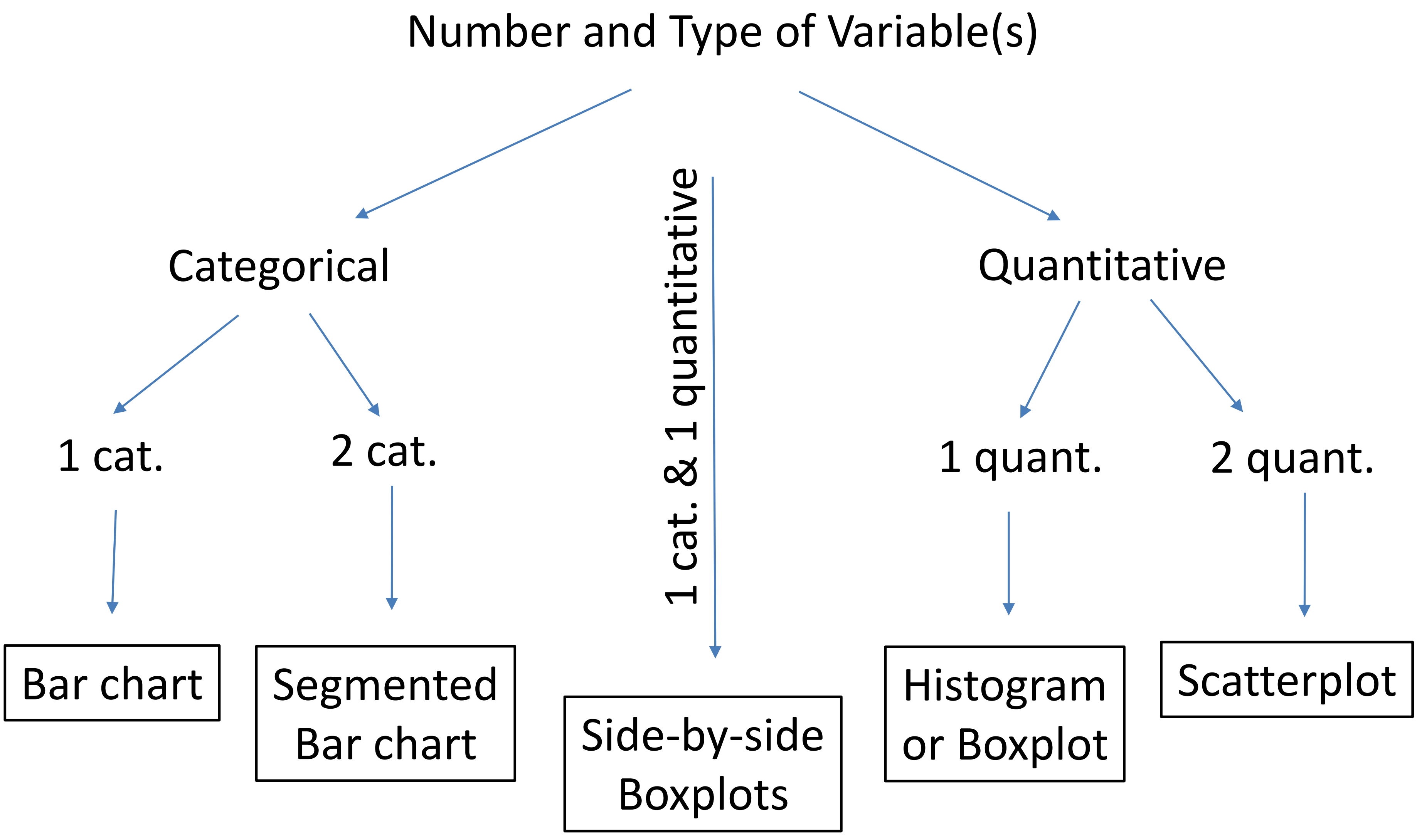
Types of Variables

Variable – any characteristic of an individual case

1) Categorical Variable - takes on categories
    - Examples: adhesive type, eye color, dog breed
    - Summarize using Counts, Proportions, or Percentages

2) Quantitative Variable - takes on numerical values on which ordinary arithmetic operations (such as averaging) make sense
    - Examples: time, weight, strength
    - Summarize using Mean & Standard deviation or Median & IQR

Graphs

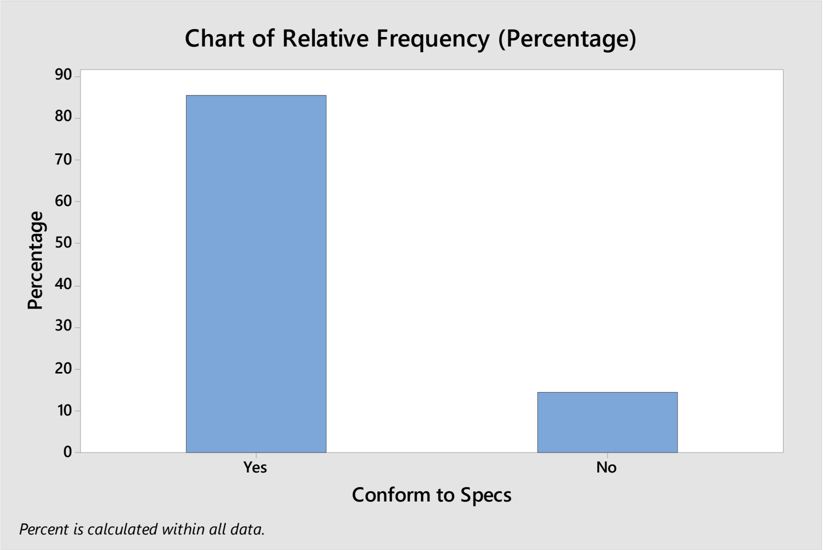

Number and Type of Variable(s)
Categorical
Quantitative
1 cat. & 1 quantitative
1 cat.
2 cat.
1 quant.
2 quant.
Bar chart
Segmented Bar chart
Side-by-side Boxplots
Histogram or Boxplot
Scatterplot

Example: Bar Chart (One Categorical Variable)

600 semiconductors wafers are categorized by whether they conform to a thickness specification.

Conforming	Nonconforming	Total
513	87	600

85.5% 15%

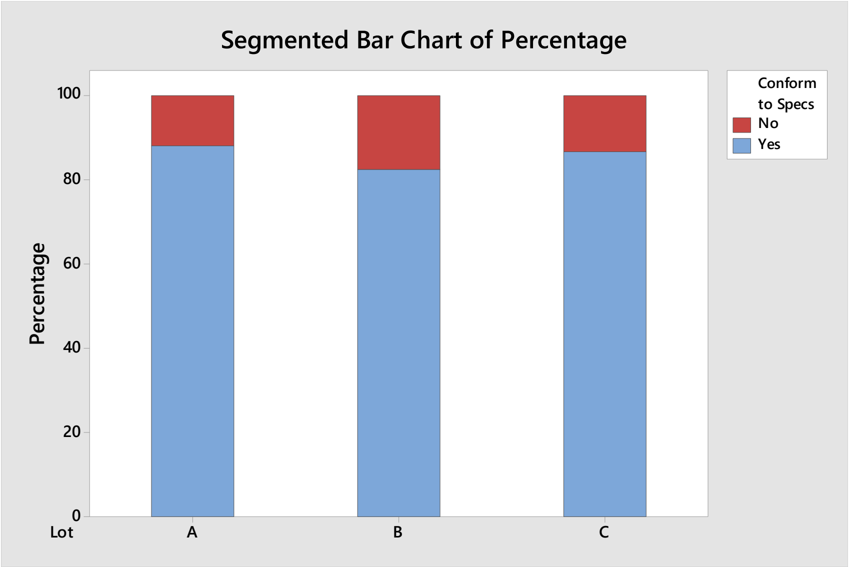


Example: Segmented Bar Chart (Two Categorical Variables)

600 semiconductors wafers are categorized by lot and whether they conform to a thickness specification.

Lot	Conforming	Nonconforming	Total
A	88	12	100
B	165	35	200
C	260	40	300
Total	513	87	600

Use Percentage if lot sizes differ

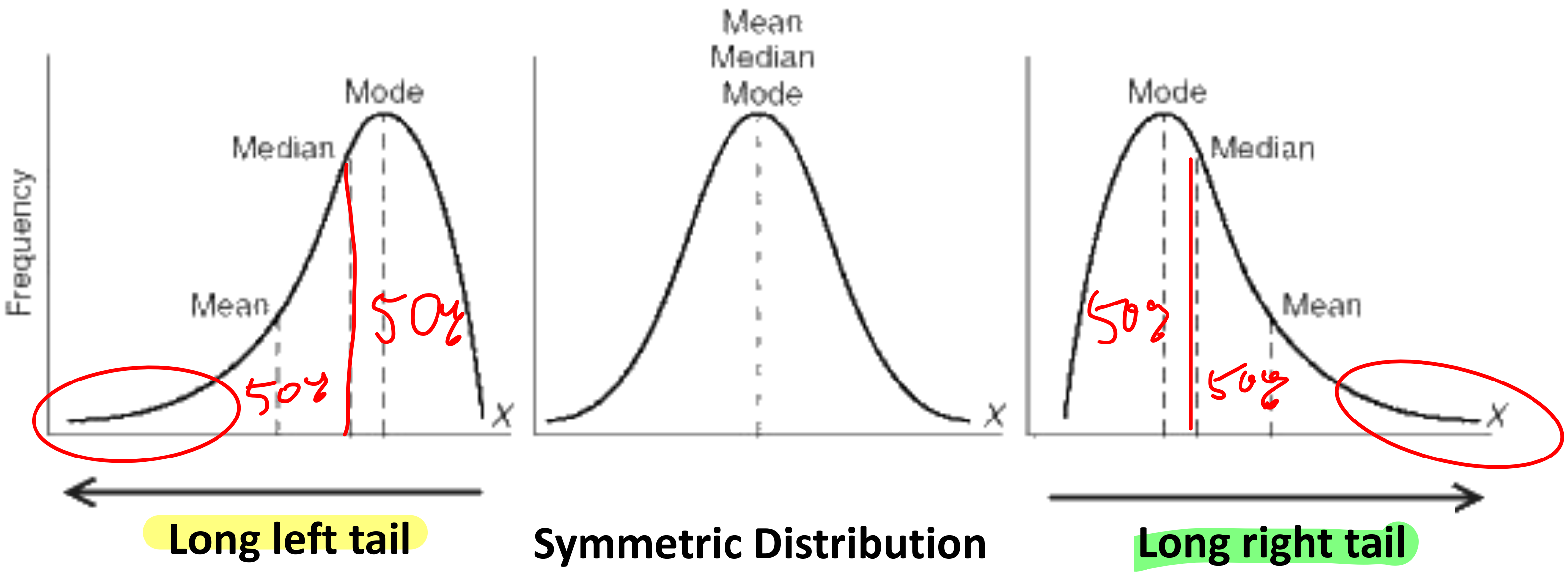


Distributions of Quantitative Variables

Features:

- Shape
- Center
- Spread (Variability)
- Outliers

Shape



(a) Left-skewed

(b) Normal distribution (One example of a symmetric distribution.)

(c) Right-skewed

IMPORTANT:

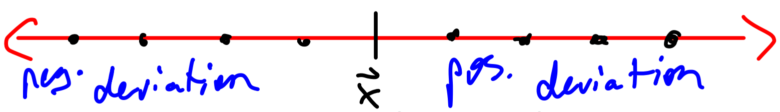
- The direction of skewness is the side with the long tail.
- The mean is pulled toward the longer tail.

Center

- **Mean** = arithmetic average = “x-bar”

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum x_i}{n}$$

  - The mean is the “balance point” – the value such that the sum of all the deviations from the mean is *zero:* $\sum \underbrace{(x_i - \bar{x})}_{\text{deviation}} = 0.$



- **Median** = middle value in ordered list = “x-tilde”
  - If n is odd, $\tilde{x} = \frac{n+1}{2}^{th}$ value in the ordered list
  - If n is even, $\tilde{x}$ is the average of the $\left(\frac{n}{2}\right)^{th}$ and $\left(\frac{n}{2}+1\right)^{th}$ values in the ordered list

Spread (Variability)

- **Sample variance:** $$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

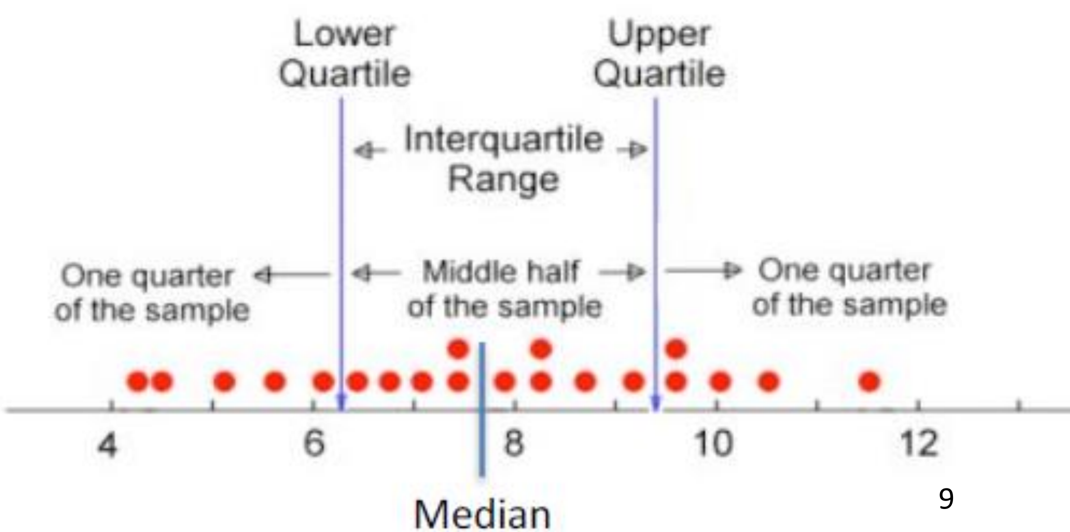
Variance and SD cannot be negative!
$s^2 \geq 0$ and $s \geq 0$

- **Standard deviation (SD):** $$s = \sqrt{s^2} = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$
  - *Roughly* the "average" distance data values are from the mean

- **Interquartile range:**

  $IQR = Q_3 - Q_1$

  - Q_3 = 75th percentile
  - Q_1 = 25th percentile



> iClicker Question:

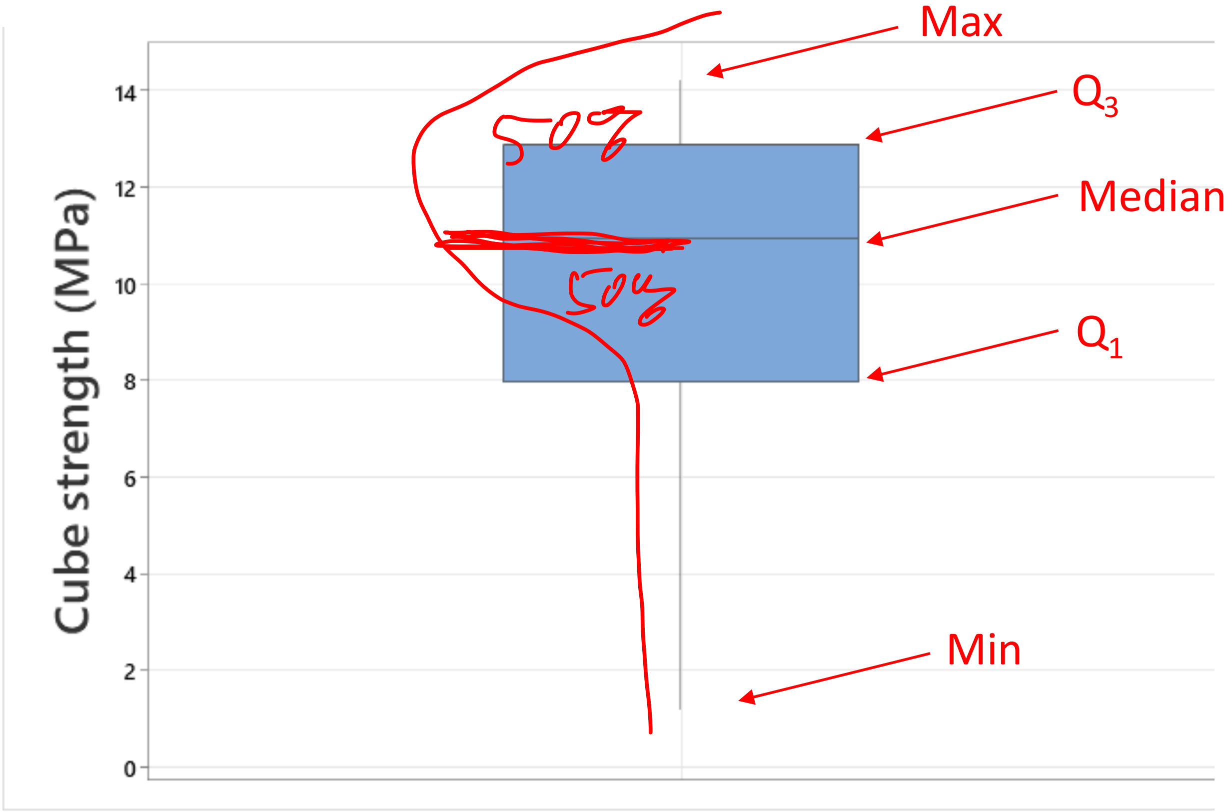
I asked 100 students how many hours they spent completing a certain homework (HW). The mean was $\bar{x} = 3$ hours. The standard deviation was s = 1.1 hours. What is the "standard deviation interpretation"?

A. On average, students in this dataset spent 1.1 hours completing the HW.
B. On average, the number of hours to complete the HW for students in this dataset is 1.1 hours above or below the mean of 3 hours.
C. On average, a student's number of hours to complete the HW is 1.1^2 hours above or below the mean of 3 hours.
D. Many students must have completed their homework exactly 1.1 hours above or below the mean of 3 hours.

Example: Boxplot

- Boxplot is a graph of the Five-number summary: Min, Q_1, Median, Q_3, Max
- Boxplots display: (1) shape, (2) center, (3) spread, and (4) outliers.

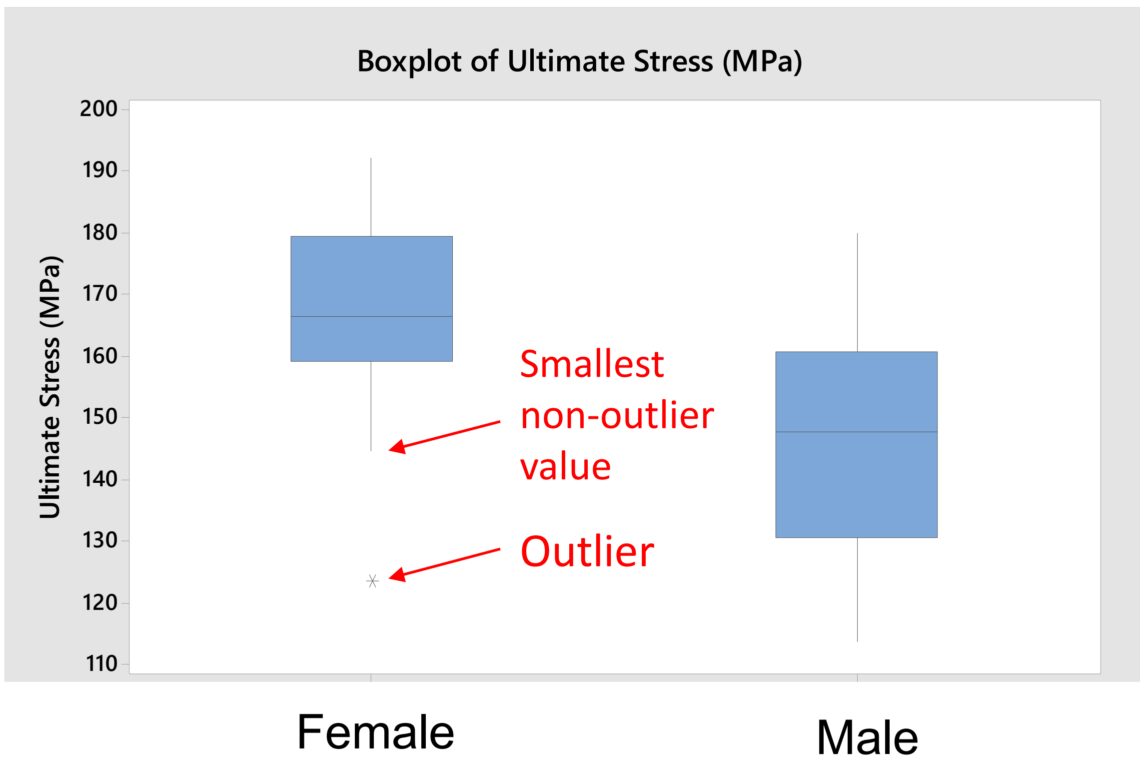
<u>Example</u>:



Example: Side-by-side Boxplots to Compare Groups

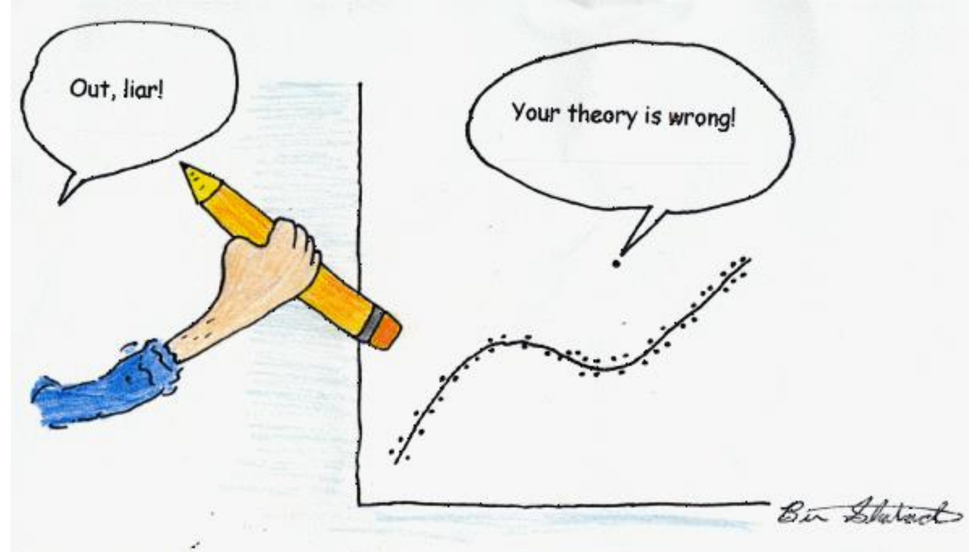
Example:

Do male or female mice have stronger bones?



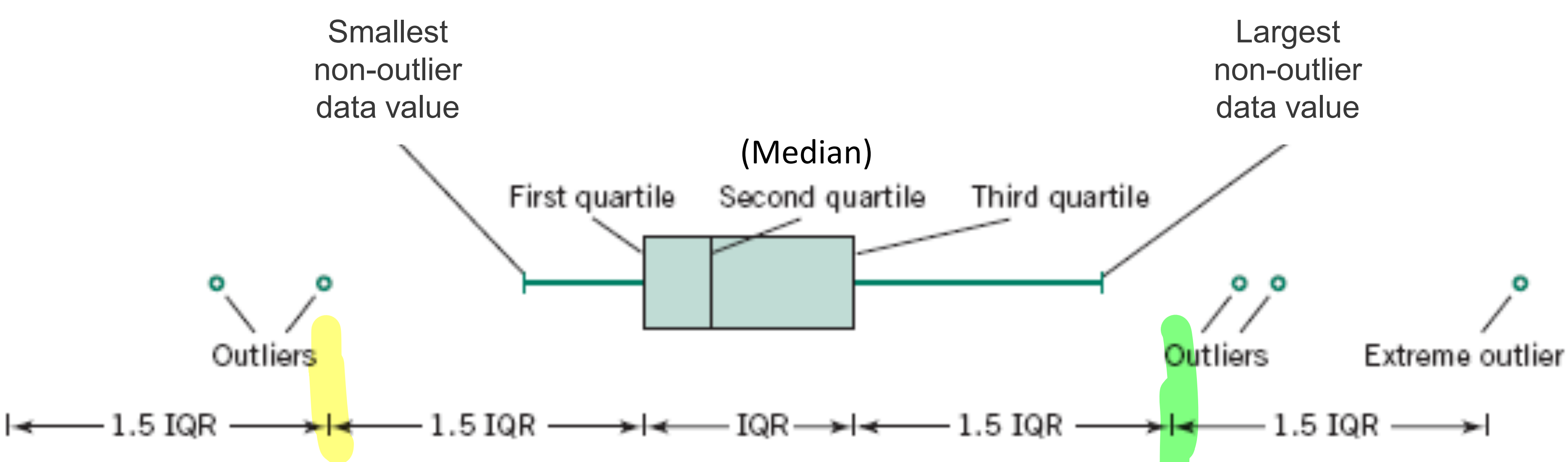
Outliers

An **outlier** is a data value that *doesn't fit the pattern of the majority of the data (i.e. much larger or smaller)*



Rule of thumb: A data value is considered an outlier if it is:

- *less than* $Q_1 - 1.5 \times IQR$ *or greater than* $Q_3 + 1.5 \times IQR$



Influence of Outliers and Shape on Mean and Median

- Outliers have a larger influence on the mean than median.

Ex: 1, 2, 3, 4, 5 → Mean = 3 Median = 3

1, 2, 3, 4, 500 → Mean = 102 Median = 3

If outliers or skewness are present use median and IQR

Otherwise, mean and S.D.

- Nearly Symmetric → Mean ≅ Median
- Skewed → Mean is pulled toward the longer tail

➢ If Mean << Median, then it's left-skewed (or has low outliers)

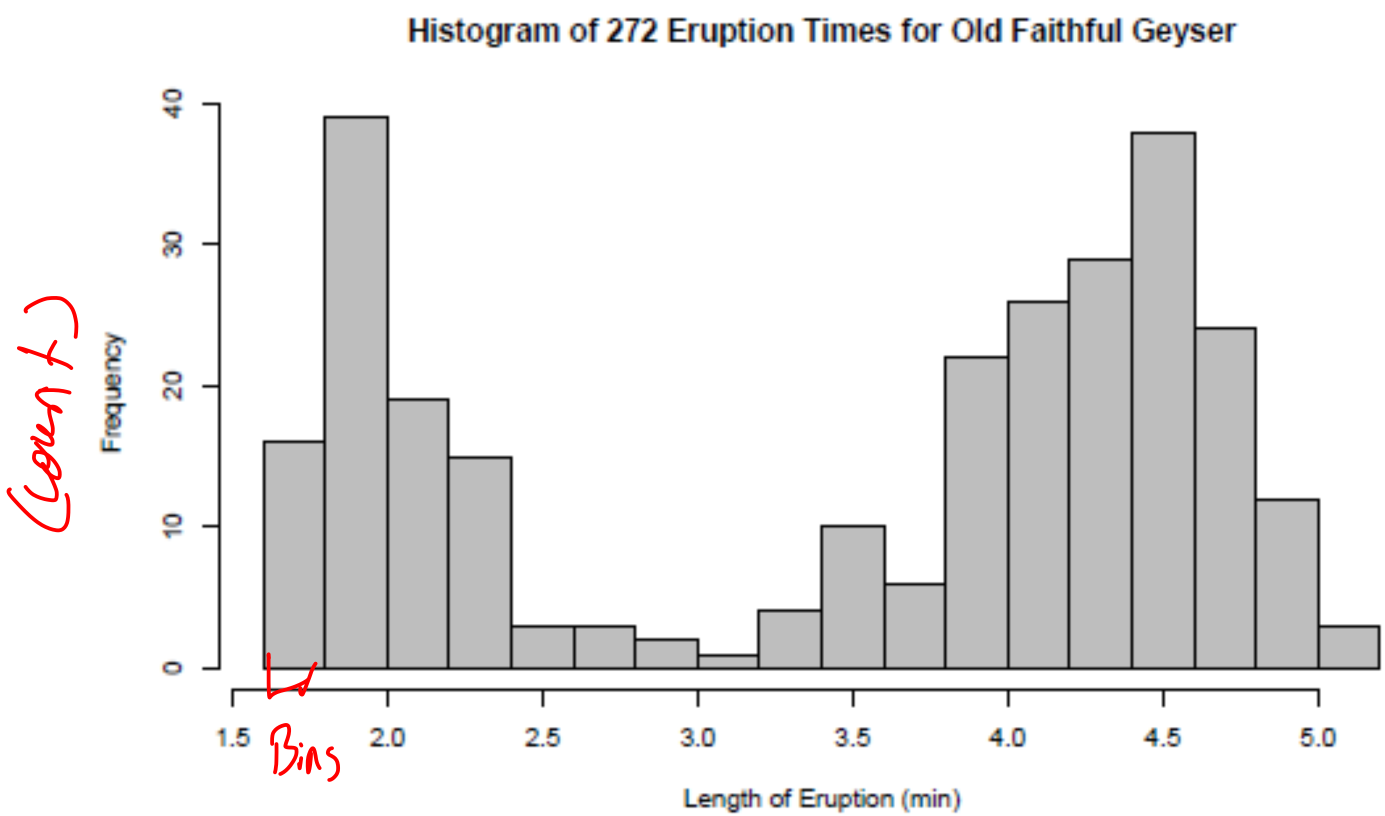
much smaller

➢ If Mean >> Median, then it's right-skewed (or has high outliers)

- The median is **resistant** to outliers (i.e. not greatly affected by outliers); the mean is *not* resistant to outliers.

Example: Histogram

Steps:
1. Divide the range of the data into intervals, called “bins”.
2. Draw a rectangle with height equal to the frequency in that bin.



EGEN350

Chapter 8

Confidence Interval for the Pop. Mean (μ)

(Skip Sections 8-3, 8-4, 8-5, 8-6, 8-7)

Overview of Statistical Inference

Goal: Estimate μ

Statistical Inference consists of two major areas:

1) Parameter estimation

single value

Ch7 • Point estimation (Ex: Use $\bar{X}$ to estimate μ)

Ch8 • Interval estimation (Ex: Form an interval around $\bar{X}$ to estimate μ with high confidence)

Ch9 2) Hypothesis testing

• Construct hypotheses about μ, collect sample data, and use $\bar{X}$ to draw a conclusion about μ.

Outline

- General Form of a Confidence Interval (CI)
- t Distribution
- CI for μ
- CI Interpretation
- CI Behavior / Trade-offs
- Sample Size Calculations

Confidence Interval (CI)

- A confidence interval is an “interval estimate” of the parameter.
- A confidence interval is an interval of plausible values for the parameter.
- A confidence interval has a level of confidence attached to it.
- For example, suppose we want to estimate the true mean diameter μ of steel rods by taking a sample. We are 95% confident that the true mean diameter value is in the interval (14.73, 14.91). pop.
- How do we find this CI?

General Form of a Two-sided CI

- The general form of a two-sided confidence interval for a parameter is: $\boxed{\text{point estimate} \pm \text{margin of error}}$

(point est. − margin of error , point est. + margin of error)

Confidence Level

- Confidence Level – a quantity (%) describing the success rate of the method (i.e. formula) being used.
- The "success" is when the confidence interval captures the parameter value.
- Commonly-used confidence levels: 90%, 95%, 99%

Two-sided* CI for μ, when σ is Known

If $\bar{X}$ is the sample mean of a random sample of size n from a normal population, a $100(1-\alpha)$ % CI for μ is:

$$\boxed{\bar{X} \pm \underbrace{Z_{\alpha/2}}_{\text{critical value}} \underbrace{\frac{\sigma}{\sqrt{n}}}_{=\sigma_{\bar{x}}}}$$

Two-sided* CI for μ, when σ is Unknown

FOCUS

If $\bar{X}$ and S are the mean and standard deviation of a random sample of size n from a normal population, a $100(1-\alpha)$ % CI for μ is:

use s to estimate σ

$$\boxed{\bar{X} \pm \underbrace{T_{\alpha/2}}_{\text{critical value}} \underbrace{\frac{S}{\sqrt{n}}}_{SE_{\bar{x}}}}$$

point estimate

margin of error

critical value

incorporates "confidence"

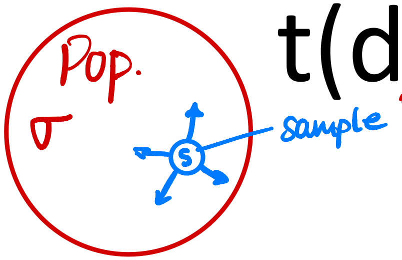
Assumptions:

1. The data must be a random sample from the population.
2. The $\bar{X}$ distribution must be at least approximately normal.

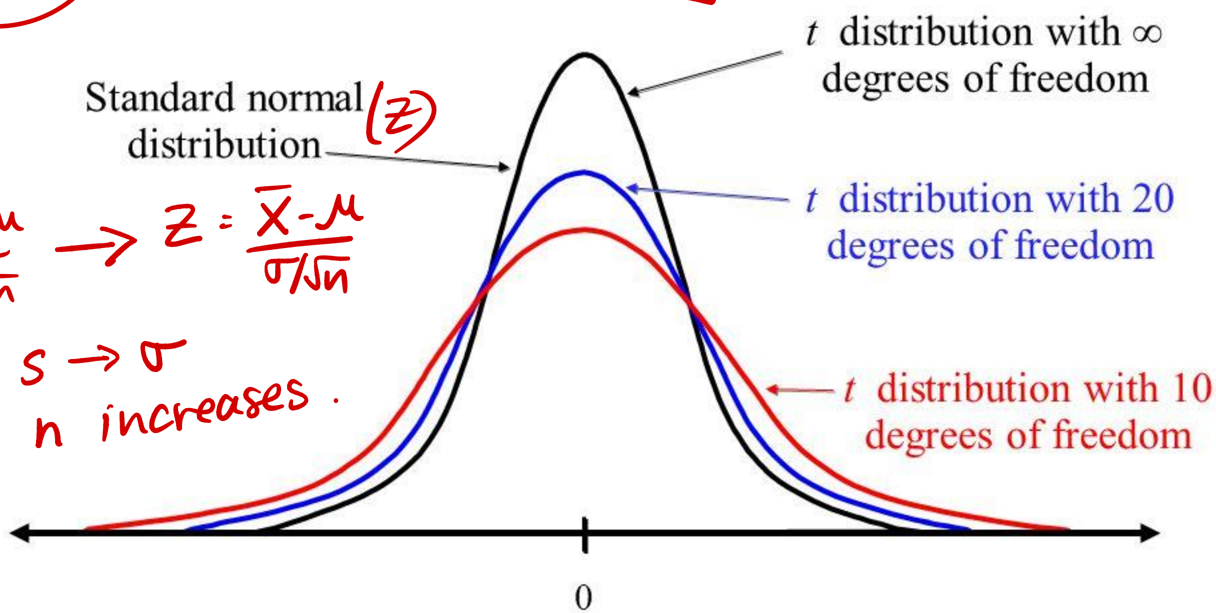
↘ "exactly normal" ($\bar{X} \sim N$) or "approximately normal" ($\bar{X} \mathrel{\dot\sim} N$)

* Footnote: One-sided CIs exist. Change $Z_{\alpha/2}$ to Z_α and use + or - , not both.

t(df) Distribution



$t(n-1)$



$T = \frac{\bar{X}-\mu}{s/\sqrt{n}} \rightarrow Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$

b/c $s \rightarrow \sigma$ as n increases.

Ex: Data and Deviations

x_i	$x_i - \bar{x}$
10	10 - 12 = -2
8	8 - 12 = -4
18	18 - 12 = 6
12	12 - 12 = 0
10	10 - 12 = -2
14	14 - 12 = 2

$\bar{x} = 12$ Sum = 0

Only n-1 deviations can "freely vary". The n^{th} is determined b/c sum = 0

- t dist is defined by its "degrees of freedom" (Degrees of freedom is the number of data values that can freely vary.)
- t dist is centered at 0
- t dist has more prob in its tails than the Z dist
- t dist approaches the Z dist as $n \rightarrow \infty$

T-table:

Left margin: Degrees of freedom (ν)

Top margin: Greater than prob. $P(T>t)$

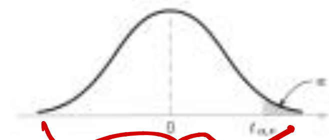


TABLE V Percentage Points $t_{\alpha,\nu}$ of the t Distribution

ν	.40	.25	.10	.05	.025	.01	.005	.0025	.001	.0005
1	.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	.289	.816	1.886	2.920	4.303	6.965	9.925	14.089	23.326	31.598
3	.277	.765	1.638	2.353	3.182	4.541	5.841	7.453	10.213	12.924
4	.271	.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	.267	.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	.265	.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	.263	.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	.262	.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	.261	.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	.260	.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	.260	.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	.259	.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	.259	.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	.258	.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	.258	.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	.258	.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	.257	.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	.257	.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	.257	.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	.257	.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	.257	.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	.256	.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	.256	.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	.256	.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	.256	.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	.256	.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	.256	.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	.256	.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	.256	.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	.256	.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	.255	.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
60	.254	.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
120	.254	.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	.253	.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

ν = degrees of freedom.

Interior t values

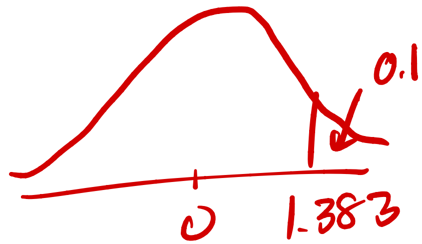
How to Read the t Table?

$t_{\alpha, \nu}$

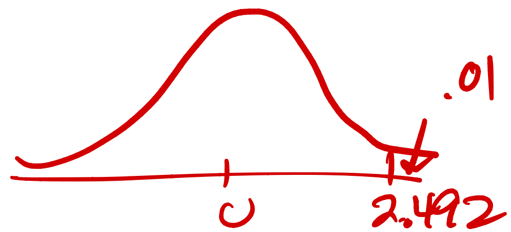
Greater than prob. (column heading)

Degrees of freedom (row heading)

- $t_{0.10, 9}$ = 1.383

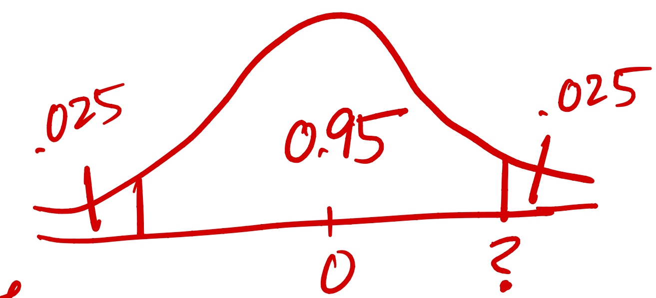


- $t_{0.01, 24}$ = 2.492



- $t_{\alpha, 5}$ = 2.571
  Find α = 0.025

- 95% two-sided CI, $n = 21$



$df = n-1 = 20$

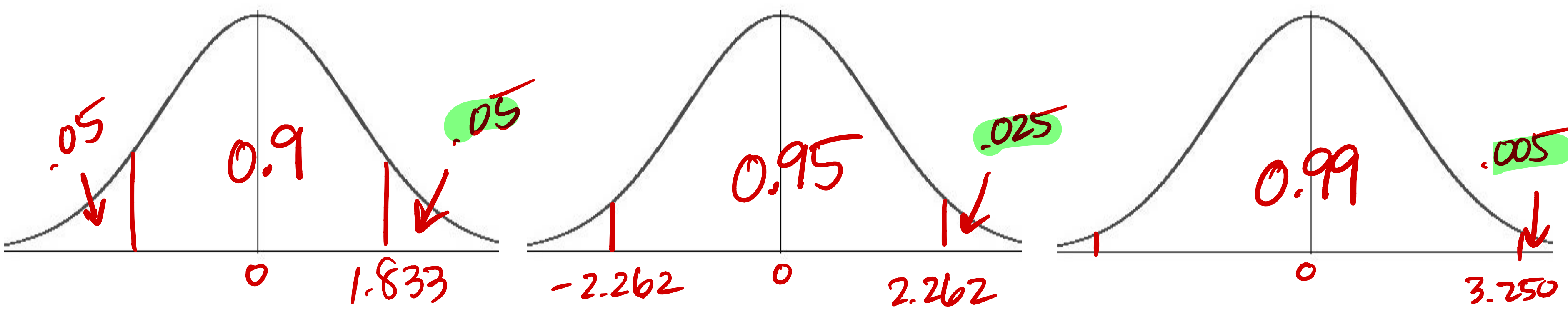
$t_{.025, 20} = 2.086$

IMPORTANT: If the desired degrees of freedom (df) is not in the table, use the next smaller df.

How to Find "Critical Value(s)"

Find $T_{\alpha/2}$ such that: $P(T < T_{\alpha/2}) = 1 - \frac{\alpha}{2}$

& $P(T > T_{\alpha/2}) = \frac{\alpha}{2}$



NOTE:
$T_{\alpha/2,9}$ is a positive number!

Confidence Level (C)	$T_{\alpha/2,9}$
90% (C=0.90, $\alpha = 0.10$)	$t_{.05,9} = 1.833$
95% (C=0.95, $\alpha = 0.05$)	$t_{.025,9} = 2.262$
99% (C=0.99, $\alpha = 0.01$)	$t_{.005,9} = 3.250$

Use df=9 for this exercise!

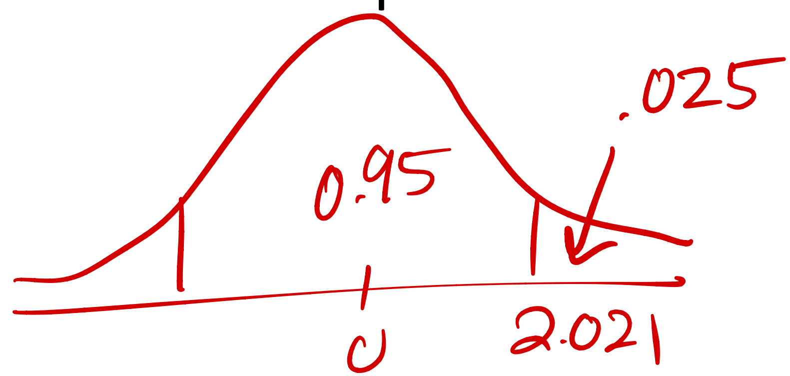
> iClicker Question:

Suppose a city plans to pay for a new parking garage through parking fees. For a random sample of 44 days, daily fees collected were \$126, on average, with a standard deviation of \$15. What is the "critical value" for a 95% CI for μ (the true mean daily income the parking garage will generate). Use 3 decimal places.

$t_{\alpha/2, n-1}$

$t_{.025, 43} = 2.021$

use df = 40 instead



Example: Steel Rods

A random sample of 100 steel rods manufactured on a certain extrusion machine is taken. The sample mean diameter was $\bar{x} = 14.82$ cm and the sample standard deviation was 0.54 cm. Construct a 95% confidence interval for the population mean μ.

FORMULA : $\bar{X} \pm T_{\alpha/2, n-1} \cdot \frac{S}{\sqrt{n}}$

$\bar{x} = 14.82$

$s = 0.54$

$14.82 \pm 2.000 \left[\frac{0.54}{\sqrt{100}} \right]$

$t_{.025, 99} = 2.000$

use 60 instead

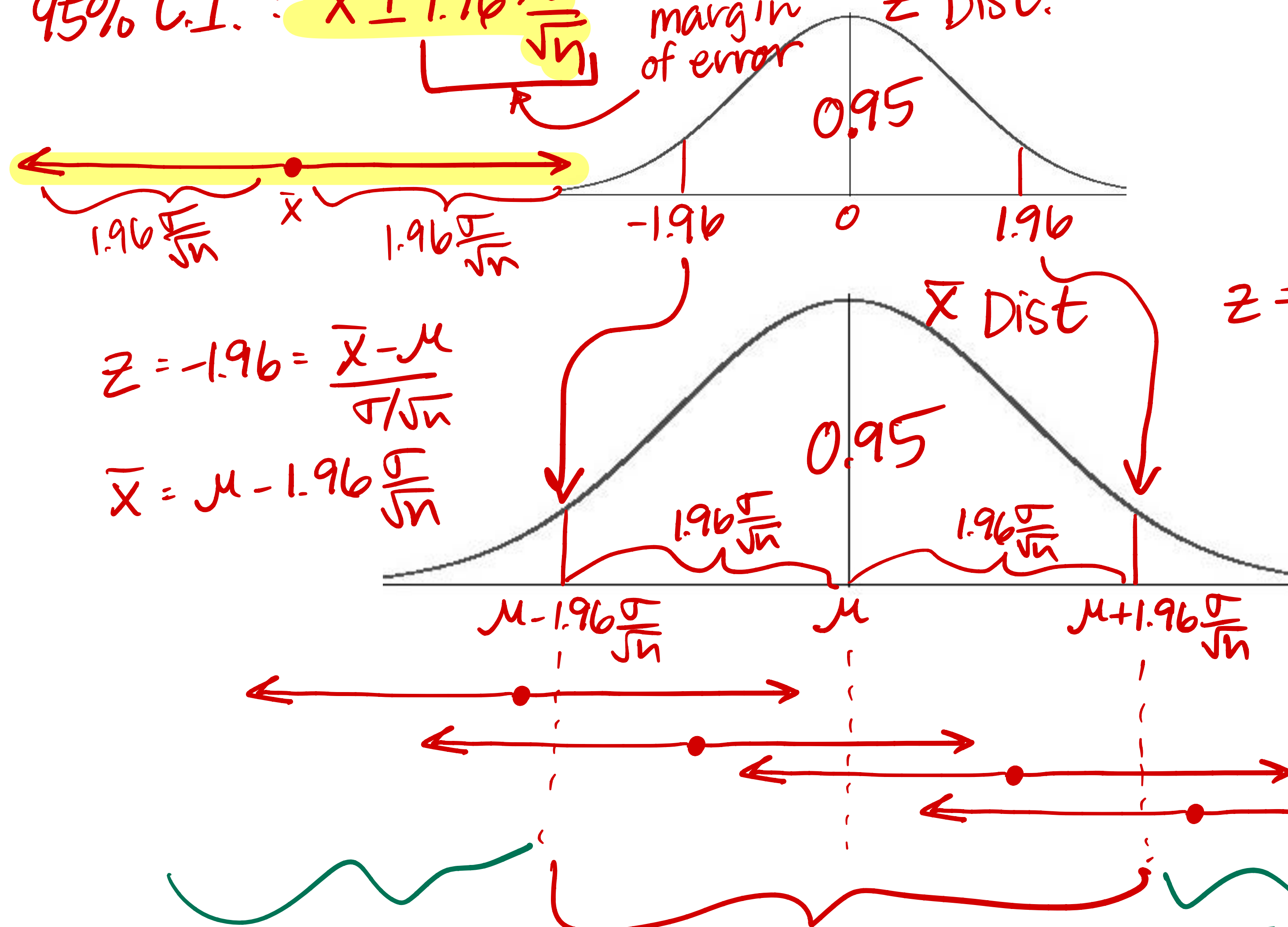
$= (14.712, 14.928)$

How do we interpret this interval?

Next up, we'll examine the theoretical behavior of all 95% CIs …

$$95\%\ C.I.:\ \bar{X} \pm \underbrace{1.96\cdot\frac{\sigma}{\sqrt{n}}}_{\text{margin of error}}$$

Z Dist.



$$z = -1.96 = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

$$\bar{X} = \mu - 1.96\frac{\sigma}{\sqrt{n}}$$

$$z = 1.96 = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

Solve for $\bar{X}$

$$\bar{X} = \mu + 1.96\cdot\frac{\sigma}{\sqrt{n}}$$

$\bar{X}$'s in central 95% will form CI's that capture μ

$\bar{X}$'s in tails will form CIs that do not capture μ

DITTO

Confidence in the Method

True for any level of confidence, such as: 80%, 90%, 99%, …

- If a 95% CI for μ was computed from each possible random sample drawn from the population, then:
  - ✓ 95% of these intervals <u>will contain</u> the true value of μ
  - ✓ 5% of these intervals <u>will not contain</u> the true value of μ

- BUT typically only <u>*one*</u> sample is selected from the pop. and therefore only <u>*one*</u> confidence interval is computed.

- What is the correct interpretation of our <u>one</u> CI?

Correct CI Interpretation

- The CORRECT INTERPRETATION of a *two-sided CI* is:

We are 95 % confident that μ lies between 42 and 48.

written in the context of the problem.

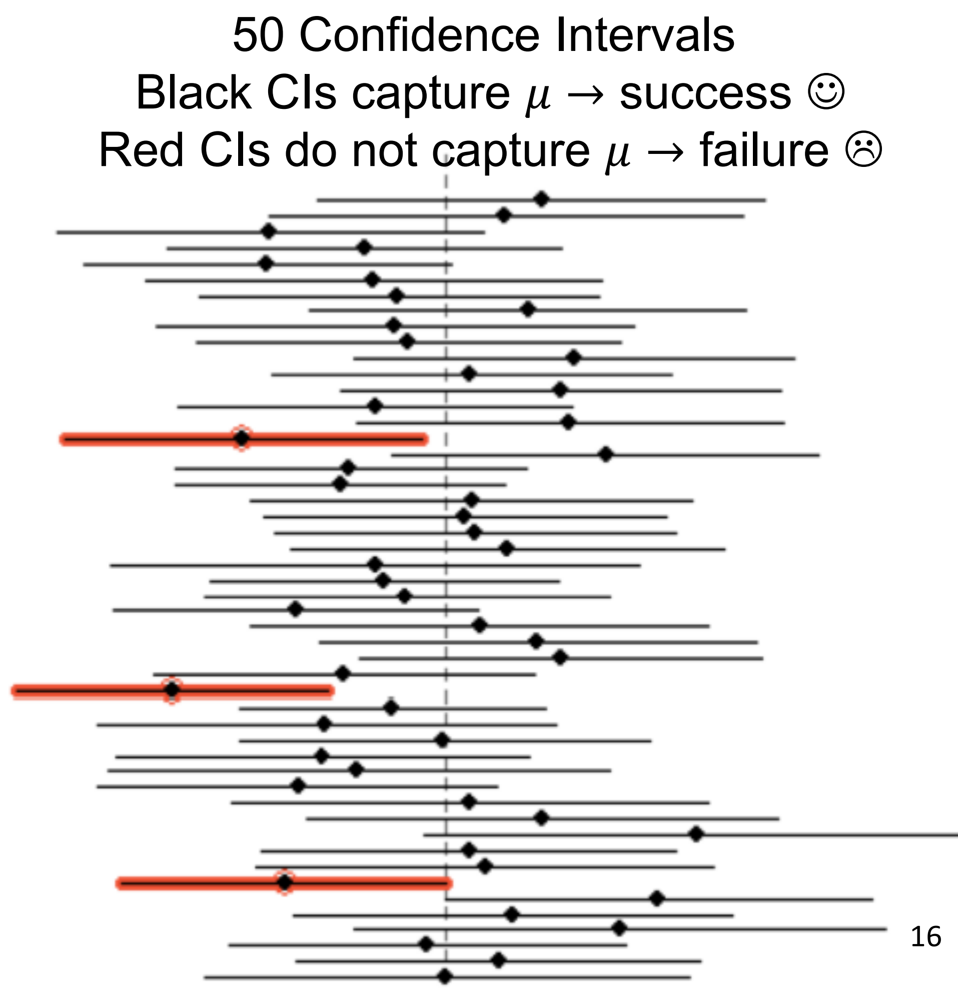
C.I.

BEWARE: Incorrect CI Interpretation

- An INCORRECT INTERPRETATION for a CI is:
  There is a 95% *chance* that μ is between 42 and 48.

- The true value of μ is unknown. The prob. of a CI capturing μ is either 0 and 1, we just don't know which it is for our problem.

- Our *one* CI either will or will not contain the true value of μ, so NEVER attach a prob. (like 0.95) to this occurrence.



Probability vs. Confidence (Timeline)

BEFORE

Coin is flipped.

AFTER

PROBABILITY	CONFIDENCE
There is a 50% chance of getting a head.	We are 50% confident it is a head.

Sample is selected.

PROBABILITY	CONFIDENCE
There is a 95% chance of getting a 95% CI that contains μ.	We are 95% confident the one computed 95% CI contains μ.

Example: Steel Rods (cont.)

A random sample of 100 steel rods manufactured on a certain extrusion machine is taken. The sample mean diameter was $\bar{x} = 14.82$ cm and the sample standard deviation was 0.54 cm. Recall the 95% CI: (14.71, 14.93)

Interpret the 95% CI in terms of the problem.

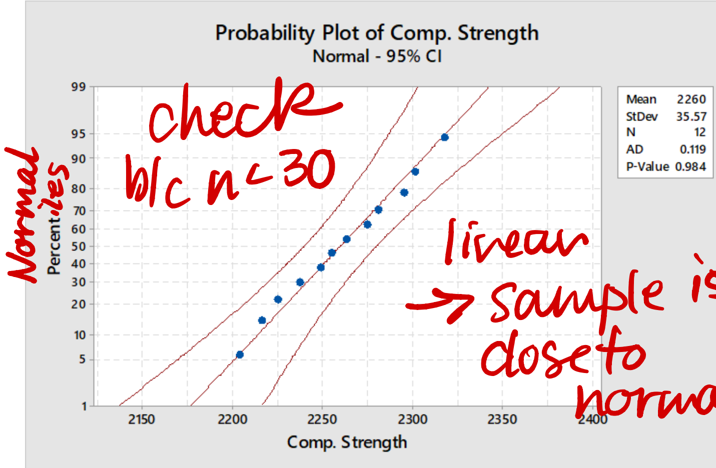
We are 95% confident that true mean diameter of steel rods manufactured on a certain machine is between 14.71 and 14.93 cm.

μ

in context of the problem

Example: Concrete

The compressive strength of concrete is being tests by a civil engineer who tested 12 specimens and obtains the following descriptive statistics. Find and interpret the 90% CI on the true mean compressive strength (psi).



$t_{.05,11} = 1.796$

Statistics

Variable	N (n)	N*	Mean ($\bar{x}$)	SE Mean ($\frac{s}{\sqrt{n}}$)	StDev (s)	Minimum	Q1	Median	Q3	Maximum
Comp. Strength	12	0	2259.9	10.3	35.6	2204.0	2228.0	2259.0	2291.5	2318.0

$$\bar{x} \pm T_{\alpha/2,\,n-1} \cdot \frac{s}{\sqrt{n}}$$

$$2259.9 \pm 1.796 \cdot \left(\frac{35.6}{\sqrt{12}}\right) = (2241.44,\ 2278.36)$$

We are 90% confident that the true (pop.) mean compressive strength of concrete is between 2241.44 and 2278.36 psi.

> iClicker Question:

involves confidence

$$E = Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$E = T_{\alpha/2}\frac{S}{\sqrt{n}}$$

The margin of error (E) determines the length of the CI.

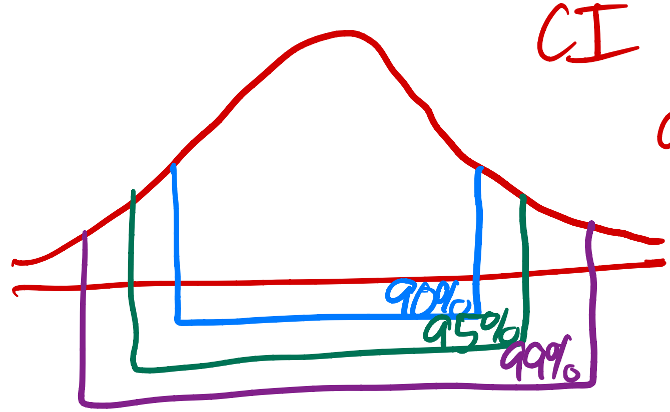
(A narrower interval is more precise. A wider interval is less precise.)

Assuming the other values remain constant,

if the confidence level (C) increases, then E and length ________.

A. increases

B. decreases



CI width increases as confidence increases

> iClicker Question:

$$E = Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$E = T_{\alpha/2} \frac{S}{\sqrt{n}}$$

The margin of error (E) determines the length of the CI.

(A narrower interval is more precise. A wider interval is less precise.)

Assuming the other values remain constant, if the variability (σ or S) increases, then E and length _________.

A. increases
B. decreases

Greater variability results in wider intervals.

> iClicker Question:

$$E = Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$E = T_{\alpha/2} \frac{S}{\sqrt{n}}$$

The margin of error (E) determines the length of the CI.

(A narrower interval is more precise. A wider interval is less precise.)

Assuming the other values remain constant,

if the sample size (n) increases, then E and length _________.

A. increases

B. decreases

Larger sample (more data) will result in a narrower interval.

It's a trade-off!

- We *want* to be highly confident that the parameter falls within a narrow interval (i.e. small margin of error) for decision making purposes.
- BUT, increasing confidence increases the length of the interval! Hence, the trade-off!
- Is there a way to get both?
  1. High confidence AND
  2. Narrow interval

Increase sample size!

Sample Size Calculation

- In situations when sample size can be controlled, we can choose n so that we have a specified % confidence that the margin of error is less than a specified bound, E.

$$\bar{X} \pm \underbrace{Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}}_{=E}$$

Sample Size Formula
$$n = \left(\frac{Z_{\alpha/2} \cdot \sigma}{E}\right)^2$$

$$E = Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

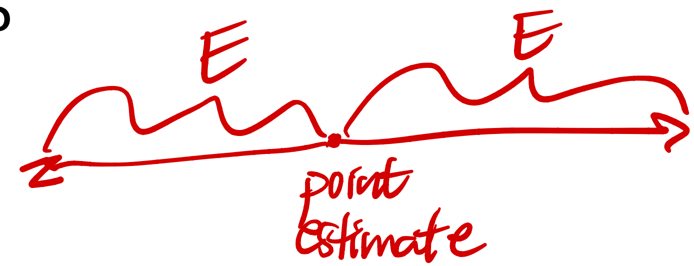
Solve for n

- If the sample size answer is not a whole number, it must be rounded up to the next larger whole number. This will ensure that the level of confidence does not fall below the specified confidence level.

Ex:
$n = 16.01$
Round up to $\boxed{n = 17}$

Example: Piston Rings

A manufacturer produces piston rings for an automobile engine. It is known that ring diameter is normally distributed with $\boxed{\sigma = 0.133}$ millimeters. How large must n be if the length of the 99% two-sided CI is to be 0.1?



FORMULA: $n = \left(\dfrac{z_{\alpha/2} \cdot \sigma}{E}\right)^2$

$= \left(\dfrac{2.576 \cdot .133}{.05}\right)^2$

$= 46.95$

$\boxed{\text{Use } n = 47}$

Length $= 0.1$

$2 \cdot E = 0.1$

$E = .05$

99% confidence

$z_{.005} = 2.576$

Summary

- A <u>confidence interval (CI)</u> is an "interval estimate" of the parameter. A CI is an <u>interval of *plausible* values for the parameter</u>
- <u>Confidence Level</u> – a quantity (typically stated as a percentage) that describes the success rate of the method (i.e. formula) being used. A "success" is when the confidence interval captures the parameter value.
- For example, if a 95% CI for μ was computed from <u>each possible</u> random sample of size n drawn from the population, then 95% of these intervals <u>will contain</u> the true value of μ and 5% of these intervals <u>will not contain</u> the true value of μ.
- The $100(1-\alpha)$ % two-sided CI to estimate μ is: $\bar{X} \pm T_{\alpha/2} \frac{S}{\sqrt{n}}$
- Use Z critical values in the CI formula only when the value of σ is known.
- <u>Assumptions</u>:
  1. The data must be a random sample from the population.
  2. The $\bar{X}$ distribution must be at least approximately normal.

Summary

- The t dist is symmetric, bell-shaped and centered at 0, similar to the Z dist, but the t dist has more prob in its tails than the Z dist. The t dist approaches the Z dist as $n \to \infty$. The t dist is defined by its "degrees of freedom" (df), which is the number of data values that can freely vary.
- How to find T critical values:
  - Find $T_{\alpha/2}$ such that $P(T < T_{\alpha/2}) = 1 - \frac{\alpha}{2}$ & $P(\mathrm{T} > T_{\alpha/2}) = \frac{\alpha}{2}$.
  - **<u>Note</u>: $\boldsymbol{T_{\alpha/2}}$ is a positive number.**
- The correct interpretation of a <u>two-sided</u> CI is:
  - *We are ___% confident that μ lies between ___ and ___. (written in context)*

Summary

- Confidence Interval Behavior:
  - If the confidence level (C) increases, then E and length of CI increases. If C decreases, then E and length decreases.
  - If the variability of the observations (σ) increases, then E and length increases. If σ decreases, then E and length decreases.
  - If the sample size (n) increases, then E and length decreases. If n decreases, then E and length increases.
- Sample size calculations:

CI:	One-sided	Two-sided
Sample size Formula:	$n = \left(\frac{Z_\alpha \cdot \sigma}{E}\right)^2$	$n = \left(\frac{Z_{\alpha/2} \cdot \sigma}{E}\right)^2$

  - We can choose n so that we have a specified % confidence that the margin of error is less than a specified bound, E.
  - If the sample size answer is not an integer, it must be rounded up to the next larger whole number.

Example (Cube data)

Michael is making a new kind of building material. He tested 12 cubes and found the following strengths, in megapascals.

Stat > Edit to enter data

Stat > calc > 1-var Stats to find summaries

$\bar{X}$ = mean (average)

S_x = standard deviation

Strength (MPa)
10.5
11.4
11.5
12.2
13.5
1.2
13.1
8.5
7.8
6.7
14.2
9.4

> iClicker Question:

What is the mean (average) of the data?

Strength (MPa)
10.5
11.4
11.5
12.2
13.5
1.2
13.1
8.5
7.8
6.7
14.2
9.4

> iClicker Question:

What is the standard deviation of the data?

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

Strength (MPa)
10.5
11.4
11.5
12.2
13.5
1.2
13.1
8.5
7.8
6.7
14.2
9.4

> iClicker Question:

What is the 5-number summary (Min, Q_1, Median, Q_3, Max) of the data?

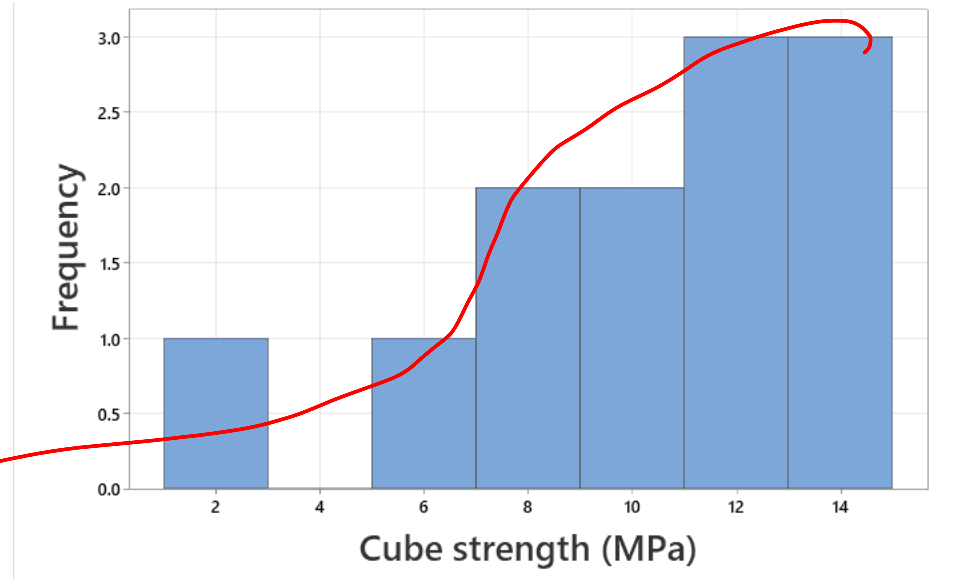
Ordered data: 1.2 6.7 7.8 8.5 9.4 10.5 11.4 11.5 12.2 13.1 13.5 14.2

A. 1.2, 7.8, 10.5, 12.2, 14.2
B. 1.2, 8.15, 10.5, 12.65, 14.2
C. 1.2, 8.15, 10.95, 12.65, 14.2
D. 1.2, 7.8, 10.95, 12.2, 14.2

> iClicker Question:

What is the shape of the histogram?

A. Symmetric
B. Right-skewed
C. Left-skewed

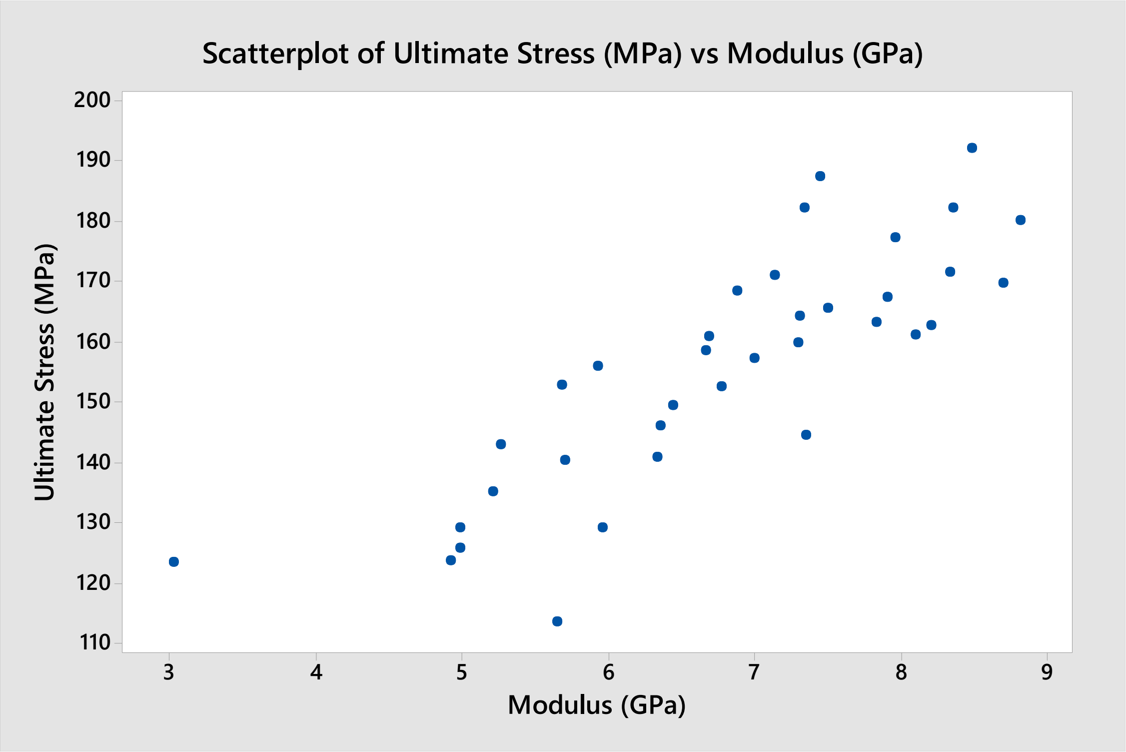


Long tail defines skewness

Example: Scatterplot

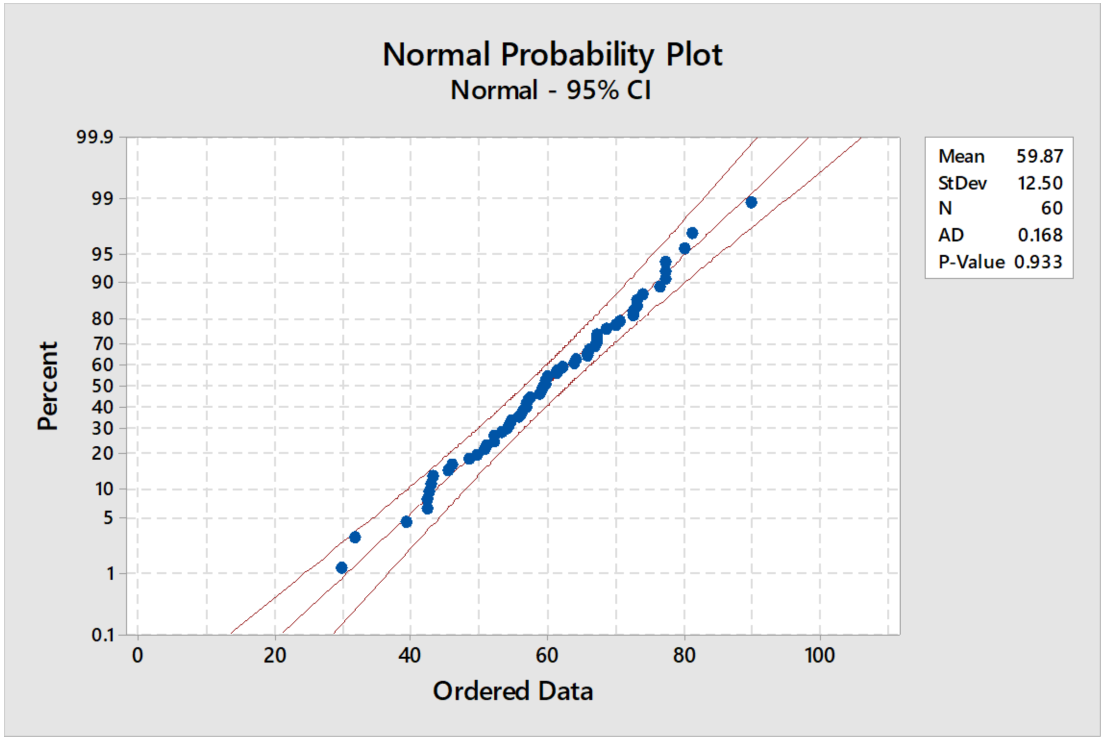
- A scatterplot displays the relationship between two quantitative variables.

- Example: Ultimate stress vs. modulus for mouse femurs



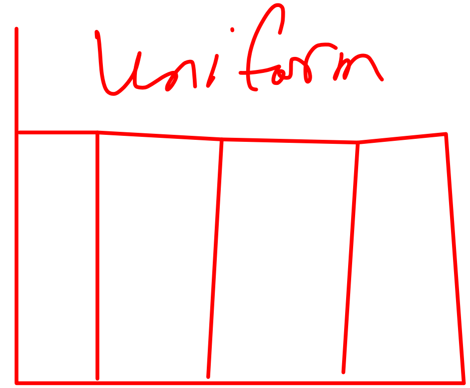
Example: Normal Probability Plot

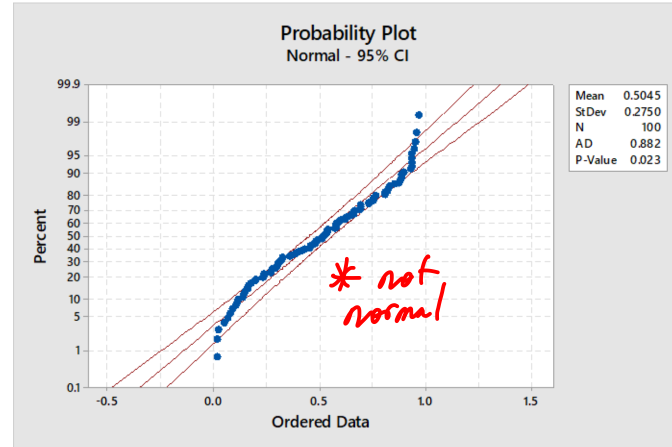
- Determines if the normal dist is a reasonable model for the data.
- The ordered sample values are plotted against the ordered standard normal values.
- If the plotted points fall approximately along a straight line, then the normal dist adequately describes the data.
- If the points deviate significantly from a straight line, then the normal dist <u>does not</u> adequately describe the data.

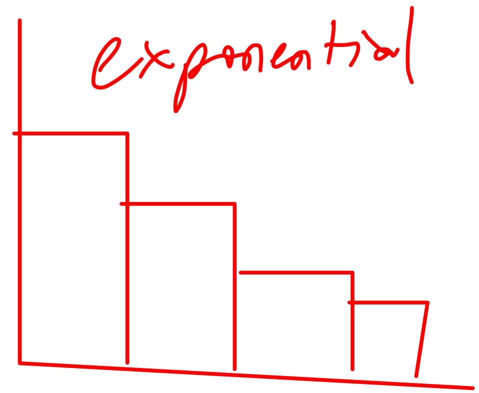


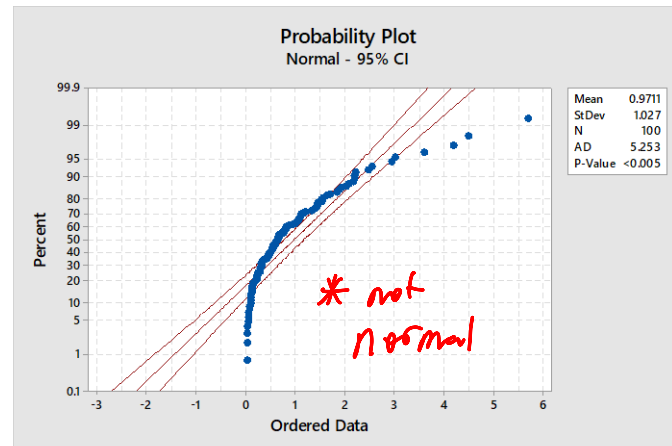
A special type of scatterplot to check the data closely follow a normal dist.

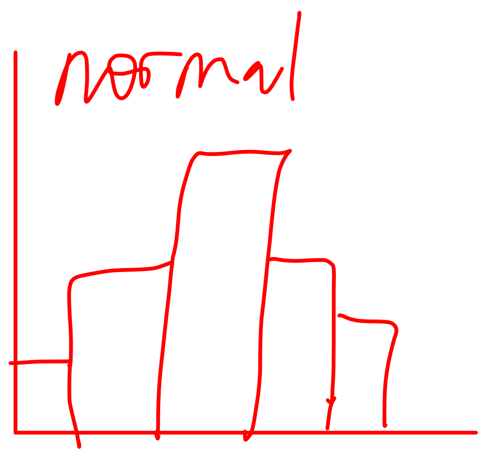
Is it reasonable to assume the data came from a normal population dist?

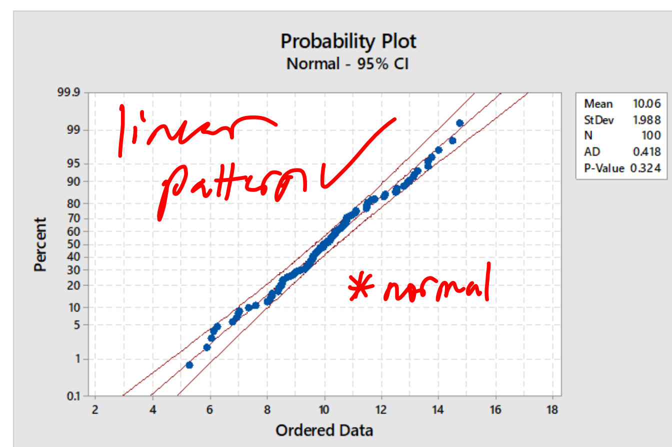








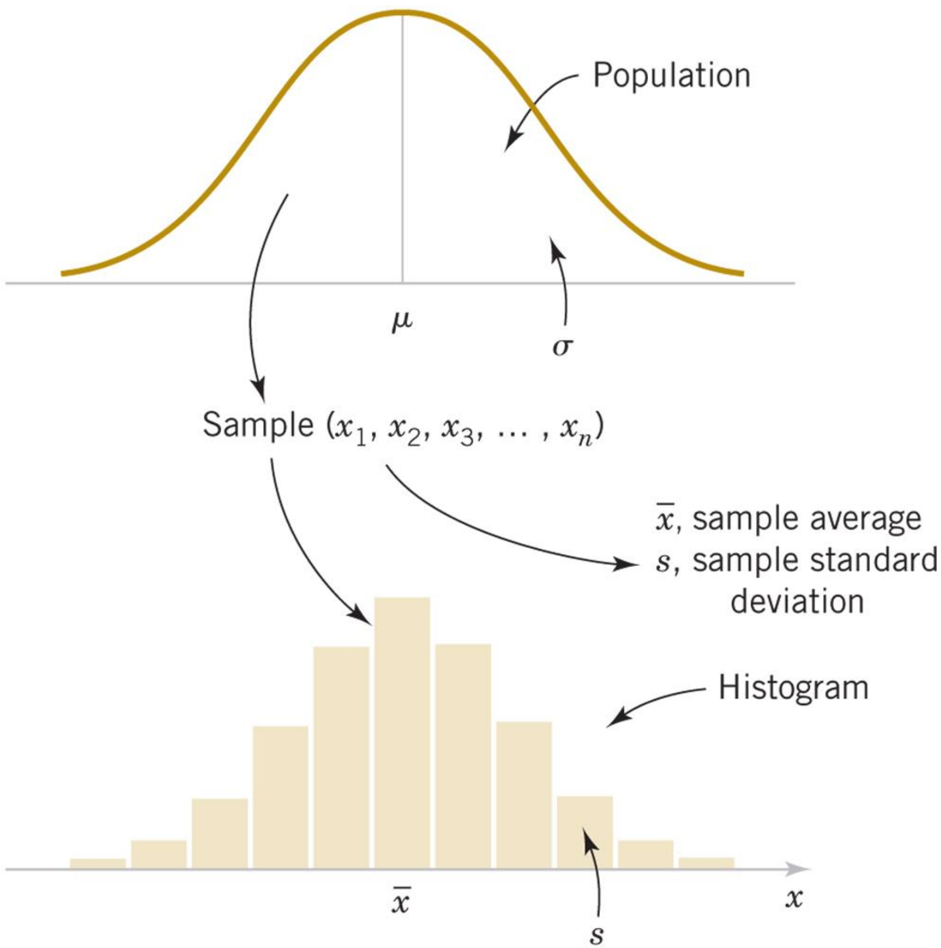




Population vs. Sample

<u>Population</u>: the entire collection of individuals about which information is sought.

<u>Sample</u>: the set of individuals from which data is collected. A sample is a subset of a pop.



Population Mean vs. Sample Mean

For a finite population with N equally-likely values, the population mean is:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

σ_x

For a sample of size n, the sample mean is:

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

s_x

Population Variance vs. Sample Variance

For a finite population with N equally-likely values, the population variance is:

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

For a sample of size n, the sample variance is:

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

Excel: =var.p()
=stdev.p()

=var.s()
=stdev.s()

TI Calculator: Use STAT > EDIT to enter data; Use STAT > CALC to find summary calculations such as mean, median, SD, etc.

Summary

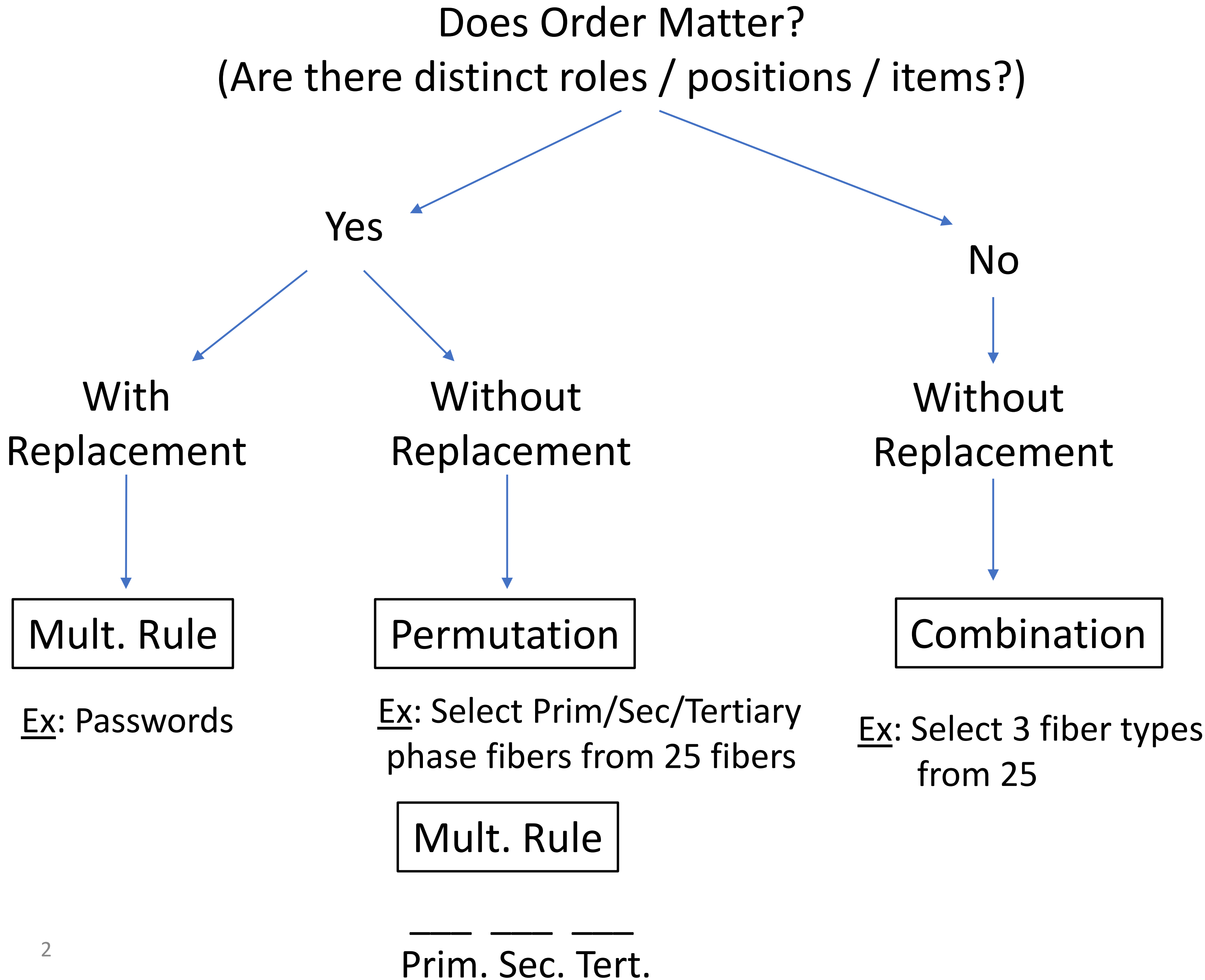
- A <u>variable</u> is any characteristic of an individual.
- There are two types of variables: <u>categorical</u> and <u>quantitative</u>
- Categorical variables are summarized using counts or proportions.
- Quantitative variables are summarized using Mean(Average)/Median, Variance/Standard deviation/IQR/Range, etc.
- <u>Graphs</u>: A <u>bar chart</u> is a graph of 1 categorical variable. A <u>segmented bar chart</u> is a graph of 2 categorial variables. A <u>side-by-side boxplot</u> is a graph of 1 categorical and 1 quantitative variable. A <u>histogram</u> or <u>boxplot</u> can be used to graph 1 quantitative variable. A <u>scatterplot</u> is a graph of 2 quantitative variables.
- An <u>outlier</u> is a data value that doesn't fit the pattern of the majority of the data (either larger or smaller). Outliers can drastically affect mean, variance, SD, and range. Median and IQR are resistant to outliers.
- <u>Normal Probability Plot</u> is a scatterplot of the ordered data values vs. percentiles from the Z distribution. It is used to access the whether the data closely follow a normal distribution.

Exam 1 Review

CHAPTER 2 – PROBABILITY:

<u>Probability Rules</u>:

1. $0 \leq \mathrm{P}(x_i) \leq 1$ where x_i is one possible outcome
2. Total probability of all possible outcomes = 1
3. <u>Additive</u>: $\mathrm{P}(x_1 \text{ or } x_2) = \mathrm{P}(x_1) + \mathrm{P}(x_2)$
4. <u>Complement</u>: $P(x_1') = 1 - P(x_1)$


Does Order Matter?
(Are there distinct roles / positions / items?)
Yes
No
With Replacement
Without Replacement
Without Replacement
Mult. Rule
Permutation
Combination
Ex: Passwords
Ex: Select Prim/Sec/Tertiary phase fibers from 25 fibers
Ex: Select 3 fiber types from 25
Mult. Rule
____ ____ ____
Prim. Sec. Tert.

Counting Techniques:

Multiplication Rule, Combination, or Permutation

Ex: There are 4 marbles in a bag (colors: red, blue, green, yellow). In how many ways can you select 2 marbles from the bag?

$C^4_2 = \binom{4}{2} = 6$ 4nCr2 = 6

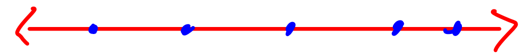
Ex: (Marbles, cont.) You are the winner of a new car! First marble selected will be the exterior color. Second marble selected will be the interior color. How many different outcomes are possible?

$P^4_2 = 12$

Ex: In an optics kit there are 6 concave lenses, 4 convex lenses, and 3 prisms. In how many ways can you choose one of each?

$\underset{\text{concave}}{\underline{6}} \times \underset{\text{convex}}{\underline{4}} \times \underset{\text{prisms}}{\underline{3}} = 72$

CHAPTER 3 – DISCRETE RANDOM VARIABLES:



Probability mass function (pmf)

- A valid pmf, $f(x)$, must satisfy: (1) $f(x_i) = P(X = x_i) \geq 0$

  (2) $\sum_{i=1}^{n} f(x_i) = 1.$

Ex: A chemical supply company currently has a certain chemical in stock, which it sells to customers in 5-lb lots.
X = the number of lots ordered by a randomly-chosen customer.

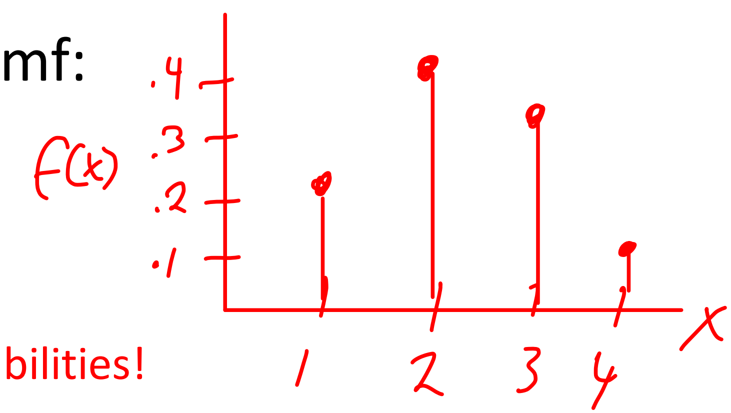
X	1	2	3	4
$P(X = x)$	0.2	0.4	0.3	? .1

- must sum to 1

.9

- $P(X < 4) =$ .9
- $P(X \leq 4) =$ 1

Graph of pmf:



* Pay attention to the equal sign for discrete probabilities!

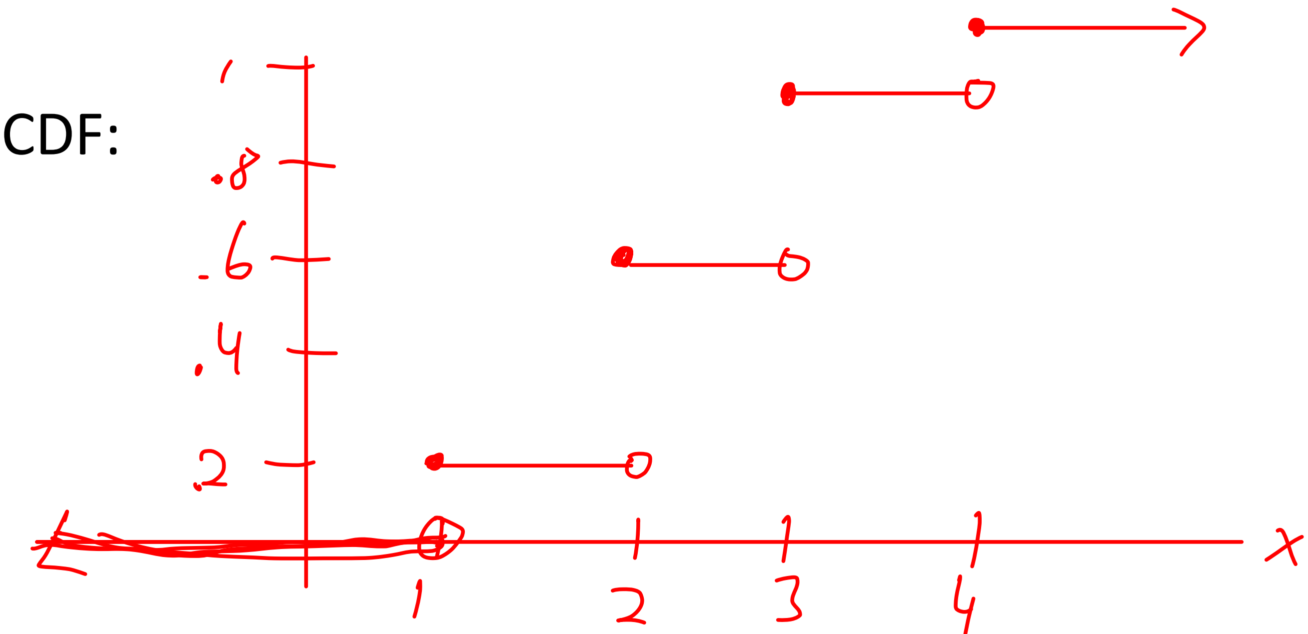
Cumulative distribution function (CDF)

- A valid CDF, F(x), must satisfy: (1) $F(x) = P(X \leq x_i) = \sum_{x \leq x_i} f(x_i)$
  (2) $0 \leq F(x) \leq 1$
  (3) If $x \leq y$, then $F(x) \leq F(y)$.
  Non-decreasing function

Ex: Chemical supply (cont.)

X	1	2	3	4
$P(X = x)$	0.2	0.4	0.3	? 0.1

Graph of CDF:



Binomial Distribution

X = number of successes in n trials

Probability formula: $P(X=x) = \underbrace{\binom{n}{x}}_{\text{combin.}} (p)^x (1-p)^{n-x}$

# of successes (above x) — # of failures (above $n-x$)

# of ways to choose x successes out of n trials

prob. of success

prob. of failure

Ex: 10% of people are left-handed. We select 4 people at random.

(a) What is the probability that none are left-handed?

$P(X=0) = \binom{4}{0}(.1)^0(.9)^4$

$(4nCr0)(.1)^0(.9)^4 = .6561$

(b) What is the probability that at least one of the 4 is left-handed?

$P(X \geq 1) = 1 - P(X=0)$

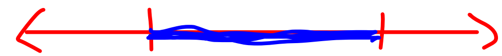
$1 - .6561 = .3439$

Possible X's

0, 1, 2, 3, 4

CHAPTER 4 – CONTINUOUS RANDOM VARIABLES:

<u>Probability density function (pdf)</u>



- A valid pdf, $f(x)$, must satisfy: (1) $f(x) \geq 0$ and
  (2) $\int_{-\infty}^{\infty} f(x)dx = 1.$

- Probability is found by integrating the pdf:

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

- For continuous RVs, $P(X \leq a) = P(X < a)$. because $P(X=a)=0$

<u>Ex</u>: $f(x) = 3x^2 \quad 0 < x < 1$

Use the pdf to find:

- $P(X < 0.5) = \int_0^{.5} 3x^2 dx$
- $P(X > 0.2) = \int_{.2}^{1} 3x^2 dx$
- $P(0.4 < X < 0.6) = \int_{.4}^{.6} 3x^2 dx$

Cumulative density function (CDF)

- The CDF is: $F(x) = P(\underset{\text{variable}}{X} \le \underset{\text{value}}{x}) = \underbrace{\int_{-\infty}^{x} f(u)\,du}$

The upper bound is x

Use a variable other than x here.

Ex: $f(x) = 3x^2 \quad 0 < x < 1$

- Find the CDF, for all x:

$\int_0^x 3u^2\,du = \frac{3u^3}{3}\Big|_0^x = x^3 - 0$

$$F(x) = \begin{cases} 0 & x \le 0 \\ x^3 & 0 < x < 1 \\ 1 & x \ge 1 \end{cases}$$

Use CDF to find:

- $P(X < 0.5) =$ $F(.5) = .5^3$
- $P(X > 0.2) =$ $1 - P(X \le .2) = 1 - F(.2) = 1 - .2^3$
- $P(0.4 < X < 0.6) =$ $P(X \le .6) - P(X \le .4) = F(.6) - F(.4)$
  $= (.6)^3 - (.4)^3$

Normal Distribution $\qquad$ Standard Normal Distribution

$X \sim N(\mu, \sigma^2)$ $\xrightarrow{\text{Standardize } X \text{ to } Z \text{ using:}}$ $Z \sim N(0,1)$

$$Z = \frac{X - \mu}{\sigma}$$

$\mu \qquad \sigma^2 \qquad \sigma = 2.5$

Ex: Heights of women: $X \sim N(64.5, 2.5^2)$

- $P(X > 67) = P\left(Z > \frac{67-64.5}{25}\right)$

  $= P(Z > 1) = 1 - P(Z \le 1) = 1 - .831435$

- $P(62 < X < 66) = P\left(\frac{62-64.5}{2.5} < Z < \frac{66-64.5}{2.5}\right)$

  $= P(-1 < Z < .6)$

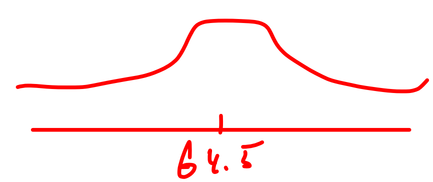
  $= P(Z \le .6) - P(Z \le -1) = .725747 - .158655$

- $P(X < x) = 0.02$

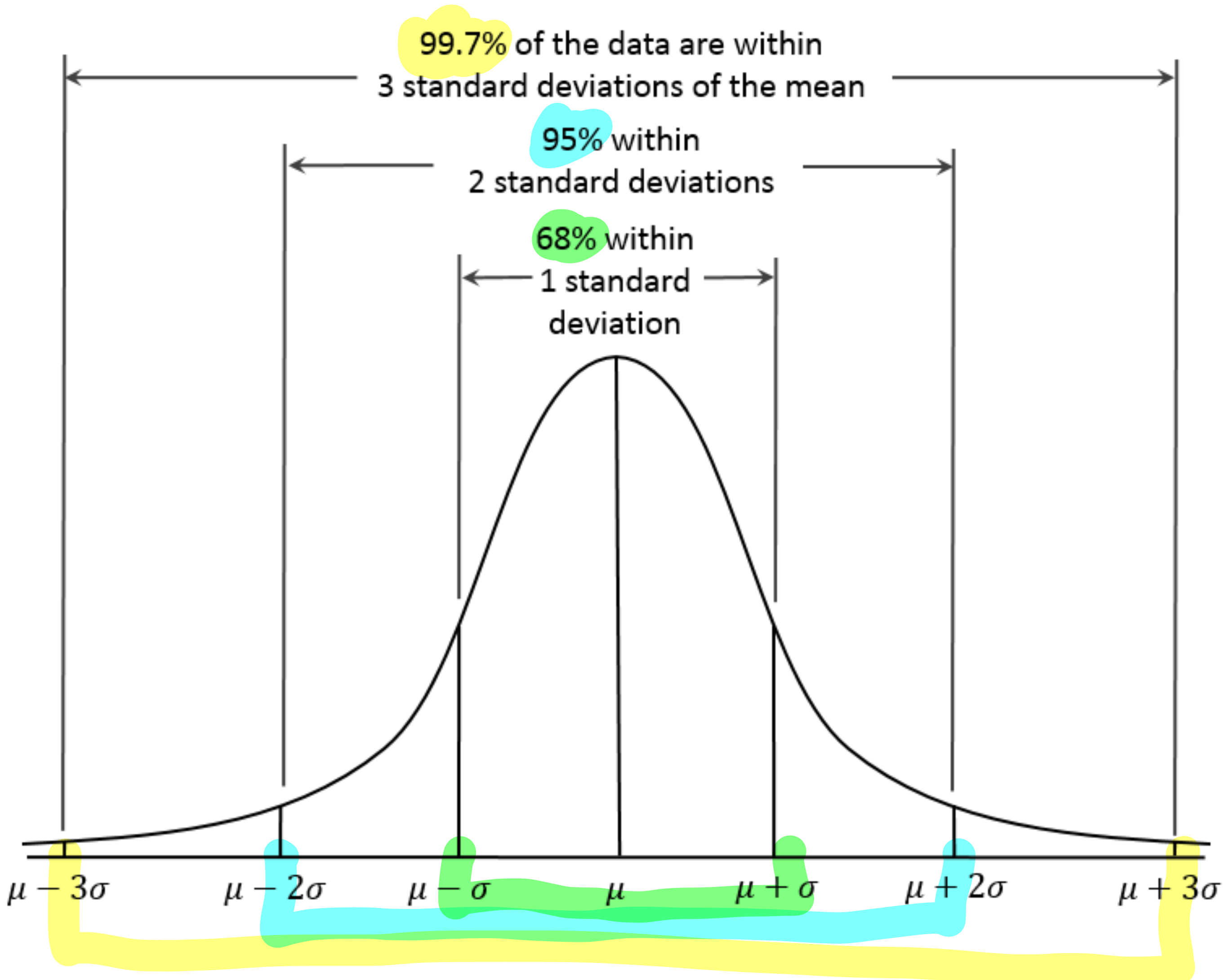
$P(Z - z) = .02$

$\hookrightarrow Z = -2.05 = \frac{X - 64.5}{2.5}$

$\rightarrow x = 59375$



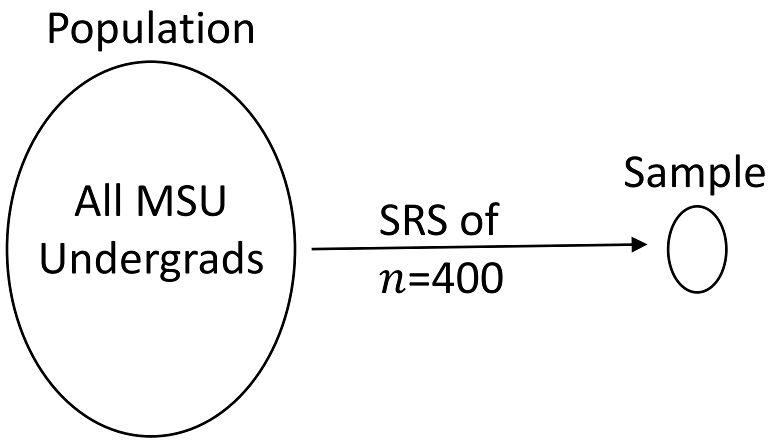
68-95-99.7% Rule for Normal Distributions



CHAPTER 1 – SAMPLING AND EXPERIMENTAL DESIGNS:

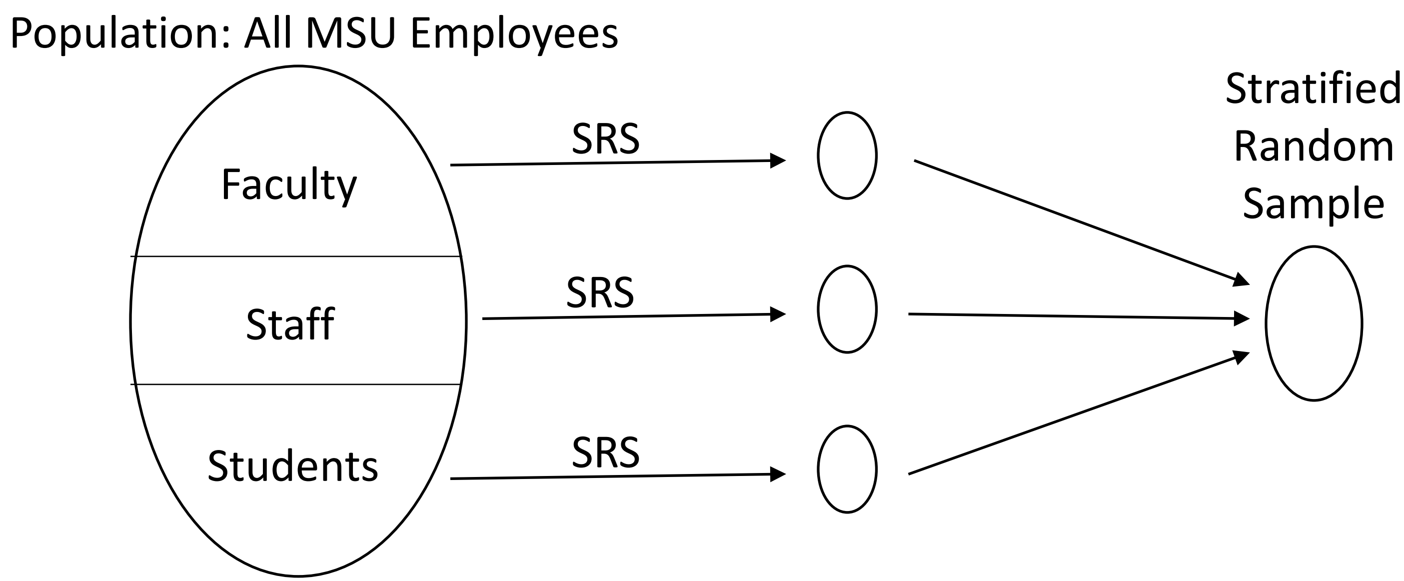
- Simple Random Sample (SRS)

Ex:



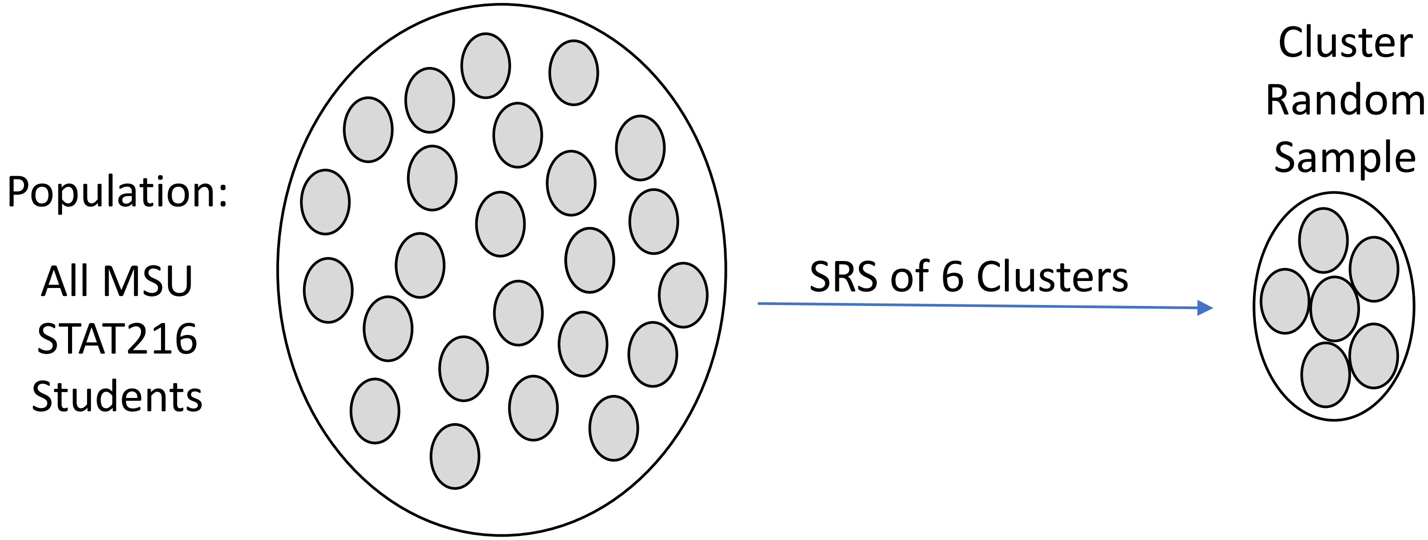
- Stratified Random Sample

Ex:



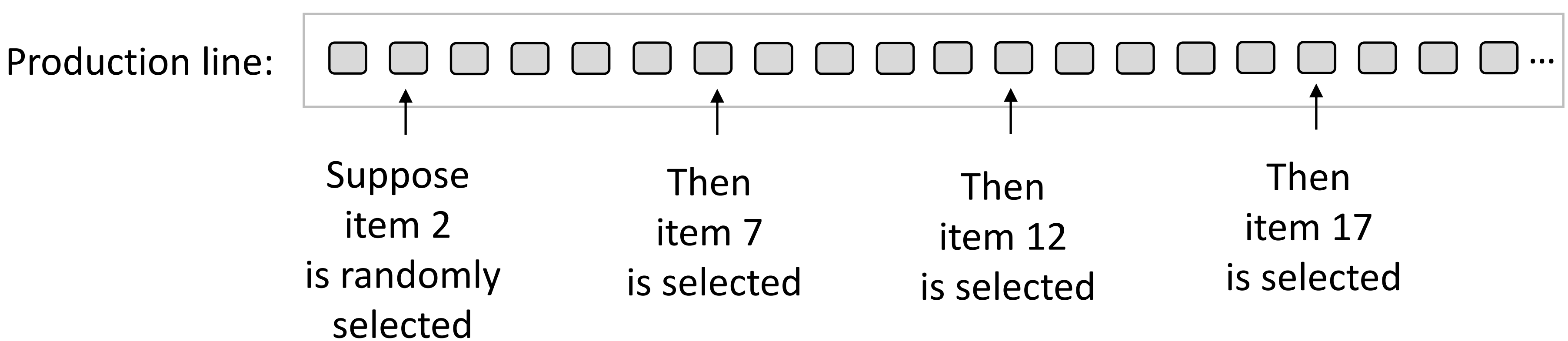
- Cluster Sample

<u>Ex</u>:



- Systematic Sample

<u>Ex</u>: Suppose k=5. Randomly select one of first 5 items, then continue to select every 5^{th} item thereafter …



Scope of Inference:

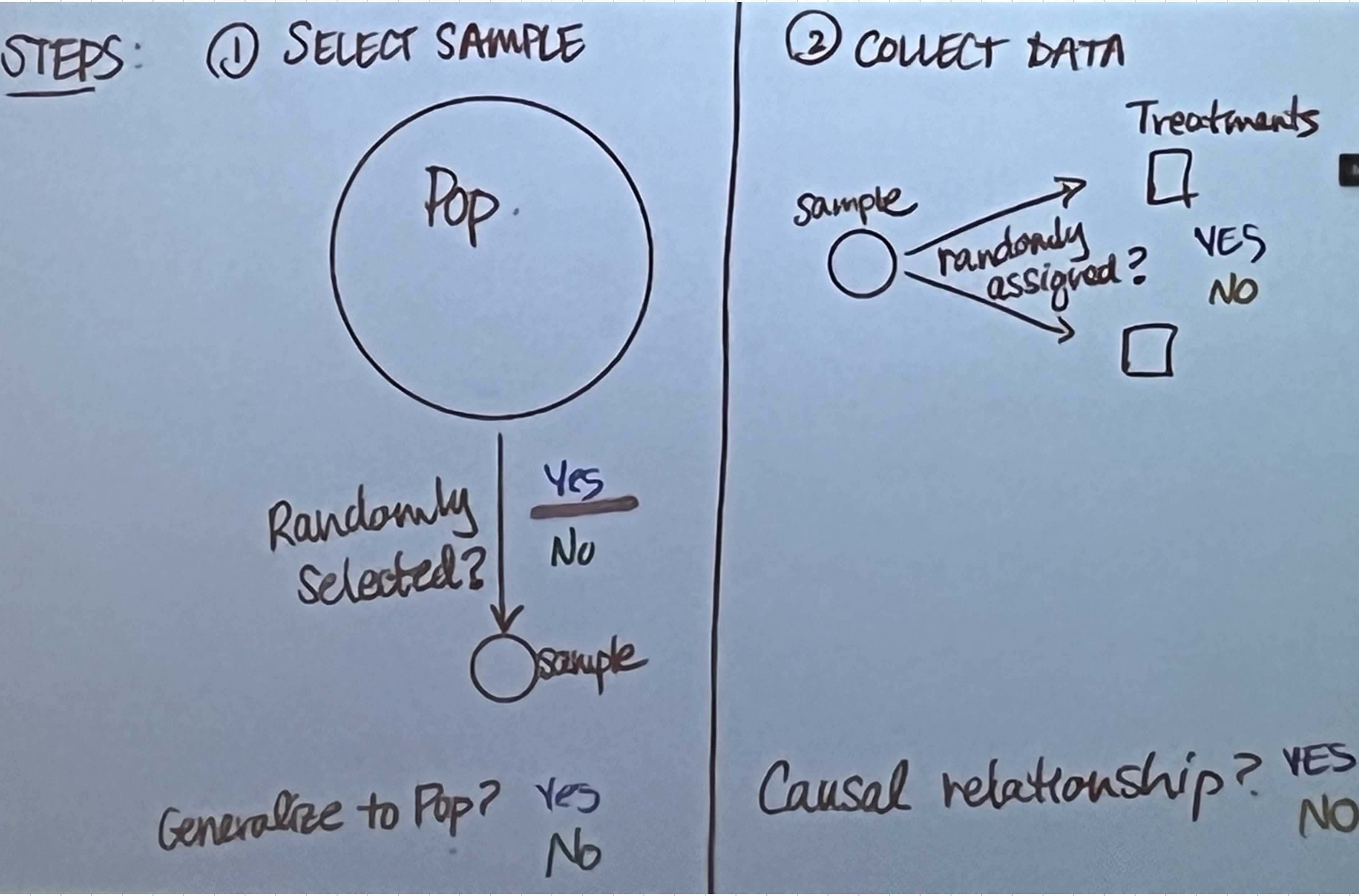
	Study Type	
Type of Sample	Randomized Experiment	Observational Study
Random sample	**Causal relationship, and can generalize results to population.**	**Cannot conclude causal relationship, but can generalize results to population.**
Non-random sample	**Causal relationship, but cannot generalize results to a pop.**	**Cannot conclude causal relationship, and cannot generalize results to a pop.**

Infer to population

Can only infer to the sample (or those similar to the sample)

Can infer causation

Can only infer association due to possible extraneous variables


STEPS:
① SELECT SAMPLE
Pop.
Randomly Selected?
Yes
No
sample
Generalize to Pop?
Yes
No
② COLLECT DATA
Treatments
sample
randomly assigned?
YES
NO
Causal relationship?
YES
NO

CHAPTER 6 – NUMERICAL AND GRAPHICAL SUMMARIES:

What graph(es) display the following?

- Gender vs. Eye Color
- Age vs. Weight
- House Prices
- House prices vs. State
- College Majors

Choices:

Bar chart

Segmented Bar chart

Histogram

Boxplot

Side-by-side boxplots

Scatterplot

What numerical summaries are best for the following?

- House prices (right-skewed)
- Person's height (normal)

Choices:

Mean, SD

Median, IQR

Which are resistant to outliers / skewness?
(Resistant means "not greatly affected by.")

Measures of Center

Mean

Median

Measures of Spread

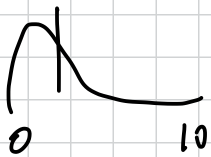
Variance

SD

$\boxed{IQR = Q_3 - Q_1}$

Range = Max – Min

1. False
2. False
3. ~~3.~~ True
4. * 4. true
5. ~~5.~~ True



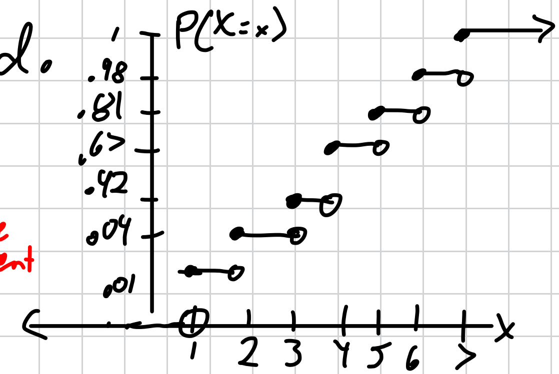
6. false
7. True
8. true
9. True
10. D
11. A
12. ~~12.~~ B
13. C (and b)
14. C
15. A

16.
a. $\boxed{B}$ ______
* b. $\boxed{C}$

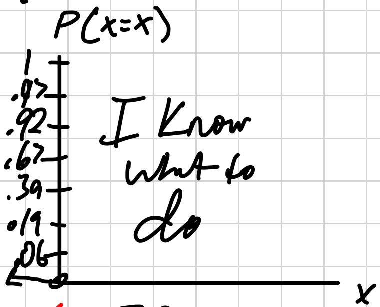
17. B
18. B
19. B
20. B
21. A
22. ~~22.~~ $_{12}C_5 = 792$
order matters because Qs are different
23. $3! = 6$
Consider adding to notes
24. ~~24.~~ $\cancel{_{18}C_2}$ 2^{18} (mult. rule)
25. $\frac{8}{\text{shapes}} \times \frac{4}{\text{sizes}} \times \frac{3}{\text{in}} \times \frac{4}{\text{out}} = 384$
26. $_{45}C_6 = 8145060$

27
a. $_8P_3 = 336$
b. $_8C_3 = 56$

28
a. $P(x=4) = .25$
b. .17
c. .58
d.



29. a.



Consider adding to notes

~~b.~~ $.39 - .19 = .2$

c. $1 - .67 = .33$

~~d.~~ $.58$ $(.97 - \cancel{.34}\ .19)$

E. $.92 - .39 = .53$

f.

X	0	1	2	3	4	5	6
P(X=x)	.06	.13	.2	.28	.25	.05	.03

30. $\binom{12}{4}(.4)^4(.6)^8 = .213$

Consider adding

~~31.~~ $\binom{5}{1}(.6)^1(.4)^4 = .0768$

32. $1 - \int_0^{.8} 1.25(1-x^4) = .08192$

33. a. $\int_3^{\infty} 2x^{-3} = -x^{-2}\Big|_3^{\infty} = 0 + \boxed{1/9}$

b. $\int_1^x 2u^{-3} = -u^{-2}\Big|_1^x \quad \frac{-1}{x^2} + 1$

c. $\frac{-1}{x^2} + 1 = .5 \quad x = \sqrt{2}$

34.

a. B

b. incidence of heart attacks

c. iron levels

d. No, there could be extraneous variables involved/not listed.

35.

a. A

b. bond strength

c. adhesive/curing pressure combos

D. $2 \times 3 \times 4 = 24$ specimens

E. We can say causation because treatments were randomly assigned. There were also limited opportunities for extraneous variables.

$$\begin{cases} 0 & x < 1 \\ \frac{-1}{x^2} + 1 & x \geq 1 \end{cases}$$

36\. a. $\mu = 2.505$ $\sigma = .008$

$$P\left(\frac{2.44-2.505}{.008} \le z \le \frac{2.51-2.505}{.008}\right)$$

$$-1.875 \quad z \le .625$$

$$z(.63) - z(-1.88) = .705599$$

b. $z = 1.64$

$$1.64 = \frac{x-2.505}{.008} \qquad x = 2.51812$$

37\. a. $1 - F(70) = .246$

b. take d/dx

38\. $\mu = .4$ $\sigma = .05$

a. $P(x > .5)$ $\quad 1 - P(X = .5)$ $\quad \frac{.5-.4}{.05} = z = 2$

$$1 - .977250 = \boxed{.02275}$$

b. $P(x = .5) - P(x = .4)$

$$\cancel{x} = 1.28 = \frac{x - .4}{.05} = \boxed{.464}$$

Read Q's very carefully

EGEN350

Chapter 7
Point Estimation of Parameters and Sampling Distributions
(Skip Section 7-4)

Overview of $\boxed{\text{Statistical Inference}}$

* Sample results are infered to pop.

Statistical Inference consists of two major areas:

single value

1) Parameter estimation

Ch7 • $\boxed{\text{Point}}$ estimation (Ex: Use $\bar{X}$ to estimate μ)

Ch8 • Interval estimation (Ex: Form an interval around $\bar{X}$ to estimate μ with high confidence)

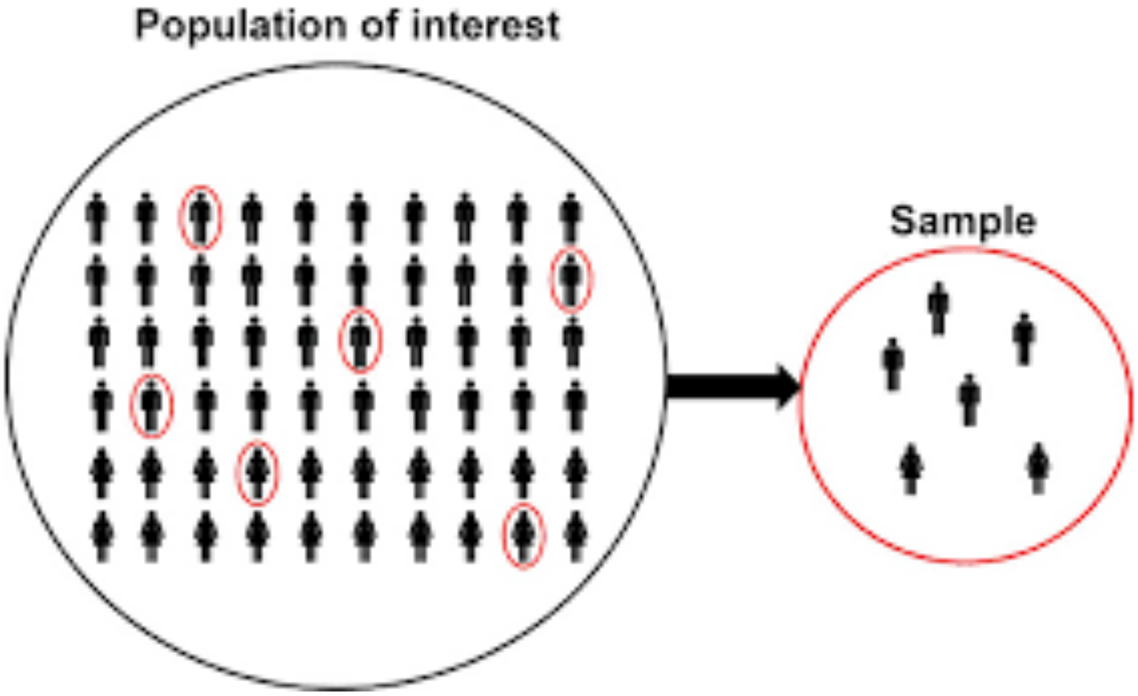
2) Hypothesis testing

Ch9 • Construct hypotheses about μ, collect sample data, and use $\bar{X}$ to draw a conclusion about μ.

Population vs. Sample

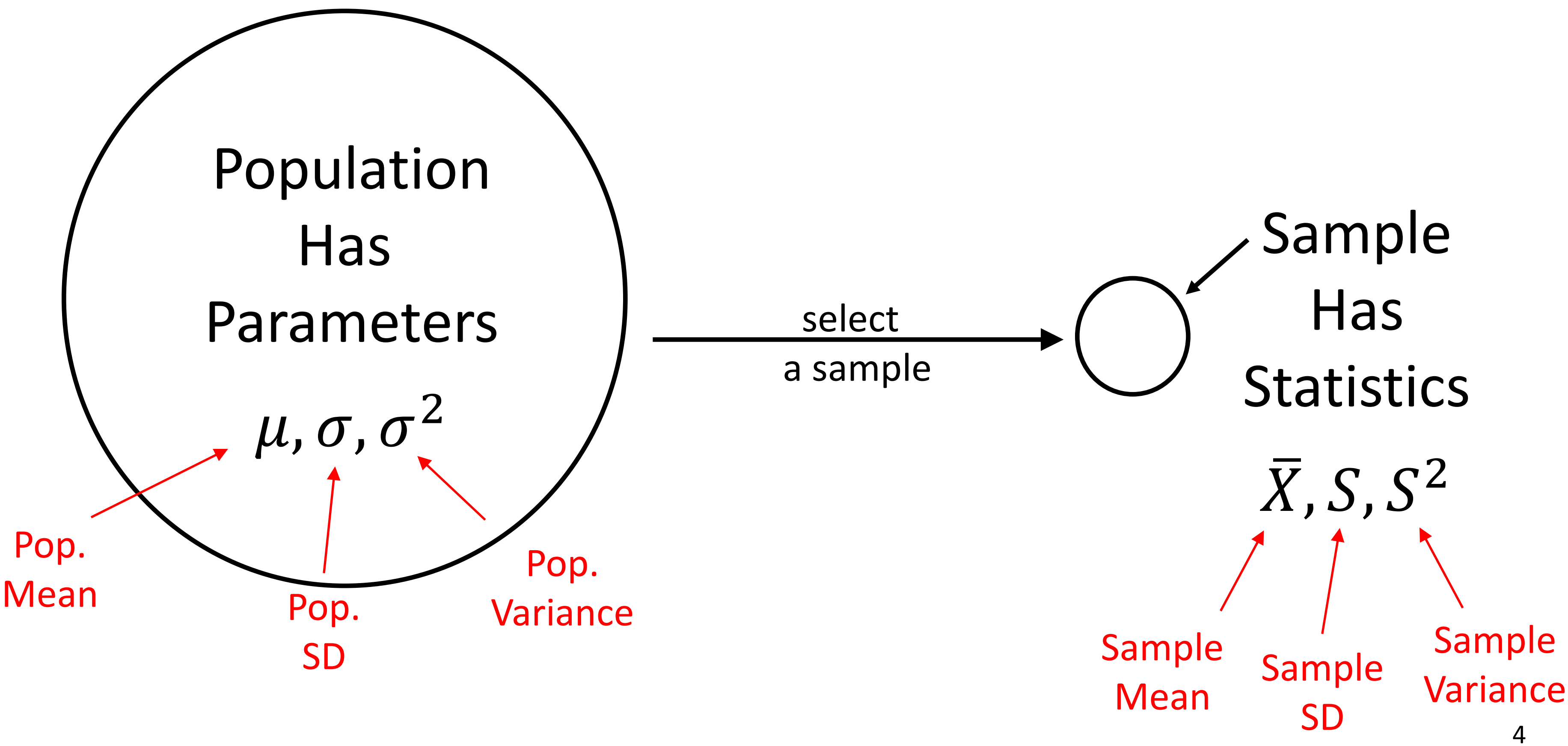
<u>Population</u>: the entire collection of individuals about which information is sought.

<u>Sample</u>: the set of individuals from which data is collected. A sample is a subset of a pop.



Parameter vs. Statistic

- Parameter: a numerical summary of a population
- Statistic: a numerical summary of a sample



Parameter vs. Statistic Notation

(greek) (english)

Parameter	Statistic	
	Random Variable (Estimator)	Observed Value (Estimate)
μ	$\bar{X}$	$\bar{x}$
σ^2	S^2	s^2
σ	S	s

$X_1, X_2, \cdots, X_n$
are the responses
(i.e. individual values)
in the sample!

Formulas:

$$\bar{X} = \frac{X_1+X_2+\cdots+X_n}{n}$$

$$S^2 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}$$

$$S = \sqrt{S^2}$$

Values:

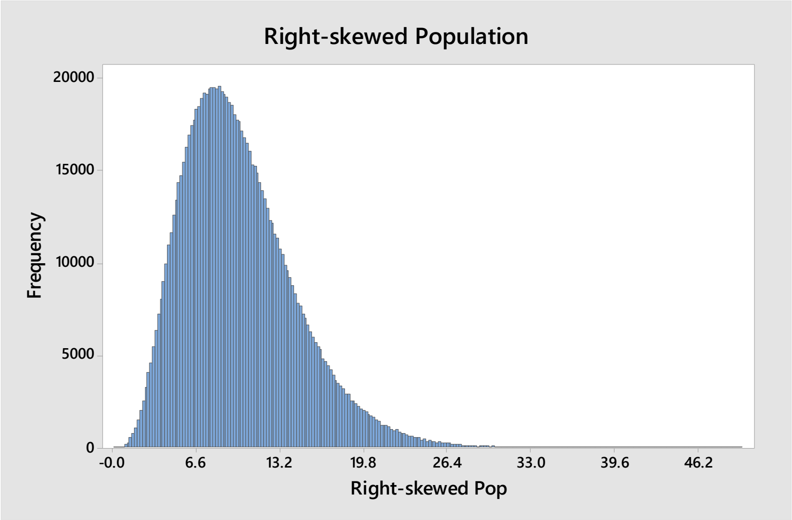
Ex: $\bar{x} = 42$

Ex: $s^2 = 16$

Ex: $s = 4$

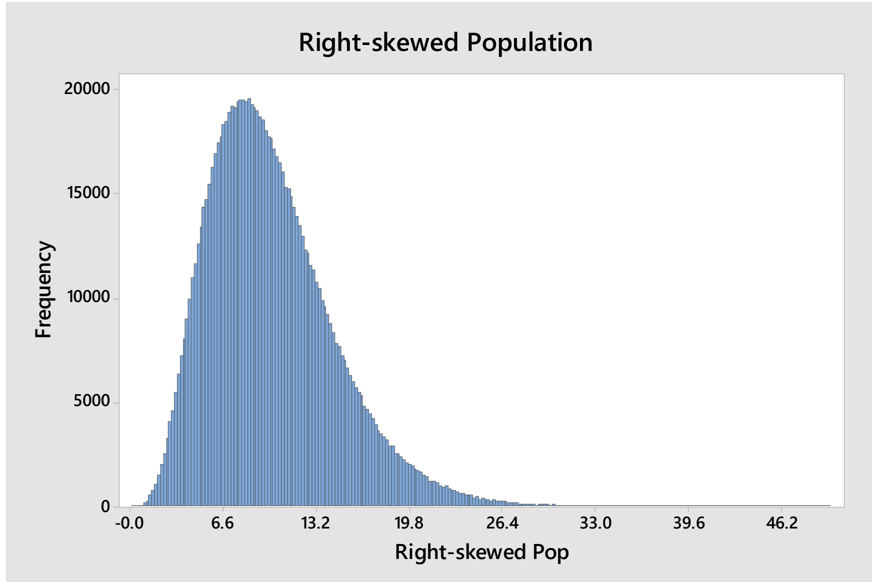
> iClicker Question:

Suppose you select a random sample of size 100 from the following population.

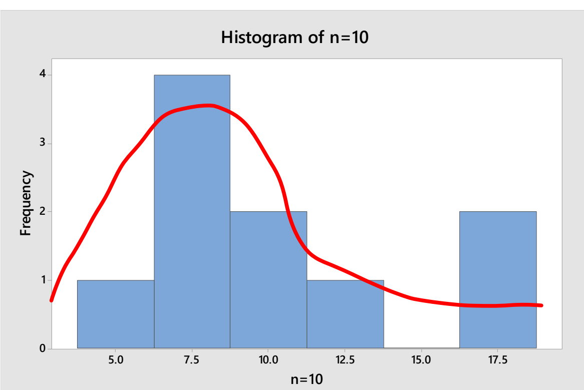


What will be the shape of the histogram of the individual values in the random sample?

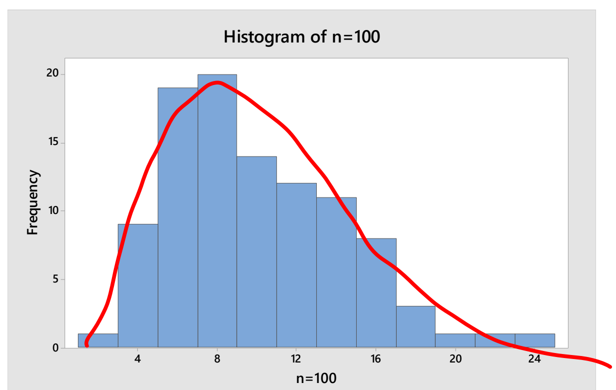
A. Right-skewed
B. Left-skewed
C. Normal



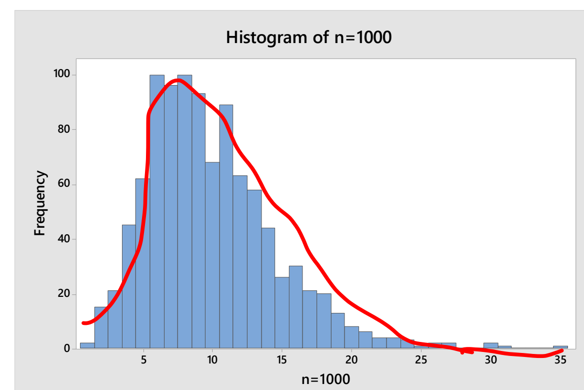
n = 10



n = 100



n = 1000



Sampling Distribution

"Sampling Variability"

Motivation: The value of a statistic varies from sample-to-sample. Thus, a statistic is a variable and has a distribution called a *sampling distribution*.

Sampling Distribution: The distribution of values of the statistic from all possible samples of size n.

Combination

$\binom{n}{p}$ ← pop size ← Check annotate notes

How to Construct a Sampling Distribution:

1. Select all possible samples of size n from the population.
2. Compute the value of the statistic for each sample.
3. Display the distribution of statistic values as a table, graph, or equation.

Name: ______________________

EGEN350 Worksheet 7 (Chapter 7)

Exploring the Sampling Distribution of the Sample Mean

Suppose our population consists of 5 people and our goal is to find the mean height of the population of 5 people.

Person	1	2	3	4	5
Height (inches)	64.5	69	72	70.5	67

Population Mean is $\mu =$ 68.6

In real-life situations, are we usually able to calculate the population mean or other parameter values? No

Imagine we could not calculate the population mean, so instead we use a sample mean (i.e., a statistic) to *estimate* the population mean (i.e., a parameter).

$\binom{5}{2} = 10$

Samples of Size n=2

Samples of size n=2	Heights	Sample Mean $\bar{x}$
1, 2	64.5, 69	66.75
1, 3	64.5, 72	68.25
1, 4	64.5, 70.5	67.5
1, 5	64.5, 67	65.75
2, 3	69, 72	70.5
2, 4	69, 70.5	69.75
2, 5	69, 67	68
3, 4	72, 70.5	71.25
3, 5	72, 67	69.5
4, 5	70.5, 67	68.75

Mean of the $\bar{x}$ values is $\mu_{\bar{x}} =$ 68.6

Sampling variability is the variability we expect to see from one random sample to another.

Did you notice the sample means were different from one another? Yes

$\binom{5}{4} = 5$

Name: ______________________________

Samples of Size n=4

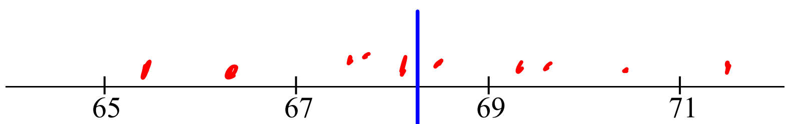
Samples of size n=4	Heights	Sample Mean $\bar{x}$
1, 2, 3, 4	64.5, 69, 72, 70.5	69
1, 2, 3, 5	64.5, 69, 72, 67	68.125
1, 2, 4, 5	64.5, 69, 70.5, 67	67.75
1, 3, 4, 5	64.5, 72, 70.5, 67	68.5
2, 3, 4, 5	69, 72, 70.5, 67	69.625

Mean of the $\bar{x}$ values is $\mu_{\bar{X}} =$ 68.6

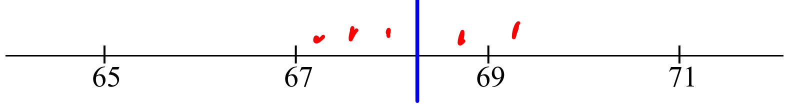
What was $\mu_{\bar{X}}$ when n=2? 68.6 What was $\mu_{\bar{X}}$ when n=4? 68.6 What is μ? 68.6 This is not a coincidence!

How did sample size affect sampling variability? Draw a dot above the axis for each $\bar{x}$.

Dotplot of the $\bar{x}$ values (sample size n=2):



Dotplot of the $\bar{x}$ values (sample size n=4):



Which one is more spread out? $\bar{x}$ values when n=2? Or when n=4?

Mark the value of μ on the dotplots. Were the values of $\bar{x}$ clustered more closely around μ for n=2 or n=4?

Important Observations:

$(\bar{x})$

1. The sampling distribution of the sample mean is always centered at the value of the population mean (μ).
2. Increasing the sample size will decrease the spread of the sampling distribution of the sample mean.

For a random variable X and a given sample size, the distribution of the random variable $\bar{X}$ is called a **sampling distribution of the sample mean**. In other words, it is the distribution of all possible sample means for samples of a given size.

See Worksheet Example: Sampling Dist.

Population of 5 people

Person	1	2	3	4	5
Height (inches)	64.5	69	72	70.5	67

Samples of Size n=2

Samples of size n=2	Heights	Sample Mean $\bar{x}$
1, 2	64.5, 69	
1, 3	64.5, 72	
1, 4	64.5, 70.5	
1, 5	64.5, 67	
2, 3	69, 72	
2, 4	69, 70.5	
2, 5	69, 67	
3, 4	72, 70.5	
3, 5	72, 67	
4, 5	70.5, 67	

$\mu_{\bar{X}}$, $\sigma_{\bar{X}}^2$, and $\sigma_{\bar{X}}$

Facts: $\mu_{\bar{X}} = \mu_X$ $\sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n}$ $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$

These facts are true for any pop. dist shape and any sample size!

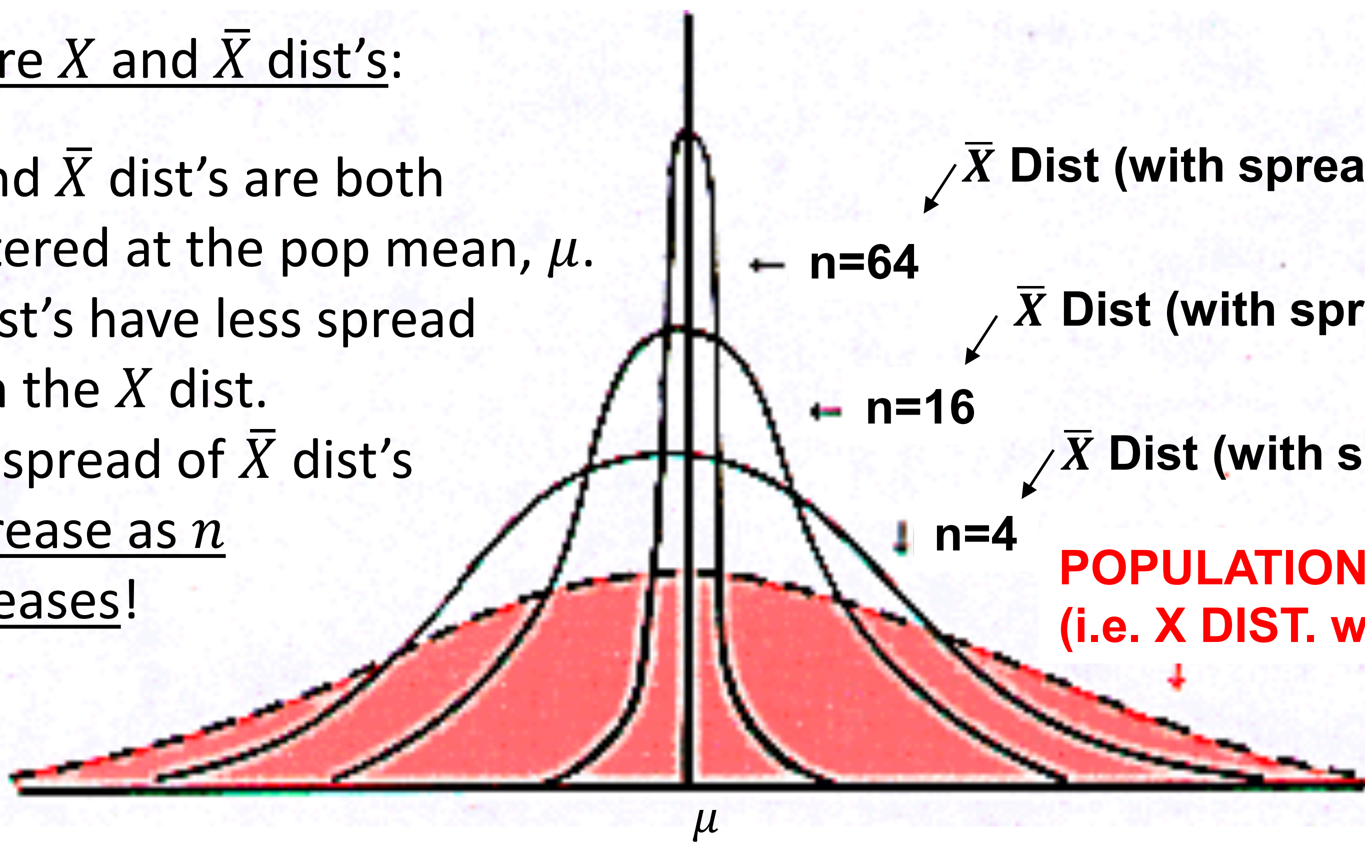
Normal Distribution Fact

Pop. Dist. Sampling Dist. of $\bar{X}$

If $X \sim N(\mu, \sigma^2)$, then $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ for any sample size n.

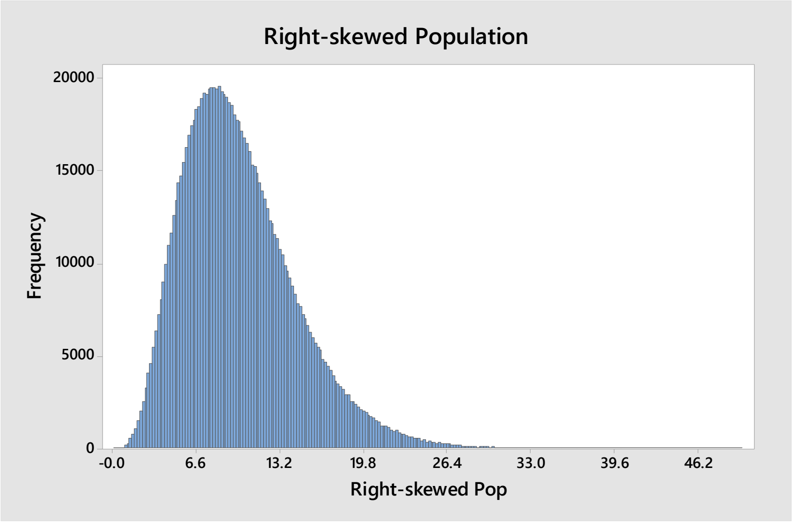
Compare X and $\bar{X}$ dist's:

- X and $\bar{X}$ dist's are both centered at the pop mean, μ.
- $\bar{X}$ dist's have less spread than the X dist.
- The spread of $\bar{X}$ dist's decrease as n increases!



> iClicker Question:

Suppose you select all possible random samples of size 40 from the following population.



The sample mean ($\bar{X}$) is calculated from all random samples. What is the shape of the distribution of the <u>sample means, $\bar{X}$'s</u>?

A. Right-skewed
B. Left-skewed
C. Normal

Question: But what if X does not have a normal dist?

Answer: The Central Limit Theorem *saves the day*!

Central Limit Theorem (CLT) for $\bar{X}$

If $X_1, \ldots, X_n$ is a random sample* from a pop. with mean μ and known variance σ^2, then

$$\bar{X} \dot{\sim} \mathrm{N}(\mu, \frac{\sigma^2}{n}) \text{ for large } n \; (n \geq 30^{**}).$$

approximately distributed

* The random variables $X_1, \ldots X_n$ are a random sample of size n if (a) the X_i's are independent random variables, and (b) every X_i has the same probability dist.

** If the variance σ^2 is unknown, then S^2 is used to estimate σ^2. Thus, a larger sample size ($n \geq 40$) is required to invoke the CLT.

Illustration of CLT, part 1

		Die 2					
		1	2	3	4	5	6
Die 1	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

"normalizing effect"

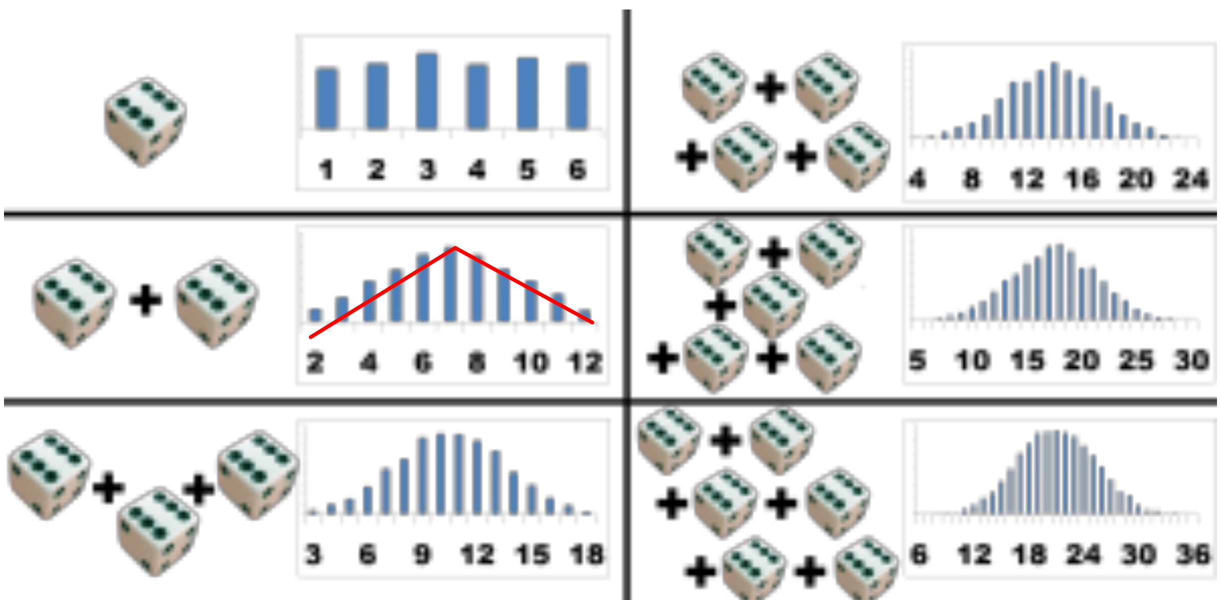


Illustration of CLT, part 2

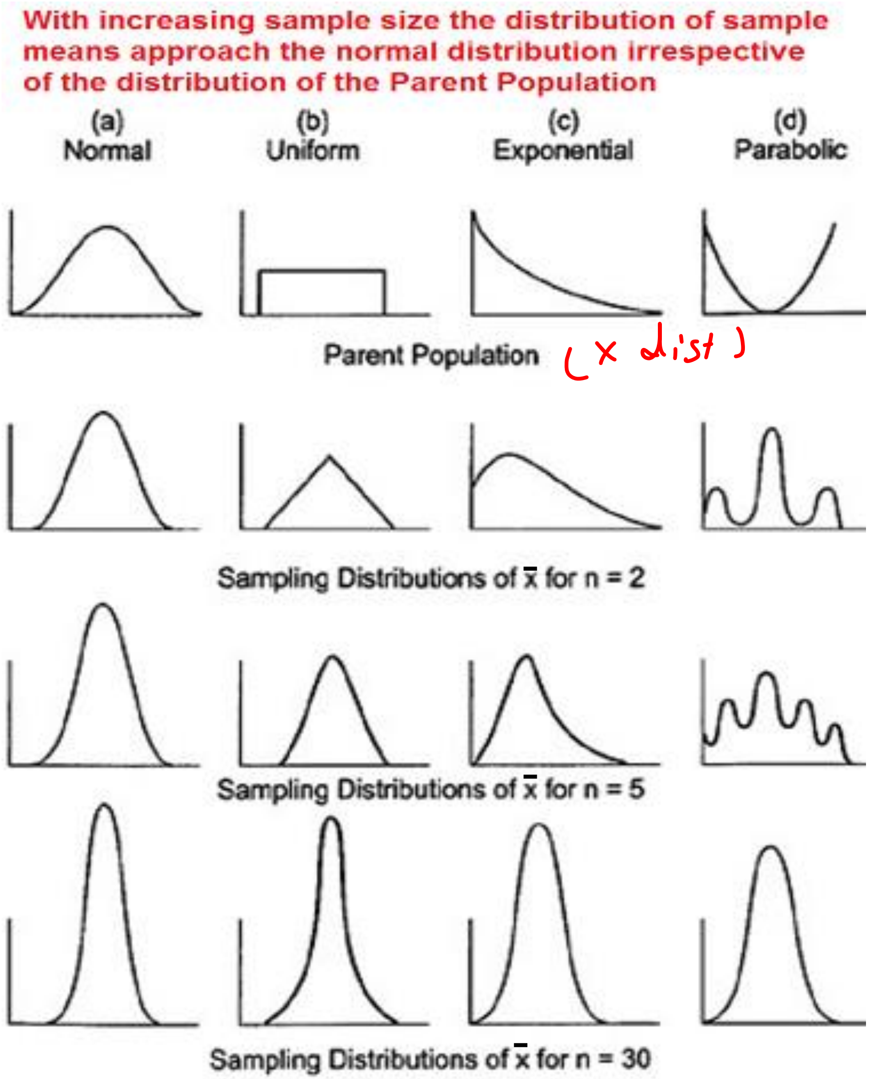


Illustration of CLT, part 3

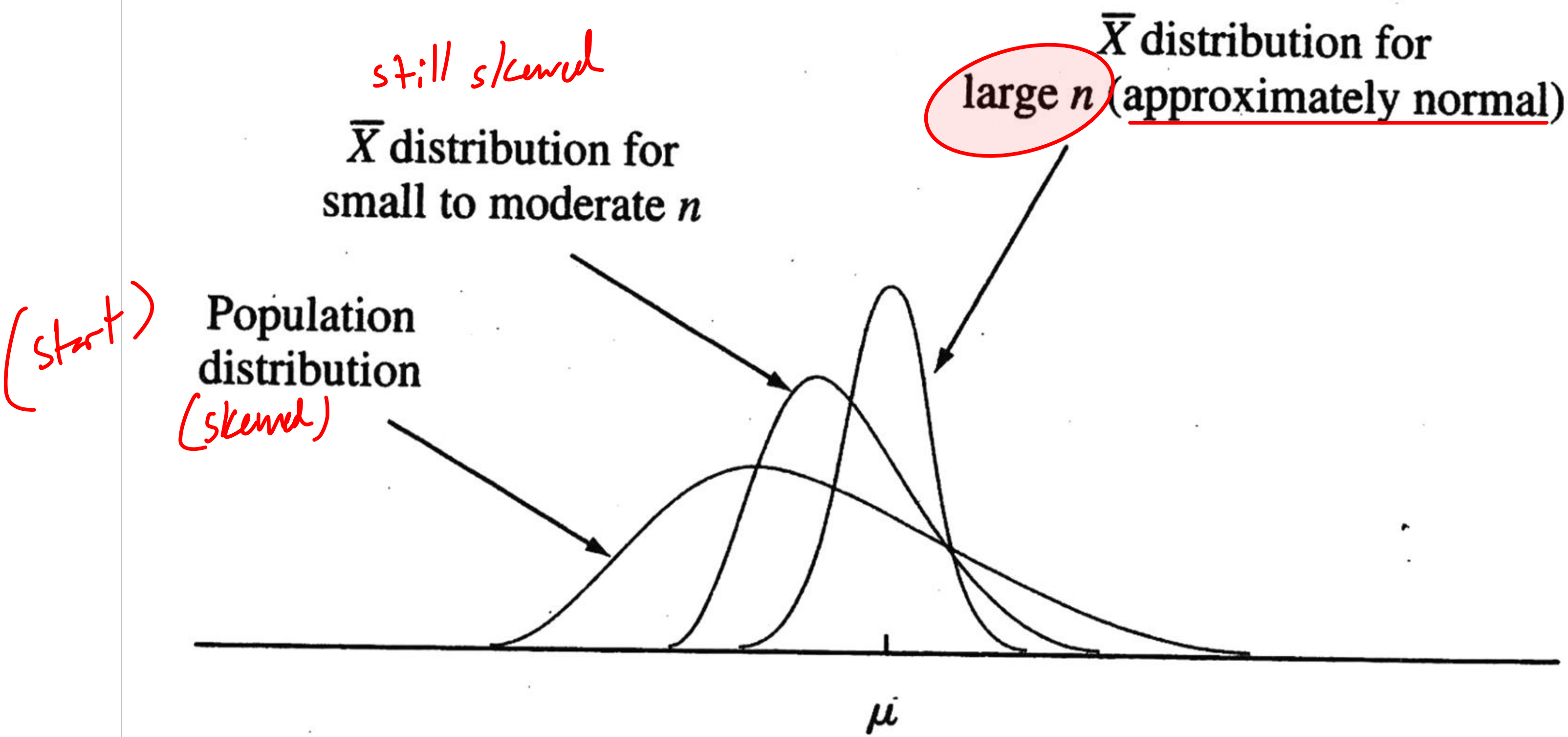


Figure 5.15 The Central Limit Theorem illustrated

How to Standardize $\bar{X}$

<u>Facts</u>: $\mu_{\bar{X}} = \mu_X$ $\sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n}$ $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$

<u>Recall from Ch4</u>: Standardize a normal X:

$$Z = \frac{X - \mu_X}{\sigma_X}$$

<u>Now</u>: Standardize a normal $\bar{X}$:

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \longrightarrow \boxed{Z = \frac{\bar{X} - \mu_X}{\sigma_X/\sqrt{n}}}$$

<u>Note</u>: $\sigma_X = \sigma$

- If $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$, then $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ where $Z \sim N(0,1)$.

(Pop. dist.)

X Dist

Normal

$X \sim N(\mu, \sigma^2)$

Normal distribution fact

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

Not Normal (or Unknown)

$n < 30$

Check the Normal Prob Plot

If sample is close to normal, then

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

$n \geq 30$

C.L.T.

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

Standard Deviation vs. Standard Error

- The standard error of a statistic is the *estimated* standard deviation of the statistic.

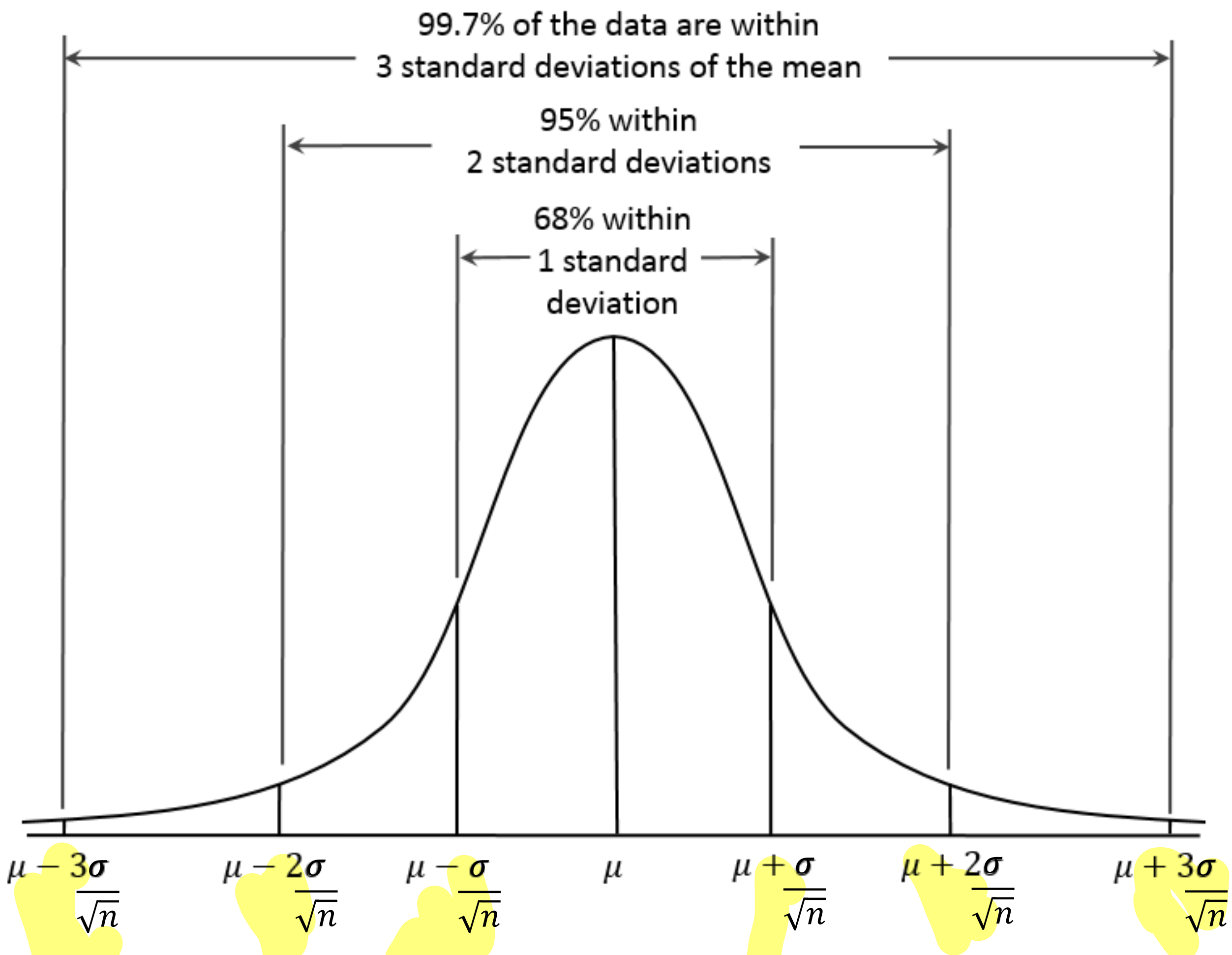
- Standard deviation of $\bar{X}$: $SD(\bar{X}) = \sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}}$

- Standard error of $\bar{X}$: $SE(\bar{X}) = \dfrac{S}{\sqrt{n}}$

Plug in sample S.D. (s) in place of σ

- Question: If we don't know pop. mean μ, then how could we know the population standard deviation (SD) σ?
- Answer: Oftentimes we don't! That's why we use S to estimate σ! Other times, we may have prior data indicating the value of the pop. SD but there has been a process shift, so we know σ but not μ.

Normal $\bar{X}$ dist: 68 – 95 – 99.7 Rule



- 68% of the $\bar{X}$ values lie within 1 $\frac{\sigma}{\sqrt{n}}$ of μ.
- 95% of the $\bar{X}$ values lie within 2 $\frac{\sigma}{\sqrt{n}}$ of μ.
- 99.7% of the $\bar{X}$ values lie within 3 $\frac{\sigma}{\sqrt{n}}$ of μ.

Point estimates are good, BUT confidence intervals are better!

- A point estimate is our best guess of the parameter value, BUT the point estimate is *almost never* exactly equal to the parameter value.

- Alternatively, we could create an interval of values around the point estimate, in which we are highly confident our interval contains the parameter value. This is a <u>confidence interval</u>!

- The confidence interval incorporates a measure of variability in our point estimate.

Summary

- Parameter: a numerical summary of a population
- Statistic: a numerical summary of a sample
- Point estimation – a "single value" estimate of a parameter (i.e. a statistic is used to estimate a parameter)
- The value of a statistic varies from sample-to-sample. Thus, a statistic has a sampling distribution.
- Sampling Distribution: The distribution of values of the statistic from all possible samples of size n.
- The sample mean ($\bar{X}$) is used as the point estimate for the pop. mean (μ).
- Facts: $\mu_{\bar{X}} = \mu_X$ $\quad \sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n}$ $\quad \sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$
- Normal Distribution Fact: If $X \sim N(\mu, \sigma^2)$, then $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ for any sample size n.
- X and $\bar{X}$ distributions are centered at the same value, μ.
- $\bar{X}$ distributions have less spread (smaller variance and SD) than the X dist.
- The variance and SD of $\bar{X}$ distributions decrease as n increases!

Summary

- Central Limit Theorem: If $X_1, \ldots, X_n$ is a random sample from a pop. with mean μ and known variance σ^2, then $\bar{X} \dot{\sim}$ N(μ , $\frac{\sigma^2}{n}$) for large n ($n \geq 30$).
- To standardize the Normal RV ($\bar{X}$) to the Standard Normal RV (Z), use $Z = \frac{\bar{X} - \mu_X}{\sigma_X / \sqrt{n}}$.
- The standard error of a statistic is the estimated standard deviation of the statistic.
- Standard deviation of $\bar{X}$: $SD(\bar{X}) = \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$.
- Standard error of $\bar{X}$: $SE(\bar{X}) = \frac{s}{\sqrt{n}}$
- 68%, 95%, and 99.7% of the values of an $\bar{X}$ normal distribution lie within 1, 2, and 3 $\frac{\sigma_X}{\sqrt{n}}$'s of μ, respectively.
- Point estimates are good, BUT confidence intervals are better! The confidence interval incorporates a measure of variability in the point estimate.

EGEN350

Chapter 9

Hypothesis Testing for the Pop. Mean (μ)

(Skip Sections 9-4 through 9-11)

Overview of Statistical Inference

Statistical Inference consists of two major areas:

single value

1) Parameter estimation

Ch7 • Point estimation (Ex: Use $\bar{X}$ to estimate μ)

Ch8 • Interval estimation (Ex: Form an interval around $\bar{X}$ to estimate μ with high confidence)

$\bar{X} \pm Z_{\alpha/2} \cdot \boxed{\frac{\sigma}{\sqrt{n}}}$ intervals incorporate the variability of $\bar{X}$'s.

Ch9 2) Hypothesis testing

• Construct hypotheses about μ, collect sample data, and use $\bar{X}$ to draw a conclusion about μ.

Purpose of CI vs HT

- The <u>purpose of a confidence interval</u> (CI) is to <u>ESTIMATE</u> the value of a parameter.
  (μ)

- The <u>purpose of a hypothesis test</u> (HT) is to <u>DECIDE</u> between two competing claims (i.e. hypotheses) about the value of a parameter.
  (μ)

The Scientific Method

1. Observe some phenomenon
2. State a hypothesis explaining the phenomenon
3. Collect data
4. Analyze the data
5. Test: Do the data support the hypothesis?
6. Conclusion: If the test fails, go back to step 2.

*Statistical procedures are part of the Scientific Method!

PREVIEW

Steps in Hypothesis Testing

1. State the null hypothesis (H_0) and alternative hypothesis (H_a).
2. Check the necessary assumptions so the test is valid.
3. Compute a test statistic. (Z or T)
   The test statistic is a statistic that is used to assess the strength of evidence against H_0.
4. Compute the p-value.
   The p-value is the probability of obtaining a test statistic at least as extreme as the observed test statistic, assuming H_0 is true.
5. State the decision about H_0.
6. State the conclusion about the strength of evidence for H_a.

The Logic of Hypothesis Testing

How do we decide whether to reject H_0?

- If the sample data are consistent with H_0, then we do not reject H_0.

- If the sample data are inconsistent with H_0, but consistent with H_a, then we reject H_0 and conclude H_a.

Motivational Example

* Initially assume H_o is true

You are asked to decide if a coin is fair or not fair.

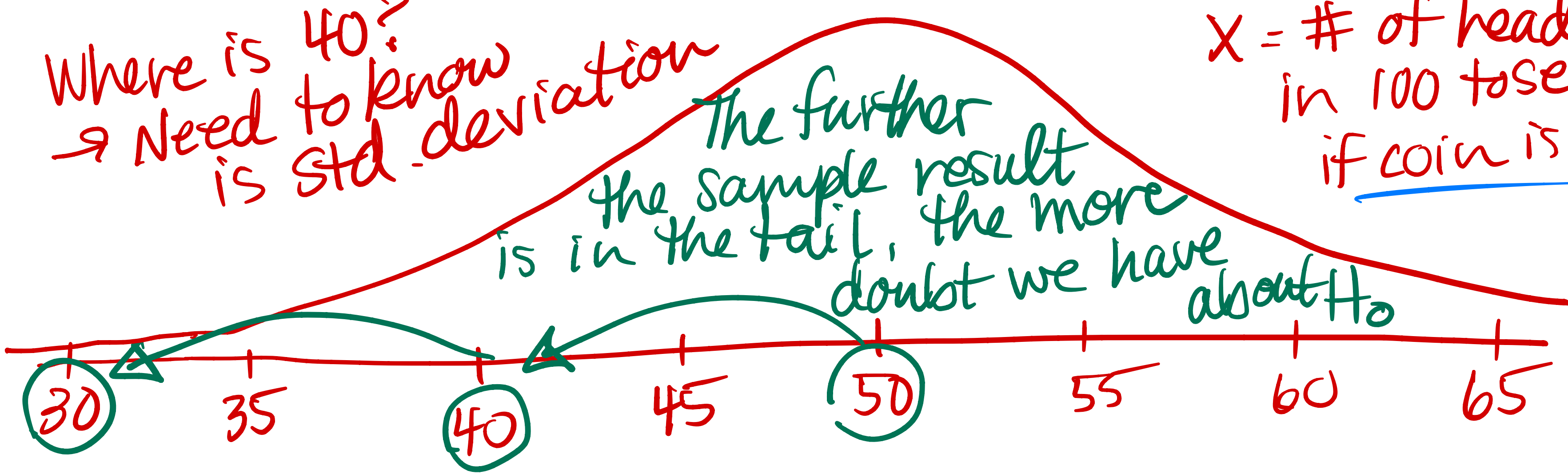
- What are the two competing claims?
  H_o: Coin is fair.
  H_a: Coin is not fair.

- We toss the coin 100 times. Do you think the coin is fair,
  - ✓ if you observe 50 heads? Yes
  - ✓ if you observe 40 heads? Maybe, due to sampling variability
  - ✓ if you observe 30 heads? NO

Where is 40?
→ Need to know is std. deviation

X = # of heads in 100 tosses, if coin is fair

The further the sample result is in the tail, the more doubt we have about H_o



> iClicker Question:

Suppose the distribution of possible sample statistics is drawn *assuming H_0 is true*. The probability of obtaining our observed sample result or a sample result even more extreme (i.e. further out into the tail of the distribution) is calculated.

Do you think a small probability or a large probability would be evidence against H_0?

A. Small probability
B. Large probability

Hypotheses

(μ)

<u>Hypotheses</u> - statements about the value of a parameter

said "H naught"

- Null Hypothesis (H_0) – Statement of no difference.

zero

- Alternative Hypothesis (H_a) – Statement we want to find evidence for.

<u>IMPORTANT</u>: Never place statistics in hypotheses.

$\cancel{H_0 : \bar{X} = 5.138}$ NO!

- Statistics are computed from sample data; there is no need to hypothesize about their value(s).

Hypothesis Testing for μ

Choose one set of hypotheses where μ_0 is a specific value.

number

H_0: $\mu \overset{=}{\cancel{\geq}} \mu_0$	H_0: $\mu \overset{=}{\cancel{\leq}} \mu_0$	H_0: $\mu = \mu_0$
H_a: $\mu < \mu_0$	H_a: $\mu > \mu_0$	H_a: $\mu \neq \mu_0$

H_a μ_0 H_0

- ✓ The two hypotheses (H_0 and H_a) cover all possibilities for μ.
- ✓ <u>BUT</u>, when H_0 is a range of values, we will use the nearest competing value to H_a, which is $\mu = \mu_0$.
- ✓ Also, note, we will initially assume H_0 is true and we need one value to use in the calculations.

$= \mu_0$

Example: Emissions

OLD PROCESS:

$X \sim N(1.45, \sigma^2)$

The Environmental Protection Agency (EPA) reports that, under the old process, the exhaust emissions for a certain car model had a normal distribution with a mean μ=1.45 grams of nitrous oxide per mile.

$X \sim N(\mu, \sigma^2)$

The car manufacturer claims their *new* process reduces the mean level of exhaust emitted for this car model but did not affect the standard deviation. A simple random sample of 28 cars produced using the new process is selected and the mean and standard deviation for this sample is $\bar{x}$=1.21 and s=0.4 grams of nitrous oxide per mile. Use α=0.05.

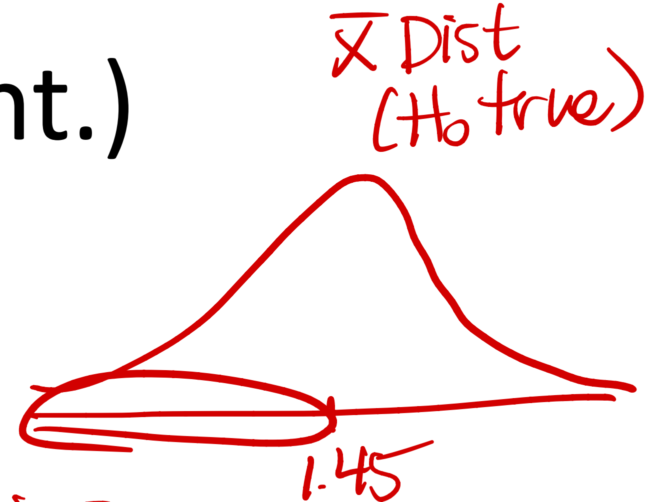
μ = true mean exhaust emission under the new process

State the null and alternative hypotheses:

$H_0: \mu = 1.45$

$H_a: \mu < 1.45$

Example: Emissions (cont.)



Old process true mean: μ_0 = 1.45

New process sample mean: $\bar{x}$ = 1.21

New process true mean: μ = ???

where does 1.21 lie?

$\bar{x}$ = 1.21 is smaller than μ = 1.45. But is it **MUCH** smaller?

QUESTION:

Ho
- Is it due to chance variation? (We know different samples selected from a population have different means!)

or

Ha
- Did the *new* process truly decrease the mean exhaust emissions?

What do you need to know to decide?

Test Statistic

A test statistic measures the difference between the observed data and what is expected if H_0 were true.

If σ is known and $\bar{X} \sim N$, the test statistic is:

$$Z_0 = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \quad \text{where } Z_0 \sim N(0, 1)$$

If σ is unknown and $\bar{X} \sim N$, the test statistic is:

Completed assuming Ho is true

$$T_0 = \frac{\bar{X} - \mu_0}{s / \sqrt{n}} \quad \text{where } T_0 \sim t(\text{n-1})$$

<u>Interpretation</u>:
Z_0 is the number of SD's ($\frac{\sigma}{\sqrt{n}}$'s) $\bar{X}$ is above/below μ

<u>Interpretation</u>:
T_0 is the number of SE's ($\frac{s}{\sqrt{n}}$'s) $\bar{X}$ is above/below μ

<u>Assumptions</u>:

1. The data must be a random sample from the population.
2. The $\bar{X}$ distribution must be at least approximately normal.

Example: Emissions (cont.)

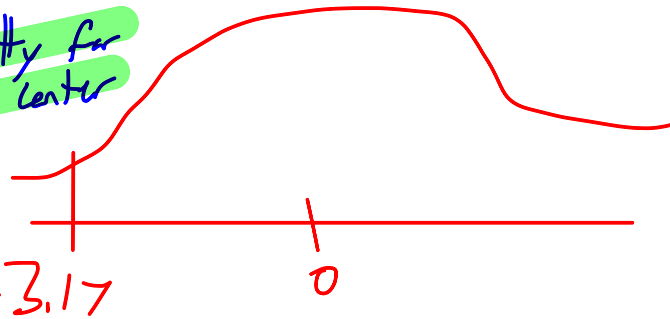
- Check the assumptions are satisfied.

1. Random Sample
2. $\bar{X} \sim N$ b/c $X \sim N$

} Both stated in description

- Calculate the value of test statistic.

Pretty far from center



Formula: $T_0 = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$

$$t_0 = \frac{1.21 - 1.45}{.4/\sqrt{28}} = -3.17$$

- Recall, the T value is the number of standard errors $\bar{X}$ is above/below μ.

(If T is far away from 0, then it is unlikely that the H_0 is true and we would reject H_0. Otherwise, we cannot reject H_0.)

Ha gives the direction of extreme

p-value

p-value: the probability of obtaining a test statistic at least as extreme as the observed test statistic value, assuming the H_0 is true.

from our sample

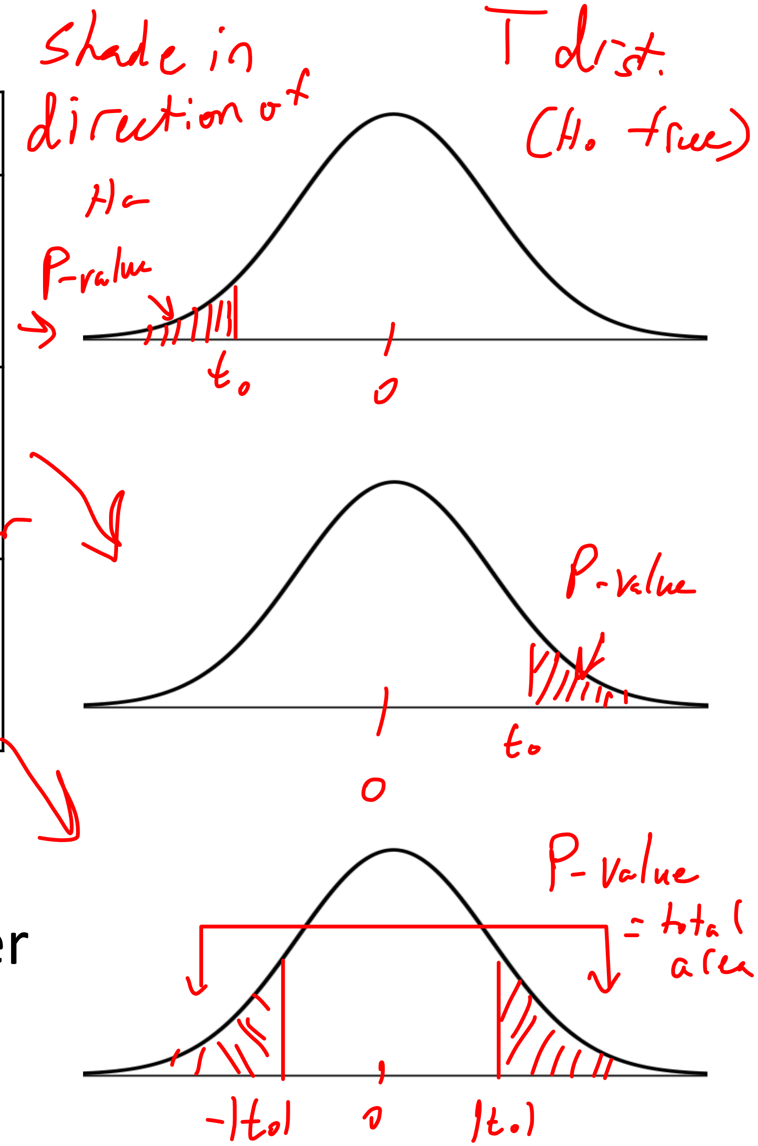
- Small p-value
  - Data are *inconsistent* with the assumption of a true H_0
  - "Strong enough" evidence against H_0 and for H_a

- Large p-value
  - Data are *consistent* with the assumption of a true H_0
  - "Not enough" evidence against H_0

IMPORTANT: The p-value is **NOT** the probability that H_0 is true! The H_0 is either true or false; the truth is just unknown to us!

H_a Gives the Direction of Extreme

Alternative Hypothesis	p-value
H_a: $\mu < \mu_0$ Lower Tail	$P(T_0 < t_0)$
H_a: $\mu > \mu_0$ upper tail	$P(T_0 > t_0)$ variable number
H_a: $\mu \neq \mu_0$	$2 \cdot P(T_0 > \|t_0\|)$



- The smaller the p-value, the stronger the evidence against H_0 and for H_a.
- BUT how small is small enough?

Example: Emissions (cont.)

Old Process Exhaust Emissions: $X \sim N(1.45, \sigma^2)$

$H_0: \mu = 1.45$ vs. $H_a: \mu < 1.45$

use direction in H_a

n=28, $\bar{x}$=1.21, s=0.4

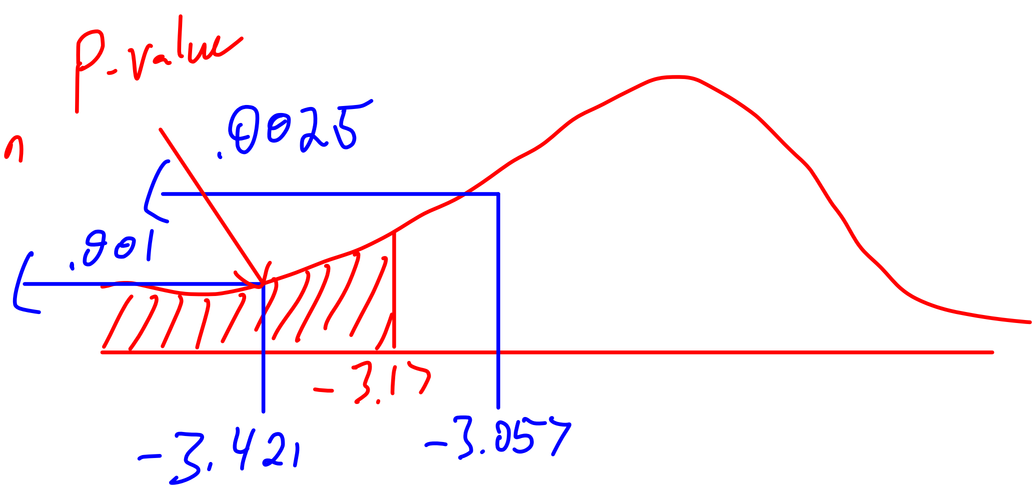
Test statistic: $\underbrace{t_0 = -3.17}_{\text{observed}}$

$P(T_0 < -3.17)$

- Find the p-value=

If Ho were true, there is a .1 to .25% chance of obtaining a sample mean of $\bar{x}$=1.21 or less.

Extremely unlikely



Name: ______________________________

EGEN350 Worksheet 11 (Chapter 9)

Directions: On problems 1 and 2, first decide if a Z_0 or T_0 test statistic is appropriate, calculate it, then find the p-value in the appropriate direction (see direction in H_a).

Alternative Hypothesis	p-value
H_a: $\mu < \mu_0$	$P(Z_0 < z_0)$ $P(T_0 < t_0)$
H_a: $\mu > \mu_0$	$P(Z_0 > z_0)$ $P(T_0 > t_0)$
H_a: $\mu \neq \mu_0$	$2 \cdot P(Z_0 > \|z_0\|)$ $2 \cdot P(T_0 > \|t_0\|)$

Pop. SD σ	Test Statistic
σ is known	$Z_0 = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$
(U	μ_0 ... $\sqrt{n}$

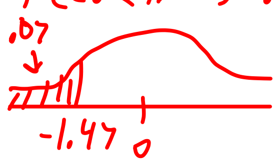
$$T_0 = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

1. ated for a certain type of car. The braking system will be installed if it can be ean stopping distance under certain controlled conditions is less than 90 ft. It is at the standard deviation of stopping distance is $\sigma = 5$ ft. A random sample of 150 stops was selected and the resulting mean stopping distance was 89.4 ft. Use the significance level $\alpha = 0.01$.
Test the hypotheses:
$H_0: \mu = 90$
$H_a: \mu < 90$
Compute test statistic:

$Z_0 = \dfrac{89.4 - 90}{5/\sqrt{150}} = -1.47$

Find the p-value:

$P(Z_0 < -1.47) = .070781$



Critical value:
Decision Rule: Reject H_0 if ______________________
Decision using critical value approach:

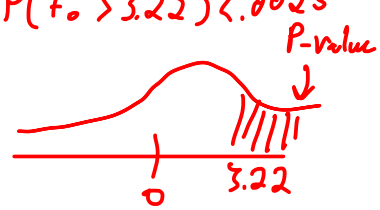
Significance level: $\alpha = 0.01$
Decision using p-value approach:

2. Suppose the EPA is in the process of monitoring the water quality in a large estuary in the eastern United States, in order to measure the polychlorinated biphenyl (PCB) concentration in parts per billion (ppb). Suppose that the random sample of size 80 has a mean of 1.59 ppb and a standard deviation of 0.25 ppb. Is there evidence that the mean PCB concentration in the estuary is higher than 1.50 ppb? Use the 5% significance level.
Test the hypotheses: $\bar{x} = 1.59$ $S = .25$
$H_0: \mu = 1.50$
$H_a: \mu > 1.50$
Compute test statistic:

$t_0 = \dfrac{1.59 - 1.5}{.25/\sqrt{80}} = 3.22$

Find the p-value:

$P(T_0 > 3.22) < .0025$



Critical value:
Decision Rule: Reject H_0 if ______________________
Decision using critical value approach:

Significance level: $\alpha = 0.05$
Decision using p-value approach:

Z rules = t rules

Name: ______________________________

Directions: First find the critical values in #3, then use those values in #4 for the decision rule "Reject H_0 if …"

	Critical value	Reject H_0 if:
H_a: $\mu < \mu_0$	$-Z_\alpha$	$Z_0 < -Z_\alpha$
H_a: $\mu > \mu_0$	Z_α	$Z_0 > Z_\alpha$
H_a: $\mu \neq \mu_0$	$\pm Z_{\alpha/2}$	$\lvert Z_0 \rvert > Z_{\alpha/2}$

3. Find the following critical values. Mark the negative and positive critical values on the Z curve provided.

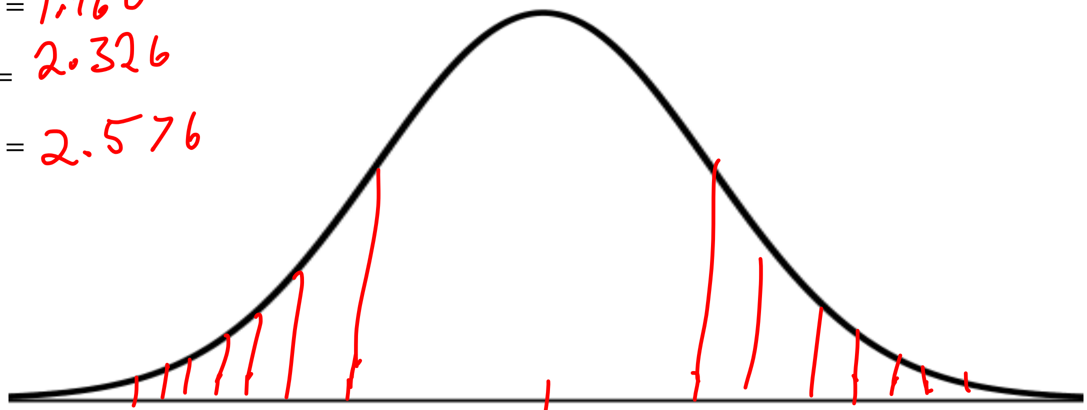
$z_{0.1} =$ 1.282

$z_{0.05} =$ 1.645

$z_{0.025} =$ 1.960

$z_{0.01} =$ 2.326

$z_{0.005} =$ 2.576



* Get annotated note for this page 10/11/23 (and previous page)

4. Suppose the value of σ is known. Find the critical value. For what test statistic values will H_0 be rejected.

- H_a: $\mu < \mu_0$ and $\alpha = 0.05$ Critical value:__________ Reject H_0 if __________
- H_a: $\mu \neq \mu_0$ and $\alpha = 0.05$ Critical values:__________ Reject H_0 if __________
- H_a: $\mu > \mu_0$ and $\alpha = 0.10$ Critical value:__________ Reject H_0 if __________
- H_a: $\mu < \mu_0$ and $\alpha = 0.01$ Critical value:__________ Reject H_0 if __________
- H_a: $\mu \neq \mu_0$ and $\alpha = 0.01$ Critical values:__________ Reject H_0 if __________
- H_a: $\mu > \mu_0$ and $\alpha = 0.05$ Critical value:__________ Reject H_0 if __________

5. Return to problems 1 and 2, first find the appropriate critical value for each, state the decision rule "Reject H_0 if …", then make a decision regarding using H_0 using (1) the critical value approach and again using (2) the p-value approach.

> iClicker Question:

The researcher is running a t-test at the 0.05 significance level.

$$H_0\text{: } \mu = 100$$
$$H_a\text{: } \mu \neq 100$$

Test statistic: $t_0 = -2.54$

What is the p-value?

A. $P(T < -2.54)$

B. $P(T \succ 2.54)$

C. $2*P(T > |-2.54|)$

D. $\alpha = 0.05$

> iClicker Question:

The researcher is running a t-test at the 0.05 significance level.

$$H_0\text{: } \mu = 50$$

$$H_a\text{: } \mu > 50$$

Test statistic: $t_0 = 4.59$

What is the p-value?

A. $P(T < 4.59)$

B. $P(T > 4.59)$

C. 2*$P(T > |4.59|)$

D. $\alpha = 0.05$

> iClicker Question:

The researcher is running a t-test at the 0.05 significance level.

H_0: $\mu = 25$

H_a: $\mu < 25$

Test statistic: $t_0 = -1.50$

What is the p-value?

A. $P(T < -1.50)$

B. $P(T > 1.50)$

C. $2 * P(T > |-1.50|)$

D. $\alpha = 0.05$

Decisions: p-value approach

- **<u>Significance level (α)</u>:** a probability <u>cut-off</u> or <u>threshold</u> to make a decision about H_0. Decision rules are below.
  - Typically, α = 0.10 (10%), α = 0.05 (5%), or α = 0.01 (1%)

If p-value $< \alpha$, Reject H_0

- "Strong enough" evidence against H_0
- Results ARE statistically significant.

If p-value $\geq \alpha$, Fail to Reject H_0

- "Not enough" evidence against H_0
- Results ARE NOT statistically significant.

> iClicker Question:

If p-value is 0.021 and the significance level is $\alpha = 0.05$, what is the appropriate decision?

A. Reject H_0
B. Fail to Reject H_0

> iClicker Question:

If p-value is 0.48 and the significance level is $\alpha = 0.05$, what is the appropriate decision?

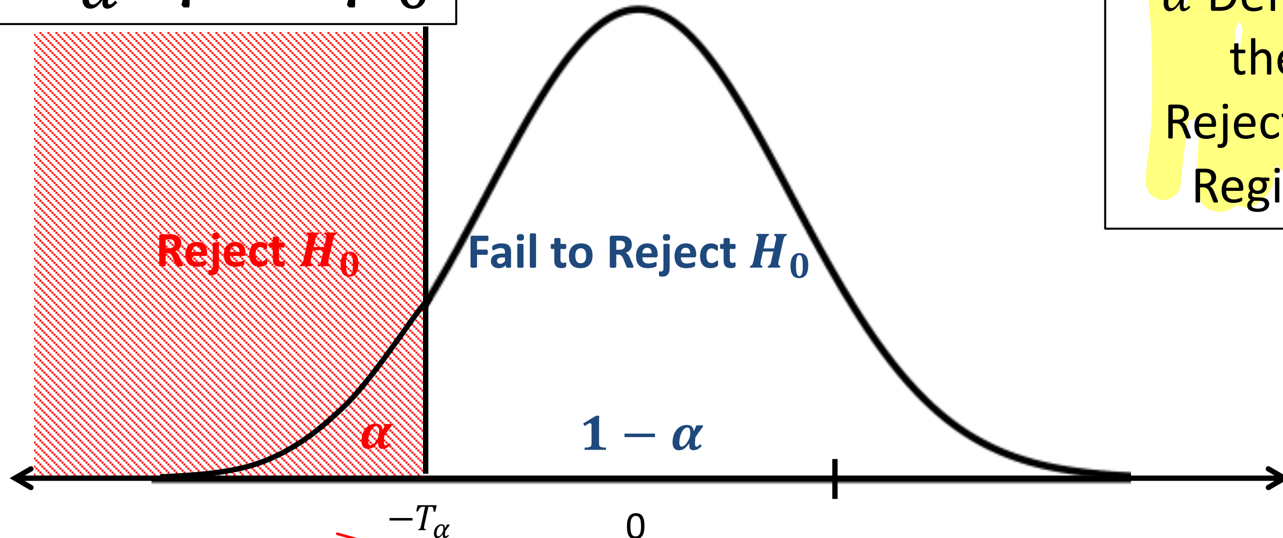
A. Reject H_0
B. Fail to Reject H_0

Decisions: Critical value approach

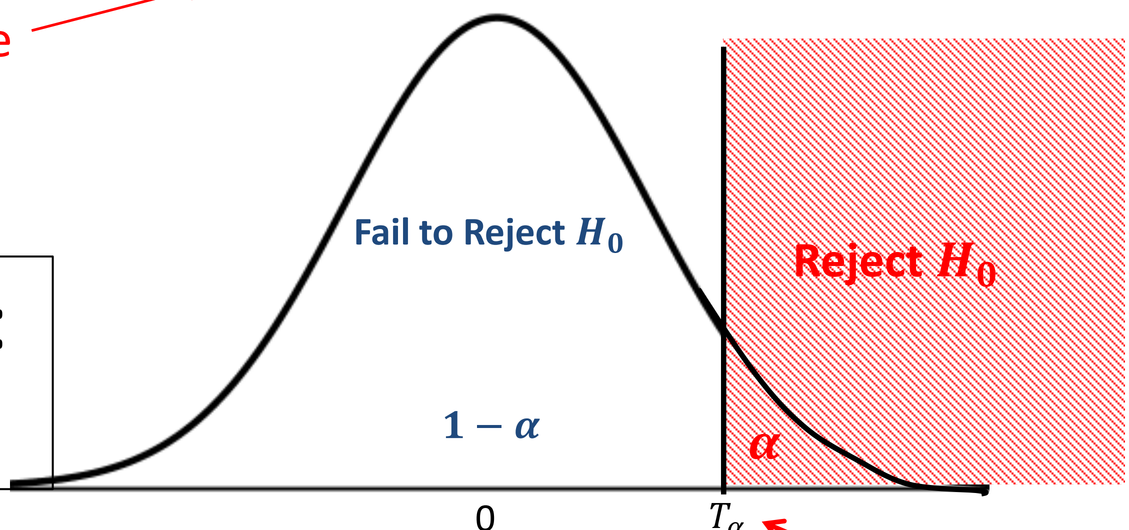
Top Dist: H_a: $\mu < \mu_0$

α Defines the Rejection Region

<u>Decisions</u>:
- Reject H_0 if $T_0 < -T_\alpha$
- FTR H_0 if $T_0 \geq -T_\alpha$



Critical value



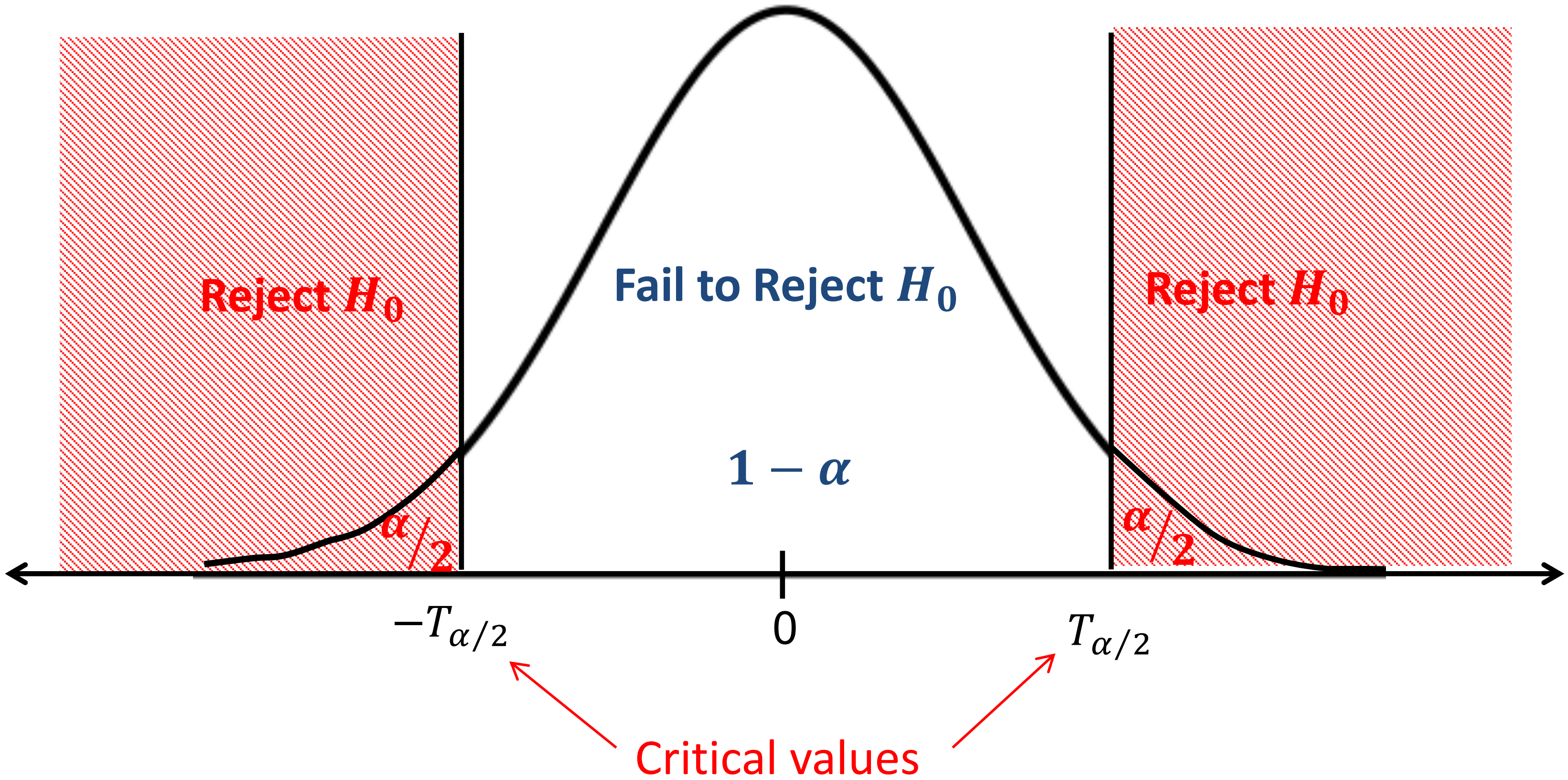
Bottom Dist:
H_a: $\mu > \mu_0$

<u>Decisions</u>:
- Reject H_0 if $T_0 > T_\alpha$
- FTR H_0 if $T_0 \leq T_\alpha$

Critical value

Decisions: Critical value approach

when H_a: $\mu \neq \mu_0$



same numbers

<u>Decisions</u>:

Reject H_0 if $|T_0| > T_{\alpha/2}$

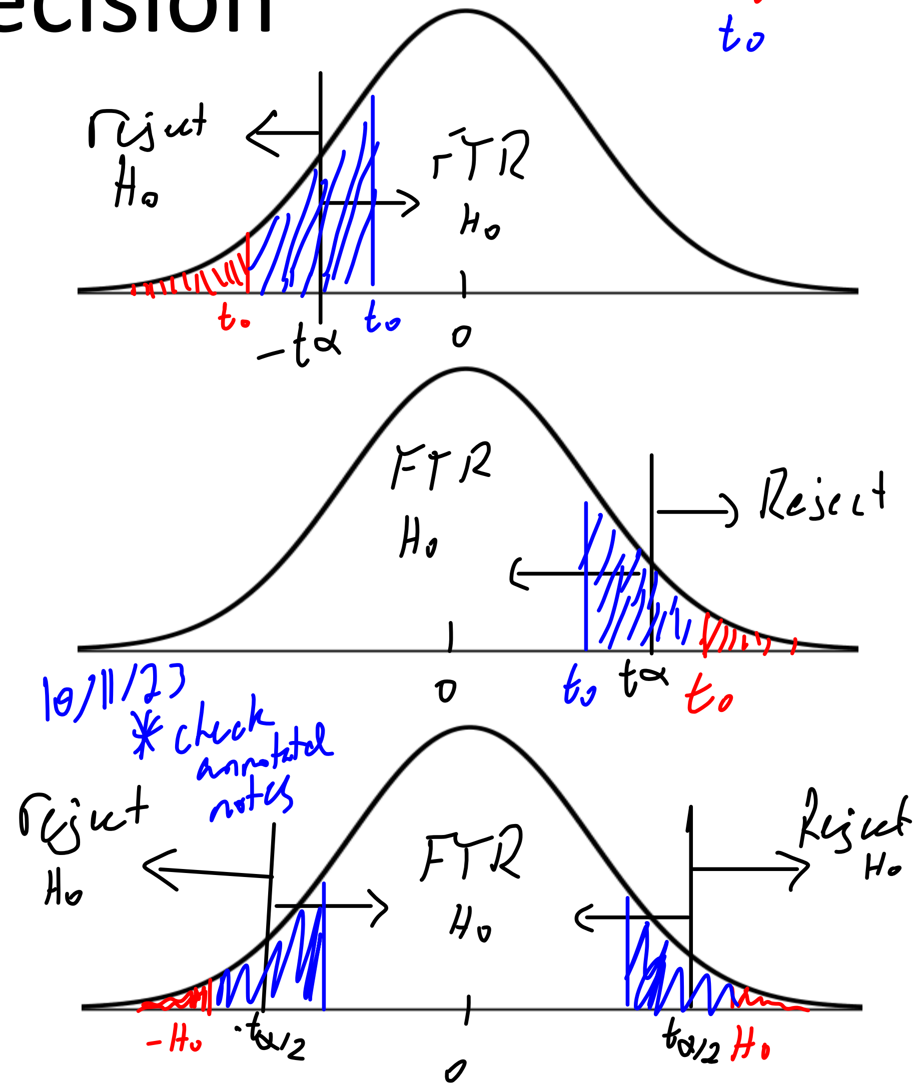
Fail to Reject H_0 if $|T_0| \leq T_{\alpha/2}$

2 Ways to Make a Decision

2 hypothetical test stats
t_0
t_0

1) Critical Value Approach

	Reject H_0 if:
H_a: $\mu < \mu_0$	$T_0 < -T_\alpha$
H_a: $\mu > \mu_0$	$T_0 > T_\alpha$
H_a: $\mu \neq \mu_0$	$\lvert T_0 \rvert > T_{\alpha/2}$



2) p-value Approach

Decision	p-value
Reject H_0	p-value $< \alpha$
Fail to Reject H_0	p-value $\geq \alpha$

P-value < α means the test statistic value lies in the rejection region

P-value ≥ α means the test statistic value lies in the fail to reject region

Example: Emissions (cont.)

Exhaust Emissions: $X \sim N(1.45, \sigma^2)$

$H_0: \mu$ = 1.45 vs. $H_a: \mu$ < 1.45

n=28, $\bar{x}$=1.21, s=0.4

Test statistic: $t_0 = -3.17$

0.001 < p-value < 0.0025

α=0.05

- What is the decision (based on the p-value)?

Reject H_0 b/c p-value < α

- What is the decision (based on the critical value)?

Critical value: $-t_{.05,27} = -1.703$

Reject H_0 b/c $-3.17 < -1.703$

Conclusion

Conclusion – a statement of the amount of evidence for H_a.

State in the context of the problem

- If you reject H_0, then you conclude:

There is enough evidence to conclude H_a.

- If you fail to reject H_0, then you conclude:

There is not enough evidence to conclude H_a.

Quote: "The absence of evidence is not the evidence of absence!"

IMPORTANT: Just because we did not find enough evidence to reject H_0, that does not imply the H_0 is true!

Example: Emissions (cont.)

Exhaust Emissions: $X \sim N(1.45, \sigma^2)$

$H_0 : \mu = 1.45$ vs. $H_a : \mu < 1.45$

Stated in words

n=28, $\bar{x}$=1.21, s=0.4

Test statistic: $t_0 = -3.17$

0.001 < p-value < 0.0025

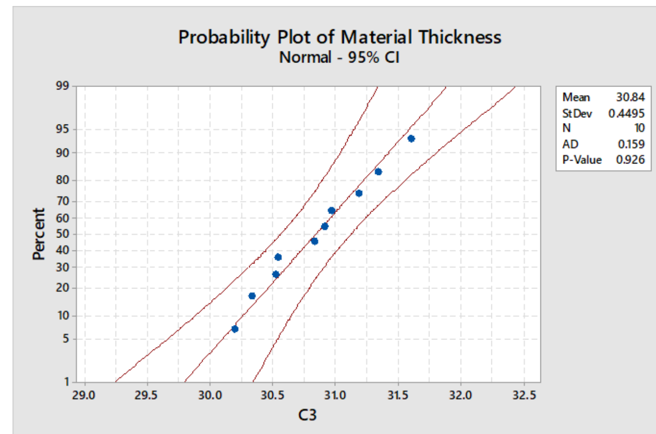
Decision: Reject H_0

- State the conclusion in the context of the problem.

There is enough evidence to suggest the true mean exhaust emission of the car model using the new process is less than 1.45 nitrous oxide per mile

Example: Material Thickness

Material manufactured continuously before being cut and wound into large rolls must be monitored for thickness (caliper). A random sample of 10 measurements were taken in millimeters: $\bar{x} = 30.91$ and $s = 2.1$. Using α=0.10, test $\boxed{H_0:\ \mu = 30}$ against the **two-sided** alternative.



1. State the hypotheses.

$H_0: \mu = 30$

$H_a: \mu \neq 30$

2. Find the critical value.

$\pm t_{.5,9} = \pm 1.833$

3. For what test statistic values will H_0 be rejected?

Reject H_0 if $|T_0| > 1.833$

4. Calculate the test statistic.

$t_0 = \dfrac{30.91-30}{2.1/\sqrt{10}} \approx 1.37$

Example: Material Thickness (cont.)

5. Find the p-value. $P(T_0 >$

6. Make a decision using the 0.10 significance level.

7. State the conclusion <u>in the context of the problem</u>.

Type I Error, Type II Error, and Power

- **Type I Error:** Rejecting H_0 when H_0 is actually true.

  P(Type I Error) = α

- **Type II Error:** Failing to Reject H_0 when H_a is actually true.

  P(Type II Error) = β

- **Power:** Probability of rejecting H_0 when H_a is true. Power = 1 - β

TRUTH TABLE

	H_0 is true	H_a is true
Reject H_0	Type I error prob = α	Correct prob = $1-\beta$
Fail to Reject H_0	Correct prob = $1-\alpha$	Type II error prob = β

sum = 1   sum =

"POWER" of the test

Trial by Jury Analogy

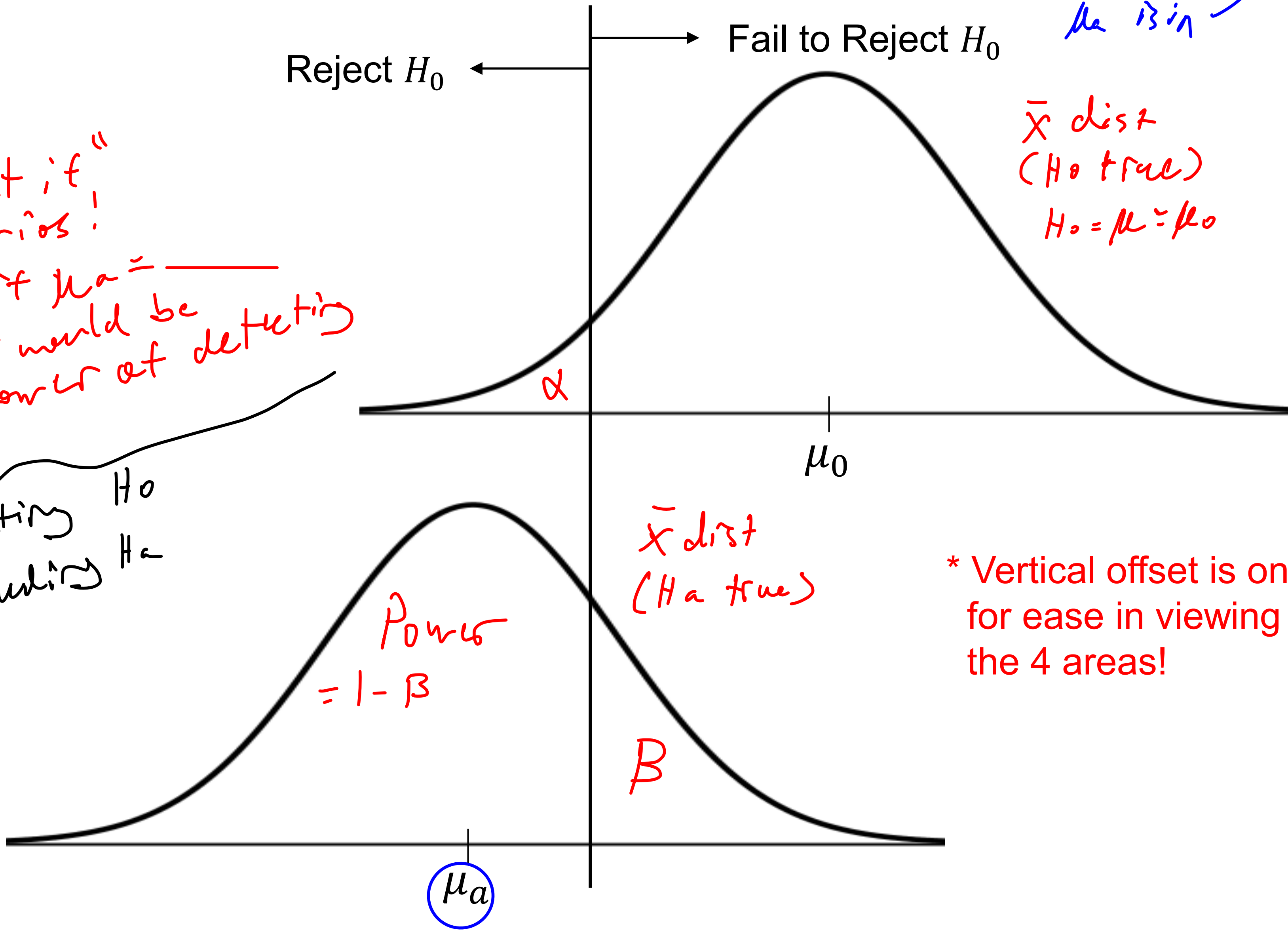
	Trial by Jury	Hypothesis Test
Who/What	Defendant on trial	H_0
Assumption	Defendant is "Innocent until proven guilty"	H_0 is assumed to be true until there is enough evidence to reject it.
Collect evidence	Collect blood, hair, DNA samples, etc.	Collect data and calculate Test Statistic and P-value
Amount of evidence	Beyond a reasonable doubt	α
Verdict / Decision	Guilty	Reject H_0
	Not guilty	Fail to Reject H_0
Error	Convicting an innocent person	Type I Error (Rejecting H_0 when H_0 is true)
	Release a guilty person	Type II Error (Failing to reject H_0 when H_0 is false)

α, β, and Power where $H_a: \mu < \mu_0$

μ_a is in

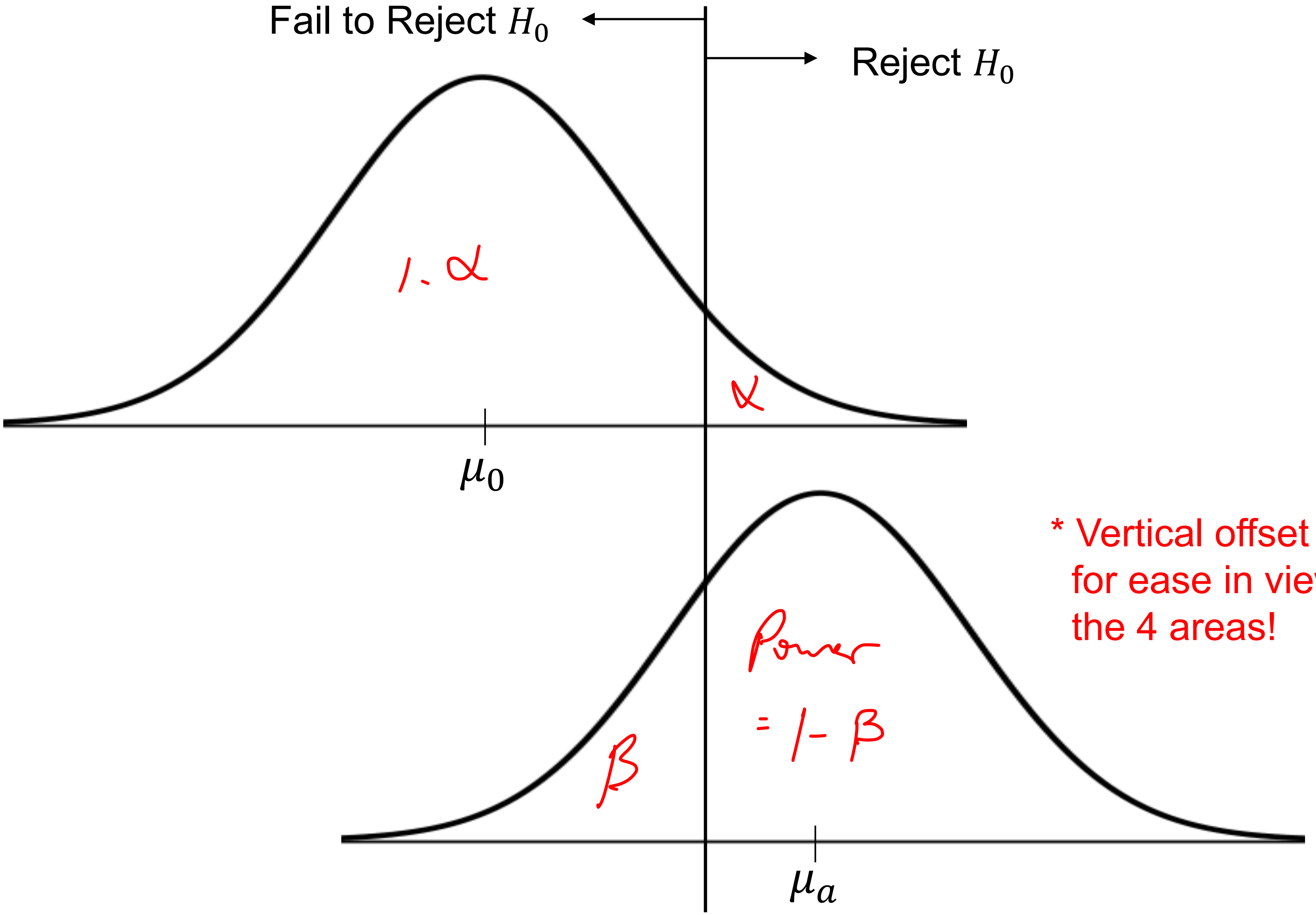
"what if" scenarios!
What if $\mu_a =$ ______
what would be the power of detecting it?

Rejecting H_0
concluding H_a



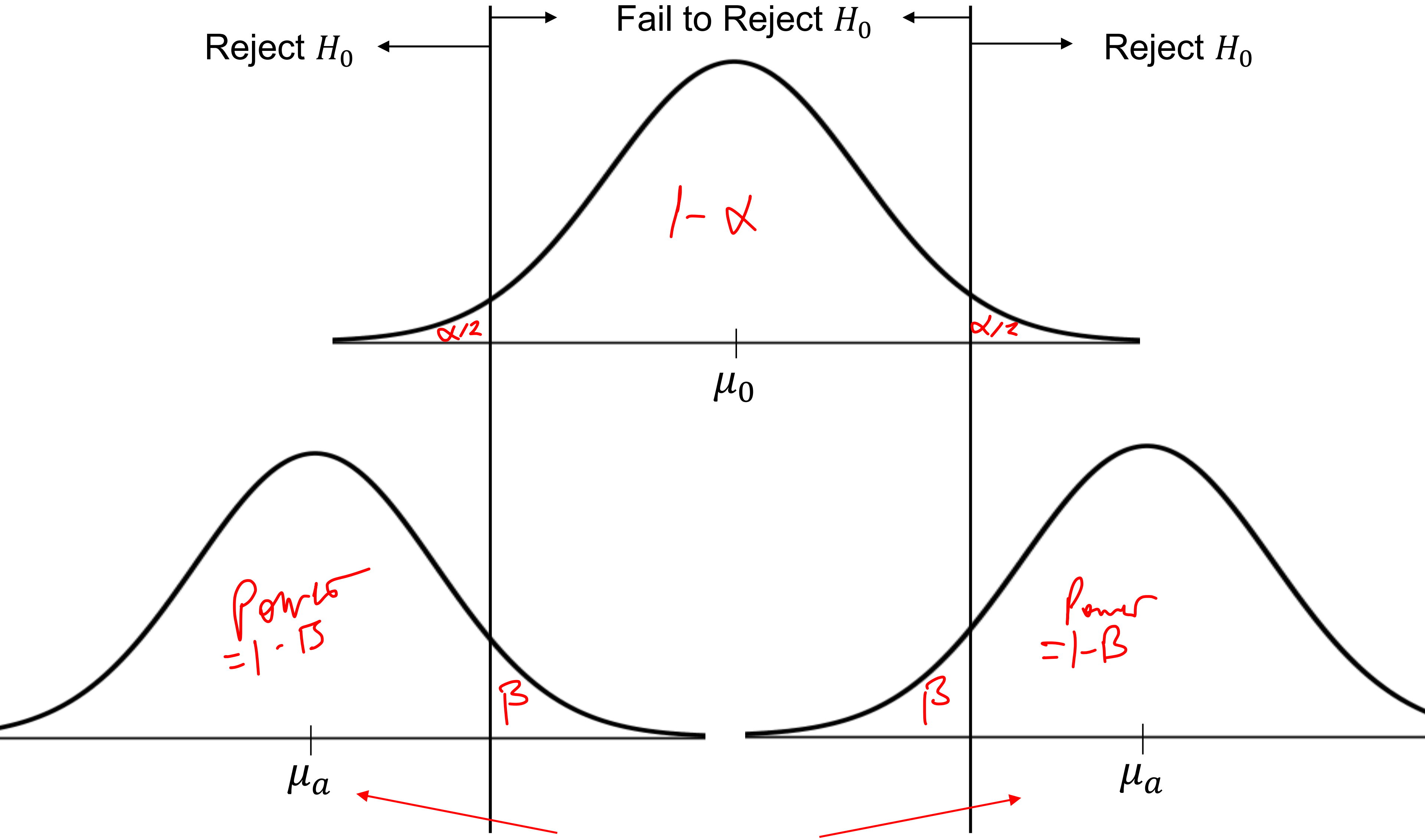
* Vertical offset is only for ease in viewing the 4 areas!

α, β, and Power where H_a: $\mu > \mu_0$



* Vertical offset is only for ease in viewing the 4 areas!

α, β, and Power where H_a: $\mu \neq \mu_0$



μ_a will either be less than or greater than μ_0, not both!

The Trade-offs

$\alpha \uparrow \ \beta \downarrow \ P \uparrow$

- If α increases, then β ___decreases___ and power ___increases___.
- If n increases, then power ___increases___.
- If σ decreases, then power ___increases___.
- If distance between μ_a and μ_0 increases, then power ___increase___.
- How could you choose a small α and still decrease β?

  Increase sample size (n)

> iClicker Question:

Holding all other values constant…
If the sample size increases, I would expect power to:

A. increase.

B. decrease.

more data, more likely to be able to detect difference

> iClicker Question:

Holding all other values constant…
If the alpha (α) increases, I would expect power to:

A. increase.
B. decrease.

$\alpha \uparrow \beta \downarrow P \uparrow$

> iClicker Question:

$\delta = \mu_a - \mu_0$ (distance between H_0 and H_a dist. centers)

Holding all other values constant…
If $|\delta|$ increases, I would expect power to:

A. increase.

B. decrease.

larger differences are easier to detect, // more likely

As α increases, power increases.

1-Sample Z Test
Testing mean = null (versus ≠ null)
Calculating power for mean = null + difference
$\boxed{\alpha = 0.01}$ Assumed standard deviation = 1

Results

Difference	Sample Size	Power
1.2	6	0.641906

1-Sample Z Test
Testing mean = null (versus ≠ null)
Calculating power for mean = null + difference
$\boxed{\alpha = 0.05}$ Assumed standard deviation = 1

Results

Difference	Sample Size	Power
1.2	6	0.836315

As n increases, power increases.

1-Sample Z Test
Testing mean = null (versus ≠ null)
Calculating power for mean = null + difference
α = 0.01 Assumed standard deviation = 1

Results

Difference	Sample Size	Power
1.2	6	0.641906

1-Sample Z Test
Testing mean = null (versus ≠ null)
Calculating power for mean = null + difference
α = 0.01 Assumed standard deviation = 1

Results

Difference	Sample Size	Power
1.2	10	0.888560

As $|\delta|$ increases, power increases.

(where $\delta = \mu_a - \mu_0$)

1-Sample Z Test
Testing mean = null (versus ≠ null)
Calculating power for mean = null + difference
α = 0.01 Assumed standard deviation = 1

Results

Difference	Sample Size	Power
$\boxed{0.8}$	10	$\boxed{0.481652}$

1-Sample Z Test
Testing mean = null (versus ≠ null)
Calculating power for mean = null + difference
α = 0.01 Assumed standard deviation = 1

Results

Difference	Sample Size	Power
$\boxed{1.2}$	10	$\boxed{0.888560}$

> iClicker Question:

Holding all other values constant…
If the power increases, I would expect the required sample size to:

Recall: $n \uparrow$ power $\uparrow$

Now: power $\uparrow$ $n \uparrow$

A. increase.

B. decrease.

> iClicker Question:

Holding all other values constant…
If the alpha (α) increases, I would expect the required sample size to:

$\alpha \uparrow$ power $\uparrow$

$n \uparrow$ Power $\uparrow$

A. increase.

B. decrease.

> iClicker Question:

$\delta = \mu_a - \mu_0$ (distance between H_0 and H_a dist. centers)

Holding all other values constant…
If $|\delta|$ increases, I would expect the required sample size to:

A. increase.
B. decrease.

Don't need as much data to detect larger differences

As power increases, sample size increases.

1-Sample Z Test
Testing mean = null (versus ≠ null)
Calculating power for mean = null + difference
α = 0.05 Assumed standard deviation = 1

Results

Difference	Sample Size	Power
1.2	6	0.836315

1-Sample Z Test
Testing mean = null (versus ≠ null)
Calculating power for mean = null + difference
α = 0.05 Assumed standard deviation = 1

Results

Difference	Sample Size	Power
1.2	8	0.924235

As α increases, sample size decreases.

(…to obtain similar power)

1-Sample Z Test
Testing mean = null (versus ≠ null)
Calculating power for mean = null + difference
$\boxed{\alpha = 0.01}$ Assumed standard deviation = 1

Results

Difference	Sample Size	Power
1.2	9	0.847123

1-Sample Z Test
Testing mean = null (versus ≠ null)
Calculating power for mean = null + difference
$\boxed{\alpha = 0.05}$ Assumed standard deviation = 1

Results

Difference	Sample Size	Power
1.2	6	0.836315

As $|\delta|$ increases, sample size decreases.

(… to obtain similar power)

1-Sample Z Test
Testing mean = null (versus ≠ null)
Calculating power for mean = null + difference
α = 0.05 Assumed standard deviation = 1

Results

Difference	Sample Size	Power
1.2	6	0.836315

1-Sample Z Test
Testing mean = null (versus ≠ null)
Calculating power for mean = null + difference
α = 0.05 Assumed standard deviation = 1

Results

Difference	Sample Size	Power
1.5	4	0.850839

Relationship between CI and HT

- You can use a $\boxed{C = 1 - \alpha}$ two-sided CI to test $H_0: \mu = \mu_0$ vs. $H_a: \mu \neq \mu_0$ at the α-level.*

* look for μ_0 in C.I

Decision Rules when using a CI

- Fail to Reject H_0 if μ_0 is within the CI
- Reject H_0 if μ_0 is NOT within the CI

* An upper CI is used to test $H_a: \mu < \mu_0$. A lower CI is used to test $H_a: \mu > \mu_0$.

> iClicker Question:

Use 90% CI = (115.5, 128.3) to test $H_0: \mu = 100$ vs. $H_a: \mu \neq 100$ at α=0.10.

$C = 1 - \alpha$

look for 100 in C.I

What is the decision regarding H_0?

A. Reject H_0

B. Fail to Reject H_0

> iClicker Question:

Use 95% CI = (55.2, 61.1) to test $H_0: \mu = 60$ vs. $H_a: \mu \neq 60$ at α=0.05.

What is the decision regarding H_0?

A. Reject H_0

B. Fail to Reject H_0 b/c 60 is within C.I.

Summary

- General Overview of Hypothesis Testing for the Population Mean (μ): Construct hypotheses about μ, collect data from a random sample drawn from the population, and use $\bar{X}$ to draw a conclusion about μ.
- The purpose of a confidence interval (CI) is to ESTIMATE the value of a parameter.
- The purpose of a hypothesis test (HT) is to DECIDE between two competing claims (i.e. hypotheses) about the value of a parameter.

- Steps for Hypothesis Testing:

  1. State the null hypothesis (H_0) and alternative hypothesis (H_a).
  2. Check the necessary assumptions so the test is valid.
  3. Compute a test statistic. (The test statistic is a statistic that is used to assess the strength of evidence *against* H_0.)
  4. Compute the p-value. (The p-value is the probability of obtaining a test statistic at least as extreme as the observed test statistic, *assuming* H_0 *is true*.)
  5. State the decision about H_0.
  6. State the conclusion about the strength of evidence for H_a.

Summary

- Hypotheses - statements about the value of a parameter
  - Null Hypothesis (H_0) – Statement of no difference.
  - Alternative Hypothesis (H_a) – Statement we want to find evidence for.

IMPORTANT: Never place statistics in hypotheses. Statistics are computed from sample data; there is no need to hypothesize about their value(s).

- Hypothesis Testing for μ:
  - Choose one set of hypotheses where μ_0 is a specified value.

$$H_0: \mu \geq \mu_0 \qquad H_0: \mu \leq \mu_0 \qquad H_0: \mu = \mu_0$$
$$H_a: \mu < \mu_0 \qquad H_a: \mu > \mu_0 \qquad H_a: \mu \neq \mu_0$$

- Test Statistic:
  - If σ is known and $\bar{X}$~N, the test statistic is $Z_0 = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$ where $Z_0 \sim N(0,1)$
  - If σ is unknown and $\bar{X}$~N, the test statistic is $T_0 = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$ where $T_0 \sim t$(n-1)

- Assumptions:
  - The data must be a random sample from the population.
  - The $\bar{X}$ distribution must be at least approximately normal.

Summary

- p-value: the probability of obtaining a test statistic at least as extreme as the observed test statistic value, assuming the H_0 is true.
  - Calculate p-value in the direction of extreme shown in H_a (see table) ->
  - The smaller the p-value, the stronger the evidence *against* H_0 and *for* H_a.

Alternative Hypothesis	p-value
H_a: $\mu < \mu_0$	$P(T_0 < t_0)$
H_a: $\mu > \mu_0$	$P(T_0 > t_0)$
H_a: $\mu \neq \mu_0$	$2 \cdot P(T_0 > \lvert t_0 \rvert)$

- Significance level (α): a probability cut-off or threshold to make a decision about H_0.
  - Typical levels are α = 0.10 (10%), α = 0.05 (5%), and α = 0.01 (1%).

- Two Ways to Make a Decision:

1) Critical Value Approach

	Reject H_0 if:
H_a: $\mu < \mu_0$	$T_0 < -T_\alpha$
H_a: $\mu > \mu_0$	$T_0 > T_\alpha$
H_a: $\mu \neq \mu_0$	$\lvert T_0 \rvert > T_{\alpha/2}$

2) p-value Approach

Decision	p-value
Reject H_0	p-value < α
Fail to Reject H_0	p-value $\geq \alpha$

Summary

- <u>Conclusion</u> – a statement of the amount of evidence <u>*for*</u> H_a.
  - ➢ If you reject H_0, then you conclude:
    "There is enough evidence to conclude $\boldsymbol{H_a}$." (written in context)
  - ➢ If you fail to reject H_0, then you conclude:
    "There is <u>not</u> enough evidence to conclude $\boldsymbol{H_a}$." (written in context)

- Errors in Hypothesis Testing:
  - ➢ **<u>Type I Error</u>:** Rejecting H_0 when H_0 is actually true.
    P(Type I Error) = α
  - ➢ **<u>Type II Error</u>:** Failing to Reject H_0 when H_a is actually true.
    P(Type II Error) = β
  - ➢ **Power** = 1 - β

- Truth Table ->

	H_0 is true	H_a is true
Reject H_0	Type I Error (prob=α)	Correct Decision called "POWER" (prob=$1-\beta$)
Fail to Reject H_0	Correct Decision (prob=$1-\alpha$)	Type II Error (prob=β)

Summary

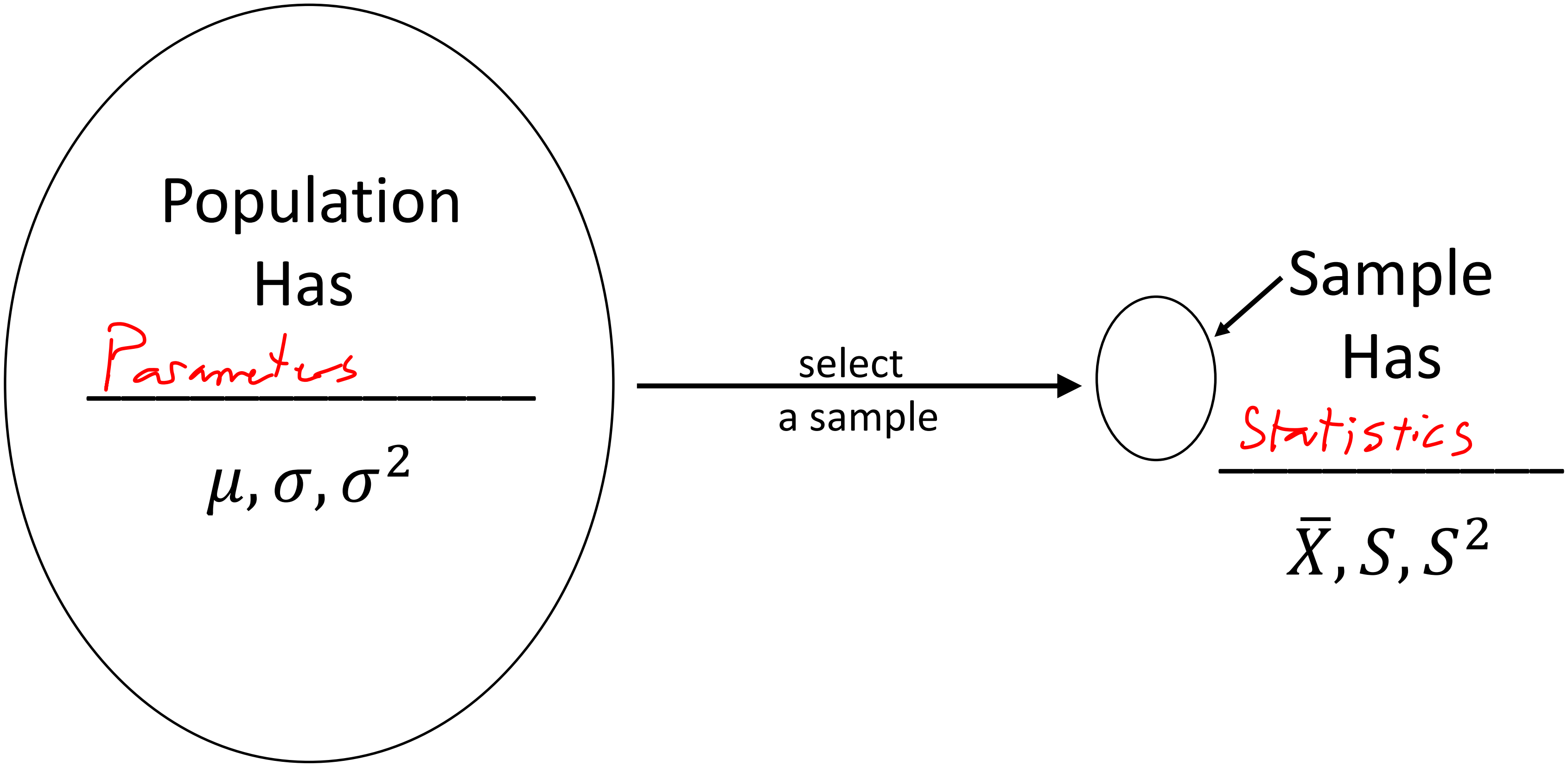
- Trade-offs between α, β, Power, n, and δ (distance between μ_0 and μ_a).
  - Power = 1 - β
  - Power increases with increased sample size (n), increased α, and increased δ.
  - Power of approximately 0.80 or higher is desirable.

- Sample Size Calculations:
  - Use Minitab to identify the required sample size, using inputs of power, α, δ, and σ.
  - Sample size increases with increased power, decreased α, and decreased δ.
  - Round up to the next whole number for the required sample size.

- A 100*C % CI can be used to make a decision about a hypothesis test at the α significance level.
  - Fail to Reject H_0 if μ_0 is within the CI
  - Reject H_0 if μ_0 is NOT within the CI

Exam 2 Review

CHAPTER 7



Terminology:

Statistics vary from sample to sample is called Sampling Variability.

The Sampling dist. of $\bar{x}$ is the distribution of sample means calculated from all possible samples of size n drawn from the pop.

These facts are true regardless of dist. shape or sample size!

<u>Facts</u>: $\mu_{\bar{X}} = \mu_x$ $\sigma_{\bar{X}}^2 = \frac{\sigma_x^2}{n}$ $\sigma_{\bar{X}} = \frac{\sigma_x}{\sqrt{n}}$

- Normal Dist. Fact:

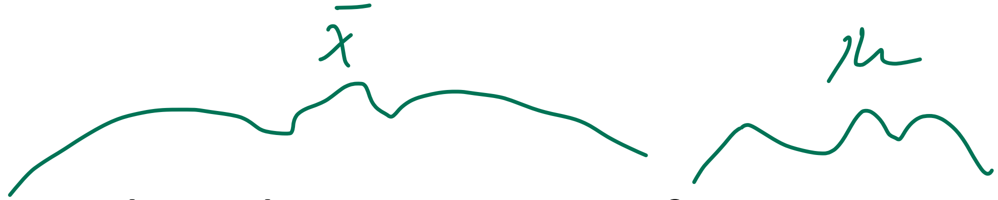
  If $X \sim N$, then $\bar{X} \sim N$ ____ for any sample size n.

- Central Limit Theorem:

  If a random sample is drawn from a non-normal pop, then $\bar{x} \mathrel{\dot\sim} N$ ____ for $n \geq 30$.

Proper Dist. Notation: $\bar{X} \sim N(\mu_x, \frac{\sigma_x^2}{n})$

CHAPTER 8



Point estimate – a single-value estimate of parameter

Interval estimate – an interval of values to estimate the parameter. CI for μ

General Form of Two-sided CI:

Point estimate $\pm$ Margin of error

Two-sided CI for μ

(When σ is known)

$$\bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Two-sided CI for μ

(When σ is unknown) use s

$$\bar{X} \pm T_{\alpha/2}\frac{s}{\sqrt{n}}$$

Margin of error:

$$E = Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$E = T_{\alpha/2} \frac{s}{\sqrt{n}}$$

Check z-table for values

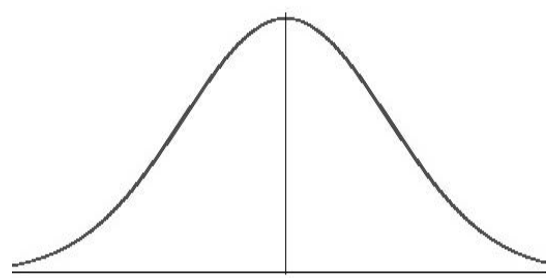
Confidence ($C = 1 - \alpha$)	$Z_{\alpha/2}$
90% ($\alpha = 0.10$)	$Z_{.05} = 1.645$
95% ($\alpha = 0.05$)	$Z_{.025} = 1.96$
99% ($\alpha = 0.01$)	$Z_{.005} = 2.576$

CI Behavior:

$C \uparrow E$ $\uparrow$

$\sigma \uparrow E$ $\uparrow$

$n \uparrow E$ $\downarrow$



Sample size calculation: $n = \left(\frac{Z_{\alpha/2} \cdot \sigma}{E}\right)^2$

* Remember to round up to next larger whole number!

* Better image can be found on D2L

BEFORE Data Collection	AFTER Data Collection

BEFORE Data Collection

- Set up hypotheses

  $H_0 : \mu = \mu_0$

  $H_a : \mu \boxed{\begin{matrix} < \\ > \\ \neq \end{matrix}} \mu_0$ — choose one!

- Choose significance level (α)
- Find critical value(s)

  → α is the area of the rejection region beyond critical value(s)

AFTER Data Collection

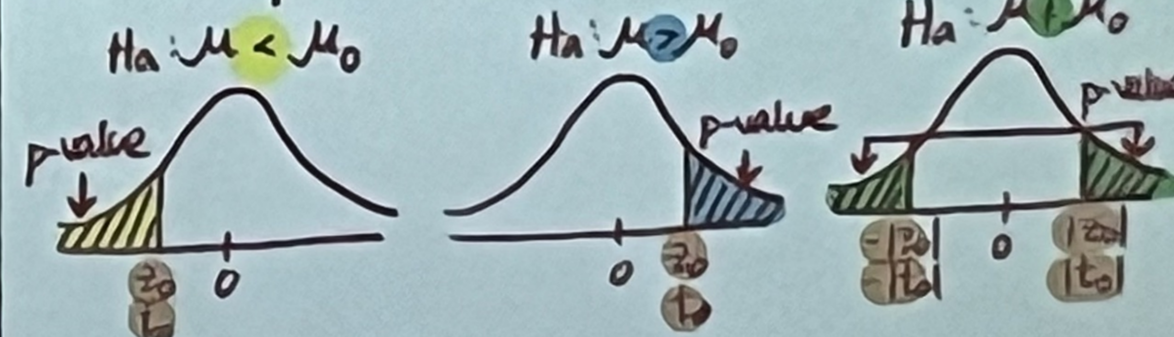
- Calculate test statistic: $z_0 = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ $\quad T_0 = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$

  Assess evidence against H_0

  - LARGE |test stat|
  - SMALL p-value

  are evidence against H_0

- Find p-value: (smaller than α)



- Make decision

  → compare p-value to α, or

  → compare test stat. to critical value(s)

- State conclusion - statement of evidence for H_a.

CHAPTER 9

<u>BEFORE DATA COLLECTION</u>:

1) Set up hypotheses

H_0:

H_a:

← described on previous slide

2) Choose α

<u>AFTER DATA COLLECTION</u>:

3) Test Statistic

4) p-value

5) Decision

6) Conclusion

Hypothesis Testing for μ

H_0: $\mu = \mu_0$

H_a: $\mu \boxed{\begin{matrix} < \\ > \\ \neq \end{matrix}} \mu_0$ - Choose one!

* Use μ in both hypotheses

* Use the same number in both hypotheses

Assumptions:

1. The data must be a random sample from the population.
2. The $\bar{X}$ distribution must be at least approximately normal.

Test Statistic:

σ is known

$$Z_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

σ is unknown

$$T_0 = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

Pre-determined cutoff values

Two ways to make a decision:

Based on sample data

1) p-value approach:	Compare	p-value	to	Sig. level α
2) critical-value approach:	Compare	Test Stat. t_0	to	(see below)

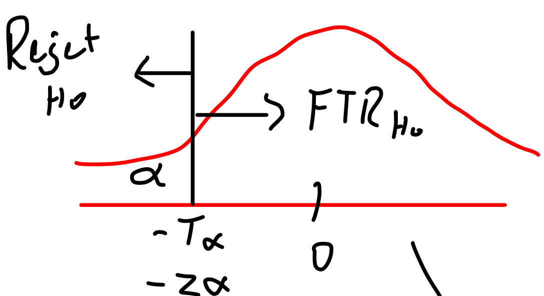
Critical Value	H_a
$-t_\alpha$	$<$
$+t_\alpha$	$>$
$\pm t_{\alpha/2}$	$\neq$

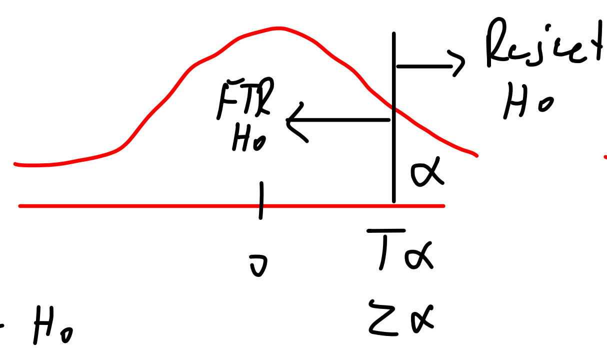
Critical values:

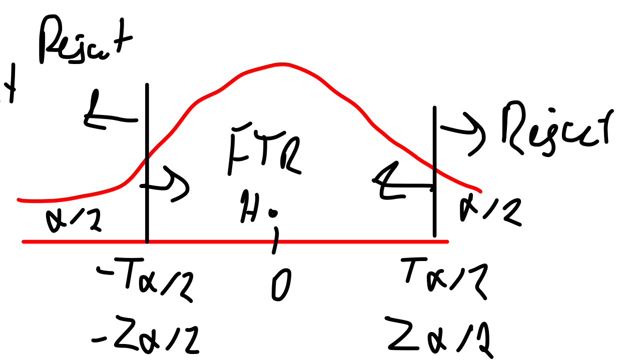
* Depend on:

1) σ known or unknown
   (Z) (T)
2) Direction in H_a
3) α

H_a: $\mu < \mu_0$ H_a: $\mu > \mu_0$ H_a: $\mu \neq \mu_0$







Decision:

(critical-value approach)

Reject H_0 if $T_0 < -T_\alpha$
$Z_0 < -Z_\alpha$

Reject H_0 if $T_0 > T_\alpha$
$Z_0 > Z_\alpha$

Reject H_0 if $|T_0| > T_{\alpha/2}$
$|Z| > Z_{\alpha/2}$

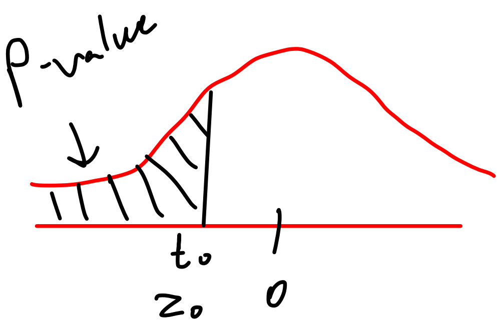
p-value:

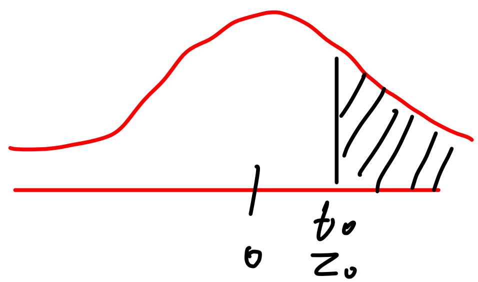
* Depends on:

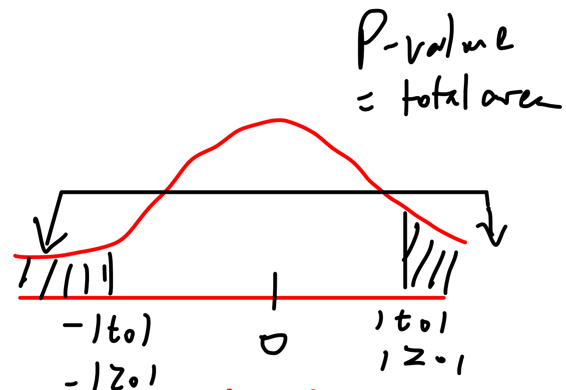
1) Test statistic

2) Direction of H_a

H_a: $\mu < \mu_0$ $\qquad$ H_a: $\mu > \mu_0$ $\qquad$ H_a: $\mu \neq \mu_0$







Decision: p-value $\quad \alpha$, Reject H_0

(p-value approach) p-value $\quad \alpha$, Fail to Reject H_0

* where α is the significance level (0.01, 0.05, or 0.10)

Conclusion: Statement of evidence for H_a

* Failing to Reject H_0 DOES NOT imply H_0 is true!

Truth Table:

	H_0 is true	H_a is true
Reject H_0	Type I Error prob = α	CORRECT prob = $1-\beta$
Fail to Reject H_0	CORRECT prob = $1-\alpha$	Type II Error prob = β
	sum = 1	sum = 1

"Power"

Power Behavior:

Use ↑ for increases.

Use ↓ for decreases.

Holding all other values constant…

- If α ↑, then β ↓ and Power $= 1-\beta$ ↑
- If σ ↑, then Power ↓
- If n ↑, then Power ↑
- If $\delta = \mu_a - \mu_0$ ↑, then Power ↑

if $H_a: \mu > \mu_0$,

→ reject right

Sample Size Behavior:

Holding all other values constant…

- If α ↑, then n ↓
- If σ ↑, then n ↑
- If Power ↑, then n ↑
- If $\delta = \mu_a - \mu_0$ ↑, then n ↓

Relationship between CI and HT

- Use a (100*C)% CI to conduct a hypothesis test at the α significance level where $C = 1 - \alpha$.

✓ If μ_0 is in CI, then Fail to reject H_0

✓ If μ_0 is not in CI, then Reject H_0

$H_0: \mu = \mu_0$

$H_a: \mu \neq \mu_0$

if sigma is given, use Z, if not given use T

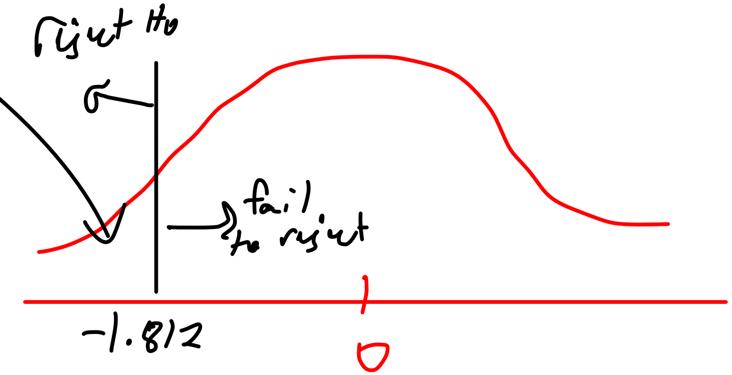
Review of How to Use T table to find Critical Value(s)

df = n-1

Use df=10

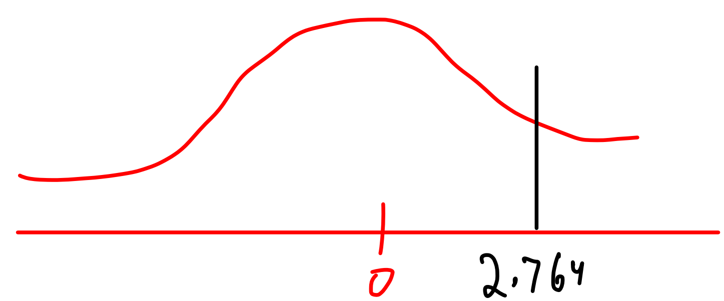
Ex: H_a: $\mu < \mu_0$ and $\alpha = \boxed{0.05}$

$-t_{.05,10} = -1.812$



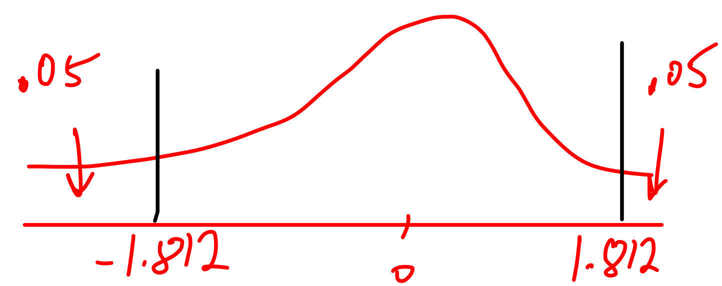
Ex: H_a: $\mu > \mu_0$ and $\alpha = 0.01$

$t_{.01,10} = 2.764$



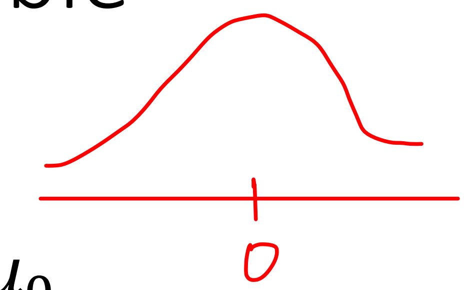
Ex: H_a: $\mu \neq \mu_0$ and $\alpha = 0.10$

$\pm t_{\frac{.1}{2},10} = \pm 1.812$



Review of How to Use T table to find p-value

Use df=10



Ex: $H_a\colon \mu > \mu_0$

$.025 < P(T > 2.1) < .05$

$P(T > 7.9) < .0005$

$P(T > 0.22) > .4$

$H_a\colon \mu < \mu_0$

$.025 < P(T < -2.1) < .05$

$P(T < -7.9) < .0005$

$P(T < -0.22) > .4$

$H_a\colon \mu \neq \mu_0$

$.05 < 2 \cdot P(T > 2.1) < .1$

$2 \cdot P(T > 7.9) < .001$

$2 \cdot P(T > 0.22) > .8$

* 5:30 review session in NAH Tues.

ch 8: "one-sample"
Z or T CIs

Ch 9: "onesmple"
Z or T tests

C 10: "two sample"
Z or T

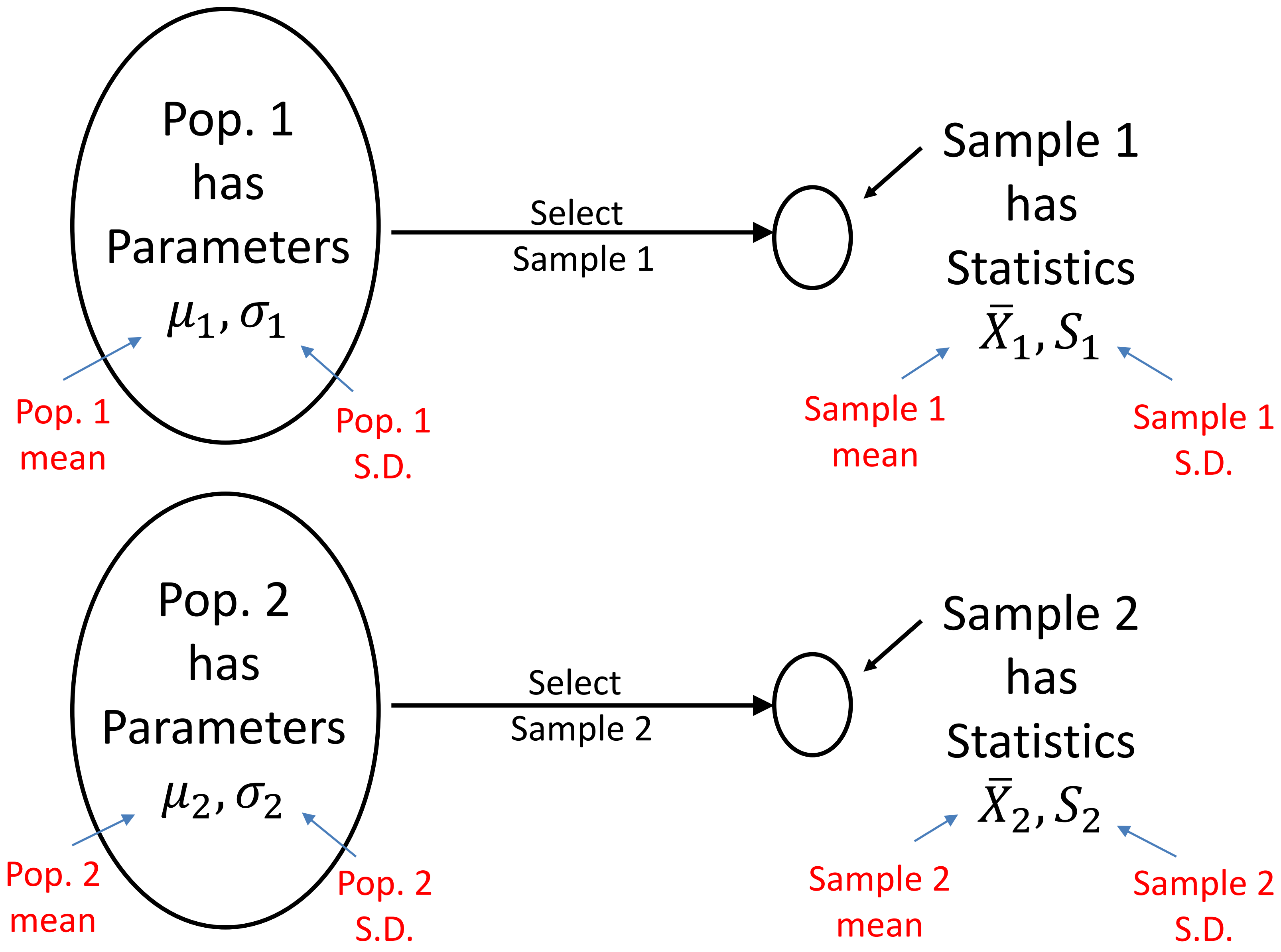
EGEN350

Chapter 10

Inference for the Diff. in Two Pop. Means ($\mu_1 - \mu_2$)

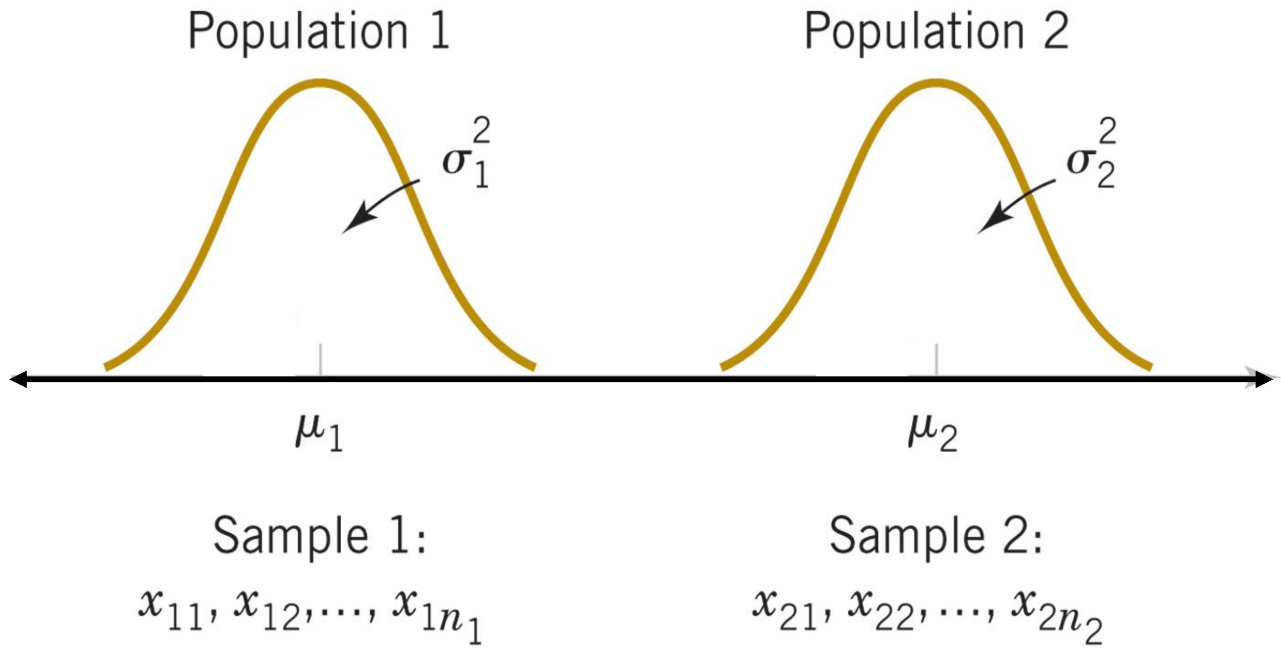
(Skip Sections 10-3, 10-5, 10-6, and 10-7)

Two Populations



> iClicker Question:

Note:
x_{ij}
$i = 1, 2$
j



When comparing the means of two populations, the parameter is $\mu_1 - \mu_2$. What statistic is the best estimator of $\mu_1 - \mu_2$?

A. $\bar{X}_1$

B. $\bar{X}_2$

C. $\bar{X}_1 - \bar{X}_2$

Sampling Distribution of $\bar{X}_1 - \bar{X}_2$

- $\bar{X}_1$: mean of the sample selected from pop. 1
- $\bar{X}_2$: mean of the sample selected from pop. 2

What is the center and spread of the $\bar{X}_1 - \bar{X}_2$ dist?

CENTER:

$$\boxed{\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2}$$

*THESE ARE TRUE, REGARDLESS OF DIST. SHAPE OR SAMPLE SIZE!

SPREAD:

$$\boxed{\sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$\boxed{\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Sampling Distribution of $\bar{X}_1 - \bar{X}_2$

What is the shape of the $\bar{X}_1 - \bar{X}_2$ dist?

- <u>Normal Distribution Fact</u>:

  If $\bar{X}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right)$ and $\bar{X}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$, then:

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \text{ for any sample size}$$

- <u>Central Limit Theorem</u>: If population distributions (X_1 and X_2) are not normal but sample sizes are at least 30, then

  $\bar{X}_1 \dot{\sim} N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right)$ and $\bar{X}_2 \dot{\sim} N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$ and

$$\bar{X}_1 - \bar{X}_2 \dot{\sim} N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \text{ for } n_1 \geq 30 \text{ and } n_2 \geq 30$$

Two Scenarios When Comparing Two Population Means

1. Independent Samples

*Sample sizes (n_1, n_2) can be equal or unequal.

- Two independent samples drawn from two independent populations

2. Dependent (Paired) Samples

*Sample sizes (n_1, n_2) MUST be equal.

- Two treatments applied to the same individuals or paired individuals (Ex: "Before/After" data)
- Two treatments applied under the same conditions

> iClicker Question:

In an experiment involving the breaking strength of a certain type of thread used in personal flotation devices, one batch of thread was subjected to a heat treatment of 60 seconds and another batch was treated for 120 seconds. The breaking strengths (in N) of 10 threads in each batch were measured.

A. Independent Samples
B. Dependent (Paired) Samples

> iClicker Question:

The manager of a fleet of automobiles is testing two brands of radial tires and assigns one tire of each brand at random to the two rear wheels of eight cars and runs the cars until the tires wear out. The data (in kilometers) were recorded.

A. Independent Samples

B. Dependent (Paired) Samples

Two-sided Confidence Interval for $\mu_1 - \mu_2$ Independent Samples, Pop. Variances Unknown

If $\bar{X}_1$ and S_1 are the mean and standard deviation of a random sample of size n_1 from a normal population 1, and $\bar{X}_2$ and S_2 are the mean and standard deviation of a random sample of size n_2 from a normal population 2, then a $100(1-\alpha)$ % CI for $\mu_1 - \mu_2$ is:

$$\boxed{\underbrace{\bar{X}_1 - \bar{X}_2}_{\text{point estimate}} \pm T_{\alpha/2} \underbrace{\sqrt{\frac{S_1{}^2}{n_1} + \frac{S_2{}^2}{n_2}}}_{SE_{\bar{x}_1 - \bar{x}_2}}}$$

$T \sim t(df)$

formula on next slide

Assumptions:

1. The data are two independent random sample(s) drawn from two independent populations.
2. The $\bar{X}_1$ and $\bar{X}_2$ dist's must be at least approximately normal.

Hypothesis Testing for $\mu_1 - \mu_2$
Independent Samples, Pop. Variances Unknown

- Use $SE_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$ to estimate $\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

$H_0: \mu_1 - \mu_2 = \Delta_0$	$H_0: \mu_1 - \mu_2 = \Delta_0$	$H_0: \mu_1 - \mu_2 = \Delta_0$
$H_a: \mu_1 - \mu_2 < \Delta_0$	$H_a: \mu_1 - \mu_2 > \Delta_0$	$H_a: \mu_1 - \mu_2 \neq \Delta_0$

where Δ_0 is a specific difference. Usually, $\Delta_0 = 0$.

- Test statistic: $\boxed{T_0 = \dfrac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\dfrac{{S_1}^2}{n_1} + \dfrac{{S_2}^2}{n_2}}}}$ where $T_0 \sim t\left(\underbrace{\dfrac{\left(\frac{{S_1}^2}{n_1} + \frac{{S_2}^2}{n_2}\right)^2}{\frac{\left({S_1}^2/n_1\right)^2}{n_1 - 1} + \frac{\left({S_2}^2/n_2\right)^2}{n_2 - 1}}}_{df}\right)$

- Assumptions:
  1. The data are two independent random sample(s) drawn from two independent populations.
  2. The $\bar{X}_1$ and $\bar{X}_2$ dist's must be at least approximately normal.

TWO SAMPLES FROM TWO NORMAL POPULATIONS

Independent Samples

Known Pop. Variances

$$\bar{X}_1 - \bar{X}_2 \pm Z_{\alpha/2}\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}$$

$$Z_0 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}}$$

Forms

Unknown Pop. Variances

use s_1^2 s_2^2

$$\bar{X}_1 - \bar{X}_2 \pm T_{\alpha/2,}\sqrt{\frac{{S_1}^2}{n_1} + \frac{{S_2}^2}{n_2}}$$

$$T_0 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{{S_1}^2}{n_1} + \frac{{S_2}^2}{n_2}}}$$

$$\mathrm{df} = \frac{\left(\frac{{S_1}^2}{n_1} + \frac{{S_2}^2}{n_2}\right)^2}{\frac{\left({S_1}^2/n_1\right)^2}{n_1 - 1} + \frac{\left({S_2}^2/n_2\right)^2}{n_2 - 1}}$$

Dependent (Paired) Samples

Example: Abrasive Wear

An experiment was performed to compare the abrasive wear (in μm) of two laminated materials. Twelve pieces of material 1 and ten pieces of material 2 were tested by exposing each piece to a machine measuring wear. Use df=17.

1. Calculate a 95% CI for $\mu_1 - \mu_2$.

*Slide in Notes

Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
Sample 1	12	85.00	4.00	1.2
Sample 2	10	81.00	5.00	1.6

2. Interpret the 95% CI in the context of the problem.

> iClicker Question:

The 95% CI for $\mu_1 - \mu_2$ is ________________.
Can we conclude, at the 0.05 significance level, that the true mean abrasive wear of material 1 differs from that of material 2?

A. Yes, because 0 is in the CI.
B. Yes, because it is an experiment.
C. No, because 0 is in the CI.
D. No, because it is an observational study.

Example: Etching Solutions

In semiconductor manufacturing, wet chemical etching is often used to remove silicon from the backs of wafers prior to metallization. The etch rate, in mils per minute, is known to follow a normal distribution. Two etching solutions have been compared using two random samples of 10 wafers for each solution. Do the data support the claim that the mean etch rates are different for the two solutions? Use α=0.05.

Sample	N	Mean	StDev	SE Mean
Solution1	10	9.970	0.422	0.13
Solution2	10	10.400	0.231	0.073

Estimation for Difference

Difference	95% CI for Difference
-0.430	(-0.759, -0.101)

Test

Null hypothesis $H_0: \mu_1 - \mu_2 = 0$
Alternative hypothesis $H_1: \mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
-2.83	13	0.014

1. State the hypotheses.

2. Compute the test statistic.

3. Find the critical value(s).

Example: Etching Solutions (cont.)

3. Find the p-value.

4. What is the decision regarding H_0? Use α=0.05.

* add annotation [illegible]

5. State the conclusion in the context of the problem.

Dependent Samples (i.e. Matched Pairs)

	Sample 1	**Sample 2**	**Difference**
Pair 1	X_{11}	X_{21}	$D_1 = X_{11} - X_{21}$
Pair 2	X_{12}	X_{22}	$D_2 = X_{12} - X_{22}$
Pair 3	X_{13}	X_{23}	$D_3 = X_{13} - X_{23}$
	⋮	⋮	⋮
Pair n	X_{1n}	X_{2n}	$D_n = X_{1n} - X_{2n}$
SUMMARIES			$\boldsymbol{\bar{X}_D, S_D}$

subscript "D" stands for differences

STEP 1: Calculate the Difference for each pair

STEP 2: Calculate the summaries of the Differences ($\boldsymbol{\bar{X}_D, S_D}$)

- ✓ $\boldsymbol{\bar{X}_D}$ is the mean of the sample differences
- ✓ $\boldsymbol{S_D}$ is the standard deviation of the sample differences

STEP 3: Use the summaries of the Differences in the CI and test stat formulas.

CI and HT for Dependent Samples Pop. Variance Unknown

- Two-sided Confidence Interval for μ_D: $\boxed{\bar{X}_D \pm T_{\alpha/2} \frac{S_D}{\sqrt{n}}}$

- Hypotheses:

$$
\begin{array}{lll}
H_0\!: \mu_D = \Delta_0 & H_0\!: \mu_D = \Delta_0 & H_0\!: \mu_D = \Delta_0 \\
H_a\!: \mu_D < \Delta_0 & H_a\!: \mu_D > \Delta_0 & H_a\!: \mu_D \neq \Delta_0
\end{array}
$$

Δ_0 is a specific difference. Usually, $\Delta_0 = 0$.

- Test statistic: $\boxed{T_0 = \frac{\bar{X}_D - \Delta_0}{S_D/\sqrt{n}}}$

where $T_0 \sim t(n-1)$

n = size of each sample

- Assumptions:
  1. The sample(s) must be random sample(s).
  2. The $\bar{X}_D$ distribution must be at least approximately normal.

Example: Steel Girders

- Each pair of observations, say (X_{1j}, X_{2j}), is taken under homogeneous conditions, but these conditions may change from one pair to another.

Each girder has the "pair" of techniques applied under the same conditions: Is there a difference?

Table 10-2 Strength Predictions for Nine Steel Plate Girders (Predicted Load/Observed Load)

Girder	Karlsruhe Method	Lehigh Method	Difference d_j
S1/1	1.186	1.061	0.119
S2/1	1.151	0.992	0.159
S3/1	1.322	1.063	0.259
S4/1	1.339	1.062	0.277
S5/1	1.200	1.065	0.138
S2/1	1.402	1.178	0.224
S2/2	1.365	1.037	0.328
S2/3	1.537	1.086	0.451
S2/4	1.559	1.052	0.507

data to be used to find CI for μ_D

Example: Steel Girders (cont.)

~~Descriptive Statistics~~

Sample	N	Mean	StDev	SE Mean
~~K Method~~	~~9~~	~~1.3401~~	~~0.1460~~	~~0.0487~~
~~L Method~~	~~9~~	~~1.0662~~	~~0.0494~~	~~0.0165~~

9 diffs

Estimation for Paired Difference

Mean	StDev	SE Mean	90% CI for μ_difference (μ_D)
$\bar{x}_D =$ 0.2739	$s_D =$ 0.1351	~~0.0450~~	(0.1901, 0.3576)

µ_difference: mean of (K Method - L Method)

Test

Null hypothesis H₀: µ_difference = 0
Alternative hypothesis H₁: µ_difference ≠ 0

T-Value	P-Value
6.08	0.000

(a) Calculate a 90% CI for μ_D.

$$\bar{x}_D \pm t_{.05,8} \cdot \frac{s_D}{\sqrt{n}} = .2739 \pm 1.86\left(\frac{.1351}{\sqrt{9}}\right) = \boxed{(.19, .36)}$$

(b) Provide an interpretation of the 90% CI.

We are 90% confident the true mean strength prediction for the Karlsruhe method is between .19 and .36 units greater than that of the Lehigh method.

Example: Emissions

Eight vehicles were chosen at random and particulate matter emissions were measured under both highway driving and stop-and-go driving conditions. The differences (stop-and-go emission – highway emission) were computed and found: $\bar{x}_D$=190.5 and s_D=281.1. Can we conclude that the mean level of emissions is less for highway driving than for stop-and-go driving, using α=0.05?

(a) State the hypotheses.

$H_0: \mu_D = 0$

$H_a: \mu_D > 0$

Aside:

Diff = stop/go - Hwy

$\mu_D = \mu_{stop/go} - \mu_{Hwy}$

If Hwy is less, then

$\mu_D = \mu_{stop/go} - \mu_{Hwy} > 0$

(b) Compute the test statistic.

$$T_0 = \frac{\bar{x}_D - \Delta_0}{S_D/\sqrt{n}} = \frac{190.5 - 0}{281.1/\sqrt{8}} = 1.92$$

Example: Emissions (cont.)

(d) Find the critical value.

$t_{.05,7} = 1.895$ Reject Ho if $T_0 > 1.895$

(e) What is the decision regarding H_0? Use α=0.05.

Reject Ho

(f) State the conclusion in terms of the problem.

There is enough evidence to suggest the true mean emission level for highway driving is less than for stop and go driving.

Formulas:

TWO SAMPLES

Independent Samples

Known Pop. Variances

$$\bar{X}_1 - \bar{X}_2 \pm Z_{\alpha/2}\sqrt{\frac{\sigma_1^{\,2}}{n_1} + \frac{\sigma_2^{\,2}}{n_2}}$$

$$Z_0 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{\sigma_1^{\,2}}{n_1} + \frac{\sigma_2^{\,2}}{n_2}}}$$

Unknown Pop. Variances

$$\bar{X}_1 - \bar{X}_2 \pm T_{\alpha/2,}\sqrt{\frac{S_1^{\,2}}{n_1} + \frac{S_2^{\,2}}{n_2}}$$

$$T_0 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{S_1^{\,2}}{n_1} + \frac{S_2^{\,2}}{n_2}}}$$

$$\mathrm{df} = \frac{\left(\frac{S_1^{\,2}}{n_1} + \frac{S_2^{\,2}}{n_2}\right)^2}{\frac{\left(S_1^{\,2}/n_1\right)^2}{n_1 - 1} + \frac{\left(S_2^{\,2}/n_2\right)^2}{n_2 - 1}}$$

Dependent (Paired) Samples

Known Pop. Variance

σ_D^2

$$\bar{X}_D \pm Z_{\alpha/2}\frac{\sigma_D}{\sqrt{n}}$$

$$Z_0 = \frac{\bar{X}_D - \Delta_0}{\sigma_D/\sqrt{n}}$$

Unknown Pop. Variance

S_D^2

$$\bar{X}_D \pm T_{\alpha/2}\frac{S_D}{\sqrt{n}}$$

$$T_0 = \frac{\bar{X}_D - \Delta_0}{S_D/\sqrt{n}}$$

$$\mathrm{df} = n - 1$$

Name: ______________________________

EGEN350 Worksheet 15 (Chapter 10)

1. A clinical trial is investigating whether a new procedure to repair a torn Anterior Cruciate Ligament (ACL), called Bridge-Enhanced ACL Reconstruction (BEAR) (a technique that uses a tissue-engineered scaffold with the patient's own cells) leads to better clinical outcomes than the traditional ACL reconstruction (ACLR). An important outcome is "Effusion Score", which is an indication of how swollen the knee is (higher score is worse). In the study, Effusion Score was measured at 1 week after surgery and again at 3 months after surgery. This was an early clinical trial, therefore only 10 patients were studied in each of the BEAR and ACLR groups. Data adapted from Murray et al., OJSM, 2016.

			Effusion Score at 1 week		Effusion Score at 3 months		Effusion Score Difference (3 month score – 1 week score)	
	Sample size		Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
BEAR	10		1.8	0.4	0.5	0.5	-1.3	0.4
ACLR	10		1.9	1.0	0.9	0.6	-1.0	0.5

a. Is there evidence to suggest that there is significant improvement in Effusion Score, on average, for the BEAR procedure?

i. State the hypotheses.

ii. Calculate the test statistic.

iii. Calculate the p-value.

iv. Make a decision regarding H_0 using the 0.05 significance level.

- Reject H_0
- Fail to Reject H_0

v. State the conclusion in the context of the problem.

Name: ____________________________________

Table (shown again)

			Effusion Score at 1 week		Effusion Score at 3 months		Effusion Score Difference (3 month score – 1 week score)	
	Sample size		Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
BEAR	10		1.8	0.4	0.5	0.5	-1.3	0.4
ACLR	10		1.9	1.0	0.9	0.6	-1.0	0.5

b. Is there evidence to suggest that Effusion Score is less at 3 months for the BEAR procedure, on average, than for the ACLR procedure?

i. State the hypotheses.

ii. Calculate the test statistic.

iii. Find the critical value using α = 0.05 and df=17. Reject H_0 if ________________________

iv. Make a decision regarding H_0 using the 0.05 significance level.

- ○ Reject H_0
- ○ Fail to Reject H_0

v. State the conclusion in terms of the problem.

Summary

- <u>Independent samples</u>: Two samples for which the individuals or objects in the first sample are selected independently from those in the second sample.
  - Parameter $\mu_1 - \mu_2$
  - Estimate: $\bar{X}_1 - \bar{X}_2$
  - Standard deviation of estimate: $\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
  - Standard error of estimate: $SE_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$

- <u>Dependent (paired) samples</u>: Two samples for which each observation in one sample is paired in a meaningful way with a particular observation in the second sample. The difference (D) in paired observation are found.
  - Parameter: μ_D
  - Estimate: $\bar{X}_D$
  - Standard deviation of estimate: $\sigma_{\bar{X}_D} = \frac{\sigma_D}{\sqrt{n}}$
  - Standard error of estimate: $SE_{\bar{X}_D} = \frac{S_D}{\sqrt{n}}$

Name: ____

EGEN350 Worksheet 16 (Chapter 10)

Two-sample and Matched-pairs Confidence Interval Interpretations

Confidence Interval Interpretations must include:

- level of confidence
- true mean (i.e., pop. mean)
- the response variable (in context)
- which pop. mean is larger / smaller than the other
- and by how much

1. Suppose we are interested in comparing the true mean battery lifetime (in hours) of brand A batteries versus brand B batteries.

   Interpret the 90% CI for $\mu_A - \mu_B$ = (-19, 28).

2. Estimate the difference in mean weight (in pounds) before and after an exercise program.
   Difference = Before - After

   Interpret the 95% CI for μ_D = (1.1, 3.6)

3. An article provides data on the absorbency rate (in g/sec) of paper towels that were produced by two different manufacturing processes (Process 1 vs. 2). Estimate the difference in the mean absorbency rate of the paper towels from process 1 vs. process 2.

   Interpret the 99% CI for $\mu_1 - \mu_2$ = (-8.1, -2.5).

Name: ____________________________________

Two-sample and Matched-pairs Hypothesis Test Conclusions

Hypothesis Test Conclusions must include:

- the amount of evidence (There is OR is not enough evidence)
- true mean (pop. mean)
- the response variable (in context)
- which pop. mean was greater / less than / differs (direction in H_a) from the other
- If there is evidence for a difference ($\neq$ H_a), then write a second sentence which states "The data suggest …" and complete the sentence with which mean was found to be larger / smaller than the other (determined by examining $\bar{X}_1$ and $\bar{X}_2$).

4. Two different analytical tests are used to determine the impurity level in steel alloys. Eight specimens are test using both procedures. Is there sufficient evidence to conclude that tests differ in the mean impurity level?

Statistics

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Ex10-56 Test1	8	0	1.4500	0.0732	0.2070	1.2000	1.3000	1.4000	1.6500	1.8000
Ex10-56 Test2	8	0	1.6625	0.0981	0.2774	1.3000	1.4250	1.6500	1.9250	2.1000
Ex10-56 Diff	8	0	-0.2125	0.0611	0.1727	-0.4000	-0.3000	-0.3000	-0.0500	0.1000

Decision: Reject H_0

Conclusion in the context of the problem:

5. An article presents data on cycles to failure of solder joints at different temperatures for different types of printed circuit boards (PCB). Failure data for two temperatures (20 and 60°C) for a copper-nickel-gold PCB are measured. Test the hypotheses: $H_0: \mu_{20°C} = \mu_{60°C}$ versus $H_a: \mu_{20°C} > \mu_{60°C}$

Decision: Fail to Reject H_0

Conclusion in the context of the problem:

Used to compare 3 or more population means

EGEN350

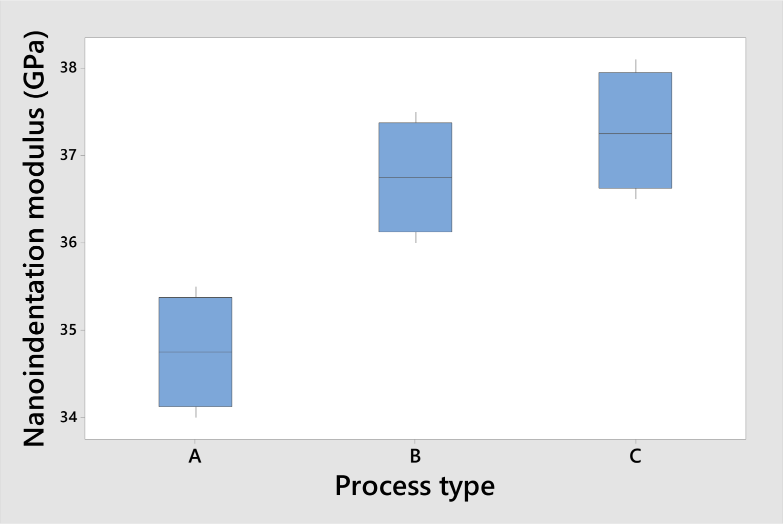
Chapter 13

Analysis of Variance

(Section 13-2 Only)

Comparing 3 (or more) population means

- Calcium carbonate crystals were made by three separate processes (A, B, C) and assessed using nanoindentation for modulus. (Modulus indicates stiffness.)
- How could we test if there is an effect of process type on nanoindentation modulus?



Should we just use multiple (3) t-tests: A vs. B, A vs. C, and B vs. C?

Should we just use multiple (3) t-tests: A vs. B, A vs. C, and B vs. C?

- No! Each time we conduct a t-test, we introduce an opportunity for Type I error. "Multiple testing problem"

"given"

Suppose $\alpha = 0.05$.
For 1 test: P(FTR H_0 | H_0 is true) = 0.95 &
P(Type I Error) = P(Reject H_0 | H_0 is true) = 0.05

For 3 tests (A vs B, B vs C, A vs C): P(all FTRs | all H_0's are true) = $(0.95)^3$ = 0.857 &
P(at least 1 Reject H_0 | all H_0's are true) = 1 - $(0.95)^3$ = 0.143

≥ 1

- Therefore, even if all 3 H_0's are true, we have a **14.3%** chance of making at least one Type I Error.

- We do not want to increase the chance of making a Type I Error!

- So, if you have 3 (or more) population means, use ANOVA, not t-tests!

"Multiple Testing Problem":
If you run 100 tests and all H_0's are true,
then you'll incorrectly Reject H_0 in the 5% of tests.

Analysis of Variance (ANOVA)

- ANOVA compares "a" population means.

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_a$$
$$H_a: \text{ not all means are equal}$$

→ do not say "all means differ"

- ANOVA is also known as "One-way ANOVA" or "Single-Factor ANOVA" because there is a single categorical variable (factor) whose categories define the populations being compared.

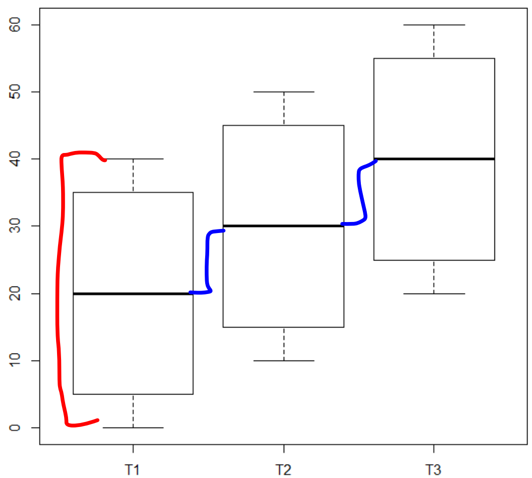
- For an experiment, the "treatments" define the populations being compared.

The Main Idea

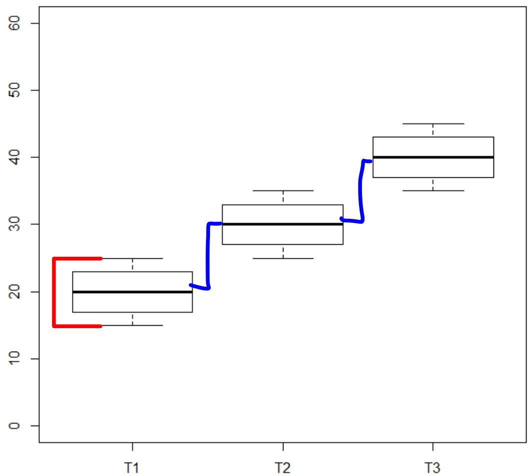
We will analyze the total variability to see:

1) How much is attributed to differences in treatment means?
2) How much is attributed to the variability of values within each treatment?

Is the variability between treatments small (or large) relative variability within treatments?



The variability between treatments is **small** compared to the variability within treatments



The variability between treatments is **large** compared to the variability within treatments

Example: Tensile Strength

A manufacturer of paper used for making grocery bags is interested in improving the tensile strength of the product. Product engineering thinks that tensile strength is a function of the hardwood concentration in the pulp and that the range of hardwood concentrations of practical interest is between 5% and 20%. A team of engineers responsible for the study decides to investigate four levels of hardwood concentration: 5%, 10%, 15%, and 20%. Use α=0.05.

Hardwood	**Observations**							
Concentration (%)	1	2	3	4	5	6	Totals	Averages
5	7	8	15	11	9	10	60	10.00
10	12	17	13	18	19	15	94	15.67
15	14	18	19	17	16	18	102	17.00
20	19	25	22	23	18	20	127	21.17
							383	15.96

<u>Response variable</u>: Tensile Strength

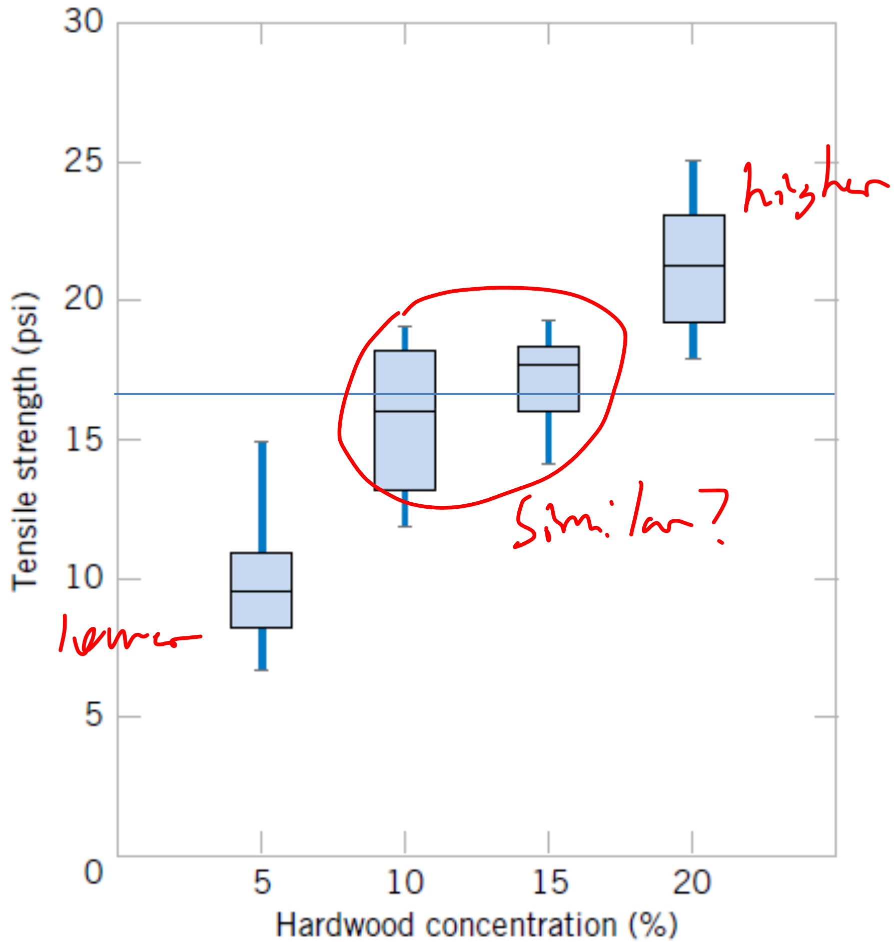
<u>Factor (Categorical explanatory variable)</u>: Hardwood concentration

<u>Factor Levels (Categories)</u>: 5%, 10%, 15%, and 20%

<u>Replicates</u>: 6 observations (replicates) were used at each factor level

Example: Tensile Strength (cont.)

- Does hardwood concentration have a significant effect on the tensile strength of the bag? Look at a plot:



Make a Guess:

Which means appear similar / different?

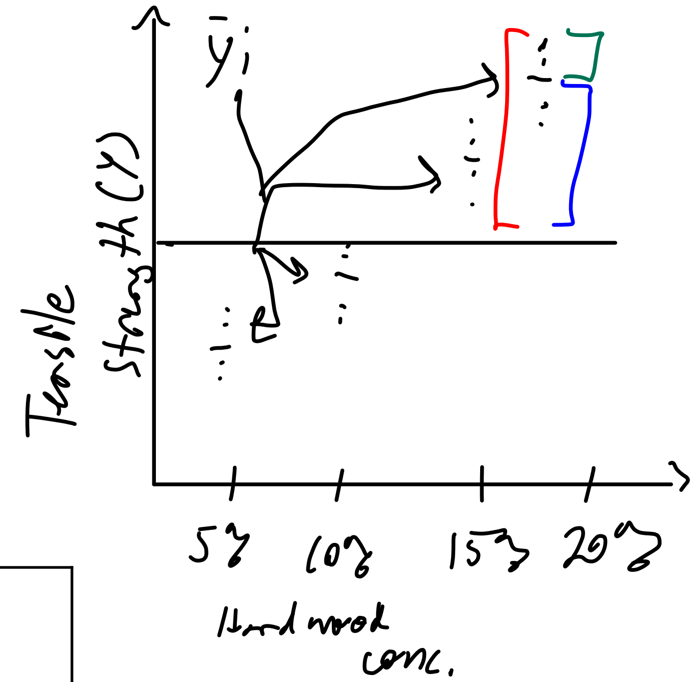
Partition the Variability

Recall:

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

← like "SS"

← like "DF"



Sources of Variation	Sum of Squares
Treatments	$SS_{Treatments} = n \sum_{i=1}^{a} (\bar{y}_{i.} - \bar{y}_{..})^2$
Error	$SS_E = \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.})^2$
Total	$SS_T = \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{..})^2$

"Variability among Treatments"
Deviations of treatment means from overall mean

"Variability within Treatments"
Deviations of observations from treatment means

"Total Variability"
Deviations of observations from overall mean

Where a = # of trts; n = sample size; y_{ij} = observation; $\bar{y}_{i.}$ = trt mean; $\bar{y}_{..}$ = overall mean

ANOVA Table

$F = t^2$

test stat.

ANOVA Table:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square $= \frac{SS}{DF}$	F_0
Treatments	$SS_{\text{Treatments}}$	$a - 1$	$MS_{\text{Treatments}}$	$\dfrac{MS_{\text{Treatments}}}{MS_E}$
Error	$+ SS_E$	$- a(n-1)$	MS_E	
Total	SS_T	$an - 1$		

ANOVA Table For Our Example:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	f_0
Hardwood concentration	382.79	3	127.60	19.60
Error	130.17	20	6.51	
Total	512.96	23		

Estimates of Pop. Variance:

(s^2)

- MS_E is always an unbiased estimator of σ^2.
- If H_0 is true, then $MS_{Treatments}$ and MS_E are an unbiased estimators of σ^2.

F Test Statistic:

Treatment-to-treatment variability

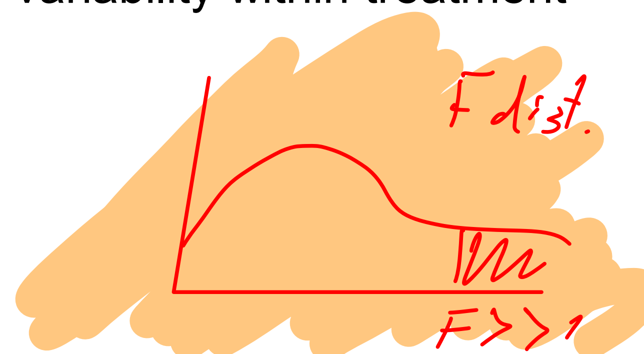
- If H_0 is true, $F_0 = \dfrac{MS_{Treatments}}{MS_E} \approx 1$

Natural variability within treatment

- If H_0 is false, $F_0 = \dfrac{MS_{Treatments}}{MS_E} \gg 1$

  Ha is true

  where $F_0 \sim F(\text{df}_{Tr}, \text{df}_E)$



- Large $F_0 = \dfrac{MS_{Treatments}}{MS_E}$ values are strong evidence <u>against</u> H_0 and <u>for</u> H_a.

  Think of this as high “signal-to-noise”.

What does ANOVA table look like in Minitab?

One-way ANOVA: Conc5, Conc10, Conc15, Conc20

Method

Null hypothesis	All means are equal
Alternative hypothesis	Not all means are equal
Significance level	α = 0.05

at least one μ differs

Equal variances were assumed for the analysis.

Reject Ho

Factor Information

Factor	Levels	Values
Factor	4	Conc5, Conc10, Conc15, Conc20

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Factor	3	382.8	127.597	19.61	0.000
Error	20	130.2	6.508		
Total	23	513.0			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2.55114	74.62%	70.82%	63.46%

S estimates σ

S^2 = MSE esitmates σ^2

Interpretation:
There is a sufficient evidence that the true mean tensile strength of grocery bags differs for at least one hardwood concentration.

Our ANOVA result does NOT tell us which hardwood concentrations have higher tensile strengths than other concentrations. To identify which treatments differ from other treatments, we need “Follow-up tests”.

Multiple Comparisons
AKA: “Follow-up Tests, or Post-hoc tests”

- If we Reject H_0 in the overall test and conclude at least 2 means are different, then perform follow-up tests to answer:

  “Which population means are different?”

- The follow-up tests are all pairwise comparisons of the means.

Control the Type I Error Rate

- The problem with performing a large number of tests is that, even if all H_0's are true, at least one of the tests will have a significant result purely by chance.
- There are several methods to control this false discovery rate. (Tukey's, Bonferroni's, Scheffe's, etc.)
- For 1 test: P(Type I Error) = P(Reject H_0 | H_0 is true) = 0.05
  P(FTR H_0 | H_0 is true) = 0.95
- For a family of 6 tests (all pairwise comparisons of 4 pop. means):
  P(all FTRs | all H_0's are true) = $(0.95)^6$ = 0.735 [assuming independence]
  P(at least 1 Reject H_0 | all H_0's are true) = 1 - $(0.95)^6$ = 0.265
- Therefore, even if all 6 H_0's are true, we have a 26.5% chance of making at least one Type I Error.
- Rather than using a 0.05 significance level for each test, we can set the family-wise error rate to 0.05, where "family" denotes the group of all multiple comparisons.

Example: Tukey's Multiple Comparisons

Tukey Simultaneous Tests for Differences of Means

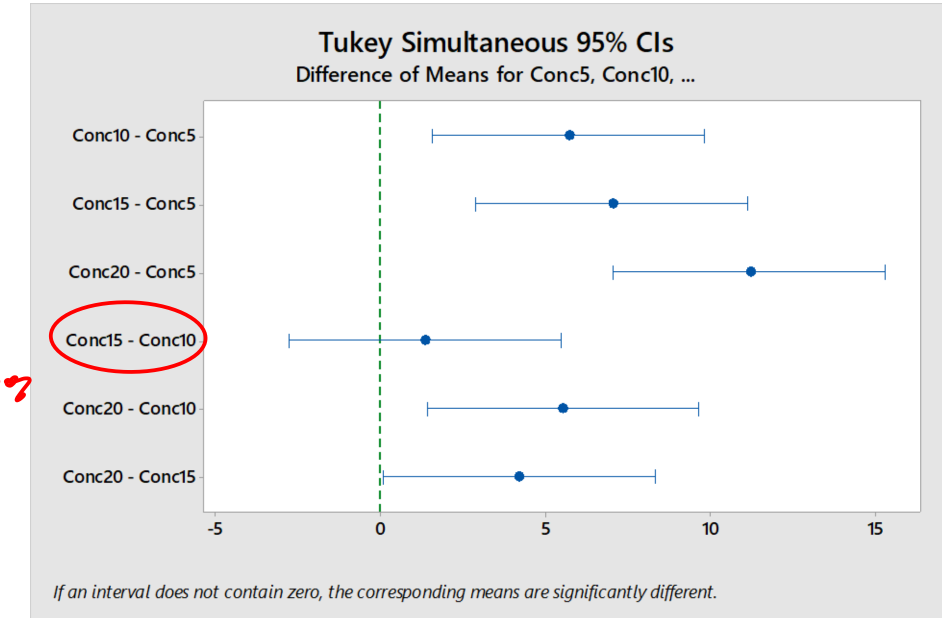
Difference of Levels	Difference of Means	SE of Difference	95% CI	T-Value	Adjusted P-Value
Conc10 - Conc5	5.67	1.47	(1.54, 9.79)	3.85	0.005
Conc15 - Conc5	7.00	1.47	(2.88, 11.12)	4.75	0.001
Conc20 - Conc5	11.17	1.47	(7.04, 15.29)	7.58	0.000
Conc15 - Conc10	1.33	1.47	$\boxed{(-2.79, 5.46)}$	$\boxed{0.91}$	$\boxed{0.802}$
Conc20 - Conc10	5.50	1.47	(1.38, 9.62)	3.73	0.007
Conc20 - Conc15	4.17	1.47	(0.04, 8.29)	2.83	0.047

Individual confidence level = 98.89%

$H_0: \mu_{15g} - \mu_{10g} = 0$

$H_a: \mu_{15g} - \mu_{10g} \neq 0$

→ CI for $\mu_{15g} - \mu_{10g}$ contains 0 so μ_{15g} and μ_{10g} are not significantly different.



Tukey Pairwise Comparisons

Grouping Information Using the Tukey Method and 95% Confidence

Factor	N	Mean	Grouping		
Conc20	6	21.17	A		
Conc15	6	17.000		B	
Conc10	6	15.67		B	
Conc5	6	10.00			C

decreasing order

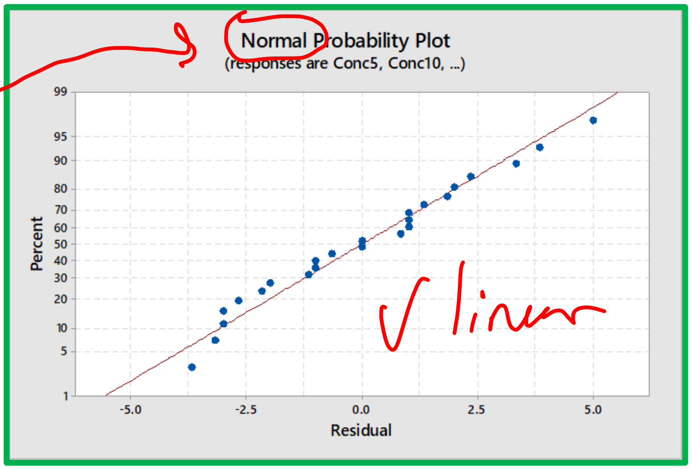
Means that do not share a letter are significantly different.

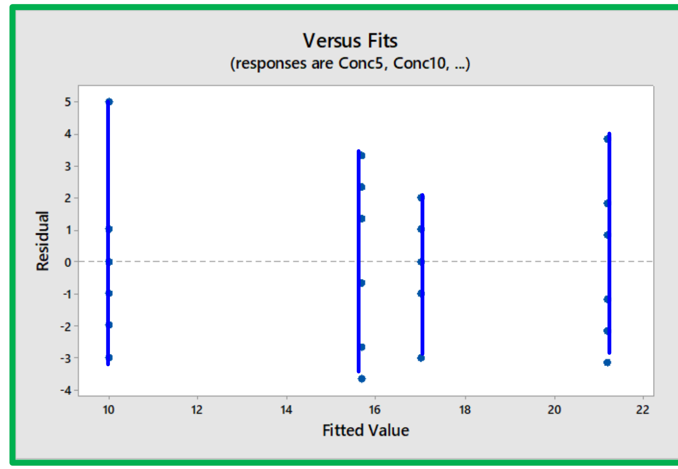
*check annotated notes for extra Q's

Assumptions

1. Independent random samples from each population.
2. Each population distribution is normal.
3. Population variance is the same for all populations.

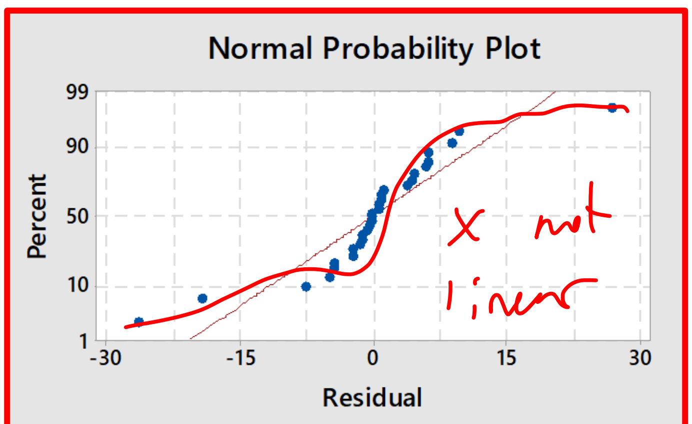
Good:

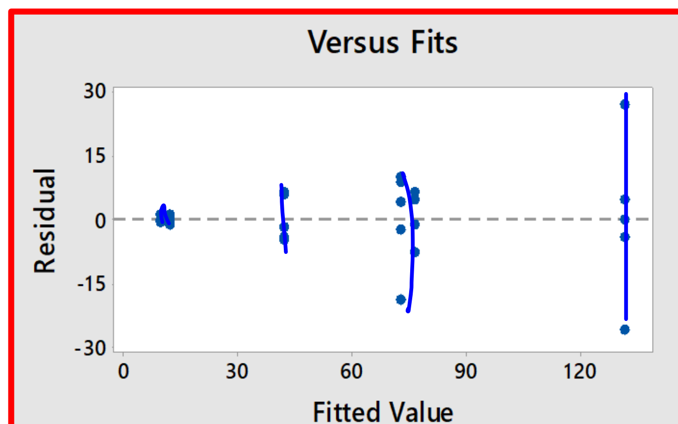




✓ similar spread

Bad:





✗ dissimilar spread

> iClicker Question:

If 50 t-tests were performed, each at $\alpha = 0.05$, then:

$$P(at\ least\ 1\ Type\ I\ Error\ |all\ H_0's\ are\ true) = ?$$

> iClicker Question:

When H_0 is true, I expect ___ test statistic & ___ p-value.

When H_a is true, I expect ___ test statistic & ___ p-value.

A. small, large and large, small
B. large, small and small, large

Summary

- The "overall test" of a one-way ANOVA is a procedure to compare the means of 3 or more populations.
- The hypotheses of the overall test are:

$$H_0\text{: } \mu_1 = \mu_2 = \cdots = \mu_a$$

$$H_a\text{: not all means are equal}$$

- The assumptions for the overall test are:

  (1) Independent random samples from each population.

  (2) Each population distribution is normal.

  (3) Population variance is the same for all populations.

- Estimates of population variance, σ^2, (i.e. estimates of variability within treatment):

  (1) MS_E is an unbiased estimator of σ^2.

  And if H_0 is true, then:

  (2) MS_{Trt} is also an unbiased estimator of σ^2.

Summary

- The F_0 test statistic compares the variability between treatments (numerator) to the variability within treatments (denominator):
  - If H_0 is true, $F_0 = \frac{MS_{Trt}}{MS_E} \approx 1$ where $F_0 \sim F(\text{df}_{Trt}, \text{df}_E)$.
  - If H_0 is false, $F_0 = \frac{MS_{Trt}}{MS_E} \gg 1$ where $F_0 \sim F(\text{df}_{Trt}, \text{df}_E)$.
  - Large $F_0 = \frac{MS_{Trt}}{MS_E}$ values are strong evidence <u>against</u> H_0 and <u>for</u> H_a.
- "Is the variability within treatments small enough to detect differences between treatments?" The answer is yes, if the p-value beyond F_0 is smaller than α, and therefore we Reject H_0 in favor of H_a.
- If we Reject H_0 in the overall test and conclude at least 2 means are different, then perform follow-up tests to answer the question: "Which population means are different?"
- The follow-up analysis examines all pairwise comparisons of the means.
- Use Tukey's multiple comparisons to control the family-wise false discovery rate (α), where the family includes all pairwise comparisons.
- If 0 is in the CI for $\mu_i - \mu_j$ then the two means are not significantly different. If 0 is NOT in the CI for $\mu_i - \mu_j$ then the two means are significantly different.

So Far: Require ① $X \sim N$ or ② $\bar{X} \sim N$ via CLT

What if your sample sizes are small and your data are not normal? Try:

1) Non-parametric Test
2) Transform your data to be normal, using ln(X), log(X), $\sqrt{X}$, $\frac{1}{X}$, etc.

$\bar{X} \sim N\left(\underbrace{\mu, \frac{\sigma^2}{n}}_{\text{parameters}}\right)$

EGEN350

Non-parametric tests — Alternatives to:

Mann-Whitney — Two-sample t-test

One-sample Wilcoxon — Matched-pairs t-test

Kruskal-Wallis — One-way ANOVA

Parametric tests: Assume that $\bar{X}$'s are at least approximately normally distributed.

- Ex: One-sample, two-sample
  Z and T tests, ANOVA
  (everything we've learned so far)

Nonparametric tests: 'Distribution-free' tests involve the median instead of the mean. Used when a small samples is drawn from non-normal populations.

- Ex: Mann-Whitney, One-sample Wilcoxon, Kruskal-Wallis (and others)

Two-sample Tests

- Independent Samples
  - $\bar{X}_1$ and $\bar{X}_2$ are $\dot{\sim}\ N$
    - Yes
      - Two-sample t-test (Chapter 10)
    - **No**
      - **Mann-Whitney Test**
- Dependent (Paired) Samples
  - $\bar{X}_D$ is $\dot{\sim}\ N$
    - Yes
      - Paired t-test (Chapter 10)
    - **No**
      - **One-sample Wilcoxon Test**

Two-sample Tests

Independent Samples

$\bar{X}_1$ and $\bar{X}_2$ are $\dot{\sim} N$

Yes → Two-sample t-test (Chapter 10)

No → **Mann-Whitney Test**

Mann-Whitney Test

Overview:

eta

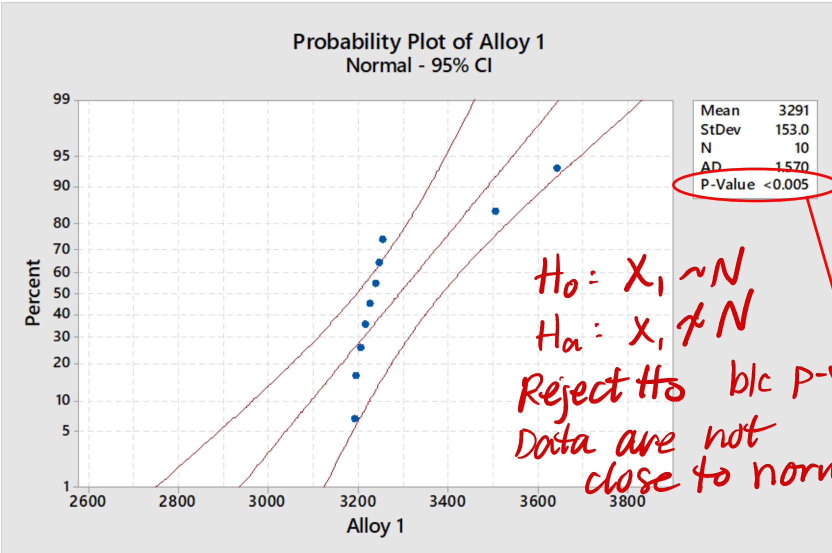
H_0: $\eta_1 = \eta_2$ $\quad$ *η = pop. median*

H_a: $\eta_1 \boxed{\begin{matrix}<\\>\\\neq\end{matrix}} \eta_2$ choose one

- Independent random samples are drawn from non-normal populations.
- Test is based on the *ranks* of observations. Thus, no requirement that $\bar{X}_1$ and $\bar{X}_2$ are normally distributed.
- Relevant for small samples.
- If sample sizes are ≥ 30, use CLT, and two-sample t-test.

Example: Motivation for Mann-Whitney Test

Check for Normality …

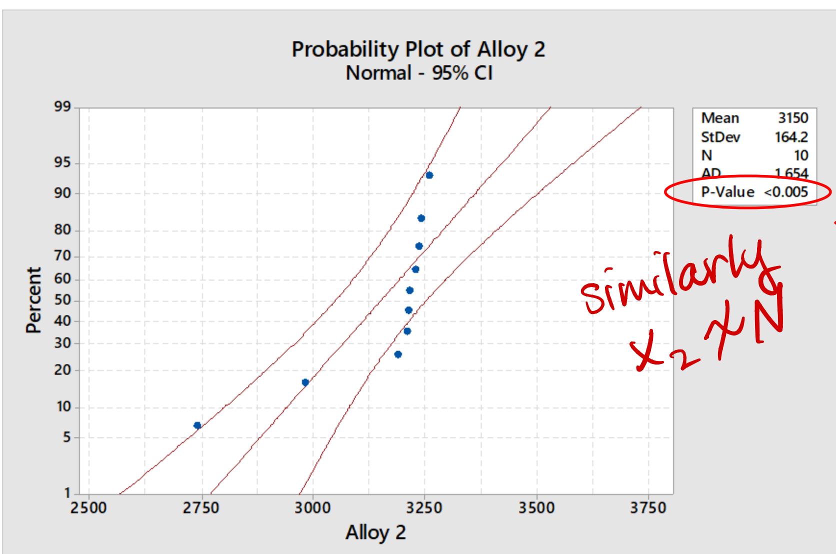


$H_0: X_1 \sim N$

$H_a: X_1 \nsim N$

Reject H_0 b/c p-value $< \alpha$

Data are not close to normal



similarly $X_2 \nsim N$

Axial stress (psi) for Alloy 1	Axial stress (psi) for Alloy 2
3238	2980
3195	3187
3246	3209
3190	3212
3204	3258
3254	2740
3640	3215
3225	3226
3504	3240
3241	3234

Neither X_1 nor X_2 are normally distributed; sample sizes are small so CLT cannot be invoked.

Example: Mann-Whitney Test

Axial stress (psi) for Alloy 1	Axial stress (psi) for Alloy 2
3238	2980
3195	3187
3246	3209
3190	3212
3204	3258
3254	2740
3640	3215
3225	3226
3504	3240
3241	3234

$n = 10$ $\quad$ $m = 10$

(1) Hypotheses: H_0: $\eta_1 = \eta_2$
H_a: $\eta_1 \neq \eta_2$

(2) Test statistic:

W = sum of *ranks* of the first sample.
(Which sample is first is arbitrary).
See next slide to learn how to calculate ranks for samples.

(3) Determine p-value

(4) Make a decision based on α

(5) State the conclusion

Example: Mann-Whitney Test

- Combine all observations from both samples
- Sort observations by rank

Axial stress (psi) for Alloy 1	Axial stress (psi) for Alloy 2
3238	2980
3195	3187
3246	3209
3190	3212
3204	3258
3254	2740
3640	3215
3225	3226
3504	3240
3241	3234

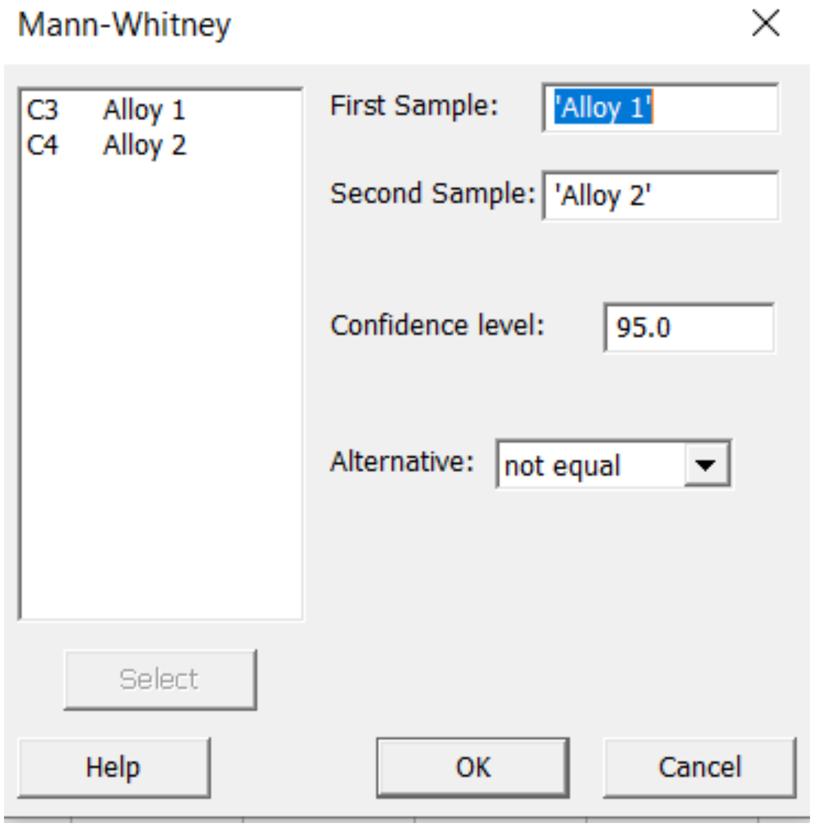
Alloy	Axial stress (psi)	Rank
2	2740	1
2	2980	2
2	3187	3
1	3190	4
1	3195	5
1	3204	6
2	3209	7
2	3212	8
2	3215	9
1	3225	10
2	3226	11
2	3234	12
1	3238	13
2	3240	14
1	3241	15
1	3246	16
1	3254	17
2	3258	18
1	3504	19
1	3640	20

increasing order

Sum = 125

Example: Mann-Whitney Test in Minitab

Stat → Nonparametrics



Note: *One*-tailed tests are allowable, although the two-tailed test is shown here.

Mann-Whitney: Alloy 1, Alloy 2

Method

η_1: median of Alloy 1
η_2: median of Alloy 2
Difference: $\eta_1 - \eta_2$

Descriptive Statistics

Sample	~~N~~ n	Median $\tilde{x}$
Alloy 1	10	3239.5
Alloy 2	10	3213.5

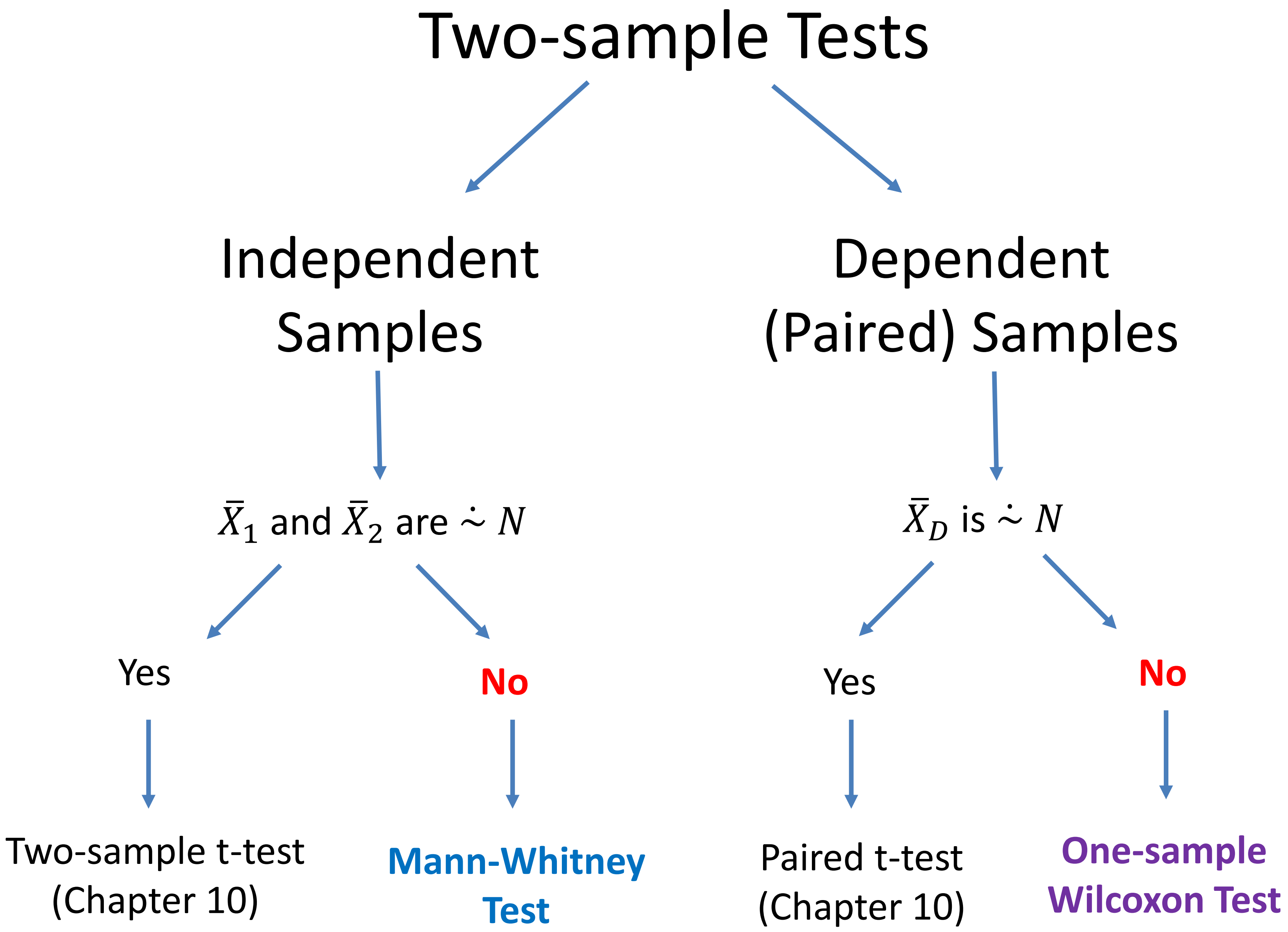
Estimation for Difference

Difference	CI for Difference	Achieved Confidence
28.5	(-11, 289)	95.48%

Test

Null hypothesis	$H_0: \eta_1 - \eta_2 = 0$
Alternative hypothesis	$H_1: \eta_1 - \eta_2 \neq 0$

W-Value	P-Value
125.00	0.140


Two-sample Tests
Independent Samples
Dependent (Paired) Samples
$\bar{X}_1$ and $\bar{X}_2$ are $\dot{\sim} N$
$\bar{X}_D$ is $\dot{\sim} N$
Yes
No
Yes
No
Two-sample t-test (Chapter 10)
Mann-Whitney Test
Paired t-test (Chapter 10)
One-sample Wilcoxon Test

One-sample Wilcoxon (Signed-Rank) Test

Overview:

Difference

H_0: $\eta_D = 0$

H_a: $\eta_D \boxed{<, >, \neq} 0$ choose one

where η_D = pop. median of the differences in paired responses

- Test is based on *ranks* of observations. Thus, no requirement that $\bar{X}_D$ is normally distributed.
- Tests the sign and rank of differences.
- Relevant for small samples.

Two-sample Tests

Dependent (Paired) Samples

$\bar{X}_D$ is $\dot{\sim} N$

Yes → Paired t-test (Chapter 10)

No → **One-sample Wilcoxon Test**

Example: One-sample Wilcoxon (Signed-Rank) Test

= Post - Pre

Pre	Post	Difference
10	12	2
8	11	3
5	15	10
3	3	0
4	8	4
8	6	-2
11	12	1
13	13	0
8	9	1
9	10	1

(1) Hypotheses: H_0: $\eta_D = 0$
H_a: $\eta_D > 0$

(2) Test statistic:

W = sum of positive *ranks*

See next slide to learn how to calculate signed-ranks.

(3) Determine p-value

(4) Make a decision based on α

(5) State the conclusion.

Example: One-sample Wilcoxon (Signed-Rank) Test

ignore 0's

Absolute Differences (ordered)		Rank	Signed Rank
~~0~~			
~~0~~			
1	1	2	2
1	2	2	2
1	3	2	2
2	4	4.5	4.5
-2	5	4.5	-4.5
3	6	6	6
4	7	7	7
10	8	8	8
		Sum of **positive** ranks: W = 31.5	

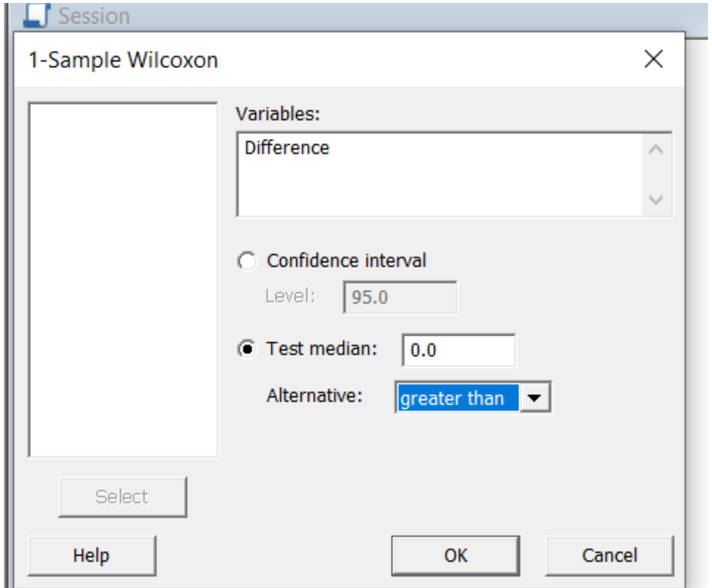
Ties!

Assign all average=2

4.5

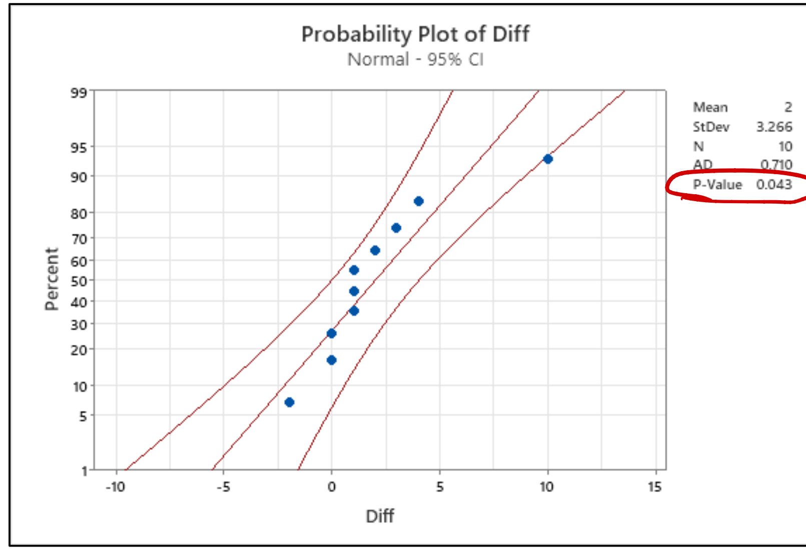
- Differences of 0 are not ranked.
- If 2 or more values share a rank, then all receive the average of the ranks. (Ex: The three values of 1 share ranks 1, 2, 3; so all receive the average rank of 2.)
- Test statistic: W = sum of **positive** ranks

Example: One-sample Wilcoxon Test in Minitab



Worksheet 4 ***

↓	C1	C2	C3	C4	C5
	Pre	Post	Difference		
1	10	12	2		
2	8	11	3		
3	5	15	10		
4	3	3	0		
5	4	8	4		
6	8	6	-2		
7	11	12	1		
8	13	13	0		
9	8	9	1		
10	9	10	1		



Wilcoxon Signed Rank Test: Difference

Method

η: median of Difference

Descriptive Statistics

Sample	N	Median
Difference	10	1.5

Test

Null hypothesis	H_0: η = 0
Alternative hypothesis	H_1: η > 0

Reject Ho conclude Ha

Sample	N for Test	Wilcoxon Statistic	P-Value
Difference	8	$\boxed{31.50}$	$\boxed{0.034}$

Test comparing 3 or more groups

Normality assumptions satisfied

One-way ANOVA (Chapter 13)

Normality assumptions NOT satisfied (but groups have similar distributions and variance)

there are some conditions to be met.

Kruskal – Wallis test

Test comparing 3 or more groups

Kruskal-Wallis Test

Overview:

H_0: $\eta_1 = \eta_2 = \eta_3 \ = \cdots$
H_a: not all pop. medians are equal

Test is based on *ranks*. Thus, no requirement that $\bar{X}$'s are normally distributed.

≥ 3

- Appropriate for 2+ samples.
- Relevant for small samples.
- NOT appropriate for samples with different distributions.
- NOT appropriate for samples with different variances.

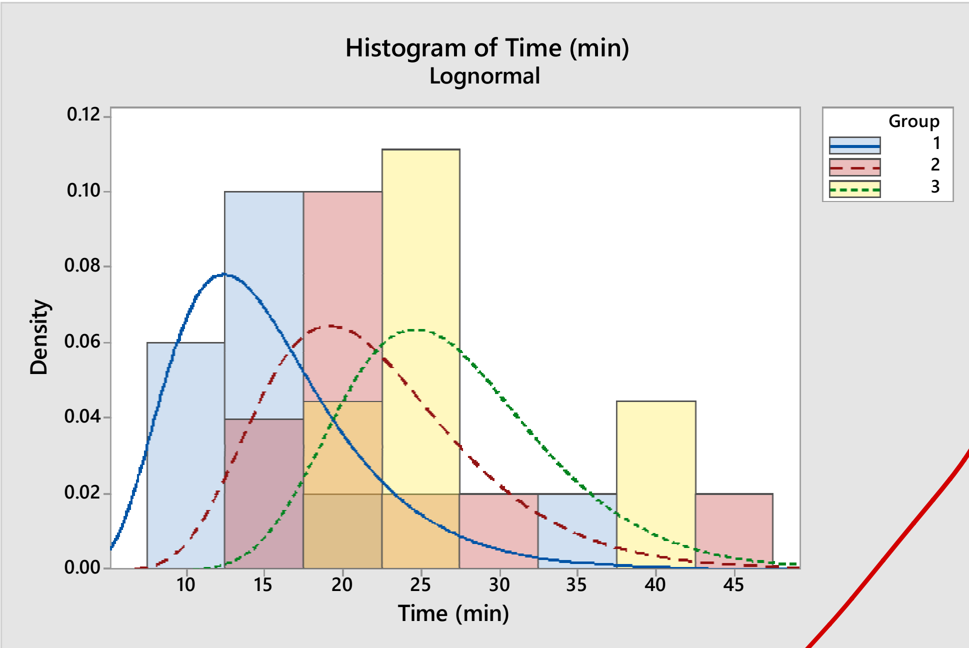
Need similar shape dist's

Need similar variances (spreads)

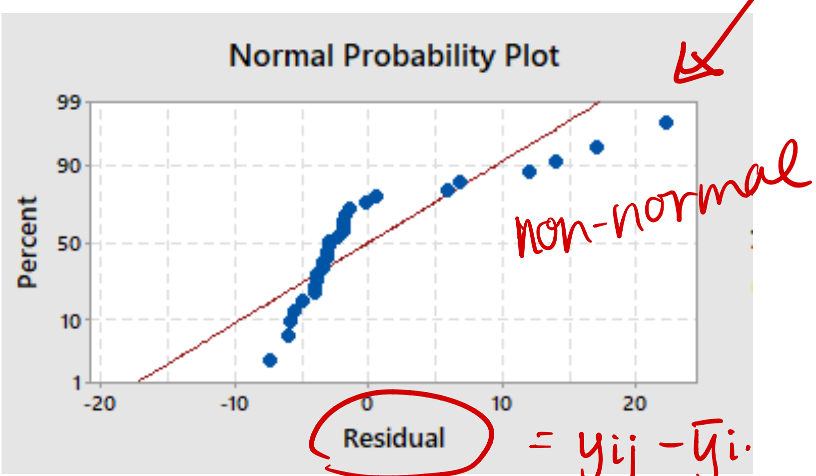
Normality assumptions NOT satisfied (but groups have similar distributions and variance)

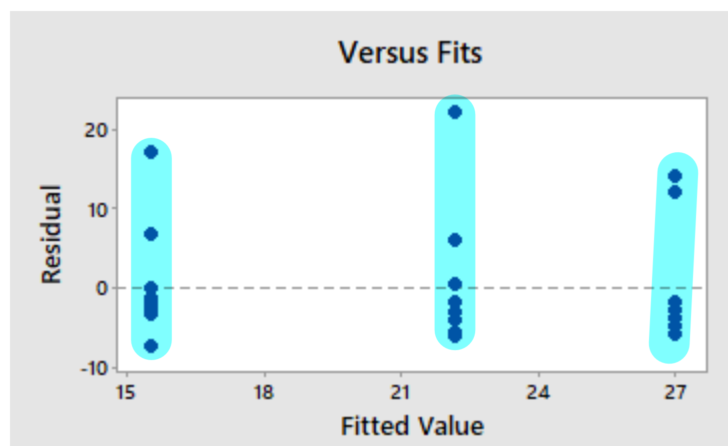
Kruskal – Wallis test

Example: Right-skewed Data



- All three groups are right-skewed.
- Diagnostic plots from ANOVA fail normality assumption; but there is reasonably similar spread of residuals for each group.

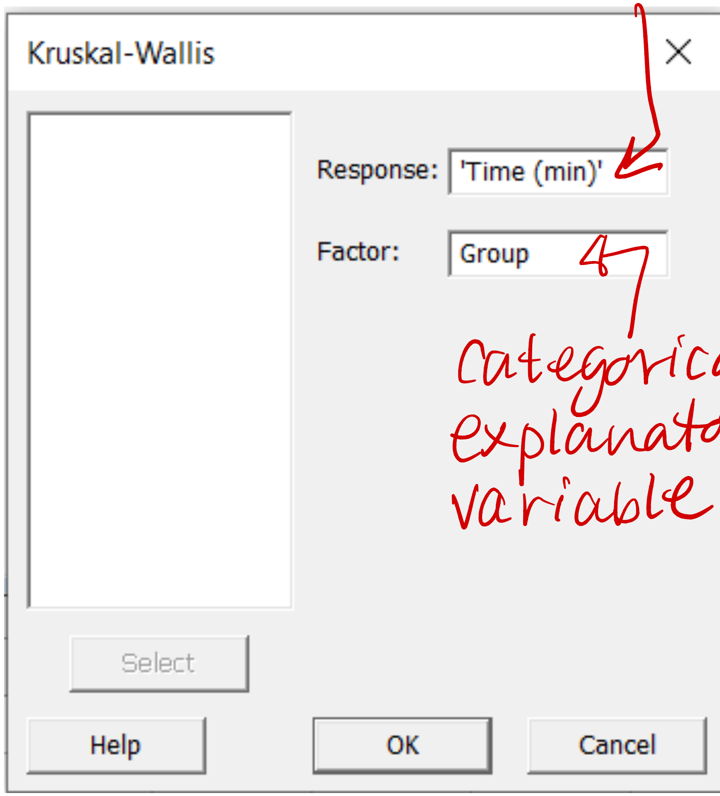




similar spread

Example: Kruskal-Wallis Test in Minitab

Quantitative Response



Categorical explanatory variable

Kruskal-Wallis Test: Time (min) versus Group

Descriptive Statistics

Group	N	Median	Mean Rank	Z-Value
1	10	13.25	8.0	-3.21
2	10	19.00	15.7	0.32
3	9	24.00	22.0	2.97
Overall	29		15.0	

Test

Null hypothesis	H_0: All medians are equal
Alternative hypothesis	H_1: At least one median is different

η_1 vs. η_2
η_1 vs. η_3
η_2 vs. η_3

Method	DF	H-Value	P-Value
Not adjusted for ties	2	12.91	0.002
Adjusted for ties	2	12.92	0.002

Reject H_0
Follow-up analysis : multiple mann-whitney tests on pairwise diff.

Note: The K-W test uses a Chi-squared distribution for determining p-values. If $n < 5$ per group, do not use the Kruskal-Wallis test.

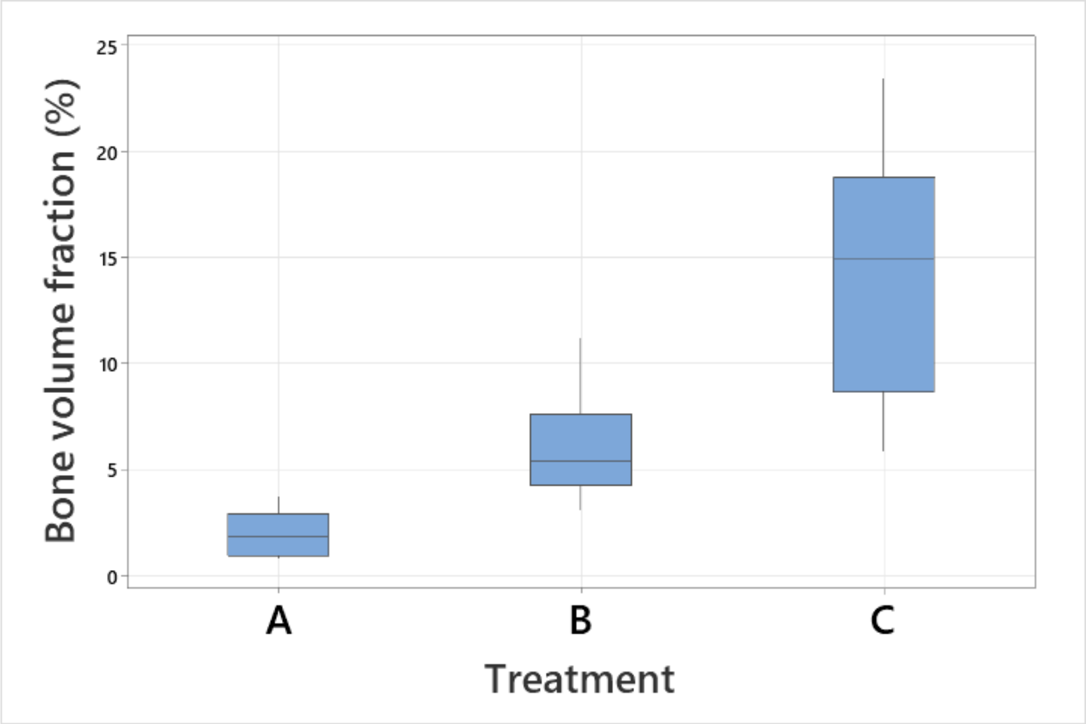
Note: If you reject H_0 in the K-W test, then Mann-Whitney tests can be performed as follow-up tests to determine which pairs of pop. medians are significantly different. Remember to control for the type I error rate!

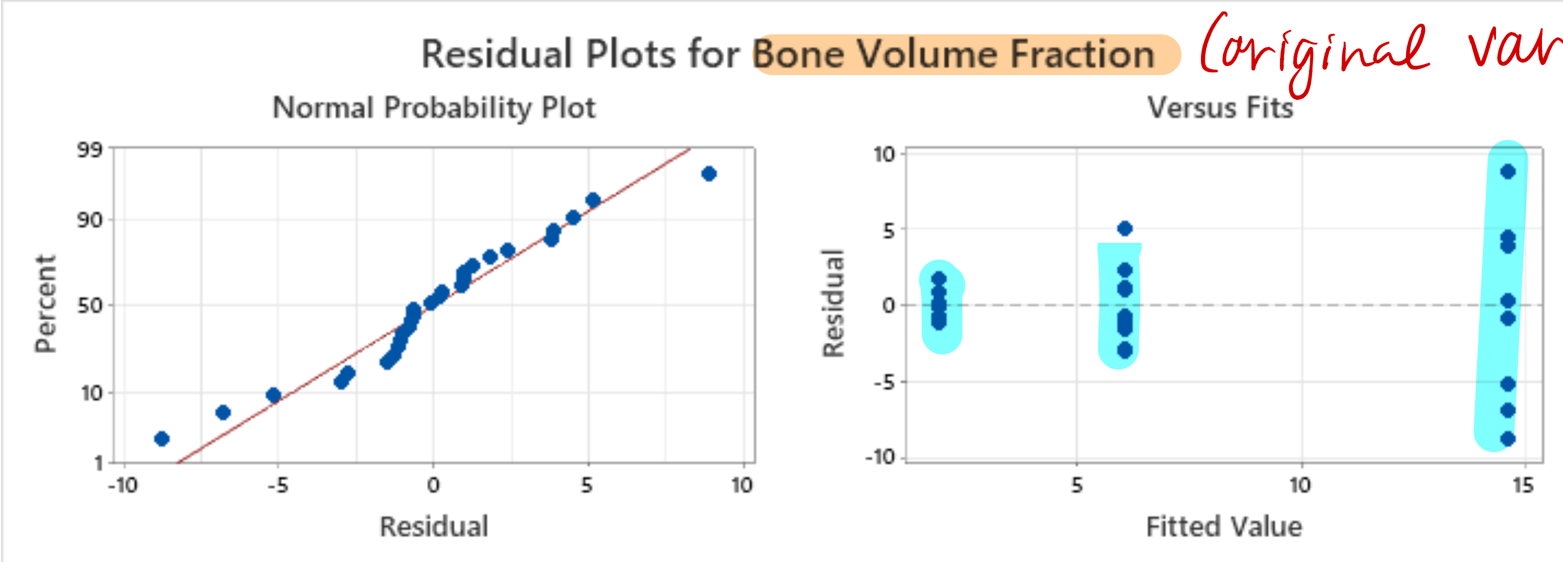
What if treatments do not have equal variance?

- Do not use a Kruskal-Wallis test.
- Try a transformation instead.

Example:

- Bone volume fraction for mice receiving drug A, B, or C.
- Equal-variance assumption fails.



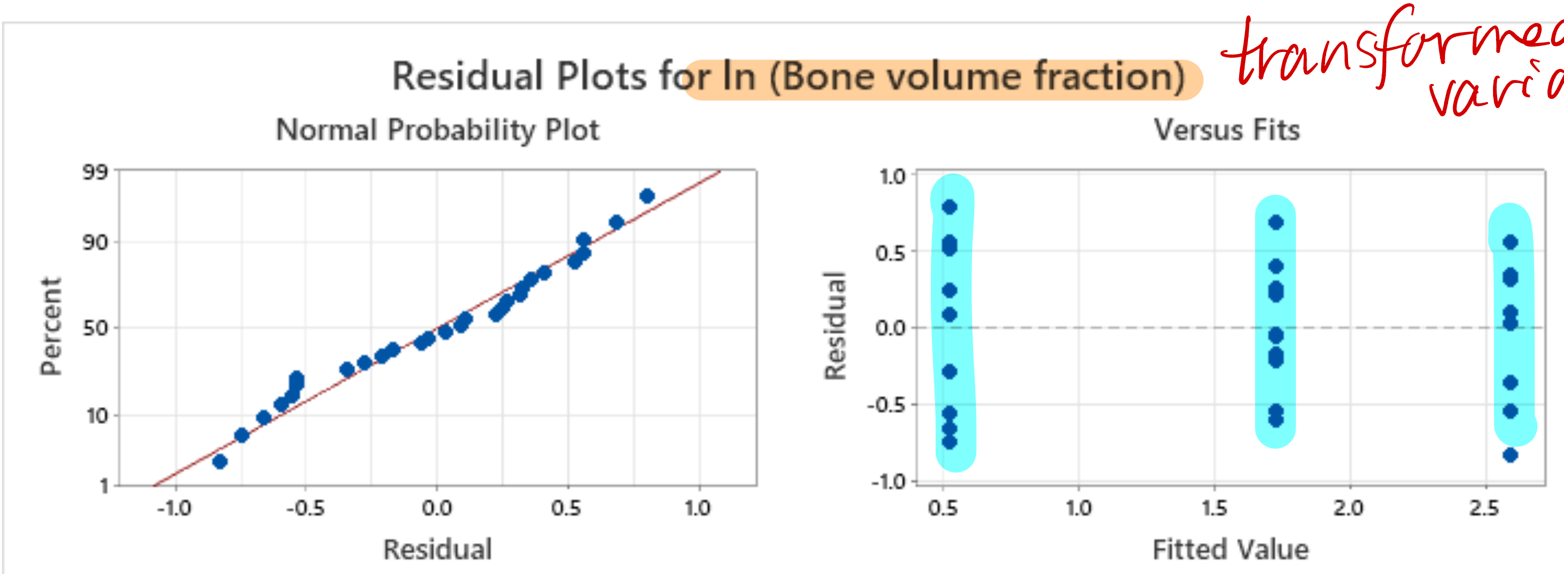


(original variable)

Transform the response using a natural log transform

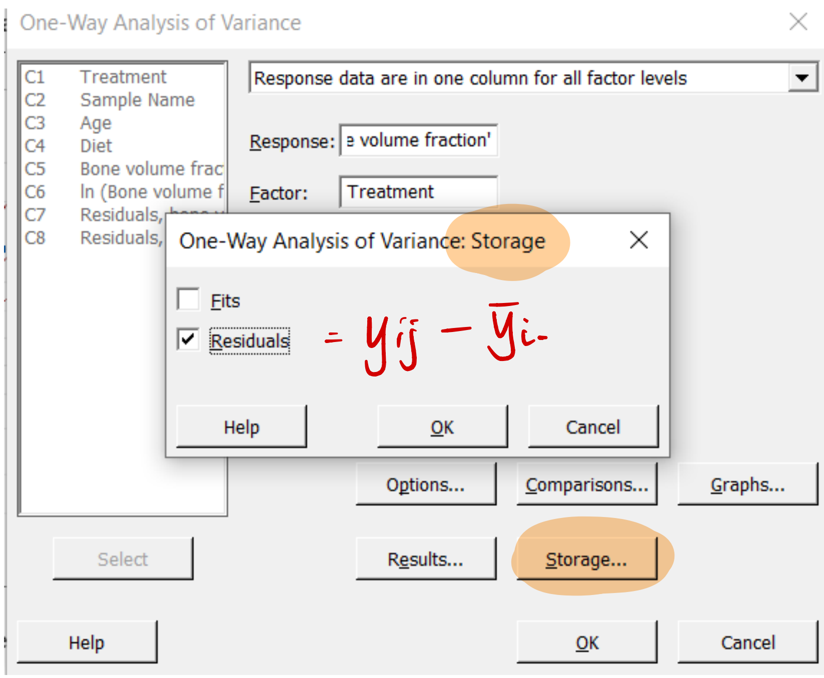
Transforming the response (bone volume fraction) improves both normality and equality of variance between groups.

transformed variable

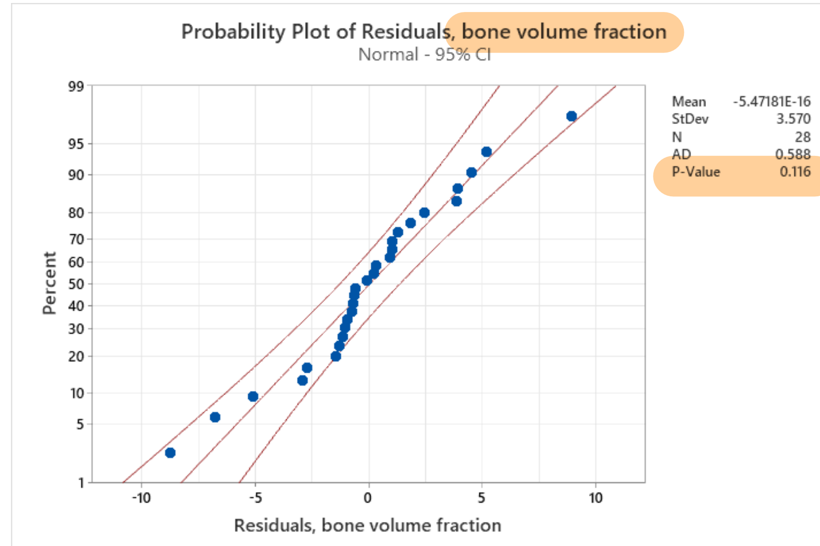


Not sure if residuals are normally distributed?

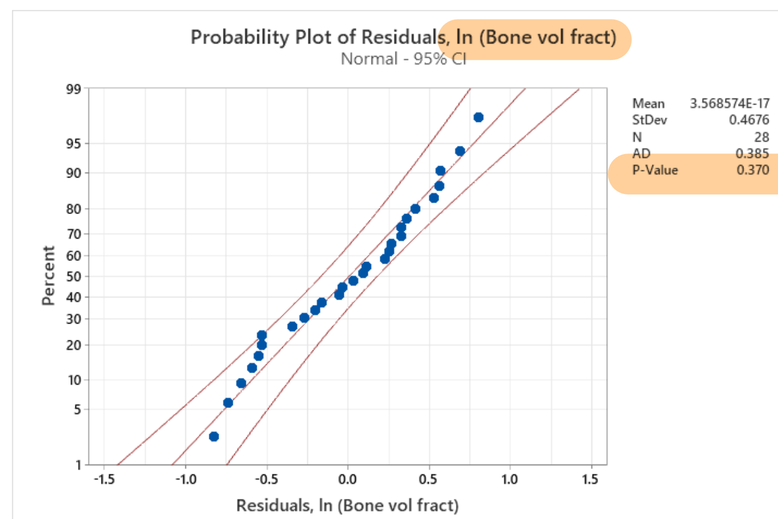
- In a one-way ANOVA, Minitab can store residuals in a column of your worksheet (see below). Create a normal probability plot by clicking Plot > Probability Plot and selecting the residual column.



Before transformation



After transformation



Guidelines to Select an Appropriate Transformation?

where Y = original response; Y' = transformed response

- $\boldsymbol{\sigma_i^2}$ **proportional to** $\boldsymbol{\mu_i}$
  - Use a <u>square root</u> transformation: $\boxed{Y' = \sqrt{Y}}$
- $\boldsymbol{\sigma_i}$ **proportional to** $\boldsymbol{\mu_i}$
  - Use a <u>log or natural log</u> transformation: $\boxed{\begin{array}{l} Y' = \log(Y) \\ Y' = \ln(Y) \end{array}}$
- $\boldsymbol{\sigma_i}$ **proportional to** $\boldsymbol{\mu_i^2}$
  - Use a <u>reciprocal</u> transformation: $\boxed{Y' = \frac{1}{Y}}$
- Examine $\frac{s_i^2}{\bar{y}_i}$, $\frac{s_i}{\bar{y}_i}$, and $\frac{s_i}{\bar{y}_i^2}$ and see if any are nearly constant across the groups (i.e numerator is proportional to denominator).

> iClicker Question:

When are non-parametric tests used?

A. Normal pop. dist(s) & small sample(s)
B. Normal pop. dist(s) & large sample(s)
C. Non-normal pop. dist(s) & small sample(s)
D. Non-normal pop. dist(s) & large sample(s)

Summary

- Parametric tests are used when normality assumptions are satisfied and involve the mean(s). The hypothesis tests up to this point in this class are all parametric.
- Nonparametric tests do not require normality assumptions and test the median(s) and are thus useful for non-normal, small samples.
- The "nonparametric" equivalent of a 2-sample t-test is a Mann-Whitney test.
- The "nonparametric" equivalent of a paired t-test is a One-sample Wilcoxon test.
- The "nonparametric" equivalent of one-way ANOVA (with equal variances) is a Kruskal-Wallis test. If variances are not equal, don't use the K-W test.
- These tests involve the population median, generally accomplished by assigning ranks to observations.
- Use nonparametric tests for non-normal small samples, or for occasions when the median is more reflective of your data than the mean (Ex: skewed data).
- Nonparametric tests generally have lower power than parametric tests, so, if normality assumptions are satisfied, use the parametric test.

EGEN350

HW10 Discussion

The questions

DATA

Use Normal Prob Plot

Normal → Use Parametric Test

Non-normal → Use Nonparametric Test

a) Can you use a parametric test on these data? Why or why not? Attach plot(s) from Minitab that support your answer.

b) What is the appropriate test?
   - One-sample T test
   - One-sample Wilcoxon test
   - Independent samples T test (Two-sample t)
   - Mann-Whitney test
   - Dependent samples (matched-pairs) T test
   - One-sample Wilcoxon test on the pair differences
   - One-way ANOVA
   - Kruskal-Wallis

c) State the appropriate null and alternative hypotheses.

d) Insert an image of the test results from Minitab. Make sure this image includes the type of test, hypotheses, test statistic, and p-value.

1st sentence

e) State your conclusion in the context of the problem.

Reject Ho or ≠Ha

2nd sentence

f) If appropriate, (1) insert follow-up analysis results and (2) state the conclusion for the follow-up analysis in the context of the problem. Otherwise, answer ‘n/a’.

How to open the HW10 data file?

Minitab WebApp:

- Go to File > Open > Worksheet > Local File and browse to the HW10 data file.

Desktop Minitab:

- Go to File > Open and browse to the HW10 data file.

How to create normal prob. plots?

Minitab WebApp:

- Go to Graph > Probability Plot > Simple (under One Y Variable) and create separate normal probability plots for each column.

Minitab Desktop:

- Go to Graph > Probability Plot > Multiple and enter all columns into Graph variables box.

Where to find each test in Minitab?

Parametric Tests

- Stat > Basic Statistics > 1-Sample t
- Stat > Basic Statistics > 2-Sample t (Independent Samples)
- Stat > Basic Statistics > Paired t (Dependent Samples)
- Stat > ANOVA > One-Way

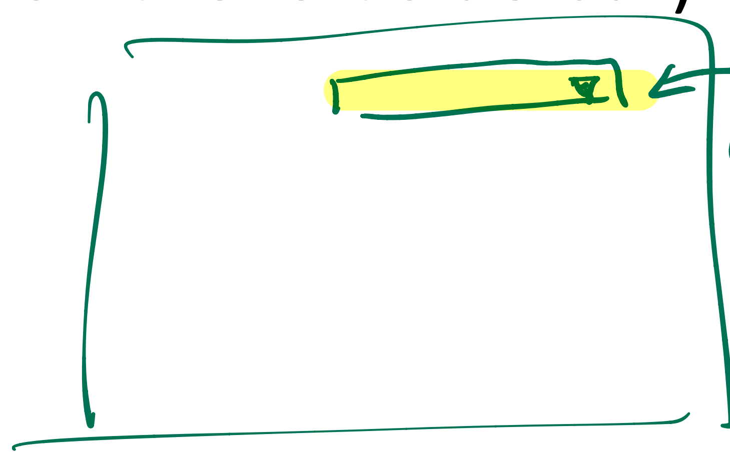
Nonparametric Tests

- Stat > Nonparametrics > 1-Sample Wilcoxon
- Stat > Nonparametrics > Mann-Whitney
- Stat > Nonparametrics > 1-Sample Wilcoxon on Differences
- Stat > Nonparametrics > Kruskal-Wallis

How to create a calculated column (e.g. Difference) in Minitab?

Minitab WebApp:

- Go to Calc > Calculator. Store result in variable for the new variable (e.g. Difference). Enter the expression as ‘C1’ – ‘C2’ (or double-click on the column names on the left side bar)



How data are arranged in worksheet

Desktop Minitab:

- Same as above.

How to type hypotheses in Gradescope?

- Use mu_1 for μ_1, etc.
- Use mu_D for μ_D, etc.
- Use eta_1 for η_1, etc.
- Use eta_D for η_D, etc.
- Use H0 for H_0
- Use Ha for H_a
- Use =/= for $\neq$

When to do a follow-up analysis?

- 1-Sample t or 1-Sample Wilcoxon:

  Reject Ho on ≠ Ha

- 2-sample or Mann-Whitney:

  Reject Ho on ≠ Ha

- Matched-pairs or 1-Sample Wilcoxon on Diffs:

  Reject Ho on ≠ Ha

- One-way ANOVA or Kruskal-Wallis:

  Reject Ho

  Use Tukey's as follow-up

  Use Mann-Whitney as follow-up

EGEN350

One predictor (explanatory variable)

Chapter 11

"Simple" Linear Regression

(Skip Sections 11-8, 11-9, 11-10)

Two or more predictor

"Multiple" Linear Regression (EC3)

Outline

- Scatterplot
- Regression Line
- Confidence Interval for the true slope β_1
- Hypothesis Testing for the true slope β_1
- Diagnostic Plots
- Correlation (R)
- Coefficient of Determination (R^2)
- Confidence Interval for the Mean Response (μ_Y)
- Prediction Interval for a Future Response (Y)

Scatterplot

Setting:

1. A random sample of size n is selected from the population.
2. Bivariate data (x_i, y_i) has been collected from the i^{th} individual in the sample.

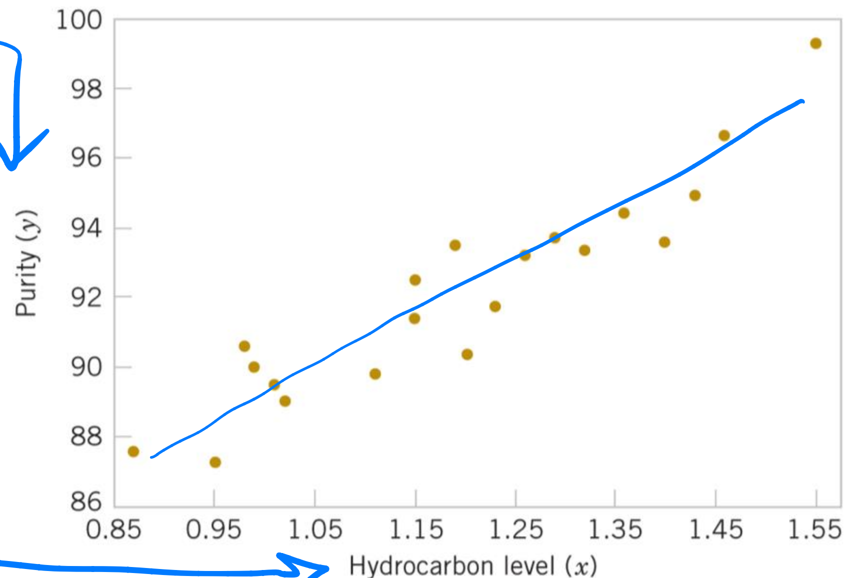
Scatterplot – displays the relationship between 2 numerical variables

Response variable – y

(AKA, Dependent variable)

Explanatory variable – x

(AKA, Independent variable or Predictor)

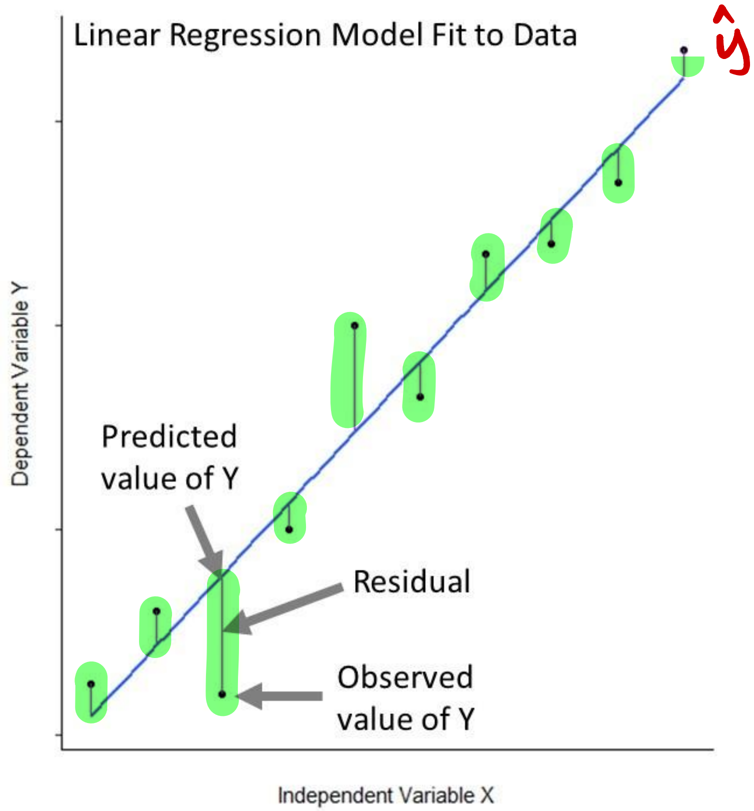


How do we find the 'best fit' line?

Least squares regression line:

$$\hat{y} = b_0 + b_1 x$$

The line that minimizes the sum of the squared residuals



points : (x_i, y_i)

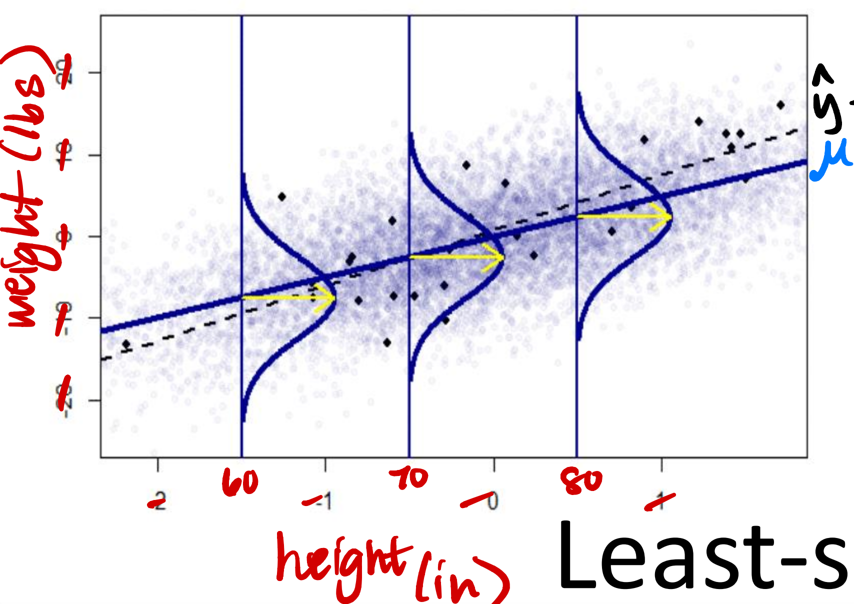
Residual:

$$e_i = y_i - \hat{y}$$

The vertical distance between a point and the least-squares regression line

blue circles - pop. (x,y)
black dots - sample (x,y)

Population Regression Line



$\hat{y}$

μ_y

$$\mu_y = \beta_0 + \beta_1 x$$

solid line

where μ_y is the true mean response

β_0 is the true y-intercept and

β_1 is the true slope

Least-squares Regression Line

$$\hat{y} = b_0 + b_1 x$$

black dashed line

At a specific x value,

$\hat{y}$ estimates μ_y

b_0 estimates β_0

b_1 estimates β_1

where $\hat{y}$ is the estimated mean response

b_0 is the estimated y-intercept and

b_1 is the estimated slope

Simple Linear Regression Model

$$y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{on pop. reg. line}} + \varepsilon_i$$

height of point

vertical distance b/w point & line

Estimated Simple Linear Regression Model

$$y_i = \underbrace{\overbrace{b_0 + b_1 x_i}^{\hat{y}}}_{\text{on the least-squares reg. line}} + e_i$$

Residual

$e_i = y_i - \hat{y}$

> iClicker Question:

When testing for a “linear relationship” between x and y, the parameter of interest is the true slope, β_1. What statistic is the best estimator of β_1?

A. b_0

B. b_1

C. μ_Y

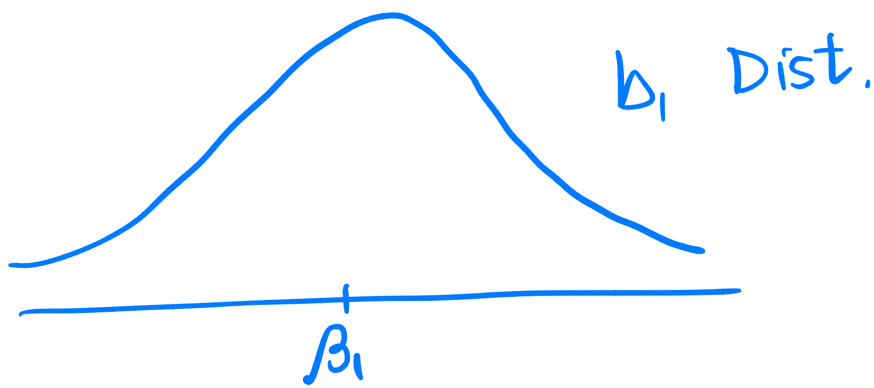
D. σ

Sampling Distribution of b_1

CENTER: $\mu_{b_1} = \beta_1$

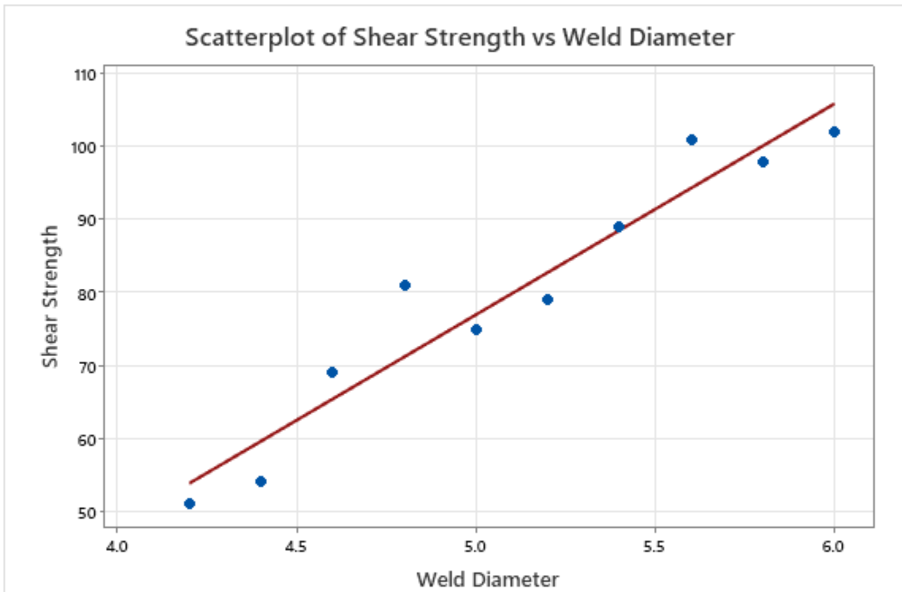
SPREAD: ${\sigma_{b_1}}^2 = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$

SHAPE: $b_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum (x_i - \bar{x})^2}\right)$



Ex: Weld Diameter vs. Shear Strength

Can you predict the shear strength of a weld, based on its diameter? Use data from a random sample of 10 welds to answer the question "Is a weld's diameter (in mm) linearly related to the shear strength (in kN/mm)?"



y-intercept

Coefficients b_0

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-67.7	15.2	-4.46	$\boxed{0.002}$	
Weld Diameter (x)	28.94	$\boxed{2.96}$	$\boxed{9.78}$	$\boxed{0.000}$	1.00

b_1 SE_{b_1} t_0 p-value

Tests: $H_0: \beta_0 = 0$, $H_a: \beta_0 \neq 0$ (not of interest)

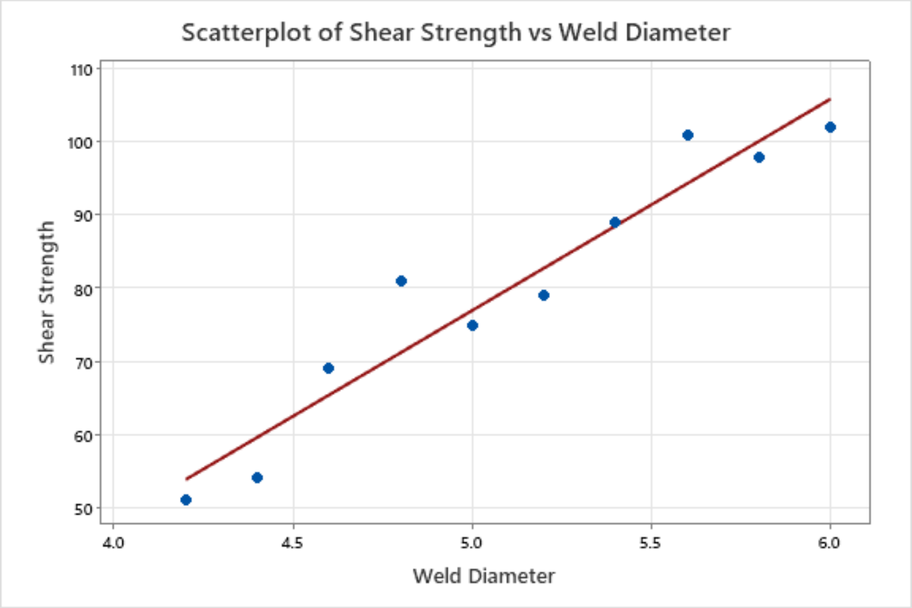
Tests: $H_0: \beta_1 = 0$, $H_a: \beta_1 \neq 0$ (of interest!)

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
$\boxed{5.37573}$	$\boxed{92.28\%}$	~~91.32%~~	~~88.24%~~

S_e R^2

> iClicker Question:



What is the response variable? What is the explanatory variable?

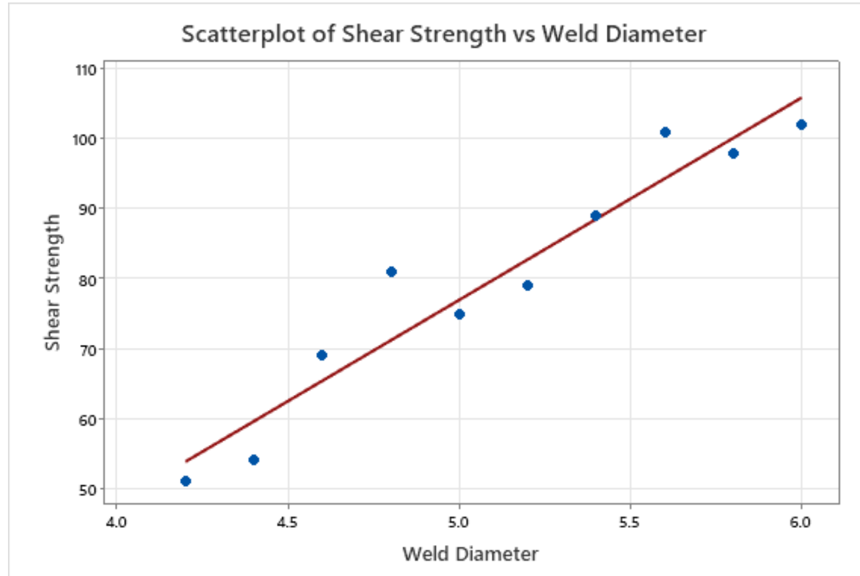
A. Response variable: Shear Strength
   Explanatory variable: Weld Diameter

B. Response variable: Weld Diameter
   Explanatory variable: Shear Strength

Ex: Interpretations of b_0 and b_1

L-S reg. line

$\hat{y} = \underline{-67.7} + \underline{28.94} \cdot x$



Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-67.7	15.2	-4.46	0.002	
Weld Diameter	28.94	2.96	9.78	0.000	1.00

In general:

- b_0 - the value of $\hat{y}$ when $x = 0$.

- b_1 - the change in $\hat{y}$ for a 1 unit increase in x.

> iClicker Question:

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-67.7	15.2	-4.46	0.002	
Weld Diameter	28.94	2.96	9.78	0.000	1.00

Interpret b_0 = -67.7:

A. For a weld with a diameter of -67.7mm, the estimated mean shear strength is 0 kN/mm.

B. For a weld with a diameter of 0mm, the estimated mean shear strength is -67.7 kN/mm.

$(x = 0)$ y

C. As the weld diameter increases by 1mm, the estimated mean shear strength decreases by 67.7 kN/mm.

D. As the weld diameter decreases by 67.7 mm, the estimated mean shear strength increases 1 kN/mm.

> iClicker Question:

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-67.7	15.2	-4.46	0.002	
Weld Diameter	28.94	2.96	9.78	0.000	1.00

Interpret b_1 = 28.94:

A. For a weld with a diameter of 0 mm, the estimated mean shear strength increases by 28.94 kN/mm.

B. For a weld with a diameter of 28.94 mm, the estimated mean shear strength increases by 1 kN/mm.

x '' by 1

C. As the weld diameter increases by 1 mm, the estimated mean shear strength increases by 28.94 kN/mm.

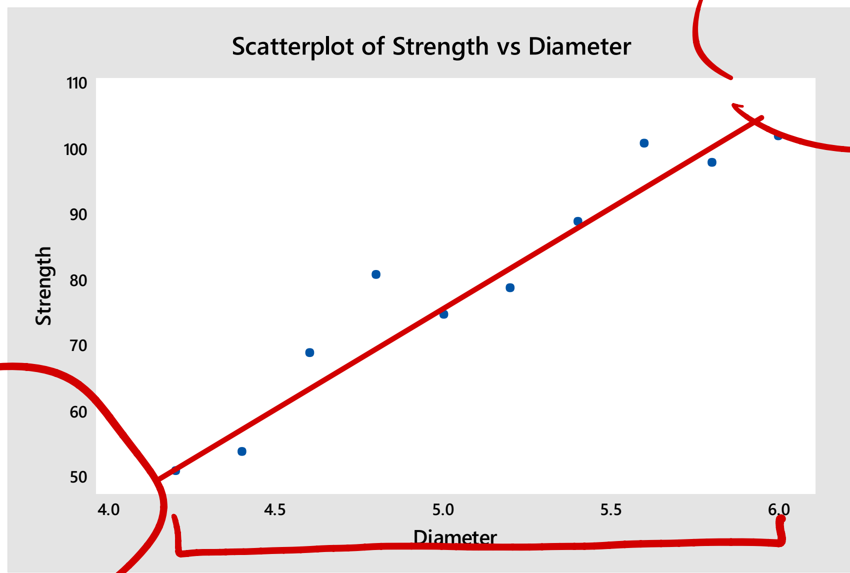
y

D. As the weld diameter increases by 28.94 mm, the estimated mean shear strength increases 1 kN/mm.

Do not extrapolate!

Extrapolation – using the least-squares regression line to estimate the mean response at an x value outside the range of x's used to construct the line.

Don't know the relationship out here



Confidence Interval for β_1

$$b_1 \pm T_{\underbrace{\alpha/2,n-2}_{\text{d.f.}}} SE_{b_1}$$

$(+\beta_1)$ $(+1)$

$\mu_y = \beta_0 + \beta_1 x$

<u>Example</u>:

Calculate and interpret the 95% CI for β_1.

$b_1 \pm t_{.025,8} \cdot SE_{b_1} = 28.94 \pm 2.306(2.96)$

$= (22.11, 35.77)$

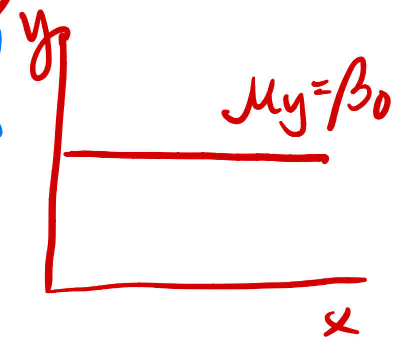
We are 95% confident for each additional mm in weld diameter, the true mean shear strength increases by between 22.11 and 35.77 KN/mm.

Hypothesis Testing for β_1

FOCUS!

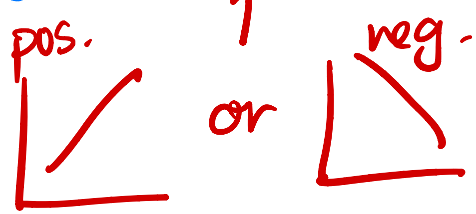
$H_0: \beta_1 = 0$	$H_0: \beta_1 = 0$	$\boxed{H_0: \beta_1 = 0}$
$H_a: \beta_1 < 0$	$H_a: \beta_1 > 0$	$\boxed{H_a: \beta_1 \neq 0}$

$H_0: \beta_1 = 0$ $\quad [\mu_y = \beta_0 + \cancel{\beta_1 x}]$ 0



there is no linear relationship b/w x & y.

$H_a: \beta_1 \neq 0$ $\quad [\mu_y = \beta_0 + \beta_1 x]$



there is a linear relationship b/w x & y.

Test Statistic: $\boxed{T_0 = \frac{b_1 - 0}{SE_{b_1}}}$ where $T_0 \sim t(n-2)$

Example: Hypothesis Test for β_1

Is there a significant linear relationship between weld's diameter and shear strength? Test hypotheses at α=0.05.

$H_0: \beta_1 = 0$ vs. $H_a: \beta_1 \neq 0$

$$t_0 = \frac{b_1 - 0}{SE_{b_1}} = \frac{+28.94 - 0}{2.96} = \boxed{9.78}$$

p-value: $2 \cdot \underbrace{P(T > |9.78|)}_{< .0005} < .001$

Reject H_0

There is adequate evidence to suggest there's a linear relationship b/w weld diameter and true mean shear strength. The data indicate that as the weld diameter increases, so does shear strength, on average.

Assumptions for CI and HT for β_1

1. A random sample of size n is selected from the population.

2. There is a linear relationship between x and y.

SAME SHAPE

SAME SPREAD

3. For a fixed value x, the response is normally distributed: $y_i \sim N(\mu_y, \boxed{\sigma^2})$
   Thus, the residuals $\varepsilon_i = y_i - \mu_y$ are normally distributed: $\varepsilon_i \sim N(0, \boxed{\sigma^2})$

   subtracts off center value

4. For every x, the variance of the response is σ^2. This is called *"Homogeneity of Variance."*

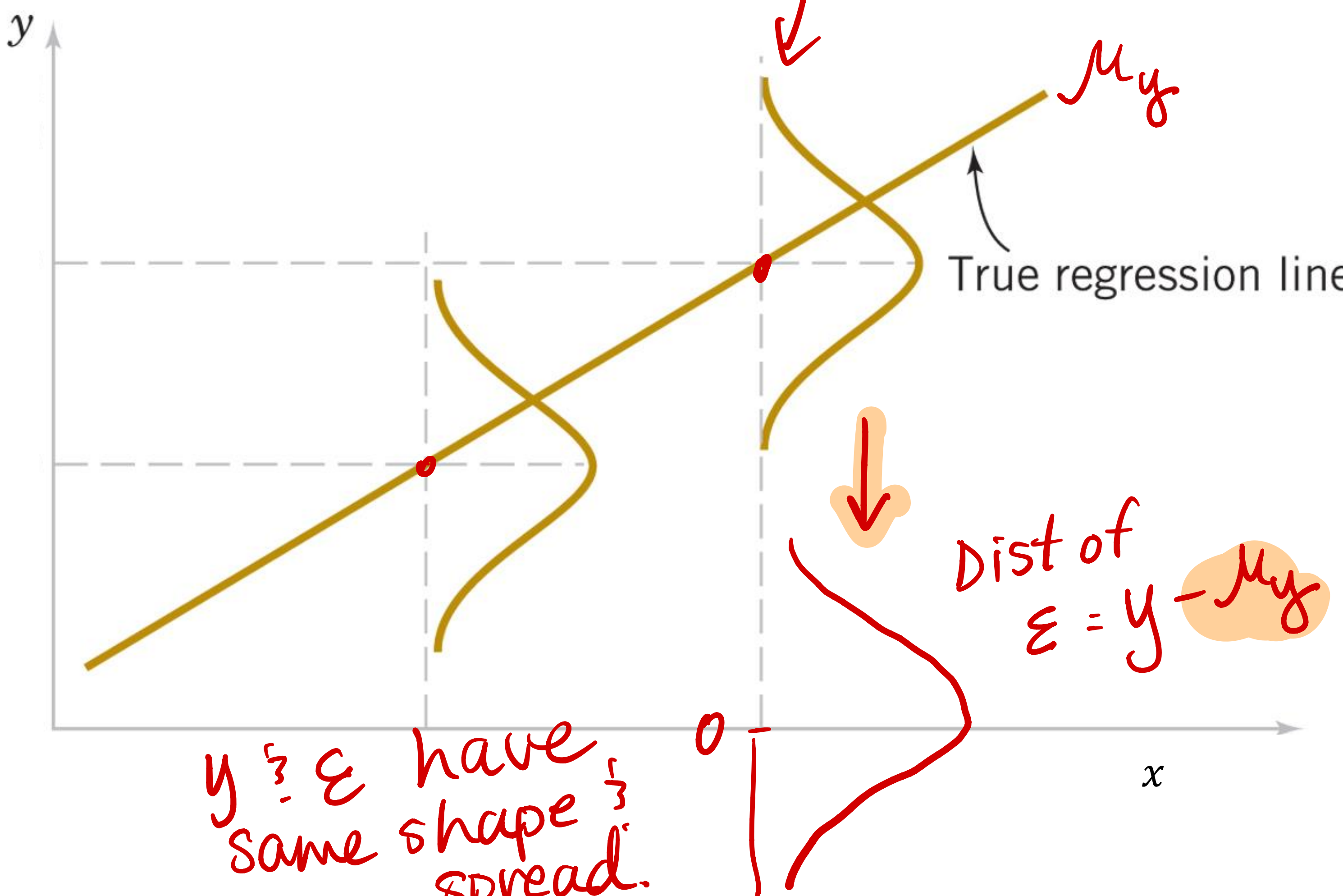
   AKA. Constant Variance

NOTE: S_e^2 estimates σ^2

Illustration of Assumptions #3 and #4

3-D plot

Dist. of y for a fixed x



y & ε have same shape & spread.

Assessing the Fit of the Line

Is the line a good fit to the data?

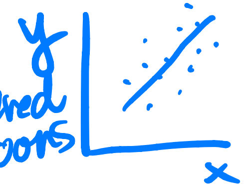
1. Diagnostic Plots

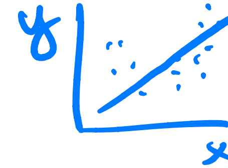
2. Correlation Coefficient (R)

3. Coefficient of Determination (R^2)

Diagnostic Plots – diagnose problems w/ assumptions

Tethered Balloons

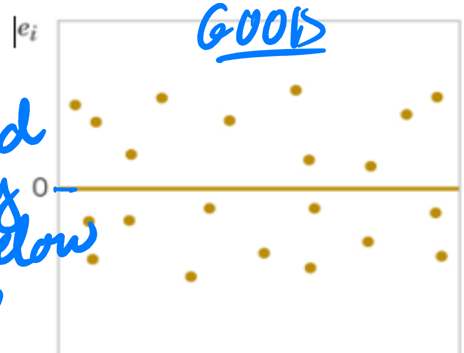




- Use resid. vs. fitted value plot to check:
  ① linear relationship between x and y

WANT points scattered randomly above/below resid =0 line

GOOD

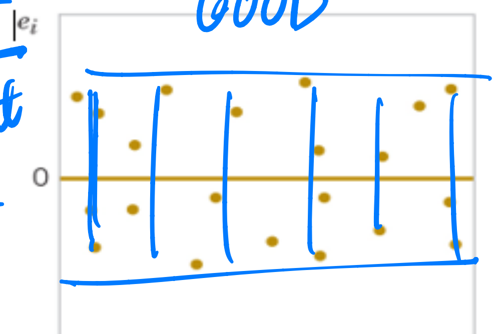


BAD
x & y are non-linear

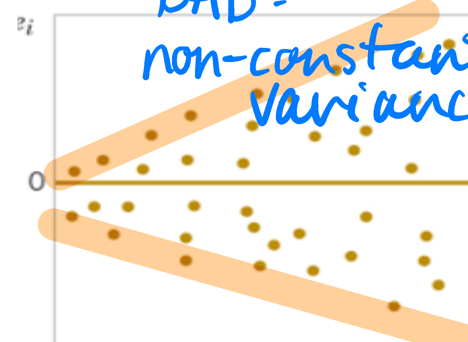
- Use resid. vs. fitted value plot to check:
  ② Constant variance

WANT constant vertical spread

GOOD



BAD: non-constant variance.

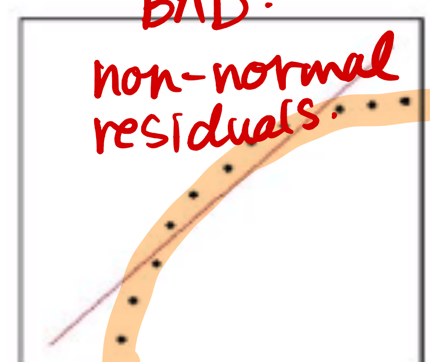


- Use Normal Prob. plot to check:
  ③ Normal residuals

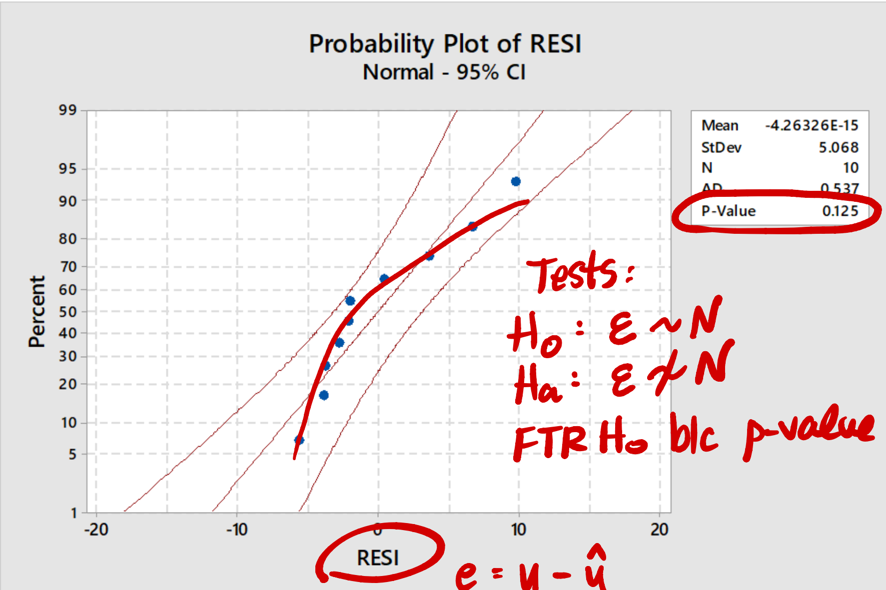
WANT points in straight line
→ no violation of normality.

GOOD:

BAD: non-normal residuals.



Example: Diagnostic Plots



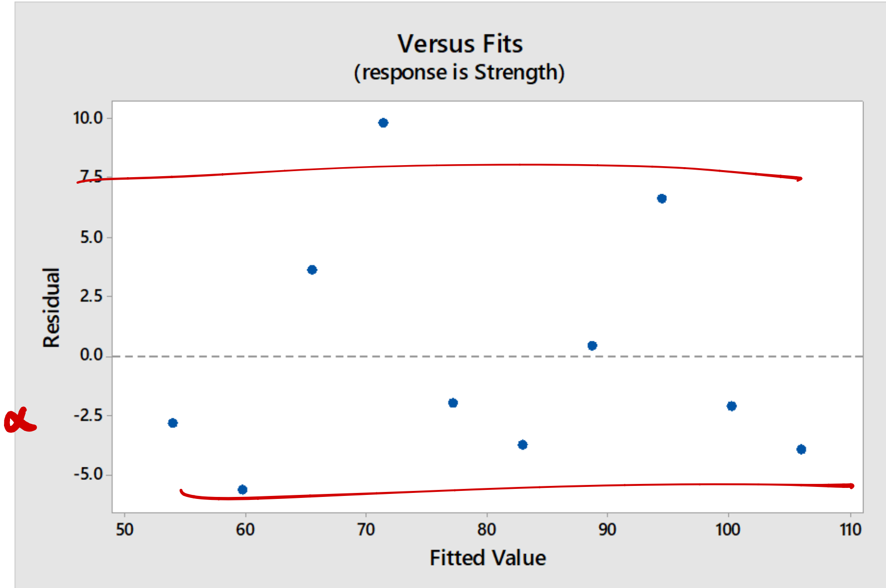
Tests:

$H_0: \varepsilon \sim N$

$H_a: \varepsilon \not\sim N$

FTR H_0 b/c p-value > α

$e = y - \hat{y}$



close enough to linear

→ No violation of the normal residuals assumption.

similar spread

→ No violation of constant variance assumption.

Correlation (R)

- Correlation (R) – measures the *direction* & *strength* of a linear relationship

R sign

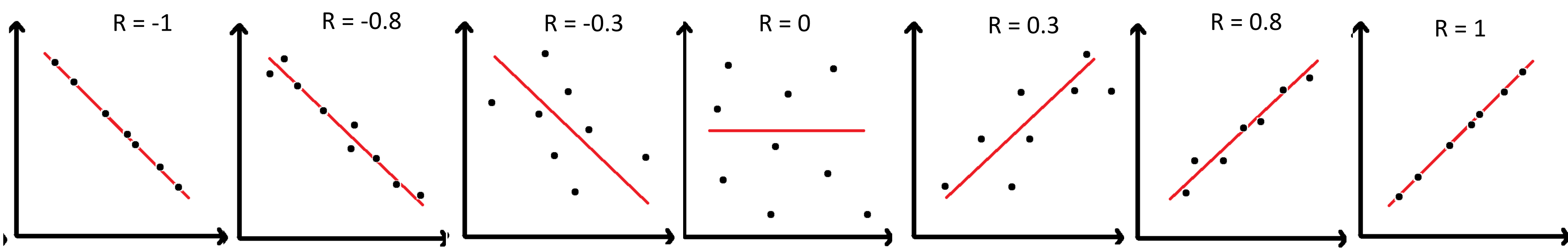
magnitude

"how close points are to the line"

- Possible values for R: $\boxed{-1 \leq R \leq 1}$

Note:

Correlation & slope always have the same sign.



Correlation Does Not Imply Causation!

- A positive correlation means that *large values* of one variable tend to occur with *large values* of the other variable.
- A negative correlation means that *large values* of one variable tend to occur with *small values* of the other variable.
- BUT… Correlation ***does not mean*** that changes in one variable **causes** changes in the other variable.

… due to the possibility of extraneous variables.

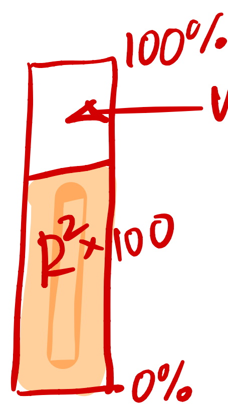
Coefficient of Determination (R^2)

or percentage if multiply by 100%

Coefficient of Determination: proportion of variability in y that is explained by the linear relationship between x and y

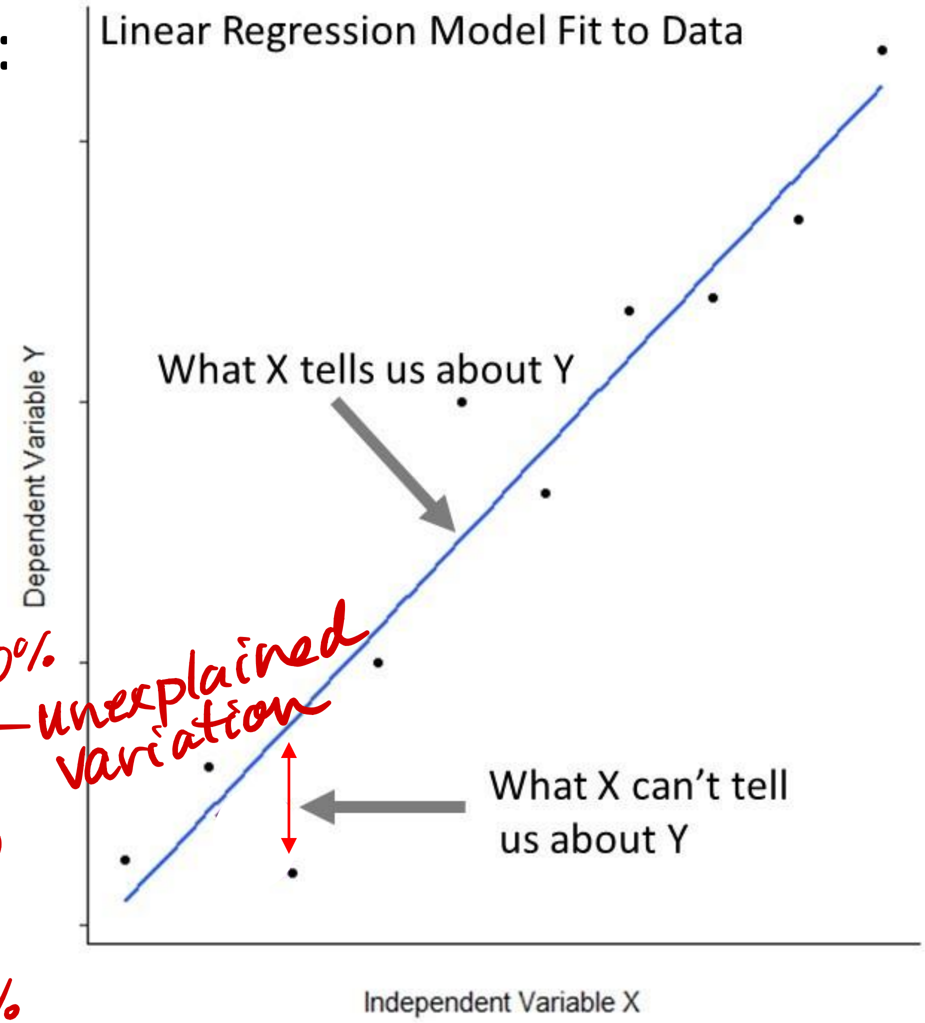
Possible values for R^2:

$\boxed{0 \leq R^2 \leq 1}$



explained by line

unexplained variation



> iClicker Question:

Interpret R^2=0.9228 (92.28%)

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
5.37573	$\boxed{92.28\%}$	~~91.32%~~	~~88.24%~~

A. 92.28% of the variation in weld diameter is explained by the linear relationship with shear strength.

(y)

B. 92.28% of the variation in shear strength is explained by the linear relationship with weld diameter.

(x)

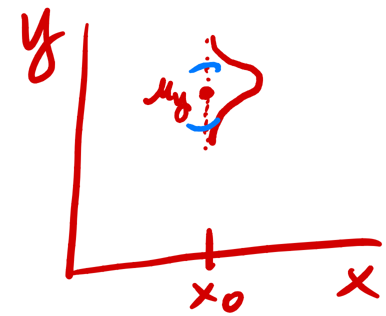
C. 92.28% of the variation in weld diameter is explained by the relationship with shear strength.

D. 92.28% of the variation in shear strength is explained by the relationship with weld diameter.

Confidence Interval for the Mean Response μ_Y @ x_0

centered at the same value

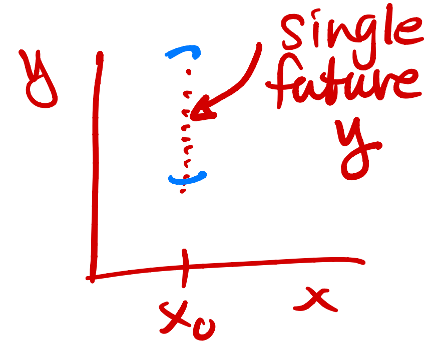
$$\hat{y}_0 \pm T_{\alpha/2, n-2} * \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right]}$$



Prediction Interval (PI) for a Future Response Y_0 @ x_0

$$\hat{y}_0 \pm T_{\alpha/2, n-2} * \sqrt{\hat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right]}$$

PI is always wider CI



where $\hat{y}_0 = b_0 + b_1 x_0$

x_0 is the specified value of x.

Note:

- CI for μ_Y and PI for Y are centered at the same estimate ($\hat{y}_0$).
- PI for Y is always wider the CI for μ_Y

Example: CI for μ_Y & PI for future Y_0 @ x_0

Prediction for Strength

Regression Equation

$\underset{\hat{y}}{\text{Strength}} = -67.7 + 28.94 \underset{x}{\text{Diameter}}$

Settings

Variable	Setting
Diameter	$\boxed{5.1} = x_0$

$\hat{y} = -67.7 + 28.94(5.1)$

Prediction

Fit	SE Fit	95% CI (for μ_y)	95% PI (for future y)
79.9	1.69996	(75.9799, 83.8201)	(66.8985, 92.9015)

μ_y

Interpret the 95% CI for μ_Y, when weld diameter=5.1:

We are 95% confident the true mean shear strength is between 75.98 and 83.82 kN/mm, for welds with diameter of 5.1 mm.

single future y

Interpret the 95% PI for a future response Y_0, when weld diameter=5.1:

We are 95% confident the shear strength will be between 66.90 and 92.90 kN/mm for a weld with a diameter of 5.1 mm.

Summary

- Scatterplot – graph of two quantitative variables
- Response variable – the variable of interest
- Explanatory variable – the variable used to explain changes in the response
- Least-squares regression – a method for obtaining the best fit line by minimizing the squared residuals. (A residual is the vertical distance between an (x, y) point and the line.)
- Population regression line: $\mu_y = \beta_0 + \beta_1 x$
- Least-squares regression line: $\hat{y} = b_0 + b_1 x$
  - μ_y is the true mean response at a given x value, and β_0, β_1 are the true intercept and slope.
  - $\hat{y}$ is the estimated mean response at a given x value, and b_0, b_1 are the estimated intercept and slope.
  - $\hat{y}$ estimates μ_y, b_0 estimates β_0, and b_1 estimates β_1
- Formula for the CI for β_1: $b_1 \pm T_{\alpha/2, n-2} SE_{b_1}$
- Formula for the test stat in a hypothesis test for β_1: $T_0 = \frac{b_1 - 0}{SE_{b_1}}$
- Assumptions and how to check the assumptions on diagnostic plots
- Correlation (R) – measures the *direction & strength* of a linear relationship
- Coefficient of Determination (R^2) – proportion of variability in y that is explained by the linear relationship between x and y
- CI for μ_Y vs. PI for future Y – similarities and differences

Final Exam Review

CHAPTER 10

- Independent Samples
  - Two-sample t test and CI for $\mu_1 - \mu_2$
- Dependent Samples
  - Matched-pairs t test and CI for μ_D

CHAPTER 13

- One-way ANOVA
  - Overall Test for 3 (or more) population means
  - Follow-up Tukey's Multiple Comparisons, if Overall test is significant. Reject H_0

CHAPTER 11

- Simple Linear Regression
  - Test and CI for β_1

+ Nonparametrics

CHAPTER 10:

*add more info from annotated notes structured review/12/4

X_1 and X_2 Distributions

Normal

$X_1 \sim N(\mu_1, \sigma_1^2)$

$X_2 \sim N(\mu_2, \sigma_2^2)$

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

Not Normal (or unknown)

- at least one $n < 30$ → Check Normal Prob Plot
  - If normal (straight line) →
  - If not normal (no straight line) → use nonparametrics.
- $n_1 \geq 30$, $n_2 \geq 30$ →

$$\bar{X}_1 - \bar{X}_2 \dot{\sim} N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

by CLT

TWO-SAMPLE T CI

Confidence Interval (CI) for $\mu_1 - \mu_2$:

$$\bar{X}_1 - \bar{X}_2 \pm T_{\alpha/2, df}\sqrt{\frac{{S_1}^2}{n_1}+\frac{{S_2}^2}{n_2}} = (-\_\_\_, +\_\_\_)$$

Generic CI Interpretation:

"We are 95% confident that the true mean response for pop. 1 is between lower limit and upper limit units $\underbrace{\text{less than}}_{(-)}$ / $\underbrace{\text{greater than}}_{(+)}$ that of pop. 2."

*** Remember to write the interpretation in the context of the problem.**

*** Remember to state which mean is larger / smaller than the other and by how much.**

TWO-SAMPLE T TEST

<u>Hypothesis Testing for $\mu_1 - \mu_2$</u>:

usually zero

$$H_0: \mu_1 - \mu_2 = \Delta_0$$
$$H_a: \mu_1 - \mu_2 \ \boxed{< \ / \ > \ / \ \neq} \ \Delta_0$$

choose 1

* Use $\mu_1 - \mu_2$ in both hypotheses

* Use the same number in both hypotheses

<u>Assumptions</u>:

1. The data are two independent random sample(s) drawn from two independent populations.
2. The $\bar{X}_1$ and $\bar{X}_2$ dist's are at least approximately normal.

$\bar{X}_1 - \bar{X}_2 \sim N$

or $\bar{x}_1 - \bar{x}_2 \dot{\sim} N$ (by CLT)

<u>Test Statistic</u>: $T_0 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{S_1{}^2}{n_1} + \frac{S_2{}^2}{n_2}}}$ where $T_0 \sim t\Bigg(\underbrace{\frac{\left(\frac{S_1{}^2}{n_1}+\frac{S_2{}^2}{n_2}\right)^2}{\frac{\left(S_1{}^2/n_1\right)^2}{n_1-1}+\frac{\left(S_2{}^2/n_2\right)^2}{n_2-1}}}_{\text{Big DF Formula}}\Bigg)$

<u>p-value</u>:

Test stat. value

Alternative Hypothesis	p-value
$H_a\colon \mu < \mu_0$	$P(T_0 < t_0)$
$H_a\colon \mu > \mu_0$	$P(T_0 > t_0)$
$H_a\colon \mu \neq \mu_0$	$2 \cdot P(T_0 > \|t_0\|)$

Decision:

Test stat. Critical value

1. Critical-value approach

	Critical value	Reject H_0 if:
H_a: $\mu < \mu_0$	$-T_\alpha$	$T_0 < -T_\alpha$
H_a: $\mu > \mu_0$	T_α	$T_0 > T_\alpha$
H_a: $\mu \neq \mu_0$	$\pm T_{\alpha/2}$	$\lvert T_0 \rvert > T_{\alpha/2}$

2. p-value approach

p-value	Decision
p-value < α	Reject H_0
p-value $\geq \alpha$	Fail to Reject H_0

- Generic Conclusion: "There is (not) enough evidence to conclude that the true mean response for pop. 1 is less/greater/different than that of pop. 2."

H_a direction

*** If you reject H_0 on a two-sided test, include a second sentence stating which mean was found to be larger / smaller than the other.** Look at $\bar{x}_1$ and $\bar{x}_2$

$\neq$

*** Failing to Reject H_0 DOES NOT imply H_0 is true!**

MATCHED-PAIRS T CI

diff :

Pre Post

$\boxed{\text{Pre - post}}$

Confidence Interval (CI) for μ_D:

$$\bar{X}_D \pm T_{\alpha/2, n-1} \frac{S_D}{\sqrt{n}}$$

diffs

Generic CI Interpretation:

"We are 95% confident that the true mean response of pop. 1 is between lower limit and upper limit units less than / greater than that of pop. 2."

*** Remember to write the interpretation in the context of the problem.**

*** Remember to state which mean is larger / smaller than the other and by how much.**

Look at $\bar{x}_D$

and order of subtraction

MATCHED-PAIRS T TEST

Hypothesis Testing for μ_D

usually zero

$$H_0: \mu_D = \Delta_0$$

$$H_a: \mu_D \begin{array}{|c|}\hline < \\ > \\ \neq \\ \hline\end{array} \Delta_0$$

choose 1

* Use μ_D in both hypotheses

* Use the same number in both hypotheses

Assumptions:

1. The sample(s) must be random sample(s).
2. The $\bar{X}_D$ distribution must be at least approximately normal.

Test Statistic: $T_0 = \frac{\bar{X}_D - \Delta_0}{s_D/\sqrt{n}}$ where $T_0 \sim t(n-1)$

d.f.

p-value:

Make a Decision

Generic Conclusion:

Alternative Hypothesis	p-value
$H_a\colon \mu < \mu_0$	$P(T_0 < t_0)$
$H_a\colon \mu > \mu_0$	$P(T_0 > t_0)$
$H_a\colon \mu \neq \mu_0$	$2 \cdot P(T_0 > \lvert t_0 \rvert)$

"There is (not) enough evidence to conclude that the true mean response for pop. 1 is less / greater / different than that of pop. 2."

*** If you reject H_0 on a two-sided test, include a second sentence stating which mean was found to be larger / smaller than the other.**

*** Failing to Reject H_0 DOES NOT imply H_0 is true!**

CHAPTER 13:

"Overall" Hypothesis Test for One-way ANOVA

$$H_0:\ \mu_1 = \mu_2 = \cdots = \mu_a$$
$$H_a:\ not\ all\ the\ means\ are\ equal$$

Assumptions:

1. Independent random samples from each population.
2. Each population distribution is normal.
3. Population variance is the same for all populations.

Test Statistic: $F_0 = \dfrac{MSTr}{MSE}$ where $F_0 \sim F(DFTr, DFE)$

p-value: $P(F > F_0)$

Make a Decision

(option)

Generic Conclusion: "There is (not) enough evidence to conclude the true mean response differs for at least one the populations."

- If you reject H_0 on the overall test, then for the follow-up analysis, perform Tukey's pairwise comparisons to determine which population means significantly differ and which do not.
  - If 0 is contained in the CI for $\boxed{\mu_i - \mu_j,}$ then the two pop. means are not significantly different.
  - If 0 is NOT contained in the CI for $\boxed{\mu_i - \mu_j,}$ then the two pop. means are significantly different.

Tests for Normality: $H_0: \varepsilon \sim N$ small p-value $\rightarrow$ violation of normality
$H_a: \varepsilon \nsim N$ Large p-value $\rightarrow$ no violation of normality

Diagnostic plots

$\varepsilon_{ij} = y_{ij} - \mu_i$

Used to "diagnose" problems with ANOVA assumptions

- Normal probability plot (Points following a straight line indicate no problem with assumption of normal residuals.)

  $e_{ij} = y_{ij} - \bar{y}_i$

- Residual vs. Fitted Response (Fits) (Similar vertical spread across the groups indicates no problem with the homogeneity of variance (constant variance) assumption. If largest s / smallest s < 2, the constant variance assumption is satisfied.)

Section 9.9: Non-parametric Tests

Z, T, F

- Parametric tests require normality. Non-parametric tests do not.

- Non-parametric tests are most useful for small samples from non-normal populations.

- Non-parametric tests generally have lower power than parametric tests, so, if normality assumptions are satisfied, use the parametric test.

Parametric	Non-parametric
2-sample t-test	Mann-Whitney test
Paired t-test	1-sample Wilcoxon test
One-way ANOVA	Kruskal-Wallis test

add something here ↓ from annotation notes

Non-parametric Hypotheses use η = population median

Mann-Whitney test

$$H_0\colon \eta_1 = \eta_2$$

$$H_a\colon \eta_1 \boxed{\begin{matrix} < \\ > \\ \neq \end{matrix}} \eta_2$$

choose 1

Wilcoxon Signed-Rank test

$$H_0\colon \eta_D = 0$$

$$H_a\colon \eta_D \boxed{\begin{matrix} < \\ > \\ \neq \end{matrix}} 0$$

choose 1

Kruskal-Wallis Test - Like one-way anova

$$H_0\colon \eta_1 = \eta_2 = \eta_3 = \cdots$$

H_a: not all pop. medians are equal

CHAPTER 11:

add stuff from annotated notes
12/4

Confidence Interval for β_1:

L-S reg. line

$\hat{y} = b_0 + b_1 x$

$$b_1 \pm T\alpha_{/2, n-2} SE_{b_1} \approx (- \quad , + \quad)$$

Generic CI Interpretation:

"We are 95% confident that for each 1 unit increase in x, the true mean response increases / decreases by between lower limit and upper limit units."

Pop. Reg line

$\mu_y = B_0 + B_1 x$

$(+B_1)$ $(+1)$

Hypothesis Testing for β_1

(Check on handwritten notes for this page)

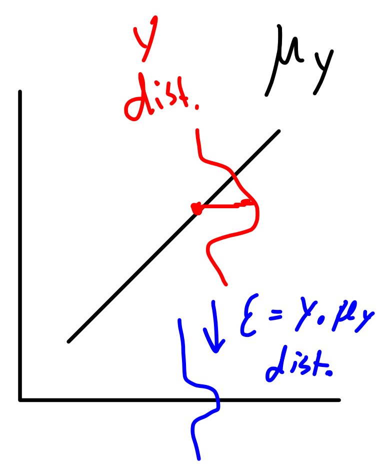
$H_0: \beta_1 = 0$

$H_a: \beta_1 \neq 0$

* Use β_1 in both hypotheses
* Use the same number in both hypotheses

Assumptions:

1. A random sample of size n is selected from the population.
2. For a fixed value x, the response is normally distributed: $y_i \sim N\left(\mu_y, \sigma^2\right)$
   Thus, the residuals $\varepsilon_i = y_i - \mu_y$ are normally distributed: $\varepsilon_i \sim N(0, \sigma^2)$.



3. For every x, the variance of the response is σ^2. This is called the Homogeneity of Variance.

Test Statistic: $T_0 = \frac{b_1 - 0}{SE_{b_1}}$ where $T_0 \sim t(n-2)$

p-value

Alternative Hypothesis	p-value
$H_a\!: \mu < \mu_0$	$P(T_0 < t_0)$
$H_a\!: \mu > \mu_0$	$P(T_0 > t_0)$
$H_a\!: \mu \neq \mu_0$	$2 \cdot P(T_0 > \lvert t_0 \rvert)$

Make a Decision

Generic Conclusion:
"There is (not) enough evidence to conclude there is a linear relationship between y and x."

- If you reject H_0 then write a second sentence stating what direction of the relationship (positive / negative) was found.

Look at sign on b_1

- Diagnostic plots – diagnose problems with regression assumptions
  - Normal probability plot (Points following a straight line indicate no violation of the normality assumption.)
  - Residual vs. Predicted (Fitted) Response (A random scatter above and below the residual=0 line indicates no violation of the linearity assumption. Curved pattern would indicate a problem. Similar vertical spread across the plot indicates no problem with the homogeneity of variance assumption. A fan-shaped pattern would indicate a problem.

- Correlation (R) – measures the *direction* & *strength* of a *linear* relationship

  +/- Magnitude

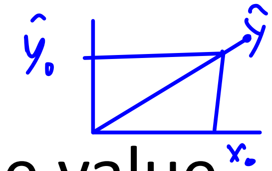
  - Possible values for R: $-1 \leq R \leq 1$

  Can be used for linear or nonlinear

- Coefficient of Determination (R^2) – proportion of variance in y that is explained by the linear relationship between x and y

  $R^2 = \frac{SSR}{SST} = \frac{SSM}{SST}$

CI for μ_Y vs. PI for future Y



- Similarity: CI and PI are centered at the same value ($\hat{y}$ at x=#)
- Difference: PI will always be wider than CI, because there is more variability in individual responses than in mean responses.

Generic Interpretation of a CI for μ_Y: at $x=\#$

"We are 95% confident that the true mean response is between lower limit and upper limit units, when x = # units."

Generic Interpretation of a PI for a future Y: at $x=\#$

"We are 95% confident that an individual's response will be between lower limit and upper limit units, when x = # units."

GENERAL FORM OF TEST STATISTICS:		GENERAL FORM OF TWO-SIDED CONFIDENCE INTERVALS:	
$Z_0 = \frac{estimate - null\ \ value}{SD_{estimate}}$	$T_0 = \frac{estimate - null\ value}{SE_{estimate}}$	$estimate\ \pm Z_{\alpha/2} \cdot SD_{estimate}$	$estimate\ \pm T_{\alpha/2} \cdot SE_{estimate}$

Type of Analysis	Parameter	(Statistic) Estimate	(σ) $SD_{Estimate}$ or $SE_{Estimate}$	Hypotheses	Test Statistic	Two-sided CI
One-sample Z Test and CI	μ	$\bar{x}$	$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$	$H_0: \mu = \mu_0$ $H_a: \mu \boxed{< / > / \neq} \mu_0$ choose 1	$Z_0 = \frac{\bar{x} - \mu_0}{\sigma_x / \sqrt{n}}$	$\bar{x} \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$
One-sample T Test and CI	μ	$\bar{x}$	$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$	$H_0: \mu = \mu_0$ $H_a: \mu \boxed{< / > / \neq} \mu_0$ choose 1	$T_0 = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$	$\bar{x} \pm T_{\alpha/2, n-1} \cdot \frac{s}{\sqrt{n}}$
Two-sample Z Test and CI	$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	usually 0 $H_0: \mu_1 - \mu_2 = \Delta_0$ $H_a: \mu_1 - \mu_2 \boxed{< / > / \neq} \Delta_0$ choose 1	$Z_0 = \frac{\bar{x}_1 - \bar{x}_2 - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	$\bar{x}_1 - \bar{x}_2 \pm Z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
Two-sample T Test and CI	$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	usually 0 $H_0: \mu_1 - \mu_2 = \Delta_0$ $H_a: \mu_1 - \mu_2 \boxed{< / > / \neq} \Delta_0$ choose 1	$T_0 = \frac{\bar{x}_1 - \bar{x}_2 - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	$\bar{x}_1 - \bar{x}_2 \pm T_{\alpha/2, \nu} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
Paired Z Test and CI	μ_D	$\bar{x}_D$	$\sigma_{\bar{x}_D} = \frac{\sigma_D}{\sqrt{n}}$	usually 0 $H_0: \mu_D = \Delta_0$ $H_a: \mu_D \boxed{< / > / \neq} \Delta_0$ choose 1	$Z_0 = \frac{\bar{x}_D - \Delta_0}{\sigma_D / \sqrt{n}}$	$\bar{x}_D \pm Z_{\alpha/2} \cdot \frac{\sigma_D}{\sqrt{n}}$
Paired T Test and CI	μ_D	$\bar{x}_D$	$SE_{\bar{x}_D} = \frac{s_D}{\sqrt{n}}$	usually 0 $H_0: \mu_D = \Delta_0$ $H_a: \mu_D \boxed{< / > / \neq} \Delta_0$ choose 1	$T_0 = \frac{\bar{x}_D - \Delta_0}{s_D / \sqrt{n}}$	$\bar{x}_D \pm T_{\alpha/2, n-1} \cdot \frac{s_D}{\sqrt{n}}$
Simple Linear Regression	β_1	b_1	SE_{b_1}	$H_0: \beta_1 = 0$ $H_a: \beta_1 \neq 0$	$T_0 = \frac{b_1 - 0}{SE_{b_1}}$	$b_1 \pm T_{\alpha/2, n-2} \cdot SE_{b_1}$