

# Overview

## Disclaimers + notes

- **Made by Yuxi Qin ([yuxiqin.ca](http://yuxiqin.ca))**
- These notes are derived from Sharon King-Yu's notes (STATS 3Y03 Fall 2022), ty Dr. King-Yu :D
- I'm not perfect, there will be mistakes in these notes. Call it a learning experience lol
- There are Childsmath questions/solutions sprinkled into corresponding lecture topics ;)
- **DO NOT SUBSTITUTE THESE FOR YOUR LECTURES!! THIS IS SIMPLY A TOOL!! I AM NOT LIABLE FOR YOUR GRADES**
- Please ignore any random comments, they were part of the learning experience
- Past midterms/exams + sample tests are VERY GOOD review content
- Feel free to connect with me on [Linkedin](#) and send me any co-op opportunities for embedded systems :D
- Okay that's it enjoy!

(Fall 2022)

Midterm 1: Up to and including Lec 10

Midterm 2: Up to and including Lec 22, focus on post-midterm 1

Exam: Up to and including Lec 24, focus on post-midterm 2

1 - 2.1 - Sample Spaces and Events

2 - 2.1 (Cont)

3 - 2.2, 2.3 - Properties of Probability

4 - 2.4, 2.5, 2.6 - Conditional Probability and the Law of Total Probability

5 - 2.7, 2.9, 3.1 - Independence, Random Variables, and Discrete Random Variable

6 - 3.1, 3.2, 3.3 - Probability Distributions, Mean and Variance of a Discrete Random Variable

7 - 3.5, 3.6 - Binomial, Geometric, Negative Binomial Distributions

8 - 3.7, 3.8 - Hypergeometric Distribution, Poisson Distribution

9 - 4.1 - Continuous Random Variables, and Probability Density Functions

10 - 4.2, 4.3 - Cumulative Distribution Functions, Mean and Variance of a Continuous Random Variable

11 - 4.5 - Normal Distribution

12 - 4.6 - Normal Approximation to the Binomial, OMIT normal approx to the posson)

13 - 4.7, 5.1 - exponential distribution, two or more random variables, omit discrete rvs, omit conditional prob distributions, omit more than two random vars)

14 - 5.1 (Cont)

15 - 5.4, 5.6 - covariance and correlation, linear functions of random variables

16 - 6.1, 6.2 - numerical summaries, stem and leaf plots

17 - 6.3, 6.4 - histograms and boxplots

18 - 6.7 - probability plots

19 - 7.1, 7.2 - point estimation, central limit theorem

20 - 7.3, 8.1 - concepts of point estimation, confidence interval for the mean, variance known

21 - 8.1 (Cont)

22 - 8.2, 8.4 - Confidence Interval for the mean, variance unknown, confidence interval for a population proportion

23 - 9.1 - hypothesis testing

24 - 9.1 (Cont)

25 - 9.2 - tests on the mean, variance known

26 - 9.2, 9.3 - tests on the mean, variance unknown, omit 9.3.2

27 - 9.5 - tests on a population proportion

28 - 10.2 - inference on the difference of means, variances unknown, omit 10.2.2

29 - 11.1, 11.2 - linear regression

30 - 11.3, 11.4 - properties of least squares estimators, hypothesis tests

31 - 11.5, 11.6 - confidence and prediction intervals, omit the CI

32 - 11.7, 11.8 - Adequacy of regression model, correlation, omit the test and CI for p

33 - 13.2 - analysis of variance, omit 13.2.5

34 - 13.2 (Cont)

## BASIC PROBABILITY

Sample Space - set of all possible outcomes (ex. Roll of die: { 1- 6 })

Event - Any subset of the sample space (ex. Roll of die, odd numbers {1,3,5})

Probability - quantifying uncertainty of outcome of an experiment

- If E is an event, probability of E occurring is  $P(E)$

For an experiment with possible outcomes  $E_1, E_2, E_3$  (events 1-3), the probability of the event must be as followed:

- $0 \leq P(E_i) \leq 1$  for all  $E_i$
- $P(E_1) + P(E_2) + \dots + P(E_n) = 1$
- $P(\emptyset) = 0$
- The OR law for mutually exclusive events (more on this later).

Suppose an experiment has n possible outcomes, r of which satisfy some event E:

$$P(E) = r/n$$

## OR RULE (Mutually Exclusive Events)

Mutually exclusive - when two or more events cannot occur at the same time

$E$  and  $F$  are mutually exclusive if  $E \cap F = \emptyset$

OR Rule:

$$P(E \text{ or } F) = P(E \cup F) = P(E) + P(F)$$

$\cup$  = union of two sets

## AND RULE

$$P(E \text{ and } F) = P(E \cap F) = P(E)P(F | E),$$

where  $F | E$  means the occurrence of an event  $F$  given that an event  $E$  has already occurred.

## STATISTICAL INDEPENDENCE

Taking two interesting points from previous examples:

- AND and OR rules can be extended to more than two events
- Notation  $P(E, F)$  also means  $P(E \text{ and } F)$

Statistical Independence is often defined in terms of the AND rule

Two events  $E$  and  $F$  are said to be **statistically independent** if

$$P(E \cap F) = P(E)P(F).$$

- AKA - the outcome of one has no effect on the outcome of the other

If  $E$  and  $F$  are statistically independent, then the following are true:

$$\begin{aligned} P(F | E) &= P(F) \\ P(E | F) &= P(E) \end{aligned}$$

- This is Not mutually exclusive - they're very different!

If  $E$  and  $F$  are mutually exclusive, then the following are true:

$$\begin{aligned} P(F | E) &= 0 \\ P(E | F) &= 0 \\ P(E \cap F) &= 0 \end{aligned}$$

From now on: independent == statistically independent

## USEFUL TRICKS (+EX.)

For any event  $E$ , we use

$E$  or  $E^c$  to denote 'not  $E$ '.

- A coin is flipped 20 times, what is the probability that it shows heads at least once?

Let  $H$  be the event of a head, then,

$$P(\text{at least one } H) = 1 - P(\text{no } H) = 1 - \left(\frac{1}{2}\right)^{20} = \frac{1048575}{1048576}$$

CM 1.8

**Problem #8:** Heart failures are due to either natural occurrences (90%) or outside factors (10%). Outside factors are related to induced substances (78%) or foreign objects (22%). Natural occurrences are caused by arterial blockage (57%), disease (22%), and infection (21%).

- (a) Determine the probability that a failure is due to an induced substance.

- (b) Determine the probability that a failure is due to disease or infection.

$$a) 0.10(0.78) = 0.078 = \frac{39}{500}$$

$$b) 0.90(0.22) + 0.99(0.21) = 0.387 = \frac{387}{1000}$$

- A fair die is rolled (the experiment). What is the probability of it showing an even number? In this case,  $E$  is the event of an even number, there are  $n = 6$  possible outcomes, and there are  $r = 3$  even numbers, so

$$P(E) = \frac{r}{n} = \frac{3}{6} = \frac{1}{2}$$

- A fair die is rolled (the experiment). What is the probability of it showing a 5? In this case "5" is the event of rolling a 5, there are  $n = 6$  possible outcomes and there is  $r = 1$  number 5, so

$$P(5) = \frac{r}{n} = \frac{1}{6}$$

- A fair coin is tossed (the experiment). What is the probability of it showing a head? In this case  $H$  is the event of a head, there are  $n = 2$  possible outcomes, and there is  $r = 1$  head, so

$$P(H) = \frac{r}{n} = \frac{1}{2}$$

- A card is selected from a well shuffled pack (the experiment). What is the probability of it being a heart? In this case  $H$  is the event of a heart, there are  $n = 52$  possible outcomes, and there are  $r = 13$  hearts, so

$$P(H) = \frac{r}{n} = \frac{13}{52} = \frac{1}{4}$$

- A card is selected from a well shuffled pack (the experiment). What is the probability of it being a jack? In this case  $J$  is the event of a jack, there are  $n = 52$  possible outcomes, and there are  $r = 4$  jacks, so

$$P(J) = \frac{r}{n} = \frac{4}{52} = \frac{1}{13}$$

- A card is selected from a well shuffled pack. What is the probability of it being a jack or a 5? Since a single card cannot be both a jack and a 5, the events are mutually exclusive. Let  $J$  be the event of a jack and  $5$  be the event of rolling a 5, then

$$P(J \cup 5) = P(J) + P(5) = \frac{4}{52} + \frac{4}{52} = \frac{8}{52} = \frac{2}{13}$$

- A fair die is rolled. What is the probability of it landing on 1 or 6? A die cannot land on both 1 and 6 so the events are mutually exclusive. Let 1 be the event of rolling a 1 and 6 be the event of rolling a 6, then

$$P(1 \cup 6) = P(1) + P(6) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$$

- Two cards are selected from a well shuffled pack. What is the probability that they are both jacks? Let  $J_1$  be the event that the first card is a jack, and let  $J_2$  be the event that the second card is a jack, then

$$P(J_1 \cap J_2) = P(J_1)P(J_2 | J_1) = \frac{4}{52} \cdot \frac{3}{51} = \frac{12}{2652} = \frac{1}{221}$$

- There are 3 red and 4 green socks in a bag. Two socks are selected consecutively, at random. What is the probability that first sock is red and the second green? Let  $R_1$  be the event that the first sock is red, and let  $G_2$  be the event that the second sock is green, then

$$P(R_1 \cap G_2) = P(R_1)P(G_2 | R_1) = \frac{3}{7} \cdot \frac{4}{6} = \frac{2}{7}$$

- A card is selected from a well shuffled pack and a die is rolled. What is the probability of obtaining a club and a 3? Let  $C$  be the event that the card is a club and let 3 be the event that the die shows 3, then

$$P(C \cap 3) = P(C)P(3) = \frac{13}{52} \cdot \frac{1}{6} = \frac{1}{24}$$

- A card is selected from a well shuffled pack and a die is rolled. What is the probability of obtaining a red card and an even number? Let  $R$  be the event that the card is red and let  $E$  be the event that the die shows an even number, then

$$P(R \cap E) = P(R)P(E) = \frac{26}{52} \cdot \frac{3}{6} = \frac{1}{4}$$

Exercises:

- 1 A card is selected from a well shuffled pack (the experiment). What is the probability of it being red?

$$\frac{1}{2}$$

- 2 A sock is chosen (at random) from a bag containing 3 red and 4 blue socks. What is the probability that the chosen sock is blue?

$$\frac{4}{7}$$

- 3 Two fair dice are rolled.

- a) What is the probability that both numbers are odd? [Hint: write out all of the possible outcomes; there are 36.]  $\frac{9}{36}$
- b) What is the probability that both numbers are the same?

$$3.1 \quad \frac{3}{6} \cdot \frac{3}{6} = \frac{9}{36} = \frac{1}{4}$$

$$3.2 \quad \frac{1}{6}$$

- 4 There are 3 red, 4 green and 7 blue socks in a bag. Three socks are selected consecutively, at random. What is the probability that the first is a red, the second green and the third blue?

$$\frac{3}{14} \cdot \frac{4}{13} \cdot \frac{7}{12} = \frac{84}{2744}$$

- 5 A card is selected from a well shuffled pack and a coin is tossed. What is the probability of obtaining a queen and a tail?

$$\frac{1}{13} \cdot \frac{1}{2} = \frac{1}{26}$$

**GENERAL OR RULE**

For any two events E and F, the probability of E OR F occurring is given by

$$P(E \cup F) = P(E) + P(F) - P(E \cap F).$$

Note that if E and F are mutually exclusive then  $P(E \cap F) = 0$  and we have the familiar formula

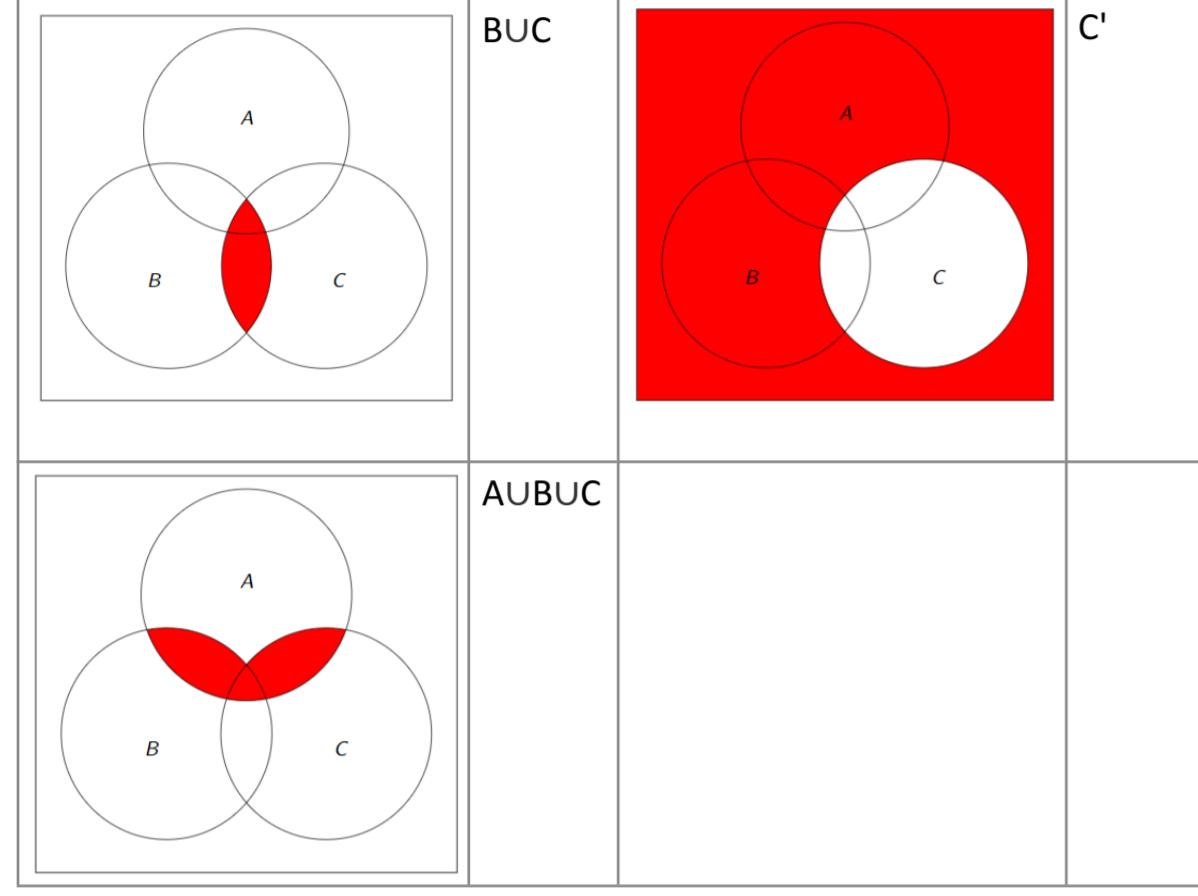
$$P(E \cup F) = P(E) + P(F).$$

**RELATIONSHIP TO SET THEORY**

When sets intersect, adding the elements of each set together results in certain elements being accounted for twice. To get the correct answer, we need to add the elements of each set and subtract the elements in the intersection.

**VENN DIAGRAMS**

These are useful tools for depicting each of the scenarios we've learned so far

**INTRODUCTION TO COUNTING**

We need to count different permutations to solve problems, starting with arranging items in a line.

**FACTORIAL NOTATION**The number of ways of arranging  $n$  DISTINCT objects in a line is:

$$n! = n \cdot (n-1) \cdot (n-2) \cdots 3 \cdot 2 \cdot 1.$$

 $n!$  is called an **n factorial**The number of ways of arranging  $n$  objects, of which  $m$  are identical is:

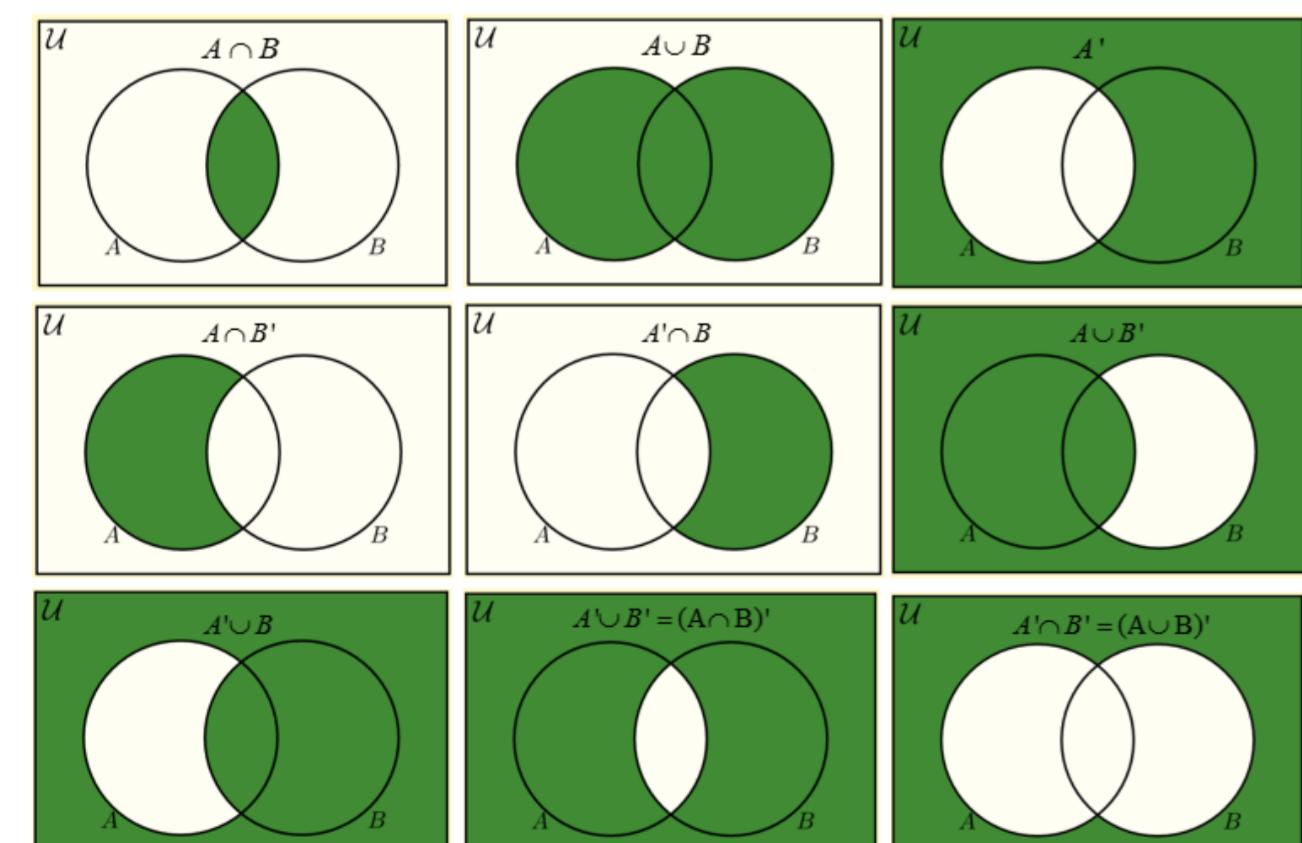
$$\frac{n!}{m!}.$$

**COMBINATORICS**The number of ways of choosing  $r$  items from  $n$  distinct items (in any order) is:

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

If a task can be accomplished  $n$  ways and then a second task may be accomplished  $m$  ways, then the first task followed by the second task may be accomplished in  $nm$  different ways.**PROBABILITY + COUNTING**There are three common probability distributions that use  $nCr$ :

- Hypergeometric distribution
- Binomial distribution
- Negative binomial distribution



1. A card is selected from a well shuffled pack. What is the probability of it being a jack or a red card? Let  $J$  be the event that the card is a Jack and let  $R$  be the event that the card is red. Since there are two red jacks, the answer is

$$P(R \cup J) = P(R) + P(J) - P(R \cap J) = \frac{26}{52} + \frac{4}{52} - \frac{2}{52} = \frac{28}{52} = \frac{7}{13}.$$

2. A card is selected from a well shuffled pack. What is the probability of it being an ace or a spade? Let  $A$  be the event that the card is an Ace. Let  $S$  be the event that the card is a spade. Since there is one ace of spades, the answer is

$$P(A \cup S) = P(A) + P(S) - P(A \cap S) = \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52} = \frac{4}{13}.$$

3. A die is rolled. What is the probability that it shows 5 or an odd number? Let  $5$  be the event of rolling a 5 and let  $O$  be the event of rolling an odd number. Since 5 is an odd number, the answer is

$$P(5 \cup O) = P(5) + P(O) - P(5 \cap O) = \frac{1}{6} + \frac{3}{6} - \frac{1}{6} = \frac{3}{6} = \frac{1}{2}.$$

- All of these answers make sense if we write out all of the possible outcomes; the possible outcomes for the last example are  $\{1, 2, 3, 4, 5, 6\}$ , three of which are either 5 or odd.

$$\text{(a)} P(Q \cup R) = P(Q) + P(R) - P(Q \cap R) \\ = \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52}$$

$$\text{(b)} P(3 \cup B) = P(3) + P(B) - P(3 \cap B) \\ = \frac{4}{52} + \frac{26}{52} - \frac{52}{52} = \frac{28}{52}$$

$$\text{(c)} P(O \cup D) = P(O) + P(D) - P(O \cap D) \\ = \frac{20}{52} + \frac{13}{52} - \frac{5}{52} = \frac{28}{52}$$

1. A card is chosen from a well shuffled pack. What is the probability that the card is a queen or a heart?
2. A card is chosen from a well shuffled pack. What is the probability that the card is a 3 or a black card?
3. A card is chosen from a well shuffled pack. What is the probability that the card is an odd numbered card or a diamond?

1. A card is chosen from a well shuffled pack. What is the probability that it is:

- (1) A queen or a heart?  
(2) A 3 or a black card?  
(3) An odd numbered card or a diamond?

2. A bag contains nine tickets numbered 1 – 9; tickets 1 – 3 are red, tickets 4 – 6 are blue and tickets 7 – 9 are green. A ticket is selected at random. What is the probability that:

- (1) It is blue?  
(2) It has an odd number on it?  
(3) It is blue or it has an odd number on it?

$$\text{(a)} \frac{1}{3} \quad \text{(b)} \frac{5}{9} \quad \text{(c)} P(B \cup O) = P(B) + P(O) - P(B \cap O) \\ = \frac{3}{9} + \frac{5}{9} - \frac{1}{9} \\ = \frac{7}{9}$$

1. How many eight letter 'words' can be formed from the letters HAMILTON?

$$8!$$

2. There are 14 teams in a soccer league. In how many different orders can the teams finish? You may assume that two teams cannot be tied.

$$14!$$

1. How many ways can I order four soccer players in the middle-field position?

$$4!$$

2. How many ways can the letters STATISTICS be arranged to make a 10 letter 'word'?

$$\frac{10!}{3! 3! 2!}$$

3. How many ways can the letters SCIENCE be arranged if all letters must be used?

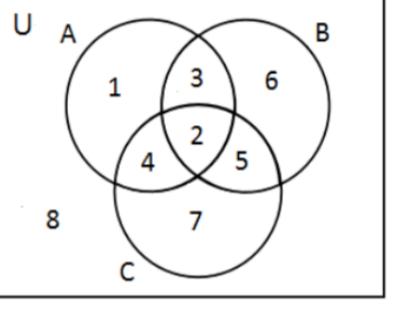
$$\frac{7!}{2! 2!}$$

4. How many ways can the letters ONTARIO be arranged if all letters must be used?

$$\frac{7!}{2!}$$

**CM 1.2**

**Problem #2:** Consider the Venn diagram given to the right. In each part, determine which of the regions corresponds to the given statements.



(a)  $A \cup (B \cap C)$

(b)  $A \cup B$

(c)  $A' \cap C'$

(d)  $A \cup B'$

a)  $A \cup (B \cap C)$

⇒ 1, 2, 3, 4, 5

b)  $A \cup B$

⇒ 1, 2, 3, 4, 5, 6

c)  $A' \cap C'$

⇒ 6, 8

d)  $A \cup B'$

⇒ 1, 2, 3, 4, 7, 8

union ↘

intersect. ↗

c)  $A' \cap C'$

⇒ 6, 8

union ↘

d)  $A \cup B'$

⇒ 1, 2, 3, 4, 7, 8

union ↘

intersect. ↗

c)  $A' \cap C'$

⇒ 6, 8

union ↘

d)  $A \cup B'$

⇒ 1, 2, 3, 4, 7, 8

union ↘

intersect. ↗

c)  $A' \cap C'$

⇒ 6, 8

union ↘

d)  $A \cup B'$

⇒ 1, 2, 3, 4, 7, 8

union ↘

intersect. ↗

c)  $A' \cap C'$

⇒ 6, 8

union ↘

d)  $A \cup B'$

⇒ 1, 2, 3, 4, 7, 8

union ↘

intersect. ↗

c)  $A' \cap C'$

⇒ 6, 8

union ↘

d)  $A \cup B'$

⇒ 1, 2, 3, 4, 7, 8

union ↘

intersect. ↗

c)  $A' \cap C'$

⇒ 6, 8

union ↘

d)  $A \cup B'$

⇒ 1, 2, 3, 4, 7, 8

union ↘

intersect. ↗

c)  $A' \cap C'$

⇒ 6, 8

union ↘

d)  $A \cup B'$

⇒ 1, 2, 3, 4, 7, 8

union ↘

intersect. ↗

c)  $A' \cap C'$

⇒ 6, 8

union ↘

d)  $A \cup B'$

⇒ 1, 2, 3, 4, 7, 8

union ↘

intersect. ↗

c)  $A' \cap C'$

⇒ 6, 8

union ↘

d)  $A \cup B'$

⇒ 1, 2, 3, 4, 7, 8

union ↘

intersect. ↗

c)  $A' \cap C'$

# Lec 3 - Permutations, Conditional Prob, Partition Theorem

Monday, September 12, 2022 2:04 PM

Yuxi Qin  
[yuxiqin.ca](http://yuxiqin.ca)

## COMBINATION VS PERMUTATION

Permutation: an ordered subset

Combination: an unordered subset

Formally, the number of permutations (ways of arranging)  $n$  distinct items that are taken  $r$  at a time is:

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

## CONDITIONAL PROBABILITY

AND Rule: if  $E$  and  $F$  are two events then

$$P(E \cap F) = P(E)P(F | E),$$

where  $F | E$  means the occurrence of event  $F$  given that event  $E$  has already occurred

Dividing both sides of the above equation by  $P(E)$  gives the definition of

**conditional probability:** the probability that event  $F$  occurs given that event  $E$  has already occurred

$$P(F | E) = \frac{P(E \cap F)}{P(E)}.$$

We can now rewrite the term  $P(E)$  in equation 1...

$$P(F | E) = \frac{P(E | F)P(F)}{P(E)}$$

## THE PARTITION THEOREM

This approach can be generalized to get a general formula for  $P(E)$  in terms of **conditional probabilities**

**Partition Theorem:** Suppose the outcome of an event  $E$  depends on an event  $F$  which has possible outcomes  $F_1, F_2, \dots, F_n$ , then

$$P(E) = \sum_{i=1}^n P(E | F_i)P(F_i)$$

In the lung cancer example above, using the Partition Theorem;

$$E = C, F_1 = S, F_2 = S'$$

## MORE LAWS

**De Morgan's Laws:** for any event  $A$  and  $B$ :

$$\begin{aligned} ① \quad & \overline{A \cup B} = \overline{A} \cap \overline{B} \\ ② \quad & \overline{A \cap B} = \overline{A} \cup \overline{B} \end{aligned}$$

For any events  $A$  and  $B$  with  $P(B) > 0$

$$P(A|B) + P(\overline{A}|B) = 1$$

If  $A$  and  $B$  are independent events,  $A$  and  $B'$  are also independent

## CM 1.4

**Problem #4:** Suppose that a code (similar to a postal code) is of the form LDL DLD, where 'L' is an uppercase letter from A to R (i.e., 18 possible letters) and 'D' is a digit from 0 to 7. Suppose that such a code is randomly generated.

- (a) Find the probability that the code has no repeated digits.
- (b) Find the probability that the code either starts with an 'A' or ends with an even digit (note that 0 is even).
- (c) Find the probability that the code starts with an 'A' and does not contain any 'B's.

$$\begin{aligned} a) \quad & \frac{8 \cdot 7 \cdot 6}{8^3} = \frac{21}{32} & b) \quad & \frac{1}{18} + \frac{4}{8} - \frac{1}{18} \left( \frac{4}{8} \right) = \frac{19}{36} & c) \quad & \left( \frac{1}{18} \right) \left( \frac{17 \cdot 17}{18^2} \right) = \frac{289}{5832} \\ & \text{1st letter } \downarrow \quad \text{last } \downarrow \quad \text{intersection of both} & & & & \text{A } \quad \text{2 possible places for B (letter)} \end{aligned}$$

## CM 1.6

**Problem #6:** A production facility employs 15 workers on the day shift, 12 workers on the swing shift, and 8 workers on the graveyard shift. A quality control consultant is to randomly select 7 of these workers for in-depth interviews.

- (a) What is the probability that all 7 selected workers will be from the same shift?
- (b) What is the probability that at least two different shifts will be represented among the selected workers?
- (c) What is the probability that exactly 2 of the workers in the sample come from the day shift?

$$\begin{aligned} a) \quad & p(\text{same shift}) = \frac{\binom{15}{7} + \binom{12}{7} + \binom{8}{7}}{\binom{35}{7} \text{ total } \#} = \frac{1447}{1344907} = 0.00107591 \\ b) \quad & 1 - p(\text{all same shift}) = 1 - \frac{1447}{1344907} = 0.99892408 \\ c) \quad & \frac{\binom{15}{2} + \binom{20}{5}}{\binom{35}{7}} = \frac{2394}{9889} = 0.24208716 \end{aligned}$$

1. a) How many ways can we arrange the four letters  $a, b, c$ , and  $d$ ? From last lecture, there would be  $4! = 24$  ways to arrange the four letters.

- b) How many ways can we arrange only two of the four letters? From the previous slide, there would be  ${}^4 P_2 = \frac{4!}{2!} = 12$  ways to arrange four of the letters taken 2 at a time.

- c) How many ways can we pick two out of the four letters? From last lecture, there would be  $\binom{4}{2} = {}^4 C_2 = \frac{4!}{2!2!} = 6$  ways to choose 2 distinct letters from the four.

- ① Consider the lottery question from Lecture 2. If 6 numbers must be selected, and the order of the numbers matters. For example, a line reading 2,3,4,5,6,7 is different from 3,2,4,5,6,7. What is the probability of the winning the lottery if a single ticket is purchased?

$$P(\text{winning}) = \frac{1}{{}^9 P_6} = \frac{1}{10068347520}$$

- So, why is it the case that  $P(C) = P(C \cap S) + P(C \cap \bar{S})$ ?

- There are two ways to rationalize this answer.

1. There are two groups that suffer from cancer – smokers and non-smokers – so we add each group's probabilities

2. Or, in a more technical language, the set  $C$  is partitioned by  $S$  and  $\bar{S}$ , giving the result.

1. In a region, 31% of people are smokers, 19% of smokers develop lung cancer and 2% of non-smokers develop lung cancer. What is the probability that a person chosen at random will develop lung cancer; i.e. what proportion of this region will suffer from lung cancer?

- Let  $S$  be the event that a person smokes, and  $C$  be the event that a person develops lung cancer.

- So,  $P(S) = 0.31$ , it follows that  $P(\bar{S}) = 1 - 0.31 = 0.69$ . Furthermore, we know that  $P(C | S) = 0.19$  and  $P(C | \bar{S}) = 0.02$ . We deduce  $P(C)$  as follows:

$$\begin{aligned} P(C) &= P(C \cap S) + P(C \cap \bar{S}) = P(C | S)P(S) + P(C | \bar{S})P(\bar{S}) \\ &= (0.19)(0.31) + (0.02)(0.69) = 0.0589 + 0.0138 = 0.0727 \end{aligned}$$

Therefore, the answer is 7.27%.

- So, we need to find the probability that the van man will return with 'X' and the probability that the van man will return with 'Y'. From the partition theorem, we have

$$\begin{aligned} P(X) &= P(X | A)P(A) + P(X | B)P(B) + P(X | C)P(C) \\ &= (10/40)(0.6) + (30/50)(0.2) + (30/40)(0.2) \\ &= 0.15 + 0.12 + 0.15 = 0.42 \end{aligned}$$

Similarly,

$$\begin{aligned} P(Y) &= P(Y | A)P(A) + P(Y | B)P(B) + P(Y | C)P(C) \\ &= (30/40)(0.6) + (20/50)(0.2) + (10/40)(0.2) \\ &= 0.45 + 0.08 + 0.05 = 0.58 \end{aligned}$$

Therefore, it is more likely that the van man returns with the 'Y' roll.

1. A building contract requires a roll of roofing felt. There are three suppliers ( $A, B$ , and  $C$ ) in the area and the probabilities that the contractor will instruct his van man to visit a particular supplier are:  $P(A) = 0.6$ ,  $P(B) = 0.2$ , and  $P(C) = 0.2$ .

- Each supplier stocks roofing felt produced by one of two manufacturers,  $X$  and  $Y$ . The stock situation at each of the suppliers is shown in the table below.

Supplier	# of 'X' rolls	# of 'Y' rolls
A	10	30
B	30	20
C	30	10

Given that the van man will be told which supplier to visit, which type of roofing felt is the van man most likely to return with?

2. Personal computers are assembled on two production lines, 60% are assembled on Line 1 and 40% on Line 2. QC records show that both lines are not equally reliable: 95% of units assembled by Line 1 require no rework, while the figure for Line 2 is 88%.

- What percentage of all computers require rework?
- If a computer is found to require rework, what is the probability that it came from Line 1?

- 2a)  $RW \rightarrow \text{computer reworks}$   
 $A_1 \rightarrow \text{computer assembled on line 1}$   
 $A_2 \rightarrow \text{computer assembled on line 2}$

$$\begin{aligned} P(A_1) &= 0.60 & P(RW | A_1) &= 1 - 0.95 = 0.05 \\ P(A_2) &= 0.40 & P(RW | A_2) &= 1 - 0.88 = 0.12 \end{aligned}$$

$$\begin{aligned} P(RW) &= P(RW \cap A_1) + P(RW \cap A_2) \\ &= P(RW \cap A_1)P(A_1) + P(RW \cap A_2)P(A_2) \\ &= 0.05 \cdot 0.60 + 0.12 \cdot 0.40 \\ &= 0.078 \end{aligned}$$

## CM 1.10

**Problem #10:** Suppose that an operating room needs to schedule 4 knee, 5 hip, and 5 shoulder surgeries. Assume that all schedules are equally likely.

$$n = 14$$

- (a) Find the probability that all of the knee surgeries are completed first.

- (b) Find the probability that the schedule begins with a hip surgery, given that all of the shoulder surgeries are last.

$$a) \quad \frac{4}{14} \left( \frac{13}{13} \right) \left( \frac{12}{12} \right) \left( \frac{1}{11} \right) = \frac{1}{1001} \quad \text{KKKK } \underbrace{\text{XXXX XXXXXX}}_{\text{don't care}}$$

$$b) \quad \frac{5}{14} \quad \begin{array}{c} 4K \ 5H \ X \\ \swarrow \quad \searrow \\ n=9 \end{array} \quad P(H) = \frac{5}{9}$$

# Lec 4 - Examples, Probability

Wednesday, September 14, 2022 1:30 PM

Yuxi Qin  
yuxiqin.ca

## EXAMPLES

1. A printed circuit board has 8 different locations in which a component can be placed. If 5 identical components are to be placed on the board, how many different designs are possible?

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

$$\frac{n=8}{r=5} \quad nC_r = \frac{8!}{5!(8-5)!} = 56$$

2. A bin of 50 manufactured parts contains 3 defective parts and 47 non-defective parts. A sample of 6 parts is selected from 50 parts without replacement.

- a: How many different samples are there of size 6 that contain exactly 2 defective parts?

$$50 \text{ total} \quad \text{select 2 def parts from 3?}$$

$$\text{def: } 3 \quad \binom{3}{2} = 3C_2 = \frac{3!}{2!(3-2)!} = 3$$

$$\text{non def: } 47 \quad \binom{47}{4} = \frac{47!}{4!(47-4)!} = 178,365$$

$$n_1 C_{n_1} \cdot n_2 C_{n_2} = 3(178,365)$$

$$a/b = 535,095$$

- b: What is the probability that exactly 2 defective parts are selected in the sample?

$$P(2D) = \frac{\# \text{ of samples w/ 2 defective}}{\text{total # of size 6}}$$

$$nC_r (\text{total}) = \frac{50!}{6!(50-6)!} = 15,890,700$$

$$P(2D) = \frac{535,095}{15,890,700} = 0.034$$

## PERMUTATIONS OF SIMILAR OBJECTS

The number of permutations for

$$n = n_1 + n_2 + \dots + n_r$$

Objects of where  $n_1$  are one type,  $n_2$  of another type, etc.. Is

$$\frac{n!}{n_1! n_2! \dots n_r!}$$

## CONDITIONAL PROBABILITY

Consider the question: A bin of 50 manufactured parts contains 3 defective parts and 47 non-defective parts.

## THE PARTITION THEOREM

P47 of 6th edition textbook - ex. 2.28

**Example 2-28 Semiconductor Failures** Continuing with semiconductor manufacturing, assume the following probabilities for product failure subject to levels of contamination in manufacturing:

Probability of Failure	Level of Contamination
0.10	High
0.01	Medium
0.001	Low

In a particular production run, 20% of the chips are subjected to high levels of contamination, 30% to medium levels of contamination, and 50% to low levels of contamination. What is the probability that a product using one of these chips fails? Let

- $H$  denote the event that a chip is exposed to high levels of contamination
- $M$  denote the event that a chip is exposed to medium levels of contamination
- $L$  denote the event that a chip is exposed to low levels of contamination

Then,

$$P(F) = P(F|H)P(H) + P(F|M)P(M) + P(F|L)P(L)$$

$$= 0.10(0.20) + 0.01(0.30) + 0.001(0.50) = 0.0235$$

The calculations are conveniently organized with the tree diagram in Fig. 2-17.

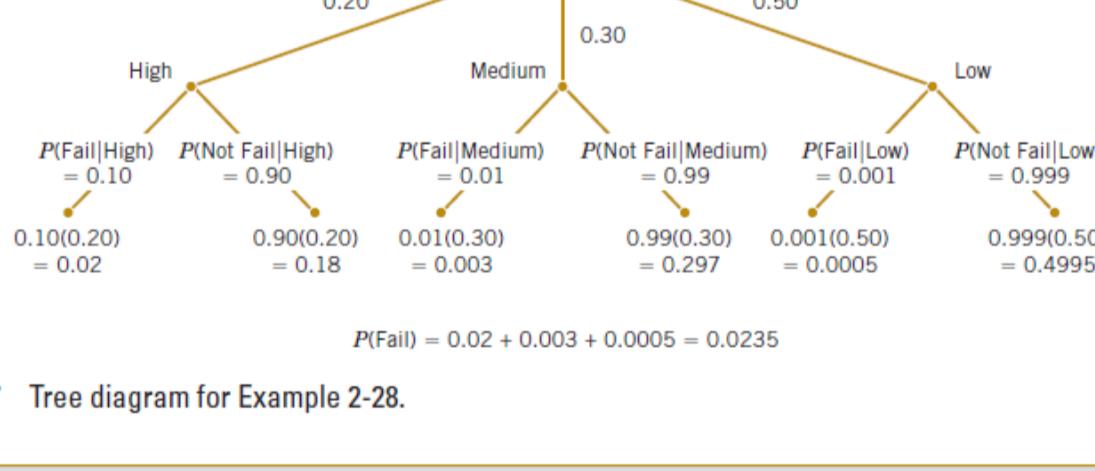


FIGURE 2-17 Tree diagram for Example 2-28.

**Example 2-12 Bar Code 39** Code 39 is a common bar code system that consists of narrow and wide bars (black) separated by either wide or narrow spaces (white). Each character contains nine elements (five bars and four spaces). The code for a character starts and ends with a bar (either narrow or wide) and a (white) space appears between each bar. The original specification (since revised) used exactly two wide bars and one wide space in each character. For example, if  $b$  and  $B$  denote narrow and wide (black) bars, respectively, and  $w$  and  $W$  denote narrow and wide (white) spaces, a valid character is  $bwbBwbwb$  (the number 6). One character is held back as a start and stop delimiter. How many other characters can be coded by this system? Can you explain the name of the system?

1 character contains 9 "elements" → 5 bars, 4 spaces  
 $\boxed{b\ b\ b\ B\ b\ b\ b}$  → each char has 2 wide bars, 1 wide space

How many chars can be coded?

let  $b = \text{narrow black}$   
 $B = \text{wide black}$   
 $w = \text{narrow white}$   
 $W = \text{wide white}$

1. Calculate # of permutations of 5 black bars

$$B=2 \quad b=3 \quad n = \frac{n!}{n_1! n_2! \dots n_n!}$$

$$\frac{n=5}{n_1=2} \quad \frac{n_2=3}{n_3=2} \quad = \frac{5!}{2! 3!} = 10$$

2. White spaces: 3 w & 1 W

$$w=4 \quad w_1=3 \quad w_2=1 \quad n = \frac{4!}{3! 1!} = 4$$

$$\text{Ans: } 10 \times 4 = 40 \text{ codes (- delimiter)}$$

$$= 40 - 1 = 39 \text{ codes}$$

A hospital operating room needs to schedule three knee surgeries and two hip surgeries in a day. In how many different orders can the operations take place?

$$n=5 \quad n_1 = 3 \text{ (knee)} \quad n_2 = 2 \text{ (hip)}$$

$$n = \frac{n!}{n_1! n_2!} = \frac{5!}{3! 2!} = 10$$

## CM 1.1

In the layout of a printed circuit board for an electronic product, 13 different locations can accommodate chips.

- (a) If 5 chips of different types are to be placed on the board, how many different layouts are possible?  
(b) If all of the locations are to be filled with chips, 2 of which are of one type, 6 of which are another type, and all others different, how many different layouts are possible?

- (c) If 5 chips of the same type are to be placed on the board, how many different layouts are possible?

$$\text{a) } {}^{13}P_5 = \frac{13!}{(13-5)!} = \frac{13!}{8!} = 154,440$$

↳ on calculator:  ${}^{13}P_5$

$$\text{b) } \frac{13!}{2! 6!} = 4,324,320$$

$$\text{c) } {}^{13}C_5 = \frac{13!}{(13-5)! 5!} = 12,870$$

↳ on calculator:  ${}^{13}C_5$

## CM 1.9

**Problem #9:** A lot of 93 semiconductor chips contains 20 that are defective.

- (a) Two are selected, one at a time and without replacement from the lot. Determine the probability that the second one is defective.

- (b) Three are selected, one at a time and without replacement. Find the probability that the first one is defective and the third one is not defective.

$$\text{a) } \frac{20}{93} \left( \frac{19}{92} \right) + \frac{73}{93} \left( \frac{20}{92} \right) = \frac{20}{93} \quad \text{DD + ND}$$

$$\text{b) } \frac{20}{93} \left( \frac{19}{92} \right) \left( \frac{73}{91} \right) + \left( \frac{20}{93} \right) \left( \frac{73}{92} \right) \left( \frac{72}{91} \right) = \frac{365}{2139} \quad \text{DDN + DNN}$$

1. What is the probability that the second part is defective given that the first part is defective?

$$P(B|A) \quad 3 \text{ defective, } 47 \text{ non defective } \quad j/n = 50$$

1st piece is defective —  
49 pieces left, 2 of which are defective :

$$\text{ans: } P(B|A) = \frac{2}{49}$$

2. What is the probability that the first two parts are defective and the third is not defective?

$$P(B|A) = \frac{2}{49}$$

48 pieces left, 1 defective →  $\frac{47}{48}$

$$P(D, D, N) = \frac{3}{50} \cdot \frac{2}{49} \cdot \frac{47}{48} = 0.0024$$

**Example 2-12 Bar Code 39** Code 39 is a common bar code system that consists of narrow and wide bars (black) separated by either wide or narrow spaces (white). Each character contains nine elements (five bars and four spaces). The code for a character starts and ends with a bar (either narrow or wide) and a (white) space appears between each bar. The original specification (since revised) used exactly two wide bars and one wide space in each character. For example, if  $b$  and  $B$  denote narrow and wide (black) bars, respectively, and  $w$  and  $W$  denote narrow and wide (white) spaces, a valid character is  $bwbBwbwb$  (the number 6). One character is held back as a start and stop delimiter. How many other characters can be coded by this system? Can you explain the name of the system?

1 character contains 9 "elements" → 5 bars, 4 spaces  
 $\boxed{b\ b\ b\ B\ b\ b\ b}$  → each char has 2 wide bars, 1 wide space

How many chars can be coded?

let  $b = \text{narrow black}$   
 $B = \text{wide black}$   
 $w = \text{narrow white}$   
 $W = \text{wide white}$

1. Calculate # of permutations of 5 black bars

$$B=2 \quad b=3 \quad n = \frac{n!}{n_1! n_2! \dots n_n!}$$

$$\frac{n=5}{n_1=2} \quad \frac{n_2=3}{n_3=2} \quad = \frac{5!}{2! 3!} = 10$$

2. White spaces: 3 w & 1 W

$$w=4 \quad w_1=3 \quad w_2=1 \quad n = \frac{4!}{3! 1!} = 4$$

$$\text{Ans: } 10 \times 4 = 40 \text{ codes (- delimiter)}$$

$$= 40 - 1 = 39 \text{ codes}$$

# Lec 5 - Independence, Random Variables, and Discrete Random Variables

Thursday, September 22, 2022 1:24 PM

## STATISTICAL INDEPENDENCE AND rules for events E and F:

$$P(E \text{ and } F) = P(E \cap F) = P(E)P(F | E)$$

- $F | E$  means the occurrence of an event F given that an event E has already occurred (F assuming E)

The notation  $P(E, F)$  also means  $P(E \text{ and } F)$

- There's a special case of the AND rule, where the events are independent

E and F are said to be statistically independent if

$$P(E \cap F) = P(E)P(F)$$

- Two events are independent if the outcome of one has no effect on the outcome of the other

If E and F are statistically independent, then these rules apply:

$$\begin{aligned} P(F | E) &= P(F) \\ P(E | F) &= P(E) \end{aligned}$$

- Very different from mutually exclusive

If E and F are mutually exclusive, then these rules apply:

$$\begin{aligned} P(F | E) &= 0 \\ P(E | F) &= 0 \\ P(E \cap F) &= 0 \end{aligned}$$

From now on, INDEPENDENT = STATISTICALLY INDEPENDENT

## RANDOM VARIABLES

Interval Data

- Continuous** data arises when all values are possible inside some interval
  - Ex. Volume of water inside a glass, distance between cities, height (any value between 0 and 2.3), age (any value between 0 and 110 years), weight (any value between 0.3 and 220kg), weight (could take any value between 0.3 and 220kg)
- Discrete** data arises when there are a finite number of possible values
  - Ex. Number of bacteria on a piece of raw meat, hurricanes that hit land each year

A random variable is a function from the sample space to the real numbers (random experiment)

- Ex. Tossing a coin 6 times - we have a random variable representing the number of heads observed
- $P(X=4) \rightarrow X$  is called a discrete random variable

Ex. Temperature data

- Consider the following univariate data set measuring the air temperature at LaGaurdia airport
  - Univariate - involves one variable quantity
  - It is composed of "continuous" interval data type
  - Has 25 observations in the temperature data set

Observation	Temperature (°F)	Observation	Temperature (°F)
1	85	14	81
2	87	15	79
3	79	16	74
4	86	17	79
5	88	18	77
6	68	19	61
7	81	20	92
8	86	21	72
9	58	22	76
10	64	23	72
11	57	24	79
12	83	25	83
13	66		

CM 2.2

**Problem #2:** Consider purchasing a system of audio components consisting of a receiver, a pair of speakers, and a CD player. Let  $A_1$  be the event that the receiver functions properly throughout the warranty period. Let  $A_2$  be the event that the speakers function properly throughout the warranty period. Let  $A_3$  be the event that the CD player functions properly throughout the warranty period. Suppose that these events are (mutually) independent with  $P(A_1) = 0.98$ ,  $P(A_2) = 0.82$ , and  $P(A_3) = 0.74$ .

- (a) What is the probability that at least one component needs service during the warranty period?  
(b) What is the probability that exactly one of the components needs service during the warranty period?

$$P(A_1) = 0.98 \quad P(\bar{A}_1) = 0.02$$

$$P(A_2) = 0.82 \quad P(\bar{A}_2) = 0.18$$

$$P(A_3) = 0.74 \quad P(\bar{A}_3) = 0.26$$

$$\begin{aligned} a) P(\text{at least one}) &= 1 - P(\text{all work}) \\ &= 1 - 0.98(0.82)(0.74) \\ &= 1 - 0.594664 \\ &= 0.405336 \quad \checkmark \end{aligned}$$

$$\begin{aligned} b) P(\text{exactly one}) &= P(\text{only receiver}) + P(\text{only speaker}) + P(\text{only CD}) \\ &= P(\bar{A}_1)P(A_2)P(A_3) + P(A_1)P(\bar{A}_2)P(A_3) + P(A_1)P(A_2)P(\bar{A}_3) \\ &= 0.02(0.82)(0.74) + 0.98(0.18)(0.74) + 0.98(0.82)P(0.26) \\ &= 0.012136 + 0.130536 + 0.208936 \\ &= 0.351608 \quad \checkmark \end{aligned}$$

CM 2.3

**Problem #3:** A box consists of 13 components, 7 of which are defective.

13 total  
 $\hookrightarrow$  7 defective (D)  
 $\hookrightarrow$  6 non-defective (N)

- (a) Components are selected and tested one at a time, without replacement, until a non-defective component is found. Let  $X$  be the number of tests required. Find  $P(X=5)$ .

- (b) Components are selected and tested, one at a time without replacement, until two consecutive non-defective components are obtained. Let  $X$  be the number of tests required. Find  $P(X=5)$ .

$$a) \text{Find } P(X=5) \rightarrow \text{DDDDN}$$

$$= \frac{7}{13} \left( \frac{6}{12} \right) \left( \frac{5}{11} \right) \left( \frac{4}{10} \right) \left( \frac{6}{9} \right)$$

- diff. orders  
of N & D

- it takes 5 components to  
get NN, so we know it's  
at the end.

$$b) \text{Find } P(X=5) \quad N = \text{Non defective} \quad D = \text{Defective}$$

$$\rightarrow \text{DDDN} \quad \rightarrow \text{NDDNN} \quad \rightarrow \text{DNDNN}$$

$$= \frac{7}{13} \left( \frac{6}{12} \right) \left( \frac{5}{11} \right) \left( \frac{4}{10} \right) \quad = \frac{6}{13} \left( \frac{7}{12} \right) \left( \frac{6}{11} \right) \left( \frac{5}{10} \right) \quad = \frac{7}{13} \left( \frac{6}{12} \right) \left( \frac{5}{11} \right) \left( \frac{4}{10} \right)$$

$$= \frac{35}{858} \quad = \frac{14}{429} \quad = \frac{14}{429}$$

$$P(X=5) = \frac{35}{858} + \frac{14}{429} + \frac{14}{429} = \frac{7}{66} \quad \checkmark$$

Yuxi Qin  
[yuxiqin.ca](http://yuxiqin.ca)

## EXAMPLE 2.21 | Sampling with Replacement

Consider the inspection described in Example 2.11. Six parts are selected randomly from a bin of 50 parts, but assume that the selected part is replaced before the next one is selected.

The bin contains 3 defective parts and 47 nondefective parts. What is the probability that the second part is defective given that the first part is defective?

In shorthand notation, the requested probability is  $P(B | A)$ , where  $A$  and  $B$  denote the events that the first and second parts are defective, respectively. Because the first part

w/ replacement!

is replaced prior to selecting the second part, the bin still contains 50 parts, of which 3 are defective. Therefore, the probability of  $B$  does not depend on whether or not the first part is defective. That is,

$$P(B | A) = \frac{3}{50}$$

Also, the probability that both parts are defective is

$$P(A \cap B) = P(B | A)P(A) = \frac{3}{50} \cdot \frac{3}{50} = \frac{9}{2500}$$

chances of A given B

The probability that a wafer contains a particle of contamination is 0.01, and wafer production are independent events.

If 15 wafers are analyzed, what is the probability no particles are found?

$$\begin{aligned} P(C) &= 0.01 \rightarrow \text{given} & \text{independent events} &\rightarrow \text{multiply probabilities} \\ P(\bar{C}) &= 0.99 \\ n &= 15 \\ P(\bar{C}_{15}) &= (0.99)^{15} = 0.86 \end{aligned}$$

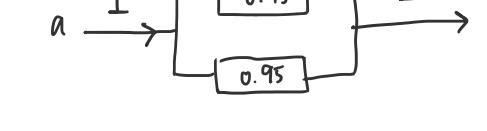
Note: independence (multiple events)

- If  $E_1, E_2, \dots, E_n$  are independent then

$$P(E_1 \cap E_2 \cap \dots \cap E_n) = P(E_1)P(E_2) \dots P(E_n)$$

Circuit in parallel -> only one component needs to work (for this arrangement)

- Devices fail independently



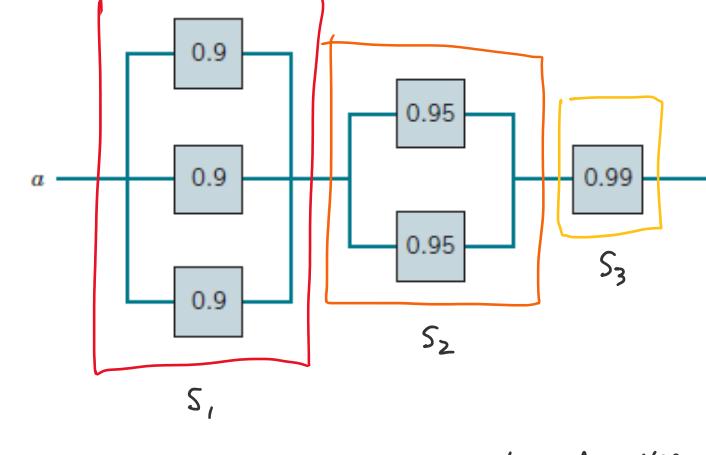
$$P(W) = 0.95 \quad P(F) = 1 - 0.95 = 0.05 \rightarrow \text{each device has a 0.05 chance of failure}$$

What is the probability the circuit works?

$$1 - P(\text{circuit fails}) = 1 - (0.05^2) = 0.9975$$

## EXAMPLE 2.25 | Advanced Circuit

The following circuit operates only if there is a path of functional devices from left to right. The probability that each device functions is shown on the graph. Assume that devices fail independently. What is the probability that the circuit operates?



$S_1, S_2, S_3$  must all work for the current to flow

$\rightarrow$  work out prob. each  $S_i$  fails

$$P(\bar{S}_i) = S_i \text{ fails}$$

$$P(S_1)P(S_2)P(S_3)$$

$$= (1 - 0.1^3)(1 - 0.05^2)(0.99)$$

$$= 0.987$$

The solution can be obtained from a partition of the graph into three columns. Let  $L$  denote the event that there is a path of functional devices only through the three units on the left. From the independence and based on the previous example,

$$P(L) = 1 - 0.1^3$$

Similarly, let  $M$  denote the event that there is a path of functional devices only through the two units in the middle. Then,

$$P(M) = 1 - 0.05^2$$

The probability that there is a path of functional devices only through the one unit on the right is simply the probability that the device functions, namely, 0.99. Therefore, with the independence assumption used again, the solution is

$$(1 - 0.1^3)(1 - 0.05^2)(0.99) = 0.987$$

$$\begin{aligned} P(\bar{S}_1) &= (0.1)^3 \\ P(S_1) &= (1 - (0.1)^3) \end{aligned} \rightarrow (1 - 0.1)(1 - 0.1)(1 - 0.1) = (0.1)^3$$

$$\begin{aligned} P(\bar{S}_2) &= (0.05)^2 \\ P(S_2) &= (1 - (0.05)^2) \end{aligned} \rightarrow (1 - 0.95)(1 - 0.95) = (0.05)^2$$

$$P(\bar{S}_3) = 0.99$$

$$\begin{aligned} P(A_1) &= 0.72 \\ P(A_2) &= 0.73 \end{aligned} \quad \text{OR}$$

$$\begin{aligned} P(B_1) &= 0.88 \\ P(B_2) &= 0.81 \end{aligned} \quad \text{AND}$$

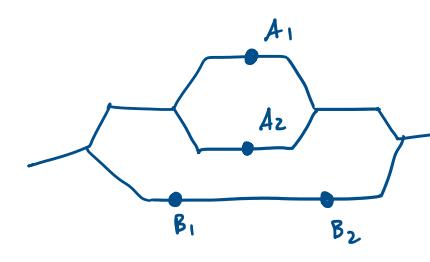
let  $A, B$  be these systems work

$$P(\text{system works}) = P(A \cup B)$$

$$= P(A) + P(B) - P(A)P(B)$$

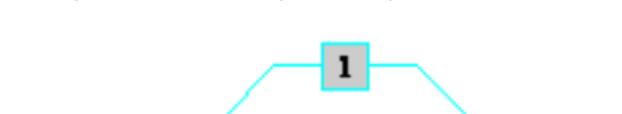
$$= P(A_1 \cup A_2) + P(B_1 \cup B_2) - P(A_1)P(A_2) - P(B_1)P(B_2)$$

$$\begin{aligned} &= 0.72 + 0.73 - 0.72(0.73) + 0.88(0.81) - P(A)P(B) \\ &= 0.9244 + 0.7128 - 0.9244(0.7128) \\ &= 0.978288 \quad \checkmark \end{aligned}$$



CM 2.1

**Problem #1:** Consider the system of components connected as in the figure below. Components 1 and 2 are connected in parallel, so that subsystem works if and only if either 1 or 2 works. Since 3 and 4 are connected in series, that subsystem works if and only if both 3 and 4 work. Suppose that  $P(1 \text{ works}) = 0.72$ ,  $P(2 \text{ works}) = 0.73$ ,  $P(3 \text{ works}) = 0.88$ , and  $P(4 \text{ works}) = 0.81$ . Find the probability that the system works.



# Lec6 - Probability Distributions, mean, and variance of a discrete random variable

Thursday, September 22, 2022 3:20 PM

Yuxi Qin  
yuxiqin.ca

## RANDOM VARIABLES

A **random variable** is a function from the sample space to the real numbers (random experiment)

- Ex. Tossing a coin 6 times - we have a random variable representing the number of heads observed
- $P(X=4)$ ?  $\rightarrow X$  is called a **discrete random variable**

## DISCRETE PROBABILITY DISTRIBUTION

If a discrete random variable  $X$  can take values  $x_1, x_2, \dots, x_n$  with probabilities  $p_1, p_2, \dots, p_n$  such that

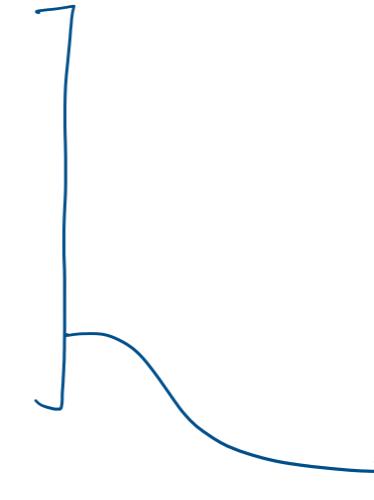
$$p_1 + p_2 + \dots + p_n = 1, \text{ and}$$

$$p_i \geq 0 \text{ for all } i,$$

- Then this defines a **discrete probability distribution for  $X$**

For example - consider an experiment where we toss 2 fair coins

- Let  $X$ , a discrete random variable, represent the number of tails
    - $X$  can take values 0, 1, or 2 with probabilities:
- |                 |
|-----------------|
| $P(X=0) = 0.25$ |
| $P(X=1) = 0.5$  |
| $P(X=2) = 0.25$ |
- These probabilities sum to 1, describing a discrete probability distribution for  $X$



## DEFINITIONS

**Sample Space:** set of all possible outcomes of an experiment, usually denoted by

$$S = \{s_1, s_2, \dots\}$$

Ex.  $S = \{\text{HH}, \text{HT}, \text{TH}, \text{TT}\}$

**Random Variable:** function from the sample space to the real numbers

$$\text{Ex. } X(\text{HH}) = 1$$

**Support** of a discrete random variable is define **informally** as the "set of all possible values of the random variable  $X$ ", usually denoted by  $S_x$

$$\text{Ex. } S_x = \{0, 1, 2\} \text{ or } x = 0, 1, 2$$

## CUMULATIVE DISTRIBUTIONS

**Probability mass function (pmf)** of a discrete random variable is defined for every number  $x$  by:

$$p(x) = P(X=x)$$

For a discrete random variable  $X$ , which can take values  $x_1, x_2, \dots, x_n$ , a pmf is a function such that:

$$\begin{aligned} f(x_i) &\geq 0 \\ \sum_{i=1}^n f(x_i) &= 1 \\ f(x_i) &= P(X=x_i) \end{aligned}$$

The **cumulative distribution function (cmf)**:

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

## DEFINITIONS

The **Expected Value** of a discrete random variable  $X$  is given by

$$\mathbb{E}[X] = \sum_x x P(X=x).$$

**Variance** of a random variable  $X$  is given by

$$\text{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

**Standard Variation** of a random variable  $X$  is given by

$$\text{SD}[X] = \sqrt{\text{Var}[X]}, \text{ i.e., } \sigma = \sqrt{\sigma^2}$$

## CM 2.4

**Problem #4:** An assembly consists of two mechanical components. Suppose that the probabilities that the first and second components meet specifications are 0.90 and 0.81, respectively. Let  $X$  be the number of components in the assembly that meet specifications.

- (a) Find the mean of  $X$ .
- (b) Find the variance of  $X$ .

only 2 components

$$P(X=2) = p(1) \cdot p(2) = 0.90 \cdot 0.81 = 0.729$$

$$\begin{aligned} X &= 0, 1, 2 \\ a) \quad P(X=0) &= \overline{p(1)} \cdot \overline{p(2)} = 0.10 \cdot 0.19 = 0.019 \\ P(X=1) &= \overline{p(1)} \cdot p(2) + p(1) \cdot \overline{p(2)} \\ &= 0.10 \cdot 0.81 + 0.90 \cdot 0.19 \\ &= 0.272 \end{aligned}$$

$$\begin{aligned} \mathbb{E}[X] &= \sum_x x P(X=x) \\ &= 0(0.019) + 1(0.272) + 2(0.729) \\ &= 1.71 \end{aligned}$$

$$\begin{aligned} b) \quad \sigma^2 &= \sum_x (x - \mu)^2 P(X=x) \\ &= (0 - 1.71)^2 (0.019) + (1 - 1.71)^2 (0.272) + (2 - 1.71)^2 (0.729) \\ &= (2.9241)(0.019) + 0.5041(0.272) + 0.0841(0.729) \\ &= 0.0555579 + 0.1270332 + 0.0617089 = 0.24389789 \end{aligned}$$

1. Two socks are selected at random and removed in succession from a drawer containing five brown socks and three green socks. Listed below are the elements of the sample space, the corresponding probabilities, the corresponding values of  $X$ , a discrete random variable representing the number of brown socks. What is the  $P(X=1)$  and the  $P(X \geq 1)$ ?

Elements of $S$	$P(X=X)$	$x$
BB	$\frac{5}{8} \cdot \frac{4}{7} = \frac{20}{56}$	2
BG	$\frac{5}{8} \cdot \frac{3}{7} = \frac{15}{56}$	1
GB	$\frac{3}{8} \cdot \frac{5}{7} = \frac{15}{56}$	1
GG	$\frac{3}{8} \cdot \frac{2}{7} = \frac{6}{56}$	0

$$\begin{aligned} P(X=1) &= \frac{15}{56} + \frac{15}{56} = \frac{30}{56} \\ P(X \geq 1) &= \frac{15}{56} + \frac{15}{56} + \frac{20}{56} = \frac{50}{56} \\ &= 1 - P(X=0) = 1 - \frac{6}{56} \end{aligned}$$

1	2	$x = \# \text{ of tails}$
T	T	
T	H	
H	H	
H	T	

## EXAMPLE 3.7 | Digital Channel

In Example 3.3, 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  $\{0, 1, 2, 3, 4\}$ . Based on a model for the errors presented in the following section, probabilities for these values will be determined. Suppose that the probabilities are

$$\begin{aligned} P(X=0) &= 0.6561 \quad P(X=2) = 0.0486 \quad P(X=4) = 0.0001 \\ P(X=1) &= 0.2916 \quad P(X=3) = 0.0036 \end{aligned}$$

Now

$$\begin{aligned} \mu &= E(X) = 0f(0) + 1f(1) + 2f(2) + 3f(3) + 4f(4) \\ &= 0(0.6561) + 1(0.2916) + 2(0.0486) + 3(0.0036) \\ &\quad + 4(0.0001) \\ &= 0.4 \end{aligned}$$

Although  $X$  never assumes the value 0.4, the weighted average of the possible values is 0.4.

To calculate  $V(X)$ , a table is convenient.

$X = \# \text{ of error bits in a 4 bit transmission}$

$X$	$f(x)$	$E(x) ? = \mu \longrightarrow E(x) = \text{expected value of } x$
0	0.6561	$E(x) = \sum_x x f(x)$
1	0.2916	$= \sum_x x f(x)$
2	0.0486	$= 0(0.6561) + 1(0.2916) + 2(0.0486) + 3(0.0036) + 4(0.0001)$
3	0.0036	
4	0.0001	

$$E(x) = 0.4$$

$\sigma^2 ? \longrightarrow \text{variance}$

$$\sigma^2 = \mathbb{E}[(X - \mu)^2]$$

$$= \sum_x (x - \mu)^2 f(x)$$

$$= \sum_x (x - \mu)^2 f(x)$$

all the same thing?

$$\begin{aligned} \sigma^2 &= (0 - 0.4)^2 (0.6561) + (1 - 0.4)^2 (0.2916) + (2 - 0.4)^2 (0.0486) + (3 - 0.4)^2 (0.0036) + (4 - 0.4)^2 (0.0001) \\ &= 0.105 + 0.105 + 0.1244 + 0.0243 + 0.0013 \end{aligned}$$

$$\sigma^2 = 0.36$$



The range of a r.u.  $x$  is  $[0, 1, 2, 3, x]$ , where  $x$  is unknown

If each value of  $X$  is **EQUALLY LIKELY**, and the mean of  $X$  is 6, determine  $x$ .

$$\mu = E(x) = 6$$

$$\text{Prob of } x_i \text{ is } \frac{1}{8} = 0.125$$

$$6 = 0(0.125) + 1(0.125) + 2(0.125) + 3(0.125) + x(0.125)$$

$$6 = 1.2 + 0.2x$$

$$x = 24$$

# Lec7 - Binomial and Geometric Distribution

Thursday, September 22, 2022 7:52 PM

## BERNOULLI TRIALS

A Bernoulli trial is an experiment with two possible outcomes, success and failure

If  $X \sim \text{Bernoulli}(p)$ , then

- $P(\text{success}) = P(X=1) = p$ , and
- $P(\text{failure}) = P(X=0) = 1-p$ .

- Ex. Tossing a coin and looking for a head is a Bernoulli trial - getting head is success and getting tail is a failure
- Rolling a dice and looking for a 6 is also a Bernoulli trial (boolean)

## DISTRIBUTION TERMINOLOGY

**Probability Density Function:** function used to calculate probabilities and to specify the probability distribution of a continuous random variable

**Probability Distribution:** for a sample space, a description of the set of possible outcomes along with a method to determine probabilities. For a random variable, it is a description of the range along with a method to determine probabilities.

**Probability Mass Function:** A function that provides probabilities for the values in the range of a discrete random variable.

## THE BINOMIAL DISTRIBUTION

**Binomial Probability Distribution** comes from a result of  $n$  independent Bernoulli trials, each trial having success probability  $p$

The probability of obtaining  $x$  successes in these  $n$  trials is given by

$$P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Looking at this formula from an intuitive viewpoint, it makes sense:

there are  $\binom{n}{x}$  ways of getting  $x$  successes from  $n$  trials,  
 $p^x$  is the probability of success,  $x$  times and  
 $(1-p)^{n-x}$  is the probability of failure,  $n-x$  times.

## BINOMIAL DISTRIBUTION PARAMETER

The probability of success  $p$  is called a parameter, because it could be any possible value in  $[0, 1]$  and the given function would still be a valid **probability mass function (pmf)**

## BINOMIAL DISTRIBUTION, MEAN, AND VARIANCE

If  $X$  is a binomial random variable with parameters  $p$  and  $n$ ,

$$\mu = E(X) = np$$

$$\sigma^2 = V(X) = np(1-p)$$

## GEOMETRIC DISTRIBUTION

In a series of Bernoulli trials, the random variable  $X$  that equals the number of trials until the first success is a geometric random variable with parameters  $0 < p < 1$  and

$$f(x) = (1-p)^{x-1} p$$

- For  $x = 1, 2, \dots$

## THE GEOMETRIC DISTRIBUTION, MEAN, AND VARIANCE

If  $X$  is a geometric random variable with parameter  $p$ ,

$$\mu = E(X) = 1/p$$

$$\sigma^2 = V(X) = (1-p)/p^2$$

Examples:

- ① A fair die is rolled 10 times.

- ① What is the probability that it shows the number 6 exactly 5 times?
- ② What is the probability that it does not show the number 1 at all?
- ③ What is the probability that it shows the number 4 less than 2 times?

- ② A couple decides to keep having children until they have a boy, then they stop.

- ① What is the probability that the couple have 5 children? (Geometric Distribution)
- ② Eight couples take this approach. What is the probability that more than 2 of these couples have 5 children?

## CM 2.5

**Problem #5:** A manufacturing process has 53 customer orders to fill. Each order requires one component part that is purchased from a supplier. However, typically 5% of the components are identified as defective, and the components can be assumed to be independent.

- (a) If the manufacturer stocks 56 components, what is the probability that the 53 orders can be filled without reordering components?
- (b) Let  $X$  be the number of good (i.e., non-defective) components among the 56 in stock. Find the mean of  $X$ .
- (c) Find the variance of  $X$  [from part (b)].

### Binomial Distribution

$$P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$\begin{aligned} a) n &= 56 && \text{chances of failure} \\ k &= 0, 1, 2, 3 && P(X \leq 3) = P(X=0) + P(X=1) + P(X=2) + P(X=3) \\ P(X=0) &= 0.05 && = \binom{56}{0} (0.05)^0 (0.95)^{56} + \binom{56}{1} (0.05)^1 (0.95)^{55} + \binom{56}{2} (0.05)^2 (0.95)^{54} + \binom{56}{3} (0.05)^3 (0.95)^{53} \\ &= 0.056561627 + 0.166703954 + 0.241287828 + 0.228588467 \\ &= 0.693145878 \end{aligned}$$

$$b) \sigma^2 = \# \text{ of ND among 56 in stock} - \text{find } \mu$$

$$X = ND = 1 - 0.05 = 0.95$$

$$\mu = np = 56(0.95) = 53.2$$

$$c) \sigma^2 = np(1-p) = \sqrt{np(1-p)} = \sqrt{53.2(0.05)} = 2.66$$

## CM 2.7

**Problem #7:** A geologist has collected 21 specimens of basaltic rock and 15 specimens of granite. The geologist instructs a laboratory assistant to randomly select 8 of the specimens for analysis.

- (a) What is the probability that at least 2 of the selected specimens are granite?
- (b) What is the expected number of granite specimens in the sample?
- (c) If this same process is repeated every day, how many days (on average) will it take before getting a sample consisting entirely of granite?

$$a) P(z \geq 8) = 1 - P(z \leq 7)$$

$$= 1 - \left[ P(0B) + P(1B) + P(2B) + P(3B) \right]$$

$$= 1 - \left[ \binom{21}{0} \left(\frac{15}{36}\right)^0 \left(\frac{21}{36}\right)^{21} + \binom{21}{1} \left(\frac{15}{36}\right)^1 \left(\frac{21}{36}\right)^{20} + \binom{21}{2} \left(\frac{15}{36}\right)^2 \left(\frac{21}{36}\right)^{19} + \binom{21}{3} \left(\frac{15}{36}\right)^3 \left(\frac{21}{36}\right)^{18} \right]$$

$$= 1 - 0.047364445$$

$$= 0.935635554$$

$$b) \mu = \frac{ns}{N} = \frac{8(15)}{36} = \frac{10}{3}$$

$$= \frac{1}{3}$$

$$c) P(8B) = \frac{\binom{21}{8} \left(\frac{15}{36}\right)^8 \left(\frac{21}{36}\right)^{13}}{\binom{36}{8}}$$

$$= \frac{61132}{61132} = 1$$

$$E[X] = \frac{1}{p} = \frac{61132}{13} = 4702.461538$$

Yuxi Qin

[yuxiqin.ca](http://yuxiqin.ca): A fair coin is tossed six times. What is the probability of obtaining exactly four heads?

Here,  $X$  is a random variable representing the number of heads obtained,  $n = 6$  is the number of trials,  $x = 4$  is the number of successes, and  $p = 1/2$  is the probability of observing a success.

$$\begin{aligned} P(X=x) &= \binom{n}{x} p^x (1-p)^{n-x} \\ P(X=4) &= \binom{6}{4} \left(\frac{1}{2}\right)^4 \left(1-\frac{1}{2}\right)^{6-4} \\ &= 0.234375 \end{aligned}$$

2. A fair coin is tossed five times. What is the probability of obtaining at least four heads?

Here,  $X$  is a random variable representing the number of heads obtained,  $n = 5$  is the number of trials,  $x = 4$  and  $x = 5$  are the number of successes, and  $p = 1/2$  is the probability of observing a success.

$$\begin{aligned} P(X=x) &= \binom{n}{x} p^x (1-p)^{n-x} \\ P(X \geq 4) &= P(X=4) + P(X=5) \\ P(X \geq 4) &= \binom{5}{4} \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^1 + \binom{5}{5} \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right)^0 \\ &= 0.15625 + 0.03125 \\ &= 0.1875 \end{aligned}$$

Ex. 2 possible outcomes, success or failure.

- Success: bit received in error
- Constant probability of success = 0.1
- N independent trials - fixed number of trials

$N = 4$ ; number of bits transmitted

$P = 0.1$ ; probability a bit will be received in error

What is the expected value and the variance?

$$\begin{aligned} E(X) &= np = 4(0.1) = 0.4 \\ \sigma^2 &= np(1-p) \\ &= 4(0.1)(0.9) \\ &= 0.36 \end{aligned}$$

Transmission of bits:

$P = 0.1$ ; probability of error

What is the mean number of transmissions until the 1st error?

$$M = E(X) = \frac{1}{p} = \frac{1}{0.1} \rightarrow 10$$

What is the standard deviation of the number of transmissions until the 1st error?

$$\sigma^2 = \frac{(1-p)}{p^2} = \frac{(1-0.1)}{(0.1)^2} = \sqrt{\frac{0.9}{(0.1)^2}} = 9.49$$

/ - Binomial: probability of getting  $x$  successes in  $n$  trials.

$$P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}$$

a) Given:  $p = 1/6$ ;  $n = 10$ ;  $x = 5$ ;

$$P(X=5) = \binom{10}{5} \left(\frac{1}{6}\right)^5 \left(1-\frac{1}{6}\right)^{10-5}$$

$$\approx 0.013$$

b)  $P(X=0) = P(X=1) = p = 1/6$ ;  $n = 10$

$$P(X<2) = \binom{10}{0} \left(\frac{1}{6}\right)^0 \left(1-\frac{1}{6}\right)^{10-0} + \binom{10}{1} \left(\frac{1}{6}\right)^1 \left(1-\frac{1}{6}\right)^{10-1}$$

$$\approx 0.485$$

c)  $P(X \geq 2) = 1 - P(X<2) = 1 - 0.485 = 0.515$

2. Geometric:  $x = \text{number of trials until 1st success}$

$$f(x) = (1-p)^{x-1} p$$

a)  $P(\text{boy}) = 0.5$

$$P(\text{1st success on 5th try}) = (1-0.5)^4 (0.5) = 0.03125$$

b) 1st success on 5th try: 0.03125

- we have binomial type Q:

$$\text{Binomial: } P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$P(X \geq 2) = 1 - P(X \leq 1)$$

$$= 1 - \left[ P(X=0) + P(X=1) \right]$$

Binomial dist.: prob of obtaining  $x$  successes

in  $n$  Bernoulli trials.

$$\begin{aligned} &= 1 - \left[ \binom{8}{0} (0.03125)^0 (1-0.03125)^{8-0} \right. \\ &\quad \left. + \binom{8}{1} (0.03125)^1 (1-0.03125)^{8-1} \right. \\ &\quad \left. + \binom{8}{2} (0.03125)^2 (1-0.03125)^{8-2} \right] \\ &= 1 - 0.98477 \\ &\approx 0.0152 \end{aligned}$$

## CM 2.7

**Problem #7:** A geologist has collected 21 specimens of basaltic rock and 15 specimens of granite. The geologist instructs a laboratory assistant to randomly select 8 of the specimens for analysis.

- (a) What is the probability that at least 2 of the selected specimens are granite?
- (b) What is the expected number of granite specimens in the sample?
- (c) If this same process is repeated every day, how many days (on average) will it take before getting a sample consisting entirely of granite?

$$a) P(z \geq 8) = 1 - P(z \leq 7)$$

$$= 1 - \left[ P(0B) + P(1B) + P(2B) + P(3B) \right]$$

$$= 1 - \left[ \binom{21}{0} \left(\frac{15}{36}\right)^0 \left(\frac{21}{36}\right)^{21} + \binom{21}{1} \left(\frac{15}{36}\right)^1 \left(\frac{21}{36}\right)^{20} + \binom{21}{2} \left(\frac{15}{36}\right)^2 \left(\frac{21}{36}\right)^{19} + \binom{21}{3} \left(\frac{15}{36}\right)^3 \left(\frac{21}{36}\right)^{18} \right]$$

$$= 1 - 0.047364445$$

# Lec8 - Negative Binomial Dist. And Hypergeometric Dist.

Thursday, September 22, 2022 8:47 PM

Yuxi Qin  
[yuxiqin.ca](http://yuxiqin.ca)

## NEGATIVE BINOMIAL DISTRIBUTION

In a series of Bernoulli trials (independent trials with constant probability  $p$  of success), the random variable  $X$  that equals the number of trials until  $r$  successes occur is a negative binomial random variable, with parameters  $0 < p < 1$  and  $r = 1, 2, 3, \dots$ . And

$$P(X=x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}$$

$$x = r, r+1, r+2$$

(1)

## SAMPLING WITH REPLACEMENT

- Sampling such that each unit is replaced before the next sample is drawn is called **sampling with replacement**
- The binomial distribution will be applicable in cases where we sample with replacement
- We have already seen some examples, such as tossing a coin or rolling dice

What if we don't sample with replacement, such as in a lottery or a raffle?

## SAMPLING WITHOUT REPLACEMENT

- Sampling such that each unit is not replaced before the next sample is drawn is called **sampling without replacement**
- In this case, the binomial distribution will not be applicable

Ex. A lottery involves drawing 6 numbers from 42, without replacement. Prizes are awarded to players who match 4, 5, or 6 of their numbers to the ones drawn. What is the probability of matching exactly 4 numbers?

- To answer this, we'll need hypergeometric distribution.

## THE HYPERGEOMETRIC DISTRIBUTION

- Broadly speaking, consider a situation where we're looking for  $x$  defects in units
- Suppose that there are  $N$  units in total and  $M$  of these are defective, such that  $M \leq N$
- If we sample  $n$  of these units then the probability that  $x$  are defective is given by:

$$P(X=x) = \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}}, \quad (2)$$

- Where  $x$ , an integer, satisfies  $\max(0, n - N + M) \leq x \leq \min(n, M)$

- So, the numerator of (2) is the particular outcome that we 'want' and the denominator is the total number of possible outcomes

$$E(X) = n \left( \frac{S}{N} \right) \quad \begin{array}{l} S \rightarrow \# \text{ of successes available} \\ N \rightarrow \text{population size} \end{array}$$

## OTHER SUGGESTED EXAMPLES:

- 800 men are studies. Suppose 30% carry a gene that indicates an increased risk of high blood pressure.

- 10 men selected, what is the probability that exactly 1 man has the marker?

$$\begin{aligned} X &= \# \text{ of men carrying marker} & 800(0.30) &= 240 \\ N &= 800; M = 240; n = 10 & P(X=1) &= \frac{\binom{240}{1} \binom{800-240}{10-1}}{\binom{800}{10}} = 0.1201 \end{aligned}$$

- 10 men selected, what is the probability that more than 1 has the marker?

$$\begin{aligned} P(X>1) &= 1 - P(X \leq 1) \\ &= 1 - [P(X=0) + P(X=1)] \\ \text{need } P(X=0) &\\ P(X=0) &= \frac{\binom{240}{0} \binom{800-240}{10-0}}{\binom{800}{10}} = 0.0276 \\ \text{back to solve} &\\ &= 1 - (0.0276 + 0.1201) \\ &= 0.8523 \end{aligned}$$

CM 2.6

**Problem #6:** The probability that a randomly selected box of a certain type of cereal has a particular prize is 0.17. Suppose that you purchase box after box until you have obtained 3 of these prizes.

- What is the probability that you purchase exactly 8 boxes?
- What is the probability that you purchase at least 12 boxes?
- How many boxes would you expect to purchase, on average?

**Negative Binomial Distribution**

$$P(X=x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}$$

$$p = 0.17$$

$$r = 3$$

$$p = 0.17, 1-p = 0.83$$

$$X = 8, x = 5$$

$$r = 3$$

$$a) P(X=8) = \binom{8-1}{3-1} (0.17)^3 (0.83)^5$$

$$= 0.040640264 \quad \checkmark$$

$$b) P(X \geq 12) = 1 - P(X \leq 11)$$

$$r = 3$$

$$r =$$

# Lec9 - Continuous Random Variables, and Probability Density Functions

Monday, September 26, 2022 1:30 PM

Yuxi Qin  
[yuxiqin.ca](http://yuxiqin.ca)

## POISSON DISTRIBUTION

Let  $X$  be a random variable that follows a Poisson distribution with mean  $\mu$ . Then:

$$P(X=x) = \frac{e^{-\mu} \cdot \mu^x}{x!}, \quad x = 0, 1, 2, \dots$$

- The poisson distribution is very powerful, both in its own right and as an approximation to the binomial distribution
- The poisson distribution to the binomial distribution is achieved by taking  $u=np$  and is very useful given the computational cost of computing  $nCx$   
for  $n \rightarrow \infty, p \rightarrow 0 \Rightarrow np=1$  remains constant

A random variable is a function from the sample space to the real numbers (random experiment)

- Ex. Tossing a coin 6 times - we have a random variable representing the number of heads observed
- $P(X=4) \rightarrow X$  is called a **discrete random variable**

A random variable  $X$  is said to be **continuous** if its distribution function takes the form

$$P(X \leq x) = F(x) = \int_{-\infty}^x f(t) dt.$$

We call  $F(x)$  the **cumulative distribution function** of  $X$ , and  $f(x)$  the **probability density function** of  $X$ , which satisfies

- $f(x) \geq 0$  for all  $x$ , and

$$\int_{-\infty}^{\infty} f(x) dx = 1.$$

## PROBABILITIES

- For a continuous random variable,  $X$ , the following hold:

- $P(X=x) = 0$  for all  $x$ ,
- $P(X \leq x) = P(X < x)$  for all  $x$ ,
- $P(a \leq X \leq b) = F(b) - F(a) = \int_a^b f(x) dx$ ,

- Where  $a$  and  $b$  are constants

### EXAMPLE 4.1 | Electric Current

Let the continuous random variable  $X$  denote the current measured in a thin copper wire in milliamperes. Assume that the range of  $X$  is  $[4.9, 5.1]$  mA, and assume that the probability density function of  $X$  is  $f(x) = 5$  for  $4.9 \leq x \leq 5.1$ . What is the probability that a current measurement is less than 5 milliamperes?

The probability density function is shown in Figure 4.4. It is assumed that  $f(x) = 0$  wherever it is not specifically

defined. The shaded area in Figure 4.4 indicates the probability.

$$P(X < 5) = \int_{4.9}^5 f(x) dx = \int_{4.9}^5 5 dx = 0.5$$

As another example,

$$P(4.95 < X < 5.1) = \int_{4.95}^{5.1} f(x) dx = 0.75$$

What proportion of parts is between 12.5 and 12.6 millimeters? Now,

$$P(12.5 < X < 12.6) = \int_{12.5}^{12.6} f(x) dx = -e^{-20(x-12.5)} \Big|_{12.5}^{12.6} = 0.865$$

Because the total area under  $f(x)$  equals 1, we can also calculate  $P(12.5 < X < 12.6) = 1 - P(X > 12.6) = 1 - 0.135 = 0.865$ .

*Practical Interpretation:* Because 0.135 is the proportion of parts with diameters greater than 12.60 mm, a large proportion of parts is scrapped. Process improvements are needed to increase the proportion of parts with dimensions near 12.50 mm.

### EXAMPLE 4.2 | Hole Diameter

Let the continuous random variable  $X$  denote the diameter of a hole drilled in a sheet metal component. The target diameter is 12.5 millimeters. Most random disturbances to the process result in larger diameters. Historical data show that the distribution of  $X$  can be modeled by a probability density function  $f(x) = 20e^{-20(x-12.5)}$ , for  $x \geq 12.5$ .

If a part with a diameter greater than 12.60 mm is scrapped, what proportion of parts is scrapped? The density function and the requested probability are shown in Figure 4.5. A part is scrapped if  $X > 12.60$ . Now,

$$P(X > 12.60) = \int_{12.6}^{\infty} f(x) dx = \int_{12.6}^{\infty} 20e^{-20(x-12.5)} dx = -e^{-20(x-12.5)} \Big|_{12.6}^{\infty} = 0.135$$

### CM 2.8

**Problem #8:** Data from the Central Hudson Laboratory determined that the mean number of insect fragments in 225-gram chocolate bars was 14.4 (<http://www.centralhudsonlab.com/chocolates.shtml>). In a 28-gram bar the mean number of insect fragments would then be 1.79. Assume that the number of insect fragments follows a Poisson distribution.

- If you eat a 28-gram chocolate bar, find the probability that you will have eaten at least 3 insect fragments.
- If you eat a 28-gram chocolate bar every week for 14 weeks, find the probability that you will have eaten no insect fragments in exactly 3 of those weeks.

$$\text{Poisson Distribution: } P(X=x) = \frac{e^{-\lambda} \cdot \lambda^x}{x!}$$

a) at least :

$$\begin{aligned} P(X \geq 3) &= 1 - P(X \leq 2) \\ &= 1 - [P(X=0) + P(X=1) + P(X=2)] \\ &= 1 - \left[ \frac{e^{-1.79} (1.79)^0}{0!} + \frac{e^{-1.79} (1.79)^1}{1!} + \frac{e^{-1.79} (1.79)^2}{2!} \right] \\ &= 1 - (0.166960169 + 0.298858703 + 0.267474539) \\ &= 1 - 0.733297411 \\ &= 0.266702588 \end{aligned}$$

### CM 2.9

**Problem #9:** Let  $X$  denote the vibratory stress (psi) on a wind turbine blade at a particular wind speed in a wind tunnel.

Suppose that  $X$  has the following Rayleigh pdf.

$$f(x) = \begin{cases} (x/\theta^2) e^{-x^2/(2\theta^2)} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

- If  $\theta = 100$ , find the probability that the vibratory stress is between 97 and 235.

- If  $\theta = 100$ , then 72% of the time the vibratory stress is greater than what value?

Probability Distribution Function (PDF)

↳ Probabilities  $P(a \leq X \leq b) = F(b) - F(a)$

$$= \int_a^b f(x) dx$$

$$u = \frac{x^2}{20000}, \quad du = \frac{x}{10000} dx$$

$$= - \int e^u du$$

$$= -e^u + C$$

$$\text{Sub in } u$$

$$= -e^{\frac{x^2}{20000}} \Big|_{97}^{235}$$

$$= 0.561508377 \quad \checkmark$$

$$\begin{aligned} a) \quad P(97 \leq X \leq 235) &= \int_{97}^{235} \left( \frac{x}{10000} \right) e^{-\frac{x^2}{20000}} dx \\ &= \frac{1}{10000} \int_{97}^{235} x e^{-\frac{x^2}{20000}} dx \\ &= -\frac{x^2}{20000} \Big|_{97}^{235} \\ &= -\int e^u du \\ &= -e^u \Big|_{97}^{235} \\ &= -e^{-\frac{97^2}{20000}} + e^{-\frac{235^2}{20000}} \\ &= 0.561508377 \quad \checkmark \end{aligned}$$

$$\begin{aligned} b) \quad 72\% \text{ of the time } \geq a \\ &\Rightarrow 1 - 0.72 \leq a \\ &\Rightarrow \text{So we can start @ } a > 0 \\ &1 - 0.72 = \int_0^a \left( \frac{x}{10000} \right) e^{-\frac{x^2}{20000}} dx \\ &= \frac{1}{10000} \int_0^a x e^{-\frac{x^2}{20000}} dx \\ &= -\frac{x^2}{20000} \Big|_0^a \\ &= -\int e^u du \\ &= -e^u \Big|_0^a \\ &= -e^{-\frac{a^2}{20000}} \Big|_0^a \\ &= 1 - e^{-\frac{a^2}{20000}} \quad \checkmark \end{aligned}$$

$$SRT \quad BS \quad a = \pm 100 \sqrt{2 \ln(\frac{25}{18})} \quad \text{take positive } (a > 0) \quad \text{round to 2 dec.}$$

$$= \pm 81.056 \quad = 81.056 \quad \checkmark$$

- Chocolates are manufactured so that the number of surface blemishes on any one chocolate is described by a Poisson distribution with mean  $\mu = 0.05$  blemishes per chocolate bar. What is the probability a chocolate bar contains two or more surface blemishes?

$$\begin{aligned} P(X=x) &= \frac{e^{-0.05} 0.05^x}{x!} \quad \text{and } P(X \geq 2) = 1 - P(X < 2) \\ &= 1 - [P(X=0) + P(X=1)] \\ &= 1 - \left[ \frac{e^{-0.05} 0.05^0}{0!} + \frac{e^{-0.05} 0.05^1}{1!} \right] \\ &= 1 - 1.05e^{-0.05} = 1 - 0.9988 = 0.0012 \end{aligned}$$

- Coliform bacteria are randomly distributed in river water at an average concentration of 1 per 25cc of water. What is the probability of finding more than two bacteria in a sample of 10cc of river water?

$$\mu = 0.04$$

$$\begin{aligned} P(X \geq 2) &= 1 - (P(X=0) + P(X=1) + P(X=2)) \\ &= 1 - \frac{e^{-0.04} (0.04)^0}{0!} - \frac{e^{-0.04} (0.04)^1}{1!} - \frac{e^{-0.04} (0.04)^2}{2!} \\ &= 1 - 0.67032 - 0.268182 - 0.0536256 \approx 0.0079 \end{aligned}$$

- Each morning, after opening her e-mail account, Lucy has to discard, on average, ten spam messages. If the number of spam messages may be described by a Poisson distribution, what is the probability that on any given morning Lucy will receive less than four spam messages?

$$\mu = 10$$

$$\begin{aligned} P(X \geq 4) &= P(X=0) + P(X=1) + P(X=2) + P(X=3) \\ &= \frac{e^{10} (10)^0}{0!} + \frac{e^{10} (10)^1}{1!} + \frac{e^{10} (10)^2}{2!} + \frac{e^{10} (10)^3}{3!} \\ &= 0.000454 + 0.00454 + 0.0227 + 0.0078 \\ &= 0.0103 \end{aligned}$$

- The pdf of a continuous random variable  $X$  is given by

$$f(x) = \begin{cases} cx & \text{if } 0 \leq x \leq 1 \\ 0 & \text{elsewhere} \end{cases}$$

where  $c$  is constant. Find  $c$  and  $P(0.2 < X < 0.5)$ .

Now,  $f(x)$  is a pdf and so

$$\begin{aligned} \int_{-\infty}^{\infty} f(x) dx &= \int_0^1 cx dx = \left[ \frac{cx^2}{2} \right]_0^1 = \frac{c}{2} = 1 \Rightarrow c = 2 \\ P(0.2 < X < 0.5) &= \int_{0.2}^{0.5} f(x) dx = \int_{0.2}^{0.5} 2x dx = \left[ x^2 \right]_{0.2}^{0.5} \\ &= (0.5)^2 - (0.2)^2 \\ &= 0.21 \end{aligned}$$

- The pdf of a continuous random variable  $X$  is given by

$$f(x) = \begin{cases} 2x & \text{if } 0 \leq x \leq 1 \\ 0 & \text{elsewhere} \end{cases}$$

Find  $F(x)$ , the cdf of  $X$ , and  $P(0.2 < X < 0.5)$ .

$$F(x) = \int_{-\infty}^x f(t) dt = \int_0^x 2t dt = \left[ 2 \frac{t^2}{2} \right]_0^x = x^2$$

$$\begin{aligned} P(0.2 < X < 0.5) &= F(0.5) - F(0.2) = (0.5)^2 - (0.2)^2 \\ &= 0.21 \end{aligned}$$

# Lec10 - Cumulative Distribution Functions, Mean and Variance of a Continuous Random Variable

Tuesday, September 27, 2022 12:58 PM

## DEFINITIONS

A random variable  $X$  is said to be **continuous** if its distribution function takes the form

$$P(X \leq x) = F(x) = \int_{-\infty}^x f(t) dt.$$

- We call  $F(x)$  the **cumulative distribution function** of  $X$
- We call  $f(x)$  the **probability density function** of  $X$  and it satisfies:

$$\circ \quad f(x) \geq 0 \quad \text{for all } x,$$

$$\circ \quad \int_{-\infty}^{\infty} f(x) dx = 1.$$

## PROBABILITIES

- For a continuous random variable,  $X$ , the following hold:

$$\circ \quad P(X = x) = 0 \quad \text{for all } x,$$

$$\circ \quad P(X \leq x) = P(X < x) \quad \text{for all } x,$$

$$\circ \quad P(a \leq X \leq b) = F(b) - F(a) = \int_a^b f(x) dx,$$

- Where  $a$  and  $b$  are constants

## EXPECTED VALUES AND VARIANCE

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x) dx$$

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f(x) dx$$

- Where  $f(x)$  is the pdf of  $X$ , and

$$\text{Var}[X] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx,$$

where  $\mu = \mathbb{E}[X]$ .

$$\bullet \text{ Note that: } \text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

## EXAMPLES

Ex. Suppose that the cumulative distribution function of the r.u.  $X$  is:

$$F(x) = \begin{cases} 0 & x < -2 \\ 0.25x + 0.5 & -2 \leq x < 2 \\ 1 & x \geq 2 \end{cases}$$

Determine the following:

$$1. \quad P(X < 1.8) = F(1.8) = 0.25(1.8) + 0.5$$

$$= 0.95$$

$$2. \quad P(X > -1.5) = 1 - P(X < -1.5) = 1 - F(-1.5) \\ = 1 - [0.25(-1.5) + 0.5] \\ = 1 - 0.125 = 0.875$$

$$3. \quad P(X < -2) = 0$$

$$4. \quad P(-1 < X < 1) = F(1) - F(-1) \\ = [0.25(1) + 0.5 - (0.25(-1) + 0.5)] \\ = (0.75 - 0.25) \\ = 0.5$$

Ex. Suppose that the probability density function for the following cumulative distribution function is:

$$F(x) = \begin{cases} 0 & x < -2 \\ 0.25x + 0.5 & -2 \leq x < 1 \\ 0.5x + 0.25 & 1 \leq x < 1.5 \\ 1 & x \geq 1.5 \end{cases}$$

$$f(x) = \int_{-\infty}^x f(x) dx \Rightarrow f(x) = \frac{d}{dx} F(x)$$

$$f(x) = \begin{cases} 0.25 & -2 < x < 1 \\ 0.50 & 1 \leq x < 1.5 \end{cases}$$

## CM 3.1

**Problem #1:** Suppose that the random variable  $X$  has the following cumulative distribution function.

$$F(x) = \begin{cases} 0 & x < 2 \\ \frac{1}{32}(x^2 - 4) & 2 \leq x < 6 \\ 1 & x > 6 \end{cases}$$

$$f(x) = \frac{d}{dx} F(x) \\ = \frac{d}{dx} \left( \frac{1}{32}(x^2 - 4) \right) \\ = \frac{1}{32} \cdot \frac{d}{dx} (x^2 - 4) \\ = \frac{1}{16} (2x) \\ = \frac{x}{16}$$

- (a) Find  $P(X > 2.44)$   
 (b) Find  $P(0.99 < X < 2.22)$   
 (c) Find the mean of  $X$ .  
 (d) Find the variance of  $X$ .

Yuxi Qin  
yuxiqin.ca Again, consider a random variable  $X$  with pdf:

$$f(x) = \begin{cases} 2x & \text{if } 0 \leq x \leq 1 \\ 0 & \text{elsewhere} \end{cases}$$

Find  $E[X]$  and the  $\text{Var}[X]$

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x) dx = \int_0^1 x(2x) dx = \int_0^1 2x^2 dx = \left[ 2 \frac{x^3}{3} \right]_0^1 = \frac{2}{3}.$$

$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 f(x) dx = \int_0^1 x^2(2x) dx = \int_0^1 2x^3 dx = \left[ 2 \frac{x^4}{4} \right]_0^1 = \frac{1}{2}.$$

$$\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{1}{2} - \left( \frac{2}{3} \right)^2 = 0.0556.$$

EXAMPLE 1:

The pdf of a continuous random variable  $X$  is given by:

$$f(x) = \begin{cases} cx^2 & \text{if } 0 \leq x \leq 2 \\ 0 & \text{elsewhere} \end{cases}$$

Where  $c$  is constant.

- Find  $c$ .
- Find the cdf,  $F(x)$ .
- What is  $P(1 \leq X \leq 1.5)$ ?
- Find  $\mathbb{E}[X]$ .
- Find  $\text{Var}[X]$ .

a)  $\int_0^2 c \cdot x^2 dx = 1$   
 $c \int_0^2 x^2 dx = 1$   
 $c \left( \frac{x^3}{3} \right) \Big|_0^2 = 1$   
 $c \left( \frac{8}{3} - 0 \right) = 1 \Rightarrow c = \frac{3}{8}$

b)  $\int_0^x \frac{3}{8} t^2 dt$   
 $= \frac{3}{8} \frac{t^3}{3} \Big|_0^x = \frac{x^3}{8} - 0$   
 $= \frac{x^3}{8}$

c)  $P(1 \leq X \leq 1.5) ?$   
 $F(1.5) - F(1)$   
 $= \frac{1}{8} \left( \frac{3}{2} \right)^3 - \frac{1}{8} \left( \frac{1}{2} \right)^3$   
 $= \frac{1}{8} \left( \frac{27}{8} \right) - \frac{1}{8}$   
 $= \frac{27}{64} - \frac{8}{64} = \frac{19}{64}$

d)  $\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x) dx$   
 $\int_0^2 x \cdot \frac{3}{8} \cdot x^2 dx = \int_0^2 \frac{3}{8} x^3 dx$   
 $= \frac{3}{32} x^4 \Big|_0^2 = \frac{3}{25} (2^4) = \frac{3}{2}$

e) Find  $\text{Var}[X]$   
 $E[X^2] = \int_0^2 x^2 \cdot \frac{3}{8} \cdot x^2 dx$   
 $= \int_0^2 \frac{3}{8} x^4 dx = \frac{3}{40} x^5 \Big|_0^2$   
 $= \frac{3}{40} (2^5) = \frac{12}{5}$

$\text{Var}[X] = E[X^2] - \mathbb{E}[X]^2$   
 $= \frac{12}{5} - \left( \frac{3}{2} \right)^2 = \frac{3}{20}$

EXAMPLE 2:

The cdf of a continuous random variable  $X$  is given by

$$F(x) = \begin{cases} 1 - e^{-x} & \text{if } x > 0 \\ 0 & \text{elsewhere} \end{cases}$$

- What is  $P(X \leq 2.6)$ ?
- What is  $P(1 < X < 4)$ ?
- Find the pdf  $f(x)$ .

a)  $P(X \leq 2.6) = F(2.6)$   
 $= 1 - e^{-2.6}$   
 $\approx 0.926$

b)  $P(1 < X < 4) = F(4) - F(1)$   
 $= 1 - e^{-4} - (1 - e^{-1})$   
 $\approx 0.350$

c)  $f(x) = \frac{d}{dx} F(x) = \frac{d}{dx} (1 - e^{-x})$   
 $= e^{-x}$

CM 3.1

**Problem #1:** Suppose that the random variable  $X$  has the following cumulative distribution function.

$$a) P(X > 2.44) \quad c) \mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) dx \quad e) E(X^2) = \int_2^6 \frac{x^3}{16} dx$$

$$F(2.44) = 1 - \frac{1}{32}(2.44^2 - 4) \quad = \frac{x^4}{64} \Big|_2^6 \quad = \frac{x^4}{64} - \frac{2^4}{64}$$

$$= 0.93895 \quad \checkmark$$

$$b) P(0.99 < X < 2.22) \quad d) \mathbb{E}(X^2) = \int_2^6 \frac{x^2}{16} dx \quad = \frac{6^4}{64} - \frac{2^4}{64}$$

$$= F(2.22) - F(0.99) \quad = \frac{1280}{64} = 20$$

$$= \frac{1}{32}(2.22^2 - 4) - 0 \quad = \frac{6^3}{48} - \frac{2^3}{48} = \frac{208}{48} \quad \checkmark$$

$$= 0.0290125 \quad \checkmark$$

$$d) \text{Var}(X) = E(X^2) - \mathbb{E}(X)^2$$

$$= 20 - \left( \frac{208}{48} \right)^2$$

$$= \frac{11}{9} \quad \checkmark$$

## BACKGROUND

- The normal distribution is the most frequently used continuous probability distribution
- Many measurements can be well approximated by a normal distribution, and is characterized by a **bell shaped curve**
- Also referred to as the **Gaussian distribution**

## DENSITY FUNCTION

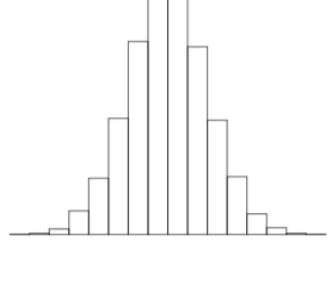
If a random variable  $X$  follows a normal distribution with parameters  $\mu$  and  $\sigma$ , then we write  $X \sim N(\mu, \sigma^2)$  and the pdf of  $X$  is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \quad -\infty < x < \infty.$$

- $\mu$  = mean of  $X$
- $\sigma$  = standard deviation of  $X$
- $\sigma^2$  = variance of  $X$

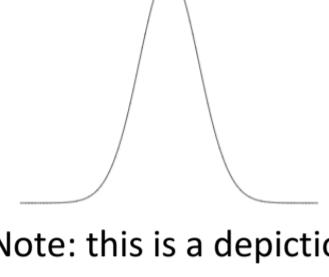
## HISTOGRAM

- The normal distribution is symmetric about the mean
- If we generate 10000 values from a normal distribution, we may get a histogram like the one below



## THE BELL SHAPED CURVE

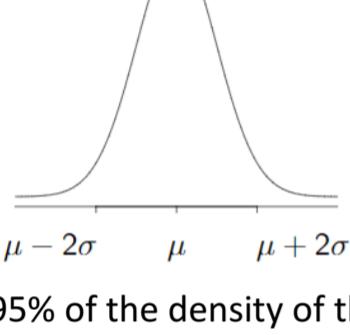
- These 10k values can also be represented like this:



- Note: this is a depiction of the normal distributions' pdf - "bell shaped" and "symmetric"

## NORMAL DISTRIBUTION PARAMETERS

- The normal distribution is usually characterized by its mean  $\mu$  and standard deviation  $\sigma$



- 95% of the density of the distribution lies within two (1.96 to be precise) standard deviations of the mean

## NORMALLY DISTRIBUTED DATA

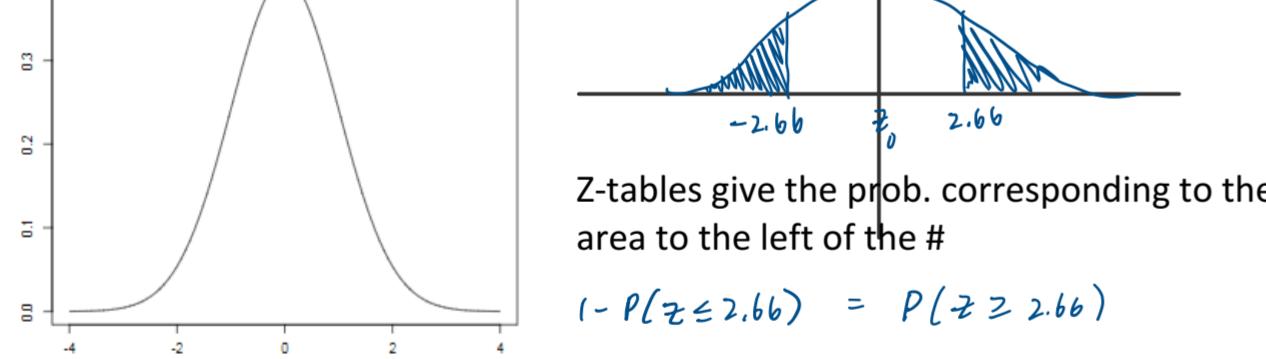
Consider the following information - we measure the heights of women in Ireland. We assume their heights are normally distributed with mean 1.62m and standard deviation 0.11m. It follows that:

- The median and mode of the height of women is 1.62m
- The same amount of women are 1.52m tall as are 1.72m tall
- The same amount of women are 1.60m tall as are 1.64m tall
- 95% of women are between 1.40m and 1.84m tall

Now, if we want to find the probability that a woman from this population will have a height above, below, or between a certain value(s), we can use the **standard normal distribution**

## STANDARD NORMAL DISTRIBUTION

The **standard normal distribution (SND)** is a normal distribution with mean 0 and standard deviation 1.



Z-tables give the prob. corresponding to the area to the left of the #

$$1 - P(z \leq 2.66) = P(z \geq 2.66)$$

- Again, 95% of the density of the distribution (the area under the curve) lies between  $\pm 1.96$  of the mean

## FROM NORMAL TO STANDARD NORMAL

- Let  $X \sim N(\mu, \sigma^2)$  and  $Z \sim N(0, 1)$ , then

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

- Therefore, to compute  $P(X \leq x)$  we can use the fact that

$$P(X \leq x) = P\left(Z \leq \frac{x - \mu}{\sigma}\right)$$

And then consult the standard normal tables

## CM 3.4

**Problem #4:** Suppose that 21% of all steel shafts produced by a certain process are nonconforming but can be reworked (rather than having to be scrapped).

- (a) In a random sample of 235 shafts, find the approximate probability that between 42 and 58 (inclusive) are nonconforming and can be reworked.

- (b) In a random sample of 235 shafts, find the approximate probability that at least 52 are nonconforming and can be reworked.

a)  $n=235$   
 $p=0.21$  for binomial random var:  $\mu=np$   
 $\sigma^2=np(1-p)$

$\mu=(235)(0.21)=49.35$   
 $\sigma=\sqrt{(235)(0.21)(0.79)}=6.24391704$

$P(42 \rightarrow 58 \text{ can be reworked})$

$P\left(\frac{41.5-\mu}{\sigma} < z < \frac{58.5-\mu}{\sigma}\right)$

$= P(-1.26 < z < 1.47)$

$= P(z < 1.47) - [1 - P(z < -1.26)]$

$= 0.929219 - (1 - 0.896165)$

$= 0.825384 \checkmark$

approx. prob  $\rightarrow$  normal approx. to binomial dist.

$\mu=49.35$   
 $\sigma=\sqrt{6.24391704}$

$\mu=49.35$   
 $\sigma=6.24391704$

# Lec12 - Normal Approximation to the Binomial, OMIT normal approx to the fish)

Monday, October 3, 2022 12:39 AM

Yuxi Qin  
[yuxiqin.ca](http://yuxiqin.ca)

## NORMAL APPROX. TO THE BINOMIAL DISTRIBUTION

- If  $X$  is a binomial random variable with parameters  $n$  and  $p$

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

is approximately a standard normal random variable. To approximate a binomial probability with a normal distribution, a continuity correction is applied:

$$P(X \leq x) = P(X \leq x + 0.5) \approx P\left(Z \leq \frac{x + 0.5 - np}{\sqrt{np(1-p)}}\right)$$

$$P(x \leq X) = P(x - 0.5 \leq X) \approx P\left(\frac{x - 0.5 - np}{\sqrt{np(1-p)}} \leq Z\right)$$

$\pm 0.5$

- Holds up for  $np > 5$  and  $n(1-p) > 5$

### RECALL:

- That for a binomial random variable,  $X$ ,  $E(X) = np$  and  $V(X) = np(1-p)$
- The approximation can be used when  $n$  is large relative to  $p$ 
  - Ex. P81, ex. 4.14, 4.15

## CONTINUITY CORRECTION: $\pm 0.5$ accordingly

- Applied when you want to use a continuous distribution to approximate a discrete distribution

using binomial distribution	using normal distribution w continuity correction
$X = 45$	$44.5 < X < 45.5$
$X \leq 45$	$X < 45.5$
$X < 45$	$X < 44.5$
$X \geq 45$	$X > 44.5$
$X > 45$	$X > 45.5$

## EXAMPLE 4.15 | Normal Approximation to Binomial

Again consider the transmission of bits in Example 4.14. To judge how well the normal approximation works, assume that only  $n = 50$  bits are to be transmitted and that the probability of an error is  $p = 0.1$ . The exact probability that two or fewer errors occur is

$$P(X \leq 2) = \binom{50}{0} 0.9^{50} + \binom{50}{1} 0.1(0.9^{49}) + \binom{50}{2} 0.1^2(0.9^{48}) = 0.112$$

Based on the normal approximation,

$$\begin{aligned} P(X \leq 2) &= P\left(\frac{X - 5}{\sqrt{50(0.1)(0.9)}} \leq \frac{2.5 - 5}{\sqrt{50(0.1)(0.9)}}\right) \\ &\approx P(Z < -1.18) = 0.119 \\ \rightarrow P(X \leq 2) &\approx P\left(z \leq \frac{2.5 - 5 - 50(0.1)}{\sqrt{50(0.1)(1-0.1)}}\right) \\ &\approx P\left(z \leq \frac{-2.5}{\sqrt{4.5}}\right) \\ &\approx P(z \leq -1.199) \\ &= 0.119 \end{aligned}$$

We can also approximate  $P(X = 5)$  as

$$\begin{aligned} P(5 \leq X \leq 5) &= P(4.5 \leq X \leq 5.5) \\ &\approx P\left(\frac{4.5 - 5}{2.12} \leq Z \leq \frac{5.5 - 5}{2.12}\right) \\ &= P(-0.24 \leq Z \leq 0.24) = 0.19 \end{aligned}$$

and this compares well with the exact answer of 0.1849.

Practical Interpretation: Even for a sample as small as 50 bits, the normal approximation is reasonable, when  $p = 0.1$ .

## EXAMPLE 4.13

Assume that in a digital communication channel, the number of bits received in error can be modeled by a binomial random variable, and assume that the probability that a bit is received in error is  $1 \times 10^{-5}$ . If 16 million bits are transmitted, what is the probability that 150 or fewer errors occur?

Let the random variable  $X$  denote the number of errors. Then  $X$  is a binomial random variable and

### EXAMPLE 4.14

The digital communication problem in Example 4.13 is solved as follows:

$$\begin{aligned} P(X \leq 150) &= P(X \leq 150.5) \\ &= P\left(\frac{X - 160}{\sqrt{160(1 - 10^{-5})}} \leq \frac{150.5 - 160}{\sqrt{160(1 - 10^{-5})}}\right) \\ &\approx P(Z \leq -0.75) = 0.227 \end{aligned}$$

$$P(X \leq 150) = \sum_{x=0}^{150} \binom{16,000,000}{x} (10^{-5})^x (1 - 10^{-5})^{16,000,000-x}$$

Practical Interpretation: Clearly, this probability is difficult to compute. Fortunately, the normal distribution can be used to provide an excellent approximation in this example.

Because  $np = (16 \times 10^6)(1 \times 10^{-5}) = 160$  and  $n(1-p)$  is much larger, the approximation is expected to work well in this case.

Practical Interpretation: Binomial probabilities that are difficult to compute exactly can be approximated with easy-to-compute probabilities based on the normal distribution.

Q: what is the probability that 150 or fewer errors occur?

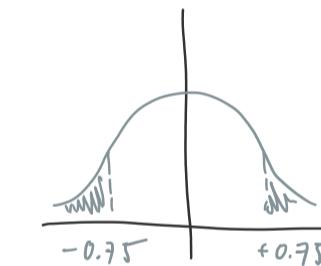
$$P(X \leq 150)$$

If  $X$  is a binomial random variable (RV) with parameters  $n$  and  $p$

$$z = \frac{X - np}{\sqrt{np(1-p)}} \rightarrow \text{apply continuity correction}$$

$$\begin{aligned} P(X \leq x) &= P(X \leq x + 0.5) \\ &\approx P\left[z \leq \frac{x + 0.5 - np}{\sqrt{np(1-p)}}\right] \end{aligned}$$

$$\begin{aligned} P(X \leq 150) &= P\left(z \leq \frac{150 + 0.5 - 160}{\sqrt{160(1 - 10^{-5})}}\right) \\ &= P(z \leq -0.75) \end{aligned}$$



$$P(z < -0.75) = 0.77337$$

$$1 - 0.77337 = 0.227$$

$$\therefore P(X \leq 150) = 0.227$$

### CM 3.3 — also on midterm 1

Problem #3: The weight of a sophisticated running shoe is normally distributed with a mean of 15 ounces.

(a) What must the standard deviation of weight be in order for the company to state that 95% of its shoes weight less than 16 ounces?

(b) Suppose that the standard deviation is actually 0.89. If we sample 8 such running shoes, find the probability that exactly 2 of those shoes weigh more than 16 ounces.

$$\text{use } z = \frac{x - \mu}{\sigma}$$

$$\mu = 15$$

$$n) ?$$

$$P(z < \frac{x - \mu}{\sigma}) = 0.95 \quad \text{from z-tables: 1.64}$$

$$x = 16$$

$$\frac{16 - 15}{\sigma} = 1.64$$

$$\text{desired}$$

$$\text{rearrange}$$

$$\sigma = \frac{16 - 15}{1.64} = 0.609756 \quad \checkmark$$

$$\text{binomial dist: } f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$b) \sigma = 0.89; x = 16$$

$$P(2 \text{ of } 8 > 16) \quad \text{use binomial dist}$$

1. prob of 1 shoe weighing > 16 ounces → success prob. used in z

2. prob of exactly 2 of 8 shoes succeed

$$\begin{aligned} i) P(z > \frac{16 - 15}{0.89}) &= 1 - P(z > 1.12) \quad \text{from z-tables:} \\ &\downarrow 1.1235 \\ &= 1 - 0.868643 \\ &= 0.131357 \quad \text{→ prob. of success} \end{aligned}$$

$$ii) \binom{8}{2} (0.131357)^2 (1 - 0.131357)^6$$

$$= 28(0.017254661)(0.42958349)$$

$$= 0.20754489 \quad \checkmark$$

# Lec13 - exponential distribution, two or more random variables, omit discrete rvs, omit conditional prob distributions, omit more than two random vars

Friday, October 21, 2022 8:00 PM

Yuxi Qin  
yuxiqin.ca

## THE EXPONENTIAL DISTRIBUTION

- If  $X$  follows an **exponential distribution**, then

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0, \lambda > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbb{E}[X] = 1/\lambda \text{ and } \text{Var}[X] = 1/\lambda^2$$

## MEMORY-LESS PROPERTY

$X \sim \exp(\lambda)$  adheres to the "memoryless property." That is,

$$P(X \geq t + t_0 \mid X \geq t_0) = P(X \geq t),$$

where  $t \geq t_0 \geq 0$  are real-valued constants. You can consider  $t_0$  and  $t$  to be the amount of time that has passed until some event occurs.

Now,  $P(X \geq t) = 1 - F(t)$ , where  $F(t)$  is the cdf for the exponential distribution. Therefore,  $P(X \geq t) = 1 - (1 - e^{-\lambda t}) = e^{-\lambda t}$ .

It follows that

$$\begin{aligned} P(X \geq t + t_0 \mid X \geq t_0) &= \frac{P(X \geq t + t_0 \cap X \geq t_0)}{P(X \geq t_0)} & (1) \\ &= \frac{P(X \geq t_0)}{P(X \geq t_0)} & (2) \\ &= \frac{1 - F(t_0)}{1 - F(t_0)} & (3) \\ &= \frac{1 - (1 - e^{-\lambda(t+t_0)})}{1 - (1 - e^{-\lambda t_0})} & (4) \\ &= \frac{e^{-\lambda(t+t_0)}}{e^{-\lambda t_0}} & (5) \\ &= e^{-\lambda((t+t_0)-t_0)} & (6) \\ &= e^{-\lambda t} & (7) \end{aligned}$$

Therefore,  $P(X \geq t + t_0 \mid X \geq t_0) = P(X \geq t)$  as required.



## POISSON DISTRIBUTION

- A discrete frequency distribution that gives you the probability of a number of **independent events occurring in a fixed time interval**
- Eg. Number of patients arriving at a waiting room between 10am and 11am

## EXPONENTIAL DISTRIBUTION

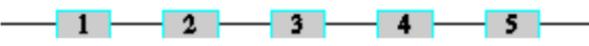
- Describes the time between events in a Poisson process
- I.e. Events occur continuously and independently, at a constant average rate
- $\lambda$  = number of events per unit time-rate parameter

## MEMORY-LESS

- Light bulb example
  - Given a bulb has survived  $A$  time units, the chance that it lasts another  $B$  time units is the same as that of a new bulb
  - Past history doesn't matter!
- Eg. Time between successive failures of a device
- The occurrence of one event does not affect the probability that a second event will occur!!

## CM 3.5

**Problem #5:** A system consists of five components connected in series as shown below.



As soon as one component fails, the entire system will fail. Assume that the components fail independently of one another.

(a) Suppose that each of the first two components have lifetimes that are exponentially distributed with mean 93 weeks, and that each of the last three components have lifetimes that are exponentially distributed with mean 130 weeks. Find the probability that the system lasts at least 55 weeks.

(b) Now suppose that each component has a lifetime that is exponentially distributed with the same mean. What must that mean be (in years) so that 97% of all such systems lasts at least one year?

$$\begin{aligned} \text{a) } & \begin{array}{c} \textcircled{1} \textcircled{2} \\ \lambda = 93 \\ \textcircled{3} \textcircled{4} \textcircled{5} \\ \lambda = 130 \end{array} \quad \begin{array}{l} f(x) = \frac{1}{\lambda} \\ \text{Var}(x) = \frac{1}{\lambda^2} \end{array} \\ & \begin{array}{l} \mathbb{P}(x \geq 55) \\ = \left[ e^{-\frac{1}{93}(55)} \right]^2 \left[ e^{-\frac{1}{130}(55)} \right]^3 \\ = (0.3064208)(0.28104729) \\ = 0.08611887 \end{array} \quad \begin{array}{l} 1 \text{ year} = 52 \text{ weeks} \\ \text{lasts 1 year} \end{array} \\ & \begin{array}{l} 0.97 = \left( e^{-\frac{1}{\lambda}} \right)^5 \\ \sqrt[5]{0.97} = e^{-\frac{1}{\lambda}} \\ \ln\left(\sqrt[5]{0.97}\right) = -\frac{1}{\lambda} \\ \frac{1}{\lambda} = -0.00609184 \\ \lambda = 164.154 \end{array} \end{aligned}$$

## Example 1.

Let  $X$  equal the amount of time between two successive arrivals at the drive-up = 1 window of a local bank. If  $X$  has an exponential distribution with  $\lambda = 1$ , compute the following:

- A. The expected time between two successive arrivals

$$\mathbb{E}(X) = \frac{1}{\lambda} = \frac{1}{1} = 1$$

- B. The standard deviation of the time between successive arrivals

$$\text{Var}[X] = \frac{1}{\lambda^2} = \frac{1}{1^2} = 1 \quad \text{SD}[X] = \sqrt{\text{Var}[X]} = 1$$

- D.  $P(2 \leq x \leq 5)$

$$F(x) = \int_0^x f(t)dt = \int_0^x e^{-t} dt = \left[ -e^{-t} \right]_0^x = 1 - e^{-x}$$

$$P(2 \leq x \leq 5) = F(5) - F(2) = (1 - e^{-5}) - (1 - e^{-2}) = 0.129$$

→ solve using calculator:

$$\int(e^{-x}, 2, 5) = 0.128597$$

## Example 2.

Logons modelled by a poisson process with

$\lambda = 25$  log-ons per hour 60 mins

Q: What is the probability that there are no log-ons in an interval of 6 minutes?

$$\lambda = 25 \text{ per 60 mins}$$

$$\frac{60}{6} = 10$$

There are 10, 6-minute intervals in an hr

∴ our interval: 0.1 hr

random variable:  $X = \text{distance b/w successive events}$

$$\begin{aligned} f(x) &= \lambda e^{-\lambda x} \\ P(X > 0.1) &= \int_{0.1}^{\infty} f(x) dx \quad \text{prob. that 1st event occurs after 6 mins} \\ &= \int e^{-\lambda x} dx \\ &= \frac{1}{\lambda} e^{-\lambda x} \end{aligned}$$

$$\Rightarrow \lambda \int e^{-\lambda x} dx = \lambda \left[ -\frac{1}{\lambda} e^{-\lambda x} \right]$$

$$\begin{aligned} \lambda &= 25 \\ P(X > 0.1) &= -\frac{25}{25} e^{-25 \times 0.1} \end{aligned}$$

$$P(X > 0.1) = 0 - (-0.082)$$

$$= 0.082$$

Q: What is the probability that the next log-on is between 2 and 3 minutes?

- Note: there are 60, 1 minute intervals in 1 hour

$$\frac{2}{60} = 0.033 \quad \begin{array}{l} 2 \text{ min int. is} \\ 0.033 \text{ of 1 hr} \end{array} \quad \frac{3}{60} = 0.05 \quad \begin{array}{l} 3 \text{ min int. is} \\ 0.05 \text{ of 2 hr} \end{array}$$

$$\begin{aligned} P(0.033 < x < 0.05) &\rightarrow -e^{-25x} \Big|_{0.033}^{0.05} \\ &= \left[ -e^{-25(0.05)} \right] - \left[ -e^{-25(0.033)} \right] \\ &= -0.287 - (-0.438) \\ &= -0.287 + 0.438 = 0.151 \end{aligned}$$

Q: Determine the time interval such that the probability that no log-on occurs in the interval is 0.9

$$\begin{aligned} P(X > x) &= 0.9 \quad \begin{array}{l} \text{prob of 0.9 that the 1st} \\ \text{event occurs after time } x \end{array} \\ -e^{-25x} \Big|_{x}^{\infty} &= 0.9 \\ e^{-25x} &= 0.9 \quad \begin{array}{l} \text{natural log BS, solve for } x \\ \ln(e^{-25x}) = x \end{array} \\ -25x &= \ln(0.9) \quad \begin{array}{l} \ln(0.9) = -0.105 \\ x = \frac{-0.105}{-25} \end{array} \\ x &= 0.00421 \end{aligned}$$

Q: What is the mean time until the next log-on?

$$M = \frac{1}{\lambda} = \frac{1}{25} = 0.04 \text{ hrs}$$

Q: What is the standard deviation until the next log-on?

$$\sigma^2 = \frac{1}{\lambda^2} = \frac{1}{25^2} = 0.0016$$

$$\sigma = \sqrt{0.0016} = 0.04 \text{ hrs}$$

## JOINT DISTRIBUTIONS

- In many situations, we will be interested in studying more than one random variable
- To utilize information from multiple random variables, we form **joint distributions**

## JOINT PROBABILITY DENSITY FUNCTION

- Consider two continuous random variables X and Y

The function  $f_{X,Y}(x,y)$ , for the pair of continuous random variables X and Y, is called the **Joint Probability Density Function** if

$$\begin{aligned} f_{X,Y}(x,y) &\geq 0 \quad \text{for all } (x,y) \in \mathbb{R}^2 \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx dy &= 1 \end{aligned}$$

and then for any region  $A \subseteq \mathbb{R}^2$  we have

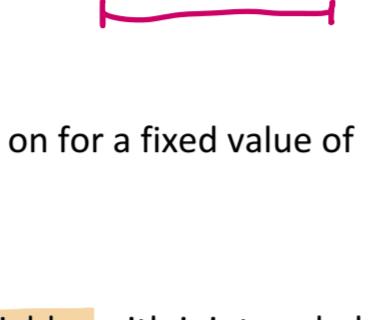
$$P(X, Y \in A) = \iint_{(x,y) \in A} f(x,y) dx dy$$

## CALCULATING PROBABILITIES

- We will be interested in calculating probabilities of the form  $P(X, Y \in A)$ .

- If A is a rectangle with corner  $(a,c), (a,d), (b,c)$ , and  $(b,d)$  then

$$P(X, Y \in A) = P(a < X < b, c < Y < d) = \int_a^b \int_c^d f_{X,Y}(x,y) dx dy$$



- If A takes on a different shape, then we have

$$P(X, Y \in A) = \int_a^b \int_{c(x)}^{d(x)} f_{X,Y}(x,y) dy dx$$

Where  $c(x)$  and  $d(x)$  are the limits that Y can take on for a fixed value of  $(x, y) \in A$ .

## MARGINAL PROBABILITY DENSITY FUNCTION

- Suppose that X and Y are continuous random variables with joint probability density function  $f_{X,Y}(x,y)$
- You might be interested in the probability distribution of each variable individually, which is called the **Marginal Probability Distribution**
- The **Marginal Probability Density Functions** for X and Y are given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy \quad \text{and} \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx$$

## INDEPENDENT CONTINUOUS RANDOM VARIABLES

- Suppose that X and Y are two continuous random variables with joint probability density function  $f_{X,Y}(x,y)$  and marginal pdfs  $f_X(x)$  and  $f_Y(y)$
- Then X and Y are said to be independent if, and only if,

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

for every pair of x and y values

- If this does not hold for every pair  $(x,y)$ , then the random variables are said to be dependent
- If they are independent, then their marginal probability distributions can be used to calculate the probabilities

individual probabilities!

## INDEPENDENCE: EXAMPLES (P106 - EX 5.9, 5.10)

## EXAMPLE 5.9 | Independent Random Variables

Suppose that Example 5.2 is modified so that the joint probability density function of X and Y is  $f_{X,Y}(x,y) = 2 \times 10^{-6} \exp(-0.001x - 0.002y)$  for  $x \geq 0$  and  $y \geq 0$ . Show that X and Y are independent and determine  $P(X > 1000, Y < 1000)$ .

Note that the range of positive probability is rectangular so that independence is possible but not yet demonstrated. The marginal probability density function of X is

$$\begin{aligned} f_X(x) &= \int_0^{\infty} 2 \times 10^{-6} e^{-0.001x - 0.002y} dy \\ &= 0.001e^{-0.001x} \quad \text{for } x > 0 \\ &= 2 \cdot 10^{-6} \left[ -\frac{1}{0.002} e^{-0.001x - 0.002y} \right]_0^{\infty} \\ &= 2 \cdot 10^{-6} \left[ 0 - \left( -\frac{1}{0.002} e^{-0.001x - 0} \right) \right] \\ &= \frac{2 \cdot 10^{-6}}{0.002} e^{-0.001x} \\ &= 0.001e^{-0.001x} \end{aligned}$$

The marginal probability density function of Y is

$$\begin{aligned} f_Y(y) &= \int_0^{\infty} 2 \times 10^{-6} e^{-0.001x - 0.002y} dx \\ &= 2 \cdot 10^{-6} \left[ \frac{1}{0.001} e^{-0.001x - 0.002y} \right]_0^{\infty} \\ &= 2 \cdot 10^{-6} \left[ 0 - \left( \frac{1}{0.001} e^{-0.001x - 0.002y} \right) \right] \\ &= \frac{2 \cdot 10^{-6}}{0.001} e^{-0.002y} \\ &= 0.002e^{-0.002y} \end{aligned}$$

$$\begin{aligned} \text{To show } f_{X,Y}(x,y) &= f_X(x)f_Y(y) \\ &= 0.001e^{-0.001x} (0.002e^{-0.002y}) \\ &= 2 \cdot 10^{-6} e^{-0.001x - 0.002y} \\ &= f_X(x)f_Y(y) \end{aligned}$$

## EXAMPLE 5.10 | Machined Dimensions

Let the random variables X and Y denote the lengths of two dimensions of a machined part, respectively. Assume that X and Y are independent random variables, and further assume that the distribution of X is normal with mean 10.5 millimeters and variance 0.0025 ( $\text{mm}^2$ ) and that the distribution of Y is normal with mean 3.2 millimeters and variance 0.0036 ( $\text{mm}^2$ ). Determine the probability that  $10.4 < X < 10.6$  and  $3.15 < Y < 3.25$ .

Because X and Y are independent,

$$\begin{aligned} P(10.4 < X < 10.6, 3.15 < Y < 3.25) &= P(10.4 < X < 10.6) \times P(3.15 < Y < 3.25) \\ &= P\left(\frac{10.4 - 10.5}{0.05} < Z < \frac{10.6 - 10.5}{0.05}\right) \times P\left(\frac{3.15 - 3.2}{0.06} < Z < \frac{3.25 - 3.2}{0.06}\right) \end{aligned}$$

- X and y are independent random variables
- Given mean and variance
- Both x and y are normally distributed

$$= P(-2 < Z < 2) P(-0.833 < Z < 0.833)$$

$$= 0.568$$

where Z denotes a standard normal random variable.

Practical Interpretation: If random variables are independent, probabilities for multiple variables are often much easier to compute.

$z = \text{standard normal random variable}$

$$M_x = 10.5 \rightarrow z^2 = 0.0025$$

$$M_y = 3.2 \rightarrow z^2 = 0.0036$$

- draw diagrams
- look up z-tables
- use positive z-values — remember that the tables give you area/prob. to the left of the value

→ use domains of x, y to determine bounds for integrals, go inside out!

Note: because the random variables are independent, you can multiply the probabilities

- Mae sure you use the correct distribution to calculate the probabilities

CM 3.6 — Also on midterm 2 (variant can be found in 2019 MT2)

**Problem #6:** Suppose that the random variables X and Y have the following joint probability density function.

$$f(x,y) = ce^{-7x-3y}, \quad 0 < y < x.$$

(a) Find the value of c.

(b) Find  $P(X < \frac{1}{6}, Y < 2)$

$$\text{joint density function} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx dy = 1$$

$$f(x,y) = ce^{-7x-3y}, \quad 0 < y < x$$

$$a) c = ?$$

$$\int_0^{\infty} \int_0^x ce^{-7x-3y} dy dx = 1$$

$$1 = \int_0^{\infty} -ce^{-7x} \left( e^{-3y} \right) \Big|_0^x dx$$

$$1 = \frac{c}{70} \left( e^{-7x} - 1 \right) \Big|_0^{\infty}$$

$$1 = \frac{c}{70} (1 - 0) \quad \text{what happens here?}$$

$$b) P(X < \frac{1}{6}, Y < 2)$$

$$\int_0^{\frac{1}{6}} \int_0^x ce^{-7x-3y} dy dx$$

Symbolic & desmos calc

$$= 0.4024984936 \quad \checkmark$$

mean

$$E[X] = 3 \quad V[X] = 0.7$$

$$E[Y] = 5 \quad V[Y] = 1.1$$

$$E[ax+by] = aE[X] + bE[Y] \quad \text{Var}[ax+by] = a^2 \text{Var}[X] + b^2 \text{Var}[Y]$$

$$a) E(5X_1 + 9X_2) = 5E[X_1] + 9E[X_2]$$

$$= 5(3) + 9(5) = 60 \quad \checkmark$$

$$b) V(5X_1 + 9X_2) = V(5X_1) + V(9X_2) = 5^2(0.7) + 9^2(1.1) = 12.25 + 98.01 = 110.26 \quad \text{Variance}$$

$$c) P(X_1 > 7) = P\left(Z > \frac{7-3}{\sqrt{0.7}}\right) = 1 - P(Z \leq \frac{4}{\sqrt{0.7}}) = 1 - 0.85341 = 0.146859 \quad \checkmark$$

$$\sigma = \sqrt{0.7} = 0.866 \quad \checkmark$$

$$\sigma = \sqrt{110.26} = 10.50047618 \quad \checkmark$$

$$\text{Std dev.} = \sqrt{110.26} = 10.50047618 \quad \checkmark$$

## CALCULATING PROBABILITIES: EXAMPLE 1

- Given the joint probability density function

$$f_{X,Y}(x,y) = \begin{cases} \frac{3}{5}x(y+x) & \text{for } 0 < x < 1, 0 < y < 2 \\ 0 & \text{elsewhere} \end{cases}$$

- Find the  $P[(X \cap Y) \in A]$ , where A is defined as the region  $(x,y) \mid 0 < x < \frac{1}{2}, 1 < y < 2$ . → domain

$$\begin{aligned} P[(X \cap Y) \in A] &= P\left(0 < x < \frac{1}{2}, 1 < y < 2\right) \\ &= \int_1^2 \int_0^{\frac{1}{2}} \frac{3}{5}x(y+x) dx dy \\ &= \int_1^2 \frac{3x^2y}{10} + \frac{3x^3}{15} \Big|_0^{\frac{1}{2}} dy \\ &= \int_1^2 \left( \frac{3y}{40} + \frac{1}{40} \right) dy = \frac{3y^2}{80} + \frac{y}{40} \Big|_1^2 = \frac{11}{80} \\ &= \frac{12}{80} + \frac{4}{80} - \left[ \frac{3(1)^2}{80} + \frac{1}{40} \right] \\ &= \frac{16}{80} - \left[ \frac{3}{80} + \frac{1}{40} \right] \\ &= \frac{16}{80} - \frac{5}{80} = \frac{11}{80} \end{aligned}$$

## MARGINAL DISTRIBUTION: EXAMPLE 1

- Given the joint probability density function

$$f_{X,Y}(x,y) = \begin{cases} 4xy & \text{for } 0 < x < 1, 0 < y < 1 \\ 0 & \text{elsewhere} \end{cases}$$

- Find the marginal distribution of X and Y

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy = \int_0^1 4xy dy = 2xy \Big|_0^1 = 2x \\ f_Y(Y) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx = \int_0^1 4xy dx = 2x^2 y \Big|_0^1 = 2y \end{aligned}$$

marginal — use equations, leave in terms of desired variable, and use domain/bounds of other variable

## EXERCISES

- Given the values of the point probability distribution of X and Y shown in the table

	X
-1	1
Y	0
1	0

- A. What is the marginal distribution of X?

$$P(X=-1) = \sum_{y \in S_Y} = \frac{1}{8} + 0 + \frac{1}{8} = \frac{1}{4}$$

$$P(X=1) = \sum_{y \in S_Y} = \frac{1}{2} + \frac{1}{4} + 0 = \frac{3}{4}$$

- B. What is the marginal distribution of Y?

$$P(Y=-1) = \sum_{x \in S_X} = \frac{1}{8} + \frac{1}{2} = \frac{5}{8}$$

$$P(Y=0) = \sum_{x \in S_X} = 0 + \frac{1}{4} = \frac{1}{4}$$

$$P(Y=1) = \sum_{x$$





# Lec17 - histograms and boxplots

Yuxi Qin  
yuxiqin.ca

Monday, October 24, 2022 6:29 PM

Histograms - section 6.3, box-plots - 6.4

## BACKGROUND: HISTOGRAMS

- Histograms work best for large data-sets, preferably greater than 100 observations
- Histograms displays utilize frequency distribution
- The vertical scale shows frequency or relative frequency, where the sum of the heights = 1 ("density" is used when the total area under the histogram = 1)
- Data is divided into intervals, usually called "bins"
- Ideally, bins should be of equal width
- For unequal bin width, the rectangle area is proportional to the bin frequency
- There are many methods of selecting bin width. Most commonly used is:  
**Number of bins = square root of number of observations**
- For cumulative frequency plots, the height of each rectangle is that **number of observations that are <= the upper limit of that rectangle**

## SYMMETRIC, BELL-SHAPED HISTOGRAM

### Histogram of the Table 6-2 Data

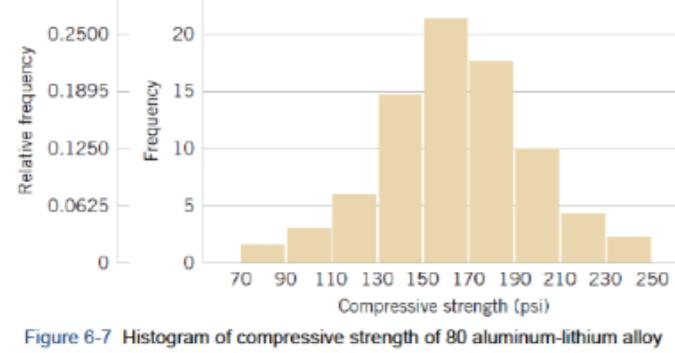


Figure 6-7 Histogram of compressive strength of 80 aluminum-lithium alloy specimens. Note these features – (1) horizontal scale, bin boundaries & labels with units, (2) vertical scale measurements and labels, (3) histogram title at top or in legend.

## CUMULATIVE DISTRIBUTION PLOT

### Cumulative Frequency Plot

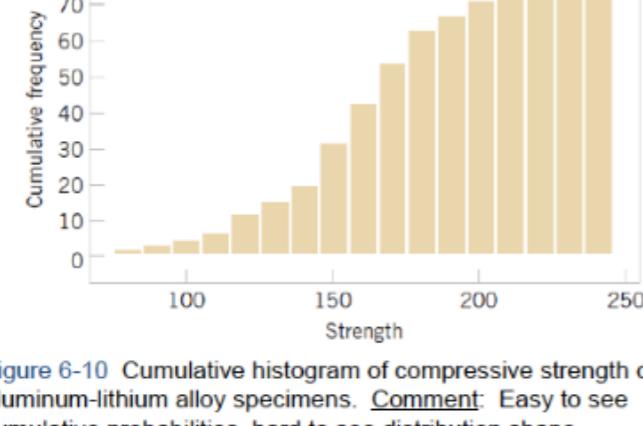


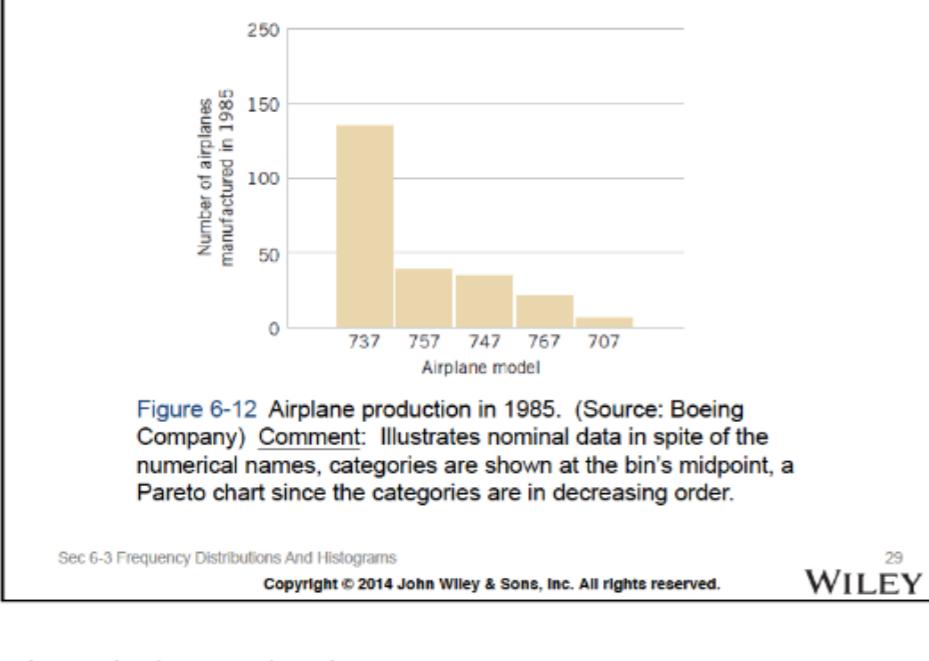
Figure 6-10 Cumulative histogram of compressive strength of 80 aluminum-lithium alloy specimens. Comment: Easy to see cumulative probabilities, hard to see distribution shape.

## NOTES ON SHAPES OF HISTOGRAMS

- Unimodal:** describes the shape of a histogram of some data
- Other terms used to describe shape: symmetric, positively or negatively skewed
- If a distribution looks symmetric, the median and mean should approximately coincide
- For a unimodal symmetric distribution, the mean, median, and mode all coincide

## CATEGORICAL HISTOGRAMS

- Histograms can also be used to display categorical or qualitative data
- In this case, bin sizes must be equal
- Categorical histograms are sometimes called Pareto charts, if the categories are ordered in terms of frequency



## BOXPLOTS: REVISE IQR

- The **median** is the 'middle' observation, or the 50th percentile
- The **interquartile range** is the size of the gap between the first and third quartile
  - Ie. The 'distance' over which the 'middle half' of the data is spread
  - Or the 'distance' between the 25th and 75th percentiles

## BOXPLOTS

- Boxplots are also called box-and-whisker plots
- They display the center, spread, symmetry, and outliers of a dataset
- They also include the interquartile range, max and min of the data set
- The centre 'box' marks the interquartile range
- The 'whiskers' extend to data points within 1.5 times the interquartile range

## Box Plot or Box-and-Whisker Chart

- A box plot is a graphical display showing center, spread, shape, and outliers (SOCS).
- It displays the 5-number summary: **min,  $q_1$ , median,  $q_3$ , and max.**

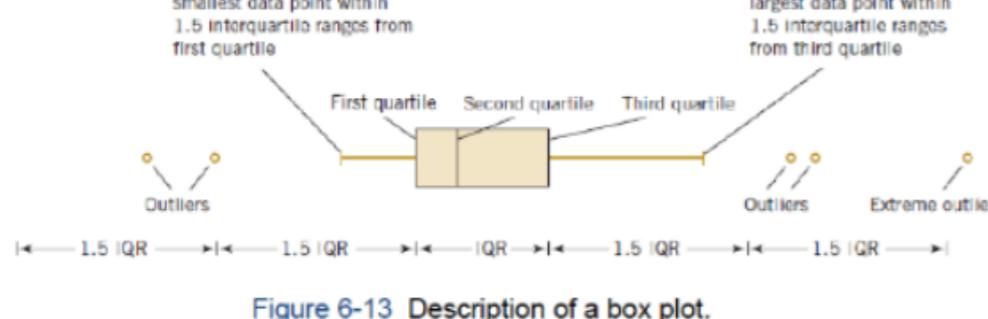


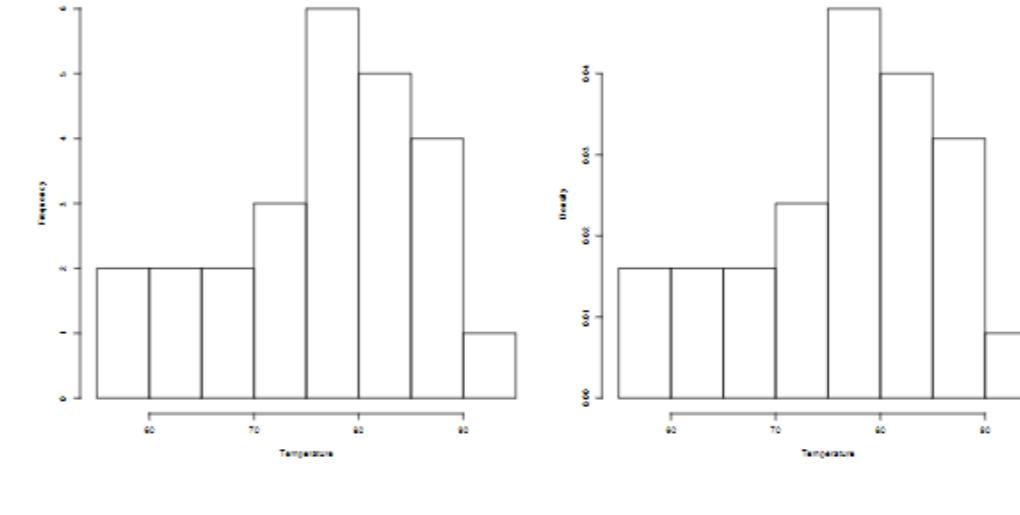
Figure 6-13 Description of a box plot.

## Ex. Temperature data

- Consider the following **univariate** (has only one random variable) data set measuring the air temperature at LaGuardia airport

Observation	Temperature (°F)	Observation	Temperature (°F)
1	85	14	81
2	87	15	79
3	79	16	74
4	86	17	79
5	88	18	77
6	68	19	61
7	81	20	92
8	86	21	72
9	58	22	76
10	64	23	72
11	57	24	79
12	83	25	83
13	66		

- Here are two histograms for the temperature data



- We can see that the bins are equally spaced and we can confirm that there is a single mode

- This is a stem-and-leaf display for the temperature data

depth	
2	7 8
6	1 4 6 8
(9)	2 2 4 6 7 9
10	1 1 3 3 5 6 6 7 8
1	2 92

- Min: 57; max: 92, median and mode: 79

## BOXPLOT (TEMPERATURE DATA)

- Draw a boxplot for the temperature data

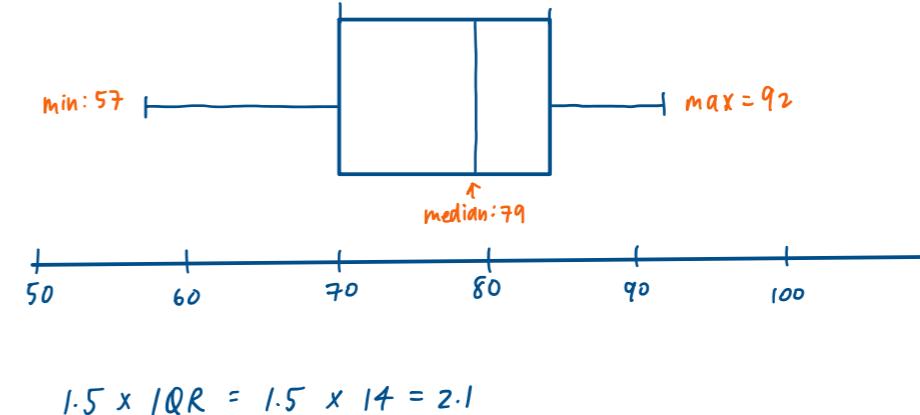
→ 25 data points

12 data pts. 13 data pt. 12 data pts.  
min → median max  
= 50% = 79 → 13th value

Find  $Q_1 = 25^{\text{th}} \text{ percent}$   
 $Q_1 = \frac{n+1}{4} = \frac{25+1}{4} = 6.5$  data pt 6: 68  
 $68 + (0.5)(72-68) = 70$  data pt 7: 72

Find  $Q_3 = 75^{\text{th}} \text{ percent}$   
 $Q_3 = \frac{3(n+1)}{4} = \frac{3(25+1)}{4} = 19.5$  data pt 19: 83  
 $83 + (0.5)(85-83) = 84$  data pt 20: 85

1QI:  $84 - 70 = 14$

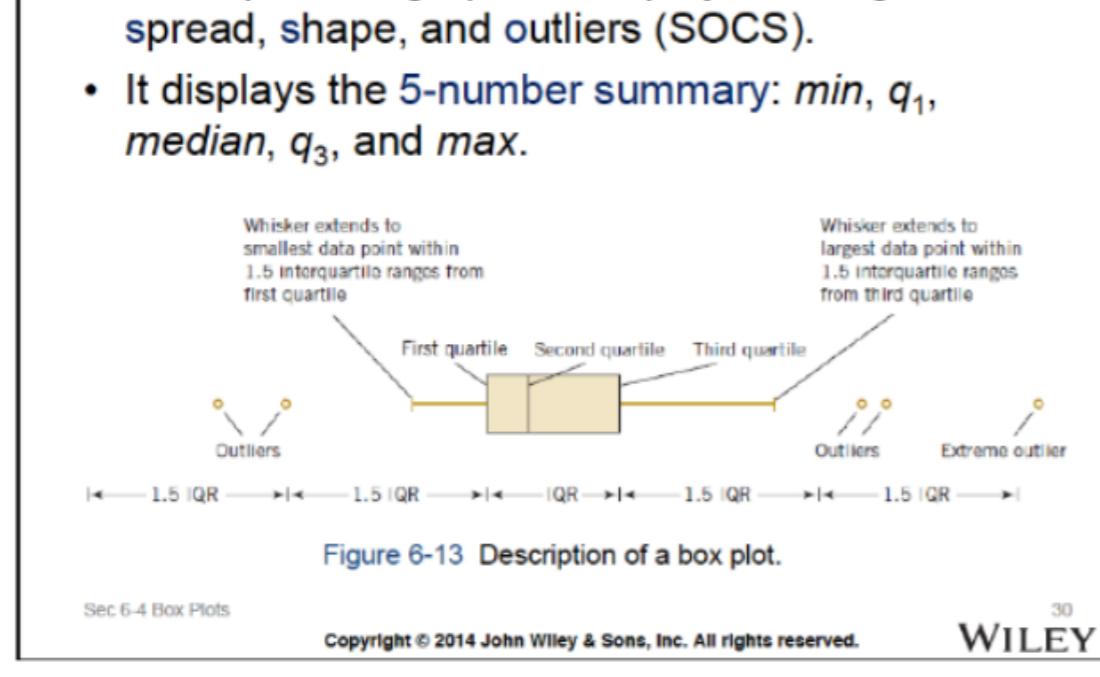


$1.5 \times 1QI = 1.5 \times 14 = 21$

→ whiskers extend to furthest data point that is still within 1.5 IQR of the 1st or 3rd quartile

## Box Plots

- Boxplots are also called box-and-whisker plots
- They display the center, spread, symmetry, and outliers of a dataset
- They also include the interquartile range, max and min of the data set
- The centre 'box' marks the interquartile range
- The 'whiskers' extend to data points within 1.5 times the interquartile range



# Lec18 - probability plots

Monday, October 24, 2022 7:56 PM

Yuxi Qin  
[yuxiqin.ca](http://yuxiqin.ca)

## Probability plots - 6.7

### BACKGROUND

- Probability plots are a visual way of assessing whether or not a data set conforms to a given distribution
- The scales on the probability plot axes will depend on what distribution you're using
- We will focus on the **normal probability plot**

### CONSTRUCTING A PROBABILITY PLOT

- Select the correct scaling for your distribution
- Rank observations from smallest to largest
- Plot the ordered observations ( $x_j$ ) against their observed cumulative frequency
  - For ordered observation ( $x_j$ ), the **observed cumulative frequency** is

$$\frac{j-0.5}{n}$$

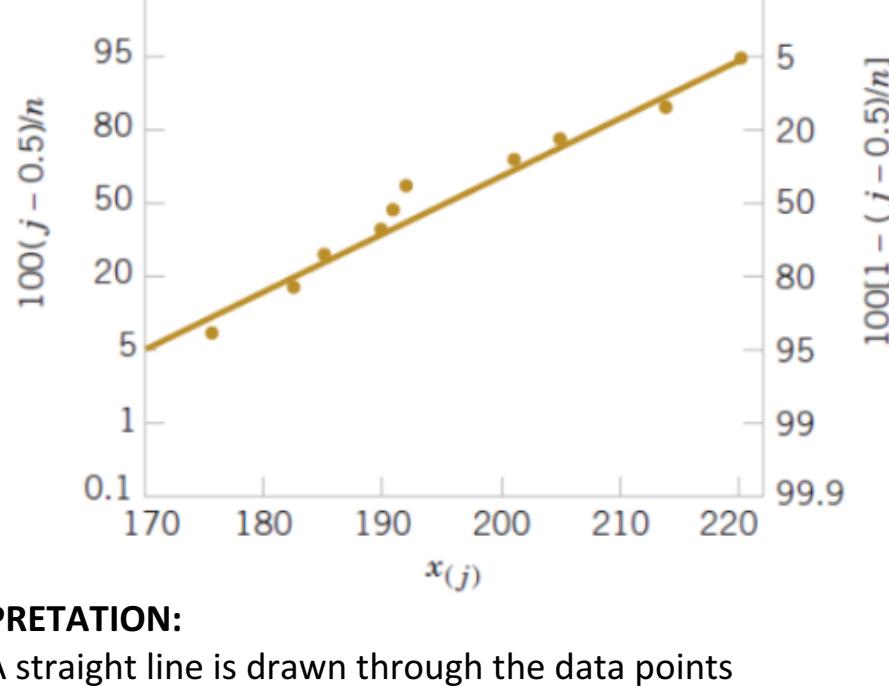
### INTERPRETING A PROBABILITY PLOT

$$100 \left[ \frac{j-0.5}{n} \right]$$

- Is usually plotted on the Y-axis and the variable is plotted on the X-axis
- If the data obeys the distribution, the data points will line up in an approximately straight line
- Often, you'll see deviation from Normality at the beginning and end of the line, which corresponds to heavy or light tailed data

Example:

#### Normal Probability Plot for Battery Life



Ex. P145 6.7

Given 10 numbers representing battery life, test if battery life is adequately represented by a N distribution

Using  $x_j = \frac{j-0.5}{n}$  and  $100 \left( \frac{j-0.5}{n} \right)$

Here,  $n=10$

$$\hookrightarrow 100 \left( \frac{j-0.5}{10} \right) = 10 (j-0.5)$$

*calculate*

plot  $10(j-0.5)$  on y-axis  
& plot  $x$  on x-axis

j	$x_j$	$\frac{10(j-0.5)}{10}$
1	176	$10(1-0.5) = 5$
2	183	$10(2-0.5) = 15$
3	185	.
4	190	$\approx 35$
5	191	.
6	192	$\approx 45$
7	201	.
8	205	$\approx 55$
9	214	.
10	220	$10(10-0.5) = 95$

### INTERPRETATION:

- A straight line is drawn through the data points
- The interquartile range should usually have the strongest influence on line position, ie. Position the line between the 25th and 75th percentile points
- Fig 6.22 on p145 represents a situation where it is appropriate to model the data with a normal distribution

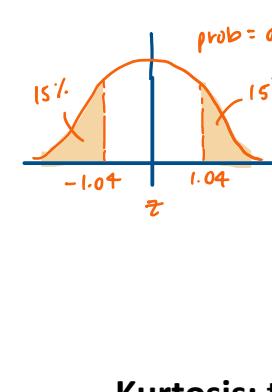


TABLE 6.6 Calculation for Constructing a Normal Probability Plot			
j	$x_j$	$(j-0.5)/10$	$z_j$
1	176	0.05	-1.64
2	183	0.15	-1.04
3	185	0.25	-0.67
4	190	0.35	-0.39
5	191	0.45	-0.13
6	192	0.55	0.13
7	201	0.65	0.39
8	205	0.75	0.67
9	214	0.85	1.04
10	220	0.95	1.64

### STANDARDIZED NORMAL PROBABILITY PLOT

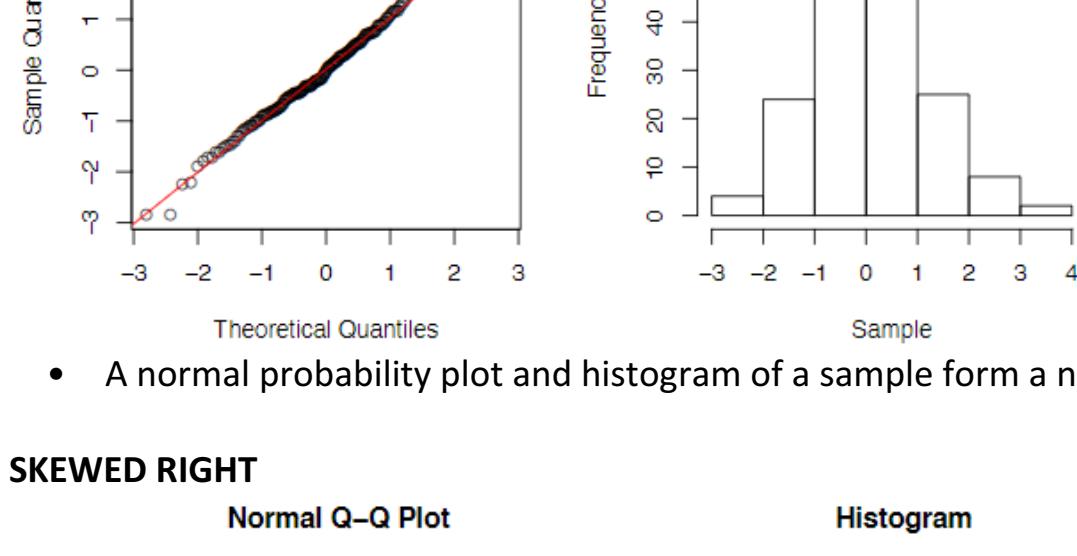
- A normal probability plot can also be constructed by plotting z-scores on the Y-axis and the variable on the X-axis

$$\frac{j-0.5}{n} = P(Z \leq z_j) = \phi(z_j) \rightarrow \phi^{-1} \text{ normal quartile function}$$

### INTERPRETING NORMALITY AND NON-NORMALITY

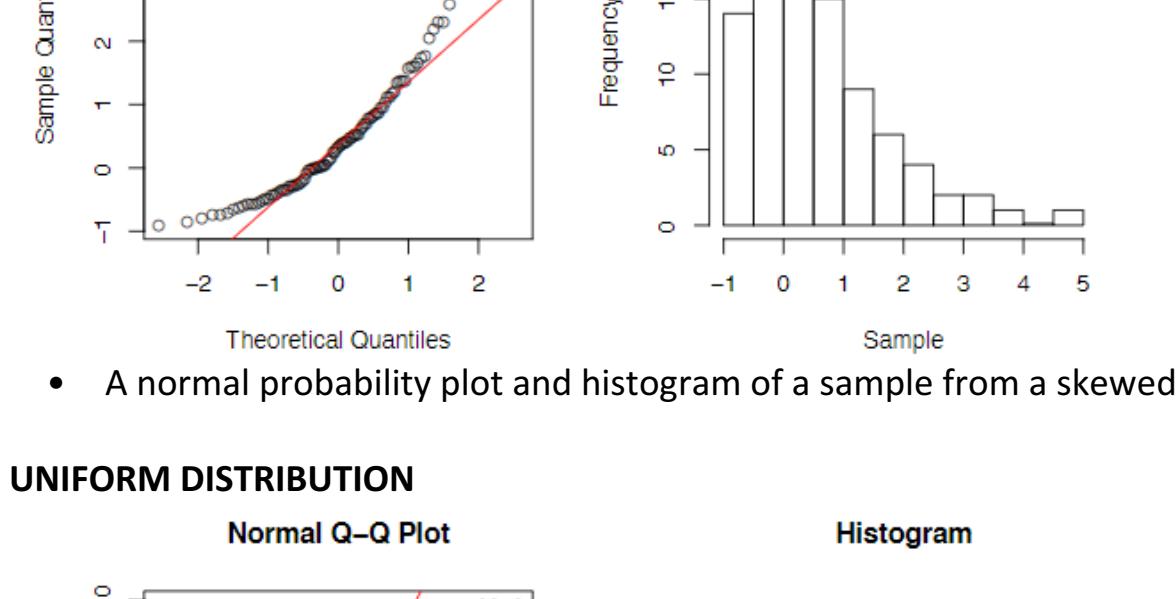
- Probability plots can identify deviations from normality such as skewness and light/heavy tails
- The larger the sample size, the more sample should conform to normality (if the underlying distribution really is normal)

### NORMAL PLOT



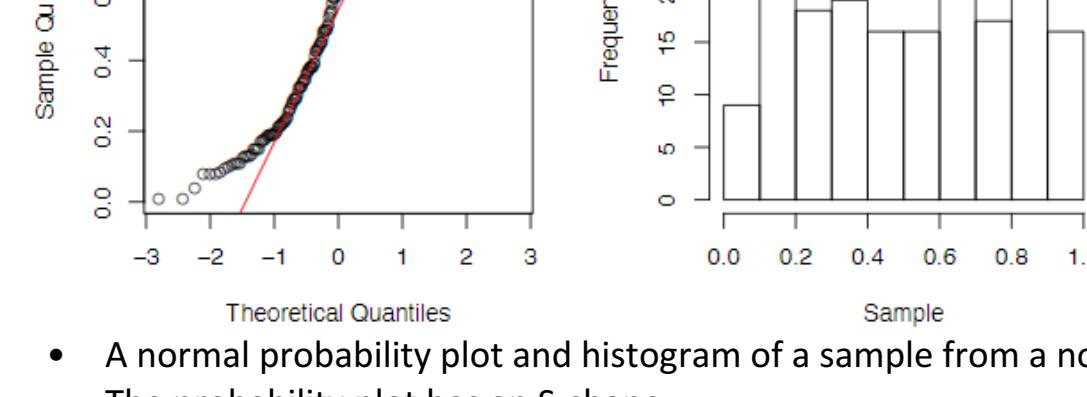
- A normal probability plot and histogram of a sample form a normal distribution

### SKEWED RIGHT



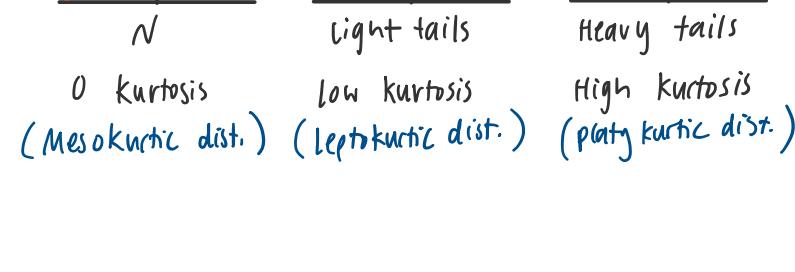
- A normal probability plot and histogram of a sample from a skewed distribution

### UNIFORM DISTRIBUTION



- A normal probability plot and histogram of a sample from a normal distribution. The probability plot has an S-shape.

**Kurtosis:** the sharpness of the peak of a frequency-distribution curve



**INTRO TO POINT/UNBIASED/MIN VARIANCE ESTIMATION**

- Suppose that we observe or collect a random sample from some population
- The parameter  $\theta$  (theta), is an unknown characteristic or parameter of the distribution
- There are many ways to estimate the value of  $\theta$ , and each use the observed/collected sample

**RECALL: ESTIMATORS AND ESTIMATES**

- A statistic is a function of a random sample:  $X_1, \dots, X_n$
- As such, the estimator of a distributions' parameter  $\theta$  is simply a statistic
- Therefore, an estimator is also a random variable
- So we can call the observed value of that random variable an estimate of  $\theta$
- Again, the observed value is based on the sample

**PROPERTIES OF ESTIMATORS**

- We can derive the properties of an estimator using its sampling distribution
- Of particular interest are the expected value (mean), variance, and standard deviation of each estimators' sampling distribution
- Note: it is desirable to have the expected value of the estimator equal to the true value of the parameter
- It is convenient when the estimators' variability around the truth values is small

**UNBIASED ESTIMATORS**

- Suppose that  $X_1, \dots, X_n$  is a random sample from a population, whose distribution is characterized by an unknown parameter,  $\theta$
- Let  $\hat{\theta}$  be a point estimator of  $\theta$
- Then  $\hat{\theta}$  is an unbiased estimator of  $\theta$  if:

$$\mathbb{E}[\hat{\theta}] = \theta$$

- For every possible value of  $\theta$

- Note: if  $\hat{\theta}$  is not unbiased then the bias estimate of  $\hat{\theta}$  is defined to be the difference

$$\mathbb{E}[\hat{\theta}] - \theta$$

**TWO UNBIASED ESTIMATORS****PROPOSITIONS:**

- Suppose that  $X_1, \dots, X_n$  is a random sample from a population with mean  $\mu$  and sample mean

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

- The estimator  $\bar{X}$  is an unbiased estimator of  $\mu$ .

- Suppose that  $X$  has a binomial distribution with known  $n$  and unknown success probability  $p$

- Then an unbiased estimator of  $p$  is

$$\hat{p} = \frac{X}{n}$$

**PROOFS**

- The estimator  $\bar{X}$  is an unbiased estimator of  $\mu$

$$\mathbb{E}[\bar{X}] = \mathbb{E}\left[\frac{\sum_{i=1}^n X_i}{n}\right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i] = \frac{1}{n} n\mu = \mu$$

- An unbiased estimator of  $p$  is  $\hat{p} = \frac{x}{n}$

$$\mathbb{E}[\hat{p}] = \mathbb{E}\left[\frac{X}{n}\right] = \frac{1}{n} \mathbb{E}[X] = \frac{1}{n} np = p$$

**PRINCIPLE OF UNBIASED ESTIMATION**

- Not all reasonable estimators are unbiased
- There is a principle stating that given a choice of estimators, we should prefer the unbiased one
- This guarantees that no matter what the true value of the unknown parameter is, the sampling distribution of the estimator will be centered at that value
- If unbiased estimation is not possible, we would want the bias to be small as the sample size  $n$  gets large.

**UNBIASED ESTIMATION OF THE VARIANCE**

- Suppose that  $X_1, \dots, X_n$  is a random sample from some population with mean  $\mu$  and variance  $\sigma^2 < \infty$
- Then an unbiased estimator of  $\sigma^2$  is the sample variance

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

- We could also consider the estimator

$$\hat{\sigma}_1^2 = S_1^2 = \frac{1}{n} \sum_{i=1}^N (X_i - \bar{X})^2;$$

- However, this estimator is biased but the bias goes to 0 as  $n$  gets bigger

**MINIMUM VARIANCE ESTIMATION**

- As well as requiring an estimator to be unbiased, we would like it to have small variability about its expected value
- There can be multiple unbiased estimators, so looking at the variability may help us choose between them
- The principle of Minimum Variance Unbiased Estimation (MVUE) states that among all unbiased estimators of  $\theta$ , we should select the one with smallest variance

**THE MEAN OF A GAUSSIAN DISTRIBUTION****THEOREM:**

- Suppose that  $X_1, \dots, X_n$  is a random sample from a Gaussian distribution with mean  $\mu$  and standard deviation  $\sigma$
- Then  $\hat{\mu} = \bar{X}$  is the minimum variance unbiased estimator of  $\mu$ 
  - There is a lot of research done into MVUE
  - There are theorems that show how to find such estimators in a wide variety of situations, but that work is beyond the scope of this course

**THE STANDARD ERROR**

- The standard error of an estimator is its standard deviation
- Generally, the standard deviation of an estimator is unknown since it depends on the parameters of the population distribution, which are oftentimes unknown
- Since we don't know the standard deviation, we will usually estimate it by replacing any unknown parameters with their point estimates
- The resulting quantity is called the estimated standard error
- We will generally denote the estimated standard error by  $s_{\hat{\theta}}$

**MEAN SQUARED ERROR OF AN ESTIMATOR**

- The Mean Squared Error of an Estimator  $\hat{\theta}$  of a parameter  $\theta$  is

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

- It follows

$$MSE(\hat{\theta}) = E[\hat{\theta} - E(\hat{\theta})]^2 + [\theta - E(\hat{\theta})]^2$$

$$MSE = V(\hat{\theta}) + (bias)^2$$

- The Mean Squared Error can be used to compare two estimators to see which one is better (we want the smallest error possible)

- The relative efficiency of  $\hat{\theta}_2$  to  $\hat{\theta}_1$  is

$$\frac{MSE(\hat{\theta}_1)}{MSE(\hat{\theta}_2)}$$

- An estimator with a mean squared error  $\leq$  mean squared error of any estimator, for any value of  $\theta$ , is the optimal estimator of  $\theta$

**Ex. 1 POINT ESTIMATOR**

We observe a sample of 20 observations from a normal distribution with mean  $\mu=30$  and standard deviation  $\sigma=5$

- The values in the sample are:

17.91	22.40	25.15	25.62	25.69	25.96	27.40	27.79	28.30	30.01
31.07	31.12	31.33	32.71	32.91	34.97	35.13	38.39	38.54	38.67

Consider the following point estimates for  $\mu$ :

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{601.0474}{20} = 30.052$$

$$\text{median}(x) = \tilde{x} = \frac{(30.01 + 31.07)}{2} = 30.541$$

$$\frac{[\min(x_i) + \max(x_i)]}{2} = \frac{17.91 + 38.67}{2} = 28.29$$

Example of Standard Error: p159, ex 7.4

**EXAMPLE 7.4 | Thermal Conductivity**

An article in the *Journal of Heat Transfer* (Trans. ASME, Sec. C, 96, 1974, p. 59) described a new method of measuring the thermal conductivity of Armco iron. Using a temperature of 100°F and a power input of 550 watts, the following 10 measurements of thermal conductivity (in Btu/hr-ft°F) were obtained:

$$41.60, 41.48, 42.34, 41.95, 41.86,$$

$$42.18, 41.72, 42.26, 41.81, 42.04$$

A point estimate of the mean thermal conductivity at 100°F and 550 watts is the sample mean or

$$\bar{x} = 41.924 \text{ Btu/hr-ft°F}$$

The standard error of the sample mean is  $\sigma_{\bar{X}} = \sigma/\sqrt{n}$ , and because  $\sigma$  is unknown, we may replace it by the sample standard deviation  $s = 0.284$  to obtain the estimated standard error of  $\bar{X}$  as

$$SE(\bar{X}) = \sigma_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{0.284}{\sqrt{10}} = 0.0898$$

**Practical Interpretation:** Notice that the standard error is about 0.2 percent of the sample mean, implying that we have obtained a relatively precise point estimate of thermal conductivity. If we can assume that thermal conductivity is normally distributed, two times the standard error is  $2\sigma_{\bar{X}} = 2(0.0898) = 0.1796$ , and we are highly confident that the true mean thermal conductivity is within the interval  $41.924 \pm 0.1796$  or between 41.744 and 42.104.

Ex. If  $X_1, \dots, X_n$  is a random sample from a Gaussian distribution with mean  $\mu$  and variation  $\sigma^2$ . Show that:

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

Is an unbiased estimator for  $\sigma^2$

$$\mathbb{E}(S^2) = \mathbb{E}\left[\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2\right]$$

$$= \mathbb{E}\left[\frac{1}{n-1} \left( \sum_{i=1}^n X_i^2 - \left( \sum_{i=1}^n X_i \right)^2 \right)\right]$$

$$= \frac{1}{n-1} \sum_{i=1}^n \mathbb{E}[X_i^2] - \frac{1}{n} \mathbb{E}\left[\left(\sum_{i=1}^n X_i\right)^2\right]$$

$$\text{Recall: } \text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \Rightarrow \mathbb{E}[X^2] = \text{Var}[X] + \mathbb{E}[X]^2$$

$$= \frac{1}{n-1} \left( \sum_{i=1}^n \mathbb{E}[X_i^2] - \frac{1}{n} \mathbb{E}\left[\left(\sum_{i=1}^n X_i\right)^2\right] \right)$$

$$= \frac{1}{n-1} \left\{ \sum_{i=1}^n (\text{Var}[X_i] + \mathbb{E}[X_i]^2) - \frac{1}{n} \left( \text{Var}\left[\sum_{i=1}^n X_i\right] + \mathbb{E}\left[\sum_{i=1}^n X_i\right]^2 \right) \right\}$$

$$= \frac{1}{n-1} \left\{ \sum_{i=1}^n (\sigma^2 + \mu^2) - \frac{1}{n} (n\sigma^2 + (n\mu)^2) \right\}$$

$$= \frac{1}{n-1} \left\{ n\sigma^2 + n\mu^2 - \frac{1}{n} n\sigma^2 - \frac{1}{n} (n\mu)^2 \right\}$$

$$= \frac{1}{n-1} n\sigma^2$$

$$= \frac{1}{n-1} \sigma^2$$

$$= \sigma^2$$

**THE MEAN OF A GAUSSIAN DISTRIBUTION****THEOREM:**

- Suppose that  $X_1, \dots, X_n$  is a random sample from a Gaussian distribution with mean  $\mu$  and standard deviation  $\sigma$
- Then  $\hat{\mu} = \bar{X}$  is the minimum variance unbiased estimator of  $\mu$ 
  - There is a lot of research done into MVUE
  - There are theorems that show how to find such estimators in a wide variety of situations, but that work is beyond the scope of this course

**THE STANDARD ERROR**

- The standard error of an estimator is its standard deviation
- Generally, the standard deviation of an estimator is unknown since it depends on the parameters of the population distribution, which are oftentimes unknown
- Since we don't know the standard deviation, we will usually estimate it by replacing any unknown parameters with their point estimates
- The resulting quantity is called the estimated standard error
- We will generally denote the estimated standard error by  $s_{\hat{\theta}}$

**MEAN SQUARED ERROR OF AN ESTIMATOR**

- The Mean Squared Error of an Estimator  $\hat{\theta}$  of a parameter  $\theta$  is

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

- It follows

$$MSE(\hat{\theta}) = E[\hat{\theta} - E(\hat{\theta})]^2 + [\theta - E(\hat{\theta})]^2$$

$$MSE = V(\hat{\theta}) + (bias)^2$$

- The Mean Squared Error can be used to compare two estimators to see which one is better (we want the smallest error possible)

- The relative efficiency of  $\hat{\theta}_2$  to  $\hat{\theta}_1$  is

$$\frac{MSE(\hat{\theta}_1)}{MSE(\hat{\theta}_2)}$$

- An estimator with a mean squared error  $\leq$  mean squared error of any estimator, for any value of  $\theta$ , is the optimal estimator of  $\theta$

# Lec20 - Central Limit Theorem, Ch7 examples

sample taken from larger population

## DEFINITION: CENTRAL LIMIT THEOREM

- If  $X_1, X_2, \dots, X_n$  is a random sample of size  $n$  taken from a population with mean  $\mu$  and finite variance  $\sigma^2$ , and  $\bar{X}$  is the sample mean, the limiting form of the distribution of

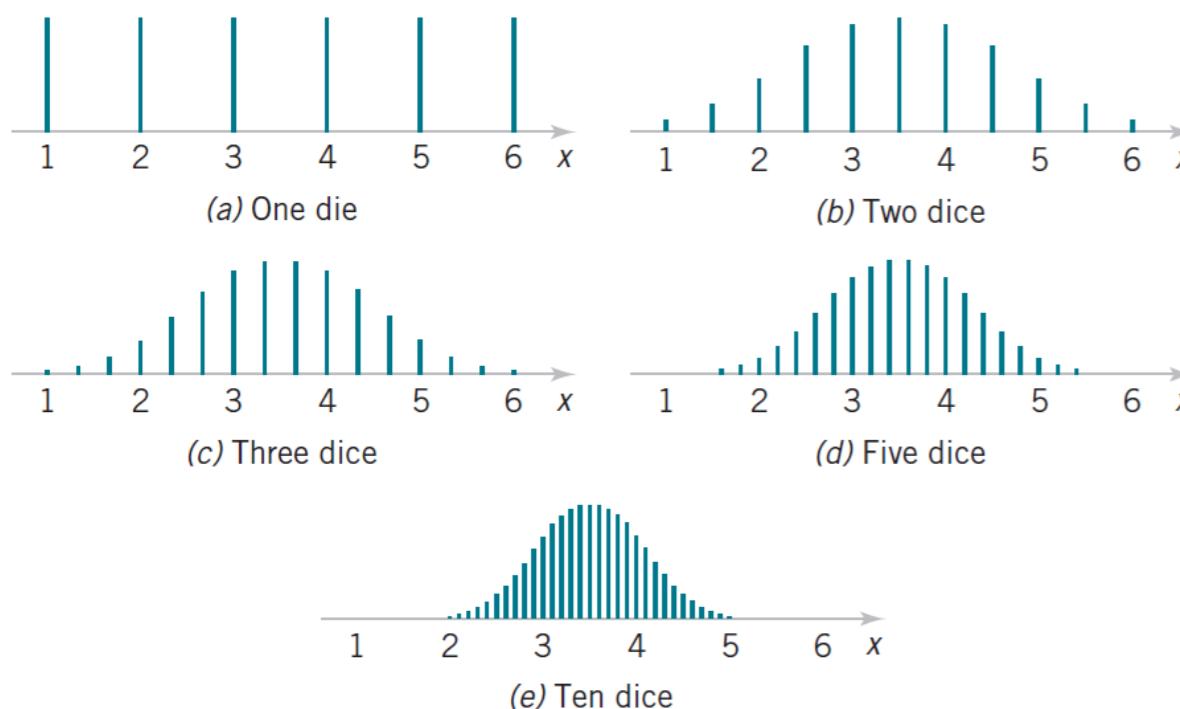
$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

as  $n \rightarrow \infty$ , is the standard normal distribution

## CONTENT

- According to the Central Limit Theorem, the distribution of the sample average  $\bar{X}$  is normal, even if the distribution of the population is non-normal
- The normal approximation for  $\bar{X}$  depends on the sample size  $n$
- If the population approximates a normal distribution, then small values of  $n$  will suffice
- If the population distribution is very non-normal, then  $n > 30$  is usually used

## DISTRIBUTION OF AVERAGE SCORES FROM THROWING DICE (P154, FIG 7.3)



## APPROXIMATE SAMPLING DIST. OF A DIFF. IN SAMPLE MEANS

- For two independent populations with  $\mu_1$  and  $\mu_2$  and variances  $\sigma_1^2$  and  $\sigma_2^2$  and if  $\bar{X}_1$  and  $\bar{X}_2$  are the sample means of two independent random samples of sizes  $n_1$  and  $n_2$  from these populations, then the sampling distribution of

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

is approximately standard normal if the conditions of the central limit theorem apply. If the two populations are normal, the sampling distribution of  $Z$  is exactly standard normal.

Ex. A synthetic fiber used in manufacturing carpet has tensile strength that is normally distributed with mean 75.5 psi and standard deviation 3.5 psi. Find the probability that a random sample of  $n = 6$  fiber specimens will have sample mean tensile strength that exceeds 75.75 psi.

$$\begin{aligned} M_{\bar{X}} &= 75.5 & P(\bar{X} \geq 75.75) \\ \sigma &= 3.5 & = 1 - P(\bar{X} \leq 75.75) \\ n &= 6 & = 1 - P(z \leq 0.175) \\ && = 1 - 0.56945 \\ && = 0.43055 \end{aligned}$$

Ex. A random sample of size  $n_1 = 16$  is selected from a normal population with a mean of 75 and a standard deviation of 8. A second random sample of size  $n_2 = 9$  is taken from another normal population with mean 70 and standard deviation 12. Let  $X_1$  and  $X_2$  be the two sample means. Find:

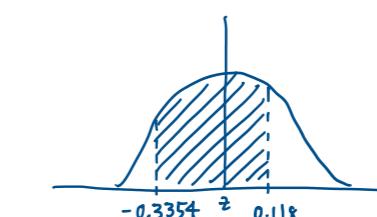
- The probability that  $X_1 - X_2$  exceeds 4
- The probability that  $3.5 \leq X_1 - X_2 \leq 5.5$

$$\begin{aligned} n_1 &= 16 & n_2 &= 9 & a) \quad \bar{X}_1 - \bar{X}_2 &> 4 : \\ \mu_1 &= 75 & \mu_2 &= 70 & z &= \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{4 - (75 - 70)}{\sqrt{\frac{8^2}{16} + \frac{12^2}{9}}} = \frac{-1}{\sqrt{\frac{64}{16} + \frac{144}{9}}} = \frac{-1}{\sqrt{4 + 16}} = \frac{-1}{\sqrt{20}} \\ \sigma_1 &= 8 & \sigma_2 &= 12 & & = -0.2236 \end{aligned}$$

$$\begin{aligned} P(\bar{X}_1 - \bar{X}_2 > 4) &= 1 - P(z \leq -0.2236) & \bar{X}_1 - \bar{X}_2 &\sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}) \\ &= 1 - 0.414 & \sim N(75 - 70, \frac{8^2}{16} + \frac{12^2}{9}) \\ &= 0.586 & \sim N(5, 20) \\ & & M &= \sigma \end{aligned}$$

$$b) \text{ prob. } 3.5 \leq \bar{X}_1 - \bar{X}_2 \leq 5.5$$

$$\begin{aligned} z &= \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \\ P\left[\frac{3.5 - 5}{\sqrt{20}} \leq z \leq \frac{5.5 - 5}{\sqrt{20}}\right] &= P[-0.3354 \leq z \leq 0.118] \end{aligned}$$



$$P(3.5 \leq \bar{X}_1 - \bar{X}_2 \leq 5.5) = P(z \leq 0.118) - P(z \leq -0.3354)$$

$$\begin{aligned} &\downarrow \\ 1 - P(z \leq 0.3354) &= 1 - 0.633 \\ &= 0.3687 \end{aligned}$$

$$P(3.5 \leq \bar{X}_1 - \bar{X}_2 \leq 5.5) = 0.5445 - 0.3687 = 0.1759$$

memory aid: eqn sheet 23

# Lec21 - Confidence intervals for a normal random variable

Thursday, November 10, 2022 4:43 PM

- Point estimates, confidence intervals - 95% and others, finding the sample size

## INTRODUCTION

- If we want to work out the average height of men on campus, do we measure all of the men on campus
- The answer is: "probably not". Instead, we take a sample and use the mean's properties to make conjectures about the true value.
- This idea is similar to the process of sampling from a production line to determine whether the average product weight is within specified bounds

## SAMPLING FROM A NORMAL DISTRIBUTION - EX 1:

- If we take a random sample of size  $n$  from a normal distribution with mean  $\mu$  and standard deviation  $\sigma$ , then we know that the distribution of the sample mean is

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

## PREVIOUS RELATIONSHIP:

- Did the process in the previous example look familiar?

Recall: if  $X \sim N(\mu, \sigma^2)$  then  
 mean, stddev  

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

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

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

## ACCESSING THE POINT ESTIMATE:

- A point estimate is very useful, but it's only a single number
- In most cases, the probability that the estimator *actually* equals the true value is equal to 0
- That is, eg., the  $P(\bar{x} = \mu) = 0$  most of the time. To assess how confident we are in our point estimate, we can construct a **confidence interval** for  $\mu$

## CONFIDENCE INTERVAL: CONSTRUCTION

- Suppose that we collect or observe a random sample,  $x_1, \dots, x_n$  from a normal population
- We estimate the population mean  $\mu$  and we know the value of the population standard deviation is  $\sigma$
- Given the sampling distribution  $\bar{X}$ , a range of possible values for  $\mu$  might be all values such that  $P(a \leq \mu \leq b) = 1 - \alpha$
- Typically, we take  $\alpha = 0.05$  and consider only intervals that are symmetric about  $\mu$
- These intervals can be written as

$$P\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

$$1 - \alpha = 0.95$$

$$\alpha = 0.05$$

$$\Rightarrow \frac{\alpha}{2} = 0.025$$

$$0.95 + 0.025 = 0.975$$

## A 95% CONFIDENCE INTERVAL

- COROLLARY:** A 95% Confidence Interval for  $\mu$  is the random interval

$$\left(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right)$$

- PROOF:** A standard normal distribution has the property

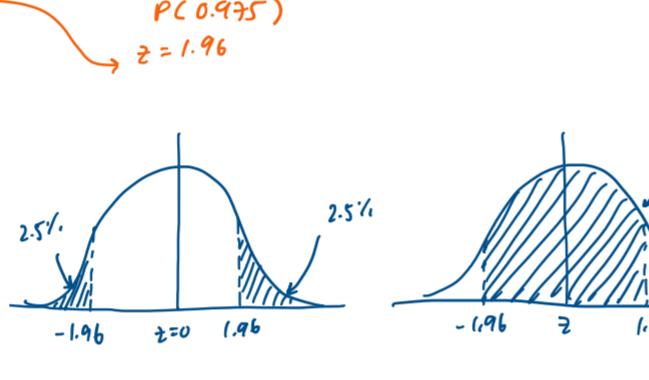
$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$$

$$P(-1.96 \leq Z \leq 1.96) = 0.95$$

$$P\left(-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right) = 0.95$$

$$P\left(-1.96 \frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq 1.96 \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$P\left(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right) = 0.95$$



There's a prob. of  $1 - \alpha$  of selecting a sample for which the CI will contain the true value of  $\mu$

- Therefore we can say with 95% confidence that  $\mu$  lies in the interval

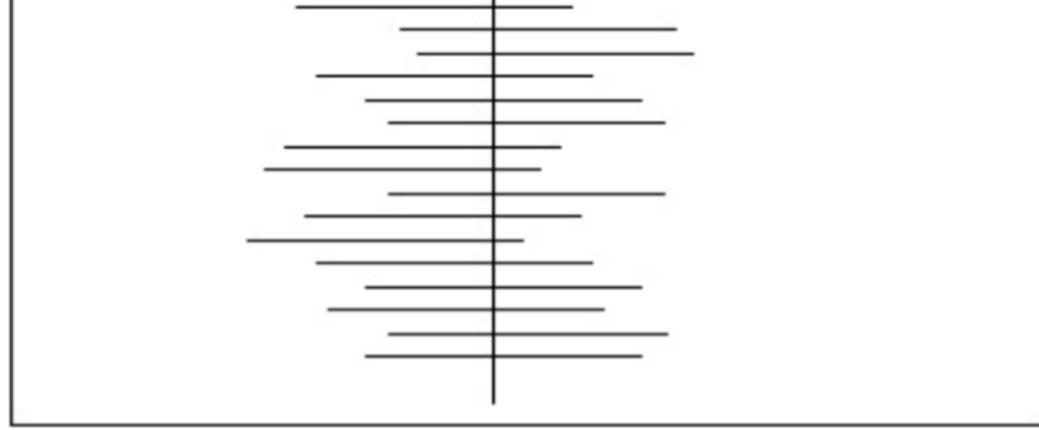
$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

## INTERPRETING A CONFIDENCE INTERVAL

- A confidence interval is a random interval
- The actual location of the interval will vary depending on the actual sample drawn since it will always be centered at  $\bar{x}$
- The value of  $\mu$  does not change
- The probability is interpreted based on what would happen if we repeatedly sampled many times
- About 95% of the confidence intervals we calculated over many, many samples would contain the true population mean  $\mu$
- For a single sample, however, we don't know if the interval does or does not contain  $\mu$

## Ex. Repeated Sampling-

- 95% CIs for  $\mu$  based on repeated sampling ( $k=20$ )



## OTHER LEVELS OF CONFIDENCE

- The confidence interval we found above is a 95% confidence interval because we started with the statement

$$P\left(-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right) = 0.95$$

- For other confidence levels, we replace the 1.96 with different quantiles of the standard normal distribution
- If we want a confidence level of  $100(1-\alpha)\%$ , we use

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

- The most common confidence levels and corresponding quantiles are:

$1 - \alpha$	0.80	0.90	0.95	0.98	0.99
$z_{\alpha/2}$	1.28	1.645	1.96	2.326	2.576

## ABOUT THE VALUE OF $Z_{\alpha/2}$

- Usually, you will be asked to calculate the 95% confidence intervals, so it is recommended that you memorize that  $Z_{\alpha/2} = Z_{0.025} = 1.96$
- However, you may be asked to calculate confidence intervals with different significance levels
- As previously mentioned, this amounts to just setting  $Z_{\alpha/2}$  equal to the appropriate value

## WIDTH OF A CONFIDENCE INTERVAL

- While maintaining a high confidence level, we would like the CIs to be as narrow as possible
- The width of the confidence interval is

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

- We can make this smaller by doing one of two things:
  - Change  $\alpha$  to make  $Z_{\alpha/2}$  smaller
  - Make  $n$  larger
- The first of these options comprises the required confidence level, since making  $Z_{\alpha/2}$  smaller will decrease  $1 - \alpha$

- Note:  $\frac{Z_{\alpha/2} \sigma}{\sqrt{n}}$  is called the Margin of Error

## CHOOSING THE SAMPLE SIZE

- Suppose that we fix the confidence level
- Then the only thing we can alter is the sample size
- If we specify that we would like the width of the interval to be equal to some value, say  $q$ , then we need:

$$2 \times z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq q$$

$$\frac{1}{\sqrt{n}} \leq \frac{q}{2z_{\alpha/2}\sigma}$$

$$\sqrt{n} \geq 2 \times z_{\alpha/2} \frac{\sigma}{q}$$

$$n \geq \left(2 \times z_{\alpha/2} \frac{\sigma}{q}\right)^2$$

Ex. Suppose that the height of men on campus is normally distributed with mean 1.78m and standard deviation 0.0m. What is the probability that the mean of a sample of 25 men will be less than 1.8m?

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

$$P(\bar{x} < 1.8) = P\left(z < \frac{1.8 - 1.78}{\frac{0.09}{\sqrt{25}}}\right) = P(z < 1.11) = 0.8665$$

Ex. Suppose that we take a sample of 25 men on campus and the average height of men in this sample is 1.8m. Given that the standard deviation of the population is 0.09m, calculate a 95% confidence interval for  $\mu$ .

$$n=25 \quad \mu=1.8 \quad \sigma=0.09$$

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}} = 1.8 \pm 1.96 \left(\frac{0.09}{\sqrt{25}}\right) = 1.8 \pm 0.035$$

$$= (1.765, 1.835)$$

Ex. It has been established over a long period of time that the breaking strength of a type of rope is normally distributed with standard deviation 2.4kg. Given  $n=25$  and  $\bar{x}=39$ kg, construct a 95% confidence interval for  $\mu$ .

95% confidence interval for  $\mu$  based on  $\bar{x}$ :

$$39 \pm 1.96 \left(\frac{2.4}{\sqrt{25}}\right) = 39 \pm 0.94 = (38.06, 39.94)$$

→ 95% confident that based on the sample mean, that the population mean lies b/w 38.06 kg and 39.94 kg

→ this is one of the two interpretations of the 95% confidence interval

Ex. A 90% confidence interval for  $\mu$

Suppose that for our sample of 25 men on campus, we're asked to calculate a 90% confidence interval for  $\mu$

$$\bar{x} \pm 1.645 \frac{\sigma}{\sqrt{n}} = 1.8 \pm 1.645 \frac{0.09}{\sqrt{25}} = 1.8 \pm 0.0030 = (1.770, 1.830)$$

Ex. A population is normally distributed with a mean of 5.6. How many observations do we need if we want to construct a 95% CI with width equal to 2?

$$n \geq \left(2 \times \frac{Z_{\alpha/2} \frac{\sigma}{2}}{2}\right)^2$$

$$n \geq \left(2 \times 1.96 \left(\frac{1.6}{2}\right)\right)^2$$

$$n \geq 120.4726$$

When choosing or finding sample size, always round up!

Ex. A sample of size 50 is drawn from a population with= 10. The sample mean is 20.26. Calculate and interpret a 95% confidence interval for  $\mu$ .

$$\bar{x} \pm \frac{Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}}{2} = 20.26 \pm 1.96 \frac{10}{\sqrt{50}} = 20.26 \pm 2.77 = (17.49, 22.03)$$

Ex. A population is normally distributed with=5.6, but  $\mu$  is unknown. A random sample of 10 individuals is drawn, and  $\bar{x}=25.4$ . Calculate a 90% confidence interval for  $\mu$ .

$$\bar{x} \pm \frac{Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}}{2} = 25.4 \pm 1.645 \frac{5.6}{\sqrt{10}} = 25.4 \pm 2.913 = (22.487, 28.313)$$

Ex. Suppose a person suspected Heinz was systematically under filling their ketchup bottles, obtained a random sample of 40 bottles, and found a sample mean of 496.1 grams. Assuming the population standard deviation is 8.0 grams, calculate a 95% confidence interval for the population mean weight in ketchup bottles of this size.

$$\bar{x} \pm \frac{Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}}{2} = 496.1 \pm 1.96 \frac{8}{\sqrt{40}} = 496.1 \pm 2.48 = (493.62, 498.58)$$

Ex. What is the minimum sample required to calculate a 95% confidence interval of width 1.5 given the standard deviation given in Exercise 3.

$$n \geq \left(2 \times \frac{Z_{\alpha/2} \frac{\sigma}{2}}{1.5}\right)^2 \Rightarrow n \geq \left(2 \times 1.96 \frac{8}{15}\right)^2 = 437.09$$

$$\therefore n=438$$

# Lec22 - Confidence Interval for the mean, variance unknown, confidence interval for a population proportion

Thursday, November 10, 2022 7:07 PM

## 1-SIDED CONFIDENCE INTERVAL

- Replace  $z_{\alpha/2}$  by  $z_{\alpha}$

Upper bound for  $\mu$  is

$$\mu \leq \bar{X} + z_{\alpha} \sigma / \sqrt{n}$$

Lower bound for  $\mu$  is

$$\bar{X} - z_{\alpha} \sigma / \sqrt{n} \leq \mu$$

## LARGE SAMPLE CONFIDENCE INTERVAL FOR $\mu$

- When  $n$  is large

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

Can be replaced by

$$Z = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

Hence

$$\bar{X} - z_{\alpha/2} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{s}{\sqrt{n}}$$

Is a large sample confidence interval for  $\mu$

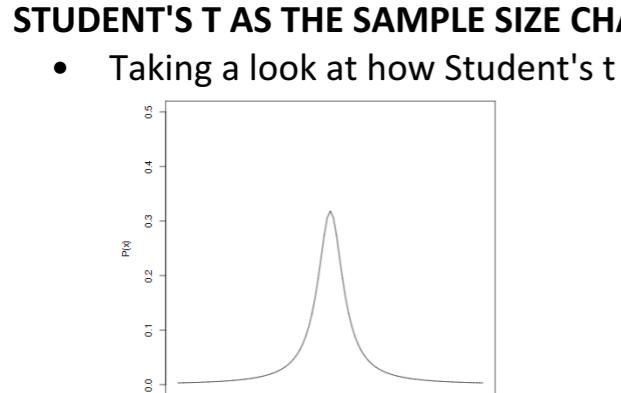
## RECAP:

- Last lecture, we learned how to calculate a confidence interval when the standard deviation is known
- However, we may come across situations where the standard deviation is unknown. We will have to estimate it when computing a confidence interval, but more often, we'll take the sample standard deviation as our estimate
- However, we also need to make one other adjustment to our confidence interval:

## STUDENT'S T-DISTRIBUTION: BACKGROUND

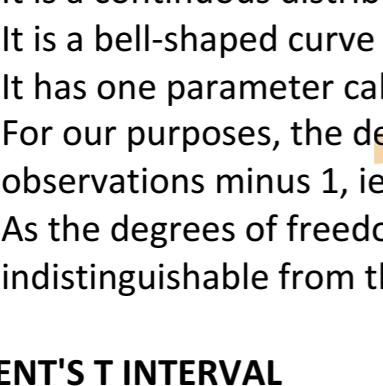
- First published by WS Gosset in 1908
- Gosset was an Oxford-trained chemist who worked in the Guinness brewery in Dublin, Ireland
- He wasn't allowed to publish his work under his own name, so he used the pseudo name 'Student'
- The t-distribution looks more and more like the normal distribution as the sample size gets larger
- For smaller sample sizes, it has relatively thick tails
- It is the backbone of the 't-test', which is one of the most frequently used statistical tests

## STUDENT'S T VS NORMAL DISTRIBUTION



## STUDENT'S T AS THE SAMPLE SIZE CHANGES

- Taking a look at how Student's t distribution changes with  $n$



## PROPERTIES:

- It is a continuous distribution defined on the whole real line
- It is a bell-shaped curve symmetric about 0
- It has one parameter called the **degrees of freedom**
- For our purposes, the degrees of freedom is equal to the number of observations minus 1, ie.  $V=n-1$
- As the degrees of freedom gets large, the Student's t becomes indistinguishable from the standard normal

## STUDENT'S T INTERVAL

- Provided that the underlying population is normal, Gosset (Student) showed

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

Has a t distribution with  $n-1$  degrees of freedom

- We can derive an interval for the mean,  $\mu$  by replacing the  $z_{\alpha/2}$  by the corresponding critical values of the  $t_{n-1}$  distribution

## A STUDENT'S T CONFIDENCE INTERVAL

- The  $100(1-\alpha)\%$  Student's t confidence interval for the mean  $\mu$  of a normal population based on a sample size  $n$  is

$$\left( \bar{X} - t_{n-1, \alpha/2} \frac{s}{\sqrt{n}}, \bar{X} + t_{n-1, \alpha/2} \frac{s}{\sqrt{n}} \right)$$

Has a t distribution with  $n-1$  degrees of freedom

- Therefore, given a sample mean  $\bar{X}$  and a sample standard deviation  $s$ , one can construct a  $100(1-\alpha)\%$  confidence interval for  $\mu$ , as follows:

$$\bar{X} \pm t_{n-1, \alpha/2} \frac{s}{\sqrt{n}}$$

- Here,  $t_{\alpha/2}$  is the value of the t-distribution corresponding to a given significance level  $\alpha$  and a given degrees of freedom,  $V$ .

## DISCUSSION

- In general, we will always use a Student's t confidence interval when  $\sigma$  is unknown
- When  $n$  is large, a student's t distribution gives a very good approximation to the normal distribution
- For both of the confidence intervals discussed so far, we have assumed the data is normally distributed

## PROPORTIONS

- We have discussed proportions before
- In general, given a sample size of size  $n$ , the proportion of times  $X$  occurs is given by

$$\hat{p} = \frac{X}{n}$$

- We consider  $\hat{p}$  to be a point estimate of a population modelled by parameter  $p$
- Therefore, we can also construct a confidence interval for  $p$  using  $\hat{p}$

## A 95% CONFIDENCE INTERVAL FOR A PROPORTION

- In order to construct confidence intervals for the parameter  $p$ , we need to know something about the sampling distribution of  $\hat{p}$

Recall: the  $E[\hat{p}] = p$

- So we need to find the standard deviation of  $\hat{p}$ . It follows that

$$\text{Var}[\hat{p}] = \text{Var}\left[\frac{X}{n}\right] = \frac{1}{n^2} \text{Var}[X] = \frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n}$$

- Therefore:

$$SD[\hat{p}] = \sqrt{\frac{p(1-p)}{n}}$$

## ADJUSTING FOR P

- We can see that the  $SD[\hat{p}]$  uses the true parameter  $p$
- Of course, in practice, we won't know the true value
- So we can write

$$SD[\hat{p}] = \sqrt{\frac{p(1-p)}{n}}$$

As long as

$$n\hat{p} \geq 15 \quad \text{and} \quad n(1-\hat{p}) \geq 15$$

It follows that a confidence interval for  $\hat{p}$  is given by

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

Where  $\hat{q} = 1 - \hat{p}$

## CHOICE OF SAMPLE SIZE FOR A PROPORTION

$$n = \left[ \frac{z_{\alpha/2}}{E} \right]^2 p(1-p)$$

- Where  $E$  is the error

## Student's t

Ex. Using the student's t table (p745)

- What is  $P(t > 1.771)$  when  $n = 14$ ?

$$1. V = n-1 \\ V = 14-1 = 13$$



- What is  $P(t < 1.711)$  when  $n = 25$ ?

$$2. V = 25-1 = 24 \\ P(t > 1.711) = 0.05$$



- What is  $P(t > 2.06)$  when  $n = 17$ ?

$$3. P(t > 2.06) \\ n=17 \\ V=17-1=16$$



- $P(t > 2.06) = 0.05$

$$(-0.05 = 0.95) \\ P(t > 1.711) = 0.95$$

- $V=17-1=16$

$$ans = 0.25 < \alpha < 0.5$$

- $t_{0.05} = 2.06$

# Lec23 - hypothesis tests for a normal random variable

Saturday, November 12, 2022 1:30 AM

- A general introduction to hypothesis testing
- An outline of how to conduct a hypothesis test

## INTRO:

- We often have some theory that we want to examine based on a sample of data
- That theory may be expressed in terms of a parameter of the underlying population distribution
- We can use a sample to determine if some hypothesized value is plausible given the observed data
- We generally do this by looking for evidence that the hypothesized value is unreasonable
- If we don't find any such evidence, then we can conclude that the hypothesized value is plausible

## HYPOTHESIS TESTING

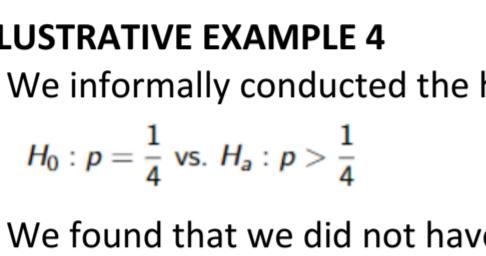
- Generally speaking, hypothesis testing is the practice of translating a scientific question into a hypothesis about the value of a population parameter
- Examples of questions that hypothesis testing may help answer:
  - Do more than half the adults in a certain area favour the legalization of marijuana?
  - Is the mean highway fuel consumption of a new model of car different from what the manufacturer claims?
  - Will a coin really land heads up 50% of the time?

## AN ILLUSTRATIVE EXAMPLE 1

- Jay claims he can guess the suit of a randomly selected playing card more than 25% of the time (on avg)
- We perform an experiment where we make Jay guess the suit of a randomly selected card 100 times
- If Jay was just guessing, on average he'd guess the right suit 25% of the time
- In this experiment, he guesses the right suit 28 times
- Does this result give evidence that Jay will correctly guess the correct suit more than 25% of the time?

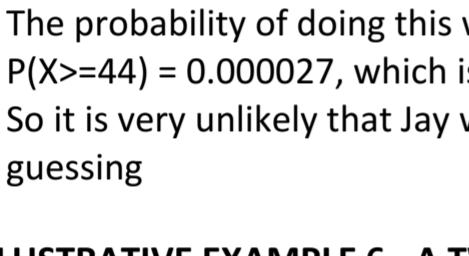
## AN ILLUSTRATIVE EXAMPLE 2

- Given a standard deck of cards, the number of correct guesses will follow a binomial distribution with parameters  $n=100$  and  $p=1/4$
- So the distribution of the number of correct guesses for 100 randomly selected playing cards looks like:



## AN ILLUSTRATIVE EXAMPLE 3

- What is the probability of getting 28 or more correct when guessing randomly?
- Given  $X \sim \text{Bin}(100, 0.25)$ , the probability that Jay would do as well as he did or even better is  $P(X \geq 28) = 0.278$



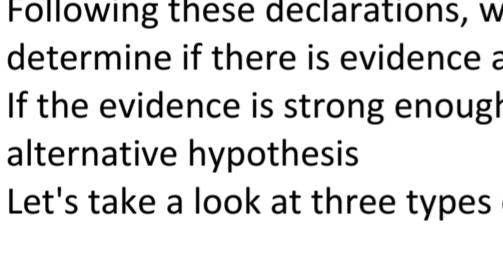
- Since the probability isn't small, we may be inclined to say that it's not unlikely to get this many correct due to chance

## AN ILLUSTRATIVE EXAMPLE 4

- We informally conducted the hypothesis test:  
 $H_0: p = \frac{1}{4}$  vs.  $H_a: p > \frac{1}{4}$
- We found that we did not have strong evidence against this null hypothesis
- The **null hypothesis ( $H_0$ )** is that Jay does just the same as someone who is randomly guessing (on average)
- The **alternative hypothesis ( $H_a$ )** is that Jay does better than someone who is simply guessing (on average)

## AN ILLUSTRATIVE EXAMPLE 5 - A TWIST

- Suppose that Jay had correctly guessed the suit 44 times



- The probability of doing this well or better is equal to  $P(X \geq 44) = 0.000027$ , which is about 1 in 37k
- So it is very unlikely that Jay would do this well or better if he was randomly guessing

## AN ILLUSTRATIVE EXAMPLE 6 - A TWIST

- Since this probability is very low, one or two things occurred
  - Jay is simply guessing with probability of success equal to 0.25 and we witnessed a very unusual event due to chance
  - Jay is truly guessing the suit more often than 0.25 of the time, on avg

- TLDR;
  - The null hypothesis is true and we witnessed a very unusual event
  - Or the null hypothesis is false

## DISCUSSION

- This example illustrated the need for a formal way to assess the value of our point estimate
- Last lecture, we learned to construct confidence intervals to do this
- Hypothesis is closely related
- We will learn how to test hypotheses when the data is from a normal dist.

## HYPOTHESIS TESTING: FORMAL DEFINITION

- We formulate a research question of interest and turn it into the appropriate null and alternative hypotheses
  - The **null hypothesis ( $H_0$ )** is the hypothesis of no effect or no difference ( $H_0$  is sometimes referred to as the status quo hypothesis)
  - The **alternative hypothesis ( $H_a$ )** is the hypothesis the researcher is *hoping* to show ( $H_a$  is sometimes referred to as the research hypothesis)
- Following these declarations, we calculate an appropriate test statistic and determine if there is evidence against  $H_0$
- If the evidence is strong enough, we reject the null hypothesis in favour of the alternative hypothesis
- Let's take a look at three types of hypothesis tests

## STATISTICAL HYPOTHESES

- TWO TAILED TEST:**  $H_0$  = parameter equals value,  $H_a$  = it doesn't  
 $H_0: \mu = \mu_0$  vs.  $H_a: \mu \neq \mu_0$
- UPPER TAILED TEST:**  $H_0$  = parameter equals value,  $H_a$  = parameter is higher  
 $H_0: \mu = \mu_0$  vs.  $H_a: \mu > \mu_0$
- LOWER TAILED TEST:**  $H_0$  = parameter equals value,  $H_a$  = parameter is lower  
 $H_0: \mu = \mu_0$  vs.  $H_a: \mu < \mu_0$

## EXAMPLES OF ALTERNATIVE HYPOTHESIS

- A researcher wishes to test if the mean fat content (in grams) of a type of hotdog is different from the desired 25g  
 $H_0: \mu_{\text{fat}} = 25$  vs.  $H_a: \mu_{\text{fat}} \neq 25$
- A researcher wishes to test the mean fat content of a type of hotdog exceeds 25g  
 $H_0: \mu_{\text{fat}} = 25$  vs.  $H_a: \mu_{\text{fat}} > 25$
- A researcher wants to test if the mean fat content of a type of hotdog < 25g  
 $H_0: \mu_{\text{fat}} = 25$  vs.  $H_a: \mu_{\text{fat}} < 25$

## TEST STATISTICS

- Once the hypothesis has been decided, a sample is taken
- The statistician then calculates a **test statistic**
- The value of the test statistic is compared to the sampling distribution that we would get if the null hypothesis is true
- If the observed test statistic seems to be an unusual value for that sampling distribution, then we take this as evidence against  $H_0$
- To determine how unusual the value is, we need to determine 3 terms:
  - Rejection region, significance level, and critical value

## REJECTION REGION

- The **rejection region** is the set of all values of the **test statistic** for which we **reject  $H_0$** , meaning the rejection region is dependent on  $H_a$
- For a TWO TAILED TEST, the rejection region will usually be in two parts:
  - Positive direction  $\oplus$
  - Negative direction  $\ominus$

- For an UPPER TAILED TEST, the rejection region will be in the pos direction only  $\oplus$
- For a LOWER TAILED TEST, the rejection region will be in the neg direction only  $\ominus$

- The **rejection region** is defined by the **significance level** and corresponding critical value

## SIGNIFICANCE LEVEL ( $\alpha$ )

- The **significance level** is the maximum probability of a type I error that we allow
- Generally, this is set to 5%, but could be 1% or 10% in different settings
- Again, the rejection region depends on the significance level we choose
- A higher significance level means a larger rejection region

## CRITICAL VALUES

- The boundary between the rejection and non-rejection region is called the **critical value**
- The **critical value** is chosen such that if the null hypothesis is true, the probability of the **test statistic** being in the rejection region is equal to  $\alpha$
- For a TWO TAILED test, we look for  $\alpha/2$  in each tail
- For a ONE TAILED test, we take the full value of  $\alpha$  in each tail of interest

## CONCLUSIONS FROM HYPOTHESIS TESTING

- If the **test statistic** is in the **rejection region** then we **reject  $H_0$** 
  - In this case, there is strong evidence against  $H_0$
- If the **test statistic** is not in the **rejection region** then we **fail to reject  $H_0$** 
  - In this case, there is no evidence against  $H_0$
- It is very important to note that **WE NEVER ACCEPT  $H_0$**
- Even if we fail to reject it, that doesn't mean it's true
- We're only looking for evidence against  $H_0$ , not in favour of  $H_a$

## SUMMARY STEPS IN HYPOTHESIS TESTING

- Formulate  $H_0$  and  $H_a$
- Decide on the value of  $\alpha$  and find the **critical value** from the tables
- Compute the value of the **test statistic**
- Make a decision about  $H_0$
- State your conclusions in terms of the original question

## TEST FOR A POPULATION MEAN WHEN $\sigma$ IS KNOWN

- Suppose that we're interested in a continuous measurement and we wish to test a hypothesis about a population mean
- We construct a test of the following form:

$$H_0: \mu = \mu_0 \quad \text{vs.} \quad H_a: \mu > \mu_0$$

- For now, assume we know the value of  $\sigma$

$$\text{For } H_0: \mu = \mu_0$$

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$$

## DECISION RULES FOR HYPOTHESIS TESTS WHEN $\sigma$ IS KNOWN

- For all three hypothesis tests, we begin calculating the test statistic

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

- UPPER TAILED TEST:** Reject  $H_0$  if  $z > z_\alpha$

$$H_0: \mu = \mu_0 \quad \text{vs.} \quad H_a: \mu > \mu_0$$

- LOWER TAILED TEST:** Reject  $H_0$  if  $z < -z_\alpha$

$$H_0: \mu = \mu_0 \quad \text{vs.} \quad H_a: \mu < \mu_0$$

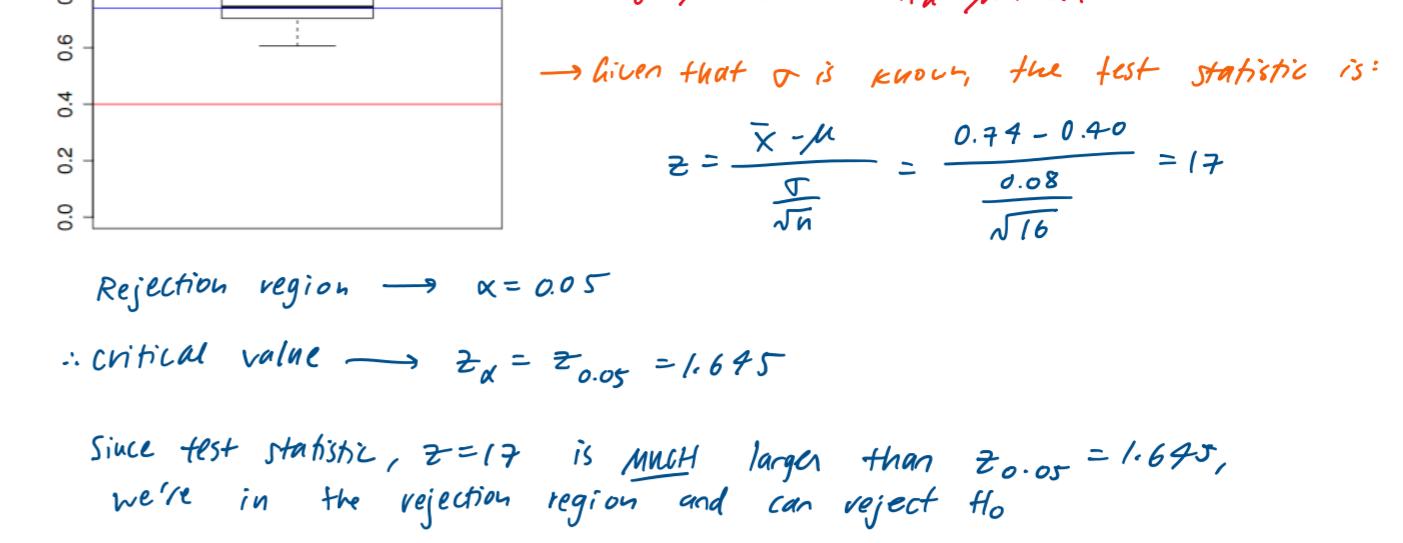
- TWO TAILED TEST:** Reject  $H_0$  if  $|z| > z_{\alpha/2}$

$$H_0: \mu = \mu_0 \quad \text{vs.} \quad H_a: \mu \neq \mu_0$$

## Comparing test statistic to critical value:

### Ex. Hypothesis test about $\mu$

- A tuna supplier collects a random sample of 16 pieces of tuna. The average mercury content of the sample is 0.74 ppm. Does this yield strong evidence that the true mean mercury is greater than 0.40 ppm, when  $\sigma = 0.08$  and  $\alpha=0.05$ ?



Rejection region  $\rightarrow \alpha = 0.05$

critical value  $\rightarrow z_\alpha = z_{0.05} = 1.645$

Since test statistic,  $z=17$  is MUCH larger than  $z_{0.05} = 1.645$ , we're in the rejection region and can reject  $H_0$

→ After deciding the hypotheses & calculating the test statistic, where did the critical value come from?

→ There's very strong evidence against  $H_0: \mu = 0.40$ , meaning that the true mean of mercury is greater than 0.40 ppm

→ we can also use the p-value approach to compare

The National Funeral Director's Association reported that the average cost of a full-service funeral in 2005 was \$6,500. A random sample of 36 funeral homes in 2006 gave  $\bar{x} = 6,819$  and  $\sigma = 1,265$ . Test the claim that the true mean is not equal to 6,500.

$$H_0: \mu = 6,500 \quad \text{vs.} \quad H_a: \mu \neq 6,500$$

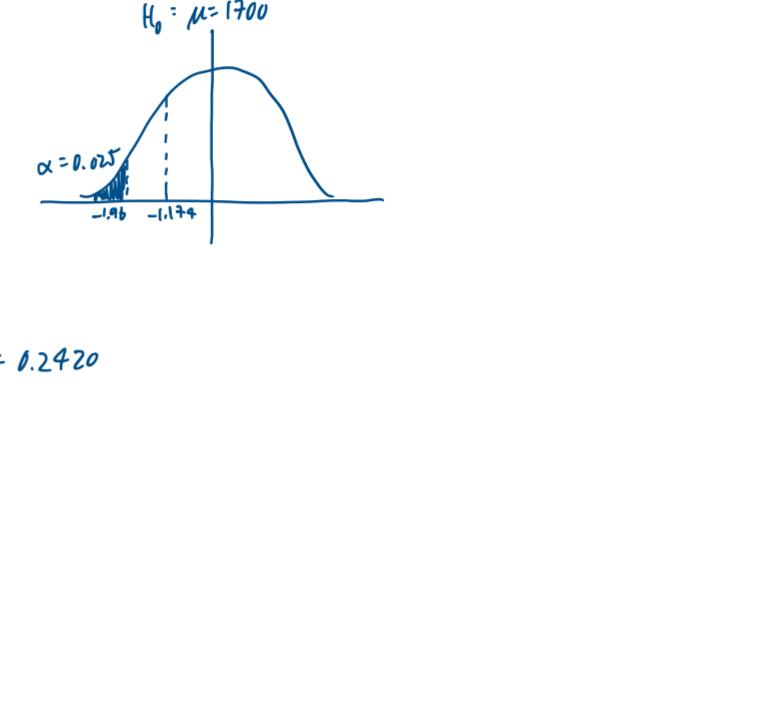
$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{6,819 - 6,500}{1,265/\sqrt{36}} = 1.513$$

Critical value is  $z_{\alpha/2} = z_{0.025} = 1.96$

∴ because  $|1.513| < 1.96$ , do not reject  $H_0$

The p-value for this is equal to  $2(0.065) = 0.1310$

∴ because  $0.1310 > 0.05$ , do not reject  $H_0$



Suppose the mean student debt for all 2nd year students at the University of Western Ontario (UWO) is \$1700. Suppose, in a sample of 15 UWO 2nd year students that  $\bar{x} = \$1600$  and that  $\sigma = \$330$ .

Based on this sample, is there evidence that the amount of debt for a 2nd student is different from \$1700? Is there evidence that it is less than \$1700?

$$H_0: \mu = 1700 \quad \text{vs.} \quad H_a: \mu \neq 1700$$

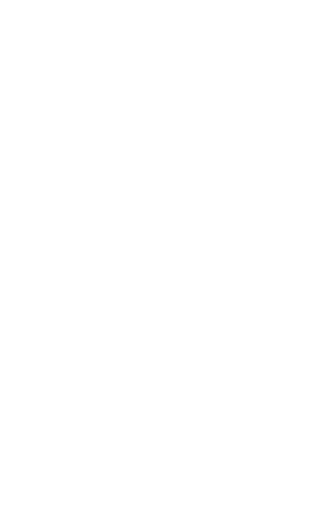
$$t = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{1600 - 1700}{330/\sqrt{15}} = -1.134$$

The critical value is  $z_{\alpha/2} = z_{0.025} = 1.96$

∴ because  $|-1.134| < 1.96$ , do not reject  $H_0$

The p-value for this is equal to  $2(0.1210) = 0.2420$

∴ because  $0.2420 > 0.05$ , do not reject  $H_0$



# Lec24 - Hypothesis tests a for a student-t random variable

Saturday, November 26, 2022 5:36 PM

- P-value (sec 9-1.4)
- T-test (sec 9.3)

## INTRODUCTION:

- Last lecture, we learned how to conduct a hypothesis test
- The underlying assumptions of that test were normality and that  $\sigma$  was known
- Of course, in practice, we usually don't know the true value of  $\sigma$
- We will introduce a hypothesis test that does not require  $\sigma$ , but first, let's look at another way to make a conclusion about a given hypothesis test

## P-VALUE

- Formally, the p-value is the probability of getting a test statistic more extreme than the one observed, assuming the null hypothesis is true
- So we can think of it as a measure of the strength of the evidence against them null hypothesis
- To calculate a p-value we use the techniques we've learned so far to find area under the normal (and Student's t) curves
- The regions under the curve that we're looking for also depends on the form of the alternative hypothesis
- Let's take a look at a few examples

## HYPOTHESIS TESTING WHEN $\sigma$ IS UNKNOWN

- To construct confidence intervals when  $\sigma$  is unknown, we calculate a test statistic using the sample standard deviation
- The rejection region, or significance level and decision rules do not change
- The critical value and test statistic do change
- The critical value is now found using Student's t distribution

## STUDENT'S T TEST

- If  $\sigma$  is unknown to the results of WS Gosset (Student)
- When conducting a hypothesis test, if the  $\sigma$  is unknown, population is normal, and  $u=0$ , then
- $$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$
- So we will now use the t-table to find the critical value and p-value
- Note: to use the Student's t test, we assume that the underlying population is normally distributed

## DECISION RULES FOR HYPOTHESIS TESTS BASED ON STUDENT'S T

- For all three hypothesis tests we first calculate the test statistic
- $$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$
- TWO TAILED TEST:** Reject  $H_0$  if  $|t| > t_{n-1, \alpha}$   
 $H_0: \mu = \mu_0$  versus  $H_a: \mu \neq \mu_0$
- UPPER TAILED TEST:** Reject  $H_0$  if  $t > t_{n-1, \alpha}$   
 $H_0: \mu = \mu_0$  versus  $H_a: \mu < \mu_0$
- LOWER TAILED TEST:** Reject  $H_0$  if  $t < -t_{n-1, \alpha}$   
 $H_0: \mu = \mu_0$  versus  $H_a: \mu \neq \mu_0$

## DISCUSSION

- That brings us to the end of our discussion on how to conduct hypothesis tests for  $\mu$
- Next, we will discuss hypothesis tests for the proportion
- Finally, we will discuss the errors associated with hypothesis testing

Ex. A call centre claims that the average wait time for their customers is no more than 2.00 minutes. You feel that the wait times are longer than 2.00 minutes on average. To investigate, you make 70 calls at randomly selected times and find an average wait time of 2.438 minutes. Does this provide strong evidence that the true mean wait time is greater than two minutes? Assume that the population is normally distributed with  $\sigma = 1.2$  and use a significance level of  $\alpha = 0.05$ . Formally, we're testing:

$$H_0: \mu = 2 \text{ vs. } H_a: \mu > 2$$

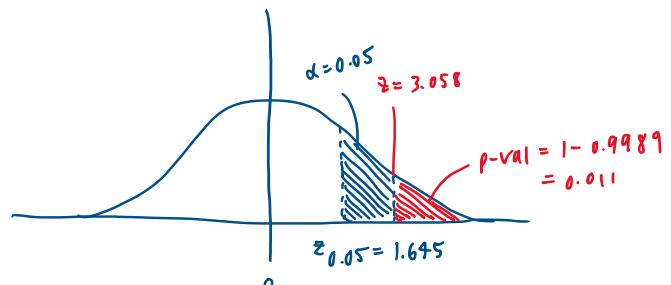
The test statistic is

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{2.438 - 2.00}{1.2/\sqrt{70}} = 3.0538$$

What is the p-value?

If  $p\text{-value} < \alpha \Rightarrow$  reject  $H_0$  otherwise do not reject.

Since  $0.001 < (\alpha = 0.05)$  we can reject  $H_0$  and conclude that there is evidence that the average wait time is greater than 2 mins.



Ex. Suppose it is known that the average starting wage for a recent MBA graduate in Canada is \$71k. Do graduates from the MBA warehouse have a different mean starting salary? A random sample of 50 recent MBA grads from the school had a mean starting wage of \$66.5k. Carry out an appropriate hypothesis test. Assume  $\sigma = \$12k$ ,  $\alpha = 0.05$ . Formally, we're testing:

$$H_0: \mu = 71,000 \text{ vs. } H_a: \mu \neq 71,000$$

The test statistic is given by

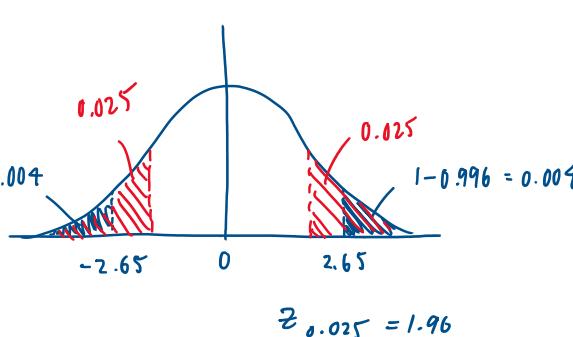
$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{66500 - 71000}{12000/\sqrt{50}} = 2.6517$$

What is the p-value?

$\therefore p\text{-value} = 0.004 + 0.004 = 0.008$

Leave out: 95% CI =  $66500 \pm 1.96 \left( \frac{12000}{\sqrt{50}} \right)$

because  $p\text{-value} < (\alpha = 0.05) \Rightarrow$  reject  $H_0$   
or since  $|z| = 2.65 > z_{0.025} = 1.96 \Rightarrow$  reject  $H_0$



## Students t-test: Ex.

A manufacturer produces bolts that are supposed to have a weight of 2.50 grams. A random sample of 20 bolts yielded a sample mean of 2.53 grams, with a standard deviation of 0.04 grams. Test whether the population mean weight is different from 2.50, using a 1% significance level. Formally, we're testing:

$$H_0: \mu = 2.50 \text{ vs. } H_a: \mu \neq 2.50$$

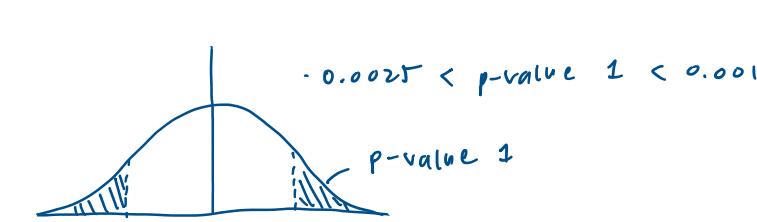
The test statistic is given by

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{2.53 - 2.50}{0.04/\sqrt{20}} = 3.354$$

Let's take a look at how to make conclusions based off this test.

Since  $|t| = 3.354 > 2.861 \Rightarrow$  reject  $H_0$

## P-VALUE APPROACH:



$$\begin{aligned} p\text{-value} &= p\text{-value}_1 + p\text{-value}_2 \\ &= 2(0.0025 < p\text{-value} < 0.001) \\ &= 0.005 < p\text{-value} < 0.002 \end{aligned}$$

We have  $\alpha = 0.1 \therefore$  since  $p\text{-value}$  is less than  $\alpha$   $\Rightarrow$  reject  $H_0$

## Students t-test: Ex.

Experimental determinations of spring constants were made for three spring types; Type 1 (4 inch, tsc=1.86), Type 2 (6 inch, tsc=2.63) and Type 3 (6inch, tsc=2.63). Note that tsc means "theoretical spring constant". The following summary statistics were obtained. Six springs were used in each case and you may assume normality.

Statistic	Type 1	Type 2	Type 3
$X_{\bar{}}\bar{}$	2.03	2.55	2.34
S	0.068	0.084	0.064

We can test whether or not the long run mean for Type 1 springs equals the tsc of 1.86.

Formally, we're testing

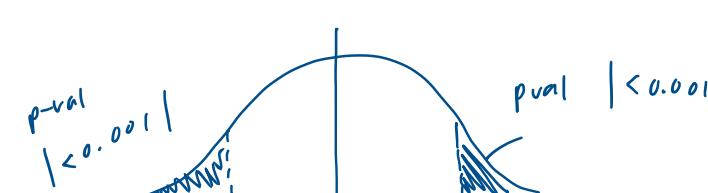
$$H_0: \mu_1 = 1.86 \text{ vs. } H_1: \mu_1 \neq 1.86 \text{ at } \alpha = 0.05$$

The test statistic is given by

$$t = \frac{\bar{X}_1 - 1.86}{S_1/\sqrt{n_1}} = \frac{2.03 - 1.86}{0.068/\sqrt{6}} = 6.12.$$

Consulting the t-distribution tables, we see that the critical value for  $\alpha = 0.05$  and  $v = n-1 = 5$  is  $t_{\alpha/2, 5} = t_{0.025, 5} = 2.571$

What is the conclusion of this test? Test whether or not the long-run mean for Type 1 springs equals the tsc of 1.86



$\therefore p\text{-value} < 0.002$  and we can reject  $H_0$

- ① The National Funeral Director's Association reported that the average cost of a full-service funeral in 2005 was \$6,500. A random sample of 36 funeral homes in 2006 gave  $\bar{x} = 6.819$  and  $s = 1.265$ . Test the claim that the true mean is not equal to 6.5.

$$H_0: \mu = 6.5 \text{ vs. } H_a: \mu \neq 6.5 \quad n=36$$

$$t = \frac{\bar{X}_0 - \mu_0}{S/\sqrt{n}} = \frac{6.819 - 6.5}{1.265/\sqrt{36}} = 1.513 = \text{test statistic}$$

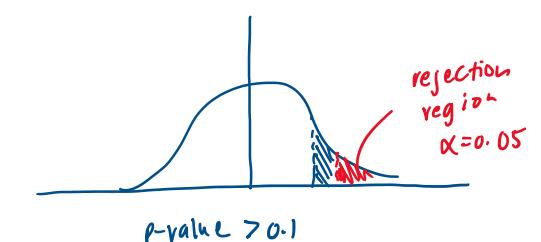
- critical value: presume 5% significance level

$$t_{\alpha/2, n-1} = t_{0.025, 35} = 2.042 \quad (\text{take lower v-value when between } n \text{ values})$$

$\therefore |1.513| < 2.042 \Rightarrow$  do not reject  $H_0$

$$\begin{aligned} \text{p-value for this test } &= 2(0.1) > \text{p-value} > 0.05 \\ &0.20 > \text{p-value} > 0.1 \end{aligned}$$

We have  $\alpha = 0.05$ ,  $\text{p-val} > 0.1 \Rightarrow \therefore$  do not reject  $H_0$



- ② Reconsider Students t test: Example 2. Test the appropriate two-sided null hypotheses for the second and third types of springs. Compare the results to the confidence intervals constructed last class.

$$H_0: \mu_2 = 1.86 \text{ vs. } H_a: \mu_2 \neq 1.86 \quad t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

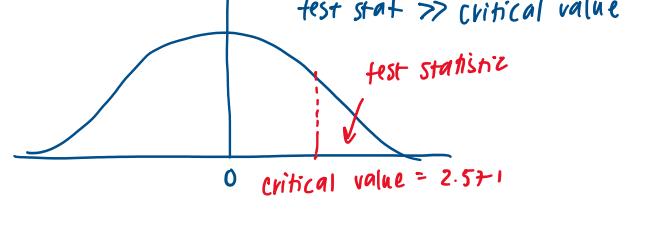
$$\alpha = 0.05$$

$$\text{Test statistic: } t = \frac{2.55 - 1.86}{0.025} = 20.12$$

Reject the null hypothesis  
There is significant evidence that the mean of the second spring type is different from 1.86

$$\begin{aligned} \text{Critical value for } \alpha = 0.05, n=6 &= 2.571 \\ v = n-1 = 5 & \\ \text{Critical value} &= 2.571 \end{aligned}$$

- compare to previous class on CIS — the 95% confidence intervals do not include the tested value



- ③ Suppose the mean student debt for all 2nd year students at the University of Western Ontario (UWO) is \$1700. Suppose, in a sample of 15 UWO 2nd year students that  $\bar{x} = \$1600$  and  $s = \$330$ . Based on this sample, is there evidence that the amount of debt for a 2nd student is different from \$1700? Is there evidence that it is less than \$1700?

$$H_0: \mu = 1700 \text{ vs. } H_a: \mu \neq 1700 \quad \text{Test statistic}$$

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{1600 - 1700}{330/\sqrt{15}} = -1.174$$

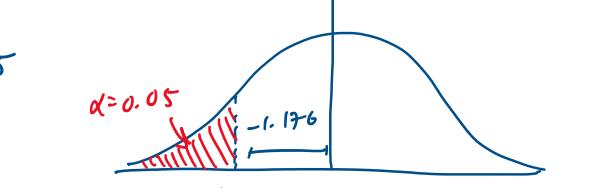
$$\text{Critical value: } t_{\alpha/2, n-1} = t_{0.025, 14} = 2.145$$

$|-1.174| < 2.145 \Rightarrow$  Do not reject  $H_0$

$$H_0: \mu = 1700, H_a: \mu < 1700$$

$$t = -1.174 \quad \text{Reject if } t < -t_{\alpha/2, n-1}(\alpha, \alpha)$$

$\Rightarrow$  Do not reject  $H_0$



$$\text{P-value calculation: } H_0: \mu = 1700 \text{ vs. } H_a: \mu \neq 1700 \quad \text{test statistic}$$

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{1600 - 1700}{330/\sqrt{15}} = -1.174$$

$$|-1.174| < 2.145 \quad \text{critical value for } t_{\alpha/2, n-1} = t_{0.025, 14}$$

$\Rightarrow$  Do not reject  $H_0$

Calculate the p-value

- go to t-tables  
- look up  $\alpha$ -value, corresponding to t-value of 1.174, v=14  
- this gives  $\alpha_{\text{test}}$  of

$$\begin{aligned} 0.1 &< \alpha_{\text{test}} < 2.5 \\ 0.2 &< \text{p-value} < 0.5 \end{aligned}$$

$$\begin{aligned} \text{p-value is } &> 0.2 \\ 0.2 &> 0.5, \alpha = 0.05 = \text{significance value} \end{aligned}$$

Total p-value  $> 0.2$

$$\text{Total } \alpha_{\text{significance}} = 0.5$$

# Lec25 - Hypothesis testing for proportions

Monday, November 28, 2022 10:33 AM

- Tests for proportions (sec 9.5)
- Examples
- Relationship between confidence intervals and hypothesis tests

## HYPOTHESIS TESTS FOR PROPORTIONS

- Just as with population means, we may wish to test population proportions
- There are the same three situations that we will examine:
  - Two-tailed Test:**  $H_0: p = p_0$  vs  $H_a: p \neq p_0$
  - Upper-tailed Test:**  $H_0: p = p_0$  vs  $H_a: p > p_0$
  - Lower-tailed Test:**  $H_0: p = p_0$  vs  $H_a: p < p_0$
- As in the case of the population mean,  $p_0$  and the hypotheses must be specified before looking at the data

## TEST STATISTIC

- Our point estimate of the unknown population proportion  $p$  is the sample proportion

$$\hat{p} = \frac{x}{n}$$

Where  $x$  is the number in the sample with the characteristic of interest

- From our work on confidence intervals, we know that

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Approximately has a standard normal distribution provided

$$np > 15 \text{ and } n(1-p) > 15$$

- If we assume that  $p = p_0$  then we have that

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Approximately has a standard normal distribution provided

$$np_0 > 5 \text{ and } n(1-p) > 5$$

- This gives us the test statistic to use:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

## CONFIDENCE INTERVALS AND TESTS I

- There is a relationship between confidence intervals and hypothesis tests
- To see this, consider the  $100(1-\alpha)\%$  confidence interval given by:

$$\left( \bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

- A two-tailed test of  $H_0: \mu = \mu_0$  has rejection region

$$\left| \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \right| > z_{\alpha/2}$$

## CONFIDENCE INTERVALS AND TESTS II

- We will **not reject**  $H_0$  if, and only if

$$\left| \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \right| \leq z_{\alpha/2}$$

- But this implies we do not reject if

$$\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

- In other words, we do not reject  $H_0$  if  $\mu_0$  is in the confidence interval and we do reject if  $\mu_0$  is not in the confidence interval.
- The same is true of all other tests and the corresponding intervals

## DISCUSSION

- So we can see how a confidence interval and hypothesis test, whose significance and confidence levels are the same, agree
- If we perform a two-sided hypothesis test when  $\alpha=0.05$  and construct a 95% confidence interval using the same sample, the conclusion will be the same

### Test of a population proportion: Ex 1

The 2006 census revealed that approximately 23% of Canadians between 25 and 64 years of age have a university degree. In a certain area, a random sample of 200 adults in this age group revealed that 10 had a university degree. Test whether the proportion with a degree in this area differs from the rest of Canada. Use a 5% significance level.

Formally, we're testing:

$$H_0: p = 0.23 \text{ vs. } H_a: p \neq 0.23$$

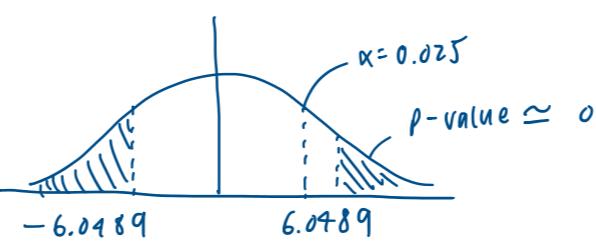
The test statistic is given by:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.05 - 0.23}{\sqrt{\frac{0.23(1-0.23)}{200}}} = -6.0489$$

Now, let's take a look at how to make our conclusion:

$$H_0: p = 0.23 \quad \hat{p} = \frac{x}{n} = \frac{10}{200} = -0.05$$

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.05 - 0.23}{\sqrt{\frac{0.23(1-0.23)}{200}}} = -6.0489$$



$$\text{Prob for } z = 6.0489 \approx 0 \\ z (\approx 0) < 0.05$$

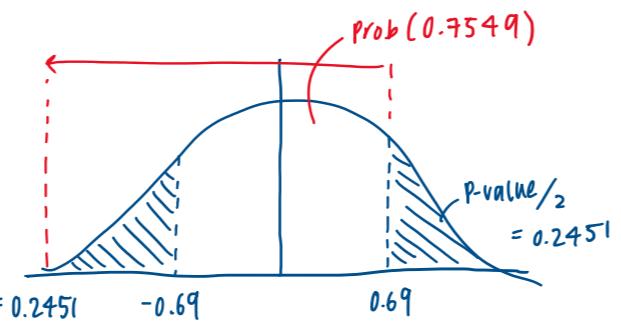
$$\text{Pvalue} < \alpha, \alpha/2 \Rightarrow \text{reject } H_0$$

∴ There is evidence to suggest that the % of Canadians with a university degree is different from 23%.

### Test of a population proportion: Ex 2

Angina pectoris is a chronic heart condition in which the sufferer has periodic attacks of chest pain. In a study to evaluate the effectiveness of the drug Timolol in preventing angina attacks, patients were randomly allocated to receive a daily dose of either Timolol or a placebo for 28 weeks. The numbers of patients who became completely free from angina attacks are shown below.

	Timolol	Placebo
Angina Free	44	19
Not Angina Free	116	128
Total	160	147



Suppose we want to test the hypothesis that 30% of patients would become angina free if Timolol were to be routinely prescribed. Our best estimate of  $p$  is  $p^ = 44/160 = 0.275$ .

$$H_0: p = 0.3 \text{ vs. } H_a: p \neq 0.3$$

$$z = \frac{0.275 - 0.3}{\sqrt{\frac{0.3(0.7)}{160}}} = -0.690$$

$$z_{0.69} \Rightarrow \text{prob} = 0.7549$$

$$1 - 0.7549 = 0.2451$$

$$5\% \text{ sig. level: } z_{\alpha/2} = 1.96$$

Now take a look at conclusion:

$$2(0.2451) = 0.4902$$

∴ 0.4902 > 0.05

⇒ do not reject  $H_0$

→ no evidence to suggest that 30% of the population who take Timolol will not be Angina free

$$\begin{aligned} CI &= \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \\ &= 39 \pm 1.96 \left( \frac{2.4}{\sqrt{25}} \right) \\ &= 39 \pm 0.94 \\ &= [38.06, 39.94] \end{aligned}$$

### Rope Example I

Consider a rope with breaking strength normally distributed with mean 41kg and standard deviation 2.4kg. A sample of 25 pieces is taken and has a mean breaking strength of 39kg; is this figure consistent with the population mean of 41kg?

So, we're testing:

$$H_0: \mu = 41 \text{ vs. } H_a: \mu \neq 41$$

$$z = \frac{39 - 41}{2.4/\sqrt{25}} = \frac{-2}{0.48} = -4.17$$

-4.17 is a significant result at  $\alpha = 0.05$ , since its absolute value is greater than the critical value  $Z_{0.025} = 1.96$

### Rope Example II

- A 95% Confidence Interval for  $\mu$  based on  $\bar{x}$

$$39 \pm 1.96 \frac{2.4}{\sqrt{25}} = 39 \pm 0.94 = (38.06, 39.94).$$

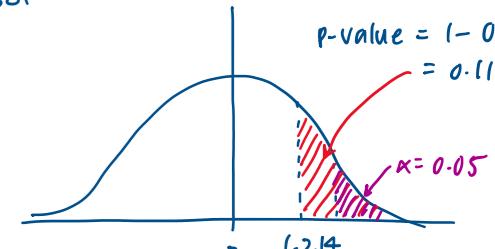
- Therefore, we are 95% confident that, based on the sample mean, the population mean lies between 38.06kg and 39.94kg
- This test and the interval are related, the fact that 41 does not lie inside the 95% confidence interval for  $\mu$ , i.e. (38.06, 39.94), is equivalent to rejecting the null hypothesis that  $\mu=41$  at significance level  $\alpha=0.05$

- ② Wagenknecht et al. (A-20) collected data on a sample of 301 Hispanic women living in San Antonio, Texas. One variable of interest was the percentage of subjects with impaired fasting glucose (IFG). IFG refers to a metabolic stage intermediate between normal glucose homeostasis and diabetes. In the study, 24 women were classified in the IFG stage. The article cites population estimates for IFG among Hispanic women in Texas as 6.3 percent. Is there sufficient evidence to indicate that the population of Hispanic women in San Antonio has a prevalence of IFG higher than 6.3 percent? Is there significant evidence that the prevalence of IFG is different from 6.3 percent? Construct a 95% confidence interval for the prevalence of IFG. Compare your results.

a)  $H_0: p = 0.063 \text{ vs. } H_a: p > 0.063$

$$\hat{p} = \frac{x}{n} = \frac{24}{301} = 0.08$$

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.08 - 0.063}{\sqrt{\frac{0.063(1-0.063)}{301}}} = 1.214$$

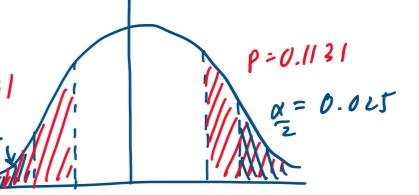


b)  $H_0: p = 0.063 \text{ vs. } H_a: p \neq 0.063$

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.08 - 0.063}{\sqrt{\frac{0.063(1-0.063)}{301}}} = 1.214$$

$$2\text{-sided critical value: } z_{\alpha/2} = \frac{z}{2} = 0.025$$

$$2\text{-test statistic: } z = 1.214 \Rightarrow \text{Prob} = 0.8868 \quad 1 - 0.8868 = 0.1131$$



$$2(0.1131) = 0.2262 \text{ total p-value}$$

∴ (p-value = 0.2262) > ( $\alpha = 0.05$ ), do not reject  $H_0$   
→ this result suggests that the true proportion is not different from 0.063, based on this sample

c) confidence interval for propor.  $\hat{p} = 0.08$   
 $\hat{p} \pm z_{\alpha/2} \frac{\hat{p}(1-\hat{p})}{n}$

$$= 0.08 \pm 1.96 \sqrt{\frac{0.08(1-0.08)}{301}}$$

$$= 0.08 \pm 1.96 (0.0156)$$

$$= 0.08 \pm 0.0307$$

∴ 95% confident that the true proportion is b/w  $[0.0493, 0.1107]$

∴ both the CI test and hypothesis test suggest  $p$  is not diff. from 0.063

# Lec26 /27 - tests on the mean, variance unknown, omit 9.3.2, tests on a population proportion

Tuesday, November 29, 2022 4:04 AM

Yuxi Qin  
[yuxiqin.ca](http://yuxiqin.ca)

- Types of error (sec 9-2,2)
- Courtroom analogy
- Finding type 2 error
- Sample size selection

## ERRORS IN HYPOTHESIS TESTING

- Two types of errors are: **TYPE I ERROR** and **TYPE II ERROR**
- A Type I Error occurs if we reject the null hypothesis when it's actually true
- A Type II error occurs if we don't reject the null hypothesis when it's actually false

## COURTROOM ANALOGY

- There's a clear analogy between a hypothesis test and a courtroom trial
- Mistakes can happen if either
  - An innocent person is convicted, or
  - A guilty person is set free
- We can summarize a **Type I Error** and a **Type II Error** using the table shown:

	$H_0$ True	$H_0$ False
Reject $H_0$	Type I Error	
Fail to reject $H_0$		Type II Error

## POWER AND THE PROBABILITY OF COMMITTING AN ERROR

- The **power** of a test is the probability that the null hypothesis is rejected and it's false
- This is the compliment of the **P(Type II Error)**
- Therefore, we can write that **P(Type II Error) =  $\beta$** , so the power is  $1-\beta$
- We will look at an example of computing the **P(Type II Error)** in a moment
- But before we do, let's talk about the **P(Type I error)**
- **P(Type I Error) =  $\alpha$** , where  $\alpha$  is the significance level

## PROBABILITY OF A TYPE II ERROR, 2-SIDED TEST

$$\beta = \Phi\left(z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right) - \Phi\left(-z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right)$$

- Where  $\beta$  is the probability of a type II error,  $\Phi(z)$  is the probability to the left of  $z$  in the standard normal distribution, and  $\delta = \mu - \mu_0$

## SAMPLE SIZE SELECTION

- The sample size required to produce a specified type II error with probability  $\beta$  given  $\delta$  and  $\alpha$ :
  - 2-sided test on the mean, variance known:
 
$$n \approx \frac{(z_{\alpha/2} + z_{\beta})^2 \sigma^2}{\delta^2}$$
  - 1-sided test on the mean, variance known:
 
$$n \approx \frac{(z_{\alpha} + z_{\beta})^2 \sigma^2}{\delta^2}$$

Where  $\delta = \mu - \mu_0$
- Remember to round up the sample size if  $n$  is not a whole number

- 9.5. A textile fiber manufacturer is investigating a new drapery yarn, which the company claims has a mean thread elongation of 12 kilograms with a standard deviation of 0.5 kilograms. The company wishes to test the hypothesis  $H_0: \mu = 12$  against  $H_1: \mu < 12$ , using a random sample of four specimens.
- What is the type I error probability if the critical region is defined as  $\bar{x} < 11.5$  kilograms?
  - Find  $\beta$  for the case in which the true mean elongation is 11.5 kilograms.
  - Find  $\beta$  for the case in which the true mean is 11.5 kilograms.

$$a) \alpha = P\left(\bar{z} \leq \frac{11.5 - 12}{0.5/\sqrt{4}}\right) = P(\bar{z} \leq -2) = 1 - P(\bar{z} \geq 2) = 1 - 0.977 = 0.2275$$

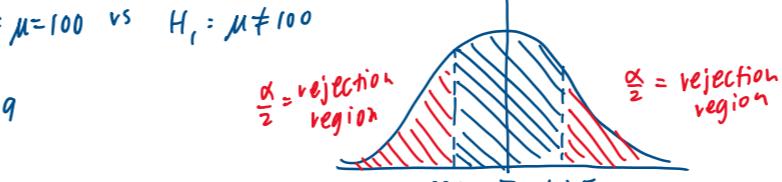
The prob. of rejecting the null hypothesis when it is true is 0.2275

$$b) \beta = P(\text{accept } H_0 \text{ when } \mu = 11.25) = P(\bar{z} > 11.5 \mid \mu = 11.25)$$

$$= P\left(\bar{z} > \frac{11.5 - 11.25}{0.5/\sqrt{4}}\right) = P\left(\bar{z} > \frac{0.25}{0.25}\right) = P(\bar{z} > 1) = 1 - P(\bar{z} \leq 1) = 1 - 0.84134 = 0.15866$$

$$c) \text{Power} = 1 - \beta = 1 - 0.15866 = 0.84134$$

- 9.10. The heat evolved in calories per gram of a cement mixture is approximately normally distributed. The mean is thought to be 100, and the standard deviation is 2. You wish to test  $H_0: \mu = 100$  versus  $H_1: \mu \neq 100$  with a sample of  $n = 9$  specimens.
- If the acceptance region is defined as  $98.5 \leq \bar{x} \leq 101.5$ , find the type I error probability  $\alpha$ .
  - Find  $\beta$  for the case in which the true mean heat evolved is 103.
  - Find  $\beta$  for the case where the true mean heat evolved is 105. This value of  $\beta$  is smaller than the one found in part (b). Why?



$$\alpha = P(\bar{z} \leq 98.5) + P(\bar{z} > 101.5) = P(\bar{z} \leq 98.5) + (1 - P(\bar{z} \leq 101.5))$$

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \quad \mu_0 = \mu = 100, \quad \sigma = 2, \quad n = 9$$

$$\alpha = P\left(\bar{z} \leq \frac{98.5 - 100}{2/\sqrt{9}}\right) + \left[1 - P\left(\bar{z} \leq \frac{101.5 - 100}{2/\sqrt{9}}\right)\right]$$

$$\alpha = P(\bar{z} \leq -2.25) + [1 - P(\bar{z} \leq 2.25)]$$

$$\alpha = 1 - P(\bar{z} \leq 2.25) + [1 - P(\bar{z} \leq -2.25)]$$

$$\alpha = 0.0244$$

P(Type I error) =  $\alpha = 0.0244$

b)  $\mu = 103$

-  $\beta$  is the probability of a type II, failing to reject null hypothesis when it's actually false

- acceptance region is defined as:

$$98.5 \leq \bar{x} \leq 101.5, \quad \mu = 103$$



$$= P(\bar{x} \leq 101.5) - P(\bar{x} \leq 98.5)$$

∴ accept one region

$$= P\left(\bar{z} \leq \frac{101.5 - 103}{2/\sqrt{9}}\right) - P\left(\bar{z} \leq \frac{98.5 - 103}{2/\sqrt{9}}\right)$$

$$= P\left(\bar{z} \leq \frac{-1.5}{0.666}\right) - P\left(\bar{z} \leq \frac{-4.5}{0.666}\right)$$

$$= P(\bar{z} \leq -2.25) - P(\bar{z} \leq -6.75)$$

$$= 0.0122 - [1 - P(\bar{z} \leq -6.75)]$$

$$= 0.0122 - [1 - 1] = 0.0122$$

$$c) \text{Power} = 1 - \beta = 1 - 0.0122 = 0.9878$$

- same method as part b

$$\beta = 0$$

$\beta$  is smaller in c because the true mean ( $\mu = 105$ ) is further from the acceptance region than  $\mu = 103$  is → there is a bigger difference b/w the true mean and the hypothesized mean

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

$$H_0: \mu = 12$$

$$H_1: \mu < 12$$

$$\bar{x} < 11.5$$

$$\sigma = 0.5$$

$$n = 4$$

## Sample Test #3

13. A university library ordinarily has a complete shelf inventory done once every year. Because of new shelving rules instituted the previous year, the head librarian believes it may be possible to save money by postponing the inventory. The librarian decides to select at random 1000 books from the library's collection and have them searched in a preliminary manner. If evidence indicates strongly that the true proportion of misshelved or unlocatable books is less than .02, then the inventory will be postponed. Among 1000 books searched, 15 were misshelved or unlocatable. Test the relevant hypothesis and advise the librarian what to do (use  $\alpha = 0.05$ ).

- Do not reject  $H_0$  since  $-1.13$  is not less than  $-1.64$ . The inventory should not be postponed.
- Do not reject  $H_0$  since  $-1.13$  is not less than  $-1.64$ . The inventory should be postponed.
- Do not reject  $H_0$  since  $-1.13$  is not less than  $-1.96$ . The inventory should not be postponed.
- Do not reject  $H_0$  since  $-1.13$  is not less than  $-1.96$ . The inventory should be postponed.
- Do not reject  $H_0$  since  $-1.48$  is not less than  $-1.64$ . The inventory should not be postponed.

$$H_0: p = 0.02 \quad \hat{p} = \frac{15}{1000} = 0.015$$

$$H_1: p < 0.02 \quad n = 1000$$

$$\alpha = 0.05 \quad \text{signif. level}$$

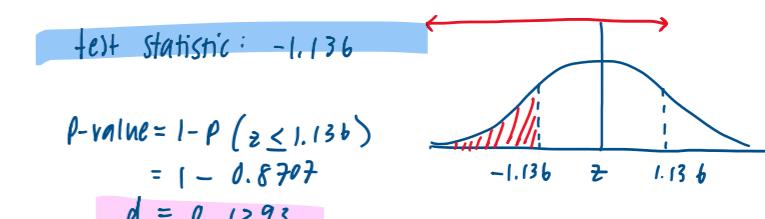
$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0.015 - 0.02}{\sqrt{0.02(0.98)/1000}} = -1.136 \quad \text{test statistic}$$

1 sided test,  $\alpha = 0.05$ , lower tailed  
z corresponding to  $P(0.95) = 1.64$

∴ do not reject  $H_0$  since  $-1.13 < -1.64$ , the inventory should not be postponed

14. (Continuation of 13.) Find the p-value.

- (a) .5168 (b) .0694 (c) .2584 (d) .1292 (e) .1883

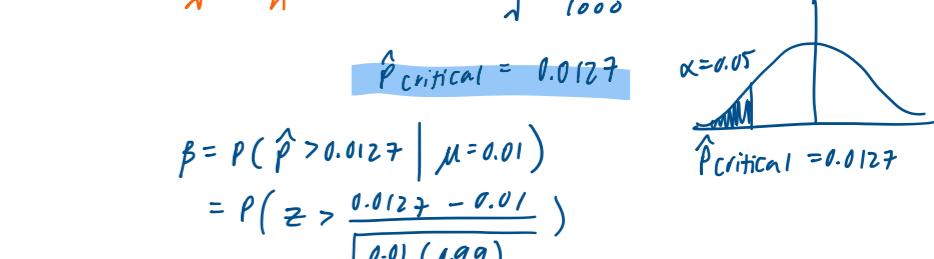
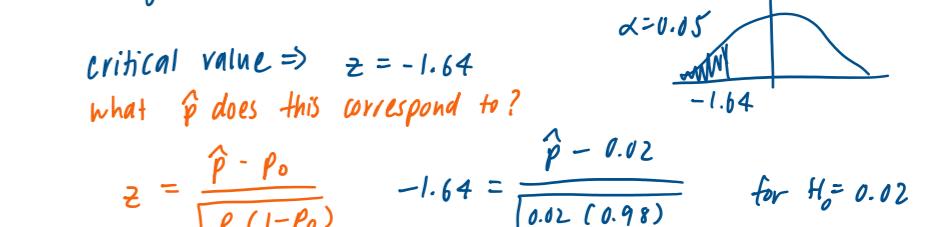


15. (Continuation of 13.) If the true proportion of misshelved and lost books is actually .01, what is the probability that the inventory will be (unnecessarily) taken?

- (a) .192 (b) .176 (c) .142 (d) .119 (e) .103

$H_0: p = 0.02$  — fail to reject, then "not postponed"

"unnecessarily taken" = not postponed when it should have been  
⇒ not reject the null hypothesis when it's actually false  
⇒ Type II error



— we want to have a large for "Power", which means we have a small chance at making a Type II error

Power of the test:  $1 - \beta = 0.80785$

— if asked to find power of a test, you need to know/find  $\beta$ , which is the probability of a type II error

# Lec28 - inference on the difference of means, variances unknown, omit 10.2.2

Tuesday, November 29, 2022 4:18 AM

Yuxi Qin  
yuxiqin.ca

- Inference on the difference in means of two normal distributions, variances unknown (10-2)

## PRELIMINARIES

- Assume that we have data from two **independent** samples, A and B
- We assume that the data in each group is **normally distributed**
- We also assume that the groups have **equal variance**; that is, we assume:  $\sigma_A^2 = \sigma_B^2$
- We then test hypotheses about the quantity  $\mu_A - \mu_B$  (or  $\mu_B - \mu_A$ )
- This makes sense because if  $\mu_A - \mu_B = 0$  then  $\mu_A = \mu_B$

## COMPUTING THE STANDARD ERROR

- Working out the standard error  $\bar{X}_A - \bar{X}_B$  requires some thought
- As seen in other tests, the standard error is used to produce a **standardized test statistic**

- In the case with known  $\sigma$

$$\text{Var}[\bar{X}] = \text{Var}\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n^2} \text{Var}\left[\sum_{i=1}^n X_i\right]$$

$$= \frac{1}{n^2} \sum_{i=1}^n \text{Var}[X_i] = \frac{1}{n^2} (n\sigma^2) = \frac{\sigma^2}{n}.$$

- And so the standard error of  $\bar{X}$  is given by

$$\text{SE}[\bar{X}] = \frac{\sigma}{\sqrt{n}}.$$

- In the test of the t-test, we do not know  $\sigma$  and so we use  $s$  to approximate it, giving:

$$\text{SE}[\bar{X}] = \frac{s}{\sqrt{n}}.$$

- Recall that the sample standard deviation is given by

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

- What about the standard error of  $\bar{X}_A - \bar{X}_B$

- Recall that we assume

$$\sigma_A^2 = \sigma_B^2 = \sigma^2$$

$$\begin{aligned} \text{Var}[\bar{X}_A - \bar{X}_B] &= \text{Var}[\bar{X}_A] + \text{Var}[\bar{X}_B] \\ &= \text{Var}\left[\frac{1}{n_A} \sum_{i=1}^{n_A} X_{A(i)}\right] + \text{Var}\left[\frac{1}{n_B} \sum_{i=1}^{n_B} X_{B(i)}\right] \\ &= \frac{1}{n_A^2} \text{Var}\left[\sum_{i=1}^{n_A} X_{A(i)}\right] + \frac{1}{n_B^2} \text{Var}\left[\sum_{i=1}^{n_B} X_{B(i)}\right] \\ &= \frac{1}{n_A^2} \sum_{i=1}^{n_A} \text{Var}[X_{A(i)}] + \frac{1}{n_B^2} \sum_{i=1}^{n_B} \text{Var}[X_{B(i)}] \\ &= \frac{1}{n_A^2} (n_A \sigma_A^2) + \frac{1}{n_B^2} (n_B \sigma_B^2) = \frac{\sigma^2}{n_A} + \frac{\sigma^2}{n_B}. \end{aligned}$$

- But for the independent groups t-test, we don't use  $\sigma_A$  or  $\sigma_B$
- So if  $n_A = n_B = n$ , then we use the quantity

$$s_p^2 = \frac{s_A^2 + s_B^2}{2}$$

To approximate  $\sigma^2$ , this is sometimes called the **pooled estimate of the variance**

- Therefore, for  $n_A = n_B = n$ ,

$$\text{SE}(\bar{X}_A - \bar{X}_B) = \sqrt{\frac{s_p^2}{n} + \frac{s_p^2}{n}} = \sqrt{\frac{2s_p^2}{n}}.$$

- Now if  $n_A \neq n_B$ , then we approximate  $\sigma^2$  using  $s_p^2$ , where

$$s_p^2 = \frac{(n_A-1)s_A^2 + (n_B-1)s_B^2}{n_A + n_B - 2}$$

- Then,

$$\text{SE}(\bar{X}_A - \bar{X}_B) = \sqrt{\frac{s_p^2}{n_A} + \frac{s_p^2}{n_B}} = \sqrt{s_p^2 \left( \frac{1}{n_A} + \frac{1}{n_B} \right)}$$

## THE INDEPENDENT GROUPS T-TEST

- The independent groups t-test tests of the form **hypotheses**

$$H_0: \mu_A - \mu_B = 0 \text{ versus } H_a: \mu_A - \mu_B \neq 0.$$

- The test statistic is given by

$$t = \frac{(\bar{X}_A - \bar{X}_B) - 0}{\text{SE}(\bar{X}_A - \bar{X}_B)}.$$

- For this hypothesis test, our critical value will have degrees of freedom:

$$V = (n_A-1) + (n_B-1) = n_A + n_B - 2$$

- Note that if  $n_A = n_B = n$ , then  $V = 2(n-1)$

## CONFIDENCE INTERVAL FOR $\mu_A - \mu_B$

- Given  $\bar{X}_A - \bar{X}_B$ , one can construct a  $(1-\alpha)\%$  confidence interval for  $\mu_A - \mu_B$  as follows:

$$(\bar{X}_A - \bar{X}_B) \pm t_{n_A+n_B-2, \alpha/2} \text{SE}(\bar{X}_A - \bar{X}_B).$$

- Once again,  $t_{n_A+n_B-2, \alpha/2}$  is the value of the t-distribution corresponding to a given  $\alpha$  and  $V = n_A + n_B - 2$

## DISCUSSION

- The last procedure assumed that  $\sigma_A^2 = \sigma_B^2$
- However, it will not always be reasonable to assume that this is true
- Therefore, there are two questions that arise from this last topic
- What do we do if  $\sigma_A^2 \neq \sigma_B^2$ , "variances unequal"

## EQUATION SHEET - HOW TO ANSWER QUESTIONS ON THIS TOPIC

- Equation 26 -  $S_p^2$  is the pooled estimator of variance
- Equation 27,  $V$  is the degrees of freedom
- Equation 28, 29 confidence interval:

$$26. \text{ Variances equal: } \bar{X}_1 - \bar{X}_2 \pm t_{\alpha/2, n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \quad s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

$$27. \text{ Variances unequal: } \bar{X}_1 - \bar{X}_2 \pm t_{\alpha/2, V} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \quad V = \left( \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2 / \left( \left( \frac{s_1^2}{n_1} \right)^2 + \left( \frac{s_2^2}{n_2} \right)^2 \right)$$

$$28. \text{ t test for comparing two means (variances equal): } t_{n_1+n_2-2} = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$29. \text{ t test for comparing two means (variances unequal): } t_V = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left( \frac{s_1^2}{n_1} \right) + \left( \frac{s_2^2}{n_2} \right)}}$$

- 10-19. Go Tutorial** The diameter of steel rods manufactured on two different extrusion machines is being investigated. Two random samples of sizes  $n_1 = 15$  and  $n_2 = 17$  are selected, and the sample means and sample variances are  $\bar{x}_1 = 8.73$ ,  $s_1^2 = 0.35$ ,  $\bar{x}_2 = 8.68$ , and  $s_2^2 = 0.40$ , respectively. Assume that  $\sigma_1^2 = \sigma_2^2$  and that the data are drawn from a normal distribution.

- (a) Is there evidence to support the claim that the two machines produce rods with different mean diameters? Use  $\alpha = 0.05$  in arriving at this conclusion. Find the P-value.
- (b) Construct a 95% confidence interval for the difference in mean rod diameter. Interpret this interval.

$$\begin{array}{ll} n_1 = 15 & n_2 = 17 \\ \bar{x}_1 = 8.73 & \bar{x}_2 = 8.68 \\ s_1^2 = 0.35 & s_2^2 = 0.40 \end{array}$$

$$\text{Assume } \sigma_1^2 = \sigma_2^2$$

The parameter of interest is mean rod diameter,  $\mu_1 - \mu_2$

$$a) H_0: \mu_1 - \mu_2 = 0 \quad \text{or} \quad \mu_1 = \mu_2 \\ H_a: \mu_1 - \mu_2 \neq 0 \quad \text{or} \quad \mu_1 \neq \mu_2$$

$$\text{Test statistic: } t_{n_1+n_2-2} = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\text{Calculate } s_p \text{ first: } s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

$$s_p = \sqrt{\frac{(15-1)(0.35) + (17-1)(0.4)}{15+17-2}} = \sqrt{\frac{14(0.35) + 16(0.4)}{30}} = 0.614$$

$$t = \frac{8.73 - 8.68}{0.614 \sqrt{\frac{1}{15} + \frac{1}{17}}} = \frac{0.05}{0.614 \sqrt{0.0833}} = 0.230$$

$$\text{Rejection region: } t > t_{\frac{\alpha}{2}, n_1+n_2-2}$$

$$t < -t_{\frac{\alpha}{2}, n_1+n_2-2}$$

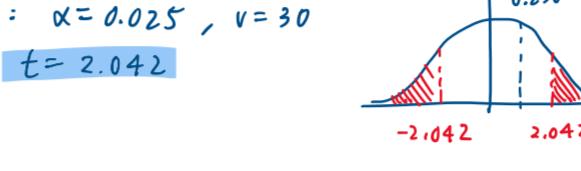
use  $\frac{\alpha}{2}$  because 2-sided interval

$$\alpha = 0.05$$

$$\frac{\alpha}{2} = 0.025$$

$$n_1+n_2-2 = 15+17-2 = 30$$

$$t\text{-tables: } \alpha = 0.025, V = 30$$



∴ fail to reject the null hypothesis

→ there is insufficient evidence to conclude that the two machines produce diff. mean diameters at  $\alpha=0.05$

$$t = 0.23 \rightarrow \text{look @ tables, } V = 30$$

$$p\text{-value} = 2( > 0.4 )$$

$$p\text{-value} = 0.8$$

$$b) 95\% \text{ CI: } t_{0.025, 30} = 2.042$$

$$\bar{X}_1 - \bar{X}_2 \pm t_{\frac{\alpha}{2}, n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$8.73 - 8.68 \pm 2.042 (0.614) \sqrt{\frac{1}{15} + \frac{1}{17}}$$

$$= 0.05 \pm 1.254 (0.3542)$$

$$= 0.05 \pm 0.444$$

$$= [-0.394, 0.494]$$

$$\text{or: } [-0.394 \leq \mu_1 - \mu_2 \leq 0.494]$$

Interpretation: because  $0$  is contained in the interval, there is insufficient evidence to conclude that the two machines produce rods with different mean diameters

- 10-27.** Two companies manufacture a rubber material intended for use in an automotive application. The part will be subjected to abrasive wear in the field application, so you decide to compare the material produced by each company in a test. Twenty-five samples of material from each company are tested in an abrasion test, and the amount of wear after 1000 cycles is observed. For company 1, the sample mean and standard deviation of wear are  $\bar{x}_1 = 20$  milligrams/1000 cycles and  $s_1 = 2$  milligrams/1000 cycles, and for company 2, you obtain  $\bar{x}_2 = 15$  milligrams/1000 cycles and  $s_2 = 8$  milligrams/1000 cycles.

- (a) Do the data support the claim that the two companies produce material with different mean wear? Use  $\alpha = 0.05$ , and assume that each population is normally distributed but that their variances are not equal. What is the P-value for this test?

- (b) Do the data support a claim that the material from company 1 has higher mean wear than the material from company 2? Use the same assumptions as in part (a).

- (c) Construct confidence intervals that will address the questions in parts (a) and (b) above.

Test statistic:

$$\bar{X}_1 = 20 \quad \bar{X}_2 = 15$$

$$s_1 = 2 \quad s_2 = 8$$

$$n_1 = 25 \quad n_2 = 25$$

$$t_0 = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$t_0 = 3.03$$

$$\bar{X}_1 = 20 \quad \bar{X}_2 = 15$$

$$s_1 = 2 \quad s_2 = 8$$

$$n_1 = 25 \quad n_2 = 25$$

$$t_0 = 3.03$$

$$t_{0.05, 27} = 2.052$$

critical value

Because  $(\text{test statistic} = 3.03) > (\text{critical value} = 2.052)$ , reject the null hypothesis.

X div 2 bc it's ied by 2 ∴ 0.05 < Pvalue < 0.01

$$b) H_0: \mu_1$$

# Lec28 - inference on the difference of means, variances unknown, omit 10.2.2

Tuesday, November 29, 2022 4:18 AM

- Inference on the difference in means of two normal distributions, variances unknown (10-2)

## PRELIMINARIES

- Assume that we have data from two **independent** samples, A and B
- We assume that the data in each group is **normally distributed**
- We also assume that the groups have **equal variance**; that is, we assume:  $\sigma_A^2 = \sigma_B^2$
- We then test hypotheses about the quantity  $\mu_A - \mu_B$  (or  $\mu_B - \mu_A$ )
- This makes sense because if  $\mu_A - \mu_B = 0$  then  $\mu_A = \mu_B$

## COMPUTING THE STANDARD ERROR

- Working out the **standard error**  $\bar{X}_A - \bar{X}_B$  requires some thought
- As seen in other tests, the standard error is used to produce a **standardized test statistic**
- In the case with known  $\sigma$

$$\text{Var}[\bar{X}] = \text{Var}\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n^2} \text{Var}\left[\sum_{i=1}^n X_i\right]$$

$$= \frac{1}{n^2} \sum_{i=1}^n \text{Var}[X_i] = \frac{1}{n^2} (n\sigma^2) = \frac{\sigma^2}{n}$$

- And so the standard error of  $\bar{X}$  is given by

$$\text{SE}[\bar{X}] = \frac{\sigma}{\sqrt{n}}$$

- In the test of the t-test, we do not know  $\sigma$  and so we use  $s$  to approximate it, giving:

$$\text{SE}[\bar{X}] = \frac{s}{\sqrt{n}}$$

- Recall that the sample standard deviation is given by

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

- What about the standard error of  $\bar{X}_A - \bar{X}_B$

- Recall that we assume  $\sigma_A^2 = \sigma_B^2 = \sigma^2$

$$\begin{aligned} \text{Var}[\bar{X}_A - \bar{X}_B] &= \text{Var}[\bar{X}_A] + \text{Var}[\bar{X}_B] \\ &= \text{Var}\left[\frac{1}{n_A} \sum_{i=1}^{n_A} X_{A(i)}\right] + \text{Var}\left[\frac{1}{n_B} \sum_{i=1}^{n_B} X_{B(i)}\right] \\ &= \frac{1}{n_A^2} \text{Var}\left[\sum_{i=1}^{n_A} X_{A(i)}\right] + \frac{1}{n_B^2} \text{Var}\left[\sum_{i=1}^{n_B} X_{B(i)}\right] \\ &= \frac{1}{n_A^2} \sum_{i=1}^{n_A} \text{Var}[X_{A(i)}] + \frac{1}{n_B^2} \sum_{i=1}^{n_B} \text{Var}[X_{B(i)}] \\ &= \frac{1}{n_A^2} (n_A \sigma_A^2) + \frac{1}{n_B^2} (n_B \sigma_B^2) = \frac{\sigma^2}{n_A} + \frac{\sigma^2}{n_B}. \end{aligned}$$

- But for the independent groups t-test, we don't use  $\sigma_A$  or  $\sigma_B$

- So if  $n_A = n_B = n$ , then we use the quantity

$$s_p^2 = \frac{s_A^2 + s_B^2}{2}$$

To approximate  $\sigma^2$ . this is sometimes called the **pooled estimate of the variance**

- Therefore, for  $n_A = n_B = n$ ,

$$\text{SE}(\bar{X}_A - \bar{X}_B) = \sqrt{\frac{s_p^2}{n} + \frac{s_p^2}{n}} = \sqrt{\frac{2s_p^2}{n}}$$

- Now if  $n_A \neq n_B$ , then we approximate  $\sigma^2$  using  $s_p^2$ , where

$$s_p^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}$$

- Then,

$$\text{SE}(\bar{X}_A - \bar{X}_B) = \sqrt{\frac{s_p^2}{n_A} + \frac{s_p^2}{n_B}} = \sqrt{s_p^2 \left( \frac{1}{n_A} + \frac{1}{n_B} \right)}$$

## THE INDEPENDENT GROUPS T-TEST

- The independent groups t-test tests of the form **hypotheses**

$$H_0: \mu_A - \mu_B = 0 \text{ versus } H_a: \mu_A - \mu_B \neq 0.$$

- The **test statistic** is given by

$$t = \frac{(\bar{X}_A - \bar{X}_B) - 0}{\text{SE}(\bar{X}_A - \bar{X}_B)}$$

- For this hypothesis test, our critical value will have degrees of freedom:

$$V = (n_A - 1) + (n_B - 1) = n_A + n_B - 2$$

- Note that if  $n_A = n_B = n$ , then  $V = 2(n-1)$

## CONFIDENCE INTERVAL FOR $\mu_A - \mu_B$

- Given  $\bar{X}_A - \bar{X}_B$ , one can construct a  $(1-\alpha)\%$  confidence interval for  $\mu_A - \mu_B$  as follows:

$$(\bar{X}_A - \bar{X}_B) \pm t_{\alpha/2, n_A + n_B - 2, \alpha/2} \text{SE}(\bar{X}_A - \bar{X}_B).$$

- Once again,  $t_{\alpha/2, n_A + n_B - 2, \alpha/2}$  is the value of the t-distribution corresponding to a given  $\alpha$  and  $V = n_A + n_B - 2$

## DISCUSSION

- The last procedure assumed that  $\sigma_A^2 = \sigma_B^2$
- However, it will not always be reasonable to assume that this is true
- Therefore, there are two questions that arise from this last topic
- What do we do if  $\sigma_A^2 \neq \sigma_B^2$ , "variances unequal"

## EQUATION SHEET - HOW TO ANSWER QUESTIONS ON THIS TOPIC

- Equation 26 -  $S_p^2$  is the pooled estimator of variance
- Equation 27,  $V$  is the degrees of freedom
- Equation 28, 29 **confidence interval**:

$$26. \text{ Variances equal: } \bar{X}_1 - \bar{X}_2 \pm t_{\alpha/2, n_1 + n_2 - 2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

$$27. \text{ Variances unequal: } \bar{X}_1 - \bar{X}_2 \pm t_{\alpha/2, V} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \quad V = \left( \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2 / \left( \frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1} \right)$$

- $t$  test for comparing two means (variances equal):  $t_{n_1 + n_2 - 2} = \frac{\bar{X}_1 - \bar{X}_2}{\text{SE}(\bar{X}_1 - \bar{X}_2)}$

- $t$  test for comparing two means (variances unequal):  $t_V = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}}$

Yuxi Qin

yuxiqin.ca

- 10-19. + Go Tutorial** The diameter of steel rods manufactured on two different extrusion machines is being investigated. Two random samples of sizes  $n_1 = 15$  and  $n_2 = 17$  are selected, and the sample means and sample variances are  $\bar{x}_1 = 8.73$ ,  $s_1^2 = 0.35$ ,  $\bar{x}_2 = 8.68$ , and  $s_2^2 = 0.40$ , respectively. Assume that  $\sigma_1^2 = \sigma_2^2$  and that the data are drawn from a normal distribution.

- (a) Is there evidence to support the claim that the two machines produce rods with different mean diameters? Use  $\alpha = 0.05$  in arriving at this conclusion. Find the  $P$ -value.  
(b) Construct a 95% confidence interval for the difference in mean rod diameter. Interpret this interval.

$$a) H_0: \mu_1 - \mu_2 = 0 \quad \text{or} \quad \mu_1 = \mu_2$$

$$H_a: \mu_1 - \mu_2 \neq 0 \quad \text{or} \quad \mu_1 \neq \mu_2$$

$$\text{Test statistic: } t_{n_1 + n_2 - 2} = \frac{\bar{x}_1 - \bar{x}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Calculate  $S_p$  first:

$$S_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}}$$

$$S_p = \sqrt{\frac{(15-1)(0.35) + (17-1)(0.4)}{15+17-2}}$$

$$S_p = 0.614$$

$$t = \frac{8.73 - 8.68}{0.614 \sqrt{\frac{1}{15} + \frac{1}{17}}} = 0.230$$

Rejection region:  $t_0 > t_{\alpha/2, n_1 + n_2 - 2}$

$$t_0 < -t_{\alpha/2, n_1 + n_2 - 2}$$

use  $\frac{\alpha}{2}$  because 2-sided interval

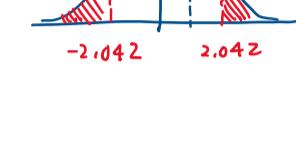
$$\alpha = 0.05$$

$$\frac{\alpha}{2} = 0.025$$

$$h_1 + h_2 - 2 = 15 + 17 - 2 = 30$$

$$t\text{-tables: } \alpha = 0.025, v = 30$$

$$t = 2.042$$



∴ fail to reject the null hypothesis

→ there is insufficient evidence to conclude that the two machines produce diff. mean diameters at  $\alpha=0.05$

$$t = 0.23 \rightarrow \text{look @ tables, } v = 30$$

$$p\text{-value} = 2( > 0.4 )$$

$$p\text{-value} = > 0.8$$

$$b) 95\% \text{ CI: } t_{0.025, 30} = 2.042$$

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2, n_1 + n_2 - 2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$8.73 - 8.68 \pm 2.042 (0.614) \sqrt{\frac{1}{15} + \frac{1}{17}}$$

$$= 0.05 \pm 1.254 (0.3542)$$

$$= 0.05 \pm 0.447$$

$$= [-0.394, 0.494]$$

Interpretation: because 0 is contained in the interval, there is insufficient evidence to conclude that the two machines produce rods with different mean diameters

- 10-27.** Two companies manufacture a rubber material intended for use in an automotive application. The part will be subjected to abrasive wear in the field application, so you decide to compare the material produced by each company in a test. Twenty-five samples of material from each company are tested in an abrasion test, and the amount of wear after 1000 cycles is observed. For company 1, the sample mean and standard deviation of wear are  $\bar{x}_1 = 20$  milligrams/1000 cycles and  $s_1 = 2$  milligrams/1000 cycles, and for company 2, you obtain  $\bar{x}_2 = 15$  milligrams/1000 cycles and  $s_2 = 8$  milligrams/1000 cycles.

- (a) Do the data support the claim that the two companies produce material with different mean wear? Use  $\alpha = 0.05$ , and assume that each population is normally distributed but that their variances are not equal. What is the  $P$ -value for this test?

- (b) Do the data support a claim that the material from company 1 has higher mean wear than the material from company 2? Use the same assumptions as in part (a).

- (c) Construct confidence intervals that will address the questions in parts (a) and (b) above.

Test statistic:

$$\bar{x}_1 = 20 \quad \bar{x}_2 = 15$$

$$s_1 = 2 \quad s_2 = 8$$

$$n_1 = 25 \quad n_2 = 25$$

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{20 - 15}{\sqrt{\frac{4}{25} + \frac{64}{25}}} = 3.03$$

$$t_0 = 3.03$$



Because  $(\text{test statistic} = 3.03) > (\text{critical value} = 2.052)$ , reject the null hypothesis.

$$b) H_0: \mu_1 - \mu_2 = 0 \quad \text{test stat} = 3.03$$

$$H_1: \mu_1 - \mu_2 > 0 \quad \text{crit. value} = t_{0.025, 27} = 1.703$$

$3.03 > 1.703$

↪ reject null hypothesis

c) unequal variances:

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2, V} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$\text{part a} - t_{\alpha/2, V} = 2.052 \quad \text{part b} - t_{\alpha/2, V} = 1.703$$

$$CI = [1.616, 8.384]$$

$$CI = [2.191, 7.819]$$

$$s_1 = 2 \quad s_2 = 8$$
</div

# Lec29 - Simple Linear Regression

Tuesday, November 29, 2022 4:38 AM

Yuxi Qin  
yuxiqin.ca

- Least squares estimates (sec 11.1, 11.2)
- Prediction

## MODELLING DATA VIA A LINE I

- Consider the following data set:

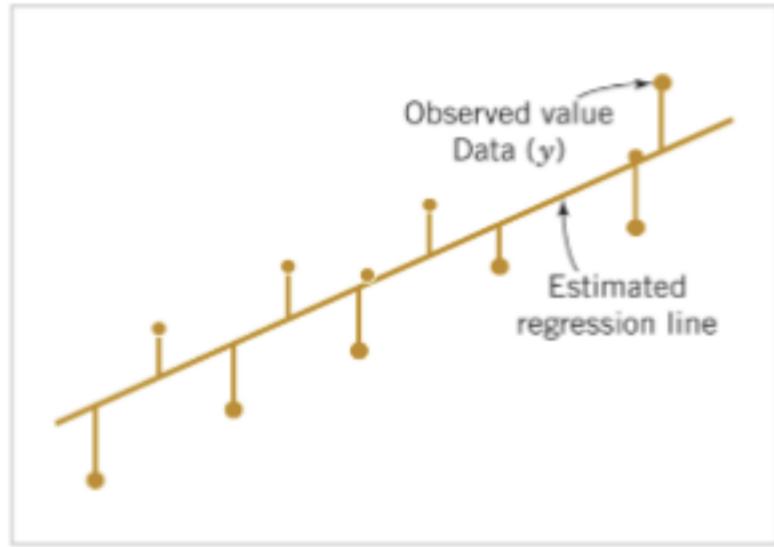
Sleep (hours)	Midterm Mark
4.6	68
8.5	83
3.2	51
5.5	76
7.0	89
6.2	75
6.4	84
5.9	71
1.5	52
10.5	91
7.3	94
5.4	74
4.9	68
3.7	56
2.2	45
Sum	77.544
	457.66

## MODELLING DATA VIA A LINE II

- We can try and model a student's midterm mark based on the number of hours they slept before...
- This will allow us to **predict** a student's midterm mark given the number of hours they slept the night before
- There are a lot of different models we could consider
- For now, let's focus on fitting a **Simple Linear Regression (SLR) model**
- We fit a SLR model using the **Least Squares Technique**

## THE LEAST SQUARES TECHNIQUE

- The least squares technique involves minimizing the sum of the squared vertical distance from each point to the line, or curve
- You can imagine this as moving the line around until the sum of the squared vertical distances is minimized
- Textbook, p285, fig 11-3



## SIMPLE LINEAR REGRESSION

- Fitting a SLR model is actually a special case of a very versatile technique called **regression**
- In general, regression refers to the process of modelling some 'dependent' variable using one or more 'explanatory' variables
- The equation of a SLR model has only one explanatory variable, which is why it's called a **Simple Linear Regression**
- Time to take a look at some finer details, and return to the Midterm Mark vs Sleep

## THE SIMPLE LINEAR REGRESSION MODEL

- Given a paired data set  $(Y_1, X_1), (Y_2, X_2), \dots, (Y_n, X_n)$ , the SLR model is given by:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \text{ for } i = 1, \dots, n,$$

Where:

- $Y_i$  is the dependent variable at  $i$ ,
- $X_i$  is the predictor variable at  $i$ ,
- $\beta_0$  is the (theoretical) intercept,
- $\beta_1$  is the (theoretical) slope,
- $\epsilon_i \sim N(0, \sigma^2)$  iid, is the (theoretical) error at  $i$ ,
- $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$ .

## FITTING THE MODEL

- In reality, we don't know the theoretical values, ie. The true values of  $\beta_0$  and  $\beta_1$
- Using these estimates in a SLR model returns the "line of best fit", ie. These estimates minimize the squared vertical distance, as required

## INTERPRETING THE PARAMETERS

$$\text{haha } y = mx + b !$$

### The slope $\beta_1$ :

- In general, for every 1 unit increase in  $X$ , the response variable  $Y$  increases by  $\beta_1$  units
- For example 1, for each additional hour of sleep a student gets, their midterm mark increases by 5.902% on average

### The y-intercept $\beta_0$ :

- In general, when  $X$  is 0 when the predicted value of  $Y$  is  $\beta_0$
- For example 1, the predicted midterm mark of students who get no sleep the night before is 39.221%

Note: use caution when interpreting the y-intercept because it may have no practical meaning

## ESTIMATING VARIANCE

- $SS_E = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ , Residual sum of squares

- $\sigma^2 = \frac{SS_E}{n-2}$ , Estimator of Variance

- $SS_E = SS_T - \hat{\beta}_1 S_{xy}$ ,

$$\text{where } SS_T = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2$$

## Ex 1. Fitting a SLR

Consider the Midterm Mark vs Sleep data - what are the estimates of  $\beta_0$  and  $\beta_1$ ?

- To answer this, we need the following summary statistics:

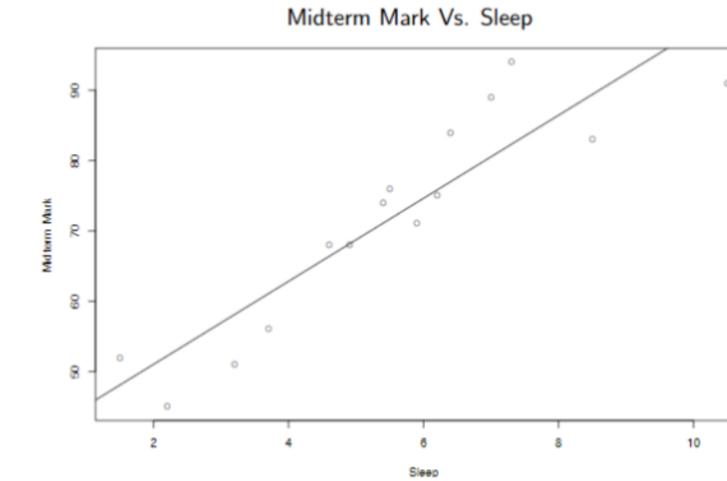
$x_i$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$y_i$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
4.6	(4.6 - 5.52)	(-0.92) <sup>2</sup>	68	(68 - 71.8)	3.496
8.5	(8.5 - 5.52)	(2.98) <sup>2</sup>	83	(83 - 71.8)	33.376
3.2	(3.2 - 5.52)	(-2.32) <sup>2</sup>	51	(51 - 71.8)	48.256
5.5	⋮	⋮	76	⋮	⋮
7.0	⋮	⋮	89	⋮	⋮
6.2	⋮	⋮	75	⋮	⋮
6.4	⋮	⋮	84	⋮	⋮
5.9	⋮	⋮	71	⋮	⋮
1.5	⋮	⋮	52	⋮	⋮
10.5	⋮	⋮	91	⋮	⋮
7.3	⋮	⋮	94	⋮	⋮
5.4	⋮	⋮	74	⋮	⋮
4.9	⋮	⋮	68	⋮	⋮
3.7	⋮	⋮	56	⋮	⋮
2.2	⋮	⋮	45	⋮	⋮
Sum	(2.2 - 5.52)	(-3.32) <sup>2</sup>	45	(45 - 71.8)	88.976
		77.544			457.66

∴ Our least square estimators are:

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{457.66}{77.544} = 5.902$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 71.8 - 5.902(5.52) = 39.221$$

∴ Our SLR model is given by:  $\hat{y}_i = 39.221 + 5.902 x_i$   
→ Let's take a look at this SLR model



## Prediction:

- We can use the simple linear regression model to make predictions
- For example, what is the predicted midterm mark for a student who got 4 hours of sleep the night before the midterm?

$$\hat{y}_i = 39.221 + 5.902 x_i = 39.221 + 5.902(4) = 62.83$$

- Therefore, based on the fitted SLR model, we predict that a student who sleeps 4 hours, the night before the midterm will get a mark of 62.83%

11-3. + An article in *Concrete Research* ["Near Surface Characteristics of Concrete: Intrinsic Permeability" (1989, Vol. 41)] presented data on compressive strength  $x$  and intrinsic permeability  $y$  of various concrete mixes and cures. Summary quantities are  $n = 14$ ,  $\sum y_i = 572$ ,  $\sum y_i^2 = 23,530$ ,  $\sum x_i = 43$ ,  $\sum x_i^2 = 157.42$ , and  $\sum x_i y_i = 1697.80$ . Assume that the two variables are related according to the simple linear regression model.

- (a) Calculate the least squares estimates of the slope and intercept. Estimate  $\sigma^2$ . Graph the regression line.

- (b) Use the equation of the fitted line to predict what permeability would be observed when the compressive strength is  $x = 4.3$ .

- (c) Give a point estimate of the mean permeability when compressive strength is  $x = 3.7$ .

- (d) Suppose that the observed value of permeability at  $x = 3.7$  is  $y = 46.1$ . Calculate the value of the corresponding residual.

$$\begin{aligned} n &= 14 \\ \sum y_i &= 572 \\ \sum y_i^2 &= 23,530 \\ \sum x_i &= 43 \\ \sum x_i^2 &= 157.42 \\ \sum x_i y_i &= 1697.80 \end{aligned}$$

$$\begin{aligned} \text{a) Slope?} \quad \hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{1697.80 - 14(43)(40.857)}{157.42 - 14(43)^2} = -2.33 \\ \bar{x} &= \frac{\sum x_i}{n} = \frac{43}{14} = 3.0714 \\ \bar{y} &= \frac{\sum y_i}{n} = \frac{572}{14} = 40.857 \\ S_{xy} &= \sum x_i y_i - n\bar{x}\bar{y} = 1697.80 - 14(43)(40.857) = -59.04 \\ S_{xx} &= \sum x_i^2 - n\bar{x}^2 = 157.42 - 14(43)^2 = 25.348 \\ \hat{\beta}_1 &= \frac{-59.04}{25.348} = -2.33 \quad (\text{slope}) \end{aligned}$$

$$\begin{aligned} \text{intercept: } \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} = 40.857 - (-2.33)(40.857) = 48.013 \quad (\text{intercept}) \\ SS_E &= 159.71 - (-2.33)(-59.04) = 22.123 \\ \sigma^2 &= \frac{SS_E}{n-2} = \frac{22.123}{14-2} = 1.8436 \end{aligned}$$

$$\begin{aligned} \text{b) } \hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x \\ \hat{y} &= 48.013 + (-2.33)(4.3) \\ \hat{y} &= 37.99 \end{aligned}$$

$$\text{c) } x = 3.7$$

$$\begin{aligned} \hat{y} &= 48.013 + (-2.33)(3.7) \\ \hat{y} &= 39.39 \end{aligned}$$

d) for  $x = 3.7$ , point estimate  $\hat{y} = 39.39$   
based on model

Given observed value  $y = 46.1$  of  $y$  is Q:  
calculate the residual:

$$e_i = y_i - \hat{y}_i = 46.1 - 39.39 = 6.71$$

# Lec30 - properties of least squares estimators, hypothesis tests

Tuesday, November 29, 2022 3:21 PM  
 • Least square estimators (section 11-3)  
 • Hypothesis tests (section 11-4)

## TESTING THE LINEAR RELATIONSHIP I

- If there is a linear relationship between two variables, say X and Y, then the slope will be non-zero
- To check if the slope is non-zero, we can use a t-test
- Formally, the t-test looks like this:

$$H_0: \beta_1 = 0 \text{ vs. } H_a: \beta_1 \neq 0$$

$$t = \frac{\hat{\beta}_1 - 0}{\widehat{SE}(\hat{\beta}_1)}.$$

- The critical value is t-distributed with  $v=n-2$  degrees of freedom for some significance level,  $\alpha$
- If  $H_0$  is rejected, then there is evidence of a linear relationship between X and Y

## TESTING THE INTERCEPT

- The test for the intercept is:

$$H_0: \beta_0 = 0 \text{ vs. } H_a: \beta_0 \neq 0$$

$$t = \frac{\hat{\beta}_0 - 0}{\widehat{SE}(\hat{\beta}_0)}.$$

## STANDARD ERRORS

- To perform a test or construct a confidence interval, we will need the standard error of  $\hat{\beta}_0$  or  $\hat{\beta}_1$ . These are given by:

$$\widehat{SE}(\hat{\beta}_0) = \sqrt{s^2 \left( \frac{1}{n} + \frac{\bar{X}^2}{\sum_i (X_i - \bar{X})^2} \right)} \text{ and } \widehat{SE}(\hat{\beta}_1) = \sqrt{\frac{s^2}{\sum_i (X_i - \bar{X})^2}},$$

Where

$$s^2 = \frac{\sum_i (Y_i - \hat{Y}_i)^2}{n-2}$$

## ALGEBRAIC DECOMPOSITION OF SUMS OF SQUARES

- For the ANOVA procedure, the Analysis of Variance Identity is given by

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

- Ie.  $SS_T = SS_R + SSE$

Where:

- $SS_T$  = Total corrected sum of squares
- $SS_R$  = Regression sum of squares
- $SSE$  = Error sum of squares

## F-TEST

- The F-test can be used to test for significance of regression
- It is equivalent to the t-test approach used earlier, ie. It will lead to the same conclusion
- However, F-tests can only be used for 2-sided tests
- If the null hypothesis  $H_0: \beta_1 = 0$  is true, the statistic

$$F_0 = \frac{SS_R/1}{SS_E/(n-2)} = \frac{MS_R}{MS_E}$$

Follows the  $F_{1, n-2}$  distribution

- Reject  $H_0$  if  $F_{\alpha/2, 1, n-2}$
- MS stands for Mean Square
- The F-distribution has 1 numerator and  $n-2$  denominator, degrees of freedom

11.4.1 [WP] Recall the regression of percent body fat on BMI from Exercise 11.2.1.

- a. Estimate the error standard deviation.

$$\bar{x} = \frac{\sum x_i}{n}$$

$$\bar{x} = 25.899$$

- b. Estimate the standard deviation of the slope.

$$\bar{y} = \frac{\sum y_i}{n}$$

$$\bar{y} = 19.03$$

- c. What is the value of the t-statistic for the slope?

- d. Test the hypothesis that  $\beta_1 = 0$  at  $\alpha = 0.05$ . What is the P-value for this test?

$$\begin{aligned} \sum_{i=1}^n x_i &= 6322.28 & \sum_{i=1}^n x_i^2 &= 162674.18 \\ \sum_{i=1}^n y_i &= 4757.90 & \sum_{i=1}^n y_i^2 &= 107679.27 \\ \sum_{i=1}^n x_i y_i &= 125471.10 & \hat{\beta}_0 &= \frac{\sum y_i}{n} = 19.03 \\ \hat{\beta}_1 &= \frac{\sum x_i y_i - \bar{x} \sum y_i}{\sum x_i^2 - n \bar{x}^2} = \frac{125471.10 - 25.899 \cdot 4757.90}{162674.18 - 25.899^2} = 1.846 \\ a) \hat{\sigma}^2 &= \frac{SSE}{n-2} = \frac{SS_T - \hat{\beta}_1 S_{xy}}{n-2} = \frac{17128.82 - 1.846 \cdot 125471.10}{248} = 30.756 \\ \text{still need to calculate } SS_T &= \sum y_i^2 - \bar{y}^2 = 107679.27 - 25.899^2 = 17128.82 \\ SS_T &= 17128.82 \end{aligned}$$

- b) Standard deviation of the slope:

$$\widehat{SE}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2}} = \sqrt{\frac{30.756}{2789.282}} = \sqrt{\frac{30.756}{2789.282}} = 0.105$$

- c) T-statistic for the slope:  $H_0: \beta_1 = 0$

$$t = \frac{\hat{\beta}_1 - 0}{\widehat{SE}(\hat{\beta}_1)} = \frac{1.846}{0.105} = 17.58$$

- d)  $\alpha = 0.05$ ,  $v = n-2 = 248$

p-value  $\approx 0$ , for  $t = 17.58$ ,  $v = 248$   
 ∴ reject  $H_0$

Yuxi Qin  
[yuxiqin.ca](http://yuxiqin.ca)

A hypothesis test: ex 1

Using the same dataset from last class (sleep data) to illustrate what was learned.

A) Is there evidence of a linear relationship between a student's midterm mark and the number of hours they slept the night before?

	$x_i$	$y_i$	$\hat{y}_i$	$Y_i - \hat{y}_i$	$(Y_i - \hat{y}_i)^2$
4.6	68	66.37	1.63	2.66	
8.5	83	89.39	-6.39	40.80	
3.2	51	58.11	-7.11	50.52	
5.5	76	71.68	4.32	18.65	
7.0	89	80.53	8.47	71.66	
6.2	75	75.81	-0.81	0.66	
6.4	84	76.99	7.01	49.09	
5.9	71	74.04	-3.04	9.26	
1.5	52	48.07	3.93	15.41	
10.5	91	101.19	-10.19	103.87	
7.3	94	82.31	11.69	136.76	
5.4	74	71.09	2.91	8.46	
4.9	68	68.14	-0.14	0.02	
3.7	56	61.06	-5.06	25.59	
2.2	45	52.21	-7.20	51.92	
Sum					585.32

First thing we need to do - compute standard error of  $\hat{\beta}_1$

$$\text{Estimate } s^2 \rightarrow s^2 = \frac{\sum_i (Y_i - \hat{y}_i)^2}{n-2} = \frac{585.32}{13} = 45.0242$$

Now compute the standard error of  $\hat{\beta}_1$ :

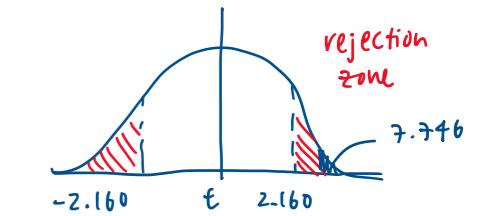
$$\widehat{SE}(\hat{\beta}_1) = \sqrt{\frac{s^2}{\sum_i (x_i - \bar{x})^2}} = \sqrt{\frac{45.0242}{77.544}} = 0.762$$

Now perform the hypothesis test about  $\hat{\beta}_1$

$$H_0: \hat{\beta}_1 = 0 \text{ vs. } H_a: \hat{\beta}_1 \neq 0, \text{ at } \alpha = 0.05$$

for this test —

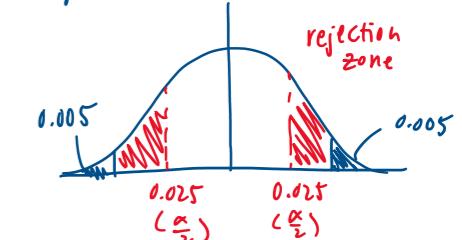
$$t = \frac{\hat{\beta}_1 - 0}{\widehat{SE}(\hat{\beta}_1)} = \frac{5.902 - 0}{0.762} = 7.746$$



Critical value is  $t_{\alpha/2, n-2} = t_{0.025, 13} = 2.160$

∴ since  $|7.746| > 2.160$ , we can reject  $H_0$  and conclude that there is evidence of linear relationship

Using p-values:  
 $v = n-2 = 13$   
 $t = 7.746$   
 $P(\text{value} < z) < 0.0005$   
 $\alpha = 0.05$   
 $< 0.001$



∴ since  $(p\text{-value} < 0.001) < (\text{sig. level} = 0.05)$ , reject  $H_0$

## EXAMPLE 11.3 | Oxygen Purity ANOVA

We now use the analysis of variance approach to test for significance of regression using the oxygen purity data model from Example 11.1. Recall that  $SS_T = 173.38$ ,  $\hat{\beta}_1 = 14.947$ ,  $S_{xy} = 10.17744$ , and  $n = 20$ . The regression sum of squares is

$$SS_R = \hat{\beta}_1 S_{xy} = (14.947) 10.17744 = 152.13$$

and the error sum of squares is

$$SS_E = SS_T - SS_R = 173.38 - 152.13 = 21.25$$

use ANOVA approach to test for significance of regression:

$$F_0 = \frac{SS_R}{SS_E} = \frac{152.13}{21.25} = 7.13$$

$$SS_R = \hat{\beta}_1 S_{xy} = (14.947)(10.17744) = 152.13$$

$$SS_E = SS_T - SS_R = 173.38 - 152.13 = 21.25$$

F-table for  $V_1 = 1, V_2 = 18$

$$F = 4.41$$

$F > F$

∴ Reject  $H_0$  since we conclude that  $\beta_1$  is not zero

TABLE 11.2 Software Output for the Oxygen Purity Data in Example 11.1						
Purity = 74.3 + 14.9 HC Level						
Predictor	Coef	SE Coef	T	P	Residual error	Total
Constant	74.283	1.593	46.62	0.000		
HC level	14.947	1.317	11.35	0.000		
S = 1.087					R-Sq = 87.7%	R-Sq (adj) = 87.1%
Analysis of Variance						
Source	DF	SS	MS	F		
Regression	1	152.13	152.13	128.86		
Residual error	18	21.25	21.25			
Total	19	173.38				
Predicted Values for New Observations						

# Lec31 - confidence and prediction intervals, omit the CI

Tuesday, November 29, 2022 5:03 PM

Yuxi Qin  
yuxiqin.ca

- Confidence intervals (section 11-5)
- Prediction (section 11-6)

## TESTING THE LINEAR RELATIONSHIP

- If there is a linear relationship between two variables, say X and Y, then the slope will be non-zero
- To check if the slope is non-zero, we can use a t-test
- Formally, the t-test looks like this:

$$H_0 : \beta_1 = 0 \text{ vs. } H_a : \beta_1 \neq 0$$

$$t = \frac{b_1 - 0}{\widehat{SE}(b_1)}.$$

- The critical value is t-distributed with  $v=n-2$  degrees of freedom for some significance level  $\alpha$
- If  $H_0$  is rejected, then there is evidence of a linear relationship between X and Y
- We can also build a  $(1-\alpha)\%$  confidence interval for  $\beta_1$  to determine if a linear relationship exists
- This interval has the form:

$$b_1 \pm t_{\alpha/2, n-2} \widehat{SE}(b_1).$$

- Similarly, we can also perform a hypothesis test and construct a confidence interval for  $\beta_0$

## STANDARD ERRORS

- We have outlined how to conduct a hypothesis tests and build a confidence interval for  $\beta_1$
- In addition, we have learned that we can do the same for  $\beta_0$
- To perform a test or construct a CI, we will need the standard error of  $b_0$  or  $b_1$  - these are given by

$$\widehat{SE}(b_0) = \sqrt{s^2 \left( \frac{1}{n} + \frac{\bar{X}^2}{\sum_i (X_i - \bar{X})^2} \right)} \text{ and } \widehat{SE}(b_1) = \sqrt{\frac{s^2}{\sum_i (X_i - \bar{X})^2}},$$

where

$$s^2 = \frac{\sum_i (Y_i - \hat{Y}_i)^2}{n-2}$$

## CONFIDENCE INTERVALS

- A confidence interval for  $\beta_1$  is given by

$$b_1 \pm t_{\alpha/2, n-2} \widehat{SE}(b_1)$$

- A confidence interval for  $\beta_0$  is given by

$$b_0 \pm t_{\alpha/2, n-2} \widehat{SE}(b_0)$$

## > CONTINUE EXAMPLE 1 WITH CONFIDENCE INTERVALS

### COMMENTS:

- We have now seen one example illustrating how to compute a confidence interval and conduct a hypothesis test about  $\beta_1$
- Again, we can also conduct inference about  $\beta_0$
- Notably, we can also perform directional, ie. One-sided, hypothesis tests and construct confidence intervals at different levels
- We can now look at a "prediction interval"

## PREDICTION INTERVALS

- We now know how to make a prediction based on a SLR model ( $y = mx + b$ )
- For this predicted value, we can compute a **Prediction Interval (PI)**
- A  $(1-\alpha)\%$  prediction interval for a single new observation  $Y$  at  $X = X_h$  is given by:

$$\hat{Y}_h \pm t_{\alpha/2, n-2} \sqrt{s^2 \left( 1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_i (X_i - \bar{X})^2} \right)},$$

Where  $\hat{Y}_h$  is the predicted value at  $X = X_h$

- It follows that this interval gives us a lower and upper bound that we can be confident that the true value of  $Y$  falls between

A hypothesis test: ex 1

Using the same dataset from last class (sleep data) to illustrate what was learned.

A) Is there evidence of a linear relationship between a student's midterm mark and the number of hours they slept the night before?

	$x_i$	$y_i$	$\hat{y}_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
4.6	68	66.37	1.63	2.66	
8.5	83	89.39	-6.39	40.80	
3.2	51	58.11	-7.11	50.52	
5.5	76	71.68	4.32	18.65	
7.0	89	80.53	8.47	71.66	
6.2	75	75.81	-0.81	0.66	
6.4	84	76.99	7.01	49.09	
5.9	71	74.04	-3.04	9.26	
1.5	52	48.07	3.93	15.41	
10.5	91	101.19	-10.19	103.87	
7.3	94	82.31	11.69	136.76	
5.4	74	71.09	2.91	8.46	
4.9	68	68.14	-0.14	0.02	
3.7	56	61.06	-5.06	25.59	
2.2	45	52.21	-7.20	51.92	
Sum				585.32	

First thing we need to do - compute standard error of  $b_1$

$$\text{Estimate } s^2 \rightarrow s^2 = \frac{\sum_i (Y_i - \hat{Y}_i)^2}{n-2} = \frac{585.32}{13} = 45.0242$$

Now compute the standard error of  $b_1$ :

$$\widehat{SE}(b_1) = \sqrt{\frac{s^2}{\sum_i (X_i - \bar{X})^2}} = \sqrt{\frac{45.0242}{77.544}} = 0.762$$

Now perform the hypothesis test about  $\beta_1$

$$H_0 : \beta_1 = 0 \text{ vs. } H_a : \beta_1 \neq 0, \text{ at } \alpha = 0.05$$

for this test -

$$t = \frac{b_1 - \beta_1}{\widehat{SE}(b_1)} = \frac{5.902 - 0}{0.762} = 7.746$$

Critical value is  $t_{\alpha/2, n-2} = t_{0.025, 13} = 2.160$

∴ since  $|7.746| > 2.160$ , we can reject  $H_0$  and conclude that there is evidence of linear relationship

Continued from Lec 30 - Ex 1 - Interval for true increase

Now let's compute the 95% confidence interval for  $\beta_1$ . A 95% confidence interval for  $\beta_1$  is given by:

$$b_1 \pm t_{\alpha/2, n-2} \widehat{SE}(b_1)$$

$$5.902 \pm 2.160 (0.762)$$

$$5.902 \pm 1.646$$

Therefore we can be 95% confident that the true increase in a student's midterm mark is between 4.256 and 7.548 percent, for every extra hour they sleep.

## Prediction Interval: Example

Calculate a 95% prediction interval for the mean midterm mark when a student gets 5 hours of sleep.

When a student gets 5 hours of sleep, we predict that their midterm mark will be  $\hat{y}_i = 39.221y + 5.902(5) = 68.731$

Therefore:

$$\begin{aligned} \hat{y}_h &\pm t_{\alpha/2, n-2} \sqrt{s^2 \left( 1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_i (X_i - \bar{X})^2} \right)} \\ 68.731 &\pm t_{\alpha/2, n-2} \sqrt{45.0241 \left( 1 + \frac{1}{15} + \frac{(5 - 7.544)^2}{77.544} \right)} \\ 68.731 &\pm 2.160 \sqrt{6.941377} \\ 68.731 &\pm 14.99337 \end{aligned}$$

∴ we can be 95% confident that a student who gets 5 hrs of sleep will have a midterm mark between 53.73% and 83.72%.

— similar to 11.3 (Lec 30)

- 11-45. Using the regression from Exercise 11-1,
- Find a 95% confidence interval for the slope.
  - Find a 95% confidence interval for the mean percent body fat for a man with a BMI of 25.
  - Find a 95% prediction interval for the percent body fat for a man with a BMI of 25.
  - Which interval is wider, the confidence interval or the prediction interval? Explain briefly.

$$\bar{X} = 25.899$$

$$\bar{Y} = 10.03$$

$$S_{xx} = 2789.28$$

$$S_{xy} = 5147.996$$

$$\beta_1 = 1.846$$

$$SS_T = 17128.82$$

$$\sigma = 5.546$$

$$\widehat{SE}\beta_1 = 0.105$$

$$t = 17.58$$

$$n = 250$$

$$\begin{aligned} \beta_0 &= \bar{Y} - \beta_1 \bar{X} \\ &= 10.032 - (1.846)(25.899) \\ \beta_0 &= -27.643 \end{aligned}$$

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

$\hat{y} = -27.643 + 1.846x$

regression model

a) 95% confidence interval on  $\beta_1$ :

$$\hat{\beta}_1 \pm t_{\alpha/2, n-2} SE(\hat{\beta}_1)$$

$$= 1.846 \pm t_{0.025, 248} (0.105)$$

$$= 1.846 \pm 1.96 (0.105)$$

$$= 1.846 \pm 0.2058$$

$$= [1.639 \leq \beta_1 \leq 2.05]$$

b) 95% CI for mean when  $X_0 = 25$

$$\hat{\mu} \pm t_{\alpha/2, n-2} \sqrt{\hat{\sigma}^2 \left( \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_i (X_i - \bar{X})^2} \right)}$$

→ calculate  $\hat{\mu}$  from regression model

$$\hat{\mu} = -27.643 + (1.846)(25)$$

$$\hat{\mu} = 18.498, t_{0.025, 248}$$

c) 95% prediction interval on  $X_0 = 25$

$$\hat{y} \pm t_{\alpha/2, n-2} \sqrt{\hat{\sigma}^2 \left( 1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_i (X_i - \bar{X})^2} \right)}$$

(note:  $\hat{\sigma}^2 = S_{yy}$ )

→ find  $\hat{y}$  using regression model:

$$\hat{y} = -27.643 + 1.846(25)$$

$$\hat{y} = 18.5$$

→  $18.5 \pm 1.96 \sqrt{30.756 \left( 1 + \frac{1}{250} + \frac{(25 - 25.899)^2}{2789.28} \right)}$

[7.553, 29.443]

Confidence interval: over many experiments, you expect the mean to fall in the confidence interval, 95% of the time - this gives you a good idea of the true population mean

Prediction interval: predicts where the next observation will be.

You construct a prediction interval using a sample

Then sample one more value from the population (you can expect that the new observation will lie in the prediction interval 95% of the time)

- d) The prediction interval is wider because it includes the variability from a measurement at  $X_0$
- o The prediction interval has to account for the uncertainty in knowing the value of the population mean, plus data scatter
  - o The prediction interval is always wider than a confidence interval

# Lec32 - Adequacy of regression model, correlation, omit the test and CI for p

Tuesday, November 29, 2022 4:18 AM

Yuxi Qin  
yuxiqin.ca

- Coefficient of determination
- Model assumptions

## THE COEFFICIENT OF DETERMINATION

- The coefficient of determination

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

- Where
  - $SS_T$  = Total corrected sum of squares
  - $SS_R$  = Regression sum of squares
  - $SS_E$  = Error sum of squares

- Is it used to judge the adequacy of a regression model. It is a measure of the amount of variability in the data that is explained by the model.

## SLR ASSUMPTIONS

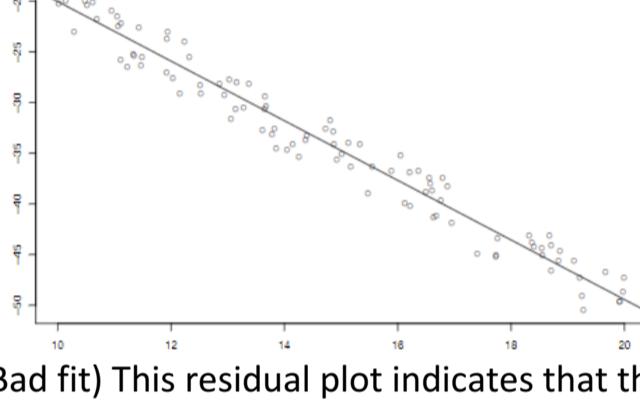
- When conducting inferences about  $\beta_1$  and  $\beta_0$ , the following model assumptions must hold:
  - Normality:** for each value of the explanatory variable ( $X$ ) there is a subpopulation of response variables ( $Y$ ) that is normally distributed
  - Linearity:** the means of the subpopulation for each value of  $X$  that falls on the straight line  $\beta_0 + \beta_1 X$
  - Constant variance:** all subpopulations have the same standard deviation  $\sigma$
  - Independence:** All observations are independent
- Note: no assumptions are needed for the distribution of  $X$

## TESTING THE ASSUMPTIONS

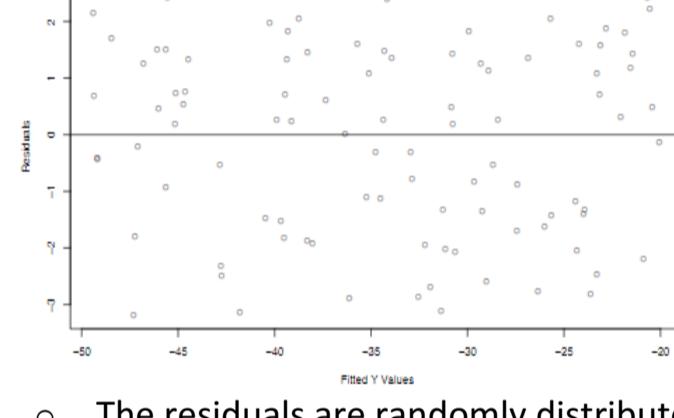
- Normality assumption
  - Construct a normal probability plot (QQ) plot of the residuals
  - The plot should have an approximately straight line if the residuals are normally distributed
- Linearity
  - Plot a scatter plot of  $Y$  vs  $X$  and superimpose the regression line
  - The estimated regression line should fit the data well
  - Plot the residual vs fitted values
  - The residuals should be randomly distributed around 0
- Constant variance
  - Plot the scatter plot of  $Y$  vs  $X$
  - Variance of  $Y$  should be constant for all  $X$
  - Plot the residuals vs fitted values of  $Y$  ie.  $\hat{Y}$
  - The variance of the residuals should be constant for all fitted values
- Independence
  - Consider the study, is it reasonable to assume that experimental units are independent?
  - Note: we can use the residual plots to test the other three assumptions

## EVALUATING ASSUMPTIONS: EXAMPLE OF A GOOD FIT

- (Good fit) This scatter plot shows a clear linear trend between  $X$  and  $Y$

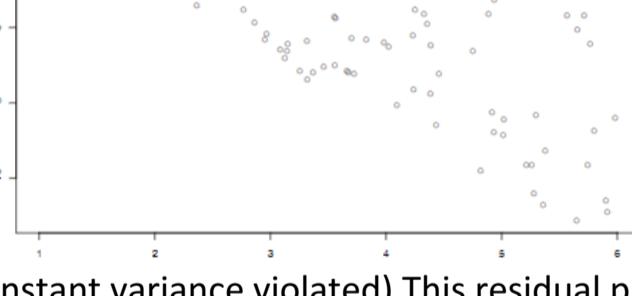


- (Bad fit) This residual plot indicates that the assumptions of linearity and constant variance do not appear to be violated

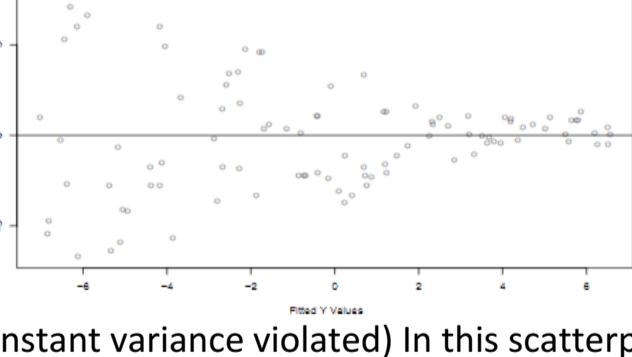


- The residuals are randomly distributed about 0
- Furthermore, the variance of the residuals is approx. const. for all fitted values

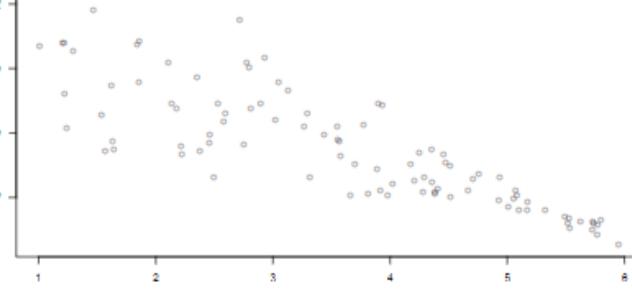
- (Constant variance violated) This scatter plot shows a decreasing trend, but shows that  $Var[Y]$  increases as  $X$  increases



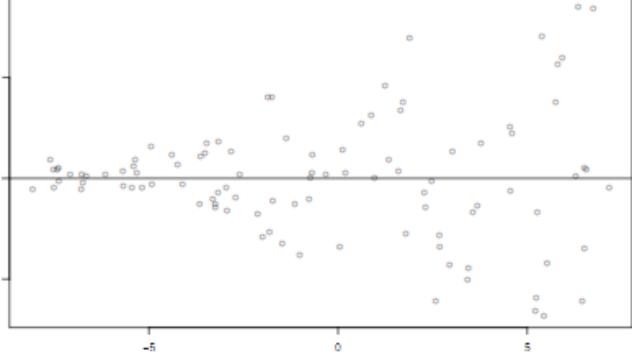
- (constant variance violated) This residual plot shows that the residual variance,  $Var[\hat{e}]$  decreases as the fitted values  $\hat{Y}$  increase, and therefore do not satisfy the assumption of constant variance



- (constant variance violated) In this scatterplot, it shows that  $Var[Y]$  decreases as  $X$  increases

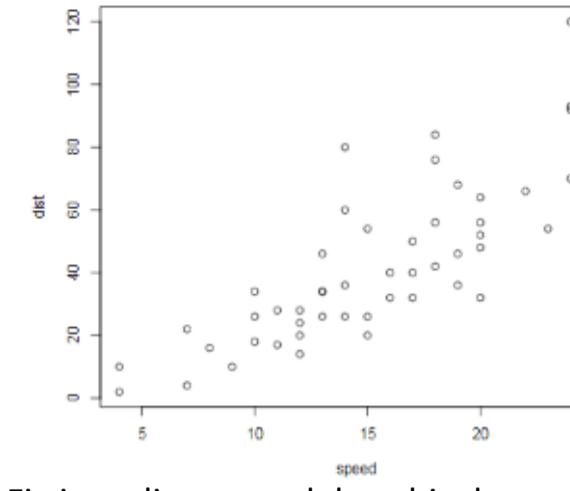


- (constant variance violated) This residual plot provides us with evidence that the assumption of constant variances is violated



## Ex. Cars Data

R contains a built-in data set called cars that contains 50 measurements on two variables: speed (in mph) and stopping distance (in feet). The following is a plot of the data.



Fitting a linear model to this data using R gives this output:

Residuals:

Min	1Q	Median	3Q	Max
-29.069	-9.525	-2.272	9.215	43.201

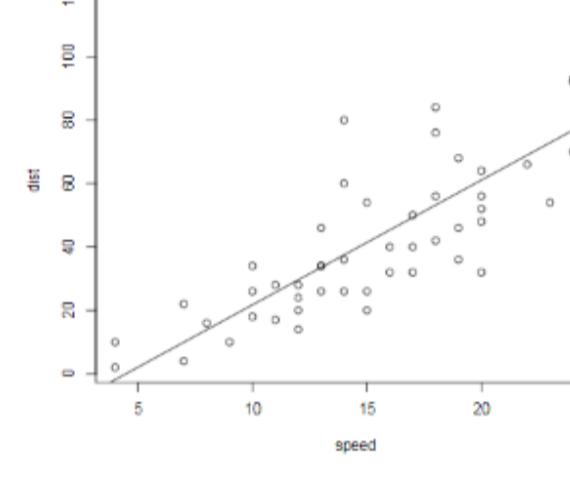
Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-17.5791	6.7584	-2.601	0.0123 *
speed	3.9324	0.4155	9.464	1.49e-12 ***

Signif. codes: 0 \*\*\* 0.001 \*\* 0.01 \* 0.05 . 0.1

Residual standard error: 15.38 on 48 degrees of freedom  
Multiple R-squared: 0.6511, Adjusted R-squared: 0.6438  
F-statistic: 89.57 on 1 and 48 DF, p-value: 1.490e-12

Here's the live we've calculated superimposed over the scatterplot:



## EXERCISES:

### ① Part I

- Explain the model on which this output is based.

The general SLR model is given by

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

where

$Y_i$  is the dependent variable (or response variable),  
 $X_i$  is the independent variable (or explanatory variable),  
 $\beta_0$  is the theoretical intercept parameter,  
 $\beta_1$  is the theoretical slope parameter,  
 $\epsilon_i \sim N(0, \sigma^2)$  are the theoretical errors for observation  $i$ , and  
 $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$

- Derive the least squares estimates of the intercept and slope in this model.

From the lecture notes, let  
 $Q = \sum_i c_i^2 = \sum_i (Y_i - \hat{Y}_i)^2 = \sum_i (Y_i - b_0 - b_1 X_i)^2$ .  
Take the partial derivatives of  $Q$  with respect to the model parameters to obtain:

$$\begin{aligned}\frac{\partial Q}{\partial b_0} &= -2 \sum_i (Y_i - b_0 - b_1 X_i) \\ \frac{\partial Q}{\partial b_1} &= -2 \sum_i X_i (Y_i - b_0 - b_1 X_i).\end{aligned}$$

Set each partial derivative equal to zero and solve for  $\beta_0$  and  $\beta_1$ . This gives:

$$\begin{aligned}b_0 &= \bar{Y} - b_1 \bar{X} \\ b_1 &= \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2}\end{aligned}$$

- Carefully explain the interpretation of the model parameters in the context of this analysis.

The slope:  $\beta_1$

- For every mile per hour that the speed of a car increase, the stopping distance increases by 3.9324 feet on average.

The y-intercept:  $\beta_0$

- The predicted stopping distance for a car travelling zero miles per hour is -17.5791 feet.

- Obtain a 95% confidence interval for the slope parameter and explain to an untrained person how this confidence interval should be interpreted.

$$b_1 \pm t_{\alpha/2, n-2} \widehat{SE}(b_1) = 3.9324 \pm 4.025 \cdot 0.4155 = 3.9324 \pm 2.011 (0.4155)$$

Therefore, we are 95% confident that for every extra mile per hour a car travels, the stopping distance will increase between 3.097 feet and 4.768 feet.

- Two p-values are given in the output. What do they represent?

The first p-value, 0.0123 results from the test  
 $H_0 : \beta_0 = 0$  vs.  $H_a : \beta_0 \neq 0$ . Therefore, at a significance level of  $\alpha = 0.05$  we can reject  $H_0$  and conclude that there is evidence that the intercept is statistically different from zero. However, given the nature of this data – the interpretation of the intercept is not realistic.

The second p-value,  $1.49e-12$ , results from the test  
 $H_0 : \beta_1 = 0$  vs.  $H_a : \beta_1 \neq 0$ . Therefore, at a significance level of  $\alpha = 0.05$  we can reject  $H_0$  and conclude that there is overwhelming evidence that the slope is statistically different from zero and there is a linear relationship between stopping distance miles per hour.

## Ex. Use the summary statistics below to calculate $R^2$

$$\sum y_i = 572 \quad \sum x_i = 43 \quad \sum x_i y_i = 1697.80$$

$$\sum y_i^2 = 23530 \quad \sum x_i^2 = 157.42 \quad n = 14$$

$$R^2 = \frac{SS_R}{SS_T} \longrightarrow \text{on each sheet: } SS_R = \hat{\beta}_1 S_{xy}$$

$$S_{xx} = \sum x_i^2 - n\bar{x}^2 = 157.42 - 14(3.0714)$$

$$S_{xx} = 25.348$$

$$S_{xy} = \sum x_i y_i - n\bar{x}\bar{y} = 1697.80 - 14(3.0714)(40.857)$$

$$S_{xy} = -59.04$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-59.04}{25.348}$$

$$\hat{\beta}_1 = -2.33$$

$$SS_T = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$

$$SS_T = 23530 - \frac{(572)^2}{14}$$

$$SS_T = 159.71$$

$$R^2 = \frac{SS_R}{SS_T} = \frac{\hat{\beta}_1 S_{xy}}{S_{yy}} = \frac{(-2.33)(-59.04)}{159.71}$$

$$R^2 = 0.861$$

Provide a practical interpretation of this quantity

- The model accounts for 86.1% of the variability in the data

# Lec33 - analysis of variance (ANOVA), omit 13.2.5

Tuesday, November 29, 2022 5:04 PM

Yuxi Qin  
yuxiqin.ca

## INTRODUCTION

- To compare more than two means, we need to use a technique based on the **algebraic decomposition of sums of squares**, called the **analysis of variance (ANOVA)**
- Although it uses the term **variance**, this test is all about the **means**
- As with the t-tests, we assume normality in each group, or sample

## TERMINOLOGY

- Suppose we have  $i=1, \dots, I$  laboratories or groups, producing data
- And suppose we have  $j=1, \dots, J$  replicates from each laboratory
- We then have to consider two types of variance: within-lab and between-lab
- ANOVA allows us to consider both types of variance when comparing the lab means

## ALGEBRAIC DECOMPOSITION OF SUMS OF SQUARES

- For the ANOVA procedure, the total sum of squares is given by

$$\sum_{i,j} (Y_{ij} - \bar{Y})^2$$

- It's true that:

$$(Y_{ij} - \bar{Y}) = (Y_{ij} - \bar{Y}_{i\cdot}) + (\bar{Y}_{i\cdot} - \bar{Y}),$$

And that if we square both sides of the equation and sum over  $I$  (from  $i=1$  to  $n$ ), we get

$$\sum_{i,j} (Y_{ij} - \bar{Y})^2 = \sum_{i,j} (Y_{ij} - \bar{Y}_{i\cdot})^2 + \sum_{i,j} (\bar{Y}_{i\cdot} - \bar{Y})^2$$

- This is the **algebraic decomposition of sums of squares**, and is very important in statistics

## TERMINOLOGY

- This decomposition of the sums of squares has its own set of terminology

$\sum_{i,j} (Y_{ij} - \bar{Y})^2$	Sum of squares total (SST)
$\sum_{i,j} (Y_{ij} - \bar{Y}_{i\cdot})^2$	Sum of squares error (SSE)
$\sum_{i,j} (\bar{Y}_{i\cdot} - \bar{Y})^2$	Sum of squares treatment (SST <sub>Trt</sub> )

$$SSTO = SSE + SST_{Trt}$$

## DEGREES OF FREEDOM

- Each element of this decomposition of the sums of squares has a degree of freedom (DoF) attached to it
- We can write the DoF along with the decomposition and terminology, through:

$$\begin{aligned} \sum_{i,j} (Y_{ij} - \bar{Y})^2 &= \sum_{i,j} (Y_{ij} - \bar{Y}_{i\cdot})^2 + \sum_{i,j} (\bar{Y}_{i\cdot} - \bar{Y})^2 \\ SSTO &= SSE + SST_{Trt} \\ IJ - 1 &= I(J - 1) + I - 1 \end{aligned}$$

## AN ANOVA MODEL

- We can summarize the ANOVA procedure using a linear model (to be talked about)
- The ANOVA model is given by:

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

For  $i=1, \dots, I$  and  $j=1, \dots, J$

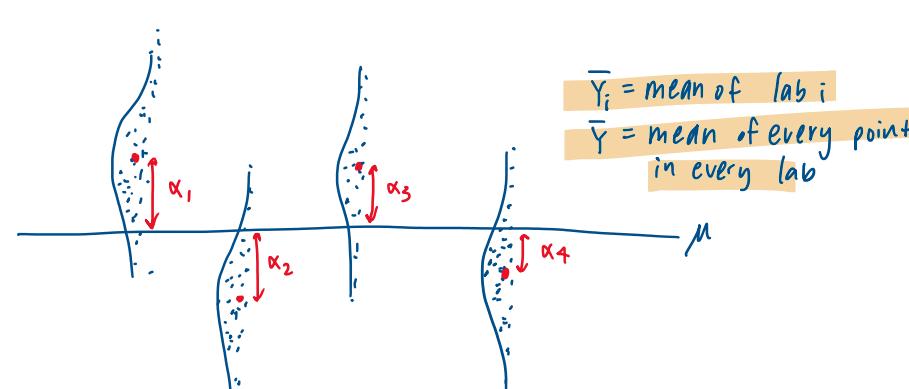
- $\mu$  and  $\alpha_i$  are fixed, unknown constants with:

$$\sum_{i=1}^I \alpha_i = 0.$$

- The error term is

$$\epsilon_{ij} \sim N(0, \sigma^2)$$

- ANOVA results are usually outputted as a table by software packages



## THE ANOVA TABLE

- The purpose of ANOVA is to **test the hypothesis that the group means are equal** vs the alternative that **a difference exists**
- This is achieved through an F-test, which uses an ANOVA table

## THE F-TEST FOR ANOVA

- The F-test associated with ANOVA is given by

$$H_0: \mu_i = \mu \quad \forall i \text{ vs. } H_a: \text{not all } \mu_i \text{ are the same}$$

- The test statistic is given by:

$$F = \frac{MS_{Trt}}{MS_E},$$

Where MS stands for **Mean Square**

- We compare the test statistic to a critical value with notation

$$F_{I-1, I(J-1)}$$

- Rejecting  $H_0$  means **there is evidence that at least two means are significantly different** from one another

TABLE VI Percentage P			
$V_1$	1	2	
1	161.4	199.5	215.7
2	18.51	19.00	19.16
3	10.13	9.55	9.28
4	7.71	6.94	6.59
5	6.61	5.79	5.41
6	5.95	5.14	4.76
7	5.59	4.74	4.36
8	5.32	4.46	4.07
9	5.12	4.26	3.86
10	4.96	4.10	3.71
11	4.84	3.98	3.59
12	4.73	3.89	3.49
13	4.67	3.81	3.41
14	4.60	3.74	3.34
15	4.54	3.68	3.29
16	4.49	3.63	3.24
17	4.45	3.58	3.16
18	4.41	3.55	3.16
19	4.38	3.52	3.13
20	4.35	3.49	3.10
21	4.32	3.47	3.07
22	4.30	3.44	3.05
23	4.28	3.42	3.03
24	4.26	3.40	3.01

## Ex. The F-Table

Since we're now conducting a F-test, we have to get used to using a new table, the F-table.

$$1. \text{ What is } F_{3, 24, \alpha=0.05} ? \quad F_{3, 24, \alpha=0.05} = 3.01$$

## Ex. An ANOVA table

Here's an ANOVA table from a study conducted with 30 observations, split evenly into 3 groups.

Source of variations	Degrees of freedom	Sum of squares	Mean square	F-value	p-value
Treatments	2	180.067	90.0335	4.6619	$0.025 > p > 0.01$
Error	27	522.46	19.313		
Total	29	702.527			

$$F_{\text{value}} = 4.6619$$

↳ off the  $\alpha=0.05$  table range  
↳ check  $\alpha=0.025$  and  $\alpha=0.1$  tables

$$0.025 > p > 0.01$$

3 groups of 10:

$$\begin{aligned} D.o.F &= i-1 = 3-1 = 2 \\ &= i(j-1) = 3(10-1) = 27 \end{aligned}$$

$$\text{Mean square} = \frac{\text{sum of squares}}{D.o.F}$$

$$\Rightarrow MS_{Trt} = \frac{180.067}{2} = 90.0335$$

$$\Rightarrow MS_E = \frac{522.46}{27} = 19.313$$

$$F = \frac{MS_{Trt}}{MS_E} = \frac{90.0335}{19.313} = 4.6619$$

↳ MS → mean square  
tr → treatment  
E → error

## Ex. Filling in an ANOVA table

Suppose there were 54 observations equally divided among 3 groups. Complete the following ANOVA table

Source of variations	Degrees of freedom	Sum of squares	Mean square	F-value	p-value
Treatments	2	536.385	268.1925	295.1925	$p < 0.01$
Error	51	46.234	0.906549		
Total	53	582.619			

3 groups of 18:

$$\begin{aligned} D.o.F : i-1 &= 3-1 = 2 \\ i(j-1) &= 3(18-1) = 51 \end{aligned} \quad \rightarrow \text{total error} = i+j = 53$$

$$\begin{aligned} SS_{Trt} &= 582.619 - 46.234 \\ SS_{Trt} &= 536.385 \end{aligned}$$

$$\begin{aligned} MS = \frac{\text{sum of square}}{D.o.F} \\ (\text{Mean Square}) \end{aligned}$$

$$\Rightarrow MS_{Trt} = \frac{536.385}{2} = 268.1925$$

$$\Rightarrow MS_E = \frac{46.234}{51} = 0.906549$$

$$\begin{aligned} F_{\text{value}} &= \frac{MS_{Trt}}{MS_E} = \frac{268.1925}{0.906549} \\ F_{\text{value}} &= 295.1925 \end{aligned}$$

$$\begin{aligned} P\text{-value} &= F_{2, 51, \alpha=0.01} \\ &\Rightarrow P < 0.01 \end{aligned}$$

## Ex. Filling in an ANOVA table

Suppose there were 6 observations per group and 48 observations total. Complete the following ANOVA table.

Source of variations	Degrees of freedom	Sum of squares	Mean square	F-value	p-value
Treatments	7	563.286	538.08	5.98	$p < 0.01$
Error	40	80.47	13.452		
Total	47	643.756	551.532		

8 groups of 6

i j

D.o.F<sub>trt</sub> = i-1 = 8-1 = 7

D.o.F<sub>E</sub> = i(j-1) = 8(6-1) = 40

MS<sub>Trt</sub> =  $\frac{SS_{Trt}}{D.o.F} = \frac{563.286}{7} =$

**INTRODUCTION**

- We have learned how to assess the results of the ANOVA procedure using an ANOVA table and the F-table
- For the ANOVA procedure, if we reject  $H_0$  then we have evidence that a difference exists, but we don't know where the difference exists
- To determine where the difference is, we can use a **multiple comparison (MC)** procedure
- There are many multiple comparison procedures
- Today we'll learn about the **Fisher's LSD**

**FISHER'S LSD**

- Say we perform an ANOVA procedure for some data set
- If  $H_0$  is rejected, then we have evidence that a difference between at least two means exists
- To determine where the difference is located, we calculate all possible pairwise confidence intervals
- These confidence intervals have the form

$$(\bar{x}_i - \bar{x}_j) \pm t_{n-l,\alpha/2} \sqrt{s_p^2 \left( \frac{1}{n_i} + \frac{1}{n_j} \right)},$$

Where  $s_p^2 = \text{MSE}$ , from the ANOVA table

=====

**Ex. Fisher's LSD**

Consider a study using female mice, where each mouse is randomly assigned to one of six possible treatment groups and we are measuring size:

- NP: Mice this group ate as much as they pleased of a non-purified, standard diet for laboratory mice.
- N/N85: This group was fed normally both before and after weaning.
- N/R50: This group was fed a normal diet before weaning and a reduced-calorie diet of 50kcal/wk after weaning.
- R/R50: This group was fed a reduced-calorie diet of 50kcal/wk both before and after weaning.
- N/R50 lopro: This group was fed a normal diet before weaning, a restricted diet of 50kcal/wk after weaning and had dietary protein content decreased with advancing age.
- N/R40: This group was fed normally before weaning and was given a severely reduced diet of 40kcal/wk after weaning.

Consider the following summary statistics and inferential results for each group of mice:

Group	Summary Stats		
	n	$\bar{x}$	s
NP	49	27.4	6.1
N/N85	57	32.7	5.1
N/R50	71	42.3	7.8
R/R50	56	42.9	6.7
N/R50 lopro	56	39.7	7.0
N/R40	60	45.1	6.7
		(25.6, 29.2)	(31.3, 34.1)
		(40.5, 44.1)	(41.1, 44.7)
		(37.8, 41.6)	(43.4, 46.8)

- Researchers want to know if there's evidence of a difference between each groups' population mean, and if so, where the difference lies
- Test the following claim:  
 $H_0 : \mu_i = \mu$  vs.  $H_a : \text{at least one } \mu_i \neq \mu_k, \text{ for } i, k = 1, \dots, 6$
- To make a conclusion about  $H_0$ , we need to compute the test statistic

$$F = \frac{MS_{Trt}}{MS_E}$$

- Compute this test statistic using an ANOVA table

Source of variations	Degrees of freedom	Sum of squares	Mean square	F-value	p-value
Treatments	5	12733.94	2546.8	57.1	p-value < 0.01
Error	343	15297.42	44.6		
Total	348				

$$N_{\text{total}} = 348, \quad \alpha = 0.05$$

$$\text{D.o.F.}_{Trt} = i-1 = 6-1 = 5 \quad \text{D.o.F.}_{E} = (j-1) = 348 - 6 = 343 \quad \rightarrow \text{Total D.o.F.} = 348 \\ \hookrightarrow \text{each treatment contributes } n_i - 1 \text{ degree of freedom}$$

$$MS_{Trt} = \frac{\text{S of Sq}}{\text{D.o.F.}} = \frac{12733.94}{5} = 2546.8$$

Critical value:

$$F_{0.05}^{0.05} = F_{5,343}^{0.05} = 2.29$$

$$F = \frac{MS_{Trt}}{MS_E} = \frac{2546.8}{44.6} = 57.1$$

Knowing p-value < 0.01 and  $\alpha = 0.05$ , we can reject  $H_0$ , conclude there is evidence of a difference.

$$@ F_{5,343, 57.1} \rightarrow p-value < 0.01$$

Since (test-statistic = 57.1) > (crit. value = 2.29), we can reject  $H_0$ .

Construct CI for every pairwise combination of means, so for this experiment we have:

$$I(I-1)/2 = 6(6-1)/2 = 15$$

=====

**Ex. Fisher's LSD**

- For illustrative purposes, consider N/N85 and N/R50 groups

Group	n	$\bar{x}$
N/N85	57	32.7
N/R50	71	42.3

- Compare these two means by building a 95% CI in the form

$$(\bar{x}_i - \bar{x}_j) \pm t_{n-l,\alpha/2} \sqrt{s_p^2 \left( \frac{1}{n_i} + \frac{1}{n_j} \right)},$$

Where  $s_p^2 = \text{MSE}$ , from the ANOVA table

- To interpret this CI, we just need to check if it includes 0 or not
  - Includes 0: no evidence to suggest means are sig. diff
  - Doesn't include 0: have evidence to suggest means are sig. diff

- Woooooo time to compute the CI for these groups :pain:

Comparison of N/N85 and N/R50 groups:

$$(\bar{x}_i - \bar{x}_j) \pm t_{n-l,\alpha/2} \sqrt{s_p^2 \left( \frac{1}{n_i} + \frac{1}{n_j} \right)}$$

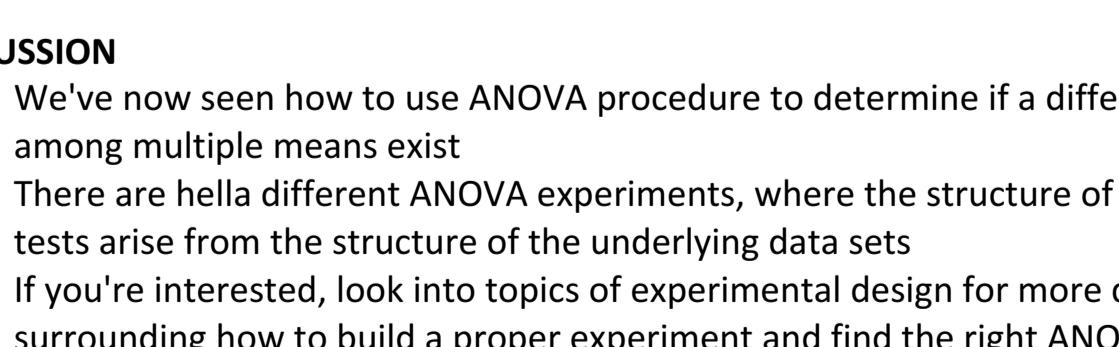
$$(32.7 - 42.3) \pm t_{343, 0.025} \sqrt{44.6 \left( \frac{1}{71} + \frac{1}{57} \right)}$$

$$-9.6 \pm t_{\infty, 0.025} (1.187697)$$

$$-9.6 \pm 1.96 (1.187697)$$

$$-9.6 \pm 2.327886$$

- Doesn't include 0, there's evidence that a difference exists
- Still need to check the other 14 pairwise combinations
- But we can just visualize the confidence intervals using a plot
- Mmmm we can use R to visualize it like so



- We can see that 10 pairwise combinations give evidence of a statistically significant difference
- Specifically, we can see that the mean of both the N/N85 and N/R50 is sig. diff. from the means of every other group

**DISCUSSION**

- We've now seen how to use ANOVA procedure to determine if a difference among multiple means exist
- There are hella different ANOVA experiments, where the structure of the tests arise from the structure of the underlying data sets
- If you're interested, look into topics of experimental design for more details surrounding how to build a proper experiment and find the right ANOVA procedure

LAST BIT OF CONTENT WOOOOOOOOOOOOOOOO

