Overview

Disclaimers + notes

- **Made by Yuxi Qin (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 34, 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)

Lec1 - Basic Probability

Yuxi Qin
yuxiqin.ca

Wednesday, September 7, 2022 1:41 PM

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 E1, E2, E3 (events 1-3), the probability of the event must be as followed:
1. $0 \leq P(E_i) \leq 1$ for all E_i
2. $P(E_1) + P(E_2) + \cdots + P(E_n) = 1$
3. $P(\emptyset) = 0$
4. 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 move events cannot occur at the same time

E and F are mutually exclusive if E ∩ F = 0

OR Rule:

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

$\cup$ = union of two sets

AND RULE

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

where $F \mid E$ 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 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:
P(F|E) = P(E)
P(E|F) = P(E)
- This is Not mutually exclusive - they're very different!

If E and F are mutually exclusive, then the following are true:
P(F|E) = 0
P(E|F) = 0
P(E∩F) = 0

From now on: independent === statistically independent

USEFUL TRICKS (+EX.)
For any event E, we use
$\overline{E}$ or E^c to denote '*not* E'.

1. 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,

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

Basic

1. 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
$$P(E) = \frac{r}{n} = \frac{3}{6} = \frac{1}{2}$$
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
$$P(5) = \frac{r}{n} = \frac{1}{6}$$
3. 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
$$P(H) = \frac{r}{n} = \frac{1}{2}$$
4. 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
$$P(H) = \frac{r}{n} = \frac{13}{52} = \frac{1}{4}$$
5. 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
$$P(J) = \frac{r}{n} = \frac{4}{52} = \frac{1}{13}$$

Or

1. 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
$$P(J \cup 5) = P(J) + P(5) = \frac{4}{52} + \frac{4}{52} = \frac{8}{52} = \frac{2}{13}$$
2. 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
$$P(1 \cup 6) = P(1) + P(6) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$$

And

1. Two cards are selected from a well shuffled pack. What is the probability that they are both jacks? Let $J1$ be the event that the first card is a jack, and let $J2$ be the event that the second card is a jack, then
$$P(J1 \cap J2) = P(J1)P(J2 \mid J1) = \frac{4}{52} \cdot \frac{3}{51} = \frac{12}{2652} = \frac{1}{221}$$
2. 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 $R1$ be the event that the first sock is red, and let $G2$ be the event that the second sock is green, then
$$P(R1 \cap G2) = P(R1)P(G2 \mid R1) = \frac{3}{7} \cdot \frac{4}{6} = \frac{2}{7}$$

Stat Independence

1. 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
$$P(C \cap 3) = P(C)P(3) = \frac{13}{52} \cdot \frac{1}{6} = \frac{1}{24}$$
2. 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
$$P(R \cap E) = P(R)P(E) = \frac{26}{52} \cdot \frac{3}{6} = \frac{1}{4}$$

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

Exercises:

1. A card is selected from a well shuffled pack (the experiment). What is the probability of it being red? 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? 4/7
3. Two fair dice are rolled.
   1. What is the probability that both numbers are odd? [Hint: write out all of the possible outcomes; there are 36.] no
   2. What is the probability that both numbers are the same?

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

3.2 $\frac{1}{6}$

1. 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?

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

1. 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?

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

Lec2 - Intro to Counting

Yuxi Qin
yuxiqin.ca

Friday, September 9, 2022 1:17 AM

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∩F)=0 and we have the familiar formula

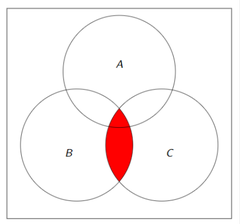
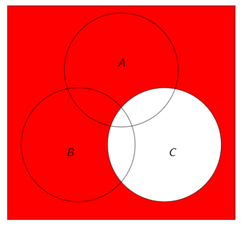
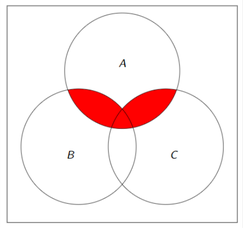
$$P(E \cup F) = P(E) + P(F).$$

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

	BUC		C'
	AUBUC		

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:

$$n! = n.(n-1).(n-2)\ldots 3.2.1.$$

n! is called an **n factorial**

The number of ways of arrange *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:

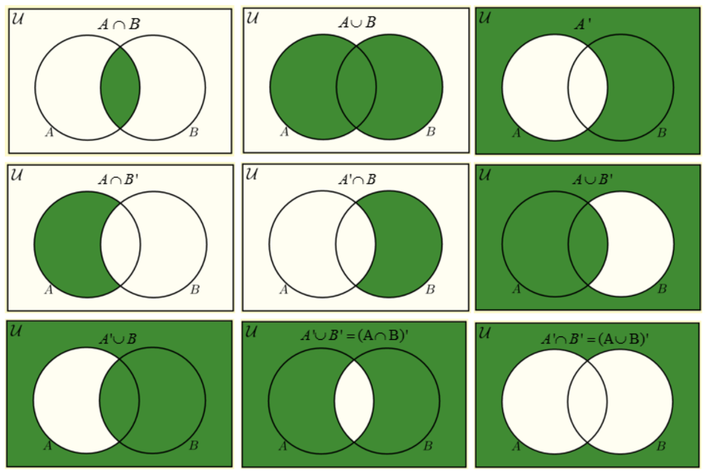
$$\binom{n}{r} = {}^nC_r = \frac{n!}{r!(n-r)!}.$$

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 is 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.

1. How many ways can the letters CANDLE be arranged to give a six letter 'word'? The answer is

$$6! = 6.5.4.3.2.1 = 720 \text{ ways.}$$

2. How many ways can 8 horses finish a race (assuming all finish and there are no ties)? The answer is

$$8! = 8.7.6.5.4.3.2.1 = 40,320 \text{ ways.}$$

3. How many ways can the letters CIARA be arranged if all must be used? There are two identical letters, so the answer is

$$\frac{5!}{2!} = 5.4.3 = 60 \text{ ways.}$$

4. How many ways can the letters ENGINEER be arranged if all must be used? There are two sets of identical letters, so

$$\frac{8!}{3!2!} = 3,360 \text{ ways.}$$

1. How many ways can a committee of 6 people be chosen from 10 people? The answer is

$$\binom{10}{6} = {}^{10}C_6 = \frac{10!}{6!4!} = \frac{10.9.8.7.6.5.4.3.2.1}{6.5.4.3.2.1.4.3.2.1} = \frac{10.9.8.7}{4.3.2.1} = \frac{5040}{24} = 210 \text{ ways}$$

2. How many ways can a committee of 4 people be chosen from 6 men and 4 women if there must be 2 people of each gender on the committee? The answer is

$$\binom{6}{2}\binom{4}{2} = {}^6C_2\,{}^4C_2 = (15)(6) = 90 \text{ ways}$$

3. In a lottery, 6 numbers are drawn from a drum with 49 numbers. How many different winning lines are possible?

$$\binom{49}{6} = {}^{49}C_6 = 13,983,816 \text{ lines}$$

- We can use the solution to Example 3 to work out the probability of winning Lotto 6/49 (if one line is played) as follows.

$$P(\text{Winning the Lotto}) = \frac{1}{{}^{49}C_6} = \frac{1}{13,983,816}.$$

CM 1.3

Problem #3: A manufacturer of front lights for automobiles tests lamps under a high-humidity, high-temperature environment using intensity and useful life as the responses of interest. The following table shows the performance of 146 lamps.

		Useful Life: Satisfactory (A)	Useful Life: Unsatisfactory (A′)
Intensity	Satisfactory (B)	110	9
	Unsatisfactory (B′)	12	15

Suppose that one of the above lamps is randomly selected. Let *A* be the event that is has a satisfactory useful life and let *B* be the event that it has a satisfactory intensity. Calculate the following probabilities.

(a) $P(A \cup B')$
(b) $P(A \cap B')$
(c) $P(A \cap B)$
(d) $P(A) \times P(B)$

a) $\frac{137}{146}$ b) $\frac{12}{146}$ c) $\frac{110}{146}$ d) $\left(\frac{122}{146}\right)\left(\frac{119}{146}\right) = \frac{7259}{10658}$

CM 1.5

Problem #5: A maintenance firm has gathered the following information regarding the failure mechanism for air conditioning systems.

		Gas Leaks: Yes (A)	Gas Leaks: No (A′)
Evidence of electrical failure	Yes (B)	67	14
	No (B′)	30	16

Cond. prob. $P(B|A) = \frac{P(A \cap B)}{P(A)}$

If this is a representative sample of AC failure, find the following probabilities.

(a) That there was a gas leak, given that there was evidence of electrical failure.
(b) That there was evidence of electrical failure, given that there was a gas leak.
(c) Let *A* be the event that there is a gas leak and let *B* be the event that there is evidence of electrical failure. Find $P(A \mid B')$.

a) $P(A|B) = \frac{67}{81}$ c) $P(A|B') = \frac{30}{46}$

b) $P(B|A) = \frac{67}{107}$

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?

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

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

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

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?

a) $\frac{1}{3}$ b) $\frac{5}{9}$ 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?

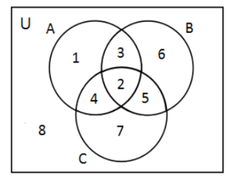
$\frac{7!}{2!2!}$

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

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

CM 1.2

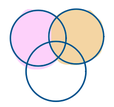
Problem #2: Consider the Venn diagram given to the right. In each part, determine which of the regions correspond to the given statements.



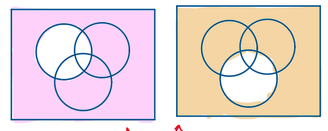
(a) $A \cup (B \cap C)$
(b) $A \cup B$
(c) $A' \cap C'$
(d) $A \cup B'$

a) A ∪ (B ∩ C) (union, intersect.) ⇒ 1, 2, 3, 4, 5

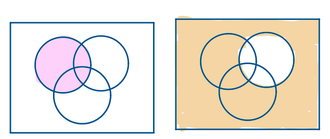
b) A ∪ B (union) ⇒ 1, 2, 3, 4, 5, 6



c) A′ ∩ C′ (intersect.) ⇒ 6, 8



d) A ∪ B′ (union) ⇒ 1, 2, 3, 4, 7, 8



CM 1.7

Problem #7: A certain system can experience three different types of defects. Let A_i (i = 1,2,3) denote the event that the system has a defect of type i. Suppose that

$P(A_1) = 0.33$, $P(A_2) = 0.28$, $P(A_3) = 0.37$,
$P(A_1 \cup A_2) = 0.55$, $P(A_1 \cup A_3) = 0.59$,
$P(A_2 \cup A_3) = 0.6$, $P(A_1 \cap A_2 \cap A_3) = 0.02$

(a) Find the probability that the system has exactly 2 of the 3 types of defects.
(b) Find the probability that the system has a type 1 defect given that it does not have a type 2 defect and does not have a type 3 defect.

a) $\left[P(A_1) + P(A_2) - P(A_1 \cup A_2) - P(A_1 \cap A_2 \cap A_3)\right] + \left[P(A_1) + P(A_3) - P(A_1 \cup A_3) - P(A_1 \cap A_2 \cap A_3)\right] + \left[P(A_2) + P(A_3) - P(A_2 \cup A_3) - P(A_1 \cap A_2 \cap A_3)\right]$
$= 2P(A_1) + 2P(A_2) + 2P(A_3) - P(A_1 \cup A_2) - P(A_1 \cup A_3) - P(A_2 \cup A_3) - 3P(A_1 \cap A_2 \cap A_3)$
$= 0.16 = \frac{4}{25}$

b) $\frac{P(A_1) - P(A_1 \cap A_2 \cap A_3) - \left[P(A_1) + P(A_2) - P(A_1 \cup A_2)\right] - P(A_1 \cap A_2 \cap A_3) - \left[P(A_1) + P(A_3) - P(A_1 \cup A_3)\right] - P(A_1 \cap A_2 \cap A_3)}{1 - P(A_2 \cup A_3)}$

$= \frac{P(A_1 \cup A_2) + P(A_1 \cup A_3) - P(A_1) - P(A_2) - P(A_3) - 3P(A_1 \cap A_2 \cap A_3)}{1 - P(A_2 \cup A_3)}$

$= \frac{9}{20}$

Lec 3 - Permutations, Conditional Prob, Partition Theorem

Yuxi Qin
yuxiqin.ca

Monday, September 12, 2022 2:04 PM

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:

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

CONDITIONAL PROBABILITY

AND Rule: if E and F are two events then

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

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

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

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

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

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 F1, F2, ... , Fn, then

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

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

E = C, F1 = S, F2 = S'

MORE LAWS

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

1. $\overline{A \cup B} = \overline{A} \cap \overline{B}$
2. $\overline{A \cap B} = \overline{A} \cup \overline{B}$

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

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 $^4P_2 = \frac{4!}{2!} = 12$ 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 $\binom{4}{2} = {}^4C_2 = \frac{4!}{2!2!} = 6$ ways to choose 2 distinct letters from the four.

2. 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?

$$P(\text{winning}) = \frac{1}{{}^{49}P_6} = \frac{1}{10068347520}$$

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 $P(\overline{S}) = 1 - 0.31 = 0.69$.
   Furthermore, we know that $P(C \mid S) = 0.19$ and $P(C \mid \overline{S}) = 0.02$.

   We deduce $P(C)$ as follows:

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

   Therefore, the answer is 7.27%.

- So, why is it the case that $P(C) = P(C \cap S) + P(C \cap \overline{S})$?
- 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 groups probabilities
2. Or, in a more technical language, the set C is **partitioned** by S and $\overline{S}$, giving the result.

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?

- 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

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

Similarly,

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

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

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%.

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

2a) RW → computer reworks
A1 → computer assembled on line 1
A2 → " " line 2

$P(A1) = 0.60$ $P(RW \mid A1) = 1 - 0.95 = 0.05$

$P(A2) = 0.40$ $P(RW \mid A2) = 1 - 0.88 = 0.12$

$$\begin{aligned} P(RW) &= P(RW \cap A1) + P(RW \cap A2) \\ &= P(RW \cap A1)P(A1) + P(RW \mid A2)P(A2) \\ &= 0.05 \cdot 0.60 + 0.12 \cdot 0.40 \\ &= 0.078 \end{aligned}$$

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.

a) $\frac{8(7)(6)}{8^3} = \frac{21}{32}$

b) $\frac{1}{18} + \frac{4}{8} - \frac{1}{18}\left(\frac{4}{8}\right) = \frac{19}{36}$ (1st letter; last #; intersection of both)

c) $\left(\frac{1}{18}\right)\left(\frac{17 \cdot 17}{18^2}\right) = \frac{289}{5832}$ (A; 2 possible places for B (letter))

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?

a) $P(\text{same shift}) = \frac{\binom{15}{7} + \binom{12}{7} + \binom{8}{7}}{\binom{35}{7}} = \frac{1447}{1344907} = 0.00107591$ (15C7; OR; $\binom{35}{7}$ total #)

b) $1 - P(\text{all same shift}) = 1 - \frac{1447}{1344907} = 0.99892408$

c) $\frac{\binom{15}{2} + \binom{20}{5}}{\binom{35}{7}} = \frac{2394}{9889} = 0.24208716$ (specific; other, don't care)

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) $\frac{4}{14}\left(\frac{13}{13}\right)\left(\frac{2}{12}\right)\left(\frac{1}{11}\right) = \frac{1}{1001}$ KKKK xxxx xxxxx (don't care)

b) $\frac{5}{9}$ 4K 5H ~~5S~~ (last) → $n = 9$, $P(H) = \frac{5}{9}$

Lec 4 - Examples, Probability

Yuxi Qin
yuxiqin.ca

Wednesday, September 14, 2022 1:30 PM

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?

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

$n = 8$, $r = 5$ $\quad {}^8C_5 = \frac{8!}{5!(8-5)!} = 56$

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 total
def: 3
non def: 47

Select 2 def parts from 3 ?

$\binom{3}{2} = {}^3C_2 = \frac{3!}{2!(3-2)!} = 3$

$\binom{47}{4} = \frac{47!}{4!(47-4)!} = 178,365$

${}_{n_1}C_{r_1} \cdot {}_{n_2}C_{r_2} = 3(178365)$

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

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

Objects of where n1 are one type, n2 of another type, etc.. Is

$$\frac{n!}{n_1!n_2!\ldots.n_r!}$$

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

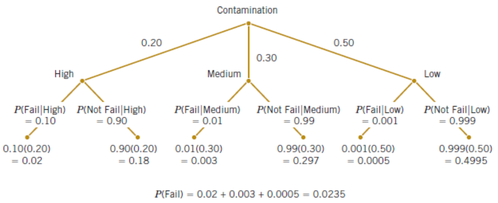
In a particular production run, 20% of the chips are subjected to high levels of contamination, 30% to medium levels of contamination, and 50% to low levels of contamination. What is the probability that a product using one of these chips fails? Let

- 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.

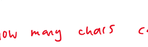


$P(\text{Fail}) = 0.02 + 0.003 + 0.0005 = 0.0235$

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 *bwBwBWbwb* (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



⟶ each char has 2 wide bars, 1 wide space

How many chars can be coded?

Let b = narrow black
B = wide black
w = narrow white
W = wide white

→ * 4 white spaces occur btn 5 black bars

1. Calculate # of permutations of 5 black bars

B=2
b=3

$n = 5$, $n_1 = 2$, $n_2 = 3$ $\quad n = \frac{n!}{n_1!\,n_2!\ldots n_n!} = \frac{5!}{2!\,3!} = 10$

2. White spaces : 3 w & 1 W

$n = 4$, $n_1 = 3$, $n_2 = 1$ $\quad n = \frac{4!}{3!\,1!} = 4$

delimiter as highlighted in question

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

$= 40 - 1 = 39$ 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?

$n = 5$
$n_1 = 3$ (knee)
$n_2 = 2$ (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?

a) ${}^{13}P_5 = \frac{13!}{(13-5)!} = \frac{13!}{8!} = 154440$

↳ on calculator: 13 P 5 (shift + x)

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

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

↳ on calculator: 13 C 5 (shift + ÷)

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. D

(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. N

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

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

❶ What is the probability that the second part is defective given that the first part is defective?

$P(B|A)$ 3 defective, 47 non defective ; $n = 50$

1st piece is defective —

49 pieces left, 2 of which are defective :

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

❷ 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 ⟶ $\frac{47}{48}$

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

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

Yuxi Qin
yuxiqin.ca

Thursday, September 22, 2022 1:24 PM

STATISTICAL INDEPENDENCE
AND rules for events E and F:

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

- 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 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:
$P(F \mid E) = P(F)$
$P(E \mid F) = P(E)$
  - Very different from mutually exclusive

If E and F are **mutually exclusive**, then these rules apply:
$P(F \mid E) = 0$
$P(E \mid F) = 0$
$P(E \cap F) = 0$

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)? -> 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		

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 $P(B \mid A)$, where A and B denote the events that the first and second parts are defective, respectively. Because the first part 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,

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

Also, the probability that both parts are defective is

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

w/ replacement!

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?

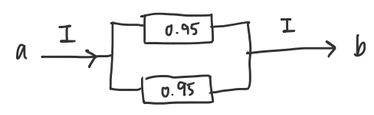
$P(C) = 0.01$ → given
$P(nC) = 0.99$
$n = 15$

independent events → multiply probabilities

$P(nC_{15}) = (0.99)^{15} = 0.86$

Note: independence (multiple events)
- If E1, E2 … En are independent then
  $P(E_1 \cap E_2 \cdots \cap E_n) = P(E_1)P(E_2)\cdots P(E_n)$

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



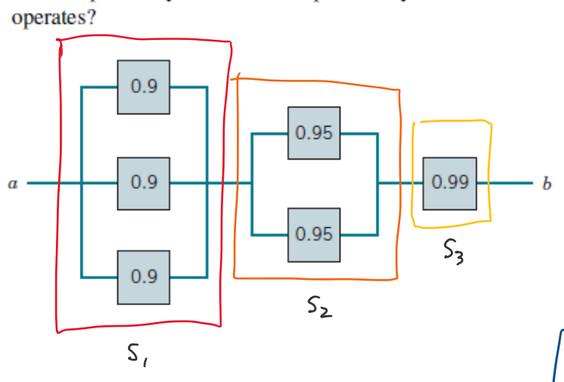
$P(W) = 0.95$
$P(F) = 1 - 0.95 = 0.05$ → each device has a 0.05 chance of failure

What is the probability the circuit works?

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

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
→ work out prob. each S_i fails
$(\bar{S}_i) = S_i$ fails

$P(\bar{S}_1) = (0.1)^3$
$P(S_1) = (1 - (0.1)^3)$ → $(1-0.9)(1-0.9)(1-0.9) = (0.1)^3$

$P(\bar{S}_2) = (0.05)^2$
$P(S_2) = (1 - (0.05)^2)$ → $(1-0.95)(1-0.95) = (0.05)^2$

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

$P(S_1)P(S_2)P(S_3)$
$= (1-0.1^3)(1-0.05^2)(0.99)$
$= 0.987$

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

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

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?

$P(A_1) = 0.98$ $P(\bar{A}_1) = 0.02$
$P(A_2) = 0.82$ $P(\bar{A}_2) = 0.18$
$P(A_3) = 0.74$ $P(\bar{A}_3) = 0.26$

a) $P(\text{at least one}) = 1 - P(\text{all work})$
$= 1 - 0.98(0.82)(0.74)$
$= 1 - 0.594664$
$= 0.405336$ ✓

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

CM 2.3

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

13 total
↳ 7 defective (D)
↳ 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) Find $P(X=5)$ → DDDDN
$= \frac{7}{13}\left(\frac{6}{12}\right)\left(\frac{5}{11}\right)\left(\frac{4}{10}\right)\left(\frac{6}{9}\right)$
$= \frac{14}{429}$ ✓

- diff. orders of N & D
- it takes 5 components to get NN, so we know it's at the end.

b) Find $P(X=5)$ N = non defective, D = Defective

→ DDDNN
$= \frac{7}{13}\left(\frac{6}{12}\right)\left(\frac{5}{11}\right)\left(\frac{6}{10}\right)\left(\frac{5}{9}\right)$
$= \frac{35}{858}$

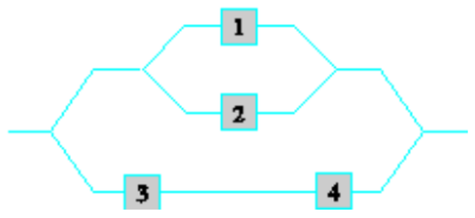
→ NDDNN
$= \frac{6}{13}\left(\frac{7}{12}\right)\left(\frac{6}{11}\right)\left(\frac{5}{10}\right)\left(\frac{4}{9}\right)$
$= \frac{14}{429}$

→ DNDNN
$= \frac{7}{13}\left(\frac{6}{12}\right)\left(\frac{6}{11}\right)\left(\frac{5}{10}\right)\left(\frac{4}{9}\right)$
$= \frac{14}{429}$

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

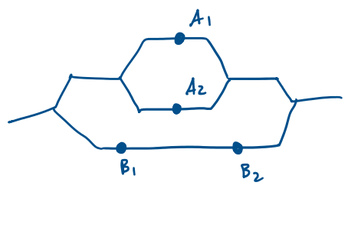
CM 2.1

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.



$P(A_1) = 0.72$
$P(A_2) = 0.73$
OR
$P(B_1) = 0.88$
$P(B_2) = 0.81$
AND

Let A, B be these systems work



A or B
$P(\text{system works}) = P(A \cup B)$
$= P(A) + P(B) - P(A)P(B)$
$= P(A_1 \cup A_2) + P(B) - P(A)P(B)$
$= P(A_1) + P(A_2) - P(A_1)P(A_2) + P(B_1)P(B_2) - P(A)P(B)$
$= 0.72 + 0.73 - 0.72(0.73) + 0.88(0.81) - P(A)P(B)$
$= 0.9244 + 0.7128 - 0.9244(0.7128)$
$= 0.978288$ ✓

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

Yuxi Qin
yuxiqin.ca

Thursday, September 22, 2022 3:20 PM

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)? -> X is called a **discrete random variable**

DISCRETE PROBABILITY DISTRIBUTION
If a discrete random variable X can take values x1, x2, … , xn with probabilities p1, p2, … , pn such that
$p_1 + p_2 + \cdots + p_n = 1$, and
$p_i \geq 0$ 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 = {s1, s2, … }
Ex. S = {HH, HT, TH, TT}

Random Variable: function from the sample space to the real numbers
Ex. X(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 Sx
Ex. Sx = {0, 1, 2} 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 x1, x2, … , xn, a pmf is a function such that:

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

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

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

Variance of a random variable X is given by

$$\mathbb{V}ar[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$$

Standard Variation of a random variable X is given by

$$\text{SD}[X] = \sqrt{\mathbb{V}ar[X]}, \quad \text{i.e., } \sigma = \sqrt{\sigma^2}$$

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 $\mathcal{S}$	$P(X = X)$	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

$P(x=1) = \frac{15}{56} + \frac{15}{56} = \frac{30}{56}$

$P(x \geq 1) = \frac{15}{56} + \frac{15}{56} + \frac{20}{56} = \frac{50}{56}$

$= 1 - P(x=0) = 1 - \frac{6}{56}$

1	2
T	T
T	H
H	H
H	T

x = # of tails

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

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

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	$x - 0.4$	$(x - 0.4)^2$	$f(x)$	$f(x)(x - 0.4)^2$
0	−0.4	0.16	0.6561	0.104976
1	0.6	0.36	0.2916	0.104976
2	1.6	2.56	0.0486	0.124416
3	2.6	6.76	0.0036	0.024336
4	3.6	12.96	0.0001	0.001296

$$V(X) = \sigma^2 = \sum_{i=1}^{5} f(x_i)(x_i - 0.4)^2 = 0.36$$

The alternative formula for variance could also be used to obtain the same result.

Practical Interpretation: The mean and variance summarize the distribution of a random variable. The mean is a weighted average of the values, and the variance measures the dispersion of the values from the mean. Different distributions may have the same mean and variance.

X = # of error bits in a 4 bit transmission

X	f(x)
0	0.6561
1	0.2916
2	0.0486
3	0.0036
4	0.0001

E(x)? = μ → E(x) = expected value of x

$E(x) = \sum_x x P(X = x)$

$= \sum_x x f(x)$

$= 0(0.06561) + 1(0.2916) + 2(0.0486) + 3(0.0036) + 4(0.0001)$

$E(x) = 0.4$

σ^2? → variance

$\sigma^2 = E[(x - E[x])^2]$

$= \sum_x (x - E(x))^2 \cdot p(x)$

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

all the same thing?

$\sigma^2 = (0 - 0.4)^2(0.6561) + (1 - 0.4)^2(0.2916) + (2 - 0.4)^2(0.0486) + (3 - 0.4)^2(0.0036) + (4 - 0.4)^2(0.0001)$

$= 0.105 + 0.105 + 0.1244 + 0.0243 + 0.0013$

$\sigma^2 = 0.36$

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
X = 0, 1, 2

a)
$P(x=0) = \overline{P(1)} \cdot \overline{P(2)} = 0.10 \cdot 0.19 = 0.019$
$P(x=1) = \overline{P(1)} \cdot P(2) + P(1) \cdot \overline{P(2)}$
$= 0.10 \cdot 0.81 + 0.90 \cdot 0.19$
$= 0.252$
$P(x=2) = P(1) \cdot P(2) = 0.90 \cdot 0.81 = 0.729$

$E[x] = \sum_x x P(x=x)$
$= 0(0.019) + 1(0.252) + 2(0.729)$
$\mu = 1.71$ ✓

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

b) $\sigma^2 = \sum_x (x-\mu)^2 f(x)$
$= (0 - 1.71)^2(0.019) + (1 - 1.71)^2(0.252) + (2 - 1.71)^2(0.729)$
$= (2.9241)(0.019) + 0.5041(0.252) + 0.0841(0.729)$
$= 0.055579 + 0.1270332 + 0.06130089$ $= 0.24389789$ ✓

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.

$\mu = E(x) = 6$

Prob of x_i is $\frac{1}{8} = 0.20$

$6 = 0(0.2) + 1(0.2) + 2(0.2) + 3(0.2) + x(0.2)$
$6 = 1.2 + 0.2x$
$x = 24$

Lec7 - Binomial and Geometric Distribution

Thursday, September 22, 2022 7:52 PM

Yuxi Qin
yuxiqin.ca

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, ...

THE GEOMETRIC DISTRIBUTION, MEAN, AND VARIANCE

If X is a geometric random variable with parameter p,

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

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

1. 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.

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

$P(X=4) = \binom{6}{4} \left(\frac{1}{2}\right)^4 \left(1 - \frac{1}{2}\right)^{6-4}$

$= 0.234375$

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.

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

$P(X \geq 4) = P(X=4) + P(X=5)$

$P[X \geq 4) = \binom{5}{4}\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^1 + \binom{5}{5}\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^0$

$= 0.15625 + 0.03125$

$= 0.1875$

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?

$E(x) = np = n(0.1) = 0.4$

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

$= 4(0.1)(0.9)$

$= 0.36$

Transmission of bits:
P = 0.1 ; probability of error

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

$M = E(x) = \frac{1}{p} = \frac{1}{0.1} \longrightarrow 10$

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

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

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?

1. 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;

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

$\simeq 0.013$

b) X = 0; p = 1/6 (probability of rolling a "1"

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

$\simeq 0.162$

c) P(x < 2)= P(x = 0)+ P(x = 1); p = 1/6; n = 10

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

$\simeq 0.485$

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?

2. Geometric: x = number of trials until 1st success

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

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

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

$= 0.03125$

b) 1st success on 5^{th} try: 0.03125

— we have binomial type Q's

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

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

$= 1 - [P(x=0) + P(x=1) + P(x=2)]$

Binomial dist.: prob of obtaining x successes in n Bernoulli trials.

$= 1 - \left[\binom{8}{0}(0.03125)^0(1-0.03125)^{8-0} + \binom{8}{1}(0.03125)^1(1-0.03125)^{8-1} + \binom{8}{2}(0.03125)^2(1-0.03125)^{8-2}\right]$

$= 1 - 0.98477$

$\simeq 0.01523$

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

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

a) n = 56; x = 0, 1, 2, 3; P(0) = 0.05

chances of failure

$P(x \leq 3) = P(x=0) + P(x=1) + P(x=2) + P(x=3)$

$= \binom{56}{0}(0.05)^0(0.95)^{56} + \binom{56}{1}(0.05)^1(0.95)^{55} + \binom{56}{2}(0.05)^2(0.95)^{54} + \binom{56}{3}(0.05)^3(0.95)^{53}$

$= 0.05656\,1627 + 0.16670\,3954 + 0.24128\,7828 + 0.22855\,8469$

$= 0.69314\,5878$ ✓

b) σ^2 = # of ND among 56 in stock — find μ

X = ND = 1 - 0.05 = 0.95

$\mu = np = 56(0.95) = 53.2$ ✓

c) $\sigma^2 = np(1-p) = 56(0.95)(0.05)$

$= 2.66$ ✓

CM 2.7

Problem #7: A geologist has collected 21 specimens of basaltic rock and 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) $P(\geq 2\text{G}) = 1 - P(\leq 1\text{G})$

$= 1 - \left[P(0\text{G}) + P(1\text{G})\right]$

$= 1 - \left[\frac{\binom{15}{0}\binom{21}{8}}{\binom{36}{8}} + \frac{\binom{15}{1}\binom{21}{7}}{\binom{36}{8}}\right]$

$= 1 - 0.06436\,84445$

$= 0.93563\,55554$ ✓

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

c) $P(8\text{G}) = \frac{\binom{15}{8}\binom{21}{0}}{\binom{36}{8}} = \frac{13}{61132} = p$

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

$= 4702$ ✓

Lec8 - Negative Bionomial Dist. And Hypergeometric Dist.

Yuxi Qin
yuxiqin.ca

Thursday, September 22, 2022 8:47 PM

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 .....** 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 <= N

- If we sample n of these units then the probability that x are defective is given by:

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

(2)

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

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

$E(X)=n\left(\frac{S}{N}\right)$ $n\longrightarrow$ # of trials; $S\longrightarrow$ # of successes available; $N\longrightarrow$ population site

OTHER SUGGESTED EXAMPLES:

1. 800 men are studies. Suppose 30% carry a gene that indicates an increased risk of high blood pressure.
   a. 10 men selected, what is the probability that exactly 1 man has the marker?

X = # of men carrying marker
N = 800 ; M = 240 ; n = 10 ← 800(0.30) = 240

$$P(X=1)=\frac{\binom{240}{1}\binom{800-240}{10-1}}{\binom{800}{10}}=0.1201$$

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

$P(X>1)=1-P(X\le 1)=1-[P(X=0)+P(X=1)]$

→ need $P(X=0)$

$$P(X=0)=\frac{\binom{240}{0}\binom{800-240}{10-0}}{\binom{800}{10}}=0.0276$$

→ back to solving

$=1-(0.0276+0.1201)$

$=0.8523$

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.

(a) What is the probability that you purchase exactly 8 boxes?
(b) What is the probability that you purchase at least 12 boxes?
(c) 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) $P(X=8)=\binom{8-1}{3-1}(0.17)^3(0.83)^5$
$=0.040640264$ ✓

b) $P(X\ge 12)=1-P(X\le 11)$ → plug into wolfram <3

$=1-\sum_{n=2}^{10}\left(nC_2\cdot(0.17)^3\cdot(1-0.17)^{x-2}\right)$

1-Sum[((x choose 2)*(0.17^3)*((1-0.17)^(x-2))),{x,2,10}]

$=1-0.283925321$
$=0.716075$ ✓

c) $\mu=\frac{r}{p}=\frac{3}{0.17}=17.64705882$ ✓

---

(1) NEGATIVE BINOMIAL DISTRIBUTION:

1. A company fabricates special-purpose robots, and records show that the probability is 0.10 that one of its new robots will require repairs during confirmation tests. What is the probability that the 12^{th} robot built in a month is the second to require repairs?

   Here, X is a random variable representing the number of new robots that are built in the month, and r is the number of robots that require repairs. Therefore, we have $X=12$, $r=2$, and $p=0.10$.

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

$P(X=12)=\binom{12-1}{2-1}(0.10)(1-0.10)^{12-2}$

$=0.0384$

❶ When exposed to a certain disease, the probability that a child will become infected is 0.30. What is the probability that the tenth child exposed to the disease will be fourth to catch it?

X = # of B-trials until r successes
x = # of children exposed
r = 4 (10th child is the 4th "success" (exposure))

X → random variable (# of children exposed)
r → # of children diseased

X = 10, r = 4, p = 0.30

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

$P(X=10)=\binom{10-1}{4-1}(0.3)^4(1-0.3)^{10-4}$

$=84(0.0081)(0.1177)=0.08$

(2) HYPERGEOMETRIC DISTRIBUTION

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?

We have $N=42$ possible number to choose from, we are interested in determining the probability of matching $X=4$ of these numbers, we drew $n=6$ numbers and there are $M=6$ "correct" numbers that we could potentially match.

$$P(X=4)=\frac{\binom{6}{4}\binom{42-6}{6-4}}{\binom{42}{6}}=0.0018$$

2. As part of an air-pollution survey, an inspector decides to examine the exhaust of six of a company's 24 trucks. If four of the company's trucks emit excessive amounts of pollutants, what is the probability that none of them will be included in the inspector's sample?

X = 0 ← want!
n = 6 ← sample size
N = 24 ← total
M = 4 ← unwanted

$$P(X=0)=\frac{\binom{4}{0}\binom{24-4}{6}}{\binom{24}{6}}=0.2880$$

∴ 28.8% chance none of the trucks will emit excessive pollutants in the inspector's sample

❶ An athlete conceals two performance enhancing tablets in a bottle containing eight vitamin pills that look similar. If a drug surveillance scheme is in operation that involves randomly sampling three of these pills, what is the probability that cheating will be detected? Assume that the analysis of the tablets is not error prone.

2 illegal tablets + 8 vitamin pills, pull 3

P(illegal tablet is selected?)

N = 10 total tablets
M = 2 total illegal
n = 3 select 3

$P(X\ge 1)$?

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

$=1-\frac{\binom{2}{0}\binom{10-2}{3-0}}{\binom{10}{3}}$

$=1-0.46$

$\simeq 0.53$

❷ Eight cards are chosen at random from a well-shuffled pack. What is the probability of obtaining three spades, two hearts, two diamonds and one club?

3 spades
2 hearts
2 diamonds
1 club
→ 8 cards

- also a counting problem
- denominator is total # of outcomes

$$P(3S,2H,2D,1C)=\frac{\binom{13}{3}\binom{13}{2}\binom{13}{2}\binom{13}{1}}{\binom{52}{8}}\simeq 0.03$$

Lec9 - Continuous Random Variables, and Probability Density Functions

Yuxi Qin
yuxiqin.ca

Monday, September 26, 2022 1:30 PM

POISSON DISTRIBUTION
Let X be a random variable that follows a Poisson distribution with mean u. Then:

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

- 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 $n \to \infty$, $p \to 0$ $\Rightarrow$ $np = \lambda$ 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)? -> X is called a **discrete random variable**

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

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

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

1. $f(x) \ge 0$ for all x, and
2. $\int_{-\infty}^{\infty} f(x)dx = 1.$

PROBABILITIES

- For a continuous random variable, X, the following hold:
  1. $P(X = x) = 0$ for all x,
  2. $P(X \le x) = P(X < x)$ for all x,
  3. $P(a \le X \le b) = F(b) - F(a) = \int_a^b f(x)dx,$
- 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 $4.9 \le x \le 5.1$. 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,

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

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 $x \ge 12.5$.

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,

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

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.

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.

(a) If you eat a 28-gram chocolate bar, find the probability that you will have eaten at least 3 insect fragments.

(b) 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.

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

a) at least:

$P(X \ge 3) = 1 - P(X \le 2)$
$= 1 - [P(X=0) + P(X=1) + P(X=2)]$
$= 1 - \left[\frac{e^{-1.79}(1.79^0)}{0!} + \frac{e^{-1.79}(1.79)^1}{1!} + \frac{e^{-1.79}(1.79)^2}{2!}\right]$
($x=0$, $x=1$, $x=2$)
$= 1 - (0.166960169 + 0.298858703 + 0.267478539)$
$= 1 - 0.733297411$
$= 0.266702588$ ✓

b) $P(X=0) = \frac{e^{-1.79}(1.79^0)}{0!} = 0.166960169$ p

$P(X!=0) = 1 - P(X=0) = 0.83303983$ $1-p$

Binomial distribution: → x successes given n trials
$P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}$

$P(X=3) = \binom{14}{3} (0.166960169)^3 (0.83303983)^{11}$
$= (364)(0.00465413)(0.134067476) = 0.227124156$ ✓

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.

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

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

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

Probability Distribution Function (PDF)
↳ Probabilities
$P(a \le X \le b) = F(b) - F(a) = \int_a^b f(x)dx$

a) $P(97 \le X \le 235)$
$= \int_{97}^{235} \left(\frac{x}{100^2}\right) e^{\frac{-x^2}{20000}} dx$
$= \frac{1}{10000} \int_{97}^{235} x e^{\frac{-x^2}{20000}}$
$u = -\frac{x^2}{20000}$ $du = -\frac{x}{10000}dx$
$= -\int e^u du$
$= -e^u + C$
sub in u
$= -e^{\frac{-x^2}{20000}}\Big|_{97}^{235}$
$= 0.561508377$ ✓

b) 72% of the time $\ge a$
$= 1 - 0.72 \le a$
↳ so we can start @ $a > 0$

$P(0 < x < a)$
$1 - 0.72 = \int_0^a \left(\frac{a}{\theta^2}\right) e^{-\frac{a^2}{2\theta^2}} d$
$\frac{7}{25} = \int_0^a \frac{ae^{-\frac{a^2}{20000}}}{10000} da$

$\frac{1}{10000}\int_0^a xa\,da$
$u = -\frac{a^2}{20000}$ $du = -\frac{a}{10000}dx$
$= -\int e^u du$
sub u
$= -e^{-\frac{a^2}{20000}}\Big|_0^a$
$= -e^{-\frac{a^2}{20000}} - (-e^0)$
$= 1 - e^{-\frac{a^2}{20000}}$

$1 - e^{-\frac{a^2}{20000}} = \frac{7}{25}$
subtract 1, BS $\times -1$
$e^{-\frac{a^2}{20000}} = \frac{18}{25}$
take reciprocal of BS
$e^{\frac{a^2}{20000}} = \frac{25}{18}$
ln BS → $\ln(e^x) = x$
$\frac{a^2}{20000} = \ln\left(\frac{25}{18}\right)$
$a^2 = 20000 \ln\left(\frac{25}{18}\right)$
SQRT BS
$a = \pm 100\sqrt{2\ln\left(\frac{25}{18}\right)}$ take positive ($a > 0$) & round to 2 dec.
$= \pm 81.056$ $= 81.06$ ✓

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

Tuesday, September 27, 2022 12:58 PM

Yuxi Qin
yuxiqin.ca

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:
  - $f(x) \geq 0 \quad \text{for all } x,$
  - $\int_{-\infty}^{\infty} f(x)dx = 1.$

PROBABILITIES

- For a continuous random variable, X, the following hold:
  - $P(X = x) = 0 \quad \text{for all } x,$
  - $P(X \leq x) = P(X < x) \quad \text{for all } x,$
  - $P(a \leq X \leq b) = F(b) - F(a) = \int_a^b f(x)dx,$
- Where a and b are constants

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

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

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

- Note that: $\mathbb{V}ar[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

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

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

Determine the following:

1. P(X < 1.8) $= F(1.8) = 0.25(1.8) + 0.5$ $= 0.95$
2. P(X > -1.5) $= 1 - P(< -1.5) = 1 - F(-1.5)$ $= 1 - [0.25(-1.5) + 5]$ $= 1 - 0.125$ $= 0.875$
3. P(X < -2) $= 0$
4. 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:

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

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

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

CM 3.1

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

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

(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.

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

a) $P(X > 2.44)$

$F(2.44) = 1 - \frac{1}{32}(2.44^2 - 4)$

$= 0.93895$ ✓

b) $P(0.99 < X < 2.22)$

$= F(2.22) - F(0.99)$

$= \frac{1}{32}(2.22^2 - 4) - 0$

$= 0.0290125$ ✓

c) $\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x)\,dx$

$= \int_2^6 \frac{x^2}{16}\,dx$

$= \frac{x^3}{48}\,dx \Big|_2^6$

$= \frac{6^3}{48} - \frac{2^3}{48} = \frac{208}{48}$ ✓

$\mathbb{E}(X^2) = \int_2^6 \frac{x^3}{16}\,dx$

$= \frac{x^4}{64}\Big|_2^6$

$= \frac{6^4}{64} - \frac{2^4}{64}$

$= \frac{1280}{64} = 20$

d) $V(X) = E(X^2) - E(X)^2$

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

$= \frac{11}{9}$ ✓

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 Var[X]

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

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

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

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.

1. Find c.
2. Find the cdf, $F(x)$.
3. What is $P(1 \leq X \leq 1.5)$?
4. Find $\mathbb{E}[X]$.
5. Find $\mathbb{V}\text{ar}[X]$.

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) $P(1 \leq X \leq 1.5)$?

$F(1.5) - F(1)$

$= \frac{1}{8}\left(\frac{3}{2}\right)^3 - \frac{1}{8}\left(\frac{3}{2}\right)^3$

$= \frac{1}{8}\left(\frac{27}{8}\right) - \frac{1}{8}$

$= \frac{27}{64} - \frac{8}{64} = \frac{19}{64}$

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

$\int_0^2 x\frac{3}{8}\cdot x^2 dx = \int_0^2 \frac{3}{8}x^3\,dx$

$= \frac{3}{32}x^4\Big|_0^2 = \frac{3}{25}(2^4) = \frac{3}{2}$

e) Find $\text{Var}[X]$

$E[X^2] = \int_0^2 x^2\frac{3}{8}\cdot x^2 dx$

$= \int_0^2 \frac{3}{8}x^4 dx = \frac{3}{40}x^5\Big|_0^2$

$= \frac{3}{40}(2^5) = \frac{12}{5}$

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

$= \frac{12}{5} - \left(\frac{13}{2}\right)^2 = \frac{3}{20}$

EXAMPLE 2:

The cdf of a continuous random variable X is given by

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

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

a) $P(X \leq 2.6) = F(2.6)$

$= 1 - e^{-2.6}$

≈ 0.926

b) $P(1 < x < 4)$

$= F(4) - F(1)$

$= 1 - e^{-4} - (1 - e^{-1})$

≈ 0.350

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

$= e^{-x}$

Lec11 - Normal Distribution

Yuxi Qin
yuxiqin.ca

Friday, September 30, 2022 9:57 PM

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

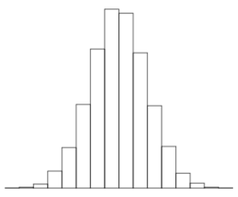
If a random variable X follows a normal distribution with parameters **μ** and **σ** , then we write **X ~ N (μ, σ²)** and the pdf of X is

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

- μ = **mean** of X
- σ = **standard deviation** of X
- σ^2 = **variance** of X

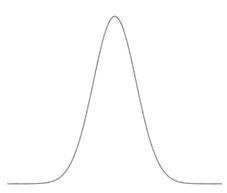
HISTOGRAM

- 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

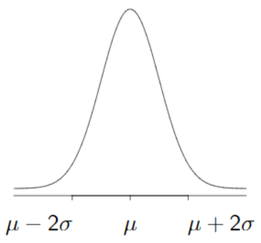
- These 10k values can also be represented like this:



- 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 μ and standard deviation σ



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

NORMALLY DISTRIBUTED DATA

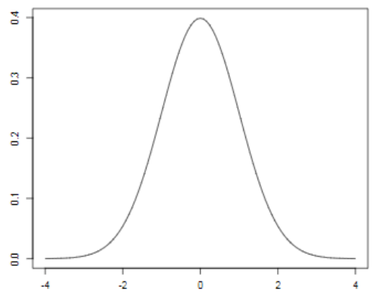
Consider the following information - we measure the heights of women in Ireland. We assume their heights are normally distributed with mean 1.62m and standard deviation 0.11m. It follows that:

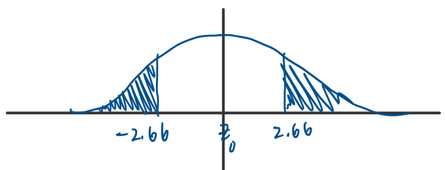
- The median and mode of the height of women is 1.62m
- The same amount of women are 1.52m tall as are 1.72m tall
- The same amount of women are 1.60m tall as are 1.64m tall
- 95% of women are between 1.40m and 1.84m tall

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 +- 1.96 of the mean

FROM NORMAL TO STANDARD NORMAL

- Let X ~ N (μ, σ^2) and Z ~ N(0.1), then

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

- Therefore, to compute P(X <= 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

NORMAL DISTRIBUTION: EXAMPLE 1:

1. Let $X \sim \mathcal{N}(2, 0.15^2)$. Find the $P(X \leq 2.1)$.

$\mu = 2$
$\sigma^2 = 0.15^2$
$\sigma = 0.15$

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

$P(X \leq 2.1) = P\left(z \leq \frac{2.1 - 2}{0.15}\right)$
$= P(z \leq 0.67)$
$= 0.748571$

0.67

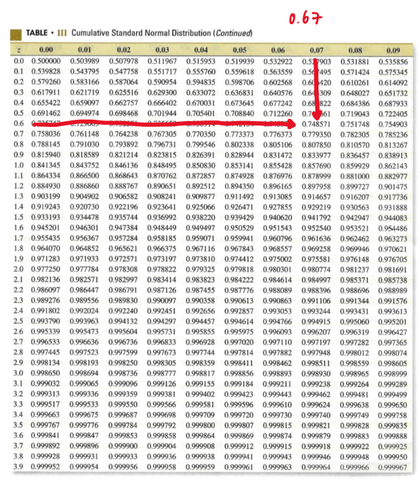


TABLE • III Cumulative Standard Normal Distribution (*Continued*)

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500000	0.503989	0.507978	0.511967	0.515953	0.519939	0.523922	0.527903	0.531881	0.535856
0.1	0.539828	0.543795	0.547758	0.551717	0.555760	0.559618	0.563559	0.567495	0.571424	0.575345
0.2	0.579260	0.583166	0.587064	0.590954	0.594835	0.598706	0.602568	0.606420	0.610261	0.614092
0.3	0.617911	0.621719	0.625516	0.629300	0.633072	0.636831	0.640576	0.644309	0.648027	0.651732
0.4	0.655422	0.659097	0.662757	0.666402	0.670031	0.673645	0.677242	0.680822	0.684386	0.687933
0.5	0.691462	0.694974	0.698468	0.701944	0.705401	0.708840	0.712260	0.715661	0.719043	0.722405
0.6	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.748571	0.751748	0.754903
0.7	0.758036	0.761148	0.764238	0.767305	0.770350	0.773373	0.776373	0.779350	0.782305	0.785236
0.8	0.788145	0.791030	0.793892	0.796731	0.799546	0.802338	0.805106	0.807850	0.810570	0.813267
0.9	0.815940	0.818589	0.821214	0.823815	0.826391	0.828944	0.831472	0.833977	0.836457	0.838913
1.0	0.841345	0.843752	0.846136	0.848495	0.850830	0.853141	0.855428	0.857690	0.859929	0.862143
1.1	0.864334	0.866500	0.868643	0.870762	0.872857	0.874928	0.876976	0.878999	0.881000	0.882977
1.2	0.884930	0.886860	0.888767	0.890651	0.892512	0.894350	0.896165	0.897958	0.899727	0.901475
1.3	0.903199	0.904902	0.906582	0.908241	0.909877	0.911492	0.913085	0.914657	0.916207	0.917736
1.4	0.919243	0.920730	0.922196	0.923641	0.925066	0.926471	0.927855	0.929219	0.930563	0.931888
1.5	0.933193	0.934478	0.935744	0.936992	0.938220	0.939429	0.940620	0.941792	0.942947	0.944083
1.6	0.945201	0.946301	0.947384	0.948449	0.949497	0.950529	0.951543	0.952540	0.953521	0.954486
1.7	0.955435	0.956367	0.957284	0.958185	0.959071	0.959941	0.960796	0.961636	0.962462	0.963273
1.8	0.964070	0.964852	0.965621	0.966375	0.967116	0.967843	0.968557	0.969258	0.969946	0.970621
1.9	0.971283	0.971933	0.972571	0.973197	0.973810	0.974412	0.975002	0.975581	0.976148	0.976705
2.0	0.977250	0.977784	0.978308	0.978822	0.979325	0.979818	0.980301	0.980774	0.981237	0.981691
2.1	0.982136	0.982571	0.982997	0.983414	0.983823	0.984222	0.984614	0.984997	0.985371	0.985738
2.2	0.986097	0.986447	0.986791	0.987126	0.987455	0.987776	0.988089	0.988396	0.988696	0.988989
2.3	0.989276	0.989556	0.989830	0.990097	0.990358	0.990613	0.990863	0.991106	0.991344	0.991576
2.4	0.991802	0.992024	0.992240	0.992451	0.992656	0.992857	0.993053	0.993244	0.993431	0.993613
2.5	0.993790	0.993963	0.994132	0.994297	0.994457	0.994614	0.994766	0.994915	0.995060	0.995201
2.6	0.995339	0.995473	0.995604	0.995731	0.995855	0.995975	0.996093	0.996207	0.996319	0.996427
2.7	0.996533	0.996636	0.996736	0.996833	0.996928	0.997020	0.997110	0.997197	0.997282	0.997365
2.8	0.997445	0.997523	0.997599	0.997673	0.997744	0.997814	0.997882	0.997948	0.998012	0.998074
2.9	0.998134	0.998193	0.998250	0.998305	0.998359	0.998411	0.998462	0.998511	0.998559	0.998605
3.0	0.998650	0.998694	0.998736	0.998777	0.998817	0.998856	0.998893	0.998930	0.998965	0.998999
3.1	0.999032	0.999065	0.999096	0.999126	0.999155	0.999184	0.999211	0.999238	0.999264	0.999289
3.2	0.999313	0.999336	0.999359	0.999381	0.999402	0.999423	0.999443	0.999462	0.999481	0.999499
3.3	0.999517	0.999533	0.999550	0.999566	0.999581	0.999596	0.999610	0.999624	0.999638	0.999650
3.4	0.999663	0.999675	0.999687	0.999698	0.999709	0.999720	0.999730	0.999740	0.999749	0.999758
3.5	0.999767	0.999776	0.999784	0.999792	0.999800	0.999807	0.999815	0.999821	0.999828	0.999835
3.6	0.999841	0.999847	0.999853	0.999858	0.999864	0.999869	0.999874	0.999879	0.999883	0.999888
3.7	0.999892	0.999896	0.999900	0.999904	0.999908	0.999912	0.999915	0.999918	0.999922	0.999925
3.8	0.999928	0.999931	0.999933	0.999936	0.999938	0.999941	0.999943	0.999946	0.999948	0.999950
3.9	0.999952	0.999954	0.999956	0.999958	0.999959	0.999961	0.999963	0.999964	0.999966	0.999967

2. Let $X \sim \mathcal{N}(2, 0.15^2)$. Find the $P(X \geq 1.6)$.

$\mu = 2$
$\sigma^2 = 0.15^2$
$\sigma = 0.15$

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

$P(z \geq 1.6) = 1 - P(z \leq 1.6)$
$= 1 - P\left(z \leq \frac{1.6-2}{0.15}\right)$
$= 1 - P(z \leq -2.66)$
$= 1 - 0.996$
$= 0.04 = P(X \leq 1.6)$

we don't have access to ⊖ z-table, so cope & do an extra step

Alternative:
$P(x \geq 1.6) = 1 - P(X \leq 1.6)$
$= 1 - 0.004$
$= 0.996$

3. The substrate concentration (mg/cm^3) of influent to a reactor is normally distributed with $\mu = 0.30$ and $\sigma = 0.06$.

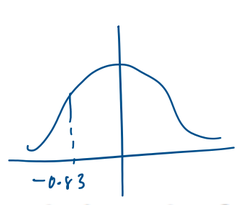
   a. What is the probability that the concentration exceeds 0.25?

   b. What is the probability that the concentration is at most 0.10?

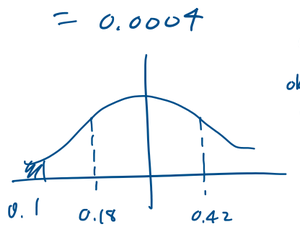
a) $N \sim (0.3, 0.06^2)$ $\mu = 3$, $\sigma = 0.06$

$P(x \geq 0.25) = 1 - P\left(z \leq \frac{0.25 - 0.30}{0.06}\right)$
$= 1 - P(z \leq -0.83)$
$= 1 - 0.2033$
$= 0.7967$

$= 1 - P(z \leq 0.83)$
$= 1 - 0.7967$
$= 0.2033$



b) $P\left(z \leq \frac{0.1 - 0.3}{0.06}\right)$
$= P(z \leq -3.33)$
$= 1 - P(z \leq 3.33)$
$= 1 - 0.99958$
$= 0.0004$



95% of observations are within 2σ of μ

❶ The height of women in a particular region follows a normal distribution with mean 1.55m and standard deviation 0.17m.

1. What proportion of women are smaller than 1.61m?
2. What proportion of women are taller than 1.8m?
3. What proportion of women are between 1.45m and 1.6m tall?

a) $P(x < 1.61) = P\left(z < \frac{1.61 - 1.55}{0.17}\right) = P(z < 0.3529) \approx 0.6368$

b) $P(x > 1.80) = 1 - P\left(z < \frac{1.80 - 1.55}{0.17}\right) = 1 - P(z < 1.4705) \approx 0.0707$

c) $\mu = 1.55$
$\sigma = 0.17$
$P = (x \leq x) = z\left(\frac{x-\mu}{\sigma}\right)$
$P(X \leq x) = P\left(z \leq \frac{x-\mu}{\sigma}\right)$

$P(x < 1.45)$
$P\left(z \leq \frac{1.45 - 1.55}{0.17}\right)$
$z \leq -0.588$

$P(z < 0.588) = 0.722$
$\therefore P(z < -0.588) = 1 - 0.722$
$= 0.278$

$(X \leq 1.6) = P\left(z \leq \frac{1.6 - 1.55}{0.17}\right)$
$= P(z \leq 0.294)$
$= 0.614$

$P(x < 1.6) - P(x < 1.45)$
$= 0.614 - 0.278$
≈ 0.336

❷ The wingspans of the males of a certain species of bird of prey form a normal distribution with mean 162.50cm and standard deviation 6.0cm. What is the probability that the wingspan of a randomly selected male will exceed 170cm?

Let X be the wingspan of a certain species of bird of prey

$P(X > 170) = 1 - P\left(z < \frac{170 - 162.50}{6.0}\right)$
$= 1 - P(z < 1.25)$
$= 1 - 0.8894$
≈ 0.1056

CM 3.2

Problem #2: IQs are known to be normally distributed with mean 100 and standard deviation 15.

(a) What percentage of people have an IQ lower than 88?

(b) Fill in the blank. 75% of the population have an IQ that is greater than _____ ?

NOTE: Do **not** use the first half of the normal table (i.e., page 742 in the textbook, with negative *z*-values) because it will **not** be provided with the tests.

$N(\mu = 100, \sigma = 15)$

a) $P(IQ < 88)$ $z = \frac{x-\mu}{\sigma}$
$= P\left(z < \frac{88 - 100}{15}\right)$
$= P(z < -0.8) = 0.788145$

$P(IQ < 88)$
$= 1 - P(z < -0.8) = 1 - 0.788145$
$= 0.211855$
$= 21.19$ ✓

b) $P(X > k) = 0.75$
$P\left(z > \frac{k-100}{15}\right) = 0.75$ → look @ z-tables & take closest value, don't interpolate!
$P(z > -0.67) = 0.75$ → take the negative version — note how we had to do 1 − (z-value) for other questions
$\frac{k-100}{15} = -0.67$
$k = 15(-0.67) + 100$
$= 89.95$ ✓

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.

approx. prob → normal approx. to binomial dist.

a) $n = 235$
$p = 0.21$

for binomial random var X: $\mu = np$, $\sigma^2 = np(1-p)$, $\sigma = \sqrt{np(1-p)}$ → $z = \frac{x - np}{\sqrt{np(1-p)}}$ → $z = \frac{x-\mu}{\sigma}$

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

$P(42 \to 58$ can be reworked$)$
$P(\overset{41.5}{\cancel{42}} \leq x \leq \overset{58.5}{\cancel{58}})$ → apply continuity correction?
$\Rightarrow P\left(\frac{41.5-\mu}{\sigma} < x < \frac{58.5-\mu}{\sigma}\right)$
$-1.2572 \to 1.26$, $1.4654 \to 1.47$
$= P(-1.26 < z < 1.47)$
$= P(z < 1.47) - [1 - P(z < 1.26)]$
$= 0.929219 - (1 - .896165)$
$= 0.825384$ ✓

at least = $P(X \geq 52)$

b) approx prob → $x \geq 52$
$P(X \geq \overset{51.5}{\cancel{52}})$ → use continuity correction
$= 1 - P\left(z < \frac{51.5 - \mu}{\sigma}\right)$
$= 1 - P(z < 0.344)$
$= 1 - 0.633072$
$= 0.366928$ ✓

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

Yuxi Qin
yuxiqin.ca

Monday, October 3, 2022 12:39 AM

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:

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

&

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

± 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: +/- 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 <= 45	X < 45.5
X < 45	X < 44.5
X >= 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

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

Based on the normal approximation,

$$P(X \le 2) = P\left(\frac{X - 5}{\sqrt{50(0.1)(0.9)}} \le \frac{2.5 - 5}{\sqrt{50(0.1)(0.9)}}\right) \approx P(Z < -1.18) = 0.119$$

$P(X \le 2) \simeq P\left(z \le \frac{2.5 - 5 - 50(0.1)}{\sqrt{50(0.1)(1-0.1)}}\right)$

$\simeq P\left(z \le \frac{-2.5}{\sqrt{4.5}}\right)$

$\simeq P(z \le -1.179)$

$= 0.119$

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

$$P(5 \le X \le 5) = P(4.5 \le X \le 5.5) \approx P\left(\frac{4.5 - 5}{2.12} \le Z \le \frac{5.5 - 5}{2.12}\right) = P(-0.24 \le Z \le 0.24) = 0.19$$

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 1×10^{-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

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

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

EXAMPLE 4.14

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

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

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?

$P(X \le 150)$

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

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

$P(X \le x) = P(X \le x + 0.5)$

$\simeq P\left[z \le \frac{x + 0.5 - np}{\sqrt{np(1-p)}}\right]$ → $np = (16 \cdot 10^6)(1 \cdot 10^{-5}) = 160$

$P(X \le 150) = P\left(z \le \frac{150 + 0.5 - 160}{\sqrt{160(1-10^{-5})}}\right)$

$= P(z \le -0.75)$

$P(z <$ $= 0.77337$

$1 - 0.7733$ 0.227

$\therefore P(X \le 150)$ 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.

Use $z = \frac{x - \mu}{\sigma}$

$\mu = 15$

a) σ ?

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

$x = 16$ ↑ desired

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

rearrange

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

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

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

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

1. prob of 1 shoe weighing > 16 ounces → success prob. used in 2)
2. prob of exactly 2 of 8 shoes succeedig

1) $P\left(z > \frac{16 - 15}{0.89}\right) = 1 - P(z > 1.12)$ → from z-tables:

1.1235

$= 1 - 0.868643$

$= 0.131357$ → prob. of success

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

$= 28(0.017254661)(0.429583349)$

$= 0.20754489$ ✓

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

Yuxi Qin
yuxiqin.ca

Friday, October 21, 2022 8:00 PM

THE EXPONENTIAL DISTRIBUTION

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

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

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

MEMORY-LESS PROPERTY

$X \backsim \exp(\lambda)$ 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

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

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*
- Ie. Events occur continuously and independently, at a constant average rate
- λ = 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 is 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?

a) ① ② ③ ④ ⑤ $\lambda = 93$ $\lambda = 130$

$E(X) = \frac{1}{\lambda}$ $\text{Var}(X) = \frac{1}{\lambda^2}$

$P(X \geq 55)$

$= \left[e^{(-\frac{1}{93})(55)}\right]^2 \left[e^{-\frac{1}{130}(55)}\right]^3$

$= (0.30642708)(0.28104729)$

$= 0.08611887$ ✓

b) 1 year = 52 weeks lasts 1 year

$0.97 = \left(e^{-\frac{1}{\lambda}}\right)^5$

$\sqrt[5]{0.97} = e^{-\frac{1}{\lambda}}$

$\ln\left(\sqrt[5]{0.97}\right) = -\frac{1}{\lambda}$

$-\frac{1}{\lambda} = -0.00609184$

$\lambda = 164.154$ ✓

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

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

B. The standard deviation of the time between successive arrivals

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

C. P(X <=4)

$P(X \leq 4) = \int_0^4 f(x)\,dx = \int_0^4 e^{-x}\,dx = \left[-e^{-x}\right]_0^4 \longrightarrow \int dx \rightarrow \int(e^{-x}, 0, 4) = 0.98168$

$= -e^{-4} - (-e^0) = 1 - e^{-4} = 0.982$

D. P(2<=x<=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:

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

Example 2.
Logons modelled by a poisson process with
λ = 25 log-ons per hour (60 mins)

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

$\lambda = 25$ per 60 mins

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

There are 10, 6-minute intervals in an hr

$\therefore$ our interval : 0.1 hr

random variable : X = distance b/n successive events

$f(x) = \lambda e^{-\lambda x}$

$P(X > 0.1) = \int_{0.1}^{\infty} f(x)\,dx$ ← prob. that 1st event occurs after 6 mins

$= \int e^a dx$

$= \frac{1}{c} e^{cx}$

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

$\lambda = 25$ $\quad e^{-\infty} \rightarrow 0$

$P(X > 0.1) = -\frac{25}{25} e^{-25x} \Big|_{0.1}^{\infty}$

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

$= 0.082$

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

$\frac{2}{60} = 0.033$ 2 min int. is 0.033 of 1 hr $\quad \frac{3}{60} = 0.05$ 3 min int. is 0.05 of 1 hr

$P(0.033 < x < 0.05) \longrightarrow -e^{-25x} \Big|_{0.033}^{0.05}$

$= \left[-e^{-25(0.05)}\right] - \left[-e^{-25(0.033)}\right]$

$= -0.287 - (-0.438)$

$= -0.287 + 0.438 = 0.151$

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

$P(X > x) = 0.9$ prob of 0.9 that the 1st event occurs after time x

$-e^{-25x} \Big|_x^{\infty} = 0.9$

$0 - (-e^{-25x}) = 0.9$

$e^{-25x} = 0.9$

natural log BS, solve for x → $\ln(e^x) = x$

$-25x = \ln(0.9) \longrightarrow x = \frac{\ln(0.9)}{-25}$

$x = 0.00421$

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

$\mu = \frac{1}{\lambda} = \frac{1}{25} = 0.04$ hrs

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

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

$\sigma = \sqrt{0.016} = 0.04$ hrs

Lec14 - joint probability distributions

Yuxi Qin
yuxiqin.ca

Friday, October 21, 2022 8:10 PM

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

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

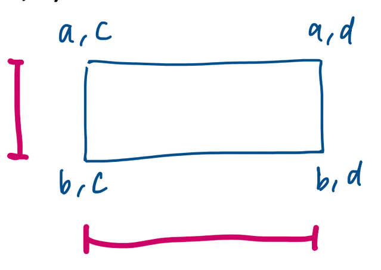
and then for any region $A \subseteq \mathbb{R}^2$ we have

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

CALCULATING PROBABILITIES

- We will be interested in calculating probabilities of the form $P(X,Y \in A)$.
- If A is rectangle with corner (a,c), (a,d), (b,c), and (b,d) then

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



- 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

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

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 $f_Y(y)$ → probability of each individual variable
- 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 dependent
- If they are independent, then their marginal probability distributions can be used to calculate the probabilities

individual probabilities!

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 $(x,y) \mid 0 < x < \frac{1}{2}, 1 < y < 2$. ⟶ domains

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

$$\int_1^2 \frac{3(\frac{1}{2})^2 y}{10} + \frac{3(\frac{1}{2})^3}{15} - [0+0]\cdot dy$$

$$\frac{3(2)^2}{80} + \frac{2}{40} - \left[\frac{3(1)^2}{80} + \frac{1}{40}\right] = \frac{12}{80} + \frac{4}{80} - \left[\frac{3}{80} + \frac{2}{80}\right] = \frac{16}{80} - \frac{5}{80} = \frac{11}{80}$$

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

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

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	X = 1
Y	−1	$\frac{1}{8}$	$\frac{1}{2}$
	0	0	$\frac{1}{4}$
	1	$\frac{1}{8}$	0

A. What is the marginal distribution of X?

$$P(X=-1) = \sum_{y \in S_y} = \frac{1}{8} + 0 + \frac{1}{8} = \frac{2}{8} = \frac{1}{4}$$
$$P(X=1) = \sum_{y \in S_y} = \frac{1}{2} + \frac{1}{4} + 0 = \frac{3}{4}$$

B. What is the marginal distribution of Y?

$$P(Y=-1) = \sum_{x \in S_x} = \frac{1}{8} + \frac{1}{2} = \frac{5}{8}$$
$$P(Y=0) = \sum_{x \in S_x} = 0 + \frac{1}{4} = \frac{1}{4}$$
$$P(Y=1) = \sum_{x \in S_x} = \frac{1}{8} + 0 = \frac{1}{8}$$

C. If the joint probability density of X and Y is given by

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{4}(2x+y) & \text{for } 0 < x < 1, 0 < y < 2 \\ 0 & \text{elsewhere} \end{cases}$$

Find the marginal distribution of Y

$$f_y(Y) = \int_{-\infty}^{\infty} f_{x,y}(x,y)dx = \int_0^1 \frac{1}{4}(2x+y)dx = \frac{1}{4}(x^2+xy)\Big|_0^1 = \frac{1+y}{4}$$

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 $f_{XY}(x,y) = 2 \times 10^{-6}\exp(-0.001x - 0.002y)$ 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

$$f_X(x) = \int_0^\infty 2 \times 10^{-6}e^{-0.001x-0.002y}dy = 0.001e^{-0.001x} \quad \text{for } x > 0$$

$$= 2\cdot10^{-6}\left[-\frac{1}{0.002}e^{-0.001x-0.002y}\right]\Big|_0^\infty = 2\cdot10^{-6}\left[0-\left(\frac{-1}{0.002}e^{-0.001x-0}\right)\right] = \frac{2\cdot10^{-6}}{0.002}e^{-0.001x} = 0.001e^{-0.001x}$$

The marginal probability density function of Y is

$$f_Y(y) = \int_0^\infty 2 \times 10^{-6}e^{-0.001x-0.002y}dx = 0.002e^{-0.002y} \quad \text{for } y > 0$$

$$= 2\cdot10^{-6}\left[\frac{1}{0.001}e^{-0.001x-0.002y}\right]\Big|_0^\infty = 2\cdot10^{-6}\left[0-\left(\frac{-e^{-0x-0.002y}}{0.001}\right)\right] = \frac{2\cdot10^{-6}}{-0.001}\left(e^{-0.002y}\right) = 0.002e^{-0.002y}$$

Therefore, $f_{XY}(x,y) = f_X(x)f_Y(y)$ for all x and y, and X and Y are independent.

To determine the probability requested, property (4) of Equation 5.8 can be applied along with the fact that each random variable has an exponential distribution. Therefore,

$$P(X > 1000, Y < 1000) = P(X > 1000)P(Y < 1000) = e^{-1}(1-e^{-2}) = 0.318$$

To show $f_{xy}(x,y) = f_x(x)f_y(y)$

$$= 0.001e^{-0.001x}(0.002e^{-0.002y}) = 2\cdot10^{-6}e^{-0.001x-0.002y} = f_x(x)f_y(y)$$

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 (mm^2) and that the distribution of Y is normal with mean 3.2 millimeters and variance 0.0036 (mm^2). Determine the probability that $10.4 < X < 10.6$ and $3.15 < Y < 3.25$.

Because X and Y are independent,

$$\begin{aligned} P(10.4 < X < 10.6, 3.15 < Y < 3.25) &= P(10.4 < X < 10.6) \times P(3.15 < Y < 3.25) \\ &= P\left(\frac{10.4-10.5}{0.05} < Z < \frac{10.6-10.5}{0.05}\right) \times P\left(\frac{3.15-3.2}{0.06} < Z < \frac{3.25-3.2}{0.06}\right) \\ &= P(-2 < Z < 2)P(-0.833 < Z < 0.833) \\ &= 0.568 \end{aligned}$$

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 = standard normal random variable
$\mu_x = 10.5 \to \sigma^2 = 0.0025$
$\mu_y = 3.2 \to \sigma^2 = 0.0036$

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

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

→ 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

Q: determine the value of x such that f(x,y)=cxy for:
0<x<3; 0<y<3
Is a joint probability density function

$$c\int_0^3 yx^2\Big|_0^3 = c\int_0^3 y\frac{(9)}{2} - 0$$

integrate w.r.t. x — "c" is outside all this because it's a constant

$$c\int_0^3 y\left(\frac{9}{2}\right) = c\left[\frac{y^2}{2}\left(\frac{9}{2}\right)\Big|_0^3\right] = c\left(\frac{9}{2}\right)\left(\frac{9}{2}\right) - 0$$

integrate w.r.t. y

$$c\left(\frac{81}{4}\right) = 1$$
$$c = \frac{4}{81}$$

$$\int_{y\text{-lower}}^{y\text{-upper}}\int_{x\text{ lower}}^{x\text{ upper}} f(x)\,dx\,dy$$

E(X)?

$$E(X) = \int_{-\infty}^{\infty} x f(x)\,dx$$

$$\begin{aligned} E(x) &= \frac{4}{81}\int_0^3\int_0^3 x(xy)\,dx\,dy \\ &= \frac{4}{81}\int_0^3\int_0^3 x^2y\,dxdy \\ &= \frac{4}{81}\int_0^3 \frac{x^3}{3}y\,dy\Big|_0^3 \\ &= \frac{4}{81}\int_0^3 9y\,dy \\ &= \frac{4}{81}\frac{9y^2}{2}\Big|_0^3 \\ &= \frac{4}{81}\left(\frac{9(9)}{2} - 0\right) = 2 \end{aligned}$$

Marginal Probability distribution of x?

$$\begin{aligned} f_x(x) &= \int_0^3 \frac{4}{81}(xy)\,dy \\ &= \frac{4}{81}x\frac{y^2}{2}\Big|_0^3 \\ &= \frac{4}{81}\left(\frac{9x}{2} - 0\right) \\ &= \frac{36x}{162} = \frac{2}{9}x \end{aligned}$$

$f_x(x)$ or $f_y(y)$ marginal probability: $\int_{lower}^{upper} f(x)\,dx$

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

joint density function — $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{x,y}(x,y)dxdy = 1$

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

a) $c = ?$

$$\int_0^\infty\int_0^x ce^{-7x-3y}\,dydx = 1$$
$$1 = \int_0^\infty -\frac{ce^{-3y}(e^{-7x-1})}{3}dx$$
$$1 = \frac{c}{70}$$

→ what happens here?

$c = 70$ ✓

b) $P(x < \frac{1}{6}, y < 2)$

$$\int_0^{\frac{1}{6}}\int_0^x 70e^{-7x-3y}\,dydx$$

→ mmm Symbolab & desmos calc

$= 0.4026989936$ ✓

CM 3.7

Problem #7: 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 $P(X < 2, Y < \frac{1}{6})$.
(b) Find the marginal probability distribution of X.

probability joint density

$c = 70$ $\quad f(x,y) = ce^{-7x-3y}$, $0 < y < x$

$y < x < 2$
$0 < y < \frac{1}{6}$

a) $P(x < 2, y < \frac{1}{6})$

$$= \int_0^{\frac{1}{6}}\int_0^x 70e^{-7x-3y}\,dydx + \int_{\frac{1}{6}}^2\int_0^{\frac{1}{6}} 70e^{-7x-3y}\,dydx$$

~ mm desmos ~

$= 0.8111233065559$ ✓

b) $f_x(x) = \int_0^x f(x,y)\,dy$

$$= \int_0^x 70e^{-7x-3y}\,dy$$

~ mmm symbolab … ~

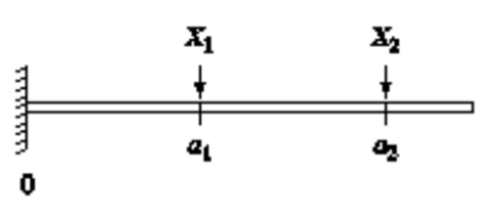
$$= -\frac{70}{3}e^{-7x}(e^{-3x}-1)$$ ✓

Desmos graphing:

```
\int_{0}^{\frac{1}{6}}\int_{0}^{x}70e^{\left(-7x-3y\right)}dydx+\int_{\frac{1}{6}}^{2}\int_{0}^{\frac{1}{6}}70e^{\left(-7x-3y\right)}dydx
```

CM 3.9

Problem #9: If two loads are applied to a cantilever beam as shown in the figure below, the bending moment at 0 due to the loads is $a_1X_1 + a_2X_2$.



Suppose that X_1 and X_2 are independent random variables with means 3 and 5 kips, respectively, and standard deviations 0.7 and 1.1 kips, respectively. Suppose that $a_1 = 5$ ft and $a_2 = 9$ ft.

(a) Find the expected value of the bending moment.
(b) Find the standard deviation of the bending moment.
(c) If X_1 and X_2 are normally distributed, what is the probability that the bending moment will exceed 71 kip-ft?

$a_1 = 5 \quad E(X_1) = 3 \quad V(X_1) = 0.7$
$a_2 = 9 \quad E(X_2) = 5 \quad V(X_2) = 1.1$

$E(aX+bY) = aE[X] + bE[Y] \qquad Var[aX+bY] = a^2Var[X] + b^2Var[Y]$

a) $E(5X_1 + 9X_2)$
$= E(5X_1) + E(9X_2)$
$= 5(3) + 9(5)$
$= 60$ ✓ mean

b) $V(5X_1 + 9X_2)$
$= V(5X_1) + V(9X_2)$
$= 5^2(0.7)^2 + 9^2(1.1)^2$
$= 12.25 + 98.01$
$= 110.26$ variance

$\sigma = \sqrt{\sigma^2}$
$\sigma = \sqrt{110.26}$
$= 10.50047618$ ✓ std dev.

c) $P(X > 71)$ $\qquad z = \frac{x-\mu}{\sigma}$
$P(5X_1 + 9X_2 > 71)$
$= 1 - P\left(z \leq \frac{71-60}{10.5005}\right)$ $\quad$ 1.04757 → 1.05
$= 1 - 0.853141$
$= 0.146859$ ✓

Lec15 - covariance and correlation, linear functions

Yuxi Qin
yuxiqin.ca

Saturday, October 22, 2022 1:39 AM

X, Y are random variables!

EXPECTATION

CONTINUOUS CASE:

- Suppose that X and Y are two continuous random variables with joint probability density function $f_{X,Y}(x,y)$. Let h be a scalar function of two variables. Then the expected value of h(X,Y) is given by

$$E[h(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y)f(x,y)dxdy$$

COVARIANCE

- Suppose that X and Y are two random variables with means $E[X] = \mu_X$ and $E[Y] = \mu_Y$.
- The covariance of X and Y is defined to be

$$\text{Cov}[X,Y] = E[(X-\mu_X)(Y-\mu_Y)]$$

- Furthermore, for the random variables X and Y - the covariance can be found using: (expanded out — these are equivalent statements)

$$\text{Cov}(X,Y) = E[XY] - E[X]E[Y]$$

- Covariance is a measure of how the two variables X and Y are related
- If in one experiment, X and Y are such that both tend to be larger/smaller than their mean, then the covariance will be positive (proportional)
- If X tends to be large when Y is small, or X tends to be small when Y is large, then the covariance will by negative (inversely proportional)
- The actual value of the covariance depends on the units of the measurements of X and Y

CORRELATION COEFFICIENT

- If X and Y are two random variables then the correlation coefficient of X and Y is

$$\text{Corr}(X,Y) = \rho = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

- It can be shown that -1 <= p <= 1
- The correlation coefficient between two random variables measures the strength of the linear relationship between them
- |p| = 1 iif Y=aX+b with a!=0
- Changing the units of X and/or Y does not change the correlation

Theorem: if X and Y are two independent random variables, g and h are scalar functions of the variables, then

$$E[g(X),h(Y)] = E[g(X)]E[h(Y)]$$

- It can be shown -1 <= p <= 1

Corollary - if X and Y are two independent random variables then their correlation coefficient is equal to 0

- Note: the converse of the Corollary is not true
- It is possible for two random variables to have a p=0 but not be independent

EXPECTATION AND VARIANCE

- Suppose that X and Y are two random variables and a and b are fixed constants then

$$E[aX+bY] = aE[X] + bE[Y]$$

$$\text{Var}[aX+bY] = a^2\text{Var}[X] + b^2\text{Var}[Y] + 2ab\text{Cov}[X,Y]$$

$$\text{Var}[aX-bY] = a^2\text{Var}[X] + b^2\text{Var}[Y] - 2ab\text{Cov}[X,Y]$$

Joint possibility density — double integrals
— get individual possibilities through marginal

EXERCISES:

If the joint probability density of X and Y is given by

$$f_{X,Y}(x,y) = \begin{cases} \frac{2}{7}(x+2y) & \text{for } 0<x<1,\ 1<y<2 \\ 0 & \text{elsewhere} \end{cases}$$

- Find the expected value of $g(X,Y) = X/Y^3$

f(x)

$$E[X/Y^3] = \int_1^2\int_0^1 \frac{2x(x+2y)}{7y^3}dxdy = \frac{2}{7}\int_1^2\left(\frac{1}{3y^3}+\frac{1}{y^2}\right)dy = \frac{15}{84}$$

Sub into calculator — $\int dx \to \frac{2}{7}\int\left(\frac{1}{3y^3}+\frac{1}{y^2}\right), 1, 2$ $= \frac{5}{28}$

If the joint probability density of X and Y is given by

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

- Find the E[X/Y]

$$E[X/y] = \int_0^1\int_0^y \frac{x}{y^2}dxdy$$

$$= \int_0^1 \frac{x^2}{2y^2}\Big|_0^y dy = \frac{y}{2}\Big|_0^1 = \frac{1}{2}$$

Q: X and Y are independent, random variables with:
E(X) = 2; V(X) = 5; E(Y) = 6; V(Y) = 8

Determine the following:

A. E(3x+2y)

aE bE

$= a\,E(x) + b\,E(y)$

$= 3(2) + 2(6) = 18$

B. V(3x+2y)

$= a^2\,Var[x] + b^2\,Var[x]$

$= 9(5) + 4(8) = 77$

μ σ^2

C. P(3x+2y<18) $\simeq N(18, 77)$

$= P\left(z \le \frac{18-18}{\sqrt{77}}\right) = P(z \le 0)$

$= 0.50$

D. P(3x+2y<28)

$\bar{x} = 28$

$\mu = 18$

$\sigma^2 = 77$

$= P\left(z \le \frac{28-18}{\sqrt{77}}\right)$

$= P(z \le 1.1396)$

$= 0.873$

5-65. A plastic casing for a magnetic disk is composed of two halves. The thickness of each half is normally distributed with a mean of 2 millimeters, and a standard deviation of 0.1 millimeter and the halves are independent.

(a) Determine the mean and standard deviation of the total thickness of the two halves.

(b) What is the probability that the total thickness exceeds 4.3 millimeters?

T = total thickness = X + Y

→ halves are independent

E(T) = E(X) + E(Y)

$\mu = 2 = E(X) = E(Y)$

$\sigma^2 = 0.1 = Var(x) = Var(y)$

a) E(T) = 2+2 = 4 mean

$V(T) = a^2\,Var(X) + b^2\,Var(y)$

$= (0.1)^2 + (0.1)^2$

$\sigma^2 = 0.02$

$\sigma = \sqrt{0.02} = 0.1414$ variance

b) $P(T > 4.3) = 1 - P(T < 4.3)$

$= 1 - P\left(z \le \frac{4.3-4}{0.1414}\right)$

$= 1 - P(z \le 2.12)$

$= 1 - 0.983$

$= 0.17$

$\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$

$4 \pm 1.96\,\sigma$

4 ± 0.277

$(3.723,\ 4.277)$

CM 3.8 — Also on MT2 ! (Find variant in 2019 MT2)

Problem #8: Suppose that X and Y have the following joint probability density function.

$$f(x,y) = \frac{3}{220}y, \quad 0<x<6,\ y>0,\ x-2<y<x+2$$

(a) Find $E(XY)$.

(b) Find the covariance between X and Y.

$f(x,y) = \frac{3}{220}y$

$0 < x < 6$

$y > 0$

$x-2 < y < x+2$

a) $E(XY) = \int_0^2\int_0^{x+2} xy\,\frac{3y}{220}\,dy\,dx + \int_2^6\int_{x-2}^{x+2} xy\,\frac{3y}{220}\,dy\,dx = \frac{5218}{275}$ ✓

b) $E(X) = \int_0^2\int_0^{x+2} x\,\frac{3y}{220}\,dy\,dx + \int_2^6\int_{x-2}^{x+2} x\,\frac{3y}{220}\,dy\,dx = \frac{433}{110}$

$E(Y) = \int_0^2\int_0^{x+2} y\,\frac{3y}{220}\,dy\,dx + \int_2^6\int_{x-2}^{x+2} y\,\frac{3y}{220}\,dy\,dx = \frac{239}{55}$

$\text{Cov}(X,Y) = \frac{5218}{275} - \left(\frac{433}{110}\right)\left(\frac{239}{55}\right) = 1.869256198$ ✓

E(XY) - E(X)E(Y)

plug into wolfram!

Wolfram:

Integrate[xy*Divide[3y,332],{y,0,x+4},{x,0,4}] + Integrate[xy*Divide[3y,332],{y,x-4,x+4},{x,4,5}]

Wolfram:

Integrate[xDivide[3y,332],{y,0,x+4},{x,0,4}] + Integrate[xDivide[3y,332],{y,x-4,x+4},{x,4,5}]

Integrate[yDivide[3y,332],{y,0,x+4},{x,0,4}] + Integrate[yDivide[3y,332],{y,x-4,x+4},{x,4,5}]

Lec16 - numerical summaries of data, stem and leaf diagrams

Yuxi Qin
yuxiqin.ca

Saturday, October 22, 2022 2:31 AM

TYPES OF DATA
- There are many types of data, which can be categorical or interval
- Each corresponds to different types of information

CATEGORICAL DATA
- Binary (categorical data that can take exactly one of two possible values)
  - Are you diabetic? Yes/no
- Nominal (labeled variables without providing quantitative data)
  - Eye colour: blue, green, brown, gray
  - Nationality: irish, korean, french, polish, canadian, etc.
- Ordinal (categorical data with set order or scale)
  - Dose: less tha 10mg; 10mg to 20mg; more than 20mg
  - Height: less than 1.5m; 1.5m to 1.8m; more than 1.8m

INTERVAL DATA
- **Continuous** data arises when all values are possible inside some interval (included open-ended intervals) — floats/decimals
  - Volume of water inside a glass
  - Distance between cities
- **Discrete** data arises when there are a finite number of possible values (in closed intervals) or "space" between values that are unobtainable — integers
  - Number of bacteria on a piece of raw meat
  - Number of hurricanes that will hit land each year
- Other examples:
  - Height - could be any value between 0 and 2.3m
  - Age - could take any value between 0 and 110 years
  - Weight - could take any value between 0.3 and 220kg

SUMMARIZING DATA
- Looking at data is a very useful first step
- Often, a picture is the best summary of data
- Common visualization techniques include: steam and leaf displays, histograms, and boxplots
- We can also summarize data using a number of quantitative measurements

DESCRIPTIVE STATISTICS I
- Suppose that we observe data in the form: x1, x2, … , xn
- We call this observed data a **sample**
- Oftentimes, the **sample** comes from a **population**
- The **mean** (average) of the sample, denoted X' is given by $\bar{X}$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}.$$

- The **standard deviation** of this sample, denoted s, is given by σ, s

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

σ^2, s^2

- The **variance** is the standard deviation squared, s^2
- The **range** of a variable is the 'distance' between the minimum and maximum values
- The **mode** is the most frequently occurring observation
- 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
  - That is, it's the 'distance' over which the 'middle half' of the data is spread,
  - Or the 'distance' between the 25th and 75th percentiles

MORE EXAMPLES:

TABLE • 6-2 Compressive Strength (in psi) of 80 Aluminum-Lithium Alloy Specimens

105	221	183	186	121	181	180	143
97	154	153	174	120	168	167	141
245	228	174	199	181	158	176	110
163	131	154	115	160	208	158	133
207	180	190	193	194	133	156	123
134	178	76	167	184	135	229	146
218	157	101	171	165	172	158	169
199	151	142	163	145	171	148	158
160	175	149	87	160	237	150	135
196	201	200	176	150	170	118	149

Stem	Leaf	Frequency
7	6	1
8	7	1
9	7	1
10	5 1	2
11	5 8 0	3
12	1 0 3	3
13	4 1 3 5 3 5	6
14	2 9 5 8 3 1 6 9	8
15	4 7 1 3 4 0 8 8 6 8 0 8	12
16	3 0 7 3 0 5 0 8 7 9	10
17	8 5 4 4 1 6 2 1 0 6	10
18	0 3 6 1 4 1 0	7
19	9 6 0 9 3 4	6
20	7 1 0 8	4
21	8	1
22	1 8 9	3
23	7	1
24	5	1

Stem: Tens and hundreds digits (psi); Leaf: Ones digits (psi).

FIGURE 6-4 Stem-and-leaf diagram for the compressive strength data in Table 6-2.

Stem-and-leaf of Strength

N = 80	Leaf	Unit = 1.0
1	7	6
2	8	7
3	9	7
5	10	1 5
8	11	0 5 8
11	12	0 1 3
17	13	1 3 3 4 5 5
25	14	1 2 3 5 6 8 9 9
37	15	0 0 1 3 4 4 6 7 8 8 8 8
(10)	16	0 0 0 3 3 5 7 7 8 9
33	17	0 1 1 2 4 4 5 6 6 8
23	18	0 0 1 1 3 4 6
16	19	0 3 4 6 9 9
10	20	0 1 7 8
6	21	8
5	22	1 8 9
2	23	7
1	24	5

FIGURE 6-6 A typical computer-generated stem-and-leaf diagram.

Figure 6-6 is a typical computer-generated stem-and-leaf display of the compressive strength data in Table 6-2. The software uses the same stems as in Fig. 6-4. Note also that the computer orders the leaves from smallest to largest on each stem. This form of the plot is usually called an **ordered stem-and-leaf diagram**. This is not usually used when the plot is constructed manually because it can be time-consuming. The computer also adds a column to the left of the stems that provides a count of the observations at and above each stem in the upper half of the display and a count of the observations at and below each stem in the lower half of the display. At the middle stem of 16, the column indicates the number of observations at this stem.

CM 5.7

Problem #7: Consider the data set that is summarized in the R Output below.

```
leaf unit: 1
        n:13
  1    1  5
  2    2  7
  2    3
  4    4  17
 (6)   5  339999
  3    6  357
```

(a) Find the values of Q_1 and Q_3.

(b) Find the median

(c) Find the adjacent values.
(**Note:** Read this file for the relevant definitions, and for an example.)

(d) Which of the following is a correct modified boxplot for this data set?

$Q_1 = \frac{n+1}{4} = \frac{13+1}{4} = \frac{14}{4} = 3.5 \longrightarrow 41 + 0.5(47-41) = 44$

$Q_3 = \frac{3(13+1)}{4} = \frac{42}{4} = 10.5 \longrightarrow 59 + 0.5(63-59) = 61$

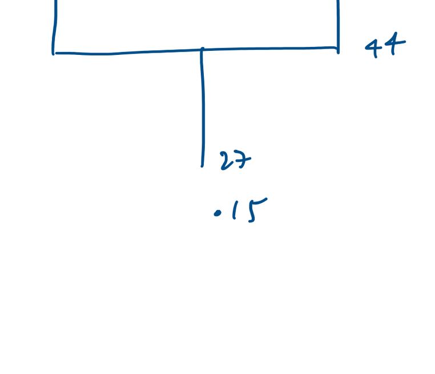
$Q_1 = 44$
$Q_3 = 61$
$a_1 = 27$
$a_2 = 67$

$IQR = Q_3 - Q_1 = 77$
$Q_1 - 1.5(IQR) = 18.5$
$Q_3 + 1.5(IQR) = 86.5$

Median = 59

a_1 = smallest non-outlier = 27

a_2 = largest non-outlier = 67



a) 44, 61 ✓
b) 59 ✓
c) 27, 67 ✓
d) G ✓

Problem #10: An 1868 paper by German physician Carl Wunderlich reported, based on over a million body temperature readings, that healthy adult body temperatures are approximately normally distributed with mean 98.6 degrees Farenheit and standard deviation 0.6. In a random sample of 45 healthy adults, find the probability that the average body temperature is between 98.48 and 98.72.

$P(98.48 \le x \le 98.72)$ $\quad z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

$P\left(\frac{98.48 - 98.6}{\frac{0.6}{\sqrt{45}}} \le z \le \frac{98.72 - 98.6}{\frac{0.6}{\sqrt{45}}}\right)$

$P(-1.34164 \le z \le 1.34164)$ → $P(-1.34164 \le z)$
$= 1 - P(1.34164)$
$= 1 - 0.909877$
$= 0.090123$

$= P(z \le 1.34164) - 0.090123$
$= 0.909877 - 0.090123$
$= 0.819754$
$= 0.8198$ ✓

Ex. Temperature data
- Consider the following **univariate** (has only one random variable) data set measuring the air temperature at LaGaurdia 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		

- The temperature data is composed of 'continuous' interval type data
- There are 25 observations in the temperature data set
- The sample mean, or average, is equal to 76.52 and the sample deviation is approx. 9.468

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

5	7	8							
6	1	4	6	8					
7	2	2	4	6	7	9	9	9	9
8	1	1	3	3	5	6	6	7	8
9	2								

- The data is now ordered
- Min. val. = 57, max. val. = 92, median and mode are both 79
- We also have evidence that the data is **unimodal** (has a single mode)

Exercises

1. Consider the following data: 2, 3, 4. What are the mean and standard deviation?

$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + x_3}{n} = \frac{2+3+4}{3} = \frac{9}{3} = 3$

$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2}{n-1}} = \sqrt{\frac{(2-3)^2 + (3-3)^2 + (4-3)^2}{2}}$

$= \sqrt{\frac{1+0+1}{2}} = \sqrt{2/2} = 1$

2. The mean of four numbers is 15. Three of these numbers are 3, 6, and 22. What is the other value?

$15 = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + x_3 + x_4}{n} = \frac{3+6+22+x_4}{4}$

$15 = \frac{31 + x_4}{4}$

$60 - 31 = x_4 = 29$

3. Consider the following data:
2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22.
What is the range, median, and IQR?

range: max - min: 22 - 2 = 20; median: 12

- 25th percentile: $Q_1 = \frac{n+1}{4}$
  pos. of $Q_1 = \frac{11+1}{4} = 3 \longrightarrow 6$
- 75th percentile: $Q_3 = \frac{3(n+1)}{4}$
  pos of $Q_3 = \frac{3(11+1)}{4} = 9 \longrightarrow 18$

→ IQR: 18 - 6 = 12

4. Consider the following data:
0.13, 0.25, 0.31, 0.44, 0.49, 0.51, 0.55,
0.59, 0.70, 0.81.
What is the IQR?

- 25th percentile: $Q_1 = \frac{n+1}{4}$
  pos of Q_1: $\frac{10+1}{4} = 2.75$ → interpolate!
  $0.25 + (0.75)(0.31 - 0.25) = 0.25 + 0.045 = 0.295$
  (0.25 → 2nd #; 0.75 → 75% the dist. from 2nd to 3rd; 0.31 - 0.25 → 3rd # - 2nd #)
  $Q_1 = 0.295$
- 75th percentile: $Q_3 = \frac{3(n+1)}{4}$
  pos of $Q_3 = \frac{3(10+1)}{4} = 8.25$ → interpolate!
  $0.59 + 0.25(0.7 - 0.59) = 0.59 + 0.0275 = 0.6175$
  (0.59 → 8th #; 0.7 - 0.59 → 9th # - 8#)
  $Q_3 = 0.6175$

→ IQR = 0.6175 - 0.295 = 0.3225

CM 5.5

Problem #5: Note that this question (along with all other problems that don't have the R symbol beside them) should be done **by hand**.

Consider the following data,

41, 66, 64, 43, 65, 47, 52, 60, 69, 57.

(a) Draw a stem and leaf plot **including depths** using intervals of length 5. (See the bottom of this example for an explanation of depths.) Enter the value of the depths **only** in the answer box below.
For example, if your stem-and-leaf plot looked like the one on at the bottom of this example then you would enter the following numbers into the answer box:
`8, 36, 20, 44, 21, 5, 1`

(b) Calculate the mean.

(c) Calculate the sample variance.

(d) Calculate the range.

1. 41, 43, 47, 52, 57, 60, 64, 65, 66, 69

n = 10

depth	stem	leaf
2	4	13
3	4	7
4	5	2
(1)	5	7
(5)	6	04
3	6	569

a) 2, 3, 4, 1, 5, 3

b) $\bar{x} = 56.4$ ✓

c) $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{(41-56.4)^2}{9} + \frac{(43-56.4)^2}{9} + \frac{(47-56.4)^2}{9} + \frac{(52-56.4)^2}{9} + \frac{(57-56.4)^2}{9} + \frac{(60-56.4)^2}{9} + \frac{(64-56.4)^2}{9} + \frac{(65-56.4)^2}{9} + \frac{(66-56.4)^2}{9} + \frac{(69-56.4)^2}{9}$

$= 26.35 + 19.95 + 9.82 + 2.15 + 0.04 + 1.44 + 6.42 + 8.22 + 10.24 + 17.64$

$s^2 = 102.27$ ✓

d) 69 - 41 = 28 ✓

Alternative (smarter) way — use calculator!
enter stats mode: MODE, MODE, 1 SD
add data: 41, M+, 43, M+, … etc → must press M+ (above AC) to add value to dataset
get mean: shift, 2 (S-var), 1 ($\bar{x}$), ENTER
⇒ 56.4 → must press = to get value
get sample variance: shift, 2 (S-var), 3 (Sx), ENTER, x^2 ⇒ 102.267
(it only gives you S, but sample variance is s^2)
get range: do it urself lol xd

Problem #9: Consider the following data.

23, 23, 39, 32, 37, 24, 29

(a) If you were to construct a normal probability plot by hand for the above data what are the numerical values (in order) of the **first five** numbers that would go on the x-axis? Separate your answers with a comma. (See this example for a description of the normal probability plot.)
For example, if your data was the same as this example then you would enter the following numbers from Step 4 (on the last page) into the answer box (note that spaces don't matter):
`4, 4, 7, 7, 9`

(b) If you were to construct a normal probability plot by hand for the above data what are the numerical values (in order) of the **first five** numbers that would go on the y-axis? Separate your answers with a comma. (See this example for a description of the normal probability plot.)
For example, if your data was the same as the data on the last page of this example then you would enter the following numbers from Step 4 (on the last page) into the answer box (note that spaces don't matter):
`-1.83, -1.28, -0.97, -0.73, -0.52`

23, 23, 24, 29, 32, 37, 39 $\quad n = 7$

a) 23, 23, 24, 29, 32 ✓

b) $\alpha_1 = \frac{1-0.5}{7} = 0.0714$ $\quad \alpha_4 = \frac{4-0.5}{7} = 0.5$

$\alpha_2 = \frac{2-0.5}{7} = 0.2143$ $\quad \alpha_5 = \frac{5-0.5}{7} = 0.6429$

$\alpha_3 = \frac{3-0.5}{7} = 0.3571$

$z_1 = -1.47$
$z_2 = -0.79$
$z_3 = -0.37$
$z_4 = 0$
$z_5 = 0.37$

-1.47, -0.79, -0.37, 0, 0.37 ✓

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

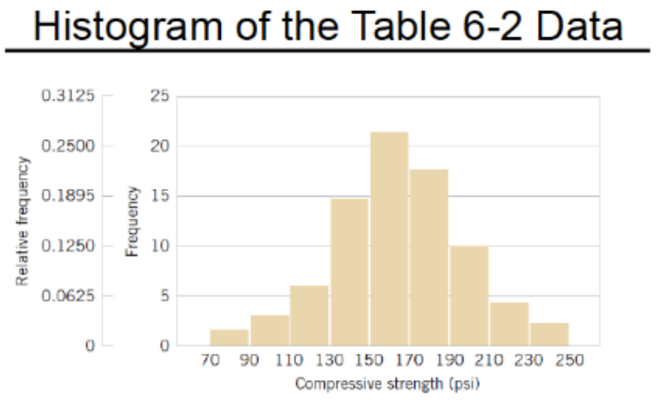


Figure 6-7 Histogram of compressive strength of 80 aluminum-lithium alloy specimens. Note these features – (1) horizontal scale bin boundaries & labels with units, (2) vertical scale measurements and labels, (3) histogram title at top or in legend.

CUMULATIVE DISTRIBUTION PLOT

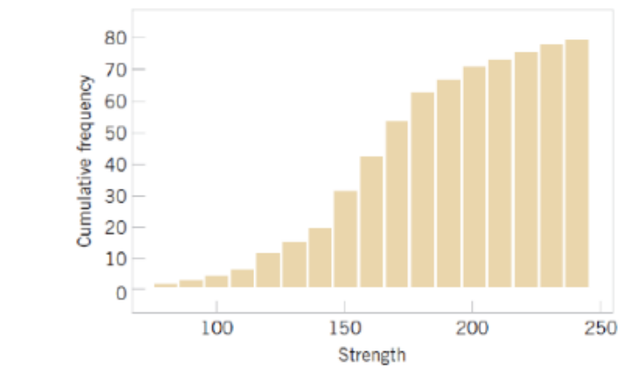


Figure 6-10 Cumulative histogram of compressive strength of 80 aluminum-lithium alloy specimens. Comment: Easy to see cumulative probabilities, hard to see distribution shape.

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

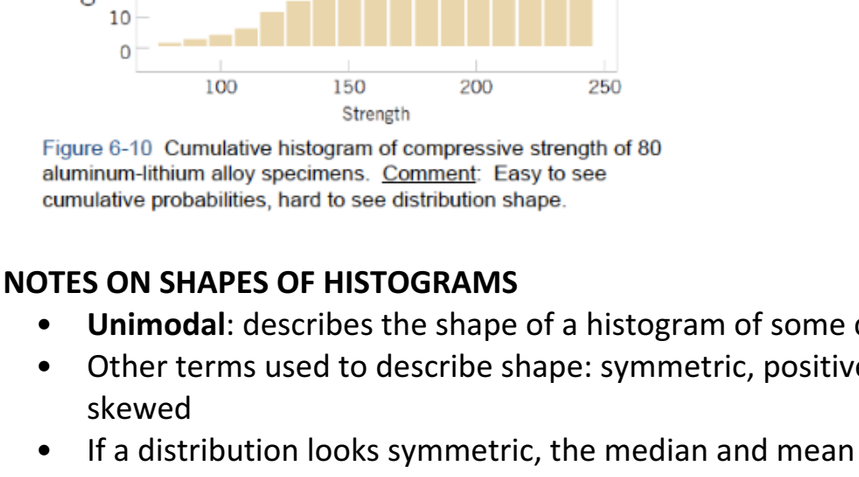


Figure 6-12 Airplane production in 1985. (Source: Boeing Company) Comment: Illustrates nominal data in spite of the numerical names, categories are shown at the bin's midpoint, a Pareto chart since the categories are in decreasing order.

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

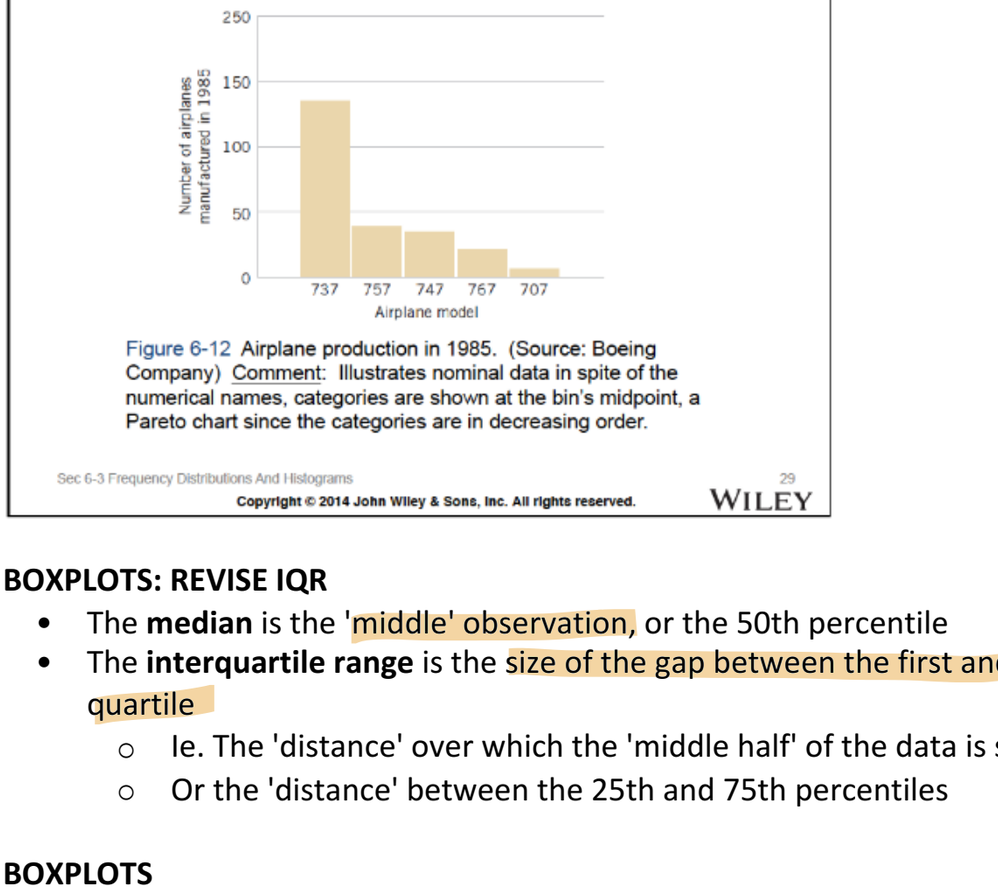


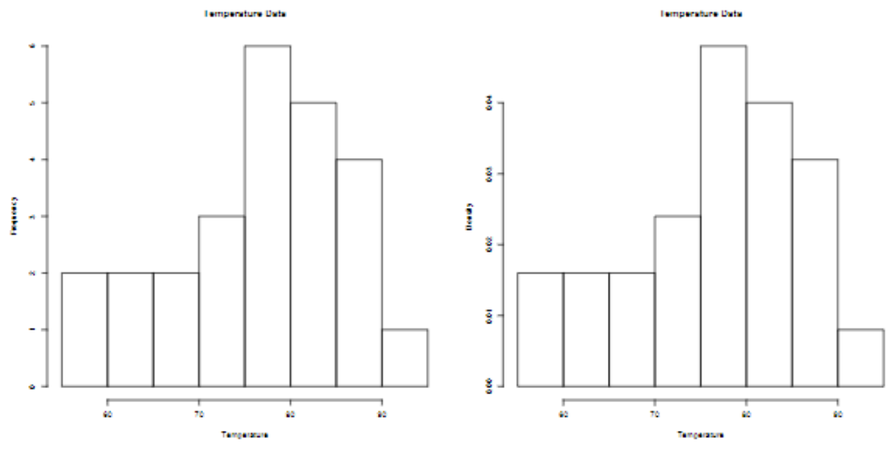
Figure 6-13 Description of a box plot.

Ex. Temperature data

- Consider the following **univariate** (has only one random variable) data set measuring the air temperature at LaGaurdia 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 | 5 | 7 8 — 57, 58
6 | 6 | 1 4 6 8 — 61, 64, 66, 68
(9) | 7 | 2 2 4 6 7 9 (9) 9 9 etc — (13th value — median)
10 | 8 | 1 1 3 3 5 6 6 7 8 etc
1 | 9 | 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 | 13 → median | max
= 50%
= 79 → 13th value

Find Q_1 = 25th percent

$Q_1 = \frac{n+1}{4} = \frac{25+1}{4} = 6.5$

data pt 6: 68
" " 7: 72

$68 + (0.5)(72-68) = 70$

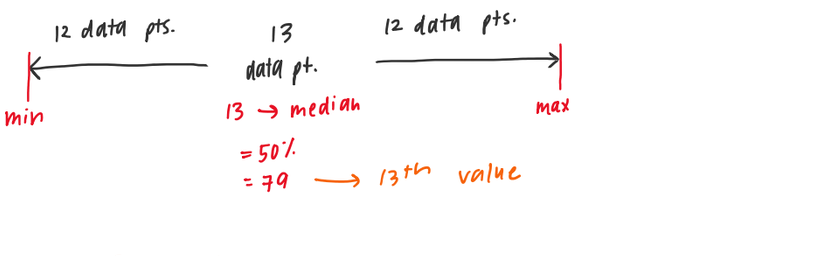
Find Q_2: 75th percent

$Q_3 = \frac{3(n+1)}{4} = \frac{3(25+1)}{4} = 19.5$

data pt 19: 83
" " 20: 85

$83 + (0.5)(85-83) = 84$

IQR: $84 - 70 = 14$



$1.5 \times IQR = 1.5 \times 14 = 21$

→ whiskers extend to furthest data point that is still within 1.5 IQR of the 1st or 3rd quartile

Lec18 - probability plots

Yuxi Qin
yuxiqin.ca

Monday, October 24, 2022 7:56 PM

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

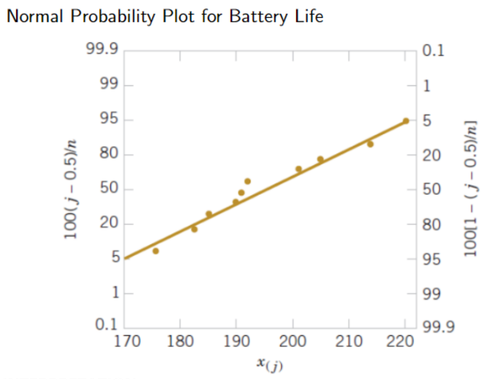
$$\frac{j-0.5}{n}$$

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:



← 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, $n=10$
↳ $100\left(\frac{j-0.5}{10}\right) = 10(j-0.5)$ ← calculate

plot $10(j-0.5)$ on y-axis & plot x on x-axis

j	x_j	$10(j-0.5)$
1	176	$10(1-0.5) = 5$
2	183	$10(2-0.5) = 15$
3	185	$= 25$
4	190	$= 35$
5	191	$= 45$
6	192	$= 55$
7	201	$= 65$
8	205	$= 75$
9	214	$= 85$
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 7th percentile points
- Fig 6.22 on p145 represents a situation where it is appropriate to model the data with a normal distribution

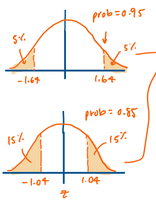


TABLE 6.6 Calculation for Constructing a Normal Probability Plot

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

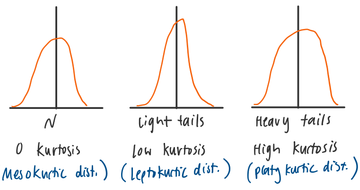
- A normal probability plot can also be constructed by plotting z-scores on the Y-axis and the variable on the X-axis

$$\frac{j-0.5}{n} = P(Z \le z_j) = \phi(z_j) \longrightarrow \phi^{-1} \longrightarrow \text{normal quartile function}$$

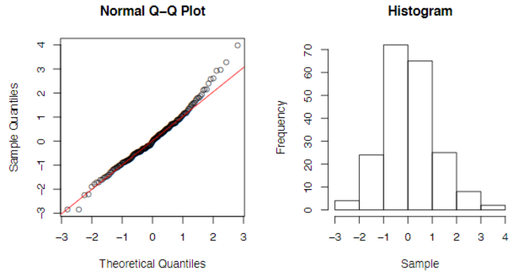
INTERPRETING NORMALITY AND NON-NORMALITY

- 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)

Kurtosis: the sharpness of a the peak of a frequency-distribution curve

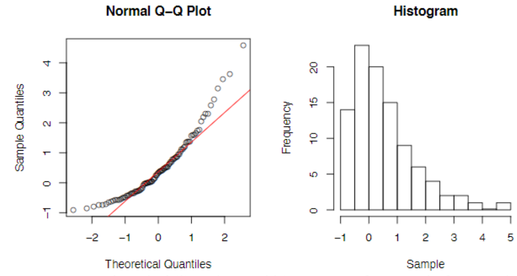


NORMAL PLOT



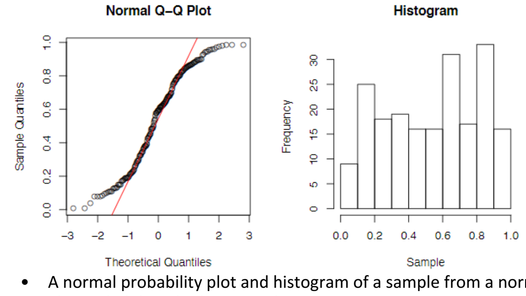
- A normal probability plot and histogram of a sample form a normal distribution

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.

Lec19 - point estimation, central limit theorem

Yuxi Qin
yuxiqin.ca

Thursday, October 27, 2022 5:41 PM

INTRO TO POINT/UNBIASED/MIN VARIANCE ESTIMATION

- Suppose that we observe or collect a random sample from some population
- The parameter θ (theta), is an unknown characteristic or parameter of the distribution
- There are many ways to estimate the value of θ, and each use the observed/collected sample

***RECALL:* ESTIMATORS AND ESTIMATES**

- A *statistic* is a function of a random sample: X1, … , Xn
- As such, the estimator of a distributions' parameter θ 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 θ**
- 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 X1, …, Xn is a random sample from a population, whose distribution is characterized by an unknown parameter, θ
- Let $\hat{\theta}$ be a point estimator of θ
- Then $\hat{\theta}$ is an **unbiased estimator** of θ if:

$$\mathbb{E}\left[\hat{\theta}\right] = \theta$$

  - For every possible value of θ
- Note: if $\hat{\theta}$ is not unbiased then the **bias** estimate of $\hat{\theta}$ is defined to be the difference

$$\mathbb{E}\left[\hat{\theta}\right] - \theta$$

TWO UNBIASED ESTIMATORS

PROPOSITIONS:

1. Suppose that X1,…,Xn is a random sample from a population with mean μ and sample mean

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

   - The estimator $\bar{x}$ is an unbiased estimator of μ

2. 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

1. The estimator $\bar{x}$ is an unbiased estimator of μ

$$\mathbb{E}\left[\bar{X}\right] = \mathbb{E}\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i\right] = \frac{1}{n}n\mu = \mu$$

2. An unbiased estimator of p is $\hat{p} = \frac{x}{n}$

$$\mathbb{E}\left[\hat{p}\right] = \mathbb{E}\left[\frac{X}{n}\right] = \frac{1}{n}\mathbb{E}\left[X\right] = \frac{1}{n}np = p$$

--

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 X1,…,Xn is a random sample from some population with mean μ and variance $\sigma^2 < \infty$
- Then an unbiased estimator of σ^2 is the sample variance

$$\hat{\sigma}^2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2.$$

- We could also consider the estimator

$$\hat{\sigma}_1^2 = S_1^2 = \frac{1}{n}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2;$$

- 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 ounbiased 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 θ, we should select the one with smallest variance

THE MEAN OF A GAUSSIAN DISTRIBUTION

THEOREM:

- Suppose that X1,…,Xn is a random sample from a Gaussian distribution with mean μ and standard deviation σ
- Then $\hat{\mu} = \bar{x}$ is the minimum variance unbiased estimator of μ
  - 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 θ 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 <= mean squared error of any estimator, for any value of θ, is the **optimal estimator** of θ

Ex. 1 POINT ESTIMATOR

We observe a sample of 20 observations from a normal distribution with mean μ=30 and standard deviation σ=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 μ:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{601.0479}{20} = 30.052$$

$$\text{median}(x) = \tilde{x} = \frac{(30.01 + 31.07)}{2} = 30.541$$

$$\frac{[\min(X_i) + \max(X_i)]}{2} = \frac{17.91 + 38.67}{2} = 28.29$$

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 σ is unknown, we may replace it by the sample standard deviation $s = 0.284$ to obtain the estimated standard error of $\bar{X}$ as

$$SE(\bar{X}) = \hat{\sigma}_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{0.284}{\sqrt{10}} = 0.0898$$

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 $2\hat{\sigma}_{\bar{X}} = 2(0.0898) = 0.1796$, and we are highly confident that the true mean thermal conductivity is within the interval 41.924 ± 0.1796 or between 41.744 and 42.104.

Ex. If X1,…,Xn is a random sample from a Gaussian distribution with mean μ and variation σ^2. Show that:

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

Is an unbiased estimator for σ^2

value at i
mean

$$\mathbb{E}(S^2) = \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]$$

$$= \mathbb{E}\left[\frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)\right]$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}\left[x_i^2\right] - \frac{1}{n}\mathbb{E}\left[\left(\sum_{i=1}^{n} x_i\right)^2\right]$$

Recall: $Var[X] = \mathbb{E}[X^2] - E[X]^2 \Rightarrow \mathbb{E}[X^2] = Var[X] + E[X]^2$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n}\mathbb{E}\left[x_i^2\right] - \frac{1}{n}\mathbb{E}\left[\left(\sum_{i=1}^{n} x_i\right)^2\right]\right)$$

$$= \frac{1}{n-1}\left\{\sum_{i=1}^{n}\left(Var[x_i] + E[x_i]^2\right) - \frac{1}{n}\left(Var\left[\sum_{i=1}^{n} x_i\right] + \mathbb{E}\left[\sum_{i=1}^{n} x_i^2\right]\right)\right\}$$

$$= \frac{1}{n-1}\left\{\sum_{i=1}^{n}\left(\sigma^2 + \mu^2\right) - \frac{1}{n}\left(n\sigma^2 + (n\mu)^2\right)\right\}$$

$$= \frac{1}{n-1}\left\{n\sigma^2 + n\mu^2 - \frac{1}{n}n\sigma^2 - \frac{1}{n}(n\mu)^2\right\}$$

$$= \frac{1}{n-1}\left\{n\sigma^2 - \sigma^2\right\}$$

$$= \frac{1}{n-1}\sigma^2\{n-1\}$$

$$= \sigma^2$$

Lec20 - Central Limit Theorem, Ch7 examples

Yuxi Qin
yuxiqin.ca

Sunday, October 30, 2022 1:58 AM

→ sample taken from larger population

DEFINITION: CENTRAL LIMIT THEOREM

- If X1, X2, … , Xn is a random sample of size n taken from a population with mean μ and finite variance σ^2, and $\bar{X}$ is the sample mean, the limiting form of the distribution of

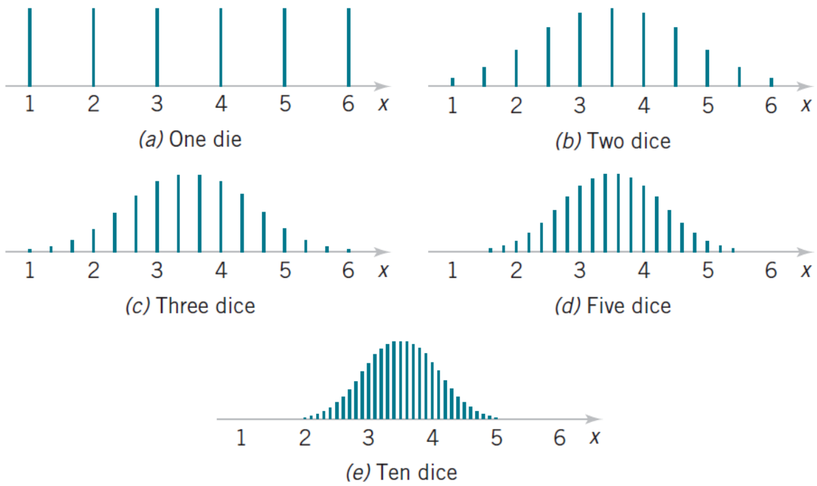
$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

as n-> inf, 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 distributino 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 μ_1 and μ_2 and variances σ_1^2 and σ_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.

$\mu_{\bar{x}} = 75.5$, $\sigma = 3.5$, $n = 6$

$P(\bar{x} \geq 75.75) = 1 - P(\bar{x} \leq 75.75)$

$= 1 - P(z \leq 0.175)$

$= 1 - 0.56945$

$= 0.43055$

$z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \rightarrow \frac{3.5}{\sqrt{6}} = 1.429$

$z = \frac{75.75 - 75.5}{1.429} = 0.175$

Ex. A random sample of size n1 = 16 is selected from a normal population with a mean of 75 and a standard deviation of 8. A second random sample of size n2 = 9 is taken from another normal population with mean 70 and standard deviation 12. Let X1 and X2 be the two sample means. Find:
a. The probability that X1 − X2 exceeds 4
b. The probability that 3.5 ≤ X1 − X2 ≤ 5.5

$n_1 = 16$, $\mu_1 = 75$, $\sigma_1 = 8$; $n_2 = 9$, $\mu_2 = 70$, $\sigma_2 = 12$

a) $\bar{X}_1 - \bar{X}_2 > 4$:

$z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{4 - (75-70)}{\sqrt{\frac{8^2}{16} + \frac{12^2}{9}}} = \frac{-1}{\sqrt{\frac{64}{16} + \frac{144}{9}}} = \frac{-1}{\sqrt{4+16}} = \frac{-1}{\sqrt{20}}$

$= -0.2236$

$\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$

$\sim N(75-70, \frac{8^2}{16} + \frac{12^2}{9})$

$\sim N(5, 20)$

$\mu \quad \sigma$

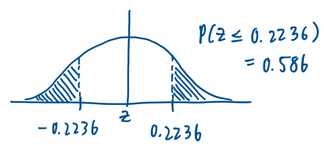
$P(\bar{X}_1 - \bar{X}_2 > 4) = 1 - P(z \leq -0.2236)$

$= 1 - 0.414$

$= 0.586$

$1 - 0.586 = 0.414$

$P(z \leq 0.2236) = 0.586$



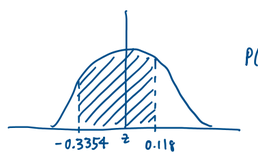
$\frac{\sigma^2}{n} = 20$ $\quad \frac{\sigma}{\sqrt{n}} = \sqrt{20}$

b) prob. $3.5 \leq \bar{X}_1 - \bar{X}_2 \leq 5.5$

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

$P\left[\frac{3.5-5}{\sqrt{20}} \leq z \leq \frac{5.5-5}{\sqrt{20}}\right]$

$= P[-0.3354 \leq z \leq 0.1118]$



$P(3.5 \leq x \leq 5.5) = P(z \leq 0.118) - P(z \leq -0.3354)$

$1 - P(z \leq 0.3354) = 1 - 0.633$

$= 0.3687$

$P(3.5 \leq X \leq 5.5) = 0.5445 - 0.3687 = 0.1759$

memory aid: eqn sheet 23

Lec21 - Confidence intervals for a normal random variable

Yuxi Qin
yuxiqin.ca

Thursday, November 10, 2022 4:43 PM

— uses central limit theorem as a basis

- 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 μ and standard deviation σ, 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 μ

CONFIDENCE INTERVAL: CONSTRUCTION

- Suppose that we collect or observe a random sample, x1,…,xn from a normal population
- We estimate tha the population mean μ and we know the value of the population standard deviation is σ
- Given the sampling distribution $\bar{X}$, a range of possible values for μ might be all values such that $p(a \le \mu \le b) = 1 - \alpha$
- Typically, we take $\alpha = 0.05$ and consider only intervals that are symmetric about μ
- These intervals can be written as

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

$1 - \alpha = 0.95$
$\alpha = 0.05$
$\Rightarrow \frac{\alpha}{2} = 0.025$
$0.95 + 0.025 = 0.975$

A 95% CONFIDENCE INTERVAL

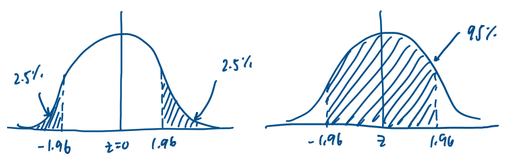
- **COROLLARY:** A **95% Confidence Interval** for μ is the random interval

$$\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}},\ \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

Go to z-table, look at z-value corresponding to PC 0.975) → $z = 1.96$

- **PROOF:** A standard normal distribution has the property

$$\begin{aligned}
P\left(-z_{\alpha/2} \le Z \le z_{\alpha/2}\right) &= 1 - \alpha \\
P\left(-1.96 \le Z \le 1.96\right) &= 0.95 \\
P\left(-1.96 \le \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \le 1.96\right) &= 0.95 \\
P\left(-1.96\frac{\sigma}{\sqrt{n}} \le \bar{X} - \mu \le 1.96\frac{\sigma}{\sqrt{n}}\right) &= 0.95 \\
P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) &= 0.95
\end{aligned}$$



There's a prob. of $1-\alpha$ of selecting a sample for which the CI will contain the true value of μ

- Therefore we can say with 95% confidence that μ lies in the interval

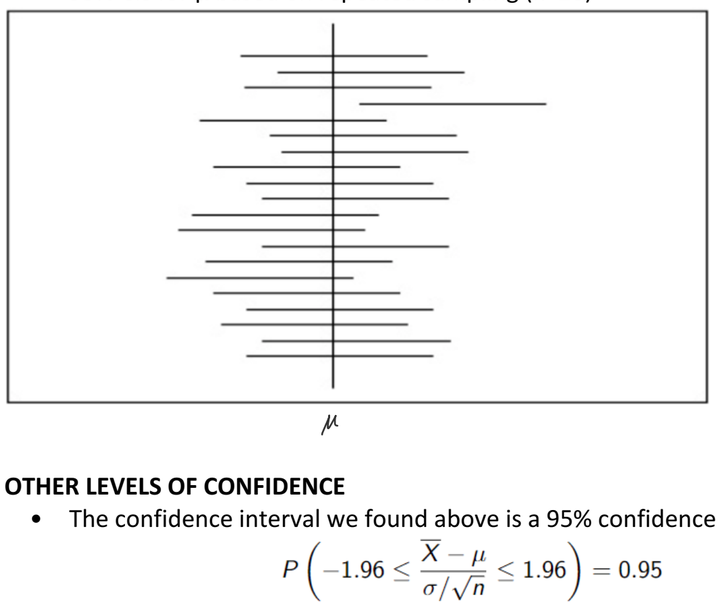
$$\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

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 μ 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 μ
- For a single sample, however, we don't know if the interval does or does not contain μ

Ex. Repeated Sampling-

- 95% CIs for μ 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

$$P\left(-1.96 \le \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \le 1.96\right) = 0.95$$

- 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-α)%, we use

$$P\left(-z_{\alpha/2} \le \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right) = 1 - \alpha$$

ex. $0.93 = 1 - \alpha$
$\alpha = 0.07$
$\frac{\alpha}{2} = 0.035$
$(1-\alpha) + \frac{\alpha}{2} = 0.965$
z-value @ 0.965 = 1.82

→ $1 - \alpha = 0.93$, $z_{\frac{\alpha}{2}} = 1.82$

- 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

$(1-\alpha) + \frac{\alpha}{2}$ → z-table

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

ex. finding with 90% confidence:

$1 - \alpha = 0.90$
$\alpha = 0.10$
$\frac{\alpha}{2} = 0.05$
$z_{\frac{\alpha}{2}} = z_{0.05}$

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 α 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-α

- Note: $z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$ 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:

$$\begin{aligned}
2 \times z_{\alpha/2}\frac{\sigma}{\sqrt{n}} &\le q \\
\frac{1}{\sqrt{n}} &\le \frac{q}{2z_{\alpha/2}\sigma} \\
\sqrt{n} &\ge 2 \times z_{\alpha/2}\frac{\sigma}{q} \\
n &\ge \left(2 \times z_{\alpha/2}\frac{\sigma}{q}\right)^2
\end{aligned}$$

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?

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

$P(\bar{x} < 1.8) = P\left(z < \frac{1.8 - 1.78}{\frac{0.09}{\sqrt{25}}}\right) = P(z < 1.11) = 0.8665$

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 μ.

$n = 25$; $\mu = 1.8$; $\sigma = 0.09$

$\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}} = 1.8 \pm 1.96\left(\frac{0.09}{\sqrt{25}}\right)$
$= 1.8 \pm 0.035$
$= (1.765, 1.835)$

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 $\bar{x}$= 39kg, construct a 95% confidence interval for μ.

95% confidence interval for μ 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/n 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 μ
Suppose that for our sample of 25 men on campus, we're asked to calculate a 90% confidence interval for μ

$\bar{X} \pm 1.645\frac{\sigma}{\sqrt{n}} = 1.8 \pm 1.645\frac{0.009}{\sqrt{25}} = 1.8 \pm 0.0030 = (1.770, 1830)$

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?

$n \ge \left(2 \times z_{\frac{\alpha}{2}}\frac{\sigma}{2}\right)^2$
$n \ge \left(2 \times 1.96\left(\frac{5.6}{2}\right)\right)^2$
$n \ge 120.4726$
$= 121$

to attain our goal, we need to collect 121 people

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 μ.

$\bar{X} \pm z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} = 20.26 \pm 1.96\frac{10}{\sqrt{25}} = 20.26 \pm 2.77 = (17.49, 23.03)$

Ex. A population is normally distributed with=5.6, but μ is unknown. A random sample of 10 individuals is drawn, and X= 25.4. Calculate a 90% confidence interval for μ.

$\bar{X} \pm z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} = 25.4 \pm 1.645\frac{5.6}{\sqrt{10}} = 25.4 \pm 2.913 = (22.487, 28.313)$

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.

$\bar{X} \pm z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} = 496.1 \pm 1.96\frac{8}{\sqrt{40}} = 496.1 \pm 2.48 = (493.62, 498.58)$

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.

$n \ge \left(2 \times z_{\frac{\alpha}{2}}\frac{\sigma}{2}\right)^2 \Rightarrow nz\left(2 \times 1.96\frac{8}{15}\right)^2 = 437.09$

$\therefore n = 438$

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_α
  - Upper bound for μ is
    $\mu \le \bar{X} + z_\alpha \sigma/\sqrt{n}$
  - Lower bound for μ is
    $\bar{X} - z_\alpha \sigma/\sqrt{n} \le \mu$

LARGE SAMPLE CONFIDENCE INTERVAL FOR μ

- 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

$$\bar{X} - z_{\alpha/2}\frac{s}{\sqrt{n}} \le \mu \le \bar{X} + z_{\alpha/2}\frac{s}{\sqrt{n}}$$

Is a large sample confidence interval for μ

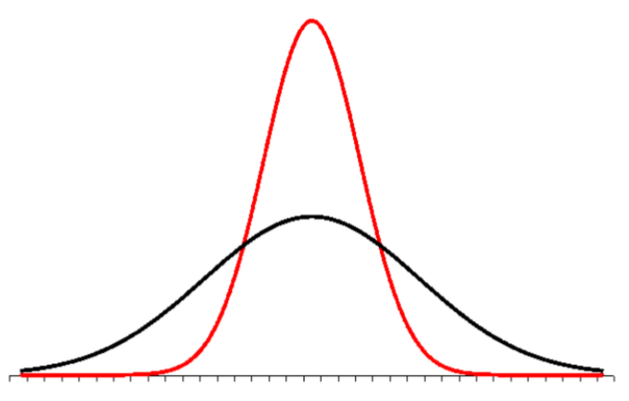
RECAP:

- 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

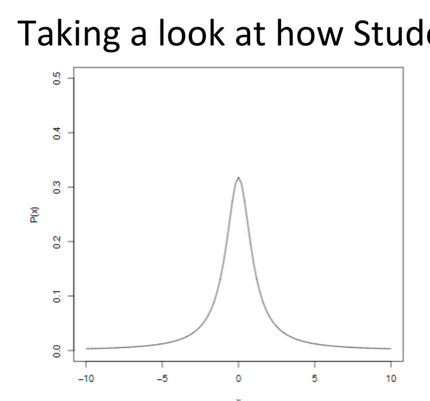
- First published by WS Gosset in 1908
- Gosset was an oxford-trained chemist who worked in the Guinness brewery in Dublin, Ireland
- He wasn't allowed to publish his work under his own name, so he used the pseudo name 'Student'
- The t-distribution looks more and more like the normal distribution as 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, μ by replacing the $z_{\frac{\alpha}{2}}$ by the corresponding critical values of the t_{n-1} distribution

A STUDENT'S T CONFIDENCE INTERVAL

- The 100(1-α)% Student's t confidence interval for the mean μ 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-α)% confidence interval for μ, as follows:

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

- Here, t_V^{α} is the value of the t-distribution corresponding to a given significance level α and a given degrees of freedom, V.

DISCUSSION

- In general, we will always use a Student's t confidence interval when σ 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

$$\mathbb{V}ar[\hat{p}] = \mathbb{V}ar\left[\frac{X}{n}\right]\frac{1}{n^2}\mathbb{V}ar[X] = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$$

- 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

$n\hat{p} \ge 15$ and $n(1-\hat{p}) \ge 15$

It follows that a confidence interval for $\hat{p}$ is given by

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

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

CM 6.3

Problem #3: A sample of 143 hypertensive people were given an anti-hypertensive drug, and the drug was found to be *effective* in 49 of those people. (By *effective*, we mean that their diastolic blood pressure is lowered by at least 10 mm Hg as judged from a repeat measurement taken 1 month after taking the drug.)

(a) Find a 92% confidence interval for the true proportion of the sampled population for which the drug is effective.

(b) Using the results from the above mentioned survey, how many people should be sampled to estimate the true proportion of hypertensive people for which the drug is effective to within 3% with 99% confidence?

(c) If no previous estimate of the sample proportion is available, how large of a sample should be used in **(b)**?

$N = 143$, $n = 49$ → 34.266 %

a) $\hat{p} = \frac{x}{n} = \frac{49}{143} = 0.343$

$\sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{0.343(1-0.343)}{143}} = 0.039697$

$1-\alpha = 0.92$, $\alpha = 0.08$, $\frac{\alpha}{2} = 0.04$

$z(0.92+0.04) = 1.75$ (0.96)

$\hat{p} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.343 \pm 1.75(0.039697)$

$= [0.2732035511,\ 0.4121111342]$

$= [0.273, 0.412]$ ✓

b) within 3% = 0.03

99% confidence: $1-\alpha = 0.99$, $\frac{\alpha}{2} = 0.005$

$z_{\frac{\alpha}{2}} = z(0.99 + 0.005) = 2.576$ (0.995)

$n = \left(\frac{z_{\frac{\alpha}{2}}}{E}\right)^2 p(1-p) = \left(\frac{2.576}{0.03}\right)^2 (0.343)(1-0.343) = 1660.737784 = 1661$ ✓ (round up) (0.343)

c) Conservative estimate — use p=0.50

$n = \left(\frac{z_{\frac{\alpha}{2}}}{E}\right)^2 p(1-p) = \left(\frac{2.576}{0.03}\right)^2 (0.50)(1-0.50) = 1843.271111 = 1844$ ✓

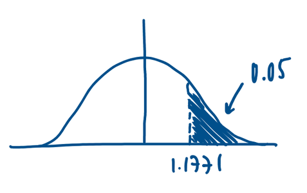
Yuxi Qin
yuxiqin.ca

Student t's

Ex. Using the student's t table (p745)

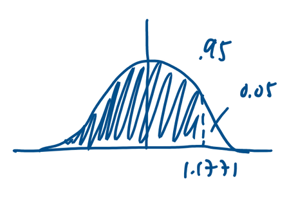
1. What is $P(t > 1.771)$ when $n = 14$?
2. What is $P(t < 1.711)$ when $n = 25$?
3. What is $P(t > 2.06)$ when $n = 17$?

1. $v = n-1$, $v = 14-1 = 13$



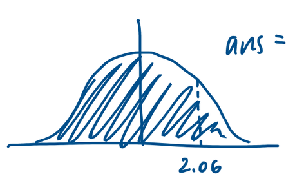
$P(t > 1.771) = 0.05$

2. $v = 25-1 = 24$, $P(t > 1.711) = 0.05$, $1 - 0.05 = 0.95$



$P(t > 1.711) = 0.95$

3. $P(t > 2.06)$ $n = 17$, $v = 17-1 = 16$



ans = $0.25 < p < 0.5$

df	0.1	0.05
1	3.078	6.314
2	1.886	2.920
3	1.638	2.353
4	1.533	2.132
5	1.476	2.015
6	1.440	1.943
7	1.415	1.895
8	1.397	1.860
9	1.383	1.833
10	1.372	1.812
11	1.363	1.796
12	1.356	1.782
13	1.350	1.771 ①

Ex. 8.3 (p176)

Construct a lower, 1-sided 95% confidence interval for the mean.

$x = 64.46$, $\sigma = 1$, $n = 10$

$\bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}} \le \mu$

Find z_α → just read $z \sim 0.95$ from the z-table

$z = 1.65$

$64.46 - 1.65\left(\frac{1}{\sqrt{10}}\right) \le \mu$

$63.94 \le \mu$

1. 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 (6 inch, 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
$\bar{x}$	2.03	2.55	2.34
s	0.068	0.084	0.064

Calculate a 95% confidence interval for μ_1.

Type 1:

$\bar{x} \pm t_{n-1,\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$

$\Rightarrow 2.03 \pm 2.57\frac{0.068}{\sqrt{6}}$

$\Rightarrow 2.03 \pm 0.07$

$\Rightarrow (1.96, 2.10)$

based on sample mean, we're 95% confident that the avg for Type 1 springs is between 1.96 and 2.10

Finding $\frac{\alpha}{2}$'s — Applies to Type 1, 2, and 3

$1-\alpha = 0.95$, $\alpha = 0.05$, $\frac{\alpha}{2} = 0.025$

$v = n-1$, $v = 6-1 = 5$

df	0.1	0.05	0.025
1	3.078	6.314	12.706
2	1.886	2.920	4.303
3	1.638	2.353	3.182
4	1.533	2.132	2.776
5	1.476	2.015	2.571

$t_{n-1,\frac{\alpha}{2}} = 2.57$

Type 2: $x \pm t_v^{\alpha}\frac{s}{\sqrt{n}}$

$\Rightarrow 2.55 \pm 2.57\frac{0.084}{\sqrt{6}}$

$\Rightarrow 2.55 \pm 0.088$

$\Rightarrow$

Type 3: $\bar{x} \pm t_v^{\alpha}\frac{s}{\sqrt{n}}$

$2.34 \pm 2.57\frac{0.064}{\sqrt{6}}$

2.34 ± 0.063

Consider a sample:
$x = (1.10, -11.00, 12.00, 0.52, 4.30, -6.00, 5.90, 2.90, 7.60, -2.20, 3.80, -2.30, 6.50, -4.30, 9.30, 0.23, 15.00, -1.60, -2.20, -1.20)$
drawn from a normal population. Find a 95% CI for μ.

$\bar{x} \pm t_{n-1,\frac{\alpha}{2}}\frac{s}{\sqrt{n}}$ → need to calculate $\bar{x}$ and s

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{38.35}{20} = 1.92$ and $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = 6.30$

$\therefore \bar{x} \pm t_{n-1,\frac{\alpha}{2}}\frac{s}{\sqrt{n}} = 1.92 \pm 2.09\frac{6.30}{\sqrt{20}} = 1.92 \pm 2.95$

$= (-1.03, 4.87)$

Finding $t_{n-1,\frac{\alpha}{2}}$:

$n = 20$, $v = 19$, $\alpha = 0.05$, $\frac{\alpha}{2} = 0.025$

df	0.1	0.05	0.025
1	3.078	6.314	12.706
2	1.886	2.920	4.303
3	1.638	2.353	3.182
4	1.533	2.132	2.776
5	1.476	2.015	2.571
6	1.440	1.943	2.447
7	1.415	1.895	2.365
8	1.397	1.860	2.306
9	1.383	1.833	2.262
10	1.372	1.812	2.228
11	1.363	1.796	2.201
12	1.356	1.782	2.179
13	1.350	1.771	2.160
14	1.345	1.761	2.145
15	1.341	1.753	2.131
16	1.337	1.746	2.120
17	1.333	1.740	2.110
18	1.330	1.734	2.101
19	1.328	1.729	2.093

look up $t_{n-1,\frac{\alpha}{2}}$ for $v = 19$, $\frac{\alpha}{2} = 0.025$

$= 2.09$

1. What is $P(t > 1.363)$ when $n = 12$?

$P(t > 1.363) = 0.100$

2. A restaurant has potatoes delivered in 25 kg bags. They decide to investigate whether there are actually 25 kg of potatoes in the bags. A random sample of 10 bags yields the following weights: 26.7, 24.6, 25.3, 22.8, 23.8, 24.2, 23.2, 24.5, 25.1, 26.4. Calculate a 95% confidence interval for μ.

$\bar{x} \pm t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}} = \bar{x} \pm t_{0.025,9}\frac{s}{\sqrt{n}}$

$= 24.66 \pm 2.262\frac{1.263}{\sqrt{10}}$

$= 24.66 \pm 0.903$

3. For the information given in 2, calculate a 99% confidence interval.

$\bar{x} \pm t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}} = \bar{x} \pm t_{0.005,9}\frac{s}{\sqrt{n}}$

$= 24.66 \pm 3.250\frac{1.263}{\sqrt{10}}$

$= 24.66 \pm 1.298$

4. Compare the intervals found in 2 and 3. What do you notice? Would you expect a 90% confidence interval to be wider or narrower than the intervals calculated in 2 and 3? Verify your assumption.

$\bar{x} \pm t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}} = \bar{x} \pm t_{0.05,9}\frac{s}{\sqrt{n}}$

$= 24.66 \pm 1.833\frac{1.263}{\sqrt{10}}$

$= 24.66 \pm 0.732$

Ex. 8.1.2., p62

Consider the 1-sided CI expression for a mean of a N population.

A) what values of z_α would result in a 90% CI?
B) what values of z_α would result in a 95% CI?
C) what values of z_α would result in a 99% CI?

a = 1.29 → 90% CI
b = 1.65 → 95% CI
c = 2.33 → 9.99 CI

Ex. For a N population with known σ^2, what value of $z_{\frac{\alpha}{2}}$ gives (2 sided interval) ?

a) 98% CI b) 80% CI c) 75% CI

a) $1-\alpha = 0.98$, $\alpha = 0.02$, $\frac{\alpha}{2} = 0.01$; $0.98 + 0.01 = 0.99$, $z_{\frac{\alpha}{2}} = 2.33$

b) $1-\alpha = 0.80$, $\alpha = 0.20$, $\frac{\alpha}{2} = 0.10$; $0.80 + 0.10 = 0.90$, $z_{\frac{\alpha}{2}} = 1.29$

c) $1-\alpha = 0.75$, $\alpha = 0.25$, $\frac{\alpha}{2} = 0.125$; $0.75 + 0.125 = 0.875$, $z_{\frac{\alpha}{2}} = 1.15$

Use value of $z_{\frac{\alpha}{2}}$ in:

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

to find the 2 sided confidence interval

Proportions

1. Suppose that 680 of 1000 Canadians said that cruises were good value for the money. Give a 95% confidence interval for the proportion of all Canadians who feel this way.

$\hat{p} = \frac{x}{n} = \frac{680}{1000} = 0.68$

$\sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{0.68(0.32)}{1000}} = 0.01475$

$\therefore \hat{p} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.68 \pm 1.96(0.01475)$

$= 0.68 \pm 0.0289$

→ 95% confident that the population proportion is b/w 0.651 and 0.709

→ when calculating confidence values for a proportion, use standard normal value

2. A recent poll of 280 small business owners in Canada found that 244 said they had been victimized by crime in the past year. Calculate a 90% confidence interval for p.

$\hat{p} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.87 \pm 1.645(0.02009)$

$= 0.87 \pm 0.0330$

→ conclusion: 90% confident that the population proportion is b/w 0.837 and 0.903

1. Suppose 95 of 180 third year students pass there third term test. Compute a 95% confidence interval for p.

$\hat{p} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.53 \pm 1.96\sqrt{\frac{0.53(0.47)}{180}}$

$= 0.53 \pm 1.96(0.037)$

$= 0.53 \pm 0.073$

2. In a sample of 100 ball bearings, 60 are found to be defective. Compute a 95% confidence interval for the true proportion of non-defective ball bearings.

$\hat{p} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.40 \pm 1.96\sqrt{\frac{0.4(0.6)}{100}}$

$= 0.53 \pm 1.96(0.049)$

$= 0.53 \pm 0.096$

3. A 95% confidence interval for the true proportion of male third year engineering students at McMaster is (0.35,0.55). Interpret this interval.

→ can be 95% confident that the true proportion of 3rd yr male eng students at McMaster is b/w 0.35 and 0.55

4. Calculate $\hat{p}$ and the margin of error for question 4. Since the interval is symmetric about $\hat{p}$, $\hat{p} = 0.45$.

→ Margin of error is 0.10

Ex. P293 - error example

In a sample of n=85 bearings, 10 have surface finish rougher than allowed. A point estimate of the proportion of bearings in the population that exceeds specifications. How large of a sample is required if we want to be 95% confidence that the error in using p^ to estimate p is less than 0.05?

$\hat{p} = \frac{x}{n} = \frac{10}{85} = 0.12$

$z_{\frac{\alpha}{2}}$ for 95% CI = 1.96

$E = 0.05$

$n = \left(\frac{z_{\frac{\alpha}{2}}}{E}\right)^2 p(1-p)$

$= \left(\frac{1.96}{0.05}\right)^2 (0.12)(0.88)$

$n = 163$

Aside - we are at least (100-α) confident that the error in estimate p by p^ is less than E if the sample size is:

$n = \left(\frac{z_{\frac{\alpha}{2}}}{E}\right)^2 (0.25)$

As sample size will always be a maximum for p=0.5

CM 6.1

Problem #1: **(a)** The following data shows the number of hours that 8 hospital patients slept following the administration of a certain anesthetic.

5, 5, 12, 9, 11, 6, 7, 10,

Find a 95% confidence interval for the average hours slept following the administration of the anesthetic for the sampled population.

(b) Which of the following statements is true regarding part **(a)**?

A)

```
data<-c(5,5,12,9,11,6,7,10)
t.test(data, conf.level=0.95)
```

(A) The population must be normal. **(B)** The population must follow a t-distribution.

(C) The population mean must be inside the confidence interval. **(D)** The population does not need to be normal.

(E) The population standard deviation σ must be known.

A)

Lec23 - hypothesis tests for a normal random variable

Saturday, November 12, 2022 1:30 AM

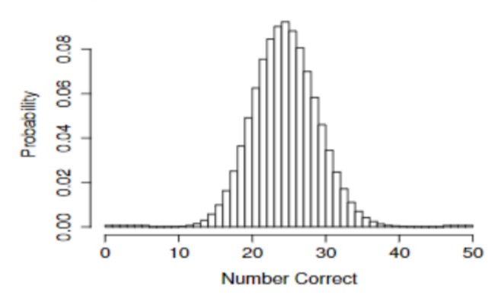
- A general introduction to hypothesis testing
- An outline of how to conduct a hypothesis test

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

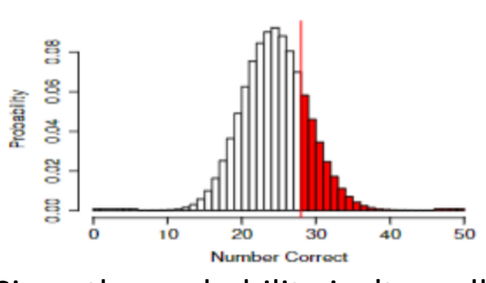
HYPOTHESIS TESTING
- Generally speaking, hypothesis testing is the practice of translating a scientific question into a hypothesis about the value of a population parameter
- Examples of questions that hypothesis testing may help answer:
  - Do more than half the adults in a certain area favour the legalization 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~Bin(100, 0.25), the probability that Jay would do as well as he did or even better is P(X>=28) = 0.278



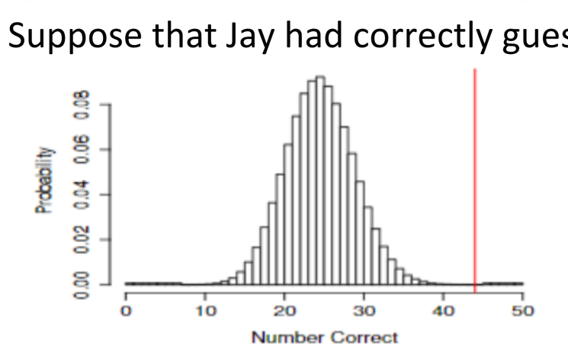
- Since the probability isn't small, we may be inclined to say that it's not unlikely to get this many correct due to chance

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>=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

  $H_0 : \mu_{fat} = 25$ vs. $H_a : \mu_{fat} \neq 25$
- A researcher wishes to test the mean fat content of a type of hotdog exceeds 25g

  $H_0 : \mu_{fat} = 25$ vs. $H_a : \mu_{fat} > 25$
- A researcher wants to test if the mean fat content of a type of hotdog <25g

  $H_0 : \mu_{fat} = 25$ vs. $H_a : \mu_{fat} < 25$

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 ⊕
  - Negative direction ⊖
- For an UPPER TAILED TEST, the rejection region will be in the pos direction only ⊕
- For a LOWER TAILED TEST, the rejection region will be in the neg direction only ⊖
- The **rejection region** is defined by the **significance level** and corresponding critical value

SIGNIFICANCE LEVEL (α)
- 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 α
- For a TWO TAILED test, we look for $\alpha/2$ in each tail
- For a ONE TAILED test, we take the full value of α 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 α and find the **critical value** from the tables
- Computer 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 σ 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:

  $H_0 : \mu = \mu_0$ vs. $H_a : \mu > \mu_0$,
- For now, assume we know the value of σ
- For H_0: μ = μ_0

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

DECISION RULES FOR HYPOTHESIS TESTS WHEN σ 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 > Zα

  $H_0 : \mu = \mu_0$ vs. $H_a : \mu > \mu_0$,
- LOWER TAILED TEST: Reject H_0 if z < -Zα

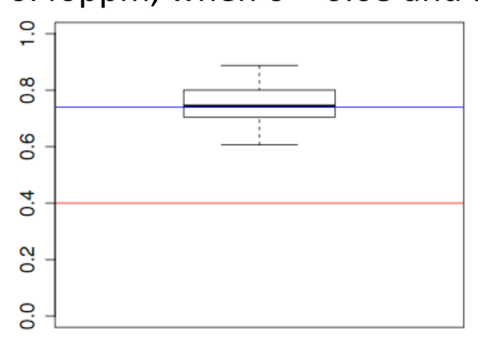
  $H_0 : \mu = \mu_0$ vs. $H_a : \mu < \mu_0$,
- TWO TAILED TEST: Reject H_0 if |z|> Zα/2

  $H_0 : \mu = \mu_0$ vs. $H_a : \mu \neq \mu_0$.

Yuxi Qin
yuxiqin.ca

comparing test statistic to critical value:

Ex. Hypothesis test about μ
- 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 0.40ppm, when σ = 0.08 and α=0.05?



→ we want to test $H_0 : \mu = 0.40$ vs $H_a : \mu > 0.40$

→ Given that σ is known, the test statistic is:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{0.74 - 0.40}{\frac{0.08}{\sqrt{16}}} = 17$$

Rejection region → $\alpha = 0.05$

∴ critical value → $z_\alpha = z_{0.05} = 1.645$

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 declaring the hypotheses & calculating the test statistic, where did the critical value come from?

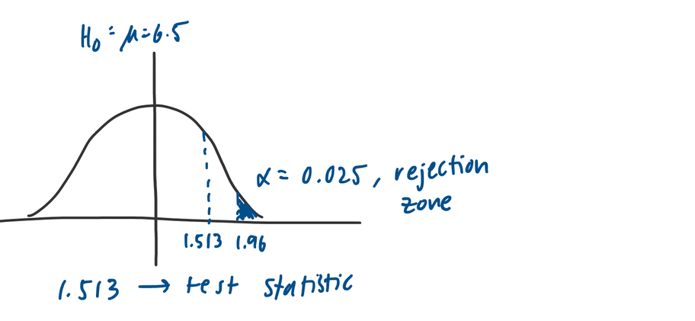
→ There's very strong evidence against $H_0 : \mu = 0.40$, meaning that the true mean of mercury is greater than 0.40 ppm

→ 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 6.5.

$H_0 : \mu = 6.5$ vs. $H_a : \mu \neq 6.5$

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{6.819 - 6.5}{\frac{1.265}{\sqrt{36}}} = 1.513$$



critical value is $z_{\alpha/2} = z_{0.025} = 1.96$

∴ because $|1.513| < 1.96$, do not reject H_0.

$\alpha = 0.025 \Rightarrow z = 1.96$, rejection region/zone

"z" critical value

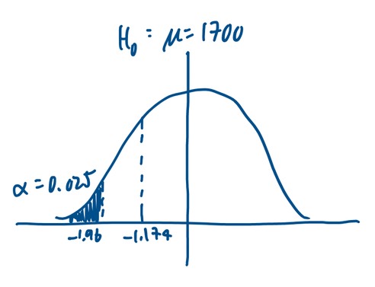
The p-value for this is equal to $2(0.0655) = 0.1310$

∴ 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 σ = $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?

$H_0 : \mu = 1700$ vs $H_a : \mu = 1600$

$$t = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{1600 - 1700}{\frac{330}{\sqrt{15}}} = -1.174$$



The critical value is $z_{\alpha/2} = z_{0.025} = 1.96$

∴ bc $|1.174| < 1.96$, do not reject H_0

The p-value for this is equal to $2(0.1210) = 0.2420$

∴ bc $0.2420 > 0.05$, do not reject H_0

CM 3.6

Problem #6: **(a)** Erythromycin is a drug that has been proposed to possibly lower the risk of premature delivery. A related area of interest is its association with the incidence of side effects during pregnancy. Assume that 30% of all pregnant women complain of nausea between the 24th and 28th week of pregnancy. Furthermore, suppose that of 182 women who are taking erythromycin regularly during this period, 65 complain of nausea. Find the *p*-value for testing the hypothesis that incidence rate of nausea for the erythromycin group is greater than for a typical pregnant woman.

(b) At the 10% significance level, what is the conclusion of the above hypothesis test?

$\hat{p} = \frac{x}{n} = \frac{65}{182}$ $p_0 = 0.3$ $n = 182$

$q_0 = 1 - p_0 = 0.7$

a) $z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = 1.682240165$

p-value $= p(z\text{-value}) = 0.95374$

$1 - p(z) = 0.04626 \rightarrow \times 2$

$= 0.09252$

$= 0.0925$ ½

b) $H_0 : \mu = \mu_0$

$H_a : \mu > \mu_0$

$\alpha = 0.10$

∴ because $(p\text{-value} = 0.09252) < (\alpha = 0.10)$, we can conclude that the rate of nausea in erythromycin group is greater than the typical pregnant woman.

→ reject H_0

∴ D ✓

CM 3.7

Problem #7: In a 1868 paper, German physician Carl Wunderlich reported based on over a million body temperature readings that the mean body temperature for healthy adults is 98.6° F. However, it is now commonly believed that the mean body temperature of a healthy adult is less than what was reported in that paper. To test this hypothesis a researcher measures the following body temperatures from a random sample of healthy adults.

98.1, 98.3, 98.8, 98.4, 97.7, 98.1, 98.2

(a) Find the value of the test statistic.

(b) Find the 10% critical value.

(c) Work through this example on R, and then use R to find the *p*-value for the relevant hypothesis test.

(d) Since the sample size is less than 30, the above analysis requires that the population follows a normal distribuiton. What method could be used to check this assumption?

$\mu_0 = 98.6$

$H_0 : \mu = \mu_0$

$H_a : \mu < \mu_0$

Reject H_0 if $t_0 < -t_{\alpha, n-1}$

$\bar{x} = 98.22857143$

$\sigma = 0.3352326839$

$n = 7$

a) Test statistic : (using student-t) $t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n-1}}} = -2.9314$

b) $\alpha = 0.10$

$\frac{\alpha}{2} = 0.05$

$t_{\alpha, n-1} = t_{0.10, 6} = -1.440$

$P(T < crit) = 0.10$

$P(T > -1.28) = 0.10$

c) Rstudio → 0.01312

```
temp_data<-c(98.1, 98.3, 98.8, 98.4, 97.7, 98.1, 98.2)
t.test(temp_data,mu=98.6,alt="less",conf.level=.90)
```

Lec24 - Hypothesis tests a for a student-t random variable

Yuxi Qin
yuxiqin.ca

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 σ was known
- Of course, in practice, we usually don't know the true value of σ
- We will introduce a hypothesis test that does not require σ, 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 then 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 σ IS UNKNOWN
- To construct confidence intervals when σ 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 σ is unknown to the results of WS Gosset (Student)
- When conducting a hypothesis test, if the σ is unknown, population is normal, and u=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 $H_a : \mu > \mu_0$
- **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 $|t| > t_{n-1,\alpha/2}$
  $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$

DISCUSSION
- That brings us to the end of our discussion on how to conduct hypothesis tests for u
- Next, we will discuss hypothesis tests for the proportion
- Finally, we will discuss the errors associate with hypothesis testing

CM 6.5

Problem #5: Suppose that we want to test the hypothesis that mothers with low socioeconomic status (SES) deliver babies whose birthweights are different than "normal". To test this hypothesis, a list of birthweights from 80 consecutive, full-term, live-born deliveries from the maternity ward of a hospital in a low-SES area is obtained. The mean birghweight is found to be 115 oz with a sample standard deviation of 21 oz. Suppose that we know from nationwide surveys based on millions of deliveries that the mean birthweight in the United States is 120 oz.

At α = .07, can it be concluded that the average birthweight from this hospital is different from the national average?

(a) Find the value of the test statistic for the above hypothesis.
(b) Find the critical value.
(c) Find the *p*-value.
(d) What is the correct way to draw a conclusion regarding the above hypothesis test?

2-sided!

National: $\mu_0 = 120$ SES: $\sigma = 21$, $n = 80$, $\bar{x} = 115$, $\alpha = 0.07$, $\frac{\alpha}{2} = 0.035$

a) Test statistic $= z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{115 - 120}{\frac{21}{\sqrt{80}}} = -2.12958855 = -2.13$ ✓

b) Critical value: $z_{\frac{\alpha}{2}} = z_{0.035} = z(0.93 + 0.035) = z(0.965) = 1.81$ ✓

c) P-value: $p(\underbrace{-2.12958855}_{\text{test statistic}}) = 0.0166$ p-value $= 2(0.0166) = 0.0332$ ✓

d) Since (p-value $= 0.0332$) $<$ ($\alpha = 0.07$), reject H_0 and conclude that there is evidence that the average birthweight is smaller/different from national average. (4c, significance level 7%)

∴ C ✓

(C) If the answer in (c) is less than 0.07 then we conclude at the 7% significance level that the average birthweight from this hospital is different from the national average.

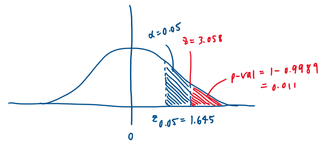
Ex. A call centre claims that the average wait time for their customers is no more than 2.00 minutes. You feel that the wait times are longer than 2.00 minutes on average. To investigate, you make 70 calls at randomly selected times and find an average wait time of 2.438 minutes. Does this provide strong evidence that the true mean wait time is greater than two minutes? Assume that the population is normally distributed with σ = 1.2 and use a significance level of σ = 0.05. Formally, we are testing:

$H_0 : \mu = 2$ 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-value $< \alpha \Rightarrow$ reject H_0 otherwise do not reject.
Since 0.0011 < (α=0.05) we can reject H_0 and conclude that there is evidence that the average wait time is greater than 2 mins.

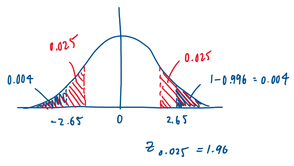
Ex. Suppose it is known that the average starting wage for a recent MBA graduate in Canada is $71k. Do graduates from the MBA warehouse have a different mean starting salary? A random sample of 50 recent MBA grads from the school had a mean starting wage of $66.5k. Carry out an appropriate hypothesis test. Assume σ = $12k, α=0.05. Formally, we're testing:

$H_0 : \mu = 71,000$ 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?



∴ p-value = 0.004 + 0.004 = 0.008

Leave out: 95% CI = $66500 \pm 1.96\left(\frac{12000}{\sqrt{50}}\right)$

because p-value $< (\alpha = 0.05) \Rightarrow$ reject H_0
or since $|z| = 2.65 > z_{0.025} = 1.96 \Rightarrow$ reject H_0

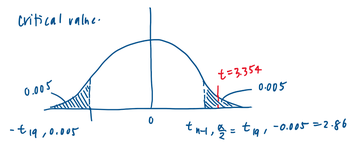
Students t-test: Ex.
A manufacturer produces bolts that are supposed to have a weight of 2.50 grams. A random sample of 20 bolts yielded a sample mean of 2.53 grams, with a standard deviation of 0.04 grams. Test whether the population mean weight is different from 2.50, using a 1% significance level. Formally, we're testing:

$H_0 : \mu = 2.50$ vs. $H_a : \mu \neq 2.50$

The test statistic is given by

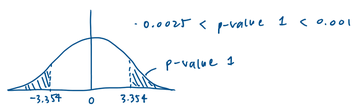
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{2.53 - 2.50}{0.04/\sqrt{20}} = 3.354$$

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:



p-value = pvalue 1 + pvalue 2
= 2(0.0025 < p-value < 0.001)
= 0.005 < p-value < 0.002

We have α=0.1 ∴ since p-value is less than α ⇒ 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	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

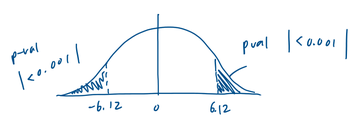
$H_0 : \mu_1 = 1.86$ vs. $H_1 : \mu_1 \neq 1.86$ at $\alpha = 0.05$

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
α = 0.05 and v = n-1 = 5 is tα/2, n-1 = t0.025,5 = 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



∴ p-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.

$H_0 : \mu = 6.5$ vs $H_a : \mu \neq 6.5$ n=36

$t = \frac{\mu_0}{\sqrt{n}} = \frac{6.819 - 6.5}{\frac{1.265}{\sqrt{36}}} = 1.5113$ = test statistic

- critical value: presume 5% significance level

$t_{\frac{\alpha}{2}, n-1} = t_{0.025, 35} = 2.042$ (take lower v-value when between n values)

∴ $|1.5113| < 2.042$, do not reject H_0

p-value for this test is:
$2(0.1 > \text{p-value} > 0.05)$
$0.20 > \text{p-value} > 0.1$

we have α=0.05, p-val > 0.1 ⟶ ∴ 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.

$H_0 : \mu_2 = 1.86$ vs $H_a : \mu_2 \neq 1.86$ $\alpha = 0.05$ $t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$

Test statistic: $t = \frac{2.55 - 1.86}{\frac{0.084}{\sqrt{6}}} = 20.12$

⇒ Reject the null hypothesis
⇒ There is significant evidence that the mean of the second spring type is different from 1.86

Critical value for $\alpha = 0.05$, $n = 6$ two-sided, so $\frac{\alpha}{2} = 0.025$

$v = n - 1 = 5$
Critical value = 2.571

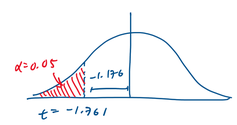
- Compare to previous class on CIs — the 95% confidence intervals do not include the tested value

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?

$H_0 : \mu = 1700$ vs $H_a : \mu \neq 1700$

$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{1600 - 1700}{\frac{330}{\sqrt{15}}} = -1.174$ Test statistic

Critical value: $t_{\frac{\alpha}{2}, n-1} = t_{0.025, 14} = 2.145$



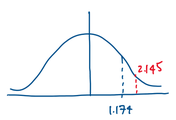
$|-1.174| < 2.145 \Rightarrow$ Do not reject H_0

$H_0 : \mu = 1700$, $H_a : \mu < 1700$

$t = -1.174$, $\alpha = 0.05$ ⇒ Reject if $t < -t_{\text{critical}(n-1, \alpha)}$

⇒ Do not reject H_0

P-value calculation:
$H_0 : \mu = 1700$ vs $H_a : \mu \neq 1700$
$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{1600 - 1700}{\frac{330}{\sqrt{15}}} = -1.174$ Test statistic

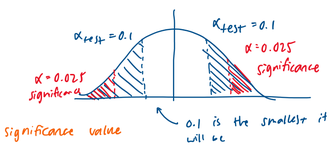


$|1.174| < 2.145$
↑ critical value for $t_{\frac{\alpha}{2}, n-1} = t_{0.025, 14}$

⇒ Do not reject H_0

Calculate the p-value
- go to t-table)
- look up α-value, corresponding to t-value of 1.174, v=14
- this gives α_{test} of

$0.1 < \alpha_{test} < 2.5$
$0.2 < \text{p-value} < 0.5$



p-value is > 0.2
$0.2 > 0.5$, $\alpha = 0.05$ = significance value

Total p value > 0.2

Total $\alpha_{significance} = 0.5$

I have no idea what all of this means ;-;

Lec25 - Hypothesis testing for proportions

Yuxi Qin
yuxiqin.ca

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 > 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

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

  Approximately has a standard normal distribution provided
  np > 15 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 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-α)% 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 H0: μ = μ_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

  $$\left|\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right| \le z_{\alpha/2}$$
- But this implies we do not reject if (falls within range)

  $$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu_0 \le \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$
- In other words, we do not reject H_0 if μ_0 is in the confidence interval and we do reject if μ_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 α=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

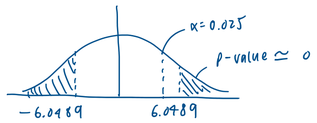
The 2006 census revealed that approximately 23% of Canadians between 25 and 64 years of age have a university degree. In a certain area, a random sample of 200 adults in this age group revealed that 10 had a university degree. Test whether the proportion with a degree in this area differs from the rest of Canada. Use a 5% significance level.

Formally, we're testing:

$H_0 : p = 0.23$ 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
H_a: p ≠ 0.23

$\hat{p} = \frac{x}{n} = \frac{10}{200} = 0.05$

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.05 - 0.23}{\sqrt{\frac{0.23(1-0.23)}{200}}}$$

Prob for $z = -6.0489 \simeq 0$
$z\ (\simeq 0) < 0.05$

Pvalue < α, $\frac{\alpha}{2}$ 0.05 ⇒ reject H_0

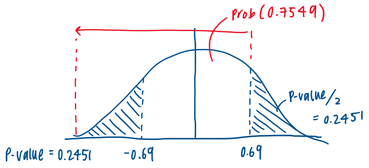
∴ There is evidence to suggest that the % of Canadians with a university degree is different from 23%

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 0.275.



$H_0 : p = 0.3$ vs $H_a : p \neq 0.3$

$$z = \frac{0.275 - 0.3}{\sqrt{\frac{0.3(0.7)}{160}}} = -0.690$$

$z_{0.69}$ ⇒ prob = 0.7549
1 − 0.7549 = 0.2451

5% sig. level: $(z_{\frac{\alpha}{2}} = 1.96)$

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

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:

$H_0 : \mu = 41$ 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 α = 0.05, since its absolute value is greater than the critical value $Z_{0.025}$ = 1.96

$$CI = \bar{x} \pm z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$
$$= 39 \pm 1.96\left(\frac{2.4}{\sqrt{25}}\right)$$
$$= 39 \pm 0.94$$
$$= [38.06, 39.94]$$

$\bar{x} = 39$
$\sigma = 2.4$
$n = 25$
$z_{\frac{\alpha}{2}}$ 95% CI = 1.96

Rope Example II

- A 95% Confidence Interval for μ based on $\bar{x}$

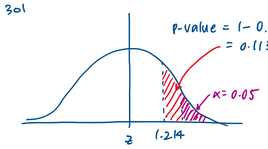
  $$39 \pm 1.96\frac{2.4}{\sqrt{25}} = 39 \pm 0.94 = (38.06, 39.94).$$
- Therefore, we are 95% confident that, based on the sample mean, the population mean lies between 38.06kg and 39.94kg
- This test and the interval are related, the fact that 41 does not lie inside the 95% confidence interval for μ, ie. (38.06, 39.94), is equivalent to rejecting the null hypothesis that μ=41 at significance level α=0.05

❶ 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? → a
Is there significant evidence that the prevalence of IFG is different from 6.3 percent? → b
Construct a 95% confidence interval for the prevalence of IFG. → c
Compare your results.

a) H_0: p = 0.063 vs H_a: p > 0.063

proportion: $\hat{p} = \frac{x}{n} = \frac{24}{301} = 0.08$

$$z = \frac{p - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.08 - 0.063}{\sqrt{\frac{0.063(1-0.063)}{301}}}$$



z test statistic = 1.214 → Prob = 0.8869
p-value = 1 − 0.8869
= 0.1131

∴ because (p-value = 0.1131) > (α = 0.05), do not reject H_0

b) H_0: p = 0.063 vs H_a: p ≠ 0.063

$$z = \frac{p - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.08 - 0.063}{\sqrt{\frac{0.063(1-0.063)}{301}}} = 1.214$$

2-sided critical value: $z_{\frac{\alpha}{2}} = z_{0.025}$

z-test statistic = 1.214
⇒ Prob ≃ 0.8868
1 − 0.8868 = 0.1131

2(0.1131) = 0.2262
total p-value

∴ (p-value = 0.2262) > (α = 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. $\hat{p} = 0.08$, $p_0 = 0.063$, $n = 301$

$$\hat{p} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$
$$= 0.08 \pm 1.96\sqrt{\frac{0.8(1-0.08)}{301}}$$
$$= 0.08 \pm 1.96\,(0.0156)$$
$$= 0.08 \pm 0.0307$$

∴ 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

- 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)** = **β**, so the power is **1-β**
- 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) = α**, where **α** 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 β is the probability of a type II error, ϕ(z) is the probability to the left of z in the standard normal distribution, and δ=μ-μ_0

SAMPLE SIZE SELECTION

- The sample size required to produce a specified type II error with probability β given δ and α:
  - 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 δ=μ-μ_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.
(a) What is the type I error probability if the critical region is defined as $\bar{x} < 11.5$ kilograms?
(b) Find β for the case in which the true mean elongation is 11.25 kilograms.
(c) Find β for the case in which the true mean is 11.5 kilograms.

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

$H_0: \mu = 12$
$H_a: \mu < 12$
$\bar{x} < 11.5$
$\sigma = 0.5$
$n = 4$

a) $\alpha = P\left(z \le \frac{11.5-12}{\frac{0.5}{\sqrt{4}}}\right) = P(z \le -2) = 1 - P(z \le 2) = 1 - 0.977 = 0.2275$

The prob. of rejecting the null hypothesis when it is true is 0.2275

b) $\beta = P(\text{accept } H_0 \text{ when } \mu = 11.25) = P(\bar{x} > 11.5 \mid \mu = 1.5)$

$P\left(\bar{x} > 11.5 \mid \mu = 11.25\right) = P\left(z > \frac{11.5-11.5}{\frac{0.5}{\sqrt{4}}}\right)$

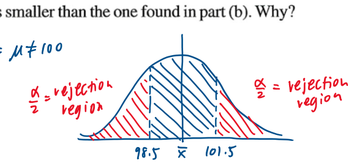
The probability of failing to reject the null hypothesis when it is false is 0.15866

$= P(z > 1) = 1 - P(z \le 1)$
$= 1 - 0.84134 \rightarrow \beta = 0.15866$

c) Power $= 1 - \beta$
$= 0.84134$

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.
(a) If the acceptance region is defined as $98.5 \le \bar{x} \le 101.5$, find the type I error probability α.
(b) Find β for the case in which the true mean heat evolved is 103.
(c) Find β for the case where the true mean heat evolved is 105. This value of β is smaller than the one found in part (b). Why?

$H_0: \mu = 100$ vs $H_1: \mu \neq 100$
$n = 9$



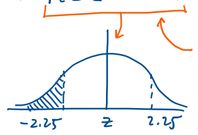
$\alpha = P(\bar{x} \le 98.5) + P(\bar{x} > 101.5)$
$= P(\bar{x} \le 98.5) + (1 - P(x \le 101.5))$

$z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$ $\quad \mu_0 = \mu = 100$, $\sigma = 2$, $n = 9$

$\alpha = P\left(z \le \frac{98.5-100}{\frac{2}{\sqrt{9}}}\right) + \left[1 - P\left(z \le \frac{101.5-100}{\frac{2}{\sqrt{9}}}\right)\right]$

$\alpha = P(z \le -2.25) + [1 - P(z \le 2.25)]$

$P(z \le -2.25) = 1 - P(z \le 2.25)$



$\alpha = 1 - P(z \le 2.25) + [1 - P(z \le 2.25)]$

$\alpha = 0.0244$

P(Type I error) = α = 0.0244

b) $\mu = 103$
- β is the probability of a type II, failing to reject null hypothesis when it's actually false
- acceptance region is defined as:
  $98.5 \le \bar{x} \le 101.5$, $\mu = 103$

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

$= P(x \le 101.5) - P(\bar{x} \le 98.5)$
∴ accept one region
$= P\left(z \le \frac{101.5-103}{\frac{2}{\sqrt{9}}}\right) - P\left(z \le \frac{98.5-103}{\frac{2}{\sqrt{9}}}\right)$
$= P\left(z \le \frac{-1.5}{0.666}\right) - P\left(z \le \frac{-9.5}{0.666}\right)$
$= P(z \le -2.25) - P(z \le -6.75)$
$= 0.0122 - [1 - P(z \le 6.75)]$
$= 0.0122 - [1 - 1]$
$\beta = 0.0122$

c) Power $= 1 - \beta = 1 - 0.0122 = 0.9878$
— same method as part b
$\beta = 0$
β 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

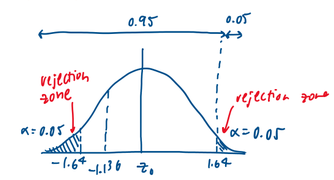
Sample Test #3

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 $\alpha = .05$).

(a) Do not reject H_0 since −1.13 is not less than −1.64. The inventory should not be postponed.
(b) Do not reject H_0 since −1.13 is not less than −1.64. The inventory should be postponed.
(c) Do not reject H_0 since −1.13 is not less than −1.96. The inventory should not be postponed.
(d) Do not reject H_0 since −1.13 is not less than −1.96. The inventory should be postponed.
(e) Do not reject H_0 since −1.48 is not less than −1.64. The inventory should not be postponed.

$H_0: p = 0.02$
$H_a: p < 0.02$
$n = 1000$
$\hat{p} = \frac{x}{n} = \frac{15}{1000} = 0.015$
$p_0 = 0.02$
$\alpha = 0.05$ ← signif. level

$z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.015-0.02}{\sqrt{\frac{0.02(0.98)}{1000}}} = -1.136$
↳ test statistic



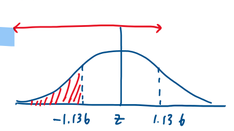
1 sided test, $\alpha = 0.05$, lower tailed
z corresponding to P(0.95) = 1.64

∴ 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

test statistic: −1.136



p-value = $1 - P(z \le 1.136)$
$= 1 - 0.8707$
$d = 0.1293$

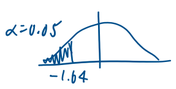
15. (Continuation of **13.**) If the true proportion of misshelved and lost books is actually .01, what is the probability that the inventory will be (unnecessarily) taken?

(a) .192 **(b)** .176 **(c)** .142 **(d)** .119 **(e)** .103

$H_0: 0.02$ — fail to reject, then "not postponed"

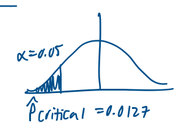
"unnecessarily taken" = not postponed when it should have been
⇒ not reject the null hypothesis when it's actually false
⇒ Type II error

critical value ⇒ $z = -1.64$
what $\hat{p}$ does this correspond to?



$z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ $\quad -1.64 = \frac{\hat{p}-0.02}{\sqrt{\frac{0.02(0.98)}{1000}}}$ for $H_0 = 0.02$

$\hat{p}_{critical} = 0.0127$



$\beta = P(\hat{p} > 0.0127 \mid \mu = 0.01)$
$= P\left(z > \frac{0.0127-0.01}{\sqrt{\frac{0.01(0.99)}{1000}}}\right)$
$= P(z > 0.8726)$
$= 1 - P(z < 0.8726)$
$= 1 - 0.80785$
$\beta = 0.192$

— we want to have a large for "Power", which means we have a small chance at making a Type II error

if does this mean 101

Power of the test: $1 - \beta = 0.80785$

— If asked to find power of a test, you need to know/find β, 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 σ

$$\mathbb{V}ar[\bar{X}] = \mathbb{V}ar\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\mathbb{V}ar\left[\sum_{i=1}^{n} X_i\right]$$
$$= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}ar\left[X_i\right] = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}.$$

- And so the standard error of $\bar{X}$ is given by

$$SE[\bar{X}] = \frac{\sigma}{\sqrt{n}}.$$

- In the test of the t-test, we do not know σ and so we use *s* to approximate it, giving:

$$SE[\bar{X}] = \frac{s}{\sqrt{n}}.$$

- Recall that the sample standard deviation is given by

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

- What about the standard error of $\bar{X}_A - \bar{X}_B$

- Recall that we assume

  $\sigma_A^2 = \sigma_B^2 = \sigma^2$

$$\mathbb{V}ar[\bar{X}_A - \bar{X}_B] = \mathbb{V}ar[\bar{X}_A] + \mathbb{V}ar[\bar{X}_B]$$
$$= \mathbb{V}ar\left[\frac{1}{n_A}\sum_{i=1}^{n_A} X_{A(i)}\right] + \mathbb{V}ar\left[\frac{1}{n_B}\sum_{i=1}^{n_B} X_{B(i)}\right]$$
$$= \frac{1}{n_A^2}\mathbb{V}ar\left[\sum_{i=1}^{n_A} X_{A(i)}\right] + \frac{1}{n_B^2}\mathbb{V}ar\left[\sum_{i=1}^{n_B} X_{B(i)}\right]$$
$$= \frac{1}{n_A^2}\sum_{i=1}^{n_A}\mathbb{V}ar\left[X_{A(i)}\right] + \frac{1}{n_B^2}\sum_{i=1}^{n_B}\mathbb{V}ar\left[X_{B(i)}\right]$$
$$= \frac{1}{n_A^2}(n_A\sigma_A^2) + \frac{1}{n_B^2}(n_B\sigma_B^2) = \frac{\sigma^2}{n_A} + \frac{\sigma^2}{n_B}.$$

- But for the independent groups t-test, we don't use σ_A or σ_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 σ^2. this is sometimes called the **pooled estimate of the variance**
- Therefore, for $n_a = n_b = n$,

$$SE(\bar{X}_A - \bar{X}_B) = \sqrt{\frac{s_p^2}{n} + \frac{s_p^2}{n}} = \sqrt{\frac{2s_p^2}{n}}.$$

- Now if $n_a \neq n_b$, then we approximate σ^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,

$$SE(\bar{X}_A - \bar{X}_B) = \sqrt{\frac{s_p^2}{n_A} + \frac{s_p^2}{n_B}} = \sqrt{s_p^2\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}.$$

THE INDEPENDENT GROUPS T-TEST

- The independent groups t-test tests of the form **hypotheses**

  $H_0 : \mu_A - \mu_B = 0$ versus $H_a : \mu_A - \mu_B \neq 0.$
- The test statistic is given by

$$t = \frac{(\bar{x}_A - \bar{x}_B) - 0}{SE(\bar{x}_A - \bar{x}_B)}.$$

- 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-α)% confidence interval for $\mu_A - \mu_B$ as follows:

$$(\bar{x}_A - \bar{x}_B) \pm t_{n_A+n_B-2,\alpha/2} SE(\bar{x}_A - \bar{x}_B).$$

- Once again, $t_{nA+nB-2,\ \alpha/2}$ is the value of the t-distribution corresponding to a given α and $v=n_A+n_B-2$

DISCUSSION

- The last procedure assumed that σ_A^2 - σ_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. 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. Variances unequal: $\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2,\nu}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$, $\quad \nu = \left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2 \Big/ \left(\frac{(s_1^2/n_1)^2}{n_1-1}+\frac{(s_2^2/n_2)^2}{n_2-1}\right)$

28. *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. *t* test for comparing two means (variances unequal): $t_\nu = \frac{\bar{x}_1-\bar{x}_2}{\sqrt{(s_1^2/n_1)+(s_2^2/n_2)}}$.

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 $n_1 = 17$ 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 α = 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.

$n_1 = 15$, $\bar{x}_1 = 8.73$, $s_1^2 = 0.35$ $\quad$ $n_2 = 17$, $\bar{x}_2 = 8.68$, $s_2^2 = 0.4$

Assume $\sigma_1^2 = \sigma_2^2$

The parameter of interest is mean rod diameter, $\mu_1 - \mu_2$

a) $H_0: \mu_1 - \mu_2 = 0$ or $\mu_1 = \mu_2$
$H_a: \mu_1 - \mu_2 \neq 0$ or $\mu_1 \neq \mu_2$

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:

$S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}$

$S_p = \sqrt{\frac{(15-1)(0.35) + (17-1)(0.4)}{15+17-2}}$

$S_p = 0.614$

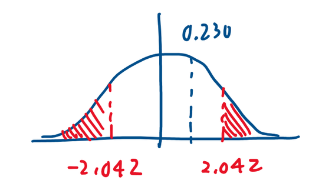
$t = \frac{8.73 - 8.68}{0.614\sqrt{\frac{1}{15}+\frac{1}{17}}}$

$t = 0.230$

Rejection region: $t_0 > t_{\frac{\alpha}{2}, n_1+n_2-2}$
$t_0 < -t_{\frac{\alpha}{2}, n_1+n_2-2}$

use $\frac{\alpha}{2}$ because 2-sided interval
$\alpha = 0.05$
$\frac{\alpha}{2} = 0.025$ $\quad$ $n_1 + n_2 - 2 = 15 + 17 - 2 = 30$

t-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 α=0.05

$t = 0.23$ ⟶ look @ tables, $v = 30$
p-value = 2(> 0.4)
p-value = > 0.8

b) 95% CI: $t_{0.025, 30} = 2.042$

$\bar{x}_1 - \bar{x}_2 \pm t_{\frac{\alpha}{2}, n_1+n_2-2}\, S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$

$8.73 - 8.68 \pm 2.042(0.614)\sqrt{\frac{1}{15}+\frac{1}{17}}$

$= 0.05 \pm 1.254(0.3542)$
$= 0.05 \pm 0.444$

$= [-0.394, 0.494]$

or: $[-0.394 \leq \mu_1 - \mu_2 \leq 0.494]$

Interpretation: because o is contained in the interval, there is insufficient evidence to conclude that the two machines produce rods with different mean diameters

TABLE • V Percentage Points $t_{\alpha,v}$ of the *t* Distribution

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

v = degrees of freedom.

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 α = 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.

$H_0: \mu_1 - \mu_2 = 0$ or $\mu_1 = \mu_2$
$H_a: \mu_1 - \mu_2 \neq 0$ or $\mu_1 \neq \mu_2$

Test statistics (variances unequal)
$\frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\left(\frac{S_1^2}{n_1}\right) + \left(\frac{S_2^2}{n_2}\right)}}$

$\alpha = 0.05$ $\quad$ what is the critical value?
$v = ?$

$V = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$ $\quad$ $S_1 = 2$, $S_2 = 8$, $n_1 = 25$, $n_2 = 25$

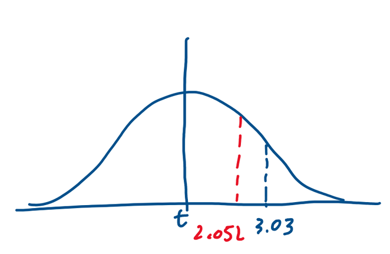
$v = 27$

$t_{0.025, 27} = 2.052$ critical value

Test statistic:
$\bar{x}_1 = 20$, $s_1 = 2$, $n_1 = 25$ $\quad$ $\bar{x}_2 = 15$, $s_2 = 8$, $n_2 = 25$

$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$

$t_0 = 3.03$



Because (test statistic = 3.03) > (critical value = 2.052), reject the null hypothesis.

P-value: $\left(\frac{P}{2}\right)$ is b/w 0.005 and 0.0025 for a test statistic of 3.03 and v=27.
x 2 bc it's divided by 2
∴ 0.005 < P_{value} < 0.01

b) $H_0: \mu_1 - \mu_2 = 0$ $\quad$ test stat = 3.03
$H_1: \mu_1 - \mu_2 > 0$ $\quad$ crit. value = $t_{0.05, 27}$ = 1.703

3.03 > 1.703
↳ reject null hypothesis

c) unequal variances:
$\bar{x}_1 - \bar{x}_2 \pm t_{\frac{\alpha}{2}, v}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_1}}$ $\quad$ $\bar{x}_1 = 20$, $S_1 = 2$, $n_1 = 25$ $\quad$ $\bar{x}_2 = 15$, $S_2 = 8$, $n_2 = 25$

Part a — $t_{\frac{\alpha}{2}, v} = 2.052$ $\quad$ Part b — $t_{\frac{\alpha}{2}, v} = 1.703$

CI = [1.616, 8.384] $\quad$ CI = [2.191, 7.809]

Lec29 - Simple Linear Regression

Yuxi Qin
yuxiqin.ca

Tuesday, November 29, 2022 4:38 AM

- Least squares estimates (sec 11.1, 11.2)
- Prediction

MODELLING DATA VIA A LINE I

- Consider the following data set:

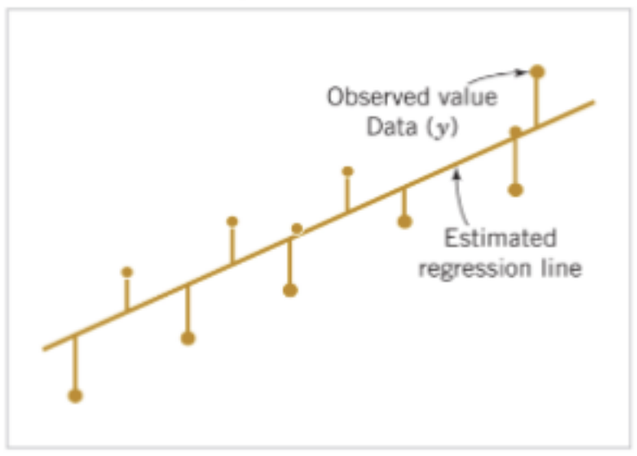
Sleep (hours)	Midterm Mark
4.6	68
8.5	83
3.2	51
5.5	76
7.0	89
6.2	75
6.4	84
5.9	71
1.5	52
10.5	91
7.3	94
5.4	74
4.9	68
3.7	56
2.2	45

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,
- β_0 is the (theoretical) intercept,
- β_1 is the (theoretical) slope,
- $\epsilon_i \backsim \mathcal{N}(0, \sigma^2)$ iid, is the (theoretical) error at i,
- $Y_i \backsim \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2)$.

FITTING THE MODEL

- In reality, we don't know the theoretical values, ie. The true values of β_0 and β_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 haha $y=mx+b$!

The slope β_1:

- In general, for every 1 unit increase in X, the response variable Y increases by β_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 β_0:

- In general, when X is 0 when the predicted value of Y is β_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
- $SS_E = SS_T - \hat{\beta}_1 S_{xy}$,

  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 β_0 and β_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)	$(-0.92)^2$	68	(68-71.8)	3.496
	8.5	(8.5 - 5.52)	$(2.98)^2$	83	(83-71.8)	33.376
	3.2	(3.2 - 5.52)	$(-2.32)^2$	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	(2.2 - 5.52)	$(-3.32)^2$	45	(45-71.8)	88.976
Sum			77.544			457.66

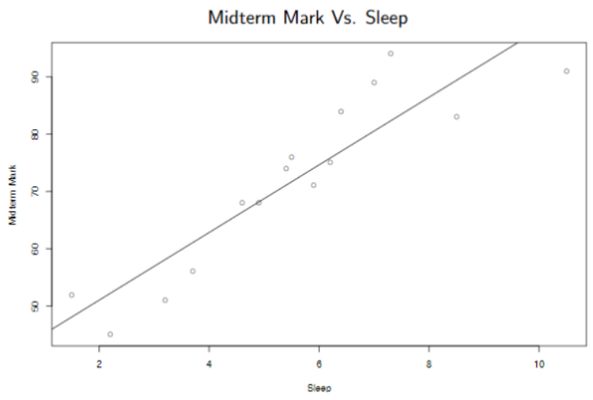
∴ Our least square estimators are:

$$b_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{457.66}{77.544} = 5.902$$

$$b_0 = \bar{y} - b_1 \bar{x} = 71.8 - 5.902(5.52) = 39.221$$

∴ Our SLR model is given by: $\hat{y}_i = 39.221 + 5902\, x_i$

→ 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?

$$\hat{y}_i = 39.221 + 5902\, x_i = 39.221 + 5.902(4) = 62.83$$

- 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 σ^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.

$n=14$
$\sum y_i = 572$
$\sum y_i^2 = 23530$
$\sum x_i = 43$
$\sum x_i^2 = 157.42$
$\sum x_i y_i = 1687.80$

a) Slope?

$b_1 = \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$

$\bar{x} = \frac{\sum x_i}{n}$, $\bar{x} = \frac{43}{14} = 3.0714$

$\bar{y} = \frac{\sum y_i}{n}$, $\bar{y} = \frac{572}{14} = 40.857$

$S_{xx} = \sum x_i^2 - n\bar{x}^2$
$= 157.42 - 14(3.0714)$
$S_{xx} = 25.348$

$S_{xy} = \sum x_i y_i - n\bar{x}\bar{y}$
$= 1697.8 - 14(3.0714)(40.857)$
$S_{xy} = -59.04$

$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-59.04}{25.348}$

$\hat{\beta}_1 = -2.33$ (slope)

intercept: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$\beta_0 = (40.857) - (-2.330)(3.0714)$

$\beta_0 = 48.013$ (intercept)

$\sigma^2 = ?$

$= \frac{SS_E}{n-2}$ not on eqn. sheet

$SS_E = SS_T - \hat{\beta}_1 S_{xy}$

$SS_T = \sum y_i^2 - n\bar{y}^2$
$= 23{,}530 - 14(40.857)^2$
$SS_T = 159.71$

$SS_E = 159.71 - (-2.33)(-59.04)$
$SS_E = 22.123$

$\sigma^2 = \frac{SS_E}{n-2} = \frac{22.123}{14-2}$

$\sigma^2 = 1.8436$

b) $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
$\hat{y} = 48.013 + (-2.33)(4.3)$
$\hat{y} = 37.99$

$\hat{\beta}_0 = 48.013$
$\hat{\beta}_1 = -2.33$
$x = 4.3$

c) $x = 3.7$:

$\hat{y} = 48.013 + (-2.33)(3.7)$
$\hat{y} = 39.39$

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:

$e_i = y_i - \hat{y}_i = 46.1 - 39.39$

$e_i = 6.71$

Lec30 - properties of least squares estimators, hypothesis tests

Yuxi Qin
yuxiqin.ca

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:

$$H_0 : \beta_1 = 0 \text{ vs. } H_a : \beta_1 \neq 0$$

$$t = \frac{b_1 - 0}{\widehat{SE}(b_1)}.$$

- The critical value is t-distributed with **v=n-2** degrees of freedom for some significance level, α
- 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:

$$H_0 : \beta_0 = 0 \text{ vs. } H_a : \beta_0 \neq 0$$

$$t = \frac{b_0 - 0}{\widehat{SE}(b_0)}.$$

STANDARD ERRORS

- To perform a test or construct a confidence interval, 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

$$s^2 = \frac{\sum_i (Y_i - \hat{Y}_i)^2}{n-2}$$

ALGEBRAIC DECOMPOSITION OF SUMS OF SQUARES

- For the ANOVA procedure, the **Analysis of Variance Identity** is given by

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

- Ie. $SS_T = SS_R + SS_E$
  Where:
  - SS_T = Total corrected sum of squares
  - SS_R = Regression sum of squares
  - SS_E = 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 $F_{\alpha, 1, n-2}$
- MS stands for **Mean Square**
- The F-distribution has 1 numerator and n-2 denominator, degrees of freedom

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

Estimate s^2 → $s^2 = \frac{\sum_i (Y_i - \hat{Y}_i)}{n-2} = \frac{585.32}{13} = 45.0242$

Now compute the standard error of b_1:

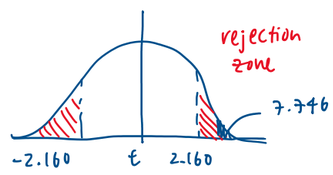
$\widehat{SE}(b_1) = \sqrt{\frac{s^2}{\sum_i (X_i - \bar{X})^2}} = \sqrt{\frac{45.0241}{77.544}} = 0.762$

Now perform the hypothesis test about β_1

$H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$, at $\alpha = 0.05$

for this test —

$t = \frac{b_1 - \beta_1}{\widehat{SE}(b_1)} = \frac{5.902 - 0}{0.762} = 7.746$

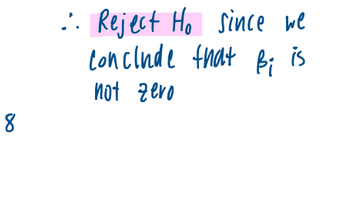


critical value is $t_{\frac{\alpha}{2}, n-2} = t_{0.025, 13} = 2.160$

∴ since $|7.746| > 2.160$, we can reject H_0 and conclude that there is evidence of linear relationship

Using P-values: $v = n-2 = 13$, $t = 7.746$, $\alpha = 0.05$

Pvalue $< 2(0.0005) < 0.001$



∴ since (p-value < 0.001) $<$ (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

$$SS_E = SS_T - SS_R = 173.38 - 152.13 = 21.25$$

The analysis of variance for testing H_0: $\beta_1 = 0$ is summarized in the computer output in Table 11.2. The test statistic is $f_0 = MS_R/MS_E = 152.13/1.18 = 128.86$, for which we find that the P-value is $P \simeq 1.23 \times 10^{-9}$, so we conclude that β_1 is not zero.

Frequently computer packages have minor differences in terminology. For example, sometimes the regression sum of squares is called the "model" sum of squares and the error sum of squares is called the "residual" sum of squares.

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
Constant	74.283 ← $\hat{\beta}_0$	1.593	46.62	0.000
HC level	14.947 ← $\hat{\beta}_1$	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	P
Regression	1	152.13	152.13	128.86	0.000
Residual error	18	21.25 ← SS_E	1.18 ← $\hat{\sigma}^2$		
Total	19	173.38			

Predicted Values for New Observations

New obs	Fit	SE Fit	95.0% CI	95.0% PI
1	89.231	0.354	(88.486, 89.975)	(86.830, 91.632)

Values of Predictors for New Observations

New obs	HC Level
1	1.00

Use ANOVA approach to test for significance of regression:

$SS_T = 173.38$
$\hat{\beta}_1 = 14.947$
$S_{xy} = 10.17744$
$n = 20$

$SS_R = \hat{\beta}_1 S_{xy} = (14.947)(10.17744)$
$SS_R = 152.13$

$SS_E = SS_T - SS_R = 173.38 - 152.13$
$SS_E = 21.25$

$F_0 = \frac{SS_R}{\frac{SS_E}{n-2}} = \frac{152.13}{\frac{21.25}{20-2}}$
$F_0 = 128.86$

F-table for $v_1 = 1$, $v_2 = 18$
$F = 4.41$

$F_0 > F$
∴ Reject H_0 since we conclude that β_1 is not zero

11.4.1 WP Recall the regression of percent body fat on BMI from Exercise 11.2.1.

a. Estimate the error standard deviation.
b. Estimate the standard deviation of the slope.
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?

$\sum_{i=1}^{n} x_i = 6322.28$ $\sum_{i=1}^{n} x_i^2 = 162674.18$

$\sum_{i=1}^{n} y_i = 4757.90$ $\sum_{i=1}^{n} y_i^2 = 107679.27$

$\sum_{i=1}^{n} x_i y_i = 125471.10$

$\hat{\sigma}$

$\bar{x} = \frac{\sum x_i}{n}$ $\bar{x} = 25.899$

$\bar{y} = \frac{\sum y_i}{n}$ $\bar{y} = 19.03$

$S_{xx} = \sum x_i^2 - n\bar{x}^2 = 162674.18 - (250)(25.289)^2$
$S_{xx} = 2789.28$

$S_{xy} = \sum x_i y_i - n\bar{x}\bar{y} = 125471.1 - (250)(25.289)(19.03)$
$S_{xy} = 5147.996$

$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{5147.996}{2789.28}$
$\hat{\beta}_1 = 1.846$

a) $\hat{\sigma}^2 = \frac{SS_E}{n-2} = \frac{SS_T - \hat{\beta}_1 S_{xy}}{n-2}$

— still need to calculate SS_T

$SS_T = \sum y_i^2 - n\bar{y}^2 = 107679.27 - 250(19.03)^2$
$SS_T = 17128.82$

$\hat{\sigma}^2 = \frac{17128.82 - 1.846(5147.996)}{248}$

$\hat{\sigma}^2 = 30.756$
$\hat{\sigma} = 5.546$

b) Standard deviation of the slope:

$\widehat{SE}(b_1) = \sqrt{\frac{\sigma^2}{\sum (x_i - \bar{x})^2}} = \sqrt{\frac{\sigma^2}{S_{xx}}} = \sqrt{\frac{30.756}{2789.282}}$

$\widehat{SE}(b_1) = 0.105$

c) T-statistic for the slope: H_0: $\beta_1 = 0$

$t = \frac{b_1 - 0}{\widehat{SE}(b_1)} = \frac{1.846}{0.105}$ $t = 17.58$

d) $\alpha = 0.05$, $v = n-2 = 248$
p-value ≈ 0, for $t = 17.48$, $v = 248$
∴ reject H_0

Lec31 - confidence and prediction intervals, omit the CI

Yuxi Qin
yuxiqin.ca

Tuesday, November 29, 2022 5:03 PM

- 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 α
- If H_0 is rejected, then there is evidence of a linear relationship between X and Y
- We can also build a (1-α)% confidence interval for β_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 β_0

STANDARD ERRORS

- We have outlined how to conduct a hypothesis tests and build a confidence interval for β_1
- In addition, we have learned that we can do the same for β_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 β_1 is given by

$$b_1 \pm t_{\alpha/2,n-2}\widehat{SE}(b_1)$$

- A confidence interval for β_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 β_1
- Again, we can also conduct inference about β_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 (mmm y=mx+b)
- For this predicted value, we can compute a **Prediction Interval (PI)**
- A (1-α)% 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

Estimate s^2 → $s^2 = \frac{\sum_i (Y_i - \hat{Y}_i)}{n-2} = \frac{585.32}{13} = 45.0242$

Now compute the standard error of b_1:

$$S\hat{E}(b_1) = \sqrt{\frac{s^2}{\sum_i (X_i - \bar{X})^2}} = \sqrt{\frac{45.0291}{77.544}} = 0.762$$

Now perform the hypothesis test about β_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 $t_{\frac{\alpha}{2},n-2} = t_{0.025,13} = 2.160$

∴ 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 β_1
A 95% confidence interval for β_1 is given by:

$$b_1 \pm t_{\frac{\alpha}{2},n-2}\,\widehat{SE}(b_1)$$

$$5.902 \pm 2.160\,(0.762)$$
$$5.902 \pm 1.646$$

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:

$$\hat{y}_h \pm t_{\frac{\alpha}{2},n-2}\sqrt{s^2\left(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_i (X_i - \bar{X})^2}\right)}$$

$$68.731 \pm t_{\frac{\alpha}{2},n-2}\sqrt{45.0241\left(1 + \frac{1}{15} + \frac{(5-5.52)^2}{77.544}\right)}$$

$$68.731 \pm 2.160\sqrt{6.941377}$$
$$68.731 \pm 14.99337$$

∴ we can be 95% confident that a Student who gets 5hrs 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,
(a) Find a 95% confidence interval for the slope.
(b) Find a 95% confidence interval for the mean percent body fat for a man with a BMI of 25.
(c) Find a 95% prediction interval for the percent body fat for a man with a BMI of 25.
(d) Which interval is wider, the confidence interval or the prediction interval? Explain briefly.

$\bar{x} = 25.899$
$\bar{y} = 19.03$
$S_{xx} = 2789.28$
$S_{xy} = 5147.996$
$\beta_1 = 1.846$
$SS_T = 17128.82$
$\hat{\sigma} = 5.546$
$\widehat{SE}\beta_1 = 0.105$
$t = 17.58$
$n = 250$

$\beta_0 = \bar{y} - \beta_1 \bar{x}$
$= 19.032 - (1.846)(25.289)$
$\beta_0 = -27.643$

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
$\hat{y} = -27.643 + 1.846x$ ← regression model

a) 95% confidence interval on β_1:

$\hat{\beta}_1 \pm t_{\frac{\alpha}{2},n-2}\, SE(\hat{\beta}_1)$
$= 1.846 \pm t_{0.025,248}\,(0.105)$
$= 1.846 \pm 1.96\,(0.105)$
$= 1.846 \pm 0.2058$
$= [1.639 \leq \beta_1 \leq 2.05]$

b) 95% CI for mean when $X_0 = 25$

$\hat{\mu} \pm t_{\frac{\alpha}{2},n-2}\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}}\right)}$

→ calculate $\hat{\mu}$ from regression model

$\hat{\mu} = -27.643 + (1.846)(25)$
$\hat{\mu} = 18.498$, $t_{0.025,248}$

↳ $18.498 \pm 1.97\sqrt{30.756\left(\frac{1}{250} + \frac{(25-25.289)^2}{2789.28}\right)}$

$= 18.498 \pm 0.693$
$= [17.807 \leq \mu \leq 19.191]$

c) 95% prediction interval on $X_0 = 25$

$\hat{y} \pm t_{\frac{\alpha}{2},n-2}\sqrt{s^2\left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_i (X_i - \bar{X})^2}\right)}$

(note: $\sum_i (X_i - \bar{X})^2 = S_{xx}$)

→ find $\hat{y}$ using regression model:
$\hat{y} = -27.643 + 1.846\,(25)$
$\hat{y} = 18.5$

$18.5 \pm 1.96\sqrt{30.756\left(1 + \frac{1}{250} + \frac{(25-25.289)^2}{2789.282}\right)}$

$[7.553, 29.443]$

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
- The prediction interval has to account for the uncertainty in knowing the value of the population mean, plus data scatter
- The prediction interval is always wider than a confidence interval

Lec32 - Adequacy of regression model, correlation, omit the test and CI for p

Yuxi Qin
yuxiqin.ca

Tuesday, November 29, 2022 4:18 AM

- Coefficient of determination
- Model assumptions

THE COEFFICIENT OF DETERMINATION

- The co-efficient 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

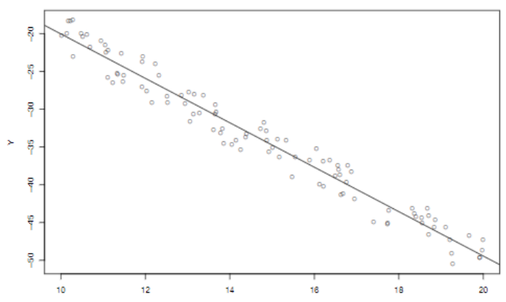
- When conducting inferences about β_1 and β_0, the following model assumptions must hold:
  - **Normality:** for each value of the explanatory variable (X) there is a subpopulation of response variables (Y) that is normally distributed
  - **Linearity:** the means of the subpopulation for each value of X that falls on the straight line $\beta_0 + \beta_1 X$
  - **Constant variance:** all subpopulations have the same standard deviation σ
  - **Independence**: All observations are independent
- Note: no assumptions are needed for the distribution of X

TESTING THE ASSUMPTIONS

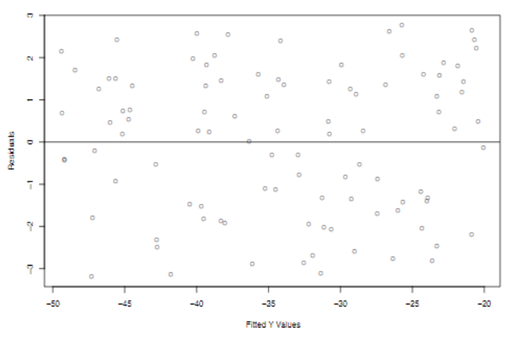
- Normality assumption
  - Construct a normal probability plot (QQ) plot of the residuals
  - The plot should have an approximately straight line if the residuals are normally distributed
- Linearity
  - Plot a scatter plot of Y vs X and superimpose the regression line
  - The estimated regression line 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

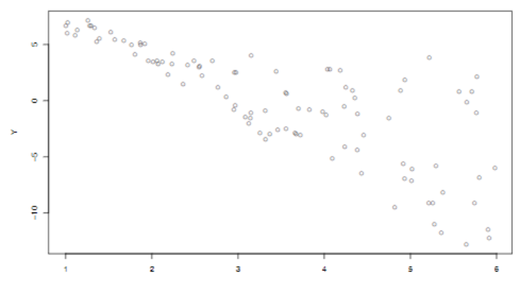


- (Bad fit) This residual plot indicates that the assumptions of linearity and constant variance do not appear to be violated

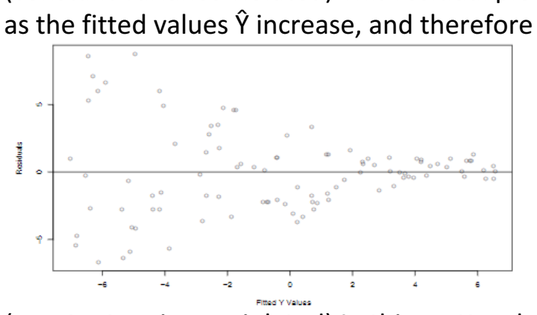


  - The residuals are randomly distributed about 0
  - Furthermore, the variance of the residuals is approx. const. for all fitted values

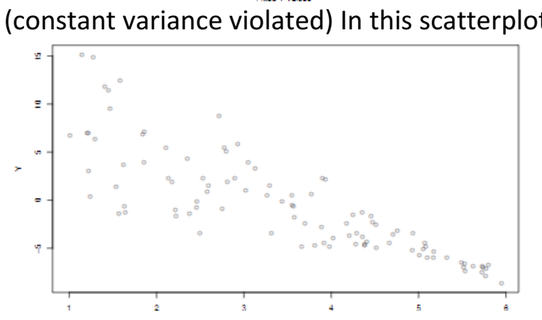
- (Constant variance violated)This scatter plot shows a decreasing trend, but shows that Var[Y] increases as X increases



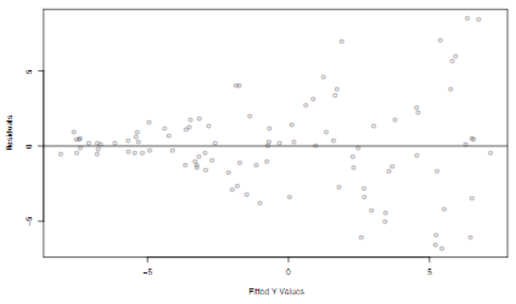
- (constant variance violated) This residual plot show that the residual variance, Var[Ê] decreases as the fitted values $\hat{Y}$ increase, and therefore do not satisfy the assumption of constant variance



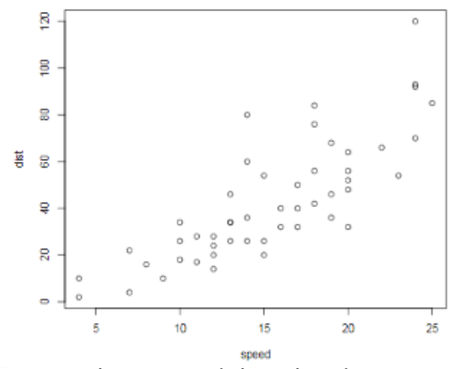
- (constant variance violated) In this scatterplot, it shows that Var[Y] decreases as X increases



- (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.



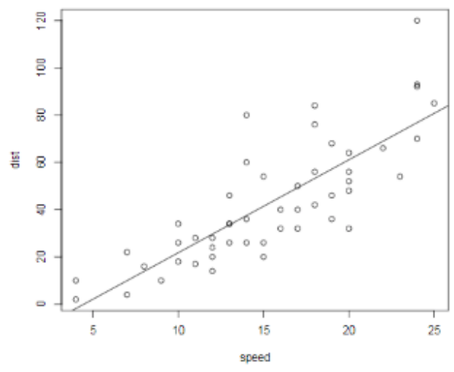
Fitting a linear model to this data using R gives this output:

```
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   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
   1. 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),
      β_0 is the theoretical intercept parameter,
      β_1 is the theoretical slope parameter,
      $\epsilon_i \backsim \mathcal{N}(0, \sigma^2)$ are the theoretical errors for observation i, and
      $Y_i \backsim \mathcal{N}(\beta_0, \beta_1 X_i, \sigma^2)$

   2. Derive the least squares estimates of the intercept and slope in this model.

      From the lecture notes, let
      $Q = \sum_i e_i^2 = \sum_i (Y_i - \hat{Y}_i)^2 = \sum_i (Y_i - b_0 - b_1 X_i)^2$.
      Take the partial derivatives of Q with respect to the model parameters to obtain:

      $$\frac{\partial Q}{\partial b_0} = -2 \sum_i (Y_i - b_0 - b_1 X_i)$$
      $$\frac{\partial Q}{\partial b_1} = -2 \sum_i X_i (Y_i - b_0 - b_1 X_i).$$

      Set each partial derivative equal to zero and solve for β_0 and β_1. This gives:

      $$b_0 = \bar{Y} - b_1 \bar{X}$$
      $$b_1 = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2}$$

   3. Carefully explain the interpretation of the model parameters in the context of this analysis.

      The slope: β_1
      1. For ever mile per hour that the speed of a car increase, the stopping distance increases by 3.9324 feet on average.

      The y-intercept: β_0
      1. The predicted stopping distance for a car travelling zero miles per hour is -17.5791 feet.

   4. Obtain a 95% confidence interval for the slope parameter and explain to an untrained person how this confidence interval should be interpreted.

      $b_1 \pm t_{\alpha/2, n-2} \widehat{SE}(b_1) = 3.9324 \pm t_{0.025,48} 0.4155 = 3.9324 \pm 2.011(0.4155)$

      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.

   5. 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

$\sum y_i = 572$ $\sum x_i = 43$ $\sum x_i y_i = 1697.80$
$\sum y_i^2 = 23530$ $\sum x_i^2 = 157.42$ $n = 14$

$R^2 = \frac{SS_R}{SS_T}$ ⟶ on eqn sheet: $SS_R = \hat{\beta}_1 S_{xy}$

Solving for SS_R:

$S_{xx} = \sum x_i^2 - n\bar{x}^2$
$= 157.42 - 14(3.0714)$
$S_{xx} = 25.348$

$\bar{x} = \frac{\sum x_i}{n}$
$\bar{x} = 3.0714$

$S_{xy} = \sum x_i y_i - n\bar{x}\bar{y}$
$= 1697.8 - 14(3.0714)(40.857)$
$S_{xy} = -59.04$

$\bar{y} = \frac{\sum y_i}{n}$
$\bar{y} = 40.857$

$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-59.04}{25.348}$

$\hat{\beta}_1 = -2.33$

Solving for SS_T:

$SS_T = \sum (Y_i - \bar{Y})^2 = \sum Y_i^2 - \frac{(\sum y_i)^2}{n} = S_{yy}$

$SS_T = 23530 - \frac{(572)^2}{14}$

$SS_T = 159.71$

$R^2 = \frac{SS_R}{SS_T} = \frac{\hat{\beta}_1 S_{xy}}{S_{yy}} = \frac{(-2.33)(-59.04)}{159.71}$

$R^2 = 0.861$

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

Yuxi Qin
yuxiqin.ca

Tuesday, November 29, 2022 5:04 PM

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,…, I laboratories or groups, producing data
- And suppose we have j = 1,…, 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 DECONOMPOSITION OF SUMS OF SQUARES

- For the ANOVA procedure, the total sum of squares is given by

$$\sum_{i,j}(Y_{ij}-\overline{Y})^2$$

- It's true that:

$$(Y_{ij}-\overline{Y})=(Y_{ij}-\overline{Y}_{i\cdot})+(\overline{Y}_{i\cdot}-\overline{Y}),$$

And that if we square both sides of the equation and sum over I (from i=1 to n), we get

$$\sum_{i,j}(Y_{ij}-\overline{Y})^2=\sum_{i,j}(Y_{ij}-\overline{Y}_{i\cdot})^2+\sum_{i,j}(\overline{Y}_{i\cdot}-\overline{Y})^2$$

- 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

$\sum_{i,j}(Y_{ij}-\overline{Y})^2$	Sum of squares total (SST)
$\sum_{i,j}(Y_{ij}-\overline{Y}_{i\cdot})^2$	Sum of squares error (SSE)
$\sum_{i,j}(\overline{Y}_{i\cdot}-\overline{Y})^2$	Sum of squares treatment (SSTrt)

SSTO = SSE + SSTrt

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:

$$\sum_{i,j}(Y_{ij}-\overline{Y})^2=\sum_{i,j}(Y_{ij}-\overline{Y}_{i\cdot})^2+\sum_{i,j}(\overline{Y}_{i\cdot}-\overline{Y})^2$$

SSTO	=	SSE	+	SSTrt
$IJ-1$	=	$I(J-1)$	+	$I-1$

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,…,I and j=1,…,J

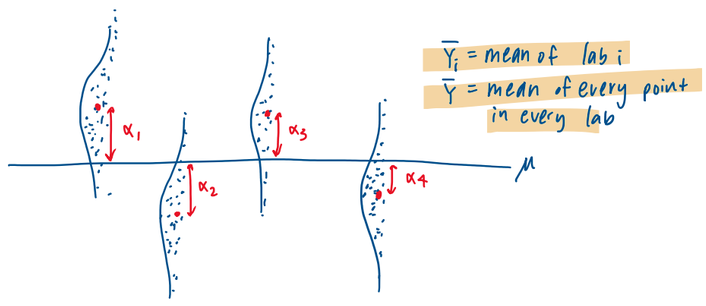
- μ and α_i are fixed, unknown constants with:

$$\sum_{i=1}^{I}\alpha_i=0.$$

- 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:

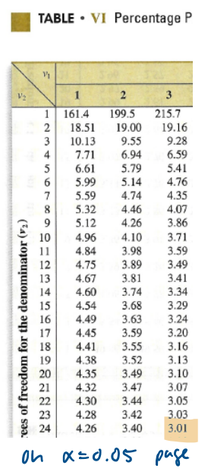
$$F=\frac{\text{MSTrt}}{\text{MSE}},$$

Where MS stands for **Mean Square**

- We compare the test statistic to a critical value with notation

$$F^{\alpha}_{I-1,I(J-1)}$$

- Rejecting H_0 means there is evidence that at least two means are significantly different from one another



on $\alpha=0.05$ page

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. What is $F_{3,\,24,\,\alpha=0.05}$? $F_{3,24,\alpha=0.05}=3.01$

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			

$F_{value}=4.6619$
↳ 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: $i=3$, j:
D.o.F = $i-1=3-1=2$
$=i(j-1)=3(10-1)=27$

$F=\frac{MS\,Tr}{MSE}=\frac{90.0335}{19.313}=4.6619$

Mean square = $\frac{\text{sum of squares}}{\text{D.o.F}}$
$\Rightarrow MS_{tr}=\frac{180.067}{2}=90.0335$
$\Rightarrow MS_E=\frac{522.46}{27}=19.313$
↳ 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 :
D.o.F : $i-1=3-1=2$
$i(j-1)=3(18-1)=51$
→ total error = $i+j=53$

$SS_{Trt}=582.619-46.234$
$SS_{Trt}=536.385$

MS (mean Square) = $\frac{\text{Sum of Square}}{\text{D.o.F}}$
$\Rightarrow MS_{Trt}=\frac{536.385}{2}=268.1925$
$\Rightarrow MS_E=\frac{46.234}{51}=0.906549$

$F\text{-value}=\frac{MS_{Tr}}{MS_E}=\frac{268.1925}{0.906549}$
F-value = 295.1925

P-value = $F_{2,51,\alpha=0.01}$
$\Rightarrow p<0.01$

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
C J
D.o.$F_{tr}=i-1=8-1=7$
D.o.$F_E=i(j-1)=8(6-1)=40$
→ each treatment contributes n-1 degrees of error

$MS_{q\,Tr}=\frac{SS_{Trt}}{\text{D.o.F}}=\frac{563.286}{7}=80.47$

$MS_{q\,E}=\frac{SS_E}{\text{D.o.F}}=13.452$

$SS_E=40(13.452)=538.08$

$F=\frac{MS_{Tr}}{MS_E}=\frac{80.47}{13.452}$
F-value = 5.98

P-value = $F_{7,40,\alpha=0.01}$
P-value < 0.01

Ex. An ANOVA table in R

The data set PlantGrowth is available in R. It contains weight measurements on 30 plants grown under 3 different experimental conditions (10 in each condition). Here is the output from the **aov** command in R:

```
            Df  Sum Sq Mean Sq F value  Pr(>F)
group        2  3.7663  1.8832  4.8461 0.01591 *
Residuals   27 10.4921  0.3886
```

- What is the conclusion of the ANOVA hypothesis test when α=0.05?
- A significance level of 0.05 indicates a 5% risk of concluding that a difference exists when there is no actual difference

Lec34 - Fisher's Least Significant Difference (LSD)

Yuxi Qin
yuxiqin.ca

Wednesday, December 7, 2022 1:39 PM

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

$$(\bar{x}_i - \bar{x}_j) \pm t_{n-I,\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_i}+\frac{1}{n_j}\right)},$$

Where s_p^2 = 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:

1. **NP**: Mice in this group ate as much as they pleased of a non-purified, standard diet for laboratory mice.
2. **N/N85**: This group was fed normally both before and after weaning.
3. **N/R50**: This group was fed a normal diet before weaning and a reduced-calorie diet of 50kcal/wk after weaning.
4. **R/R50**: This group was fed a reduced-calorie diet of 50kcal/wk both before and after weaning.
5. **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.
6. **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:

	Summary Stats			
Group	n	$\bar{x}$	s	95% CI
NP	49	27.4	6.1	(25.6, 29.2)
N/N85	57	32.7	5.1	(31.3, 34.1)
N/R50	71	42.3	7.8	(40.5, 44.1)
R/R50	56	42.9	6.7	(41.1, 44.7)
N/R50 lopro	56	39.7	7.0	(37.8, 41.6)
N/R40	60	45.1	6.7	(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 : at least one $\mu_i \neq \mu_k$, for $i, k = 1, \ldots, 6$

- To make a conclusion about H_0, we need to compute the test statistic

$$F = \frac{MSTrt}{MSE}$$

- 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				

$n_{total} = 349$, $\alpha = 0.05$

D.o.F.$_{Tr}$ = $i - 1 = 6 - 1 = 5$

D.o.F.$_E$ = $i(j-1) = 349 - 6 = 343$ → Total D.o.F. = 348

↳ each treatment contributes $n_T - 1$ degree of freedom

$MS_{Trt} = \frac{S\ of\ Sq}{D.o.F} = \frac{12733.42}{5} = 2546.8$

$MS_E = \frac{S\ of\ Sq}{D.o.F} = \frac{15297.42}{343} = 44.6$

$F = \frac{MS_{Trt}}{MS_E} = \frac{2546.8}{44.6} = 57.1$

@ $F_{5,343,\ 57.1}$ → p-value < 0.01

Critical value:

$F^{0.05}_{5,343} = F^{0.05}_{5,120} = 2.29$

Knowing p-value < 0.01 and $\alpha = 0.05$, we can reject H_0, conclude there is evidence of a difference.

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

$$(\bar{x}_i - \bar{x}_j) \pm t_{n\ \ I,\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_i}+\frac{1}{n_j}\right)},$$

Where s_p^2 = 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:

$$(\bar{x}_i - \bar{x}_j) \pm t_{n\ \ I,\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

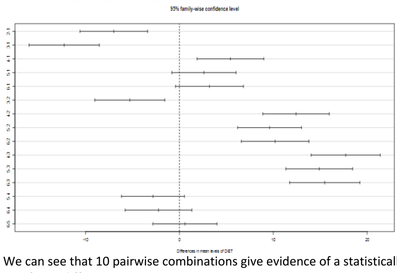
$$(32.7 - 42.3) \pm t_{343,0.025}\sqrt{44.6\left(\frac{1}{71}+\frac{1}{57}\right)}$$

$$-9.6 \pm t_{\infty,0.025}\,(1.187697)$$

$$-9.6 \pm 1.96\,(1.187697)$$

$$-9.6 \pm 2.327886$$

- 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 WOOOOOOOOOOOOO