

MINE AUTOMATION AND DATA ANALYTICS



SWAYAM NPTEL COURSE ON

MINE AUTOMATION AND DATA ANALYTICS

By

Prof. Radhakanta Koner

Department of Mining Engineering

Indian Institute of Technology (Indian School of Mines) Dhanbad



CONCEPTS COVERED

1. Compute and Interpret numerical Summaries of Data

- Compute and Interpret measures of dispersion: Variance, Standard Deviation
- Compute and Interpret percentiles and Interquartile Range (IQR)
- Compute and Interpret five number summary

2. Association between two variables

- Understanding the association between numerical variables through a scatter plot
- Compute and interpret Covariance and Correlation.



Variance

- In contrast to the range, the variance considers all the observations.
- One way of measuring the variability of a data set is to consider the deviations of the data values from a central value.



Population Variance and Sample Variance

- When we refer to a dataset from a population, we assume the dataset has N observations. In contrast, when referring to a data set from a sample, we assume the data set has n observations.
- The variation is computed using the following formulae

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

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



Units of Variance

- The sample variance is expressed in units of square units of the original variable.



JAN 2024

Example

	Data	Deviation from Mean $(x_i - \bar{x})$	Squared Deviation $(x_i - \bar{x})^2$
1	9	3.2	10.24
2	5	-0.8	0.64
3	8	2.2	4.84
4	3	-2.8	7.84
5	4	-1.8	3.24
Total	29	0	26.8

$$\text{Sample Variance} = \frac{26.8}{4} = 6.7$$

$$\text{Population Variance} = \frac{26.8}{5} = 5.36$$



Adding/Multiplying a constant

- Let $y_i = x_i + c$ where c is a constant then ***new variance = old variance***
- ***In general, adding a constant does not change dataset variability. Hence, it is the same.***
- Let $y_i = x_i * c$ where c is a constant then ***new variance = $c^2 * \text{old variance}$***



Standard Deviation

Another handy measure of dispersion is the standard deviation.

Definition

The quantity, the square root of the sample variance, is the sample **standard deviation**.

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



Example

	Data	Deviation from Mean $(x_i - \bar{x})$	Squared Deviation $(x_i - \bar{x})^2$
1	9	3.2	10.24
2	5	-0.8	0.64
3	8	2.2	4.84
4	3	-2.8	7.84
5	4	-1.8	3.24
Total	29	0	26.8

$$\text{Sample Variance} = \frac{26.8}{4} = 6.7$$

$$\text{Population Variance} = \frac{26.8}{5} = 5.36$$

$$\text{Sample Standard Deviation} = \sqrt{6.7} = 2.58$$

$$\text{Population Standard Deviation} = \sqrt{5.36} = 2.31$$



Units of Standard Deviation

- The sample standard deviation is measured in the same units as the original data.



Adding/Multiplying a constant

- Let $y_i = x_i + c$ where c is a constant then *new standard deviation = old standard deviation*
- *In general, adding a constant does not change dataset variability. Hence, it is the same.*
- Let $y_i = x_i * c$ where c is a constant then *new standard deviation = $c * standard deviation$*



Quartiles

Definition

The sample 25th percentile is called the first quartile. The sample 50th percentile is called the median or the second quartile. The sample 75th percentile is called the third quartile.



In other words, the quartiles break a data set into four parts, with about 25 percent of the data values being less than the first (lower) quartiles, about 25 percent being between the first and second quartiles, about 25 percent being between the second and third (upper) quartiles, and about 25 percent being larger than the third quartile



Interquartile Range (IQR)

Definition

The interquartile range (IQR) is the difference between the first and third quartiles.

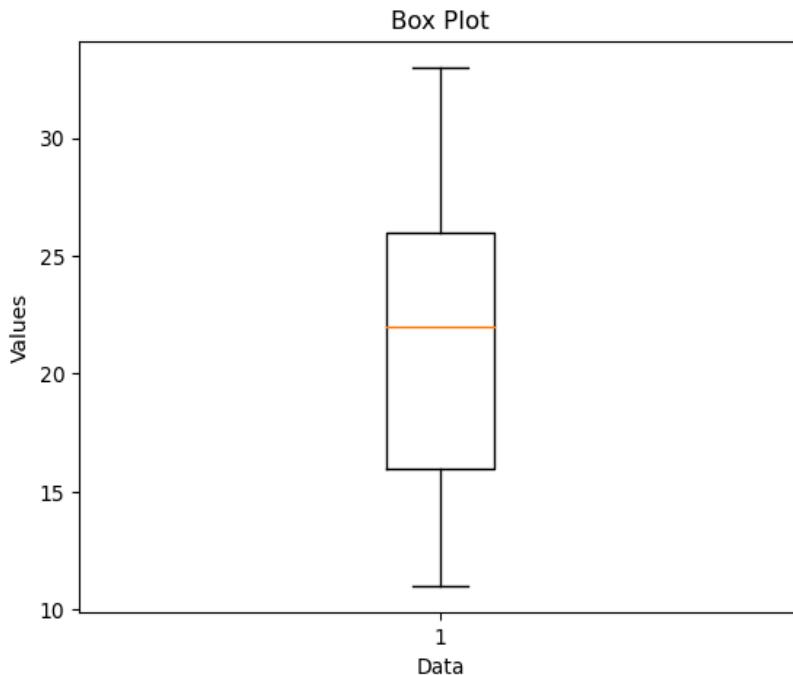
$$IQR = Q_3 - Q_1$$



The Five Number Summary

- i) Minimum
- ii) Q_1 : First Quartile or Lower Quartile
- iii) Q_2 : Second Quartile or Median
- iv) Q_3 : Third Quartile or upper quartile
- v) Maximum

Data = [11, 22, 15, 29, 33, 15, 17, 22, 19, 25, 27]



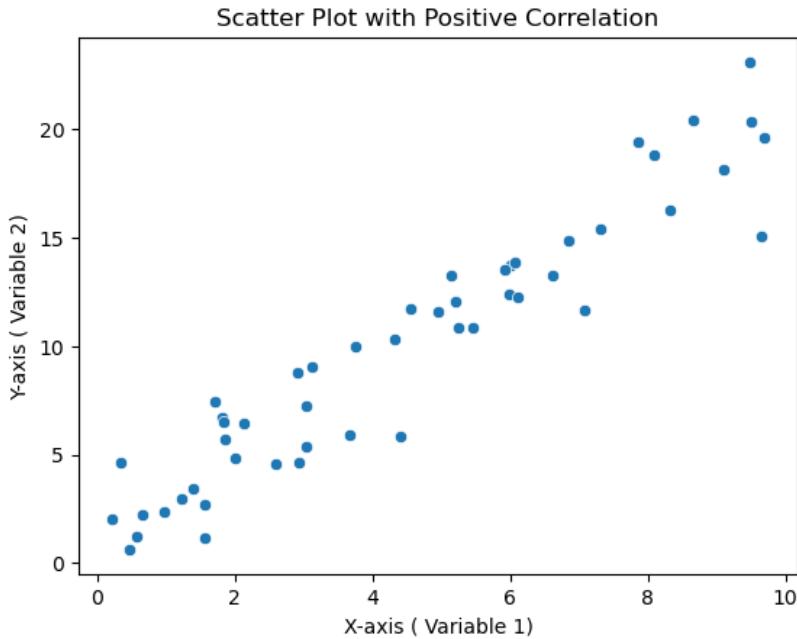
Association

The pattern of data in one variable occurs in a particular manner related to the pattern of data in one or several other variables.



Scatter Plot

- A scatter plot is a graphical representation of the relationship between two numerical variables.
- It allows you to visually inspect the pattern of the data points and understand the association or correlation between the variables.

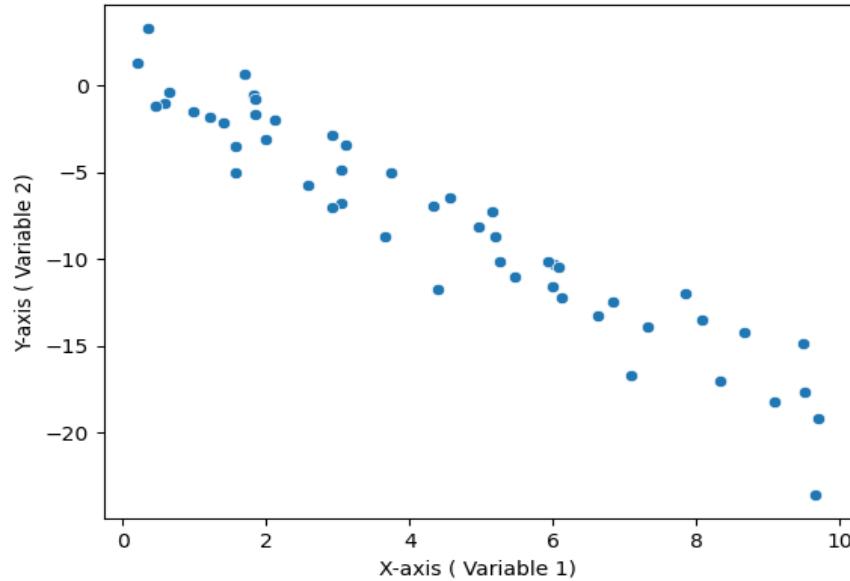


Positive Relationship:

If the points on the scatter plot generally form an upward-sloping pattern from left to right, it indicates a positive correlation. This means that as one variable increases, the other also tends to increase.

Scatter Plot

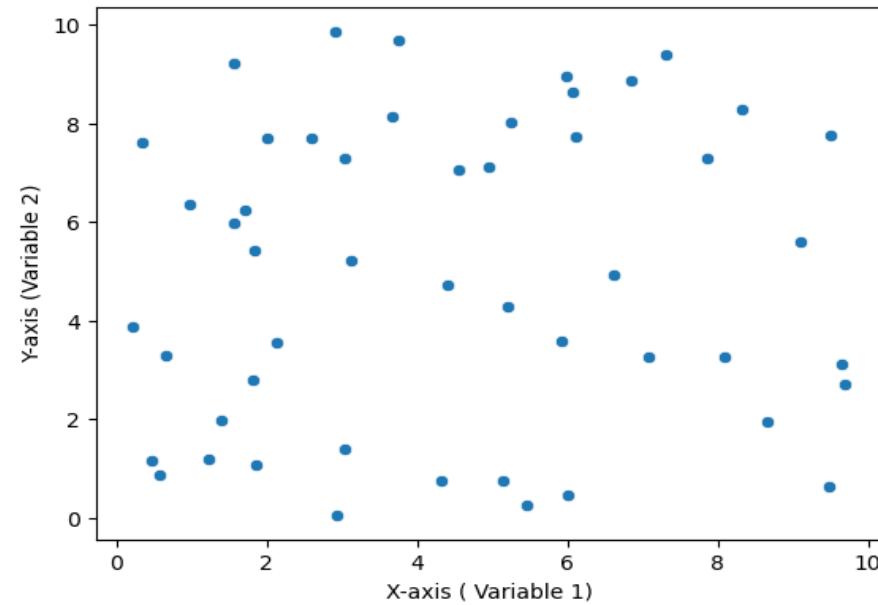
Scatter Plot with Negative Correlation



Negative Relationship:

Conversely, if the points on the scatter plot form a downward-sloping pattern from left to right, it indicates a negative correlation. This means that as one variable increases, the other tends to decrease.

Scatter Plot with No Correlation



No Relationship:

If the points appear randomly scattered without any clear pattern, it suggests no solid linear relationship between the variables. However, other types of relationships might still exist, such as non-linear or complex associations.



Measures of association

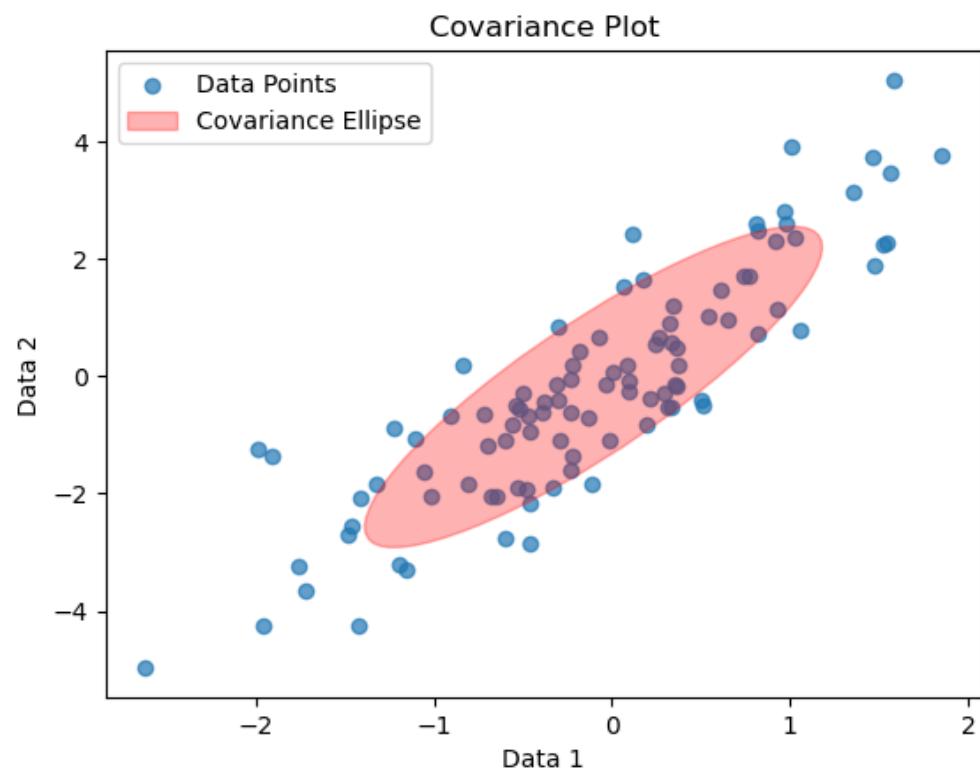
How do we measure the strength of association between two variables i.e. from earlier plots ?

1. Covariance
2. Correlation



Covariance

- Covariance quantifies the strength of the linear association between two numerical variables.



Covariance

- **Definition**
- **Let x_i denote the i^{th} observation of variable x and y_i is the i^{th} observation of variable y . Let (x_i, y_i) be the i^{th} paired observation of a population (sample) dataset having $N(n)$ observations. The covariance between the variables x and y is given by**

$$\text{Population Variance : } \text{Cov}(x, y) = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{N}$$

$$\text{Sample Variance : } \text{Cov}(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$



Example

- Recall , the association between variable x and variable y of a person

x	y	Deviation of x $(x_i - \bar{x})$	Deviation of y $(y_i - \bar{y})$
2	5	2 – 17.67	5 – 25.5
8	12	8 – 17.67	8 – 25.5
18	18	18 – 17.67	18 – 25.5
20	23	20 – 17.67	20 – 25.5
28	45	28 – 17.67	28 – 25.5
30	50	30 – 17.67	30 – 25.5

Example

x	y	Deviation of x $(x_i - \bar{x})$	Deviation of y $(y_i - \bar{y})$
2	5	2 – 17.67	5 – 25.5
8	12	8 – 17.67	12 – 25.5
18	18	18 – 17.67	18 – 25.5
20	23	20 – 17.67	23 – 25.5
28	45	28 – 17.67	45 – 25.5
30	50	30 – 17.67	50 – 25.5

Number of observations = 6
Mean of X = 17.67
Mean of Y = 25.5

Cov (X, Y) =

$$\begin{aligned} & \left(\frac{1}{6} \right) [(2 - 17.67)(5 - 25.5) + (8 - 17.67)(12 - 25.5) + (18 - 17.67)(18 - 25.5) + (20 - 17.67)(23 - 25.5) + (28 - 17.67)(45 - 25.5) + (30 - 17.67)(50 - 25.5)] \\ & = 157.83 \end{aligned}$$



Units of Covariance

- The size of the covariance, however, is difficult to interpret because the covariance has units.
- The units of the covariance are those of the x variable times of the y variable.



Correlation

- A more easily interpreted measure of linear association between two numerical variables in correlation.
- It is derived from covariance
- To find the correlation between two numerical variables, x and y, divide the covariance between x and y by the product of the standard deviations of x and y.
- The Pearson correlation coefficient (r) between x and y is given by

$$r = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{Cov(x, y)}{s_x s_y}$$



Units of Correlation

- The units of the standard deviation cancel out the units of covariance
- It can be shown that the correlation measure always lies between -1 and +1



Example

- Recall , the association between x variable and y variable

x	y	Deviation of x	Deviation of y
20	60	20 – 37.5	60 – 67.5
25	60	25 – 37.5	60 – 67.5
30	70	30 – 37.5	70 – 67.5
40	73	40 – 37.5	73 – 67.5
50	67	50 – 37.5	67 – 67.5
60	75	60 – 37.5	75 – 67.5

Example

- Recall , the association between x variable and y variable

x	y	Deviation of x $(x_i - \bar{x})$	$(x_i - \bar{x})^2$	Deviation of y $(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
20	60	-17.5	306.25	-7.5	56.25	131.25
25	60	-12.5	156.25	-7.5	56.25	93.5
30	70	-7.5	56.25	2.5	6.25	-18.75
40	73	-2.5	6.25	5.5	30.25	13.75
50	67	12.5	156.25	-0.5	0.25	-6.25
60	75	22.5	506.25	7.5	56.25	168.75
			= 1187.5		= 205.5	382.5



$$Cov(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

$$Cov(x, y) = \frac{382.5}{5} = 76.5$$

$$\text{Variance of } x = \frac{1187.5}{5} = 237.5, S_x = \sqrt{237.5} = 15.41$$

$$\text{Variance of } y = \frac{205.5}{5} = 41.1, S_y = \sqrt{41.1} = 6.41$$

$$r = \frac{Cov(x, y)}{S_x S_y} = \frac{76.5}{15.41 * 6.41} = 0.774$$



REFERENCES

- Introduction to Probability and Statistics for Engineers and Scientists, Sixth Edition, Sheldon M. Ross
- Statistical Methods Combined Edition (Volume I& II), N G Das



CONCLUSION

1. Numerical Summaries

- ❖ Variance
- ❖ Standard Deviation
- ❖ Interquartile Range (IQR)
- ❖ Interpret five number summary using Box Plot

2. Association between two variables

- ❖ Scatter plot
- ❖ Covariance
- ❖ Correlation



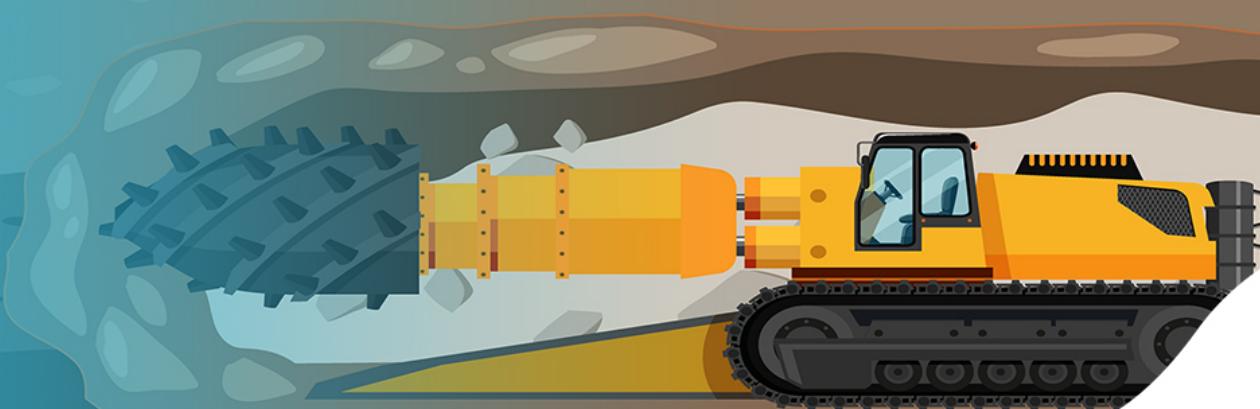


THANK YOU



JAN 2024

MINE AUTOMATION AND DATA ANALYTICS



MINE AUTOMATION AND DATA ANALYTICS





SWAYAM NPTEL COURSE ON

MINE AUTOMATION AND DATA ANALYTICS

By

Prof. Radhakanta Koner

Department of Mining Engineering
Indian Institute of Technology (ISM) Dhanbad



Module 7: Probability

Lecture 18A : Introduction to
Probability and its associated terms

CONCEPTS COVERED

- Basic Terms of Probability
- Types of Probability
- Conditional Probability
- Marginal Probability
- Joint Probability



JAN 2024

Probability

Definition:

- Probability is a measure of the likelihood that a given event will occur.
- To do a certain task, we have certain chances of success. These chances of success increase by increasing the number of attempts

Importance:

Used in various fields, such as statistics, finance, science, and more.

Real-world examples:

Coin toss, weather forecasting.



Probability

Random Experiment:

- We know possible outcomes but do not know the exact outcomes.

Example:

- i. Toss a coin: Possible Outcome: {H ,T}, **Exact outcome ??**
- ii. Rolling a dice: Possible Outcome: {1,2,3,4,5,6}, **Exact outcome ??**
- iii. Box containing R,G,B balls. The possible outcome is {R,G,B}. A ball is picked @ random, **Exact outcome ??**



Probability

Sample Space:

Set of all possible outcomes of a random experiment

Toss a coin: Possible Outcome: {H ,T}

Rolling a dice: Possible Outcome: {1,2,3,4,5,6},

Box containing Red, Green, and Blue balls. A ball is picked @ random. The possible outcome is {R, G, B}.



Probability

Event:

- Subset of sample space which is our area of interest

Toss a coin: Possible Outcome: {H,T}

$$E_1 = \{H\}$$

$$E_2 = \{T\}$$

E1 and E2 are sample spaces

Since **Event and Sample Space** both are sets, so all set operations can be applied to them like Union, Intersection, Set difference, etc.,

Sample space = {1,2,3,4,5,6}

$$E1 = \{1,2\}$$

$$E2 = \{2,3,4\}$$

$$E1 \cap E2 = \{2\}$$

$$E1 \cup E2 = \{1,2,3,4\}$$

$$E1 - E2 = \{1\}$$

$$E1 \cap E2 = \{2\}$$

$$E1^C = \{3,4,5,6\}$$



Probability

Mutually Exclusive Events

- Two events E_1 and E_2 are said to be mutually exclusive iff $E_1 \cap E_2 = \{\}$ or \emptyset

$$E_1 = \{H\}$$

$$E_2 = \{T\}$$

E_1 and E_2 are Mutually Exclusive Events because $E_1 \cap E_2 = \{\}$



Basic Probability Terms

Sample Space:

- **Definition:** The set of all possible outcomes of an experiment.
- **Example:** Coin toss (Heads, Tails).

Event:

- **Definition:** A subset of the sample space.
- **Example:** Getting a Head in a coin toss.

Probability of an Event:

- **Definition:** The likelihood of an event occurring.
- **Formula:** $P(\text{Event}) = \text{Number of favorable outcomes} / \text{Total outcomes}$.



Types of Probability

Classical Probability:

1. Definition: Based on equally likely outcomes.
2. Formula: $P(E) = \text{Number of favorable outcomes} / \text{Total number of outcomes}$.
3. Example: Rolling a fair six-sided die.

Empirical Probability:

1. Definition: Based on observed outcomes.
2. Formula: $P(E) = \text{Number of times event E occurs} / \text{Total number of trials}$.
3. Example: Probability of rain based on historical data.

Subjective Probability:

1. Definition: Based on personal judgment or opinion.
2. Example: Probability of winning a game based on a person's intuition.



Probability

Probability:

- Probability is a measure of the likelihood that a given event will occur
- Probability = **favorable outcomes / total number of outcomes** = E / SS
- Ex: Roll a dice , then probability of getting even number?
- $E = \{ 2, 4 ,6\}$ and $SS = \{1,2 3,4,5,6 \}$
- Probability = $E / SS = 3/6 = 0.5$
- Ex: Roll a die; probability of getting a number 5?
- $E = \{ 5 \}$ and $SS = \{1,2 3,4,5,6 \}$
- Probability = $E / SS = 1/6$
- Ex: Toss 2 coins. Probability of getting 2 heads or 2 tails?
- $E = \{ TT,HH \}$ and $SS = \{TT,HH,TH,HT\}$
- Probability = $E / SS = 2/4 = 0.5$



Probability Example

- Suppose you roll a fair six-sided die. What is the probability of rolling a 3?
- Sample Space (S): {1, 2, 3, 4, 5, 6}
- Event (E): Rolling a 3
- Probability (P(E)): $1/6$ (because there is 1 favorable outcome out of 6 possible outcomes)



Probability

- Probability of a Sample Space?
- $E = SS$
- $P = E/SS = SS/SS = 1$
- $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$
- $P(E_1 \cup E_2 \cup E_3) = P(E_1) + P(E_2) + P(E_3) - P(E_1 \cap E_2) - P(E_2 \cap E_3) - P(E_3 \cap E_1) + P(E_1 \cap E_2 \cap E_3)$
- **E₁ and E₂ are Mutually Exclusive Events, $P(E_1 \cup E_2) = ?$ $P(E_1 \cup E_2 \cup E_3) = ?$**
- $P(E_1 \cup E_2) = P(E_1) + P(E_2)$ Since $P(E_1 \cap E_2) = 0$
- $P(E_1 \cup E_2 \cup E_3) = P(E_1) + P(E_2) + P(E_3)$



Probability

Roll a dice $SS = \{1, 2, 3, \dots, 6\}$, $E_1 = \{1, 2, 3\}$, $E_2 = \{4, 5\}$, $P(E_1 \cup E_2) = ?$

$P(E_1 \cup E_2) = P(E_1) + P(E_2)$ Since $P(E_1 \cap E_2) = 0$

$$P(E_1 \cup E_2) = 3/6 + 2/6 = 5/6$$

Toss 2 coins $E_1 = \{HH\}$, $E_2 = \{TT\}$, $P(E_1 \cup E_2) = ?$

$P(E_1 \cup E_2) = P(E_1) + P(E_2)$ Since $P(E_1 \cap E_2) = 0$

$$P(E_1 \cup E_2) = 1/4 + 1/4 = 1/2$$

$$P(E^C) = ?$$

$$E \cup E^C = SS$$

$$P(E \cup E^C) = P(SS)$$

$$P(E) + P(E^C) = 1$$

$$P(E^C) = 1 - P(E)$$

Q: Probability of getting good rank in GATE is 0.9, Probability of not getting a good rank is ??

$$P(E^C) = 1 - P(E)$$

$$P(E^C) = 1 - 0.9 = 0.1$$



Conditional Probability

- 2 Events A and B are there.
 - $P(A|B)$ = Probability of A given B = What is the probability of happening A given B
 - $P(B|A)$ = probability of B given A = What is the probability of happening B given A
-
- A = getting a prime number from a dice = {2,3,5}
 - B = getting a even number from a dice = {2,4,6}
 - $P(A|B) = ?$
 - B already happened = {2,4,6} , Now sample space would be B = {2,4,6}
 - Probability of happening A given B already happened = 1/3
 - $P(A|B) = P(A \cap B) / P(B)$
 - $P(B|A) = P(B \cap A) / P(A)$



Conditional Probability

- **Definition:** Probability of an event given that another event has occurred.
- **Formula:** $P(A|B) = P(A \text{ and } B) / P(B)$.
- **Example:** Probability of getting a Head in the second toss given that the first toss resulted in a Tail.



Conditional Probability

Steps to solve the questions of Conditional Probability $P(A/B)$

- Find Sample Space
- Finding Which is Event A
- Finding Which is Event B
- Finding Probabilities $P(A)$, $P(B)$, $P(A \cap B)$
- Calculating required one $P(A/B) = P(A \cap B) / P(B)$ or $P(B/A) = P(B \cap A) / P(A)$



Example of Conditional Probability

- Q) If 2 dice are rolled, and the 1st dice shows 4, then what is the probability that the sum is 6?
- A) $|SS| = 36$, 1st dice = {1,2,3,4,5,6} , 2nd dice = {1,2,3,4,5,6}
- A = dice showed 4 = { (4,1),(4,2),(4,3),(4,4),(4,5),(4,6) }
- B = sum is 6 = {(1,5),(2,4),(3,3),(4,2),(5,1)} , $|B| = 5$
- $A \cap B = \{(4,2)\}$
- $P(A) = 6/36 = 1/6$; $P(B) = 5/36$; $P(A \cap B) = 1/36$
- $P(A/B) = P(A \cap B) / P(B) = (1/36) / (5/36) = 1/5$
- $P(B/A) = P(B \cap A) / P(A) = (1/36) / (6/36) = 1/6$

What is the correct solution to this question = ?

It is $P(B/A) = 1/6$

What if the question asked: the probability that dice is showing 4 given the dices sum is 6?

$$= P(A/B)$$

$$= 1/5$$



Example of Conditional Probability

Q) Two dice are rolled, and both show odd numbers. Then the probability of getting the sum 6?

$$|SS| = 36$$

$$A = \{(1,1), (3,3), (5,5), (1,3), (1,5), (3,1), (3,5), (5,1), (5,3)\} ; |A| = 9$$

$$B = \{(1,5), (2,4), (3,3), (4,2), (5,1)\} ; |B| = 5$$

$$A \cap B = \{(1,5), (3,3), (5,1)\} ; |A \cap B| = 3$$

$$P(A) = |A| / |SS| = 9/36 ; P(B) = |B| / |SS| = 5/36 ; P(A \cap B) = |A \cap B| / |SS| = 3 / 36$$

- $P(A/B) = P(A \cap B) / P(B) = (3/36) / (5/36) = 9/5$
- $P(B/A) = P(B \cap A) / P(A) = (3/36) / (9/36) = 3/9 = 1/3$

- $P(B/A)$ is our solution = 1/3



Properties of Conditional Probability

1) $A \subseteq B$

$$P(A|B) = P(A \cap B) / P(B) = P(A) / P(B)$$

$$P(B|A) = P(B \cap A) / P(A) = 1$$

2) If A and B are Mutually Exclusive Events

$$P(A|B) = P(A \cap B) / P(B) = 0$$

$$P(B|A) = P(B \cap A) / P(A) = 0$$

3) $P(A \cup B | C) = P(A | C) + P(B | C) - P(A \cap B | C)$

4) $P(A^C | B) = 1 - P(A | B)$



Marginal Probability

- To understand marginal probability, let's consider a scenario with two random variables, A and B.
- The joint probability $P(A \cap B)$ represents the probability that events A and B occur together. The marginal probability of A, denoted as $P(A)$, focuses solely on the probability of event A happening, regardless of the occurrence or non-occurrence of event B.
- Mathematically, marginal probability is obtained by summing (or integrating, In the case of continuous random variables) the joint probabilities over all possible values of the other variable.

The formulas are as follows,

$$P(A) = \sum_{\text{all possible } B} P(A \cap B)$$

$$P(A) = \int_{\text{all possible } B} P(A \cap B) dB$$

- In simpler terms, you "marginalize" or "sum out" the unwanted variable (in this case, variable B) to obtain the probability distribution for the variable of interest (variable A).



Marginal probability

- Marginal probability refers to the probability of a single event or outcome occurring without considering the occurrence of other events.
- It is derived from a joint probability distribution, which describes the probabilities of combinations of events.
- Marginal probability is a fundamental concept in probability theory and is used in various statistical analyses and machine learning algorithms, especially when dealing with multiple variables and their interactions.



Joint Probability

- Joint probability is a concept in probability theory that describes the likelihood of two or more events occurring simultaneously.
- It is denoted as $P(A \cap B)$, where A and B are events. Joint probability is used to quantify the probability of the intersection of events.
- Mathematically, the joint probability of events A and B is calculated as follows: $P(A \cap B)$
- For independent events, where the occurrence of one event does not affect the occurrence of the other, the joint probability simplifies to the product of the individual probabilities:

$$P(A \cap B) = P(A) \cdot P(B)$$

Joint Probability

- However, when events are dependent, meaning the occurrence of one event affects the occurrence of the other, the joint probability is calculated using the conditional probability formula:

$$P(A \cap B) = P(A | B) \cdot P(B)$$

- Or, equivalently:

$$P(A \cap B) = P(B | A) \cdot P(A)$$

The terms involved:

$P(A \cap B)$: Joint probability of events $P(A)$ and $P(B)$.

$P(A)$ and $P(B)$: Marginal probabilities of events $P(A)$ and $P(B)$, respectively.

$P(A|B)$: Conditional probability of event $P(A)$ given that event $P(B)$ has occurred.

$P(B|A)$: Conditional probability of event $P(B)$ given that event $P(A)$ has occurred



Solved Example

Suppose you have two fair coins, one is red (R) and the other is blue (B). We want to calculate the conditional probability, joint probability, and marginal probability for the outcomes.

Conditional Probability:

Let's say we're interested in the probability of getting a red coin (R) given that we flipped two heads (HH).

$P(R|HH) = (\text{Number of outcomes where both coins are heads and one is red}) / (\text{Total number of outcomes where both coins are heads})$

Number of outcomes where both coins are heads and one is red = 1 (RR)

Total number of outcomes where both coins are heads = 1 (HH)

- So, $P(R|HH) = 1/1 = 1$



Solved Example - 1

Suppose you have two fair coins, one is red (R) and the other is blue (B). We want to calculate the conditional probability, joint probability, and marginal probability for the outcomes.

Joint Probability:

The joint probability is the probability of two events happening together.

Let A be the event of getting a red coin (R) and B be the event of getting two heads (HH).

$$P(A \cap B) = P(R \text{ and } HH)$$

Number of outcomes where both coins are heads, and one is red = 1 (RR)

$$\text{So, } P(A \cap B) = 1/4$$



Solved Example - 1

Suppose you have two fair coins, one is red (R) and the other is blue (B). We want to calculate the conditional probability, joint probability, and marginal probability for the outcomes.

Marginal Probability:

The marginal probability is the probability of a single event occurring without reference to any other event.

Let's calculate the marginal probability of getting a red coin (R).

$$P(R) = (\text{Number of outcomes where one coin is red}) / (\text{Total number of outcomes})$$

Number of outcomes where one coin is red = 2 (RR, RB)

Total number of outcomes = 4 (RR, RB, BR, BB)

$$\text{So, } P(R) = 2/4 = 1/2$$



Solved Example - 1

Suppose you have two fair coins, one is red (R) and the other is blue (B). We want to calculate the conditional probability, joint probability, and marginal probability for the outcomes.

In summary:

Conditional Probability: $P(R|HH) = 1$

Joint Probability: $P(R \text{ and } HH) = 1/4$

Marginal Probability: $P(R) = 1/2$



REFERENCES

- Introduction to Probability and Statistics for Engineers and Scientists, Sixth Edition, Sheldon M. Ross
- Statistical Methods Combined Edition (Volume I& II), N G Das



CONCLUSION

1. Basic Terms of Probability like Random Experiment, Sample Space, Event, and Mutually Exclusive events are being discussed
2. Types of Probability.
3. Discussed Conditional Probability with examples.
4. Discussed Marginal Probability.
5. Discussed Joint Probability.
6. Solved example involving Conditional , Marginal , and Joint Probability.





THANK YOU



JAN 2024

MINE AUTOMATION AND DATA ANALYTICS



MINE AUTOMATION AND DATA ANALYTICS





SWAYAM NPTEL COURSE ON

MINE AUTOMATION AND DATA ANALYTICS

By

Prof. Radhakanta Koner

Department of Mining Engineering
Indian Institute of Technology (ISM) Dhanbad

Module 8: Probability



Lecture 18B: Introduction to
Probability and its associated terms

CONCEPTS COVERED

- Independent Events
- Multiplication Theorem
- Total Probability
- Bayes Theorem
- Applications of Probability in Mining Industry



Independent Events

Two events are independent if **knowledge of the happening of one event does not affect the happening of the other event.**

Let A & B are independent events:

$$\begin{aligned} P(A/B) &= P(A) \\ P(B/A) &= P(B) \end{aligned}$$

$$P(A/B) = P(A \cap B) / P(B)$$

$$E_1 \cap E_2 \quad P(B/A) = P(B \cap A) / P(A)$$

$$P(B \cap A) = P(B).P(A)$$

$$P(A/B) = P(A \cap B) / P(B)$$

$$P(A \cap B) = P(A / B) . P(B)$$

$$P(B \cap A) = P(B / A) . P(A)$$

multiplication theorem for independent events

multiplication theorem for dependent events



Independent Events

Definition

Independent events are events where the occurrence of one event does not affect the occurrence of another.

- Mathematically, two events, A and B, are independent if and only if

$$P(A \cap B) = P(A) \cdot P(B)$$



Solved Example - 1 of Independent Events

Q) One coin is tossed, and one dice is rolled; E1: appearing head on coin, E2: appearing 3 on dice; Check whether the E1 and E2 are independent events.

$$P(E1 \cap E2) = P(E1).P(E2)$$

$$P(E1 \cap E2) = (1/2).(1/6) = 1/12$$

1 coin tossed AND 1 dice rolled:

$\{(H,1),(H,2),(H,3),(H,4),(H,5),(H,6),(T,1),(T,2),(T,3),(T,4),(T,5),(T,6)\}$

E1: appearing head on coin ; E2: appearing 3 on dice

$$P(E1 \cap E2) = 1/12$$

$$P(E1 \cap E2) = P(E1).P(E2)$$

E1 and E2 are two independent random events



Solved Example - 2 of Independent Events

Person A: The probability of hitting a target = $P(A) = 3/4$

Person B: The probability of hitting a target = $P(B) = 4/5$

The probability of both A and B hitting a target $P(E1 \cap E2) = ?$

A) since both events are independent events

$$P(A \cap B) = P(B).P(A)$$

$$P(A \cap B) = 3/4 \cdot 4/5 = 12/20 = 3/5$$



Multiplication Theorem

$$P(A \cap B \cap C) = P(A) \cdot P(B/A) \cdot P(C / A \cap B)$$

Where in independent events

$$P(A \cap B \cap C) = P(A) \cdot P(B) \cdot P(C)$$

So $P(A \cap B \cap C \cap D \cap E \dots \cap K) = P(A) \cdot P(B) \cdot P(C) \dots P(K)$

Bag A contain 3R & 2G balls, Bag B contain 3R & 5G balls , Bag C contain 1R & 4G balls

If one ball is drawn from each bag, then find the probability of getting Red from Bag A, Green from Bag B, and Red from Bag C. $P(E1 \cap E2 \cap E3) = ?$

A) $P(E1 \cap E2 \cap E3) = P(E1) \cdot P(E2) \cdot P(E3)$

$$\begin{aligned}P(E1 \cap E2 \cap E3) &= 3/5 \cdot 5/8 \cdot 1/5 \\&= 15/200 \\&= 3/40\end{aligned}$$



Total Probability

Let $A_1, A_2, A_3 \dots$ be events that form a partition of the sample space s . Let B be any event, then.

$$P(B) = P(B \cap A_1) + P(B \cap A_1) + P(B \cap A_1) + \dots$$

$$P(B) = P(A_1) \cdot P(B / A_1) + P(A_2) \cdot P(B / A_2) + P(A_3) \cdot P(B / A_3) + \dots$$

Total Probability

Q) Total Probability theorem is applicable for both dependent and independent events?

A) Yes

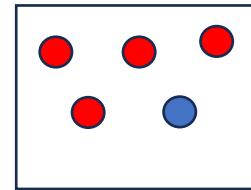
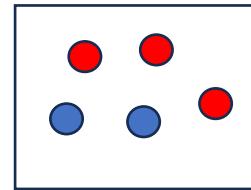
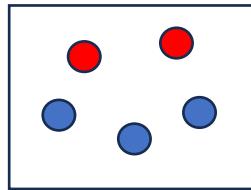
Q) The number of subsets of Sample Space that form a partition can be finite or infinite.

A) Yes



Example 1 of Total Probability

There are three boxes. A ball is picked @ random from a box. What is the probability that it is red? Assuming there is an equal likelihood of selecting a box.



$$P(B) = P(A_1) \cdot P(B / A_1) + P(A_2) \cdot P(B / A_2) + P(A_3) \cdot P(B / A_3) + \dots$$

$$P(\text{red}) = P(\text{Box A1}) \cdot P(\text{red} / \text{Box A1}) + P(\text{Box A2}) \cdot P(\text{red} / \text{Box A2}) + P(\text{Box A3}) \cdot P(\text{red} / \text{Box A3})$$

$$P(\text{red}) = (1/3)(2/5) + (1/3)(3/5) + (1/3)(4/5) = 9/15$$



Example 2 of Total Probability

A box containing 5 fair coins and 3 unfair coins ($P(H) = 1/3$; $P(T) = 2/3$). A coin is picked and tossed. Find the probability of getting a head from a picked coin.

$$\begin{aligned}P(H) &= P(F)(PH/F) + P(UF)(H/UF) \\P(H) &= (5/8)(1/2) + (3/8)(1/3)\end{aligned}$$



Example 3 of Total Probability



JAN 2024

13

Bayes Theorem

$$P(A|B) = P(A \cap B) / P(B) = P(A).P(B / A) / P(B)$$

$$P(A|B) = P(A).P(B / A) / P(B)$$



Example of Bayes Theorem

Suppose you are reaching home in three ways: 1) Bus, 2) Car, 3)Scooty

$$P(\text{late/Bus}) = 0.5, P(\text{Late/Car}) = 0.1, P(\text{Late/Scooty}) = 0.2$$

$$P(\text{Bus}) = 0.2, P(\text{Car}) = 0.7, P(\text{Scooty}) = 0.1$$

$$P(\text{Bus / late}) = ?, P(\text{Car / late}) = ?, P(\text{Scooty / late}) = ?$$

Ans) $P(B/L) = P(B \cap L) / P(L) = P(L/B).P(B) / P(L) = (0.5)(0.2) / P(L)$

$$\begin{aligned} P(L) &= P(B \cap L) + P(C \cap L) + P(S \cap L) \\ &= P(B).P(L/B) + P(C).P(L/C) + P(S).P(L/S) \\ &= (0.2)(0.5) + (0.7)(0.1) + (0.1)(0.2) = 0.19 \end{aligned}$$

$$P(B/L) P(B \cap L) / P(L) = P(L/B).P(B) / P(L) = (0.5)(0.2) / 0.19 = 10/19$$

$$P(C/L) = P(C \cap L) / P(L) = P(L/C). P(C) / P(L) = (0.1)(0.7) / 0.19 = 7/19$$

$$P(S/L) = P(S \cap L) / P(L) = P(L/S). P(S) / P(L) = (0.2)(0.1) / 0.19 = 2/19$$



Probability Distributions

- **Definition:**

Describes the likelihood of different outcomes in a random experiment.

- **Examples:**

- Uniform distribution: Equal probability for all outcomes.
- Normal distribution: Bell-shaped curve, common in many natural phenomena.



Applications of Probability

Probability theory plays a significant role in various aspects of mining engineering. Here are some applications:

Resource Estimation: Probability theory is extensively used in estimating the reserves of minerals or ores in a given area. Techniques such as kriging and geostatistics rely on probabilistic models to interpolate and extrapolate data from sampling points to estimate the quantity and quality of mineral resources.

Risk Assessment: Mining projects involve various risks, including geological uncertainties, market fluctuations, and operational hazards. Probability theory helps in quantifying these risks through techniques like Monte Carlo simulation, which evaluates the potential outcomes of different scenarios based on probabilistic inputs.

Safety Analysis: Probability theory is employed in assessing safety risks associated with mining operations. By analyzing historical data and identifying potential hazards, engineers can calculate probabilities of accidents or failures, allowing them to implement preventive measures and design safety protocols accordingly.



Applications of Probability

Probability theory plays a significant role in various aspects of mining engineering. Here are some applications:

Equipment Reliability: Mining equipment reliability is crucial for maintaining productivity and minimizing downtime. Probability theory is used to model the reliability and availability of equipment, predicting failure rates and optimizing maintenance schedules to ensure continuous operation.

Environmental Impact Assessment: Probability theory assists in assessing the environmental impact of mining activities. By analyzing probabilistic models of pollutant dispersion, groundwater contamination, and ecosystem disruption, engineers can evaluate the potential consequences of mining operations on the environment and devise mitigation strategies.

Exploration Decision Making: Probability theory aids in decision-making during mineral exploration. Through techniques like Bayesian inference, engineers can update their beliefs about the presence and characteristics of mineral deposits based on new data, guiding the allocation of exploration resources more efficiently.



Applications of Probability

Probability theory plays a significant role in various aspects of mining engineering. Here are some applications:

Grade Control: Probability theory is employed in grade control strategies to optimize ore extraction and minimize waste. By incorporating probabilistic models of ore grade variability, mining engineers can design sampling protocols and ore blending strategies to maximize the economic value of extracted material.

Financial Analysis: Probability theory is used in financial modeling and risk analysis of mining projects. By evaluating the probabilistic distribution of costs, revenues, and commodity prices, stakeholders can assess the financial viability of investments, make informed decisions, and manage investment risks effectively.

Overall, probability theory serves as a **fundamental tool in various aspects** of mining engineering, helping professionals make informed decisions, manage risks, and optimize resource utilization throughout **the lifecycle of mining projects**.



REFERENCES

- Introduction to Probability and Statistics for Engineers and Scientists, Sixth Edition, Sheldon M. Ross
- Statistical Methods Combined Edition (Volume I& II), N G Das



CONCLUSION

- Discussed Independent Events
- Discussed Multiplication Theorem
- Discussed Total Probability
- Discussed Bayes Theorem
- Discussed the Applications of Probability in Mining Industry





THANK YOU



JAN 2024

MINE AUTOMATION AND DATA ANALYTICS



MINE AUTOMATION AND DATA ANALYTICS





SWAYAM NPTEL COURSE ON

MINE AUTOMATION AND DATA ANALYTICS

By

Prof. Radhakanta Koner

Department of Mining Engineering
Indian Institute of Technology (ISM) Dhanbad

Module 8: Inferential Statistics

Lecture 19A: Discrete Random Variable
Part I



CONCEPTS COVERED

- Definition of Random Variable with an example
- Types of Random Variables: Discrete and Random Variable
- Probability Mass Function, graph and examples.
- Cumulative Distribution Function, graph and examples.
- Definition of Expectation and its Properties.
- Variance: Definition



Random Variable

- In probability experiments, our focus often lies not in every detail of the experiment's outcome, but rather in the numerical value of certain quantities derived from the result.
- For instance, when rolling a dice twice, we may only be concerned with the sum of the outcomes rather than the specific values on each individual dice. This means we might only care about knowing that the sum is seven, without being interested in whether the actual outcome was (1,6), (2,5), (3,4), (4,3), (5,2), or (6,1).
- These quantities of interest, or more formally, these real-valued functions defined on the sample space, are referred to as random variables.
- Since the outcome of the experiment determines the value of a random variable, we can assign probabilities to the possible values of the random experiment.



Random Variable

Rolling a dice: Sample Space

- The sample space for this experiment, denoted as S , consists of all possible outcomes when a dice is rolled twice.
- $\{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),$
 $(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),$
 $(3,1),(3,2),(3,3),(3,4),(3,5),(3,6),$
 $(4,1),(4,2),(4,3),(4,4),(4,5),(4,6),$
 $(5,1),(5,2),(5,3),(5,4),(5,5),(5,6),$
 $(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}$
- Given the probabilities associated with the question, we need to determine how many outcomes will yield a sum of 3.
- It's important to note that the experiment and sample space used to address this question remain the same.



Random Variable

- Let X represent the sum of outcomes from the two rolls.
- Therefore, X can take on values ranging from 2 to 12.

$$P(X=2)=P[1,1]=1/36,$$

$$P(X=3)=P[(1,2),(2,1)]=2/36$$

$$P(X=4)=P[(1,3),(2,2),(3,1)]=3/36$$

$$P(X=5)=P[(1,4),(2,3),(3,2),(4,1)]=4/36$$

$$P(X=6)=P[(1,5),(2,4),(3,3),(4,2),(5,1)]=5/36$$

$$P(X=7)=P[(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)]=6/36$$

$$P(X=8)=P[(2,6),(3,5),(4,4),(5,3),(6,2)]=5/36$$

$$P(X=9)=P[(3,6),(4,5),(5,4),(6,3)]=4/36$$

$$P(X=10)=P[(4,6),(5,5),(6,4)]=3/36$$

$$P(X=11)=P[(5,6),(6,5)]=2/36,$$

$$P(X=12)=P[6,6]=1/36$$



Discrete and Continuous Random Variables

Definition:

A discrete random variable is characterized by its ability to assume, at most, a countable number of possible values.

Consequently, any random variable capable of adopting either a finite number or a countably infinite number of distinct values qualifies as a discrete random variable.

It's worth noting that there are also random variables whose set of potential values is uncountably infinite.

Definition:

Continuous random variables pertain to scenarios where outcomes of random events are numerical, yet cannot be enumerated and are infinitely divisible.



Discrete Random Variable

- A discrete random variable is characterized by having possible values that are distinct points along the real number line.
- Discrete random variables are often associated with counting scenarios

Continuous Random Variable

- A continuous random variable is defined by its possible values spanning an interval along the real number line.
- Continuous random variables typically involve measurement scenarios.



Discrete and Continuous Random Variable Examples

Examples of discrete random variables include:

- The number of people in a house
- The number of languages a person can speak
- The number of times a student takes a particular test before qualifying
- The number of collisions at an intersection
- The number of spelling mistakes in a document

Examples of Continuous random variables include:

- Temperature of a patient.
- Height of an athlete
- Speed of a vehicle.
- Time taken by a person to come home from the office.



Probability Mass Function (p.m.f)

- A random variable characterized by its ability to assume, at most, a countable number of potential values is termed a discrete random variable.
- Let X be a discrete random variable, and suppose that it has n possible values, which we will label x_1, x_2, \dots, x_n . For a discrete random variable X , we can define the probability mass function $p(x)$ of X by

$$P(x_i) = P(X = x_i)$$

- Represent in Tabular Form

X	x_1	x_2	x_3	x_n
$P(X = x_i)$	$P(x_1)$	$P(x_2)$	$P(x_3)$	$P(x_n)$



Key Properties of p.m.f

- The probability mass function $p(x)$ is positive for, at most, a countable number of x values.
- if X must assume one of the values x_1, x_2, \dots, x_n .
 - $p(x_i) \geq 0, i = 1, 2, 3 \dots n$
 - $p(x_i) = 0$, for all other x values
- Since X must take one of the values x_i , we have

$$\sum_{i=1}^{\infty} p(x_i) = 1$$

X	x_1	x_2	x_3	x_n
$P(X = x_i)$	$P(x_1)$	$P(x_2)$	$P(x_3)$	$P(x_n)$



Example 1 of p.m.f

X is a random variable that assumes three values: 0, 1, and 2, with the corresponding probabilities as follows:

$$P(X=0): 1/3$$

$$P(X=1): 1/3$$

$$P(X=2): 1/3$$

1. Each probability is greater than or equal to 0, i.e., non-zero.
 2. Sum of probabilities = $1/3 + 1/3 + 1/3 = 1$
-
- Two key properties are satisfied.
 - It is p.m.f.



Example 2 of p.m.f

- Flipping a coin three times.
- Sample Space $S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$
- Let X denote the random variable representing the count of heads in the tosses.
- What is the Probability Mass Function?

Solution:

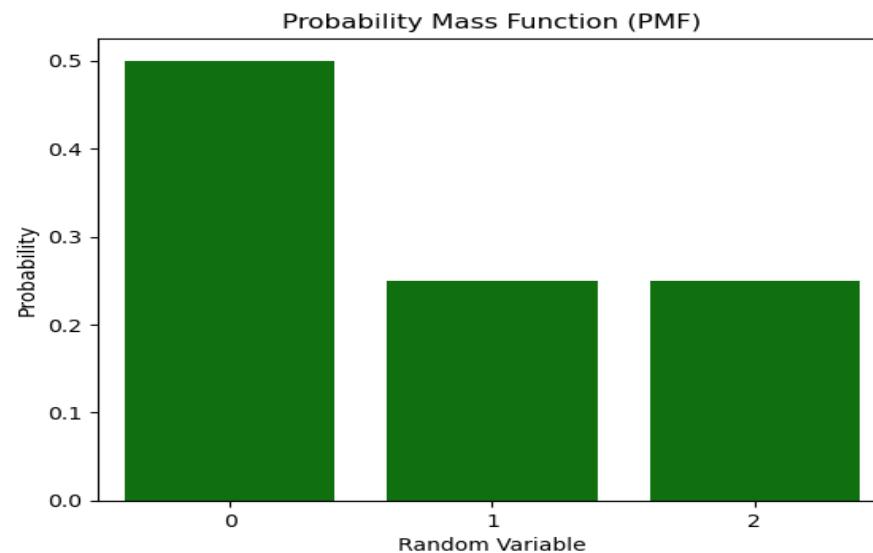
X	0	1	2	3
$P(X=x_i)$	$1/8$	$3/8$	$3/8$	$1/8$

- Key properties are satisfied.
- Hence it is Probability Mass Function (p.m.f)



Graph of Probability Mass function

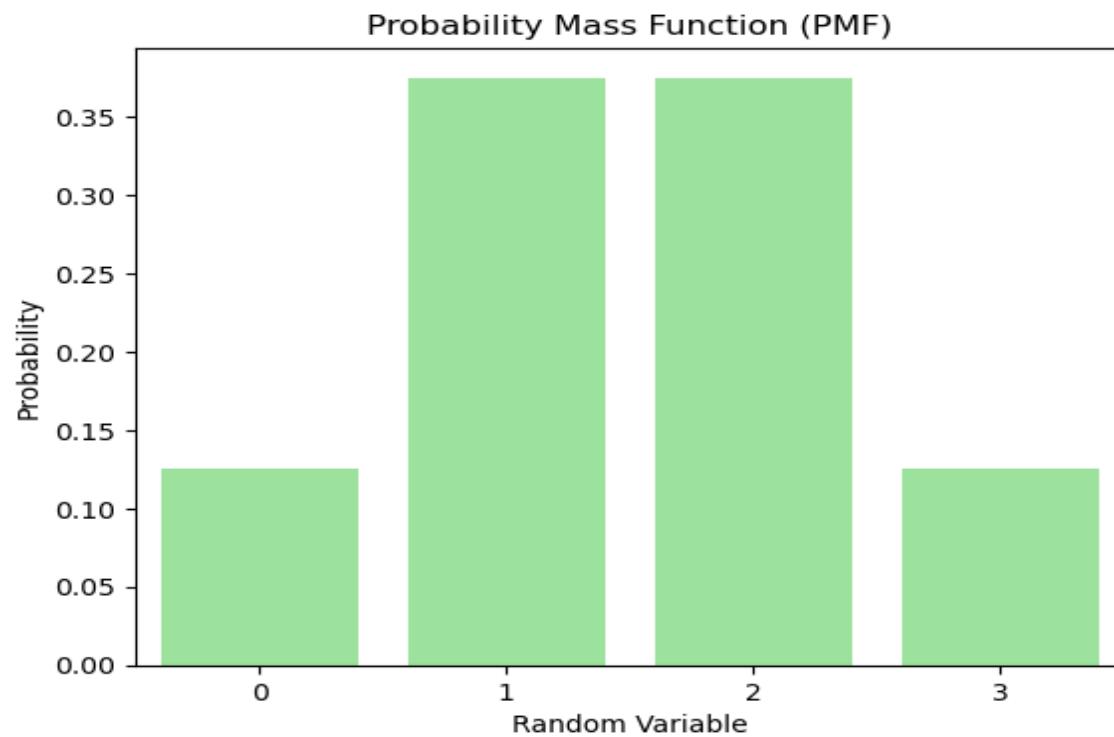
- Illustrating the probability mass function graphically can be beneficial, often done by plotting $P(X=x_i)$ on the y-axis against x_i on the x-axis.



X	0	1	2
$P(X = x_i)$	1/2	1/4	1/4

Tossing a coin thrice X = the number of heads

X	0	1	2	3
$P(X=x_i)$	1/8	3/8	3/8	1/8



Cumulative distribution function

- The cumulative distribution function (CDF), denoted as F , can be represented as follows:
$$F(a) = P(X \leq a)$$
- If X is a discrete random variable whose possible values are x_1, x_2, x_3, \dots , where $x_1 < x_2 < x_3 \dots$, then a step function will be the distribution function F of X .



Cumulative distribution function

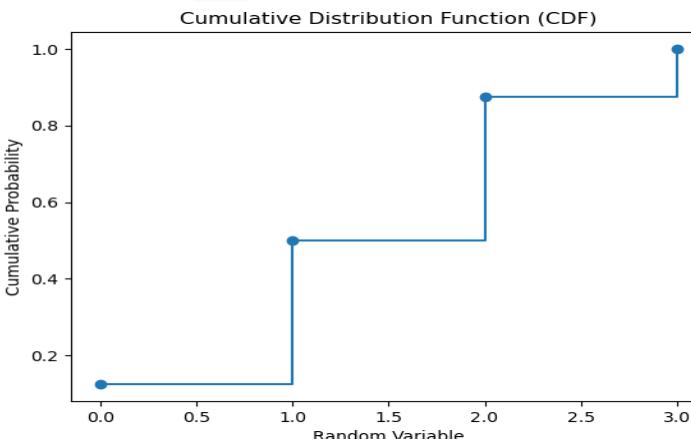
Step Function

- Let X be a discrete random variable with the following probability mass function.

X	0	1	2	3
P(X = x ₁)	1/8	3/8	3/8	1/8

- The cumulative distribution function of X is given by

$$\bullet F(a) = \begin{cases} 0 & a < 0 \\ \frac{1}{8} & 0 \leq a < 1 \\ \frac{1}{2} & 1 \leq a < 2 \\ \frac{7}{8} & 2 \leq a < 3 \\ 1 & 3 \leq a < \infty \end{cases}$$



Note that the step size at any of the values 0, 1, 2, and 3 corresponds to the probability that X assumes that specific value.



Expectation of Random Variable

- Let X be a discrete random variable taking values x_1, x_2, \dots, x_n . The expected value of X denoted by $E(X)$ and referred to as Expectation of X is given by

$$E(X) = \sum_{i=1}^{\infty} x_i P(x_i)$$

- The Expectation of a random variable can be interpreted as the "long-run average" value of the random variable across repeated independent observations.



Expectation Example

- Consider X as a discrete random variable with the following probability mass function.

X	0	1	2	3
$P(X = x_i)$	1/8	3/8	3/8	1/8

- Expectation of X can be calculated :

$$E(X) = \sum_{i=1}^{\infty} x_i P(x_i)$$

$$= (0 \times 1/8) + (1 \times 3/8) + (2 \times 3/8) + (3 \times 1/8)$$

$$= 12/18$$

$$= 2/3$$

Bernoulli random variable

- A random variable that can assume either the value 1 or 0 is referred to as a Bernoulli random variable.
- Let X represent a Bernoulli random variable that assumes the value 1 with probability p .
- The probability distribution of this random variable is as follows:

X	0	1
$P(X = x_i)$	$1 - p$	p

- Expected value of a Bernoulli random variable: $E(X) = \sum_{i=1}^{\infty} x_i P(x_i)$
$$E(X) = 0 \times (1 - p) + 1 \times p = p$$

Discrete Uniform random variable

- Let X denote a random variable that is equally probable to assume any of the values $1, 2, 3, \dots, n$.
- Probability mass function is

X	1	2	...	n
$P(X = x_i)$	$1/n$	$1/n$...	$1/n$

$$E(X) = \sum_{i=1}^{\infty} x_i P(x_i)$$

- $E(X) = (1 \times 1/n) + (2 \times 1/n) + \dots + (n \times 1/n) = n(n+1)/2n$
- $E(X) = (n+1)/2$



Properties of Expectation

- Let X represent a discrete random variable with values x_i and its corresponding probability mass function ($P(X = x_i)$).
- Let h be any real values function; the expected value of $g(X)$ is

$$E(h(x)) = \sum_i h(x_i)P(X = x_i)$$

- If a and b are constants,

$$E(aX + b) = aE(X) + b$$

Note: $E(x^2) \neq (E(x))^2$



Expectation of sum of two random variables

- The expected value of the sum of random variables equals the sum of the individual expected values.
- In other words, let X and Y be two random variables. Then,

$$E(X+Y) = E(X) + E(Y)$$

Expectation of sum of many random variables

- The result stating that the expected value of the sum of random variables equals the sum of the expected values holds true not only for two but for any number of random variables.
- Let X_1, X_2, \dots, X_n be k discrete random variables. Then,

$$E\left(\sum_{i=1}^k X_i\right) = \sum_{i=1}^k E(X_i)$$



Need for Variance

- The expected value of a random variable provides a weighted average of its potential values, but it does not provide information about the variation or spread of these values. For example, consider the random variables X, Y, and Z, with their respective values and probabilities as follows:
- $X = 0$ *with probability 1*
- $Y = \begin{cases} -3 & \text{with probability } 1/3 \\ 3 & \text{with probability } 1/3 \end{cases}$
- $Z = \begin{cases} -500 & \text{with probability } 1/2 \\ 500 & \text{with probability } 1/2 \end{cases}$
- $E(X) = E(Y) = E(Z) = 0.$
- However, it's evident that the spread of Z is greater than that of Y, and Y's spread is greater than that of X.



Variance of a random variable

- Let's denote the expected value of a random variable X by the Greek letter μ .
- If X is a random variable with an expected value μ , then the variance of X , denoted by $\text{Var}(X)$ or $V(X)$, is defined by:

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

- In essence, the variance of a random variable X quantifies the squared difference between the random variable and its mean μ on average.



Computational formula for $\text{Var}(X)$

- $\text{Var}(X) = E(X - \mu)^2$
- $(X - \mu)^2 = X^2 + \mu^2 - 2\mu X$
- **Using properties of Expectation , We know**

$$\begin{aligned} E(X - \mu)^2 &= E(X^2 + \mu^2 - 2\mu X) \\ &= E(X^2) + \mu^2 - 2\mu E(X) \\ &= E(X^2) + \mu^2 - 2\mu^2 \\ &= E(X^2) - \mu^2 \\ \boxed{\text{Var}(X) = E(X^2) - (E(X))^2} \end{aligned}$$



Rolling a dice once

- **Random Experiment:** Roll a dice once
- **Sample Space:** $S = \{1, 2, 3, 4, 5, 6\}$
- **Random variable X is the outcome of the roll.**
- **The probability distribution is given by**

X	1	2	3	4	5	6
X^2	1	4	9	16	25	36
$P(X = x_i)$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$

$$E(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$$

$$E(X^2) = 1 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 9 \cdot \frac{1}{6} + 16 \cdot \frac{1}{6} + 25 \cdot \frac{1}{6} + 36 \cdot \frac{1}{6} = 15.167$$

$$Var(X) = E(X^2) - (E(X))^2 = 15.167 - (3.5)^2 = 2.917$$



Bernoulli random variable

- A random variable that can assume either the value 1 or 0 is referred to as a Bernoulli random variable.
- Let X be a Bernoulli random variable that takes on the value 1 with probability p .
- The probability distribution of the random variable is as follows:

X	0	1
X^2	0	1
$P(X = x_i)$	$1 - p$	p

Expected value of a Bernoulli random variable:

$$E(X) = 0 \times (1 - p) + 1 \times p = p$$

$$E(X^2) = 0 \times (1 - p) + 1 \times p = p$$

$$E(X) = \sum_{i=1}^{\infty} x_i P(x_i)$$

Var (X) = Variance of Bernoulli's random variable

$$= E(X^2) - (E(X))^2 = p - p^2 = p(1 - p)$$



Discrete Uniform random variable

- Let X represent a random variable that has an equal likelihood of assuming any of the values 1, 2, 3, ..., n .
- Probability mass function

X	1	2	...	n
X^2	1	4	...	n^2
$P(X = xi)$	$1/n$	$1/n$...	$1/n$

- $E(X) = (1 \times 1/n) + (2 \times 1/n) + \dots + (n \times 1/n) = (n+1)/2$
- $E(X^2) = (1 \times 1/n) + (4 \times 1/n) + \dots + (n^2 \times 1/n) = (n+1)(2n+1)/6$
- $Var(X) = E(X^2) - (E(X))^2 = (n^2-1)/12$



REFERENCES

- Introduction to Probability and Statistics for Engineers and Scientists, Sixth Edition, Sheldon M. Ross
- Statistical Methods Combined Edition (Volume I& II), N G Das



CONCLUSION

- Discussed about Random Variable
- Discussed types of Random Variables: Discrete and Continuous
- Discussed Probability Mass Function, graph, and examples.
- Discussed Cumulative Distribution Function, graph, and examples.
- Discussed Expectation: Definition and its Properties
- Discussed and calculated Variance for different random variables.





THANK YOU



JAN 2024

MINE AUTOMATION AND DATA ANALYTICS



MINE AUTOMATION AND DATA ANALYTICS





SWAYAM NPTEL COURSE ON MINE AUTOMATION AND DATA ANALYTICS

By

Prof. Radhakanta Koner

Department of Mining Engineering
Indian Institute of Technology (ISM) Dhanbad

Module 8: Inferential Statistics

**Lecture 19B: Discrete Random Variable
Part II**



CONCEPTS COVERED

In Continuation:

- Properties of Variance and Standard Deviation.
- Bernoulli Distribution
- Binomial Distribution
- Uniform Distribution



Properties of Variance

Let X be a random variable, let c be a constant, then

$$\begin{aligned} \text{Var}(cx) &= c^2 \text{Var}(X) \\ \text{Var}(X + c) &= \text{Var}(X) \end{aligned}$$

If a and b are constants, $\text{V}(aX + b) = a^2\text{V}(X)$

Proof.

We know $E(ax + b) = a\mu + b$,
 $\text{Var}(ax + b) = E(ax + b - a\mu - b)^2 = a^2E(X - \mu)^2 = a^2\text{Var}(X)$



Variance of sum of two random variables

- The expected value of the sum of random variables equals the sum of the individual expected values. In other words, let X and Y be two random variables. Then,

$$E(X + Y) = E(X) + E(Y)$$

- Where as $Var(X + X) = Var(2X) = 4Var(X) \neq Var(X) + Var(X)$
- $4Var(X) \neq 2 Var(X)$
- Is this statement always true in all cases?



Independent random variables

Definition

Random variables X and Y are considered independent if the knowledge of the value of one of them does not alter the probabilities associated with the other.

Example: Roll a dice twice. $S = \{(1, 1), \dots, (6, 6)\}$

- **X = the outcome of the first dice.**
- **Y = the outcome of the second dice**
- Knowing $X = i$ does not change the probability of Y taking any value of 1,2,..,6 .
- **X and Y are independent random variables.**



Variance of sum of independent random variables

Result

Let X and Y be independent random variables. Then

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$



Example: Rolling a dice twice

- Let X represent the outcome of one fair dice, and let Y represent the outcome of another fair dice.
- We observe that $E(X) = E(Y) = 3.5$.
- The sum of outcomes of both dice rolled together, denoted as $X+Y$, is calculated to have $\text{Var}(X) = \text{Var}(Y) = 2.917$.
- Since X and Y are independent, we compute $\text{Var}(X+Y) = 2.917 + 2.917 = 5.83$, which aligns with the result obtained by applying the computational formula.



Variance of sum of many independent random variables

- The outcome stating that the variance of the sum of independent random variables equals the sum of the variances applies not only to two but to any number of random variables.
- Let X_1, X_2, \dots, X_k be k discrete random variables. Then,

$$Var\left(\sum_{i=1}^k X_i\right) = \sum_{i=1}^k Var(X_i)$$



Standard Deviation (SD) of a random variable

Definition: The quantity $SD(X) = \sqrt{Var(X)}$ is the standard deviation of X.

Therefore, the standard deviation (SD) is defined as the positive square root of the variance.

Similar to the expected value, the standard deviation is expressed in the same units as the random variable.



Properties of Standard Deviation

- Let X be a random variable, let c be a constant, then

$$\begin{aligned}SD(cX) &= c \cdot SD(X) \\SD(X + C) &= SD(X)\end{aligned}$$



Binomial Distribution

Bernoulli Trail

- A trial or experiment, where the outcome can be categorized as either a success or a failure, is referred to as a Bernoulli trial.
- The sample space $S = \{\text{Success, Failure}\}$
- Let X represent a random variable that takes the value 1 if the outcome is a success and 0 if the outcome is a failure.
- X is referred to as a Bernoulli random variable.



Examples of Bernoulli trials

Experiment 1:

Tossing a coin: $S = \{\text{Head, Tail}\}$

Success : Head

Failure : Tail

Experiment 2:

Rolling a dice: $S = \{1,2,3,4,5,6\}$

Success : Getting a six.

Failure : Getting any other number



Non Bernoulli trial

- Experiment: Selecting a person at random and inquiring about their age.
- This experiment does not qualify as a Bernoulli trial because it does not entail only two possible outcomes.



Bernoulli random variable

- A random variable that can assume either the value 1 or 0 is termed a Bernoulli random variable.
- Let X be a Bernoulli random variable that takes on the value 1 with probability p .
- The probability distribution of this random variable is as follows:

X	0	1
$P(X = xi)$	$1 - p$	p

- Expected value of a Bernoulli random variable:

$$E(X) = \sum_{i=1}^{\infty} x_i P(x_i)$$

$$E(X) = 0 \times (1 - p) + 1 \times p = p$$

- Variance of a Bernoulli random variable : $V(X) = p - p^2 = p(1-p)$



Variance of Bernoulli Distribution

- The maximum variance occurs when $p = \frac{1}{2}$, when success and failure are equally probable.
- In simpler terms, the most uncertain Bernoulli trials, characterized by the largest variance, resemble the tossing of a fair coin.



Independent and Identically distributed Bernoulli trials

- $N = 3$ independent trials
- Let $n = 3$ independent bernoulli trials.
- Let p is the probability of success.
- The probability of outcomes of the independent trials are

<i>Sl. No</i>	<i>Outcome</i>	<i>Number of successes</i>	<i>Probabilities</i>
1	(s, s, s)	3	$p \cdot p \cdot p$
2	(s, s, f)	2	$p \cdot p \cdot (1 - p)$
3	(s, f, s)	2	$p \cdot (1 - p) \cdot p$
4	(s, f, f)	1	$p \cdot (1 - p) \cdot (1 - p)$
5	(f, s, s)	2	$(1 - p) \cdot p \cdot p$
6	(f, s, f)	1	$(1 - p) \cdot p \cdot (1 - p)$
7	(f, f, s)	1	$(1 - p) \cdot (1 - p) \cdot p$
8	(f, f, f)	0	$(1 - p) \cdot (1 - p) \cdot (1 - p)$



N = 3 independent trials, X = number of successes

- Let $n = 3$ independent Bernoulli trials
- Let p is the probability of success.
- Let X = number of successes in 3 independent trials.
- The probability distribution of X

X	0	1	2	3
$P(X = i)$	$(1 - p)^3$	$3p(1 - p)^2$	$3p^2(1 - p)$	p^3



N independent trials , X = number of successes

- Consider any outcome that results in a total of i successes.
- The outcome will have a total of i successes and $(n - i)$ failures.
- Probability of i success and $(n - i)$ failures = $p^i \cdot (1 - p)^{n-i}$
- There number of different outcomes that result in i successes and $(n - i)$ failures = $\binom{n}{i}$
- The probability of i success in n trials is given by

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



Binomial Random Variable

Definition

- Let X be a binomial random variable with parameters n and p, which denotes the number of successes in n independent Bernoulli trials, where each trial has a success probability of p
- X takes values 0,1, 2,3,...,n with the probability.

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



Example: Tossing a coin thrice

$$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

- Success = head, Failure = tails
- X is the random variable which counts the number of heads in the tosses.
- N = 3, P = 0.5
- Probability Mass function

X	0	1	2	3
$P(X=x_i)$	$\binom{3}{0} (1/2)^0 (1/2)^3$ $= 1/8$	$\binom{3}{1} (1/2)^1 (1/2)^2$ $= 3/8$	$\binom{3}{2} (1/2)^2 (1/2)^1$ $= 3/8$	$\binom{3}{3} (1/2)^3 (1/2)^0$ $= 1/8$



Shape of the pmf for same n different p

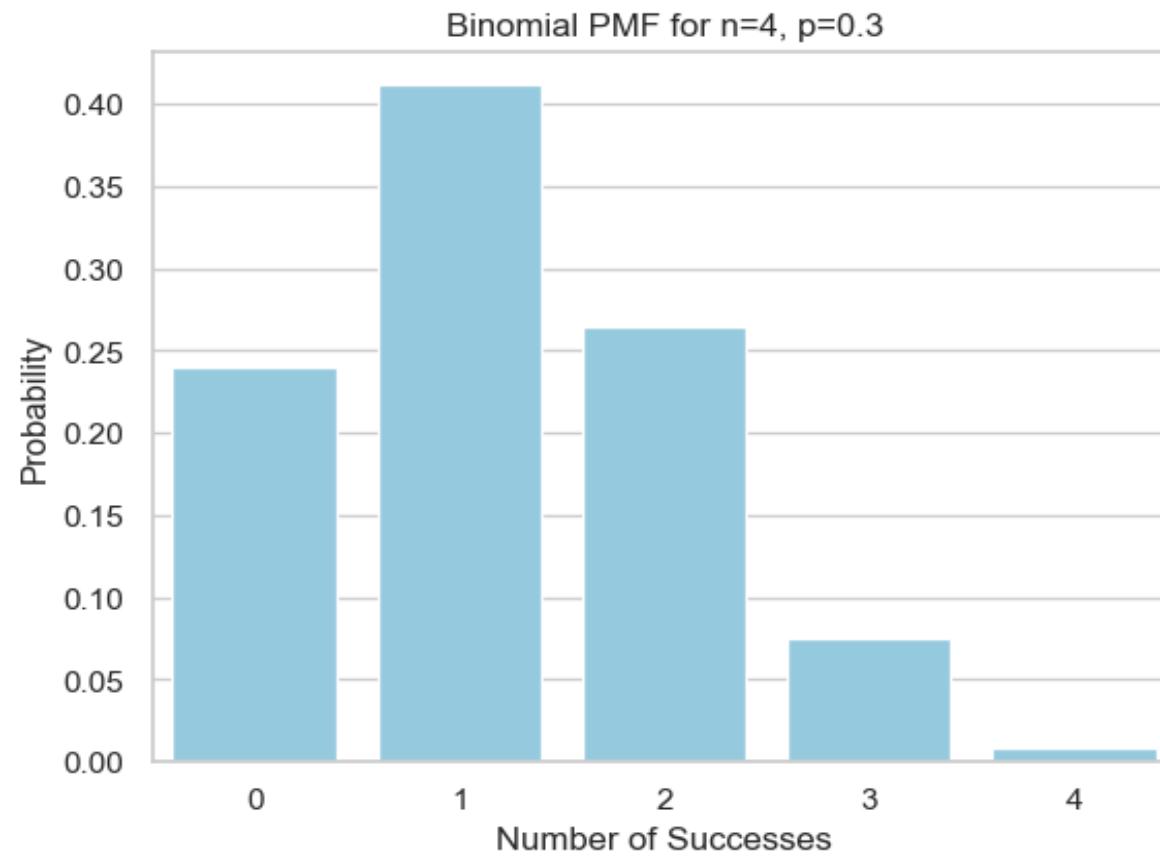
A binomial distribution is

- Right skewed if $p < 0.5$
- symmetric if $p = 0.5$
- Left skewed if $p > 0.5$
- We demonstrate the same for $n = 4$ and different p

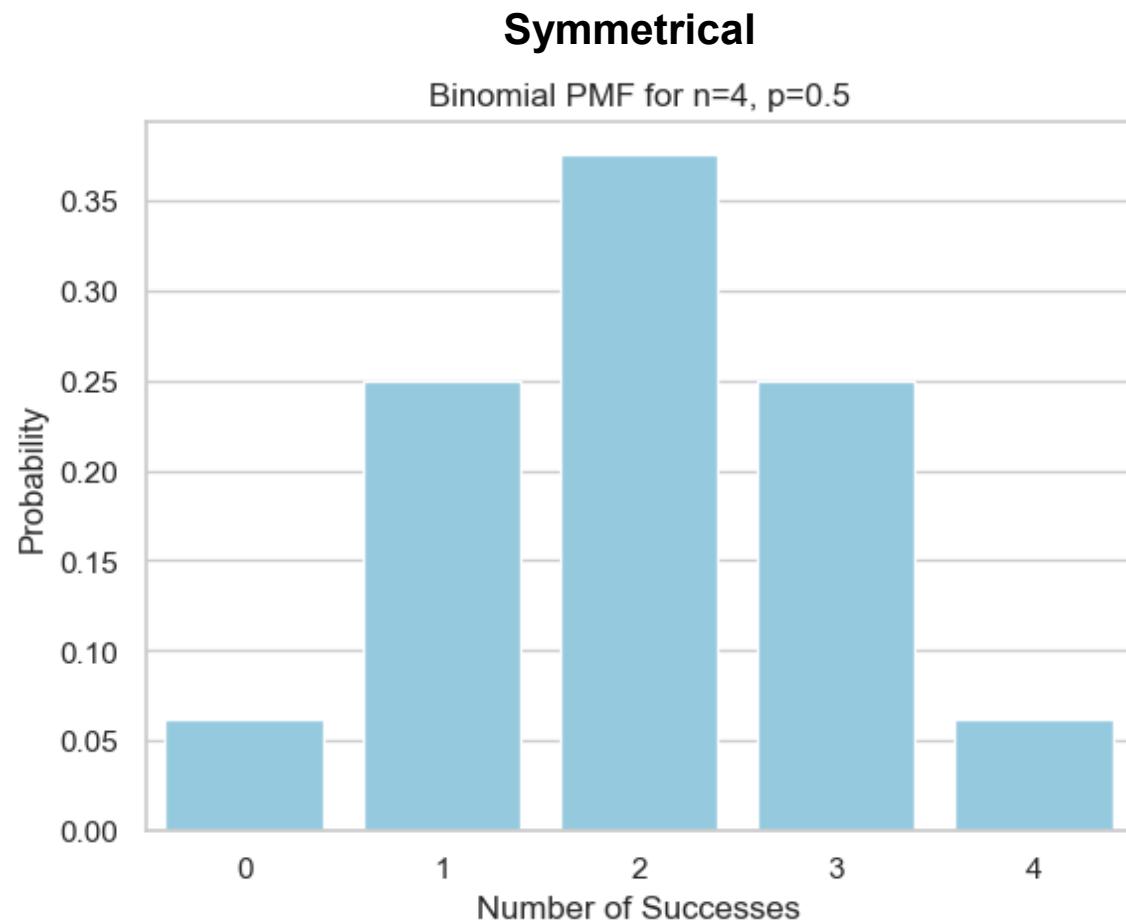


$N = 4$, $p = 0.3$, $X = \text{number of success}$

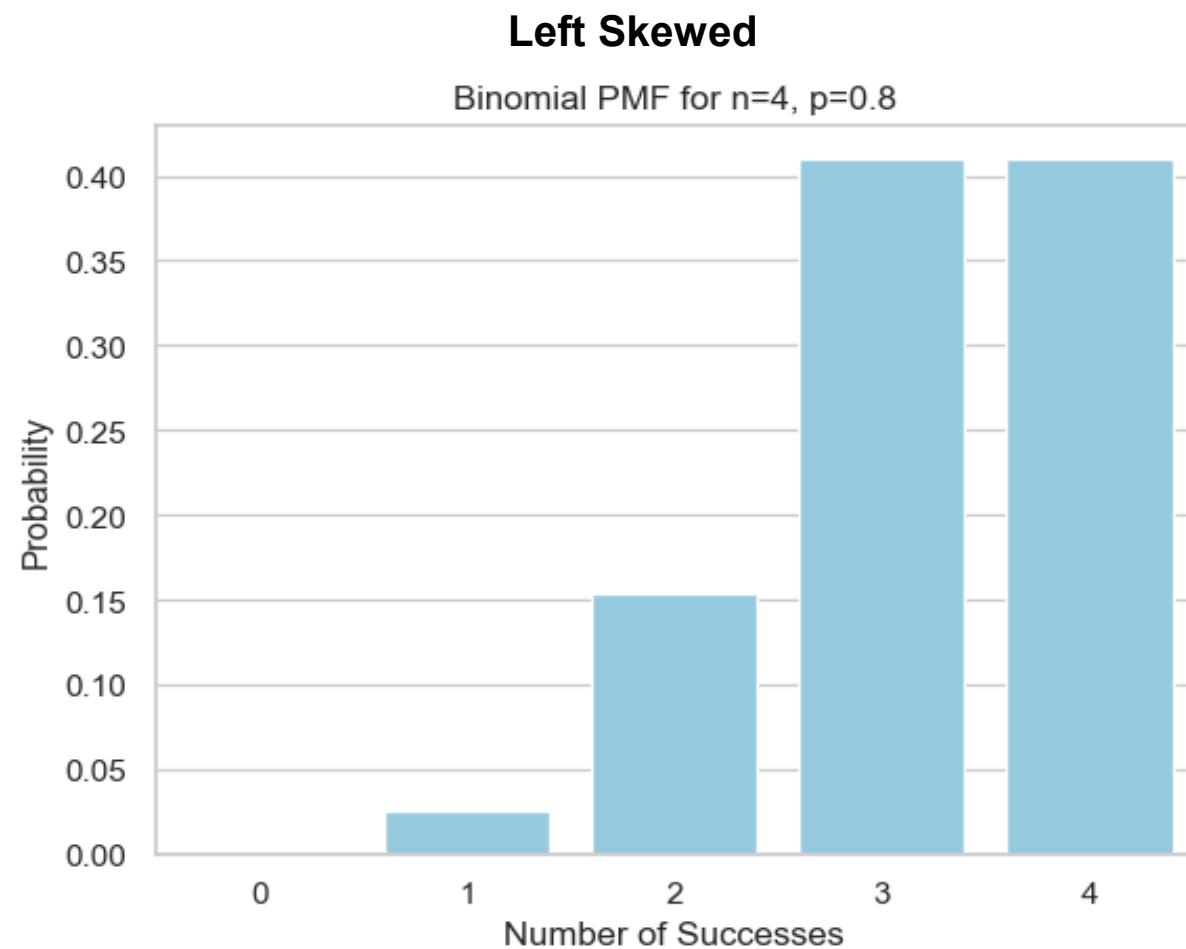
Right Skewed



$N = 4, p = 0.5, X = \text{number of success}$

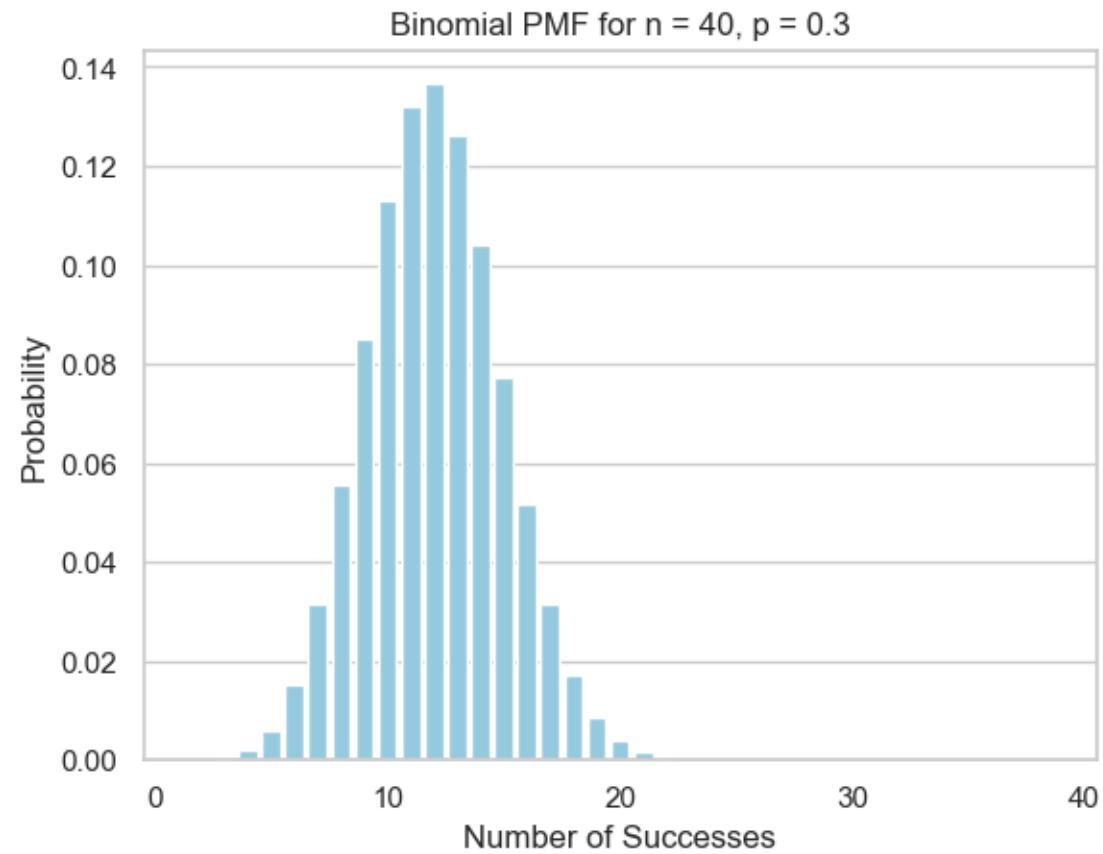


$N = 4$, $p = 0.8$, X = number of success



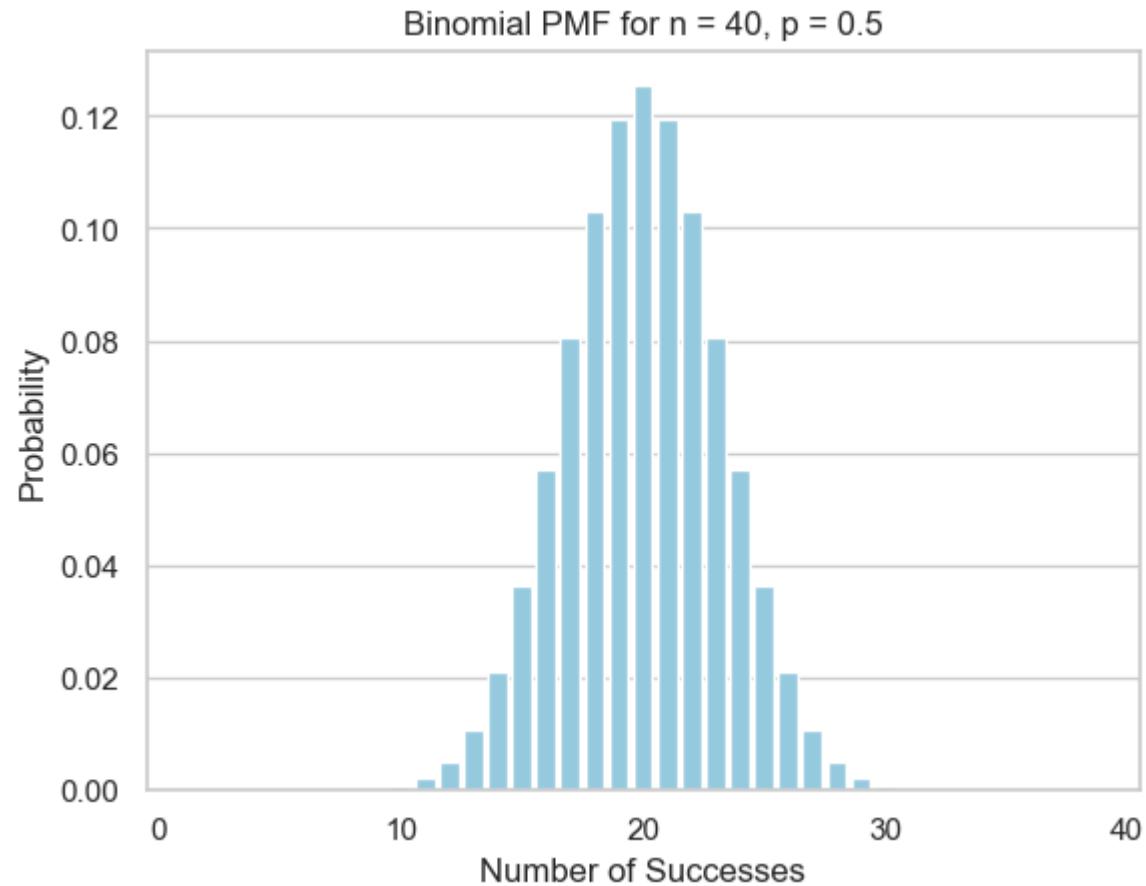
Graph of pmf of Binomial distribution

Right Skewed – for larger n

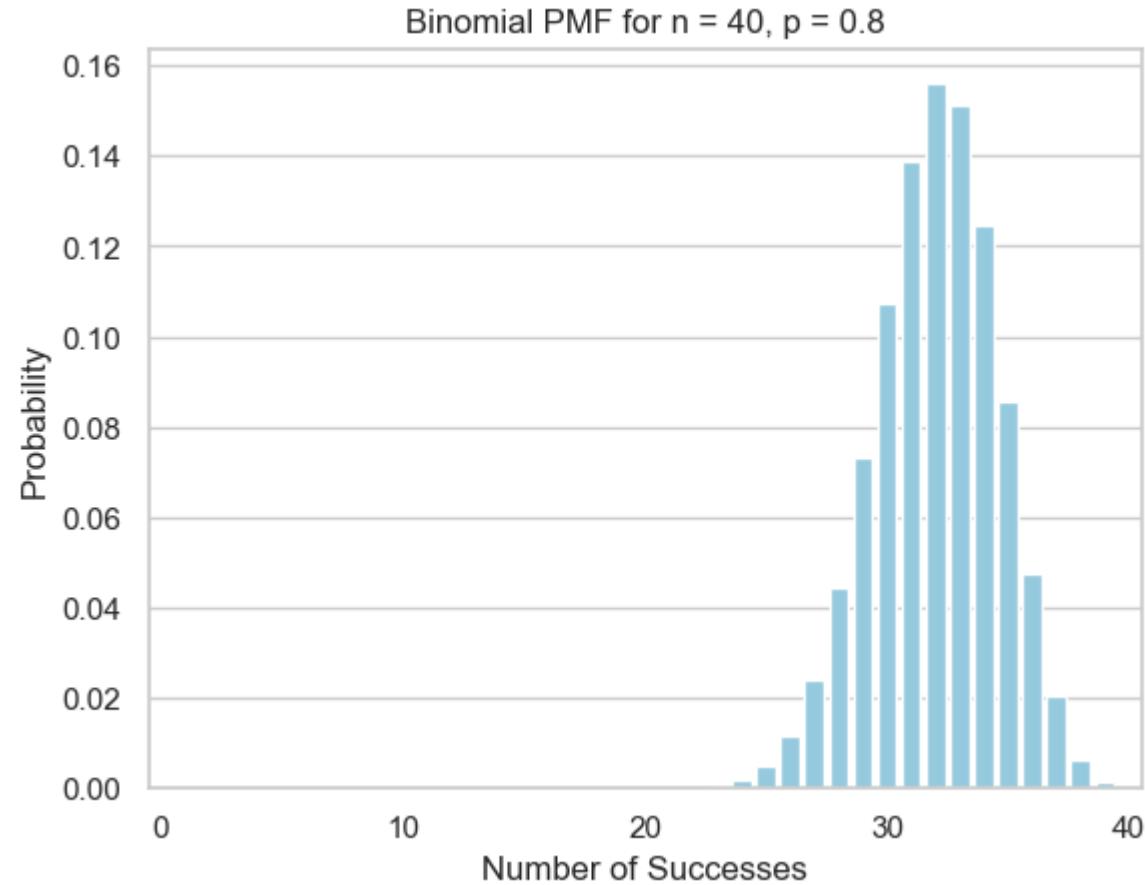


Graph of pmf of Binomial distribution

Symmetrical – for larger n



Graph of pmf of Binomial distribution left skewed – for larger n



Effect of n and p on the shape of the distribution

- For small n and small p, the distribution is right-skewed.
- For small n and large p, the distribution is left-skewed.
- For small n and $p=0.5$, the distribution is symmetric.
- As n becomes large, the binomial distribution tends towards symmetry.



Expectation and Variance of Binomial Random Variable

- A binomial random variable $X \sim \text{Bin}(n, p)$ equals the number of successes in n independent trials when each trial is a success with probability p .
- We can represent X as

$$X = X_1 + X_2 + \dots + X_n$$

- Where X_i is equal to 1 if trail i is a success and is equal to 0 if trail i is a failure.

$$P(X_i = 1) = p$$

$$P(X_i = 0) = 1 - p$$



Expectation and Variance of Binomial Random Variable

$$X = X_1 + X_2 + \dots + X_n$$

- $E(X) = E(X_1 + X_2 + \dots + X_n) = E(X_1) + E(X_2) + \dots + E(X_n)$
 $= p + p + \dots + p = np$
- $V(X) = V(X_1 + X_2 + \dots + X_n) = V(X_1) + V(X_2) + \dots + V(X_n)$
 $= p(1-p) + p(1-p) + \dots + p(1-p) = p(1-p)$

Result:

- The expectation of a binomial random variable : $E(X) = np$
- The variance of a binomial random variable : $V(X) = np(1-p)$



Example: Tossing a coin 500 times

- If a fair coin is tossed 5000 times, what is the standard deviation of the number of times a head appears?
- Let X = the number of heads in 5000 fair coin tosses. Then $X \sim Bin(500, 1/2)$
- $E(X) = np = (5000)(1/2) = 2500$
- $V(X) = np(1 - p) = 5000(1/2)(1 - (1/2)) = 1250$
- $S(X) = \sqrt{V(X)} = 35.35$

Finding probability given expectation and n

In a series of 10 coin tosses, the expected number of heads is 6. What is the probability of obtaining 8 heads?

Given $E(X) = 6$

- We already know the probability of getting a fair head each in 10 independent coin tosses = $\frac{1}{2}$
- But we don't know whether it is a fair coin.
- $np = 6, p = 0.6$
- $Bin(n = 10, p = 1/6)$
- $P(X = 8) ?$

$$\begin{aligned}P(X = i) &= \binom{n}{i} \cdot p^i \cdot (1 - p)^{n-i} \\&= \binom{10}{8} \cdot (0.6)^8 \cdot (1 - 0.6)^2 = 0.121\end{aligned}$$



Uniform Random Variable

- A discrete uniform random variable is a type of random variable that characterizes outcomes from a finite set of equally probable values.
- In essence, it indicates a scenario where each potential outcome has an identical likelihood of occurrence.
- The term "uniform" is employed because the probabilities are uniformly distributed across the range of possible values.



Uniform Random Variable

- Let X be a random variable that is equally likely to take any of the values $1, 2, 3, \dots, n$
- The probability mass function (PMF) of a discrete uniform random variable is given by

X	1	2	...	n
$P(X = x_i)$	$1/n$	$1/n$...	$1/n$

- $E(X) = (1 \times 1/n) + (2 \times 1/n) + \dots + (n \times 1/n) = (n+1)/2$
- $E(X^2) = (1 \times 1/n) + (4 \times 1/n) + \dots + (n^2 \times 1/n) = (n+1)(2n+1)/6$
- $\text{Var}(X) = E(X^2) - (E(X))^2 = (n^2-1)/12$



Example of a Uniform random variable

Example 1

- Let's take a fair six-sided die as an example. The potential outcomes when rolling the die are 1, 2, 3, 4, 5, and 6.
- Because each face has an equal probability of landing face up, the random variable representing the outcome of the die roll adheres to a discrete uniform distribution with $n=6$.
- The probability of obtaining any specific number (such as 3) is $1/6$

Example 2

- Another instance is selecting a card randomly from a thoroughly shuffled standard deck of 52 cards, where each card carries an equal probability of $1/52$



REFERENCES

- Introduction to Probability and Statistics for Engineers and Scientists, Sixth Edition, Sheldon M. Ross
- Statistical Methods Combined Edition (Volume I& II), N G Das



CONCLUSION

- We have discussed the properties of Variance and Standard Deviation.
- We have discussed about Bernoulli Distribution
- We have discussed Binomial Distribution
- We have discussed about Uniform Distribution





THANK YOU



JAN 2024

MINE AUTOMATION AND DATA ANALYTICS



MINE AUTOMATION AND DATA ANALYTICS





SWAYAM NPTEL COURSE ON MINE AUTOMATION AND DATA ANALYTICS

By

Prof. Radhakanta Koner

Department of Mining Engineering

Indian Institute of Technology (Indian School of Mines) Dhanbad

Module 8: Inferential Statistics

Lecture 20 B: Continuous Random Variable
Part II



CONCEPTS COVERED

1. Standard Normal Distribution
2. T distribution
3. Chi-Squared Distribution



JAN 2024

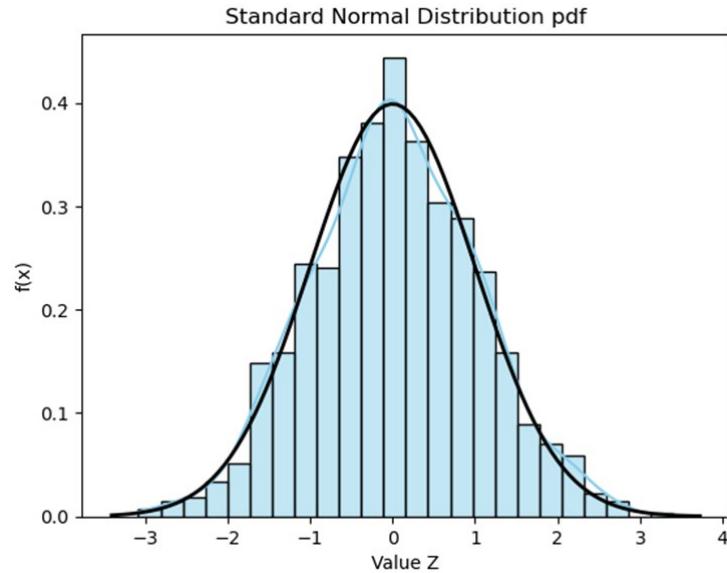
Standard Normal Distribution

$Z \sim N(0,1)$

A normal continuous random variable with an expected value (mean) of 0 and a standard deviation of 1 is referred to as a standard normal continuous random variable.

The density curve associated with it is termed the standard normal curve.

This type of random variable is symbolized by Z.



Standard Normal Distribution

$Z \sim N(0,1)$

Z-Score Formula

The statistical formula for a value's z-score is calculated using the following formula:

$$z = (x - \mu) / \sigma$$

Where:

z = Z-score

x = the value being evaluated

μ = the mean

σ = the standard deviation



Standard Normal Distribution

$Z \sim N(0,1)$

How to Calculate Z-Score

Calculating a z-score requires that you first determine the mean and standard deviation of your data. Once you have these figures, you can calculate your z-score. So, assume you have the following variables:

$$x = 57$$

$$\mu = 52$$

$$\sigma = 4$$

You would use the variables in the formula:

$$z = (57 - 52) / 4$$

$$z = 1.25$$

So, your selected value has a z-score that indicates it is 1.25 standard deviations from the mean.

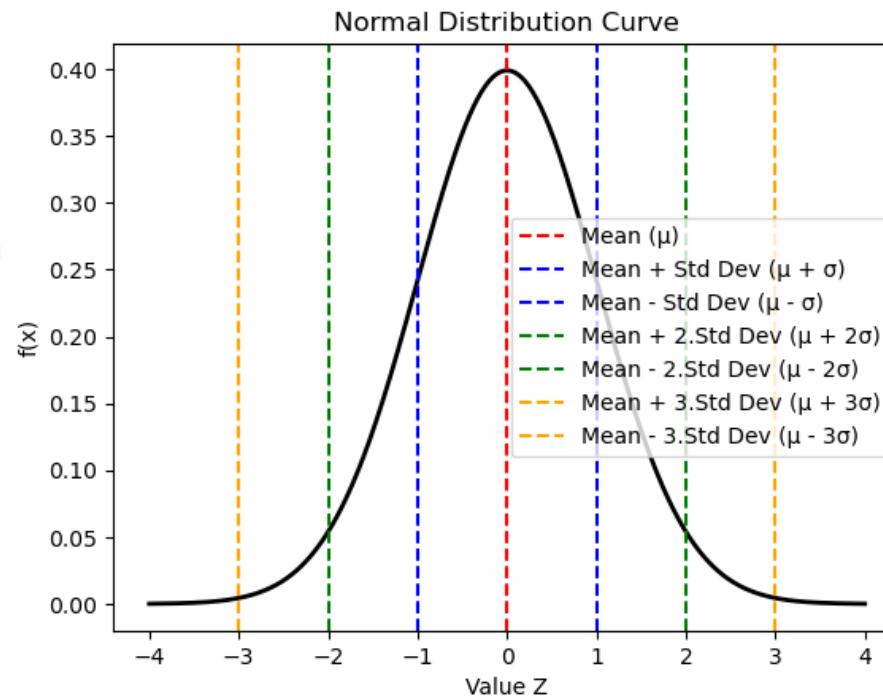


Approximate Rule to Standard Normal Distribution

$Z \sim N(0,1)$

For a continuous random variable following a normal distribution with mean μ and standard deviation σ :

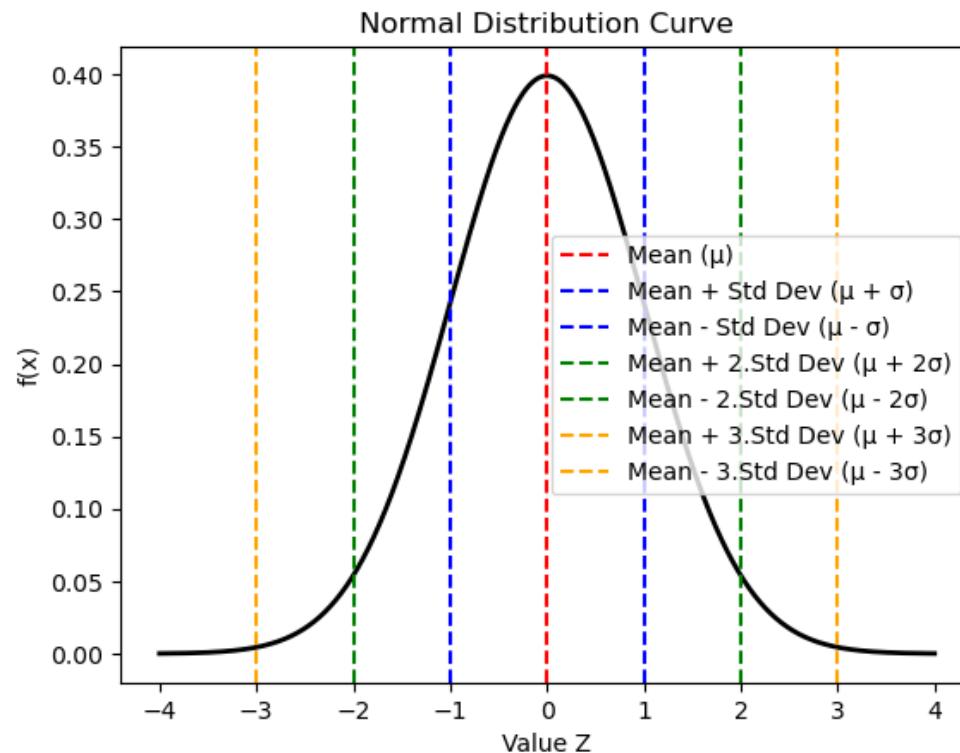
- Between $\mu + \sigma$ and $\mu - \sigma$, approximately 68% of the distribution is covered.
- Between $\mu + 2\sigma$ and $\mu - 2\sigma$, approximately 95% of the distribution is encompassed.
- Between $\mu + 3\sigma$ and $\mu - 3\sigma$, approximately 99.7% of the distribution is included.



Approximate Rule to Standard Normal Distribution

$Z \sim N(0,1)$

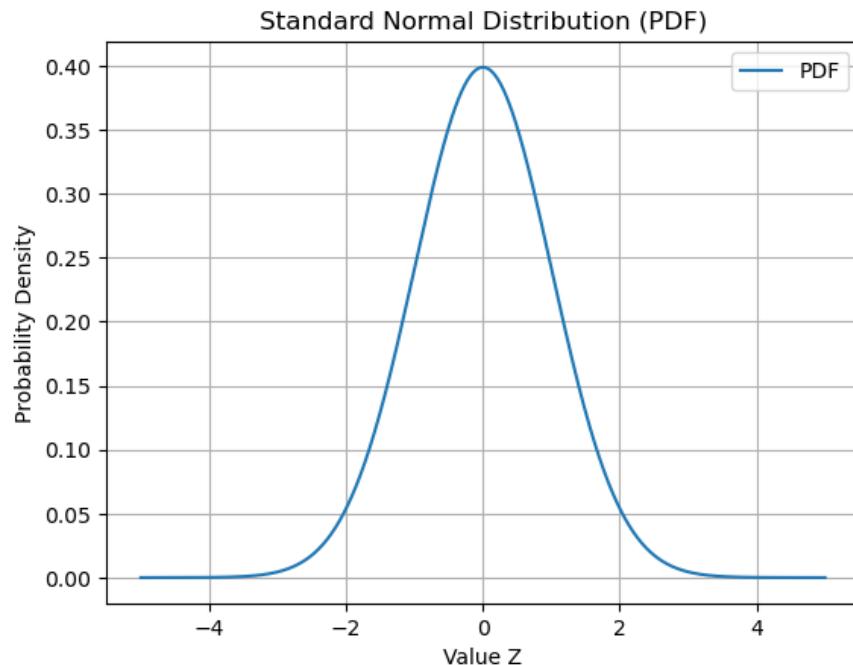
- 1) $P(-1 < Z < 1) = ?$ 0.68
- 2) $P (-2 < Z < 2) = ?$ 0.95
- 3) $P(Z > -1) = ?$ 0.84
- 4) $P(Z < -3) = ?$ 0
- 5) $P(Z < 2) = 0.975$



Properties of Z

1) 1st Property

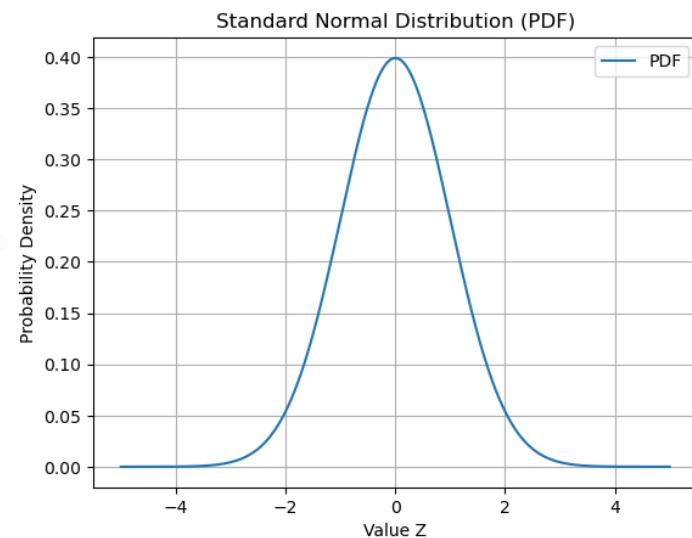
$$P(Z < -x) = 1 - P(Z < x)$$



Properties of Z

1) 2nd Property

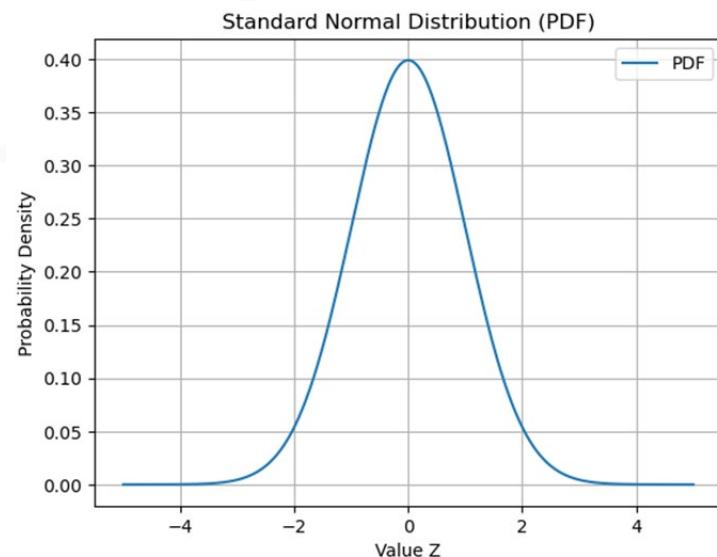
$$P(a < Z < b) = P(Z < b) - P(Z < a)$$



Properties of Z

1) 3rd Property

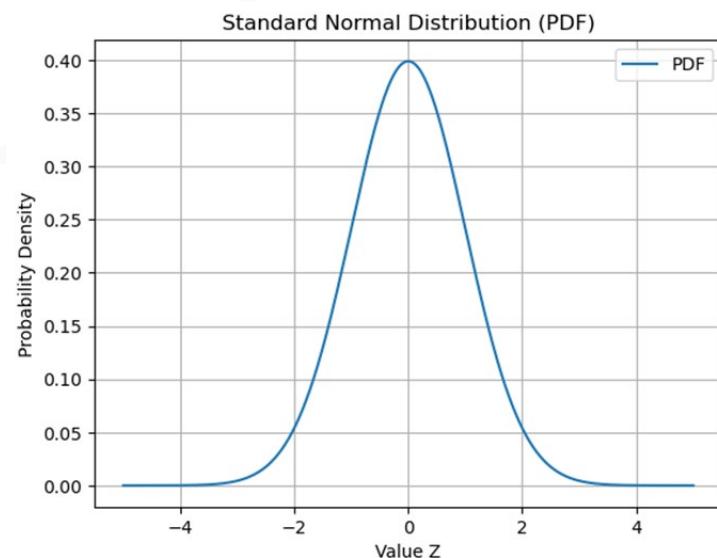
$$(|Z| > a) = 2(1 - P(Z < a))$$



Properties of Z

1) 4th Property

$$(|Z| < a) = 2 P(Z < a) - 1$$



Standard Normal Table

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361



Standard Normal Table

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361

1. $P(Z < 1.09) = 0.86214$

2. $P(Z > 2.1) = 0.01786$

**3. $P(0 < Z < 1)$
 $= P(Z < 1) - P(Z < 0)$
 $= 0.84134 - 0.5$
 $= 0.34134$**



Standard Normal Table

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361

$$\begin{aligned}
 1. P(-0.9 < Z < 2.3) &= P(Z < 2.3) - P(Z < -0.9) \\
 &= P(Z < 2.3) - P(Z > 0.9) \\
 &= 0.98928 - (1 - P(Z < 0.9)) \\
 &= 0.98928 - (0.18406) \\
 &= 0.80522
 \end{aligned}$$

$$2. P(Z > -0.96) = P(Z < 0.96)$$

$$3. P(Z < -0.53) = P(Z > 0.53)$$



Standard Normal Table

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361

$$1. P(|Z| < 0.2) = 2 \cdot P(Z < 0.2) - 1$$



What Is a T-Distribution?

- The t-distribution, also known as the Student's t-distribution, is a type of probability distribution that is similar to the normal distribution with its bell shape but has heavier tails.
- It is used for estimating population parameters for small sample sizes or unknown variances.
- T-distributions have a greater chance for extreme values than normal distributions, and as a result have fatter tails.
- The t-distribution is the basis for computing t-tests in statistics

KEY TAKEAWAYS

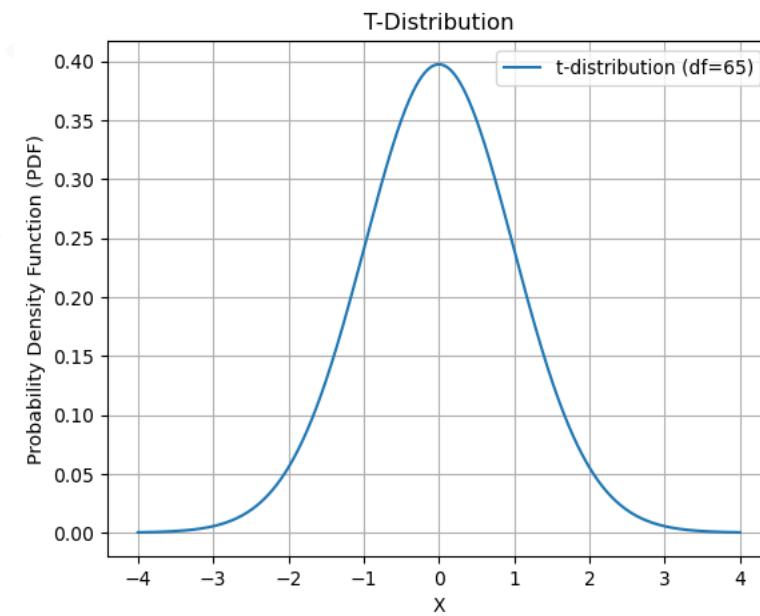
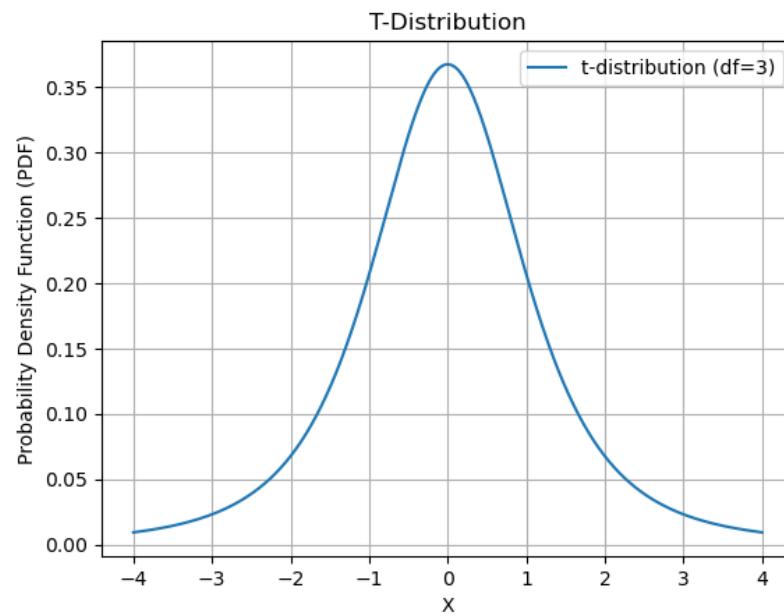
- The t-distribution is a continuous probability distribution of the z-score when the estimated standard deviation is used in the denominator rather than the true standard deviation.
- The t-distribution, like the normal distribution, is bell-shaped and symmetric, but it has heavier tails, which means that it tends to produce values that fall far from its mean.
- T-tests are used in statistics to estimate significance.



What Is a T-Distribution?

What Does a T-Distribution Tell You?

Tail heaviness is determined by a parameter of the t-distribution called degrees of freedom, with smaller values giving heavier tails, and with higher values making the t-distribution resemble a standard normal distribution with a mean of 0 and a standard deviation of 1.



What Is a T-Distribution?

When a sample of n observations is taken from a normally distributed population having mean (M) and standard deviation (D), the **sample mean (m)** and the **sample standard deviation (d)** will differ from **M** and **D** because of the randomness of the sample.

A z-score can be calculated with the population standard deviation as :

$$Z = (x - M)/D$$

The value Z has the normal distribution with mean 0 and standard deviation 1.

But when using the estimated standard deviation, a t-score is calculated as :

$$T = (m - M)/\{d/\sqrt{n}\}$$

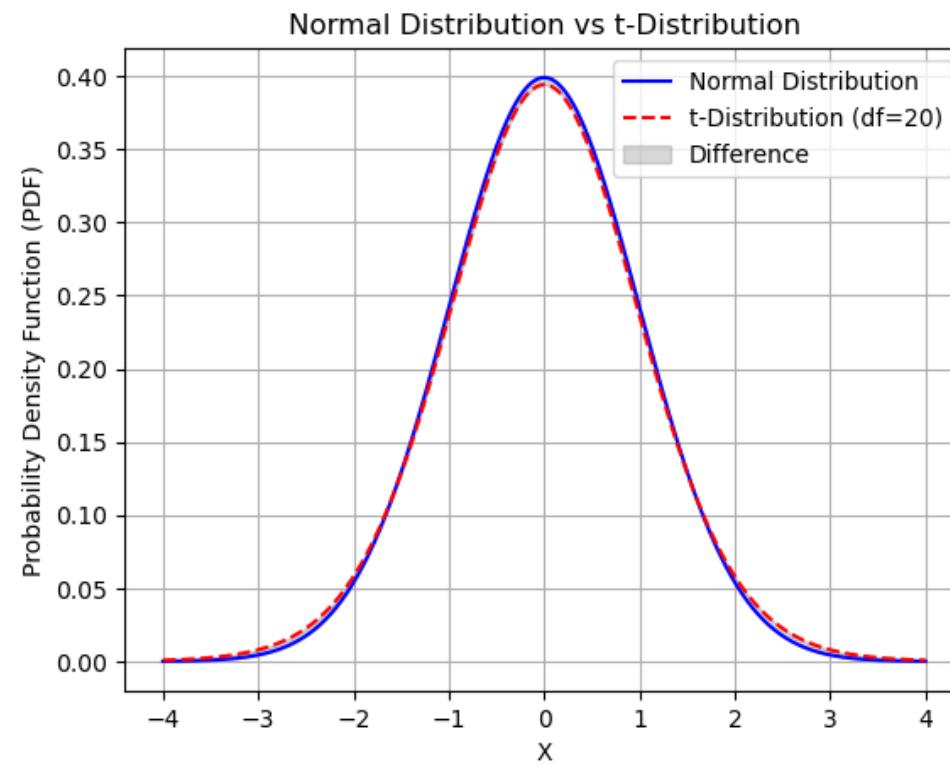
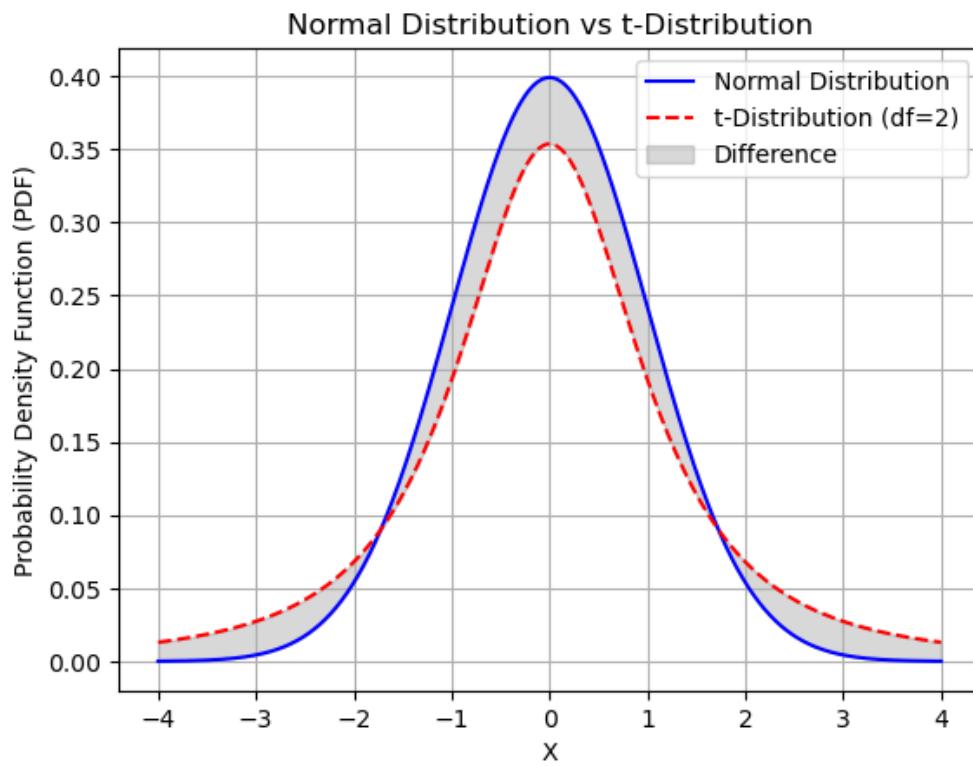
and the difference between d and D makes the distribution a t-distribution with $(n - 1)$ degrees of freedom rather than the normal distribution with mean 0 and standard deviation 1.

T-Distribution vs. Normal Distribution

- Normal distributions are used when the population distribution is assumed to be normal.
- The t-distribution is similar to the normal distribution, just with fatter tails. Both assume a normally distributed population.
- T-distributions thus have higher kurtosis than normal distributions.
- The probability of getting values very far from the mean is larger with a t-distribution than a normal distribution.



T-Distribution vs. Normal Distribution



Important Note: Because the t-distribution has fatter tails than a normal distribution, it can be used as a model for financial returns that exhibit excess kurtosis, which will allow for a more realistic calculation of Value at Risk (VaR) in such cases.



Limitations of Using a T-Distribution

- The t-distribution can skew exactness relative to the normal distribution.
- Its shortcoming only arises when there's a need for perfect normality.
- The t-distribution should only be used when the population standard deviation is not known.
- If the population standard deviation is known and the sample size is large enough, the normal distribution should be used for better results.

When should the t-distribution be used?

The t-distribution should be used if the population sample size is small and the standard deviation is unknown. If not, then the normal distribution should be used.

The Bottom Line

The t-distribution is used in statistics to estimate the significance of population parameters for small sample sizes or unknown variations. Like the normal distribution, it is bell-shaped and symmetric. Unlike normal distributions, it has heavier tails, which result in a greater chance for extreme values.

Note: We will look into the main application of t distribution in the Hypothesis Testing lecture (i.e., T-test).



Chi-squared distribution

- The chi-squared distribution is a continuous probability distribution that arises in statistical inference, particularly in hypothesis testing and confidence interval construction.
- It is used in various statistical tests, such as the chi-squared test for independence and the chi-squared goodness-of-fit test.
- Definition: The chi-squared distribution is a continuous probability distribution of the sum of squared standard normal deviates.
- Symbol: χ^2
- It is widely used in statistical hypothesis testing.



Chi-squared distribution

The probability density function of the chi-squared distribution is given by:

$$f(x; k) = \frac{1}{2^{k/2}\Gamma(k/2)}x^{k/2-1}e^{-x/2}$$

where k is the degrees of freedom , and Γ is the gamma function.



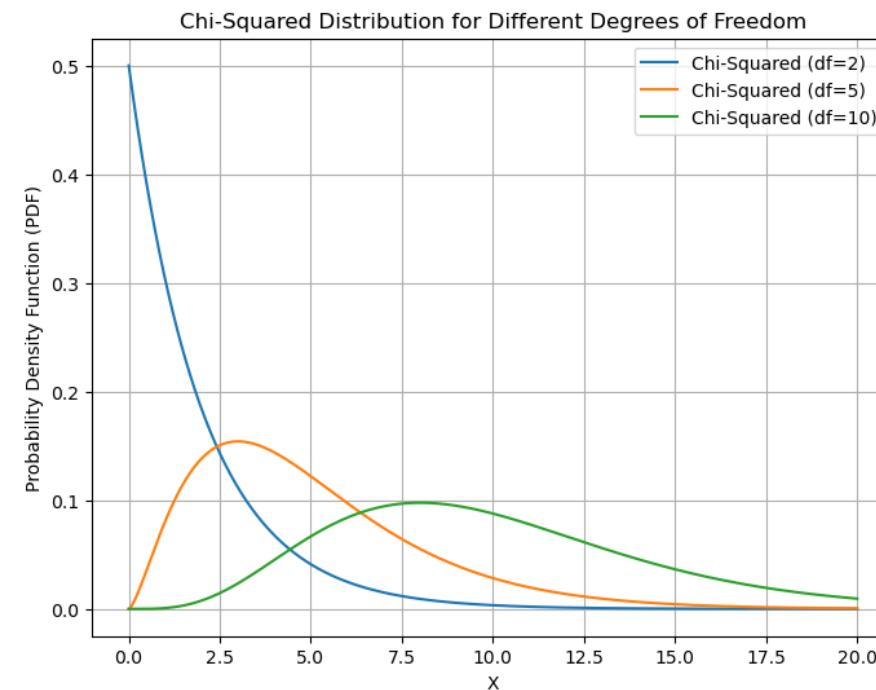
Chi-squared distribution

Degrees of freedom (k) in the chi-squared distribution determine its shape.

As k increases, the distribution becomes more symmetric and approaches normality.

Mean: $E(X) = k$

Variance: $\text{Var}(X) = 2k$



Chi-squared distribution

Chi-Squared Test for Independence

One of the main applications is the chi-squared test for independence.

Used to determine if there is a significant association between two categorical variables.

Chi-Squared Goodness-of-Fit Test

Another application is the chi-squared goodness-of-fit test.

Tests whether an observed frequency distribution matches an expected distribution

Relationship with Normal Distribution

The chi-squared distribution is a special case of the gamma distribution.

As k increases, the chi-squared distribution approaches a normal distribution.

Note: We will look into the main application examples of Chi-Squared distribution in the Hypothesis Testing lecture (in the Chi-Square test).

REFERENCES

- Introduction to Probability and Statistics for Engineers and Scientists, Sixth Edition, Sheldon M. Ross
- Statistical Methods Combined Edition (Volume I& II), N G Das



CONCLUSION

- Discussed the Standard Normal distribution and its Approximation Rule
- Discussed about the T distribution
- Discussed about the Chi-Squared distribution





THANK YOU



JAN 2024

MINE AUTOMATION AND DATA ANALYTICS





MINE AUTOMATION AND DATA ANALYTICS



SWAYAM NPTEL COURSE ON MINE AUTOMATION AND DATA ANALYTICS

By

Prof. Radhakanta Koner

Department of Mining Engineering

Indian Institute of Technology (Indian School of Mines) Dhanbad

Module 9 : Hypothesis Testing



Lecture 21A : Hypothesis Testing - I

CONCEPTS COVERED

- 1) What is Hypothesis Testing?
- 2) Size and Power of a Test.
- 3) Neyman - Pearson Paradigm of Hypothesis Testing
- 4) Types of hypothesis testing
- 5) Motivation example for hypothesis testing

“You can't prove a hypothesis; you can only improve or disprove it.” – Christopher Monckton



What is Hypothesis Testing

- Motivating Example: Is a coin fair or unfair?
- A fair coin is said to have a probability of getting head $P(H) = 0.5$
- An unfair coin is said to have a probability of getting head $P(H) = 0.6$
- Let us suppose you have a coin that could be fair or unfair. You may toss the coin multiple times and observe the results. **How would you test whether the coin is fair or unfair?**

(i) Null Hypothesis (H_0):

The null hypothesis (H_0) is a statement about a population parameter or effect that is assumed to be true unless evidence suggests otherwise.

It represents the status quo or a baseline assumption.

Formally, the null hypothesis is denoted as H_0 and is typically expressed as equality.

(ii) Alternative Hypothesis (H_A):

The alternative hypothesis (H_A) is a statement that contradicts the null hypothesis.

It represents what the researcher is trying to provide evidence for.



What is Hypothesis Testing

- Motivating Example: Is a coin fair or unfair?
- A fair coin is said to have a probability of getting head $P(H) = 0.5$
- An unfair coin is said to have a probability of getting head $P(H) = 0.6$
- Let us suppose you have a coin that could be fair or unfair. You may toss the coin multiple times and observe the results. **How would you test whether the coin is fair or unfair?**

- Hypothesis Testing:
- Using samples, decide between a null hypothesis (H_0) and an alternative hypothesis (H_A)
- Fair Coin Example:
$$H_0 : P(H) = 0.5$$
$$H_A : P(H) = 0.6$$
- One of the most important statistical analysis methods with a wide range of applications.



What is Hypothesis Testing

- In summary, the null hypothesis represents the assumption to be tested, while the alternative hypothesis represents the researcher's claim or the possibility of an effect or difference.
- The goal of hypothesis testing is to gather evidence from sample data to decide whether to reject the null hypothesis in favor of the alternative hypothesis.



Accepting or Rejecting the Null Hypothesis

➤ Motivating Example: Is a coin fair or unfair?

- Suppose we toss the coin 3 times.
- Possible outcomes are HHH, HHT, . . . , TTT
- For some outcomes, we will accept H_0 and others, we will reject H_0
- Let A be the set of all outcomes for which we accept H_0
- Every acceptance subset A corresponds to a test



Size and Power of a Test

- Metric 1: Significance level (also called size) of a test, denoted α .
- Type I Error: Reject H_0 when H_0 is true
- **Size of a test = $\alpha = P(\text{ Type I error}) = P(\text{Reject } H_0 \mid H_0 \text{ is true})$**

- Metric 2: Power of a test, $1 - \beta$
- Type II error: Accept H_0 when H_A is true
- $\beta = P(\text{Type II error}) = P(\text{Accept } H_0 \mid H_A \text{ is true})$
- **Power = $1 - \beta = P(\text{ Reject } H_0 \mid H_A \text{ is true})$**



Computing the Size and Power for Unfair Coin Example

$$H_0 : P(H) = 0.5$$

$$H_A : P(H) = 0.6$$

- Toss 3 times = { HHH , HHT, HTH, THH, THT , HTT , TTH , TTT}
- If acceptance subset A = \emptyset
 - Always reject H_0
 - $\alpha = 1$, $\beta = 0$
- If acceptance subset A = { HHH, HHT, HTH, THH, THT, HTT, TTH, TTT}
 - Always accept H_0
 - $\alpha = 0$, $\beta = 1$
- If acceptance subset A = { HHT, HTH, THH, THT, HTT, TTH }
 - $\alpha = P(A^C | P(H) = 0.5) = 2/8 = 0.25$
 - $\beta = P(A | P(H) = 0.6) = 3(0.4)^2(0.6) + 3(0.4)(0.6)^2 = 0.72$
- The value α , called the level of significance of the test, is usually set in advance, with commonly chosen values being $\alpha = .1, .05, .005$.



Neyman-Pearson Paradigm of Hypothesis Testing

$X_1, X_2, X_3, \dots, X_n \sim \text{iid } X$

- H_0 : Null Hypothesis on the distribution of X , H_A : Alternative Hypothesis
- Test: Defined by an acceptance set A
- If samples fall in A , accept H_0 ; otherwise, reject H_0
- Two Errors:
 - Type I Error: Reject H_0 when H_0 is true
 - Type II error: Accept H_0 when H_A is true
- Two Metrics
 - Size of a test = $\alpha = P(\text{Type I error}) = P(\text{Reject } H_0 | H_0 \text{ is true})$
 - Power = $1 - \beta = P(\text{Reject } H_0 | H_A \text{ is true})$



Types of Hypothesis Testing

Simple Hypothesis:

- A hypothesis that completely specifies the distribution of the samples is called a **simple hypothesis**.
- Example:
 - 1) Coin Toss ; $P(\text{Heads}) = 0.5$, $P(\text{Tails}) = 0.5$
 - 3) Normal (μ , σ^2) samples ; $\mu = 1$, $\mu = -1$ etc.,
- **Simple null vs simple alternative**



Types of Hypothesis Testing

Composite Hypothesis

- A hypothesis that does not specify the distribution of the samples is called a **Composite hypothesis**.

Example 1: Coin Toss ;

- Null: $P(\text{Heads}) = 0.5$ (coin is fair), simple
- Alternative: $P(\text{Heads}) \neq 0.5$ (coin is unfair), composite

Example 2: Normal ($\mu, 3$) samples ;

- Null: $\mu = 0$ (some effect is not present), simple
- Alternative: $\mu > 1$ (the effect is present), composite
- **Simple / Composite Null vs Composite Alternative**



Types of Hypothesis Testing

Standard Tests: One Sample

$X_1, X_2, X_3, \dots, X_n \sim \text{iid } X$

$$E(X) = \mu ; \text{Var}(X) = \sigma^2$$

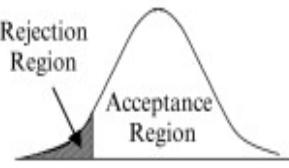
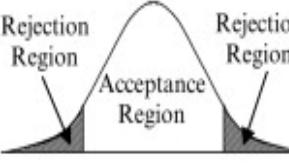
- Testing for mean,

Null $H_0 : \mu = c$

Alternative :

- Right tail test, $H_A : \mu > c$
- Left tail test, $H_A : \mu < c$
- Two tail test, $H_A : \mu \neq c$

Two Cases: known or unknown variance

One-Tailed Test (Left Tail)	Two-Tailed Test	One-Tailed Test (Right Tail)
$H_0 : \mu_X = \mu_0$ $H_1 : \mu_X < \mu_0$	$H_0 : \mu_X = \mu_0$ $H_1 : \mu_X \neq \mu_0$	$H_0 : \mu_X = \mu_0$ $H_1 : \mu_X > \mu_0$
		



Z-score values for Rejection Regions

99% Confidence level (i.e alpha = 0.01):

Left-tailed test: $z = -2.33$

Two-tailed test: $z = \pm 2.55$

(the critical z-values are $+2.55$ and -2.55)

Right-tailed test: $z = +2.33$

95% Confidence level (i.e alpha = 0.05):

Left-tailed test: $z = -1.65$

Two-tailed test: $z = \pm 1.96$

(the critical z-values are -1.96 and 1.96)

Right-tailed test: $z = +1.65$

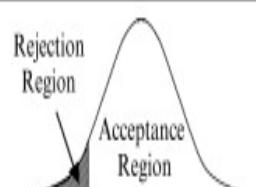
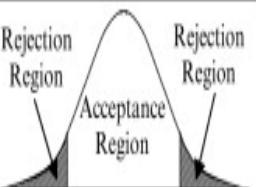
90% Confidence level (i.e alpha = 0.1):

Left-tailed test: $z = -1.2$

Two-tailed test: $z = \pm 1.65$

(the critical z-values are -1.65 and 1.65)

Right-tailed test: $z = +1.2$

One-Tailed Test (Left Tail)	Two-Tailed Test	One-Tailed Test (Right Tail)
$H_0: \mu_X = \mu_0$ $H_1: \mu_X < \mu_0$	$H_0: \mu_X = \mu_0$ $H_1: \mu_X \neq \mu_0$	$H_0: \mu_X = \mu_0$ $H_1: \mu_X > \mu_0$
		

Standard Tests: One Sample

$X_1, X_2, X_3, \dots, X_n \sim \text{iid } X$

$$E(X) = \mu ; \text{Var}(X) = \sigma^2$$

- Testing for variance,

Null $H_0 : \sigma = c$

Alternative :

- Right tail test, $H_A: \sigma > c$
- Left tail test, $H_A: \sigma < c$
- Two-tail test, $H_A: \sigma \neq c$



Standard Tests: Two Sample

$$X_1, X_2, X_3, \dots, X_n \sim \text{iid } X$$

$$Y_1, Y_2, Y_3, \dots, Y_n \sim \text{iid } Y$$

$$E(X) = \mu_1 ; \text{Var}(X) = \sigma_1^2$$

$$E(Y) = \mu_2 ; \text{Var}(Y) = \sigma_2^2$$

- Testing to compare means

Null $H_0 : \mu_1 = \mu_2$

Alternative: $H_A : \mu_1 \neq \mu_2$

- Testing to compare variances

Null $H_0 : \sigma_1 = \sigma_2$

Alternative: $H_A : \sigma_1 \neq \sigma_2$



Goodness of fit testing

Samples: $X_1, X_2, X_3, \dots, X_n$

- Problem: Do the samples follow a certain distribution?
- Examples :
- Integer Samples $X_i \in \{ 0, 1, 2, \dots \}$. Is the distribution Poisson?
- Continuous Samples $X_i \in (-\infty, \infty)$. Is the distribution normal?



Observations

- In all examples, the questions seem to be reasonably posed in a statistical hypothesis testing framework.
- In most cases, the null and/or alternative are composite
- In all cases, the **confidence** of the testing is very important.

How do you quantify confidence?

- With the help of an alpha value concept (or)
- A notion called **P - value** is used to quantify confidence.
- The P-value is known as the probability value.
- It is defined as the probability of getting a result that is either the same or more extreme than the actual observations.
- The P-value is known as the level of marginal significance within the hypothesis testing that represents the probability of occurrence of the given event.
- The P-value is used as an alternative to the rejection point to provide the least significance at which the null hypothesis would be rejected.
- If the P-value is small, then there is stronger evidence in favor of the alternative hypothesis.



Covering concepts (through an example)

Suppose that a construction firm has just purchased a large supply of cables that have been guaranteed to have an average breaking strength of at least 7,000 psi.

To verify this claim, the firm has decided to take a random sample of 10 of these cables to determine their breaking strengths. They will then use the result of this experiment to ascertain whether or not they accept the cable manufacturer's hypothesis that the population mean is at least 7,000 pounds per square inch.

A statistical hypothesis is usually a statement about a set of parameters of a population distribution. It is called a hypothesis because it is not known whether or not it is true.

A primary problem is to develop a procedure for determining whether or not the values of a random sample from this population are consistent with the hypothesis

For instance,
consider a particular normally distributed population having an unknown mean value θ and known variance 1.

The statement “ θ is less than 1” is a statistical hypothesis that we could try to test by observing a random sample from this population.

If the random sample is deemed to be consistent with the hypothesis under consideration, we say that the hypothesis has been “**accepted**”; otherwise, we say that it has been “**rejected**”

Important Note: In accepting a given hypothesis, we are not actually claiming that it is true but rather we are saying that the resulting data appear to be consistent with it.

For instance,
in the case of a normal $(\theta, 1)$ population,

If a resulting sample of size 10 has an average value of 1.25, then although such a result cannot be regarded as being evidence in favor of the hypothesis " $\theta < 1$," it is not inconsistent with this hypothesis, which would thus be accepted.

On the other hand, if the sample of size 10 has an average value of 3, then even though a sample value that large is possible when $\theta < 1$, it is so unlikely that it seems inconsistent with this hypothesis, which would thus be rejected.



Significance Levels

Consider a population having distribution F_θ , where θ is unknown, and suppose we want to test a specific hypothesis about θ .

We shall denote this hypothesis by H_0 and call it the null hypothesis. For example, if F_θ is a normal distribution function with mean θ and variance equal to 1, then two possible null hypotheses about θ are

- (a) $H_0 : \theta = 1$
- (b) $H_0 : \theta \leq 1$

Thus, the first of these hypotheses states that the population is normal with mean 1 and variance 1, whereas the second states that it is normal with variance 1 and a mean less than or equal to 1.

Note: the null hypothesis in (a), when true, completely specifies the population distribution, whereas the null hypothesis in (b) does not. A hypothesis that, when true, completely specifies the population distribution is called a *simple hypothesis*; one that does not is called a *composite hypothesis*.



Suppose now that in order to test a specific null hypothesis H_0 , a population sample of size n — say X_1, \dots, X_n — is to be observed. Based on these n values, we must decide whether or not to accept H_0 .

A test for H_0 can be specified by defining a region C in n -dimensional space with the proviso that the hypothesis is to be rejected if the random sample X_1, \dots, X_n turns out to lie in C and accepted otherwise. The region C is called the **critical region**. In other words, the statistical test determined by the critical region C is the one that

$$\begin{array}{ll} \text{accepts } H_0 & \text{if } (X_1, X_2, \dots, X_n) \notin C \\ \text{and} & \\ \text{rejects } H_0 & \text{if } (X_1, X_2, \dots, X_n) \in C \end{array}$$

For instance, a common test (**will cover in next lecture**) of the hypothesis that θ , the mean of a normal population with variance 1, is equal to 1 has a critical region given by

$$C = \{(X_1, \dots, X_n) : |\bar{X} - 1| > 1.96 / \sqrt{n}\}$$

Thus, this test calls for rejection of the null hypothesis that $\theta = 1$ when the sample average differs from 1 by more than 1.96 divided by the square root of the sample size.



Important Note:

When developing a procedure for testing a given null hypothesis H_0 that, in any test, two different types of errors can result.

1. The first of these, called a type I error, is said to result if the test incorrectly calls for rejecting H_0 when it is indeed correct.
2. The second, called a type II error, results if the test calls for accepting H_0 when it is false.

The objective of a statistical test of H_0 is not to explicitly determine whether or not H_0 is true but rather to determine if its validity is consistent with the resultant data.

Hence, with this objective, it seems reasonable that H_0 should only be rejected if the resultant data are very unlikely when H_0 is true.

The classical way of accomplishing this is

- (i) First, specify a value α and
- (ii) then require the test to have the property that whenever H_0 is true its probability of being rejected is never greater than α .

The value α , called the level of significance of the test, is usually set in advance, with commonly chosen values being $\alpha = .1, .05, .005$.

In other words,

the classical approach to testing H_0 is to fix a significance level α and then require that the test have the property that the probability of a type I error occurring can never be greater than α .



REFERENCES

- Introduction to Probability and Statistics for Engineers and Scientists, Sixth Edition, Sheldon M. Ross
- Statistical Methods Combined Edition (Volume I& II), N G Das



CONCLUSION

- **Defined the Hypothesis Testing with examples.**
- **Discussed how to compute the Size and Power of a Test.**
- **Discussed the Neyman-Pearson Paradigm of Hypothesis Testing.**
- **Types of hypothesis testing**
 - a. **Standard test – one sample**
 - b. **Standard tests – two sample**
 - c. **Goodness of fit testing**
- **Motivation example for hypothesis testing with significance level (alpha).**





THANK YOU



JAN 2024

MINE AUTOMATION AND DATA ANALYTICS





MINE AUTOMATION AND DATA ANALYTICS



SWAYAM NPTEL COURSE ON MINE AUTOMATION AND DATA ANALYTICS

By

Prof. Radhakanta Koner

Department of Mining Engineering

Indian Institute of Technology (Indian School of Mines) Dhanbad

Module 9 : Hypothesis Testing



Lecture 21B : Hypothesis Testing - II

CONCEPTS COVERED

- 1) Z test: Tests concerning the mean of a normal population
- 2) Solved Example using Z test: 1
- 3) Solved Example using Z test: 2
- 4) Solved Example using Z test: 3
- 5) Solved Example using Z test : 4



Tests concerning the mean of a normal population

Z test

1. Case of Known Variance

Suppose that X_1, \dots, X_n is a sample of size n from a normal distribution having an unknown mean μ and a known variance σ^2 and suppose we are interested in testing the null hypothesis

$$H_0: \mu = \mu_0$$

against the alternative hypothesis

$$H_1: \mu \neq \mu_0$$

Since $\bar{X} = \sum_{i=1}^n \frac{X_i}{n}$ is a natural point estimator of μ , it seems reasonable to accept H_0 if \bar{X} is not too far from μ_0 . That is, the critical region of the test would be of the form

$$C = \{X_1, \dots, X_n : |\bar{X} - \mu_0| > c\}$$

for some suitably chosen value c



If we desire that the test has significance level α , then we must determine the critical value c that will make the type I error equal to α . That is, c must be such that

$$P_{\mu_0}\{|\bar{X} - \mu_0| > c\} = \alpha$$

where we write P_{μ_0} to mean that the preceding probability is to be computed under the assumption that $\mu = \mu_0$. However, when $\mu = \mu_0$, X will be normally distributed with mean μ_0 and variance σ^2/n and so Z , defined by

$$Z \equiv \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} \quad \text{will have a standard normal distribution}$$

$$P\left\{|Z| > \frac{c\sqrt{n}}{\sigma}\right\} = \alpha$$

$$2P\left\{Z > \frac{c\sqrt{n}}{\sigma}\right\} = \alpha$$



where Z is a standard normal random variable. However, we know that

$$P\{Z > z_{\alpha/2}\} = \alpha/2$$

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

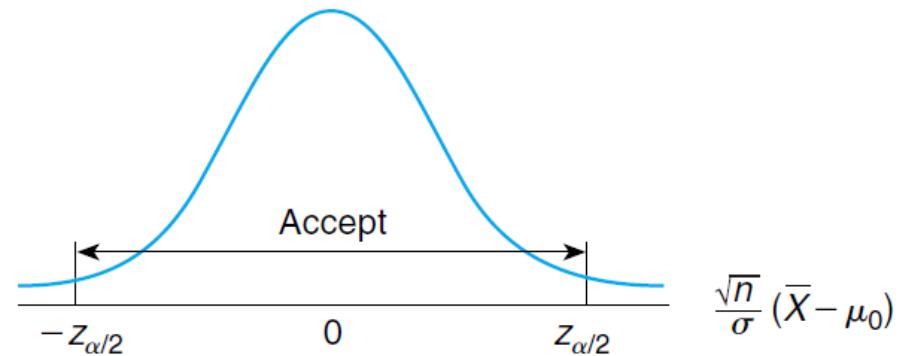
$$c = \frac{z_{\alpha/2}\sigma}{\sqrt{n}}$$

Thus, the significance level α test is to reject H_0 if $|\bar{X} - \mu_0| > z_{\alpha/2}\sigma/\sqrt{n}$ and accept otherwise; or, equivalently, to

reject H_0 if $\frac{\sqrt{n}}{\sigma}|\bar{X} - \mu_0| > z_{\alpha/2}$

accept H_0 if $\frac{\sqrt{n}}{\sigma}|\bar{X} - \mu_0| \leq z_{\alpha/2}$





we have superimposed the standard normal density function [which is the density of the test statistic $\frac{\sqrt{n}}{\sigma} (\bar{X} - \mu_0)$ when H_0 is true].

Determining Critical Values ($Z\alpha/2$)

What is the critical value ($Z\alpha/2$) for a 95% confidence level (for $\alpha = 0.05$) , assuming a two-tailed test?

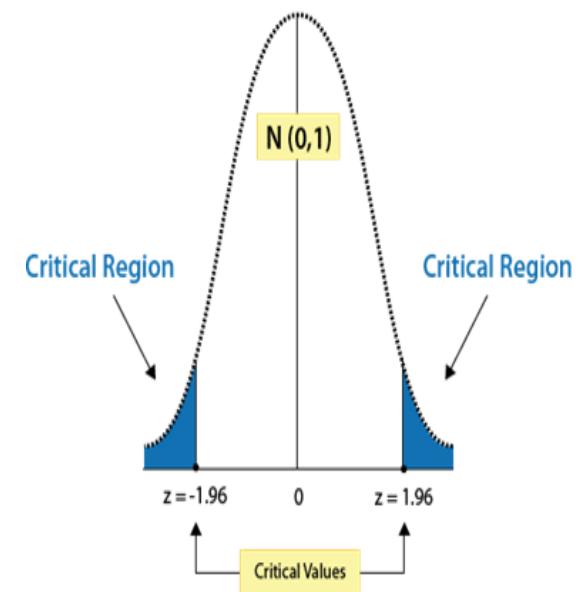
A 95% confidence level means that a total of 5% of the area under the curve is considered the critical region.

Since this is a two-tailed test, $\frac{1}{2} (5\%) = 2.5\%$ of the values would be in the left tail, and the other 2.5% would be in the right tail.

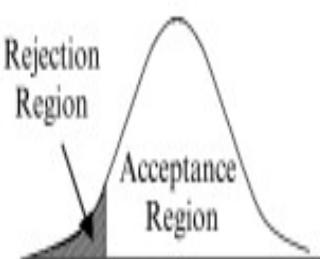
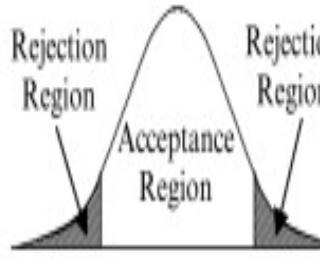
Looking up the Z-score associated with 0.025 on a reference table, we find 1.96. Therefore, +1.96 is the critical value of the right tail, and -1.96 is the critical value of the left tail.

The critical value for a 95% confidence level is $Z\alpha/2 = \pm 1.96$

Critical Regions for a Two-Tailed z Test



Rejection regions for different tailed Z test

One-Tailed Test (Left Tail)	Two-Tailed Test	One-Tailed Test (Right Tail)
$H_0 : \mu_X = \mu_0$ $H_1 : \mu_X < \mu_0$	$H_0 : \mu_X = \mu_0$ $H_1 : \mu_X \neq \mu_0$	$H_0 : \mu_X = \mu_0$ $H_1 : \mu_X > \mu_0$
		



Z-score values for common confidence levels of a normal distribution

99% Confidence level (i.e alpha = 0.01):

Left-tailed test: $Z\alpha = -2.33$

Two-tailed test: $Z\alpha/2 = +/- 2.55$ (the critical z-values are +2.55 and -2.55)

Right-tailed test: $Z\alpha = +2.33$

95% Confidence level (i.e alpha = 0.05):

Left-tailed test: $Z\alpha = -1.65$

Two-tailed test: $Z\alpha/2 = +/- 1.96$ (the critical z-values are -1.96 and 1.96)

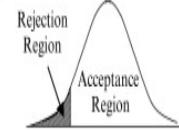
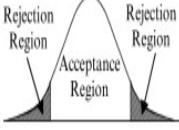
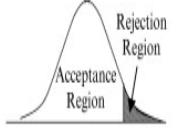
Right-tailed test: $Z\alpha = +1.65$

90% Confidence level (i.e alpha = 0.1):

Left-tailed test: $Z\alpha = -1.2$

Two-tailed test: $Z\alpha/2 = +/- 1.65$ (the critical z-values are -1.65 and 1.65)

Right-tailed test: $Z\alpha = +1.2$

One-Tailed Test (Left Tail)	Two-Tailed Test	One-Tailed Test (Right Tail)
$H_0 : \mu_x = \mu_0$ $H_1 : \mu_x < \mu_0$	$H_0 : \mu_x = \mu_0$ $H_1 : \mu_x \neq \mu_0$	$H_0 : \mu_x = \mu_0$ $H_1 : \mu_x > \mu_0$
		



If a signal of value μ is sent from location A, then the value received at location B is normally distributed with mean μ and standard deviation 2. The random noise added to the signal is an $N(0, 4)$ random variable. There is reason for the people at location B to suspect that the signal value $\mu = 8$ will be sent today. Test this hypothesis if the same signal value is independently sent five times and the average value received at location B is $\bar{X} = 9.5$.

$$\frac{\sqrt{n}}{\sigma} |\bar{X} - \mu_0| = \frac{\sqrt{5}}{2} (1.5) = 1.68$$

Since this value is less than $z_{.025} = 1.96$, the hypothesis is accepted. In other words, the data are not inconsistent with the null hypothesis in the sense that a sample average as far from the value 8 as observed would be expected, when the true mean is 8, over 5 percent of the time.

Note: however, that if a less stringent significance level were chosen — say $\alpha = 0.1$ then the null hypothesis would have been rejected. This follows since $z_{.05} = 1.645$, which is less than 1.68.

Hence, if we had chosen a test that had a 10 percent chance of rejecting H_0 when H_0 was true, then the null hypothesis would have been rejected.



The “correct” level of significance to use in a given situation depends on the individual circumstances involved in that situation.

For instance, if rejecting a null hypothesis H_0 would result in large costs that would thus be lost if H_0 were indeed true, then we might elect to be quite conservative and so choose a significance level of .05 or .01.

Also, if we initially feel strongly that H_0 was correct, then we would require very stringent data evidence to the contrary for us to reject H_0 . (That is, we would set a very low significance level in this situation).



Example (1/4) of Z test

Suppose a manufacturer claims that the mean weight of their product is 500 grams. To test this claim, a random sample of 36 products is selected, and their weights are recorded. The sample mean weight is found to be 495 grams, with a sample standard deviation of 10 grams. Assume the weights follow a normal distribution. Using a significance level of 0.05, test the manufacturer's claim.

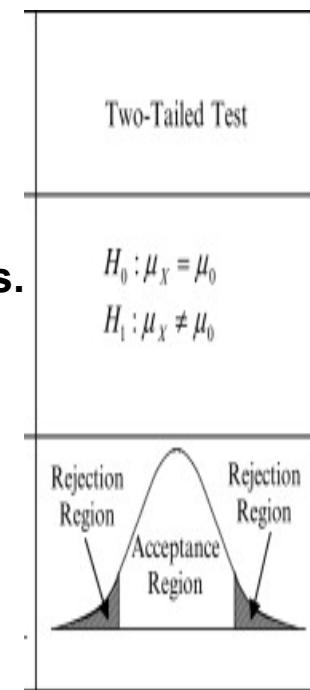
1) State the hypothesis:

Null Hypothesis (H_0): The mean weight of the product is 500 grams.

Alternative Hypothesis (H_1): The mean weight of the product is not 500 grams.

$$H_0: \mu = 500$$

$$H_1: \mu \neq 500$$



2) Determine the significance level (α):

$$\alpha = 0.05$$

3) Calculate the test statistic (z-score):
$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

The formula for the z-test statistic for the mean is:

Where:

\bar{x} is the sample mean

μ is the population mean under the null hypothesis

σ is the population standard deviation

n is the sample size



Given:

$\bar{x} = 495$ grams

$\mu = 500$ grams

$\sigma = 10$ grams

$n = 36$

$$z = \frac{495 - 500}{\frac{10}{\sqrt{36}}} = \frac{-5}{\frac{10}{6}} = -3$$



Z-score values for common confidence levels of a normal distribution

99% Confidence level (i.e alpha = 0.01):

Left-tailed test: $Z\alpha = -2.33$

Two-tailed test: $Z\alpha/2 = \pm 2.55$ (the critical z-values are +2.55 and -2.55)

Right-tailed test: $Z\alpha = +2.33$

95% Confidence level (i.e alpha = 0.05):

Left-tailed test: $Z\alpha = -1.65$

Two-tailed test: $Z\alpha/2 = \pm 1.96$ (the critical z-values are -1.96 and 1.96)

Right-tailed test: $Z\alpha = +1.65$

90% Confidence level (i.e alpha = 0.1):

Left-tailed test: $Z\alpha = -1.2$

Two-tailed test: $Z\alpha/2 = \pm 1.65$ (the critical z-values are -1.65 and 1.65)

Right-tailed test: $Z\alpha = +1.2$



4) Determine the critical value:

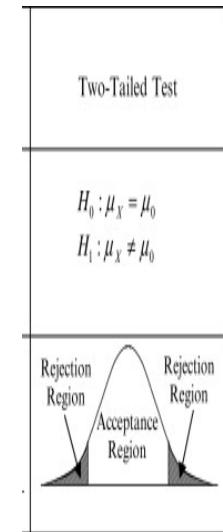
Since it's a two-tailed test, the critical z-values are -1.96 and 1.96 at a significance level of 0.05.

5) Decision rule:

If the absolute value of the z-score is greater than 1.96, we reject the null hypothesis.

6) Make a decision:

The calculated z-value (-3) falls in the rejection region (less than -1.96), so we reject the null hypothesis.



7) Conclusion:

Since we reject the null hypothesis, we have sufficient evidence to conclude that the mean weight of the product is not 500 grams.

Therefore, based on the sample data, there is enough evidence to suggest that the manufacturer's claim is not correct at the 0.05 significance level.

Example (2/4) of Z test

Suppose a manufacturer claims that the average lifespan of their light bulbs is at least 1000 hours. You believe that the average lifespan is actually less than that. To test this claim, you collect a sample of 50 light bulbs and find that the average lifespan is 980 hours with a standard deviation of 40 hours. You want to test whether the average lifespan is significantly less than 1000 hours at a 5% significance level.

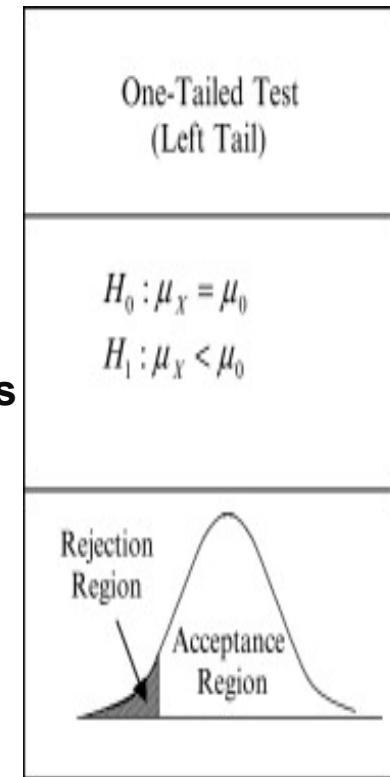
State the hypothesis:

Null Hypothesis (H₀): The average lifespan of the light bulbs is at least 1000 hours.

Alternative Hypothesis (H₁): The average lifespan of the light bulbs is less than 1000 hours.

$$H_0: \mu = 1000$$

$$H_1: \mu < 1000$$



Given:

1. Sample mean (\bar{x}) = 980 hours
2. Population mean (μ) = 1000 hours
3. Sample standard deviation (σ) = 40 hours
4. Sample size (n) = 50
5. Significance level (α) = 0.05

Calculate the test statistic (z-score):

$$Z = \frac{980 - 1000}{\frac{40}{\sqrt{50}}}$$

$$Z = \frac{-20}{5.6568}$$

$$Z \approx -3.54$$



Z-score values for common confidence levels of a normal distribution

99% Confidence level (i.e alpha = 0.01):

Left-tailed test: $Z\alpha = -2.33$

Two-tailed test: $Z\alpha/2 = \pm 2.55$ (the critical z-values are +2.55 and -2.55)

Right-tailed test: $Z\alpha = +2.33$

95% Confidence level (i.e alpha = 0.05):

Left-tailed test: $Z\alpha = -1.65$

Two-tailed test: $Z\alpha/2 = \pm 1.96$ (the critical z-values are -1.96 and 1.96)

Right-tailed test: $Z\alpha = +1.65$

90% Confidence level (i.e alpha = 0.1):

Left-tailed test: $Z\alpha = -1.2$

Two-tailed test: $Z\alpha/2 = \pm 1.65$ (the critical z-values are -1.65 and 1.65)

Right-tailed test: $Z\alpha = +1.2$



Determine the critical value:

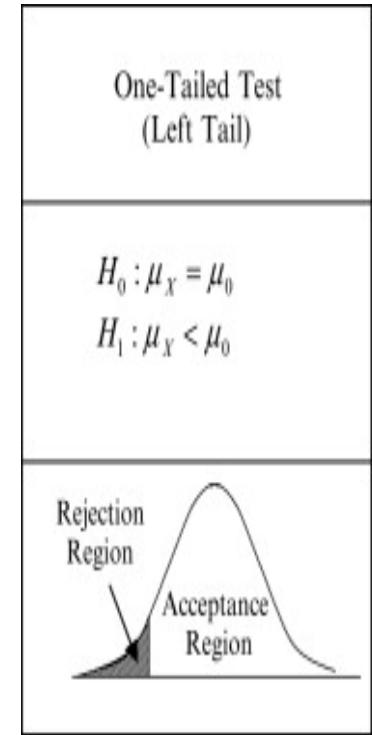
Since this is a left-tailed test and the significance level is 0.05, we find the critical z-value from the standard normal distribution table. At $\alpha = 0.05$, the critical value is approximately -1.645..

Make a decision:

Since the calculated z-value (-3.54) is less than -1.645 (the critical value for a 5% significance level), we reject the null hypothesis.

Conclusion:

There is enough evidence to suggest that the average lifespan of the light bulbs is significantly less than 1000 hours.



Example (3/4) of Z test

Suppose a company claims that the average response time for their customer service hotline is no more than 3 minutes. You believe that the average response time is actually longer than that. To test this claim, you collect a sample of 40 calls to the hotline and find that the average response time is 3.5 minutes with a standard deviation of 0.8 minutes. You want to test whether the average response time is significantly greater than 3 minutes at a 5% significance level.

State the hypothesis:

Null Hypothesis (H_0):

The average response time for the customer service hotline is no more than 3 minutes..

Alternative Hypothesis (H_1):

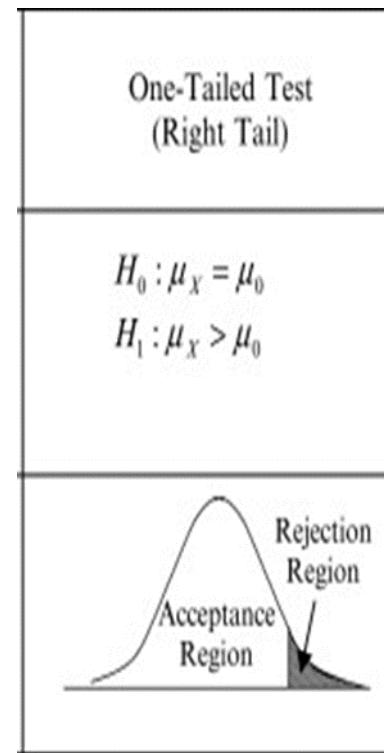
The average response time for the customer service hotline is greater than 3 minutes..

$$H_0: \mu = 3$$

$$H_1: \mu > 3$$

Set Significance Level:

Let's choose a significance level (α) of 0.05.



Calculate the test statistic (z-score):

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

Where:

- \bar{x} is the sample mean (3.5 minutes).
- μ is the population mean under the null hypothesis (3 minutes).
- σ is the population standard deviation (0.8 minutes).
- n is the sample size (40).

$$Z = \frac{3.5 - 3}{\frac{0.8}{\sqrt{40}}}$$

$$Z = \frac{0.5}{\frac{0.8}{\sqrt{40}}}$$

$$Z \approx \frac{0.5}{0.1265}$$

$$Z \approx 3.95$$



Z-score values for common confidence levels of a normal distribution

99% Confidence level (i.e alpha = 0.01):

Left-tailed test: $Z\alpha = -2.33$

Two-tailed test: $Z\alpha/2 = \pm 2.55$ (the critical z-values are +2.55 and -2.55)

Right-tailed test: $Z\alpha = +2.33$

95% Confidence level (i.e alpha = 0.05):

Left-tailed test: $Z\alpha = -1.65$

Two-tailed test: $Z\alpha/2 = \pm 1.96$ (the critical z-values are -1.96 and 1.96)

Right-tailed test: $Z\alpha = +1.65$

90% Confidence level (i.e alpha = 0.1):

Left-tailed test: $Z\alpha = -1.2$

Two-tailed test: $Z\alpha/2 = \pm 1.65$ (the critical z-values are -1.65 and 1.65)

Right-tailed test: $Z\alpha = +1.2$



Determine the critical value:

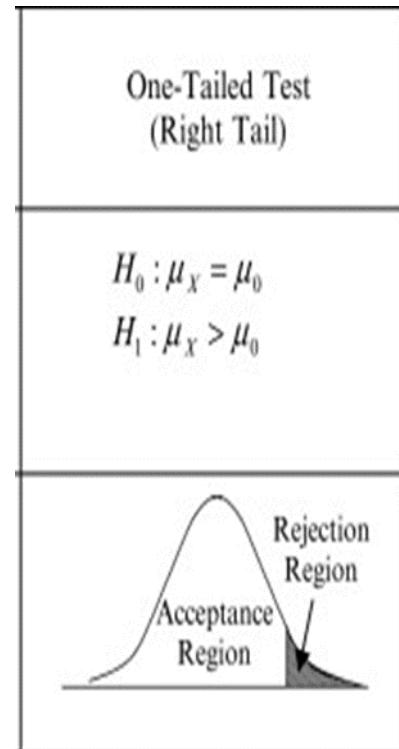
Since this is a right-tailed test and the significance level is 0.05, we find the critical z-value from the standard normal distribution table. At $\alpha = 0.05$, the critical value is approximately 1.645.

Make a decision:

Since the calculated z-value (3.95) is greater than 1.645 (the critical value for a 5% significance level), we reject the null hypothesis.

Conclusion:

There is enough evidence to suggest that the average response time for the customer service hotline is significantly greater than 3 minutes.



Example (4/4) of Z test

An educational institute claims that the average score of its students on a standardized test is 75. A random sample of 50 students is selected, and their scores are recorded. The sample mean score is found to be 72, with a sample standard deviation of 8. Test the institute's claim at a significance level of 0.01.

State the hypothesis:

Null Hypothesis (H_0): The average score of the institute's students is 75.

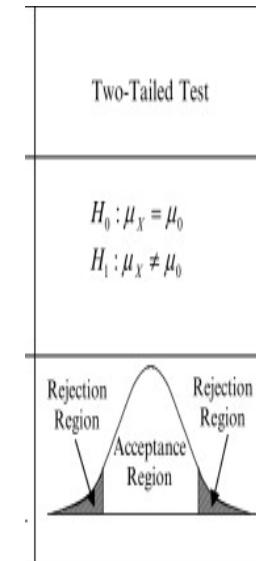
Alternative Hypothesis (H_1): The average score of the institute's students is not 75.

$$H_0: \mu = 75$$

$$H_1: \mu \neq 75$$

Given:

1. Sample mean (\bar{x}) = 72
2. Population mean (μ) = 75
3. Sample standard deviation (σ) = 8
4. Sample size (n) = 50
5. Significance level (α) = 0.01



Calculate the test statistic (z-score):

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

$$z = \frac{72 - 75}{\frac{8}{\sqrt{50}}} = \frac{-3}{1.131} \approx -2.65$$



Z-score values for common confidence levels of a normal distribution

99% Confidence level (i.e alpha = 0.01):

Left-tailed test: $Z\alpha = -2.33$

Two-tailed test: $Z\alpha/2 = \pm 2.55$ (the critical z-values are +2.55 and -2.55)

Right-tailed test: $Z\alpha = +2.33$

95% Confidence level (i.e alpha = 0.05):

Left-tailed test: $Z\alpha = -1.65$

Two-tailed test: $Z\alpha/2 = \pm 1.96$ (the critical z-values are -1.96 and 1.96)

Right-tailed test: $Z\alpha = +1.65$

90% Confidence level (i.e alpha = 0.1):

Left-tailed test: $Z\alpha = -1.2$

Two-tailed test: $Z\alpha/2 = \pm 1.65$ (the critical z-values are -1.65 and 1.65)

Right-tailed test: $Z\alpha = +1.2$



Determine the critical value:

Since it's a two-tailed test at $\alpha = 0.01$, the critical z-values are ± 2.58 (rounded from z-table).

Decision rule:

If the absolute value of the z-score is greater than 2.58, we reject the null hypothesis.

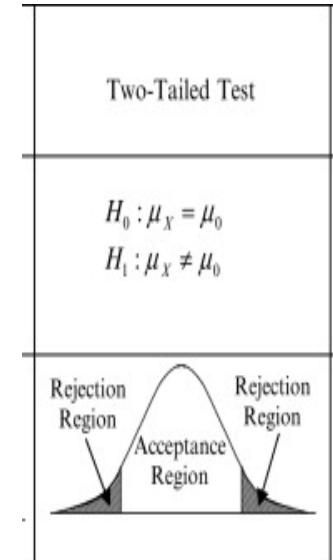
Make a decision:

The absolute value of the calculated z-value (2.65) falls in the rejection region (greater than 2.58), so we reject the null hypothesis.

Conclusion:

Since we reject the null hypothesis, we have sufficient evidence to conclude that the average score of the institute's students is not 75 at a significance level of 0.01.

Therefore, based on the sample data, there is enough evidence to suggest that the institute's claim of the average score being 75 is not supported.



REFERENCES

- Introduction to Probability and Statistics for Engineers and Scientists, Sixth Edition, Sheldon M. Ross
- Statistical Methods Combined Edition (Volume I& II), N G Das



CONCLUSION

- We have discussed the Z test hypothesis testing in detail, along with four different examples.





THANK YOU



JAN 2024

MINE AUTOMATION AND DATA ANALYTICS





MINE AUTOMATION AND DATA ANALYTICS



SWAYAM NPTEL COURSE ON MINE AUTOMATION AND DATA ANALYTICS

By

Prof. Radhakanta Koner

Department of Mining Engineering

Indian Institute of Technology (Indian School of Mines) Dhanbad



Module 9 : Hypothesis Testing

Lecture 22A : t - test

CONCEPTS COVERED

- 1) Introduction to t-test (for Population unknown variance)
 - Motivation Example about t-test
 - Assumptions of t-test
- 2) One Sample t-test
- 3) Two Sample t-test.
 - (i) Assuming Populations with equal Variance Case
 - (ii) Assuming Populations with unequal Variance Case



Overview of t test: Mathematically

$X_1, X_2, X_3, \dots, X_n \sim \text{iid Normal} (\mu, \sigma^2)$

σ^2 is unknown

$$E(X) = \mu ; \text{Var}(X) = \sigma^2$$

Testing for mean,

Null hypothesis (H_0) : $\mu = \mu_0$,

Alternative hypothesis (H_A) : $\mu > \mu_0$



t-test

- T-test is a statistical method used to determine if there is a significant difference between the means of two groups or between the mean of a sample and a known value.
- It helps you assess whether any observed differences between the groups are likely to have occurred by chance or if they are statistically significant.
- The t-test is based on the t-distribution, which is a mathematical distribution similar to the normal distribution but with heavier tails.
- The test calculates a t-statistic, which measures the difference between the means of the two groups in terms of the standard error of the difference.
- The larger the t-statistic, the more likely it is that the difference between the groups' means is not due to random chance.



When Should We Perform a t-test?

Let's first understand where a t-test can be used before we dive into its different types and their implementations. The best way to learn a concept is by visualizing it through an example. So, let's take a simple example to see where a t-test comes in handy.

Consider a telecom company that has two service centers in the city. The company wants to find out whether the average time required to service a customer is the same in both stores.

The company measures the average time taken by 50 random customers in each store. Store A takes 22 minutes, while Store B averages 25 minutes. Can we say that Store A is more efficient than Store B in terms of customer service?

It does seem that way, doesn't it? However, we have only looked at 50 random customers out of the many people who visit the stores. Simply looking at the average sample time might not be representative of all the customers who visit both stores.

This is where the t-test comes into play. It helps us understand if the difference between two sample means is actually real or simply due to chance.



Assumptions for Performing a t-test

There are certain assumptions we need to heed before performing a t-test:

- The data should follow a continuous or ordinal scale (the IQ test scores of students, for example)
- The observations in the data should be randomly selected
- The data should resemble a bell-shaped curve when we plot it, i.e., it should be normally distributed.
- Large sample size should be taken for the data to approach a normal distribution (although a t-test is essential for small samples as their distributions are non-normal)
- Variances among the groups should be equal (for independent two-sample t-test).
- Variances among the groups should not be equal (for Welch's t-test).



One sample t-test

The one-sample t-test is a statistical method used to determine whether the mean of a single sample is statistically different from a known or hypothesized population mean.

Background:

The one-sample t-test is employed when you have a single sample and want to assess whether its mean differs significantly from a known or hypothesized population mean.

It's particularly useful in situations where you're interested in evaluating the effectiveness of a treatment, comparing sample data to a theoretical expectation, or testing a hypothesis about a population parameter.



One sample t-test

Hypotheses:

The null hypothesis (H_0) typically states that there is no difference between the sample mean and the population mean, while the alternative hypothesis (H_1) states that there is a significant difference.

- Null Hypothesis (H_0): $\mu = \mu_0$ (The sample mean is equal to the population mean)
- Alternative Hypothesis (H_1): $\mu \neq \mu_0$ (The sample mean is not equal to the population mean)

One sample t-test

Test Statistic:

The test statistic for the one-sample t-test is calculated as:

$$t = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Where:

\bar{X} is the sample mean.

μ_0 is the population mean under the null hypothesis.

s is the sample standard deviation.

n is the sample size.



One sample t-test

Assumptions:

Random Sampling: The sample is randomly selected from the population.

Normality: The data are approximately normally distributed or the sample size is large enough for the Central Limit Theorem to apply.

Independence: The observations in the sample are independent of each other.



One sample t-test

Decision Rule:

To make a decision about the null hypothesis,

- we compare the calculated t-value to the critical t-value from the t-distribution with $n-1$ degrees of freedom, where n is the sample size.
- Alternatively, we can use the p-value associated with the test statistic.



One sample t-test

Conclusion:

- If the calculated t-value is greater than the critical t-value (or if the p-value is less than the significance level, commonly 0.05), we reject the null hypothesis. This indicates that there is a statistically significant difference between the sample mean and the population mean.
- If the calculated t-value is less than the critical t-value (or if the p-value is greater than the significance level), we fail to reject the null hypothesis. This suggests that there is insufficient evidence to conclude that there is a difference between the sample mean and the population mean.

Interpretation:

- If we reject the null hypothesis, it means that the observed sample mean is unlikely to have occurred by random chance alone, and there is evidence to support the alternative hypothesis, indicating a difference between the sample mean and the population mean.



t-test critical values

t Table

cum. prob.	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
Z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										



Example of one sample t-test

Example: Suppose you have a sample of 20 students, and you want to test if their average score is significantly different from the population mean of 70. The sample mean is 72, and the sample standard deviation is 8. Test whether the sample mean is significantly different from the population mean at a 5% significance level.

Solution:

1. Formulate Hypotheses:

1. $H_0: \mu = 70$ (No significant difference)
2. $H_1: \mu \neq 70$ (Significant difference)

2. Choose Significance Level:

1. $\alpha = 0.05$

3. Collect and Analyze Data:

1. $\bar{x} = 72$
2. $s = 8$
3. $n = 20$



4. Calculate the Test Statistic:

$$t = \frac{72 - 70}{\sqrt{\frac{8}{20}}} \approx 1.58$$

5. Determine Degrees of Freedom:

1. $df = 20 - 1 = 19$

6. Find Critical Value or P-value:

1. At $\alpha/2=0.025$ and $df=19$, $t(\alpha/2, df) \approx 2.093$. (With the help of the **t distribution table shown earlier**)

7. Make a Decision:

1. Since $|1.58| < 2.093$ and the p-value is greater than 0.05, it fails to reject the null hypothesis.

8. Interpret the Results:

1. There is not enough evidence to suggest that the average score of the sample is significantly different from the population mean at the 5% significance level.

Two-Tailed Test	
$H_0: \mu_X = \mu_0$	$H_1: \mu_X \neq \mu_0$
Rejection Region	Rejection Region
Acceptance Region	



Two sample t-test cases - Overview

Assumption: Is the variance for two populations equal?

Yes

$$t_{cal} = \frac{(\bar{x}_1 - \bar{x}_2)}{s_p \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$df = n_1 + n_2 - 2$$

No

$$t_{cal} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)}}$$

$$df = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right]^2}{\frac{(s_1^2/n_1)^2}{(n_1 - 1)} + \frac{(s_2^2/n_2)^2}{(n_2 - 1)}}$$



Two sample t-test

Equal Variance Case

Hypothesis testing using a t-test for two samples is a statistical method used to determine if there is a significant difference between the means of two independent groups. It is commonly used in research and experimentation to compare means from different populations or treatments.

Here's a detailed explanation of the process:

1. Define the Hypotheses:

Null Hypothesis (H_0): There is no significant difference between the means of the two groups.

$$H_0: \mu_1 = \mu_2$$

Alternative Hypothesis (H_1 or H_a): There is a significant difference between the means of the two groups.

$$H_a: \mu_1 \neq \mu_2 \text{ (two-tailed test)}$$

$$H_a: \mu_1 > \mu_2 \text{ or } \mu_1 < \mu_2 \text{ (one-tailed test)}$$



Two sample t-test

Equal Variance Case

2. Collect Data:

Obtain data from two independent samples, each with its own set of observations.

3. Verify Assumptions:

Both samples are independent.

Both populations follow a normal distribution.

Homogeneity of variances (the variances of the two populations are equal).

4. Calculate the Test Statistic:

The test statistic formula for a two-sample t-test is calculated using the difference between the sample means divided by the standard error of the difference between the means. Here's the formula:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{pooled} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$



Two sample t-test

Equal Variance Case

Where:

- \bar{x}_1 and \bar{x}_2 are the sample means of the two groups.
- s_{pooled} is the pooled standard deviation, calculated as:

$$s_{pooled} = \sqrt{\frac{(n_1-1) \times s_1^2 + (n_2-1) \times s_2^2}{n_1+n_2-2}}$$

- s_1 and s_2 are the sample standard deviations of the two groups.
- n_1 and n_2 are the sample sizes of the two groups.

5. Determine the Degrees of Freedom:

The test statistic t follows a t-distribution with degrees of freedom.

Degrees of freedom (df) for the t-test is calculated using the formula:

$$df = (n_1 - 1) + (n_2 - 1)$$



Example: Two sample t-test (1/2)

Problem: Suppose we want to compare the exam scores of two different classes (Class A and Class B) to see if there's a significant difference between their mean scores. Assuming equal variance for the population.

Class A: Sample size $n_1 = 30$, Mean Score $\bar{x}_1 = 75$, Standard deviation $s_1 = 8$

Class B : Sample size $n_2 = 25$, Mean Score $\bar{x}_2 = 72$, Standard deviation $s_2 = 7$

Solution:

Null Hypothesis (H_0): There is no significant difference between the mean scores of Class A and Class B.

$$H_0 : \mu_1 = \mu_2$$

Alternative Hypothesis (H_a): There is a significant difference between the mean scores of Class A and Class B.

$$H_a : \mu_1 \neq \mu_2$$

We will use a significance level of $\alpha=0.05$.



Example: Two sample t-test (1/2)

Calculate the Test Statistic:

$$s_{pooled} = \sqrt{\frac{(n_1-1) \times s_1^2 + (n_2-1) \times s_2^2}{n_1+n_2-2}}$$

$$s_{pooled} = \sqrt{\frac{(30-1) \times 8^2 + (25-1) \times 7^2}{30+25-2}}$$

$$s_{pooled} \approx 7.40$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{pooled} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$t = \frac{75 - 72}{7.40 \times \sqrt{\frac{1}{30} + \frac{1}{25}}}$$

$$t \approx \frac{3}{1.99} \approx 1.51$$



t-test critical values

t Table

cum. prob one-tail two-tails	$t_{.50}$ 0.50	$t_{.75}$ 0.25	$t_{.80}$ 0.20	$t_{.85}$ 0.15	$t_{.90}$ 0.10	$t_{.95}$ 0.05	$t_{.975}$ 0.025	$t_{.99}$ 0.01	$t_{.995}$ 0.005	$t_{.999}$ 0.001	$t_{.9995}$ 0.0005
df	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										



Example: Two sample t-test (1/2)

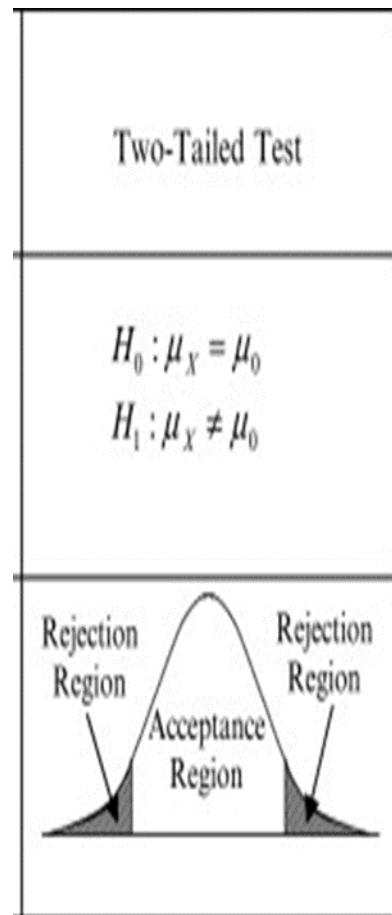
Degrees of Freedom: $df = 30+25-2=53$

Critical Values:

From the t-distribution table, with $df=53$ and $\alpha=0.05$,

$t_{critical}$ is approximately ± 2.004 .

Since $|1.51| < 2.004$, we fail to reject the null hypothesis.



Example: Two sample t-test (2/2)

Problem: Suppose a school district is considering implementing a new teaching method (Method A) for teaching mathematics. To evaluate its effectiveness, they conduct a study comparing it to the traditional teaching method (Method B). Assuming the equal variance for the population. They randomly select two groups of students from the same grade level.

Group A: Students taught using Method A

Sample size $n_1 = 35$, Mean Score $\bar{x}_1 = 85$, Standard deviation $s_1 = 10$

Group B: Students taught using Method B

Sample size $n_2 = 40$, Mean Score $\bar{x}_2 = 80$, Standard deviation $s_2 = 8$

Solution:

Null Hypothesis (H_0): There is no significant difference between the mean scores of students taught using Method A and Method B.

$$H_0: \mu_1 = \mu_2$$

Alternative Hypothesis (H_a): There is a significant difference between the mean scores of students taught using Method A and Method B.

$$H_a: \mu_1 \neq \mu_2$$



Example: Two sample t-test (2/2)

We will use a **significance level** of $\alpha=0.05$.

Calculate the Test Statistic:

$$s_{pooled} = \sqrt{\frac{(35-1) \times 10^2 + (40-1) \times 8^2}{35+40-2}}$$

$$s_{pooled} \approx 9.04$$

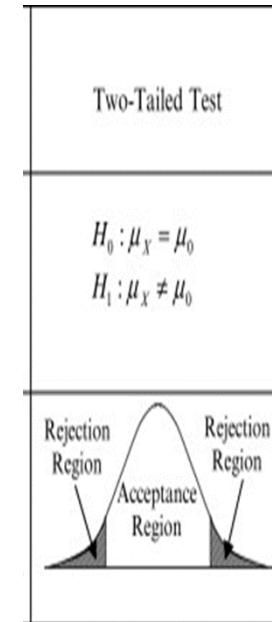
$$t = \frac{85-80}{9.04 \times \sqrt{\frac{1}{35} + \frac{1}{40}}}$$
$$t \approx \frac{5}{2.303} \approx 2.17$$

Degrees of Freedom: $df = 35+40-2=73$

Critical Values:

From the t-distribution table, with $df = 73$ and $\alpha=0.05$, $t_{critical}$ is approximately ± 1.994 .

Since $|2.17| > 1.994$, we reject the null hypothesis.



t-test critical values

t Table

cum. prob one-tail	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$	
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
df												
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62	
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599	
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924	
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610	
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869	
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959	
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408	
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041	
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781	
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587	
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437	
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318	
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221	
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140	
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073	
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015	
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965	
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922	
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883	
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850	
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819	
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792	
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768	
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745	
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725	
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707	
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690	
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674	
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659	
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646	
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551	
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460	
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416	
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390	
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300	
Z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291	
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%	
	Confidence Level											



Two sample t-test

Unequal Variance Case

the hypothesis testing process for a two-sample t-test when the population variances are not assumed to be equal, also known as the Welch's t-test.

Here's a detailed explanation of the process:

1. Define the Hypotheses:

Null Hypothesis (H_0): There is no significant difference between the means of the two groups.

$$H_0: \mu_1 = \mu_2$$

Alternative Hypothesis (H_1 or H_a): There is a significant difference between the means of the two groups.

$$H_a: \mu_1 \neq \mu_2 \text{ (two-tailed test)}$$

$$H_a: \mu_1 > \mu_2 \text{ or } \mu_1 < \mu_2 \text{ (one-tailed test)}$$



Two sample t-test

Unequal Variance Case

2. Collect Data:

Obtain data from two independent samples, each with its own set of observations.

3. Verify Assumptions:

Both samples are independent.

Both populations follow a normal distribution.

The populations do not need to have equal variances.

4. Calculate the Test Statistic:

Use the Welch's t-test formula to calculate the test statistic:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$



Two sample t-test

Unequal Variance Case

Where:

- \bar{x}_1 and \bar{x}_2 are the sample means of the two groups.
- s_1^2 and s_2^2 are the sample standard deviations of the two groups.
- n_1 and n_2 are the sample sizes of the two groups.

5. Determine the Degrees of Freedom:

The test statistic t follows a t-distribution with degrees of freedom.

Degrees of freedom (df) for this t-test is calculated using the formula:

$$df \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{\left(\frac{s_1^2}{n_1} \right)^2}{n_1-1} + \frac{\left(\frac{s_2^2}{n_2} \right)^2}{n_2-1}}$$

This equation is used to approximate the degrees of freedom for the t-distribution when the sample sizes and variances of the two groups being compared are different. It is used in situations where the assumption of equal variances (as assumed in the traditional Student's t-test) does not hold.



Two sample t-test

Unequal Variance Case

6. Find Critical Values or P-value:

Look up the critical value of t from the t -distribution table based on the significance level (α) and degrees of freedom.

Alternatively, calculate the p-value using software or statistical tools.

7. Make a Decision:

If the calculated t -value is greater than the critical value (or if the p-value is less than α), reject the null hypothesis.

If the calculated t -value is less than the critical value (or if the p-value is greater than α), fail to reject the null hypothesis.



Two sample t-test

Unequal Variance Case

8. Interpretation:

If the null hypothesis is rejected, it indicates that there is a significant difference between the means of the two groups.

If the null hypothesis is not rejected, it suggests that there is insufficient evidence to conclude a significant difference between the means of the two groups.

This procedure allows for hypothesis testing when the assumption of equal population variances is violated, making it applicable in a wider range of scenarios where the population variances may differ.

Example

Question 1:

Suppose you want to investigate whether there is a significant difference in the average scores between two teaching methods. You have two groups of students: Group A, taught using Method 1, and Group B, taught using Method 2. Assume Variance of Populations are not equal.

You collect the following data:

- Group A (Method 1): Sample size (n_1) = 25, Sample mean (\bar{x}_1) = 78, Sample standard deviation (s_1) = 7
- Group B (Method 2): Sample size (n_2) = 30, Sample mean (\bar{x}_2) = 82, Sample standard deviation (s_2) = 9

Test whether there is a significant difference in the average scores between the two teaching methods at a 5% significance level.

Solution:

1. Formulate Hypotheses:

1. $H_0: \mu_1 = \mu_2$ (No significant difference between teaching methods)
2. $H_1: \mu_1 \neq \mu_2$ (Significant difference between teaching methods)

2. Choose Significance Level:

1. $\alpha = 0.05$

3. Collect and analyze Data:

1. Group A: $\bar{x}_1 = 78$, $s_1 = 7$, $n_1 = 25$
2. Group B: $\bar{x}_2 = 82$, $s_2 = 9$, $n_2 = 30$



Example

4. Calculate the Test Statistic:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

On Substituting values: t value : -2.06

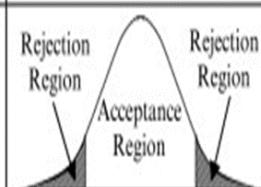
5. Determine Degrees of Freedom:

Degrees of freedom (df) are calculated using the formula:

$$\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{\left(\frac{s_1^2}{n_1} \right)^2}{n_1-1} + \frac{\left(\frac{s_2^2}{n_2} \right)^2}{n_2-1}}$$

$df \approx 51.69$ (rounded down to the nearest whole number, $df = 51$)

Two-Tailed Test
$H_0: \mu_X = \mu_0$
$H_1: \mu_X \neq \mu_0$



t-test critical values

t Table

cum. prob one-tail two-tails	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$	
	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005	
	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001	
df												
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62	
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599	
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924	
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610	
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869	
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959	
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408	
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041	
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781	
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587	
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437	
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318	
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221	
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140	
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073	
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015	
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965	
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922	
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883	
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850	
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819	
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792	
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768	
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745	
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725	
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707	
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690	
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674	
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659	
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646	
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551	
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460	
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416	
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390	
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300	
z		0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
Confidence Level												
0% 50% 60% 70% 80% 90% 95% 98% 99% 99.8% 99.9%												



Example

6. Find Critical Value or P-value:

1. At $\alpha/2=0.025$ and $df = 51$, $t_{\alpha/2, df}$ is approximately ± 2.009 (using a t-table or statistical software).

7. Make a Decision:

1. Since $-2.06 < -2.009$, reject the null hypothesis.

8. Interpret the Results:

1. There is enough evidence to suggest that there is a significant difference in the average scores between the two teaching methods at the 5% significance level.



REFERENCES

- Introduction to Probability and Statistics for Engineers and Scientists, Sixth Edition, Sheldon M. Ross
- Statistical Methods Combined Edition (Volume I& II), N G Das



CONCLUSION

- We discussed the t-test (for Population unknown variance)
 - Motivation Example about t-test
 - Assumptions of t-test
- We discussed one Sample t-test along with solved examples.
- We discussed about Two Sample t-test along with solved examples.
 - (i) Assuming Populations with equal Variance
 - (ii) Assuming Populations with unequal Variance





THANK YOU



JAN 2024

MINE AUTOMATION AND DATA ANALYTICS





MINE AUTOMATION AND DATA ANALYTICS



SWAYAM NPTEL COURSE ON MINE AUTOMATION AND DATA ANALYTICS

By

Prof. Radhakanta Koner

Department of Mining Engineering

Indian Institute of Technology (Indian School of Mines) Dhanbad



Module 9 : Hypothesis Testing

Lecture 22B : Chi-Squared Test

CONCEPTS COVERED

- 1) Introduction to chi-squared test
- 2) Assumptions of chi-squared test
- 3) Introduction to chi-squared test for independence
 - Solved Example of the above chi-squared test - 1
- 4) Introduction to chi-squared goodness of fit test
 - Solved Example of the above chi-squared test -1
 - Solved Example of the above chi-squared test -2



Chi-squared test

The Chi-squared test is a statistical test used to determine if there is a significant association between two categorical variables. It's particularly useful when you want to compare observed frequencies with expected frequencies to see if there is a significant difference between them.

There are two main types of Chi-Squared tests:

- 1) Chi-squared test of independence 2) Chi-squared goodness of fit test.

Note: These two tests are the same mathematically. However, they are utilized for distinct goals; we generally conceive them as separate tests.



Chi-squared test

First, we will discuss the Chi-Squared test for independence.

We use the chi-squared test for independence to determine whether there is a significant association between two categorical variables. This test is particularly useful when we want to examine the relationship between two variables to see if they are related or independent of each other.

1. Formulate the Hypotheses:

Null Hypothesis (H_0): There is no significant association between the two categorical variables.

Alternative Hypothesis (H_1): There is a significant association between the two categorical variables.

2. Choose the Significance Level (α):

Common choices are 0.05, 0.01, or 0.10.

3. Collect and Organize Data:

Organize the data into a contingency table, which displays the frequencies of each combination of the two categorical variables.



Chi-squared test

4. Calculate Expected Frequencies:

Calculate the expected frequency for each cell in the contingency table under the assumption that the variables are independent. The expected frequency (E) for a cell is given by

$$E = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$$

5. Calculate the Test Statistic:

Calculate the Chi-Squared test statistic (χ^2) using the formula:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where O_{ij} is the observed frequency, and E_{ij} is the expected frequency for each cell.



Chi-squared test

6. Determine Degrees of Freedom:

Degrees of freedom (df) is given by (Number of Rows-1) × (Number of Columns-1)

7. Find Critical Value or P-value:

Look up the critical value from the Chi-Squared distribution table or use statistical software to find the p-value.

8. Make a Decision:

If $\chi^2 > \chi_{\alpha, df}^2$ or if the p-value $< \alpha$, reject the null hypothesis.

If $\chi^2 \leq \chi_{\alpha, df}^2$ and p-value $\geq \alpha$, fail to reject the null hypothesis.

9. Interpret the Results:

If the null hypothesis is rejected, it suggests that there is a significant association between the two categorical variables.



Chi-squared Critical Values

Degrees of freedom (df)	Significance level (α)							
	.99	.975	.95	.9	.1	.05	.025	.01
1	-----	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892



Assumptions of Chi-squared test

The chi-squared test relies on several assumptions to ensure its validity and reliability. These **assumptions** include:

Independence of Observations: The observations or data points used in the chi-squared test should be independent of each other. In other words, the occurrence or value of one observation should not influence the occurrence or value of another observation.

Random Sampling: The data should be collected through a random sampling process to ensure that the sample is representative of the population from which it is drawn. This helps to minimize bias and ensure that the results are generalizable.

Categorical Data: The chi-squared test is designed for categorical data, meaning that the variables being analyzed are divided into distinct categories or groups. It is not appropriate for numerical data or continuous variables.

Expected Frequencies: The expected frequencies in each cell of the contingency table (or cross-tabulation) should be greater than or equal to 5. When expected frequencies are too small, the chi-squared test may produce unreliable results. In such cases, alternative tests like Fisher's exact test may be more appropriate.



Assumptions of Chi-squared test

Large Sample Size: While there is no strict requirement for sample size, the chi-squared test tends to perform better with larger sample sizes. As the sample size increases, the distribution of the test statistic approaches a chi-squared distribution, making the test results more reliable.

Mutual Exclusivity and Exhaustiveness: The categories within each variable should be mutually exclusive (i.e., each observation should belong to only one category) and exhaustive (i.e., all possible categories should be represented in the analysis). This ensures that every observation is accounted for and avoids ambiguity in interpretation.

No Cell Count Should Be Zero: None of the cells in the contingency table should have an observed or expected frequency of zero. A zero count in any cell can lead to undefined results or computational issues when calculating the test statistic.

Adhering to these assumptions helps to ensure that the chi-squared test provides accurate and meaningful results when assessing the association between categorical variables. If any of these assumptions are violated, the reliability and validity of the test may be compromised, and alternative methods or adjustments may be necessary.



Example of Chi-squared test

A human resources department wants to investigate whether there is a significant association between employees' educational attainment (high school diploma, bachelor's degree, or master's degree) and their reported level of job satisfaction (satisfied or dissatisfied). They collect data from a sample of 500 employees and categorize them based on their educational attainment and job satisfaction level.

	Satisfied	Dissatisfied	Total
High School Diploma	50	100	150
Bachelor's Degree	150	100	250
Master's Degree	100	0	100
Total	300	200	500

1. Formulate Hypotheses:

1. H_0 : There is no significant association between educational attainment and job satisfaction.
2. H_1 : There is a significant association between educational attainment and job satisfaction.

2. Choose Significance Level:

1. $\alpha = 0.05$



3. Calculate Expected Frequencies:

Using the same approach as before, we calculate the expected frequencies for each cell by multiplying the row total by the column total and dividing by the overall total.

contingency table (with observed frequencies)

	Satisfied	Dissatisfied	Total
High School Diploma	50	100	150
Bachelor's Degree	150	100	250
Master's Degree	100	0	100
Total	300	200	500

$$E_{ij} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$$

contingency table (with expected frequencies)

	Satisfied	Dissatisfied	Total
High School Diploma	$(150 * 300) / 500 = 90$	$(150 * 200) / 500 = 60$	150
Bachelor's Degree	$(250 * 300) / 500 = 150$	$(250 * 200) / 500 = 100$	250
Master's Degree	$(100 * 300) / 500 = 60$	$(100 * 200) / 500 = 40$	100
Total	300	200	500



4. Calculate the Test Statistic:

1. Use the Chi-Squared formula to calculate χ^2 .

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$\chi^2 = [(50 - 90)^2 / 90] + [(100 - 60)^2 / 60] + [(150 - 150)^2 / 150] + [(100 - 150)^2 / 150] + [(100 - 60)^2 / 60] + [(0 - 40)^2 / 40]$$

$$\chi^2 = (20.00) + (40.00) + (0.00) + (16.67) + (40.00) + (100.00)$$

$$\chi^2 \approx 216.67$$

contingency table (with observed frequencies)

	Satisfied	Dissatisfied	Total
High School Diploma	50	100	150
Bachelor's Degree	150	100	250
Master's Degree	100	0	100
Total	300	200	500

contingency table (with expected frequencies)

	Satisfied	Dissatisfied	Total
High School Diploma	$(150 * 300) / 500 = 90$	$(150 * 200) / 500 = 60$	150
Bachelor's Degree	$(250 * 300) / 500 = 150$	$(250 * 200) / 500 = 100$	250
Master's Degree	$(100 * 300) / 500 = 60$	$(100 * 200) / 500 = 40$	100
Total	300	200	500

Chi-squared Critical Values

Degrees of freedom (df)	Significance level (α)							
	.99	.975	.95	.9	.1	.05	.025	.01
1	-----	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892



5. Determine Degrees of Freedom:

Degrees of freedom (df) = (number of rows - 1) * (number of columns - 1) = (3 - 1) * (2 - 1) = 2

6. Find Critical Value or P-value:

At a significance level of $\alpha = 0.05$ and 2 degrees of freedom, the critical value from the chi-squared distribution table is approximately 5.99.

7. Compare χ^2 to Critical Value:

Since χ^2 (216.67) is much greater than the critical value (5.99), we reject the null hypothesis.

8. Interpretation:

We conclude that there is a significant association between employees' educational attainment and their reported level of job satisfaction in the company.

This example demonstrates the application of the chi-squared independence test to analyze the relationship between two categorical variables (educational attainment and job satisfaction) and interpret the results accordingly.



Chi-squared test

Chi-Square Goodness of Fit test (Second main type of Chi-Squared test)

The chi-square goodness of fit test is used to determine whether an observed frequency distribution matches an expected frequency distribution for a categorical variable.

This test is often employed when you want to compare observed frequencies to expected frequencies for one categorical variable, rather than comparing two categorical variables as in the chi-square test for independence.



Example - 1 of Chi-squared test (Goodness of fit)

Suppose we have observed eye color frequencies for a sample of 200 individuals:

Blue eyes: 50 individuals

Brown eyes: 100 individuals

Green eyes: 30 individuals

Gray eyes: 20 individuals

We want to test whether these observed frequencies match the expected distribution of eye colors in the population, hypothesizing that eye colors are distributed equally:

Blue eyes: 25%

Brown eyes: 50%

Green eyes: 15%

Gray eyes: 10%



1. Formulate Hypotheses:

H_0 : The observed eye color frequencies match the expected distribution.

H_1 : The observed eye color frequencies do not match the expected distribution.

2. Choose Significance Level:

1. $\alpha = 0.05$



3. Calculate Expected Frequencies:

Since we expect the eye colors to be distributed equally, we can calculate the expected frequencies as follows:

- Expected frequency for blue eyes: $E_{\text{blue}} = 0.25 \times 200 = 50$
- Expected frequency for brown eyes: $E_{\text{brown}} = 0.50 \times 200 = 100$
- Expected frequency for green eyes: $E_{\text{green}} = 0.15 \times 200 = 30$
- Expected frequency for gray eyes: $E_{\text{gray}} = 0.10 \times 200 = 20$



4. Calculate the Test Statistic:

1. Use the Chi-Squared formula to calculate χ^2 .

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$\chi^2 = [(50 - 50)^2 / 50] + [(100 - 100)^2 / 100] + [(30 - 30)^2 / 30] + [(20 - 20)^2 / 20]$$

$$\chi^2 = 0 + 0 + 0 + 0$$

$$\chi^2 \approx 0$$

Chi-squared Critical Values

Degrees of freedom (df)	Significance level (α)							
	.99	.975	.95	.9	.1	.05	.025	.01
1	-----	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892



JAN 2024

5. Determine Degrees of Freedom:

Since we have 4 categories of eye color, the degrees of freedom (df) = 4 – 1 = 3

6. Find Critical Value or P-value:

At a significance level of 0.05 and 3 degrees of freedom, the critical value from the chi-square distribution table is approximately 7.815.

7. Compare χ^2 to Critical Value:

Since χ^2 (= 0) is less than the critical value of 7.815, we fail to reject the null hypothesis.

8. Interpretation:

We conclude that there is not enough evidence to suggest that the observed eye color frequencies differ significantly from the expected distribution. Thus, we accept the hypothesis that eye colors are distributed equally in the population.

In this case, the chi-square goodness of fit test indicates that the observed frequencies match the expected frequencies, supporting the hypothesis of an equal distribution of eye colors in the population.



Example - 2 of Chi-squared test (Goodness of fit)

Suppose a chocolate manufacturer claims that their assorted box of chocolates contains four flavors in the following proportions:

Milk Chocolate: 30%

Dark Chocolate: 25%

White Chocolate: 20%

Caramel: 25%

To verify this claim, a quality control team randomly selects a sample of 200 chocolates from the assorted box and records the number of chocolates of each flavor.

Observed frequencies from the sample:

Milk Chocolate: 60 chocolates

Dark Chocolate: 40 chocolates

White Chocolate: 50 chocolates

Caramel: 50 chocolates

We will conduct a chi-square goodness of fit test to determine whether the observed distribution of chocolate flavors in the sample matches the claimed distribution by the manufacturer.



1. Formulate Hypotheses:

Null Hypothesis (H0): The observed distribution of chocolate flavors matches the claimed distribution by the manufacturer.

Alternative Hypothesis (H1): The observed distribution of chocolate flavors does not match the claimed distribution by the manufacturer.

2. Choose Significance Level:

$$1. \alpha = 0.05$$



3. Calculate Expected Frequencies:

Based on the claimed proportions, we calculate the expected frequencies for each flavor:

- Expected frequency for Milk Chocolate : $E_{\text{milk}} = 0.3 \times 200 = 60$
- Expected frequency for Dark Chocolate : $E_{\text{dark}} = 0.25 \times 200 = 50$
- Expected frequency for White Chocolate : $E_{\text{white}} = 0.20 \times 200 = 40$
- Expected frequency for Caramel : $E_{\text{caramel}} = 0.25 \times 200 = 50$



4. Calculate the Test Statistic:

1. Use the Chi-Squared formula to calculate χ^2 .

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$\chi^2 = [(60 - 60)^2 / 60] + [(40 - 50)^2 / 50] + [(50 - 40)^2 / 40] + [(50 - 50)^2 / 50]$$

$$\chi^2 = 0 + 2 + 2.5 + 0$$

$$\chi^2 \approx 4.5$$



Chi-squared Critical Values

Degrees of freedom (df)	Significance level (α)							
	.99	.975	.95	.9	.1	.05	.025	.01
1	-----	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892



5. Determine Degrees of Freedom:

Since we have 4 categories of chocolate flavors, the degrees of freedom (df) is $4 - 1 = 3$

6. Find Critical Value or P-value:

At a significance level of 0.05 and 3 degrees of freedom, the critical value from the chi-square distribution table is approximately 7.815.

7. Compare χ^2 to Critical Value:

Since $\chi^2 (= 4.5)$ is less than the critical value of 7.815, we fail to reject the null hypothesis.

8. Interpretation:

We conclude that there is not enough evidence to suggest that the observed distribution of chocolate flavors significantly differs from the claimed distribution by the manufacturer. Thus, we accept the manufacturer's claim regarding the distribution of flavors in the assorted box of chocolates.

In this example, the chi-square goodness of fit test indicates that the observed frequencies align with the expected frequencies based on the manufacturer's claim.



REFERENCES

- Introduction to Probability and Statistics for Engineers and Scientists, Sixth Edition, Sheldon M. Ross
- Statistical Methods Combined Edition (Volume I& II), N G Das



CONCLUSION

- We discussed the chi-squared test along with computation steps.
- We discussed the assumptions of chi-squared test
- We discussed the chi-squared test for independence, along with one example
- We discussed the chi-squared goodness of fit test, along with two solved examples





THANK YOU



JAN 2024

MINE AUTOMATION AND DATA ANALYTICS

