Certificate Program in Machine Learning & Artificial Intelligence: Batch-3

NEENA PANDEY, IIMV

CERTIFICATE PROGRAM IN INTEGRATED OPERATIONS & SUPPLY CHAIN MANAGEMENT

Session Objective

- Random Variable
- Probability Distributions
  - Discrete Distributions
  - Continuous Distributions
- Normal & Standard Normal Distribution
- Estimation – Point & Interval
- Hypothesis Testing

Random Variables & Probability Distribution

- A numerical measure of the outcome of an experiment
- Random variables and probability distributions are models for populations of data
- May be discrete or continuous
- Examples?
  - No. of days that a supplier takes to supply raw material
  - No. of defects in a batch
  - Time between two customer arrivals
  - People income; Percentage of project completed after six months
- In supply chain, most models are created with an assumption of continuous demand

Probability Distributions - Discrete

- Discrete Probability Function
  - Probability Distributions - Discrete
- Discrete Uniform Probability Distribution
- Discrete Uniform Probability Function

$$f(x) = 1/n$$

where: n = the number of values the random variable may assume

The values of the random variable are equally likely.

Probability Distributions - Discrete

- Binomial Probability Distribution
  - 'n' identical independent trials, two possible outcomes in each trial
  - The probability of both events does not change from trial to trial
  - Trials are independent
  - Example:
    - Showing an advertisement to 100 families, each either purchases a policy or not; Families are selected randomly (independent trials); Probability that at least 20 families buy the policy
    - Past experience – probability that any customer will buy from the store is 0.30. What is the probability that two out of the next three customers will buy

$$P(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{(n-x)!\,x!} p^x q^{n-x}$$

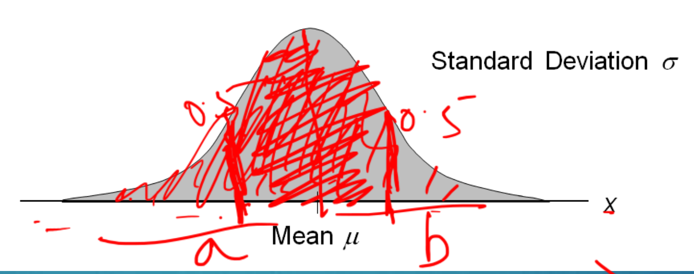
Probability Distributions - Continuous

- Continuous Uniform Probability Distribution
  - Probability is proportional to the interval's length
  - Probability Density Function
  - a, b: smallest, largest value the variable can assume
  - E.g., Flight time of an airplane traveling from Delhi to Mumbai
  - E.g., Battery life of iPad Mini is uniformly distributed between 8.5 and 12 hours – what's the probability that the battery life will be more than 11 hours?
  - Area as a measure of probability
    - Probability that x takes value between x1 and x2 = area under the curve f(x)

$$f(x) = 1/(b-a) \quad \text{for } a \le x \le b$$
$$= 0 \quad \text{elsewhere}$$

Probability Distributions – Continuous - Normal

- Normal Probability Density Function: $$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e - \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2$$
- Normal Distribution Curve:



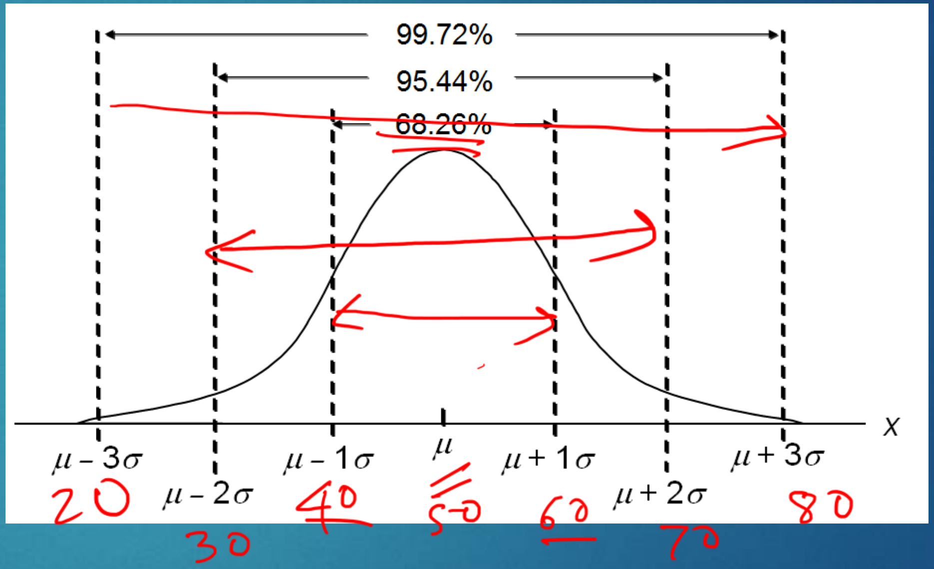
- Total area under the curve = 1
- Excel Functions to calculate cumulative probabilities as well as inverse function to calculate the value of x

Normal Probability Distribution – N(Mean, Std. Dev)

Chebyshev →

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

- **Empirical Rule**
- 68.26% of values of a normal random variable are within ± 1 standard deviation of its mean.
- 95.44% of values of a normal random variable are within ± 2 standard deviations of its mean.
- 99.72% of values of a normal random variable are within ± 3 standard deviations of its mean.

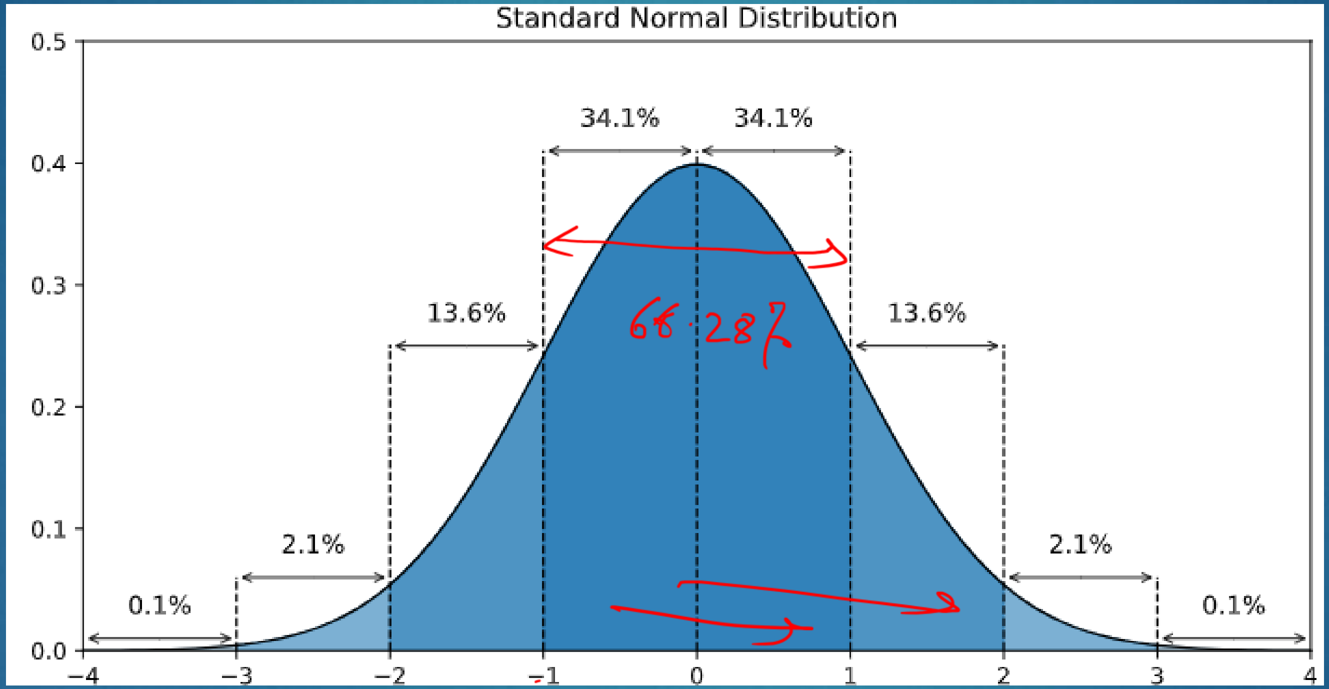


$\mu = 50$

$\sigma = 10$

$X \sim N(50, 100)$

Standard Normal Distribution – N(0,1)



Estimation: Point & Interval

- Sampling: Simple random sampling (with replacement, without replacement), stratified sampling, systematic, Convenience
- Point estimation – use sample data to estimate population parameter
  - sample mean, sample standard deviation etc.
- Interval estimation – includes margin of error
  - Point estimate +- Margin of error
  - Confidence interval
    - To have a higher confidence, width of confidence interval will be larger

Hypothesis Testing

- Performed to determine if a statement about the value of a population parameter should or should not be rejected
  - Null Hypothesis: H_0: tentative assumption about a population parameter
  - Alternate Hypothesis: H_1: Opposite of Null Hypothesis
- Developing Null & Alternate Hypothesis is very crucial
- Type 1 Error
  - Rejecting Null Hypothesis when it is true
  - The probability of Type 1 error – Level of significance
- Type II Error
  - Accepting Null Hypothesis when it is false

Type 1 & Type 11 Errors

	Population Condition	
Conclusion	H_0 True	H_0 False
Accept H_0	Correct Conclusion	Type II Error
Reject H_0	Type 1 Error	Correct Conclusion

$\beta_1 = 0$

α → level of significance

$\alpha = 0{\cdot}05$

$\alpha = 0{\cdot}01$

$\alpha = 0{\cdot}001$

The Importance of p-value

- P-value: Support provided by the sample for the Null hypothesis
- If p-value is less than or equal to the level of significance (α Type 1 error), the value of the test statistic is in the rejection region
  - Reject $H_0(\beta=0)$, if the p-value <= α
- Guidelines for interpreting p-values
  - Less than 0.01: Overwhelming evidence to conclude H_1 is true.
  - Between 0.01 & 0.05: Strong evidence to conclude H_1 is true.
  - Between .05 and .10: Weak evidence to conclude H_1 is true.
  - Greater than .10: Insufficient evidence to conclude H_1 is true

Example of Linear Regression (for p-value)

- We first check the distributions of the data collected (both dependent and independent variables)
- We check assumptions for setting a linear regression model
- On running linear regression, you get a model – coefficients for independent variable
- Then you need to check the significance of those coefficients – and hence we need the p-value
- Compare the p-value with the level of significance you have decided
  - A low p-value indicates that coefficient is likely not equal to zero (i.e., Reject H0)
  - A high p-value means we cannot conclude that the independent variable affects the dependent variable

p= 0.1 β=