Day 3 – Hypothesis Testing

Hypotheis Testing - Introduction

- Hypothesis testing is a process used for either rejecting or retaining a null hypothesis.

- The claim is usually about population parameters such as mean or proportion and we seek evidence from a sample for the support of the claim.

Blackout Babies

- On 9 November 1965 there was a power failure that resulted in blackout for approximately 12 hours in New York and surrounding areas.
- Nine months later, in August 1966, New York Times published a series of three articles in which it claimed that the birth rates in August 1966 was higher than normal based on interviews with city doctors.
- **The babies were nicknamed 'blackout babies'.**
- The articles published by the New York Times raised an interesting question on whether power failures result in procreation?

- Izenman and Zabell (1981) using time series data analysis claimed that there is not enough evidence to suggest that the 1965 power failure resulted in increased birth rate nine months after the blackout.

Hypothesis Testing

- Hypothesis is a claim or belief, hypothesis testing is a statistical process of either rejecting or retaining a claim or belief or association related to a business context, product, service, processes, etc.

- Hypothesis testing consists of two complementary statements called **null hypothesis** and **alternative hypothesis**, and only one of them is true.

Hypothesis Testing – Business Cases

- Children who drink the health drink Complan (a health drink owned by the company Heinz in India) are likely to grow taller.
- If you drink Horlicks, you can grow taller, stronger, and sharper (3 in 1).
- Using fair and lovely (fair and handsome) cream can make one fair and lovely (fair and handsome).
- Wearing perfume (such as Axe) will help to attract opposite gender (known as Axe effect).
- Women use camera phone more than men.
- Beautiful people are likely to have girl child. This is one of my favorite hypotheses since I have a daughter I can claim that I am good looking.
- Married people are happier than singles, especially those who married their best friend (many married people may not agree!).
- Vegetarians miss few flights.
- Smokers are better sales people.

Setting up a Hypothesis Test

- Data analysis in general can be classified as exploratory data analysis (detective analytics) or confirmatory data analysis.

- **In exploratory data analysis**, the idea is to look for new or previously unknown hypothesis or suggest hypotheses.

- **In the case of confirmatory data analysis**, the objective is to test the validity of a hypothesis (confirm whether the hypothesis is true or not) using techniques such as hypothesis testing and regression.

Hypothesis Testing – Steps involved

1. **Describe the hypothesis in words.**
   Hypothesis is described using a population parameter (such as mean, standard deviation, proportion, etc.) about which a claim (hypothesis) is made. Example claims (hypothesis) are :

   (a) Average time spent by women using social media is more than men.
   (b) On average women upload more photos in social media than men.
   (c) Customers with more than one mobile handsets are more likely to churn.

Hypothesis Testing – Steps involved

2. Based on the claim made in step 1, define null and alternative hypotheses.

Initially we believe that the null hypothesis is true.

In general, null hypothesis means that there is no relationship between the two variables under consideration (for example, null hypothesis for the claim 'women use social media more than men' will be 'there is no relationship between gender and the average time spent in social media').

Null and alternative hypotheses are defined using a population parameter.

Hypothesis Testing – Steps involved

3. Identify the test statistic to be used for testing the validity of the null hypothesis.

Test statistic will enable us to calculate the evidence in support of null hypothesis.

The test statistic will depend on the probability distribution of the sampling distribution:

For example, if the test is for mean value and the mean is calculated from a large sample and if the population standard deviation is known, then the sampling distribution will be a normal distribution and the test statistic will be a Z-statistic (standard normal statistic).

Hypothesis Testing – Steps involved

4. Decide the criteria for rejection and retention of null hypothesis.

* This is called significance value traditionally denoted by symbol **α**.
* The value of a will depend on the context and usually 0.1**, 0.05,** and 0.01 are used.

5. Calculate the p-value (probability value), which is the conditional probability of observing the test statistic value when the null hypothesis is true. In simple terms, **p-value is the evidence in support of the null hypothesis.**

6. Take the decision to reject or retain the null hypothesis based on the p-value and significance value α.

The null hypothesis is rejected when p-value is less than α and the null hypothesis is retained when p-value is greater than or equal to α.

Null and Alternate Hypothesis

Null hypothesis (H_0) refers to the statement that there is no relationship or no difference between different groups with respect to the value of a population parameter.

- Null hypothesis is the claim that is assumed to be true initially.
- **That is at the beginning we assume that the null hypothesis is true** and try to retain it unless there is strong evidence against null hypothesis.

Alternative hypothesis (H_A) is the complement of null hypothesis.

Alternative hypothesis is what the researcher believes to be true and would like to reject the null hypothesis.

Null and Alternate Hypothesis Examples

Hypothesis statement to definition of null and alternative hypothesis

S. No.	Hypothesis Description	Null and Alternative Hypothesis
1	Average annual salary of machine learning experts is different for males and females. (In this case, the null hypothesis is that there is no difference in male and female salary of machine learning experts)	$H_0: \mu_m = \mu_f$ $H_A: \mu_m \neq \mu_f$ μ_m and μ_f are average annual salary of male and female machine learning experts, respectively.
2	On average people with Ph.D. in analytics earn more than people with Ph.D. in engineering.	$H_0: \mu_a \leq \mu_e$ $H_A: \mu_a > \mu_e$ μ_a = Average annual salary of people with Ph.D. in analytics. μ_e = Average annual salary of people with Ph.D. in engineering. It is essential to have the equal sign in null hypothesis statement.

Test Statistic

- Test statistic is the standardized difference between the estimated value of the parameter being tested calculated from the sample(s) and the hypothesis value (that is, standardized difference between X- and μ in the case of testing mean) in order to establish the evidence in support of the null hypothesis.

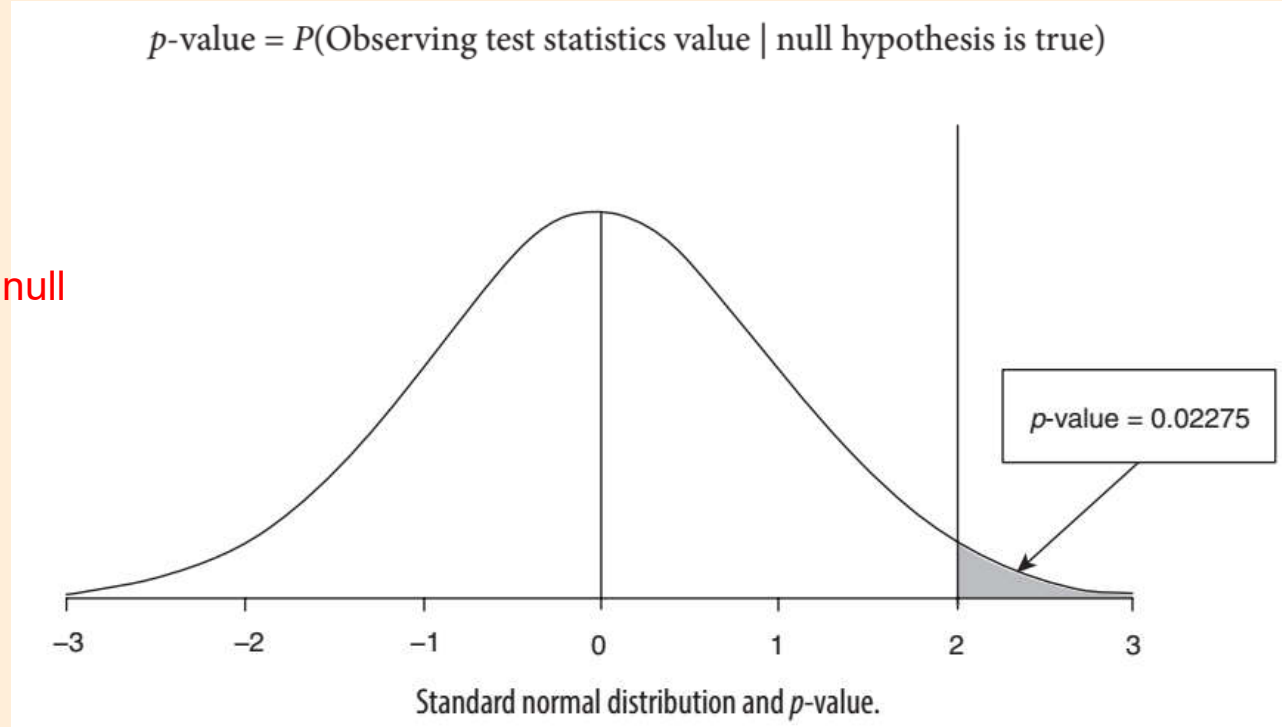
- Test statistic is the standardized value used for calculating the p-value (probability value) in support of null hypothesis.

- Since test statistic is a standardized value, it measures the standardized distance (measured in terms of number of standard deviations) between the value of the parameter estimated from the sample(s) and the value of the null hypothesis.

P-value

- The p-value is the conditional probability of observing the statistic value when the null hypothesis is true.

P-value is the evidence in support of null hypothesis.



Decision Criteria – Significance Value

- Significance level, usually denoted by α, is the criteria used for taking the decision regarding the null hypothesis (reject or retain) based on the calculated p-value.
- **The significance value a is the maximum threshold for p-value.** The decision to reject or retain will depend on whether the calculated p-value crosses the threshold value a or not.

$$\text{Significance value } \alpha = P(\text{Rejecting a null hypothesis} \mid \text{null hypothesis is true})$$

Decision making under hypothesis testing

Criteria	Decision
p-value $< \alpha$	Reject the null hypothesis
p-value $\geq \alpha$	Retain (or fail to reject) the null hypothesis

One-Tailed and Two-Tailed Tests

Consider the following three hypotheses:

1. Salary of machine learning experts on average is at least US $100,000.
2. Average waiting time at the London Heathrow airport security check is less than 30 minutes.
3. Average annual salaries of male and female MBA students are different at the time of graduation.

STATEMENT 1 Salary of machine learning experts on average is at least US $100,000:

The null and alternative hypotheses in this case are given by

$$H_0: \mu_m \leq 100{,}000$$

$$H_A: \mu_m > 100{,}000$$

where μ_m is the average annual salary of machine learning experts.

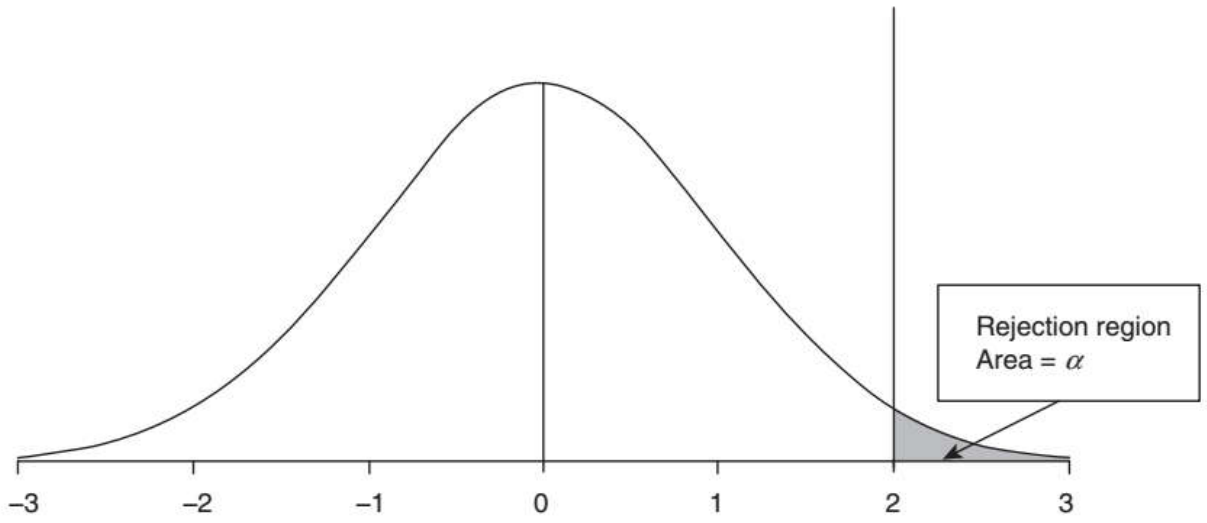


FIGURE 6.2 Right-tailed hypothesis test's rejection region.

STATEMENT 2 Average waiting time at the London Heathrow airport security check is less than 30 minutes:

The null and alternative hypotheses in this case are given by

$$H_0: \mu_w \geq 30$$
$$H_A: \mu_w < 30$$

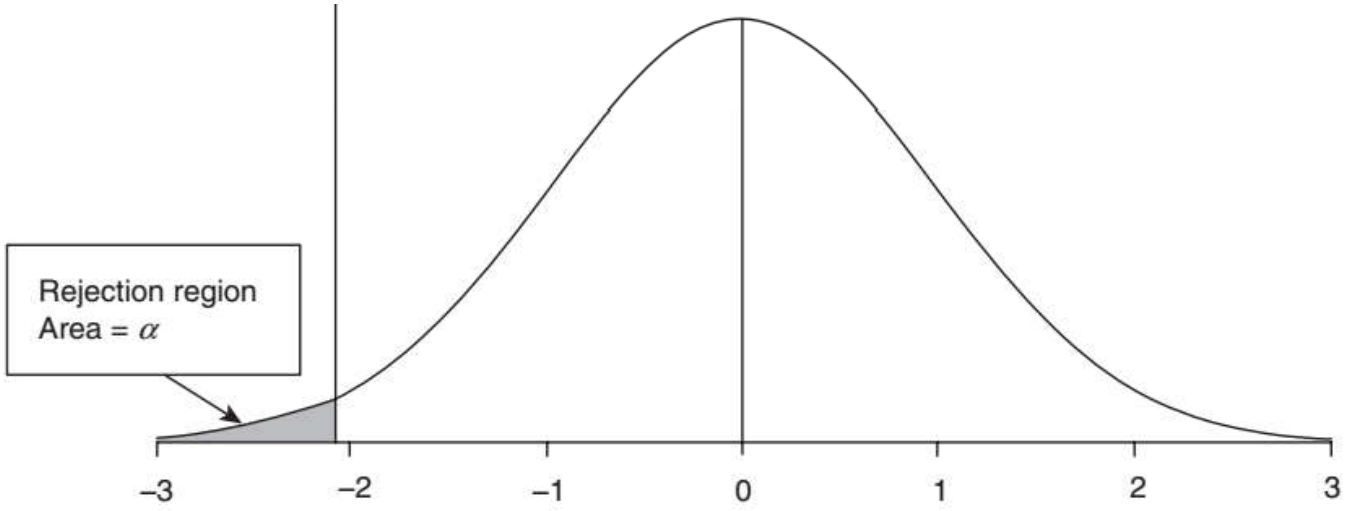


FIGURE 6.3 Rejection region in case of left-sided test.

STATEMENT 3 Average salary of male and female MBA students at graduation is different:

The null and alternative hypotheses in this case are given by

$$H_0: \mu_m = \mu_f$$

$$H_A: \mu_m \neq \mu_f$$

where μ_m and μ_f are the average salaries of male and female MBA students,

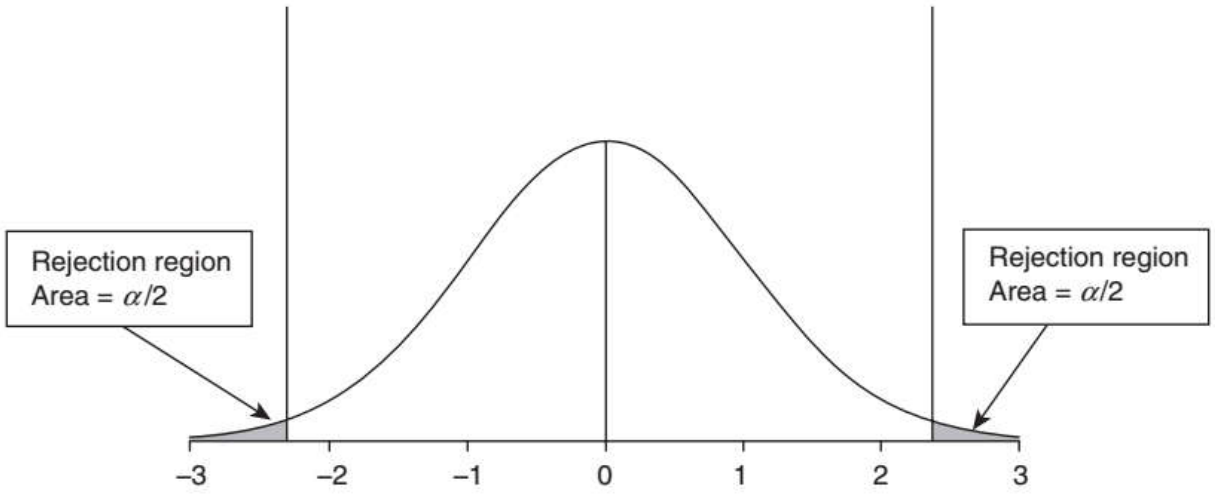


FIGURE 6.4 Rejection region in case of two-tailed test.

Type-I and Type-II Errors

Type I Error: Conditional probability of rejecting a null hypothesis when it is true is called Type I Error or False Positive (falsely believing that the claim made in alternative hypothesis is true). The significance value a is the value of Type I error.

$$\text{Type I Error} = \alpha = P(\text{Rejecting null hypothesis} \mid H_0 \text{ is true})$$

Type II Error: Conditional probability of failing to reject a null hypothesis (or retaining a null hypothesis) when the alternative hypothesis is true is called Type II Error or False Negative (falsely believing that there is no relationship).

$$\text{Type II Error} = \beta = P(\text{Retain null hypothesis} \mid H_0 \text{ is false})$$

$$\text{Power of the test} = 1 - \beta = 1 - P(\text{Retain null hypothesis} \mid H_0 \text{ is false})$$

$$\text{Alternatively the power of test} = 1 - \beta = P(\text{Reject null hypothesis} \mid H_0 \text{ is false})$$

Type-I and Type-II Errors

Description of type I error, type II error, and the power of test

Actual Value of H_0	Decision made about Null Hypothesis Based on the Hypothesis Test	
	Reject H_0	Retain H_0
H_0 is true	Type I error	Correct Decision
	$P(\text{Reject } H_0 \mid H_0 = \text{true}) = \alpha$	$P(\text{Retain } H_0 \mid H_0 = \text{true}) = (1 - \alpha)$
H_0 is false	Correct Decision (Power of test)	Type II Error
	$P(\text{Reject } H_0 \mid H_0 = \text{false}) = 1 - \beta$	$P(\text{Retain } H_0 \mid H_0 = \text{false}) = \beta$

Z-Tests

- Z-test (also known as one-sample Z-test) is used when a claim (hypothesis) is made about the population parameter such as population mean or proportion when population variance is known.

- Since the hypothesis test is carried out with just one sample, this test is also known as one-sample Z-test.

Actual Value of H_0	Decision made about Null Hypothesis Based on the Hypothesis Test	
	Reject H_0	Retain H_0
H_0 is true	Type I error $P(\text{Reject } H_0 \mid H_0 = \text{true}) = \alpha$	Correct Decision $P(\text{Retain } H_0 \mid H_0 = \text{true}) = (1 - \alpha)$
H_0 is false	Correct Decision (Power of test) $P(\text{Reject } H_0 \mid H_0 = \text{false}) = 1 - \beta$	Type II Error $P(\text{Retain } H_0 \mid H_0 = \text{false}) = \beta$

Central Limit Theorem (CLT)

- Central limit theorem (CLT) for sampling distribution of mean, the sampling distribution of mean from an independent and identically distributed population for large sample follows a normal distribution with mean μand standard deviation $\sigma / \sqrt{n}$.

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

One-Sample Z-Test

One-sample Z-test is used when

1. *Testing the value of population mean when population standard deviation is known.*
2. *The population is a normal distribution and the population variance is known.*
3. *The sample size is large and the population variance is known. That is, the assumption of normal distribution can be relaxed for large samples ($n > 30$).*

Critical value for different values of α

	Approximate Critical Values		
α	Left-Tailed Test	Right-Tailed Test	Two-Tailed Test
0.1	−1.28	1.28	−1.64 and 1.64
0.05	−1.64	1.64	−1.96 and 1.96
0.01	−2.33	2.33	−2.58 and 2.58

One Sample Z-Test

Condition for rejection of null hypothesis H_0

Type of Test	Condition	Decision
Left-tailed test	Z-statistic $<$ Critical value	Reject H_0
	Z-statistic $\geq$ Critical value	Retain H_0
Right-tailed test	Z-statistic $>$ Critical value	Reject H_0
	Z-statistic $\leq$ Critical value	Retain H_0
Two-tailed test	$\lvert Z\text{-statistic}\rvert > \lvert\text{Critical Value}\rvert$	Reject H_0
	$\lvert Z\text{-statistic}\rvert \leq \lvert\text{Critical Value}\rvert$	Retain H_0

Exercise - 1

$H_0: \mu \leq 4200$

$H_A: \mu > 4200$

An agency based out of Bangalore claimed that the **average** monthly disposable income of families living in Bangalore is greater than INR 4200 with a standard deviation of INR 3200.

From a random **sample** of 40,000 families, the **average** disposable income was estimated as INR 4250.

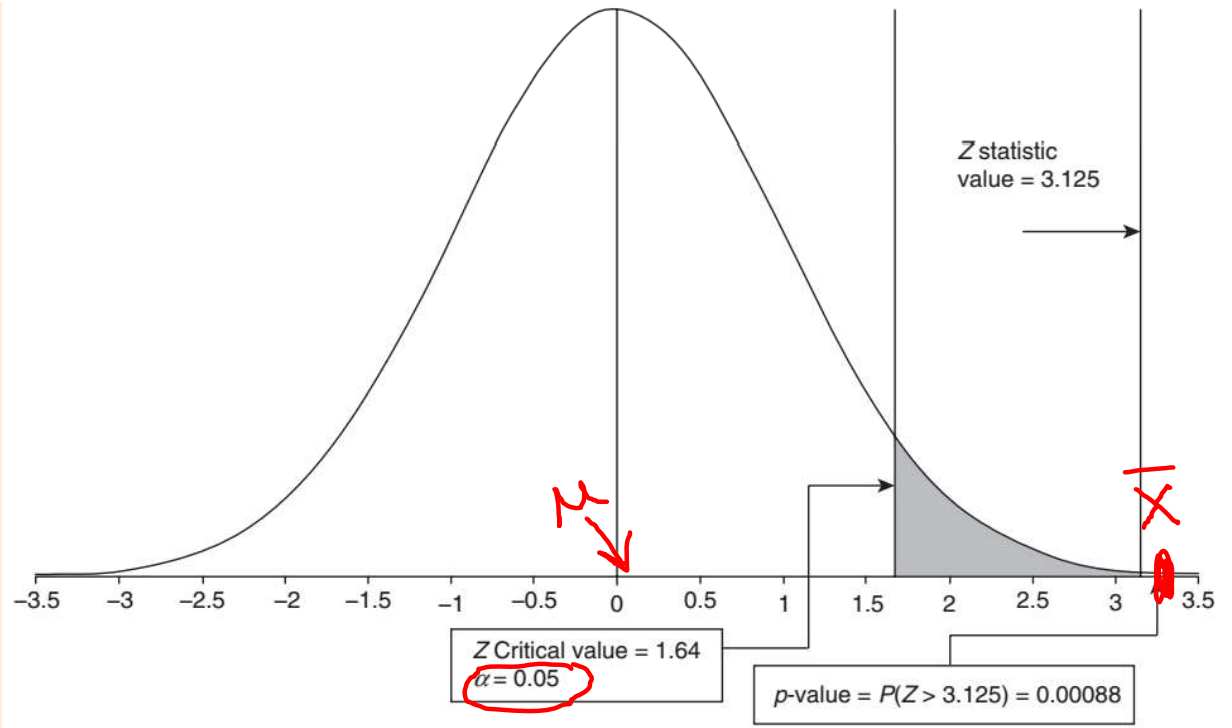
Assume that the **population standard deviation** is INR 3200.

Conduct an appropriate hypothesis test at 95% confidence level (α = 0.05) to check the validity of the claim by the agency.

Exercise – 1: Solution

$$H_0: \mu \leq 4200$$
$$H_A: \mu > 4200$$

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{4250 - 4200}{3200 / \sqrt{40000}} = 3.125$$



Exercise – 2:

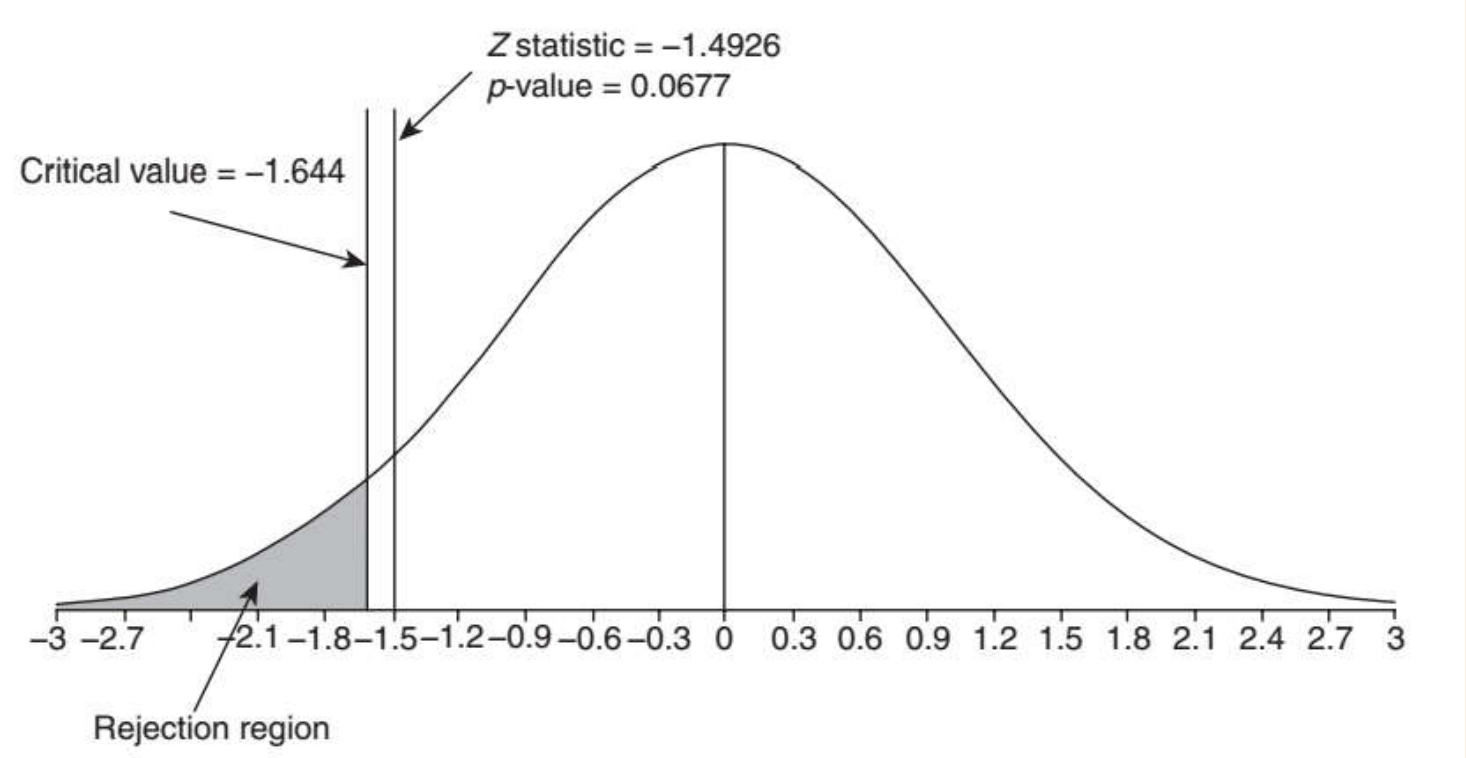
A passport office claims that the passport applications are processed within 30 days of submitting the application form and all necessary documents. Table shows processing time of 40 passport applicants.

The population standard deviation of the processing time is 12.5 days.

Conduct a hypothesis test at significance level a = 0.05 to verify the claim made by the passport office.

$$H_0: \mu \geq 30$$
$$H_A: \mu < 30$$

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{27.05 - 30}{12.5 / \sqrt{40}} = -1.4926$$



Example - 3

According to the company IQ Research, the average Intelligence Quotient (IQ) of Indians is 82 derived based on a research carried out by Professor Richard Lynn, a British Professor of Psychology, using data collected from 2002 to 2006.

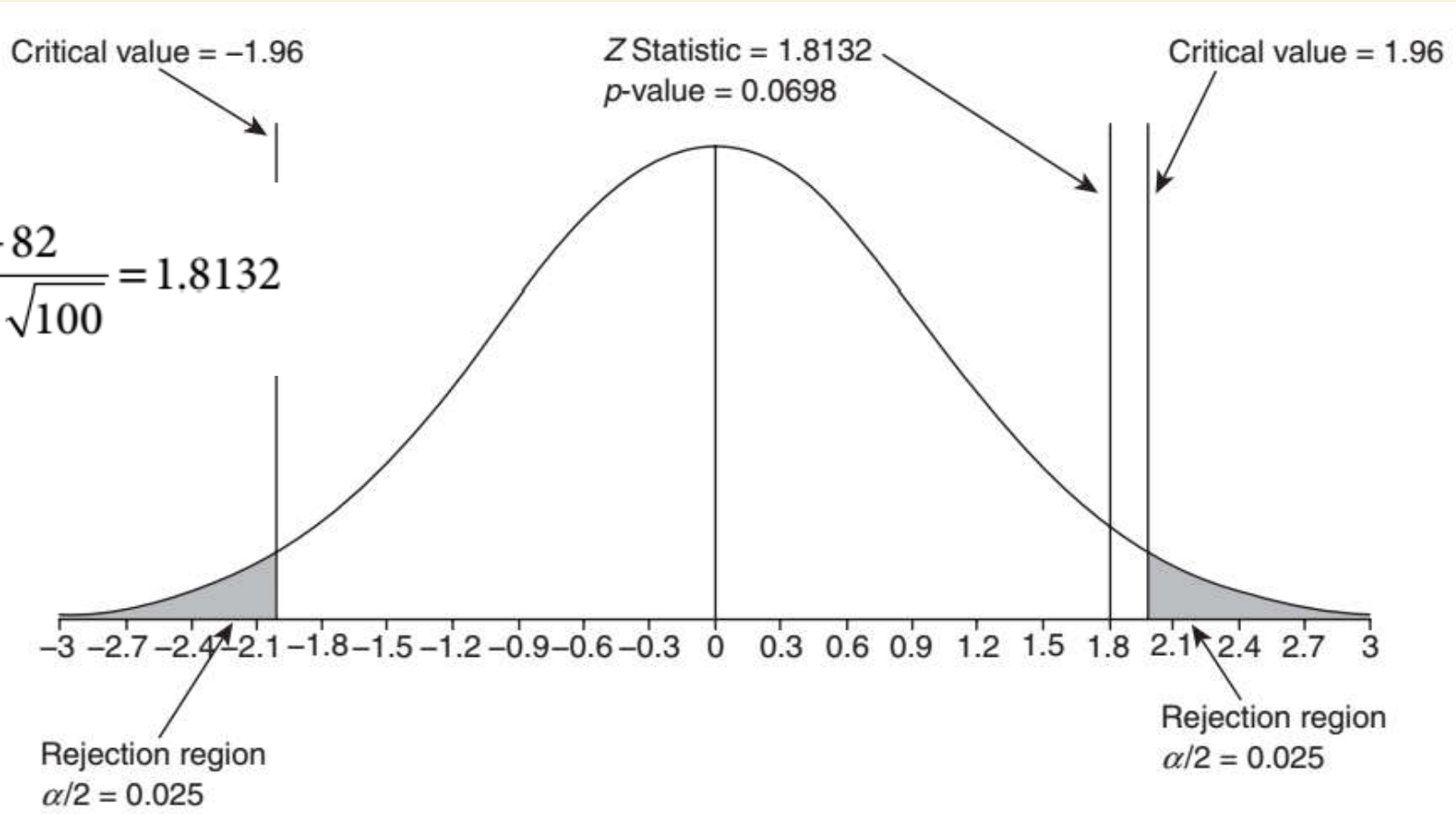
The population standard deviation of IQ is estimated as 1.103. Based on a sample of 100 people from India, the sample IQ was estimated as 84.

(a) Conduct an appropriate hypothesis test at a = 0.05 to validate the claim of IQ Research (that average IQ of Indians is 82).
(b) Ministry of education believes that the IQ is more than 82. If the actual IQ (population mean) of Indians is 86, calculate the Type II error and the power of hypothesis test.

$$H_0: \mu = 82$$

$$H_A: \mu \neq 82$$

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{84 - 82}{11.03 / \sqrt{100}} = 1.8132$$



$H_0: \mu \leq 82$

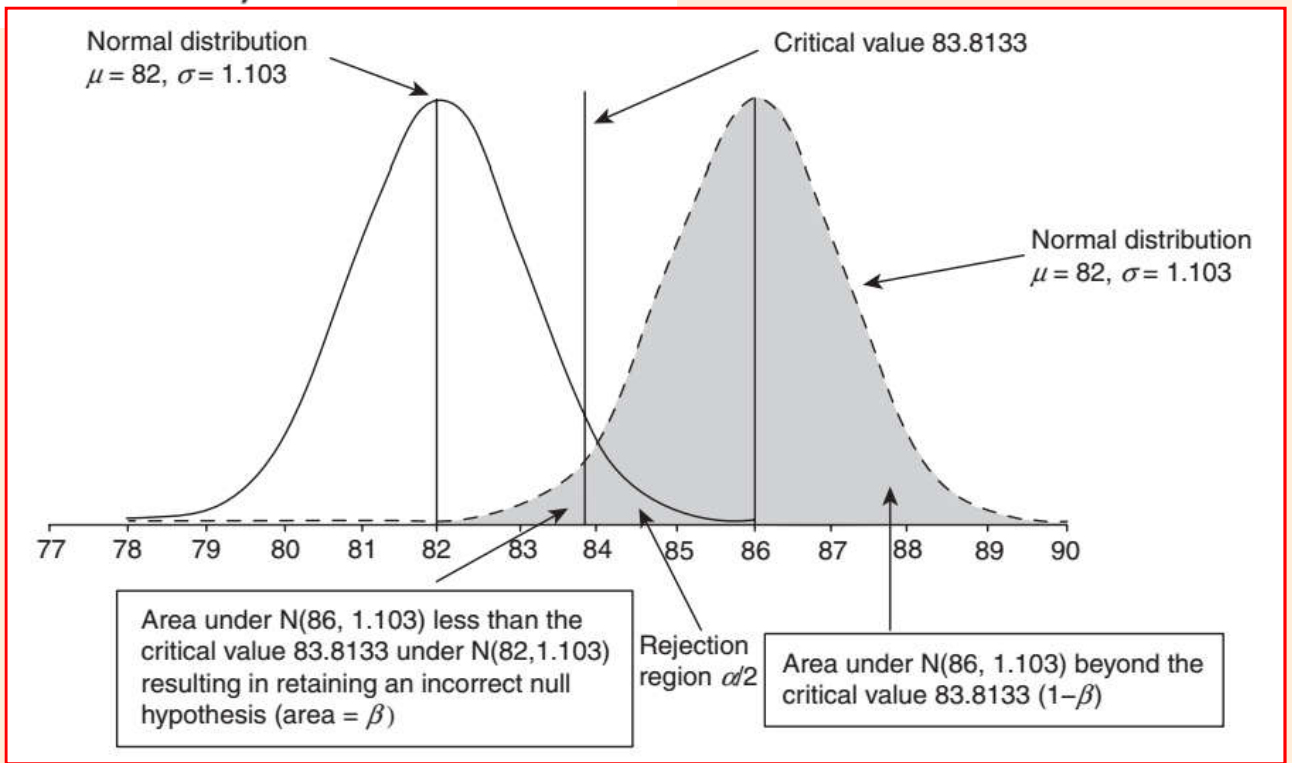
$H_A: \mu > 82$

$$X_{\text{critical}} = \mu + Z_\alpha \times \sigma / \sqrt{n} = 82 + 1.644 \times 1.103 = 83.8133$$

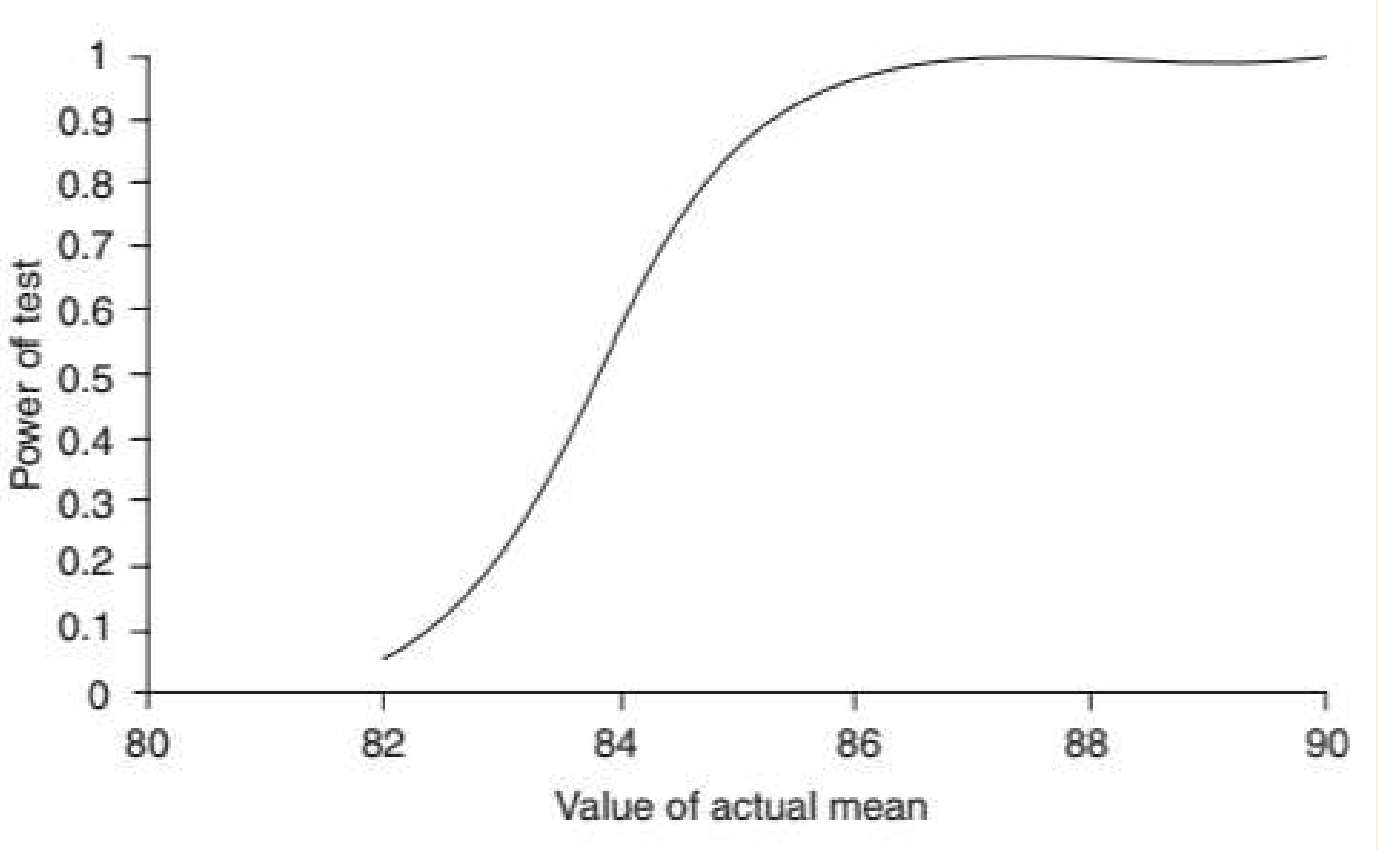
$$P(X \leq 83.8133) = P\left(Z \leq \frac{83.8133 - 86}{1.103}\right) = 0.0237$$

That is, the Type II error $\beta = 0.0237$

The power of test, $1 - \beta = 1 - 0.0237 = 0.9763$



Power of Test and the Power Function



HYPOTHESIS TEST FOR POPULATION MEAN UNDER UNKNOWN POPULATION VARIANCE

- The **t-test** is used when the population follows a normal distribution and the population standard deviation **σ** is unknown and is estimated from the sample.
- **t-test** is a robust test for violation of normality of the data as long as the data is close to symmetry and there are no outliers.

$$t\text{-statistic} = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

- The t-statistics will follow a **t-distribution with (n − 1) degrees of freedom** if the sample is drawn from a population that follows a normal distribution.

Exercise - 4

- Aravind Productions (AP) is a newly formed movie production house based out of Mumbai, India. AP was interested in understanding the production cost required for producing a Bollywood movie.
- The industry believes that the production house will require at least INR 500 million (50 crore) on average.
- It is assumed that the Bollywood movie production cost follows a normal distribution.
- Production cost of 40 Bollywood movies in millions of rupees are given in the Excel Sheet.
- **Conduct an appropriate hypothesis test at a = 0.05 to check whether the belief about average production cost is correct.**

Solution

$n = 40$, $\bar{X} = 429.55$, and $S = 195.0337$

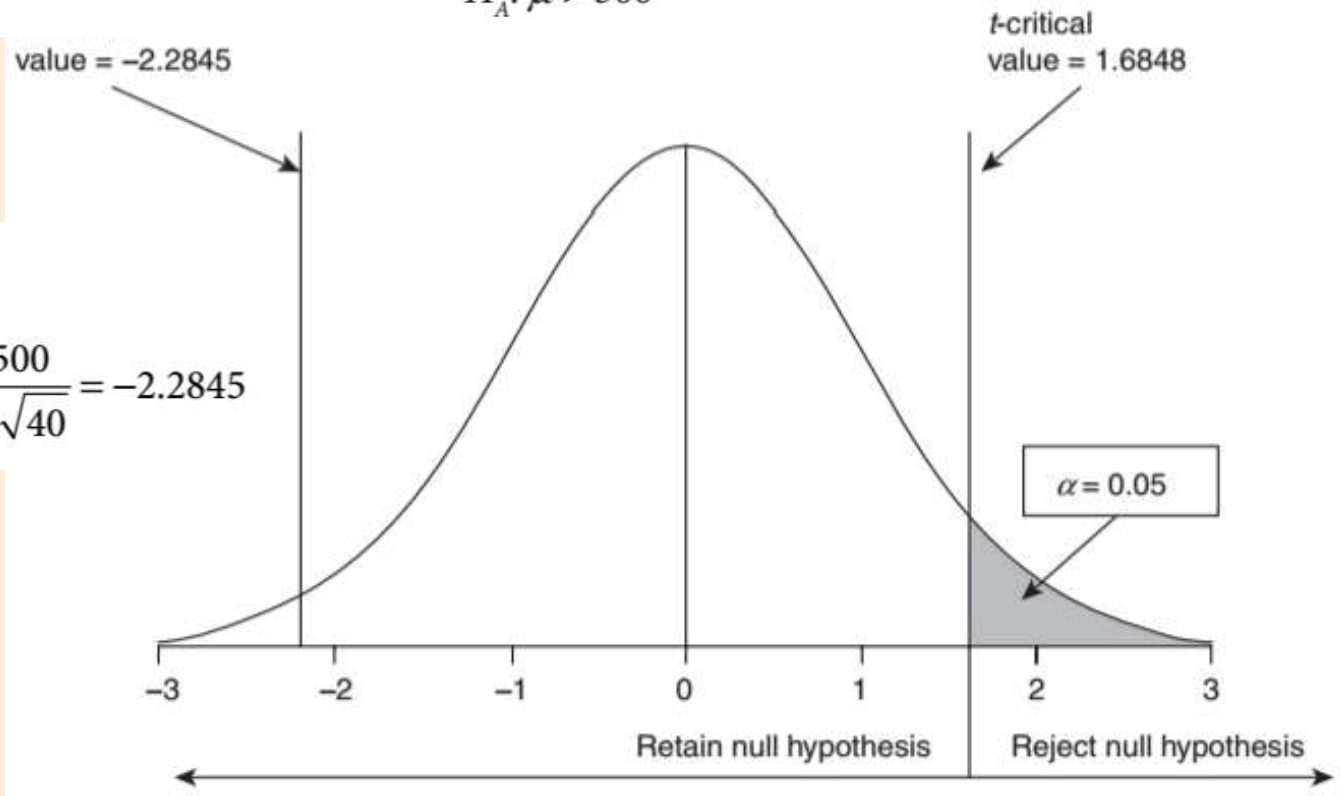
The null and alternative hypotheses are

$$H_0: \mu \leq 500$$

$$H_A: \mu > 500$$

The corresponding test statistic is

$$t\text{-statistic} = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{429.55 - 500}{195.0337/\sqrt{40}} = -2.2845$$



Non-Parametric Tests: Chi-Square Tests

- In the previous sections, we discussed methods of testing hypothesis which are about population parameters and make certain assumptions about the population distribution.
- We assumed that the test statistic follows standard normal distribution, or t-distribution.
- Tests such as Z-test and t-test are called parametric tests since the objective is to infer about a population parameter such as mean in case of single sample tests or compare population parameters in the case of two sample tests.

- **To conduct Z-test and t-test we need summary statistics (mean and/or standard deviation), and not necessarily the entire distribution.**

Non-Parametric Tests: Chi-Square Tests

- In this section, we will be discussing non-parametric tests (also known as distribution free tests since they **do not have assumptions about distribution of the population**).

- Nonparametric tests imply that the tests are not based on the **assumptions that the data is drawn from a probability distribution** defined through parameters such as mean, proportion and standard deviation.

Non-Parametric Tests: Chi-Square Tests

- A **major difference** between parametric and non-parametric tests is that in a parametric test we need only values of the parameter and the knowledge about the distribution, **whereas in case of non-parametric test we use the entire distribution of the data.**

- Importantly, the data may not follow any parametric distribution such as normal distribution.
- **Also, the test is not about the population parameter but about characteristics of the entire distribution** (for example, whether the data follows a normal distribution or not).

Non-Parametric Tests: Chi-Square Tests

A non-parametric method for hypothesis tests is used when one or more of the following conditions exist in the test:

1. The test is not about the population parameter such as mean and standard deviation.

2. The method does not require assumptions about population distribution (such as population follows normal distribution).

Chi-Square Goodness of Fit Tests

The null and alternative hypotheses in chi-square goodness of fit tests are :

H0: There is no statistically significant difference between the observed frequencies and the expected frequencies from a **hypothesized** distribution.

HA: There is a statistically significant difference between the observed frequencies and the expected frequencies from a hypothesized distribution.

Chi-Square Goodness of Fit Tests

$$\chi^2 \text{ statistic} = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- **Thus, chi-square test is always a right-tailed test.**
- In goodness of fit tests, degrees of freedom is k - c - 1, where k is the number of groups, c is the number of parameters estimated from the data.
- **Ideally, expected frequencies in each group should be at least 5.**

Exercise - 5

- Hindustan Airlines (HA) operated daily flights to several Indian cities. One of the problems HA faces is the food preferences by the passengers.
- Captain Cook, the operations manager of HA, **believes** that *35% of their passengers prefer vegetarian food, 40% prefer non-vegetarian food, 20% low calorie food, and 5% request for diabetic food.*
- A sample of **500 passengers** was chosen to analyse the food preferences and the data is shown in Table below.

Sample preferences of 500 customers

Food Type	Vegetarian	Non-Vegetarian	Low Calorie	Diabetic
Number of Passengers	190	185	90	35

- Conduct a chi-square test to check whether Captain Cook's belief is true at a = 0.05.

Solution:

The null and alternative hypotheses in this case are given as

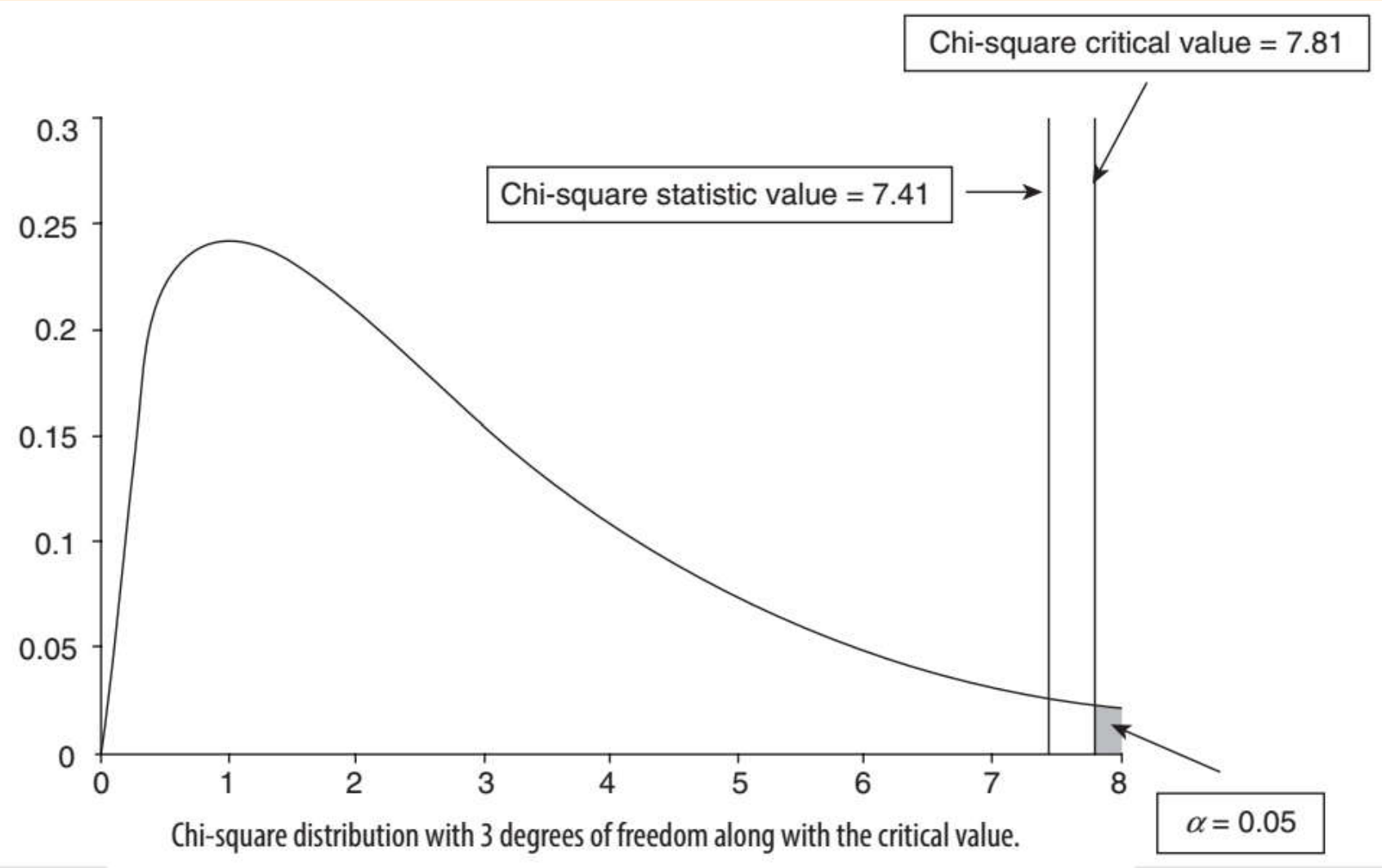
H_0: Probability distribution of the food preference is P(Vegetarian) = 0.35; P(Non-Vegetarian) = 0.40; P(Low Calorie) = 0.20, and P(Diabetic) = 0.05

H_A: Probability distribution of the food preference is not as defined in null hypothesis

Food Type	Observed Frequency (O_i)	Expected Frequency (E_i)	$\frac{(O_i - E_i)^2}{E_i}$
Vegetarian	190	175	1.285
Non-Vegetarian	185	200	1.125
Low Calorie	90	100	1
Diabetic	35	25	4

The chi-square statistic value is given by

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 1.285 + 1.125 + 1 + 4 = 7.410$$



Chi-square distribution with 3 degrees of freedom along with the critical value.

Exercise - 6

- Peter is the chief operating officer at Airmobile, a mobile phone service provider based out of Coimbatore, India. Peter was interested in finding the probability distribution of the call duration**. His friend Erlang suggested Peter that call duration is most likely to be an exponential distribution**.

- To check whether the call duration actually follows an exponential distribution, Peter collected a sample of 50 calls and the duration of call in minutes is shown in Table.

- **Using goodness of fit test, check whether the data follows an exponential distribution.**

Exercise – 6 Data

Sample call duration (in minutes) data									
2.47	4.23	5.41	3.49	4.17	10.09	18.78	0.68	2.28	16.16
0.28	2.97	4.01	5.88	20.32	26.88	19.07	0.22	6.37	10.38
4.2	10.17	1.84	21.88	9.42	0.01	6.15	4.99	3.07	18.6
1.54	10.23	3.99	6.17	0.39	11.03	9.38	1.57	6.91	2.49
5.52	11.53	7.64	8.8	7.17	3.26	6.74	16.32	10	7.45

Solution: The null and alternative hypotheses are

H0: There is no difference between the observed frequencies and expected frequencies from an exponential distribution.

HA: There is a difference between the observed frequencies and expected frequencies from an exponential distribution.

Exercise – 6 Solution:

- We can group the data into 6 groups;
- the minimum and maximum values of the call duration are 0.01 and 26.88.
- The range is 26.87 and the length of the interval is 4.48.
- For exponential distribution, the expected frequency is given by:

$$E_i = n \times [\exp(-\lambda \times L_i) - \exp(-\lambda \times U_i)]$$

where the scale parameter $\lambda = 1/\text{mean call duration}$.

- In this case the mean call duration is 7.652 and the corresponding value of λ= 0.1306.

Exercise – 6 Solution:

Observed, expected, and chi-square statistic values

Group	Group Range		O_i	E_i	$\frac{(O_i - E_i)^2}{E_i}$
1	0.01	4.49	20	22.08	0.20
2	4.5	8.98	13	12.30	0.04
3	8.99	13.47	9	6.85	0.67
4	13.48	17.96	2	3.82	0.87
5	17.97	22.45	5	2.13	3.88
6	22.46	26.94	1	1.18	0.03

Exercise – 6 Solution:

The chi-square statistic value is given by

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 5.7$$

- The chi-square critical value for a = 0.05 (here the degrees of freedom is 4) is 9.48.
- **Since the chi-square statistic value is less than the chi-square critical values**, we retain the null hypothesis and conclude that the data follows an exponential distribution.

Prashant Sahu

Introduction to Probability

Prashant Sahu

Random Experiment

- Random experiment is an experiment in which the outcome is not known with certainty.
- That is, the output of a random experiment cannot be predicted with certainty.
- Predictive analytics mainly deals with random experiments such as predicting quarterly revenue of an organization, customer churn (whether a customer is likely to churn or how many customers are likely to churn before next quarter), demand for a product at a future time period, number of views for an YouTube video, outcome of a football match (win, draw or lose), etc.

Sample Space

- Sample space is the universal set that consists of all possible outcomes of an experiment.
- Sample space is usually represented using the letter ‘S’ and individual outcomes are called the elementary events.
- The sample space can be finite or infinite.

Experiment: Predicting customer churn at an individual customer level

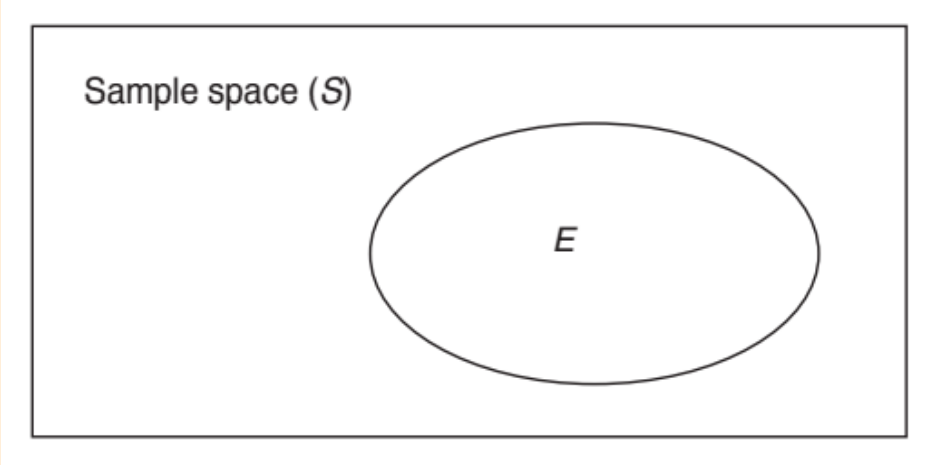
Sample Space = S = {Churn, No Churn}

Experiment: Predicting percentage of customer churn

Sample Space = $S = \{X \mid X \in R, 0 \leq X \leq 100\}$, that is X is a real number that can take any value between 0 and 100 percentage.

Event

Event (E) is a subset of a sample space and probability is usually calculated with respect to an event.



Probability Estimation using Relative Frequency

- The classical approach to probability estimation of an event is based on the relative frequency of the occurrence of that event.
- According to frequency estimation, the probability of an event X, P(X), is given by :-

$$P(X) = \frac{\text{Number of observations in favour of event } X}{\text{Total number of observations}} = \frac{n(X)}{N}$$

- For example, say a company has 1000 employees and every year about 200 employees leave the job. Then the probability of attrition of an employee per annum is 200/1000 = 0.2.

Exercise - 1

A website displays 10 advertisements and the revenue generated by the website depends on the number of visitors to the site clicking on any of the advertisements displayed on the website. The data collected by the company has revealed that out of 2500 visitors, 30 visitors clicked on 1 advertisement, 15 clicked on 2 advertisements, and 5 clicked on 3 advertisements. Remaining did not click on any of the advertisements. Calculate:

(a) The probability that a visitor to the website will click on an advertisement. 50/2500 = 0.02

(b) The probability that the visitor will click on at least two advertisements. 20/2500 = 0.008

(c) The probability that a visitor will not click on any advertisements. 1-0.02 = 0.98

Algebra of Events

- Assume that X, Y and Z are three events of a sample space. Then the following algebraic relationships are valid and are useful while deriving probabilities of events:

Commutative rule: $X \cup Y = Y \cup X$ and $X \cap Y = Y \cap X$

Associative rule: $(X \cup Y) \cup Z = X \cup (Y \cup Z)$ and $(X \cap Y) \cap Z = X \cap (Y \cap Z)$

Distributive rule: $X \cup (Y \cap Z) = (X \cup Y) \cap (X \cup Z)$

$X \cap (Y \cup Z) = (X \cap Y) \cup (X \cap Z)$

- **DeMorgan's Laws** on complementary sets are useful while deriving probabilities:

$$(X \cup Y)^C = X^C \cap Y^C$$
$$(X \cap Y)^C = X^C \cup Y^C$$

where X^C and Y^C are the complementary events of X and Y, respectively.

AXIOMS OF PROBABILITY

1. The probability of event E always lies between 0 and 1. That is, $0 \leq P(E) \leq 1$.
2. The probability of the universal set S is 1. That is, $P(S) = 1$.
3. $P(X \cup Y) = P(X) + P(Y)$, where X and Y are two mutually exclusive events.

AXIOMS OF PROBABILITY

1. For any event A, the probability of the complementary event, written A^C, is given by

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

If $P(A)$ is a probability of observing a fraudulent transaction at an e-commerce portal, then $P(A^C)$ is the probability of observing a genuine transaction.

2. The probability of an empty or impossible event, ϕ, is zero:

$$P(\phi) = 0$$

3. If occurrence of an event A implies that an event B also occurs, so that the event class A is a subset of event class B, then the probability of A is less than or equal to the probability of B:

$$P(A) \leq P(B)$$

If A is students with more than 3.5 CGPA (cumulative grade point average) out of 4 and B is students with a CGPA of more than 3.0, then $P(A) \leq P(B)$.

AXIOMS OF PROBABILITY

4. The probability that either events A or B occur or both occur is given by

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

5. If A and B are mutually exclusive events, so that $P(A \cap B) = 0$, then

$$P(A \cup B) = P(A) + P(B)$$

6. If $A_1, A_2, \ldots, A_n$ are n events that form a partition of sample space S, then their probabilities must add up to 1:

$$P(A_1) + P(A_2) + \cdots + P(A_n) = \sum_{i=1}^{n} P(A_i) = 1$$

Joint Probability

- Let A and B be two events in a sample space. Then the joint probability of the two events, written as P(A ∩ B), is given by

$$P(A \cap B) = \frac{\text{Number of observations in } A \cap B}{\text{Total number of observations}}$$

Exercise - 2

At an e-commerce customer service centre a total of 112 complaints were received. 78 customers complained about late delivery of the items and 40 complained about poor product quality.

(a) Calculate the probability that a customer complaint will be about both late delivery and product quality.

(b) What is the probability that a complaint is **only** about poor quality of the product?

Exercise - 2

Solution:

Let A = Late delivery and B = Poor quality of the product. Let $n(A)$ and $n(B)$ be the number of cases in favour of A and B. So $n(A) = 78$ and $n(B) = 40$. Since the total number of complaints is 112 (here complaints is treated as the sample space), hence

$$n(A \cap B) = 118 - 112 = 6$$

Probability of a complaint about both delivery and poor product quality is

$$P(A \cap B) = \frac{n(A \cap B)}{\text{Total number of complaints}} = \frac{6}{112} = 0.0535$$

$$\text{Probability that the complaint is only about poor quality} = 1 - P(A) = 1 - \frac{78}{112} = 0.3035$$

Exercise - 3

Table below describes loan default status at a bank and their marital status. Calculate the marital status that has maximum joint probability of default.

Joint and marginal probability

Marital Status	Loan Status		Total
	Default	Non-Default	
Single	42	258	300
Married	60	590	650
Divorced	13	37	50
Total	115	885	1000

Exercise – 3: Solution

Joint and marginal probability

Marital Status	Loan Status		Total
	Default	Non-Default	
Single	42	258	300
Married	60	590	650
Divorced	13	37	50
Total	115	885	1000

$$P(\text{Single} \cap \text{Default}) = 0.042$$

$$P(\text{Married} \cap \text{Default}) = 0.06$$

$$P(\text{Divorced} \cap \text{Default}) = 0.013$$

The maximum joint probability is for $P(\text{Married} \cap \text{Default})$.

Marginal Probability

- Marginal probability is simply a probability of an event X, denoted by P(X**), without any conditions.**

Let,

X1 = Loan Status Default

X2 = Loan Status Non-default

Y1 = Marital Status Single

Y2 = Marital Status Married

Y3 = Marital Status Divorced

Then marginal probabilities are

$$P(X_1) = \frac{115}{1000} = 0.115 \, , P(X_2) = \frac{885}{1000} = 0.885$$

$$P(Y_1) = \frac{300}{1000} = 0.3 \, , P(Y_2) = \frac{650}{1000} = 0.65 \, , P(Y_3) = \frac{50}{1000} = 0.05$$

Independent Events

- Two events A and B are said to be independent when occurrence of one event (say event A) does not affect the probability of occurrence of the other event (event B).

- Mathematically, two events A and B are independent when $P(A \cap B) = P(A) \times P(B)$.

For example, winning the toss by the Indian cricket team captain in consecutive matches are independent events.

- Whereas, let event A be life of an equipment exceeding 100 hours and event B be life of the equipment exceeding 200 hours.
- Then events A and B are dependent events.

Independent events are useful property of events since it simplifies calculating probability values.

Conditional Probability

- If A and B are events in a sample space, then the conditional probability of the event B given that the event A has already occurred, denoted by P(B|A), is defined as

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}, \quad P(A) > 0$$

The conditional probability symbol P(B|A) is read as the probability of B given A.

It is necessary to satisfy the condition that P(A) > 0, because it does not make sense to consider the probability of B given that event A is impossible.

Exercise

Joint and marginal probability

Marital Status	Loan Status		Total
	Default	Non-Default	
Single	42	258	300
Married	60	590	650
Divorced	13	37	50
Total	115	885	1000

the conditional probability of default given divorced is

P(Default|Divorced) = 0.013/0.05 = 0.26

P(Default| Divorced) = P(Divorced^Default) / Divorced)

and similarly probability of default given single is

P(Default|Single) = 0.042/0.3 = 0.14

P(Default | Single) = P(Single^Default) / P(Single)

Conditional Prob. = Joint Prob. / Marginal Prob.

BAYES' THEOREM

- Bayes' theorem is one of the most important concepts in analytics since several problems are solved using Bayesian statistics. Consider two events A and B. We can write the following two conditional probabilities:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \text{ and } P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

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

Bayes' theorem helps the data scientists to update the probability of an event (B) when any additional information is provided.

1. P(B) is called the prior probability (estimate of the probability without any additional information).
2. P(B|A) is called the posterior probability (that is, given that the event A has occurred, what is the probability of occurrence of event B). That is, post the additional information (or additional evidence) that A has occurred, what is estimated probability of occurrence of B.
3. P(A|B) is called the likelihood of observing evidence A if B is true.
4. P(A) is the prior probability of A.

Exercise - 5

- Black boxes used in aircrafts are manufactured by three companies A, B and C. 75% are manufactured by A, 15% by B, and 10% by C.
- The defect rates of black boxes manufactured by A, B, and C are 4%, 6%, and 8%, respectively.
- If a black box tested randomly is found to be defective, **what is the probability that it is manufactured by company A?**

- P(A), P(B), P(C) .. > Prob. Of blackbox being manuf. By A,B,C. >> Prior prob. >> 0.75, 0.15, 0.1

- P(D) >> Prob. Of blackbox being defective.

- P(D) = P(D|A)*P(A) + P(D|B)*P(B) + P(D|C)*P(C) = 0.04*0.75 + 0.06*0.15 + 0.08*0.1 = 0.047

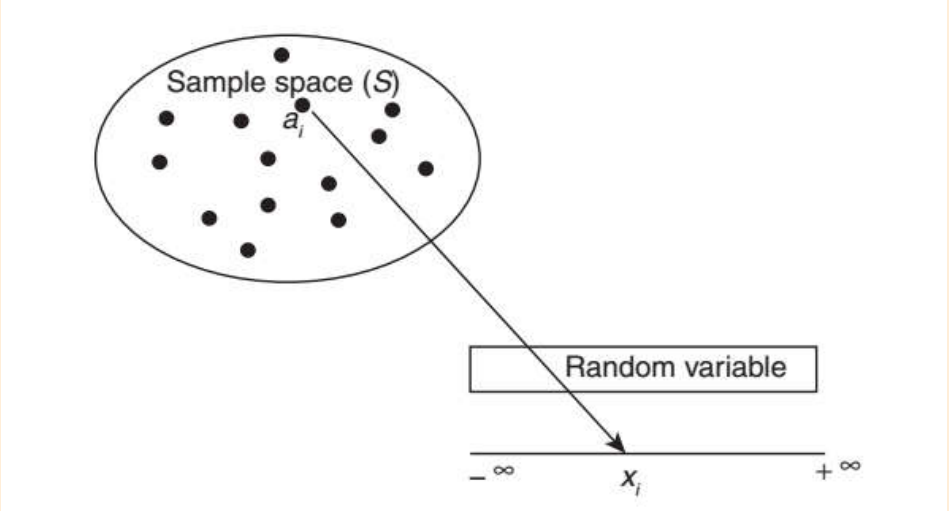
Bayes Theorem:

- P(A | D) >> P(D | A)*P(A) / P(D) >> 0.04*0.75 / 0.047 **= 0.638**

Probability Distributions

RANDOM VARIABLES

- Random variable is a function that maps every outcome in the sample space to a real number.
- Random variables provide robustness required while developing probabilistic models since the outcome of a random experiment may be recorded in different format.
- Random variables can be classified as discrete or continuous depending on the values that the random variable can take.



Discrete Random Variables

- **If the random variable X can assume only a finite or countably infinite set of values, then it is called a discrete random variable.**
- There are very many situations where the random variable X can assume only finite or countably infinite set of values.

1. Credit rating (usually classified into different categories such as low, medium and high or using labels such as AAA, AA, A, BBB, etc.).

2. Number of orders received at an e-commerce retailer which can be countably infinite.

3. Customer churn [the random variables take binary values: (a) Churn and (b) Do not churn].

4. Fraud [the random variables take binary values: (a) Fraudulent transaction and (b) Genuine transaction].

5. Any experiment that involves counting (for example, number of returns in a day from customers of e-commerce portals such as Amazon, Flipkart; number of customers not accepting job offers from an organization).

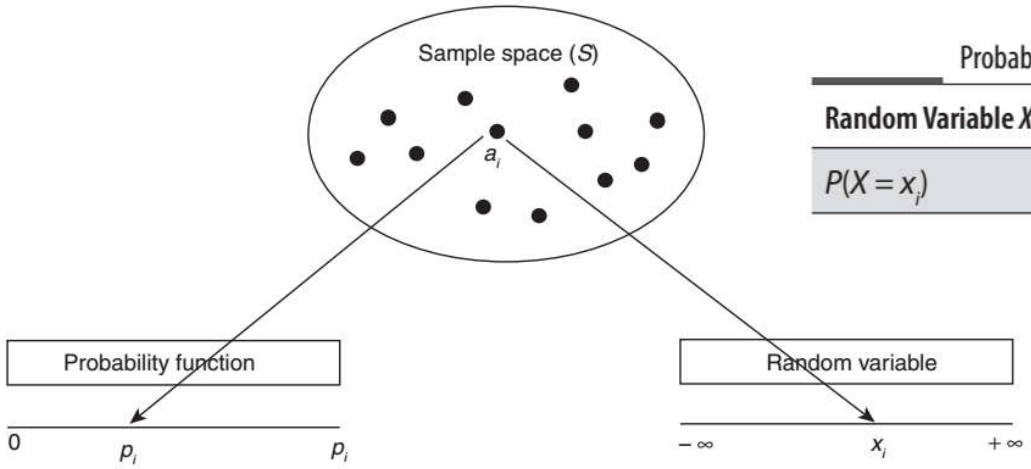
Continuous Random Variables

- **A random variable X which can take a value from an infinite set of values is called a continuous random variable.**
- Examples of continuous random variables are listed below:

1. Market share of a company (which take any value from an infinite set of values between 0 and 100%).
2. Percentage of attrition among employees of an organization.
3. Time to failure of engineering systems.
4. Time taken to complete an order placed at an e-commerce portal.
5. Time taken to resolve a customer complaint at call and service centers. In many situations, a continuous variable may be converted to a discrete random variable for modelling purpose.

Probability Mass Function of a Discrete Random Variable

- For a discrete random variable, the probability that a random variable X taking a specific value xi, P(X = xi), is called the **probability mass function P(xi).**



X	x_1	x_2	.....	x_n
P	p_1	p_2	.....	p_n

Probability distribution of a random variable

Probability mass function

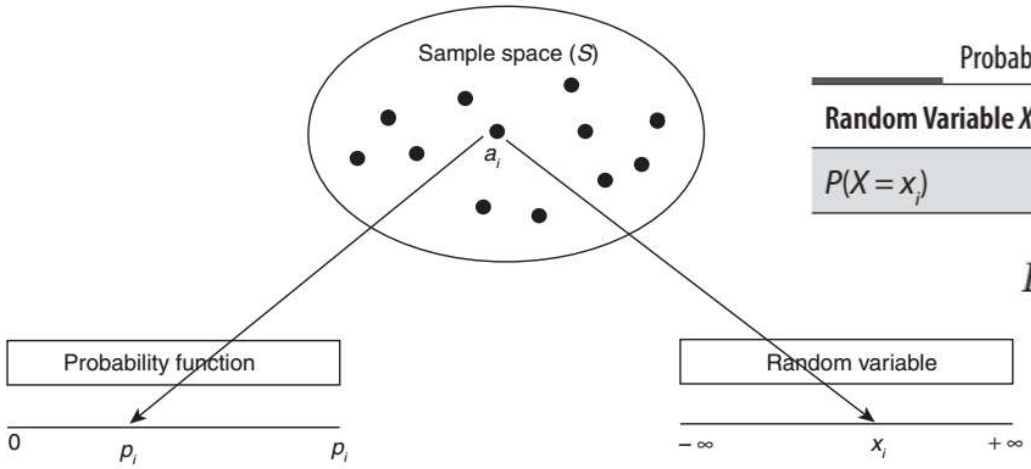
Random Variable X (X = number of fraudulent transactions)	$x_i = 0$	$x_i = 1$	$x_i = 2$	$x_i = 3$	$x_i = 4$
$P(X = x_i)$	0.20	0.15	0.25	0.25	0.15

The probability mass function, P(xi) satisfies the following conditions:

1. $P(x_i) \geq 0.$
2. $\sum_{x_i} P(x_i) = 1$

Cumulative Distribution Function of a Discrete Random Variable

- Cumulative distribution function, F(xi), is the probability that the random variable X takes values **less than or equal xi.** That is, F(xi) = P(X ≤ xi).



X	x_1	x_2	.....	x_n
P	p_1	p_2	.....	p_n

Probability distribution of a random variable

Probability mass function

Random Variable X (X = number of fraudulent transactions)	$x_i = 0$	$x_i = 1$	$x_i = 2$	$x_i = 3$	$x_i = 4$
$P(X = x_i)$	0.20	0.15	0.25	0.25	0.15

$$F(2) = P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = 0.60$$

Expected Value, Variance, and Standard Deviation of a Discrete Random Variable

- Expected value (or mean) of a discrete random variable is given $E(X) = \sum_{i=1}^{n} x_i P(x_i)$

Where xi is the specific value taken by a discrete random variable X and P(xi) is the corresponding probability, that is, P(X = xi).

- For example, expected monetary value (EMV) forms the basis for selecting an alternative from several possible alternatives in a decision tree approach.
- **Variance of a discrete random variable is given by**

$$\mathrm{Var}(X) = \sum_{i=1}^{n} \left[x_i - E(X) \right]^2 \times P(x_i)$$

- Standard deviation of a discrete random variable is given by $\sigma = \sqrt{\mathrm{Var}(X)}$

PROBABILITY DENSITY FUNCTION (PDF)

The probability density function, f(xi), is defined as probability that the value of random variable X lies between an infinitesimally small interval defined by xi and xi + dx and its mathematical expression

$$f(x) = \lim_{\delta x \to 0} \frac{P(x_i \leq X \leq x_i + \delta x)}{\delta x}$$

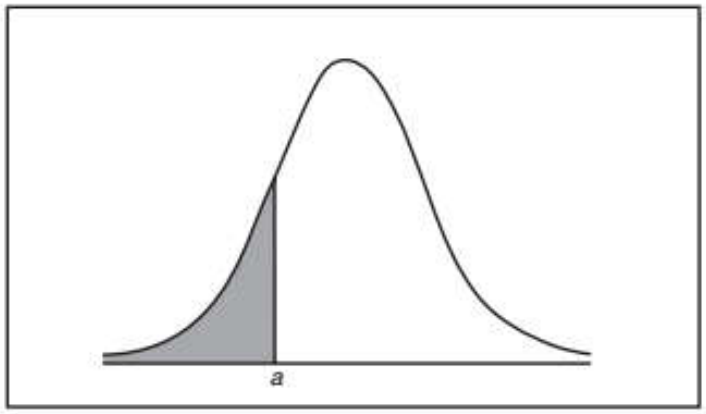
Probability density function reflects how dense is the likelihood of a continuous random variable X taking a value in an infinitesimally small interval around value x.

CUMULATIVE DISTRIBUTION FUNCTION (CDF)

The cumulative distribution function (CDF) of a continuous random variable is defined by

$$F(a) = P(X \leq a) = \int_{-\infty}^{a} f(x)dx$$

- Cumulative distribution function F(a) is the area under the probability density function up to X = a.



BINOMIAL DISTRIBUTION

- Binomial distribution is one of the most important discrete probability distribution due to its applications in several contexts. A random variable X is said to follow a Binomial distribution when
    - The random variable can have only two outcomes success and failure (also known as Bernoulli trials).
    - The objective is to find the probability of getting k successes out of n trials.
    - The probability of success is p and thus the probability of failure is (1 − p).
    - The probability p is constant and does not change between trials.
- **Success and failure are generic terminologies used in binomial distribution; based on the context.**

1. Customer churn where the outcomes are: (a) Customer churn and (b) No customer churn.
2. Fraudulent insurance claims where the outcomes are: (a) Fraudulent claim and (b) Genuine claim.
3. Loan repayment default by a customer where the outcomes are: (a) Default and (b) No default.

Binomial Distribution

- The PMF of the Binomial distribution (probability that the number of success will be exactly x out of n trials) is given by

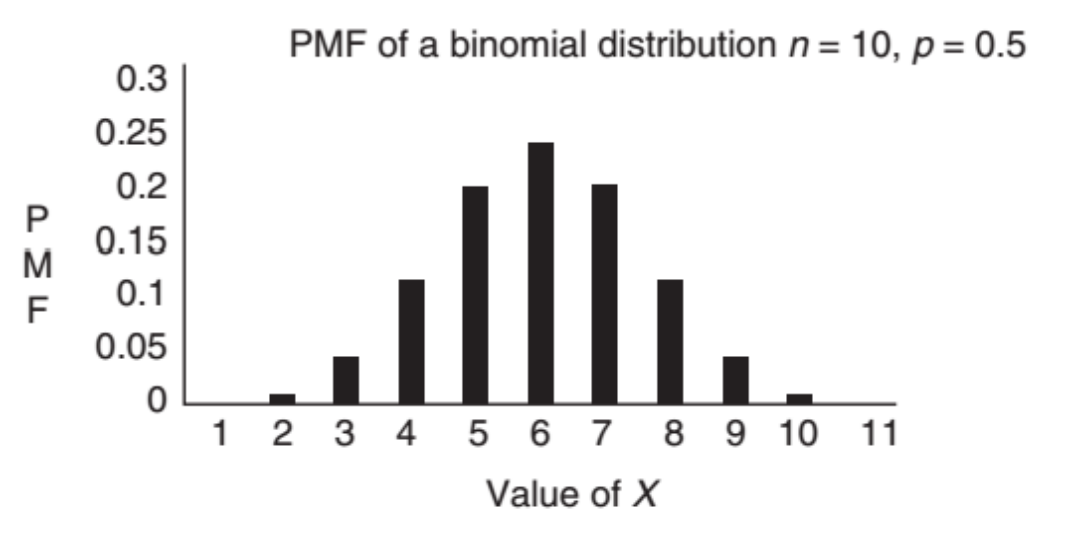
$$PMF(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \ 0 \le x \le n$$

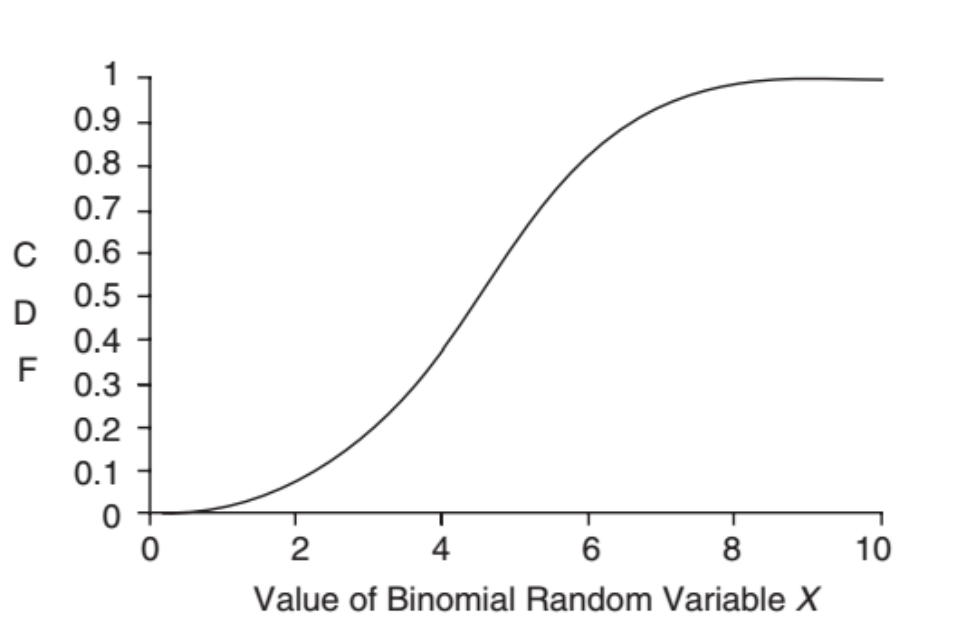
$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

- CDF of a binomial distribution function, F(a), representing the probability that the random variable X takes value less than or equal to a, is given by

$$F(a) = P(X \le a) = \sum_{k=0}^{a} P(X = k) = \sum_{k=0}^{a} \binom{n}{k} p^k (1-p)^{n-k}$$

Binomial Distribution





Mean and Variance of Binomial Distribution

Mean of a binomial distribution is given by

$$\text{Mean} = E(X) = \sum_{x=0}^{n} x \times \text{PMF}(x) = \sum_{x=0}^{n} x \times \binom{n}{x} p^x (1-p)^{n-x} = np$$

The variance of a binomial distribution is given by

$$\text{Var}(X) = \sum_{x=0}^{n} [x - E(X)]^2 \times \text{PMF}(x) = \sum_{x=0}^{n} [x - E(X)]^2 \times \binom{n}{x} p^x (1-p)^{n-x} = np(1-p)$$

Approximation of Binomial Distribution using Normal Distribution

If the number of trials (n) in a binomial distribution is large, then it can be approximated by normal distribution with mean np and variance npq, where $q = 1 - p$.

Example – Binomial Distribution

- Die Another Day (DAD) hospital recruits nurses frequently to manage high attrition among the nursing staff. Not all job offers from DAD hospital are accepted. Based on the past recruitment data, it was estimated that only 70% of offers rolled out by DAD hospital are accepted.

(a) If 10 offers are made, what is the probability that more than 5 and less than 8 candidates will accept the offer from DAD hospital?

(b) During March 2017, DAD required 14 new nurses to manage attrition. What should be the number of offers made by DAD hospital so that the average numbers of nurses accepting the offer is 14?

Example – Binomial Distribution

- Die Another Day (DAD) hospital recruits nurses frequently to manage high attrition among the nursing staff. Not all job offers from DAD hospital are accepted. Based on the past recruitment data, it was estimated that only 70% of offers rolled out by DAD hospital are accepted.

(a) If 10 offers are made, what is the probability that more than 5 and less than 8 candidates will accept the offer from DAD hospital?

(a) Probability that the number of accepted offers will be greater than 5 and less than 8 out of 10 offers is given by

$$P(5<X<8)=P(X=6)+P(X=7)=\binom{10}{6}\times 0.7^6\times 0.3^4+\binom{10}{7}\times 0.7^7\times 0.3^3=0.4669$$

(b) During March 2017, DAD required 14 new nurses to manage attrition. What should be the number of offers made by DAD hospital so that the average numbers of nurses accepting the offer is 14?

$$n\times p=14\Rightarrow n=\frac{14}{p}=\frac{14}{0.7}=20$$

That is, the hospital should make 20 offers to ensure that the expected number of accepted offers is 14.

Poisson Distribution

- In many situations, we may be interested in calculating the number of events that may occur over a period of time (or corresponding unit of measurement).

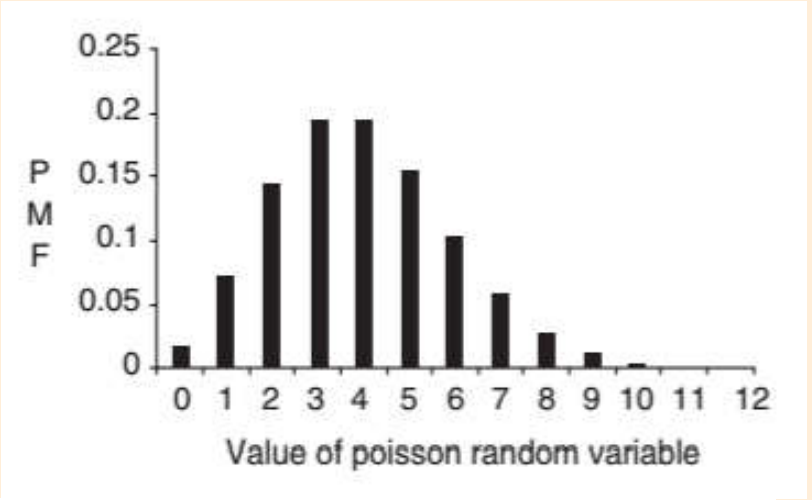
- For example, number of cancellation of orders by customers at an e-commerce portal, number of customer complaints, number of cash withdrawals at an ATM, number of typographical errors in a book, number of potholes on Bangalore roads, etc.

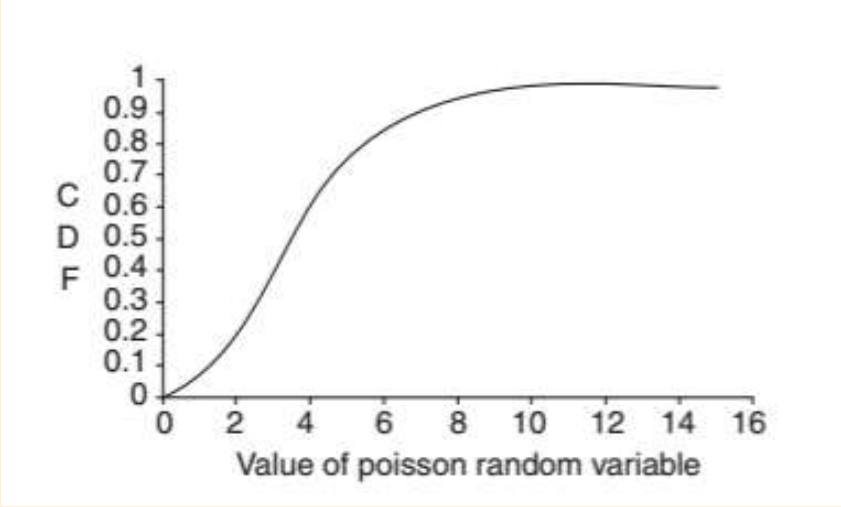
- When we have to find the probability of number of events, we use Poisson distribution.

Poisson Distribution

- PMF $P(X=k)=\frac{e^{-\lambda}\times\lambda^{k}}{k!},\quad k=0,1,2,\ldots$
- CDF: $P[X\le k]=\sum_{i=0}^{k}\frac{e^{-\lambda}\times\lambda^{i}}{i!}$
- Mean and Variance:

$$E(X)=\lambda \text{ and } \mathrm{Var}(X)=\lambda$$





UNIFORM DISTRIBUTION

Uniform distribution is one of the simplest continuous distributions. Its probability density function and cumulative distribution functions are given by

$$f(x)=\begin{cases}\dfrac{1}{b-a}, & x\in[a,b]\\ 0, & \text{otherwise}\end{cases}$$

$$F(x)=\begin{cases}0, & x<a\\ \dfrac{x-a}{b-a}, & a\le x<b\\ 1, & x\ge b\end{cases}$$

Mean and variance of uniform distribution are

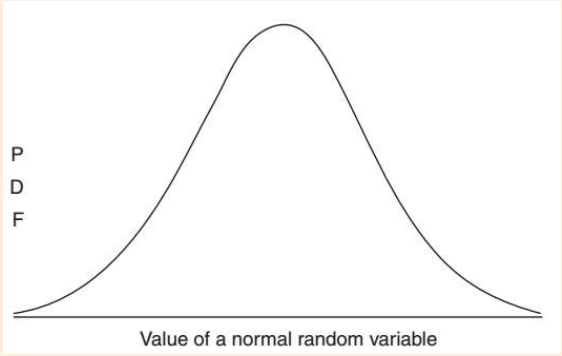
$$E(X)=\frac{1}{2}(a+b) \text{ and } \operatorname{Var}(X)=\frac{1}{12}(b-a)^2$$

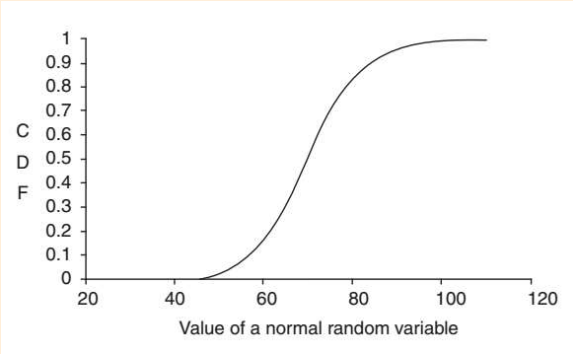
NORMAL DISTRIBUTION

- Normal distribution, also known as Gaussian distribution, is one of the most popular continuous distribution in the field of analytics especially due to its use in multiple contexts.
- Normal distribution is observed across many naturally occurring measures such as birth weight, height, intelligence, etc.
- The probability density function and the cumulative distribution function are given by :-

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \qquad -\infty < x < +\infty$$

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt, \qquad -\infty < x < +\infty$$





Properties of Normal Distribution

- Theoretical normal density functions are defined between $-\infty$ and $+\infty$.
- It is a two parameter distribution, where the parameter m is the mean (location parameter) and the parameter s is the standard deviation (scale parameter).
- All normal distributions have symmetrical bell shape around mean μ (thus it is also median)
- μ is also the mode of the normal distribution, that is, μ is the mean, median as well as the mode.
- For any normal distribution, the areas between specific values measured in terms of μ and s are given by:

Value of Random Variable	Area under the Normal Distribution (CDF)
$\mu-\sigma \le X \le \mu+\sigma$ (area between one sigma from the mean)	0.6828
$\mu-2\sigma \le X \le \mu+2\sigma$ (area between two sigma from the mean)	0.9545
$\mu-3\sigma \le X \le \mu+3\sigma$ (area between three sigma from the mean)	0.9973

Exercise:

According to a survey on use of smart phones in India, the smart phone users spend 68 minutes in a day on average in sending messages and the corresponding standard deviation is 12 minutes. Assume that the time spent in sending messages follows a normal distribution.

(a) What proportion of the smart phone users are spending more than 90 minutes in sending messages daily?

(b) What proportion of customers are spending less than 20 minutes?

(c) What proportion of customers are spending between 50 minutes and 100 minutes?

USE: NORMSDIST(x,mu,sigma,false) >> RETURNS PDF

USE: NORMSDIST(x,mu,sigma,TRUE) >> RETURNS CDF

Exercise:

According to a survey on use of smart phones in India, the smart phone users spend 68 minutes in a day on average in sending messages and the corresponding standard deviation is 12 minutes. Assume that the time spent in sending messages follows a normal distribution.

(a) What proportion of the smart phone users are spending more than 90 minutes in sending messages daily?

Mean = 68min; std=12 min

Proportion of the smart phone users are spending more than 90 minutes

= P(X>=90) = 1 - P(X<=90) = 1 – F(90) = 1 - NORMSDIST(90,68,12,TRUE)

= 0.0334

Exercise:

According to a survey on use of smart phones in India, the smart phone users spend 68 minutes in a day on average in sending messages and the corresponding standard deviation is 12 minutes. Assume that the time spent in sending messages follows a normal distribution.

(b) What proportion of customers are spending less than 20 minutes?

P(X<=20) = F(20) = NORMDIST(20,68,12,TRUE)

Exercise:

According to a survey on use of smart phones in India, the smart phone users spend 68 minutes in a day on average in sending messages and the corresponding standard deviation is 12 minutes. Assume that the time spent in sending messages follows a normal distribution.

(c) What proportion of customers are spending between 50 minutes and 100 minutes?

F(100) – F(50)

Summary

1. The concept of probability, random variables and probability distributions are foundations of data science. Knowledge of these concepts is important for framing and solving analytics problems.

2. Random variable is a function that maps an outcome of a random experiment to a real number and plays an important role in analytics since many key performance indicators used across industries are random variables

Summary

3. Basic probability concepts such as joint events, independent events, conditional probability and Bayes' theorem are useful for predicting probability of an event of importance. These concepts are used in algorithms such as association rule learning which is used in solving analytics problems such as market basket analysis and recommender systems.

4. Discrete probability distributions such as binomial distribution, Poisson distribution and geometric distribution are used for modelling discrete random variables

Summary

5. Continuous distributions such as normal distribution, chi-square distribution, t-distribution and F-distribution play an important role in hypothesis testing.