$$v_6' = \frac{52-58}{19.47} = -0.308$$

$$v_7' = \frac{56-58}{19.47} = -0.1027$$

$$v_8' = \frac{60-58}{19.47} = -0.1027$$

$$v_9' = \frac{63-58}{19.47} = 0.256$$

$$v_{10}' = \frac{70-58}{19.47} = 0.616$$

$$v_{11}' = \frac{70-58}{19.47} = 0.616$$

$$v_{12}' = \frac{110-58}{19.47} = 2.670$$

Contingency Table

→ A contingency table (or cross-tabulation) is a matrix that displays the frequency distribution of variables.

→ It is especially useful when both variables are categorical.

Example

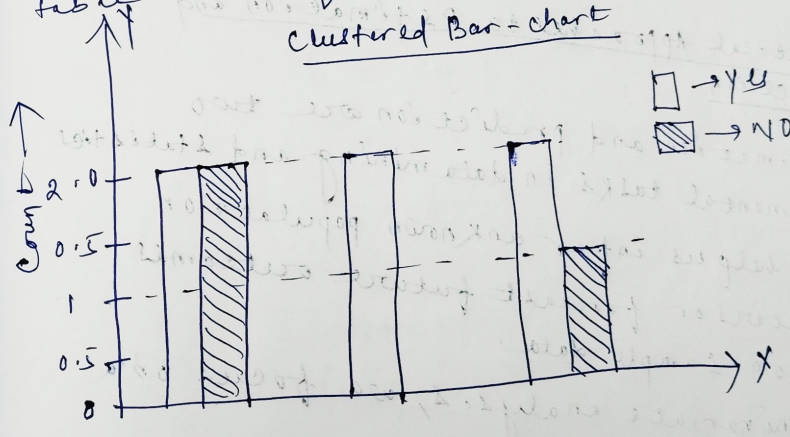
Weather	play
Sunny	Yes
Sunny	No
Overcast	yes
Rainy	yes
Rainy	No
Overcast	yes
Sunny	yes
Rainy	yes
Sunny	yes

The contingency table is

Weather	play = Yes	play = No	Total
Sunny	2	2	4
Overcast	2	0	2
Rainy	2	1	3
Total	6	3	9

cluster bar chart for the above contingency table is as follows: —

Clustered Bar-chart



Chapter-5

Univariate Statistical Analysis

Data Mining Tasks in Discovering Knowledge in Data

* Description
* Estimation
* Prediction
* classification
* clustering
* Association.

Statistical Approaches to Estimation and Prediction

* Estimation and prediction are two fundamental tasks in data mining and statistics
* They help us infer unknown population parameters or forecast future outcomes based on sample data.
* In univariate analysis, we focus on a single variable at a time.

I) Estimation

* Estimation involves using sample data to approximate population parameters.
* point-estimation provides a single guess of a parameter.

Example: Sample mean $\bar{x}$ as an estimation of population mean μ.

i.e $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

* Interval Estimation (Confidence Intervals) provides a range of plausible values for the parameter.

Example: 95% confidence interval for mean

i.e $\bar{x} \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

~~it captures uncertainty in estimation.~~

~~* proper~~ where $\bar{x}$ = sample mean

σ = population standard deviation (or sample standard deviation if σ unknown)

n = sample size

~~α~~ $\alpha = 0.05$ (as 95% confidence)

$\alpha/2 = 0.025$

$Z_{\alpha/2}$ = critical value from the standard normal distribution (for 95%, $Z_{0.025} \approx 1.96$).

* It captures uncertainty in estimation.

* Estimator properties

a) Unbiasedness: Expected value equals the true parameter.

b) Consistency: Estimator converges to the true value as $n \to \infty$

c) Efficiency: minimum variance among unbiased estimators.

II - Prediction

a) Regression - Based Prediction:

* It uses sample data to forecast outcomes.
* In univariate analysis, prediction often uses simple linear regression:

$$\hat{y} = \beta_0 + \beta_1 x$$

* Prediction Intervals quantify uncertainty in forecasts. It is wider than confidence intervals because they account for both estimation error and random variation in future values.
* Prediction error is the difference between observed and predicted values.
* Prediction error is measured using metrics like Mean Squared Error (MSE)

Example: actual = [3, 5, 7]
and predicted = [2, 6, 8]

$$MAE = \frac{|3-2| + |5-6| + |7-8|}{3}$$

$$MSE = \frac{(3-2)^2 + (5-6)^2 + (7-8)^2}{3}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad \text{(predicted)}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Application of Estimation and Prediction in Data Mining

* Estimation: Customer average spending, defect rate in manufacturing.
* Prediction: forecasting sales, predicting loan defaults.
* Both tasks underpin classification models (predicting categories) and regression models (predicting continuous values).

Confidence Interval Estimation of the Mean

* When we collect a sample from a population, we usually compute a point estimate (like the sample mean $\bar{x}$) to approximate the population mean μ.
* But a single number doesn't tell us how reliable that estimate is.
* Confidence Intervals (CIs) provide a range of possible values for the population parameter incorporating the uncertainty due to sampling.
* Sampling introduces variability as different samples will have different estimates.
* A CI expresses this variability by giving a range rather than a single point.
* Example: Instead of saying "average height = 165cm" we say "average height is between 162 - 168 cm with 95% confidence".
* CI is more informative than point estimates as a point estimate alone doesn't show precision but CI shows both the estimates and the margin of error.
* Narrow CI indicates high precision whereas wider CI indicates low precision.
* CI helps to determine whether a population parameter is significantly different from a benchmark.

Example

If a 95% CI for mean exam score is (72, 78) and the passing threshold is 70, we can be confident the true mean exceeds the threshold.

* Confidence intervals and hypothesis tests are closely related.
If a hypothesized value (e.g. $\mu = 0$) lies outside the CI, it would be rejected in a hypothesis test at the same confidence level.

* Practical Applications
a) Medical Studies: CI for drug effectiveness shows the likely range of improvement.
b) Business Analytics: CI for average customer spending helps in forecasting revenue.
c) Predictive modelling: CI around predictions quantifies reliability.

* Key Insights
1) Confidence interval (e.g. 95%) means: If we repeated sampling many times, 95% of the intervals would contain the true mean.
2) Large samples → narrower CI
↓
more precise estimates
3) Trade-off: Higher confidence → wider interval (more certainty, less precision)

Example

* A CI provides a range of possible values for a population mean (μ) based on sample data.
* It reflects: both the point estimate (sample mean $\bar{x}$) and the uncertainty due to sampling variability.
* The confidence level (e.g. 95%) indicates the proportion of intervals that would contain the true mean if repeated sampling were done.

General formula

* For a sample of size n, mean $\bar{x}$ and standard deviation s:

$$CI = \bar{x} \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

* When population standard deviation (σ) is known → use Z-distribution
* When σ is unknown → use t-distribution with $n-1$ degrees of freedom

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

Steps in Constructing a CI

1. Identify sample statistics: $\bar{x}, s, n$
2. Choose confidence level (commonly 90%, 95% or 99%)
3. Find critical value
   a) $Z_{\alpha/2}$ for large samples or known σ
   b) $t_{\alpha/2, n-1}$ for small samples or unknown σ.

4. Compute margin of error (ME):

$ME = \text{critical value} \times \frac{s}{\sqrt{n}}$

5. Construct Interval

$CI = (\bar{x} - ME, \bar{x} + ME)$

$\Gamma(k+\frac{1}{2}) = \frac{2k!}{4^k k!}\sqrt{\pi}$

Example

Suppose

→ sample mean, $\bar{x} = 50$

→ sample standard deviation, $s = 10$

→ sample size, $n = 25$

→ confidence level = 95%.

find out CI.

Ans:

$df = n - 1 = 25 - 1 = 24$

$\alpha = 1 - 0.95 = 0.05$ → as (95%)

$\alpha/2 = 0.05/2 = 0.025$

So critical value $= t_{\alpha/2, n-1}$

$= t_{0.025, (25-1)}$

$= t_{0.025 \times 24}$

So in the t-table, you would use $df = 24$ and look for the $\alpha/2 = 0.025$ column. The value is approximately

$= 2.064$.

i.e $t_{0.025, 24} \approx 2.064$

then $ME = \text{critical value} \times \frac{s}{\sqrt{n}}$

$= 2.064 \times \frac{10}{\sqrt{25}}$

$= 2.064 \times 2 = 4.128$

so $CI = (\bar{x} - ME, \bar{x} + ME)$

$= (50 - 4.128, 50 + 4.128)$

$= (45.87, 54.13)$

N.B If we are using a z-table, then for a two-tailed confidence interval, divide α by 2 ($\alpha/2 = 0.025$ for 95% interval). Look up this value in the z-table to find the corresponding critical value which is typically 1.96 for a 95% confidence interval.

How to Reduce the Margin of Error

$\text{Margin of Error (ME)} = \text{Critical value} \times \frac{s}{\sqrt{n}}$

→ So the margin of error ME contains three quantities which are as follows:-

1) $t_{\alpha/2}$, which depends on the confidence level and the sample size.

2) the sample standard deviation s, which is a characteristic of the data and may not be changed.

3) n, the sample size.

→ Thus we may decrease our margin of error in two ways, which are as follows:-

1) By decreasing the confidence level, which reduced the value of $z_{\alpha/2}$, and therefore reduced ME. But it is not recommended

2) By increasing the sample size. It is recommended. Increasing the sample size is the only way to decrease the margin of error while maintaining a constant level of confidence.

Confidence Interval Estimation of the proportion

Chapter-3 shows that 483 of 3333 customers had churned, so that an estimate of the population proportion π of all the company's customers who churn is

$$p = \frac{\text{number}}{}$$

* In data mining and predictive analytics we often estimate population parameters from sample data.
* A proportion (P) represents the portion of items in a population with a certain characteristic (e.g. proportion of customers who churn)
* The confidence interval for the population proportion is given

$$CI = p \pm z_{\alpha/2}\sqrt{\frac{p\cdot(1-p)}{n}}$$

where $p = \frac{x}{n}$ = no. of success in sample / sample size

Example $p = \frac{\text{number who churn}}{\text{sample size}} = \frac{x}{n} = \frac{483}{3333} = 0.1449$

* $z_{\alpha/2}$ → critical value from the standard normal distribution (e.g. 1.96 for 95% confidence, $\alpha = 0.05$, $\alpha/2 = 0.025$)

* $z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$ represents the margin of error.

* This z-interval for the population proportion may be used whenever both $np \geq 5$ and $n(1-p) \geq 5$ → Thumb rule.

Example

Let $n = 200$, $x = 60$, confidence = 95%

$p = \frac{x}{n} = 60/200 = 0.30$

$\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.30 \times (1-0.30)}{200}} \simeq 0.032$

$\alpha = 0.05$ since 95% confidence.

So $\alpha/2 = 0.025$, So $z_{\alpha/2} = 1.96$

So $z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} = 1.96 \times 0.032$

$= 0.0627$

So for 95% CI

$CI = p \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$

$= 0.30 \pm 0.0627 = (0.237, 0.363)$

Interpretation

We are 95% confident that the true population proportion lies between 23.7% and 36.3%.

* Wider intervals = more uncertainty
* Narrower intervals = more precision.
* Increasing sample size reduces uncertainty

Hypothesis Testing for the Mean

→ Hypothesis testing is a procedure which claims about the value of a population parameter (such as mean or standard deviation) may be considered using the evidence from the sample.

→ Two competing statements or hypotheses are crafted about the parameter value. which are as follows:—

a) The null hypothesis H_0: it refers to any hypothesis we wish to test and is denoted by H_0. It is the status quo hypothesis

b) The alternative hypothesis or research hypothesis H_a represents an alternative claim about the value of the parameter.

→ The rejection of H_0 leads to the acceptance of an alternative hypothesis, denoted by H_1.

→ The alternative hypothesis H_1 usually represents the question to be answered or the theory to be tested and thus it's specification is crucial.

→ The two possible conclusions are i) reject H_0 and b) donot reject H_0.

Example

H_0: Defendant is innocent

H_a: Defendant is guilty

→ The following table illustrates the four possible outcomes of the criminal trial with respect to the jury's decision and what is true in reality.

		Reality	
		H_0 true: Defendant did not commit crime	H_0 false: Defendant did not commit crime
Jury's Decision	Reject H_0: find defendant guilty	Type I error	Correct decision
	Do not reject H_0: find defendant not guilty	Correct decision	Type II error

* Type I error: Reject H_0 when H_0 is true. The jury convicts an innocent person.
* Type II error: Donot reject H_0 when H_0 is false. The jury acquits a guilty person.
* Correct decisions:
  → Reject H_0 when H_0 is false. The jury convicts a guilty person.
  → Donot reject H_0 when H_0 is true: The jury acquits an innocent person.

→ The probability of a Type I error is denoted as α. While the probability of a Type II error is denoted as β.

→ For a constant sample size, a decrease in α is associated with an increase in β, and vice versa.

→ In statistical analysis, α is usually fixed at some small value such as 0.05 and called the level of significance

	H_0 is true	H_0 is false
Do not reject H_0	correct decision	Type II error
Reject H_0	Type I error	correct decision

Forms of Hypothesis Testing

a) one Tailed test

* A test of any statistical hypothesis where the alternative is one sided such as $H_0: \theta = \theta_0$,

$H_a: \theta > \theta_0$ → alternative is one sided ($<$ or $>$).

or perhaps $H_0: \theta = \theta_0$,

$H_a: \theta < \theta_0$,

is called one-tailed test

* one-tailed test may be left tailed or right tailed
* $H_a: \theta < \theta_0$ is called left-tailed test.
* $H_a: \theta > \theta_0$ is called right-tailed test.

* A test of any statistical hypothesis where the alternative is two sided, such as $H_0: \theta = \theta_0$,

$H_a: \theta \neq \theta_0$

is called a two-tailed test.

where θ_0 represents the hypothesized value of θ.

---

on previous class

Estimation :→ Estimation is the process of using sample data to infer the numerical value of an unknown population parameter such as mean or proportion.

→ point Estimation → point estimation refers to the use of a single value of a statistic to estimate the associated population parameters. The observed value of the statistic is called the point estimate.

Sampling Error :→

* A sampling error is a statistical error that occurs when a sample does not represent the entire population.
* Sampling error refers to the difference between estimates obtained from a sample and the true population value.

e.g. sampling error for mean $= |\bar{x} - \mu|$

Confidence Interval

* In statistics, a confidence interval is a range of values used to estimate an unknown statistical parameter, such as a population mean.
* A confidence interval estimate of a population parameter consists of an interval of numbers produced by a point estimator together with an associated confidence level specifying the probability that the interval contains the parameter.
* The general form of confidence interval (CI)

CI = point estimate ± margin of error

Margin of Error

* A measure of the precision of the interval estimation.
* Smaller margin of error indicate greater precision.

Q. The duration customer service calls to an insurance company is normally distributed with mean 20 minute and standard deviation 5 minutes. for the following sample sizes, construct a 95% confidence interval for the population mean duration of customer service calls (a) $n=25$ (b) $n=100$ (c) $n=400$, calculate cumulative probability.

Ans:- confidence level = 95%, mean = 20 minute

significance level $\alpha = 1 - \frac{95\%}{100\%} = 0.05$

$\alpha/2 = 0.025$

cumulative probability $= 1 - \alpha/2$
$= 1 - 0.05/2 = 0.975$

a) $n = 25$

Degree of freedom $= 25 - 1 = 24$, $s = 5$, mean = 20

~~$t_{\alpha/2, n-1}$~~.

t-value at 0.025 for degree of freedom 24

i.e $t_{\alpha/2, n-1} = 2.064$

so confidence interval $= 20 \pm 2.064 \times \frac{5}{\sqrt{25}}$

$= 20 \pm 2.064$

confidence interval is $(20 - 2.064, 20 + 2.064)$

i.e $(17.936, 22.064)$ minute

b) $n = 100$

so $t_{\alpha/2, n-1} = t_{0.025, 99} = 1.984$

so CI $= 20 \pm 1.984 \times \frac{5}{\sqrt{100}} = 20 \pm 0.992$

CI $= (20 - 0.992, 20 + 0.992)$

$= (19.008, 20.992)$

c) $n = 400$, $t_{\alpha/2, n-1} = t_{0.025, 399} = 1.966$

so CI $= 20 \pm 1.966 \times \frac{5}{\sqrt{400}} = 20 \pm 0.4915$

CI $= (20 - 0.4915, 20 + 0.4915)$

$= (19.5085, 20.4915)$

Q. calculate and interpret margin of error for the previous question for $n = 25$

Ans:- Margin of error $= t_{\alpha/2, n-1} \times \left(\frac{s}{\sqrt{n}}\right)$

$= 2.064 \times \frac{5}{\sqrt{25}} = 2.064$

Interpretation

We expect that the true population mean will be located no more and no less than 2.064 minutes from the sample mean of 20 minutes.

Forms of Hypothesis Testing

→ A common treatment of hypothesis testing for the mean is to restrict the hypothesis to the following three forms:

i) Left-tailed test, actual mean / hypothesized mean

$H_0 : \mu \geq \mu_0$ versus $H_a : \mu < \mu_0$

↳ always take alternative hypothesis to know the difference.

ii) Right-tailed test

$H_0 : \mu \leq \mu_0$ versus $H_a : \mu > \mu_0$

iii) Two-tailed test

$H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$

where μ_0 represents hypothesized value of μ.

* When the sample size is large or the population is normally distributed, the test statistics (more than 30)

$$t_{data} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

follows a t-distribution with $(n-1)$ degree of freedom.

* t_{data} — no. of standard error above/below the hypothesized mean μ_0 that the sample mean $\bar{x}$ resides.

* $s/\sqrt{n}$ — standard error

When the value of t_{data} is extreme, this indicates a conflict between the null-hypothesis and the observed data.

* As the data represents empirical evidence whereas the null hypothesis represents merely a claim, such conflicts are resolved in favour of the data, so that when t_{data} is extreme, the null hypothesis H_0 is rejected. This extremeness is measured using the p-value.

* The p-value is the probability of observing a sample statistic (such as $\bar{x}$ or t_{data}) at least as extreme as the statistic actually observed, if we assume that the null hypothesis is true.

* As the p-value represents a probability, its value must always fall below 0.1.

Table: How to calculate p-value

Form of Hypothesis Test	p-value
1) Left-tailed test $H_0 : \mu \geq \mu_0$ versus $H_a : \mu < \mu_0$	$P(t < t_{data})$
2) Right-tailed test $H_0 : \mu \leq \mu_0$ versus $H_a : \mu > \mu_0$	$P(t > t_{data})$
3) Two-tailed test $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$	if $t_{data} < 0$, then p-value $= 2 \times P(t < t_{data})$ if $t_{data} > 0$, then p-value $= 2 \cdot P(t > t_{data})$

* A small p-value indicate conflict between the data and the null hypothesis. Thus we will reject H_0 if the p-value is small. We consider the p-value to be small if it is less than α.

Reject H_0 if the p-value $< \alpha$

e.g. H_0 : mean = 2.4 , H_1 : mean $\neq$ 2.4
$\bar{x} = 1.607$, $s = 1.892$, $n = 28$, $\mu_0 = 2.4$
$\alpha = 0.05$
perform hypothesis testing

Ans:- H_0 : mean = 2.4, H_1 : mean $\neq$ 2.4

so it is two tailed test

$$t_{data} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{1.607 - 2.4}{1.892/\sqrt{28}}$$

$$= \frac{-0.793}{0.359}$$

$$= -2.2178$$

$\downarrow$

$t_{data} < 0$

So p-value $= 2 \times P(t < t_{data})$

$\downarrow$ since t table for $df = n-1 = 27$ and 2.2178

take average of 27 + 2.25 and next

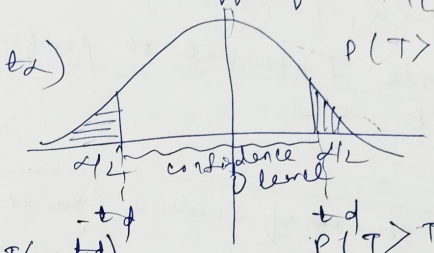
$= 2 \times 0.0178 = 0.035$

$\alpha = 0.05$

Since p-value $< \alpha$, so manual hypothesis H_0 is rejected.

(always our aim is to reject null hypothesis and accept alternate hypothesis)

Note:- mean no differ from 2.4 (H_1 is true)



$P(T < -2.2178) = P(T > 2.2178)$

Q-1 A random sample of 100 recorded deaths in the United States during the past years showed an average life span of 71.8 years. Assuming a population standard deviation of 8.9 years, does this seem to indicate that the mean life span today is greater than 70 years? Use 0.05 level of significance.

Ans:- H_0 : $\mu \leq 70$ | $\alpha = 0.05$, $n = 100$ | $\sigma = 8.9$, $\bar{x} = 71.8$

H_1 : $\mu > 70$

Since H_1 containing greather than, so it is right tailed.

Since no. of samples = 100 (>30) and given population SD, so we have to use Z-distribution

$$\text{So } Z_{data} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{71.8 - 70}{8.9/\sqrt{100}} = \frac{1.8}{8.9/10} = \frac{1.8}{0.89} \simeq 2.12359$$

So p-value $= p(Z > Z_{data}) = P(Z > 2.12359) = 0.983$ (because Z-distribution gives always left)

But for right tail we need $Z > Z_{data}$

so p-value $= P(Z > Z_{data}) = 1 - 0.983 = 0.017$

Since p-value $< \alpha$, so H_0 is rejected

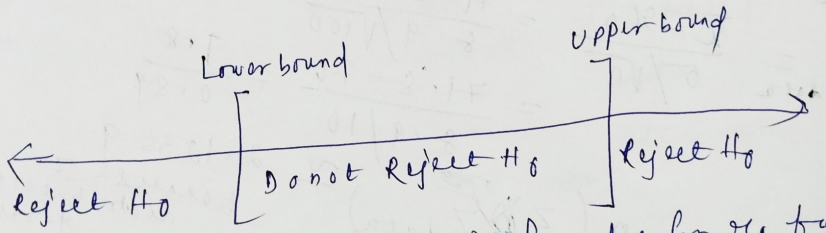
skip for students

Using confidence Interval to perform Hypothesis Test

p-value	Strength of Evidence Against H_0
p-value ≤ 0.001	Extremely strong evidence
$0.001 <$ p-value ≤ 0.01	Very strong evidence
$0.01 <$ p-value ≤ 0.05	Solid evidence
$0.05 <$ p-value ≤ 0.10	Mild evidence
$0.10 <$ p-value ≤ 0.15	Slight evidence
$0.15 <$ p-value	No evidence

Confidence Level $100(1-\alpha)\%$	Level of significance (α)
90%	0.10
95%	0.05
99%	0.01

$\therefore \alpha = 1 - 0.95 = 0.05$



* If a certain hypothesized value for μ_0 falls outside the confidence level with confidence level $100(1-\alpha)\%$, then the two-tailed hypothesis test with level of significance α will reject H_0 for that value of μ_0.

* If the hypothesized value for μ_0 falls inside the confidence interval with confidence level $100(1-\alpha)\%$ then the two-tailed hypothesis test with level of significance α will not reject H_0 for that value of μ_0

e.g. 95% confidence interval for the population mean no. of customer service calls for all customers who have the international plan and voice mail plan and who have more than 220 minutes of day use is

(lower bound, upper bound) = (0.875, 2.339)

$\alpha = 0.05$

For any no. of possible values of μ_0

a) 0.5 b) 1.0 c) 2.4

a) $H_0: \mu = 0.5$
$H_1: \mu \neq 0.5$

b) $H_0: \mu = 1.0$
$H_1: \mu \neq 1.0$

c) $H_0: \mu = 2.4$
$H_1: \mu \neq 2.4$

use α to find CI. then see the given μ_0 values inside the CI or not. if inside $\rightarrow$ accept otherwise $\rightarrow$ reject

We construct the 95% confidence interval and place the hypothesized value μ_0 on the number line

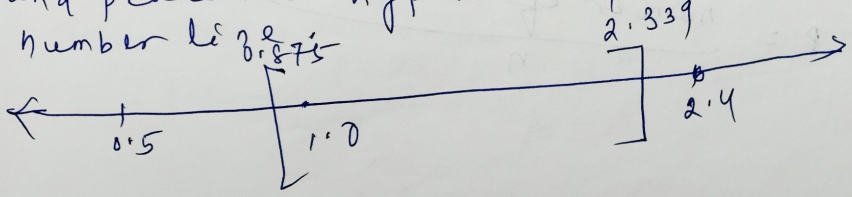


Table: conclusion for three hypothesis tests using confidence level

μ_0	Hypothesis with $\alpha = 0.05$	position in relation to CI	conclusion
0.5	$H_0: \mu = 0.5$ Vs. $H_a: \mu \neq 0.5$	outside	Reject H_0
1.0	$H_0: \mu = 1.0$ Vs. $H_a: \mu \neq 1.0$	Inside	Do not Reject H_0
2.4	$H_0: \mu = 2.4$ Vs. $H_a: \mu \neq 2.4$	outside	Reject H_0

↓ as boundary is 0.877 to 2.339

Hypothesis Testing for the Proportion

Hypothesis tests may also be performed about the population proportion π. The test statistic is

$$Z_{data} = \frac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$$

→ population proportion
p → sample proportion

where π_0 is the hypothesized value of π and p is the sample proportion.

$$p = \frac{\text{number of successes}}{n}$$

Table: Hypotheses and p-values for Hypothesis Tests about π

Hypotheses with $\alpha = 0.05$	p-value
Left tailed test: $H_0: \pi \geq \pi_0$ Vs. $H_a: \pi < \pi_0$	$P(Z < Z_{data})$
Right tailed test: $H_0: \pi \leq \pi_0$ Vs. $H_a: \pi > \pi_0$	$P(Z > Z_{data})$
Two tailed test: $H_0: \pi = \pi_0$ Vs. $H_a: \pi \neq \pi_0$	if $Z_{data} < 0$, then p-val $= 2 \cdot P(Z < Z_{data})$ if $Z_{data} > 0$, then p-val $= 2 \cdot P(Z > Z_{data})$

Example

483 of 3333 customers have changed

$$p = \frac{\text{n.o of customers who churn}}{\text{sample size}} = \frac{x}{n} = \frac{483}{3333} = 0.449$$

Suppose we would like to test using level of significance $\alpha = 0.10$ whether π differs from 0.15. The hypotheses are:

$H_0: \pi = 0.15$ versus $H_a: \pi \neq 0.15$

The test statistic is

$$Z_{data} = \frac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} = \frac{0.1449 - 0.15}{\sqrt{0.15(1-0.15)/3333}}$$

$$= -0.8246$$

As $Z_{data} < 0$, the p-value $= 2 \times P(Z < Z_{data})$

$= 2 \times P(Z < -0.8246)$ ← since p-value gives left area we got left. because -ve

$= 2 \times 0.2048$

$= 0.4096$.

p-value $= 0.4096 > \alpha (= 0.10)$

fail to reject H_0

Next Solve 4 Problems from exercise

Q. of 1000 customers who received promotional materials for a marketing campaign, 100 responed to the promotion. for the following confidence levels, construct a confidence interval for the population proportion who would respond to the promotion.

a. 90% b. 95% c. 99%

Ans:- $p = \frac{\text{respondents}}{\text{sample size}(n)} = \frac{100}{1000} = 0.10$ ← population proportion

since population proportion is given also sample size ≥ 30, so population is 1000), so z-distribution will be used

a) As confidence = 90%

So $\alpha = 1 - \frac{90}{100} = 1 - 0.9 = 0.10$, $\alpha/2 = 0.10 = 0.05$

so critical value $= Z_{\alpha/2} = Z_{0.05} \approx 1.645$

So Margin of Error (ME) = critical value $\times \sqrt{\frac{p(1-p)}{n}}$

$= 1.645 \times \sqrt{\frac{0.10(1-0.10)}{1000}}$

$= 1.645 \times 0.03$

$= 0.0494$

so the confidence Interval (CI)

$= p \pm ME$

$= 0.10 \pm 0.494 = (0.0506, 0.1494)$

b) 95% confidence level

$\alpha = 1 - \frac{95}{100} = 1 - 0.95 = 0.05$

So critical value $= Z_{\alpha/2} = Z_{0.025} \approx 1.960$

Margin of Error (ME) = critical value $\times \sqrt{\frac{p(1-p)}{n}}$

$= 1.960 \sqrt{\frac{0.10 \times (1-0.10)}{1000}}$

$= 0.0588$

So confidence Interval (CI)

$= p \pm ME = 0.10 \pm 0.0588 = (0.0412, 0.1588)$

c) Similarly for 99% confidence

Q. A sample of 100 donors to a charity has a mean donation amount of $55 with a sample standard deviation of $20. Test using $\alpha = 0.05$ whether the population mean donation amount exceeds $50.

a. provide the hypotheses. state the meaning of it

b. what is the rejection rule.

c. what is the conclusion and interpret the conclusion.

Ans:-

Ans:- a) The hypotheses are

i) Null hypothesis H_0: The population mean μ is less than or equal to the hypothesized mean $\mu_0 = \$50$

ii) Alternative hypothesis H_a: The population mean μ is greater than the hypothesized mean $\mu_0 = \$50$.

So, it is a right-tailed test

b) Rejection Rule

We will reject the null hypothesis H_0 if the p-value for the test statistics is less than the significance level α. For our problem, we would reject the null hypothesis if the p-value $< \alpha = 0.05$

c) Since sample of donors and sample standard deviation given

$$t_{data} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

$\bar{x} = 55$
$\mu_0 = 50$
$s = 25$
$n = 100$

$$= \frac{55 - 50}{25/\sqrt{100}} = 2$$

So for right-tailed test,

p-value $= 2p(t > t_{data})$

$= P(t > 2)$ with 99 degree of freedom

$= 0.02412 \rightarrow$ (Solid evidence against the null hypothesis)

Since p-value $< \alpha = 0.05$, so the null-hypothesis is rejected.

— Hypothesis test for CI

25 to 50 not in exam

Refer to the first confidence interval you calculated for the population mean duration of customer service calls. Use this confidence interval to test whether this population mean differs from the following values, using level of significance $\alpha = 0.05$

a. 15 minutes b. 20 minutes c. 25 minutes

Ans:- We can leverage a confidence interval to perform hypothesis testing as it is complementary to a two-tailed test. Our hypotheses for this type of test would be formulated as:

$H_0: \mu = \mu_0$, $H_a: \mu \neq \mu_0$

Since our confidence level is 95%, our significance level α for these tests is calculated as follows:—

$\alpha = (1 - 95/100) = 0.05$

The first confidence interval we calculated for population mean duration of customer call was in Exercise 17 a and was based on a sample size $n = 25$, sample mean of 20 minutes and sample standard deviation of 5 minutes and calculated as follows:—

CI $= 20 \pm 2.064 = (17.936, 22.064)$ minutes

a. 15 minutes

Since 15 minutes lies outside of (i.e below), the confidence interval, we would reject the null hypothesis for this test and conclude that the population mean duration is not 15 minutes.

b. 20 minutes

Since 20 minutes lies within the confidence interval, we would fail to reject the null—

hypothesis for this test and conclude that there is not enough evidence to indicate that the population mean duration is not 20 minutes.

c. 25 minutes

since 25 minutes fall outside (i.e. above) the confidence interval, we would reject the null hypothesis for this test and conclude that the population mean duration is not 15 minutes

— hypothesis for proportion

26. In a sample of 1000 customers, 240 churned when the company raised rates. Test whether the population proportion of churners is less than 25%, using level of significance $\alpha = 0.01$

Ans! - $p = \frac{x}{n} = \frac{\text{respondents}}{\text{sample size}} = \frac{240}{1000} = 0.24$

So the ~~Alternate hypothesis~~ : Null hypothesis!
population proportion $\pi \geq 0.25$

Alternative Hypothesis: Ha : population proportion $\pi < 0.25$

→ So it is one-tailed test
Since, it is given population proportion.
So Z-distribution will be used

→ So $Z_{data} = \frac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} = \frac{0.24 - 0.25}{\sqrt{\frac{0.25 \times 0.75}{1000}}}$

$= \frac{-0.01}{0.01369} = -0.7305$

So for ~~or~~ left tailed test
p-value = $P(Z < Z_{data}) = P(Z < -0.7305)$
$= 0.232542$

since the p-value $> \alpha = 0.01$, we fail to reject the null hypothesis concluding that there is no evidence to dispute it.
In other words, there is no evidence to indicate that the population proportion of churners will remain less than 25%

Chapter-6

Multivariate Statistical

→ So far we have discussed inference methods for one variable at a time.

→ Data Analysts are also interested in multivariate inferential methods, where the relationship between two variables or between one target variable and a set of predictor variables are analyzed.

→ In bivariate analysis, we have two independent samples and wish to test for significant differences in the mean or proportion of the two samples

→ In this chapter, we illustrate how the data is partitioned into training dataset and a test dataset for cross-validation purpose.

→ Data miners can use the hypothesis tests shown here to determine whether significant differences exist between the means of various variables in the training and test data sets.

→ If such differences exist, then cross-validation is invalid, because the training data set is non-representative of the test dataset.

Types of Multivariate Analysis (Statistics)

* For a continuous variable, use the two-sample t-test for the difference in means.
* For a flag variable, use the two-sample Z-test for the differences in proportions
* For a multinomial variable, use the test for the homogeneity of proportions.

Two-Sample T-test for Differences in Means

→ To test for the differences in population means, we use the following test statistic

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

which follows an approximate t distribution with degrees of freedom the smaller of $n_1 - 1$ and $n_2 - 1$, whenever either both populations are normally distributed or both samples are large.

Example

We partitioned the churn data set into a training set of 2529 records and a test set of 804 records. We would like to assess the validity of the proportion by testing whether the population mean number of customer service calls differs between the two data sets. The summary statistics is

Data Set	Sample Mean	Sample SD	Sample Size
Training set	$\bar{x}_1 = 1.5714$	$s_1 = 1.3126$	$n_1 = 2529$
Test set	$\bar{x}_2 = 1.5361$	$s_2 = 1.3251$	$n_2 = 804$

So the hypothesis are

$H_0: \mu_1 = \mu_2$ Vs. $H_a: \mu_1 \neq \mu_2$

The test statistic is

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

$$= \frac{1.5714 - 1.5361}{\sqrt{(1.3126)^2/2529 + ((1.3251)^2/804)}}$$

$$= 0.6595$$

Since the given test is two-tailed

So p-value $= 2P(t > t_{data})$ as $t_{data} > 0$

$$= 2P(t > 0.6595).$$

$$= 0.5098$$

Since the p-value is large, there is no evidence that the mean no. of customer service calls differs between the training dataset and the test data set, for this variable. at least, the partition seems valid.

Two-Sample Z-test for Differences in proportions

→ Of course not all variables are numeric, like customer service calls.

→ So for the flag variables with 0/1 variable like membership in the voice Mail plan and we wish to test whether the proportions of records with value 1 differ between the training data set and test data set?

→ So we could turn to the two-sample Z-test for the difference in proportions.

↳ The test statistic is

$$Z_{data} = \frac{p_1 - p_2}{\sqrt{p_{pooled}(1 - p_{pooled})((1/n_1) + (1/n_2))}}$$

where $p_{pooled} = \frac{x_1 + x_2}{n_1 + n_2}$ and x_i and p_i represents the number of and proportion of records with value 1 for sample i, respectively.

For example, our partitioned resulted in $x_1 = 707$ of $n_1 = 2529$ customers in the training set belonging to the voice Mail plan, while $x_2 = 215$ of $n_2 = 804$ customers in the test set belonging, so that $p_1 = \frac{x_1}{n_1}$

$$= \frac{707}{2529} = 0.2796$$

$$p_2 = \frac{x_2}{n_2} = \frac{215}{804} = 0.2674$$

$$p_{pooled} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{707 + 215}{2529 + 804} = 0.2766$$

The hypothesis are:-

$H_0: \pi_1 = \pi_2$ Vs. $H_a: \pi_1 \neq \pi_2$

The test statistic is

$$Z_{data} = \frac{p_1 - p_2}{\sqrt{p_{pooled} \cdot (1 - p_{pooled})((1/n_1) + (1/n_2))}}$$

$$= \frac{0.2796 - 0.2674}{\sqrt{0.2766 \times 0.7234\ ((1/2529) + (1/804))}} = 0.6736$$

The p-value is
p-value $= 2 \times P(Z > Z_{data})$ (for two tailed $Z_{data} > 0$)

$= 2 \times P(Z > 0.6736)$

$= 0.5006$.

Since p-value $> \alpha$, ~~so null hypothesis~~ So we fail to reject the null hypothesis. So there is no evidence that the proportion of voice mail plan members differed between the training and test data sets. For this variable, the partition is valid.

Test for the Homogeneity of Proportions

Multinomial data is an extension of binomial data to $K > 2$ categories. For example, suppose a multinomial variable marital status takes the values married, single and other. Suppose we have a training set of 1000 people and a test set of 250 people, with the frequencies shown in Table

Dataset	Married	Single	other	total
Training set	410	340	250	1000
Test set	95	85	70	250
Total	505	425	320	1250

To determine whether significant differences exist between the multinomial proportions of the two data sets, we could turn to the test for the homogeneity of proportions. The hypothesis are:

H_0: $p_{married,training} = p_{married,test}$,

$p_{single,training} = p_{single,test}$,

$p_{other,training} = p_{other,test}$

H_a: At least one of the claims in H_0 is wrong.

To determine whether these observed frequencies represent proportions that are significantly different for the training and test data sets, we compare these observed frequencies with the expected frequencies that we would expect if H_0 were true. For example, to find the expected frequency for the number of married people in the training set, we (i) find the overall proportion of married people in both the training and test sets, $\frac{505}{1250}$ and (ii) we multiply this overall proportion by the number of people in the training set, 1000, giving us the expected proportion of married people in the training set to be

$$\text{Expected frequency} = \frac{(\text{row total})(\text{column total})}{\text{grand total}}$$

so Expected frequency$_{married,training}$

$$= \frac{1000 \times 505}{1250} = 404$$

We use the overall proportion of 1 (i) because H_0 states that the training and test proportions are equal.

Applying this formula to each cell in the table gives us the table of expected frequencies as follows:-

Table: Expected Frequency

Data set	Married	single	other	Total
Training set	404	340	256	1000
Test set	101	85	64	250
Total	505	425	320	1250

→ The observed frequencies (O) and the expected frequencies (E) are compared using a test statistic from the χ^2 (chi-square) distribution

$$\chi^2_{data} = \sum \frac{(O-E)^2}{E}$$

Cell	Observed frequency	Expected frequency	$\frac{(Obs - Exp)^2}{Exp}$
Married, training	410	404	$(410-404)^2/404 = 0.09$
Married, test	95	101	$(95-101)^2/101 = 0.36$
single, training	340	340	$(340-340)^2/340 = 0$
single, test	85	85	$(85-85)^2/85 = 0$
other, training	250	256	$(250-256)^2/256 = 0.14$
other, test	70	64	$\frac{(70-64)^2}{64} = 0.56$
			Sum $= \chi^2_{data} = 1.15$

p-value = is the area to the right of χ^2_{data} under the χ^2 curve with degrees of freedom equal to (number of rows − 1)(number of columns − 1)

$= (1)(2) = 2$

so p-value $= P(\chi^2 > \chi^2_{data})$

$= P(\chi^2 > 1.15) = 0.5627$

with df = 2

since p-value $> \alpha$ and is very large, so there is no evidence that the observed frequencies represent proportions that are significantly different for the training and test data sets. In other words, for this variable, the partition is valid.

Chi-Square Test for Goodness of Fit of Multinomial Data

→ Use chi-square goodness-of-fit for one categorical variable to see if its observed counts match an expected distribution

→ Use a chi-square test of Homogeneity when you have one categorical variable but across two or more different groups/populations, testing if proportions for that variable are the same (homogenous) on each group.

Example

Suppose a multinomial variable marital status takes the values married, single and other and other, and suppose that we know that 40% of the individuals in the population are married, 35% are single and 25% report another marital status.

→ We are taking a sample and would like to determine the sample is representative of the population. We could turn to the χ^2 (chi-square) goodness of fit test.

→ The hypothesis for this χ^2 goodness of fit test would be as follows: —

H_0 : $p_{married} = 0.40$, $p_{single} = 0.35$, $p_{married} = 0.25$

H_a : At least one of the proportions in H_0 is wrong

→ Our sample of size $n = 100$, yields the following observed frequencies (represented by the letter O)

$O_{married} = 36$, $O_{single} = 35$, $O_{other} = 29$

→ To determine whether these counts represent proportions that are significantly different from those expressed in H_0, we compare these observed frequencies with the expected frequencies that we would expect if H_0 were true. If H_0 were true, then we would expect 40% of our sample of 100 individuals to be married, that is the expected frequency for married is

$E_{married} = n \cdot p_{married} = 100 \times (0.40) = 40$

Similarly,

$E_{single} = n \cdot p_{single} = 100 \times 0.35 = 35$

$E_{other} = n \cdot p_{other} = 100 \times 0.25 = 25$

These frequencies are compared using the test statistic: —

$$\chi^2_{data} = \sum \frac{(O-E)^2}{E}$$

→ Again, large differences between the observed and expected frequencies, and thus a large value for χ^2_{data}, will lead to a small p-value and a rejection of the null hypothesis.

→ The test statistic is calculated as follows: —

Marital Status	Observed Frequency	Expected Frequency	$\frac{(Obs - Exp)^2}{Exp}$
Married	36	40	$\frac{(36-40)^2}{40} = 0.4$
single	35	35	$\frac{(35-35)^2}{35} = 0$
other	29	25	$\frac{(29-25)^2}{25} = 0.64$
			Sum = $\chi^2_{data} = 1.04$

→ The p-value is the area to the right of χ^2_{data} under the χ^2 curve with $K-1$ degrees of freedom, where K is the no. of categories (here $K = 3$)

p-value $= P(\chi^2 > \chi^2_{data}) = P(\chi^2 > 1.04)$

$= 0.5945$

Hence the p-value is very large, so there is no evidence that the observed frequencies represent proportions that differ significantly from those in the null hypothesis.
→ In other words, our sample is representative of the population.

Chapter - 7

Preparing to Model the Data

Supervised and Unsupervised Learning

→ ~~Supervised~~
→ Supervised and unsupervised learning are two main types of machine learning/data mining method.
→ In supervised learning, the model is trained with labeled data where each input has a corresponding output.
→ On the other hand, unsupervised learning involves training the model with unlabeled data which helps to uncover patterns, structures or relationships within the data without predefined outputs.

Basis	Supervised Learning	Unsupervised Learning
1) Definition	Supervised learning algorithm trains data, where every input has a corresponding output.	Unsupervised learning algorithm finds patterns in data that has no predefined labels.
2) Goal	The goal of supervised learning is to predict or classify based on input features.	The goal of unsupervised learning is to discover hidden patterns, structures and relationships.
3) I/p Data	Labeled: input data with corresponding o/p labels.	Unlabeled: input data is raw and unlabeled.
4) Human supervision	Supervised learning algorithm needs human supervision to train the model.	Unsupervised learning algorithms do not need any kind of supervision to train the model.
5) Tasks	Regression, classification	clustering, association and dimensionality reduction.
6) Complexity	Supervised methods are computationally simple.	Unsupervised methods are computationally complex.
7) Algorithms	Linear Regression, KNN, Decision Tree, Naive Bayes, SVM	K-Means clustering, DBSCAN, Autoencoders
8) Accuracy	Supervised methods are highly accurate.	Unsupervised methods are less accurate.
9) Applications	Image classification, sentiment analysis, recommendation systems	customer segmentation, Anomaly detection

Statistical Methodology and Data Mining Methodology

→ Statistical methodology focuses on confirming theories with structured data using probability/math (mean, regression) for insights, while data mining is an exploratory process finding hidden patterns in massive, often messy data using algorithms (clustering, classification) for prediction, with statistics forming the core of many data mining techniques.

Basis	Data Mining	Statistics
① Data Type	Works with any kind of data (numeric, text etc)	Works mainly with quantitative (numeric) data
② Goal	Derives insights, often through prediction.	Deduces conclusions based on probability distributions.
③ Approach	Exploratory: focused on discovering hidden patterns.	confirmatory: focuses on validating hypotheses.
④ Domain Knowledge	Heuristic (rules of thumb) plays an important role.	Relies on mathematical evidence and probability.
⑤ Data Collection	Emphasis on working with existing data not collection.	Focuses on data collection and cleaning.
⑥ Hypothesis	No a priori hypothesis	A priori hypotheses

Cross validation

→ Cross validation is a technique used to check how well a machine learning model performs on unseen data while preventing overfitting.

→ It works by:—

* Splitting the dataset into several parts.
* Training the model of some parts and testing it on the remaining part.
* Repeating this resampling process multiple times by choosing different parts of the dataset.
* Averaging the results from each validation step to get the final performance

Types of Cross validation

① **Holdout validation**

→ In holdout validation method typically 50% data is used for training & 50% for testing making it simple and quick to apply.

→ The major drawback of this method is that only 50% data is used for training the model which may miss important patterns in the other half which leads to high bias.

② **LOOCV (Leave One Out Cross validation)**

→ ~~on the entire dataset except for one data point~~

→ In this method, the model is trained on the entire dataset except for one data point which is used for testing. This process is repeated for each data point in the dataset.

→ All data points are used for training resulting in low bias.

* Testing on a single data point can cause high variance especially if the point is an outlier.
* It can be very time consuming for large datasets as it requires one iteration per data point.

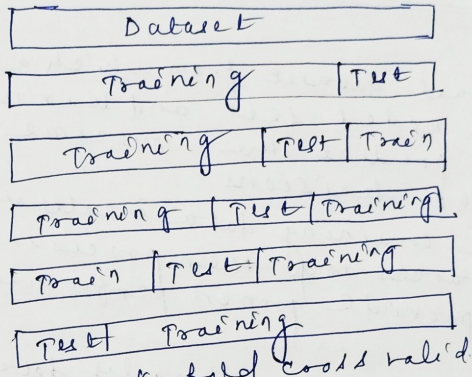
(3) Stratified Cross Validation

→ It is a technique that ensures each fold of the cross validation process has the same class distribution as the full dataset.
→ This is useful for imbalanced datasets where some classes are under-represented.
→ The dataset is divided into K-folds, keeping class proportions consistent in each fold.
→ In each iteration, one fold is used for testing and the remaining folds for training.
→ This process is repeated K times so that each fold is used once as the test set.
→ It helps classification models generalize better by maintaining balanced class representation.

(4) K-fold Cross Validation

→ K-fold cross validation splits the dataset into K equal-sized folds.
→ The model is trained on K-1 folds and tested on the remaining fold.
→ This process is repeated K-times each time using a different fold for testing.

Note:→ It is always suggested that the value of K - should be 10 as the lower value of K takes towards validation and higher value of K leads to LOOCV method.



* 1st iteration → test 20% of data is used for testing remaining 80% is used for training
* 2nd iteration – 2nd 20% of data is used for testing & the remaining data is used for training
* This process continues until each fold has been used once as the test set.

Table – Suggested hypothesis tests for validating different types of target variables.

Type of Target variable	Hypothesis Test
Continuous	Two sample t-test for difference in means
	Two sample z-test for difference in proportion
Flag	Test for homogeneity of proportions
Multinomial	

~~Testing Set~~

Training set, Test set and Validation Test

Training Set:

→ This is the actual dataset from which a model trains i.e the model sees and learns from this data to predict the outcome or to make the right decisions.

→ Most of the training data is collected from several resources and then processed and organized to provide proper performance of the model.

→ Type of training data hugely determines the ability of the model to generalize i.e the better the quality and diversity of training data, the better will be the performance of the model.

Testing Set

→ The dataset is independent of the training set but has a somewhat similar type of probability distribution of classes and is used as a benchmark to evaluate the model, used only after the training of the model is complete.

→ Testing set is usually a properly organized dataset having all kinds of data for scenarios that the model would probably be facing when used in the real world.

→ often the validation and testing set combined is used as a testing set which is not considered as a good practice.

Validation Set

→ The validation set is used to fine tune the hyper-parameters of the model and is considered a part of the training of the model.

→ The model only sees this data for evaluation but does not learn from this data, providing an objective unbiased evaluation of the model.

Over fitting and Under-fitting

* Machine learning models aim to perform well on both training data and new unseen data, and is considered "good" if: —

i) It learns patterns effectively from the training data.

ii) It generalizes well to new unseen data.

iii) It avoids memorizing the training data (overfitting) or failing to capture relevant patterns (underfitting).

→ To evaluate how well a model learns and generalizes, we monitor it's performance on both the training data and a separate validation or test dataset which is often measured by its accuracy or prediction errors.

→ However, achieving this balance can be challenging.

→ Two common issues that affect a model's performance and generalization ability are overfitting and underfitting.

→ These problem are major contributions to poor performance in machine learning models.

Bias and variance in Machine Learning

→ Bias and variance are two key sources of error in machine learning models that directly impact their performance and generalization ability.

→ **Bias** :→

* Bias is the error that happens when a machine learning model is too simple and does not learn enough details from the data.
* E.g : It is like assuming all birds can only be small and fly, so the model fails to recognize big birds like ostriches or penguins that can't fly and get biased with predictions.
* These assumptions make the model easier to train but may prevent it from capturing the underlying complexities of the data.
* High bias typically

Resampling

→ Resampling is a technique used to adjust the class distribution of an imbalanced dataset which helps to prevent machine learning models from being biased towards the majority class.

→ This is typically achieved through oversampling (increasing the minority class) or undersampling (decreasing the majority class).

→ General equation for resampling is

Rare + x = p (record + x)

where, x → required no. of resampled records.

p → desired proportion of rare values in the balanced dataset.

record → no. of records in the unbalanced data set.

rare → current no. of rare target values.

Note :→ The test dataset should never be balanced.

Challenges with Oversampling

→ oversampling aims to increase the no. of examples in the minority class.

→ the associated challenges are :—

(i) Overfitting : simple random oversampling can lead to the model overfitting the minority class, as it memorizes the repeated instances rather than learning general pattern.

(i) Introduction to noise

→ Advanced techniques like SMOTE (Synthetic Minority Oversampling Technique) generate synthetic examples via interpolation.
→ If the data is inherently noisy or the class overlap, these methods may introduce noisy or unrealistic samples that do not accurately represent the real world minority class distribution, potentially degrading performance.

(ii) Increased Computation Cost

→ Creating a larger training dataset increases the memory usage and the computational time required to train the model.

(iv) False Confidence

→ Models trained on synthetically balanced data might appear to have high performance metrics during validation but can fail when applied to real world imbalanced scenarios if the synthetic data did not accurately capture the true data distribution.

Challenges with Undersampling

→ Undersampling aims to reduce the no. of samples in the minority class.
→ The associated challenges are: —

i) Loss of Information : The most significant drawback of undersampling is the potential loss of valuable information.
→ The loss of valuable and informative data from the majority class.
→ Randomly removing instances may discard crucial patterns that are important for the model to learn the decision boundary, leading to underfitting.

ii) Risk of Bias : If the removed majority samples were critical or represented unique sub patterns, the resulting model may be biased and fail to generalize well to unseen data.

iii) Poor Minority class Recognition

→ While undersampling helps balance the classes, reducing the overall dataset size may still leave the model with insufficient data (especially if the original minority class was very small) to learn meaningful patterns in the minority class effectively.

iv) Difficulty in Generalization

→ Undersampling might create an artificial distribution that is too different from real-world conditions, hindering the model's ability to make accurate predictions in practice.

Example

oversampling – 100000 – non fraudulent records
1000 – fraudulent records.

We want to make 25% of the records fraudent with oversampling.

Rare + x = p (records + x)

⇒ 1000 + x = 0.25 × (100000 + x)

⇒ 1000 + x = 25000 + 0.25x

⇒ 0.75x = 24000

⇒ x = 24000/0.75 = 32000

Total rare records after oversampling
= 1000 + 32000 = 33000.

(Ans)

Chapter – 8

Simple Linear Regression

→ We consider the modelling between the dependent and independent variable.

→ When there is only one independent variable in the linear regression model, the model is generally termed as a simple linear regression model.

→ When there are more than one independent variable in the model, then the linear model is termed as the multiple linear regression model.

→ Consider a simple linear regression model –

$y = b_0 + b_1 x + \epsilon$

where y → dependent or study variable
x → independent or exploratory variable
b_0 & b_1 → parameters of the model (regression coefficients)
b_0 → y-intercept of the regression line
b_1 → slope of the regression line.
ϵ → Unobservable error component (it accounts for the failure of data to lie on a straight line and represents the difference between the true and observed realization of y)

→ the goal of linear regression is to find a straight line that minimizes the error (the difference) between the observed points and the predicted values.

→ this line helps us predict the dependent variable for new, unseen data.

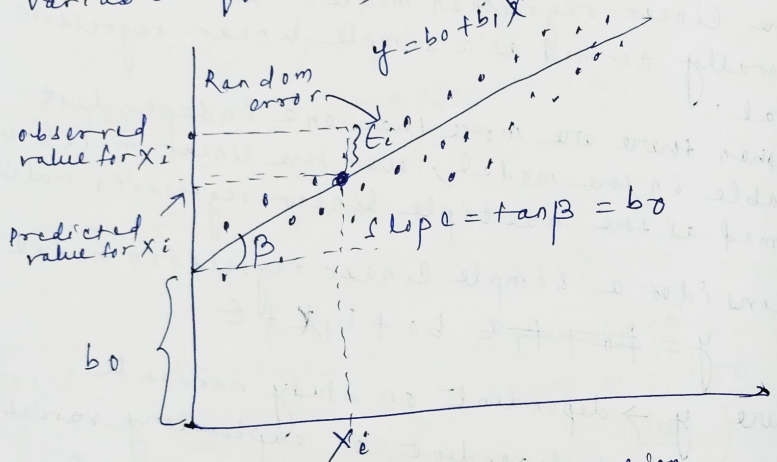


figure: Linear Regression

Equation of the Best-fit Line

→ for simple linear regression (with one independent variable), the best fit line is represented by the equation

$$y = b_0 + b_1 X$$

Minimizing the Error - The Least Square Method

→ To find the best fit line, we use a method called Least Square Estimation.

→ the idea behind this method is to minimize the sum of squared difference between the actual values (data points) and the predicted values from the line equation of regression line

$$y = b_0 + b_1 x + \epsilon \quad \text{——(1)}$$

for n observations, we have

$$y_i = b_0 + b_1 x_i + \epsilon_i, \text{ for } i = 1, 2, \cdots n \quad \text{——(2)}$$

→ the least squares line is that line that minimizes the population sum of squared error (SSE),

$$SSE_p = \sum_{i=1}^{n} \epsilon_i^2 \quad \text{——(3)}$$

where ϵ_i = actual observed value$_i$ − predicted value$_i$

$$SSE_p = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \quad \text{——(4)}$$

→ we may find values of b_0 and b_1 that minimizes $\sum_{i=1}^{n} \epsilon_i^2$ by differentiating eq^n(4) w.r.t. to b_0 and b_1 and setting the result equal to zero.

→ the partial derivations of eq^n(4) w.r.t to b_0 and b_1 are respectively −

$$\frac{d SSE_p}{d b_0} = \sum_{i=1}^{n} 2 \cdot (y_i - b_0 - b_1 x_i) \cdot \frac{d(-b_0)}{d b_0}$$

$$= -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) \quad \text{——(5)}$$

$$\frac{dSSE}{db_1} = \sum_{i=1}^{n} 2\cdot(y_i - b_0 - b_1 x_i)\cdot\frac{d}{db_1}(-b_1 x_i)$$

$$= -2\sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i) \quad \text{— (6)}$$

setting eq^n(5) & (6) to zero, we have

$$-2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) = 0$$

$$-\sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i) = 0$$

$$\Rightarrow \begin{cases} \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) = 0 \\ \sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i) = 0 \end{cases}$$

$$\Rightarrow \begin{cases} \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} b_0 - \sum_{i=1}^{n} b_1 x_i = 0 \\ \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} b_0 x_i - \sum_{i=1}^{n} b_1 x_i^2 = 0 \end{cases}$$

$$\Rightarrow \begin{cases} \sum_{i=1}^{n} y_i - n b_0 - b_1 \sum_{i=1}^{n} x_i = 0 \\ \sum_{i=1}^{n} x_i y_i - b_0 \sum_{i=1}^{n} x_i - b_1 \sum_{i=1}^{n} x_i^2 = 0 \end{cases}$$

$$\Rightarrow \begin{cases} n b_0 + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \quad \text{— (7)} \\ b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \quad \text{— (8)} \end{cases}$$

solving eq^n(7) and eq^n(8) for b_0 and b_1, we have

$$b_0 n = \sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i \quad \text{(from eq^n(7))}$$

$$\Rightarrow b_0 = \frac{\sum_{i=1}^{n} y_i}{n} - \frac{b_1 \sum_{i=1}^{n} x_i}{n}$$

$$\Rightarrow \boxed{b_0 = \bar{y} - b_1 \bar{x}} \quad \text{— (9)}$$

From eq^n(8), we have (putting b_0 value from eq^n(9) in eq^n(8))

$$\left[\frac{\sum_{i=1}^{n} y_i}{n} - \frac{b_1 \sum_{i=1}^{n} x_i}{n}\right]\sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

$$\Rightarrow \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n} - \frac{b_1 \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i}{n} + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

$$\Rightarrow \frac{\sum x_i \sum y_i}{n} - b_1 \frac{(\sum x_i)^2}{n} + b_1 \sum x_i^2 = \sum x_i y_i$$

$$\Rightarrow b_1\left(\sum x_i^2 - (\sum x_i)^2/n\right) = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$$

$$\Rightarrow \boxed{b_1 = \frac{\sum x_i y_i - (\sum x_i \sum y_i)/n}{\sum x_i^2 - (\sum x_i)^2/n}}$$

Numerical Problem on Simple Linear Regression

Consider the following set of points $\{(-2,-1),(1,1),(3,2)\}$

a. Find the least square regression line for the given data points.

Ans:- We use the following equations to find the regression line coefficients.

$$b_0 = \bar{y} - b_1\bar{x}, \quad b_1 = \frac{\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right)/n}{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2/n}$$

x	y	xy	x^2
-2	-1	2	4
1	1	1	1
3	2	6	9
$\Sigma x = 2$	$\Sigma y = 2$	$\Sigma xy = 9$	$\Sigma x^2 = 14$

Now, we can calculate the value of b_0 and b_1 (regression coefficient)

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right)/n}{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2/n}$$

$$= \frac{9 - (2\times 2)/3}{14 - (2)^2/3} = \frac{23}{38} = 0.605$$

$$b_0 = \bar{y} - b_1\bar{x}$$

$$= 2/3 - \frac{23}{38} \times \frac{2}{3} = 0.263$$

so equation of regression line is

$$y = 0.263 + 0.605x \quad (\text{Ans})$$

Q

Record No.	X	Y
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81

find out the regression line.

Ans:-

Record No.	X	Y	XY	X^2
1	43	99	4257	1849
2	21	65	1365	441
3	25	79	1975	625
4	42	75	3150	1764
5	57	87	4959	3249
6	59	81	4779	3481
	$\Sigma X = 247$	$\Sigma Y = 486$	$\Sigma XY = 20485$	$\Sigma X^2 = 11409$

So $b_1 = \frac{\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right)/n}{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2/n}$

$$= \frac{20485 - ((247 \times 486)/6)}{11409 - (247)^2/6} = 0.385$$

$$b_0 = \bar{y} - b_1\bar{x} = \frac{486}{6} - 0.385 \times \frac{247}{6} = 65.15$$

so equation of regression line

$$y = b_0 + b_1 x = 65.15 + 0.385x$$

(Ans)

Calculation of SSE, SSR, SST, r^2

Sum of Squared Error (SSE)

* It is the sum of the squared differences between each observed and the predicted value.
* It can be defined as

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

y_i → actual value
$\hat{y}_i$ → predicted value

Sum of Squares Regression (SSR)

* It is the sum of the squared differences between the predicted value and the mean of the dependent variable.
* It can be defined as

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$\bar{y}$ → mean of dependent variable

Sum of Squares Total (SST)

* It is the sum of the squared differences between the observed dependent variables and the overall mean.
* It can be defined as

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

where y_i → actual value for i^{th} data point
$\hat{y}_i$ → predicted value for i^{th} data point and $\bar{y}$ → mean of variable y

Coefficient of Determination (r^2)

* It explains how much a dependent variable varies when the independent variable is varied.
* It can be defined as

$$r^2 = \frac{SSR}{SST}$$

Numerical problem on SSE, SSR and SST

Subject	X = Time	Y = Distance	Predicted Score $\hat{y} = 6 + 2x$
1	2	10	10
2	2	11	10
3	3	12	12
4	4	13	14
5	4	14	14
6	5	15	16
7	6	20	18
8	7	18	20
9	8	22	22
10	9	25	24

Subject	X = Time	Y = Distance	Predicted Score $\hat{y} = 6 + 2x$	Error in prediction $(y - \hat{y})$	Error in prediction $(y - \hat{y})^2$
1	2	10	10	0	0
2	2	11	10	1	1
3	3	12	12	0	0
4	4	13	14	−1	1
5	4	14	14	0	0
6	5	15	16	−1	1
7	6	20	18	2	4
8	7	18	20	−2	4
9	8	22	22	0	0
10	9	25	24	1	1

$$SSE = \sum_{i=1}^{?} (y_i - \hat{y}_i)^2 = 12$$

subject	x = time	y = distance	predicted scores	$(\hat{y} - \bar{y})$	$(\hat{y} - \bar{y})^2$
1	2	10	10	-6	36
2	2	11	10	-6	36
3	3	12	12	-4	16
4	4	13	14	-2	4
5	4	14	14	-2	4
6	5	15	16	0	6
7	6	20	18	2	4
8	7	18	20	4	16
9	8	22	22	6	36
10	9	25	24	8	64
					216

$\bar{y} = 16$ $\quad SSR = \sum_{i=1}^{?} (\hat{y} - \bar{y})^2 = 216$

subject	x = time	y = Distance	$\hat{y}$	$(\hat{y} - \bar{y})$	$(\hat{y} - \bar{y})^2$
1	2	10	16	-6	36
2	2	11	16	-5	25
3	3	12	16	-4	16
4	4	13	16	-3	9
5	4	14	16	-2	4
6	5	15	16	-1	1
7	6	20	16	4	16
8	7	18	16	2	4
9	8	22	16	6	36
10	9	25	16	9	81

$\bar{y} = 16$, $\quad SST = \sum_{i=1}^{n} (y - \bar{y})^2 = 228$

or, SST can also be calculated as

$$\boxed{SST = SSR + SSE}$$

$= 12 + 216 = 228$

$$\therefore r^2 = \frac{SSR}{SST} = \frac{12}{228} = 0.0526$$


COLOMBIA
SOUTH