Statistics Worksheet - 4

(1) • The central limit theorem tells us that as sample sizes get larger, the sampling distribution of the mean will become normally distributed, even if the data within each sample are not normally distributed.

• The central limit theorem is important for statistics because it allows us to safely assume that the sampling distribution of the mean will be normal in most cases. This means that we can take advantage of statistical techniques that assumes a normal distribution.

(2) • When you conduct research about a group of people, it's rarely possible to collect data from every person in that group, instead you select a sample. The sample is a group of individuals who will actually participate in the research.

- To draw valid conclusions from your result, you have to carefully decide how you will select a sample that is representative of the group as a whole.
- There are 2 types of sampling methods:

① Probability Sampling

It involves random selection, allowing you to make strong statistical inferences about the whole group

② Non-Probability Sampling

It involves non-random selection based on convenience or other criteria allowing you to easily collect data.

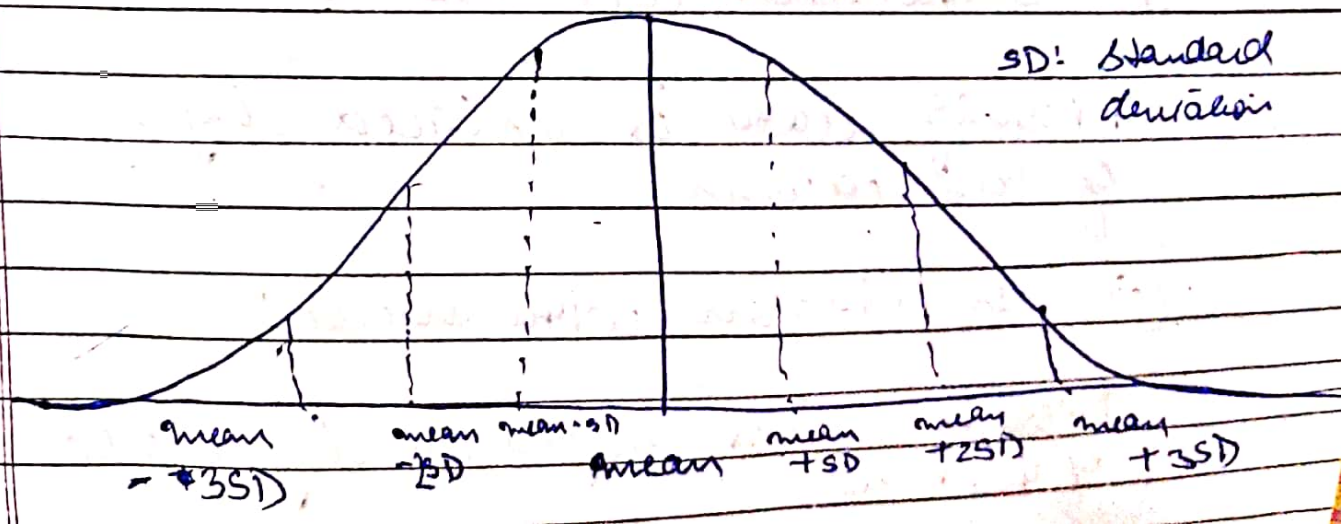
③

Type 1 error	Type 2 error
• Error caused by rejecting null hypothesis when it is true	error that occurs when null hypothesis is accepted when it is not true
• false Positive	false negative

Probability of type I error is equal to the level of significance	Probability of type 2 is equal to one mi Power of the test
It can be reduced by decreasing level of significance	It can be reduced by increasing lev of significance
It is caused by luck or chance	It is caused by a smaller sample size or a less powerful test

(4) What do you understand by the term Normal distribution?

- Normal distribution is a probability distribution that is symmetric about the mean
- The standard normal distribution has 2 parameters: mean & standard deviation
- The empirical rule states that:
  (i) 68% of values fall within 1 standard deviation of the mean
  (ii) 95% of values fall 2 standard deviation of the mean
  (iii) 99.7% of values are within 3 standard deviation of the mean
- In graph form, it appears as bell curve



(5) → Covariance

- Covariance is a measure to indicate the extent to which 2 random variables change in tandem.
- It is measure of correlation.
- It indicates the direction of linear relationship between variables.

⇒ Correlation

- It is a measure used to represent how strongly two random variables are related to each other
- It is scaled form of covariance.
- It measures both the strength and direction of the linear relationship between 2 variables.

(5) (1) Univariate Analysis

It is the simplest form of data analysis where data being analyzed contains only one variable.

(2) Bivariate Analysis.

It is used to find out if there is a relationship between two variables. If the data seems to fit a line or curve then there is relationship between the 2 variables.

(3) Multivariate analysis

It is the analysis of 3 or more variables

(7). Sensitivity is also called true positive rate the recall or probability of detection

- It is defined as the ability of a test to measure the proportion of positives that are correctly identified

Formula :-

$$\text{Sensitivity} = \frac{\text{number of True Positives}}{\text{number of true positives} + \text{number of false negatives}}$$

$$= \frac{TP}{TP + FN}$$

(8). Hypothesis testing is used to choose between two competing hypothesis about the value of a population parameter.

- H_0: It is referred as null hypothesis

The hypothesis to be tested is null hypothesis. It is assumed to be true unless there is a strong evidence

to contrary.

- H_1: The other hypothesis which is assumed to be true, when null hypothesis is false is called alternative hypothesis

- An alternative hypothesis that specified that the parameter can lie on either side of the value specified by H_0 is called two-sided test

$H_0: \mu = 100$

$H_A: \mu <> 100$

(9) Qualitative data

- It is a set of numbers collected from a group of people and it values statistical analysis.

- ~~Data~~ Qualitative data

- It is set of information which can not be measured using numbers.

- It generally consist of words, subjective

narratives. Result of an qualitative data analysis can come in form of highlighting key words, extracted information and concept elaborate

(10)

Range:

$R = \text{max} - \text{min}$

Inter Quartile Range:

$IQR = Q_3 - Q_1$

Q_3: 3rd quartile

Q_1: 1st quartile.

(17) • Normal distribution is a probability distribution that is symmetric about the mean

• The standard normal distribution has 2 parameters : mean & standard deviation

• The empirical rule states that :
(i) 68% of values fall within 1 standard deviation of the mean
(ii) 95% of values fall 2 standard deviation of the mean
(iii) 99.7% of values are within 3 standard deviation of the mean

(12) · The most effective ways to find all of your outliers is by using Inter Quartile Range (IQR)

· The IQR contains the middle bulk of data, so outliers can be easily found once you know the IQR.

· An outlier is defined as being any point of data that lies over 1.5 IQRs below the first quartile (Q_1) or above the third quartile (Q_3) in a dataset.

$$High = (Q_3) + 1.5IQR$$

$$Low = (Q_1) - 1.5IQR$$

(13) · The P-value is the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test assuming that the null hypothesis is correct.

The P-value is used as an alternative to rejection points to provide the smallest level of significance at which null hypothesis will be rejected

(14) Binomial Probability Formula:

$$P_x = \binom{n}{x} p^x q^{n-x}$$

where:-

P = binomial probability

x = number of times for a specific outcome within n trials

$\binom{n}{x}$ = number of combinations

p = Probability of success on a single trial.

q = Probability of failure on a single trial.

n = number of trials.

(15) ANOVA

Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other.

ANOVA checks the impact of one or more factors by comparing the mean of different samples.