**Assignment2 – STATISTICS**

1. **Define and briefly elaborate Central Tendency using measures with examples?**
   * central tendency is the descriptive summary of a data set
   * Through the single value from the dataset , it reflects the centre of the data distribution
   * It does not provide information regarding individual data from dataset, Where it gives summary of dataset
   * There are generally 3 measures of central tendency: Mean, Median, Mode



* + Mean is the most common measure of central tendency

Ex: The weight of 8boys in kg : 45,39,53,45,43,48,50,45

MEAN = sum of weight/ No of boys

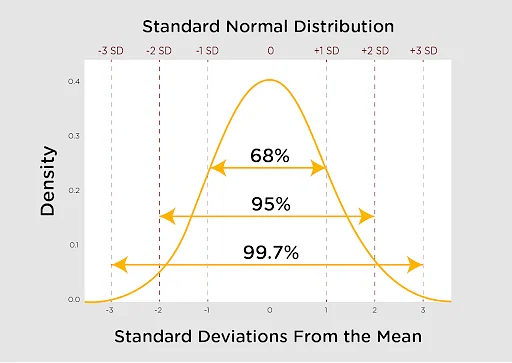
= 46

MEDIAN=45

MODE = 45

1. **What do you understand by the empirical rule of normal distribution?**

* The empirical rule also known as three-sigma rule or 68-95-99.7 rule
* It is a statistical rule that states almost all observed data for a normal distribution will fall within three standard deviations
* According to this rule, 68% of data falls within one standard deviation,95% within two standard deviations and 99.7% within three standard deviations from mean
* Graphical representation :



FORMULA : µ µ-mean ,m-multiplier,-standard deviation

EX : suppose pulse rates of 100 students are bell-shaped with mean of 75 and standard deviation of 4

Around 68% of the men have pulse rate in interval 751(4) =[71,79]

Around 95% of the men have pulse rate in interval 752(4) =[67,83]

Around 99.7% of the men have pulse rate in interval 753(4) =[63,87]

LIMITATIONS : It can only applied to a symmetric and unimodal distribution because its only applicable to normal statistical distributions

1. **Describe the Hypothesis Testing and why do we conduct it?**

* Hypothesis testing is a tool for making statistical inference about the population data.
* Its an analysis tool that tests assumptions and determines how likely something is within a given standard of accuracy
* A null hypothesis and an alternative hypothesis are setup before performing the hypothesis testing

Z = ( x̅ – μ0 ) / (σ /√n)

Here, x̅ is the sample mean,

μ0 is the population mean,

σ is the standard deviation,

n is the sample size.

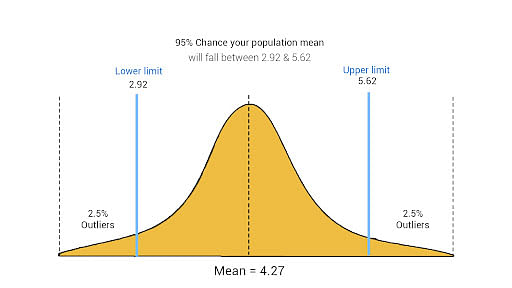
* + In hypothesis testing, there are two mutually exclusive statements about the population: the null hypothesis and the alternative hypothesis.
  + Since both of them are contradictory, we therefore must examine the sample to determine whether we have enough evidence to conclude that the research objective or the alternative hypothesis holds true or not.

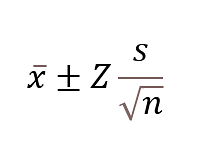
1. **What is the difference between Type-I and Type-II error.**

|  |  |  |
| --- | --- | --- |
| **s.no** | **Type-I error** | **Type-II error** |
| 1. | Non-acceptance of hypothesis which ought to be accepted | Acceptance of hypothesis which ought to be rejected |
| 2. | Incorrect rejection of true null hypothesis | Incorrect acceptance of false null hypothesis |
| 3. | Indicated by Greek letter ‘α’ | Indicated by Greek letter ‘β’ |

1. **What are Confidence Interval, Significant Level and P-Value?**

* A confidence interval shows the [probability](https://www.simplilearn.com/tutorials/statistics-tutorial/what-is-probability-distribution) that a parameter will fall between a pair of values around the mean.
* Confidence intervals show the degree of uncertainty or certainty in a [sampling method](https://www.simplilearn.com/types-of-sampling-techniques-article).
* They are constructed using confidence levels of 95% or 99%.
* The size of a 90% confidence interval for a given estimate is one method to gauge how "excellent" it is; the greater the range, the more care must be used when utilising the estimate. Confidence intervals serve as a crucial reminder of the estimates' limits.



* The 95% confidence interval is the range that you can be 95% confident that the similarly constructed intervals will contain the parameter being estimated.
* The sample mean (centre of the CI) will vary from sample to sample because of natural sampling variability. 
  + - * X bar is the sample mean.
      * Z is the number of [standard deviations](https://www.simplilearn.com/tutorials/statistics-tutorial/what-is-normal-distribution) from the sample mean.
      * S is the standard deviation in the sample.
      * n is the size of the sample.
      * The value after the ± symbol is known as the margin of error.

**SIGNIFICANCE LEVEL :**

* The [significance level](https://statisticsbyjim.com/glossary/significance-level/), also known as alpha or α, is a measure of the strength of the evidence that must be present in your [sample](https://statisticsbyjim.com/glossary/sample/) before you will reject the null hypothesis and conclude that the [effect](https://statisticsbyjim.com/glossary/effect/) is statistically significant.
* The researcher determines the significance level before conducting the experiment.
* Lower significance levels indicate that you require stronger evidence before you will reject the null hypothesis.
* Compare your [p-value](https://statisticsbyjim.com/glossary/p-value/) to your significance level.
* If the p-value is less than your significance level, you can reject the null hypothesis and conclude that the effect is statistically significant.

**P-VALUE**

* The P-value is known as the probability value.
* It is defined as the probability of getting a result that is either the same or more extreme than the actual observations.
* The P-value is known as the level of marginal significance within the hypothesis testing that represents the probability of occurrence of the given event.
* The P-value is used as an alternative to the rejection point to provide the least significance at which the [null hypothesis](https://byjus.com/maths/null-hypothesis/) would be rejected.
* If the P-value is small, then there is stronger evidence in favour of the alternative hypothesis.

|  |  |
| --- | --- |
| **P-value** | **Decision** |
| P-value > 0.05 | The result is not statistically significant and hence don’t reject the null hypothesis. |
| P-value < 0.05 | The result is statistically significant. Generally, reject the null hypothesis in favour of the alternative hypothesis. |
| P-value < 0.01 | The result is highly statistically significant, and thus rejects the null hypothesis in favour of the alternative hypothesis. |

1. **List the differences between Parametric and Non Parametric Tests**

|  |  |  |
| --- | --- | --- |
| **Properties** | **Parametric** | **Non-parametric** |
| Assumptions | Yes | No |
| central tendency Value | Mean value | Median value |
| Correlation | Pearson | Spearman |
| Probabilistic distribution | Normal | Arbitrary |
| Population knowledge | Requires | Does not require |
| Used for | Interval data | Nominal data |
| Applicability | Variables | Attributes & Variables |
| Examples | z-test, t-test, etc. | Kruskal-Wallis, Mann-Whitney |

1. **What is Central Limit Theorem**

Central limit theorem (CLT) is a statistical theory given that as sample sizes get larger, the mean of all samples will be approximately equal to the mean of the population, and the distribution of means will approach normality.

1. **Find the probability of P(x<400) given that mean is = 1000 variance is =100**

We calculate Z-score and use standard normal distribution table

Z=(X-µ)/ (mean=1000, variance=100,S.D=variance,10)

=400-1000/10

= -600/10

=-60

In standard normal distribution probability of obtaining Z-score of -60 or lower is close to zero

CONCLUSION: Probability p(X<400) for normal distribution with mean of 1000 and variance of 100 is negligible ,so observing value less than 400 is **extremely low, almost impossible**

1. **Provide a proper description as to when to apply z-test, t-test, Chi-Square and Annona, with examples.**
2. **T-Test:**

* A t-test is used to compare the means of two groups to determine if there is a significant difference between them.
* You can use a t-test when your sample size is small (n < 30) and the population standard deviation is unknown.
* There are two types of t-tests:
* A two-sample t-test
* A paired t-test.

A two-sample t-test is used to compare the means of two independent groups, while a paired t-test is used to compare the means of two related groups, such as before-and-after measurements.

1. **Z-Test:**

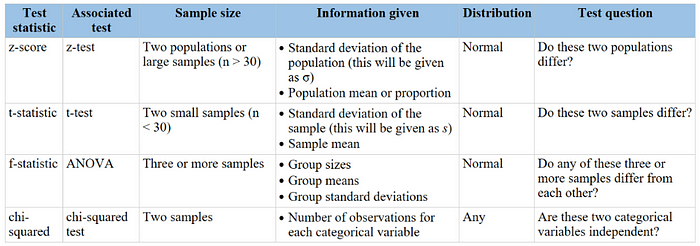
* A z-test is used to compare the mean of a single sample to a known population mean.
* Unlike a t-test, a z-test assumes that the population standard deviation is known.
* A z-test is typically used when the sample size is large (n > 30) or when the population standard deviation is known from previous studies.

1. **ANOVA:**

* ANOVA (analysis of variance) is used to compare the means of more than two groups to determine if there is a significant difference between them.
* **ANOVA is appropriate when you have continuous data and multiple groups to compare.**

1. **Chi-Square Test:**

* A chi-square test is used to compare the frequency or count of observations in two or more categories or groups.
* It is used to test the independence of two categorical variables.
* For example, you can use a chi-square test to determine if there is a relationship between gender and political affiliation.
* **The chi-square test is appropriate when you have categorical data and you want to determine if there is a relationship between two variables.**

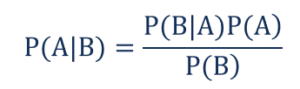


1. **Explain the Bayes Theorem using an example.**

In statistics and probability theory, the Bayes’ theorem (also known as the Bayes’ rule) is a mathematical formula used to determine the conditional probability of events. Essentially, the Bayes’ theorem describes the [probability](https://corporatefinanceinstitute.com/resources/data-science/total-probability-rule/) of an event based on prior knowledge of the conditions that might be relevant to the event.

**Formula for Bayes’ Theorem**

The Bayes’ theorem is expressed in the following formula:



Where:

* P(A|B) – the probability of event A occurring, given event B has occurred
* P(B|A) – the probability of event B occurring, given event A has occurred
* P(A) – the probability of event A
* P(B) – the probability of event B

Note that events A and B are [independent events](https://corporatefinanceinstitute.com/resources/data-science/independent-events/) (i.e., the probability of the outcome of event A does not depend on the probability of the outcome of event B).

A special case of the Bayes’ theorem is when event A is a [binary variable](https://www.fico.com/fico-xpress-optimization/docs/latest/mipform/dhtml/chap2s1.html). In such a case, the theorem is expressed in the following way:

Special Case

Where:

* P(B|A–) – the probability of event B occurring given that event A– has occurred
* P(B|A+) – the probability of event B occurring given that event A+ has occurred

In the special case above, events A– and A+ are mutually exclusive outcomes of event A.