STATISTICS

Inferential statistics

Inferantial statistics is a part of statistics where we can estimate the value of any paramater of population data based on the suitable or appropiate experiment of sample data.

In one word, here, we can extimate the information about population based on the information of sample.

The important topics of inferential statistics are —

Central Limit Theorm

Defination : The central limit theorm states that the distribution of the sample means of a large number of independent and identically distributed random variables will approach a normal distribution, regardless of the underlying distribution of the variables.

The conditions required for the CLT to hold are —

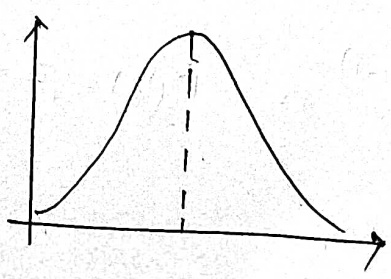
1. the sample size ≥ 30.
2. The sample is drawn from a finite population or infinite population with finite variance.
3. The random variables in the sample are independent and identically distributed.

If the population has the mean μ and variance σ^2 then the emprerical mean and variance would be μ and σ^2/n.

$$\bar{X}_i \sim N(\mu, \sigma^2/N) \qquad N = \text{sample size.}$$

Z-score

For normal distribution —



(II) Hypothesis Testing :

A statistical hypothesis test is a method of statistical inference used to decide whether the data at hand sufficiently support a particular hypothesis. Hypothesis testing allows us to make probabilistic statements about population parameters.

(III) Null hypothesis and Alternative Hypothesis :

(A) Null hypothesis :

In simple word, the null hypothesis is a statement that assumes there is no significant effect or relationship between the variables being studied.

(B) Alternative hypothesis :

Alternative hypothesis is a statement that contradicts the null hypothesis and claims that there is significant effects or relationship between the variables being studied.

It is also called Research hypothesis.

(IV) The purpose of hypothesis testing is to gather evidence (data) to either reject or fail to reject the null hypothesis in favour of alternative hypothesis.

(V) It is important to note that failing to reject the null hypothesis doesn't necessarily mean that the null hypothesis is true. It just means that there isn't enough evidence to support the alternative hypothesis.

(VI) Type-I and Type-II error

In hypothesis testing, there are two types of errors that can occur when making a decision about the null hypothesis.

(A) type-I (B) type-II

(A) **Type-I Error** :

It occurs when the sample results lead to the rejection of a null hypothesis, when it is actually true. (False positive)

(1) It is the mistake of finding a significant effect or relationship when there is none.

It is also known as significance level, denoted by α. By choosing a significance level researchers can control the risk of making a type I error.

(B) **Type-II error**

It occurs when based on the sample results, the null hypothesis is not rejected when it is actually false. This means that the researchers fail to detect a significant effect or relationship when atleast one actually exists. It is denoted by β.

		Truth about the population	
		H_0 true	H_0 false
Decision based on sample	Reject H_0	False positive (type-I) (α)	Correct Decision
	Accept H_0	Correct Decision	False Negative (type-II) (β).

(1) **Trade off between Type-I error and Type-II error** :

Let say we are decreasing our α value that is type-I error. So, we reduce the probability of rejecting null hypothesis when it is true. Indirectly we are raising the chance of happening the type-II error. (that we are accepting Null hypothesis may be when it is actually false).

⑩ One Sided and Two Sided Test

⑪ One Sided (one tailed) Test :

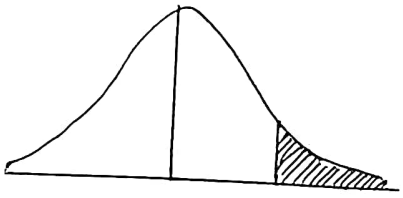
A one sided test is used when the researcher is interested in testing the effect in a specific direction (either greater than or less than)

Example — A researcher wants to test wheather a new medication increases the average recovery rate compared to the existing medication.

So, here the Null hypothesis be —

H_0 : There is no significant changes.

H_a : New medication increases the average recovery rates.



Here we have interest in the right side of graph (greater than). That is also called Right tailed test. (same logic for left tail test.

⑫ Advantage :

① More powerfull — Here the entire significance level (α) is allocated to one tail of distribution, that means that the test is more likely to detect an effect in that specified direction.

(ii) Directional Hypothesis.

⑬ Disadvantage :

① Missed effect in another side.

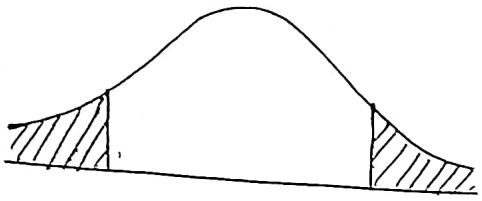
(n) Increase the risk of type-I error.

(ii) **Two-side or two tailed test** :

A two-sided test is used when the researcher is interested in testing the effect in both directions.

Example — A researcher wants to test a new medication has different average recovery rate compare to existing one.

So, here the researcher wants to test in both directions. (May be it is less than or greater than).



Null hypothesis $H_0 : \mu_{before} = \mu_{after}$

Alternative hypothesis $H_0 : \mu_{before} \neq \mu_{after}$

(it would be $\mu_{before} > \mu_{after}$ or $\mu_{before} < \mu_{after}$) that's why it is called two-tailed or two sided Test.

(iii) **Advantage** :

(i) Detects effects in both directions.

(ii) **More conservative** — because the significance level is splited between both tails, this reduces the type-I errors in case where the direction of effect is uncertain.

(iv) **Disadvantage** :

(i) **less powerfull** : It increase the type-II error as the α is splited into two direction of distribution.

(ii) Not appropiate for directional hypothesis.

P-Value :

P-Value is the probability of getting a sample as or more extreme (having more evidence against (H_0)) than our own sample given the Null hypothesis (H_0) is true.

Suppose we have done a experiment of tossing a coin 100 times, and we state that our null hypothesis is that the coin is fair. ($No_H = No_T$) and Alternative hypothesis is ~~the coin is not~~ number of getting head is more than 50.

Now, we did the experiment and got that it gives the result as 53 heads and 47 tails.

let suppose the corresponding P-Value of this experiment = 0.15

It means that there is 15% probability of getting 53 or more heads when we do the experiment, when our null-hypothesis is true.

In simple words, p-value is a measure of the strength of the evidence against Null hypothesis that is provided by our sample data.

Interpreting p-value with significance of level :

If p-value < 0.01 $\Rightarrow$ we have strong evendence against the Null hypothesis.

if $0.01 \leq p < 0.05$ $\Rightarrow$ modarate evidence against the null hypothesis.

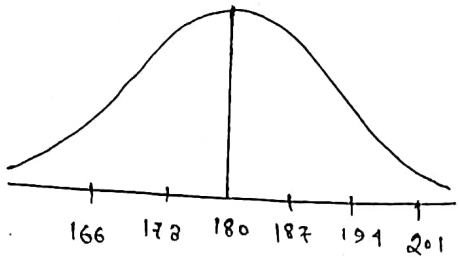
if $0.05 \leq p < 0.1$ $\Rightarrow$ indicates weak evidence against the null hypothesis. (but there is still some level of uncertainty).

if $p \geq 0.1$ $\Rightarrow$ indicates weak or no evidence against the null hypothesis.

Z-score and it's Application

Z-score is a statistic by which we can find for a data point how much far it is from mean.

For example let say we have a dataset whose mean = 180 and s.d = 7.



Now, if I want to know that ~~189~~ How much far 189 from it's mean, we apply Z score on it.

So, formula of **Z-score = (point − mean) / s.d**

Here — Z-score = $\frac{189 - 180}{7} = 1.28$

so, 189 is +1.28 s.d far from it's mean (180). (Ans)

Application

① Standarization

Let suppose we have the following two data columns having different units value.

Age (Y)	Weight (kg)
25	78
20	92
35	65
14	75
26	80

(1) Standarization is a common processing steps that aims to make the input data more suitable for certain machine learning algorithms, specially those that are sensitive to the scale and distribution of the features.

Here we can see that for age column mean = 24.

$$Var(Age) = \frac{1}{n}\sum(x_i - \bar{x})^2$$

$$= \frac{1}{5}\left[1^2 + 4^2 + 11^2 + 10^2 + 2^2\right] = 48.4$$

$$S.D = \sqrt{Var(Age)} = 6.95$$

in similarly we can find mean = 78 and SD = 8.6948 for weight column.

Now, if we do scaling with respect to these two column our columns looks like —

Age	Weight
0.144	0
-0.575	1.611
1.583	-1.495
-1.439	-0.345
0.288	0.230

It is only done by the formula of Z-score (For each datapoint (x_i)

$$Z\text{-}Sc = \frac{x_i - \bar{x}}{\sigma}$$

(2) **Compare any score**:

Z score is very useful in comparison of two score. (anything)

For example let say in 2020 India's avg run in T20 was 181 with S.D 12 and for 2021, it was 182 with S.D 5. Now, India scored 187 and 185 in the final of 2020 and 2021 respectively. Now, if we compare in which year India score well in final, we can't say directly. so Z score helps us in this situation.

(10) For 2020 —

$$Z\ score\ (187) = \frac{187 - 181}{12} = 0.5$$

For 2021

$$Z\ score\ (185) = \frac{185 - 182}{5} = 0.6$$

So, From here we see that Z score (for 2021) > Z score (2020).

So, although India scored more run in final-2020, but the better final performance was in 2021. (Proved)

⑩ Significance level :

Significance level (α), is a predetermined threshold used in hypothesis testing to determine weather the null hypothesis should be rejected or not. ~~When it actually~~ It represents the probability of rejecting the null hypothesis when it is actually true. (type-I error)

⑪ Type -I error and Type-II error.

⑫ Application of Hypothesis Testing :

① testing the effectiveness of interventions or treatement.

② Comparing means or proportions

③ Analysing relationship between variables

④ Evaluating the goodness of fit

⑤ Testing the independent of categorical variables

⑥ A/B testing