**Simple Hypothesis**

- Exactly one independent/cause and Exactly one dependent variable/effect. e.g:

a) eating more vegetables leads to weight loss

b) brushing teeth everyday reduces cavities

**Complex Hypothesis**

- Can have more than one independent and more than one dependent variables:

**Null Hypothesis (no difference)**

- Whatever was happening will continue to happen the same way.

a) Average score of a student is 60, will he continue to get 60 score based on the data.

**Alternate Hypothesis**

- Whatever was happening will NOT continue to happen the same way.

a) One tailed Test (when a person is criminal in front of court)

b) Two tailed Test (when our hypothesis says average score will be less than or greater than 60)

**Statistical Hypothesis**

- Performing statistical inference on data and using these statistical calculations for Null Hypothesis

|  |  |  |
| --- | --- | --- |
| **Actual** | **Predicted (By Judge)** |  |
| Criminal | Criminal |  |
| Innocent | Innocent |  |
| Innocent | Criminal | **α**: Type-1 Error (When you end up rejecting the Null Hypothesis) |
| Criminal | Innocent | **β**: Type-2 Error (When you end up accepting the Null Hypothesis) |

**H0**: NULL Hypothesis: (Every Person is innocent until proved guilty)

**H1**: Alternate Hypothesis

**α**: Type-1 Error (When you end up rejecting the Null Hypothesis)

**β**: Type-2 Error (When you end up accepting the Null Hypothesis)

**Population Data** -> Entire Dataset

**Sample Data** -> subset of population

**μ** -> Mean of Population data

**σ** -> Standard deviation of population data

-> mean of sample data

**n** -> number of records in sample data

**N** -> number of records in population data

**Central Limit Theorem:**

1. When population data is known to follow normal distribution, the sample data also follows a normal distribution.
2. When the population data is not normally distributed, the sample follows a normal distribution, provided the sample size is large enough. where generally n>35.

**Normal Distribution**:

When the majority of data is around mean value.

**Parametric Tests** (Assuming that data is normally distributed). **Z-test, T-test, Anova Test.**

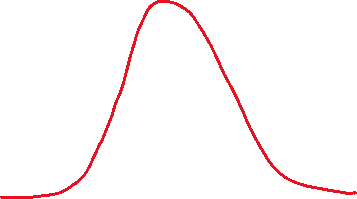
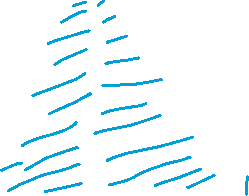
**Null Hypothesis (H0)**: A person sells 350 Toffees every day.

**Two Tailed Test:** Person could sell less than 350 and more than 350.

**One Tailed Test:** If the hypothesis would have been “Sells at least 350 toffees every day” or “Sells at most 350 toffees every day”.

Critical Points

95%



**μ** = 350

2.5% error

2.5% error

Finding the Upper and Lower Limits (Critical Points):

**Z-Score** =

If -1.96 < z-score < 1.96 , accept the Null Hypothesis.

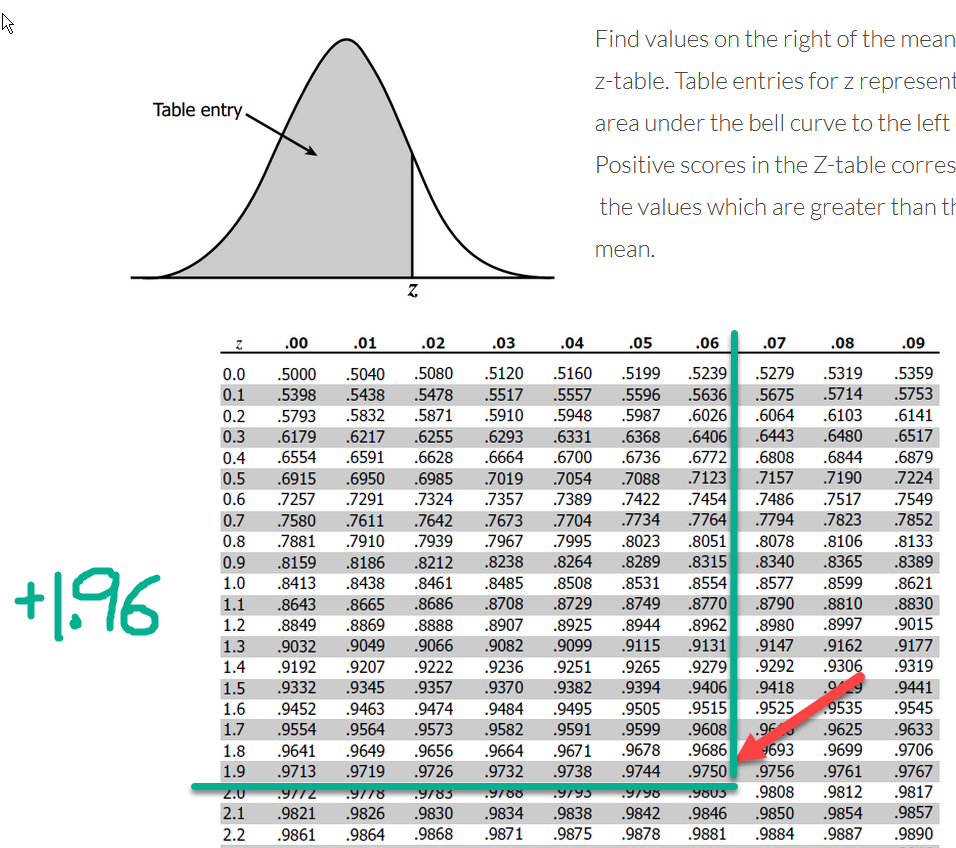
Else, Reject Null Hypothesis.

<https://www.z-table.com/#>

Find level of significance value 0.025 and 0.975 (1-0.025)

A picture containing table

Description automatically generated



It can be a challenge to get Standard deviation of population. Alternatively, we can use T-Score.

s = standard deviation of sample data.

**T-Score** =

School of 1000 students, Average IQ of students is 100.

Marketing representative of a pharmaceutical company claims that a pill can improve IQ.

H0: The Average IQ of students will remain equal to 100. (Since this is Two-tailed test, refer to 2-tail T-Table)

N = 1000

n = 25

μ = 100

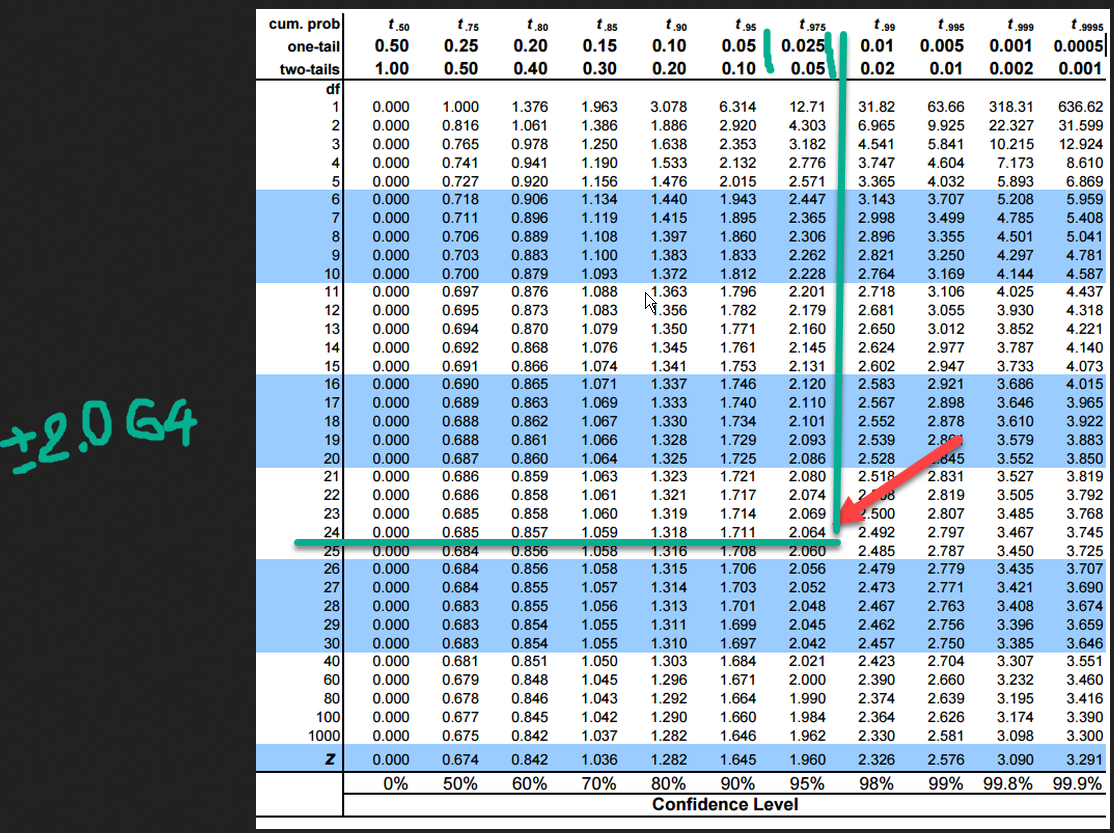
= 125

s = 18 (assumption, to be calculated from sample data)

Degree of Freedom = df = n-1 = 24

T-score = = 6.94

-- Reject the Null Hypothesis --



Anova Test:

When we want to compare more than one dataset.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Car1 | Car2 | Car3 |
| Driver1 | 65 | 69 | 67 |
| Driver2 | 35 | 38 | 33 |
| Driver3 | 56 | 54 | 58 |

Try to compare if the pollution levels of five different ponds are statistically similar.

k = 5 (number of groups)

n = 20 (total no of observations in each group)

H0: All the groups are similar in data behavior (always same for Anova Test)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Ponds | **X1** | **X2** | **X3** | **X4** | **X5** |
| 20 observations for each pond | X11  X12  X13  .  .  .  X120 | X21  X22  X23  .  .  .  X220 | X31  X32  X33  .  .  .  X320 | X41  X42  X43  .  .  .  X420 | X51  X52  X53  .  .  .  X520 |
|  |  |  |  |  |  |

Fstats =

Grand Mean = =

Variance between the groups = where *i* = 1 to k

Variance within the groups = where *i* = 1 to n

If the datasets are similar, then “**Variance between the groups**” should be **low** and “**Variance within the groups**” can be high or low.

If these ponds are to be similar, **Fstats** value should be **lower**.