# Session-17 Statistics

10 August 2024 01:50 PM

***Agenda***

- **Inferential Statistical Tests**
- **Confidence Interval**
- **Regression Analysis**
- **Hypotheses Testing**
- **T-Test I Z-Test I F-Test I Annova Test I Chi Square Test**
- **Null Hypotheses**
- **Alternate Hypotheses**
- **P - Value, Significance Level**

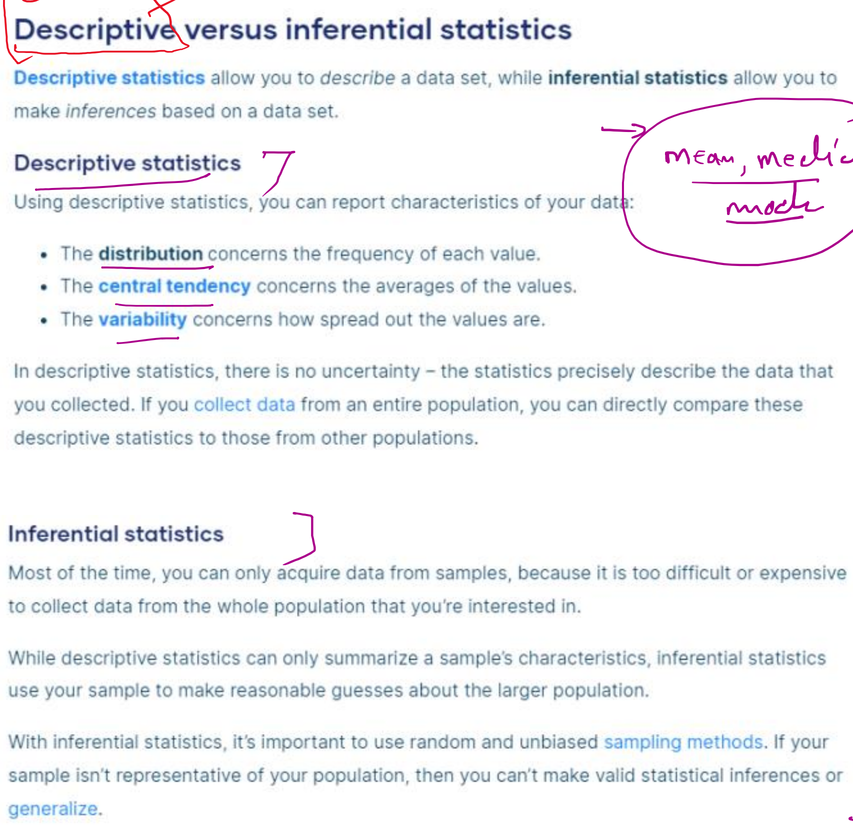
### Inferential Statistics Definition

Inferential statistics can be defined as a field of statistics that uses analytical tools for drawing conclusions about a population by examining random samples. The goal of inferential statistics is to make generalizations about a population. In inferential statistics, a statistic is taken from the sample data (e.g., the sample mean) that used to make inferences about the population parameter (e.g., the population mean).

34% → 30% → 30%

x y z

summarize



## Descriptive versus inferential statistics

**Descriptive statistics** allow you to *describe* a data set, while **inferential statistics** allow you to make *inferences* based on a data set.

### Descriptive statistics

Using descriptive statistics, you can report characteristics of your data:

- The **distribution** concerns the frequency of each value.
- The **central tendency** concerns the averages of the values.
- The **variability** concerns how spread out the values are.

In descriptive statistics, there is no uncertainty – the statistics precisely describe the data that you collected. If you collect data from an entire population, you can directly compare these descriptive statistics to those from other populations.

### Inferential statistics

Most of the time, you can only acquire data from samples, because it is too difficult or expensive to collect data from the whole population that you're interested in.

While descriptive statistics can only summarize a sample's characteristics, inferential statistics use your sample to make reasonable guesses about the larger population.

With inferential statistics, it's important to use random and unbiased sampling methods. If your sample isn't representative of your population, then you can't make valid statistical inferences or generalize.

mean, median
mode

Hypothesis testing
&
Regression Analysis

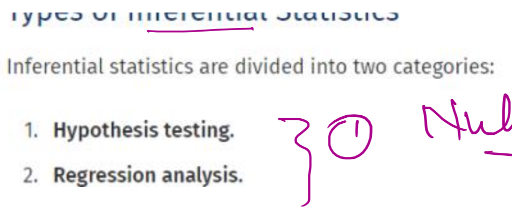
→ t - test, z-test,

## Types of Inferential Statistics

Inferential statistics are divided into two categories:

→ t

Null hypothesis → (current belief)



Types of Inferential Statistics

Inferential statistics are divided into two categories:

1. **Hypothesis testing.**
2. **Regression analysis.**

① Null Hypothesis → (current belief)

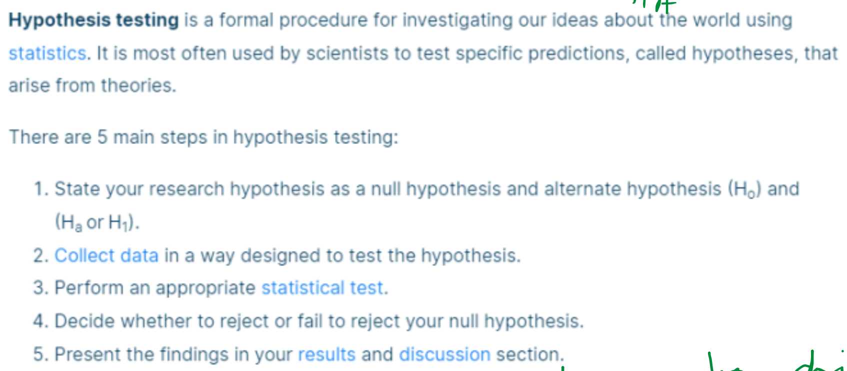
↓

(existing belief)

② Alternate Hypothesis → what we intend to establish

↓

HA



**Hypothesis testing** is a formal procedure for investigating our ideas about the world using statistics. It is most often used by scientists to test specific predictions, called hypotheses, that arise from theories.

There are 5 main steps in hypothesis testing:

1. State your research hypothesis as a null hypothesis and alternate hypothesis ($H_0$) and ($H_a$ or $H_1$).
2. Collect data in a way designed to test the hypothesis.
3. Perform an appropriate statistical test.
4. Decide whether to reject or fail to reject your null hypothesis.
5. Present the findings in your results and discussion section.

eg Null Hypothesis → Children who drink the health drink complan are likely to grow taller.

↓

Ho

### 01. Hypothesis testing

Testing hypotheses and drawing generalizations about the population from the sample data are examples of inferential statistics. Creating a null hypothesis and an alternative hypothesis, then performing a statistical test of significance are required.

A hypothesis test can have left-, right-, or two-tailed distributions. The test statistic's value, the critical value, and the confidence intervals are used to conclude. Below are a few significant hypothesis tests that are employed in inferential statistics.

- ① ***Z Test***

  greater

  When data has a normal distribution and a sample size of at least 30, the **z test** is applied to the data. When the population variance is known, it determines if the sample and population means are equal. The following setup can be used to test the right-tailed hypothesis:

  **Null Hypothesis:** $H_0: \mu=\mu_0$

  **Alternate hypothesis:** $H_1: \mu>\mu_0$

  **Test Statistic:** Z Test = $(\bar{x} - \mu) / (\sigma / \sqrt{n})$

  where,

  $\bar{x}$ = sample mean

  μ = population mean

  σ = standard deviation of the population

  n = sample size

  **Decision Criteria:** If the z statistic > z critical value, reject the null hypothesis.

  [illegible]

  ① Normal distribution

  ② sample size >30

  ③ $\boxed{\mu = \mu_0}$

  $\boxed{\alpha = 0.05}$

- ***T Test***

  When the sample size is less than 30, and the data has a student t distribution, a **t test** is utilized. The sample and population mean are compared when the population variance is unknown. The inferential statistics hypothesis test is as follows:

  **Null Hypothesis:** $H_0: \mu=\mu_0$

  **Alternate Hypothesis:** $H_1: \mu>\mu_0$

  **Test Statistic:** $t = \bar{x} - \mu / s\sqrt{n}$

  ① sample size < 30

  30

**Alternate Hypothesis:** $H_1$: $\mu > \mu_0$

30

**Test Statistic:** t = x̄−μ / s√n

The representations x̄, μ, and n are the same as stated for the z-test. The letter "s" represents the standard deviation of the sample.

**Decision Criteria:** If the t statistic > t critical value, reject the null hypothesis.

- ***F Test***
  When comparing the variances of two samples or populations, an **f test** is used to see if there is a difference. The right-tailed f test can be configured as follows:

**Null Hypothesis:** H0 :σ21 =σ22

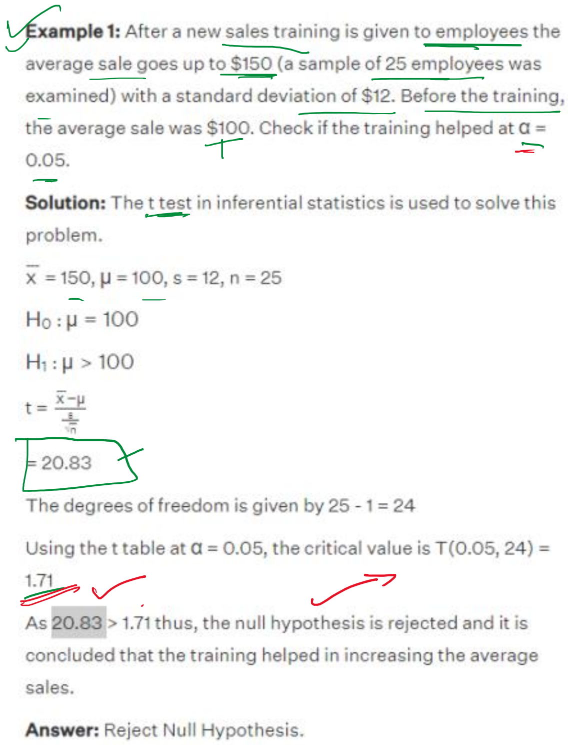
**Alternate Hypothesis:** H1 :σ21> σ22

**Test Statistic:** f = σ21 / σ22, where σ21 is the variance of the first population, and σ22 is the variance of the second population.

**Decision Criteria:** Deciding Criteria: Reject the null hypothesis if f test statistic > critical value.

right tailed

A confidence interval aids an estimation of a population's parameters. For instance, a 95% confidence interval means that 95 out of 100 tests with fresh samples performed under identical conditions will result in the estimate falling within the specified range. A confidence interval can also be used to determine the crucial value in hypothesis testing.



**Example 1:** After a new sales training is given to employees the average sale goes up to $150 (a sample of 25 employees was examined) with a standard deviation of $12. Before the training, the average sale was $100. Check if the training helped at α = 0.05.

**Solution:** The t test in inferential statistics is used to solve this problem.

$\overline{x}$ = 150, μ = 100, s = 12, n = 25

$H_0 : \mu = 100$

$H_1 : \mu > 100$

$t = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}}$

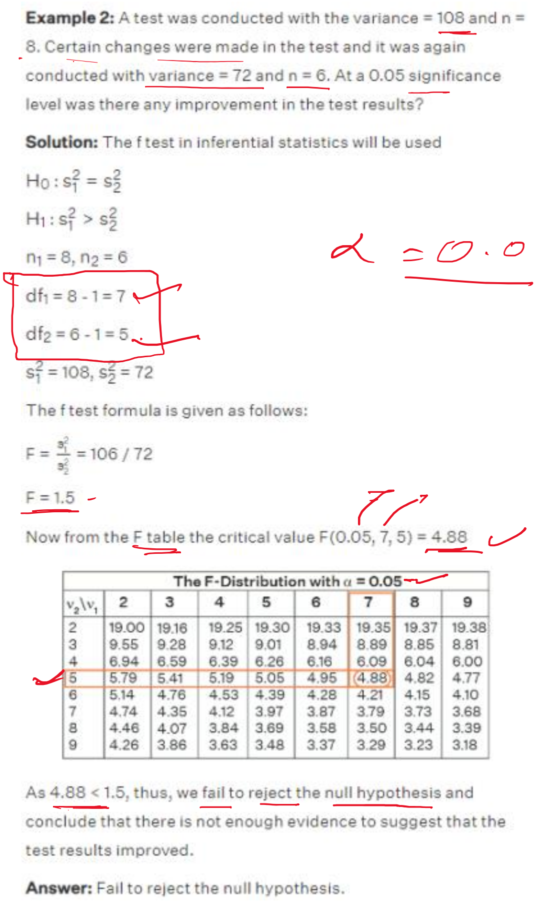
= 20.83

The degrees of freedom is given by 25 - 1 = 24

Using the t table at α = 0.05, the critical value is T(0.05, 24) = 1.71

As 20.83 > 1.71 thus, the null hypothesis is rejected and it is concluded that the training helped in increasing the average sales.

**Answer:** Reject Null Hypothesis.



**Example 2:** A test was conducted with the variance = 108 and n = 8. Certain changes were made in the test and it was again conducted with variance = 72 and n = 6. At a 0.05 significance level was there any improvement in the test results?

**Solution:** The f test in inferential statistics will be used

$H_0 : s_1^2 = s_2^2$

$H_1 : s_1^2 > s_2^2$

$n_1 = 8, n_2 = 6$

$df_1 = 8 - 1 = 7$

$df_2 = 6 - 1 = 5$

$s_1^2 = 108, s_2^2 = 72$

α = 0.05

The f test formula is given as follows:

$F = \frac{s_1^2}{s_2^2}$ = 106 / 72

F = 1.5

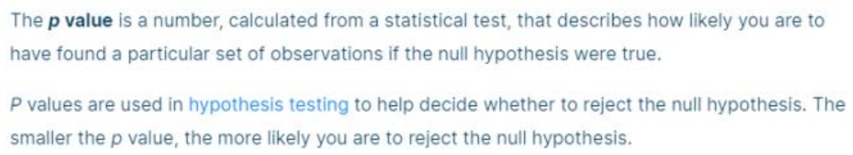
Now from the F table the critical value F(0.05, 7, 5) = 4.88

| The F-Distribution with α = 0.05 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $v_2 \backslash v_1$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 2 | 19.00 | 19.16 | 19.25 | 19.30 | 19.33 | 19.35 | 19.37 | 19.38 |
| 3 | 9.55 | 9.28 | 9.12 | 9.01 | 8.94 | 8.89 | 8.85 | 8.81 |
| 4 | 6.94 | 6.59 | 6.39 | 6.26 | 6.16 | 6.09 | 6.04 | 6.00 |
| 5 | 5.79 | 5.41 | 5.19 | 5.05 | 4.95 | 4.88 | 4.82 | 4.77 |
| 6 | 5.14 | 4.76 | 4.53 | 4.39 | 4.28 | 4.21 | 4.15 | 4.10 |
| 7 | 4.74 | 4.35 | 4.12 | 3.97 | 3.87 | 3.79 | 3.73 | 3.68 |
| 8 | 4.46 | 4.07 | 3.84 | 3.69 | 3.58 | 3.50 | 3.44 | 3.39 |
| 9 | 4.26 | 3.86 | 3.63 | 3.48 | 3.37 | 3.29 | 3.23 | 3.18 |

As 4.88 < 1.5, thus, we fail to reject the null hypothesis and conclude that there is not enough evidence to suggest that the test results improved.

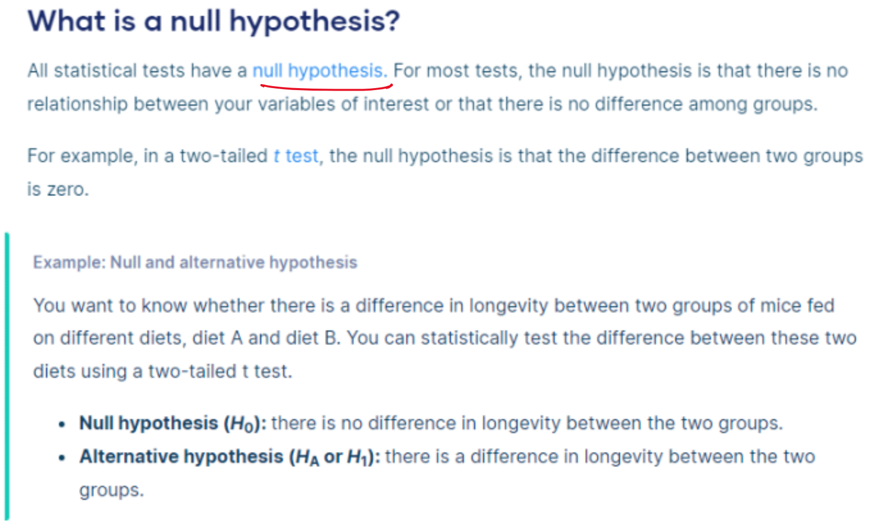
**Answer:** Fail to reject the null hypothesis.

**Confidence Interval:** A confidence interval helps in estimating the parameters of a population. For example, a 95% confidence interval indicates that if a test is conducted 100 times with new samples under the same conditions then the estimate can be expected to lie within the given interval 95 times. Furthermore, a confidence interval is also useful in calculating the critical value in hypothesis testing.



The **p value** is a number, calculated from a statistical test, that describes how likely you are to have found a particular set of observations if the null hypothesis were true.

*P* values are used in hypothesis testing to help decide whether to reject the null hypothesis. The smaller the *p* value, the more likely you are to reject the null hypothesis.

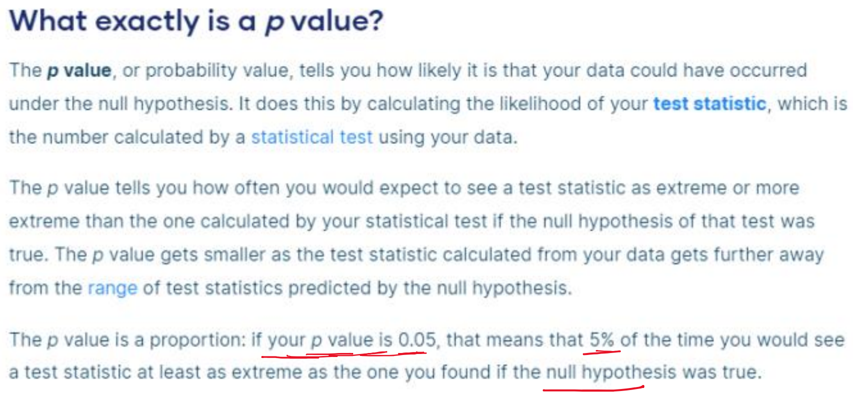


## What is a null hypothesis?

All statistical tests have a null hypothesis. For most tests, the null hypothesis is that there is no relationship between your variables of interest or that there is no difference among groups.

For example, in a two-tailed *t* test, the null hypothesis is that the difference between two groups is zero.

> Example: Null and alternative hypothesis
>
> You want to know whether there is a difference in longevity between two groups of mice fed on different diets, diet A and diet B. You can statistically test the difference between these two diets using a two-tailed t test.
>
> - **Null hypothesis ($H_0$):** there is no difference in longevity between the two groups.
> - **Alternative hypothesis ($H_A$ or $H_1$):** there is a difference in longevity between the two groups.



## What exactly is a *p* value?

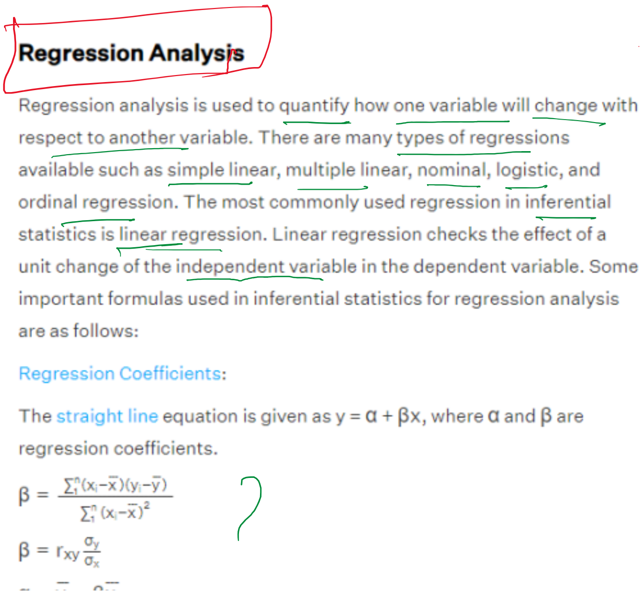
The ***p* value**, or probability value, tells you how likely it is that your data could have occurred under the null hypothesis. It does this by calculating the likelihood of your **test statistic**, which is the number calculated by a statistical test using your data.

The *p* value tells you how often you would expect to see a test statistic as extreme or more extreme than the one calculated by your statistical test if the null hypothesis of that test was true. The *p* value gets smaller as the test statistic calculated from your data gets further away from the range of test statistics predicted by the null hypothesis.

The *p* value is a proportion: if your *p* value is 0.05, that means that 5% of the time you would see a test statistic at least as extreme as the one you found if the null hypothesis was true.

$\alpha$

$\alpha = 0$



## Regression Analysis

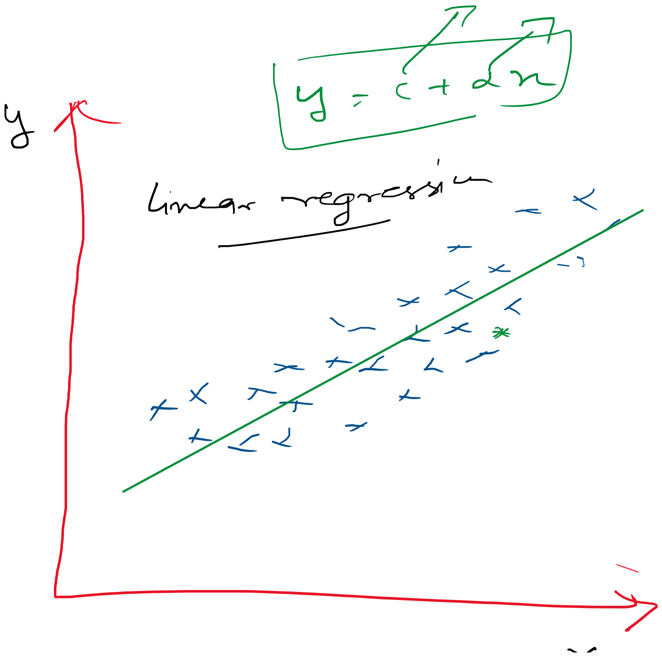
Regression analysis is used to quantify how one variable will change with respect to another variable. There are many types of regressions available such as simple linear, multiple linear, nominal, logistic, and ordinal regression. The most commonly used regression in inferential statistics is linear regression. Linear regression checks the effect of a unit change of the independent variable in the dependent variable. Some important formulas used in inferential statistics for regression analysis are as follows:

Regression Coefficients:

The straight line equation is given as y = α + βx, where α and β are regression coefficients.

$$\beta = \frac{\sum_1^n (x_i-\bar{x})(y_i-\bar{y})}{\sum_1^n (x_i-\bar{x})^2}$$

$$\beta = r_{xy}\frac{\sigma_y}{\sigma_x}$$



$\beta = \frac{}{\sum_1^n (x_i - \bar{x})^2}$

$\beta = r_{xy} \frac{\sigma_y}{\sigma_x}$

$\alpha = \bar{y} - \beta\bar{x}$

Here, $\bar{x}$ is the mean, and $\sigma_x$ is the standard deviation of the first data set. Similarly, $\bar{y}$ is the mean, and $\sigma_y$ is the standard deviation of the second data set.

## House Price Prediction

| Size | Location | , No Rooms | Price |
|---|---|---|---|
| 100 | | | — |
| 120 | | | — |
| 200 | | | — |

new