

Statistics:

statistics is the area of applied math that deals with the collection, organization, analysis, interpretation, and presentation of data.

Descriptive & Inferential Statistics:

Descriptive statistics focus on describing the visible characteristics of a dataset (a population or sample).

Inferential statistics focus on making predictions or generalizations about a larger dataset, based on a sample of those data.

Descriptive Statistics:

Distribution: Distribution shows us the frequency of different outcomes (or data points) in a population or sample. Using visualizations is common practice in descriptive statistics. It helps us more readily spot patterns or trends in a dataset.

-
- **Central tendency:** Central tendency is the name for measurements that look at the typical central values within a dataset.

Common measures of central tendency include:

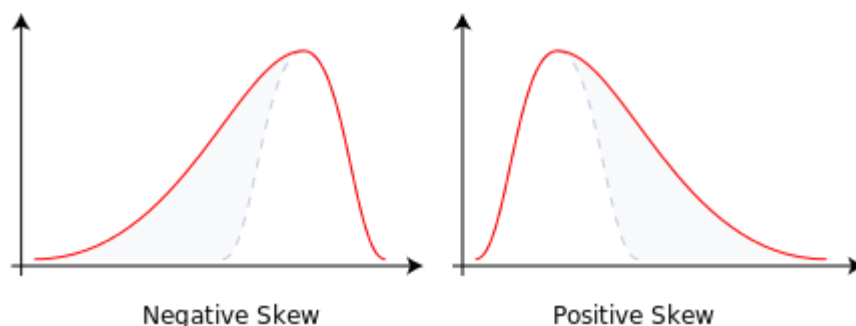
- **The mean:** The average value of all the data points.
- **The median:** The central or middle value in the dataset.
- **The mode:** The value that appears most often in the dataset
- **Variability:** The variability, or dispersion, of a dataset, describes how values are distributed or spread out.

Common measures of variability include:

- **Standard deviation:** This shows us the amount of variation or dispersion. Low standard deviation implies that most values are close

to the mean. High standard deviation suggests that the values are more broadly spread out.

- **Range:** This measures the size of the distribution of values. This can be easily determined by subtracting the smallest value from the largest.
- **Kurtosis:** This measures whether or not the tails of a given distribution contain extreme values (also known as outliers). If a tail lacks outliers, we can say that it has low kurtosis. If a dataset has a lot of outliers, we can say it has high kurtosis. Leptokurtic, Mesokurtic & Platykurtic
- **Skewness:** This is a measure of a dataset's symmetry. If you were to plot a bell-curve and the right-hand tail was longer and fatter, we would call this positive skewness. If the left-hand tail is longer and fatter, we call this negative skewness. This is visible in the following image.



•
Inferential Statistics:

Inferential statistics focus on making generalizations about a larger population based on a representative sample of that population. Because inferential statistics focuses on making predictions (rather than stating facts) its results are usually in the form of a probability.

How do we obtain a random sample?

Random sampling can be a complex process and often depends on the particular characteristics of a population. However, the fundamental principles involve:

1. Defining a population
2. Deciding your sample size
3. Randomly select a sample
4. Analyze the data Sample

Inferential Statistics involves:

- **Hypothesis testing:** Hypothesis testing involves checking that your samples repeat the results of researcher's hypothesis (or proposed explanation). The aim is to rule out the possibility that a given result has occurred by chance.
- **Confidence intervals:** Confidence intervals are used to estimate certain parameters for a measurement of a population (such as the mean) based on sample data. Rather than providing a single mean value, the confidence interval provides a range of values. This is often given as a percentage.
- **Regression and correlation analysis:** Regression and correlation analysis are both techniques used for observing how two (or more) sets of variables relate to one another.
- Regression analysis aims to determine how one dependent (or output) variable is impacted by one or more independent (or input) variables. It's often used for hypothesis testing and predictive analytics. Correlation analysis, meanwhile, measures the degree of association between two or more datasets.

Descriptive statistics:

- Describe the features of populations and/or samples.

- Organize and present data in a purely factual way.
- Present final results visually, using tables, charts, or graphs.
- Draw conclusions based on known data.
- Use measures like central tendency, distribution, and variance.

Inferential statistics:

- Use samples to make generalizations about larger populations.
- Help us to make estimates and predict future outcomes.
- Present final results in the form of probabilities.
- Draw conclusions that go beyond the available data.
- Use techniques like hypothesis testing, confidence intervals, and regression and correlation analysis.

t-Test assumptions

While t-tests are relatively robust to deviations from assumptions, t-tests do assume that:

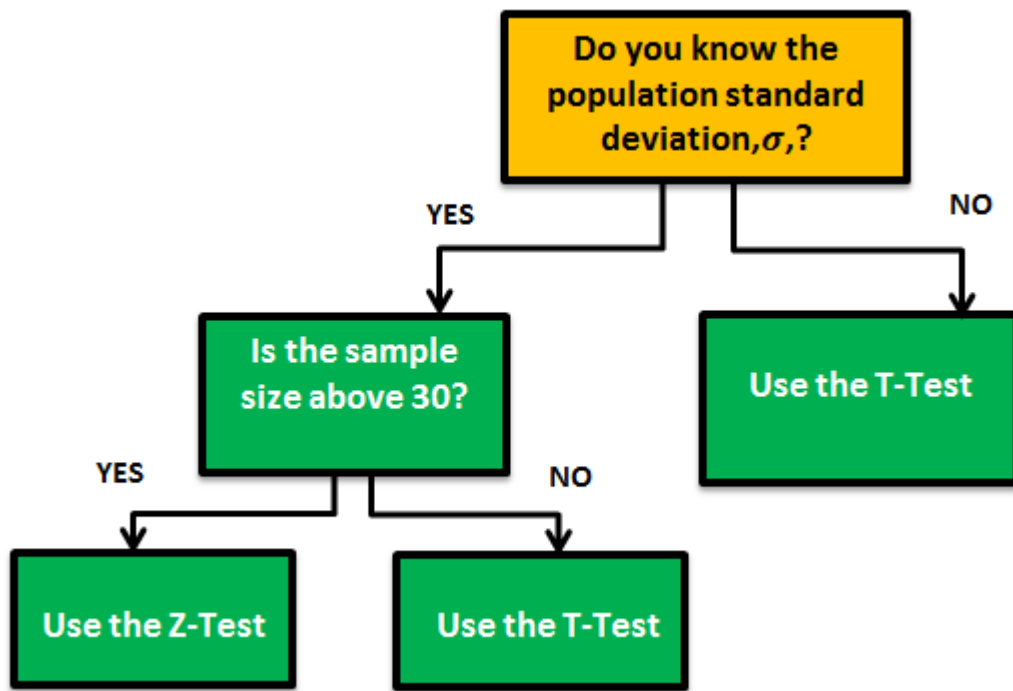
- ☐ The data are continuous.
- ☐ The sample data have been randomly sampled from a population.
- ☐ There is homogeneity of variance (i.e., the variability of the data in each group is similar).
- ☐ The distribution is approximately normal.

	<u>One-sample <i>t</i>-test</u>	<u>Two-sample <i>t</i>-test</u>	<u>Paired <i>t</i>-test</u>
--	---	---	---

Synonyms	Student's <i>t</i> -test	- Independent groups <i>t</i>-test - Independent samples <i>t</i>-test - Equal variances <i>t</i>-test - Pooled <i>t</i>-test - Unequal variances <i>t</i>-test	- Paired groups <i>t</i>-test - Dependent samples <i>t</i>-test
Number of variables	One	Two	Two
Type of variable	Continuous measurement	- Continuous measurement - Categorical or Nominal to define groups	- Continuous measurement - Categorical or Nominal to define pairing within group
Purpose of test	Decide if the population mean is equal to a specific value or not	Decide if the population means for two different groups are equal or not	Decide if the difference between paired measurements for a population is zero or not
Example: test if...	Mean heart rate of a group of people is equal to 65 or not	Mean heart rates for two groups of people are the same or not	Mean difference in heart rate for a group of people before and after exercise is zero or not
Estimate of population mean	Sample average	Sample average for each group	Sample average of the differences in paired measurements
Population standard deviation	Unknown, use sample standard deviation	Unknown, use sample standard deviations for each group	Unknown, use sample standard deviation of differences in paired measurements
Degrees of freedom	Number of observations in sample minus 1, or: $n-1$	Sum of observations in each sample minus 2, or: $n_1 + n_2 - 2$	Number of paired observations in sample minus 1, or: $n-1$

--	--	--	--

Basis	Z Test	T-Test
Basic Definition	Z-test is a kind of hypothesis test which ascertains if the averages of the 2 datasets are different from each other when standard deviation or variance is given.	The t-test can be referred to as a kind of parametric test that is applied to an identity, how the averages of 2 sets of data differ from each other when the standard deviation or variance is not given.
Population Variance	The Population variance or standard deviation is known here.	The Population variance or standard deviation is unknown here.
Sample Size	The Sample size is large.	Here the Sample Size is small.
Key Assumptions	• All data points are independent. • Normal Distribution for Z, with an average zero and variance = 1.	• All data points are not dependent. • Sample values are to be recorded and taken accurately.
Based upon (a type of distribution)	Based on Normal distribution .	Based on Student-t distribution.



F Test	T-Test
An F test is a test statistic used to check the equality of variances of two populations	The T-test is used when the sample size is small ($n < 30$) and the population standard deviation is not known.
The data follows an F distribution	The data follows a Student t-distribution
The f test is used for variances.	It is used for testing means .