**Population** : The entire group one desires information about.  
**Sample** : A subset of the population taken because the entire population is usually too large to analyze. It's characteristics are taken to be representative of the population.  
  
**Mean** : The sum of all the values in the sample divided by the number of values in the sample/population.  
  
**Median** : The median is the value separating the higher half of a data sample from the lower half.  
  
**Standard Deviation** : Square root of the variance. It measures the dispersion around the mean.  
  
**Percentiles** : An extension of median to values other than 50%.  
  
**Interquartile range (IQR)** : the difference between the 75th and 25th percentile  
**Mode** : The most frequently occuring value  
**Range** : Difference between the maximum value and the minimum value.

Notice that most of these fall into one of two categories: they capture either the center of the distribution (e.g., mean, median, mode), or its spread (e.g., variance, IQR, range). These two categories are often called **measures of central tendency** and **measures of dispersion**, respectively.

Important Distributions

**Gaussian/Normal** : We say x ∼ N (µ, σ2) to mean that x is drawn from a Gaussian (or Normal) distribution with mean µ and  
variance σ2 (or equivalently standard deviation σ). We’ll often use the standard normal distribution, or N (0, 1) (i.e., mean 0 and variance 1).

* The probability of getting a value within 1 standard deviation of the mean is about 68%. For 2 standard deviations, it’s about 95%, and for 3 standard deviations it’s about 99%. This is sometimes called the **“68-95-99 rule”**.

**Bernoulli** : A Bernoulli random variable can be thought of as the outcome of flipping a biased coin, where the probability of heads is p. To be more precise, a Bernoulli random variable takes on value 1 with probability p and value 0 with probability 1−p. Its expectation is p, and its variance is p(1 − p).

Bernoulli variables are typically used to model binary random variables.

Central Limit Theorem

The Central Limit Theorem states that the sampling distribution of the sample means approaches a normal distribution as the sample size gets larger — no matter what the shape of the population distribution. This fact holds especially true for sample sizes over 30.

Confidence Intervals

A confidence interval is how much uncertainty there is with any particular statistic. Confidence intervals are often used with a margin of error. It tells you how confident you can be that the results from a poll or survey reflect what you would expect to find if it were possible to survey the entire population.

Hypothesis Testing

Hypothesis testing in statistics is a way for you to test the results of a survey or experiment to see if you have meaningful results.  
  
Hypothesis testing steps:

* Figure out your null hypothesis,
* State your null hypothesis,
* Choose what kind of test you need to perform,
* Either support or reject the null hypothesis.

Significance Level

The significance level α is the probability of making the wrong decision when the null hypothesis is true. Alpha levels (sometimes just called “significance levels”) are used in hypothesis tests. Usually, these tests are run with an alpha level of .05 (5%), but other levels commonly used are .01 and .10.

P-value

A p value is used in hypothesis testing to help you support or reject the null hypothesis. The p value is the evidence against a null hypothesis. The smaller the p-value, the strong the evidence that you should reject the null hypothesis.

A/B testing

A statistical way of comparing two (or more) techniques, typically an incumbent against a new rival. A/B testing aims to determine not only which technique performs better but also to understand whether the difference is statistically significant. A/B testing usually considers only two techniques using one measurement, but it can be applied to any finite number of techniques and measures.

Correlation

Correlation is a statistical measure that describes the association between random variables.

Types of Correlation

* Pearson Correlation Coefficient (measures the linear association between continuous variables)
* Spearman's Correlation (special case of Pearson ρ applied to ranked (sorted) variables. appropriate to use with both continuous and discrete data.)
* Kendall's Tau (more appropriate for discrete data.)

Statistical Hypothesis Tests ⭐️⭐️

Normality Tests (Statistical tests that you can use to check if your data has a Gaussian distribution.)

* **Shapiro-Wilk Test** : Tests whether a data sample has a Gaussian distribution.

Correlation Tests (Statistical tests that you can use to check if two samples are related)

* **Pearson’s Correlation Coefficient** : Tests whether two samples have a monotonic relationship.
* **Spearman’s Rank Correlation** : Tests whether two samples have a monotonic relationship.
* **Chi-Squared Test** : Tests whether two categorical variables are related or independent.

Parametric Statistical Hypothesis Tests (Statistical tests that you can use to compare data samples.)

* **Student’s t-test** : Tests whether the means of two independent samples are significantly different.
* **Paired Student’s t-test** : Tests whether the means of two paired samples are significantly different.
* **Analysis of Variance Test (ANOVA)** : Tests whether the means of two or more independent samples are significantly different.
* **Repeated Measures ANOVA Test** : Tests whether the means of two or more paired samples are significantly different.

Nonparametric Statistical Hypothesis Tests

* + **Mann-Whitney U Test** : Tests whether the distributions of two independent samples are equal or not.
  + **Wilcoxon Signed-Rank Test** : Tests whether the distributions of two paired samples are equal or not.
  + **Kruskal-Wallis H Test** : Tests whether the distributions of two or more independent samples are equal or not.
  + **Friedman Test** : Tests whether the distributions of two or more paired samples are equal or not.