**STATISTICS INTERVIEW QUESTIONS**

**What is statistics, and what are branches of statistics?**

Statistics is mathematical science pertaining to the collection, presentation, analysis, and interpretation of data. It is used for:

* Analyze Primary Data
* Build a Statistical Model
* Predict the outcomes

The two main branches of statistics are [descriptive statistics](https://explorable.com/descriptive-statistics) and [inferential statistics](https://explorable.com/inferential-statistics).

**Please explain descriptive statistics.**

Helps organize data and focuses on the main characteristics of the data. It provides a summary of data numerically or graphically. It describes the basic features of the data in a study, such as mean, median, mode, SD, correlation.

Descriptive statistics are either measures of central tendency or measures of spread.

* Measures of central tendency: mean, median, and mode
* Measures of spread: standard deviation, variance, the minimum and maximum variables, range and the kurtosis and skewness.

## What are the data types? Please explain.

1. Numerical data represents some quantifiable information that is measurable and is further divided into two subcategories:

Discrete data, which is integer based (e.g. number of people)

Continuous data, which is decimal based (e.g. price, distance, temperature).

2. Categorical data is qualitative data that is used to classify data into categories (think of an enumeration in programming). For example, gender, car brands, country of residence etc.

3. Ordinal data represents discrete and ordered units, e.g. champions league rank (1st, 2nd, 3rd), bug priority (low, critical or showstopper), or hotel rating (1–5\*).

## What is variable and what are categorical, discrete, and continuous variables?

A variable is a characteristic that describes a member of the sample.

Categorical variable

Categorical variables contain a finite number of categories or distinct groups. gender, material type, and payment method.

Discrete variable

Discrete variables are numeric variables that have a countable number of values between any two values.

Continuous variable

Continuous variables are numeric variables that have an infinite number of values between any two values.

## Please explain population and sample terms. What is the difference between a population and a sample?

Population: A population is the group from which data is to be collected.

Sample: A subset of population

A population describes all of the members of a group, while a sample is a subset of members that time and resources allow you to measure.

The CLT states that the mean values from a group of samples will be normally distributed about the population mean, even if the population itself is not normally distributed.

The central limit theorem states that, when samples from a data set with a known variance are aggregated, their mean roughly equals the population mean.

Thus, even though we might not know the shape of the distribution where our data comes from, the central limit theorem says that we can treat the sampling distribution as if it were normal.

Suppose that we are interested in estimating the average height among all people. Collecting data for every person in the world is impossible. While we can’t obtain a height measurement from everyone in the population, we can still sample some people. The question now becomes, what can we say about the average height of the entire population given a single sample. The Central Limit Theorem addresses this question exactly.

**What are the measures of variability (dispersion)? Please explain.**

Range: Range is the difference between the lowest and the highest number of a dataset. To calculate the range, we subtract the minimum from the maximum value.

It shows us how varied the dataset is, i.e. how spread it is, but again, like mean, it is really sensitive to outliers.

Variance: Variance measures how spread out the data is. It describes how far values lie from the mean. Calculated as the sum of square distances from each point to the mean.

●          Calculate mean.

●          Take difference between each value and the mean

●          Square this difference.

●          Sum all differences

●          Finally, divide by the total number of observations.

There’s a difference between the SAMPLE variance and the POPULATION variance

subject to Bessel's correction (𝒏 − 𝟏)

* For full population, we divide by the number of data points (n)
* samples, we divide by the number of data points minus 1(n - 1)

Standard Deviation: Standard deviation (represented by the greek letter sigma — σ) is just the square root of the variance. It is a measure of dispersion in terms of how many standard deviations it is away from the mean, and used to judge which data point is an outlier.

Benefit: Same units as the sample and meaningful to talk about.

## What is an interquartile range?

The IQR is a number that indicates how spread the middle half (i.e. the middle 50%) of the dataset is and can help determine outliers. It is the difference between the Q3 and Q1.

Generally speaking, outliers are those data points that fall outside from the Q1 – 1.5 x IQR and Q3 + 1.5 x IQR range.

## What are levels of measurement?

Based on the nature of variables, it is classified into four types:

Nominal Variables: They are ones which have two or more categories, and it is impossible to order the values. Predetermined categories and can’t be sorted.

Ex gender and blood group

Ordinal Variables: They have values in a logical order. However, the relative distance between two data values is not clear. Can be sorted and lacks scale.

Ex the size of a coffee cup – large, medium, and small

The ratings of a product – bad, good, and best

Interval Variable: With an interval scale, equal differences between scale values do have equal quantitative meaning. An interval scale provides more quantitative information than the ordinal scale. The interval scale does not have a true zero point. Provides scale and lacks a ‘zero’ point.

Ex. The Fahrenheit degree scale used to measure temperature. The distance between two compartments in a train.

Ratio Variable: They are similar to interval scales in that equal differences between scale values have equal quantitative meaning. It has a true zero point. Ex the system of inches used with a common ruler.

## What is the difference between covariance and correlation?

Correlation and Co-variance both are used as a measure to check how two variables change with respect to each other.

Covariance:

- measure of change of how two variables change with respect to each other.

- It is unit dependent.

- Difference in scale can output different co-variance.

- It varies from - infinity to infinity

- Example: Height vs weight(kg) and Height vs weight(lbs) will have different covariance values

Correlation :

- unit measure of change between two variables change with respect to each other. ( Basically it’s normalized covariance value)

- It is unit independent.

- Difference in scale doesn’t affect the correlation value

- It varies from -1 to 1.

- Example: Height vs weight(kg) and Height vs weight(lbs) will have same correlation values.

## What are depend or independ events?

Independent Events

An independent series of events occur when the outcome of one event has no effect on the outcome of another.

Independent are those events whose outcome does not influence the probability of the outcome of the other event. Due to this reason:

P(A|B) = P(A)

P(B|A) = P(B)

Dependent Events

A dependent event occurs when the outcome of a first event does affect the probability of a second event.

## What is conditional probability? Please explain.

Conditional probability is the likelihood of an event occurring, based on the occurrence of a previous event.

The notation for conditional probability is P(A|B), read as ‘the probability of A given B’. The formula for conditional probability is:

The probability of A given B is equal to the probability of A and B occurring over the probability of B alone occurring.

Conditional probability does not indicate that there is necessarily a causal relationship between the two events, nor that both events occur simultaneously.

## What is Bayes' Teorem? Please explain.

Bayes' theorem describes the [probability](https://en.wikipedia.org/wiki/Probability) of an [event](https://en.wikipedia.org/wiki/Event_(probability_theory)), based on prior knowledge of conditions that might be related to the event. The probability of A given B is equal to the probability of B given A times the probability of A, over the probability of B.

For example, if cancer is related to age, then, using Bayes' theorem, a person's age can be used to more accurately assess the probability that they have cancer.

A is called the proposition and B is called the evidence.

P(A) is called the prior probability of proposition and P(B) is called the prior probability of evidence.

P(A|B) is called the posterior.     P(B|A) is the likelihood.

Posterior=(Likelihood)\*(Proposition prior probability) / Evidence prior probability

The more variables that are in play, and the more certain we become about those variables, the more certain an accurate conclusion can be drawn, using conditional probabilities!

**What is poisson distributions? Specify discrete probability distributions categories.**

The Poisson distribution gives the probability of a number of events occurring in a fixed interval of time, if these events happen:

* with a known average rate and
* independently of the time since the last event.

A Poisson Distribution considers the number of successes per unit of time\* over the course of many units

\* or any other continuous unit, e.g. distance

The average number of homes sold by the Acme Realty company is 2 homes per day. What is the probability that exactly 3 homes will be sold tomorrow?   μ=2, x=3,e=2.71...

P(x; μ) = (e-μ) (μx) / x!

A classic example here is the number of phone calls received by a call centre.

 — Or —

If we know the average number of things that happen in a given time period, another use-case can be to predict the odds of getting another value instead, on a given future time. E.g. my Medium posts get an average of 1,000 views per day; I can use the Poisson probability mass function to estimate the probability of having 1,500 visits.

## What is probability?

Probability is a value between 0 and 1 that a certain event will occur

the act of flipping a coin is called a trial.

Each trial of flipping a coin can be called an experiment.

Each mutually exclusive outcome is called a simple event.

The sample space is the sum of every possible simple event.

## What is binomial distribution?

“Binomial” means there are two discrete, mutually exclusive outcomes of a trial. Let’s consider an experiment having two possible outcomes: either success or failure. Suppose the experiment is repeated several times and the repetitions are independent of each other. The total number of experiments where the outcome turns out to be a success is a random variable whose distribution is called binomial distribution.

Gives the probability of observing x successes in n trials,.

Suppose a die is tossed 5 times. What is the probability of getting exactly 2 fours?

b(x; n, P) = { n! / [ x! (n - x)! ] } \* Px \* (1 - P)n - x

x=2, n=5, p=1/6

## What is bernoulli distribution?

A Bernoulli Trial is a random experiment in which there are only two possible outcomes

- success or failure

● A series of trials n will follow a binary distribution so long as

a) the probability of success p is constant

b) trials are independent of one another

Gives the probability of observing x successes in n trials

● The probability of success on a single trial is denoted by p

● Assumes that p is fixed for all trials

## What do you understand by the term normal distribution?

It is a set of continuous variables spread across a normal curve or in the shape of a bell curve. It can be considered as a continuous probability distribution and is useful in statistics. It is the most common distribution curve and it becomes very useful to analyze the variables and their relationships when we have the normal distribution curve. The normal distribution curve is symmetrical. The non-normal distribution approaches the normal distribution as the size of the samples increases. It is also very easy to deploy the Central Limit Theorem. This method helps to make sense of data that is random by creating an order and interpreting the results using a bell-shaped graph.

Data is usually distributed in different ways with a bias to the left or to the right or it can all be jumbled up. However, there are chances that data is distributed around a central value without any bias to the left or right and reaches normal distribution in the form of a bell shaped curve. The random variables are distributed in the form of an symmetrical bell shaped curve.

Properties of Normal Distribution:

Unimodal -one mode Symmetrical -left and right halves are mirror images Bell-shaped -maximum height (mode) at the mean Mean, Mode, and Median are all located in the center Asymptotic.

## For what purpose are z and t tests used in statistics?

We use Z- and t-Distributions to answer the question “ What is the probability that two samples come from the same population?

Either test assess whether mean of two groups are statistically different from each other or not.

z-test follows normal-distribution, which is appropriate when the sample size is large, and the population standard deviation is known.

T-test follows t-distribution, which is appropriate when the sample size is small, and the population standard deviation is not known.

## What do you mean by confidence interval in statistical analysis?

It is an interval estimate for a parameter value. It is constructed in a way so that, in the long run, a given proportion of these intervals will include the unknown true parameter value. The proportion is given by the "level of confidence". For instance, you can expect that at least 90% of (a large series of) 90% confidence intervals will include the unknown true values of the parameters.

## What is a p-value? Why is it important?

When you perform a hypothesis test in statistics, a p-value can help you determine the strength of your results. P-value is a number between 0 and 1. Based on the value it will denote the strength of the results. The claim which is on trial is called Null Hypothesis.

## What are measures of central tendency?

Mean: The mean (represented by the greek letter mu— μ) is the average of a dataset. To calculate the mean, we sum up all the values and divide it by the number of values.

Median: The median is the middle of a dataset. To calculate the median, we sort all the values (in ascending or descending order) and take the one that is in the middle.

If there is an even number of data points, then we calculate the mean of the two that fall in the middle.

The median is less susceptible to outliers than the mean, and hence we need to take into consideration how the data distribution looks like, to choose which one to use. The median is much closer to most of the values in the series!

Mode: The mode is the most common/occurring value in the dataset. To calculate the mode, we locate the number that occurs more frequently.

Mode is usually only relevant to discrete numerical data — not to continuous data.

**What is hypothesis testing? What are hypothesis testing types? Please explain with examples.**

A hypothesis is some kind of assumption and hypothesis testing is an inferential statistical technique to determine whether there is enough evidence in a data sample to infer that a certain condition holds true for the entire population. It is used to test the validity of a claim that is made about a population.

* Take a random sample
* Analyze the properties of the sample
* Test whether or not the identified conclusions correctly represent the population
* A hypothesis is generated about a population parameter.

Null Hypothesis: The null hypothesis is assumed to be true unless there is a strong evidence to the contrary. This claim that’s on trial is called the [null hypothesis](https://www.dummies.com/education/math/statistics/how-to-determine-a-p-value-when-testing-a-null-hypothesis/).

Alternative Hypothesis: The alternative hypothesis is assumed to be true when the null hypothesis is proved false. The alternative hypothesis is the one you would believe if the null hypothesis is concluded to be untrue.

Framing Hypothesis:

● At the start of the experiment, the null hypothesis is assumed to be true.

● If the data fails to support the null hypothesis, only then can we look to an alternative hypothesis

\* If testing something assumed to be true, the null hypothesis can reflect the assumption:

Claim:  “Our product has an average shipping weight of 3.5kg.”

Null hypothesis: average weight = 3.5kg

Alternate hypothesis: average weight ≠ 3.5kg

\* If testing a claim we want to be true, but can’t assume, we test its opposite:

Claim:  “This prep course improves test scores.”

Null hypothesis: old scores ≥ new scores

Alternate hypothesis: old scores < new scores

The null hypothesis should contain an equality (=, ≤ ,≥)

The alternate hypothesis should not have an equality (≠,<,>)

## What do you understand by Type I vs Type II error?

Type I error is committed when the null hypothesis is true and we reject it, also known as a ‘False Positive’. Type II error is committed when the null hypothesis is false and we fail to reject it, also known as ‘False Negative'.

## What Is the Goal Of A/B Testing?

It is a statistical hypothesis testing for randomized experiment with two variables A and B. The objective of A/B Testing is to detect any changes to the web page to maximize or increase the outcome of an interest. Example: Identify the click-through rate for a banner ad.

A/B testing is a fantastic method for figuring out the best online promotional and marketing strategies for business. It can be used to test everything from website copy to sales emails to search ads An example of this could be identifying the click-through rate for a banner ad.

## Why statistical knowladge are important in data science?

As Josh Wills once said,

“Data Scientist is a person who is better at statistics than any programmer and better at programming than any statistician.”