1. **What is data and how is it defined?**

Data refers to any information, facts, or statistics that can be collected, stored, and analyzed. It can be in various forms such as text, numbers, images, audio, or video.

2. **What are the different types of data, and how are they classified?**

There are mainly four types of data: nominal, ordinal, interval, and ratio. Nominal data is arranged, ordinal data has an order, interval data has a fixed scale, and ratio data has a true zero point.

3. **What is the importance of data in decision-making processes?**

Data is essential for decision-making as it provides valuable insights into trends, patterns, and behaviors that help in making informed decisions.

4. **How is data collected, stored, and processed for analysis?**

Data can be collected through surveys, experiments, sensors, or from various online sources. It is then stored in databases, data warehouses, or in the cloud. It is processed using tools like statistical software, machine learning algorithms, and visualization tools.

5. **What are the ethical concerns surrounding data collection and usage?**

There are concerns around privacy, security, bias, and transparency when it comes to collecting, storing, and using data. It's essential to ensure that data is collected and used ethically and responsibly.

6. **What are some common challenges associated with working with data?**

Common challenges include data quality, data integration, data privacy, data security, data silos, and data governance.

7. **How do companies use data to improve their business operations?**

Companies use data to gain insights into customer behaviour, market trends, and operational performance. This information is then used to improve products and services, optimize operations, and reduce costs.

8. **What are the current trends in data analytics and data management?**

Current trends include the adoption of cloud-based analytics, the use of machine learning and artificial intelligence, the rise of data storytelling, and the growing importance of data governance.

9. **What role does data play in machine learning and artificial intelligence?**

Data is crucial in machine learning and artificial intelligence as algorithms rely on data to learn and make predictions. The quality and quantity of data can impact the accuracy and reliability of the machine learning models.

10. What is data visualization?

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. Additionally, it provides an excellent way for employees or business owners to present data to non-technical audiences without confusion.

In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.

11. What are the advantages and disadvantages of data visualization?

Something as simple as presenting data in graphic format may seem to have no downsides. But sometimes data can be misrepresented or misinterpreted when placed in the wrong style of data visualization. When choosing to create data visualization, it's best to keep both the advantages and disadvantages in mind.

Some other advantages of data visualization include:

- Easily sharing information.
- Interactively explore opportunities.
- Visualize patterns and relationships.

Disadvantages

- Biased or inaccurate information.
- Correlation doesn't always mean causation.
- Core messages can get lost in translation.

12. Why data visualization is important

The importance of data visualization is simple: it helps people see, interact with, and better understand data. Whether simple or complex, the right visualization can bring everyone on the same page, regardless of their level of expertise.

13. What is zero skew?
14. What is right skew (positive skew)?
15. What is left skew (negative skew)?
16. How to calculate skewness ?
17. What to do if your data is skewed
18. What are the four levels of measurement?

Levels of measurement tell you how precisely variables are recorded. There are 4 levels of measurement, which can be ranked from low to high:

Nominal: the data can only be categorized.

Ordinal: the data can be categorized and ranked.

Interval: the data can be categorized and ranked, and evenly spaced.

Ratio: the data can be categorized, ranked, evenly spaced and has a natural zero.

What are the three types of skewness?

The three types of skewness are:

- **Right skew (also called positive skew).** A right-skewed distribution is longer on the right side of its peak than on its left.
- **Left skew (also called negative skew).** A left-skewed distribution is longer on the left side of its peak than on its right.
- **Zero skew.** It is symmetrical and its left and right sides are mirror images.

Click to enlarge

What's the difference between the arithmetic and geometric means?

The **arithmetic mean** is the most commonly used type of mean and is often referred to simply as "the mean." While the arithmetic mean is based on adding and dividing values, the **geometric mean** multiplies and finds the root of values.

Even though the geometric mean is a less common measure of central tendency, it's more accurate than the arithmetic mean for percentage change and positively skewed data. The geometric mean is often reported for financial indices and population growth rates.

What are outliers?

Outliers are extreme values that differ from most values in the dataset. You find outliers at the extreme ends of your dataset.

When should I remove an outlier from my dataset?

It's best to remove outliers only when you have a sound reason for doing so.

Some outliers represent natural variations in the population, and they should be left as is in your dataset. These are called true outliers.

Other outliers are problematic and should be removed because they represent measurement errors, data entry or processing errors, or poor sampling.

How do I find outliers in my data?

You can choose from four main ways to detect outliers:

- Sorting your values from low to high and checking minimum and maximum values
- Visualizing your data with a box plot and looking for outliers
- Using the interquartile range to create fences for your data
- Using statistical procedures to identify extreme values

Why do outliers matter?

Outliers can have a big impact on your statistical analyses and skew the results of any hypothesis test if they are inaccurate.

These extreme values can impact your statistical power as well, making it hard to detect a true effect if there is one.

What are the different types of means?

The arithmetic mean is the most commonly used mean. It's often simply called the mean or the average. But there are some other types of means you can calculate depending on your research purposes:

- **Weighted mean:** some values contribute more to the mean than others.
- **Geometric mean:** values are multiplied rather than summed up.
- **Harmonic mean:** reciprocals of values are used instead of the values themselves.

How do I find the mean?

You can find the mean, or average, of a data set in two simple steps:

- Find the sum of the values by adding them all up.
- Divide the sum by the number of values in the data set.

This method is the same whether you are dealing with sample or population data or positive or negative numbers.

When should I use the median?

The median is the most informative measure of central tendency for skewed distributions or distributions with outliers. For example, the median is often used as a measure of central tendency for income distributions, which are generally highly skewed.

Because the median only uses one or two values, it's unaffected by extreme outliers or non-symmetric distributions of scores. In contrast, the mean and mode can vary in skewed distributions.

How do I find the median?

To find the median, first order your data. Then calculate the middle position based on *n*, the number of values in your data set.

- If *n* is an odd number, the median lies at the position $\frac{(n+1)}{2}$.
- If *n* is an even number, the median is the mean of the values at positions $\frac{n}{2}$ and $(\frac{n}{2})+1$.

Can there be more than one mode?

A data set can often have no mode, one mode or more than one mode – it all depends on how many different values repeat most frequently.

Your data can be:

- without any mode
- unimodal, with one mode,
- bimodal, with two modes,
- trimodal, with three modes, or
- multimodal, with four or more modes.

How do I find the mode?

To find the mode:

- If your data is numerical or quantitative, order the values from low to high.
- If it is categorical, sort the values by group, in any order.

Then you simply need to identify the most frequently occurring value.

When should I use the interquartile range?

The interquartile range is the best measure of variability for skewed distributions or data sets with outliers. Because it's based on values that come from the middle half of the distribution, it's unlikely to be influenced by outliers.

What are the two main methods for calculating interquartile range?

The two most common methods for calculating interquartile range are the exclusive and inclusive methods.

The exclusive method excludes the median when identifying Q1 and Q3, while the inclusive method includes the median as a value in the data set in identifying the quartiles.

For each of these methods, you'll need different procedures for finding the median, Q1 and Q3 depending on whether your sample size is even- or odd-numbered. The exclusive method works best for even-numbered sample sizes, while the inclusive method is often used with odd-numbered sample sizes.

What's the difference between the range and interquartile range?

While the range gives you the spread of the whole data set, the interquartile range gives you the spread of the middle half of a data set.

What's the difference between standard deviation and variance?

Variance is the average squared deviations from the mean, while standard deviation is the square root of this number. Both measures reflect variability in a distribution, but their units differ:

- Standard deviation is expressed in the same units as the original values (e.g., minutes or meters).
- Variance is expressed in much larger units (e.g., meters squared).

Although the units of variance are harder to intuitively understand, variance is important in statistical tests.

What is the empirical rule?

The empirical rule, or the 68-95-99.7 rule, tells you where most of the values lie in a normal distribution:

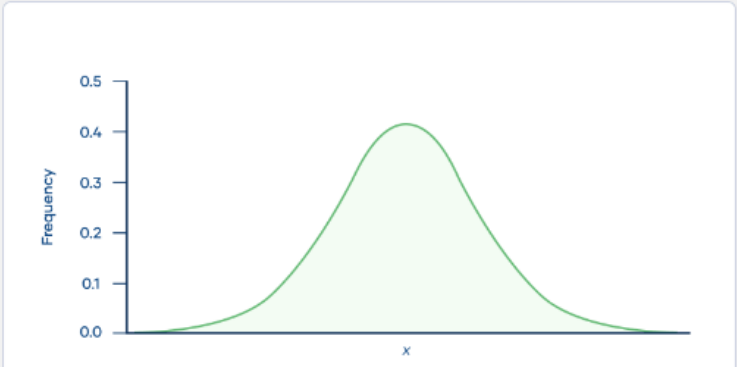
- Around 68% of values are within 1 standard deviation of the mean.
- Around 95% of values are within 2 standard deviations of the mean.
- Around 99.7% of values are within 3 standard deviations of the mean.

The empirical rule is a quick way to get an overview of your data and check for any outliers or extreme values that don't follow this pattern.

What is a normal distribution?

In a normal distribution, data are symmetrically distributed with no skew. Most values cluster around a central region, with values tapering off as they go further away from the center.

The measures of central tendency (mean, mode, and median) are exactly the same in a normal distribution.



What does standard deviation tell you?

The standard deviation is the average amount of variability in your data set. It tells you, on average, how far each score lies from the mean.

In normal distributions, a high standard deviation means that values are generally far from the mean, while a low standard deviation indicates that values are clustered close to the mean.

Can the range be a negative number?

No. Because the range formula subtracts the lowest number from the highest number, the range is always zero or a positive number.

What is the range in statistics?

In statistics, the range is the spread of your data from the lowest to the highest value in the distribution. It is the simplest measure of variability.

What's the difference between central tendency and variability?

While central tendency tells you where most of your data points lie, variability summarizes how far apart your points from each other.

Data sets can have the same central tendency but different levels of variability or vice versa. Together, they give you a complete picture of your data.

What are the 4 main measures of variability?

Variability is most commonly measured with the following descriptive statistics:

- **Range:** the difference between the highest and lowest values
- **Interquartile range:** the range of the middle half of a distribution
- **Standard deviation:** average distance from the mean
- **Variance:** average of squared distances from the mean

What is variability?

Variability tells you how far apart points lie from each other and from the center of a distribution or a data set.

Variability is also referred to as spread, scatter or dispersion.

What is the difference between interval and ratio data?

While interval and ratio data can both be categorized, ranked, and have equal spacing between adjacent values, only ratio scales have a true zero.

For example, temperature in Celsius or Fahrenheit is at an interval scale because zero is not the lowest possible temperature. In the Kelvin scale, a ratio scale, zero represents a total lack of thermal energy.

What is ordinal data?

Ordinal data has two characteristics:

- The data can be classified into different categories within a variable.
- The categories have a natural ranked order.

However, unlike with interval data, the distances between the categories are uneven or unknown.

What's the difference between nominal and ordinal data?

Nominal and ordinal are two of the four levels of measurement. Nominal level data can only be classified, while ordinal level data can be classified and ordered.

What is nominal data?

Nominal data is data that can be labelled or classified into mutually exclusive categories within a variable. These categories cannot be ordered in a meaningful way.

For example, for the nominal variable of preferred mode of transportation, you may have the categories of car, bus, train, tram or bicycle.

What is a standard normal distribution?

The standard normal distribution, also called the *z*-distribution, is a special normal distribution where the mean is 0 and the standard deviation is 1.

Any normal distribution can be converted into the standard normal distribution by turning the individual values into *z*-scores. In a *z*-distribution, *z*-scores tell you how many standard deviations away from the mean each value lies.

What's the best measure of central tendency to use?

The mean is the most frequently used measure of central tendency because it uses all values in the data set to give you an average.

For data from skewed distributions, the median is better than the mean because it isn't influenced by extremely large values.

The mode is the only measure you can use for nominal or categorical data that can't be ordered.

Which measures of central tendency can I use?

The **measures of central tendency** you can use depends on the level of measurement of your data.

- For a **nominal** level, you can only use the **mode** to find the most frequent value.
- For an **ordinal** level or ranked data, you can also use the **median** to find the value in the middle of your data set.
- For **interval** or **ratio** levels, in addition to the mode and median, you can use the **mean** to find the average value.

What are measures of central tendency?

Measures of central tendency help you find the middle, or the average, of a data set.

The 3 most common measures of central tendency are the mean, median and mode.

- The **mode** is the most frequent value.
- The **median** is the middle number in an ordered data set.
- The **mean** is the sum of all values divided by the total number of values.

Why do levels of measurement matter?

The level at which you measure a variable determines how you can analyze your data.

Depending on the level of measurement, you can perform different descriptive statistics to get an overall summary of your data and inferential statistics to see if your results support or refute your hypothesis.

What is standard error?

The standard error of the mean, or simply standard error, indicates how different the population mean is likely to be from a sample mean. It tells you how much the sample mean would vary if you were to repeat a study using new samples from within a single population.

What's the difference between standard error and standard deviation?

Standard error and standard deviation are both measures of variability. The standard deviation reflects variability within a sample, while the standard error estimates the variability across samples of a population.

What's the difference between a point estimate and an interval estimate?

Using descriptive and inferential statistics, you can make two types of estimates about the population: point estimates and interval estimates.

- A **point estimate** is a single value estimate of a parameter. For instance, a sample mean is a point estimate of a population mean.
- An **interval estimate** gives you a range of values where the parameter is expected to lie. A confidence interval is the most common type of interval estimate.

Both types of estimates are important for gathering a clear idea of where a parameter is likely to lie.

How do you calculate a confidence interval?

To calculate the confidence interval, you need to know:

- The point estimate you are constructing the confidence interval for
- The critical values for the test statistic
- The standard deviation of the sample
- The sample size

Then you can plug these components into the confidence interval formula that corresponds to your data. The formula depends on the type of estimate (e.g. a mean or a proportion) and on the distribution of your data.

What is the difference between a confidence interval and a confidence level?

The **confidence level** is the percentage of times you expect to get close to the same estimate if you run your experiment again or resample the population in the same way.

The **confidence interval** consists of the upper and lower bounds of the estimate you expect to find at a given level of confidence.

For example, if you are estimating a 95% confidence interval around the mean proportion of female babies born every year based on a random sample of babies, you might find an upper bound of 0.56 and a lower bound of 0.48. These are the upper and lower bounds of the confidence interval. The confidence level is 95%.

What is a critical value?

A critical value is the value of the test statistic which defines the upper and lower bounds of a confidence interval, or which defines the threshold of statistical significance in a statistical test. It describes how far from the mean of the distribution you have to go to cover a certain amount of the total variation in the data (i.e. 90%, 95%, 99%).

If you are constructing a 95% confidence interval and are using a threshold of statistical significance of $p = 0.05$, then your critical value will be identical in both cases.

What are independent and dependent variables?

You can think of independent and dependent variables in terms of cause and effect: an independent variable is the variable you think is the *cause*, while a dependent variable is the *effect*.

In an experiment, you manipulate the independent variable and measure the outcome in the dependent variable. For example, in an experiment about the effect of nutrients on crop growth:

- The **independent variable** is the amount of nutrients added to the crop field.
- The **dependent variable** is the biomass of the crops at harvest time.

Defining your variables, and deciding how you will manipulate and measure them, is an important part of experimental design.

What's the definition of an independent variable?

An independent variable is the variable you manipulate, control, or vary in an experimental study to explore its effects. It's called "independent" because it's not influenced by any other variables in the study.

Independent variables are also called:

- Explanatory variables (they explain an event or outcome)
- Predictor variables (they can be used to predict the value of a dependent variable)
- Right-hand-side variables (they appear on the right-hand side of a regression equation).

What's the definition of a dependent variable?

A dependent variable is what changes as a result of the independent variable manipulation in experiments. It's what you're interested in measuring, and it "depends" on your independent variable.

In statistics, dependent variables are also called:

- Response variables (they respond to a change in another variable)
- Outcome variables (they represent the outcome you want to measure)
- Left-hand-side variables (they appear on the left-hand side of a regression equation)

How do you make quantitative observations?

To make **quantitative observations**, you need to use instruments that are capable of measuring the quantity you want to observe. For example, you might use a ruler to measure the length of an object or a thermometer to measure its temperature.

What is a frequency distribution?

The **frequency** of a value is the number of times it occurs in a dataset. A **frequency distribution** is the pattern of frequencies of a variable. It's the number of times each possible value of a variable occurs in a dataset.

Types of frequency distributions

There are four types of frequency distributions:

- **Ungrouped frequency distributions:** The number of observations of each **value** of a variable.
  - You can use this type of frequency distribution for categorical variables.
- **Grouped frequency distributions:** The number of observations of each **class interval** of a variable. Class intervals are ordered groupings of a variable's values.
  - You can use this type of frequency distribution for quantitative variables.

- **Relative frequency distributions:** The proportion of observations of each value or class interval of a variable.
    - You can use this type of frequency distribution for **any type of variable** when you're more interested in **comparing frequencies** than the actual number of observations.
- **Cumulative frequency distributions:** The sum of the frequencies less than or equal to each value or class interval of a variable.
    - You can use this type of frequency distribution for **ordinal or quantitative variables** when you want to understand **how often observations fall below certain values**.

A **population** is the entire group that you want to draw conclusions about.

A **sample** is the specific group that you will collect data from. The size of the sample is always less than the total size of the population.

In research, a population doesn't always refer to people. It can mean a group containing elements of anything you want to study, such as objects, events, organizations, countries, species, organisms, etc.

What is Python

Python is a general-purpose, dynamically typed, high-level, compiled and interpreted, garbage-collected, and purely object-oriented programming language that supports procedural, object-oriented, and functional programming.

History of Python

Python was created by Guido van Rossum. In the late 1980s, Guido van Rossum, a Dutch programmer, began working on Python while at the Centrum Wiskunde & Informatica (CWI) in the Netherlands. He wanted to create a successor to the **ABC programming language** that would be easy to read and efficient.

In February 1991, the first public version of Python, version 0.9.0, was released. This marked the official birth of **Python as an open-source project**. The language was named after the British comedy series "**Monty Python's Flying Circus**".

Features of Python:

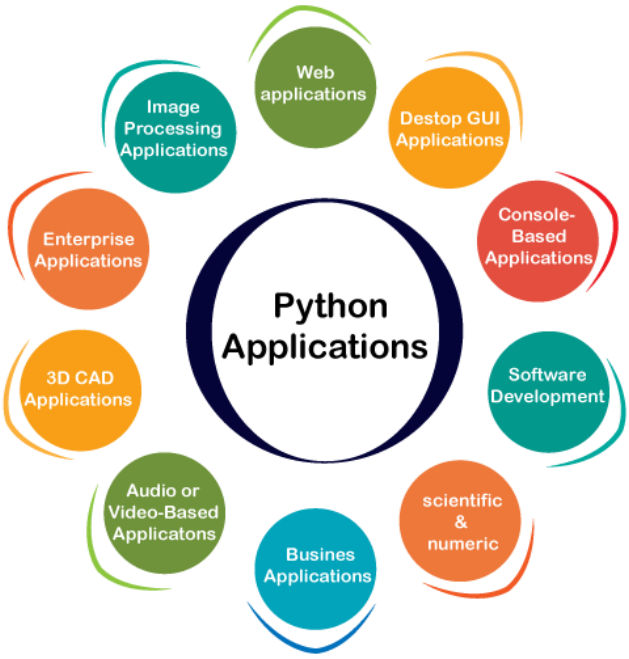
- **Easy to use and Read -** Python's syntax is clear and easy to read, making it an ideal language for both beginners and experienced programmers. This simplicity can lead to faster development and reduce the chances of errors.
- **Dynamically Typed** - The data types of variables are determined during run-time. We do not need to specify the data type of a variable during writing codes.
- **High-level** - High-level language means human readable code.
- **Compiled and Interpreted** - Python code first gets compiled into bytecode, and then interpreted line by line. When we download the Python in our system form org we download the default implement of Python known as CPython. CPython is considered to be Complied and Interpreted both.
- **Garbage Collected** - Memory allocation and de-allocation are automatically managed. Programmers do not specifically need to manage the memory.
- **Purely Object-Oriented** - It refers to everything as an object, including numbers and strings.
- **Cross-platform Compatibility** - Python can be easily installed on Windows, macOS, and various Linux distributions, allowing developers to create software that runs across different operating systems.
- **Rich Standard Library** - Python comes with several standard libraries that provide ready-to-use modules and functions for various tasks, ranging from **web development** and **data manipulation** to **machine learning** and **networking**.
- **Open Source** - Python is an open-source, cost-free programming language. It is utilized in several sectors and disciplines as a result.

Python Applications

← Prev Next →

Python is known for its general-purpose nature that makes it applicable in almost every domain of software development. Python makes its presence in every emerging field. It is the fastest-growing programming language and can develop any application.

Here, we are specifying application areas where Python can be applied.



1) Web Applications

We can use Python to develop web applications. It provides libraries to handle internet protocols such as HTML and XML, JSON, Email processing, request, beautifulSoup, Feedparser, etc. One of Python web-framework named Django is used on **Instagram**. Python provides many useful frameworks, and these are given below:

2) Desktop GUI Applications

The GUI stands for the Graphical User Interface, which provides a smooth interaction to any application. Python provides a **Tk GUI library** to develop a user interface. Some popular GUI libraries are given below.

- Tkinter or Tk
- wxWidgetM
- Kivy (used for writing multitouch applications)
- PyQt or Pyside

4) Software Development

Python is useful for the software development process. It works as a support language and can be used to build control and management, testing, etc.

- **SCons** is used to build control.
- **Buildbot** and **Apache** Gumps are used for automated continuous compilation and testing.
- **Round** or **Trac** for bug tracking and project management.

5) Scientific and Numeric

This is the era of Artificial intelligence where the machine can perform the task the same as the human. Python language is the most suitable language for Artificial intelligence or machine learning. It consists of many scientific and mathematical libraries, which makes easy to solve complex calculations.

Implementing machine learning algorithms require complex mathematical calculation. Python has many libraries for scientific and numeric such as Numpy, Pandas, Scipy, Scikit-learn, etc. If you have some basic knowledge of Python, you need to import libraries on the top of the code. Few popular frameworks of machine libraries are given below.

- SciPy
- Scikit-learn
- NumPy
- Pandas
- Matplotlib

6) Business Applications

Business Applications differ from standard applications. E-commerce and ERP are an example of a business application. This kind of application requires extensively, scalability and readability, and Python provides all these features.

7) Audio or Video-based Applications

Python is flexible to perform multiple tasks and can be used to create multimedia applications. Some multimedia applications which are made by using Python are **TimPlayer, cplay,** etc. The few multimedia libraries are given below.

- Gstreamer
- Pyglet
- QT Phonon

9) Enterprise Applications

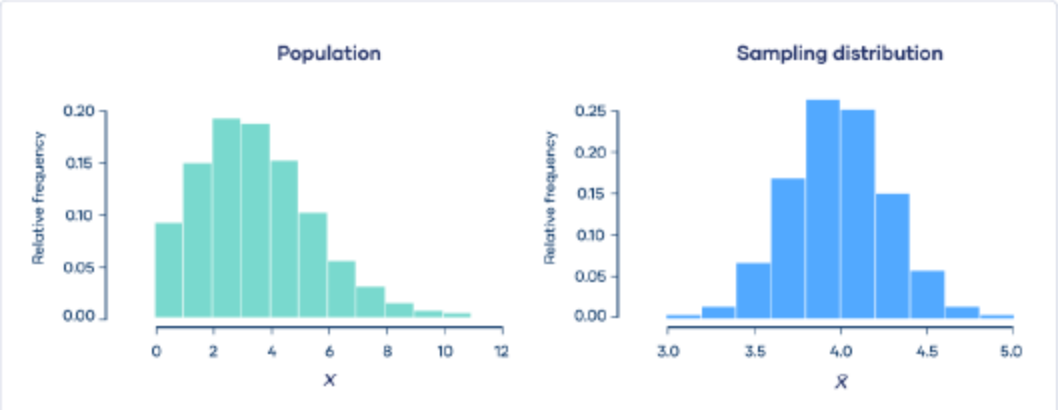
Python can be used to create applications that can be used within an Enterprise or an Organization. Some real-time applications are OpenERP, Tryton, Picalo, etc.

Central Limit Theorem

The **central limit theorem** states that if you take sufficiently large samples from a population, the samples' means will be normally distributed, even if the population isn't normally distributed.

Example: Central limit theorem

A **population** follows a **Poisson distribution** (left image). If we take 10,000 **samples** from the population, each with a sample size of 50, the sample means follow a normal distribution, as predicted by the **central limit theorem** (right image).



1. **What is standard error?**

The standard error of the mean, or simply standard error, indicates how different the population mean is likely to be from a sample mean. It tells you how much the sample mean would vary if you were to repeat a study using new samples from within a single population.

2. **What's the difference between standard error and standard deviation?**

Standard error and standard deviation are both measures of variability. The standard deviation reflects variability within a sample, while the standard error estimates the variability across samples of a population.

3. aa