Unit-II
Statistical Concepts

Population and Sample

- 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.

Population vs sample

Population	Sample
Advertisements for IT jobs in the Netherlands	The top 50 search results for advertisements for IT jobs in the Netherlands on May 1, 2020
Songs from the Eurovision Song Contest	Winning songs from the Eurovision Song Contest that were performed in English
Undergraduate students in the Netherlands	300 undergraduate students from three Dutch universities who volunteer for your psychology research study
All countries of the world	Countries with published data available on birth rates and GDP since 2000

Collecting data from a population

KIIT administrator wants to analyze the final exam scores of all graduating seniors to see if there is a trend. Since they are only interested in applying their findings to the graduating seniors at KIIT university, they use the whole population dataset.

Collecting data from a sample

KIIT want to study political attitudes in young people. KIIT population is arround the 30000 undergraduate students. Because it's not practical to collect data from all of them, you use a sample of 300 undergraduate volunteers from different dept who meet your inclusion criteria. This is the group who will complete your online survey.

Types of data: Quantitative vs categorical variables

Data is a specific measurement of a variable – it is the value you record in your data sheet. Data is generally divided into two categories:

Quantitative data **represents amounts**

Categorical data **represents groupings**

A variable that contains quantitative data is a quantitative variable; a variable that contains categorical data is a categorical variable. Each of these types of variables can be broken down into further types.

Quantitative variables

When you collect quantitative data, the numbers you record represent real amounts that can be added, subtracted, divided, etc. There are two types of quantitative variables: discrete and continuous.

Quantitative variables

When you collect quantitative data, the numbers you record represent real amounts that can be added, subtracted, divided, etc. There are two types of quantitative variables: **discrete and continuous.**

Discrete vs continuous variables

Type of variable	What does the data represent?	Examples
Discrete variables (aka integer variables)	Counts of individual items or values.	• Number of students in a class • Number of different tree species in a forest
Continuous variables (aka ratio variables)	Measurements of continuous or non-finite values.	• Distance • Volume • Age

Categorical variables

Categorical variables represent groupings of some kind. They are sometimes recorded as numbers, but the numbers represent categories rather than actual amounts of things.

There are three types of categorical variables: binary, nominal, and ordinal variables.

Binary vs nominal vs ordinal variables

Type of variable	What does the data represent?	Examples
Binary variables (aka dichotomous variables)	Yes or no outcomes.	• Heads/tails in a coin flip • Win/lose in a football game
Nominal variables	Groups with no rank or order between them.	• Species names • Colors • Brands
Ordinal variables	Groups that are ranked in a specific order.	• Finishing place in a race • Rating scale responses in a survey, such as Likert scales*

This example sheet is color-coded according to the type of variable: nominal, continuous, ordinal, and binary.

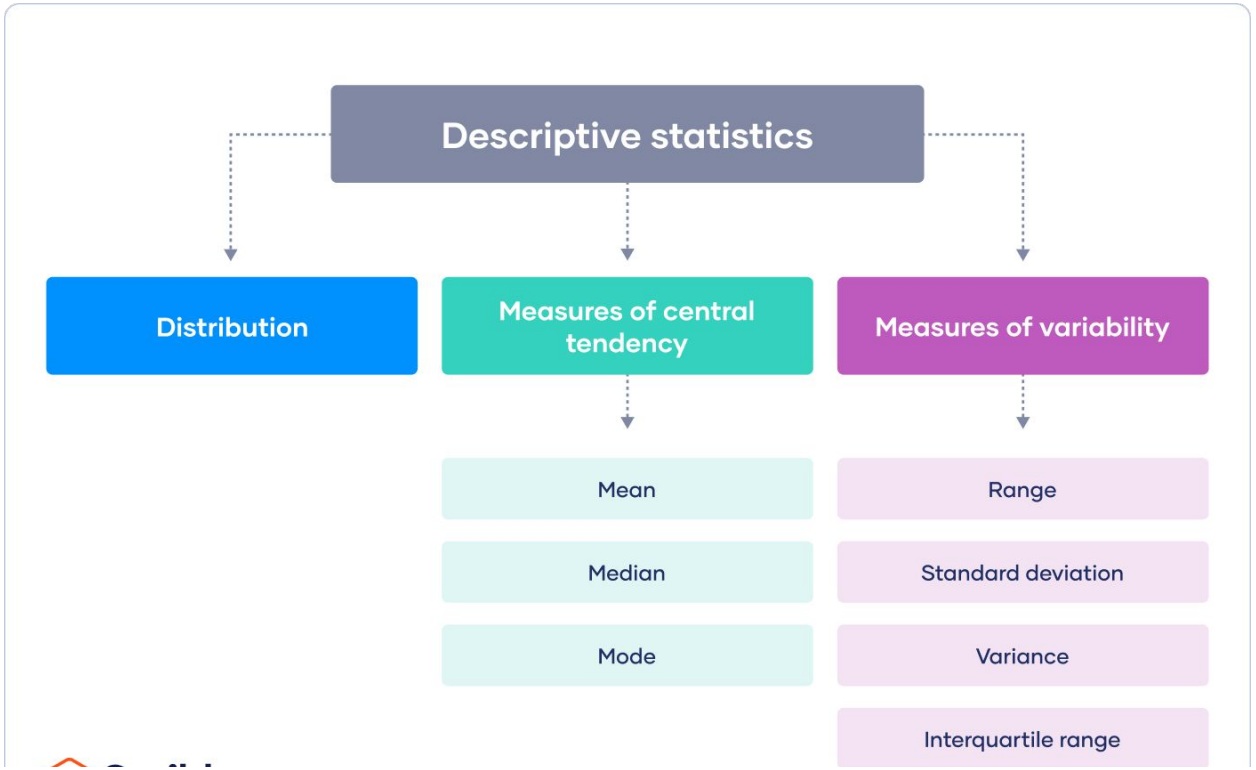
Sample #	Plant Species	Salt Added (mg/L water)	Starting Height (cm)	Growth (cm) (current height – starting height)	Wilting (rank 0-10)	Survival (1=survived, 0=died)
1	A	0	12			
2	A	100	13			
3	A	250	11			
4	B	0	25			
5	B	100	26			
6	B	250	25			

Descriptive Statistics

Descriptive statistics summarize and organize characteristics of a data set. A data set is a collection of responses or observations from a sample or entire population.

In quantitative research, after collecting data, the first step of **statistical analysis** is to describe characteristics of the responses, such as the average of one variable (e.g., age), or the relation between two variables (e.g., age and creativity).

Types of descriptive statistics



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

Frequency distribution

A data set is made up of a distribution of values, or scores. In tables or graphs, you can summarize the frequency of every possible value of a variable in numbers or percentages. This is called a frequency distribution.

Gender	Number
Male	182
Female	235
Other	27

Measures of central tendency

Measures of central tendency estimate the center, or average, of a data set. The mean, median and mode are 3 ways of finding the average.

Mean number of library visits

Data set	15, 3, 12, 0, 24, 3
Sum of all values	15 + 3 + 12 + 0 + 24 + 3 = 57
Total number of responses	*N* = 6
Mean	Divide the sum of values by *N* to find *M*: 57/6 = **9.5**

Mode number of library visits

Ordered data set	0, 3, 3, 12, 15, 24
Mode	Find the most frequently occurring response: **3**

Median number of library visits

Ordered data set	0, 3, 3, 12, 15, 24
Middle numbers	3, 12
Median	Find the mean of the two middle numbers: (3 + 12)/2 = **7.5**

Measures of variability

Measures of variability give you a sense of how spread out the response values are. The range, standard deviation and variance each reflect different aspects of spread.

Range

The range gives you an idea of how far apart the most extreme response scores are. To find the range, simply subtract the lowest value from the highest value.

Range of visits to the library in the past year

Ordered data set: 0, 3, 3, 12, 15, 24

Range: 24 − 0 = **24**

Standard deviation

The standard deviation (s or SD) is the average amount of variability in your dataset. It tells you, on average, how far each score lies from the mean. The larger the standard deviation, the more variable the data set is.

There are six steps for finding the standard deviation:

1. **List each score and find their mean.**
2. **Subtract the mean from each score to get the deviation from the mean.**
3. **Square each of these deviations.**
4. **Add up all of the squared deviations.**
5. **Divide the sum of the squared deviations by N – 1.**
6. **Find the square root of the number you found.**

Raw data	Deviation from mean	Squared deviation
15	15 – 9.5 = 5.5	30.25
3	3 – 9.5 = -6.5	42.25
12	12 – 9.5 = 2.5	6.25
0	0 – 9.5 = -9.5	90.25
24	24 – 9.5 = 14.5	210.25
3	3 – 9.5 = -6.5	42.25
M = 9.5	Sum = 0	Sum of squares = 421.5

Step 5: 421.5/5 = 84.3

Step 6: √84.3 = 9.18

From learning that ***s* = 9.18**, you can say that on average, each score deviates from the mean by 9.18 points.

Variance

The variance is the average of squared deviations from the mean. Variance reflects the degree of spread in the data set. The more spread the data, the larger the variance is in relation to the mean.

To find the variance, simply square the standard deviation. The symbol for variance is s2.

Variance of visits to the library in the past year

Data set: 15, 3, 12, 0, 24, 3

$s = 9.18$

$s^2 = \mathbf{84.3}$

Univariate descriptive statistics

Univariate descriptive statistics focus on only one variable at a time. It's important to examine data from each variable separately using multiple measures of distribution, central tendency and spread. Programs like SPSS and Excel,python can be used to easily calculate these.

	Visits to the library
N	6
Mean	9.5
Median	7.5
Mode	3
Standard deviation	9.18
Variance	84.3
Range	24

Bivariate descriptive statistics

If you've collected data on more than one variable, you can use **bivariate or multivariate descriptive statistics** to explore whether there are relationships between them.

In bivariate analysis, you simultaneously study the frequency and variability of two variables to see if they vary together. You can also compare the central tendency of the two variables before performing further statistical tests.

Scatter plots

A scatter plot is a chart that shows you the relationship between two or three variables. It's a visual representation of the strength of a relationship.

In a scatter plot, you plot one variable along the x-axis and another one along the y-axis. Each data point is represented by a point in the chart.

Frequency Distribution

A frequency distribution describes the number of observations for each possible value of a variable. Frequency distributions are depicted using graphs and frequency tables.

Medal	Frequency
Gold	8
Silver	10
Bronze	7

Quartiles & Quantiles

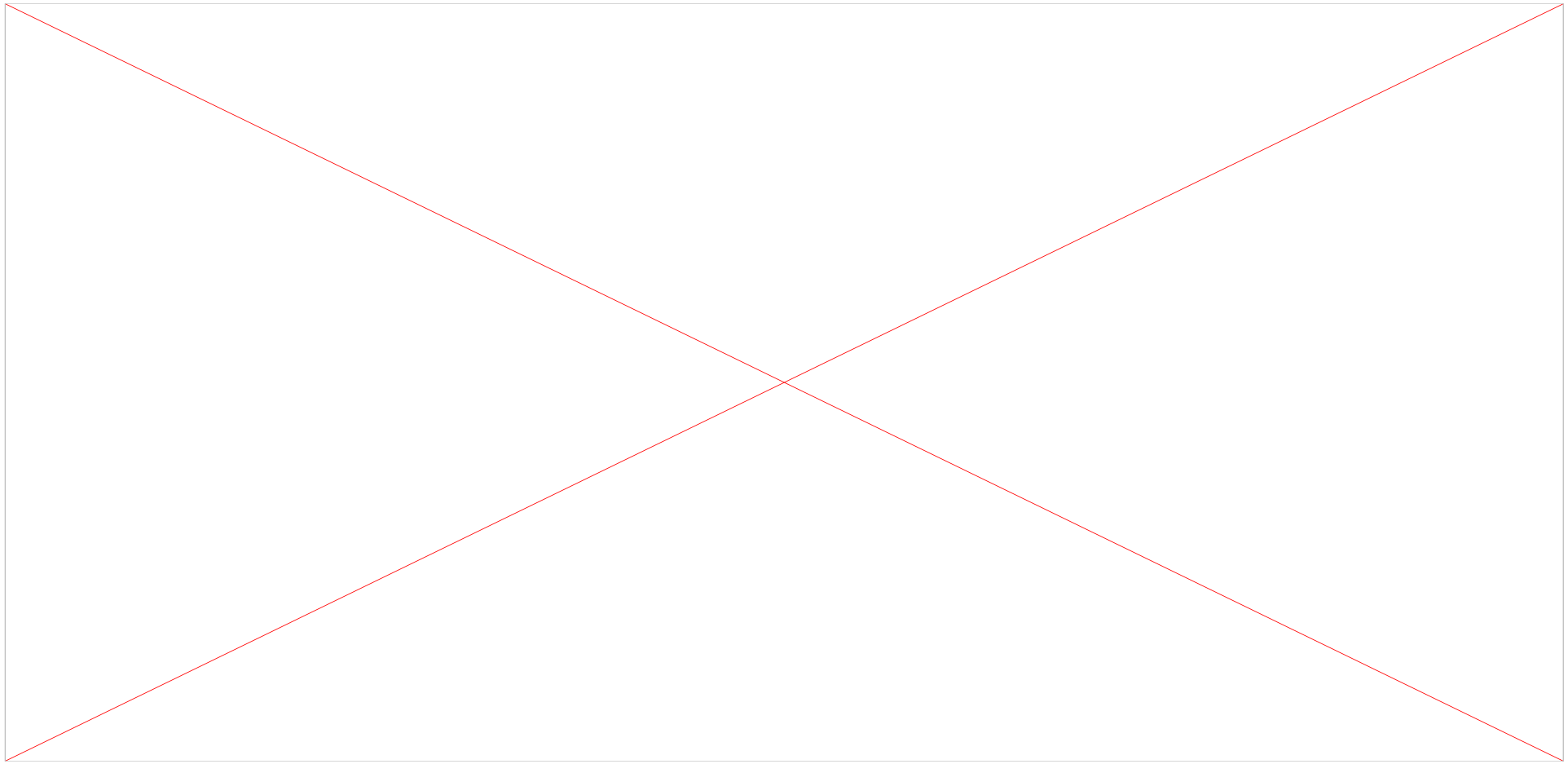
Quartiles are three values that split sorted data into four parts, each with an equal number of observations. Quartiles are a type of quantile.

First quartile: Also known as Q1, or the lower quartile. This is the number halfway between the lowest number and the middle number.

Second quartile: Also known as Q2, or the median. This is the middle number halfway between the lowest number and the highest number.

Third quartile: Also known as Q3, or the upper quartile. This is the number halfway between the middle number and the highest number.

Quartiles



Probability Distribution

Probability Distribution

A probability distribution is a mathematical function that describes the probability of different possible values of a variable.

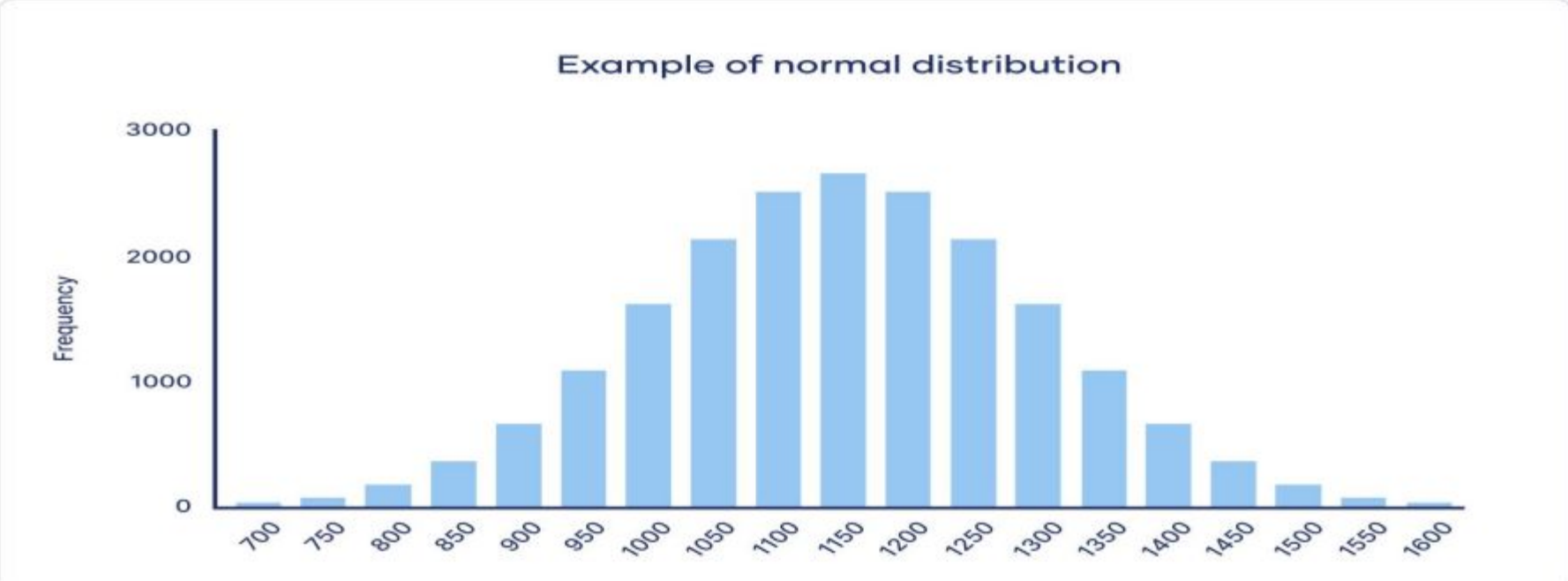
Probability distributions are often depicted using graphs or probability tables.

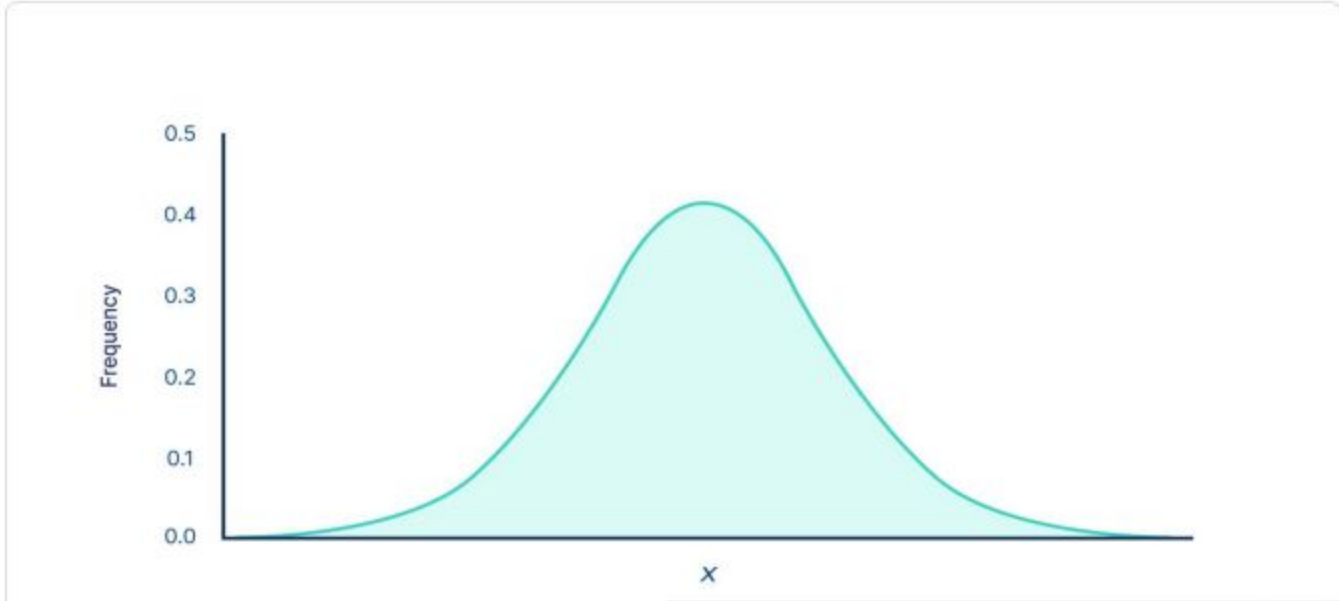
Outcome	Probability
Heads	Tails
.5	.5

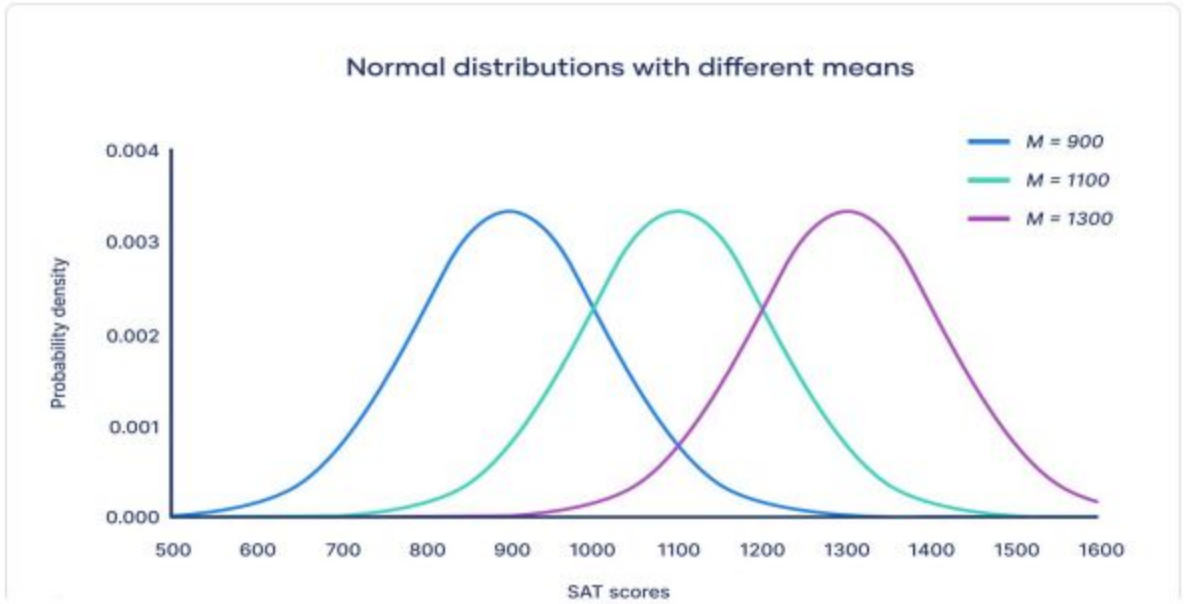
Normal Distribution

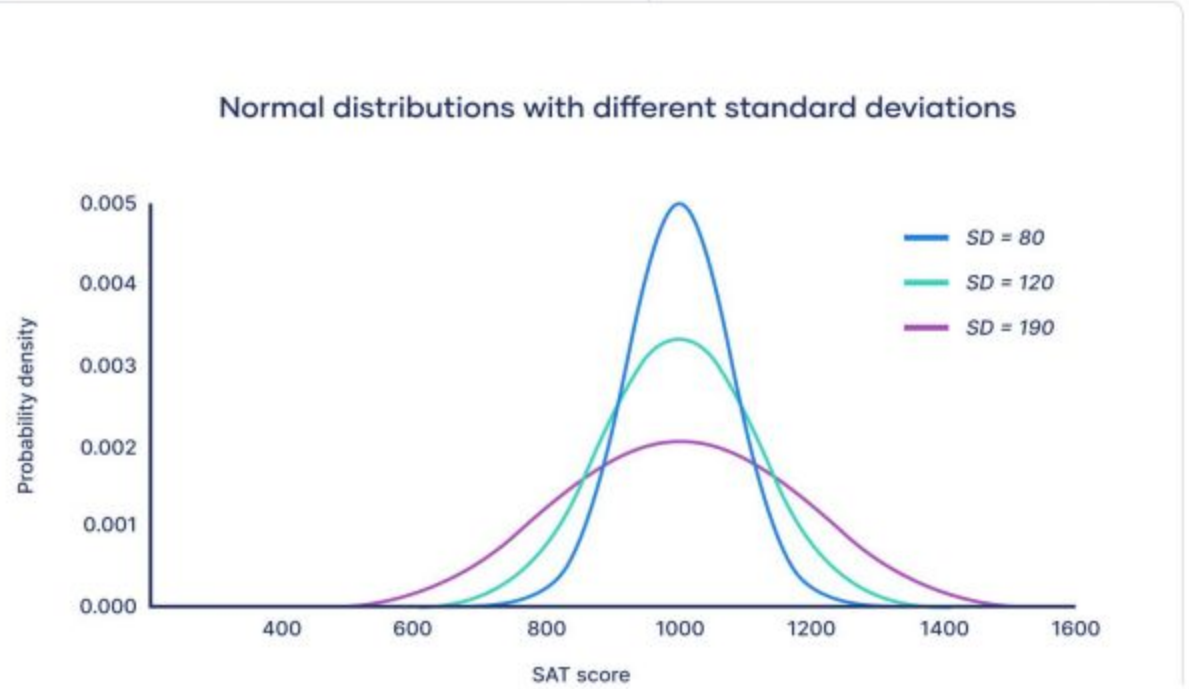
In a **normal distribution**, data is **symmetrically** distributed with **no skew**. When plotted on a graph, the data follows a bell shape, with **most values clustering around a central region** and tapering off as they go further away from the center.

Normal distributions are also called Gaussian distributions or bell curves because of their shape




0.5
0.4
0.3
0.2
0.1
0.0
Frequency
x

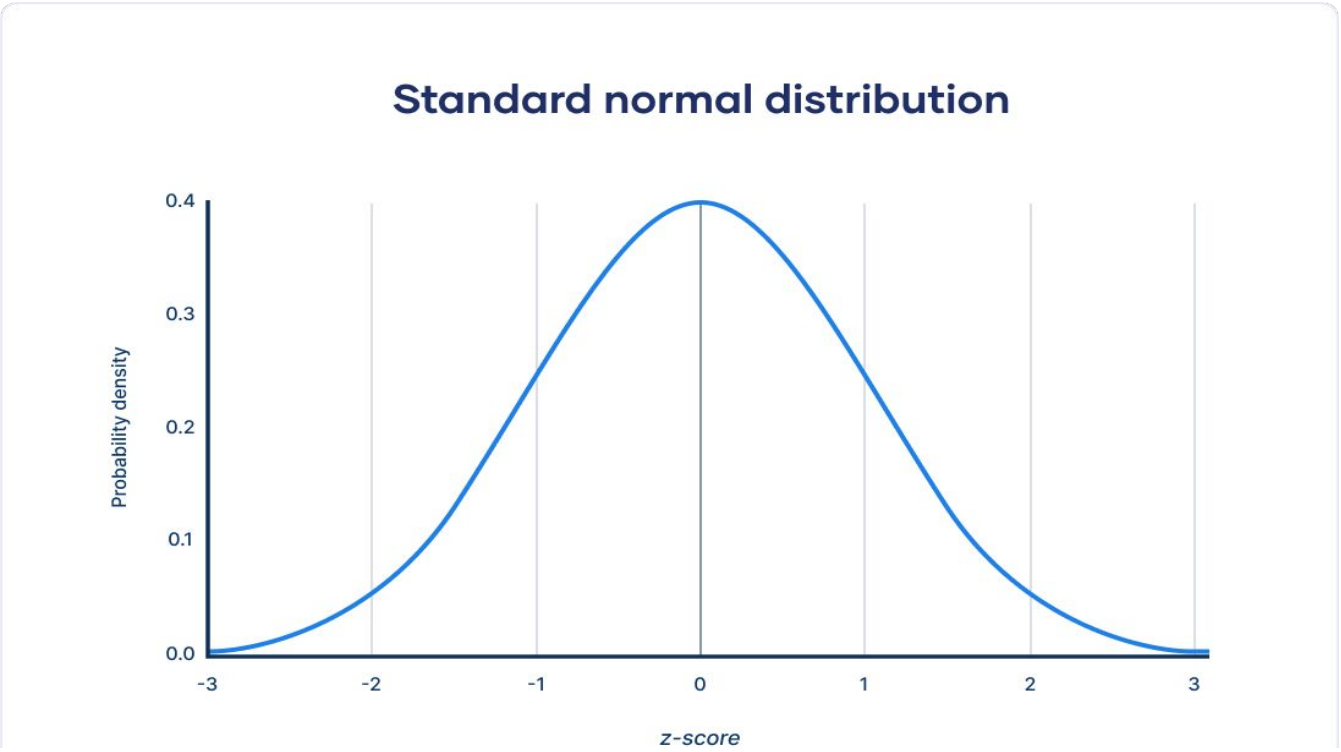

Normal distributions with different means
M = 900
M = 1100
M = 1300
0.004
0.003
0.002
0.001
0.000
Probability density
500
600
700
800
900
1000
1100
1200
1300
1400
1500
1600
SAT scores


Normal distributions with different standard deviations
SD = 80
SD = 120
SD = 190
0.005
0.004
0.003
0.002
0.001
0.000
Probability density
400
600
800
1000
1200
1400
1600
SAT score

The 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.**

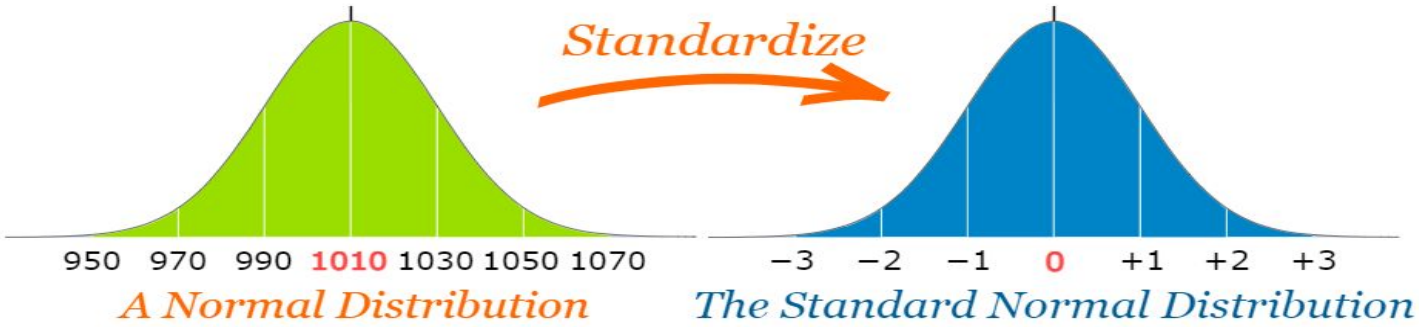
Any normal distribution can be standardized by converting its values into z scores. **Z scores** tells how many **standard deviations from the mean each value lies.**



Converting a **normal distribution** into a z-distribution allows you to **calculate the probability of certain values occurring** and **to compare different data sets.**

Normal distribution vs the Standard normal distribution

- All **normal distributions**, like the standard normal distribution, are unimodal and symmetrically distributed with a bell-shaped curve. However, a **normal distribution** can take on **any value** as its **mean** and standard deviation.
  - In the **standard normal distribution**, the mean and standard deviation are always fixed.



Standardizing a normal distribution

When you standardize a normal distribution, the **mean becomes 0** and the **standard deviation becomes 1.** This allows you to easily calculate the probability of certain values occurring in your distribution, or to compare data sets with different means and standard deviations.

While **data points** are referred to as x in a normal distribution, they **are called z or z scores** in the **z distribution**. A z score is a standard score that tells you how many standard deviations away from the mean an individual value (x) lies:

A **positive z score** means that your x value is greater than the mean.

A **negative z score** means that your x value is less than the mean.

A **z score of zero** means that your x value is equal to the mean.

How to calculate a z score

To standardize a value from a normal distribution, convert the individual value into a z-score:

- Subtract the mean from your individual value.
- Divide the difference by the standard deviation.

Z-score formula	Explanation
$z = \frac{x - \mu}{\sigma}$	• x = individual value • μ = mean • σ = standard deviation

Example: Finding a z score

The SAT scores from students in a new test preparation course are collected. The data follows a normal distribution with a mean score (M) of 1150 and a standard deviation (SD) of 150. **You want to find the probability that SAT scores in your sample exceed 1380.**

To standardize your data, you first find the z score for 1380. The z score tells you how many standard deviations away 1380 is from the mean.

Step 1: Subtract the mean from the *x* value.	$x = 1380$ $M = 1150$ $x - M = 1380 - 1150 = 230$
Step 2: Divide the difference by the standard deviation.	$SD = 150$ $z = 230 \div 150 = 1.53$

The *z* score for a value of 1380 is **1.53**. That means 1380 is 1.53 standard deviations from the mean of your distribution.

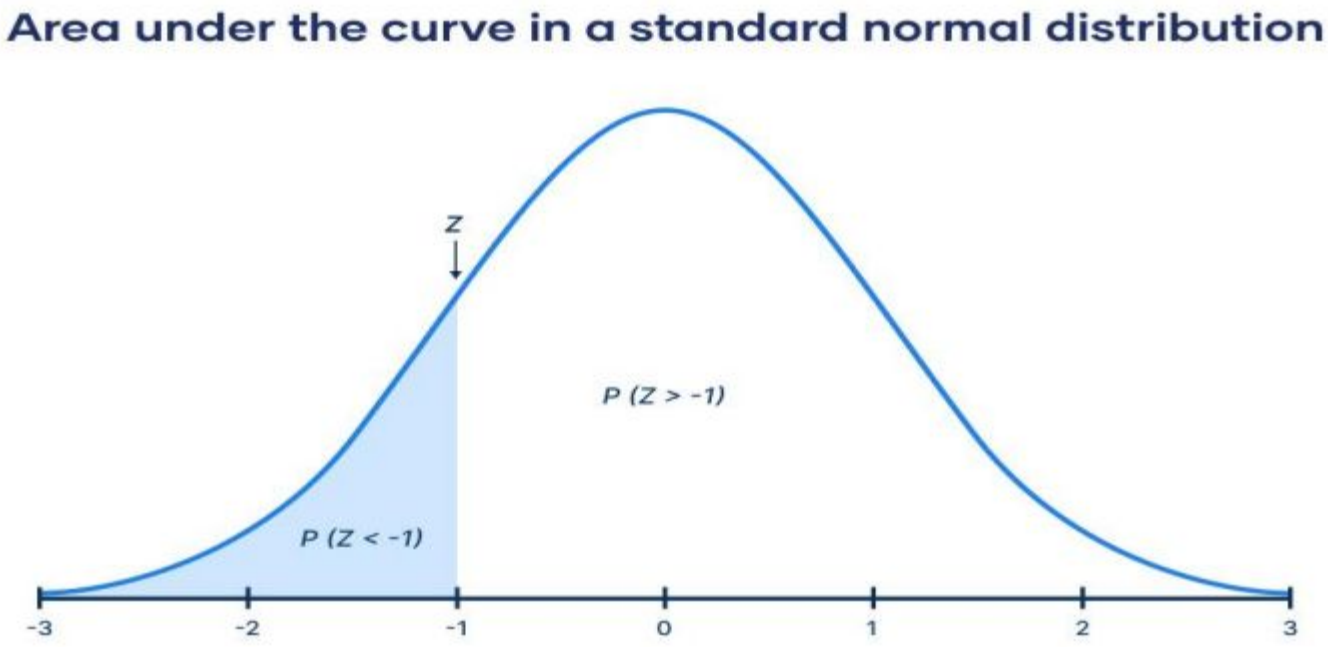
Next, we can find the probability of this score using a *z* table.

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
+0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
+0.1	.53983	.54380	.54776	.55172	.55567	.55966	.56360	.56749	.57142	.57535
+0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
+0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
+0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
+0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
+0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
+0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
+0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
+0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
+1	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
+1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
+1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
+1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91308	.91466	.91621	.91774
+1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
+1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
+1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
+1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
+1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
+1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
+2	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
+2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
+2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
+2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
+2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361
+2.5	.99379	.99396	.99413	.99430	.99446	.99461	.99477	.99492	.99506	.99520
+2.6	.99534	.99547	.99560	.99573	.99585	.99598	.99609	.99621	.99632	.99643
+2.7	.99653	.99664	.99674	.99683	.99693	.99702	.99711	.99720	.99728	.99736
+2.8	.99744	.99752	.99760	.99767	.99774	.99781	.99788	.99795	.99801	.99807
+2.9	.99813	.99819	.99825	.99831	.99836	.99841	.99846	.99851	.99856	.99861
+3	.99865	.99869	.99874	.99878	.99882	.99886	.99889	.99893	.99896	.99900
+3.1	.99903	.99906	.99910	.99913	.99916	.99918	.99921	.99924	.99926	.99929
+3.2	.99931	.99934	.99936	.99938	.99940	.99942	.99944	.99946	.99948	.99950
+3.3	.99952	.99953	.99955	.99957	.99958	.99960	.99961	.99962	.99964	.99965
+3.4	.99966	.99968	.99969	.99970	.99971	.99972	.99973	.99974	.99975	.99976
+3.5	.99977	.99978	.99978	.99979	.99980	.99981	.99981	.99982	.99983	.99983
+3.6	.99984	.99985	.99985	.99986	.99986	.99987	.99987	.99988	.99988	.99989
+3.7	.99989	.99990	.99990	.99990	.99991	.99991	.99992	.99992	.99992	.99992
+3.8	.99993	.99993	.99993	.99994	.99994	.99994	.99994	.99995	.99995	.99995
+3.9	.99995	.99995	.99996	.99996	.99996	.99996	.99996	.99996	.99997	.99997
+4	.99997	.99997	.99997	.99997	.99997	.99997	.99998	.99998	.99998	.99998

Use the standard normal distribution to find probability

The **standard normal distribution** is a **probability distribution**, so the **area under the curve between two points tells you the probability of variables taking on a range of values**. The total area under the curve is 1 or 100%.

Every **z score has an associated p value that tells you the probability of all values below or above that z score occurring.** This is the area under the curve left or right of that z score.



Z tests and p values

- The **z score is the test statistic used in a z test.** The **z test** is used to **compare the means of two groups**, or to **compare the mean of a group to a set value**. Its null hypothesis* typically assumes no difference between groups.

- The **area under the curve** to the right of a **z score** is the **p value**, and it's the likelihood of your observation occurring if the null hypothesis is true.

- Usually, a **p value of 0.05 or less means that your results are unlikely to have arisen by chance; it indicates a statistically significant effect.**

- By converting a value in a **normal distribution into a z score**, **p value for a z test can be obtained.**

*suggests that no statistical relationship and significance exists in a set of given single observed variable, between two sets of observed data and measured phenomena.

Example: As a sleep researcher, you're curious about how sleep habits changed during COVID-19 lockdowns. You collect sleep duration data from a sample during a full lockdown.

Before the lockdown, the population mean was 6.5 hours of sleep and standard deviation is 0.5. The lockdown sample mean is 7.62.

To assess whether your sample mean significantly differs from the pre-lockdown population mean, you perform a z test:

First, you calculate a z score for the sample mean value.

Then, you find the p value for your z score using a z table.

Step 1: Calculate a z-score

To compare sleep duration during and before the lockdown, you convert your lockdown sample mean into a z score using the pre-lockdown population mean and standard deviation.

Formula	Explanation	Calculation
$z = \frac{x - \mu}{\sigma}$	x = sample mean μ = population mean σ = population standard deviation	$x = 7.62$ $\mu = 6.5$ $\sigma = 0.5$ $z = \frac{7.62 - 6.5}{0.5} = 2.24$

A z score of 2.24 means that your sample mean is 2.24 standard deviations greater than the population mean.

Step 2: Find the p value

To find the probability of your sample mean z score of 2.24 or less occurring, you use the z table to find the value at the intersection of row 2.2 and column +0.04.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.9834 1	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.986 10	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361

The table tells you that the **area under the curve up to or below your z score is 0.9874.**
This means that **sample's mean sleep duration is higher than about 98.74% of the population's mean sleep duration pre-lockdown.**
To find the **p value to assess whether the sample differs from the population,** calculate the area under the curve above or to the right of your z score. Since the total area under the curve is 1, by subtracting the area under the curve below your **z score from 1.**

A p value of less than 0.05 or 5% means that the sample significantly differs from the population.

Probability of z > 2.24 = 1 – 0.9874 = 0.0126 or 1.26%

With a *p*-value of less than 0.05, it concludes that average sleep duration in the COVID-19 lockdown was significantly higher than the pre-lockdown average.

Q: 300 college student’s exam scores are tallied at the end of the semester. Eric scored 800 marks in total out of 1000. The average score for the batch was 700 and the standard deviation was 180. Find out how well Eric scored compared to his batch mates.

What is a Poisson distribution?

A Poisson distribution is a **discrete probability distribution**, meaning that it gives the probability of a discrete **(i.e., countable)** outcome. For Poisson distributions, the discrete outcome is the **number of times an event occurs**, represented by k.

You can use a Poisson distribution to **predict or explain the number of events occurring within a given interval of time or space. "Events"** could be anything from disease cases to customer purchases to mete or strikes. The interval can be any specific amount of time or space, such as 10 days or 5 square inches.

Examples of Poisson distributions

- Text messages per hour
- Machine malfunctions per year
- Website visitors per month
- Influenza cases per year

Poisson distribution formula

The probability mass function of the Poisson distribution is:

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

Where:

- X is a random variable following a Poisson distribution
- k is the number of times an event occurs
- $P(X = k)$ is the probability that an event will occur k times
- e is Euler's constant (approximately 2.718)
- λ is the average number of times an event occurs
- ! is the factorial function

Example: Applying the Poisson distribution formula

An average of 0.61 soldiers died by horse kicks per year in each Prussian army corps. You want to **calculate the probability that exactly two soldiers** died in the VII Army Corps in 1898, assuming that the number of horse kick deaths per year follows a Poisson distribution.

Calculation

k = 2 deaths by horse kick

λ = 0.61 deaths by horse kick per year

e = 2.718

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

$$P(X = 2) = \frac{(2.718^{-0.61})(0.61^2)}{2!}$$

$$P(X = 2) = \frac{(0.54339)(0.3721)}{2}$$

$$P(X = 2) = 0.101$$

The probability that exactly two soldiers died in the VII Army Corps in 1898 is 0.101.

What is a chi-square distribution?

Chi-square (X2) distributions are a family of continuous probability distributions. A chi-square (χ^2) statistic is **a measure of the difference between the observed and expected frequencies of the outcomes of a set of events or variables**.

Chi-square test statistics (formula)

Formula	Explanation
$X^2 = \sum \frac{(O-E)^2}{E}$	Where

- X^2 is the chi-square test statistic
- $\sum$ is the summation operator (it means "take the sum of")
- O is the observed frequency
- E is the expected frequency

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 how much the sample mean would vary if you were to repeat a study using new samples from within a single population.

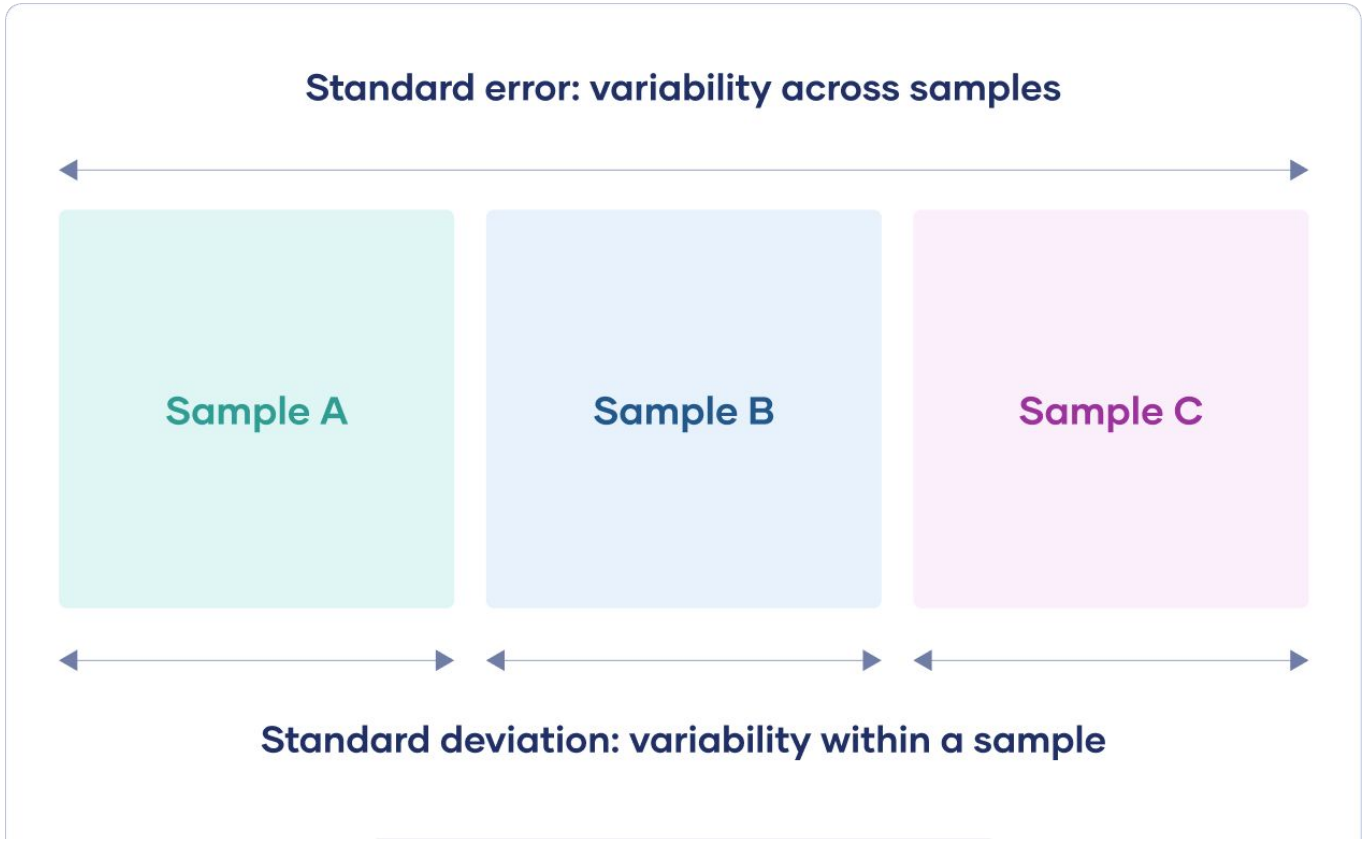
The **standard error of the mean** (SE or SEM) is the most commonly reported type of standard error.

Standard error vs standard deviation

Standard error and standard deviation are both measures of **variability**:

The **standard deviation** describes **variability within** a **single** sample.

The **standard error** estimates the **variability** across **multiple** samples of a population.



Example: Standard error vs standard deviation

In a **random sample** of **200** students, the **mean math SAT score** is **550**. In this case, the sample is the 200 students, while the population is all test takers in the region.
The standard deviation of the math scores is **180.** This number reflects on average how much each score **differs from the sample mean score of 550.**

The **standard error** of **the math scores,** on the other hand, tells you how much the sample **mean score of 550 differs from other sample mean scores**, in samples of equal size, in the population of all test takers in the region.

Standard error formula

When population parameters are known

Formula	Explanation
$SE = \frac{\sigma}{\sqrt{n}}$	• SE is standard error • σ is population standard deviation • n is the number of elements in the sample

When population parameters are unknown

Formula	Explanation
$SE = \frac{s}{\sqrt{n}}$	• SE is standard error • s is sample standard deviation • n is the number of elements in the sample

Example: Using the standard error formula

To estimate the standard error for math SAT scores, you follow two steps.

First, find the square root of your sample size (n).

Formula	Calculation
$\sqrt{n}$	$n = 200$
	$\sqrt{n} = \sqrt{200} = 14.1$

Next, divide the sample standard deviation by the number you found in step one.

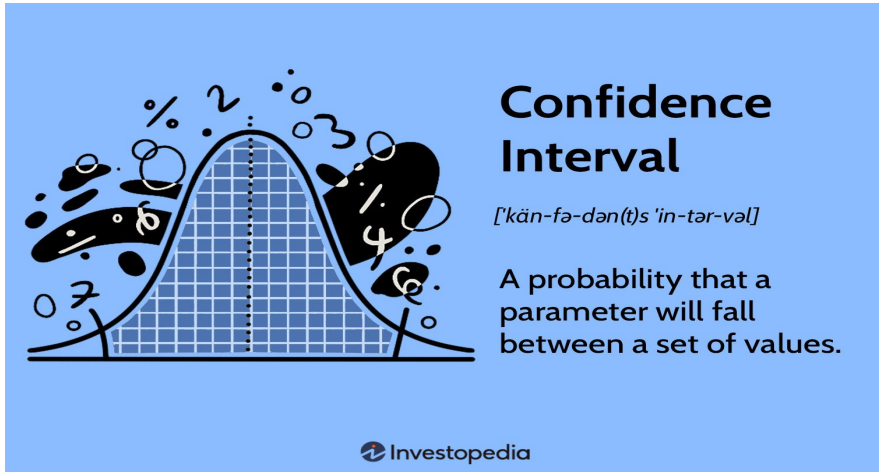
Formula	Calculation
$SE = \frac{s}{\sqrt{n}}$	$s = 180$
	$\sqrt{n} = 14.1$
	$\frac{s}{\sqrt{n}} = \frac{180}{14.1} = 12.8$

Confidence Intervals

A confidence interval, in statistics, refers to the probability that a population parameter will fall **between a set of values** for a **certain proportion of times.**

A confidence interval is the mean of your estimate plus and minus the variation in that estimate.

Confidence, in statistics, is another way to describe probability. For example, if you construct a confidence interval with a 95% confidence level, you are confident that 95 out of 100 times the estimate will fall between the upper and lower values specified by the confidence interval.

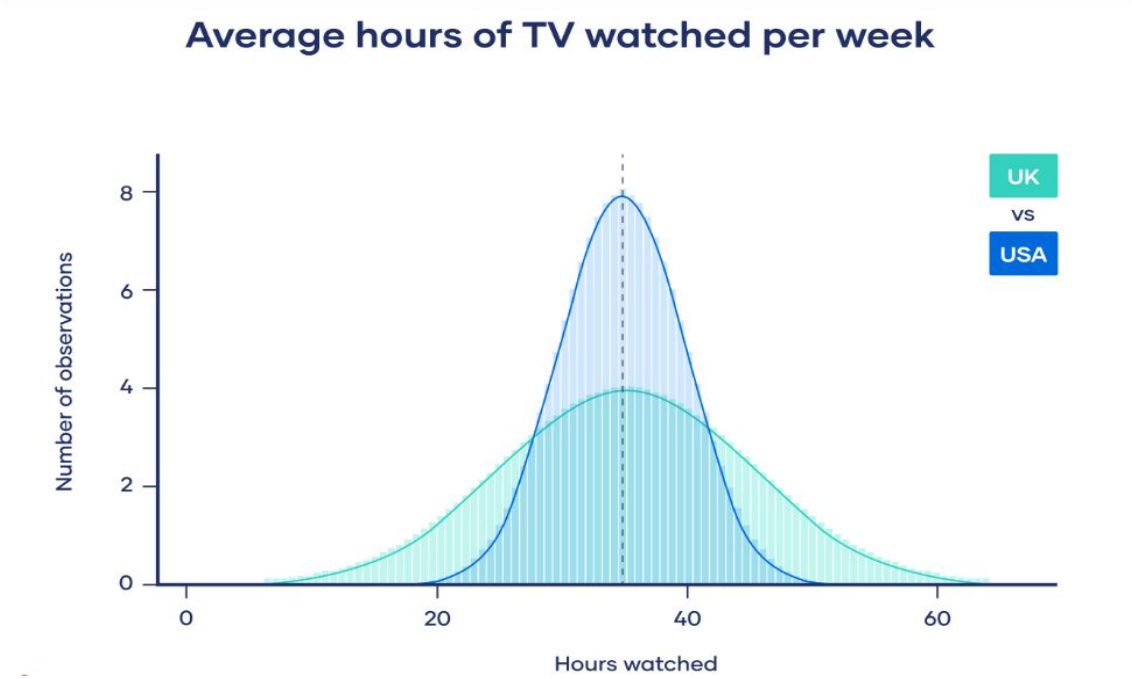


Example: Variation around an estimate

You survey 100 Brits and 100 Americans about their television-watching habits, and find that both groups watch an average of 35 hours of television per week.

However, the British people surveyed had a wide variation in the number of hours watched, while the Americans all watched similar amounts.

Even though both groups have the same point estimate (average number of hours watched), the British estimate will have a wider confidence interval than the American estimate because there is more variation in the data.



Calculating a confidence interval: what 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

Confidence Interval Formulas:

If $n \geq 30$, Confidence Interval = $\bar{x} \pm z_c(\sigma/\sqrt{n})$

If $n<30$, Confidence Interval = $\bar{x} \pm t_c(S/\sqrt{n})$

Where,

n = Number of terms
$\bar{x}$ = Sample Mean
σ = Standard Deviation
z_c = Value corresponding to confidence interval in z table
t_c = Value corresponding to confidence interval in t table

Example 1: A random sample of **30 apples** was taken from a large population. On measuring their diameter the mean diameter of the sample was **91 millimeters** with a **standard deviation of 8 mm**. Calculate the 85% confidence limits for the mean diameter of the whole population of apples.

For 85% confidence, $Z = 1.440$ We have: $\bar{x} = 91$, $s = 8$, $Z = 1.440$

and $n = 30$

Substitute into confidence interval formula:

$$\bar{x} \pm Z \frac{s}{\sqrt{n}}$$

$$\bar{x} \pm z_c(\sigma/\sqrt{n})$$

Therefore, the 85% confidence limits are:

$$= 91 \pm 1.440 \times \frac{8}{\sqrt{30}}$$

$$= 91 \pm 1.440 \times \frac{8}{5.477\ldots}$$

$$= 91 \pm 2.1$$

Answer: The 85% confidence limits are $= 91 \pm 2.1$

Example 2: A random sample of ten scores obtained by the students in a Math test are as follows: 2, 16, 3, 10, 11, 4, 6, 7, 9, 12. What will be the 90% confidence limits for the mean of the whole sample?

First, calculate the sample mean:

$$\bar{x} = \frac{2+3+4+6+7+9+10+11+12+16}{10} = \frac{80}{10} = 8$$

Now the sample standard deviation:

Because this is a sample, we use the formula $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

with $\bar{x} = 8$ and $n - 1 = 9$

Therefore $s =$

$$\sqrt{\frac{1}{9}((2-8)^2 + (3-8)^2 + (4-8)^2 + (6-8)^2 + (7-8)^2 + (9-8)^2}$$

$$= \sqrt{\frac{1}{9}(36 + 25 + 16 + 4 + 1 + 1 + 4 + 9 + 16 + 64)}$$

$$= \sqrt{\frac{1}{9} \times 176}$$

$$= \sqrt{19.555\ldots}$$

$$= 4.4221\ldots$$

For 90% confidence, $Z = 1.645$

Therefore, the 90% confidence limits are $\bar{x} \pm Z\frac{s}{\sqrt{n}}$

$$= 8 \pm 1.645 \times \frac{4.4221\ldots}{\sqrt{10}}$$

$$= 8 \pm 1.645 \times \frac{4.4221\ldots}{3.1622\ldots} = 8 \pm 2.3$$

Answer: The 90% confidence limits are 8 ± 2.3

Hypothesis Testing

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 (Ho) and (Ha or H1).
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.

Step 1: State your null and alternate hypothesis

After developing your initial research hypothesis (the prediction that you want to investigate), it is important to restate it as a null (Ho) and alternate (Ha) hypothesis so that you can test it mathematically.

The alternate hypothesis is usually your initial hypothesis that predicts a relationship between variables. The null hypothesis is a prediction of no relationship between the variables you are interested in.

Hypothesis testing example

You want to test whether there is a relationship between gender and height. Based on your knowledge of human physiology, you formulate a hypothesis that men are, on average, taller than women. To test this hypothesis, you restate it as:

H0: Men are, on average, not taller than women.

Ha: Men are, on average, taller than women.

Step 2: Collect data

For a statistical test to be valid, it is important to perform sampling and collect data in a way that is designed to test your hypothesis. If your data are not representative, then you cannot make statistical inferences about the population you are interested in.

Hypothesis testing example

To test differences in average height between men and women, your sample should have an equal proportion of men and women, and cover a variety of socio-economic classes and any other control variables that might influence average height.

You should also consider your scope (Worldwide? For one country?) A potential data source in this case might be census data, since it includes data from a variety of regions and social classes and is available for many countries around the world.

Step 3: Perform a statistical test

There are a variety of statistical tests available, but they are all based on the comparison of within-group variance (how spread out the data is within a category) versus between-group variance (how different the categories are from one another).

If the between-group variance is large enough that there is little or no overlap between groups, then your statistical test will reflect that by showing a low p-value. This means it is unlikely that the differences between these groups came about by chance.

Alternatively, if there is high within-group variance and low between-group variance, then your statistical test will reflect that with a high p-value. This means it is likely that any difference you measure between groups is due to chance.

Your choice of statistical test will be based on the type of variables and the level of measurement of your collected data.

Hypothesis testing example

Based on the type of data you collected, you perform a one-tailed t-test to test whether men are in fact taller than women.

This test gives you:

an estimate of the difference in average height between the two groups.

a p-value showing how likely you are to see this difference if the null hypothesis of no difference is true.

Your t-test shows an average height of 175.4 cm for men and an average height of 161.7 cm for women, with an estimate of the true difference ranging from 10.2 cm to infinity. The p-value is 0.002.

Step 4: Decide whether to reject or fail to reject your null hypothesis

Based on the outcome of your statistical test, you will have to decide whether to reject or fail to reject your null hypothesis.

In most cases you will use the p-value generated by your statistical test to guide your decision. And in most cases, your predetermined level of significance for rejecting the null hypothesis will be 0.05 – that is, when there is a less than 5% chance that you would see these results if the null hypothesis were true.

In some cases, researchers choose a more conservative level of significance, such as 0.01 (1%). This minimizes the risk of incorrectly rejecting the null hypothesis (Type I error).

Hypothesis testing example

In your analysis of the difference in average height between men and women, you find that the p-value of 0.002 is below your cutoff of 0.05, so you decide to reject your null hypothesis of no difference.

Step 5: Present your findings

The results of hypothesis testing will be presented in the results and discussion sections of your research paper, dissertation or thesis.

In the results section you should give a brief summary of the data and a summary of the results of your statistical test (for example, the estimated difference between group means and associated p-value). In the discussion, you can discuss whether your initial hypothesis was supported by your results or not.

In the formal language of hypothesis testing, we talk about rejecting or failing to reject the null hypothesis. You will probably be asked to do this in your statistics assignments.

Stating results in a statistics assignment

In our comparison of mean height between men and women we found an average difference of 13.7 cm and a p-value of 0.002; therefore, we can reject the null hypothesis that men are not taller than women and conclude that there is likely a difference in height between men and women.

Stating results in a research paper

We found a difference in average height between men and women of 14.3cm, with a p-value of 0.002, consistent with our hypothesis that there is a difference in height between men and women.