STATISTICS

Statistics Introduction

Statistics is used in all kinds of science and business applications.

Statistics gives us more accurate knowledge which helps us make better decisions.

Statistics can focus on making predictions about what will happen in the future. It can also focus on explaining how different things are connected.

Note: Good statistical explanations are also useful for predictions.

Typical Steps of Statistical Methods

The typical steps are:

1. Gathering data

2. Describing and visualizing data

3. Making conclusions

How is Statistics Used?

Statistics can be used to explain things in a precise way. You can use it to understand and make conclusions about the group that you want to know more about. This group is called the population.

Gathering data about the population will give you a sample. This is a part of the whole population. Statistical methods are then used on that sample.

The results of the statistical methods from the sample is used to make conclusions about the population.

Note: The word 'statistic' can also refer to specific bits of knowledge; like the average value of something.

Note: Data from a proper sample is often just as good data from the whole population, as long as it is representative! A good sample allows you to make accurate conclusions about the whole population.

Descriptive Statistics

Descriptive statistics is also useful for guiding further analysis, giving insight into the data, and finding what is worth investigating more closely.

Statistical Inference

Probability theory is used to calculate the certainty that those statistics also apply to the population.

Uncertainty is often expressed as confidence intervals.

Confidence intervals are numerical ways of showing how likely it is that the true value of this statistic is within a certain range for the population.

Hypothesis testing is a another way of checking if a statement about a population is true. More precisely, it checks how likely it is that a hypothesis is true is based on the sample data.

Some examples of statements or questions that can be checked with hypothesis testing:

People in the Netherlands taller than people in Denmark

Do people prefer Pepsi or Coke?

Does a new medicine cure a disease?

Note: Confidence intervals and hypothesis testing are closely related and describe the same things in different ways. Both are widely used in science.

Causal Inference

Causal inference is used to investigate if something causes another thing.

For example: Does rain make plants grow?

Note: Good experimental design is often difficult to achieve because of ethical concerns or other practical reasons.

Prediction

Predictions about future events are called forecasts. Not all predictions are about the future.

Some predictions can be about something else that is unknown, even if it is not in the future.

Explanation

Making conclusions about causality should be done carefully.

Population and Samples

Population: Everything in the group that we want to learn about.

Sample: A part of the population.

Parameters and Statistics

Parameter: A number that describes something about the whole population.

Sample statistic: A number that describes something about the sample.

Sample statistics gives us estimates for parameters.

Some Important Examples

Parameter	**Sample statistic**
Mean	Sample mean
Median	Sample median

Different Types of Sampling Methods

Random Sampling

A random sample is where every member of the population has an equal chance to be chosen

Note: Every other sampling method is compared to how close it is to a random sample - the closer, the better.

Convenience Sampling

A convenience sample is where the participants that are the easiest to reach are chosen.

Systematic Sampling

A systematic sample is where the participants are chosen by some regular system.

For example:

* The first 30 people in a queue

* Every third on a list

Stratified Sampling

A stratified sample is where the population is split into smaller groups called 'strata'.

The 'strata' can, for example, be based on demographics, like:

* Different age groups

* Professions

Clustered Sampling

A clustered sample is where the population is split into smaller groups called 'clusters'.

The clusters are usually natural, like different cities in a country.

Different types of data

Qualitative Data

Information about something that can be sorted into different categories that can't be described directly by numbers. With categorical data we can calculate statistics like proportions.

Examples:

* Brands

* Nationality

Quantitative Data

Information about something that is described by numbers. With numerical data we can calculate statistics like the average.

Examples:

* Income

* Age

Measurement Levels

Nominal Level

Categories (qualitative data) without any order.

Examples:

* Brand names

* Countries

Ordinal level

Categories that can be ordered (from low to high), but the precise "distance" between each is not meaningful.

Examples:

* Letter grade scales from F to A

* Military ranks

Interval Level

Data that can be ordered and the distance between them is objectively meaningful. But there is no natural 0-value where the scale originates.

Examples:

* Years in a calendar

* Temperature measured in Fahrenheit

Ratio Level

Data that can be ordered and there is a consistent and meaningful distance between them. And it also has a natural 0-value.

Examples:

* Money

* Age

Statistics - Descriptive Statistics

Key Features to Describe about Data

The Center of the Data

The center of the data is where most of the values are concentrated.

Different kinds of averages, like mean, median and mode, are measures of the center.

The Variation of the Data

The variation of the data is how spread out the data are around the center.

Statistics like standard deviation, range and quartiles are measures of variation.

The Shape of the Data

The shape of the data can refer to the how the data are bunched up on either side of the center.

Statistics like skew describe if the right or left side of the center is bigger. Skew is one type of shape parameters.

Frequency Tables

One typical of presenting data is with frequency tables.

A frequency table counts and orders data into a table. Typically, the data will need to be sorted into intervals.

Visualizing Data

Different types of graphs are used for different kinds of data. For example:

* Pie charts for qualitative data

* Histograms for quantitative data

* Scatter plots for bivariate data

* box plots show where the quartiles are.

Average

The Center of the Data

The center of the data is where most of the values in the data are located.

There are different types of averages. The most commonly used are:

* Mean

* Median

* Mode

Note: In statistics, averages are often referred to as 'measures of central tendency'.

Mean

The mean is the sum of all the values in the data divided by the total number of values in the data.

Calculating the Mean

You can calculate the mean for both the **population** and the **sample**.

Calculating the **population mean** (μ) and sample mean (x) is done with this formula:

Population Mean	Sample Mean
$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$
N = number of items in the population	n = number of items in the sample

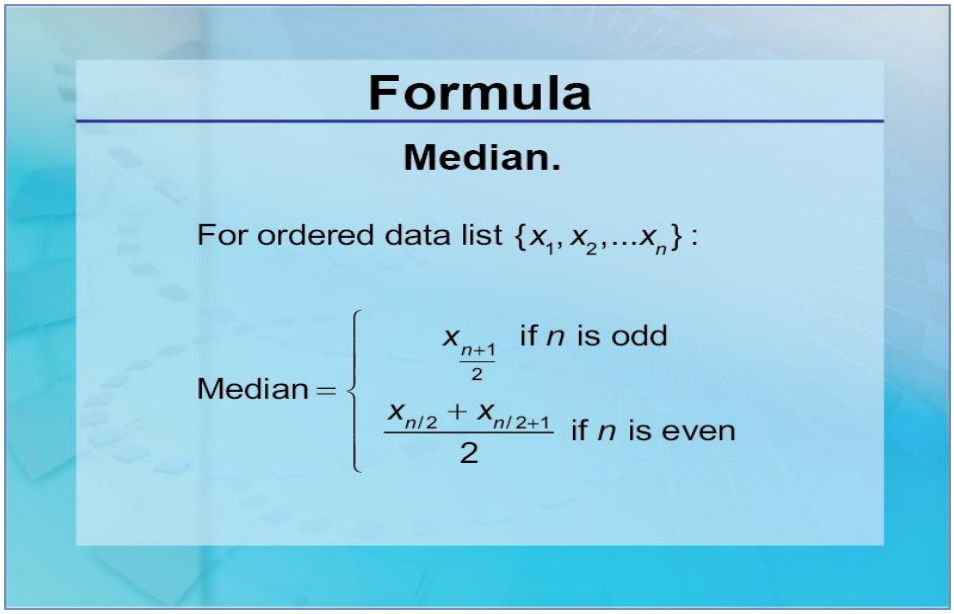
Calculation with Programming

In [1]:
```
1 import numpy as np
2 values = [4,11,25,3]
3 mean = np.mean(values)
4 print(mean)
```

```
10.75
```

Median

The median is the **middle** value in a data set ordered from low to high. The median is a type of average value, which describes where the center of the data is located.



Finding the Median with Programming

In [2]:
```
1 import numpy
2 values = [13,21,21,40,42,48,55,72]
3 median = np.median(values)
4 print(median)
```

```
41.0
```

Mode

The mode is the value(s) that appears most often in the data.

A distribution of values with only one mode is called **unimodal**.

A distribution of values with two modes is called **bimodal**. In general, a distribution with more than one mode is called **multimodal**.

Finding the Mode with Programming

In [6]:
```
1 import statistics
2 values = [4,7,3,8,11,10,19,6,9,12,12]
3 print(statistics.mode(values))
```

```
12
```

Variation-

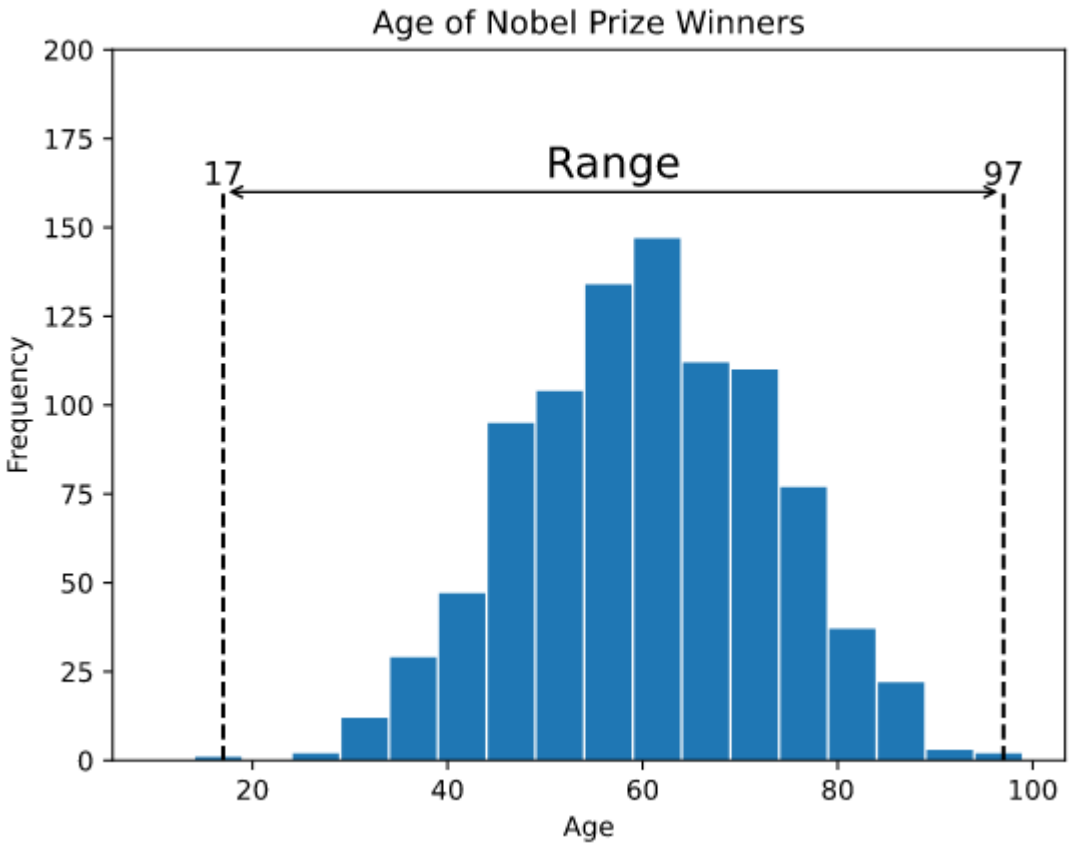
Variation is a measure of how spread out the data is around the center of the data.

There are different measures of variation. The most commonly used are:

- Range
- Quartiles and Percentiles
- Interquartile Range
- Standard Deviation

Range

The range is the difference between the smallest and the largest value of the data.



Calculating the Range with Programming

In [7]:
```
1 import numpy
2 values = [13,21,21,40,48,55,72]
3 x = np.ptp(values)
4 print(x)
```

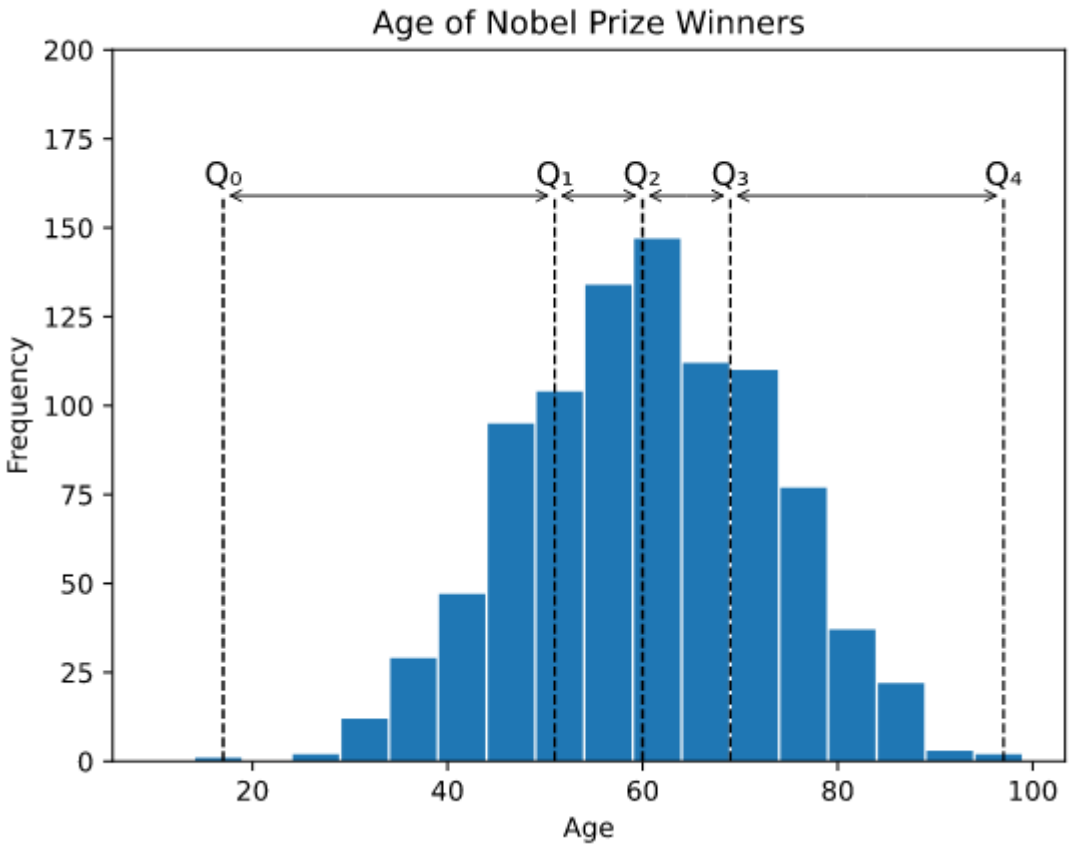
```
59
```

Quartiles and Percentiles

Quartiles and percentiles are both types of **quantiles**.

Quartiles

Quartiles are values that separate the data into four equal parts.



- Q_0 is the smallest value in the data.
- Q_1 is the value separating the first quarter from the second quarter of the data.
- Q_2 is the middle value (median), separating the bottom from the top half.
- Q_3 is the value separating the third quarter from the fourth quarter
- Q_4 is the largest value in the data.

Calculating Quartiles with Programming

In [9]:
```
1 import numpy
2 values = [13,21,21,40,42,48,55,72]
3 x = np.quantile(values,[0,0.25,0.50,0.75])
4 print(x)
```

```
[13.   21.   41.   49.75]
```

Percentiles

Percentiles are values that separate the data into 100 equal parts.

The 25th percentile ($P_{25\%}$) is the same as the first quartile (Q_1).

The 50th percentile ($P_{50\%}$) is the same as the second quartile (Q_2) and the median.

The 75th percentile ($P_{75\%}$) is the same as the third quartile (Q_3)

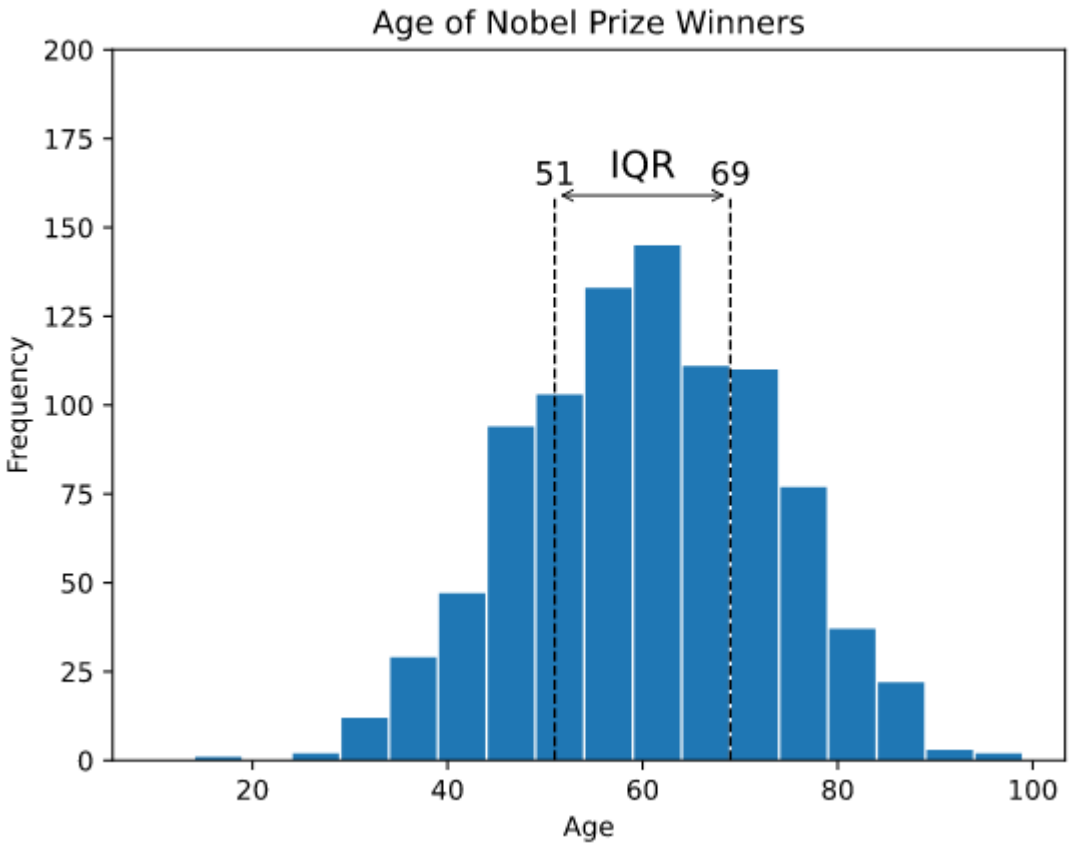
Calculating Percentiles with Programming

In [10]:
```
1 import numpy
2 values = [13,21,21,40,42,48,55,72]
3 x = np.percentile(values,65)
4 print(x)
```

```
45.3
```

Interquartile Range

Interquartile range is the difference between the first and third quartiles (Q_1 and Q_3).



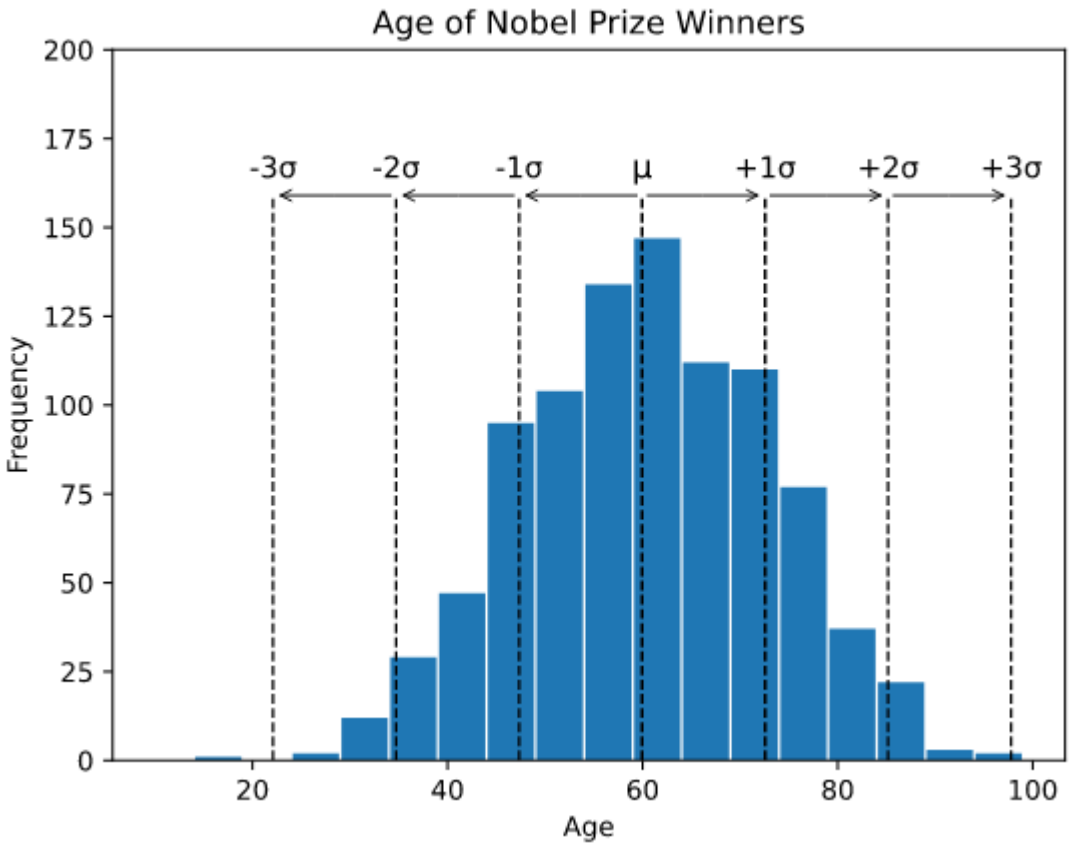
Calculating the Interquartile Range with Programming

In [11]:
```
1 from scipy import stats
2 values = [13,21,21,40,42,48,55,72]
3 x = stats.iqr(values)
4 print(x)
```

```
28.75
```

Standard Deviation

Standard deviation (σ) measures how far a 'typical' observation is from the average of the data (μ).



If the data is **normally distributed:**

- Roughly 68.3% of the data is within 1 standard deviation of the average (from μ-1σ to μ+1σ)
- Roughly 95.5% of the data is within 2 standard deviations of the average (from μ-2σ to μ+2σ)
- Roughly 99.7% of the data is within 3 standard deviations of the average (from μ-3σ to μ+3σ)

Note: A **normal** distribution has a "bell" shape and spreads out equally on both sides.

Standard Deviation Formula	
Population	**Sample**
$\sigma = \sqrt{\frac{\sum(X-\mu)^2}{N}}$	$s = \sqrt{\frac{\sum(X-\bar{x})^2}{n-1}}$
X – The Value in the data distribution *μ – The population Mean* *N – Total Number of Observations*	*X – The Value in the data distribution* *$\bar{x}$ – The Sample Mean* *n - Total Number of Observations*

Calculating the Standard Deviation with Programming

```
3 import numpy
4 values = [4,11,7,14]
5 x = np.std(values)
6 print(x)
```

```
3.8078865529319543
```

Statistics- Inferential Statistics

Estimation

Statistics from a sample are used to estimate population parameters.The

most likely value is called a **point estimate**.

There is **always** uncertainty when estimating.

The uncertainty is often expressed as **confidence intervals** defined by a likely lowest and highest value for the parameter.

Hypothesis Testing

Hypothesis testing is a method to check if a claim about a population is true. More precisely, it checks how likely it is that a hypothesis is true is based on the sample data.

The steps of the test depends on:

- Type of data (categorical or numerical)
- If you are looking at:
  - A single group
  - Comparing one group to another
  - Comparing the same group before and after a change

Normal Distribution

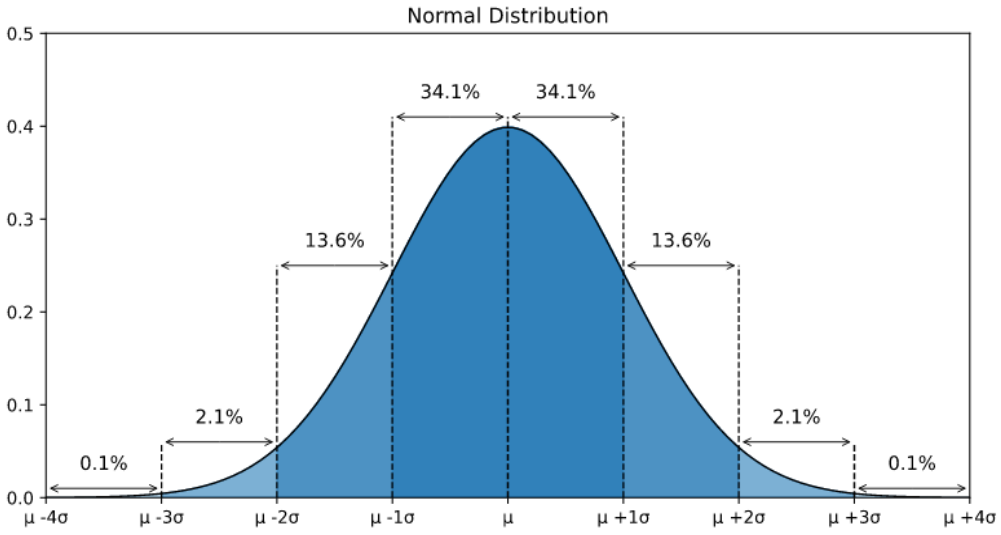
The normal distribution is described by the mean (μ) and the standard deviation (σ).

The normal distribution is often referred to as a 'bell curve' because of it's shape:

- Most of the values are around the center (μ)
- The median and mean are equal
- It has only one mode
- It is symmetric, meaning it decreases the same amount on the left and the right of the center

The area under the curve of the normal distribution represents probabilities for the data.

The area under the whole curve is equal to 1, or 100%



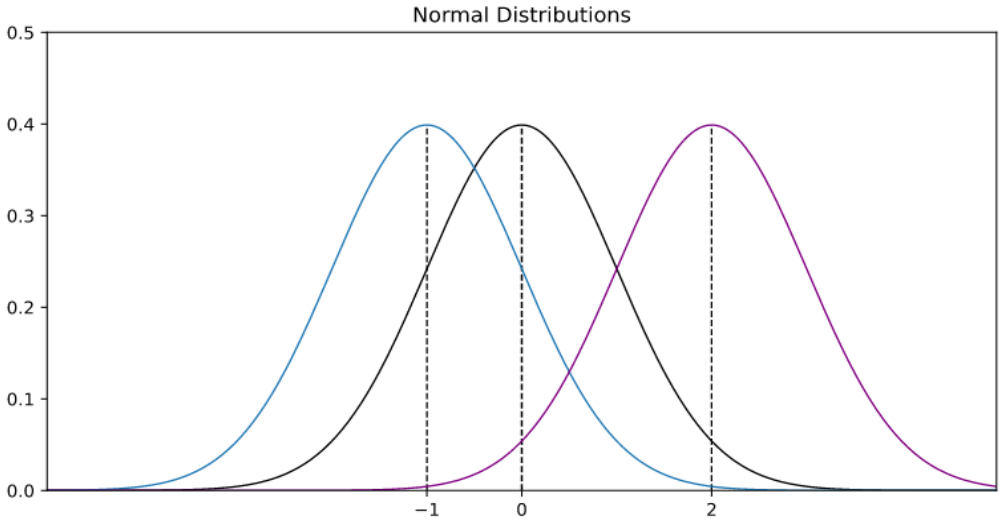
Note: Probabilities of the normal distribution can only be calculated for intervals (between two values).

Different Mean and Standard Deviations

The mean describes where the center of the normal distribution is.

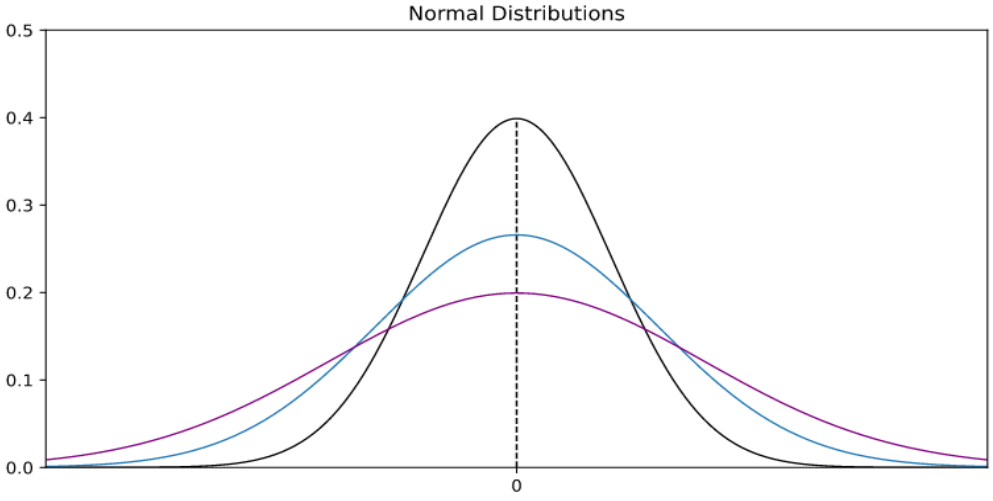
Here is a graph showing three different normal distributions with

the **same** standard deviation but different means.



The standard deviation describes how spread out the normal distribution is.

Here is a graph showing three different normal distributions with
the **same** mean but different standard deviations.



The purple curve has the biggest standard deviation and the black curve has the smallest standard deviation.

The area under each of the curves is still 1, or 100%.

Probability Distributions

Probability distributions are functions that calculates the probabilities of the outcomes of random variables.

Typical examples of random variables are coin tosses and dice rolls.

Standard Normal Distribution

The standard normal distribution is a normal distribution where the mean is 0 and the standard deviation is 1.

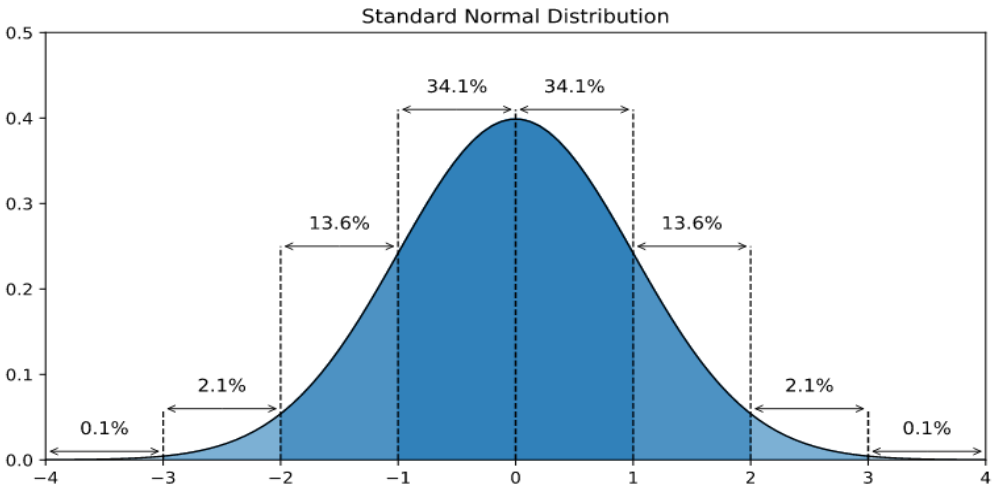
Normally distributed data can be transformed into a standard normal distribution.

Standardizing normally distributed data makes it easier to compare different sets of data.

The standard normal distribution is used for:

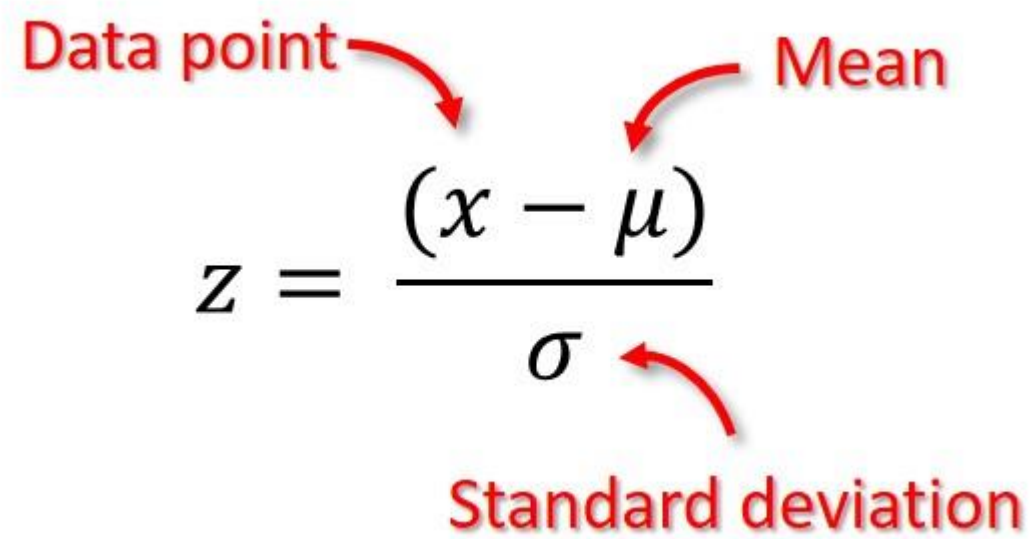
- Calculating confidence intervals
- Hypothesis tests

Here is a graph of the standard normal distribution with probability values (p-values) between the standard deviations:



The standard normal distribution is also called the 'Z-distribution' and the values are called 'Z-values' (or Z-scores).

FORMULA OF Z-SCORE-



Z-Values

Z-values express how many standard deviations from the mean a value is.

`The mean height of people in Germany is 170 cm (μ)`

`The standard deviation of the height of people in Germany is 10 cm (σ)`

`Bob is 200 cm tall (x)`

Bob is 30 cm taller than the average person in Germany.

30 cm is 3 times 10 cm. So Bob's height is 3 standard deviations larger than mean height in Germany.

Using the formula:

$Z = x - \mu / \sigma = 200 - 170 / 10 = 30 / 10 = 3$

The Z-value of Bob's height (200 cm) is 3.

Finding the P-value of a Z-Value

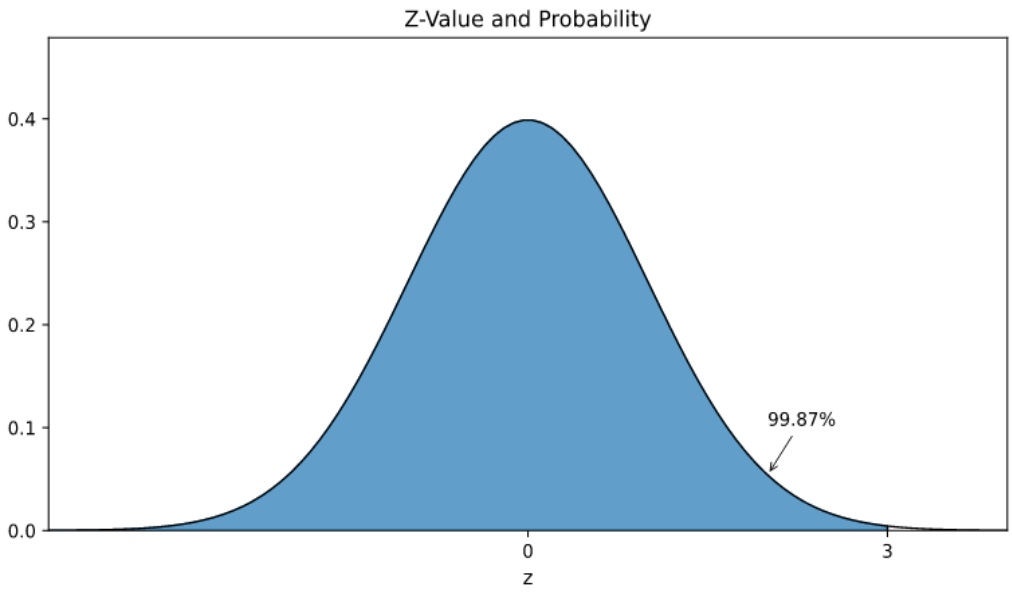
Using a Z-table or programming we can calculate how many people Germany are shorter than Bob and how many are taller.

```
In [13]:
1 import scipy.stats as stats
2 print(stats.norm.cdf(3))
```

```
0.9986501019683699
```

Using either method we can find that the probability is ≈0.9987, or 99.87%

Which means that Bob is taller than 99.87% of the people in Germany.



Finding P-value

To find the p-value above the z-value we can calculate 1 minus the probability.

So in Bob's example, we can calculate 1 - 0.9987 = 0.0013, or 0.13%.

T Distribution

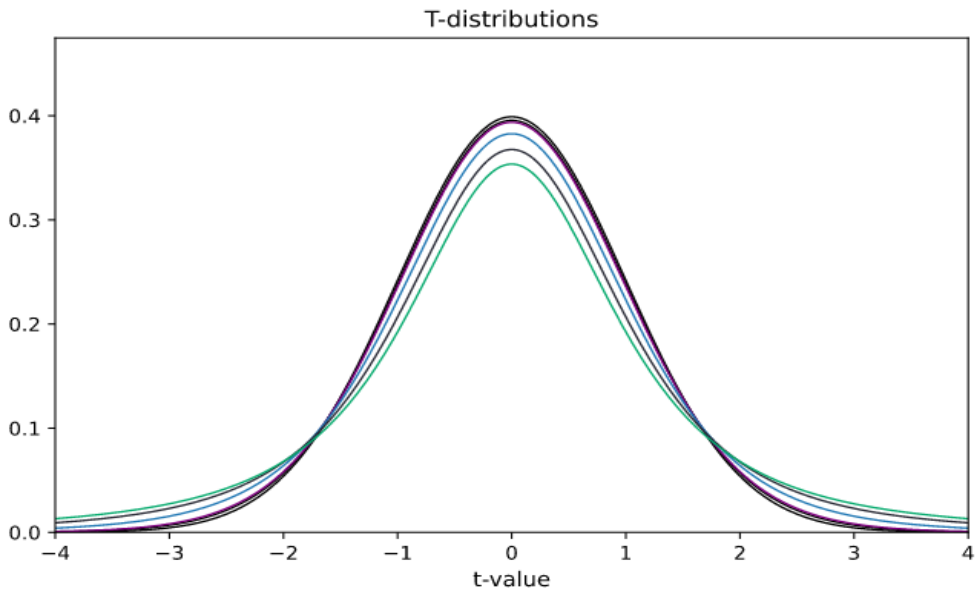
The t-distribution is used for estimation and hypothesis testing of a population mean (average).

The t-distribution is adjusted for the extra uncertainty of estimating the mean.

If the sample is small, the t-distribution is wider. If the sample is big, the t-distribution is narrower.

The bigger the sample size is, the closer the t-distribution gets to the standard normal distribution.

Below is a graph of a few different t-distributions.



The t-distribution is used to find critical t-values **and** p-values **(probabilities) for estimation and hypothesis testing.**

Note: **Finding the critical t-values and p-values of the t-distribution is similar z-values and p-values of the standard normal distribution. But make sure to use the correct degrees of freedom.**

Finding the P-Value of a T-Value

```
In [14]:  1 import scipy.stats as stats
          2 print(stats.t.cdf(2.1,29))
```

```
0.9777290209818548
```

Finding the T-value of a P-Value

```
In [15]:  1 import scipy.stats as stats
          2 print(stats.t.ppf(.75,9))
```

```
0.7027221467513188
```

Estimation

Point estimates are the most likely value for a population parameter.

Confidence intervals express the uncertainty of an estimated population parameter.

The Point Estimate

A point estimate is calculated from a sample.

The point estimate depends on the type of data:

- **Categorical data**: the number of occurrences divided by the sample size.
- **Numerical data**: the mean (the average) of the sample.

One example could be:

```
The point estimate for the average height of people in Denmark is 180 cm.
```

Estimates are always **uncertain**. This uncertainty can be expressed with a **confidence interval**.

Confidence Intervals

The confidence interval is defined by a **lower bound** and an **upper bound**.

This gives us a range of values that the true parameter is likely to be between.

For example that:

```
The average height of people in Denmark is between 170 cm and 190 cm.
```

Here, 170 cm is the lower bound, and 190 cm is the upper bound.

The lower and upper bounds of a confidence interval is based on the **confidence level**.

The Confidence Level

Confidence levels can be expressed as percentages or decimal numbers, and the most commonly used are:

- 90% (0.90)
- 95% (0.95)
- 99% (0.99)

The higher the confidence level, the bigger the interval will be.

For example, the confidence intervals for the average height of people in Denmark might be:

```
90% confidence level: between 175 cm and 185 cm.

95% confidence level: between 170 cm and 190 cm.

99% confidence level: between 160 cm and 200 cm.
```

We use this confidence level together with a probability distribution to decide how large the **margin of error** is.

The Margin of Error

The margin of error is the distance between the point estimate and the lower and upper bounds.

The margin of error is based on the confidence level and the data we have from the sample.

For example, if the point estimate for the average height of people in Denmark is 180 cm:

```
5 cm margin of error: between 175 cm and 185 cm.
```

Steps for Calculating the Confidence Interval

The following steps are used to calculate a confidence interval:

1. Check the conditions
2. Find the point estimate
3. Decide the confidence level
4. Calculate the margin of error
5. Calculate the confidence interval

One **condition** is that the sample is randomly selected from the population.

The other conditions depends on what type of parameter you are calculate the confidence interval for.

Commonly estimated parameters are:

- Proportions (for qualitative data)
- Mean values (for numerical data)

1. Checking the Conditions

The conditions for calculating a confidence interval for a proportion are:

- The sample is randomly selected
- There is only two options:
    - Being in the category
    - Not being in the category
- The sample needs at least:
    - 5 members in the category

- 5 members not in the category

In our example, we randomly selected 6 people that were born in the US.

The rest were not born in the US, so there are 24 in the other category.

Note: It is possible to calculate a confidence interval without having 5 of each category. But special adjustments need to be made.

2. Finding the Point Estimate

The point estimate is the sample proportion (p^).

The formula for calculating the sample proportion is the number of occurrences (x) divided by the sample size (n):

$$\hat{p} = \frac{x}{n}$$

In our example, 6 out of 30 were born in the US: x is 6, and n is 30.

So the point estimate for the proportion is:

p^=x/n=630=0.2—=20%

So 20% of the sample were born in the US.

3. Deciding the Confidence Level

The confidence level is expressed with a percentage or a decimal number.

For example, if the confidence level is 95% or 0.95:

The remaining probability (α) is then: 5%, or 1 - 0.95 = 0.05.

Commonly used confidence levels are:

- 90% with α = 0.1
- 95% with α = 0.05
- 99% with α = 0.01

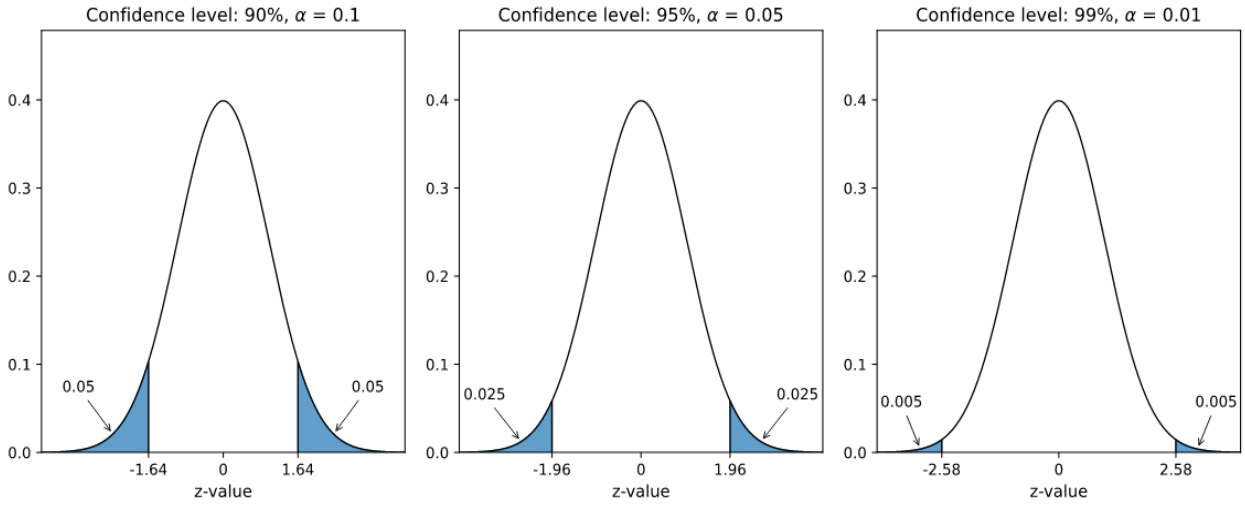
Note: A 95% confidence level means that if we take 100 different samples and make confidence intervals for each:

The true parameter will be inside the confidence interval 95 out of those 100 times.

We use the standard normal distribution to find the **margin of error** for the confidence interval.

The remaining probabilities (α) are divided in two so that half is in each tail area of the distribution.

The values on the z-value axis that separate the tails area from the middle are called **critical z-values**.



4. Calculating the Margin of Error

The margin of error is the difference between the point estimate and the lower and upper bounds.

The margin of error (E) for a proportion is calculated with a critical z-value and the **standard error**:

$$E = Z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

The critical z-value Zα/2 is calculated from the standard normal distribution and the confidence level.

The standard error $\hat{p}(1-\hat{p})/n$ is calculated from the point estimate ($\hat{p}$) and sample size (n).

In our example with 6 US-born Nobel Prize winners out of a sample of 30 the standard error is:

$\hat{p}(1-\hat{p})/n = 0.2(1-0.2)30 = 0.2 \cdot 0.830 = 0.1630 = 0.00533 \approx 0.073$

If we choose 95% as the confidence level, the α is 0.05.

So we need to find the critical z-value

In [16]:
```
import scipy.stats as stats
print(stats.norm.ppf(1-0.025))
```

```
1.959963984540054
```

5. Calculate the Confidence Interval

The lower and upper bounds of the confidence interval are found by subtracting and adding the margin of error (E) from the point estimate ($\hat{p}$).

In our example the point estimate was 0.2 and the margin of error was 0.143, then:

The lower bound is:

$\hat{p} - E = 0.2 - 0.143 = 0.057 —$

The upper bound is:

$\hat{p} + E = 0.2 + 0.143 = 0.343 —$

The confidence interval is:

[0.057,0.343] or [5.7%,34.3%]

Hypothesis Testing

A **hypothesis** is a claim about a population parameter.

A **hypothesis test** is a formal procedure to check if a hypothesis is true or not.

Examples of claims that can be checked:

`The average height of people in Denmark is` **`more`** `than 170 cm.`

The Null and Alternative Hypothesis

Hypothesis testing is based on making two different claims about a population parameter.

The **null** hypothesis (H0) and the **alternative** hypothesis (H1) are the claims.

The two claims needs to be **mutually exclusive**, meaning only one of them can be true.

The alternative hypothesis is typically what we are trying to prove.

For example, we want to check the following claim:

`"The average height of people in Denmark is more than 170 cm."`

In this case, the **parameter** is the average height of people in Denmark (μ).

The null and alternative hypothesis would be:

`Null hypothesis``: The average height of people in Denmark` **`is`** `170 cm.`

`Alternative hypothesis``: The average height of people in Denmark is` **`more`** `than 170 cm.`

The claims are often expressed with symbols like this:

`H0: μ=170cm`

`H1: μ>170cm`

If the data supports the alternative hypothesis, we **reject** the null hypothesis and **accept** the alternative hypothesis.

If the data does **not** support the alternative hypothesis, we **keep** the null hypothesis.

Note: The alternative hypothesis is also referred to as H_{A}

The Significance Level

The significance level (α) is the **uncertainty** we accept when rejecting the null hypothesis in the hypothesis test.

The significance level is a percentage probability of accidentally making the wrong conclusion.

Typical significance levels are:

- α=0.1 (10%)
- α=0.05 (5%)
- α=0.01 (1%)

A lower significance level means that the evidence in the data needs to be stronger to reject the null hypothesis.

There is no "correct" significance level - it only states the uncertainty of the conclusion.

Note: A 5% significance level means that when we reject a null hypothesis:

We expect to reject a **true** null hypothesis 5 out of 100 times.

The Test Statistic

The test statistic is used to decide the outcome of the hypothesis test.

The test statistic is a **standardized** value calculated from the sample.

Standardization means converting a statistic to a well known **probability distribution**.

The type of probability distribution depends on the type of test.

he Critical Value and P-Value Approach

There are two main approaches used for hypothesis tests:

- The **critical value** approach compares the test statistic with the critical value of the significance level.
- The **p-value** approach compares the p-value of the test statistic and with the significance level.

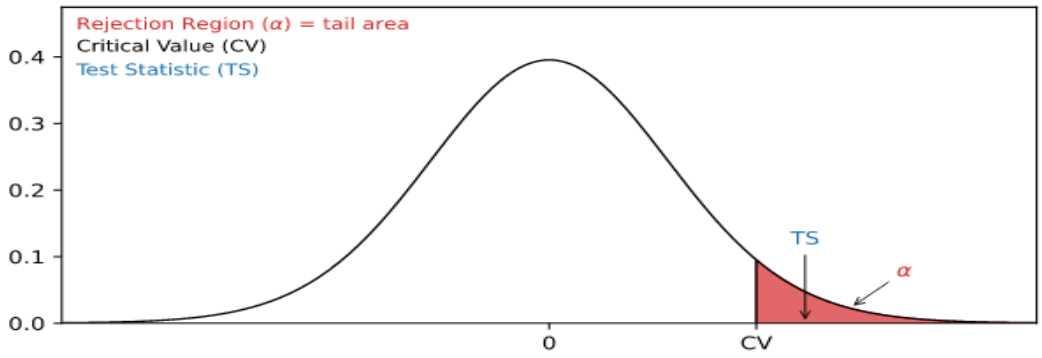
The Critical Value Approach

The critical value approach checks if the test statistic is in the **rejection region**.

The rejection region is an area of probability in the tails of the distribution.

The size of the rejection region is decided by the significance level (α).

The value that separates the rejection region from the rest is called the **critical value**.

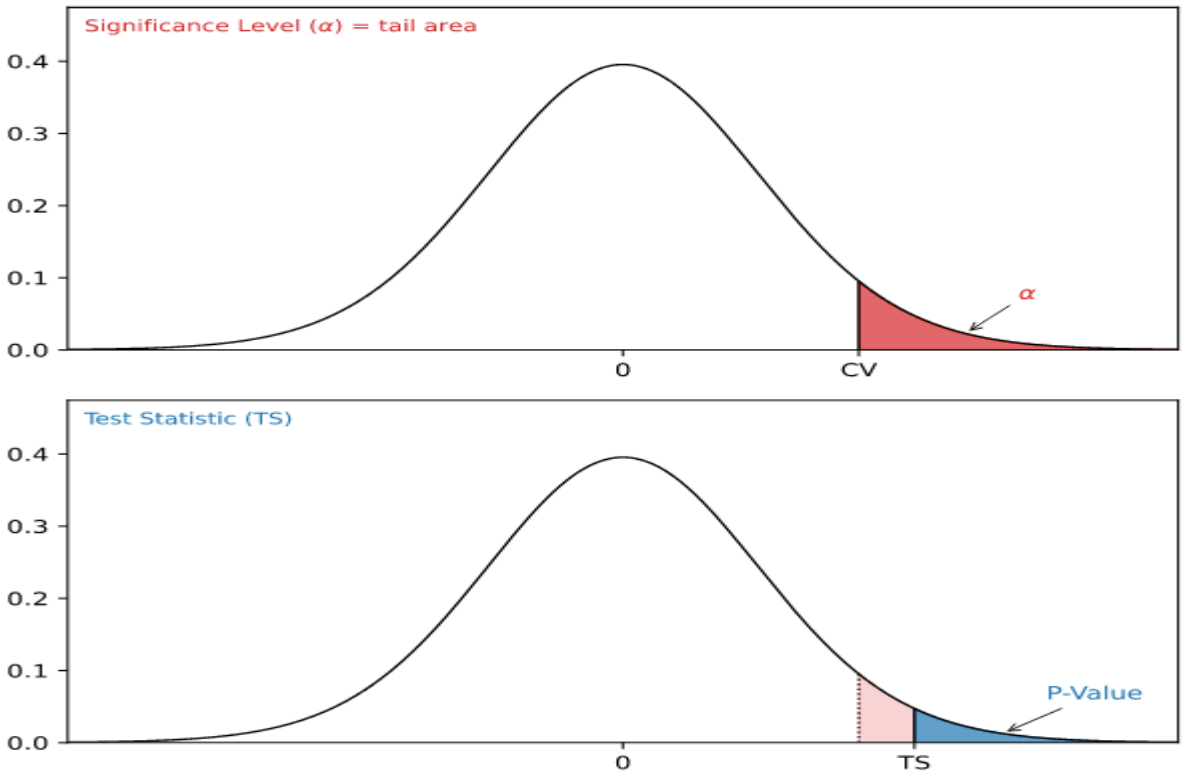


If the test statistic is **inside** this rejection region, the null hypothesis is **rejected**.

The P-Value Approach

The p-value approach checks if the p-value of the test statistic is **smaller** than the significance level (α).

The p-value of the test statistic is the area of probability in the tails of the distribution from the value of the test statistic.



If the p-value is **smaller** than the significance level, the null hypothesis is **rejected**.

The p-value directly tells us the **lowest significance level** where we can reject the null hypothesis.

Steps for a Hypothesis Test

The following steps are used for a hypothesis test:

1. Check the conditions
2. Define the claims
3. Decide the significance level
4. Calculate the test statistic
5. Conclusion

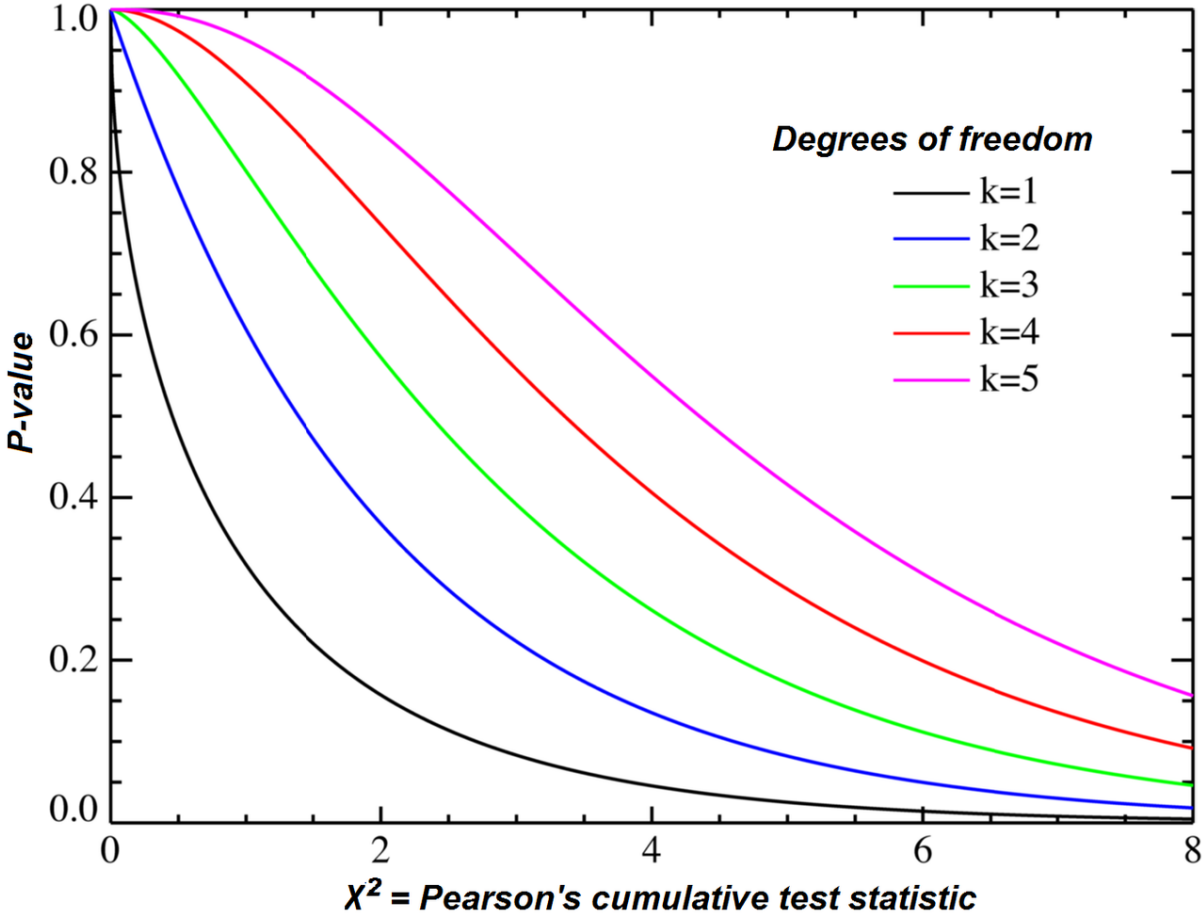
Chi Square

A **chi-squared test** (symbolically represented as χ^2) is basically a data analysis on the basis of observations of a random set of variables. Usually, it is a comparison of two statistical data sets. A chi-square test is a statistical test used **to compare observed results with expected results**. The purpose of this test is to determine if a difference between observed data and expected data is due to chance, or if it is due to a relationship between the variables you are studying. It gives the probability of independent variables.

Formula

$$X^2 = \sum \frac{(\text{Observed value} - \text{Expected value})^2}{\text{Expected value}}$$

Note: Chi-squared test is applicable only for categorical data, such as men and women falling under the categories of Gender, Age, Height, etc.



Anova Test

- Analysis of variance, or ANOVA, is a statistical method that separates observed variance data into different components to use for additional tests.

- A one-way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables.
- If no true variance exists between the groups, the ANOVA's F-ratio should equal close to 1.

when to use Anova test ?

1. It is only conducted when there is no relationship between the subjects in each sample. this means that subjects in the first group cannot also be in the second group ie independent samples between groups.
2. Groups must have equal sample size.

The Formula for ANOVA is

$$F = MST/MSE$$

where:

F=ANOVA coefficient

MST=Mean sum of squares due to treatment

MSE=Mean sum of squares due to error

Type of Anova

There are two main types of ANOVA:

• One-way (or unidirectional) and

• Two-way.

There also variations of ANOVA. For example, MANOVA (multivariate

ANOVA) differs from ANOVA as the former tests for multiple dependent

variables simultaneously while the latter assesses only one dependent

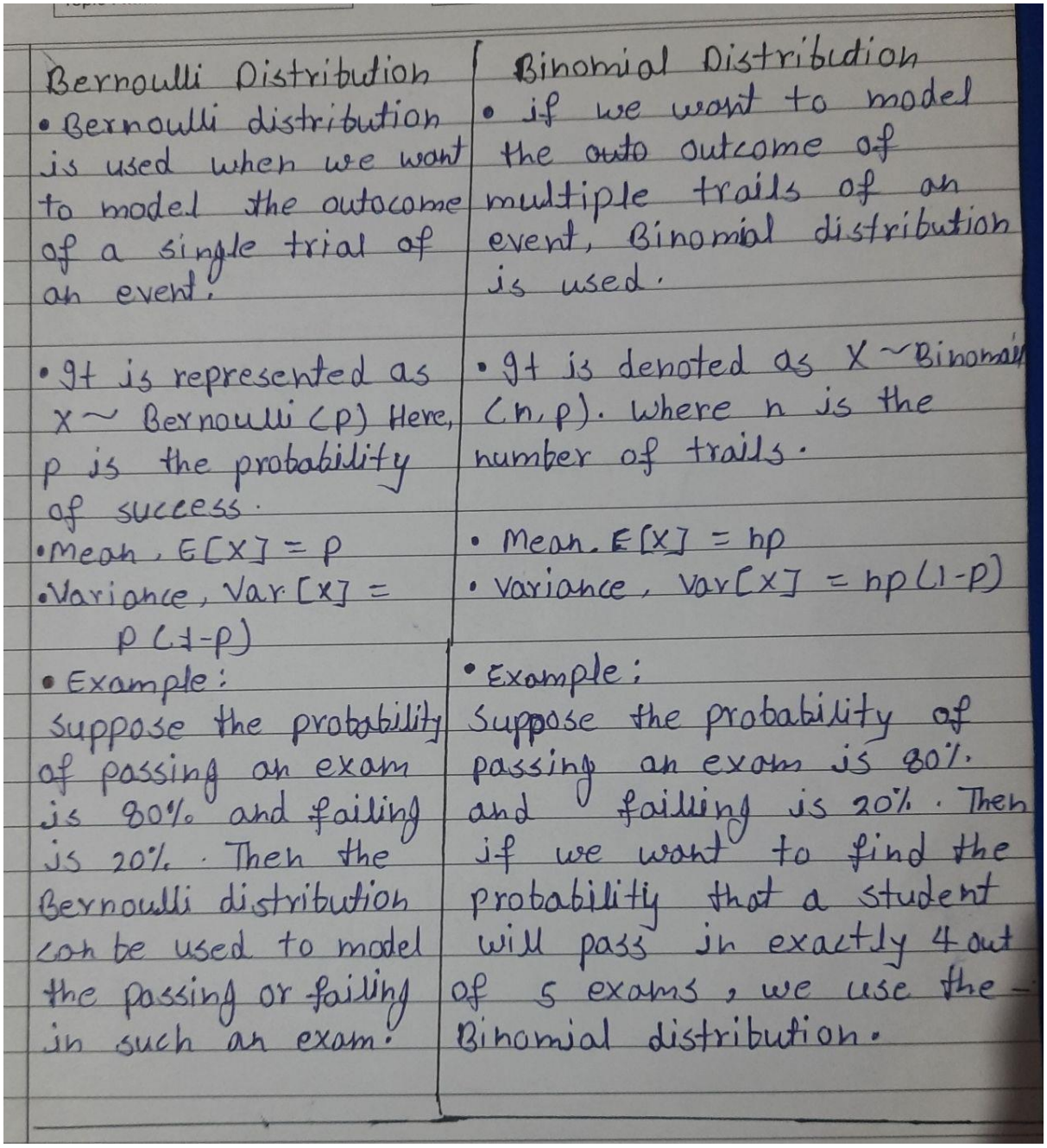
variable at a time. One-way or two-way refers to the number of independent variables in your analysis of variance test.

A one-way ANOVA evaluates the impact of a sole factor on a sole response variable. It determines whether all the samples are the same. The one-way ANOVA is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups.

A two-way ANOVA is an extension of the one-way ANOVA. With a one-way, you have one independent variable affecting a dependent variable. With a two-way ANOVA, there are two independents. For example, a two-way ANOVA allows a company to compare worker productivity based on two independent variables, such as salary and skill set. It is utilized to observe the interaction between the two factors and tests the effect of two factors at the same time.

Example: A grocery chain wants to know if three different types of advertisements affect mean sales differently. They use each type of advertisement at 10 different stores for one month and measure total sales for each store at the end of the month.

Difference between Bernoulli and Binomial Distribution

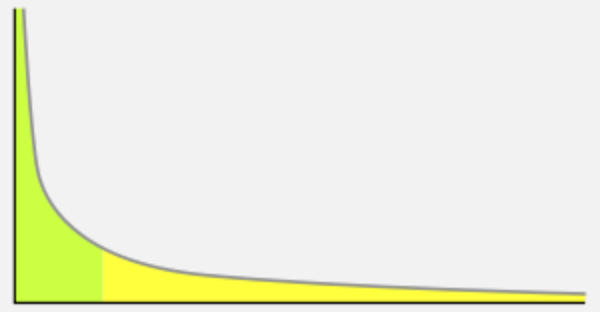


Bernoulli Distribution	Binomial Distribution
• Bernoulli distribution is used when we want to model the outcome of a single trial of an event.	• if we want to model the outo outcome of multiple trails of an event, Binomial distribution is used.
• It is represented as $X \sim$ Bernoulli (p) Here, p is the probability of success.	• It is denoted as $X \sim$ Binomial (n, p). where n is the number of trails.
• Mean, $E[X] = p$	• Mean, $E[X] = np$
• Variance, $Var[X] = p(1-p)$	• Variance, $Var[X] = np(1-p)$
• Example: Suppose the probability of passing an exam is 80% and failing is 20%. Then the Bernoulli distribution can be used to model the passing or failing in such an exam.	• Example; Suppose the probability of passing an exam is 80%. and failing is 20%. Then if we want to find the probability that a student will pass in exactly 4 out of 5 exams, we use the Binomial distribution.

Power Law Distribution/Pareto Distribution

A power law distribution has the property that large numbers are rare, but smaller numbers are more common. So it is more common for a person to make a small amount of money versus a large amount of money.

The **Pareto** distribution is a continuous power law distribution that is based on the observations that Pareto made.



An example power-law graph that demonstrates ranking of popularity. To the right is the long tail, and to the left are the few that dominate (also known as the 80–20 rule).