

AGO CET

Statistic for Data Science

Learning Outcomes

Understand different methods in the Descriptive Statistic

Understand different methods in the Inferential Statistic

Statistics Basics

Introduction to Statistics

What is Descriptive Statistic?

Types of Data based on Measurement Scale

Central Tendency

Skewness/Kurtosis

Correlation

What is Statistics?

Statistics is a set of mathematical methods and tools that enable us to answer important questions about data. It is divided into two categories:

- ▶ Descriptive Statistics - this offers methods to summarise data by transforming raw observations into meaningful information that is easy to interpret and share.
- ▶ Inferential Statistics - this offers methods to study experiments done on small samples of data and chalk out the inferences to the entire population (entire domain).

Uses of Statistics

From Data to Knowledge

- In isolation, raw observations are just data.
- We use descriptive statistics to transform these observations into insights that make sense.
- Then we can use inferential statistics to study small samples of data and extrapolate our findings to the entire population.

Uses of Statistics

Statistics helps answer questions like...

- What features are the most important?
- How should we design the experiment to develop our product strategy?
- What performance metrics should we measure?
- What is the most common and expected outcome?
- How do we differentiate between noise and valid data?

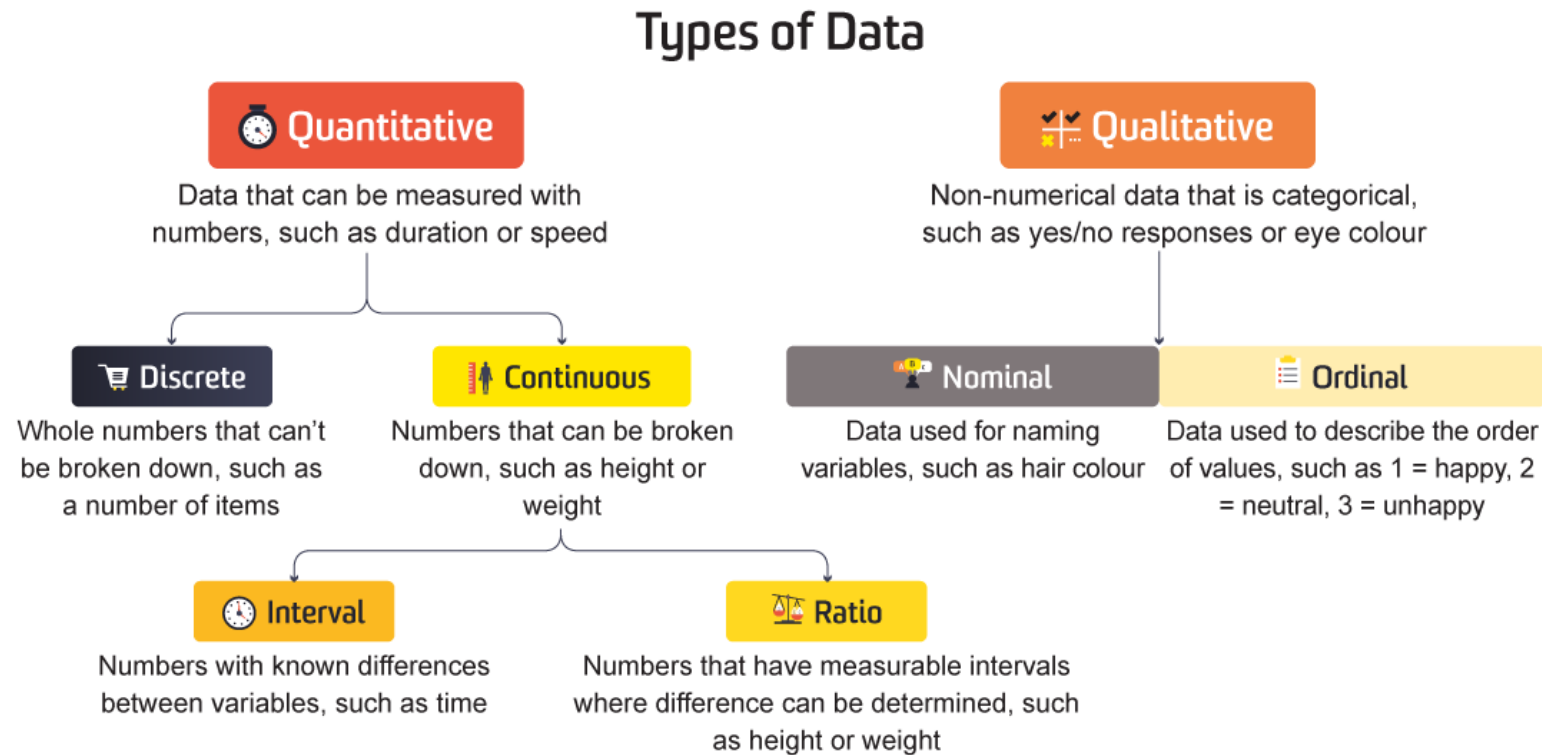
Descriptive Statistics

What is Descriptive Statistics?

- It's about summative information about the given dataset. It tells us what the data set's overall properties.
 - ▷ Types of Data based on Measurement Scale
 - ▷ Central Tendency

Descriptive Statistics

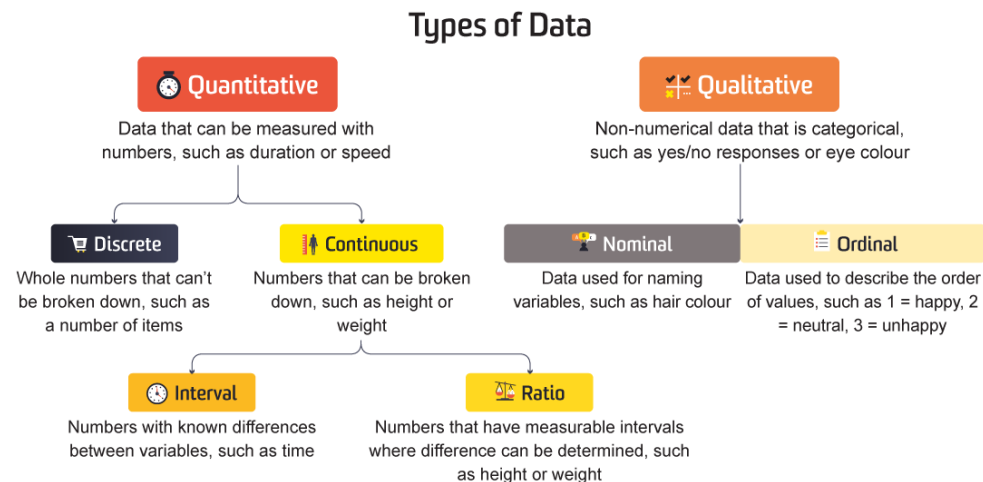
Types of Data based on Measurement Scale:



Descriptive Statistics

Types of Data based on Measurement Scale:

Product ID	SupplierID	Cost	Last Order Date	Size
P001	KSC	\$1.5	10-1-2017	Small
P101	PDS	\$3.25	21-3-2017	Medium
P333	KING	\$16.0	1-8-2017	Large
Nominal	Nominal	Ratio	Interval	Ordinal



Graphs and tables

Frequency distribution tables

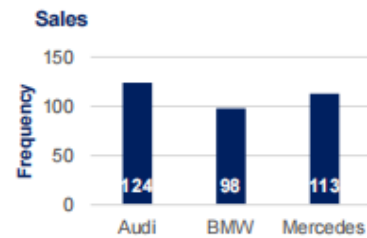
Bar charts

Pie charts

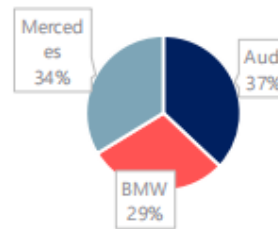
Pareto diagrams

Frequency	
Audi	124
BMW	98
Mercedes	113
Total	335

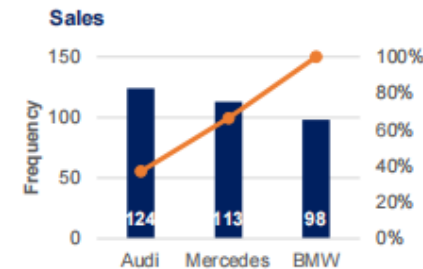
Frequency distribution tables show the category and its corresponding absolute frequency.



Bar charts are very common. Each bar represents a category. On the y-axis we have the absolute frequency.



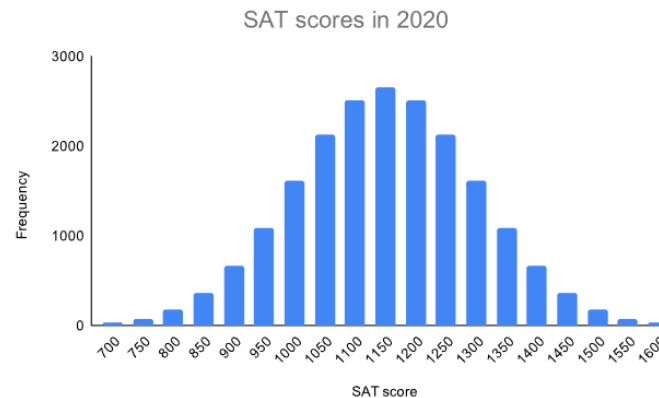
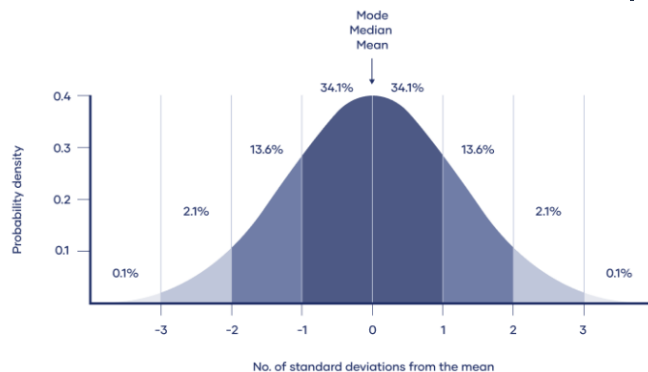
Pie charts are used when we want to see the share of an item as a part of the total. Market share is almost always represented with a pie chart.



The Pareto diagram is a special type of bar chart where the categories are shown in descending order of frequency, and a separate curve shows the cumulative frequency.

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.



Descriptive Statistics

- Measures of central tendency:
 - ▷ Mean
 - ▷ Median
 - ▷ Mode

Measures of Central Tendency - Mean

■ Mean - the average

nums = {872, 432, 397, 427, 388, 782, 397}

mean = $\sum \text{nums} / |\text{nums}|$

Measures of Central Tendency - Median

■ Median - the center value in the ordered list

nums = {872, 432, 397, 427, 388, 782, 397}
= {388, 397, 397, 427, 432, 782, 872}

median = 427

nums = {872, 432, 397, 388, 782, 397}
= {388, 397, 397, 432, 782, 872}

median = (397 + 432)/2

Measures of Central Tendency - Mean

- Mode - the most frequent observation. If there is no repetition, no mode exists.

nums = {872, 432, 432, 432, 388, 782, 388}

Mode = 432

Outlier

■ What is Outlier?

Suppose we have a group of data [1, 3, 5, 5, 5, 7, 29]

What is the mean, medium and mode of this group of data?

Mean = 7.8571

Medium = 5

Mode = 5

What causes the mean value to be so different from the medium and mode?

Variance

- Variance (σ^2) in statistics is a measurement of the spread between numbers in a data set. That is, it measures how far each number in the set is from the mean and therefore from every other number in the set.

$$\text{variance } \sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

where:

x_i = the i^{th} data point

\bar{x} = the mean of all data points

n = the number of data points

Standard Deviation

- Variance (σ^2) in statistics is a measurement of the spread between numbers in a data set. That is, it measures how far each number in the set is from the mean and therefore from every other number in the set.

$$\text{variance } \sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

where:

x_i = the i^{th} data point

\bar{x} = the mean of all data points

n = the number of data points

Standard Deviation

- The standard deviation measures the spread of the data about the mean value. It is useful in comparing sets of data which may have the same mean but a different range.

$$\text{Standard Deviation} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

where:

x_i = Value of the i^{th} point in the data set

\bar{x} = The mean value of the data set

n = The number of data points in the data set

Measurement of Variability

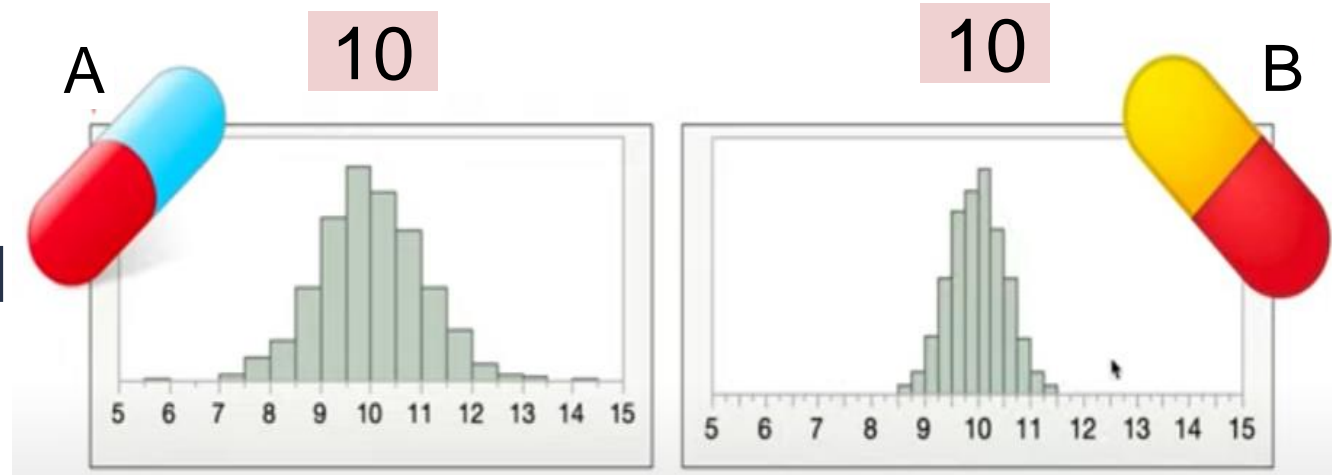
■ Measurement of Variability

While the central tendency, or average, tells you where most of your points lie, variability summarizes how far apart they are. This is important because the amount of variability determines how well you can generalize results from the sample to your population.

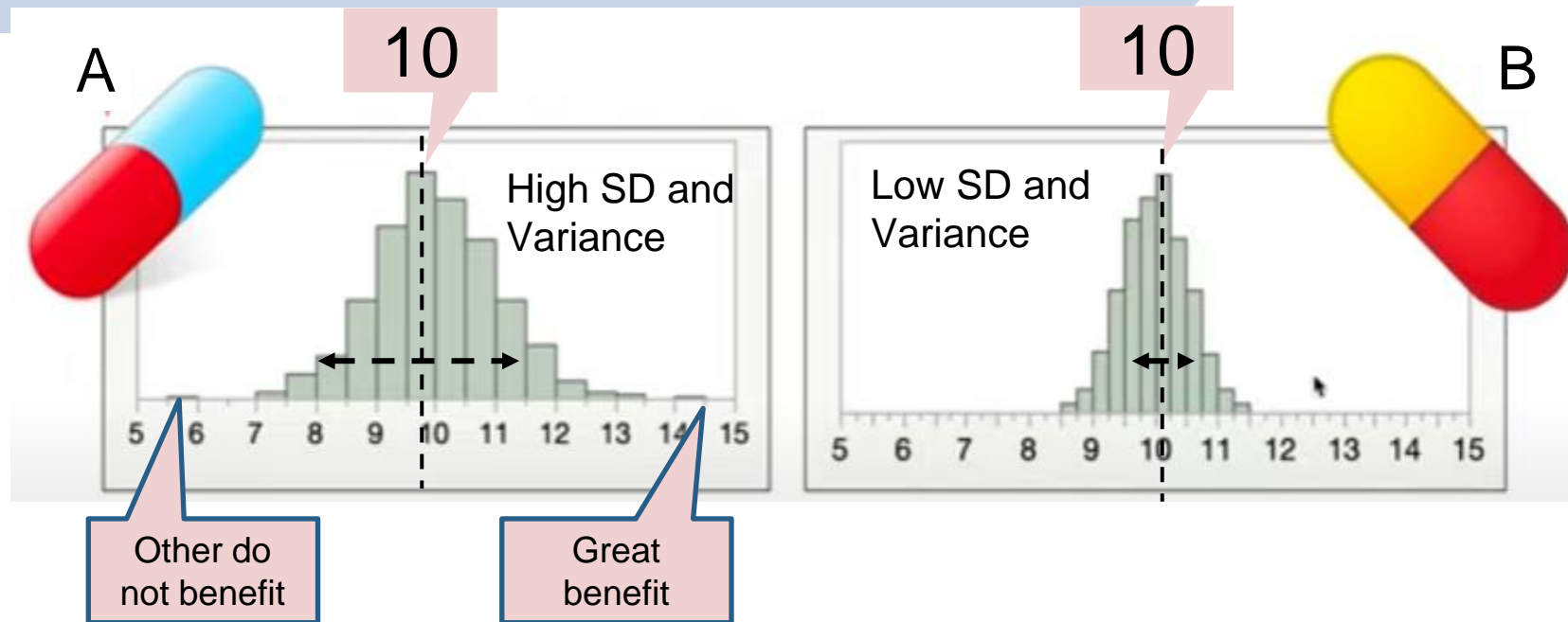
Low variability is ideal because it means that you can better predict information about the population based on sample data. High variability means that the values are less consistent, so it's harder to make predictions.

Measurement of Variability

- Both medicines are tested with the patients. The result show the effective score of the medicines toward the patients.
- 5 with the lowest effectiveness and 15 the highest effectiveness.
- Both medicines have a mean effectiveness score of 10. Which medicine is more effective?



Measurement of Variability

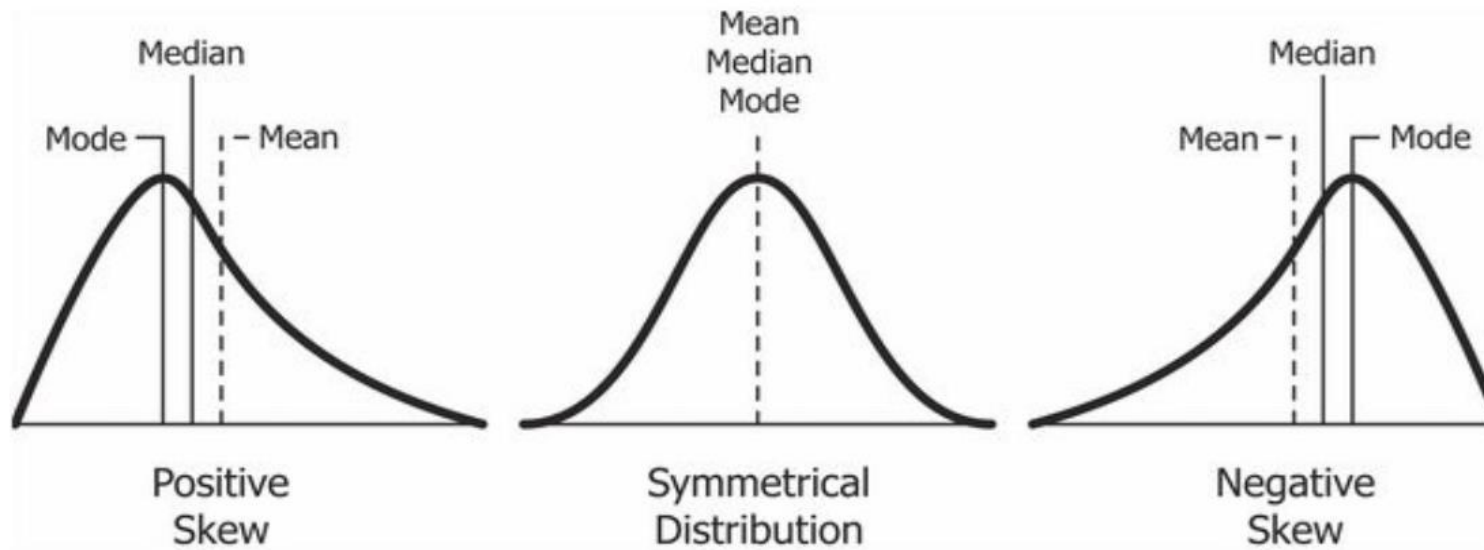


- High variability means that the values are less consistent, so it's harder to make predictions.

- Low variability is ideal because it means that you can better predict information about the population based on sample data.

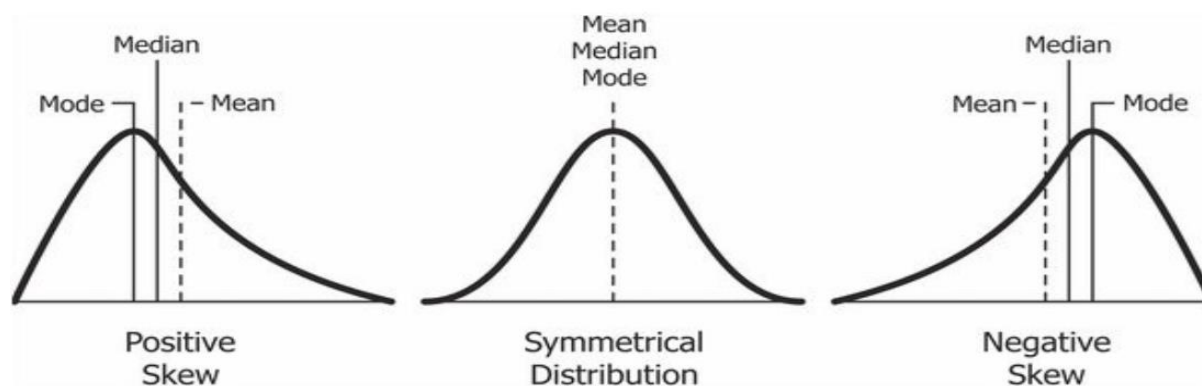
Skewness

- Skewness measures the symmetry of a distribution.
- If $\text{Mode} < \text{Median} < \text{Mean}$ then the distribution is positively skewed.
- If $\text{Mode} > \text{Median} > \text{Mean}$ then the distribution is negatively skewed.



Skewness

- Pearson mode skewness, also called Pearson's first coefficient of skewness, is a way to figure out the skewness of a distribution.
- If you have a distribution and you know the mean, mode, and standard deviation (σ), then the Pearson mode skewness formula is:
- $(\text{mean}-\text{mode})/\sigma$ or $3(\text{mean}-\text{median})/\sigma$



0 means no skewness at all.

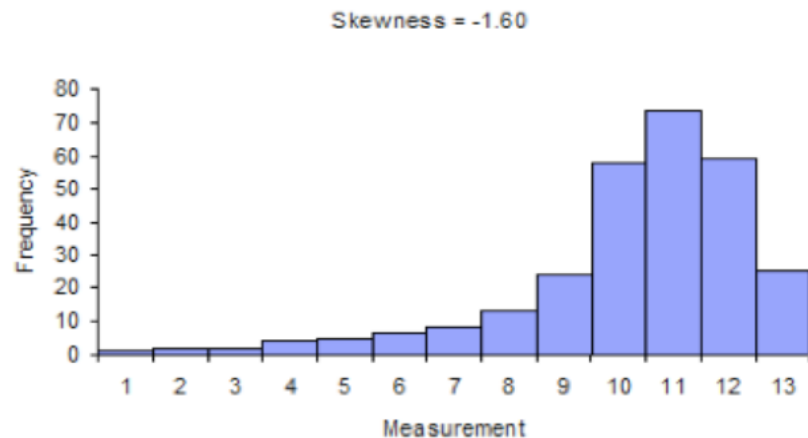
- value means the distribution is negatively skewed.

+value means the distribution is positively skewed.

Skewness

As a general rule of thumb:

1. If skewness is less than -1 or greater than 1, the distribution is highly skewed.
2. If skewness is between -1 and -0.5 or between 0.5 and 1, the distribution is moderately skewed.
3. If skewness is between -0.5 and 0.5, the distribution is approximately symmetric.



Skewness

Pandas: `DataFrame.skew(axis=None, skipna=None, level=None, numeric_only=None, **kwargs)`

- Parameters :
 - `axis` : {index (0), columns (1)}
 - `skipna` : Exclude NA/null values when computing the result.
 - `level` : If the axis is a MultiIndex (hierarchical), count along a particular level, collapsing into a Series
 - `numeric_only` : Include only float, int, boolean columns. If None, will attempt to use everything, then use only numeric data. Not implemented for Series.
 - Return : `skew` : Series or DataFrame (if level specified)
- `axis = 0`: column-wise = along the rows
 - `axis = 1`: row-wise = along the columns

Skewness

```
# importing pandas as pd
import pandas as pd
```

```
# Creating the dataframe
df = pd.read_csv("nba.csv")
```

```
# skewness along the index axis
df.skew(axis = 0, skipna = True)
```

```
Number    1.668386
Age        0.626349
Weight     0.113788
Salary     1.576321
dtype: float64
```

	Name	Team	Number	Position	Age	Height	Weight	College	Salary
0	Avery Bradley	Boston Celtics	0.0	PG	25.0	6-2	180.0	Texas	7730337.0
1	Jae Crowder	Boston Celtics	99.0	SF	25.0	6-6	235.0	Marquette	6796117.0
2	John Holland	Boston Celtics	30.0	SG	27.0	6-5	205.0	Boston University	NaN
3	R.J. Hunter	Boston Celtics	28.0	SG	22.0	6-5	185.0	Georgia State	1148640.0
4	Jonas Jerebko	Boston Celtics	8.0	PF	29.0	6-10	231.0	NaN	5000000.0
5	Amir Johnson	Boston Celtics	90.0	PF	29.0	6-9	240.0	NaN	12000000.0
6	Jordan Mickey	Boston Celtics	55.0	PF	21.0	6-8	235.0	LSU	1170960.0
7	Kelly Olynyk	Boston Celtics	41.0	C	25.0	7-0	238.0	Gonzaga	2165160.0
8	Terry Rozier	Boston Celtics	12.0	PG	22.0	6-2	190.0	Louisville	1824360.0

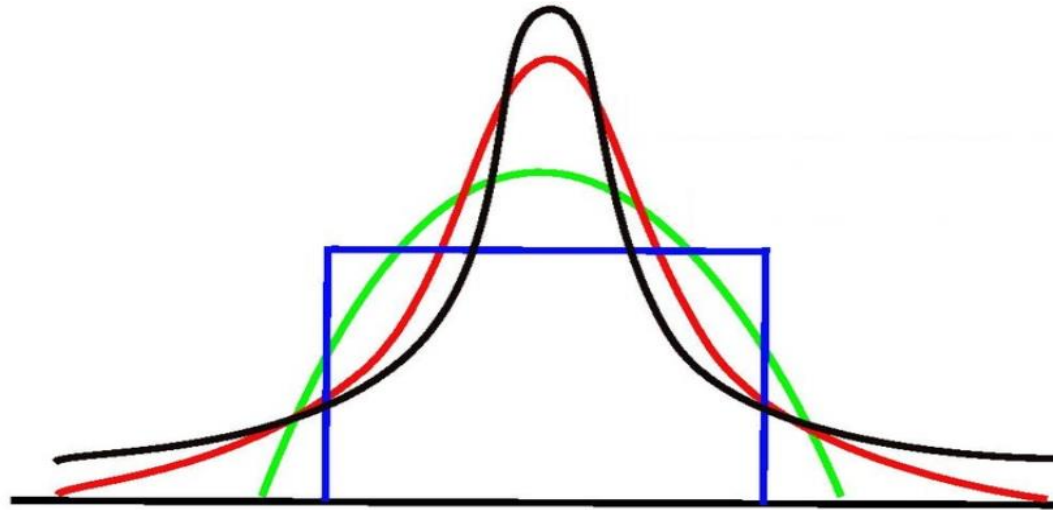
Skewness

Why is Skewness important?

- The linear models assume that the distribution of the independent variable and the dependent(target) variable are similar. Thus, knowing the Skewness of data helps us in creating better models.
- Suppose we have *positively skewed distributed data*. So that means it has a higher number of data points having low values. So during the model training on this type of data, it will perform better at predicting lower values than those with higher values.
- Skewness helps us know the direction of outliers. In the case of a positively skewed distribution, most outliers are present on the right side of the distribution. In contrast, most outliers are present on the left side of the distribution in the case of negatively skewed data. Skewness does not tell us about the frequency of outliers. It just tells us the direction.

Kurtosis

- Kurtosis is all about the tails of the distribution – not the peakedness or flatness. It is used to describe the extreme values in one versus the other tail. It is actually the measure of outliers present in the distribution.



Kurtosis

■ Here, \bar{x} is the sample mean

$$k_1 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^2}.$$

K1 without bias correction

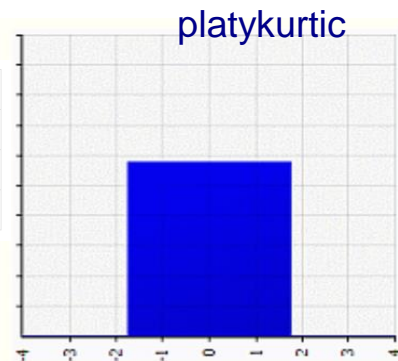
K0 with bias correction

$$k_0 = \frac{n-1}{(n-2)(n-3)} ((n+1)k_1 - 3(n-1)) + 3.$$

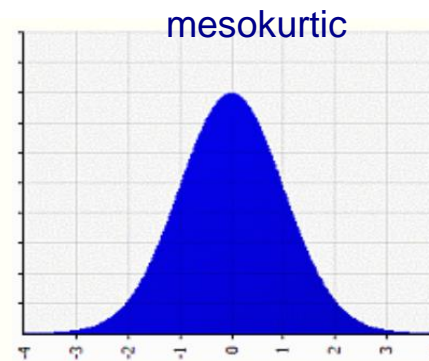
Kurtosis

- A distribution with kurtosis <3 (excess kurtosis <0) is called platykurtic. Compared to a normal distribution, its tails are shorter and thinner, and often its central peak is lower and broader.
- A normal distribution has kurtosis exactly 3 (excess kurtosis exactly 0). Any distribution with kurtosis ≈ 3 (excess ≈ 0) is called mesokurtic.
- A distribution with kurtosis >3 (excess kurtosis >0) is called leptokurtic. Compared to a normal distribution, its tails are longer and fatter, and often its central peak is higher and sharper.

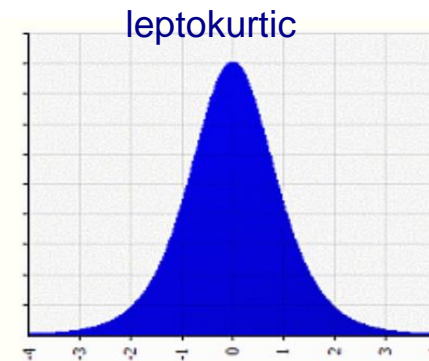
	Kurtosis	Excess Kurtosis
Leptokurtic	>3	>0
Platykurtic	<3	<0
Mesokurtic	$=3$	$=0$



Uniform($\min=-\sqrt{3}$, $\max=\sqrt{3}$)
kurtosis = 1.8, excess = -1.2



Normal($\mu=0$, $\sigma=1$)
kurtosis = 3, excess = 0



Logistic($\alpha=0$, $\beta=0.55153$)
kurtosis = 4.2, excess = 1.2

Kurtosis

```
Series.kurtosis(axis=None, skipna=None, level=None,  
numeric_only=None, **kwargs)
```

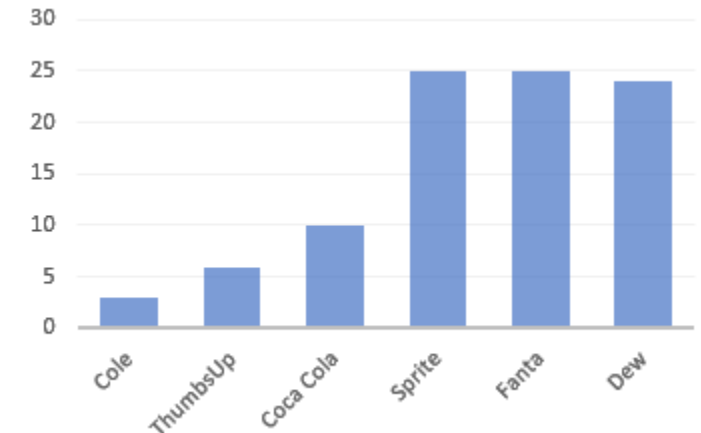
```
import pandas as pd # importing pandas as pd
```

```
sr = pd.Series([10, 25, 3, 25, 24, 6]) # Creating the Series
```

```
index_ = ['Coca Cola', 'Sprite', 'Coke', 'Fanta', 'Dew', 'ThumbsUp'] #  
Create the Index
```

```
sr.index = index_ # set the index
```

```
result = sr.kurtosis() # return the kurtosis
```



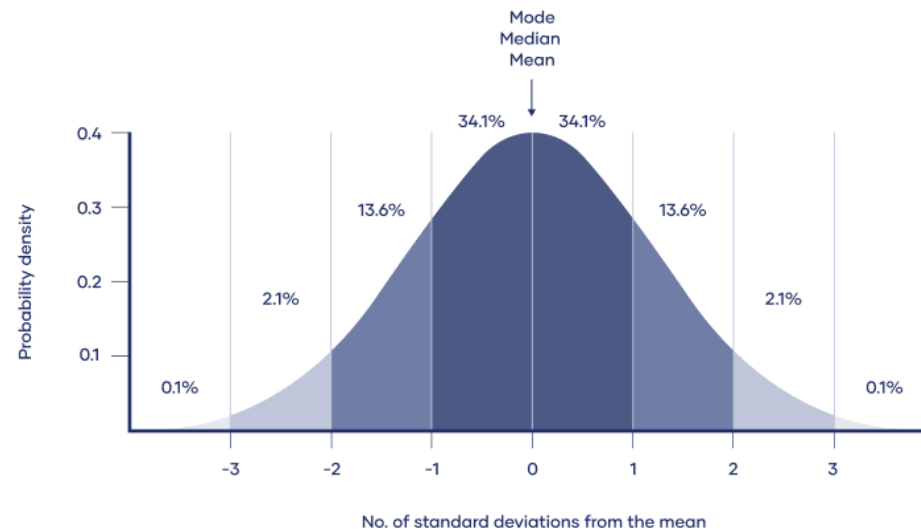
```
Coca Cola    10  
Sprite       25  
Coke         3    -2.818014750138433  
Fanta        25  
Dew          24  
ThumbsUp     6  
dtype: int64
```


Skewness and Kurtosis

How do skewness and kurtosis affect the normality of data?

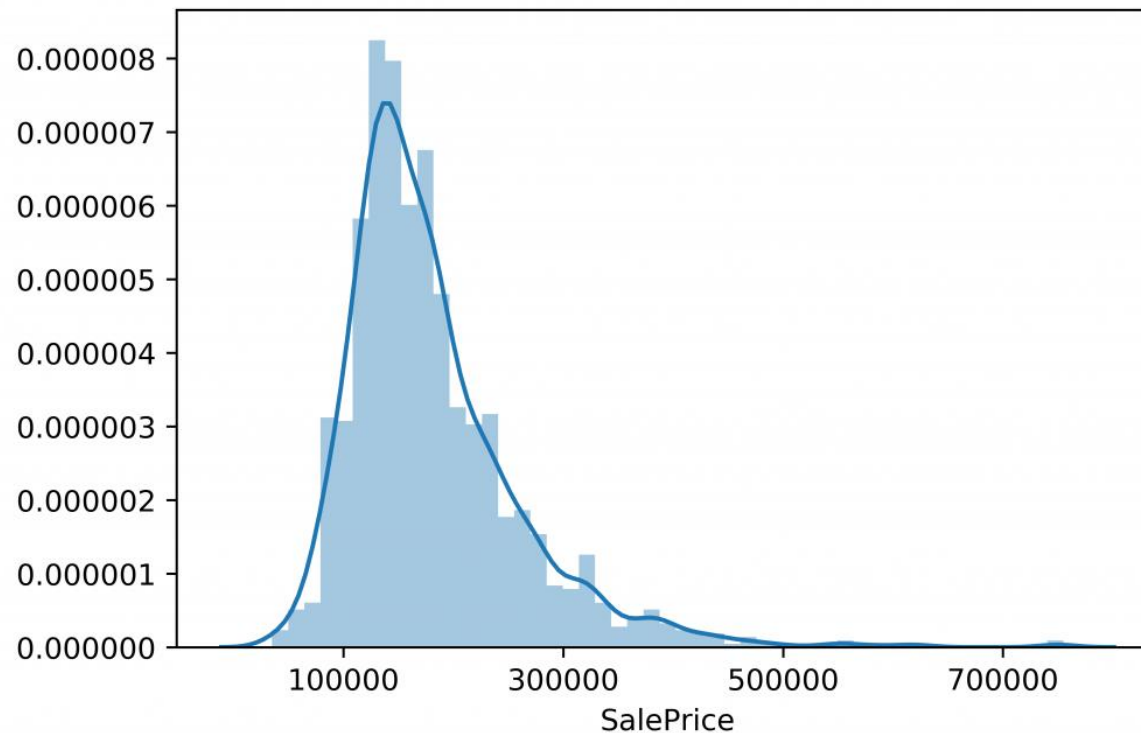
Statistically, two numerical measures of shape – Skewness and excess kurtosis can be used to test for normality. Normality tests determine whether a data set is designed for normal distribution.

Standard normal distribution



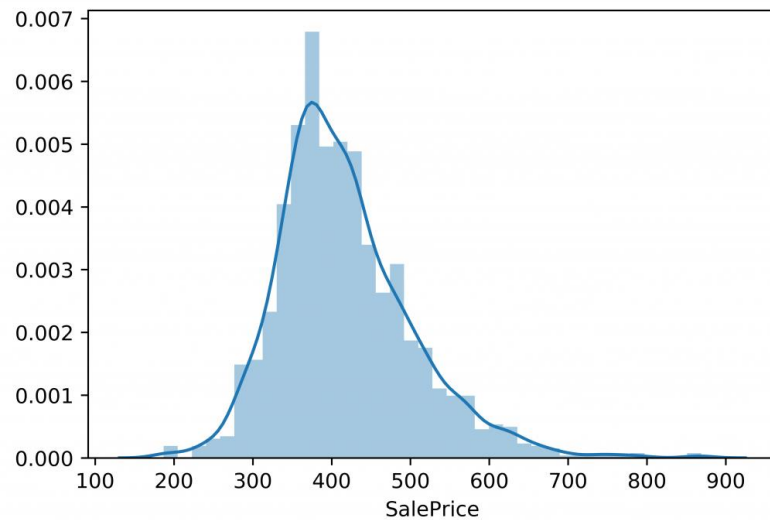
Transform Skewed Data

Skewed data is common in data science; skew is the degree of distortion from a normal distribution. For example, below is a plot of the house prices that is right skewed, meaning there are a minority of very large values.

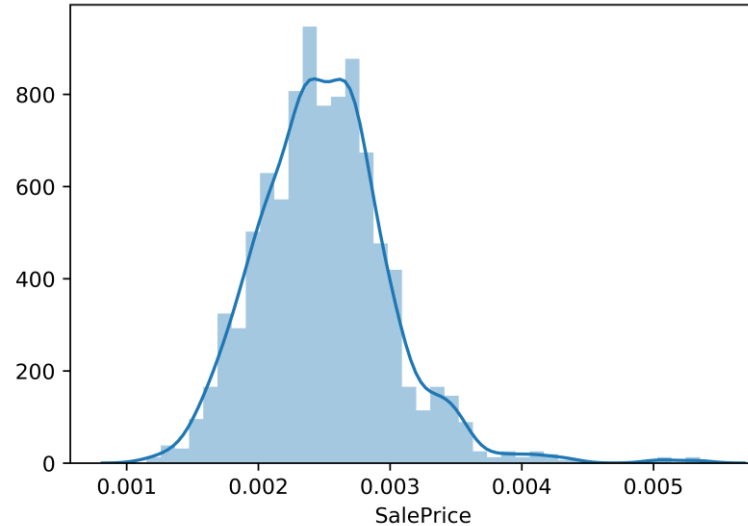


Transform Skewed Data

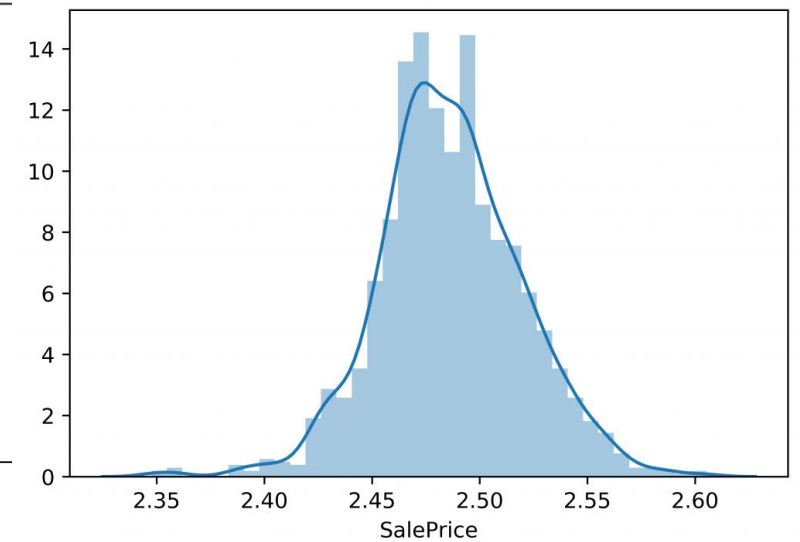
Square Root Transformation



Reciprocal Transformation



Log Transformation



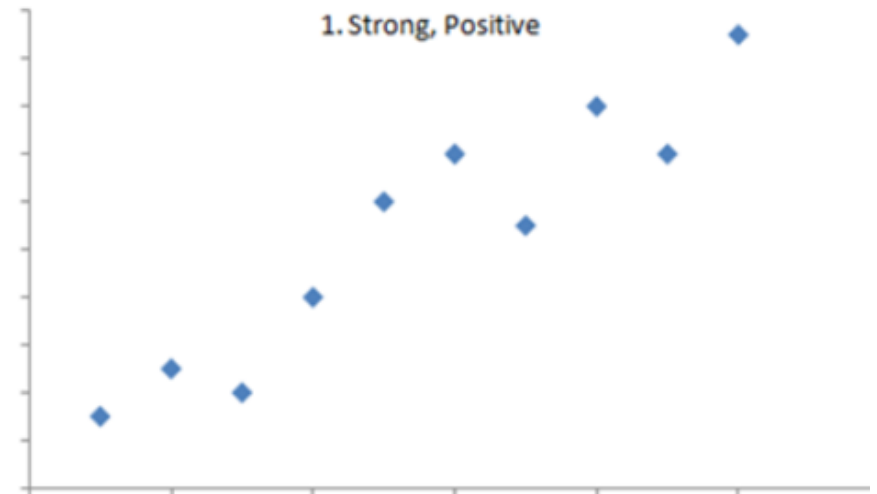
Correlation Analysis

- In correlation analysis, we estimate a sample correlation coefficient, more specifically the Pearson Product Moment correlation coefficient.
- The sample correlation coefficient, denoted r , ranges between -1 and +1 and quantifies the direction and strength of the linear association between the two variables.
- The correlation between two variables can be positive (i.e., higher levels of one variable are associated with higher levels of the other) or negative (i.e., higher levels of one variable are associated with lower levels of the other).
- The sign of the correlation coefficient indicates the direction of the association. The magnitude of the correlation coefficient indicates the strength of the association.

Correlation Analysis

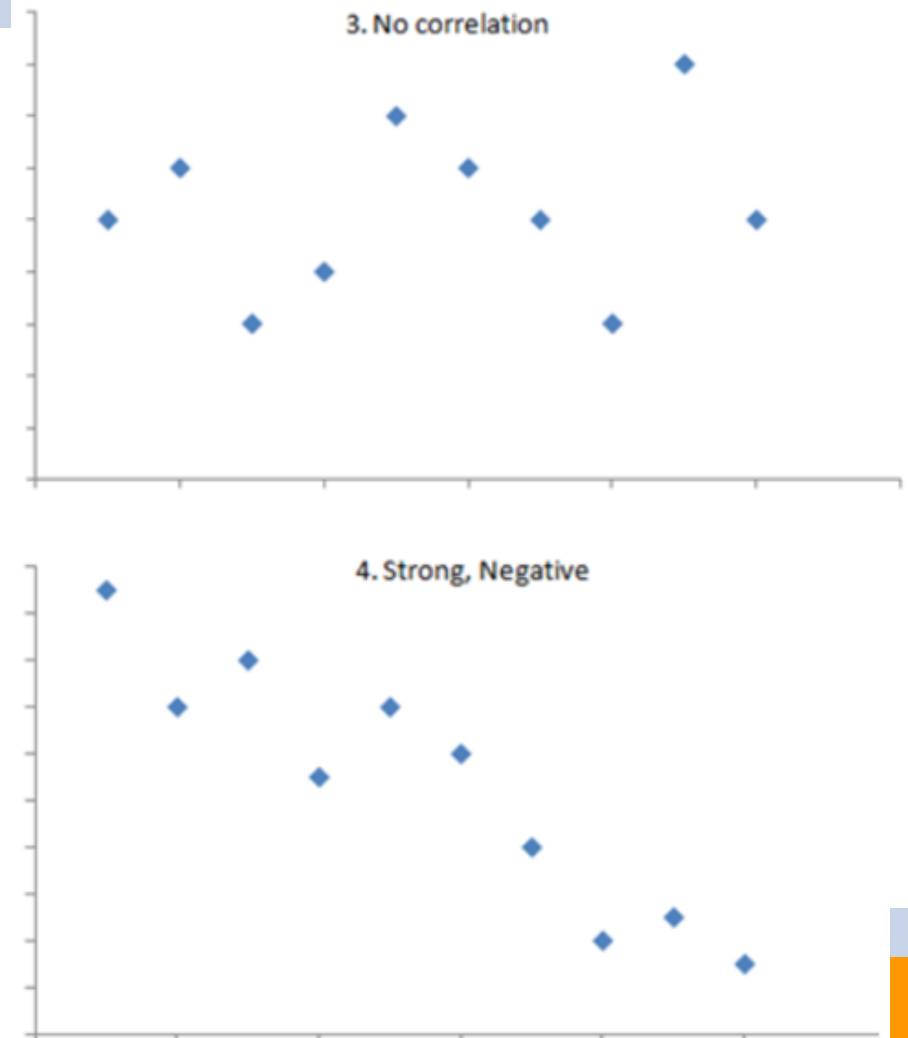
The figure below shows four hypothetical scenarios in which one continuous variable is plotted along the X-axis and the other along the Y-axis.

- Scenario 1 depicts a strong positive association ($r=0.9$), similar to what we might see for the correlation between infant birth weight and birth length.
- Scenario 2 depicts a weaker association ($r=0.2$) that we might expect to see between age and body mass index (which tends to increase with age).



Correlation Analysis

- Scenario 3 might depicts the lack of association (r approximately 0) between the extent of media exposure in adolescence and age at which adolescents initiate sexual activity.
- Scenario 4 might depicts the strong negative association ($r = -0.9$) generally observed between the number of hours of aerobic exercise per week and percent body fat.



Correlation Analysis

- Python is a great language for doing data analysis, primarily because of the fantastic ecosystem of data-centric python packages. *Pandas* is one of those packages and makes importing and analyzing data much easier.
- Pandas `dataframe.corr()` is used to find the pairwise correlation of all columns in the dataframe. Any null values are automatically excluded. For any non-numeric data type columns in the dataframe it is ignored.

Correlation Analysis

To find the correlation among
the columns using pearson method
df.corr(method='pearson')

	Name	Team	Number	Position	Age	Height	Weight	College	Salary
0	Avery Bradley	Boston Celtics	0.0	PG	25.0	6-2	180.0	Texas	7730337.0
1	Jae Crowder	Boston Celtics	99.0	SF	25.0	6-6	235.0	Marquette	6796117.0
2	John Holland	Boston Celtics	30.0	SG	27.0	6-5	205.0	Boston University	NaN
3	R.J. Hunter	Boston Celtics	28.0	SG	22.0	6-5	185.0	Georgia State	1148640.0
4	Jonas Jerebko	Boston Celtics	8.0	PF	29.0	6-10	231.0	NaN	5000000.0
5	Amir Johnson	Boston Celtics	90.0	PF	29.0	6-9	240.0	NaN	12000000.0
6	Jordan Mickey	Boston Celtics	55.0	PF	21.0	6-8	235.0	LSU	1170960.0
7	Kelly Olynyk	Boston Celtics	41.0	C	25.0	7-0	238.0	Gonzaga	2165160.0
8	Terry Rozier	Boston Celtics	12.0	PG	22.0	6-2	190.0	Louisville	1824360.0
9	Marcus Smart	Boston Celtics	36.0	PG	22.0	6-4	220.0	Oklahoma State	3431040.0

	Number	Age	Weight	Salary
Number	1.000000	0.028724	0.206921	-0.112386
Age	0.028724	1.000000	0.087183	0.213459
Weight	0.206921	0.087183	1.000000	0.138321
Salary	-0.112386	0.213459	0.138321	1.000000

Statistics Basics

What is Inferential Statistic?
Types of Hypothesis Test
Features Engineering

Inferential Statistics

- Inferential statistics can be defined as **a field of statistics that uses analytical tools for drawing conclusions about a population by examining random samples**. The goal of inferential statistics is to make generalizations about a population.
- Hypothesis testing
- Regression

Frequency vs Probability Distribution

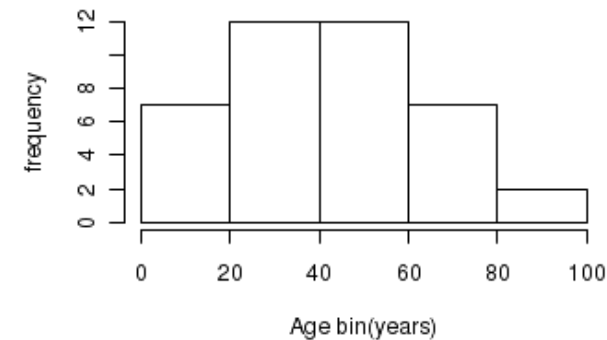
- Frequency distribution is related to probability distribution.
- While a frequency distribution gives the exact frequency or the number of times a data point occurs,
- a probability distribution gives the probability of occurrence of the given data point.
- When the number of test cases are large, the frequency distribution and the probability distributions are similar in shape.

Frequency vs Probability Distribution

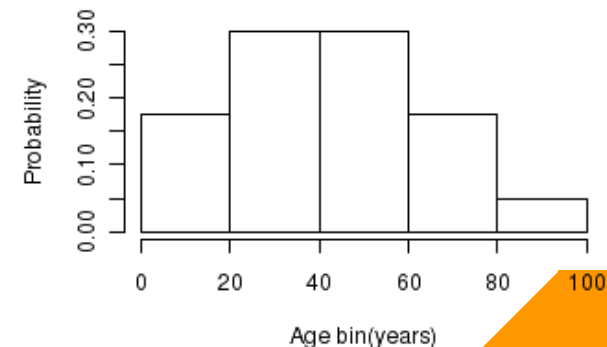
- As an example, let us look at the following imaginary data set on the age(in years) of the male patients visiting a clinic over a period of three days:

age bin (years)	frequency	empirical probability
0-10	3	0.075
10-20	4	0.1
20-30	5	0.125
30-40	6	0.15
40-50	7	0.175
50-60	6	0.15
60-70	4	0.1
70-80	3	0.075
80-90	2	0.05
	sum = 40	

Frequency histogram

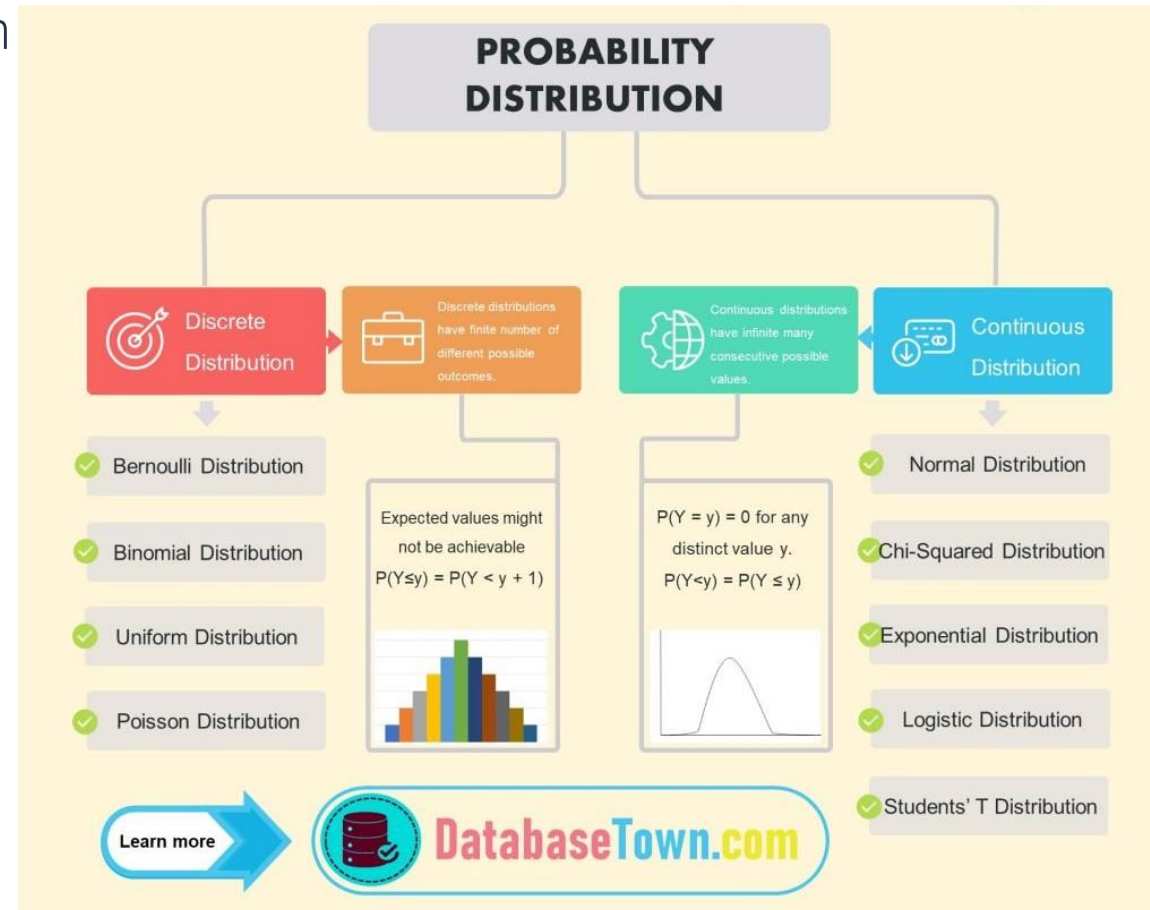


Empirical probability histogram



Type of Probability Distribution

■ Probability Distribution



Source" DatabaseTowm.com

Hypothesis testing

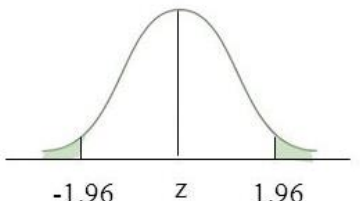
What is hypothesis testing ?

- Hypothesis testing is a Inferential statistical method that is used in making statistical decisions using experimental data. Hypothesis Testing is basically an assumption that we make about the population parameter.
- Example : you say average student in class is 40 or a boy is taller than girls.
- All those example we assume need some statistic way to prove those. We need some mathematical conclusion whatever we are assuming is true.

Hypothesis testing

Step 1: $H_0: \mu_{\text{GRE after Kaplan Course}} = 500$ ($\mu = 500$ g, $\sigma = 100$ g)
 (There is no effect of the Kaplan training course on average GRE scores)
 $H_1: \mu_{\text{GRE after Kaplan Course}} \neq 500$
 (There is an effect...)

Step 2: Set criteria

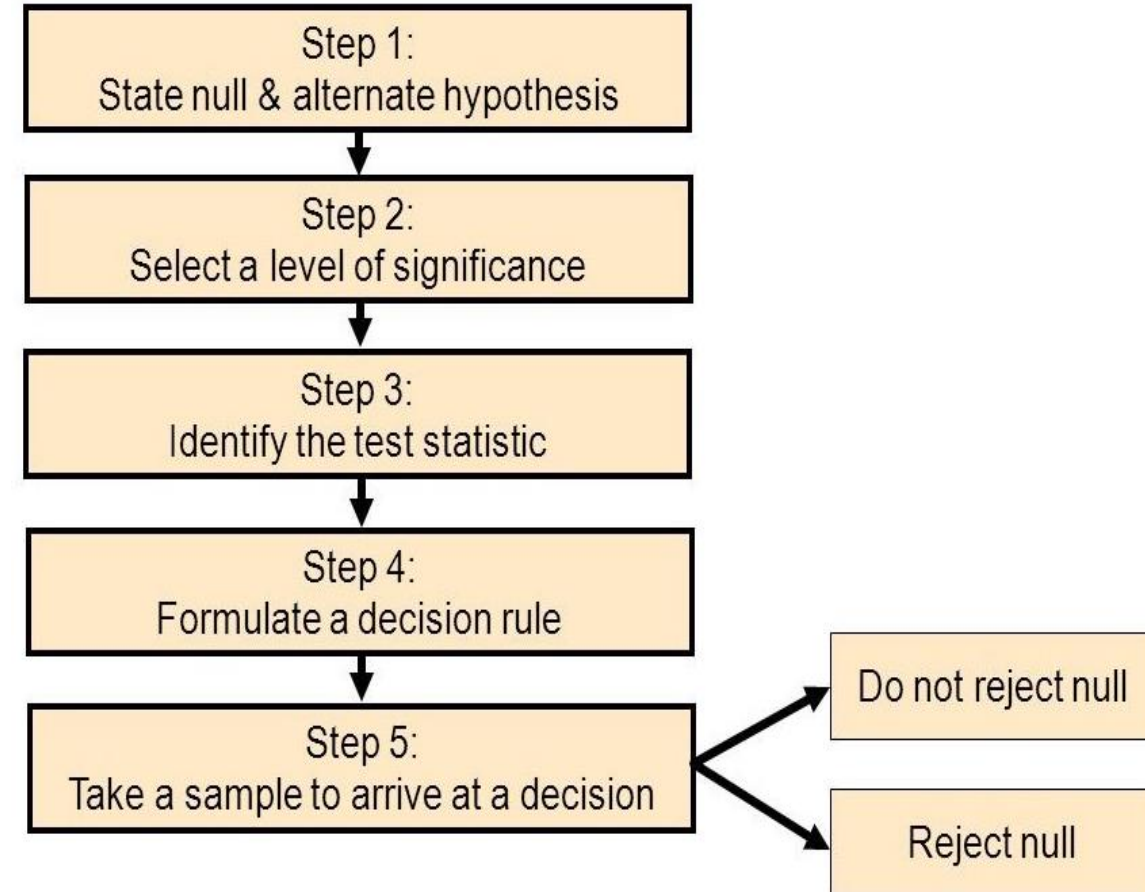


Critical Region
 $z > 1.96$
 or
 $z < -1.96$
 $\alpha = 0.05$

Step 3: $n = 100$, $\bar{X} = 525$, $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{100}{\sqrt{100}} = \frac{100}{10} = 10$
 $Z_{\text{obt}} = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{525 - 500}{10} = \frac{25}{10} = 2.5$

Step 4: Reject H_0 because Z_{obt} of 2.5 is in the critical region.

Step 5: Conclusion. The Kaplan training course significantly increased GRE scores on average, $z = 2.5$, $p < .05$.



Hypothesis testing - Step 1

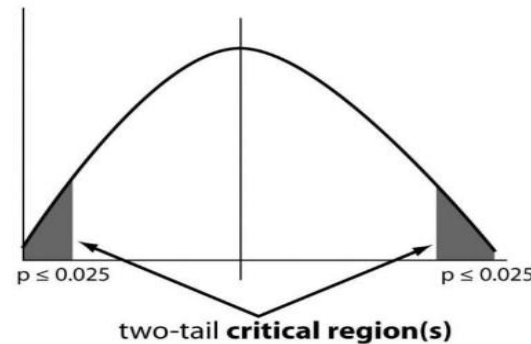
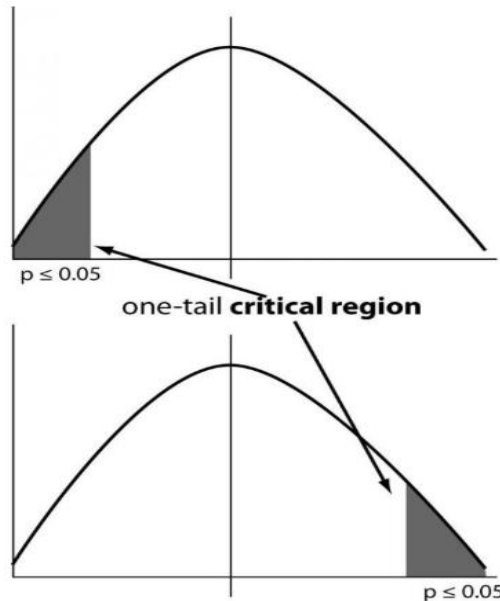
We need to define two opposing statements,

- Null hypothesis (same, not related, not effective, etc).
- Alternative hypothesis (different, related, effective, etc)

Hypothesis testing – Step 2

We need to define a critical value α . (significant level)

- A significance level denotes how much chance we accept, (ranging from 0 to 1)



Hypothesis testing – Step 3

some of widely used hypothesis testing type :

- Test (Student T test)
- Z Test
- ANOVA Test

If the p-value is less than what is tested at, most commonly 0.05, one can reject the null hypothesis.

Hypothesis testing(T-Test)

- A t-test is a type of inferential statistic which is used to determine if there is a significant difference between the means of two groups which may be related in certain features.
- It is mostly used when the data sets, like the set of data recorded as outcome from flipping a coin a 100 times, would follow a normal distribution and may have unknown variances.
- T test is used as a hypothesis testing tool, which allows testing of an assumption applicable to a population.
- T-test has 2 types :
 - ▶ one sampled t-test
 - ▶ two-sampled t-test.

Hypothesis testing(One sampled T-Test)

One sampled t-test

- The One Sample t Test determines whether the sample mean is statistically different from a known or hypothesised population mean. The One Sample t Test is a parametric test.
- Example :- you have 10 ages and you are checking whether avg age is 30 or not.

$$t = \frac{(\bar{x} - \mu_0)}{\frac{s}{\sqrt{n}}}$$
$$df = n - 1$$

—————→ P-value

Hypothesis testing (One Sampled T-Test)

One sample t-test :

This test makes the following assumptions:

- Independence: The observations in the sample should be independent.
- Random Sampling: The observations should be collected using a random sampling method to maximize the chances that the sample is representative of the population of interest.
- Normality: The observations should be roughly normally distributed.

Hypothesis testing(Two Sampled T-Test)

Two sampled T-test

- The Independent Samples t Test or 2-sample t-test compares the means of two independent groups in order to determine whether there is statistical evidence that the associated population means are significantly different.
- The Independent Samples t Test is a parametric test. This test is also known as: Independent t Test.
- Example : is there any association between week1 and week2

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1} \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1} \left(\frac{s_2^2}{n_2}\right)^2}$$

Hypothesis testing(Two Sampled T-Test)

Two sampled T-test :

This test makes the following assumptions:

- Independence: The observations in each sample should be independent.
- Random Sampling: The observations in each sample should be collected using a random sampling method.
- Normality: Each sample should be roughly normally distributed.
- Equal Variance: Each sample should have approximately the same variance.

Hypothesis testing(Paired Sampled T-Test)

Paired sampled t-test

- The paired sample t-test is also called dependent sample t-test. It's an univariate test that tests for a significant difference between 2 related variables. An example of this is if you were to collect the blood pressure for an individual before and after some treatment, condition, or time point.

H0 :- means difference between two sample is 0

H1:- mean difference between two sample is not 0

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$
$$df = n_1 + n_2 - 2$$

Hypothesis testing(Z-Test)

You would use a Z test if:

- Your sample size is greater than 30. Otherwise, use a t test.
- Data points should be independent from each other. In other words, one data point isn't related or doesn't affect another data point.
- Your data should be normally distributed. However, for large sample sizes (over 30) this doesn't always matter.
- Your data should be randomly selected from a population, where each item has an equal chance of being selected.
- Sample sizes should be equal if at all possible.

Hypothesis testing(Z-Test)

Z test

$z = (x - \mu) / (\sigma / \sqrt{n})$, where

x = sample mean

μ = population mean

σ / \sqrt{n} = population standard deviation

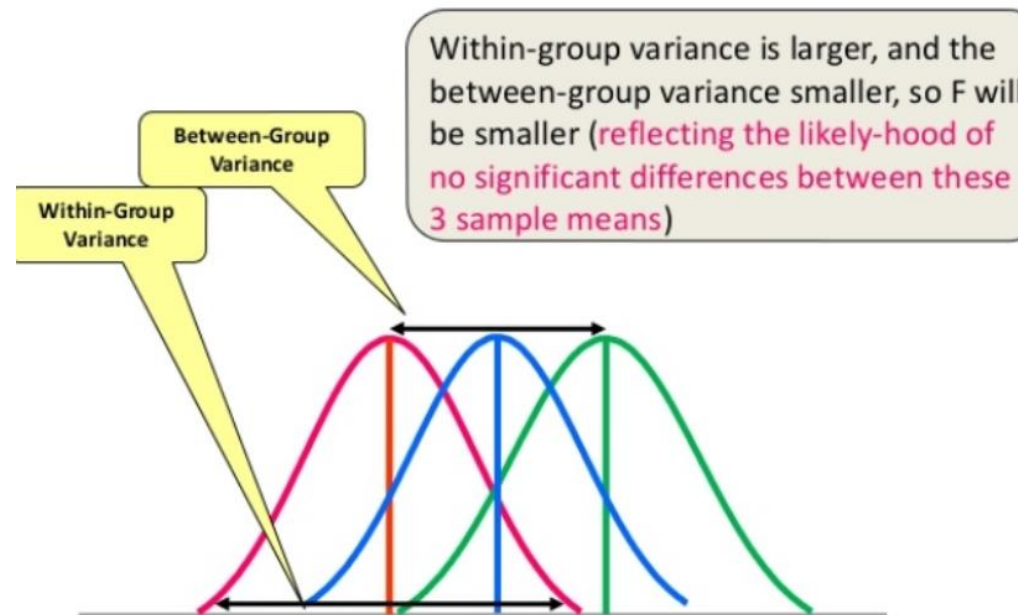
Hypothesis testing(F-Test)

ANOVA (F-TEST)

- The t-test works well when dealing with two groups, but sometimes we want to compare more than two groups at the same time.
- For example, if we wanted to test whether voter age differs based on some categorical variable like race, we have to compare the means of each level or group the variable.
- We could carry out a separate t-test for each pair of groups, but when you conduct many tests you increase the chances of false positives.
- The analysis of variance or ANOVA is a statistical inference test that lets you compare multiple groups at the same time.

Hypothesis testing(F-Test)

- $F = \text{Between group variability} / \text{Within group variability}$
- Unlike the z and t-distributions, the F-distribution does not have any negative values because between and within-group variability are always positive due to squaring each deviation.



Hypothesis testing(F-Test)

One Way F-test(Anova)

- It tell whether two or more groups are similar or not based on their mean similarity and f-score.
- Example : there are 3 different category of plant with their weight. We need to check whether all 3 groups are similar or not

Hypothesis testing(Chi square-Test)

Chi Square Test

- The Chi-square test is a statistical test used to determine the relationship between the categorical variables/columns in the dataset. It examines the correlation between the variables which do not contain the continuous data.
- Example :

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

χ^2 = chi-square

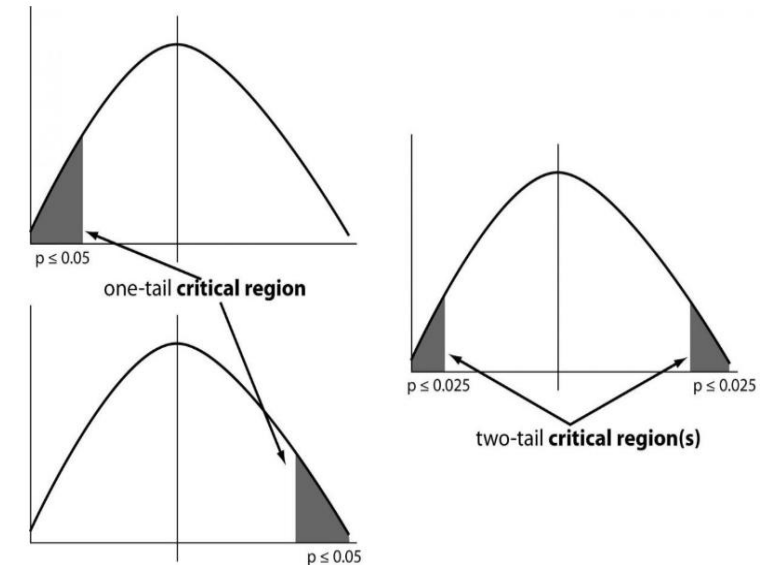
O_i = observed value

E_i = expected value

Hypothesis testing – Step 4/5

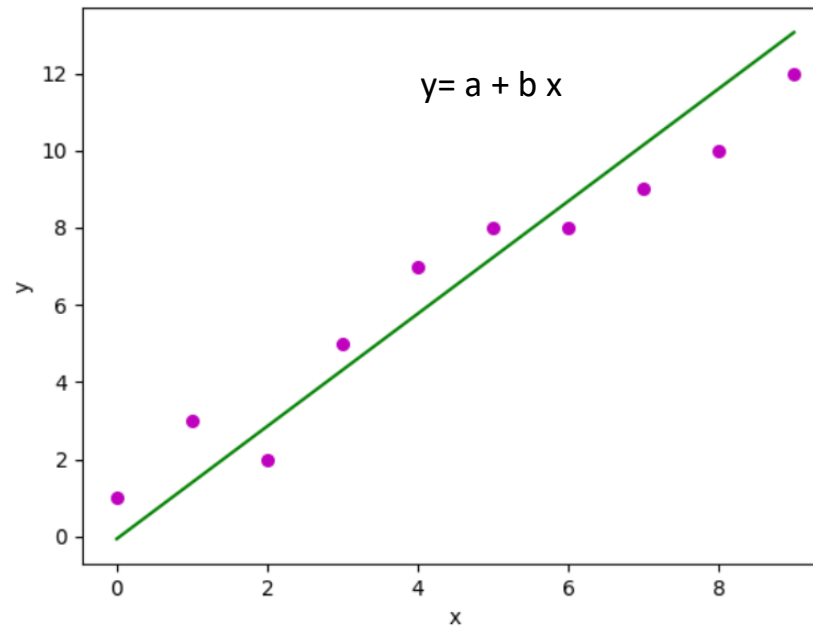
Formulate a decision rule

- We compare the test result (p-value) against the α to decide whether we should accept or reject the null hypothesis. When $p\text{-value} \leq \alpha$ we reject the null hypothesis.
- Typically for most of the project α would be 0.05 (i.e. 5% chances).
- For mission critical systems α should be much lower



Linear Regression

- Hypothesis test on Linear regression is a statistical approach for modelling relationship between a dependent variable with a given set of independent variables.



Linear Regression-T Test

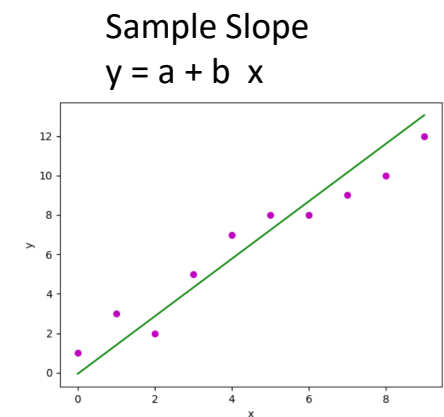
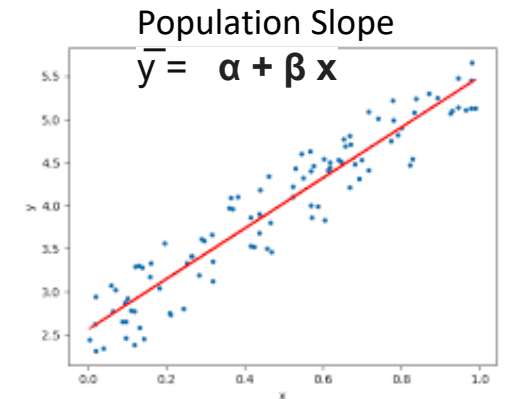
If there is a significant linear relationship between the independent variable X and the dependent variable Y , the slope will *not* equal zero. $H_0: \beta = 0$ $H_a: \beta \neq 0$

The null hypothesis states that the slope is equal to zero, and the alternative hypothesis states that the slope is not equal to zero.

Test statistic. The test statistic is a t statistic (t) defined by the following equation.

$$t = b / SE$$

where b_1 is the slope of the sample regression line, and SE is the standard error of the slope.

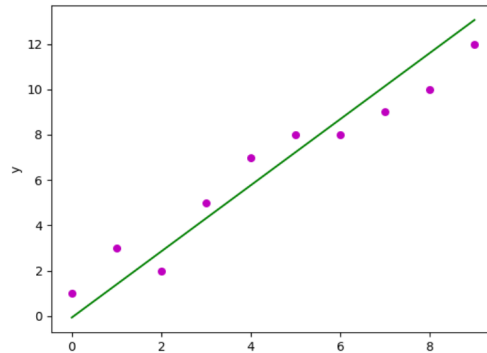


Predictor	Coef	SE Coef	T	P
X	35	20	1.75	0.04
	(b)	(SE)		

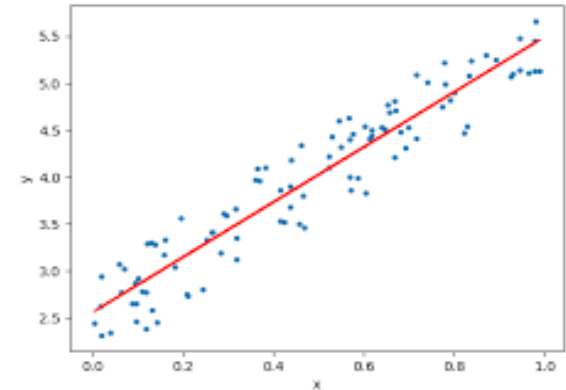
Linear Regression-Confidence interval

Introduction to sample slopes and using them to make confidence intervals or do a test about the population slope in least-squares regression

Sample slope
 $y = a + b x$



Population Slope
 $\bar{y} = \alpha + \beta x$



Here's the formula for a t interval estimating slope:

$$(\text{statistic}) \pm \left(\begin{array}{c} \text{critical} \\ \text{value} \end{array} \right) \left(\begin{array}{c} \text{standard deviation} \\ \text{of statistic} \end{array} \right)$$

$$b \pm t_{n-2}^* (SE_b)$$

Linear Regression

Cars lose value the further they are driven. Heidi collected data about the mileage (in thousands of kilometres) and value (in thousands of dollars) of a random sample of 11 cars of the same make and model. Here is computer output from a least-squares regression analysis on her sample:

Predictor	Coef	SE Coef	T	P
Constant	39.575	0.765	51.77	0.00
Mileage	-0.246	0.013	-18.87	0.00
S = 1.349 R-sq = 97.26%				

What is a 99%, percent confidence interval for the slope of the least squares regression line?

Linear Regression

What is a 99%, percent confidence interval for the slope of the least squares regression line?

Compute alpha (α):

$$\alpha = 1 - (\text{confidence level} / 100)$$

$$\alpha = 1 - 99/100 = 0.01$$

find the critical probability (p^*):

$$p^* = 1 - \alpha/2 = 1 - 0.01/2 = 0.995$$

Find the degrees of freedom (df):

$$df = n - 2 = 11 - 2 = 9.$$

Here's the formula for a t interval estimating slope:

$$(\text{statistic}) \pm \left(\begin{array}{c} \text{critical} \\ \text{value} \end{array} \right) \left(\begin{array}{c} \text{standard deviation} \\ \text{of statistic} \end{array} \right)$$

$$b \pm t_{n-2}^* (SE_b)$$

$$-0.246 \pm 3.25(0.013)$$

<https://www.socscistatistics.com/pvalues/tdistribution.aspx>

Linear Regression

T Table (Two Tail)

P value

Degree of
freedom

T value

DF	A = 0.2	0.1	0.05	0.02	0.01	0.002	0.001
∞	ta = 1.282	1.645	1.96	2.326	2.576	3.091	3.291
1	3.078	6.314	12.706	31.821	63.656	318.289	636.578
2	1.886	2.92	4.303	6.965	9.925	22.328	31.6
3	1.638	2.353	3.182	4.541	5.841	10.214	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.61
5	1.476	2.015	2.571	3.365	4.032	5.894	6.869
6	1.44	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.86	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.25	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.93	4.318
13	1.35	1.771	2.16	2.65	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.14
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.12	2.583	2.921	3.686	4.015
17	1.333	1.74	2.11	2.567	2.898	3.646	3.965
18	1.33	1.734	2.101	2.552	2.878	3.61	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883

Feature Engineering

- Feature Engineering is the way of extracting features from data and transforming them into formats that are suitable for analysis
- It is divided into 3 broad categories:
 - ▷ Feature Selection
 - ▷ Feature Transformation
 - ▷ Feature Extraction

Feature Engineering

Feature Selection

- We select those attributes which best explain the relationship of an independent variable with the target variable. There are certain features which are more important than other features to the accuracy of the model.
- The methods of Feature Selection are Hypothesis test, correlation, Least Squares regression , etc.

Feature Engineering

Feature Transformation:

- It means transforming our original feature to the functions of original features. Scaling, discretization, binning and filling missing data values are the most common forms of data transformation.
- Example to reduce right skewness of the data, we use log transform.

Feature Engineering

Feature Extraction

- When the data to be processed through an algorithm is too large, it's generally considered redundant. Analysis with a large number of variables uses a lot of computation power and memory, therefore we should reduce the dimensionality of these types of variables. It is a term for constructing combinations of the variables.
- For tabular data, we use PCA to reduce features. For image, we can use line or edge detection. Use correlation to extract variables with strong influence of the output.

Normalization

- Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.
- Here's the formula for normalization:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Normalization

- For example, consider a data set containing two features, age(x_1), and income(x_2).
- Where age ranges from 0–100, while income ranges from 0–20,000 and higher. Income is about 1,000 times larger than age and ranges from 20,000–500,000. So, these two features are in very different ranges.
- When we do further analysis, like multivariate linear regression, for example, the attributed income will intrinsically influence the result more due to its larger value. But this doesn't necessarily mean it is more important as a predictor.

Summary

Apply different methods in the Descriptive Statistic

Apply different methods in the Inferential Statistic