Data Science Cheat Sheet

Diogo Centeno

This document serves as a way to record things that I thought important blablabla….

Contents

Combinatorics 5
Combination symmetry 5
Bayesian Inference 5
Exclusivity 6
Dependency 6
Conditional probability 6
Law of total probability 6
Additive law 6
Multiplication rule 7
Bayes' Law 7
Distributions 8
Discrete vs Continuous distributions 8
Discrete distributions 9
Uniform distribution 9
Bernoulli distribution 10
Binomial distribution 10
Poisson distribution 11
Continuous distributions 12
Normal distribution 12
Students' T 13
Chi-Squared 13
Exponential Distribution 14
Logistic distribution 15
Descriptive Statistics 16
Types of data 16
Levels of measurement 16
Visual representations for categorical data 16
Visual representation for numerical data 19

Freedman-Diaconis rule 19
Relations between variables 20
Population vs Sample 21
Difference in symbols 22
Mean, median and mode 22
Skewness 23
Variance and standard deviation 24
Covariance 24
Correlation 25
Inferential statistics 26
The Central Limit theorem 26
Estimators and estimates 26
Estimator 26
Estimates 27
Confidence Intervals and Margin of Error 27
Hypothesis testing 30
Decisions 30
Level of significance and type of test 30
Statistical errors 31
P-value 31
Formulae 32
Linear regression 33
Linear regression model 33
Linear regression equation 33
Geometrical representation 34
Correlation vs Regression 34
Regression methods 34
OLS assumptions 35
Logistic Regression 35
Understanding the model values 36

Cluster analysis..... 37
K-means clustering..... 37
Euclidian distance 38
Pros and Cons of K-means 39

Combinatorics

Order matters?	Elements can repeat?	Formula
Yes	Yes	n^r
Yes	No	$\frac{n!}{(n-r)!}$
No	No	$\frac{n!}{r!\,(n-r)!}$
No	Yes	$\frac{(n+r-1)!}{r!\,(n-1)!}$

Table 1 - Permutations and combinations

n = size of set r = positions

Combination symmetry

Picking r out of n is the same as **not** picking $n-r$ out of n.

Ex.: ${}^{10}C_6 = {}^{10}C_{10-6} = {}^{10}C_4$

Bayesian Inference

Ø - Null set

$x \in A$ - x is element of A

$A \ni x$ - A contains x

This can be inverted by doing a $\notin or \not\ni$

$\forall x$ - For all x

: $x\ is\ (condition)$ - such that x is (condition)

$A \subseteq B$ - A is subset of B

$A \cup B$ - A union with B

$A \cap B$ - A interception with B

$A|B$ - A given B

Exclusivity

If $A \cap B = \emptyset$ then they are mutually exclusive since they have no overlap.

Dependency

If $P(A|B) = P(A)$ then A and B are independent events since B happening does not affect A from happening.

Conditional probability

$P(A|B) = \frac{P(A \cap B)}{P(B)}$ since the event is guaranteed to have happened in B and we now need to know that the probability of A inside B is.

Law of total probability

$P(A) = P(A|B_1) * P(B_1) + P(A|B_2) * P(B_2) + \cdots + P(A|B_n) * P(B_n)$ since this calculates the sum of the intersection of A in all other sets.

Additive law

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$, we have to subtract $P(A \cap B)$ since it would be duplicated otherwise.

Multiplication rule

$P(A \cap B) = P(A|B) * P(B)$, this is basically a rewrite of the Conditional Probability rule above.

Bayes' Law

$P(A|B) = \frac{P(B|A)*P(A)}{P(B)}$, this is basically the Conditional Probability rule but applying the Multiplication rule in the numerator.

Distributions

Y - outcome

y - one of the possible outcomes

$P(Y = y)$ is the same as $p(y)$

When working with a population (full set of data) or a sample (partial set of data) there are some differences in notation as shown below:

	Population	Sample
Mean	μ	$\bar{x}$
Variance	σ^2	s^2
Standard deviation	σ	s

Table 2 - Population vs Sample

Discrete vs Continuous distributions

There are 2 types of distributions, **Discrete** and **Continuous**, their main characteristics are the following.

Discrete:

- Have finite amount of outcomes;
- Can add up values in an interval to determine its probability;
- Can be expressed with tables and graphs;
- Expected values could be unattainable;
- Graphs consist of bars lined up;
- $P(Y \leq y) = P(Y < y + 1)$ since there are no values in between.

Continuous:

- Have infinite amount of possible values;
- Can't add the values that make up an interval since there is an infinite number of them;
- Can be expressed using graphs and continuous functions;
- Graphs consist of smooth curves;
- Need to use intervals to calculate probability;
- $P(Y = y) \approx 0$ for any y since the probability the exact y happening in the infinite number of values available is gargatuously low;
- $P(Y < y) = P(Y \leq y)$ for the same reason above.

Discrete distributions

Uniform distribution

$$U(a, b)$$

Characteristics:

- All outcomes have the same probability;
- Expected values and variance hold no value.

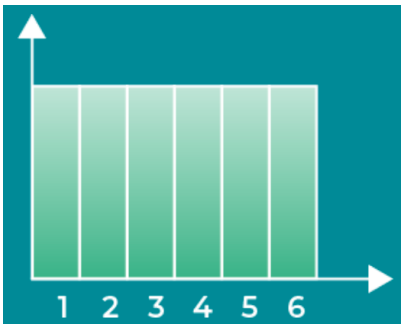

Figure 1 - Uniform distribution

Examples could be rolling a die.

Bernoulli distribution

$$Bern(p)$$

Characteristics:

- Only 1 trial;
- Only 2 outcomes;
- Used on binary situations like guessing True/False;
- Expected value is $E(Y) = p$ and the variance is $Var(Y) = p(1-p)$.

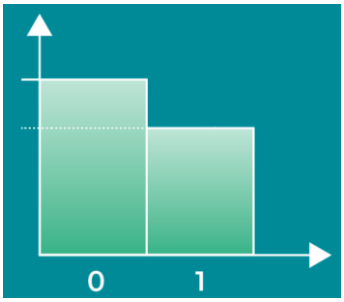

Figure 2 - Bernoulli distribution

Examples could be guessing heads or tails on a coin flip.

Binomial distribution

$$B(n,p)$$

Characteristics:

- Is a sequence of identical Bernoulli events;
- $P(Y = y) = C(y,n) * p^y * (1-p)^{n-y}$ where n is the number of trials;
- $E(Y) = n * p$ and $Var(Y) = n * p * (1-p)$.

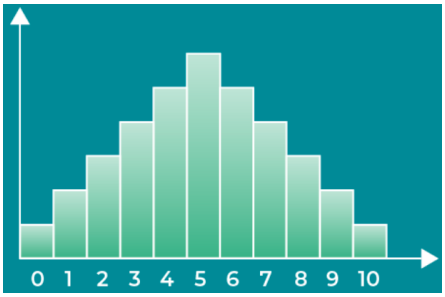

Figure 3 - Binomial distribution

Example could be how many times could we expect to hit tails if we flipped a coin n times.

Poisson distribution

$$Pois(\lambda)$$

Characteristics:

- $E(Y) = \lambda$;
- Measures over an interval number of time or distance with non-negative values only;
- $P(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!}$;
- $Var(Y) = \lambda$.

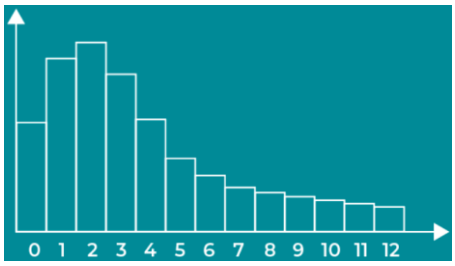

Figure 4 - Poisson distribution

Example could be determining the probability of a number a calls y received per-minute in a call centre, knowing that on average the number of calls is λ.

Continuous distributions

Normal distribution

$$N(\mu, \sigma^2)$$

Characteristics:

- Bell-shaped, symmetric and thin tails;
- $E(Y) = \mu$;
- $Var(Y) = \sigma^2$;
- Follows the 68-95-99.7 (or empirical) rule which means that 68%, 95% and 99.7% of the values are located 1, 2 and 3 standard deviations from the mean respectively.

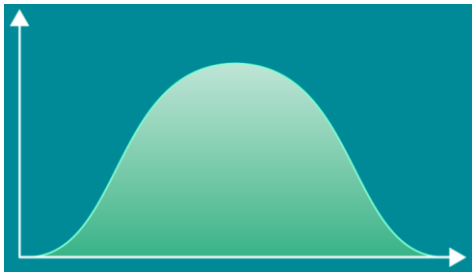

Figure 5 - Normal distribution

Example could be the weight of animals in the wild.

Standardization

Normal distributions can also be standardized in order to use a z-table, for this we have to make the mean become 0 and both the variance and standard deviation become 1, the formula is the following:

$$z = \frac{y - \mu}{\sigma}$$

The new z variable is used to represent how many standard deviations from the mean the value is.

Students' T

$$t(k)$$

Characteristics:

- Smaller sample size compared to the normal distribution;
- Bell-shaped, symmetric and has fatter tails;
- Is better than the normal distribution at handling extreme values;
- For $k > 2$: $E(Y) = \mu$ and $Var(Y) = s^2 * \frac{k}{k-2}$.

Figure 6 - Students' T distribution

Examples could be samples of Normal Distributions.

Chi-Squared

$$X^2(k)$$

Characteristics:

- Asymmetric with skewness to the right;
- $E(Y) = k$ and $Var(Y) = 2k$.

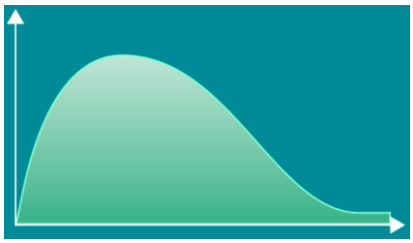

Figure 7 - Chi-Squared distribution

Example could be to test **goodness of fit**.

Exponential Distribution

$$Exp(\lambda)$$

Characteristics:

- Both the PDF and CDF plateau at the same point;
- Often uses the natural logarithm to transform values since we don't have a table of known values;
- $E(Y) = \frac{1}{\lambda}$ and $Var(Y) = \frac{1}{\lambda^2}$.

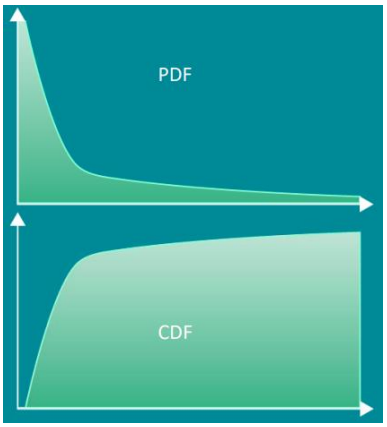

Figure 8 - Exponential distribution

Example could be the number of views a video gets from the day it gets published.

Logistic distribution

$$Logistic(\mu, s)$$

Characteristics:

- The CDF peaks when near the mean;
- The smaller the scale (s) parameter, the quicker it gets close to 1;
- $E(Y) = \mu$ and $Var(Y) = \frac{s^2 * \pi^2}{3}$.

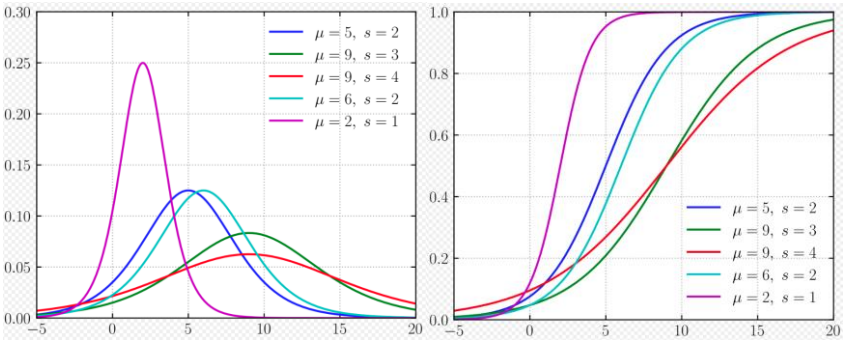

Figure 9 - Logistic distribution PDF (left) and CDF (right)

Descriptive Statistics

Types of data

Categorical:

- Data that represents a group or category, such as brands, gender and yes/no questions.

Numerical:

- Discrete: usually countable and finite, such as number of students in a classroom, multiple choice test scores.
- Continuous: infinite intervals and impossible to count, such as weight and height.

Levels of measurement

Qualitative (for categorical data):

- Nominal: categories that don't have any valuable order, such as car brands and the seasons of the year.
- Ordinal: categories that can be ordered and hold value in their order, such as ratings based on sentiment like bad, average, good and perfect.

Quantitative (for numerical data):

- Interval: represents numbers but **doesn't have** a true zero, such as degrees Celsius and Fahrenheit.
- Ratio: represents numbers that **have** a true zero, such as degrees Kelvin and weight.

Visual representations for categorical data

Usually the visual representation for categorical data is linked to their frequency (amount of times it appears) in the data available, the most common representations are **frequency distribution tables**, **bar charts**, **pie charts** and **Pareto diagrams**.

Example:

We have a data set with 3 different dog breeds, their frequency is as follows: 17 Doberman's, 11 German Shepherds and 21 Corgis. Now let's represent this data using the different methods available.

Frequency distribution table

	Frequency	Relative frequency
German Shepherd	11	0.22
Dobberman	17	0.35
Corgi	21	0.43
Total	**49**	**1**

Figure 10 - Frequency distribution table example

Bar chart

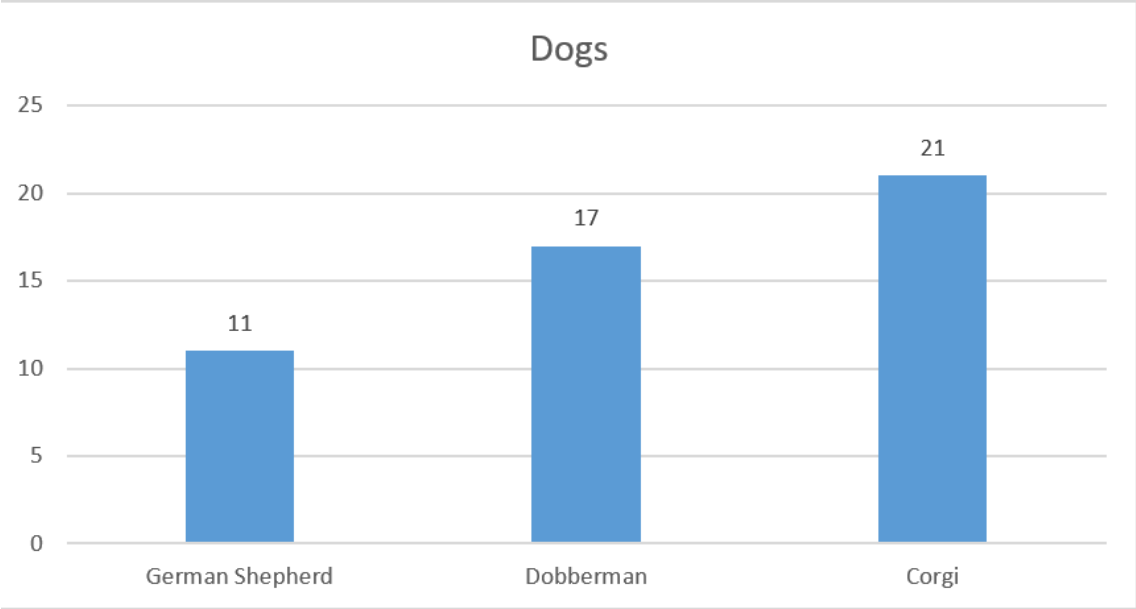

Figure 11 - Bar chart

Pie chart

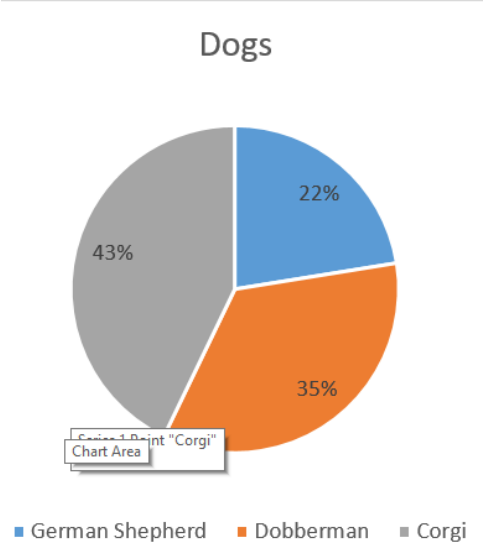

Figure 12 - Pie chart example

Pareto diagram

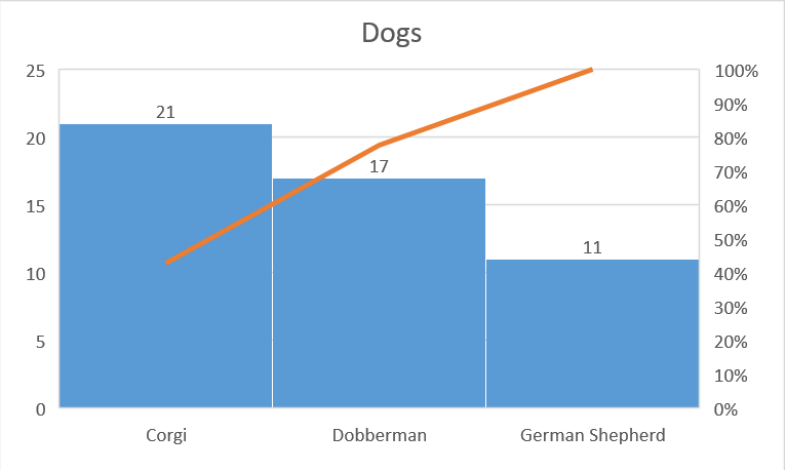

Figure 13 - Pareto diagram example

Note: the orange line represents the cumulative percentage from the biggest to the smallest, in this example it shows us that the 2 most frequent dog breeds in this data represent close to 80% of it.

Visual representation for numerical data

Numerical data can be represented by a large variety of ways depending on size and relation with other variables, for single variables the most popular method is the **histogram**.

Example:

We'll generate random data that has 1 column, 30 rows and values ranging from 1 to 50 and represent them using the methods mentioned above.

Histogram

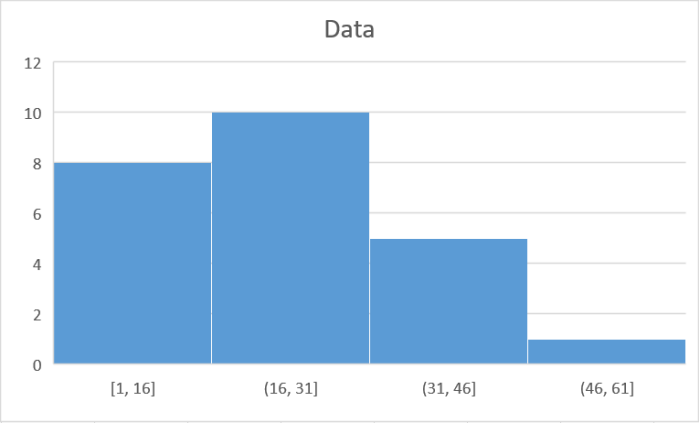

Figure 14 - Histogram example

Note: the bars are touching to represent continuity between intervals, in this example the histogram has 4 bins, usually the most robust method to calculate the bin size is the ***Freedman-Diaconis***, but it is not strictly necessary and it could instead just be played around with.

Freedman-Diaconis rule

The *Freedman-Diaconis* rule states that the bin-width should be $2 * \frac{IQR(x)}{\sqrt[3]{n}}$ where $IQR(x)$ is the interquartile range of the data and n the size of the

data, and that the number of bins is $(max - min)/h$ where h is the bin-width, the overall combination of this would be:

$$n\ of\ bins = (max - min)/(2 * \frac{IQR(x)}{\sqrt[3]{n}})$$

Relations between variables

We can also represent relations between multiple variables instead of just one, as an example we'll see the **Side-by-side bar chart** for categorical data and the **Scatter plot** for numerical data.

Side-by-side bar chart

In this example we have 17 Doberman's, 11 German Shepherds and 21 Corgis distributed randomly between 2 shelters.

For clarification this is the cross table:

Dog breed/Shelter	Shelter 1	Shelter 2	Total
German Shepherd	8	3	11
Dobberman	9	8	17
Corgi	8	13	21
Total	**25**	**24**	**49**

Figure 15 - Cross table example

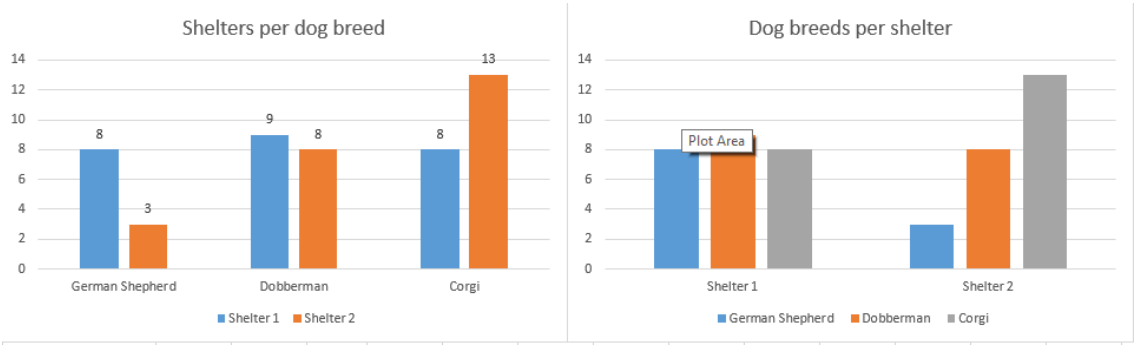

Figure 16 - Side-by-side chart example

Note: the side-by-side chart depends on the relation of the cross table, although both charts represent the same data the visualization of **dogs per shelter** is different from **shelters per dog**, choose ate your own discretion the one you believe is best.

Scatter plot

In this example we'll use random data that has 2 columns, 30 rows and values ranging from 1 to 50.

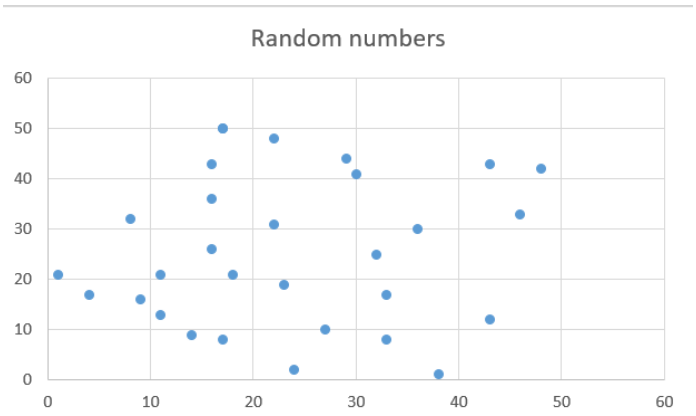

Figure 17 - Scatter plot example

Note: in this case the scatter plot doesn't hold much value since they are random numbers without any relation, usually if two variables have a relation between them we can observe progressions or clusters in the scatter plot.

Population vs Sample

x – dataset

n – dataset size

When using a dataset, it is important to know whether we are using a sample or a population, the population represents the entirety of the data we are trying to analyse and the sample is a sample of that, usually it is not realistic to use the population due to lack of data or hardware limitations.

Example:

If we want to know the average salary for people in a country we would use a sample of the people instead of the whole population since we probably don't have the means to get that information and even if we do we might not be able to process it.

Difference in symbols

	Population	Sample
Data	N	n
Mean	μ	$\bar{x}$
Variance	σ^2	s^2
Standard deviation	σ	s
Covariance	σ_{xy}	s_{xy}
Correlation	ρ	r

Table 3 - Population vs Sample

Mean, median and mode

The mean, median and mode are very important statistical values they will be explained bellow.

Mean

The mean is simply the average of the dataset, because it is only the average it can be highly affected by outliers, the formula is:

$$\frac{\Sigma_{i=1}^{n} x_i}{n}$$

Median

The median is the middle of the ordered dataset, it can be very useful since it is not affected by outliers like the mean, although it may not represent the whole dataset as well the mean. It has different formulas whether the dataset size is odd or even.

If n is odd, $median(x) = x_{(n+1)/2}$

If n is even, $median(x) = \frac{x_{n/2} + x_{(n+1)/2}}{2}$, this may seem complicated but is simply the average of the 2 middle values.

Mode

The mode is the most repeated element in the dataset, this is calculated by finding the highest frequency. It doesn't give much insight about the dataset as a whole specially when we are dealing with continuous values, unless we have a mode of intervals, which in this case would be the highest bar on the histogram.

Skewness

Skewness is a measure of asymmetry that indicates where the data in the dataset is concentrated, the closer to 0 the closer it is to the middle, if negative it means the data is more concentrated on the right side and that the outliers are to the left, and if positive it means that the data is concentrated on the left side and the outliers are to the right, bellow is the example of a positive skew:

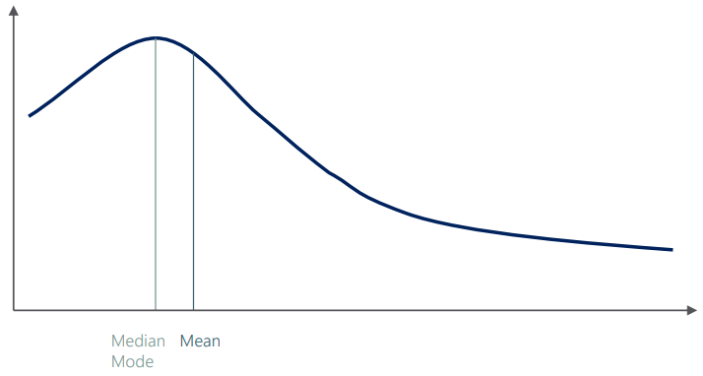

Figure 18 - Positive skew example

Usually software is used to calculate the skewness but the formula is as follows:

$$\frac{\frac{1}{n}\Sigma_{i=1}^{n}(x_i - \bar{x})^3}{\sqrt{\frac{1}{n-1}\Sigma_{i=1}^{n}(x_i - \bar{x})^2}^3}$$

Note: feeling dumb right now.

Variance and standard deviation

The variance and standard deviation are methods of measuring the dispersion of data around the mean value, they are different whether we are working with samples or populations due to needing to correct the bias (*Bessels's correction*), the formulas are as follows:

Variance:

$$s^2 = \frac{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

$$\sigma^2 = \frac{\Sigma_{i=1}^{N}(x_i - \mu)^2}{N}$$

Standard deviation:

$$s = \sqrt{s^2}$$

$$\sigma = \sqrt{\varrho^2}$$

Covariance

Covariance is the measure of joint variety of two variables, a positive covariance means that the variables move together, a negative means they are opposite and a zero means they are independent, covariance takes values between $-\infty$ and $+\infty$ so is hard to put it into perspective, a problem solved by correlation, the formulas are as follows:

$$s_{xy} = \frac{\Sigma_{i=1}^{n}(x_i - \bar{x}) * (y_i - \bar{y})}{n-1}$$

$$\sigma_{xy} = \frac{\Sigma_{i=1}^{N}(x_i - \mu_x) * (y_i - \mu_y)}{N}$$

Correlation

Correlation solves the problem that covariance has by making the value obtained be between -1 and 1, the logic is the same as covariance, the formulas are as follows:

$$r = \frac{s_{xy}}{s_x s_y}$$

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Inferential statistics

The Central Limit theorem

The Central Limit theorem is a statistical insight that states that no matter the distribution of a dataset, the sampling distribution of the means would approximate a normal distribution, also, the mean of the sampling of the means would be the mean of the original distribution and the variance would be n times smaller, where n is the size of the samples.

This theorem allows us to assume normality for many different variables and is very useful when doing confidence intervals, hypotheses testing and regressions.

Summarized, any distribution when sampled multiple times has the mean of the means for those samples approximate the original mean, the more and the bigger our samples are, the closer it will be to normal distribution.

Estimators and estimates

Estimator

An estimator is the term used to call a mathematical function that approximates a population parameter given only sample information.

Term	Estimator	Parameter
Mean	$\bar{x}$	μ
Variance	s^2	σ^2
Correlation	r	ρ

Table 4 - Estimator vs Parameter

Estimators have 2 important properties:

- **Bias**: an unbiased estimator is the population parameter. A biased estimator is $parameter + b$ where b is the bias.
- **Efficiency**: the smaller the variance of an estimator the more efficient it is.

Estimates

An estimate is the output generated by an estimator, there are 2 types of estimates, point estimates and confidence intervals.

Point estimates are single values while confidence intervals are intervals between values. Confidence intervals are much more precise than point estimates and are the preferred method when doing inferences.

Confidence Intervals and Margin of Error

A confidence interval is an interval in which we are confident using a percentage that the population parameter will fall within. We make a confidence interval around a point estimate.

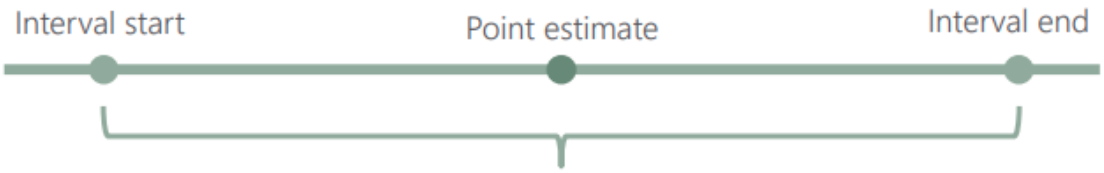

Figure 19 - Confidence Interval example

The level of confidence is represented by $(1-\alpha)$ where $(1-\alpha)*100$ is our confidence percentage such as 90%, 95%, 99%. Common α values are 0.1, 0.05 and 0.01.

The general formula for confidence intervals is $[\bar{x}-ME, \bar{x}+ME]$, the formula for ME is:

$$ME = reliability\ factor * \frac{standard\ deviation}{\sqrt{sample\ size}}$$

This formula chances whether we are dealing with populations or samples and whether we are working with 1 or 2 populations, the $reliability\ factor$ is given by the respective values on the *z-table* or *t-table*.

One population

If we only have 1 population the formulas depend on whether the population variance is known or not.

Population variance	Statistic	Variance	Formula
known	z	σ^2	$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
unknown	t	s^2	$\bar{x} \pm t_{n-1,\alpha/2} \frac{\sigma}{\sqrt{n}}$

Table 5 - Single population confidence intervals

Two populations

For two populations we have much different ways of calculating both the variance and the confidence intervals.

Dependent samples

Statistic	Variance	Formula
t	s_d^2	$\bar{d} \pm t_{n-1,\alpha/2} \frac{s_d}{\sqrt{n}}$

Table 6 - Two populations, dependent samples

Note: d represents the difference from sample 1 to 2

Independent Samples with known population variance

Statistic	Variance	Formula
z	σ_x^2, σ_y^2	$(\bar{x} - \bar{y}) \pm z_{\alpha/2} \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$

Table 7 - Two populations, known variance

Independent Samples with unknown variances

Population variance	Variance	Formula
Assumed equal	$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$	$(\bar{x} - \bar{y}) \pm t_{n_x+n_y-2,\alpha/2} \sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}$
Assumed different	s_x^2, s_y^2	$(\bar{x} - \bar{y}) \pm t_{v,\alpha/2} \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$

Table 8 - Two populations, unknown variances

Note that when we assume different population variance the degrees of freedom are represented by v, the formula for v can be calculated using *Welch's t-test* and it haunts me:

$$v = \frac{(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y})^2}{\frac{(\frac{s_x^2}{n_x})^2}{n_x - 1} + \frac{(\frac{s_y^2}{n_y})^2}{n_y - 1}}$$

Hypothesis testing

Hypothesis are ideas that can be put to the test, this is based on evidence as a starting point for investigation.

A hypothesis consists of 2 types:

- Null hypothesis (H_0), this is the hypothesis we want to test, it represents the status-quo and we assume it true unless we have evidence that states otherwise (innocent until proven guilty).
- Alternative hypothesis (H_1 or H_A), this is the change that is contesting the status-quo, usually it is our own opinion.

Example:

Your friend tells you that the average salary for a senior data scientist is at least $70k, you don't share that opinion and want to disprove him. In this example the hypothesis would be:

$H_0 - Average\ Salary\ \geq \$70.000$ (your friend's opinion, status-quo)

$H_1 - Average\ Salary\ < \$70.000$ (your opinion, the challenge of the status-quo)

Decisions

When testing a hypothesis there are 2 possible outcomes, to either **accept** or **reject** the null hypothesis. To **accept** means we don't have enough information support the change to **challenge the status-quo**. To **reject** means that we have enough statistical information to state that the **status-quo is not representative of the truth**.

Level of significance and type of test

The level of significance is represented by α and it represents the probability of making the error of rejecting a null hypothesis that is true

(false positive). Common significance levels are 0.1, 0.05 and 0.01 (10%, 5% and 1% respectively).

There are 2 types of tests we can make, the **Two-sided/tailed** test, when the null hypothesis contains a $=$ or $\neq$ sign, and **One-sided/tailed** test when the null hypothesis doesn't ($<, \leq, >, \geq$).

Figure 20 - Type of test example

Statistical errors

There are 2 types of errors we can make while testing a hypothesis, **Type I** (False positive, rejecting the null when it is true) which is equal to the significance level (α), and **Type II** (False negative, accepting the null when it is false) which is equal to the beta (β).

To learn about how to calculate β go to https://www.statology.org/beta-level/.

The table for these values would look like the following:

	Reject H_0	**Accept H_0**
H_0 is true	Type I error (α)	Correct decision ($1-\alpha$)
H_0 is false	Correct decision ($1-\beta$)	Type II error (β)

Table 9 - Error table

P-value

The p-value is the smallest possible significance level where we can reject the null hypothesis given our sample.

Notable p-values are numbers that start with 0.000, meaning that when testing a hypothesis, we always strive for a number that contains 3 zeroes after the dot since we can reject the null at all significance levels, another

is 0.05 since it is often the cut-off line and if our p-value is bigger than 0.05 we usually accept the null hypothesis.

Calculating a p-value can be done using this website https://www.socscistatistics.com/pvalues/.

Formulae

There are different formulae for hypothesis testing depending on population numbers, variance and of the relation between the samples, they are as follows:

(maybe replace with my own table in the future)

# populations	Population variance	Samples	Statistic	Variance	Formula for test statistic
One	known	-	z	σ^2	$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
One	unknown	-	t	s^2	$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$
Two	-	dependent	t	$s^2_{difference}$	$T = \frac{\bar{d} - \mu_0}{s_d/\sqrt{n}}$
Two	Known	independent	z	σ_x^2 , σ_y^2	$Z = \frac{(\bar{x} - \bar{y}) - \mu_0}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$
Two	unknown, assumed equal	independent	t	$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$	$T = \frac{(\bar{x} - \bar{y}) - \mu_0}{\sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}}$

Figure 21 - Hypothesis testing formulae

The decision rules to reject the null are the following:

- $|test\ statistic| > |critical\ value|$
- $p\ value < significance\ level$

Note: do not confuse the critical values which are represented by either t or z depending on the appropriate statistic with the test statistic represented by T or Z, the lower case values are extracted from either the t or z-table and the upper case values are calculated with the above formulae.

Linear regression

Regression analysis is one of the most used methods for prediction, it's the most fundamental machine learning method and the starting point for advanced analytical learning.

A linear regression is the approximation of the relationship of 2 or more variables. Like many other statistical techniques linear regressions help us make prediction about the population with sample data.

Linear regression model

The model representation for the linear regression is $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, where:

- Y_i is the dependent variable
- β_0 is the constant
- β_1 is the slope
- X_i is the independent variable
- ε_i is the error

Note: we are using Greek letters since we are referring to the population.

Linear regression equation

The equation for the linear regression is $\hat{y} = b_0 + b_1 * x_1$ where:

- $\hat{y}$ (pronounced y-hat) is the predicted value
- b_0 is the constant estimation
- b_1 is the coefficient, which is sort of the influence that the variable is going to have in the curve
- x_1 is the data for the independent sample variable

Geometrical representation

$$\hat{y} = b_0 + b_1 * x_i$$

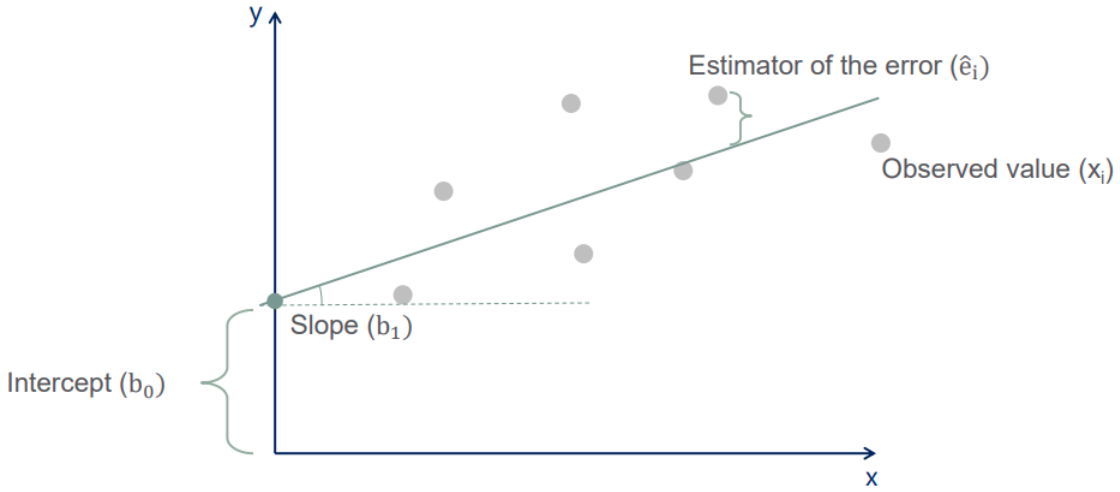

Figure 22 - Geometrical representation of linear regression

Note: a linear regression can only be represented if we have at max 2 independent samples since that is going to be a 3D representation, after that we just use the formulas. Also $\hat{e}_i$ is not included in the equation since on average it is expected to be 0.

Correlation vs Regression

Correlation	**Regression**
Represents the relationship between 2 variables	Represents the relationship between 2 or more variables
Shows that 2 variables move together (no matter the direction)	Shows cause and effect (one variable is affected by another)
Symmetrical $\boldsymbol{\rho(x,y) = \rho(y,x)}$	One way, only one variable is causally dependent
A single point (a number)	A 2D space line

Regression methods

The simplest and often sufficient method to estimate the regression line is OLS (ordinary least squares), although there are other methods that are

more appropriate for specific datasets and problems which are GLS (generalized least squares), MLE (maximum likelihood estimation), Bayesian regression, Kernel regression and Gaussian progress regression.

OLS assumptions

OLS's simplicity sometimes doesn't allow us to use it, in order to use it we have to meet some assumptions beforehand if we want to rely on this method, they are as follows:

- Linearity, the model must represent a linear relationship
- No endogenity, the independent variables shouldn't be correlated with the error term
- Normality and homoscedasticity, the variance of the errors should be consistent across observations
- No autocorrelation, no identifiable relationships should be found between the values of the error term
- No multicollinearity, no predictor variable should be almost perfectly explained by other predictors

Logistic Regression

Logistic regressions are used to predict outcomes that are categorical rather than numerical where the dependent variable is between 1 and 0, representing the probability of the event.

The model representation for the Logistic Regression model (or Logit model) is $p(X) = \frac{e^{(\beta_0+\beta_1 X_1+\cdots+\beta_i X_i)}}{1+e^{(\beta_0+\beta_1 X_1+\cdots+\beta_i X_i)}}$ where the outcome is the probability of an event ocurring.

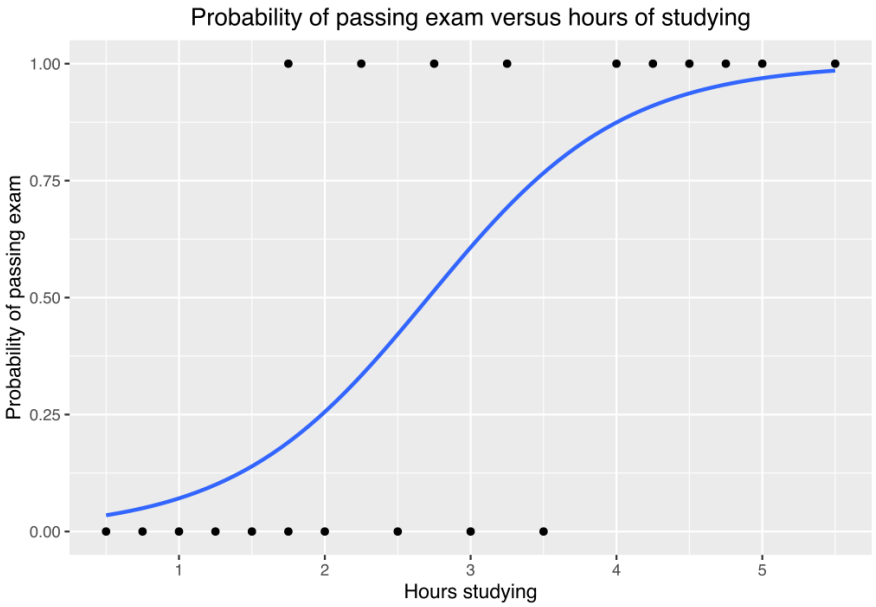

Figure 23 - Logistic regression example

Understanding the model values

Dep. Variable:	y	**No. Observations:**	518
Model:	Logit	**Df Residuals:**	516
Method:	MLE	**Df Model:**	1
Date:	Wed, 14 Feb 2024	**Pseudo R-squ.:**	0.2121
Time:	17:41:45	**Log-Likelihood:**	-282.89
converged:	True	**LL-Null:**	-359.05
Covariance Type:	nonrobust	**LLR p-value:**	5.387e-35

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.7001	0.192	-8.863	0.000	-2.076	-1.324
duration	0.0051	0.001	9.159	0.000	0.004	0.006

Figure 24 - Logistic regression table example

Above is a Logistic Regression table made by the “statsmodels” module in Python, in this example we can observe multiple fields, they are as follows:

- Dependent variable: the variable we are trying to predict

- Converged: Boolean value that indicates if we successfully found a solution, if it is False it means that the variables were not significant enough
- Pseudo R-Squared: according to McFadden, the favoured range for this is between 0.2 and 0.4
- Log-Likelihood: always negative and we aim that it is as high as possible
- Log-Likelihood-Null (LL-Null): is the Log-Likelihood for a model without independent variables, used to benchmark the worst possible model
- Log-Likelihood Ration p-value (LLR p-value): measures how different our model statistically is to the worst possible model

Note: on a Logistic regression the dependent variables coefficient contributes to the **log odds** therefore cannot be interpreted directly.

Cluster analysis

Cluster analysis is a statistical technique that separates data into different groups based on their features, the goal is to maximize similarity within a cluster and make different clusters be strongly distinct from each other.

Cluster analysis is also an example of unsupervised learning, meaning we have no input data and just try to group data points based on similarity without knowing what the outcome is.

K-means clustering

K-means clustering is a very popular and simple clustering model, to work with this model we have some steps to take:

1. We have to define k which will be the number of clusters we want, this is a manual operation but there are methods such as the *Elbow Method* that can help us find the most optimal number of clusters

2. Define the centroid (centre point of a cluster) seed, this can be chosen at random but there are algorithms such as *kmeans++* that can help to choose a more optimal seed
3. Assign each data point to a centroid, this is based on Euclidian distance proximity
4. Adjust the centroids to be in the centre of their data points, if all centroids remain in the same spot then the best clustering solution was found, otherwise repeat step 3

Euclidian distance

Euclidian distance simple represents the distance between 2 points, bellow is a representation of the distance for a 2D space:

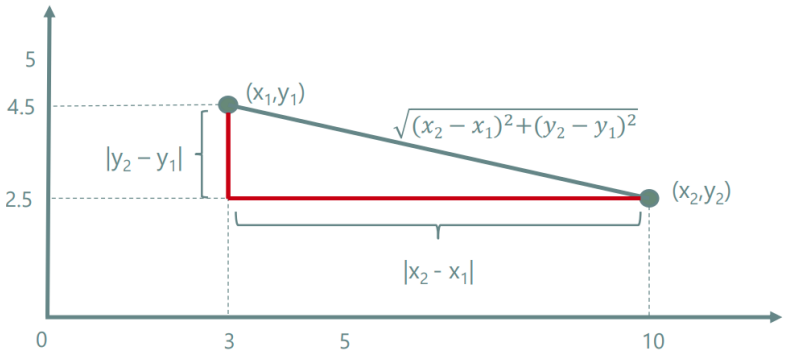

Figure 25 - Euclidian distance 2D example

Euclidian distance can be visualised and understood in our reality for 2D and 3D environments, but it can be used for any amount of N dimensions we desire, for example the 2D Euclidian distance formula is :

$$d(A,B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

and the 3D counterpart is:

$$d(A,B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$$

Noticing the pattern in here we can verify that just by adding the $(a_n - b_n)^2$ inside the radical we can calculate the distance for any dimension size possible, this is important since every independent variable we add to our clustering adds another dimension.

Note: $d(A,B) = d(B,A)$

Pros and Cons of K-means

Pros:

- Simple implementation
- Computationally efficient
- Widely used and in demand
- Always provides a result

Cons:

- We need to pick a K without context of how many clusters we would ideally have (methods such as the *Elbow Method* can help)
- Sensitive to initialization (methods such as *kmeans++* can help with this)
- Sensitive to outliers (the biggest issue with k-means by far, could be helped with previous data cleaning)
- Spherical solutions since it is based on Euclidian distance (not really fixable for this model)
- Always provides results (yes, both a pro and con) which can be decieving