Data Science Cheat Sheet

Diogo Centeno

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

Contents

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

$\emptyset$ - 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.

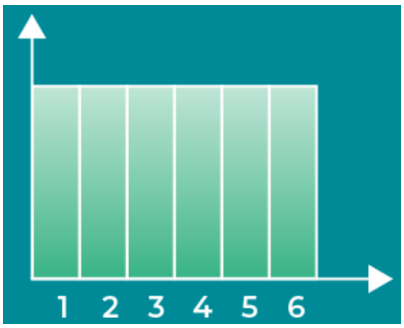

Figure 1 - Uniform distribution

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 $Var(Y) = p(1-p)$.

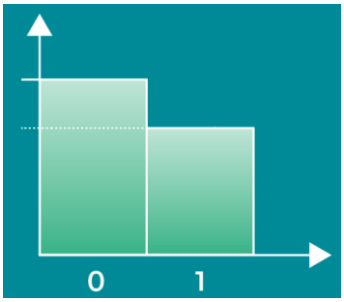

Figure 2 - Bernoulli distribution

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 $Var(Y) = n * p * (1-p)$.

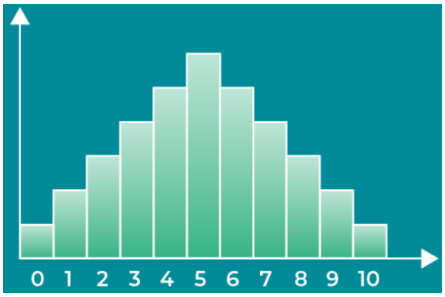

Figure 3 - Binomial distribution

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;
- $P(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!}$;
- $Var(Y) = \lambda$.

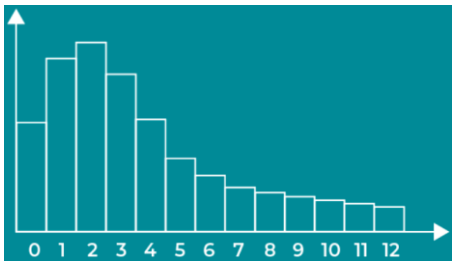

Figure 4 - Poisson distribution

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.

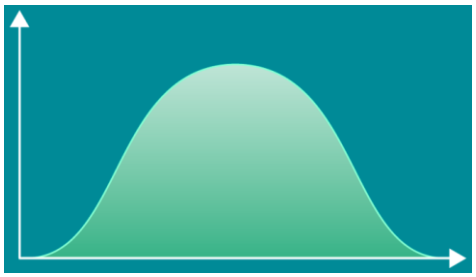

Figure 5 - Normal distribution

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 $Var(Y) = s^2 * \frac{k}{k-2}$.

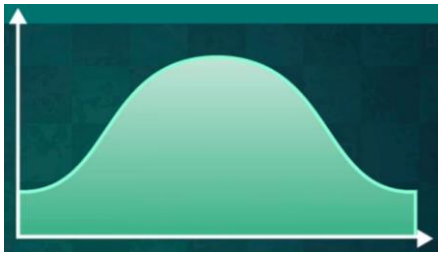

Figure 6 - Students' T distribution

Examples could be samples of Normal Distributions.

Chi-Squared

$$X^2(k)$$

Characteristics:

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

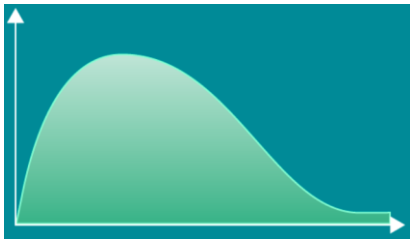

Figure 7 - Chi-Squared distribution

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 $Var(Y) = \frac{1}{\lambda^2}$.

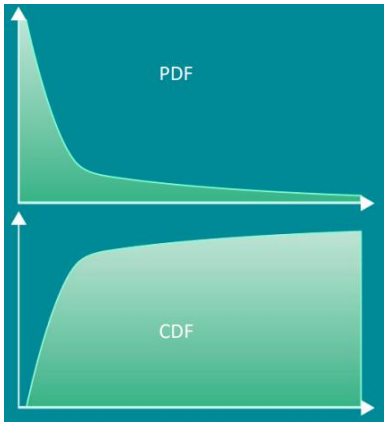

Figure 8 - Exponential distribution

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 $Var(Y) = \frac{s^2 * \pi^2}{3}$.

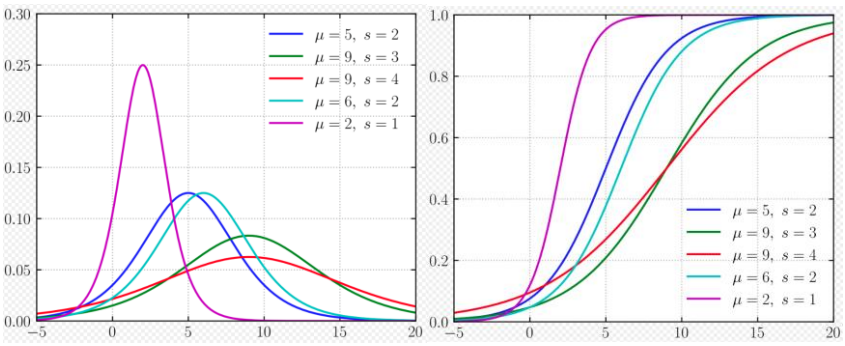

Figure 9 - Logistic distribution PDF (left) and CDF (right)

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

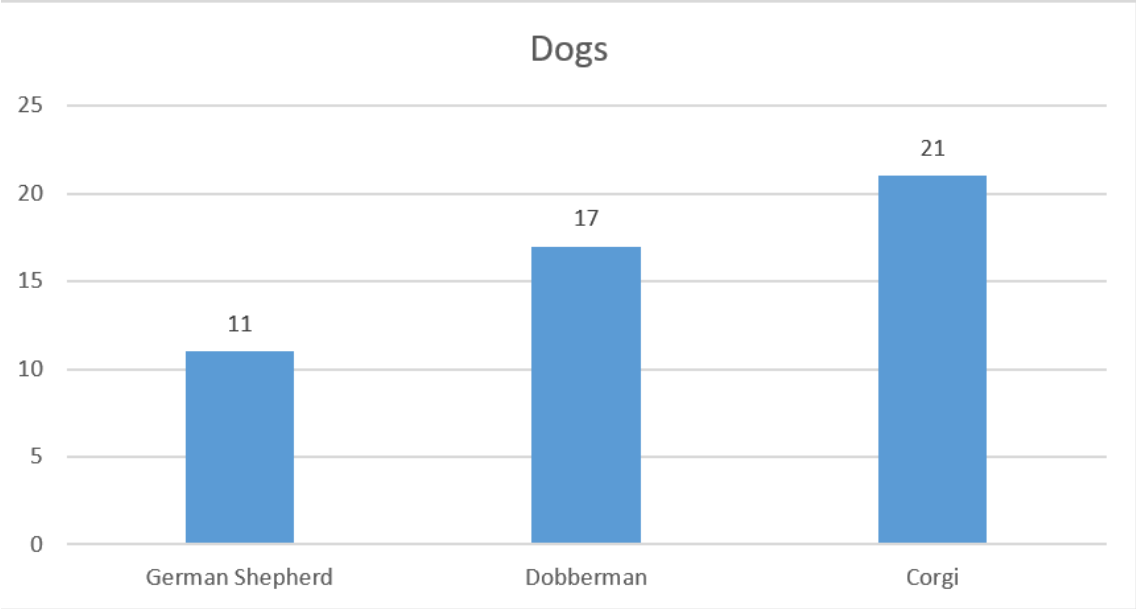

Figure 11 - Bar chart

Pie chart

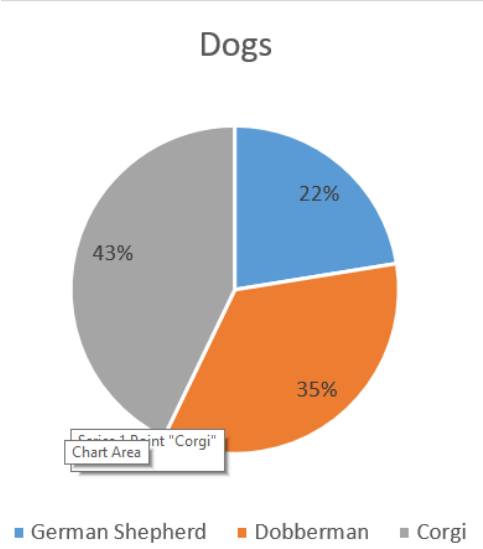

Figure 12 - Pie chart example

Pareto diagram

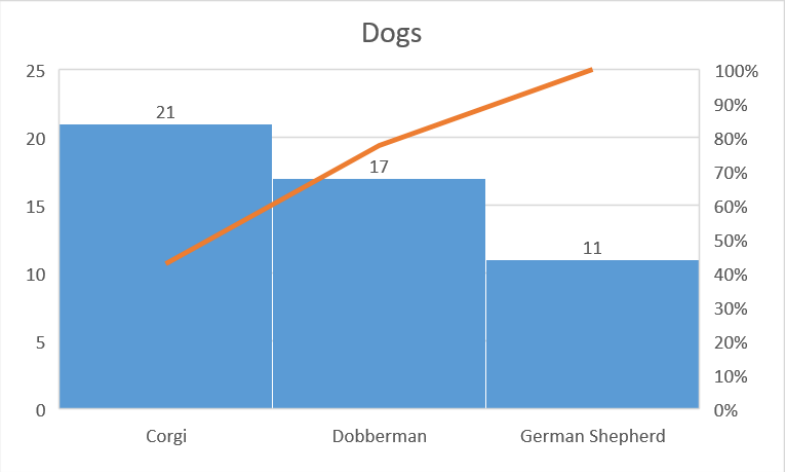

Figure 13 - Pareto diagram example

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

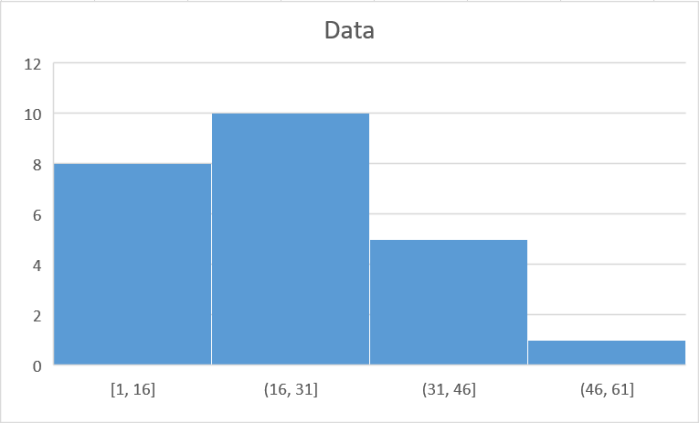

Figure 14 - Histogram example

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

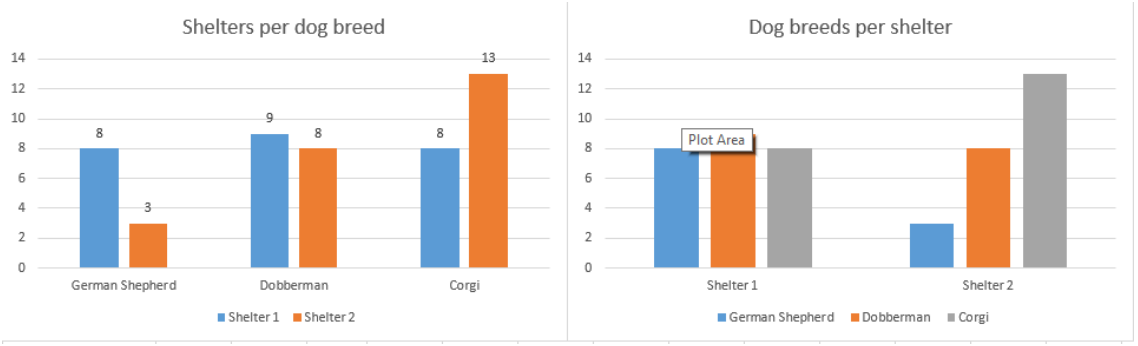

Figure 16 - Side-by-side chart example

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.

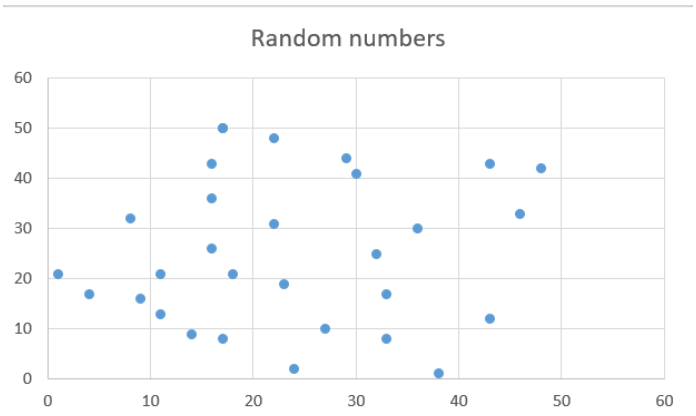

Figure 17 - Scatter plot example

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:

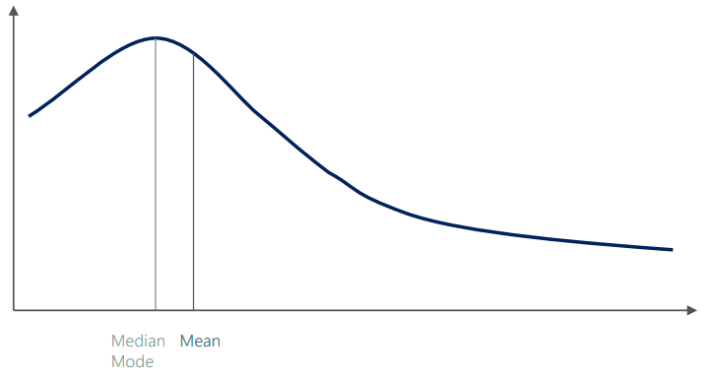

Figure 18 - Positive skew example

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:

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

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