Statistics for Data Science

Overview

Based on work by Janani Ravi (loonycorn.com) and several other resources

Definition

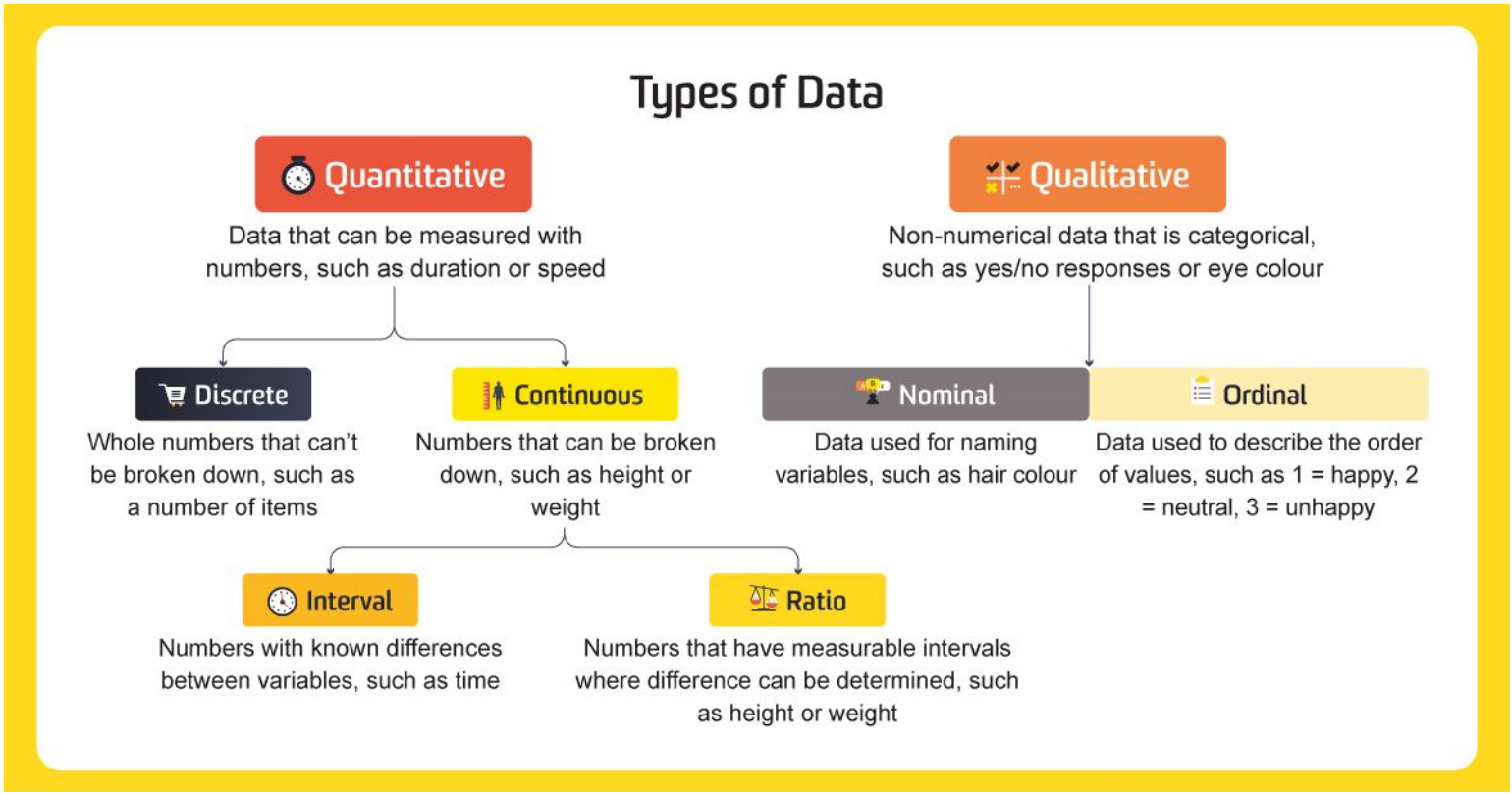
Statistics is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data. Conventionally we begin with a statistical population or a statistical model to be studied. Populations can be diverse groups of people or objects such as "all people living in a country" or "every atom composing a crystal". Statistics deals with every aspect of data, including the planning of data collection in terms of the design of surveys and experiments.

A descriptive statistic is a summary statistic that quantitatively describes or summarizes features from a collection of information, it can also be thought of as the process of using and analysing those statistics. We also use descriptive statistics to detect outliers and prepare data for further analysis such as feature engineering. This begins our exploratory analysis and descriptive visualization.

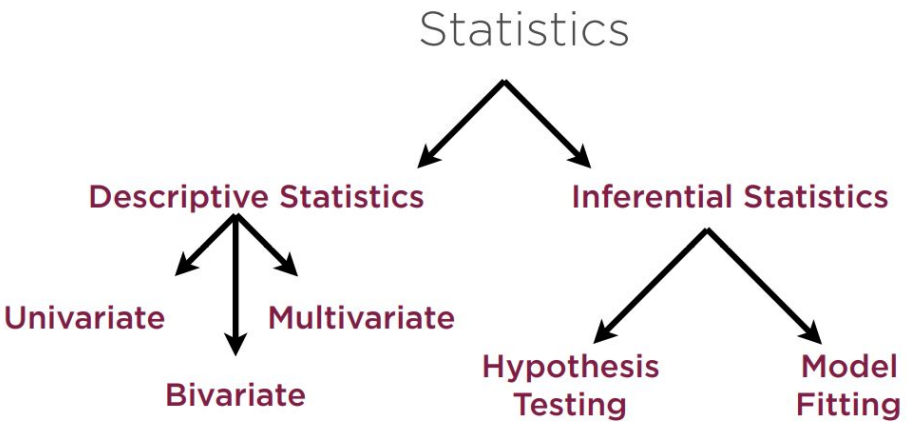
Inferential statistics is the process of using data analysis to deduce properties of an underlying distribution of probability from a model or a population by testing hypotheses and deriving estimates. (It is assumed that the observed data set is sampled from a larger population.)

Types of Variables - Categorical and Numerical

- Nominal variables are *categorical* variables that have two or more categories, but which do not have an intrinsic order. For example, houses, condos, co-ops or bungalows do not have an intrinsic order. Other examples include gender and hair color.
- Ordinal variables are *categorical* variables that have two or more categories just like nominal variables only the categories can also be ordered or ranked. A scale such as “terrible”, “not bad”, “reasonable”, “good”, and “great” can be ordered or ranked.
- Interval data is on a *numeric* scale. We know both the order and the exact differences between the values, however zero does not mean a “lack of”. An example of this is the temperature.
- Ratio data is on a *numeric* scale. It has everything that interval data has, and it has a “true zero”.



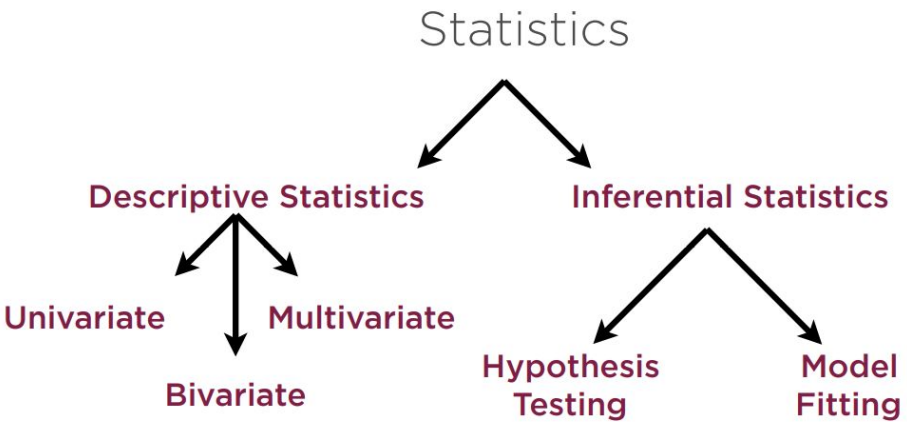
https://studyonline.unsw.edu.au/blog/types-of-data



Univariate data describes observations on only a single characteristic or attribute. For instance, the gender of attendees to an event, the amount of salaries in our company, or the type of product we are studying.

Bivariate data is data where each value of one of the variables is paired with a value of the other variable. We typically investigate the associations between the variables via a tables, graphs, or via sample statistics. For interval or ratio type variables can use a scatter plot, correlation, or regression model to quantify the association. For nominal or ordinal type variables a can use a contingency table or a test of independence.

Multivariate statistics studies the simultaneous observation and analysis of more than one outcome variable.



A statistical hypothesis is a hypothesis that is testable on observed data, modeled as the realised values of a collection of random variables.

Model fitting is the measure of how well a (machine learning) model generalizes data similar to the training data. A good fit means that it accurately approximates the output when it is provided with unseen inputs.

Fitting refers to adjusting the parameters in the model to improve accuracy. The process involves running an algorithm on data for which the target variable is known. The model's outcomes are compared to the real, observed values of the target variable to determine the accuracy.

Topics in Inferential Statistics

Many of these topics are often not covered (and are not crucial) for Data Science models.

However, certain ideas are occasionally used or referenced in a model, so it is an good idea to be familiar with them.

Inferential Statistics

42: Parameters, Statistics And Sampling Error
43: Distribution Of The Sample Mean
44: The Central Limit Theorem

45: Sample Proportions
46: Confidence Intervals About The Mean With Population Standard Deviation Known
47: Calculating Required Sample Size To Estimate Population Mean

48: Student's t-Distribution
49: Confidence Intervals About The Mean, Population Standard Deviation Unknown
50: Confidence Intervals For Population Proportions
51: Calculating Required Sample Size To Estimate Population Proportions

52: Null And Alternative Hypotheses
53: Type I and Type II Errors
54: One-Tailed And Two-Tailed Tests

55: Effect Size
56: Power
57: Statistical Vs. Practical Significance
58: Independent And Dependent Samples

59: One Sample z-Test
60: One Sample z-Test for Proportions
61: One Sample t-Test

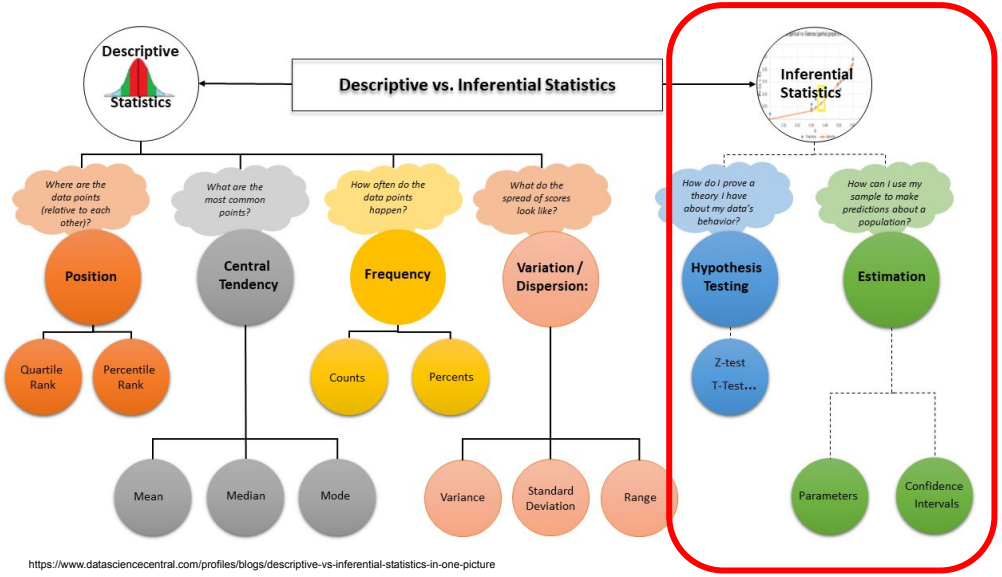
62: Independent Samples t-Test
63: Confidence Intervals For Independent Samples t-Test
64: Effect Size For Independent Samples t-Test
65: Dependent Samples t-Test

66: Confidence Intervals For Dependent Samples t-Test
67: Effect Size For Dependent Samples t-Test
68: z-Test for Proportions, Two Samples
69: Confidence Intervals For The Difference Of Two Proportions

70: Introduction To Analysis Of Variance (ANOVA)
71: One-Way ANOVA
72: Effect Size For One-Way ANOVA
73: Post-Hoc Tests For One-Way ANOVA
74: Repeated-Measures ANOVA
75: Factorial ANOVA, Two Independent Factors
76: Factorial ANOVA, Two Dependent Factors
77: Factorial ANOVA, Two Mixed Factors

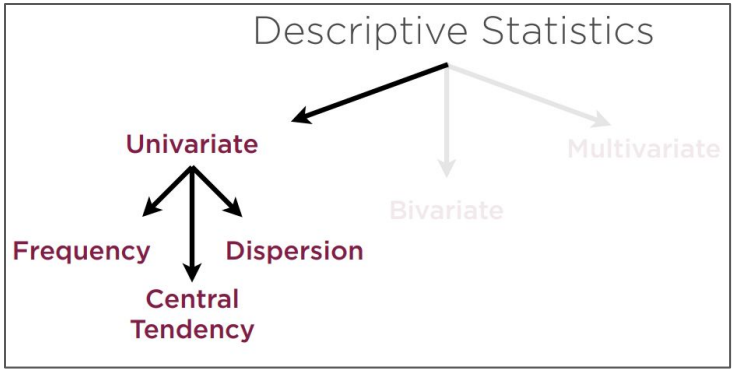
78: Chi-Square Test for Goodness of Fit
79: Chi-Square Test for Independence
80: Mann-Whitney U-Test
81: Wilcoxon Signed-Ranks Test
82: The Kruskal-Wallis Test
83: The Friedman Test

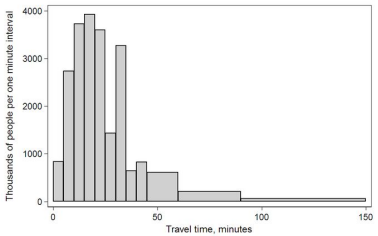
https://www.onlinemathlearning.com/statistics.html

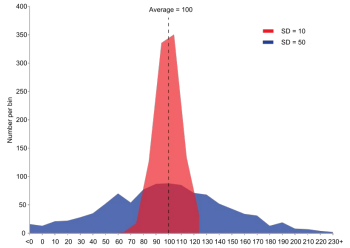


https://www.datasciencecentral.com/profiles/blogs/descriptive-vs-inferential-statistics-in-one-picture

Descriptive Statistics







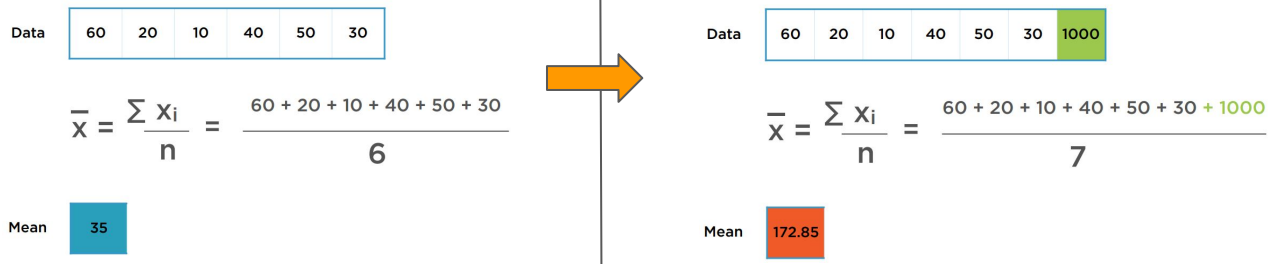
The frequency of an event *i* is the number n_i times the observation occurred in an experiment or study. It is often represented as a histogram.

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. Examples of this include harmonic mean, geometric mean, arithmetic mean, mode, median, etc.

Dispersion (also called variability, scatter, or spread) is the extent to which a distribution is stretched or squeezed. Examples of measures of dispersion include the variance, standard deviation, and interquartile range.

Measures of Central Tendency

Arithmetic Mean



The mean is a *model* of your data set. The mean is not often one of the actual observations in your data set. However, it minimizes error in the prediction of any one value in your data set. That is, it is the value that produces the lowest amount of error from all other values in the data set. It is a “balance point” in the data set.

It includes every value in your data set as part of the calculation. In addition, the mean is the only measure of central tendency where the sum of the deviations of each value from the mean is always zero. It should not be used for data that has large outliers or is skewed.

Harmonic Mean

The harmonic mean H of the positive real numbers $x_1, x_2, \ldots, x_n$ is

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \left(\frac{\sum_{i=1}^{n} x_i^{-1}}{n} \right)^{-1}.$$

$$\left(\frac{1^{-1} + 4^{\ 1} + 4^{\ 1}}{3} \right)^{-1} = \frac{3}{\frac{1}{1} + \frac{1}{4} + \frac{1}{4}} = \frac{3}{1.5} = 2$$

Geometric Mean

$$\left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}} = \sqrt[n]{x_1 x_2 \cdots x_n}$$

$$\sqrt[3]{4 \cdot 1 \cdot 1/32} = 1/2.$$

- The *arithmetic mean* is used when the values have the same units
- The *geometric mean* is used when the values have differing units. The geometric mean does not accept negative or zero values, although we can shift the data and thereby use such values. It is often used when data is exponential, such as when dealing with compound interest.
- The *harmonic mean* is used when the data values are ratios of two variables with different measures, called rates. An example of this is if we were to travel distances at different rates and we wanted to find a weighted average of the trip.

Median

The median is the middle score for a set of data that has been arranged in order of magnitude. The median is less affected by outliers and skewed data. In order to calculate the median, suppose we have the data below:

65	55	89	56	35	14	56	55	87	45	92

We first need to rearrange that data into order of magnitude (smallest first):

14	35	45	55	55	**56**	56	65	87	89	92

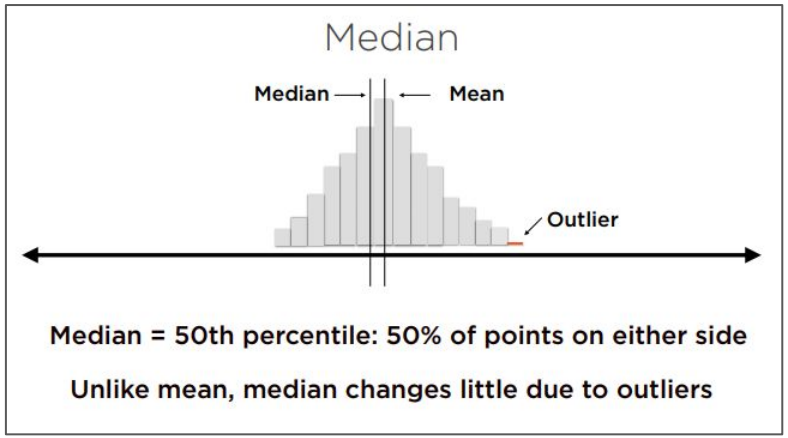
Our median mark is the middle mark - in this case, 56 (highlighted in bold). It is the middle mark because there are 5 scores before it and 5 scores after it. This works fine when you have an odd number of scores, but what happens when you have an even number of scores? What if you had only 10 scores? Well, you simply have to take the middle two scores and average the result. So, if we look at the example below:

65	55	89	56	35	14	56	55	87	45

We again rearrange that data into order of magnitude (smallest first):

14	35	45	55	**55**	**56**	56	65	87	89

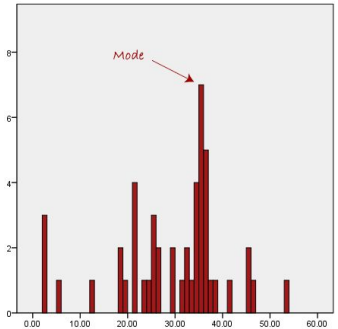
Only now we have to take the 5th and 6th score in our data set and average them to get a median of 55.5.



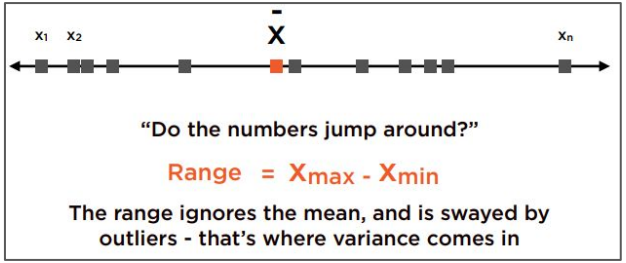
https://statistics.laerd.com/statistical-guides/measures-central-tendency-mean-mode-median.php

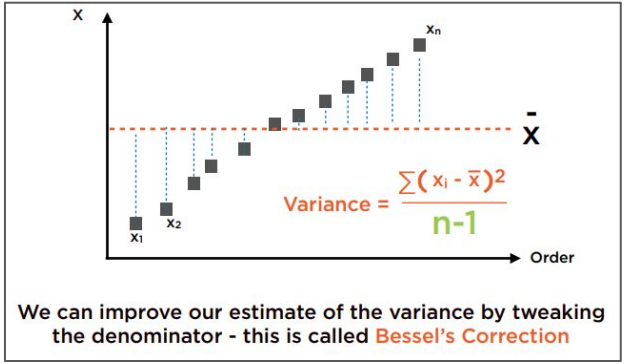
MODE

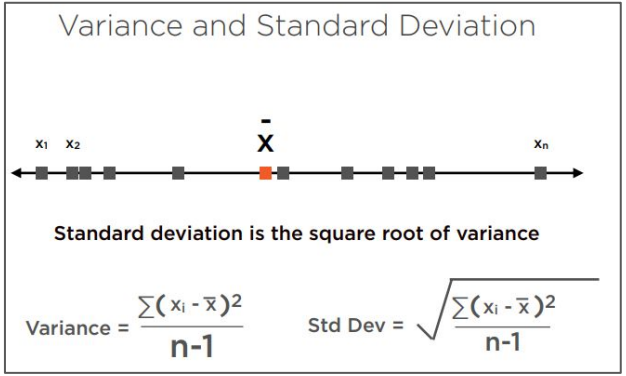
The mode is the most frequent score in our data set. We generally use it with discrete categorical data. There may be more than one mode, which may render the statistic meaningless.



Measures of Dispersion



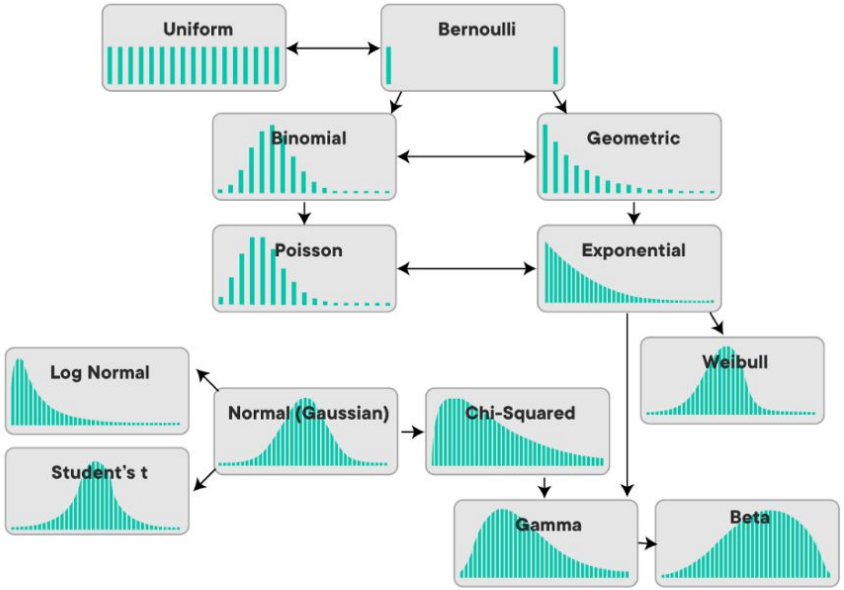




	A	B	C	D	E	F	G
1	DATA		AVERAGE	(Xi-AVG)	(Xi-AVG)^2	SUM(Xi-AVG)^2)	SQRT(SUM(Xi-AVG)^2/(n-1))
2	100		103.6714	-3.67141	13.47923721	66.09129952	1.865070896
3	100.8518			-2.81962	7.950254975		
4	105.3955			1.724051	2.972352875		
5	105.2107			1.539316	2.369495088		
6	105.9844			2.313002	5.349976286		
7	104.8436			1.172157	1.373951523		
8	106.6469			2.975514	8.853683054		STDEV.S
9	103.3541			-0.31731	0.100687446		1.865070896
10	103.2768			-0.39456	0.155680227		STDEV.P
11	100.7317			-2.93969	8.641781104		1.817846247
12	101.1523			-2.51909	6.345821859		
13	103.3346			-0.33679	0.113428102		
14	103.4622			-0.20918	0.043754713		
15	104.3601			0.688668	0.474263902		
16	104.4587			0.787323	0.619876914		
17	104.2989			0.627468	0.393715915		
18	102.6271			-1.0443	1.090569634		
19	105.8521			2.180731	4.755588191		
20	104.4924			0.820963	0.673979442		
21	103.0942			-0.57724	0.333201061		

Standard Deviation - the second moment, a measure of spread

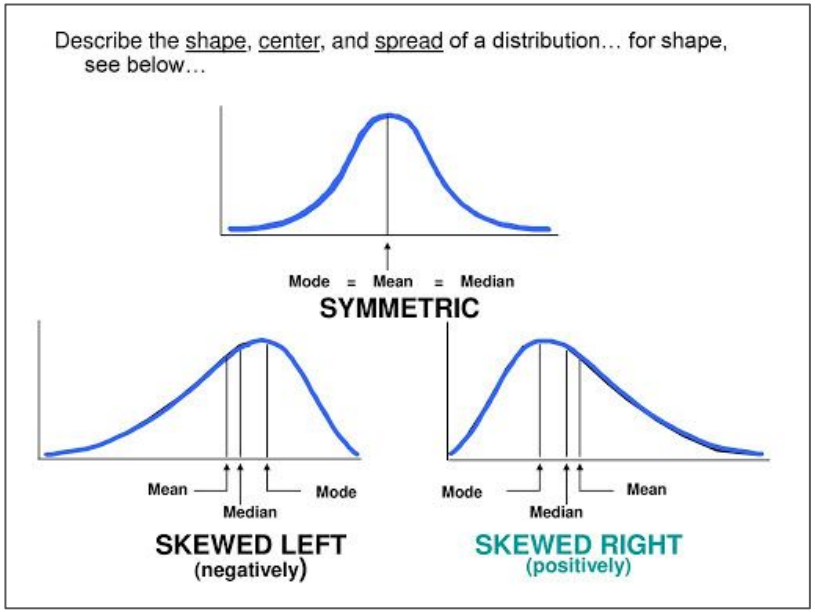
Distributions

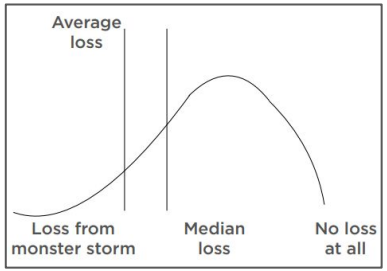


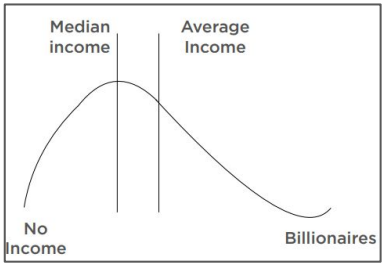
https://medium.com/mytake/understanding-different-types-of-distributions-you-will-encounter-as-a-data-scientist-27ea4c375eec

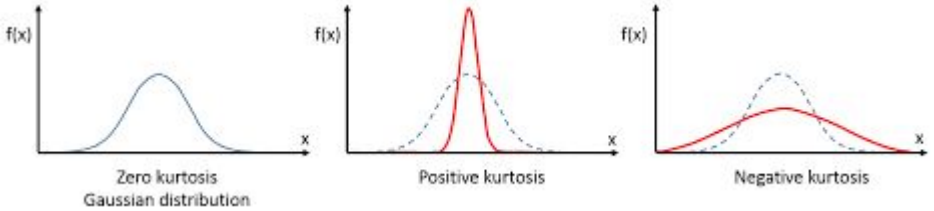
Moments

Moment ordinal	Moment		
	Raw	Central	Standardized
1	Mean	0	0
2	–	Variance	1
3	–	–	Skewness
4	–	–	(Non-excess or historical) kurtosis
5	–	–	Hyperskewness
6	–	–	Hypertailedness
7+	–	–	–









http://ceadserv1.nku.edu/longa/classes/sta205/days/day16.html

Binomial Distribution Formula

$$P(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{(n-x)!\,x!} p^x q^{n-x}$$

where
n = the number of trials (or the number being sampled)
x = the number of successes desired
p = probability of getting a success in one trial
$q = 1 - p$ = the probability of getting a failure in one trial

Normal Distribution Formula

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$\mu =$ mean of x
$\sigma =$ standard deviation of x
$\pi \approx 3.14159 \ldots$
$e \approx 2.71828 \ldots$

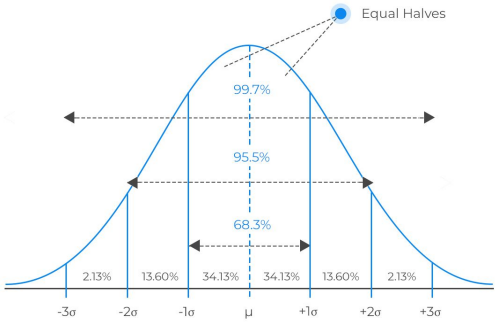
Poisson Distribution Formula

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

where
x = 0, 1, 2, 3, …
λ = mean number of occurrences in the interval
e = Euler's constant ≈ 2.71828

https://www.onlinemathlearning.com/

Shape of the normal distribution

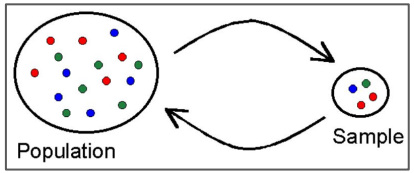


No. of standard deviations from the mean

https://analystprep.com/

Sampling Distributions

Sampling from a Population



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

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

A sampling distribution is a distribution of a statistic obtained from a large (enough) number of samples drawn from a population. The sampling distribution of a given population is the distribution of frequencies of a range of different outcomes that could possibly occur for a statistic of a population.

The sample mean is the best unbiased estimator of the true mean.

The Central Limit Theorem:

- As the sample size (n) gets larger, the sample means tend to follow a normal probability distribution
- As the sample size (n) gets larger, the sample means tend to cluster around the true population mean
- The above is true regardless of the distribution of the population from which the sample was drawn

Standard Error

- As the sample size increases, the distribution of sample means tends to converge closer together – to cluster around the true population mean
- Therefore, as the sample size increase, the standard deviation of the sample means decreases
- The **standard error of the mean** is the standard deviation of the sample means

- Standard error can be calculated as follows:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Where:

$\sigma_{\bar{X}}$ = the standard deviation of the sample means (standard error)

σ = the standard deviation of the population

$\sqrt{n}$ = the sample size

- In many applications, the true value of σ (the SD of the population) is unknown

- SE can be estimated using the sample SD

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Where:

$SE_{\bar{x}}$ = the standard deviation of the sample means (standard error)

s = the sample standard deviation (the sample based estimate of the SD of the population

$\sqrt{n}$ = the sample size

https://www.jcu.edu.au/__data/assets/pdf_file/0008/115478/Basic-Statistics-6_Sample-vs-Population-Distributions.pdf

Confidence Intervals

Confidence Intervals

A confidence interval is an estimate computed from the statistics of the *observed* data. This proposes a range of plausible values for an unknown parameter. The interval has an associated confidence level that the true parameter is in the proposed range. The confidence level is chosen by the investigator. For a fixed sample, higher degrees of confidence require a wider (less precise) confidence interval. In general terms, a confidence interval for an unknown parameter is based on sampling the distribution of a corresponding estimator.

The confidence level represents the theoretical long-run frequency (i.e., the proportion) of confidence intervals that contain the true value of the unknown population parameter. In other words, 90% of confidence intervals computed at the 90% confidence level contain the parameter, 95% of confidence intervals computed at the 95% confidence level contain the parameter, 99% of confidence intervals computed at the 99% confidence level contain the parameter, etc.

Confidence intervals are often misinterpreted. The logic behind them may be a bit confusing. Remember that when we're constructing a confidence interval we are estimating a population parameter when we only have data from a sample. We don't know if our sample statistic is less than, greater than, or approximately equal to the population parameter. And, we don't know for sure if our confidence interval contains the population parameter or not.

The correct interpretation of a 95% confidence interval is that "we are 95% confident that the population parameter is between X and X."

Example: Correlation Between Height and Weight

At the beginning of the Spring 2017 semester a sample of World Campus students were surveyed and asked for their height and weight. In the sample, Pearson's r = 0.487. A 95% confidence interval was computed of [0.410, 0.559].

The correct interpretation of this confidence interval is that we are 95% confident that the correlation between height and weight in the population of all World Campus students is between 0.410 and 0.559.

Example: Seatbelt Usage

A sample of 12th grade females was surveyed about their seatbelt usage. A 95% confidence interval for the proportion of all 12th grade females who always wear their seatbelt was computed to be [0.612, 0.668].

The correct interpretation of this confidence interval is that we are 95% confident that the proportion of all 12th grade females who always wear their seatbelt in the population is between 0.612 and 0.668.

Example: IQ Scores

A random sample of 50 students at one school was obtained and each selected student was given an IQ test. These data were used to construct a 95% confidence interval of [96.656, 106.422].

The correct interpretation of this confidence interval is that we are 95% confident that the mean IQ score in the population of all students at this school is between 96.656 and 106.422.

Wiki Warnings

Misunderstandings [edit]

See also: § Counter-examples, and Misunderstandings of p-values

Confidence intervals and levels are frequently misunderstood, and published studies have shown that even professional scientists often misinterpret them.[7][8][9][10][11]

- A 95% confidence level does not mean that for a given realized interval there is a 95% probability that the population parameter lies within the interval (i.e., a 95% probability that the interval covers the population parameter).[12] According to the strict frequentist interpretation, once an interval is calculated, this interval either covers the parameter value or it does not; it is no longer a matter of probability. The 95% probability relates to the reliability of the estimation procedure, not to a specific calculated interval.[13] Neyman himself (the original proponent of confidence intervals) made this point in his original paper:[5]

> "It will be noticed that in the above description, the probability statements refer to the problems of estimation with which the statistician will be concerned in the future. In fact, I have repeatedly stated that the frequency of correct results will tend to α. Consider now the case when a sample is already drawn, and the calculations have given [particular limits]. Can we say that in this particular case the probability of the true value [falling between these limits] is equal to α? The answer is obviously in the negative. The parameter is an unknown constant, and no probability statement concerning its value may be made..."

Deborah Mayo expands on this further as follows:[14]

> "It must be stressed, however, that having seen the value [of the data], Neyman–Pearson theory never permits one to conclude that the specific confidence interval formed covers the true value of 0 with either $(1 - \alpha)100\%$ probability or $(1 - \alpha)100\%$ degree of confidence. Seidenfeld's remark seems rooted in a (not uncommon) desire for Neyman–Pearson confidence intervals to provide something which they cannot legitimately provide; namely, a measure of the degree of probability, belief, or support that an unknown parameter value lies in a specific interval. Following Savage (1962), the probability that a parameter lies in a specific interval may be referred to as a measure of final precision. While a measure of final precision may seem desirable, and while confidence levels are often (wrongly) interpreted as providing such a measure, no such interpretation is warranted. Admittedly, such a misinterpretation is encouraged by the word 'confidence'."

- A 95% confidence level does not mean that 95% of the sample data lie within the confidence interval.
- A confidence interval is not a definitive range of plausible values for the sample parameter, though it may be understood as an estimate of plausible values for the population parameter.
- A particular confidence level of 95% calculated from an experiment does not mean that there is a 95% probability of a sample parameter from a repeat of the experiment falling within this interval.[11]

Calculating the Confidence Interval

How It Is Calculated

The Confidence Interval formula is

$$\bar{X} \pm Z\frac{s}{\sqrt{(n)}}$$

Where:

- $\bar{X}$ is the mean
- **Z** is the Z-value from the table below
- **s** is the standard deviation
- **n** is the number of observations

	Z
80%	1.282
85%	1.440
90%	1.645
95%	1.960
99%	2.576
99.5%	2.807
99.9%	3.291

https://www.mathsisfun.com/data/confidence-interval-calculator.html

Confidence Interval Formula

$$\text{Confidence Interval} = \left(\bar{X} - Z \times \frac{\sigma}{\sqrt{n}}\right) \text{ to } \left(\bar{X} + Z \times \frac{\sigma}{\sqrt{n}}\right)$$

$$\text{Confidence Interval} = \bar{X} \pm Z \times \frac{\sigma}{\sqrt{n}}$$

Confidence Intervals for μ

- For n ≥ 30

$$X \pm z\frac{s}{\sqrt{n}}$$

Use the Z table for the standard normal distribution.

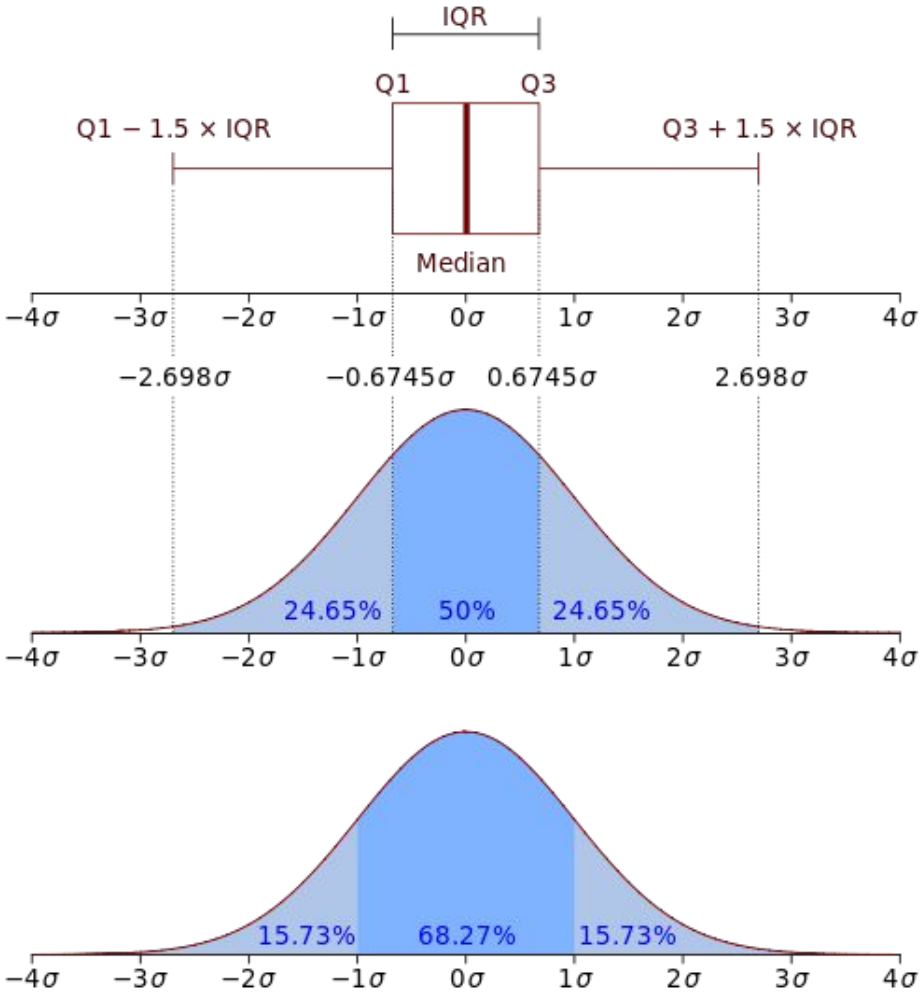
- For n<30

$$\bar{X} \pm t\frac{s}{\sqrt{n}}$$

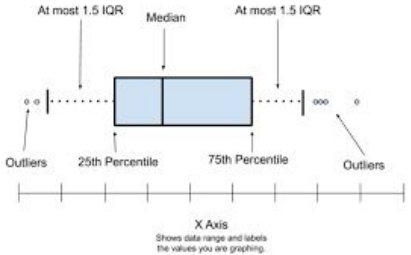
Use the t table with df=n-1

https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Confidence_Intervals/BS704_Confidence_Intervals_print.html

Boxplot



Wikipedia



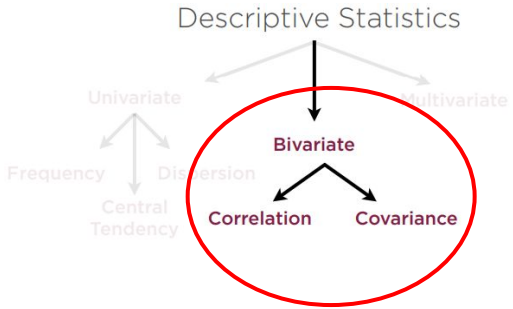

```
                    +-----+-+
          * |-------|     | |-------|
                    +-----+-+

+---+---+---+---+---+---+---+---+---+---+---+---+---+   number line
0   1   2   3   4   5   6   7   8   9   10  11  12
```

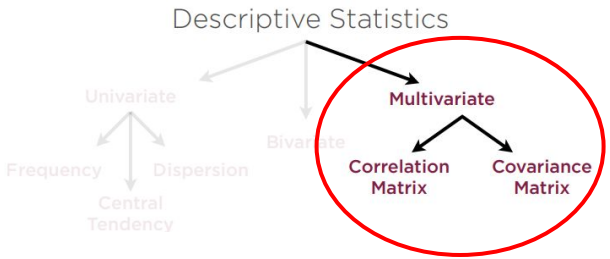
For the data set in this box plot:

- lower (first) quartile $Q_1 = 7$
- median (second quartile) $Q_2 = 8.5$
- upper (third) quartile $Q_3 = 9$
- interquartile range, IQR = Q_3 - Q_1 = 2
- lower 1.5*IQR whisker = Q_1 - 1.5 * IQR = 7 - 3 = 4. (If there is no data point at 4, then the lowest point greater than 4.)
- upper 1.5*IQR whisker = Q_3 + 1.5 * IQR = 9 + 3 = 12. (If there is no data point at 12, then the highest point less than 12.)

This means the 1.5*IQR whiskers can be uneven in lengths.

Bivariate and Multivariate Descriptive Statistics





Covariance

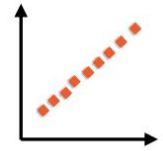
Measures relationship between two variables, specifically whether greater values of one variable correspond to greater values in the other.

$$\operatorname{cov}(X,Y) = \operatorname{E}\left[(X - \operatorname{E}[X])(Y - \operatorname{E}[Y])\right]$$

Correlation

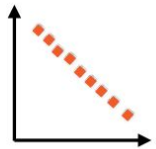
Similar to covariance; measures whether greater values of one variable correspond to greater values in the other. Scaled to always lie between +1 and -1.

$$\rho_{X,Y} = \operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\operatorname{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$



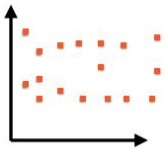
Correlation = +1

As X increases, Y increases linearly



Correlation = -1

As X increases, Y decreases linearly



Correlation = 0

Changes in X independent* of changes in Y