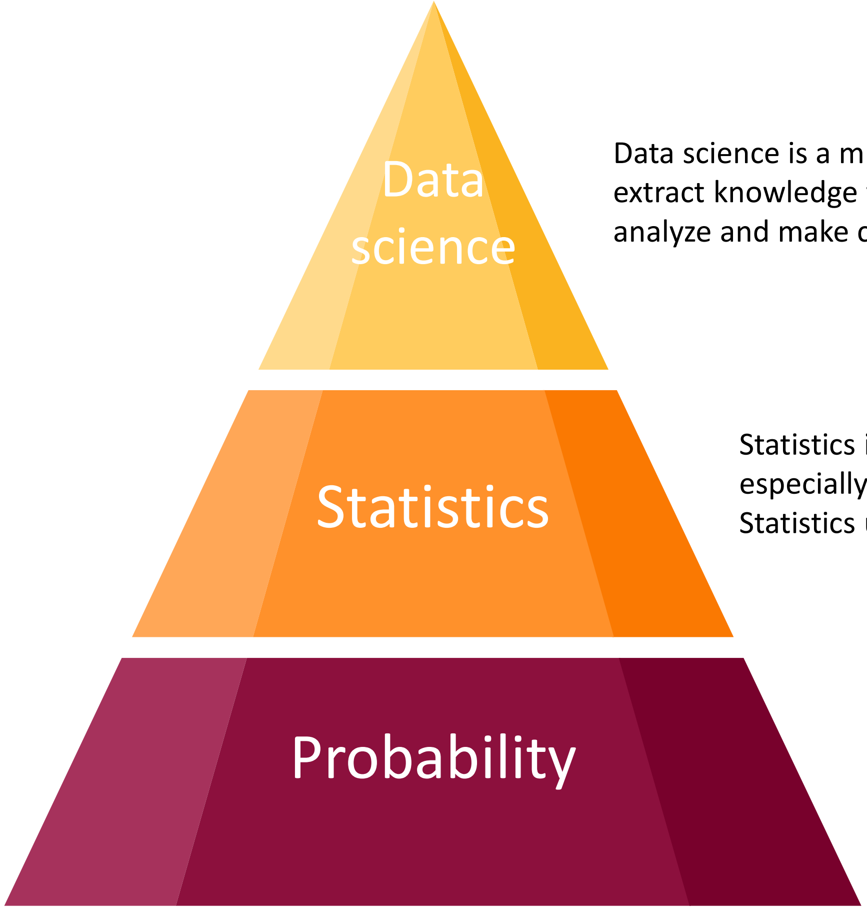
Probability and Statistics



Data science

Statistics

Probability

Data science is a multidisciplinary field which uses scientific methods, processes, and systems to extract knowledge from data in a range of forms. Statistics provides the methodology to collect, analyze and make conclusions from data.

Statistics is the practice or science of collecting and analysing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample. Statistics uses probability theory to draw conclusions from data.

Probability theory is a branch of mathematics concerned with probability. Probability is a numerical description of the likelihood of an event.

Probability example:
You have a fair coin (equal probability of heads or tails). You will toss it 100 times.
What is the probability of 60 or more heads?

Statistics example:
You have an unknown coin. You toss it 100 times and count 60 heads.
Your job as a statistician is to draw a conclusion (inference) from this data. Is it a fair coin or not?

Experimental probabilities

Trial: observing an event occur and recording the outcome (es. flipping a coin and recording the outcome).

Experiment: a collection of one or multiple trials (es. flipping a coin 20 times and recording the 20 individual outcomes).

Sample space is a set of all possible outcomes from an experiment.

An **event** is a set of outcomes of an experiment.

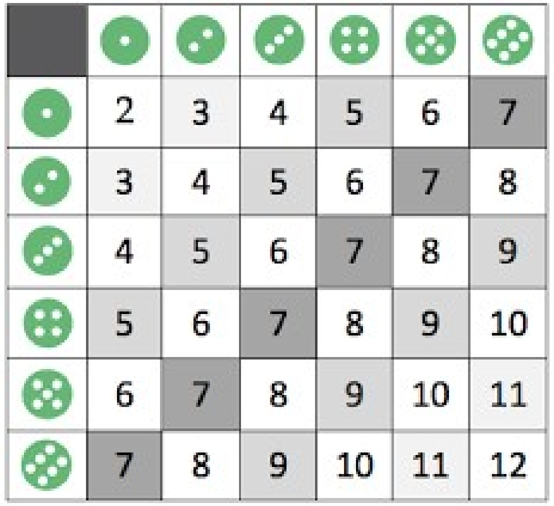
Events can be:

- **Independent**, which means that they are not affected by other events. For example, if you toss a fair coin, the chance that it lands on “heads” is 1/2 no matter what.
- **Dependent** namely, they are affected by other events. For example, as we remove cards from a deck, the probability of us choosing a king is becoming higher and higher.
- **Mutually exclusive** that is they can’t happen at the same time. E.g. you can’t turn left and right at the same time.

Experimental probability: probability of an event based on the experiment (preferred outcomes / sample space)

Expected value

Expected value: the average outcome we expect if we run an experiment many times.



	⚀	⚁	⚂	⚃	⚄	⚅
⚀	2	3	4	5	6	7
⚁	3	4	5	6	7	8
⚂	4	5	6	7	8	9
⚃	5	6	7	8	9	10
⚄	6	7	8	9	10	11
⚅	7	8	9	10	11	12

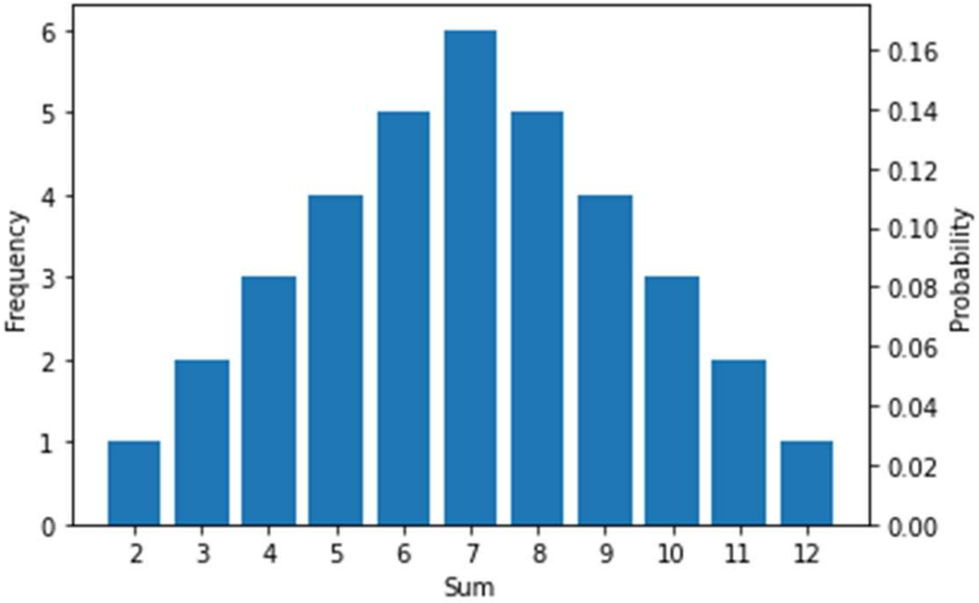
$$\boxed{E(X) = \sum_{i=1}^{n} p_i n_i}$$

$$E(X) = P(2)*2 + P(3)*3 + \ldots + P(12)*12$$

Expected value

Expected value: the average outcome we expect if we run an experiment many times.

Sum	Frequency	Probability
2	1	0.028
3	2	0.056
4	3	0.083
5	4	0.111
6	5	0.139
7	6	0.167
8	5	0.139
9	4	0.111
10	3	0.083
11	2	0.056
12	1	0.028

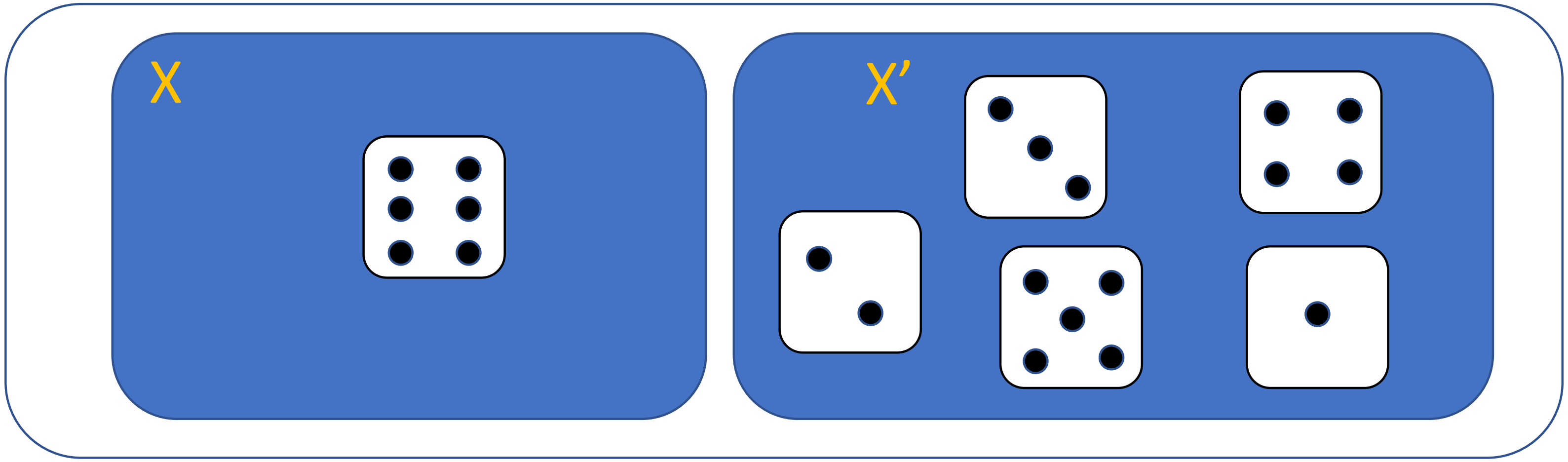


E(X) = P(2)*2 + P(3)*3 + … + P(12)*12 = 7

Complementary events

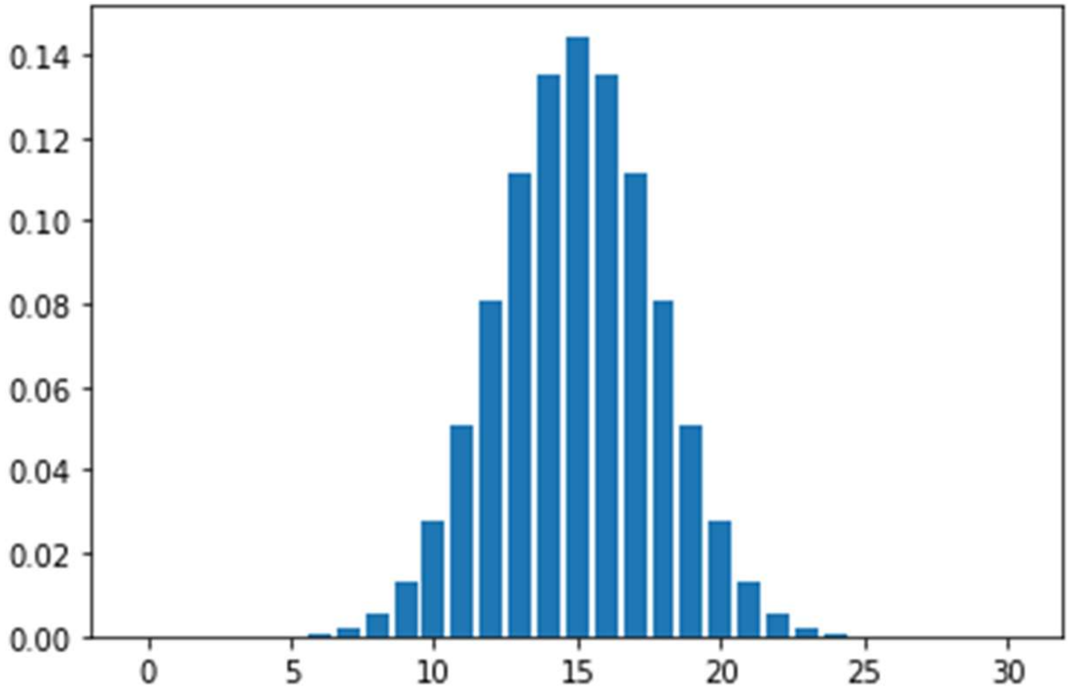
Two events are said to be complementary when one event occurs if and only if the other does not. The probabilities of two complimentary events add up to 1.

$$P(X') = 1 - P(X)$$



Probability distributions

A probability distribution describes all the possible values and likelihoods that a random variable can take within a given range.



Mean:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Variance:

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

Standard deviation:

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

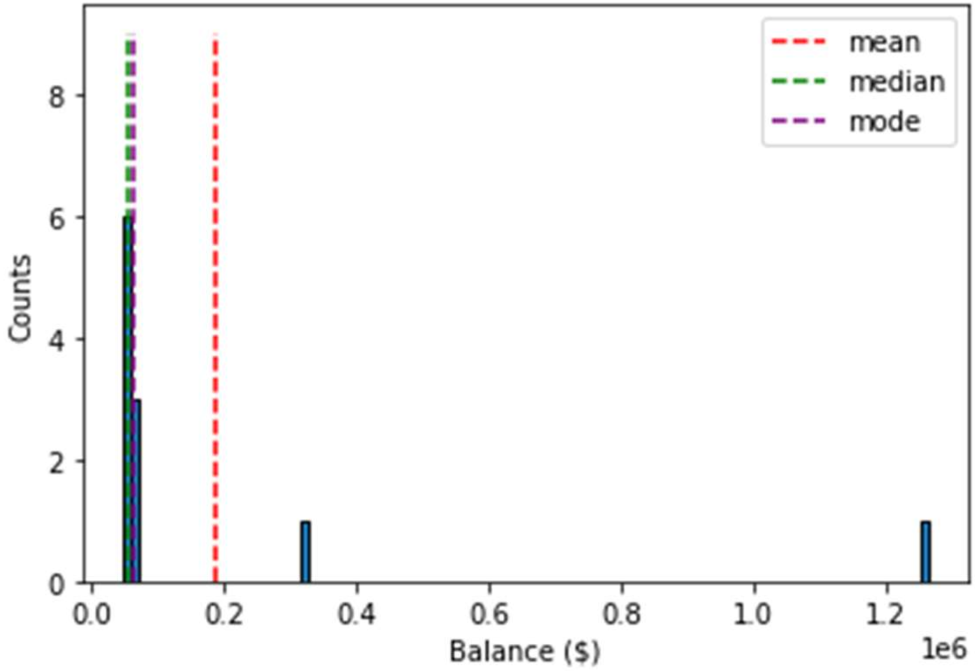
	Population data In statistics, a population is the pool of individuals from which a statistical sample is drawn for a study. Any selection of individuals grouped together by a common feature can be said to be a population.	**Sample data** A sample is a statistically significant portion of a population, not an entire population. The size of the sample is always less than the total size of the population.
mean	$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
variance	$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$	$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$
standard deviation	$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$	$\mathrm{s} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$
covariance	$\sigma_{xy} = \frac{\sum_{i=1}^{N} (x_i - \mu_x)(y_i - \mu_y)}{N}$	$s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$

Measures of central tendency

- **Mean:** sum of values divided by the number of values
- **Median:** the middle number in an ordered dataset
- **Mode:** the most frequent value

Annual income
$ 62,000.00
$ 64,000.00
$ 49,000.00
$ 324,000.00
$ 1,264,000.00
$ 54,330.00
$ 64,000.00
$ 51,000.00
$ 55,000.00
$ 48,000.00
$ 53,000.00

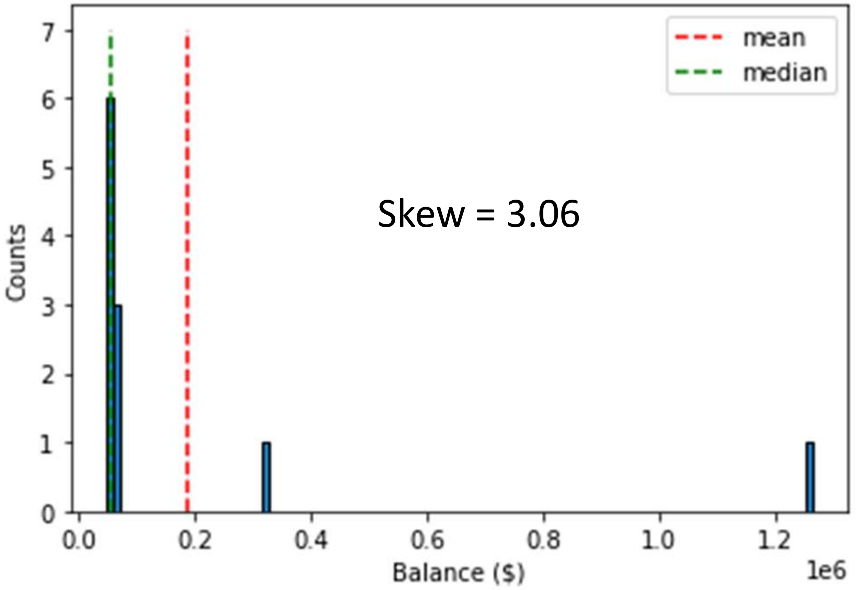
	Annual income
Mean	$ 189,848.18
Median	$ 55,000.00
Mode	$ 64,000.00



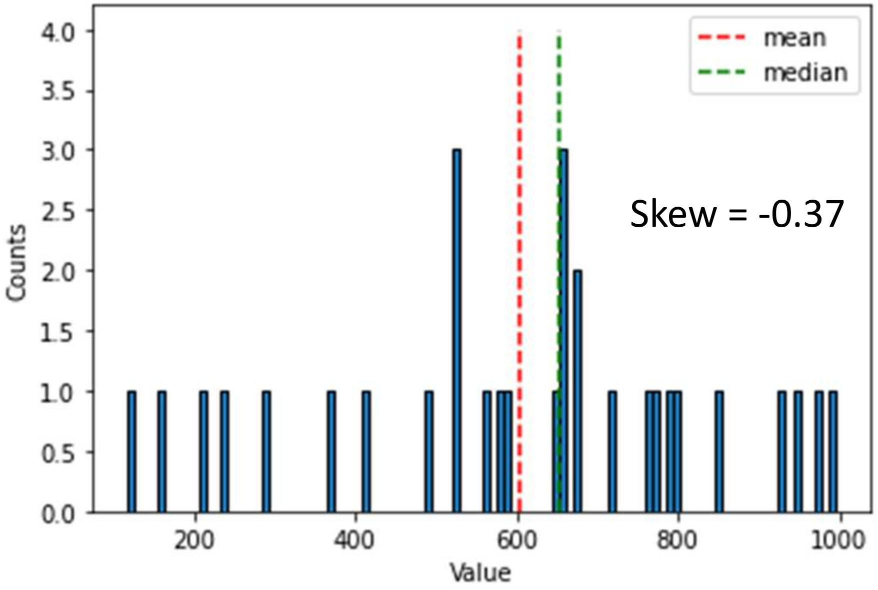
Measures of asimmetry

Skewness indicates whether the data is concentrated on one side.

mean > median → positive skew

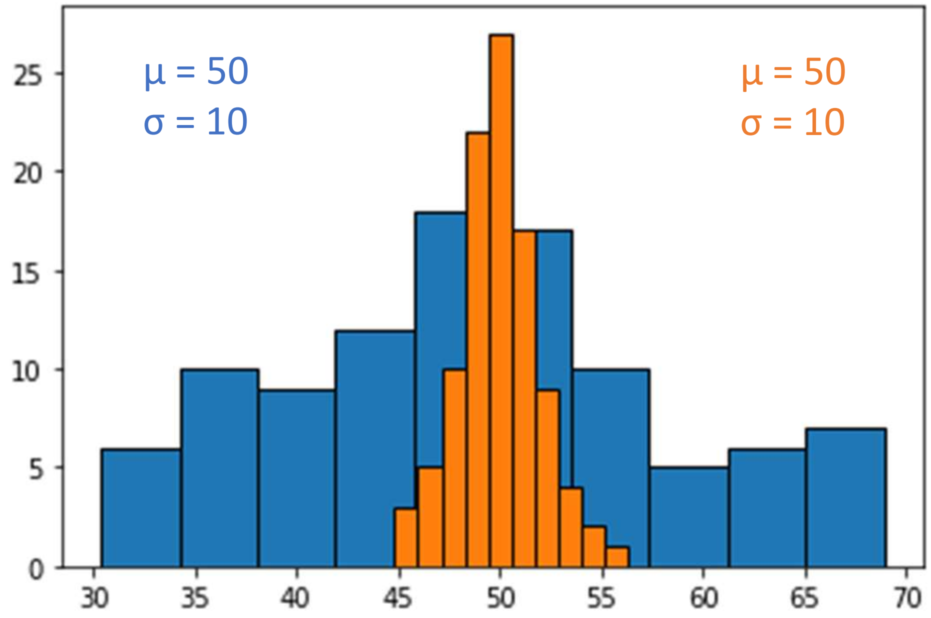


mean < median → positive skew



Measures of variability

	Population	Sample
variance	$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$	$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
standard deviation	$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$	$\mathrm{s} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$



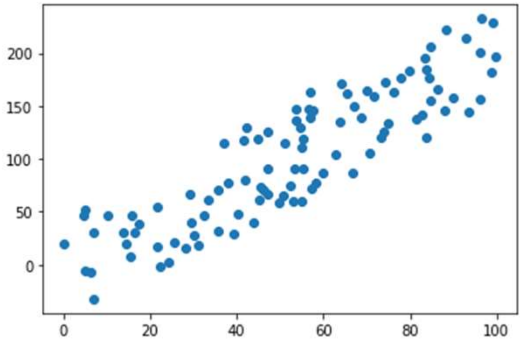
Measures of relationship between variables

Covariance

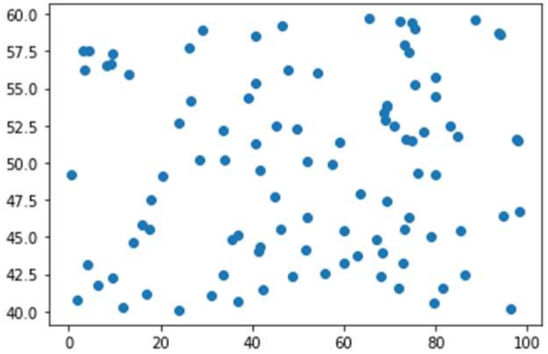
$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Linear correlation coefficient

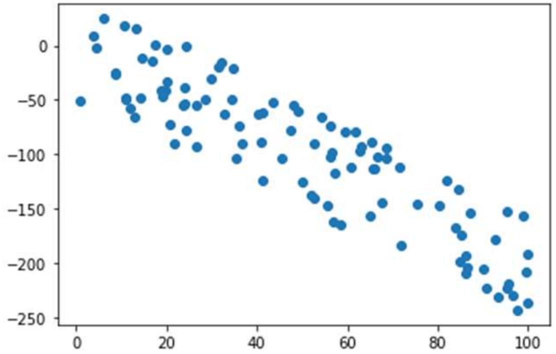
$$corr_{xy} = \frac{s_{xy}}{s_x s_y}$$



corr = 0.89



corr = 0.08

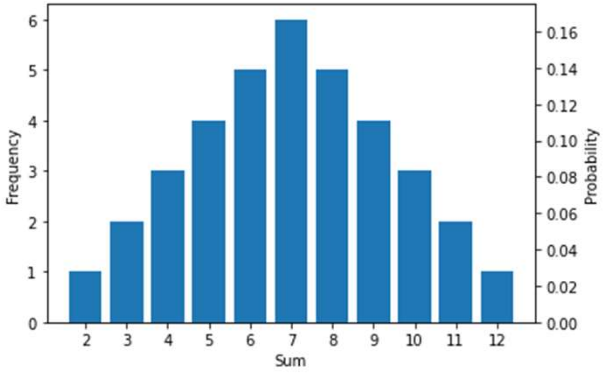


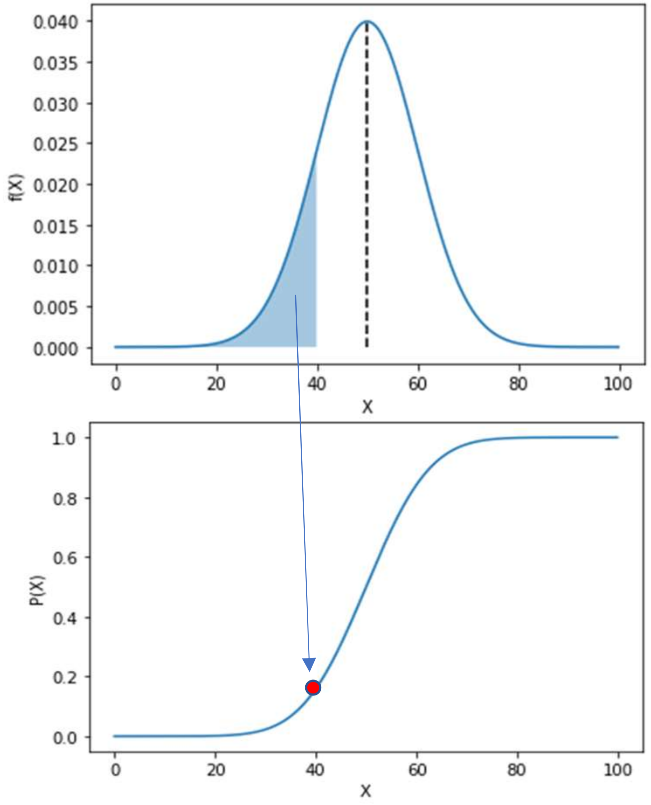
corr = -0.90

Correlation does not imply causation!

Discrete distribution	Continuous distribution
Every unique outcome has a probability assigned to it	We cannot record the frequecy (or probability) of each distinct value
Finite sample space	Infinite sample space

Sum	Frequency	Probability
2	1	0.028
3	2	0.056
4	3	0.083
5	4	0.111
6	5	0.139
7	6	0.167
8	5	0.139
9	4	0.111
10	3	0.083
11	2	0.056
12	1	0.028



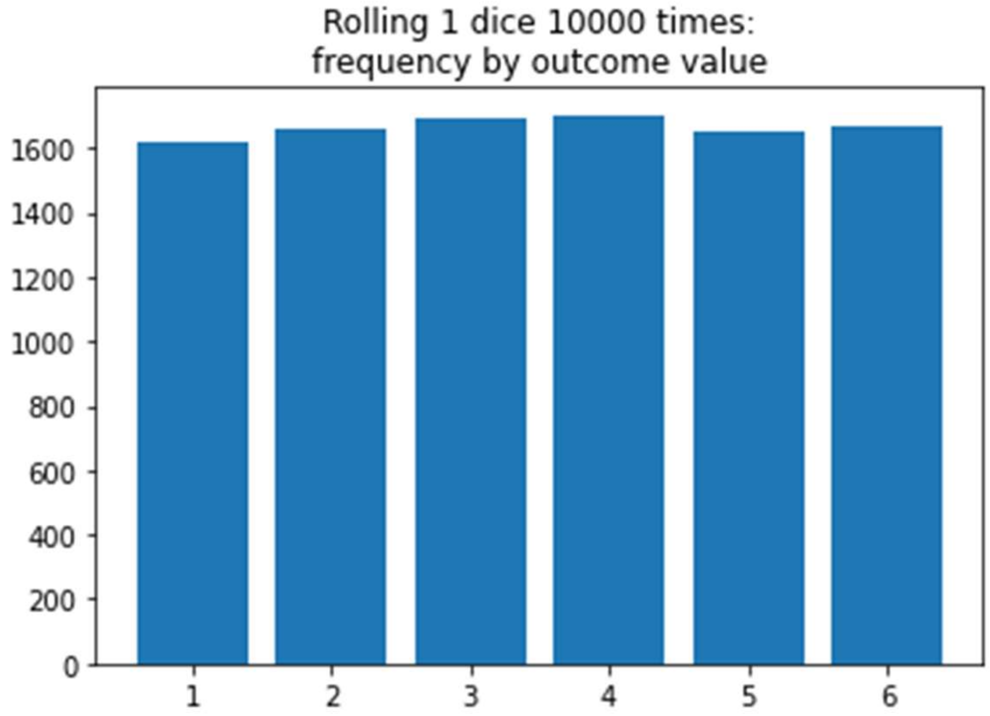


Probability density function

Cumulative distribution function

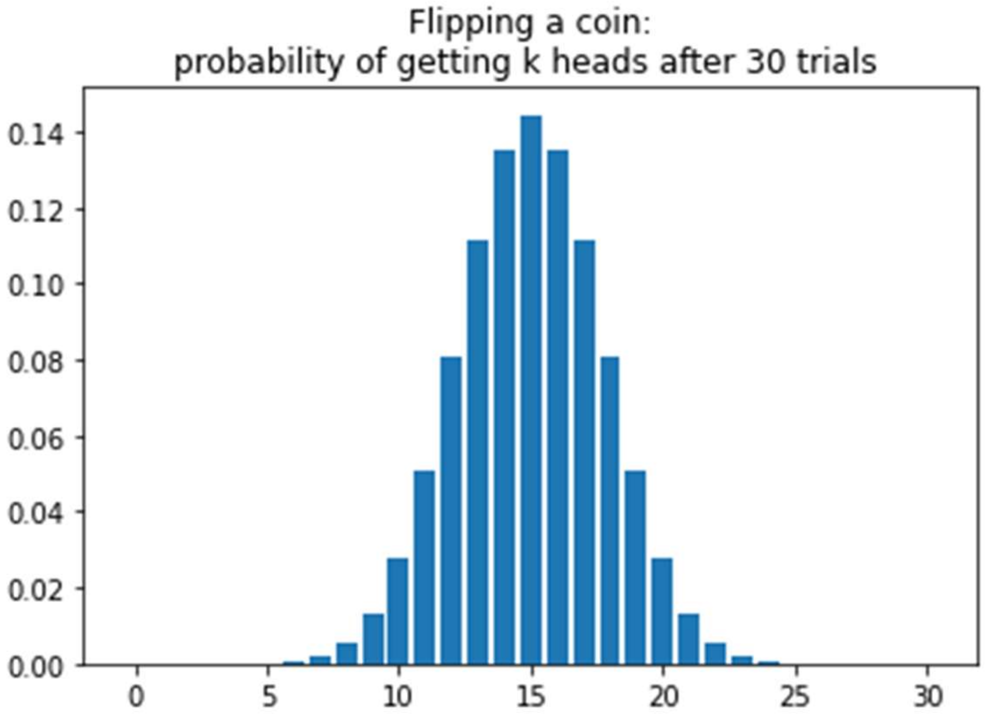
Uniform distribution

describes an experiment where there is an arbitrary outcome that lies between certain bounds.



Binomial distribution

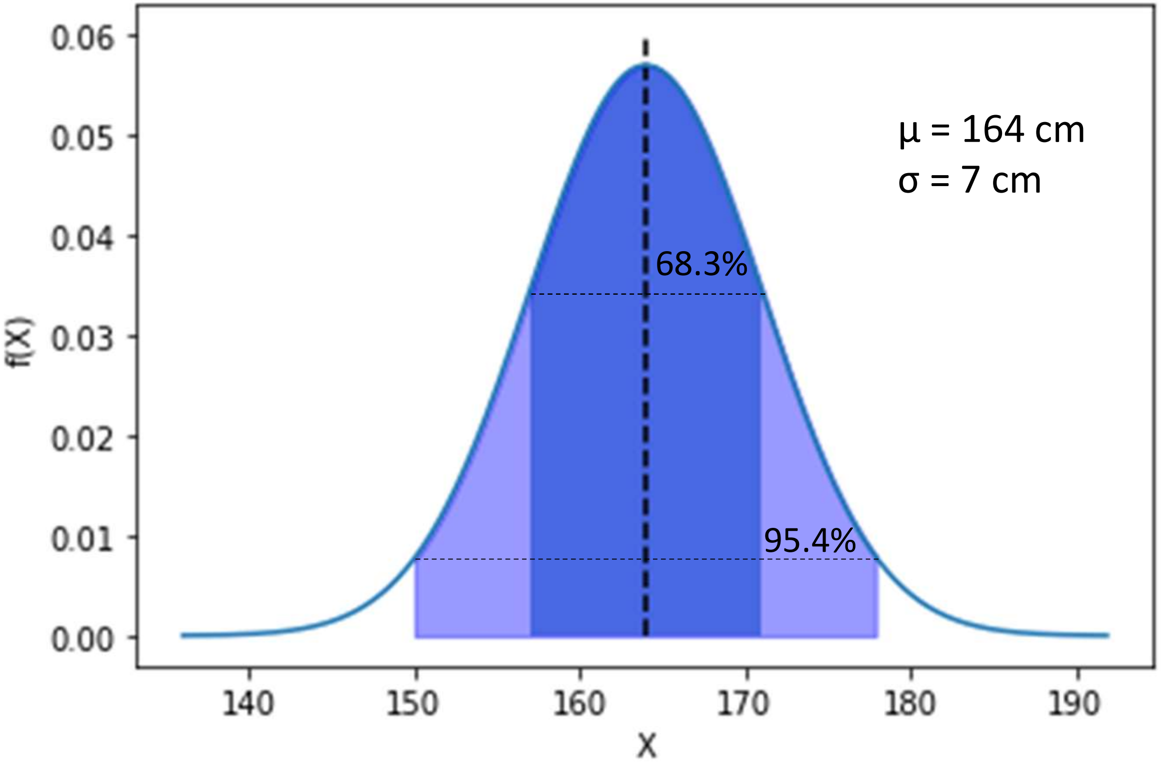
distribution of the possible number of successful outcomes in a given number of trials in each of which there is the same probability of success.



MC9

Normal distribution

Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.



* 2019, height distribution of women born in Italy in 2000

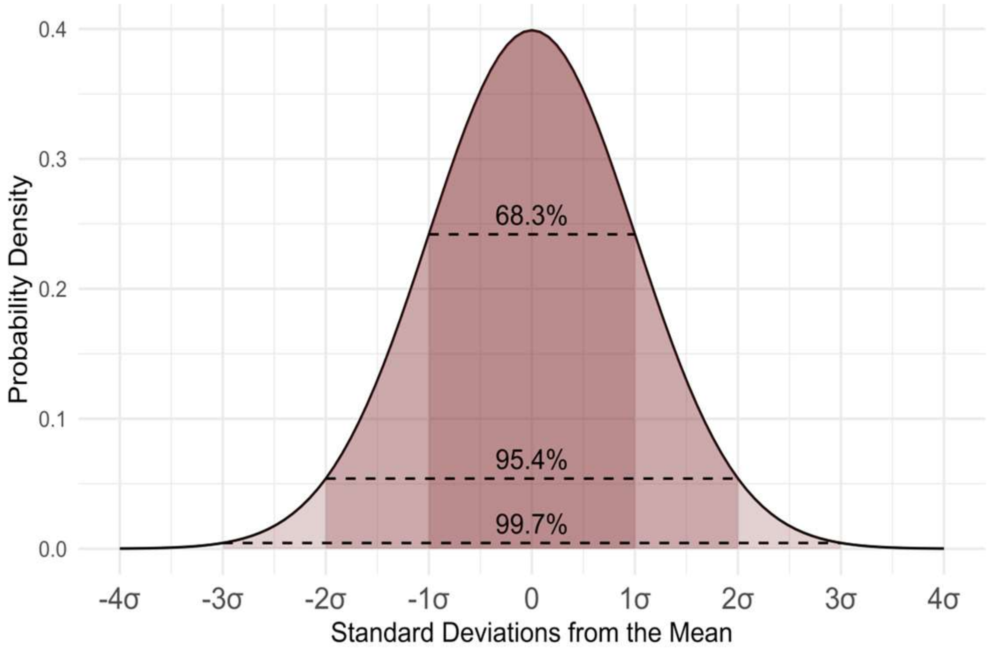
Slide 17

MC9 Independent of the mean and std, same number of std from mean --> same probability

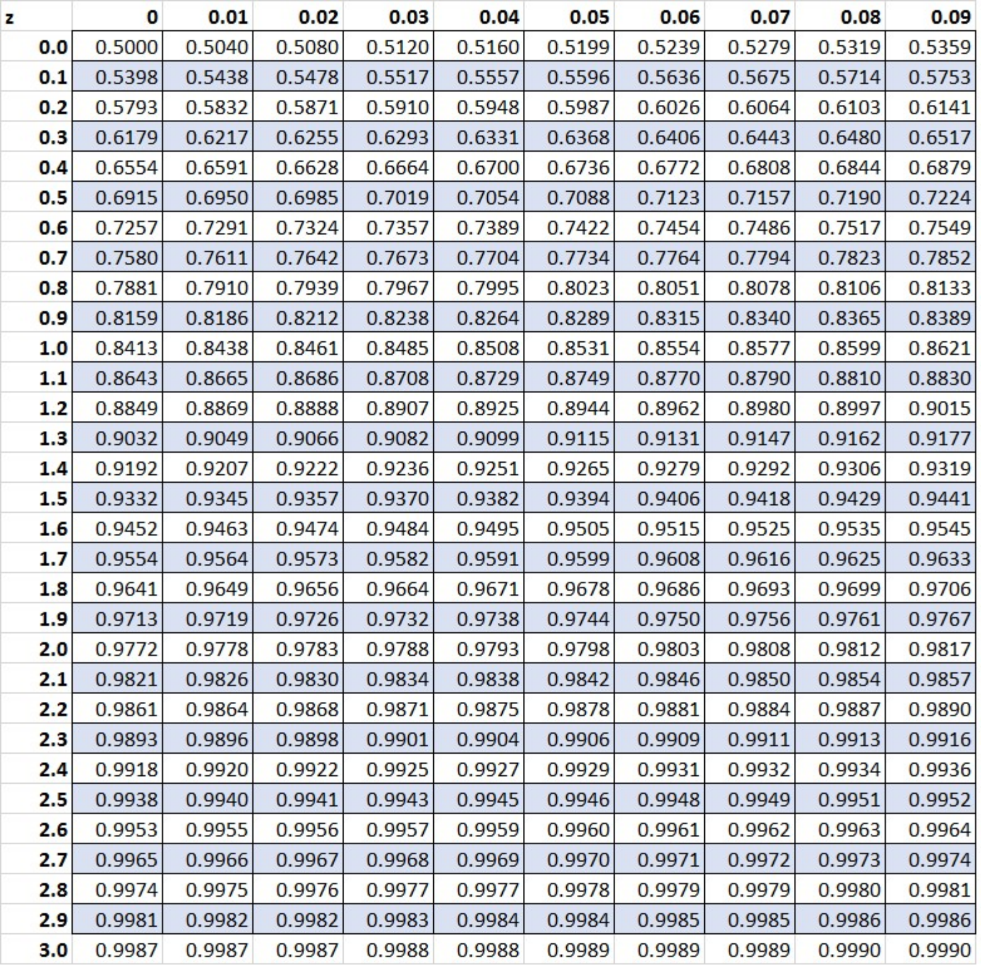
Marco Calbucci, 10/05/2022

Standard normal distribution

$$f(X) \rightarrow f(z), \qquad z = \frac{X - \mu}{\sigma}$$



Confidence interval provides a range that is likely to contain the unknown value and a degree of confidence (**confidence level**) that the unknown value lies within that range.

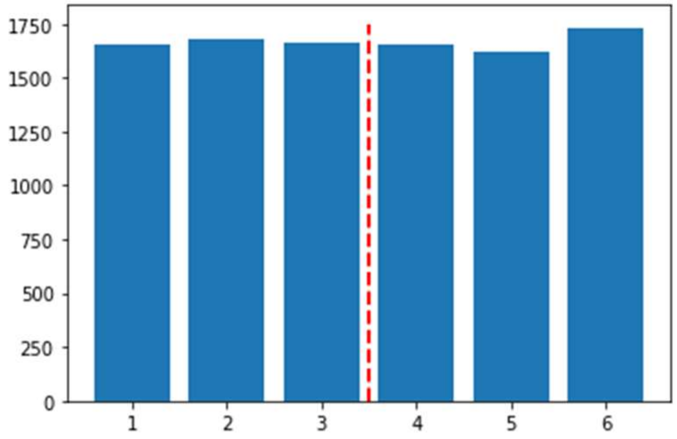


z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Central limit theorem

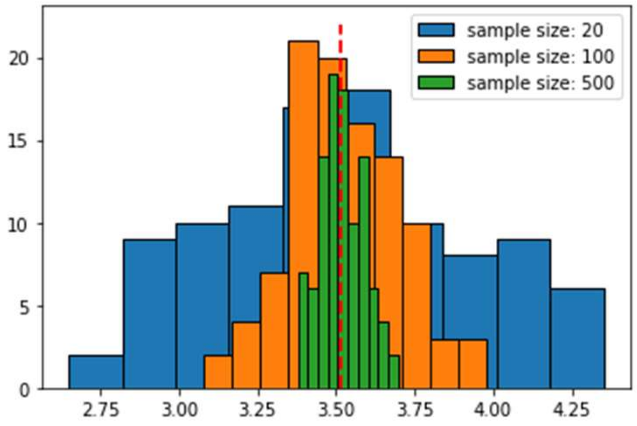
- Given a set of sufficiently large samples drawn from the same population, the means of the samples will be approximately normally distributed.
- The normal distribution will have a mean close to the mean of the population.
- The variance of the sample mean will be close to the variance of the population divided by the sample size.

Population with 10000 elements
following uniform distribution



Take 100 samples with a fixed size and calulate
the average of each sample

Take 100 samples with a fixed size and calulate
the average of each sample



The primary value of the CLT is that it allows us to compute confidence levels and intervals even when the underlying population distribution is not normal.

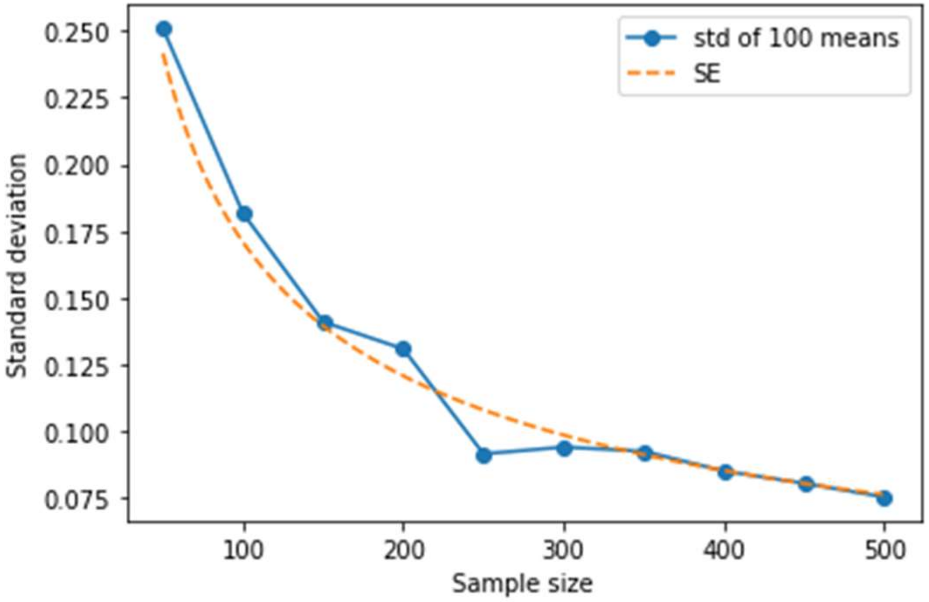
MC10

Standard error

The standard error for a sample size n is the standard deviation of the means of an infinite number of samples of size n drawn from the population.

$$SE = \frac{\sigma}{\sqrt{n}}$$

If all we have is a single sample, we don't know the standard deviation of the population. Typically, we assume that the standard deviation of the sample, is a reasonable proxy for the standard deviation of the population. In practice, people use the standard deviation in place of the unknown population standard deviation to estimate SE.



Slide 20

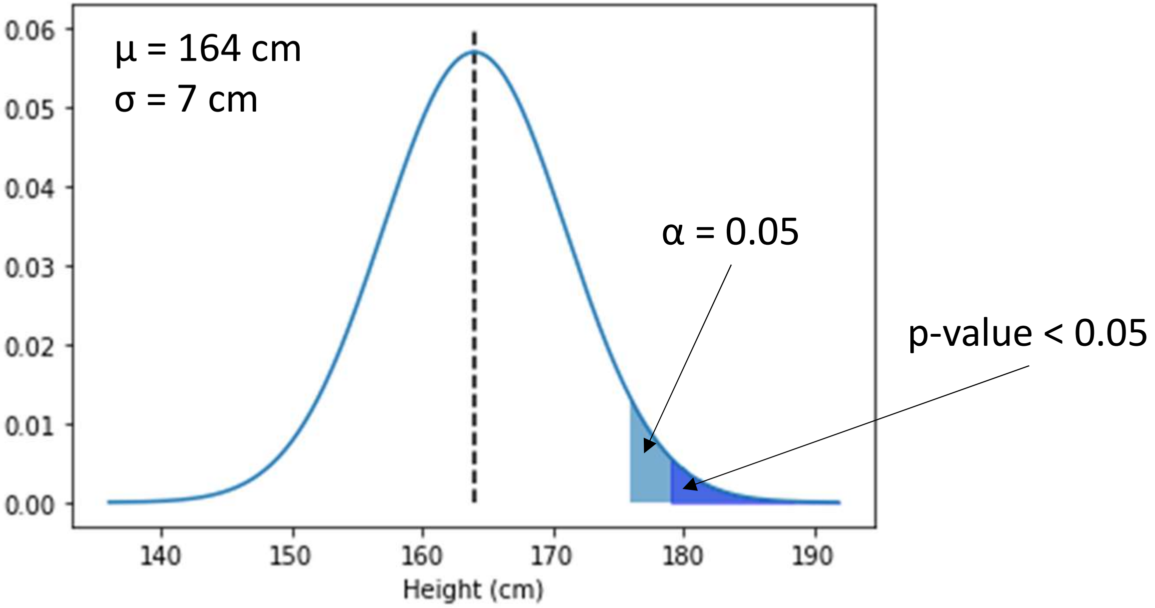
MC10 Esempio maratoneti

Marco Calbucci, 10/05/2022

Hypothesis testing

In any experiment that involves drawing samples at random from a population there is always the possibility that an effect occurred purely by chance. We must set a threshold for **statistical significance α**.

- State a **null hypothesis** and an **alternative hypothesis**.
- Compute the probability (**p-value**) of the **test statistic** under null hypothesis.
- Decide whether that probability is sufficiently small (p-value < α) to reject the null hypothesis.



A/B testing

The process of A/B testing is identical to the process of hypothesis.
It requires analysts to conduct some initial research to understand what is happening and determine what feature needs to be tested. At this point, the analyst can also determine what are the success and tracking metrics because they would have used these statistics to understand the trend of the observations. After this, the hypotheses will be formulated. Without these hypotheses, the testing campaign will be directionless. Next, variations of the testing feature will be randomly assigned to users. Results are then collected and analyzed, and the successful variant will be deployed.

Frequentists address the probability as a measure of the frequency of various outcomes of an experiment. For example, if we have a fair coin (50% probability of landing heads) we expect that the half number of experiments will land on heads.

Bayesians address the probability as an abstract concept that measures a state of knowledge or a degree of belief in a given proposition. This means that probability has a range of values that may be true, rather than a single one.