Mathematical statistics

Normal distribution

Normal Distribution - ND

The normal distribution is the most important probability distribution in statistics because it fits many natural phenomena. It is also known as the Gaussian distribution and the bell curve.

ND

It is a symmetric distribution where most of the observations cluster around the central peak and the probabilities for values further away from the mean taper off equally in both directions. Extreme values in both tails of the distribution are similarly unlikely.

Normal Distribution - ND

Many things closely follow a normal distribution: heights of people, size of things produced by machine, errors in measurements, blood pressure, marks on a test etc.

Normal Distribution - ND

The normal distribution has two parameters, the mean μ and standard deviation σ. The shape of random variable distribution changes based on the parameter values. The mean is the central tendency of the distribution. It defines the location of the peak for normal distributions. Most values cluster around the mean. The standard deviation is a measure of variability. It defines the width of the normal distribution. The standard deviation determines how far away from the mean the values tend to fall. It represents the typical distance between the observations and the average.

ND

A random variable *X* has a normal distribution with parameters *μ* and *σ*

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

if the density function:

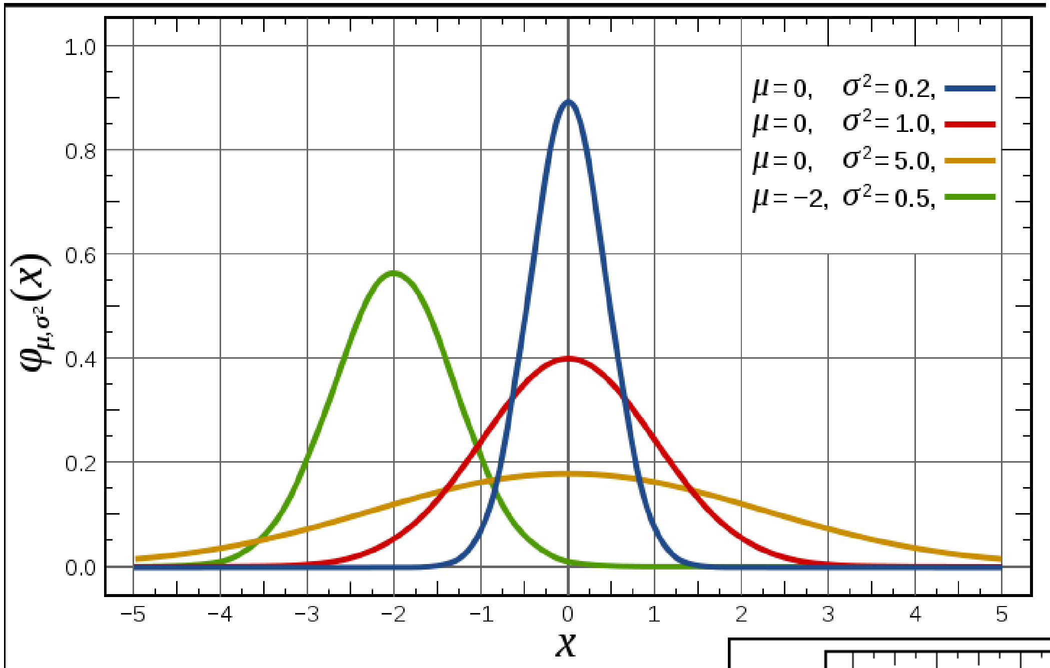
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}, -\infty < x < \infty, \sigma > 0$$

ND, expected value and variance

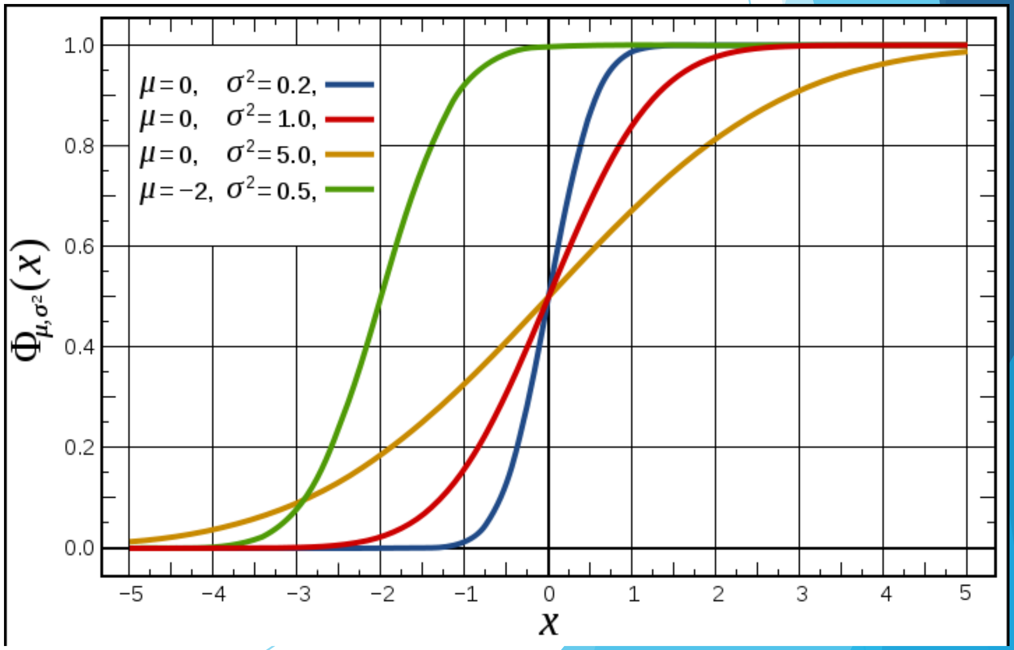
$$F(x) = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{(t-m)^2}{2\sigma^2}}\,dt$$

$$E(x) = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} xe^{-\frac{(x-m)^2}{2\sigma^2}}\,dx = m$$

$$D^2(x) = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x} (x-m)^2 e^{-\frac{(x-m)^2}{2\sigma^2}}\,dx = \sigma^2$$

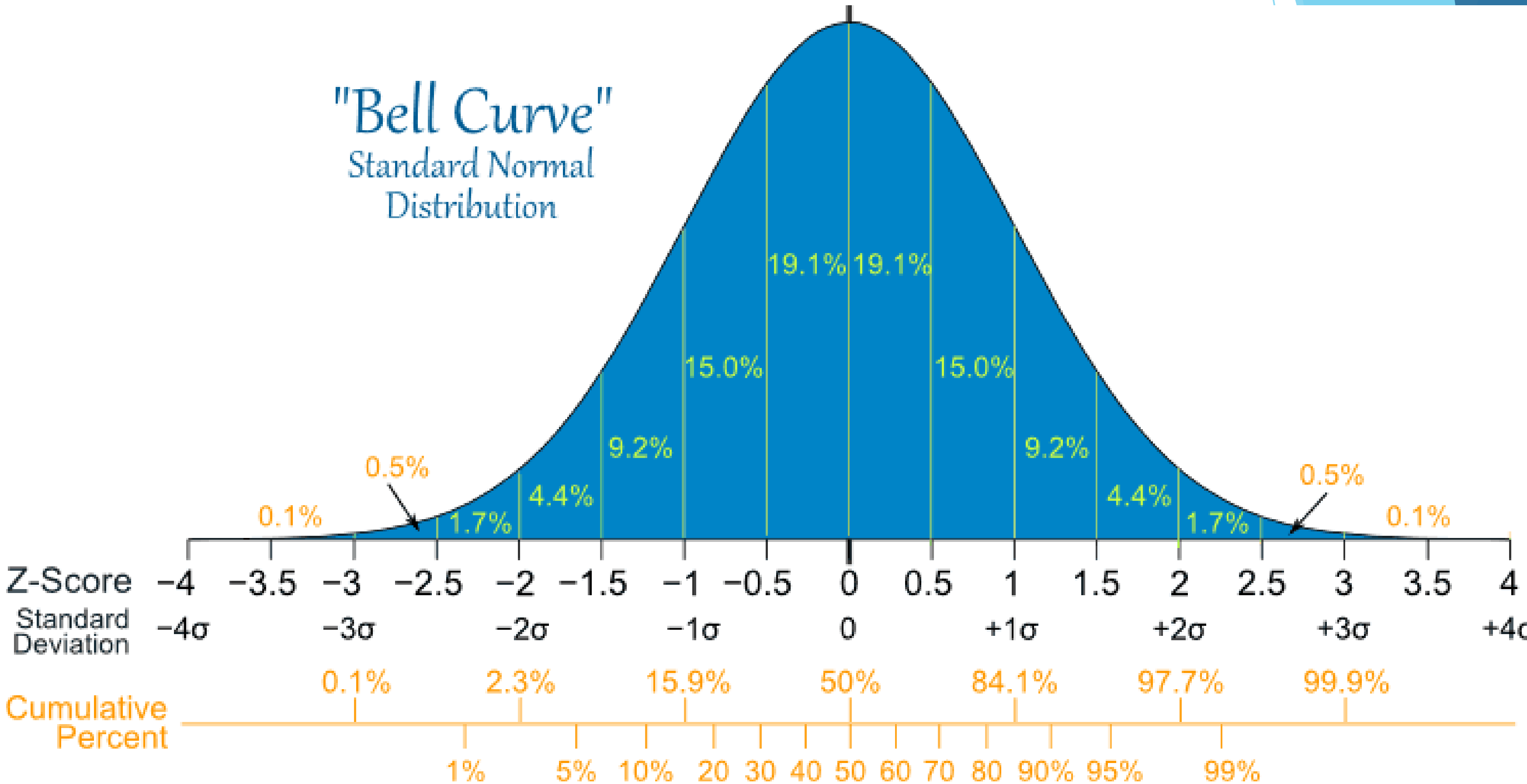


$\mu=0, \quad \sigma^2=0.2,$
$\mu=0, \quad \sigma^2=1.0,$
$\mu=0, \quad \sigma^2=5.0,$
$\mu=-2, \quad \sigma^2=0.5,$
$\varphi_{\mu,\sigma^2}(x)$
1.0
0.8
0.6
0.4
0.2
0.0
−5 −4 −3 −2 −1 0 1 2 3 4 5
x



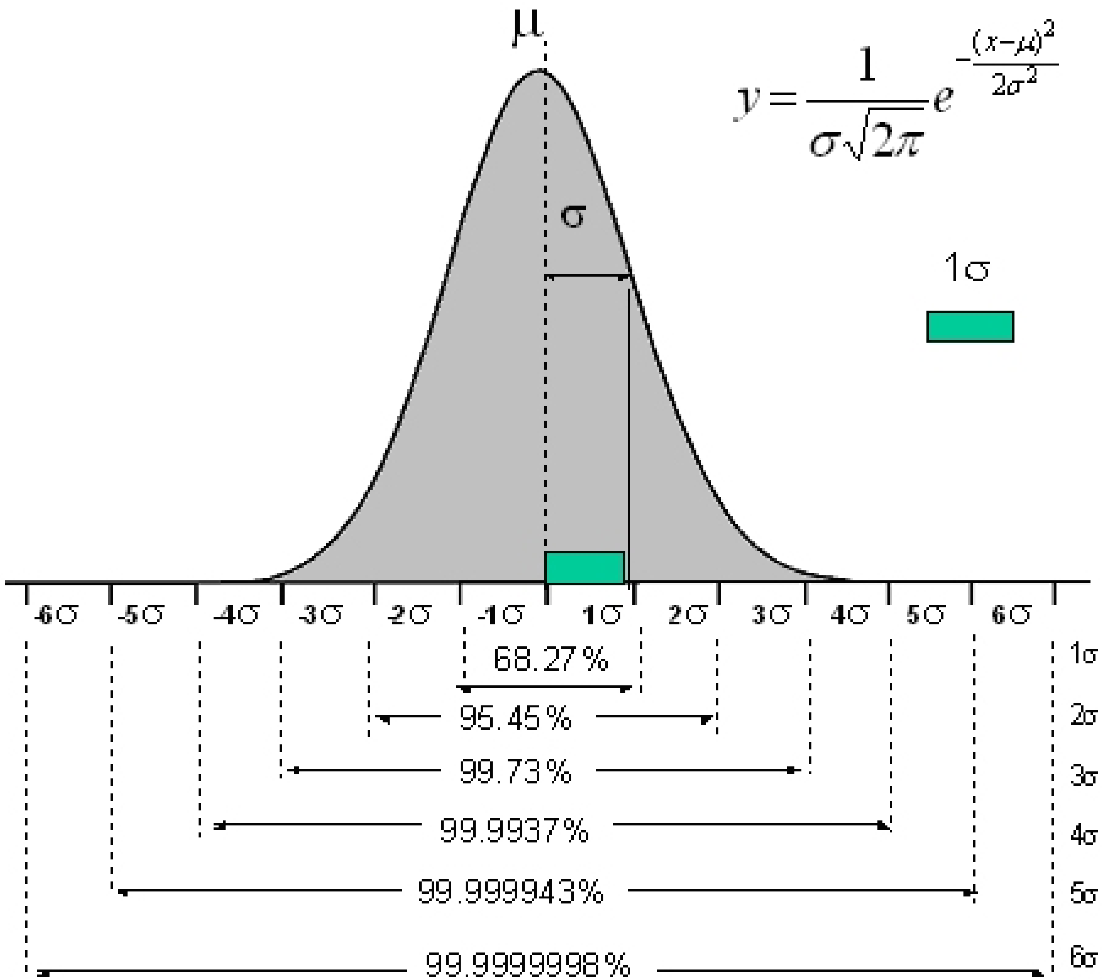
$\mu=0, \quad \sigma^2=0.2,$
$\mu=0, \quad \sigma^2=1.0,$
$\mu=0, \quad \sigma^2=5.0,$
$\mu=-2, \quad \sigma^2=0.5,$
$\Phi_{\mu,\sigma^2}(x)$
1.0
0.8
0.6
0.4
0.2
0.0
−5 −4 −3 −2 −1 0 1 2 3 4 5
x

Standard Normal Distribution



https://www.mathsisfun.com/data/standard-normal-distribution.html

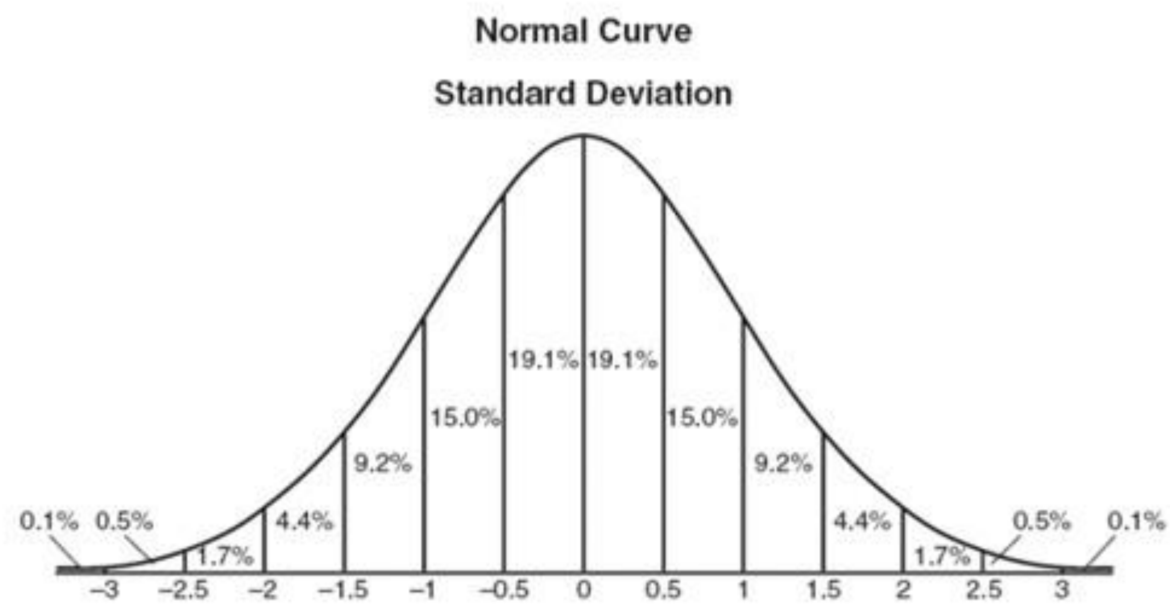
3 sigma rule



http://icpartnership.com/sixsigma.html

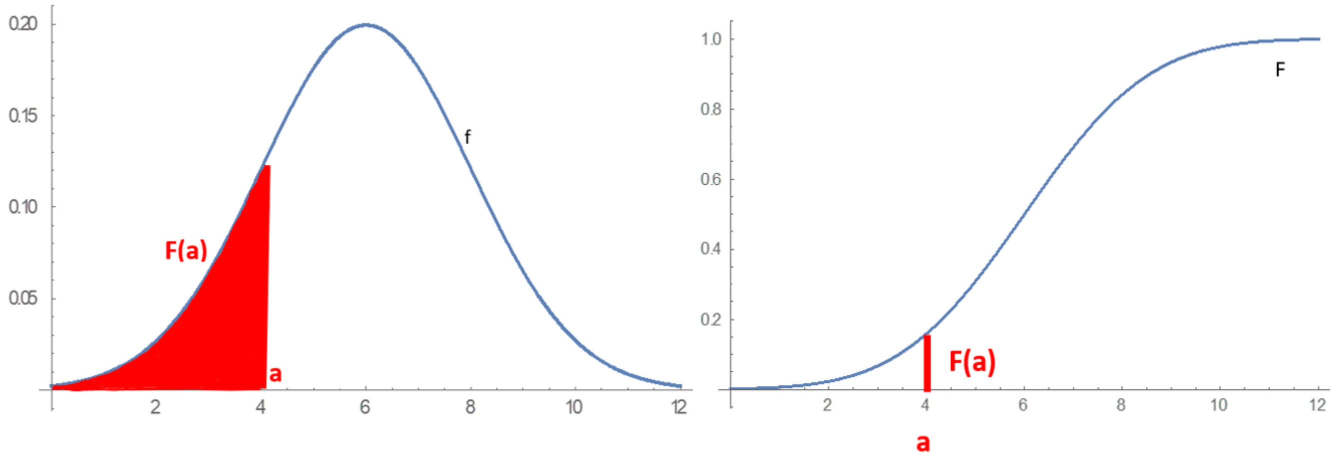
Standard normal distribution $X \sim N(0,1)$

Standard normal distribution is characterized by μ=0 and σ=1.



https://www.mathplanet.com/education/algebra-2/quadratic-functions-and-inequalities/standard-deviation-and-normal-distribution

ND



https://en.wikipedia.org/wiki/Probability_distribution#/media/File:Combined_Cumulative_Distribution_Graphs.png

ND

In statistics, the standard score is the number of standard deviations by which the value of a raw score (i.e., an observed value or data point) is above or below the mean value of what is being observed or measured. Raw scores above the mean have positive standard scores, while those below the mean have negative standard scores.

ND

It is calculated by subtracting the population mean from an individual raw score and then dividing the difference by the population standard deviation. This process of converting a raw score into a standard score is called standardizing.

ND

Standard scores are most commonly called *Z*-scores; the two terms may be used interchangeably. Other terms include *Z*-values, normal scores, and standardized variables.

ND

Computing a *Z*-score requires knowing the mean and standard deviation of the complete population to which a data point belongs.

Standard normal distribution $X \sim N(0,1)$

If the population mean and population standard deviation are known, a raw score x is converted into a standard score by[1]

$$Z = \frac{X - \mu}{\sigma}$$

where:

μ is the mean of the population.

σ is the standard deviation of the population.

Calculating z using this formula requires the population mean and the population standard deviation, not the sample mean or sample deviation.

The absolute value of Z represents the distance between that raw score X and the population mean in units of the standard deviation. **Z** is negative when the raw score is below the mean, positive when above.

ND

Example:

In a population of 60 year old males in whom BMI was normally distributed and had a mean value 29 and a standard deviation 6, what is the probability that a randomly selected male from this population would have a BMI less than 30.

Calculations:

How many standard deviation it is away from the mean:

$$Z = \frac{30 - 29}{6} = 0.17$$

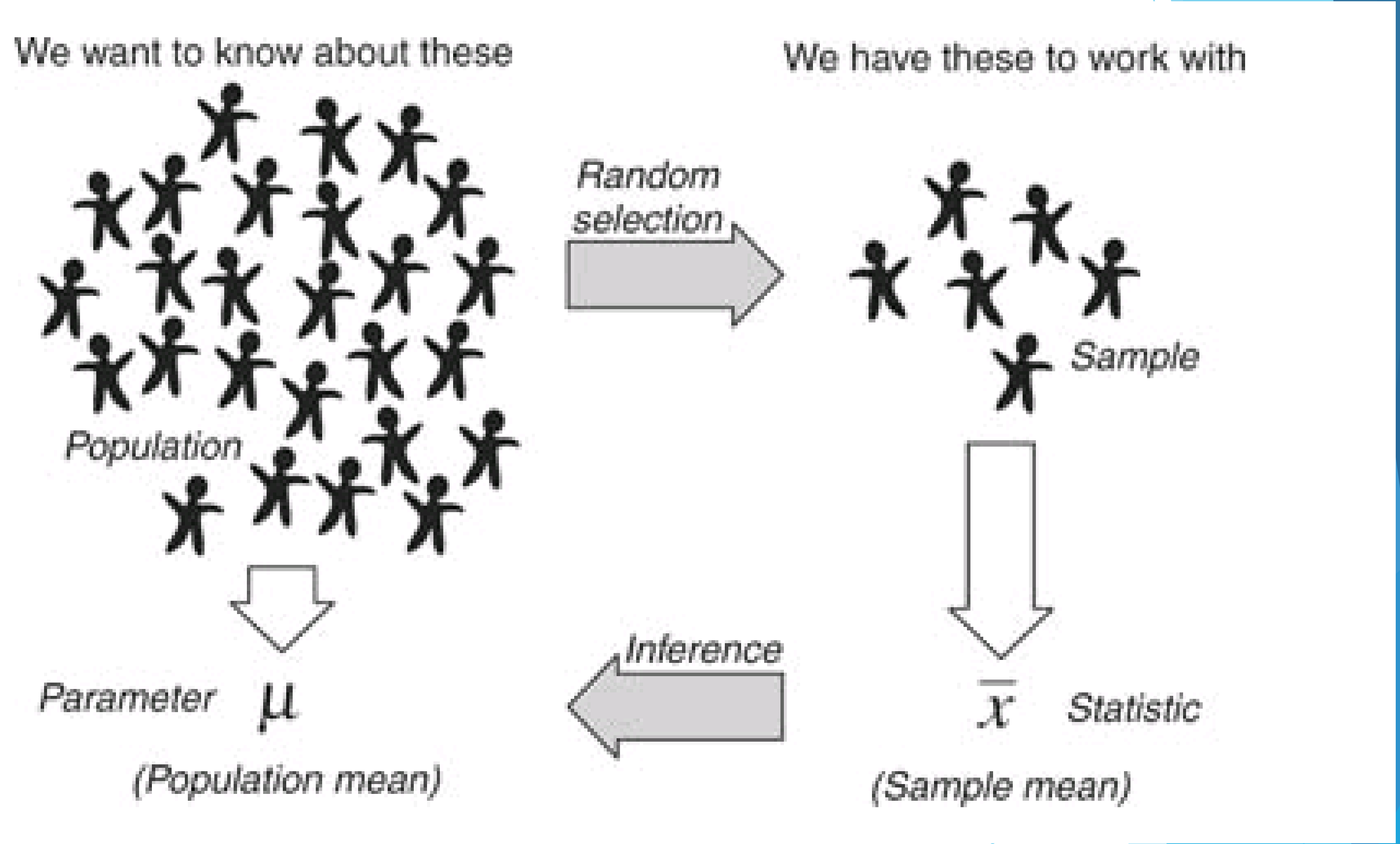
ND

x	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,50000	0,50399	0,50798	0,51197	0,51595	0,51994	0,52392	0,52790	0,53188	0,53586
0,1	0,53983	0,54380	0,54776	0,55172	0,55567	0,55962	0,56356	0,56749	0,57142	0,57535
0,2	0,57926	0,58317	0,58706	0,59095	0,59483	0,59871	0,60257	0,60642	0,61026	0,61409
0,3	0,61791	0,62172	0,62552	0,62930	0,63307	0,63683	0,64058	0,64431	0,64803	0,65173
0,4	0,65542	0,65910	0,66276	0,66640	0,67003	0,67364	0,67724	0,68082	0,68439	0,68793
0,5	0,69146	0,69497	0,69847	0,70194	0,70540	0,70884	0,71226	0,71566	0,71904	0,72240
0,6	0,72575	0,72907	0,73237	0,73565	0,73891	0,74215	0,74537	0,74857	0,75175	0,75490
0,7	0,75804	0,76115	0,76424	0,76730	0,77035	0,77337	0,77637	0,77935	0,78230	0,78524
0,8	0,78814	0,79103	0,79389	0,79673	0,79955	0,80234	0,80511	0,80785	0,81057	0,81327
0,9	0,81594	0,81859	0,82121	0,82381	0,82639	0,82894	0,83147	0,83398	0,83646	0,83891
1,0	0,84134	0,84375	0,84614	0,84849	0,85083	0,85314	0,85543	0,85769	0,85993	0,86214
1,1	0,86433	0,86650	0,86864	0,87076	0,87286	0,87493	0,87698	0,87900	0,88100	0,88298
1,2	0,88493	0,88686	0,88877	0,89065	0,89251	0,89435	0,89617	0,89796	0,89973	0,90147
1,3	0,90320	0,90490	0,90658	0,90824	0,90988	0,91149	0,91308	0,91466	0,91621	0,91774
1,4	0,91924	0,92073	0,92220	0,92364	0,92507	0,92647	0,92785	0,92922	0,93056	0,93189
1,5	0,93319	0,93448	0,93574	0,93699	0,93822	0,93943	0,94062	0,94179	0,94295	0,94408
1,6	0,94520	0,94630	0,94738	0,94845	0,94950	0,95053	0,95154	0,95254	0,95352	0,95449
1,7	0,95543	0,95637	0,95728	0,95818	0,95907	0,95994	0,96080	0,96164	0,96246	0,96327
1,8	0,96407	0,96485	0,96562	0,96638	0,96712	0,96784	0,96856	0,96926	0,96995	0,97062
1,9	0,97128	0,97193	0,97257	0,97320	0,97381	0,97441	0,97500	0,97558	0,97615	0,97670
2,0	0,97725	0,97778	0,97831	0,97882	0,97932	0,97982	0,98030	0,98077	0,98124	0,98169
2,1	0,98214	0,98257	0,98300	0,98341	0,98382	0,98422	0,98461	0,98500	0,98537	0,98574
2,2	0,98610	0,98645	0,98679	0,98713	0,98745	0,98778	0,98809	0,98840	0,98870	0,98899
2,3	0,98928	0,98956	0,98983	0,99010	0,99036	0,99061	0,99086	0,99111	0,99134	0,99158
2,4	0,99180	0,99202	0,99224	0,99245	0,99266	0,99286	0,99305	0,99324	0,99343	0,99361
2,5	0,99379	0,99396	0,99413	0,99430	0,99446	0,99461	0,99477	0,99492	0,99506	0,99520
2,6	0,99534	0,99547	0,99560	0,99573	0,99585	0,99598	0,99609	0,99621	0,99632	0,99643

ND example

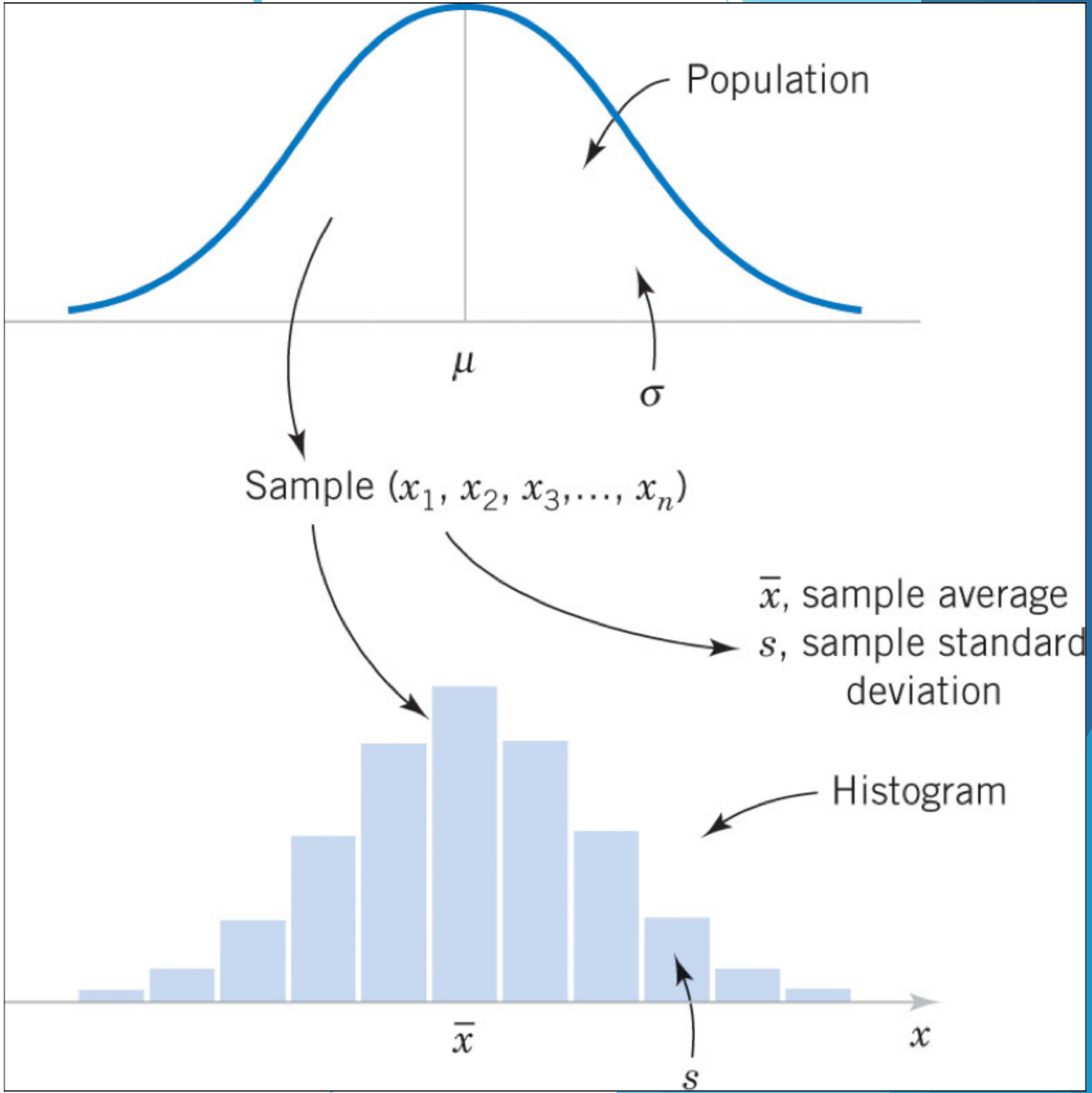
The weight of the orange is a normally distributed random variable with an average of 195.6 g and a variance of 16.3. Calculate the probability that at least one of the four randomly selected fruits will weigh over 200 g.

A population, a sample



https://rchsbowman.wordpress.com/2009/12/08/statistics-notes-sampling-distribution/

A population, a sample



http://cmaps.cmappers.net/rid=1KW8MPZ6V-S2LSQ4-YZ5/Fig%206-3%20Population%20vs%20Sample.png

The likelihood function

The likelihood function measures the goodness of fit of a statistical model to a sample of data for given values of the unknown parameters. It is formed from the joint probability distribution of the sample, but viewed and used as a function of the parameters only, thus treating the random variables as fixed at the observed values. In the estimation process, based on the sample $x_1, x_2, \ldots, x_n$ the parameters describing the assumed probability distribution can be determined. In the estimation process, the parameters should be selected to maximize the probability of the sample used to determine them. The likelihood function can be called the probability product for *n* available samples. Maximum likelihood estimation is a method of estimating the parameters of a probability distribution by maximizing a likelihood function, so that under the assumed statistical model the observed data is most probable.

The likelihood function

For n independent observations x_1 … x_n with the distribution given by the density function p_θ depending on the unknown parameter θ the function:

$$L(x_1, \dots, x_n, \theta) = p_\theta(x_1) \cdot p_\theta(x_2) \dots \cdot p_\theta(x_n)$$

$$l = \ln(L) = \ln\left(\sum_{i=1}^{n} p_\theta(x_1)\right) = \sum_{i=1}^{n} \ln p_\theta(x_i)$$

this is the likelihood function.

The parameter θ is assumed to be for which the likelihood function reaches the highest value.

The likelihood function

Let $x_1 \ldots x_n$ be variables derived from the normal distribution $N\,(\mu, \sigma^2)$ then:

$$L(x_1, \ldots, x_n, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_1-\mu)^2}{2\sigma^2}} \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_2-\mu)^2}{2\sigma^2}} \cdot \cdots \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}}$$

$$L(x_1, \ldots, x_n, \mu, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{\left[\sum_{i=1}^{n}\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right]}$$

$$lnL(x_1, \ldots, x_n, \mu, \sigma) = ln\left[\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{\left[\sum_{i=1}^{n}\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right]}\right]$$

The likelihood function

$$\ln(x \cdot y) = lnx + lny$$

$$lnL(x_1, \dots, x_n, \mu, \sigma) = ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n + lne^{\left[\sum_{i=1}^{n}\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right]}$$

$$\ln e^x = x$$

$$lnL(x_1, \dots, x_n, \mu, \sigma) = ln\left(\sigma\sqrt{2\pi}\right)^{-n} + \sum_{i=1}^{n}\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

a function reaches a minimum when its derivative is zero

The likelihood function

$$lnL(x_1, \ldots, x_n, \mu, \sigma) = ln\left(\sigma\sqrt{2\pi}\right)^{-n} + \sum_{i=1}^{n}\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

$$\frac{d\left(-nln\left(\sigma\sqrt{2\pi}\right) + \frac{-1}{2\sigma^2}\sum_{i=1}^{n}\left((x_i-\mu)^2\right)\right)}{d\mu} = 0$$

$$\frac{d\left[\sum_{i=1}^{n}\left((x_i-\mu)^2\right)\right]}{d\mu} = 0$$

$$\frac{d\left(\sum_{i=1}^{n} x_i^2 - 2\mu\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \mu^2\right)}{d\mu} = 0$$

The likelihood function

$$\frac{d\left(\sum_{i=1}^{n} x_i^2 - 2\mu \sum_{i=1}^{n} x_i + n\mu^2\right)}{d\mu} = 0$$

$$0 - 2\sum_{i=1}^{n} x_i + n2\mu = 0$$

$$n\mu = \sum_{i=1}^{n} x_i$$

μ estimator is:

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

The likelihood function

$$lnL(x_1, \dots, x_n, \mu, \sigma) = ln\left(\sqrt{2\pi\sigma^2}\right)^{-n} + \sum_{i=1}^{n}\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

$$\frac{d\left(-nln(2\pi\sigma^2)^{\frac{1}{2}} + \frac{-1}{2\sigma^2}\sum_{i=1}^{n}\left((x_i-\mu)^2\right)\right)}{d\sigma^2} = 0$$

$$\frac{d\left(-\frac{1}{2}nln\sigma^2 - \frac{1}{2}nln\left(\sqrt{2\pi}\right) + \frac{-1}{2\sigma^2}\sum_{i=1}^{n}\left((x_i-\mu)^2\right)\right)}{d\sigma^2} = 0$$

$$-\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2}\frac{1}{\sigma^4}\sum_{i=1}^{n}\left((x_i-\mu)^2\right) = 0$$

The likelihood function

$$-n + \frac{1}{\sigma^2}\sum_{i=1}^{n}\left((x_i - \mu)^2\right) = 0$$

Estymator of σ^2 is:

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

Confidence intervals

Confidence Intervals - CI

- Estimation is the estimation of values such as the mean, standard deviation, variance, fractions for the entire population based on a sample.
- Estimation allows for the generalization of the collected results from the sample to the entire population.

CI

Point estimation

μ estimator is:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Estymator of σ^2 is:

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

CI

fraction (percentage of the population that meets the given condition)

k - number of favorable events

n - number of all events

$$\bar{p} = \frac{k}{n}$$

CI

- The confidence interval for a given statistical measure informs that the real value sought is within a certain interval with the assumed probability.

Example for the mean:

- tests on a sample provide an average value of a certain feature, on its basis, it is possible to determine a confidence interval in which the value of the average for the entire population falls with the assumed probability

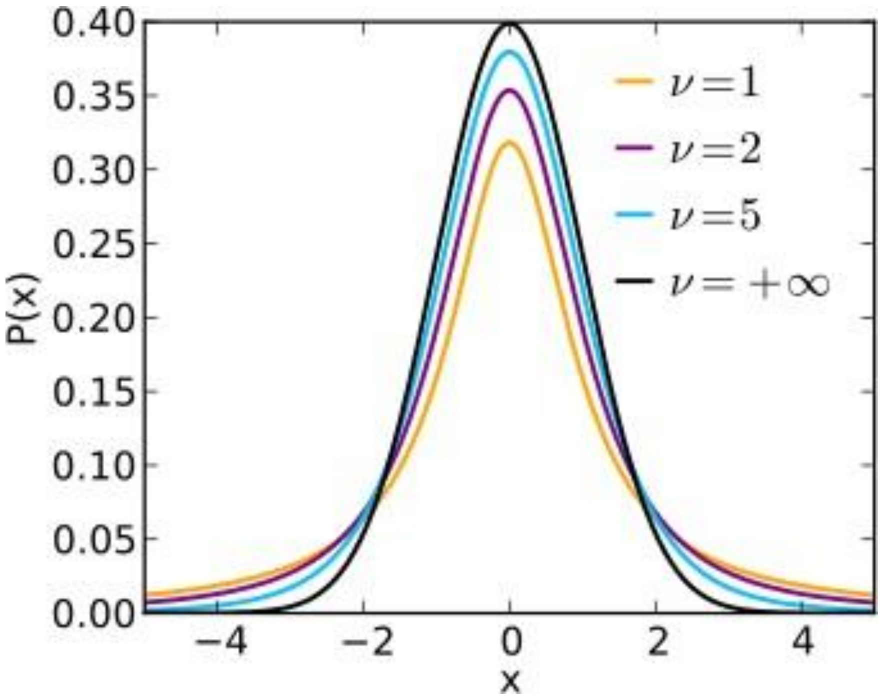
CI

Useful continuous distributions:

t-Student distribution

chi square distribution

T-Student's



http://en.wikipedia.org/wiki/Student%27s_t-distribution#mediaviewer/File:Student_t_pdf.svg

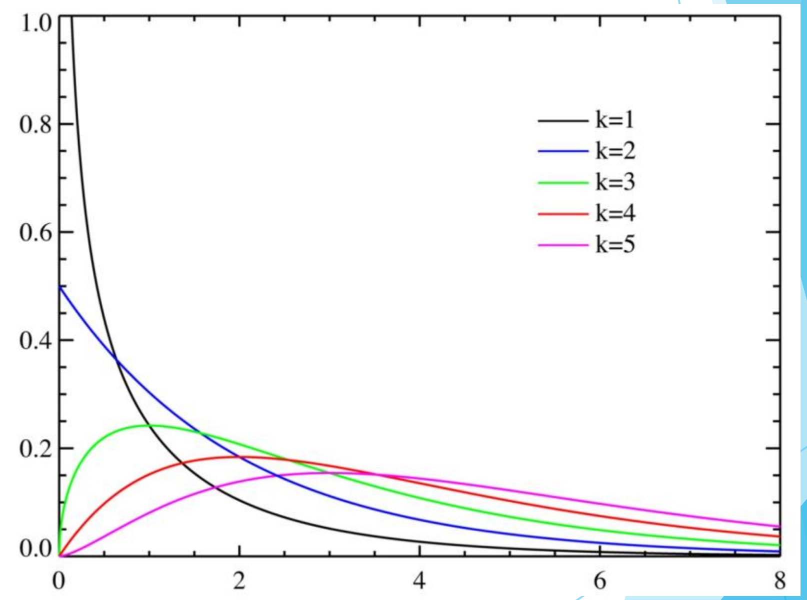
χ^2

Distribution χ2 (chi square) - distribution of a random variable, which is the sum of k squares of independent random variables with a standard normal distribution. A natural number k is called the number of degrees of freedom in the distribution of a random variable.

$$k = 1 \qquad X^2 \qquad X \sim N(0,1)$$

$$k = 2 \qquad X_1^2 + X_2^2 \qquad X \sim N(0,1)$$

$$k = 3 \qquad X_1^2 + X_2^2 + X_3^2 \qquad X \sim N(0,1)$$



https://pl.wikipedia.org/wiki/Rozk%C5%82ad_chi_kwadrat#/media/Plik:Chi-square_distributionPDF.png

Confidence interval μ

- X~N(μ, σ^2), μ, σ^2 unknown

$$\mu \epsilon \left\langle \bar{x} - t_{\alpha,\nu} \frac{s}{\sqrt{n}} ; \bar{x} + t_{\alpha,\nu} \frac{s}{\sqrt{n}} \right\rangle$$

Confidence interval σ^2

- $X \sim N(\mu, \sigma^2)$, μ, σ^2 unknown

$$\sigma^2 \epsilon \left\langle \frac{s^2(n-1)}{\chi^2_{\frac{\alpha}{2},\nu}} ; \frac{s^2(n-1)}{\chi^2_{1-\frac{\alpha}{2},\nu}} \right\rangle$$

Confidence interval *p*

- X~B(*n*, *p*), *p* unknown

$$p \in \left\langle \bar{p} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}; \bar{p} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \right\rangle$$

Confidence interval for difference of means

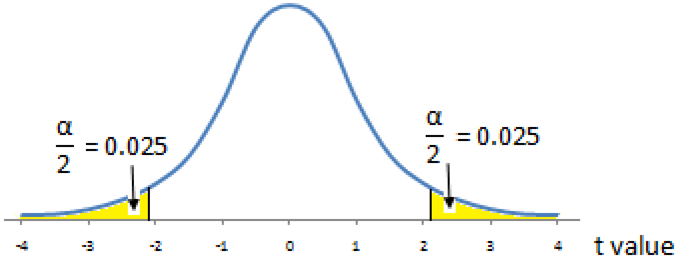
$$\mu_1 - \mu_2 \in \langle (\bar{x}_1 - \bar{x}_2) - t_{\alpha,\nu} \cdot s_r ; (\bar{x}_1 - \bar{x}_2) + t_{\alpha,\nu} \cdot s_r \rangle$$

$$s_r = \sqrt{s_e^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

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

Student's t Distribution Table

For example, the t value for 18 degrees of freedom is 2.101 for 95% confidence interval (**2-Tail** α = 0.05).



	90%	95%	97.5%	99%	99.5%	99.95%	1-Tail Confidence Level
	80%	90%	95%	98%	99%	99.9%	2-Tail Confidence Level
	0.100	0.050	0.025	0.010	0.005	0.0005	1-Tail Alpha
df	0.20	0.10	0.05	0.02	0.01	0.001	2-Tail Alpha
1	3.0777	6.3138	12.7062	31.8205	63.6567	636.6192	
2	1.8856	2.9200	4.3027	6.9646	9.9248	31.5991	
3	1.6377	2.3534	3.1824	4.5407	5.8409	12.9240	
4	1.5332	2.1318	2.7764	3.7469	4.6041	8.6103	
5	1.4759	2.0150	2.5706	3.3649	4.0321	6.8688	
6	1.4398	1.9432	2.4469	3.1427	3.7074	5.9588	
7	1.4149	1.8946	2.3646	2.9980	3.4995	5.4079	
8	1.3968	1.8595	2.3060	2.8965	3.3554	5.0413	
9	1.3830	1.8331	2.2622	2.8214	3.2498	4.7809	
10	1.3722	1.8125	2.2281	2.7638	3.1693	4.5869	
11	1.3634	1.7959	2.2010	2.7181	3.1058	4.4370	
12	1.3562	1.7823	2.1788	2.6810	3.0545	4.3178	
13	1.3502	1.7709	2.1604	2.6503	3.0123	4.2208	
14	1.3450	1.7613	2.1448	2.6245	2.9768	4.1405	
15	1.3406	1.7531	2.1314	2.6025	2.9467	4.0728	
16	1.3368	1.7459	2.1199	2.5835	2.9208	4.0150	
17	1.3334	1.7396	2.1098	2.5669	2.8982	3.9651	
18	1.3304	1.7341	2.1009	2.5524	2.8784	3.9216	
19	1.3277	1.7291	2.0930	2.5395	2.8609	3.8834	

http://growingknowing.com/GKStatsBookStudentTTable.html

Critical values for χ^2

v \ a	0,995	0,990	0,975	0,950	0,900	0,100	0,050	0,025	0,010	0,005
1	$0{,}0^4393$	0,0002	0,0010	0,0039	0,0158	2,7055	3,8415	5,0239	6,6349	7,8794
2	0,0100	0,0201	0,0506	0,1026	0,2107	4,6052	5,9915	7,3778	9,2104	10,5965
3	0,0717	0,1148	0,2158	0,3518	0,5844	6,2514	7,8147	9,3484	11,3449	12,8381
4	0,2070	0,2971	0,4844	0,7107	1,0636	7,7794	9,4877	11,1433	13,2767	14,8602
5	0,4118	0,5543	0,8312	1,1455	1,6103	9,2363	11,0705	12,8325	15,0863	16,7496
6	0,6757	0,8721	1,2373	1,6354	2,2041	10,6446	12,5916	14,4494	16,8119	18,5475
7	0,9893	1,2390	1,6899	2,1673	2,8331	12,0170	14,0671	16,0128	18,4753	20,2777
8	1,3444	1,6465	2,1797	2,7326	3,4895	13,3616	15,5073	17,5345	20,0902	21,9549
9	1,7349	2,0879	2,7004	3,3251	4,1682	14,6837	16,9190	19,0228	21,6660	23,5893
10	2,1558	2,5582	3,2470	3,9403	4,8652	15,9872	18,3070	20,4832	23,2093	25,1881
11	2,6032	3,0535	3,8157	4,5748	5,5778	17,2750	19,6752	21,9200	24,7250	26,7569
12	3,0738	3,5706	4,4038	5,2260	6,3038	18,5493	21,0261	23,3367	26,2170	28,2997
13	3,5650	4,1069	5,0087	5,8919	7,0415	19,8119	22,3620	24,7356	27,6882	29,8193
14	4,0747	4,6604	5,6287	6,5706	7,7895	21,0641	23,6848	26,1189	29,1412	31,3194
15	4,6009	5,2294	6,2621	7,2609	8,5468	22,3071	24,9958	27,4884	30,5780	32,8015
16	5,1422	5,8122	6,9077	7,9616	9,3122	23,5418	26,2962	28,8453	31,9999	34,2671

CI - example

It can be assumed that the skull length (in mm) in the population of Tatra chamois has a normal distribution with unknown parameters. Calculations were made for a random sample of 15 skulls and the result was:

$$\sum_i x_i = 2971.5 \quad \sum_i x_i^2 = 591888.71$$

Determine the following characteristics of the studied trait in the population:

average score,

scoring of variance,

standard deviation score,

95% and 99% confidence interval for the mean,

95% and 99% confidence intervals for the variance,

95% and 99% confidence intervals for the standard deviation.

CI fraction - example

10,000 butterflies, including 5,433 females, were caught. Give an estimate of the female fraction in the butterfly population:

95% and 99% confidence interval

Libre Calc

Critical values

- =T.INV.2T(0.05,14)
- =T.INV.2T(0.01,14)
- =NORM.S.INV(0.95)
- =NORM.S.INV(0.99)