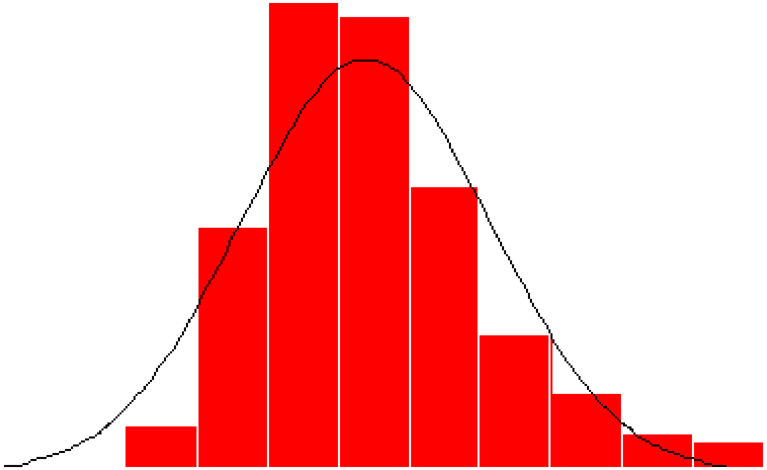
Biostatistics

Author: **Krisztina Boda** PhD

University of Szeged
Department of Medical Physics and Informatics

www.model.u-szeged.hu
www.szote.u-szeged.hu/dmi

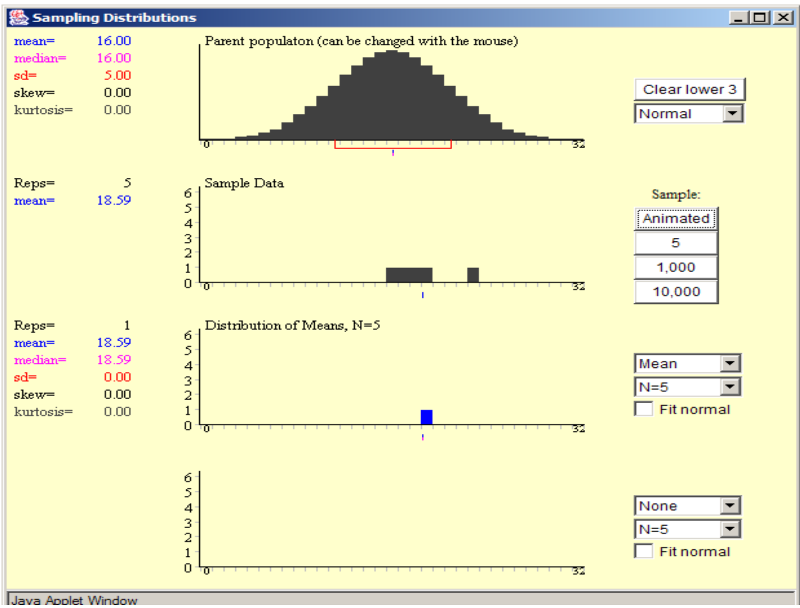
Statistical estimation, confidence intervals

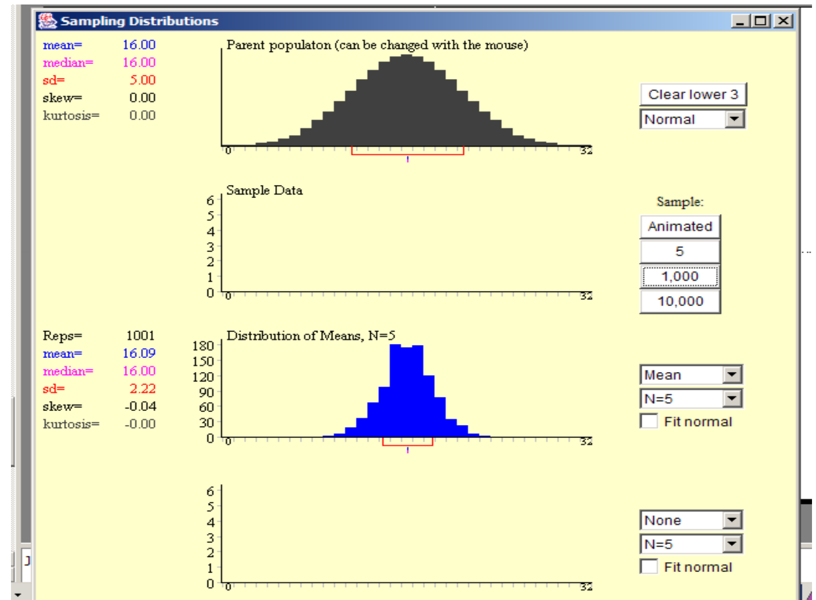


The central limit theorem

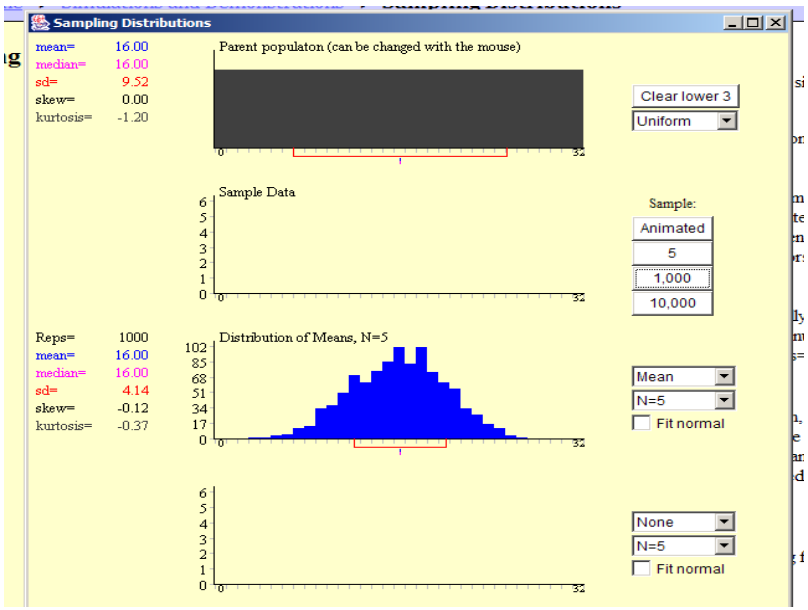
Distribution of sample means

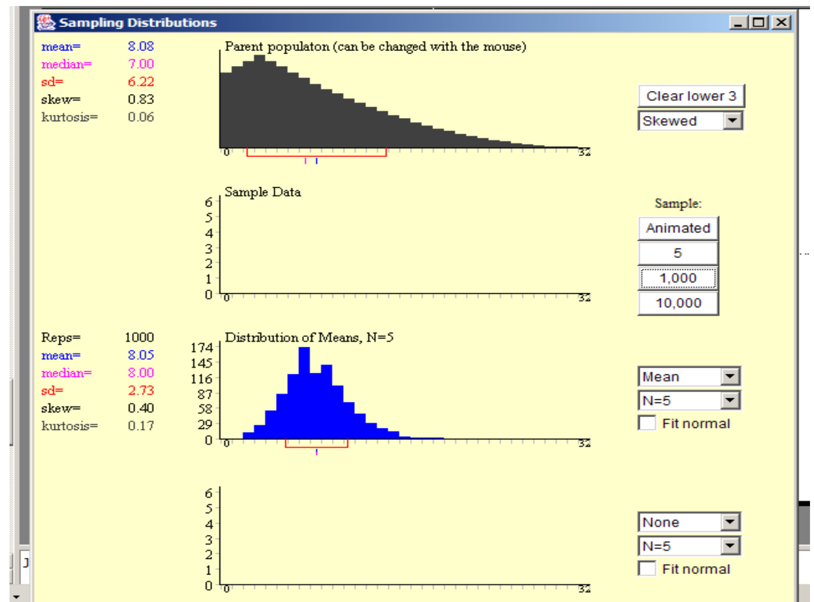
http://onlinestatbook.com/stat_sim/sampling_dist/index.html





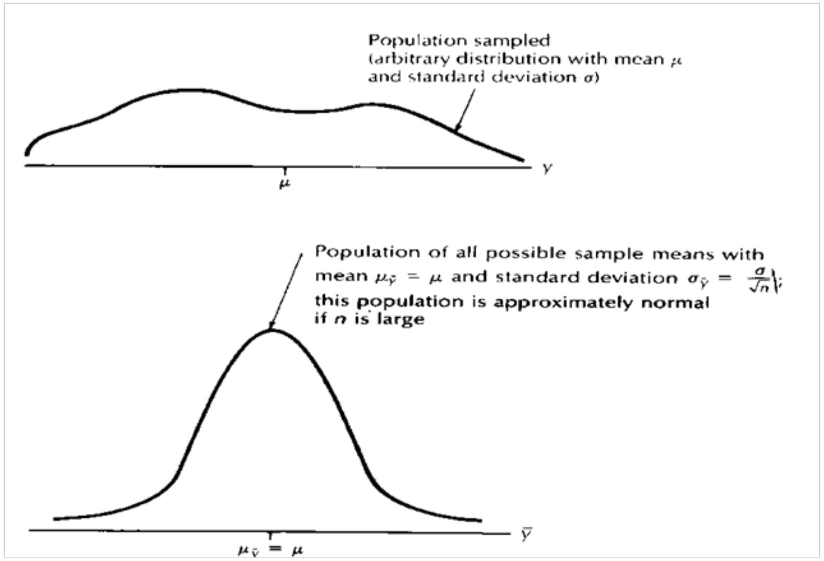
The population is not normally distributed





The central limit theorem

- If the sample size *n* is large (say, at least 30), then the population of all possible sample means approximately has a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$ no matter what probability describes the population sampled



The prevalence of normal distribution

- Since real-world quantities are often the balanced sum of many unobserved random events, this theorem provides a partial explanation for the prevalence of the normal probability distribution.

Tha standard error of mean (SE or SEM)

- $\frac{\sigma}{\sqrt{n}}$ is called the standard error of mean
- Meaning: the dispersion of the sample means around the (unknown) population mean.

Calculation of the standard error from the standard deviation when σ is unknown

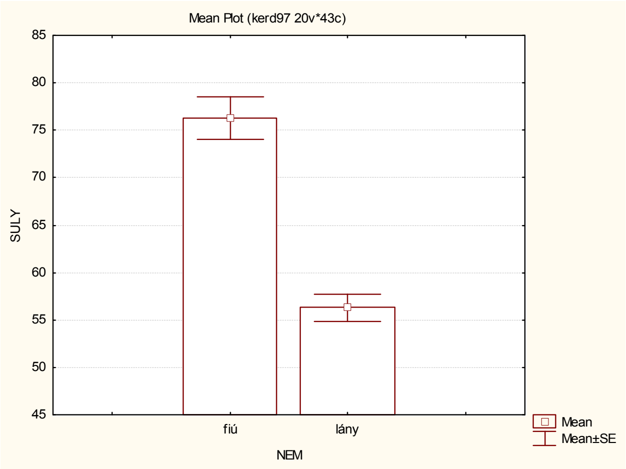
- Given $x_1, x_2, x_3, \ldots, x_n$ statistical sample, the stadard error can be calculated by

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

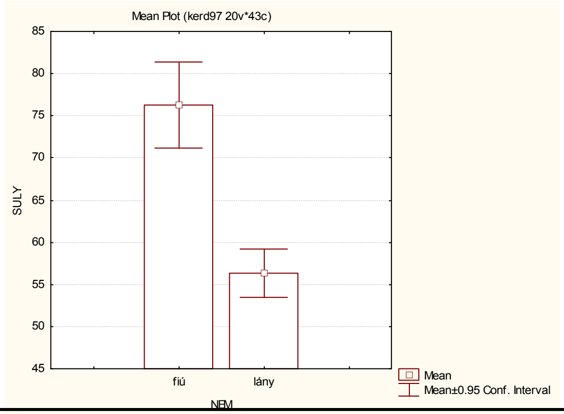
- It expresses the dispersion of the sample means around the (unknown) population mean.

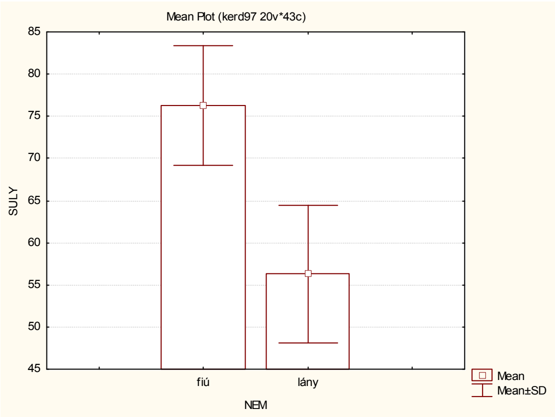
Mean-dispersion diagrams

- **Mean + SD**
- **Mean + SE**
- **Mean + 95% CI**



Mean ± SE





Mean ± 95% CI

Mean ± SD

Statistical estimation

Statistical estimation

- A **parameter** is a number that describes the population (its value is not known).
- For example:
  - μ and σ are parameters of the normal distribution $N(\mu,\sigma)$
  - n, p are parameters of the binomial distribution
  - λ is parameter of the Poisson distribution
- **Estimation:** based on sample data, we can calculate a number that is an approximation of the corresponding parameter of the population.
- A **point estimate** is a single numerical value used to approximate the corresponding population parameter.
  - For example, the sample mean is an estimation of the population's mean, μ.

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$ approximates μ

$$SD = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$ approximates σ

Interval estimate, confidence interval

- Interval estimate: a range of values that we think includes the true value of the population parameter (with a given level of certainty) .
- **Confidence interval: an interval which contains the value of the (unknown) population parameter with high probability.**
- The higher the probability assigned, the more confident we are that the interval does, in fact, include the true value.
- The probability assigned is the <u>confidence level</u> (generally: 0.90, 0.95, 0.99)

Interval estimate, confidence interval (cont.)

- „**high**” probability:
  the probability assigned is the <u>confidence level</u>
  (generally: 0.90, 0.95, 0.99).
- „**small**” probability:
  the „error” of the estimation (denoted by α)
  according to the confidence level is

  1-0.90=0.1, 1-0.95=0.05, 1-0.99=0.01
- The most often used confidence level is

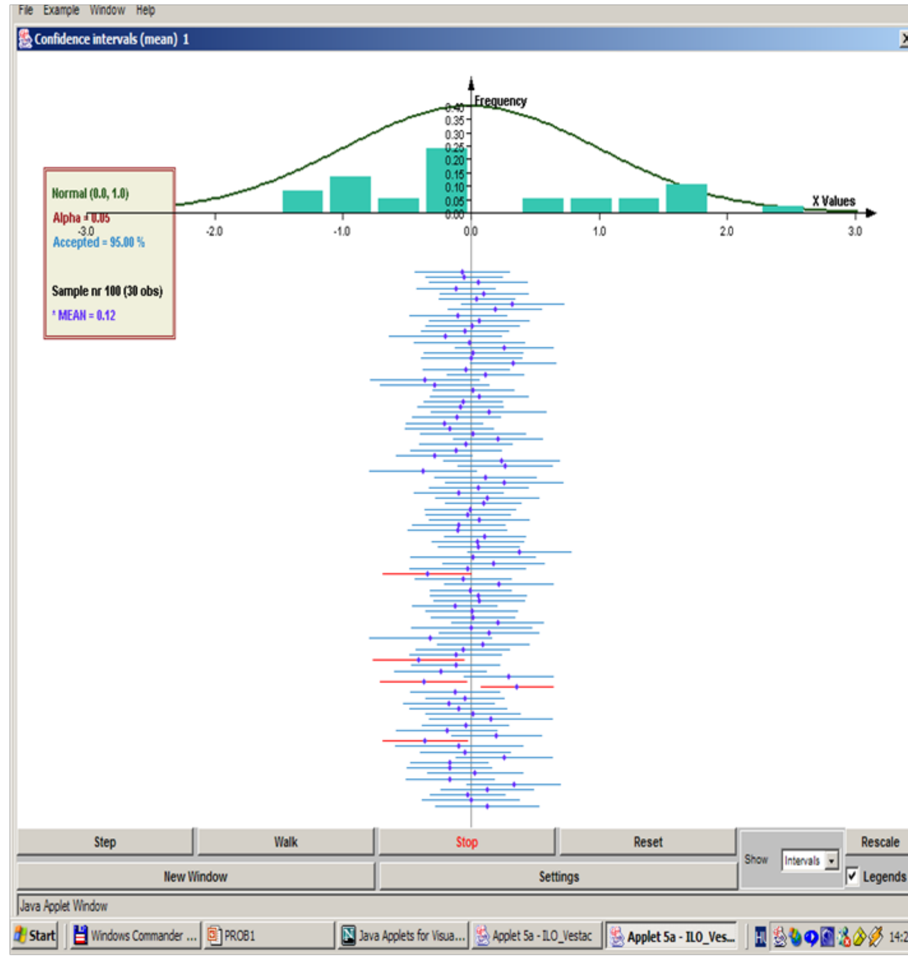
  95% (0.95),
- so the most often used value for α is

  $\boldsymbol{\alpha=0.05}$

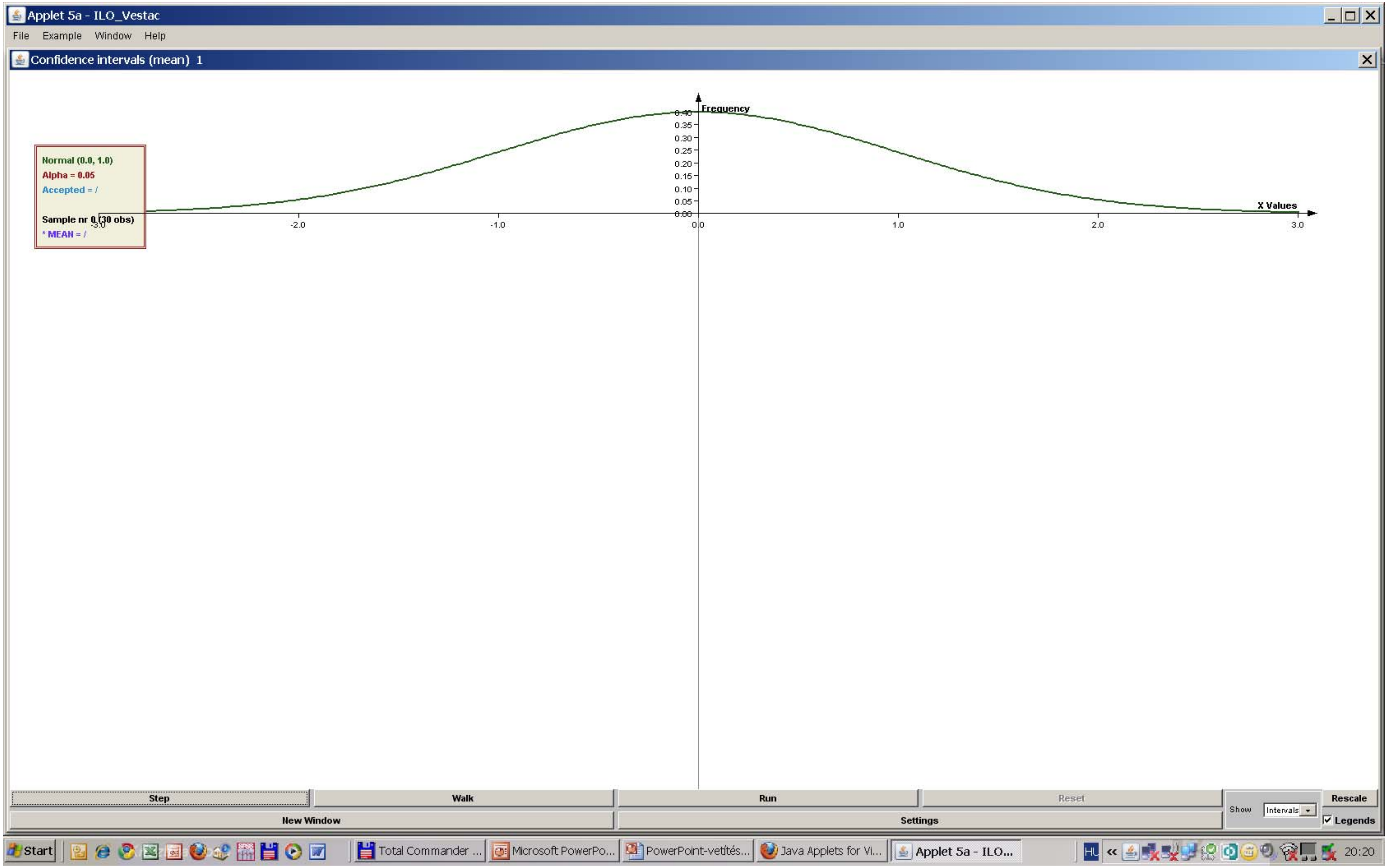
The confidence interval is based on the concept of repetition of the study under consideration

http://www.kuleuven.ac.be/ucs/java/index.htm

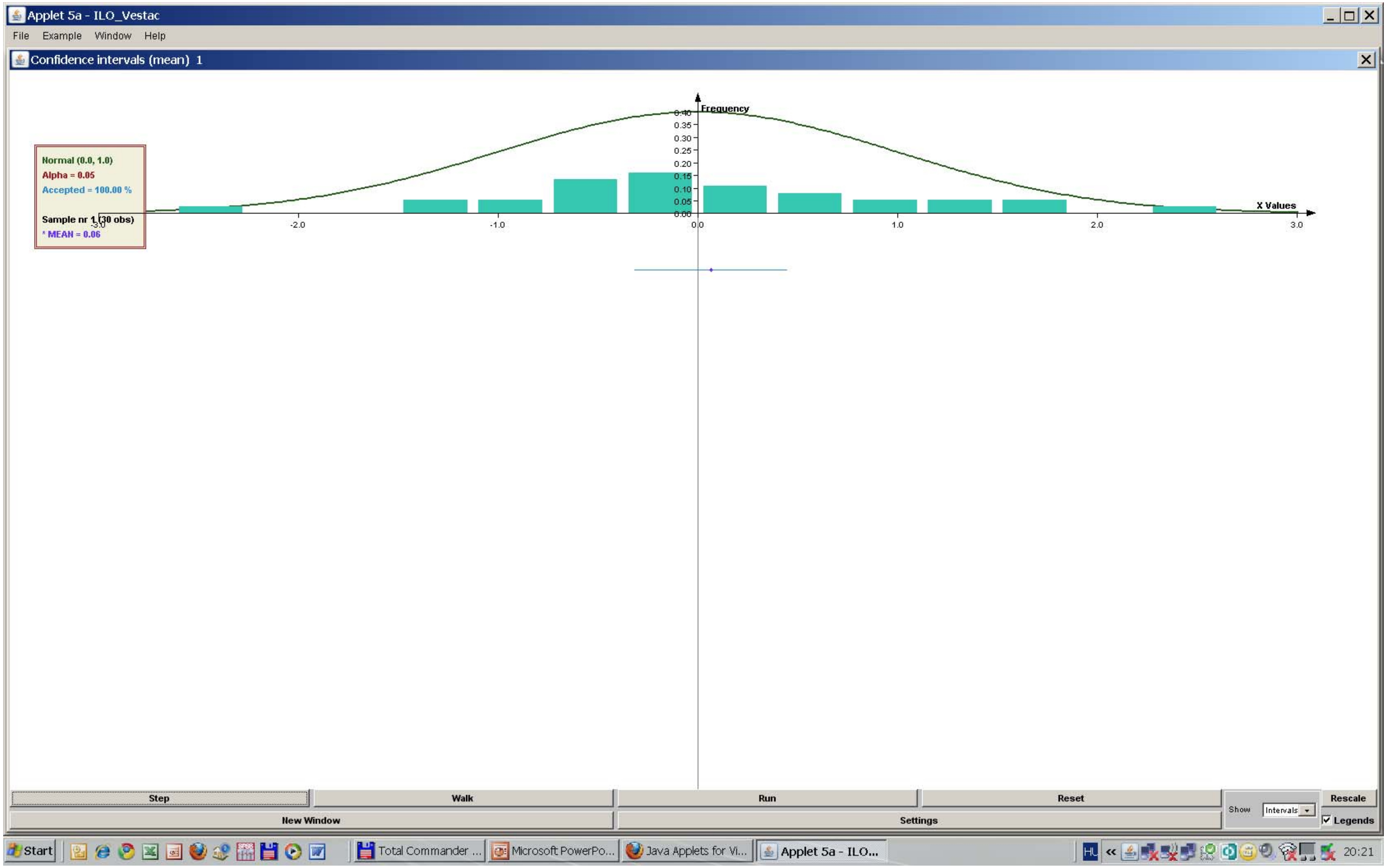
- If the study were to be repeated 100 times, of the 100 resulting 95% confidence intervals, we would expect 95 of these to include the population parameter.



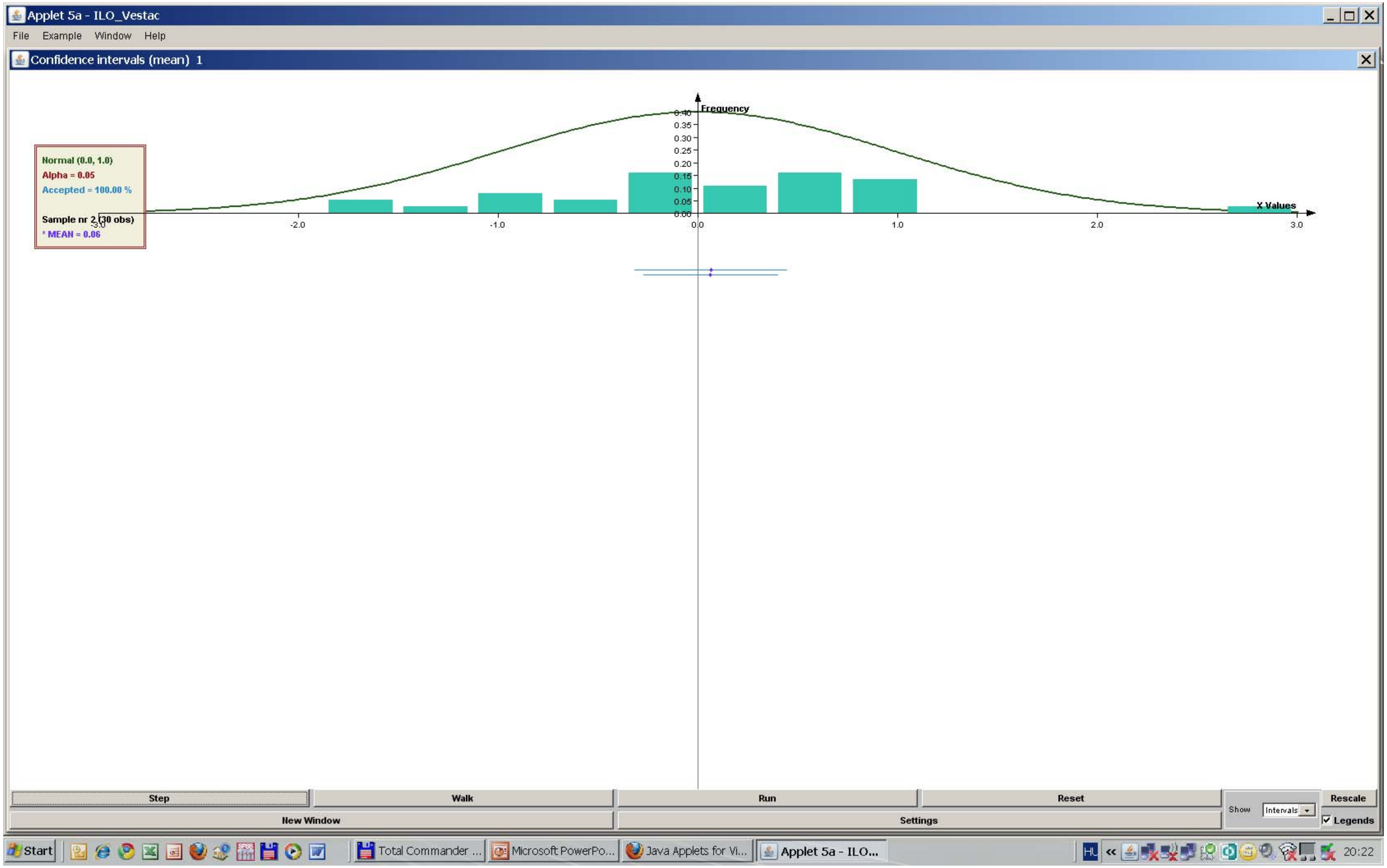
The distribution of the population



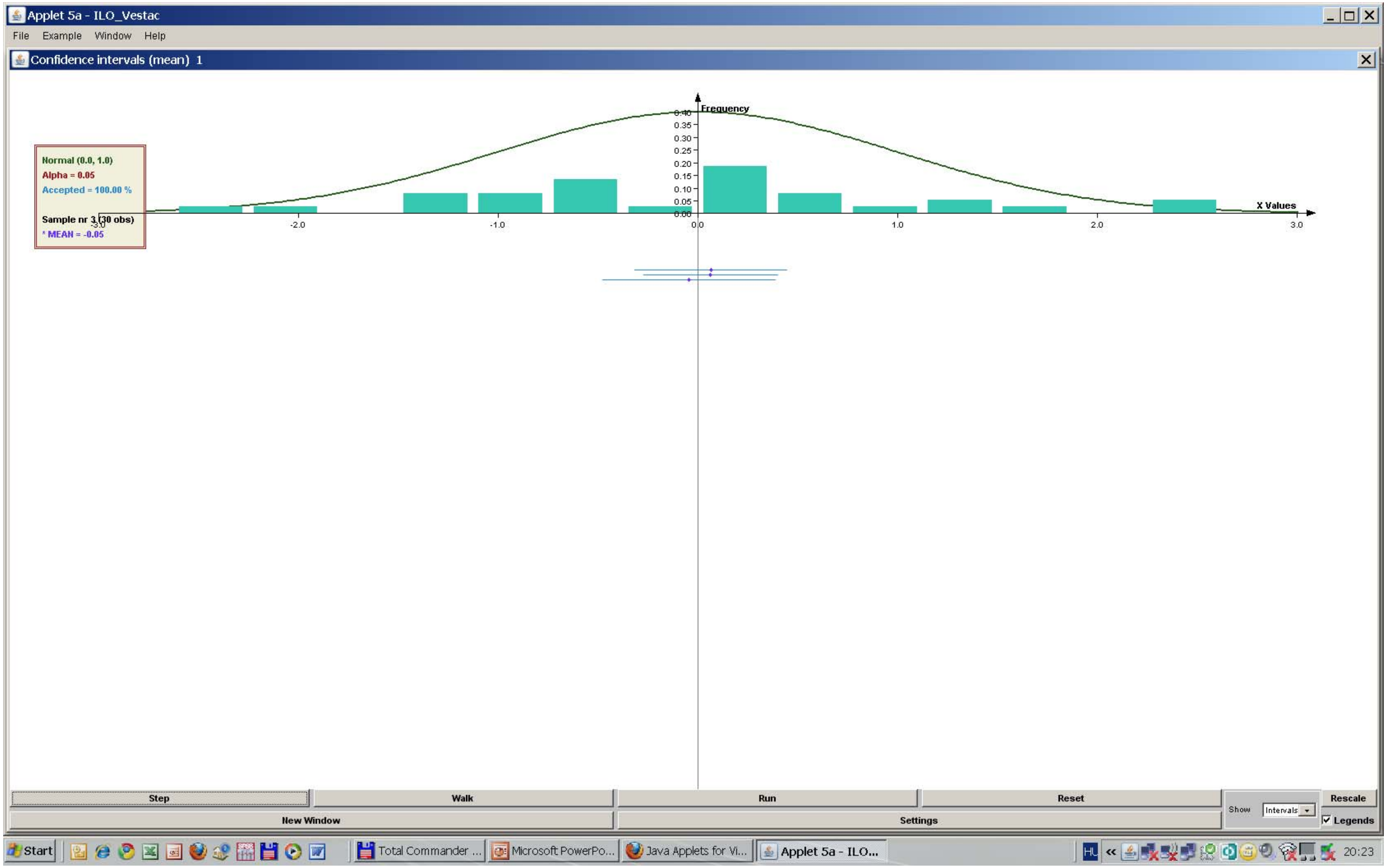
The histogram, mean and 95% CI of a sample drawn from the population



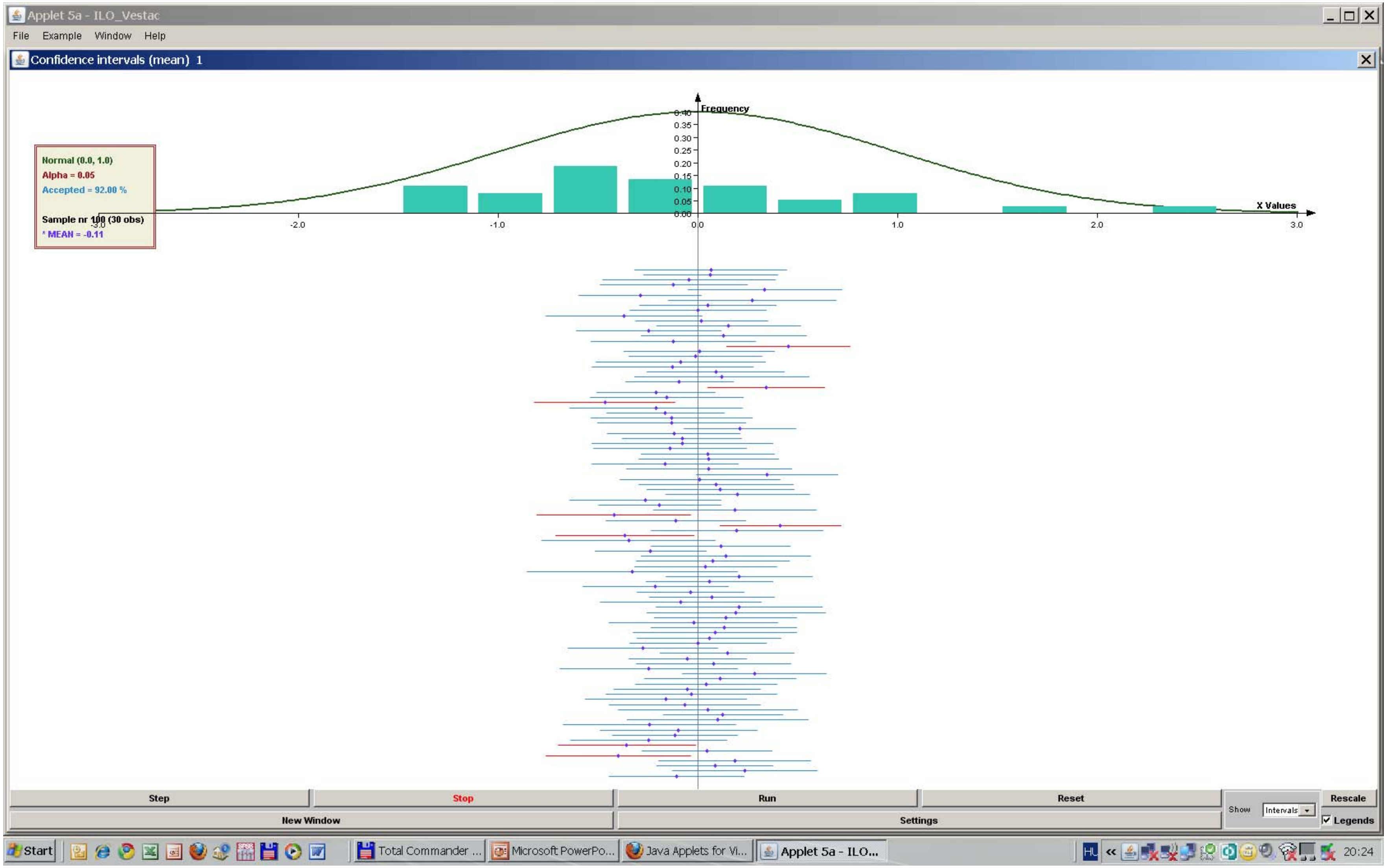
The histogram, mean and 95% CI of a 2nd sample drawn from the population



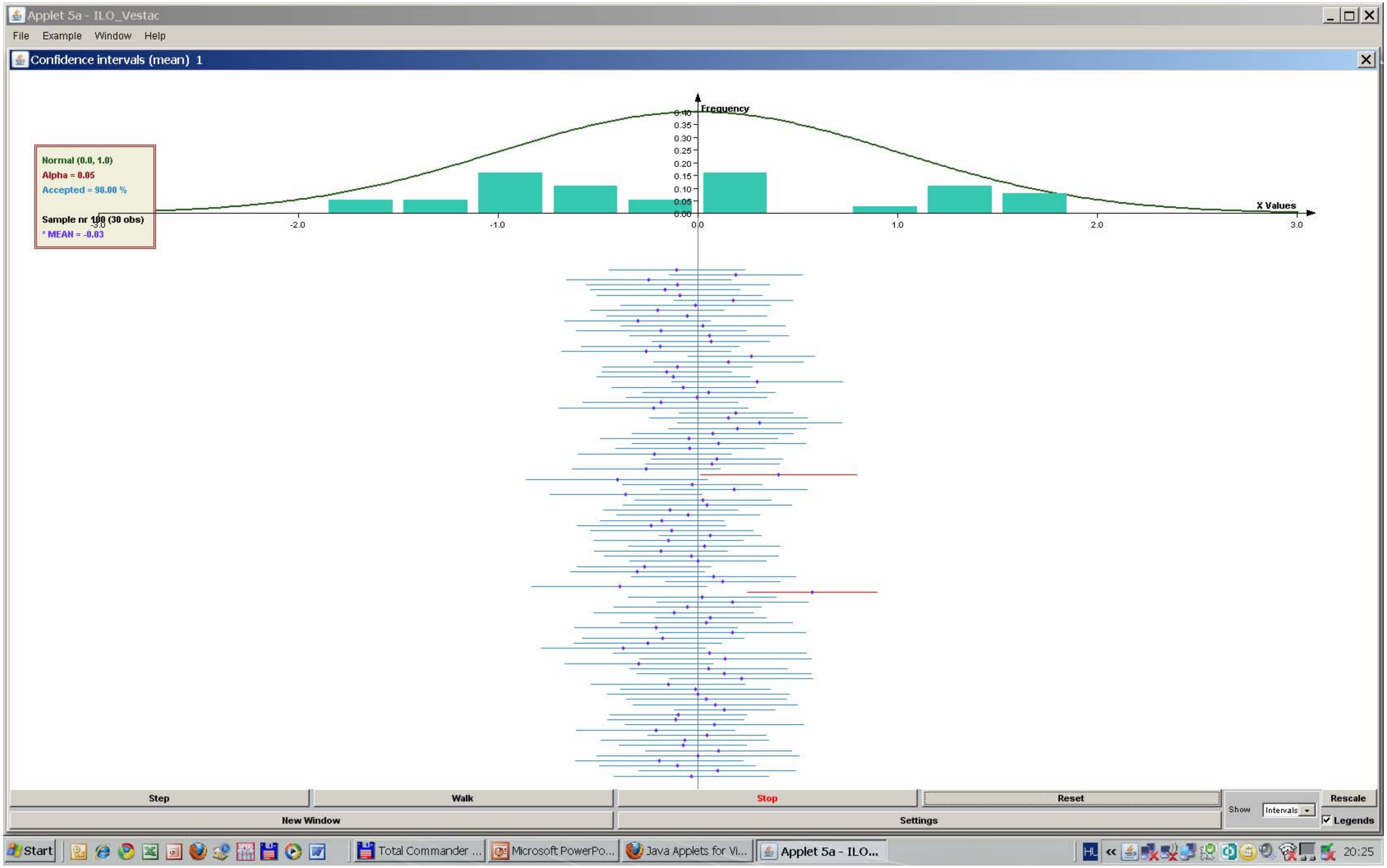
The histogram, mean and 95% CI of a 3rd sample drawn from the population



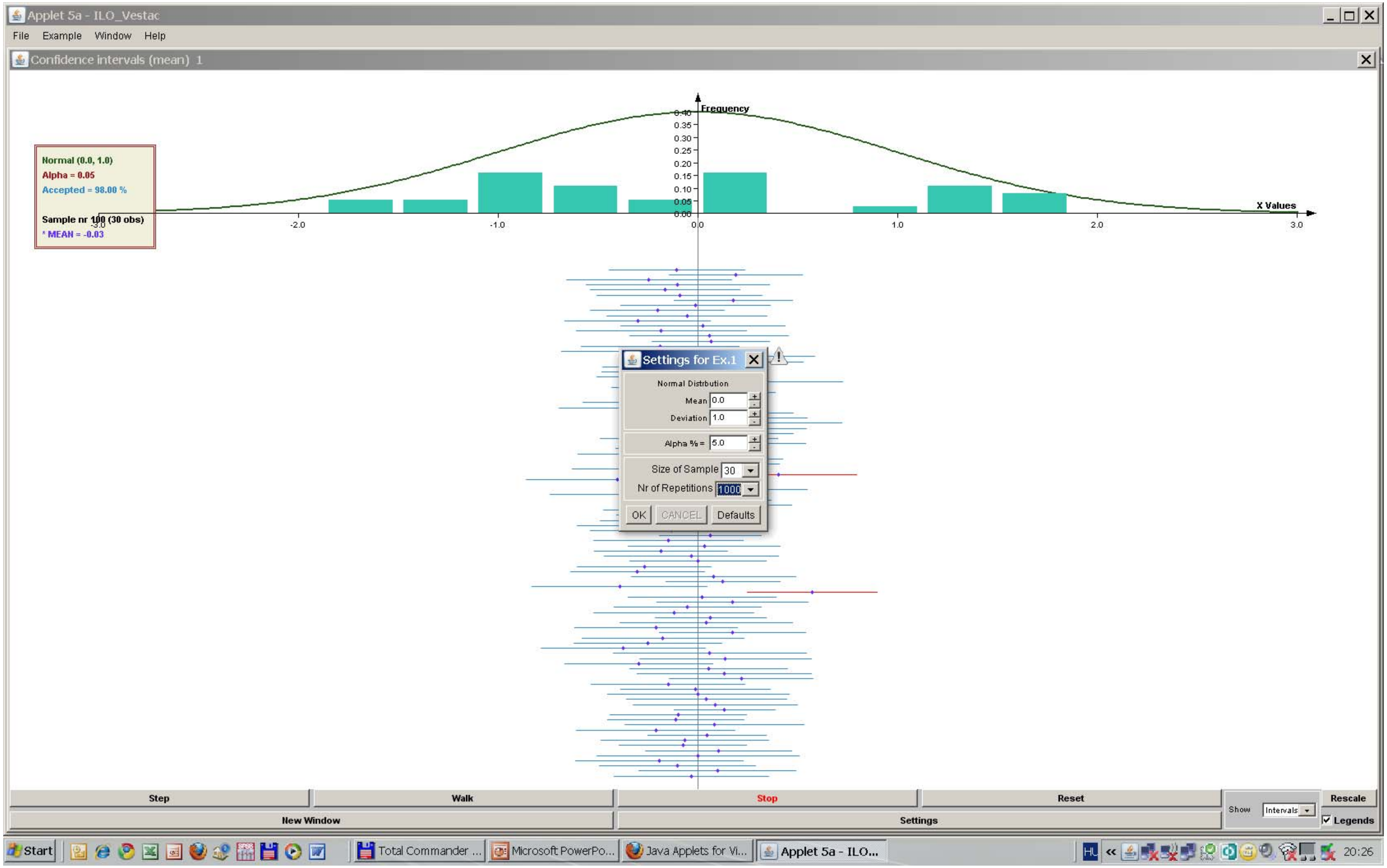
The histogram, mean and 95% CI of 100 samples drawn from the population



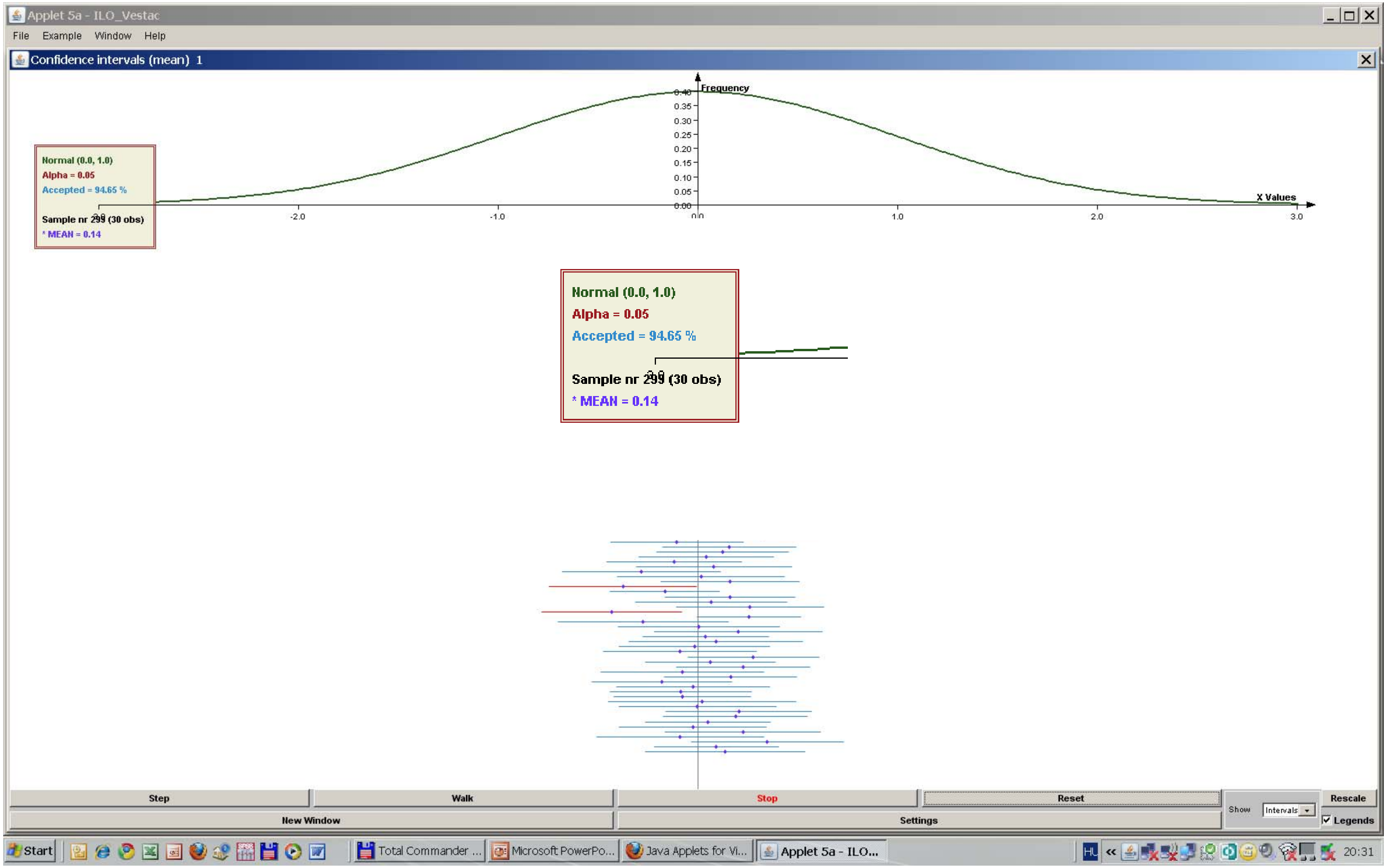
The histogram, mean and 95% CI of another 100 samples drawn from the population



Settings: 1000 samples



Result of the last 100



Formula of the confidence interval for the population's mean μ when σ is known

- It can be shown that

$$\left(\bar{x} - u_{\alpha/2}\frac{\sigma}{\sqrt{n}},\quad \bar{x} + u_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

  is a $(1-\alpha)100\%$ confidence interval for μ.
  - $u_{\alpha/2}$ is the $\alpha/2$ critical value of the standard normal distribution, it can be found in standard normal distribution table

    for $\alpha=0.05$ $u_{\alpha/2}=1.96$

    for $\alpha=0.01$ $u_{\alpha/2}=2.58$
- 95%CI for the population's mean

$$\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}},\quad \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

The standard error of mean (SE or SEM)

- $\dfrac{\sigma}{\sqrt{n}}$ is called the standard error of mean
- Meaning: the dispersion of the sample means around the (unknown) population's mean.
- When σ is unknown, the standard error of mean can be estimated from the sample by: $\dfrac{SD}{\sqrt{n}}$

$$\frac{\sigma}{\sqrt{n}} \approx \frac{SD}{\sqrt{n}} = \frac{\text{standard deviation}}{\sqrt{n}}$$

Example

- We wish to estimate the average number of heartbeats per minute for a certain population
  - Based on the data of 36 patients, the sample mean was 90, and the sample standard deviation was 15.5 (supposed to be known). Assuming that the heart-rate is normally distributed in the population, we can calculate a 95 % confidence interval for the population mean:

- $\alpha=0.05$, $u_{\alpha/2}=1.96$, $\sigma=15.5$
- The lower limit

  $90 - 1.96 \cdot 15.5/\sqrt{36}=90-1.96 \cdot 15.5/6=90-5.063=84.937$
- The upper limit

  $90 + 1.96 \cdot 15.5/\sqrt{36}=90+1.96 \cdot 15.5/6=90+5.063=95.064$
- The 95% confidence interval is
  - (84.94, 95.06)
- We can be 95% confident from this study that the true mean heart-rate among all such patients lies somewhere in the range 84.94 to 95.06, with 90 as our best estimate. This interpretation depends on the assumption that the sample of 36 patients is representative of all patients with the disease.

Formula of the confidence interval for the population's mean when σ is unknown

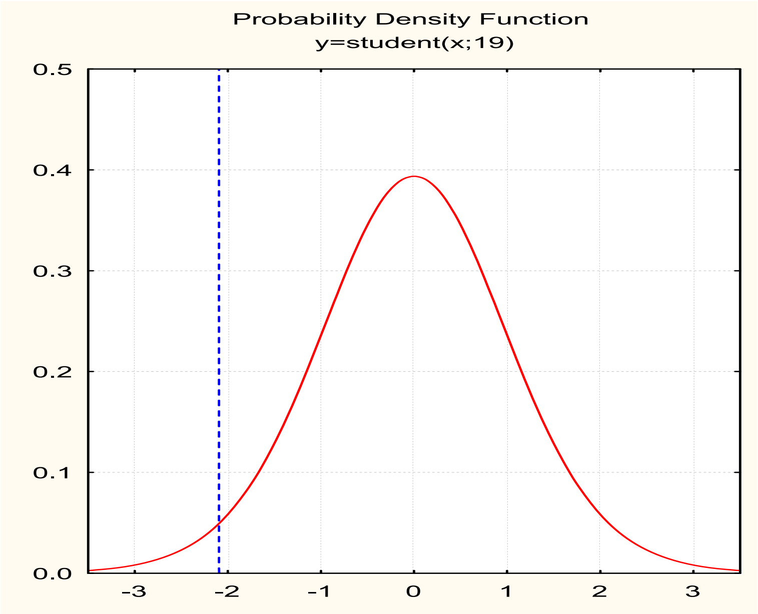
- When σ is unknown, it can be estimated by the sample SD (standard deviation). But, if we place the sample SD in the place of σ, $u_{\alpha/2}$ is no longer valid, it also must be replace by $t_{\alpha/2}$. So

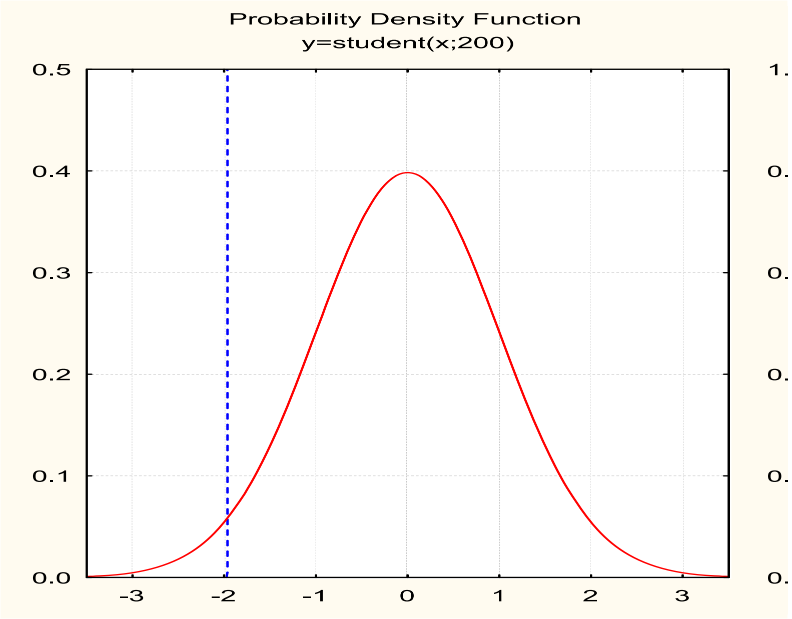
$$\left(\bar{x} - t_{\alpha/2}\frac{SD}{\sqrt{n}},\quad \bar{x} + t_{\alpha/2}\frac{SD}{\sqrt{n}}\right)$$

is a $(1-\alpha)100$ confidence interval for μ.

- $t_{\alpha/2}$ is the two-tailed α critical value of the Student's t statistic with n-1 degrees of freedom (see next slide)

t-distributions (Student's *t*-distributions)



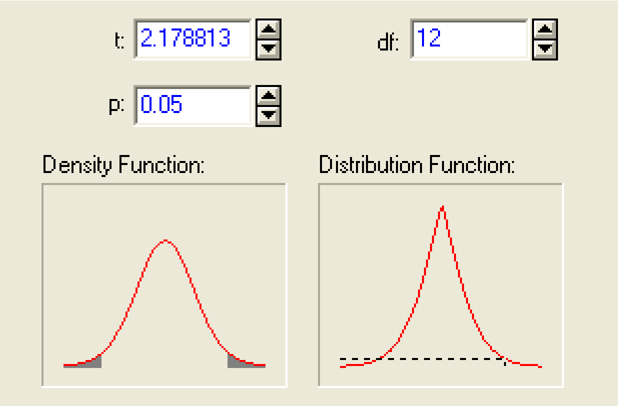


df=19

df=200

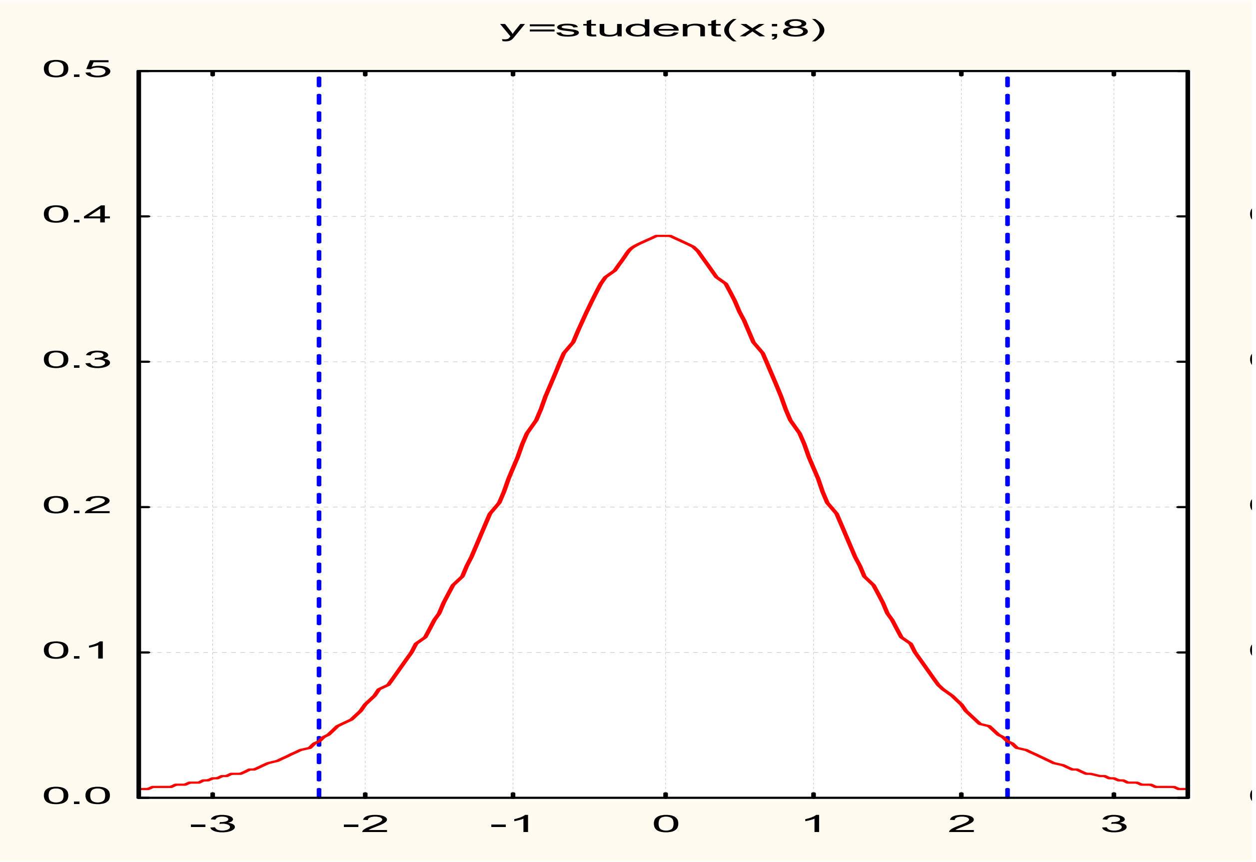
The Student's t-distribution

	Two-sided alfa					
df	**0.2**	**0.1**	**0.05**	**0.02**	**0.01**	**0.001**
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.599
3	1.638	2.353	3.182	4.541	5.841	12.924
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.869
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.408
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.768
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
∞	1.282	1.645	1.960	2.326	2.576	3.291



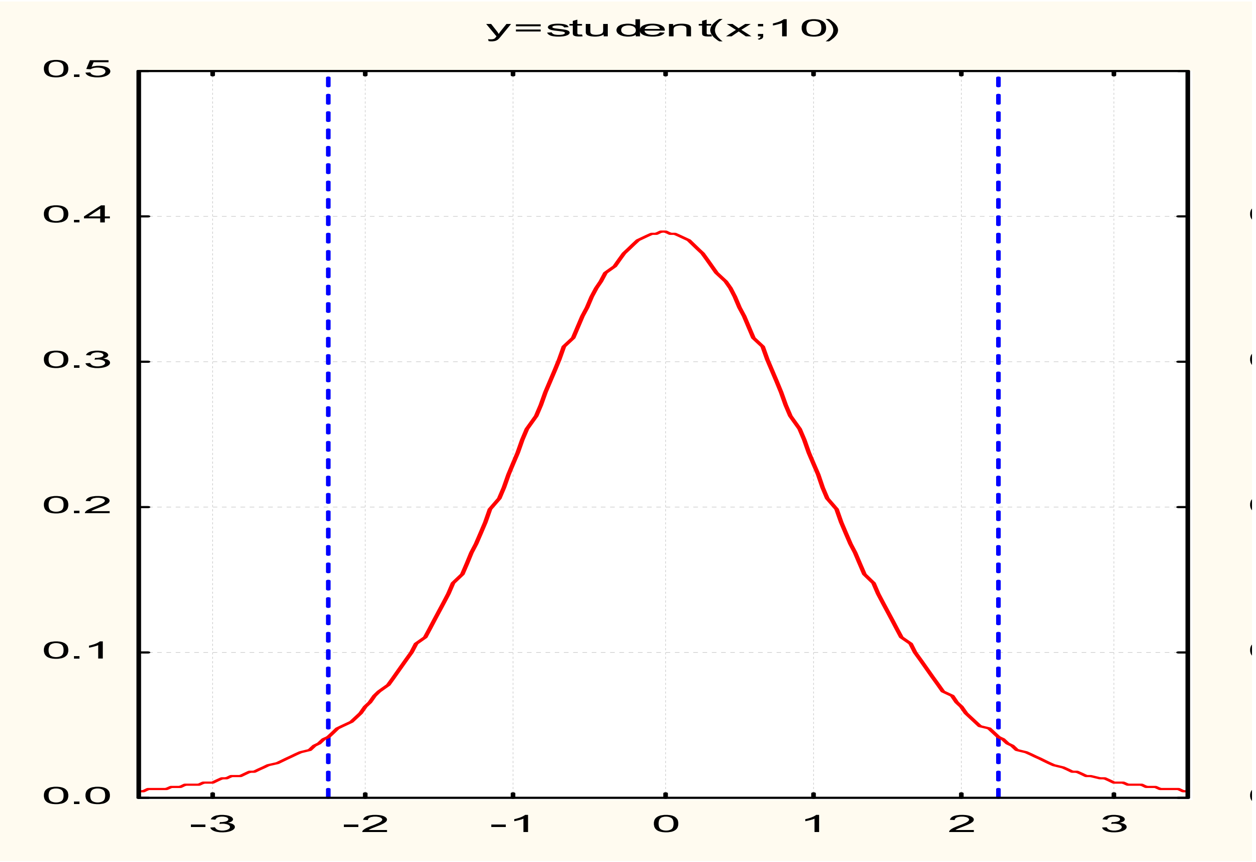
For α=0.05 and df=12,
the critical value is $t_{\alpha/2}$ =2.179

Student’s *t*-distribution



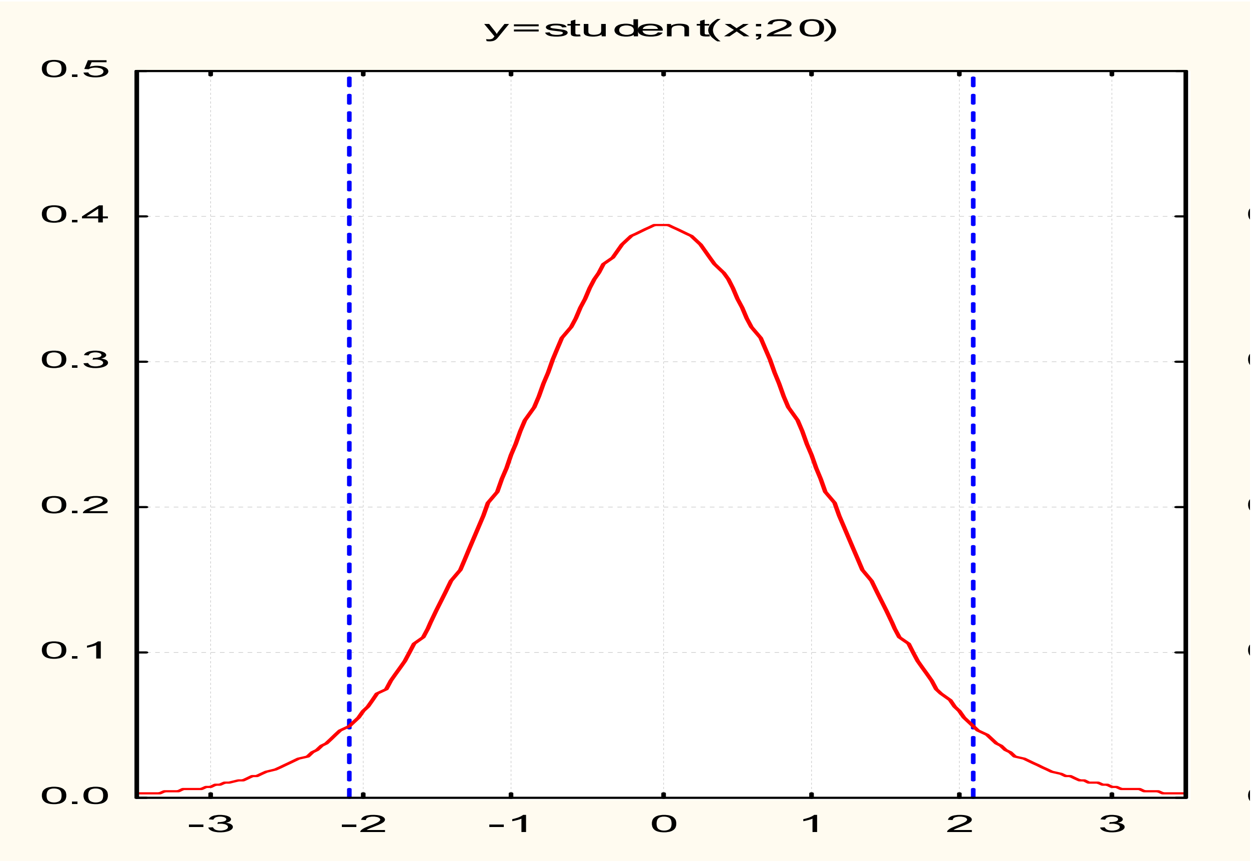
Degrees of freedom: 8

Student’s *t*-distribution



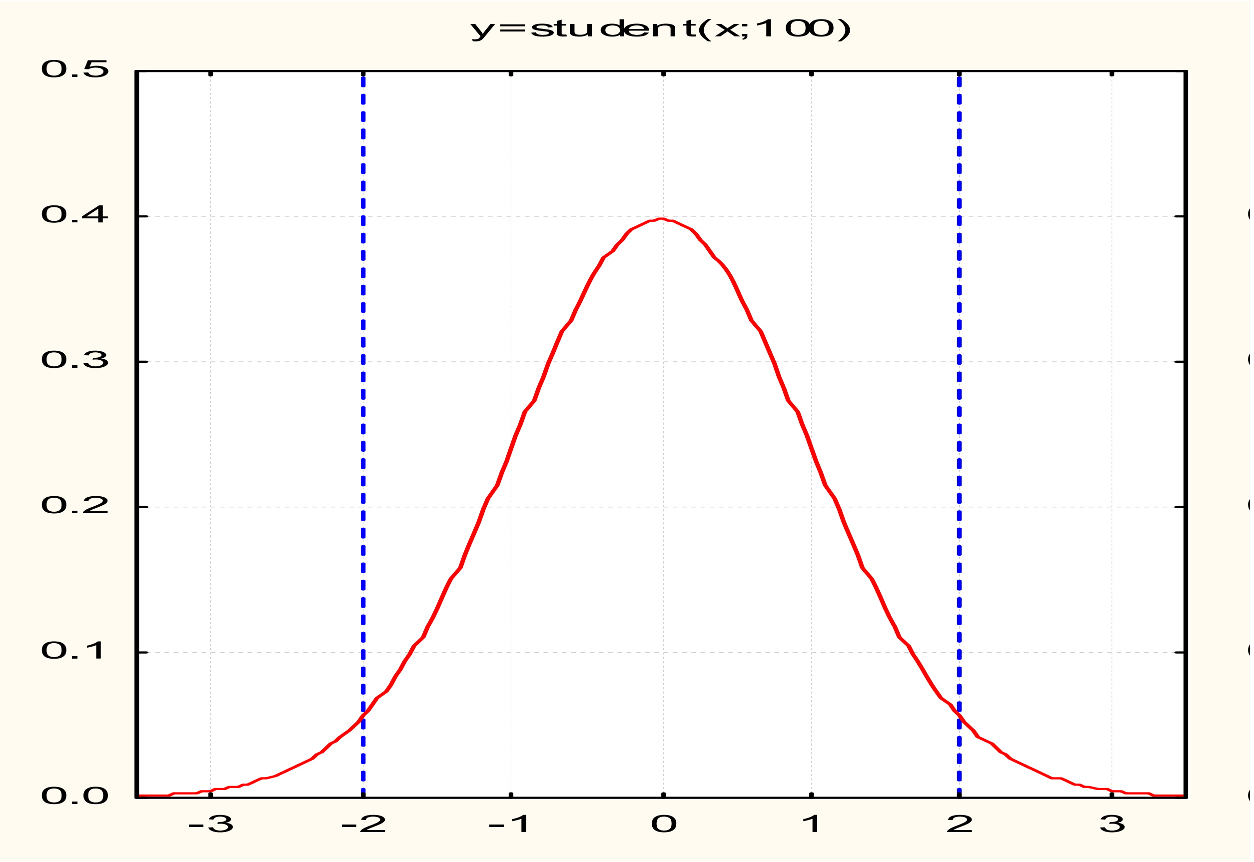
Degrees of freedom: 10

Student's *t*-distribution



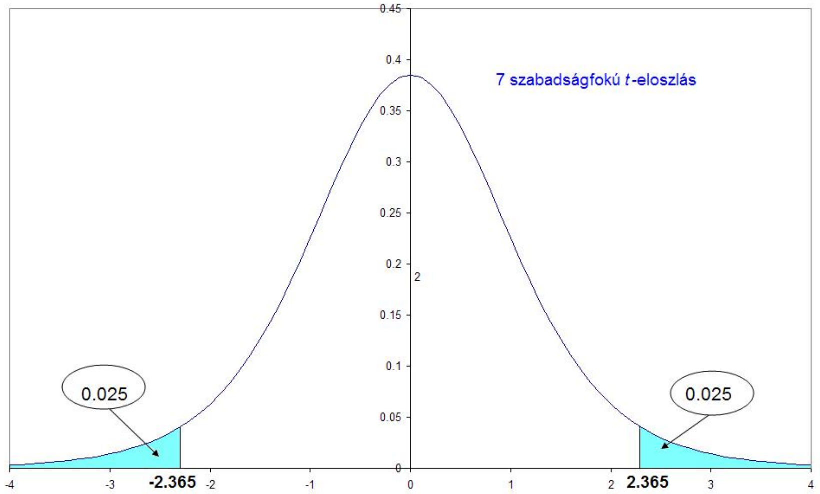
Degrees of freedom: 20

Student's *t*-distribution



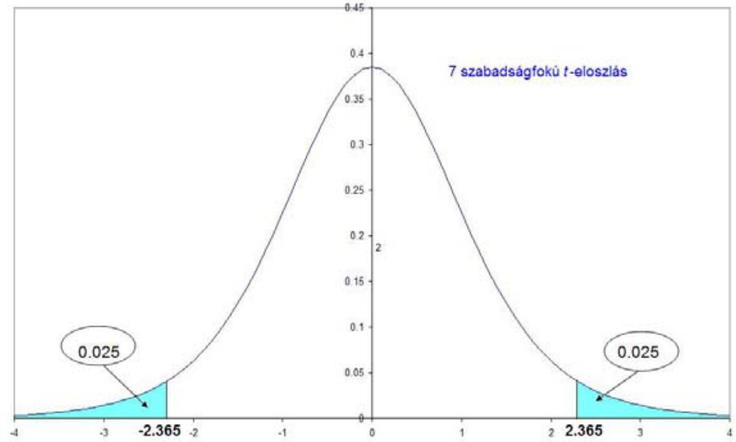
Degrees of freedom: 100

Student's *t*-distribution table



Degrees of freedom	Two sided alfa 0.2	0.1	0.05	0.02	0.01
1	3.077683537	6.313752	12.7062	31.82052	63.65674
2	1.885618083	2.919986	4.302653	6.964557	9.924843
3	1.637744352	2.353363	3.182446	4.540703	5.840909
4	1.533206273	2.131847	2.776445	3.746947	4.604095
5	1.475884037	2.015048	2.570582	3.36493	4.032143
6	1.439755747	1.94318	2.446912	3.142668	3.707428
7	1.414923928	1.894579	2.364624	2.997952	3.499483
8	1.39681531	1.859548	2.306004	2.896459	3.355387
9	1.383028739	1.833113	2.262157	2.821438	3.249836
10	1.372183641	1.812461	2.228139	2.763769	3.169273
11	1.363430318	1.795885	2.200985	2.718079	3.105807

Student’s *t*-distribution table



degrees of freedom	0.2	0.1	two sided alfa **0.05**	0.02	0.01	0.001
1	3.077683537	6.313752	**12.7062**	31.82052	63.65674	636.6192
2	1.885618083	2.919986	**4.302653**	6.964557	9.924843	31.59905
3	1.637744352	2.353363	**3.182446**	4.540703	5.840909	12.92398
4	1.533206273	2.131847	**2.776445**	3.746947	4.604095	8.610302
5	1.475884037	2.015048	**2.570582**	3.36493	4.032143	6.868827
6	1.439755747	1.94318	**2.446912**	3.142668	3.707428	5.958816
7	1.414923928	1.894579	**2.364624**	2.997952	3.499483	5.407883
...	…	…	**…**	…	…	…
100	1.290074761	1.660234	**1.983971**	2.364217	2.625891	3.390491
...	…	…	**…**	…	…	…
500	1.283247021	1.647907	**1.96472**	2.333829	2.585698	3.310091
...	…	…	**…**	…	…	…
1000000	1.281552411	1.644855	**1.959966**	2.326352	2.575834	3.290536

Example 1.

- **We wish to estimate the average number of heartbeats per minute for a certain population.**
- The mean for a sample of 13 subjects was found to be 90, the standard deviation of the sample was SD=15.5. Supposed that the population is normally distributed the 95 % confidence interval for μ:
- α=0.05, SD=15.5
- Degrees of freedom: df=n-1=13 -1=12
- $t_{\alpha/2}$ =2.179
- The lower limit is

  $90 - 2.179 \cdot 15.5/\sqrt{13} = 90 - 2.179 \cdot 4.299 = 90 - 9.367 = 80.6326$
- The upper limit is

  $90 + 2.179 \cdot 15.5/\sqrt{13} = 90 + 2.179 \cdot 4.299 = 90 + 9.367 = 99.367$
- **The 95% confidence interval for the population mean is**

  (80.63, 99.36)
- **It means that the true (but unknown) population means lies it the interval (80.63, 99.36) with 0.95 probability. We are 95% confident the true mean lies in that interval.**

Example 2.

- We wish to estimate the average number of heartbeats per minute for a certain population.
- The mean for a sample of 36 subjects was found to be 90, the standard deviation of the sample was SD=15.5. Supposed that the population is normally distributed the 95 % confidence interval for μ:
- α=0.05, SD=15.5
- Degrees of freedom: df=n-1=36-1=35
- t α/2=2.0301
- The lower limit is

  $90 - 2.0301 \cdot 15.5/\sqrt{36} = 90 - 2.0301 \cdot 2.5833 = 90 - 5.2444 = 84.755$
- The upper limit is

  $90 + 2.0301 \cdot 15.5/\sqrt{36} = 90 + 2.0301 \cdot 2.5833 = 90 + 5.2444 = 95.24$
- The 95% confidence interval for the population mean is

  (84.76, 95.24)
- It means that the true (but unknown) population means lies it the interval (84.76, 95.24) with 0.95 probability. We are 95% confident that the true mean lies in that interval.

Comparison

- **We wish to estimate the average number of heartbeats per minute for a certain population.**
- The mean for a sample of 13 subjects was found to be 90, the standard deviation of the sample was SD=15.5. Supposed that the population is normally distributed the 95 % confidence interval for μ:
- $\alpha=0.05$, SD=15.5
- Degrees of freedom: df=n-1=13 -1=12
- $t_{\alpha/2}=2.179$
- The lower limit is

  $90 - 2.179 \cdot 15.5/\sqrt{13} = 90 - 2.179 \cdot 4.299 = 90 - 9.367 = 80.6326$
- The upper limit is

  $90 + 2.179 \cdot 15.5/\sqrt{13} = 90 + 2.179 \cdot 4.299 = 90 + 9.367 = 99.367$
- **The 95% confidence interval for the population mean is**

  (80.63, 99.36)

- **We wish to estimate the average number of heartbeats per minute for a certain population.**
- The mean for a sample of 36 subjects was found to be 90, the standard deviation of the sample was SD=15.5. Supposed that the population is normally distributed the 95 % confidence interval for μ:
- $\alpha=0.05$, SD=15.5
- Degrees of freedom: df=n-1=36-1=35
- $t\ \alpha/2=2.0301$
- The lower limit is

  $90 - 2.0301 \cdot 15.5/\sqrt{36} = 90 - 2.0301 \cdot 2.5833 = 90 - 5.2444 = 84.755$
- The upper limit is

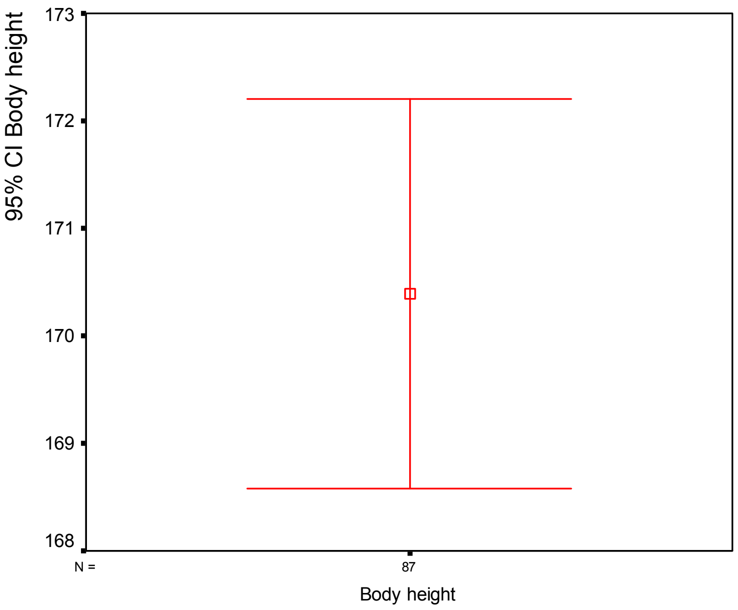
  $90 + 2.0301 \cdot 15.5/\sqrt{36} = 90 + 2.0301 \cdot 2.5833 = 90 + 5.2444 = 95.24$
- **The 95% confidence interval for the population mean is**

  (84.76, 95.24)

Example

Descriptives

			Statistic	Std. Error
Body height	Mean		170.3908	.91329
	95% Confidence Interval for Mean	Lower Bound	168.5752	
		Upper Bound	172.2064	
	5% Trimmed Mean		170.2886	
	Median		170.0000	
	Variance		72.566	
	Std. Deviation		8.51859	
	Minimum		152.00	
	Maximum		196.00	
	Range		44.00	
	Interquartile Range		11.0000	
	Skewness		.274	.258
	Kurtosis		.270	.511



Presentation of results

Table 2 | Primary and secondary outcomes according to treatment in the 502 randomised children according to allocation to new treatment (oral co-amoxiclav) or standard treatment (intravenous ceftriaxone followed by oral co-amoxiclav). Figures are means (SD) unless specified otherwise

Parameter	New treatment (n=244)	Standard treatment (n=258)	Mean difference (95% CI)
Short term outcomes			
Time to defervescence (hours)	36.9 (19.7) (n=241)	34.3 (20) (n=253)	2.6 (−0.9 to 6)
White cell count (×10^9/l)*	9.8 (3.5) (n=230)	9.5 (3.1) (n=243)	0.3 (−0.3 to 0.9)
Neutrophils (×10^9/l)*	3.0 (2.2) (n=207)	2.8 (1.9) (n=217)	0.2 (−0.2 to 0.6)
Erythrocyte sedimentation rate (mm in first hour)*	50.8 (32) (n=170)	52.6 (27.9) (n=168)	−1.8 (−8.2 to 4.7)
C reactive protein (mg/l)*†	9.3 (20.9) (n=235)	8.2 (15.4) (n=251)	1.1 (−2.6 to 4.1)
Sterile urine	185/186 (99.45%)	203/204 (99.5%)	−0.05% (−1.5% to 1.4%)
Primary outcome			
Scar on renal scan at 12 months	27/197 (13.7%)	36/203 (17.7%)	−4% (−11.1% to 3.1%)

*Parameters obtained 72 hours after start of antibiotic treatment.
†Ratio between obtained value and upper limit of normal reference values for each laboratory.

BMJ | ONLINE FIRST | bmj.com

BMJ

Antibiotic treatment for pyelonephritis in children: multicentre randomised controlled non-inferiority trial

Giovanni Montini, Antonella Toffolo, Pietro Zucchetta, Roberto Dall'Amico, Daniela Gobber, Alessandro Calderan, Francesca Maschio, Luigi Pavanello, Pier Paolo Molinari, Dante Scorrano, Sergio Zanchetta, Walburga Cassar, Paolo Brisotto, Andrea Corsini, Stefano Sartori, Liviana Da Dalt, Luisa Murer and Graziella Zacchello

BMJ 2007;335;386-; originally published online 4 Jul 2007;
doi:10.1136/bmj.39244.692442.55

Review questions and problems

- The central limit theorem
- The meaning and the formula of the standard error of mean (SE)
- The meaning of a confidence interval
- The confidence level
- Which is wider, a 95% or a 99% confidence interval?
- Calculation of the confidence interval for the population mean in case of unknown standard deviation
- Studenst's t-distribution
- In a study, systolic blood pressure of 16 healthy women was measured. The mean was 121, the standard deviation was SD=8.2. Calculate the standard error.
- In a study, systolic blood pressure of 10 healthy women was measured. The mean was 119, the standard error 0.664. Calculate the 95% confidence interval for the population mean!
  (α=0.05, t_{table}=2.26).