2024 Biophysics 2

Biostatistics

Károly Liliom

topics

1. descriptive statistics (what data are)
2. hypothesis testing (comparing data)
3. correlation and regression analysis

Recommended readings:

- Medical Biophysics Practices – ed. M. Kellermayer, 3rd ed., Semmelweis Publisher: Appendix: Biostatistics

Further readings:

- Harvey Motulsky: Intuitive Biostatistics – A Nonmathematical Guide to Statistical Thinking, Oxford University Press

- Nature Collection: Statistics for Biologists

“There are three types of lies -- lies, damn lies, and statistics.”
— **Benjamin Disraeli**

“Facts are stubborn things, but statistics are pliable.”
— **Mark Twain**

“If your experiment needs a statistician, you need a better experiment.”
— **Ernest Rutherford**

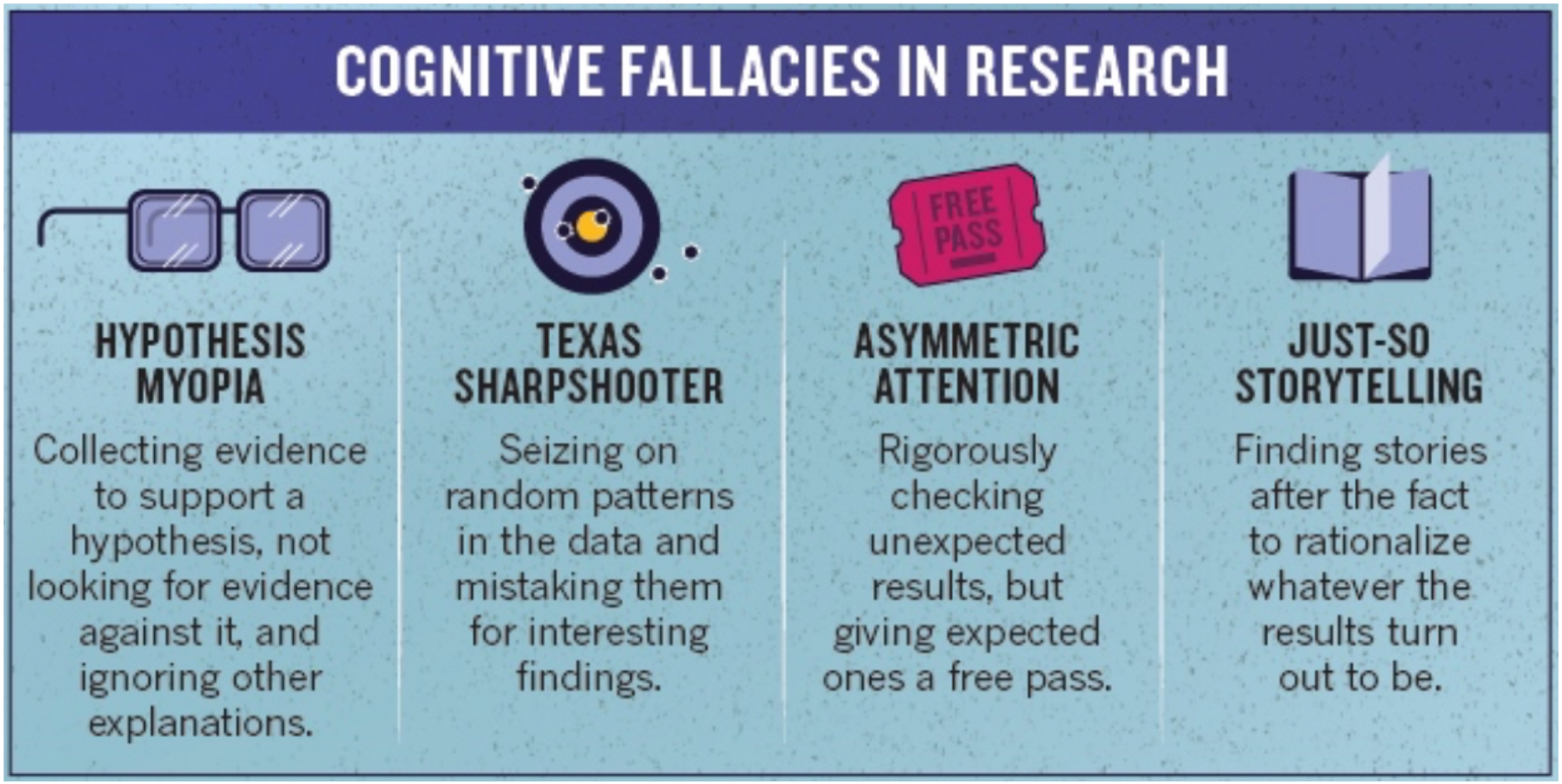
When analyzing data, our goal is simple: we wish to make the strongest possible conclusion from limited amounts of data.

To achieve this, we need to overcome two problems:

- Important findings can be obscured by biological variability and experimental imprecision. This makes it difficult to distinguish real differences from random variation.

- The human brain excels at finding patterns, even in random data. Our natural inclination (especially with our own data) is to conclude that differences are real and to minimize the contribution of random variability. Statistical rigor prevents us from making this mistake.

How Scientist fool themselves



Nature Special Collection: Reproducibility

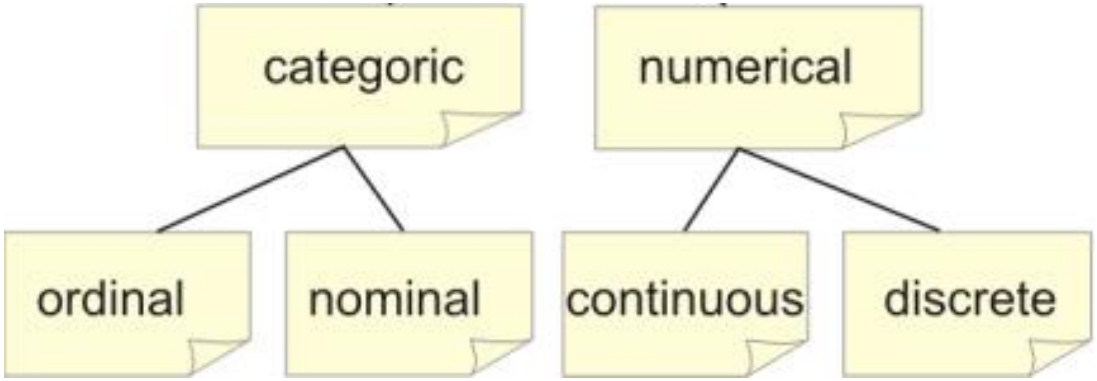
Ronald Fisher:

"To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of."

For statistical analyses to be interpretable, it is essential that these three statements be true:

- All analyses were planned.
- All planned analyses were conducted exactly as planned and then reported.
- All the analyses are taken into account when interpreting the results.

Data collection = data types



I. **Categoric**

I/a. Nominal: Qualitative not sortable data, e.g. the blood groups: A, B, AB, 0.

I/b. Ordinal: Qualitative sortable data , e.g. the severity of the disease: modest, medium, strong.

II. Numerical

II/a. Discrete: Quantitative data which can have only certain values, e.g. the number of children in a family.

II/b. Continuous: Quantitative data which can have any value in a certain interval, e.g. blood pressure, weight of people...

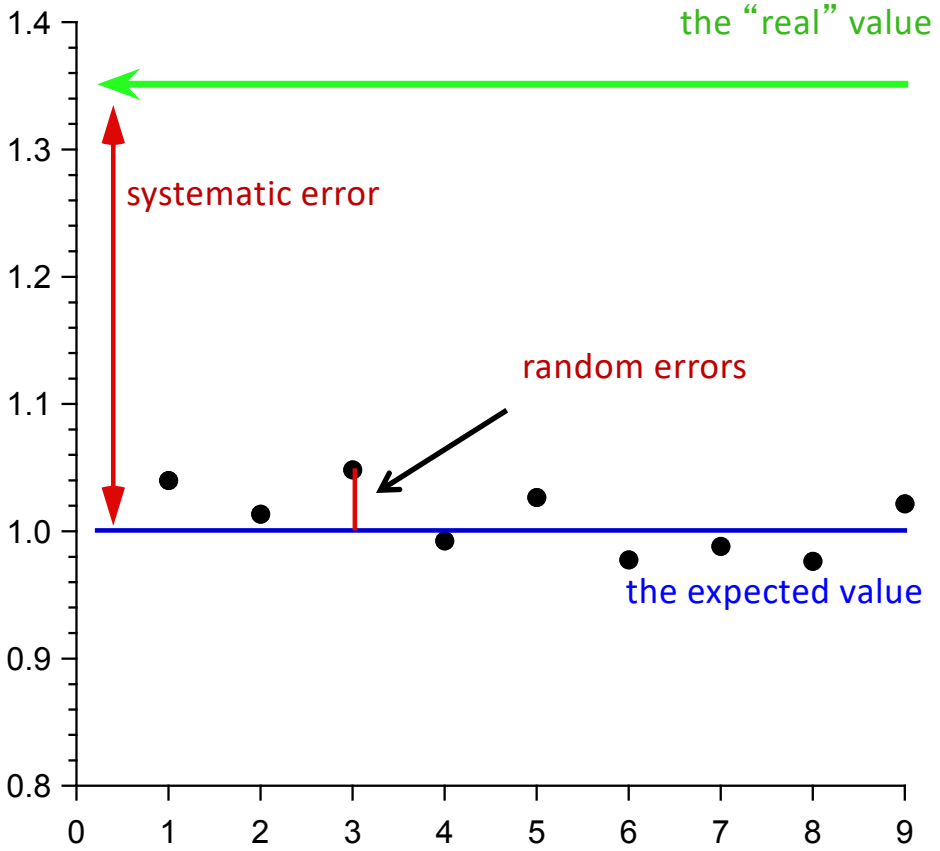
Data collection = measurement

- During measurement the quantity to be measured should be compared to a standard (measurement unit) $\Rightarrow$ the results of measurements are never precise!
- As a consequence, the results of measurements (data) are governed by the rules of chances = every piece of data is a stochastic variable.
- Data may contain systematic errors and for sure contains random (stochastic) errors.
- Systematic errors can be corrected - stochastic errors are inherent to the measurements.

Data are stochastic variables

- Given set of elements together with the numerical values of their properties under study form a statistical manifold (distribution). Usually, we want to determine this distribution of values in the set. To estimate the parameters describing the distribution we have to do sampling of the distribution.
- Sampling means random selection of a subset (“n” elements) of the statistical distribution.
- A measurement is always sampling!
- In laboratory practice, the obtained value of a measurement is a continuously distributed stochastic variable.
- “n” independent measurements (parallels) result in “n” independent stochastic variables with the same probability distribution.
- The sampling is representative if every element of the distribution can be selected by equal chance.

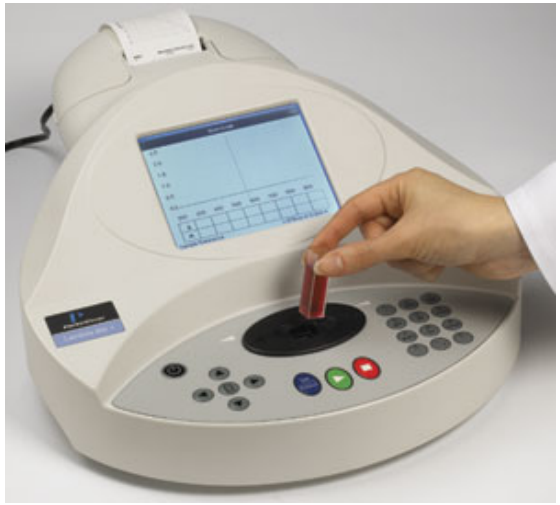
The outcome of the measurement is the expected value of the distriution



Random errors are almost always normally distributed!

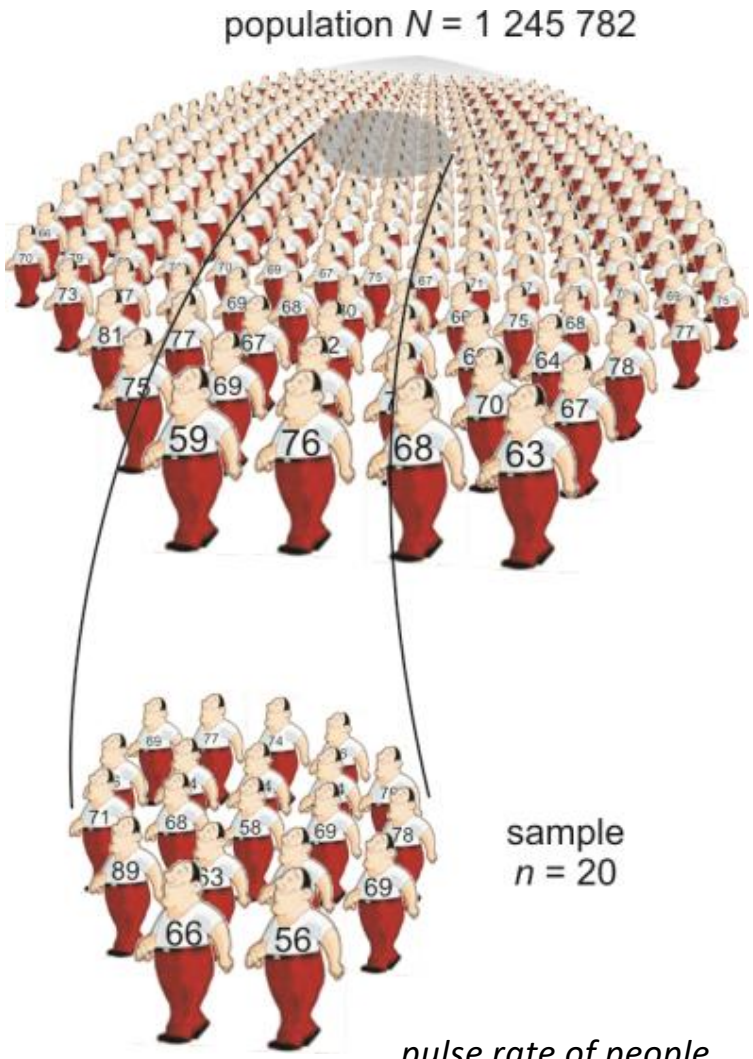
“n” number of independent measurements result in

“n” independent stochastic variables with the same probability distribution.



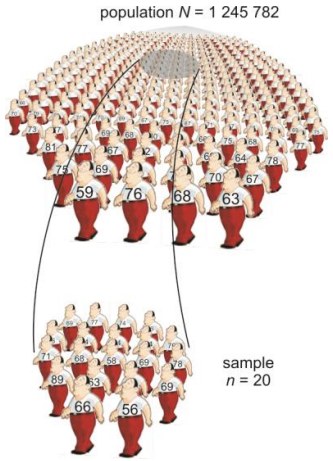
How many parallels do we have by measuring a sample "n"-times?

Population and sample



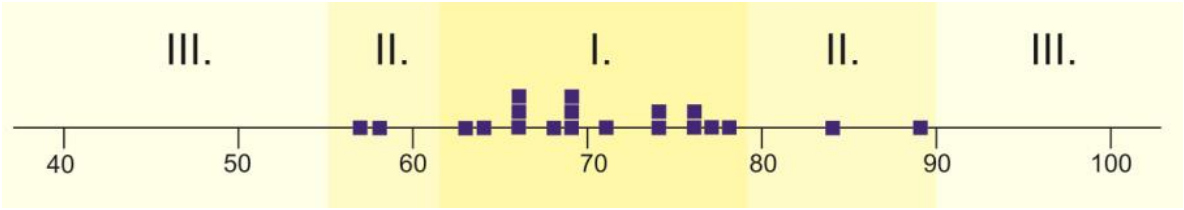
pulse rate of people

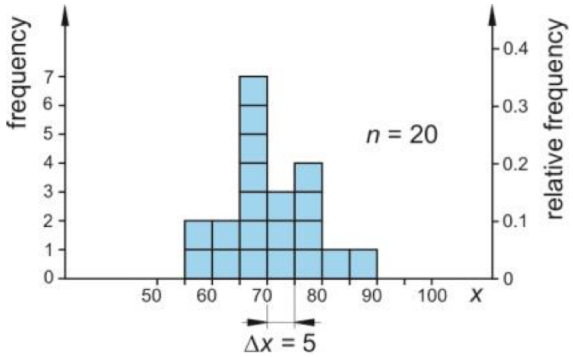
Frequency Distribution, Histogram

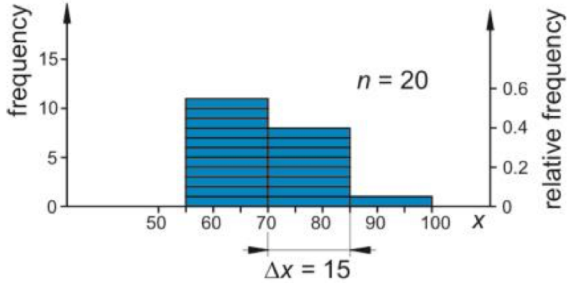


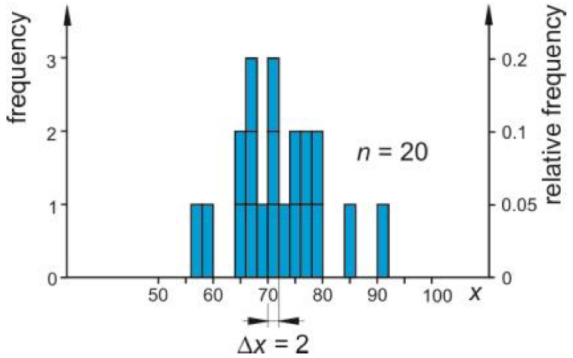
66	56	89	63	66	69	71	68	58	69
78	66	64	84	74	76	69	77	74	76

pulse rates of a sample

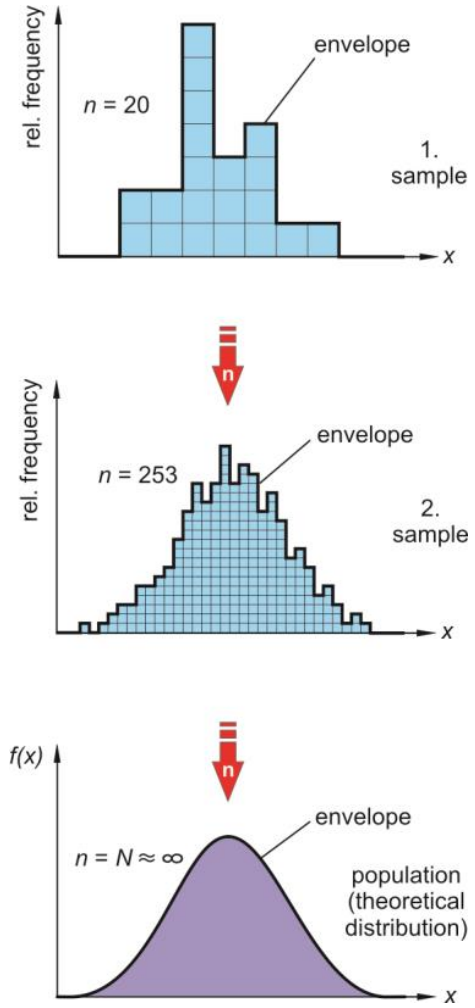






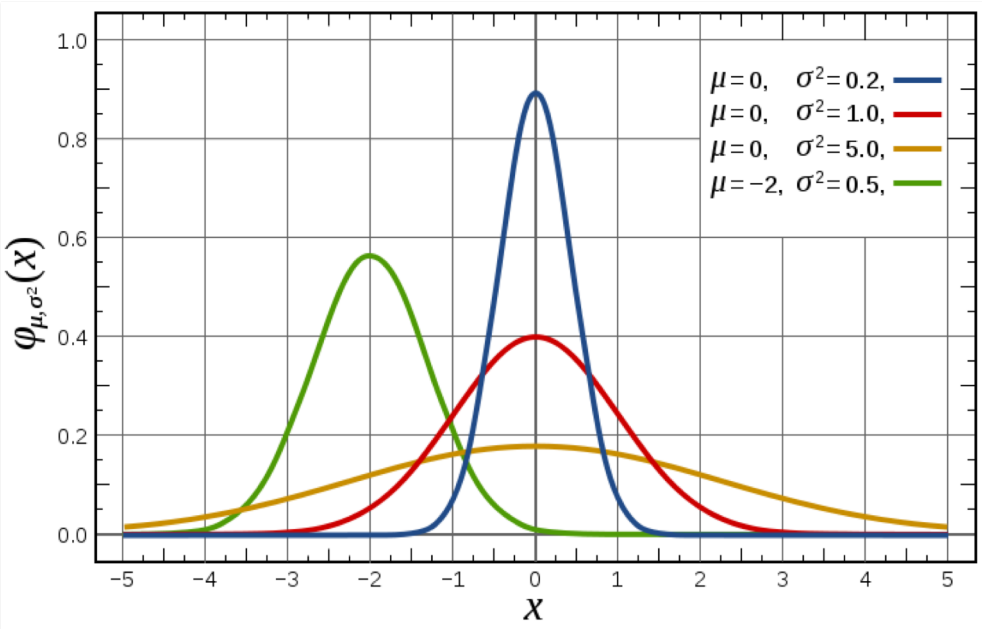


Frequency Distribution →

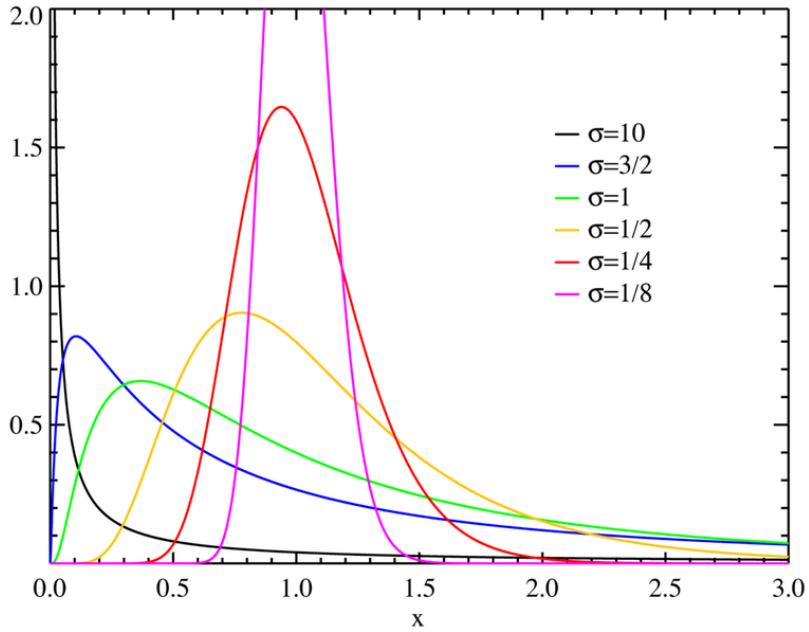


Probability Distribution

Data from life science experiments usually follow normal or lognormal distribution. These distributions are completely determined by the expected values and variances.

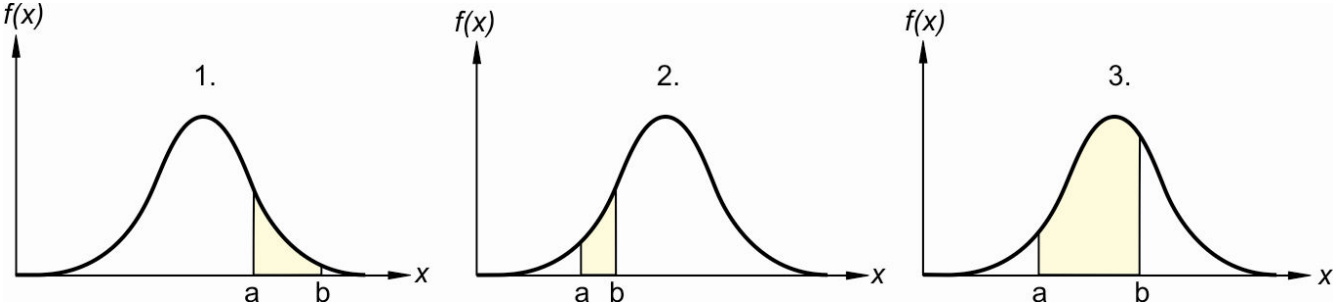


$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\mu - x)^2}{2\sigma^2}}$$

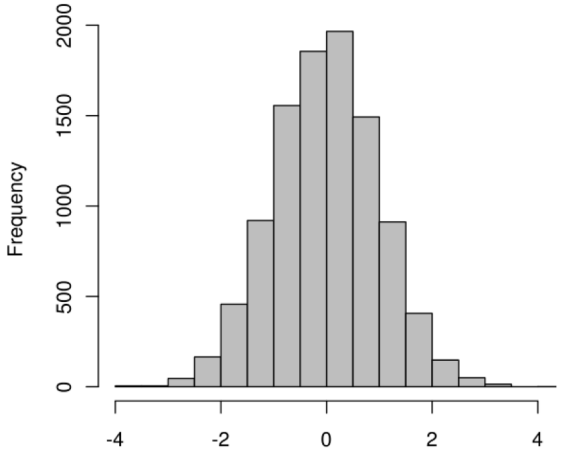


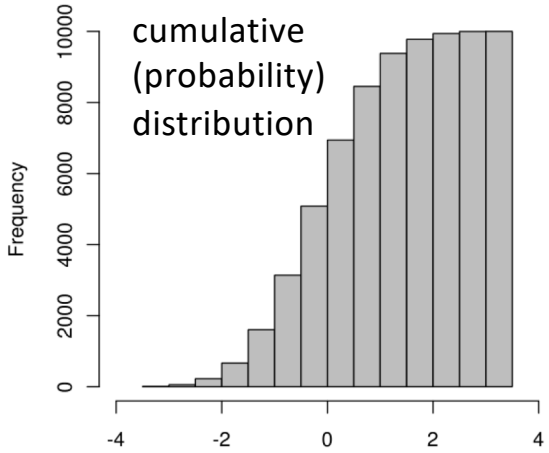
$$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln x - \mu)^2}{\sqrt{2\sigma^2}}}$$

What is the meaning of the distribution function?

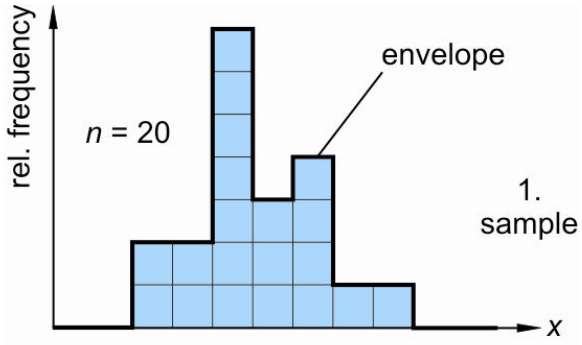


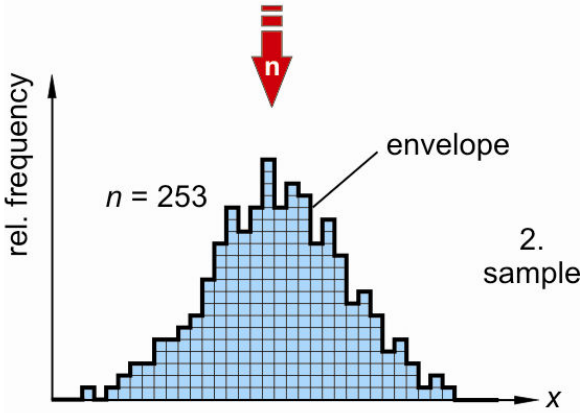
The area under the distribution function in the (a, b) interval is the probability that a randomly chosen sample fells into this interval, if the area under the full function is set to 1.

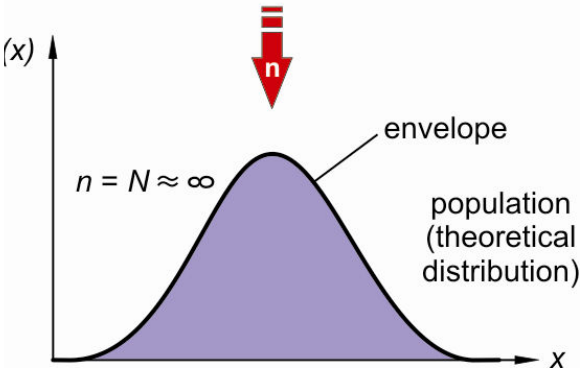




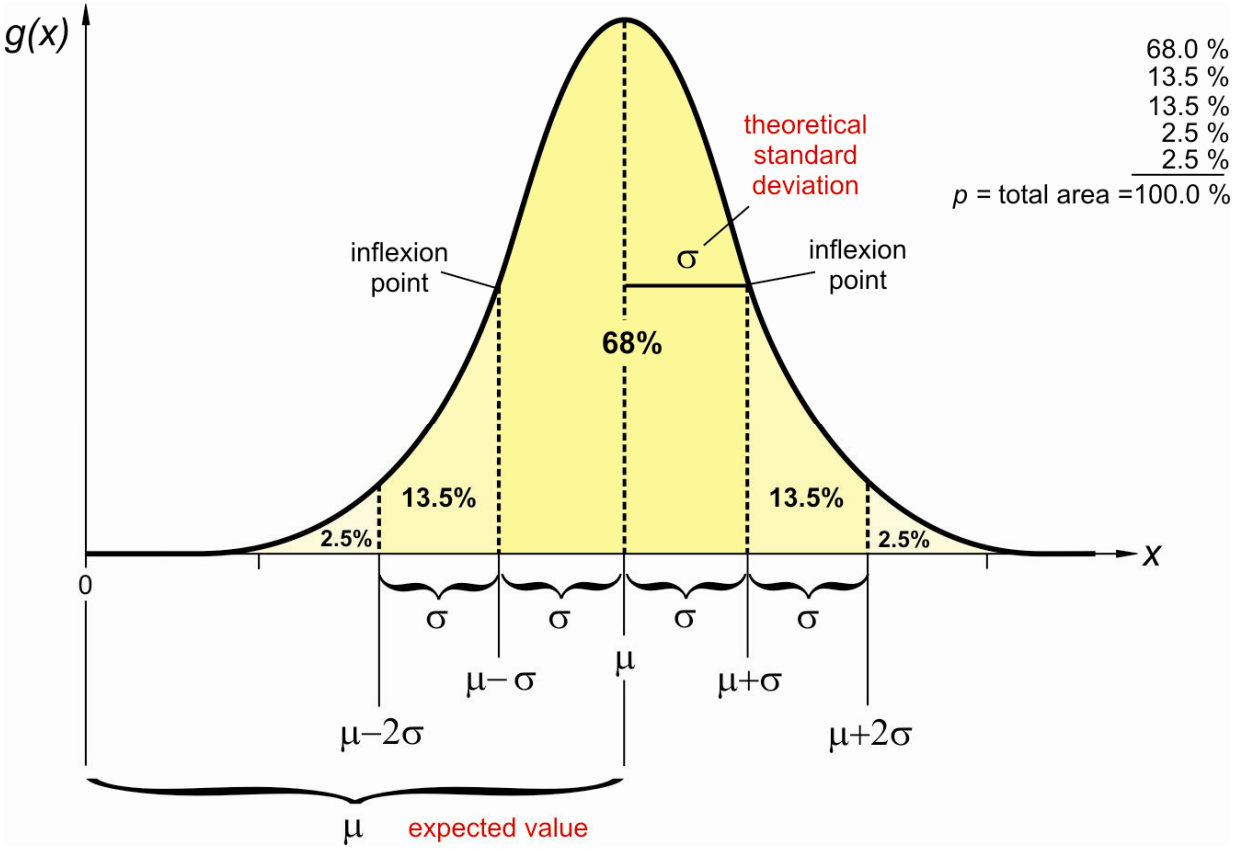
Principal theorem in statistics: by increasing the sample size, the experimental distribution function (envelop of the histogram) will approximate the envelop of theoretical distribution function.







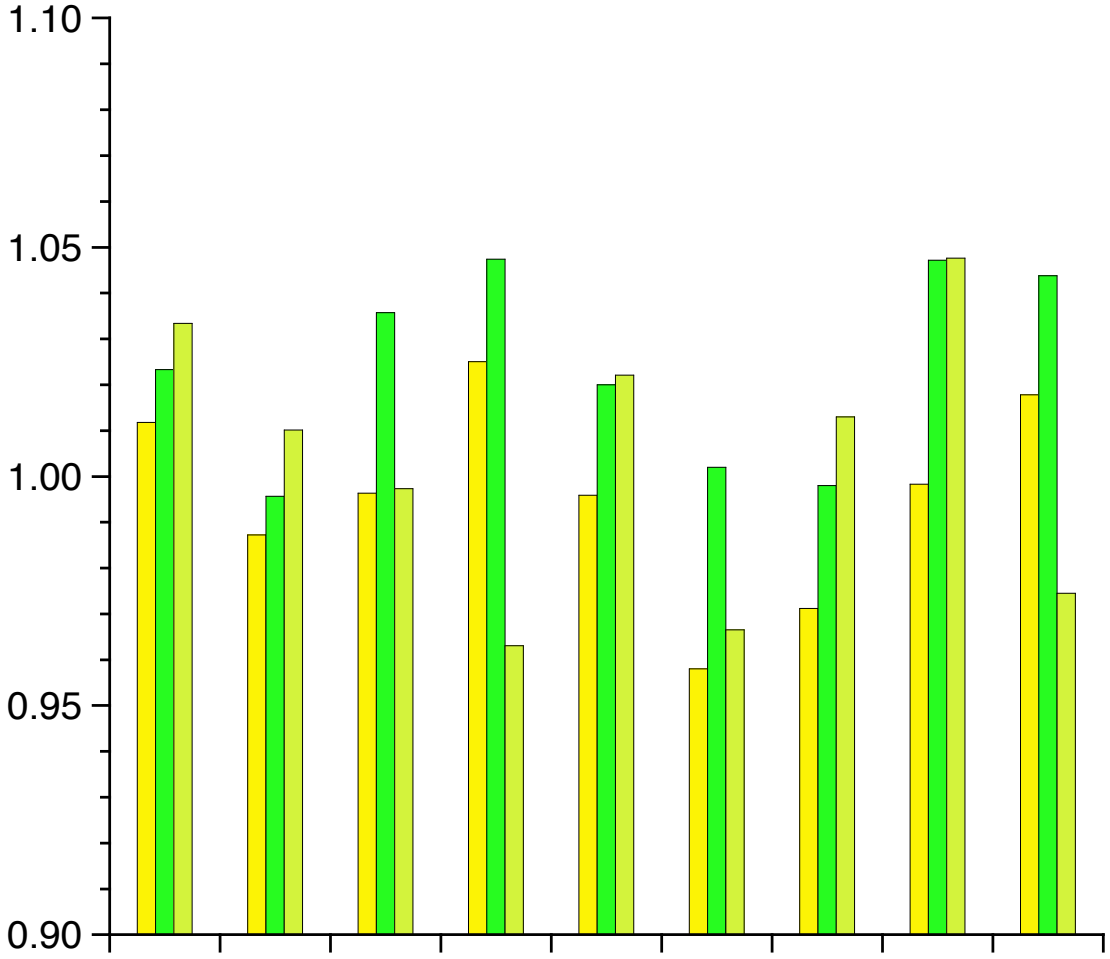
What is the meaning of the distribution function?



Yellow areas: ~95% of data obtained by random sampling from a normal distribution fell within ± 2σ range around the expected value.

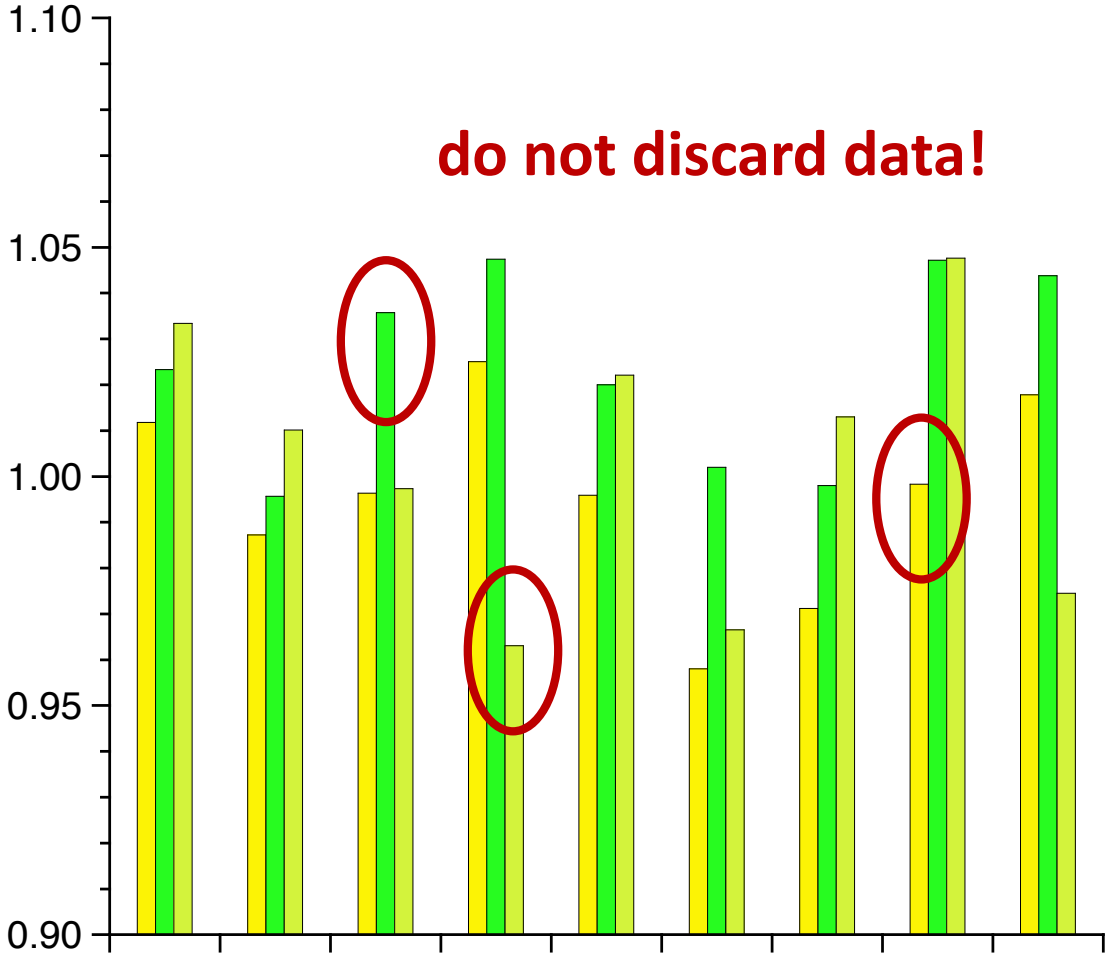
Consequently, data from a small sample very likely fell into this interval. **It is unlikely that we find outliers in a small sample!**

Triplicates from a normal distribution



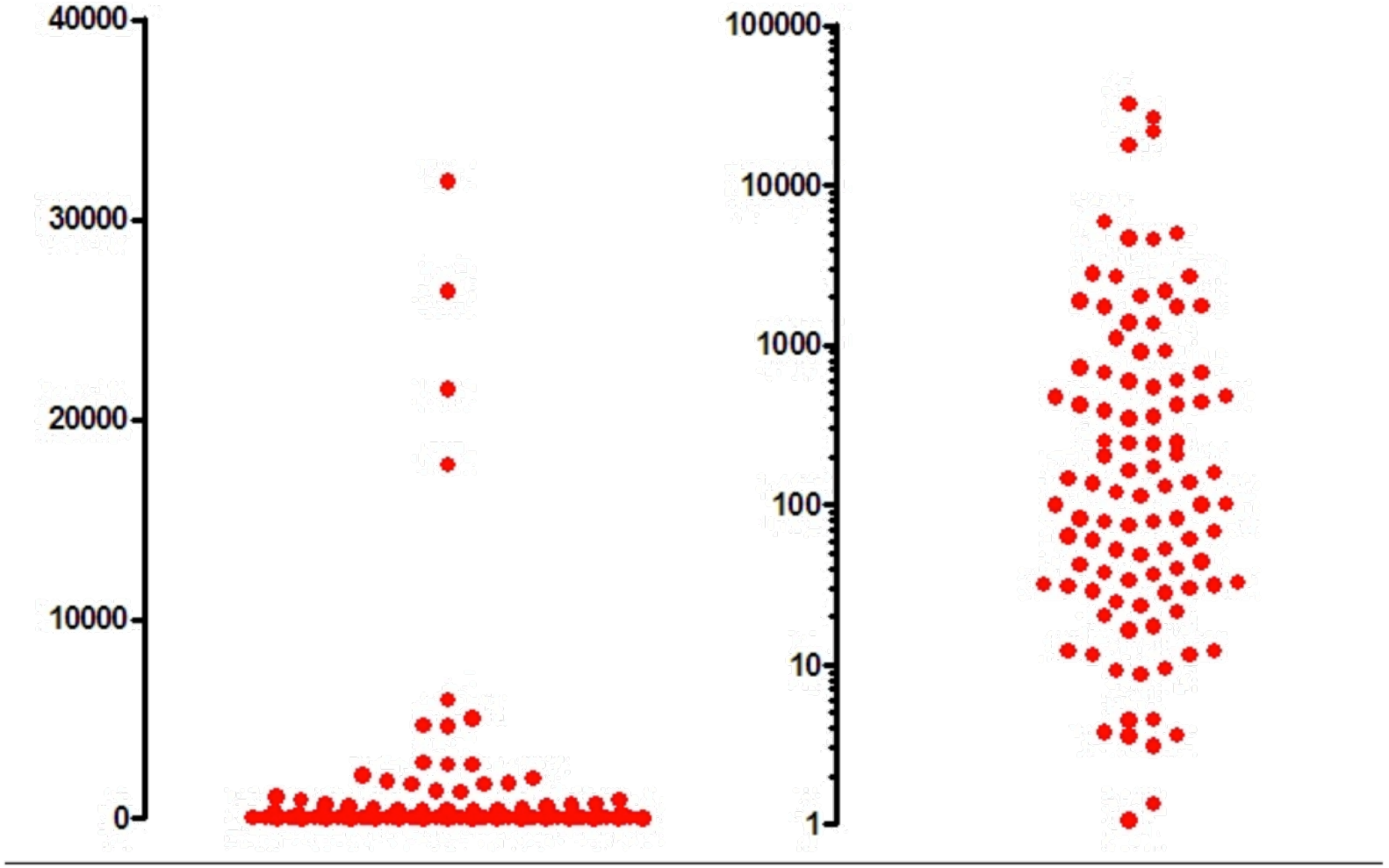
Do not look for pattern!

Be suspicious of yourself - not of your data!



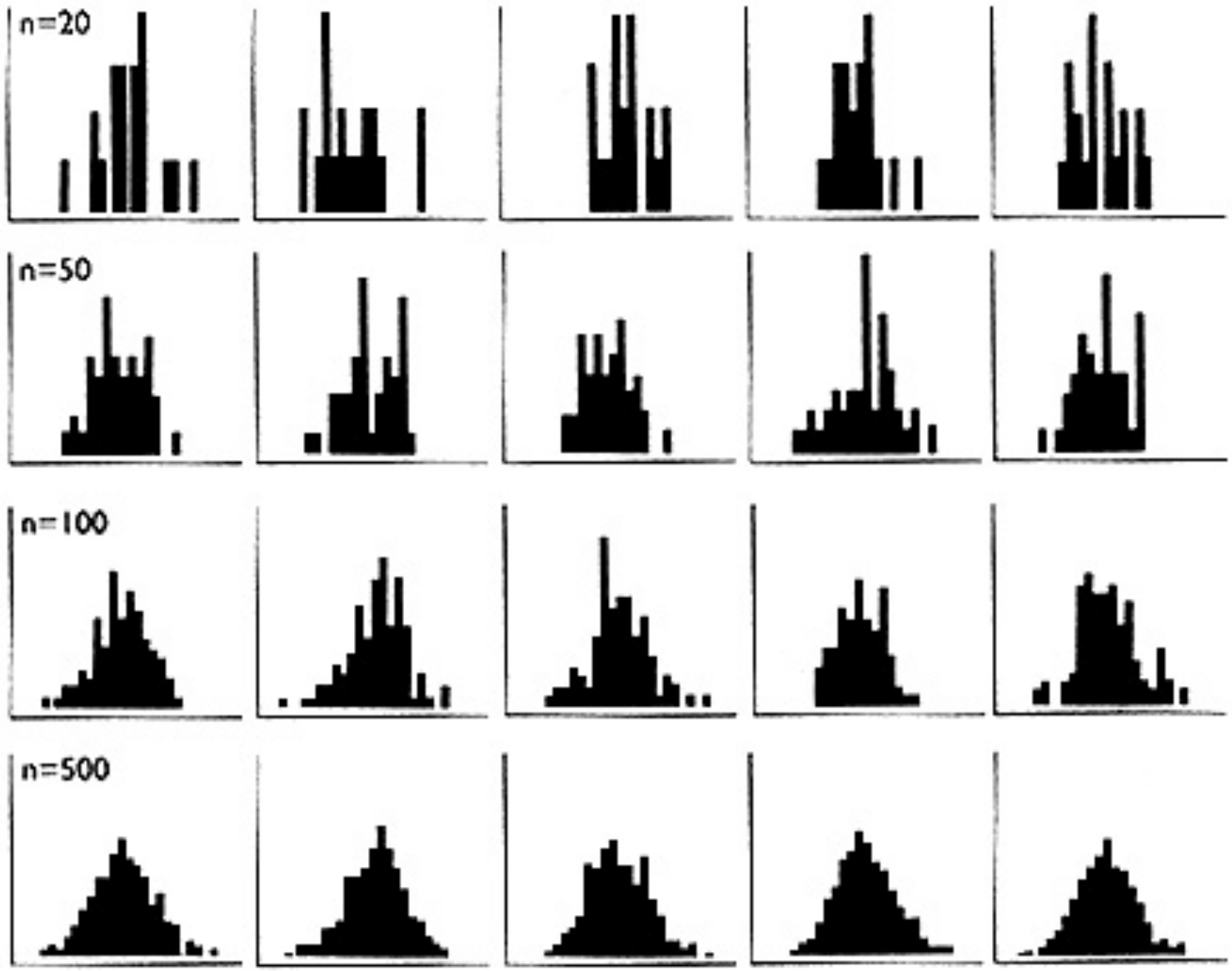
It is unlikely to find outliers among 3-5 data points!

What if data are from lognormal distribution?



Check the distribution by applying logarithmic scaling…

Random sampling from a normal distribution



In life science experiments, n tipically of 3-5 !!!

How to characterize the normal distribution?

we need the expected value and the standard deviation

The mean value is the unbiased estimate of the expected value of a distribution:

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

The unbiased estimate of the theoretical standard deviation is the square root of the variance:

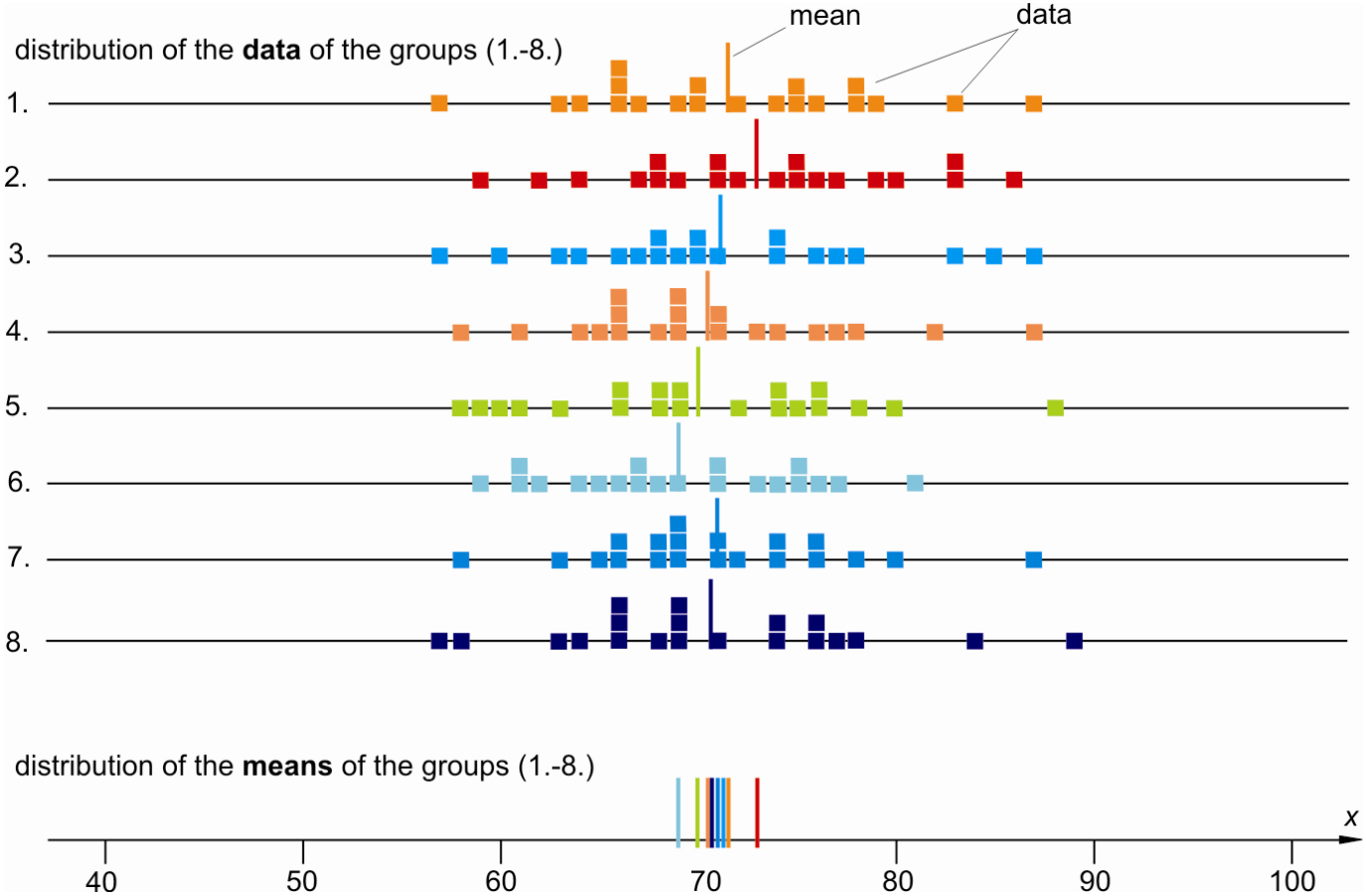
Variance: $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - < x >)^2$$

The sample standard deviation: $s = \sqrt{(s^2)}$ (why n-1?)

Every estimates of the theoretical values are calculated from the sampling.

The sample values are statistical variables. Every quantity calculated from statistical variables must also be statistical variable!

How to characterize the mean value?



spreading of mean (standard error): $S_{\bar{X}} = \frac{S}{\sqrt{N}}$

How to characterize the mean value?

The standard error : **$SE = s/\sqrt{n}$**

(other common abbreviation is S.E.M. = standard error of mean)

The standard error is the sample standard deviation of the mean!!!

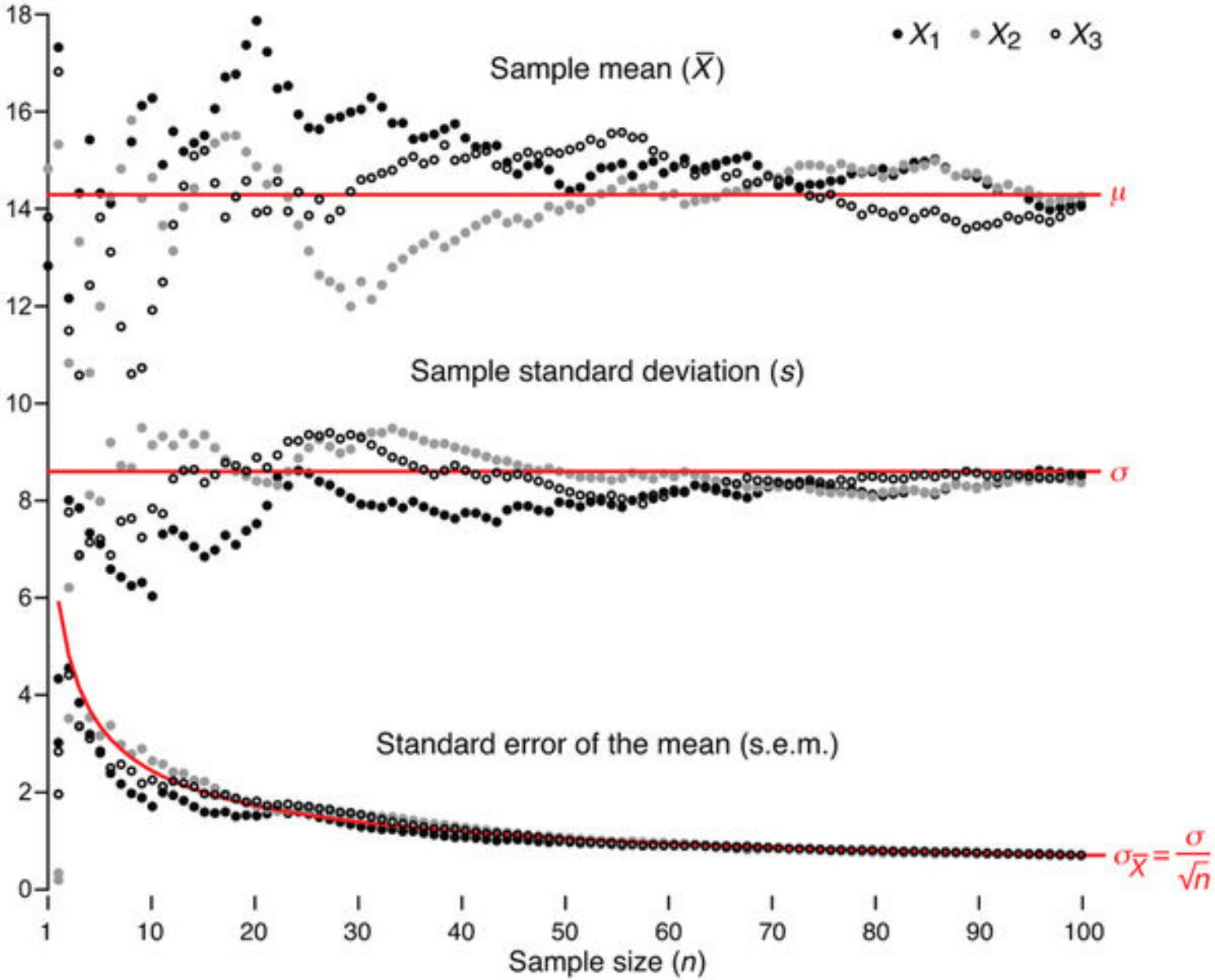
It describes the fluctuation of the mean value around the theoretical expected value of the distribution (and not the dispersion of the data)!

The standard error decreases by increasing the sample size!

Remember: standard deviation decribes the dispersion of data, while standard error is the fluctuation of the mean around the expected values.

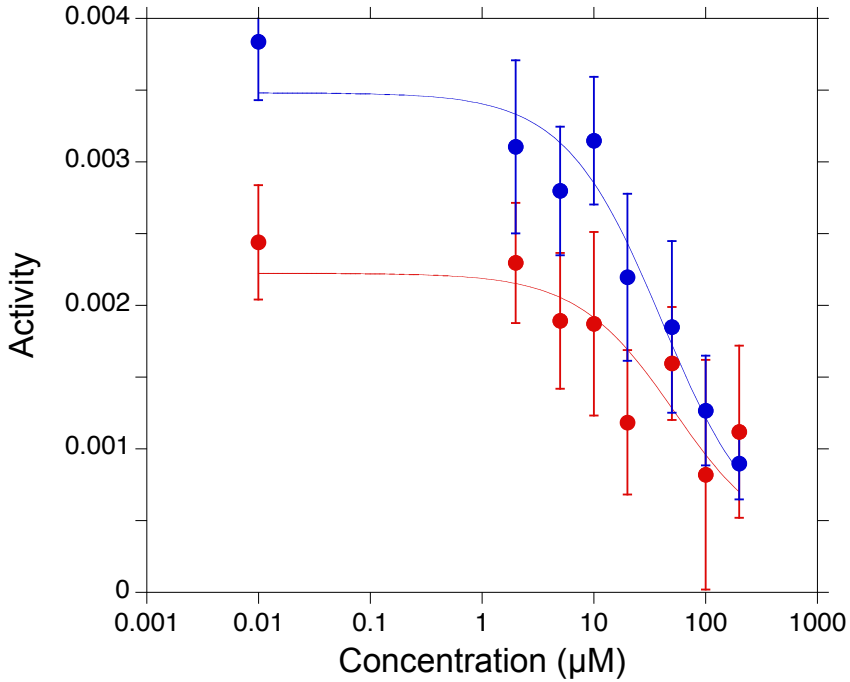
Use standard error when displaying data means, unless it is important to show up the dispersion of the original data!

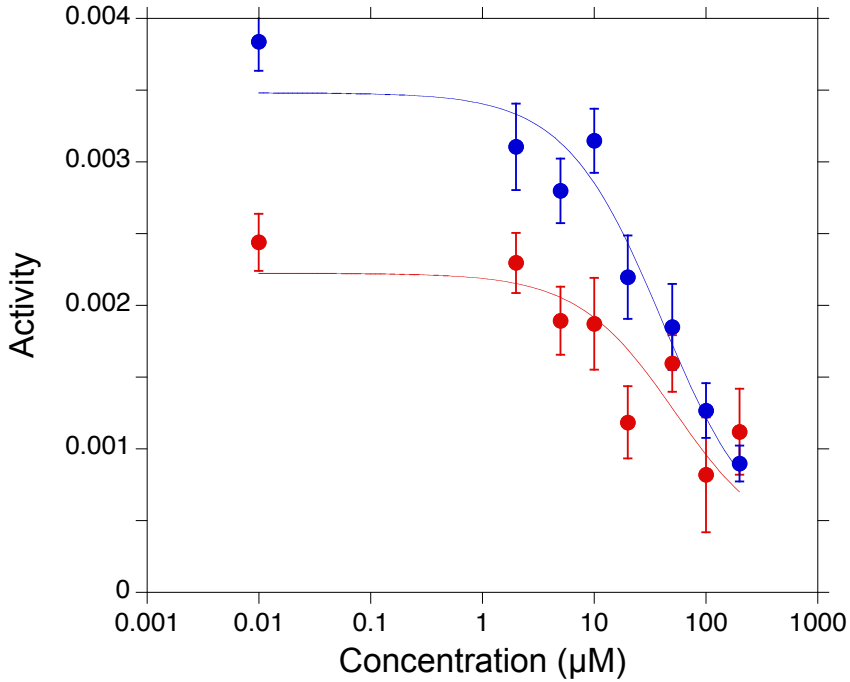
Mean, standard deviation, standard error ~ n



Graphing mean values with SD or SE

standard deviation n=4 standard error





dispersion of the data

dispersion of the mean

Pooled data from “k” data sets

The mean value of k data sets does not depend on the order of averaging:

$$< x >= \frac{1}{k}\sum_{j=1}^{k} < x >_j = \frac{1}{k}\sum_{j=1}^{k}\left(\frac{1}{n_j}\sum_{i=1}^{n_j} x_i\right) = \frac{1}{\sum_{j=1}^{k} n_j}\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_i$$

The sample standard deviation from the set of k parallels of n_j data points each:

$s_k = \sqrt{s_k^2}$, where the combined variance is: $$s_k^2 = \frac{\sum_{j=1}^{k}(n_j - 1)s^2}{\sum_{j=1}^{k}(n_j - 1)}$$

The standard error of the mean of k pooled data sets: $SE_k = s_k / \sqrt{(\Sigma\, n_j - 1)}$

Propagation of errors

Sample standard deviation of the functions of x_i, $y=f(x_i)$:

$$s^2 = \sqrt{\sum (\frac{\partial y}{\partial x_i} s_i)^2} = \sqrt{(\frac{\partial y}{\partial x_1} s_1)^2 + (\frac{\partial y}{\partial x_2} s_2)^2 + \ldots}$$

Consequence: the variances of all experimental errors are additive!

$$s = \sqrt{(s_1^2 + s_2^2 + s_3^2 + \ldots)}$$

The coefficient of variation is the relative error of the sample:

C_V = standard error/mean = s/<x>

Propagation of errors

background substraction and normalization:

1) when subtracting background, the sample standard deviation becomes:

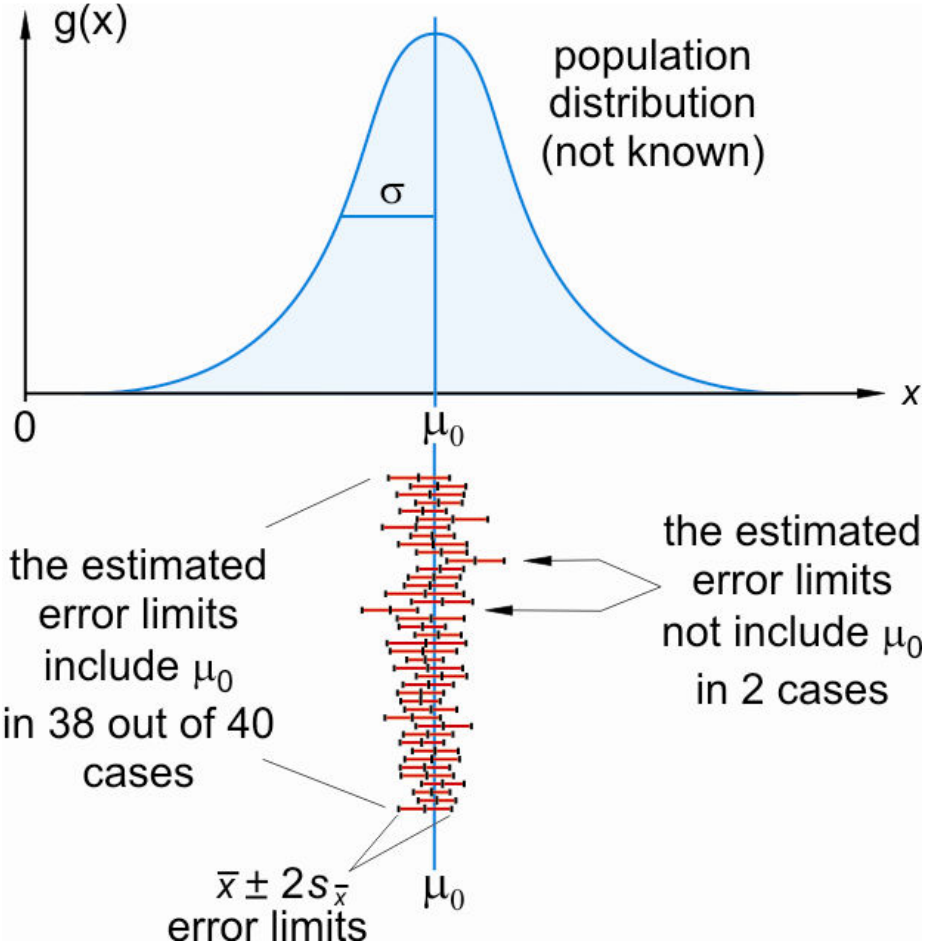
$$s_{korr} = \sqrt{s^2 + s_0^2}$$

2) calculate with the relative standard deviation, when normalizing data:

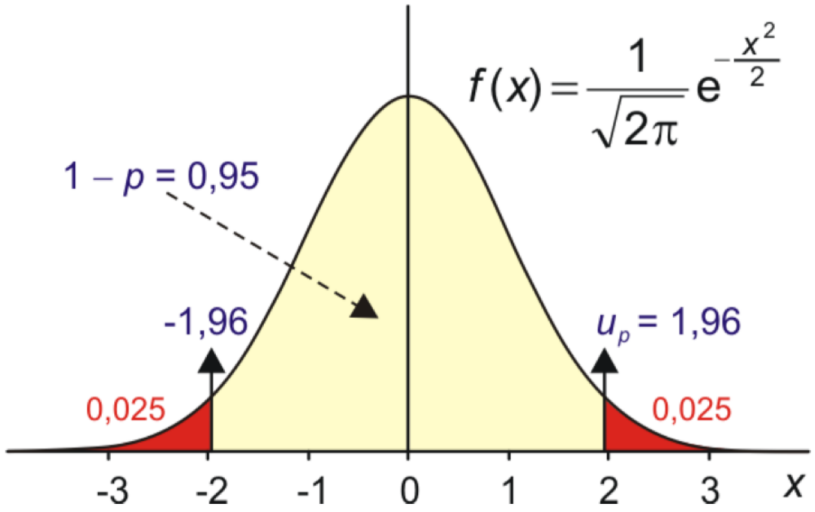
$$s_R = \frac{s}{< x >_0} \Rightarrow s_R = \sqrt{s_R^2 + s_{Ro}^2}$$

Confidence interval

What is the distance of the sample mean and the theoretical expected value?



We chose α error level (let's say 5%). If n is large, than (1-α)100% is the probability that the expected mean value is within the ± 2SE interval of the experimental mean value. We call this as the 95% confidence interval.



The use of 95% confidence level is highly recommended!

Confidence interval for any value of "n"

If we set the error level α, than (1-α)100% is the probability that the expected mean value is within the $\langle x \rangle \pm t_\alpha SE$ interval, where t_α is the Student's t-distribution value at α and (n-1) degree of freedom.

The results of "n" parallel measurements should be given as the mean ± its confidence interval:

$$\langle x \rangle \pm t_\alpha s/\sqrt{n} = \langle x \rangle \pm t_\alpha SE$$

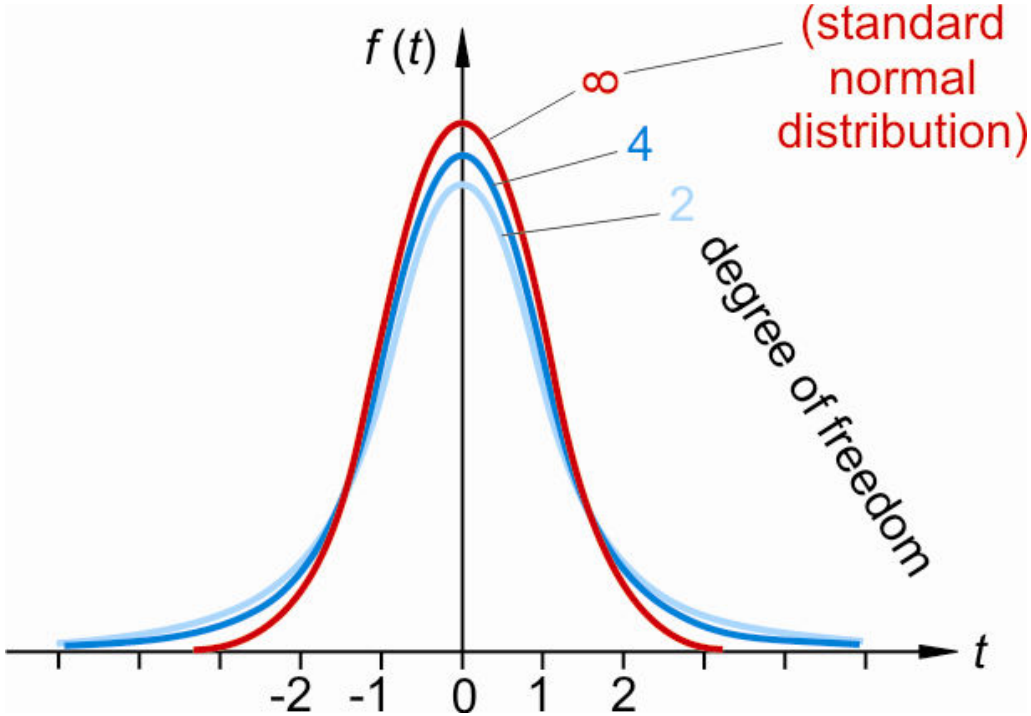
Suggestion: set α always to 0.05 (5%)

How big is the confidence interval?

(1-α)100% confidence level and (n-1) degree of freedom:

$$\langle x \rangle \pm t_\alpha \times SE$$

	$t_{5\%}$	$t_{1\%}$
DF = (n-1)	95%	99%
1	12.71	63.66
2	4.30	9.93
3	3.18	5.84
4	2.78	4.60
9	2.26	3.25
∞	1.96	2.58



Student's t-distribution

How to estimate the necessary number of parallels?

We denote the desired precisity by "h" (h = <x> - μ), "s" is the sample standard deviation, and α is the critical error:

$$\mathbf{n = (t_\alpha s/h)^2} \quad \text{or} \quad \mathbf{n = (t_\alpha s\% / h\%)^2}$$

s	h	95% **n** 99%	
10%	10%	4	7
10%	5%	15	27
10%	1%	384	666
20%	10%	15	27

Rule of thumb at 95% confidence level: $\mathbf{n \approx 4 \times (s\% / h\%)^2}$

Hypothesis testing

Null hypothesis or initial hypothesis (H_0) – we assume a statement is true for the statistical manifold. We must also define an alternative hypothesis H_1 against H_0.

	If H_0 is:	
	accepted	rejected
H_0 is true	right decision	Type-I error
H_0 is false	Type-II error	right decision

The lingo "null hypothesis" comes from the frequent situation that the expected difference is zero (e.g. the sample mean and the expected mean values are the same, there is no effect of a treatment, etc). The alternative hypothesis usually opposing the null hypothesis.

The probability of Type-I error is α (the critical error level). The probability of Type-II error can not be given in general, it depends on “n” and the form of H_1.

How a hypothesis testing works?

• We define H_0 and calculate the test-statistics value from our data and compare it to the theoretical value at the maximal α error level (the probability of making a type-I error) at the given degree of freedom. We always have to define H_1 against H_0 as well.

• If H_0 is true, then the probability that the empirical value of the test statistics exceeds the theoretical value is α.

• **If the relation "empirical (calculated) value" < "theoretical (critical) value" is true, than we can say: our data do not contradict to H_0 at (1-α)100% confidence (or probability) level.**

• **Never say that a test statistics proves your assumption!!!**

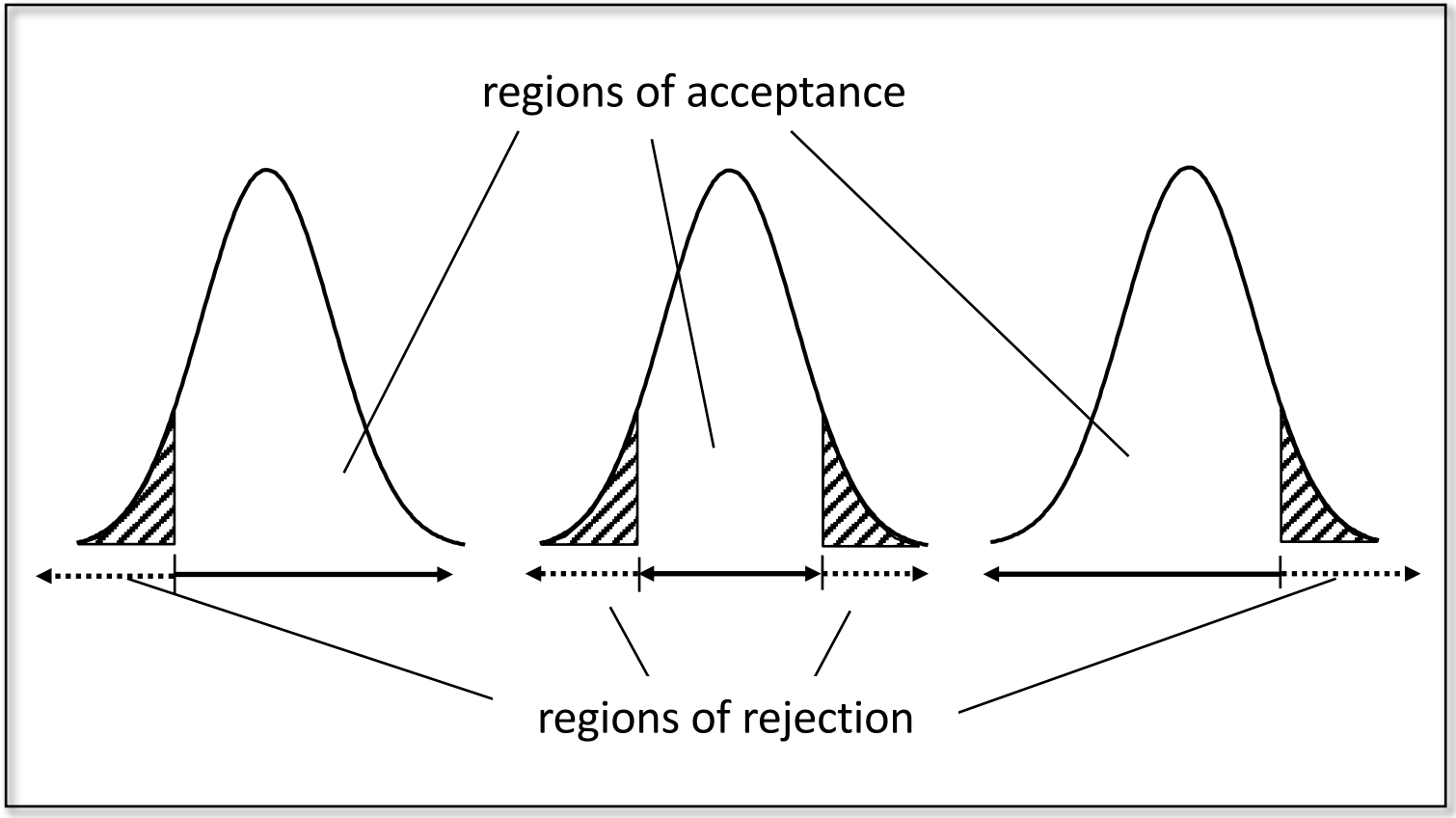
• Always define precisely H_0 és H_1 as pairs! For example, these two hypothesis-pairs define different tests:

H_0: $\mu 1-\mu 2=0$ and H_1: $\mu 1-\mu 2>0$ or H_0: $\mu 1-\mu 2=0$ and H_1: $\mu 1-\mu 2\neq 0$

How a hypothesis testing works?

H_0: μ1-μ2=0 and H_1: μ1-μ2>0 or H_0: μ1-μ2=0 and H_1: μ1-μ2≠0

one-tailed test two-tailed test



Unless there is a clear indication for a one-tailed test (e.g. testing the effect of antipyretic or antihypertensive drugs, etc), always use the two-tailed test.

The most common statistical tests are:

- Comparing means - Student's t test / ANOVA
- Comparing variances - Fischer's F test
- Assumptions about a distribution

  or comparing distributions - χ^2 tests

Student's t test

$$t = \frac{< x > - \mu}{s}\sqrt{n} = \frac{< x > - \mu}{SE}$$

If <x> = μ, than t = 0 ! But the mean is a statistical variable, its value fluctuates around μ. For any α error level the critical value of the t test (t_α^{crit}) are calculated (tabulated).

If $|t| \leq t_\alpha^{crit}$, we accept H_0 at (1-α)100% confidence level.

Choosing the H_0: <x> - μ = 0 and H_1: <x> - μ ≠ 0 hypothesis pair at 95% confidence level is highly recommended!

Comparing two groups – significant difference

$$t = \frac{< x >_1 - < x >_2}{s_d}$$

$n_1 \neq n_2$

$$s_d = \sqrt{\frac{(n_1 - 1)s_1{}^{*2} + (n_2 - 1)s_2{}^{*2}}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

DF = $n_1 + n_2 - 2$

$n_1 = n_2 = n$

$$s_d = \sqrt{\frac{s_1{}^{*2} + s_2{}^{*2}}{n}}$$

DF = 2n - 2

For the H_0: μ1-μ2 = 0 and H_1: μ1-μ2 ≠ 0 hypothesis pair: If the empirical t value in absolute term larger than the critical t-value (**| t | > t_α**) than we say the difference is significant at (1-α)100% confidence.

Again, it is important to emphasize that a statistical test never proves or disproves our hypothesis, only contradict or do not contradict to it…

Comparing variances: Fischer's F-test

Just like anything else derived from experimental data, variance (or sample standard deviation) is also a statistical variable. To test whether the difference in two variances is due to sampling (random errors) - do a Fischer's F test:

H_0: $\sigma_1 = \sigma_2$ or H_0: $\sigma_1/\sigma_2 = 1$ (with H_1 being the opposite statement)

The test statistics is: $F = s_1/s_2$ (use indexing that $s_1 > s_2$)

Compare the calculated F value to the critical ones given in F tables. If the empirical F is smaller then the critical value in absolute term then we can say that these two samples are from distributions of equal variances at $(1-\alpha)100\%$ confidence. This is true this way to normal distribution only. Note that the degree of freedom for the nominator and denominator can be different (n_1-1 and n_2-1).

Assumptions of Student's t test

The basic assumption of Student's t test that data are obtained by sampling normal distribution with equal variances. Usually, this requirement holds because of the experimental design. This assumption can be tested by F test.

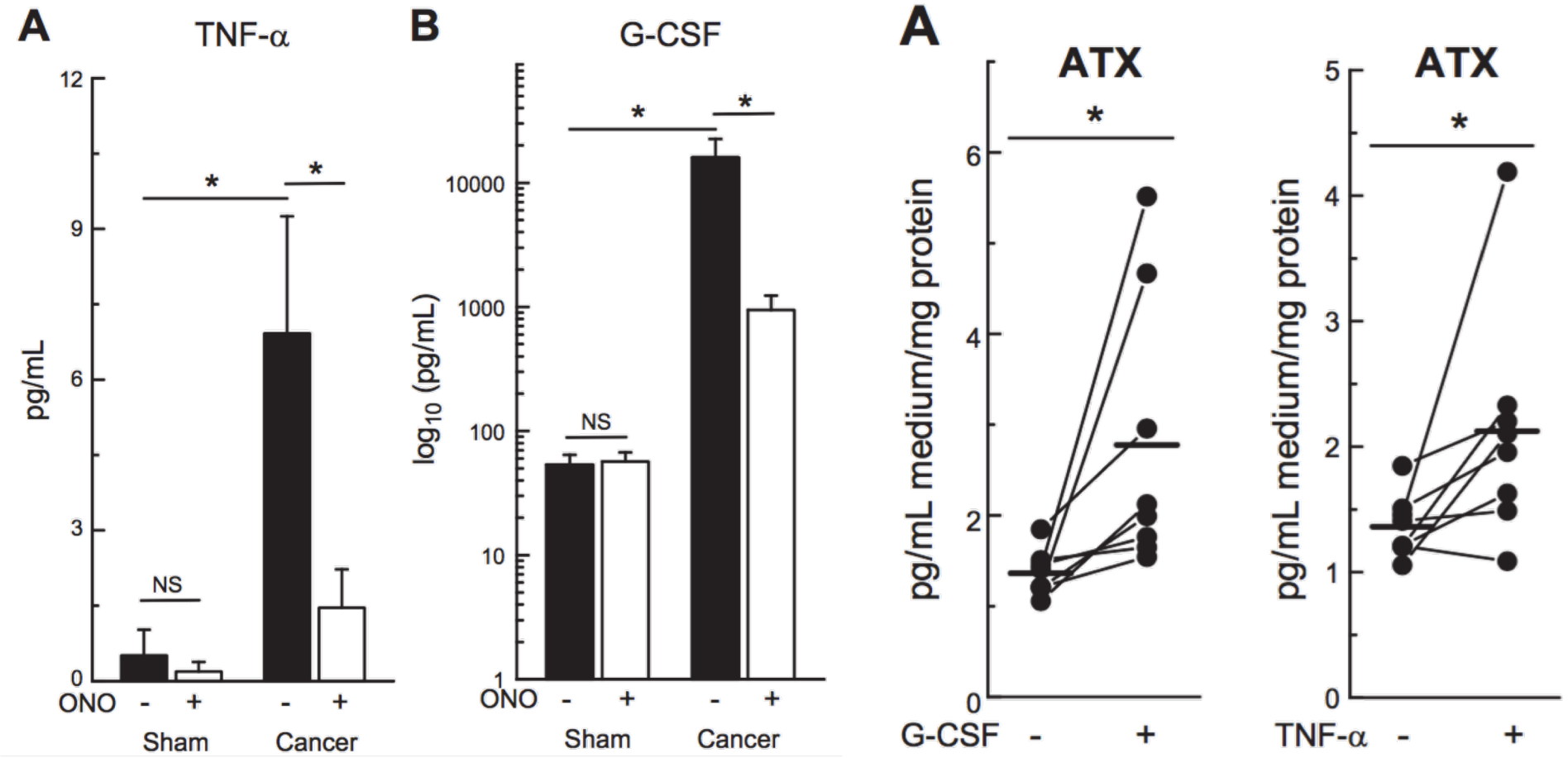
In case the variances are different (e.g. treatment changes the distribution, or variance changes from preparation to preparation, etc), but the samples are from normal distribution, the Welch-corrected t test can be used.

If the distribution of samples are not known, or data are not sampled from normal distribution, than we must choose the non-parametric version of the t test = the Mann-Whitney test. This test compare ranks only and does not assume normal distribution (i.e. any distribution) for the samples.

Varieties of Student's t test

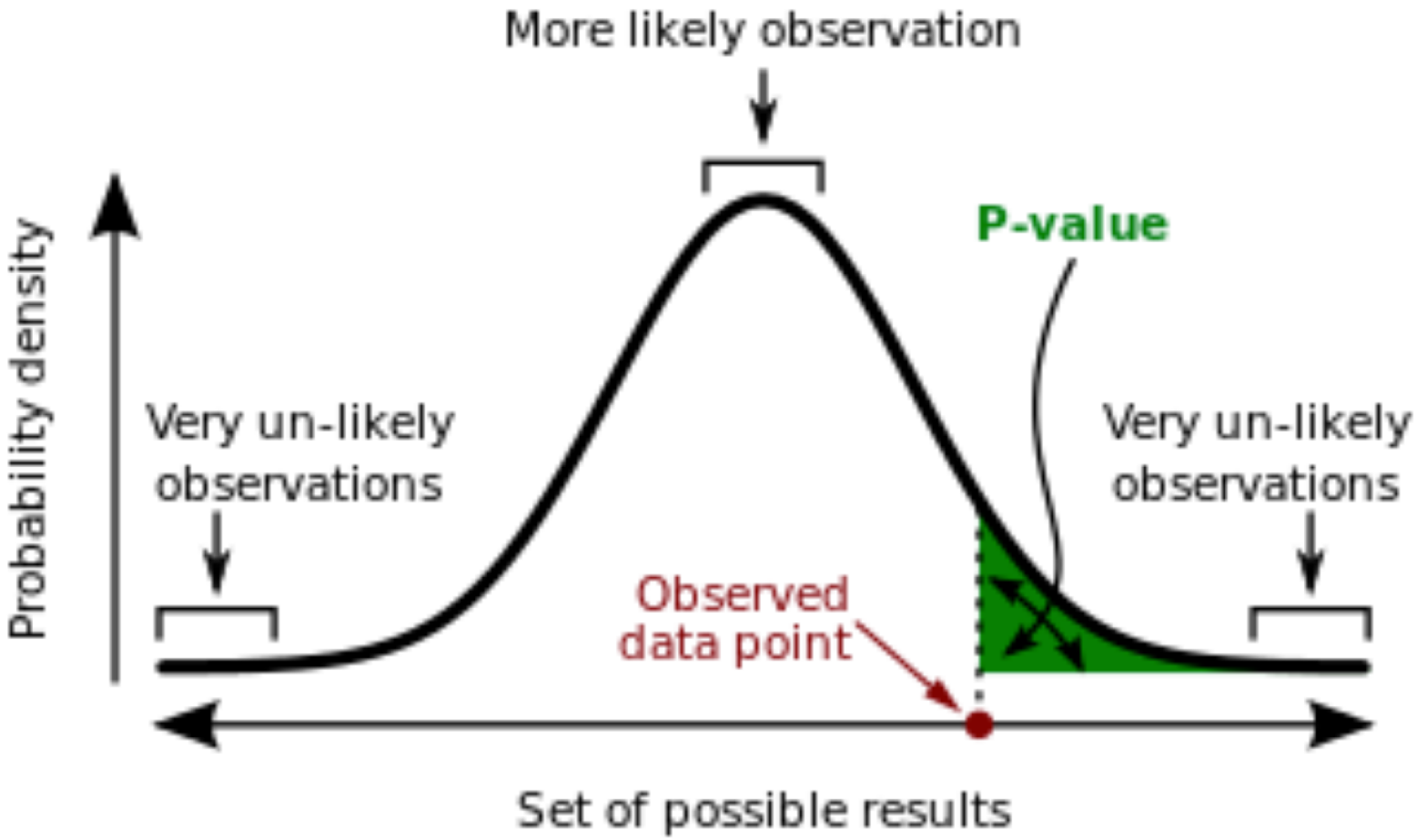
unpaired groups

paired groups



If data are paired (the sample is measured, then treated, then measured again), than do a regular unpaired t test for the differences (or ratios) with H_0: $\langle \Delta x \rangle = 0$.

p value



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

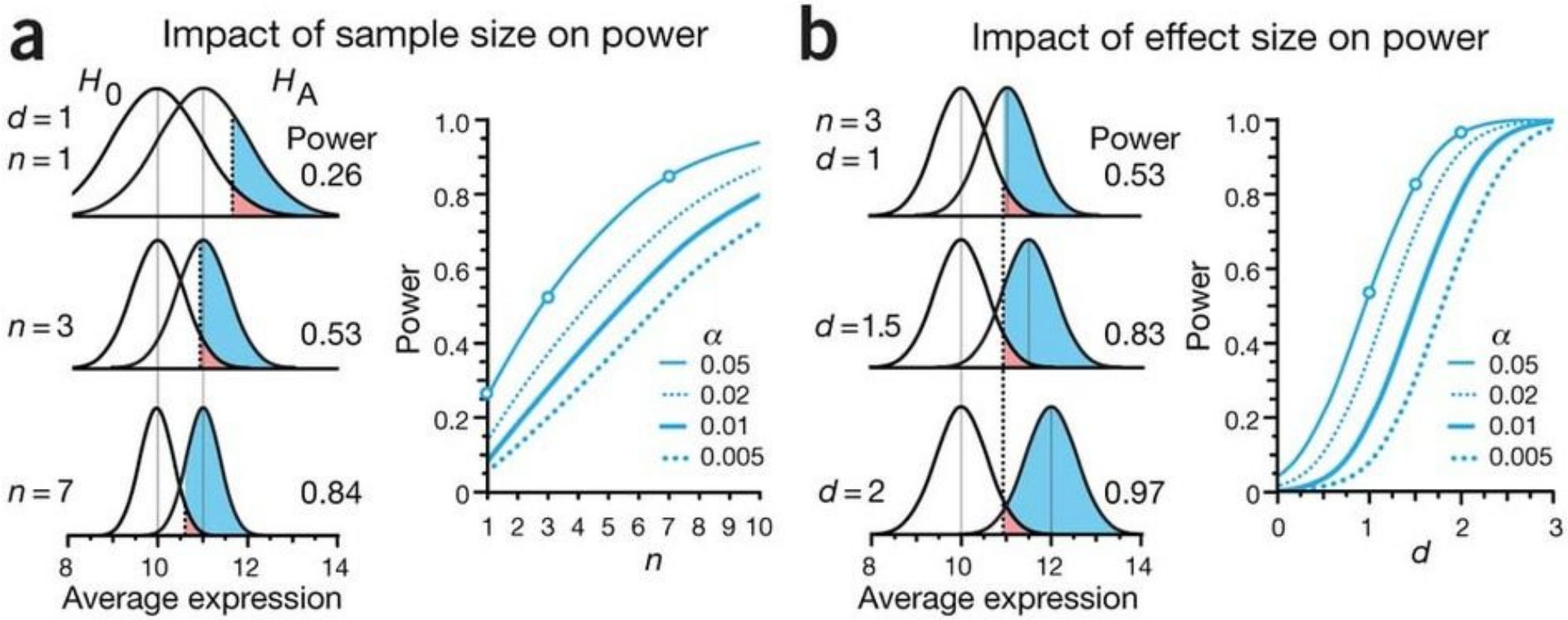
Never rely only on p-value…

The irony is that when UK statistician Ronald Fisher introduced the *P* value in the 1920s, he did not mean it to be a definitive test. He intended it simply as an informal way to judge whether evidence was significant in the old-fashioned sense: worthy of a second look. The idea was to run an experiment, then see if the results were consistent with what random chance might produce. Researchers would first set up a 'null hypothesis' that they wanted to disprove, such as there being no correlation or no difference between two groups. Next, they would play the devil's advocate and, assuming that this null hypothesis was in fact true, calculate the chances of getting results at least as extreme as what was actually observed. This probability was the *P* value. The smaller it was, suggested Fisher, the greater the likelihood that the straw-man null hypothesis was false.

The P value was never meant to be used the way it is used today!

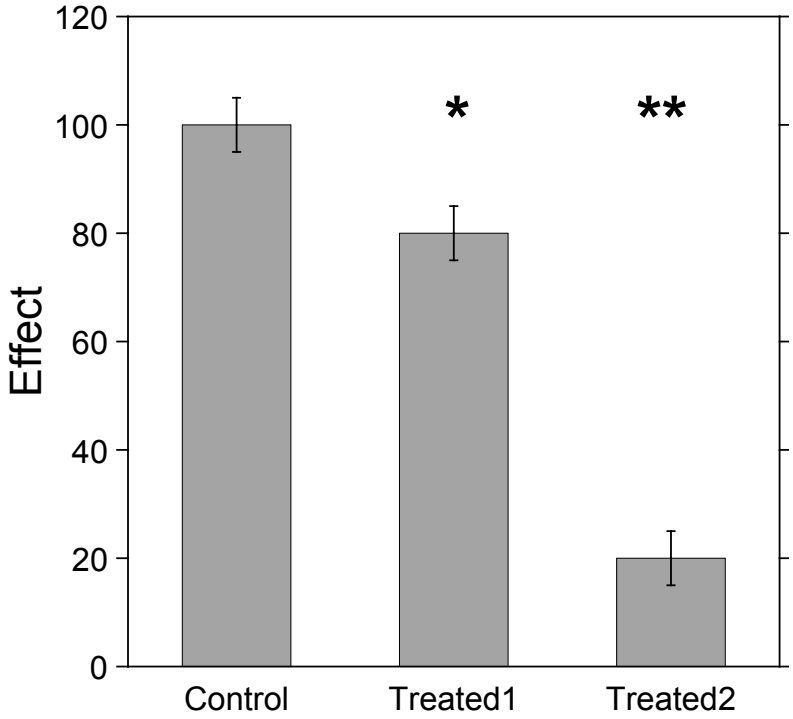
statistical power

What is the percentage of cases when a test leads to the acceptance of true null hypothesis - it depend on α, "n", and on the effect size



effect size, d = (μ_0 - μ_A)/ σ

A popular mistake: “more” significant



Significance is like homology: either exist or does not exist at a certain error level. Do not attempt to quantify the significance of the differences with a statistical test parameters, like the p value. Choose α for type-I error (e.g. α = 0.05) and either calculate the empirical *t* value and compare it to its critial value, or display the *p* value but make statistical deceison only by comparing p to α !

In a set of experiment is unlikely that parts of it have different statistical power!

* p<0.05 = significant, **=p<0.01 = more significant – NEVER say so

If we use α = 0.05 error level, which is highly recommended in biology, then p=0.04 or p=0.002 have the same meaning, the difference can be significant at 95% confidence level.

χ^2 test

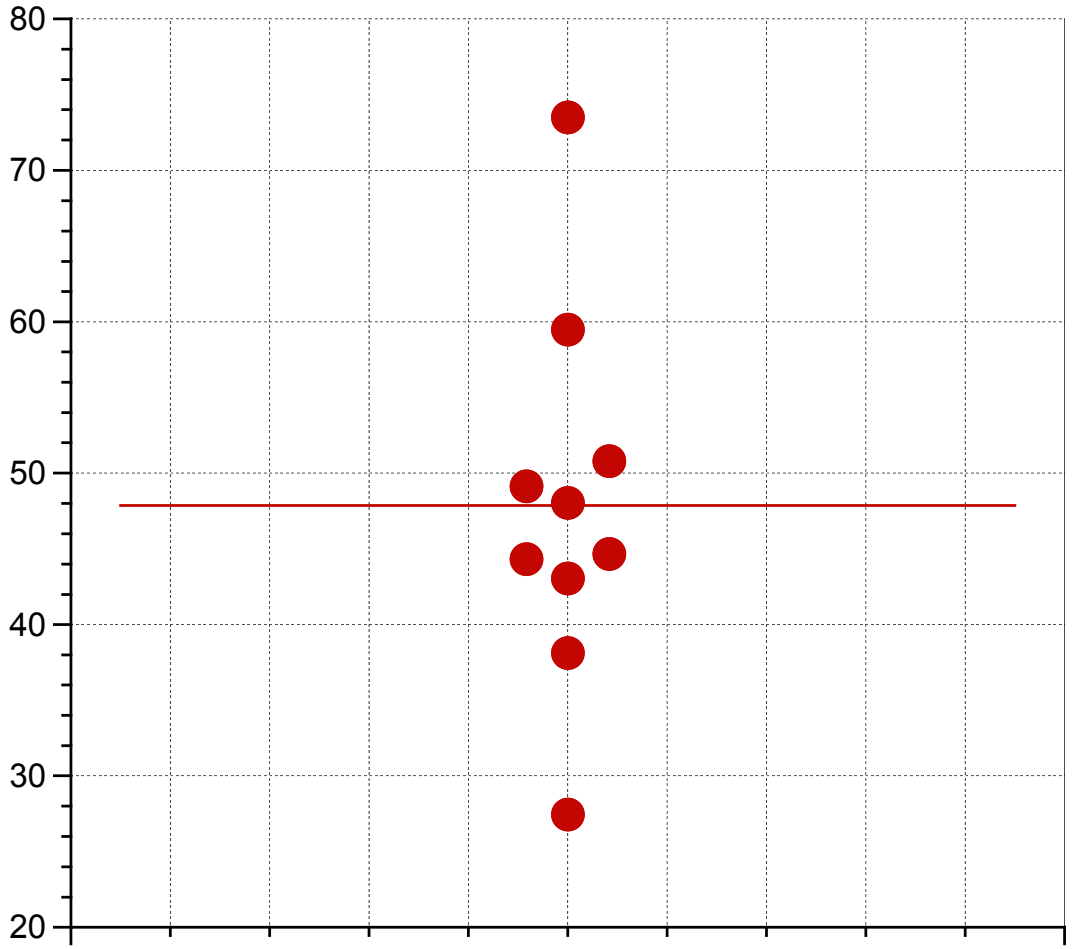
$$\chi^2 = \xi_1^2 + \xi_2^2 + \cdots + \xi_n^2$$

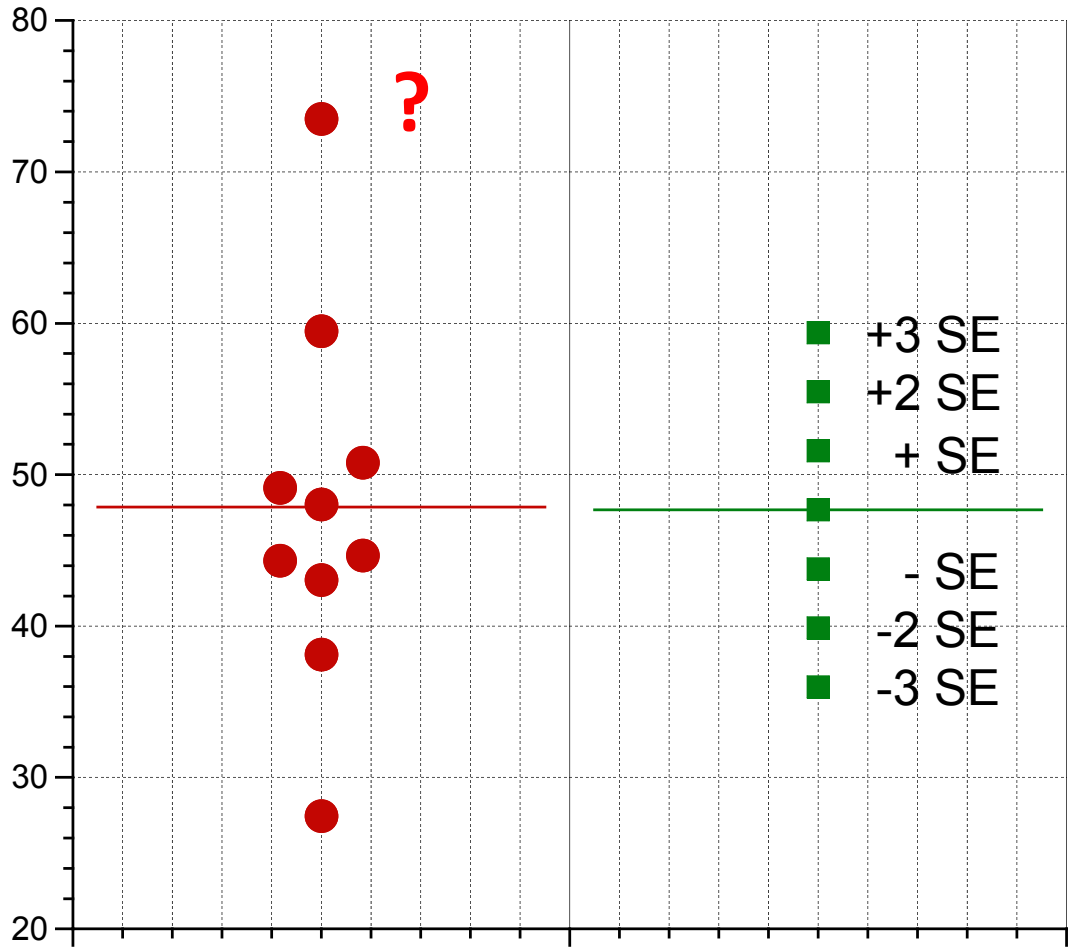
To check that our data are from a normal distribution, or are from the same (normal) distribution.

In regression analysis χ^2 test can be used to estimate the goodness of fit
- If the regression model we fitted to the data is correct then the residuals should follow a normal distribution with zero expected value:

$$\chi^2 = \sum_{i=1}^{n} (\Delta y_i)^2$$

Example1: Is there any outlier in this data set?





It looks very likely by eye! What to do to decide?

Use the Gauss' g-statistics: Denote the suspicious point by x_0, calculate <x> and s from the rest of the data. Then:

$$g = \frac{|x_0 - < x >|}{s}$$

If calculated g-value exceeds the critical value in the table, the x_0 probably is an outlier…

n	g
3	46.7
4	10.1
5	6.51
6	5.31
7	4.73
8	4.40
9	4.18
10	**4.04**

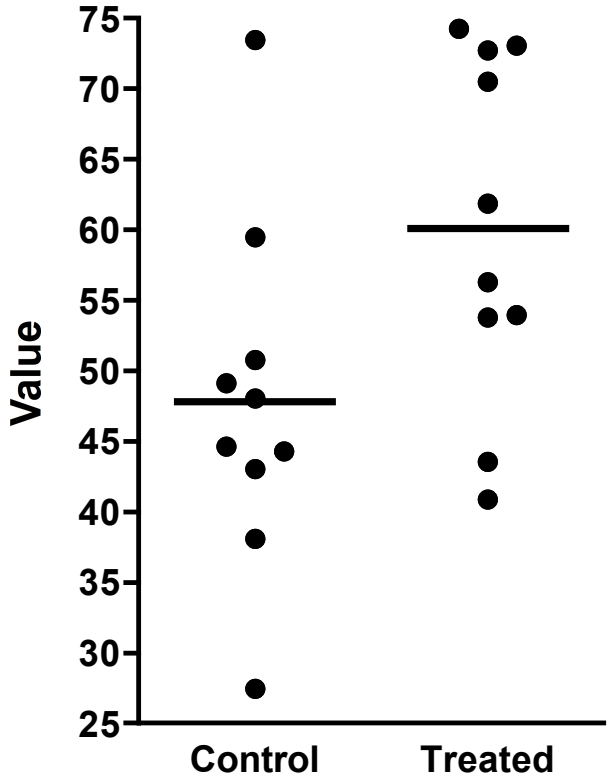
50.791
44.672
48.050
73.504
59.496
27.462
38.133
49.142
43.052
44.323

$$g = \frac{|73.504 - 45.014|}{8.869} = 3.21$$

It is probably NOT an outlier, though it looked so!

Example2: two sets of 10 samples, one received placebo, the other received treatment

Student's t test (unpaired, two-tailed, α=0.05)



	Control	Treated
Mean	47.86	60.11
Std. Deviation	12.29	12.36

Mean ± SEM of column A	47.86 ± 3.887 N=10
Mean ± SEM of column B	60.11 ± 3.908 N=10
Difference between means	-12.24 ± 5.512
95% confidence interval	-23.83 to -0.6637

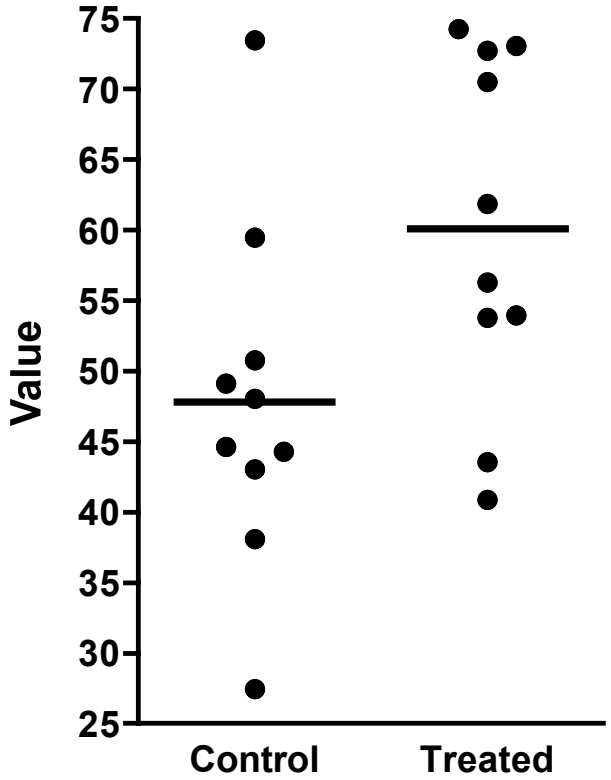
P value	0.0394
P value summary	*
Are means signif. different? (P < 0.05)	Yes
One- or two-tailed P value?	Two-tailed
t, df	t=2.221 df=18

Table Analyzed	TTEST.TXT
Column A	Control
vs	vs
Column B	Treated
Unpaired t test	
P value	0.0394
P value summary	*
Are means signif. different? (P < 0.05)	Yes
One- or two-tailed P value?	Two-tailed
t, df	t=2.221 df=18
How big is the difference?	
Mean ± SEM of column A	47.86 ± 3.887 N=10
Mean ± SEM of column B	60.11 ± 3.908 N=10
Difference between means	-12.24 ± 5.512
95% confidence interval	-23.83 to -0.6637
R squared	0.2152
F test to compare variances	
F,DFn, Dfd	1.011, 9, 9
P value	0.9874
P value summary	ns
Are variances significantly different?	No

Graphpad Prism

What if the same sets of data were obtained from a paired experiment:

Student's t test (paired, two-tailed, α=0.05)

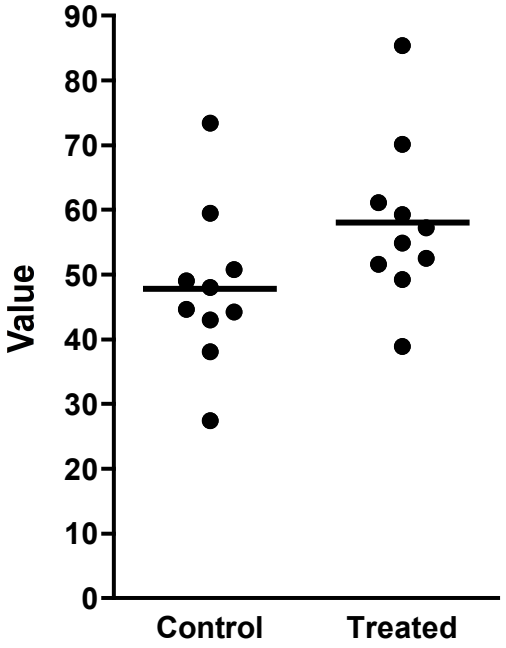


	Control	Treated
Mean	47.86	60.11
Std. Deviation	12.29	12.36

How big is the difference?	
Mean of differences	-12.24
95% confidence interval	-25.04 to 0.5498
R squared	0.3424

P value	0.0586
P value summary	ns
Are means signif. different? (P < 0.05)	No
One- or two-tailed P value?	Two-tailed
t, df	t=2.165 df=9

What if the data are obtained from real paired sets:

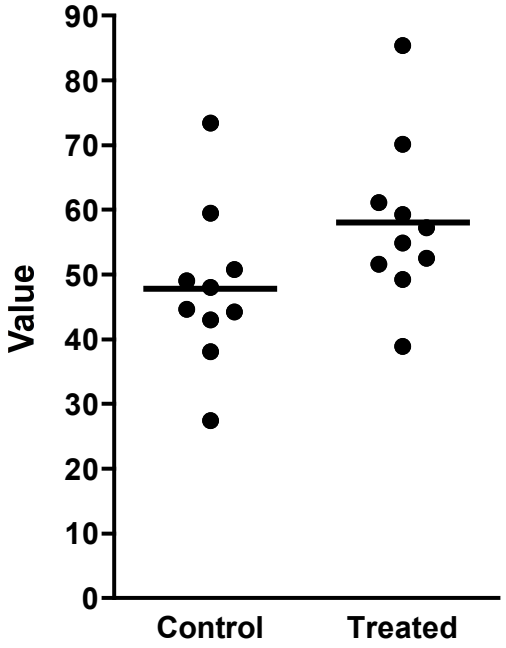


	Control	Treated
Mean	47.86	58.09
Std. Deviation	12.29	12.63

How big is the difference?	
Mean of differences	-10.23
95% confidence interval	-11.50 to -8.962
R squared	0.9737

P value	P<0.0001
P value summary	***
Are means signif. different? (P < 0.05)	Yes
One- or two-tailed P value?	Two-tailed
t, df	t=18.27 df=9

paired t test



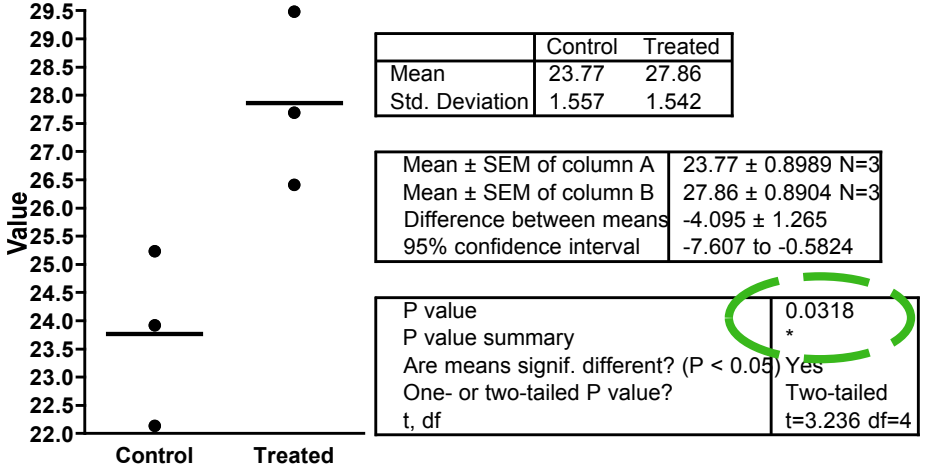
	Control	Treated
Mean	47.86	58.09
Std. Deviation	12.29	12.63

Mean ± SEM of column A	47.86 ± 3.887 N=10
Mean ± SEM of column B	58.09 ± 3.995 N=10
Difference between means	-10.23 ± 5.574
95% confidence interval	-21.94 to 1.483

P value	0.0831
P value summary	ns
Are means signif. different? (P < 0.05)	No
One- or two-tailed P value?	Two-tailed
t, df	t=1.835 df=18

unpaired t test

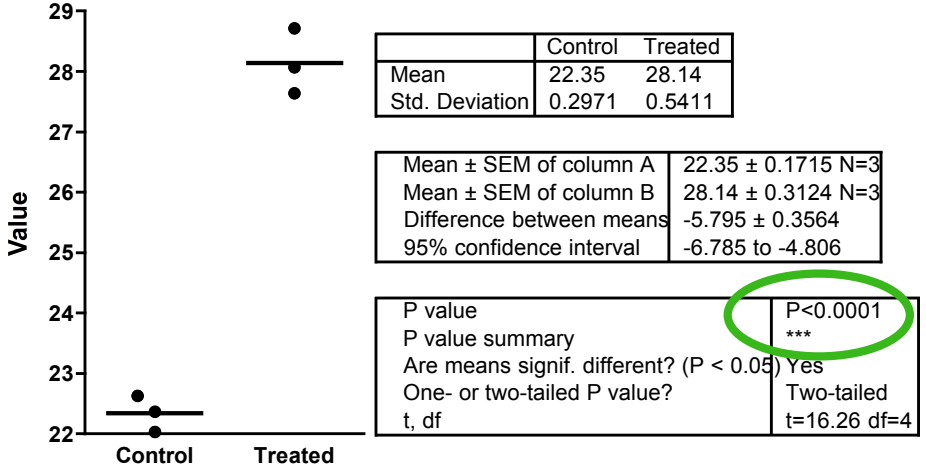
Example 3: triplicates of placebo and treated, compare one-by-one



	Control	Treated
Mean	23.77	27.86
Std. Deviation	1.557	1.542

Mean ± SEM of column A	23.77 ± 0.8989 N=3
Mean ± SEM of column B	27.86 ± 0.8904 N=3
Difference between means	-4.095 ± 1.265
95% confidence interval	-7.607 to -0.5824

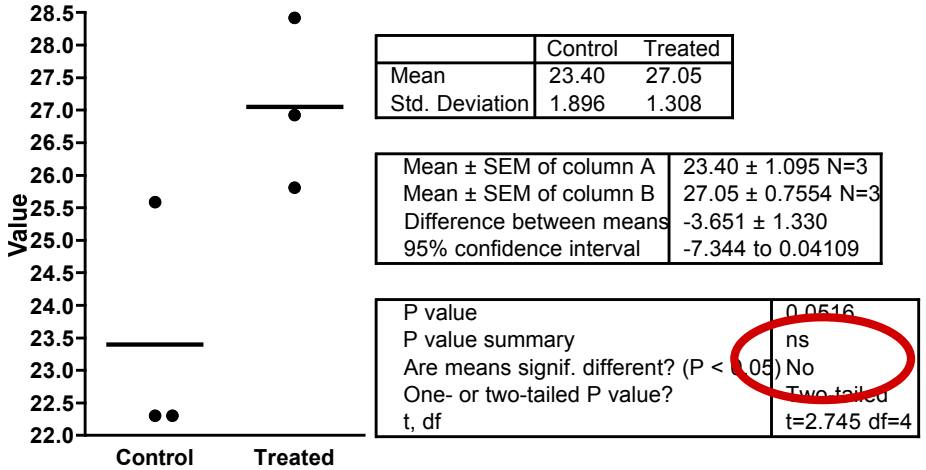
P value	0.0318
P value summary	*
Are means signif. different? (P < 0.05)	Yes
One- or two-tailed P value?	Two-tailed
t, df	t=3.236 df=4



	Control	Treated
Mean	22.35	28.14
Std. Deviation	0.2971	0.5411

Mean ± SEM of column A	22.35 ± 0.1715 N=3
Mean ± SEM of column B	28.14 ± 0.3124 N=3
Difference between means	-5.795 ± 0.3564
95% confidence interval	-6.785 to -4.806

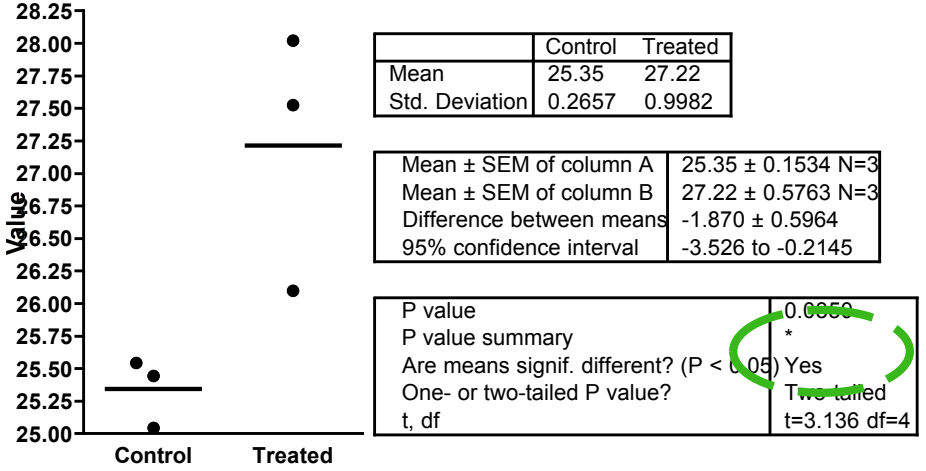
P value	P<0.0001
P value summary	***
Are means signif. different? (P < 0.05)	Yes
One- or two-tailed P value?	Two-tailed
t, df	t=16.26 df=4



	Control	Treated
Mean	23.40	27.05
Std. Deviation	1.896	1.308

Mean ± SEM of column A	23.40 ± 1.095 N=3
Mean ± SEM of column B	27.05 ± 0.7554 N=3
Difference between means	-3.651 ± 1.330
95% confidence interval	-7.344 to 0.04109

P value	0.0516
P value summary	ns
Are means signif. different? (P < 0.05)	No
One- or two-tailed P value?	Two-tailed
t, df	t=2.745 df=4



	Control	Treated
Mean	25.35	27.22
Std. Deviation	0.2657	0.9982

Mean ± SEM of column A	25.35 ± 0.1534 N=3
Mean ± SEM of column B	27.22 ± 0.5763 N=3
Difference between means	-1.870 ± 0.5964
95% confidence interval	-3.526 to -0.2145

P value	0.0350
P value summary	*
Are means signif. different? (P < 0.05)	Yes
One- or two-tailed P value?	Two-tailed
t, df	t=3.136 df=4

Muliple comparisons increase the prevalence of type-I error!

Let's pool the data for placebo and treated:

$$t = \frac{< x >_1 - < x >_2}{s_d} \qquad s_d = \sqrt{\frac{{s_1}^{*2} + {s_2}^{*2}}{n}} \qquad {s^*}^2_k = \frac{\sum_{j=1}^{k} (n_j - 1) s^{*2}}{\sum_{j=1}^{k} (n_j - 1)}$$

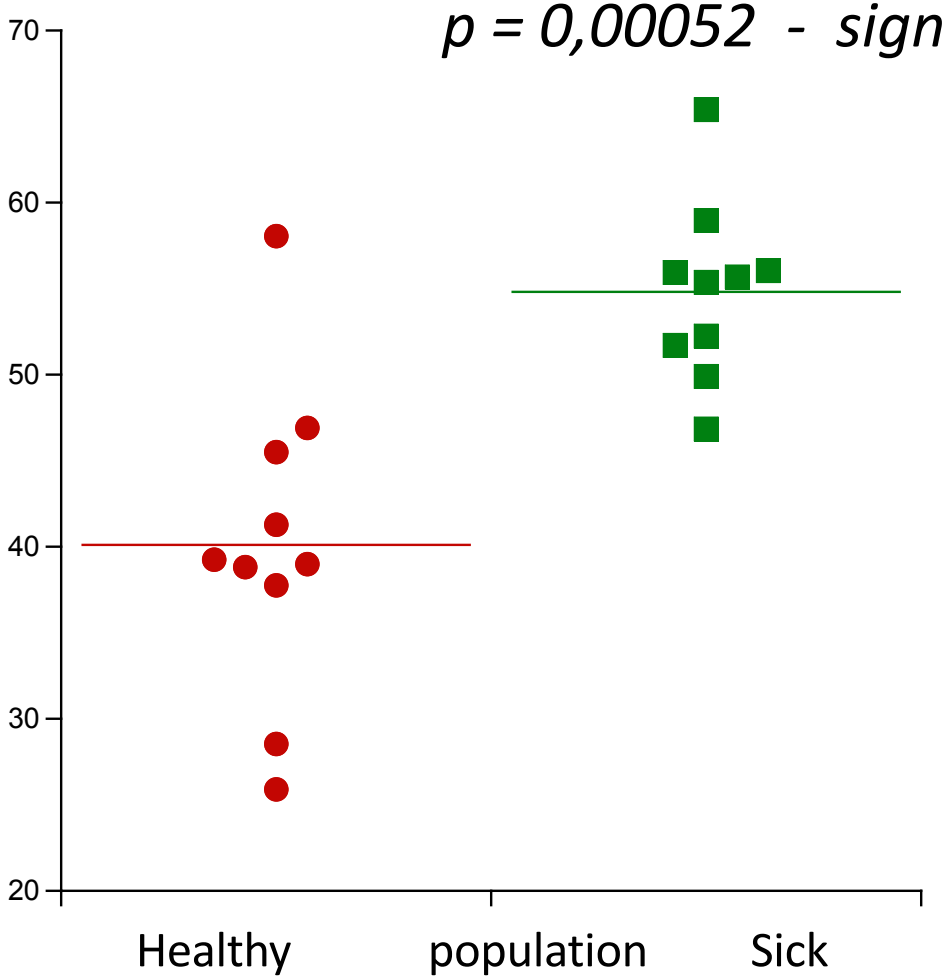
t = (27.569-23.716)/√((1.3451/2+1.5445/2)) = 2.667

Degree of freedom = 24 -8 -2 = 14

t(crit) = 2.15 (95%) |t| > t(crit) = the difference is significant

Potencial biomarkers – diagnostic value !?

p = 0,00052 - significant difference



Diagnostic value: if confidence intervalls do not overlap!

no diagnostic value!

Multiple independent H_0 (comparisons)

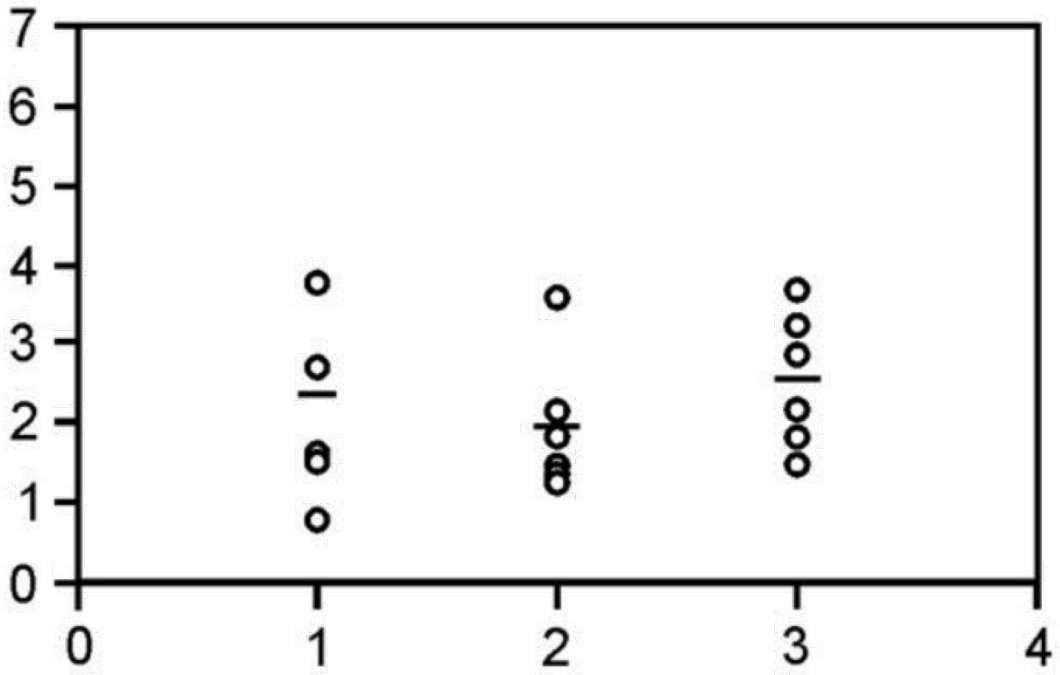
p valódi értéke > 0.05

Number of independent H_0	probabilty of p<0.05	critical “p” for α = 0.05
1	5%	0.0500
2	10%	0.0253
3	14%	0.0170
4	19%	0.0127
5	23%	0.0102
6	26%	0.0085
7	30%	0.0073
8	34%	0.0064
9	37%	0.0057
10	40%	0.0051
20	64%	0.0026
50	92%	0.0010
100	99%	0.0005
n	$100(1-0.95^n)$	$1-0.95^{1/n}$

ANOVA = Analysis of variance

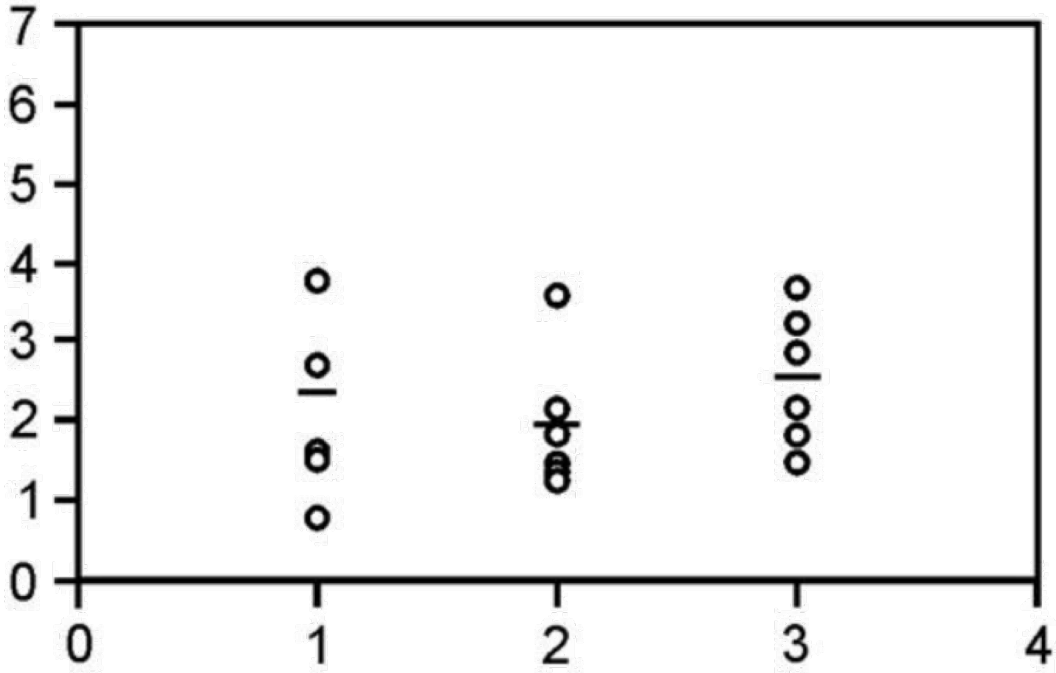
For multiple comparisons regular t test is not proper, because the occurrence of type-I error accmulates = we would reject the true null hypothesis too frequently.

ANOVA is a generalization of t test for multiple comparisons. It compares mean values but utilizing variances of different groupings. The variance of the whole set of data is examined to see that is is solely due to random sampling (statistical fluctuations) or partly caused by real differences among means...



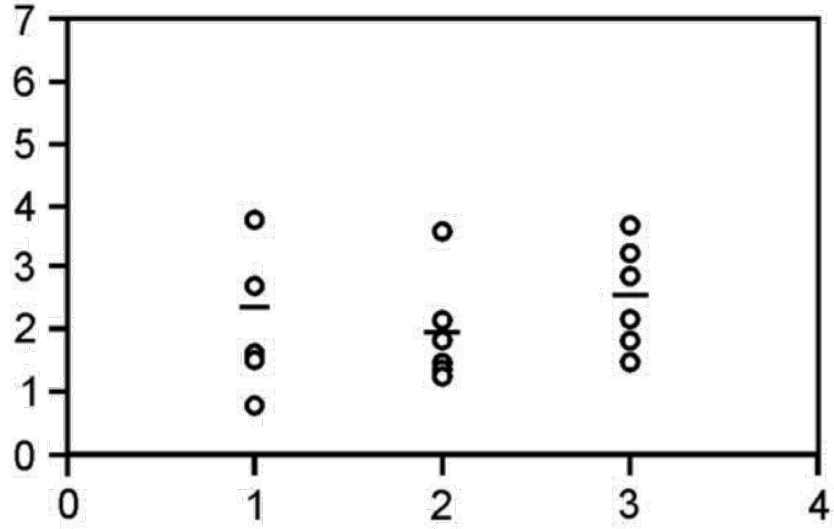
ANOVA = Analysis of variance

Basic idea: if all the samples were from the same distribution, than their grand-average would be the same irrespective of the sequence of calculation.



Let's compare the variances:
– for the whole data set;
– for the individual groups (within groups);
– for the means values (among groups).

ANOVA = Analysis of variance



variance	sum of squares	DF	variances	F test
Among	$SS_k = \sum_{j=1}^{k} n_j (< x_j > - < x >)^2$	k-1	$s_k^2 = SS_k/k-1$	$F = s_k^2/s_b^2$
Within	$SS_b = \sum_{j=1}^{k} \sum_{l=1}^{nk} (x_l - < x_j >)^2$	N-k	$s_b^2 = SS_b/(N-k)$	
Whole	$SS_t = \sum_{j=1}^{k} \sum_{l=1}^{nk} (x_l - < x_j >)^2$	N-1		

ANOVA - which group is different?

If $F = s_k^2/s_b^2 > F_{krit}$ ⬇ reject H_o, there is at least one different group ⬇

How to figure it out without the multiple comparisons? - post hoc tests!

– Dunnett's test: one group is the control and compare all the others to this by Student's t test ⬇ k-1 comparisons only, instead of k(k-1)/2

– Tukey's test: corrected Student's t-test: $(\langle x_1 \rangle - \langle x_2 \rangle) / SE$ $(\langle x_1 \rangle > \langle x_2 \rangle)$
(k*(k-1)/2 comparisons)

– Bonferroni's test: correct the value of α: $\alpha' = \alpha / n$ (n comparisos)

"Effect size" - how to quantitate the effect

- correlation type measures:

  Pearson corelation coefficient = R

  determination coefficient = R^2

  $\eta^2 = SS_c / SS_t$ (overestimates the effect)

  $\omega^2 = (SS_c - DF_c * MS_b) / (SS_t + MS_b)$

SS = sum of squares
DF = degree of freedom
MS = mean square
SD = standard deviation

- difference type measure:

  Cohen: $d = \langle x_1 \rangle - \langle x_2 \rangle / sd$

  $Z' = 1 - 3*(SD_p + SD_n) / (\langle x \rangle_p - \langle x \rangle_n)$

  ("p" is a positive control, "n"is a negative control)

  if $Z' \approx 1$ ⬇ pefect assay

  if $0,5 < Z' < 1$ ⬇ great assay

  if $0 < Z' < 0,5$ ⬇ poor assay

  if $Z' < 0$ ⬇ useless assay

Final remark…

Herbert George Wells, writer:

Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.

(he was great in science fiction ☺)