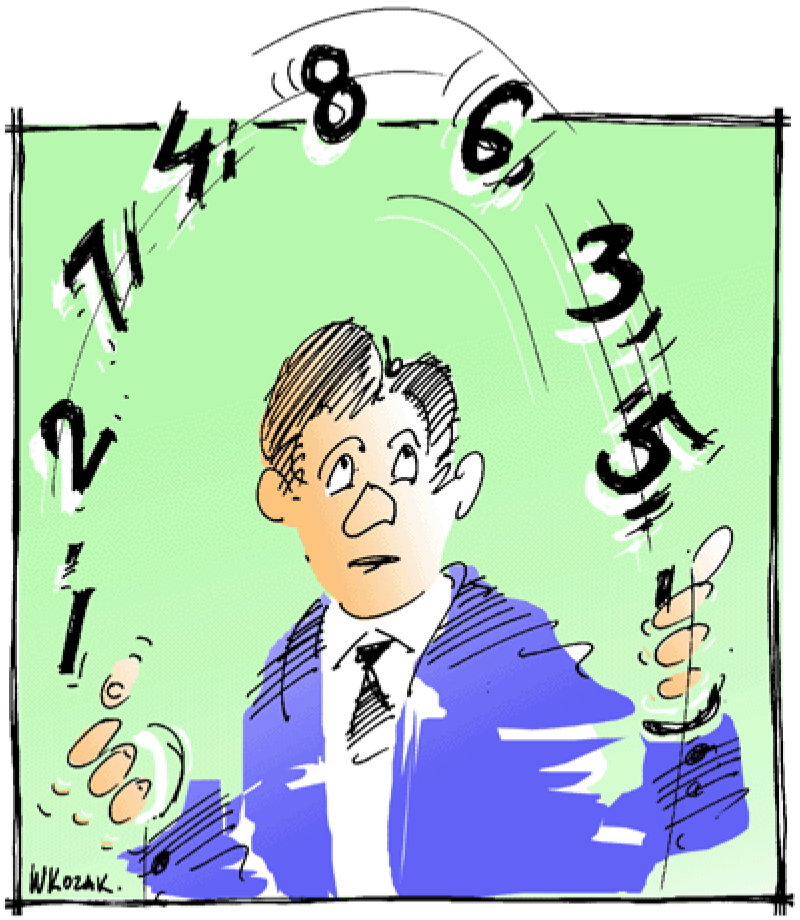
Principles of statistical testing

(1) simple lie
(2) treacherous lie
(3) Statistics

Benjamin Disraeli

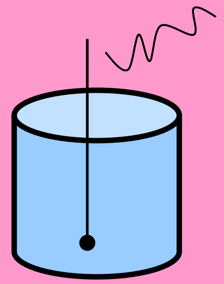


What is statistics?

- the way data are collected, organised, presented, analysed and interpreted
- statistics **helps to decide**
  - **descriptive**
    - basic characteristics of the data
  - **inductive**
    - characterisation of the sample or population studied, which make possible to interfere characteristics of the whole population (entire “sample”)

Why do we need statistics? → variability!

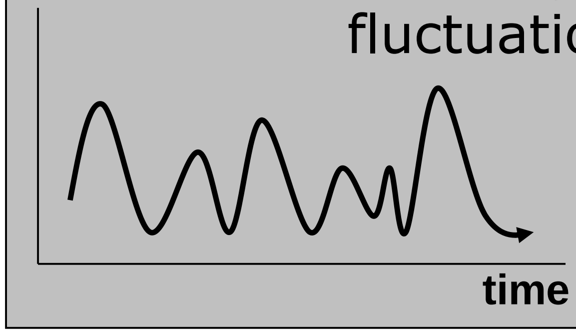
repeated measurements
of temperature



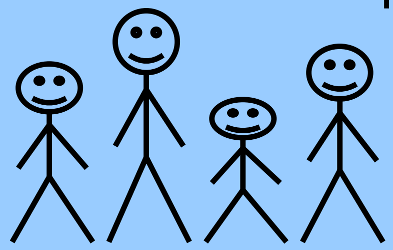
18.2°C
18.5°C
19.1°C
18.7°C

diversity in biological
populations
inter-population or ethnical
differences
= **BIODIVERZITY**

temporal changes/
fluctuations



variability of height
in population



180 cm
175 cm
165 cm
157 cm

statistics is about variability !!!

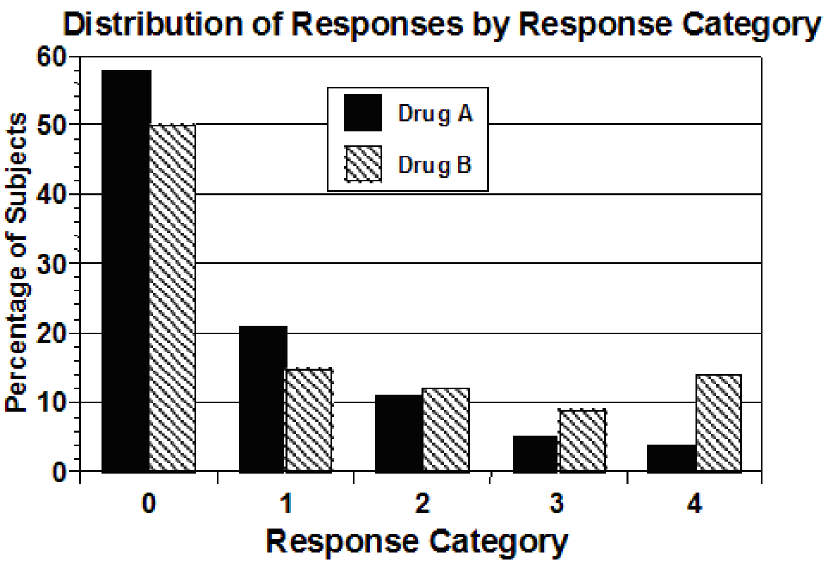
Type of data

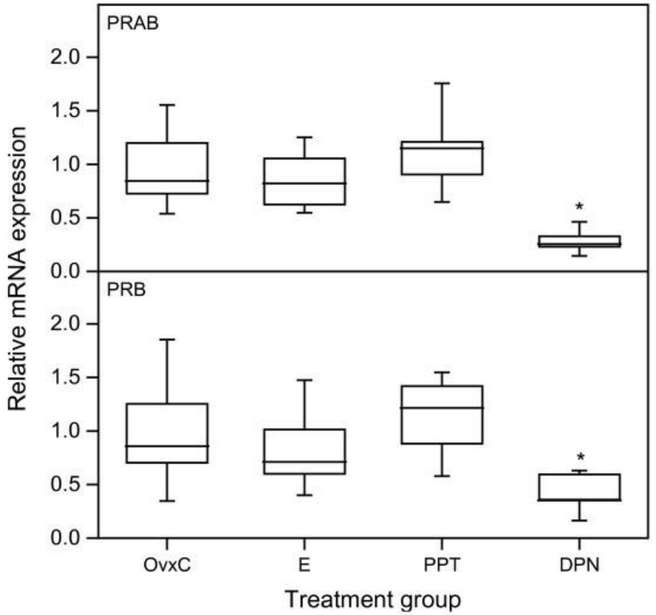
- data, measures
  - qualitative = descriptive
    - nominal, binary
      - ☛e.g. blood groups A, B, 0, AB or Rh^+, Rh^-
    - ordinal, categorical
      - ☛e.g. grades NYHA I, II, III, IV or TNM system (cancer)
  - quantitative = measurable on scale
    - directly measured values
    - interval (how much more?)
    - ratios (how many times?)

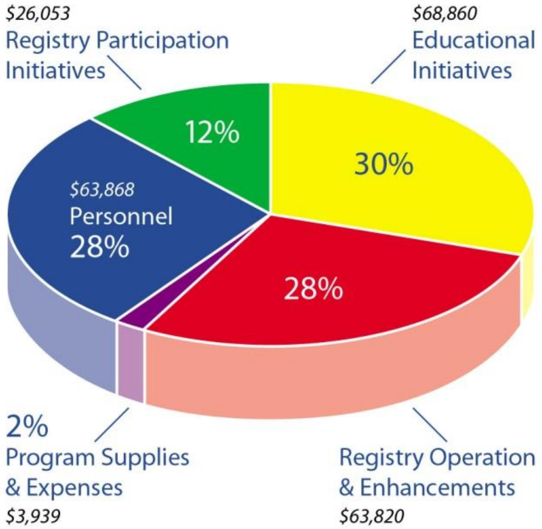
Raw data – not too clear

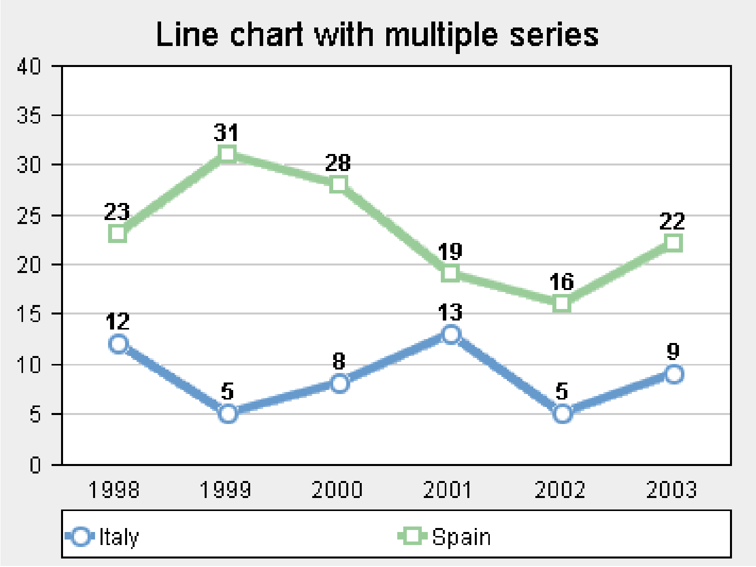
DNA	DN_kod	UREA	KREATININ	glom_filt	sRAGE
HER0087	3	7.6	97	1.172	9660.3
HER0037	3	7.6	139	0.574	5843
HER0009	3	6	118	1.502	5753.5
HER0012	3	17.3	274	0.442	5400
HER0118	3	22.6	156	0.463	5386.7
HER0094	3	10.8	234	0.812	5312.4
HER0144	3				5200
KRUS002	3	25.9	309	0.393	4947.8
HER0006	3	7.5	118	1.028	4944.5
HER0007	3	4.7	84	0.764	4917.8
HER0122	3	28.4	295	0.308	4627.1
HER0128	3	7.2	123	1.048	4503.5
KRUS50	3	37.8	525	0.284	4446
HER0035	3	7.1	111	0.739	4404
HER0001	3	14.2	188	0.557	4395.1
HER0057	3	21.8	281	0.703	4389.2
HER0015	3	7.2	75	2.703	4263.3
HER0111	3	13.7	131	0.954	4188.9
KRUS042	3	4.4	104	0.983	4127
HER0047	3	26	333	0.244	4101.9
HER0062	3	22.8	169	0.42	3852.7
HER0002	3	6.9	135	0.999	3815.3
HER0115	3	18.3	152	0.396	3741.2
KRUS045	3	4.4	85	1.7	3693.3
KRUS001	3	20.5	178	0.861	3621.5
M__0136	2				3606.9
HER0086	3	24.7	300	0.237	3577.7
HER0132	3	13	154	0.608	3409.8
HER0010	3	6.4	64	1.4	3398
HER0032	3	7.3	73	1.839	3325.5
HER0005	3	3.9	89	2.074	3318.7
KRUS016	2	6	105	2.38	3243.2
HER0071	3	7.3	120	0.769	3234.5
KRUS009	3	10.8	188	0.89	3212.6
M__0164	1	7.3	59		3203.9
OLS0008	2				3203.9
HER0061	3	18.2	241	0.277	3080.6
HER0065	3	7.2	116	0.953	3072.3
HER0058	3	16.8	158	0.668	3066
HER0014	3	14.6	187	0.0765	3047.4

Graphical data description

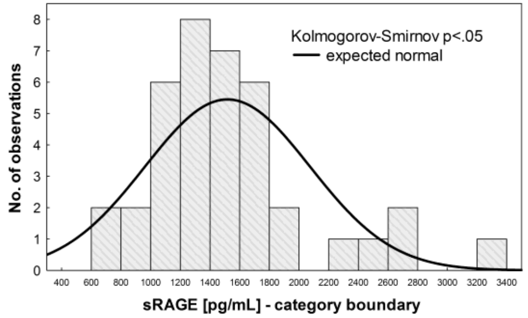


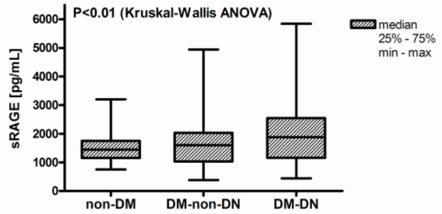


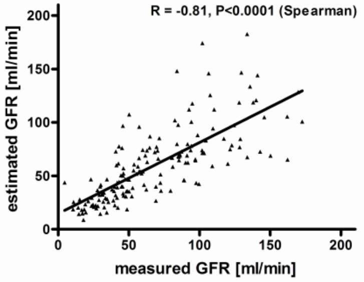


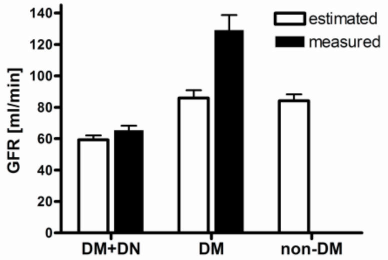


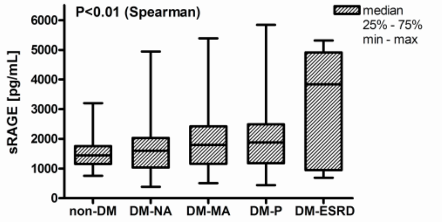
Examples of real data

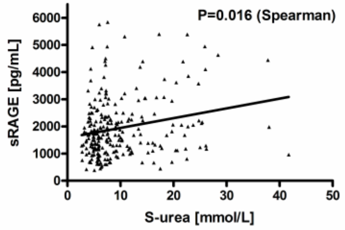


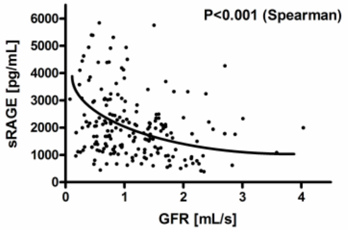


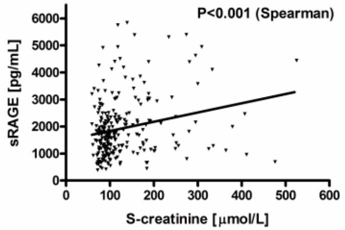


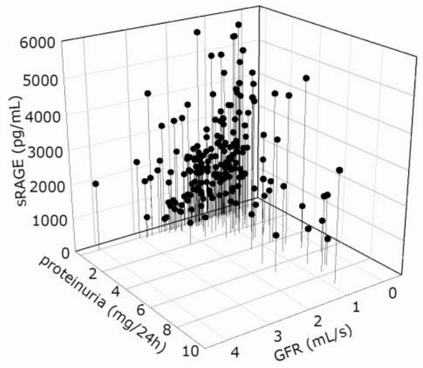


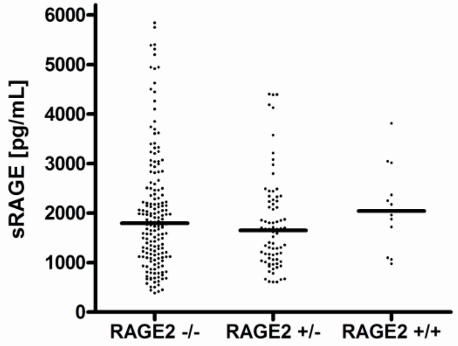


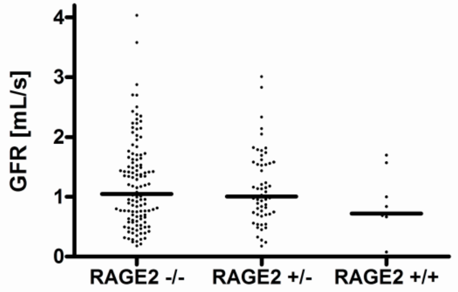








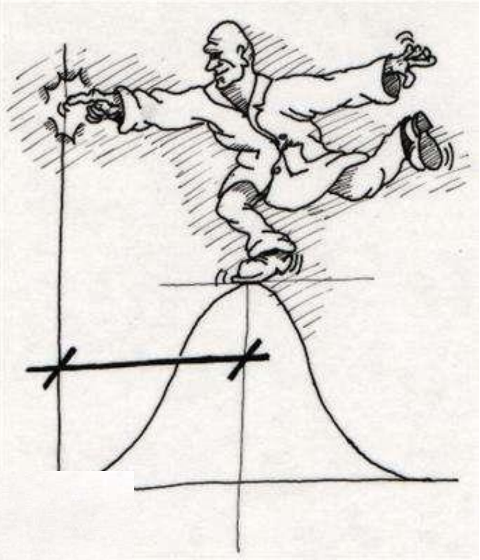


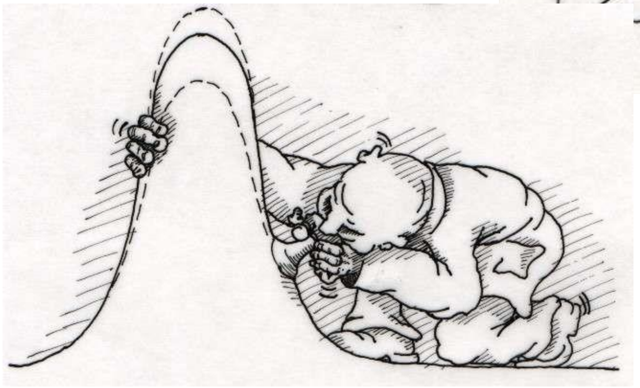



This pie chart shows how much pie I ate while making this chart.
BIZARRO.COM
©DAN PIRARO 12-1-04

Data description

- **position measures (central tendency measures)**
  - mean (μ)
  - median (= 50% quintile)
    - frequency middle
  - quartiles
    - upper 25%, median, lower
  - mode
    - the most frequent value
- **variability measures**
  - variance (σ^2)
  - standard deviation (SD, σ)
  - standard error of mean (SEM)
  - coefficient of variance (CV= σ/μ)
  - min-max (= range)
  - skewness
  - kurtosis
- **distribution**





Data description

- frequency (polygon, histogram)

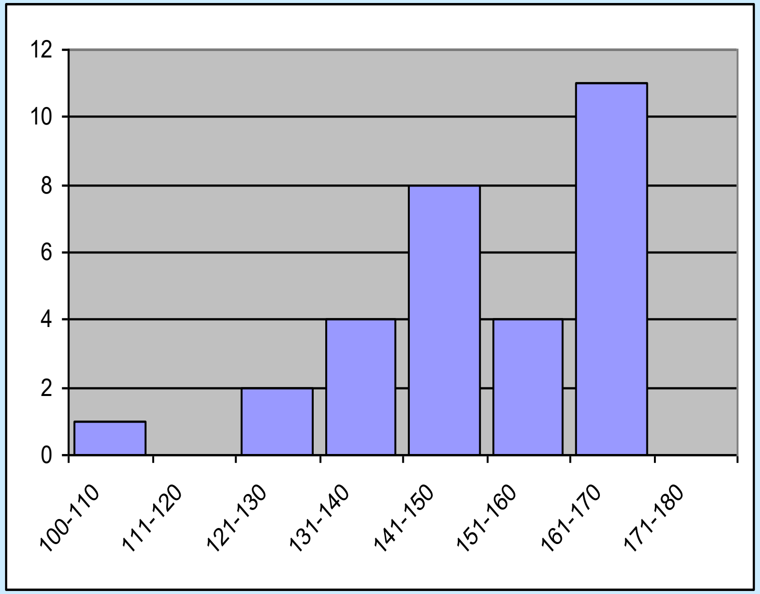
original data

115
135
120
140
125
130
150
145
.
.
.

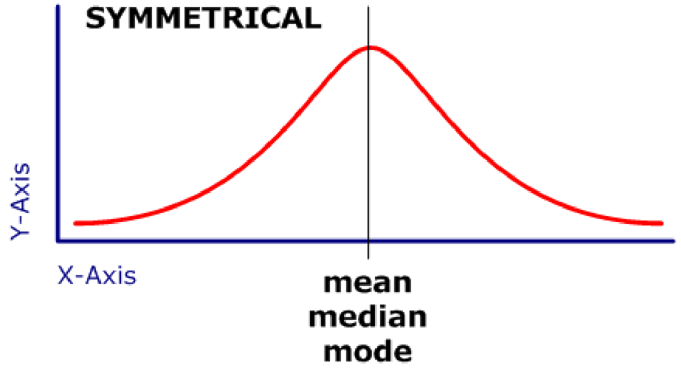
ordered data

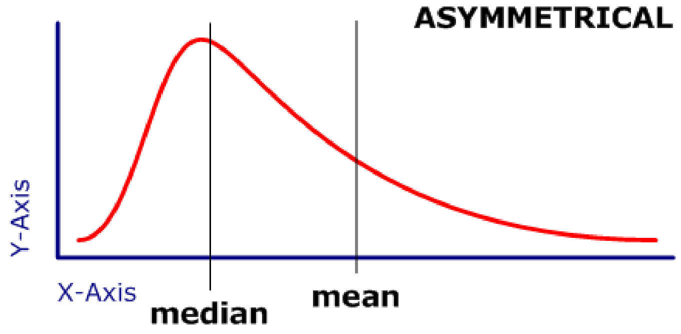
<100: 0
100-110: 1
111-120: 0
121-130: 2
131-140: 4
141-150: 8
151-160: 4
161-170: 11
>171: 0

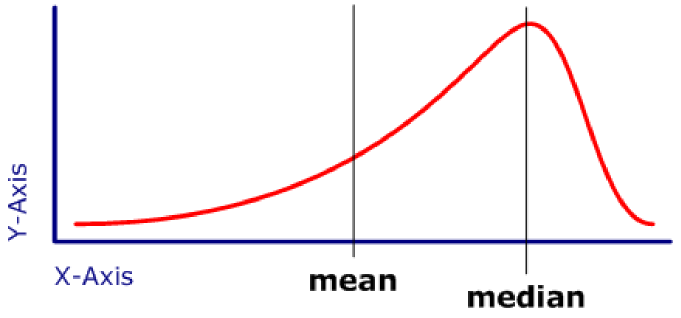
histogram



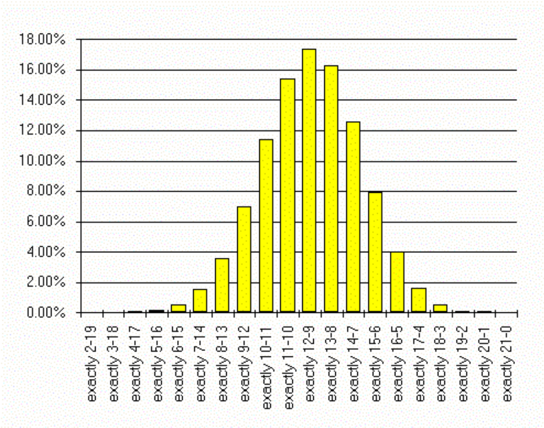
Distribution



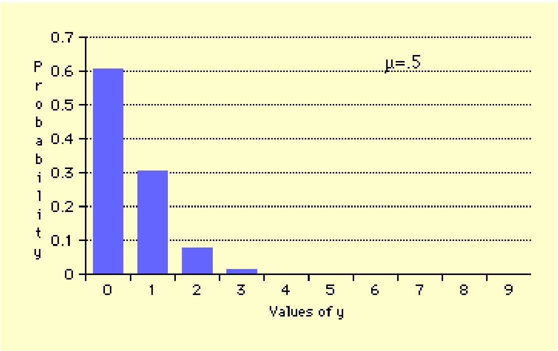




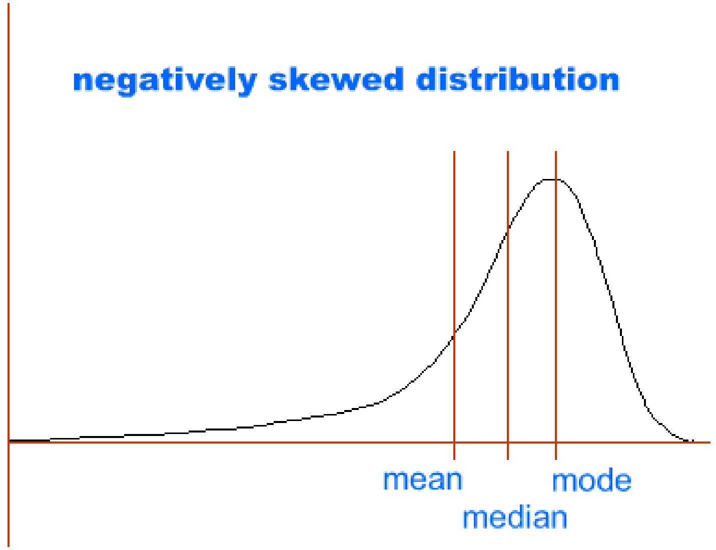
- continuous
  - normal
  - asymmetrical
  - exponential
  - log-normal

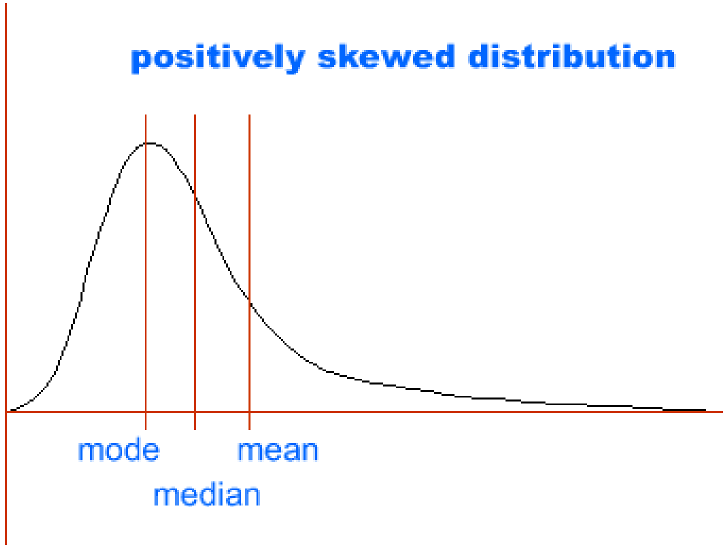


- discrete
  - binomial
  - Poisson



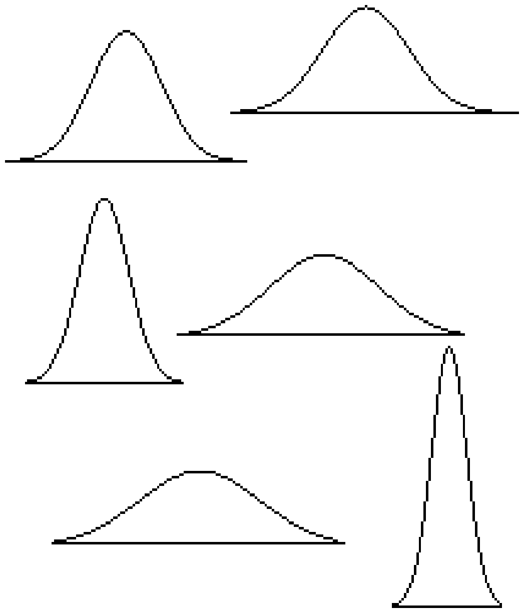
Mean vs. median vs. mode(s)





- numbers:13, 18, 13, 14, 13, 16, 14, 21, 13
  - **x** = (13 + 18 + 13 + 14 + 13 + 16 + 14 + 21 + 13) ÷ 9 = **15**
  - **median** = (9 + 1) ÷ 2 = 10 ÷ 2 = 5. číslo = **14**
  - **mode** = **13**
  - **range =** 21 – 13 = **8**

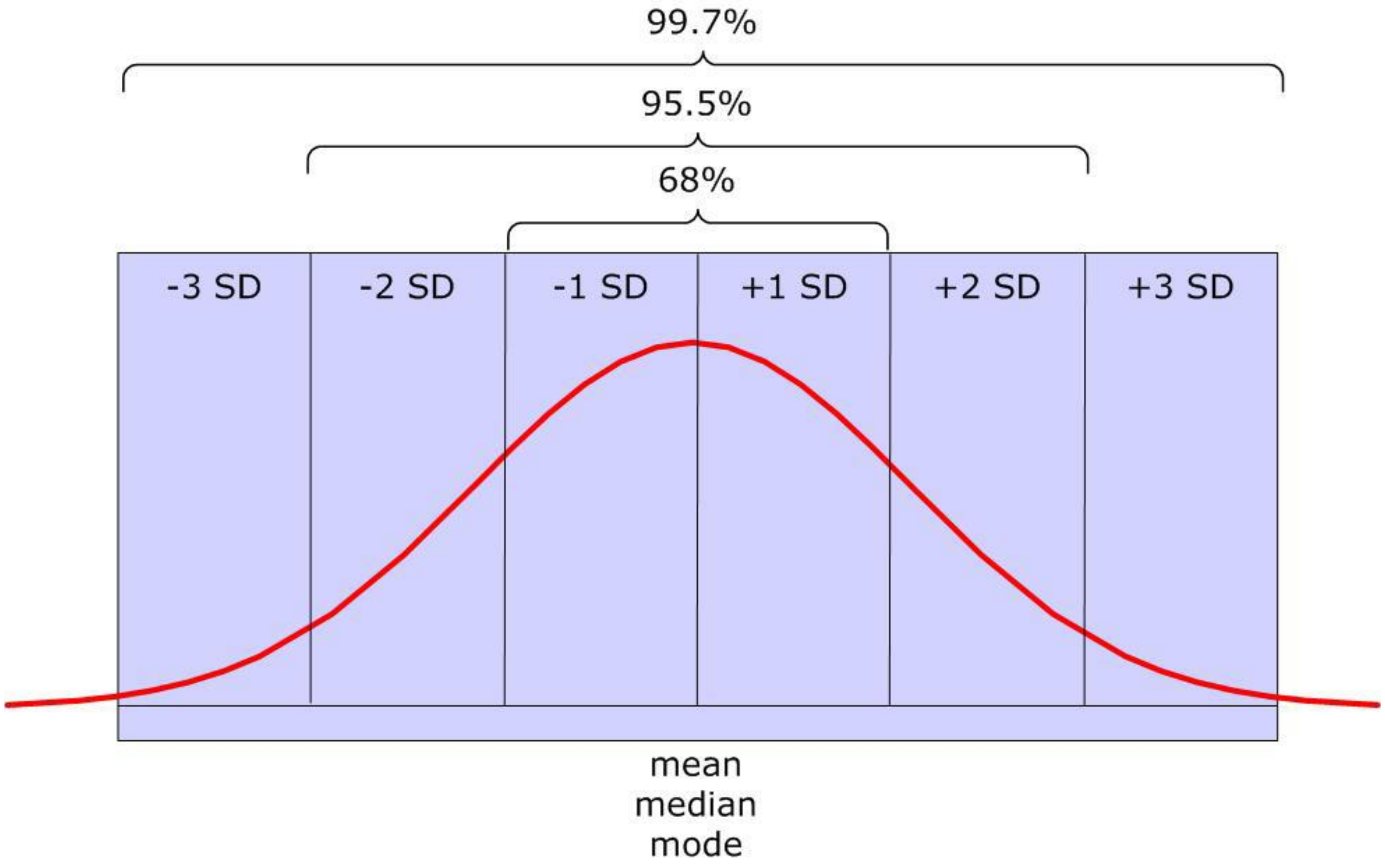
Normal (Gaussian) × Student × symmetrical distribution



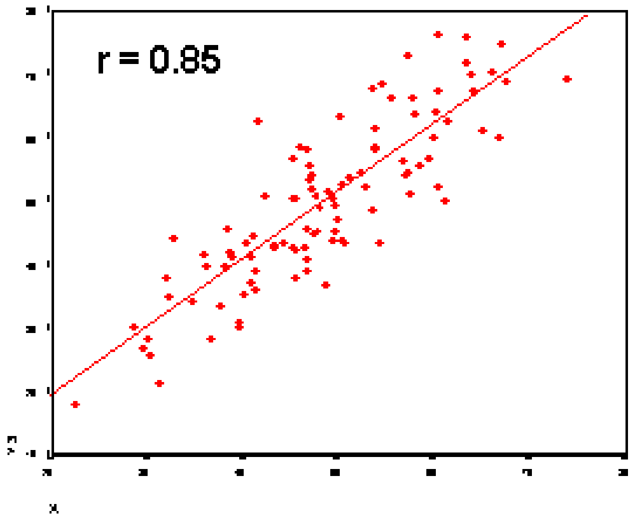
$$f(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(z)^2}$$

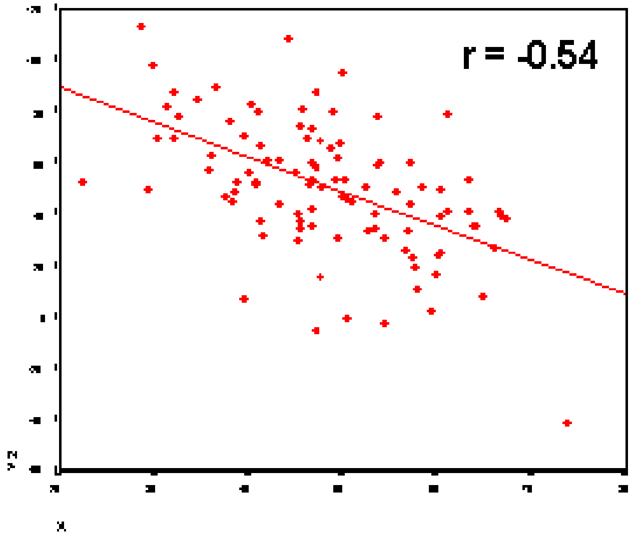
- not every symmetrical distribution has to be normal !!
  - there are several conditions that have to be fulfilled
    - interval density of frequencies
    - distribution function
    - skewness = 0, kurtosis = 0
  - data transformation
    - mathematical operation that makes original data normally distributed
- Student distribution is an approximation of the normal distribution for smaller sets of data
- test of normality
  - Kolmogorov-Smirnov
  - Shapiro-Wilks
    - null hypothesis: distribution tested is not different from the normal one

Normal distribution

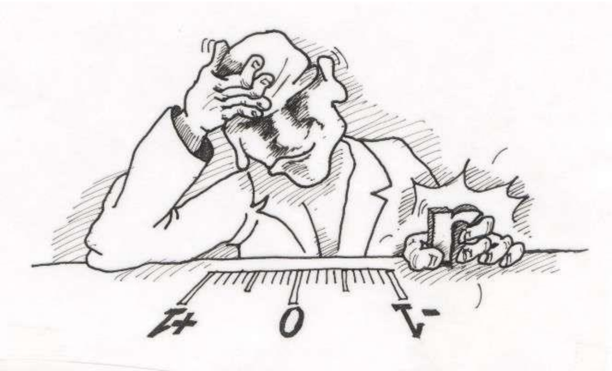


Relationship between variables



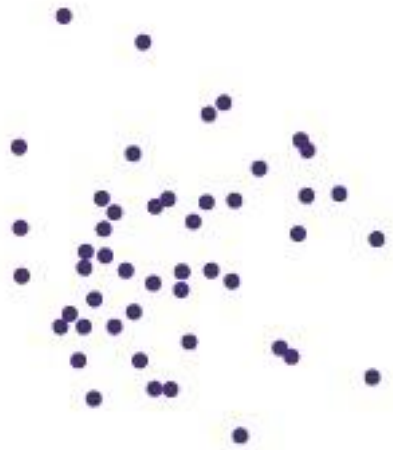


- **Correlation** = relationship (dependence) between the two variables
  - correlation coefficient = degree of (linear) dependence of the two variables X and Y
    - Pearson (parametric)
    - Spearman (non-parametric)

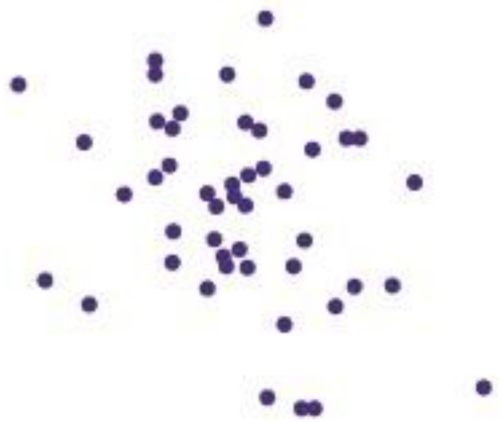


- **Regression** = functional relationship between variables (i.e. equation)
  - one- or multidimensional
  - linear vs. logistic
  - interpretation: assessment of the value (or probability) of one parameter (event) when knowing the value of the other one

Examples



Correlation $r = 0$

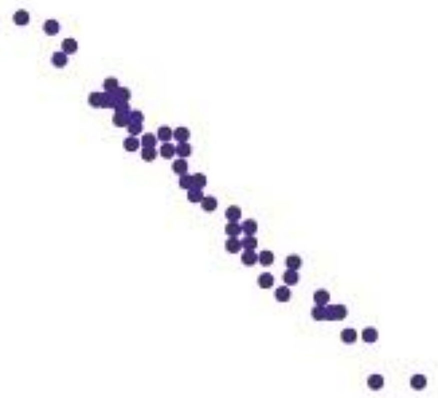


Correlation $r = -0.3$

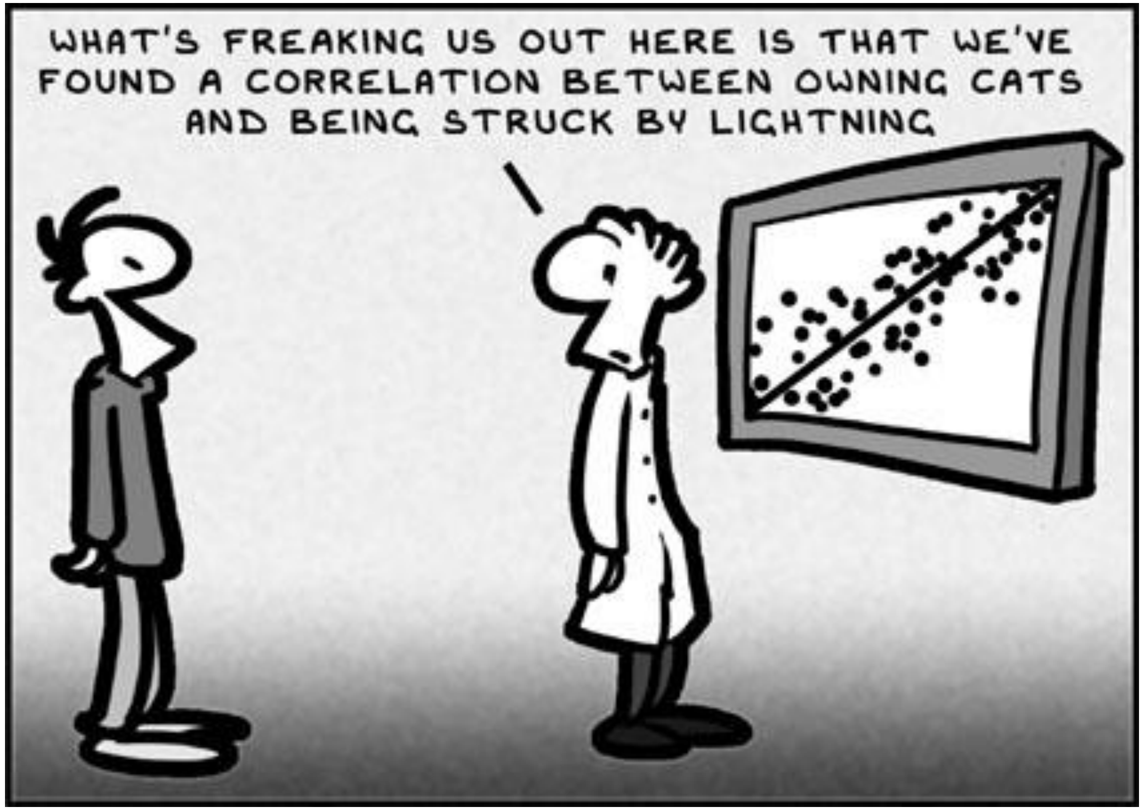
Correlation $r = 0.5$

Correlation $r = -0.7$

Correlation $r = 0.9$



Correlation $r = -0.99$


WHAT'S FREAKING US OUT HERE IS THAT WE'VE FOUND A CORRELATION BETWEEN OWNING CATS AND BEING STRUCK BY LIGHTNING

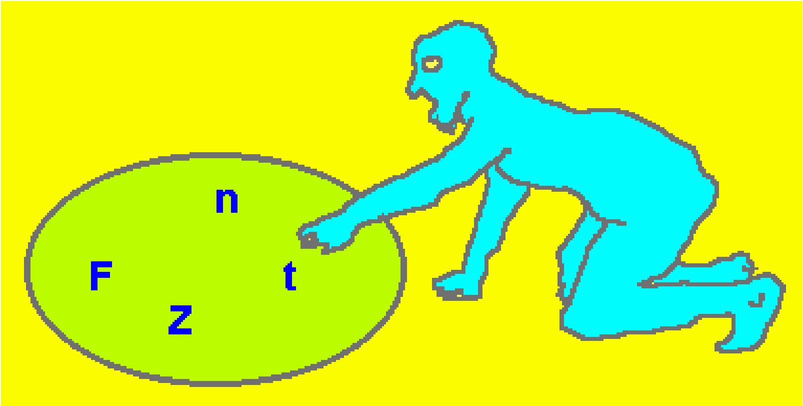
Principles of statistical thinking

- inferences about the **whole population** (sample) based on the results obtained from the **limited study sample**
  - whole population (sample)
    - e.g. entire living human population
    - we want to know facts applying to this whole population and use them (e.g. in medicine)
  - selection
    - no way we van study every single member of the whole population or sample
    - we have to select “represen-tative” sub-set which will serve to obtain results valid for the whole population
  - random sample
    - every subject has an equal chance to be selected

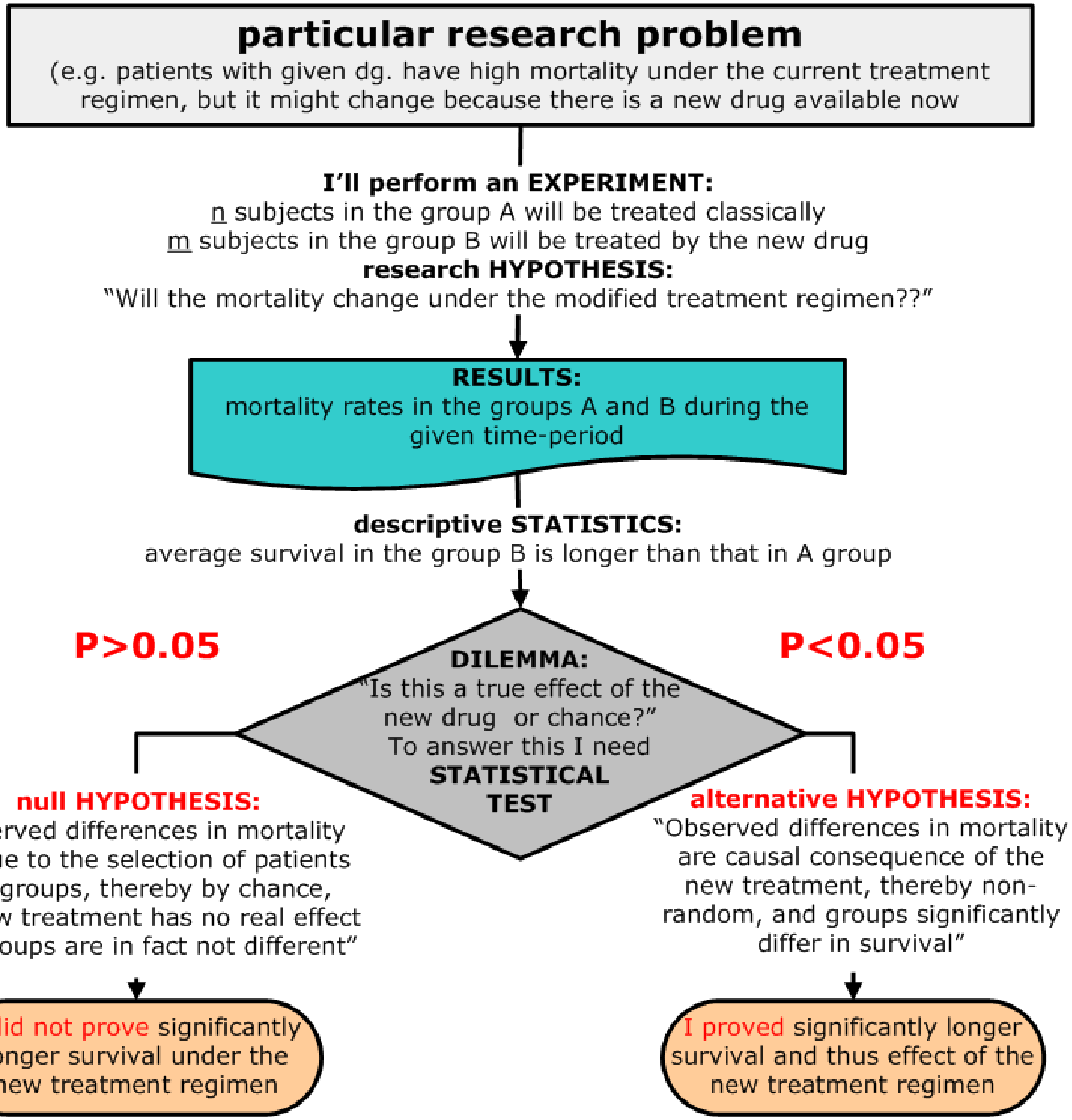

BUT THEY WERE SELECTED RANDOMLY

Statistical hypothesis

- our personal **research hypothesis**
  - e.g. "*We think that due to the effects of the newly described drug (…) on blood pressure lowering our proposed treatment regimen – tested in this study – will offer better hypertension therapy compared to the current one*".
- **statistical hypothesis** = mathematical formulation of our research hypothesis
  - the question of interest is simplified into two competing claims / hypotheses between which we have a choice
    - null hypothesis (H_0): e.g. there is no difference on average in the effect of an "old" and "new" drug
      - ☛ $\mu_1 = \mu_2$ (equality of means)
      - ☛ $\sigma_1 = \sigma_2$ (equality of variance)
    - alternative hypothesis (H_1): there is a difference
      - ☛ $\mu_1 \neq \mu_2$ (inequality of means)
      - ☛ $\sigma_1 \neq \sigma_2$ (inequality of variance)
- the **outcome of a hypothesis testing** is:
  - "reject H_0 in favour of H_1"
  - "do not reject H_0"



Hypothesis testing




HIRAM S. DUDSON
1930 - 1993
Member,
Placebo Group
D. Reilly

Statistical errors

- to perform hypothesis testing there is a large number of statistical tests, each of which is suitable for the particular problem
  - selection of proper test (respecting its limitation of use) is crucial!!!
- when deciding about which hypothesis to accept there are 2 types of errors one can make:
  - type 1 error
    - α = probability of incorrect rejection of valid H_0
    - **statistical significance P = true value of α**
  - type 2 error
    - β = probability of not being able to reject false H_0
    - **1 – β = power of the test**

	True state of the null hypothesis	
Statistical decision made	H_0 true	H_0 false
Reject H_0	**type I error**	correct
Don’t reject H_0	correct	**type II error**

Statistical significance

- In normal **English**, "significant" means important, while in **statistics** "significant" means probably true (= not due to the chance)
  - however, research findings may be true without being important
    - when statisticians say a result is "highly significant" they mean it is very probably true, they do not (necessarily) mean it is highly important

- **Significance levels show you how likely a result is due to chance**

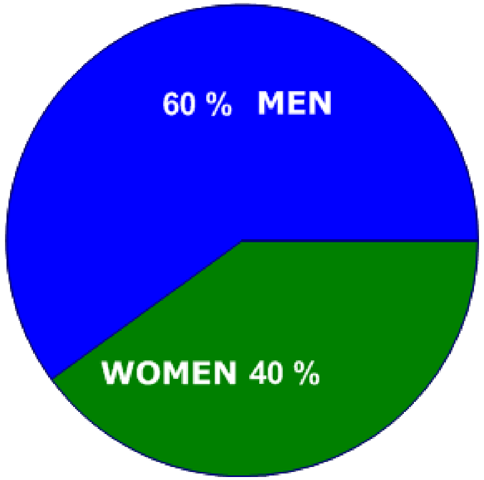
Statistical tests for quantitative (continuous) data, 2 samples

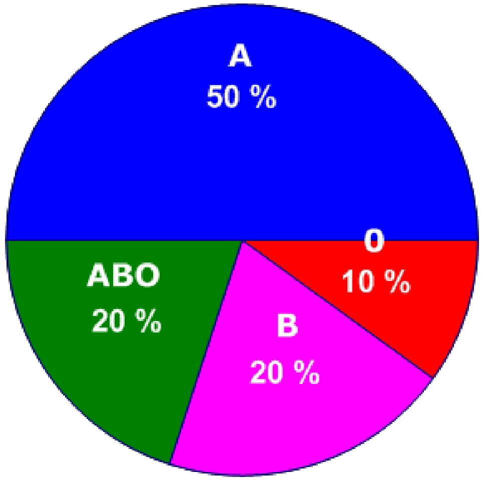
test	unpaired	paired
PARAMETRIC (for normally or near normally distributed data)	1. two-sample t-test	1. one-sample t-test dependent
NON-PARAMETRIC (for other than normal distribution)	1. Mann-Whitney U-test (synonym Wilcoxon two-sample)	1. Wilcoxon one-sample 2. sign test
	comparison of parametrs between 2 independent groups (e.g. cases × controls)	comparison of parametrs in the same group in time sequence (e.g. before × after treatment)

Statistical tests for quantitative (cont.) data, multiple samples

test	unpaired	paired
PARAMETRIC (normal distribution, equal variances)	1. Analysis of variance (ANOVA)	1. modification of ANOVA
NON-PARAMETRIC (other than normal distribution)	1. Kruskal-Wallis test 2. median test	1. modification of ANOVA (Friedman sequential ANOVA)
	H_0: all of <u>n</u> compared samples have equal distribution of variable tested	

Statistical tests for binary and categorical data





- binary variable
  - 1/0, yes/no, black/white, …
- categorical variable
  - category (from – to) I, II, III
- contingency tables <u>n</u> × <u>n</u> or <u>n</u> × <u>m.</u> resp.
  - Fisher exact testy
  - chi-square

	diseased	healthy
mutation	50	2
no	4	48

WELCOME TO ... STATISTICS
POPULATION: 387
SUZY, I HATE TO BREAK THE NEWS TO YOU LIKE THIS ... BUT THERE'S A GOOD REASON WHY YOUR NEW BOYFRIEND DON'T FIT IN AMONGST THE TOWNSFOLK TOO WELL ....
YOU MEAN ...?
YES SUZY. HE'S A NOTORIOUS OUTLIER.

Thank you for your attention