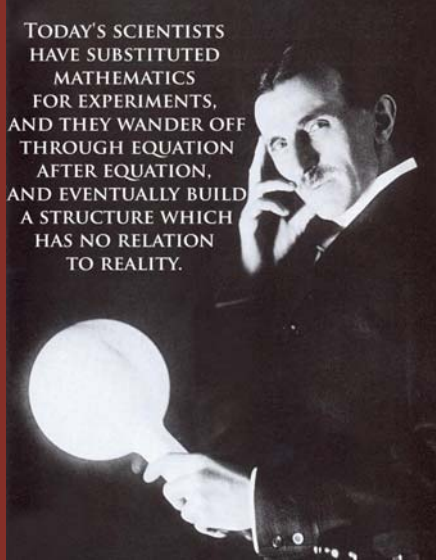


## Statistics in research

TODAY'S SCIENTISTS  
HAVE SUBSTITUTED  
MATHEMATICS  
FOR EXPERIMENTS,  
AND THEY WANDER OFF  
THROUGH EQUATION  
AFTER EQUATION,  
AND EVENTUALLY BUILD  
A STRUCTURE WHICH  
HAS NO RELATION  
TO REALITY.



## Statistics in research



**If your experiment needs statistics,  
you ought to have done a better experiment**

*Ernest Rutherford*

## Why statistics in research?

Gives credit to your claim (**statistically significant** probability)

To identify the connections

(organisms with environmental conditions and each other, food sources to animals, experimental treatments to subjects e.g. doses to enzyme activity...)

To identify the differences

(locations, habitats, communities, treatment effects...)

**REQUIREMENTS:** parameters must be measurable - data must be quantified

## Identifying the causality of the observed events

**Indirectly - establishing correlations**  
(correlation coefficients)

**Perfect correlation = 1**

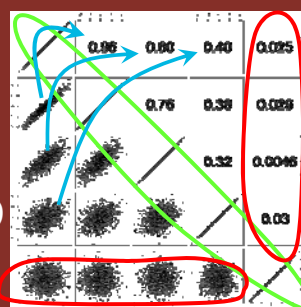
(change of a parameter is paired with change of another in same proportion)

**No correlation = 0**

(change of a parameter is NOT paired with a consistent change of another)

**Negative correlation**

(increase of a parameter is paired with decrease of another)



**Most causalities will also yield correlations. NOT VICE VERSA!**

**Correlation coefficients - correlation isn't causality**

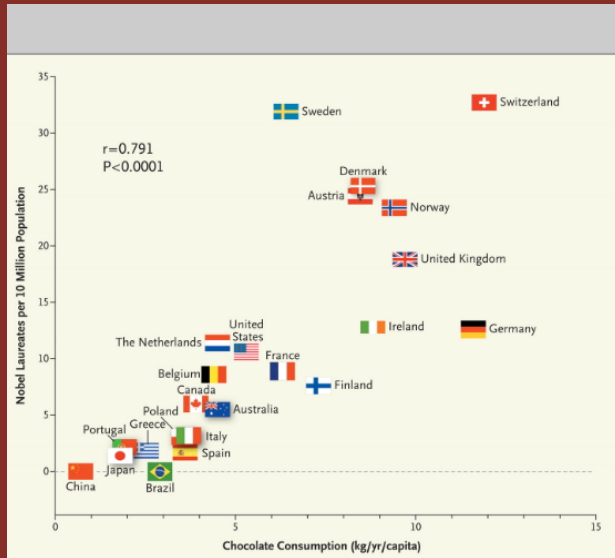


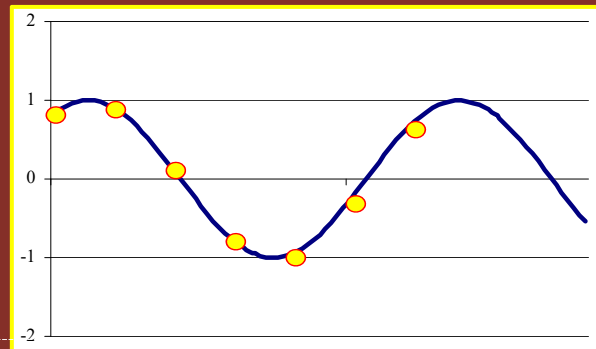
Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

**There are three kinds of lies: lies, damned lies and statistics**

*Benjamin Disraeli*

**Correlation coefficients - no correlation isn't no causality**

X	Y
1	0.84
2	0.91
3	0.14
4	-0.76
5	-0.96
6	-0.28
7	0.66



$r = 0.06$

$R_s = 0.04$

### Identifying the difference between the observed events

By comparing dispersal of data around a mean

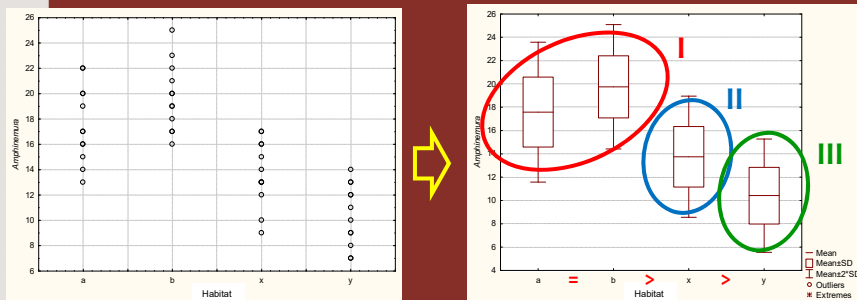
n	X	X-Xs	n	X	X-Xs
1	29	-6	1	1	-34
2	30	-5	2	10	-25
3	33	-2	3	15	-20
4	35	0	4	35	0
5	38	3	5	55	20
6	39	4	6	60	25
7	41	6	7	69	34
Σ	245		Σ	245	
$\bar{X}$	35		$\bar{X}$	35	

Variance  $\rightarrow s^2 = \frac{\Sigma (X - Xs)^2}{n - 1}$  = 21      = 727

Standard deviation  $\rightarrow \sigma = \sqrt{s^2}$  = 4.58      = 26.96

### Identifying the difference between the observed events

By comparing dispersal of data around a mean



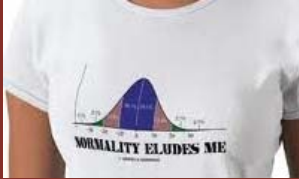
Variance:  $s^2 = \frac{\Sigma (X - Xs)^2}{n - 1}$

	<p><b>Categorization of data:</b></p> <p><b>1. Qualitative (categorical) and 2. Quantitative</b></p> <p><b>1. a) Nominal</b>            Cannot be ordered (not noted with numbers)            Descriptive            Gender: M/F            Location: A/B/C... or 1/2/3... E.g.?            Habitat: benthos / moss mat...            Treatment: control / treatment 1 / treatment 2...</p> <p><b>1. b) Ordinal</b>            Can be ordered but with vague boundaries            Upstream, downstream;            Fast flow, slow flow E.g.?            Grades ... (2,3,4,5)</p>


	<p><b>Categorization of data:</b></p> <p><b>2. Quantitative</b></p> <p><b>2. a) Discontinuous (discrete)</b>            integer values            number of individuals, students, photons... E.g.?</p> <p><b>2. b) Continuous</b>            any value (within a range)            temperature: <math>-273.15-10^{17}</math> °C or <math>0-10^{17}</math> K            pH: 0-14</p>

**Distribution of data:**

**Normal**  
Most values near a median, few extremes



**Non-normal**



Methods for achieving normality  
Log (x+1)  
or  
 $X^3, 2, 1/2, 1/3...$

Normality tests: Shapiro-Wilk's, Kolmogorov-Smirnov...

**Classification of commonly used methods in data processing:**

**Parametric:**  
Assume normal distribution

**Nonparametric:**  
Do not assume normal distribution  
small data sets (around 10)

Measured value	Rank
4.361	2
9.553	8
4.588	3
5.451	4
6.855	6
5.807	5
0.983	1
17.736	9
8.239	7

### Common methods in the processing of environmental data:

#### Parametric:

Pearson's correlation coefficient  
Analysis of variance ANOVA; (t-test)

#### Nonparametric:

Spearman correlation coefficient  
Kruskal-Wallis ANOVA; Mann-Whitney U-test

### Common methods in the processing of environmental data:

Task	Number of data per group (category)	Analysis type that should be used	Number of data groups (categories)	Analysis that should be used
Detect differences	<10	Nonparametric	3 or more	Kruskal-Wallis
	>10	Parametric*	2	Man-Whitney
Detect relationships	<10	Nonparametric	-	ANOVA
	>10	Parametric*	-	Spearman R
				Pearson r (basic)

\* Potential need for normalization (e.g.  $\log(x + 1)$ )

### Null hypothesis ( $H_0$ ):

The assumption that the analysis is testing:

For:

Correlations  $\rightarrow H_0 =$  there is NO correlation

Analyses of variance  $\rightarrow H_0 =$  data groups are NOT different

Tests of normality  $\rightarrow H_0 =$  distribution IS normal

### Alternative hypothesis ( $H_a$ ):

The assumption opposite to  $H_0$

Generally - an assumption that the researcher (you) believes is true!

Apart from the results of analyses algorithms, a **level of probability that  $H_0$  is true**  $\rightarrow$  **value 'p'** is always reported

### Statistical significance ( $p$ ):

Measured data or their relationships are not coincidental;  
They are not a result of chance



Repetition of the result/measurement can be expected  
under same circumstances



**The result is TRUE**

$$p < 0.05$$

For ANOVA,  $p = 0.01$  means that the probability that  $H_0$  is true (no differences among datasets) is 1 %;  
Therefore we can be **99 % certain that  $H_a$  is true** and that there are significant differences among datasets.



## Errors

Type I (rejection of a correct  $H_0$  - false positive)

Type II (acceptance of a false  $H_0$  - false negative)

### Genesis 18

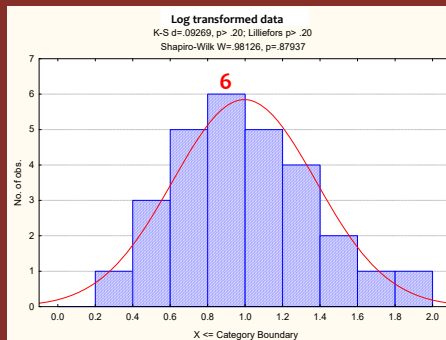
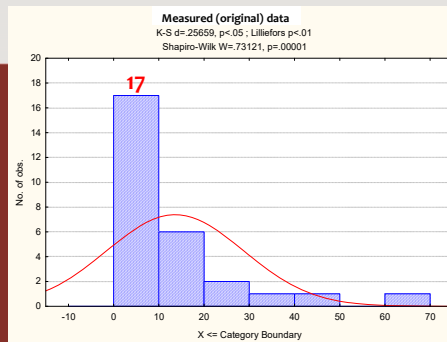
23: And Abraham drew near, and said, Wilt thou also destroy the righteous with the wicked?

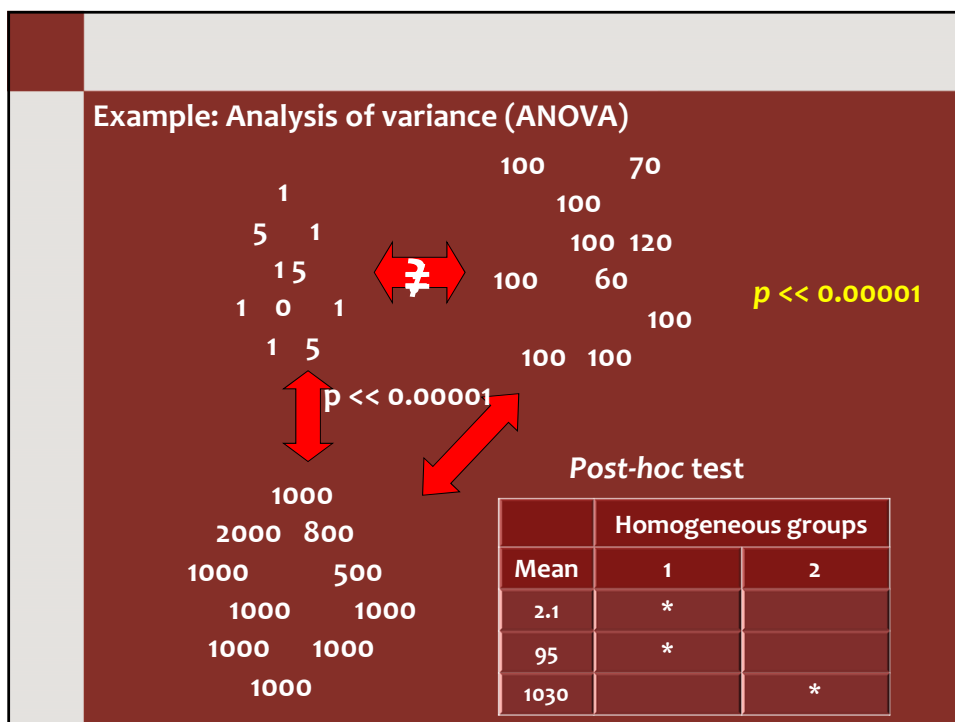
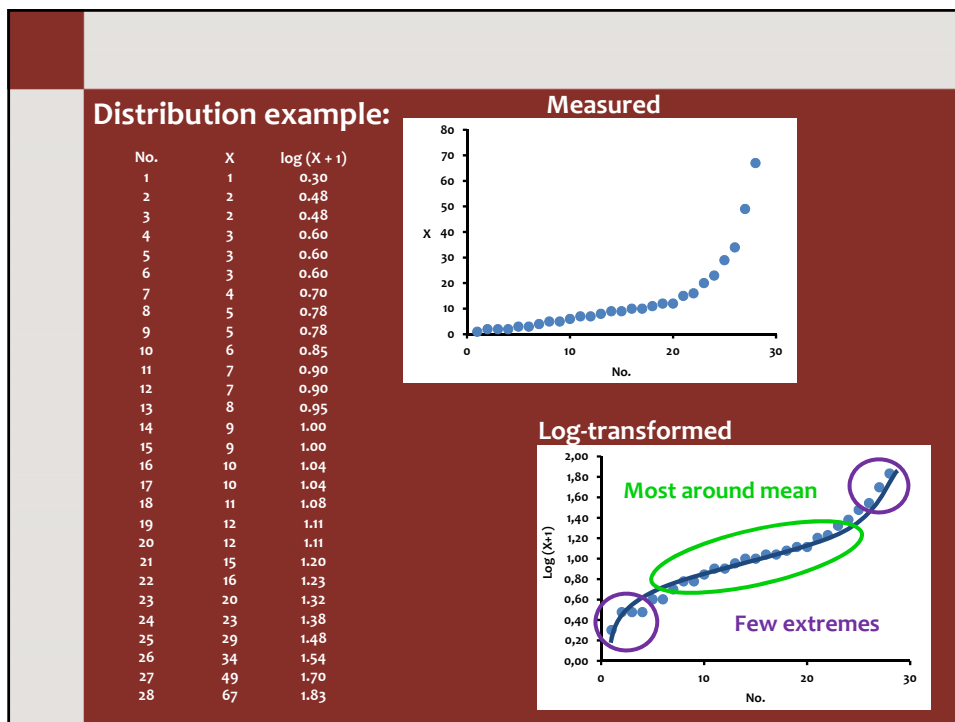
24: Peradventure there be fifty righteous within the city: wilt thou also destroy and not spare the place for the fifty righteous that are therein?

32: ...peradventure ten shall be found there? And He [The Lord] said, I will not destroy it for the ten's sake.

## Distribution example:

No.	X	$\log(X + 1)$
1	1	0.30
2	2	0.48
3	2	0.48
4	2	0.48
5	3	0.60
6	3	0.60
7	4	0.70
8	5	0.78
9	5	0.78
10	6	0.85
11	7	0.90
12	7	0.90
13	8	0.95
14	9	1.00
15	9	1.00
16	10	1.04
17	10	1.04
18	11	1.08
19	12	1.11
20	12	1.11
21	15	1.20
22	16	1.23
23	20	1.32
24	23	1.38
25	29	1.48
26	34	1.54
27	49	1.70
28	67	1.83





**Example: Analysis of variance (ANOVA)**  
**Logarithmic transformation**

1                    100    70  
 5   1                    100  
   15                    100 120  
 1 0 1                100    60  
     1                                    100  
     100 100

1000  
 2000 800  
 1000    500  
   1000    1000  
 1000    1000  
   1000

**Example: Analysis of variance (ANOVA)**  
**Logarithmic transformation**

0.3                    2    1.9  
 0.7   0.3                    2  
   0.3 0.7                    2 2.1  
 0.3 0 0.3                2    1.8  
   0.3 0.7                                    2  
     2 2

3  
 3.3 2.9  
 3    2.7  
   3    3  
   3    3  
     3

**Post-hoc test**

	Homogeneous groups		
Mean	1	2	3
0.41	*		
1.97		*	
2.99			*

Some individuals use statistics as a drunk man uses lamp post - for support rather than for illumination.

Andrew Lang



### Examples of titles and interpretations of analyses

Relationships among number of taxa and environmental factors expressed as Pearson's correlation coefficient;  
Statistically significant correlations ( $p < 0.05$ ) are bold.

	Lumbriculidae	Ancylus	Amphinemura
T	0.28	0.20	-0.28
O <sub>2</sub>	<b>-0.31</b>	-0.22	0.24
KPK	<b>-0.32</b>	0.24	<b>0.45</b>
pH	<b>-0.32</b>	<b>0.72</b>	0.25
Chl a	-0.14	<b>0.98</b>	0.16

Statistically significant negative correlations ( $p < 0.05$ ) were found between the number of individuals of Lumbriculidae and oxygen concentration, quantity of dissolved organic matter given as chemical oxygen demand (COD) and pH, respectively. Positive correlations were found between population size of Ancylus and pH and amount of chlorophyll a respectively, and the number of individuals of the genus Amphinemura and COD.

### Examples of titles and interpretations of analyses

Analysis of variance comparison of habitats with different flow velocity in respect to the content of detritus.

	SS	Df	MS	F	p
Flow velocity	0.08	3	0.08	6.56	0.014

*Analysis of variance showed statistically significant difference ( $p = 0.014$ ) in amount of detritus between (among) different habitats flow velocity of water.*

NOTE: The results of this analysis does not reveal where more detritus is deposited so the conclusion is general - the difference we have found, but the character of the difference we did not.

In the case of only two habitats - a reference to mean values would be enough that we can conclude specifically e.g.: The habitat x accumulated significantly more detritus than the habitat y.

But completely correct way and the only way if more than two habitats are viewed is to do a post-hoc test.

### Examples of titles and interpretations of analyses

Post-hoc Tukey HSD test for the mass of accumulated detritus among four habitats with different flow velocities.

Flow velocity	Mean detritus mass	Homogenous groups	
		1	2
< 30 cm s <sup>-1</sup>	0.51	x	
30-60 cm s <sup>-1</sup>	0.65	x	
60-90 cm s <sup>-1</sup>	0.71	x	
>90 cm s <sup>-1</sup>	1.25		x

*The results of post-hoc Tukey HSD test showed that significantly more detritus was accumulated in the habitat with the fastest water flow. Among other habitats no statistically significant difference in the amount of accumulated detritus was found (as they were all grouped within the same homogenous group).*