# Experimental Data Analysis

*in ©MATLAB*

## Lecture 4:

## Correlations, normality of data testing, parametric vs. non-parametric tests

Jan Rusz
Czech Technical University in Prague

# Motivation

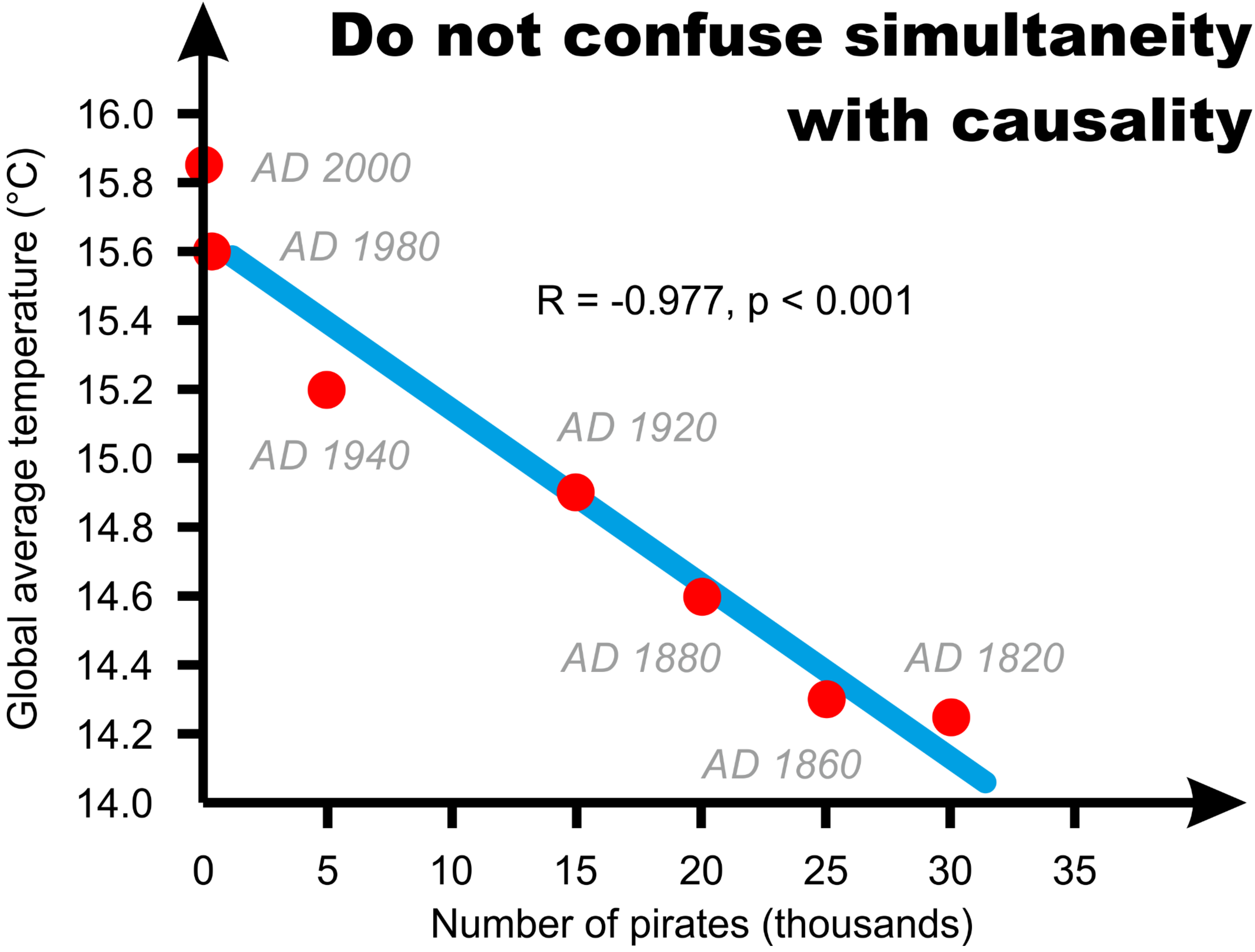
## Association

Question: Can be increased blood pressure associated with stress?
Answer: Correlation analysis.

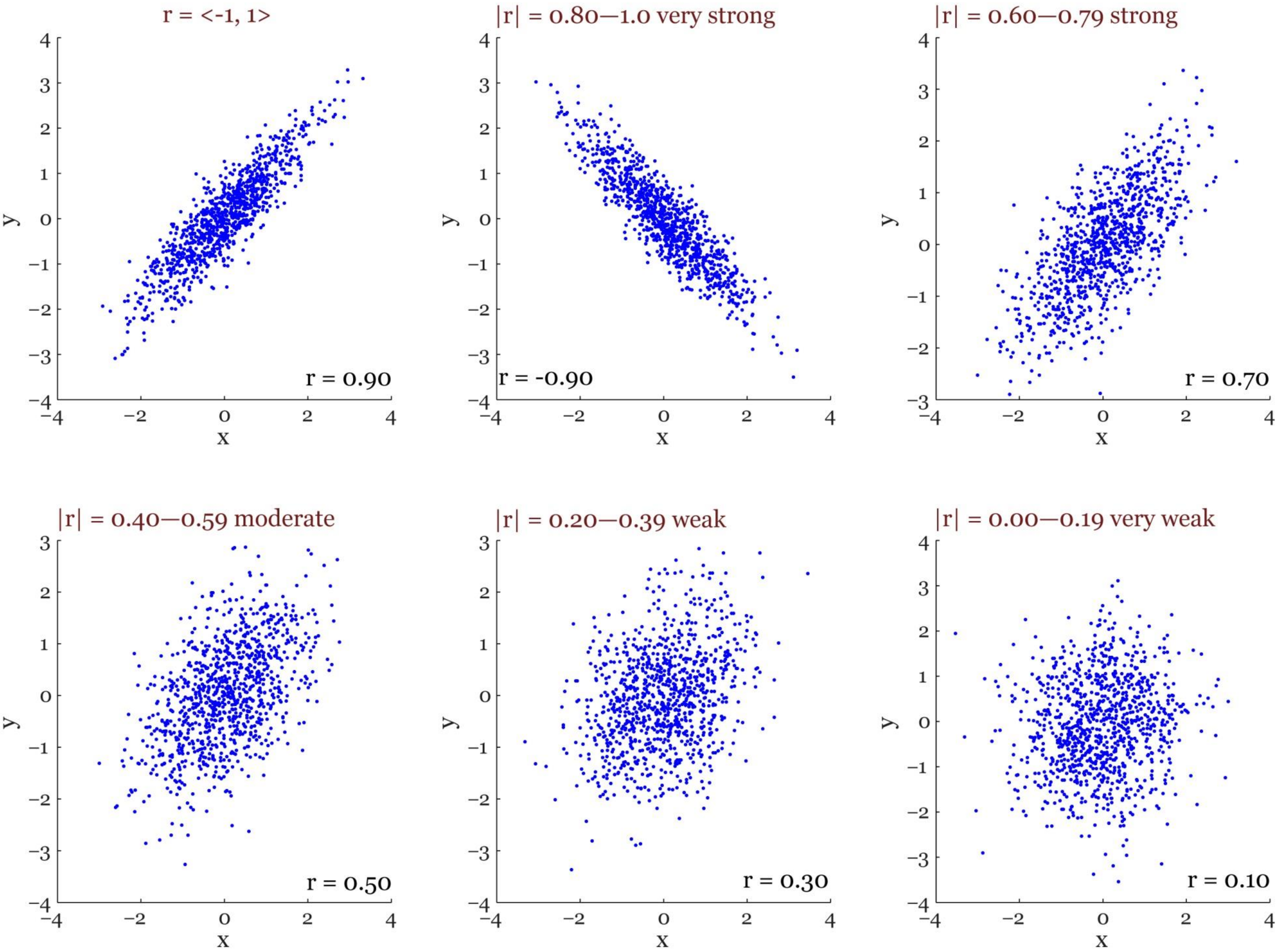
## Connection

- Correlation indicates a relationship not causality.
- We need to find a connection to say that relationship is causal (i.e. examine that hormonal response to stress can elevate blood pressure).

# Do not confuse simultaneity with causality



# Correlation values



r = <-1, 1>

|r| = 0.80—1.0 very strong

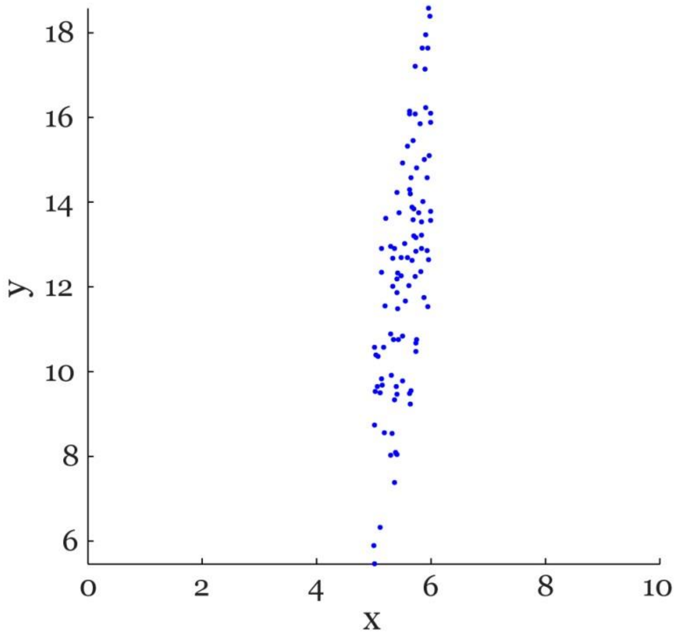
|r| = 0.60—0.79 strong

|r| = 0.40—0.59 moderate

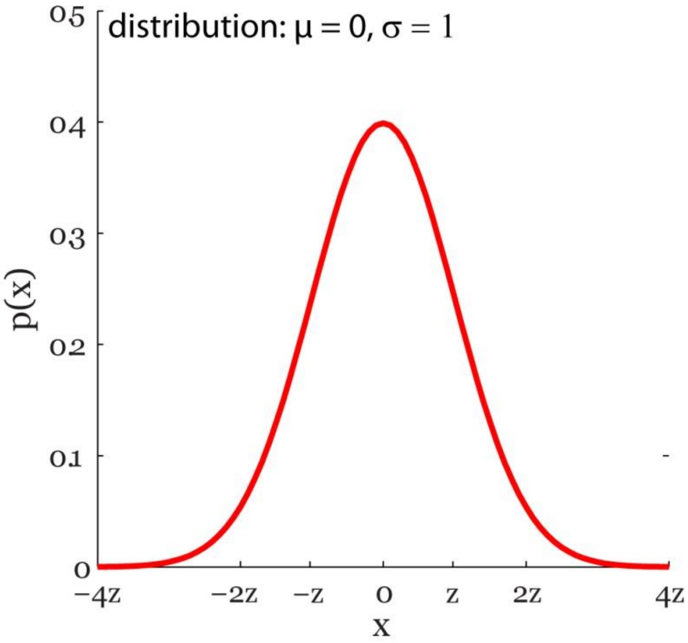
|r| = 0.20—0.39 weak

|r| = 0.00—0.19 very weak

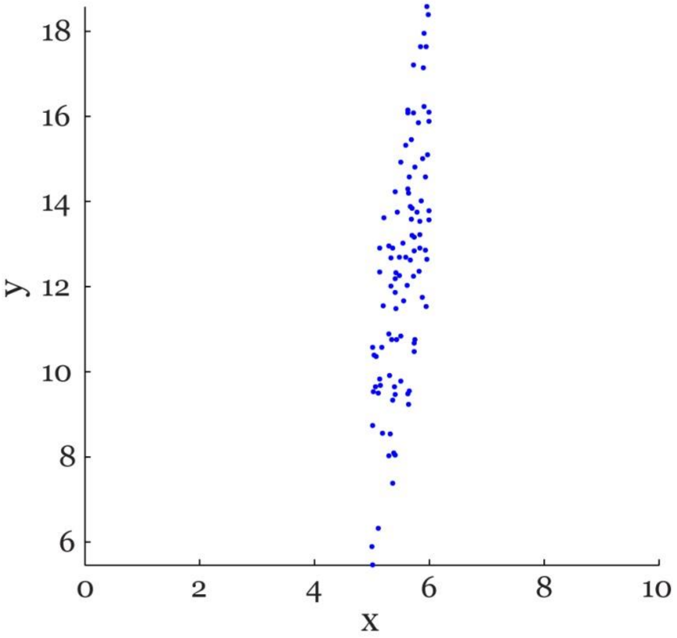
## Data (raw)



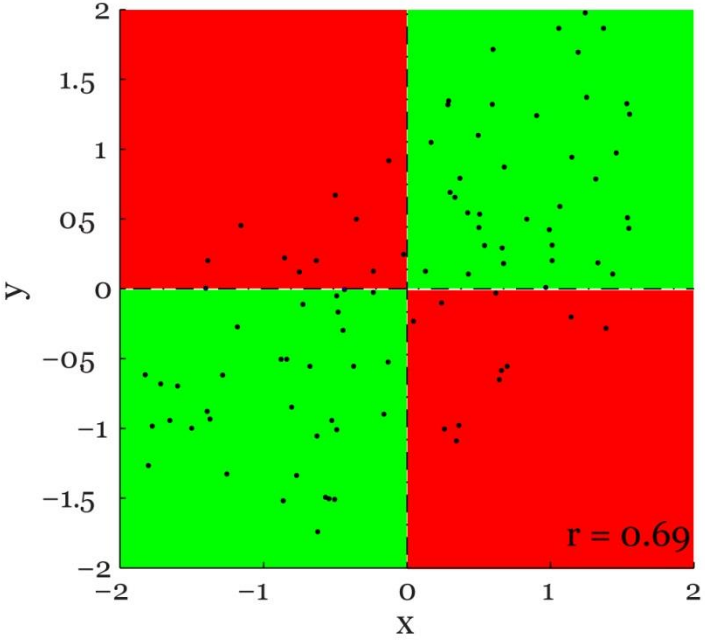
## Data (z-scored)



## Data (raw)

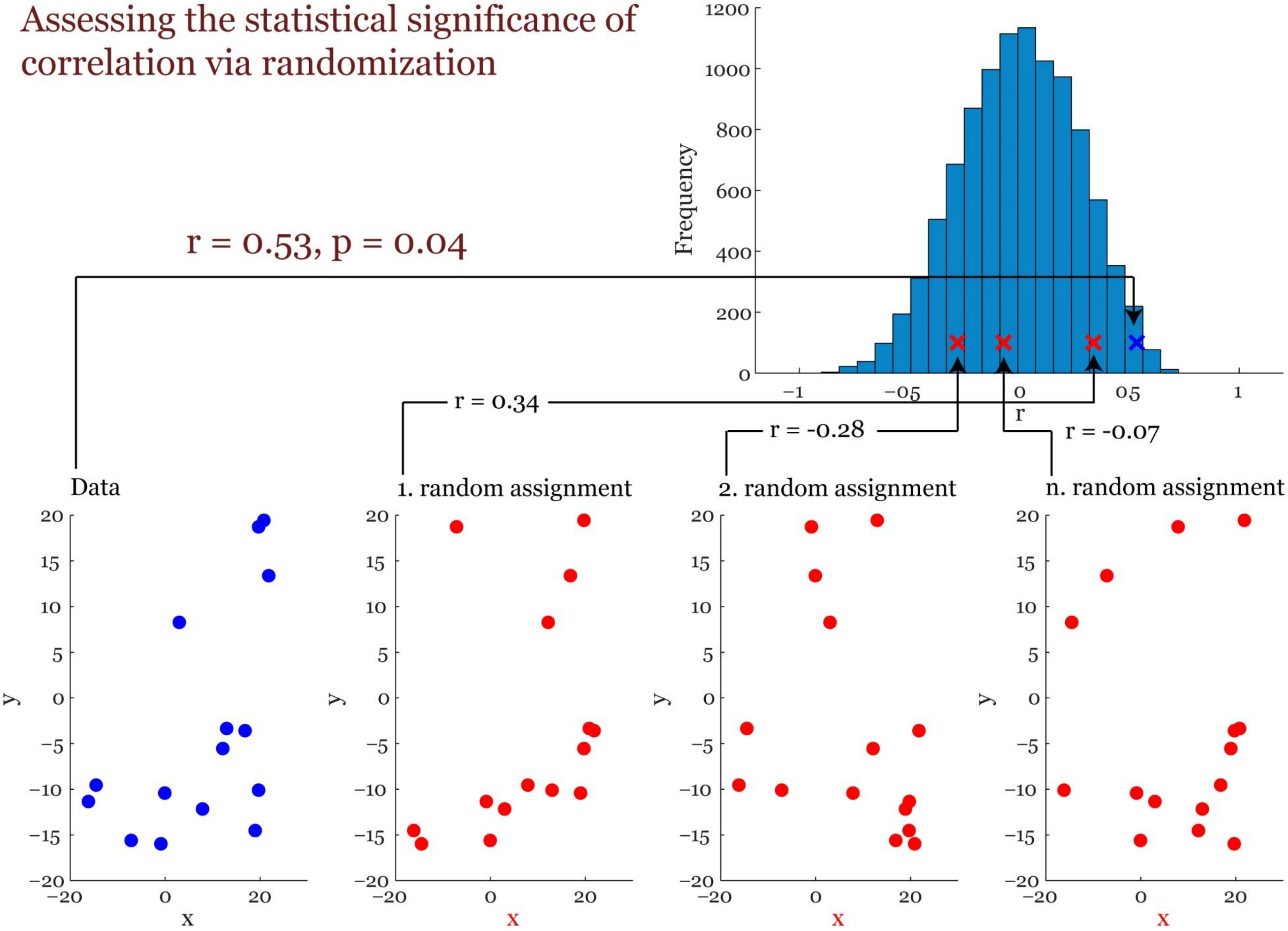


## Data (z-scored)



$$r = \frac{\sum_{i=1}^{n}\left(\frac{x_i - \mu_x}{\sigma_x}\right)\left(\frac{y_i - \mu_y}{\sigma_y}\right)}{n}$$

# Assessing the statistical significance of correlation via randomization



r = 0.53, p = 0.04

r = 0.34

r = -0.28

r = -0.07

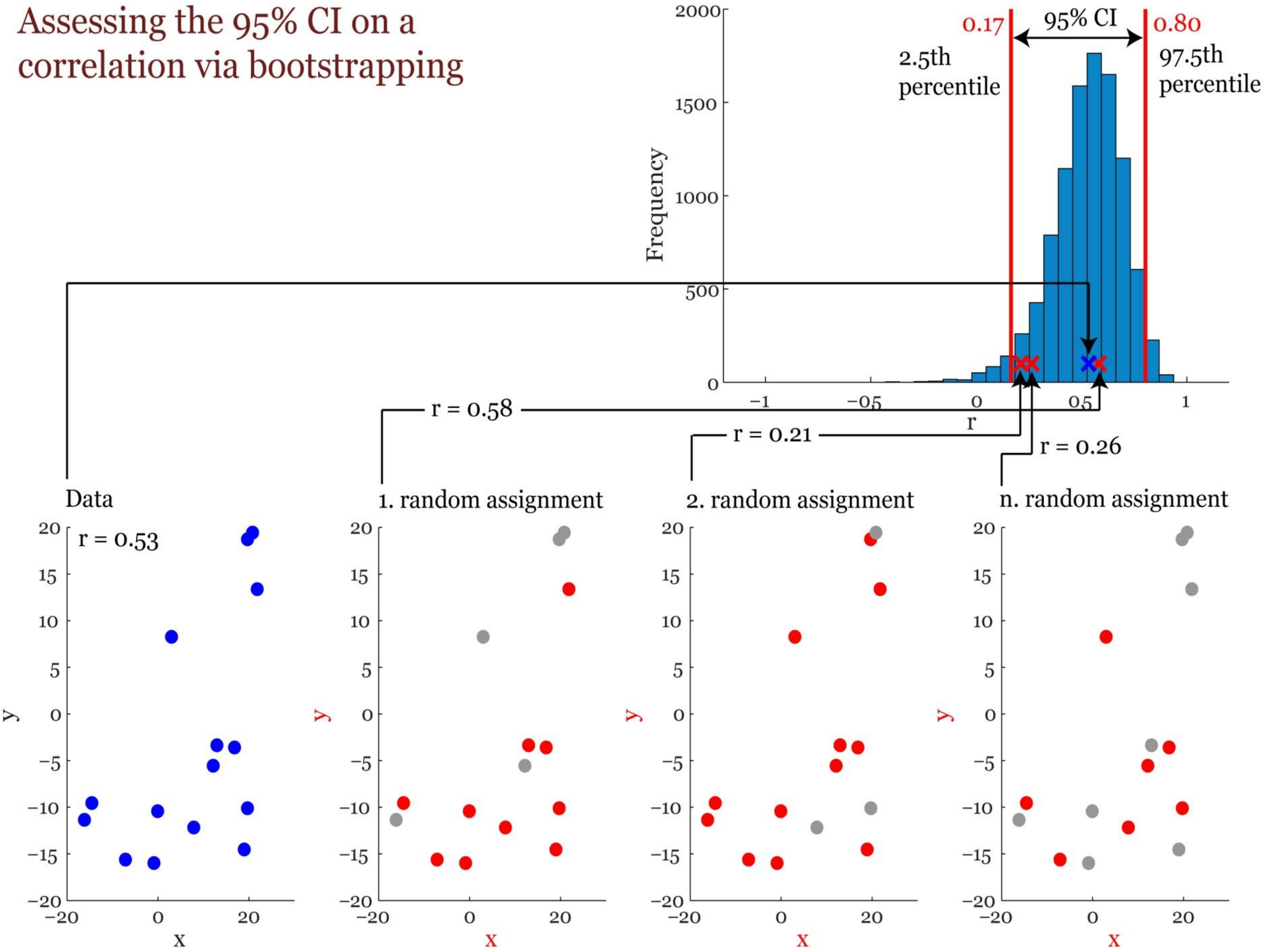
Data

1. random assignment

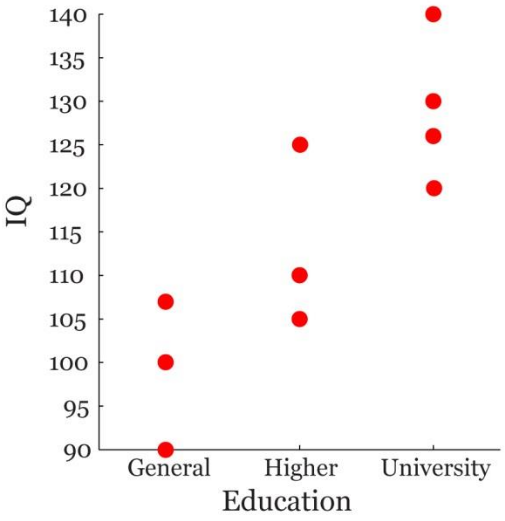
2. random assignment

n. random assignment

# Assessing the 95% CI on a correlation via bootstrapping



# Spearman rank correlation



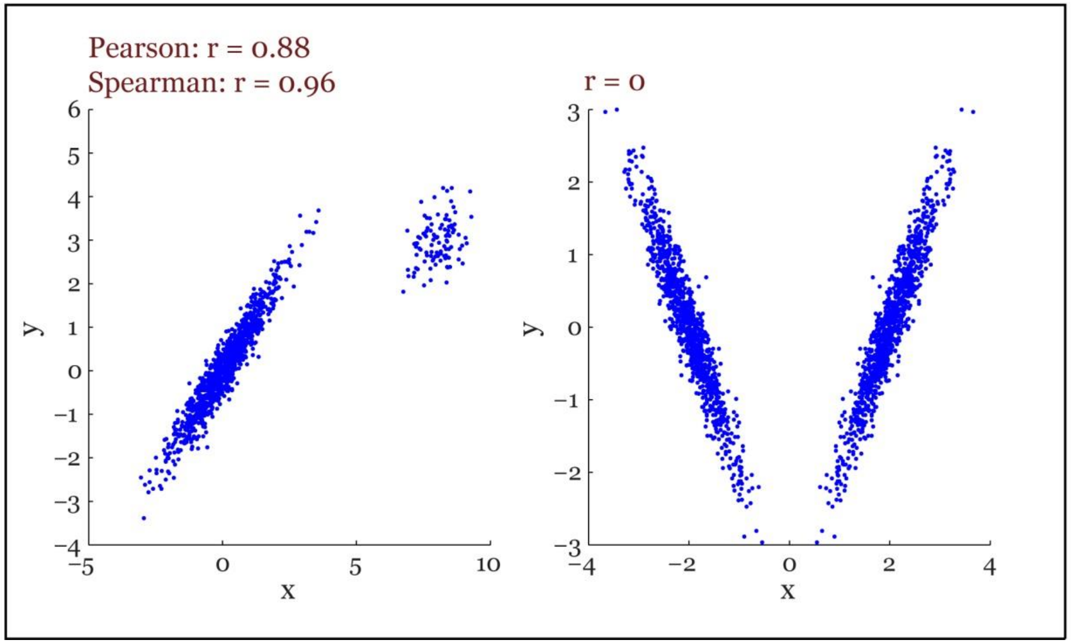
| Raw data | | Ranks | | Difference | |
|---|---|---|---|---|---|
| IQ | Education | $x_r$ | $y_r$ | $d = x_r - y_r$ | $d^2$ |
| 100 | General | 2 | 2 | 0 | 0 |
| 105 | Higher | 3 | 5 | -2 | 4 |
| 140 | University | 10 | 8.5 | 1.5 | 2.25 |
| 125 | Higher | 7 | 5 | 2 | 4 |
| 110 | Higher | 5 | 5 | 0 | 0 |
| 130 | University | 9 | 8.5 | 0.5 | 0.25 |
| 90 | General | 1 | 2 | -1 | 1 |
| 107 | General | 4 | 2 | 2 | 4 |
| 120 | University | 6 | 8.5 | -2.5 | 6.25 |
| 126 | University | 8 | 8.5 | -0.5 | 0.25 |

$\sum d^2 = 22$

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

$$r_s = 1 - \frac{6 \times 22}{10(100 - 1)}$$

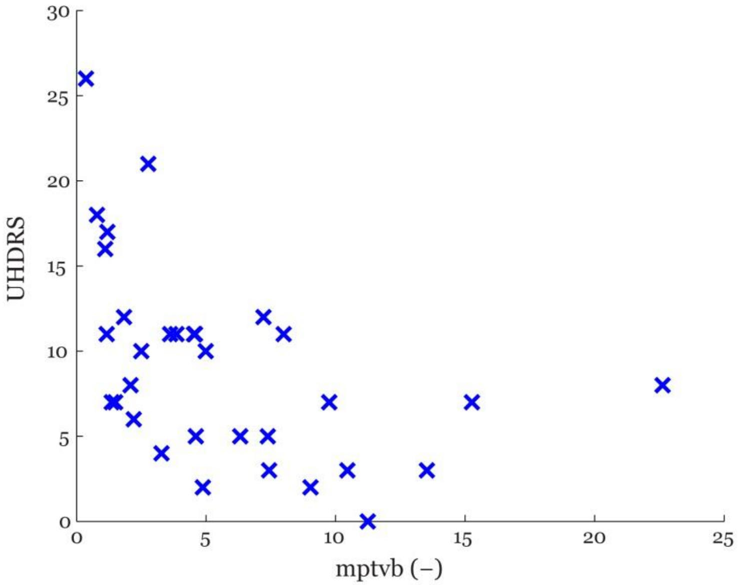
$$r_s = 0.87$$



## Correlations

Matlab example 1

Normal speaker is able to perform sustained vowel phonation for several seconds without voice breaks that represent impaired function of vocal folds. To verify if vocal fold function disability in patients with Huntington's disease (HD) corresponds to overall motor disability, researcher collected sustained phonations from 32 speakers with HD and performed analysis of maximum phonation time until voice breaks  (mptvb). He also assessed every patient using clinical motor scale of Unified Huntington's Disease Raring Scale (UHDRS).



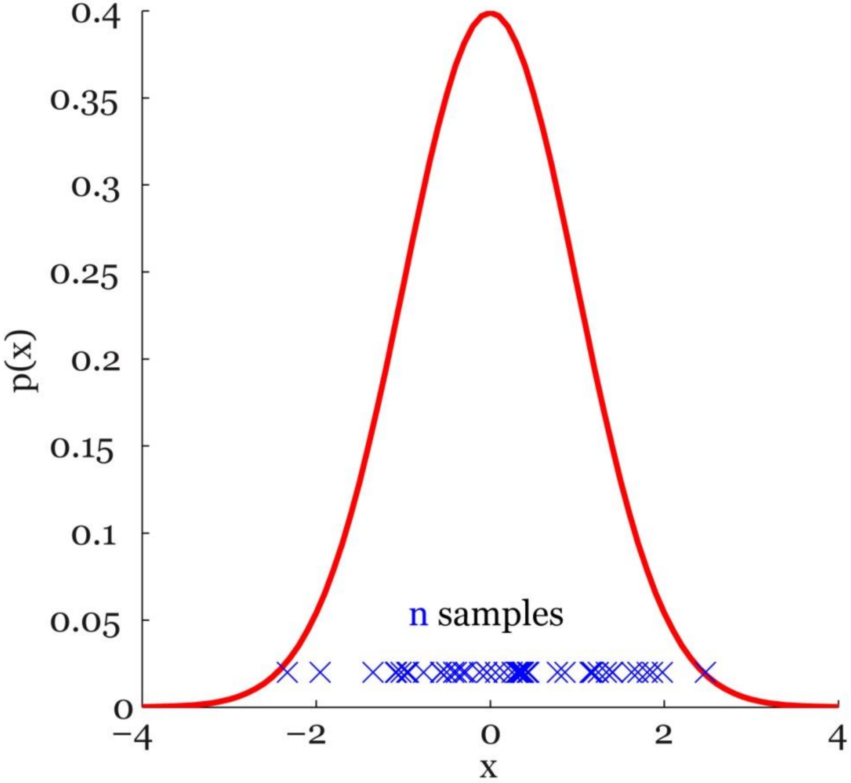
Pearson:
$r = -0.47, p = 0.01$

Spearman:
$r = -0.61, p < 0.001$

Spearman correlation is more powerful due to non-normally distributed data

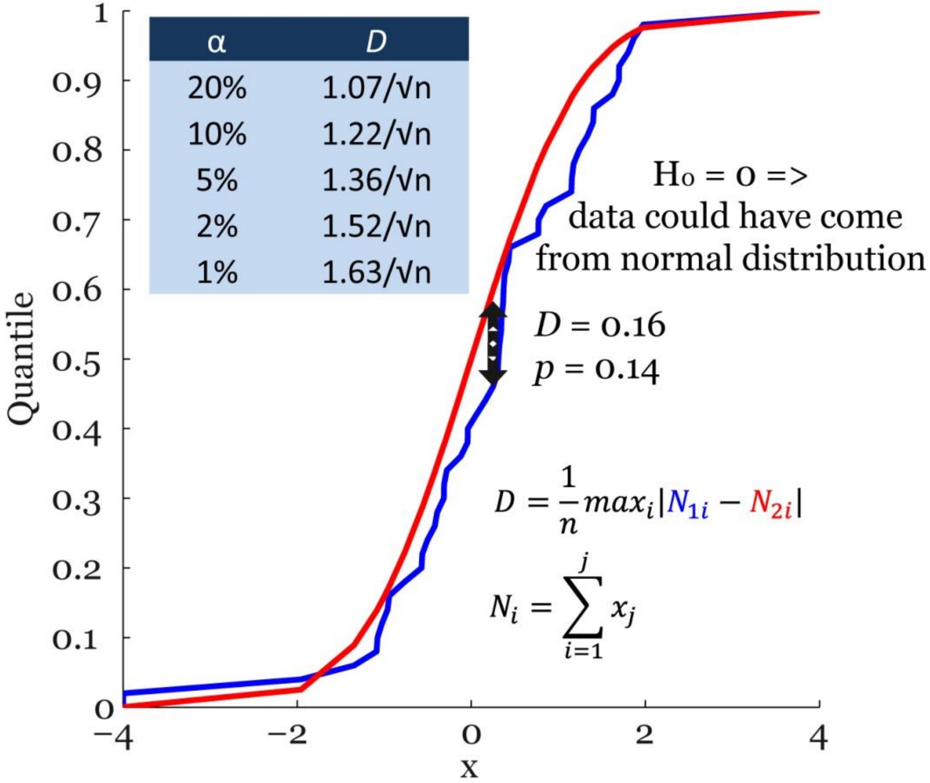
# Testing normality of the data using goodness-of-fit test: Kolmogorov-Smirnov

## Probability density function (PDF)



n samples

## Cumulative distribution function (CDF)



| α | *D* |
|---|---|
| 20% | 1.07/√n |
| 10% | 1.22/√n |
| 5% | 1.36/√n |
| 2% | 1.52/√n |
| 1% | 1.63/√n |

H₀ = 0 =>
data could have come
from normal distribution

$D = 0.16$
$p = 0.14$

$$D = \frac{1}{n} max_i |N_{1i} - N_{2i}|$$

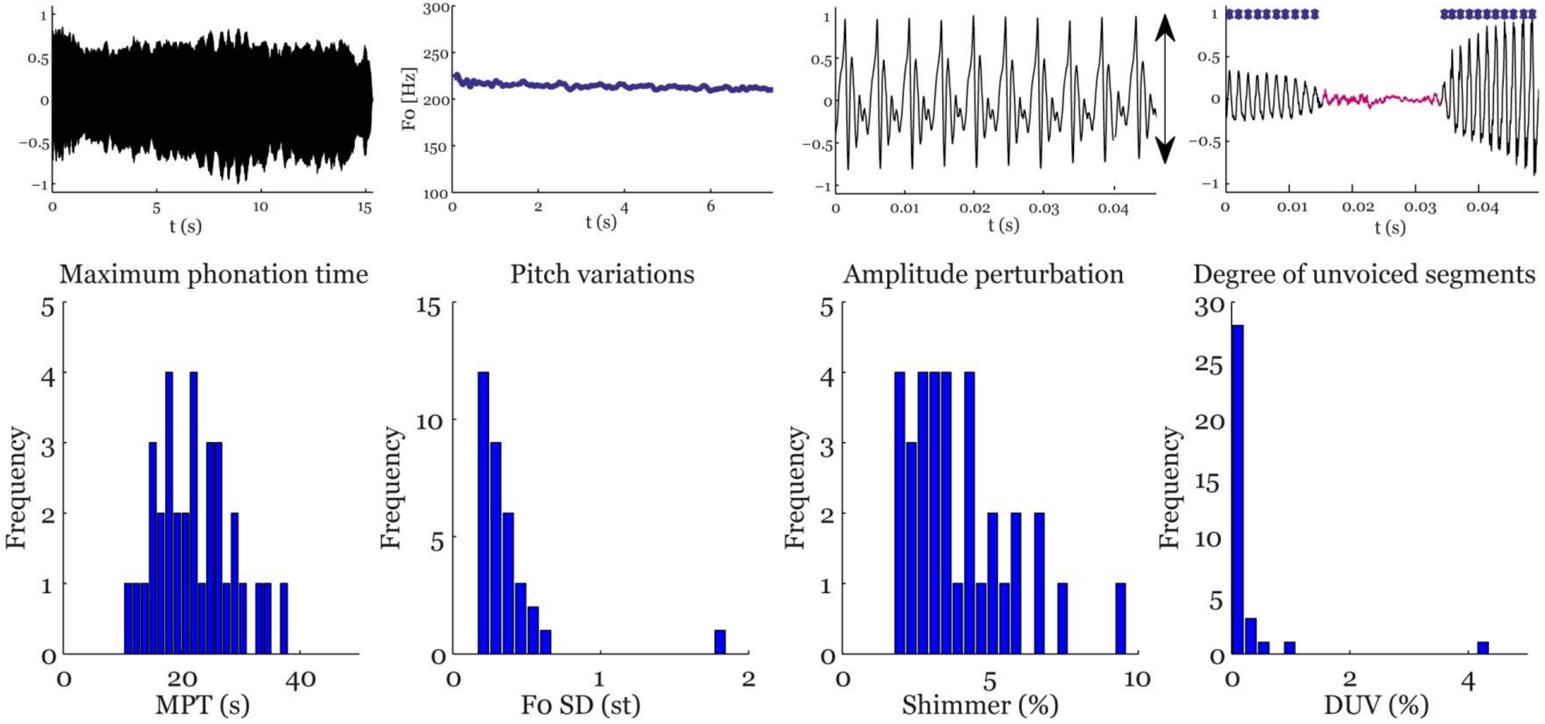
$$N_i = \sum_{i=1}^{j} x_j$$

CDF is integral of PDF evaluated from -∞ to x

# Testing the data normality using Kolmogorov-Smirnov test

Matlab example **2**

Researcher wants to verify the normality of data based upon 4 measures extracted from sustained phonation of 34 healthy speakers.



Maximum phonation time

Pitch variations

Amplitude perturbation

Degree of unvoiced segments

Normal distribution

$D = 0.11$, $p = 0.76$
Do not reject $H_0$!

Normal distribution with outliers

$D = 0.25$, $p = 0.02$
Reject $H_0$!

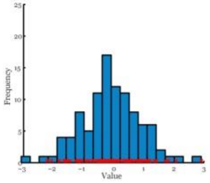
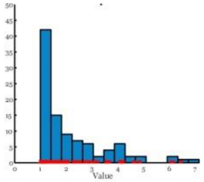
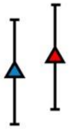
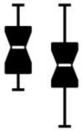
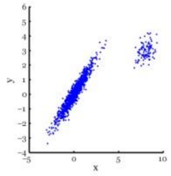
Log-normal distribution

$D = 0.14$, $p = 0.52$
Do not reject $H_0$!

Non-normal distribution

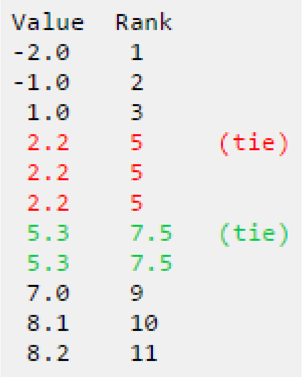
$D = 0.39$, $p < 0.001$
Reject $H_0$!

# 1) Parametric vs. non-parametric data

| | Parametric | | Non-parametric | |
|---|---|---|---|---|
| Expected distribution: | Normal |  | No limits |  |
| Expected variance: | Homogeneous |  | No limits |  |
| Typical scale: | Interval | $y = x + d$<br>dates (years), temperature (°C), IQ scale | Nominal | Non-comparable<br>YES/NO, colours, gender, phone numbers |
| | Ratio | $y = k . x + d$<br>velocities, lengths, age... | Ordinal | Can be sorted in terms of *„greater“* or *„less“*<br>education |
| Central measures: | Mean (SD) | | Median (IQR) | |
| Advantages: | More powerful | | Less vulnerable to outliers |  |
| Disadvantages: | More samples needed | | Less powerful | |

# 2) Tests for normality

Matlab example 3



```
Value   Rank
-2.0     1
-1.0     2
 1.0     3
 2.2     5      (tie)
 2.2     5
 2.2     5
 5.3     7.5    (tie)
 5.3     7.5
 7.0     9
 8.1     10
 8.2     11
```

**Visual inspection**

- Relatively robust

**Chi-Square test** (chi2gof)

- Ties may be problematic
- Unsuitable for small samples

**Kolmogorov-Smirnov test** (kstest)

- Work for small samples (optimal > 50)
- Ties are not problem
- Lower power

**Shapiro-Wilk test** (not available in Matlab)

- Highest power among all tests of normality
- Most commonly used
- Samples < 50

**Liliefors test** (lillietest)

- Higher power than KS test
- Not effective for small samples (n >80)

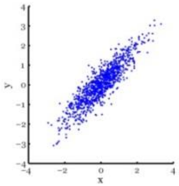
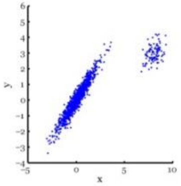
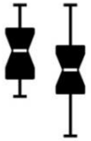
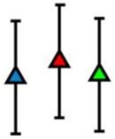
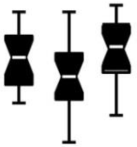
**Anderson-Darling test** (adtest)

- Higher power than KS test

**Cramér-von-Mises test** (not available in Matlab)

- Higher power than KS test

# 3) Parametric vs. non-parametric tests

| | Parametric | Non-parametric |
|---|---|---|
| Correlations: | Pearson<br>*corr*<br> | Spearman<br>*corr*<br> |
| Independent<br>2 groups: | Independent t-test<br>*ttest2*<br> | Mann-Whitney U test<br>*ranksum*<br> |
| Paired<br>2 „levels“: | Paired t-test<br>*ttest* | Wilcoxon signed-rank test<br>*signrank* |
| Independent<br>> 2 groups: | Analysis of variance<br>*anova1*<br> | Kruskal-Wallis test<br>*kruskalwallis*<br> |
| Paired<br>> 2 „levels“: | Repeated measures analysis of variance<br>*ranova* | Friedman test<br>*friedman* |