VETERINARY PUBLIC HEALTH INSTITUTE

UNIVERSITÄT BERN

Non-Normality

Normality Tests and QQ-plots

Dr. Beatriz Vidondo

Veterinary Public Health Institute UniBe

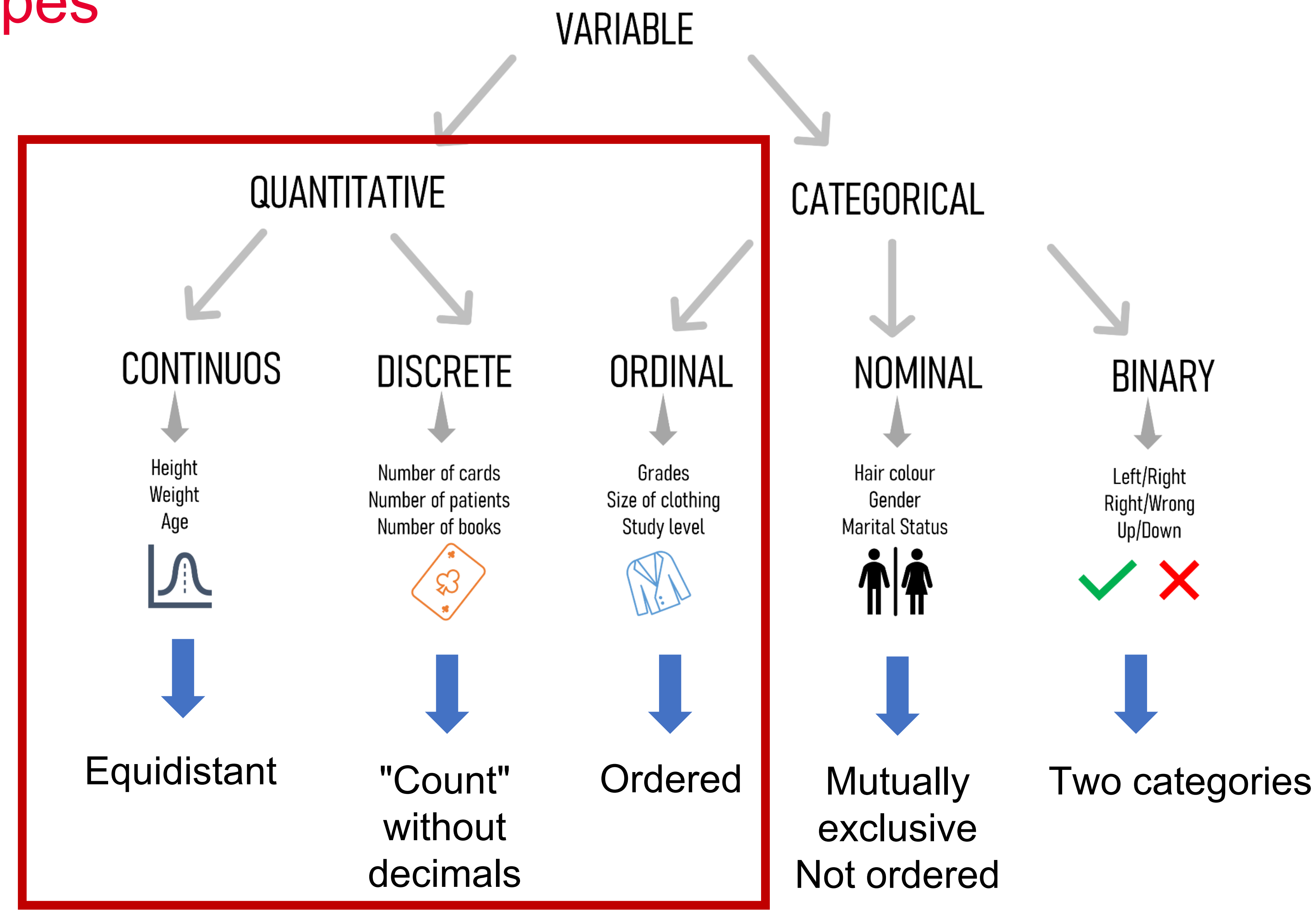
Objectives

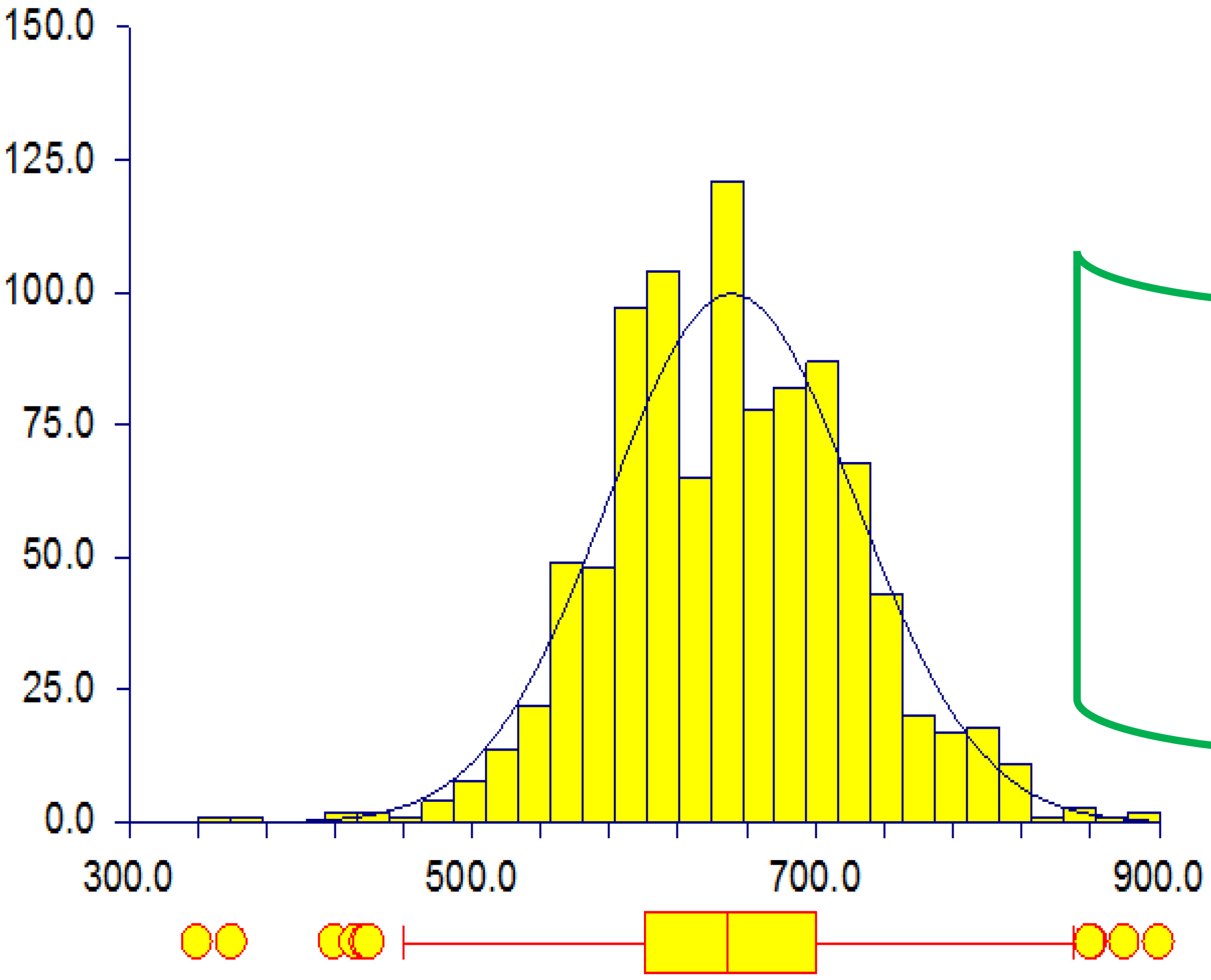
Identify non-Normal data visually and by testing for it

Compare non-Normal groups

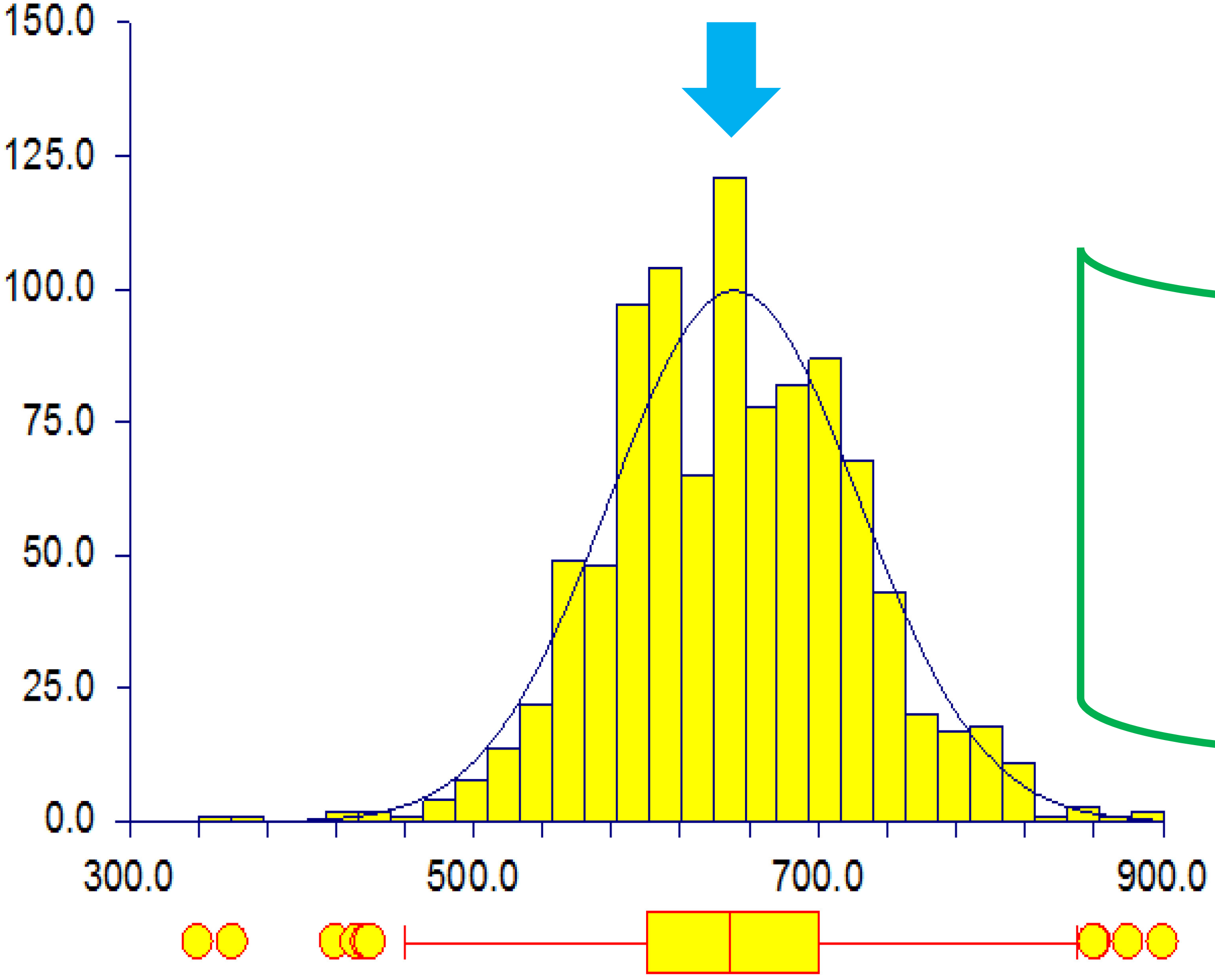
Multiple comparison corrections

Data Types

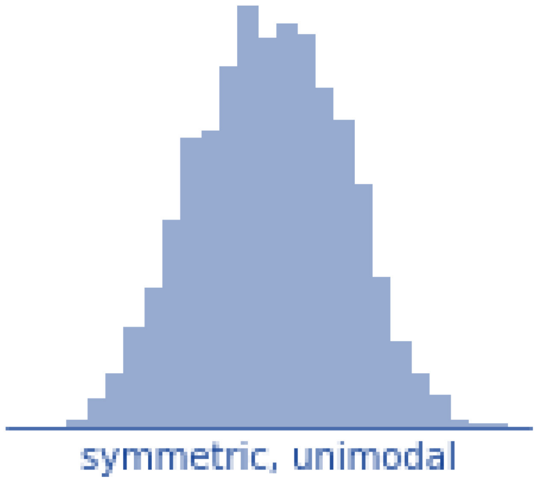

VARIABLE
QUANTITATIVE
CATEGORICAL
CONTINUOS
Height
Weight
Age
Equidistant
DISCRETE
Number of cards
Number of patients
Number of books
"Count" without decimals
ORDINAL
Grades
Size of clothing
Study level
Ordered
NOMINAL
Hair colour
Gender
Marital Status
Mutually exclusive
Not ordered
BINARY
Left/Right
Right/Wrong
Up/Down
Two categories

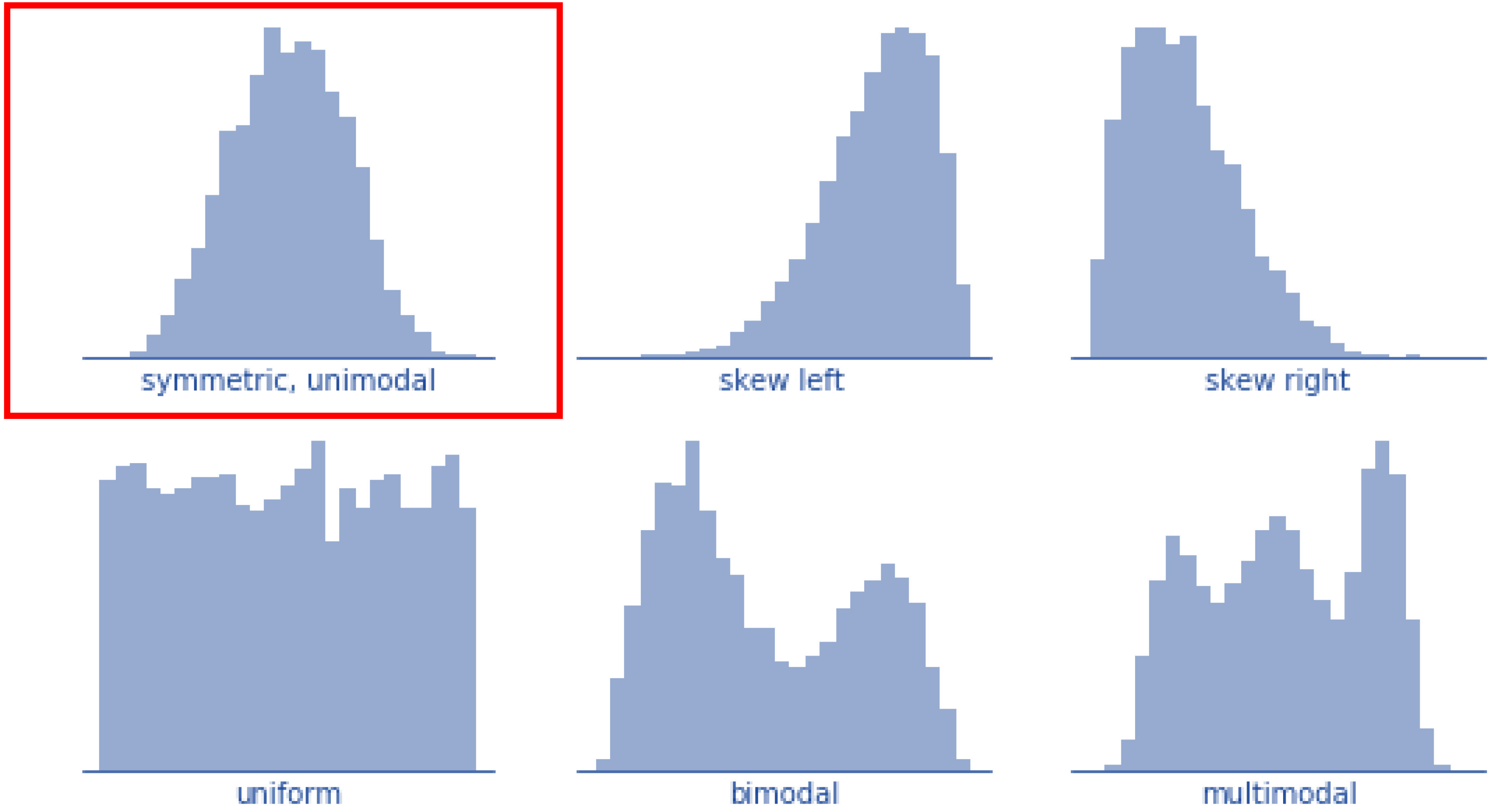


Which three statistics are the same?

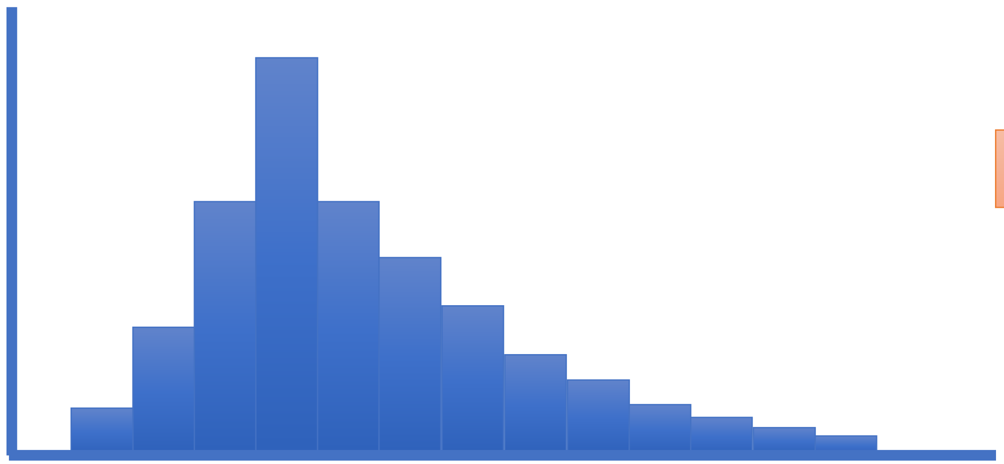


Which three statistics are the same?
mean = median = mode


symmetric, unimodal

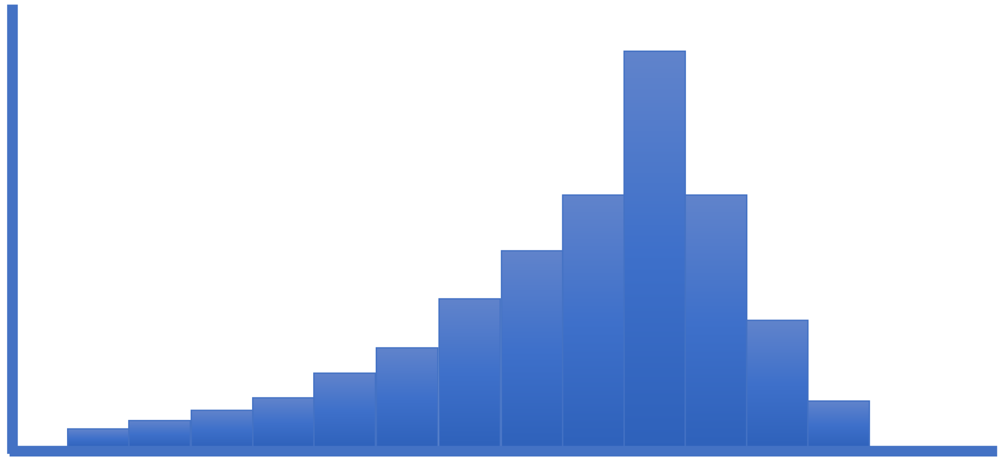

symmetric, unimodal
skew left
skew right
uniform
bimodal
multimodal

Skew



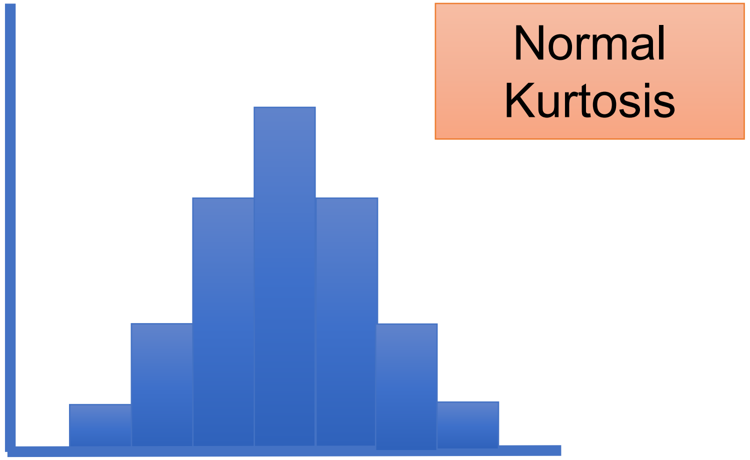
Positive Skew

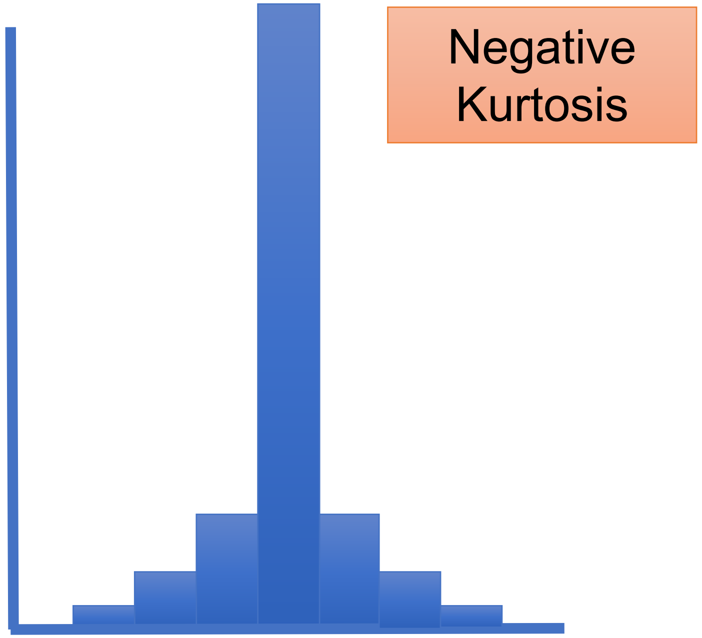
Normal distribution:
Skewness = 0 [-3,3]

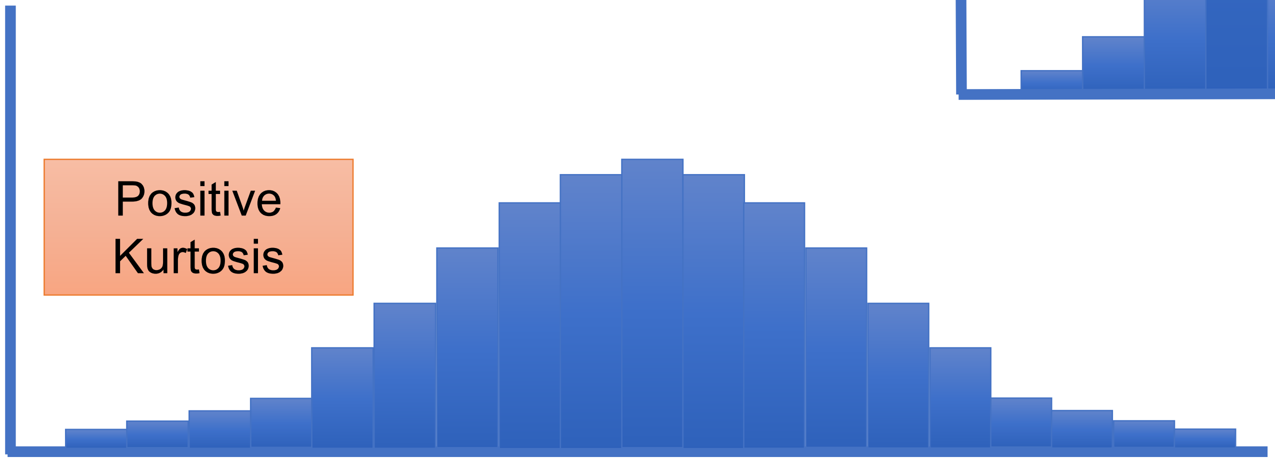


Negative Skew

Kurtosis

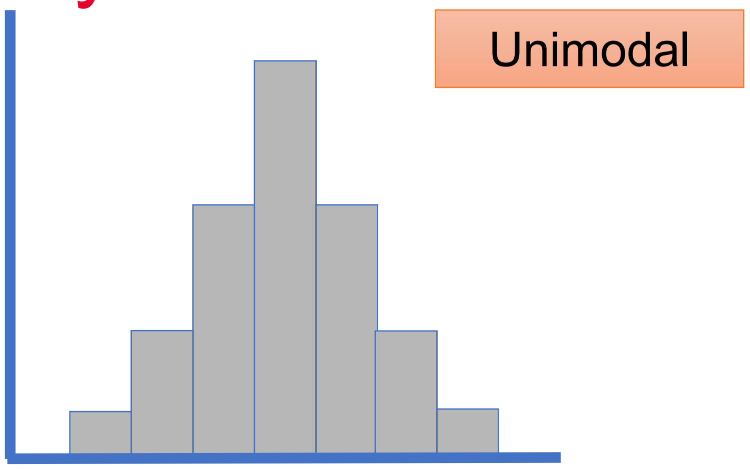


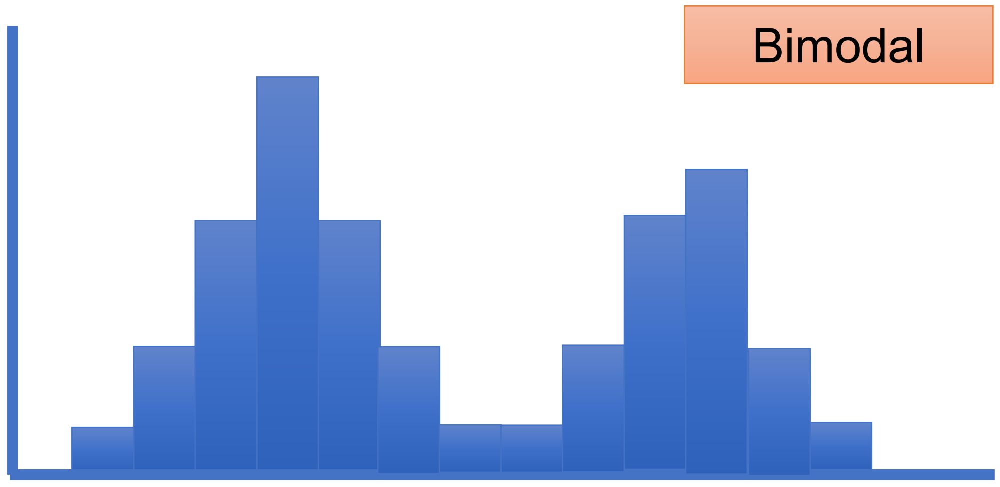


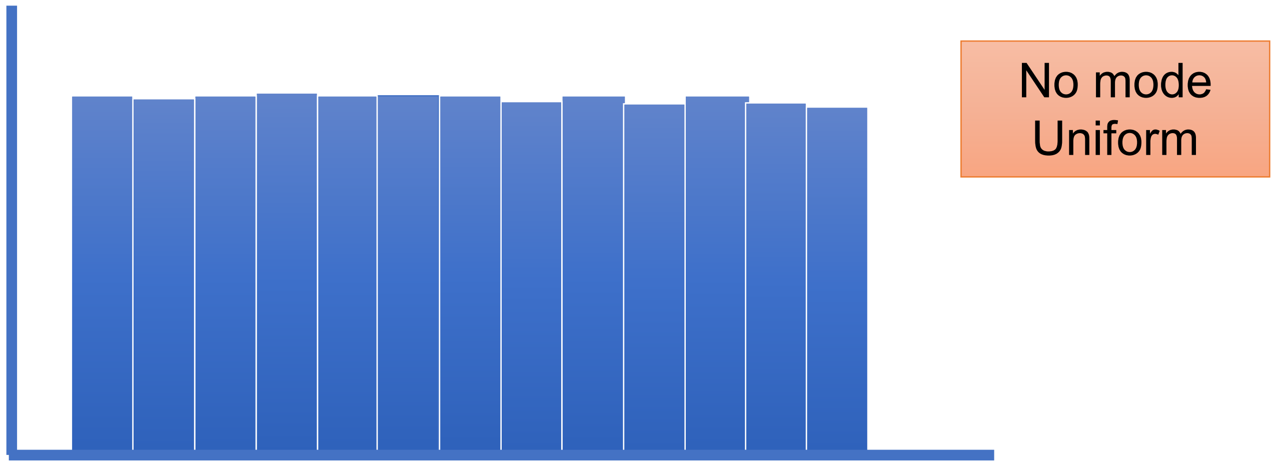


Normal distribution:
Kurtosis = 3 [2,4]

Modality

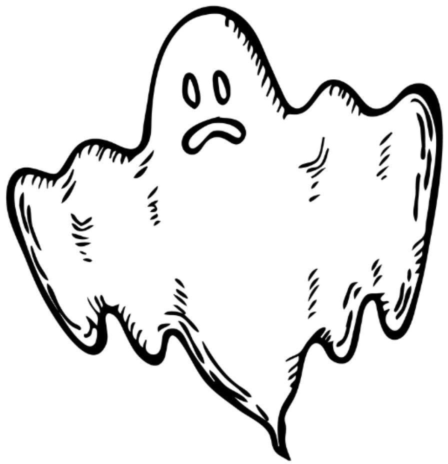


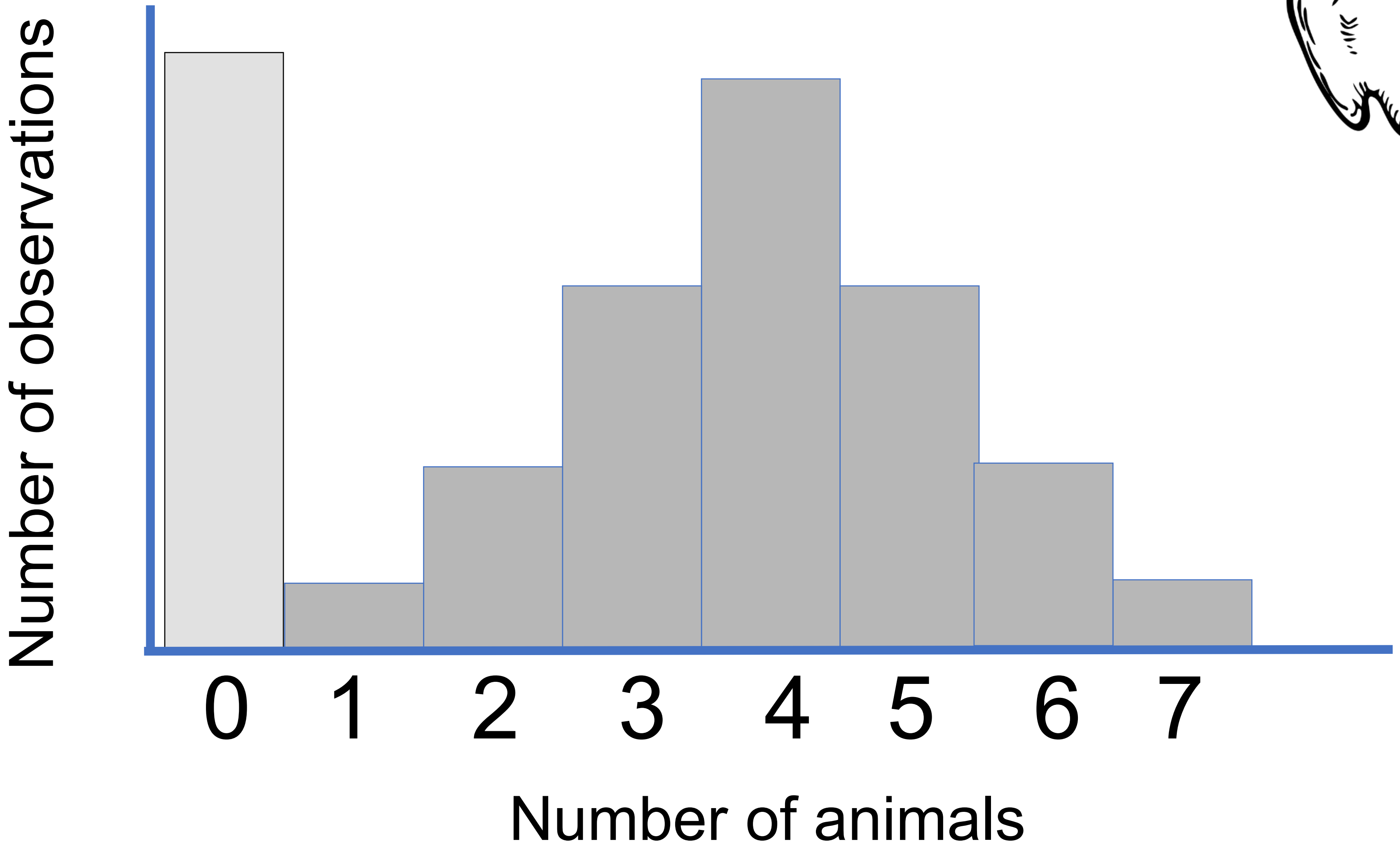




https://pixabay.com/de/illustrations/halloween-und-ghost-geister-4451903/

Zero Inflation



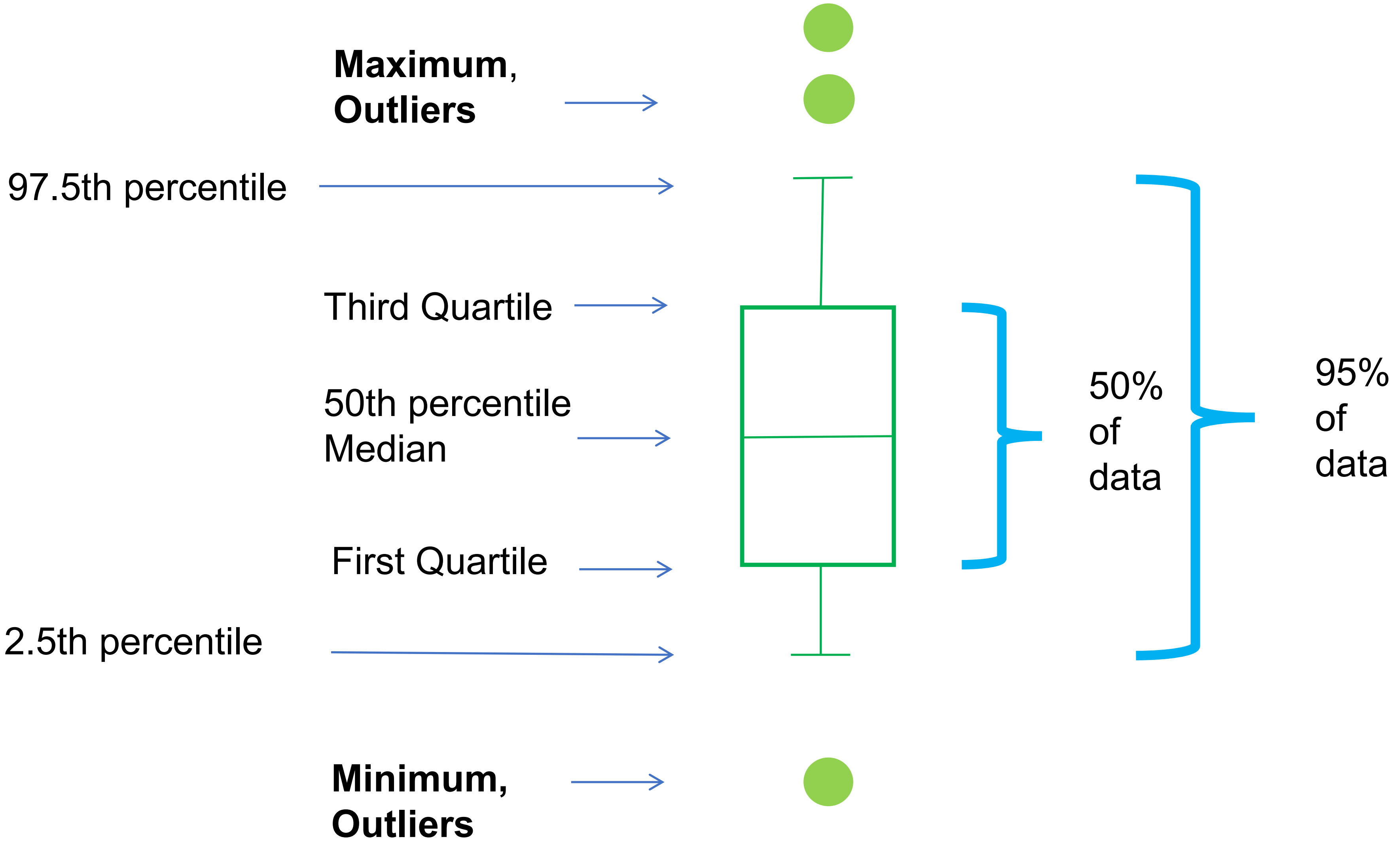


Who you Want to Call?

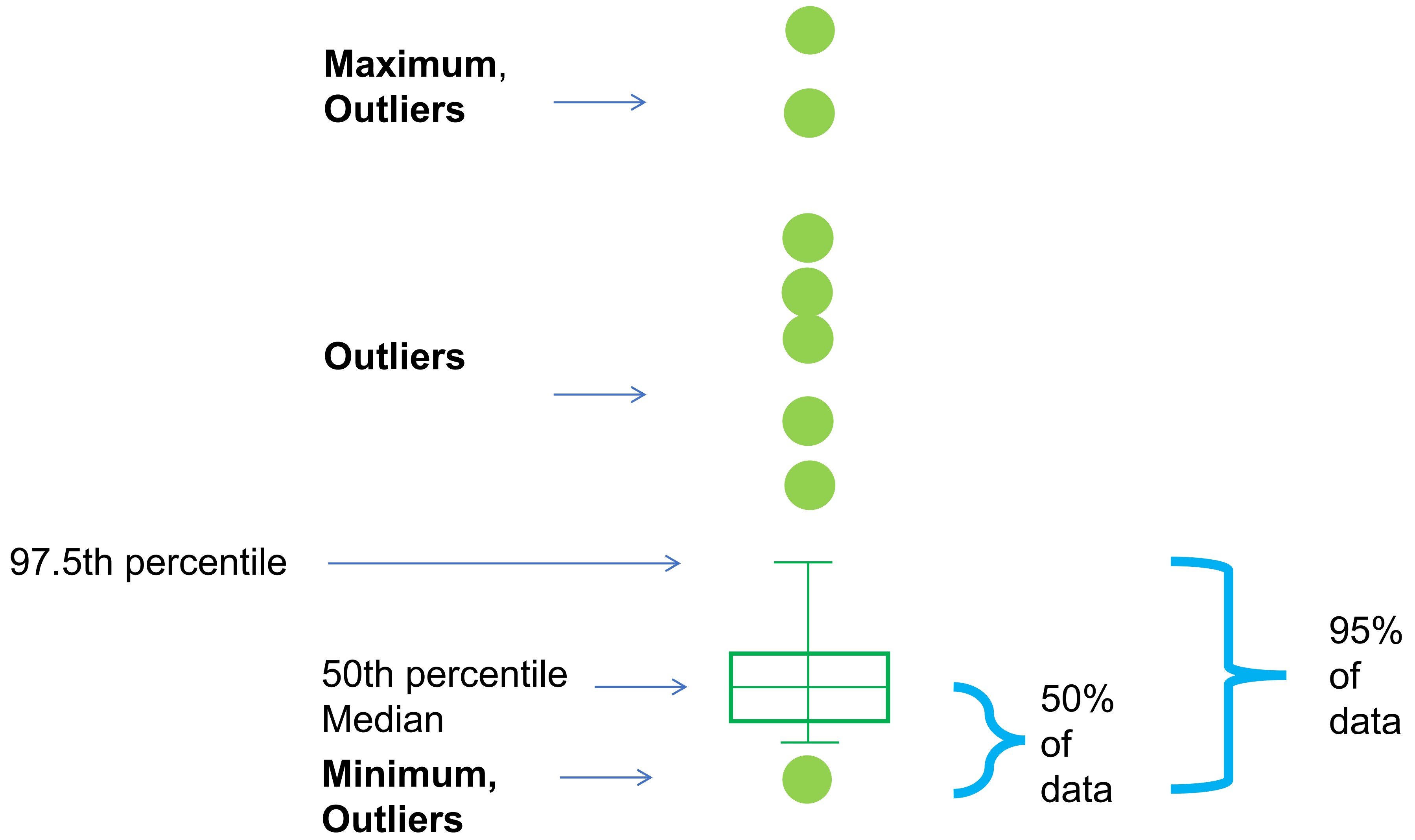
Who you Want to Call?

Percentiles!

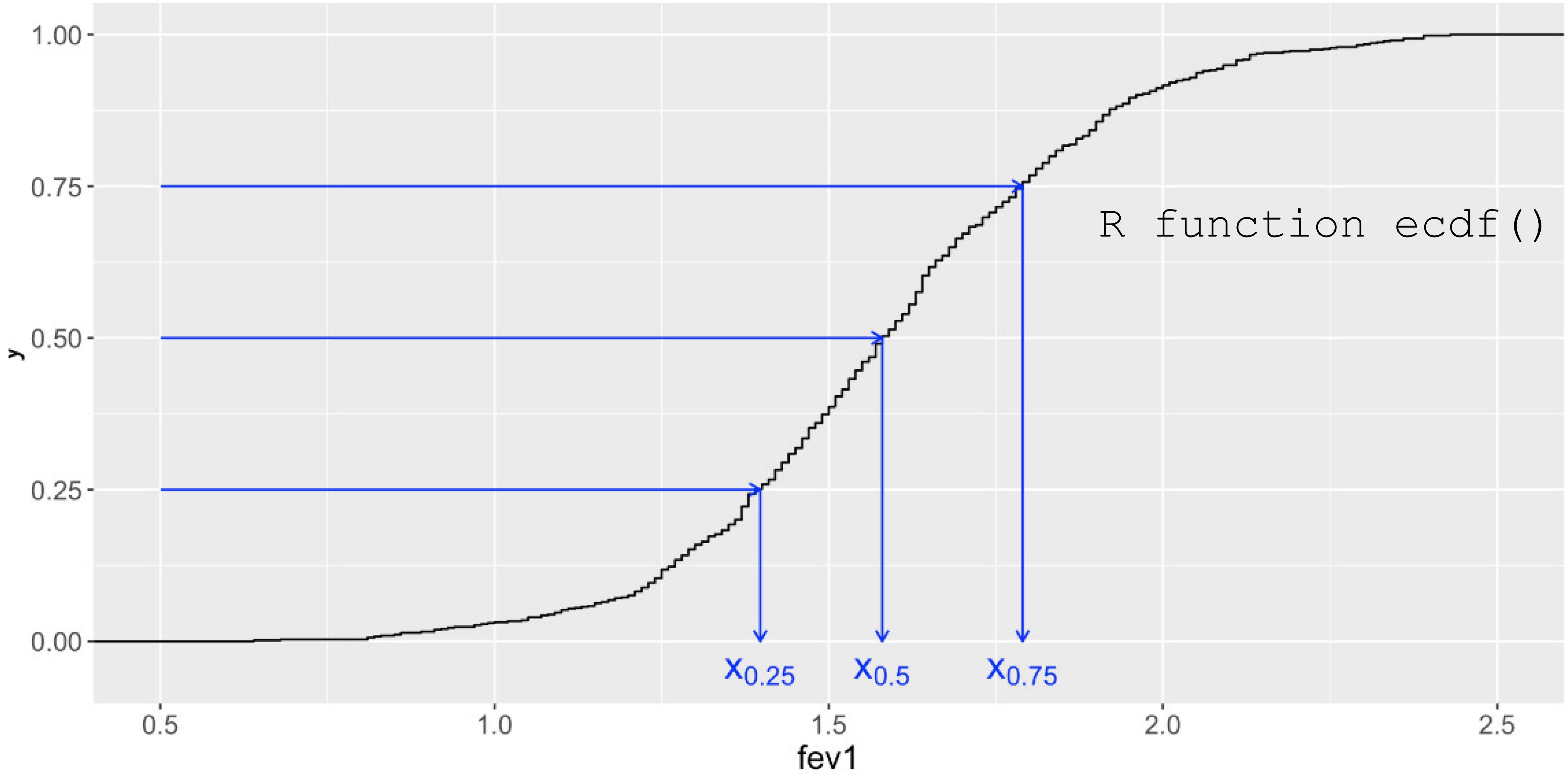
Box and Whiskers Plot (Normal)



Box and Whiskers Plot (Skewed)



Empirical Cumulative Distribution Function



Percentiles

- Sort observations from min to max
- Take 100 segments (1%-segments)
- 25th percentile is the value below which 25% of the data can be found

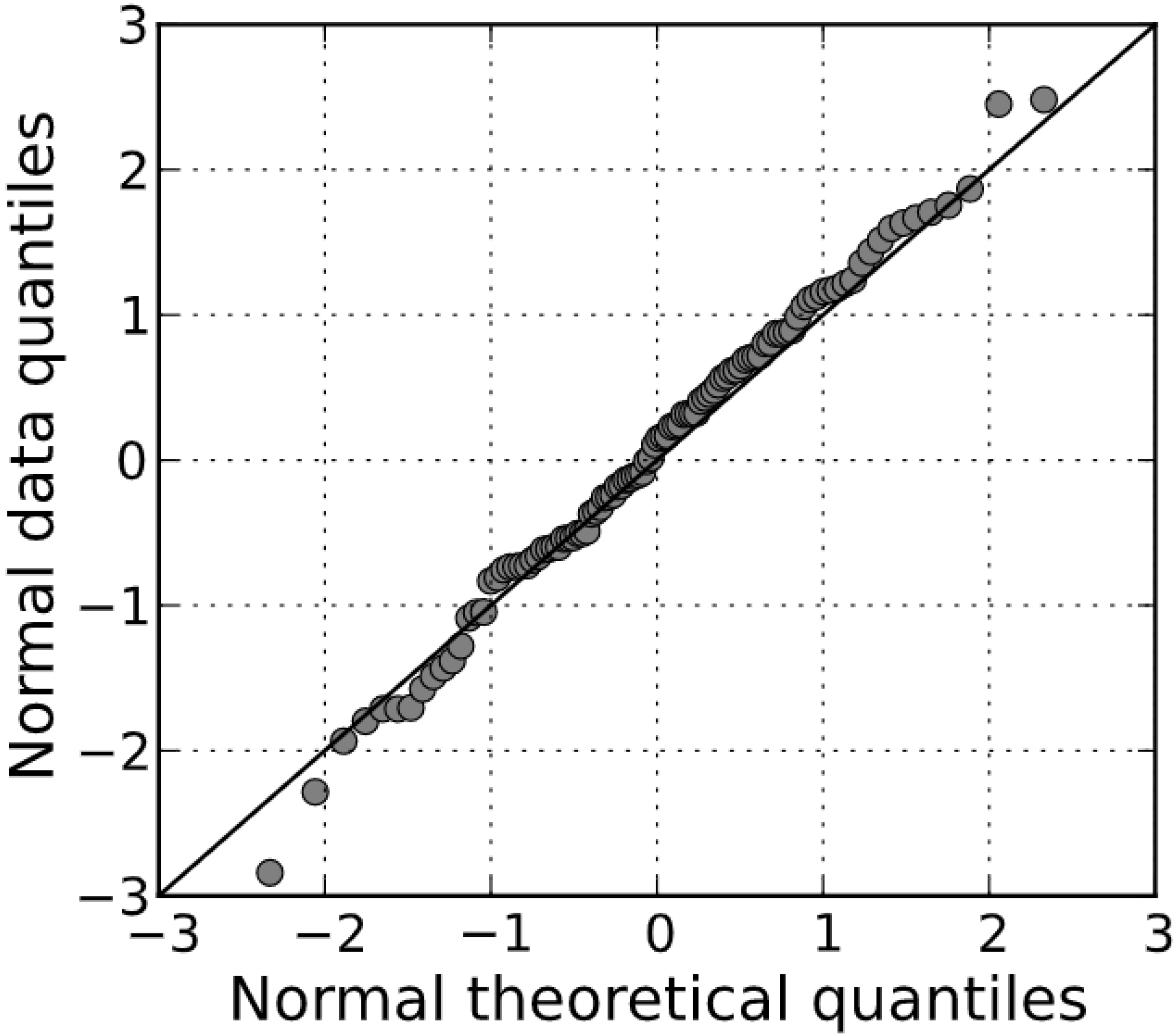
80th percentile is the value below which 80% of the data is found

For someone as tall as the 80th percentile, 80% of people are shorter

Quantile Quantile Plot QQ-Plot

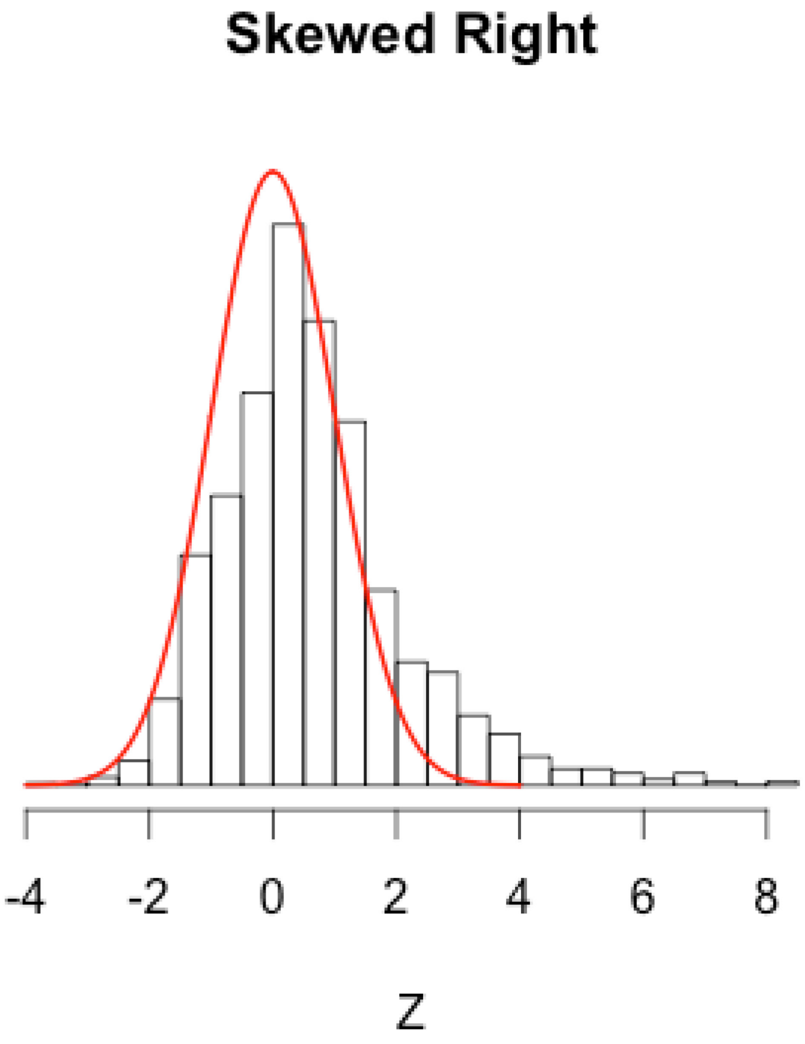
"a graphical method for comparing two probability distributions by plotting their quantiles against each other"

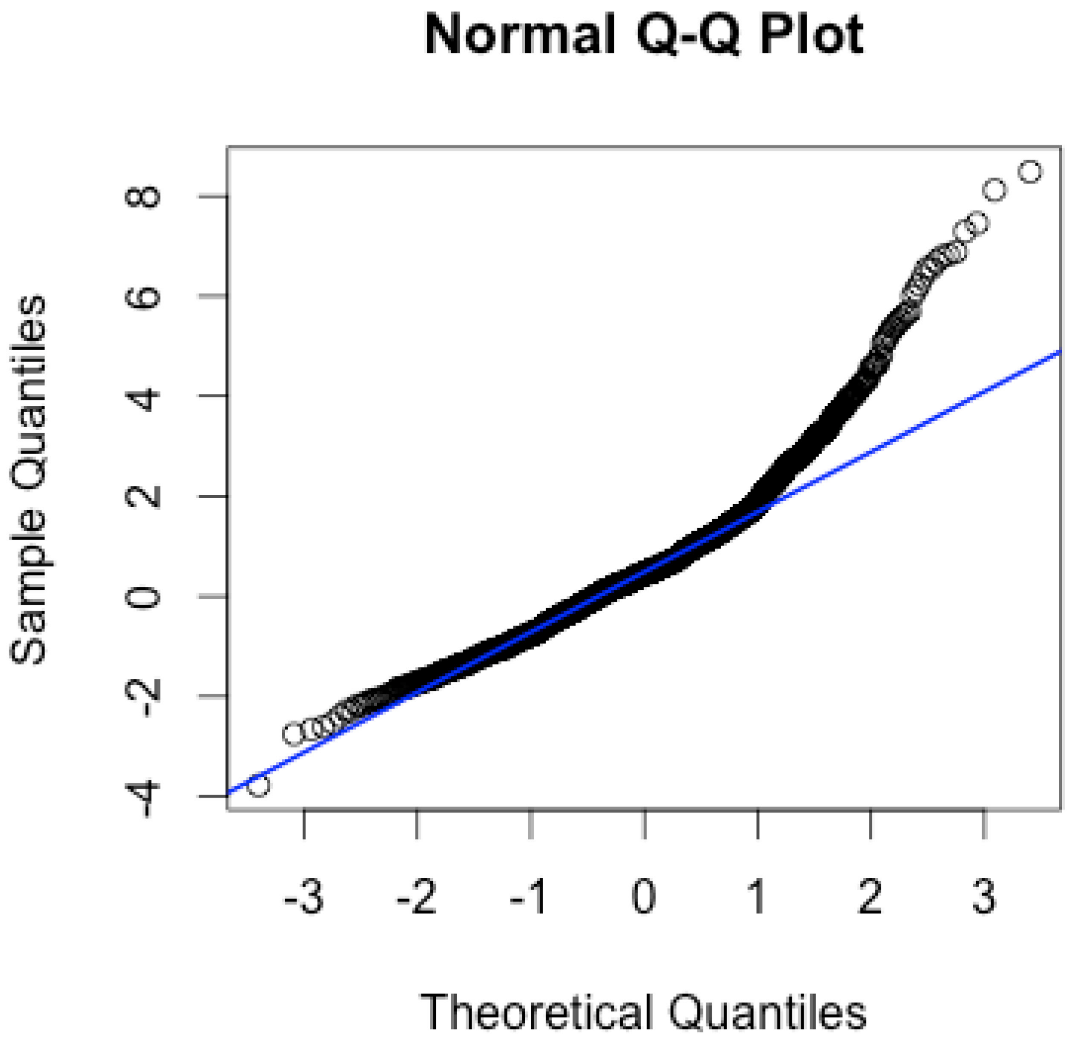
_*Wilk, M.B.; Gnanadesikan, R. (1968)*



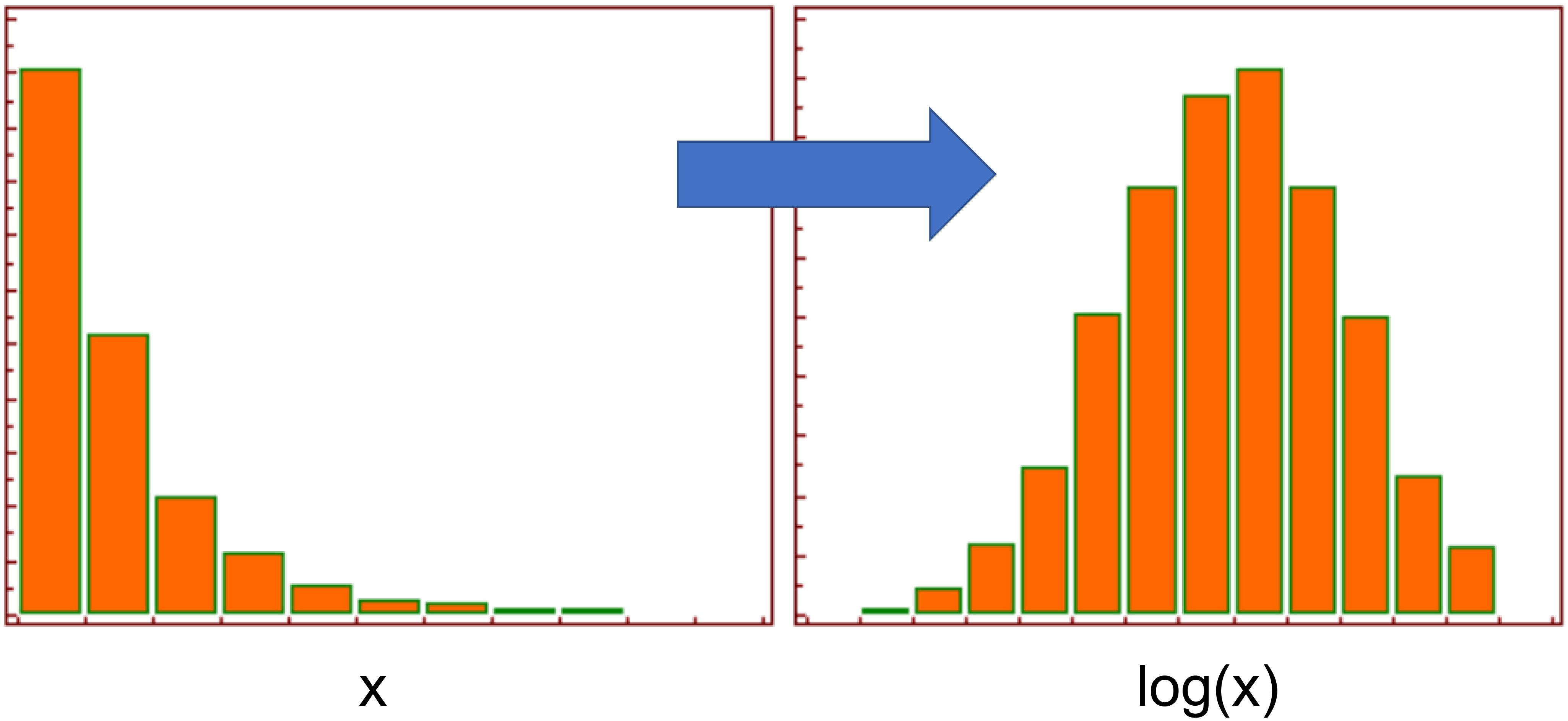
Skewed Right

https://seankross.com/2016/02/29/A-Q-Q-Plot-Dissection-Kit.html





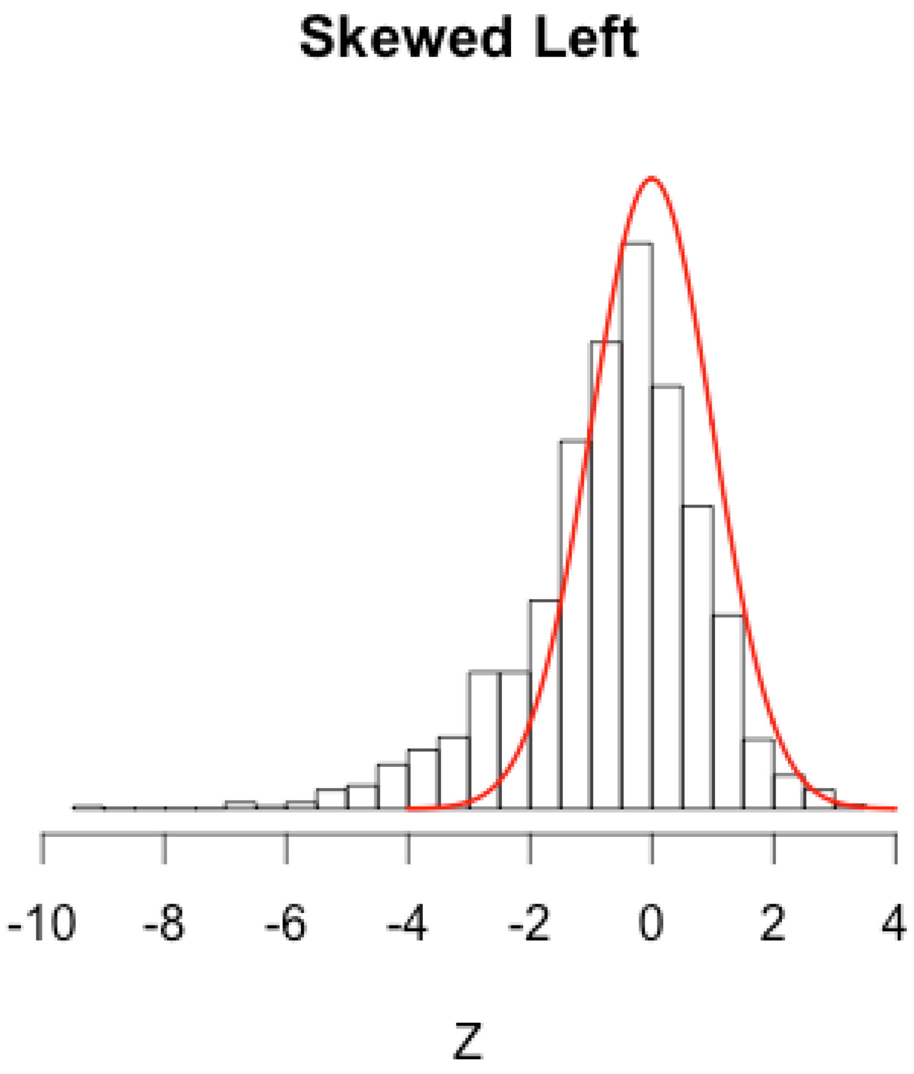
Right Skew and Log Transformation

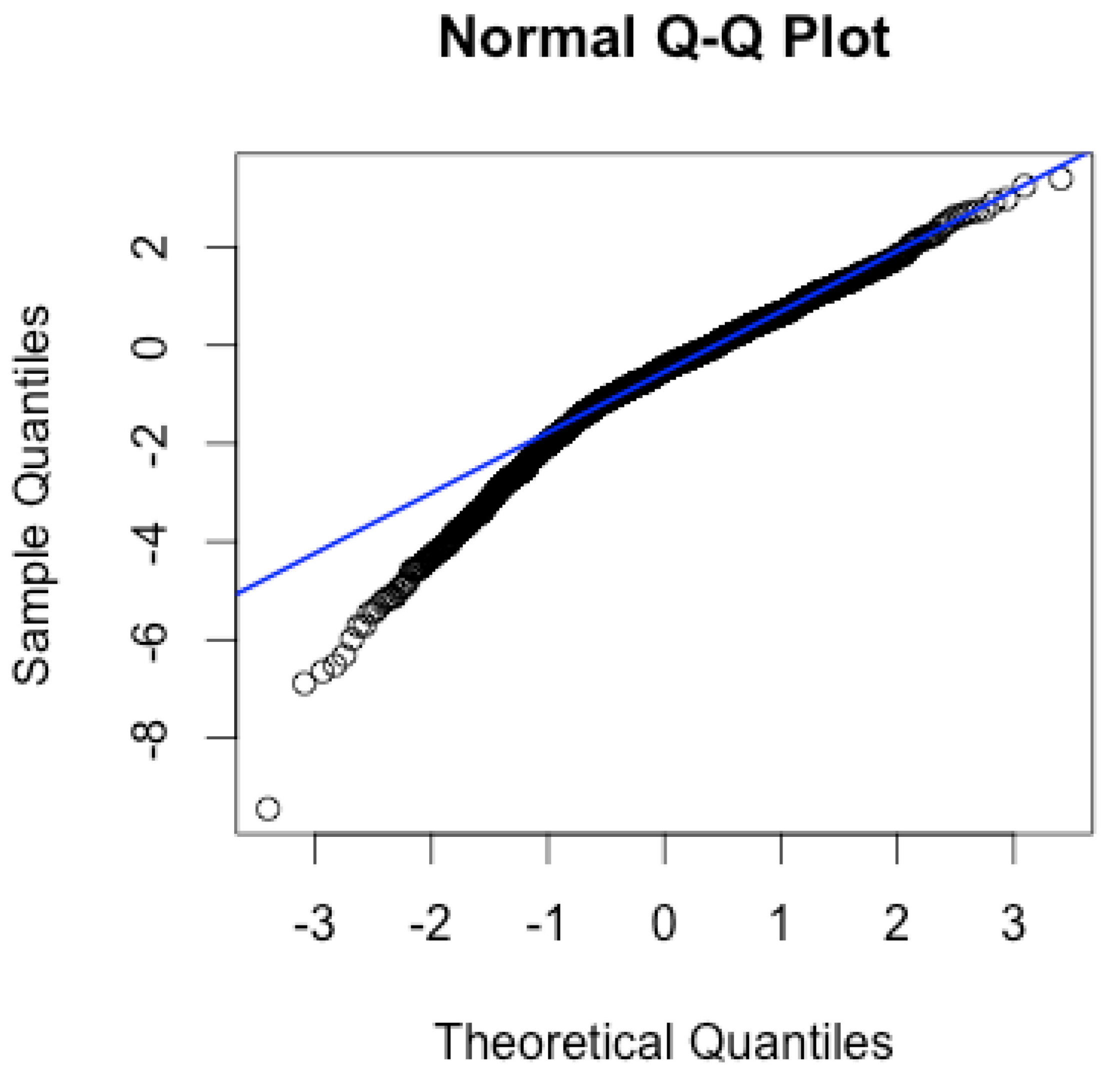


https://www.medcalc.org/manual/log-transformation.php

Skewed Left

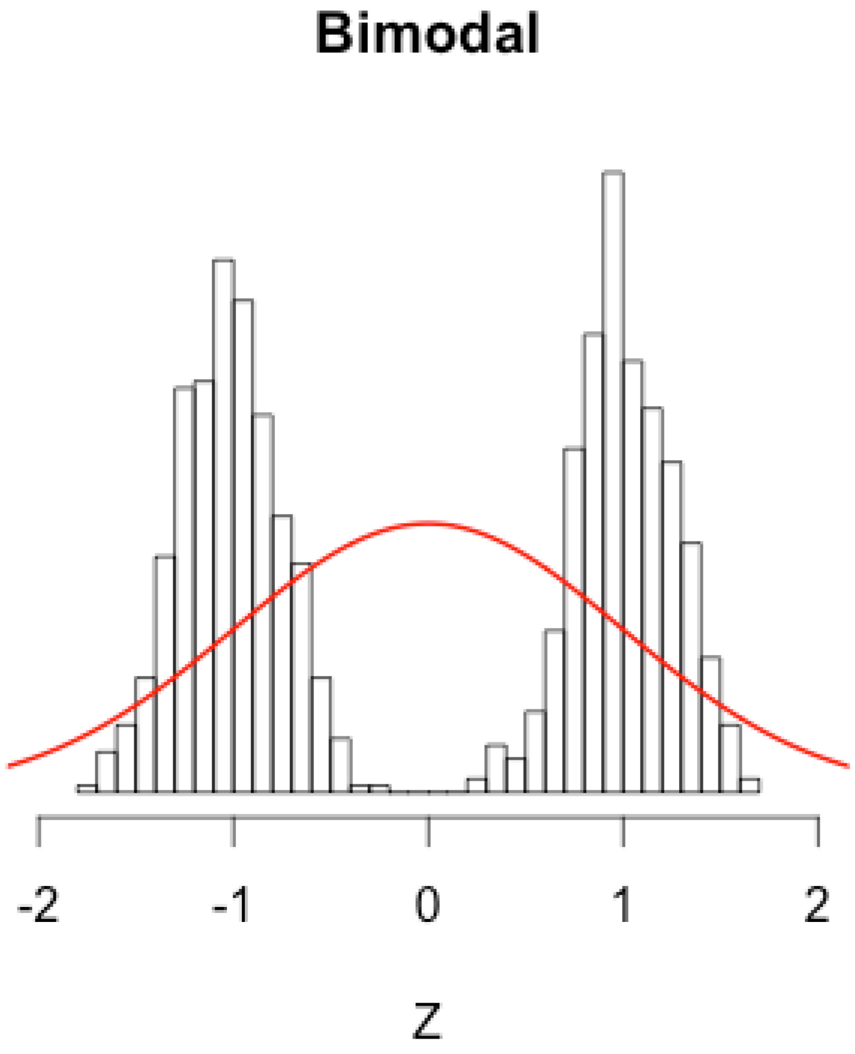
https://seankross.com/2016/02/29/A-Q-Q-Plot-Dissection-Kit.html

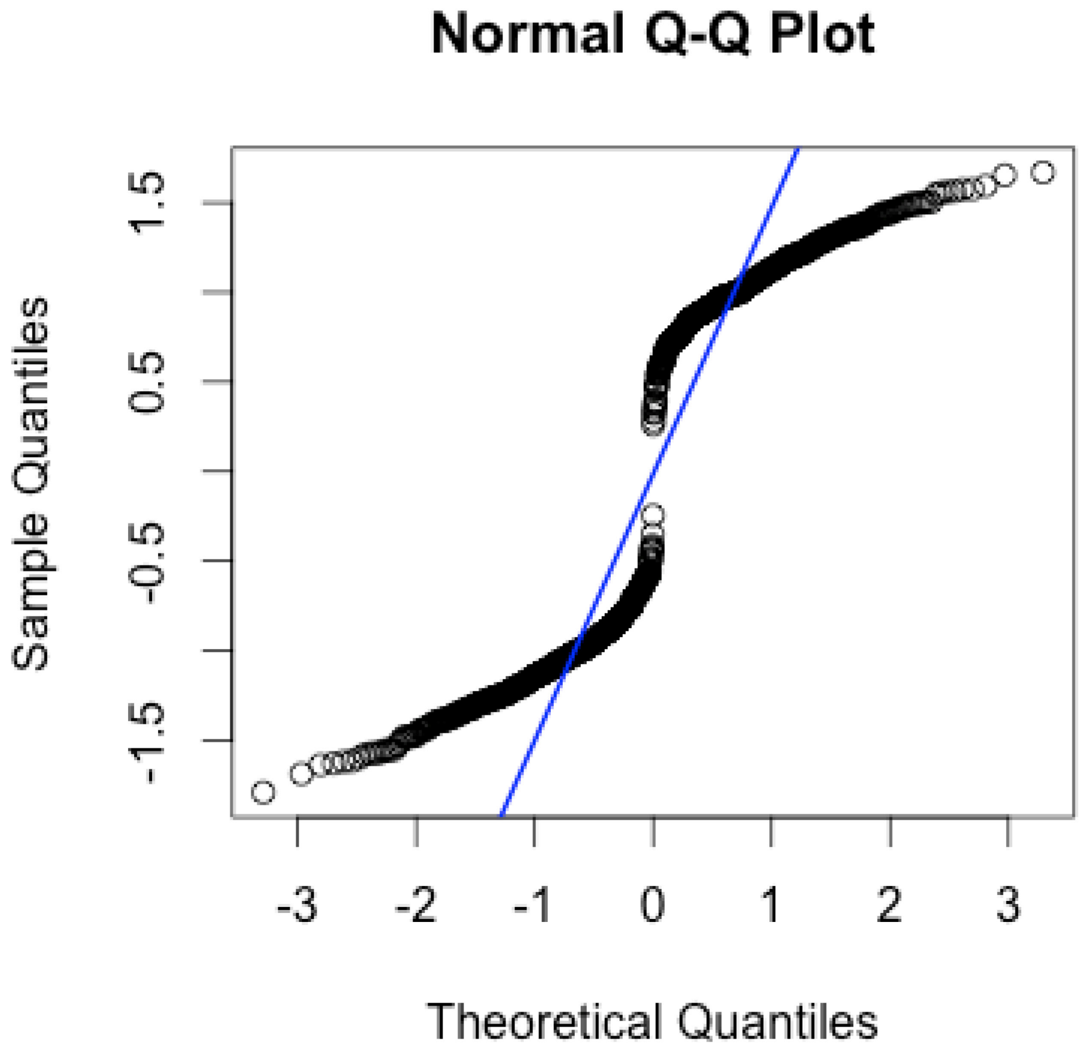




Bimodal

https://seankross.com/2016/02/29/A-Q-Q-Plot-Dissection-Kit.html



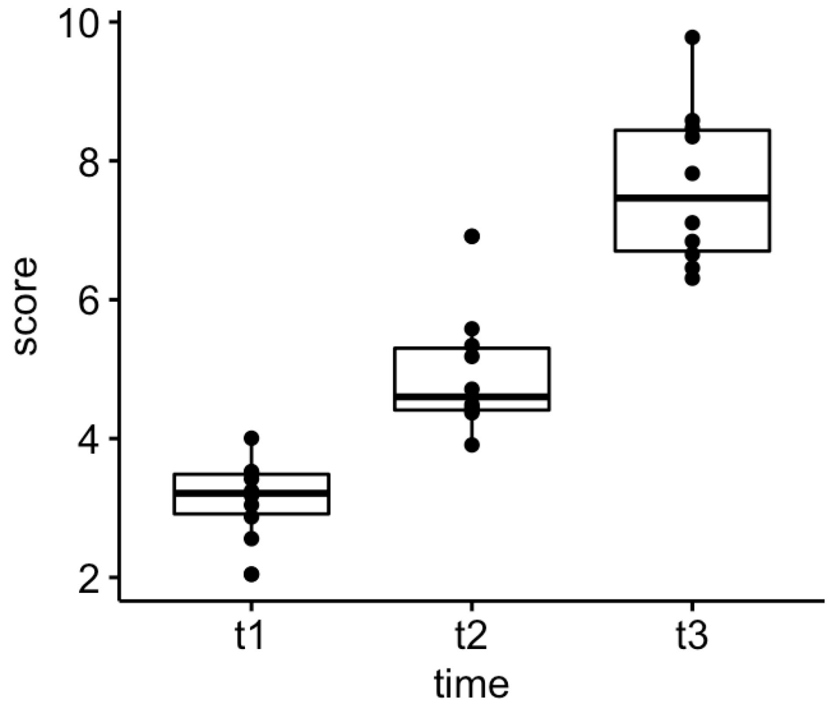


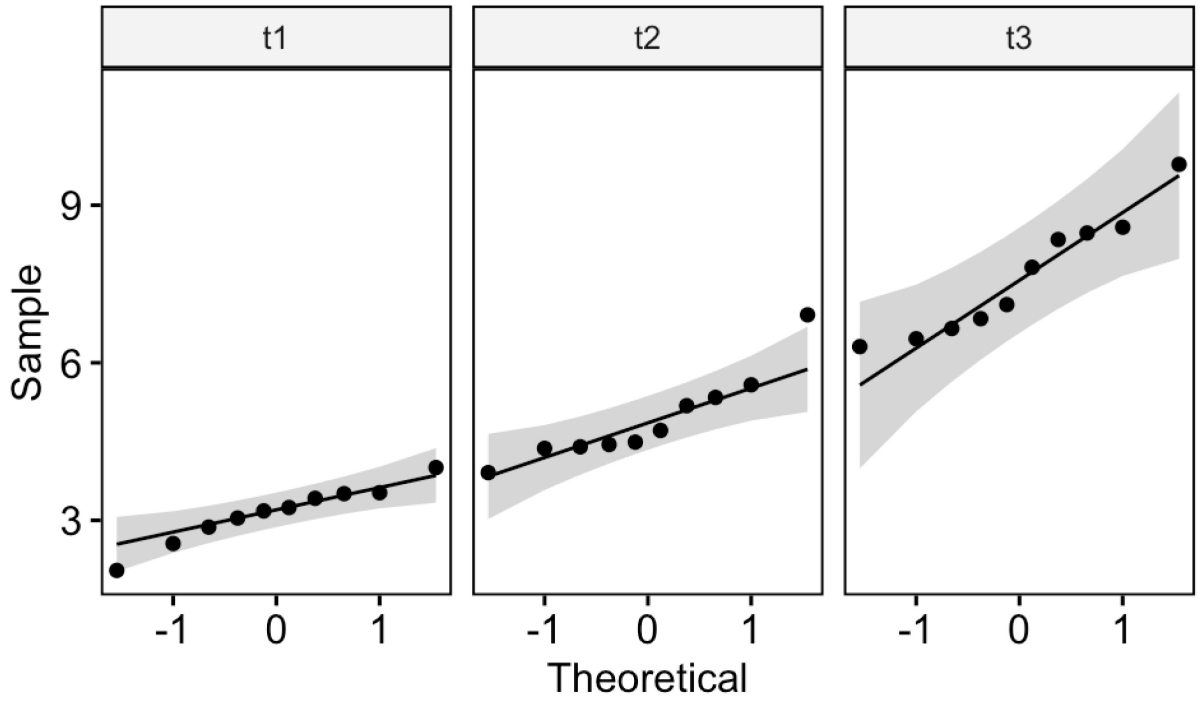
Normality Tests – Each Group Separately

When our research question is about comparing two groups

We need to test each group separately

https://www.datanovia.com/





Normality Tests

Shapiro-Wilk

`R function shapiro.test()`

Ho: data is Normally distributed

If p-value > 0.05 we can not reject Ho

Kolmogorov-Smirnov

`R function ks.test()`

One-sample test to compare to Normal or a two-samples test

Ho: two samples were drawn from the same distribution

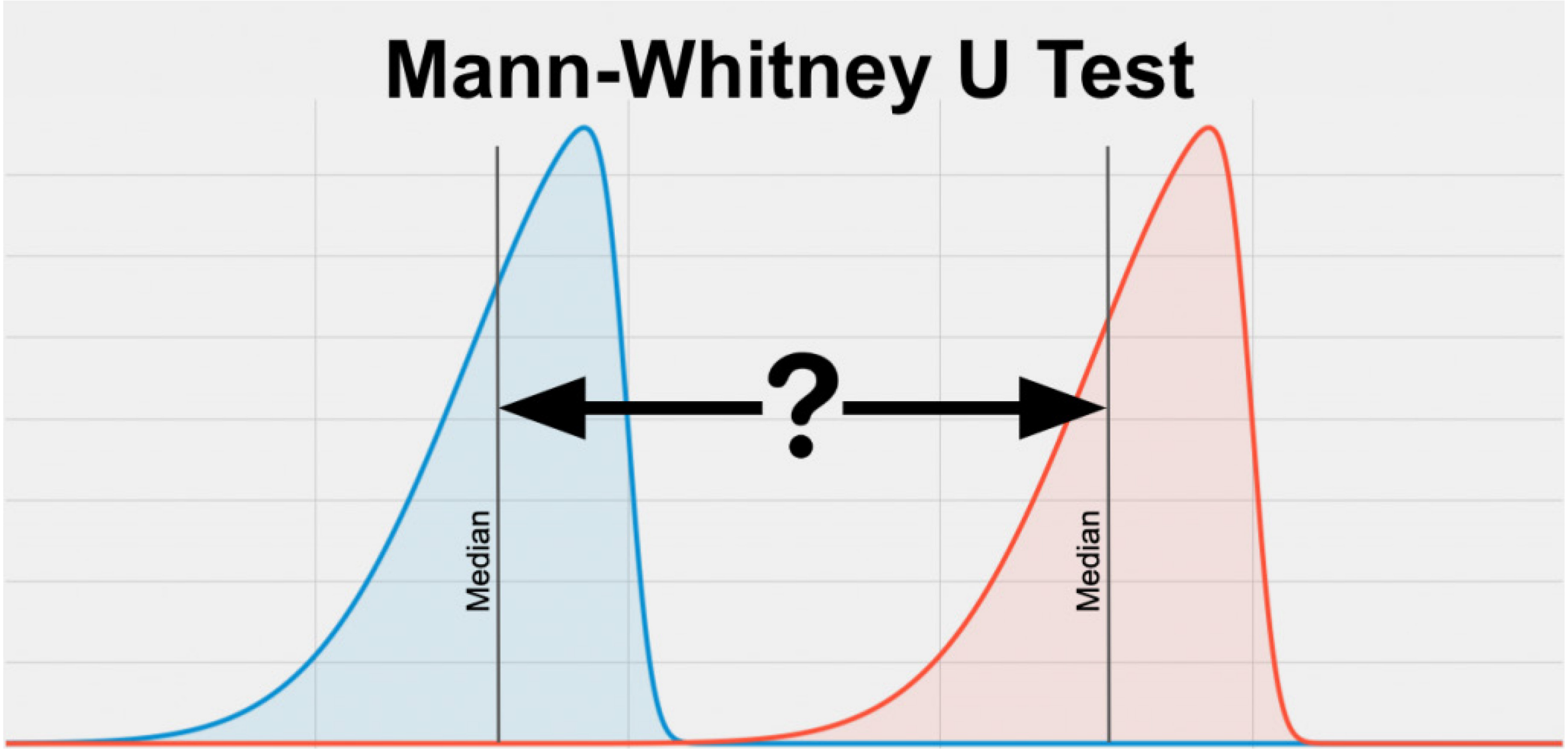
```
x <- rnorm(50) ; y <- runif(30)
# Do x and y come from the same distribution?
ks.test(x, y)
```

https://www.statology.org/test-for-normality-in-r/

Comparing Independent non-Normal Distributions

Mann–Whitney U Test

https://www.statstest.com/mann-whitney-u-test/

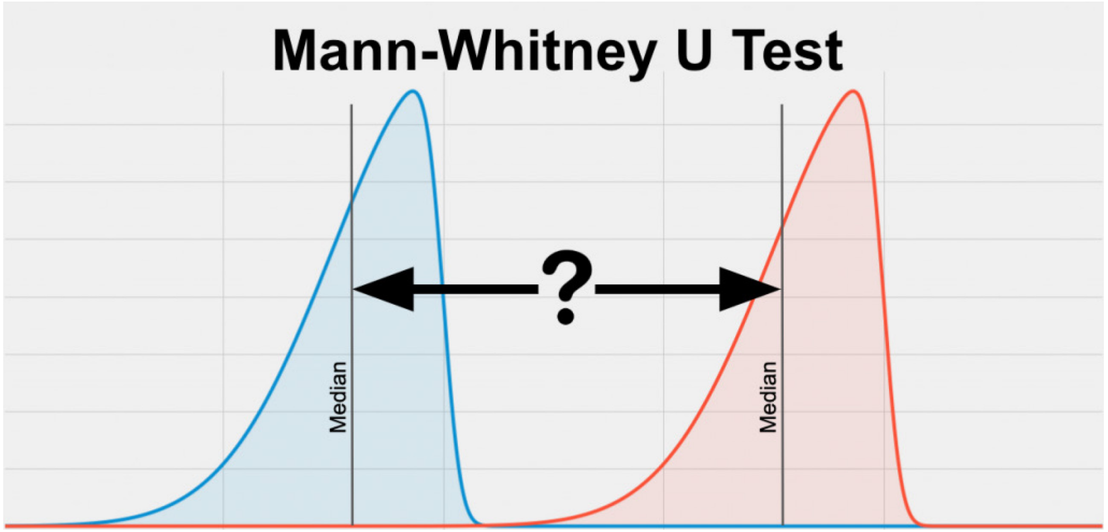


Given two identically shaped and scaled distributions,

Ho: are the medians different?

Mann–Whitney U Test

https://www.statstest.com/mann-whitney-u-test/



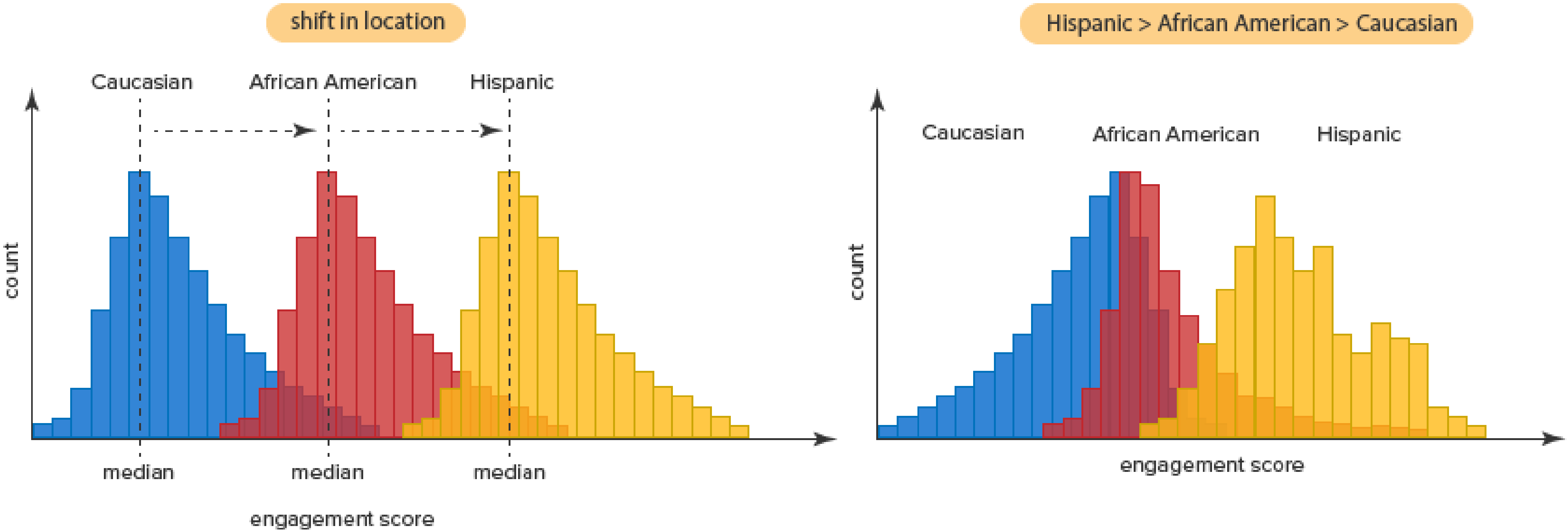
Wilcoxon 1945

Mann & Whitney 1947

Mann–Whitney–Wilcoxon

R function `wilcox.test(y~x, paired = FALSE)`

Kruskal Wallis Test: More than Two Groups



https://statistics.laerd.com/spss-tutorials/kruskal-wallis-h-test-using-spss-statistics.php

Kruskal Wallis Test

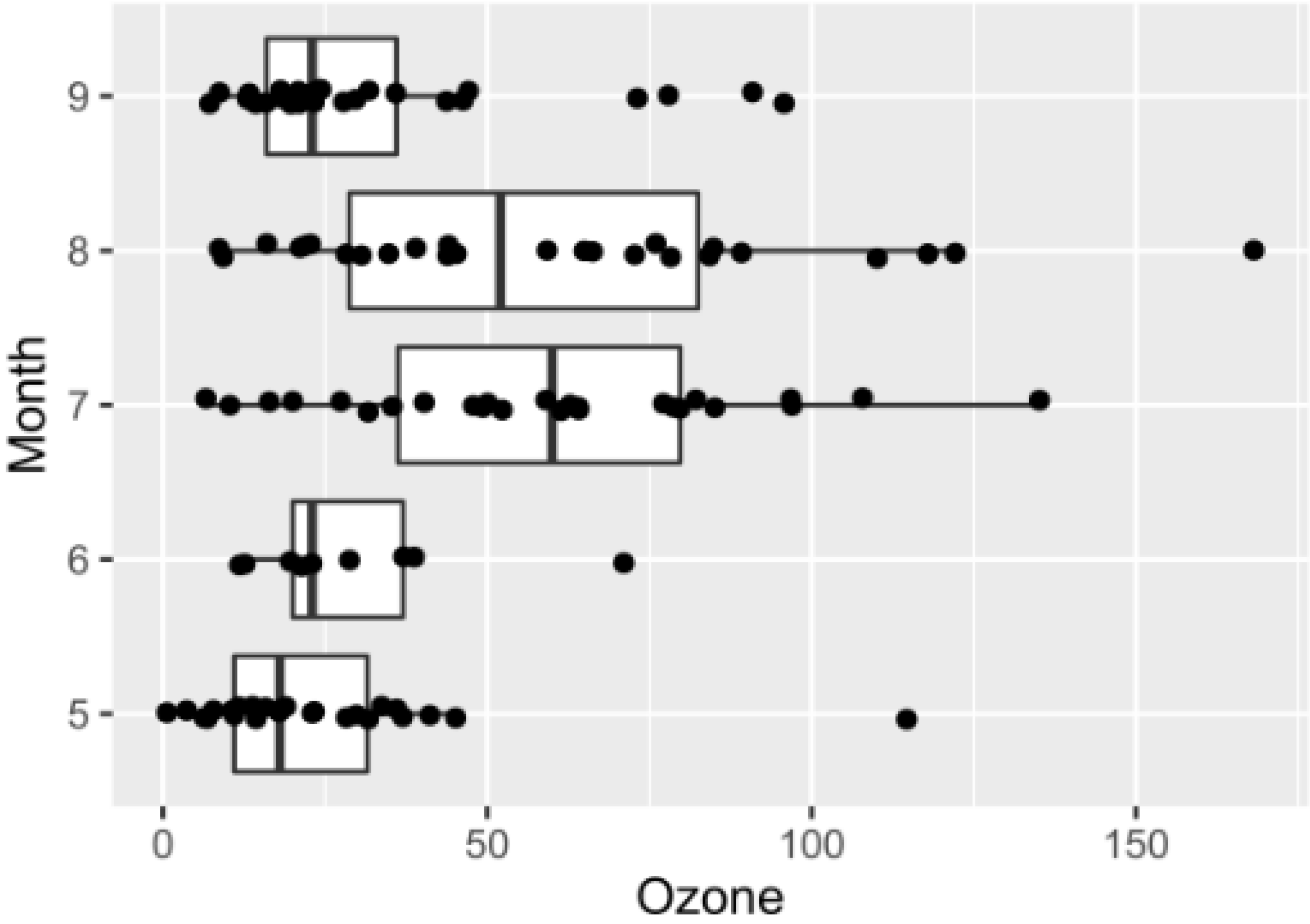
W. H. Kruskal, W. A. Wallis 1952

One-way ANOVA on ranks

Similar assumption as Mann-Whitney U

For identically shaped and scaled distribution for all groups,
Ho: are the medians different?

Kruskal Wallis Test



R function

```
kruskal.test()
```

https://en.wikipedia.org/wiki/Kruskal%E2%80%93Wallis_one-way_analysis_of_variance

Multiple Comparisons Corrections

The more comparisons we make,
the higher the chances of rejecting Ho

Multiple Comparisons Corrections

One comparison
Significance level = 0.05

Reduce the level of significance (alfa)
for each comparison
So that the probability of getting
one wrong is the same as if only one
comparison was made

Multiple Comparisons Corrections

Goal: compare each pair of medians/means (or the one of interest) WHILE achieving an overall Type I error < 5%

`R function dunn.test()`

```
dunn.test(y~x, method = "bonferroni")
```

Performs a Kruskal Wallis test with Bonferroni correction

Multiple Comparisons Corrections

For Non-Normal (independent) groups comparisons

```
dunn.test(y ~ x, method = "bonferroni")
pairwise.wilcox.test(y,x, p.adjust.method = "bonferroni“)
pairwise_wilcox_test(y,x, p.adjust.method = "bonferroni“)
```

For Normal (independent) group comparisons

```
stats::pairwise.t.test(y ~ x, p.adjust.method = "bonferroni“)
rstatix::pairwise_t_test(y ~ x, p.adjust.method = "bonferroni“)
```

https://rpkgs.datanovia.com/rstatix/

Reasons Not to Correct p-values

- The sample size (n) calculation depends on both type I (alfa) and type II (beta) errors
- So if we decrease alfa and don't increase n, then beta error increases (to keep the equation)
- Result: We loose statistical power
- We would need to increase the sample size to keep the same power
- This is a problem when sample size is small (occurs often)

- Solution: focus on fewer comparisons

Summary

- Skewness, kurtosis, modality, zero-inflation
- Percentiles (Ghostbusters! ☺) and ecdf
- QQplots and Normality tests (Shapiro-Wilk, Kolmogorov-Smirnov)
- Mann Whitney U test to compare 2 groups (`wilcox.test()`)

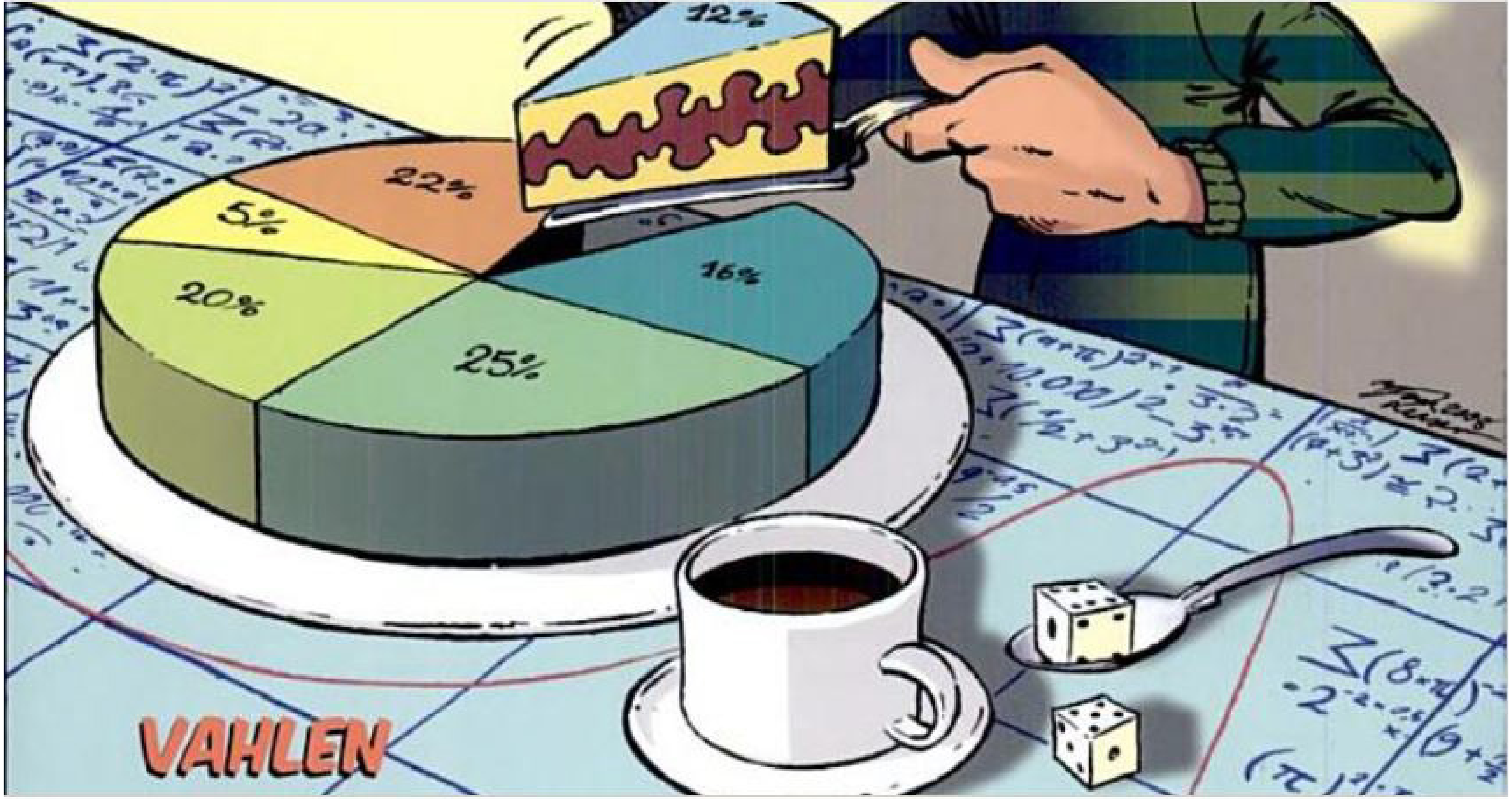
- Tests that include multiple comparison corrections (pairwise comparisons with Bonferroni correction)

Table to Choose the Right Test

1.Measurement scale and distribution of dependent (outcome) variable (-> COLUMNS)
2.Study objective related to type and scale of independent (often group) variable (->ROWS)
3.Independent or dependent observations between subjects?

	Type of dependent variable ("outcome" or y-variable)			
Study objective	**Interval data (normally distributed measurements)**	**Ranks, Scores, or non-normally distributed measurements**	**Binary outcome (two levels)**	Survival Time
Describe a single study group (summaries, frequencies)	**Mean**, median, mode, **SD**, SEM, percentiles	**Median, range,** interquartile range, percentiles	Freq., Proportion (Prevalence) + 95% CI	Kaplan Meier survival curve
Compare a single study group to a fixed (population) value	Single sample *t* **test**	Wilcoxon **Signed-Rank test**	Single sample **proportion test**	
Compare two independent study groups	Independent sample *t* **test**	Two-sample **Wilcoxon** (Rank Sum) test	Two-sample proportions test (Chisquare, FET)	Log-rank test or Mantel-Haenszel
Compare two paired (dependent) sample study groups	Paired sample *t* **test**	Paired sample **Wilcoxon** (Signed-Rank) test	**McNemar's test,** Kappa statistic	Conditional proportional hazards regression
Compare three or more unmatched (independ.) study groups	One-way **ANOVA**	**Kruskal-Wallis** Test	Proportions test (Chisquare) Logistic Regression	Cox proportional hazard regression
Compare three or more matched (dependent) groups	Multiway **ANOVA**	**Friedman test**	Cochrane Q	Conditional proportional hazards regression
Quantify correlation between two variables	Pearson **correlation**	Spearman Rank **correlation**	Contingency coefficients	
Predict outcome value from another interval or categorical variable	Generalized linear Models: Simple **Linear regression** or **ANOVA**	Nonparametric regression	Cross tabulation (Odds ratio), **Simple logistic regression**	Cox proportional hazard regression
Predict outcome value from several measured (interval), categorical or binomial variables	Generalized linear Models: Multivariable **Linear regression** or **ANOVA**	Generalized linear models accommodating nonparametric components	Stratified cross tables, **multiple logistic regression**	Cox proportional hazard regression

Questions



Thanks for your attention

b
UNIVERSITÄT
BERN

VETERINARY PUBLIC HEALTH INSTITUTE

b
UNIVERSITÄT BERN

Correlation Coefficients

Association Continuous Variables

Dr. Beatriz Vidondo
Veterinary Public Health Institute UniBe

Objectives

Understand correlation coefficients and know when to apply which type of correlation coefficient

Take appropriate choices when deleting/excluding missing values to calculate correlation

Interpret correlation correctly

Correlation Coefficient Definition

numerical measure of some type of association, meaning a statistical relationship between two variables

Do they vary “together”?

[Dancing statistics: explaining the statistical concept of correlation through dance – YouTube](https://www.youtube.com/watch?v=VFjaBh12C6s)

https://www.youtube.com/watch?v=VFjaBh12C6s

Correlation Coefficient Definition

numerical measure of some type of association, meaning a statistical relationship between two variables

Variance of one variable: $\mathrm{Var}(\mathrm{Y}) = \sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2$

Correlation Coefficient Definition

numerical measure of some type of association, meaning a statistical relationship between two variables

Variance of one variable: $\mathrm{Var}(\mathrm{Y}) = \sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2$

Covariance between 2 vars: $Cov\,(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$

Correlation Coefficient Definition

numerical measure of some type of association,
or a statistical relationship between two variables

Variance of one variable: $\mathrm{Var(Y)} = \sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2$

Covariance between 2 vars: $Cov\,(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$

Correlation Coefficient (Pearson): $r_p = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$

Different Correlation Coefficients

Pearson - a measure of the strength and direction of the *linear* relationship between two variables

Spearman - a measure of how well the relationship between two variables can be described by a monotonic function

Kendall - measure of the portion of ranks that match between two variables

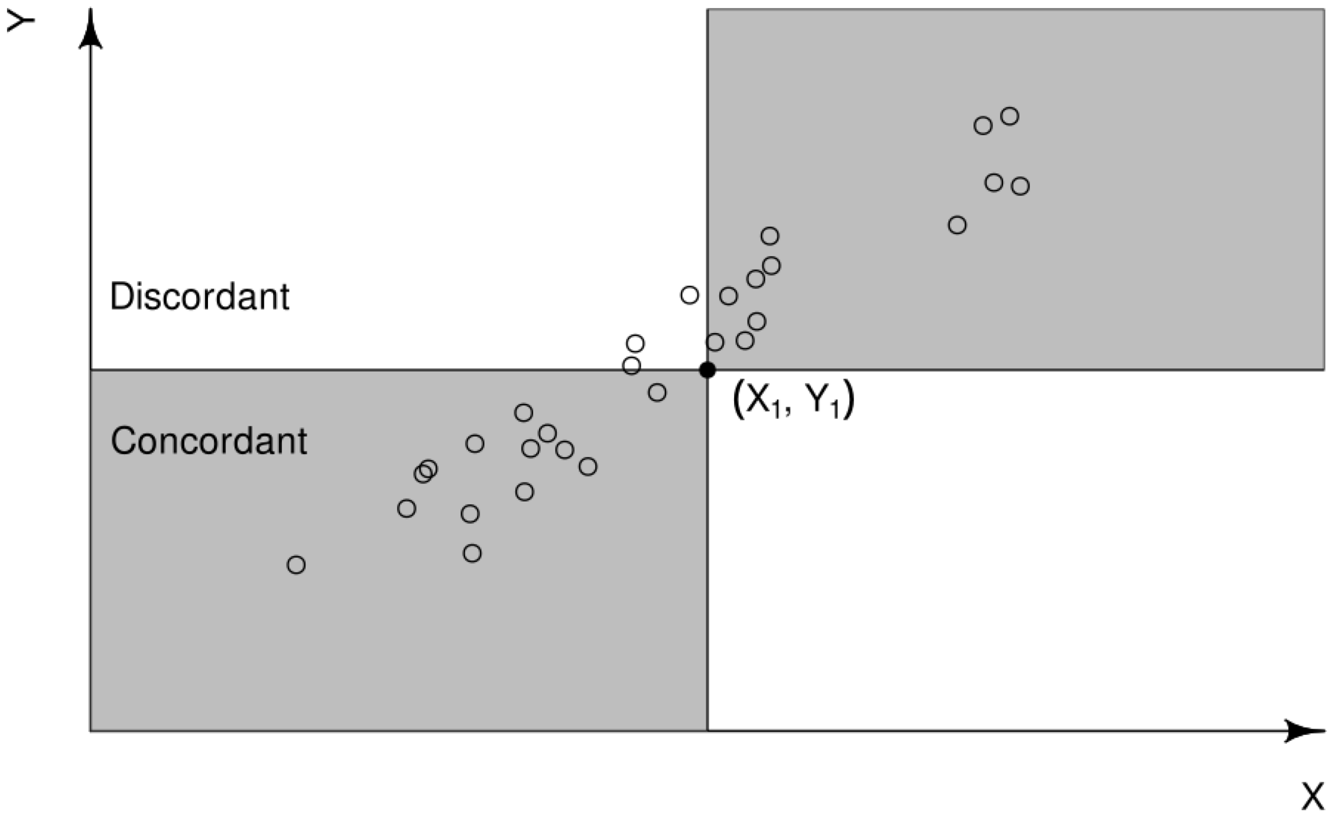
Pearson vs. Spearman

Spearman correlation coefficient uses the ranks R of the variables instead of the variables

$$r_p = \frac{cov(X,Y)}{\sigma_X \sigma_Y}; \qquad r_s = \frac{cov(R(X),R(Y))}{\sigma_{R(X)} \sigma_{R(Y)}}$$

Kendall Rank Correlation Coefficient

$$\tau_s = \frac{number\ of\ concordant\ pairs\ - number\ of\ discordant\ pairs}{total\ number\ of\ possible\ pairs}$$



395 concordant points

(grey areas)

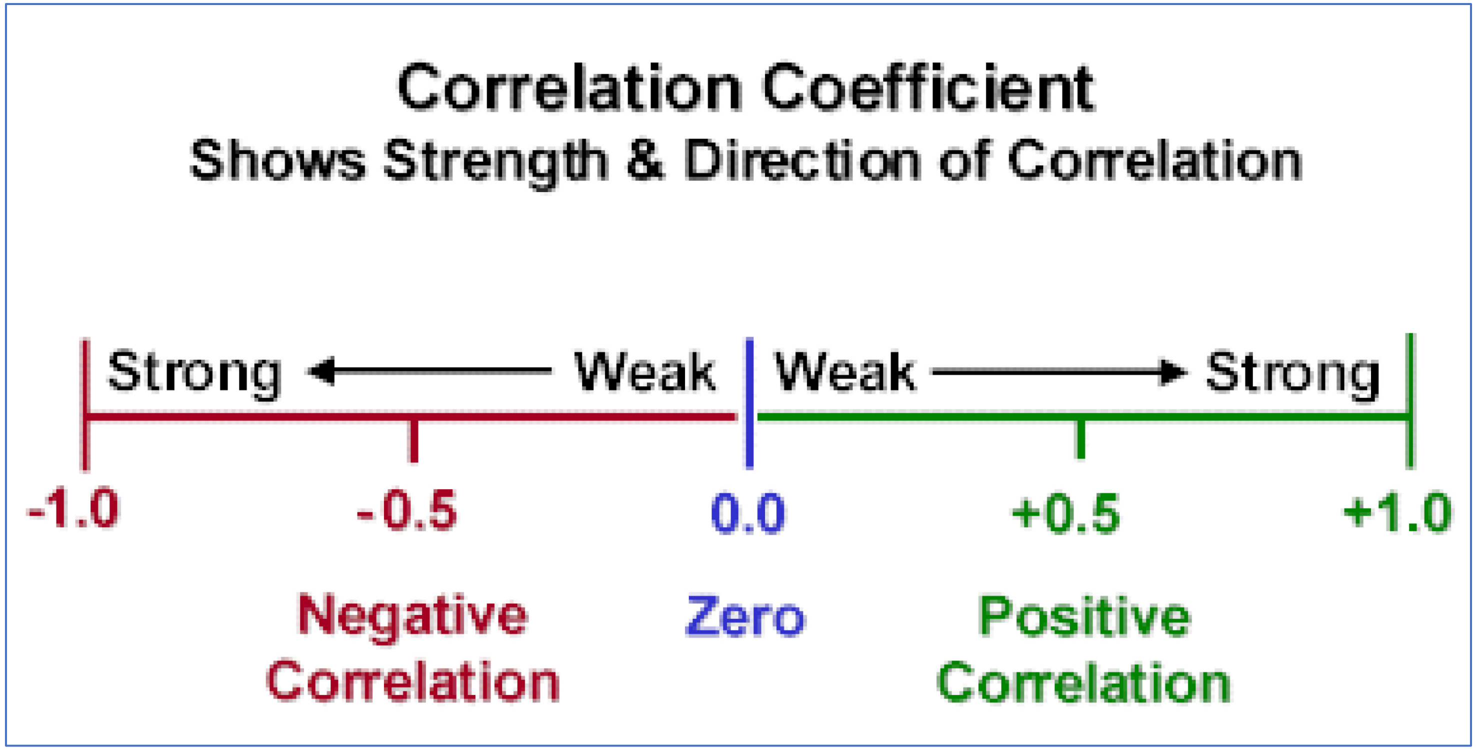
40 discordant points

(white areas)

Kendall tau = 0.816

https://en.wikipedia.org/wiki/Kendall_rank_correlation_coefficient

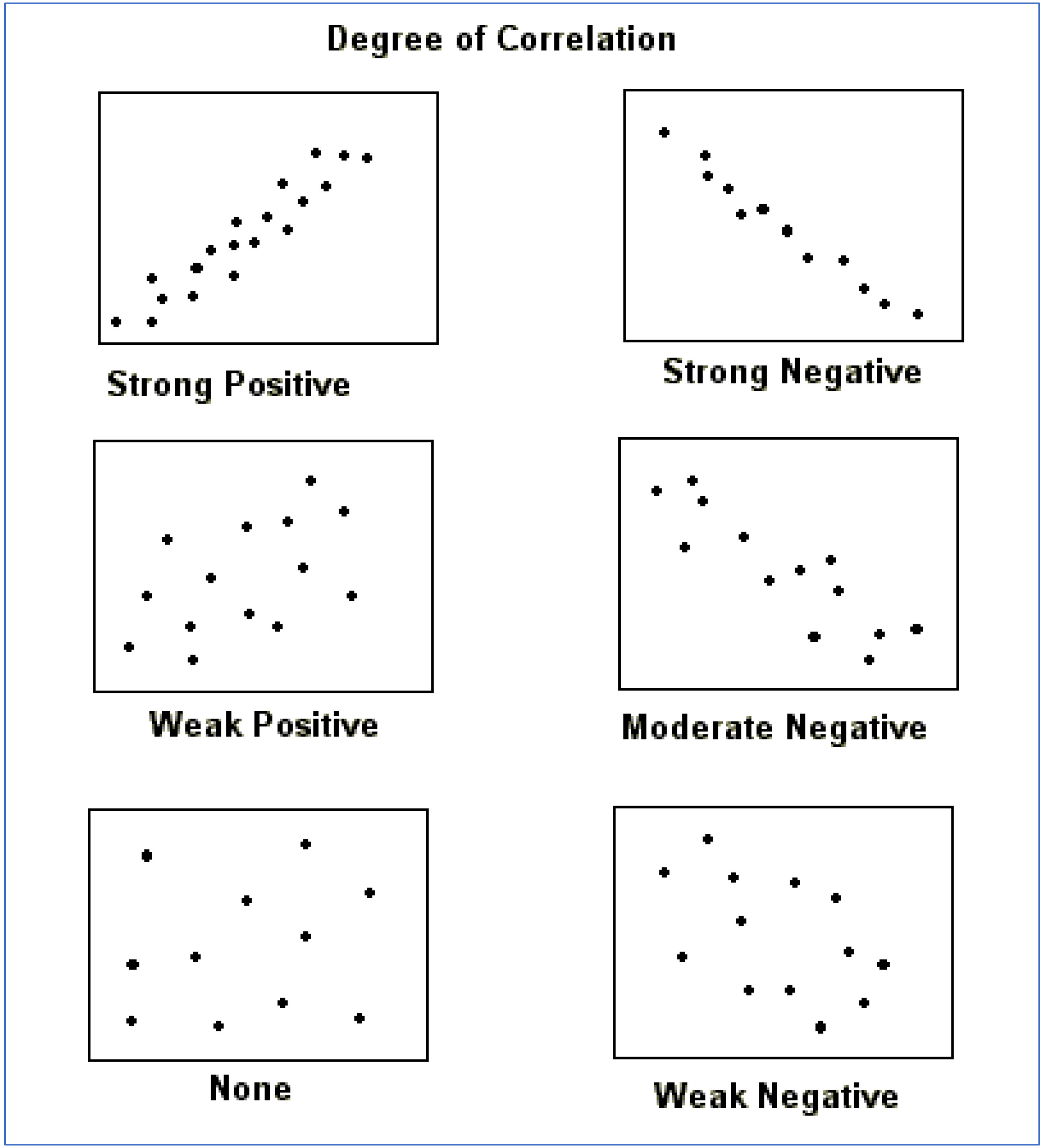
Interpretation

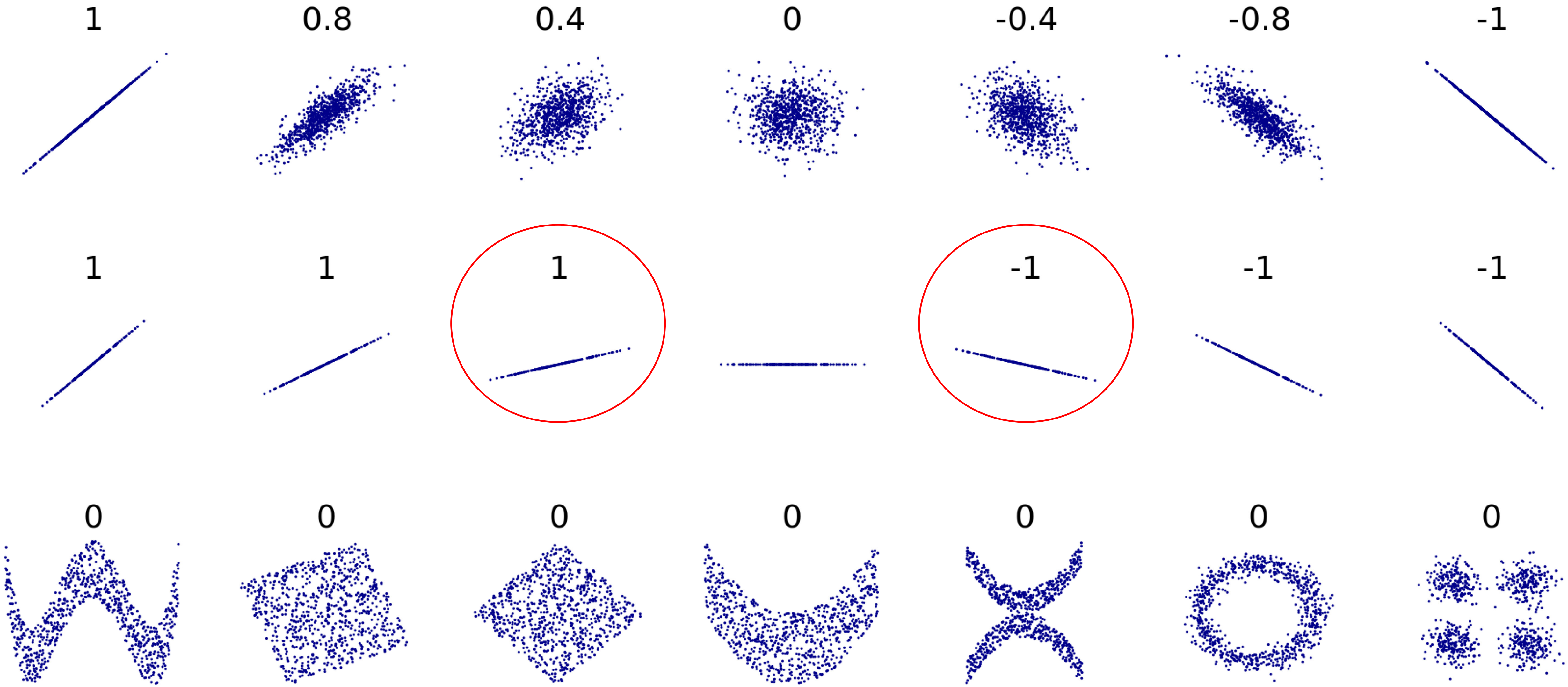


Renders values in the range from −1 to +1,
where ±1 indicates the strongest possible agreement and
0 the strongest possible disagreement

Interpretation

To be able to interpret correctly, always plot a scatterplot





1
0.8
0.4
0
-0.4
-0.8
-1
1
1
1
-1
-1
-1
0
0
0
0
0
0
0

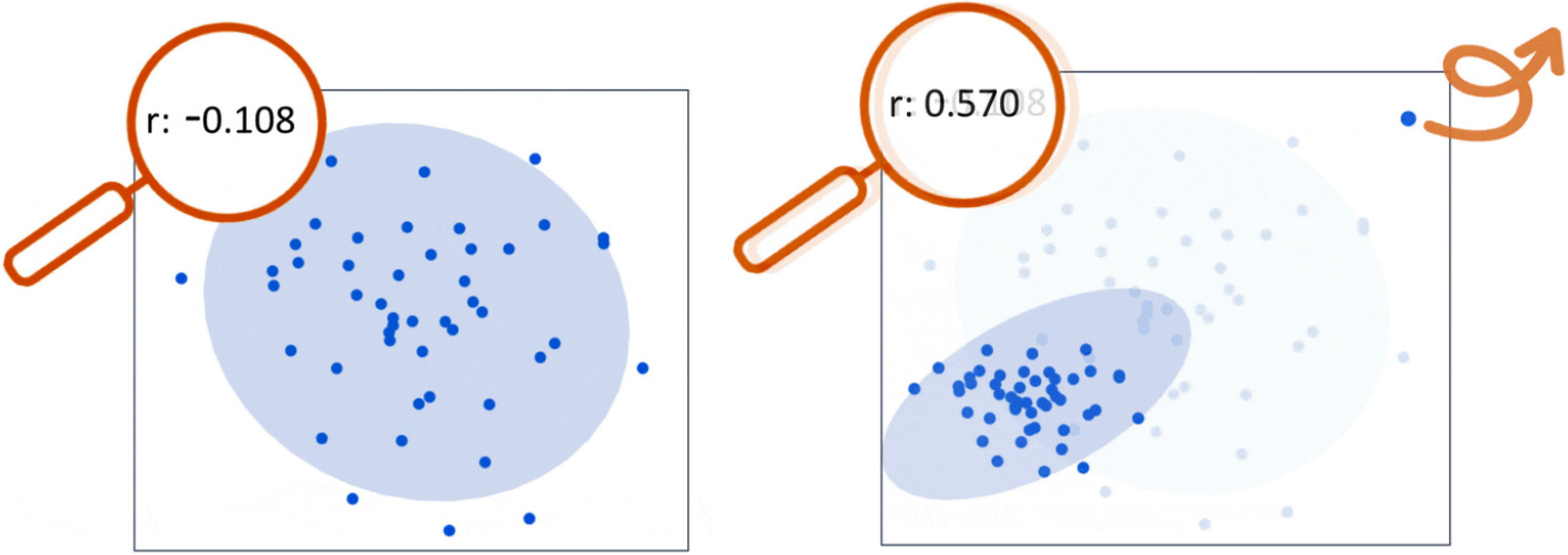
Interpreting Correlation

1. Distortion
    1. Distortion by outliers
    2. Distortion by missing values

2. Non-Linear and Non-monotonic relationships

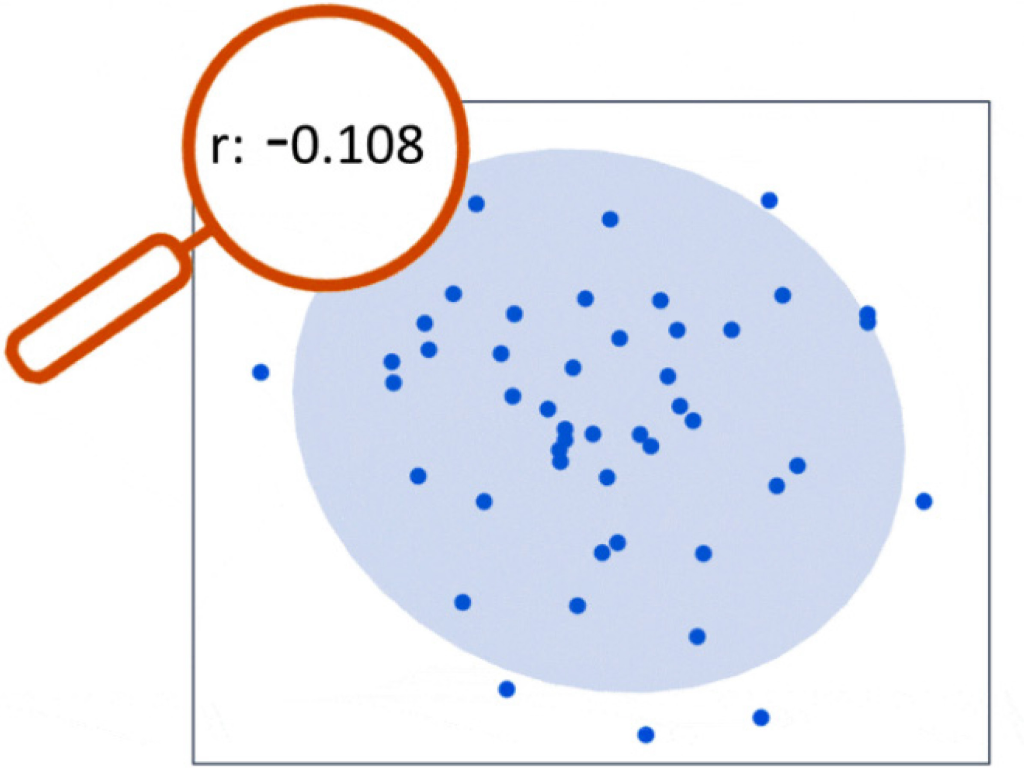
3. Correlation does not mean causation

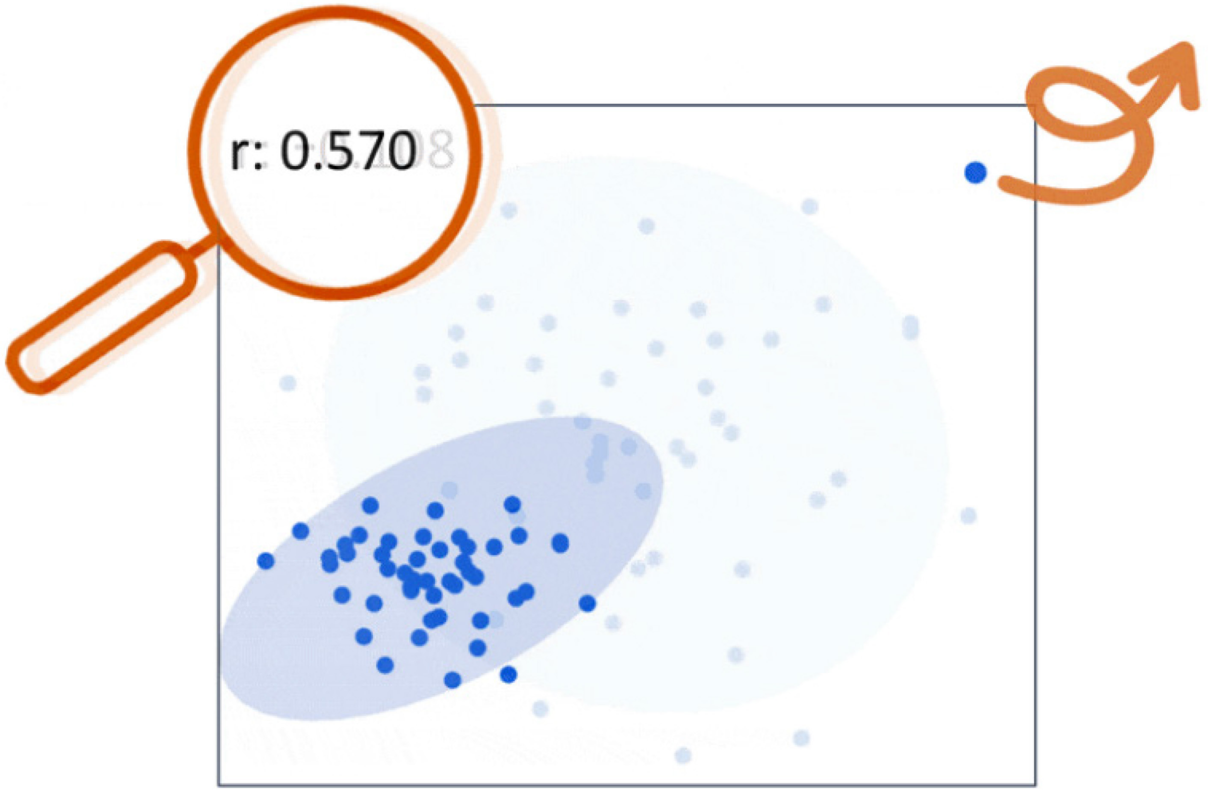
Distortion by Outliers



https://www.jmp.com/en_hk/statistics-knowledge-portal/what-is-correlation/correlation-coefficient.html

Always Calculate With and Without Outliers





https://www.jmp.com/en_hk/statistics-knowledge-portal/what-is-correlation/correlation-coefficient.html

Distortion by Missing Values

Assumption: Missing Values Completely at Random (MCAR)

Need to delete/exclude missing values NA -> otherwise ERROR

Deletion/exclusion causes bias in results

Descriptive analysis report % of missing for each variable

Consider to discard variables with % missing > 10%

Missings in the «outcome variable» need to be discarded listwise (the full row, or the full ‘case’)

Deleting Missing Values

Important when making large **Correlation Matrices**

Listwise deletion (complete-case analysis)

List = observation = row = case = patient

removes the whole row if any variable value is missing

(casewise) => fewer cases => better to delete some variables

Pairwise deletion (available-case analysis)

For any two pair of variables, delete rows with missings

makes the most of the data available, but uses a different **sample size** every time => might result in biased results

Options To Delete Missing Values

`R function cor()`

`na.rm = TRUE/FALSE` should NA's be removed?

`use = "everything", "all.obs", "complete.obs", "na.or.complete", or "pairwise.complete.obs‟`

https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/cor

Options To Delete Missing Values

- “everything” renders cor=NA whenever there are missings
- "na.or.complete" renders cor=NA if there are no complete cases

- "all.obs“ only works when no missing values, otherwise `ERROR`

- "complete.obs“ = listwise/casewise deletion (=> less data)
- "pairwise.complete.obs“ = pairwise deletion (=> diff sample sizes)

https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/cor

Option Missing Value Deletion

`R function cor()`

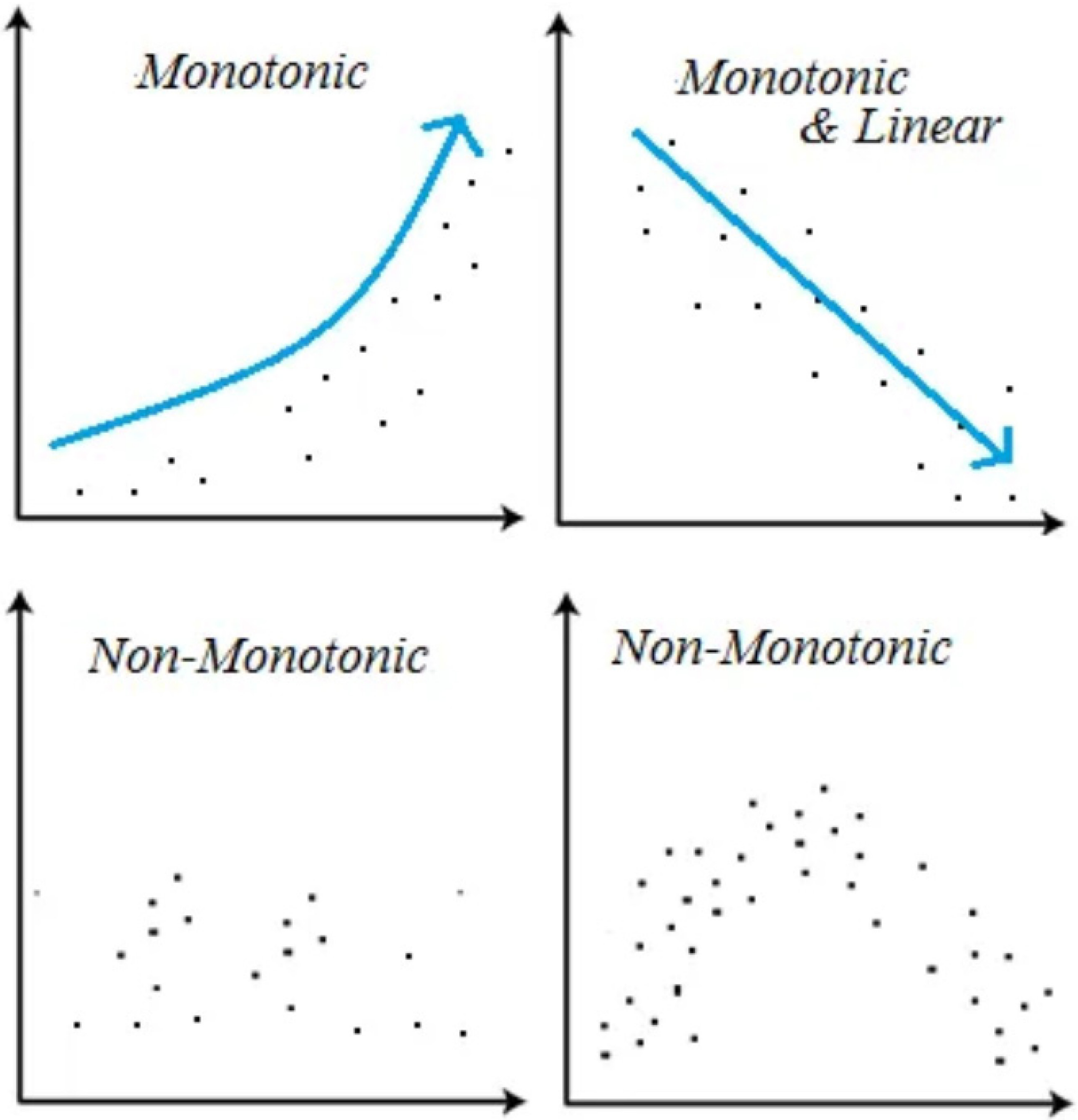
- `na.rm = TRUE/FALSE` should NA's be removed?
- `na.rm = TRUE` is equivalent to `use = "na.or.complete"`
- `na.rm = FALSE` is equivalent to `use = "everything"`

https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/cor

Pearson vs Spearman

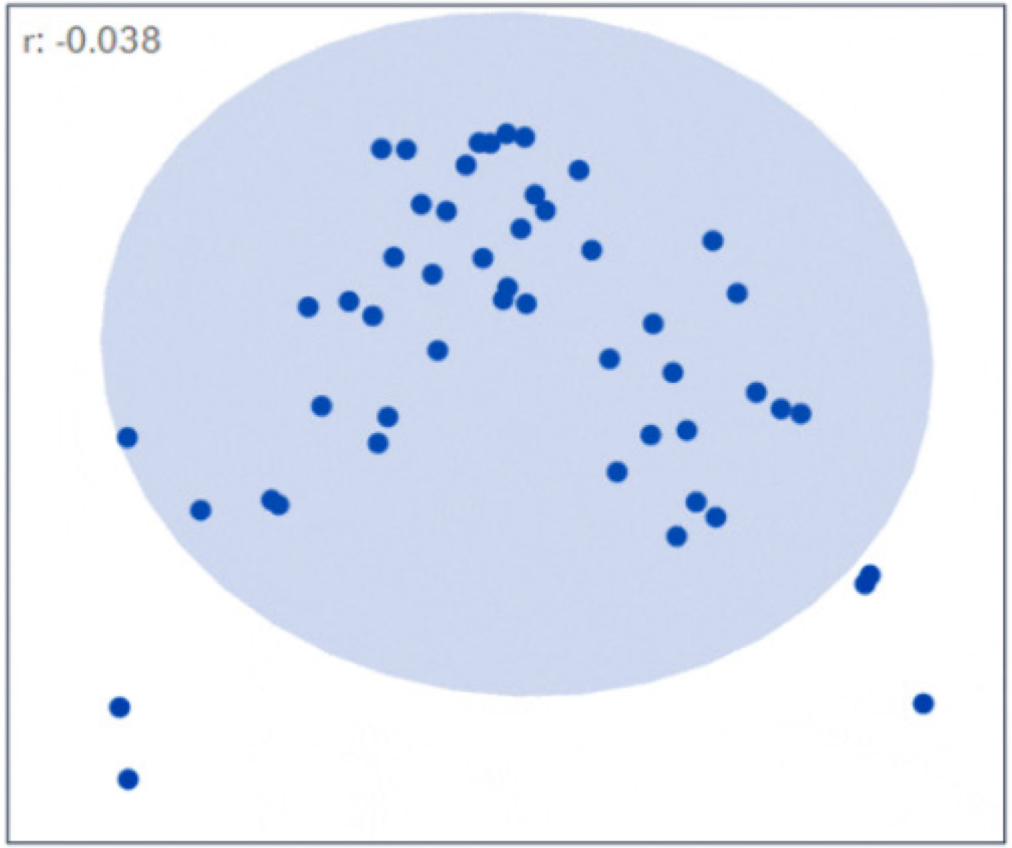
Pearson measures Linearity
Spearman measures Monotonicity

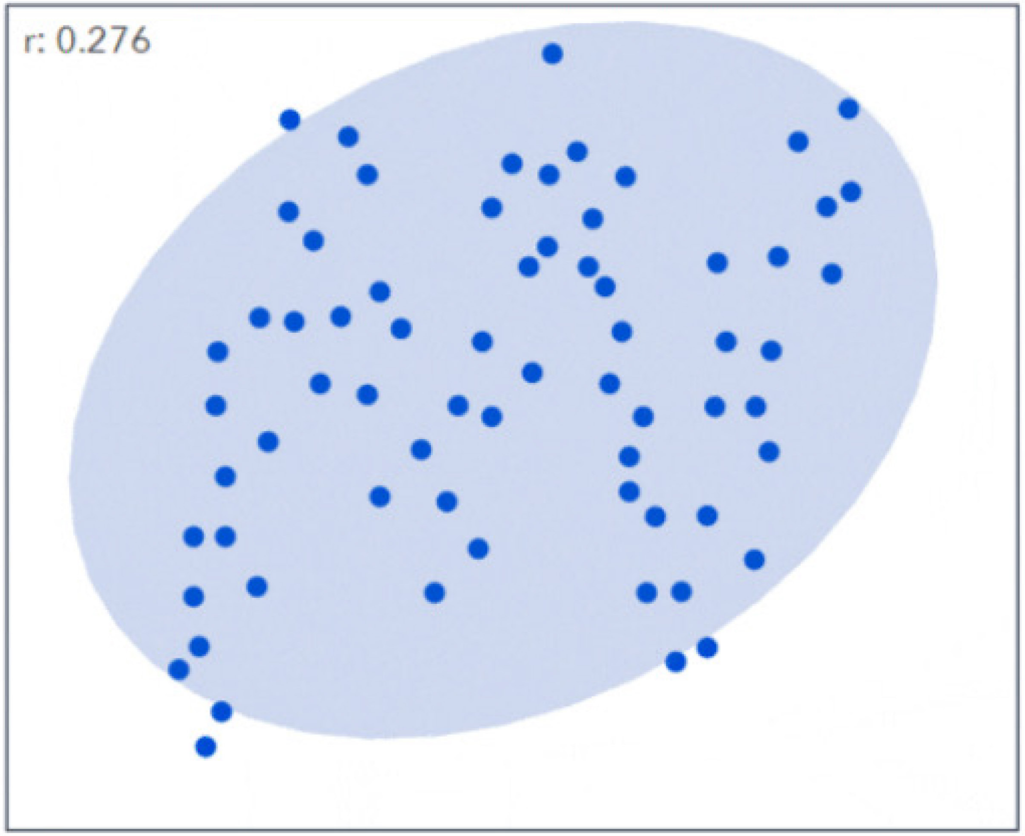
Non-Monotonic associations require other specific non-linear models to capture the pattern



https://anyi-guo.medium.com/correlation-pearson-vs-spearman-c15e581c12ce

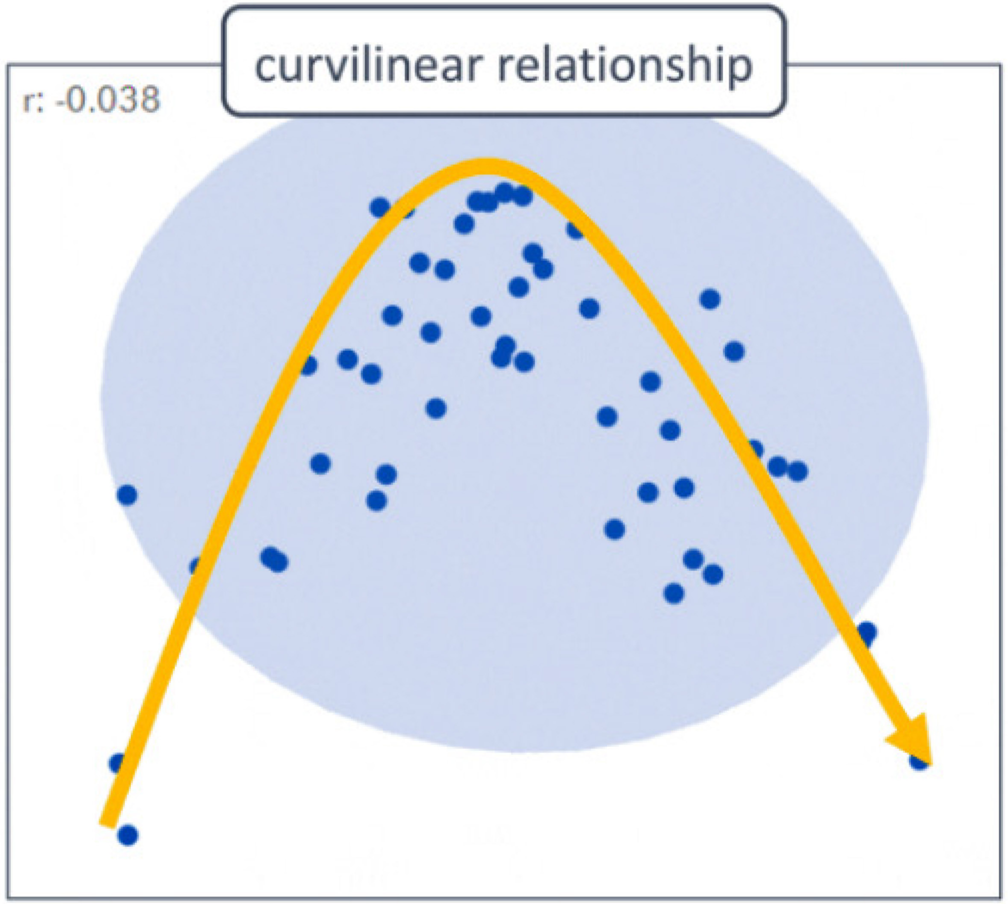
Non-Linear and Non-Monotonic Associations

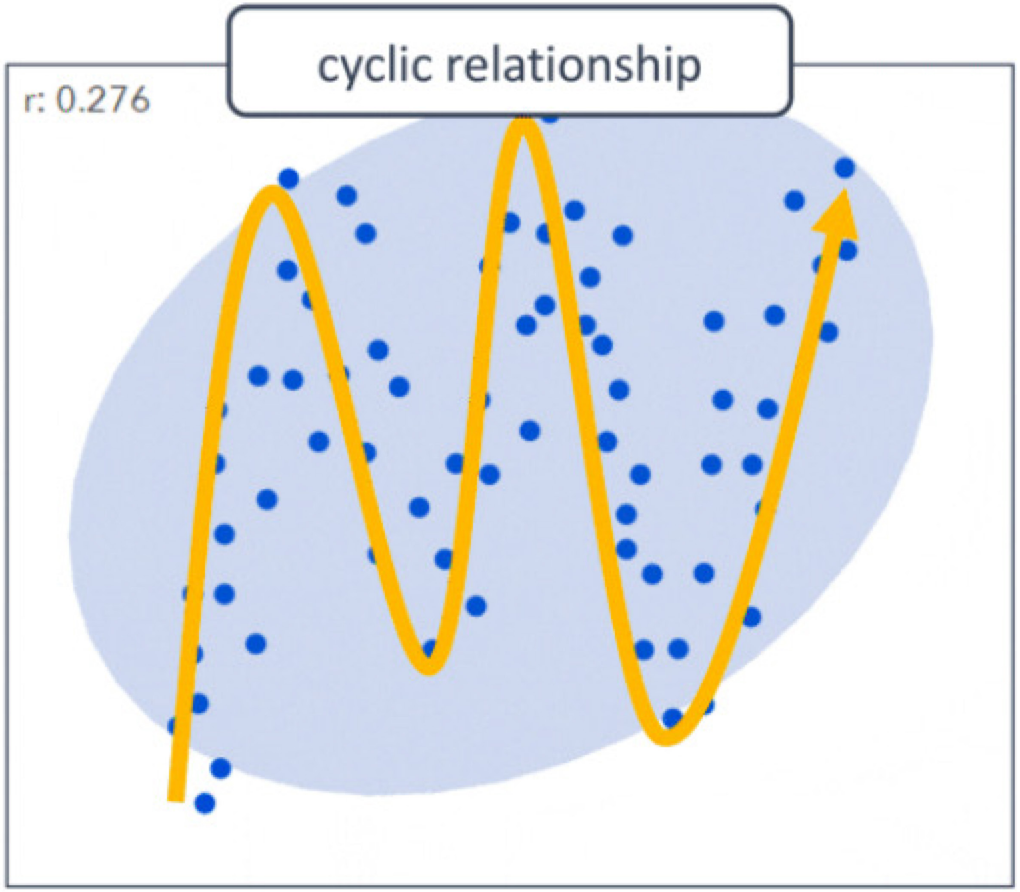




https://www.jmp.com/en_hk/statistics-knowledge-portal/what-is-correlation/correlation-coefficient.html

Require Specific Models and Additional Knowledge





https://www.jmp.com/en_hk/statistics-knowledge-portal/what-is-correlation/correlation-coefficient.html

Correlation Does **Not** Imply Causation

Causation requires knowledge of the causal mechanism: applying heat makes water boil

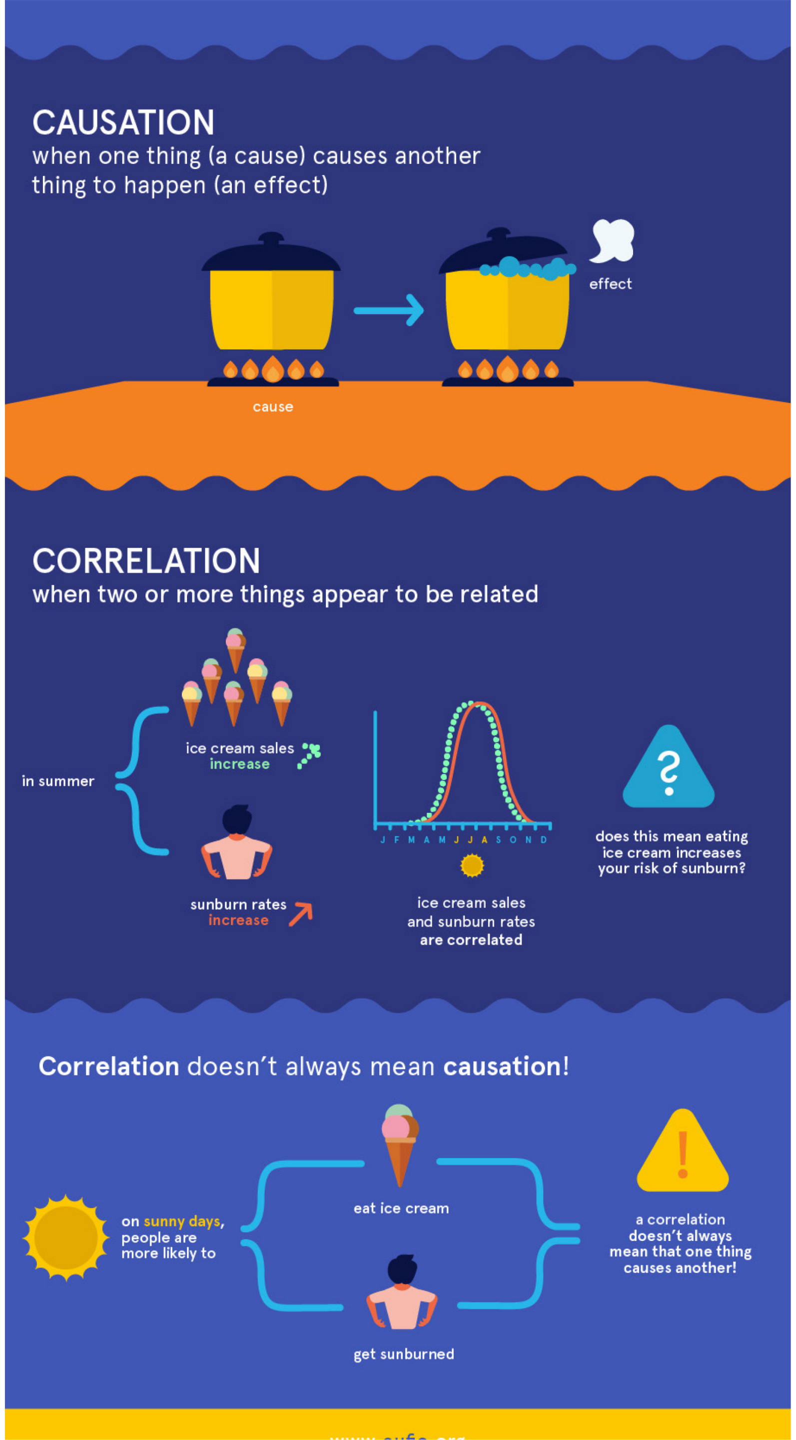
Correlation is not multivariate, no control for confounders

«On sunny days, one eats ice cream

On sunny days one gets sun burn

No! ice cream causes sun burn»

https://www.eufic.org/en/understanding-science/article/correlation-vs.-causation-infographic



Summary

Pearson correlation coefficient measures linear associations

Spearman rank correlation measures monotonic associations

Always plot histograms, scatter plots and check for outliers
Always check proportion of missing values

Evidence for causation has to do with study design

Correlation does not imply causation

(regression takes into account confounding factors)

R Exercise

Plot a matrix of scatterplots and histograms

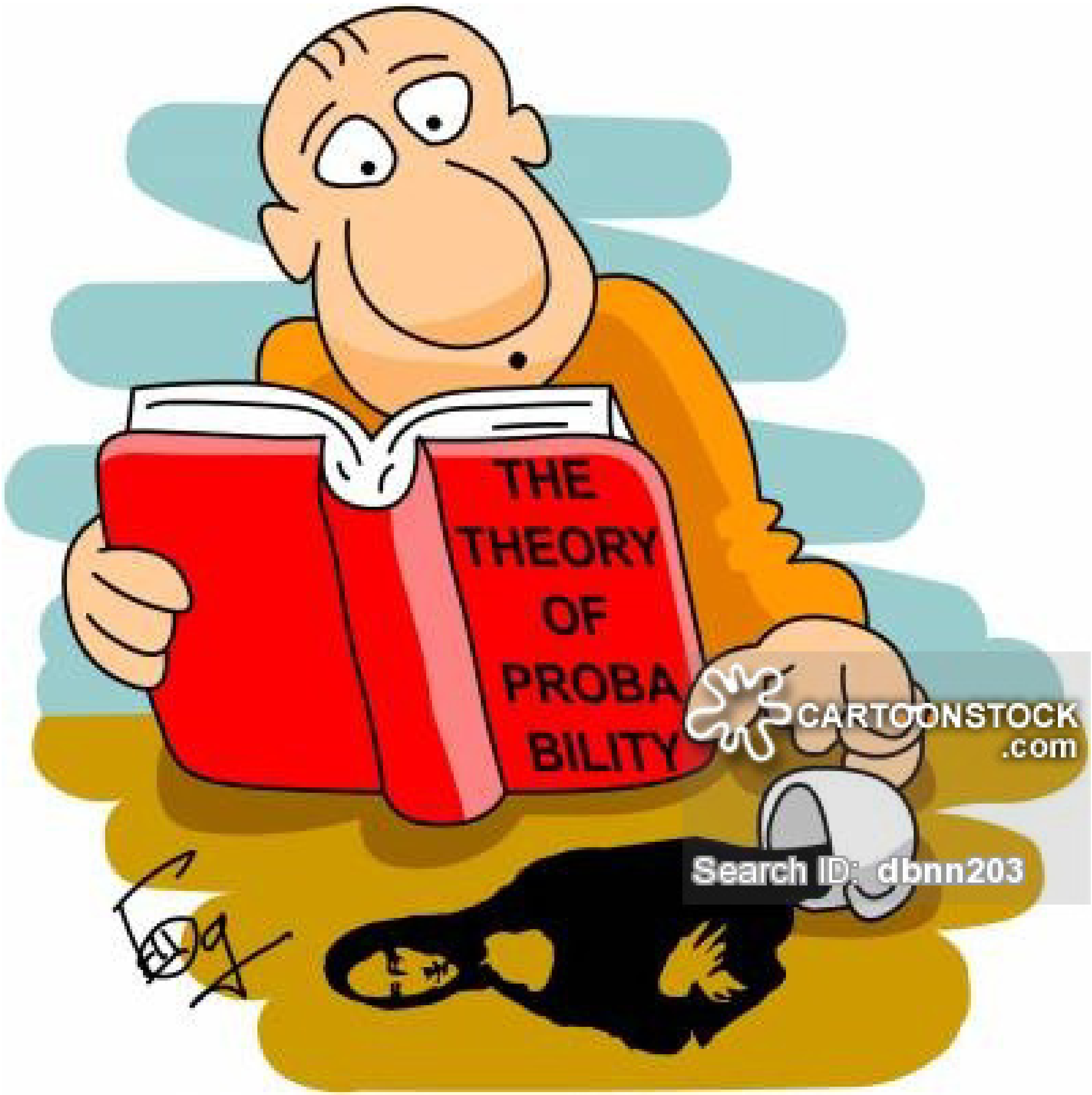
Calculate a correlation matrix

Missing value deletion: pairwise vs list-wise

Choose between Pearson or Spearman

Interpret results

Questions


THE THEORY OF PROBA BILITY
CARTOONSTOCK .com
Search ID: dbnn203

Thanks for your attention

b
UNIVERSITÄT
BERN

Dependent Data – Paired, Clustered, Panel Data

Dependent Events

Dr. Beatriz Vidondo

Veterinary Public Health Institute, UniBe

Objectives

Identify paired data and plot it adequately

Name tests adequate to compare paired groups

Understand between group/cluster variability and Intraclass Correlation Coefficient

Paired t Test

Two samples of measures on the same individuals

The paired measurements are not independent

Paired-sample t test is a **one sample t-test** performed on the pairwise ***differences*** between the two measurements

1. Calculate the difference (d) between both measures
2. Calculate the mean (m) and the standard deviation (s) of d
3. Compare the average difference to 0 (one-sample t test)

Wilcoxon Tests

Frank Wilcoxon described two tests with his name in 1945

"Individual comparisons by ranking methods" Biometrics Bulletin

1) Wilcoxon Rank Sum Test (2 non-normal independent groups)

`R function wilcox.test(y~x) default option paired = FALSE`

(Mann-Whitney U)

2) Wilcoxon Signed Rank to test differences of paired data

`R function wilcox.test(y~x, paired = TRUE)`

Wilcoxon Signed Rank Test

Uses data ranks (ordering) instead of the actual data

Rank is a score to sort results

Two ‘paired’ samples lead to a series of differences, some of which are positive (+) and some negative (-)

Signed ranks take the corresponding signs (+) or (-)

Ranking Data

For 8 Blocks = experiment repetitions, n=8 groups

Create an ordinal variable or score 1,2,3,4… n

Block	*A*	*B*	*A-B*	*Rank*
1	209	151	58	8
2	200	168	32	7
3	177	147	30	6
4	169	164	5	1
5	159	166	—7	—3
6	169	163	6	2
7	187	176	11	5
8	198	188	10	4

Signed Ranks:

A-B=5 gets the smallest rank 1

6 is the next rank 2

-7 is next rank but gets -3

…

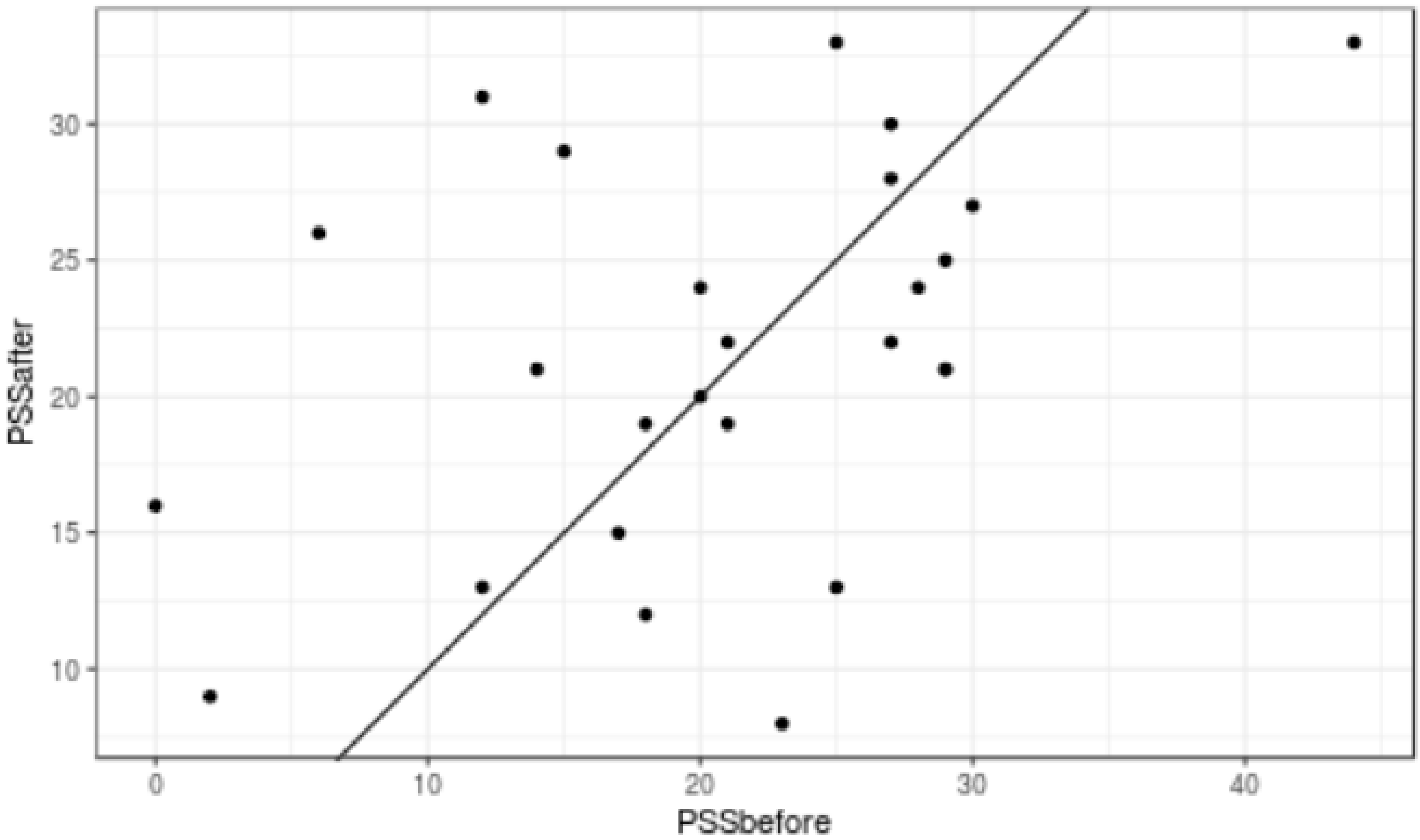
Wilcoxon’s Probability Table

TABLE II

For Determining the Significance of Differences in Paired Experiments

Number of Paired Comparisons	Sum of rank numbers, + or —, whichever is less	Probability of this total or less
7	0	0.016
7	2	0.047
8	0	0.0078
8	2	0.024

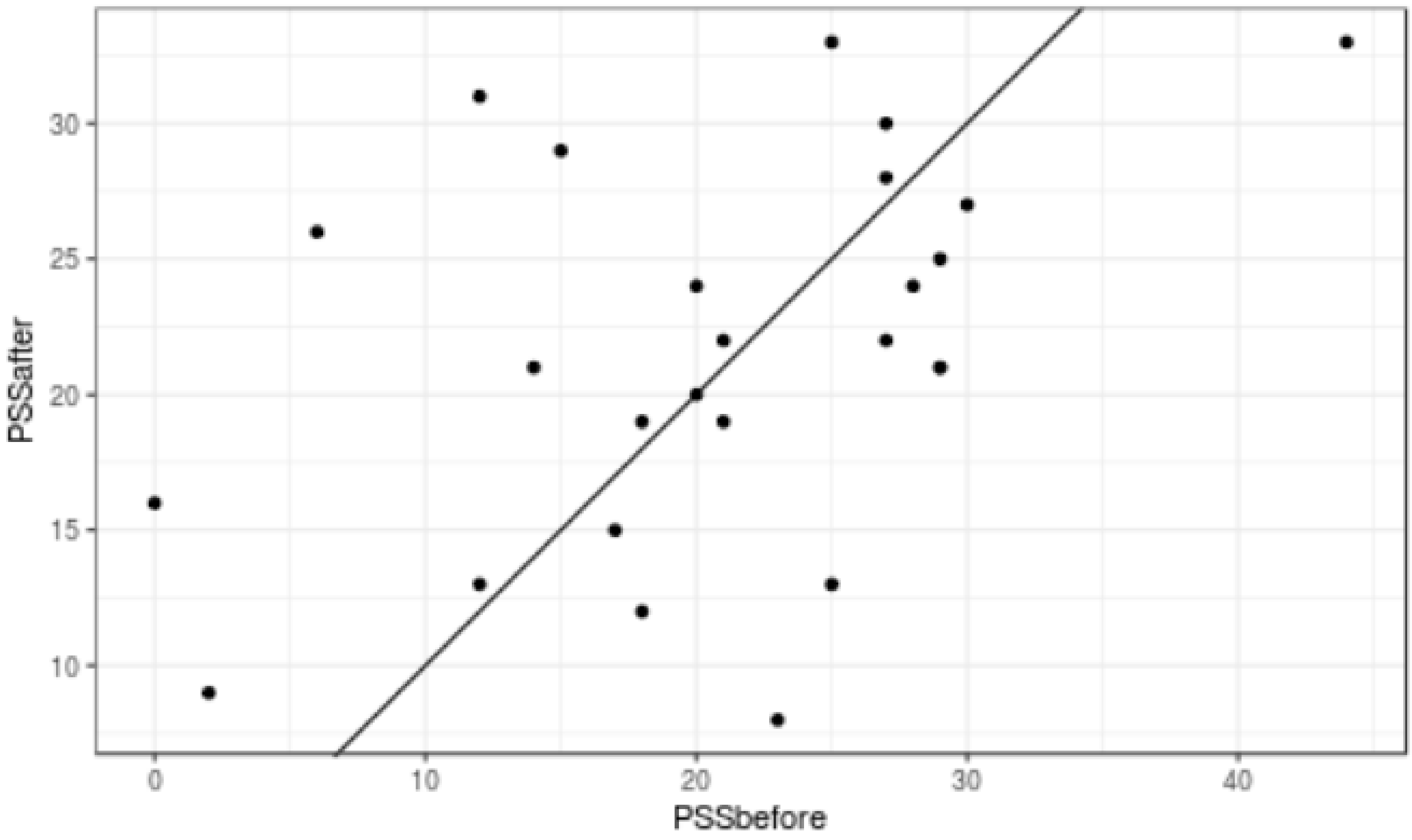
Plotting Paired Data

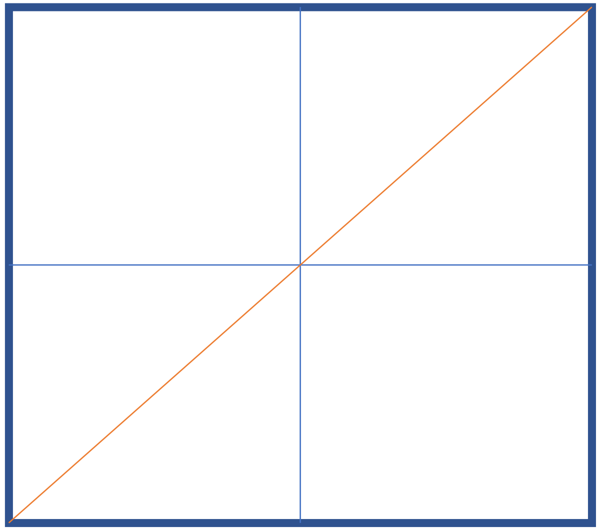


Stress Score at entry (before) and exit (after)

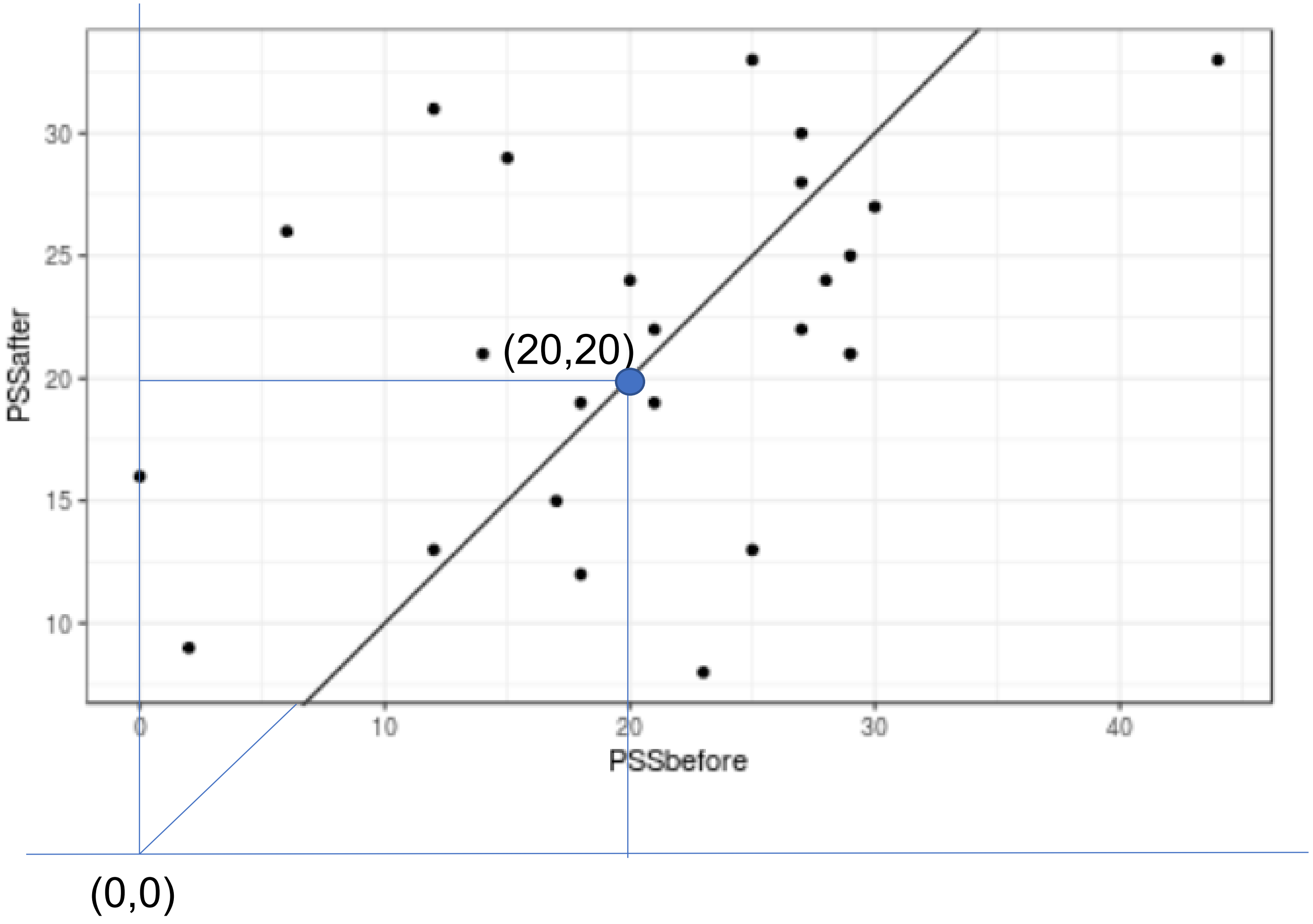
Can you tell what this line is?

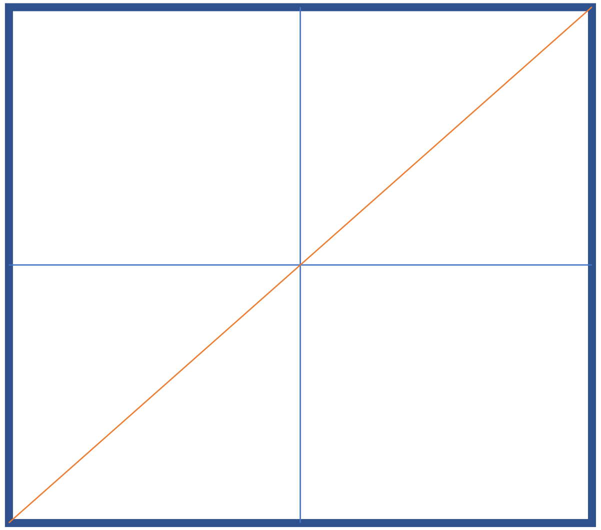
Plotting Paired Data



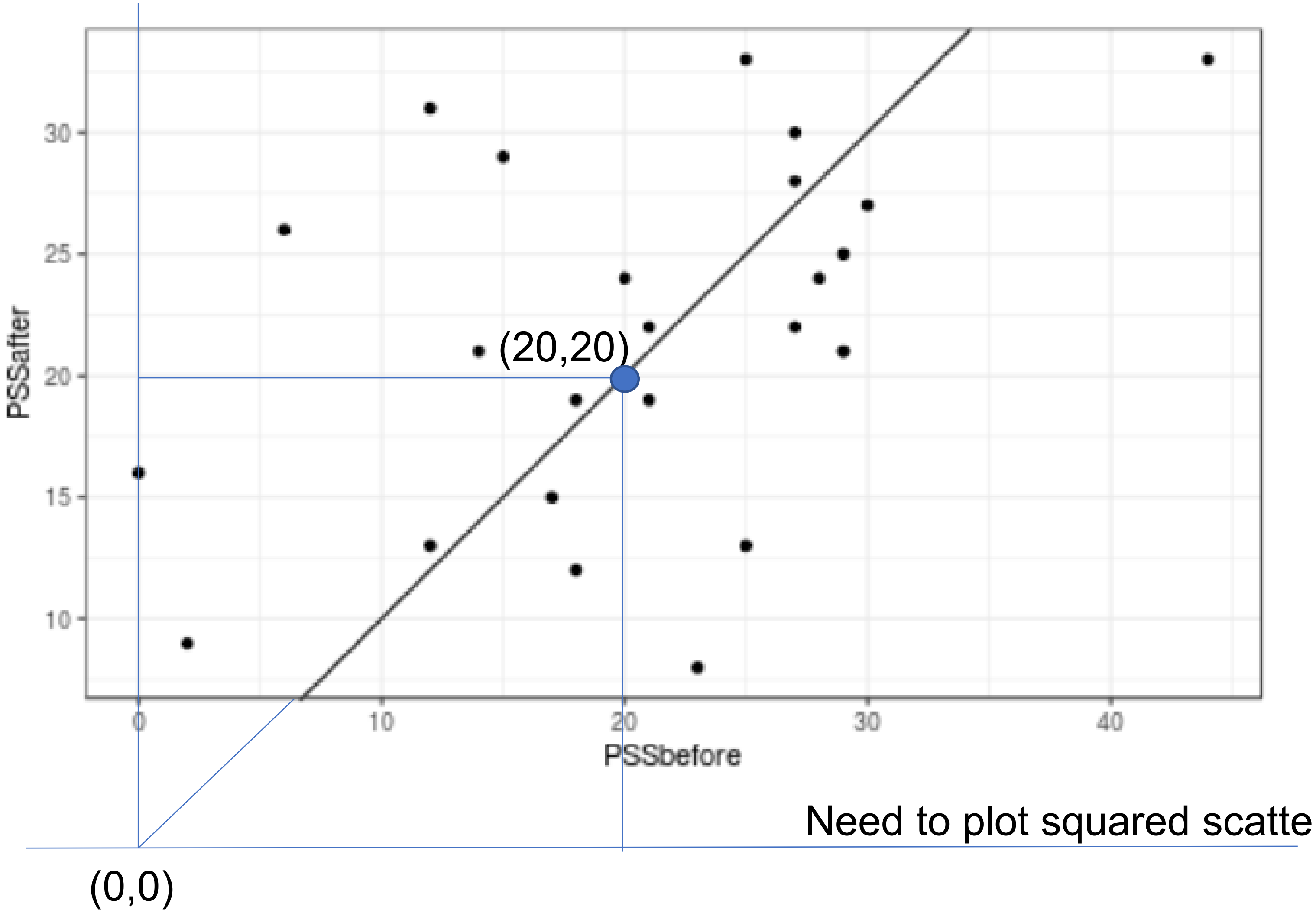


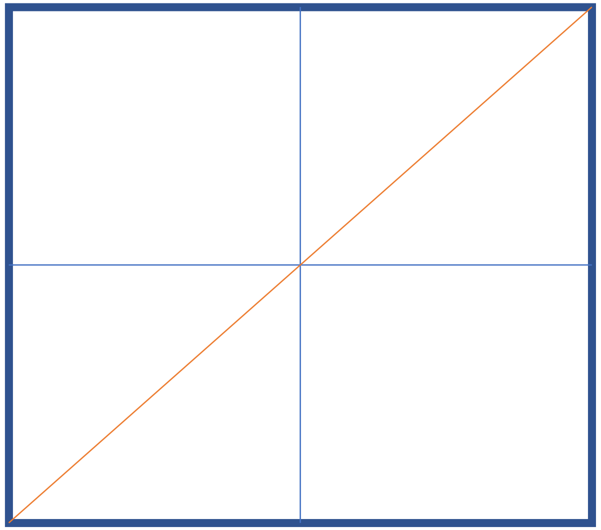
Plotting Paired Data





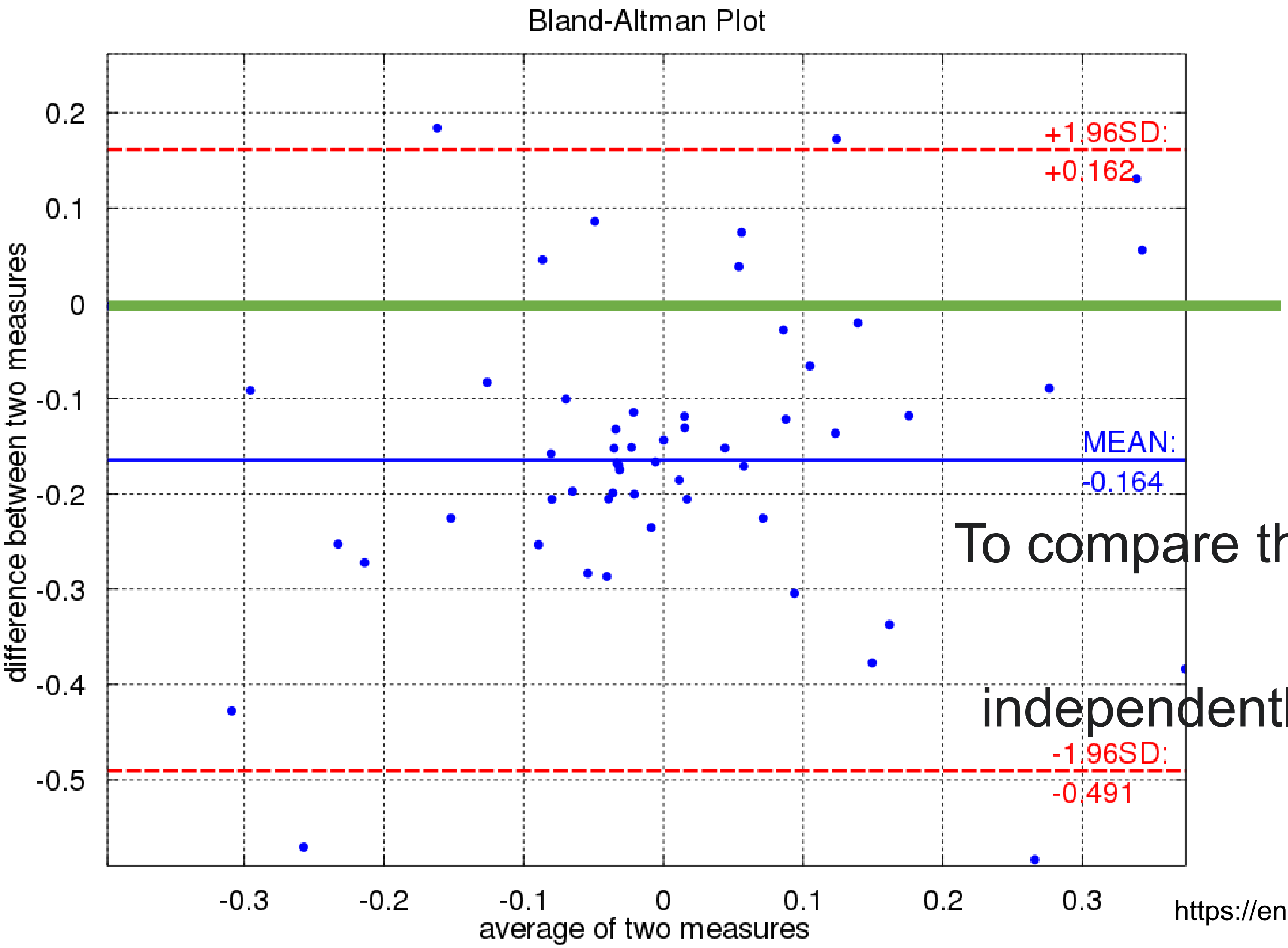
Plotting Paired Data





line y=x is a reference line
Need to plot squared scatterplot (same units for both axes)

Agreement: Bland-Altman Plot

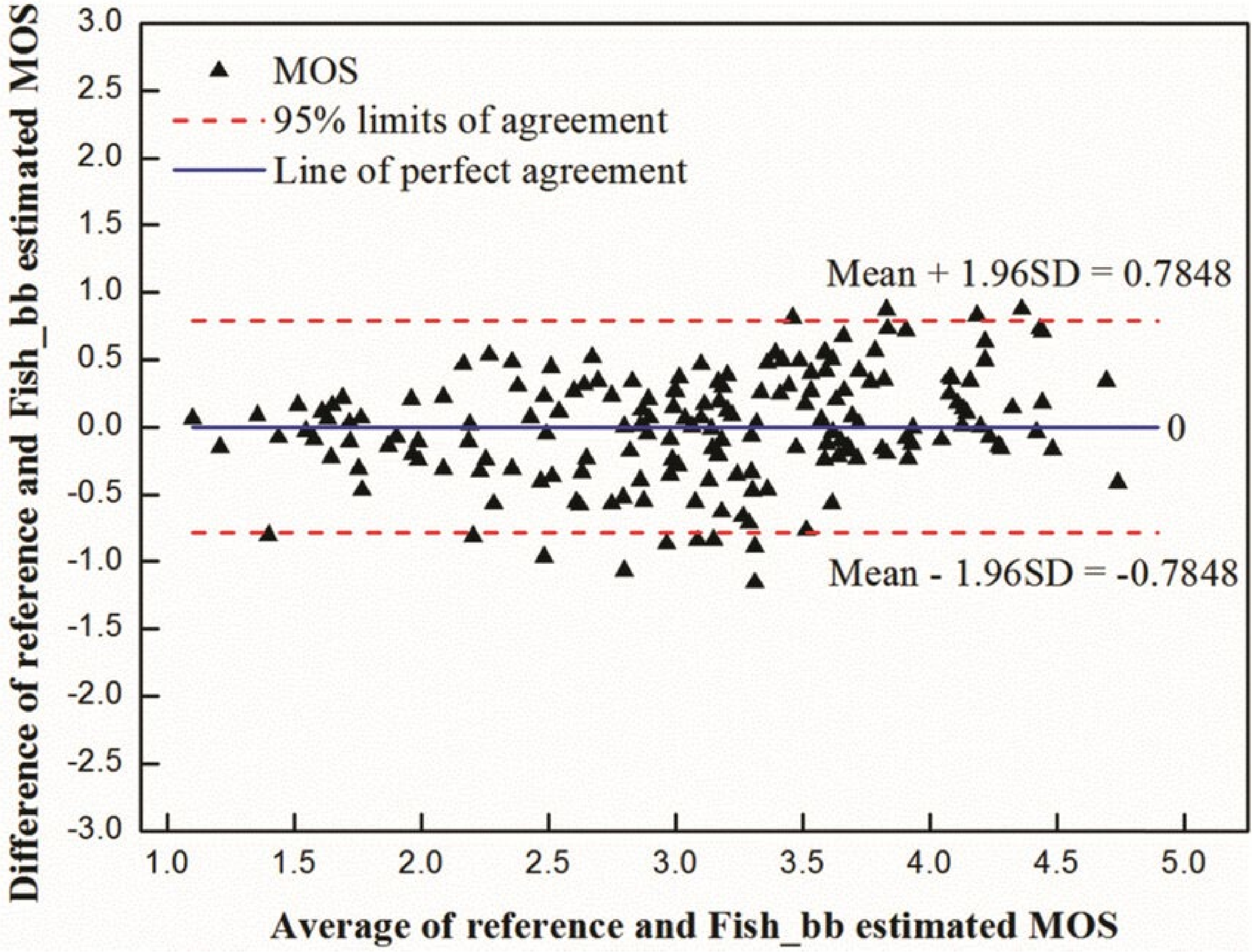


Reference line at 0,0

To compare the dissimilarities between the two sets of samples independently from their mean values

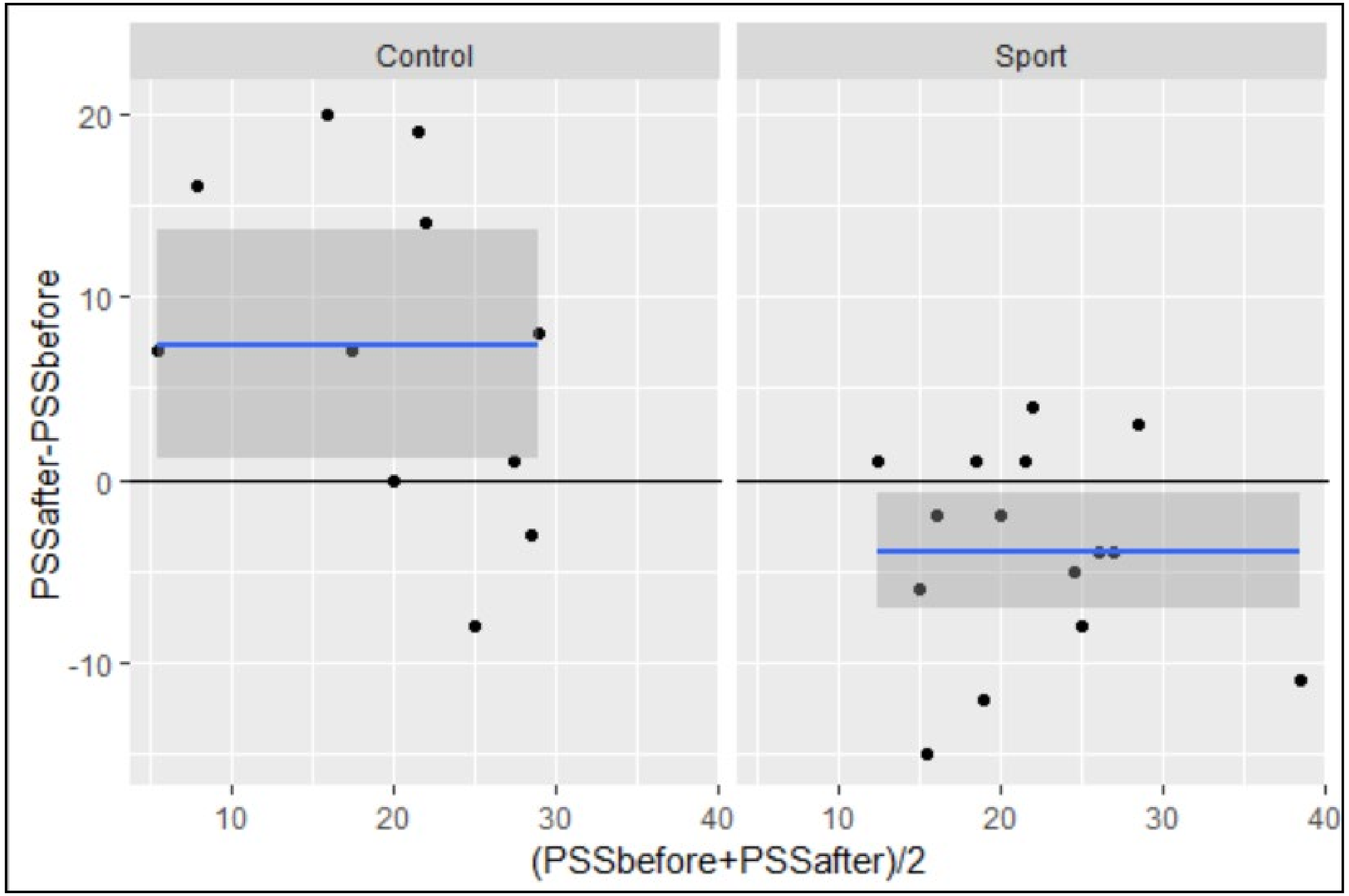
Image from https://en.wikipedia.org/wiki/Bland%E2%80%93Altman_plot

Agreement: Bland-Altman Plot



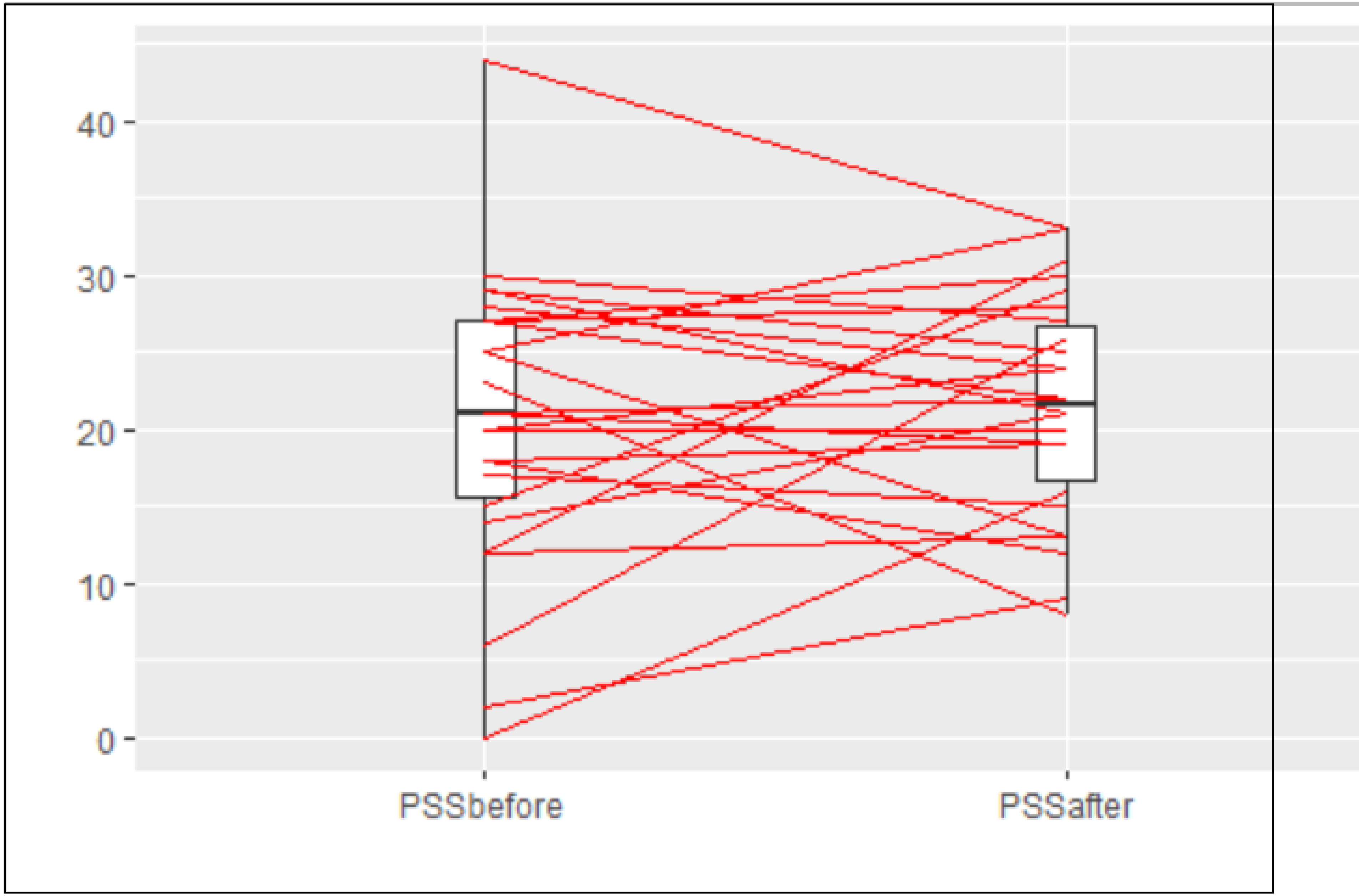
https://www.researchgate.net/figure/Bland-Altman-plot-N-181-image-samples-of-the-difference-against-average-for-THz-IQA_fig4_322907411

Bland-Altman Plot & Treatment Effects



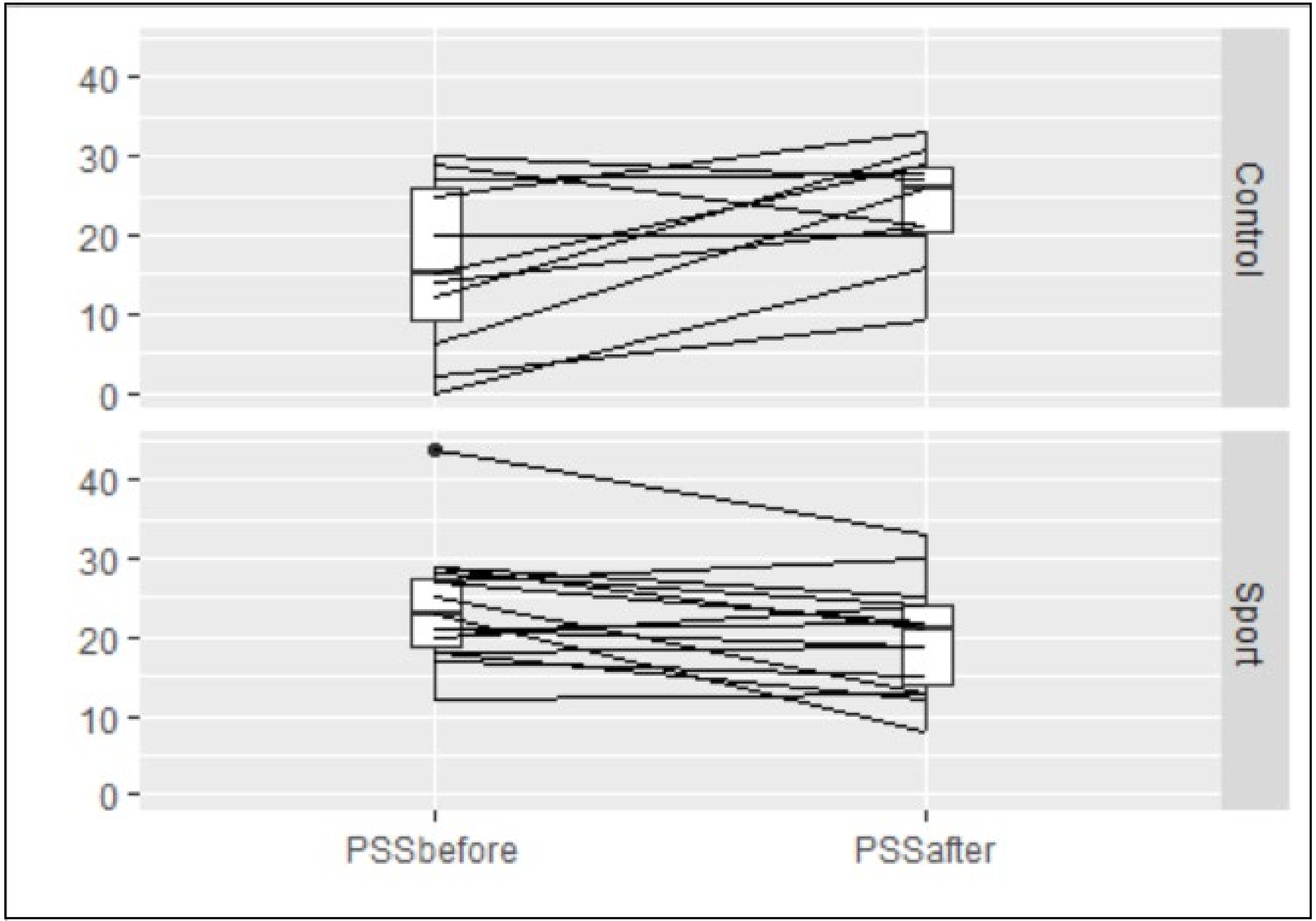
Tukey mean difference or Bland Altman plot

Profile Plot

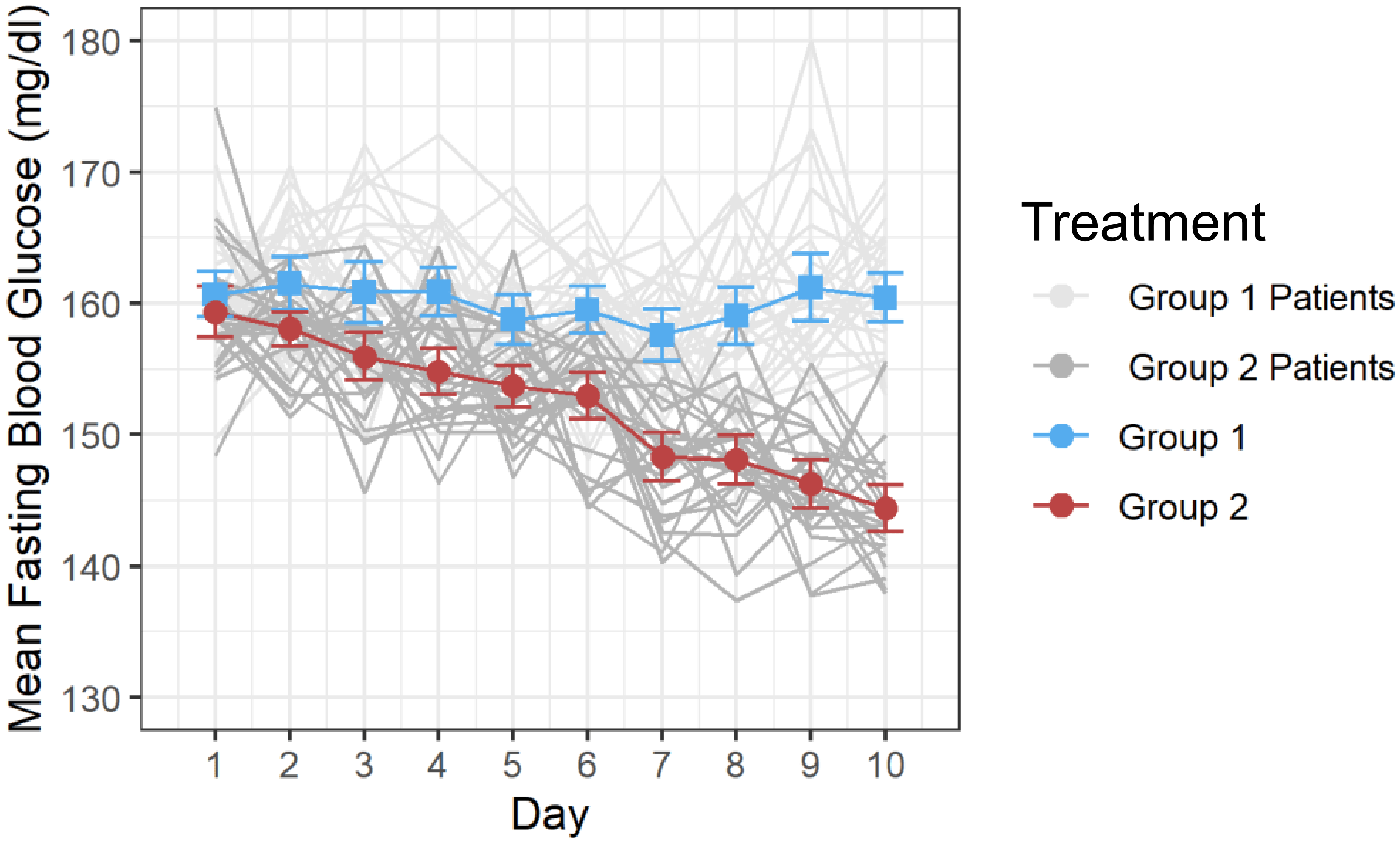


Profile Plot

Are the Control and Sport groups here dependent or independent?

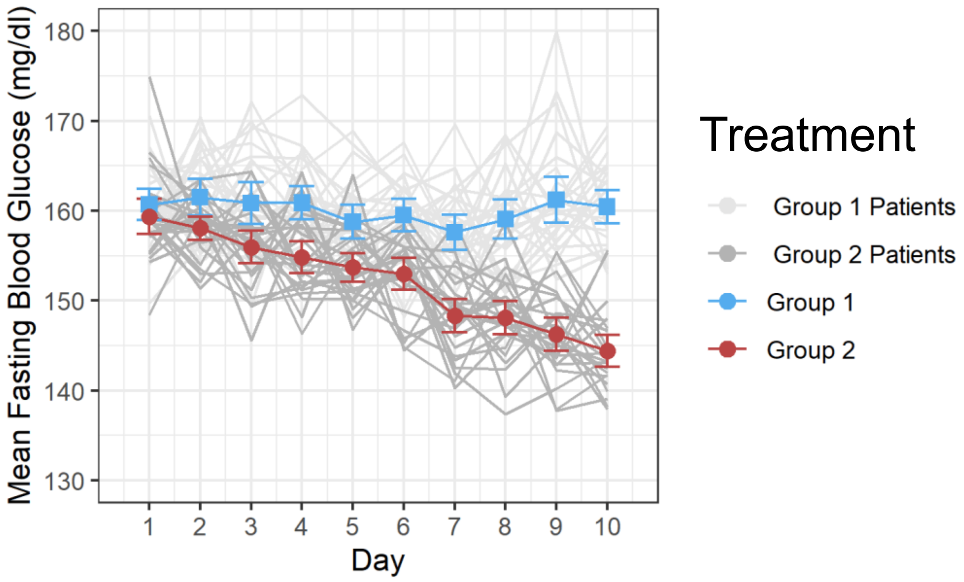


Panel Data Several Time Points



https://rgraphs.com/mean-profile-plot-in-r/

Panel Data Several Time Points



Every patient is a "cluster" of measures

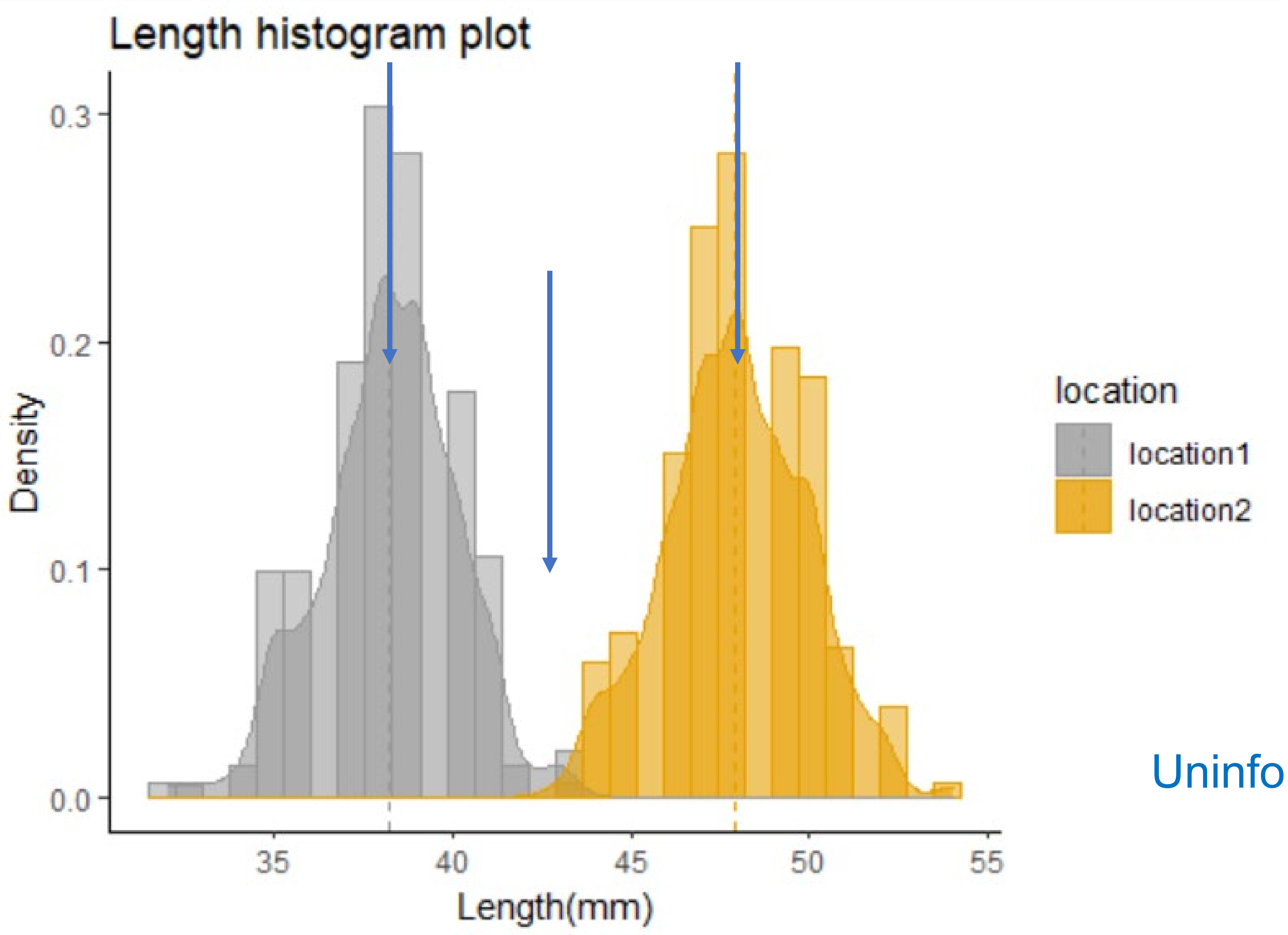
Treatment groups independent

Cross-over designs => Treatment groups dependent

https://rgraphs.com/mean-profile-plot-in-r/

Clustered Data

Clustered Data



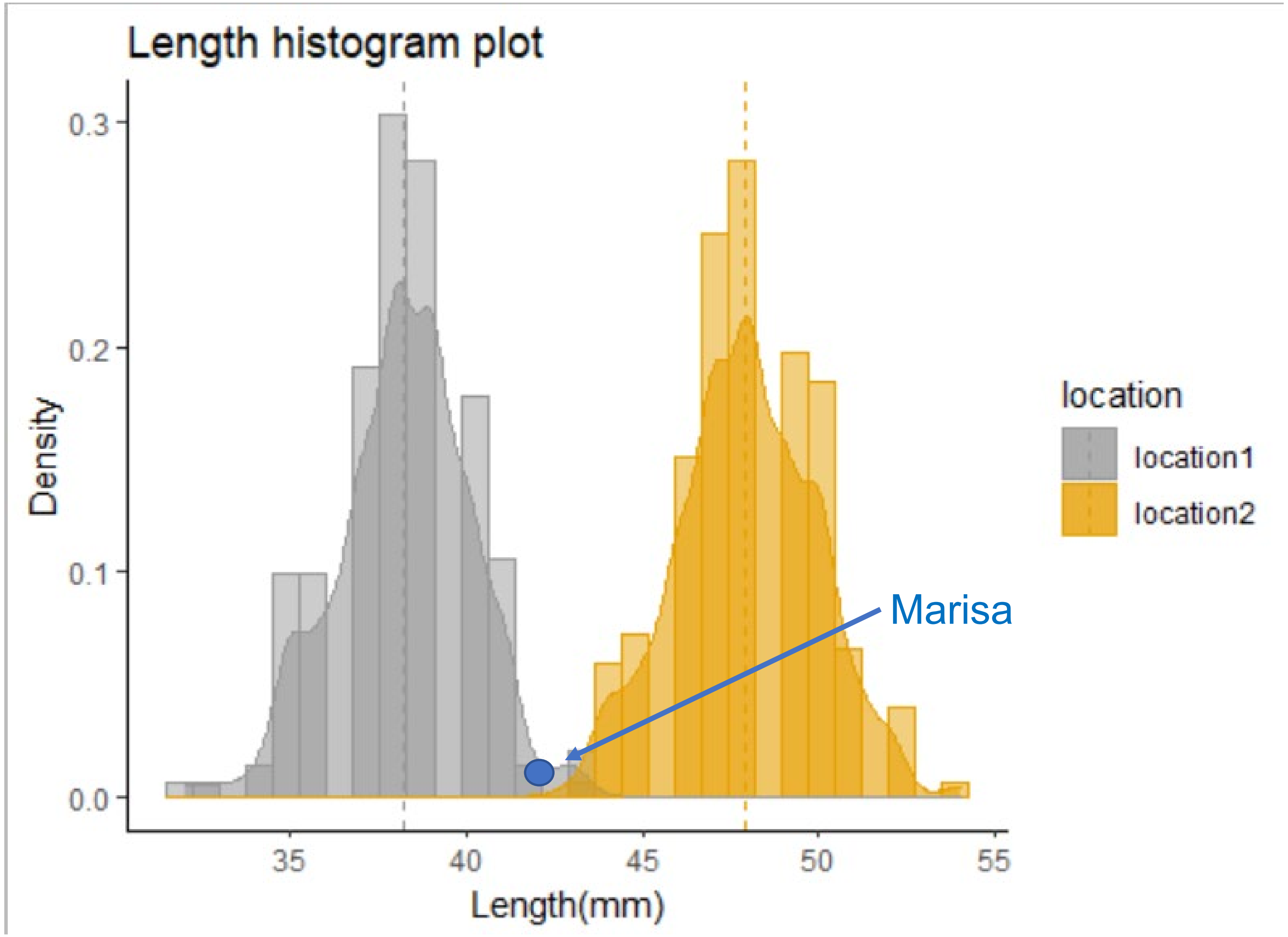
Mean Location 1 = 38

Mean Location 2 = 48

Uninformative overall mean = 43

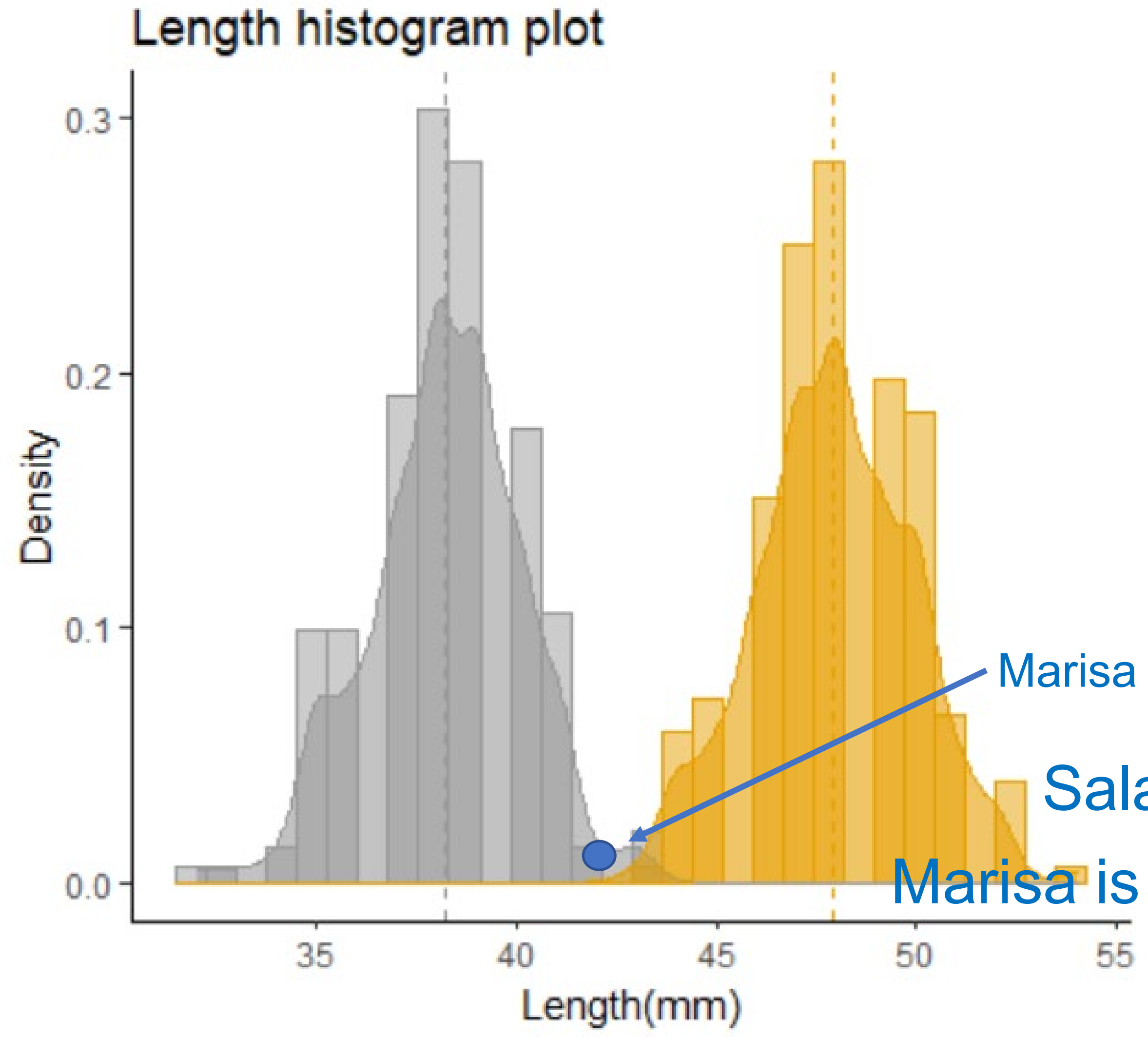
https://scholars.duke.edu/person/stephen.vaisey

Clustered Data



https://scholars.duke.edu/person/stephen.vaisey

Between and Within Statements

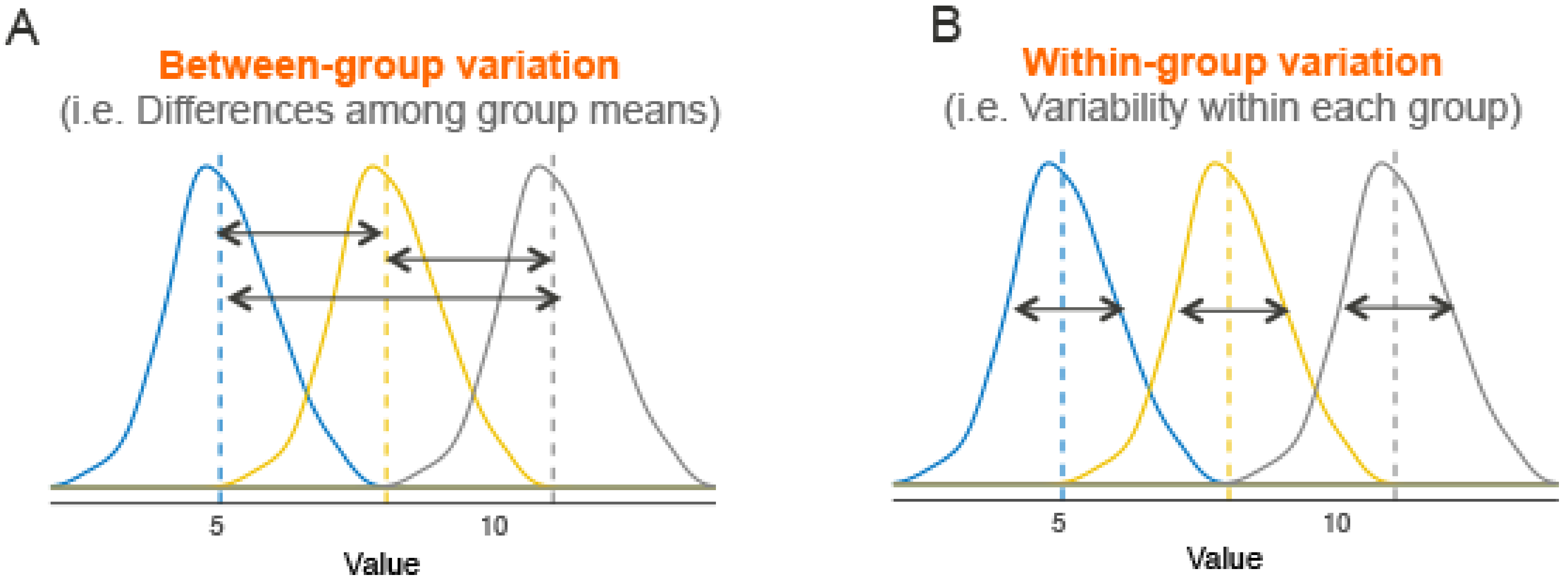


Salamanders at location 2 are larger

Marisa is one of the largest at her location

https://scholars.duke.edu/person/stephen.vaisey

Between and Within Variance, Cluster = Group



https://www.datanovia.com/en/lessons/anova-in-r/

Between and Within Cluster/Group Variance

Between variance = average distance of (each) cluster/group means to overall mean

Overall mean

Cluster/group means

Within variance = average variance of observations to their cluster/group means

Variance of each cluster/group separately

Average of these variances

https://www.statology.org/within-between-group-variation-anova

Intraclass Correlation Coefficient ICC

$$ICC = \frac{Between\ cluster/group\ variance}{Total\ variance}$$

$$\rho = \frac{\tau^2}{\tau^2 + \sigma^2}$$

τ^2 (tao) is the between cluster/group variance

σ^2 is the within cluster/group variance

https://scholars.duke.edu/person/stephen.vaisey

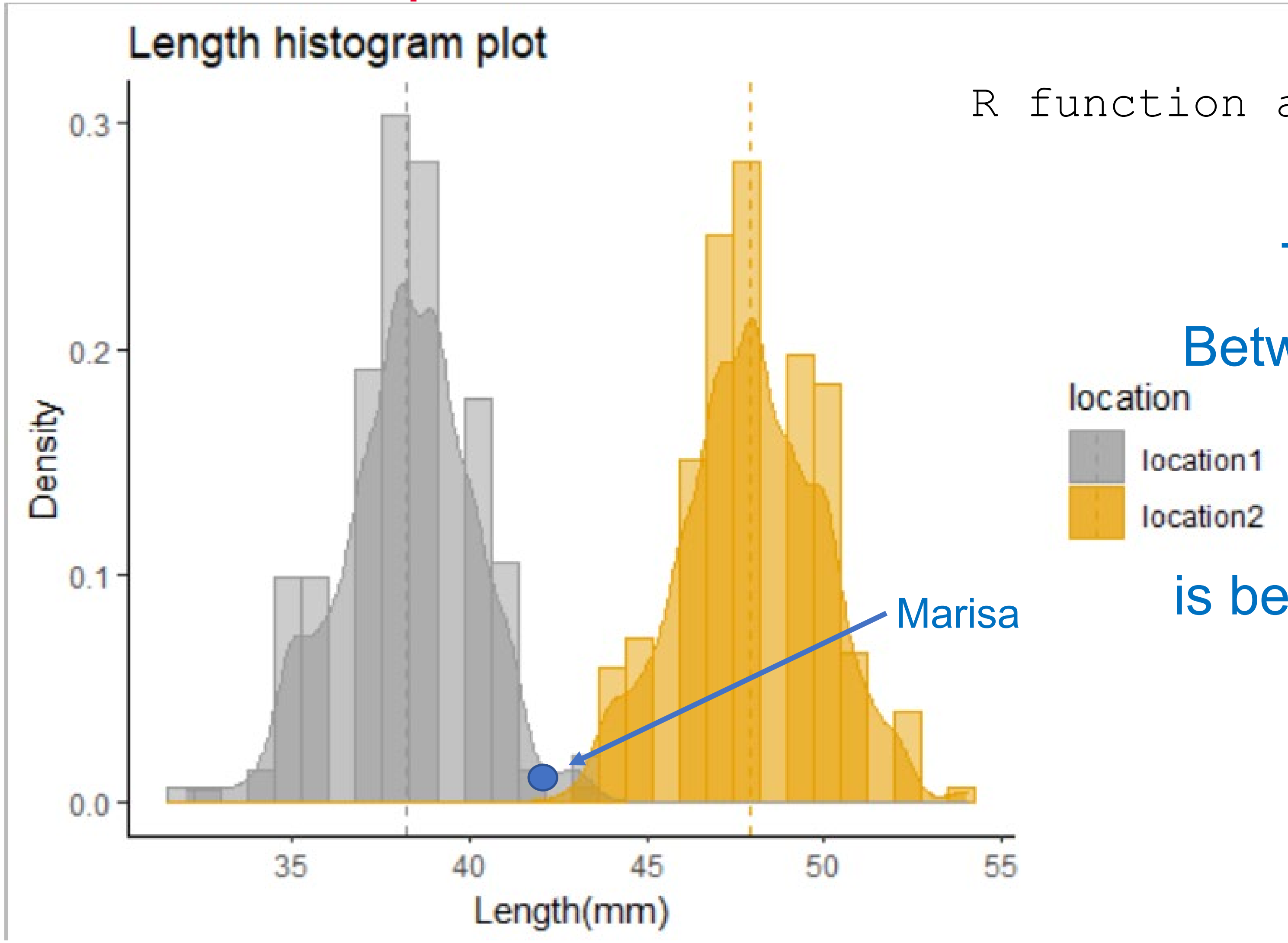
ICC

ICC is a new descriptive statistic that describes where the variance "lives"

A new descriptive statistic that becomes available when you deal with clustered or panel data

With an ICC of 100% individual characteristics are defined by its group/cluster

Cluster Example




```
R function aov(length~location)
```

Total Variance = 27.33

Between Variance = 3.694

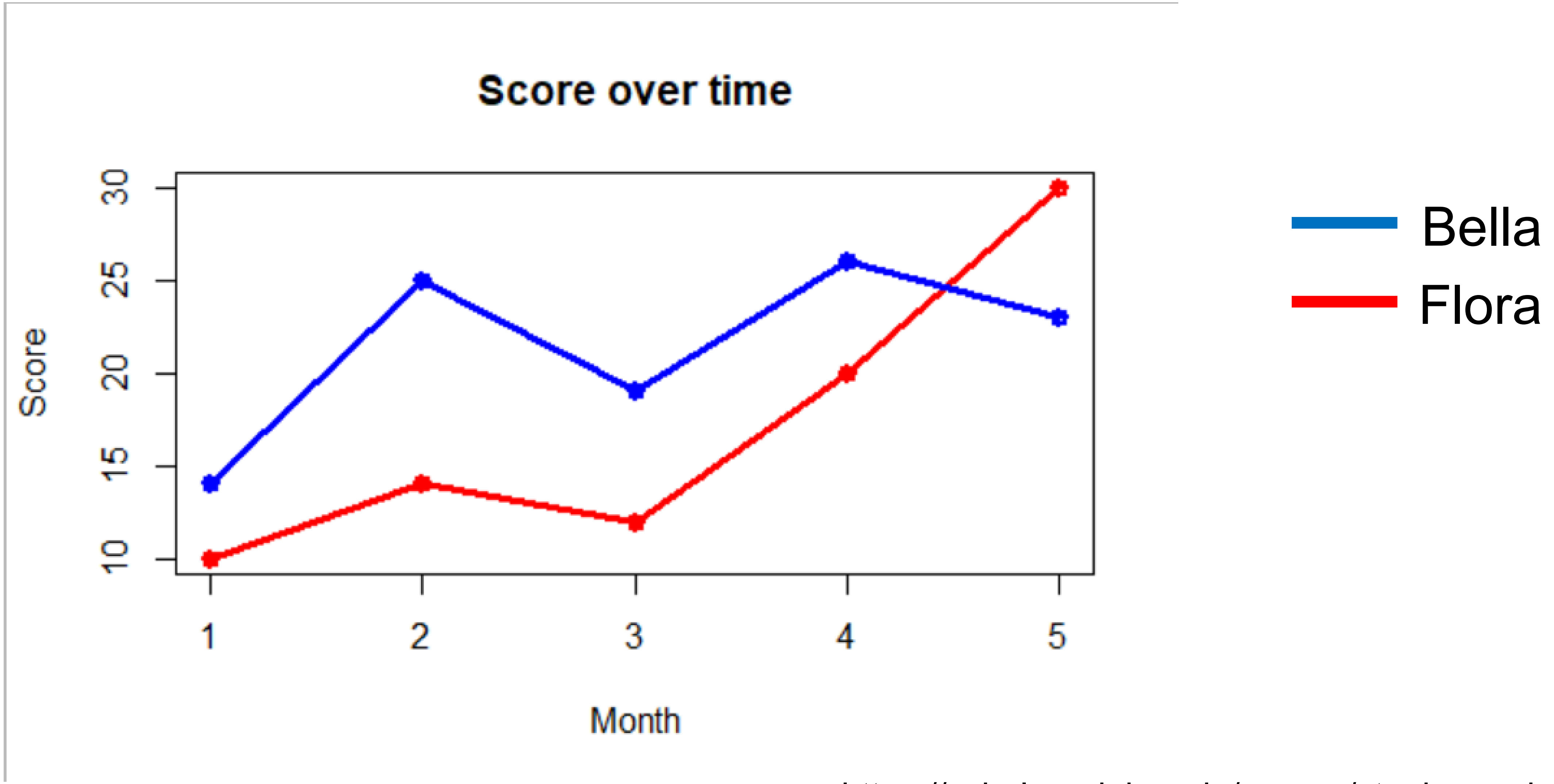
ICC = 0.865

87% of variance

is between clusters/groups

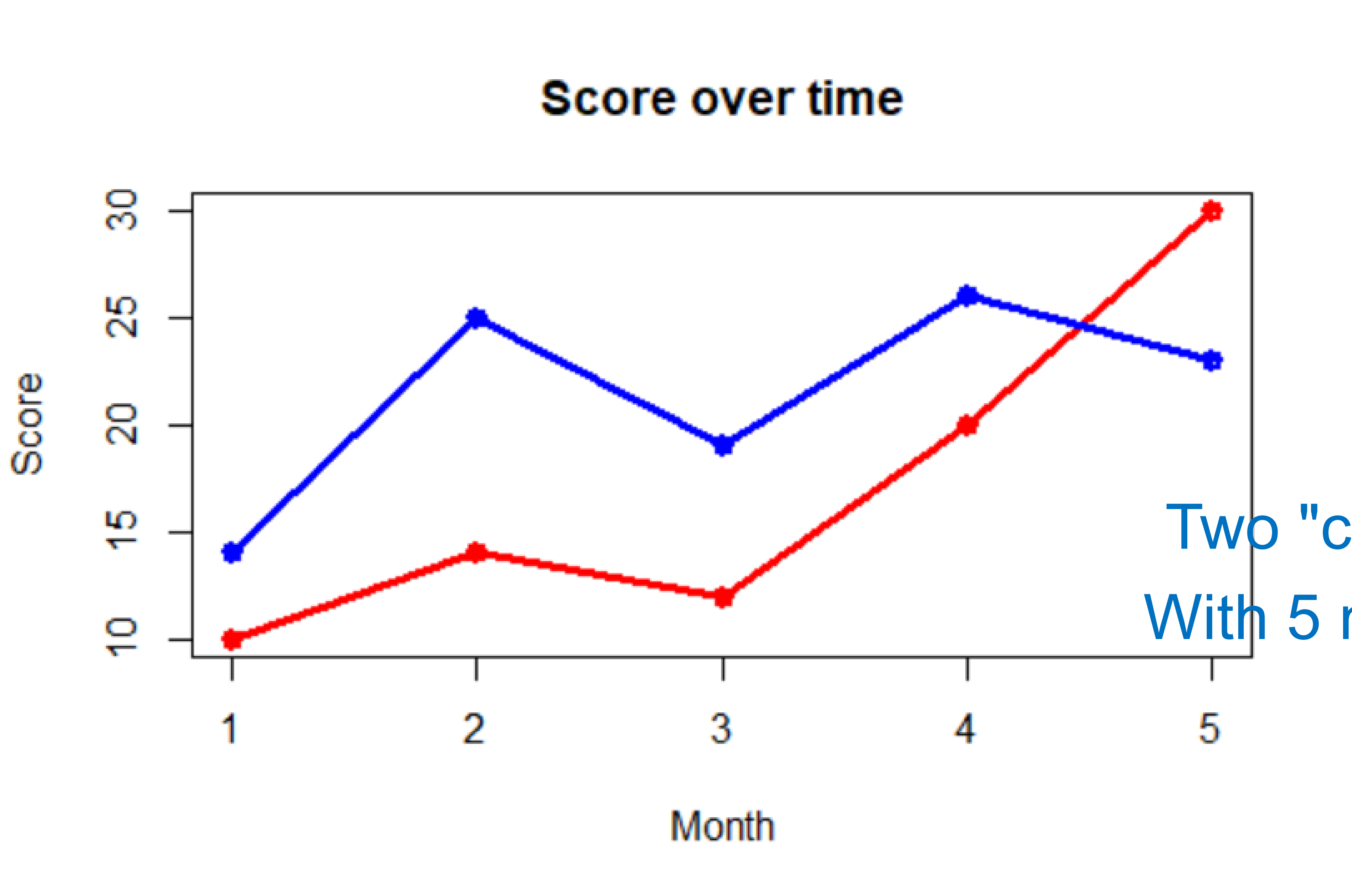
https://scholars.duke.edu/person/stephen.vaisey

Panel Example



https://scholars.duke.edu/person/stephen.vaisey

Panel Example



Bella

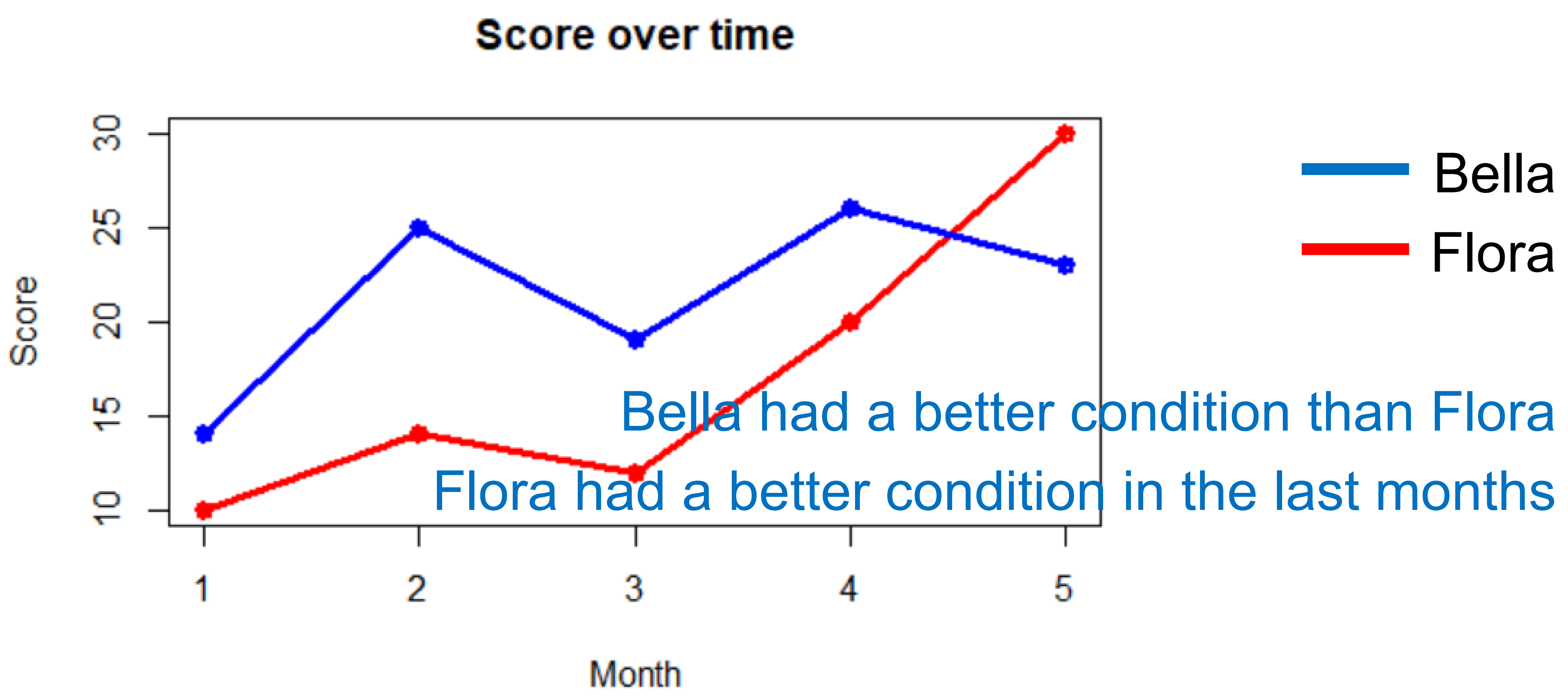
Flora

Two "clusters" of data

With 5 measures each

https://scholars.duke.edu/person/stephen.vaisey

Between and Within Statements

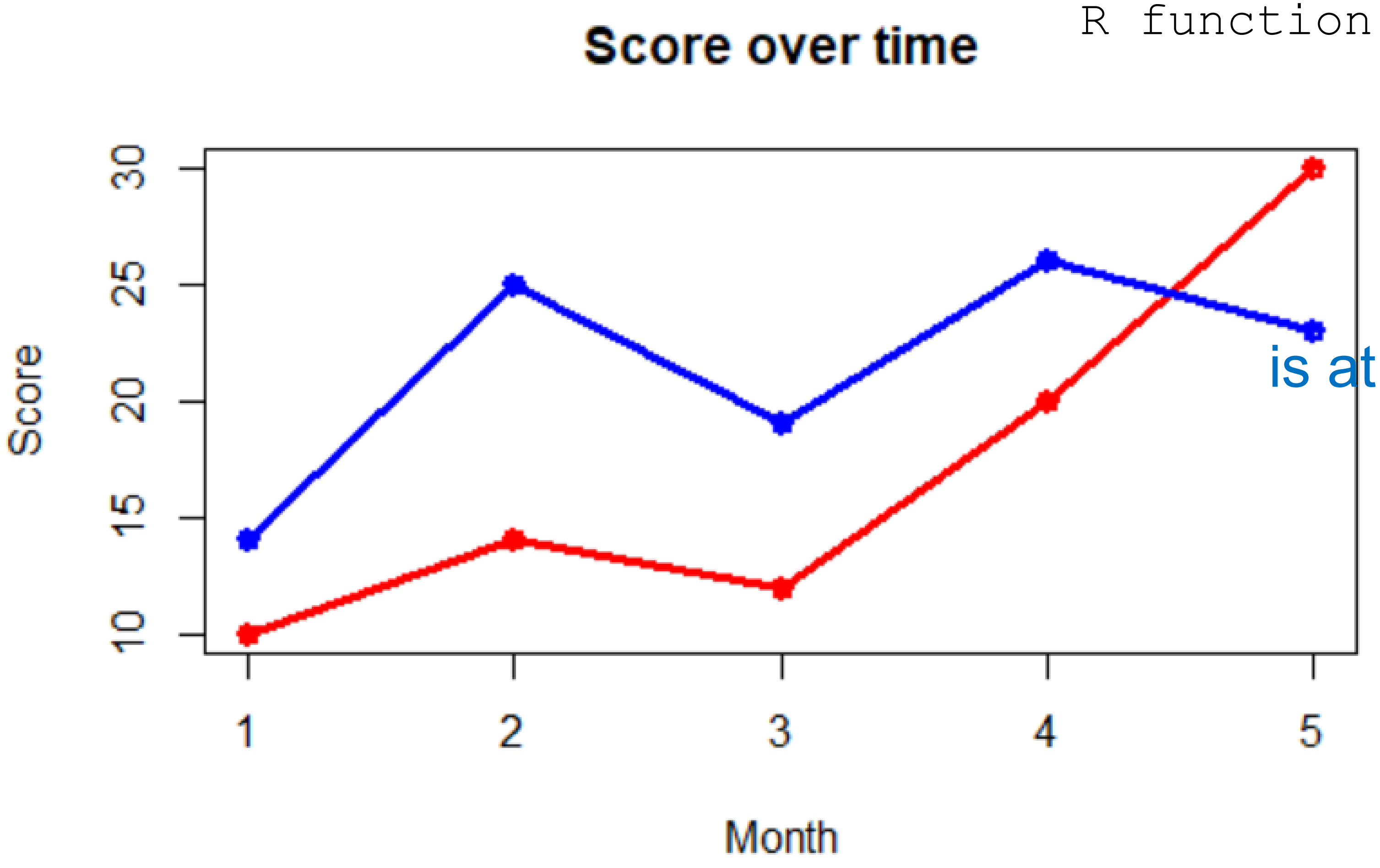


Bella had a better condition than Flora

Flora had a better condition in the last months

https://scholars.duke.edu/person/stephen.vaisey

Panel Example



`R function aov(score~subject)`

ICC = 0.11

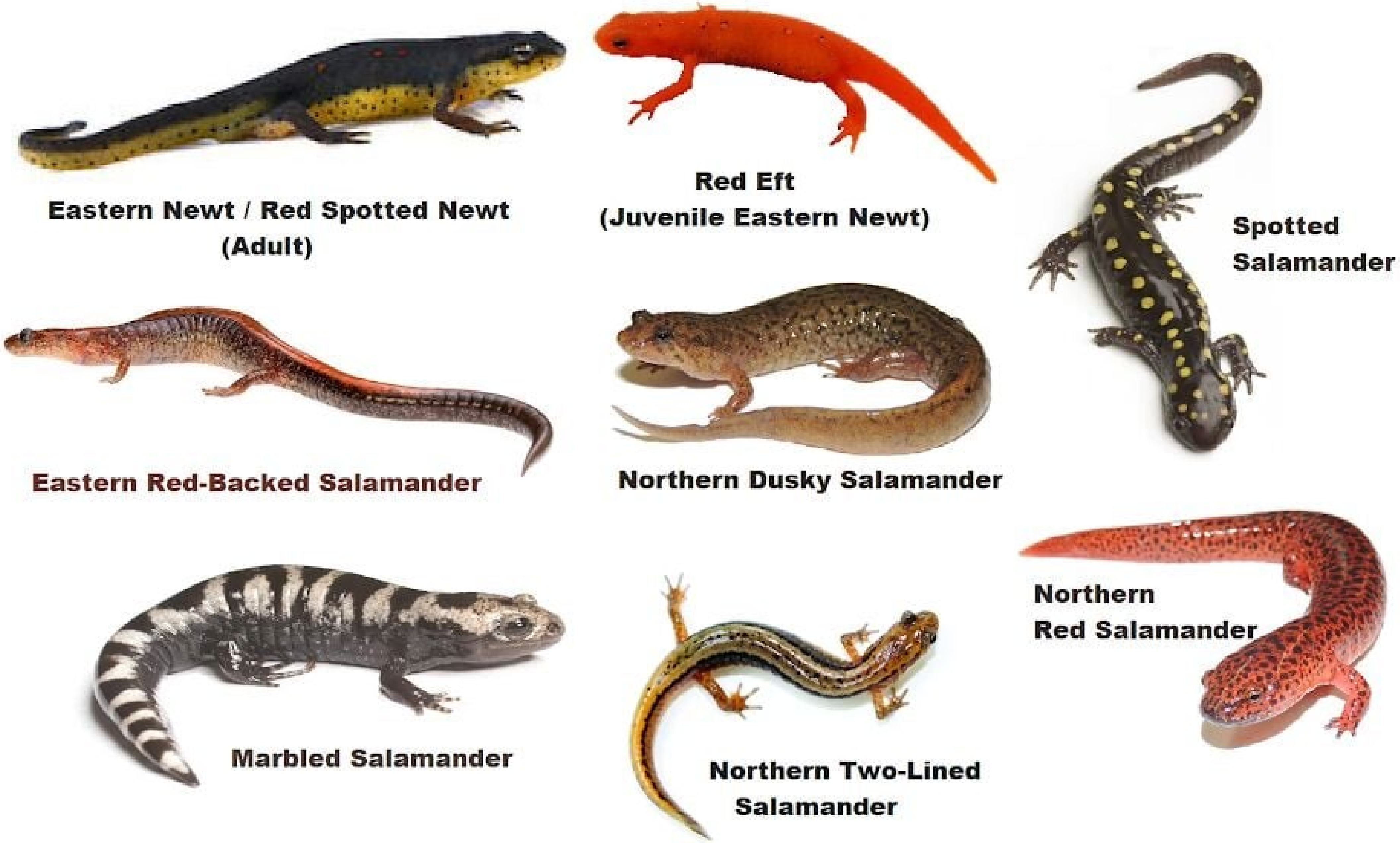
11% of variance
is at subject (group) level

https://scholars.duke.edu/person/stephen.vaisey

Questions

Quantifying Variance: Solving the Riddle

https://scholars.duke.edu/person/stephen.vaisey



https://www.shipnc.com/columns/stream_of_consciousness/article_30fb5d9a-7ec4-11ec-b9a2-37dc22ab5df2.html

Summary

Simple tests, Wilcoxon Signed Rank Test

Plotting paired data

Bland-Altman plot (agreement between methods)

Between cluster/group variance

Intraclass Correlation Coefficient – where the variance lives!

Exercises

Plots (profile, Bland-Altman)

Perform Wilcoxon signed rank

Calculate and interpret ICC

Thanks for your attention

b
UNIVERSITÄT
BERN

Variance

$$\text{Var(Y)} = \sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{y})^2$$

i Subscript for every observation

$\sum_{i=1}^{N}$ Sum of content in () from 1 zu N

N is the number of observations

$\bar{y}$ is the overall mean

y_i is every observation

Variance is the average distance from observations to their mean

What is the average of this data set?

36, 37, 38, 39, 40

https://scholars.duke.edu/person/stephen.vaisey

What is the average of this data set?

36, 37, 38, 39, 40

The average is 38

What is the variance?

https://scholars.duke.edu/person/stephen.vaisey

36, 37, 38, 39, 40

36-38, 37-38, 38-38, 39-38, 40-38

-2, -1, 0, 1, 2

-2^2, -1^2, 0^2, 1^2, 2^2

The variance is the average of this new data series 2.5

https://scholars.duke.edu/person/stephen.vaisey

Between Group Variance

i Subscript for every observation

j Subscript for every group

n_j is the sample size of every group

M is the number of groups

$\bar{y}_j$ is the mean of each group

$\bar{y}$ is the overall mean

https://www.statology.org/within-between-group-variation-anova

Between Group Variance

i Subscript for every observation

j Subscript for every group

n_j is the sample size of every group

M is the number of groups

$\bar{y}_j$ is the mean of each group

$\bar{y}$ is the overall mean

$$\text{Between } Variation(Y_{ij}) = \sum_{j=1}^{M} n_j(\bar{y}_j - \bar{y})^2$$

https://www.statology.org/within-between-group-variation-anova

Within Group Variance

i Subscript for every observation

j Subscript for every group

n_j is the sample size of every group

M is the number of groups

$\bar{y}_j$ is the mean of each group

$\bar{y}$ is the overall mean

y_{ij} is the ith observation of group j

$$\text{Within } Variation(Y_{ij}) = \sum_{j=1}^{M} \sum_{ij=1}^{N_{ij}} (y_{ij} - \bar{y}_j)^2$$

https://www.statology.org/within-between-group-variation-anova

Analysis of Variance ANOVA Table

SS = Sum of Squares = Sum of Squared Differences

ANOVA will be covered on the regression course in October

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	**207.2**	2	103.6	7.6952	0.0023	3.3541
Within Groups	**363.5**	27	13.4630			
Total	570.7	29				

https://www.statology.org/within-between-group-variation-anova

Variance Partition Table

SS = Sum of Squares = Sum of Squared Differences

calculate it regardless of independent/dependent data

Provides descriptive statistics at two levels

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	**207.2**	2	103.6	7.6952	0.0023	3.3541
Within Groups	**363.5**	27	13.4630			
Total	570.7	29				

https://www.statology.org/within-between-group-variation-anova