VETERINARY PUBLIC HEALTH INSTITUTE

UNIVERSITÄT BERN

Non-Normality

Normality Tests and QQ-plots

Dr. Beatriz Vidondo

Veterinary Public Health Institute UniBe

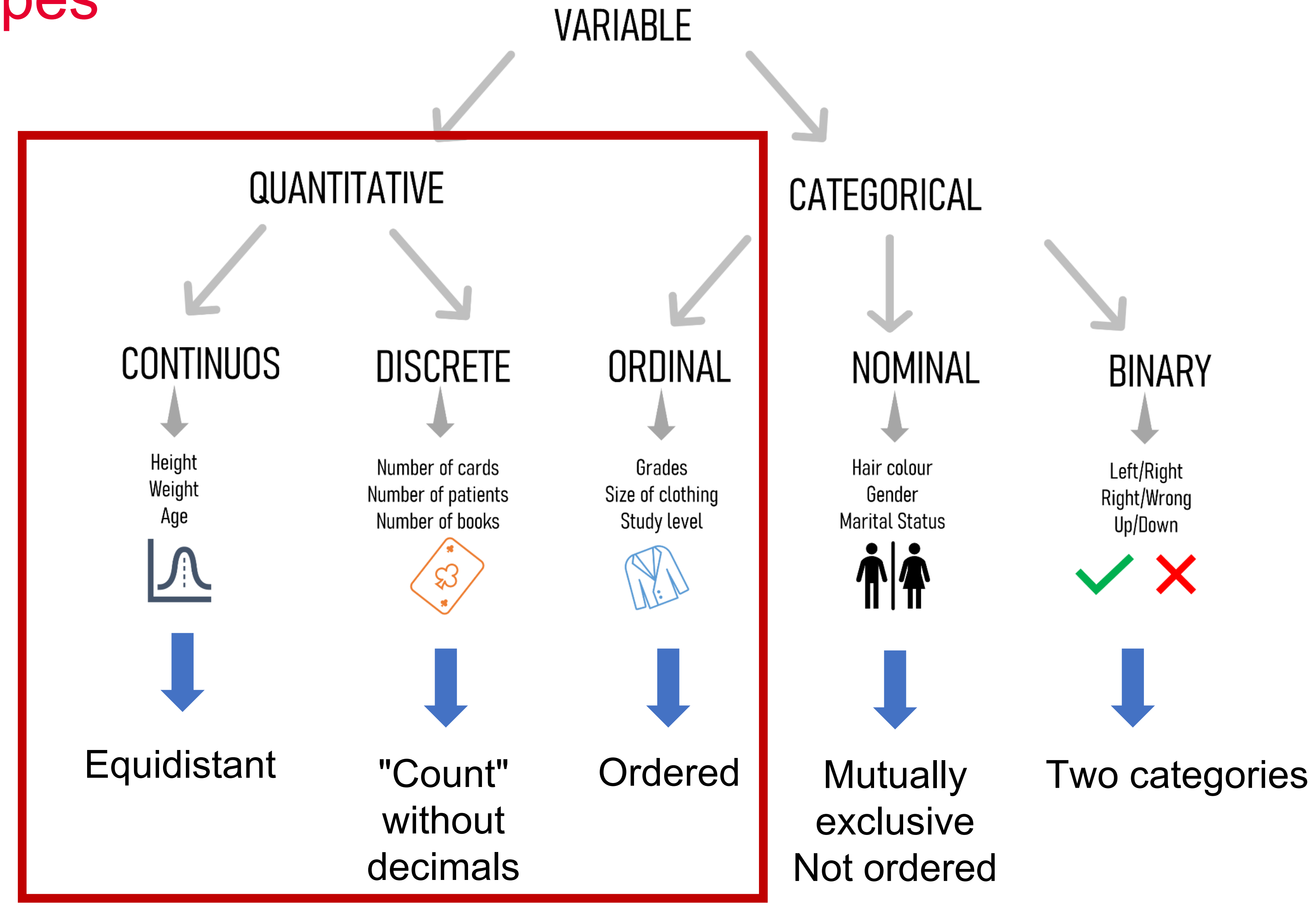
Objectives

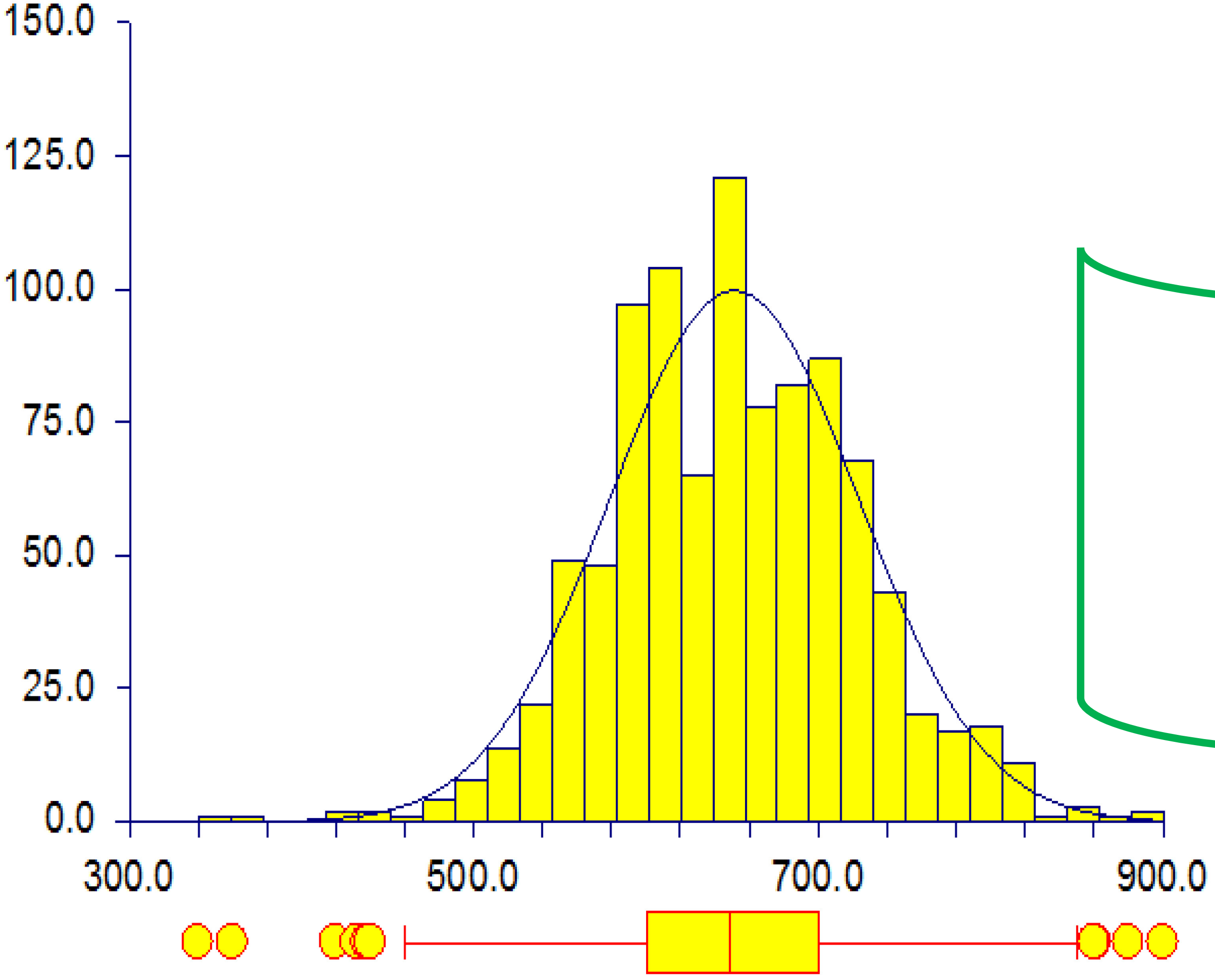
Identify non-Normal data visually and by testing for it

Compare non-Normal groups

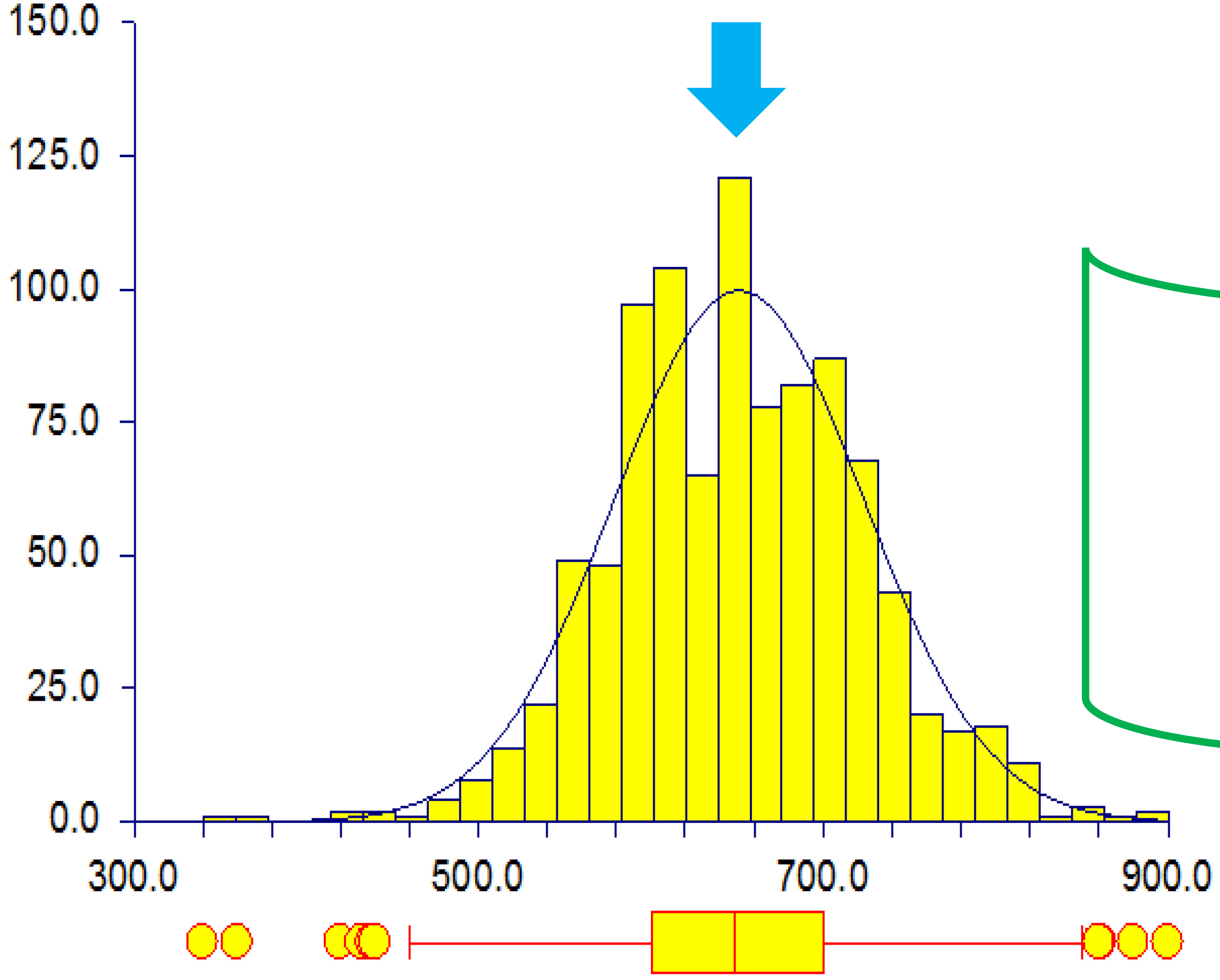
Multiple comparison corrections

Data Types

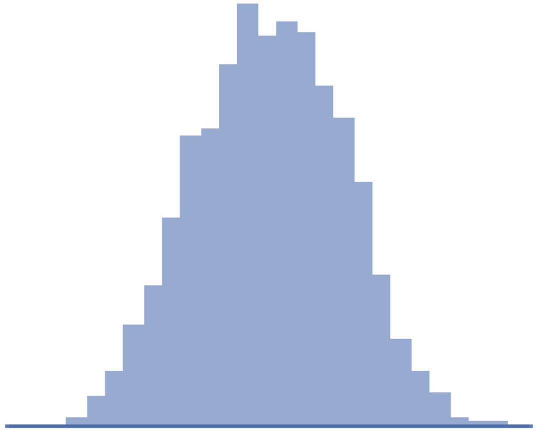

VARIABLE
QUANTITATIVE
CATEGORICAL
CONTINUOS
DISCRETE
ORDINAL
NOMINAL
BINARY
Height
Weight
Age
Number of cards
Number of patients
Number of books
Grades
Size of clothing
Study level
Hair colour
Gender
Marital Status
Left/Right
Right/Wrong
Up/Down
Equidistant
"Count" without decimals
Ordered
Mutually exclusive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



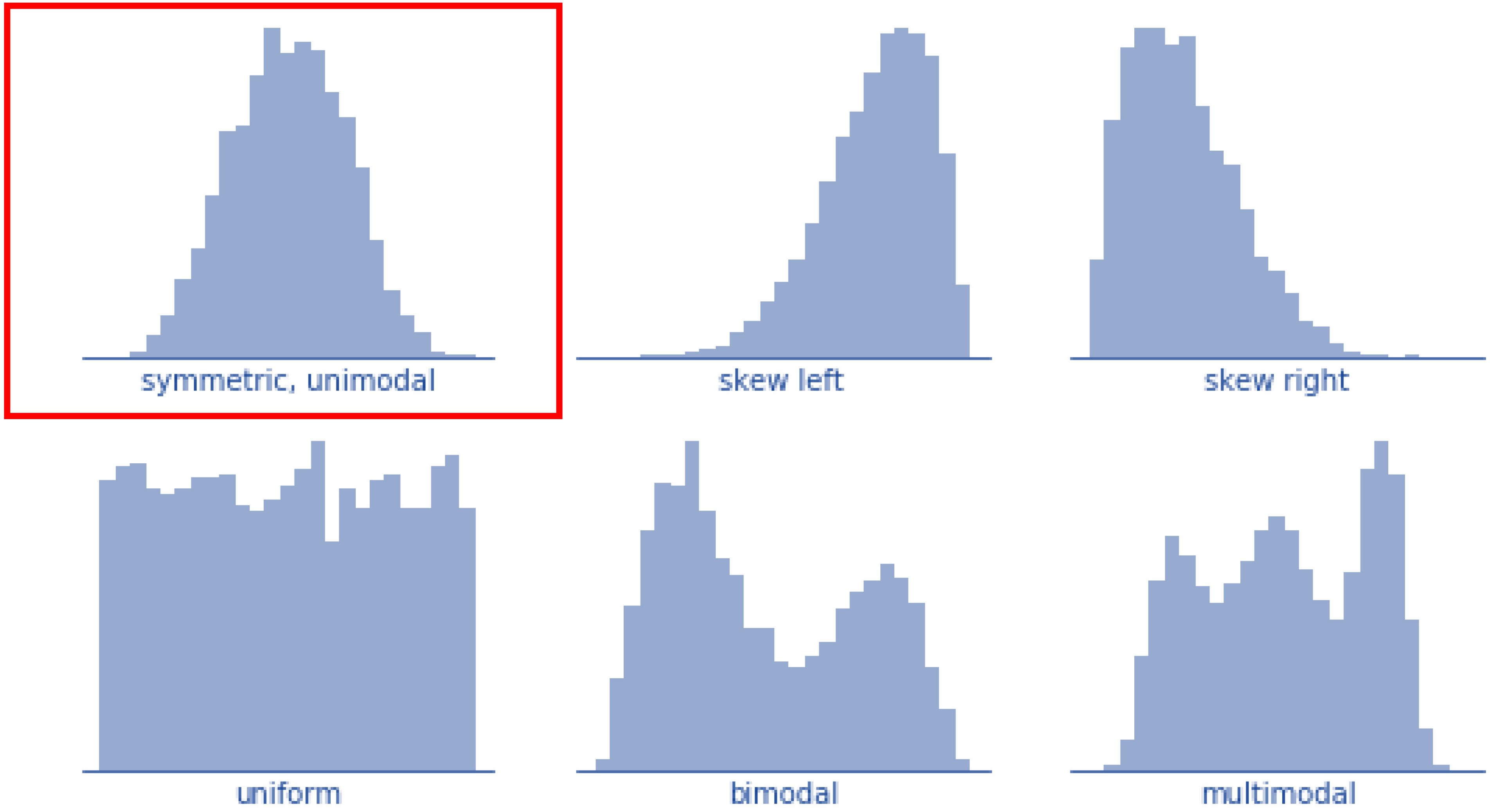
Which three statistics are the same?



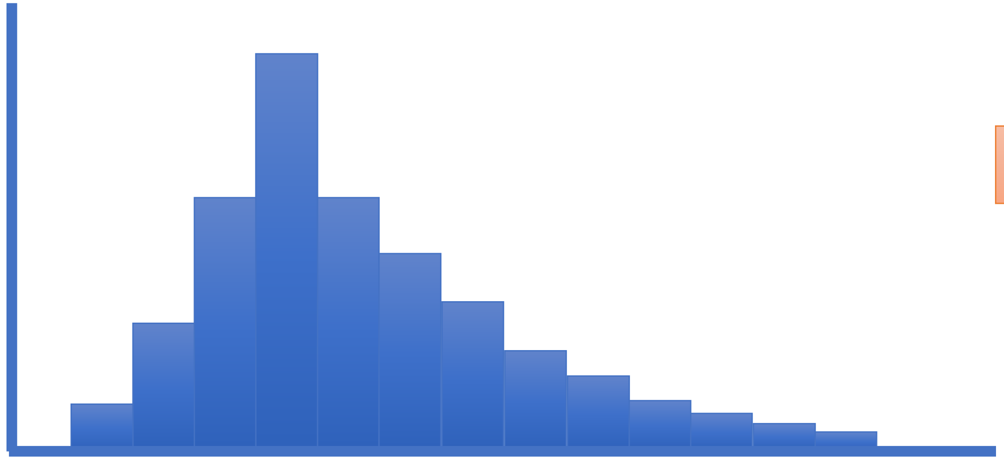
Which three statistics are the same?
mean = median = mode



symmetric, unimodal

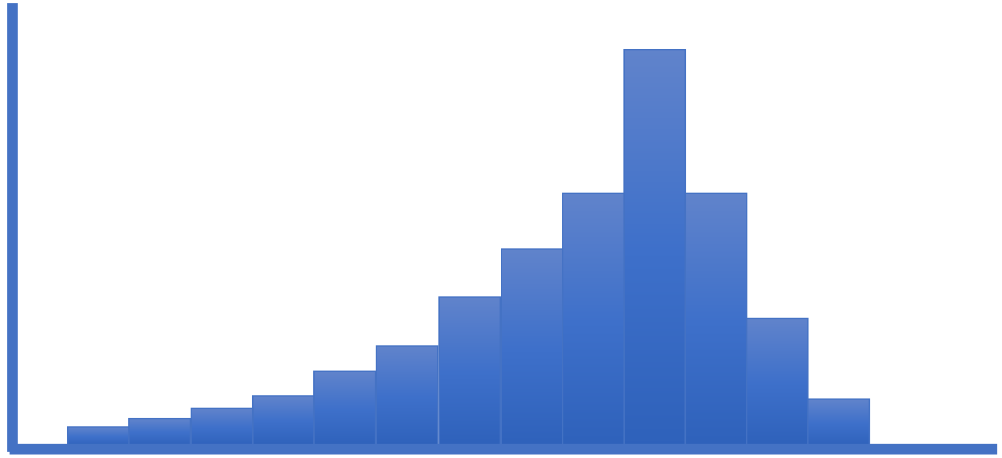

symmetric, unimodal
skew left
skew right
uniform
bimodal
multimodal

Skew



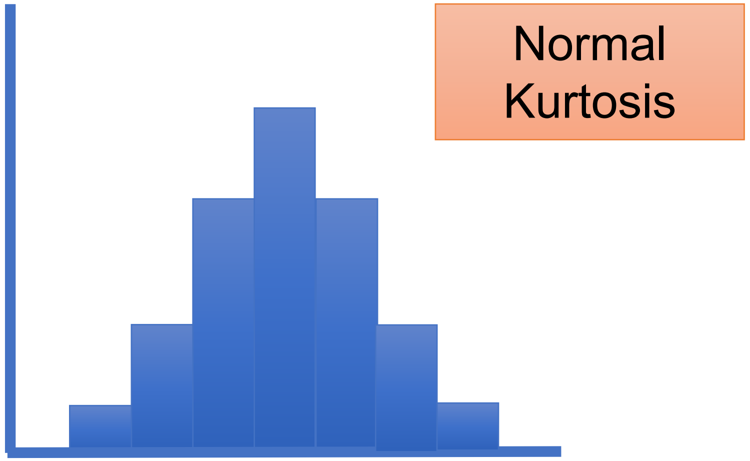
Positive Skew

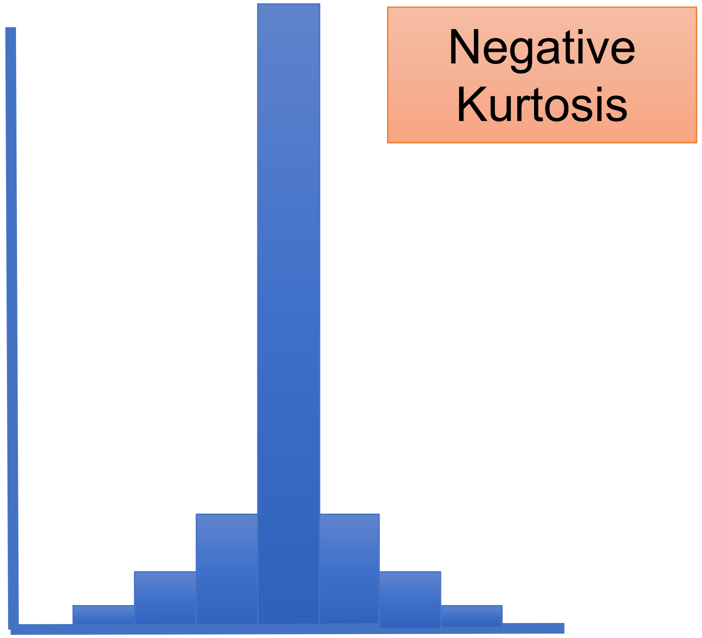
Normal distribution:
Skewness = 0 [-3,3]

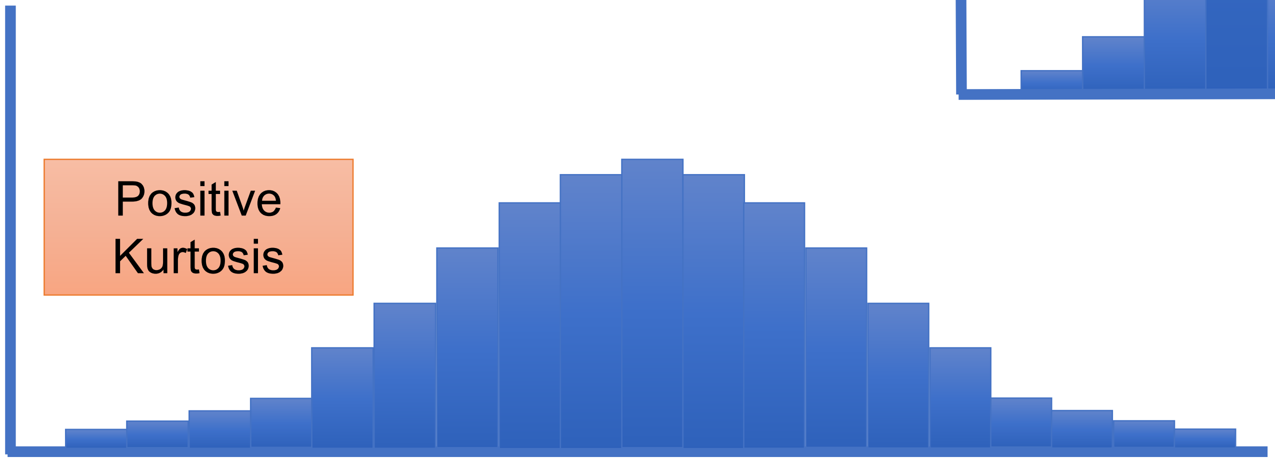


Negative Skew

Kurtosis

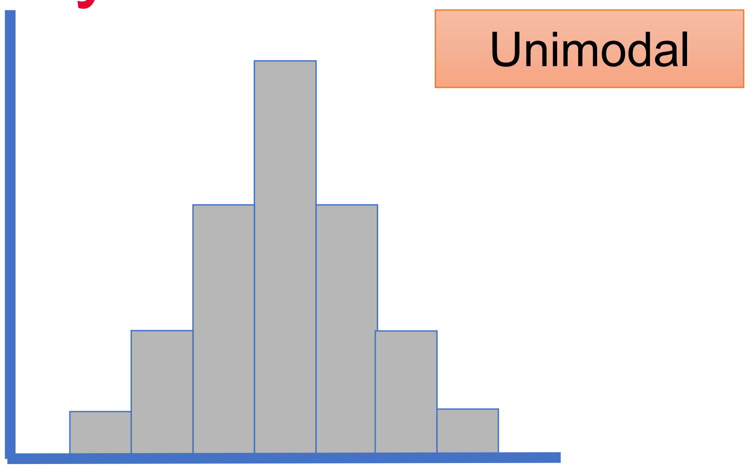


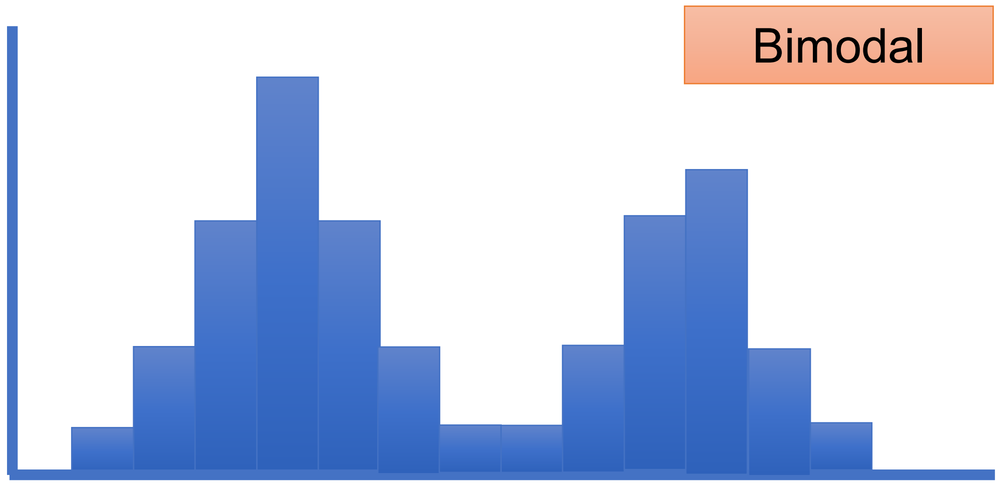


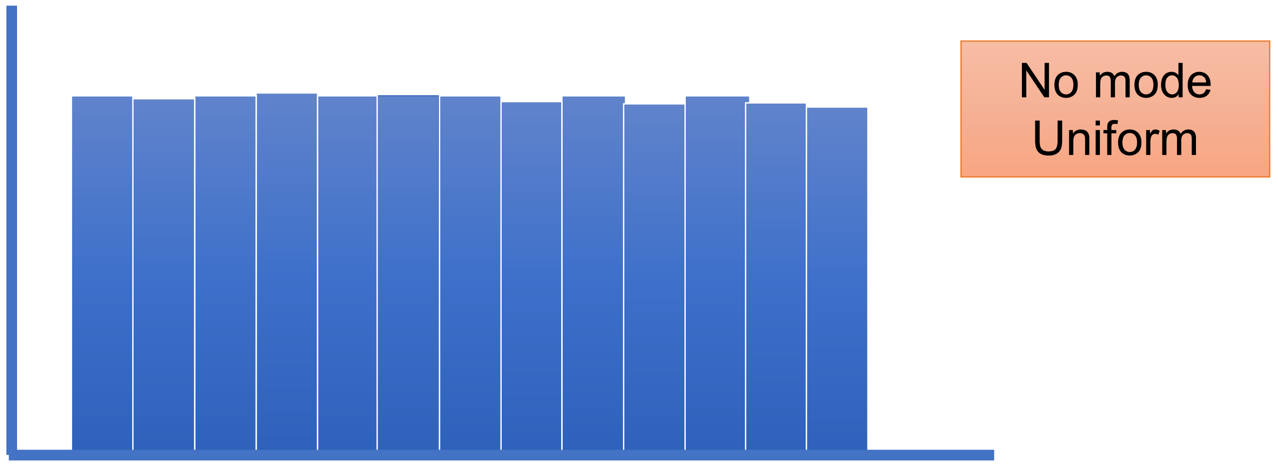


Normal distribution:
Kurtosis = 3 [2,4]

Modality

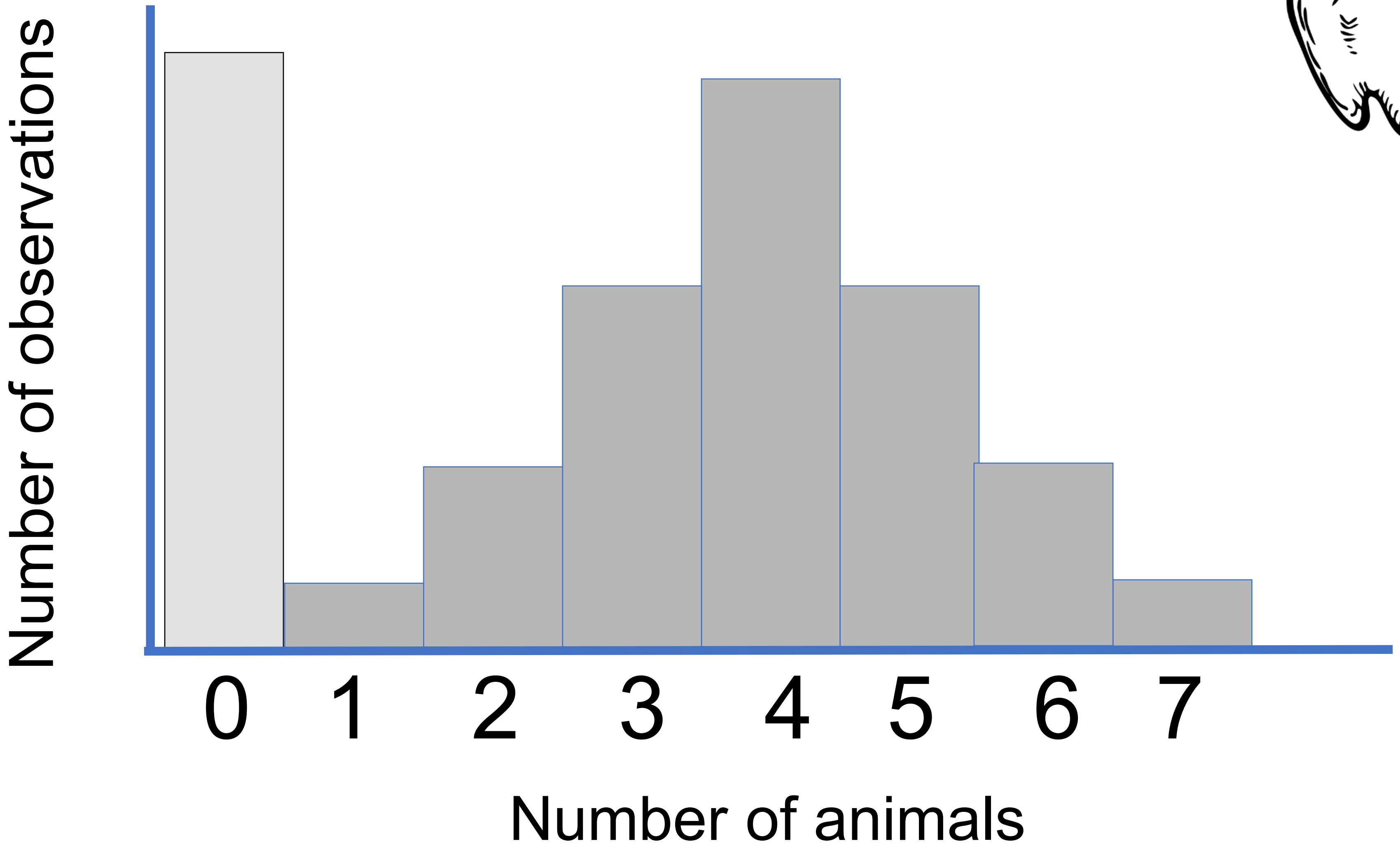


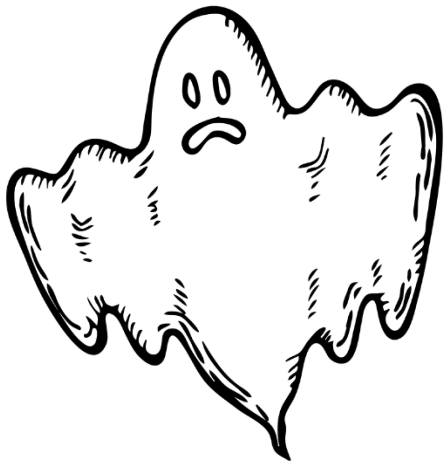




https://pixabay.com/de/illustrations/halloween-und-ghost-geister-4451903/

Zero Inflation



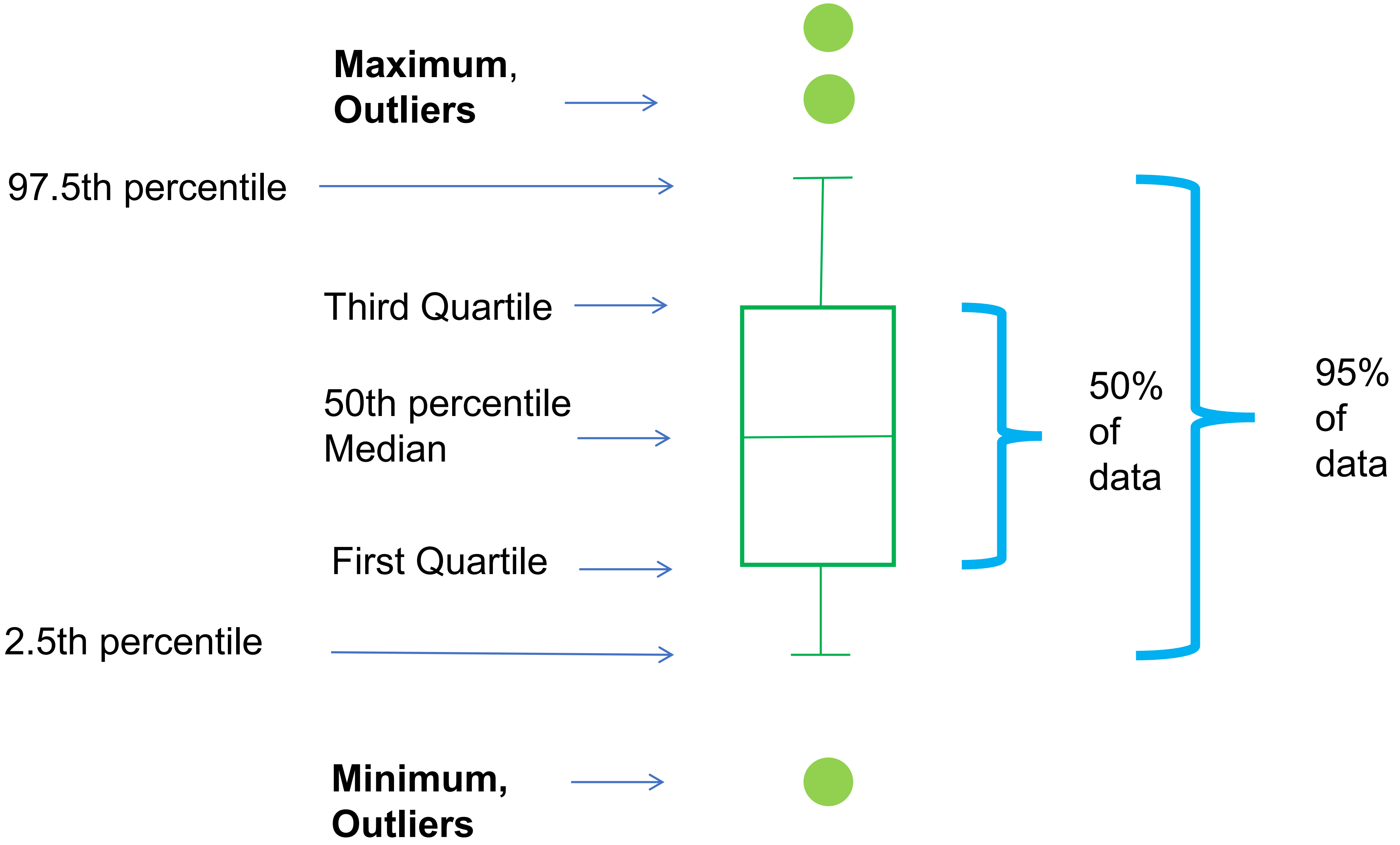


Who you Want to Call?

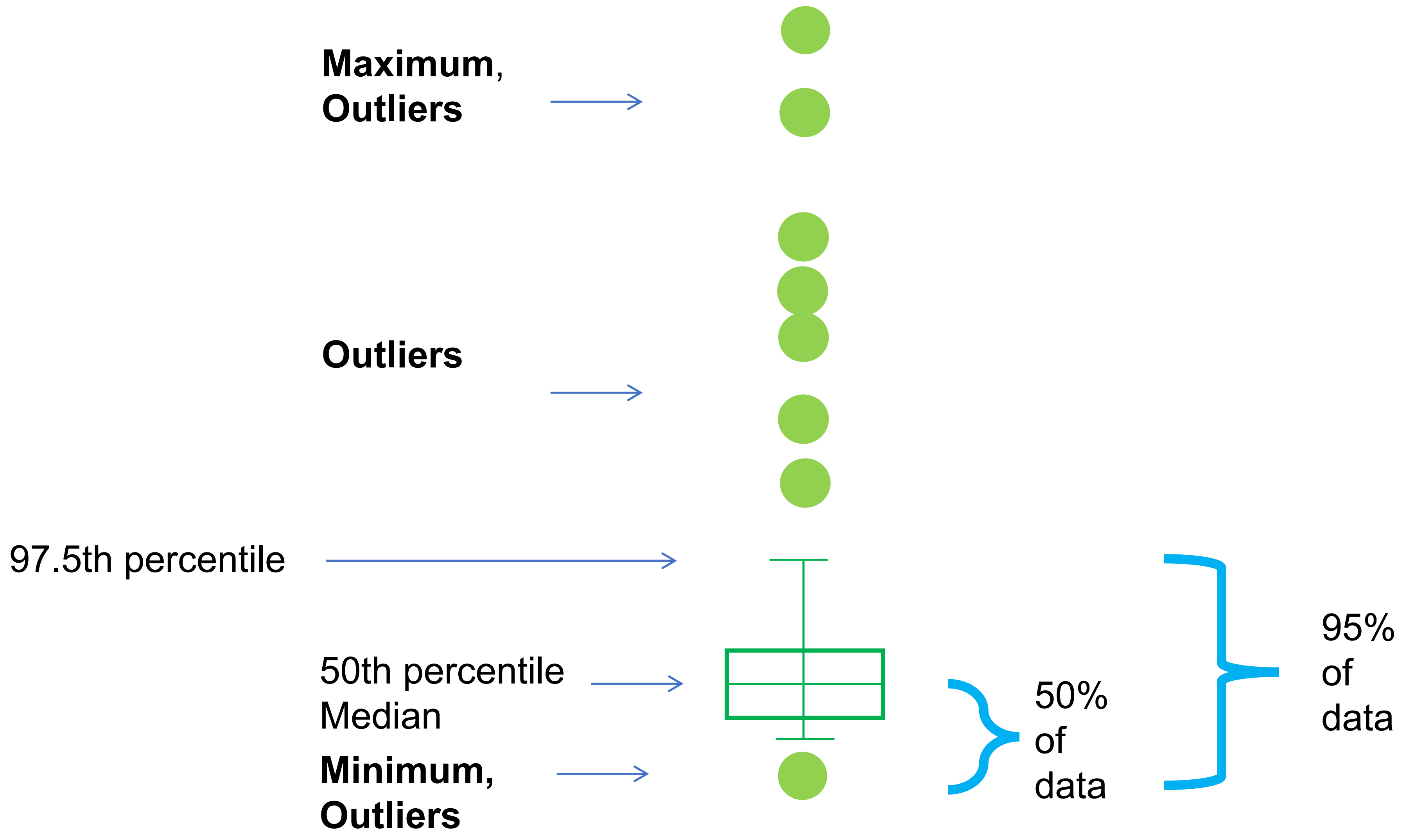
Who you Want to Call?

Percentiles!

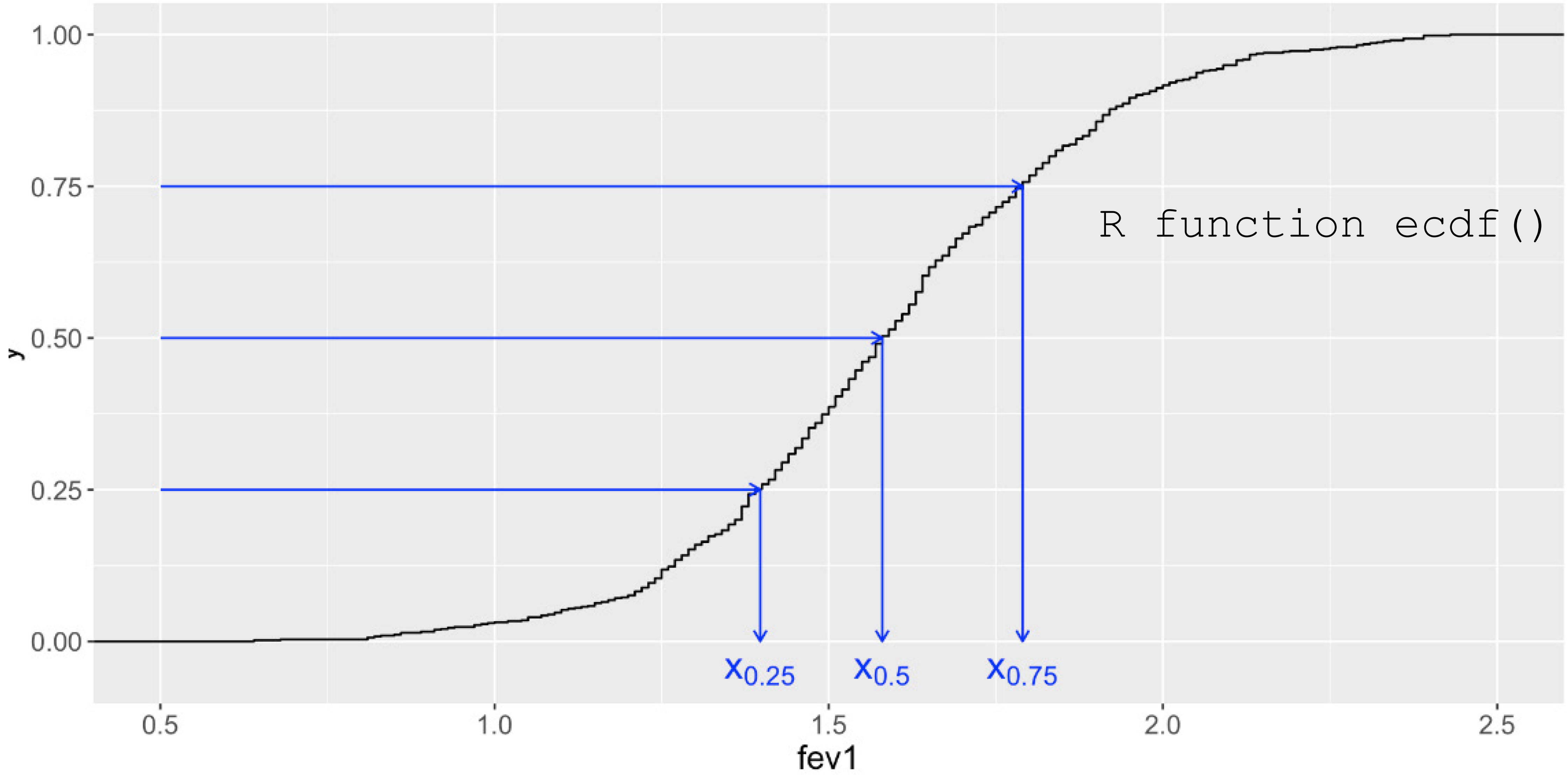
Box and Whiskers Plot (Normal)



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

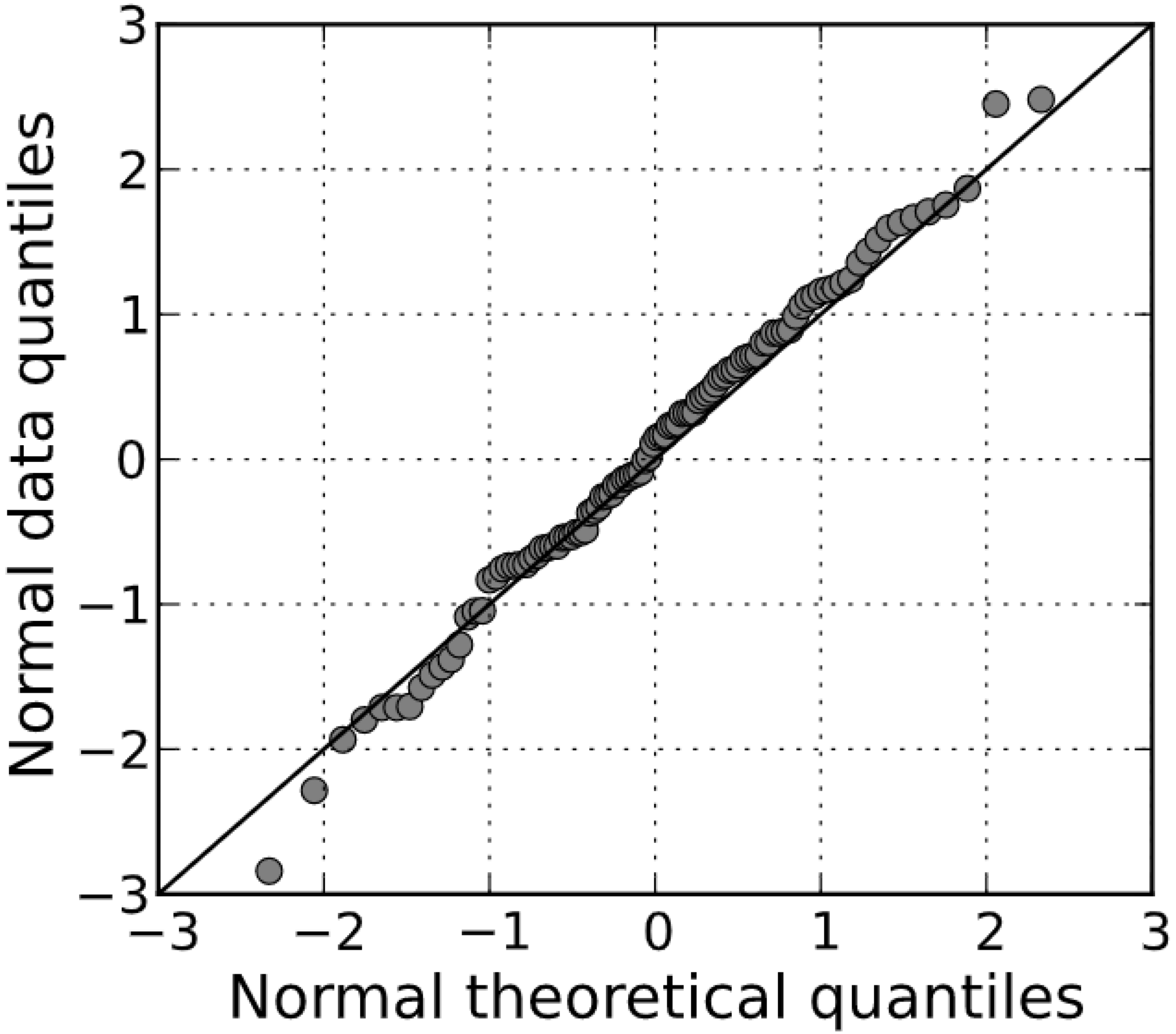
80th percentile is the value below which 80% of the data is found

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

Quantile Quantile Plot QQ-Plot

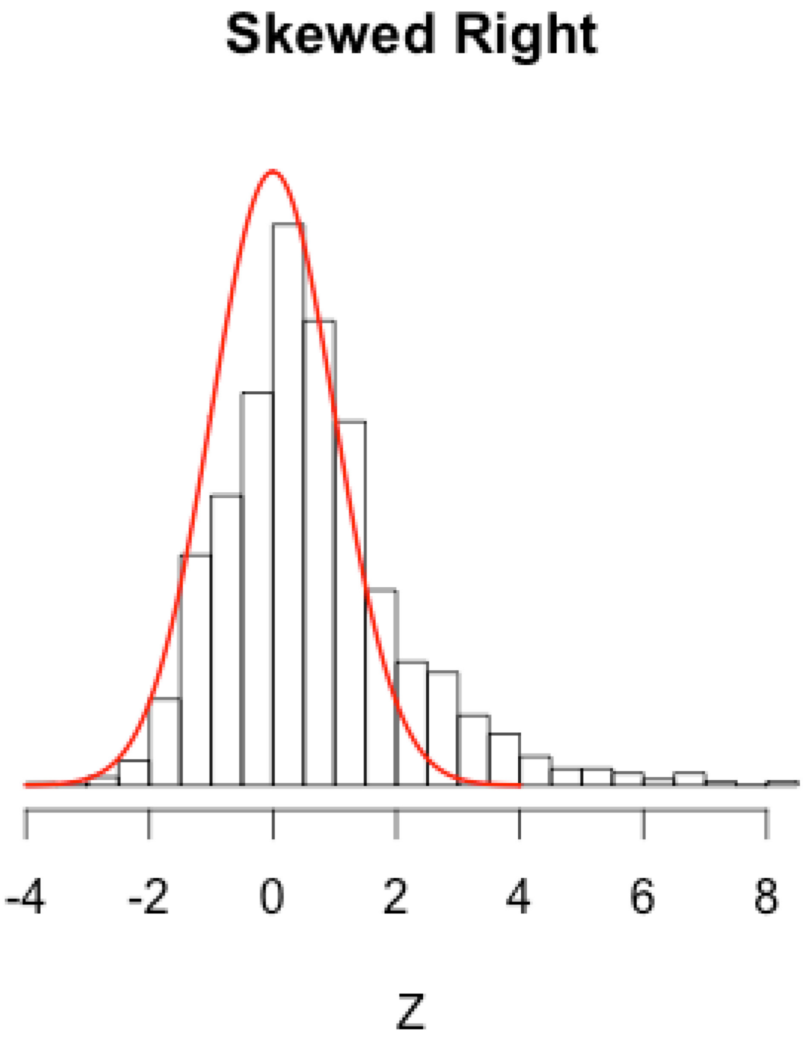
"a graphical method for comparing two probability distributions by plotting their quantiles against each other"

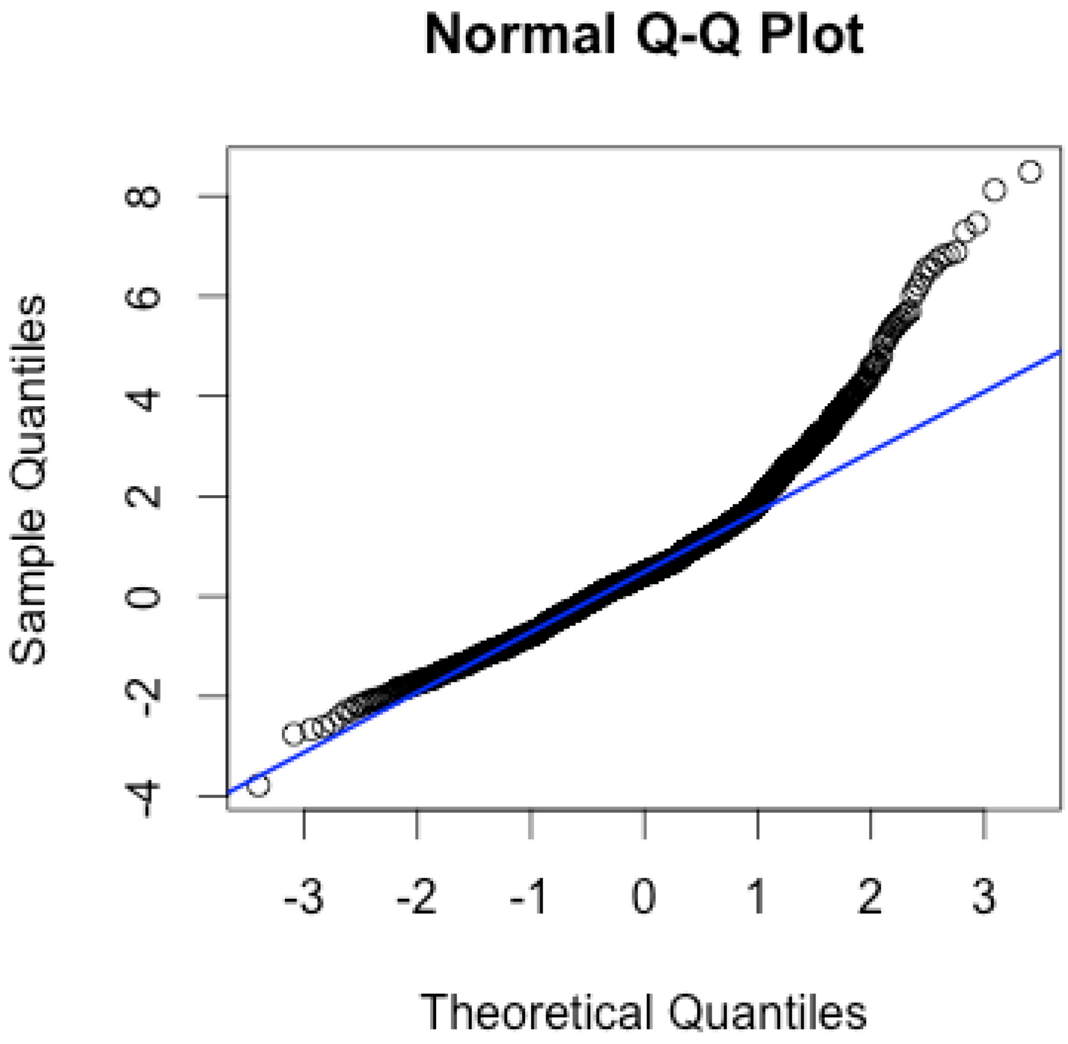
_*Wilk, M.B.; Gnanadesikan, R. (1968)*



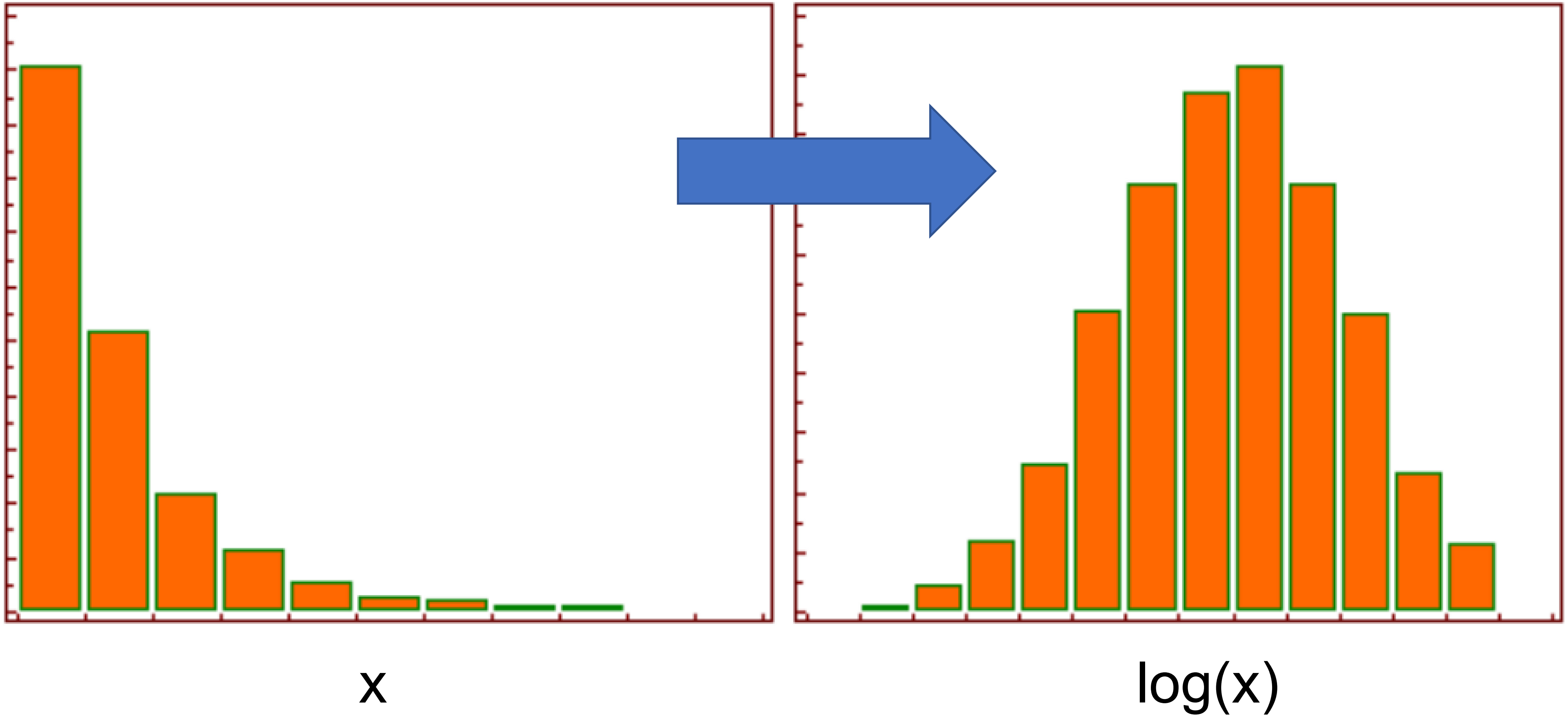
Skewed Right

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





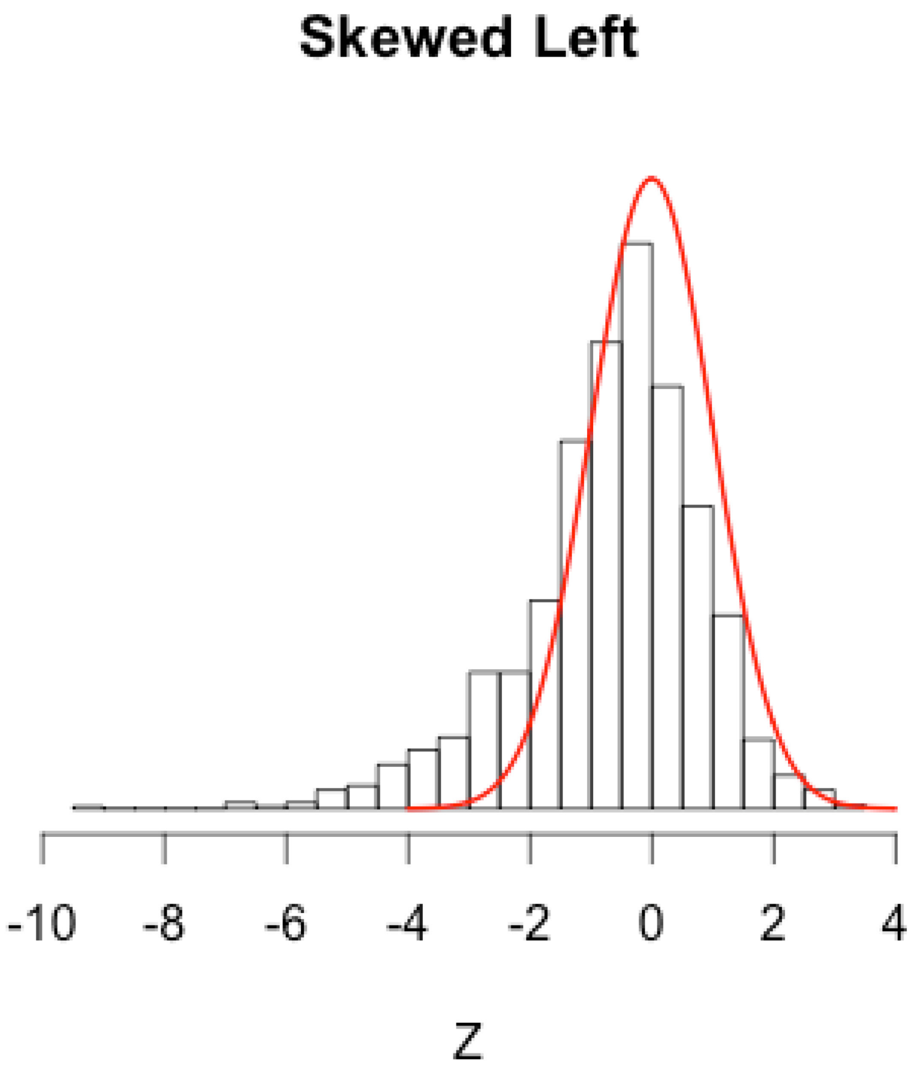
Right Skew and Log Transformation

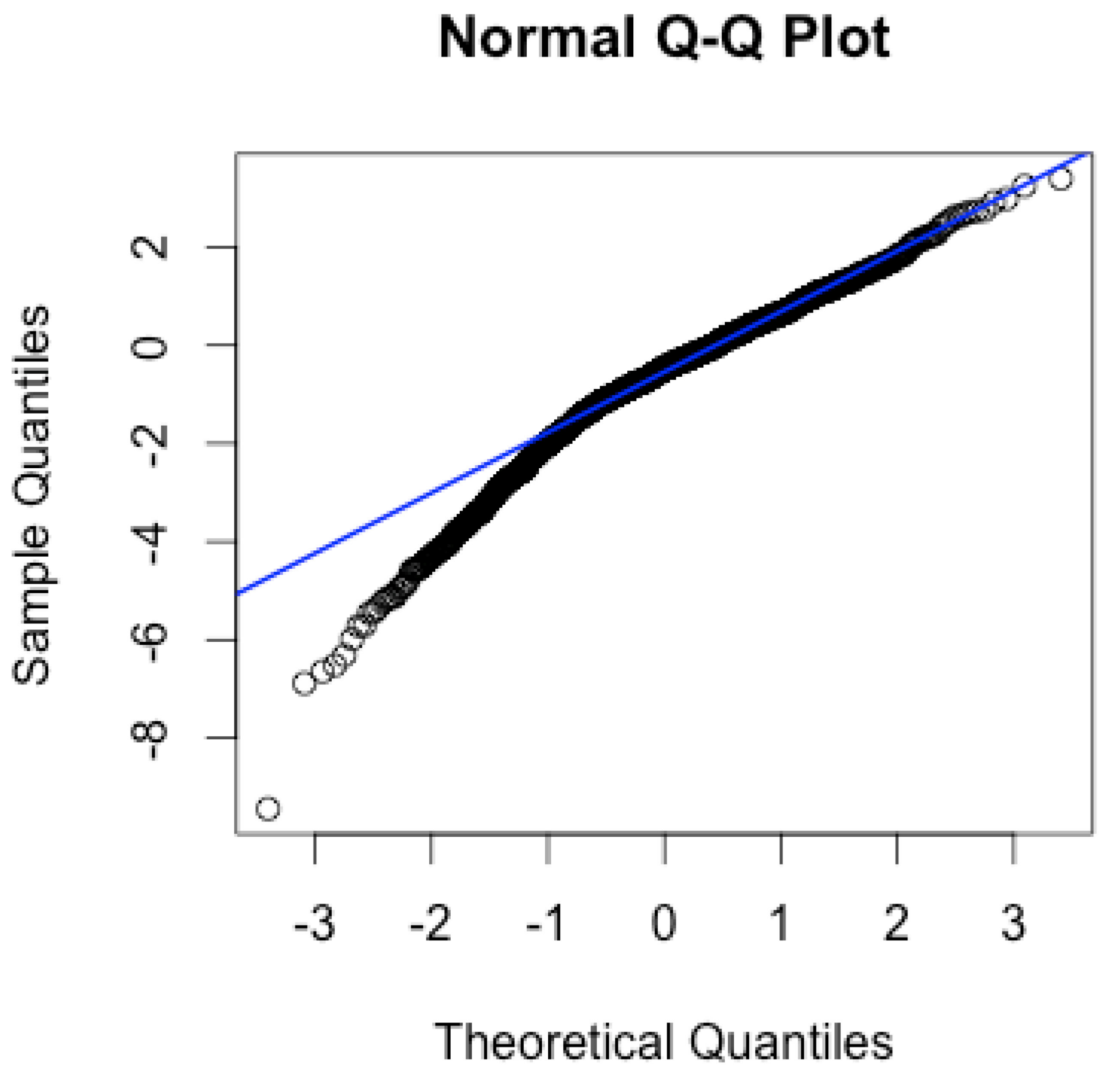


https://www.medcalc.org/manual/log-transformation.php

Skewed Left

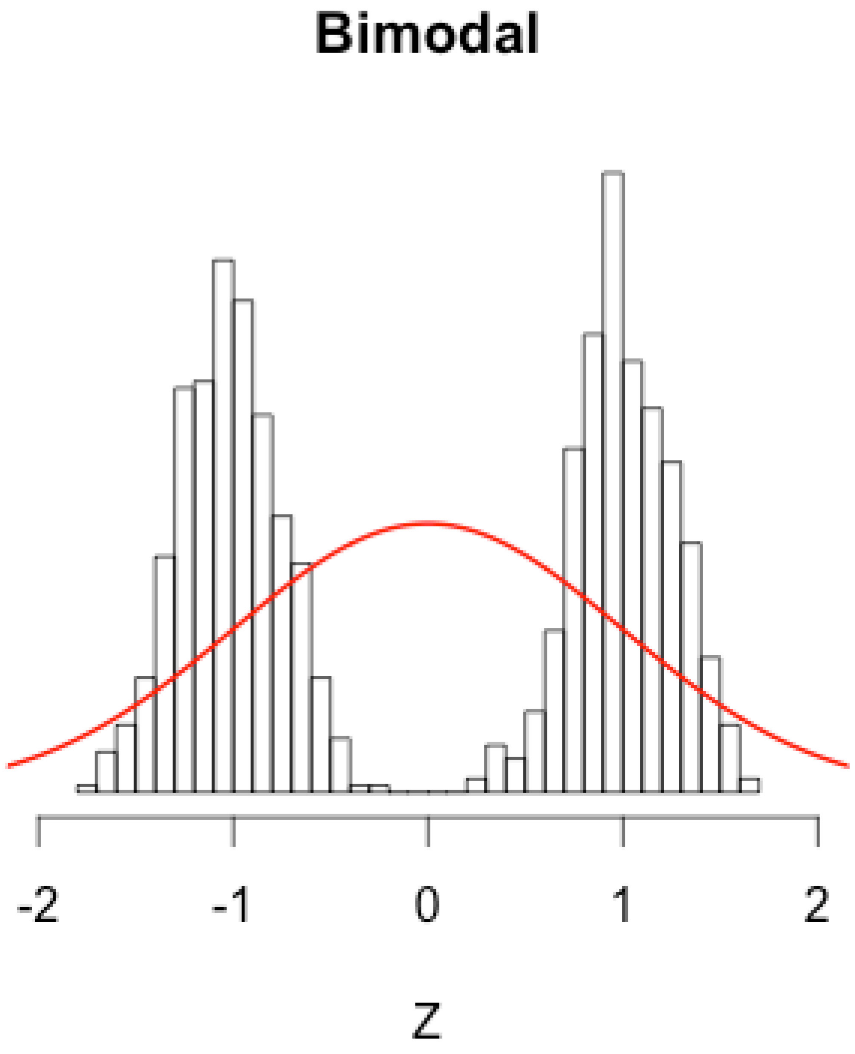
https://seankross.com/2016/02/29/A-Q-Q-Plot-Dissection-Kit.html

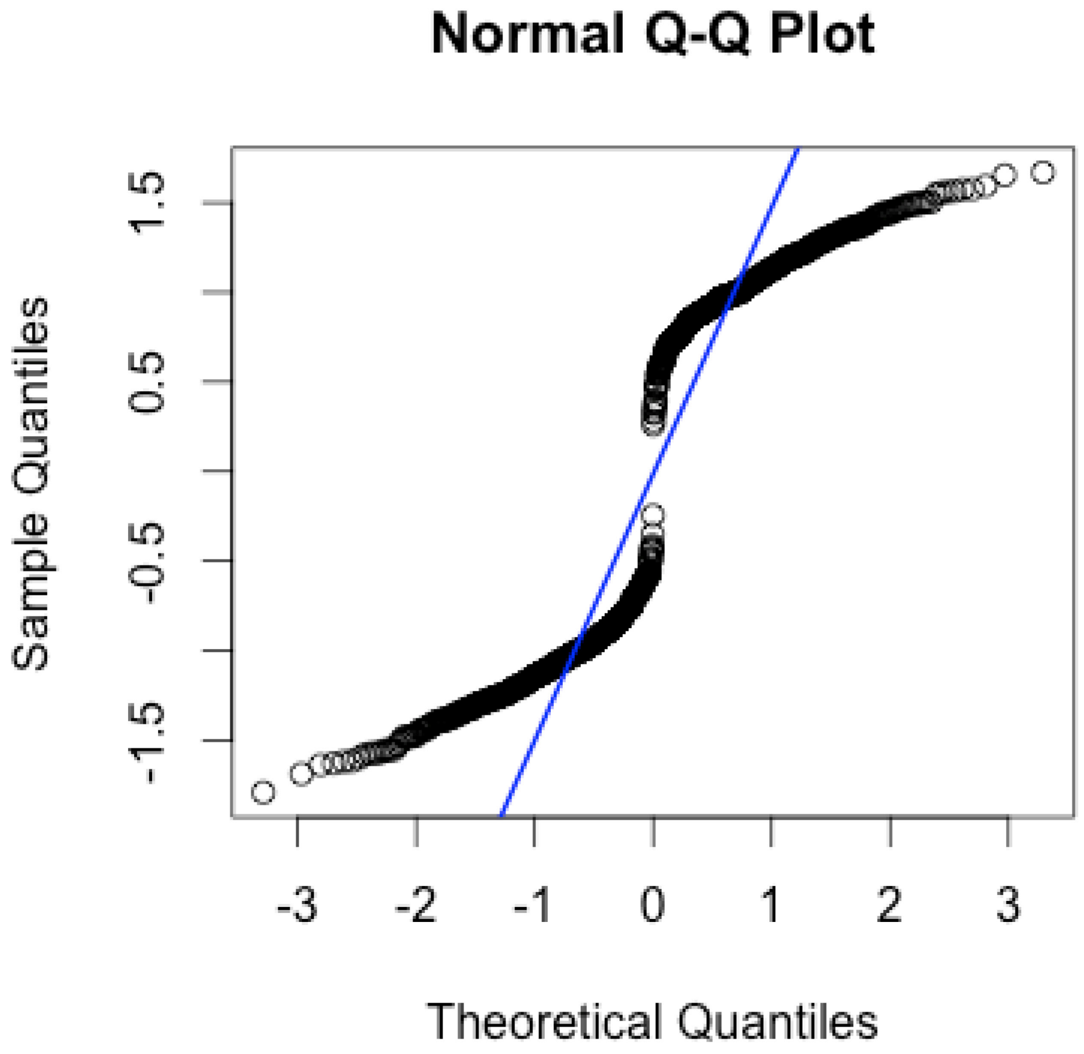




Bimodal

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



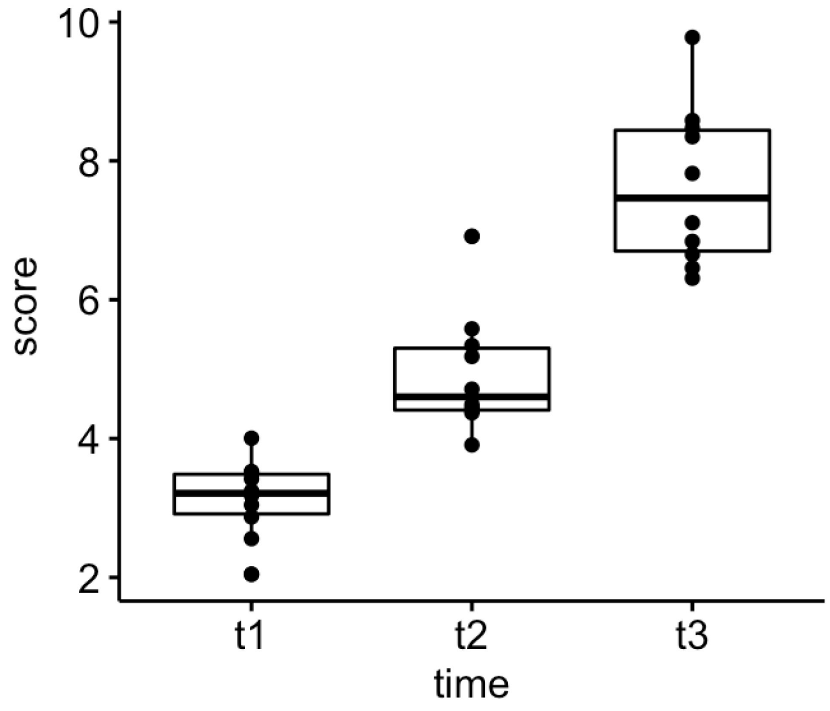


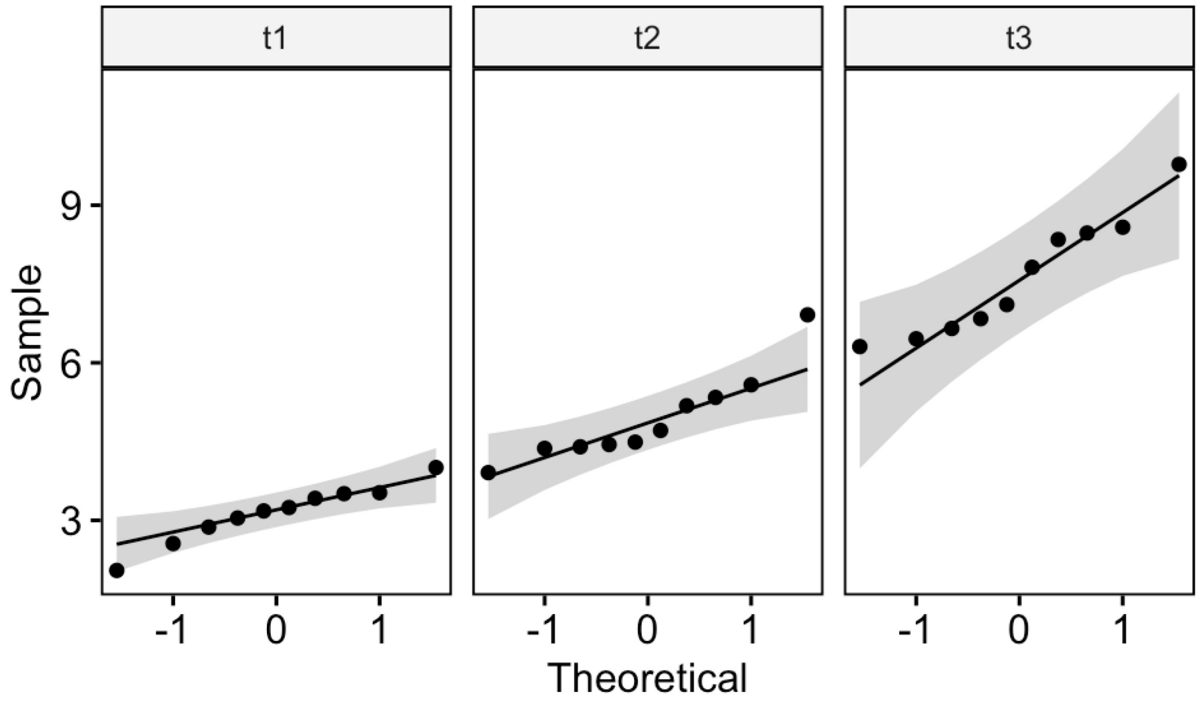
Normality Tests – Each Group Separately

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

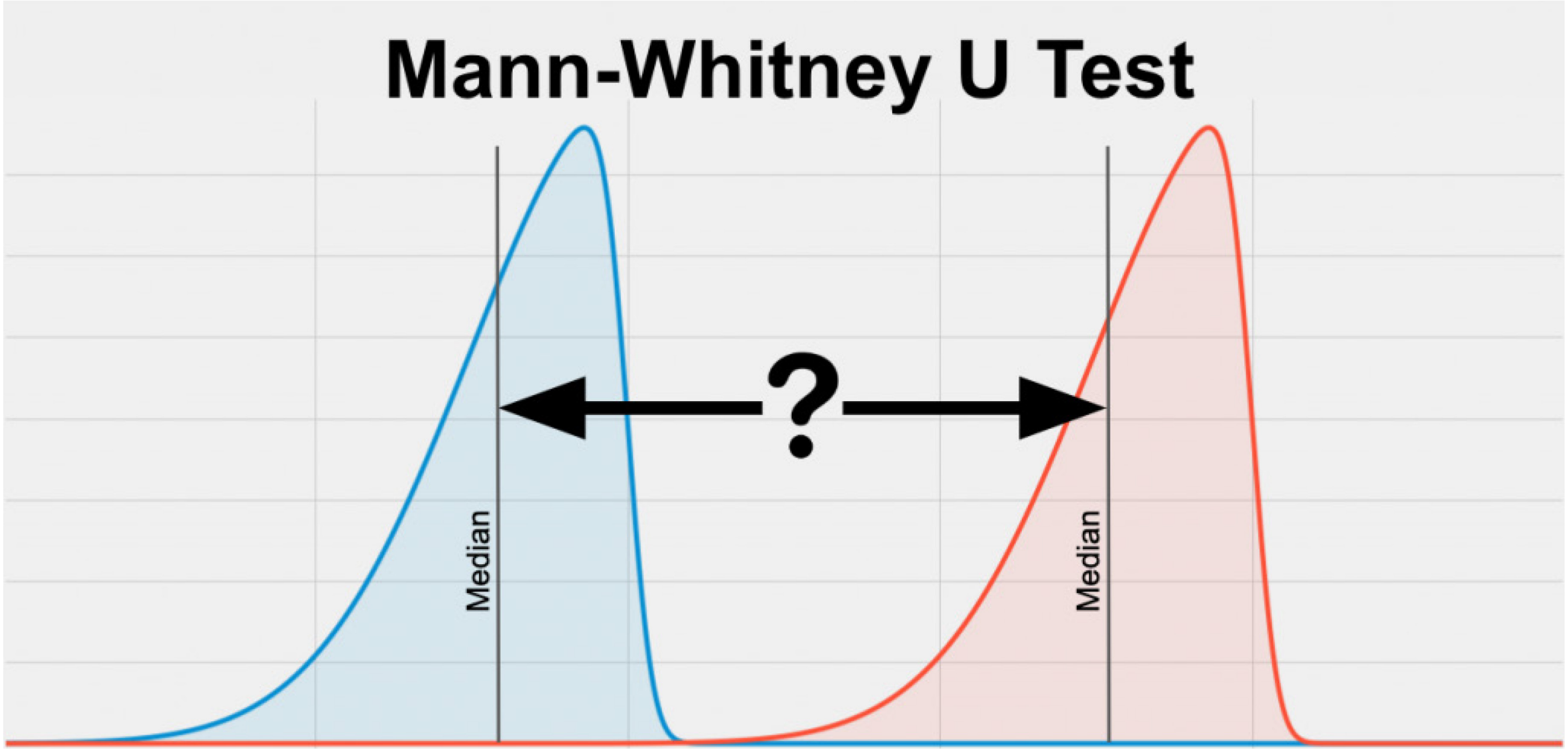
```
x <- rnorm(50) ; y <- runif(30)
# Do x and y come from the same distribution?
ks.test(x, y)
```

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/

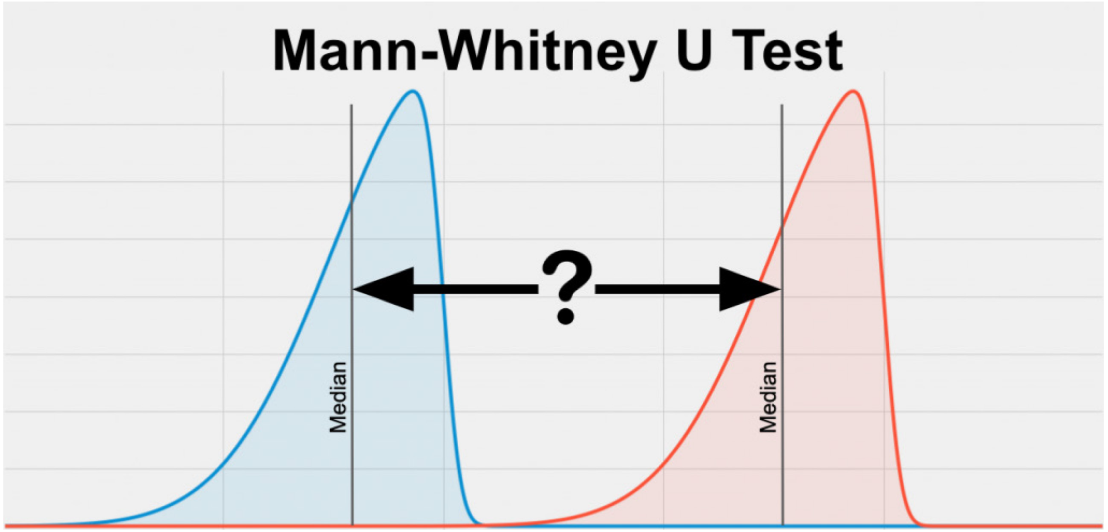


Given two identically shaped and scaled distributions,

Ho: are the medians different?

Mann–Whitney U Test

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



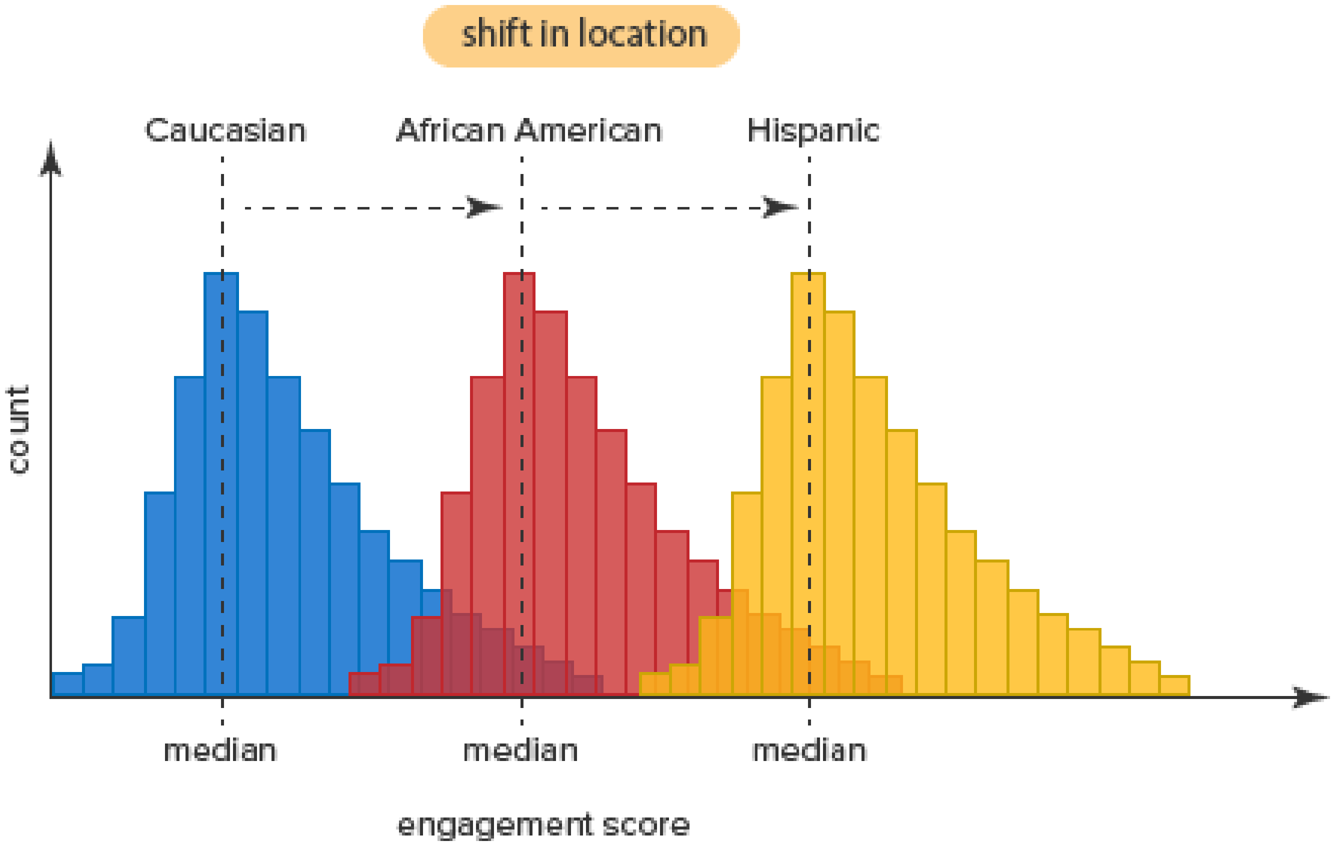
Wilcoxon 1945

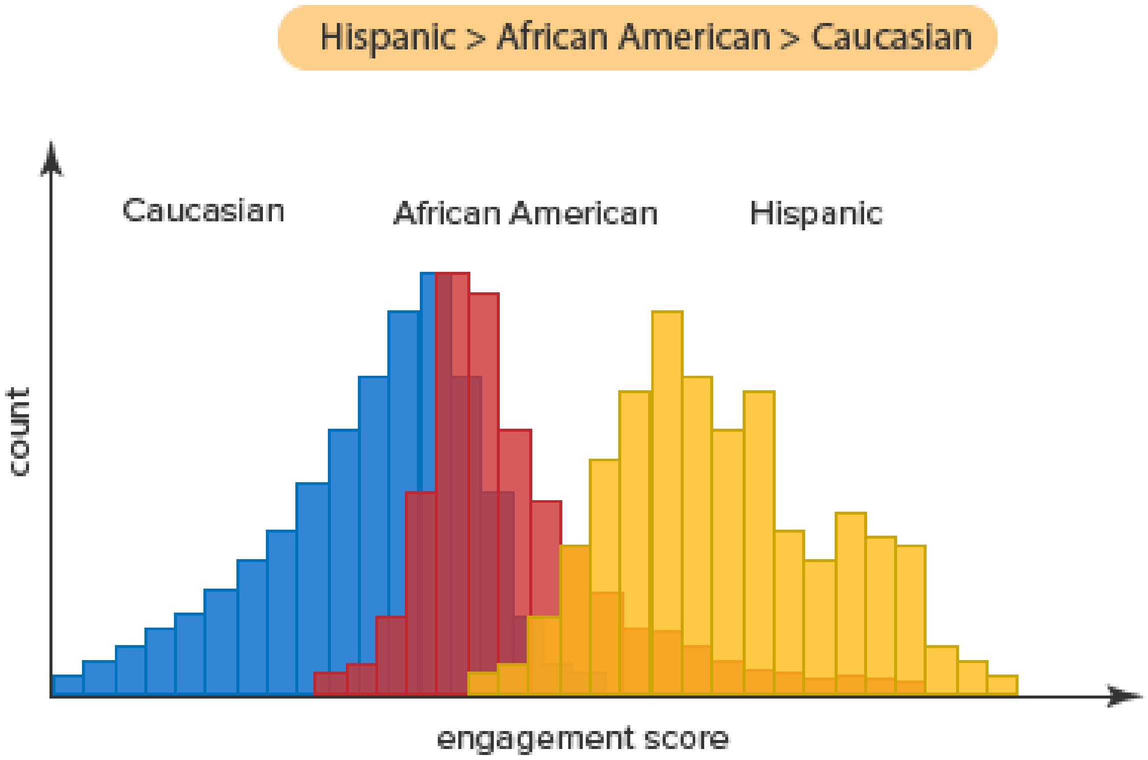
Mann & Whitney 1947

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

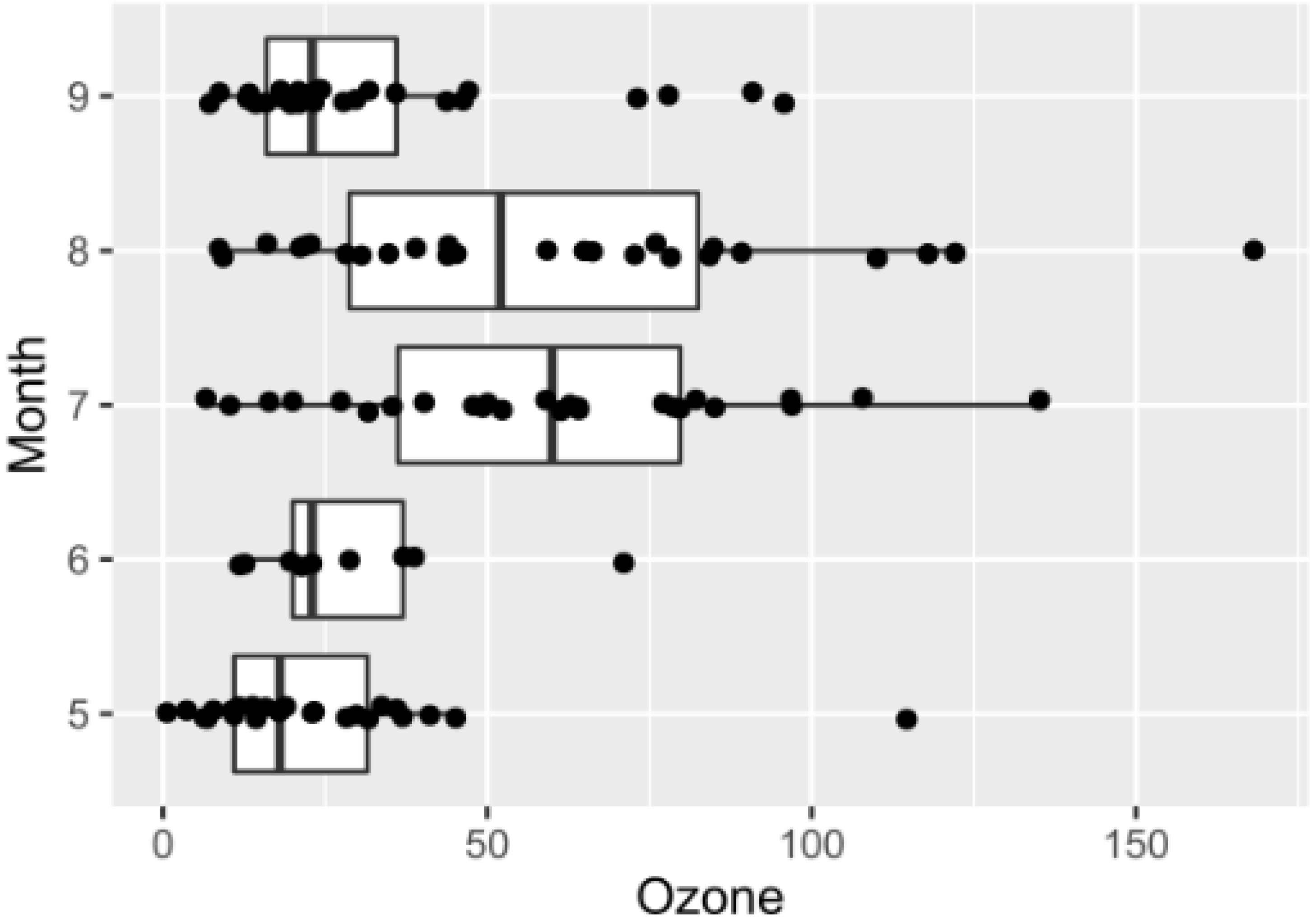
W. H. Kruskal, W. A. Wallis 1952

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

<u>Reduce</u> the level of significance (alfa)
<u>for each comparison</u>
So that the probability of getting
<u>one wrong</u> 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()`)

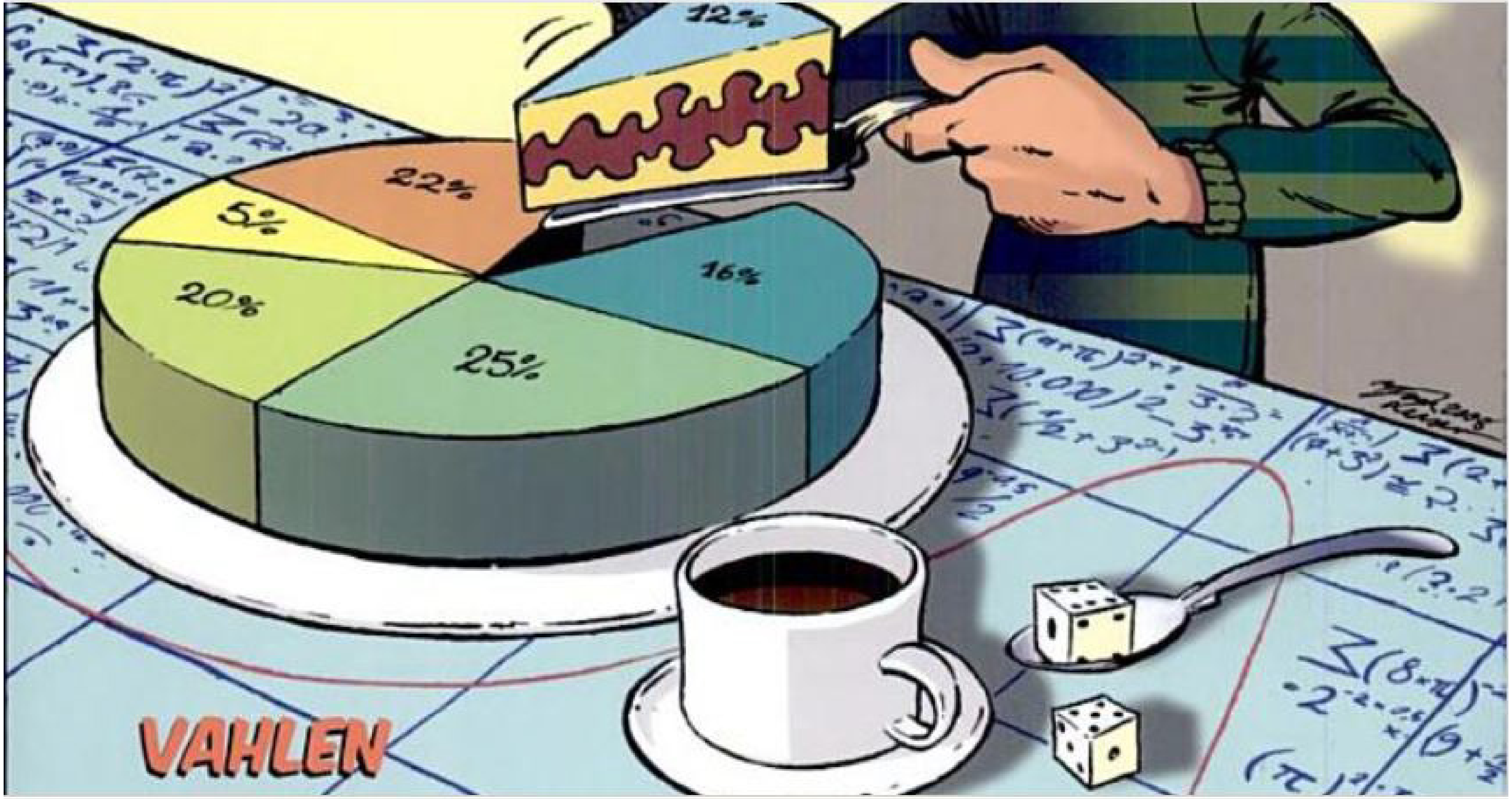
- Tests that include multiple comparison corrections (pairwise comparisons with Bonferroni correction)

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