Concise but significant* handbook to M5 statistics

Hotel Management School Maastricht

* p = 0.01

What measurement level?

Variable
data

Categorical
ABC
Labels, making groups, distinct categories

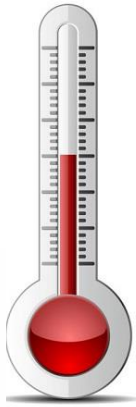
Nominal
Categories are equal (= and ≠)

Ordinal
There is an order to the categories (> and <)

Continuous
123
Measured, scale, ranges, values

Interval
Equal distance between the values (+ and -)

Ratio
*There is an absolute 0 point (* and /)*



Which chart to use?

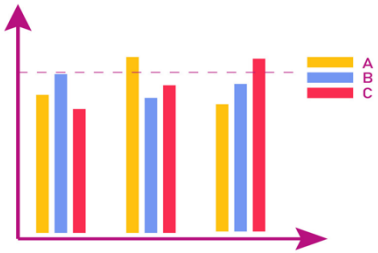
Variable
data

Categorical **ABC**
Labels, making groups, distinct categories

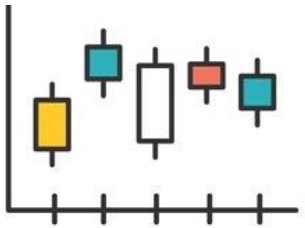
Continuous
123
Measured, scale, ranges, values

1 variable
Bar chart or pie chart

2 variables
Clustered or stacked bar chart



1 of each
Various options: Bar chart, error bar plot, box chart



1 variable
Histogram

2 variables
Scatterplot, or line graph (over time)

Making hypotheses

A research hypothesis is a concise statement about the expected result of an experiment or project. Two types exist:

- the ***null hypothesis*** H_0 Assumed to be factual, normal, ´what there is´ in the current world.
- the ***alternative*** **hypothesis** H_1 / H_a What could or might be, what we look for out of the ordinary.

Example: A researcher is testing the efficacy of a new drug. Normally, a random pill or treatment doesn´t help, so normal reality H_0 is: the drug has no benefits and doesn´t work. If the drug doesn´t make a difference, there are no differences to expect in the mean measurements of test group and control group, so:

H_0 = the drug has no effect ALSO $\mu_{test} = \mu_{control}$

In an alternative world, however, the new drug might do something, so H_1 is ´the drug makes a difference´:

H_1 = the drug does something ALSO $\mu_{test} \neq \mu_{control}$

The hypotheses above do not have any direction. They simply deal with a possible difference in μ´s. This test is **two-sided**: any result could be more or less, better or worse. However, we could also formulate a hypothesis with a direction, in which case we could say: the drug improves health. Now, we test **one-sided**: the result of the test group should be higher/better.

H_1 = the drug has a positive effect ALSO $\mu_{test} > \mu_{control}$

Hypotheses, μ, p, and distribution

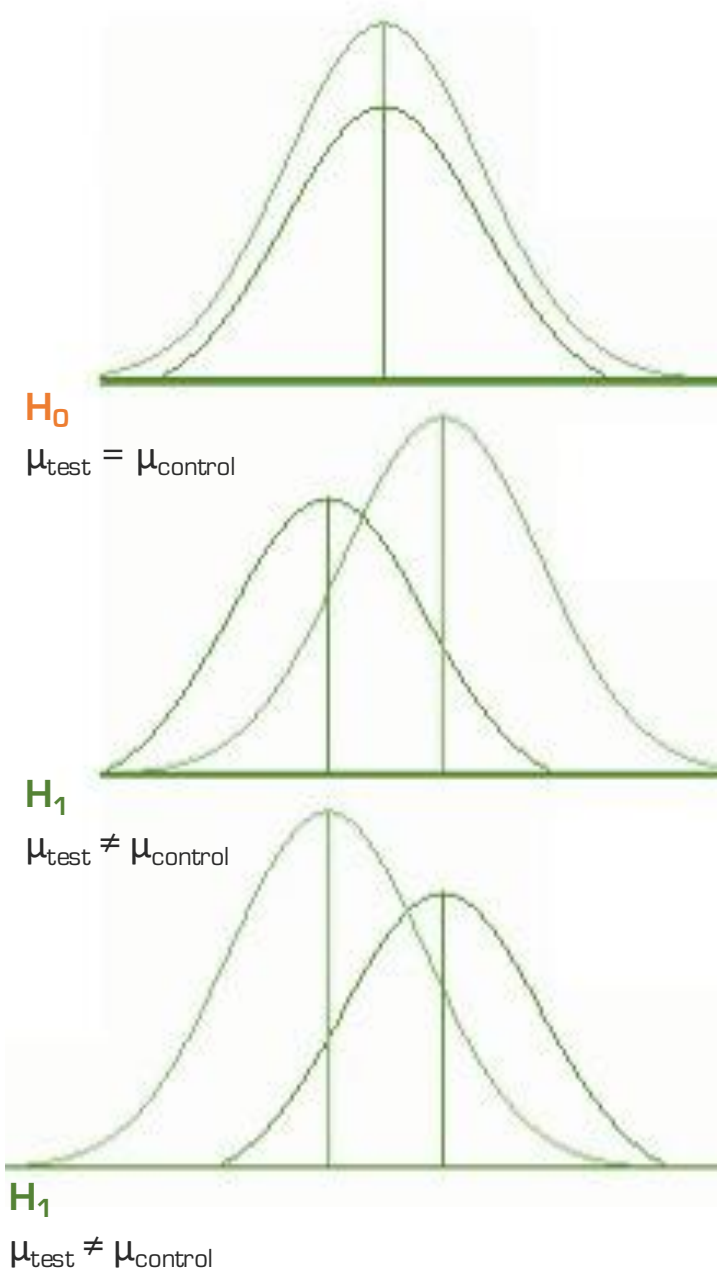
Statistical tests – simply put – calculate the probability that the H_0 is true – that is: the probability that you would find your study´s observations in the data if H_0 would be true.

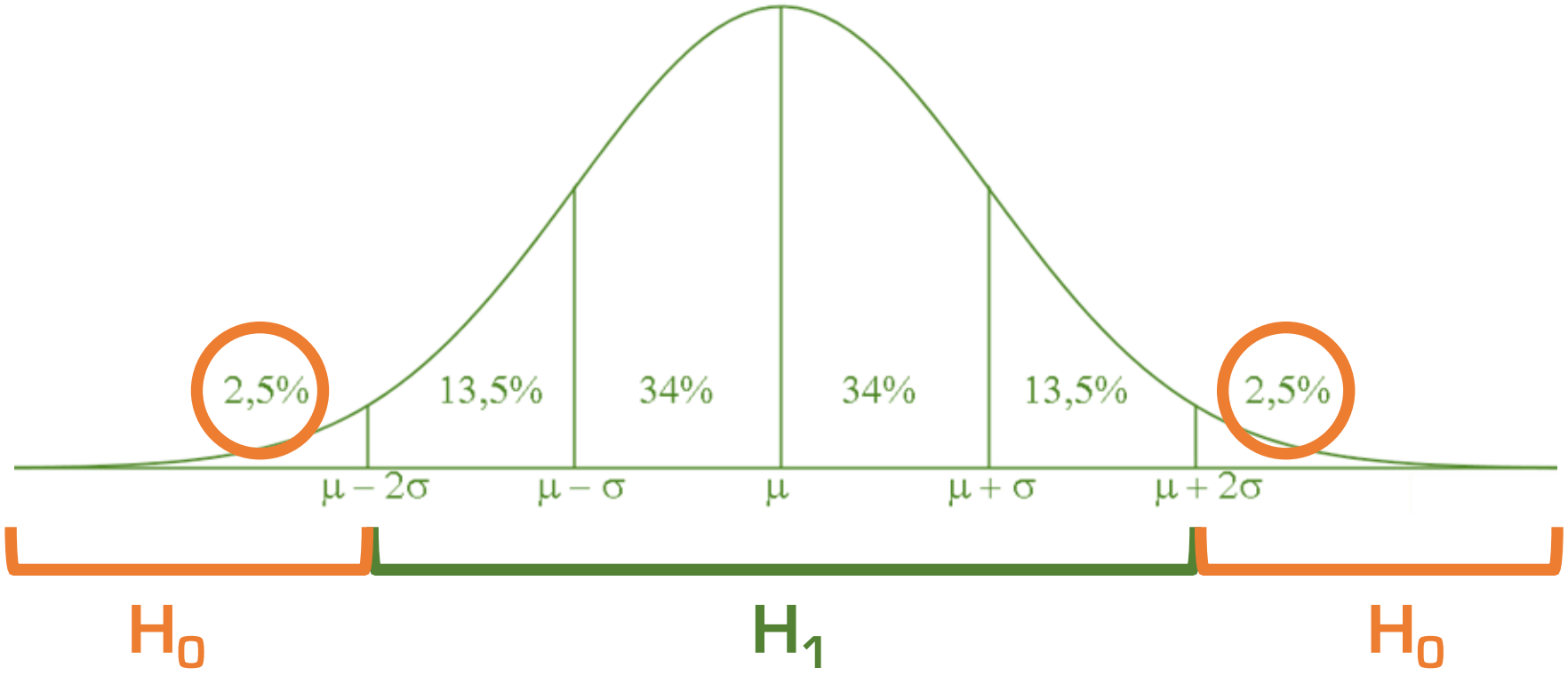
Why? Because if we want there to be something ´special and different´ for H_1, then we **do** want the data of the groups to differ enough and we **do not** want too many observations that suggest no difference.

Normally, we use a critical p-value of 0.05, which means that we want at least 95% of your observed data to fall within the confidence interval of 95% to assume a significant difference and thus assume H_1. That also means allowing a chance of max. 5% for the possibility that you would find your observed data anyway and that the data are not that different at all, which would mean that we would stick to H_0.

NOTE: When testing for differences in means, we use ´the Greek m´: μ. When testing for differences in proportions (parts, shares), we use ´the Greek p´: π. Hypotheses are formulated using either μ or π.

If the probability p of H_0 is below the critical value p (e.g. 0.05), we reject H_0 and accept H_1.





If the resulted probability p is below the chosen critical value p (e.g. 0.05), then
there are limited cases in the H_0 region and we can reject H_0
&
we assume there are enough cases in the H_1 region to assume a difference in μ´s and support H_1, so we accept H_1.

! **If you test one-sided, then the test´s p-value has to be divided by 2 (you can´t ´use´ the other side of the distribution). In Jamovi you can tick boxes for testing either two-sided or one-sided, in SPSS the results table reports both.**

Types of error

If you use statistics to make decisions on hypotheses, the chance exists that you draw false conclusions. Options here are:

Type 1-error, or a false positive

The rejection of the H_0 when it is actually true and should be assumed

Thus also meaning: accepting H_1 while you shouldn´t.

Type 2-error, or a false negative

Not rejecting H_0 while it is actually false and should be rejected

Thus also meaning: not accepting H_1 while you actually should.

Type 3-error, for the enthusiasts

Rejecting H_0 because the numbers say so, but based on wrong reasoning/methodology

E.g.1 Concluding ´*the drug* ***improves*** *health*´ with two-tailed testing

E.g.2 The drug has a positive effect and people feel better, but it doesn´t concern the theoretical effect and illness that was studied

Validity and Reliability

Validity refers to the accuracy of a measure: do you measure what is supposed to be measure?

e.g. A new questionnaire is used and the results align with those of previously used questionnaires on the same topic among a comparable population 👍

e.g. A thermometer is placed next to the door in a controlled laboratory fridge 👎

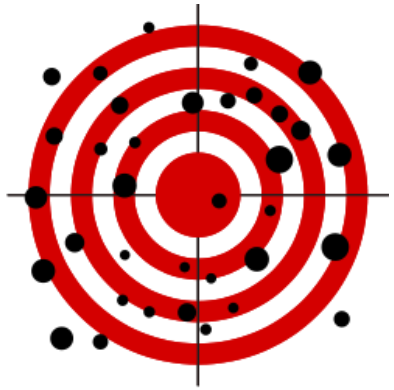
Reliability refers to the consistency of a measure and data: can results be reproduced under the same conditions, and if the dataset size N increases, does the outcome of the test increasingly approaches the correct outcome?

e.g. A respondent answers ´yes´ to *Do you eat meat?* and ´no´ to *Are you a vegetarian?* 👍

e.g.2 The more respondents (higher N), the more significant differences are found, which strengthens the outcome

→ Internal consistency is usually expressed with Cronbach's alpha α

Unreliable & Unvalid



Unreliable, But Valid

Reliable, Not Valid

Both Reliable & Valid

Validity

Face Validity

Expert judgement

Does it make sense at first sight, also to participants?

Content Validity

Completeness

Does the test cover everything it should?

Criterion Validity

Does one measurement corrrelate with another?

Does a test accurately predict or correlate with the criterion outcome?

Construct Validity

Am I measuring what I want to measure?

Does a set of indicators represent a concept that is not directly measurable?

Internal Validity

Causality X -> Y without Z influencing the relation

Does the study design, conduct, and analysis answer the research question without bias?

External validity

Can you generalise your results?

Can the study findings can be generalized to other contexts?

Reliability

- **Sample Size (N)**
- **Test-retest**
- **Standardization**
- **Pilots**
- **Peer Feedback**
- **Reporting and justification**

Quantitative

- **Intersubjectivity**
- **Triangulation**
- **Iteration**

Qualitative

Sampling

→ Set criteria for your target population, e.g. age limits or country
→ Get a sample of subjects that is as <u>random</u> and <u>representative</u> as possible.

Random: the word says it all. Don´t cherry-pick or go around asking only friends and family.
Representative: The sample should contain the best mix of characteristics *within your target population*.

PROBABILITY SAMPLING:	based on the principle of randomization: random selection or chance within population or subpopulation
NON-PROBABILITY SAMPLING:	Selection is influenced by subjective boundaries and/or convenient picking from population or subpopulation

Sampling: options

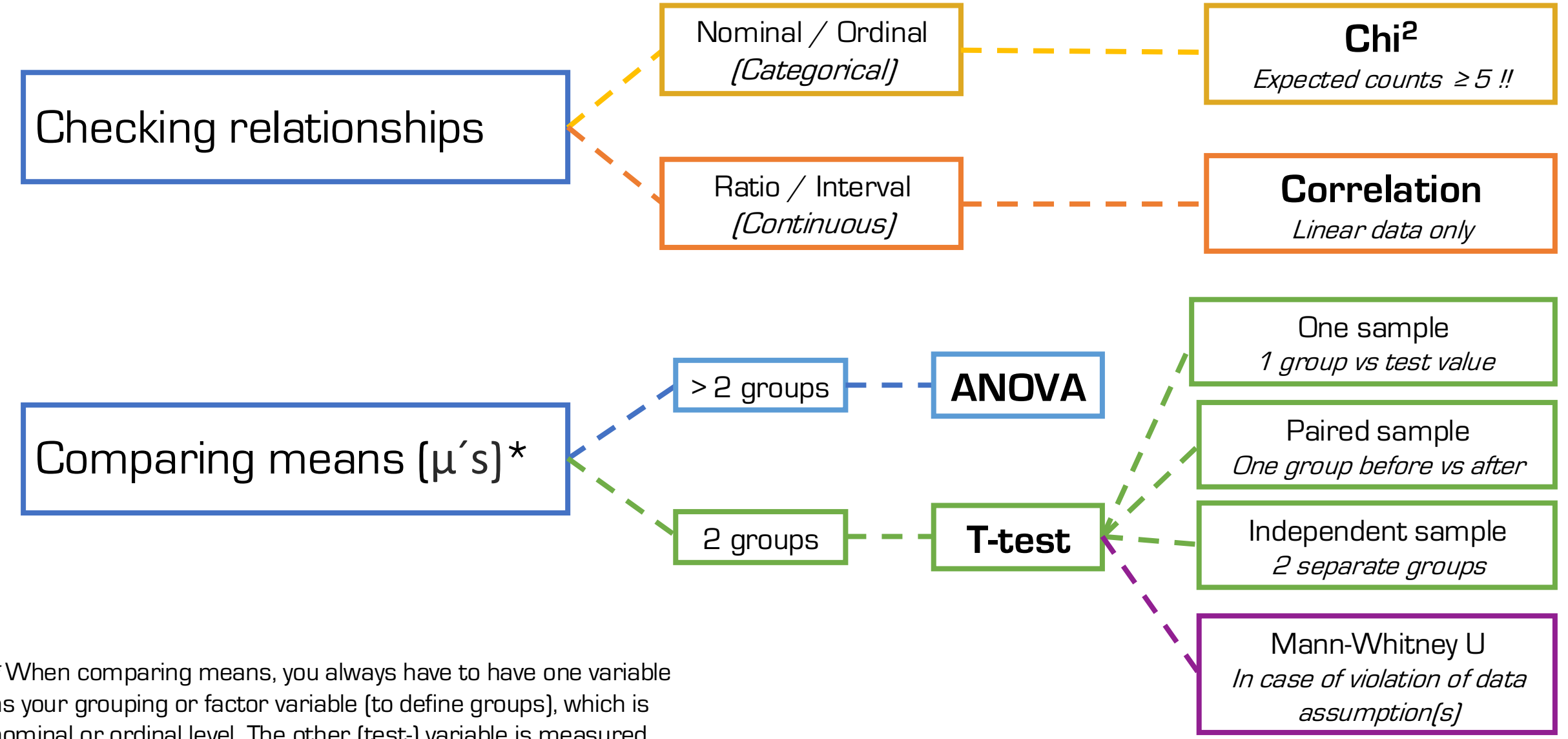
PROBABILITY SAMPLING OPTIONS

- (SIMPLE) RANDOM — Any subject within the target population (random from database/list
- STRATIFIED — Randomly sampling within ***relative subpopulations***
  e.g. if a student population has 50% NL, 20% Germany, 20% Belgium and 10% French students, you randomly point out people in those same proportions (so 500 NL, 200 DE, 200 BE and 100 FR)
- CLUSTERED — Dividing a large population into ***representative clusters***, then selecting clusters
  e.g. you want a sample of all language students in the NLs, but cannot go to all uni´s and hbo´s, so you make clusters (which are language departments from all institutions), pick one cluster from every institution, and talk to to students
- SYSTEMATIC — Using some order or pattern to select, such as every 10th guest

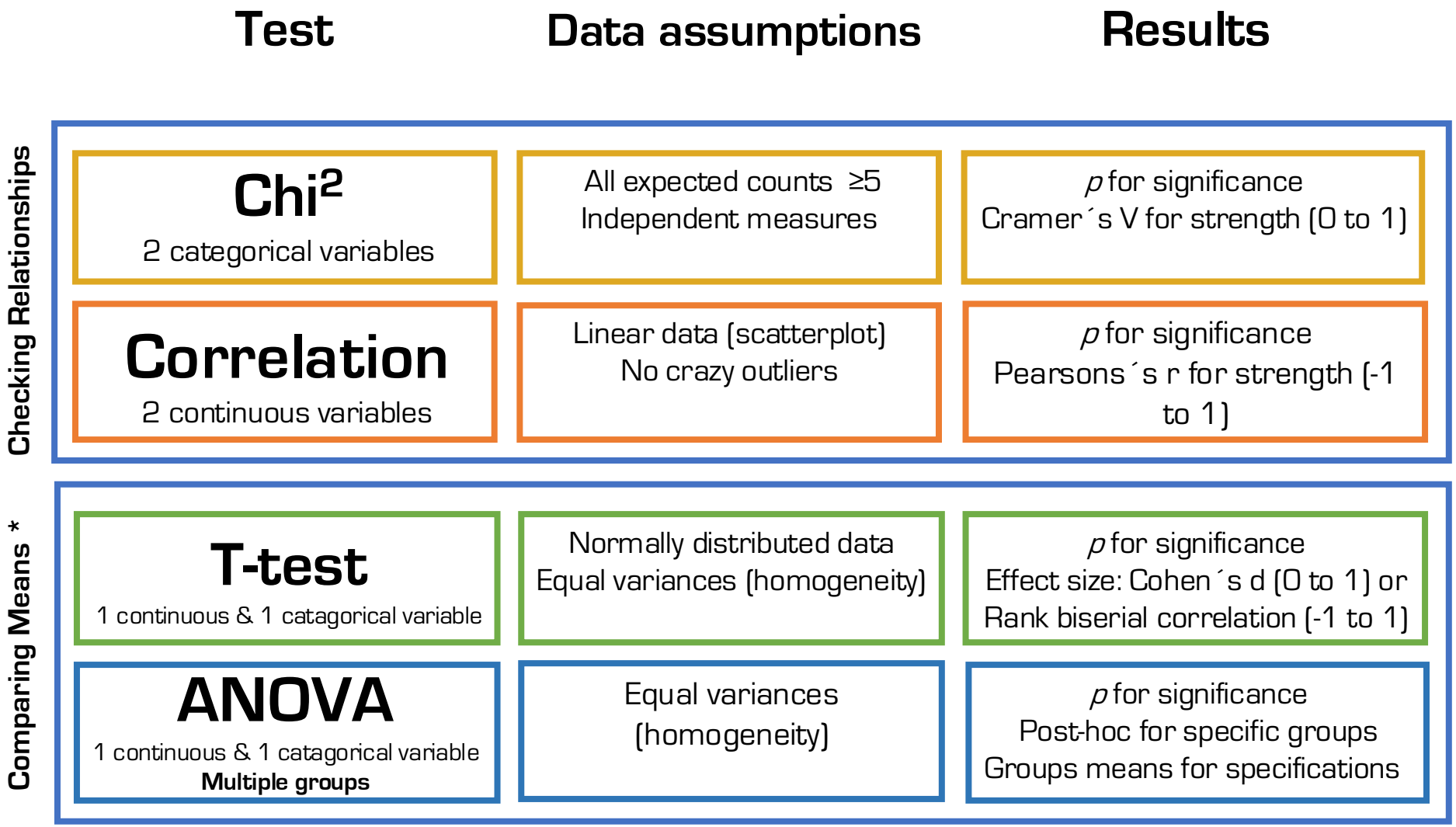
NON-PROBABILITY SAMPLING OPTIONS

- SNOWBALL — Any subject can select a next subject
- QUOTA — Sampling within relatively set subpopulations, but not in a random way
  e.g. if a hotel has 80% rooms and 20% suites, you simply pick 40 room-people and 10 suite-people. You only stop when you reach your quota, so when you have those 40 and 10.
- SELF-SELECTION — People put themselves forward for participation
- CONVENIENCE — Quick and convenient: any subject that you can find within a certain period or area. Very ´not random´.

What test to use?



*When comparing means, you always have to have one variable as your grouping or factor variable (to define groups), which is nominal or ordinal level. The other (test-) variable is measured at either interval or ratio level.



* When comparing means, you always have to have one variable as your grouping or factor variable (to define groups), which is nominal or ordinal level (e.g. gender or age group). The other (test-) variable(s) is/are measured at continuous level.

CHI²

- The Chi² test, or χ^2 test, is used to see if there is a relationship between two categorical variables.
- Because categorical variables don´t have numerical measures and therefore cannot be analysed using means, Chi² works with frequencies to see if there are differences between observed outcomes (your measured counts) and expected counts.
- The expected outcomes for all cells should be at least 5, and are calculated using the formula (rowtotal x columntotal) / N, for example (28x32) /106 for cell **1-NO** in the table below
- You find the Chi² test in Jamovi under Analyses > Frequencies > Independent Samples $\chi 2$ of association
- Jamovi provides a contingency table, the χ^2 test outcomes, and Cramer´s V (if selected under *Statistics*)

Contingency Tables

qualification		healthplan No	healthplan Yes	Total
1	Observed	12	16	28
	Expected	8.45	19.55	28.00
2	Observed	18	27	45
	Expected	13.58	31.42	45.00
3	Observed	0	8	8
	Expected	2.42	5.58	8.00
4	Observed	2	18	20
	Expected	6.04	13.96	20.00
5	Observed	0	5	5
	Expected	1.51	3.49	5.00
Total	Observed	32	74	106
	Expected	32	74	106

χ^2 Tests

	Value	df	p
χ^2	13.7	4	0.008
N	106		

Nominal

	Value
Phi-coefficient	NaN
Cramer's V	0.359

CORRELATION

- The correlation test is used to see if there is a relationship between two continuous variables
- To check data assumptions, in Jamovi use Exploration > Scatterplot to see if data are reasonably linear or not
- You find the correlation test in Jamovi under Analyses > Regression > Correlation Matrix
- Jamovi provides a Correlation Matrix with the outcomes, among which Pearson´s r

Correlation Matrix

		age	seniority
age	Pearson's r	—	
	df	—	
	p-value	—	
seniority	Pearson's r	0.882	—
	df	104	—
	p-value	<.001	—

T-TEST

- The T-test is used compare the means (μ´s) of **two groups** for a continuous dependent variable, to see if there is a significant difference between the groups. There are 3 types of T-test.
  - **One sample t-test**. Used to compare the μ of data you gathered to an already existing μ (e.g. from literature)
  - **Independent t-test**. Used to compare the μ´s of two groups you observed
  - **Paired samples t-test**. Used to compare the μ´s of data you gathered within one group, but at two different moments (e.g. μ heart rate before and μ after drinking coffee). ¡This requires the same people in both groups!
- Here, you have to set clear one-sided or two-sided hypotheses to choose the right settings and interpret results correctly. *In case of one-sided testing, only half of the chosen p-value is used*. Jamovi automatically changes the p when you tick one-sided; in SPSS you would have to read to column ´one-sided´.
- You find T-tests in Jamovi under Analyses > T-tests.
- To check data assumptions, in Jamovi tick *Homogeneity test* (Levene´s) and *Normality test* (Shapiro-Wilk).
  - If both p-values are **higher** than 0.05 you can move forward with a normal T-test.
  - If either one of them is violated or both are, you have to tick the *Mann-Whitney U* option and then continue. (The *Mann-Whitney U* uses medians instead of means)
- For effect size, tick *Effect Size* under *Additional Statistics*. This will add Cohen´s d (applies when doing a normal t-test) and Rank biserial correlation (applies when doing the Mann-Whitney U).

Normality Test (Shapiro-Wilk)

	W	p
perf_quantity	0.979	0.094

Note. A low p-value suggests a violation of the assumption of normality

Homogeneity of Variances Test (Levene's)

	F	df	df2	p
perf_quantity	0.256	1	104	0.614

Note. A low p-value suggests a violation of the assumption of equal variances

Independent Samples T-Test

		Statistic	df	p
perf_quantity	Student's t	1.43	104	0.157

Note. $H_a\ \mu_2 \neq \mu_1$

	Effect Size
Cohen's d	0.383
Rank biserial correlation	−0.206

ANOVA

- ANOVA is used to compare means (μ´s) of **three or more groups** for a continuous dependent variable, to see if there is a significant difference between those groups. You have to know 2 options for ANOVA, for which the choice is based on Levene´s test of homogeneity (as a data assumption).
  - **ANOVA Fisher´s test**. Used if homogeneity is met (so: Levene´s p > 0.05)
  - **ANOVA Welch test**. Used if homogeneity is **not** met (so: Levene´s p < 0.05)
- You find ANOVA in Jamovi under Analyses > ANOVA > One-way ANOVA.
- To check homogeneity data assumption, in Jamovi tick *Homogeneity test* (Levene´s) within the ANOVA test

However.... if a significant difference is found, we still have to see **between which groups**, for which we use post-hoc tests. These post-hoc tests show p-values for specific pairs of groups so you can report outcomes more precisely. To also report which groups scores higher than another, use desriptives to see groups´ means. Post-hoc options are:

- For an ANOVA Welch test we use the Games-Howell
- For an ANOVA Fisher´s test we use the Bonferroni or Tukey, for which the choice depends on acceptable risks:
  - Bonferroni is quite strict, which means a lower chance of a type I error, yet a bigger chance of a type II error
  - Tukey is less strict, which means a higher chance of a type I error, yet a smaller chance of a type II error. Jamovi only has Tukey.

Assumption Checks

Homogeneity of Variances Test (Levene's)

	F	df1	df2	p
perf_quantity	1.68	4	101	0.160

One-Way ANOVA

		F	df1	df2	p
perf_quantity	Welch's	8.80	4	28.6	<.001
	Fisher's	4.18	4	101	0.004

Tukey Post-Hoc Test – perf_quantity

		1	2	3	4	5
1	Mean difference	—	0.365	1.132	−0.439	−1.136
	p-value	—	0.734	0.156	0.742	0.324
2	Mean difference		—	0.767	−0.804	−1.501
	p-value		—	0.486	0.117	0.081
3	Mean difference			—	−1.571	−2.268
	p-value			—	0.024	0.014
4	Mean difference				—	−0.697
	p-value				—	0.789
5	Mean difference					—
	p-value					—

NOTES

Hotel Management School Maastricht

ZU YD

HMSM Team Research Skills (2026). Concise but significant handbook to M5 statistics. Version 0226ENG.