EPSE 592: Design & Analysis of Experiments

Ed Kroc

University of British Columbia

ed.kroc@ubc.ca

March 26 & April 2, 2020

Nonparametric procedures

- Analyzing categorical response data:
  - Chi-squared tests
  - Fisher's exact test
  - Sign and McNemar tests (paired data)
- Analyzing continuous *or* categorical response data:
  - Mann-Whitney and Wilcoxon tests
  - Kruskal-Wallis nonparametric one-way "ANOVA"
  - Friedman test (nonparametric RM-ANOVA)
  - Permutation tests

Nonparametric Statistics

- So far, all statistical tests that we have considered (e.g. t-tests, F-tests, ANOVAs) have been examples of *parametric* procedures. A parametric test is one that makes a *distributional* assumption about the data in some way.
  - t-tests assume data are *normally distributed*.
  - F-tests assume data are *normally distributed*.
  - Traditional ANOVAs (and regressions) assume that residuals are *normally distributed*.
- In contrast, a *nonparametric procedure* does *not* make any distributional assumptions about the data.

When to use nonparametric procedures

Consider using nonparametric procedures in the following contexts:

- When you are worried about severe violations of assumptions of robust parametric procedures (e.g. t-tests, ordinary ANOVA).
- When you are worried about mild violations of assumptions of sensitive parametric procedures (e.g. RM-ANOVA).
- When you have only categorical response data (e.g. voting data).
- When you have outliers in your data (robust procedures may also be available).
- When you have too little data to reasonably check assumptions of parametric procedures (remember: checking those assumptions requires enough data in *all* groups in order to have enough power to detect violations).

Chi-squared Tests

- Chi-squared (χ^2) tests are often used to test for the presence of relationships between **categorical variables**.
- More specifically, chi-squared tests always ask if the probability distribution of a categorical variable is the same across the levels of another categorical variable.
- Put another way, chi-squared tests are used to test if the observed sample proportions of a categorical response variable are similiar across different groups/treatments.

Chi-squared Tests: example 4

- Suppose we would like to test if patient survival is independent of the type of drug used in treatment. Here, we have data on 31 patients taking Drug A, and 59 patients taking Drug B. After 5 years, we have the following counts:

	Drug A	Drug B
Death	14	22
Survival	17	37

- Here, testing independence amounts to testing if the likelihood of survival (or death) is the same for both drugs. Thus, our particular null hypothesis is:

$$H_0 : \Pr(Survival \mid DrugA) = \Pr(Survival \mid DrugB)$$

Chi-squared Tests: example 4

Contingency Tables

Survival		Treatment: Drug A	Treatment: Drug B	Total
No	Observed	14	22	36
	Expected	12.400	23.600	
Yes	Observed	17	37	54
	Expected	18.600	35.400	
Total	Observed	31	59	90
	Expected	31	59	

χ^2 Tests

	Value	df	p
χ^2 continuity correction	0.248	1	0.618
Fisher's exact test	1.380		0.503

- Here, we find no evidence against the null hypothesis. Note the Fisher's exact test statistic at the bottom that seems to be corroborating the result of the χ^2 test.

Fisher's exact test

- Fisher's exact test is a truly nonparametric alternative (i.e. assumes no asymptotic parametric relationships) to the χ^2 test.
- It (or a more general version) can be applied in all the situations previously discussed, though Jamovi will only allow for easy implementation in the 2×2 contingency table case (as in Ex. 4).
- Unlike the χ^2 test, Fisher's exact test is valid (and should be used) when:
  - Total sample sizes are insufficient to apply a χ^2 test.
  - Observed or expected cell counts are too small. In particular, if any cell counts are as small as 0 or 1, and usually when any cell counts are less than 5.

Fisher's exact test rationale

- For simplicity, we will describe the procedure for the case of 2×2 count data as in Example 4.
- Suppose we have the following 2×2 contingency table:

	$X = 1$	$X = 2$	total
response 1	a	b	$a + b$
response 2	c	d	$c + d$
total	$a + c$	$b + d$	$n = a + b + c + d$

- We hypothesize that X should not affect the response variable, call it Y. That is, we hypothesize

$$H_0 : \Pr(Y = 1 \mid X = 1) = \Pr(Y = 1 \mid X = 2)$$

Fisher's exact test rationale

- There is an easy way to calculate just how likely these observed data are, given the null hypothesis and the fixed row and column totals. In our case, the probability of observing our data is given by

$$\Pr(\textit{data}) = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}},$$

where $\binom{\alpha}{\beta}$ is a *binomial coefficient*, or *choose* function, defined as

$$\binom{\alpha}{\beta} := \frac{\alpha!}{\beta!(\alpha-\beta)!},$$

where $\alpha! = \alpha \cdot (\alpha-1) \cdot (\alpha-2) \cdots 2 \cdot 1$.

- The quantity $\binom{\alpha}{\beta}$ is the *number of unique ways to select β objects from a pool of α objects*; e.g. $\binom{4}{2} = 6$ is the number of ways to select 2 people from a group of 4 total people.

Fisher's exact test rationale

- For the Fisher's exact test, our test statistic are *the data themselves*. Thus, to calculate a p-value, we need to calculate the probability of observing any 2×2 contingency table *as or more extreme* than the one we observed (i.e. which tables are more *unbalanced* than the one we observed).

Fisher's exact test rationale

For example, if our data are:

	$X = 1$	$X = 2$	total
response 1	1	10	11
response 2	9	13	22
total	10	23	33

then there is only one 2×2 table more extreme (i.e. more unbalanced) as the one for our data:

	$X = 1$	$X = 2$	total
response 1	0	10	11
response 2	10	13	22
total	10	23	33

Thus, our p-value for the above data (the test statistic) would just be the sum of the probabilities of observing each of these two tables, using the formula from the previous slide.

Return to example 4, Fisher's exact test

- Suppose we would like to test if patient survival is independent of the type of drug used in treatment. Here, we have data on 31 patients taking Drug A, and 59 patients taking Drug B. After 5 years, we have the following counts:

	Drug A	Drug B
Death	14	22
Survival	17	37

- Here, testing independence amounts to testing if the likelihood of survival (or death) is the same for both drugs. Thus, our particular null hypothesis is:

$$H_0 : \Pr(Survival \mid DrugA) = \Pr(Survival \mid DrugB)$$

Return to example 4, Fisher's exact test

Contingency Tables

Survival		Treatment: Drug A	Treatment: Drug B	Total
No	Observed	14	22	36
	Expected	12.400	23.600	
Yes	Observed	17	37	54
	Expected	18.600	35.400	
Total	Observed	31	59	90
	Expected	31	59	

χ^2 Tests

	Value	df	p
χ^2 continuity correction	0.248	1	0.618
Fisher's exact test	1.380		0.503

- Here, we find no evidence against the null hypothesis using Fisher's exact test.

Fisher's exact test vs. chi-squared tests

A reminder:

- Both tests can be used to test for the presence of relationships between categorical variables.
- Both tests assume all observations are independent (won't work for paired data, though analogues do exist).
- χ^2 tests rely on *asymptotics*; i.e. they require sufficient sample sizes, overall and within each cell of the contingency table.
- Fisher's exact test applies *regardless* of sample size, overall or cell-wise.

What to do with paired or repeated measures data

- Since both χ^2 and Fisher's exact tests rely on an assumption of *independence* of all data, they do not apply for repeated measures (or paired) comparisons.
- Example: we present 30 people with a choice of two brands of soda (A or B). Each person tastes **both** sodas and records their preference. We would like to determine if there is evidence that one brand is preferred over the other.
  - Sign, McNemar's, or Cochran's tests
    - Used when paired data can assume only two possible categories
    - Used when paired data are nominal (i.e. categories are not ordered)
  - Friedman's test
    - Used when the (ordinal) paired data have more than 2 factor levels.
    - Also used as a nonparametric analogue to repeated measures ANOVA.

Tests based on ranks

- Chi-squared and Fisher's exact tests (etc.) allow us to compare a categorical response variable over different levels of a categorical explanatory variable.
- But their tests of hypotheses consider whether or not *sample proportions* are similar. Thus, they are not directly comparable to t-tests and ANOVAs that test hypotheses about similarity of *means.*
- Moreover, *they do not directly propose a model*. Thus, unlike ANOVA, it is unclear how to examine the relationship of a categorical response variable to *multiple* categorical explanatory variables.
- Instead, there are many methods based on *rank statistics* that allow for nonparametric generalizations of the ANOVA (and t-test) framework.

The Mann-Whitney (or Mann-Whitney-Wilcoxon or Wilcoxon rank-sum) test rationale

- Suppose we have a random sample of ordinal or continuous (independent) observations from two groups. Observations from Group 1: $X_1, \ldots, X_n$, and observations from Group 2: $Y_1, \ldots, Y_m$.
- Now, we may *rank* all the $n + m$ observations, from smallest to largest, assigning an average rank to observations if there are any ties.
- Notationally, we let $R(X_i)$ denote the rank assigned to X_i, and $R(Y_j)$ denote the rank assigned to Y_j for each X_i and Y_j in our sample.
- If the mean of the $R(X_i)$'s are close to the mean of the $R(Y_j)$'s, then the "typical" X value should be close to the "typical" Y value.
- We will return to what "typical" means soon.

The Mann-Whitney test rationale

- More precisely, we can test the null hypothesis

$$H_0 : \Pr(X > Y) = \Pr(X < Y),$$

by considering the natural test statistic

$$U = \frac{1}{n} \sum_{i=1}^{n} R(X_i)$$

where $R(X_i)$ denotes the rank assigned to observation X_i.

- Under the null hypothesis, the average of the ranks in the X group should be *very close* to the average of the ranks in the Y group; i.e.

$$\text{Assuming } H_0 : \frac{1}{n} \sum_{i=1}^{n} R(X_i) \approx \frac{1}{m} \sum_{j=1}^{m} R(Y_j)$$

The Mann-Whitney test rationale

- For reasons of mathematical convenience, the actual test statistic used in the Mann-Whitney test is:

$$U_1 = \sum_{i=1}^{n} R(X_i) - \frac{n(n+1)}{2}.$$

  Note that this statistic contains the *exact same information* as the more natural test statistic above.

- Under the null hypothesis, U_1 should be *very close* to the complementary test statistic for the other group of observations:

$$U_2 = \sum_{j=1}^{m} R(Y_j) - \frac{m(m+1)}{2}.$$

- Note: these U-statistics change slightly if there are ties in the data.

The Mann-Whitney test rationale

- The Mann-Whitney U-statistic follows a known probability distribution for which we can calculate probabilities; thus, we can calculate p-values.
- The Mann-Whitney test can give us evidence that two groups of observations have different "typical" values. But what does "typical" mean?
- Literally, here, the "typical" observation in one group is the *mean of the sample ranks* derived from pooling and ranking all the data from both groups.
- In many cases, this typical value can be thought of as the *median*, or 50th percentile, (a rank statistic in its own right), or at least acting very much like a median.
- However, in general, the mean of the sample ranks can be different between the two groups while the sample medians are identical.

The Mann-Whitney test: Example 1

- In Jamovi, you can run a Mann-Whitney test by using the standard 'Independent Samples t-test' procedure and then selecting the 'Mann-Whitney' option.

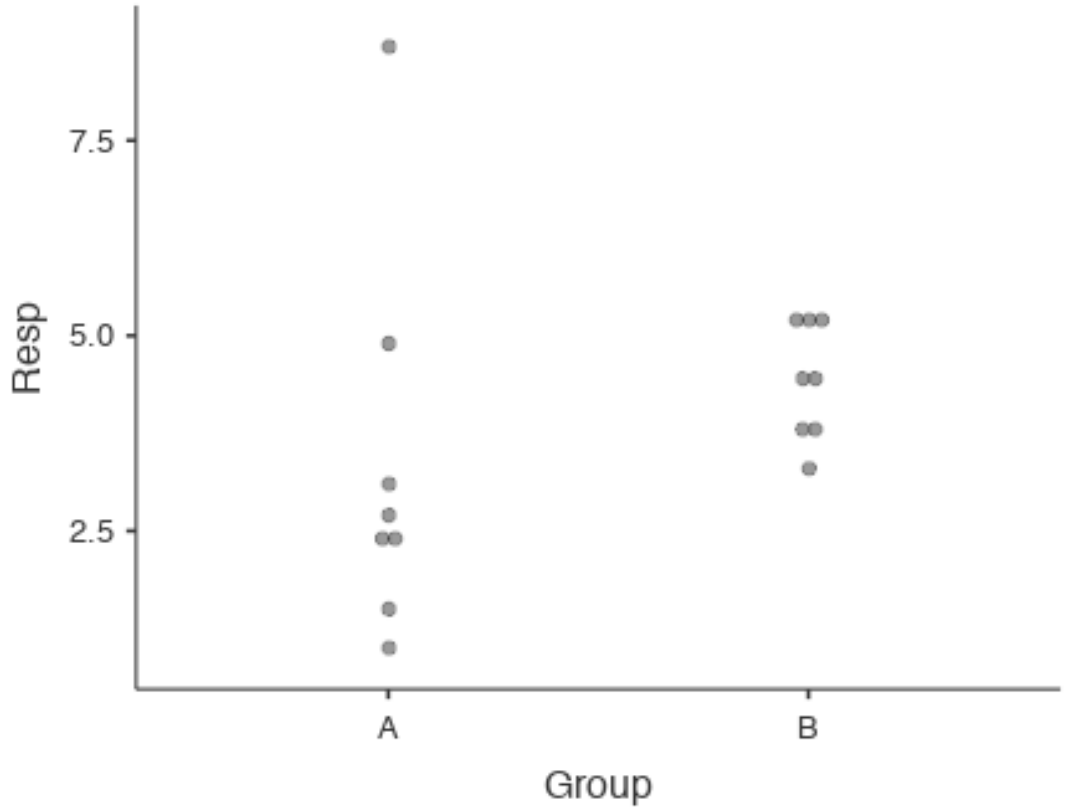
Independent Samples T-Test

		statistic	df	p	Mean difference	SE difference
Resp	Student's t	−1.198	14.000	0.251	−1.087	0.908
	Welch's t	−1.198	8.279	0.264	−1.087	0.908
	Mann-Whitney U	13.000		0.050	−1.800	

- Notice: no mean difference detected by ordinary or robust t-tests
- However, Mann-Whitney indicates evidence of a difference in "typical" values between the two groups.

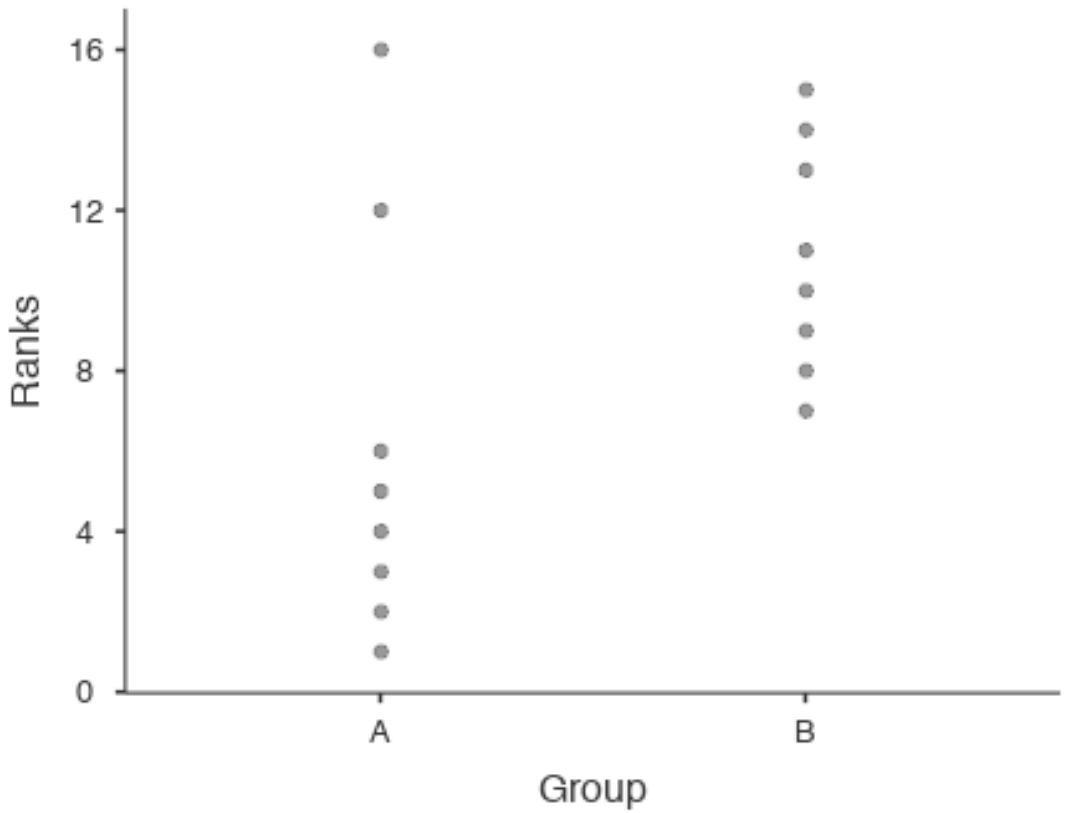
The Mann-Whitney test: Example 1

- These data have the following 'typical' values:
  - Group A mean: 3.34, Group A median: 2.60
  - Group B mean: 4.43, Group B median: 4.45



The Mann-Whitney test: Example 1

- These data have the following 'typical' values for the ranks:
  - Group A mean rank: 6.13
  - Group B mean rank: 10.88
- Remember: MW-test compares the **group mean ranks**.



The Mann-Whitney test: Example 2

- Now we'll change *a single data point* so that the group means become the same, but the group medians remain different. We can accomplish this by making the 'outlier' in the A group more extreme.

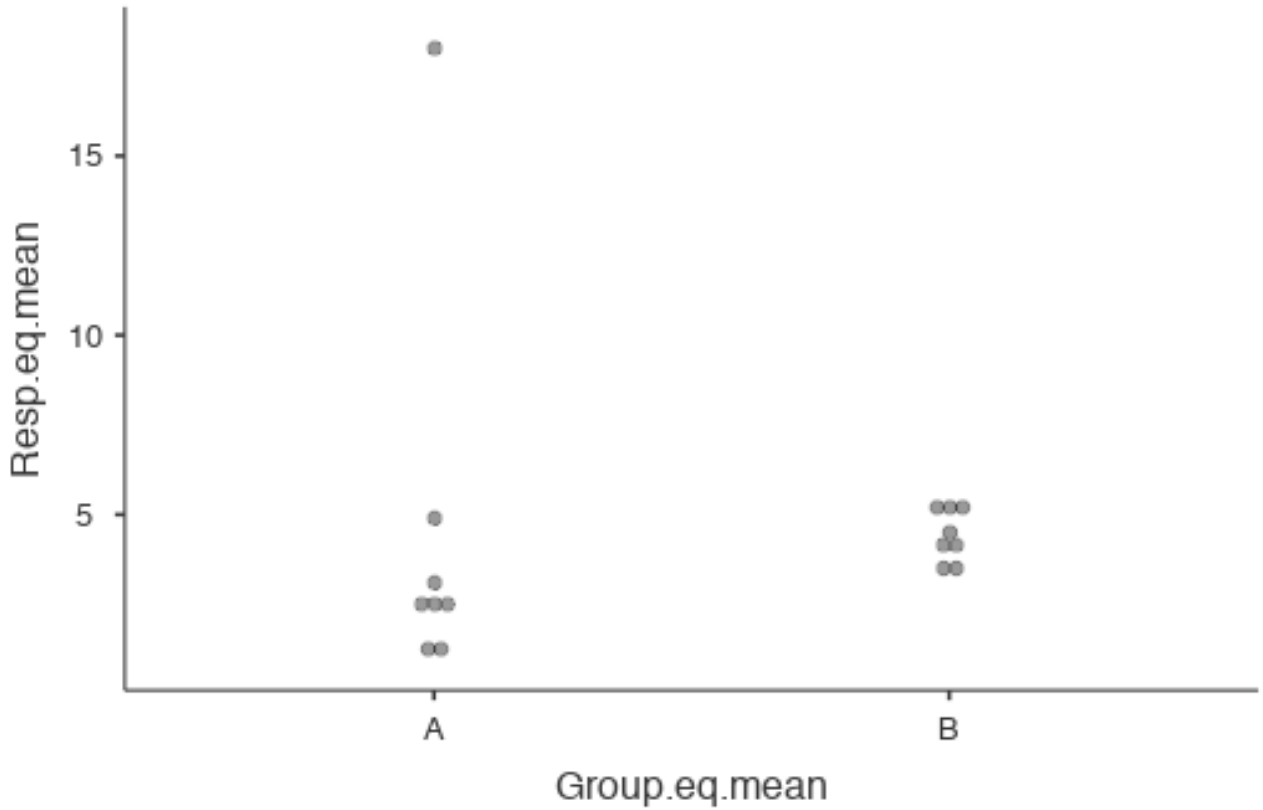
Independent Samples T-Test

		statistic	df	p
Resp.eq.mean	Student's t	0.038	14.000	0.970
	Welch's t	0.038	7.250	0.971
	Mann-Whitney U	13.000		0.050

- Notice: no mean difference detected by ordinary or robust t-tests
- However, Mann-Whitney indicates evidence of a difference in "typical" values between the two groups. Note too that the Mann-Whitney statistic is *identical* to the previous one.

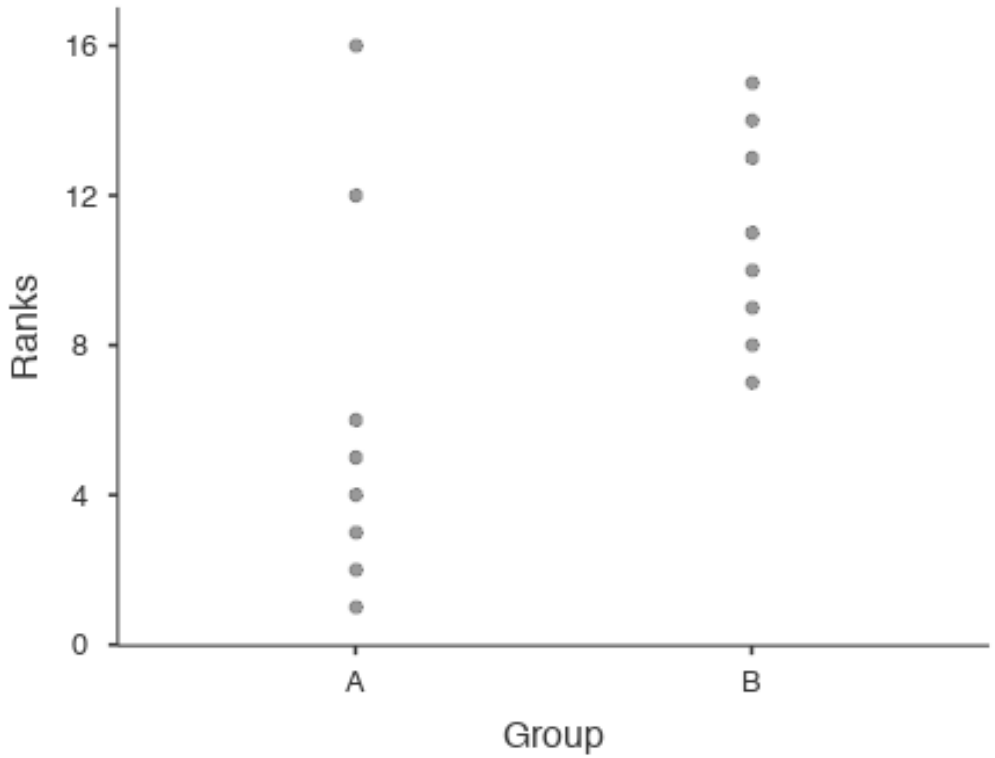
The Mann-Whitney test: Example 2

- These data have the following 'typical' values (note: medians are the same as previous example):
  - Group A mean: 4.50, Group A median: 2.60
  - Group B mean: 4.43, Group B median: 4.45



The Mann-Whitney test: Example 2

- These data have the following 'typical' values for the ranks. Notice that these are *exactly the same* as the previous example:
  - Group A mean rank: 6.13
  - Group B mean rank: 10.88
- Remember: MW-test compares the **group mean ranks**.



The Mann-Whitney test: Example 3

- Now we'll change the data again to make the group means and medians nearly the same.

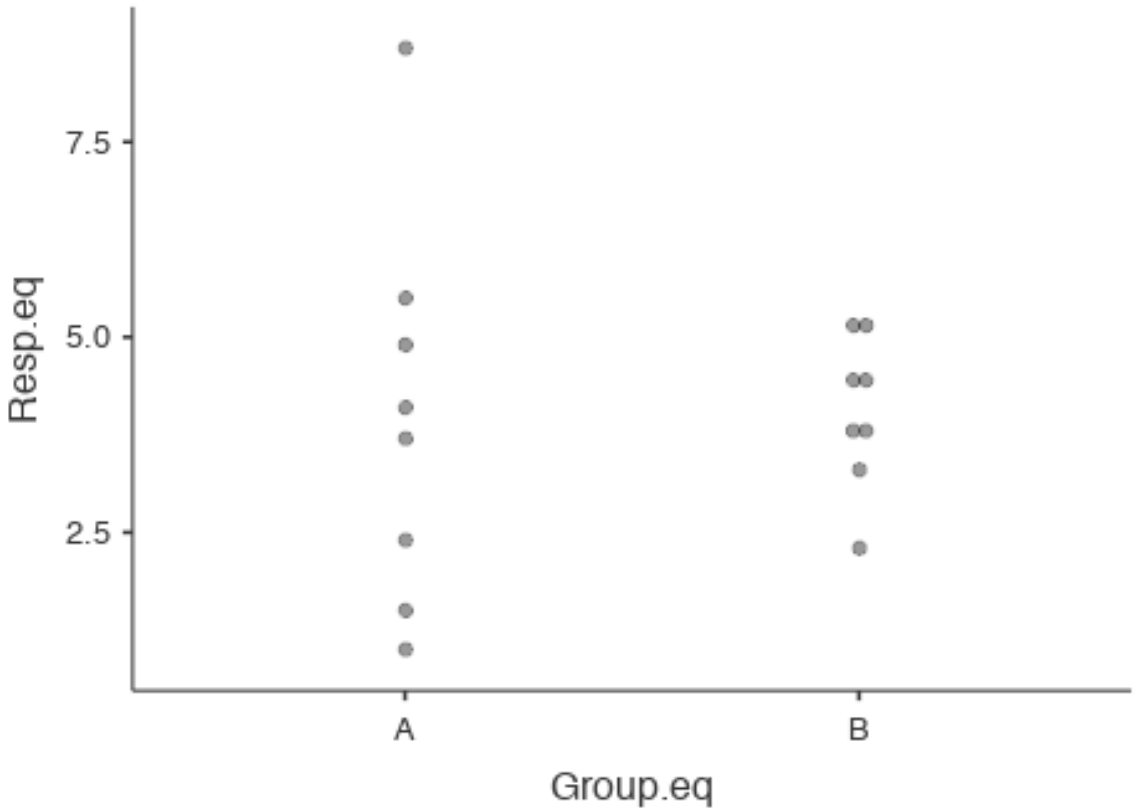
Independent Samples T-Test

		statistic	df	p
Resp.eq	Student's t	–0.080	14.000	0.938
	Welch's t	–0.080	9.073	0.938
	Mann-Whitney U	29.500		0.834

- Notice: no mean difference detected by ordinary or robust t-tests
- Now, Mann-Whitney indicates no evidence of a difference in "typical value" between the two groups.

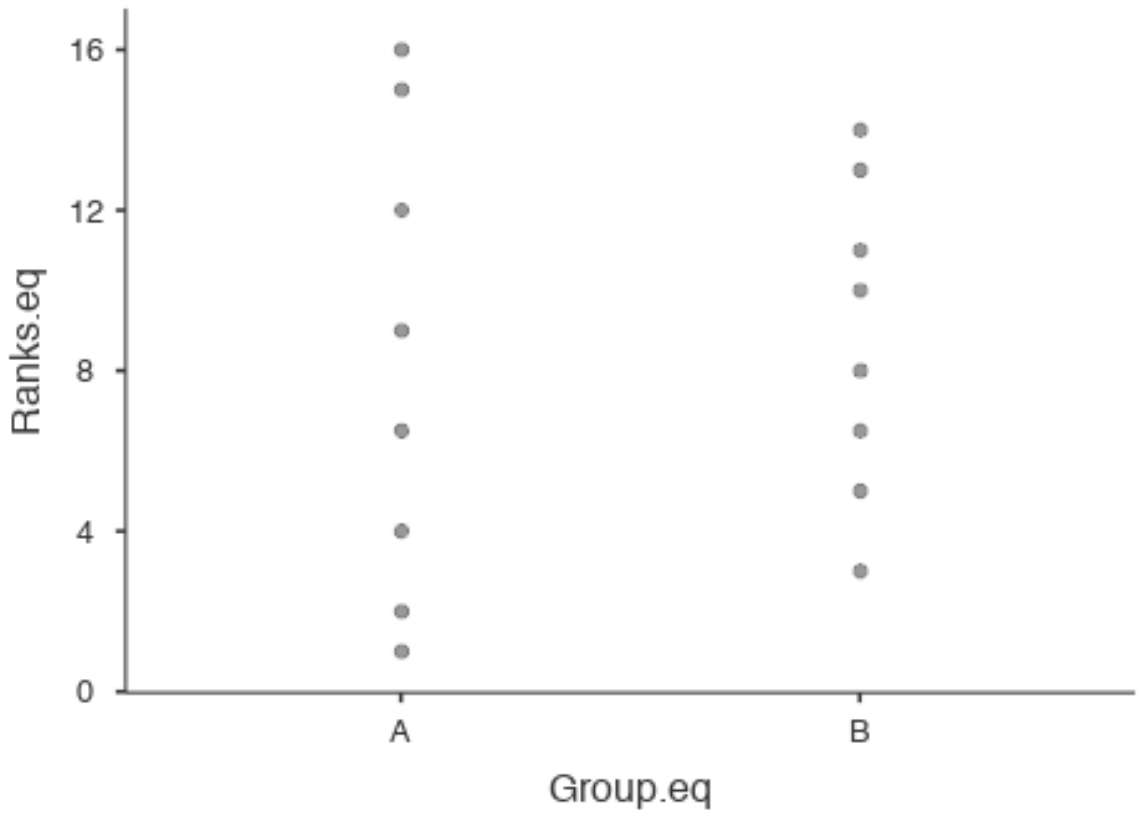
The Mann-Whitney test: Example 3

- These data have the following 'typical' values:
  - Group A mean: 3.98, Group A median: 3.90
  - Group B mean: 4.05, Group B median: 4.15



The Mann-Whitney test: Example 3

- These data have the following 'typical' values for the ranks.
  - Group A mean rank: 8.19
  - Group B mean rank: 8.81
- Remember: MW-test compares the **group mean ranks**.



The Mann-Whitney test: reporting

- The Mann-Whitney U-statistic tests a null hypothesis about the *difference in a probability distribution over groups*:

$$H_0 : \Pr(X > Y) = \Pr(Y > X)$$

- Under this H_0, the mean ranks (and the median ranks) should be about the same between the two groups.
- So what effect size should be reported?
  - Traditionally, people report either the *sample medians* of the raw data or the *difference* in these sample medians along with a MW test.
  - It is *rare* to see anyone report *sample means or medians of the ranks*.
- Heuristically, one can think of the Mann-Whitney test as the nonparametric analogue of the independent samples t-test, although as we have seen they do *not* test exactly the same thing (Example 2).

The Mann-Whitney test: when to use

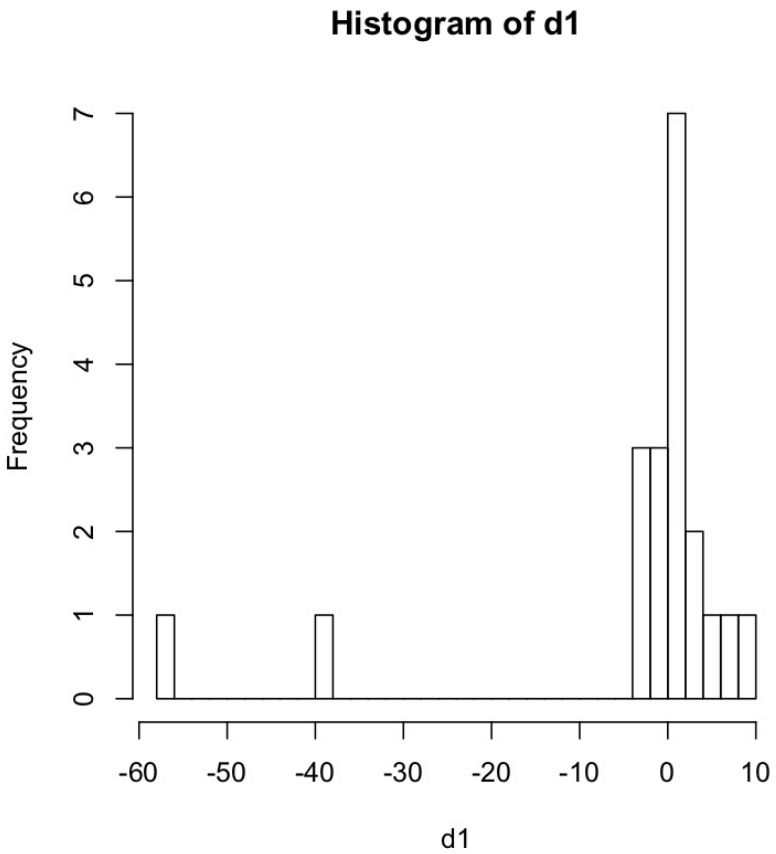
- MW-test can be used when:
  - comparing typical values of ordinal categorical variables.
  - comparing typical values (heuristically, the medians) of continuous variables.
- MW-test is **more powerful** than (traditional or Welch's) t-tests when:
  - data are not normally distributed.
  - data from different groups are distributed differently.
  - data are heteroskedastic.
  - data contain outliers.
- However, MW-test cannot detect *all* differences between comparison groups. [E.g. will not detect unequal variances; use an F-test instead.]

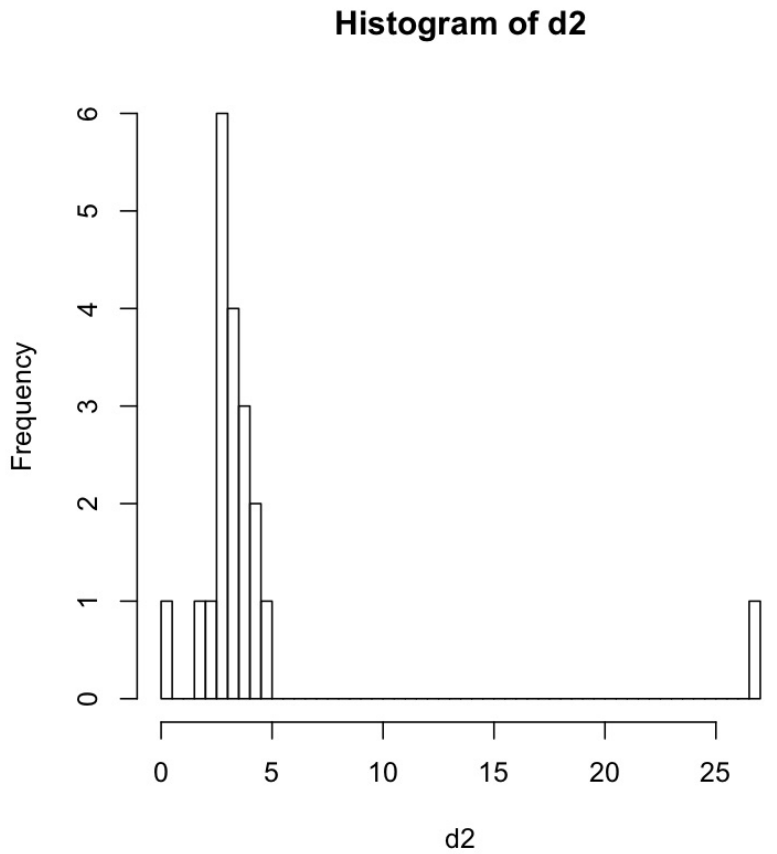
The Mann-Whitney test: it's not all about power

- The MW-test is much more powerful than the *t*-test when outliers are present, but remember: we don't just run statistical tests to tell us if there is evidence of an average difference between groups.
- In practice, *we report and interpret effect sizes*; e.g. sample means. These are the *treatment effects* on which we then base clinical decisions.
- Means are highly sensitive to outliers. Thus, when outliers are present, *means can be a bad measure of centrality*.
- Medians (or rank statistics) are *not* sensitive to outliers, as percentiles (ranks) do not use information about *distances* between observations, only *ordering* of observations (compare Examples 1 and 2).

The Mann-Whitney test: it's not all about power

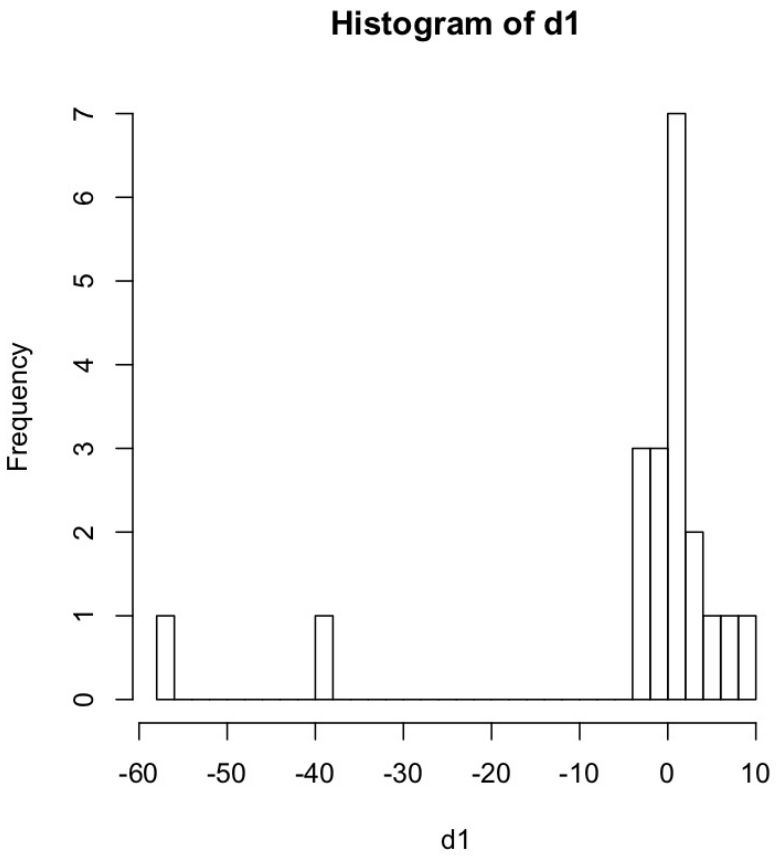
- Consider the following data on two groups (20 data points each). These are generated from $t(2)$ random variables with mean=median=0 for Group 1 and mean=median=3 for Group 2.

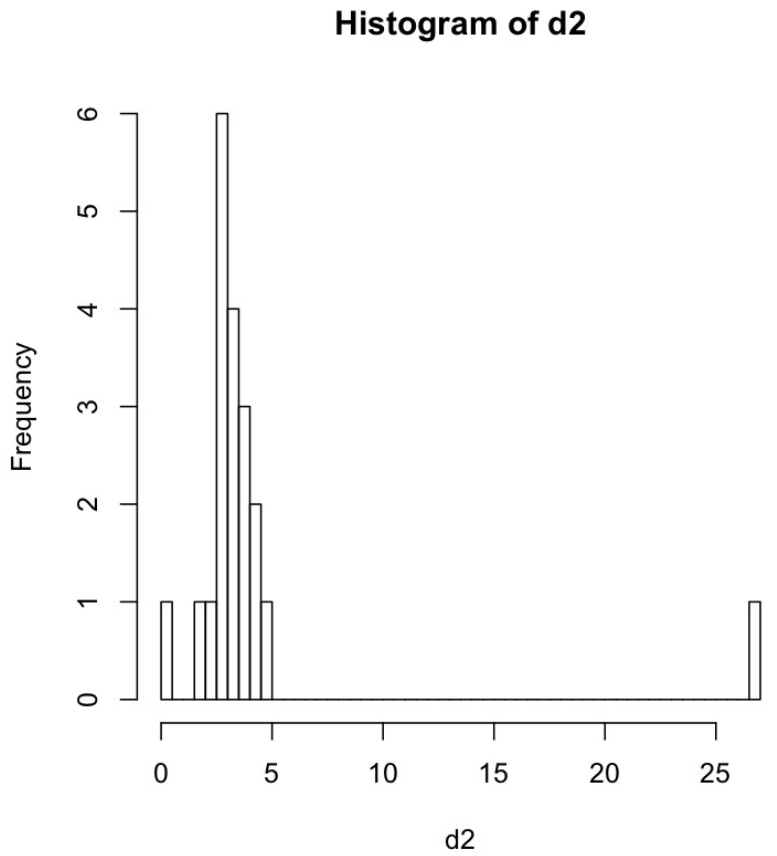




The Mann-Whitney test: it's not all about power

- However, the *sample mean* for Group 1 is -3.8 and the *sample mean* for Group 2 is 4.2. These estimates are quite bad, due to the presence of the outliers (the distributions are "heavy tailed").





The Mann-Whitney test: it's not all about power

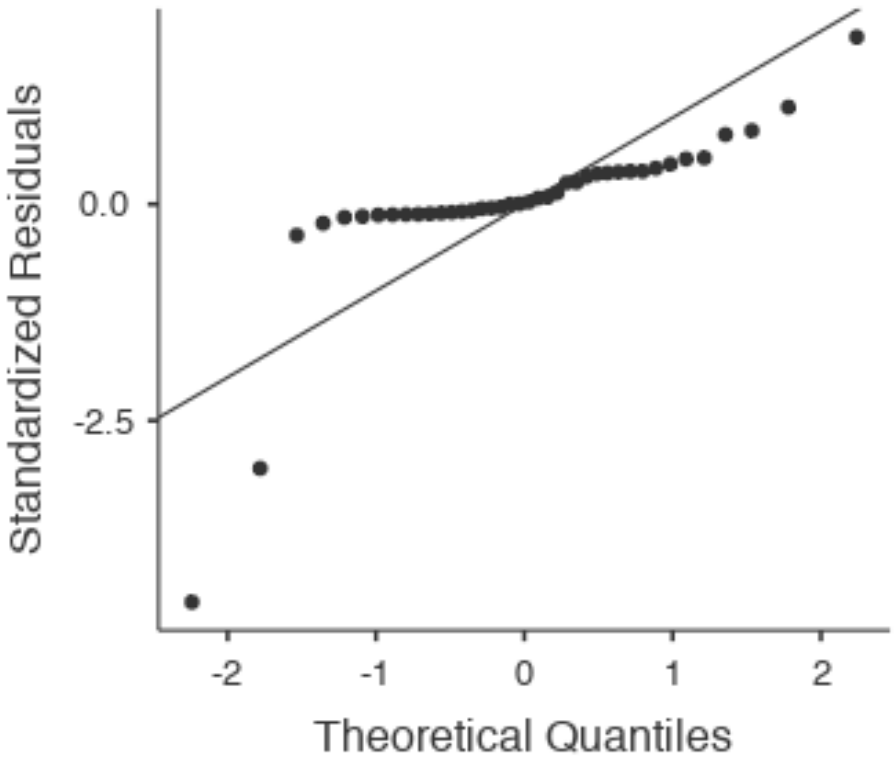
- The traditional and robust *t*-tests suggest evidence for a difference in means. But the raw effect size (i.e. mean difference) is nearly *twice as big as it should be*.
- Note too that the 95% confidence intervals are way off: [-15.5,-0.4]

	statistic	df	p	Mean difference	SE difference
Student's t	−2.152[a]	38.000	0.038	−7.998	3.717
Welch's t	−2.152	23.354	0.042	−7.998	3.717
Mann-Whitney U	68.000		<.001	−2.836	

- However, the comparison of ranks (in this case, same as comparing medians) is quite accurate: a sample difference of 3 (95% CI of [-4.2,-1.8]). This is an example of why *means should never be compared when your data contain substantial outliers*.

The Mann-Whitney test: it's not all about power

- The QQ-plot for these data is quite bad:



- Note the distinct "shelf" in the plot. This is characteristic of data that are "heavy tailed". This is a *very bad* kind of non-normality.

The Kruskal-Wallis one-way nonparametric "ANOVA"

- The MW-test is the nonparametric analogue to the independent samples *t*-test.
- But what if we want to compare *more than two groups*?
- The parametric approach would be a one-way ANOVA.
- The nonparametric analogue is the Kruskal-Wallis test.
  - Rationale is same as MW-test (based on ranks).
  - Implemented in Jamovi under "ANOVA" tab.
  - Same advantages over parametric one-way ANOVA as MW-test has over *t*-tests.
  - However, **cannot be generalized to more complex ANOVA models**. In particular, cannot be directly generalized to two-way ANOVA, with or without interactions.

Kruskal-Wallis example

- Recall our survey asking students about their current satisfaction with their academic program: 63 undergrads, 61 Master's students, and 73 PhD students responding on a 5-point Likert scale.
- We used a χ^2-test to analyze these ordinal data, and found weak evidence that academic satisfaction was related to degree program.
- We could apply a Kruskal-Wallis test to these same data to see if there is evidence that the "typical" Likert response is different between degree programs.

Kruskal-Wallis example

Contingency Tables

	Degree.Program			
Likert	U	M	P	Total
1	9	7	11	27
2	13	9	15	37
3	16	10	29	55
4	17	20	10	47
5	8	15	8	31
Total	63	61	73	197

χ^2 Tests

	Value	df	p
χ^2 continuity correction	17.712	8	0.023

- χ^2-test indicates weak evidence against the null hypothesis; i.e. sample proportions may be different for at least one of the degree programs.

Kruskal-Wallis example

- Now perform a Kruskal-Wallis test:

Kruskal-Wallis

	χ^2	df	p
Likert.resp	8.287	2	0.016

Dwass-Steel-Critchlow-Fligner pairwise comparisons

Pairwise comparisons - Likert.resp

		W	p
P	U	1.338	0.344
P	M	4.007	0.005
U	M	2.643	0.062

- Because we find evidence inconsistent with the null, we then ask for post hoc pairwise comparisons (note: p-values have been adjusted here to account for multiple comparisons).

A word of caution: example

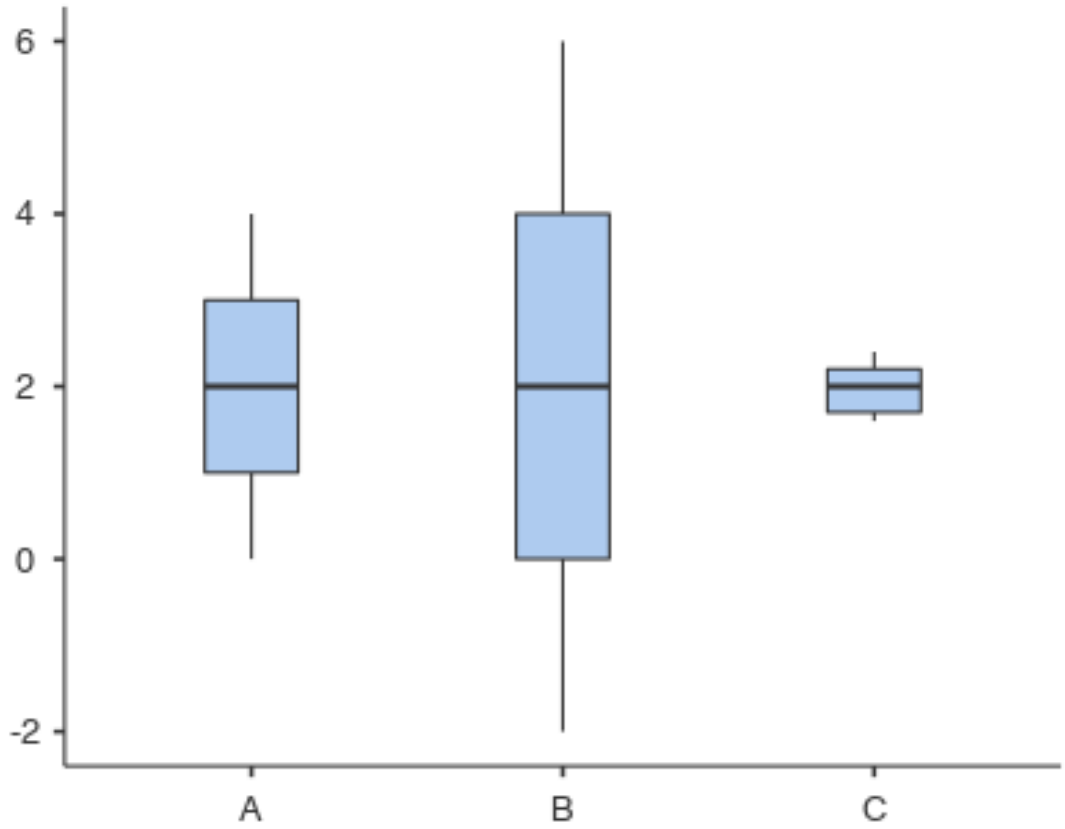
- While neither the MW-test nor the KW-test *require* homoskedasticity, they will also *not* be able to necessarily *detect departures* from homoskedasticity.
- This is because we can have groups with the same "typical" values, but very different variances.
- Consider the example comparison on 3 groups below: no evidence against the null.

Kruskal-Wallis

	χ^2	df	p
G	0.004	2	0.998

A word of caution: example

- Clearly, means and medians nearly identical for all 3 groups, but variances are quite different.



The Wilcoxon signed-rank test and the Friedman test

- The nonparametric analogue of the *paired t-test* is the Wilcoxon signed-rank test.
  - Rationale is similar to MW-test (based on ranks).
  - Implemented in Jamovi under "T-tests" tab, then 'Paired Samples T-test', then select the 'Wilcoxon rank' option.
  - Same advantages over paired t-tests as MW-test has over unpaired t-tests.
  - Can be generalized to handle *repeated measures* data on more than two time points; this is called the Friedman test.
  - However, the Friedman test **cannot be generalized to more complex RM-ANOVA models**. In particular, cannot be generalized to handle between-subjects factors, with or without interactions.

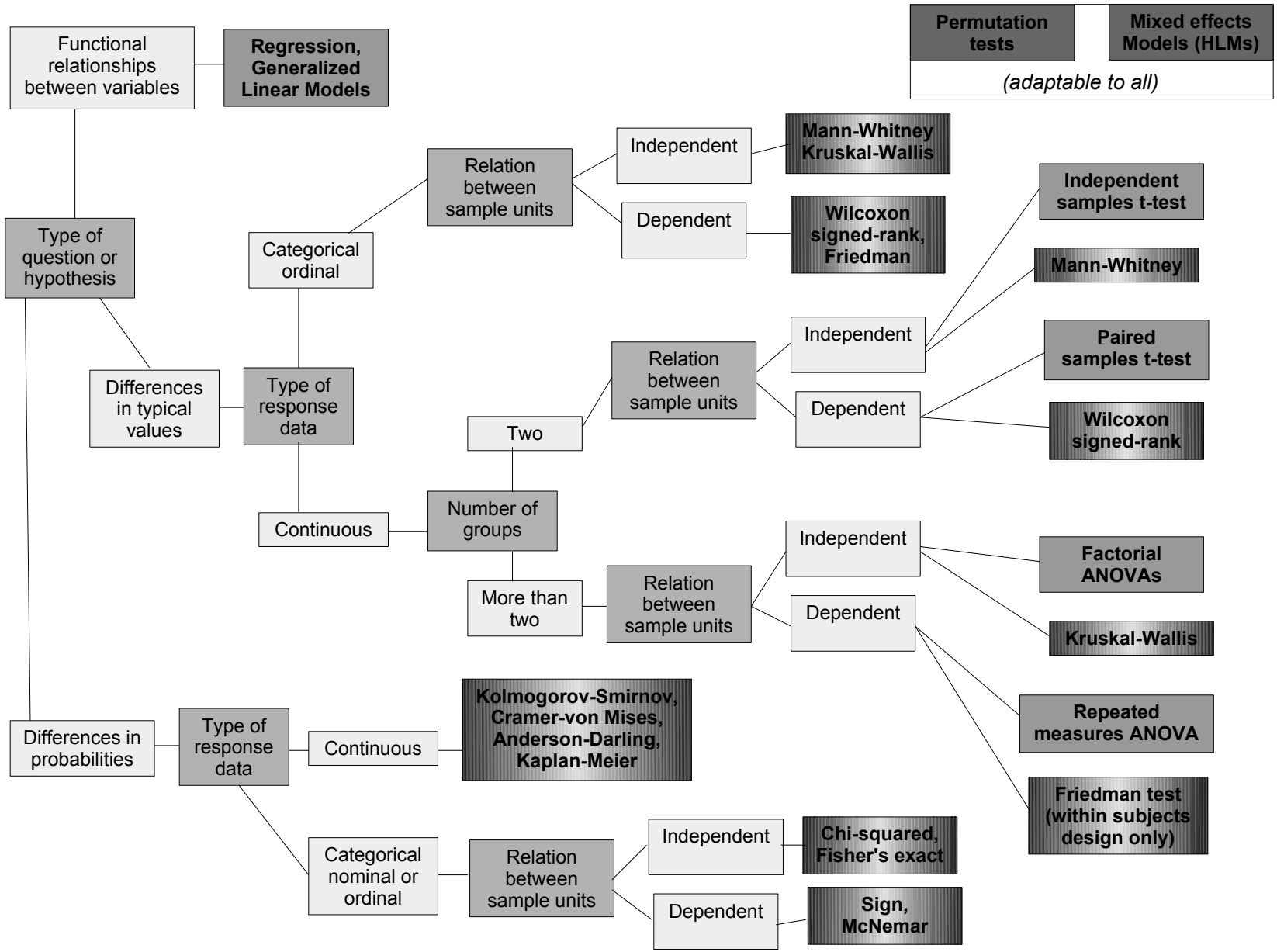
Further nonparametric procedures

- The previous nonparametric methods based solely on ranks are mathematically limited to one-way ANOVA-type situations.
- However, in practice, this problem can be circumvented by simply performing *multiple* one-way nonparametric ANOVAs, and then adjusting the overall significance level *a posteriori* (e.g. via Bonferonni).
- This is *less efficient* than performing a single, *n*-way ANOVA, but still effective if you want to use nonparametric methods with more than one explanatory variable.
- There are also more general and flexible nonparametric methods (theoretically) available for virtually *any* type of analysis: permutation tests.
- However, permutation tests can be difficult to perform and even harder to interpret.

Summary: when to consider nonparametric procedures

- When you have only **nominal or ordinal response data** (e.g. voting data).
- When you are worried about severe **violations of assumptions** of robust parametric procedures (e.g. t-tests, ordinary ANOVA of multimodal data).
- When you are worried about mild **violations of assumptions** of sensitive parametric procedures (e.g. sphericity of RM-ANOVA).
- When you have **outliers** in your data (robust procedures may also be available).
- When you have **too little data** to reasonably check assumptions of parametric procedures (remember: checking those assumptions requires enough data in *all* groups in order to have enough power to detect violations).

Statistical analysis (partial) cheat sheet (also on webpage)



Final words

My hope for your three main take-aways from this course:

- You are equipped with some common and useful **tools** for comparing *typical values* of a response variable of interest over different treatment groups and explanatory factors.
- You have gained some **technical understanding** about how these tools work and how to decide when their application is more or less appropriate.
- You have gained some **methodological confidence** to critically evaluate your own work and that of others. Remember: *do not ever let yourself be "math-washed"!*

Remember: your job is *not* to be a statistician; so when in doubt, *ask a statistician!* I'm always happy to talk stats or connect you with someone else who can help.