

Non-parametric tests

How to deal with Categorical data?

How to deal with cases where parametric assumptions are violated?

Recap

- Chi-square test
 - Goodness of fit
 - Independence of variables (2 variables) (for unrelated case)
 - Effect size
 - Median test for independence of samples
 - Log-Linear analysis (>2 categorical variables)
 - Binomial sign test (for related categorical variables)

Selecting a statistical test

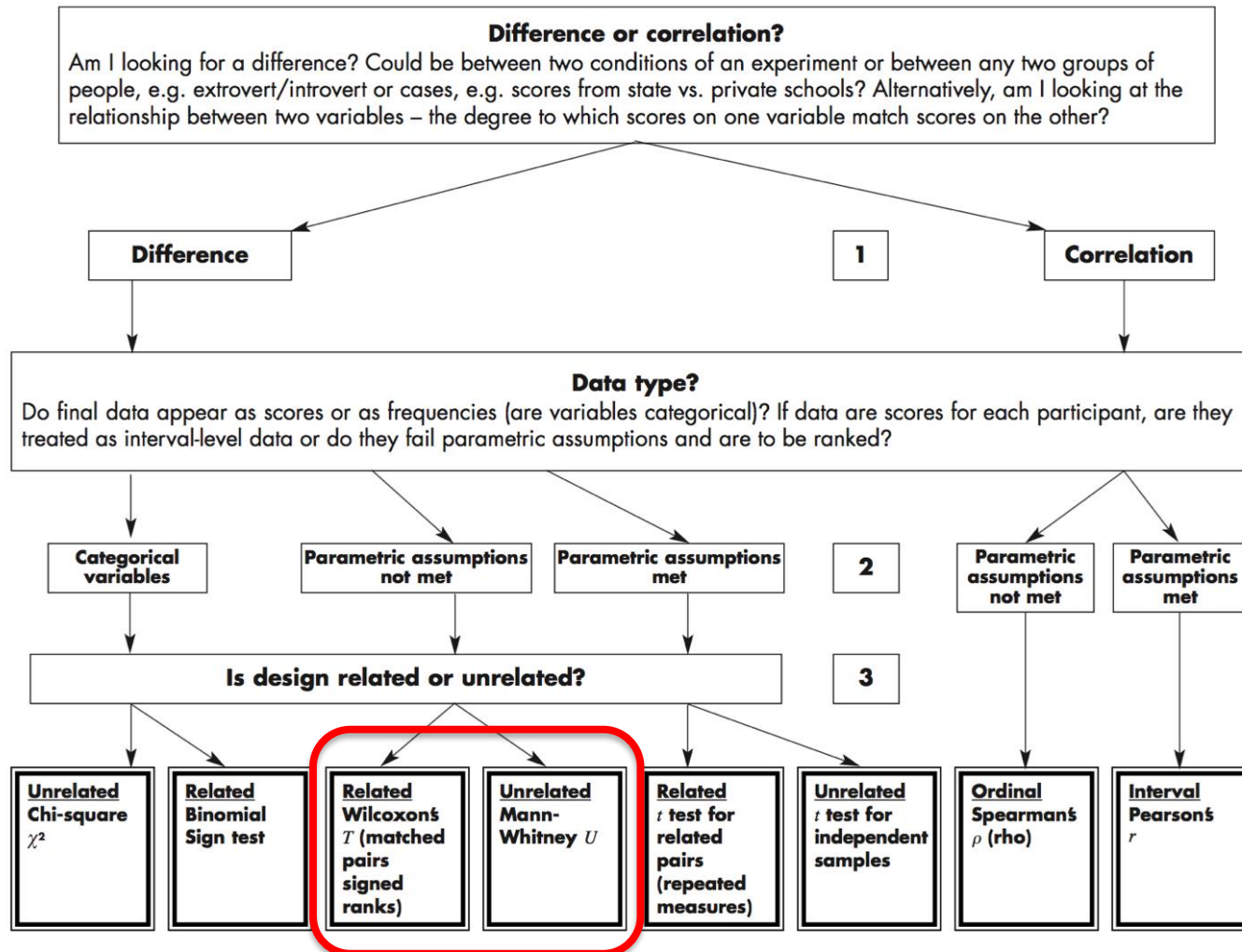


Figure 23.1 Choosing an appropriate two-sample test.

	Predictor variable	Outcome variable	Use in place of...
Spearman's <i>r</i>	<ul style="list-style-type: none"> Quantitative 	<ul style="list-style-type: none"> Quantitative 	Pearson's <i>r</i>
Chi square test of independence	<ul style="list-style-type: none"> Categorical 	<ul style="list-style-type: none"> Categorical 	Pearson's <i>r</i>
Sign test	<ul style="list-style-type: none"> Categorical 	<ul style="list-style-type: none"> Quantitative 	One-sample <i>t</i> -test
Kruskal-Wallis <i>H</i>	<ul style="list-style-type: none"> Categorical 3 or more groups 	<ul style="list-style-type: none"> Quantitative 	ANOVA
ANOSIM	<ul style="list-style-type: none"> Categorical 3 or more groups 	<ul style="list-style-type: none"> Quantitative 2 or more outcome variables 	MANOVA
Wilcoxon Rank-Sum test	<ul style="list-style-type: none"> Categorical 2 groups 	<ul style="list-style-type: none"> Quantitative groups come from 	Independent <i>t</i> -test
Wilcoxon Signed-rank test	<ul style="list-style-type: none"> Categorical 2 groups 	<ul style="list-style-type: none"> Quantitative groups come from the same population 	Paired <i>t</i> -test

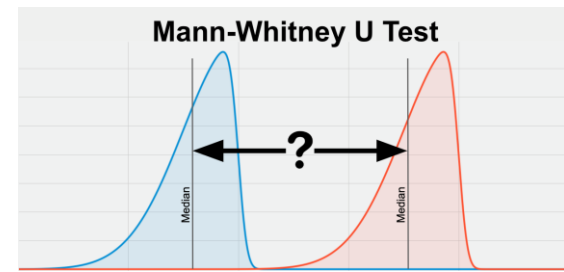
Choosing a nonparametric test

Non-parametric tests don't make as many assumptions about the data and are useful when one or more of the common statistical assumptions are violated. However, the inferences they make aren't as strong as with parametric tests.

Different types of tests

Test type	Between subjects designs (Independent samples)	Within subject designs (repeated measures/matched pairs)
Non-parametric (for categorical data)	Chi-square	<i>The binomial sign test</i>
Non-parametric (for ordinal data)	<i>Mann-Whitney U</i>	<i>Wilcoxon Signed-Rank test</i> <i>The binomial sign test</i>
Parametric	<i>Unrelated t-test (level of data: interval)</i>	<i>Related t-test (level of data: interval)</i>

Mann-Whitney U Test



- between subjects design
- skewed distribution
- used on ordinal non-normal data
- **assumption:**
 - a real difference between two populations should cause the scores in one sample to be generally larger than the other;
 - if two samples are combined and all scores are ranked, then the larger ranks should be concentrated in one sample and smaller ranks in the other
 - eg: Likert items (e.g., a 7-point scale from "strongly agree" through to "strongly disagree")

Mann-Whitney U test

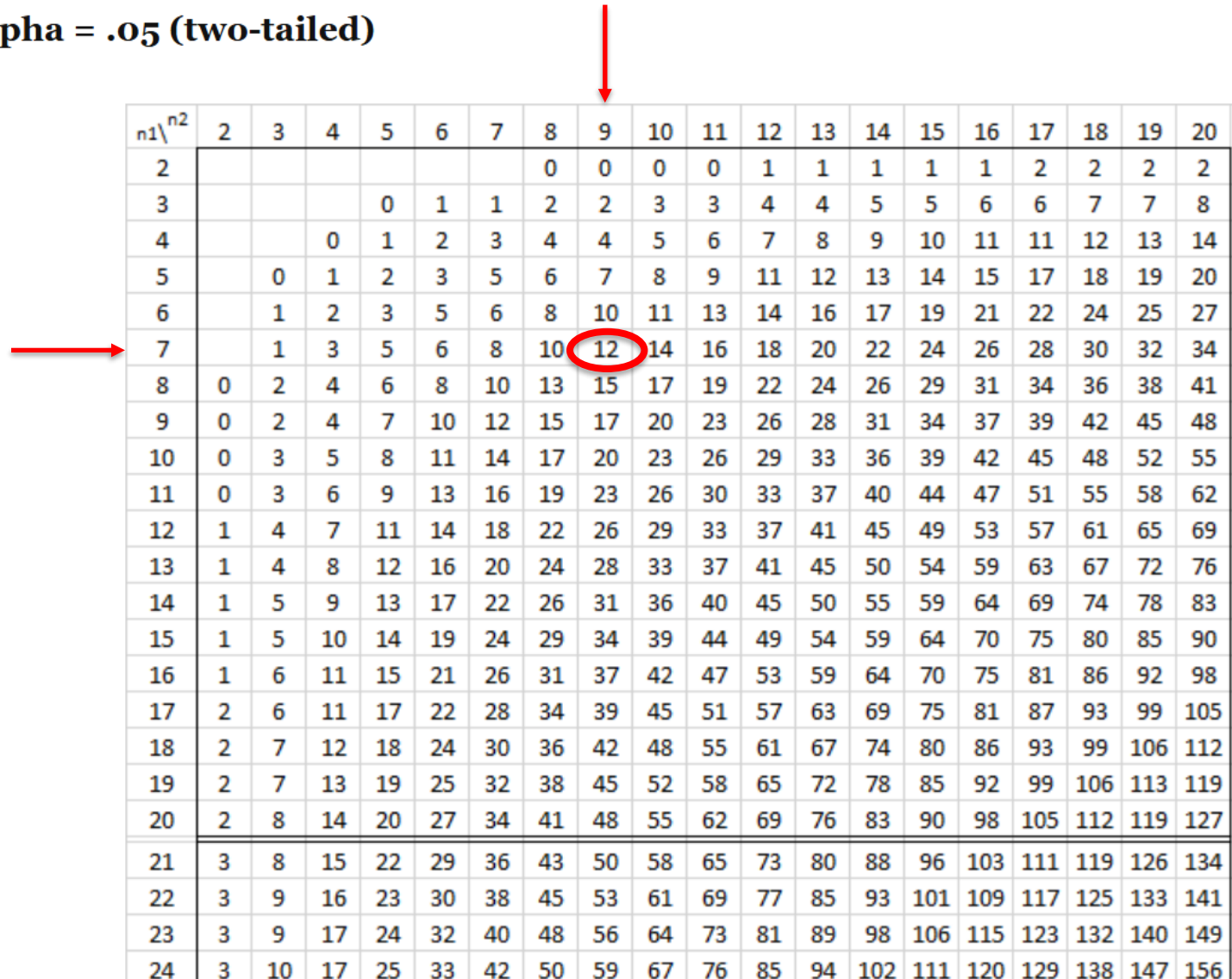
- ex: children's tendency to stereotype according to traditional gender roles if they have working mothers vs not

Full-time jobs		No job outside home	
Score	Points	Score	Points
17	9	19	6
32	7	63	0
39	6.5	78	0
27	8	29	4
58	6	39	1.5
25	8	59	0
31	7	77	0
		81	0
		68	0
Totals:	$51.5 = U_1$		$11.5 = U_2$
U is the lower of 51.5 and 11.5, so U is 11.5			

- the observed U value should be less than or equal to critical U value in order to reject H_0

Mann-Whitney U Table

Alpha = .05 (two-tailed)



$n_1 \backslash n_2$	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2							0	0	0	0	1	1	1	1	1	2	2	2	2
3				0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4			0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	14
5		0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6		1	2	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7		1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
8	0	2	4	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
9	0	2	4	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
10	0	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
11	0	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
12	1	4	7	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
13	1	4	8	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	76
14	1	5	9	13	17	22	26	31	36	40	45	50	55	59	64	69	74	78	83
15	1	5	10	14	19	24	29	34	39	44	49	54	59	64	70	75	80	85	90
16	1	6	11	15	21	26	31	37	42	47	53	59	64	70	75	81	86	92	98
17	2	6	11	17	22	28	34	39	45	51	57	63	69	75	81	87	93	99	105
18	2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99	106	112
19	2	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	119
20	2	8	14	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127
21	3	8	15	22	29	36	43	50	58	65	73	80	88	96	103	111	119	126	134
22	3	9	16	23	30	38	45	53	61	69	77	85	93	101	109	117	125	133	141
23	3	9	17	24	32	40	48	56	64	73	81	89	98	106	115	123	132	140	149
24	3	10	17	25	33	42	50	59	67	76	85	94	102	111	120	129	138	147	156

> two groups - Kruskal-Wallis test

Mann-Whitney U test

- ex: children's tendency to stereotype according to traditional gender roles if they have working mothers vs not

Full-time jobs		No job outside home	
Score	Points	Score	Points
17	9	19	6
32	7	63	0
39	6.5	78	0
27	8	29	4
58	6	39	1.5
25	8	59	0
31	7	77	0
		81	0
		68	0
Totals:	$51.5 = U_1$		$11.5 = U_2$
U is the lower of 51.5 and 11.5, so U is 11.5			

critical U value = 12

$$\alpha < .05$$

children of working mothers are less likely to use gender-role stereotypes

REJECTED
 H_0

Kruskal-Wallis Test

Aim



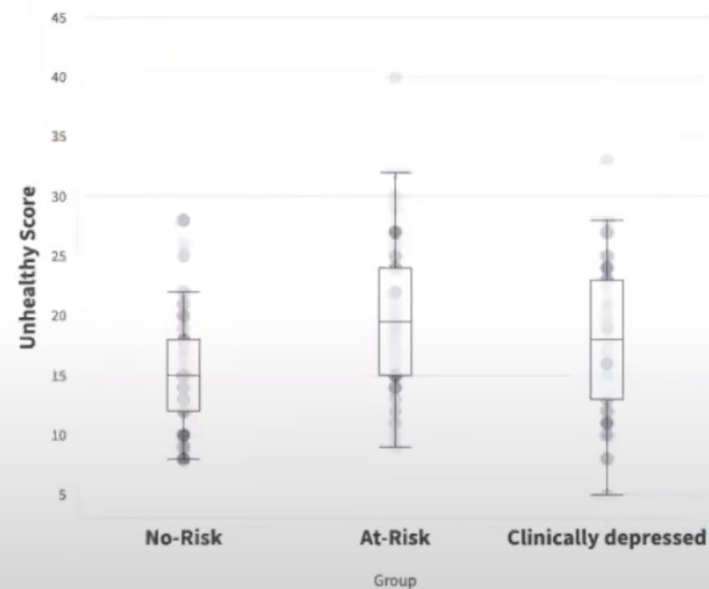
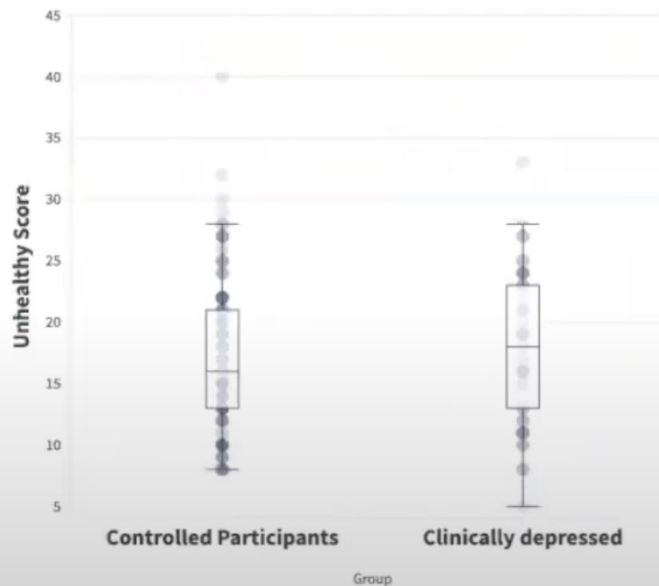
Clinically Depressed Cohort (DC)



Controlled Participants (CP)

To investigate musical engagement strategies via HUMS in DC compared to control participants (CP) from the community

Results: Group Differences for *Unhealthy* Scores



Different types of tests

Test type	Between subjects designs (Independent samples)	Within subject designs (repeated measures/matched pairs)
Non-parametric (for categorical data)	Chi-square	<i>The binomial sign test</i>
Non-parametric (for ordinal data)	<i>Mann-Whitney U</i>	<i>Wilcoxon Signed-Rank Test</i>
Parametric	<i>Unrelated t-test (level of data: interval)</i>	<i>Related t-test (level of data: interval)</i>

Wilcoxon Signed-Rank Test

- ordinal level (tests based on rank order)
- within subjects design (related, repeated-measures/matched pairs)
- null hypothesis as the claim that the two populations from which scores are sampled are identical
- most of the time this is more specifically that the two medians are equal (not means because we are working at the ordinal level)
- the observed W value should be less than or equal to critical W value in order to reject H_0

Note: Can also be used for discrete related samples.

If the paired data violate parametric assumptions,

Convert the paired scores into difference and rank them.

Wilcoxon Signed-Rank Test

- example:
 - assess if students performed better in the mock exam than the final GRE exam

H_0 : Population median difference = 0

H_1 : Population median difference > 0 (1-tail)

Wilcoxon Signed-Rank Test

Student	Mock	Real	Diff(d)	Rank
1	316	320	-4	-4.5
2	324	319	5	6
3	317	318	-1	-1.5
4	323	314	9	10
5	333	333	0	n/a
6	329	321	8	9
7	328	311	17	12
8	319	309	10	11
9	320	318	2	3
10	314	321	-7	-8
11	309	315	-6	-7
12	323	319	4	4.5
13	335	334	1	1.5

$$T_+ = 57 \quad T_- = 21$$

$$W_{\text{stat}} = \min(T_+, T_-) = 21$$

(> critical W value 17
 $\alpha < .05$)

n	Two-Tailed Test		One-Tailed Test	
	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$
5	--	--	0	--
6	0	--	2	--
7	2	--	3	0
8	3	0	5	1
9	5	1	8	3
10	8	3	10	5
11	10	5	13	7
12	13	7	17	9
13	17	9	21	12
14	21	12	25	15

H_0



Different types of tests

Test type	Between subjects designs (Independent samples)	Within subject designs (repeated measures/matched pairs)
Non-parametric (for categorical data)	Chi-square	<i>The binomial sign test</i>
Non-parametric (for ordinal data)	<i>Mann-Whitney U</i>	<i>Wilcoxon Signed-Rank Test</i> <i>The binomial sign test</i>
Parametric	<i>Unrelated t-test (level of data: interval)</i>	<i>Related t-test (level of data: interval)</i>

The Binomial Sign Test

Categorical data

- Within subjects design
- Items are dichotomous and nominal
- may be reduced from interval or ordinal level
- two dependent samples should be paired or matched

The Binomial Sign Test

A	B	C	D	E	
Client	Self-image rating before therapy	Self-image rating after 3 months' therapy	Difference (C – B)	Sign of difference	
a	3	7	4	+	
b	12	18	6	+	
c	9	5	-4	-	
d	7	7	0		
e	8	12	4	+	$S = 1$
f	1	5	4	+	
g	15	16	1	+	
h	10	12	2	+	
i	11	15	4	+	
j	10	17	7	+	

Table 17.6 Self-image scores before and after three months' therapy.

- the observed S value should be less than or equal to critical S value in order to reject H_0

Numbers in the table represent $p(X=x)$ for a binomial distribution with n trials and probability of success p .

Binomial probabilities:

$$\binom{n}{x} p^x (1-p)^{n-x}$$

		p										
n	x	0.1	0.2	0.25	0.3	0.4	0.5	0.6	0.7	0.75	0.8	0.9
8	0	0.430	0.168	0.100	0.058	0.017	0.004	0.001	0.000	0.000	0.000	0.000
	1	0.383	0.336	0.267	0.198	0.090	0.031	0.008	0.001	0.000	0.000	0.000
	2	0.149	0.294	0.311	0.296	0.209	0.109	0.041	0.010	0.004	0.001	0.000
	3	0.033	0.147	0.208	0.254	0.279	0.219	0.124	0.047	0.023	0.009	0.000
	4	0.005	0.046	0.087	0.136	0.232	0.273	0.232	0.136	0.087	0.046	0.005
	5	0.000	0.009	0.023	0.047	0.124	0.219	0.279	0.254	0.208	0.147	0.033
	6	0.000	0.001	0.004	0.010	0.041	0.109	0.209	0.296	0.311	0.294	0.149
	7	0.000	0.000	0.000	0.001	0.008	0.031	0.090	0.198	0.267	0.336	0.383
9	0	0.000	0.000	0.000	0.000	0.001	0.004	0.017	0.058	0.100	0.168	0.430
	1	0.387	0.134	0.075	0.040	0.010	0.002	0.000	0.000	0.000	0.000	0.000
	2	0.387	0.302	0.225	0.156	0.060	0.018	0.004	0.000	0.000	0.000	0.000
	3	0.172	0.302	0.300	0.267	0.161	0.070	0.021	0.004	0.001	0.000	0.000
	4	0.045	0.176	0.234	0.267	0.251	0.164	0.074	0.021	0.009	0.003	0.000
	5	0.007	0.066	0.117	0.172	0.251	0.246	0.167	0.074	0.039	0.017	0.001
	6	0.001	0.017	0.039	0.074	0.167	0.246	0.251	0.172	0.117	0.066	0.007
	7	0.000	0.003	0.009	0.021	0.074	0.164	0.251	0.267	0.234	0.176	0.045
10	0	0.000	0.000	0.001	0.004	0.021	0.070	0.161	0.267	0.300	0.302	0.172
	1	0.000	0.000	0.000	0.000	0.004	0.018	0.060	0.156	0.225	0.302	0.387
	2	0.000	0.000	0.000	0.000	0.000	0.002	0.010	0.040	0.075	0.134	0.387
	3	0.349	0.107	0.056	0.028	0.006	0.001	0.000	0.000	0.000	0.000	0.000
	4	0.387	0.268	0.188	0.121	0.040	0.010	0.002	0.000	0.000	0.000	0.000
	5	0.194	0.302	0.282	0.233	0.121	0.044	0.011	0.001	0.000	0.000	0.000
	6	0.057	0.201	0.250	0.267	0.215	0.117	0.042	0.009	0.003	0.001	0.000
	7	0.011	0.088	0.146	0.200	0.251	0.205	0.111	0.037	0.016	0.006	0.000
11	0	0.001	0.026	0.058	0.103	0.201	0.246	0.201	0.103	0.058	0.026	0.001
	1	0.000	0.006	0.016	0.037	0.111	0.205	0.251	0.200	0.146	0.088	0.011
	2	0.000	0.001	0.003	0.009	0.042	0.117	0.215	0.267	0.250	0.201	0.057
	3	0.000	0.000	0.000	0.001	0.011	0.044	0.121	0.233	0.282	0.302	0.194
	4	0.000	0.000	0.000	0.000	0.002	0.010	0.040	0.121	0.188	0.268	0.387
	5	0.000	0.000	0.000	0.000	0.000	0.001	0.006	0.028	0.056	0.107	0.349
	6	0.314	0.086	0.042	0.020	0.004	0.000	0.000	0.000	0.000	0.000	0.000
	7	0.384	0.236	0.155	0.093	0.027	0.005	0.001	0.000	0.000	0.000	0.000
12	0	0.213	0.295	0.258	0.200	0.089	0.027	0.005	0.001	0.000	0.000	0.000
	1	0.071	0.221	0.258	0.257	0.177	0.081	0.023	0.004	0.001	0.000	0.000
	2	0.016	0.111	0.172	0.220	0.236	0.161	0.070	0.017	0.006	0.002	0.000
	3	0.002	0.039	0.080	0.132	0.221	0.226	0.147	0.057	0.027	0.010	0.000
	4	0.000	0.010	0.027	0.057	0.147	0.226	0.221	0.132	0.080	0.039	0.002
	5	0.000	0.002	0.006	0.017	0.070	0.161	0.236	0.220	0.172	0.111	0.016
	6	0.000	0.000	0.001	0.004	0.023	0.081	0.177	0.257	0.258	0.221	0.071
	7	0.000	0.000	0.000	0.001	0.005	0.027	0.089	0.200	0.258	0.295	0.213
13	0	0.000	0.000	0.000	0.000	0.001	0.005	0.027	0.093	0.155	0.236	0.384
	1	0.000	0.000	0.000	0.000	0.000	0.000	0.004	0.020	0.042	0.086	0.314

[From the Binomial Distribution Table]
The p-value is 0.01, which is smaller than the **alpha-level** of 0.05. We can reject the null hypothesis and say there is a significant difference.

S-Table Lookup



<i>n</i>	One tailed, $\alpha = 0.005$	$\alpha = 0.01$	$\alpha = 0.025$	$\alpha = 0.05$
	Two tailed, $\alpha = 0.01$	$\alpha = 0.02$	$\alpha = 0.05$	$\alpha = 0.10$
8	0	0	0	1
9	0	0	1	1
10	0	0	1	1
11	0	1	1	2
12	1	1	2	2
13	1	1	2	3
14	1	2	3	3
15	2	2	3	3
16	2	2	3	4
17	2	3	4	4
18	3	3	4	5
19	3	4	4	5
20	3	4	5	5
21	4	4	5	6
22	4	5	5	6
23	4	5	6	7
24	5	5	6	7
25	5	6	6	7

The Binomial Sign Test

A	B	C	D	E	
Client	Self-image rating before therapy	Self-image rating after 3 months' therapy	Difference (C – B)	Sign of difference	
a	3	7	4	+	
b	12	18	6	+	
c	9	5	-4	-	
d	7	7	0		
e	8	12	4	+	$S = 1$
f	1	5	4	+	
g	15	16	1	+	
h	10	12	2	+	
i	11	15	4	+	
j	10	17	7	+	

Table 17.6 Self-image scores before and after three months' therapy.

critical S value = 1

$$\alpha \leq .05$$

REJECTED

H_0

	Predictor variable	Outcome variable	Use in place of...
Spearman's r	<ul style="list-style-type: none"> Quantitative 	<ul style="list-style-type: none"> Quantitative 	Pearson's r
Chi square test of independence	<ul style="list-style-type: none"> Categorical 	<ul style="list-style-type: none"> Categorical 	Pearson's r
Sign test	<ul style="list-style-type: none"> Categorical 	<ul style="list-style-type: none"> Quantitative 	One-sample t -test
Kruskal-Wallis H	<ul style="list-style-type: none"> Categorical 3 or more groups 	<ul style="list-style-type: none"> Quantitative 	ANOVA
ANOSIM	<ul style="list-style-type: none"> Categorical 3 or more groups 	<ul style="list-style-type: none"> Quantitative 2 or more outcome variables 	MANOVA
Wilcoxon Rank-Sum test	<ul style="list-style-type: none"> Categorical 2 groups 	<ul style="list-style-type: none"> Quantitative groups come from 	Independent t -test
Wilcoxon Signed-rank test	<ul style="list-style-type: none"> Categorical 2 groups 	<ul style="list-style-type: none"> Quantitative groups come from the same population 	Paired t -test

Choosing a nonparametric test

Non-parametric tests don't make as many assumptions about the data, and are useful when one or more of the common statistical assumptions are violated. However, the inferences they make aren't as strong as with parametric tests.

Permutation Tests

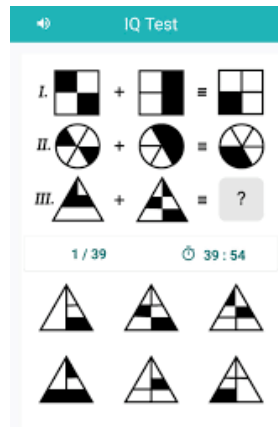
- rely on randomizations of the observed data and typically seek to quantify the null distribution in order to perform hypothesis testing
- permute the data in a way that removes some aspects of the statistical structure and evaluate how likely is the observed statistic to occur if the null hypothesis was true
- the test statistic is compared against a theoretical distribution of test statistics expected under the H_0 .
- determine the statistical significance of a model by computing a test statistic on the dataset and then for many random permutations of that data
 - > If the model is significant, the original test statistic value should lie at one of the tails of the null hypothesis distribution.

Why choose Permutation Tests?

- small sample size
- assumptions (for parametric approach) not met
- test statistic other than comparing means/medians
- difficult to estimate SE for test statistic

EXAMPLE

H_A = Engineering students have higher IQ than Art students



Group 1: CS

Group 2: Art

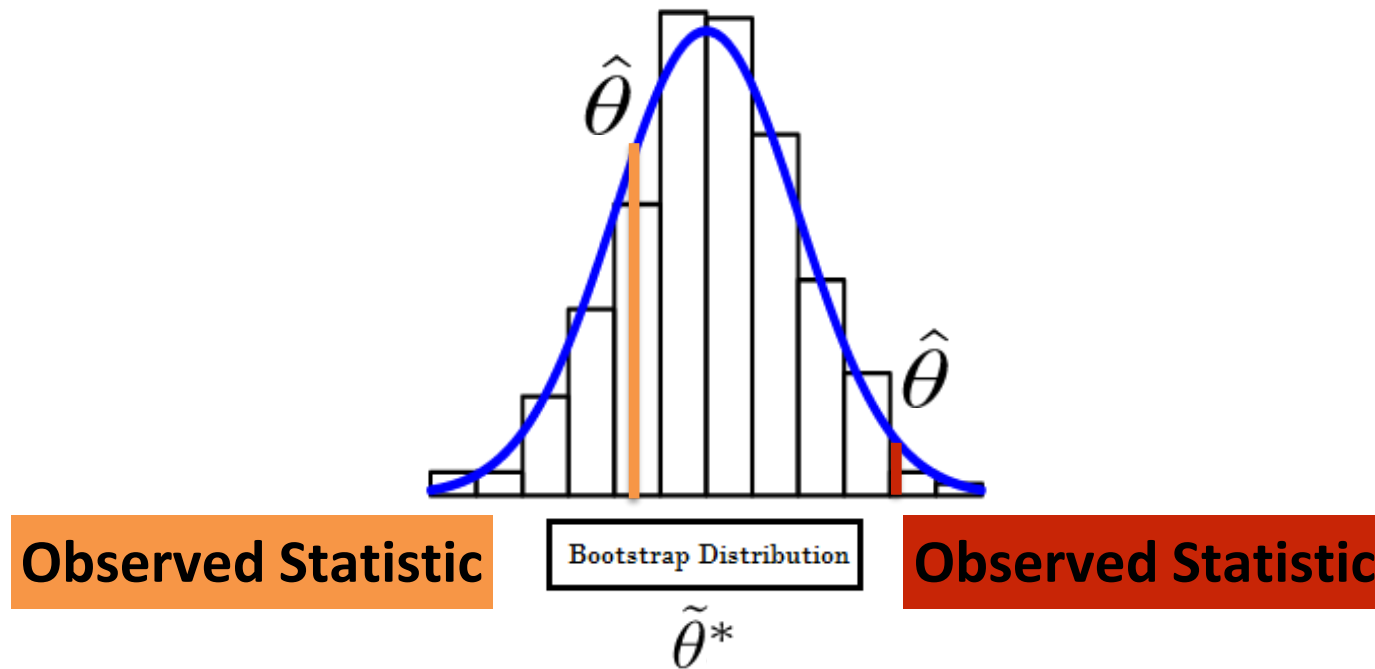
Result: $\text{mean}(\text{IQ}_{\text{CS}}) > \text{mean}(\text{IQ}_{\text{Art}})$

how certain am I that i can reject the null-hypothesis?
what is the probability that this result can appear due to chance alone?

Permutation Tests (Group Differences)

- eg: context of *difference in mean* of both groups
 1. randomly permute (or “shuffle”) the data into both groups
 2. recalculate the difference in mean
 3. repeat steps 1-2 several times to obtain the resulting samples to characterize the null distribution (i.e. the distribution we would expect if there were no statistical relationship between x and y)
 4. significance estimation: evaluate the proportion of times you get at least a value equal to or more extreme than the observed/actual statistic (*difference in mean*)

likely to accept null hypothesis or not?



Estimate

$\hat{\theta}$ = difference in medians, t-statistic, mode, std, etc..

COMMENT • 20 MARCH 2019

Scientists rise up against statistical significance

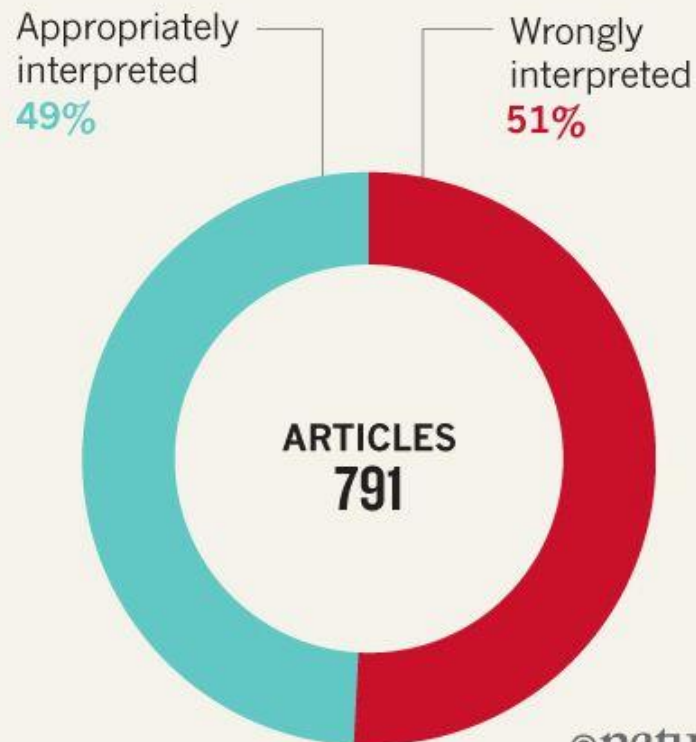
Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.

Valentin Amrhein , Sander Greenland & Blake McShane

WRONG INTERPRETATIONS

An analysis of 791 articles across 5 journals* found that around half mistakenly assume non-significance means no effect.

*Data taken from: P. Schatz *et al.* *Arch. Clin. Neuropsychol.* **20**, 1053–1059 (2005); F. Fidler *et al.* *Conserv. Biol.* **20**, 1539–1544 (2006); R. Hoekstra *et al.* *Psychon. Bull. Rev.* **13**, 1033–1037 (2006); F. Bernardi *et al.* *Eur. Sociol. Rev.* **33**, 1–15 (2017).



COMMENT • 20 MARCH 2019

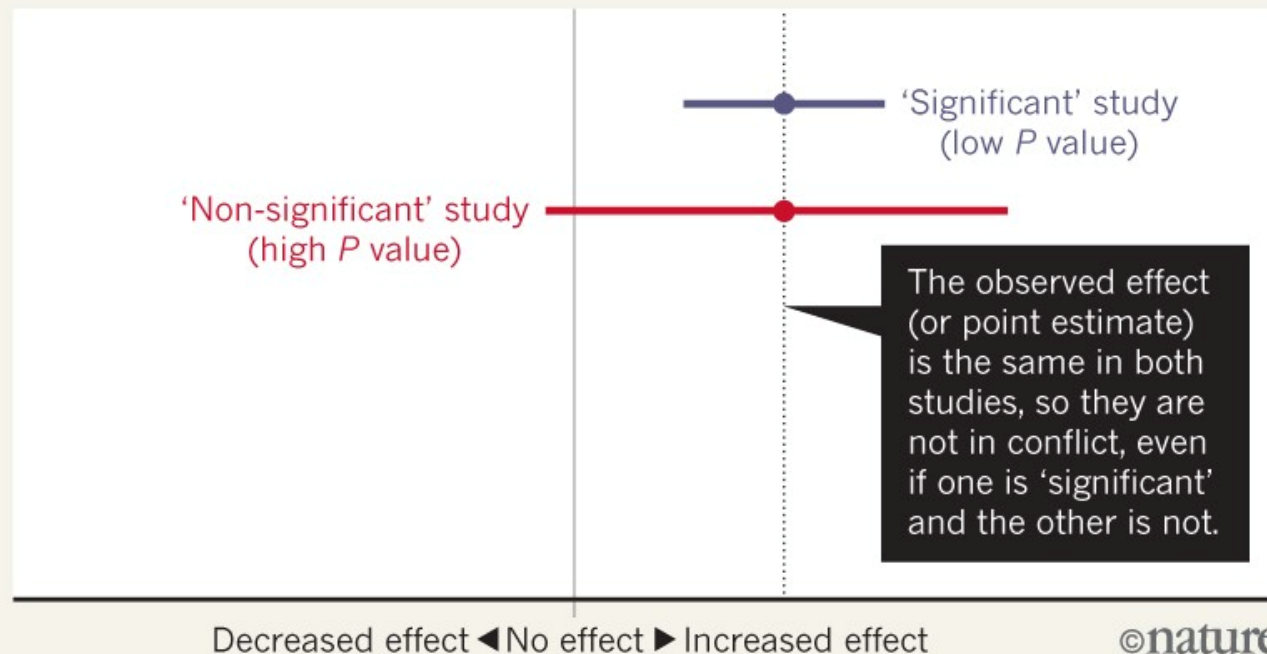
Scientists rise up against statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.

Valentin Amrhein , Sander Greenland & Blake McShane

BEWARE FALSE CONCLUSIONS

Studies currently dubbed ‘statistically significant’ and ‘statistically non-significant’ need not be contradictory, and such designations might cause genuine effects to be dismissed.



p -hacking!!

1. Stop collecting data once $p < .05$
2. Analyze many measures, but report only those with $p < .05$.
3. Collect and analyze many conditions, but only report those with $p < .05$.
4. Use covariates to get $p < .05$.
5. Exclude participants to get $p < .05$.
6. Transform the data to get $p < .05$.

Different types of tests

Test type	Between subjects designs (Independent samples)	Within subject designs (repeated measures/matched pairs)
Non-parametric (for categorical data)	Chi-square	<i>The binomial sign test</i>
Non-parametric (for ordinal data)	<i>Mann-Whitney U</i>	<i>Wilcoxon Signed-Rank Test</i>
Parametric	<i>Unrelated t-test (level of data: interval)</i>	<i>Related t-test (level of data: interval)</i>

Next Class