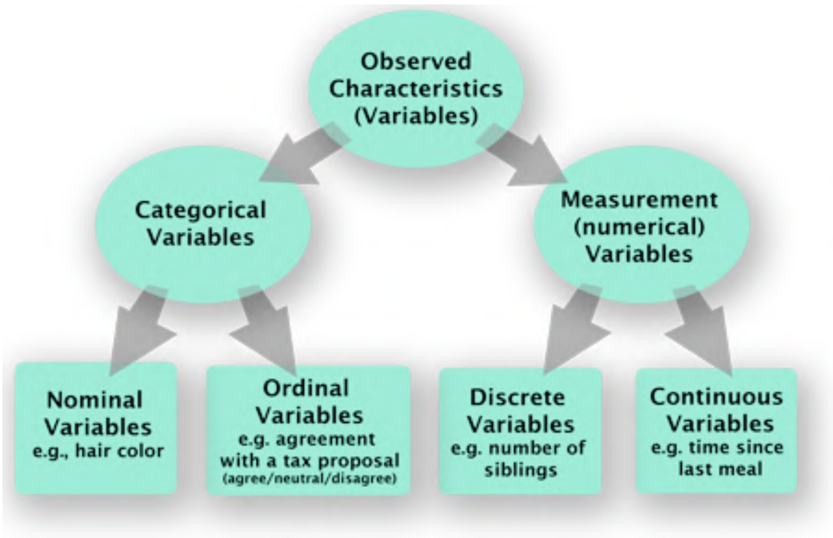
Non-parametric tests

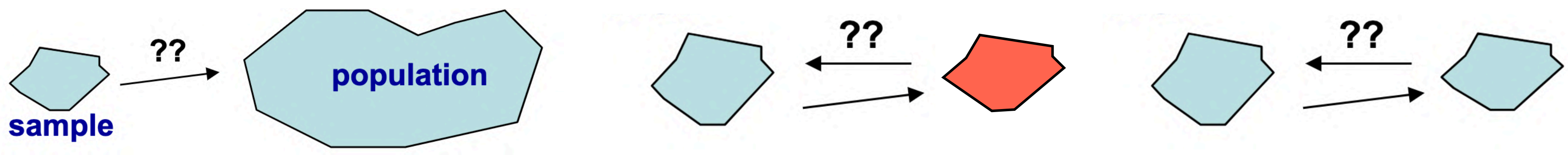
Selecting a statistical test

- Different tests are used according to the level of measurement:
  - Interval
  - Ordinal
  - Categorical

- Parametric vs non-parametric (makes no assumption on the population distribution or sample size) assumptions



Selecting a statistical test



- Different tests are used for varying amount of groups/ conditions:
  - two samples
  - > two samples
- Different tests are used for related versus unrelated designs:
  - unrelated samples = between subjects designs
  - related samples = within subjects designs & matched pairs

Selecting a statistical test

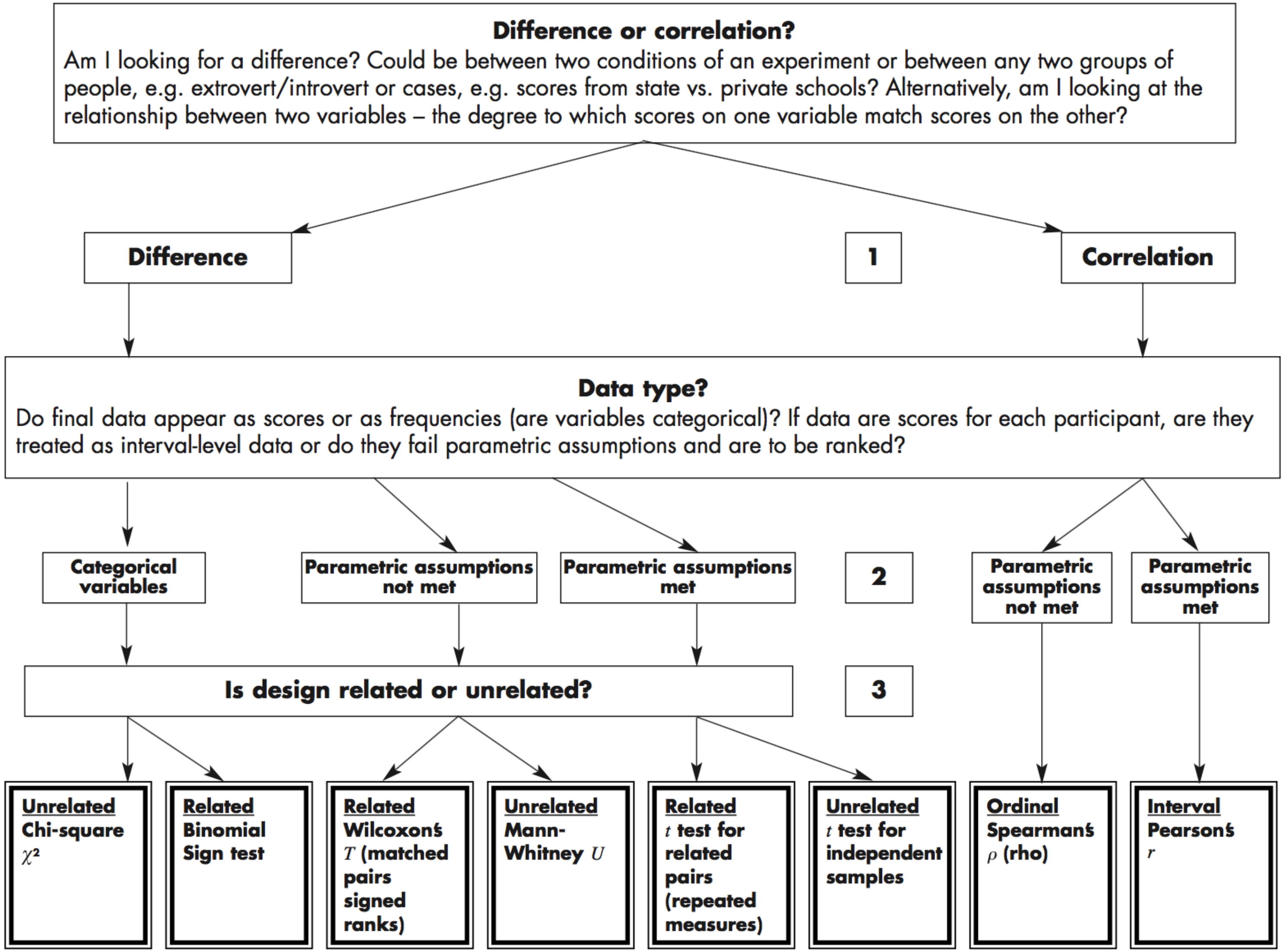


Figure 23.1 Choosing an appropriate two-sample test.

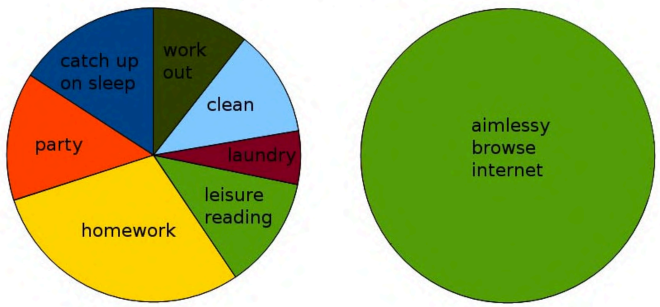
Different types of tests

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

Chi-Square Test

Theoretical
categorical disribution
vs
Observed
categorical distribution

Weekend in college



Expectation Reality

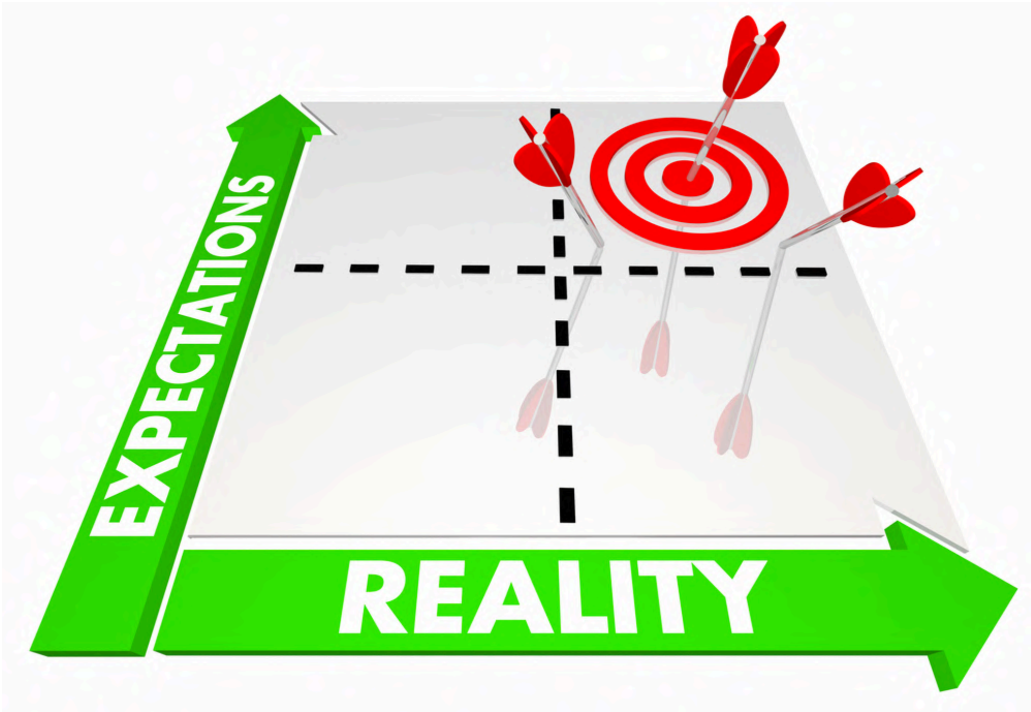
preference for one brand

Chi-Square test

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

H_0

H_A



Data collected

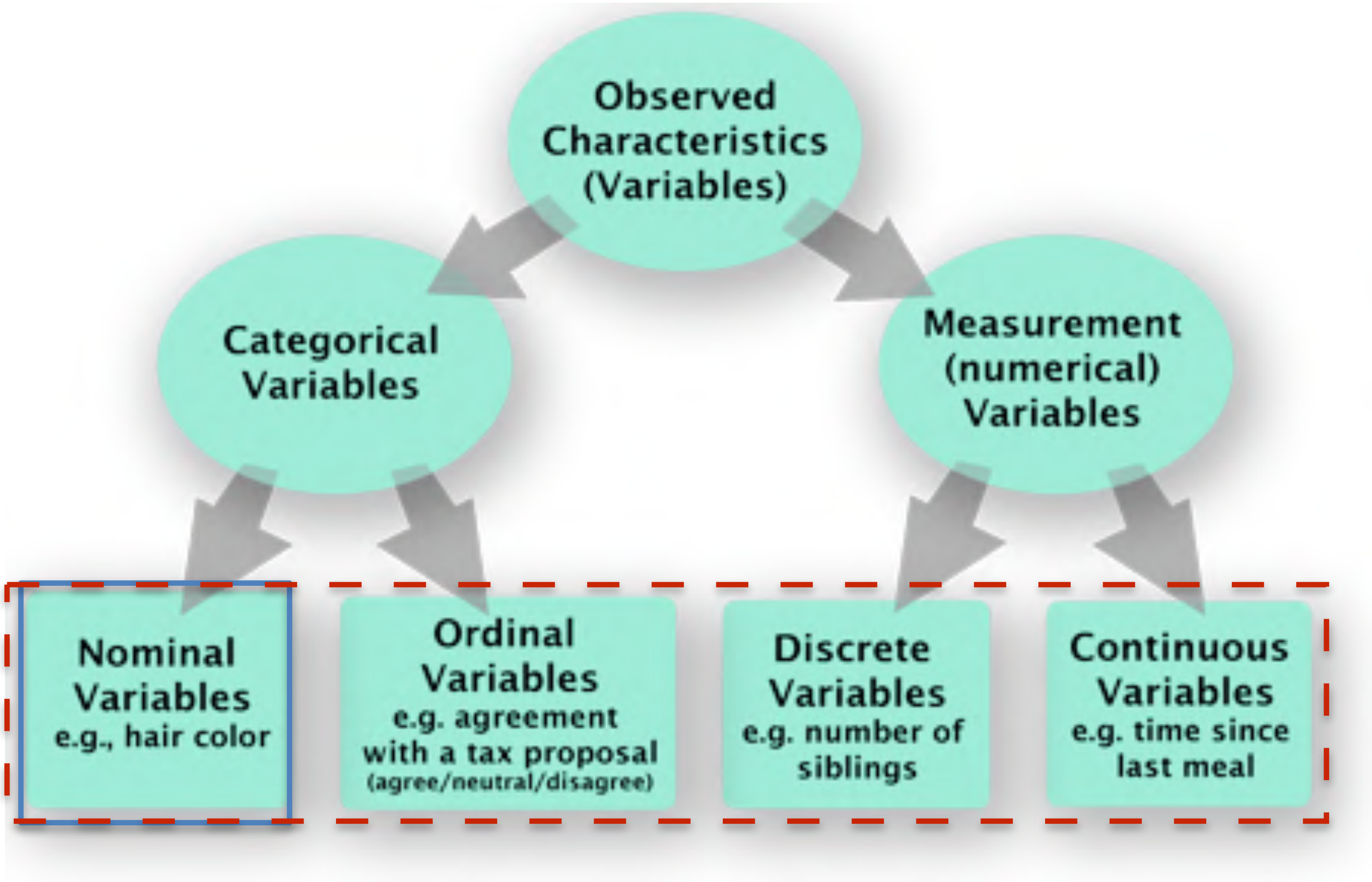
Chi-Square Test

- Goodness-of-fit:
  - compare the observed sample distribution with the expected probability distribution
  - H_0 = no difference from a known population
- Chi-Square fit test:
  - determines how well theoretical distribution fits the empirical distribution
  - H_0 = no difference, equal proportions

Chi-Square Test

- Test for Independence (for two variables):
  - test *relationship* between two separate variables
  - H_0 = there is no relationship between the variables
    - eg: females prefer pepsi more than males

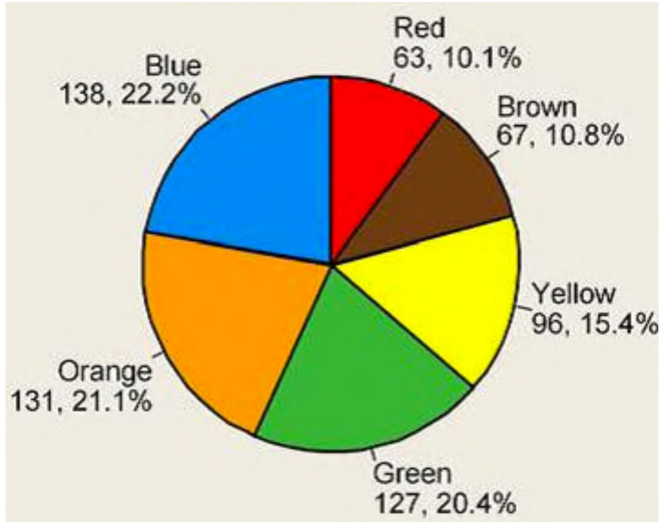
Chi-Square Test



Chi-Square test/goodness-of-fit

H_0

M&Ms Color Distribution % according to their website

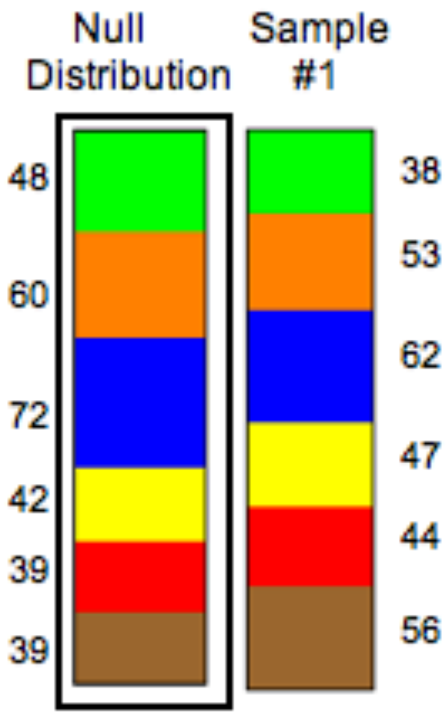


H_0: The color distribution is equal

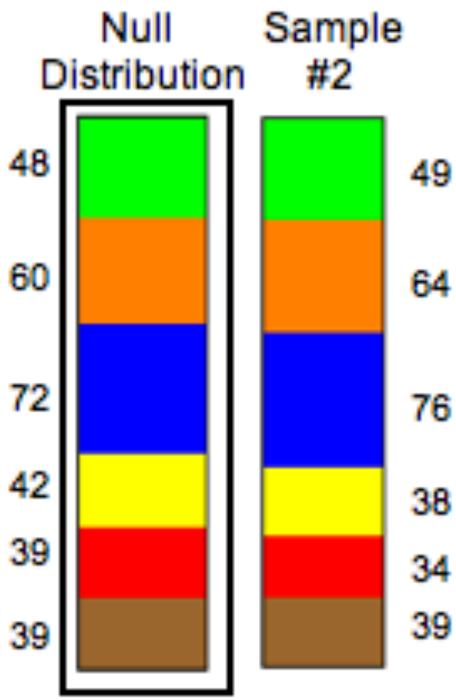
revised H_0: The color distribution is 13% brown, 13% red, 14% yellow, 24% blue, 20% orange, 16% green

H_A: The color distribution is different from 13% brown, 13% red, 14% yellow, 24% blue, 20% orange, 16% green

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$



$\chi^2 = 12.94$



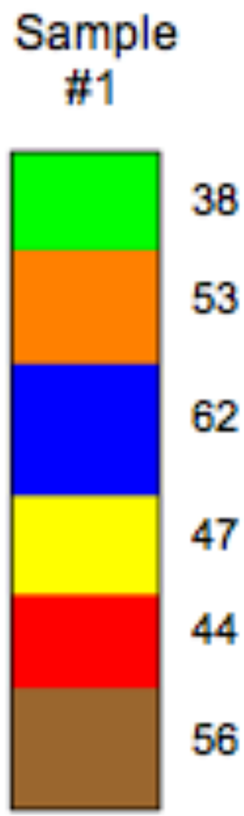
$\chi^2 = 1.53$

df = ?

H_0?

Chi-Square distribution & df

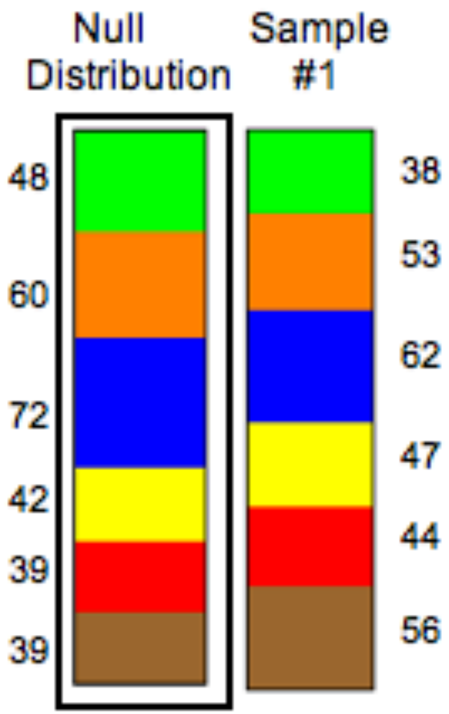
- Degrees of freedom for goodness-of-fit
  - number of cells you would need to calculate all other cell values, assuming we know marginal values
- df = *C*-1, *C* = no. of categories

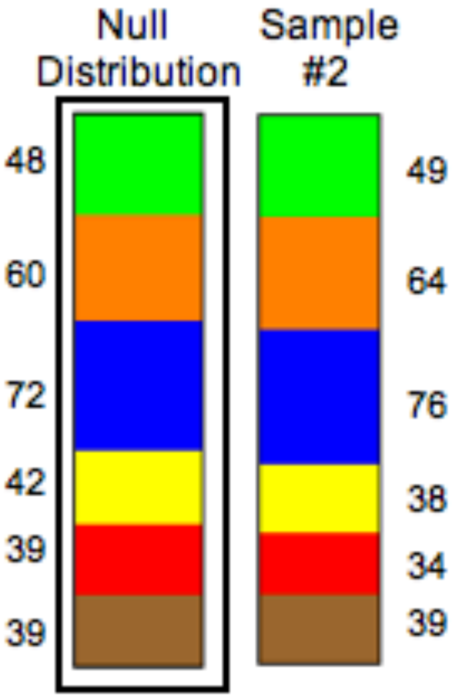


df = ?

https://www.khanacademy.org/math/statistics-probability/inference-categorical-data-chi-square-tests/chi-square-goodness-of-fit-tests/v/chi-square-distribution-introduction

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$





$\chi^2 = 12.94$

$\chi^2 = 1.53$

H_0?

df = 5

**Critical values of the Chi-square distributio
with *d* degrees of freedom**

	Probability of exceeding the critical value					
d	0.05	0.01	0.001	*d*	0.05	0.01
1	3.841	6.635	10.828	11	19.675	24.725
2	5.991	9.210	13.816	12	21.026	26.217
3	7.815	11.345	16.266	13	22.362	27.688
4	9.488	13.277	18.467	14	23.685	29.141
5	11.070	15.086	20.515	15	24.996	30.578
6	12.592	16.812	22.458	16	26.296	32.000
7	14.067	18.475	24.322	17	27.587	33.409
8	15.507	20.090	26.125	18	28.869	34.805
9	16.919	21.666	27.877	19	30.144	36.191
10	18.307	23.209	29.588	20	31.410	37.566

Degrees of Freedom (*df*)

- number of independent pieces of information that go into the estimate of a parameter
- df depends on
  - particular calculation you will be performing
  - what you already know before making calculation

https://www.youtube.com/watch?v=rATNoxKg1yA

https://www.youtube.com/watch?v=rATNoxKg1yA&ab_channel=JamesGilbert

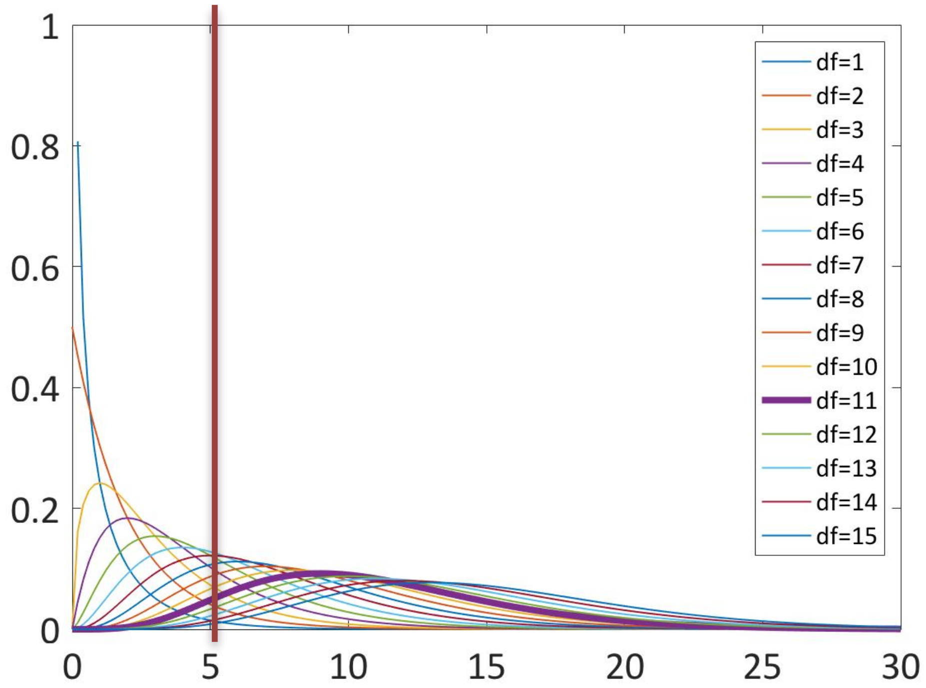
H_A: artists typically tend to be Aries or Cancer

H_0:

Category	Observed
Aries	29
Taurus	24
Gemini	22
Cancer	19
Leo	21
Virgo	18
Libra	19
Scorpio	20
Sagittarius	23
Capricorn	18
Aquarius	20
Pisces	23

256 artists

df = ?



Chi-square Distribution Table

d.f.	.995	.99	.975	.95	.9	.1	.05	.025
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12

zodiac signs are evenly distributed across artists

H_0

Chi-Square test for independence

test ***relationship*** between two separate variables

H_{01} = there is no relationship between extraversion and comfort level of dancing in public

test ***difference*** between two conditions

H_{02} = there is no difference in comfort level of dancing in public between introverts and extraverts

Chi-Square (example)

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

	Extraverts	Introverts	TOTAL
Not comfortable	10	40	50
comfortable	40	10	50
TOTAL	50	50	100

observed frequencies of Introverts and Extraverts who say they would or would not feel comfortable dancing in public

Chi-Square (example)

	Extraverts	Introverts	TOTAL
Not comfortable	10	40	50
comfortable	40	10	50
TOTAL	50	50	100

expected frequencies if the null hypothesis were true?

Chi-Square (example)

$$\frac{\text{row total} \times \text{column total}}{\text{total } n \text{ for table}}$$

	Extraverts	Introverts	TOTAL
Not comfortable	25 10	25 40	50
comfortable	25 40	25 10	50
TOTAL	50	50	100

observed and **expected** frequencies Introverts and Extraverts who say they would or would not feel comfortable dancing in public

Chi-Square (example)

	Extroverts	Introverts	TOTAL
Comfortable	25 10	25 40	50
Not Comfortable	25 40	25 10	50
TOTAL	50	50	100

$$\chi^2 = \sum \frac{(\text{observed - expected})^2}{\text{expected}}$$

$$\frac{(40-25)^2}{25} + \frac{(10-25)^2}{25} + \frac{(10-25)^2}{25} + \frac{(40-25)^2}{25}$$

$$9 \quad + \quad 9 \quad + \quad 9 \quad + \quad 9$$

$$\chi^2 = 36$$

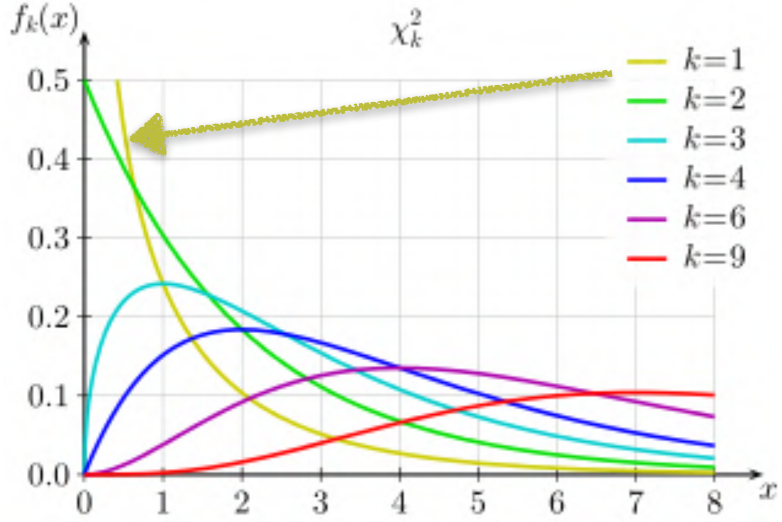
df=?

Chi-Square (example)

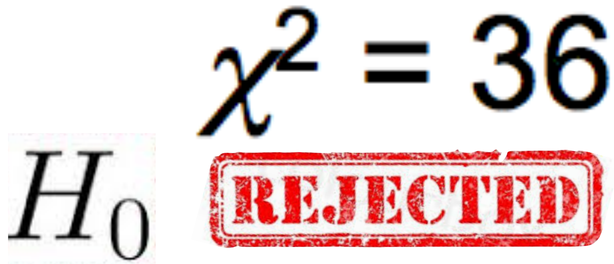
- Degrees of freedom for independence
  - number of cells you would need to calculate all other cell values, assuming we know marginal values

$$df = (R-1)(C-1)$$

$$df = (2-1)(2-1) = 1$$

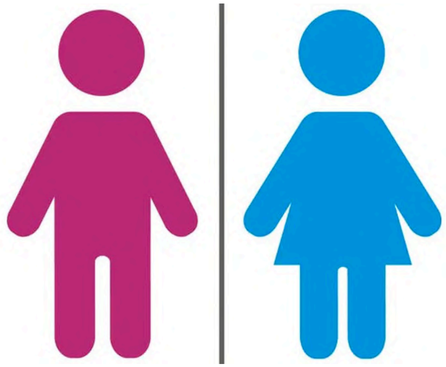


- Our chi-square is significant
  - Introverts tend to feel more comfortable dancing in public compared to Extraverts (surprise!)



EXAMPLE

Gender Study



MENTAL HEALTH FIRST AID®

H_0: There is no relationship between gender and willingness to use metal health services

H_0: The distribution of reported willingness to use mental health services has the same proportions for males and females

H_A: The distribution of reported willingness to use mental health services for males has proportions that are different from those in females

Contingency Table

$$\frac{\textbf{row total} \times \textbf{column total}}{\textbf{total } \boldsymbol{n} \textbf{ for table}}$$

Willingness to Use Mental Health Services (n=150)

	No	Maybe	Yes	**Total**
Males	17 **12**	32 **30**	11 **18**	60
Female	13 **18**	43 **45**	34 **27**	90
Total	30	75	45	150

$df = (R-1)(C-1)$ **df = 2**

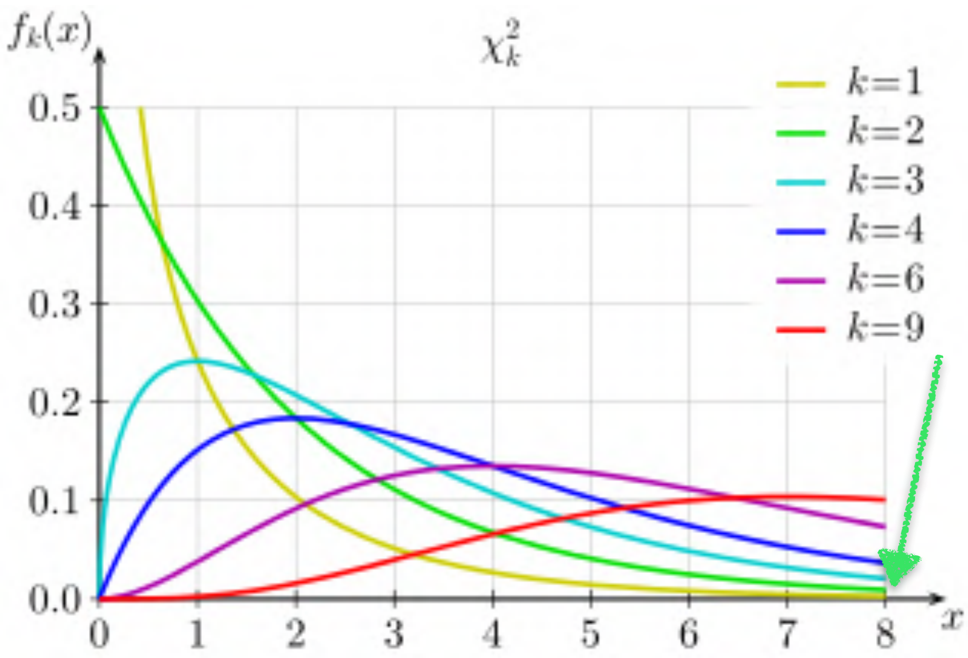
EXAMPLE

Gender Study

Willingness to Use Mental Health Services

$$\chi^2 = 8.23$$

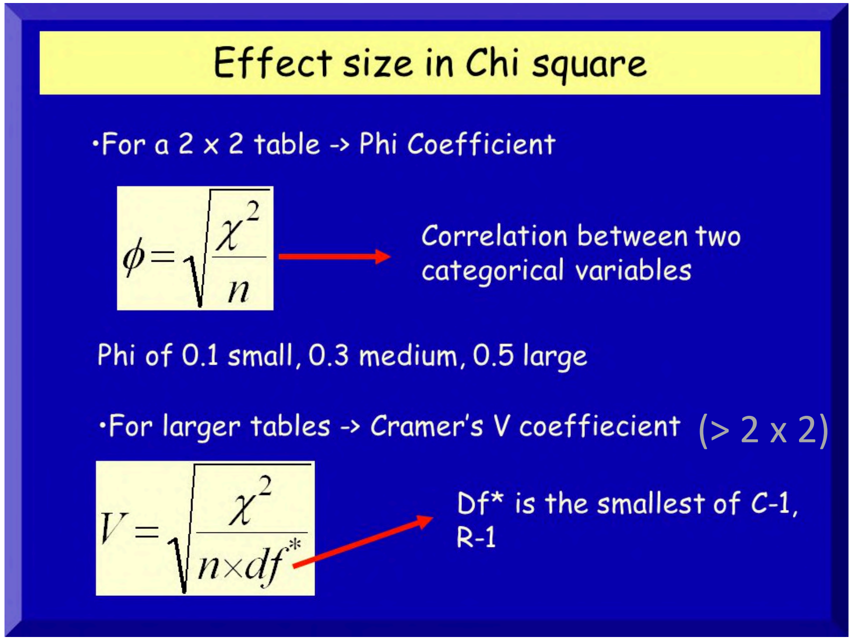
df = 2



H_0 REJECTED

Males are less willing to use Mental Health Services

Effect Size



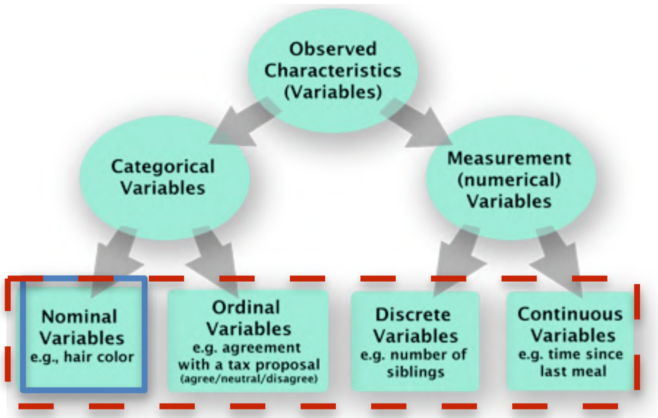
Phi = sqrt(8.23/150)
= 0.23

Results showed a significant difference between males' and females' attitude toward using metal health services,

χ^2 (2, *n = 150*) = 8.23, *p* < .05, *V* = 0.23

Chi-Square Test and Correlation

Participant	Self-Esteem X	Academic Performance Y
A	13	73
B	19	88
C	10	71
D	22	96
E	20	90
F	15	82
.	.	.
.	.	.
.	.	.



		Level of Self-Esteem			
		High	Medium	Low	
Academic Performance	High	17	32	11	60
	Low	13	43	34	90
		30	75	45	$n = 150$

Chi-square and independent measures *t* and ANOVA

Participant	Self-Esteem X	Academic Performance Y
A	13	73
B	19	88
C	10	71
D	22	96
E	20	90
F	15	82
.	.	.
.	.	.
.	.	.

		Level of Self-Esteem			
		High	Medium	Low	
Academic Performance	High	17	32	11	60
	Low	13	43	34	90
		30	75	45	$n = 150$

Median Test for Independent Samples

- non-parametric alternative to independent measures *t*-test (or ANOVA) to determine significant group differences
- H_0 = different samples come from population that share a common median

Self-Esteem Scores for Children at Three Levels of Academic Performance

High		Medium				Low	
22	14	22	13	24	20	11	19
19	18	18	22	10	16	13	15
12	21	19	15	14	19	20	16
20	18	11	18	11	10	10	18
23	20	12	19	15	12	15	11

Median Test for Independent Samples

- calculate median for combined group (n = 40)
- within each group, perform median (17) split and fill contingency table

Self-Esteem Scores for Children at Three Levels of Academic Performance

High		Medium				Low	
22	14	22	13	24	20	11	19
19	18	18	22	10	16	13	15
12	21	19	15	14	19	20	16
20	18	11	18	11	10	10	18
23	20	12	19	15	12	15	11

	Academic Performance		
	High	Medium	Low
Above Median	8	9	3
Below Median	2	11	7

Median Test for Independent Samples

	Academic Performance		
	High	Medium	Low
Above Median	8 **5**	9 **10**	3 **5**
Below Median	2 **5**	11 **10**	7 **5**

$\chi^2 = 5.4$ **df = 2** $\chi^2 = 5.99$ **(p < .05)**

—> not sufficient evidence to conclude that there are significant differences among the self-esteem for these three groups of students

Chi-Square test

Limitations

- Observations must be unique to one cell (Between subjects)
  - each person must fall into only one cell
  - not valid for within subject designs (repeated measures/ matched pairs)
- Only frequencies can be studied, not means, percentages, ratios, etc.
- Low **expected** frequencies cause problems (should be ≥ 5)
  - loss of statistical power
- No group should contain less than 10 (or 5) (try to regroup instead)
- Not apt for low sample size.
- Informs of presence or absence (probability of occurrence) of association but doesn't measure strength of association

Can J Psychiatry. 1998 Oct;43(8):837-42.

Life after chi-squared: an introduction to log-linear analysis.

Streiner DL[1], Lin E.

Author information

1 Kunin-Lunenfeld Applied Research Unit, Baycrest Centre for Geriatric Care, North York, Ontario. dstreiner@rotman-baycrest.on.ca

Abstract

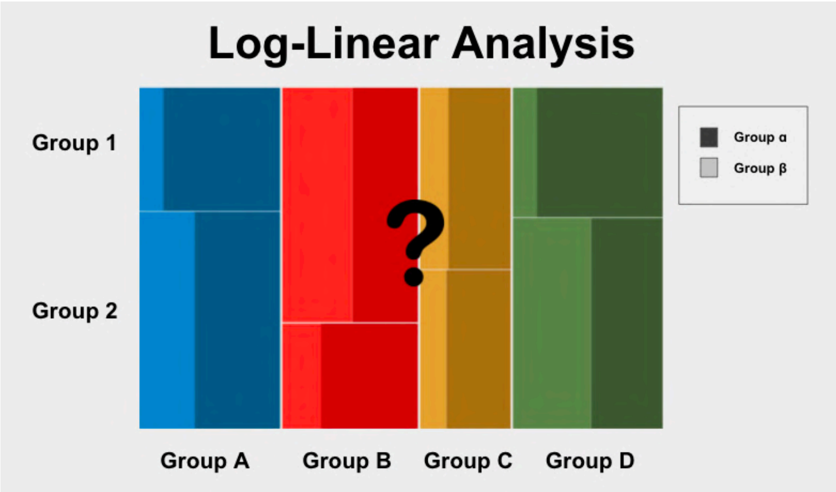
Chi-squared tests are used to examine the relationships among categorical variables. However, they are difficult to use and interpret when more than 2 variables are involved. In such cases, it is better to use a related statistic, called log-linear analysis. This article is an introduction to log-linear models, illustrating how they can be used to tease apart relationships among several variables in looking at the factors associated with photonumerophobia.

	Age category of car		
	New	Old	Total
Male drivers			
Behaviour at amber light			
Stopped	79	63	142
Did not stop	87	95	182
Total	166	158	324
Female drivers			
Behaviour at amber light			
Stopped	95	83	178
Did not stop	51	94	145
Total	146	177	323
Total old/new cars:	312	335	647

Table 18.15 Stopping behaviour of male and female drivers in old and new cars.

Log-Linear Analysis

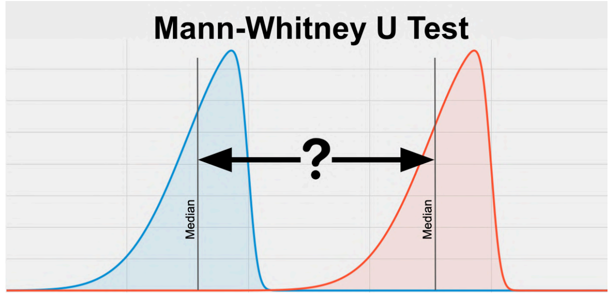
- variable of interest is proportional or categorical
- have two or more options
- no assumptions of IV or DV
- used for both hypothesis testing and model building



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	*The binomial sign test*
Non-parametric (for ordinal data)	*Mann-Whitney U*	*Wilcoxon Signed-Rank test* *The binomial sign test*
Parametric	*Unrelated t-test (level of data: interval)*	*Related t-test (level of data: interval)*

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		

critical U value =12

$\alpha < .05$

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

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

Kruskal-Wallis Test

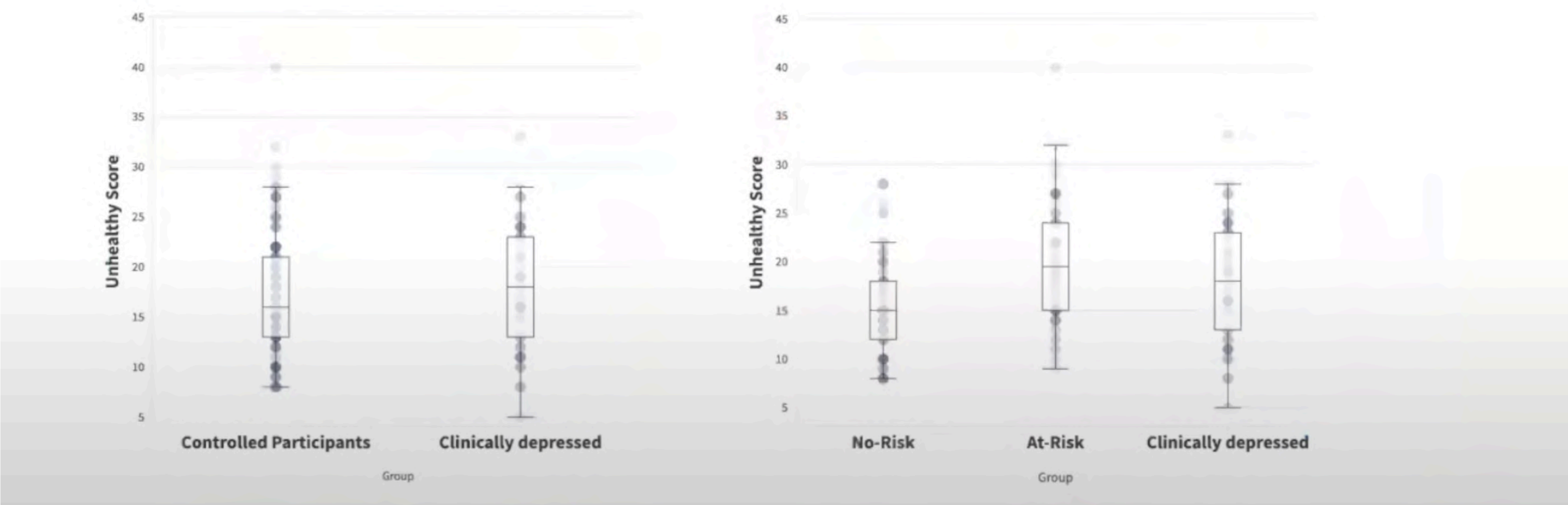
Aim

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	*The binomial sign test*
Non-parametric (for ordinal data)	*Mann-Whitney U*	*Wilcoxon Signed-Rank Test*
Parametric	*Unrelated t-test (level of data: interval)*	*Related t-test (level of data: interval)*

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

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

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

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

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

n	Two-Tailed Test		One-Tailed Test	
	α = .05	α = .01	α = .05	α = .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 ACCEPTED

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	*The binomial sign test*
Non-parametric (for ordinal data)	*Mann-Whitney U*	*Wilcoxon Signed-Rank Test* *The binomial sign test*
Parametric	*Unrelated t-test (level of data: interval)*	*Related t-test (level of data: interval)*

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	+
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.

$S = 1$

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

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	+
f	1	5	4	+
g	15	16	1	+
h	10	12	2	+
i	11	15	4	+
j	10	17	7	+

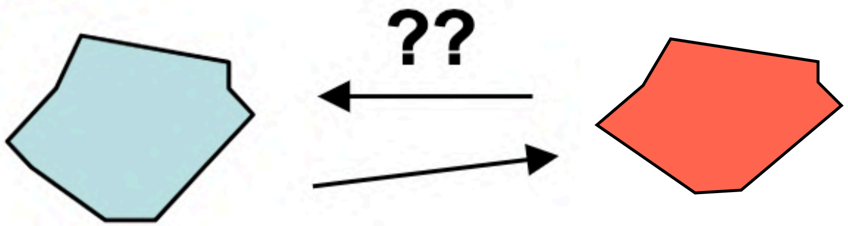
$S = 1$

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

critical S value =1

$\alpha \leq .05$

which tests can i use?



unrelated / between

---

analytic skills: **CSE vs ECE**

Brain connectivity patterns **musicians vs non-musicians**

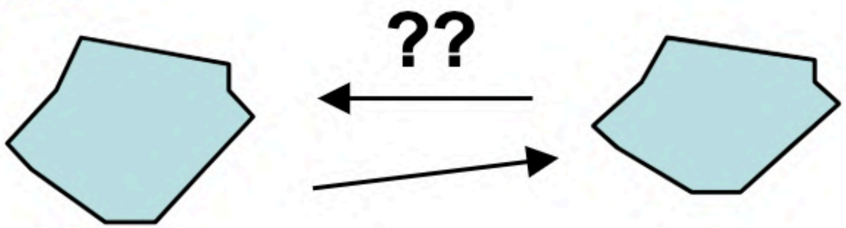
Gender differences in social media usage

EXAMPLE

Performance in **Quiz 1 vs Quiz 2**

Memory **Pre- vs Post**- Sleep depravation

Pollution level **before vs after** Diwali



related / within conditions

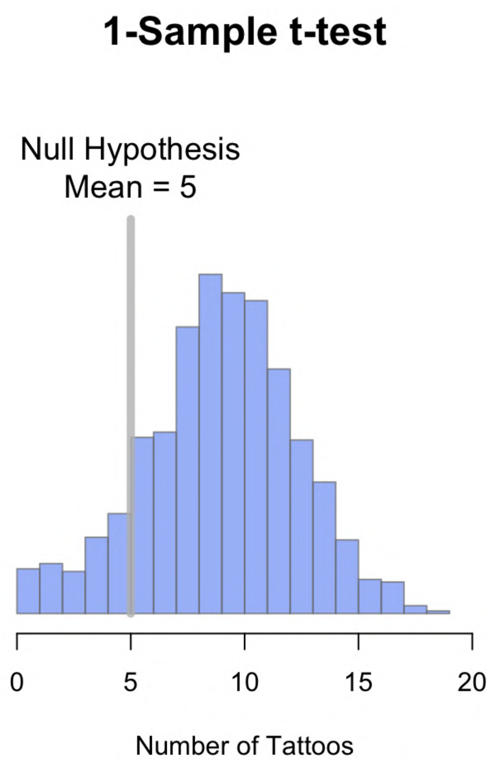
Different types of tests

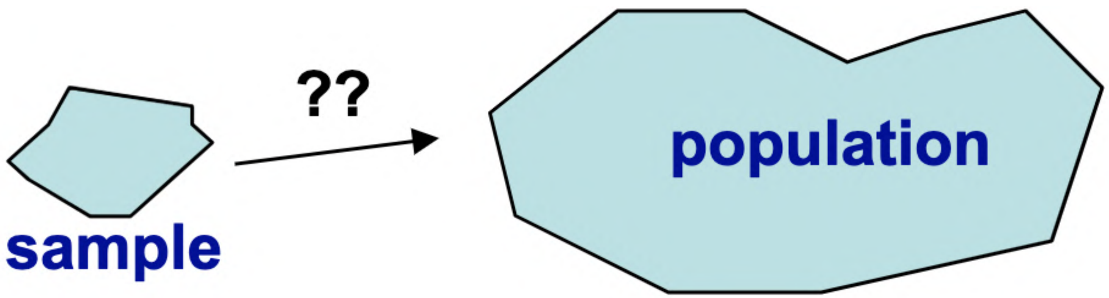
Test type	Between subjects designs (Independent samples)	Within subject designs (repeated measures/ matched pairs)
Non-parametric (for categorical data)	Chi-square	*The binomial sign test*
Non-parametric (for ordinal data)	*Mann-Whitney U*	*Wilcoxon Signed-Rank Test*
Parametric	*Unrelated t-test (level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

Single Sample t-test

Does the observed distribution come from a population with a certain mean?

How certain are we that it comes from a different population?



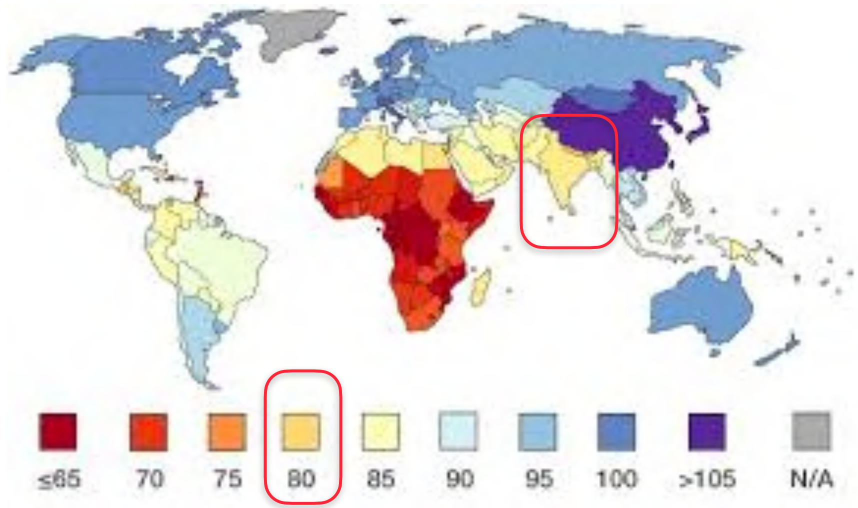


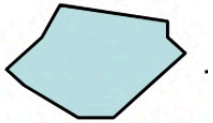
EXAMPLE

Single Sample t-test

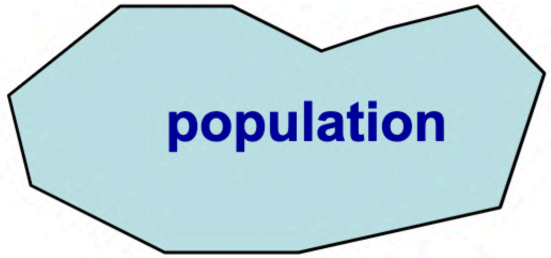
$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{N}}}$$

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD





sample



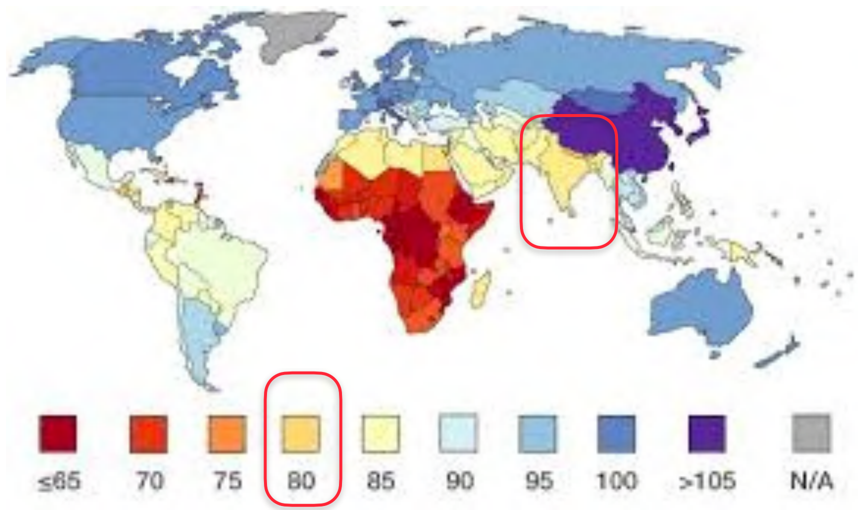
H_0:

H_A:

EXAMPLE

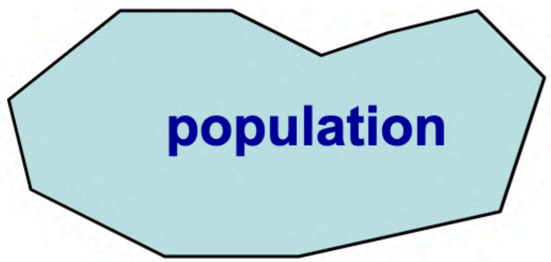
Single Sample t-test

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD



$N = 500$

$\bar{X} = 83$



$\mu = 80$

H_0

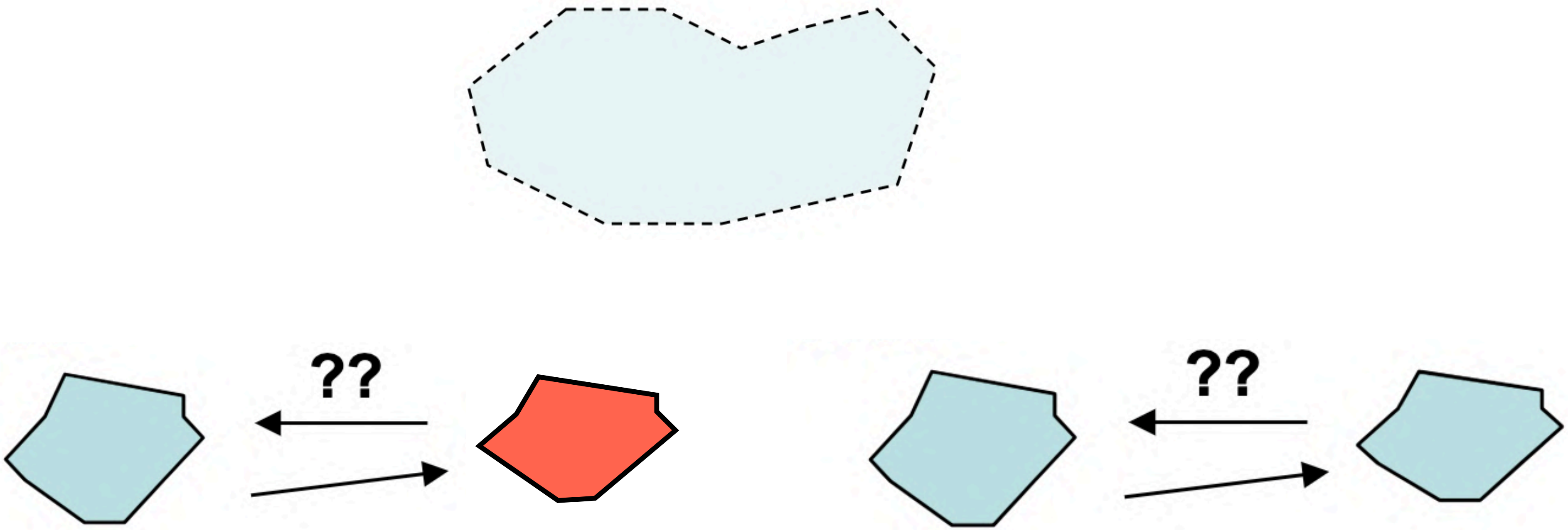
REJECTED

$t = 14.9$ $(df = 499)$, $p < .001$

Two Sample test

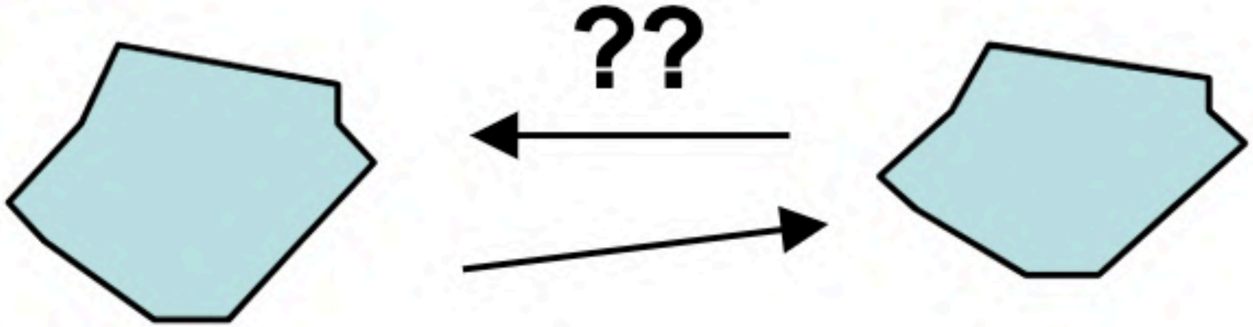
Do they come from the same population?

How certain are we that they are different?



Two Sample related *t*-test

H_A: Students perform better/worse in mock tests



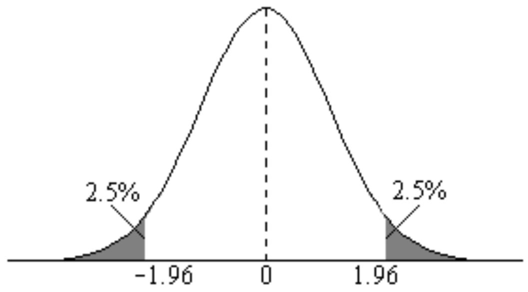
EXAMPLE

dir or **non-dir H_A?**

Two Sample related *t*-test

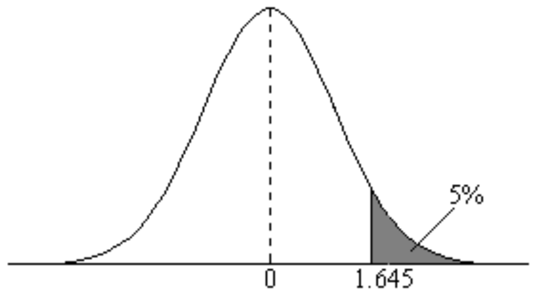
Student	Mock	Real
1	316	320
2	324	319
3	317	318
4	323	314
5	333	330
6	329	321
7	328	311
8	319	309
9	320	318
10	314	321

non-dir H_A



(b) Two-tailed test

dir H_A



(a) One-tailed test

Is t-test valid??

> 2 groups/conditions

Test type	Between subjects designs (Independent samples)	Within subjects designs (dependent samples)
Parametric	*One-way ANOVA*	*One-way Repeated measures ANOVA*
Non-parametric	*Kruskal-Wallis one way analysis of variance*	Friedman's two-way analysis of variance

MENU nature International journal of science | Subscribe | Search | Login

COMMENT · 20 MARCH 2019

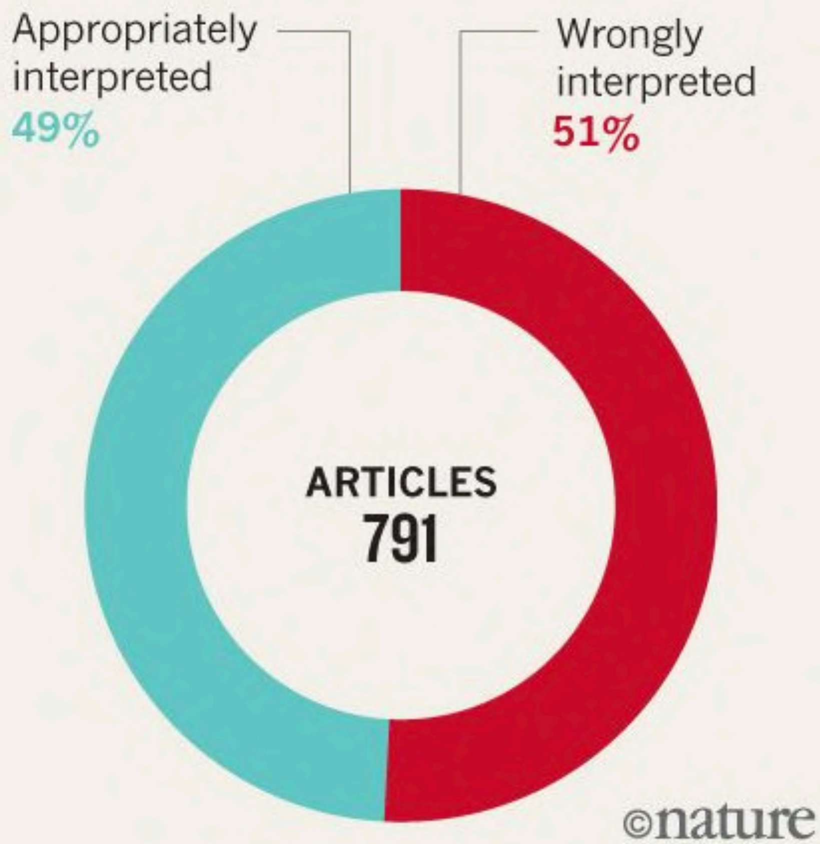
Scientists rise up against statistical significance

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).

MENU nature International journal of science Subscribe Search Login

COMMENT · 20 MARCH 2019

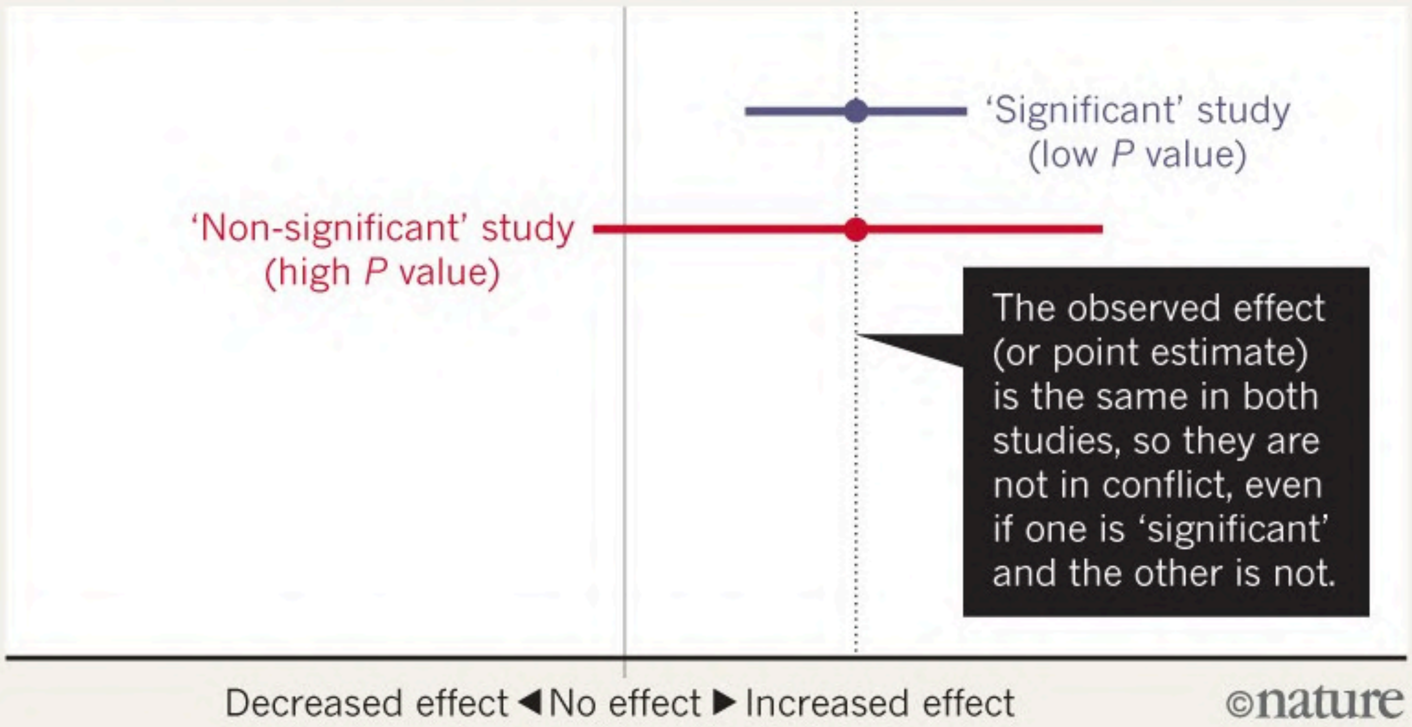
Scientists rise up against statistical significance

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 $p<.05$.