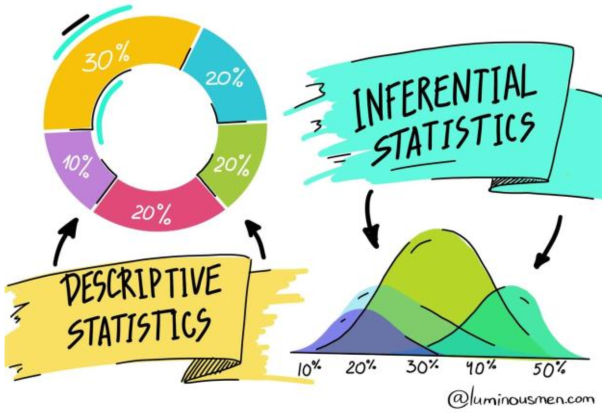
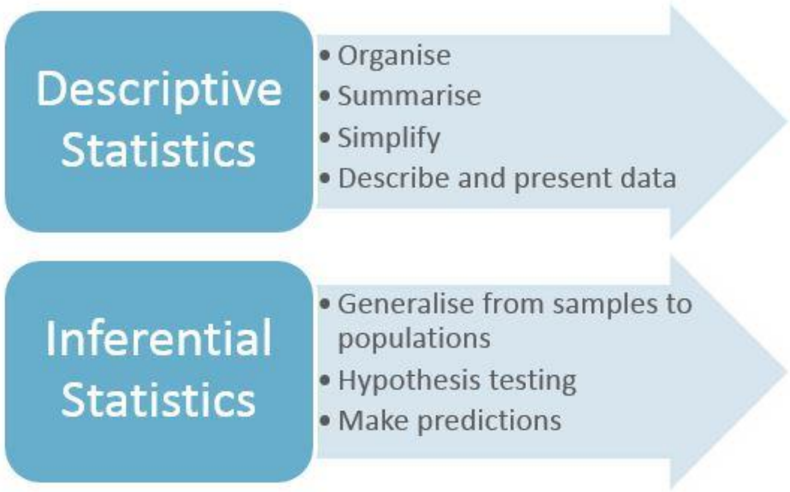
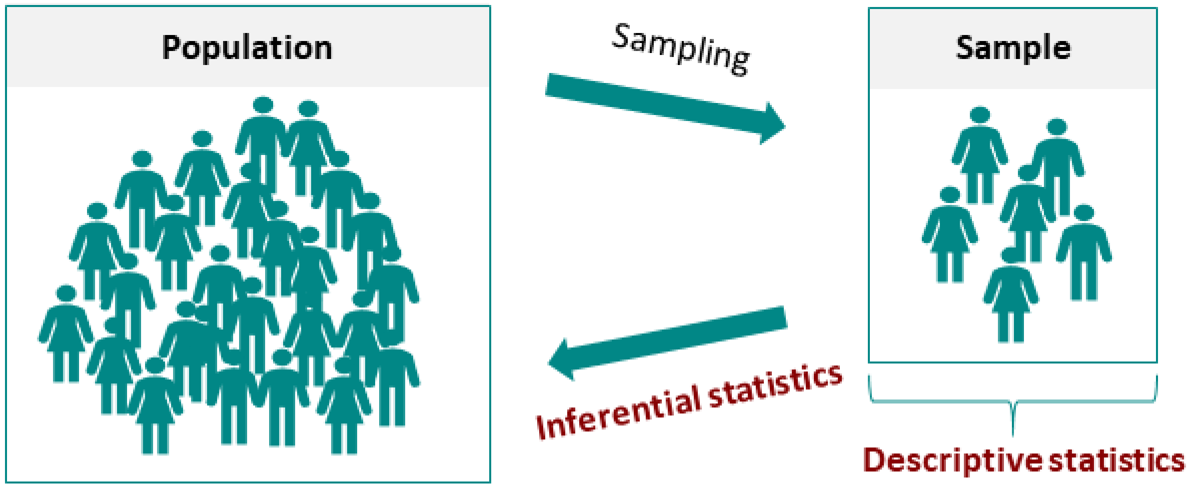
Hypothesis Testing

Why do we need inferential statistics?



Inferential statistics allow us to *infer* or generalize observations made with samples to the larger population from which they were selected.





What is a Hypothesis?

Research Question

(ideas)

A specific testable statement

(that guides an experiment)

What Is a Real Hypothesis?

- A hypothesis is an educated guess, based on observation.
- Usually, a hypothesis can be supported or refuted through experimentation or more observation.
- A hypothesis can be disproven, but not proven to be true.

Research Question – Is online teaching effective?

Hypothesis Statement – Students taught offline perform better than students taught online

(ASSUMPTION) – based on previous studies, observations, experiences, etc.

Null Hypothesis and Alternative Hypothesis

H_0 VS H_1 or H_a

Students taught online vs offline perform equally well on exams

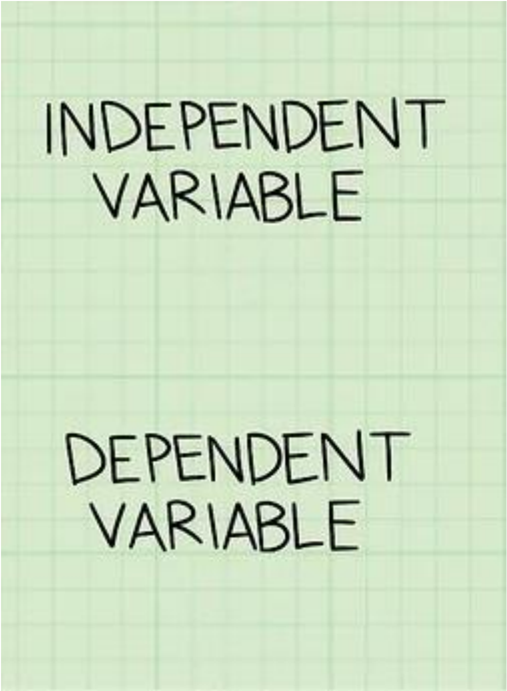
Students taught offline perform better than students taught online

Students taught online perform better than students taught offline

You perform experiments to check if the Ho holds true or not.
By disproving the Ho you accept the HA

Variables in a hypothesis

Hypothesis Statement – Students taught offline perform better than students taught online



not changed by the other variables you are trying to measure

Teaching method → online
→ offline

Value is changed or affected by the independent variable/s

Exam performance

Individuals with more years of education have higher income

H_o – No relationship between years of education and income

H_1 - Individuals with more years of education have higher income

Leopards are stronger than Tigers

H_o – Leopards and Tigers are equally strong, no difference

H_1 – Tigers are stronger than Leopards

H_2 – Leopards are stronger than Tigers

Exercise effects on anxiety

H_0 - Exercise has no effect on anxiety

H_1 - Exercise lowers anxiety

H_2 – Exercise increases anxiety

IV – Exercise (exercising, not exercising)

DV – anxiety levels

Directionality in a hypothesis

This prediction is typically based on past research, accepted theory, extensive experience, or literature on the topic.

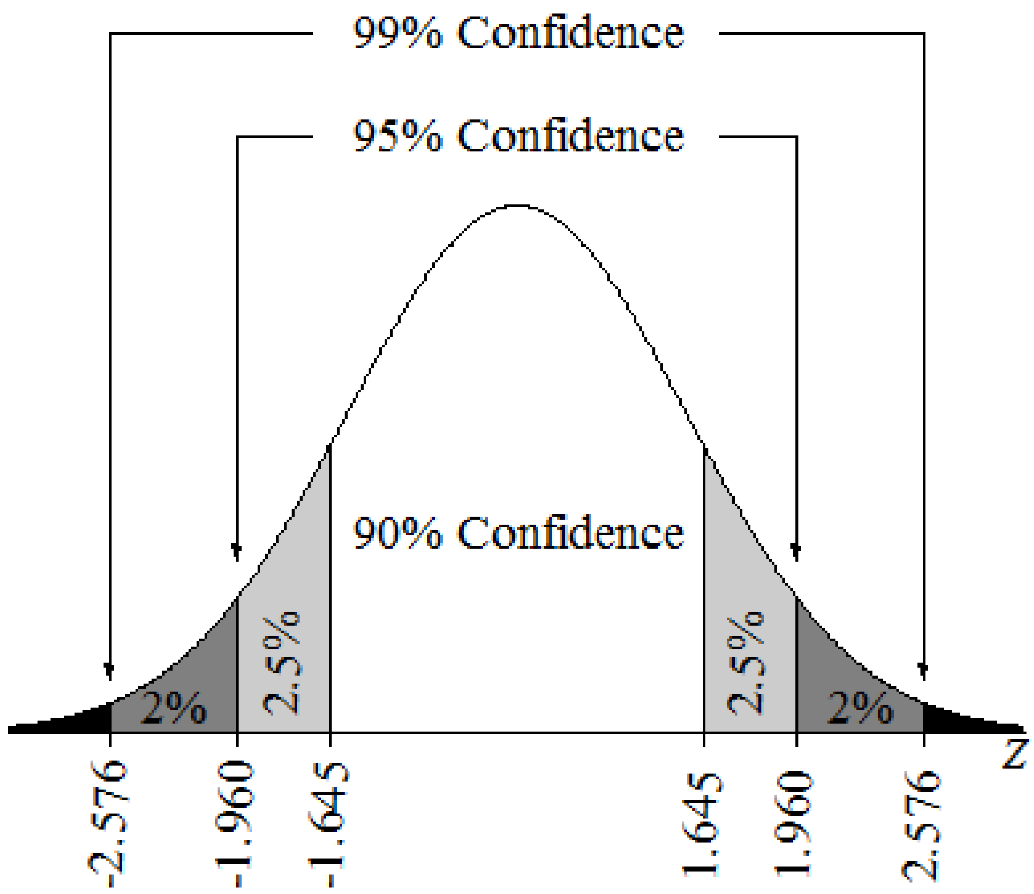
Else your statistical outcome can be misleading, by ignoring other outcomes
e.g. Does a technical degree impart technical skills?

High quality of engineering education leads to higher technical skills

H_o – Quality of engineering education has no effect on technical skills

H_1 - High quality of engineering education leads to higher technical skills

Confidence Intervals



Confidence Level	α (level of significance)	$Z_{\alpha/2}$
99%	1%	2.575
95%	5%	1.96
90%	10%	1.645

$$CI = \bar{x} \pm z\frac{s}{\sqrt{n}}$$

CI = confidence interval

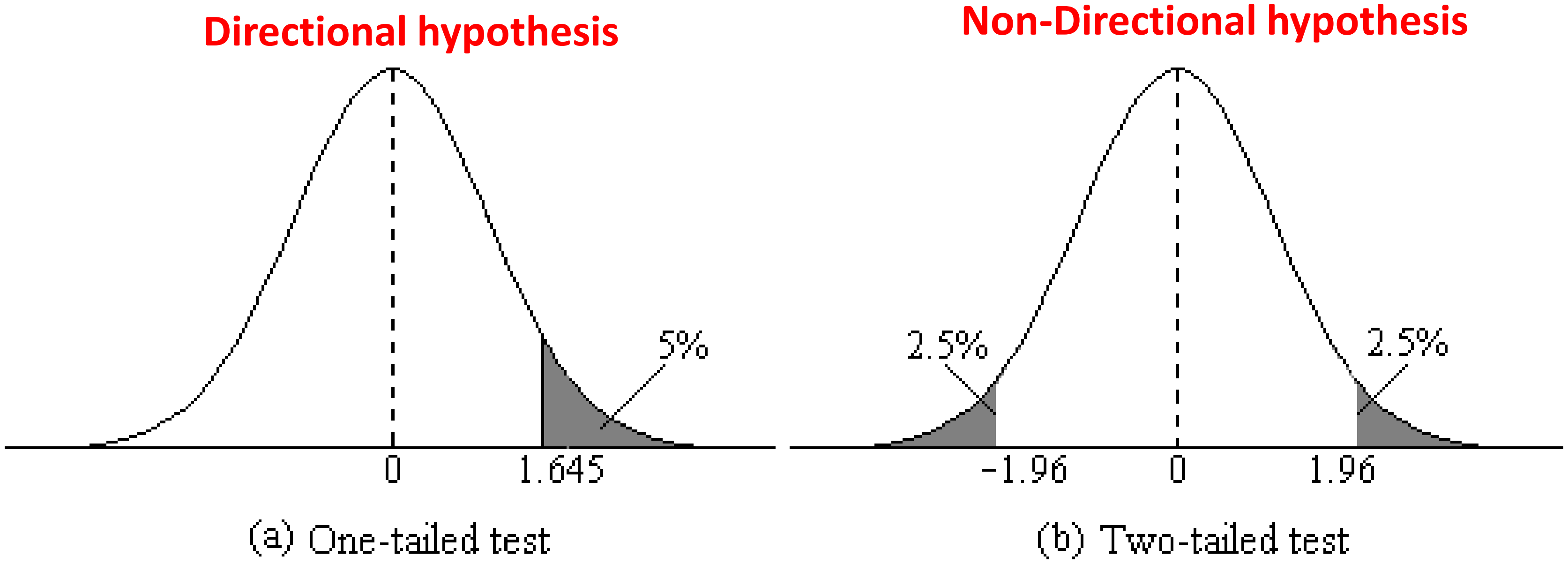
$\bar{x}$ = sample mean

z = confidence level value

s = sample standard deviation

n = sample size

Hypothesis testing



General Rule: Use two-tailed test.

Only if direction is known from prior studies (justified reason), use one-tail test.

Criterion (α) for significance – 5% (0.05) for most behavioural studies (95 % CI)

If *p* > 0.05 → Accept the Ho

If *p* ≤ 0.05 → Reject the Ho & accept H_A

One-tailed vs two-tailed test

When is a one-tailed test NOT appropriate?

- Choosing a one-tailed test for the sole purpose of attaining significance is not appropriate.

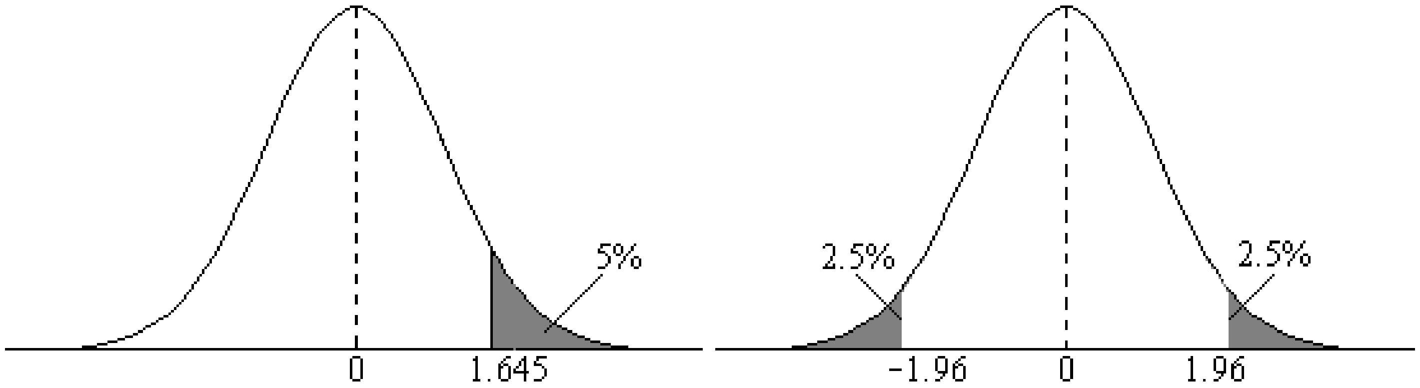
- Choosing a one-tailed test after running a two-tailed test that failed to reject the null hypothesis is not appropriate, no matter how "close" to significant the two-tailed test was.

- Using statistical tests inappropriately can lead to invalid results that are not replicable and highly questionable–a steep price to pay to show significance in your results

Exercise effects on anxiety

H_0 - Exercise has no effect on anxiety
H_1 - Exercise lowers anxiety

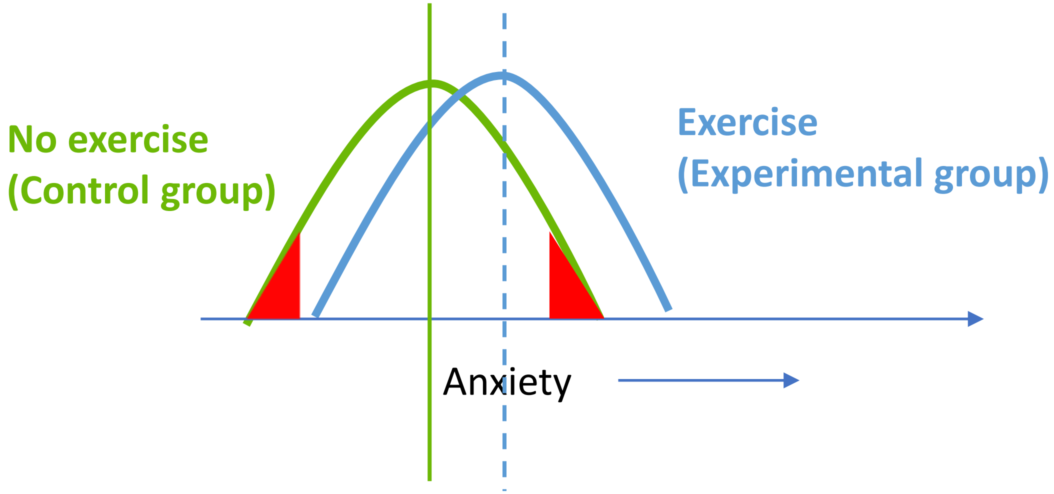
H_2 – Exercise increases anxiety?



(a) One-tailed test (b) Two-tailed test

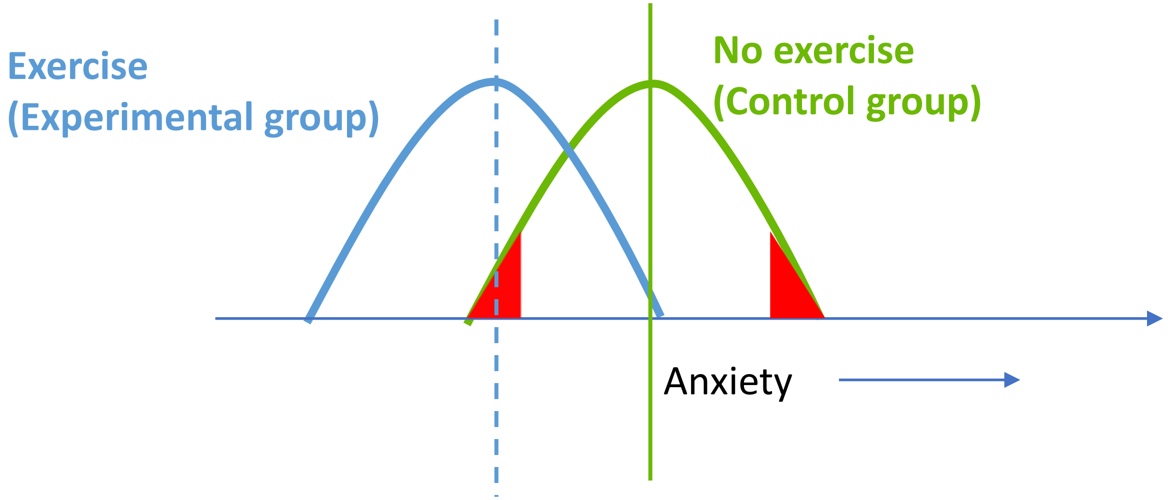
When $p > .05$, we retain the null hypothesis – there is less difference between the groups.

H0 - Exercise has no effect on anxiety

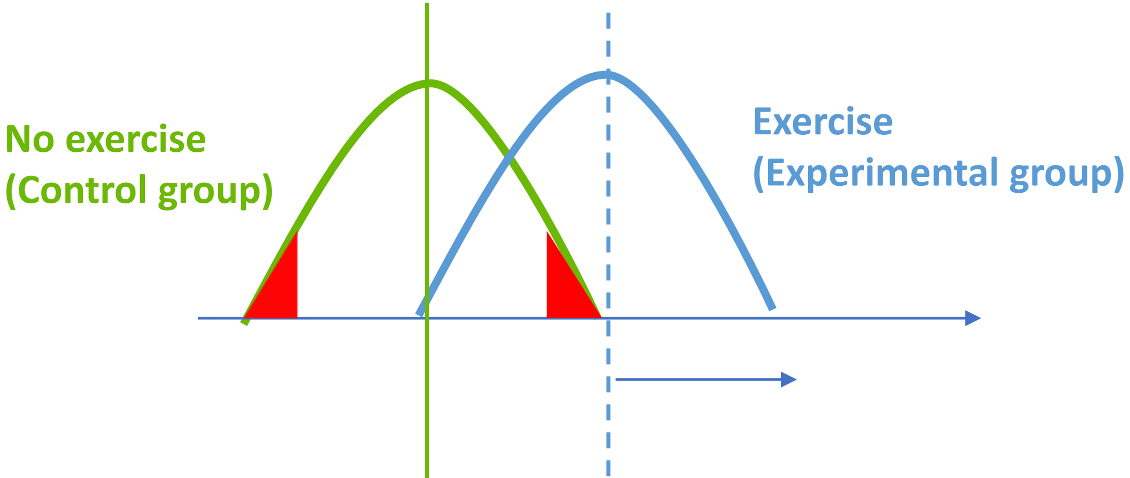


When $p \leq .05$, we reject the null hypothesis - there is a **‘significant’** difference between the two groups.

H1 - Exercise lowers anxiety



H2 - Exercise increases anxiety



Another Directional Hypothesis

You have a new drug to treat pain that is cheaper than the existing drug and you only want to confirm if the new drug is less effective than the existing drug

Whether the new drug is similar or better than the existing drug does not matter.

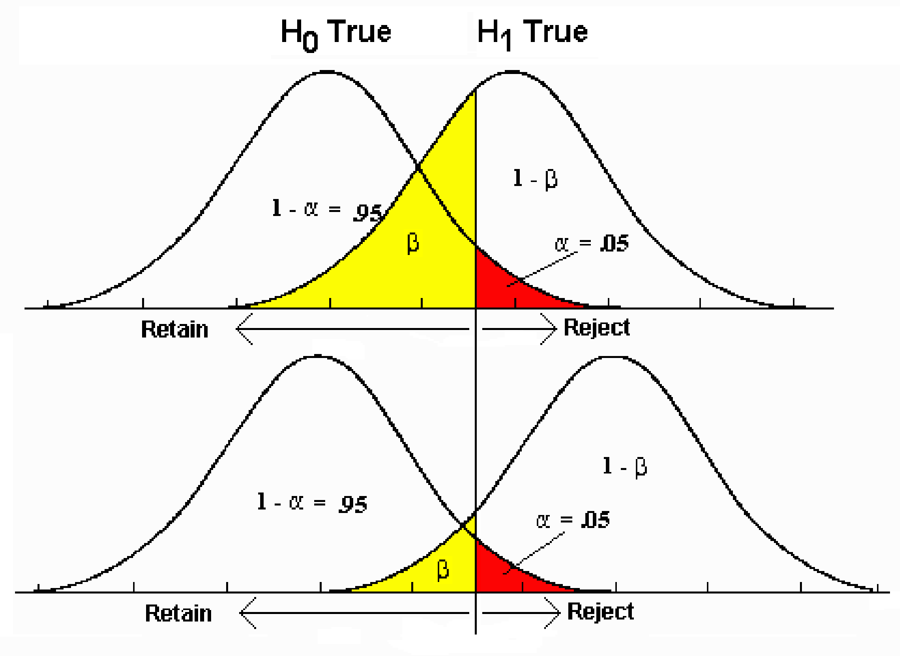
H_0 - Null hypothesis – No difference between new drug and existing drug to treat pain

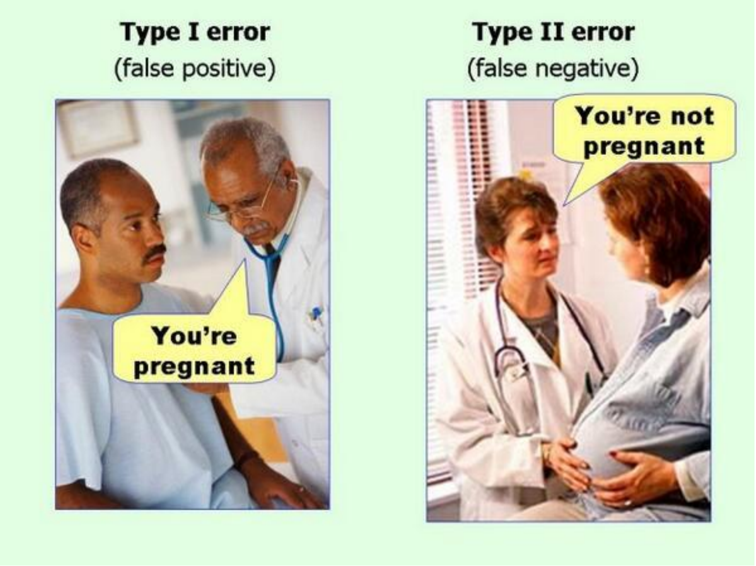
H_1 - Alternate hypothesis – Is the new drug less effective than the existing drug

- Ho, HA, IV, DV, one or two tailed test?

- Smoking is injurious to the lungs
- Videogaming can lower attention span
- Does repetition in advertising improve sales?
- Air pollution is more fatal than COVID19
- Is there a difference in leadership style between men and women?

Types of Errors in hypothesis testing





Reality:

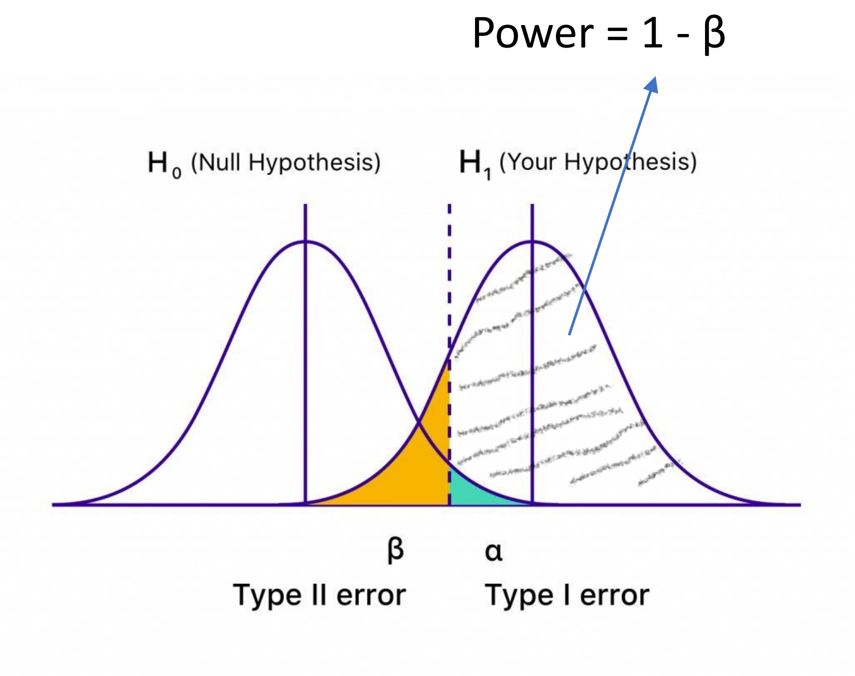
Decision from statistical tests	Ho False	Ho True	
Reject Ho	**Correct Decision** Sensitivity/Power $1- \beta$	Type 1 Error "False Positive" α	Observe difference when none exists *Overreacting!*
Accept Ho	Type 2 Error "False Negative" β	**Correct Decision** Specificity $1-\alpha$	

Fail to find a difference when there is one

Underreacting!

- Sample size it too small (high variability)
- Choosing one-tailed instead of two-tailed test
- Wrong statistical test

Power



- Power - the probability that your test will find a statistically significant difference when such a difference actually exists.
- In other words, power is the probability that you will reject the null hypothesis when you should (and thus avoid a Type II error).
- It is generally accepted that power should be .8 or greater; that is, you should have an 80% or greater chance of finding a statistically significant difference when there is one.

https://meera.snre.umich.edu/

Power

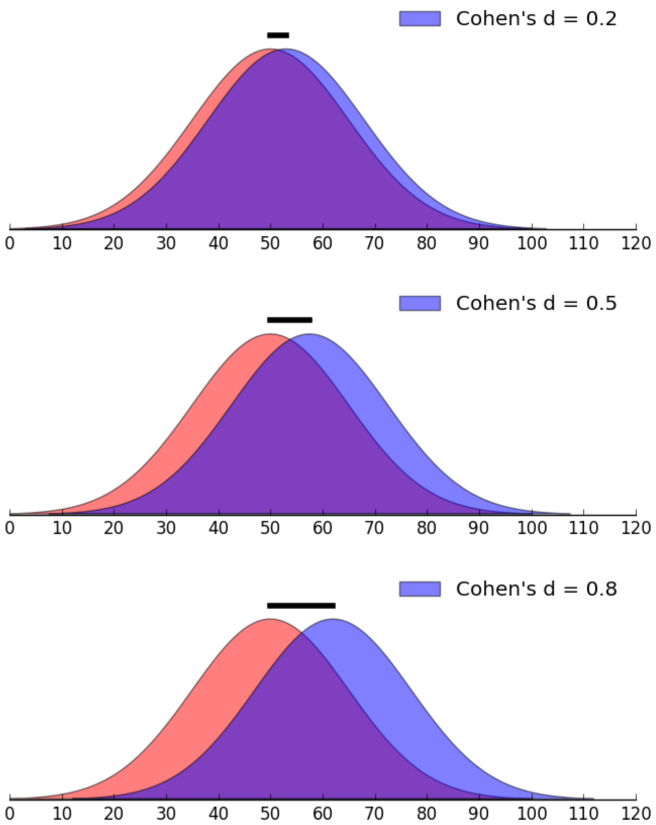
Power is calculated using statistical software. You need to know –

- What type of test you plan to use (e.g., independent t-test, paired t-test, ANOVA, correlation, regression, etc.)
- The alpha value or significance level you are using (usually 0.05 or 0.01)
- The expected effect size
- The sample size you are planning to use

As your sample size increases, so does the power of your test.

- larger sample means that you have collected more information -- which makes it easier to correctly reject the null hypothesis when you should.
- A power value is between 0 and 1.
- If the power is less than 0.8, you typically need to increase your sample size.

Effect Size



E.g. you evaluate the effect of a group discussion on student knowledge using pre and post tests on 500 students. The mean score on the pre test was 83 out of 100 while the mean score on the post test was 84.

- What if you simply found a statistical difference by virtue of a large sample size (> 1000 or 10000)?

- If you calculate the effect size – you get a standard method to defining the importance of the statistical difference

Cohen's d effect size interpretation

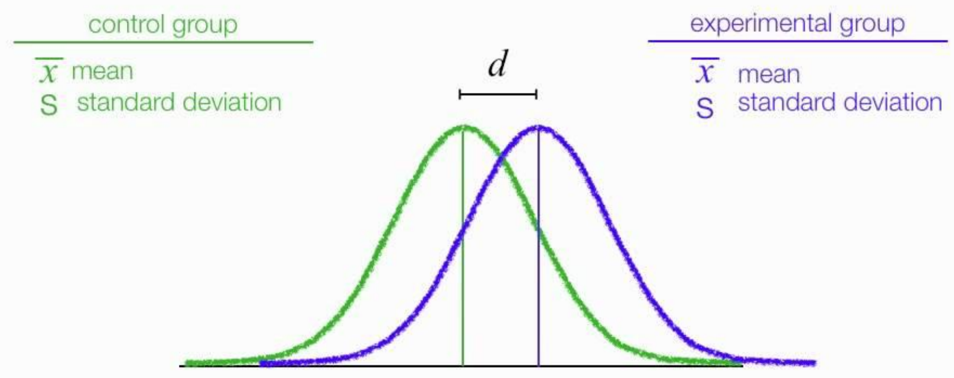
< 0.1 = trivial effect
0.1 - 0.3 = small effect
0.3 - 0.5 = moderate effect
> 0.5 = large difference effect

Effect Size

- Effect size is a quantitative measure of the *strength of a phenomenon*.
- Effect size emphasizes the **size** of the difference or relationship
- Examples:
  - the correlation between two variables (specifically r^2)
    - r=.1 weak, r=.5 moderate, r=.7 strong, r=.9 very strong
  - the regression coefficient in a regression (B_0, B_1, B_2)
    - Relative to model and field
  - the mean differences in t tests (use Cohen's D)
    - d = .2 is small; r = .5 is medium; r = .8 is large
  - The mean differences in ANOVA (use eta)
    - .01 is small, .06 medium, .14 large

$$\text{Cohen's Effect size} = \frac{(\text{Mean}_{\text{treatment}} - \text{Mean}_{\text{control}})}{\text{Standard deviation}_{\text{pooled}}}$$

$$d = \frac{\bar{x} - \bar{x}}{S}$$



Basic formula for sample size - Continuous data

$$\text{Number of samples per group (n)} = \frac{2\,x\,(Z_{(1-\alpha/2)}+Z_{\beta})^2\,x\,\sigma^2}{\Delta^2}$$

Where Δ = size of difference, minimal effect of interest

α = significance level (eg 0.05)

β = power, probability of detecting a significant result (typically 80%, 90%)

σ = SD of data

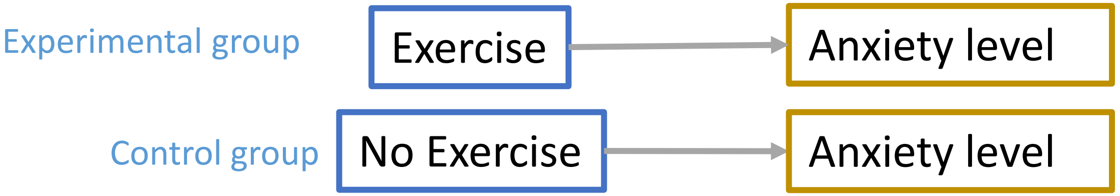
Z_p = points on normal distribution to give required power and significance

DV: Anxiety level

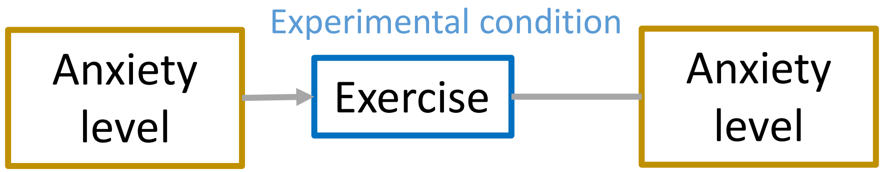
IV: Exercise

Do people who exercise have lower levels of anxiety?

Does exercise lower anxiety?

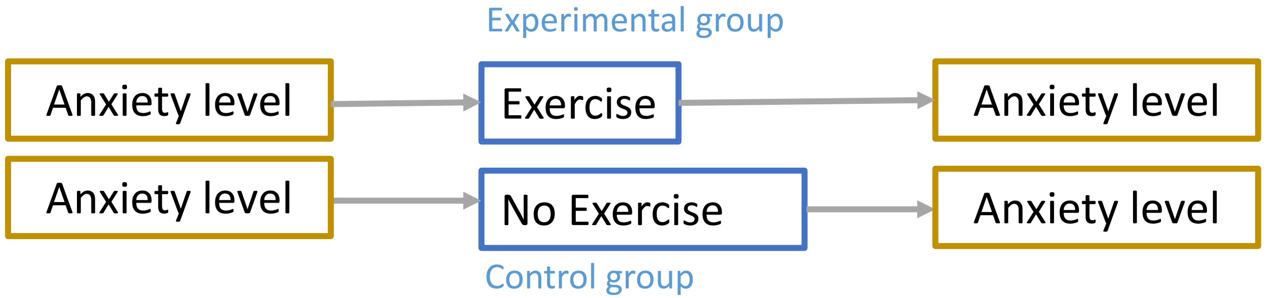


Between groups
(this does not allow you to measure change)



Within group/Repeated measures
(crossover design)

- Participant fatigue
- Longer experimental duration
- Carry over effects

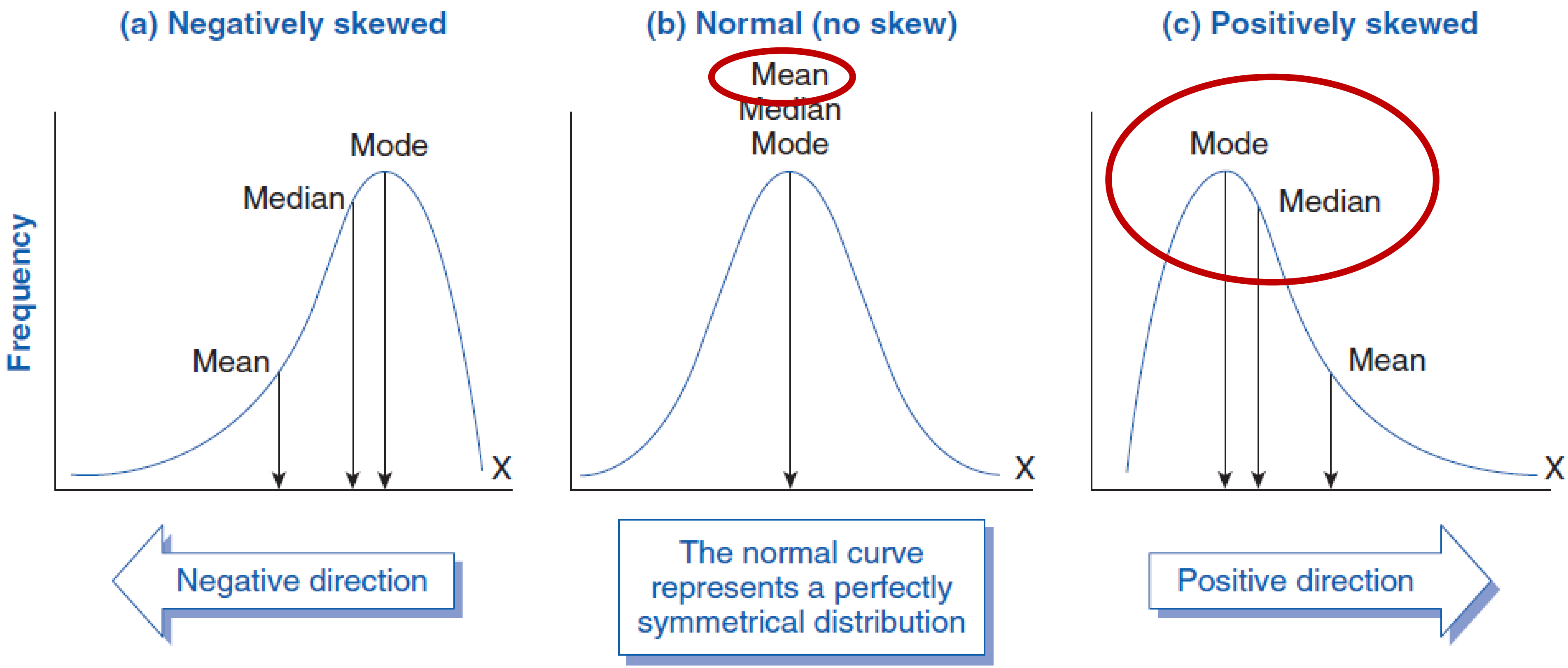


Mixed design
Between groups & Within groups

Anxiety levels

	Exercise	No -Exerci
	20	24
	23	35
	25	41
	30	21
	35	38
	29	23
	37	37
	24	44
	29	32
	31	33
	26	34
	28	42
Mean	28.08333	33.66667
SD	4.680782	7.261007

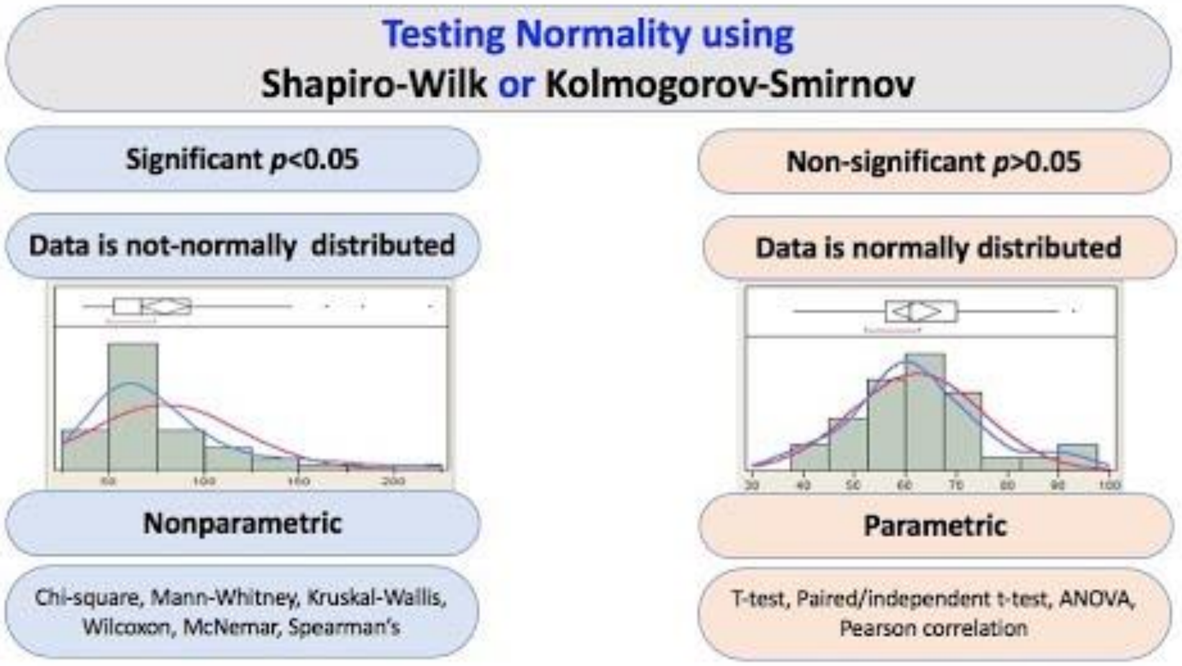
Normality?



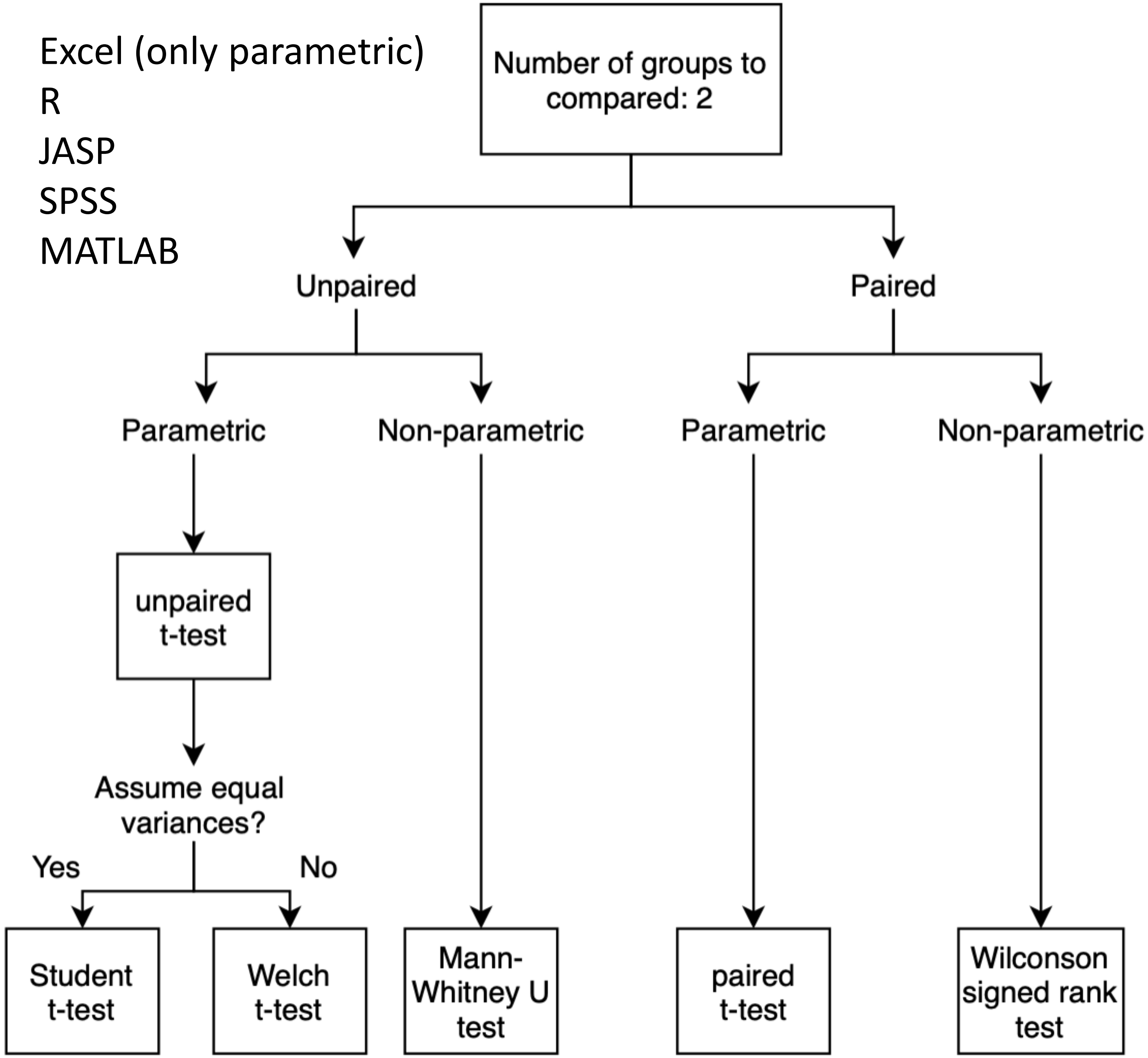
Kolmogorov–Smirnov test (n>=50)
OR
Shapiro–Wilk test (n<50)

The null hypothesis for normality → data is normally distributed

Parametric vs non-parametric



Excel (only parametric)
R
JASP
SPSS
MATLAB



Independent Sample t test

T-Test Example

People who exercise have lower levels of anxiety

Anxiety levels

	Exercise	No -Exerci
	20	24
	23	35
	25	41
	30	21
	35	38
	29	23
	37	37
	24	44
	29	32
	31	33
	26	34
	28	42
Mean	28.08333	33.66667
SD	4.680782	7.261007

	Exercise	*No -Exercise*
Mean	28.08333333	33.66666667
Variance	23.90151515	57.51515152
Observations	12	12
Pooled Variance	40.70833333	
Hypothesized Mean Diff	0	
df	22	
t Stat	-2.143519905	
P(T<=t) one-tail	0.021690748	
t Critical one-tail	1.717144374	
P(T<=t) two-tail	0.043381495	
t Critical two-tail	$\boxed{2.073873068}$	

t-value

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$\bar{x}_1$: Mean value of the first group
$\bar{x}_2$: Mean value of the second group
n_1 : Size of the first group
n_2 : Size of the second group
s_1 : Standard deviation of the first group
s_2 : Standard deviation of the second group

$$\text{Cohen's Effect size} = \frac{(\text{Mean}_{\text{treatment}} - \text{Mean}_{\text{control}})}{\text{Standard deviation}_{\text{pooled}}}$$

Cohen's d = (33.66 - 28.083) / 6.107782 = **0.913097**

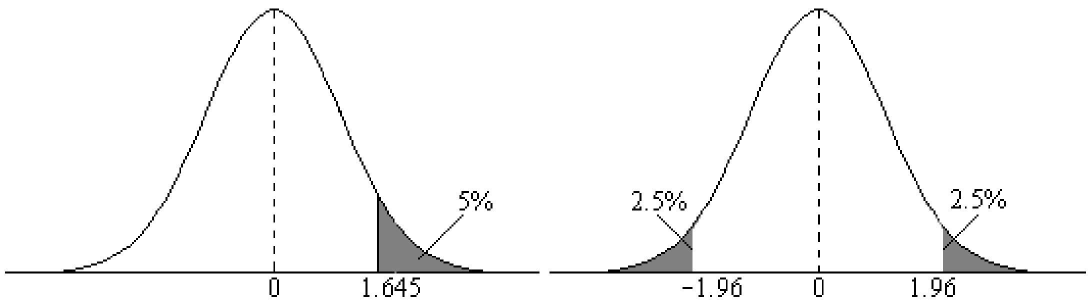
Cohen's d effect size interpretation

< 0.1 = trivial effect
0.1 - 0.3 = small effect
0.3 - 0.5 = moderate effect
> 0.5 = large difference effect

t(*df=22*) = -2.14, p=0.04, *d* = 0.9

Critical value α = 0.05

df= 22



(a) One-tailed test (b) Two-tailed test

Statistic	df	Explanation
ANOVA: Mean Sum of Squares Within (MSW)	N - k	N: total # of all data points k: # of groups
ANOVA: Mean Sum of Squares Between (MSB)	k - 1	
χ^2	n - 1	n: Sample Size
χ^2 test for Goodness of Fit	n - 1	k: # of categories
χ^2 test for Independence	(r – 1)(c – 1)	r: # of rows, c: #columns
χ^2 test for Variance	n - 1	n: Sample Size
F	$n_1 - 1$ and $n_2 - 1$	n_1 and n_2: Sizes of the 2 Samples
t	n - 1	n: Sample Size
1-Sample t-test, and Paired t-test	n - 1	
2 (Independent)-Sample t-test	$n_1 + n_2 - 2$	n_1 and n_2: Sizes of the 2 Samples

Table T Critical Values of the *t* Distribution

df	One-Tail = .4 Two-Tail = .8	.25 .5	.1 .2	.05 .1	.025 .05	.01 .02	.005 .01	.0025 .005	.001 .002	.0005 .001
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.598
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.214	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.256	0.686	1.321	1.717	$\boxed{2.074}$	2.508	2.819	3.119	3.505	3.792
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
60	0.254	0.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
120	0.254	0.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	0.253	0.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

Source: From *Biometrika Tables for Statisticians,* Vol. 1, Third Edition, edited by E. S. Pearson and H. O. Hartley, 1966, p. 146.

https://online.stat.psu.edu/stat414/lesson/26/26.4

Independent Samples T-Test

t-value

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$\bar{x}_1$: Mean value of the first group
$\bar{x}_2$: Mean value of the second group
n_1 : Size of the first group
n_2 : Size of the second group
s_1 : Standard deviation of the first group
s_2 : Standard deviation of the second group

For equal sample size

$$df = (n1 + n2 - 2)$$

For unequal sample size

$$\text{degrees of freedom, df} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

$$\text{Cohen's Effect size} = \frac{(\text{Mean}_{\text{treatment}} - \text{Mean}_{\text{control}})}{\text{Standard deviation}_{\text{pooled}}}$$

Paired Samples T-Test

Paired Samples t-tests

$$t = \frac{\Sigma(X_{pre} - X_{post})}{SE_{diff}}$$

$$t = \frac{\bar{d}}{\sqrt{s^2/n}}$$

t-value

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$\bar{x}_1$: Mean value of the first group
$\bar{x}_2$: Mean value of the second group
n_1 : Size of the first group
n_2 : Size of the second group
s_1 : Standard deviation of the first group
s_2 : Standard deviation of the second group

t-Test: Paired Two Sample for Means

	Variable 1	Variable 2
Mean	28.0833333	33.6666667
Variance	23.9015152	57.5151515
Observations	12	12
Pearson Corr	0.06701871	
Hypothesized	0	
df	11	
t Stat	-2.2120964	
P(T<=t) one-t	0.02451926	
t Critical one	1.79588482	
P(T<=t) two-t	0.04903853	
t Critical two	2.20098516	

t-Test: Two-Sample Assuming Equal Variances

	Variable 1	Variable 2
Mean	28.0833333	33.6666667
Variance	23.9015152	57.5151515
Observations	12	12
Pooled Varia	40.7083333	
Hypothesized	0	
df	22	
t Stat	-2.1435199	
P(T<=t) one-t	0.02169075	
t Critical one	1.71714437	
P(T<=t) two-t	0.0433815	
t Critical two	2.07387307	

Cohen's Effect size = $\frac{Mean_{difference}}{SD_{difference}}$