Hypothesis Tests

Week 09 Day 02

DS 3000 – Foundations of Data Science

Reminders

HW 6

Released today

Quiz 3

Released today

Outline

Dependent-samples t test

One-way analysis of variance

Effect Sizes

Hypothesis Testing: A Magical Example

Research question:

Is the Elder Wand more powerful than any in existence?

IV: Wand

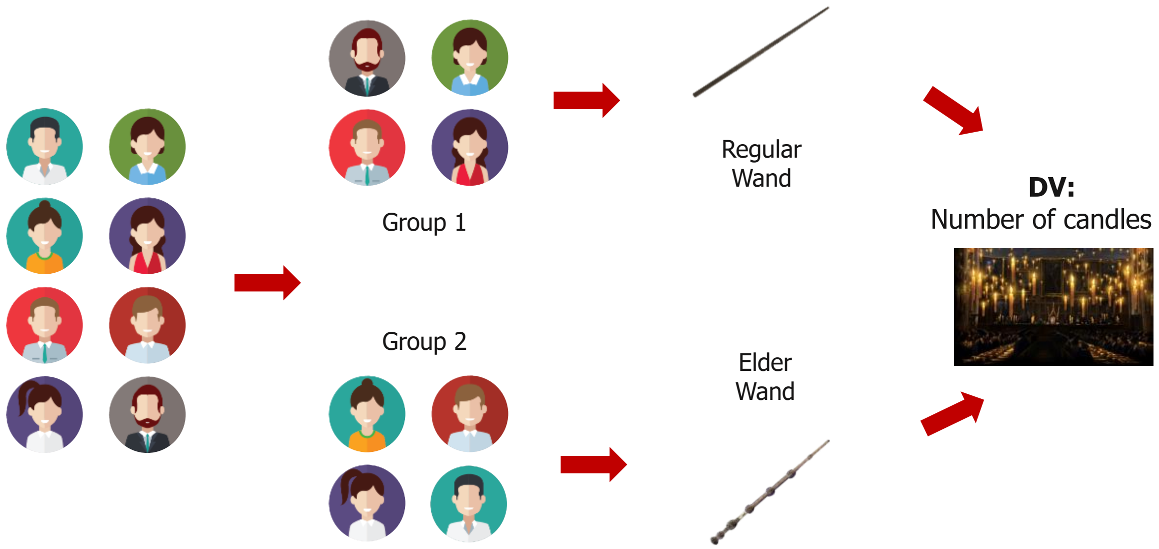
Dependent variable:

Number of candles lit

Experimental hypothesis:

Wizards will light more candles when using the Elder Wand compared to when using the regular wand

Experimental Procedure



Group 1

Group 2

Regular Wand

Elder Wand

DV:
Number of candles lit

Reporting Assumptions

For an independent-samples *t* test

Assumptions of normality, as assessed by Shapiro-Wilk's test (*p* > .05) and homogeneity (equality) of variances, as assessed by Levene's test (*p* > .05) were met.

* Usually precedes the results of the independent-samples *t* test

Reporting An Independent-Samples *t* Test

Test & Purpose

An independent sample *t* test was conducted to compare the number of candles lit by wizards using the Elder Wand and using the regular wand.

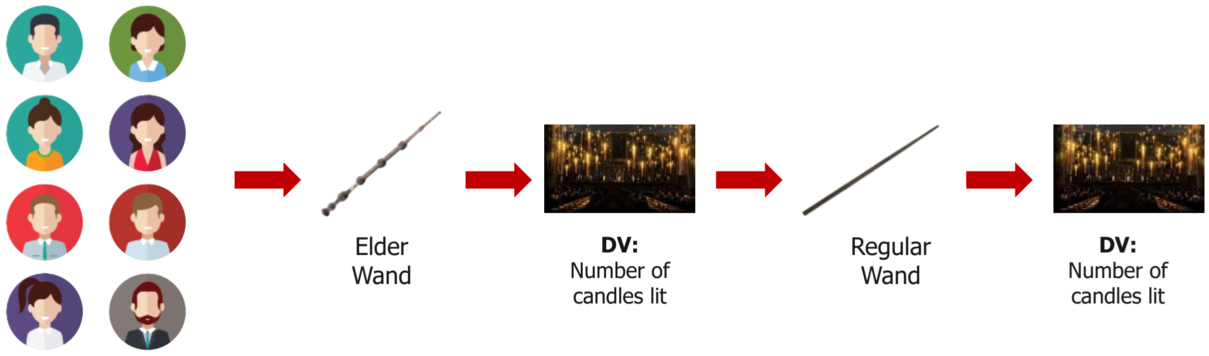
Actual results

Results showed a statistically significant difference between the wands, *t*(38) = 8.74, *p* < .001. An examination of the average number of candles lit in each wand condition revealed that the wizards who used the Elder Wand could light a greater number of candles (*n* = 20, *M* = 21.1, *SE* = .49) compared to those who used the regular wand (*n* = 20, *M* = 15.6, *SE* = .40).

Meaning

These results indicate that the Elder Wand is more powerful than a regular wand in terms of enabling wizards to perform the *Incendio* spell.

Experimental Procedure



Elder Wand

DV: Number of candles lit

Regular Wand

DV: Number of candles lit

Paired-Samples *t* Test

t Tests

Used to compare two means to see if they are **significantly** different from each other

Hypotheses

H_0: No difference between the two means ($\mu_A = \mu_B$)

H_A: The two means are different ($\mu_A \neq \mu_B$)

Two common *t* tests

Independent-samples *t* test

Paired-samples (or dependent-samples) *t* test

Paired-Samples *t* test

Used for within-subjects experimental designs

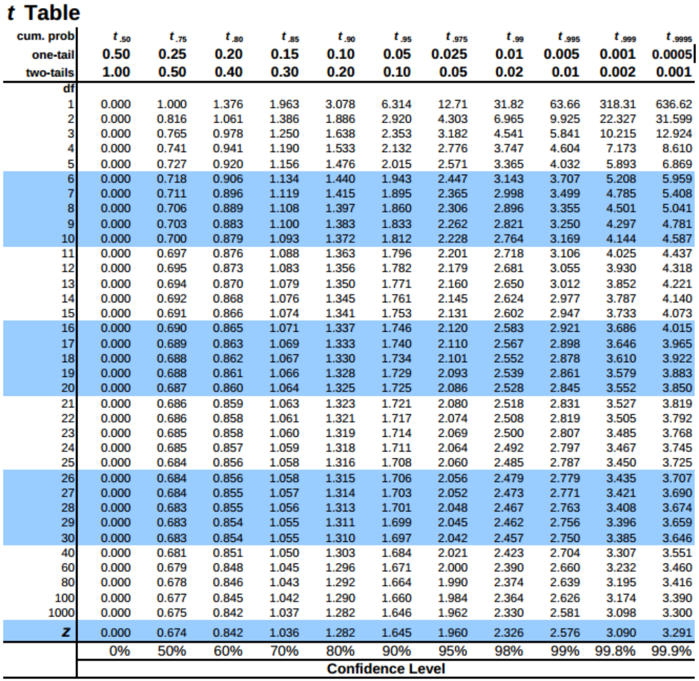
- One categorical IV with two levels
- One quantitative DV measured in two different conditions for the same group of participants

The sample mean is computer for each condition

The sample mean difference is used to test a hypothesis about the corresponding population mean difference

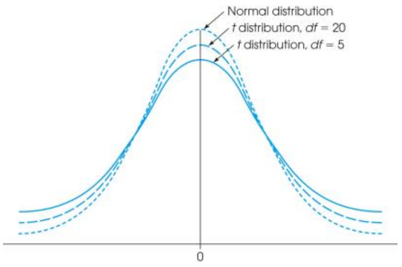
$$t = \frac{\bar{X}_D}{s_D / \sqrt{n}}$$

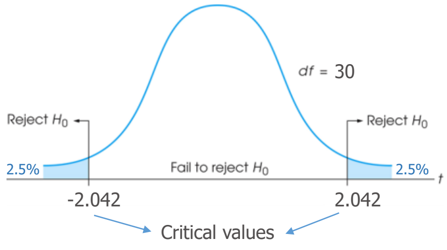
The null hypothesis states that the population mean difference is zero



***t* Table**

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										





Hypothesis Testing: A Magical Example

IV:
Wand

Research question:

Is the Elder Wand more powerful than any in existence?

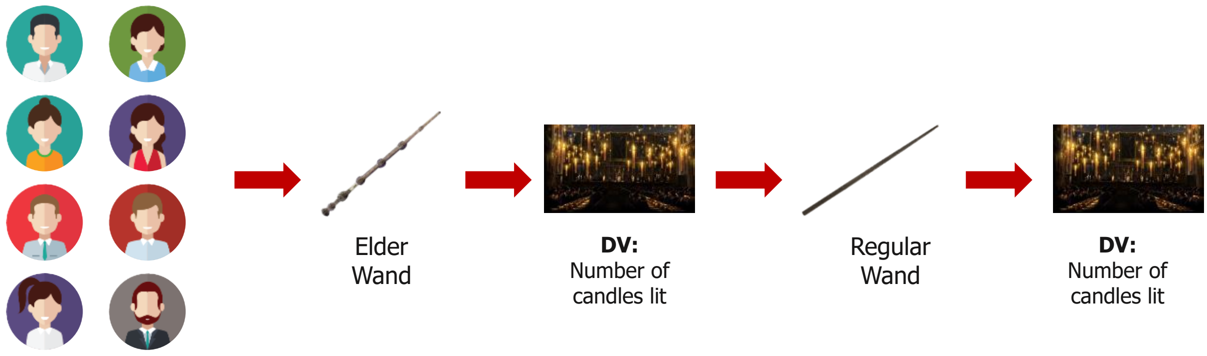
Dependent variable:

Number of candles lit

Experimental hypothesis:

Wizards will light more candles when using the Elder Wand compared to when using the regular wand

Experimental Procedure



Elder Wand

DV:
Number of candles lit

Regular Wand

DV:
Number of candles lit

Choosing Statistical Tests for NHST

Experiment Design	# of IVs	# of Conditions for each IV	Statistical Test
Between-subjects	1	2	Independent-samples *t* test
	1	3 or more	One-way ANOVA
	2 or more	2 or more	Factorial ANOVA
Within-subjects	**1**	**2**	**Paired-samples *t* test**
	1	3 or more	Repeated measures ANOVA
	2 or more	2 or more	Repeated measures ANOVA
Between- and within-subjects	2 or more	2 or more	Mixed (split-plot) ANOVA

Paired-Samples *t* Test

Assumptions

1. The levels of the IV are paired, or matched, in some way
2. The distribution of the differences in the dependent variable between the two related groups should be approximately normally distributed

SciPy allows you to check for Assumption 2

Assumption 2: Shapiro-Wilk Test of Normality

You want non-significant results from this test ($p > .05$)!

Reporting A Paired-Samples *t* Test

Report the results of checking for assumptions

Report *t* statistic, degrees of freedom (*df*), *p*-value

$t(df) = x$, $p = y$ or $p < .05$

Report the mean and standard error of the DV for each group

Group 1: ($N = a$, $M = b$, $SE = c$)

Group 2: ($N = a$, $M = y$, $SE = z$)

Reporting Assumptions

For a paired-samples *t* test

The assumption of normality, as assessed by Shapiro-Wilk's test, (*p* > .05) was met.

* Usually precedes the results of the independent-samples *t* test

Reporting A Paired-Samples *t* Test

Test & Purpose

A paired-sample *t* test was conducted to compare the number of candles lit by wizards using the Elder Wand and using the regular wand.

Actual results

Results showed a statistically significant difference between the wands, *t*(38) = 8.74, *p* < .001. An examination of the average number of candles lit using each wand revealed that wizards could light a greater number of candles (*n* = 20, *M* = 21.1, *SE* = .49) when using the Elder Wand compared to when using the regular wand (*n* = 20, *M* = 15.6, *SE* = .40).

Meaning

These results indicate that the Elder Wand is more powerful than a regular wand in terms of enabling wizards to perform the *Incendio* spell.

What if you compared three wands?

Is the Elder Wand more powerful than any in existence?

Analysis of Variance

Used to compare more than two means to see if they are **significantly** different from each other

Hypotheses

H_0: No difference among the means

H_A: There is at least one mean difference among the populations

Common ANOVAs

- One-way (single-factor) ANOVA
- Repeated measures ANOVA
- Factorial ANOVA
- Mixed ANOVA

One-way ANOVA

Used for between-subjects designs

- One categorical IV with more than two levels (three or more)
- One quantitative DV measured in each condition for different groups of participants

The mean value of DV is calculated for each group

The differences among the means from the sample data are used to test a hypothesis about the differences among the corresponding population means

The null hypothesis states that there are no differences among the population means

Hypothesis Testing: A Magical Example

Research question:

Is the Elder Wand more powerful than any in existence?

IV: Wand

Dependent variable:

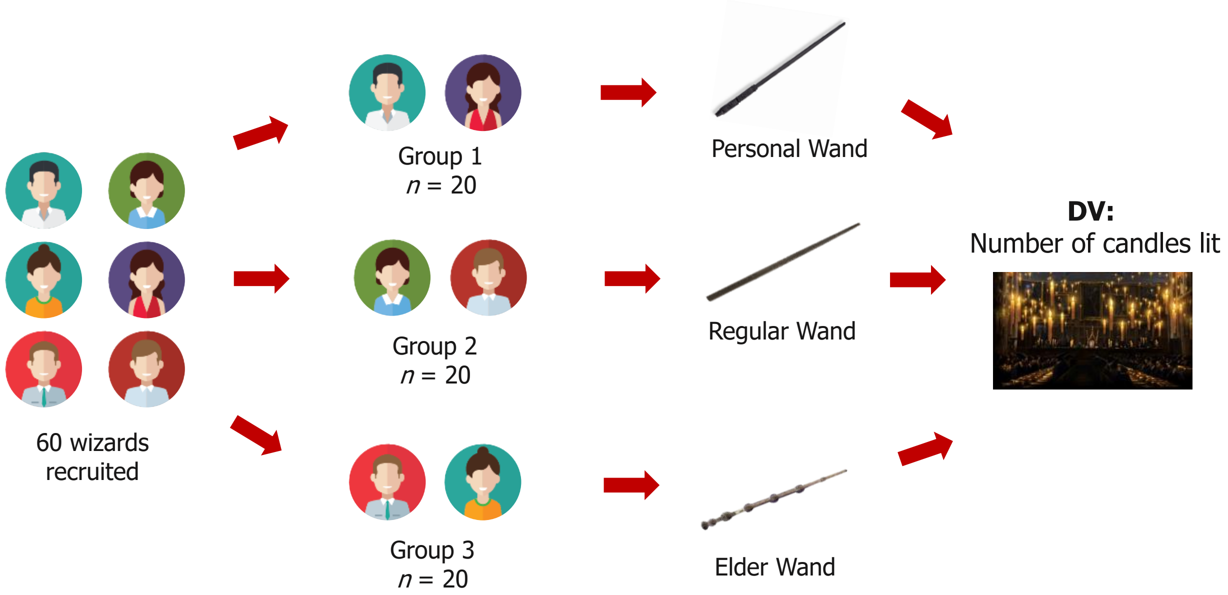
Number of candles lit

Experimental hypotheses:

Wizards will light more candles when using the Elder Wand compared to when using the regular wand and personal wand

More candles with the personal wand than with the regular wand

Experimental Procedure



60 wizards recruited

Group 1
n = 20

Group 2
n = 20

Group 3
n = 20

Personal Wand

Regular Wand

Elder Wand

DV:
Number of candles lit

Choosing Statistical Tests for NHST

Experiment Design	# of IVs	# of Conditions for each IV	Statistical Test
Between-subjects	1	2	Independent-samples *t* test
	1	**3 or more**	**One-way ANOVA**
	2 or more	2 or more	Factorial ANOVA
Within-subjects	1	2	Paired-samples *t* test
	1	3 or more	Repeated measures ANOVA
	2 or more	2 or more	Repeated measures ANOVA
Between- and within-subjects	2 or more	2 or more	Mixed (split-plot) ANOVA

Assumptions of ANOVA

Assumptions

1. The variances of the dependent variable are roughly equal across groups
2. The dependent variable is approximately normally distributed within each group

SciPy allows you to check for these assumptions

Assumption 1: Levene's Test of Equality of Variances

You want non-significant results from these tests (p > .05)!

Assumption 2: Shapiro-Wilk Test of Normality

F statistic

Compares the amount of systematic variance in the data to the amount of unsystematic variance

Based on variance instead of sample mean difference

A large *F* ratio indicates that the sample mean differences are greater than would be expected if there were no corresponding mean differences in the population

ANOVA is an omnibus test

It tests for an overall experimental effect

It tells us whether the experimental manipulation was generally successful

It does not provide specific information about which groups were affected

Post Hoc Tests

After an ANOVA you need a further analysis to find out which groups differ, if any

Post hoc tests are necessary because the original ANOVA simply establishes that mean differences exist, but does not identify exactly which means are significantly different and which are not

Post hoc tests are conducted **only after** a significant ANOVA result

Post Hoc Tests

Consist of pairwise comparisons designed to compare all different combinations of the groups of an experiment

Purpose: to determine exactly which means are significantly different by going back through the data and comparing group means two at a time

With three groups, we have three means M_1, M_2, and M_3,

A significant F-statistic tells us that $M_1 = M_2 = M_3$ is not true

Three possibilities:

$M_1 \neq M_2 \neq M_3$, or $M_1 = M_2 \neq M_3$, *or* $M_1 \neq M_2 = M_3$, *or* $M_1 = M_3 \neq M_2$

Reporting A One-Way ANOVA

Report the results of checking for assumptions

Report *F* statistic, degrees of freedom (df), *p* value, and effect size

$F(\text{df}_1, \text{df}_2) = x$, $p = y$ or $p < .05$

Report the mean and standard error of the DV for each group

Group 1: (n = a, M = b, SE = c)

Group 2: (n = x, M = y, SE = z)

Group 3: (n = x, M = y, SE = z)

Reporting A One-Way ANOVA

Test & Purpose:

A one-way analysis of variance (ANOVA) was conducted to compare the number of candles lit by wizards using the Elder Wand, the regular wand, and their personal wand.

or

A one-way analysis of variance (ANOVA) was conducted to examine the effect of the wand used on the number of candles lit by wizards using the Incendio spell. Wizards were randomly assigned to one of the wand conditions: the Elder Wand, regular wand, and personal wand.

Reporting A One-Way ANOVA

Actual Results

Results revealed a statistically significant difference among the three wands, *F*(2, 57) = 56.45, *p* < .001.

Post-hoc comparisons using the Tukey test indicated that the average number of candles lit using the Elder Wand (*M* = 21.1, *SE* = ..49) was significantly greater than the number of candles lit using the regular wand (*M* = 15.1, *SE* = .41) and personal wand (*M* = 15.65, *SE* = .43). The regular wand and personal wand did not significantly differ from each other in the number of candles lit using each wand.

Reporting A One-Way ANOVA

Interpretation

These results indicate that the Elder Wand is more powerful than both the regular wand and personal wand in terms of enabling wizards to perform the *Incendio* spell.

Results also indicate that using one's personal wand did not lead to better performance on the *Incendio* spell, compared to the regular wand, a finding incongruent with the established rules of Wandlore.

Effect Sizes

Hypothesis Testing Concerns

Hypothesis test determines whether the treatment effect is greater than chance

BUT

No measure of the size of the effect or difference is included

A very small treatment effect or difference can be statistically significant, but may not be practically significant or meaningful

Increasing the sample size increases the likelihood of obtaining a significant result

Measuring Effect Size

Purpose

To provide an objective, standardized measure of the size of a treatment effect or the difference between two means independent of sample size

Standardized measure of the magnitude of an effect

- We can compare effect sizes across different studies
- Tells us about how large the effect or difference is
- Describes practical significance of a treatment effect or difference

Measuring Effect Size with Cohen's *d*

Cohen's *d* measures effect size simply and directly in a standardized way by measuring the mean difference in terms of the standard deviation

$$Cohen's\ d = \frac{mean\ difference}{standard\ deviation}$$

***d* = 2.00**

- the mean difference is twice as big as the standard deviation

***d* = 0.5**

- the mean difference is only half as large as the standard deviation

Measuring Effect Size with Cohen’s *d*

Criteria for evaluation the size of an effect using Cohen’s *d*

Magnitude of *d*	Evaluation of Effect Size
d = 0.2	Small effect
d = 0.5	Medium effect
d = 0.8	Large effect

Need to manually calculate Cohen’s *d* for *t* tests

Cohen’s *d* for Independent-Samples t Test

$$\text{estimated } d = \frac{\text{estimated mean difference}}{\text{estimated standard deviation}} = \frac{M_1 - M_2}{\sqrt{s_p^2}}$$

$$S_P^2 = \frac{S_1^2(n_1 - 1) + S_2^2(n_2 - 1)}{(n_1 - 1) + (n_2 - 1)}$$

s_p^2 : Pooled variance

Cohen’s *d* for Paired-Samples t Test

$$estimated\ Cohen's\ d = \frac{M_D}{s}$$

$$M_D = \frac{\sum D}{n}$$

Difference score = $D = X_2 — X_1$

s = standard deviation of the differences

JASP for Hypothesis Tests

https://jasp-stats.org/

Statistical analysis package for common hypothesis tests

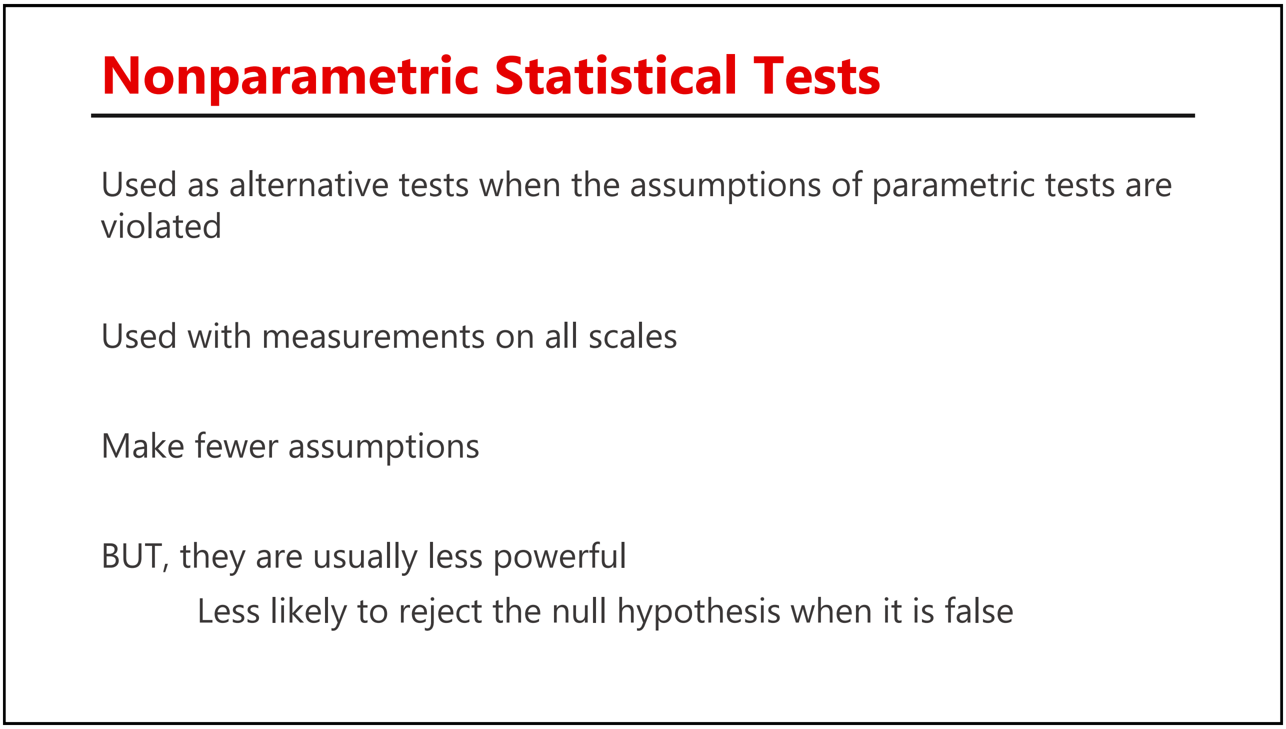
Parametric and Nonparametric Statistics

Hypothesis tests used thus far tested hypotheses about population **parameters**

Parametric tests (*t* tests, ANOVA, etc.) share several assumptions

- Normal distribution in the population
- Homogeneity of variance in the population
- Numerical score for each individual

Nonparametric tests are needed if research situation does not meet all these assumptions



Nonparametric Statistical Tests

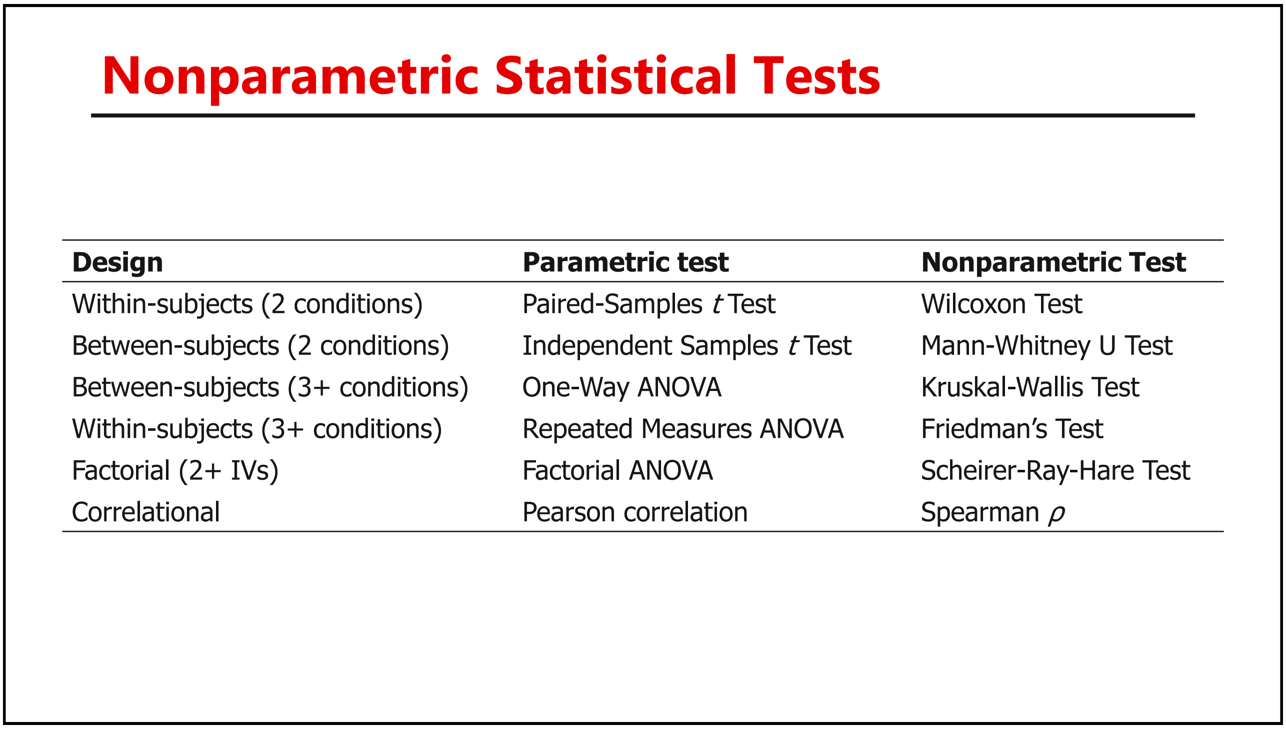
Used as alternative tests when the assumptions of parametric tests are violated

Used with measurements on all scales

Make fewer assumptions

BUT, they are usually less powerful

Less likely to reject the null hypothesis when it is false



Nonparametric Statistical Tests

Design	Parametric test	Nonparametric Test
Within-subjects (2 conditions)	Paired-Samples *t* Test	Wilcoxon Test
Between-subjects (2 conditions)	Independent Samples *t* Test	Mann-Whitney U Test
Between-subjects (3+ conditions)	One-Way ANOVA	Kruskal-Wallis Test
Within-subjects (3+ conditions)	Repeated Measures ANOVA	Friedman’s Test
Factorial (2+ IVs)	Factorial ANOVA	Scheirer-Ray-Hare Test
Correlational	Pearson correlation	Spearman ρ