Week#2 Article:

Simmons et al. (2011)-"False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis allows presenting anything as significant"

- The most costly error is a false positive, the incorrect rejection of a null hypothesis. (so, incorrectly saying that you find a sig. effect)
- Null results have many possible causes, failures to replicate previous findings are never conclusive.
- It is unacceptably easy to publish "statistically significant" evidence consistent with any hypothesis. The culprit is a construct referred to as "researcher degrees of freedom".
- It is common for researchers to explore various analytic alternatives, to search for a combination that yields "statistically significant" and then report only what "worked".
  **why? 2 reasons:
  - Ambiguity in how best to make these decisions
  - The researcher's desire to find a statistically sig. result.
    (ppl are self-serving in their interpretations of ambiguous information and adept at reaching justifiable conclusions that mash with their desires)

*How researcher **degrees of freedom** influence the probability of a false-positive result:

Flexibility in:

- Choosing among dependent variables,
- Choosing sample size
- Using covariates
- Reporting subsets of experimental conditions

Solution:

6 requirements for authors:

1) Authors must decide the rule for terminating data collection before data collection begins and report this rule in the article. (reporting power calculations or disclosing arbitrary rules such as " we decided to collect 100 observations")
2) Authors must collect at least 20 observations per cell or else provide a compelling cost-of-data collection justification
3) Authors must list all variables collected in a study. (prevents authors from reporting only a convenient subset of the many measures)
4) Authors must report all experimental conditions, including failed manipulations (prevents authors from selectively choosing only to report the condition comparisons that yield results that are consistent with their hypothesis)
5) If observations are eliminated, authors must also report what the statistical results are if those observations are included (justify the elimination of data)
6) If an analysis includes a covariate, authors must report the statistical results of the analysis without the covariate. (reporting covariate-free results makes transparent the extent to which a finding is reliable on the presence of a covariate, puts pressure to justify the covariate)

Guideline for reviewers:

1) Reviewers should ensure that authors follow the requirements.
2) Reviewers should be more tolerant of imperfections in results

3) Reviewers should require authors to demonstrate that their results do not hinge on arbitrary analytic decisions.
4) If justifications of data collection or analysis are not compelling, reviewers should require the authors to conduct an exact replication.

—----------

Lec Notes/ Descriptives and One-sample t-test:

2 sets of Analysis:

DESCRIPTIVES

① Inferential Statistics → making inferences, testing hypoth./res. qu

② Descriptive " → describing how your data look like

Exp. in JASP: each Ps given 2 types of sleeping pills; their extra hs. of sleep was recorded. → a repeat measures design (Ps are assessed repeatedly in multiple occasions)

extra	group	ID
0.7	1	1
-1.6	1	2
-1.2	1	3
0.8	1	4
⋮	2	⋮
⋮	2	⋮
⋮	1	2
⋮	⋮	3
⋮	⋮	4

Ps measured in diff. conds

↳ 2 ways to structure your dataset:

1) Long format (like this exp.) → each row is a measurement

2) Wide format → every Ps is noted by a single row

It would look like in this exp

ID	extra1	extra2
1	0.7	0.6
2	-1.2	0.3
3	⋮	⋮
4		
5		

↳ Desc. Stat.:

select units "extra"

valid	20
missing	0
mean	1.54
SD	2.01
min	-1.6
max	5.50

→ here, it says 20 cases but we know that not all of these are independent, each 2 case comes from a single particip.

↳ split by "group"

	1	2
valid	10	10
mean	0.70	5.53
SD	1.78	2.00
min	-1.62	-0.10
max	3.70	5.50
std. error of mean	0.55	0.63
variance	3.20	4.00

→ can split by 2 conds to see sleep gain after taking diff. pills

valid → * for each cond. we have 10 data points. There are no missing values.

mean → * Ps who take pill 1 on average gain 0.7 hs of sleep and who take pill 2 gain on average 5.5 hs of sleep.

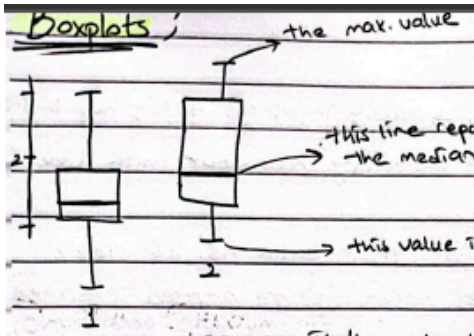
median → * middle value, If you rank all sleep assessms. from high-to-low, what is the mid point? → median

variance → * variability in data / describing how much ppl deviate from the mean, how much ppl are away from the average value / how far each of the datapoints from the mean → summary index of deviations from the mean.

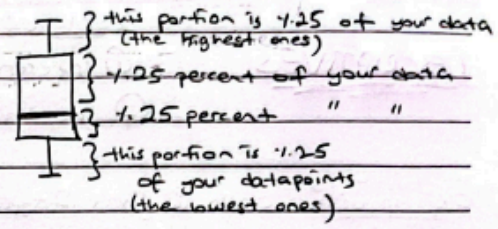
SD → * helps us to make meaning of effect sizes / how much the raw scores tend to deviate from the mean. A small SD means → raw scores deviate a little from the mean, so there is little within group variability (Ps had similar scores within each conds)

Std. error of mean → * indicates how different the population mean is likely to be from a sample mean. It tells you how much the sample mean would vary if you were to repeat a study

Boxplots;



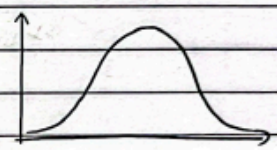
* the box creates 4 quartiles in your data



→ Is the standard deviation of the sampling distribution of mean

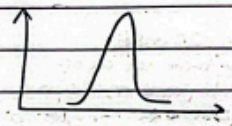
(*) Std. error of mean → the lower the sample size = the higher the std. error = wider the sampling distribution = less reliable

Exp: we run a study and mean looks like;



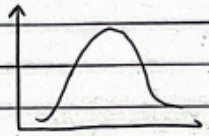
↳ repeated this study 10.000 times;

↳ ① with N=25 →



→ distribution of means is much narrow than distribution with only 5 ppl.
so lower std. error = more reliable

↳ ② with N=5 →



Running a hypothesis test; (Null Hypoth. Significance Testing)

Exp: One-Sample t-test. (It compares a sample mean to a defined point)

Ps completed 8 week weight gain program.

	Weight Before	Weight After	Difference
1	122.54	135.74	13.2
2	120.78	129.30	8.58
3	137.06	⋮	14.00
4	⋮	⋮	⋮
5			
6			

Hypoth.; Does the average weight gain at the end of the program exceed 5 pounds?

① first, look at descriptives

Mean = 10.40 (mean weight gain) → Question is; whether this mean is statistically diff. than 5 pounds

(2) Run, one-sample t-test → first, you compute a test statistic

is what you are interested in here, we are interested in the diff. btw sample mean and pre-determined value (5 pounds) (10.40)

↓

then, the null hypoth. testing paradigm will compare this test statistic to a known sampling distribution of mean.

* If test statistic lies at the extreme of this distribution → conclude that there is enough evidence for a sig. finding

! We are trying to reject the null hypoth. (sample mean = 5) and favor the alternative hypoth. (sample mean > 5)

↳ In Jasp; * Test value : 5

* select Alt. Hypoth. as "≠ Test value" → described as Two-tailed (non-directional) tests (almost always select this)

* Check "Location parameter"

this is the value of test statistic (which is what we are interested in; in this case the diff. btw sample mean and pre-determin. value 15 pounds)

= So, (10.40) − (5) = 5.40

sample mean, pre-det. value, mean difference (point assterment)

this called the point assterment; a single assterment based on your sample

So, based on your sample, the mean is 10.40, so it is 5.40 pounds more than your test value (pre-determ. value)

(3) The next step is to see whether 5.40 is statistically larger than 5 pounds. To do that; you need to compare 5.40 to the sampling distribution of mean.

↳ So we convert the point assterment (5.40) to a t-score, to see where it falls under the t-distribution

is what we call a t-distribution

↳ one sample t-test

	t	df	p	mean Difference
Difference	5.63	15	<.001	5.40

(4) If test statistic falls on the extreme ends of the distribution, your p-value will be very small. (reliable)

Here; p value is <.001, which means average weight gain (10.40) is greater than 5 pounds. So we can reject the null hypoth.

NOTE!; If you repeat the study with diff. sample, you would get a diff. point assterment. So, what is the interval of those point assterments? If I repeat the study many times, what are the plausible values of this mean difference? → this informat. is provided by Confidence Interval

(3) Confidence Interval → .95 of the time, if we repeat the study, based on the sample information I have, it's likely that the effect will be somewhere btw 3.63 to 7.45 pounds difference.

	t	df	p	Mean Difference	.95 CI for Mean Difference Lower	Upper
ference	5.63	15	<.001	5.40	3.63	7.45

(4) The Effect Size → the p-value just tells you whether the effect is statistically sig. or not.

↳ Here, we have an effect; average weight gain is signly larger than 5 pounds. (5.40 is the size of the effect)

(4) But, how big is that difference? how big is that effect? (→ diff in kg, lbs, etc.

So we need a standard effect to compare among studies.

↳ For, one sample T-test, Cohen's d is used for effect size. → means; how many SD is your effect. (it measures distance in terms of SD)

In this exp; our Cohen's d is 1.40.

Week#3: Independent and Paired Sample t-test

One Sample T-Test

	t	df	p	Mean Difference	95% CI for Mean Difference Lower	Upper	Cohen's d
Difference	5.633	15	< .001	5.409	3.362	7.455	1.408

Descriptives

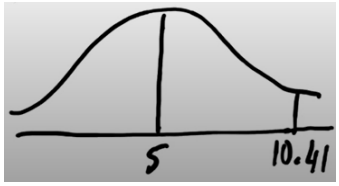
	N	Mean	SD	SE
Difference	16	10.409	3.841	0.960

The basic idea behind null hypothesis sig. testing is to formulate 2 representations of reality: null hyo. vs alternative hyo.
In this exp: the null hypothesis was “mean weight gain = 5 pounds” and the alternative hypo was “mean weight gain is not equal to 5 pounds”

How can we test that?: just compute the mean difference (test statistic) btw mean in your sample and whatever value you are testing against. Here, mean in our sample is 10.409 and predetermined value is 5 pounds so our mean difference (point estimate) is 5.40

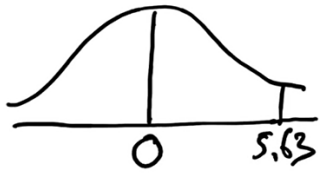
The question is whether this 5.40 is sig.ly greater than 5 pounds?
How do we do that?: comparing test statistic with sampling distribution of mean
(to see if 10.40 falls on an extreme end of the sampling distribution of mean of 5)

- Sampling distribution of mean is a normal distribution; meaning that it is symmetric around its mean.

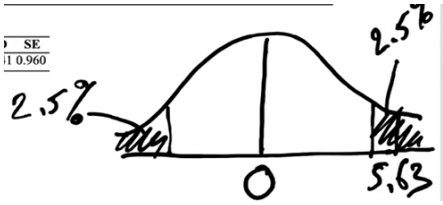


This is sampling dist. with a mean of 5.

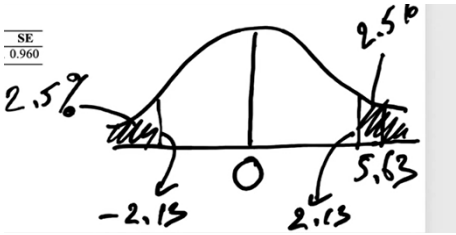
So, whether this 10.40 is far enough from 5 to conclude that it is sig.ly different value than 5?
How do we decide this?:
The sampling distr. has the characteristics of what we call a t-distribution. So we can use these characteristics of t-distribution to make our decision.
So, we have to convert our values to t-values.
How to do that: t= x-M/ SE
When we convert our values to t-values:



Now, the question is whether 5.63 is sufficiently away from the mean?
How far is far enough?
**if the value right to your test statistic does correspond to 2.5% of area under the curve, any test statistic that falls within these areas are considered as statistically significant. (2.5 + 2.5 = 5%, which corresponds to p =.05)

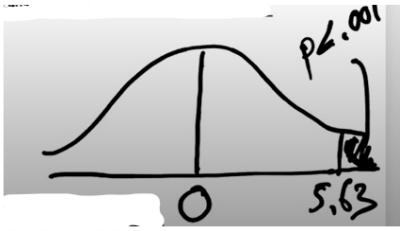


What marks 2.5%?
In this example, the critical t-value is 2.13. (depending on the shape of the t-distribution, this critical t-value changes. The shape of the t-distribution changes based on degrees of freedom)



What is the degree of freedom?: is the num of free parameters you can assign
(your sample size - 1 = df)
So, now that we know our converted test statistic (5.63) is greater than the critical t-value (2.13); the mean weight gain of 10.40 is greater than 5 pounds.

What does p value represents here?: p value is the area that falls of your obtained t-value.



P value is the likelihood that you will obtain as extreme as the score of 5.63 if the null hypo was correct. So, the probability of obtaining this result when the null hypothesis is true should be very little. (in this case it is small, less than .001)

What is confidence Interval?: the confidence interval of the mean difference will tell you if you repeat this study many times, where will your mean difference be. If I repeat the study many times, what is the plausible range of values of this mean difference? = CI (gives you lowest and highest CI)
The smaller the Stand.error, the narrower the CI (the more precise your estimation)

Formula: CI=mean difference +/- (t critical x SE)

What is effect size?: = how big is the effect? It is good to have a standardized effect size to compare across studies (such as Cohen's d).
For exp: in this exp we had pounds, but what if other studies examine this by kg. So standardize your effect.

$$\text{Cohen's } d = \frac{\text{Mean diff}}{SD} = \frac{5.41}{3.84} = 1.41$$

so, basically this diet brings about on average 1.41 SD increase in weight gain if your comparison point is 5 pounds.

Independent (between-participant) Sample t-test:

- We will have 2 sample means (we have 2 groups) and we will compare them with each other.

Exp: 24 students, randomly assigned to 2 conds. (experimental cond: they were asked to wear an invisibility cloak, control cond: not instructed) DV: the num. of mischievous behav. they engaged in. Hypoth: those who were assigned to the experim.cond would engage in more mischievous behav. (null hypot is the equality of the group means, suggesting that there is no sig. diff. Btw these groups)

Independent Samples T-Test

Independent Samples T-Test

	t	df	p	Mean Difference	SE Difference	95% CI for Mean Difference Lower	Upper	Cohen's d
Mischief	−1.713	22	0.101	−1.250	0.730	−2.763	0.263	−0.700

Note. Student's t-test.

Descriptives

Group Descriptives

	Group	N	Mean	SD	SE
Mischief	0	12	3.750	1.913	0.552
	1	12	5.000	1.651	0.477

Looking at the means we can say there is a difference, but we will examine whether this difference is statistically significant.
First, convert our test statistic (mean difference; 1.25) into a t-value (1.71) and see if that value is an extreme value on the t-distribution.
And our p value is 0.10, meaning that t-value of 1.71 roughly corresponds to an area of 0.10.
So, this difference is not statistically sig. = the value involved in the null hypot ("mean difference is 0") will be in the CI (Looking at the CI: we see that upper CI is 0.26)
**so we can interpret whether our result is sig. Without needing the p-value. For a t-test, if the CI does not include 0, then it means that the finding is sig.

Paired (Within-participant) Sample t-test:
Exp: 15 Ps, measuring their aggression on the full moon days and other days
**the Ps each provided 2 measurements.
For each Ps, it will compute the difference btw aggression levels on full moon and other days.
Then, through difference scores the analysis will provide a single sample, and run a one sample t-test.

Paired Samples T-Test

Paired Samples T-Test

Measure 1		Measure 2	t	df	p	Mean Difference	SE Difference	95% CI for Mean Difference Lower	Upper	Cohen's d
Moon	-	Other	6.452	14	< .001	2.433	0.377	1.624	3.241	1.666

Note. Student's t-test.

Descriptives

Descriptives

	N	Mean	SD	SE
Moon	15	3.022	1.499	0.387
Other	15	0.589	0.445	0.115

Mean difference of aggression btw full moon days and other days is 2.43. Converted t-value of mean difference is 6.45. And p value is <.001, suggesting a significant difference.

How to convert mean difference into t-value?:

We have 2 SDs, so first compute the pooled SD: $Pooled\ SD\ = \sqrt{\frac{(SD_1{}^2 + SD_2{}^2)}{2}}$

Then, convert it to t-value by formula: t= (M_1-M_2) / SE; SE = $P_{SD}.\sqrt{1/n1\ +\ 1/n2}$

How to find critical t-value?

Exp: For the degrees of freedom of 62 (n1: 49, n2: 15 →49+15-2), the critical t-value is something like 1.99. (use online computational tools) So our computed test statistic (3.53) falls under this area, thus we can say that two groups are significantly different from each other.

Week#4: Power analysis and Chi-square test

Null Hypothesis testing: the idea is to define 2 representations of reality.

		Null Hypothesis (H_0) is:	
		True	False
Judgment of Null (Statistical Result)	Reject H_0 ($p < .05$)	Type I Error False Positive α	Correct Inference True Positive (1 – β)
	Fail to reject H_0 ($p > .05$)	Correct Inference True Negative	Type II Error False Negative β

False Positive (alpha)/Type 1 error: you are claiming there is an effect (rejecting the null), but in reality there is none (null hypot is true)

True negative/Correct ınference: you are claiming there is no effect (fail to reject the null) and there is no effect indeed (null hypo is true).

True Positive/Correct Inference: you claim there is an effect (reject null), and there is effect indeed (null is false).

False negative (beta)/type 2 error: you claim there is no effect (fail to reject the null), but actually there is an effect (null is false).

Power is basically: 1- type 2 error (1-β). So, statistical power is probability of rejecting null hypo when it is false (probability of finding an effect when it is real).

Depending on the test you are conducting, the formula for power analysis will be different. But, we can discuss the underlying commonalities:

What affects statistical power? (regardless of the test you are using, the statistical power is the function of these 3 things)
-the alpha lvl: is typical .05
-the effect size: for larger effect size, the statistical power will be higher
-the sample size: larger the sample size the more powerful your test is

**power analysis: to determine your minimum sample size required to obtain a predetermined effect at a certain alpha level at a certain power.
(generally we aim 80% power)

When running **apriori power analysis:**
We need to know alpha level, statistical power and effect size.
We can take .05 as alpha and aim for %80 power but how to determine effect size?
-check previous literature! (effects obtained in similar studies)
-or if there are no similar literature, think about what practically most meaningful effect size would be? Based on the topic of your study you can have an idea of what the minimum practically meaningful effect is.
Or you can think about the average effect in your field of study (psy typically deals with smaller effects like .20)

Research can also use preexisting data to use in a new study: run a **sensitivity power analysis**
-here you can use sample size, alpha level and power to estimate the minimum effect size that can be reliably detected in your sample.
Exp: you have 315 Ps sample and you want to run an independent sample t-test
Run sensitivity power analysis to see what the minimum effect size that sample of 315 ppl can detect with reasonable power (%80-90) at alpha of .05.
**so estimating the minimum effect that can be reliably detected in your data.

! Sometimes num of parameters you need to run apriori analysis are too much, there will be num of variances you need to have an idea about. So, when using complex models, researches suggest complementing apriori analysis with a sensitivity analysis. Because you base your apriori analyses on too many assumptions like multiple effect sizes and variances.Once you collect your data you can run an additional sensitivity analysis.

Why power analysis Imp?
-statistical power is basically your ability to find an effect if the effect is real. So, in order to detect the effect you need statistical power.
-if your design is not sufficiently powered, sig. results are more likely to be false positives. So, statistical power is also imp for reproducibility (to ensure the effect you are obtaining is not a false positive)

Exp: lets say you run a low powered study (only have 40-50% power to detect an average effect in your field of research) and you failed to reject the null hypothesis (you did not find enough evidence to reject null). So, what does that null finding mean?

- The null finding won't tell you much because maybe the effect is there but you didnt find it as you have low power

Or you run the same study but this time find a sig. effect. What does that mean?

- The sig. finding won't tell you much because there is a risk that finding is false positive

!!So low powered study literally tells you nothing either you have a sig or no non-sig result.

On the other hand; an adequatly powered study

- If sig.: tells you its a reliable finding
- If non-sig.: still gives you clues about that there might be an effect, but it perhaps be smaller than your effect size.

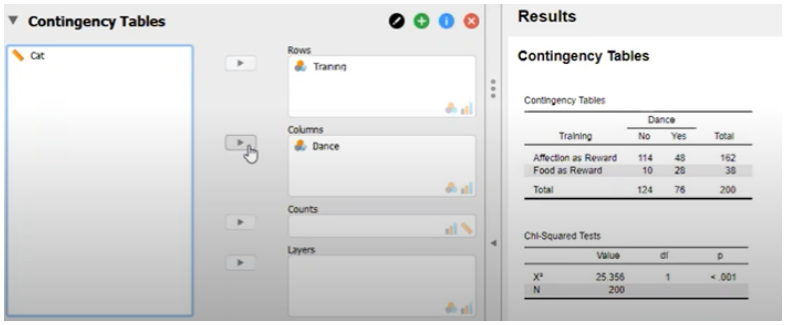
Chi-Square test:

When comparing 2 conditions or 2 groups, but you dont have any information about the means. Instead, you have num of observations in each conds.

Exp: 200 cats as Ps, either trainer used food as reward or affection as reward. the outcome measure is whether the cat learned to dance or not (which is not a continuous scale).
Does the kind of reward differ in predicting outcome variable?
(Null will be: type of reward is independent from dancing bhv)

Here we are comparing frequencies; the num of cats that learned to dance in food cond, and num of cats learned to dance in affection cond.
There are no means or sample averages, but instead counts/frequencies.
We will conceptualize the difference in frequencies and compare that difference to the known sampling distribution.

In Jasp; go to frequencies → contingency tables



You get 4 groups: affection/ dance, affection/no-dance, food/dance, food/no-dance
Out of 200 cats, 162 cats trained with affection; 38 trained with food
Out of 200, 76 learned to dance

If null is true: if reward has no effect on learning to dance, what would be the expected valued in each 4 groups?

For exp: if affection reward has no impact on dance, how many cats would we expect to see in affection/no-dance cell?
Simple probability computation: if affection reward and dance were completely independent we would expect the num of cats that fall in the affection/no-dance cell would be: (162 x 124)/ 200
or, we would expect the num of cats that fall in the affection/dance cell would be: (162x76)/200

So, we should first look at expected frequencies: (total row x total column / total sample)

Contingency Tables

Contingency Tables

Training		Dance No	Dance Yes	Total
Affection as Reward	Count	114.000	48.000	162.000
	Expected count	100.440	61.560	162.000
Food as Reward	Count	10.000	28.000	38.000
	Expected count	23.560	14.440	38.000
Total	Count	124.000	76.000	200.000
	Expected count	124.000	76.000	200.000

Whether the actual counts you obtained are different from expected counts?
For exp: if affection and dance were completely independent we would expect out of 168 cats, 61 of them to learn dance.

In the actual data, 28 cats learned to dance in food reward (twice the expected count)
And, 48 cats learned to dance in affection reward (1.5 less than expected count)

****Compare your expected counts to the sampling distribution of frequencies. Frequencies are distributed along chi-square distribution which is an asymmetrical, skewed, one-tailed distribution.

Chi-Squared Tests

	Value	df	p
X^2	25.356	1	<.001
N	200		

The chi-square here (25.35), is a numerical index describing the deviation of actual counts from expected counts. The higher the deviation, higher the chi-squared value.
And our chi-squared value is sig., meaning that food is a more effective reward than affection when teaching cats how to dance. So, type of reward and dancing is not independent from each other; type of reward has an impact on the probability of the cat learning dance.

How big is this effect?
The effect size measure will be association-based; called as variance explained effect size. The effect size described in terms of variability you explain in your outcome.
Here, the variability is whether the cat learns to dance or not (2 outcomes)

Nominal ▼

	Value
Contingency coefficient	0.335
Phi-coefficient	0.356
Cramer's V	0.356

These coefficients try to estimate the same thing: what % of variability of reward type explain dancing ability? The amount of variability you explain by your predictors
(The higher the coefficient, the closer the association or greater the dependence btw 2 events.)

Week 4 article:
Nosek et al. (2018) - The preregistration revolution

Distinction between postdiction and prediction:
Postiction: characterized by the use of data to generate hypotheses about why something occurred. The data are already known and postdiction is generated to explain why they occurred. Generating predictions is vital for discovery of possibilities not yet discovered.
Prediction: characterized by the acquisition of data to test ideas about what will occur.The data are used to confront the possibility that the prediction is wrong. Testing predictions assess the uncertainty of scientific models by observing how well the predictions account for new data.

Failing to appreciate the difference between postd vs. predict. can lead to overconfidence in post hoc explanations (postdictions) and inflate the likelihood of believing that there is evidence for a finding when there is not.
Presenting predictions as predictions can increase the attractiveness and publishability of findings by falsely reducing uncertainty, and this will decrease reproducibility.

Hindsight bias: the tendency to see outcomes as more predictable after the fact compared with before they were observed. With hindsight bias, the observer uses the data to generate an explanation (postdiction) and simultaneously perceives that they would have anticipated that explanation in advance (prediction).
Fallibility of memory, motivated reasoning and cognitive biases can occur outside of conscious awareness. And researchers may employ confirmation bias by seeking evidence consistent with their expectations.

Null hypothesis testing: designed for prediction generating hypothesis (not postdictions).
Garden of forking paths: there are a vast num of choices for analyzing data. Researchers have substantial degrees of freedom to conduct many different tests and selection of those that benefit them. If those choices are made during analysis, observing the data may make selecting some paths more likely and others less likely. This leaves the observed p-values with unknown diagnosticity, rendering them uninterpretable. So, null testing cannot be used with confidence for postdiction. (ın prediction, the problem of forking paths is avoided because he analytic pipeline is specified before observing the data)

If researchers are clear about when they are in prediction and postdiction models of research, then the benefits (limits) of statistical inference will be preserved. However, with means, motive and opportunity to misperceive postdiction as prediction and to selectively rationalize and report a biased subset of outcomes, researchers are prone to false confidence in evidence.

Solution?: pre-registration (posting the analysis plan to an independent registry before observing data) = helps researchers maintain clarity btw prediction and postdiction and preserve accurate calibration of evidence.

- Pre-registration distinguishes predic and postdic
- Hypothesizing in advance relates to increased replicability
- Reduce publication or reporting bias
- Preserves the diagnosticity of null hypothesis testing inference for predictions
- Clarifies the role of postdiction for generating possible explanations to test as predictions in the future.

Challenges of pre-registration:

- <u>Changes in procedure during study administration</u> (can be deviations that may not be predicted and these may alter the analysis plan)

- <u>Discovery of assumption violations during analysis</u> (some violations may not be identified until observation of the data such as ceiling effect or anormal distribution)
    - Blinding the dataset by scrambling some of the observations without knowing the actual outcomes, or pre-registration of decision tree (the sequence of tests and decision rules at each stage of the sequence) can be solutions.
- <u>Using data that is preexisting</u> (there are a lot of ways to be directly/indirectly influenced by observed data. Knowing some outcomes might influence decisions even if the analysis is on different variables from the dataset)
    - The researcher has to be blind about the data
- <u>Longitudinal studies and large, multivariate datasets</u> (longitudinal design amplifies the challenges of preexisting data and changes to protocols after pre-registration)
    - Pre-registraring in advance of new observations gain some benefits of blinding.
- <u>Multiple experiments</u> (pre-reg every experiment seems highly burdening)
    - Pre-regis the design of the experiments themselves, a template that defines variables and parameters for the protocol and which parameters will be changed or manipulated for each experiment.
- <u>Area of research</u> (can be high risk and most research outcomes may be null, so researchers can engage in selective reporting)
    - Preregis avoid problem of selective reporting and encourages transparency.
- <u>Having few apriori expectations</u> (postdictions can be recast as predictions, because of hindsight bias)
- <u>Having competing predictions (</u>about what will occur because of researchers distinct theoretical orientations)
- <u>Narrative inferences and conclusions</u> (researchers' narrative focus may be on the more interesting results; may lead increase in inferential error and failing to correct for multiple comparisons)

- To correct this: apply alpha corrections like Bonferroni, such that narrative focus on just positive results is not associated with inflated likelihood of false positives.

—----------

What if we have to check the data min-collection (for exp: we are collecting data for a rare case, and it takes too much effort and time; we have to check data mid-point to see at least it is worth studying it)? = use sequential analysis to avoid confirmation bias and prevent false positives.
Before starting the study:
How many times you are going to check the data
What would be the p-value at each check

—----------

Week 5: ANOVA

Use ANOVA when we have more than 2 conds we want to compare. It can be within subjects (repeated measures) or between subjects. “One-Way” means that we have 1 IV.

Between Ps ANOVA:

EXP: 134 Ps, Ps were asked to rate Facebook profiles (1:least liked, 7: the most liked). In experimental group; the num of friends in the profiles were manipulated (cond1: 102 friends, cond2: 302 friends, cond3:502 friends, cond4: 702 friends, cond5: 902 friends)
R.Q: whether num of friends in a social media account affects impressions of other people?

	Score
Valid	134
Missing	0
Mean	4.382
Std. Deviation	1.146
Minimum	1.000
Maximum	7.000

	Score				
	102	302	502	702	
Valid	24	33	26	30	21
Missing	0	0	0	0	0
Mean	3.817	4.879	4.562	4.407	3.990
Std. Deviation	0.999	0.851	1.070	1.428	1.023
Minimum	2.200	2.600	2.000	1.000	1.600
Maximum	6.000	6.400	6.800	7.000	5.200

*Idea is to see whether these means are statistically different than each other.
Idea behind ANOVA:

- to compute between group variability. (How much ppl in different groups differ from each other) so differences based on group membership
- It also computes within group variability: focuses on how much ppl in the same group differ from each other?
- And **if within group variability is greater than between group variability: then no btw group differences**

$$Score_{ij} = \mu + u_i + r_{ij}$$

Outcome score in cond i of person j= grand mean + deviation from grand mean due to person's membership + any remaining deviation that my model cannot explain (residual from the explained part)

ANOVA will see: whether if *deviation from grand mean due to being a member of a particular group* is statistically meaningful

To see how much each group's mean deviate from the grand mean:

$$\sum n \cdot (\text{group mean} - \mu)^2$$

∑ num of Ps in the group .(group mean - grand mean) ^{2} = btw group variability
(do it for all groups and sum it up, you will get btw group variability)

$$df = \#\text{ of groups} - 1$$

To look at within group variability (also called residual variability):

$$\sum (x - \text{group mean})^2$$

∑ (each Ps' score - group mean) ^{2} = within group variability

$$df = N - \#\text{ of groups}$$

(N is sample size)

To get mean squares of btw groups:

$$MS_{between} = \frac{SS_{bw}}{df_{bw}}$$

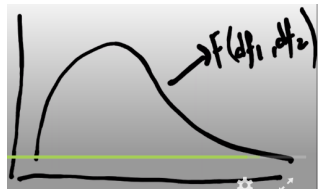
: just divide sum of squares (btw group variability) to df

To get mean squares of within groups: just divide within group variability to df

The next step is to compare mean square of btw groups and mean square of within group:

$$F = \frac{MS_{bw}}{MS_{wn}}$$ if you dive them; you will ***get F statistic***

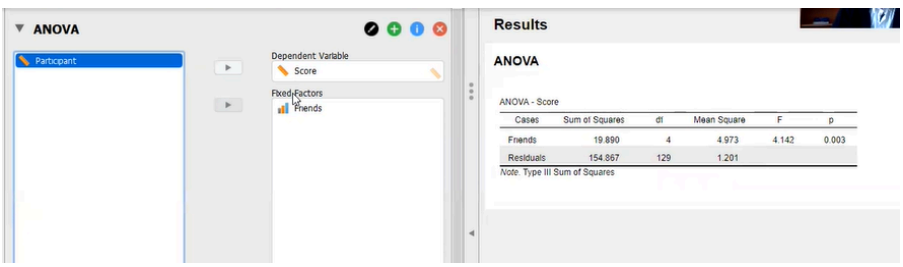
!!!The F test is always one-tailed



!!F test is skewed toward to right hand tail
!!it changes shape depending of df
!!we have 2 df (one of MS btw, one for MS within)

!!!you will check whether your test statistic is at the extreme ends of the distribution, if so, you have a sig. result.

How to do it in JASP:

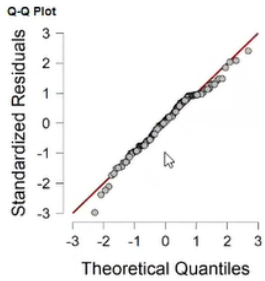


This sig. result just tells you that at least one group mean is different from the others. But we dont know which group(s)?
And how big is this effect: we often use eta squared(η 2) effect size (the variability you can explain in the outcome): how much variability does your effect explain in the DV

Eta square calculation: sum of squares betw /(sum of squares betw + sum of squares within)
!! 0.14 - 0.15 are considered big effects for human behav.

Assumption Checks: there are 3 assumptions that you meet to perform ANOVA:

- **<u>Independence of residuals:</u>** the residuals from diff Ps should not be correlated, dependent on each other (if you randomly assign Ps, then your sample will be independent) However, sometimes because of the nature of the study, you will violate this assumption, such as in couple's study. Answers from couples can be dependent as they share so much so the correlation btw their residuals would be larger than a correlation btw an 3rd residual.
- **<u>Residuals should be normally distributed</u>** (to check: looking at a Q-Q plot)



, if the dots varied from the line than the assumption of normality is violated. Here, we see some plots away from the line but it is really close still.

- **<u>The variance across the groups should be the same</u>**: in order to reliably compare group means, the group variances should be roughly be the same. Otherwise, we won't have an idea of whether the differences in distribution is coming from the variance or coming from the mean. Checked by *<u>Homogeneity (Leven's) Test:</u>* if sig.at least one variance is diff. than others

(what counts as ig.?: if Leven's test is less than .001, than check the data or do additional tests)

**If Leven's test is sig. you have to correct that: there are 2 alternative tests you can run:

- Brown-Forsythe and Welch tests
- These are also F tests
- Whenever you violate an assumption, the corrections will involve penalizing yourself by degrees of freedom. Violation of assumption leads to increased type 1 error, meaning that finding a sig. effect when the effect is actually non-existent. In order to prevent that, you penalize yourself by decreasing degrees of freedom so that finding a small t value becomes more difficult.

Which means are different from each other?
→ can do a post hoc test: compares pairs of means to each other, consider all the possible pairwise comparisons so we can see which groups are diff from each other. (basically a t-test)

Post Hoc Tests ▼

Standard ▼

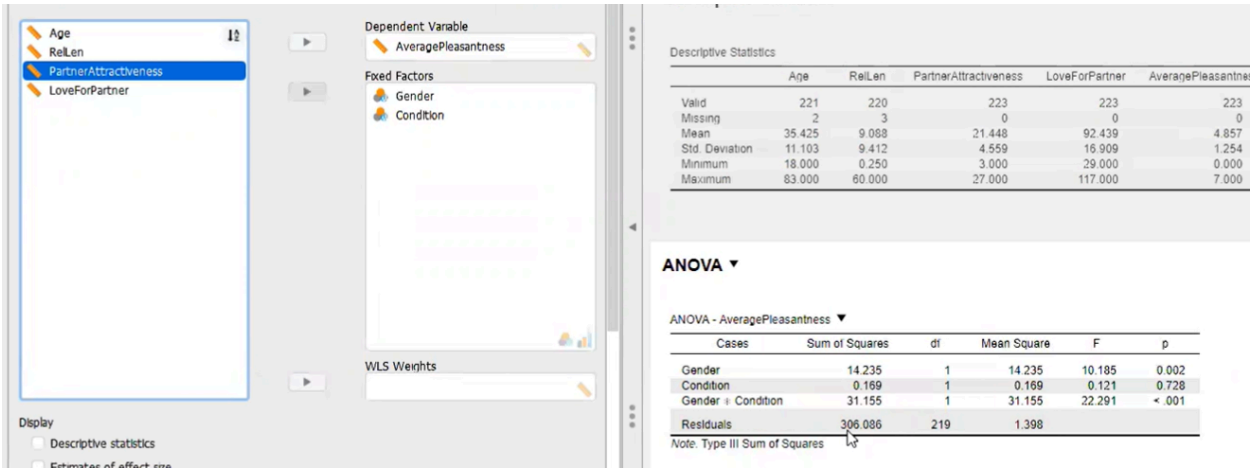
Post Hoc Comparisons - Friends ▼

		Mean Difference	SE	t	P_{tukey}
102	302	−1.062	0.294	−3.613	0.004
	502	−0.745	0.310	−2.402	0.121
	702	−0.590	0.300	−1.966	0.288
	902	−0.174	0.327	−0.531	0.984
302	502	0.317	0.287	1.104	0.804
	702	0.472	0.276	1.708	0.433
	902	0.888	0.306	2.904	0.035
502	702	0.155	0.294	0.528	0.984
	902	0.571	0.321	1.776	0.392
702	902	0.416	0.312	1.335	0.670

Note. P-value adjusted for comparing a family of 5

Two-Way ANOVA

EXP: randomly assigned male and female Ps to one of 2 conds: Nude art and Abstract art and measured how pleasant they found the photo, their feelings of love towards their romantic partners and how attractive they find their partner



*men and women had sig.different pleasantness ratings
*the effect of condition was moderated by gender
*if the interaction is sig. Run post hoc
—----------

Readings:

In experimental data, sometimes Ps may not follow instructions, may fail the manipulation check and it is customary to remove these Ps from the sample.
Caution against this practice, removing Ps from conds mess up with your random assignment (the essence of random assignm is you can not predict whic Ps go to which cond) If these Ps you have removed have a common property (such as being clumsy) then there is no random assignm anymore, you interfered with something systematic.
What to do when Ps do not follow the instructions but we cannot remove them because them we would be interfere with random assignment?

- Taking these data from the sample is common practice, but may interfere with random assignm.
- **Intend to treat**; keep them in the condition they were initially assigned to, even when they do instructions wrong. You analyze them in the group that they are assigned to initially.
- **Pilot test your experiment**: you make sure everyone understands the experiment and follows the instructions.
- Include **attention checks** before assignment into random groups; so that you can remove these Ps who failed in attention check before assigning them into groups.
- **Instrumental variable approach** can be also used.

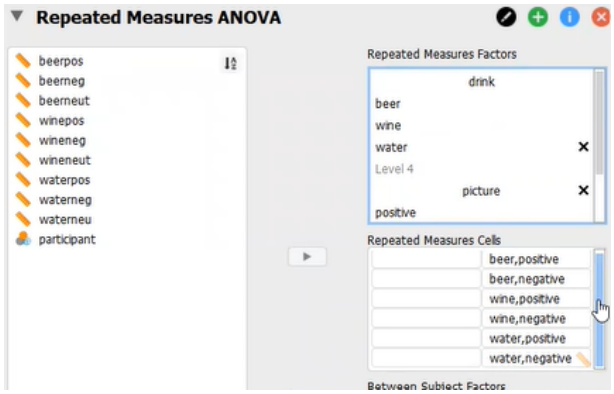
Week#6: ANOVA within/repeated measures Ps

If you have within Ps assessment, you can both estimate within Ps and btw Ps effects.
Difference btw within Ps ANOVA and btw Ps ANOVA:

- we can estimate 2 different sources of residuals; residuals due to btw person differences and due to within Ps differences
- The assumptions underlying the analysis are slightly different from each other. In btw ANOVA; we had mutually exclusive groups consistent of diff individuals. So, one assumption we needed to meet was that the variability among these different ppl in diff groups should be the same. We dont have this assumption in within ANOVA because there are no groups consistent of diff ppl, each and every cond consists of the same inds. (each Ps in your study completed every conds)
- However, we have sphericity assumption to meet in within ANOVA: the correlations btw pairs of repeated measures should be the same to each other. Exp: lets say you measured a Ps on 4 diff conds, the sphericity assumption requires that: correlation btw 1 and 2 should be equal to correlation btw 1 and 3, and that should be equal to correlation btw 1 and 4...and so on. So each and every pair of correlations should be equal to each other.
- Assumption of sphericity is hard to maintain: because of the nature of human beh. The correlation btw 1 and 2 would be naturally different than correlation btw 1 and 5, as ind would change over time. And whenever we violate this assumption we need to make a correction (because violation of assumption may increase the rate of making a type 1 error). When applied correction; it decreases your degrees of freedom which will prevent you from making type 1 error.
- How to correct: we compute a parameter called epsilon (ε). The epsilon will tell you how much you violate from the sphericity assumption. The further away the epsilon is from 1, the greater is the violation. And then we multiply the degrees of freedom by that epsilon value = you will get the corrected degrees of freedom. And we interpret that corrected degrees of freedom.
- The rest is same: you compute sum of squares, degrees of freedom, then divide sum of squares by df to compute your mean squares, then divide the mean squares by mean square residual to estimate the F statistic.

**more demand effects and carryover effects in within Ps designs.

Exp data:
20 Ps who were measured on 9 different occasions to see if Ps attitudes toward diff drinks can be changed by showing them pictures. 3x3 repeated measures experiment; we have 2 IVs with each having 3 conds.: IV1: Drink (beer, wine and water), IV2: Pic (positive, negative, neutral)
DV: attitudes
Ps complete all conditions: within Ps study.



Repeated Measures ANOVA

Within Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p	η^2
drink	2092.344[a]	2[a]	1046.172[a]	5.106[a]	0.011[a]	0.052
Residuals	7785.878	38	204.892			
picture	21628.678[a]	2[a]	10814.339[a]	122.565[a]	< .001[a]	0.535
Residuals	3352.878	38	88.234			
drink ✳ picture	2624.422	4	656.106	17.155	< .001	0.065
Residuals	2906.689	76	38.246			

Note. Type III Sum of Squares
[a] Mauchly's test of sphericity indicates that the assumption of sphericity is violated (p < .05).

Between Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p
Residuals	1920.106	19	101.058		

Note. Type III Sum of Squares

Descriptives

Descriptives

drink	picture	Mean	SD	N
beer	negative	4.450	17.304	20
	neutral	10.000	10.296	20
	positive	21.050	13.008	20
water	negative	−9.200	6.802	20

Drink and picture as a main effect found sig. And our interaction is sig.
We can see in the table that sphericity is violated: which means that the pairs of correlations are not the same. Now we should go back and correct this assumption: click assumption check→sphericity check

Assumption Checks

Test of Sphericity

	Mauchly's W	Approx. χ^2	dfSphericity	p-value	Greenhouse-Geisser ε	Huynh-Feldt ε	Lower Bound ε
drink	0.267	23.753	2	< .001	0.577	0.591	0.500
picture	0.662	7.422	2	0.024	0.747	0.797	0.500
drink ✳ picture	0.595	9.041	9	0.436	0.798	0.979	0.250

Here if the p value is less than .05 it means that at least 1 correlation is different than others (not equal) For 2 main effects we see that sphericity assumption is vio7lated. Typically we use greenhouse-geisser epsilon value for correction. Now we should interpret our output regarding greenhouse-geisser correction.

Within Subjects Effects

Cases	Sphericity Correction	Sum of Squares	df	Mean Square	F	p	η^2
drink	None	2092.344[a]	2.000[a]	1046.172[a]	5.106[a]	0.011[a]	0.052
	Greenhouse-Geisser	2092.344	1.154	1812.764	5.106	0.030	0.052
	Huynh-Feldt	2092.344	1.181	1770.939	5.106	0.029	0.052
Residuals	None	7785.878	38.000	204.892			
	Greenhouse-Geisser	7785.878	21.930	355.028			
	Huynh-Feldt	7785.878	22.448	346.836			
picture	None	21628.678[a]	2.000[a]	10814.339[a]	122.565[a]	< .001[a]	0.535
	Greenhouse-Geisser	21628.678	1.495	14468.490	122.565	< .001	0.535
	Huynh-Feldt	21628.678	1.594	13571.496	122.565	< .001	0.535
Residuals	None	3352.878	38.000	88.234			
	Greenhouse-Geisser	3352.878	28.403	118.048			
	Huynh-Feldt	3352.878	30.280	110.729			
drink ✳ picture	None	2624.422	4.000	656.106	17.155	< .001	0.065
	Greenhouse-Geisser	2624.422	3.194	821.778	17.155	< .001	0.065
	Huynh-Feldt	2624.422	3.914	670.462	17.155	< .001	0.065
Residuals	None	2906.689	76.000	38.246			
	Greenhouse-Geisser	2906.689	60.678	47.903			
	Huynh-Feldt	2906.689	74.373	39.083			

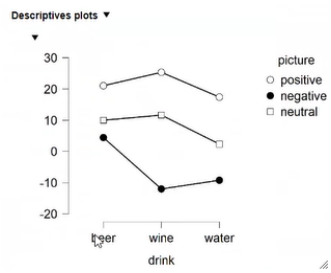
Note. Type III Sum of Squares
[a] Mauchly's test of sphericity indicates that the assumption of sphericity is violated (p < .05).

Correction reflected on the increase of p values, but still we have sig. Main and interaction effects.
So, now we can observe post hoc tests:

Post Hoc Comparisons - drink ✻ picture

		Mean Difference	SE	t	P_{bonf}
beer, positive	wine, positive	−4.300	3.063	−1.404	1.000
	water, positive	3.650	3.063	1.192	1.000
	beer, negative	16.600	2.343	7.084	< .001
	wine, negative	33.050	3.324	9.944	< .001
	water, negative	30.250	3.324	9.102	< .001
	beer, neutral	11.050	2.343	4.716	< .001
	wine, neutral	9.400	3.324	2.828	0.211
	water, neutral	18.700	3.324	5.627	< .001
wine, positive	water, positive	7.950	3.063	2.596	0.417
	beer, negative	20.900	3.324	6.289	< .001
	wine, negative	37.350	2.343	15.939	< .001
	water, negative	34.550	3.324	10.396	< .001
	beer, neutral	15.350	3.324	4.619	< .001
	wine, neutral	13.700	2.343	5.847	< .001
	water, neutral	23.000	3.324	6.920	< .001
water, positive	beer, negative	12.950	3.324	3.896	0.007
	wine, negative	29.400	3.324	8.846	< .001
	water, negative	26.600	2.343	11.352	< .001
	beer, neutral	7.400	3.324	2.227	1.000
	wine, neutral	5.750	3.324	1.730	1.000
	water, neutral	15.050	2.343	6.423	< .001
beer, negative	wine, negative	16.450	3.063	5.371	< .001
	water, negative	13.650	3.063	4.457	0.001
	beer, neutral	−5.550	2.343	−2.368	0.715
	wine, neutral	−7.200	3.324	−2.166	1.000
	water, neutral	2.100	3.324	0.632	1.000
wine, negative	water, negative	−2.800	3.063	−0.914	1.000
	beer, neutral	−22.000	3.324	−6.620	< .001
	wine, neutral	−23.650	2.343	−10.093	< .001
	water, neutral	−14.350	3.324	−4.318	0.002
water, negative	beer, neutral	−19.200	3.324	−5.777	< .001

We have a lot of results, but we dont have to go over each of them. Your theory usually tells you what the critical comparisons are. And we can run just those comparisons.
Also look at the descriptive plots:



Article: Calibrate your confidence in research findings: A tutorial on improving research methods and practices- Frost et al. (2020)

-statistical power: likelihood that a statistical test will correctly detect a true effect if it exists (i.e., the likelihood that if you are testing a real effect, your test statistic will be sig.)
-type 1 error:
(i.e., the likelihood that if you are testing a null effect, your test statistic will be sig.)
-both of these influence the positive predictive value of a finding, or how likely it is that a sig. result reflects a true effect in the population.
-the statistical power of the researcher's studies determined how many of the true effects will be detected as significant.
-if the study is powered at 40%; 4/10 studies will correctly detect a sig. Effect and 6/10 studies will fail to detect the effect that is in fact present in the population.
-if the study is powered at 90%; 9/10 studies will correctly detect the sig. effect and only 1/!0 will miss it.
-if the studies have a type 1 error rate of 30%; 3/10 null effects will be falsely detected as sig., and the other 7/10 will be correctly identified as non-sig.
-the positive predictive value of a study is low when power is low: the likelihood that a sig. result in this pool of sig. results reflects a true effect is 57%.
-the positive predictive value of a study is higher when power is higher: the likelihood that a sig. result in this pool of sig. results reflect a true effect is 75%.

- So, if we are to understand how much to trust a sig. result, we want to be able to gauge both the likely statistical power and the likely type 1 error rate of the study.

Ways to increase power without affecting type 1 error rate:

- Increasing sample size, improving the reliability of a measure

How to assess Statistical Power?

- Have the approximate sample size that is required to detect various effects.
- Develop your sense of how big such effects really are (Cohen's d: medium effect size at d = .50 to "represent an effect likely to be visible to the naked eye of a careful observer", A large effect size of d = .80 is typically an effect that even a casual observer would notice)
- If you are planning your own study; conduct a priori power analysis to decide how many Ps you need in order to achieve your desired level of power.
- Conserve resources by sequential analysis.(allow you to select a priori the largest sample size you are willing to collect if necessary as well as middle points when you would stop data collection earlier if you could.) Sequential analysis are planned ahead of time, before looking at your results, and preserve a max. type 1 error rate of %5. In contrast, if you check your data multiple times without using a formal sequential analysis, type 1 error rate inflates. You have to decide how many times you want to check your data before reaching your final sample size ahead of time. Each additional check will reduce power by a small amount in exchange for the possibility of stopping early and conserving resources.
- Consider multiple approaches to boosting power; when possible and appropriate, making a manipulation within-subjects instead of between-subjects can dramatically boost the power of an experiment. Also, you can increase power by strengthening a manipulation and by improving the reliability of your measures. Also, select ahead of time a planned covariate that correlates strongly with the dependent measure of interest. Also, consider collaborating across multiple labs and aggregating the results.

Strategies for an Existing study:

- Conduct a sensitivity analysis (you input the actual sample size used in the study of interest (e.g., N = 60), the statistical test (e.g., a t test comparing two between-subjects conditions), and a given level of power (e.g., 80%), and the program tells you the effect size the study could detect with this level of power (e.g., d = .74).
- The aim here is to provide a good sense of the range of effect sizes that an existing study was adequately powered to detect.
- Avoid post hoc or observed power analysis

How to assess Type 1 error rates?

- First, understand how data-dependent decisions inflate the type 1 error rate:
    - Perhaps most importantly, the type I error rate can inflate – often by an unknown amount – when the various decisions that a researcher makes about how to

construct their dataset and analyze their results are informed in some way by the data themselves. (such decisions are called data-dependent)

- The main problem arises because there are multiple posarcher could or does run to test their research questisible tests that a reseon. When the decision about which test to run and report is informed by knowledge of the dataset in question, the type 1 error rate starts to inflate.

- Clearly distinguish btw data-dependent and data-independent analysis with a pre-analysis plan: pre analysis plan involves selecting and writing down ahead of time the various researcher decisions that will need to be made about how to construct and analyze a dataset, before looking at the data. Also, it is important not to follow the plan blindly, and always check whether the assumptions of a statistical test are met given the data.
- Distinguish btw different varieties of preregistration and their respective functions

Table 4. Different definitions of preregistration and their intended purpose

Definition	Goal
Preanalysis plan	Distinguish data-independent vs. data-dependent analyses; constraining unintended type I error inflation
Write down as much information as you can about your study before you conduct it	Transparency: Someone else can check what you said you planned ahead of time against what you actually wrote down
Record your theoretical predictions	Theory falsification
Record your intuitive predictions	Figure out how good you are personally at guessing a study's outcome
Record the existence of your study in a centralized, searchable repository	Combat publication bias
Registered report	All of the above plus reviewer objectivity (reviewers can evaluate the methods and planned analyses without being biased by the whether the results fit their intuitions or theories)

Conclusion:

Table 5. Summary of recommendations

When planning a study:
1. Assess and maximize statistical power – Conduct an a priori power analysis. – Identify the biggest sample size worth collecting. – Use sequential analysis to conserve resources when possible. – Consider multiple approaches to boosting power.
2. Avoid unintended or invisible type I error inflation – Clearly distinguish between data-dependent and data-independent analyses with a preanalysis plan. – Distinguish between different varieties of preregistration and their respective functions. If your goal is to avoid type I error inflation, preregister a preanalysis plan that clearly constrains researcher degrees of freedom for any planned, data-independent analyses.
When evaluating a study that has already been conducted:
1. Assess the likely power of a given statistical test – Conduct a sensitivity analysis to assess power to detect a range of effect sizes. – *Do not* calculate "post-hoc" or "observed" power.
2. Assess the likely type I error rate – Understand how data-dependent decisions inflate the type I error rate. – Look for a preanalysis plan. Is it clear which analyses were data-independent? Are researcher decisions well constrained or flexible? Does the preanalysis plan account for multiple comparisons and are the analyses appropriate?

Week#7: Regression

The essence of regression analysis: predicting an outcome. You have set of predictors/determinants, and using those information you are trying to predicting an outcome.

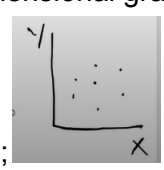
Now we will work on covariances. Since now, we are trying to predict an outcome from a set of other variables, we need to estimate their associations, covariances, whether variability in one construct/variable/factor is associated with variability in another.
The basic metric describing the association between 2 variables is called covariance.
Covariances are the building blocks of regression.
(variance is the variability of a single construct; covariance describes the association between variabilities in 2 construct)

The variance formula we used so far:

$$s^2 = \frac{\sum (X - \bar{X})^2}{N-1}$$

$$\text{cov}(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1} \qquad -\infty \leq \text{cov}(x, y) \leq \infty$$

The covariance formula:
When you are testing covariance btw 2 variables, it means that you have collected data on both x and y. So you can describe each Ps on a 2 dimensional graph (x and y).

If each Ps can be denoted on this space like this; 

the formula is that; for each Ps' score you can subtract their x value from the mean of x, subtract their y value from the mean of y, multiply these two, sum all multiplications and then dive by the degrees of freedom (n-1).
This will give you: how much these 2 variables (x and y) go together.
Using this associations we build models that predict an outcome from one or more predictors.

!since covariance has no range (-∞, +∞), it depends on the row values on which you measure your constructs. lets assume we are looking at the association between stress and depression, and lets say you measured stress on a scale 1-10 and measure depression through an interview by scores 0-10. And then you compute a covariance. Suppose that other researcher do the exact same study, but their stress measure ranges from 1-3 and their depression measure ranges from 0-5. One problem is that because of the formula, the covariance value would be highly dependent on the scale of measurement, even if the effect you detect is the same across 2 studies.

So we need a standardized index to compare these 2 covariance values.--> correlation coefficient (r) is a standardized way of describing an association. The range is limited (-1 meaning a perfect negative association, +1 meaning perfect positive association)

A standardized score is basically: Z= (x- $\bar{x}$)/SD
And here the idea is the same:

$$r_{xy} = \frac{\text{cov}(x, y)}{s_x s_y} \qquad -1 \le r \le 1$$

the formula of correlation coefficient; dividing the covariance by 2 variables' standard deviations.

<u>Simple regression model:</u>

$$Y = a + b(X) + e$$

$$b = \frac{\text{cov}(X, Y)}{\text{var}(X)}$$

$$a = \bar{Y} - b\bar{X}$$

y=outcome variable, X=predictor, a=intercept, b=slope (the covariance in terms of the variance of x, meaning that how much y increases if you increase x by 1 unite), e=residuals

One measure of effect size: the slope.
The greater the slope, the stronger the association. But it is (b value) not a standardized unit, because the slope/b value will change depending on the scale of measurement.

How to standardized: $b = r \frac{S_y}{S_x}$
Typically we report the unstandardized effect size and its confidence interval.
!slope/b value is basically a t-distribution. If you divide the slope by Stanadat error you will get a t statistic (t= b/ SE), then you can check whether it is sig. or not.

What if I have more than one predictor?: if ı have 3 predictors, my model would look like;
y= a+ B1(X1) +B2(X2) +B3(X3) + e
Each predictor has its own effect: The association btw y and x1 is described by B1. If increase X1 by 1 unit, y increases by B1 units.

A second source of effect size: how much of a good job your model is doing as a whole, when include all these 3 predictors, how much variability I can explain in the outcome behavior (y)?

So we have a predictor level effect size: b coefficient/slope
And we have a model level effect size that would represent the whole model.

How to get a model level effect size?
When we have many data points, it means that there can be many lines that can be drawn to explain the model. The line that best describes is the line that minimizes the sum of square of distance btw actual points and the predicted points. (Ordinary least squares regression model)

The models success/the goodness of fit: is measured by how much you minimize these distances. How much of the variability in y I can explain by these predictions: that would be the essence of my model level effect size. We call this r^2

$$s_Y^2 = \frac{\sum (Y_i - \bar{Y})^2}{n-1}$$

The total variability in your outcome:
Instead of using the actual score, if for each Ps you use the predicted value by the model you get the variance explained by the model: $s_{\hat{Y}}^2 = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{n-1}$

And if you divide the variance explained by the total variance you get r^2 value: $R^2 = \frac{s_{\hat{Y}}^2}{s_Y^2}$
(this is at essence an F test; we are dividing the sums of squares basically, so the distribution of r^2 is an F statistic, and depending on that F distribution we will determine whether the r^2 value is sig.ly greater than 0 or not)

***If you have only 1 variable, the predictor level effect size and model level effect size will be exactly the same. But things will change as we add more variables.

Example: predicting album sales (outcome variable) from a predictor (attractiveness of the band 1-10).

Linear Regression

Model Summary - sales

Model	R	R^2	Adjusted R^2	RMSE
H_0	0.000	0.000	0.000	80.699
H_1	0.326	0.106	0.102	76.480

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
H_1	Regression	137822.446	1	137822.446	23.563	< .001
	Residual	1.158e+6	198	5849.139		
	Total	1.296e+6	199			

Note. The intercept model is omitted, as no meaningful information can be shown.

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p	95% CI Lower	95% CI Upper
H_0	(Intercept)	193.200	5.706		33.857	< .001	181.947	204.453
H_1	(Intercept)	65.510	26.855		2.439	0.016	12.550	118.469
	attract	18.861	3.886	0.326	4.854	< .001	11.199	26.524

H0 (null hypoth) and H1 (alternative hypo). In H0 our model is an intercept only model meaning that we only have an intercept we dont have any predictors. In H1 we have 1 predictor.

H1 r^2 is roughly .11, 11% of the variance is associated with the band attractiveness.
The ANOVA table says whether this 11% is statistically significant or not. Here F statistic is sig., meaning that 11% is sig.

Looking at the coefficient table: we have intercept and slope. Slope is 18.86, meaning that for 1 unit increase in band's attractiveness, there is a 18.86 more album sales.
If you want to achieve a standardized estimate: just multiply 18.86 with SD of attractiveness (predictor) and dive that number by SD of the sales (outcome measure); in this case you will get .32. → in the simple regression framework this is the same as a correlation coefficient. You can interpret this standardized estimate of slope as correlation coefficient.

Whether this slope value is sig.ly different than 0 or not?--> look at t value and p value.

The intercept value refers to the average value of outcome when the predictor is 0; so 18.86 albums sold if the band's attractiveness was 0.
Whether 0 is a meaningful value? Our attractiveness range is 1-10, so 0 is not meaningful point in our measure anyway. If 0 is not meaningful on your variable then intercept is meaningless, do not interpret it.

Can we do smt to make the intercept a meaningful value? It is called **centring**. (transforming your variables so that they have a meaningful 0 value) (we use mean subtraction to do centring)

When intercept is something meaningful?: It corresponds to the … group in the data.

Week#8: Regression approach to ANOVA

We analyzed this example via independent sample t-test, now we will analyze it with regression
Data: Ps from hogwards, invisible cloaks, DV: num of mischief behav they engage

T test:

Independent Samples T-Test

	t	df	p
Mischief	−1.713	22	0.101

Note. Student's t-test.

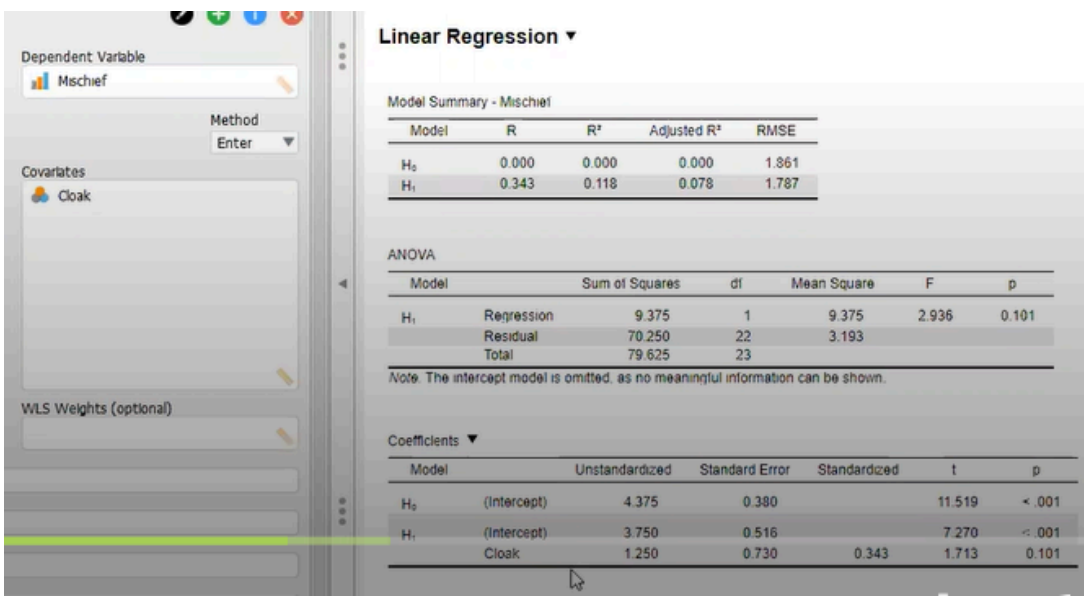
Descriptives

Group Descriptives

	Group	N	Mean	SD	SE
Mischief	0	12	3.750	1.913	0.552
	1	12	5.000	1.651	0.477

Control cond misc. Beh: 3.75, Invisible cloak cond, misch behv.: 5.
Whether this 5 is statistically different than 3.75? Looking at out t-test table: t=-1.73, sı our critical point would be somewhere around 1.96, and p value not sig.
No evidence that invisibility cloak increasing misch. Behav.

Linear Regression:



Our Iv in this case is not continuous. IV is dichotomous (can take only one value; 0 or 1). Given that 0 is a meaningful value in this case, now the intercept will be meaningful. Intercept will tell us mean num of misch beh when the IV is 0, it means the control cond.

In experimental data if you dummy code your IV (0,1), your intercept will tell you the mean value of DV in the control cond.

And the slope, the value of unstandardized slope, the effect of cloak, will tell you the diff btw control and experimental conds.(what happens to the misch. Behav when I increase the cloak 1 unite)
Which means; if I go from 0 to 1 (going from control to experimental cond) how much misch behav change. = they change 1.71 (diff btw control-experim group means)

What we do in btw Ps ANOVA, we can do it in Linear Regression as well

Data: facebook friends, num of friends was manipulated (profiles showed 102, 302, 502, 702 and 902 friends) DV: how likeable the profile is.
We are trying to see whether increasing the num of friends has any effect on likability.
Using a baseline (102 friends as baseline): Compare each of these profiles with the profile where the ind has 102 friends.

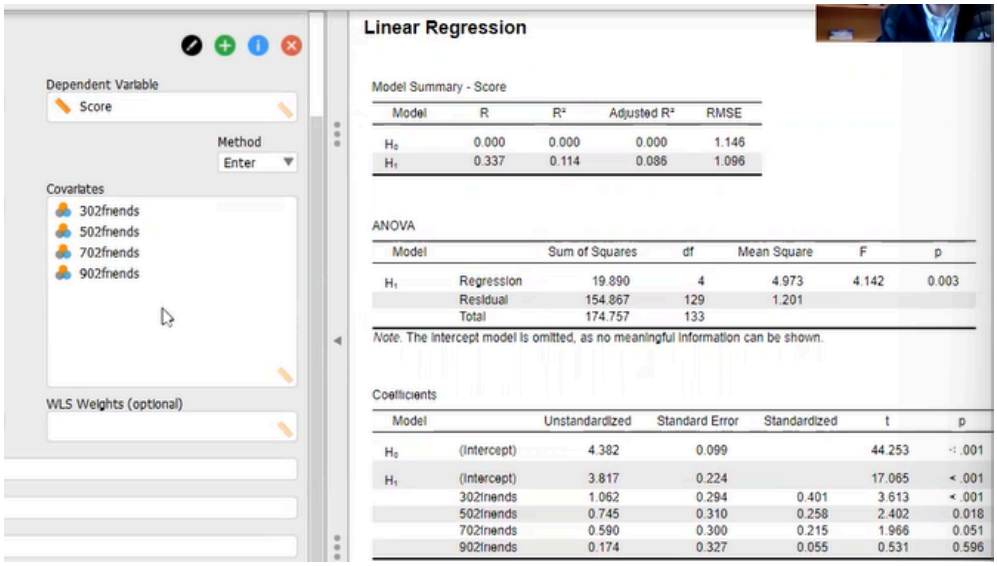
When comparing multiple groups, running multiple comparisons in regression; you need to create k-1 new variable (k is the num of conds in your grouping variable)
Her k=5, we have 5 conds; so if ı create 4 new dichotomous variables I can express this grouping variable in a way to be able to use it in the regression framework.

In these 4 new variables, your baseline variable (comparison point/cond) will always be 0. And the remaining cond will get the value 1 one of the variables.

Friends	Participant	Score	302friends	502friends	702friends	902friends
302	51	5	1	0	0	0
302	52	5.2	1	0	0	0
302	53	4.2	1	0	0	0
302	54	5	1	0	0	0
302	55	5.8	1	0	0	0
302	56	5.6	1	0	0	0
302	57	3.8	Double click to edit data	0	0	0
502	58	4.6	0	1	0	0
502	59	4	0	1	0	0
502	60	4.8	0	1	0	0
502	61	3	0	1	0	0
502	62	2	0	1	0	0
502	63	5.8	0	1	0	0
502	64	5.6	0	1	0	0
502	65	4.4	0	1	0	0
502	66	4.4	0	1	0	0
502	67	5.6	0	1	0	0

Exp: Here, “302 friends created group” gets the value of 1 only when the person is in the 302 cond group; for all other cases it gets the value 0.

Across these 4 new variables, my baseline cond always got the value 0.
When you re-express your grouping variables as a series of dichotomous variables, you can run regression analyses.



The intercept is the value of outcome when all the predictors are 0. If all of these variables are 0, I am talking about my baseline cond. (because the only cond that gets the value of 0 across all groups is my baseline cond). So the intercept tells you the average score in the 102 friends cond. And each comparison basically tells you the mean diff btw that particular cond and baseline.
(for exp: for 302 friends cond; unstandardized score (1.06) gives you the diff btw 302 friends cond and 102 friends cond)
This way we can run any btw Ps tests using regression.
***We can run every analysis we have seen so far (except repeated measures) on regression as well. Linear regression can only handle btw Ps tests.
When we have repeated measures; there is a regression for that called mixed effects model.
—----------

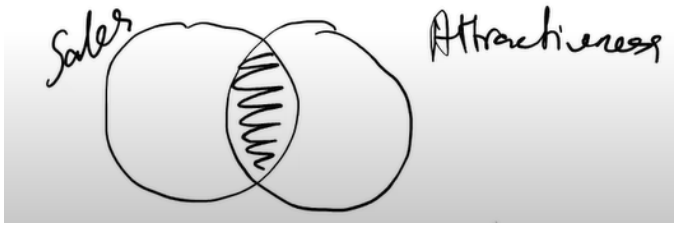
Week#9: Multiple Regression

Multiple regression framework: now we can test associations btw an outcome and multiple predictors (which can be dichotomous, continuous, categorical, etc.), can test interaction effects or mediation hypothesis.
How to fit linear models with multiple predictors?

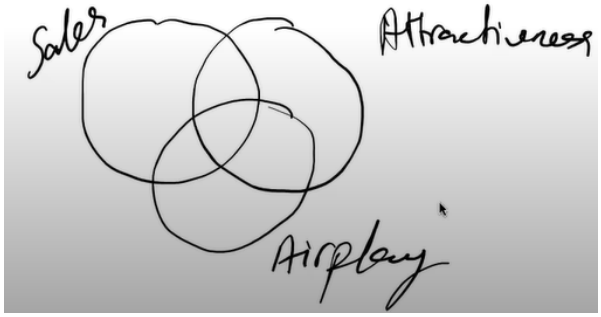
Exp: album sales of bands, sales: DV(outcome variable), predictors: attractiveness of the band (predictor1), money spend (predictor2) and airplay number (predictor3)

The circle represents: total variance of sales, given that our model will explain variability/variance in sales

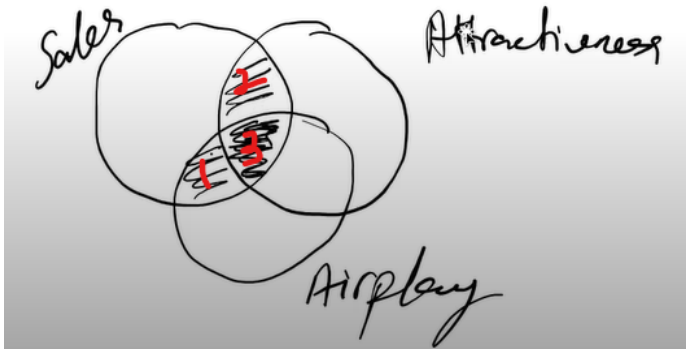
Adding out predictor (attractiveness) we first looked in the linear regression model, this shows how much variance attractiveness explained:



Now we will include another predictor (airplay) in our model:
Our predictors are not only associated with the outcome variable, but there might also be a correlation btw the predictors themselves.



So how multiple regression model handels this issue?:
The variance these 2 predictors explain are these areas:



Area 1: uniquely explained by airplay predictor
Area2: uniquely explained by attractiveness predictor
Area3: explained by both predictors
What multiple regression model does?: it estimates only unique effects, so area 3 is taken out.

In a multiple regression model, the unstandardized effect of each predictor refers to the their unique contribution in explaining variability in outcome.

- Whenever we are constructing a multiple reg model: the effect of each predictor is unique, meaning that: that effect refers to the effect of predictor on the outcome controlling for all other predictors. So these are not main effects as in the ANOVA, rather these are the adjusted/conditional effects
- In the linear model; the correlation btw the predictor and outcome was exactly the same as the slope (b coefficient = r coefficient). In multiple model: the correlation coefficient will refer to the explained variance in total for our predictors; but the regression coefficient will only estimate the unique effects. So correlation and regression coefficients start to diverge, so we should report both of them.
- Assume that airplay is variable that is very highly correlated with attractiveness

in that case; there is a large overlapping variance btw these predictors. The regression analysis will take this out, resulting in very small values for the slope (b coefficient).
!!!Whenever you have highly correlated predictors, linear regression is not the best way to go, it wont' give you accurate estimates. (this is called; multicollinearity)
!!if there is multicollinearity btw predictors, you wont get meaningful regression coefficients.
*How to check multicollinearity: check the correlation table btw our predictors, and see if there is high correlation (like .70 or .80)

- The model equation: $Sales = B_0 + B_1(att) + B_2(airplay) + B_3(adv) + e$

(To explain variability: B0: intercept value, B1: effect of attractiv., B2: effect of airplay, B3:effect of money, e: residual value)

- The multiple model assumes that these residuals (e) are normally distributed.

We have an actual sales value we measured and a predicted sales value that our model produces; the difference is the residual value (the information that ı couldnt explain by my model) So, ı will have a residual value for each and every data points. These residuals are assumed to be normally distributed. So, we check to make sure that linear regression is the appropriate approach to tackle our data.

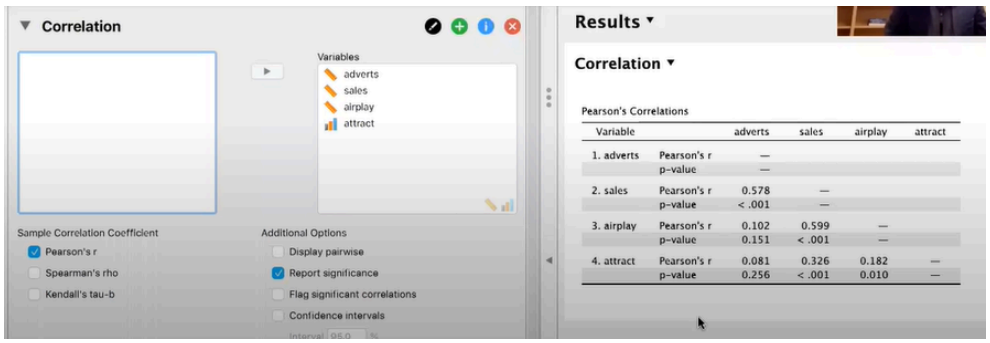
- We need to check: homoscedasticity (is the equivalent of homogeneity of variance assumption). Here, we want to make sure that whether the variability in the residuals is roughly the same across the levels of the outcome. Way to check: obtain a scatterplot charting residuals against predicted values, if there is no clear pattern it means that the variability in residuals is homogeneously distributed. If you see a cone shaped, it means

that the variability is lower or higher for some of the predicted values, meaning that variability is not homoscedastic, the association is not linear (linear reg might not be a good idea then).

- If these assumptions are met; multiple reg framework can be a way to explain the variability in our outcome variables.

JASP exmp:

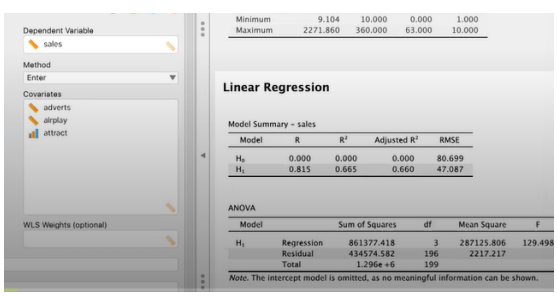
First check the correlations for multicollinearity:



Descriptive Statistics

	adverts	sales	airplay	attract
Valid	200	200	200	200
Missing	0	0	0	0
Mean	614.412	193.200	27.500	6.770
Std. Deviation	485.655	80.699	12.270	1.395
Minimum	9.104	10.000	0.000	1.000
Maximum	2271.860	360.000	63.000	10.000

Correlations seem to be modest, so no concern for multicollinearity.



Linear Regression

Model Summary – sales

Model	R	R^2	Adjusted R^2	RMSE
H_0	0.000	0.000	0.000	80.699
H_1	0.815	0.665	0.660	47.087

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
H_1	Regression	861377.418	3	287125.806	129.498	< .001
	Residual	434574.582	196	2217.217		
	Total	1.296e +6	199			

Note. The intercept model is omitted, as no meaningful information can be shown.

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
H_0	(Intercept)	193.200	5.706		33.857	< .001
H_1	(Intercept)	−26.613	17.350		−1.534	0.127
	adverts	0.085	0.007	0.511	12.261	< .001
	airplay	3.367	0.278	0.512	12.123	< .001
	attract	11.086	2.438	0.192	4.548	< .001

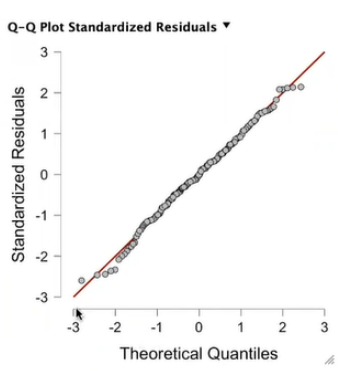
NOTE: Adjusted Rsquare: adjustment corrects for the num of predictors. Any predictor added to the model somewhat increases the Rsquare value, so if you put many predictors Rsquare value will be inflated. Adjusted Rsquare corrects for that.

-our .65 percent (Rsquare value) is sig. (p<.001 in ANOVA table)
-when we analyzed only attractiv. (unique effect controlling for other predictors): 11.086 more sales.
-Similarly the unique effect of airplay: 3.367 more sales; and advertisement effect: .085 more sales.
-we report unstandardized values, p values and CI.
-check normality of residuals:

Residuals Statistics

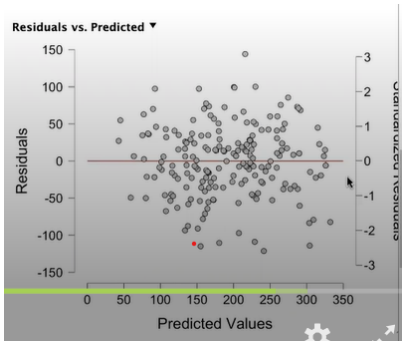
	Minimum	Maximum	Mean	SD	N
Predicted Value	42.906	332.364	193.200	65.792	200
Residual	−121.324	144.132	−1.258e −15	46.731	200
Std. Predicted Value	−2.284	2.115	−1.119e −16	1.000	200
Std. Residual	−2.629	3.093	−2.506e −4	1.003	200

Look at std. Residuals colum; If standardized mean is close to 0 and standardized SD is close to 1: thats an indication of normal distribution for the residuals. Here, we see normal dist.
Or you can check Q-Q plot:



data points line lineratl y, again suggesting that normality is met.

-check homoscedasticity: look at residuals vs. predicted plot:



we can see no cone shaped btw predicted values and residuals. So it is homoscedastic.

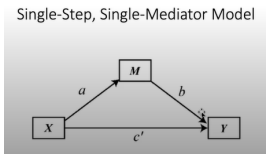
—----------

Reading:
What to do? Potential cases of covariance?
-Standard robustness checks (always run 2 analysis, one your covariance included (adjusted model), one where your covariance excluded (unadjusted model); if it appears in only once, in only one analyses, then that finding is covariate dependent, so robustness in question)
Unadj model results null; do not even run adjusted. If you get sig. in unadjusted and you run and again get sig in adjust then it is a good robust finding. A robust finding appears in both models.
-when we evaluate research as consumers; when reading a paper; there are 2 possible roads; you may be a reader or reviewer. It has diff approaches to the paper. When being a reviewer; you start to evaluate the paper whether the model was appropriate or not. It is hard to maintain this perspective as just a reader.

Week#10: Mediation

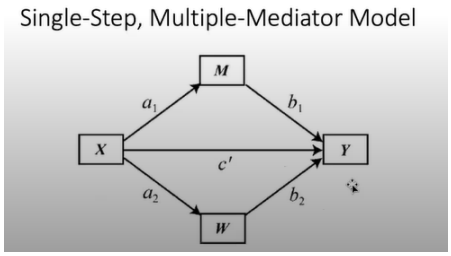
Mediation processes: answer the “why” question, what is the underlying mechanism of an effect.
Exp: association between social bonds/relationsh and well-being. This is a direct association.
But what is the mechanism? For instance researchers found that one mechanism is high quality relationships reduce stress reactivity (which in turn improve well-being).
So here, stress reactivity is the mediator btw quality of relationship and well-being.

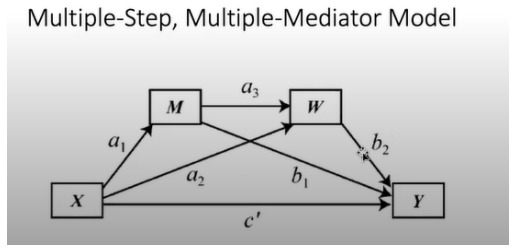
What is moderation? It answers whether the association between one IV and DV depend on the lvls on the other IV. It answers; under which conditions you observe an association btw IV and DV.
Looking at mediation models:
The goal is to estimate an indirect association through the mediator.



The indirect association can occur in the presence or absence of a direct association.
So a predictor can have both direct and indirect association.
How to test indirect association? (there are 2 effects here; a from X(predictor) to M(mediator) and b from M to Y(outcome))
-we will combine these two effects (a and b) to estimate an indirect association.

The model can be extended:





But lets keep it simple:

So our task is to estimate a and b, we can do that in 2 steps using 2 regression models.

The first step would be predicting M from X (the path a) using a simple regression model:

$$M = B_0 + B_1 X + e$$

Here the B1(slope) is basically your a.

Then, run another regression analysis (this time multiple regression)to estimate the b path; here we will have 2 predictors (multiple model) to estimate the outcome from both X and mediator. (so getting the effect of mediator controlling for X; remember the multiple reg.model gives you the unique effects of each predictor)

$$Y = B_0 + B_1 X + B_2 M + e$$

Here, B2 will be b.

Now that I know both the a and b paths, the next step is to estimate the indirect association (it is the multiplication of a and b) => Indirect effect/association = a x b → will tell you how much change occurs in the outcome as a result of one unit change in the predictor through the mediator, so if I change my predictor for 1 unit how much indirect effect will occur through the mediation in my outcome

Whether this indirect association is statistically sig.?
Multiple ways to test this:

-**<u>the Sobel tests</u>**: based on the idea that; compute the sampling distribution of test statistic (in this case it is indirect effect ("a x b")) and then compute the SD of the sampling distribution (which is SE)= thus just divide (a x b)/SE (this will give a known z-distribution on which I can make my null hypoth testing decision)

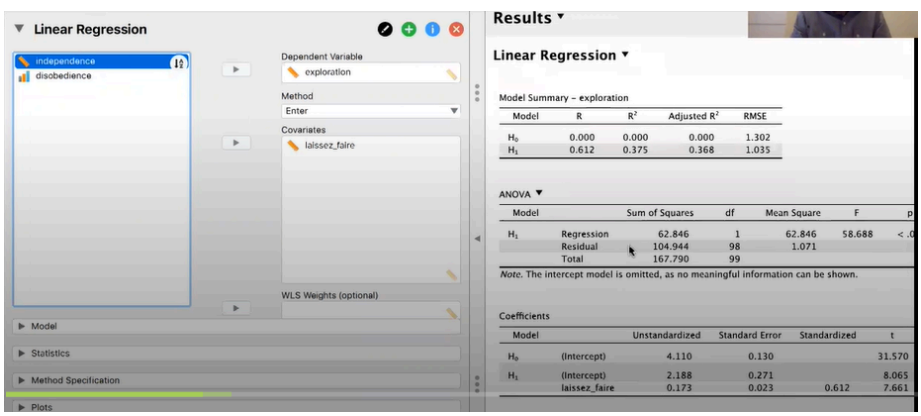
- Limitation: very powerless test, so you need very large sample size to reliably make a judgment about the presence of an indirect association.

Another approach to test indirect association:
- based on simulations or bootstrapping: the idea is that, you create random samples that resemble characteristics of your actual sample. Each of those samples you estimate indirect association. If %95 CI includes 0, it means that indirect association is not sig.

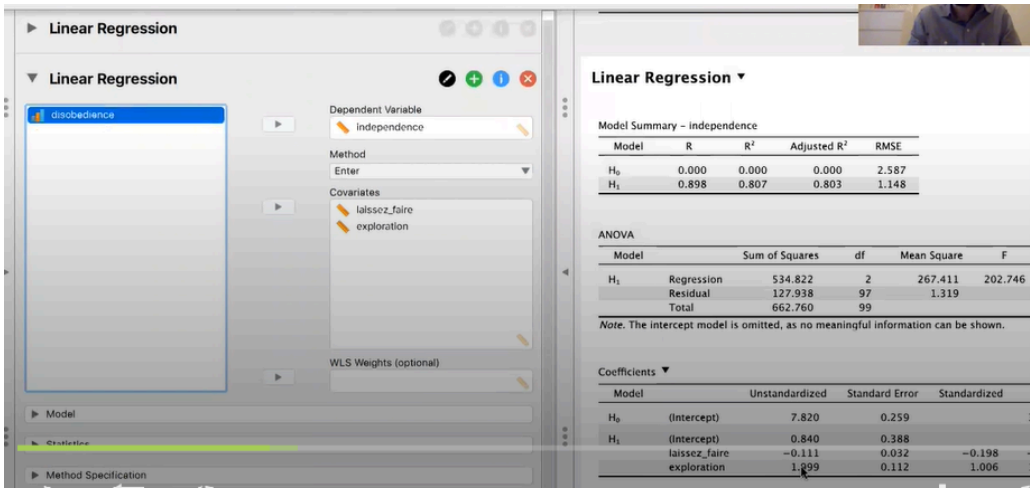
There is one case you might increase Type 1 Error risk: feel free to use simulations or bootstrapping when you are estimating indirect effects but make sure that both the a and b paths are separately, statistically significant. If one of these is non-sig. then the simulation and bootstrapping approach may increase type 1 error rate.

Example data from JASP:
-researchers examine whether there is association btw parenting attitude (laissez faire; IV) and independence (DV) of their children. Child's exploration of the new environm (mediator) and disobedience behav was measured.

Estimating a (laissez faire) and b (exploration) path: running regressions



The unstandardized a path (laissez faire) is 0.17
(laissez faire/parenting behv. (IV) is positively associated with exploration; it predicts greater exploratory behav (M))
For the b path (exploration), include multiple regres.model:



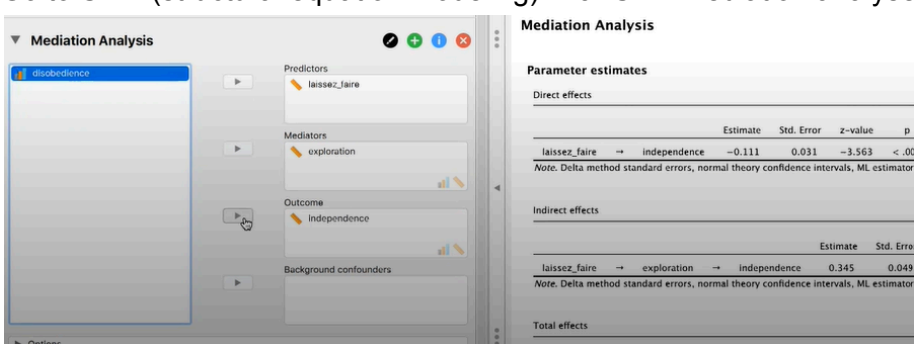
The effect of b (exploration) is 1.99
(greater exploratory behav (M) predicts greater independence (DV))

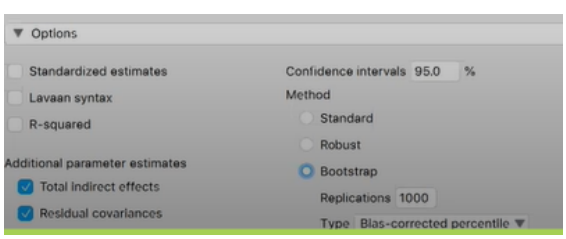
Now test whether there is an indirect association:
-*the joint significance approach: the indirect effect exists if both a and b paths are sig. (here, they are sig.), then we can talk about an indirect association.

!!!!!combine joint significance approach with the simulation/bootstrap approach; to both

- Estimate the indirect effect's sig.
- And also protect for Type 1 error

Whether this indirect association is sig.?
-can use a simulation or bootstrap approach.
Go to SEM (structural equation modeling) in JASP→mediation analyses





Results:

Mediation Analysis ▾

Parameter estimates ▾

Direct effects

			Estimate	Std. Error	z-value	p	95% Confidence Interval Lower	Upper
laissez_faire	→	independence	−0.111	0.031	−3.563	< .001	−0.175	−0.050

Note. Delta method standard errors, bias-corrected percentile bootstrap confidence intervals, ML estimator.

Indirect effects ▾

					Estimate	Std. Error	z-value	p	95% Confidence Interval Lower	Upper
laissez_faire	→	exploration	→	independence	0.345	0.049	7.116	< .001	0.261	0.454

Note. Delta method standard errors, bias-corrected percentile bootstrap confidence intervals, ML estimator.

Total effects

			Estimate	Std. Error	z-value	p	95% Confidence Interval Lower	Upper
laissez_faire	→	independence	0.234	0.051	4.596	< .001	0.154	0.338

Note. Delta method standard errors, bias-corrected percentile bootstrap confidence intervals, ML estimator.

- Direct effect: from predictor to the outcome
- Indirect effect: from the predictor to the mediator to outcome
- The indirect effect estimate is 0.35 (this is the “a x b”)
- The indirect effect estimate driven from the original data, but the CI is driven from sampling distribution (so CI are bootstrap’s CIs)
- We have a sig. Indirect effect from the predictor to the outcome.

Monte Carlo approach: to test whether the indirect association is sig or not. It wont use the actual data, instead it uses the model parameters.
Monte carlo method is based on the assumption that the regression coefficients are normally distributed, the sampling distribution of regression coefficients are normal.
It randomly generates a and b values based on the plausible set of values.
Then look at the CI, if it excludes 0, then it is sig.

Week#11: Moderation

*the idea behind moderation: testing whether the association between a focal predictor and outcome differs across the levels of a 3rd moderator variable. Testing whether the effect of focal predictor on the outcome is modulated by your moderator variable. Whether the average effects differs based on the levels of a 3rd variable.

How to construct a moderation model using regression?
-the model will include outcome (Y), predictor (X) and moderator (Mo).

$$(1)Y = a + b_1(X) + b_2(Mo)$$

In this equation there is no interaction term yet. No term to tell us whether X on Y changes depending on the levels of Mo.

We create a new variable in our dataset; which is the mathematical product of X and the Mo. That term (X.Mo) will be the interaction term. B3 tells us the extent to which the effect of X depends on the moderator.

$$(2)Y = a + b_1(X) + b_2(Mo) + b_3(X \times Mo)$$

Rearrange:

$$(3)Y = a + (b_1 + b_3Mo)(X) + b_2(Mo)$$

The effect of X is denoted by this term; b1+b3.Mo
So, b3 is the extent to which the effect of X is dependent on the Mo. If b3 effect is large it means that the effect of X changes to a great degree based on the Mo. If that b3 effect is small, it means that not much is provided by the Mo, its really the b1 that is driving the effect.

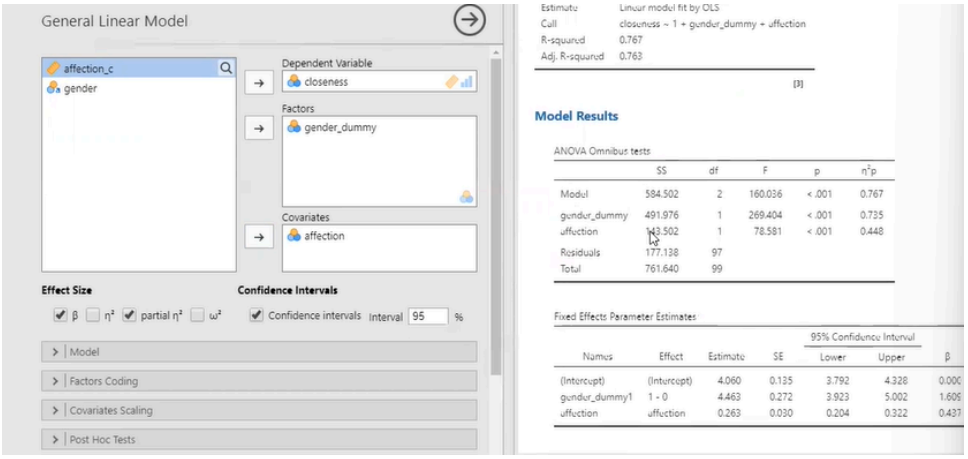
So b3 effect is the interaction effect and sig. of it tells you whether the effect of X on Y meaningfully depend on the levels of Mo.

NOTE!!!: when testing for interactions, always keep the main effects in the model. With the main effects, you are able to tell that interaction effect gives you info about above and beyond the main effects. Because remember the rule no.1 of the regression: the effect of each variable is adjusted for the other variables, so controlling for the main effects, there is additional info provided by the interaction effect (b3). Without the main effects you wont be sure whether the b3 is just a combination of main effects.

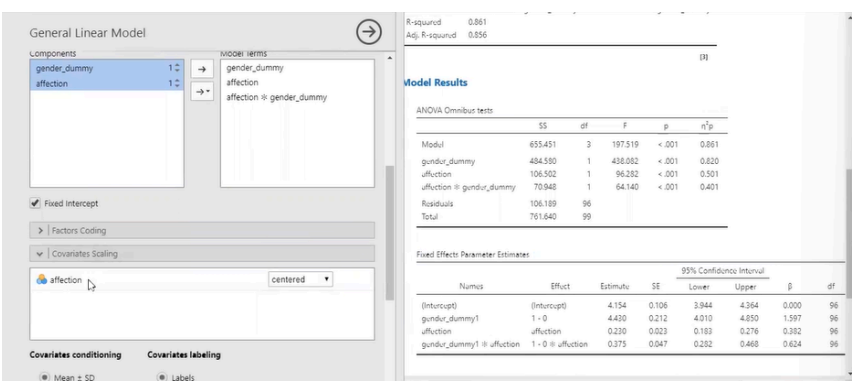
NOTE: 0 should be a meaningful variable. Prior to the analyses make sure that 0 is meaningful , create meaningful 0 value on all continuous variables. HOW?: centering (subtracting the mean value from all observations)

JAMOVI EXP:
Data: how ppl get bonded to their pets. And whether gender of the pet moderates this association. Hypo: the more affectionate behavs towards the pet predict more closeness btw pet and owner. Predictor: affection behv. Outcome: closeness. Mo: gender of the animal.

Jamovi→Linear models→general linear models



(jamovi both calculates the ANOVA and regression analyses for you, can use either of them)
Here we have affection and gender in our model, but not have an interaction varaibles yet.
Go to Model→Components: select your variables→All 2 way (it will multiply the moderator and the focal predictor variables and include the interaction term in the model)



Here, a sig. main effect of gender (Mo). As we go from 0 to 1 (boy to girl), closeness increase.
Interpreting interaction:
If the interaction effect is sig. Our next task is to see the pattern of that interaction. We will test the association between affection and closeness for male and female pets separately.
Simple effects→ set gender as moderator and affection as simple effects varaible:

Simple Effects

Simple effects of affection : Omnibus Tests

Moderator levels				
gender_dummy	F	Num df	Den df	p
0	1.384	1.000	96.000	0.242
1	192.489	1.000	96.000	< .001

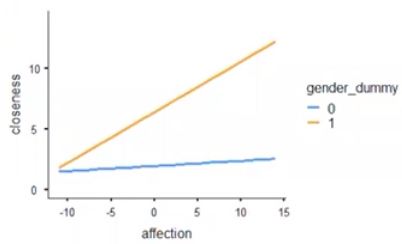
Simple effects of affection : Parameter estimates

Moderator levels			95% Confidence Interval				
gender_dummy	Estimate	SE	Lower	Upper	df	t	p
0	0.042	0.036	−0.029	0.114	96.000	1.177	0.242
1	0.418	0.030	0.358	0.477	96.000	13.874	< .001

Here we see that there's not much of an effect btw affection and closeness when the pet is male (owners of male pets, their affectionate behv does not effect their closeness); however, there is a sig. effect for female pets (the affectionate behv does effect closeness for the female pets)

We can also see this on the Plot:

Plots



What happens if your moderator was continuous? Tricky as continuous variables have infinite many levels. How to look for a moderation effect when there is no fixed set of levels?
2 approaches:
1)pick a point approach:

One SD above the mean and one SD below the mean