Statistics in experimental research

Session 3

Francesca Delogu
delogu@coli.uni-saarland.de

Overview today

- ANOVAs
  - One-way ANOVA
  - ANOVAs with more than one factor
- Recap of Hypothesis testing
- Which-test-to-use-when
- Software & books
- How to set up a psycholinguistic experiment

The independant variable

- Our example: coffee/ no coffee
- Can have more than two levels
  - coffee/ tea/ water
    - Compare 3 groups!
- There could be more than one independent variable!
  - coffee/ no coffee
  - enough sleep / sleep deprivation
    - Compare 4 groups!

Inflation of α

- If the independent variable has 3 levels
  - You would have to perform 3 t tests (for all possible pairs)

- Your chance of making a Type I error (detecting an effect when there is none) is $1-(1-\alpha)^3$, (15%, for α=.05)

- The easiest solution: Bonferroni correction
  - just divide your α by the number of comparisons you perform
  - the overall chance of making a Type I error remains α

- Bonferroni is very conservative (higher chance of Type II error) and not all comparisons may be relevant

One-way ANOVA

- Alternative strategy
  - First test for an overall effect of the variable
  - Only test the relevant pairs

- Use analysis of variance (ANOVA)

How ANOVA works

ANOVA (ANalysis Of Variance) measures two sources of variation in the data and compares their relative size:

- variation BETWEEN groups
  - for each data value it looks at the difference between its group mean and the overall mean
- variation WITHIN groups
  - for each data value it looks at the difference between that value and mean of its group

Statistical hypotheses

$$H_0: \mu_1 = \mu_2 = \mu_3 \ldots = \mu_k$$

$$H_1: \textit{At least one mean is different}$$

- Notice that ANOVA tests only for an effect of the factor, but does not tell you in which direction or between which groups
  - The kind of drink might have an effect but you don't know whether the difference between tea and coffee is significant

Statistical hypotheses

$$H_0: \mu_1 = \mu_2 = \mu_3 \ldots = \mu_k$$

$$H_1: \textit{At least one mean is different}$$

- Instead of comparing means, ANOVA compares the variance **between groups** with the variance **within groups**
- If the independent variable has an effect, the variance between groups should be larger than the variance within groups
- The test statistics is a ratio between the two sources of variance and is called F

How ANOVA works

k

n

Coffee	Tee	Water
3	5	5
2	3	6
1	4	7
$\overline{X}_1 = 2$	$\overline{X}_2 = 4$	$\overline{X}_3 = 6$

$$\overline{\overline{X}} = \frac{3+2+1+5+4+3+5+6+7}{9} = \mathbf{4}$$

$$SS_{tot} = \sum_i \sum_j (x_{ij} - \overline{\overline{X}})^2$$
$$= (3-4)^2 + (2-4)^2 + (1-4)^2$$
$$+ (5-4)^2 + (4-4)^2 + (3-4)^2$$
$$+ (5-4)^2 + (6-4)^2 + (7-4)^2 = \mathbf{30}$$

$$df_{tot} = (k \times n) - 1 = 8$$

$$SS_{within} = \sum_i \sum_j (x_{ij} - \overline{X}_j)^2$$

$$= (3-2)^2 + (2-2)^2 + (1-2)^2$$
$$+ (5-4)^2 + (3-4)^2 + (4-4)^2$$
$$+ (5-6)^2 + (6-6)^2 + (7-6)^2 = \mathbf{6}$$

$$df_{within} = k(n-1) = 6$$

$$SS_{between} = \sum_i \sum_j (\overline{X}_j - \overline{\overline{X}})^2$$

$$= 3(2-4)^2 + 3(4-4)^2 + 3(6-4)^2 = \mathbf{24}$$

$$df_{between} = k - 1 = 2$$

How ANOVA works

m

n

Coffee	Tee	Water
3	5	5
2	3	6
1	4	7
$\overline{X}_1 = 2$	$\overline{X}_2 = 4$	$\overline{X}_3 = 6$

$$SS_{tot} = 30 \qquad df_{tot} = 8$$

$$SS_{between} = 24 \qquad df_{between} = 2$$

$$SS_{within} = 6 \qquad df_{within} = 6$$

$$F = \frac{SS_{between} / df_{between}}{SS_{within} / df_{within}}$$

$$SS_{tot} = SS_{between} + SS_{within}$$

$$df_{tot} = df_{between} + df_{within}$$

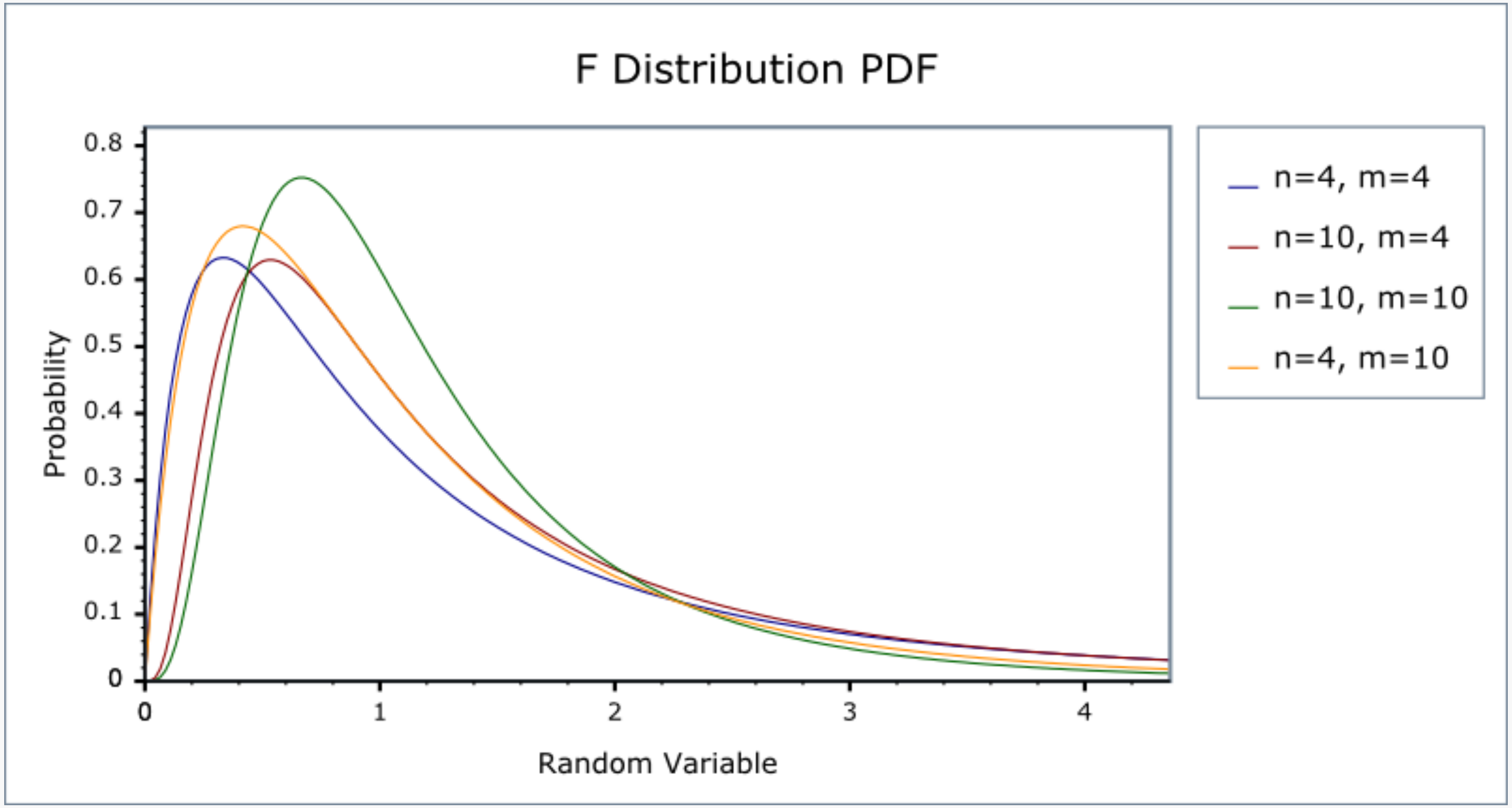
F ratio

$$F = \frac{variance\ between\text{ - }group}{variance\ within\text{ - }group} = \frac{SS_{between} / df_{between}}{SS_{within} / df_{within}} = \frac{MS_{between}}{MS_{within}}$$

- If the variance between groups is much larger than the variance within groups → large F → evidence *against* H_0
  - the difference between the means is likely due to the IV effect

- If the variance between groups is close to the variance within groups → small F → not enough evidence against H_0
  - the difference between the means is likely due to random variability

The F distribution



- $df : [k-1, N-k]$

Back to the example

$H_0: \mu_{coffee} = \mu_{tea} = \mu_{water}$

$\alpha = .05$

$H_1:$ *At least one mean is different*

Sum of Squares	Mean Squares	F statistics
$SS_{between} = 24,\ df = 2$ $SS_{within} = 6,\ df = 6$	$MS_{between} = 24/2$ $MS_{within} = 6/6$	$F = \frac{MS_{between}}{MS_{within}} = \frac{12}{1} = 12$

F - Distribution ($\alpha = 0.05$ in the Right Tail)

$df_2 \backslash df_1$ (Denominator Degrees of Freedom \ Numerator Degrees of Freedom)	1	2	3	4	5	6	7	8	9
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385
3	10.128	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867	8.8452	8.8123
4	7.7086	9.9443	6.5914	6.3882	6.2561	6.1631	6.0942	6.0410	6.9988
5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8759	4.8183	4.7725
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2067	4.1468	4.0990
7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767
8	5.3177	4.4590	4.0662	3.8379	3.6875	3.5806	3.5005	3.4381	3.3881
9	5.1174	4.2565	3.8625	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	2.9480	2.8962
12	4.7472	3.8853	3.4903	3.2592	3.1059	2.9961	2.9134	2.8486	2.7964
13	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.8321	2.7669	2.7144
14	4.6001	3.7389	3.3439	3.1122	2.9582	2.8477	2.7642	2.6987	2.6458
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066	2.6408	2.5876
16	4.4940	3.6337	3.2389	3.0069	2.8524	2.7413	2.6572	2.5911	2.5377
17	4.4513	3.5915	3.1968	2.9647	2.8100	2.6987	2.6143	2.5480	2.4943
18	4.4139	3.5546	3.1599	2.9277	2.7729	2.6613	2.5767	2.5102	2.4563
19	4.3807	3.5219	3.1274	2.8951	2.7401	2.6283	2.5435	2.4768	2.4227
20	4.3512	3.4928	3.0984	2.8661	2.7109	2.5990	2.5140	2.4471	2.3928
21	4.3248	3.4668	3.0725	2.8401	2.6848	2.5727	2.4876	2.4205	2.3660
22	4.3009	3.4434	3.0491	2.8167	2.6613	2.5491	2.4638	2.3965	2.3419
23	4.2793	3.4221	3.0280	2.7955	2.6400	2.5277	2.4422	2.3748	2.3201
24	4.2597	3.4028	3.0088	2.7763	2.6207	2.5082	2.4226	2.3551	2.3002
25	4.2417	3.3852	2.9912	2.7587	2.6030	2.4904	2.4047	2.3371	2.2821
26	4.2252	3.3690	2.9752	2.7426	2.5868	2.4741	2.3883	2.3205	2.2655
27	4.2100	3.3541	2.9604	2.7278	2.5719	2.4591	2.3732	2.3053	2.2501
28	4.1960	3.3404	2.9467	2.7141	2.5581	2.4453	2.3593	2.2913	2.2360
29	4.1830	3.3277	2.9340	2.7014	2.5454	2.4324	2.3463	2.2783	2.2229
30	4.1709	3.3158	2.9223	2.6896	2.5336	2.4205	2.3343	2.2662	2.2107
40	4.0847	3.2317	2.8387	2.6060	2.4495	2.3359	2.2490	2.1802	2.1240
60	4.0012	3.1504	2.7581	2.5252	2.3683	2.2541	2.1665	2.0970	2.0401
120	3.9201	3.0718	2.6802	2.4472	2.2899	2.1750	2.0868	2.0164	1.9588
∞	3.8415	2.9957	2.6049	2.3719	2.2141	2.0986	2.0096	1.9384	1.8799

$$F_{critical}(2,6) = 5.1433$$

$$F(2,6) = 12$$

$$p < .05$$

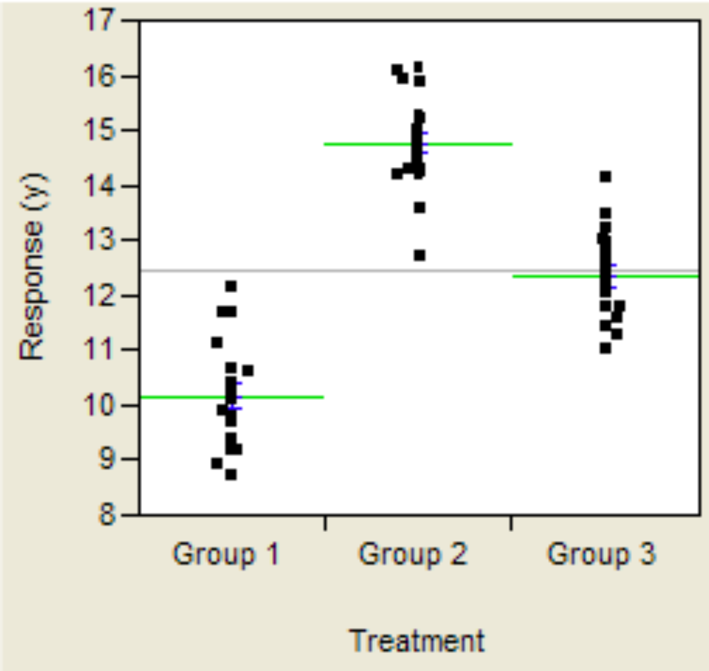
Reject H_0!

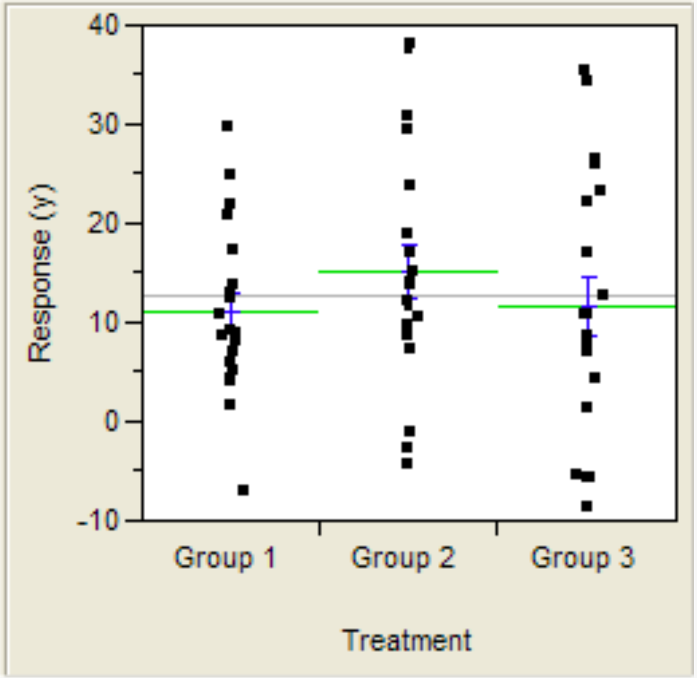
ANOVA table

Analysis of Variance

Source	DF	SS	MS	F	P
treatment	2	34.74	17.37	6.45	0.006
Error	22	59.26	2.69		
Total	24	94.00			

Variability between and within





The independant variable

- Our example: coffee/ no coffee

- Can have more levels (coffee, tea, water)
  - Compare 3 groups!

- There can be more than one independent variable
  - Coffee/ no coffee
  - Enough sleep / sleep deprivation
    - Compare 4 groups!

Why test more than 1 independent variable?

- Why not to test variables separately – in two different experiments?

- Because we expect an interaction!
  - The factors coffee and sleep influence each other

Factorial design of an experiment

- Crossing two independent variables, leads to 4 experimental conditions:

- COFFEE(coffee/nocoffee) x SLEEP(enough_sleep/not_enough_sleep)

    - Condition 1: coffee, enough_sleep
    - Condition 2: coffee, not_enough_sleep
    - Condition 3: nocoffee, enough_sleep
    - Condition 4: nocoffee, not_enough_sleep

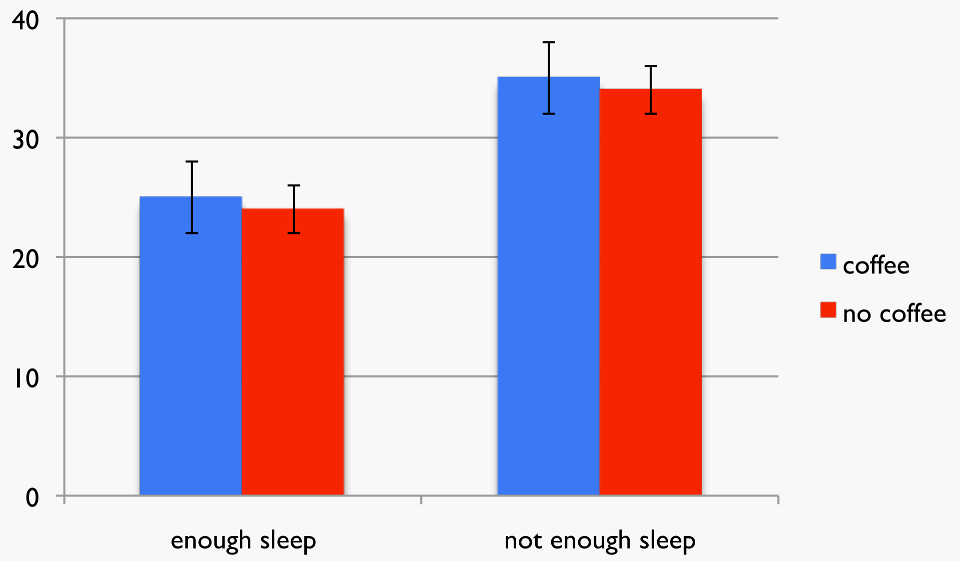
- Experiments can have more factors
  - 2 x 3 design = 6 conditions
  - 2 x 2 x 2 design = 8 conditions

What do we want from the analysis?

- Does coffee have an effect (*main effect* of coffee)?
- Does sleep have an effect (*main effect* of sleep)?
- Do coffee and sleep influence each other (*interaction* of coffee and sleep)?

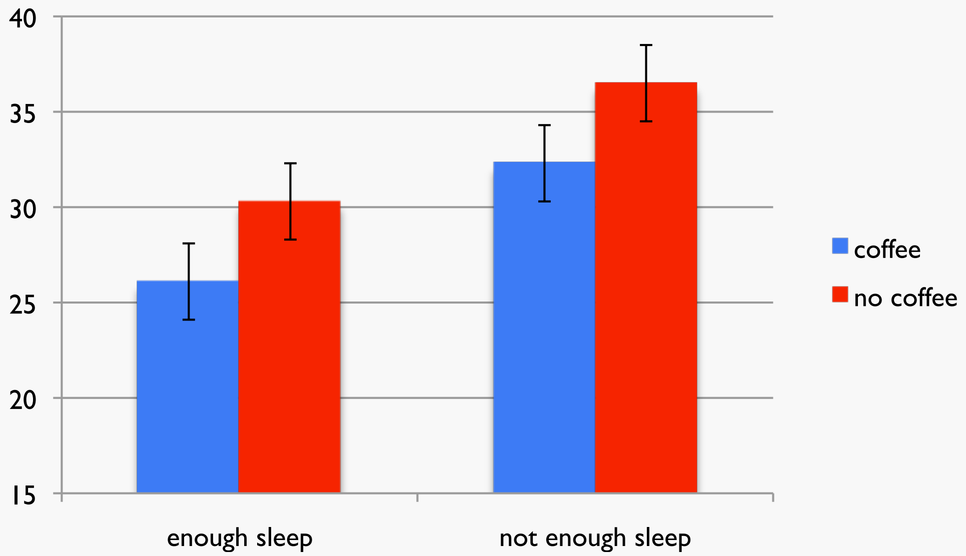
Possible outcomes

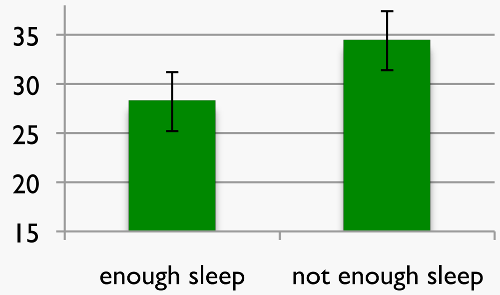
- Main effect of sleep

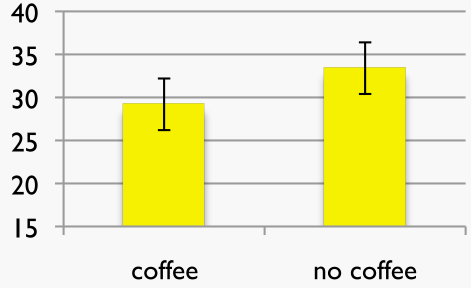


Possible outcomes

- Two main effects



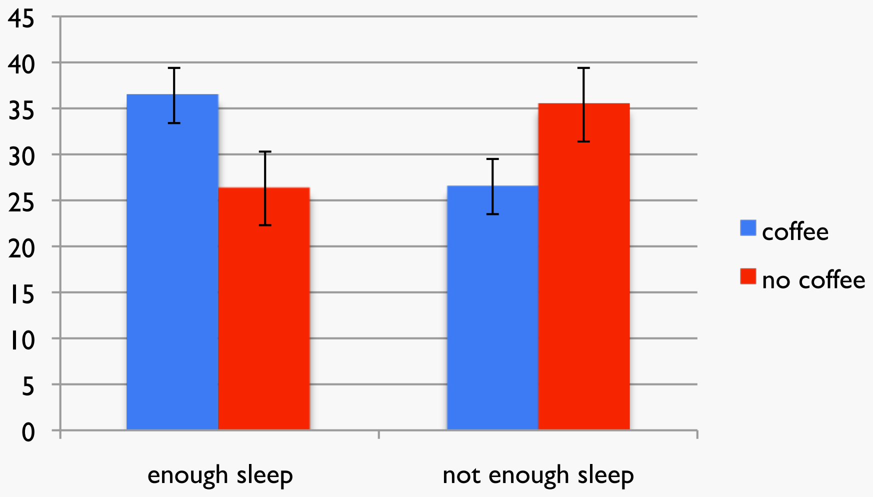


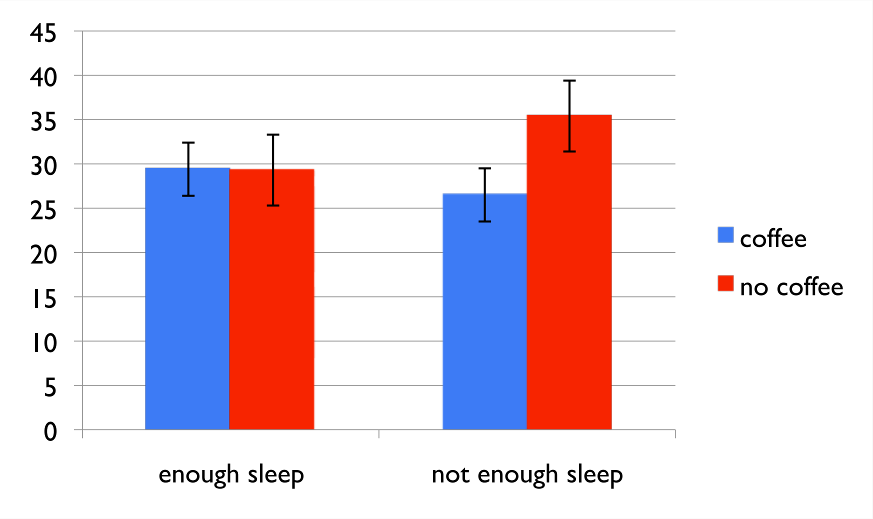


  - People are faster with coffee
  - People are faster with enough sleep

Possible outcomes II

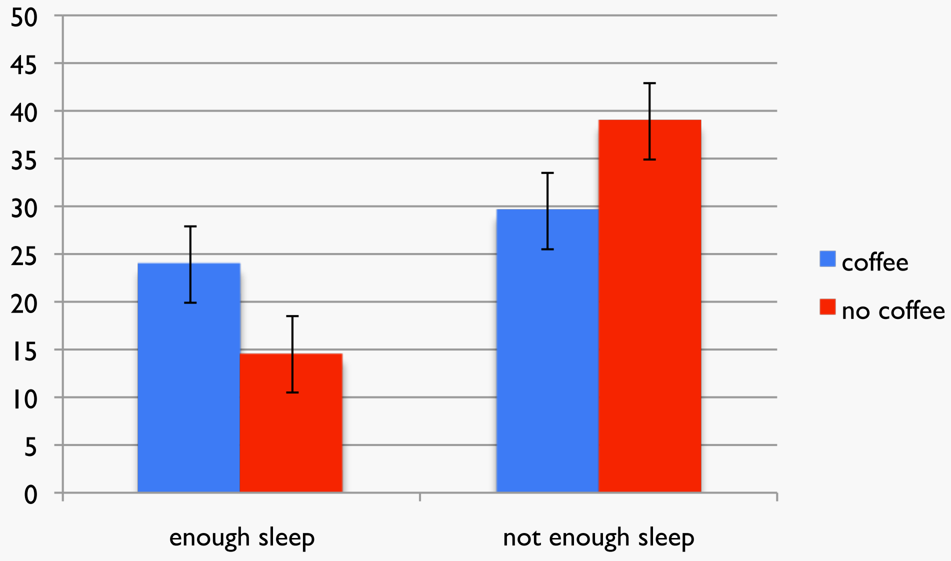
- Interaction





Possible outcomes III

- Main effect of sleep+ interaction



How do we interpret an interaction?

- The information we get from the ANOVA is that there is an interaction, not what kind of interaction
- At least we know that the two factors influence each other (are not independent)
- We don't know which differences between individual conditions are significant
- Pairwise comparisons!

Planned vs post hoc

- Planned comparisons:
  - Your hypothesis predicts a particular data pattern, e.g.

  Coffee makes students faster, but only if they are tired before

  - Difference between the conditions:
    - coffee, not_enough_sleep
    - nocoffee, not_enough_sleep
  - No Difference between the conditions:
    - coffee, enough_sleep
    - nocoffee, enough_sleep
- Perform two t tests with Bonferroni correction

Planned vs post hoc

- Post hoc tests:
  - Your hypothesis didn't state particular differences
    - possibly because you did not expect an interaction
  - Test all possible pairs!
    - Have to use a more conservative correction here
    - Tukey's Test

Different types of ANOVAs

- Between subjects design
  - Tests different participants in each condition
    - One-way ANOVA → 1 factor, independent sample
    - Factorial ANOVA → more than 1 factor, independent samples
- Within subjects design
  - Tests the same participants in all conditions
    - Repeated measure ANOVA ->1 factor, same participants in each condition
    - Repeated measure ANOVA → 2 or more factors, same participants
- Mixed design
  - Factorial design with both within and between factors

Summary ANOVA

- Dependant variable: continuous
- One or more independent variables with 2 or more levels each
- Gives significance-values for
  - Main effect (effect of one factor)
  - Interaction (influence of factors on each other)
- Usually requires additional testing
  - Planned comparisons
  - Post hoc tests

Hypothesis testing

- Identify the hypothesis
  - Be as specific as you can be!

- Define your dependant and independant variable(s)
- Classify your variables
  - Continuous or categorical?

- Do you test the same entity (person) in all conditions?
  - Use the paired or repeated measure variant

- Choose an appropriate test
  - T test, chi square test, ANOVA, something else

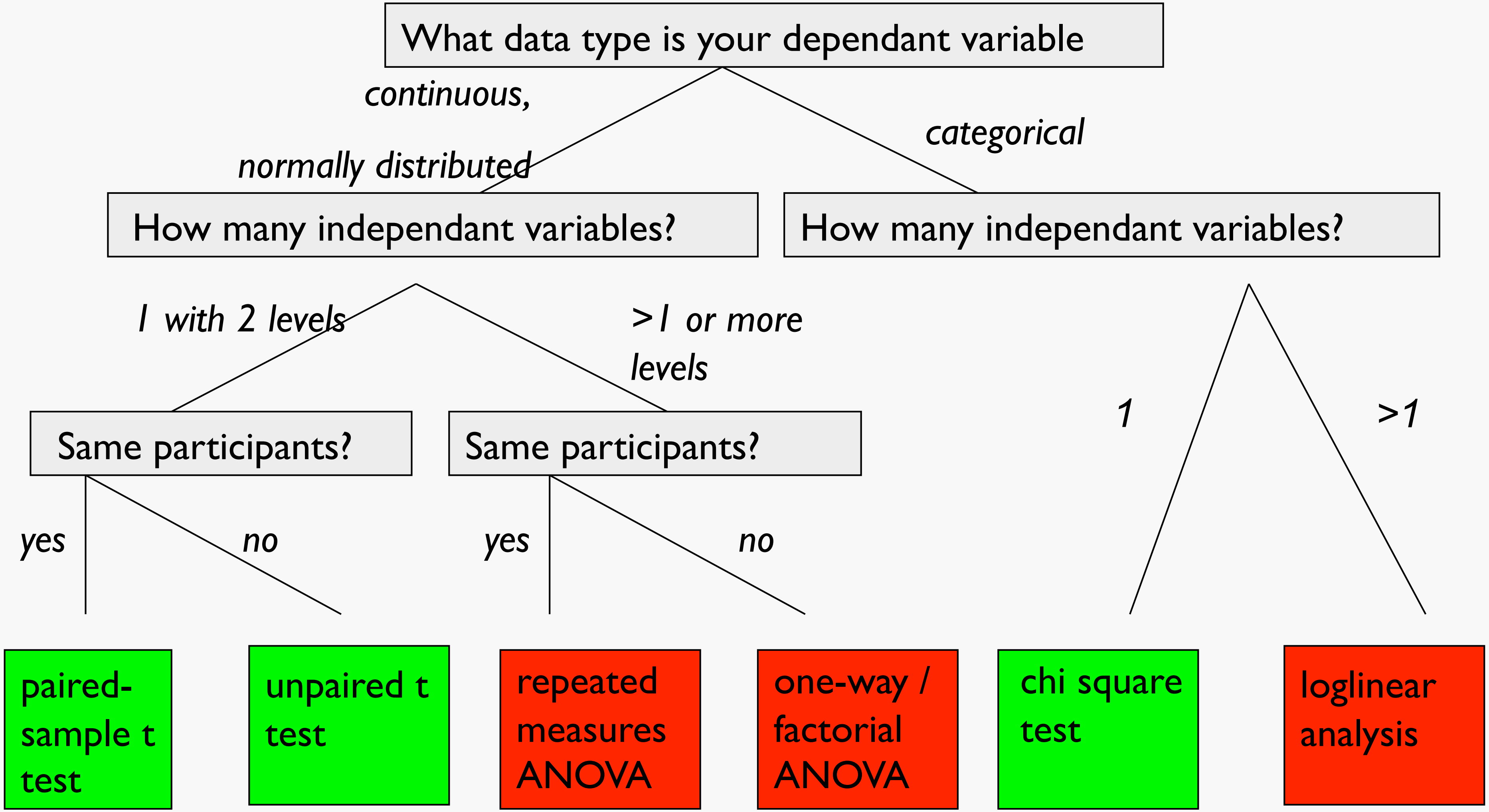
Hypothesis testing II

- Calculate the test statistic
  - Or have a programm do this for you ;)
- Compare the test statistic to the critical value depending on your α
- If the test statistic is above the critical value
  - Your result is significant, i.e.
  - The probability of observing your data if the Null-Hypothesis were true is below α ($p<0.05$)
- If the test statistic is below the critical value
  - Your result is not significant
  - You can‘t reject the Null-Hypothesis
  - DOES NOT MEAN THAT THERE IS NO DIFFERENCE!

Important concepts

- Dependant vs independant variables
- Data types: continuous vs categorical
- Level of significance:
  - α is the predefined boundary
  - p-value is the actual probability of our observation if the Null-Hypothesis is true
- Population and sample

Which test to use



Questions?

- What is α?
- What does it mean for a difference to be statistically significant?
- When do we use the t test?
- What is a continuous/categorical variable?
- What is a dependant/independent variable?
- Where do we get our hypotheses from?
- Why do statistical tests assume the Null Hypothesis (H_0)?
- What is an interaction?
- What kinds of errors can we make in hypothesis testing?

Software for statistical analysis

- Excel:
  - Chi square test, T tests
  - Descriptive: mean, variance, graphs

  \+ probably available to everybody ☺

- SPSS:
  - ANOVAs, loglinear, non-parametric tests
  - Everything that Excel can do

  \- licenses are expensive ☹

  available in the psycholinguistics department, but not on all machines

Software for statistical analysis

- R
  - Everything that Excel and SPSS can do
  - Can do other models (mixed effects models etc)
  - Without graphical user interface

    ⇒ You have to know what you are doing!

  + can be downloaded for free ☺

Helpful readings

Statistical analysis in general:

McDonald, J.H. (2008). *Handbook of Biological Statistics*. Sparky House Publishing. Baltimore: Maryland. http://udel.edu/~mcdonald/statintro.html

Statistics in SPSS:

Field, Andy (2009). *Discovering statistics using SPSS*. London, England: SAGE.

Statistics in R:

Baayen, R. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R. Cambridge: Cambridge University Press.*

Setting up an experiment

- He ate an apple. vs He ate a table.
- H_1: people take longer to read a word, if it does not match the semantic restrictions of the verb
- How do we test this?
  - Condition1: valid_object
  - Condition2: invalid_object
- Where to sample from:
  - all english speaker
  - all english nouns

Sampling from two populations

- Participants:
  - A random sample of English speakers
- Items:
  - A constructed sample of English sentences containing a selective verb and a noun

Constructing items

Usually, we want to test the same item in all conditions:

	valid	invalid
1	Peter eats an apple	Peter drives an apple
2	Paul plants a tree	Paul smokes a tree
3	Suzy reads a book	Suzy drinks a book

…

Additional variation: the verb

- control for frequency, length

Verb restrictions might be of different strength

- Use the same verb in the other condition, too!

Constructing items

- Use the verb in the other condition:

	valid	invalid
1	Peter eats an apple	Peter drives an apple
2	Paul plants a tree	Paul smokes a tree
3	Suzy reads a book	Suzy drinks a book

…

- Counterbalancing version:

	valid	invalid
1	Peter drives a car	Peter eats a car
2	Paul smokes a cigar	Paul plants a cigar
3	Suzy drinks a beer	Suzy reads a beer

…

Constructing lists

- Usually, we show each participant each item in only one condition

  - They might react differently when reading the same word again
  - 2 conditions and a counterbalancing version for the item
    - 4 experimental lists
    - Each should be tested an equal number of times

- Make sure that every condition appears equally often
- Randomize the list!

Validity

- If your participants can guess the goal of your experiment, they might behave differently!

- Don‘t tell them the purpose

- Try to distract them from the purpose
  - e.g. put in more sentences, that don‘t have anything to do with the experiment (filler items)

Summary

- You sample from two populations:
  - Participants
  - Items

- Try to eliminate as much variation in your materials as you can!
  - Controll for factors
  - Counterbalance your materials

- Try to prevent your participant from behaving strategically
  - Introduce filler items to distract from the real purpose