Statistics in experimental research

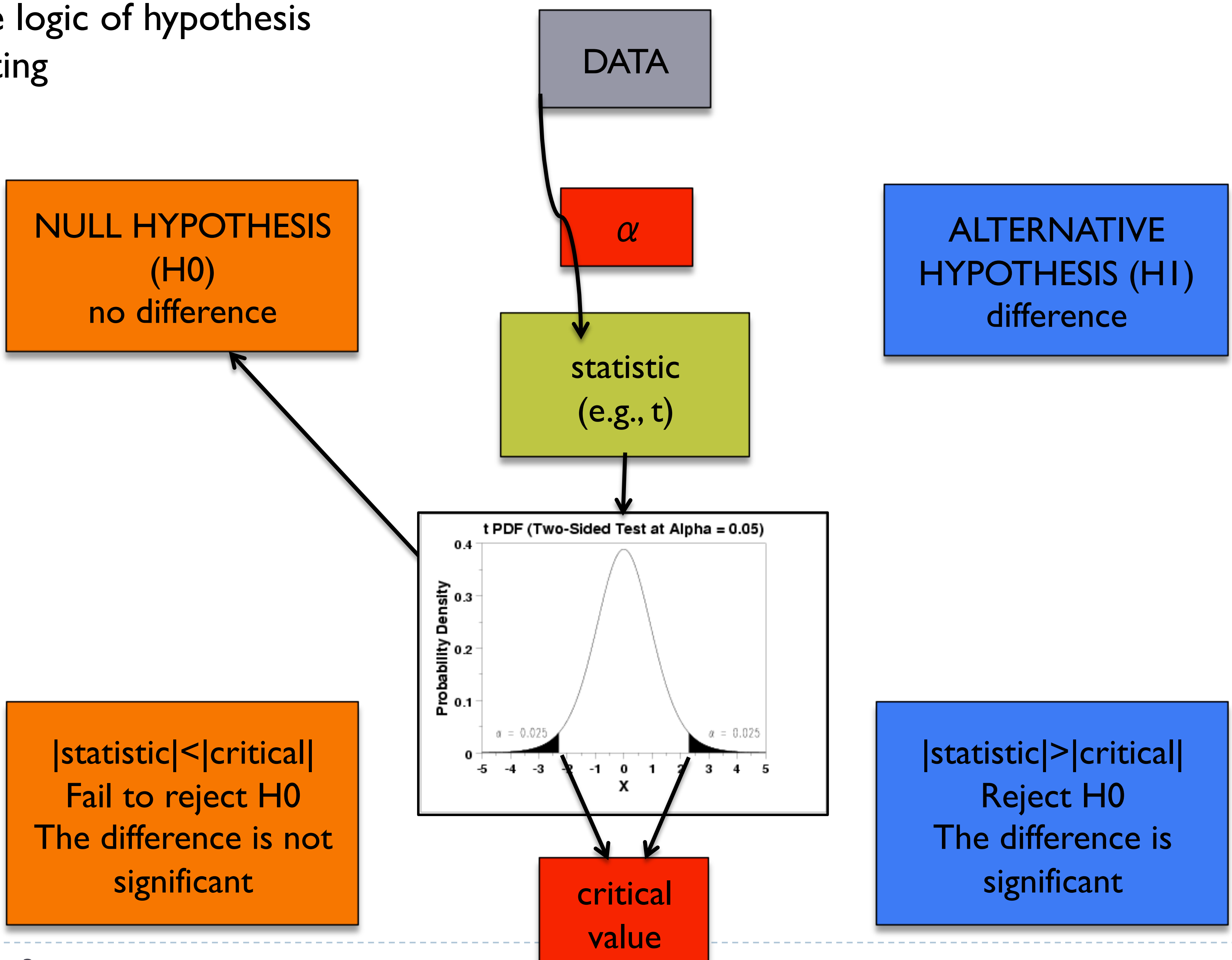
Session 2

Francesca Delogu
delogu@coli.uni-saarland.de

Overview today

- Recap
- Criteria for choosing a statistical test
  - Type of variables
  - Inflation of α
- χ^2 test
- One-way ANOVA

The logic of hypothesis testing



Why the t-test is called Student's t test

- From Wikipedia:

"The t-statistic was introduced in 1908 by William Sealy Gosset, a chemist working for the Guinness brewery in Dublin, Ireland ("Student" was his pen name). Gosset had been hired due to Claude Guinness's policy of recruiting the best graduates from Oxford and Cambridge to apply biochemistry and statistics to Guinness' industrial processes. Gosset devised the t-test as a way to cheaply monitor the quality of stout. He published the test in *Biometrika* in 1908, but was forced to use a pen name by his employer, who regarded the fact that they where using statistics as a trade secret."

Two general research strategies

Independent samples

- The two sets of data come from completely separate samples
- e.g., men and women
- an independent-measures t test is used
- between-subjects design

Related samples

- the two sets of data come from the same sample
- e.g., students before and after coffee
- a paired-samples t test is used
- within-subject design

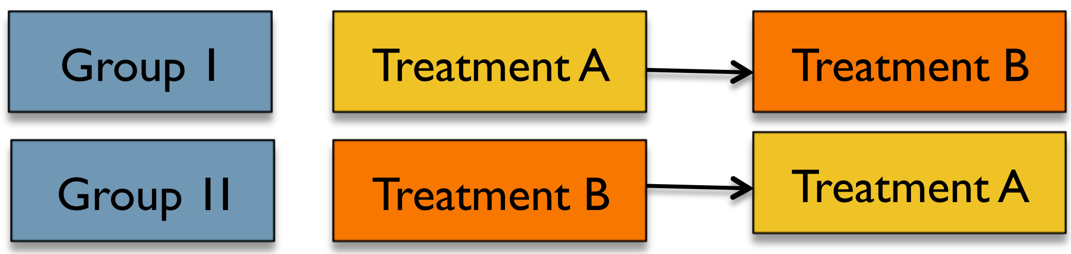
Pros and cons of within-subjects

- Pros
  - More economical (each participant tested on each condition)
  - Reduction in error variance: any factor that may effect performance on the dependent variable (e.g., sleep the night before), will be exactly the same for the two conditions.

- Contra
  - Carry-over effects (participation in one condition may effect performance in other conditions, thus creating a confounding variable)
    - Fatigue effect (in before/after experiments, performance may be worse on the second test)
    - Practice effect (in before/after experiments, performance may be better on the second test)

Counterbalancing

- Practice effects can be reduced if your conditions are counterbalanced



- Example
  - You split the sample in two, and half of them would get coffee before the first test, while the other half would get it before the second test.

Pros and cons of between-subjects

- Pros
  - No carry-over effects resulting from testing each participant twice
- Cons
  - Individual differences

Effect size

- Effect size is a measure of the strength of the effect that your IV had on your DV
- It can be measured as the standardized difference between two means
  - Cohen's *d*

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}$$

- In behavioral and social sciences the convention is:
  - .20 → small effect
  - .50 → moderate effect
  - .80 → large effect

Relationship between effect size and power

- The larger the effect size, the grater the power of a test (the probability of rejecting the null hypothesis when it is in fact false)

- The power of a test is influenced by the α level, the sample size, and the effect size

- Power analysis allows you to decide how large a sample is needed to enable reliable statistical judgments and how likely your statistical test will be to detect effects of a given size

- Power can be increased by increasing the sample size

Different data types

Reconsider yesterday‘s example:

- *After drinking coffee, students are faster to solve their homework*

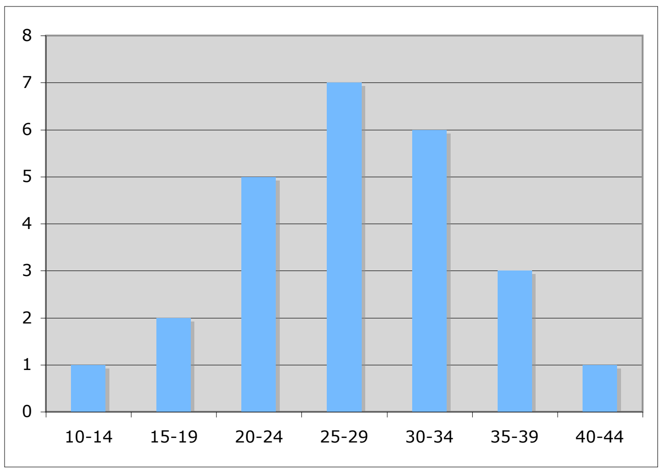
What if we want to extend the influence of coffee to performance in the exam?

- *After drinking coffee, students are more likely to pass the exam*

What is your dependent variable now?

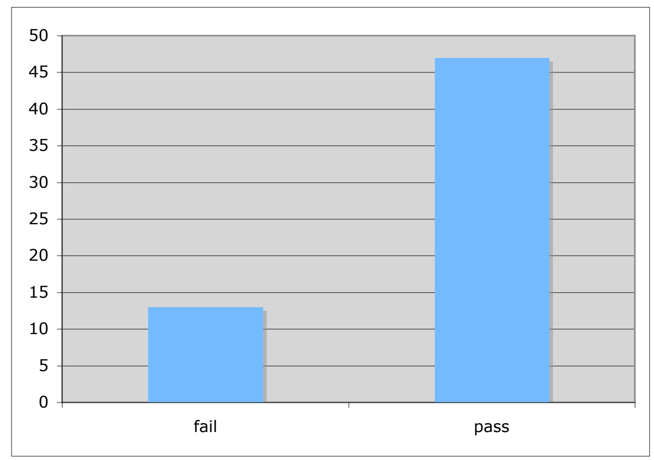
Distributions

Time to finish homework



Normal

Exam outcome



Binomial

Continuous variables

- Interval variables
  - measured along a continuous numerical scale
  - assume equal intervals between single units
  - e.g., temperature: a difference of 10 degrees is always the same no matter how hot or cold might be

- Ratio variables
  - interval variables with an absolute zero point.
  - the zero implies an absence of the thing being measured
  - e.g., weight: something could not weight a negative amount

Categorical variables

- Nominal variables
  - classify data into pre-defined categories with no intrinsic order.
  - e.g., gender (male/female); colour (green/red/blu…)

- Ordinal variables
  - rank scores in the order of being larger or smaller.
  - do not include information of the numeric difference between data point
  - e.g., the rank order of students based on their exam scores: 1^{st}, 2^{nd}, 3^{rd})

Classify the following variables

- Age
- Speed
- Month of birth
- Blood pressure
- Number of pizzas you can eat before fainting
- Whether or not you went to sleep before 12:00am
- Distance from home
- The number of letters in your last name
- Responses on a Likert scale

How do your variables look like?

- *After drinking coffee, students are more likely to pass the exam*

- IV → nominal
- DV→ nominal

categorical

The χ^2 Test (Chi Square)

Used when

- both the independent and the dependent variables are measured on a categorical scale
- All observations are independent and can appear only once in a table

The statistical hypotheses

H_0: The two variables are independent

- E.g., Consuming coffee and passing/failing an exam are independent events

H_1: The two variables are associated

- E.g., Coffee helps or hinders you to pass the exam

Some data

Student	coffee	no coffee
1	passed	
2		passed
3	failed	
4		passed
5		passed
6		passed
7		failed
8	passed	
9		passed
10		passed
11		passed
12	passed	
13		failed
14		passed
15	passed	
16	passed	
17	failed	
18		passed
19		passed
20		passed

- We want to know:
  - How many students passed the exam after coffee
  - How many students failed the exam after coffee
  - How many students passed the exam without coffee
  - How many students failed the exam without coffee

Contingency table of observed frequencies

- Step 1: transform your data into a frequency table:

	pass	fail		sum row
coffee	11	7		18
no coffee	33	9		42
sum column	44	16		60

- We want to contrast the observed frequencies in each cell with the **expected** frequencies.

- The expected frequencies represent the number of cases that would be observed in each cell if the null hypothesis were true (i.e., the two variables are independent)

Expected frequencies

How do we calculate the expected frequencies?

- $P(A, B) = P(A)\ P(B)$ if A and B are truly independent

	pass	fail		sum row
coffee				18
no coffee				42
sum column	44	16		60

Model for χ^2

$$E_{i,j} = \frac{Observed_{row_i} * Observed_{col\ j}}{N}$$

	pass	fail		sum row
coffee	13,2	4,8		18
no coffee	30,8	11,2		42
sum column	44	16		60

- Compare model to observed values!

Calculating χ^2

- Compare model to data:

The model:

	pass	fail
coffee	13,2	4,8
no coffee	30,8	11,2

The data:

	pass	fail
coffee	11	7
no coffee	33	9

The formula to calculate χ^2 :

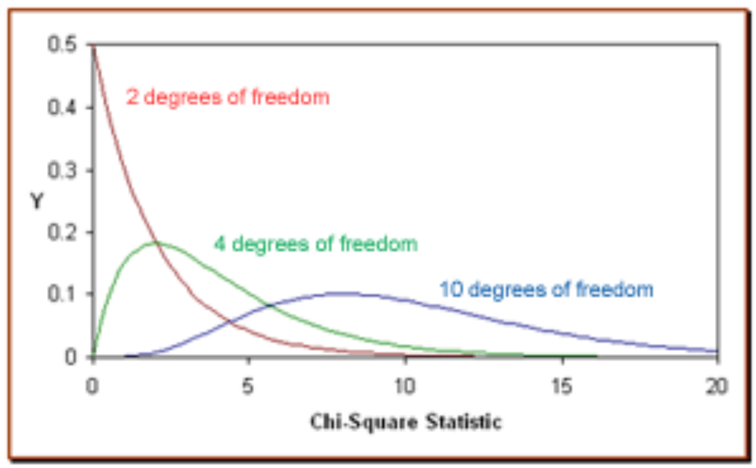
$$\chi^2 = \sum \frac{\left(Observed_{ij} - Model_{ij}\right)^2}{Model_{ij}}$$

$$\chi^2 = \frac{(11-13.2)^2}{13.2} + \frac{(7-4.8)^2}{4.8} + \frac{(33-30.8)^2}{30.8} + \frac{(9-11.2)^2}{11.2}$$

$$= 1.964$$

The χ^2 distribution

- The χ^2 distribution has one parameter
  - its degrees of freedom.
- It has a positive skew. As the df increase, the distribution approaches a normal distribution.



Finding critical value for χ^2

- Degrees of freedom:

  $(levels_{coffee}-1) \times (levels_{pass}-1) = 1$

- $\chi^2 = 1.964$

Table with critical values :

df	90%	95%	99%	…
1	2,71	3,84	6,63	…
2	4,61	5,99	9,21	…
3	6,25	7,81	11,34	…
…	…	…	…	…

T test and χ^2 test

- Very useful in comparing two groups with either continuous or categorical dependant variable

- Prerequisites:
  - T test
    - Both groups are normally distributed
    - Both groups have the same variance
  - χ^2 test
    - Expected values must be greater than 5

Other tests

Type of data / Goal	Continuous data from normal distributed population	Rank or Score (not normally distributed)	Binomial
Compare one group to a hypothetical value	One-sample t test	Wilcoxon test	Chi-square (χ^2)
Compare two unpaired groups	Unpaired t test (independant two-sample t test)	Mann-Whitney test	Chi square or Fisher's test
Compare two paired groups	Paired t test (dependant t test)	Wilcoxon test	McNemar's test

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!

Inflation of α

- 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 divides your α by the number of comparisons you do
  - Makes sure that the overall chance of making a Type I error is still α

One-way ANOVA

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

- 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

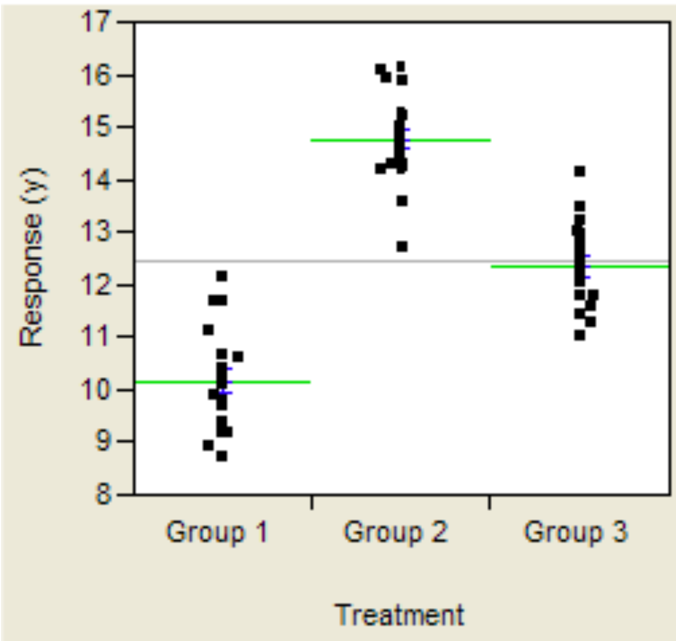
- variation BETWEEN groups
  - for each data value looks at the difference between its group mean and the overall mean (grand mean, GM)

$$\left(\bar{x}_i - \bar{x}\right)^2$$

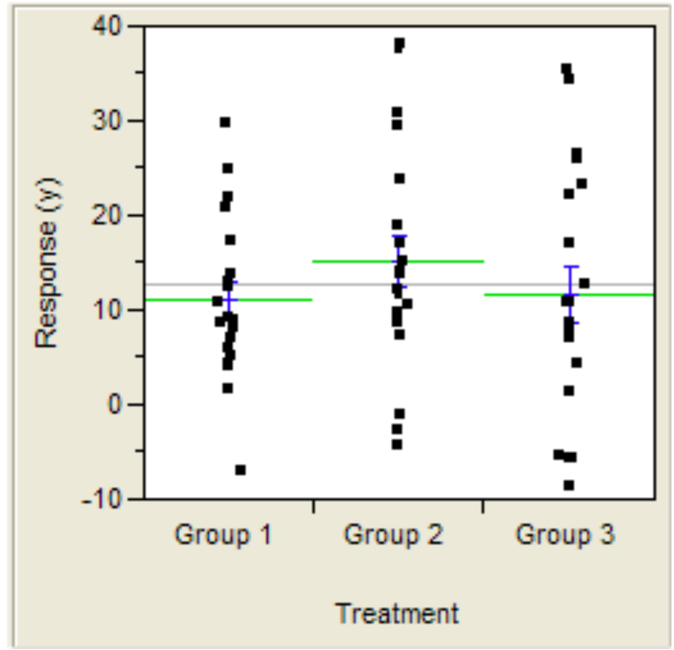
- variation WITHIN groups
  - for each data value looks at the difference between that value and mean of its group

$$\left(x_{ij} - \bar{x}_i\right)^2$$

Example



17
16
15
14
13
12
11
10
9
8
Response (y)
Group 1
Group 2
Group 3
Treatment



40
30
20
10
0
-10
Response (y)
Group 1
Group 2
Group 3
Treatment

F ratio

- The test statistic for ANOVA is called F

$$F = \frac{\text{between-group variability}}{\text{within-group variability}}.$$

- If the variability between groups is much larger than the variability within groups => the means are more likely to be different

ANOVA rationale

- ANOVA assumes:
  - Total variability = between-group variability + Error
  - Each data value = the GM + the IV effect + Error

ANOVA rationale

Model the data as

a) The mean of the whole sample
b) The mean plus an influence from the factor

- Calculate how much the observed data deviates from both models

- If the error is greatly reduced by including the factor, the factor has a significant influence

One-way ANOVA

- Used to test the effect of one factor with 2 or more levels

  - E.g., You want to know whether type of substance (coffee, tea, water) has an influence on time to finish homework.

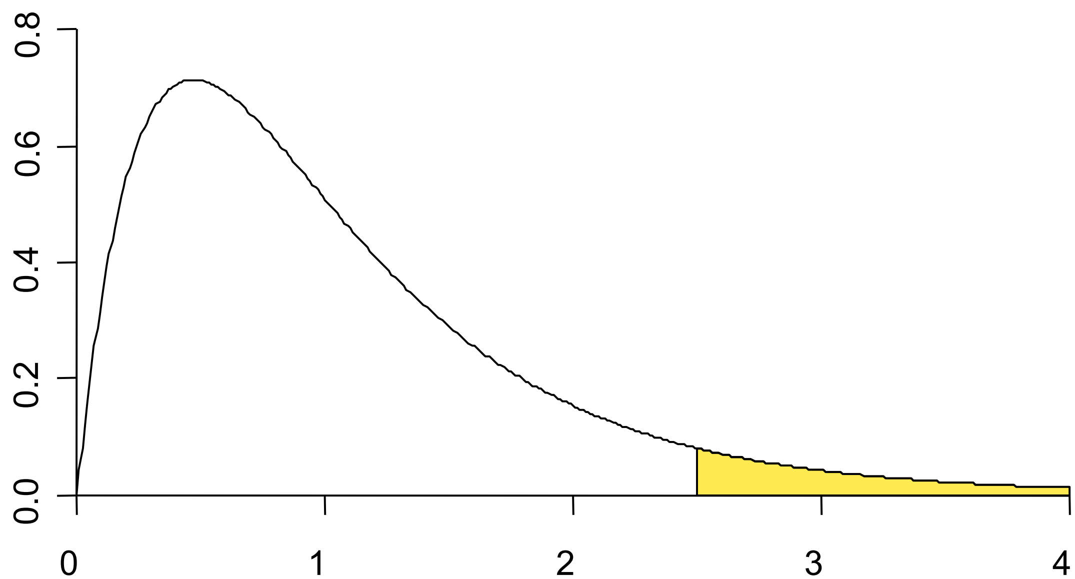
  - Factor = type of substance
  - Levels = coffee, tea, water

Statistical hypotheses

- H_0: $\mu_1 = \mu_2 = \mu_3$
- H_1: at least one of the means is different from another
- 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

The F distribution

- The f value can only be positive (so this is always a one-tailed test)
- df = (k-1) and (n-k)



- F(4,30) = 2.5, p = 0.06

ANOVA output

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			

- SS = Sum of Squares
- MS = Mean Square (Variance) = SS/DF
- $$F = \frac{MS_{Treatment}}{MS_{Error}}$$

Summary

- Criteria for choosing the test statistic
  - Data type
  - Inflation of the Type I error
  - The independent variable

- If we want to compare more than two groups, we can
  - Use Bonferroni correction
  - Use ANOVA (analysis of variance)
    - Then test the comparisons that you prefer