

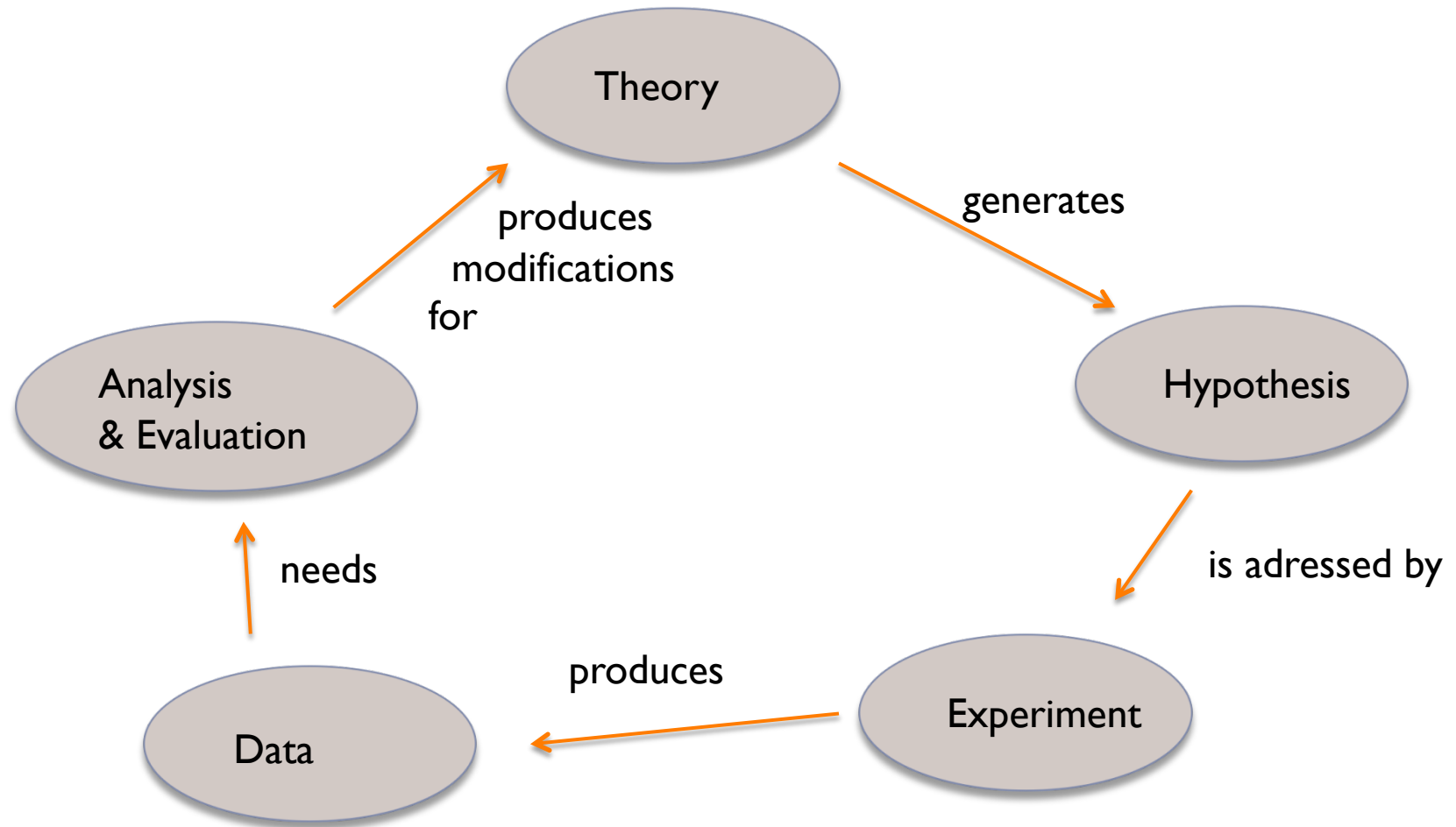
Statistics in experimental research

Session 1

Francesca Delogu

delogu@coli.uni-saarland.de

The research cycle



Overview

- ▶ **Today:**

- ▶ Statistical hypothesis testing
 - ▶ Sampling, distributions, confidence intervals
- ▶ Student's t-test

- ▶ **Thursday:**

- ▶ Different data types
- ▶ χ^2 Test

- ▶ **Friday:**

- ▶ Design
- ▶ ANOVAs

Let's start with an example

- ▶ Suppose we believe for some reason that caffeine improves cognitive performances of people
 - ▶ (e.g., memory, attention, etc.)
- ▶ How can we test this theory?
- ▶ We need a working hypothesis
 - ▶ experimental hypothesis

The experimental hypothesis

- ▶ An experimental hypothesis makes a prediction about the relationship between two (or more) events, or variables.

Example:

- ▶ Caffeine improves cognitive performances
 - ➡ People are better at recalling a text after a cup of coffee.
 - ➡ Students are faster to finish their homework after a cup of coffee.

The experimental variables

▶ **Independent variable (IV)**

the variable you manipulate in an experiment, set up independently before the experiment even begins

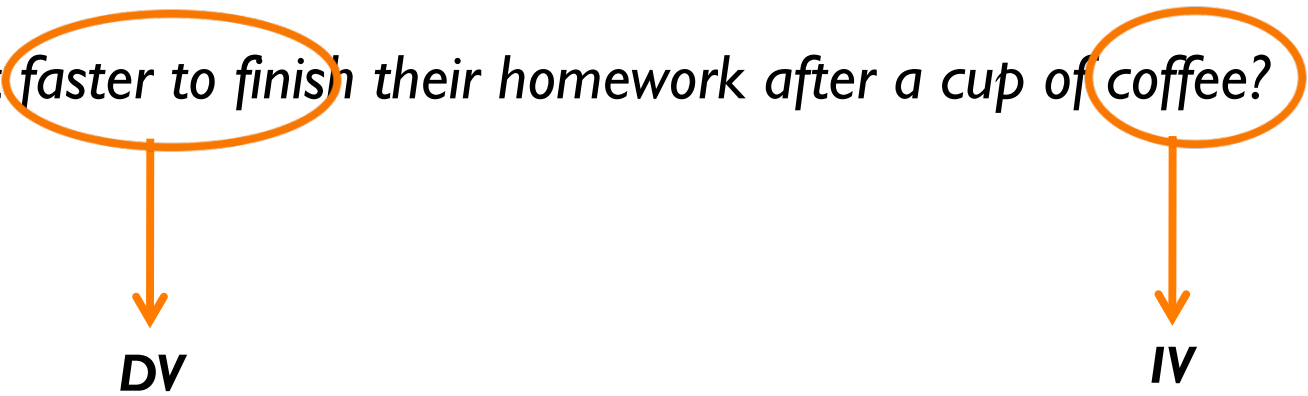
- ▶ also called factor or predictor
- ▶ can take two or more levels

▶ **Dependent variable (DV)**

the variable that you measure in an experiment, dependent on the way in which the experimenter manipulates the independent variable.

- ▶ E.g., reading times, etc.

Example

- ▶ Are student *faster to finish* their homework after a cup of *coffee*?


DV

IV
- ▶ What the experiment does is manipulating the independent variable (having or not a cup of coffee) to see whether it has an effect on the dependent variable (time to finish the homework).

The null hypothesis

- ▶ The experimental hypothesis is always tested against a **null hypothesis**.
- ▶ The null hypothesis states that any results found in the experiment will be due to chance.

Example

- ▶ **Null hypothesis (H_0):**

- ▶ Coffee has no influence on time to finish homework; any observed difference will be due to chance

- ▶ **Alternative hypothesis (H_1):**

- ▶ Coffee has an influence on time to finish homework.
 - ▶ Two-tailed hypothesis (neutral with respect to the direction of the effect)
- ▶ Coffee speeds up time to finish homework
 - ▶ One-tailed hypothesis (makes a guess on the direction of the effect)

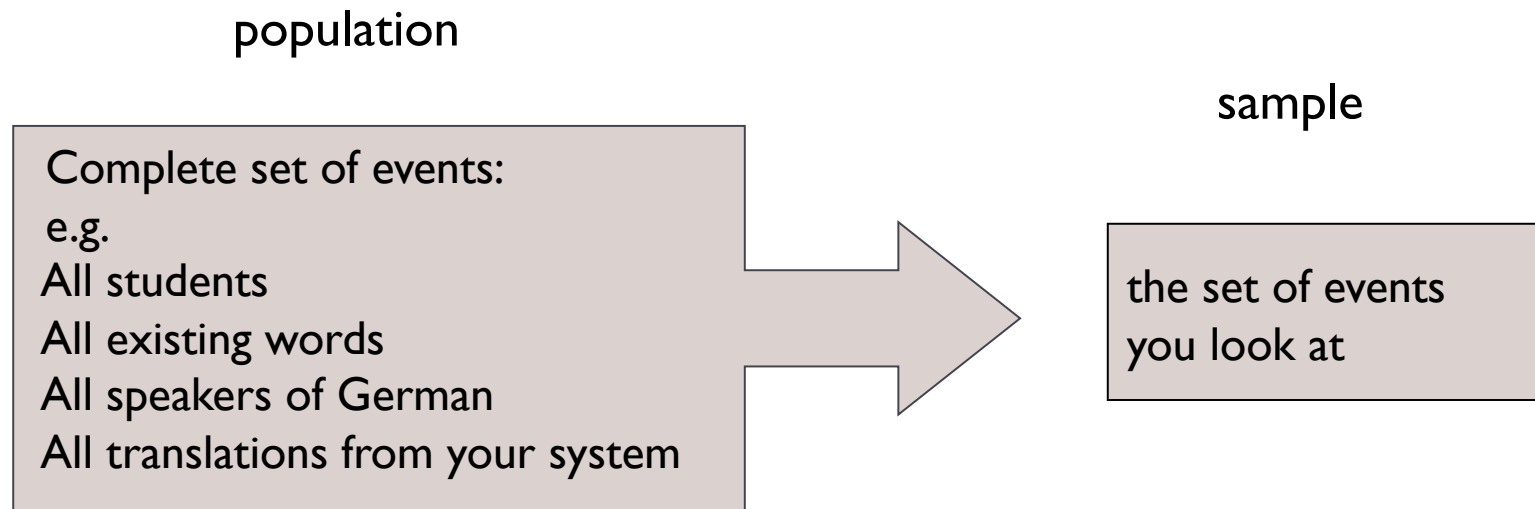
Goal of hypothesis testing

- ▶ Decide between H_0 and H_1
- ▶ In particular, the main goal of hypothesis testing is to tell us whether we have enough evidence to *reject* H_0 .
 - ▶ i.e., how probable your result is, assuming H_0 is true
- ▶ How do we do this?

Example

- ▶ *Does coffee have an influence on time to finish homework?*
- ▶ Two options:
 - ▶ look at all students and every instance where they finished their homework (impossible)
 - ▶ look at a subset of students doing one particular homework with and without coffee (sample)

Sampling



- ▶ Statistical hypotheses are always assumptions about a population parameter (e.g., the mean), which may or may not be true
- ▶ Based on the sample, we try to estimate the population parameter
- ▶ The best way to sample is random
- ▶ Sampling is always subject to bias or error

Back to the example

- ▶ *Does coffee have an influence on time to finish homework?*
- ▶ Two options:
 - ▶ Select N students and, for each of them, look at time to finish a particular homework before and after coffee (two dependent samples)
 - ▶ Select two groups of N students, give coffee to only one of them, and look at time to finish a particular homework (two independent samples)
- ▶ Let's try with the second option!

Data

| | coffee | no coffee |
|--------|--------|-----------|
| Anne | 35 | |
| Jim | 25 | |
| John | 28 | |
| Mary | 25 | |
| Peter | 19 | |
| Carl | 31 | |
| Judy | 18 | |
| Bob | 30 | |
| Liz | 26 | |
| Betty | 23 | |
| Sandra | | 30 |
| Tom | | 23 |
| Frank | | 35 |
| Kate | | 32 |
| Mark | | 26 |
| Carol | | 33 |
| Tedd | | 39 |
| Susan | | 22 |
| Helen | | 32 |
| David | | 28 |

- ▶ Mean group_{coffee} = 26 min
- ▶ Mean group_{no_coffee} = 30 min
- ▶ Diff = 4 min

- ▶ What does this difference tell us?

Nothing!

Statistical significance

- ▶ We need to test whether the 4 min difference is *statistically significant* (i.e., not likely due to chance)

Recap

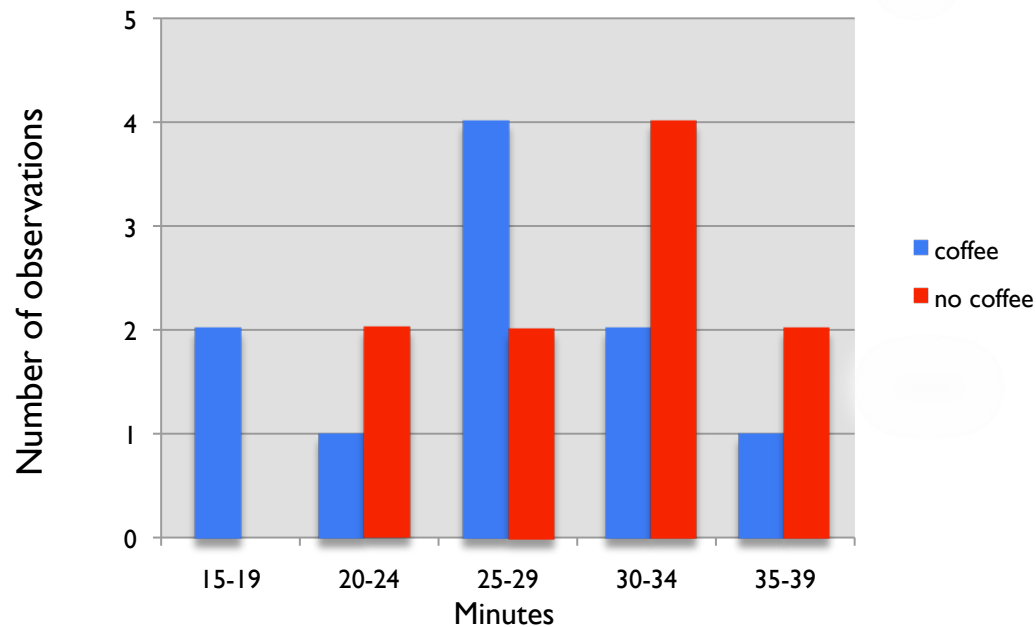
$H_0: \mu_{\text{coffee}} = \mu_{\text{no_coffee}} \rightarrow \text{population}_{\text{coffee}} = \text{population}_{\text{no_coffee}}$

$H_1: \mu_{\text{coffee}} \neq \mu_{\text{no_coffee}} \rightarrow \text{population}_{\text{coffee}} \neq \text{population}_{\text{no_coffee}}$

- ▶ We collected two samples (*coffee* vs. *no_coffee*), measured the time to finish homework, and found that the coffee-group was on average 4 min faster than the no_coffee-group.
- ▶ If H_0 is true, then the two samples are drawn from the same population (which we assume is normally distributed) and the 4 min difference is due to chance
- ▶ If H_1 is true, then the two samples are drawn from different distributions and the 4 min difference is due to the influence of caffeine

Example

- Assumption: samples are drawn from a normal distribution with a certain mean and variance



- We want to test whether the two samples are drawn from the same distribution

Statistical tests

- ▶ To test the probability that two samples are drawn from the same distribution we compute a test statistic
- ▶ Test statistics (t , χ^2 , etc.,) have well-known distributions and can tell you how likely it is for your data to come out the way they did if the null hypothesis is true
- ▶ If this probability is low enough, then we can reject the null hypothesis

Significance level: α

- ▶ α an arbitrary cutoff value representing the risk you are willing to take of rejecting the null hypothesis when it is actually true.
- ▶ If you want to be 95% confident of making the right decision (i.e., rejecting the null hypothesis when it is in fact false), then you are prepared to accept a 5% chance of making the wrong decision.
- ▶ In such a case, your α level is 0.05
- ▶ Thus, if the probability that your data came out the way they did by chance (i.e., assuming the null hypothesis is true) is less than or equal the significance level, the null hypothesis is rejected and the result is said to be statistically significant

Back to the data

| | coffee | no coffee |
|--------|--------|-----------|
| Anne | 35 | |
| Jim | 25 | |
| John | 28 | |
| Mary | 25 | |
| Peter | 19 | |
| Carl | 31 | |
| Judy | 18 | |
| Bob | 30 | |
| Liz | 26 | |
| Betty | 23 | |
| Sandra | | 30 |
| Tom | | 23 |
| Frank | | 35 |
| Kate | | 32 |
| Mark | | 26 |
| Carol | | 33 |
| Tedd | | 39 |
| Susan | | 22 |
| Helen | | 32 |
| David | | 28 |

- ▶ Mean group_{coffee} = 26 min
- ▶ Mean group_{no_coffee} = 30 min
- ▶ Diff = 4 min

- ▶ Is this difference statistically significant?

Let's test it!!


- ▶ For samples of two normally distributed populations with similar variance:

Student's t-test


The t-test

- ▶ First, we need to compute a t-value (our **statistic**):

$$t = \frac{\text{difference between group means}}{\text{variability of groups}}$$



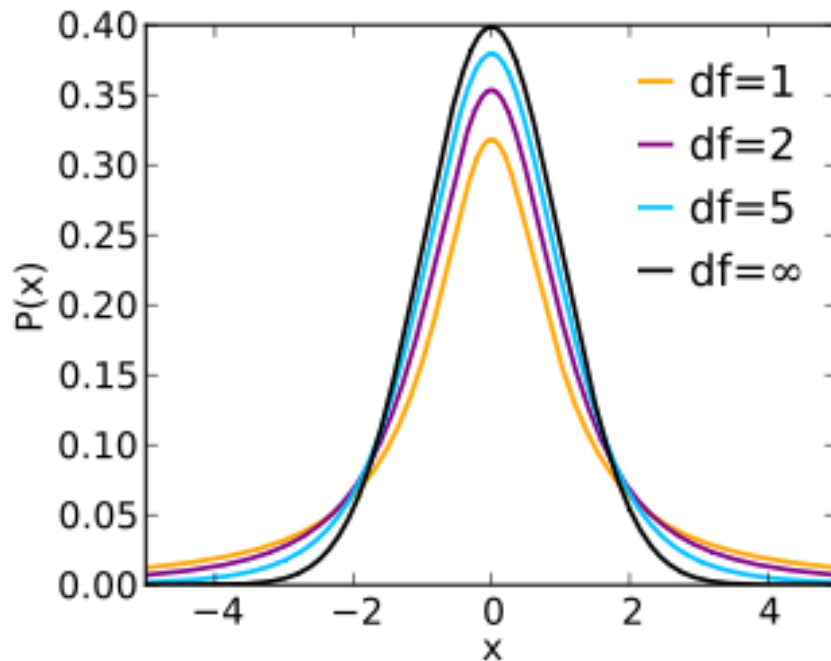
predicted variability due to independent variables



total variability due to all variables

- ▶ t is close to 0 if means are close
- ▶ t is smaller if variance is big
- ▶ t is dependent on the size of the sample N
- ▶ t is associated with a probability

The t-distribution



- ▶ The t-distribution is a family of curves the shape of which depends on the number of degrees of freedom (df)
- ▶ As df goes to infinity, the t-distribution converges to the standard normal distribution.

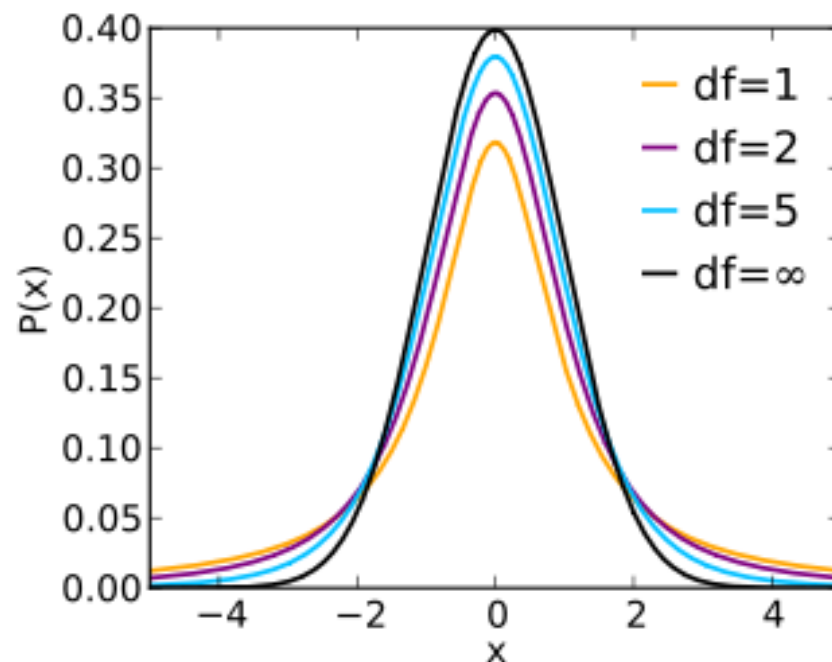
- ▶ The df for the t-test is the total number of observations in the two groups minus 2, or $n_1 + n_2 - 2$

Degrees of freedom

- ▶ The number of df is a function of both the number of observations and the number of parameters estimated.
- ▶ It is the number of values in the final calculation of a statistic that are free to vary.
- ▶ Imagine you have four numbers (a, b, c and d) that must add up to a total of m ; you are free to choose the first three numbers at random, but the fourth must be chosen so that it makes the total equal to m - thus your degree of freedom is three.

The t-distribution

- ▶ The t-distribution represents the relative frequencies of the possible t values if the null hypothesis is true.
 - ▶ Notice that the t distribution is centred at 0, which is what the t statistic would be if the difference between the means is 0



What is the probability of finding a t statistic equal or greater than 1 if the null hypothesis is true?

Unpaired-sample t test

| | coffee | no coffee |
|--------|--------|-----------|
| Anne | 35 | |
| Jim | 25 | |
| John | 28 | |
| Mary | 25 | |
| Peter | 19 | |
| Carl | 31 | |
| Judy | 18 | |
| Bob | 30 | |
| Liz | 26 | |
| Betty | 23 | |
| Sandra | | 30 |
| Tom | | 23 |
| Frank | | 35 |
| Kate | | 32 |
| Mark | | 26 |
| Carol | | 33 |
| Tedd | | 39 |
| Susan | | 22 |
| Helen | | 32 |
| David | | 28 |

$$t = \frac{\text{mean}_{\text{coffee}} - \text{mean}_{\text{nocoffee}}}{\sqrt{(s_{\text{coffee}}^2 + s_{\text{nocoffee}}^2) / N}}$$

$$s^2 = \frac{1}{N-1} \times \sum_{i=1}^N (\text{value}_i - \text{mean})^2$$

Calculate the t-value!!!

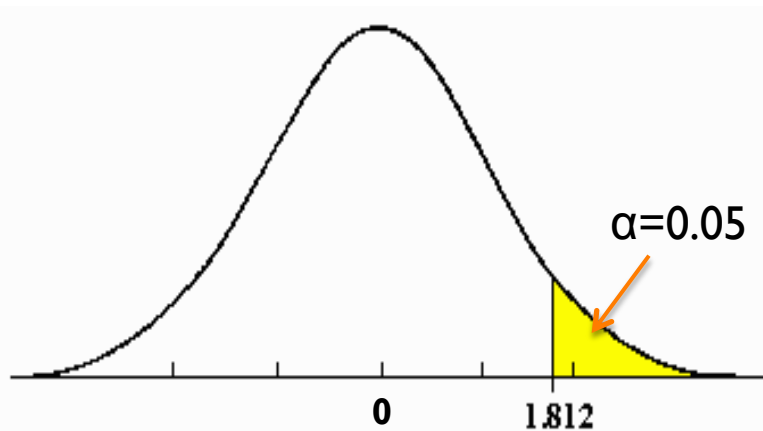
$$t = 1.68$$

t-critical

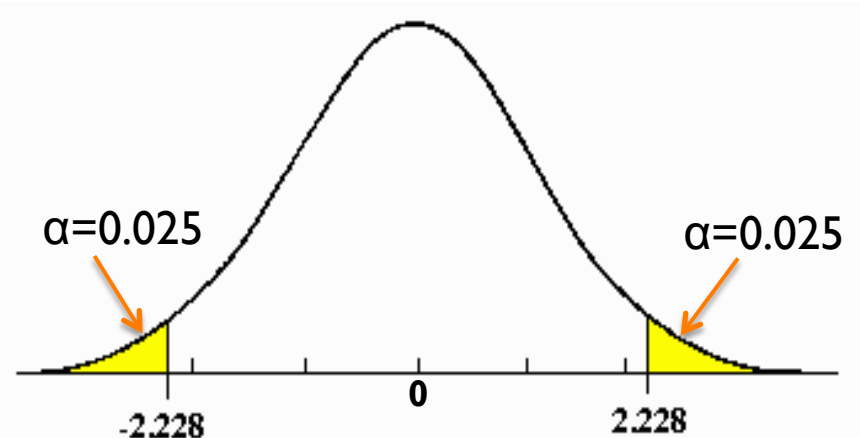
- ▶ Once we have computed the t-value, we want to know whether the probability of observing t under the null hypothesis is less than or equal to our level of significance α (0.05)
- ▶ i.e., we want to find the value that t must exceed in order for the null hypothesis to be rejected
- ▶ This value is called t-critical and depends on
 - ▶ The number of df
 - ▶ Whether H_1 is one-tailed or two-tailed

Finding t-critical

- ▶ One-tailed t-test
- ▶ 10 df
- ▶ $\alpha = 0.05$



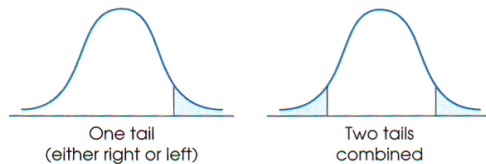
- ▶ Two-tailed t-test
- ▶ 10 df
- ▶ $\alpha = 0.05$



The t-table

TABLE B.2 THE t DISTRIBUTION

Table entries are values of t corresponding to proportions in one tail or in two tails combined.



| df | PROPORTION IN ONE TAIL | | | | | |
|-----|----------------------------------|-------|-------|--------|--------|--------|
| | 0.25 | 0.10 | 0.05 | 0.025 | 0.01 | 0.005 |
| df | PROPORTION IN TWO TAILS COMBINED | | | | | |
| | 0.50 | 0.20 | 0.10 | 0.05 | 0.02 | 0.01 |
| 1 | 1.000 | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 |
| 2 | 0.816 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 |
| 3 | 0.765 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 |
| 4 | 0.741 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 |
| 5 | 0.727 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 |
| 6 | 0.718 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 |
| 7 | 0.711 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 |
| 8 | 0.706 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 |
| 9 | 0.703 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 |
| 10 | 0.700 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 |
| 11 | 0.697 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 |
| 12 | 0.695 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 |
| 13 | 0.694 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 |
| 14 | 0.692 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 |
| 15 | 0.691 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 |
| 16 | 0.690 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 |
| 17 | 0.689 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 |
| 18 | 0.688 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 |
| 19 | 0.688 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 |
| 20 | 0.687 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 |
| 21 | 0.686 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 |
| 22 | 0.686 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 |
| 23 | 0.685 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 |
| 24 | 0.685 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 |
| 25 | 0.684 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 |
| 26 | 0.684 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 |
| 27 | 0.684 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 |
| 28 | 0.683 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 |
| 29 | 0.683 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 |
| 30 | 0.683 | 1.310 | 1.697 | 2.042 | 2.457 | 2.750 |
| 40 | 0.681 | 1.303 | 1.684 | 2.021 | 2.423 | 2.704 |
| 60 | 0.679 | 1.296 | 1.671 | 2.000 | 2.390 | 2.660 |
| 120 | 0.677 | 1.289 | 1.658 | 1.980 | 2.358 | 2.617 |
| ∞ | 0.674 | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 |

- ▶ $t\text{-value} = \pm 1.68$
- ▶ $t\text{-critical} = \pm 2.101$
- ▶ $t\text{-value} < t\text{-critical}$
- ▶ We fail to reject the null hypothesis
- ▶ The difference is not significant!

How do we report a result?

- ▶ $t(18) = 1.68, NS$ → null result
- ▶ $t(18) = 2.69, p < 0.05$ → significant result
- ▶ The **p-value** is the the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true.

Errors

- ▶ A null result does not mean that H_0 is true
- ▶ We might fail to reject the null hypothesis, although there is a difference
⇒ Type II error (false negative)
- ▶ A significant difference doesn't mean this difference is beyond doubt!
- ▶ With a 5% chance, this difference is not there
⇒ Type I error (false positive)

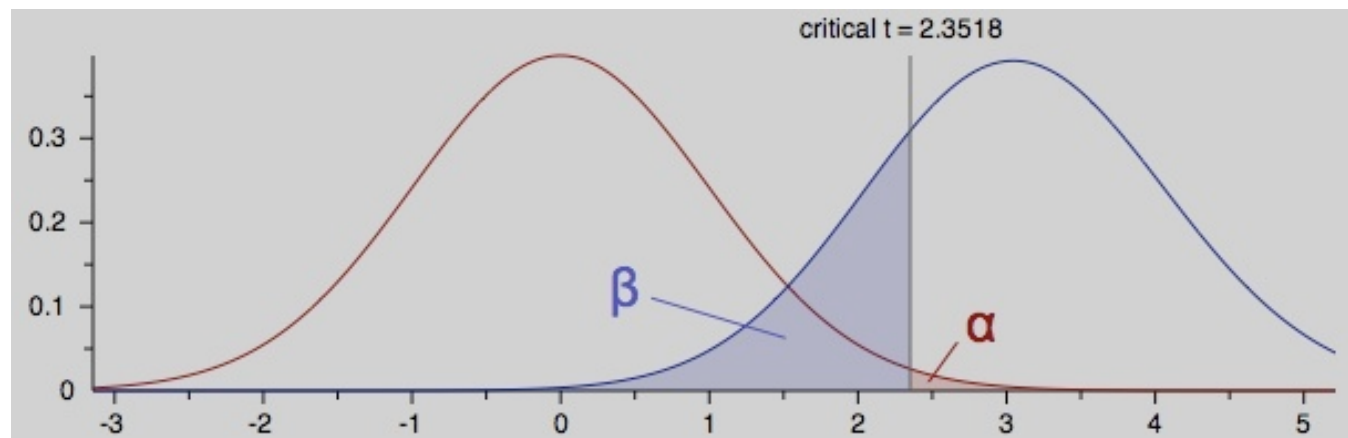
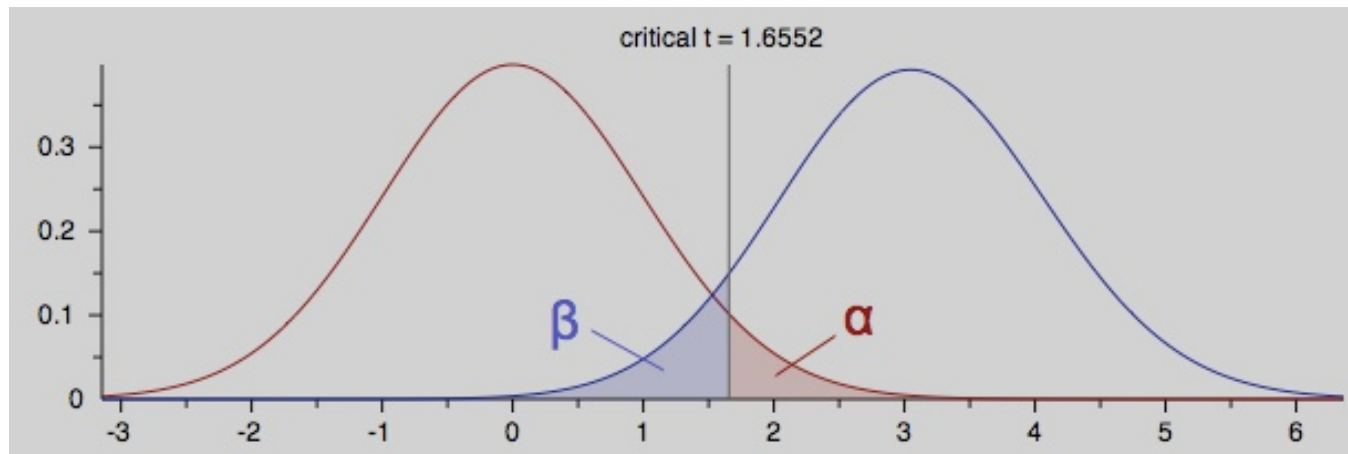
Table of errors

| | H_0 True | H_0 False |
|----------------------|------------------|-------------------|
| Reject H_0 | Type I Error | Correct Rejection |
| Fail to Reject H_0 | Correct Decision | Type II Error |

α = probability of committing a Type I error

β = probability of committing a Type II error

Relationship between α and β



Statistical power

- ▶ Power = $1 - \beta$

Probability that the test will reject the null hypothesis when the null hypothesis is actually false.

- ▶ Power is influenced by many factors:

- ▶ α level
- ▶ Effect size
- ▶ Sample size

- ▶ Samples too small or too much variability in the data decrease the power of the test, increasing the probability of committing a type II error.

- ▶ Change the experimental design!

Paired-sample T Test

- ▶ Measure the same participants in both conditions
 - ▶ More participants to test
 - ▶ Less variability associated with individual differences
- ▶ Power increases

| | coffee | | no coffee | difference |
|-------|--------|--|-----------|------------|
| Anne | 35 | | 39 | -4 |
| Jim | 25 | | 32 | -7 |
| John | 28 | | 26 | 2 |
| Mary | 25 | | 35 | -10 |
| Peter | 19 | | 23 | -4 |
| Carl | 31 | | 30 | 1 |
| Judy | 18 | | 22 | -4 |
| Bob | 30 | | 32 | -2 |
| Liz | 26 | | 28 | -2 |
| Betty | 23 | | 33 | -10 |

$$t = \frac{\text{mean}_{\text{difference}}}{\sqrt{s_{\text{difference}}^2 / N}}$$

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

Use a within-subject design whenever possible!!

Summary

1. We **sample** from the whole **population** and build a **model** of the population (e.g. mean).
2. We define the null and alternative hypotheses for the question under examination
3. We establish the level of significance (alpha level)
4. Finally, we use a **test statistic** to find out whether an observed difference is **significant**, i.e. very unlikely to be due to chance.

Exercise

- I. You have a machine translation system and want to compare it to another one

How do you apply a t test?