FLST: Cognitive Foundations

Francesca Delogu
delogu@coli.uni-saarland.de

http://www.coli.uni-saarland.de/courses/FLST/2013/

Schedule

Experimental research in psycholinguistics

- Today
  - Experimental methods in psycholinguistic research
- Wednesday
  - Principles of experimental design
  - Basic statistical concepts for data analysis
- Monday
  - Tutorial

Experimental variables

- Any experiment can be described as a study investigating the effects of some factor X on some type of behavior Y.

- Any experiment involves

  1) *varying some factor* (or factors)
  → independent variable (IV)

  2) *holding all other factors constant*
  → extraneous (confounding) variables

  3) observing *the results* of the variation
  → dependent variable (DV)

The independent variable

- The factor of interest, the one that is being studied to see if it will influence behavior

- Also called a "*manipulated*" factor because the experimenter has *complete control over it*

    - Independent variables must have a minimum of two contrasts or *levels* (also called ***conditions***)

        - At the very least, an experiment involves a comparison between two conditions

Operationalize IVs

Research question: Are focused words faster to identify than non-focused words?

- IV = focus
- Levels = focus, non-focus
  - Must clarify: Syntactic focus? Prosodic focus? Semantic focus?
  - Must operationalize: Clefting? Fronting? Other devices?

Subject variables

- Refer to already existing characteristics of the individuals participating in the study, such as gender, age, socioeconomic class, cultural group, etc.
- Subject variables are independent variables not manipulated by the experimenter
- Experiments using subject variables are sometimes called *quasi-experiments*

The dependent variable

- The variable that is measured, the outcome of the experiment

- Research question: Are focused words faster to identify than non-focused words?

  - We need measures of speed of word identification
    e.g., lexical decision, naming, reading time

- Important to choose an appropriate method

Extraneous variables

- Variables that are not of interest but which might influence the behavior in *some systematic way* (also called **confounding** variables)
- *A confound **co-varies*** with the independent variable and could provide an alternative explanation of the results
- Confounded studies are uninterpretable → extraneous variables must be controlled (i.e., held constant)

Potential confounds in psycholinguistic research

- Word frequency, word length, word predictability, verb biases, number of words in a sentence, repetition, ambiguity, etc. may affect the participant's behavior

- These variables should be kept constant (by using norming, corpus studies, etc.).

- When you can't hold them constant, make sure they are not associated (confounded) with your IV

Task

- Find the IV and DV
- Think of which other factors could influence the results

Traxler, Bybee and Pickering (1999)

Abstract

"An eye-tracking experiment investigated whether incremental interpretation applies to interclausal relationships. According to Millis and Just's (1994) *delayed-integration hypothesis*, interclausal relationships are not computed until the end of the second clause,[…].

We investigated the processing of *causal* and *diagnostic* sentences […] that contained the connective *because*. Previous research […] has demonstrated that readers have greater difficulty processing diagnostic sentences than causal sentences.

Our results indicated that difficulty processing diagnostic sentences occurred well before the end of the second clause. Thus comprehenders appear to compute interclausal relationships incrementally".

Materials

(1) Heidi could imagine and create things because she won first prize at the art show.

(2) Heidi felt very proud and happy because she won first prize at the art show.

- What factor is manipulated?
- What is measured?
- Predictions?
- Are there any confounds?

Design and predictions

- Factor: type of relationship
  - Two levels: causal, diagnostic
    - NB: to obtain the two readings, the first clause was manipulated
- DV: Eye-tracking measures (reading time)
- Predictions
  - Delayed Integration hypothesis → differences should emerge at the end of the second clause
  - Immediate integration hypothesis → differences should emerge before the end of the second clause

Controlling extraneous variables

- The critical regions are held constant

(1) Heidi could imagine and create things because she won first prize at the art show.

(2) Heidi felt very proud and happy because she won first prize at the art show.

Sturt, Pickering and Crocker (1999)

Abstract

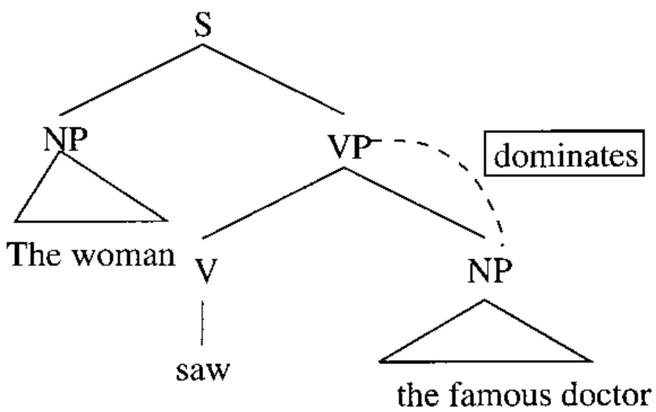
“Many theories of parsing predict that the difficulty of syntactic reanalysis depends on the type of structural change involved. […]

We report two self-paced reading experiments which demonstrate clear differences in the magnitude of garden path effects associated with different types of structural change. However, difficulty of reanalysis was not affected by the position of the head noun within the ambiguous phrase. We interpret these results in terms of theories of structural change such as Sturt and Crocker (1996)”

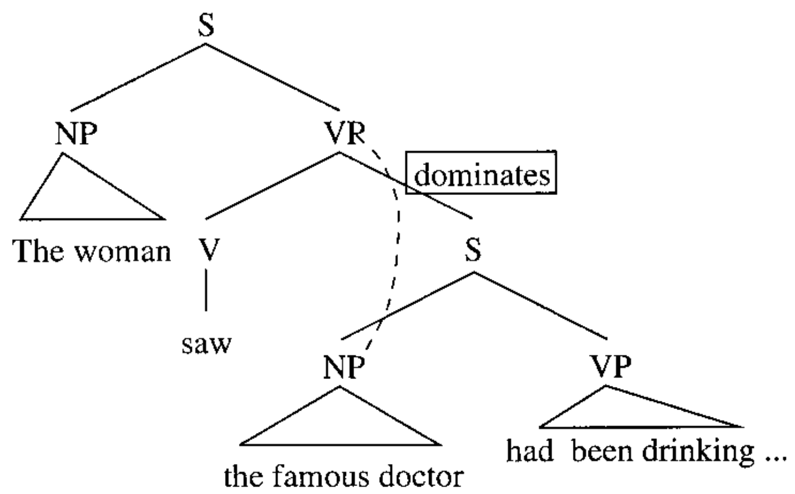
Types of structural change

1) The woman saw the famous doctor had been drinking quite a lot.

Before reanalysis



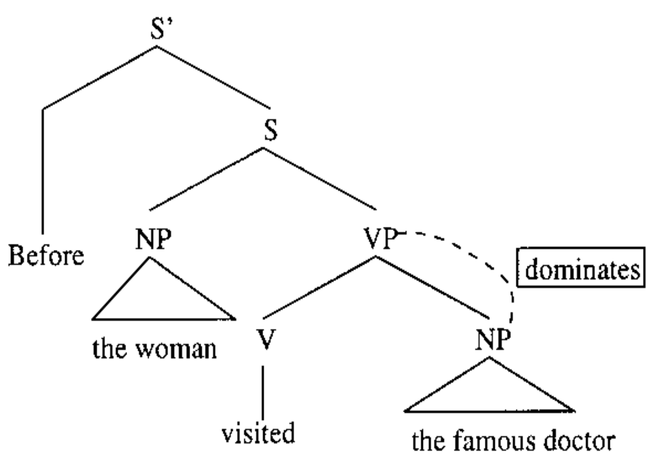
After reanalysis



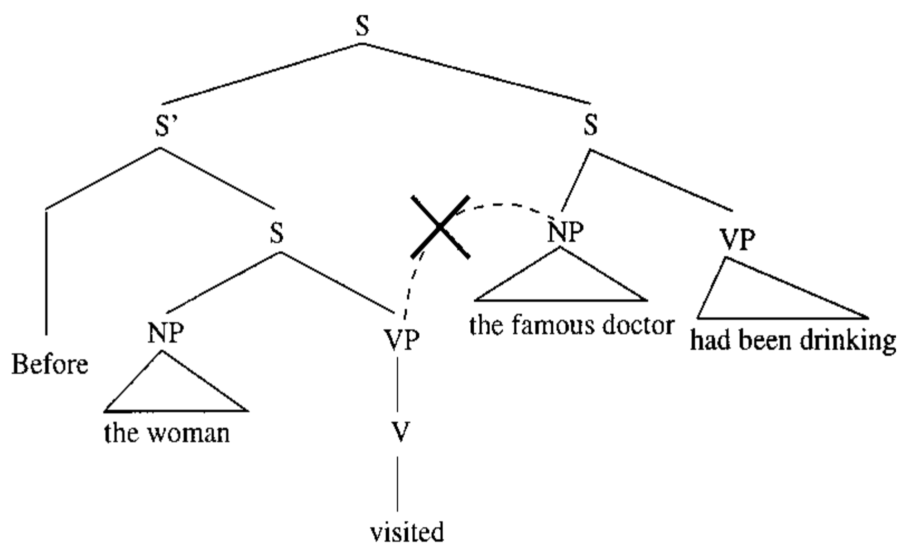
Types of structural change

2) Before the woman visited the famous doctor had been drinking quite a lot.

Before reanalysis



After reanalysis



Items

Ambiguous – NP/S

a) The Australian woman saw the famous doctor had been drinking quite a lot.

Unambiguous – NP/S

b) The Australian woman saw that the famous doctor had been drinking quite a lot.

Ambiguous – NP/Z

c) Before the woman visited the famous doctor had been drinking quite a lot.

Unambiguous – NP/Z

d) Before the woman visited, the famous doctor had been drinking quite a lot.

Pre-tests

- ➤ **Verb bias** was checked in a corpus
  - "Each item had *a verb pair* which was as balanced as possible in terms of the degree to which the NP analysis was preferred over the alternative analyses"

NP Bias of Each Verb Pair in Experiment 1

NP/S verb	NP bias	NP/Z verb	NP bias
understood	.92	negotiated	.94
accepted	.93	polished	.93
recalled	.87	scratched	.91
heard	.89	packed	.89
confirmed	.81	typed	.86
maintained	.98	built	.97
forgot	.89	painted	.94
mentioned	.94	debated	.90
found	.94	lost	.90
announced	.91	investigated	.93
discovered	.71	watched	.68
noticed	.65	knitted	.67
saw	.97	visited	.98
acknowledged	.97	questioned	.99
remembered	.97	attacked	.97
remembered	.97	invaded	.95
read	.99	edited	.98
revealed	.79	washed	.77
revealed	.79	followed	.78
doubted	.83	typed	.86

- ➤ **Plausibility judgment task**
  - to make sure each NP reading was equally plausible

Summary

- **Factors**
  - Ambiguity (two levels: ambiguous, unambiguous)
  - Verb subcategorization properties (two levels: NP/S, NP/Z)

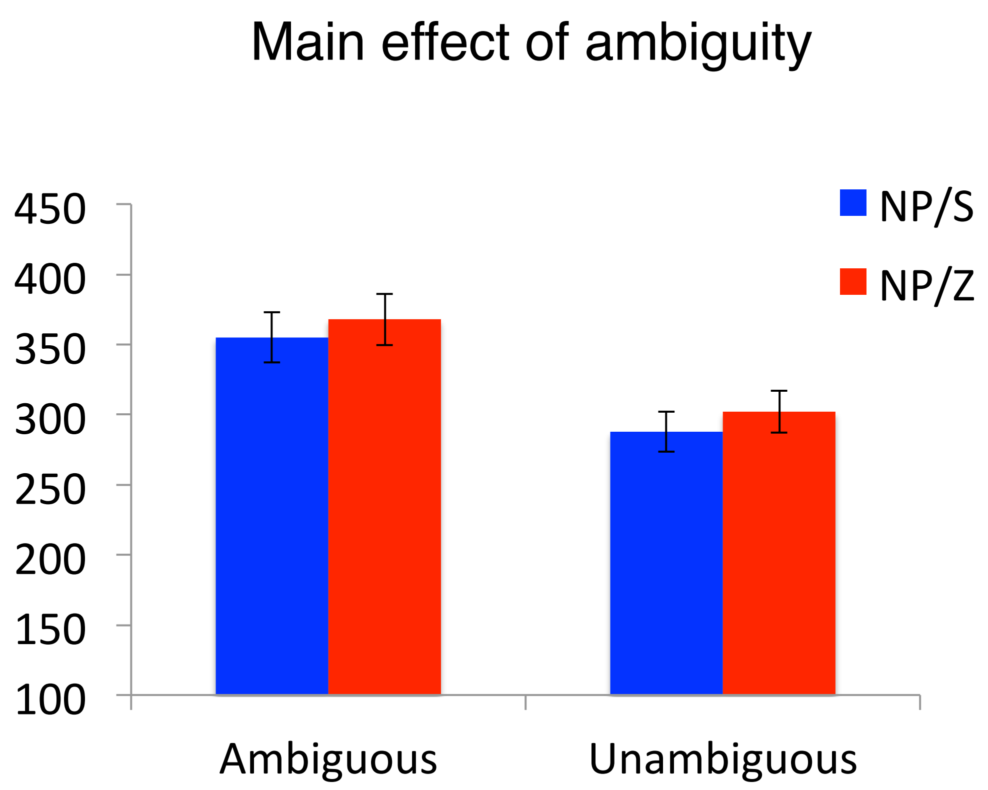
  → **2 X 2 design** = two factors, each one with two levels

- **Task (experimental method)**
  - Self-paced reading

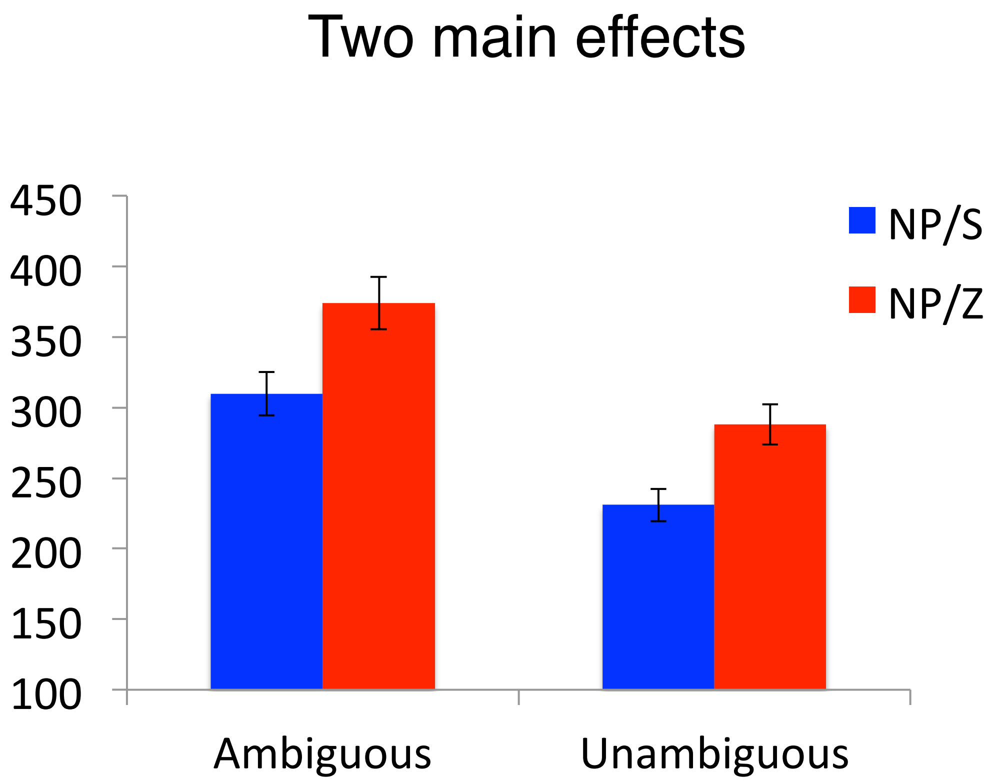
- **Dependent variable**
  - Reading time

- **Controlled variables**
  - Verb biases
  - Plausibility of the misanalysis

Possible outcomes of a 2x2 design

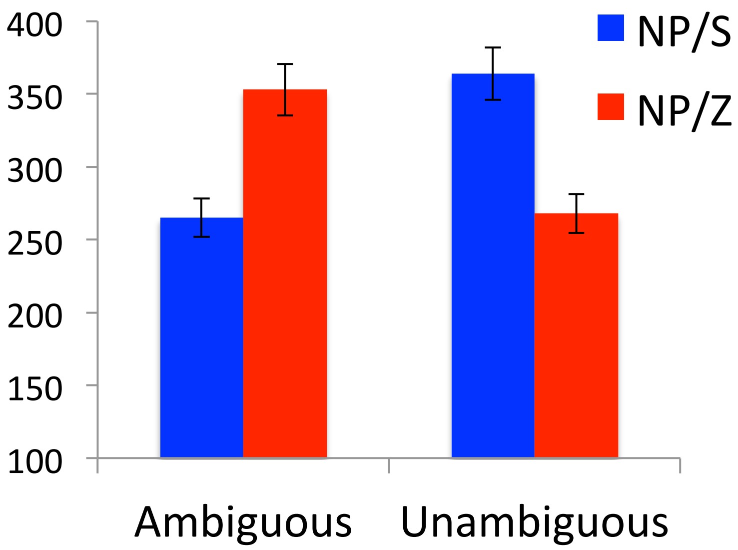


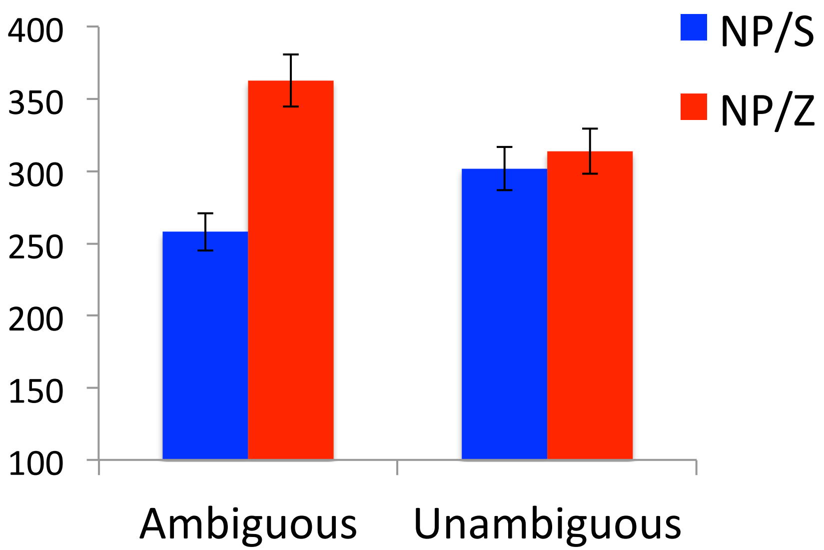
Possible outcomes of a 2 X 2 design



Possible outcomes of a 2X2 design

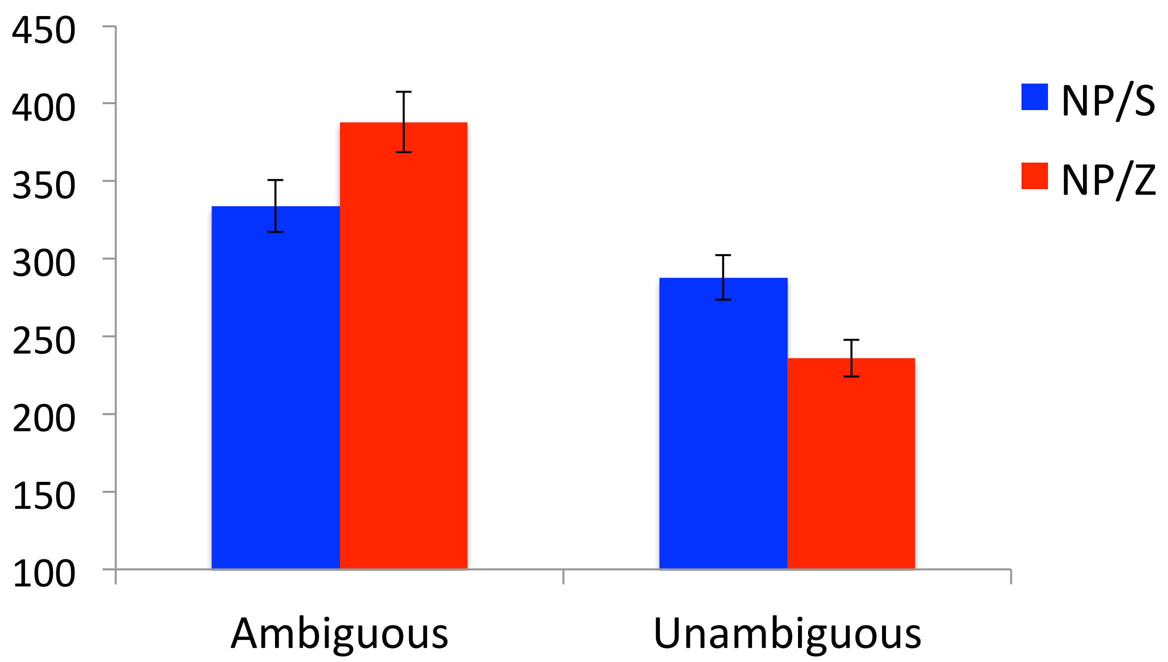
Interactions





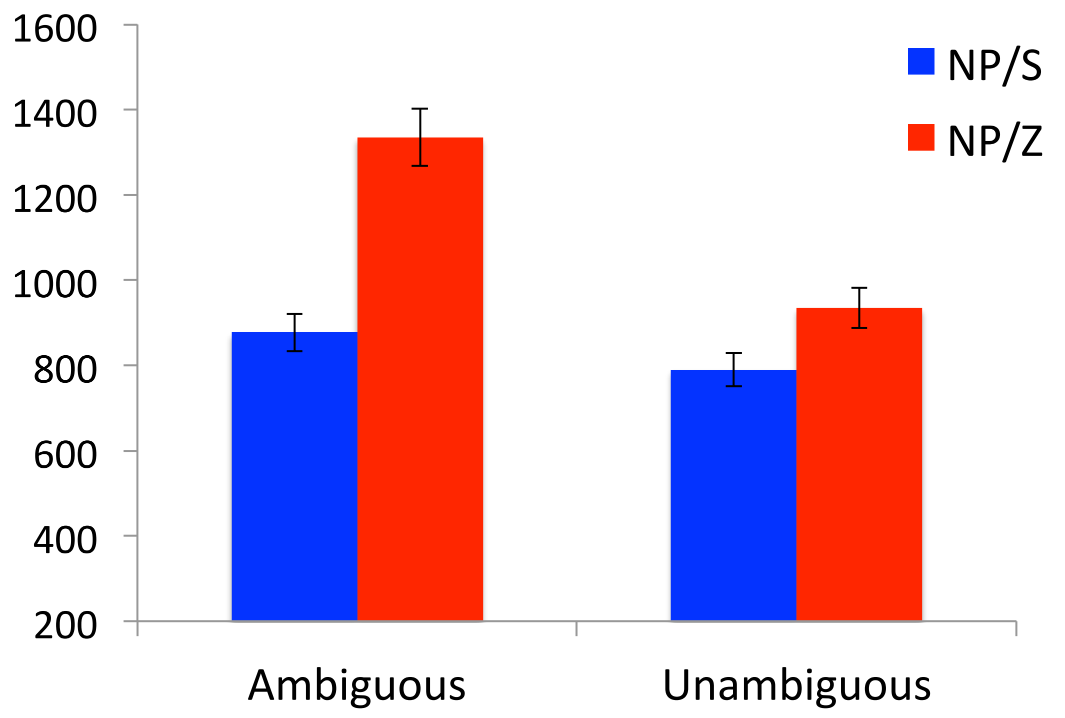
Possible outcomes of a 2X2 design

One main effect and an interaction



Results

The Australian woman saw (that) the famous doctor had been drinking quite a lot.
Before the Australian woman visited(,) the famous doctor had been drinking quite a lot.



Types of design

- **Single factor design**

  Two or more levels/conditions (e.g., word length – levels: 1 syllable, 2 syllables, 3 syllables)

- **Factorial design**

  Two or more factors, each with two or more levels

  - 2X2 design
  - 2X3 design
  - 2X2X2 design
  - 3X2X2X2 design (difficult to interpret!)

- Between or within-subjects?

Between-subjects design

- Each participant is tested in only one condition

Group A	Group B
Cond 1	Cond 2

- **Advantage**
  - Participants are less likely to guess the the purpose of the experiment

- **Disadvantages**
  - Large number of participants needed
  - Differences between conditions could reflect *individual differences* between groups

Individual differences

- Individuals may vary from each other in terms of mood, intelligence, concentration, etc.
- If one group differs from the other with respect to one of these variables, you may no longer be able to say whether the results are due to the manipulation or to differences between groups
- Create *equivalent groups* → groups that are equal to each other in every important way except for the levels of the independent variable
  - Random assignment → very participant should have an equal chance to be included in any group

Within-subject design

- Each participant is tested in each condition (also called *repeated measure design*)

- Advantages
  - More control on individual differences
  - Less subjects needed

- Disadvantages
  - Carry-over effects
  - Participants are more likely to guess the purpose of the experiment

Carry-over effects

- Carry-over effects occur when having been tested under one condition affects how participants behave in another condition

a) The Australian woman saw the famous doctor had been drinking quite a lot.
b) The Australian woman saw that the famous doctor had been drinking quite a lot.

- If you present participants with very similar sentences such as a) and b) (in this order), they may be faster to read b) because they remember a)

- Solution: counterbalancing (Latin square design)

Counterbalancing

- Simple design: n items (sets of sentences) in 2 conditions

	List 1	List 2
Item 1	cond 1	cond 2
Item 2	cond 2	cond 1
Item 3	cond 1	cond 2
Item 4	cond 2	cond 1
...		
Item n	cond 2	cond 1

- Participants are randomly assigned to lists, each participant will see each item in only one condition

Hiding the manipulation

- Include fillers (sentences with different structure)

	List 1	List 2
Item 1	cond 1	cond 2
Filler 1	Filler 1	Filler 1
Item 2	cond 2	cond 1
Filler 2	filler 2	filler 2
Filler 3	filler 3	filler 3
Item 4	cond 1	cond 2
…	…	….

- The number of fillers depend on the experiment (at least twice the number of items)

Summary: General design principles

1. Formulate your question clearly and choose appropriate independent and dependent variable to test it
2. Keep everything constant that you don’t want to vary
3. Know how to deal with unavoidable extraneous variability
4. Use a within-subject design whenever possible
5. Counterbalance your materials

Analyzing data

- Suppose we have designed and carried out an experiment to test the hypothesis that NP/S ambiguous sentences are more difficult to process (slower to read) than NP/S unambiguous sentences

Hypothetical data

Participants	Unambiguous	Ambiguous
1	312ms	325ms
2	365ms	356ms
3	200ms	224ms
4	324ms	388ms
5	356ms	412ms
6	326ms	378ms
7	279ms	299ms
…	…	…
20	323ms	340ms

- What do we do with this data set?

Descriptive statistics

- We first "describe" the data using some measure of central tendency (mean) and variability (variance and standard deviation)

Mean	$\bar{X} = \frac{\Sigma X}{N}$
Variance	$s^2 = \frac{\Sigma(X - \bar{X})^2}{N - 1}$
Standard Deviation	$s = \sqrt{s^2}$

- Variance and Standard Deviation are measures of the dispersion of the data, i.e., how individual data points are distributed around the mean

Hypothetical data

Unambiguous	Ambiguous
312ms	325ms
365ms	356ms
200ms	224ms
324ms	388ms
356ms	412ms
326ms	378ms
279ms	299ms
…	…
323ms	340ms

$$\bar{X}_{Un} = \frac{\Sigma X}{N} = \frac{313 + 365 + \ldots + 323}{20} = 320$$

$$\bar{X}_{Am} = \frac{\Sigma X}{N} = \frac{325 + 356 + \ldots + 340}{20} = 350$$

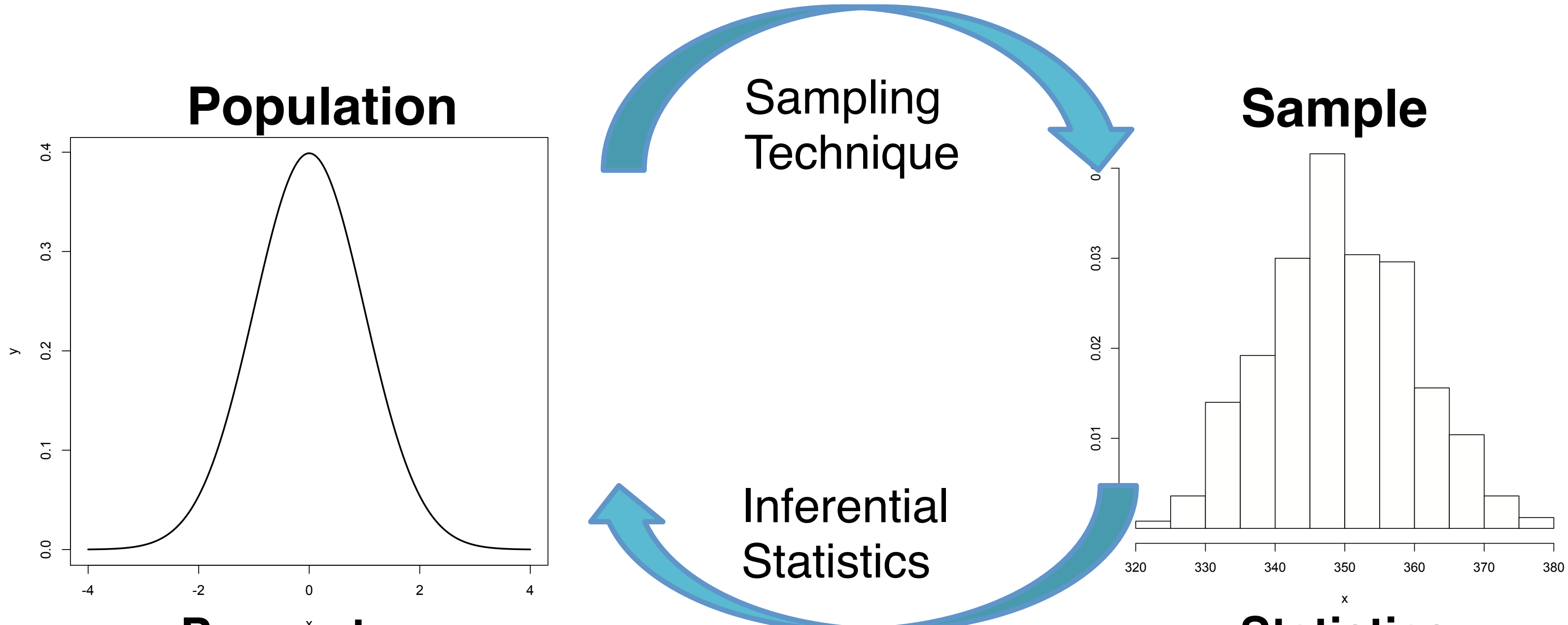
$$s_{Un} = \sqrt{\frac{(312-320)^2 + \ldots. + (323-320)^2}{20-1}} = 48$$

$$s_{Am} = \sqrt{\frac{(325-350)^2 + \ldots. + (340-350)^2}{20-1}} = 55$$

- Does the 30ms difference reflect an effect of the manipulation?

Population versus Sample

Population | **Sample**



Parameters

Mean (μ)
Standard Deviation (σ)
etc.

Statistics

Mean (x)
Standard Dev (s)
etc.

Sampling error and bias

- An estimate of a population parameter (e.g., the sample mean) is likely to be different for different samples
- Each estimate is likely to be different from the population parameter
- This variability (error) is due to
  - Chance (**sampling error**)
  - Selection of non-representative samples (**sampling bias**)
- Sampling error and bias can be reduced by using an appropriate sampling method (probability sampling)

Sampling in psycholinguistics

- In psycholinguistic experiments we sample from:

1) Speakers

- we use inferential statistics to generalize the results to the population of all speakers of a language

2) Language

- We use inferential statistics to generalize to the entire collection of linguistic items displaying a certain property (not just the items we use in the experiment)

Hypothetical data

Unambiguous	Ambiguous
312ms	325ms
365ms	356ms
200ms	224ms
324ms	388ms
356ms	412ms
326ms	378ms
279ms	299ms
…	…
323ms	340ms

$\bar{X}_{Un} = 320$ $\bar{X}_{Am} = 350$

$s_{Un} = 48$ $s_{Am} = 55$

- Does the 30ms difference reflect a true difference between the population means, or is it just due to chance?
- Is the difference significant?

Statistical Hypotheses Testing

- First step to test whether a difference is significant is to make the assumption that it is not (i.e., it is just due to chance)

Null hypothesis (H_0)

$$\mu_{Amb} = \mu_{Unamb} \Rightarrow \mu_{Amb} - \mu_{Unamb} = 0$$

- The research hypothesis states the outcome of your experiment reflects a true difference (i.e., it is due to the manipulation)

Alternative hypothesis (H_1).

$$a)\ \mu_{Amb} \neq \mu_{Unamb} \Rightarrow \mu_{Amb} - \mu_{Unamb} \neq 0$$ Two-tailed hypothesis

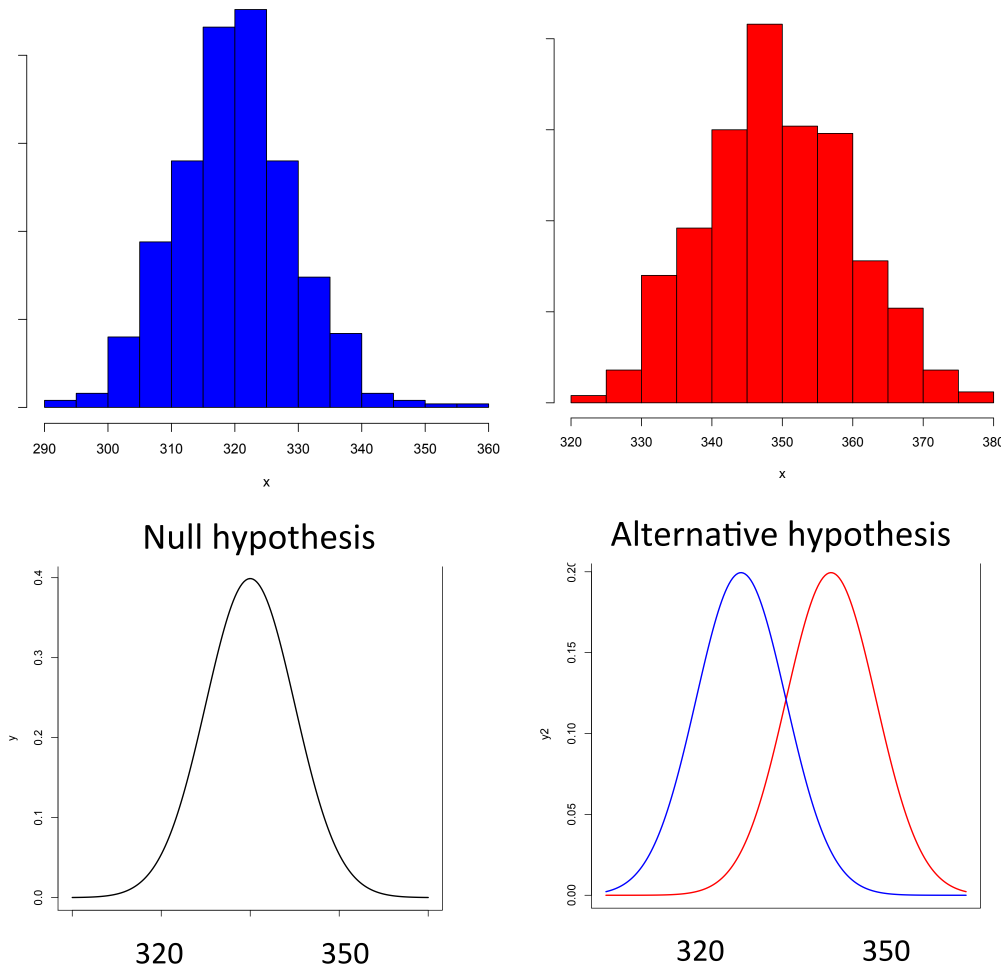
$$b)\ \mu_{Amb} - \mu_{Unamb} > 0$$ One-tailed hypothesis

Hypothetical data

Unambiguous	Ambiguous
312ms	325ms
365ms	356ms
200ms	224ms
324ms	388ms
356ms	412ms
326ms	378ms
279ms	299ms
…	…
323ms	340ms

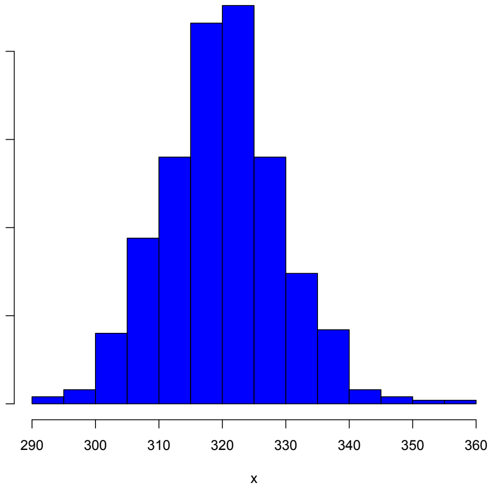
$\bar{X}_{Un} = 320$ $\bar{X}_{Am} = 350$

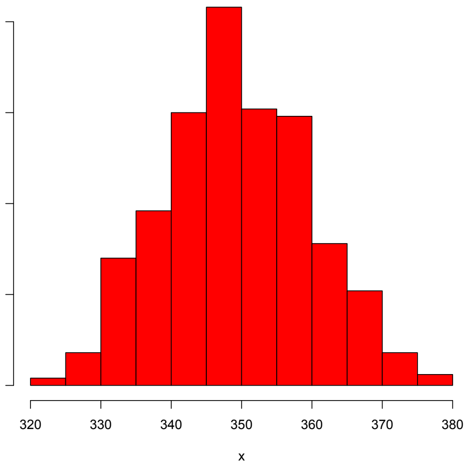
$s_{Un} = 48$ $s_{Am} = 55$



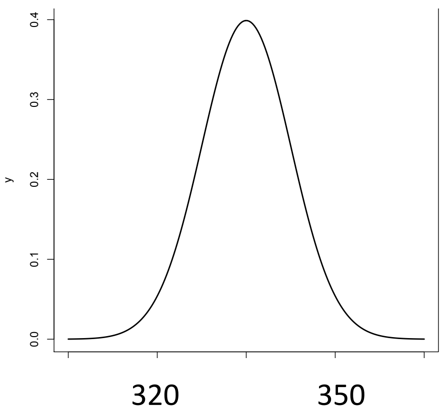
Hypothetical data

- **Statistical tests** (e.g., t-test, ANOVA) will tell you how likely it is to observe your data assuming the null hypothesis is true
- If this probability is low, you can reject the null hypothesis
- The difference is significant

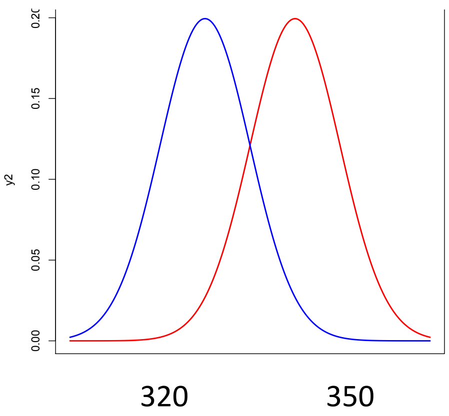




Null hypothesis



Alternative hypothesis



Variability

- There are two general types of variability in the data:

a) **Systematic**
   - the result of some identifiable factor (either the variable of interest or some factor that you've failed to control adequately)

b) **Error**
   - nonsystematic variability due to individual differences within and between groups and any number of random, unpredictable effects

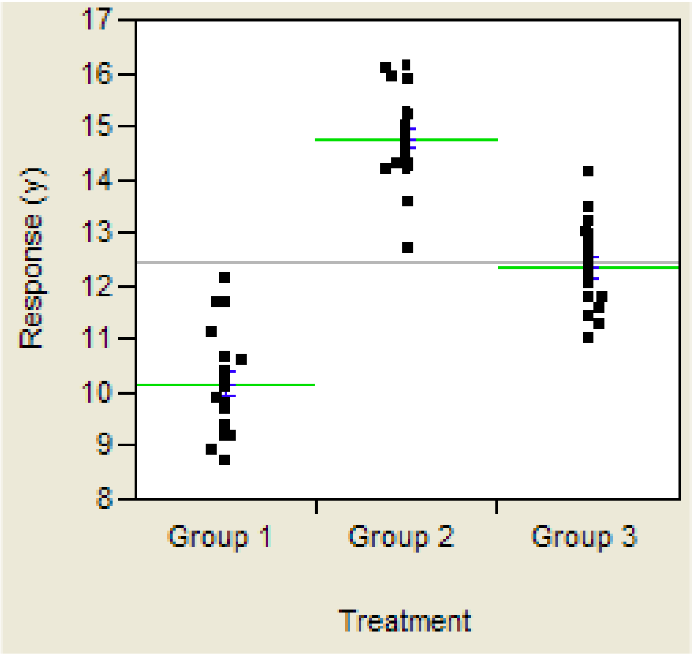
Statistical tests

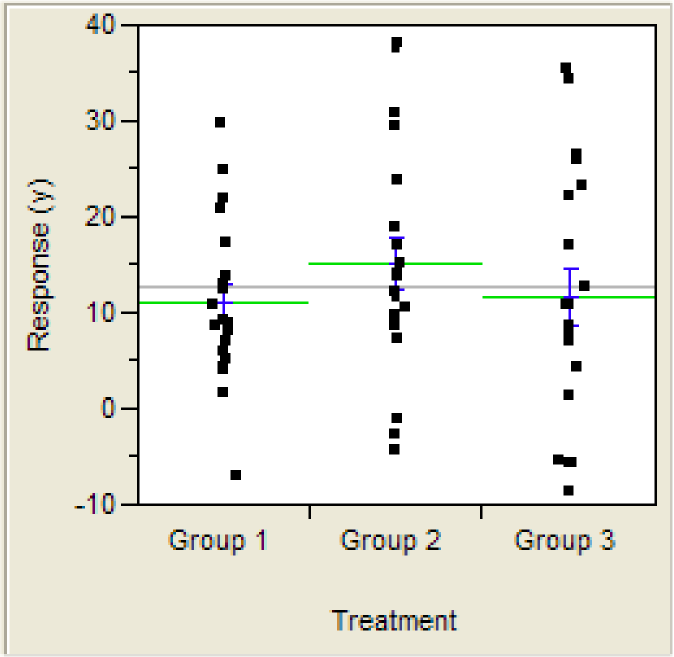
- Most statistical tests calculate a ratio that takes into account two sources of variability

$$statistic = \frac{Variability\ between\ conditions\ (systematic\ +\ error)}{Variability\ within\ conditions\ (error)}$$

- If the variability between conditions is huge and the variability within condition is relatively small => the difference between conditions is likely to be significant

Variability between and within





The t-test

- Can be used to test whether the difference between two means is significant

- Simplified formula for a repeated (dependent) measures design

Variability between conditions

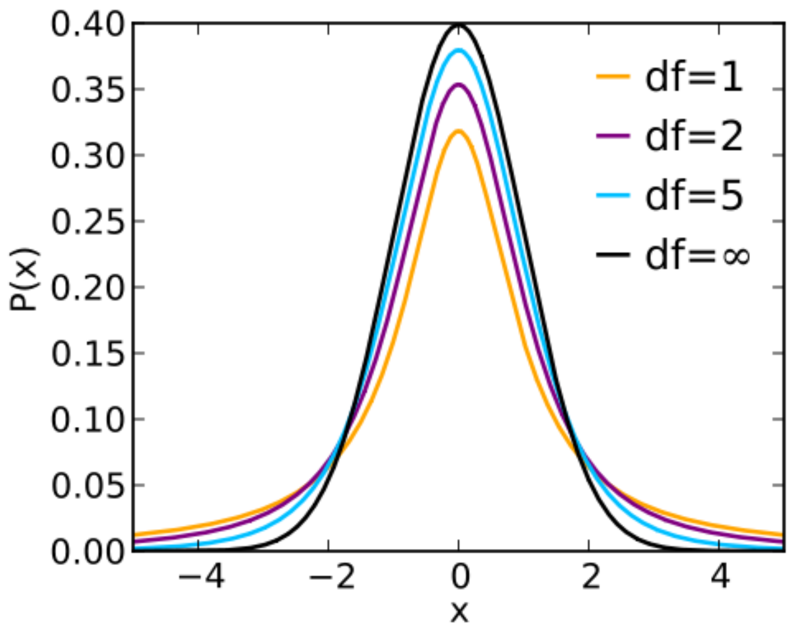
$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_{X_1 - X_2} / \sqrt{N}}$$

Variability within conditions

- The *t* statistics follows the t-distribution

T-distribution

- Continuous probability distribution



- The shape of the distribution depends on the number of **degrees of freedom** (***df***)

- As ***df*** go to infinity, the t-distribution converges to the standard **normal distribution**

Degrees of freedom (df)

- *df* are used to provide a more accurate estimate of the parameters of a population; the number of *df* is a function of both the sample size and the number of parameters estimated

- Defined as the number of values in the calculation of a statistic that are free to vary

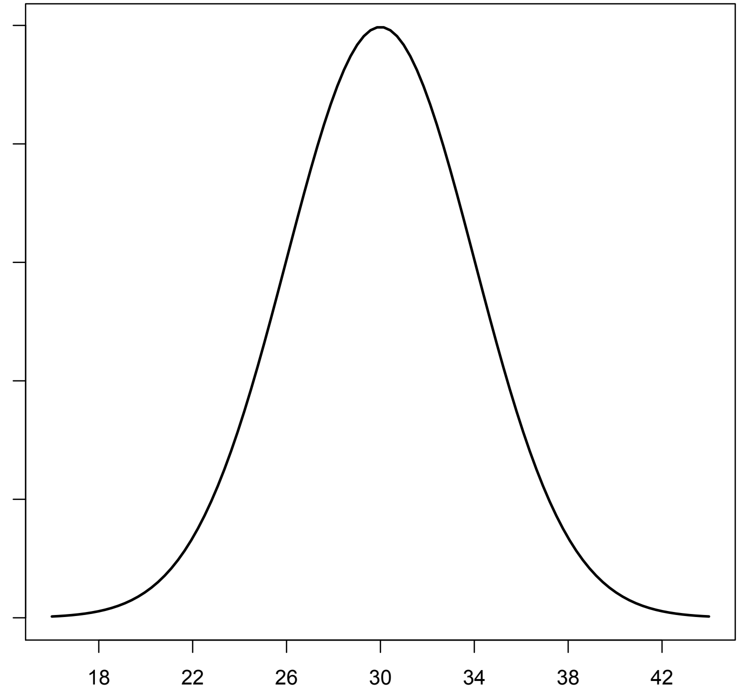
- Imagine you have four numbers (a, b, c and d) that must add up to a total of m; you are free to choose the first three numbers at random, but the fourth must be chosen so that it makes the total equal to m - thus your *df* is three

The normal distribution

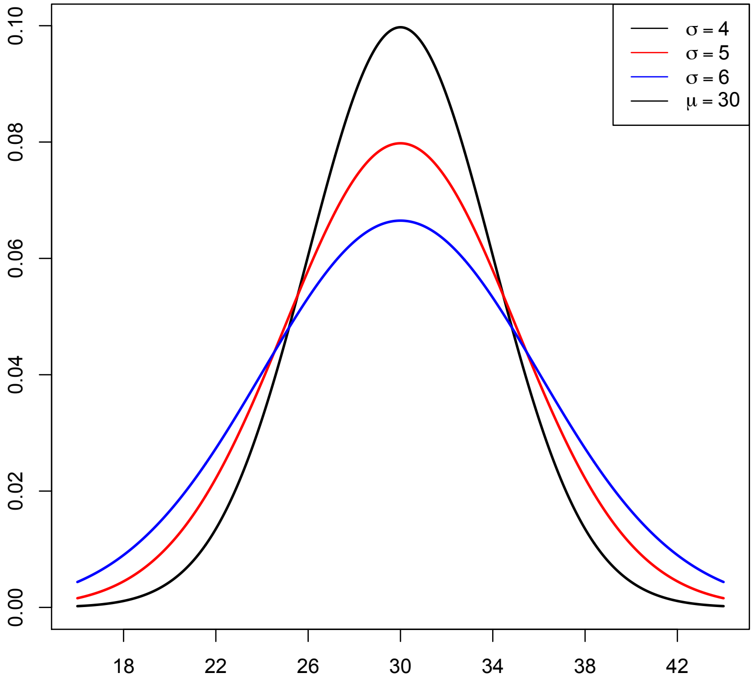
- A probability density function, symmetrical about the mean, bell-shaped, described by μ and σ

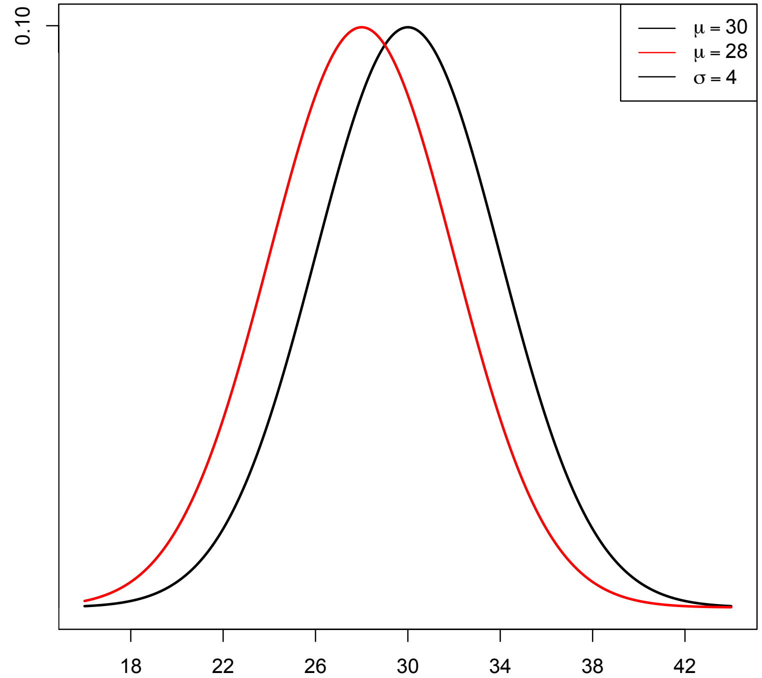
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$-\infty < x < +\infty$$



The normal distribution

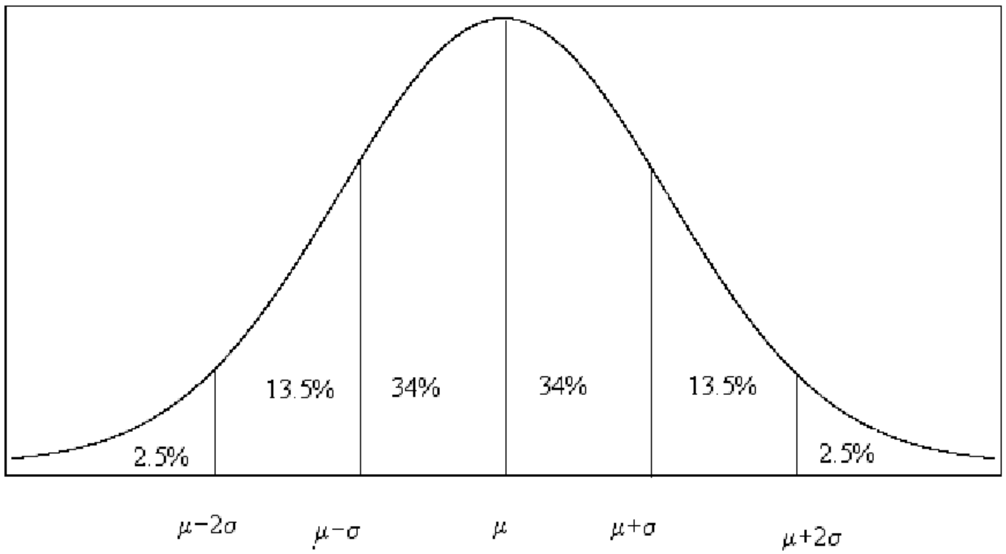




The standard normal distribution has $\mu = 0$ and $\sigma = 1$

The normal distribution

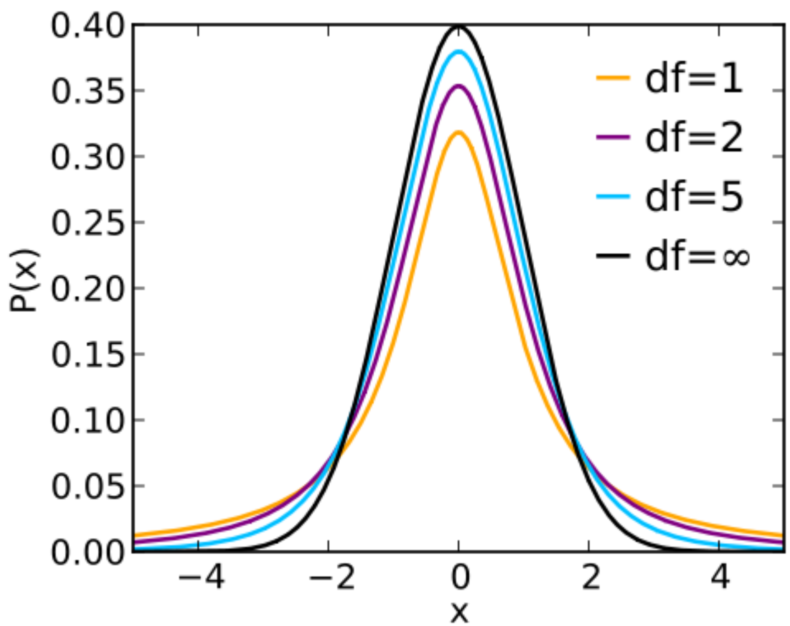
- The area under the normal curve is equal to 1



$$P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 0.68 \quad P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.95$$

T-distribution

- Continuous probability distribution



- The shape of the distribution depends on the number of **degrees of freedom** (***df***)
- As ***df*** go to infinity, the t-distribution converges to the standard normal distribution
- Intuitively, the t-distribution represents the distribution of possible t-values if the null hypothesis is true

Hypothetical data

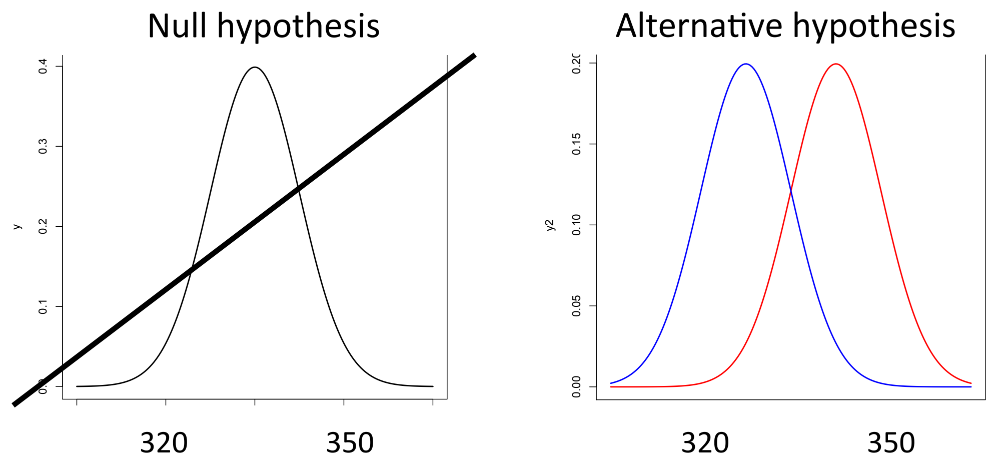
Unambiguous	Ambiguous
312ms	325ms
365ms	356ms
200ms	224ms
324ms	388ms
356ms	412ms
326ms	378ms
279ms	299ms
…	…
323ms	340ms

$\bar{X}_{Un} = 320$ $\quad \bar{X}_{Am} = 350$

$s_{Un} = 48$ $\quad s_{Am} = 55$

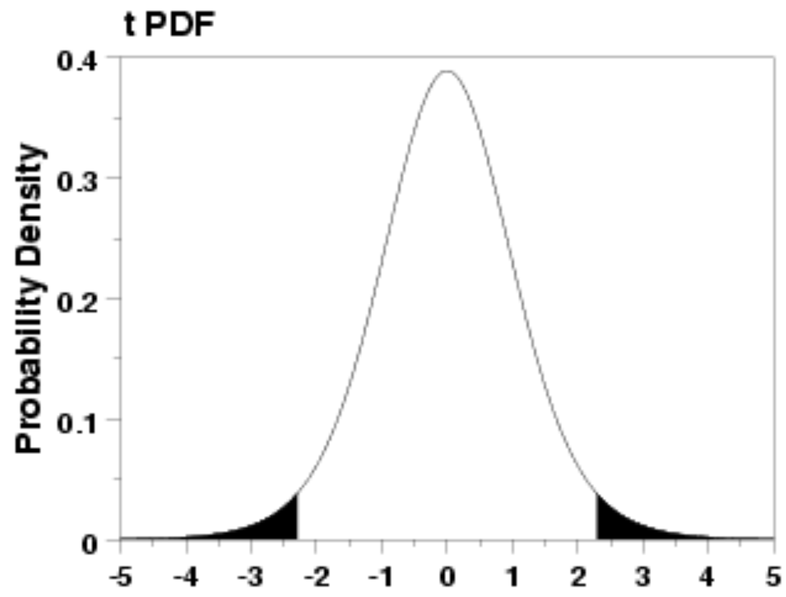
$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_{X_1 - X_2} / \sqrt{N}} = \frac{20}{S_{X_1 - X_2} / \sqrt{20}} = 2.3$$

- If the probability of observing a t statistics (at least as large) as 2.3 under the null hypothesis is low, we can reject the null hypothesis



Significance level

t = 2.3;



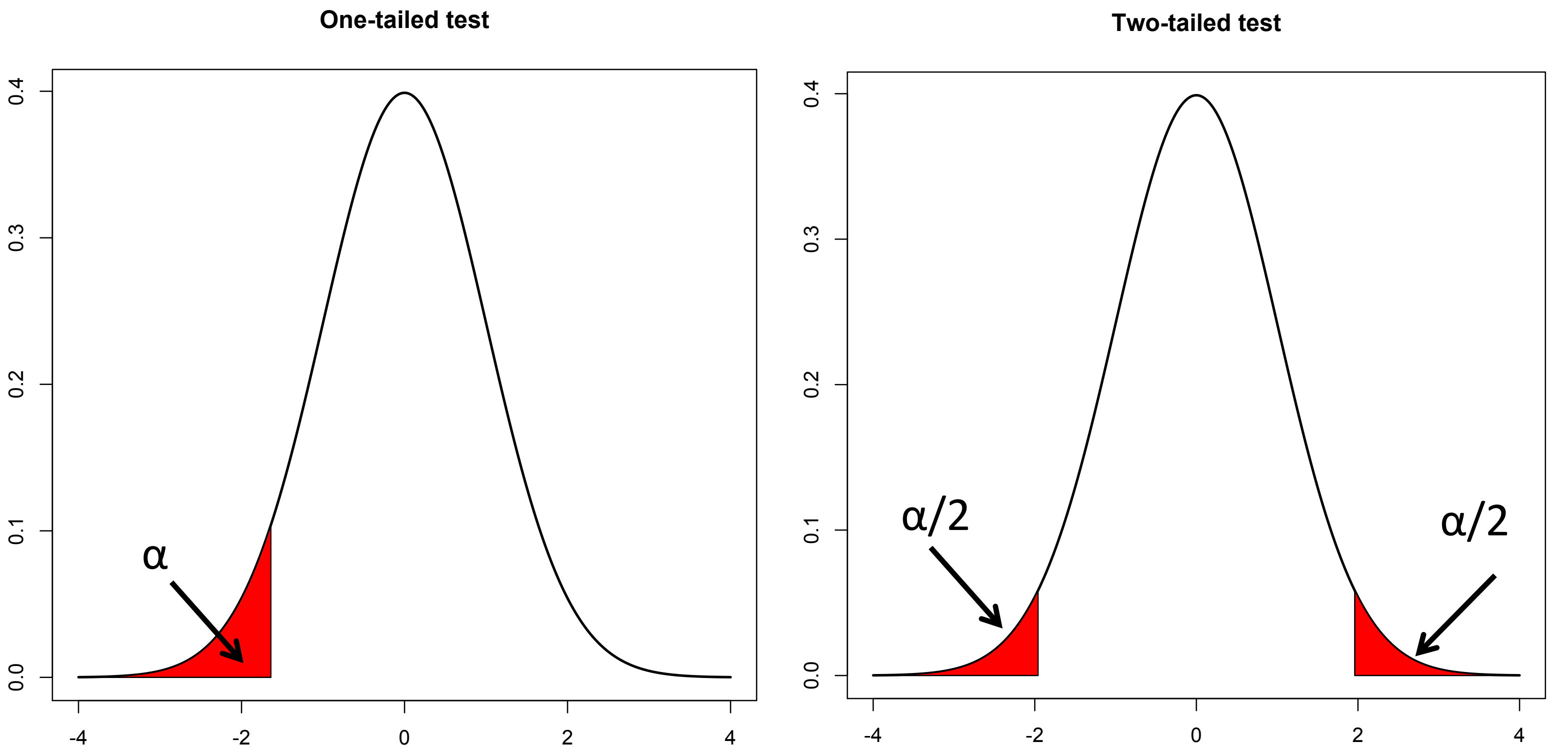
t-distribution with 19 *df*

- would you reject the null hypothesis?

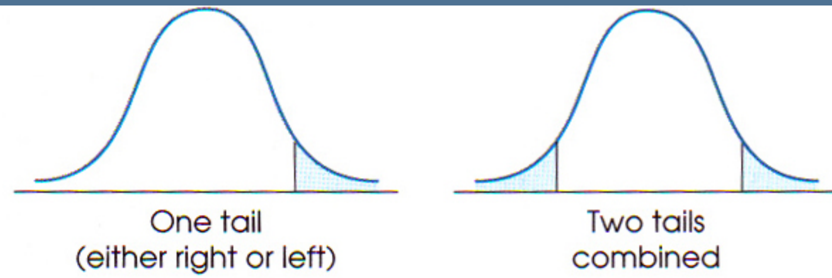
Significance level: α

- Alpha (α) is an arbitrary cutoff value representing the probability with which we are willing to reject H_0 when it is, in fact, correct.

  α levels convenyionally used: 0.05, 0.01



The t-table



	PROPORTION IN ONE TAIL					
	0.25	0.10	0.05	0.025	0.01	0.005
	PROPORTION IN TWO TAILS COMBINED					
df	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
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Hypothetical data

Unambiguous	Ambiguous
312ms	325ms
365ms	356ms
200ms	224ms
324ms	388ms
356ms	412ms
326ms	378ms
279ms	299ms
…	…
323ms	340ms

$\bar{X}_{Un} = 320$ $\bar{X}_{Am} = 350$

$s_{Un} = 48$ $s_{Am} = 55$

- Calculate t statistics
- Choose the alpha level (e.g., .05)
- If $|t| > t_\alpha$ => $p < .05$ => reject H_0
  - The difference is significant
- If $|t| \leq t_\alpha$ => $p \geq .05$ => *fail to reject* H_0
  - Null result

Possible outcomes

- A statistical test yields only two results:
  - **Reject H0** => you believe that an effect truly happened in your study and that the results can be generalized
    - You find a significant result
  - **Fail to reject H0** => the difference in the means is most likely due to chance
    - You find a null result

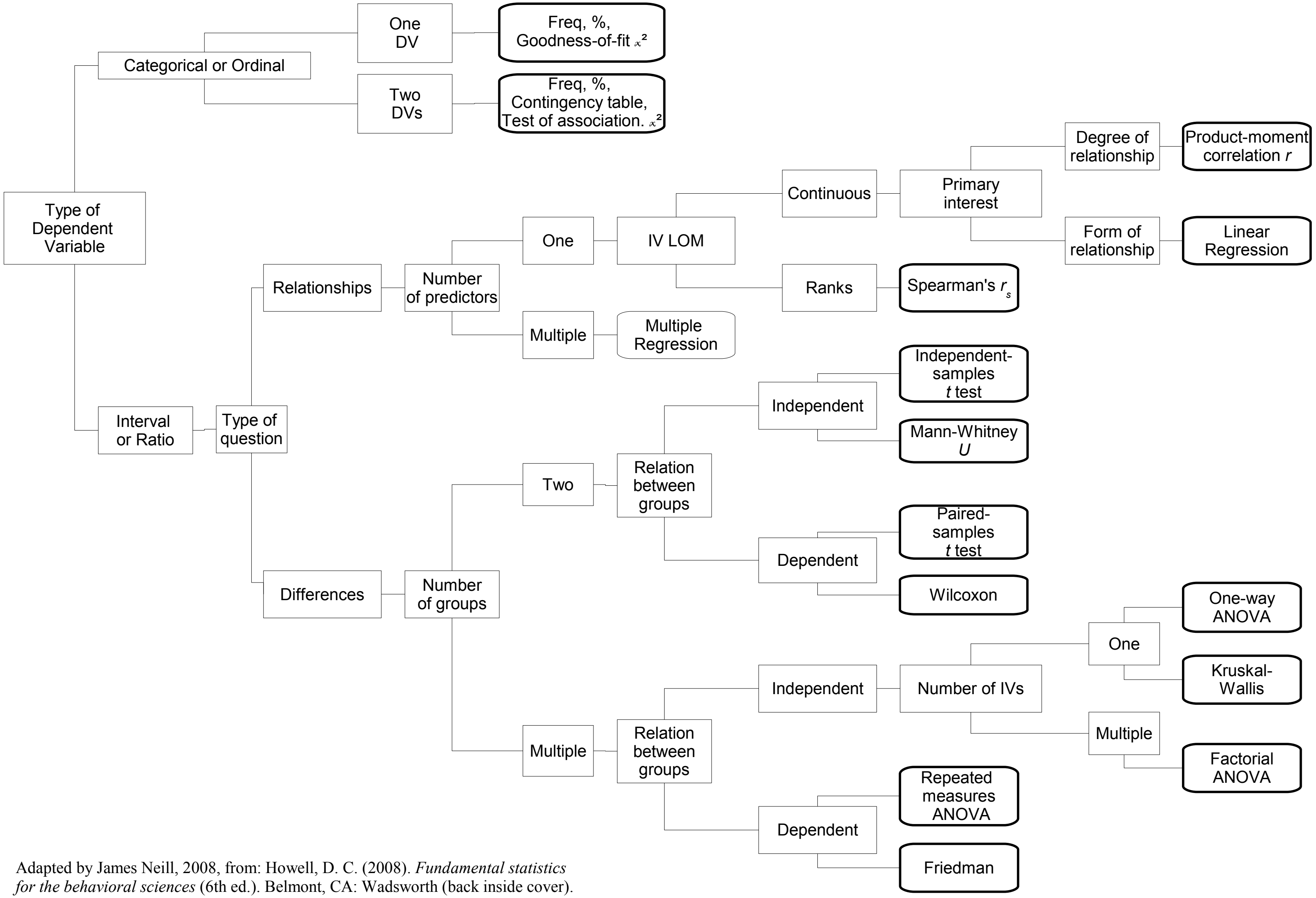
Errors

- A significant result does not prove that H_1 is true
  - If α=.05, you have a 5% chance of rejecting H_0 when it is in fact true
  - **Type I error** (false positive) → reject the H_0 when it is in fact true
- A null result does not prove that H_0 is true
  - **Type II error** (false negative) → fail to reject the null hypothesis when it is in fact false

Statistical tests

- What kind of statistical test should be used - e.g., t-test, ANOVA (F distribution), χ^2-test - depends on:
  - The type of data (Categorical vs. Continuous)
  - The assumed underlying distributions (normal, binomial, etc.)
  - Number of IVs
  - Whether the design is between- or within-subjects
- In psycholinguistics, statistical analyses are performed by subjects (e.g, t_1, F_1) and by items (t_2, F_2)

Decision tree



Adapted by James Neill, 2008, from: Howell, D. C. (2008). *Fundamental statistics for the behavioral sciences* (6th ed.). Belmont, CA: Wadsworth (back inside cover).

Summary

- State the null (H_0) and alternative hypothesis (H_1)
- Sample from a population (collect data)
- Describe the data and calculate an appropriate test statistics
- Choose an alpha level
- Make a decision (reject H_0 or fail to reject H_0)