Introduction to Psycholinguistics

Lecture 6

Experimental Methods II

Pia Knoeferle & M. W. Crocker

Department of Computational Linguistics

Saarland University

SS 2006

Overview

- Homework
- Exploring the data
  - Quantitative data: e.g., reading times
    - Bargraphs of means & confidence intervals
    - Boxplots
    - Histograms: Skew and kurtosis
    - Testing for normality and homogeneity of variance
- Inferential statistics
  - Parametric tests
    - Comparing two means: *t*-test
    - Comparing more than two means: *F*-statistic
  - An example from the eye-tracking literature

Homework

- Design an experiment
  - Theory1: There is a processing preference (e.g., subject-first) for both ambiguous and unambiguous sentences
  - Theory 2: Such a preference exists only for ambiguous sentences
    - Operationalization, hypotheses, design + example sentences, and lists (only the condition coding per list); method
    - How many factors?
    - Assume 24 items
    - How many data points per condition for 1 participant?
    - Type of data and analysis?

Homework

- Operationalization
  - If information later in the sentence (e.g., NP2) disambiguates a sentence-initial ambiguous NP, we should observe processing difficulty
    - *Hypothesis0*: Such difficulty should be observed for both initially structurally ambiguous and unambiguous sentences
    - *Hypothesis 1*: Such difficulty should only be observed for initially structurally ambiguous sentences
- Method
  - Eye tracking (self-paced reading would also be possible)
- Your independent variables are …
  - *Word order* (SVO vs. OVS) & *ambiguity* (ambiguous vs. unambig.)
- Your dependent variable is …
  - Reading times in a word region

Homework

- Design
  - (1a) Die Mutter verabschiedet den Besucher nach der Party.
  - (1b) Die Mutter verabschiedet der Besucher nach der Party.
  - (2a) Der Vater verabschiedet den Besucher nach der Party.
  - (2b) Den Vater verabschiedet der Besucher nach der Party.
- Control
  - Plausibility, e.g., pretest in form of plausibility ratings on a scale from 1 (very implausible) to 7 (highly plausible)
  - Word length (+/-2chars)
  - Frequency of lemmas (e.g., *Celex*)

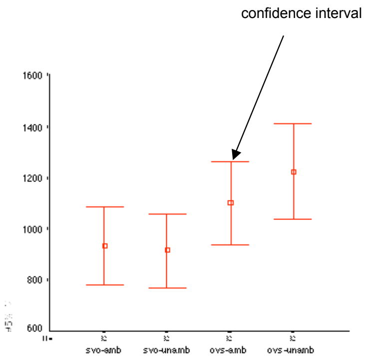
Homework

- Lists
  - For a 2x2 design with 2 levels for each factor, there are 4 exp. lists
  - One participant sees one list
  - Latin Square to ensure that there is for each list
    - Equal number of trials in each condition (24 items/4 conds: 6)
- Conducting the experiment
- Analysing the data to find out whether our manipulation (ambig. vs. unambig.) had an "effect"?
  - Exploring the data
  - Inferential statistics

Item	List1	List2	List3	List4
1	a	b	c	d
2	b	c	d	a
3	c	d	a	b
4	d	a	b	c
5	a	b	c	d
6	b	c	d	a
7	c	d	a	b
8	d	a	b	c
9	a	b	c	d
…	…	…	…	…
21	a	b	c	d
22	b	c	d	a
23	c	d	a	b
24	d	a	b	c

Exploring the data

- Quantitative data
  - Compare mean reading times
    - Bar graphs with *confidence intervals* (CI): 95% CIs
      - CIs indicate the range within which we expect the true value of the mean will fall
      - 95% of the mean values in our population fall between the range indicated by the confidence intervals
    - So what does a narrow confidence interval indicate?
      - The sample mean is close to the true mean
      - Wide confidence interval: mean could be very different from true mean

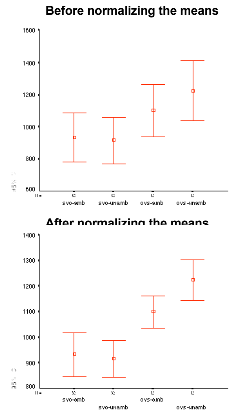


Mean regression path duration on NP2

Exploring the data

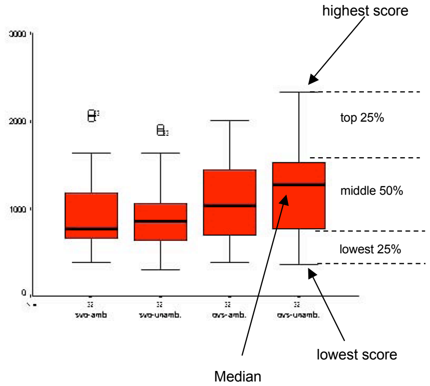
- Error bar graphs for repeated measures design
  - Stats programs treat data as if from diff. groups
  - Solution
    - Eliminate between-subjects variability
    - Normalize participants' means
    - All participants have same mean across conditions

1. Calculate mean time for each part. across conditions
2. Compute grand mean of all the participants' means
3. Calculate adjustment factor: adjust = grand mean - participant means
4. Create adjusted values for each variable: Var. + adjust



Exploring the data

- Boxplots (box-whisker diagrams)
  - Quartiles
    - Top/bottom quartile
      - Range between which top/lowest 25% of scores fall
    - Interquartile range
      - Range in which the middle 50% of the scores fall
    - Median
      - Middle score if you arranged the reading times in order (≠ mean)
  - Looking for outliers

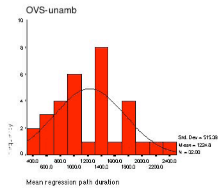


Assumptions about the data

- If we ultimately wanted to do more than just descriptively explore the data
  - We need to decide which test to use
- For our data (reading times) we typically use *parametric tests*
  - Parametric tests are based on the normal distribution
  - There are certain requirements for performing parametric tests
    - The data
      - Must be at least interval-scale data
      - Must be normally distributed
      - Variances in populations/groups/conditions roughly equal (*homogeneity of variance*)
    - Test for independent (between-subjects) design in addition assume
      - Scores that we compare are independent (i.e., from different people)
  - So we need to check first whether our data meets these requirements

Exploring the data: skew and kurtosis

- In a normal distribution, skew (lack of symmetry) and kurtosis (pointyness) should be zero
  - Positive values of skewness means left-skewed
  - Negative skewness values indicate right-skewed
  - Positive kurtosis values indicate a pointy distribution
  - Negative kurtosis indicates a flat distribution
- The further the skewness/ kurtosis values from zero, the more likely it is that the data are not normally distributed
  - Actual values for skew/kurtosis not informative
  - z-transformation $z_{skewness} = \frac{S - 0}{SE_{skewness}}$ $z_{kurtosis} = \frac{K - 0}{SE_{kurtosis}}$



Moments

N	Valid	32
	Missing	0
Mean		1224.8125
Std. Error of Mean		91.10763
Median		1276.7500
Mode		373.50(a)
Std. Deviation		515.38260
Variance		265619.22177
Skewness		.312
Std. Error of Skewness		.414
Kurtosis		-.616
Std. Error of Kurtosis		.809
Range		1962.00
Minimum		373.50
Maximum		2335.50

a Multiple modes exist. The smallest value is shown

Middle score if you ordered the obtained scores (Median)

Score that occurs most frequently in data set (Mode)

Z_s = .312/.414 = 0.755
Z_k = -.616/.809 = -0.761

Testing for normality

- Kolmogorov-Smirnov test for normality
  - Should you test the data overall or rather for each condition?

Tests of Normality

	Kolmogorov-Smirnov(a)			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
OSOS	.100	32	.200(*)	.971	32	.529

* This is a lower bound of the true significance.
a Lilliefors Significance Correction

- If the result of the K-S test are significant you cannot perform a parametric test on that data
  - Transform the data
    - E.g., log transformations squash the right tail of the distribution, and can reduce a positive skew

Testing for homogeneity of variance

- For between-subject designs
  - Levene's test
- For repeated measures
  - Sphericity assumption in repeated measures analysis of variance (ANOVA)
- Once we have explored the data in this way
  - And are sure they meet the assumptions of parametric tests
    - We can test differences between the means using *inferential statistics*

Statistical tests

- Which test should we chose?
- We distinguish between parametric and non-parametric tests
  - Parametric tests
    - For data that are based on the normal distribution (e.g., interval scale and above)
    - *T*-Test: For 1-factor designs with 2 levels
    - Analysis of Variance (ANOVA)
      - Can test the independent effect of a factor: *main effect*
      - Can test for *interactions* (relationships between effects)
  - Non-parametric tests
    - Do not assume the data are from a normal distribution (e.g., for categorical data)
      - Chi-square test
      - Log-linear models
  - For our data (inspection duration) we use parametric tests

Statistical tests

- So, how do test statistics "work"?
- Two types of variance for both dep./indep. designs
  - *Systematic variation*: result of experimental manipulation
    - E.g., SVO vs. OVS sentence condition
  - *Unsystematic variation*: variation due to random factors: e.g., age, gender
- Test statistics
  - Discover how much variation there is in performance
  - How much of this variation is *systematic* versus *unsystematic*
  - Is there more variation than without the experimental manipulation?

Data collection and variation

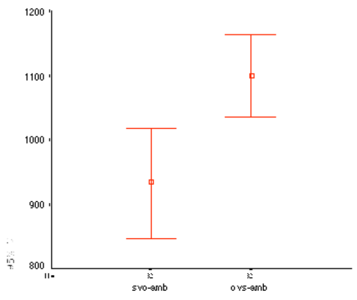
- In our decision tree, why do we get a distinction between tests for "dependent" and "independent" data collection?
- Unsystematic variation in data differs depending on the type of data collection
  - Within-subjects (dependent) design
    - One participants receives all conditions
    - So other factors (e.g., age, IQ etc.) are constant across conditions
  - Between-subjects (independent) design
    - Even in the absence of an experimental manipulation, we would find differences between the groups since these contain different participants that differ in gender, IQ, age, etc.
- Repeated measures designs are good at detecting true effects
  - Why?
    - Unsystematic variation ('noise') is kept to a minimum

Minimize unsystematic variation

- In both types of design: minimize unsystematic variation
  - Randomization: eliminates sources of systematic variation other than our manipulation
    - Repeated-measures
      - Practice effects: after 10 OVS sentences, they become easy
      - Boredom effects
    - Solution
      - Ensure that these effects produce no systematic variation between our conditions
      - Counterbalance the order in which a person participates in a condition
    - Independent designs
      - Confounding factors contribute to variation (e.g., age, IQ),
      - But: ensure they contribute to unsystematic, not systematic, variation
    - Solution
      - Allocate participants randomly to an experimental condition

Comparing two means

- Let's assume for a first test that we had an experiment with only 2 conditions (1 factor, 2 levels)
  - SVO and OVS ambiguous
  - Effect of independent variable 'sentence type' on reading times
    - Error bars for regression path duration on NP2
    - It looks as if the ovs-amb. mean is much higher than the svo-amb. mean
  - Test: Comparing two means
  - Is the difference due to **chance** (e.g., noise) or our **experimental manipulation**?
  - Statistical tests provide us with a probability (*p*) that the difference is genuine (and not due to chance)



T-Test

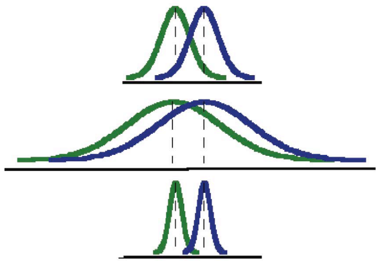
- Comparing means between two groups/conditions
  - Let's look at a simple test statistics: *T*-Test
  - Independent means t-test
    - When there are two conditions and different participants assigned to each condition (*independent measures/samples t*-test)
  - Dependent means t-test
    - Same participants took part in both conditions *(matched-pairs/paired-samples t*-test)
- We have collected data and calculated the means
- If from the same population, the means should be roughly equal
  - H0: experimental manipulation has no effect on participants, and sample means should be very similar
    - I.e., mean reading time for SVO-amb. is similar to OVS-amb.
  - Means might differ by chance
    - But: large differences should occur infrequently by chance

T-Test

- Compare difference between obtained sample means to difference between means that you would expect by chance
  - That means you need a measure of two things
    - How different the observed difference between your sample means is from the difference that you would expect in population means (if H0 is true this second diff. would be 0)
  - We further need a measure of *unsystematic* variation (i.e., noise that we would get by chance)
  - We need to know how likely it is that a difference between the means could result from the fact that for our data sample means differ a lot already by chance
- Recall the standard error (*SE*)
  - Measure of variability between sample means
    - Small *SE*: most samples should have similar means
    - Large *SE*: large differences in sample means by chance alone

Variance

- Variance is the average variability in the data (*spread*)
- medium variability
- high variability
- low variability



T-Test

- Let's assume
  - ➪ The difference between our obtained samples (SVO-amb. & OVS-amb.) is larger than the what we would expect based on the SE
    - Sample means in our population vary a lot by chance & our two samples are atypical of our population
    - The two samples came from different populations & are typical of their respective population
      - ➪ Difference between samples represents a true difference
      - ➪ As observed diff. between sample means gets larger, the more confident we can be that the second option is correct
- The result for the t-test is *t*-value that helps us decide whether we have found a true difference or not
  - ➪ The bigger the *t*, the more likely we found a true diff.

$$t = \frac{\text{Observed difference between sample means} - \text{Expected difference between population means (if null hypothesis is true)}}{\text{Estimate of the standard error of the difference between two sample means}}$$

The dependent *T*-Test

- The t-test
  - ➪ Compares the mean difference between our samples ($\bar{D}$) with the difference we would expect to find between populations means (μ_D)
    - The effect of our manipulation
  - ➪ Takes into account the standard error of the differences ($s_D/sqrt(N)$)
    - I.e., unsystematic variation

$$t = \frac{\bar{D} - \mu_D}{s_D / \sqrt{N}}$$

    - For our 1-factor (2levels) example the result is
      - ➪ $t(31) = -2.77$, $p < 0.01$
- But actually, for the 2-factor example from your homework, we need a more complicated analysis: repeated measures ANOVA

ANOVA

- Just like a *T*-Test, the ANOVA tells you whether
  - ➪ Differences between conditions are due to your manipulation
  - ➪ Due to unsystematic variation
  - ➪ The two types of variance allow us to draw inferences about means
- The ANOVA can help us analyse differences between means in more complicated designs (e.g., 2x2)
  - ➪ The result of an ANOVA analysis is a *F*-value
    - Ratio of the variance due to your experimental manipulation over unsystematic variation
    - A high *F*-value indicates a lot of the variation results from your manipulation

$$F = \frac{\text{systematic variation}}{\text{unsystematic variation}}$$

  - ➪ This is a very general formula, and the exact calculations will differ depending on your type of measurement (dependent vs. indep.)

Example study

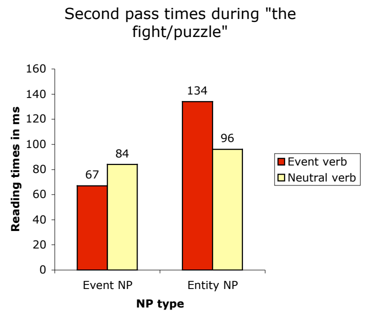
Traxler, Pickering, & McElree, 2002, *JML*

- Semantic interpretation
  - Verbs like *begin* can occur with NP-arguments of different semantic types
    - Event: *start a fight*
    - Entity: *start a puzzle*
    - Verbs like *begin* and *start* appear to prefer an event as argument
  - Coercion operation that type-shifts an entity to an event by inserting additional semantic structure
    - *The boy started **solving** the puzzle*
  - 2x2 design
    - Factor 1: NP type (entity, event)
    - Factor 2: Verb type (entity, event)
  - Target region: *the fight/puzzle*

(7a) The boy started the fight after school today. Event verb + event NP.
(7b) The boy saw the fight after school today. Neutral verb + event NP.
(7c) The boy started the puzzle after school today. Event verb + entity NP.
(7d) The boy saw the puzzle after school today. Neutral verb + entity NP.

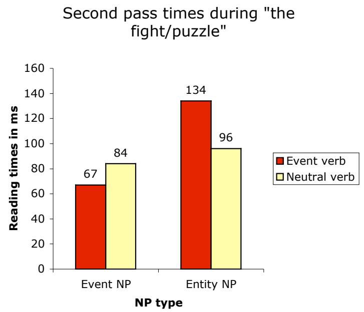
Main effect and interaction

- Main effect
  - The unique effect of an independent variable
  - Reading times for entity NP conditions are higher than for event-type NPs
  - Main effect of NP type confirms this observation
    - F1(1, 35) = 14.4, $p < 0.01$
      F2(1, 31) = 5.74, $p < 0.05$
  - *F*: signal-to-noise; the bigger the F, the stronger the effect of our manipulation
  - *p*: probability that the findings are due to chance



Main effect and interaction

- Interaction
  - The combined effect of two or more independent variables on the dependent variable
  - The verb-type factor affects reading times differently for Entity-type NPs than for Event NPs



Summary

- Homework: experiment design
- Exploring data (here: at least interval-scale)
  - Error bar graphs
  - Box plots
  - Testing for normal distribution and homogeneity of variance
- Inferential statistics
  - Comparing two means (1 factor, 2 levels): *T*-Test
  - ANOVA
  - An example reading study: main effect vs. interaction
- Reading for next week:
  - Lexical processing and the mental lexicon. In: A. Radford, M. Atkinson, D.Britain, H. Clahsen, & A. Spencer (1999). *Linguistics: an introduction* (pp. 226-239). Cambrigde, CUP.