Introduction to Psycholinguistics

Lecture 4

Experimental Methods I

Pia Knoeferle & M. W. Crocker

Department of Computational Linguistics

Saarland University

SS 2006

Overview

- What's an *experiment* and why do we run experiments*?*
- Empirical research cycle
- Building statistical models for observed data
  - Standard deviation
  - Frequency distributions
- Samples: representative of the population?
  - Standard error
- Hypothesis testing
- Choosing a statistical test

Field, 2005; Howell, 2004

Why do we run experiments?

- To answer a research question or test a theory/hypothesis
  - Non-experiment
    - Introspection
      - Is sentence A harder to understand than B?
    - Collection of data
      - Speech errors (e.g., spoonerisms, blendings)
    - ? further ideas
  - *Experiment*
    - Some definitions
      - Systematic observations of a specific behaviour under controlled circumstances
      - Set of actions and observations, performed to verify or falsify a hypothesis or research a causal relationship between phenomena
      - The act of conducting a controlled test or investigation
    - Quasi-experimental designs
    - Experimental designs

Why do we run experiments?

- Quasi-experimental designs
  - Example (made up)
    - Gender differences in using hedges versus assertions

	Hedging expressions	Assertive statements
Men	(1)	(1)
Women	(2)	(2)

- Problem
  - Gender is not a freely manipulated variable; could correlate with other variables (I.e., your findings might result from an other variable rather than gender)
- For some research questions unavoidable
  - Gender, class, intelligence

Why do we run experiments?

- Experimental designs
  - Example
    - Effects of preparation on using hedges versus assertions

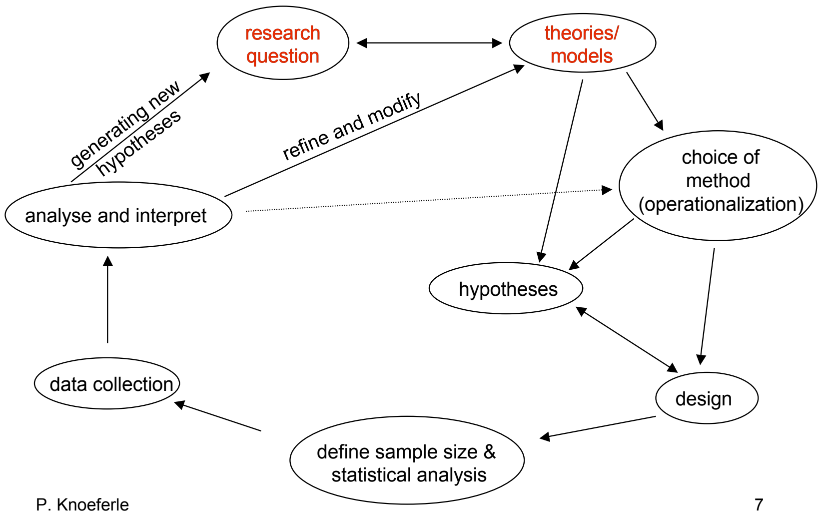
	Hedging expressions	Assertive statements
Prepared	(1)	(1)
Unprepared	(2)	(2)

  - All variables can be freely manipulated
  - Participants can randomly be assigned to the experimental conditions
  - Avoids systematic effects of other (correlated) variables through randomization

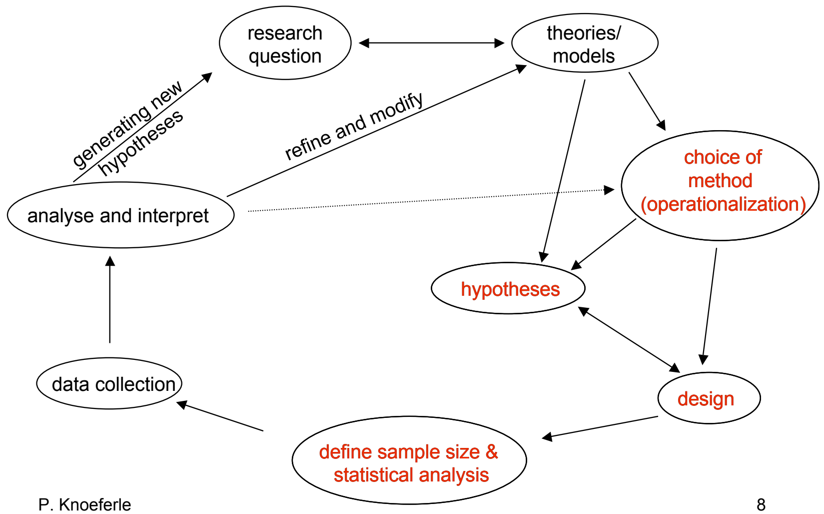
Recap: Theories and models

- Last lecture
  - Research question
    - How does the comprehension system build a syntactic analysis of a sentence?
  - Theory1 (simplest-first)
    - We follow strategies (e.g., choose the simplest analysis first)
  - Theory2
    - We do not chose the simplest analysis first

Cycle of empirical research



Cycle of empirical research



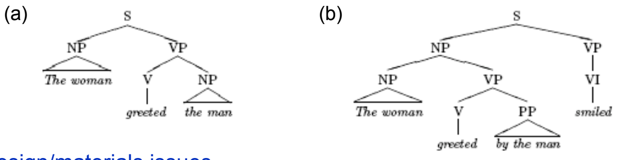
Cycle of empirical research: example

- Operationalize
  - If simplest-first theory is true, people should experience processing difficulties when simplest analysis is disconfirmed
- Hypotheses
  - H1 (*Experimental hypothesis*)
    - Processing difficulty when simplest analysis disconfirmed
  - H0 (*Null hypothesis*)
    - No processing difficulties
- Design
  - Independent variable(s)
    - What we manipulate
  - Dependent variable(s)
    - What we measure

Cycle of empirical research: example

- Design
  - 1 *Factor* design: Sentence type (i.e., our independent variable)
    - 2 *levels*: simple (a) versus complex (b) syntactic analysis

(a) The woman greeted the man with a smile.
(b) The woman greeted by the man smiled.



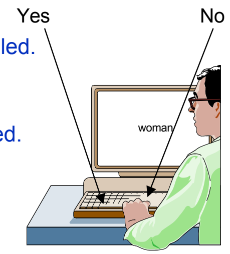
  - Design/materials issues
    - Confounds; length/frequency matching; counterbalancing

Cycle of empirical research: example

- Method?
  - Lexical decision, naming, window-methods in self-paced reading (SPR), eye tracking, neuropsychological methods
  - Lexical decision

(a) The woman greeted the man and smiled.

(b) The woman greeted by the man smiled.



  - Pros/cons of using this method?

Cycle of empirical research: example

- Method?
  - SPR (see Lecture 2 for example)
    - Pro
      - Fairly incremental - per-region reading times
    - Cons
      - Only total reading times in a region (i.e., no first-/second-pass, or regression path duration)
      - You only see one region at a time: artificial presentation that does not allow you to re-read earlier text
  - Eye tracking (see Lecture 2)
    - Pros
      - Incremental, per-region reading times
      - Fine-grained distinctions: first pass, second pass, regression path duration, total times
      - The entire sentence is presented, allowing people to re-read text
    - Cons
      - No direct evidence of neural activation

Cycle of empirical research: example

- Lists for a 1-factor within-subjects design with 2 levels

Item	List 1	List 2	List 3	List 4 ...
1	MC	RR	...	
2	RR	MC		
3	MC	RR		
4	RR ...	MC		

Cycle of empirical research: example

- Items versus fillers
- Randomize items and fillers
  - At least 1 filler in between 2 items
  - Typically at least 3 fillers at the beginning of a lists
  - Pseudo randomization, e.g., Latin Square
  - Sometimes even a practice run to illustrate the task/procedure
- What your experimental lists looks like depends on your design

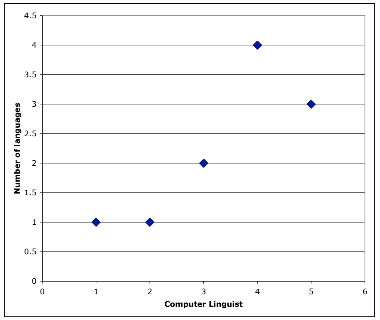
Building statistical models

- Analogy of building a house
  - Collect data about houses, their materials, quality
  - Use this information to build a model
  - Model is small-scale version of a real house
  - But try to build the model that best fits the real world based on the available data
  - Model can be used to make predictions about the real world
  - Test model under various conditions
  - *Infer* from the model about the real-world situation
  - Degree to which a model represents the collected data: *fit*

Building statistical models

- Populations and samples
  - We want our results to apply to the entire *population* of people/things
    - General/narrow populations
- Impossible to access every member of a population
- Instead, we collect a *sample*, and use the behavior within the sample to infer from it about the population
- *Random sampling*
  - Each member of a population has an equal chance of being in the sample
- Simple statistical model*: mean* of a list of numbers
  - Sum of all the members of the list / by the number of items in the list
  - Number of programming languages a CL student knows
  - Five samples: 1, 1, 2, 4, 3 languages; *Mean*: (1+1+2+4+3)/5=2.2

Building statistical models



- Fit between observed data and fitted model
  - ➪ Deviance between observed data and model

    $x_1 - \bar{x} = 1 - 2.2 = -1.2$
  - ➪ Total error: sum of deviances

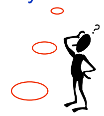
    $\sum (x_i - \bar{x}) = (-1.2) + (-1.2) + (-0.2) + (1.8) + (0.8) = 0$
  - ➪ So, no total error?
  - ➪ Sum of squared errors (*SS*)

    $\sum (x_i - \bar{x})(x_i - \bar{x}) = (-1.2)^2 + (-1.2)^2 + (-0.2)^2 + (1.8)^2 + (0.8)^2 = 1.44 + 1.44 + 0.04 + 3.24 + 0.64 = 6.8$
  - ➪ But ...?

Building statistical models

- Fit between observed data and fitted model
  - ➪ Average error: divide *SS* by the number of observations (*N*)
    - Average error of sample: divide *SS* by *N*
  - ➪ To use sample error as estimate of population error, divide *SS* by N-1
    - *Variance*: average error between mean and observations

      $\text{variance}(s^2) = \frac{SS}{N-1} = \frac{\sum (x_i - \bar{x})^2}{N-1} = \frac{6.8}{4} = 1.7$

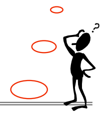


Why N-1?: Degrees of freedom

- - - *Standard deviation*: square root of the variance

      $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N-1}} = \sqrt{1.7} = 1.3$
    - Small *s* relative to mean
      - ➪ Data points are close to mean
      - ➪ The mean is an accurate representation of the data
      - ➪ *s* equal to zero would mean ?

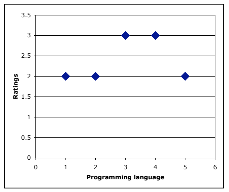
Building statistical models

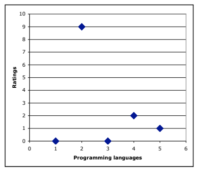


Why N-1?: Degrees of freedom

- Degrees of freedom
  - ➪ Number of observations that are free to vary
  - ➪ Example: Number of languages a CL knows
    - Sample of four observations from a population: can vary freely
    - If we use this sample to calculate standard deviation: we have to use mean of the sample as an estimate of the population mean (i.e., we hold one parameter constant)
      - ➪ Mean of sample: 4 languages
      - ➪ We assume that population mean is also 4
    - If we have four CLs, can all four vary freely?
      - ➪ Three CLS: 2, 6, 6 languages
      - ➪ The fourth CL must know 2 language to keep the mean of 4
    - The last observation is not free to vary if we hold one parameter constant: df is one less than then number of observations

The *s* as a measure of fit





- Two CLs are rated on their skills in five programming languages

$\bar{x} = \frac{2+2+3+3+2}{5} = 2.4$

$s = \sqrt{\frac{(x_i - \bar{x})^2}{N-1}} = \sqrt{\frac{(2-2.4)^2 + (2-2.4)^2 + (3-2.4)^2 + (3-2.4)^2 + (2-2.4)^2}{4}} = \sqrt{\frac{0.16+0.16+0.36+0.36+0.16}{4}} = \sqrt{\frac{1.2}{4}} = 0.55$

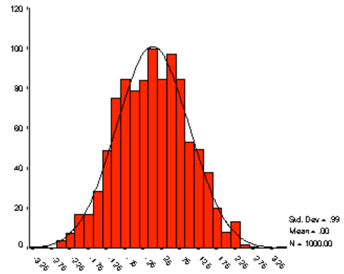
$\bar{x} = \frac{0+9+0+2+1}{5} = 2.4$

$s = \sqrt{\frac{(x_i - \bar{x})^2}{N-1}} = \sqrt{\frac{(0-2.4)^2 + (9-2.4)^2 + (0-2.4)^2 + (2-2.4)^2 + (1-2.4)^2}{4}} = \sqrt{\frac{5.76+43.56+5.76+0.16+1.96}{4}} = \sqrt{\frac{57.2}{4}} = \sqrt{14.3} = 3.78$

Building a statistical model

- $Outcome_i = (Model_i) + error_i$
  - Observed data can be predicted from model + some amount of error
- s^2 and s are measures of goodness of fit of a model
  - deviation = $\sum(\text{observed-model})^2$
- A further estimate of model-data fit
  - Using frequency distributions

The normal distribution

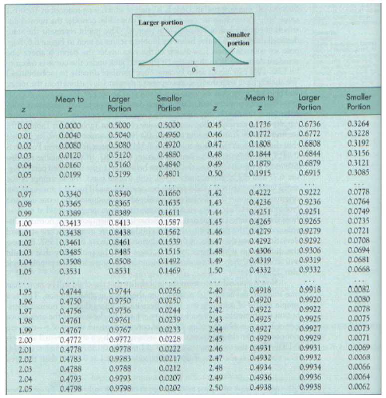


- Frequency distribution (histogram)
  - X-axis: values of observation
  - Y-axis: frequency of occurrence
- Probability distributions
  - Idealized version of the frequency distributions
  - E.g., *standard normal distribution*
    - mean=0
    - s=1

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(X-\mu)^2/(2\sigma^2)}$$

- Using distributions to get an idea of the probability that a score occurs with
- Tables of probabilities for normal distribution
  - Look up how likely a score is to occur

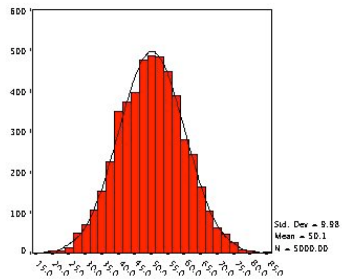
Standard normal distribution



z	Mean to z	Larger Portion	Smaller Portion	z	Mean to z	Larger Portion	Smaller Portion
0.00	0.0000	0.5000	0.5000	0.45	0.1736	0.6736	0.3264
0.01	0.0040	0.5040	0.4960	0.46	0.1772	0.6772	0.3228
0.02	0.0080	0.5080	0.4920	0.47	0.1808	0.6808	0.3192
0.03	0.0120	0.5120	0.4880	0.48	0.1844	0.6844	0.3156
0.04	0.0160	0.5160	0.4840	0.49	0.1879	0.6879	0.3121
0.05	0.0199	0.5199	0.4801	0.50	0.1915	0.6915	0.3085
...	...	...	...	...	...	...	...
0.97	0.3340	0.8340	0.1660	1.42	0.4222	0.9222	0.0778
0.98	0.3365	0.8365	0.1635	1.43	0.4236	0.9236	0.0764
0.99	0.3389	0.8389	0.1611	1.44	0.4251	0.9251	0.0749
1.00	0.3413	0.8413	0.1587	1.45	0.4265	0.9265	0.0735
1.01	0.3438	0.8438	0.1562	1.46	0.4279	0.9279	0.0721
1.02	0.3461	0.8461	0.1539	1.47	0.4292	0.9292	0.0708
1.03	0.3485	0.8485	0.1515	1.48	0.4306	0.9306	0.0694
1.04	0.3508	0.8508	0.1492	1.49	0.4319	0.9319	0.0681
1.05	0.3531	0.8531	0.1469	1.50	0.4332	0.9332	0.0668
...	...	...	...	...	...	...	...
1.95	0.4744	0.9744	0.0256	2.40	0.4918	0.9918	0.0082
1.96	0.4750	0.9750	0.0250	2.41	0.4920	0.9920	0.0080
1.97	0.4756	0.9756	0.0244	2.42	0.4922	0.9922	0.0078
1.98	0.4761	0.9761	0.0239	2.43	0.4925	0.9925	0.0075
1.99	0.4767	0.9767	0.0233	2.44	0.4927	0.9927	0.0073
2.00	0.4772	0.9772	0.0228	2.45	0.4929	0.9929	0.0071
2.01	0.4778	0.9778	0.0222	2.46	0.4931	0.9931	0.0069
2.02	0.4783	0.9783	0.0217	2.47	0.4932	0.9932	0.0068
2.03	0.4788	0.9788	0.0212	2.48	0.4934	0.9934	0.0066
2.04	0.4793	0.9793	0.0207	2.49	0.4936	0.9936	0.0064
2.05	0.4798	0.9798	0.0202	2.50	0.4938	0.9938	0.0062

- Often mean and s will not be 0 and 1 respectively (N(0,1))
- To rely on probability with which a score occurs for a sampling distribution
  - Use *linear transformation* to convert other sampling distributions to the standard normal distribution

Z-transformation



- Mean μ = 50
- Std. dev. s = 10
  - 5000 samples
  - *z*-transformation

$$z = \frac{X-\mu}{\sigma} \qquad z = \frac{X-50}{10}$$

$$for X = 70: z = \frac{70-50}{10} = \frac{20}{10} = 2$$

Samples: representative of the population?

- *Standard error* (≠ standard deviation)
  - Example: ratings of 2 computer linguists (1=good to 6=bad)
    - Let population mean μ be 3
    - We cannot obtain data from the entire population
    - But we can draw several samples (e.g., 9) of ratings
    - Each sample has a *sample mean*
    - *Sampling variation:* sample means may differ
    - *Sampling distribution:* centered at population mean (i.e, average of all sample means is 4)
      - Frequency distribution of sample means from the same population
  - Remember
    - Standard deviation measures the error **within** a sample

Samples: representative of the population?

- Sampling distribution tells us about behaviour of samples from population
- Average of sample means is population mean
- If we know accuracy of that average, then we know how likely it is that a sample is representative of the population
- Standard deviation between *sample means*
  - Measure of variability in sample means
  - *Standard Error* (SE) of the mean $SE = \frac{\sum(\bar{X}_i - \bar{X})^2}{N}$
  - Too cumbersome in real life, so we rely on approximations
    - Divide *s* by square root of the sample size

$$\sigma_{\bar{X}} = \frac{s}{\sqrt{N}}$$

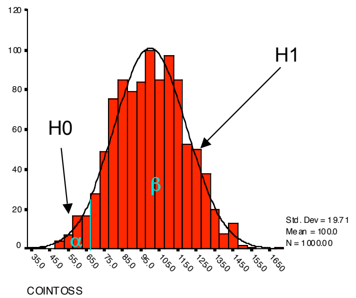
Hypothesis testing

- Test research hypothesis
  - Set up null hypothesis H0
    - Sample came from pop. with mean μ=50
      - Obtain a random sample: sample mean=55
      - Obtain sampling distribution of mean under assumption that H0 is true
      - Probability that a sample mean of 55 could reasonably arise if we had drawn in from a population with $\mu = 50$?
  - Sampling distribution can provide the answer
- Standard normal distribution (z-transformation)
  - Determine probability of obtaining a sample mean of 55 (0.3)
    - Based on probability, we decide either to reject or fail to reject H0
  - A sample with mean 55 is obtained in 30% of the time from this population, so we don't have good reason to doubt that this sample came from such a population
    - We fail to reject H0

Hypothesis testing

- Alternative: sample mean is 70
  - Probability of that mean is 0.02 - unlikely event that occurs only in 2%
  - This sample mean came from another pop.
  - We reject H0
- Fisher: Confidence level
  - Tea-cup example
  - $p < 0.05$: We are more than 95 % certain that our findings are not the result of chance

Type I and II errors



	True state of the world	
Decision	H0 True	H0 False
Reject H0	Type I error $p=\alpha$	Correct decision $p=1-\beta$=Power
Fail to reject H0	Correct decision $p=1-\alpha$	Type II error $p=\beta$

- Deciding which hypothesis
  - Critical values
    - Values of the variable that describe the boundary of the rejection region(s) (here: 67)
  - Decision rule
    - Reject H0 when result falls in the lowest 5% of distribution ($p < 0.05$ of coming from the population, μ=100 σ=20)
- Type I error
  - Reject H0 when true (α is the probability of a Type I error)
- Type II error
  - Not rejecting H0 when it is false (β is the prob. of a Type II error)

Choosing a statistical test

- Choice of inferential & descriptive statistics depends on
  - Types of data
  - Type of design

Type of data

- *Qualitative (categorical/frequency/count)* data
  - Consist of totals or frequencies for a category (e.g., the number of times people looked at the word *kitchen* compared with *ktchien*)
- *Quantitative* data
  - Result of any sort of measurement (e.g., reading speed, fixation duration)
- More detailed characterization based on *level of measurement*
  - 1. *Nominal* 2. *Ordinal* 3. *Interval* 4. *Ratio*
    - Can be classified according to their informativity from left to right
  - Qualitative data: nominal and ordinal levels
  - Quantitative data: interval and ratio levels

Type of data

- Nominal
  - Unordered set of qualitative values
  - Numbers represent names/categories
  - Descriptive statistics: relative frequencies
  - Inferential statistics: e.g., chi-square test, log-linear models
- Ordinal (Rank data)
  - Like nominal, but the values have a meaningful ordering
    - E.g., “First to last” relationship between values
  - Descriptive statistics: percentiles (One of a set of points on a scale arrived at by dividing a group into parts in order of magnitude)
  - Inferential statistics: non-parametric statistics for ordinal data

Type of data

- Interval (continuous data)
  - ➪ Arbitrary zero point and measuring unit
  - ➪ Possible to determine the relationships between differences in individual observations (intervals)
    - For true interval data on a scale from 1-10, the increase from 2 to 3 should be the same as from 9 to 10
  - ➪ Descriptive statistics: mean, variance, standard deviation
  - ➪ Inferential statistics: t-test, analysis of variance (ANOVA)
- Ratio
  - ➪ Like interval, but in addition has a true zero point
    - In addition: on scale from 1-10, 4 is twice as good as 2
  - ➪ Descriptive statistics: central tendency, geometric mean
  - ➪ Inferential statistics: like interval data

Type of design

1. Relationships versus differences
   - ➪ Differences between two or more groups
   - ➪ Relationships between two or more variables
2. Number of groups/variables
   - ➪ One
   - ➪ Two or more
3. Way of measuring
   - ➪ Dependent
     - Within subjects/items design
   - ➪ Independent
     - Between subjects/items design
   - ➪ Mixed
     - Partly between; partly within

Type of design

1. Relationships versus differences
   - ➪ Relationships
     - ➪ Between number of cigarettes smoked per day and scores on a task
     - ➪ Between working memory span and reading times
   - ➪ Differences
     - ➪ Between smokers and non-smokers on the same task
     - ➪ For reading times in readers with low compared with readers that have a high working memory span
     - ➪ Reading times for simple main clause compared with reduced relative clause sentences

Type of design

2. Number of groups/variables

- One factor
  - ➪ E.g., Word order, sentence complexity, grammaticality

Item	Subj 1	Subj 2	Subj 3	Subj4 ...
1	SVO	OVS	...	
2	OVS	SVO		
3	SVO	OVS		
4	OVS ...	SVO		

Type of design

2. Number of groups/variables

- Two or more factors
  - ➪ Word order (SVO/OVS) & Ambiguity (amb., unamb.)

Item	Subj 1	Subj 2	Subj 3	Subj4 ...
1	SVOa	OVSa	SVOu	OVSu
2	OVSa	SVOu	OVSu	SVOa
3	SVOu	OVSu	SVOa	OVSa
4	OVSu	SVOa	OVSa	SVOu
5	SVOa	OVSa	SVOu	OVSu
				...

  - ➪ How many items?

Type of design

3. Ways of measuring
   - ➪ Between-subject design
     - Designs in which different subjects serve under different treatment levels
     - Example: Order of presentation (picture-first, picture-last) x Congruence (match, mismatch)
       - ➪ Between: Order of presentation
       - ➪ Within: Congruence

see Underwood et al., 2004

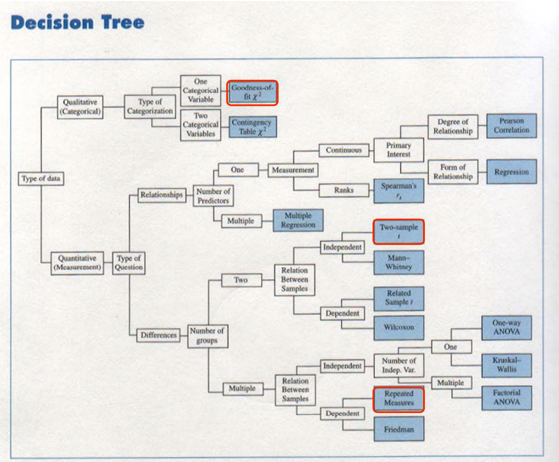
Item	Subj 1	Subj 2	Subj 3	Subj4 ...
1	Picfirst-m	Picfirst-n	Piclast-m	Piclast-n
2	Picfirst-n	Picfirst-m	Piclast-n	Piclast-m
3	Picfirst-m	Picfirst-n	Piclast-m	Piclast-n
4	Picfirst-n	Picfirst-m	Piclast-n	Piclast-m
5	Picfirst-m	Picfirst-n	Piclast-m	Piclast-n
				...

Type of design

3. Ways of measuring
   - ➪ Within-subject design (repeated measures)
     - Designs in which each subject receives all levels of at least on independent variable
       - ➪ See Example for “Two or more factors”
     - We measure reading times for SVO vs. OVS sentences
       - ➪ One subjects receives both SVO and OVS sentences
       - ➪ If, e.g., a subject is a weak reader, they will have difficulties reading both SVO and OVS, but more so for OVS

Item	Subj 1	Subj 2	Subj 3	Subj4 ...
1	SVOa	OVSa	SVOu	OVSu
2	OVSa	SVOu	OVSu	SVOa
3	SVOu	OVSu	SVOa	OVSa
4	OVSu	SVOa	OVSa	SVOu
5	SVOa	OVSa	SVO	OVSu
				...

Choice of analysis techniques



Howell, D. C. (2004). Fundamental statistics. Thomson: London.

Conclusions

- Cycle of empirical research: an example
- Building statistical models
- Hypothesis testing
- Choosing a statistical test
- Homework: Design an experiment (in writing)
  - 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?