Introduction to experimental methods for linguists

Lecture 2b

Hypothesis Testing

- hypothesis
- formulate a null hypothesis (Ho) – a statement that there is no difference between measured phenomena
- typical statistics based on tests to reject the null hypothesis in favour of the alternative hypothesis (Ha) that there is a difference
- the level of significance for a the rejection of Ho is 5% (written as $P < 0.05$)
- this means that the probability that the null hypothesis could be rejected by chance alone is 5% or less

Hypothesis Testing

- error
- bias is bad, variability (error) is expected
- type I error
  - false positive: finding something that is not there
  - incorrectly rejecting a true null hypothesis
- type II error
  - false negative: failing to find something that is there
  - incorrectly accepting a false null hypothesis

Hypothesis Testing

- error
- bias is bad, variability (error) is expected
- type I error
  - false positive: finding something that is not there
  - incorrectly rejecting a true null hypothesis
- type II error
  - false negative: failing to find something that is there
  - incorrectly accepting a false null hypothesis

Hypothesis Testing

- the Lady Tasting Tea
  - famous randomised experiment by Ronald A. Fisher (1935)
  - a lady (Muriel Bristol) claimed she could taste the difference between tea that had milk poured in the cup first, or after the null hypothesis was that she could not taste the difference
Hypothesis Testing

the Lady Tasting Tea
the lady is presented with 8
cups of tea – 4 with milk first, 4
after – in a randomised order
(there are 70 unique
combinations)
to satisfy Fisher, she had to
guess all 8 correctly (a
probability of 1/70 or 1.4%)
she did – the null hypothesis
was rejected

Participants

how will your participants be compensated?
course credit
cash
a chance to win money or other prizes
stickers and sweets (children)
nothing at all (aren’t you cheap!)
still need to pay even if they withdraw from
experiment

Participants

how will your participants be recruited?
friends and their networks
posters and emails on campus
participant recruitment pools (i.e., 1st year
undergraduate course requirement)
off-campus recruitment
websites and survey systems (e.g., MTurk)
social network

Participants

will you exclude participants?
non-native speakers & bilinguals
age (not too young, but not too old either)
failure on key questions/controls (e.g., do they pay
attention?)
behavioural problems during testing (e.g.,
restlessness)
need to have exclusion criteria in advance
cannot exclude participants because you do not like
them or a “gut” feeling - can result in sample bias
sample size should be based on number of participants
used in analysis excluding dropouts (which can be up
to 30% for difficult populations)

Participants

how will you deal with participants?
be courteous, polite and brief (don’t waste their time)
inform them of study (information sheet for
participants often required)
debrief them after the experiment
provide them with contact details for experimenter
(work email, never home) as well as ethics committee
if they have complaints
what is the best way to respond to no-shows and
difficult individuals?
Participants

how will you deal with participants?
  children and vulnerable populations require more care
  police checks are needed before being allowed to test
  need informed consent of parents and guardians
  have to be especially sensitive to needs of participants and to be respectful to guardians

Sampling

cannot test an entire population
  time (difficult to collect data from everyone quickly, though not impossible, e.g., vote, census)
  money (expensive to collect data from everyone, i.e., production costs, payment to participants)
  access (not always possible to reach all members of a population)
  sufficiency (pattern of results don’t change much even if we have data from everyone)

Sampling

how to sample
  random sample
    the gold standard
    each member of the population has an equal chance of being selected
    usually quasi-random (stratified random sample)
    systematic
draw from the population at fixed intervals
problematic in populations with a periodic function

Sampling

sample: a selection of individuals from the larger population
  for any population there are many possible samples
  sample statistics are used to infer population parameters

Sampling

how to sample
  stratified sample
    proportional: specified groups appear in numbers proportional to their size in the population
    disproportional: specified groups which are not equally represented in the population, are selected in equal proportions
  cluster sample
    researcher samples an entire group or cluster from the population of interest
Sampling

how to sample
opportunity/convenience sample
people who are easily available (e.g., museum visitors)
can lead to a biased sample (e.g., your linguist friends for a linguistic phenomenon)
snowball sampling
recruit small number of participants and then use those initial contacts to recruit further participants biases the sample, but useful if you want to recruit very specific populations

Sampling

typically for a linguistic experiment
test native speakers only
relatively monolingual (definitely so for children)
try to test people from varied socio-economic backgrounds and dialects (unless you stick to one)
do not test linguists!
sample size will depend on:
the type of method you use (check literature)
the number of conditions
> more conditions means more participants
the number of items per condition
> less items means more participants

Sampling

sample size
size matters!
sampling error can result if your sample is not large enough
trade-off between size and time/cost
factors in deciding on sample size
more conditions means more subjects
power analysis
need pilot data or effect sizes from other studies

test as many as other people do (typically 24-36 subjects for lab studies, up to several hundred for internet based studies)

Subjects Design

subjects design: the assignment of participants to experimental conditions (levels of the IV)

between-subjects / independent groups
participants each exposed to one level of the IV

within-subjects / repeated measures
participants exposed to all levels of the IV

mixed designs
mixture of between and within designs require use of different statistical tests

Subjects Design

between-subjects / independent groups
assign participants to two groups
e.g., alcohol consumption (IV = 2 levels) and memory (DV)
N=20 for alcohol and N=20 for control
Subjects Design

**between-subjects / independent groups**
how do we ensure that any differences in memory function result from alcohol intake rather than some other variable (e.g. age, gender, nutritional factors, experience with alcohol etc.)?
we cannot eliminate the effects of these other variables, but we can minimise these effects by spreading their influence across the different levels of the IV(s)

**Subjects Design**

**between-subjects / independent groups**
random allocation
ensures that each participant is equally likely to be assigned to any IV level
why randomise?
distributes the occurrence of potential moderating variables equally among experimental conditions prevents experimenters (un)intentionally biasing their results
allows use of powerful statistical tests

Subjects Design

**within-subjects / repeated measures**
participants take part in both levels (conditions) of IV e.g., alcohol consumption as before, with N=20 for alcohol and the same participants for control

**Subjects Design**

**within-subjects / repeated measures**
potentially moderating characteristics are kept equal across the levels of the IV (each participant acts as his/her own control)
requires fewer participants than between-subjects design
however, problem of order effects: once participants have been exposed to one level of the IV there’s no way to return them to their original state

Subjects Design

**within-subjects / repeated measures**
solution: counterbalancing
split the group of participants in half (A and B)
Group A can participate in Level 1 then Level 2
Group B can participate in Level 2 then Level 1
e.g., Group A: alcohol – test; then no alcohol – test
Group B: no alcohol – test; then alcohol – test
order effects will still influence Ps performance, but the effect of that influence will be evenly spread across each level of the IV

**Subjects Design**

factorial designs
experimental designs with 2 or more IVs allow us to ask:
what effect does IV1 have on the DV?
what effect does IV2 have on the DV?
what effect does the interaction of IV1 and IV2 have on the DV?
example: effects of alcohol consumption and work shift patterns on work productivity – DV: work output, IV1: shift pattern, IV2: alcohol consumption
Subjects Design

factorial designs

fully independent factorial design (between subjects)
- each participant takes part in just one experimental condition (1 level of a single IV)
- with 2 levels for each IV, that’s 4 cells, so 4 x N Ps

fully repeated measures factorial design (within subjects)
- each participant takes part in all experimental conditions (all levels of all IVs) (e.g., N = 20)

Subjects Design

factorial designs

factorial mixed design
- one IV will be between-subjects (randomly assign Ps to either group)
- the other IV will be within-subjects (each group does both levels, counterbalanced for order)

Experimental Design

Deamer et al. (2010)

<table>
<thead>
<tr>
<th>Literal context:</th>
<th>It was a forest.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sam and Mark went for a walk in a national park.</td>
<td></td>
</tr>
</tbody>
</table>

| Metaphor context: | Sam always got lost. The university was enormous. |

| Hyperbole context: | The backyard definitely needed pruning. |

Experimental Design

Deamer et al. (2010)

- reading-time study
- target sentences preceded by a literal, hyperbolic or metaphoric context
- long/short context
- what is within/between subjects?
- what do we counterbalance/randomise?
- how do we ensure the participants are paying attention?

Choosing a design

- we are often able to choose which subjects design to employ
- independent
- repeated measures
- mixed

Choosing a design

- choice depends on:
  - concerns
    - between subjects – eliminate order effects
    - within subjects – eliminate individual difference effects
  - number of participants (availability)
    - within subjects requires fewer participants
Experimental Design
Deamer et al. (2010)

- reading-time study
- target sentences preceded by a literal, hyperbolic or metaphoric context
- long/short context
trope condition within subjects —> each participant was randomly assigned to one list (each target sentence once, in one condition)
context length condition between subjects —> half participants saw the target sentences in long contexts/ half short contexts

Experimental Design

all research involves comparison
a research design which allows us to make causal inferences about the influence of one or more variables on a variable of interest
the researcher manipulates one or more independent variables (IV) and measures the effect on the dependent variable (DV)

Experimental Design

control freaks
not all comparisons can be against an established baseline or standard
placebos used in medicine controls used in other experiments
control conditions have to be chosen carefully for the variables of interest
use negative controls when the factor of interest is not supposed to appear
use positive controls for factors that can arise for reasons other than the experimental factors

Experimental Design
Behne et al. (2005b)

Understanding communicative intentions

60 14, 18 and 24-month-olds

In the context of a hiding game, an adult indicated for the child the location of a hidden toy by giving a communicative cue: either pointing or ostensive gazing toward the container containing the toy.

All ages successfully used both types of cues.
“infants (…) interpret an adult behaviour as a relevant communicative act done for them.”

Experimental Design
Behne et al. (2005b)
Understanding communicative intentions Behne et al. (2005b)

Control condition:
- Experimenter pointed (or gazed) at correct container, but in a distracted, non-communicative manner
- Children chose randomly

Experimental Design

randomisation
- an important part of experiments, especially assigning participants to groups (see below)
- randomisation can be accounted for statistically, biased choice cannot
- not the same as haphazard sampling, which can introduce bias the experimenter is not aware of (e.g., testing people who make eye contact)
- randomness can be generated by coin tosses, tables, dice or websites (randomizer.org is good)

Experimental Design

blocked design
- instead of presenting conditions or variables randomly, they are presented in groups
- e.g., 3 trials of condition A followed by 3 of B for half of the time, and 3 of B then A for the other half (randomly determined)
- this can be used to reduce confusion in subjects and decrease variability in results (especially useful with smaller sample sizes)

Experimental Design

factorial design
- varying more than one independent variable at a time
- more complex, but allows for testing interactions between variables
- the foundation for the ANOVA test (analysis of variance)

Experimental Design

2x2 design
- called “god’s design” by Tomasello
- allows testing for the simplest interactions
- e.g., Liszkowski et al (2006) presented children at 2 ages with a target or distractor (2 condition x 2 ages)
- asterisks indicate significant difference (P < 0.05)
- no interaction with age but a main effect of condition

Experimental Design
Counterbalancing

Latin square design

- for more than two conditions per treatment (e.g., 2x3 and 3x3 designs), number of combinations is large
- 2 factor design has 2 Latin squares, 3 factor has 12, 4 factor has 576, etc.
- use Latin square to simplify design: each letter occurs only once in each row and column

\[
\begin{array}{ccc}
A & B & C \\
C & A & B \\
B & C & A \\
\end{array}
\]

Experimental Design
Counterbalancing

ABA (reversal) design

- when testing the same subjects, there is the problem of order effects
- performing on one condition can have an effect on subsequent ones
- time (e.g., ageing) can also be a confound
- to eliminate this, present one condition first (A), then the second one (B), then the first one again (A)
- can also do ABAB or ABBA

Experimental Design
Acquiring figurative expressions
Deamer & Pouscoulous (in prep.)

- Can children select the metaphorical meaning of an expression when there is a ‘literal’ competitor?
- Does understanding figurative meaning of an expression - and inhibiting/suppressing its literal meaning - involve an active role of inhibition control?
- What about other forms of non-literal speech?

Experimental Design
Acquiring figurative expressions
Deamer & Pouscoulous (in prep.)

- Participants:
  - 16 3-year-olds (8 girls): 3;0-3;11 (mean: 3;6)
  - 14 4-year-olds (7 girls): 4;1-4;8 (mean: 4;4)
  - 20 5-year-olds (9 girls): 5;1-5;10 (mean: 5;5)
  - Adult control group: 10 undergraduate (18-25)

- Tasks:
  - Picture selection task: 6 metaphors (Experiment 1) + 6 hyperbole (Experiment 2)
  - 2 inhibition control tasks (Carlson & Moises, 2001)
  - Naming and pointing vocabulary task

Metaphor

This weekend Emily is staying at her granny’s house. Emily has had a bath and now she is a hedgehog. Can you find the picture in which Emily is a hedgehog?
It’s Saturday and Tom is playing outside in the rain. He is splashing in the lake outside of his house. Can you find the picture in which Tom is splashing in the lake outside his house?

Experimental Design

replication
results have to be replicated to be determine whether effects are true
lab effects – sometimes some labs consistently get positive results, some always negative
there is pressure to publish positive findings (type I error) and a bias against publishing negative findings specifically and replications generally

Experimental Design

standardisation
- to avoid inadvertent confounds, all aspects of your experiment should be controlled for,
- counterbalanced, randomised or standardised
- who are the experimenters?
greeting and instructions should always be the same
- if participants hear speakers: are they male/female?
- speak in a specific dialect?
- write a precise and exhaustive script, rehearse it and stick to it

Experimental Design

Liebal et al. (2009)

Investigated whether infants use their shared experience with an adult to determine the meaning of a pointing gesture.

- 2 adults share different activities with the infant.
- One of the adults pointed to a target object.
- 18- but not 14-month-olds responded appropriately to the pointing gesture based on the particular activity they had previously shared with that particular adult.s
- 14-month-olds succeeded in a 2nd simpler study.
Experimental Design
Liebal et al. (2009)

Experimental Design
Cooperation & human uniqueness
Warneken et al. (2007)

Reliability

• reliability: precision (consistency)
  the extent to which our measure would provide the same results under the same conditions
  required for studies in which subjective biases can influence the results

Reliability

• test-retest reliability: measures fluctuations from one time to another
  if we administered our measure to the same participants on separate occasions, would we obtain the same results?
  important for constructs which we expect to be stable (e.g., personality type)

Reliability

• inter-rater reliability: measures fluctuations between observers
  if two different raters/observers measured the variable of interest, would they obtain the same results?
  tested by having 2nd observer/coder blind to study’s design look at 20% of data
  statistics used: Kappa’s coefficient or a correlation
Reliability

parallel forms reliability
if we administer different versions of our measure to
the same participants, would we obtain the same
results?

Material

Instructions
carefully craft your Instructions
make it clear, but don’t say too much
ask participants to go fast, but not too fast
make sure the Instructions work (pilot them!)
don’t forget to ask the participants to consent to
taking part

Material

Experimental materials
how well has your phenomenon been defined (e.g.,
hyperbole vs. metaphor)
standardise your items using explicit and well-
derived criteria
make sure they do/sounds/are understood as you
want they to be
if there are speakers: male/ female voice? accent?
if they are written check font and other variable
for a linguistics experiment they almost invariably
have to be pre-tested

Material

Deamer et al. (2010)
The target sentences were designed using the formal
criteria below;
(a) In all of the descriptions (contexts and target
sentences combined), an object or concept in the context,
is described using the target word in the target sentence
(e.g. A fridge is described as a monster, or earl grey tea
is described as noxious).
(b) The hyperbolic and metaphoric descriptions
differed in one way: A metaphorical description can
never be literally true no matter how much you
manipulate the context. However, a hyperbolic
description can be literally true if the context is modified
slightly.

Material

control for the frequency of your lexical items
(especially when comparing responses on items
across conditions)
Many databases out there to provide frequency
data (i.e, BNC based: http://
www.kilgarriff.co.uk/bnc-readme.html)
check for known lexical and psycholinguistic
effects (e.g., priming effects, wrapping up, etc.)
The contexts were all written with the following considerations in mind:

a) The metaphor, hyperbole and literal contexts clearly induce either a metaphorical, hyperbolic or a literal interpretation.

b) The contexts themselves were written using only literal language. Any loose uses which may be included were very conventionalised.

c) The contexts did not just translate the target sentence into literal language.

d) All efforts were made to ensure that the degree to which the target sentence followed the context was as similar as possible in the metaphoric, hyperbolic and literal conditions.

---

**Experimental Procedure**

**Material**

Deamer et al. (2010)

The contexts were all written with the following considerations in mind:

a) The metaphor, hyperbole and literal contexts clearly induce either a metaphorical, hyperbolic or a literal interpretation.

b) The contexts themselves were written using only literal language. Any loose uses which may be included were very conventionalised.

c) The contexts did not just translate the target sentence into literal language.

d) All efforts were made to ensure that the degree to which the target sentence followed the context was as similar as possible in the metaphoric, hyperbolic and literal conditions.

---

**Experimental Procedure**

- **piloting**
  - in addition to determining which population to test, how to sample them, which subject design and which experimental design, and obtaining ethics, you need to make sure your experiment works
  - do not use the same participants in the test
  - do not do too much piloting (this can lead to type I error)
  - can keep changing procedure until it “works”
  - cannot change any aspect of the procedure once testing has begun

---

**Experimental Procedure**

- **other issues**
  - experimenter influence
  - will the presence of the experimenter influence the participants (Clever Hans)
  - social desirability
  - will answers be honest if the participant tries to anticipate what the experimenter wants?
  - leading questions
  - will the test questions influence participant answers?

---

**Experimental Procedure**

- **blind testing**
  - **single blind** – subject should be blind to the condition and the expected outcome (e.g., should not know whether in experimental or control group)
  - **double blind** – experimenter should also be blind to the condition to avoid unintentionally influencing subject (e.g., neither subject nor experimenter knows whether the trial is part of the control or experimental group)

---

Discussion
Discussion

consider your own work
what is your null hypothesis?
what is your alternate hypothesis?
who will be your participants?
how will they be recruited and compensated?
how many do you need?
what kind of subject design will work best?
what experimental design?
will you need control conditions?