

GRS LX 865

Topics in Linguistics

Week 4. Statistics etc.

Inversion in negation

- Guasti, Thornton & Wexler (BUCLD 1995) looked at doubling in negative questions.
- Previous results (Bellugi 1967, 1971, Stromswold 1990) indicated that kids tend to invert less often in negative questions.
 - First: True?
 - Second: Why?

GTW (1995)

- Elicited negative questions...
 - I heard the snail doesn't like some things to eat. Ask him what.
 - There was one place Gummi Bear couldn't eat the raisin. Ask the snail where.
 - One of these guys doesn't like cheese. Ask the snail who.
 - I heard that the snail doesn't like potato chips. Could you *ask* him if he doesn't?

GTW (1995)

- Kids got positive questions right for the most part.
 - 88% of kids' *wh*-questions had inversion
 - 96% of kids' *yes-no* questions had inversion
 - Except youngest kid (3;8), who had inversion only 42% of the time.
- Kids got negative declaratives right without exception, with *do*-support and clitic *n't*.

GTW (1995)

- Kids got lots of negative *wh*-questions **wrong**.
- **Aux-doubling**
 - What kind of bread do you don't like? (3;10)
- **Neg & Aux doubling**
 - Why can't she can't go underneath? (4;0)
- **No I to C raising (inversion)**
 - Where he couldn't eat the raisins? (4;0)
- **Not structure**
 - Why can you not eat chocolate? (4;1)

GTW (1995)

- But kids got negative **subject** *wh*-questions right.
 - which one doesn't like his hair messed up? (4;0)
- ...as well as **how-come** questions.
 - How come the dentist can't brush all the teeth? (4;2)
- Re: **Not** structure
 - Why can you not eat chocolate? (4;1)
 - Kids only do this with object and adjunct *wh*-questions—if kids just sometimes prefer *not* instead of *n't*, we would expect them to use it just as often with subject *wh*-questions.

GTW (1995)

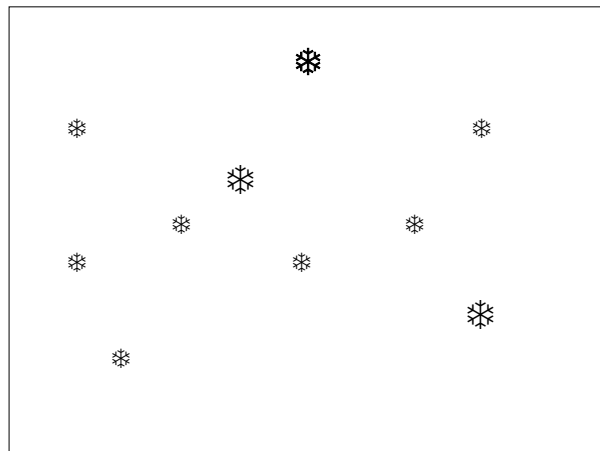
- So, in sum:
 - Kids get positive questions right
 - Kids get negative declaratives right
 - Kids get negative subject questions right.
 - Kids get negative *how-come* questions right.
- Kids make errors in negative *wh*-questions where *inversion* is required. Where inversion isn't required (or where the sentence isn't negative), they're fine.

GTW (1995)

- The kids' errors all seem to have the character of *keeping negation inside the IP*.
 - What did he didn't wanna bring to school? (4;1)
 - What she doesn't want for her witch's brew? (3;8)
 - Why can you not eat chocolate? (4;1)
 - Why can't she can't go underneath? (4;3)
- GTW propose that this is a legitimate option; citing Paduan (Italian dialect) as a language doesn't allow neg->C.

GTW (1995)

- Re: subject and *how come* questions...
- In a subject question, we don't *know* that the subject *wh*-word got out of IP—maybe kids left it in IP... heck, maybe even *adults* do.
 - Who left?
 - *Who did leave?
- *How come* questions don't require SAI in the adult language{./?}
 - How come John left?
 - *How come did John leave?



Descriptive, inferential

- Any discussion of statistics anywhere (\pm a couple) seems to begin with the following distinction:
- Descriptive statistics
 - Various measures used to describe/summarize an existing set of data. Average, spread, ...
- Inferential statistics
 - Similar-looking measures, but aiming at drawing conclusions about a population by examining a sample.

Central tendency and dispersion

- A good way to summarize a set of numbers (e.g., reaction times, test scores, heights) is to ascertain a "usual value" given the set, as well as some idea of how far values tend to vary from the usual.
- Central tendency:
 - mean (average), median, mode
- Dispersion:
 - Range, variance (S^2), standard deviation (S)

Data points relative to the distribution: z-scores

- Once we have the summary characteristics of a data set (mean, standard deviation), we can describe any given data point in terms of its position relative to the mean and the distribution using a standardized score (the z-score).
- The z-score is defined so that 0 is at the mean, -1 is one standard deviation below, and 1 is one standard deviation above:

$$z_i = \frac{x_i - M}{S}$$

Type I and Type II errors

- As a reminder, as we evaluate data sampled from the world to draw conclusions, there are four possibilities for any given hypothesis:

	Inno-cent	Guilty
Convict	Type I error	Correct
Acquit	Correct	Type II error

- The hypothesis is (in reality) either true or false
- We conclude that the hypothesis is true or false.

This leaves two outcomes that are correct, and two that are errors.

Type I and Type II errors

- The risk of making a Type I error is counterbalanced by the risk of making Type II errors; being safer with respect to one means being riskier with respect to the other.
- One needs to decide which is worse, what the acceptable level of risk is for a Type I error, and establish a **critierion**— a threshold of evidence that is needed in order to decide to convict.

	Inno-cent	Guilty
Convict	Type I error	Correct
Acquit	Correct	Type II error

You may sometimes encounter Type I errors referred to as α errors, and Type II errors as β errors.

Binomial/sign tests

- If you have an experiment in which each trial has two possible outcomes (coin flip, rolling a 3 on a die, kid picking the right animal out of 6), you can do a **binomial test**.
 - Called a **sign test** if success and failure have equal probabilities (e.g. coin toss)
- Hsu & Hsu's (1996) example: Kid asked to pick an animal in response to stimulus sentence. Picking the right animal (of 6) serves as evidence of knowing the linguistic phenomenon under investigation.
 - Random choice would yield 1 out of 6 chance (probability .17) of getting it right. Success.
 - Failure: probability 1-.17=.83
 - Chances of getting it right 4 times out of 5 by guessing = .0035. Chances of getting it right all 5 times is .0001.

Hypothesis testing

- Independent variable** is one which we control.
- Dependent variable** is the one which we measure, and which we hypothesize may be affected by the choice of independent variable.
- Summary score**: What we're measuring about the dependent variable. Perhaps number of times a kid picks the right animal.
 - H_0 : The independent variable has no effect on the dependent variable.
 - A grammatically indicated animal is not more likely to be picked.
 - H_1 : The independent variable **does** have an effect on the dependent variable.
 - A grammatically indicated animal is more likely to be picked.

Hypothesis testing

- H_0 : The independent variable has no effect on the dependent variable.
 - A grammatically indicated animal is not more likely to be picked.
- H_1 : The independent variable **does** have an effect on the dependent variable.
 - A grammatically indicated animal is more likely to be picked.
- If H_0 is true, the kid has a 1/6th chance (0.17) of getting one right in each trial.
 - So, given 5 tries, that's a 40% chance (.40) of getting one.
 - But odds of getting 3 are about 3% (0.03), and odds of getting 4 are about .4% (0.0035).
 - So, if the kid gets 3 of 5 right, the likelihood that this came about by chance (H_0) are *slim*.
- =BINOMDIST(3, 5, 0.17, false)
 - Yields 0.03. 3 is number of successes, 5 is number of tries, 0.17 is the probability of success per try. True instead of false would be probability that *at most* 3 were successes.

Criteria

- In hypothesis testing, a criterion is set for rejecting the null hypothesis.
- This is a maximum probability that, if the null hypothesis were true, we would have gotten the observed result.
- This has arbitrarily been (conventionally) set to 0.05.
- So, if the probability p of seeing what we see if H_0 were true is less than 0.05, we reject the null hypothesis.
- If the kid gets 3 animals right in 5 trials, $p=0.03$ — that is, $p<0.05$ so we reject the null hypothesis.

Measuring things

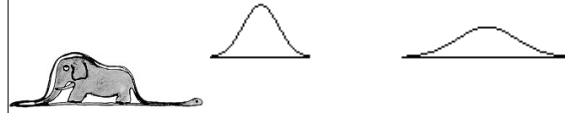
- When we go out into the world and measure something like reaction time for reading a word, we're trying to investigate *the underlying phenomenon that gives rise to the reaction time*.
- When we measure reaction time of reading I vs. $they$, we are trying to find out if there is a real, systematic difference between them (such that I is generally faster).

Measuring things

- Does it take longer to read I than $they$?
- Suppose that in principle it takes Pat A ms to read I and B ms to read $they$.
- Except sometimes his mind wanders, sometimes he's sleepy, sometimes he's hyper-caffeinated.
- Does it take longer for *people* to read I than $they$?
- Some people read/react slower than Pat. Some people read/react faster than Pat.

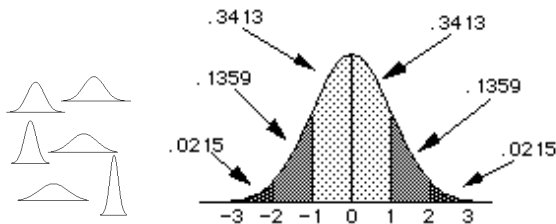
Normally...

- Many things we measure, with their noise taken into account, can be described (at least to a good approximation) by this "bell-shaped" normal distribution.
- Often as we do statistics, we implicitly assume that this is the case...



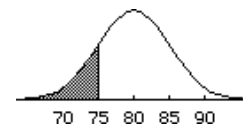
Properties of the normal distribution

- A normal distribution can be described in terms of two parameters.
 - μ = mean
 - σ = standard deviation (spread)



Interesting facts about the standard deviation

- About 68% of the observations will be within one standard deviation of the population mean.
- About 95% of the observations will be within two standard deviations of the population mean.
- Percentile (mean 80, score 75, stdev 5): 15.9



Inferential statistics

- For much of what you'll use statistics for, the presumption is that there *is* a distribution out in the world, a truth of the matter.
- If that distribution is a normal distribution, there will be a population mean (μ) and standard deviation (σ).
- By measuring a sample of the population, we can try to guess μ and σ from the properties of our sample.

A common goal

- Commonly what we're after is an answer to the question: *are these two things that we're measuring actually different?*
- So, we measure for *I* and for *they*. Of the measurements we've gotten, *I* seems to be around *A*, *they* seems to be around *B*, and *B* is a bit longer than *A*. The question is: given the inherent noise of measurement, how likely is it that we got that difference just by chance?

So, more or less, ...

- If we knew the *actual* mean of the variable we're measuring and the standard deviation, we can be 95% sure that any given measurement we do will land within two standard deviations of that mean—and 68% sure that it will be within one.
- Of course, we can't know the actual mean. But we'd like to.

Estimating

- If we take a sample of the population and compute the sample mean of the measures we get, that's the best estimate we've got of the population mean.
 - =AVERAGE(A2:A10)
- To estimate the spread of the population, we use a number related to the number of samples we took and the variance of our sample.
 - =STDEV(A2:A10)
 - If you want to *describe your sample* (that is if you have the entire population sampled), use STDEVP instead.

t-tests

- Take a sample from the population and measure it. Say you took n measurements.
 - Population estimates:
 $\mu_M = \text{AVERAGE}(\text{sample})$, $\sigma_M = \text{SQRT}(\text{VAR}(\text{sample})/n)$
- Your hypotheses determine what you expect your population mean to be if the null hypothesis is true.
 - We're actually considering variability in the *sample means* here—what is the mean mean you expect to get, and what is the variance in those means?
- You look at the distance of the sample mean from the estimated population mean (of sample means) and see if it's far enough away to be very unlikely (e.g., $p < 0.05$) to have arisen by chance.

t-tests

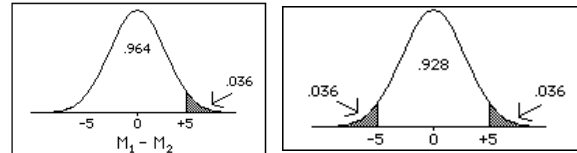
- Does caffeine affect heart rate (example from Loftus & Loftus 1988)?
- Sample 9 people, measure their heart rate pre- and post-caffeination. The measure for each subject will be the *difference* score (post-pre). This is a within-subjects design.
 - Estimate the sample mean population:
 $\mu_M = \text{AVERAGE}(B1:B10) = 4.44$
 $\sigma_M = \text{SQRT}(\text{VAR}(B1:B10) / \text{COUNT}(B1:B10)) = 1.37$
 - *t*-score (like *z*-score) is scaled (here, against estimated standard deviation), giving a measure of how "extreme" the sample mean was that we found.
- If the *t*-score (here 3.24) is higher than the criterion *t* (2.31, based on "degrees of freedom" = $n-1 = 8$) and desired α -level (0.05), we can reject the null hypothesis: caffeine affects heart rate.

t-tests: 2 sample means

- The more normal use of a *t*-test is to see if two sample means are different from one another.
 - $H_0: \mu_1 = \mu_2$
 - $H_1: \mu_1 > \mu_2$
- This is a **directional** hypothesis—we are investigating not just that they are *different*, but that μ_1 is *more* than μ_2 .
- For such situations, our *criterion t* score should be *one-tailed*. We're only looking in one direction, and μ_1 has to be *sufficiently bigger* than μ_2 to conclude that H_0 is wrong.

Tails

- If we are taking as our alternative hypothesis (H_1) that two means simply *differ*, then they could differ in either direction, and so we'd conclude that they differ if the one were far out from the other in either direction. If H_1 is that the mean will increase, then it is a directional hypothesis, and then a one-tailed criterion is called for.



t-tests in Excel

- If you have one set of data in column A, and another in column B,
- =TTEST(A1:A10, B1:B10, 1, type)
 - Type is 1 if paired (each row in column A corresponds to a row in column B), 2 if independently sampled but with equal variance, 3 if independently sampled but with unequal variance.
 - Paired is generally better at keeping variance under control.

ANOVA

- Analysis of Variance (ANOVA), finding where the variance comes from.
- Suppose we have three conditions and we want to see if the means differ.
 - We could do *t*-tests, condition 1 against condition 2, condition 1 against condition 3, condition 2 against condition 3, but this turns out to be not as good.

Finding the variance

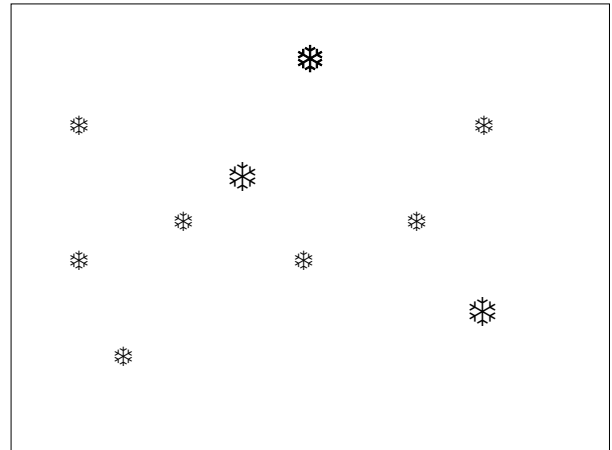
- The idea of the ANOVA is to divide up the total variance in the data into parts (to “account for the variance”):
 - Within group variance (variance that arises within a single condition)
 - Between group variance (variance that arises between different conditions)
- | ANOVA: | SS | df | MS | F | p | Fc |
|----------------|-----|----|----|------|-------|------|
| between groups | ... | 5 | .. | 2.45 | 0.045 | 2.39 |
| within groups | ... | 54 | .. | | | |
| total | | | | | | |

Confidence intervals

- As well as trying to decide if your observed sample is within what you'd expect your estimated distribution to provide, you can kind of run this logic in reverse as well, and come up with a confidence interval:
- Given where you see the measurements coming up, they must be 68% likely to be within 1 CI of the mean, and 95% likely to be within 2 CI of the mean, so the more measurements you have the better guess you can make.
- A 95% CI like $209.9 < \mu < 523.4$ means “we're 95% confident that the *real* population mean is in there”.
 - =CONFIDENCE(0.05, STDEV(sample), COUNT(sample))

Correlation and Chi square

- Correlation between two measured variables is often measured in terms of (Pearson's) r .
- If r is close to 1 or -1, the value of one variable can predict quite accurately the value of the other.
- If r is close to 0, predictive power is low.
- Chi-square test is supposed to help us decide if two conditions / factors are independent of one another or not. (Does knowing one help predict the effect of the other?)



So...

- There's still work to be done. Since I'm not sure exactly *what* work that is, once again... no lab work to do.
- Places to go:
 - <http://davidmlane.com/hyperstat/>
 - <http://www.stat.sc.edu/webstat/>