experimental design for linguists - pt. 2

HPSG 2012

PHILIP HOFMEISTER
UNIVERSITY OF ESSEX

Experimental Control

CONTROL

- Typically, an experimenter is interested in how one or more variables affect an outcome *X* (e.g. judgments, reading times, speech onset times)
  - but NOT what sorts of things affect *X*

CONTROL

- Everything that is not of interest should be kept constant as much as possible
  - Reduces chances that any observable effects are due to something besides predictor variables

CONTROL

- Two kinds of unwanted variation:
  - Variation not associated with independent variable(s)
  - Variation associated with independent variable(s) = confounds

CONTROL

- What sorts of things influence linguistic experiments (particularly judgment tasks)?
  - Order of presentation
  - Lexical factors (frequency, abstractness, collocational frequency)
  - Plausibility & context
  - Complexity

CONTROL

- Order of presentation
  - Response times almost always get faster throughout an experiment
  - Judgments for a variety of sentence types get higher with repeated exposure
    - Linguist's disease
    - Satiation
    - Priming

ORDER EFFECTS

- Such effects can be minimized by randomization
  - Note: the efficacy of randomization increases as you increase the # of participants

1	10
3	8
4	3
2	2
5	7
10	9
9	4
7	2
8	6
6	1

ORDER EFFECTS

- How do you randomize?
  - Some experimental programs will do this for you (e.g. Linger, Turkolizer)
  - You can write your own randomization script
  - Commercially available options

COUNTER-BALANCING

- Imagine the following sequence of trials:

1. *I know what who bought.*
2. *Money is tight for many people now.*
3. *I know which present who bought.*

COUNTER-BALANCING

- The response to (3) may be affected by the response to (1) since they are different conditions of the same item

1. *I know what who bought.*
2. *Money is tight for many people now.*
3. *I know which present who bought.*

COUNTER-BALANCING

- Each subject should see each item in only one condition

COUNTER-BALANCING

	List 1	List 2	List 3	List 4
Item 1	Cond1	Cond2	Cond3	Cond4
Item 2				
Item 3				
Item 4				
Item 5				
...				
Item *n*				

COUNTER-BALANCING

	List 1	List 2	List 3	List 4
Item 1	Cond1	Cond2	Cond3	Cond4
Item 2	Cond2	Cond3	Cond4	Cond1
Item 3	Cond3	Cond4	Cond1	Cond2
Item 4	Cond4	Cond1	Cond2	Cond3
Item 5	Cond1	Cond2	Cond3	Cond4
. . .	. . .	. . .	. . .	. . .
Item *n*	Cond4	Cond1	Cond2	Cond3

COUNTER-BALANCING

- This method of counterbalancing (called a Latin Square design) means each list will have an equal # of items in condition A, B, C, etc.
- Minimizes chances of list effects, but does not rule them out

COUNTER-BALANCING

- An equal # of participants should see each list; # of participants needed is a multiple of the number of condition/factor levels

CONFOUNDS

- Resumptive pronouns
  - There was a prisoner that the guard helped him/___ to make a daring escape.
  - There was a prisoner that the officials confirmed that the guard helped him/___ to make a daring escape.

CONFOUNDS

- Sample question:
  - Is a resumptive pronoun more acceptable as depth of embedding increases?

CONFOUNDS

- There was a prisoner that the officials confirmed that the guard helped him to make a daring escape.
- There was a prisoner that the guard helped him to make a daring escape.

CONFOUNDS

- There was a prisoner that the officials confirmed that the guard helped him to make a daring escape.
- There was a prisoner that the guard helped him to make a daring escape.

Sentences differ in length, meaning, & complexity

A BETTER QUESTION

- Does the difference between gaps & resumptives increase significantly with embedding?

BIASING AGAINST YOUR HYPOTHESIS

- Consider creating materials that work against your hypothesis
- e.g. longer sentences = lower judgments
  - *Which book did which student read?*
  - *What did who read?*

FILLERS

- Fillers/distractors should reduce the salience of the critical items

FILLERS

HPSG2012

FILLERS

- Imagine an experiment with only multiple wh-questions

FILLERS

- Imagine an experiment with only multiple wh-questions
  - *What did who buy?*

HPSG2012

FILLERS

- Imagine an experiment with only multiple wh-questions
  - *What did who buy?*
  - *Who saw what?*

FILLERS

- Imagine an experiment with only multiple wh-questions
  - *What did who buy?*
  - *Who saw what?*
  - *Which medicine does who get?*

FILLERS

- Imagine an experiment with only multiple wh-questions
  - *What did who buy?*
  - *Who saw what?*
  - *Which medicine does who get?*
  - *Which invention did which inventor make?*

FILLERS

- Fillers/distractors should thus reduce the salience of the critical items
  - Rule of thumb: The weirder the items, the more fillers needed

FILLERS

- Sometimes, your materials may not need any fillers (but this is the exception rather than the rule)

ANTECEDENT COMPLEXITY

- A group of military advisers met with a (ruthless military) dictator to discuss the recent election results. It had been necessary to use intimidation and violence to beat the rival political party. Some advisers suggested releasing some political prisoners as a gesture of peace, but he rejected the suggestion outright.

FILLERS

- Other notes on fillers
  - Fillers & critical items should be interleaved, e.g.
    - Item 1 = FILLER
    - Item 2 = CRITICAL ITEM
    - Item 3 = FILLER
    - etc.

FILLERS

- How many? What type?
  - No hard & fast rule, but > 2x the # of critical items is common

FILLERS

- Imagine looking for a difference between
  - *Who did you buy a picture yesterday at the market of?*
  - *Who did you buy a picture yesterday of at the market?*

FLOOR & CEILING EFFECTS

- Even if you don't ask them to, participants will partly base their ratings on their previous judgments

FILLERS

- In acceptability studies, it’s prudent to have some fillers intuitively better and worse than your critical items
  - Spreads out judgments & reduces chances of floor/ceiling effects

DEBRIEF

- Ask participants what they thought the experiment was about at the end

PARTICIPANT CONTROL

- Most psychology & linguistics experiments draw from college age (18-22) participants
  - High education
  - Young
  - Socioeconomic class

PARTICIPANT CONTROL

- The important question is: do you have reason to suspect your results will not generalize if you had chosen a different sample?

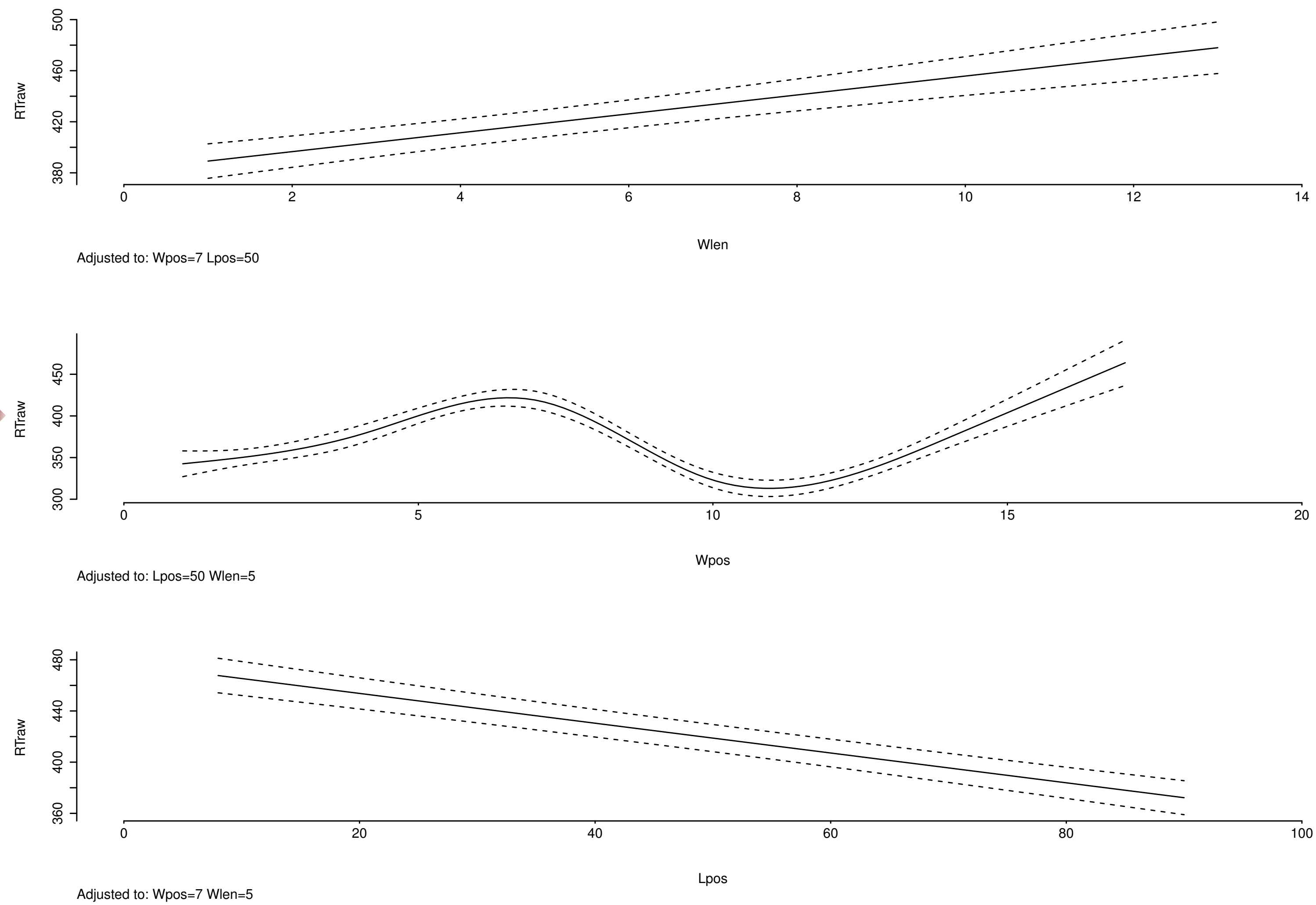
PARTICIPANT CONTROL

- Collect demographic information where possible
- Determine whether there is significant variation in the data due to individual-level characteristics

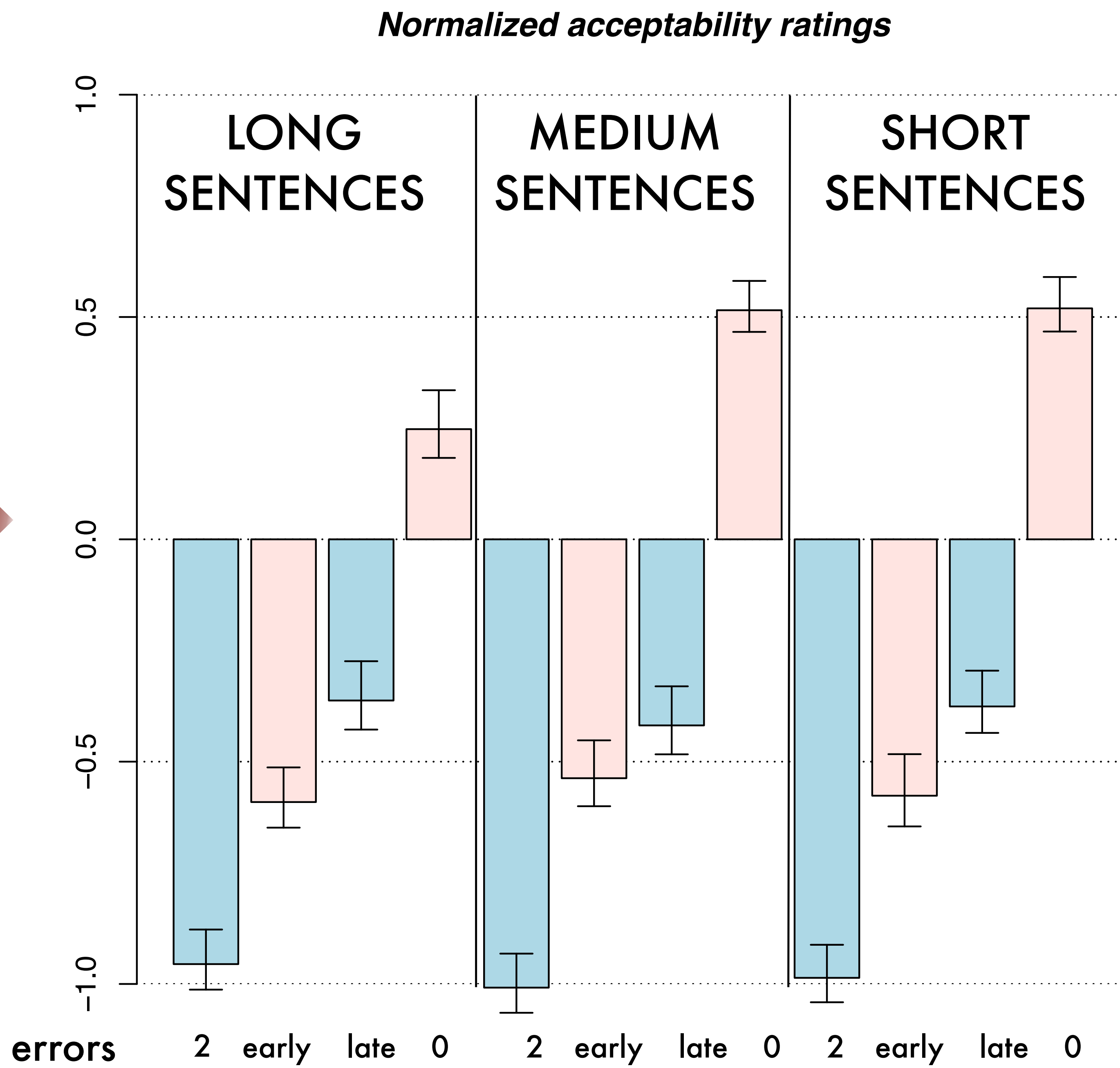
ITEM CONTROL

- There a number of frequently uncontrolled variables in linguistic experiments:
  - Plausibility = contextualizability
  - Sentence length
  - Word length
    - Slower words read longer / responded to more slowly
  - Complexity

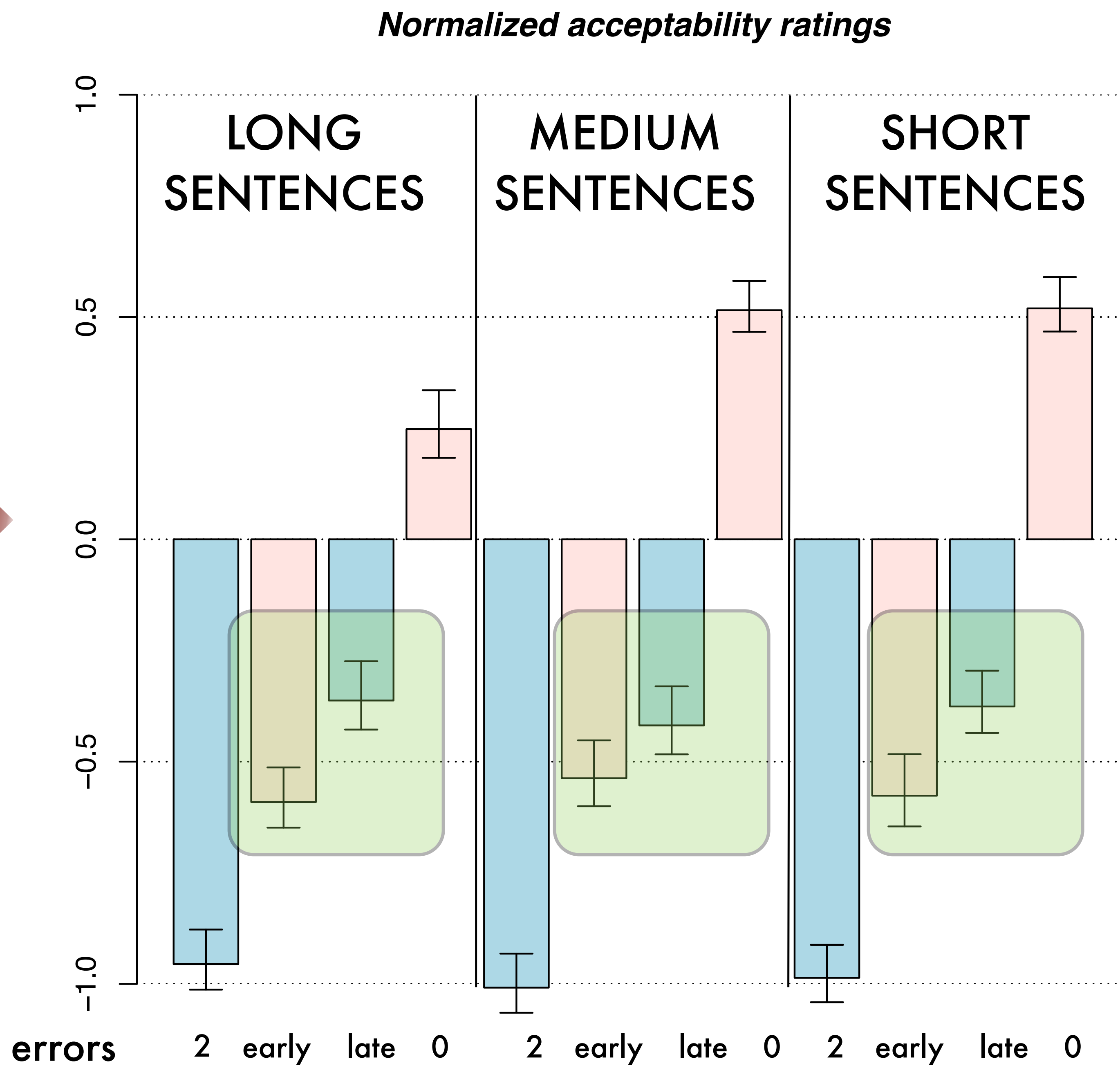
TYPICAL SOURCES OF ERROR



POSITION EFFECTS



POSITION EFFECTS



GENERALIZE

- Recall that the purpose of doing experiments with samples is to generalize to a population
  - In the case of language, we are trying to capture how people use and represent language generally
  - This means that results are more robust as the number of items increases, but also . . .

GENERALIZE

- Laboratory settings and experiments are not normal
- People don't rate sentences for acceptability in everyday life
  - It's in the researcher's interest to offset this unnaturalness as much as possible

GENERALIZE

- What can be done to increase the ecological validity of linguistic experiments?
- Where possible, use
  - Context (see Bolinger 1968, Bever 1970, Schütze 1996)
  - Attested sentences to create materials
  - Plausible examples

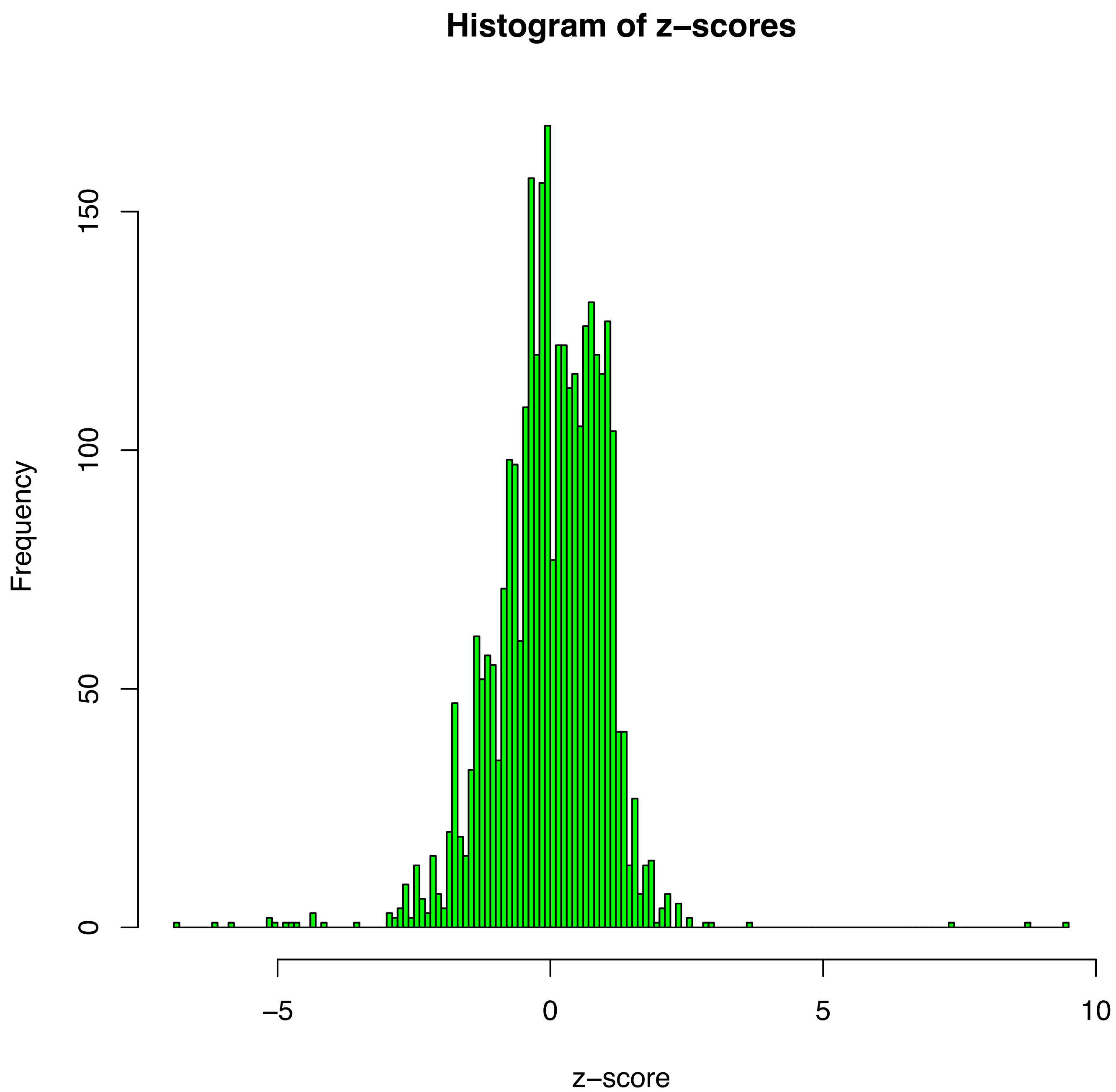
Understanding Your Data

- At some point, you need to analyze your data
- This means some statistics, but modern day statistical programs (e.g. SPSS, R) mean that you don't need to be an expert at the underlying math

OUTLIERS

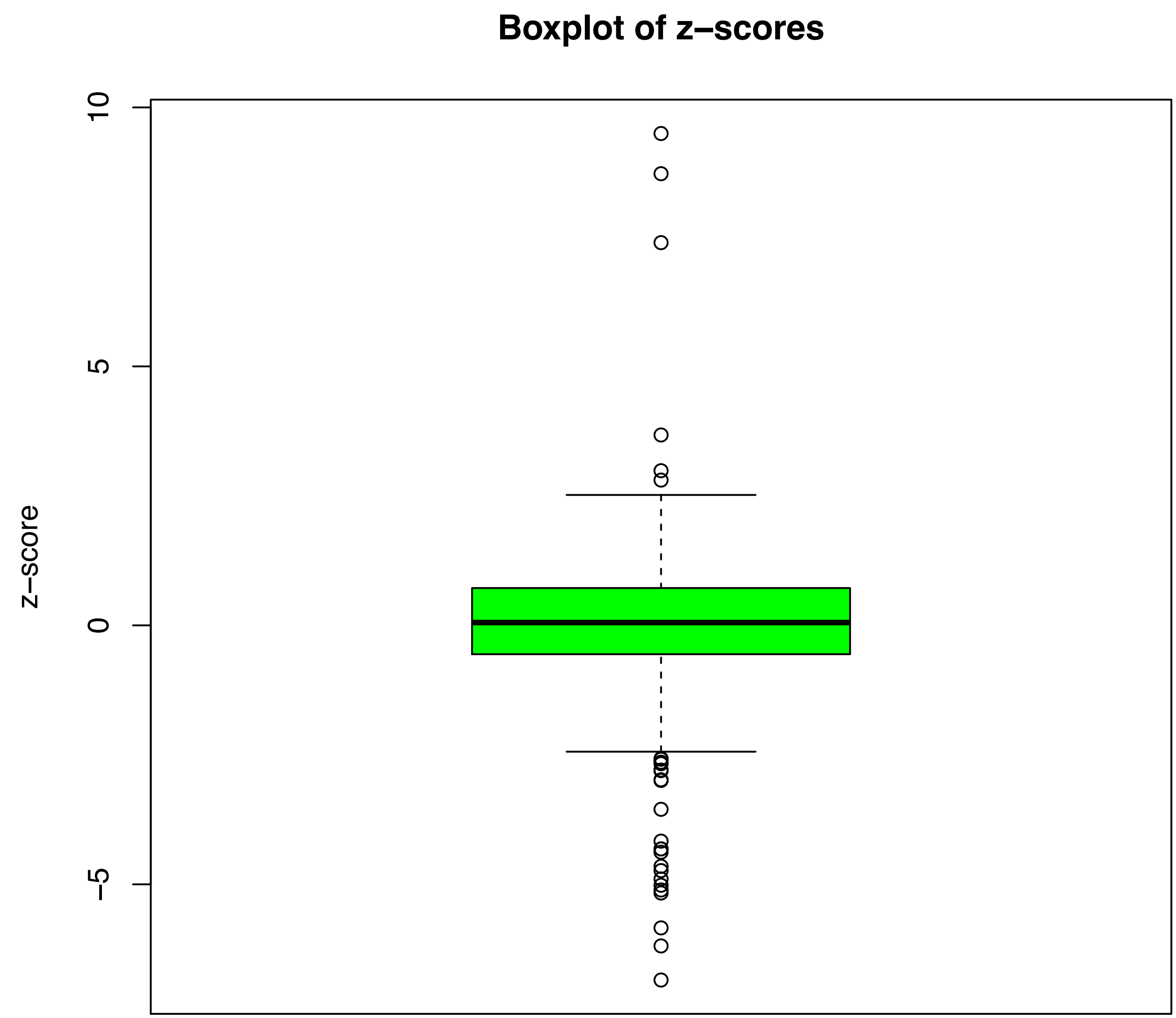
- Some data points result from
  - Distraction/lack of attention
  - Annoyance
  - Misunderstanding
  - Uncooperative participants

VISUALIZE



HPSG2012

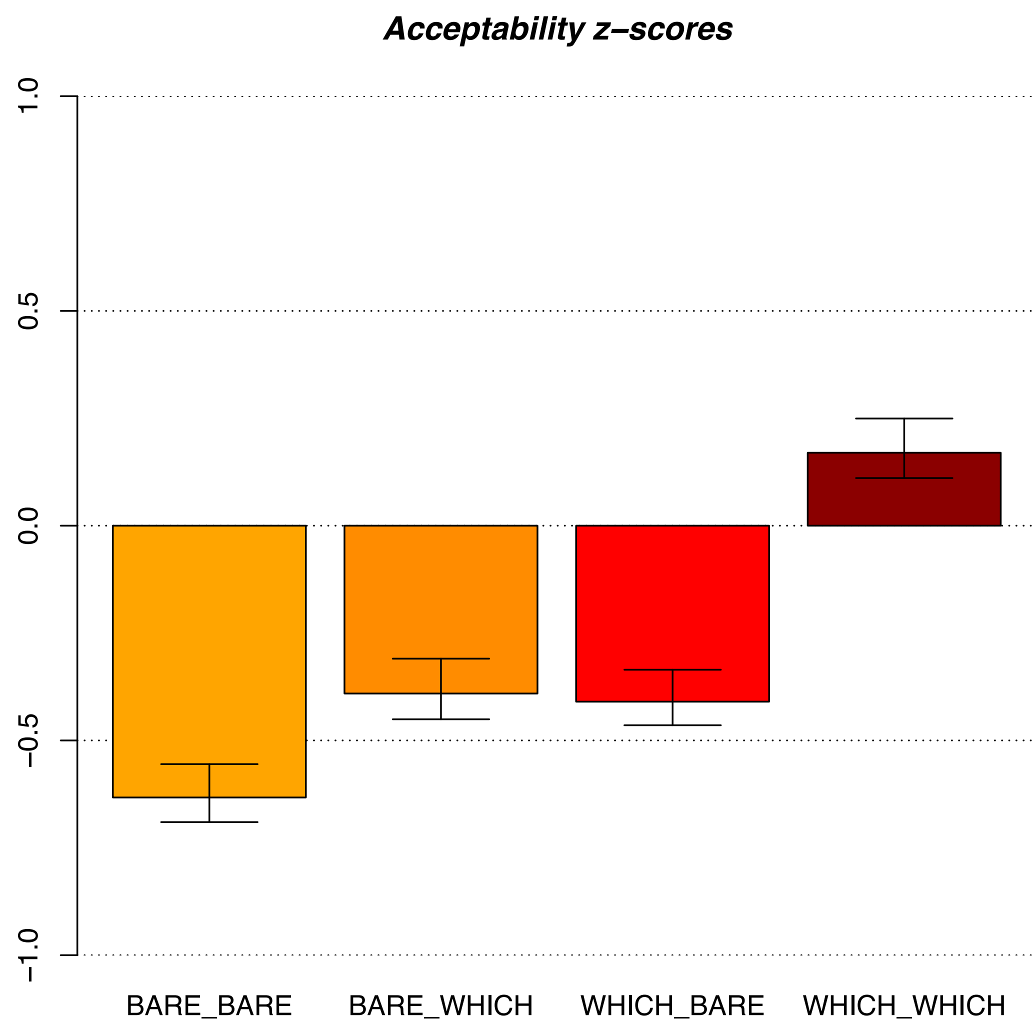
VISUALIZE



HPSG2012

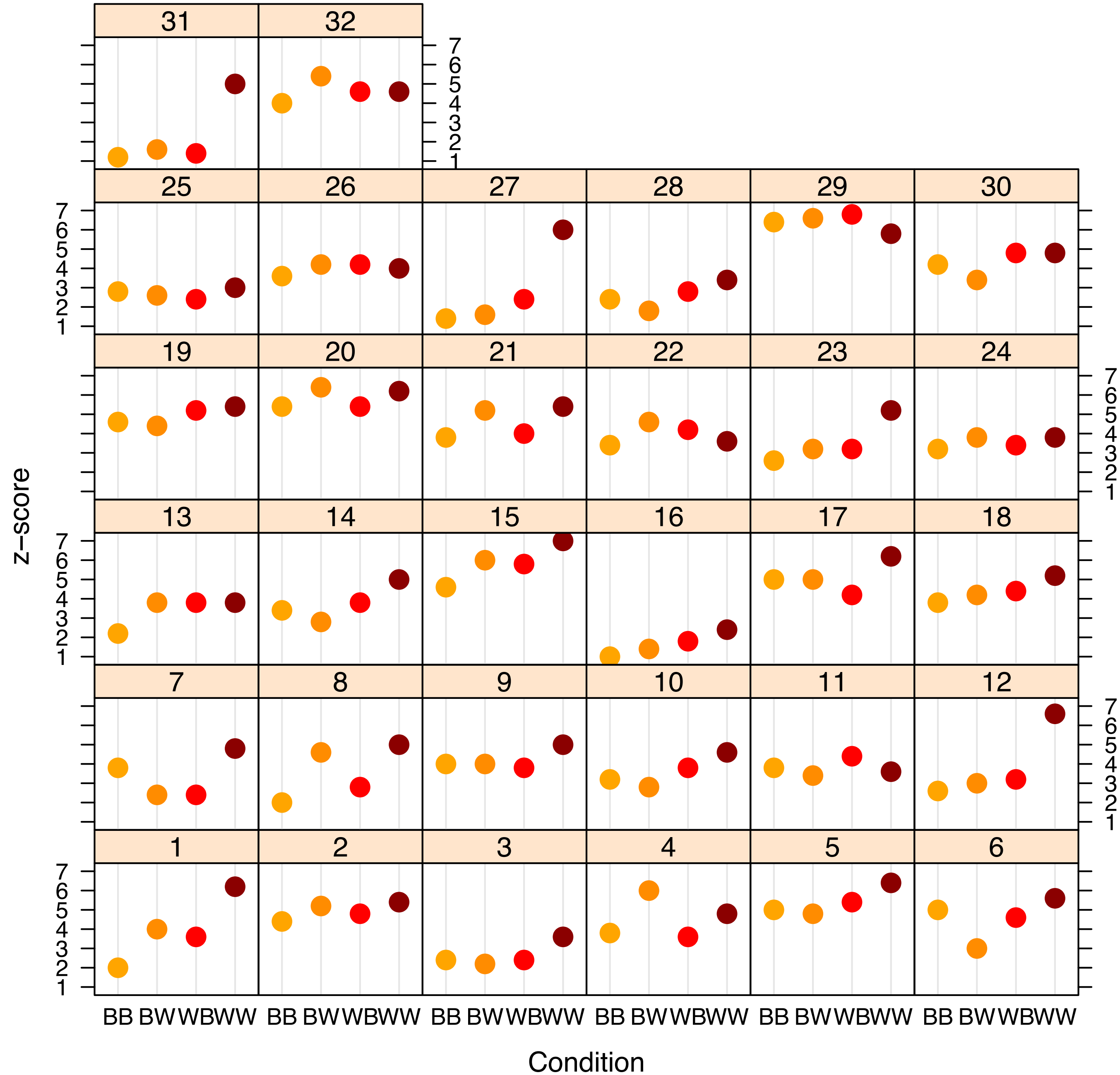
OUTLIERS

- Ideally, no data is removed, but this is often not justifiable
- Criteria for outlier removal:
  - Standard deviations
  - Cutoffs
  - Cook's distance



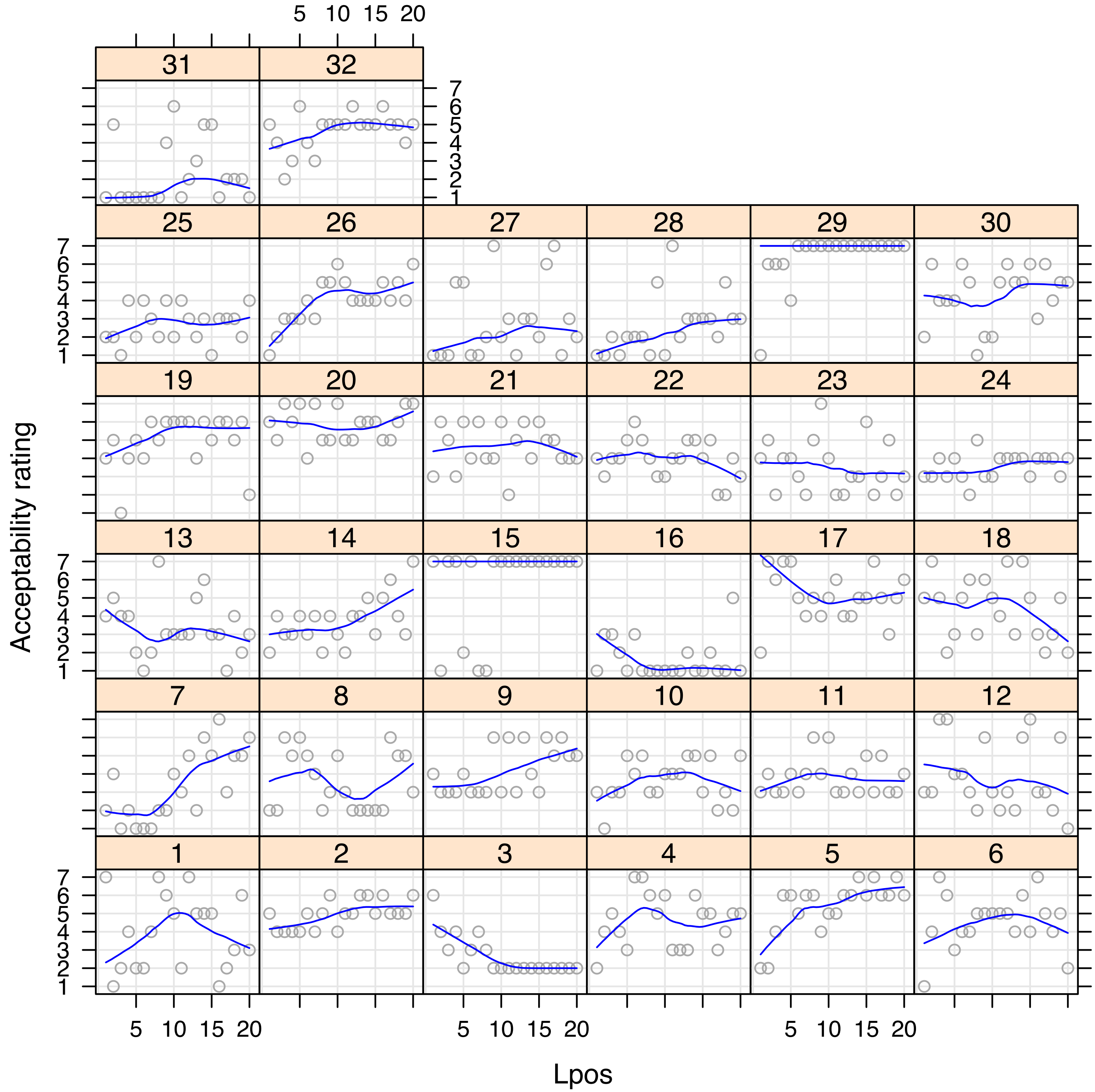
We knew what who needs.
We knew what which patient needs.
We knew which medicine who needs.
We knew which medicine which patient needs.

SUBJECTS

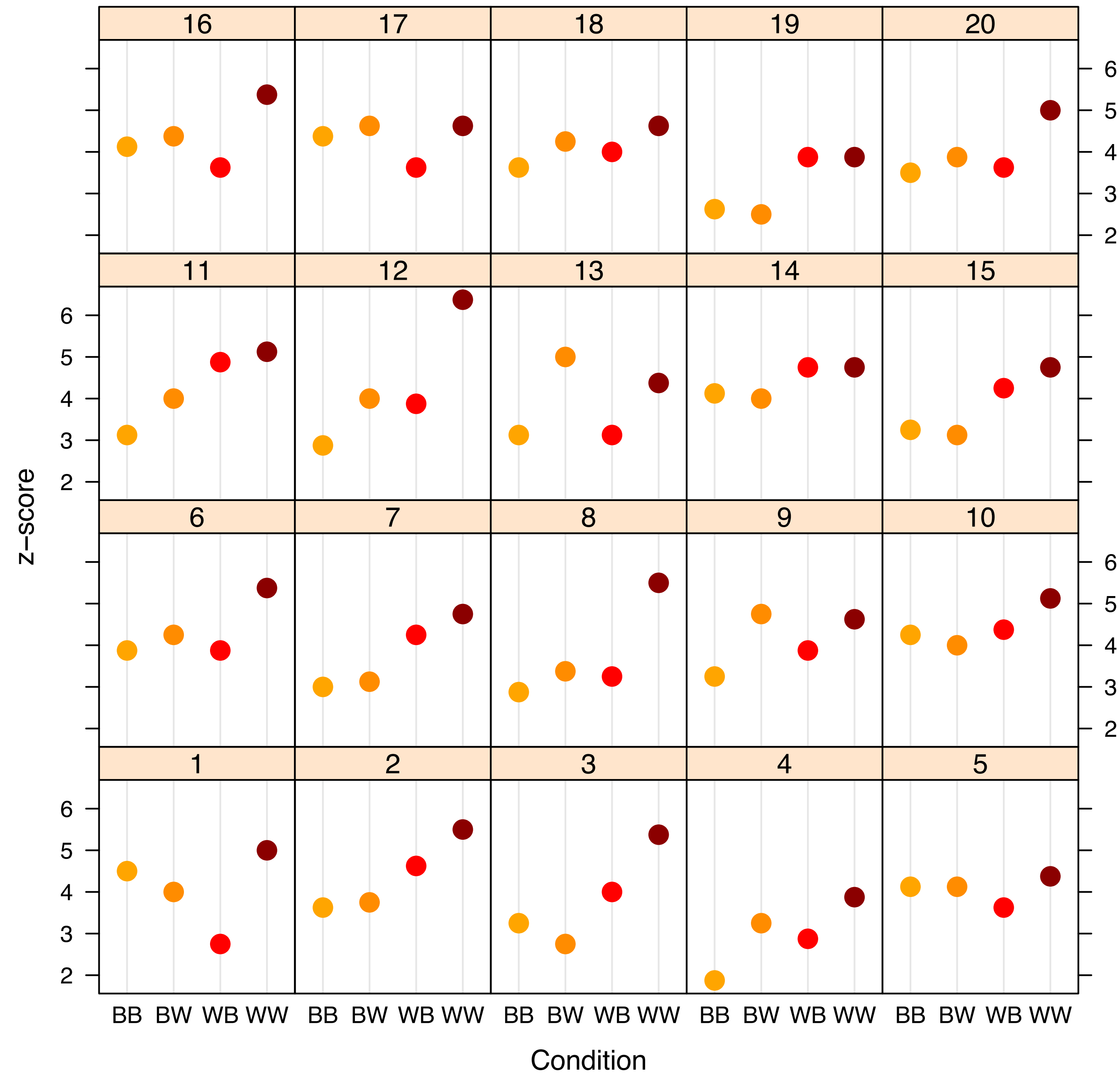


z–score
Condition
BB BW WB WW
31
32
25
26
27
28
29
30
19
20
21
22
23
24
13
14
15
16
17
18
7
8
9
10
11
12
1
2
3
4
5
6

LIST POSITION EFFECTS BY SUBJECT



ITEMS


16
17
18
19
20
11
12
13
14
15
6
7
8
9
10
1
2
3
4
5
z–score
2
3
4
5
6
BB BW WB WW
Condition

EVALUATING DATA

- For experimental syntax, we are often interested in the comparison of 2 or more conditions

ONE FACTOR W/ 3 LEVELS

cond1	cond2	cond3
5	1	7
3	3	5
4	2	5
5	3	5
7	7	7

DATA ANALYSIS

- It's straightforward to calculate MEANS for each of our conditions
  - cond1 = 24/5 = 4.8
  - cond2 = 16/5 = 3.2
  - cond3 = 28/5 = 5.6

- The question is: are these differences reliable or are they due to chance?
  - cond1 = 24/5 = 4.8
  - cond2 = 16/5 = 3.2
  - cond3 = 28/5 = 5.6

- Note, there are several things this question could mean:
  - Is there a difference due to the factor of interest GENERALLY?
  - Are specific levels different from each other?

- Is there a difference due to the factor of interest GENERALLY? = Repeated measures ANOVAs
- Are specific levels different from each other? = independent t-tests

T-TESTS

- When comparing 2 factor levels in a within-subjects design, t-tests are a common tool

T-TESTS

- Intuitively speaking, a t-test looks at the difference between 2 conditions, the observed variation around the means, and tells us the probability that the means are different in the population

- Reliability of the mean is reflected in the standard error
    - $SE = \sigma / \sqrt{n}$
    - σ = standard deviation
    - n = # of observations

- To figure out whether two condition means are reliably different
  - Compare the difference in means to the standard errors

ONE FACTOR W/ 3 LEVELS

cond1	cond2	cond3
5	1	7
3	3	5
4	2	5
5	3	5
7	7	7

$\bar{X}_1 = 4.8$ $\bar{X}_2 = 3.2$ $\bar{X}_3 = 5.8$

SD = .663 SD = 2.28 SD = 1.10

COMPARE INDIVIDUAL LEVELS

$$t = \frac{\bar{X}_2 - \bar{X}_1}{\sqrt{\frac{\sigma_2^2}{n_2} + \frac{\sigma_1^2}{n_1}}}$$

cond1	cond2	cond3
5	1	7
3	3	5
4	2	5
5	3	5
7	7	7

$\bar{X}_1 = 4.8$, SD = 1.48

$\bar{X}_2 = 3.2$, SD = 2.28

$\bar{X}_3 = 5.8$, SD = 1.10

COMPARE INDIVIDUAL LEVELS

$$t = \frac{\bar{X}_2 - \bar{X}_1}{\sqrt{\frac{\sigma_2^2}{n_2} + \frac{\sigma_1^2}{n_1}}}$$

$$t = \frac{3.2 - 4.8}{\sqrt{\frac{2.28^2}{5} + \frac{1.48^2}{5}}}$$

cond1	cond2	cond3
5	1	7
3	3	5
4	2	5
5	3	5
7	7	7

$\bar{X}_1 = 4.8$ SD = 1.48 $\bar{X}_2 = 3.2$ SD = 2.28 $\bar{X}_3 = 5.8$ SD = 1.10

COMPARE INDIVIDUAL LEVELS

$$t = \frac{\bar{X}_2 - \bar{X}_1}{\sqrt{\frac{\sigma_2^2}{n_2} + \frac{\sigma_1^2}{n_1}}}$$

$$t = \frac{3.2 - 4.8}{\sqrt{\frac{2.28^2}{5} + \frac{1.48^2}{5}}}$$

$$t = \frac{3.2 - 4.8}{\sqrt{1.040 + .438}}$$

cond1	cond2	cond3
5	1	7
3	3	5
4	2	5
5	3	5
7	7	7

$\bar{X}_1 = 4.8$, SD = 1.48

$\bar{X}_2 = 3.2$, SD = 2.28

$\bar{X}_3 = 5.8$, SD = 1.10

COMPARE INDIVIDUAL LEVELS

$$t = \frac{\bar{X}_2 - \bar{X}_1}{\sqrt{\frac{\sigma_2^2}{n_2} + \frac{\sigma_1^2}{n_1}}}$$

$$t = \frac{3.2 - 4.8}{\sqrt{\frac{2.28^2}{5} + \frac{1.48^2}{5}}}$$

$$t = \frac{3.2 - 4.8}{\sqrt{1.040 + .438}}$$

$$t = 1.32$$

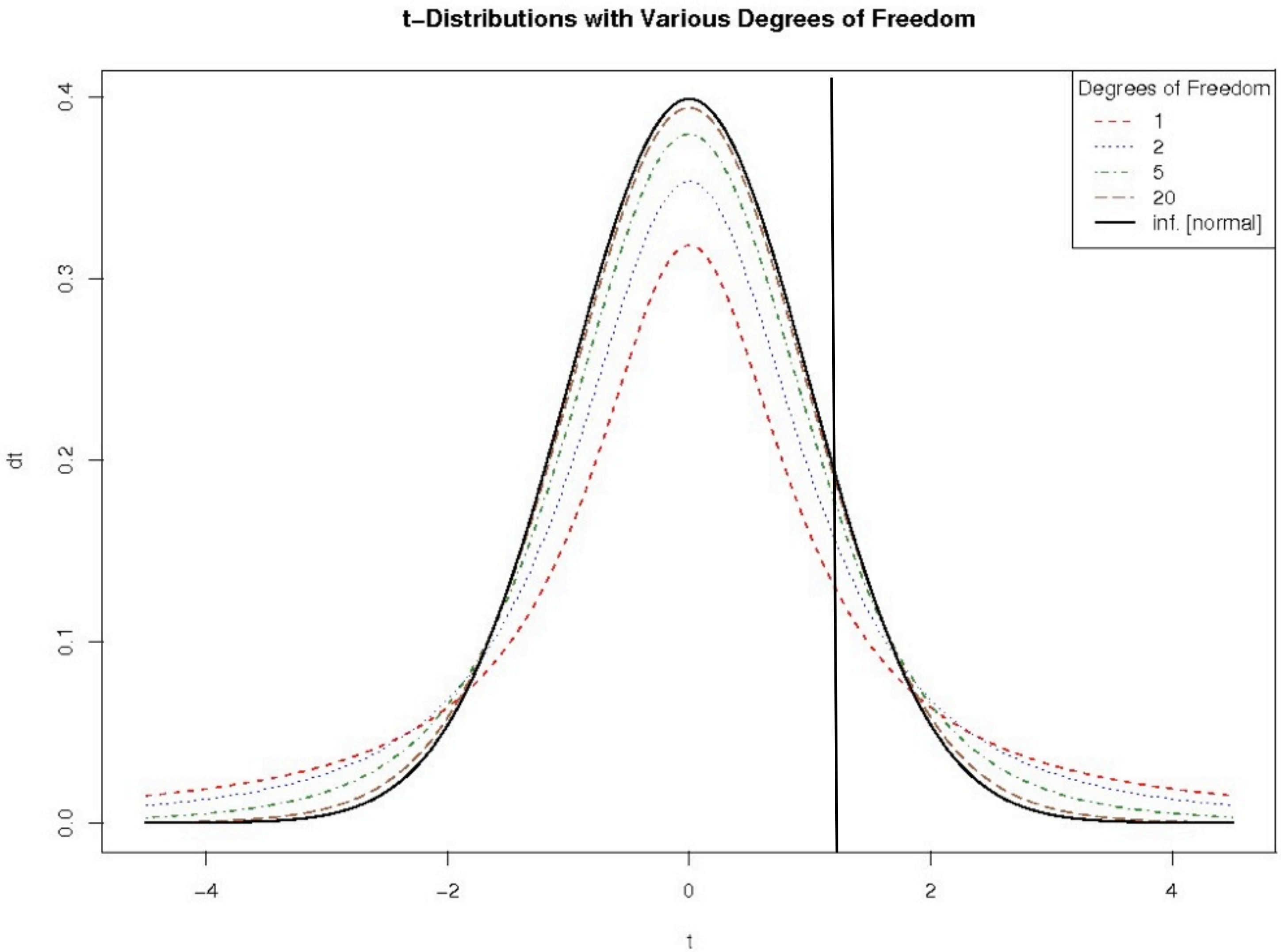
cond1	cond2	cond3
5	1	7
3	3	5
4	2	5
5	3	5
7	7	7

$\bar{X}_1 = 4.8$ SD = 1.48

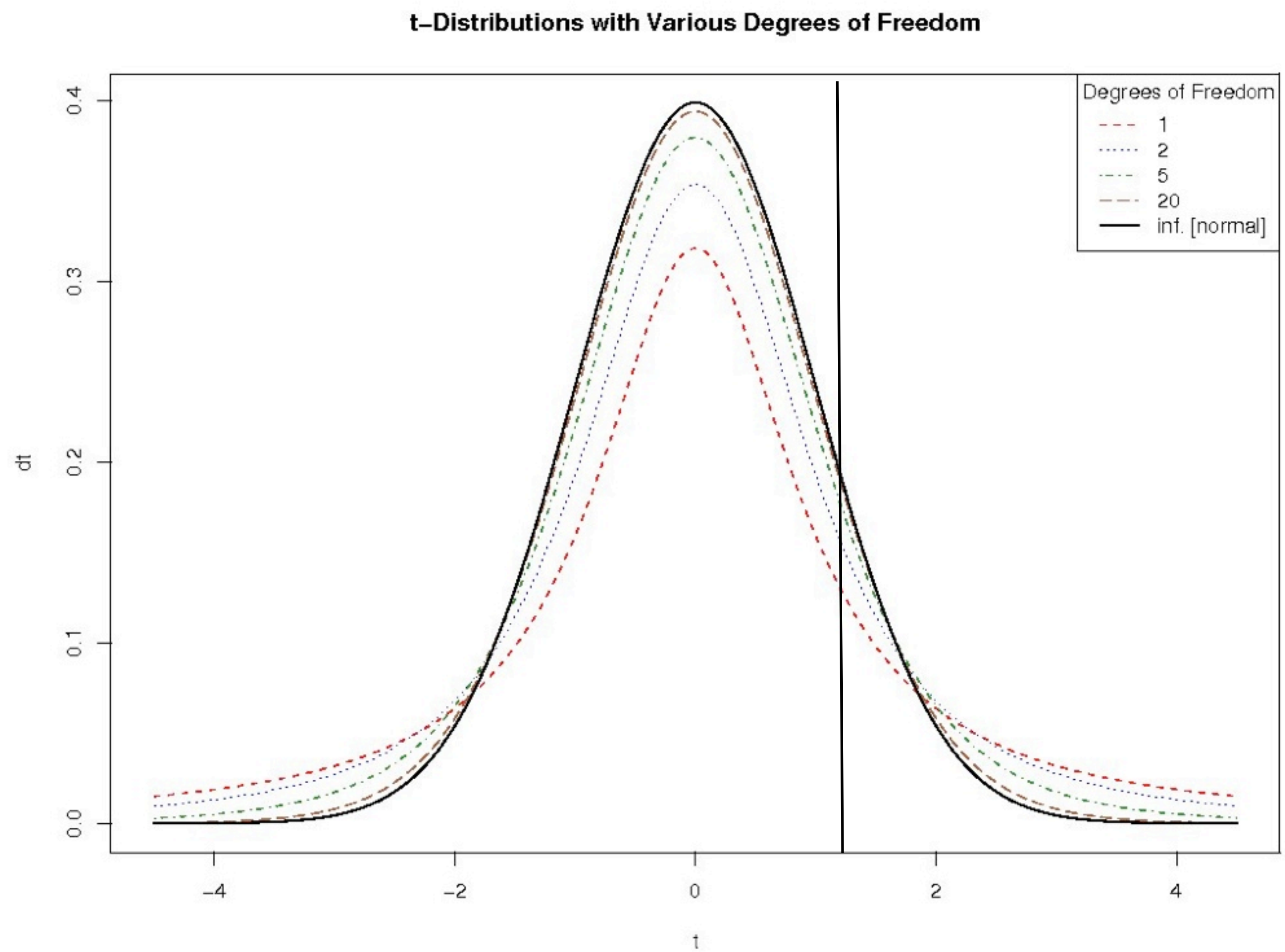
$\bar{X}_2 = 3.2$ SD = 2.28

$\bar{X}_3 = 5.8$ SD = 1.10

COMPARE INDIVIDUAL LEVELS



COMPARE INDIVIDUAL LEVELS



$$t = 1.32, df = 8, \text{p} = .22$$

THE LOGIC OF ANOVAS

- Analyses of variance is commonly used to determine whether there is an effect of a factor with three more or levels

- Several sources of possible variation
  - Variation due to independent variable
  - Variation due to error (participants or items)

cond1	cond2	cond3
2	4	6
2	4	6
2	4	6
2	4	6
2	4	6

$\bar{X}_1 = 2$ $\bar{X}_1 = 4$ $\bar{X}_1 = 6$

SD = 0 SD = 0 SD = 0

cond1	cond2	cond3
2	4	7
3	6	7
1	7	4
1	2	5
3	1	7

$\bar{X}_1 = 2$ $\bar{X}_1 = 4$ $\bar{X}_1 = 6$

SD = 1 SD = 2.55 SD = 1.41

WHY NOT JUST DO MULTIPLE T-TESTS?

- Imagine you have 100 sentence pairs (e.g. grammatica/ungrammatical) and want to tell whether there are significant differences

WHY NOT JUST DO MULTIPLE T-TESTS?

- Each t-test performed has a 1/20 (=.05) chance of returning a spuriously significant result

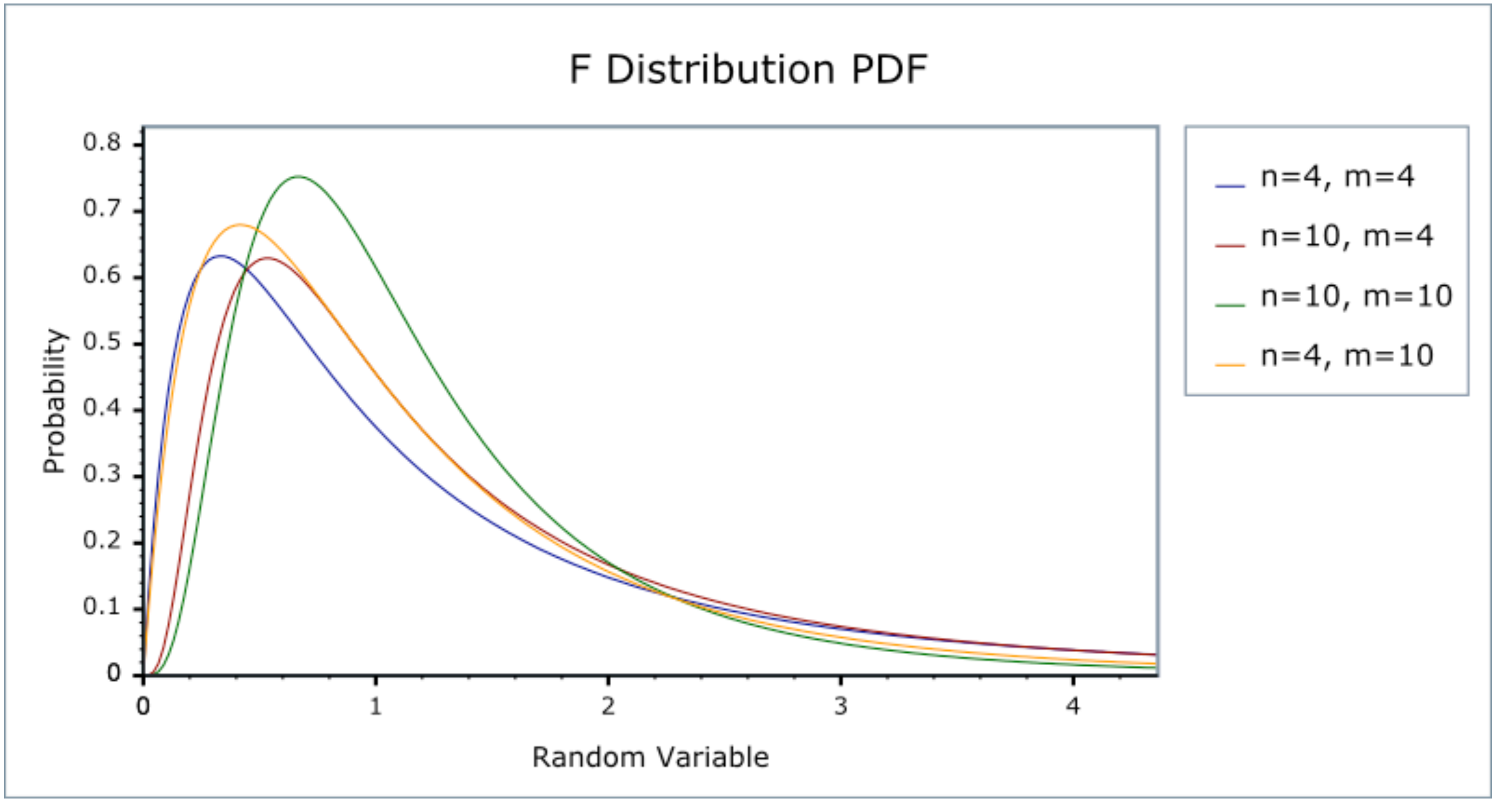
LOGIC OF ANOVAS

- Calculate how much each condition/ factor level differs from the grand mean
- Calculate how much data point differs from its condition mean

LOGIC OF ANOVAS

$$F = \frac{\text{Summed variance between conditions/factor levels}}{\text{Summed variance within conditions/factor levels}}$$

- If F is ≤ 1, we can be confident that there is no effect of treatment


F Distribution PDF
Probability
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
1
2
3
4
Random Variable
n=4, m=4
n=10, m=4
n=10, m=10
n=4, m=10

SUMMARY

- The underlying logic of many classical statistical analyses relies on comparing the difference between groups/conditions and the variance within those groups

SUMMARY

- Traditional techniques in linguistic theorizing do not allow us to gage the within-group/within-condition variance

TIPS

TIPS

- Tip #1: Record everything you can!
  - Keep a binder for each study & a runsheet for each participant
  - Note time, date, any observations about the participant
  - Months or years later, you won't remember anything about the session
  - Obviously, backup data

TIPS

- Tip #2: Get as much data as you can even if you don't plan to use it
  - Since individuals vary so much, collect as much individual data as possible
  - Consider testing subjects on standard neuropsych batteries, e.g. verbal fluency tests, reading span or other memory tests, vocabulary tests, etc.

TIPS

- Tip #3: Take the experiment yourself (or have a friend / colleague take it)
  - Confounds become most obvious when you actually sit there and see/hear the stimuli

TIPS

- Tip #4: Keep designs simple
  - In a 2 x 2 x 3 x 2 design, it’s hard to make clear predictions and there’s lots of room for random noise
  - Simpler designs = fewer subjects & items

TIPS

- Tip #5: Look at the data before analyzing it
  - How do participants differ from one another? How many show effects of the experimental manipulation? Are there certain items driving your effects?

end part 2