UCL Department of Experimental Psychology

Keeping Data Analysis Simple – Or Why I Love the `emmeans` `R` Package

Henrik Singmann

http://singmann.org/

h.singmann@ucl.ac.uk

Two Parts

Contrasts for Factors With More Than 2 Levels

A General Approach for Testing Omnibus Tests

“Two Statistics”

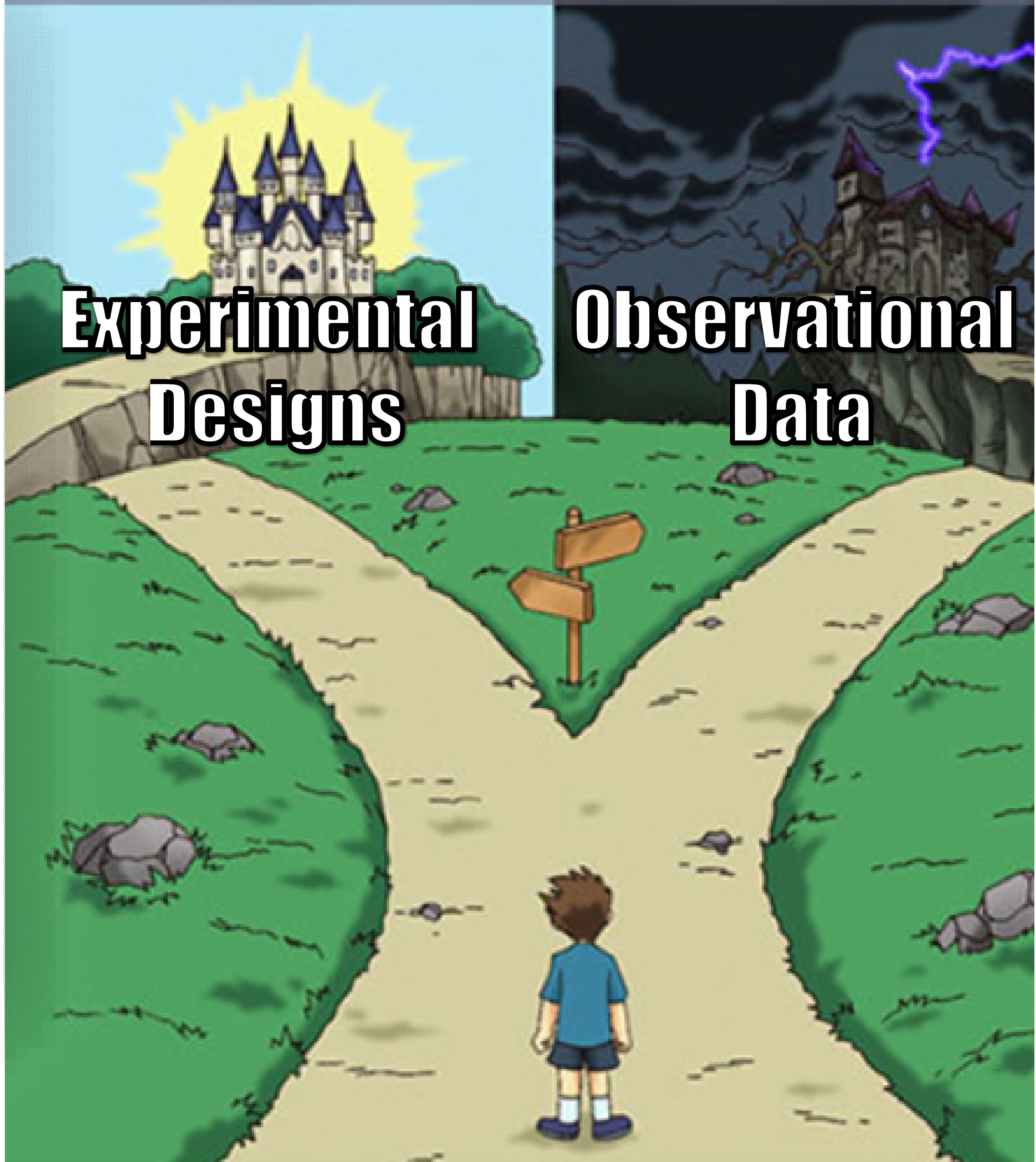
Statistics is tool that helps us answer substantive questions!

Talk applies to experimental designs:

- IVs are **categorical** (factors)
- IV is **randomly assigned** (i.e., manipulated) and DV is measured
- Inferential and substantial goal is **hypothesis test**: Is there effect of IV on DV?
- Random assignment allows **causal interpretation**: IV is cause of change in DV

Different considerations can apply for **observational data**

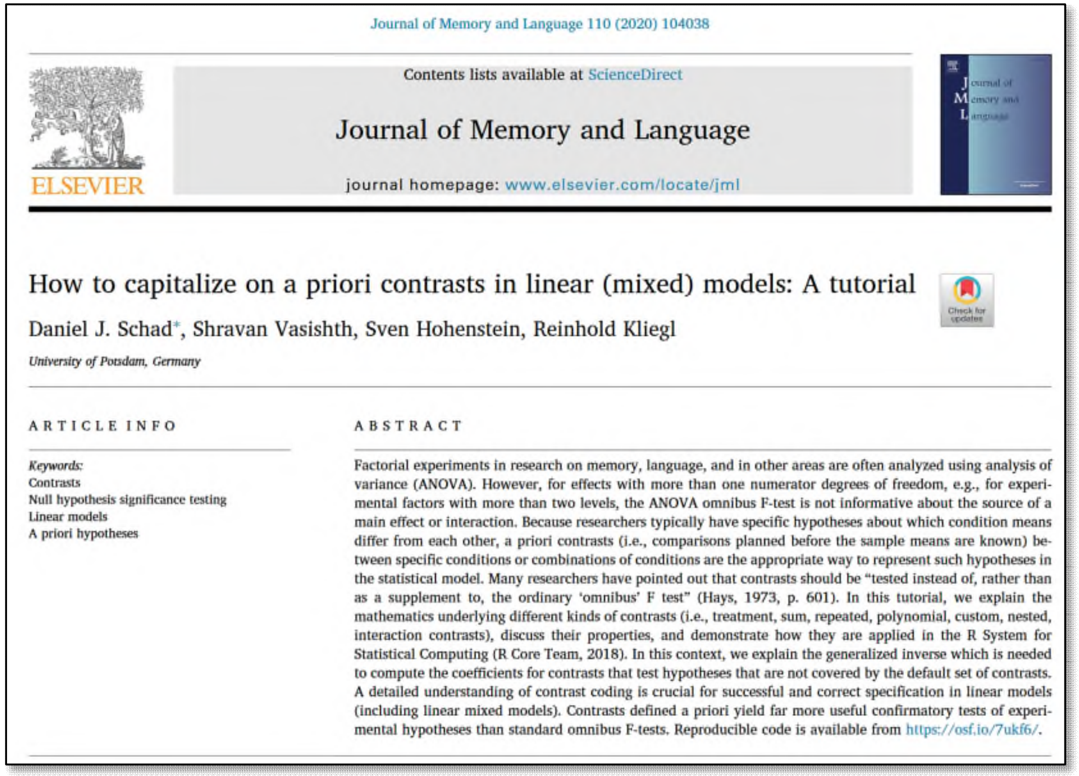
- IV is **measured** (and not manipulated) and **graded/continuous**
- Inferential goal is **estimation**: What is size and magnitude of effect of IV on DV?
- Substantive goal is **interpretation of effect** (causal or non-causal): What does effect of IV on DV mean?



Contrasts for Factors With More Than 2 Levels

My Provocative Take

The Contrast-First Approach



Journal of Memory and Language 110 (2020) 104038

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

ELSEVIER

Journal of Memory and Language

How to capitalize on a priori contrasts in linear (mixed) models: A tutorial

Daniel J. Schad*, Shravan Vasishth, Sven Hohenstein, Reinhold Kliegl

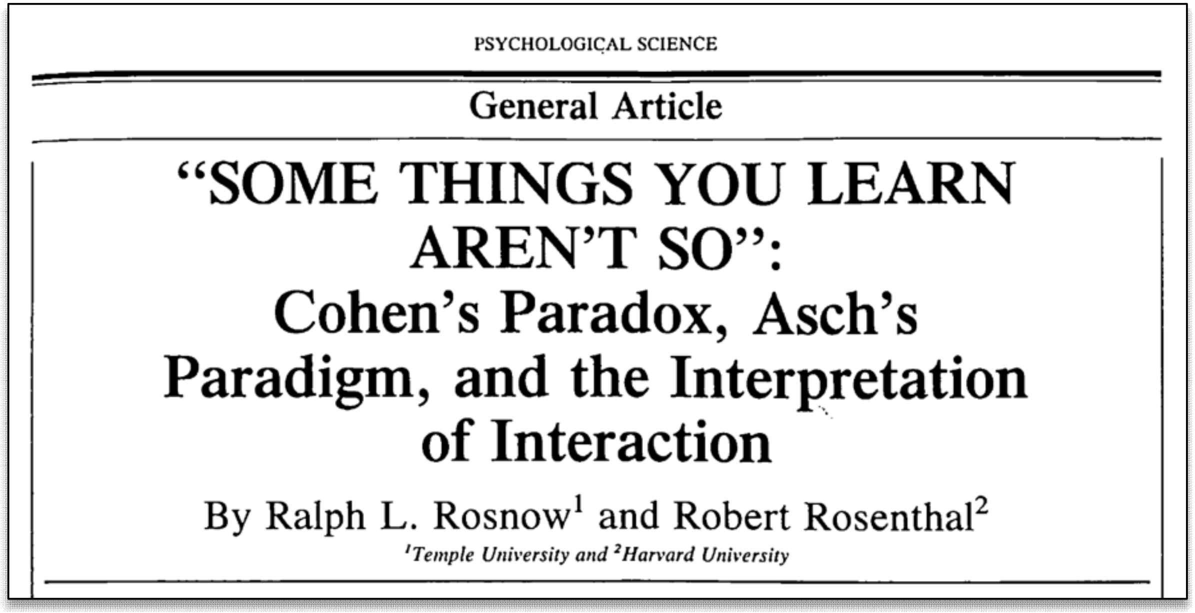
University of Potsdam, Germany

ARTICLE INFO

Keywords:
Contrasts
Null hypothesis significance testing
Linear models
A priori hypotheses

ABSTRACT

Factorial experiments in research on memory, language, and in other areas are often analyzed using analysis of variance (ANOVA). However, for effects with more than one numerator degrees of freedom, e.g., for experimental factors with more than two levels, the ANOVA omnibus F-test is not informative about the source of a main effect or interaction. Because researchers typically have specific hypotheses about which condition means differ from each other, a priori contrasts (i.e., comparisons planned before the sample means are known) between specific conditions or combinations of conditions are the appropriate way to represent such hypotheses in the statistical model. Many researchers have pointed out that contrasts should be "tested instead of, rather than as a supplement to, the ordinary 'omnibus' F test" (Hays, 1973, p. 601). In this tutorial, we explain the mathematics underlying different kinds of contrasts (i.e., treatment, sum, repeated, polynomial, custom, nested, interaction contrasts), discuss their properties, and demonstrate how they are applied in the R System for Statistical Computing (R Core Team, 2018). In this context, we explain the generalized inverse which is needed to compute the coefficients for contrasts that test hypotheses that are not covered by the default set of contrasts. A detailed understanding of contrast coding is crucial for successful and correct specification in linear models (including linear mixed models). Contrasts defined a priori yield far more useful confirmatory tests of experimental hypotheses than standard omnibus F-tests. Reproducible code is available from https://osf.io/7ukf6/.



PSYCHOLOGICAL SCIENCE

General Article

"SOME THINGS YOU LEARN AREN'T SO": Cohen's Paradox, Asch's Paradigm, and the Interpretation of Interaction

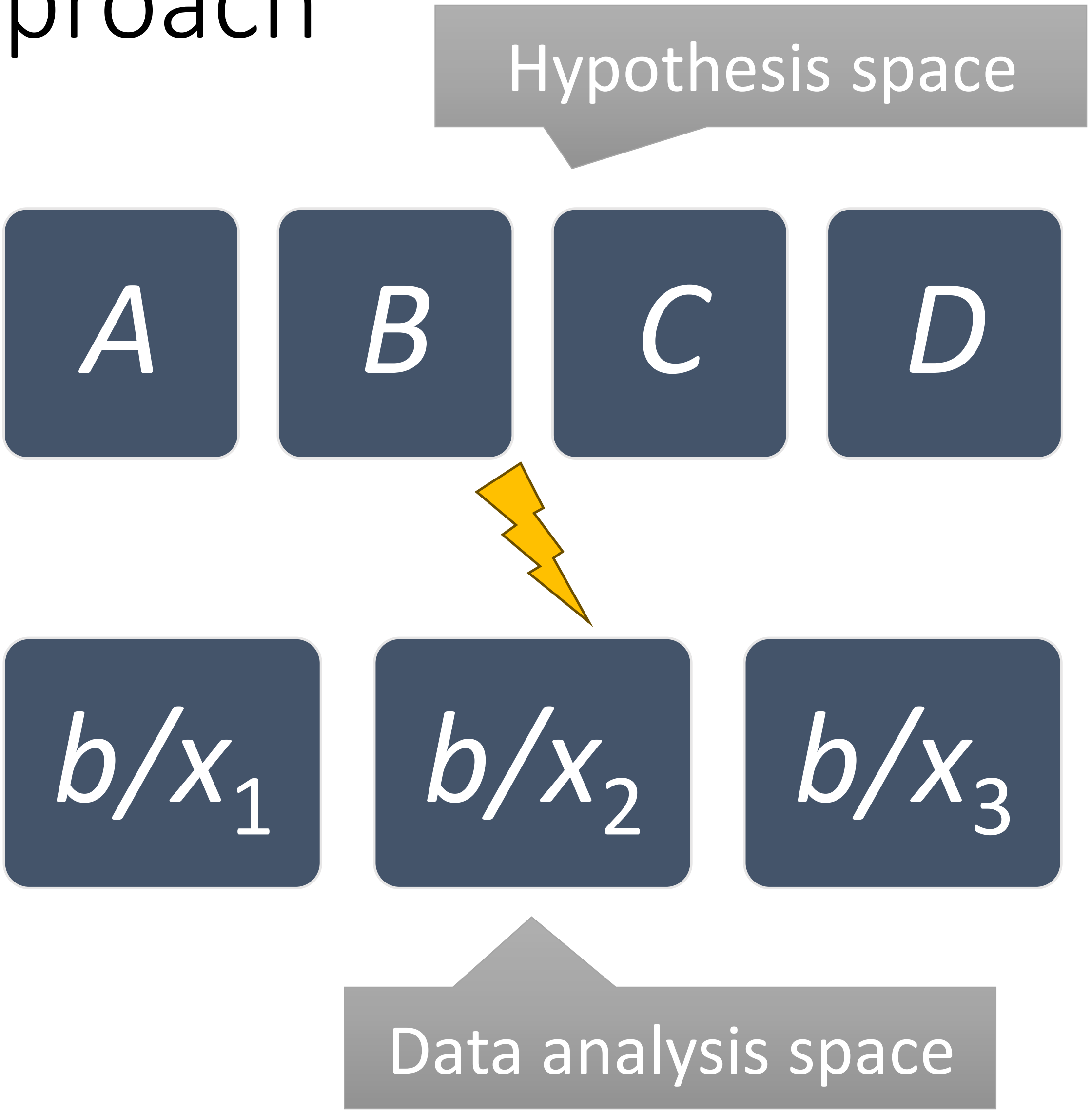
By Ralph L. Rosnow[1] and Robert Rosenthal[2]

[1]Temple University and [2]Harvard University

Steps of Contrast-First Approach

1. For factor with k levels, specify k hypotheses among cell means (based on theory/EDA)
2. Transform k hypothesis into set of $k - 1$ non-redundant contrasts
3. Estimate model with specified contrasts and inspect estimates
4. For testing additional hypotheses, set up new contrasts and refit model

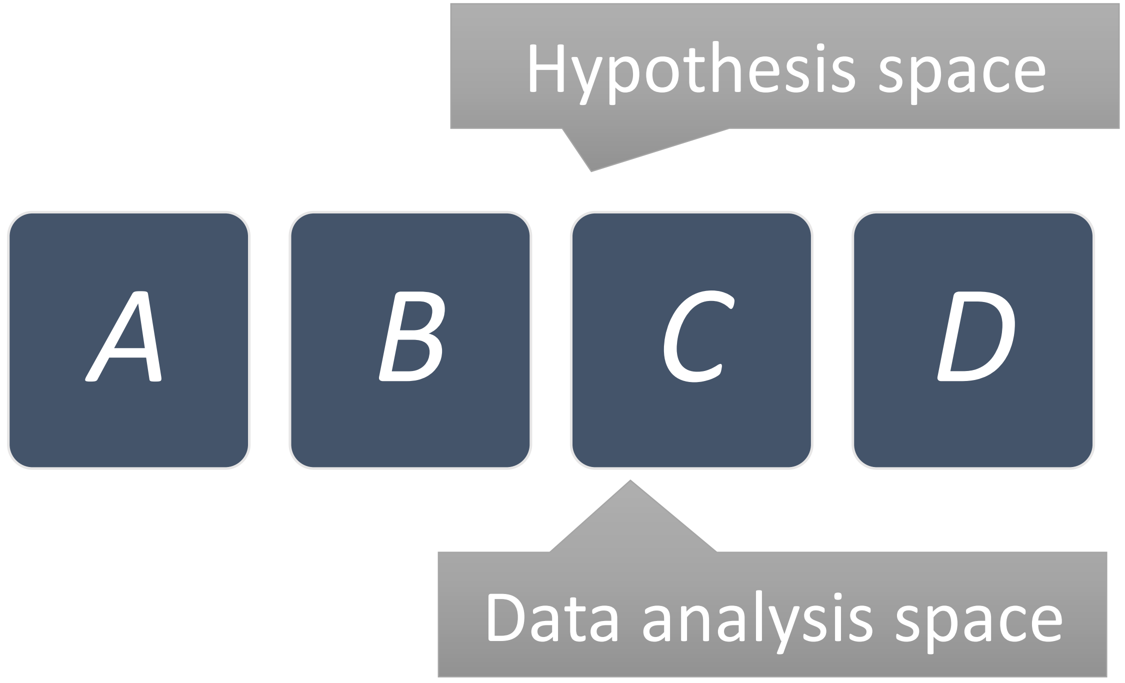
May require several model fits, which can be expensive



The Model-First Approach

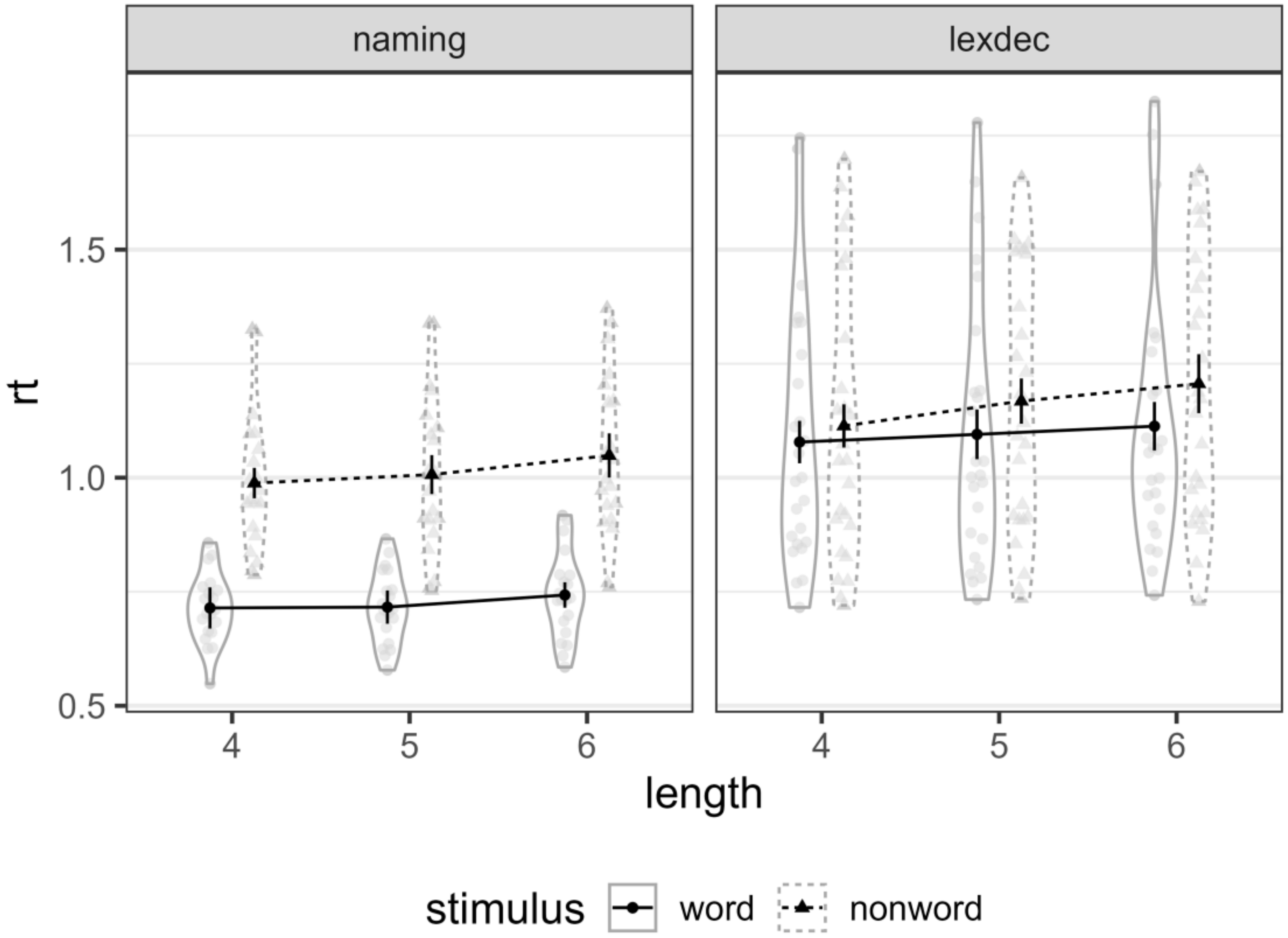
1. Set up statistical model using sum-to-zero contrasts
2. Check ANOVA table with omnibus tests
   - Don’t look at fixed-effect coefficients!
3. Inspect estimated means for model terms (main effects and interactions) of interest
4. Follow-up analysis: specify and tests contrasts on means using `emmeans`
   - consider use of multiple comparison procedure (e.g. `"holm"`)

Only requires single fit of model



Leverage *abstraction* through `emmeans`: hide contrast details behind software interface

Example Data



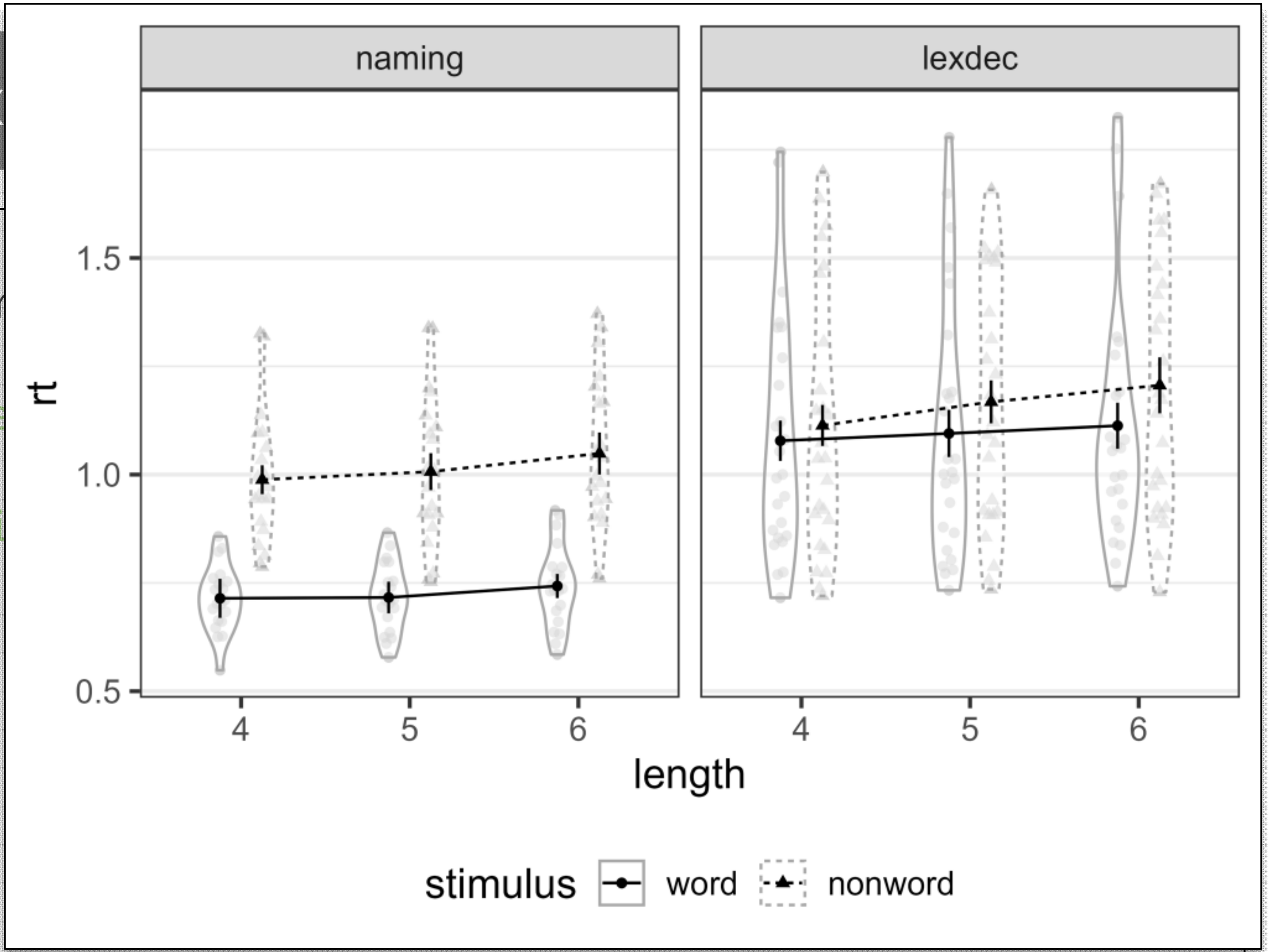
- Freeman et al. (2010, *JML*):
  - Cognitive task: 1 letter string per trial, 300 trials per participant
  - *IV_1* (between-subjects) task: Naming task vs. lexical decision task
  - *IV_2* stimulus (within-subjects): word *vs.* nonwords
  - *IV_3* length (within-subjects): 4, 5, or 6 letters length
  - *DV*: response time
  - Crossed random-effect design
    - 45 participants (20 in naming task, 25 in lexical decision task)
    - 600 items (300 words, 300 nonwords)

Steps 1 & 2: Fit model using afex::mixed() and inspect ANOVA table

```
library("afex")
m3 <- mixed(rt ~ task*stimulus*length + (stimulus||id) + (task||item), fhch, expand_re=TRUE)
m3
# Mixed Model Anova Table (Type 3 tests, S-method)
#
# Model: rt ~ task * stimulus * length + (stimulus || id) + (task || item)
# Data: fhch
#                 Effect        df         F p.value
# 1                 task  1, 44.23 15.17 ***   <.001
# 2             stimulus  1, 50.36 80.37 ***   <.001
# 3               length 2, 590.04 11.88 ***   <.001
# 4        task:stimulus  1, 58.29 29.01 ***   <.001
# 5          task:length 2, 578.59      0.52    .596
# 6      stimulus:length 2, 590.07      2.07    .127
# 7 task:stimulus:length 2, 578.60      0.14    .871
# ---
# Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘+’ 0.1 ‘ ’ 1
```

Steps 1 & 2: Fit model using afex

```
library("afex")
m3 <- mixed(rt ~ task*stimulus*length + (stim
m3
# Mixed Model Anova Table (Type 3 tests, S-me
#
# Model: rt ~ task * stimulus * length + (sti
# Data: fhch
#                  Effect        df     F
# 1                  task  1, 44.23 15.17 ***
# 2              stimulus  1, 50.36 80.37 ***
# 3                length 2, 590.04 11.88 ***
# 4         task:stimulus  1, 58.29 29.01 ***
# 5           task:length 2, 578.59      0.52
# 6       stimulus:length 2, 590.07      2.07      .127
# 7 task:stimulu
# ---
# Signif. codes:
```

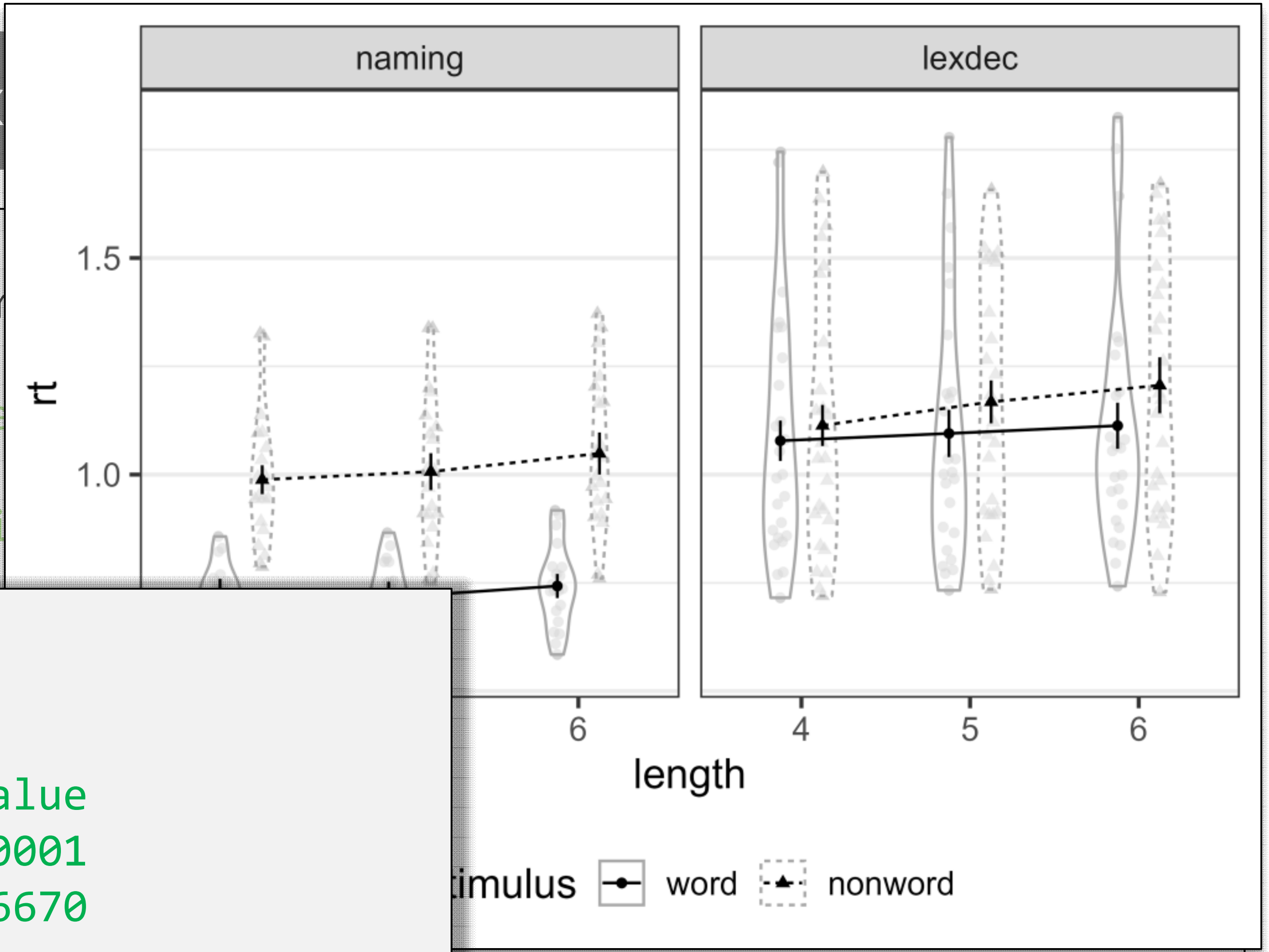


```
afex_plot(m3_nc, "length", "stimulus", "task", id = "id", error = "within",
          data_geom = list(geom_quasirandom, geom_violin),
          data_arg = list(list(width = 0.05, dodge.width = 0.5),
                          list(width = 0.5)))
```

Step 3: Inspect estimated means

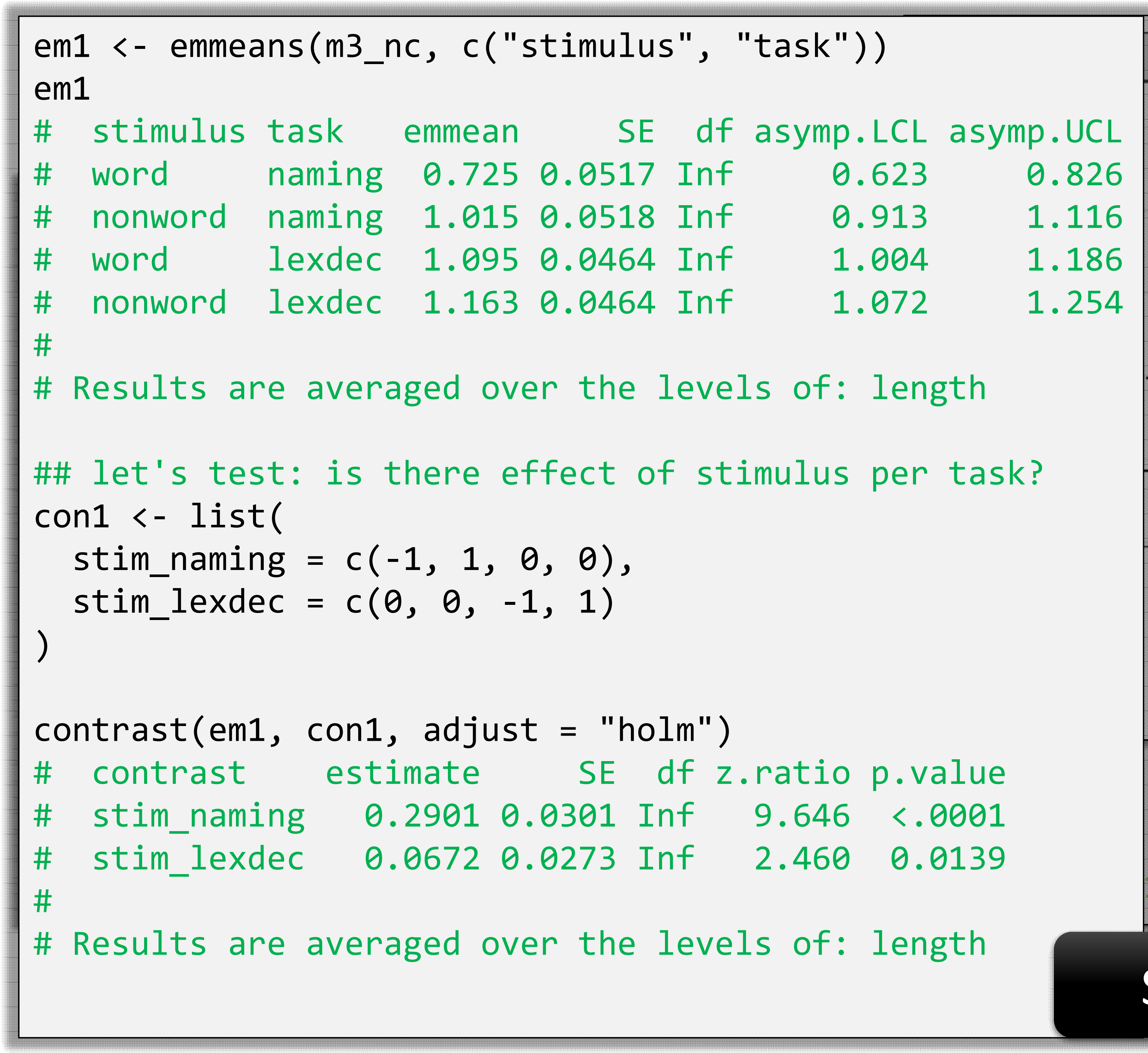
Steps 1 & 2: Fit model using afex

```
library("afex")
m3 <- mixed(rt ~ task*stimulus*length + (stim
m3
# Mixed Model Anova Table (Type 3 tests, S-me
#
# Model: rt ~ task * stimulus * length + (sti
# Data: fhch
```

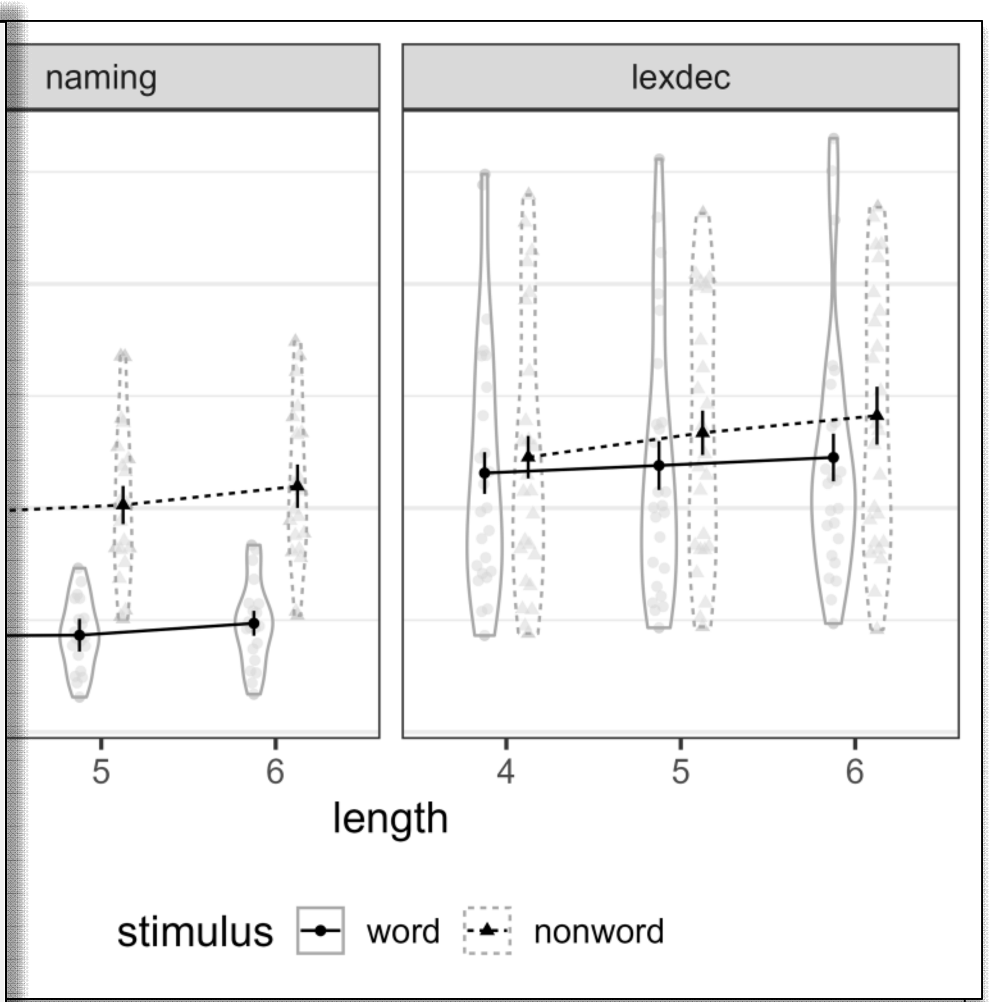


```
library("emmeans")
emmeans(m3_nc, "length") %>%
  contrast("poly")
#  contrast  estimate     SE  df z.ratio p.value
#  linear     0.05405 0.0111 Inf   4.856  <.0001
#  quadratic  0.00827 0.0192 Inf   0.430  0.6670
#
# Results are averaged over the levels of: task, stimulus
# Degrees-of-freedom method: asymptotic
```

Step 4(a): Follow-up analysis



```
em1 <- emmeans(m3_nc, c("stimulus", "task"))
em1
#  stimulus task    emmean     SE  df asymp.LCL asymp.UCL
#  word     naming   0.725 0.0517 Inf     0.623     0.826
#  nonword  naming   1.015 0.0518 Inf     0.913     1.116
#  word     lexdec   1.095 0.0464 Inf     1.004     1.186
#  nonword  lexdec   1.163 0.0464 Inf     1.072     1.254
#
# Results are averaged over the levels of: length

## let's test: is there effect of stimulus per task?
con1 <- list(
  stim_naming = c(-1, 1, 0, 0),
  stim_lexdec = c(0, 0, -1, 1)
)

contrast(em1, con1, adjust = "holm")
#  contrast    estimate     SE  df z.ratio p.value
#  stim_naming   0.2901 0.0301 Inf   9.646  <.0001
#  stim_lexdec   0.0672 0.0273 Inf   2.460  0.0139
#
# Results are averaged over the levels of: length
```



1 ‘ ’ 1

Step 4(b): Follow-up analysis

FAQ on Model-First Approach

- What is ultimately benefit of model-first approach?
  - Less cumbersome and mentally taxing, all tests are performed on design-cell space. In case you want to test more than one set of contrasts, avoids refitting model.
- What if I have a specific hypothesis or am interested in the random-effect estimates?
  - If you have very specific hypothesis and no real uncertainty of how the means will order or hypotheses regarding the random-effect estimates, the contrast-first approach is probably better, even if more cumbersome.
- Can you only test contrasts for model term if corresponding omnibus test is significant?
  - No. Whether you want to look at omnibus tests depends on hypothesis. If strong hypothesis exists, omnibus test not necessary (and can hide effects). Consult omnibus test if results can surprise you.
- Do you always have to correct for multiple testing when calculating contrasts?
  - It depends. If you only conduct $k - 1$ pre-planned contrasts (for factor with k levels), multiple comparison correction probably not necessary. For not planned contrasts or more than $k - 1$ contrasts, Holm-correct probably a good idea.
- Why should I not look at model coefficients for factors with more than 2 levels?
  - For a factor with k levels there are only $k - 1$ coefficients so there exist no 1 to 1 mapping. Especially for popular canned contrasts (especially `contr.sum` or `contr.equalprior`), coefficients cannot be interpreted.

Henrik's List of Things to Remember About Contrasts

- Appropriate contrasts for models with interactions need to sum to zero (in balanced designs): `contr.sum()`, `contr.helmert()`, `contr.sdif()`
  - Intercept needs to correspond to (unweighted) grand mean
  - For treatment contrasts and other contrasts that do not sum to zero, lower order effects correspond to simple effects and not main effects
- For Bayesian models, contrasts should have same marginal effect on all factor levels: `bayestestR::contr.equalprior()`
- For factors with more than 2 levels, avoid looking at model estimates unless contrasts were specifically chosen for this purpose

A General Approach for Testing Omnibus Tests

Wald Tests + More `emmeans` Magic

Estimation and Testing can be Easy in Frequentist Approach

- `lme4` and `MixedModels.jl`
  - LMMs (normal distribution) and GLMMs (with response distribution in the “exponential family”: binomial, Poisson, Gamma, etc.)
  - Allows estimating only conditional mean of response distribution
- `glmmTMB` (or `gamlss`)
  - Supports wider set of selected response distributions, e.g. beta, Student *t*, beta-binomial, negative binomial, Tweedie
  - Allows distributional models: can model both location and scale (with fixed-effects) or other distributional parameters (supports e.g. zero-one-inflated models)
  - Supports flexible correlation structures (compound-symmetric, autoregressive, etc.)
- Statistical testing “easy”
  - Wald tests of simple or compound hypotheses require only fitted model
  - For LMMs, some methods even allow estimating denominator *df* for small samples

Problems for Frequentist Approach

1. Convergence Issues:
   - Every additional parameter increases dimensionality of search space by one
   - If data is sparse for given model, difficult to estimate variance/covariance parameters ("singular fit"): Eager & Roy (2017) - Mixed Effects Models are Sometimes Terrible
   - Dale Barr: "Reducing random-effects structure of model introduces unknown risk of anti-conservativity, and should be done with caution"
2. Extension to other response distributions **very** difficult for models involving random effects
   - Evaluating likelihood equation requires integrating likelihood of data with respect to random effects:

     $$Pr(y \mid \beta, \theta, \sigma^2) \ = \int f(y \mid b, \beta, \theta, \sigma^2) f(b \mid \theta, \sigma^2) \, db$$
   - Simple only for normal response distribution (random effects can be integrated out)
   - Generally requires numerical methods (Gauss-Hermite method) or approximations (Laplace Approximation of integral) and model-specific approach
   - In psych/cognitive science, often interested in cognitive models with complicated likelihoods: Diffusion model, multinomial processing tree (MPT) models

Bayesian Alternatives Make Estimation Easy

- Can use MCMC integration for integrating likelihood of data with respect to random effects (i.e., no further numerical methods necessary)
- `brms`
  - Supports wide range of response distributions including 4-parameter diffusion model (Wiener model)
  - Allows extension to new response distribution via `custom_family()`
  - Fully supports estimation of distributional models with arbitrarily complex formula for each parameter of response distribution
  - Supports various advanced features: multivariate responses, monotonic effects, meta-analyses, certain mixture models, etc.
  - Estimation much more robust due to regularisation provided by priors and MCMC sampling (e.g., Bates et al., 2015, arXiv)

Bayesian Testing Generally Not Trivial

- Simplest method for testing (inspection of whether 95%-CI includes 0) only possible for simple hypothesis
- Principled testing in Bayesian framework of simple and compound hypotheses is *Bayes factor*
  - Bayes factor can be highly sensitive to parameter priors
  - For compound hypotheses computationally expensive:
    - Requires estimating two models, full and restricted model, for each test
    - `bridgesampling` package requires at least an order of magnitude more samples for testing than for estimation; see also Schad et al., 2022)
- Misleading propaganda on philosophical issues surrounding Bayes factors
  - Bayes factor only provides relative comparison of two parameterised models (which allows providing evidence for null model/hypothesis)
  - Does not control false positives (type I errors): does not provide adequate control of fooling oneself
  - Type I error control of *p*-values provides some connection to external world outside of model

Easiness and Flexibility for Mixed-Effects Models

	Frequentist	Bayesian
Estimation	◐ Fast, brittle, & difficult to extend	✓ Slower, robust, & easy to extend
Testing	✓ Fast, specification simple, control errors, prone to misunderstanding	⚠ Very slow, cumbersome model setup, prior dependence

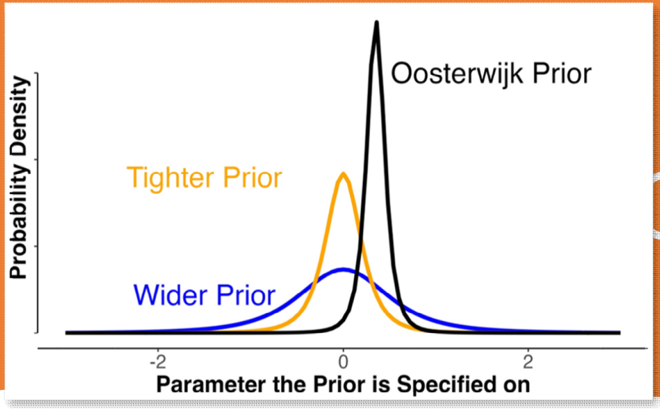
Combining Bayesian Estimation with Frequentist Testing?

emmeans

```
library("bayestestR") # for contr.equalprior_deviations
library("brms")       # for model estimation
library("emmeans")    # for calculating p-values using joint_tests()

options(contrasts = c('contr.equalprior_deviations', 'contr.poly’)) # set contrasts

mbayes <- brm(rt ~ task*stimulus*length + (stimulus*length|id) + (task|item), fhch)

joint_tests(mbayes)
 # model term            df1 df2 F.ratio  Chisq p.value
 # task                    1 Inf  13.806 13.806  0.0002
 # stimulus                1 Inf  73.654 73.654  <.0001
 # length                  2 Inf  10.220 20.440  <.0001
 # task:stimulus           1 Inf  26.615 26.615  <.0001
 # task:length             2 Inf   0.437  0.874  0.6463
 # stimulus:length         2 Inf   1.908  3.816  0.1484
 # task:stimulus:length    2 Inf   0.128  0.256  0.8796
```

Can we use Bayesian-Frequentist *p*-Values?

A Simulation Study



on Setup

Repeat 1000 times

01

Simulate Data

- Null hypothesis is true
- 2 and 3 groups
- Sample size: 20 to 100
- 3 models:
  - ANOVA ($\sigma_\varepsilon^2 = 1$)
  - Logistic β(a = 2, b = 2)
  - Logistic GLMM (σ^2 = .5 or 1)

02

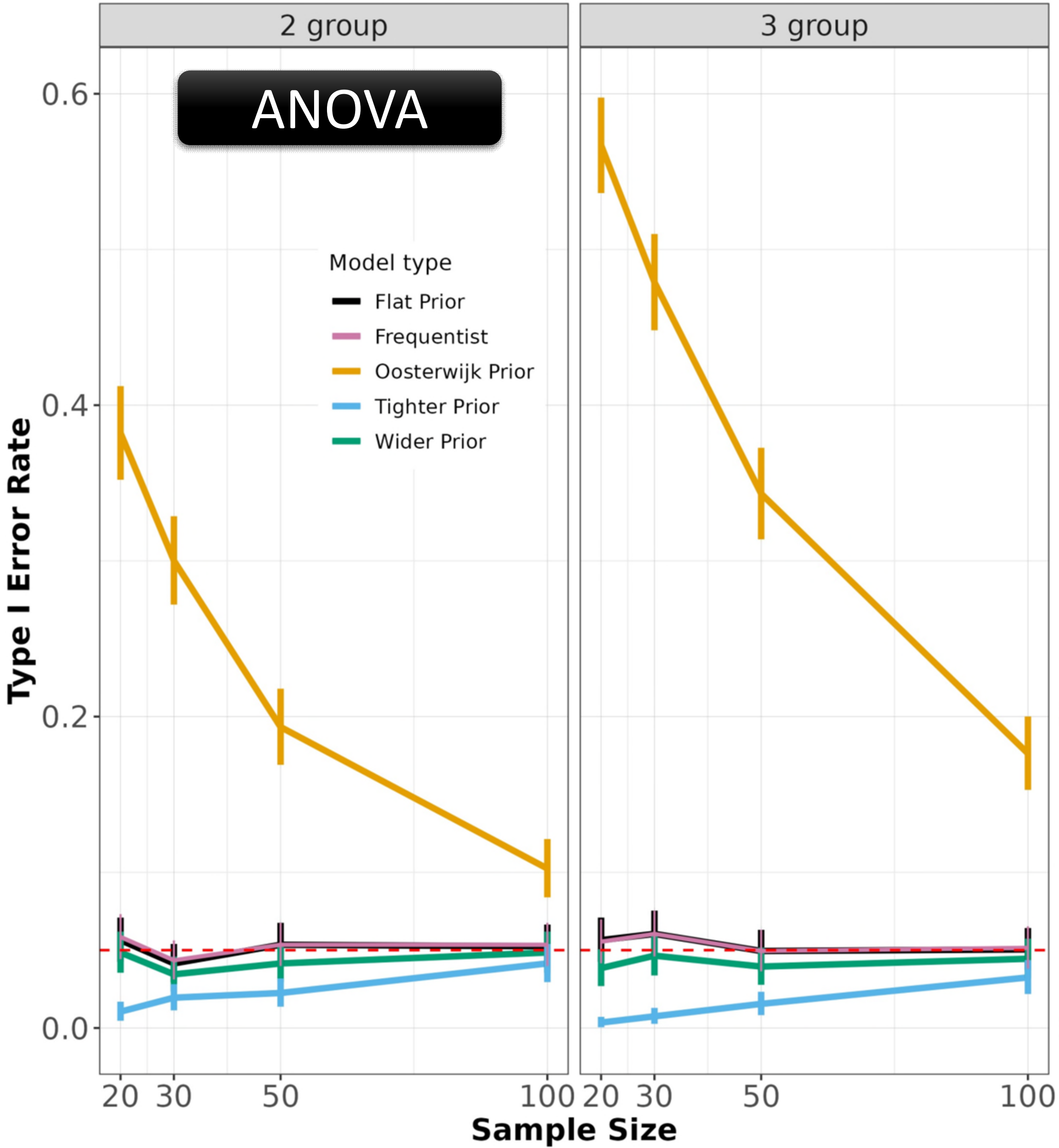
Estimate Model

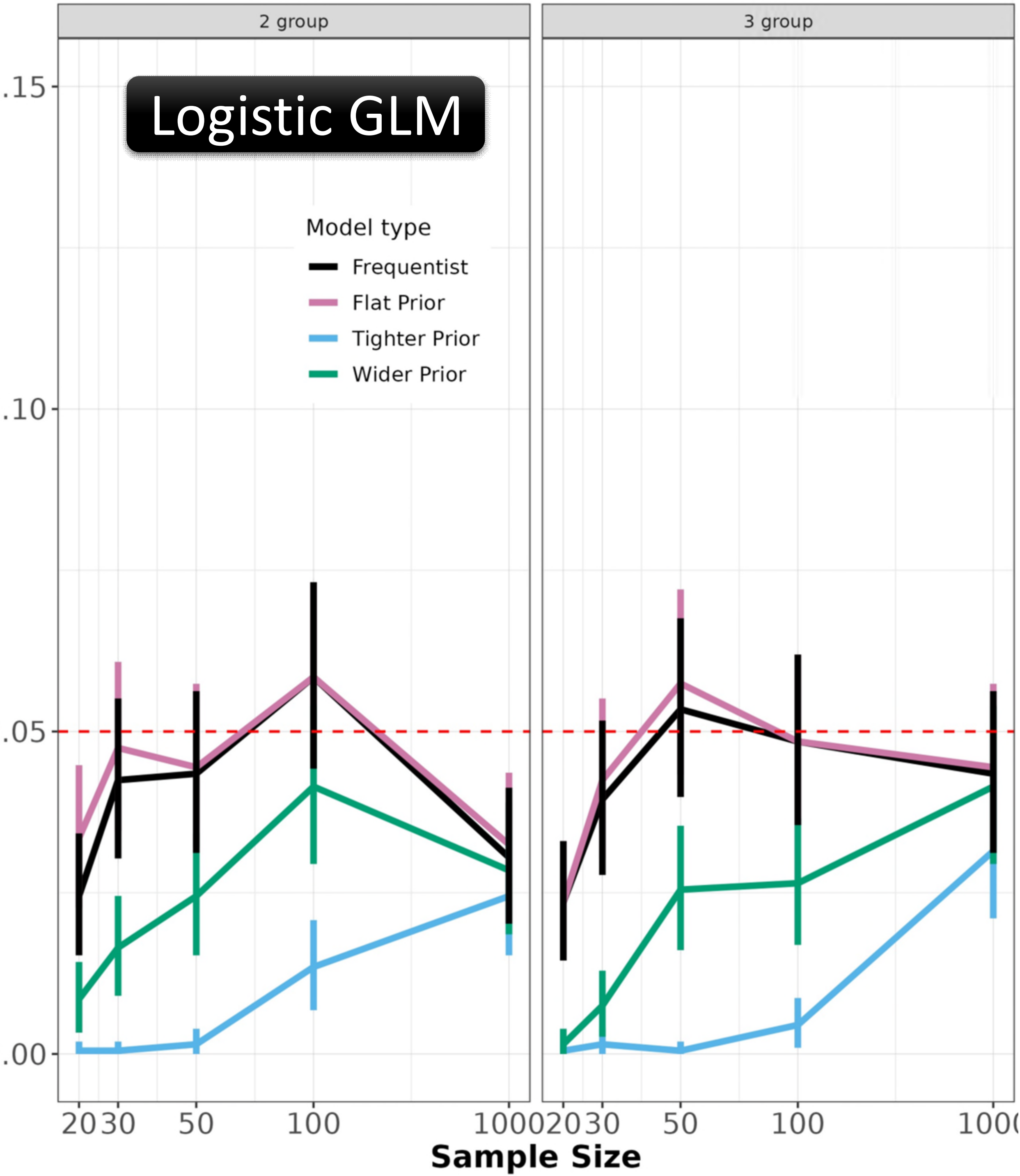
- Frequentist model
- Bayesian models:
  - Improper flat prior
  - Wide prior: *t*(3, 0, 0.5)
  - Tight prior: *t*(3, 0, 0.2)
  - [ANOVA only] Non-centred Oosterwijk prior: *t*(3, 0.35, 0.102)

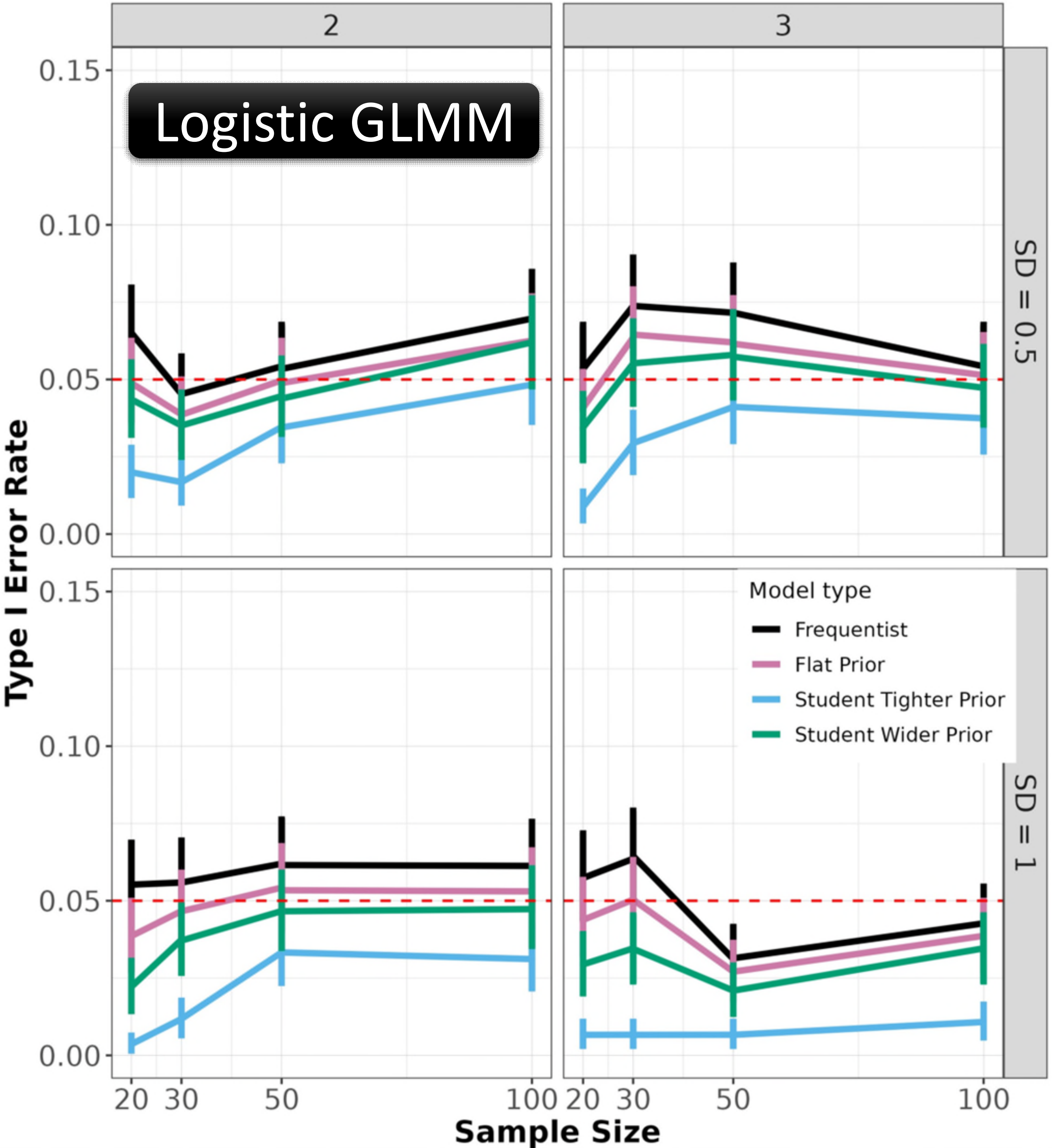
03

Calculate *p*-value

- Using `emmeans::joint_tests()`
- Record *p*-value of main effect of group


ANOVA
2 group
3 group
Type I Error Rate
0.6
0.4
0.2
0.0
Model type
Flat Prior
Frequentist
Oosterwijk Prior
Tighter Prior
Wider Prior
20
30
50
100
Sample Size


Logistic GLM
2 group
3 group
.15
.10
.05
.00
Model type
Frequentist
Flat Prior
Tighter Prior
Wider Prior
20
30
50
100
1000
Sample Size



Logistic GLMM

Simulation Results Summary

- Type I error rates of flat prior very similar to frequentist model
- Type I error rates of zero-centred priors conservative
  - wider (i.e., weakly informative) priors only slightly conservative
  - tighter priors noticeable conservative, especially for small *N*s
- Non-centred priors have highly anti-conservative Type I error rates and are unsuitable for calculation *p*-values
- We should probably run more than 1000 replicates per simulation!
- `brms` default – improper flat priors on fixed-effects – seems to provide performance like frequentist approach so is probably OK!

Henrik’s Simple Statistics

Learn emmeans

p-values are OK

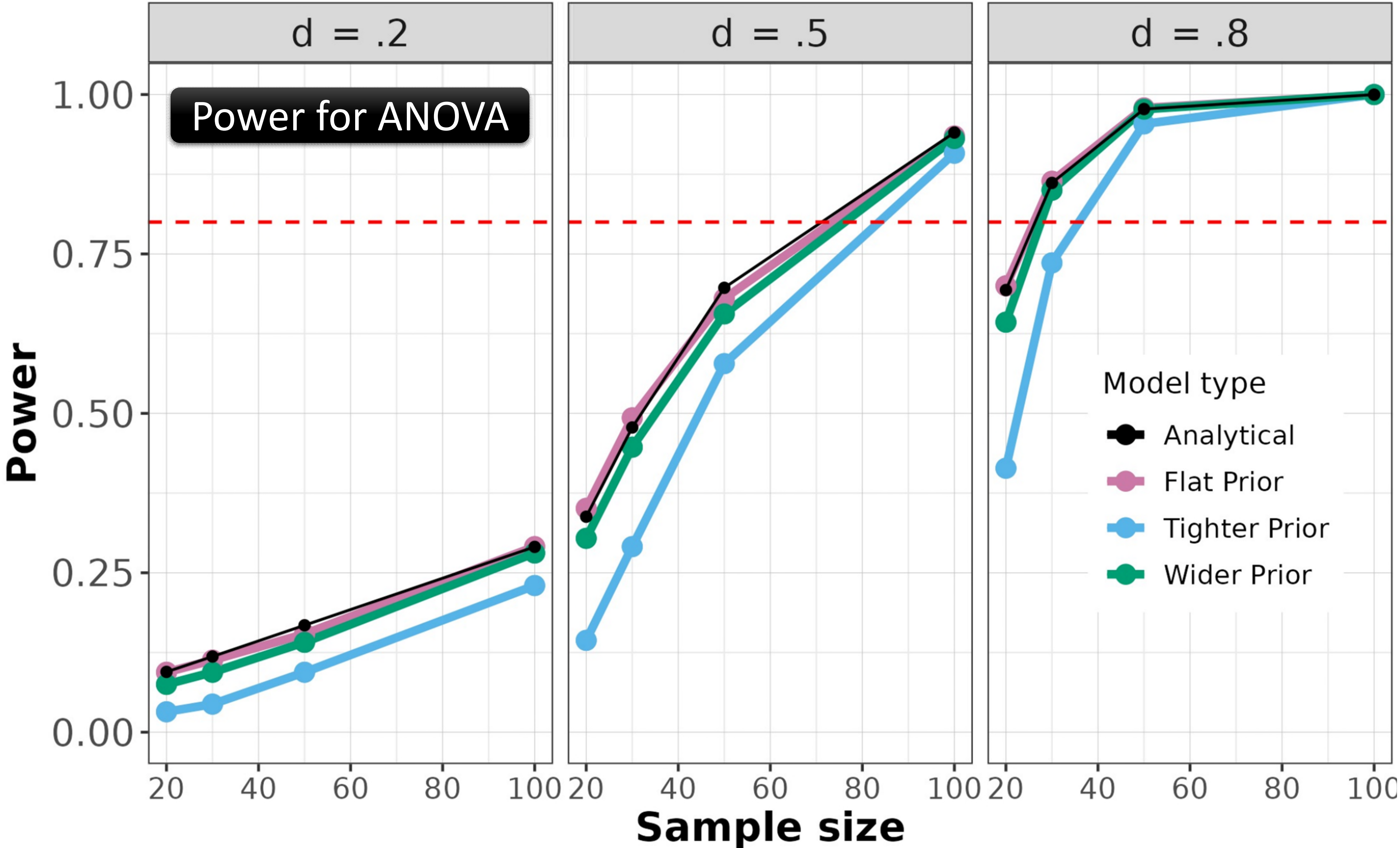
Model-First + Follow-Up Analysis Approach

1. Set up statistical model using sum-to-zero contrasts
2. Check ANOVA table with omnibus tests
   - Don’t look at fixed-effect coefficients!
3. Inspect estimated means for model terms (main effects and interactions) of interest
   - Ideally graphically (e.g., `afex_plot()`)
4. Follow-up analysis: specify and tests contrasts on means using `emmeans`
   - consider use of multiple comparison procedure (e.g. `"holm"`)

Bayesian-frequentist *p*-values

- Nominal type I errors with flat/wide priors

```
library("bayestestR")
library("brms")
library("emmeans")

options(contrasts = c('contr.equalprior',
'contr.poly’))

bayesian_model <- brm(formula, data)
joint_tests(bayesian_model)

afex::afex_plot(bayesian_model, ...)
emmeans(bayesian_model, ...)
```



Power for ANOVA

My Take on Specifying Random Effects

Specifying Random-Effects Structure

- Omitting random-effect parameters for model terms that vary within levels of a random-effect grouping factor and for which random variability exists leads to non-iid residuals (i.e., σ_ϵ) and potentially anti-conservative results (e.g., Barr, Levy, Scheepers, & Tily, 2013, *JML*).
- Safeguard is **maximal model justified by the design** (Barr et al., 2013).
  - Which factors/terms vary within levels of (i.e. are crossed with) each random-effect grouping factor?
  - Are there replicates within factor levels (or parameters/coefficients) for levels of random-effects grouping factor?
- If maximal model is overparameterized and/or contains degenerate estimates or **singular fits**, power of maximal model can be reduced and a reduced model should be used (Bates et al., 2015; Matuschek et al., 2017).
  - Start by removing correlation among random-effect parameters
  - Remove random-effect parameters with variance of 0 and/or for highest-order effects with lowest variance
  - Compare *p*-values/fixed-effect estimates across models (*p*-values from degenerate/minimal models are potentially not reliable)
  - Warning: Reducing model introduces unknown risk of anti-conservativity and should be done with caution!