LMER models: comparison and selection

Vera Demberg

Saarland University

18.1.2022

Fixed vs. random variables

Fixed:

- assumed to be "without error"
- assumed to be the same in one study compared to another

Random:

- drawn from a larger population of values
- would differ from study to study (but the distribution of random effects would be representative of the distribution in the sample)

Table of Contents

Table of Contents

Model comparison and model selection

There are many different possible models: Which fixed effects to include? Which random effects to include?

Today, we'll take a bit more time to talk about how to decide between models.
For this, we will take another look at model comparison.

Relevant sections in the Navarro book are sections 15.10 and 16.5.

Model comparison

We are often faced with the question whether one model is significantly better than another one. In order to have a model that generalizes well, we should choose the smallest model that explains the data best.

- The **degrees of freedom** tell us how many parameters we are estimating compared to how many data points are free to vary.
- The **Sum of Squares** tell us how much of the variability in the data we can vs. can't explain.

Theoretical considerations

Don't throw all kinds of possible model predictors and their interactions into the model, but think in advance:

- which ones are of theoretical interest?
- which ones might sensibly affect model fit?
- then, we do model comparisons and model selection among the sensible models

Statistics serves the scientific process, not the other way around.

Calculating improvements in model fit

We only want to add those predictors, that significantly improve model fit.

Calculating improvements in model fit

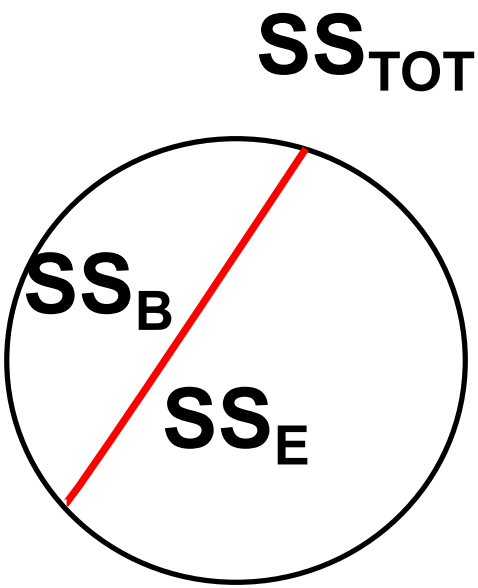
We only want to add those predictors, that significantly improve model fit.

Therefore, we need to quantify how much a predictor improves model fit, and then calculate significance.

How to achieve that?

Reminder: sum of squares

Remember that we can think of our total variability in the data in terms of the variability that we can explain with the help of a predictors (SS_B), plus the variability of the residuals (SS_E).



Testing for the influence of a predictor

```
m0<-lm(response ~ predA)
m1<-lm(response ~ predA + predB)
m2<-lm(response ~ predA + predB + predC)
m3<-lm(response ~ predA * predB)
```

which one is best?

```
m0<-lm(response ~ predA)
m1<-lm(response ~ predA + predB)
```

We can calculate the Sum of Squares (explained variance SS_B) for m0 and m1, and then calculate the difference to see how much explanatory power the inclusion of predB has gained us.

$$\begin{aligned} \mathrm{SS}_\Delta &= \mathrm{SS}_{M1} - \mathrm{SS}_{M0} \\ &= (\mathrm{SS}_T - \mathrm{SS}_{R1}) - (\mathrm{SS}_T - \mathrm{SS}_{R0}) \\ &= \mathrm{SS}_{R0} - \mathrm{SS}_{R1} \end{aligned}$$

Then we need to divide this by the number of addidional df; we also need to calculate mean squared error for the residuals of the larger model:

$$\mathrm{MS}_\Delta = \frac{\mathrm{SS}_\Delta}{\mathrm{df}_\Delta} \qquad \mathrm{MS}_{R1} = \frac{\mathrm{SS}_{R1}}{\mathrm{df}_{R1}} \qquad F = \frac{\mathrm{MS}_\Delta}{\mathrm{MS}_{R1}}$$

We obtain an **F statistic**, telling us how much of the variance we have explained as a proportion of the overall remaining variance.

Testing for the influence of a predictor

```
m0<-lm(response ~ predA)
m1<-lm(response ~ predA + predB)
m2<-lm(response ~ predA + predB + predC)
m3<-lm(response ~ predA * predB)
```

Model comparison tests allow us to test the effect of adding several parameters to the model at once, for instance we could compare the overall effect of adding predictor B as a main effect and with its interaction if comparing m0 to m3.

```
> anova( model.1, model.3 )
Analysis of Variance Table

Model 1: mood.gain ~ drug
Model 2: mood.gain ~ drug * therapy
  Res.Df   RSS Df Sum of Sq    F Pr(>F)
1     15 1.392
2     12 0.653  3     0.738 4.52  0.024 *
```

$RSS_{m1} = 1.392$

$RSS_{m2} = 0.653$

$SS_{\Delta} = 1.392 - 0.653 = 0.738$

$MS_{\Delta} = \frac{0.738}{3} = 0.246$

$RMS_{m1} = \frac{0.653}{12} = 0.054$

$F = \frac{0.246}{0.054} = 4.52$

`1 - pf(4.52, 3, 12)` $= 0.024$

Model comparison for mixed effects models

When doing model comparison for LMER models, we can still call `anova()`.

However, the output will look a bit different (more informative because the info given there takes into account the random effect structure for estimates of df).

Model comparison for mixed effects models

```
> anova(m1,m2a, m2b)
refitting model(s) with ML (instead of REML)
Data: subset(dat, RELWDINDEX == 0)
Models:
m1: WORD_TIME ~ ITEM_TYPE + (1 + ITEM_TYPE | PARTICIPANT) + (1 +
m1:     ITEM_TYPE | ITEM_ID)
m2a: WORD_TIME ~ ITEM_TYPE + itemOrder + (1 + ITEM_TYPE | PARTICIPANT) +
m2a:     (1 + ITEM_TYPE | ITEM_ID)
m2b: WORD_TIME ~ ITEM_TYPE * itemOrder + (1 + ITEM_TYPE | PARTICIPANT) +
m2b:     (1 + ITEM_TYPE | ITEM_ID)
    Df    AIC    BIC  logLik deviance   Chisq Chi Df Pr(>Chisq)
m1   9 8175.3 8214.5 -4078.6   8157.3
m2a 10 8120.2 8163.7 -4050.1   8100.2 57.1145      1  4.112e-14 ***
m2b 11 8119.5 8167.3 -4048.7   8097.5  2.7169      1    0.09929 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

deviance similar to residual sum of squares. Not to be confused with deviation!!

Model comparison for LMER models

Unfortunately, we cannot easily estimate R^2 for lmer models:

Nakagawa, Shinichi, and Holger Schielzeth. "A general and simple method for obtaining R2 from generalized linear mixed-effects models." Methods in Ecology and Evolution 4.2 (2013): 133-142.

AIC

The Akaike Information Criterion (AIC) is a commonly used measure of comparison.

$$\text{AIC} = \frac{\text{SS}_{res}}{\hat{\sigma}^2} + 2K$$

It takes into account both model fit (SS_{res}) and the number of predictor variables K.
deviance + 2×Df = AIC

$$8100.2 + 2 \times 10 = 8120.2$$

$$8097.5 + 2 \times 11 = 8119.5$$

Lower AICs are better.

Model Comparisons can only be done for nested models!

You **cannot** compare

```
m1<-lm(response ~ predA + predB)
m2<-lm(response ~ predA + predC)
```

and then use `anova(m1,m2)` to find out whether predictor B or C is better.

The significance test for such a comparison is not meaningful!

Model selection

Now that we know how to compare models, we can talk about model selection.

- forward selection
- backward selection

These two approaches don't necessarily come out to the same result.

Recommendation: Backward selection

Always report how you did model selection.

Forward Selection

- You start out with a model including only the intercept,
- and then try to add each predictor, comparing how much it improves model fit (check this with `anova()`).
- The best predictor is then added to the model,
- and you repeat the process for all predictors and their interactions.

Backward selection

- You start out with the complete model (all predictors and all (theoretically motivated) interactions).
- Then you remove first each of the interactions to find out which one contributes least,
- and remove that one, after checking interactions, you remove any predictors that don't improve model it.
- Careful, you can't remove a main effect of a predictor that's part of an interaction!

Keep it maximal

When you do hypothesis testing, it is recommended to use maximal random effects models.

Hypothesis testing vs. exploratory analysis

Q1: What combination of factors characterises my data best?
Method: data-driven model selection; exploratory analysis

Q2: Do factors A and B contribute independently or interactively to the observed response variable?
Method: Test of an interaction hypothesis; confirmatory analysis
important aspects: generalizability and replicability

Keep it maximal

When you do hypothesis testing, it is recommended to use maximal random effects models.

maximal *random* effects

- not maximal fixed effects
- maximal given logical data structure

Convergence problems

Definitely needs further investigation (especially with categorical data), but Barr et al. recommend:

- Check for potential extreme values, coding errors etc.
- Examine the random effects estimates in the non-converged model. Remove the highest-order random effect closest to zero, then refit
- Consider dropping random intercept/slope correlations. `lmer(Y~X+(1|Subject)+(0+X|Subject))`
- Even dropping close-to-zero random intercepts is worth considering. Detrimental to power but not to Type-I error.
- If all of this doesn't work, give up the idea of "crossing" random effects and perform separate by-subjects and by-item analyses, each with appropriate maximal random effect structure.

Table of Contents

Power

So far, we have mostly worried about **Type I error**, i.e. drawing incorrect conclusions in the sense of rejecting the null-hypothesis even though it is true.
As good scientists, making sure that we restrict the risk of Type I errors is an important safeguard, and the primary reason why we use statistics in the first place.

Power

So far, we have mostly worried about **Type I error**, i.e. drawing incorrect conclusions in the sense of rejecting the null-hypothesis even though it is true.
As good scientists, making sure that we restrict the risk of Type I errors is an important safeguard, and the primary reason why we use statistics in the first place.
When talking about **power**, we are focussing on **Type II error**, i.e. the risk of not being able to reliably detect an effect even though it is there.
Type II error is usually referred to as β.

Power

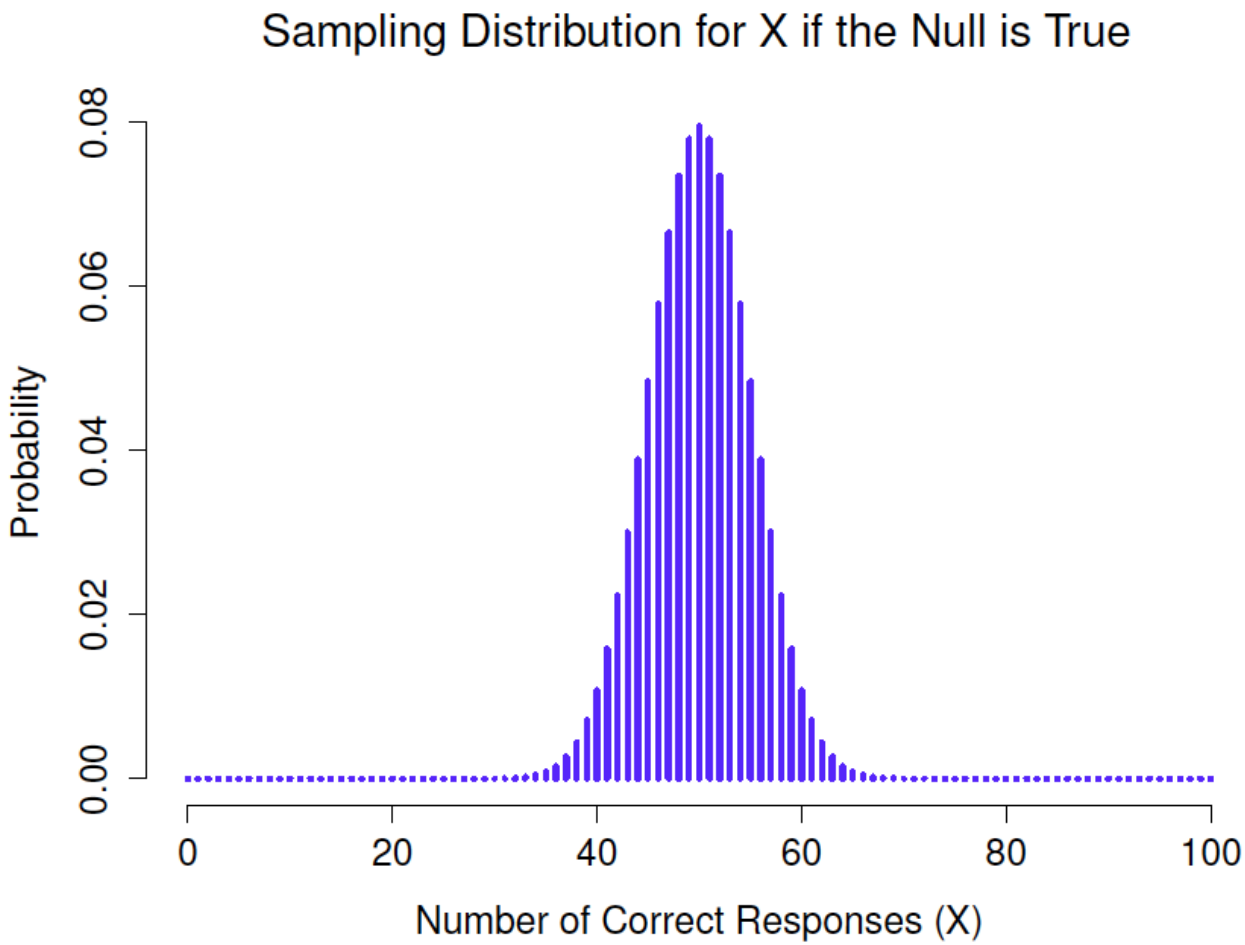
Power = 1 - β
i.e. instead of minimizing Type II error, we usually talk about maximizing power.

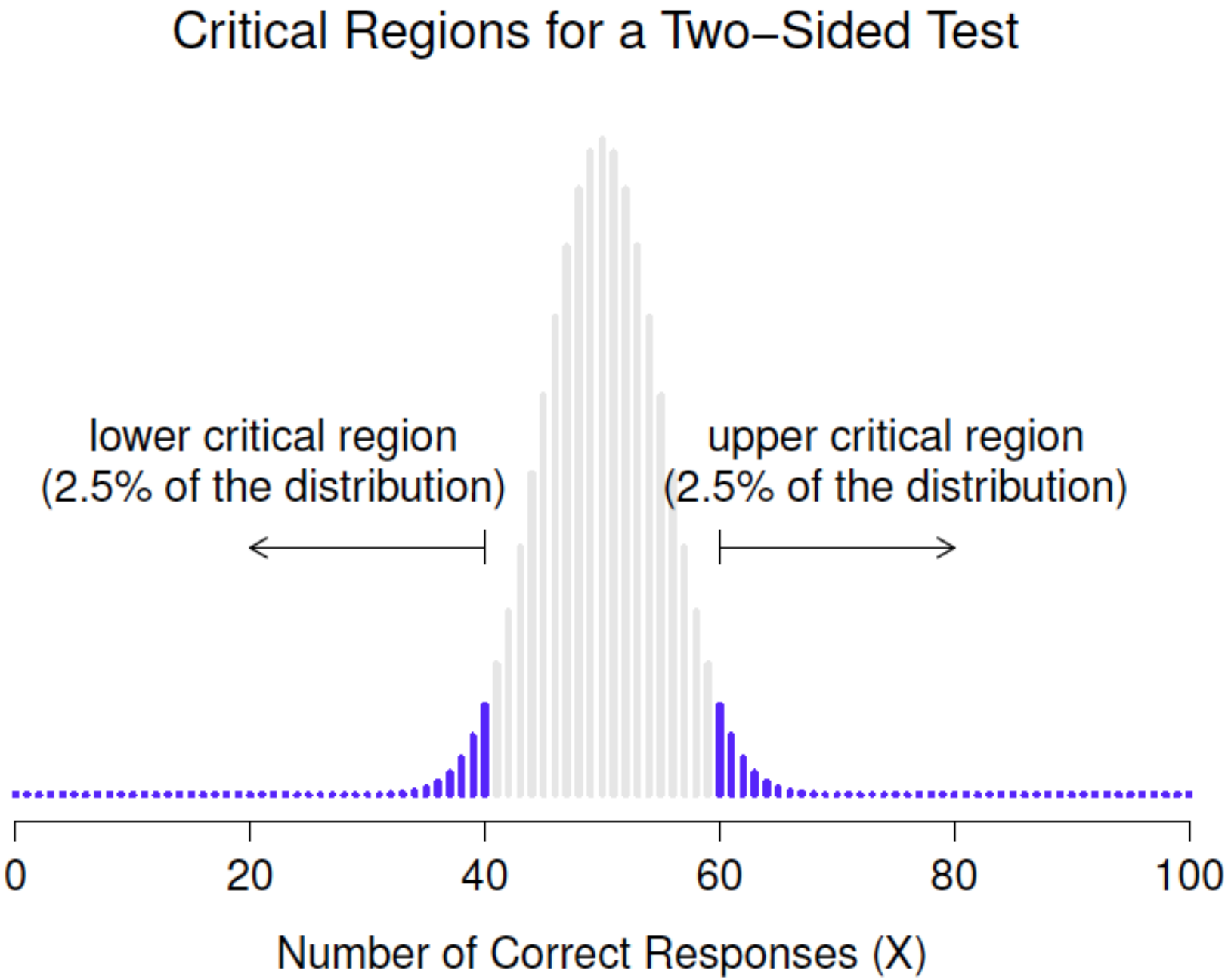
Let's take a look at an example

Dan Navarro's hypothetical "clairvoyance" experiment

Each participant sits down at a table, and is shown a card by an experimenter. The card is black on one side and white on the other. The experimenter takes the card away, and places it on a table in an adjacent room. The card is placed black side up or white side up completely at random, with the randomisation occurring only after the experimenter has left the room with the participant. A second experimenter comes in and asks the participant which side of the card is now facing upwards. It's purely a one-shot experiment. Each person sees only one card, and gives only one answer; and at no stage is the participant actually in contact with someone who knows the right answer.

100 people were asked.
Null-hypothesis: 50 times correct answer.
Observation: 57 times correct answer.



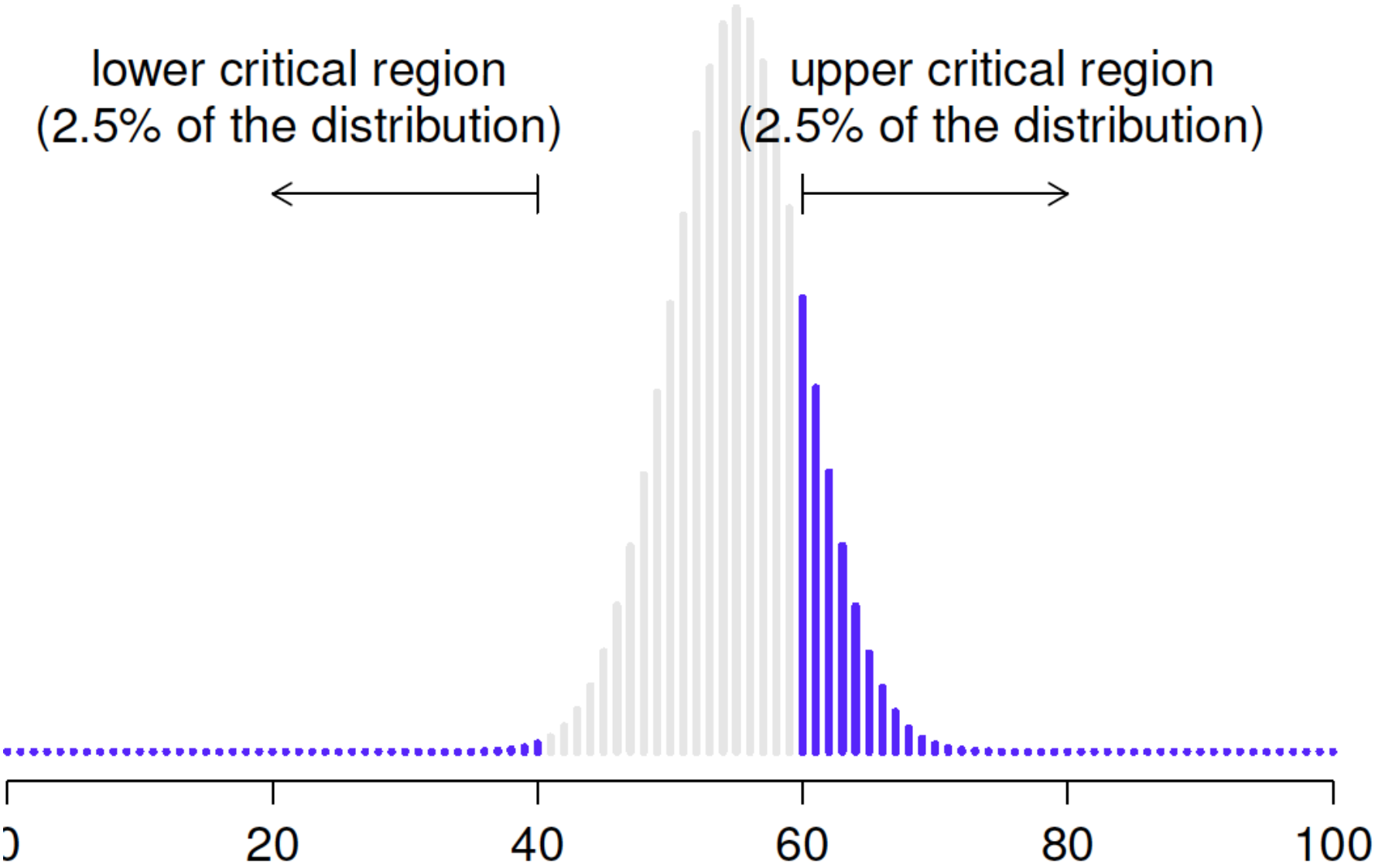

Critical Regions for a Two–Sided Test
lower critical region
(2.5% of the distribution)
upper critical region
(2.5% of the distribution)
0
20
40
60
80
100
Number of Correct Responses (X)

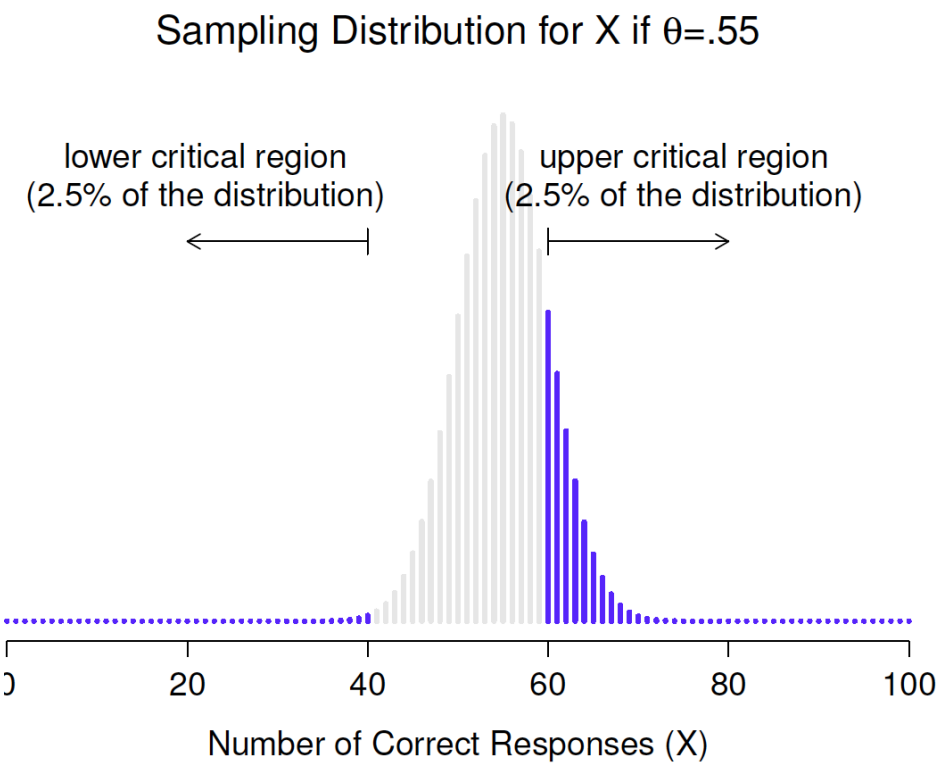
Next we will plot what the rejection region looks like for different situations in our populations.

Example:
I.e. our null hypothesis is that 50% of the time the correct side is guessed.
But let's assume for a moment that clairvoyance is real, and that therefore in the population probability of guessing right is 55%.

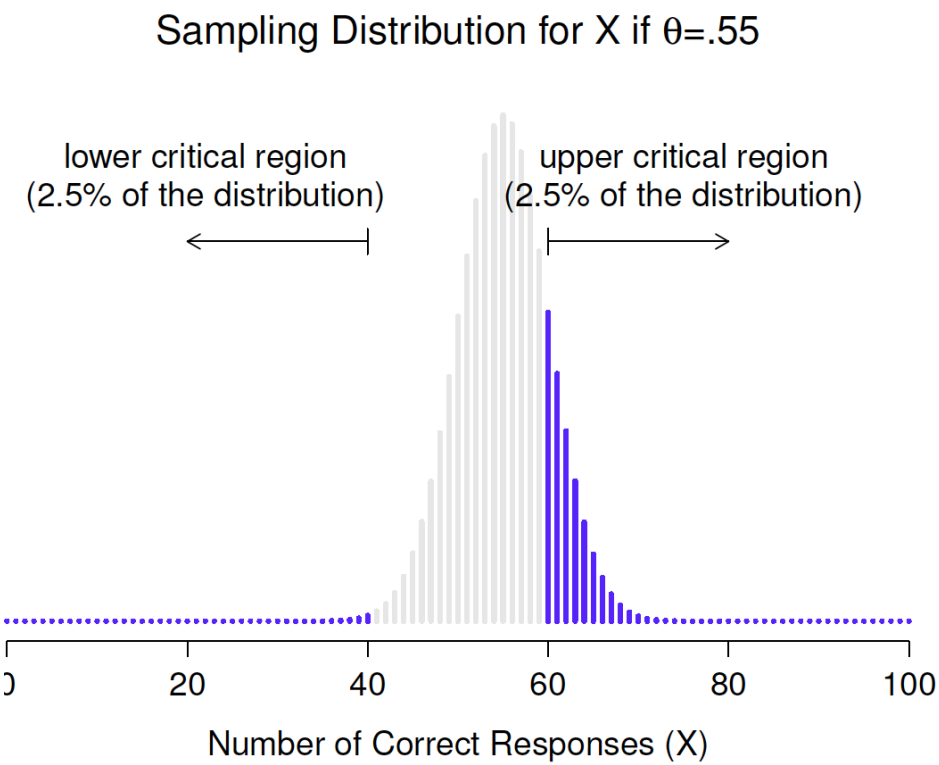
Then we can draw again our rejection regions of the null hypothesis, but for the hypothetical population data.

Sampling Distribution for X if θ=.55



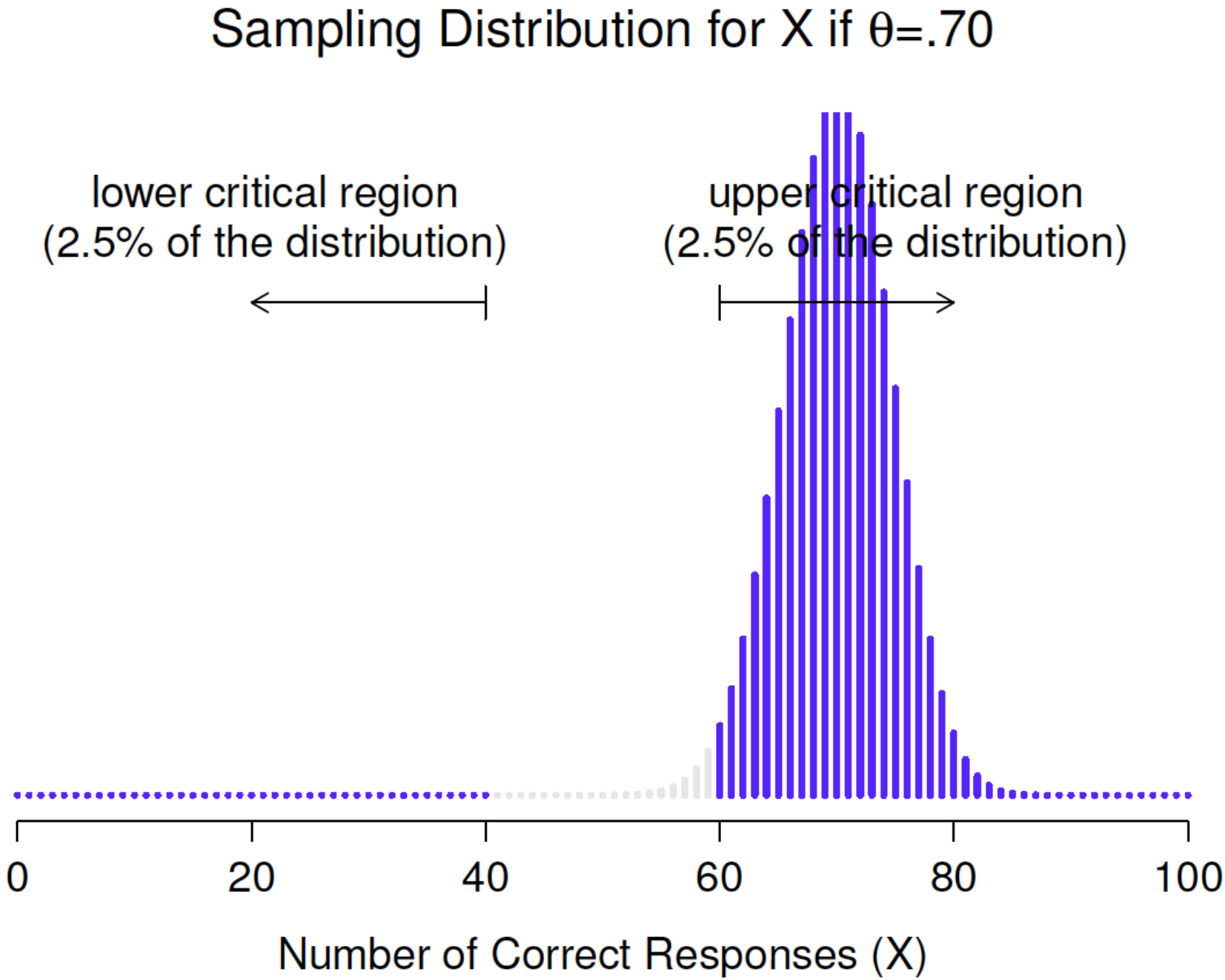


The purple region again shows for which findings we would reject the null hypothesis, together with the probability of making such findings given that the real probability of guessing correctly is 0.55.



The purple region again shows for which findings we would reject the null hypothesis, together with the probability of making such findings given that the real probability of guessing correctly is 0.55.

Of course, it could be that clairvoyance is real, and the population probability of guessing right is even as high as 70%.

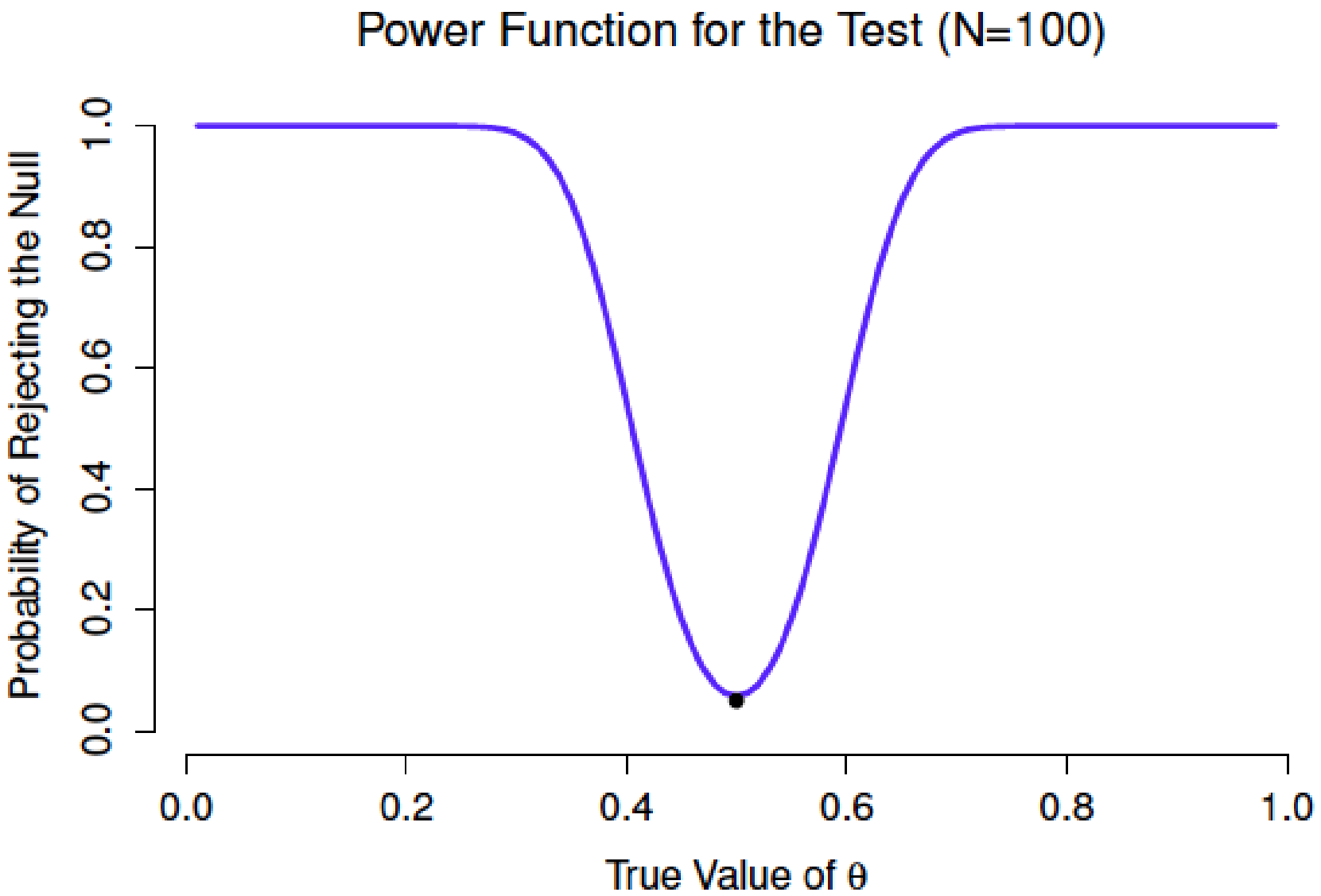

Sampling Distribution for X if θ=.70
lower critical region
(2.5% of the distribution)
upper critical region
(2.5% of the distribution)
0
20
40
60
80
100
Number of Correct Responses (X)

Effect size vs. Type II error / power

We can observe that if the effect size is larger, our Type II error probability decreases, even though the null hypothesis is identical.

We can also plot the probability of rejecting the null hypothesis for different values of θ, i.e. different "true" probabilities that might occur in the population.
$\rightarrow$ the power function

Power function



Increasing the power of your study

As scientists, we like to maximise the power of our experiments.

- we want our experiments to work
- $\rightarrow$ maximise the chance of rejecting the null hypothesis if it is false (and of course we usually want to believe that it is false!)
- increase effect size:
  In practice, what this means is that you want to design your study in such a way that the effect size gets magnified $\rightarrow$ clever experimental design is one way to boost power; because it can alter the effect size.
- increase sample size

Larger sample size increases the power of your study

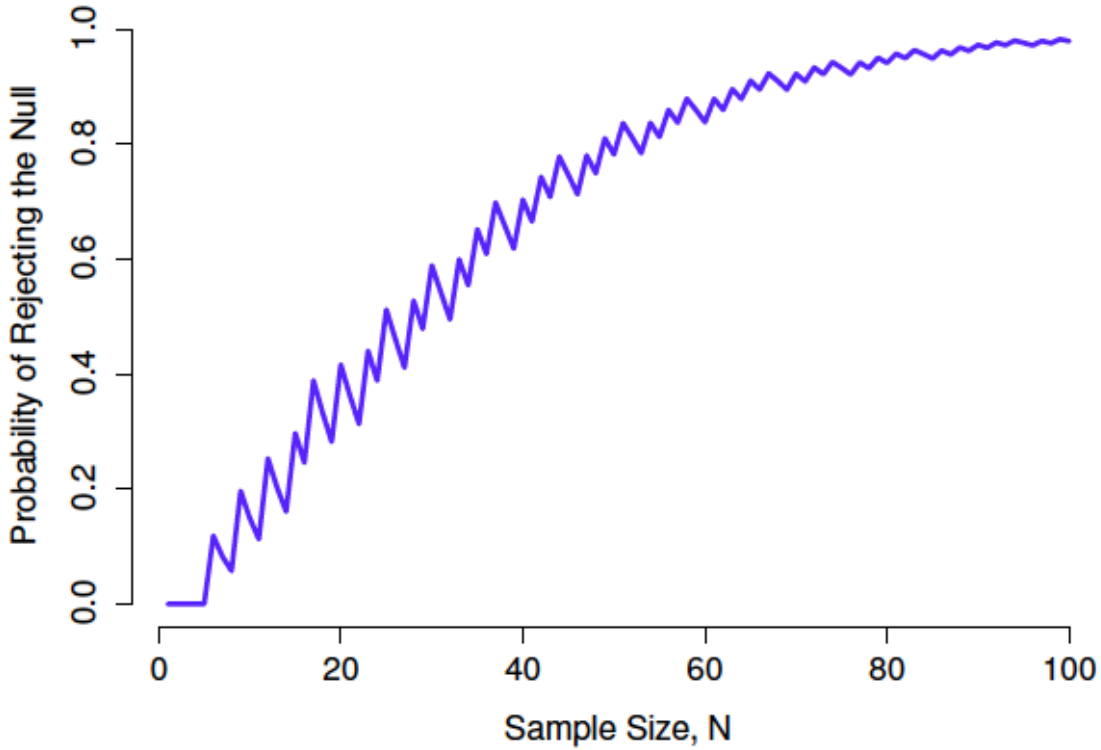


Figure 11.7: The power of our test, plotted as a function of the sample size N. In this case, the true value of θ is 0.7, but the null hypothesis is that $\theta = 0.5$. Overall, larger N means greater power. (The small zig-zags in this function occur because of some odd interactions between θ, α and the fact that the binomial distribution is discrete; it doesn't matter for any serious purpose)

Power analysis

In **power analysis**, we can sometimes estimate before the experiment how large the effect size will be (e.g. based on other similar earlier experiments), in order to calculate how many subjects / items we should be collecting data for.

A power analysis can tell you something about whether you have enough time or money to be able to run the experiment successfully with reasonable probability.

Calculating power

If you have

- target significance level
- power
- effect size

then you can calculate the necessary sample size.

Calculating power

If you have

- target significance level
- power
- effect size

then you can calculate the necessary sample size.

Or, if you have

- target significance level
- power
- sample size

you can calculate the minimal effect size you need in order to have a given probability of actually find an effect at significance level alpha.

In practice, the problem is usually the lack of knowing the effect size.

Summary of power calculation functions in R:
http://www.statmethods.net/stats/power.html

Table of Contents

Data simulation

Research in statistical models sometimes uses data simulation.

- data are generated, such that we know exactly what the intercept, slopes etc. are.
- then we observe whether the model can recover them correctly
- and whether the type I error estimates are correct

Here, we will use an example from Barr et al., 2011.

Performance of different approaches to model selection

- Monte Carlo simulation: Generate and analyze (x100000)
  - ANOVA-based approaches: $F_1, F_1 \times F_2$,min-F'
  - LMEMs with varying random effects structures
    - ★ Random-intercept-only models
    - ★ "Maximal" models
    - ★ Data-driven random effects (Model selection)
  - p values from
    - ★ "t as z" procedure
    - ★ model comparison
    - ★ MCMC simulation (only for RI-only)
- Data from a simple one-factor design
  - continuous DV
  - factor was within-subject, and either between-item (WSBI) or within-item (WSWI)
  - 24 subjects and 12 or 24 items

Method

- Parameter space

Parameter	Description	Value
β_0	grand-average intercept	$\sim U(-3,3)$
β_1	grand-average slope	0 (H_0 true) or .8 (H_1 true)
${\tau_{00}}^2$	by-subject variance of S_{0s}	$\sim U(0,3)$
${\tau_{11}}^2$	by-subject variance of S_{1s}	$\sim U(0,3)$
ρ_S	correlation between (S_{0s}, S_{1s}) pairs	$\sim U(-.8,.8)$
${\omega_{00}}^2$	by-item variance of I_{0i}	$\sim U(0,3)$
${\omega_{11}}^2$	by-item variance of I_{1i}	$\sim U(0,3)$
ρ_I	correlation between (I_{0i}, I_{1i}) pairs	$\sim U(-.8,.8)$
σ^2	residual error	$\sim U(0,3)$
$p_{missing}$	proportion of missing observations	$\sim U(.00,.05)$

- Metrics
  - Type I error rate (false rejection of H_0)
  - Power (accurate rejections of H_0)
  - Terminology:

Conservative	Type I $< \alpha$
Exact	Type I $= \alpha$
Anticonservative	Type I $> \alpha$

Data-driven selection approaches to random effect structure

- Forward selection
  - start with RI-only model, add slopes, terminate when no improvement
- Backward selection
  - start with maximal model, remove slopes as long as they don't worsen fit
- "Best path" model
  - forward/backward
  - test both slopes and follow "biggest mover"
- Varied α level for inclusion of slope

within-subj betw-item design

Type-I error

	$\alpha = .01$		$\alpha = .05$		$\alpha = .10$	
N_{items}	12	24	12	24	12	24
	Type I Error at or near α					
min-F'	.009	.009	.044	.045	.092	.093
LMEM, Maximal, χ^2_{LR}	.017	.013	.070	.058	.129	.113
$F_1 \times F_2$	.014	.019	.063	.077	.120	.137
LMEM, Maximal, t	.029	.017	.086	.065	.143	.120
	Type I Error far exceeding α					
LMEM, RI-only, χ^2_{LR}	.032	.039	.102	.111	.171	.177
LMEM, RI-only, t	.055	.051	.128	.124	.193	.189
LMEM, RI-only, MCMC	.071	.103	.173	.211	.255	.294
F_1	.297	.217	.421	.339	.497	.420

Power

	$\alpha = .01$		$\alpha = .05$		$\alpha = .10$	
N_{items}	12	24	12	24	12	24
	Type I Error at or near α					
min-F'	.079	.154	.210	.328	.311	.444
LMEM, Maximal, χ^2_{LR}	.118	.185	.267	.364	.371	.478
$F_1 \times F_2$	.106	.222	.252	.403	.355	.510
	Type I Error exceeding α					
LMEM, Maximal, t	.162	.214	.300	.382	.394	.490
LMEM, RI-only, χ^2_{LR}	.164	.279	.319	.449	.419	.548
LMEM, RI-only, t	.228	.318	.360	.472	.447	.563
LMEM, RI-only-MCMC	.252	.444	.428	.601	.524	.680
F_1	.541	.571	.671	.706	.732	.767

within-subj
within-item
design

Type-I error

	$\alpha = .01$		$\alpha = .05$		$\alpha = .10$	
N_{items}	12	24	12	24	12	24
	Type I Error at or near α					
min-F'	.004	.005	.027	.031	.061	.068
LMEM, Maximal, χ^2_{LR}	.013	.012	.059	.056	.113	.108
$F_1 \times F_2$	.012	.018	.057	.072	.112	.130
LMEM, Maximal, t	.022	.016	.072	.063	.126	.115
	Type I Error exceeding α					
F_1	.083	.059	.176	.139	.251	.210
LMEM, RI-only, χ^2_{LR}	.317	.377	.440	.498	.514	.567
LMEM, RI-only, t	.320	.379	.441	.499	.515	.568
LMEM, RI-only-MCMC	.260	.360	.390	.500	.440	.600

Power

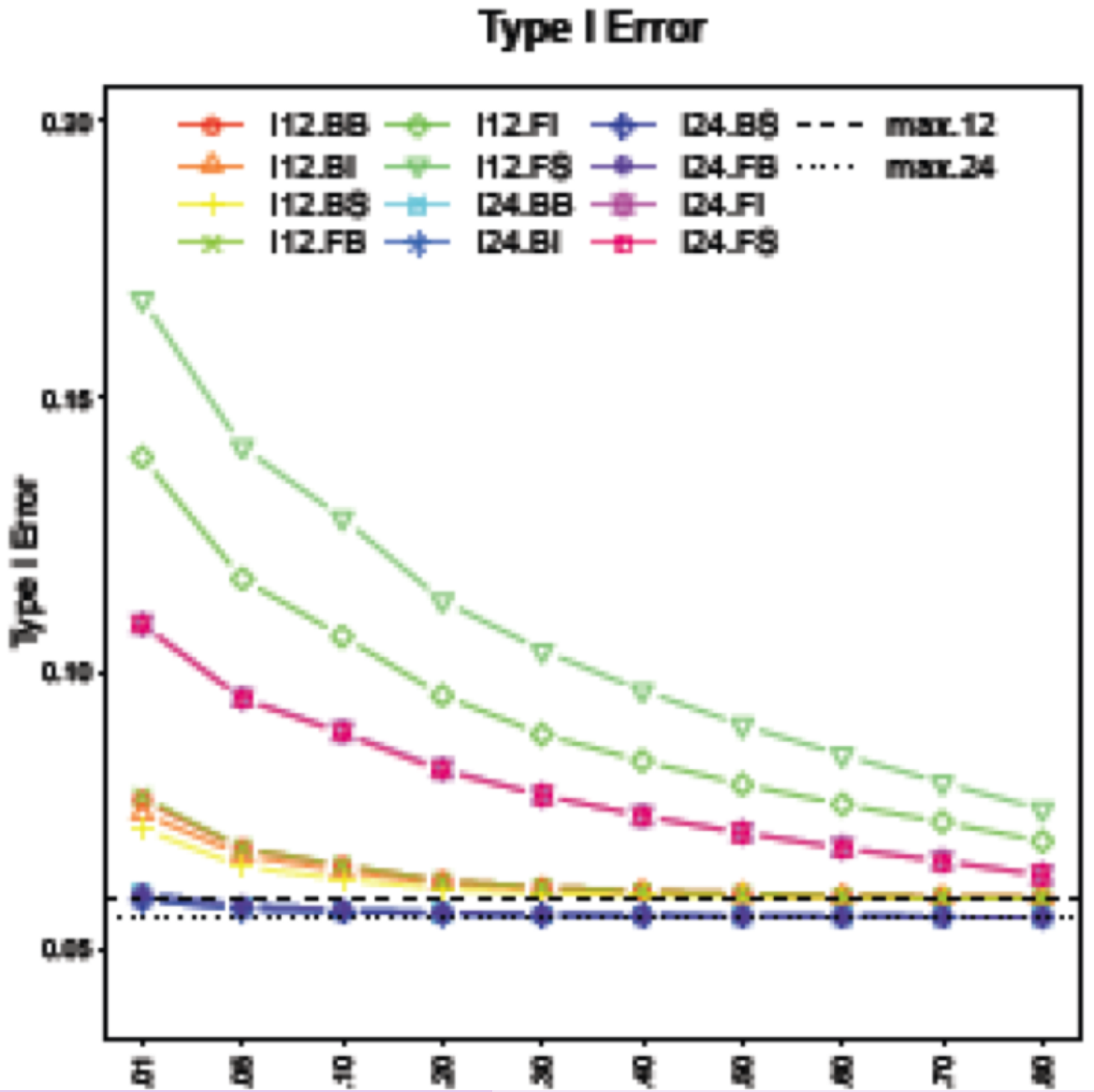
	$\alpha = .01$		$\alpha = .05$		$\alpha = .10$	
N_{items}	12	24	12	24	12	24
	Type I Error at or near α					
min-F'	.129	.266	.327	.512	.463	.643
LMEM, Maximal, χ^2_{LR}	.240	.382	.460	.610	.582	.717
$F_1 \times F_2$	.212	.410	.440	.643	.568	.746
LMEM, Maximal, t	.306	.425	.496	.629	.603	.727
	Type I Error exceeding α					
F_1	.455	.538	.640	.724	.725	.800
LMEM, RI-only, χ^2_{LR}	.787	.904	.853	.935	.883	.949
LMEM, RI-only, t	.789	.904	.854	.935	.883	.949
LMEM, RI-only-MCMC	.860	.898	.880	.918	.920	.939

Data-driven random effects selection

BB: backwards best path
BI: backwards item first
BS: backwards subj first

FB: forwards best path
FI: forwards item first
FS: forwards subj first

X axis gives alpha levels for testing inclusion of random slopes.



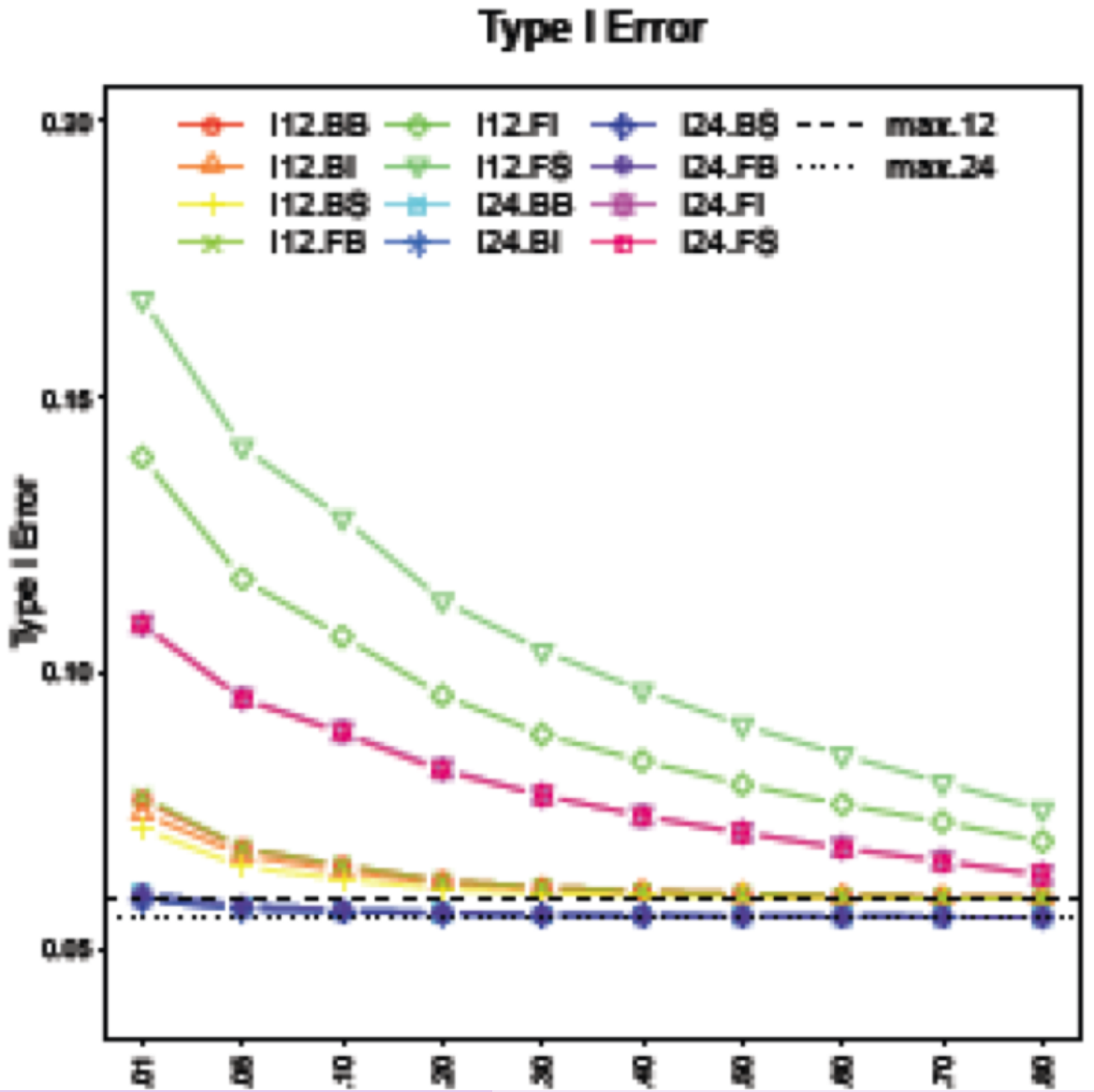
Data-driven random effects selection

BB: backwards best path
BI: backwards item first
BS: backwards subj first

FB: forwards best path
FI: forwards item first
FS: forwards subj first

X axis gives alpha levels for testing inclusion of random slopes.

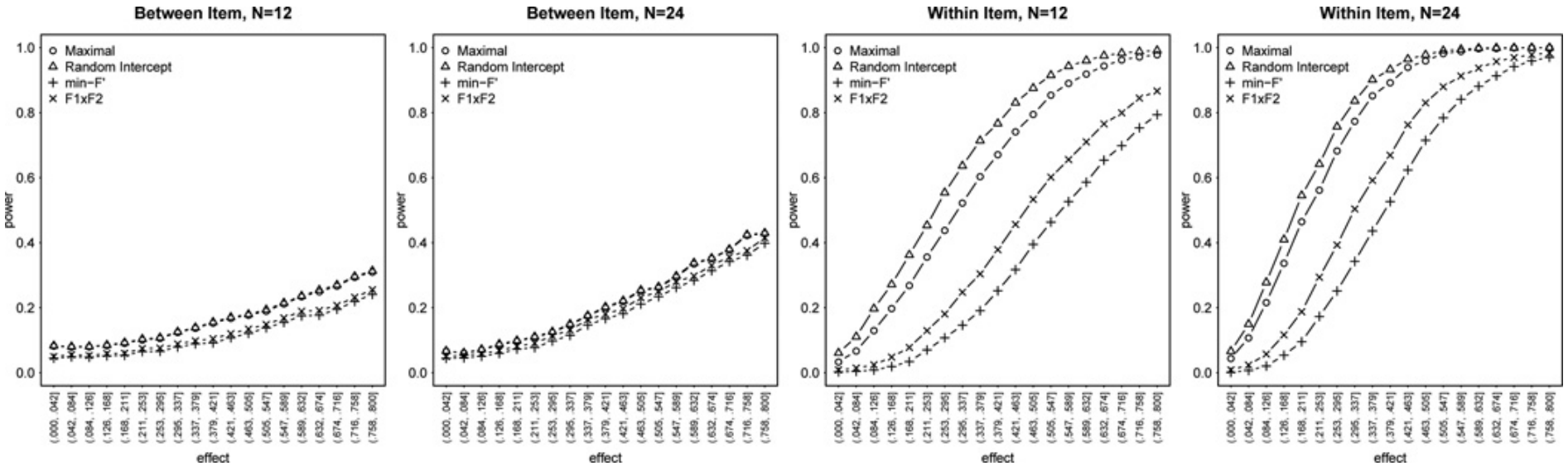
⇒ if model selection necessary for conversion, use backward selection!



Worst-case scenario for maximal models

How bad is the loss of power due to inclusion of max random effects?

Experiment: fit maximal models to data generated by a random-intercept only process.



Even in this scenario

- maximal models are not far off the power of the correct RI-only model.
- and more powerful than F1 x F2 (ANOVA)

Table of Contents

Before running the model...

If your model includes interactions or variables that have some collinearity

- you can check this via VIF (variance inflation factor)
- ... you should "standardize" your predictors
  standardize = subtract the mean and divide by sd.
- but keep in mind that this affects interpretability of the coefficients.
  to interpret them later on, you have to back-transform
  (de-standardize) them.

Summary

- compare models using AIC
- do backward selection of fixed effects structure
- forward selection on random effects structure
- if you have interactions in your data, you should standardize the data for faster convergence and better detection of effects and interactions
- back-transforming your data for interpretation
- concept of power: power = 1- Type II error
- power can be increased by larger effect size and / or larger sample size
- power analysis can help you estimate how much data to collect if you know effect size at least approximately beforehand.