POWER AND SAMPLE SIZE BY SIMULATIONS USING R

MIXED MODELLING IN R WORKSHOP

AARHUS UNIVERSITY
DEPARTMENT OF ANIMAL AND VETERINARY SCIENCES

LESLIE FOLDAGER
SENIOR RESEARCHER

BiRC

SOLIDUM PETIT IN PROFUNDIS UNIVERSITAS ARHUSIENSIS

ABOUT ME

- Senior Researcher at Dept. of Animal and Veterinary Sciences, Aarhus University
  - Mainly participating in projects
  - Consultancy (approx. 2 month/year)
  - Teaching … not much – so far
- Affiliated Bioinformatic Research Centre (BiRC), Aarhus University, Aarhus, 2008-
- MSc in statistics from Dept. of Theoretical Statistics, AU, 1992-1998
- PhD in medicine, Health, AU (psychiatric genetics), 2007-2014

25+ years of experience as (bio)statistician in a variety of fields:

- **ANIS/ANIVET, AU, Mar 2015-**
- Dept. of Public Health, AU, Jan-Feb 2015
- Psychiatric Research, Risskov, 2002-2014 (Aarhus Amt, RM 2007, AU 2013)
- **Biometry Research Unit, Danish Institute of Agricultural Sciences (DJF), 2000-2002**
- Aalborg Hospital, Clinical physiology and nuclear medicine,1999-2000
- ConStat, Hirtshals, 1998-1999 (statistical consultancy mainly in Fisheries research)

LITERATURE FOR INSPIRATION

Arnold *et al. BMC Medical Research Methodology* 2011, **11**:94
http://www.biomedcentral.com/1471-2288/11/94

BMC
Medical Research Methodology

COMMENTARY Open Access

Simulation methods to estimate design power: an overview for applied research

Benjamin F Arnold[1*†], Daniel R Hogan[2†], John M Colford Jr[1] and Alan E Hubbard[3]

http://www.biomedcentral.com/1471-2288/11/94

I will go through the `simr` paper

Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2016, **7**, 493–498 doi: 10.1111/2041-210X.12504

APPLICATION

SIMR: an R package for power analysis of generalized linear mixed models by simulation

Peter Green* and Catriona J. MacLeod

Landcare Research, Private Bag 1930, Dunedin 9054, New Zealand

https://doi.org/10.1111/2041-210X.12504

WHY ARE POWER CALCULATIONS NEEDED? E.G.,

- **Optimise use of fundings: not to waste resources or miss real effects**
  - If 10 rats per treatment is enough, then we should not use 50.
  - If 5 plots are not enough, then perhaps we should aim for more.
- **Ethical reasons**
  - If 13 pigs per treatment is enough, then we should not use sacrifice 50.
  - If 5 repeats is enough, then we should not stick a needle into the animal more times.
- **Better chance of publishing – and satisfying the editor and reviewers**
  - It is just easier to write the paper if there is something significant to report.
  - Quite often the reviewers or editor will ask what the power was.
  - Calculating post-hoc is not advisable! So do it before the experiment!

WHAT TO DO IF POWER IS TOO LOW?

If say 50 cows is not enough to have a reasonable chance to find the hypothesised effects and we cannot afford more cows, then perhaps

a) Not do the experiment
b) Pick another method having a higher precision and/or better accuracy
c) Rethink the experiment and perhaps focus on fewer outcomes
d) Get more funding so that we can afford a larger experiment
e) …

- **Increasing the sample size may help.**
- **Decreasing the variance may help.**
- **Reduce the number of tests if multiple comparison is an issue.**

TYPE I ERROR, TYPE II ERROR, AND POWER

- **Type I error (α)**
  - The mistaken rejection of a null hypothesis (H_0) that is actually true.
  - False positive
  - "an innocent person is convicted"
- **Type II error (β)**
  - The failure to reject a null hypothesis that is actually false.
  - False negative
  - "a guilty person is not convicted"
- **Power = 1 - β**
  - True positive
  - "a guilty person is convicted"

Table of error types		TRUTH Null hypothesis (H_0) is	
		True	**False**
TEST **Decision about null hypothesis (H_0)**	**Fail to reject**	Correct inference (true negative) (probability = $1-\alpha$)	**Type II error** Type II error (false negative) (probability = β)
	Reject	**Type I error** Type I error (false positive) (probability = α)	**Power** Correct inference (true positive) (probability = $1-\beta$)

Significance level

WHY USE SIMULATION

- **There are cases where we could use analytical formulas to calculate the power**
  - But these are typically either approximations
  - or require special designs – perhaps too simple to be relevant
- **Even if appropriate formulas exist,**
  - they might be difficult to understand
  - more time consuming to find and use than running simulation
- **Setting up a simulation experiment could**
  - also be (too) complicated for many researchers
- **The *simr* R package**
  - is made to provide a solution than avoid some of the complexity
  - but still allows for a lot of flexibility

HOW THE POWER CALCULATIONS WORK

- **The basic idea** is to simulate a model under the **alternative hypothesis** (H_a) a larger number of times. That is, we simulate under the assumption that the null hypothesis of no effect (H_0) is false.
- **Then count** how often the test fails to detect the effect, i.e. **errors of Type II**.
- Equivalently find **the proportion of tests** being below the chosen level of significance, i.e. the proportion of **correct rejections** of H_0, which under H_a is **the power**.
- **In `simr` the following steps are each repeated `nsim` times:**
  1) Simulate a new set of data using the fitted model provided.
  2) Refit the model to the simulated data.
  3) Apply a test to the simulated fit.

LET US SEE HOW THIS IS DONE IN PRACTICE

We will use the artificial data set `simdata` from `simr`

```
> str(simdata)
'data.frame':   30 obs. of  4 variables:
 $ y: num  8.14 7.95 9.28 7.78 5.8 ...
 $ x: int  1 2 3 4 5 6 7 8 9 10 ...
 $ g: Factor w/ 3 levels "a","b","c": 1 1
  1 1 1 1 1 1 1 1 ...
 $ z: int  3 3 3 2 3 2 3 0 2 1 ...
>
```

SIMULATION OF TRAINING DATA LIKE SIMDATA

- Instead of using the functions from `simr` we could also simulate data in the way `simdata` was generated
- . . . depending of course on the model that we wish to examine.
- This can also be a way to obtain training data to be used for the simulation using the `simr` functions.

INITIATING THE POWER ANALYSIS – FITTING A MODEL

- The power analysis in `simr` starts with a model fitted in `lme4` i.e., a model fitted with `lmer` or `glmer`. Models fitted using `lm` and `glm` (i.e. without random effects) are also supported.

```
> model1 <- glmer(z ~ x + (1|g), family="poisson", data=simdata)
> summary(model1)
# <snip>
 Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
 (Intercept)  1.54079    0.27173   5.670 1.43e-08 ***
 x           -0.11481    0.03955  -2.903   0.0037 **
```

- Let us have a look at how `simdata` was generated.

SPECIFYING AN EFFECT SIZE

- **Recall that the power is the proportion of TRUE rejections of a FALSE null hypothesis.** (table repeated on next slide)
- **We therefore want to simulate under the alternative hypothesis by specifying the effect size we wish to be able to find if it truly exists, i.e., if the null hypothesis is FALSE.**
- **Let us consider the power to detect a slope of -0.05 (on the log link scale) in the mixed effects Poisson regression that we fitted in `model1`:**

```
> fixef(model1)["x"]
        x
-0.1148147
> fixef(model1)["x"] <- -0.05
```

RECALL THE TWO-BY-TWO TABLE

Table of error types		TRUTH Null hypothesis (H_0) is	
		True	False
TEST Decision about null hypothesis (H_0)	Fail to reject	Correct inference (true negative) (probability = $1-\alpha$)	Type II error Type II error (false negative) (probability = β)
	Reject	Type I error Type I error (false positive) (probability = α)	Power Correct inference (true positive) (probability = $1-\beta$)

Significance level

OTHER CHANGES OF PARAMETERS IN THE MODEL

- It is also possible to change the parameters of the random effects and the residual variance
- See the description in the documentation here:

```
> ?modify
```

RUNNING THE POWER ANALYSIS

- It may be an idea to use `set.seed(#)` if you want to be able to reproduce the results!

```
> set.seed(123)
> powerSim(model1)
```

- Let us do this in R
- . . . it takes a couple of minutes depending on your hardware.
- It is possible choose another test, e.g., the parametric bootstrap test `PBmodcomp` from the package `pbkrtest`: `> ?tests`
- The number of simulations (`nsim=1000`) can also be changed.

```
> powerSim(model1,nsim=100)
```

INCREASING POWER BY CHANGING DESIGN

- The power was not good enough to find the effect size of -0.05 so what do we do?
- Perhaps we could increase the sample size.
- Perhaps we could increase the number of groups.
- Imagine that the current pilot data corresponded to observations from 10 weeks.
- Perhaps we could extend the experiment up to 20 weeks
- . . . let us try to do the power calculation in this scenario

```
> fixef(model1)["x"]
        x
-0.1148147
> fixef(model1)["x"] <- -0.05
```

INCREASING THE SAMPLE SIZE

- To change x so that it runs from 1 to 20 instead of 1 to 10, we can use the function `extend`:

```
> model2 <- extend(model1,along="x",n=20)
> powerSim(model2)
```

- Let us try this in R (again it takes a couple of minutes)
- The `along` argument specifies which variable is being extended, and `n` specifies how many levels to replace it with. Let us have a look:

```
> getData(model2)
```

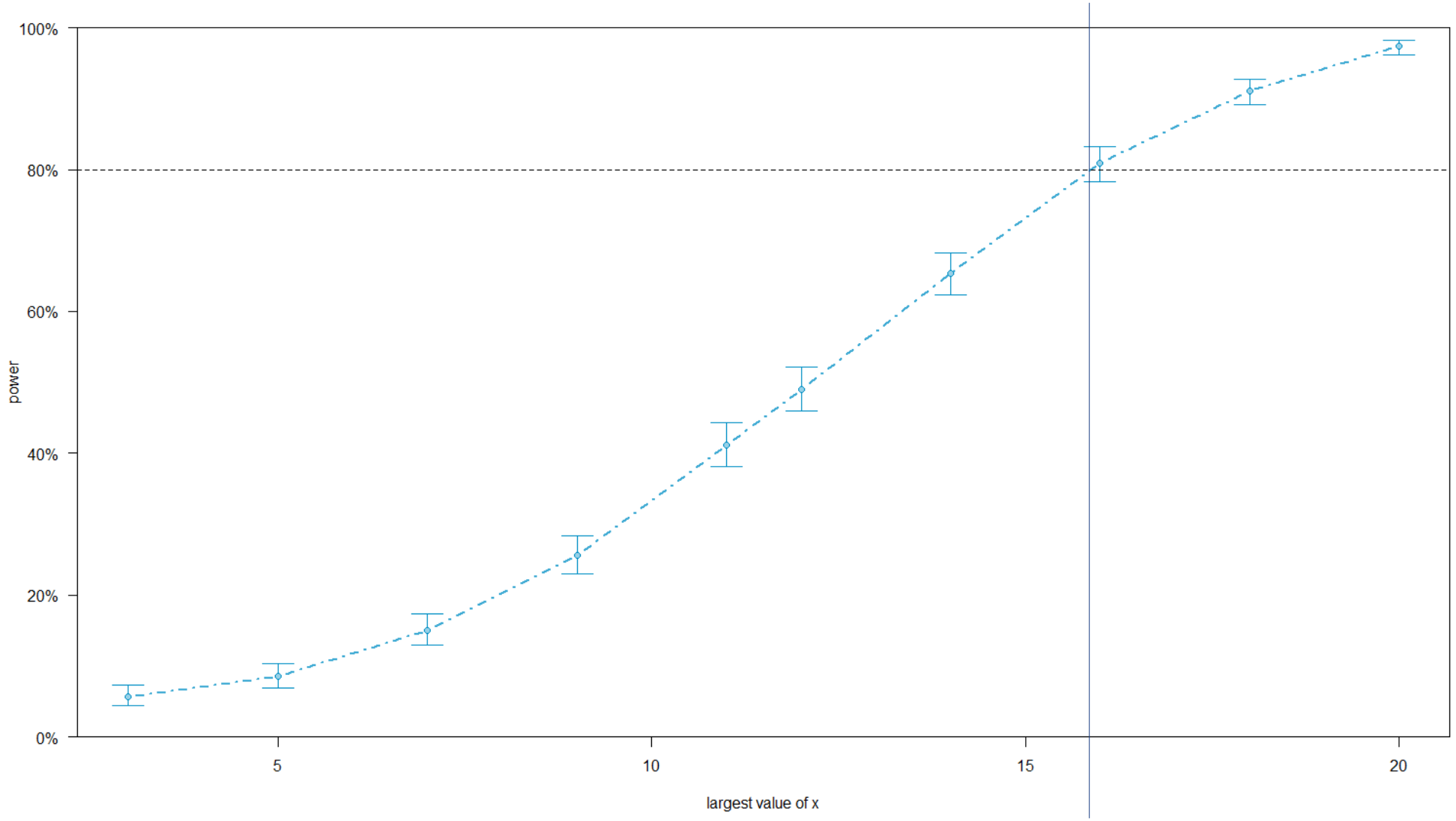
OPTIMISATION OF SAMPLE SIZE – EXAMINING A RANGE

- Perhaps a power of 97% is a bit too much and a waste of resources!?
- So let us try a range of samples sizes to figure what will be optimal, e.g., when a power of 80% is reached.
- We can use the `powerCurved` function to explore this:

```
> pc2 <- powerCurve(model2)
> print(pc2)
> plot(pc2)
```

- This takes a while (10-20 min) so let us cheat.

THE RESULTING POWER CURVE FOR `model2`



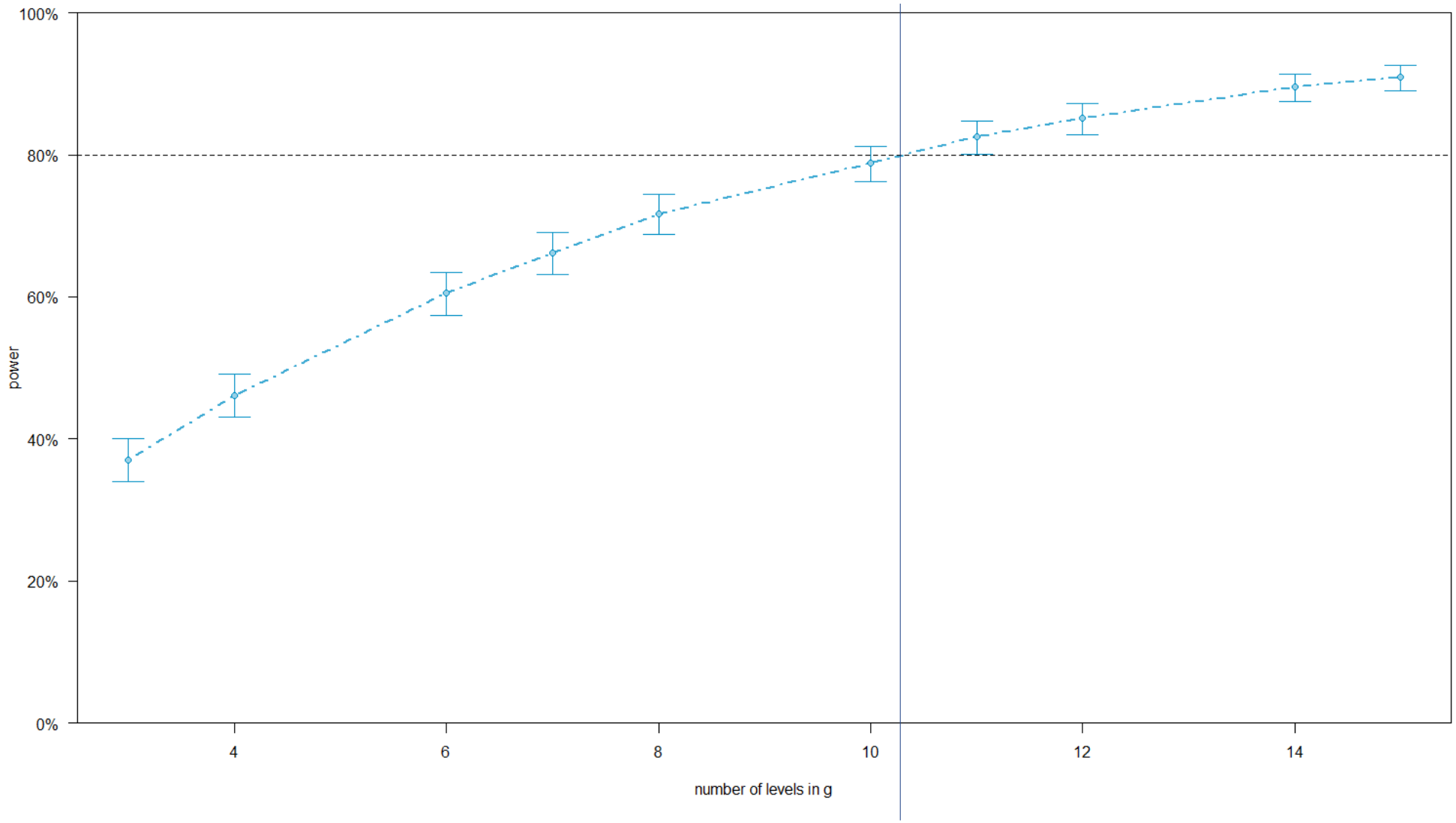
OPTIMISATION – ADDING MORE GROUPS

- Perhaps we would not have the time to wait for 16+ weeks to obtain a power of at least 80%
- So let us instead try to increase the number and the size of the groups.
- To increase the number of groups, we can again use the `extend` function with the `along` argument :

```
> model3 <- extend(model1,along="g",n=15)
> dat3 <- getData(model3)
> pc3 <- powerCurve(model3, along="g")
```

- Note the use of `along` argument in `powerCurve`.

THE RESULTING POWER CURVE FOR `model3`



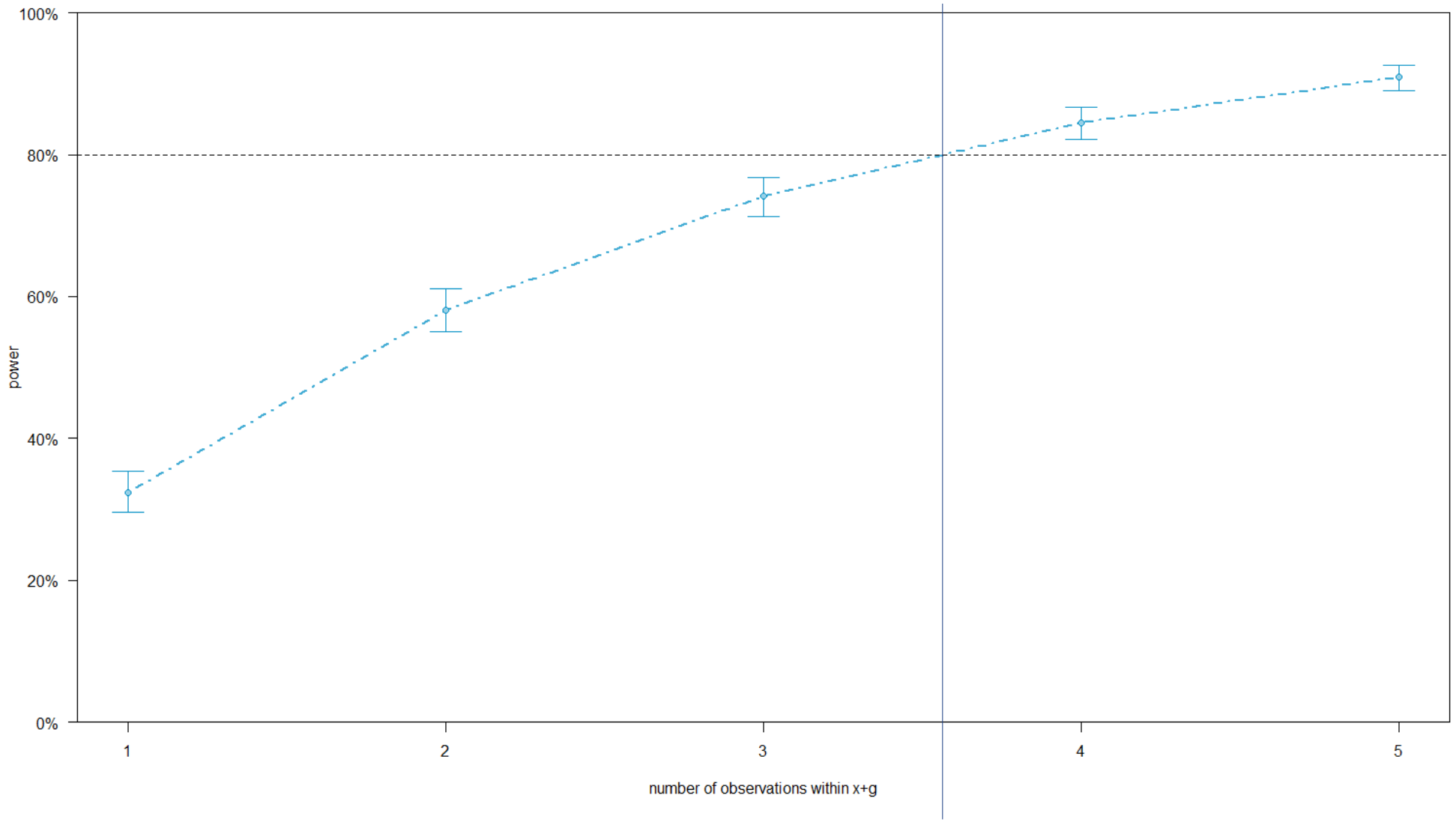
OPTIMISATION – INCREASING GROUP SIZE

- Another option might be to change the group size.
- The size within group can be changed by use of the argument `within` instead of `along` in the `extend` and `powerCurve` functions:

```
> model4 <- extend(model1,within="x+g",n=5)
> dat4 <- getData(model4)
> pc4 <- powerCurve(model4,within="x+g",breaks=1:5)
```

- Note the use of the `breaks` argument in `powerCurve`!
- This gives us 1 to 5 observations per combination of x and g.

THE RESULTING POWER CURVE FOR MODEL4



SIMULATION UNDER THE NULL HYPOTHESIS

What do you expect we will get

if we set the effect size to zero?

```
> fixef(model1)["x"] <- 0
> powerSim(model1)
```

RECALL THE TWO-BY-TWO TABLE

Table of error types		TRUTH Null hypothesis (H_0) is	
		True	False
TEST Decision about null hypothesis (H_0)	Fail to reject	Correct inference (true negative) (probability = $1-\alpha$)	Type II error Type II error (false negative) (probability = β)
	Reject	Type I error Type I error (false positive) (probability = α)	Power Correct inference (true positive) (probability = $1-\beta$)

Significance level

RESULT FOR MODEL 1 WITH X=0

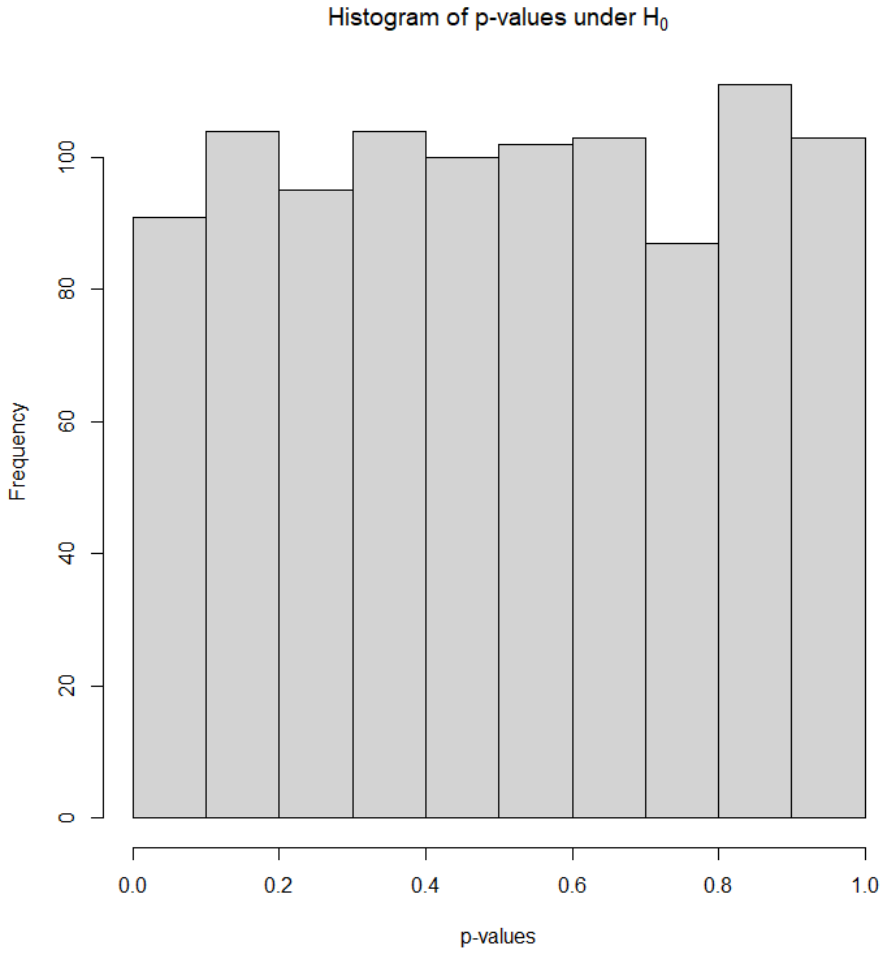
```
Power for predictor 'x', (95% confidence interval):

      4.60% ( 3.39,  6.09)

Test: z-test

     Effect size for x is 0.0

Based on 1000 simulations, (0 warnings, 0 errors)

alpha = 0.05, nrow = 30

Time elapsed: 0 h 1 m 34 s
```

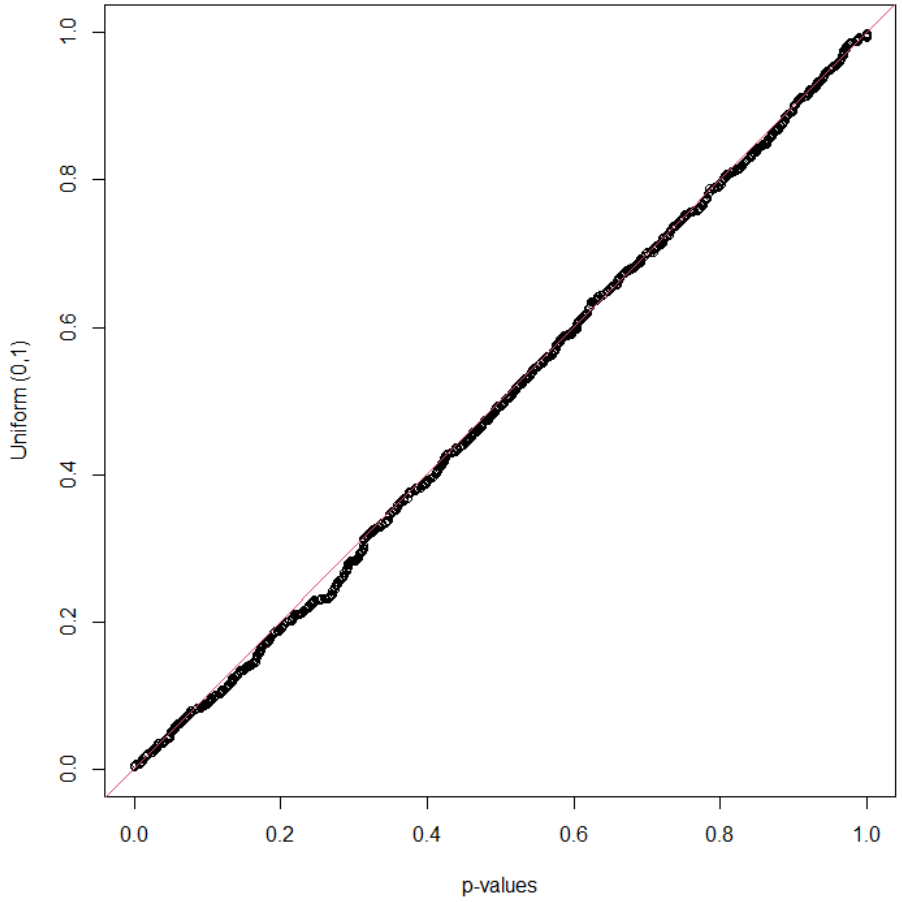
CHECK SIMULATION UNDER THE NULL HYPOTHESIS!

- **Check that the significance level is as set**
  - Here the p-values below 0.05 (say) are false positives

    . . . not true positives.
  - This is the Type I error rate.
  - A histogram or a qq-plot against the uniform distribution should show a uniform distribution of the p-values.
  - Let us see how this looks ... and another way to simulate by use of `simulate.merMod` directly.

THE RESULTING POWER CURVE FOR `model4`





WHY NOT TO DO POST-HOC POWER CALCULATIONS!

- Note the last 3 lines of 2nd column on page 494 in Green & MacLeod (2016):

  "Retrospective 'observed power' calculations, where the target effect size comes from the data, give misleading results (Hoening & Heisey 2001)"

  Hoenig, J.M. & Heisey, D.M. (2001) The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician*, 55, 19–24.
  https://doi.org/10.1198/000313001300339897

- From their Abstract: "... There is also a large literature advocating that power calculations be made whenever one performs a statistical test of a hypothesis and one obtains a statistically nonsignificant result. ... **This approach, which appears in various forms, is fundamentally flawed.** ..."

QUOTES FROM HOENIG & HEISEY (2001)

"This can be called the dilemma of the nonrejected null hypothesis: what should we do when we fail to reject a hypothesis?"

"In our experience as consulting statisticians, authors are not infrequently required to perform such calculations by journal reviewers or editors;"

MORE QUOTES FROM HOENIG & HEISEY (2001)

"... arisen due to applied scientists being heavily tradition-bound to test the usual "no impact null hypothesis," despite it not always being the relevant null hypothesis for the question at hand".

"... more emphasis on the investigator's choice of hypotheses and on the interpretation of confidence intervals."

"... suggest that introducing the concept of equivalence testing may help students understand hypothesis tests."

"OBSERVED POWER" IS 1:1 FUNCTION OF THE P-VALUE

- Let the "observed power" be the power of the test for the observed value of the test statistic. That is, assume the effects and variability are equal to the true parameter values.

  "... for any test the observed power is a 1:1 function of the p value."

- An example is given in Figure 1 of Hoenig & Heisey (2001), and they give various other arguments against post-hoc power calculations.
- **"Power calculations tell us how well we might be able to characterize nature in the future given a particular state and statistical study design, but they cannot use information in the data to tell us about the likely states of nature."**

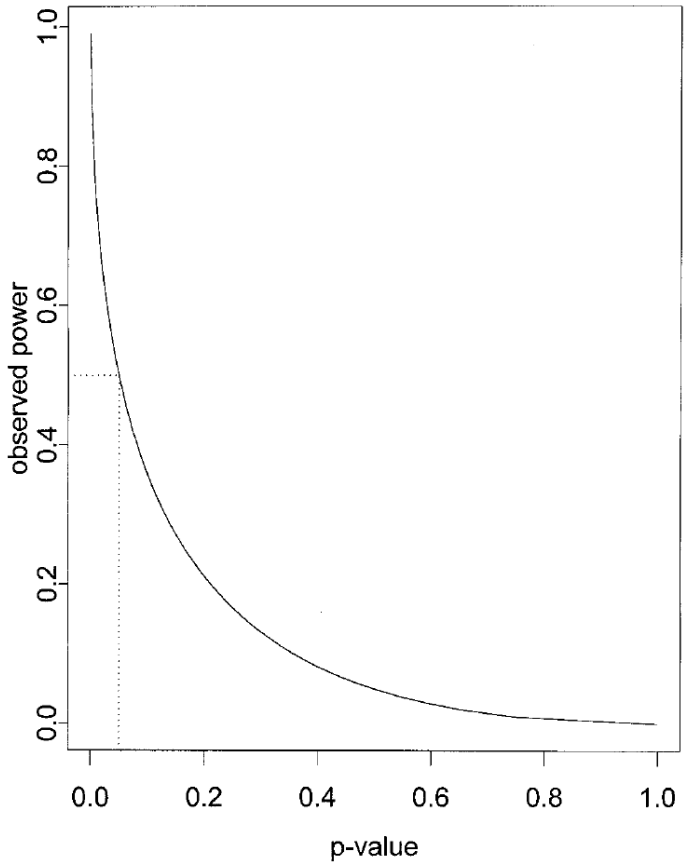


Figure 1. "Observed" Power as a Function of the p Value for a One-Tailed Z Test in Which α is Set to .05. When a test is marginally significant (P = .05) the estimated power is 50%.

RUNNING MORE EXAMPLES

- Appendix 1 of Green & MacLeod (2016) https://doi.org/10.1111/2041-210X.12504 can be found as the vignette Test examples:
- https://cran.r-project.org/web/packages/simr/vignettes/examples.html

- Appendix 2 of Green & MacLeod (2016) contains a vignette that shows how to start the power analysis from scratch, including making a pilot data set:
- https://cran.r-project.org/web/packages/simr/vignettes/fromscratch.html

Thank you for your attention

mailto://leslie@anivet.au.dk
http://anivet.au.dk/en
http://birc.au.dk
Office: 8867/K27.3205

AARHUS
UNIVERSITY
DEPARTMENT OF ANIMAL AND VETERINARY SCIENCES

BiRC

SOLIDUM PETIT IN PROFUNDIS · UNIVERSITAS ARHUSIENSIS ·