Course Business

- **Final papers** due on Canvas on Monday, Nov. 23rd (2 weeks from Monday)

- Meteyard & Davies (2020) paper provides some useful guidelines on reporting

---

**Best practice guidance for linear mixed-effects models in psychological science**

Lotte Meteyard¹, Robert A. I. Davies²

¹ School of Psychology & Clinical Language Science, University of Reading, Berkshire RG6 6AL, UK
² Department of Psychology, Lancaster University, Lancaster LA1 4YF, UK

**Abstract**

The use of Linear Mixed-effects Models (LMMs) is set to dominate statistical analyses in psychological science and may become the default approach to analyzing quantitative data. The rapid growth in adoption of LMMs has been matched by a proliferation of different approaches. Unlike traditional fixed-effect models, LMMs can be used to analyze data with arbitrary and complex covariate structures. The use of LMMs can also help to address many of the concerns raised by other statistical approaches. LMMs allow for the systematic review of papers using LMMs to be performed with more confidence in the reliability of their findings. The review of papers complements the survey, showing variation in how the models are built, how effects are evaluated and, most worryingly, how models are reported. Using these data as a departure point, we present a set of best-practice guidelines. Scoring on the reporting of LMMs. It is the authors’ intention that the paper supports a step-change in the reporting of LMMs across the psychological sciences, preventing a trajectory in which findings reported today cannot be transparently understood and used tomorrow.

**Introduction**

Linear Mixed-effects Models (LMMs) have become increasingly popular as a data analysis method in the psychological sciences. They are also known as hierarchical or multilevel or random effects models. LMMs are warranted when data are collected according to a multi-stage sampling or repeated measures design. That is, when there are likely to be correlations across the conditions of an experiment because the conditions include the same participants or participants who have some association with each other. Multi-stage sampling can arise naturally when collecting data about the behavior or attributes of participants recruited, e.g., as students from a sample of classes in a sample of schools, or as patients from a sample of clinics in a sample of regions. Repeated measures occur when participants experience or are more than one of the manipulated experimental conditions, or when all participants are presented with all stimuli. Such investigations are common in psychology. These designs yield data-sets that have a multilevel or hierarchical structure. Participant-level observations, e.g., an individual’s measured skill level or scores, can be grouped within the classes or schools, clinics or regions from which the participants are recruited. Trial-level observations, e.g., the latency of response to a stimulus word, can be grouped by the participants to whom or by the stimuli presented.

We expect that the responses made by a participant to some stimuli will be correlated, or that responses from children in the same class or school or region will be correlated, or that responses to the same stimuli from across participants will be correlated. The hierarchical structure in the data (the ways in which data can be grouped) is associated with a hierarchical structure in the error variance. LMMs allow this structure to be explicitly modelled.

We review current practice for LMMs in the psychological sciences. To begin, we present an example of a mixed-effects analysis (Section 1.1), with the aim of clearly illustrating how random effects relate to fixed effects. Researchers who are comfortable in their conceptual understanding of LMMs may wish to skip this part. Following the example, we present data from a survey of researchers (Section 2) and a review of reporting practices in papers published between 2013 and 2016 (Section 3). Our observations reveal significant concerns in the community over the implementation of LMMs, and a worrying range of reporting practices in published papers (Section 4). Using the available literature, we then present best-practice guidelines (Section 4.1) with a bullet-point summary (Section 5). To prevent two key conclusions, researchers should be measured that there is no single correct way to implement an LMM, and that the choices they make during analysis will

---

¹ Corresponding author.
E-mail address: Lmeteyard@reading.ac.uk (L. Meteyard).

https://doi.org/10.1016/j.nlm.2020.10.010
Received 18 March 2019; Received in revised form 14 January 2020; Accepted 15 January 2020
© 2020 Elsevier Inc. All rights reserved.
**Course Business**

- Last lecture & lab today!
  - Lab materials and data on Canvas
  - Package to install: `simr`

- “Bonus” lecture on Canvas on signal detection
  - Relevant for datasets where participants are making a binary *decision* (e.g., recognition memory, grammaticality judgments, moral dilemmas, etc.)

- Teaching survey (OMET) begins Monday

- Thanks for working with the challenges of this term!
Week 12.2: Power

- Statistical Power
  - Intro to Power Analysis
  - Estimating Effect Size
  - Why Do We Care About Power?
  - Determining Power
  - Power Simulations in R
  - Influences on Power

- Lab
Recap of Null Hypothesis Significance Testing

• Does “brain training” affect general cognition?
  • $H_0$: There is no effect of brain training on cognition
  • $H_A$: There is an effect of brain training on cognition
Recap of Null Hypothesis Significance Testing

• Does “brain training” affect general cognition?
  • $H_0$: There is no effect of brain training on cognition
  • $H_A$: There is an effect of brain training on cognition

These two books contain the sum total of human knowledge…
Let’s consider a world where $H_0$ is true—there is no effect of brain training on general cognition.

A Consensus on the Brain Training Industry from the Scientific Community

To date, there is little evidence that playing brain games improves underlying broad cognitive abilities, or that it enables one to better navigate a complex realm of everyday life. Some intriguing
Recap of Null Hypothesis Significance Testing

• Let’s consider a world where $H_0$ is true—there is no effect of brain training on general cognition
• Two possible outcomes…

<table>
<thead>
<tr>
<th>Actual State of the World</th>
<th>WHAT WE DID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retain $H_0$</td>
<td>GOOD!</td>
</tr>
<tr>
<td>Reject $H_0$</td>
<td>OOPS!</td>
</tr>
</tbody>
</table>

- Retain $H_0$: $H_0$ is true
  - Probability: 1-$\alpha$
- Reject $H_0$: $H_0$ is false
  - Probability: $\alpha$
Recap of Null Hypothesis Significance Testing

• What about a world where $H_A$ is true?

Cognitive Training Data Response Letter

In October 2014, the Stanford Center on Longevity released a statement titled "A Consensus on the Brain Training Industry from the Scientific Community." However, the statement did not reflect a true consensus from the community. Please see below for an open letter response signed by well over 100 neuroscientists, psychologists, and other experts in the field of neural plasticity.

An Open Letter

To the Stanford Center on Longevity:

The Controversy

Does scientific evidence show brain training works?

Yes, although not all scientists agree. That may be because the question itself is flawed: it assumes all brain training is essentially the same.
Recap of Null Hypothesis Significance Testing

- Another mistake we could make: There really is an effect, but we retained $H_0$
  - False negative / Type II error
  - Historically, not considered as “bad” as Type I
  - Probability: $\beta$

<table>
<thead>
<tr>
<th>ACTUAL STATE OF THE WORLD</th>
<th>WHAT WE DID</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$ is true</td>
<td>Retain $H_0$</td>
</tr>
<tr>
<td>$H_A$ is true</td>
<td>Reject $H_0$</td>
</tr>
</tbody>
</table>

OOPS! Type II error Probability: $\beta$
Recap of Null Hypothesis Significance Testing

Never confuse Type I and II errors again:

Just remember that the Boy Who Cried Wolf caused both Type I & II errors, in that order.

First everyone believed there was a wolf, when there wasn't. Next they believed there was no wolf, when there was.

Substitute "effect" for "wolf" and you're done.

Kudos to @danolner for the thought. Illustration by Francis Barlow "De pastoris puero et agricolis" (1687). Public Domain. Via wikimedia.org
Recap of Null Hypothesis Significance Testing

- **POWER (1-β):** Probability of correct rejection of $H_0$: detecting the effect when it really exists
- If our hypothesis ($H_A$) is right, what probability is there of obtaining significant evidence for it?

<table>
<thead>
<tr>
<th>ACTUAL STATE OF THE WORLD</th>
<th>WHAT WE DID</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$ is true</td>
<td>Retain $H_0$</td>
</tr>
<tr>
<td>$H_A$ is true</td>
<td>✅ GOOD! Probability: 1-α</td>
</tr>
<tr>
<td></td>
<td>✗ OOPS! Type II error Probability: β</td>
</tr>
</tbody>
</table>
Recap of Null Hypothesis Significance Testing

- **POWER (1-β):** Probability of correct rejection of $H_0$: detecting the effect when it really exists
- Can we find the thing we’re looking for?
Recap of Null Hypothesis Significance Testing

- **POWER (1-β):** Probability of correct rejection of **$H_0$: detecting the effect when it really exists**
  - Can we find the thing we’re looking for?
  - If our hypothesis is true, what is the probability we’ll get $p < .05$?

- We compare retrieval practice to re-reading with power = .75
  - If retrieval practice is actually beneficial, there is a 75% chance we’ll get a significant result

- We compare bilinguals to monolinguals on a test of non-verbal cognition with power = .35
  - If there is a difference between monolinguals & bilinguals, there is a 35% chance we’ll get $p < .05$
**Power Analysis**

- **Power analysis**: Do we have the power to detect the effect we’re interested in?
  - Depends on effect size, $\alpha$ (Type I error rate), and sample size.

- **In practice**:
  - We **can’t** control effect size; it’s a property of nature.
  - $\alpha$ is usually fixed (e.g., at .05) by convention.
  - But, we **can** control our sample size $n$!

- **So**:
  - Determine desired power (often .80).
  - Estimate the effect size(s).
  - Calculate the necessary sample size $n$. 
Week 12.2: Power

- Statistical Power
- Intro to Power Analysis
  - Why Do We Care About Power?
  - Determining Power
  - Power Simulations in R
  - Influences on Power
- Lab
Estimating Effect Size

- One reason we haven’t always calculated power is it requires the **effect size** of the effect we’re looking for.

- But, two ways to **estimate** effect size:
  1. Prior literature
     - What is the effect size in other studies in this domain or with a similar manipulation?
Estimating Effect Size

• One reason we haven’t always calculated power is it requires the effect size of the effect we’re looking for
• But, two ways to estimate effect size:
  1. Prior literature
  2. Smallest Effect Size Of Interest (SESOI)
     • Decide smallest effect size we’d care about
     • e.g., we want our educational intervention to have an effect size of at least .05 GPA
     • Calculate power based on that effect size
     • True that if actual effect is smaller than .05 GPA, our power would be lower, but the idea is we no longer care about the intervention if its effect is that small
Week 12.2: Power

- Statistical Power
  - Intro to Power Analysis
  - Estimating Effect Size
- Why Do We Care About Power?
  - Determining Power
  - Power Simulations in R
  - Influences on Power

- Lab
Why Do We Care About Power?

1. Efficient **use of resources**
   - A major determinant of power is **sample size** (larger = more power)
   - Power analyses tell us if our planned sample size \( n \) is:
     - Large enough to be able to find what we’re looking for
     - Not too large that we’re collecting more data than necessary
Why Do We Care About Power?

1. Efficient **use of resources**
   - A major determinant of power is **sample size** (larger = more power)
   - Power analyses tell us if our planned **sample size** ($n$) is:
     - Large enough to be able to find what we’re looking for
     - Not too large that we’re collecting more data than necessary
   - This is about good use of our resources
     - Societal resources: Money, participant hours
     - **Your** resources: Time!!
Why Do We Care About Power?

1. Efficient **use of resources**
2. Avoid **p-hacking** (Simmons et al., 2011)
   - Rate of false positive results increases if we keep collecting data whenever our effect is non-sig.
   - In the limit, *ensures* a significant result
     - Random sampling means that p-value is likely to differ in each sample

![Graph showing p-value vs. sample size](image-url)
Why Do We Care About Power?

1. Efficient **use of resources**
2. Avoid **p-hacking** (Simmons et al., 2011)
   - Rate of false positive results increases if we keep collecting data whenever our effect is non-sig.
   - In the limit, *ensures* a significant result
     - Random sampling means that p-value is likely to differ in each sample
   
   ![Graph](Image) 

   Now, $p$-value happens to be lower

   (number of observations in each of two conditions)
Why Do We Care About Power?

1. Efficient **use of resources**
2. Avoid **p-hacking** (Simmons et al., 2011)
   - Rate of false positive results increases if we keep collecting data whenever our effect is non-sig.
   - In the limit, **ensures** a significant result
     - Random sampling means that p-value is likely to differ in each sample
     - At some point, $p < .05$ by chance

![Graph showing p-value vs. sample size]
Why Do We Care About Power?

1. Efficient use of resources
2. Avoid p-hacking (Simmons et al., 2011)
   - Rate of false positive results increases if we keep collecting data whenever our effect is non-sig.
   - In the limit, ensures a significant result
     - Random sampling means that p-value is likely to differ in each sample
     - At some point, $p < .05$ by chance
     - Bias to get positive results if we stop if and only if $p < .05$

But not significant in this even larger sample
Why Do We Care About Power?

1. Efficient use of resources
2. Avoid p-hacking (Simmons et al., 2011)
   - Rate of false positive results increases if we keep collecting data whenever our effect is non-sig.
   - We can avoid this if we use a power analysis to decide our sample size in advance
Why Do We Care About Power?

1. Efficient **use of resources**
2. Avoid **p-hacking** (Simmons et al., 2011)
3. Understand **non-replication** (Open Science Collaboration, 2015)
   - Even if an effect exists in the population, we’d **expect** some non-significant results
   - Power is almost never 100%
   - In fact, many common designs in psychology have **low power** (Etz & Vandekerckhove, 2016; Maxwell et al., 2015)
   - Small to moderate **sample sizes**
   - Small **effect sizes**
Why Do We Care About Power?

1. Efficient use of resources
2. Avoid p-hacking (Simmons et al., 2011)
3. Understand non-replication (Open Science Collaboration, 2015)

What can't you reliably detect with n = 20?

- People who like spicy food are more likely to like Indian food (n = 26)
- Liberals rate social equality as more important than do conservatives (n = 34)
- Men weigh more than women (n = 46)
- People who like eggs report eating egg salad more often (n = 48)
- Smokers think smoking is less likely to kill someone than do non-smokers (n = 144)
Why Do We Care About Power?

1. Efficient **use of resources**
2. Avoid **p-hacking** (Simmons et al., 2011)
3. Understand **non-replication** (Open Science Collaboration, 2015)
   - Even if an effect exists in the population, we’d **expect** some non-significant results
     - Power is almost never 100%
     - In fact, many common designs in psychology have **low power** (Etz & Vandekerckhove, 2016; Maxwell et al., 2015)
       - Small effect sizes
       - Small to moderate sample sizes
     - Failures to replicate might be a sign of low power, rather than a non-existent effect
Why Do We Care About Power?

1. Efficient **use of resources**
2. Avoid **p-hacking** (Simmons et al., 2011)
3. Understand **non-replication** (Open Science Collaboration, 2015)
4. Understand **null results**
   - A non-significant result, by itself, doesn’t **prove** an effect doesn’t exist
   - We “fail to reject H₀” rather than “accept H₀”
Why Do We Care About Power?

1. Efficient **use of resources**
2. Avoid **p-hacking** (Simmons et al., 2011)
3. Understand **non-replication** (Open Science Collaboration, 2015)
4. Understand **null results**
   - A non-significant result, by itself, doesn’t **prove** an effect doesn’t exist
   - We “fail to reject H₀” rather than “accept H₀”
   - “Absence of evidence is not evidence of absence.”

“I looked around Schenley Park for 15 minutes and didn’t see any giraffes. Therefore, giraffes don’t exist.”
Why Do We Care About Power?

1. Efficient **use of resources**
2. Avoid **p-hacking** (Simmons et al., 2011)
3. Understand **non-replication** (Open Science Collaboration, 2015)
4. Understand **null results**
   - A non-significant result, by itself, doesn’t **prove** an effect doesn’t exist
     - We “fail to reject $H_0$” rather than “accept $H_0$”
     - “**Absence of evidence** is not **evidence of absence**.”

We didn’t find enough evidence to conclude there is a significant effect

**DOES NOT** Mean No significant effect exists
Why Do We Care About Power?

1. Efficient use of resources
2. Avoid p-hacking (Simmons et al., 2011)
3. Understand non-replication (Open Science Collaboration, 2015)
4. Understand null results
   • A non-significant result, by itself, doesn’t prove an effect doesn’t exist
     • We “fail to reject $H_0$” rather than “accept $H_0$”
     • “Absence of evidence is not evidence of absence.”
     • Major criticism of null hypothesis significance testing!
Why Do We Care About Power?

1. Efficient **use of resources**
2. Avoid **p-hacking** (Simmons et al., 2011)
3. Understand **non-replication** (Open Science Collaboration, 2015)
4. Understand **null results**
   - A non-significant result, by itself, doesn’t **prove** an effect doesn’t exist
   - But, with high power, null result is more informative
Why Do We Care About Power?

1. Efficient **use of resources**
2. Avoid **p-hacking** (Simmons et al., 2011)
3. Understand **non-replication** (Open Science Collaboration, 2015)
4. Understand **null results**
   - A non-significant result, by itself, doesn’t **prove** an effect doesn’t exist
   - But, with high power, null result is more informative
     - e.g., null effect of working memory training on intelligence with 20% power
       - Maybe brain training works & we just couldn’t detect the effect
     - But: null effect of WM on intelligence with 90% power
       - Unlikely that we just missed the effect!
Why Do We Care About Power?

1. Efficient **use of resources**
2. Avoid **p-hacking** (Simmons et al., 2011)
3. Understand **non-replication** (Open Science Collaboration, 2015)
4. Understand **null results**
   - A non-significant result, by itself, doesn’t **prove** an effect doesn’t exist
   - But, with high power, null result is more informative

A Consensus on the Brain Training Industry from the Scientific Community
Why Do We Care About Power?

1. Efficient **use of resources**
2. Avoid **p-hacking** (Simmons et al., 2011)
3. Understand **non-replication** (Open Science Collaboration, 2015)
4. Understand **null results**
5. **Granting agencies** now want to see it
   - Don’t want to fund a study with low probability of showing anything
   - e.g., Our theory predicts greater activity in Broca’s area in condition A than condition B. But our experiment has only a 16% probability of detecting the difference. Not good!
Why Do We Care About Power?

1. Efficient **use of resources**
2. Avoid **p-hacking** (Simmons et al., 2011)
3. Understand **non-replication** (Open Science Collaboration, 2015)
4. Understand **null results**
5. **Granting agencies** now want to see it
   - NIH:
     
     The updated review language directs peer reviewers to assess the scientific rigor of the experimental design, including the appropriateness of any justification for the sample size selection, under the Approach review criterion for research grant applications. The updated review language formalizes NIH's expectations that statistical power be addressed. Program staff also may, at the IC's discretion, recommend changes to experimental designs, for example, to establish conformity with adopted best practices for the research discipline.
   
   - IES:
     
     For all quantitative inferential analyses, demonstrate that the sample provides sufficient power to address your research aims.
Why Do We Care About Power?

1. Efficient **use of resources**
2. Avoid **p-hacking** (Simmons et al., 2011)
3. Understand **non-replication** (Open Science Collaboration, 2015)
4. Understand **null results**
5. **Granting agencies** now want to see it
6. **Scientific accuracy!**
   • If there is an effect, we want to know about it!
Week 12.2: Power

- Statistical Power
  - Intro to Power Analysis
  - Estimating Effect Size
  - Why Do We Care About Power?
  - Determining Power
    - Power Simulations in R
    - Influences on Power

- Lab
Determining Power

- Power for simpler tests like t-tests or ANOVAs can be simply found from tables
  - Simpler design. Only 1 random effect (at most)

- But, more complicated for mixed effect models
  - Varying number of fixed effects, random intercepts, random slopes
  - Not possible to have a table for every possible design
  - We need a different approach

### Appendix Power: Power as a Function of δ and Significance Level (α)

<table>
<thead>
<tr>
<th>δ</th>
<th>0.10</th>
<th>0.05</th>
<th>0.02</th>
<th>0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>0.26</td>
<td>0.17</td>
<td>0.09</td>
<td>0.06</td>
</tr>
<tr>
<td>1.10</td>
<td>0.29</td>
<td>0.20</td>
<td>0.11</td>
<td>0.07</td>
</tr>
<tr>
<td>1.20</td>
<td>0.33</td>
<td>0.22</td>
<td>0.13</td>
<td>0.08</td>
</tr>
<tr>
<td>1.30</td>
<td>0.37</td>
<td>0.26</td>
<td>0.15</td>
<td>0.10</td>
</tr>
<tr>
<td>1.40</td>
<td>0.40</td>
<td>0.29</td>
<td>0.18</td>
<td>0.12</td>
</tr>
<tr>
<td>1.50</td>
<td>0.44</td>
<td>0.32</td>
<td>0.20</td>
<td>0.14</td>
</tr>
<tr>
<td>1.60</td>
<td>0.48</td>
<td>0.36</td>
<td>0.23</td>
<td>0.17</td>
</tr>
<tr>
<td>1.70</td>
<td>0.52</td>
<td>0.40</td>
<td>0.27</td>
<td>0.19</td>
</tr>
<tr>
<td>1.80</td>
<td>0.56</td>
<td>0.44</td>
<td>0.30</td>
<td>0.22</td>
</tr>
<tr>
<td>1.90</td>
<td>0.60</td>
<td>0.48</td>
<td>0.34</td>
<td>0.25</td>
</tr>
<tr>
<td>2.00</td>
<td>0.64</td>
<td>0.52</td>
<td>0.37</td>
<td>0.28</td>
</tr>
<tr>
<td>2.10</td>
<td>0.68</td>
<td>0.56</td>
<td>0.41</td>
<td>0.32</td>
</tr>
<tr>
<td>2.20</td>
<td>0.71</td>
<td>0.60</td>
<td>0.45</td>
<td>0.35</td>
</tr>
<tr>
<td>2.30</td>
<td>0.74</td>
<td>0.63</td>
<td>0.49</td>
<td>0.39</td>
</tr>
<tr>
<td>2.40</td>
<td>0.78</td>
<td>0.67</td>
<td>0.53</td>
<td>0.43</td>
</tr>
<tr>
<td>2.50</td>
<td>0.80</td>
<td>0.71</td>
<td>0.57</td>
<td>0.47</td>
</tr>
<tr>
<td>2.60</td>
<td>0.83</td>
<td>0.74</td>
<td>0.61</td>
<td>0.51</td>
</tr>
<tr>
<td>2.70</td>
<td>0.85</td>
<td>0.77</td>
<td>0.65</td>
<td>0.55</td>
</tr>
<tr>
<td>2.80</td>
<td>0.88</td>
<td>0.80</td>
<td>0.68</td>
<td>0.59</td>
</tr>
<tr>
<td>2.90</td>
<td>0.90</td>
<td>0.83</td>
<td>0.72</td>
<td>0.63</td>
</tr>
<tr>
<td>3.00</td>
<td>0.91</td>
<td>0.85</td>
<td>0.75</td>
<td>0.66</td>
</tr>
<tr>
<td>3.10</td>
<td>0.93</td>
<td>0.87</td>
<td>0.78</td>
<td>0.70</td>
</tr>
<tr>
<td>3.20</td>
<td>0.94</td>
<td>0.89</td>
<td>0.81</td>
<td>0.73</td>
</tr>
<tr>
<td>3.30</td>
<td>0.95</td>
<td>0.91</td>
<td>0.84</td>
<td>0.77</td>
</tr>
<tr>
<td>3.40</td>
<td>0.96</td>
<td>0.93</td>
<td>0.86</td>
<td>0.80</td>
</tr>
<tr>
<td>3.50</td>
<td>0.97</td>
<td>0.94</td>
<td>0.88</td>
<td>0.82</td>
</tr>
<tr>
<td>3.60</td>
<td>0.98</td>
<td>0.95</td>
<td>0.90</td>
<td>0.85</td>
</tr>
<tr>
<td>3.70</td>
<td>0.98</td>
<td>0.96</td>
<td>0.92</td>
<td>0.87</td>
</tr>
<tr>
<td>3.80</td>
<td>0.98</td>
<td>0.97</td>
<td>0.93</td>
<td>0.89</td>
</tr>
<tr>
<td>3.90</td>
<td>0.99</td>
<td>0.97</td>
<td>0.94</td>
<td>0.91</td>
</tr>
<tr>
<td>4.00</td>
<td>0.99</td>
<td>0.98</td>
<td>0.95</td>
<td>0.92</td>
</tr>
<tr>
<td>4.10</td>
<td>0.99</td>
<td>0.98</td>
<td>0.96</td>
<td>0.94</td>
</tr>
<tr>
<td>4.20</td>
<td>—</td>
<td>0.99</td>
<td>0.97</td>
<td>0.95</td>
</tr>
<tr>
<td>4.30</td>
<td>—</td>
<td>0.99</td>
<td>0.98</td>
<td>0.96</td>
</tr>
<tr>
<td>4.40</td>
<td>—</td>
<td>0.99</td>
<td>0.98</td>
<td>0.97</td>
</tr>
<tr>
<td>4.50</td>
<td>—</td>
<td>0.99</td>
<td>0.99</td>
<td>0.97</td>
</tr>
<tr>
<td>4.60</td>
<td>—</td>
<td>—</td>
<td>0.99</td>
<td>0.98</td>
</tr>
<tr>
<td>4.70</td>
<td>—</td>
<td>—</td>
<td>0.99</td>
<td>0.98</td>
</tr>
<tr>
<td>4.80</td>
<td>—</td>
<td>—</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>4.90</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.99</td>
</tr>
<tr>
<td>5.00</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

Source: The entries in this table were computed by the author.
Monte Carlo Methods

• Remember the definition of power?
  • The probability of observing a significant effect in our sample if the effect truly exists in the population
  • What if we knew for a fact that the effect existed in a particular population?
  • Then, a measure of power is how often we get a significant result in a sample (of our intended $n$)
  • Observe a significant effect 10 samples out of 20 = 50% of the time = power of .50
  • Observe a significant effect 300 samples out of 1000 = 30% of the time = power of .30
  • Observe a significant effect 800 samples out of 1000 = 80% of the time = power of .80
Monte Carlo Methods

• Remember the definition of power?
  • The probability of observing a significant effect in our sample if the effect truly exists in the population
  • What if we knew for a fact that the effect existed in a particular population?
  • Then, a measure of power is how often we get a significant result in a sample (of our intended $n$)

Great, but where am I ever going to find data where I know exactly what the population parameters are?
Monte Carlo Methods

• Remember the definition of power?
  • The probability of observing a significant effect in our sample if the effect truly exists in the population
  • What if we knew for a fact that the effect existed in a particular population?
  • Then, a measure of power is how often we get a significant result in a sample (of our intended $n$)

• Solution: We create (“simulate”) the data.
Data Simulation

• Set some plausible population parameters
  (effect size, subject variance, item var., etc.)

Set population parameters
  Mean = 723 ms
  Group difference = 100 ms
  Subject var = 30

• Since we are creating the data…
  • We can choose the population parameters
  • We know we exactly what they are
Data Simulation

- Create (“simulate”) a random sample drawn from this population

  Set population parameters
  Mean = 723 ms
  Group difference = 100 ms
  Subject var = 30

  Create a random sample from these data
  $N$ subjects = 20
  $N$ items = 40

- Like most samples, the sample statistics will *not* exactly match the population parameters
  - It’s randomly generated
  - But, the difference is we know what the population is like & that there IS an effect
**Data Simulation**

- Now, fit our planned mixed-effects model to this **sample** of simulated data to get **one** result

  - **Set population parameters**
    - Mean = 723 ms
    - Group difference = 100 ms
    - Subject var = 30

  - **Create a random sample** from these data
    - \( N \) subjects = 20
    - \( N \) items = 40

  - Run our planned model and see if we get a significant result

- **Might get a significant result**
  - Correctly detected the effect in the population

- **Might get a non-significant result**
  - Type II error – missed an effect that **really exists in the population**
Monte Carlo Methods

• If we do this repeatedly, we will get multiple significance tests, each on a different sample

Outcomes:
• Sample 1: $p < .05$ (Yes)
• Sample 2: $p = .23$ (No)
• Sample 3: $p < .05$ (Yes)
• Sample 4: $p = .14$ (No)

• Detected the effect $\frac{1}{2}$ of the time: Power = .50
Monte Carlo Methods

• If we do this repeatedly, we will get multiple significance tests, each on a different sample

  Set population parameters
  Mean = 723 ms
  Group difference = 100 ms
  Subject var = 30

  Create a random sample from these data
  N subjects = 20
  N items = 40

  Run our planned model and see if we get a significant result

  Repeat with a new sample from the same population

• Hmm, that power wasn’t very good 😞
Monte Carlo Methods

• If we do this repeatedly, we will get *multiple* significance tests, each on a different sample

Set population parameters
Mean = 723 ms
Group difference = 100 ms
Subject var = 30

Create a random sample from these data
N subjects = 60
N items = 40

Run our planned model and see if we get a significant result

• Hmm, that power wasn’t very good 😞
• Let’s increase the **number of subjects** and run a new simulation to see what our **power** is like now

Repeat with a new sample from the same population
Monte Carlo Methods

• If we do this repeatedly, we will get *multiple* significance tests, each on a different sample

• Goal: Find the sample size(s) that let you detect the effect at least 80% of the time (or whatever your desired power is)
  • Will 40 subjects in each of 5 schools suffice?
  • What about 50 subjects in each of 10 schools?

Set population parameters
Mean = 723 ms
Group difference = 100 ms
Subject var = 30

Create a random sample from these data
N subjects = 60
N items = 40

Run our planned model and see if we get a significant result
Repeat with a new sample from the same population
Week 12.2: Power

- Statistical Power
  - Intro to Power Analysis
  - Estimating Effect Size
  - Why Do We Care About Power?
  - Determining Power
- Power Simulations in R
  - Influences on Power
- Lab
Power Simulations in R

• We can do these Monte Carlo simulations in R with the `simr` package.

• If we already have a model & dataset (“observed power”):
  - `powerSim(model1, test=fixed('VariableName'), nsim=200)`
  - Tells us our power to detect a significant effect of the fixed effect `VariableName` (e.g., 83%)
  - `nsim` is the number of Monte Carlo simulations. More will always give you a more accurate power analysis, but will take more time to run.
**Power Simulations in R**

- We can do these Monte Carlo simulations in R with the `simr` package.

- If we already have a model & dataset ("observed power"):
  - `powerSim(model1, test=fixed('VariableName'), nsim=200)`
  - Could also do `test=random('RandomEffectName')` to find our power to detect a significant random effect.
  - Or `test=fcompare` to compare to a specific alternative (nested) model.
Power Simulations in R

• What if I don’t have enough power?
  • We may need to add more observations—more subjects, items, classrooms, etc.

• `plot(powerCurve(model1, test=fixed('VariableName'), along='Subject', breaks=c(40,60,80,100)))`
  • Varies the number of Subjects and runs the power-analysis for each sample size
  • Specifically, tests the power with 40, 60, 80, and 100 subjects
  • Where do we hit 80% power?
Power Simulations in R
Power Simulations in R

• What if I don’t have my data yet?
  • A priori power

• Can first create a (simulated) dataset & model
  • makeLmer()}
Week 12.2: Power

- Statistical Power
  - Intro to Power Analysis
  - Estimating Effect Size
  - Why Do We Care About Power?
  - Determining Power
  - Power Simulations in R
  - Influences on Power

- Lab
Influences on Power

- So what makes for a powerful design?
- Things that *increase* power:
  - Larger **effect size estimates** for the fixed effects
    - Bigger things are easier to find
  - Larger **sample size** (at any level)
    - More data = more confidence
    - This one we can control
    - Increasing sample size at a higher level (e.g., subjects rather than time points within subjects) is more effective
  - Variance of **independent variables**
    - Easier to see an effect of income on happiness if people *vary* in their income
    - *Hard* to test effect of “number of fingers on your hand”
    - With a categorical variable, would prefer to have an equal # of observations in each condition—most information
Influences on Power

• So what makes for a powerful design?
• Things that *decrease* power:
  • Larger **variance of random effects**
    • More differences between people (noise) make it harder to see what’s consistent
  • Larger **error variance**
    • Again, more noise = harder to see consistent effects
  • May be able to *reduce* both of these if you can add covariates / control variables
Week 12.2: Power

- Statistical Power
  - Intro to Power Analysis
  - Estimating Effect Size
  - Why Do We Care About Power?
  - Determining Power
  - Power Simulations in R
  - Influences on Power

Lab