S-043/Stat-151
Analysis for Clustered and Longitudinal Data
(Multilevel & Longitudinal Models)

Lecture ST.1
Power!
(And how to obtain it.)

Roadmap for the day

What is power? What makes for good power?

Using math to calculate power.

- ★ Vanilla power (OLS)
- ★ Power and clustered data (with MLM models)

Using simulation to calculate power.

- ★ A worked case study for a treatment impact on a rate of growth (CD4 count for sick children)

What is Power?

What is Power?

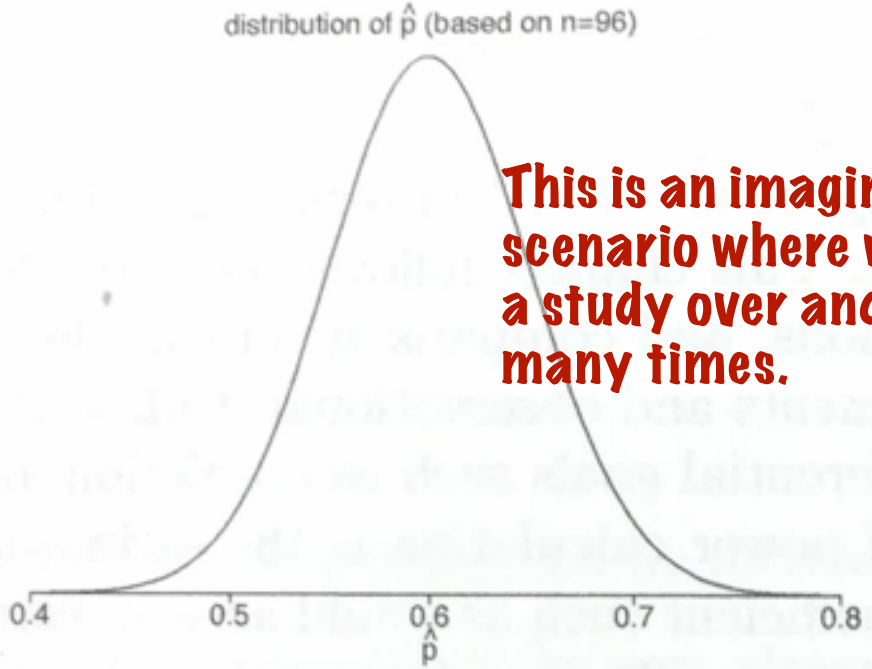
A Degree of Precision

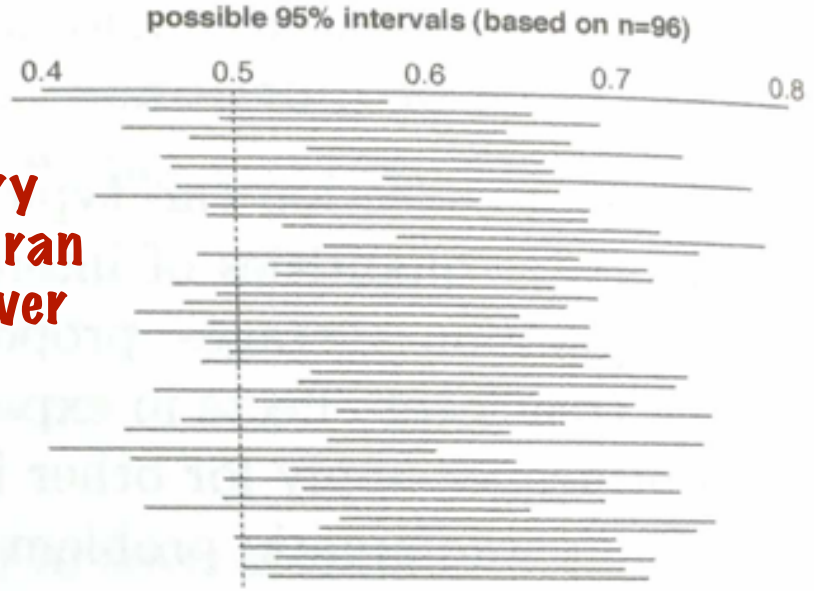
- ★ Measuring a target parameter with a specified level of uncertainty.
- ★ Express this as a *desired Standard Error*

The Probability of Rejection

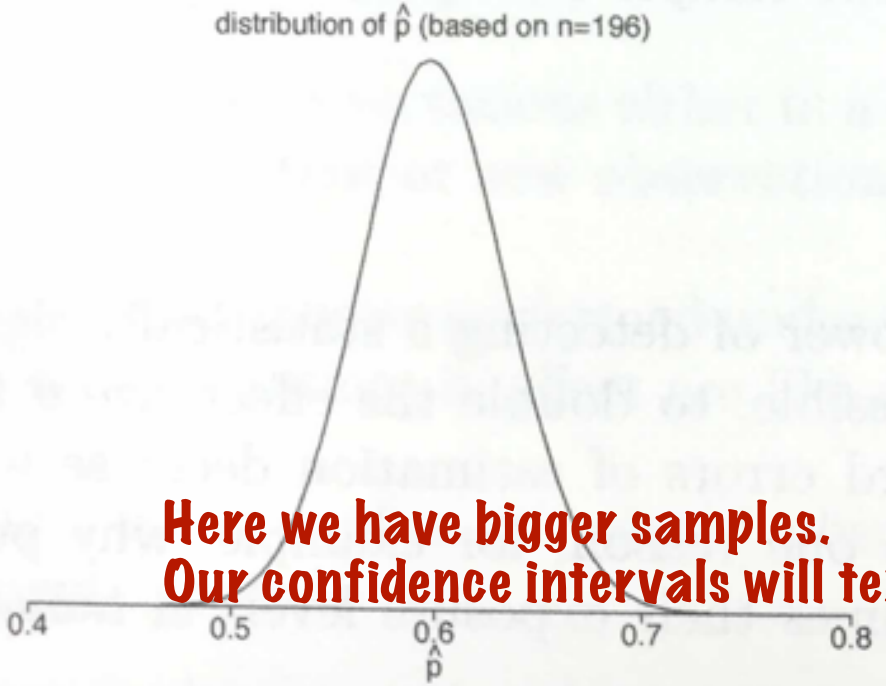
- ★ You are attempting to prove an (alternate) hypothesis.
- ★ You want to know **what the chance of rejecting the null is**, *if you ran a specific study*.
- ★ Alternatively, how big a study do you need to run to have a reasonable chance of rejecting the null?

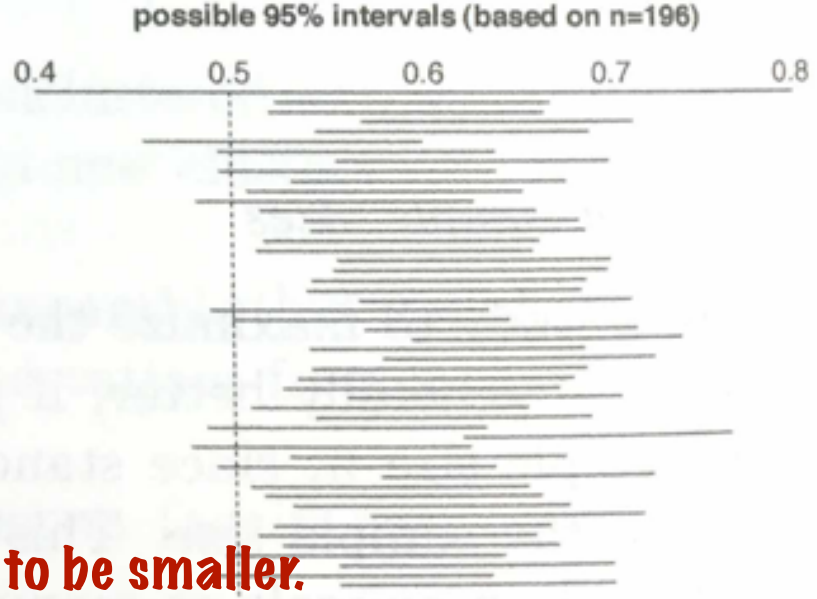
Depending on the distribution of our estimator, our confidence intervals will miss zero at different rates.





This is an imaginary scenario where we ran a study over and over many times.



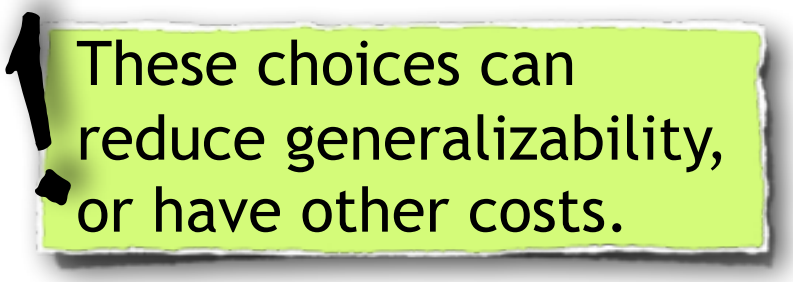


Here we have bigger samples.
Our confidence intervals will tend to be smaller.

How do I get my hands on some power?

General ways of obtaining power

1. **Get more data**
   - More individuals in a cluster
   - More clusters
2. **Find susceptible units**
   - Select units that will likely respond strongly to treatment.
3. **Seek homogeneity**
   - Reduce residual noise by finding units that are similar.
   - Alternatively, use predictive covariates to reduce unexplained variation.
4. **Be extreme**
   - Give as large a dose as you can.
   - Maximize the difference between comparison groups' experience.



Some things are easier than others

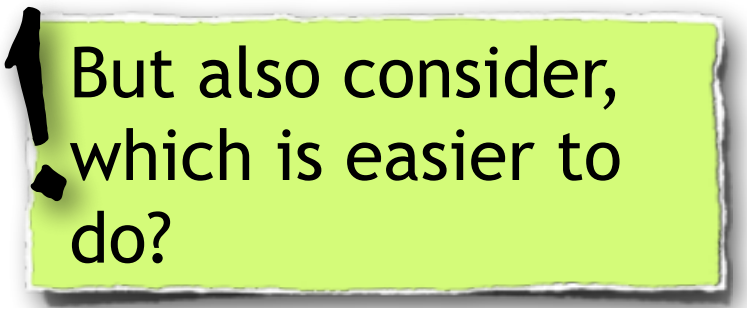
Compare:

doubling your sample size

vs.

doubling your effect size

Which will give you greater power?



A paradox: sample size is never enough

I have more data!
Let's immediately ask
more complex questions.

E.g., subgroup analyses – these examine *subsets of your data* (with smaller sample sizes again)

"Just as you never have enough money, because perceived needs increase with resources, your inferential needs will increase with your sample size."
G&H, pg 438

Power depends on your model of what the real world looks like

Power is a *prospective* question. You say:

"If the world really looks like this,

What is the chance of things working out for me?"

Or:

"If the world really looks like this,

How much work do I have to do to notice?"

To answer it you need to ***define your world***:

- ★ Define what the true data-generation process (DGP) is.
- ★ Define general nuisance aspects such as residual variation, relationship of covariates to outcome, etc.
- ★ Define how big (or small) the thing you are looking for is.

Measuring the scope of what you want

You need to define how big the thing you are interested in is.

Often this is in terms of an *effect size*.

Some examples of effect size measures:

★ Cohen's *d*: $$d \equiv \frac{\mu_1 - \mu_2}{\sigma_{pool}}$$

★ Glass's delta: $$\Delta \equiv \frac{\mu_1 - \mu_2}{\sigma_{co}}$$ ← I like this one!

★ Relative Risk:

$$RR = Pr(Y|G=1) - Pr(Y|G=0)$$

Two Approaches to Power Calculations

Formula and Calculations

- ★ Typically only for simple scenarios (equal cluster size, etc.).
- ★ Give good intuition as to how different decisions & features impact power.

Simulation

- ★ Flexible and general.
- ★ Computationally intensive (slow).
- ★ Easy to do without deep (any?) understanding of mathematical properties.
- ★ Easy to get trapped playing with them!

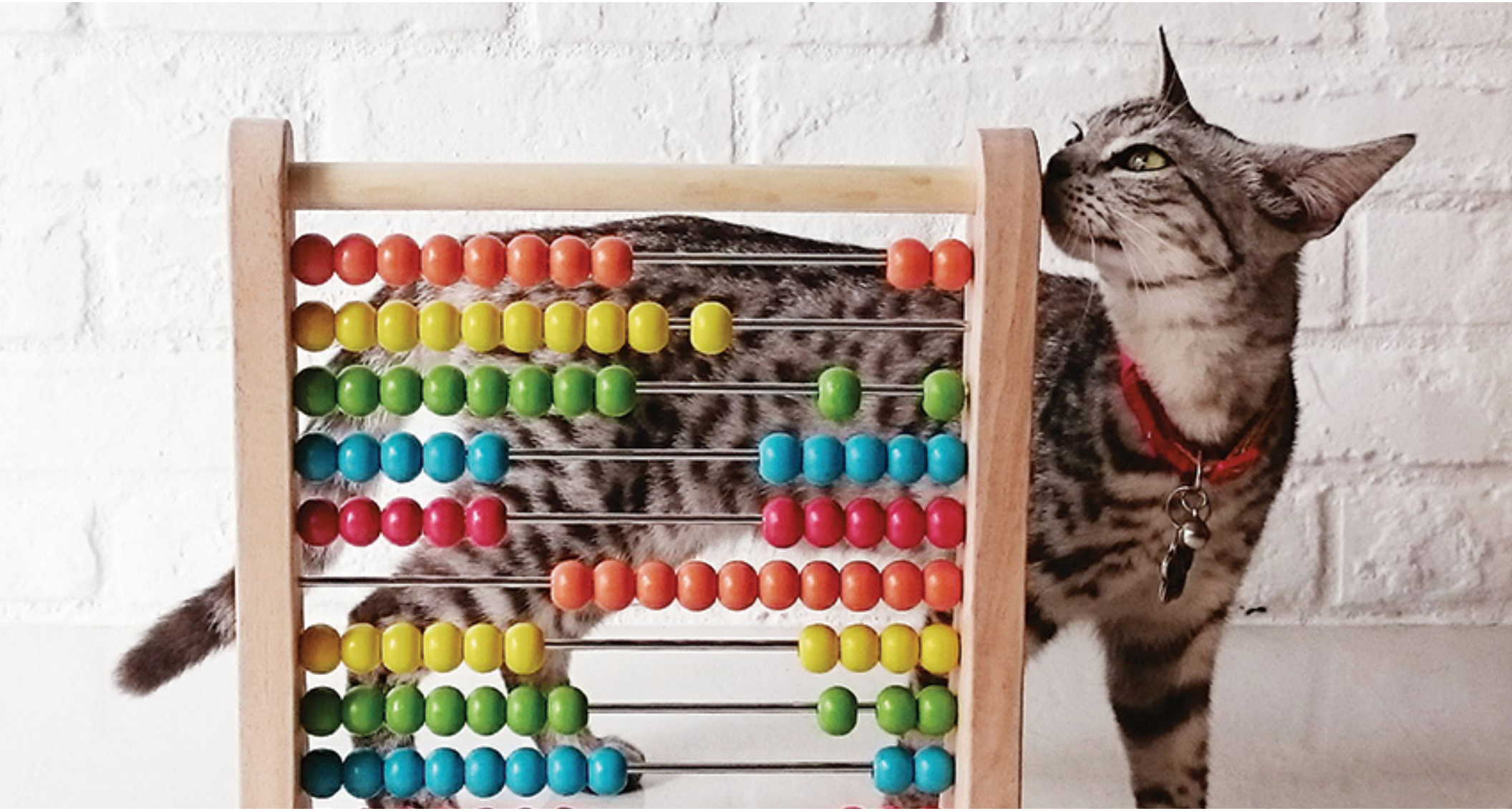
Flip it around: Minimum Detectable Effect Size (MDES)

MDES is, given a specified alternate hypothetical world and study design, the answer to the following question

What is the smallest effect that I have at least an 80% chance of detecting, given my study design.

Formula and calculators exist to give MDES given *a specified model and parameter values*.

A taste of the formula approach (no clustering)



Power to detect a simple Treatment vs. Control difference

Q: What is our experiment?

A: We will sample from two groups that we want to compare.

Q: How will we analyze our data?

A: We will calculate the average of each group and examine the difference of these averages.

Given this:

- ★ What will our standard error be?
- ★ Given our standard error, what sample size do we need, etc.

Simple comparison of means example

$$SE(\bar{y}_1 - \bar{y}_0) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}}$$

Under equal sample sizes (n/2 each side) we have

$$SE = \sqrt{\frac{2(\sigma_1^2 + \sigma_0^2)}{n}}$$

giving $n = \frac{2(\sigma_1^2 + \sigma_0^2)}{SE^2} = \left(\frac{2\sigma}{SE}\right)^2$

If the variances in the groups are assumed equal

We are already seeing lots of simplifying assumptions (equal size groups, same variance)

General power

GOLDEN RULE: To have 80% power you need your true effect to be 2.8 SEs above the null value.

So for a specific Δ ("delta", here a raw effect):

- ★ Calculate Δ/2.8 - this is what your SE needs to be (your needed precision)
- ★ Now calculate *n* based on this required SE.

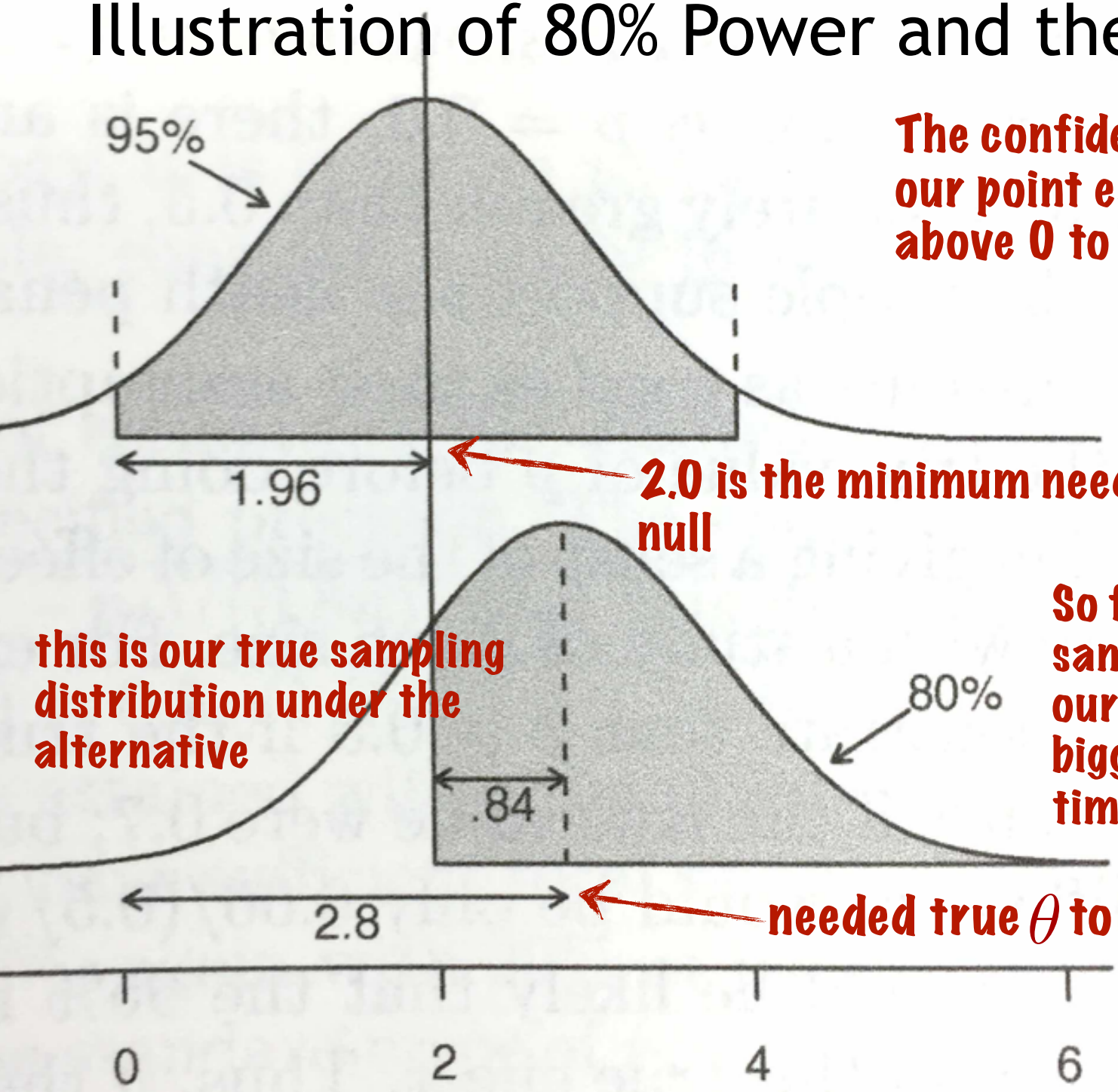
Example:

$$n = \frac{2(\sigma_1^2 + \sigma_0^2)}{\Delta^2/2.8^2} = \left(\frac{5.6\sigma}{\Delta^2}\right)^2$$

For effect size units, σ =1

How big of a sample to detect a 0.10 effect-size effect?

Illustration of 80% Power and the 2.8 SE rule



The confidence interval around our point estimate must be 2 SE above 0 to reject 0.

So for 80% power, the sampling distribution of our test statistic must be bigger than 2.0 80% of the time.

the scale is in standard errors SEs of our test statistic

The formula approach (for clustered designs)

Reading: G&H 20.4

Random Intercept Model

The estimated population mean (overall intercept) will vary as

$$SE(\bar{y}) = \sqrt{\frac{\sigma_y^2}{n} + \frac{\sigma_\alpha^2}{J}}$$

n = *Jm* is total sample size

This assumes

- ★ Equal size clusters randomly drawn from population.
- ★ We have a large super population and are sampling clusters from it, and sampling people from those clusters

Examining uncertainty further

Alternatively we can write

$$SE(\bar{y}) = \sqrt{\frac{\sigma^2_{total}}{Jm}\left(1 + (m-1)ICC\right)}$$

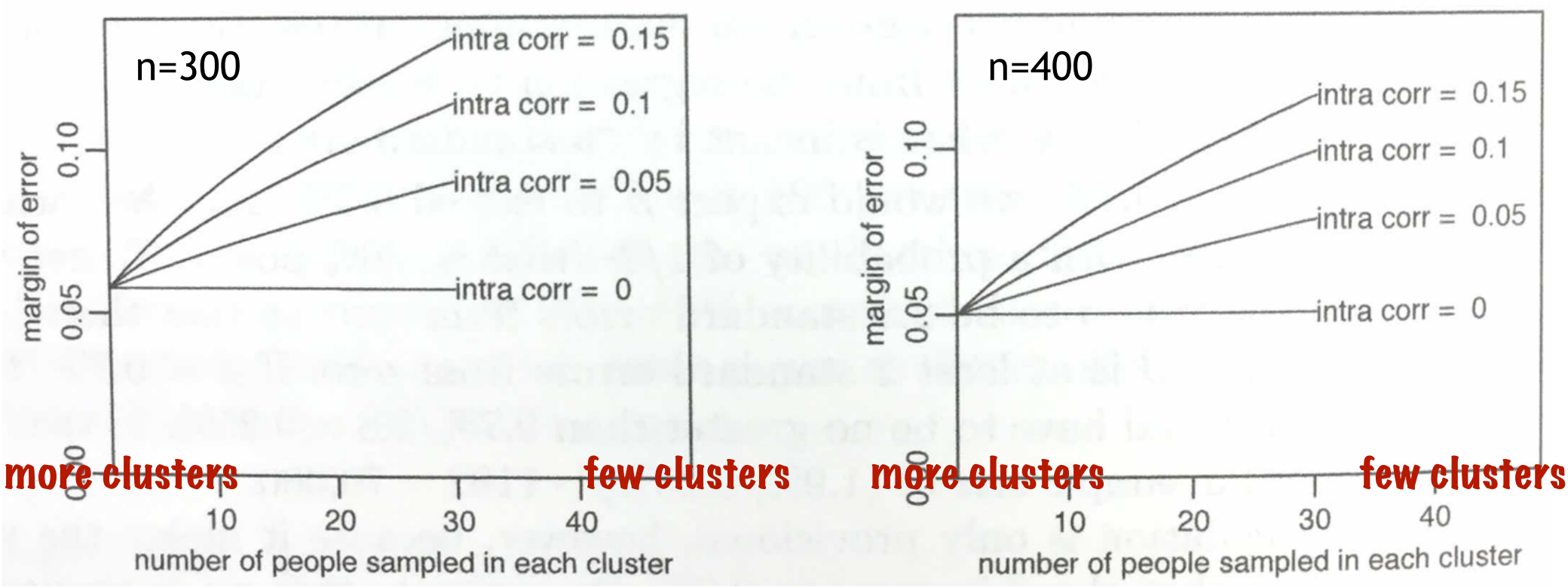
with $\sigma^2_{total} = \sigma^2_y + \sigma^2_\alpha$ and $ICC = \sigma^2_\alpha / \sigma^2_{total}$

Now we can see the trade-off between the impact of the number of clusters and the ICC.

Moral: high ICC means big SEs for a fixed sample size *n* (*n* is *total* number of individuals).

How many clusters to detect a 0.10 effect-size effect if we have 10 people/cluster?

It is more about the clusters than the people



Here we keep total sample size fixed.

We see that more clusters is good, if we have reasonable ICC.

This is because *repeated measures in a cluster tell us less than getting a new cluster.*

Resources to be aware of

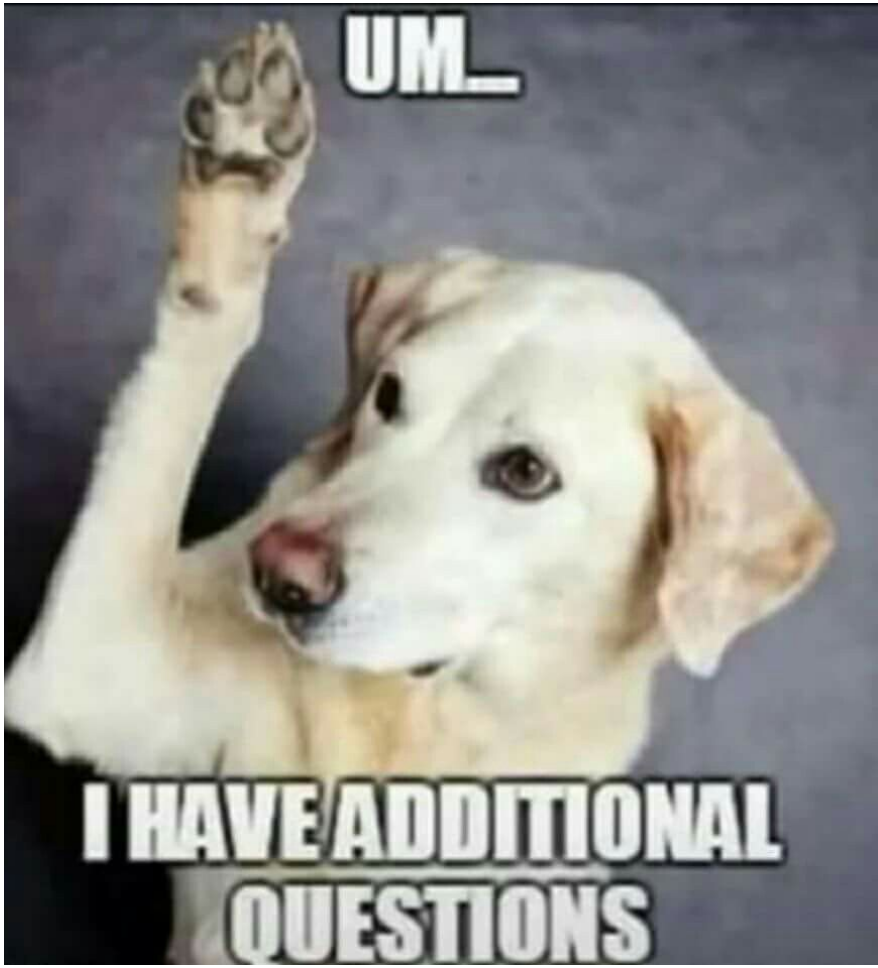
An overview paper

Hedges, Larry V., and Christopher Rhoads. "Statistical Power Analysis in Education Research. NCSER 2010-3006." National Center for Special Education Research (2010).

Power Calculators

These are excel spreadsheets or programs where you type in parameters and it gives you power estimates back for classic designs (cluster randomized experiments, for example)

How do you get your initial parameter estimates?



Prior studies can motivate guesses for parameter values

Rosado et al. (1997), Mexico

Treatment	Sample size	Avg. # episodes in a year ± s.e.
placebo	56	1.1 ± 0.2
iron	54	1.4 ± 0.2
zinc	54	0.7 ± 0.1
zinc + iron	55	0.8 ± 0.1

Ruel et al. (1997), Guatemala

Treatment	Sample size	Avg. # episodes per 100 days [95% c.i.]
placebo	44	8.1 [5.8, 10.2]
zinc	45	6.3 [4.2, 8.9]

Lira et al. (1998), Brazil

Treatment	Sample size	% days with diarrhea	Prevalence ratio [95% c.i.]
placebo	66	5%	1
1 mg zinc	68	5%	1.0 [0.72, 1.4]
5 mg zinc	71	3%	0.68 [0.49, 0.95]

Muller et al. (2001), West Africa

Treatment	Sample size	# days with diarrhea/ total # days
placebo	329	$997/49021 = 0.020$
zinc	332	$869/49086 = 0.018$

G&H Pg 444

Power through simulation

Example: cd4 and children

Imagine developing a medical treatment for children with HIV.

Initial data show HIV+ children's cd4 counts tend to decline at an average rate of -0.5/year.

We are interested in the power to detect a hypothetical treatment that would change this rate to 0 (so an effect of +0.5/year on the slope).

Power requires a stated "state of the world"

Initial data has:

- ★7 measurements across about 1.5 years for 83 kids
- ★We focus on control kids of initial study to get information to motivate our hypothetical future study.
- ★We are thus assuming we are working with similar kids, with the same control condition.

Our model for our hypothetical world

Linear growth model. β_{0i} is initial cd4 of child, β_{1i} is growth rate

$$y_{ti} = \beta_{0i} + \beta_{1i} t_i + \epsilon_{ti}$$

$$\beta_{0i} = \gamma_{00} + u_{0i}$$

No treatment on intercept. Why?

$$\beta_{1i} = \gamma_{10} + \gamma_{11} z_i + u_{1i}$$

$$\begin{pmatrix} u_{0i} \\ u_{1i} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & \tau_{10} \\ & \tau_{11} \end{pmatrix} \right]$$

The z_i are treatment indicators (at the person level)

We will simulate from this world.

A growth curve model on the control kids (fit to pilot data)

```
> M1 <- lmer( y ~ 1 + time + (1 + time | person),
                                        data=dat )
> display(M1)
            coef.est coef.se
(Intercept)  4.85     0.16
time        -0.47     0.13

Error terms:
 Groups   Name        Std.Dev. Corr
 person   (Intercept) 1.33
          time        0.68     0.15
 Residual             0.75
---
number of obs: 369, groups: person, 83
AIC = 1108.1, DIC = 1088
deviance = 1091.9
```

The time coefficient is our rate of decline.
We want our treatment to change this.

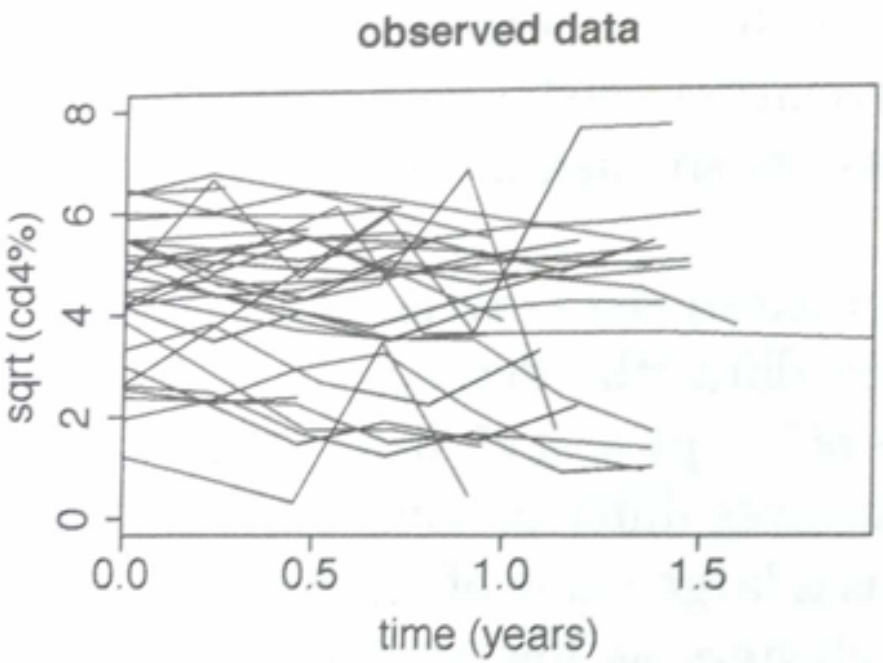
We are ignoring covariates here.

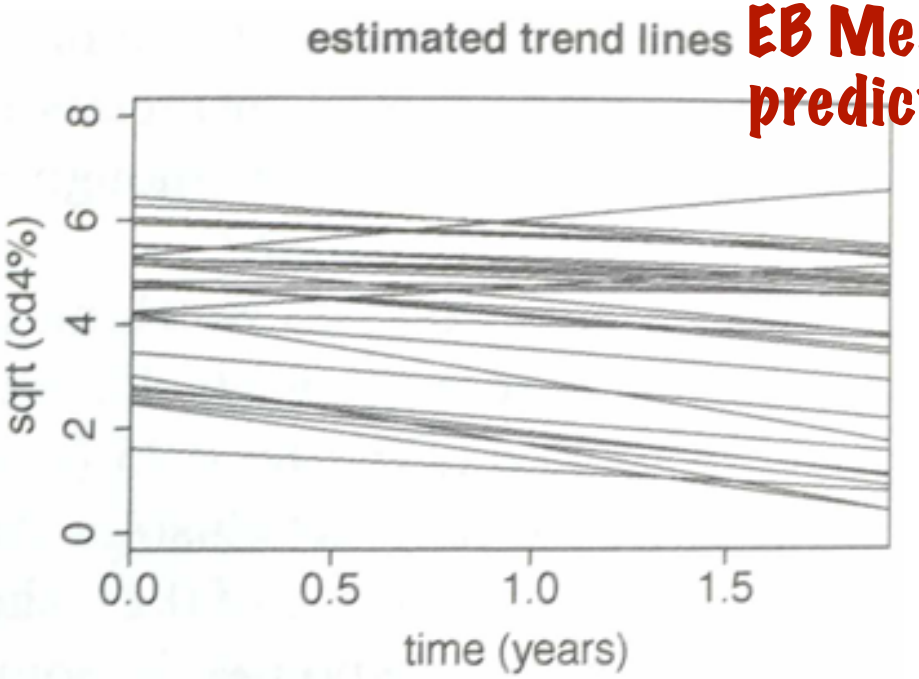
What is an effect size here?

```
> M1 <- lmer( y ~ 1 + time + (1 + time | person),
                                        data=dat )
> display(M1)
            coef.est coef.se
(Intercept)  4.85     0.16
time        -0.47     0.13

Error terms:
 Groups   Name        Std.Dev. Corr
 person   (Intercept) 1.33
          time        0.68     0.15
 Residual             0.75
---
number of obs: 369, groups: person, 83
AIC = 1108.1, DIC = 1088
deviance = 1091.9
```

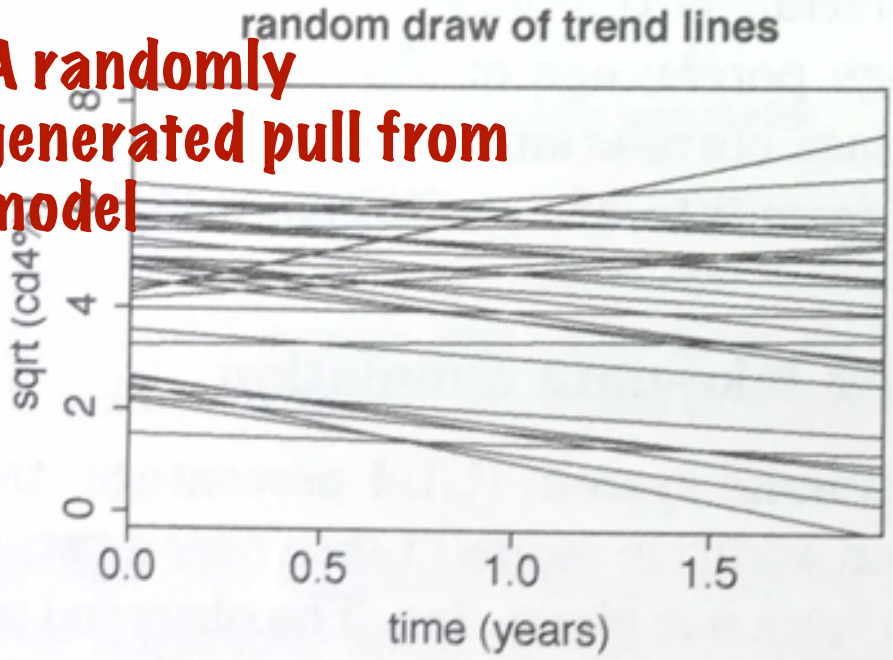
Effect size: not totally clear how to define it. Any thoughts?

A quick model check before we do power.

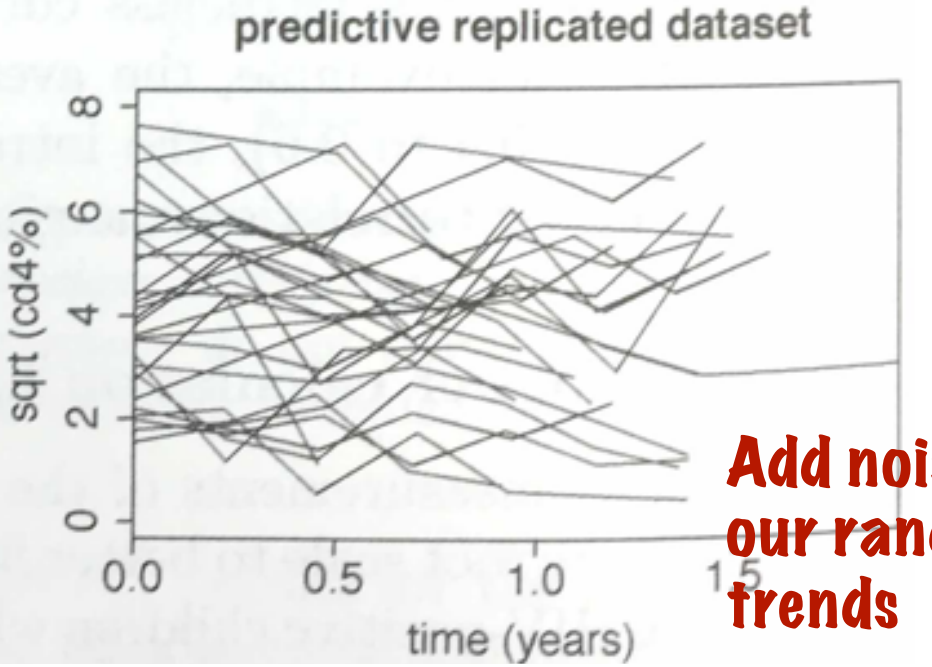




EB Mean predictions



A randomly generated pull from model



Add noise to our random trends

The Simulation Scheme

Repeat 1000 times:

1. Make a fake data set assuming our hypothesized world, experimental/survey design, and our hoped-for treatment effect.

2. Analyze the fake data set and assess:
   a. What our actual error in estimation is (compared to our known truth)
   b. What our estimated standard error is
   c. Whether we rejected the null hypothesis

Speedy Code Warning

We are about to go through a bunch of R code, but not in detail.

Focus on the big picture.

Step 1: Set our parameters

```
# from our model
mu.a <- 4.85
g.0 <- -0.47
sigma.y <- 0.75
sigma.a <- 1.33
sigma.b <- 0.68

g.1 <- 0.47

# number of time points per year
K = 7

# number of people to simulate for
J = 3
```

These come from the model on our initial data that we hope is similar to our future study data.

(We are ignoring our correlation of random intercept and slope)

Our (hoped for) treatment effect is enough to offset the decline we estimated from real data.

We will eventually vary this to see how different J impact power.

from make_data_illustration.R

Step 2: Generate our fake data

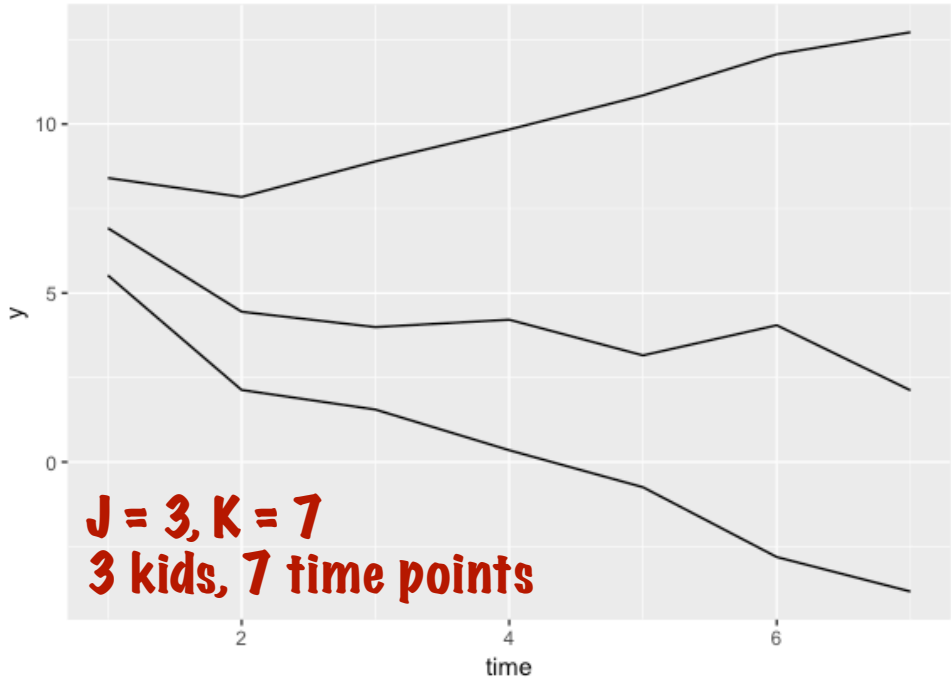
```
# make people
peeps = data.frame( id=1:J,
                    Z = as.numeric( sample(J)<=J/2 ) )

# add person-level random intercept and slope
peeps = mutate( peeps,
                a = rnorm(J, mu.a, sigma.a),
                b = rnorm(J, g.0 + g.1*Z, sigma.b) )

# make time variables
times = data.frame( time = 1:K )
peeps = merge( peeps, times )


## data
peeps = mutate( peeps,
                y.hat =  a + b * time,
                y = y.hat + rnorm( J*K, sd=sigma.y ) )
```

Our random effects are uncorrelated. Could be improved.

Step 3: Bundle our DGP into a handy function

```
make.fake.data <- function(J, K,
                           mu.a, g.0, g.1,
                           sigma.y, sigma.a, sigma.b){

    # all that code from prior page
}
fake = make.fake.data( J, K,
                       mu.a, g.0, g.1,
                       sigma.y, sigma.a, sigma.b )

ggplot( fake, aes( time, y,
                   group=id ) ) +
  geom_line()
```

To make code first write a script that does everything step by step. Then make a function so you can do it all with one single call.



Step 4: Analyze our fake data
This is a single simulation trial

```
> mod <- lmer(y ~ 1 + time + time:treatment1 +
                    (1 + time | person), data=fake)

> g.1.hat <- fixef(mod)["time:treatment1"]
> g.1.hat
[1] 0.043
```

Estimated effect

```
> g.1.se <- se.fixef(mod)["time:treatment1"]
> g.1.se
[1] 1
```

Estimated SE

```
> g.1.hat - g.1
[1] -0.43
```

Actual error

```
> signif <- g.1.hat / g.1.se >= 1.96
> signif
[1] FALSE
```

Reject null?

Step 5: Run lots of simulations!

```
> nsims = 4
```

rerun - run something multiple times

```
> rps = rerun( nsims, runOneSim( J=J, K=K,
+                      mu.a=mu.a, g.0=g.0, g.1=g.1,
+                      sigma.y=sigma.y, sigma.a=sigma.a,
                      sigma.b=sigma.b ) )

> rps
  g.1.hat g.1.se reject
1     1.7   0.88  FALSE
2     2.4   1.02   TRUE
3    -3.3   2.36  FALSE
4     1.0   0.61  FALSE
```

We get a dataframe of simulation results, 1 row per simulation run

Step 6: Bundle our simulation up too.

Print out values to help with error checking

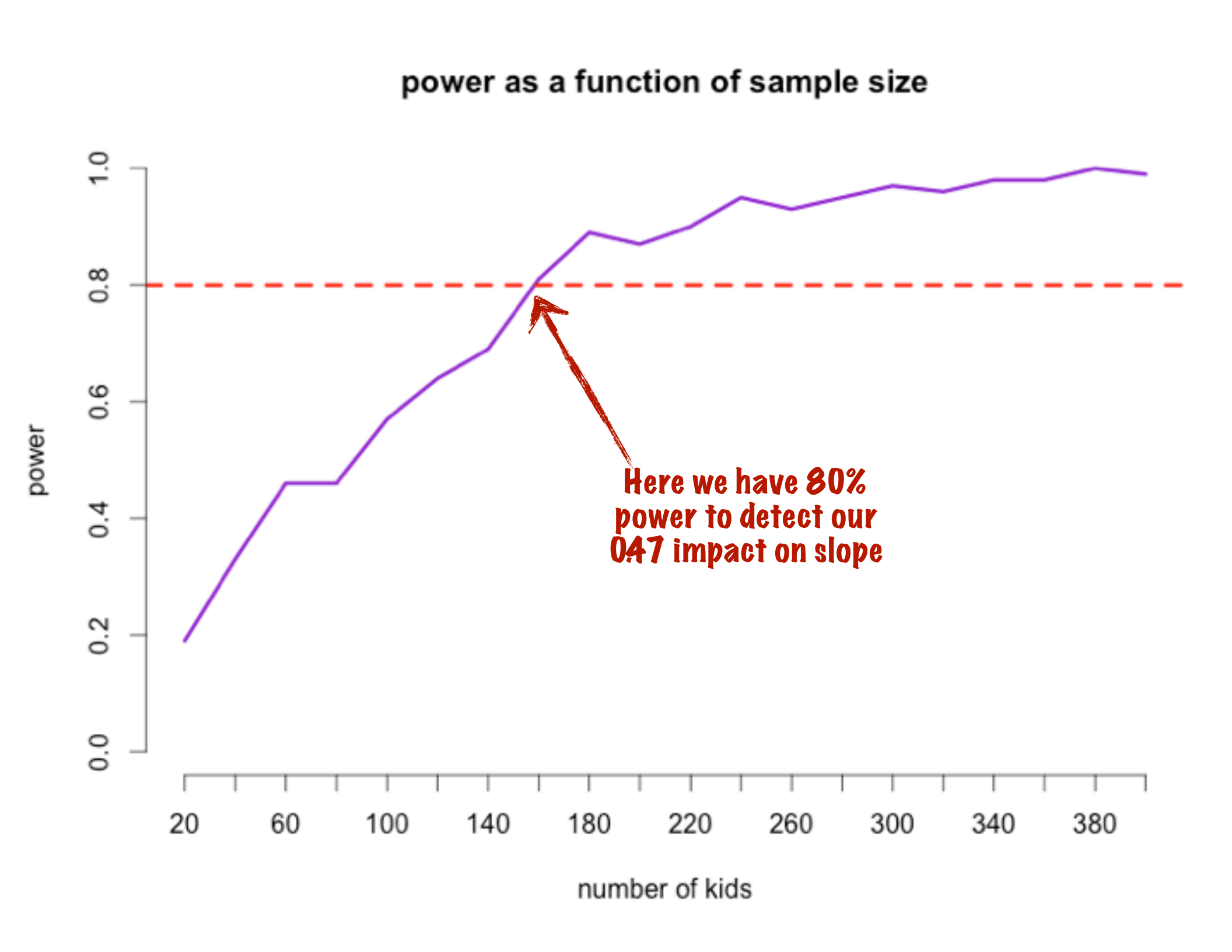
```
cd4power <- function(J, K, nsims=1000,
                     mu.a, g.0, g.1,
                     sigma.y, sigma.a, sigma.b) {
    cat( "Calculating power for J = ",J," K= ",K,"\n" )

    rps = ldply( 1:nsims, runOneSim, J=J, K=K,
                 mu.a=mu.a, g.0=g.0, g.1=g.1,
                 sigma.y=sigma.y, sigma.a=sigma.a,
                 sigma.b=sigma.b )

    power <- mean(rps$reject)
    SE.true = sd( rps$g.1.hat )
    bias = mean( rps$g.1.hat ) - g.1
    mean.SE.hat = mean( rps$g.1.se )

    data.frame( power, SE.true, mean.SE.hat, bias )
}
```

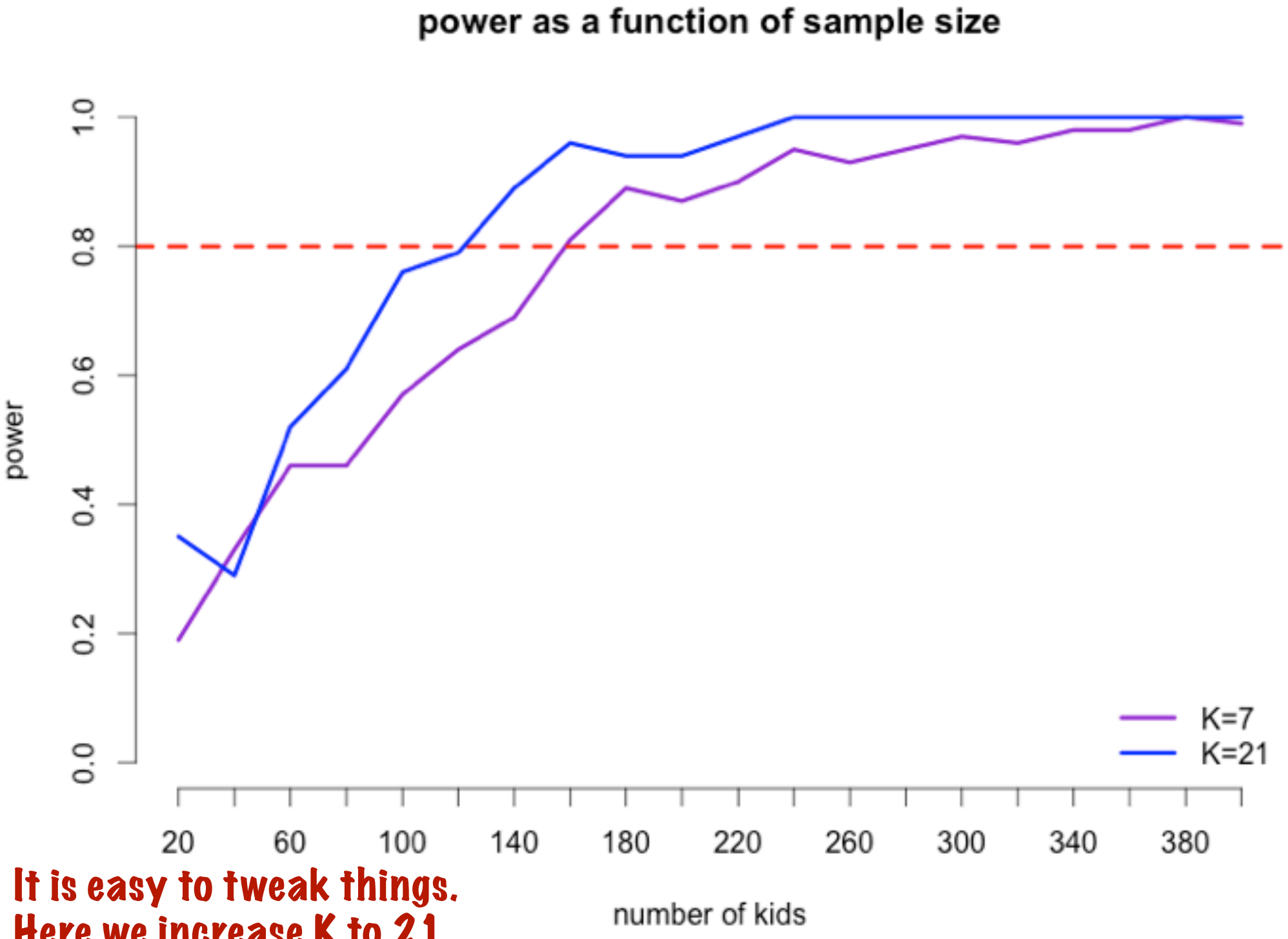
calculate different quantities of interest

Step 7: Rerun our simulation for a range of sample sizes

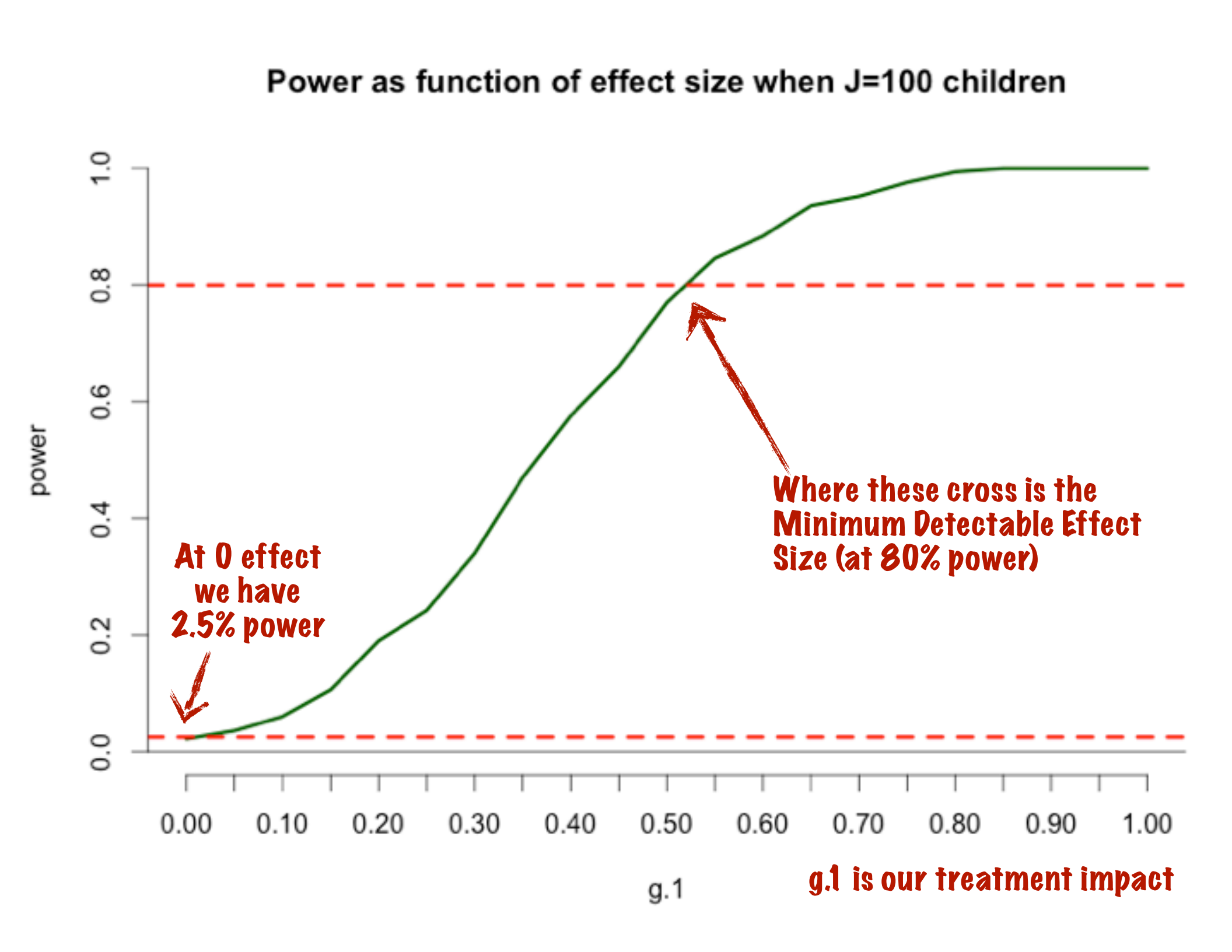
```
> sizes = seq(20, 400, by=20)
> length( sizes ) # we will run 20 simulations
20

> results = map_df( sizes, cd4power, K=7, nsims=100,
                    mu.a=mu.a, g.0=g.0, g.1=g.1,
                    sigma.y=sigma.y, sigma.a=sigma.a,
                    sigma.b=sigma.b)

> results$J = sizes
> head( results )

  power SE.true mean.SE.hat    bias   J
1  0.19    0.49         0.46 -0.0059  20
2  0.33    0.36         0.33  0.0054  40
3  0.46    0.28         0.27  0.0245  60
4  0.46    0.23         0.24 -0.0137  80
5  0.57    0.23         0.21 -0.0180 100
6  0.64    0.17         0.19 -0.0174 120
```

We have a nice table of how different choices of J will impact our precision and power


power as a function of sample size
power
0.0
0.2
0.4
0.6
0.8
1.0
20
60
100
140
180
220
260
300
340
380
number of kids
Here we have 80%
power to detect our
047 impact on slope



It is easy to tweak things.
Here we increase K to 21


Power as function of effect size when J=100 children
power
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
g.1
At 0 effect
we have
2.5% power
Where these cross is the
Minimum Detectable Effect
Size (at 80% power)
g.1 is our treatment impact

Recap

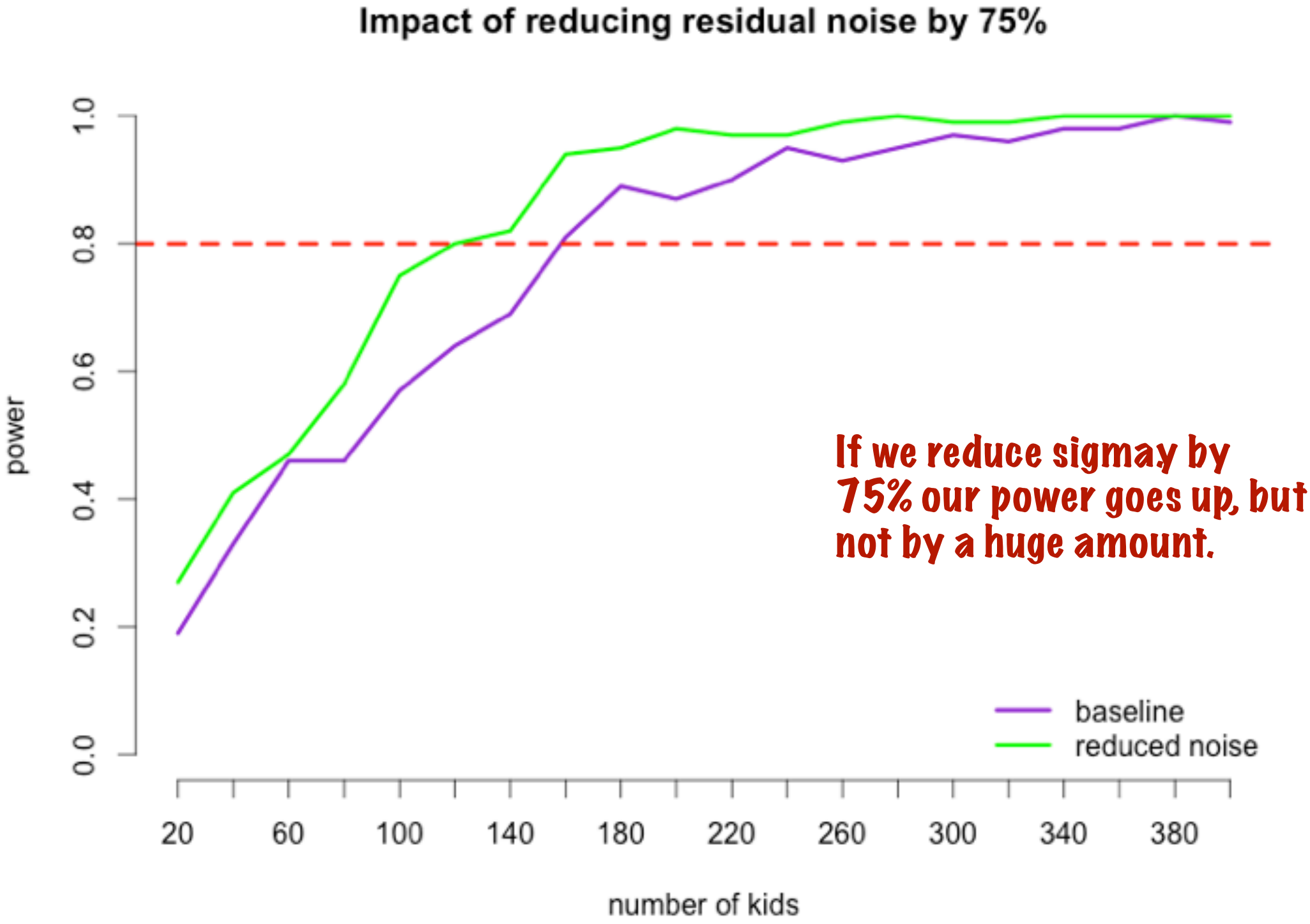
Take-aways

Check In:
http://cs179.org/lecST1

- ★ Understanding power is understanding where information comes from.
- ★ Power is always about how a specified analytic approach will work when a specified model of the world is true.
- ★ Formula and simulations are two ways of examining power.
- ★ Once you have a simulation framework, it is easy to play around and see what makes a difference.

Appendix

An extra power curve example



An extra power curve example