Statistical Modeling, Causal Inference, and Social Science

You need 16 times the sample size to estimate an interaction than to estimate a main effect

Posted on March 15, 2018 9:11 AM by Andrew

Yesterday I shared the following exam question:

In causal inference, it is often important to study varying treatment effects: for example, a treatment could be more effective for men than for women, or for healthy than for unhealthy patients. Suppose a study is designed to have 80% power to detect a main effect at a 95% confidence level. Further suppose that interactions of interest are half the size of main effects. What is its power for detecting an interaction, comparing men to women (say) in a study that is half men and half women? Suppose 1000 studies of this size are performed. How many of the studies would you expect to report a statistically significant interaction? Of these, what is the expectation of the ratio of estimated effect size to actual effect size?

Here's the solution:

If you have 80% power, then the underlying effect size for the main effect is 2.8 standard errors from zero. That is, the z-score has a mean of 2.8 and standard deviation of 1, and there's an 80% chance that the z-score exceeds 1.96 (in R, pnorm(2.8, 1.96, 1) = 0.8).

Now to the interaction. The standard of an interaction is roughly twice the standard error of the main effect, as we can see from some simple algebra:
– The estimate of the main effect is ybar_1 – ybar_2, which has standard error $\sqrt{\sigma^2/(N/2) + \sigma^2/(N/2)} = 2\sigma/\sqrt{N}$; for simplicity I'm assuming a constant variance within groups, which will typically be a good approximation for binary data, for example.
– The estimate of the interaction is (ybar_1 – ybar_2) – (ybar_3 – ybar_4), which has standard error $\sqrt{\sigma^2/(N/4) + \sigma^2/(N/4) + \sigma^2/(N/4) + \sigma^2/(N/4)} = 4\sigma/\sqrt{N}$. [algebra fixed]

And, from the statement of the problem, we've assumed the interaction is half the size of the main effect. So if the main effect is 2.8 on some scale with a se of 1, then the

interaction is 1.4 with an se of 2, thus the z-score of the interaction has a mean of 0.7 and a sd of 1, and the probability of seeing a statistically significant effect difference is pnorm(0.7, 1.96, 1) = 0.10. That's right: if you have 80% power to estimate the main effect, you have 10% power to estimate the interaction.

And 10% power is really bad. It's worse than it looks. 10% power kinda looks like it might be OK; after all, it still represents a 10% chance of a win. But that's not right at all: if you do get "statistical significance" in that case, your estimate is a huge overestimate:

```
> raw <- rnorm(1e6, .7, 1)
> significant <- raw > 1.96
> mean(raw[significant])
[1] 2.4
```

So, the 10% of results which do appear to be statistically significant give an estimate of 2.4, on average, which is over 3 times higher than the true effect.

Take-home point

The most important point here, though, has nothing to do with statistical significance. It's just this: Based on some reasonable assumptions regarding main effects and interactions, *you need 16 times the sample size to estimate an interaction than to estimate a main effect*.

And this implies a major, major problem with the usual plan of designing a study with a focus on the main effect, maybe even preregistering, and then looking to see what shows up in the interactions. Or, even worse, designing a study, not finding the anticipated main effect, and then using the interactions to bail you out. The problem is not just that this sort of analysis is "exploratory"; it's that these data are a lot noisier than you realize, so what you think of as interesting exploratory findings could be just a bunch of noise.

I don't know if all this in the textbooks, but it should be.

Some regression simulations in R

In response to a comment I did some simulations which I thought were worth including in the main post.

I played around in R to get a sense of how the standard errors depend on the parameterization.

For simplicity, all my simulations assume that the true (underlying) coefficients are 0; that's no big deal, as my point here is to work out the standard error.

I started with the basic model in which I simulate 1000 data points with two predictors, each taking on the value (-0.5, 0.5). This is the same as the model in the above post: the estimated main effects are simple differences, and the estimated interaction is a difference in differences. I've also assumed the two predictors are independent, which is how I'd envisioned the problem:

```
library("arm")
N <- 1000
sigma <- 10
y <- rnorm(N, 0, sigma)
x1 <- sample(c(-0.5,0.5), N, replace=TRUE)
x2 <- sample(c(-0.5,0.5), N, replace=TRUE)
display(lm(y ~ x1))
display(lm(y ~ x1 + x2 + x1:x2))
```

And here's the result:

```
lm(formula = y ~ x1 + x2 + x1:x2)
            coef.est coef.se
(Intercept) -0.09    0.32
x1          -0.41    0.63
x2          -0.13    0.63
x1:x2       -0.91    1.26
```

Ignore the estimates; they're pure noise. Just look at the standard errors. They go just as in the above formulas: 2*sigma/sqrt(N) = 2*10/sqrt(1000) = 0.63, and 4*sigma/sqrt(N) = 1.26.

Now let's do the exact same thing but make the predictors take on the value (0, 1) rather than (-0.5, 0.5):

```
x1 <- sample(c(0,1), N, replace=TRUE)
x2 <- sample(c(0,1), N, replace=TRUE)
display(lm(y ~ x1))
display(lm(y ~ x1 + x2 + x1:x2))
```

Here's what we see:

```
lm(formula = y ~ x1 + x2 + x1:x2)
            coef.est coef.se
(Intercept) -0.44    0.64
x1           0.03    0.89
x2           0.95    0.89
x1:x2       -0.54    1.26
```

Again, just look at the standard errors. The s.e. for the interaction is still 1.26, but the standard errors for the main effects went up to 0.89. What happened?

What happened was that the main effects are now estimated at the edge of the data: the estimated coefficient of x1 is now the difference in y, comparing the two values of x1, just at x2=0. So its s.e. is sqrt(sigma^2/(N/4) + sigma^2/(N/4)) = 2*sqrt(2)*sigma/sqrt(N). Under this parameterization, the coefficient of x1 is estimated just from the data given x2=0, only half the data so the s.e. is sqrt(2) times as big as before. Similarly for x2.

But these aren't really "main effects"; in the context of the above problem, the main effect of the treatment is the average over men and women, hence if we put the problem in a regression framework, we should be coding the predictors as (-0.5, 0.5), not (0, 1).

But here's another possibility: what about coding each predictor as (-1, 1)? What happens then? Let's take a look:

```
x1 <- sample(c(-1,1), N, replace=TRUE)
x2 <- sample(c(-1,1), N, replace=TRUE)
display(lm(y ~ x1))
display(lm(y ~ x1 + x2 + x1:x2))
```

This yields:

	coef.est	coef.se
(Intercept)	-0.23	0.31
x1	0.28	0.31
x2	-0.60	0.31
x1:x2	0.05	0.31

(Again, ignore the coefficient estimates and look at the standard errors.)

Hey---the coefficients are all smaller (by a factor of 2), and they're all equal! What happened?

The factor of 2 is clear enough: If you multiply x by 2, and x*beta doesn't change, then you have to divide beta by 2 to compensate, and its standard error gets divided by 2 as well. But what happened to the interaction? Well, that's clear too: we've multiplied x1 and x2 each by 2, so x1*x2 is multiplied by 4.

So to make sense of all these standard errors, you have to have a feel for the appropriate scale for the coefficients. In my post above, when I talked about interactions being half the size of main effects, I was thinking of differences between the two groups, which corresponds to parameterizing the predictors as (-0.5, 0.5).

As you can see from the above simulations, the exact answer will depend on your modeling assumptions.

It all depends on what you mean by "interaction half the size of the main effect"

This came up in comments.

What do I mean by "interactions of interest are half the size of main effects"? Suppose the main effect of the treatment is, say, 0.6 and the interaction with sex is 0.3, then the treatment effect is 0.45 for women and 0.75 for men. That's what I meant.

If, by "interactions of interest are half the size of main effects," you have in mind something like a main effect of 0.6 that is 0.3 for women and 0.9 for men, then you'll only need 4x the sample size, not 16x.

P.S. Further discussion here. Thanks, everyone, for all the thoughtful comments.

This entry was posted in Miscellaneous Statistics by Andrew. Bookmark the permalink [https://statmodeling.stat.columbia.edu/2018/03/15/need-16-times-sample-size-estimate-interaction-estimate-main-effect/] .

---

68 THOUGHTS ON "YOU NEED 16 TIMES THE SAMPLE SIZE TO ESTIMATE AN INTERACTION THAN TO ESTIMATE A MAIN EFFECT"

> ### Bob Kubinec
> on **March 15, 2018 9:48 AM at 9:48 am** said:
>
> I think this is also why it's really helpful to simulate these values instead of relying on one's ability to construct the correct formula. The R package randomizr has some helpful tools in that regard.
>
> > **Andrew** on **March 15, 2018 9:53 AM at 9:53 am** said:
> >
> > Bob:
> >
> > Yes, it's a good idea to do simulation and also work through the formula. The simulation is more trustworthy and more flexible; the formula is useful to help understand what's going on.
> >
> > > ### Keith O'Rourke on **March 15, 2018 10:11 AM at 10:11 am** said:
> > >
> > > Also simulation is a good check on whether you mixed up changing n versus sd ;-)

Perhaps more importantly you can simulate more realistic assumptions than fit nicely into simple formulas.

Which is more useful to help understand what's going on, I think, depends on the individual.

expr
on **March 15, 2018 10:05 AM at 10:05 am** said:

Nit pick:
sqrt(sigma^2/(N/2) + sigma^2/(N/2)) = sqrt(2*sigma^2/N + 2*sigma^2/N) = sqrt(4*sigma^2/N) = 2*sigma/sqrt(N) ?

sqrt(sigma^2/(N/4) + sigma^2/(N/4) + sigma^2/(N/4) + sigma^2/(N/4)) = sqrt(4*sigma^2/N + 4*sigma^2/N + 4*sigma^2/N + 4*sigma^2/N) = sqrt(16*sigma^2/N) = 4*sigma/sqrt(N) ?

Andrew on **March 15, 2018 10:12 AM at 10:12 am** said:

Thanks; algebra fixed. I'd remembered the factors of 1 and 2 because with binary outcomes, sigma=0.5, so 2*sigma=1 and 4*sigma=2.

sonata
on **March 15, 2018 10:14 AM at 10:14 am** said:

I have difficulties to follow the take-home message of this post.

(1) Is it reasonable to expect that interactions are always ~half the size of main effects? Because this seems to be the critical point.

(2) I have even more difficulties when I start thinking about ANOVA as a linear model of the data. I could imagine main effects as predictors, and interactions as their multiplications. At this point, I have the impression that it would be safe to think of these predictors as orthogonal, and hence wondering why should the estimation for these predictors should be more difficult for interaction predictors in comparison to main effect predictors?

Andrew on **March 15, 2018 10:21 AM at 10:21 am** said:

Sonata:

1. I think that considering interactions as half the size of main effects is a useful starting point. The analysis is there, so you can alter it as needed.

2. I agree that when your model changes, your standard error can change, so the conclusions will depend on context.

So, yes, my factor of 16 is just for one particular set of assumptions.

> **Scott Glover** on **March 15, 2018 5:05 PM at 5:05 pm** said:
>
> Hi Andrew,
>
> What do you base 1) above on though? I've asked people and they don't seem to think there's any a priori reason to assume an interaction will be half the size of a main effect.
>
> Otherwise your claim seems to reduce to smaller effects = less power, which isn't surprising.
>
> > **Andrew** on **March 15, 2018 5:24 PM at 5:24 pm** said:
> >
> > Scott:
> >
> > It's not *just* that smaller effects have less power. That's one of the factors of 4. The other factor of 4 is that the estimate of the interaction is a difference in differences, which will have twice the standard error (thus 4 times the variance) of the estimate of the main effect.
> >
> > Regarding your question about a priori reasons: I guess it depends on the application. It makes sense to me that large interactions could be as large as the main effect or even larger (for example, if the main effect is 0.6 and the effect is 0 for women and 1.2 for men, then the interaction is twice the size of the main effect!); I'd guess that *typical* interactions are of the order of half the size of the main effect. A lot depends on whether you're going in looking for one particular interaction, or if you're rooting around looking for what interactions might turn up.
> >
> > **Scott Glover** on **March 15, 2018 6:29 PM at 6:29 pm** said:
> >
> > Thanks for your reply Andrew.
> >
> > $(X11 - X12) - (X21 - X22) = (X11 + X22) - (X12 + X22)$

If you code your interaction this way, it'll be twice the size and have twice the SE as a main effect (coded as the difference between the marginal means). If you divide it by 2 it'll be exactly the same. So, (X11 + X22)/2 − (X12 + X22)/2 looks a lot like (X11 + X12)/2 − (X21 + X22)/2

So, maybe if you start off by assuming that the difference of differences (interaction) should be the same size as the difference in marginal means (main effect), you can arrive at the conclusion that the SE should be twice as large. But that's just another way of saying you expect the interaction to be small.

Andrew on **March 15, 2018 10:57 AM at 10:57 am** said:

Sonata:

I played around in R to get a sense of how the standard errors depend on the parameterization.

For simplicity, all my simulations assume that the true (underlying) coefficients are 0; that's no big deal, as my point here is to work out the standard error.

I started with the basic model in which I simulate 1000 data points with two predictors, each taking on the value (-0.5, 0.5). This is the same as the model in the above post: the estimated main effects are simple differences, and the estimated interaction is a difference in differences. I've also assumed the two predictors are independent, which is how I'd envisioned the problem:

```
library("arm")
N <- 1000
sigma <- 10
y <- rnorm(N, 0, sigma)
x1 <- sample(c(-0.5,0.5), N, replace=TRUE)
x2 <- sample(c(-0.5,0.5), N, replace=TRUE)
display(lm(y ~ x1))
display(lm(y ~ x1 + x2 + x1:x2))
```

And here's the result:

```
lm(formula = y ~ x1 + x2 + x1:x2)
            coef.est coef.se
(Intercept) -0.09    0.32
x1          -0.41    0.63
```

```
x2              -0.13      0.63
x1:x2           -0.91      1.26
```

Ignore the estimates; they're pure noise. Just look at the standard errors. They go just as in the above formulas: 2*sigma/sqrt(N) = 2*10/sqrt(1000) = 0.63, and 4*sigma/sqrt(N) = 1.26.

Now let's do the exact same thing but make the predictors take on the value (0, 1) rather than (-0.5, 0.5):

```
x1 <- sample(c(0,1), N, replace=TRUE)
x2 <- sample(c(0,1), N, replace=TRUE)
display(lm(y ~ x1))
display(lm(y ~ x1 + x2 + x1:x2))
```

Here's what we see:

```
lm(formula = y ~ x1 + x2 + x1:x2)
              coef.est coef.se
(Intercept)  -0.44      0.64
x1            0.03      0.89
x2            0.95      0.89
x1:x2        -0.54      1.26
```

Again, just look at the standard errors. The s.e. for the interaction is still 1.26, but the standard errors for the main effects went up to 0.89. What happened?

What happened was that the main effects are now estimated at the edge of the data: the estimated coefficient of x1 is now the difference in y, comparing the two values of x1, just at x2=0. So its s.e. is sqrt(sigma^2/(N/4) + sigma^2/(N/4)) = 2*sqrt(2)*sigma/sqrt(N). Under this parameterization, the coefficient of x1 is estimated just from the data given x2=0, only half the data so the s.e. is sqrt(2) times as big as before. Similarly for x2.

But these aren't really "main effects"; in the context of the above problem, the main effect of the treatment is the average over men and women, hence if we put the problem in a regression framework, we should be coding the predictors as (-0.5, 0.5), not (0, 1).

But here's another possibility: what about coding each predictor as (-1, 1)? What happens then? Let's take a look:

```
x1 <- sample(c(-1,1), N, replace=TRUE)
x2 <- sample(c(-1,1), N, replace=TRUE)
display(lm(y ~ x1))
display(lm(y ~ x1 + x2 + x1:x2))
```

This yields:

```
              coef.est coef.se
(Intercept) -0.23      0.31
x1           0.28      0.31
x2          -0.60      0.31
x1:x2        0.05      0.31
```

(Again, ignore the coefficient estimates and look at the standard errors.)

Hey---the coefficients are all smaller (by a factor of 2), and they're all equal! What happened?

The factor of 2 is clear enough: If you multiply x by 2, and x*beta doesn't change, then you have to divide beta by 2 to compensate, and its standard error gets divided by 2 as well. But what happened to the interaction? Well, that's clear too: we've multiplied x1 and x2 each by 2, so x1*x2 is multiplied by 4.

So to make sense of all these standard errors, you have to have a feel for the appropriate scale for the coefficients. In my post above, when I talked about interactions being half the size of main effects, I was thinking of differences between the two groups, which corresponds to parameterizing the predictors as (-0.5, 0.5).

As you can see from the above simulations, the exact answer will depend on your modeling assumptions.

> **sonata** on **March 15, 2018 3:27 PM at 3:27 pm** said:
>
> Thx!
> The [-1 1] case seems most reasonable given that ANOVA computes differences. Linear modeling examples you present above show that SE depends (partly) on the relative norm of the predictor vectors. Using a norm of 1 everywhere leads to both the main effect and interaction predictors to also be of unit length, and thus all SEs are the same. Makes all numbers more comparable…
>
> > **Andrew** on **March 15, 2018 4:06 PM at 4:06 pm** said:
> >
> > Sonata:
> >
> > Maybe it will help to consider specific numbers. Suppose the main effect of the treatment is, say, 0.6 and the interaction with sex is 0.3, so that the treatment effect is 0.45 for women and 0.75 for men. Then if you do the regression with the predictors on the (-0.5, 0.5) scale, you'll get estimates of 0.6 for the treatment effect and 0.3 for the interaction. If you do the regression with the predictors on the (-1, 1) scale, you'll get estimates of 0.3 for the treatment effect and 0.075 for the interaction.
> >
> > Depending on the context, maybe I'm being too pessimistic in the above post. One could easily imagine a treatment where the main effect is 0.6, with an effect of 0.3 for women and 0.9 for men. In that case, the interac-

tion is as large as the main effect, so then you only need 4 times the sample size, not 16.

Sebastian H on **March 15, 2018 11:30 AM at 11:30 am** said:

Andrew just means it as a discussion point, but I would tend to expect most interesting interactions to be much less than half the size of main effects. For example a common hypothesis is that one's sex will effect one's opinion on abortion. But actually religiosity has at least four or five times the explanatory power of sex interactions on that data (and the wrong way for most explanatory hypotheticals, women are slightly more pro life than men). Now as it turns out we've done enormous polls on the subject so we can still tease out smaller interactions but I would tend to think of a power effect of 1/2 as being an overestimate for discussion not an underestimate.

Anoneuoid
on **March 15, 2018 10:54 AM at 10:54 am** said:

Based on some reasonable assumptions regarding main effects and interactions, you need 16 times the sample size to estimate an interaction than to estimate a main effect.

Aren't you assuming zero intercept and zero sex/gender main effect in this model?

Andrew on **March 15, 2018 10:59 AM at 10:59 am** said:

Anon:

No. The values of the intercept and the "main effect" for sex won't matter here because, by construction, the four predictors in the model are statistically independent (that is, we're assuming equal numbers of men and women, or approximately equal numbers of men and women, in the two treatment groups).

Anoneuoid on **March 15, 2018 11:22 AM at 11:22 am** said:

I made a post in the other thread (not showing up yet) that fixed a bug and made some modifications to my sim (thanks Arno). Actually, none of these variants of lm duplicated your result but leaving out the sex main effect was necessary:

lm(Val ~ 0 + Group*Sex, dat)
lm(Val ~ Group*Sex, dat)

lm(Val ~ 0 + Group + Group:Sex, dat)
lm(Val ~ Group + Group:Sex, dat)

https://image.ibb.co/mjUz5H/power_Sim2.png

What would the model you are using look like in "lm format"?

> **Anoneuoid** on **March 15, 2018 11:29 AM at 11:29 am** said:
>
> Oops. The labels in the title for intercept are reversed there. Read it as "intercept dropped" but "sex main effect included". Sorry about that.

> **Arno** on **March 17, 2018 3:46 AM at 3:46 am** said:
>
> I looked into your simulation a bit more: https://pastebin.com/S7KU7vfu
>
> As you can see, using Group*Sex as factors results in slightly different values and only 70% power for the main effect (just as Howard wrote below). The SE of the interaction is also only sqrt(2) larger than SE of main (see Andrew's reasoning above). Applied to your graph, this may account for the discrepancy between your "analytic results" (dashed lines) and the actual results.
>
> If instead you encode the groups as (-0.5, 0.5) you get 80% power (sim2()). Now the z.coeffs are ~2.8 and ~0.7 (just as in Andrew's calculation) and inflation of significant interaction effects is ~3x.
>
> Which leads me to the question: why is dummy coding used in all regression packages, if it effectively diminishes power (from 80% to 70%)? And how would this look like for factors with more than two groups?

> ### David Manheim
> on **March 15, 2018 11:39 AM at 11:39 am** said:

Andrew – as a take-away for interpreting and conducting studies, would you say:

1) This result implies that while you need to include interactions as a way to improve your estimate of the main effect, those interaction estimates should be (mostly) ignored? (Or should we avoid including interaction terms if we don't have a sufficient sample to estimate them accurately?)

and/or

2) The estimates for the interactions themselves need to be a focus of a preregistered trial, and power should be computed on that basis, for those results to be worth taking seriously? (I assume we cannot say that if there is a large sample, we can trust the interaction effects estimates – because of garden of forking paths issues?)

> **Andrew** on **March 15, 2018 11:49 AM at 11:49 am** said:
>
> David:
>
> I'm not sure what my advice is. In the example above, it's not necessary to estimate inter-actions at all; you can just estimate the main effect averaging over the data. More general-ly, I think that in settings where interactions are important, I think we need to accept a large amount of posterior uncertainty. It's a mistake for people to think that the result of a study should be some attitude of near-certainty.

Howard
on **March 15, 2018 11:43 AM at 11:43 am** said:

A little wrinkle: If there is both a main effect (e.g 2.8/ N^.5) and an interaction effect in the sample, then the power of the original main effect test drops from 80% to about 70%, because of greater variance in the main sample. Do you agree? I simulated this with some R code here

```
#simulation of main effect plus interaction
studies = 1000 # Number of studies
N=1000 # Number of people in each group in study
main_effect = 2.8 /sqrt(N)
interaction_effect_male = main_effect/2/2
interaction_effect_female = – interaction_effect_male
interaction_effect = interaction_effect_male – interaction_effect_female

#Simulation of Main Effect
study_results = c(
sapply(1:studies, FUN = function(x) {mean(rnorm(N/2, main_effect + interaction_effect_male,
1))}),
sapply(1:studies, FUN = function(x) {mean(rnorm(N/2, main_effect + interaction_effect_female,
1))}))
H0_results = sapply(1:studies, FUN = function(x) {mean(rnorm(N, 0, 1))})
reject_level = quantile(H0_results, probs = 0.975)
(power = mean(study_results > reject_level))
[1] 0.691
```

> **Andrew** on **March 15, 2018 11:48 AM at 11:48 am** said:

Howard:

I think this depends on the relative sizes of the effects and the residual variance. Usually I'm thinking about examples where effect sizes are small compared to residual variance; for example if you have binary data and effects that can change the probability from 0.4 to 0.6, or even 0.3 to 0.7, then most of the variance remains unexplained.

Carlos Ungil
on **March 15, 2018 2:04 PM at 2:04 pm** said:

Conclusion: measuring the height of a fridge in terms of hands is easier than measuring its width in terms of steps.

Daniel Lakeland on **March 15, 2018 3:02 PM at 3:02 pm** said:

maybe more like its frontal surface area (interaction of height, in hands to width in steps?)

Think again
on **March 15, 2018 2:22 PM at 2:22 pm** said:

Wouldn't the direction of the interaction matter? Wouldn't a disordinal interaction that qualifies the main effect have more power than the ordinal one? Maybe it's a silly question or maybe I'm missing something but that's what I've seem happen often.

Keith O'Rourke on **March 15, 2018 2:41 PM at 2:41 pm** said:

There is variation in ways a suitable example could be discerned.

For instance I likely would have used a two-sided test for the interaction and if I've done the calculations correctly that would make the increase in sample size a factor of 20 (larger for not knowing the direction of say male effect – female effect).

Think again on **March 15, 2018 3:21 PM at 3:21 pm** said:

Hm... right... but if the question is about power and power implies knowing what the direction should be then we should have an idea of the direction right? Statistics is applied statistics right?

Keith O'Rourke on **March 16, 2018 8:55 AM at 8:55 am** said:

But one-side versus two sided issue is about choosing a rejection region which might be informed by an idea of the direction but not necessarily determined by it – recall these threads were started with the title "Classical hypothesis testing is really really hard".

bxg
on **March 15, 2018 3:56 PM at 3:56 pm** said:

Two probably silly questions...

Doesn't our prior (it is prior, right?) intent to test the interaction change the main effect test itself? (And how many other interactions will we test?). Or are you thinking of running the same study design twice?

Second, we want power for the interaction test, supposing "that interactions of interest are half the size of main effects." Are we to suppose that the interactions are half the size of our pre-experiment expectation (whatever we used in the main-effect power calcuation) or does the power calculation for the interaction test depend on what we actually learned about the main effect? It seems to be weird to get to a situation where, perhaps, don't reject the main effect null, but continue to claim that the interaction test had high power (against a value which we perhaps don't take seriously any more.)

Andrew on **March 15, 2018 4:23 PM at 4:23 pm** said:

Bxg:

In the balanced design that I was assuming for this problem (a randomly assigned treatment, thus approximately equal proportions of men and women in each group), you can estimate the interaction without changing the estimate of the main effects.

Regarding your other question: I'm not at all proposing that decisions be made based on statistical significance or rejection of a hypothesis test. The real point here is working out the standard errors.

bxg
on **March 15, 2018 4:31 PM at 4:31 pm** said:

I think I'm misunderstanding your answer, but my first question was not about estimates. For some reason we are doing hypothesis tests and looking for statistically significant results (we shouldn't, but that's the question you asked). But we are planning on doing at least two tests with the same data. So the test thresholds need to be adjusted for that, no?

> **Andrew** on **March 15, 2018 4:33 PM at 4:33 pm** said:
>
> Bxg,
>
> Oh, sure, I was just assuming you'd do all the tests together. The tests are what they are, no need to be adjusting thresholds.

Squid
on **March 15, 2018 6:01 PM at 6:01 pm** said:

Would this be bad news for difference-in-difference type estimator that's popular in applied econometrics? There the goal is to estimate an interaction and find statistical significance…

Jake Westfall
on **March 15, 2018 6:50 PM at 6:50 pm** said:

I think the blog post is misleading in a couple of different ways.

(1) Contrary to the suggestion of the title, an interaction and a main effect will have the same power given the same N and the same standardized effect size (partial eta^2). Of course, the body of the blog post doesn't actually dispute this—it's based on the explicit assumption that the interaction effect size is half that of the main effects—so I'd just say the title is misleading, given that most people who see the title's conclusion will likely interpret it as applying to situations where the standardized effect size is held constant.

(2) As noted in the addendum, there's some ambiguity about what it means for the interaction effect size to be half that of the main effect on an unstandardized scale. I'd argue that we can only sensibly compare the sizes of unstandardized effects in this way after equating them on var(X), otherwise it's sort of apples-to-oranges. The 16X conclusion violates this principle because it compares main effects on a [-1/2, +1/2] scale to an interaction that's on a [-1/4, +1/4] scale. So when you cut β_interaction in half in addition to that change of scaling, I'd argue you're really putting the interaction at a quarter the size of the main effects, not half the size. If you equate the scales first and then set β_interaction to half on that scale, the sample size multiplier is 4X, not 16X (as hinted in the addendum). This also agrees with the fact that the sample size multiplier is 4X, not 16X, when the standardized effect size (partial eta^2) for the interaction is half that for the main effects.

Andrew on **March 15, 2018 7:48 PM at 7:48 pm** said:

Jake:

I don't really understand things like standardized effect size and partial eta—actually, I've never heard of partial eta. As discussed in the addendum, the power for estimating main effect and interaction will only be equal in a setting where, for example, the main effect is 0.6, with an effect of 0 for one group and 1.2 for the other. I agree that this *can* occur but I don't think it's typical. This is related to the piranha principle that we discussed not long ago on the blog.

In any case, I hope the R code and then the addendum helped.

To elaborate on one of the points implicitly arising in our discussion: There are cases where main effects are small and interactions are large. Indeed, in general, these labels have some arbitrariness to them, as my colleagues and I realized many years ago when studying congressional elections: recode the outcome from Democratic or Republican vote share to incumbent party vote share, and interactions with incumbent party become main effects, and main effects become interactions. So, yes, the above post is in the context of main effects which are modified by interactions; there's the implicit assumption that if the main effect is positive, then it will be positive in the subgroups we look at, just maybe a bit larger or smaller. Again, I think this makes sense in most of the social science research I've seen, and I think it makes sense for most of the interactions that people look at—especially in the common setting that people look at lots of interactions, in which case I think most of them will have to be small—but it won't apply in every case.

To continue that last thought: I think it makes sense, where possible, to code variables in a regression so that the larger comparisons appear as main effects and the smaller comparisons appear as interactions. This is what we did in our paper on incumbency analysis. You might say I'm now engaged in circular reasoning, but I think that in most cases, the very nature of a "main effect" is that it's supposed to tell as much of the story as possible. When interactions are important, they're important as modifications of some main effect. Again, not always (you can have a treatment that flat-out hurts men while helping women), but in such examples it's not clear that the main-effects-plus-interaction framework is the best way of looking at things.

> **Jake Westfall** on **March 15, 2018 8:46 PM at 8:46 pm** said:
>
> I don't doubt or dispute that, more often than not, interactions tend to have smaller effect sizes than main effects (and that higher-order interactions tend to smaller effect size than lower-order interactions). This would be the "hierarchical ordering principle" discussed in some design of experiments literature. My inaugural blog post was all about reasons why hierarchical ordering is probably true most of the time, and it includes the points just mentioned:
> http://jakewestfall.org/blog/index.php/2015/05/11/the-hierarchical-ordering-principle/
>
> Rather, my objection (1) was that the title is misleading because it's easily misinterpreted by readers who don't carefully read the blog post, which is almost certainly a large majority. I don't argue that it's actually common for main effects and interactions to have the same standardized effect size (e.g., the same partial correlation with Y)–like I said, hierarchical ordering is probably true more often not–

just that the natural interpretation of your title is that it refers to such cases. The context for me here is that many in my field mistakenly believe that interactions are *inherently* harder to detect than main effects *because they are interactions*– that is, even for the same partial correlation. And your title seems to just fuel those misconceptions.

Andrew on March 15, 2018 8:52 PM at 8:52 pm said:

Jake:

I take no responsibility for readers who don't read the post. I think everyone should read the comments too.

If I just wanted to write titles, I'd be on twitter. I blog because I want to work things through.

Think again on March 16, 2018 12:49 AM at 12:49 am said:

I absolutely agree. Not sure why he wont take responsibility for the misleading title. As if this whole hysterical obsession with power hasn't caused enough damage already. Stats students now a days are so bombarded with stuff about power that they often confuse it with effect size of statistical significance (yes i know talking about statistical significance is "bad" in this forum). If people knew about the size of the effect size ahead of time they wouldn't need to conduct experiments. But that's a rant for a different day…

Joe on March 16, 2018 4:10 AM at 4:10 am said:

I absolutely agree as well. He should also take responsibility for all of the reckless posts that include cat pictures. These could be easily misleading to cat lovers who don't carefully read the blog posts.

Andrew on March 16, 2018 6:53 AM at 6:53 am said:

Joe:

I absolutely agree. As if this whole hysterical obsession with cats hasn't caused enough damage already!

Think again on **March 16, 2018 2:52 PM at 2:52 pm** said:

That's right ! Now every time I want to publish a picture of my cat the editor is refusing to do so because he wants it from at least 100 angles. No wait it's actually a bad example that has nothing to do with the subject. I guess that means that your right again of course. Yes I need 10000 people in my sample for an interaction.

Andrew on **March 16, 2018 3:03 PM at 3:03 pm** said:

Think:

As I wrote above, I understand that my statistical advice can be upsetting because I'm a bearer of bad tidings. All I can tell you is that these issues have confused a lot of people for a long time, which is one reason why the replication crisis is a crisis and is one reason why people keep being surprised that "p less than 0.05" results aren't getting replicated. Getting angry is easy but it won't help you understand the world better in a replicable way.

Sameera Daniels on **March 17, 2018 10:21 AM at 10:21 am** said:

LOL

Ben Prytherch on **March 16, 2018 3:52 PM at 3:52 pm** said:

Think again: what does "effect size of statistical significance" mean? Are you saying that too much emphasis on power is resulting in students confusing effect size with significance?

How do you think power should be taught, relative to how it is currently taught?

Ben Prytherch on **March 16, 2018 2:23 PM at 2:23 pm** said:

"Rather, my objection (1) was that the title is misleading because it's easily misinterpreted by readers who don't carefully read the blog post, which is almost certainly a large majority"

Seems a bit harsh, and a bit hard to believe. Who goes to a technical statistical methods blog and skins

Ben Prytherch on **March 16, 2018 2:23 PM at 2:23 pm** said:

…skims the titles?

(Accidentally hit submit there)

Noah Motion
on **March 16, 2018 8:44 AM at 8:44 am** said:

Another nitpick: the order of the arguments in the pnorm functions is wrong, isn't it? It should be q, mean, sd, not mean, q, sd.

Milden on **March 16, 2018 6:06 PM at 6:06 pm** said:

It was the same for me, I rather expected

1-pnorm(1.96,2.8,1)
than
pnorm(2.8,1.96,,1)

it took a while to see that the two areas under the respective bell curves are always the same,
visually just reflected to the vertical axis at (2.8-1.96)/2

Noah Motion on **March 17, 2018 11:44 AM at 11:44 am** said:

Right. Oops.

Sean
on **March 16, 2018 8:51 AM at 8:51 am** said:

One interesting addition for the x_i in [0, 1] case is when the probability of 1 differs from 0.5. I believe the s.e. should increase for the interaction as you move away from p = 0.5 in either direction too.

a reader
on **March 16, 2018 12:19 PM at 12:19 pm** said:

I think this solution is mistaken.

In particular, suppose there is a *non-zero* interaction effect (which we have to be). If we are powered at 0.80 to detect a main effect in a model in which we've excluded the interaction effect, then the interaction effect gets absorbed into sigma. So sigma_mainEffectOnly can be arbitrarily larger than sigma_withInteractions. In fact, it's theoretically possible to have power 0.80 with the main effect and power 1.00 when we include the interaction with the same sample size (i.e. the pathological case in which all the error came from excluding the interaction effect).

So under this interpretation of your problem, which I think is quite reasonable, the answer is undefined.

> **Andrew** on **March 16, 2018 1:58 PM at 1:58 pm** said:
>
> A:
>
> See comment here for what I'd been thinking of.
>
> > **a reader** on **March 16, 2018 3:55 PM at 3:55 pm** said:
> >
> > Right, but to give a statement like "You need 16 times the sample size to estimate an interaction than to estimate a main effect", you need to add in the assumption that the ratio of interaction effect over sigma approaches 0 (assuming you're looking at a Gaussian response variable)…which is a pretty depressing assumption. Also note that this assumption means the main effect size you are looking at has roughly the same ratio.
> >
> > If you're willing to make the assumption that the estimates of the main + interaction effects are normally distributed (not too strong an assumption), then I think the required sample size ratio for the interaction effect should actually be a function of the original required sample size and the distribution of interaction variable (which we have stated is binary with p = 0.5)? I.e., if a small sample size is required, then the main effect must be large relative to sigma_mainEffectOnly, and

so sigma_withInteraction should be very small (as presented in my pathological case above, you can actually make the sample size required *smaller* than the no interaction effect sample size!). But it's not really clear to me that this function would be consistent across different response distributions (i.e. I think you get a different function if you're looking a Gaussian response vs. survival analysis model etc.).

Just a little tangent, but I think this is one lesson the field of statistics has taken the very wrong approach and the field of machine learning has really done right. Don't worry, I'll blame it all on p-values. Classical frequentist methods have been very afraid of model building, because it can make it very difficult to keep valid frequentist inference while also performing model selection from the data (although there's been plenty of recent work on addressing this). This fear of invalid inference has scared many statisticians away from model building (don't look for interactions because it might take away from your power to find a main effect!), while many machine learning researchers skipped p-value day and said "wow, it's really easy to build models with high predictive power if we just survey a large number of models (or hyper parameters) and pick the one that does best on out-of-sample error!"

Andrew on March 16, 2018 4:50 PM at 4:50 pm said:

A:

Just to be clear, I never said, "don't look for interactions because it might take away from your power to find a main effect." I think interactions are very important; we just need to accept that in many settings we won't be able to attain anything like near-certainty regarding the magnitude or even direction of particular interactions. That's a message that people don't want to hear, which is too bad, because uncertainty is a core statistical principle.

a reader on March 16, 2018 5:06 PM at 5:06 pm said:

' Just to be clear, I never said, "don't look for interactions because it might take away from your power to find a main effect." '

Of course, I don't expect you to be someone who wants to protect type I error rates **at all costs**!

Andrew on March 16, 2018 5:10 PM at 5:10 pm said:

A:

I was just worried because a post with a theme, "Interactions are hard to estimate," could be taken to imply, "Don't estimate interactions." But I don't want to imply that. I do want people to estimate interactions, I just want them (a) to estimate them with Bayesian inference using prior information, and (b) to accept that their posterior inferences will have a lot of uncertainty, in particular not considering it a failure when the 95% interval includes zero.

a reader on **March 16, 2018 5:45 PM at 5:45 pm** said:

Hmm, getting back to my original point, I think I might have been mentally overestimating the difference between sigma_mainEffect vs sigma_withInteraction.

In particular, in a simple Gaussian model, for sigma_mainEffect to be greater than 10% larger than sigma_withInteraction, we need delta_interaction just about equal to sigma_withInteraction, implying delta_mainEffect = 2 * sigma_withInteraction. As delta_mainEffect gets smaller, so does the relative difference between sigma_mainEffect and sigma_withInteraction.

This (a) doesn't change the sample size ratio very much at all and (b) would considered a very strong effect in most fields.

Arno
on **March 17, 2018 4:08 AM at 4:08 am** said:

Something crazy is going on: if I change labels for Male & Female (resulting in a different 0/1 dummy coding) my *power* drops from 70% to 35%:

https://pastebin.com/XHxrCq4X

The basic function is this:

```
sim1 = function(n=40, mu=0.6343) {
a = rnorm(n, 0.0, 1)
b1 = rnorm(n/2, mu − mu/4, 1)
b2 = rnorm(n/2, mu + mu/4, 1) # interaction is mu/2
dat = data.frame(Group = c(rep("C", n), rep("T", n)),
Sex = rep(c(rep("M", n/2), rep("F", n/2)), 2),
Val = c(a, b1, b2))
fit = lm(Val ~ Group*Sex, dat)
```

```
return(summary(fit)$coefficients)
}
```

Where is my error?
It can't be that an arbitrary label (and resulting positive or negative coefficient) halves my power from 70% to 35%!

Anoneuoid on **March 17, 2018 7:23 PM at 7:23 pm** said:

Thanks for taking a look at this. I'm not sure whats going on, it seems to have something to do with how lm is treating the factors. You can change only the ordering of the factor levels and see that the estimate of the effect (and associated test stats) for the other *factor level changes:*

set.seed(1234)
n = 4
mu = 1
a = rnorm(n, 0.0, 1)
b1 = rnorm(n/2, mu – mu/4, 1)
b2 = rnorm(n/2, mu + mu/4, 1) # interaction is mu/2

dat1 = dat2 = data.frame(Group = c(rep("C", n), rep("T", n)),
Sex = rep(c(rep("F", n/2), rep("M", n/2)), 2),
Val = c(a, b1, b2))

 # Now male is the reference
dat2$Sex = factor(dat2$Sex, levels = c("M", "F"))

fit1 = lm(Val ~ Group*Sex, dat1)
fit2 = lm(Val ~ Group*Sex, dat2)

 # Results
dat1
dat2

summary(fit1)
summary(fit2)

> dat1
Group Sex Val
1 C F -1.2070657
2 C F 0.2774292
3 C M 1.0844412
4 C M -2.3456977
5 T F 1.1791247
6 T F 1.2560559
7 T M 0.6752600
8 T M 0.7033681
> dat2
Group Sex Val
1 C F -1.2070657
2 C F 0.2774292

3 C M 1.0844412
4 C M -2.3456977
5 T F 1.1791247
6 T F 1.2560559
7 T M 0.6752600
8 T M 0.7033681
>
> summary(fit1)

Call:
*lm(formula = Val ~ Group * Sex, data = dat1)*

Residuals:
1 2 3 4 5 6 7 8
-0.74225 0.74225 1.71507 -1.71507 -0.03847 0.03847 -0.01405 0.01405

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.4648 0.9346 -0.497 0.645
GroupT 1.6824 1.3218 1.273 0.272
SexM -0.1658 1.3218 -0.125 0.906
GroupT:SexM -0.3625 1.8692 -0.194 0.856

Residual standard error: 1.322 on 4 degrees of freedom
Multiple R-squared: 0.4079, Adjusted R-squared: -0.03622
F-statistic: 0.9184 on 3 and 4 DF, p-value: 0.5081

> summary(fit2)

Call:
*lm(formula = Val ~ Group * Sex, data = dat2)*

Residuals:
1 2 3 4 5 6 7 8
-0.74225 0.74225 1.71507 -1.71507 -0.03847 0.03847 -0.01405 0.01405

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.6306 0.9346 -0.675 0.537
GroupT 1.3199 1.3218 0.999 0.374
SexF 0.1658 1.3218 0.125 0.906
GroupT:SexF 0.3625 1.8692 0.194 0.856

Residual standard error: 1.322 on 4 degrees of freedom
Multiple R-squared: 0.4079, Adjusted R-squared: -0.03622
F-statistic: 0.9184 on 3 and 4 DF, p-value: 0.5081

I usually do not (pretty much never) do anovas or use R factors, but this seems pretty disturbing.

Anoneuoid on **March 17, 2018 10:01 PM at 10:01 pm** said:

I guess you are supposed to do this instead of "summary", then they are the same?

> anova(fit1)
Analysis of Variance Table

Response: Val
Df Sum Sq Mean Sq F value Pr(>F)
Group 1 4.5071 4.5071 2.5798 0.1835
Sex 1 0.2409 0.2409 0.1379 0.7292
Group:Sex 1 0.0657 0.0657 0.0376 0.8557
Residuals 4 6.9881 1.7470
> anova(fit2)
Analysis of Variance Table

Response: Val
Df Sum Sq Mean Sq F value Pr(>F)
Group 1 4.5071 4.5071 2.5798 0.1835
Sex 1 0.2409 0.2409 0.1379 0.7292
Group:Sex 1 0.0657 0.0657 0.0376 0.8557
Residuals 4 6.9881 1.7470

Perhaps related (And someone cites Gelman 2005 there):
https://stats.stackexchange.com/questions/175246/why-is-anova-equivalent-to-linear-regression

Its probably another one of those things that is "obvious" to people with good training in stats but severely disturbing to most people trying to apply it.

Thomas B
on **March 17, 2018 10:19 AM at 10:19 am** said:

In chap 13 of your book with Jennifer Hill you note that using a hierarchical Bayesian framework estimates can be derived in the posterior even when n=1 for a label or feature. Why wouldn't this also be true for interactions?

Johannes
on **March 19, 2018 4:41 AM at 4:41 am** said:

"I don't know if all this in the textbooks, but it should be."

Maxwell and Delaney have it covered ("Designing Experiments and Analyzing Data: A Model Comparison Perspective", 2nd Ed., p. 318). They refer to an Abelson insight: The ratio of the t-values for the main effect and the interaction effect is equal to (t1+t2)/(t1-t2), where t1 and t2 refer to the simple main effects. Assuming t1 and t2 to be on the same side (i.e., ordinal interaction), the effect size of the interaction will always be smaller than the main effect unless one of the simple main effects is zero. Playing around with the equation by assuming that one simple main effect is half the size of the other (t2=.5*t1) yields the solution that the effect size of the main effect must be three times that of the interaction. In many cases where disordinal interactions are implausible, t2=.5*t1 may be a reasonable assumption. In this scenario, the power for the interaction would be ~5 times lower than the power for the main effect: pnorm(2.8, 1.96, 1) / pnorm(2.8/3, 1.96, 1).

> **Gary McClelland** on **March 20, 2018 12:14 PM at 12:14 pm** said:
>
> Some nice mathematical work on the issue of constraints on size of the interaction is available in
>
> Rogers, W. M. (2002). Theoretical and mathematical constraints of interactive regression models. Organizational Research Methods, 5, 212–230.

Vainius
on **March 24, 2018 10:54 AM at 10:54 am** said:

Theoretical calculation becomes more complicated if you do two-sided tests (that are used in a regression). Not sure about first part of the question, but results to last part could change because we would also include large values with opposite sign when calculating mean of significant results.

Ewout Steyerberg
on **May 13, 2018 5:49 AM at 5:49 am** said:

Realize I'm a bit late in this discussion. Great points made on this important subject of testing and estimating subgroup effects; a key issue in the analysis and interpretation of randomized clinical trials in medicine. Thank you Andrew, and all those that commented!

Perhaps I missed some of the points above; did anyone really simulate the case study as suggested? Admittedly, estimates of effect sizes are key; and these impact on the estimated SE in a model, correct?
Under the Null, main effect has SE 0.63; interaction has SE of 1.26, with N=1000, sigma 10.

With the given example, main effect 2.8*sigma; for x2== -.5, 2.1*sigma, and for x2== .5, 3.5*sigma. This is the implication of the interaction being half the size of the main effect (1.4; -.7 and +.7 effect).

This results is 3 interesting findings:
1. SE of main effect is estimated larger as 0.70
2. SE of main effect is estimated as 0.66 if we adjust for x2
3. SE of interaction is estimated as 1.26.

My reflections:
– The only correct model is the model with interaction; and there the SE is identical to what was derived under the Null (SE 1.26).
– In practice, we will start with the main effect model, and some variance is explained by adjusting for other covariates that are associated with the outcome. This is indeed what we observed for x2, even if x2 is interacting with x1. So, this confirms the recommendation to include covariates that are associated with the outcome, more than searching for subgroup effects.
https://www.ncbi.nlm.nih.gov/pubmed/2727470
– The inclusion of prognostic covariates is beneficial in linear models as well as in generalized linear models such as logistic or Cox regression:
https://www.ncbi.nlm.nih.gov/pubmed/9620808;
https://www.ncbi.nlm.nih.gov/pubmed/10783203;
https://www.ncbi.nlm.nih.gov/pubmed/15196615;
https://www.ncbi.nlm.nih.gov/pubmed/16275011

The R script:
```
library("arm")
N <- 1000
sigma <- 10
y <- rnorm(N, 0, sigma)
x1 <- sample(c(-0.5,0.5), N, replace=TRUE)
x2 <- sample(c(-0.5,0.5), N, replace=TRUE)
display(lm(y ~ x1))
display(lm(y ~ x1 + x2 + x1:x2))
# this was with y under the Null

# now with y under the alternative of separate x1 effects for x2 values
# specifically:
# overall effect of x1 = 2.8 * sigma; for x2==-.5: 2.1 * sigma; for x2==0.5: 3.5 * sigma
y[x1==.5 & x2== -.5] <- rnorm(length(y[x1==.5 & x2==-.5]), 2.1*sigma, sigma)
y[x1==.5 & x2== .5] <- rnorm(length(y[x1==.5 & x2== .5]), 3.5*sigma, sigma)
display(lm(y ~ x1)) # SE 0.72
display(lm(y ~ x1 + x2)) # SE 0.69
display(lm(y ~ x1 + x2 + x1:x2)) # SE interaction 1.31; 1.31/0.72 equals 1.82 rather than a factor 2
```

Anon
on **July 2, 2019 9:07 AM at 9:07 am** said:

I wonder if estimating these regression using Bayesian inference will ameliorate the problem in that the standard errors estimated are typically lower.

Natalie

on **November 5, 2021 9:21 AM at 9:21 am** said:

I know it's been a long time since this post has been active, but I came accross it because I'm reading "Regression and other stories" and I can't get my head around the statement "What happened was that the main effects are now estimated at the edge of the data: the estimated coefficient of x1 is now the difference in y, comparing the two values of x1, just at x2=0. " It migth be something obvious that I'm missing but I would appreciate some clarification about why this happens when changing parameterization.

Comments are closed.