0. What's this about? I've written a number of posts in last yr or so on the value of Bayesian likelihood ratios (a heuristic cousin of the "Bayes Factor") as an "evidentiary weight" statistic generally, and its value in particular as a remedy for the inferential bareness of p-values and related NHT statistics used to implement "null hypothesis testing."

In this post I want to call attention to another virtue of using "likelihood ratios: the contribution they can make to protecting against the *type 1 error risk* associated with underpowered studies. Indeed, I'm going to try to make the case for using LRs for this purpose *instead of* a method proposed by stats legend & former Freud expert Andrew Gelman (Gelman & Carlin 2014).

As admittedly elegant, and as admittedly valuable it has been in making people aware of a serious problem, the statistical indexes that the G&C method features inject a form of ***confirmation bias*** into the practical assessment of the weight we should afford empirical studies. Using LRs to avoid the "type 1" error risk associated with underpowered studies avoids that.

Or at least that's what I think.

I must be crazy, huh?

Continue if you dare . . .

[below the fold:]

1. Underpowered studies & type 1 error? Wha?

So . . . .

Pretty much everyone knows (but sadly some don't . . .) that underpowered studies invite type 2 error—i.e., the mistaken rejection of the inference that *x* is causally related to *y*. The smaller one's sample, the greater the probability that measurement error will either hide a real effect completely or render such an effect indistinguishable from one due to chance. So if we concluded that there's no meaningful relationship between *x* and *y* based on a small sample study, we are at risk of making a mistake.

The usual insurance policy against Type 2 error of this sort consists in using a sample large enough to furnish a predetermined likelihood—usually 0.80—of observing an effect of some theoretically justifiable size at a specified "statistical significance" level (e.g., "$p < 0.05$") (e.g., Streiner 2003).

Fine, fine.

But now thanks largely (& heroically) to stats legend & former Freud expert Andrew Gelman, empirical scholars (including but not limited to social psychologists) have come to be conscious that underpowered studies that report "significant" results can also present a risk of *Type 1 error*—the mistaken acceptance of an inference that *x* *is* causally related to *y* in some meaningful way.

This seems counterintuitive to the NHT-conditioned mind. After all, "statistical significance" *is* the insurance policy against type 1 error associated with underpowered studies. If one says one won't "accept" a finding unless it is "significant at $p < 0.05$," then one is saying (supposedly) that one would regret mistakenly crediting a chance effect as "real" 19x more than mistakenly rejecting a "real" effect as "chance" one.

But as Gelman has persistently pointed out (e.g., Gelman & Weakliem 2009), this insurance policy has a hole in it.

When both the "true" effect and the sample size are small, a researcher might by "chance" get an observed effect that just clears the "$p < 0.05$" bar. If the "true" effect is indeed "small," then under those conditions the observed effect would have migrated consistently (and pretty darn rapidly) toward zero as the sample size increased.

The signature of this underpowered-design hazard, according to Gelman, is an implausibly *large* reported effect size. If we have reason to believe that the "true" effect size—if there is one at all—must be quite small, then the *only* way a study that has a *small* sample--and hence a relatively large standard error--can find an effect significant at $p < 0.05$ is when the observed effect size in that particular study is *much larger* than the "true" effect.

Indeed, Gelman (who himself hates the "Type 1" & "Type 2" terminology; he's right on that too, but since it's impossible to replace every part of a moving vehicle at once without it grinding to a halt, I'll leave that aside) calls this a "Type M error," which quantifies the ***m**agnitude* of the effect size exaggeration attributed to use of an underpowered study design.

To illustrate, Gelman his coauthor Carlin (2014) critique Durante, Arsena, and Griskevicius (2013), who report finding that the stage of a woman's menstrual cycle to be associated with a 17-percentage point difference in the probability of voting for a Democrat as opposed to a Republican presidential candidate.

Gelman actually thinks that the DAG study is filled with all sorts of methodological errors. But put all those aside—b/c that's what G&C do, in order to illustrate how their underpowered-study-type-1-error-risk protection apparatus works.

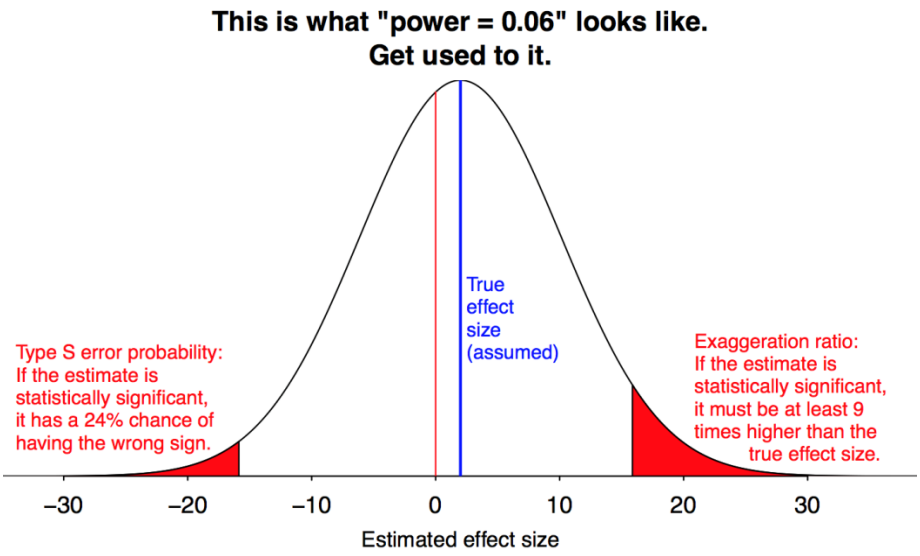
In other words let's assume DAG is otherwise a completely valid study and consider only what sort of inferential mistake we can make if we don't take account of it being "underpowered."

G&C state,

> Given the lack of evidence for large swings among any groups during the campaign, one can reasonably conclude that any average differences among women at different parts of their menstrual cycle would be small. Large differences are theoretically possible, as any changes during different stages of the cycle would cancel out in the general population, but are highly implausible given the literature on stable political preferences. Furthermore, the menstrual cycle data at hand are self-reported and thus subject to error. Putting all this together, if this study was to be repeated in the general population, *we would consider an effect size of 2 percentage points to be on the upper end of plausible differences in voting preferences*.

Based on DAG's reported $p = 0.035$, G&C determine the standard error for DAG's 17% effect size finding to be 8 percentage points. That's pretty big relative to the effect size—a consequence of the small sample in DAG (N = 134, as far as I can tell from the somewhat obscure reporting in the paper).

As illustrated in this cool graphic (one not included in G&C but made famous via Gelman's blog & stops he made in a raucous barnstorming tour over the course of 2015), G&C conclude that the "type M error" is 9.7: that is, assuming a "true" effect size of 2% and standard error of 8%, DAG would *have* had to observe an effect size over 9x larger than the "true effect" in order to obtain a "statistically significant" result, given their small sample size:



The graphic reports a study "power" calculation. For G&C, "power" is the likelihood that a "replication" of the study, using the same small sample size, would generate a "statistically significant" result given the assumed "true effect" size.

The "G&C power" for DAG is 0.06. That is, assuming the "true effect" is 0.02 (2% change in voting preference based on stage of ovulation cycle), G&C conclude that there is only a 6% chance that the DAG result would be replicated.

Finally, G&C report a "Type S" error calculation. For them, a "Type S" error is the probability that the observed effect of a study, conditional on it being "statistically significant" result, will have the wrong *sign*.

Given G&C's assumption that the "true" effect size here is 0.02, we wouldn't expect a DAG replication with $N = 134$ to generate a significant effect very often--only 6% of the time, according to G&C's "power" calculation. But by their Type S error calculation, in about 1/4 of those replications in which we do observe a "statistically significant" result, the sign will be negative!

Veeeeeery interesting – and for sure ***very helpful*** in highlighting how confused and misled we can get if we treat the "statistical significance" of a small sample study as grounds for believing that *x* causes *y*—and in particular that the *size* of the causal effect is anything close to what the study reported.

3. Hey wait. . . .

But still, I got to say, I don't like G&C's apparatus!

Indeed, I think "Type M," "Type S," and G&C "power" obligate whoever is using them to engage in a form of *confirmation bias* in assessing how much weight to give a particular study.

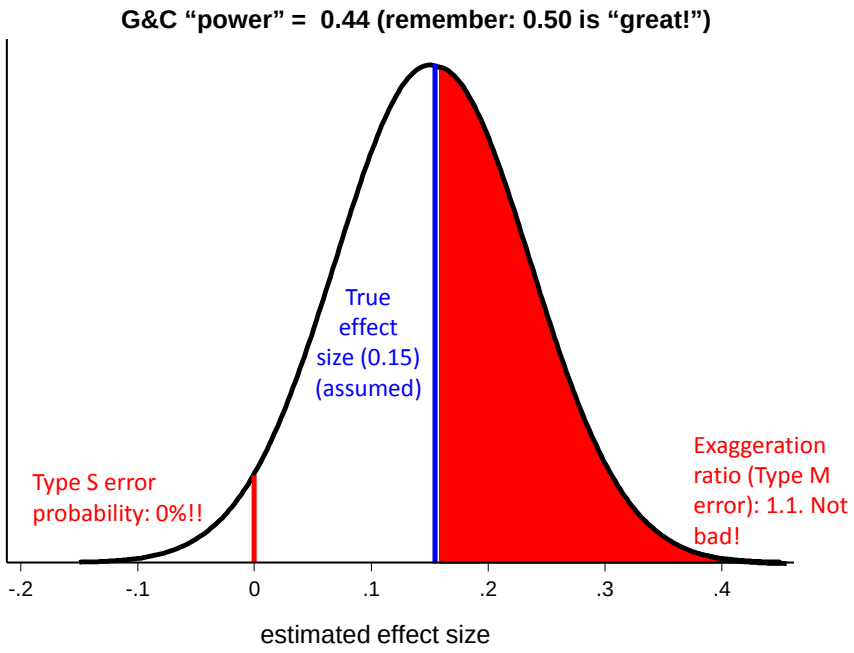
I'll illustrate first, explain later.

Imagine that Dr. W.T.F. Kredulius believes the "true" effect of a woman's menstrual cycle on her presidential voting preference is 15%.

When he looks at the DAG results, he *obviously* won't get the same "Type M," "Type S," and G&C "power" calculations that G&C, who assume the "true" effect is 0.02 do.

In fact, his calculations (derived via statistical simulation, as G&C's were) will look like this:

From Dr. W.T.F. Kredulius's point of view...
So *whose* G&C (2014) calculations should we "get used to"?



Hey--nothing for us to worry (or DAG to be embarrassed) about here!

The "power" rating might seem small if one is used to thinking " 'power should be 0.80.' "

But in fact that's the way to think (if one has an NHT-conditioned mind) when one is trying to assess the risk of "Type 2" error associated with an underpowered study.

If we are thinking about the risk of Type 1 error, then we should be satisfied so long as G&C "power" is in the vicinity of 0.50.

You wouldn't know it (more evidence of the detrimental impact of substituting NHT protocols for thinking) from all the fretting of about failure of "1/2 the studies in social psychology" to replicate at "$p < 0.05$." But b/c a p-value is a random variable, when a study finds a "significant" result at "$p = 0.05$," *we should expect 1/2 of the "replications" of a perfectly valid "$p = 0.05$" study to have p-values below and 1/2* ***above*** *$p = 0.05$. . . .*

So if we use the G&C "power" index to protect ourselves from type 1 error, we should view power = 0.50 as perfectly adequate!

If WTFK thinks 0.44 is "close enough," I think we'd have to be kind of petty to argue with him.

Also, because WTFK thinks the "true" effect is 15%, he'll calculate the Type M error as 1.1. That is, he'll calculate that when N = 134 the observed effect will have to be 1.1 times greater (about 0.16) than what he assumes to be the "true" effect (0.15) to be "significant" at "$p < 0.05$." No big deal!

And look: Type S error, by WTFK's calculation, will be 0%! The likelihood that one will observe an effect that is significant & has the wrong sign if we assume the "true" effect is 0.15 is astronomically low w/ N = 134.

But G&C will surely protest that WTFK is peddling his Type M, Type S and G&C "power" values as kosher only because his assumed "true" effect of 0.15 is *absurd*.

It might be . . . .

But maybe G&C are the ones who are making a big error in assuming the "true" effect is only 0.02?

Presumably, the whole point of the DAG study was to help get some *evidence* on whose hypothesized "true" effect size—G&C's or WTFK's—is closer to true.

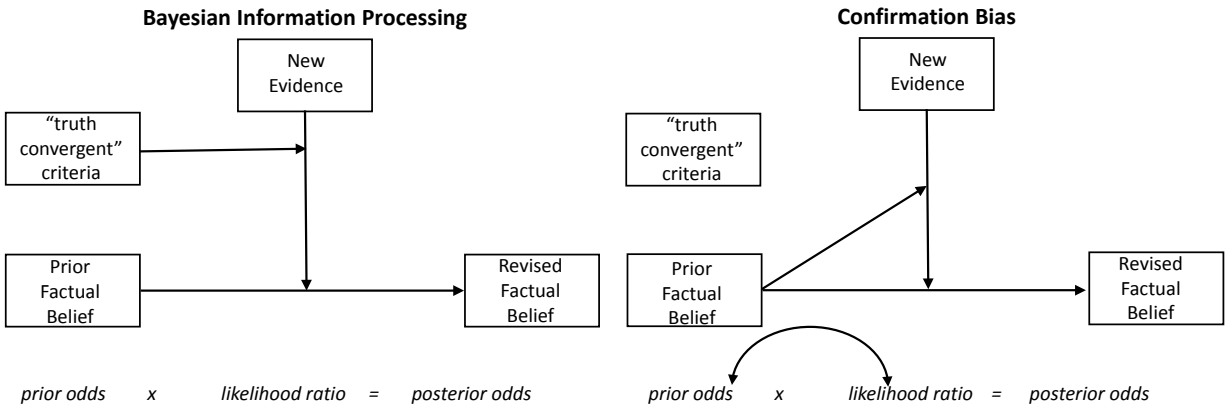
But if we have to pick sides on whose hypothesis was more likely true *in order to know whether to take DAG seriously as evidence* of the "true" effect size, then we are definitely in a bad place: it's called *confirmation bias*.

In Bayesian terms, G&C's & WTFK's alternative "assumed 'true' effect sizes" are their *priors*.

Bayesian reasoning says we can assume whatever the hell we want about the "true" effect size. That's the beauty of it! No harm will come of being wrong—really really really wrong, even—so long as we keep appropriately updating by multiplying our prior odds by a factor equivalent to the likelihood ratio of any new evidence we encounter.

But obviously that strategy for getting smarter won't work if we use "consistent with our priors" as the guide for determining the likelihood ratio (the weight, in practical terms) to be assigned new evidence. In that case, people who start w/ opposing priors will never converge b/c they will each assign an LR of 1 to anything that is inconsistent with what they already believe!

This sort of endogeneity between priors and likelihood ratio *is* the essence of confirmation bias (Stanovich 2011; Rabin & Schrag 1999). It guarantees not only persistent disagreement but persistent stasis in our understandings of how the world works: no matter how compelling the contrary evidence is, we will never update our priors—precisely b/c its being contrary to our priors will convince us not to take the evidence seriously.



To avoid confirmation bias, we need to determine the LR or weight to assign evidence on the basis of criteria *independent of our priors*.

The G&C apparatus for detecting which studies present a underpowered-study-type-1-error risk are inconsistent with this basic, core element of sound empirical inference, b/c calculating Type S error, Type M error, and G&C power all commit us to using our priors to critically assess the weight to assign particular study results.

4. Likelihood ratios to the rescue!

Wow. So what are we going to do?

G&C definitely are right that we need to avoid being sucked into giving undue weight to studies that by virtue of being underpowered seduce us into Type 1 error—crediting *x* causes *y* claims that are "wrong," either b/c *x* doesn't cause *y* at all or b/c whatever causal effect exists is much much much weaker than what is being represented by the study in question.

But the apparatus G&C have created to protect us from this risk—the measures of Type S and Type M error & their distinctive type-1-error "power" index—all commit us to a form of reasoning biased toward discounting evidence because it challenges what we already "assume" to be true.

Well, how about a solution that reflects the genius of Bayesian reasoning's ecumenical stance toward opposing priors? It's inspiring philosophy that we "can all just get along" & *get smarter* at the same time so long as we keep updating appropriately when confronted with new evidence?

As I just explained, to get the benefit of this inferential strategy, we need to have a valid means of determining the "weight" of the evidence—the conceptual equivalent of the likelihood ratio, in Bayesian terms—that is independent of our priors.

Well, the use of "Bayesian likelihood ratios" as a "weight of the evidence" statistic gives us exactly that!

The "Bayesian likelihood ratio/weight of the evidence" statistic—let's call it the "Good-Rozeboom-Goodman score" (Good 1984, 1994; Rozeboom 1960; Goodman 2005, 1999)--is the workpersonlike cousin of the Bayes Factor. Because unlike a real Bayes Factor, it makes certain simplifying assumptions about how to model the distribution of "observed" effects in relation to the the "true" effect, & doesn't attempt to integrate over the the interval of effects that inhabit the space between competing hypotheses, we should view it as essentially a **heuristic** to gauge the practical weight of the evidence (of course, some people convincingly treat even a kosher Bayes Factors as heuristics; the turtle lady would probably tell us that the world just is a heuristic sitting ontop of a heuristic sitting ontop of a heuristic understanding of evidentiary weight . . . .).

One of the things the GRG LR can do when used in this way is protect us from Gelman's very appropriate concern about the risk of Type 1 error associated with underpowered studies.

But the cool thing is that it can do this in a way that *avoids* the confirmation bias built into the G&C apparatus.

Essentially, the guiding insight of the GRG score is that every experimental result (or equivalent empirical observation) is itself a random variable.

The reason is the inescapability of measurement error in our experimental machinery. In effect, every time we do a **valid** experiment (or engage in any **valid** observational study equivalent), we are randomly drawing from a probability density distribution of results that has the "true effect" as its mean.

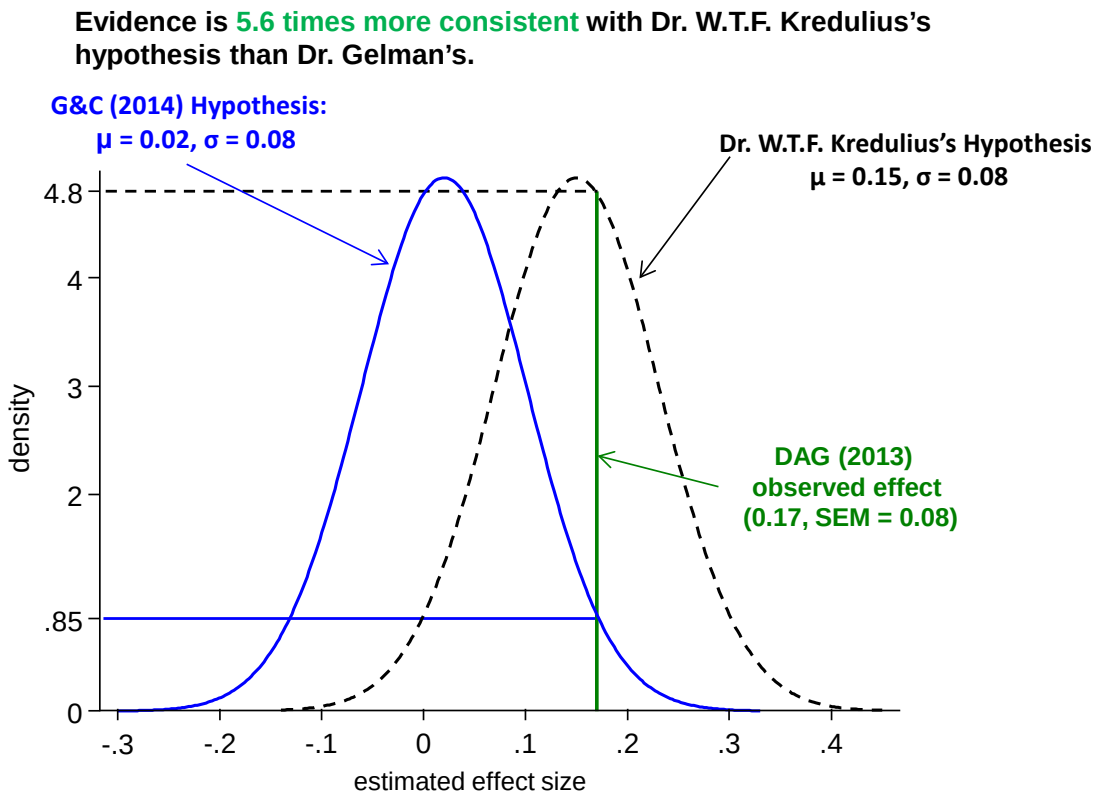
Of course, the occasion for experimental testing is the existence of competing hypotheses on the "true" effect size. What we want, then, is a way to figure out how much weight to assign any experiment in relation to competing hypotheses of interest, given that any experimental observation is a random variable. And of course, we want a way to do that that doesn't simply "assume" the correctness of our own favored hypothesis!

GRG gives us a strategy for doing that.

It says we should simply compute the relative likelihood of the observed effect conditional on each hypothesis *recognizing that* the observed effect is a random variable in regard to *each* of the relevant competing hypotheses.

Here's how that strategy would work for DAG, given the competing hypotheses of G&C and WTFK:

This is what a GRG "likelihood ratio" alternative to G&C (2014) looks like. What's wrong with it?



There are two probability density distributions: one for G&C's hypothesis—"true" effect = 2%; and one for WTFK's—"true" effect = 0.15.

The standard errors for each are 0.08.

That's the precision of the estimate that DAG obtained in their N = 134 experiment. Given the inevitability of measurement error, we will assume that G&C's and WTFK's "hypotheses" actually comprise *the range of values* associated with probability distributions that have means equivalent to their respective hypothesized "true" effect sizes and the DAG standard error.

This is the move, BTW, that sets the GRG LR statistic apart from the Bayes Factor™, which demands additional assumptions about the properties of the density distribution associated with competing hypotheses. If you don't like that simplification, I understand. But note that G&C's apparatus uses exactly the same simplification. So as between these two strategies—GRG and G&C—this simplifying decision is a wash.

So treating the G&C & WTFK hypotheses as opposing probability distributions of experimental outcomes, we can now compute a Bayesian LR for DAG in relation to these two competing hypotheses. All we have to do is compare the relative likelihood of observing the DAG result under the each hypothesis's probability density distribution.

As the Figure illustrates, DAG's result is 5.6x more consistent with G&C's.

Likely they'd both be taking the Rozeboom-Goodman score a bit too seriously if they did this, but both could now update their priors on their respective hypothesis rather than its rival being true by a factor equal to this LR.

I'm pretty sure G&C view the odds of 2% rather than 15% being the "true" effect of women's menstrual cycles on their presidential voting behavior as some astronomically high number to 1. I'll assume that WTFK likewise believes the odds are astronomically high in favor of 15% vs. 2%.

Accordingly, once both apply a 5.6 LR to their priors, they'll still basically believe that the odds are some astronomically high number to 1 in favor of their respective hypotheses.

It's really just not a big deal, in other words, for anyone to take DAG on board, because the LR associated with DAG is pretty modest given the strength of the priors of those who harbor one or another hypothesis on the "true" effect size of women's menstrual cycles on their voting behavior.

*This—the relatively inert effect of DAG on scholars' assessment of the probability of the competing hypotheses—**is** the penalty for the low power of DAG's study!*

Remember, GRG says that we should treat any observed effect as having been drawn from the probability density distribution associated with a random variable. The breadth of that distribution will be the standard error associated with the experiment in question. The bigger the standard error, the more probable the observed effect will be with multiple competing hypotheses.

So if one wants evidence that has really strong weight, GRG says *increase your friggin' sample size!* That will reduce the standard error of your measurement, shrink the range of the probability density distributions associated with competing hypotheses, and thus *increase* the size of the LR associated with any observed effect.

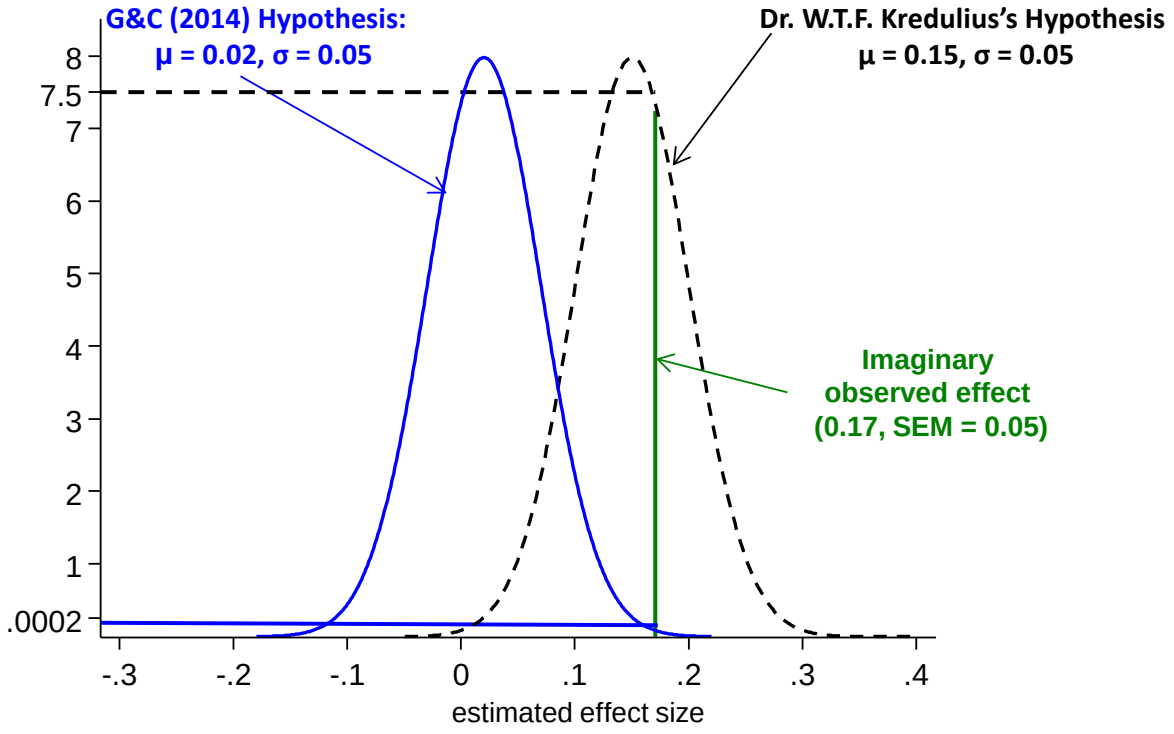
To illustrate, imagine DAG had received the same result—a 17% shift in the presidential candidate preference of women depending on their menstrual cycle—in a study with 400 subjects. Taking the 0.08 SEM for 134 subjects as given, the SEM for 400 subjects would be about 0.05.

Here is what the GRG LR assessment of the weight of the evidence would then look like:

This is what a GRG "likelihood ratio" alternative to G&C (2014) would look like if we got same result in DAG w/ N ≈ 400.

We'll just have to deal with that (if it ever happens)!

Evidence is 3.75 x 10^3 times more consistent with Dr. W.T.F. Kredulius's hypothesis than Dr. Gelman's.



Whoa! I think WTFK would then have something to chortle about—and G&C, being the good scientists that they are, would shrug their shoulders (tip their 1969 Mets caps in G's case), update their priors accordingly, and move on.

Of course, that *isn't* the weight that the real DAG study bears—because it was *underpowered* with respect to the evidence needed to start achieving convergence among those who have competing hypotheses this far apart.

If DAG *did* a study w/ N = 400, I'm sure G&C think it is ***inconceivable*** that they'd get a result like the one they got in their N = 134 study. So G&C are perfectly entitled under the GRG LR test to remain immensely skeptical of the DAG result.

But they aren't--and none of the rest of us would be either--if that skeptical orientation rests on a methodological apparatus that directs us to dismiss D&G's study results out of hand (presumably G&C expect journal editors & reviewers to use something like their methods to assess publishability) b/c the observed effect is just too far out of line with what we all "already know" or "assume" is the "true" effect size.

We should just wait & see. . . .

Indeed, by now, others, using larger sample sizes have reported finding nothing even remotely resembling the results in DAG, just as G&C expected (Scott & Pound 2015). All G&C, WTFK, and all the rest of us have to do is take all the results on board, give them the weight they are due in a GRG LR or equivalent sense, and the risk of Type 1 error associated with underpowered studies won't hurt us one bit.

5. A proviso: internal validity

Now, as I said, my analysis is assuming—just as G&C did--that the DAG study is *otherwise valid,* i.e., that the only complaint one could make about treating the result seriously in attempting to draw inferences on the basis of it is its small sample size. The GRG LR test, whether used to protect us from the risk of "type 1 error" associated with underpowered studies or to do any of the other great things it can do, *presupposes* that a study design is internally valid.

I'm sure G&C don't believe the DAG study is internally valid, and frankly I don't either. Its internal invalidity could very well warrant rejection of publication and dismissal of its results out of hand, etc.

But if G&C or anyone else believe that we should dismiss DAG's result on that basis, then demonstrating its internal invalidity—and not just dismissing it as "underpowered"—is the burden they should carry in scholarly debate over how the world works.

6. A last point: The problems of NHT can't be fixed with more NHT!!!

The occasion for G&C's very important paper, and Gelman's campaign generally to alert people to the Type 1 error associated with small sample size, is the inferential deficiency of NHT. Only those who confuse "significant at p< 0.05" for "*this* is the 'true' effect size" or "*x* therefore is the cause of *y*" or "it has been established by science . . ." would ever make the mistake of thinking that a result like DAG's, because it is "significant," "proves" that women's menstrual cycles influence their presidential candidate voting or even furnishes any meaningful "weight" on the balance of evidence on one side of which sits such a silly hypothesis.

But the problem w/ G&C's remedy is that it presupposes NHT! It is designed to help supply a set of instructions to be followed, a new set of buttons to be pushed, by those who have substituted the "which button do I push" mentality of NHT for thought in doing statistical analysis.

This is a corrective strategy that is doomed to implode from the weight of its internal contradictions!

The solution to the problems of NHT is not "more NHT" – but less, as in zero.

GRG LR ratios are one of the many Bayesian devices that are intended to replace NHT and restore actual *thinking* to the process of empirical inquiry, in the social sciences & elsewhere!

References

Durante, K.M., Rae, A. & Griskevicius, V. The Fluctuating Female Vote Politics, Religion, and the Ovulatory Cycle. *Psychol Sci*, 0956797612466416 (2013).
Gelman, A. & Carlin, J. Beyond Power Calculations Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science* **9**, 641-651 (2014).
Good, I. Causal Tendency, Necessitivity and Sufficientivity: an updated review. in *Patrick Suppes: Scientific Philosopher* 293-315 (Springer, 1994).

Good, I.J. Weight of evidence: A brief survey. in *Bayesian statistics 2: Proceedings of the Second Valencia International Meeting* (ed. J.M. Bernardo, M.H. DeGroot, D.V. Lindley & A.F.M. Smith) 249-270 (Elsevier, North-Holland, 1985).
Goodman, S.N. Introduction to Bayesian methods I: measuring the strength of evidence. *Clin Trials* **2**, 282 - 290 (2005).
Goodman, S.N. Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of internal medicine* **130**, 1005-1013 (1999).
Rabin, M. & Schrag, J.L. First Impressions Matter: A Model of Confirmatory Bias. *The Quarterly Journal of Economics* **114**, 37-82 (1999).
Rozeboom, W.W. The fallacy of the null-hypothesis significance test. *Psychological bulletin* **57**, 416 (1960).
Scott, I.M. & Pound, N. Menstrual cycle phase does not predict political conservatism. PLoS ONE 10(4): e0112042, doi:10.1371/journal.pone.0112042 (2015).
Stanovich, K.E. *Rationality and the reflective mind* (Oxford University Press, New York, 2011).
Streiner, D.L. Unicorns Do Exist: A Tutorial on "Proving" the Null Hypothesis. *Canadian Journal of Psychiatry* **48**, 756-761 (2003).