power analysis is "logically doomed" (p. 22), while endorsing a more nuanced use of both tests and intervals as in a severity assessment.

Our next exhibit looks at retrospective power in a different manner, and in relation, not to insignificant, but to significant results. It's not an objection to power analysis, but it appears to land us in a territory at odds with severity (as well as CIs and tests).

Exhibit (vii): Gelman and Carlin (2014) on Retrospective Power. They agree with the critiques of performing post-experiment power calculations (which are really shpower calculations), but consider "retrospective design analysis to be useful . . . in particular when apparently strong (statistically significant) evidence for nonnull effects has been found" (ibid., p. 2). They worry about "magnitude error," essentially our fallacy of making mountains out of molehills (MM). Unlike shpower, they don't compute power in relation to the observed effect size, but rather "on an effect size that is determined from literature review or other information external to the data at hand" (ibid.). They claim if you reach a just statistically significant result, yet the test had low power to detect a discrepancy from the null that is known from external sources to be correct, then the result "exaggerates" the magnitude of the discrepancy. In particular, when power gets much below 0.5, they say, statistically significant findings tend to be much larger in magnitude than true effect sizes. By contrast, "if the power is this high [.8], . . . overestimation of the magnitude of the effect will be small" (ibid., p. 3).

From the MM Fallacy, if POW(μ_1) is high then a just significant result is *poor* evidence that $\mu > \mu_1$; while if POW(μ_1) is low it's good evidence that $\mu > \mu_1$. Is their retrospective design analysis at odds with severity, *P*-values, and confidence intervals? Here's one way of making their assertion true using test T+: If you take the observed mean $\overline{x}_\alpha$ as the estimate of μ, and you happen to know the true value of μ is smaller than $\overline{x}_\alpha$ – between $\mu = \mu_0$ and $\mu = \overline{x}_\alpha$ (where the power ranges from α to 0.5.) – then obviously $\overline{x}_\alpha$ exceeds ("exaggerates") μ. Still I'm not sure this brings agreement.

Let's use our water plant accident testing $\mu \leq 150$ vs. $\mu > 150$ (with $\sigma = 10$, $\sigma/\sqrt{n} = 1$). The critical value for $\alpha = 0.025$ is $d_{0.025} = 1.96$, or $\overline{x}_{0.025} = 150 + 1.96(1) = 151.96$. You observe a *just* statistically significant result. You reject the null hypothesis and infer $\mu > 150$. Gelman and Carlin write:

> [An] unbiased estimate will have 50% power if the true effect is 2 standard errors away from zero, it will have 17% power if the true effect is 1 standard error away from 0, and it will have 10% power if the true effect is 0.65 standard errors away from 0. (ibid., p. 4)

These correspond to μ =152, μ =151, and μ =150.65. It's odd to talk of an estimate having power; what they mean is that the test T+ has a power of 0.5 to detect a discrepancy 2 standard errors away from 150, and so on. The "unbiased estimate" here is the statistically significant $\overline{x}$. To check that we match their numbers, compute POW(μ = 152), POW(μ = 151), and POW(μ = 150.65)[4]:

(a) $\Pr(\overline{X} \geq 151.96;\ \mu = 152) = \Pr(Z \geq 0.04) = 0.51$;
(b) $\Pr(\overline{X} \geq 151.96;\ \mu = 151) = \Pr(Z \geq 0.96) = 0.17$;
(c) $\Pr(\overline{X} \geq 151.96;\ \mu = 150.65) = \Pr(Z \geq 1.31) = 0.1$.

They appear to be saying that there's better evidence for $\mu \geq 152$ than for $\mu \geq 151$ than for $\mu \geq 150.65$, since the power assessments go down. Nothing changes if we write >. Notice that the SEV computations for $\mu \geq 152$, $\mu \geq 151$, $\mu \geq 150.65$ are the complements of the corresponding powers 0.49, 0.83, 0.9. So the lower the power for μ_1 the stronger the evidence for $\mu > \mu_1$. Thus there's disagreement. But let's try to pursue their thinking.

Suppose we observe $\overline{x} = 152$. Say we have excellent reason to think it's too big. We're rather sure the mean temperature is no more than ~150.25 or 150.5, judging from previous cooling accidents, or perhaps from the fact that we don't see some drastic effects expected from water that hot. Thus 152 is an *overestimate*. The observed mean "exaggerates" what you know on good evidence to be the correct mean (< 150.5). No one can disagree with that, although they measure the exaggeration by a ratio.[5] Is this "power analytic" reasoning? No, but no matter. Some remarks:

First, the inferred estimate would not be 152 but rather the lower confidence bounds, say, $\mu > (152 - 2\sigma_{\overline{X}})$, i.e., $\mu > 150$ (for a 0.975 lower confidence bound). True, but suppose the lower bound at a reasonable confidence level is still at odds with what we assume is known. For example, a lower 0.93 bound is $\mu > 150.5$. What then? Then we simply have a conflict between what these data indicate and assumed background knowledge.

Second, do they really want to say that the statistically significant $\overline{x}$ fails to warrant $\mu \geq \mu_1$ for any μ_1 between 150 and 152 on grounds that the power in this range is low (going from 0.025 to 0.5)? If so, the result surely couldn't warrant values larger than 152. So it appears no values would be able to be inferred from the result.

[4] You can obtain these from the severity curves in Section 5.4.
[5] There are slight differences from their using a two-sided test, but we hardly add anything for the negative direction: For (a), $\Pr(\overline{X} < -2;\ \mu = 2) = \Pr(Z < -4) \simeq 0$. The severe tester would not compute power using both directions once she knew the result.

A way to make sense of their view is to construe it as saying the observed mean is so out of line with what's known that we suspect the assumptions of the test are questionable or invalid. Suppose you have considerable grounds for this suspicion: signs of cherry picking, multiple testing, artificiality of experiments, publication bias, and so forth – as are rife in both examples given in Gelman and Carlin's paper. *You have grounds to question the result* because you *question the reported error probabilities*. Indeed, no values can be inferred if the error probabilities are spurious, the severity is automatically low.

One reasons, if the assumptions are met, and the error probabilities approximately correct, then the statistically significant result *would* indicate $\mu > 150.5$, *P*-value 0.07, or severity level 0.93. But you happen to know that $\mu \leq 150.5$. Thus, that's grounds to question whether the assumptions are met. You suspect it would fail an audit. In that case put the blame where it belongs.[6]

Recall the (2010) study purporting to show genetic signatures of longevity (Section 4.3). Researchers found the observed differences suspiciously large, and sure enough, once reanalyzed, the data were found to suffer from the confounding of batch effects. When results seem out of whack with what's known, it's grounds to suspect the assumptions. That's how I propose to view Gelman and Carlin's argument; whether they concur is for them to decide.

5.6 Positive Predictive Value: Fine for Luggage

Many alarming articles about questionable statistics rely on alarmingly questionable statistics. Travelers on this cruise are already very familiar with the computations, because they stem from one or another of the "*P*-values exaggerate evidence" arguments in Sections 4.4, 4.5, and 5.2. They are given yet another new twist, which I will call the diagnostic screening (DS) criticism of significance tests. To understand how the DS criticism tests really took off, we should go back to a paper by John Ioannidis (2005):

> Several methodologists have pointed out that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles should

[6] The point can also be made out by increasing power by dint of sample size. If $n = 10{,}000$, $(\sigma/\sqrt{n}) = 0.1$. Test T+ ($n = 10{,}000$) rejects H_0 at the 0.025 level if $\overline{X} \geq 150.2$. A 95% confidence interval is [150, 150.4]. With $n = 100$, the just 0.025 significant result 152 corresponds to the interval [150, 154]. The latter is indicative of a larger discrepancy. Granted, sample size must be large enough for the statistical assumptions to pass an audit.