A/B Testing Intuition Busters

Common Misunderstandings in Online Controlled Experiments

Ron Kohavi
Kohavi
Los Altos, CA
ronnyk@live.com

Alex Deng
Airbnb Inc
Seattle, WA
alexdeng@live.com

Lukas Vermeer
Vista
Delft, The Netherlands
lukas@lukasvermeer.nl

ABSTRACT

A/B tests, or online controlled experiments, are heavily used in industry to evaluate implementations of ideas. While the statistics behind controlled experiments are well documented and some basic pitfalls known, we have observed some seemingly intuitive concepts being touted, including by A/B tool vendors and agencies, which are misleading, often badly so. Our goal is to describe these misunderstandings, the "intuition" behind them, and to explain and bust that intuition with solid statistical reasoning. We provide recommendations that experimentation platform designers can implement to make it harder for experimenters to make these intuitive mistakes.

CCS CONCEPTS

General and Reference → Cross-computing tools and techniques → Experimentation; Mathematics of computing → Probability and statistics → Probabilistic inference problems → Hypothesis testing and confidence interval computation

KEYWORDS

A/B Testing, Controlled experiments, Intuition busters

ACM Reference format:

Ron Kohavi, Alex Deng, Lukas Vermeer. A/B Testing Intuition Busters: Common Misunderstandings in Online Controlled Experiments. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22)*, August 14-18, 2022, Washington DC, USA. ACM, New York, NY, USA. 10 pages. https://doi.org/10.1145/3534678.3539160

1. Introduction

Misinterpretation and abuse of statistical tests, confidence intervals, and statistical power have been decried for decades, yet remain rampant. A key problem is that there are no interpretations of these concepts that are at once simple, intuitive, correct, and foolproof

-- Greenland et al (2016)

A/B tests, or online controlled experiments (see appendix for references), are heavily used in industry to evaluate implementations of ideas, with the larger companies starting over 100 experiment treatments every business day (Gupta, et al. 2019). While the statistics behind controlled experiments are well documented and some pitfalls were shared (Crook, et al. 2009, Dmitriev, Frasca, et al. 2016, Kohavi, Tang and Xu 2020, Dmitriev, Gupta, et al. 2017), we see many erroneous applications and misunderstanding of the statistics, including in books, papers, and software. The appendix shows the impact of these misunderstood concepts in courts and legislation.

The concepts we share appear intuitive yet hide unexpected complexities. Although some amount of abstraction leakage is usually unavoidable (Kluck and Vermeer 2015), our goal is to share these common intuition busters so that experimentation platforms can be designed to make it harder for experimenters to misuse them. Our contributions are as follows:

- We share a collection of important intuition busters. Some well-known commercial vendors of A/B testing software have focused on "intuitive" presentations of results, resulting in incorrect claims to their users instead of addressing their underlying faulty intuitions. We believe that these solutions exacerbate the situation, as they reinforce incorrect intuitions.
- We drill deeply into one non-intuitive result, which to the best of our knowledge has not been studied before: the distribution of the treatment effect under non-uniform assignment to variants. Non-uniform assignments have been suggested in the statistical literature. We highlight several concerns.
- We provide recommendations as well as deployed examples for experimentation platform designers to help address the underlying faulty intuitions identified in our collection.

2. Motivating Example

You win some, you learn some
-- Jason Mraz

GuessTheTest is a website that shares "money-making A/B test case studies." We believe such efforts to share ideas evaluated using A/B tests are useful and should be encouraged. That said,

KDD '22, August 14–18, 2022, Washington, DC, USA.

ACM ISBN 978-1-4503-9385-0/22/08...$15.00.
https://doi.org/10.1145/3534678.3539160

some of the analyses could be improved with the recommendations shared in this paper (indeed, some were already integrated into the web site based on feedback from one of the authors). This site is not unique and represents common industry practices in sharing ideas. We are using it as a concrete example that shows several patterns where the industry can improve.

A real A/B test was shared on December 16, 2021, in *GuessTheTest's* newsletter and website with the title: "Which design radically increased conversions 337%?" (O'Malley 2021). The A/B test described two landing pages for a website (the specific change is not important). The test ran for 35 days, and traffic was split 50%/50% for maximum statistical power. The surprising results are shown in Table 1 below.

Table 1: Results of a real A/B Test

Variant	Visitors	Conversions	Conversion rate	Lift
Control	82	3	3.7%	--
Treatment	75	12	16.0%	337%

The analysis showed a massive lift of 337% for the Treatment with a p-value of 0.009 (using Fisher's exact test, which is more appropriate for small numbers, the p-value is 0.013), which the article said is "far below the standard < 0.05 cut-off," and with observed power of 97%, "well beyond the accepted 80% minimum."

Given the data presented, we strongly believe that this result should not be trusted, and we hope to convince the readers and improve industry best practices so that similar experiment results will not be shared without additional validation. Based on our feedback and feedback from others, GuessTheTest added that the experiment was underpowered and suggested doing a replication run.

3. Surprising Results Require Strong Evidence—Lower P-Values

Extraordinary claims require extraordinary evidence" (ECREE)
-- Carl Sagan

Surprising results make great story headlines and are often remembered even when flaws are found, or the results do not replicate. Many of the most cited psychology findings failed to replicate (Open Science Collaboration 2015). Recently, the term Bernoulli's Fallacy has been used to describe the issue as a "logical flaw in the statistical methods" (Clayton 2021).

While controlled experiments are the gold standard in science for claiming causality, many people misunderstand p-values. A very common misunderstanding is that a statistically significant result with p-value 0.05 has a 5% chance of being a false positive (Goodman 2008, Greenland, Senn, et al. 2016, Vickers 2009). A common alternative to p-values used by commercial vendors is "confidence," which is defined as (1-p-value)*100%, and often misinterpreted as the probability that the result is a true positive.

Vendors who sell A/B testing software and should know better, get this concept wrong. For example, Optimizely's documentation equates p-value of 0.10 with "10% error rate" (Optimizely 2022):

> ...to determine whether your results are statistically significant: how confident you can be that the results actually reflect a change in your visitors' behavior, not just noise or randomness... In statistical terms, it's 1-[p-value]. If you set a significance threshold of 90%...you can expect a 10% error rate.

Book authors about A/B Testing also get it wrong. The book *A/B Testing: The Most Powerful Way to Turn Clicks Into Customers* (Siroker and Koomen 2013) incorrectly defines p-value:

> ...we can compute the probability that our observed difference (-0.007) is due to random chance. This value, called the p-value...

The book *You Should Test That: Conversion Optimization for More Leads, Sales and Profit* (Goward 2012) incorrectly states

> ...when statistical significance (that is, it's unlikely the test results are due to chance) has been achieved.

Even Andrew Gelman, a Statistics professor at Columbia University, has gotten it wrong in one of his published papers (due to an editorial change) and apologized (Gelman 2014).

The above examples, and several more in the appendix, show that p-values and confidence are often misunderstood, even among experts who should know better. What is the p-value then? The p-value is the probability of obtaining a result equal to or more extreme than what was observed, assuming that *all* the modeling assumptions, including the null hypothesis, H_0, are true (Greenland, Senn, et al. 2016). Conditioning [1] on the null hypothesis is critical and most often misunderstood. In probabilistic terms, we have

$$\text{p-value} = P(\Delta \text{ observed or more extreme} | H_0 \text{ is true}) .$$

This conditional probability is not what is being described in the examples above. All the explanations above are variations of the opposite conditional probability: what is the probability of the null hypothesis given the delta observed:

$$P(H_0 \text{ is true } | \Delta \text{ observed})$$

Bayes Rule can be used for inverting between these two, but the crux of the problem is that it requires the prior probability of the null hypothesis. Colquhoun (2017) makes a similar point and writes that "we hardly ever have a valid value for this prior." However, in companies running online controlled experiments at scale, we *can* construct good prior estimates based on historical experiments.

One useful metric to look at is the False Positive Risk (FPR), which is the probability that the statistically significant result is a false positive, or the probability that H_0 is true (no real effect) when the

[1] Some authors prefer to use the semicolon notation; see discussion at: https://statmodeling.stat.columbia.edu/2013/03/12/misunderstanding-the-p-value/#comment-143481

test was statistically significant (Colquhoun 2017). Using the following terminology:

- **SS** is a statistically significant result
- α is the threshold used to determine statistical significance (SS), commonly 0.05 for a two-tailed t-test.
- β is the type-II error (usually 0.2 for 80% power)
- π is the prior probability of the null hypothesis, that is $P(H_0)$

Using Bayes Rule, we can derive the following (Wacholder, et al. 2004, Ioannidis 2005, Kohavi, Deng and Longbotham, et al. 2014, Benjamin, et al. 2017):

$$P(H_0|SS) = P(SS|H_0) * \frac{P(H_0)}{P(SS)}$$

$$= \frac{P(SS|H_0)*P(H_0)}{P(SS|H_0)*P(H_0)+P(SS|\neg H_0)*P(\neg H_0)}$$

$$= \frac{\alpha * \pi}{\alpha * \pi + (1-\beta) * (1-\pi)}$$

Several estimates of historical success rates (what the org believes true improvements to the Overall Evaluation Criterion) have been published. These numbers may involve different accounting schemes, and we never know the true rates, but they suffice as ballpark estimates. The table below summarizes the corresponding implied FPR, assuming $\pi = 1 - \text{success-rate}$, experiments were properly powered at 80%, and using a p-value of 0.05 but plugging in 0.025 into the above formula because only statistically significant improvements are considered successful in two-tailed t-tests. In practice, some results will have a significantly lower p-value than the threshold, and those have a lower FPR, while results close to the threshold have a higher FPR, as this is the overall FPR for p-value <= 0.05 in a two-tailed t-test (Goodman and Greenland 2007). Also, other factors like multiple variants, iterating on ideas several times, and flexibility in data processing increase the FPR due to multiple hypothesis testing.

What Table 2 summarizes is how much more likely it is to have a false positive stat-sig result than what people intuitively think. Moving from the industry standard of 0.05 to 0.01 or 0.005 aligns with the threshold suggested by the 72-author paper (Benjamin, et al. 2017) for "claims of new discoveries." Finally, if the result of an experiment is highly unusual or surprising, one should invoke Twyman's law—any figure that looks interesting or different is usually wrong (Kohavi, Tang and Xu 2020)—and only accept the result if the p-value is very low.

In our motivating example, the lift to overall conversion was over 300%. We have been involved in tens of thousands of A/B tests that ran at Airbnb, Booking, Amazon, and Microsoft, and have never seen any change that improves conversions anywhere near this amount. We think it's appropriate to invoke Twyman's law here. In the next section, we show that the pre-experiment power is about 3% (highly under-powered). Plugging that number in, even with the highest success rate of 33% from Table 2, we end up with an FPR of 63%, so likely to be false. Alternatively, to override such low power, if we want the false positive probability, $P(H_0|SS)$ to be 0.05, we would need to set the p-value threshold as follows:

$$\alpha/2 = \frac{0.05 * (1-\beta) * (1-\pi)}{0.95 * \pi}$$

= 0.0016, much lower than the 0.009 reported.

Table 2: False Positive Risk given the Success Rate, p-value threshold of 0.025 (successes only), and 80% power

Company/ Source	Success Rate	**FPR**	Reference
Microsoft	33%	5.9%	(Kohavi, Crook and Longbotham 2009)
Avinash Kaushik	20%	11.1%	(Kaushik 2006)
Bing	15%	15.0%	(Kohavi, Deng and Longbotham, et al. 2014)
Booking.com, Google Ads, Netflix	10%	22.0%	(Manzi 2012, Thomke, Experimentation Works: The Surprising Power of Business Experiments 2020, Moran 2007)
Airbnb Search	8%	26.4%	https://www.linkedin.com/in/ronnyk[2]

We recommend that experimentation platforms show the FPR or estimates of the posterior probability in addition to p-values, and that surprising results be replicated. At Microsoft, the experimentation platform, ExP, provides estimates that the treatment effect is not zero using Bayes Rule with priors from historical data. In other organizations, FPR was used to set α.

4. Experiments with Low Statistical Power NOT Trustworthy

When I finally stumbled onto power analysis...
it was as if I had died and gone to heaven
-- Jacob Cohen (1990)

Statistical power is the probability of detecting a meaningful difference between the variants when there really is one, that is, rejecting the null when there is a true difference of δ. When running controlled experiments, it is recommended that we pick the sample size to have sufficient statistical power to detect a minimum delta of interest. With an industry standard power of 80%, and p-value threshold of 0.05, the sample size for each of two equally sized variants can be determined by this simple formula (van Belle 2002):

$$n = \frac{16\,\sigma^2}{\delta^2}$$

Where n is the number of users in each variant, and the variants are assumed to be of equal size, σ^2 is the variance of the metric of

[2] Permission to include statistic was given by Airbnb

interest, and δ is the sensitivity, or the minimum amount of change you want to detect.

The derivation of the formula is useful for the rest of the section and the next section, so we will summarize its derivation (van Belle 2002). Given two variants of size n each with a standard deviation of σ, we reject the null hypothesis that there is no difference between Control and Treatment (treatment effect is zero) if the observed value is larger than $Z_{1-\alpha/2}$*SE (e.g., $Z_{1-\alpha/2}$ for $\alpha = 0.05$ in a two-tailed test is $Z_{0.975} = 1.96$); SE, the standard error for the difference is $\sigma\sqrt{2/n}$. We similarly reject the alternative hypothesis that the difference is δ if the observed value is smaller than $Z_{1-\beta}$*SE from δ. (Without loss of generality, we evaluate the left tail of a normal distribution centered on a positive δ as the alternative; the same mirror computation can be made with a normal centered on $-\delta$.) The critical value is, therefore, when these two rejection criteria are equal (the approximation ignores rejection based on the wrong tail, sometimes called type III error, a very reasonable and common approximation):

$$Z_{1-\alpha/2} * \text{SE} = \delta - Z_{1-\beta}*\text{SE} \quad \text{Equation 1}$$

$$\text{SE} = \delta/(Z_{1-\beta} + Z_{1-\alpha/2}) \quad \text{Equation 2}$$

$$\sigma\sqrt{2/n} = \delta/(Z_{1-\beta} + Z_{1-\alpha/2})$$

$$n = 2\sigma^2(Z_{1-\beta} + Z_{1-\alpha/2})^2/\delta^2$$

For 80% power, $\beta = 0.2$, $Z_{1-\beta} = 0.84$, and $Z_{1-\alpha/2} = 1.96$, so the numerator is $15.68\sigma^2$, conservatively rounded to 16. Another way to look at Equation 2, is that with 80% power, the detectable effect, δ, is 2.8SE (0.84SE+1.96SE).

From our *GuessTheTest* motivating example, a conservative pre-test statistical power calculation would be to detect a 10% relative change. In Optimizely's survey (2021) of 808 companies, about half said experimentation drove 10% uplift in revenue over time from multiple experiments. At Bing, monthly improvements in revenue from multiple experiments were usually in the low single digits (Kohavi, Tang and Xu 2020, Figure 1.4). A large relative percentage, such as 10% for a single experiment, is conservative in that it will require a smaller sample than attempting to detect smaller changes. Assuming historical data showed 3.7% as the conversion rate (what we see for Control), we can plug-in

$$\sigma^2 = p * (1 - p) = 3.7\% * (1 - 3.7\%) = 3.563\% \text{ and}$$

$$\delta = 3.7\% * 10\% = 0.37\%$$

The sample size recommended for each variant to achieve 80% power is therefore:

$$16\sigma^2/\delta^2 = 16 * 3.563\%/(0.37\%)^2 = 41{,}642\,.$$

The above-mentioned test was run with about 80 users per variant, and thus grossly underpowered even for detecting a large 10% change.

The power for detecting a 10% relative change with 80 users in this example is 3% (formula in the next section). With so little power, the experiment is meaningless. Gelman et al. (2014) show that when power goes below 0.1, the probability of getting the sign wrong (e.g., concluding that the effect is positive when it is in fact negative) approaches 50% as shown in Figure 1.

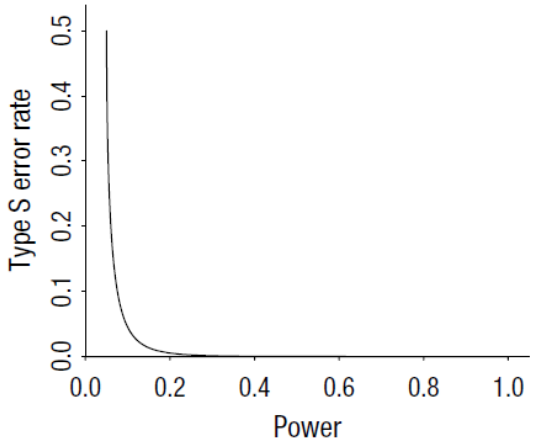

Figure 1: Type S (sign) error of the treatment effect as a function of statistical power (Gelman and Carlin 2014)

The general guidance is that A/B tests are useful to detect effects of reasonable magnitudes when you have, at least, thousands of active users, preferably tens of thousands (Kohavi, Deng and Frasca, et al. 2013).

Table 3 shows the False Positive Risk (FPR) for different levels of power. Running experiments at 20% power with similar success rate to Booking.com, Google ads, Netflix, or Airbnb search, more than half of your statistically significant results will be false positives!

Table 3: False Positive Risk as in Table 2, but with 80% power, 50% power, and 20% power

Company/ Source	Success Rate	FPR @ 80% Power	FPR @ 50% Power	FPR @ 20% Power
Microsoft	33%	5.9%	9.1%	20.0%
Avinash Kaushik	20%	11.1%	16.7%	33.3%
Bing	15%	15.0%	22.1%	41.5%
Booking.com, Google Ads, Netflix	10%	22.0%	31.0%	52.9%
Airbnb search	8%	26.4%	36.5%	59.0%

Ioannidis (2005) made this point in a highly cited paper: *Why Most Published Research Findings Are False*. With many low statistical power studies published, we should expect many false positives when studies show statistically significant results. Moreover, power is just one factor; other factors that can lead to incorrect findings include: flexibility in designs, financial incentives, and simply multiple hypothesis testing. Even if there is no ethical concern, many researchers are effectively p-hacking.

A seminal analysis of 78 articles in the Journal *Abnormal and Social Psychology* during 1960 and 1961 showed that researchers had only 50% power to detect medium-sized effects and only 20% power to detect small effects (Cohen 1962). With such low power, it is no wonder that published results are often wrong or exaggerated. In a superb paper by Button et al. (2013), the authors analyzed 48 articles that included meta-analyses in the neuroscience domain. Based on these meta-analyses, which evaluated 730 individual studies published, they were able to assess the key parameters for statistical power. Their conclusion: the median statistical power in neuroscience is conservatively estimated at 21%. With such low power, many false positive results are to be expected, and many true effects are likely to be missed!

The Open Science Collaboration (2015) attempted to replicate 100 studies from three major psychology journals, where studies typically have low statistical power. Of these, only 36% had significant results compared to 97% in the original studies.

When the power is low, the probability of detecting a true effect is small, but another consequence of low power, which is often unrecognized, is that a statistically significant finding with low power is likely to highly exaggerate the size of the effect. The winner's curse says that the "lucky" experimenter who finds an effect in a low power setting, or through repeated tests, is cursed by finding an inflated effect (Lee and Shen 2018, Zöllner and Pritchard 2007, Deng, et al. 2021). For studies in neuroscience, where power is usually in the range of 8% to 31%, initial treatment effects found are estimated to be inflated by 25% to 50% (Button, et al. 2013).

Gelman and Carlin (2014) show that when power is below 50%, the exaggeration ratio, defined as the expectation of the absolute value of the estimate, divided by the true effect size, becomes so high as to be meaningless, as shown in Figure 2.

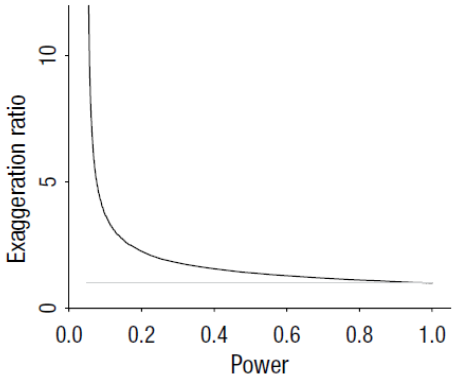

Figure 2: Exaggeration ratio as a function of statistical power (Gelman and Carlin 2014)

Our recommendation is that experimentation platforms should discourage experimenters from starting underpowered experiments. With high probability, nothing statistically significant will be found, and in the unlikely case (e.g., by multiple running iterations) a statistically significant result is obtained, it is likely to be a false positive with an overestimated effect size.

5. Post-hoc Power Calculations are Noisy and Misleading

This power is what I mean when I talk of reasoning backward
-- Sherlock Holmes, A Study in Scarlet

Given an observed treatment effect δ, one can assume that it is the true effect and compute the "observed power" or "post-hoc power" from Equation 1 above as follows:

$$Z_{1-\beta}\text{*SE} = \delta - Z_{1-\alpha/2} * \text{SE}$$
$$Z_{1-\beta} = \delta/\text{SE} - Z_{1-\alpha/2}$$
$$1-\beta = \Phi(\delta/\text{SE} - Z_{1-\alpha/2})$$

The term δ/SE is the observed Z-value used for the test statistic. It is hence $Z_{1-pval/2}$, and we can derive the ad-hoc power as

$$1-\beta = \Phi(Z_{1-pval/2} - Z_{1-\alpha/2}).$$

Note that power is thus fully determined by the p-value and α, and the graph is shown in Figure 3. If the p-value is greater than 0.05, then the power is less than 50% (technically as noted above, this ignores type-III errors, which are tiny).

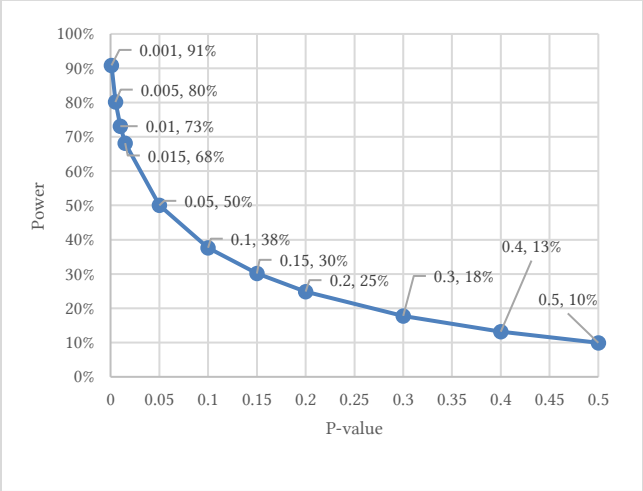

Figure 3: post-hoc power is determined by p-value

In our motivating example, the p-value was 0.009, translating into Z of 2.61. Subtracting 1.96 gives 0.65, which translates into 74% post-power, which may seem reasonable.

However, compare this number to the calculation in Section 4, where the pre-experiment power was estimated at 3%. In low-power experiments, the p-value has enormous variation, and translating it into post-hoc power results in a very noisy estimate (a video of p-values in a low power simulation is at https://tiny.cc/dancepvals). Gelman (2019) wrote that "using observed estimated of effect size is too noisy to be useful." Greenland (2012) wrote: "for a study as completed (observed), it is analogous to giving odds on a horse race after seeing the outcome"

and "post hoc power is unsalvageable as an analytic tool, despite any value it has for study planning."

A key use of statistical power is to claim that for a non-significant result, the true treatment effect is bounded by a small region of $\mp\varepsilon$ because otherwise there is a high probability (e.g., 80%) that the observation would have been significant. This claim holds true for pre-experiment power calculations, but it fails spectacularly for post-hoc, or observed power, calculations.
In *The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis* (Hoenig and Heisey 2001), the authors share what they call a "fatal logical flow" and the "power approach paradox" (PAP). Suppose two experiments gave rise to nonrejected null hypotheses, and the observed power was larger in the first than the second. The intuitive interpretation is that the first experiment gives stronger support favoring the null hypothesis, as with high power, failure to reject the null hypothesis implies that it is probably true. However, this interpretation is only correct for pre-experiment power. As shown above, post-hoc power is determined by the p-value and α, so the first experiment has a *lower* p-value, providing stronger support *against* the null hypothesis!

Experimenters who get a non-significant result will sometimes do a post-hoc power analysis and write something like this: the non-significant result is due to a small sample size, as our power was only 30%. This claim implies that they believe they have made a type-II error and if only they had a larger sample, the null would be rejected. This is catch-22—the claim cannot be made from the data using post-hoc power, as a non-significant result will *always* translate to low post-hoc power.

Given the strong evidence that post-hoc power is a noisy and misleading tool, we strongly recommend that experimentation systems (e.g., https://abtestguide.com/calc) not show it at all. Instead, if power calculations are desired, such systems should encourage their users to pre-register the minimum effect size of interest ahead of experiment execution, and then base their calculations on this input rather than the observed effect size. At Booking.com, the deployed experimentation platform—Experiment Tool—asks users to enter this information when creating a new experiment.

6. Minimize Data Processing Options in Experimentation Platforms

Statistician: you have already calculated the p-value?
Surgeon: yes, I used multinomial logistic regression.
Statistician: Really? How did you come up with that?
Surgeon: I tried each analysis on the statistical software dropdown menus, and that was the one that gave the smallest p-value
-- Andrew Vickers (2009)

In an executive review, a group presented an idea that, they said, was evaluated in an A/B test and resulted in a significant increase to a key business metric. When one of us (Kohavi) asked to see the scorecard, and the metric's p-value was far from significant. Why did you say it was statistically significant, he asked? The response was that it was statistically significant once you turn on the option for extreme outlier removal. We had inadvertently allowed users to do multiple-comparisons and inflate type-I error rates.
Outlier removal must be blind to the hypothesis. André (2021) showed that outlier removal within a variant (e.g., removal of the 1% extreme values, determined for each variant separately), rather than across the data, can result in false-positive rates as high as 43%.

Optimizely's initial A/B system was showing near-real-time results, so their users peeked at the data and chose to stop when it was statistically significant, a procedure recommended by the company at the time. This type of multiple testing significantly inflates the type-I error rates (Johari, et al. 2017).

Flexibility in data collection, analysis, and reporting dramatically increases actual false-positive rates (Simmons, Nelson and Simonsohn 2011). The culprit is researcher *degrees of freedom*, which include:

1. Should more data be collected, or should we stop now?
2. Should some observations be excluded (e.g., outliers, bots)?
3. Segmentation by variables (e.g., gender, age, geography) and reporting just those as statistically significant.

The authors write that "In fact, it is unacceptably easy to publish 'statistically significant' evidence consistent with *any* hypothesis."

Gelman and Loken (2014) discuss how data-dependent analysis, called the "garden of forking paths," leads to statistically significant comparisons that do not hold up. Even without intentional p-hacking, researchers make multiple choices that lead to a multiple-comparison problem and inflate type-I errors. For example, Bem's paper (2011) providing evidence of extrasensory perception (ESP) presented nine different experiments and had multiple degrees of freedom that allowed him to keep looking until he could find what he was searching for. The author found statistically significant results for erotic pictures, but performance could have been better overall, or for non-erotic pictures, or perhaps erotic pictures for men but not women. If results were better in the second half, one could claim evidence of learning; if it's the opposite, one could claim fatigue.

For research, preregistration seems like a simple solution, and organizations like the Center for Open Science support such preregistrations.

For experimentation systems, we recommend that data processing should be standardized. If there is a reason to modify the standard process, for example, outlier removal, it should be pre-specified as part of the experiment configuration and there should be an audit trail of changes to the configuration, as is done at Booking.com. Finally, the benefit in doing A/B testing in software is that replication is much cheaper and easier. If insight leads to a new hypothesis about an interesting segment, pre-register it and run a replication study.

7. Beware of Unequal Variants

The difference between theory and practice is larger in practice than the difference between theory and practice in theory
-- Benjamin Brewster

In theory, a single control can be shared with several treatments, and the theory says that a larger control will be beneficial to reduce the variance (Tang, et al. 2010). Assuming equal variances, the effective sample size of a two-sample test is the harmonic mean $1/(\frac{1}{N_T}+\frac{1}{N_C})$. When there is one control taking a proportion x of users and k equally sized treatments with size $\frac{1-x}{k}$, the optimal control size should be chosen by minimizing the sum $\frac{k}{1-x}+\frac{1}{x}$. We differentiate to get

$$\frac{k}{(1-x)^2}-\frac{1}{x^2}.$$

The optimal control proportion x is the positive solution to

$$(k-1)x^2+2x-1=0\text{ , which is }\frac{1}{\sqrt{k}+1}.$$

For example, when k = 3, instead of using 25% of users for all four variants, we could use 36.6% for control and 21.1% for the treatments, making control more than 1.5x larger. When k = 9, control would get 25% and each treatment only 8.3%, making control 3 times the size of treatment.

Ramp-up is another scenario leading to more extreme unequal treatment vs. control sample size. When a treatment starts at a small percentage, say 2%, the remaining 98% traffic may seem to be the obvious control.

There are several reasons why this seemingly intuitive direction fails in practice:

1. Triggering. As organizations scale experimentation, they run more triggered experiments, which give a stronger signal for smaller populations, great for testing initial ideas and for machine learning classifiers (Kohavi, Tang and Xu 2020, Chapter 20, Triggering). It is practically too hard to share a control and compute for each treatment whether to trigger, especially for experiment treatments that start at different times and introduce performance overhead (e.g., doing inference on both control and treatment to determine if the results differ in order to trigger).
2. Because of cookie churn, unequal variants will cause a larger percentage of users in the smaller variants to be contaminated and be exposed to different variants (their probability of being re-randomized into a larger variant is higher than to their original variant). If there are mechanisms to map multiple cookies to users (e.g., based on logins), this mapping will cause sample-ratio mismatches (Kohavi, Tang and Xu 2020, Fabijan, et al. 2019).
3. Shared resources, such as Least Recently Used (LRU) caches will have more cache entries for the larger variant, giving it a performance advantage (Kohavi, Tang and Xu 2020).

Here we raise awareness of an important statistical issue mentioned in passing by Kohavi et al (2012). When distributions are skewed, in an unequal assignment, the t-test cannot maintain the nominal Type-I error rate on both tails. When a metric is positively skewed, and the control is larger than the treatment, the t-test will over-estimate the Type-I error on one tail and under-estimate on the other tail because the skewed distribution convergence to normal is different. But when equal sample sizes are used, the convergence is similar and the Δ(observed delta) is represented well by a Normal- or t-distribution.

Two common sources of skewness are 1) heavy-tailed measurements such as revenue and counts, often zero-inflated at the same time; and 2) binary/conversion metric with very small positive rate. We ran two simulated A/A studies. In the first study, we drew 100,000 random samples from a heavy-tailed distribution, D1, of counts, like nights booked at a reservation site. This distribution is both zero inflated (about 5% nonzero) and a skewed non-zero component, with a skewness of 35. The second study drew 1,000,000 samples from a Bernoulli distribution, D2, with a small p of 0.01%, which implies a skewness of 100.

In each study, we allocated 10% samples to the treatment. We then compared two cases: in one, the control also allocated 10%; in the second, the remaining 90% were allocated to the control. We did 10,000 simulation trials and counted number of times H_0 was rejected at the right tail and left tail at 2.5% level for each side (5% two-sided). Skewness of Δ and metric value from the 10% treatment group are also reported.

Table 4 shows the results with the following observations:

1. The realized Type-I error is close to the nominal 2.5% rate when control is the same size as treatment.
2. When control is larger, Type-I error at the left tail is greater than 2.5%, while smaller than 2.5% at the right tail.
3. Skewness of the Δ is very close to 0 when control and treatment are equally sized. It is closer to the skewness of treatment metric when control is much larger.

Table 4: Type I errors at left and right tails from 10,000 simulation runs for two skewed distributions

Distri-bution	Variants	Type-I Left tail	Type-I Right tail	Skewness ofΔ	Skewness of 10% variant
D1	10%/10%	2.35%	2.30%	0.0142	0.36
	10%/90%	5.42%	0.85%	0.2817	0.36
D2	10%/10%	2.63%	2.63%	-0.0018	0.32
	10%/90%	5.75%	0.96%	0.2745	0.32

Skewness of a metric decreases with the rate of $\sqrt{n}$ as the sample size increases. Kohavi, Deng, et al. (2014) recommended that sample sizes for each variant large enough such that the skewness of metrics be no greater than $1/\sqrt{355}$ = 0.053. Because the skewness of Δ is more critical for the t-test, note how in equally sized variants, the skewness is materially smaller. Table 4 shows that even when the skewness of the metric itself is above 0.3, the skewness of these Δ for equal sized cases were all smaller than 0.053. Because the ratio of skewness is so high (e.g., 0.2817/0.0142=~19.8), achieving the same skewness, that is, convergence to normal, with unequal variants requires $19.8^2 \approx 400$ times more users.

For experiment ramp-up, where the focus is to reject at the left tail so we can avoid degradation of experiences to users, using a much larger control can lead to higher-than-expected false rejections, so a correction should be applied (Boos and Hughes-Oliver 2000). For using a shared control to increase statistical power, the real statistical power can be lower than the expected. For a number of treatments ranging from two to four, the reduced variance from using the optimal shared control size is less than 10%. We do not think this benefit justifies all the potential issues with unequal variants, and therefore recommend against the use of a large (shared) control.

8. Summary

We shared five seemingly intuitive concepts that are heavily touted in the industry, but are very misleading. We then shared our recommendations for how to design experimentation platforms to make it harder for experimenters to be misled by these. The recommendations were implemented in some of the deployed platforms in our organizations.

ACKNOWLEDGMENTS

We thank Georgi Georgiev, Somit Gupta, Roger Longbotham, Deborah O'Malley, John Cutler, Pavel Dmitriev, Aleksander Fabijan, Matt Gershoff, Adam Gustafson, Bertil Hatt, Michael Hochster, Paul Raff, Andre Richter, Nathaniel Stevens, Wolfe Styke, and Eduardo Zambrano for valuable feedback.

9. References

André, Quentin. 2021. "Outlier exclusion procedures must be blind to the researcher's hypothesis." *Journal of Experimental Psychology: General.* doi:https://psycnet.apa.org/doi/10.1037/xge0001069.

Bem, Daryl J. 2011. "Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect." *Journal of Personality and Social Psychology* 100 (3): 407-425. doi:https://psycnet.apa.org/doi/10.1037/a0021524.

Benjamin, Daniel J., James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, et al. 2017. "Redefine Statistical Significance." *Nature Human Behaviour* 2 (1): 6-10. https://www.nature.com/articles/s41562-017-0189-z.

Boos, Dennis D, and Jacqueline M Hughes-Oliver. 2000. "How Large Does n Have to be for Z and t Intervals?" *The American Statistician*, 121-128.

Button, Katherine S, John P.A. Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma S.J. Robinson, and Marcus R Munafò. 2013. "Power failure: why small sample size undermines the reliability of neuroscience." *Nature Reviews Neuroscience* 14: 365-376. https://doi.org/10.1038/nrn3475.

Clayton, Aubrey. 2021. *Bernoulli's Fallacy: Statistical Illogic and the Crisis of Modern Science.* Columbia University Press.

Cohen, Jacob. 1962. "The Statistical Power for Abnormal-Social Psychological Research: A Review." *Journal of Abnormal and Social Psychology* 65 (3): 145-153. https://psycnet.apa.org/doi/10.1037/h0045186.

Cohen, Jacob. 1990. "Things I have Learned (So Far)." *American Psychologist* 45 (12): 1304-1312. https://www.academia.edu/1527968/Things_I_Have_Learned_So_Far_.

Colquhoun, David. 2017. "The reproducibility of research and the misinterpretation of p-values." *Royal Society Open Science* (4). https://doi.org/10.1098/rsos.171085.

Crook, Thomas, Brian Frasca, Ron Kohavi, and Roger Longbotham. 2009. "Seven Pitfalls to Avoid when Running Controlled Experiments on the Web." *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1105-1114.

Deng, Alex, Yicheng Li, Jiannan Lu, and Vivek Ramamurthy. 2021. "On Post-Selection Inference in A/B Tests." *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2743-2752.

Dmitriev, Pavel, Brian Frasca, Somit Gupta, Ron Kohavi, and Garnet Vaz. 2016. "Pitfalls of long-term online controlled experiments." *IEEE International Conference on Big Data.* Washington, DC. 1367-1376. doi:https://doi.org/10.1109/BigData.2016.7840744.

Dmitriev, Pavel, Somit Gupta, Dong Woo Kim, and Garnet Vaz. 2017. "A Dirty Dozen: Twelve Common Metric Interpretation Pitfalls in Online Controlled Experiments." *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2017).* Halifax, NS, Canada: ACM. 1427-1436. http://doi.acm.org/10.1145/3097983.3098024.

Fabijan, Aleksander, Jayant Gupchup, Somit Gupta, Jeff Omhover, Wen Qin, Lukas Vermeer, and Pavel Dmitriev. 2019. "Diagnosing Sample Ratio Mismatch in Online Controlled Experiments: A Taxonomy and Rules of Thumb for Practitioners." *KDD '19: The 25th SIGKDD International Conference on Knowledge Discovery and Data Mining.* Anchorage, Alaska, USA: ACM.

Gelman, Andrew. 2019. "Don't Calculate Post-hoc Power Using Observed Estimate of Effect Size." *Annals of Surgery* 269 (1): e9-e10. doi:10.1097/SLA.0000000000002908.

—. 2014. "I didn't say that! Part 2." *Statistical Modeling, Causal Inference, and Social Science.* October 14. https://statmodeling.stat.columbia.edu/2014/10/14/didnt-say-part-2/.

Gelman, Andrew, and Eric Loken. 2014. "The Statistical Crisis in Science." *American Scientist* 102 (6): 460-465. doi:https://doi.org/10.1511/2014.111.460.

Gelman, Andrew, and John Carlin. 2014. "Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors." *Perspectives on Psychological Science* 9 (6): 641 –651. doi:10.1177/1745691614551642.

Goodman, Steven. 2008. "A Dirty Dozen: Twelve P-Value Misconceptions." *Seminars in Hematology.* doi:https://doi.org/10.1053/j.seminhematol.2008.04.003.

Goodman, Steven, and Sander Greenland. 2007. *Assessing the unreliability of the medical literature: a response to "Why most published research findings are false".* Johns Hopkins University, Department of Biostatistics. https://biostats.bepress.com/cgi/viewcontent.cgi?article=1135&context=jhubiostat.

Goward, Chris. 2012. *You Should Test That: Conversion Optimization for More Leads, Sales and Profit or The Art and Science of Optimized Marketing*. Sybex.

Greenland, Sander. 2012. "Nonsignificance Plus High Power Does Not Imply Support for the Null Over the Alternative." *Annals of Epidemiology* 22 (5): 364-368. doi:https://doi.org/10.1016/j.annepidem.2012.02.007.

Greenland, Sander, Stephen J Senn, Kenneth J Rothman, John B Carlin, Charles Poole, Steven N Goodman, and Douglas G Altman. 2016. "Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations." *European Journal of Epidemiology* 31: 337-350. https://doi.org/10.1007/s10654-016-0149-3.

Gupta, Somit, Ronny Kohavi, Diane Tang, Ya Xu, Reid Anderson, Eytan Bakshy, Niall Cardin, et al. 2019. "Top Challenges from the first Practical Online Controlled Experiments Summit." 21 (1). https://bit.ly/ControlledExperimentsSummit1.

Hoenig, John M, and Dennis M Heisey. 2001. "The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis." *American Statistical Association* 55 (1): 19-24. doi:https://doi.org/10.1198/000313001300339897.

Ioannidis, John P. 2005. "Why Most Published Research Findings Are False." *PLoS Medicine* 2 (8): e124. doi:10.1371/journal.pmed.0020124.

Johari, Ramesh, Leonid Pekelis, Pete Koomen, and David Walsh. 2017. "Peeking at A/B Tests." *KDD '17: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Halifax, NS, Canada: ACM. 1517-1525. doi:https://doi.org/10.1145/3097983.3097992.

Kaushik, Avinash. 2006. "Experimentation and Testing: A Primer." *Occam's Razor by Avinash Kaushik*. May 22.

Kluck, Timo, and Lukas Vermeer. 2015. "Leaky Abstraction In Online Experimentation Platforms: A Conceptual Framework To Categorize Common Challenges." *The Conference on Digital Experimentation (CODE@MIT)*. Boston, MA. https://arxiv.org/abs/1710.00397.

Kohavi, Ron, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann. 2013. "Online Controlled Experiments at Large Scale." *KDD 2013: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. http://bit.ly/ExPScale.

Kohavi, Ron, Alex Deng, Roger Longbotham, and Ya Xu. 2014. "Seven Rules of Thumb for Web Site Experimenters." *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '14)*. http://bit.ly/expRulesOfThumb.

Kohavi, Ron, Diane Tang, and Ya Xu. 2020. *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*. Cambridge University Press. https://experimentguide.com.

Kohavi, Ron, Thomas Crook, and Roger Longbotham. 2009. "Online Experimentation at Microsoft." *Third Workshop on Data Mining Case Studies and Practice Prize*. http://bit.ly/expMicrosoft.

Lee, Minyong R, and Milan Shen. 2018. "Winner's Curse: Bias Estimation for Total Effects of Features in Online Controlled Experiments." *KDD 2018: The 24th ACM Conference on Knowledge Discovery and Data Mining*. London: ACM.

Manzi, Jim. 2012. *Uncontrolled: The Surprising Payoff of Trial-and-Error for Business, Politics, and Society*. Basic Books.

Moran, Mike. 2007. *Do It Wrong Quickly: How the Web Changes the Old Marketing Rules* . IBM Press.

O'Malley, Deborah. 2021. "Which design radically increased conversions 337%?" *GuessTheTest*. December 16. https://guessthetest.com/test/which-design-radically-increased-conversions-337/?referrer=Guessed.

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349 (6251). doi:https://doi.org/10.1126/science.aac4716.

Optimizely. 2022. "Change the statistical significance setting." *Optimizely Help Center*. January 10. https://support.optimizely.com/hc/en-us/articles/4410289762189.

—. 2021. "How to win in the Digital Experience Economy." *Optimizely*. https://www.optimizely.com/insights/digital-experience-economy-report/.

Simmons, Joseph P, Leif D Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22 (11): 1359-1366. https://journals.sagepub.com/doi/full/10.1177/0956797611417632.

Siroker, Dan, and Pete Koomen. 2013. *A/B Testing: The Most Powerful Way to Turn Clicks Into Customers*. Wiley.

Tang, Diane, Ashish Agarwal, Deirdre O'Brien, and Mike Meyer. 2010. "Overlapping Experiment Infrastructure: More, Better, Faster Experimentation." *Proceedings 16th Conference on Knowledge Discovery and Data Mining*. https://ai.google/research/pubs/pub36500.

Thomke, Stefan H. 2020. *Experimentation Works: The Surprising Power of Business Experiments*. Harvard Business Review Press.

van Belle, Gerald. 2002. *Statistical Rules of Thumb*. Wiley-Interscience.

Vickers, Andrew J. 2009. *What is a p-value anyway? 34 Stories to Help You Actually Understand Statistics*. Pearson.

Wacholder, Sholom, Stephen Chanock, Montserrat Garcia-Closas, Nathaniel Rothman, and Laure Elghormli. 2004. "Assessing the Probability That a Positive Report is False: An Approach for Molecular Epidemiology Studies." *Journal of the National Cancer Institute*. https://doi.org/10.1093/jnci/djh075.

Zöllner, Sebastian, and Jonathan K Pritchard. 2007. "Overcoming the Winner's Curse: Estimating Penetrance Parameters from Case-Control Data." *The American Journal of Human Genetics* 80 (4): 605-615. https://doi.org/10.1086/512821.

A/B Testing Intuition Busters: Appendix

Introduction

This appendix provides additional support and useful references to several sections in the main paper.

There are many references for A/B tests, or online controlled experiments (Kohavi, Tang and Xu 2020, Luca and Bazerman 2020, Thomke 2020, Georgiev 2019, Kohavi, Longbotham, et al. 2009, Goward 2012, Siroker and Koomen 2013); (Box, Hunter and Hunter 2005, Imbens and Rubin 2015, Gerber and Green 2012).

Statistical concepts that are misunderstood have not only caused businesses to make incorrect decisions, hurting user experiences and the businesses themselves, but have also resulted in innocent people being convicted of murder and serving years in jail.
In courts, incorrect use of conditional probabilities is called the Prosecutor's fallacy and "The use of p-values can also lead to the prosecutor's fallacy" (Fenton, Neil and Berger 2016). Sally Clark and Angela Cannings were convicted of the murder of their babies, in part based on a claim presented by eminent British pediatrician, Professor Meadow, who incorrectly stated that the chance of two babies dying in those circumstances are 1 in 73 million (Hill 2005). The Royal Statistical Society issued a statement saying that the "figure of 1 in 73 million thus has no statistical basis" and that "This (mis-)interpretation is a serious error of logic known as Prosecutor's Fallacy" (2001).

In the US, right turn on red was studied in the 1970s but "these studies were underpowered" and the differences on key metrics were not statistically significant, so right turn on red was adopted; later studies showed "60% more pedestrians were being run over, and twice as many bicyclists were struck" (Reinhart 2015).

Surprising Results Require Strong Evidence—Lower P-Values

Eliason (2018) shares 16 popular myths that persist despite evidence they are likely false. In the *Belief in the Law of Small Numbers* (Tvesrky and Kahneman 1971), the authors take the reader through intuition busting exercises in statistical power and replication.

Additional examples where concepts are incorrectly stated by people or organizations in the field of A/B testing include:

Until December 2021, Adobe's documentation stated that

> The confidence of an experience or offer represents the probability that the lift of the associated experience/offer over the control experience/offer is "real" (not caused by random chance). Typically, 95% is the recommended level of confidence for the lift to be considered significant.

This statement is wrong and was likely fixed after a LinkedIn post from one of us that highlighted this error.

The book *Designing with Data: Improving the User Experience with A/B Testing* (King, Churchill and Tan 2017) incorrectly states

> p-values represent the probability that the difference you observed is due to random chance

GuessTheTest defined confidence incorrectly (GuessTheTest 2022) as

> A 95% confidence level means there's just a 5% chance the results are due to random factors -- and not the variables that changed within the A/B test

The owner is in the process of updating its definitions based on our feedback.

The web site AB Test Guide (https://abtestguide.com/calc/) uses the following incorrect wording when the tool is used, and the result is statistically significant:

> You can be 95% confident that this result is a consequence of the changes you made and not a result of random chance

The industry standard threshold of 0.05 for p-value is stated in medical guidance (FDA 1998, Kennedy-Shaffer 2017).

Minimize Data Processing Options in Experimentation Platforms

Additional discussion of ESP following up on Bem's paper (2011) are in Schimmack et. al. (2018).

In *Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results* (Silberzhan, et al. 2018), the authors shared how 29 teams involving 61 analysts used the same data set to address the same research question. Analytic approaches varied widely, and estimated effect sizes ranged from 0.89 to 2.93. Twenty teams (69%) found a statistically significant positive effect, and nine teams (31%) did not. Many subjective decisions are part of the data processing and analysis and can materially impact the outcome.

In the online world, we typically deal with a larger number of units than in domains like psychology. Simmons et al. (2011) recommend at least 20 observations per cell, whereas in A/B testing we recommend thousands to tens of thousands of users (Kohavi, Deng, et al. 2013). On the one hand, this larger sample size results in less dramatic swings in p-values because experiments are adequately powered, but on the other hand online experiments offer more opportunities for optional stopping and post-hoc segmentation, which suffer from multiple hypothesis testing.

Resources for Reproducibility

The key tables and simulations are available for reproducibility at https://bit.ly/ABTestingIntuitionBustersExtra .

References

Bem, Daryl J. 2011. "Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect." *Journal of Personality and Social Psychology* 100 (3): 407-425. doi:https://psycnet.apa.org/doi/10.1037/a0021524.

Box, George E.P., J Stuart Hunter, and William G Hunter. 2005. *Statistics for Experimenters: Design, Innovation, and Discovery.* 2nd. John Wiley & Sons, Inc.

Eliason, Nat. 2018. *16 Popular Psychology Myths You Probably Still Believe.* July 2. https://www.nateliason.com/blog/psychology-myths.

FDA. 1998. "E9 Statistical Principles for Clinical Trials." *U.S. Food & Drug Administration.* September. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/e9-statistical-principles-clinical-trials.

Fenton, Norman, Martin Neil, and Daniel Berger. 2016. "Bayes and the Law." *Annual Review of Statistics and Its Application* 3: 51-77. doi:https://doi.org/10.1146/annurev-statistics-041715-033428.

Georgiev, Georgi Zdravkov. 2019. *Statistical Methods in Online A/B Testing: Statistics for data-driven business decisions and risk management in e-commerce.* Independently published. https://www.abtestingstats.com/.

Gerber, Alan S, and Donald P Green. 2012. *Field Experiments: Design, Analysis, and Interpretation.* W. W. Norton & Company.

Goward, Chris. 2012. *You Should Test That: Conversion Optimization for More Leads, Sales and Profit or The Art and Science of Optimized Marketing.* Sybex.

GuessTheTest. 2022. "Confidence." *GuessTheTest.* January 10. https://guessthetest.com/glossary/confidence/.

Hill, Ray. 2005. "Reflections on the cot death cases." *Significance* 13-16. doi:https://doi.org/10.1111/j.1740-9713.2005.00077.x.

Imbens, Guido W, and Donald B Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press.

Kennedy-Shaffer, Lee. 2017. "When the Alpha is the Omega: P-Values, "Substantial Evidence," and the 0.05 Standard at FDA." *Food Drug Law J.* 595-635. https://pubmed.ncbi.nlm.nih.gov/30294197.

King, Rochelle, Elizabeth F Churchill, and Caitlin Tan. 2017. *Designing with Data: Improving the User Experience with A/B Testing.* O'Reilly Media.

Kohavi, Ron, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann. 2013. "Online Controlled Experiments at Large Scale." *KDD 2013: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining.* http://bit.ly/ExPScale.

Kohavi, Ron, Diane Tang, and Ya Xu. 2020. *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing.* Cambridge University Press. https://experimentguide.com.

Kohavi, Ron, Roger Longbotham, Dan Sommerfield, and Randal M. Henne. 2009. "Controlled experiments on the web: survey and practical guide." *Data Mining and Knowledge Discovery* 18: 140-181. http://bit.ly/expSurvey.

Luca, Michael, and Max H Bazerman. 2020. *The Power of Experiments: Decision Making in a Data-Driven World.* The MIT Press.

Reinhart, Alex. 2015. *Statistics Done Wrong: The Woefully Complete Guide.* No Starch Press.

Royal Statistical Society. 2001. *Royal Statistical Society concerned by issues raised in Sally.* London, October 23. https://web.archive.org/web/20110824151124/http://www.rss.org.uk/uploadedfiles/documentlibrary/744.pdf.

Schimmack, Ulrich, Linda Schultz, Rickard Carlsson, and Stefan Schmukle. 2018. "Why the Journal of Personality and Social Psychology Should Retract Article DOI: 10.1037/a0021524 "Feeling the Future: Experimental evidence for anomalous retroactive influences on cognition and affect" by Daryl J. Bem." *Replicability-Index.* January 30. https://replicationindex.com/2018/01/05/bem-retraction/.

Silberzhan, R, E L Uhlmann, D P Martin, and et. al. 2018. "Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results." *Advances in Methods and Practices in Psychological Science* 1 (3): 337-356. doi:https://doi.org/10.1177%2F2515245917747646.

Simmons, Joseph P, Leif D Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22 (11): 1359-1366. https://journals.sagepub.com/doi/full/10.1177/0956797611417632.

Siroker, Dan, and Pete Koomen. 2013. *A/B Testing: The Most Powerful Way to Turn Clicks Into Customers.* Wiley.

Thomke, Stefan H. 2020. *Experimentation Works: The Surprising Power of Business Experiments.* Harvard Business Review Press.

Tvesrky, Amos, and Daniel Kahneman. 1971. "Belief in the Law of Small Numbers." *Psychological Bulletin* 76 (2): 105-110. https://psycnet.apa.org/record/1972-01934-001.