Statistical Power Calculations Revisited: Incorporating Beliefs About Effect Sizes

Melany Gualavisi melanygd@amazon.com
Ryan Kessler rykessle@amazon.com
Lorenzo Masoero masoerl@amazon.com

1. Overview In A/B testing, statistical power depends on both the variance of estimated impacts and the distribution of true impacts. A low variance metric can have low power if true impacts on the metric tend to be small, while a high variance metric can have high power if true impacts on the metric tend to be large.

Traditional power calculations, however, focus solely on the variance of estimated impacts. They compute the probability of detecting a fixed effect size or the smallest effect size that can be detected with high probability (i.e., the "minimum detectable effect" or MDE). While such calculations capture the role of the variance of estimated impacts, they do not provide a way to measure expected power taking into account uncertainty or beliefs about the distribution of true impacts.

In this paper, we present two approaches to connecting power calculations to beliefs about the distribution of true impacts. First, we show how frequentists can compute "prior-informed average power" by taking a weighted average of conventional power across different effect sizes, with weights based on how likely that effect size is believed to occur. Second, we show how Bayesians can compute "Bayesian decision power" by taking a weighted average of the probability of meeting a launch or dial down criteria across different effect sizes, with weights again based on how likely that effect size is believed to occur. When true impacts are assumed to be normally distributed, both approaches yield simple closed-form expressions that can be computed using data readily available in most A/B testing tools.

These approaches enable A/B testing tools to provide more realistic and informative assessments of statistical power. By incorporating beliefs about the distribution of true impacts, they can better inform experiment design decisions such as traffic allocation and duration by leveraging the relative power of different metrics. This is especially valuable given that many large A/B testing tools already estimate beliefs regarding the distribution of true impacts via empirical Bayes methods but rarely leverage them in thinking about power. We provide a simple way to close the gap, aligning power calculations with the same beliefs regarding true impacts used in Bayesian inference.

2. Theory We are interested in measuring the impact of a feature change. We run an A/B test, exposing the feature change to a random subset of traffic. The impact of the feature change Δ is a random variable, whose realized values we denote by δ. The A/B test delivers an estimator $\hat{\Delta}$ of Δ, whose realized values we denote by $\hat{\delta}$, and an estimate τ^2 of its sampling variance, which we treat as a known constant. Motivated by the randomization of the feature change and the central limit theorem, we assume that $\hat{\Delta}$ is normally distributed, with mean δ and variance τ^2:

$$\hat{\Delta} \mid \Delta = \delta, \tau^2 \sim \mathcal{N}(\delta, \tau^2) \tag{1}$$

We assume that the true impact Δ is distributed according to some assumed distribution G.

Given this setup, our goal is to measure our ability to detect impacts, taking into account the sampling variance τ^2 and the likelihood of observing different effect sizes under G.

Frequentist power: Traditional power analyses compute the probability of rejecting the null hypothesis given an assumed effect size $\Delta = \delta$, sampling variance τ^2, and significance level α. For a two-sided t-test for equality in means, the usual formula approximating the power of the test is:

$$\Pi(\delta, \tau^2, \alpha) = 1 - \Phi\left(\Phi^{-1}\left(1 - \alpha/2\right) - \frac{\delta}{\tau}\right) + \Phi\left(-\Phi^{-1}(1 - \alpha/2) - \frac{\delta}{\tau}\right). \tag{2}$$

The primary limitation of eq. (2) is that it conditions on a single assumed effect size $\Delta = \delta$, ignoring the distribution of the true impact G. This can be misleading in two ways. First, power may be significantly understated or overstated if the assumed effect size lies far from likely values under G. Second, it can distort assessments of relative power when the distribution G varies significantly across metrics — for example, when true impacts on some metrics tend to be significantly larger than those on others.

Prior-informed average power: An alternative approach, aimed more toward frequentist practitioners, is to measure average power, marginalizing over the distribution of the true impact G. For any assumed distribution G, sampling variance τ^2, and significance level α, we define "prior-informed average power" as:

$$\bar{\Pi}(G, \tau^2, \alpha) = \int \Pi(\delta, \tau^2, \alpha) \mathrm{d}G(\delta). \tag{3}$$

Equation (3) explicitly links standard frequentist notions of statistical power to the beliefs regarding the distribution of the true impact.

We can estimate prior-informed average power via Monte Carlo simulations as follows:

(i) Draw Δ from G

(ii) Draw $\hat{\Delta}$ from $N(\delta, \tau^2)$ given $\Delta = \delta$

(iii) Compute $\Pi(\delta, \tau^2, \alpha)$

(iv) Repeat steps (i)-(iii) a sufficiently large number of times and estimate $\bar{\Pi}(G, \tau^2, \alpha)$ as the average of $\Pi(\delta, \tau^2, \alpha)$ across replications

We prove in appendix A that in the special case where $G = N(\mu, \sigma^2) = G^N$, $\bar{\Pi}(G, \tau^2, \alpha)$ has a simple closed-form expression:

$$\bar{\Pi}(G^N, \tau^2, \alpha) = 1 - \Phi\left(\frac{\tau\Phi^{-1}(1 - \alpha/2) - \mu}{\sqrt{\tau^2 + \sigma^2}}\right) + \Phi\left(\frac{-\tau\Phi^{-1}(1 - \alpha/2) - \mu}{\sqrt{\tau^2 + \sigma^2}}\right). \tag{4}$$

Bayesian decision power: Another approach, aimed more toward Bayesian experimenters, is to compute the probability of meeting a launch or dial-down criterion for their experiment. This is accomplished by marginalizing over the distribution of the true impact G. We consider again the special case where $G = N(\mu, \sigma^2) = G^N$ and we define $w = \sigma^2(\sigma^2 + \tau^2)^{-1}$. Further assuming normality for the estimated impact $\hat{\Delta} = \hat{\delta}$ and sampling variance τ^2 via the CLT, also the posterior distribution of Δ is Gaussian, with mean $\tilde{\mu} = w\hat{\delta} + (1 - w)\mu$ and variance $\tilde{\sigma}^2 = w\tau^2$. Launch and dial-down decisions can then be made based on whether the posterior probability that Δ is positive, $\Phi(\tilde{\mu}/\tilde{\sigma})$, exceeds threshold α or falls below threshold $1 - \beta$, respectively.

We prove in appendix A that given the distribution of the true impact G, sampling variance τ^2, and launch/dial-down thresholds (α, β), the probability of launching or dialing down (or "Bayesian decision power") is given by:

$$\tilde{\Pi}(G^N, \tau^2, \alpha, \beta) = 1 - \Phi\left(\frac{\mu\sqrt{\tau^2 + \sigma^2}}{\sigma^2} - \frac{\tau\Phi^{-1}(1 - \beta)}{\sigma}\right) + \Phi\left(\frac{\mu\sqrt{\tau^2 + \sigma^2}}{\sigma^2} - \frac{\tau\Phi^{-1}(\alpha)}{\sigma}\right). \tag{5}$$

Equation (5) allows a Bayesian decision maker to directly compute the probability of meeting launch or dial-down criteria while accounting for both measurement precision and beliefs regarding the distribution of the true impact.

During the design phase, experimenters can use prior-informed average power or Bayesian decision power to help better inform design decisions such as how to allocate traffic across treatments and determine experiment duration or how to choose across different metrics.

3. Simulations We illustrate the above concepts via stylized simulations. We assume $G = N(0, \sigma^2) = G^N$ and consider metrics with different specifications of the sampling variance τ^2 and prior variance σ^2. For the first metric (metric A), we consider relatively small sampling variance and effect size, with $\tau = 0.002$ and $\sigma = 0.001$. For the second metric (metric B), we consider relatively large sampling variance and effect size, with $\tau = 0.005$ and $\sigma = 0.01$.

Figure 1 shows that despite metric A's smaller sampling variance (and therefore smaller minimum detectable effect (MDE) size), it achieves lower prior-informed average power and Bayesian decision power than metric B. This is because metric A's true effects, while more precisely estimated, tend to be smaller than those of metric B. This highlights how traditional power analysis, focused solely on sampling variances or MDEs, can mislead practitioners when thinking about the ability to detect impacts.

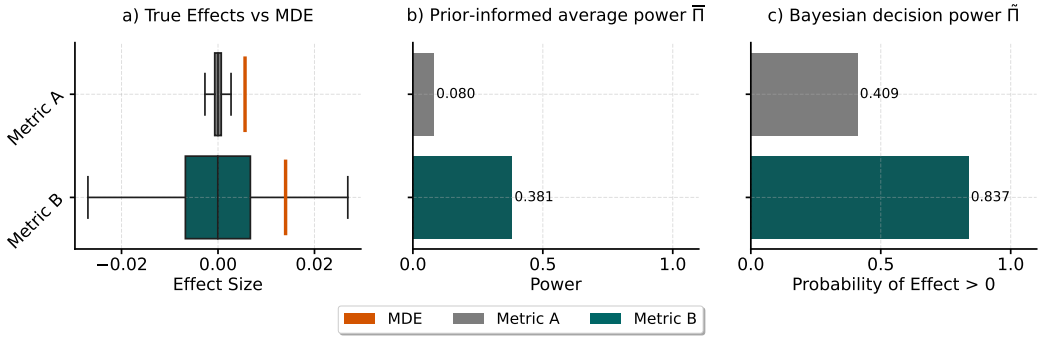

Figure 1: Shortcomings of standard power analysis and proposed alternatives.

4. Conclusions In this paper, we present two approaches to connecting power calculations to beliefs about the distribution of true impacts. The prior-informed average power approach enables frequentists to move beyond MDEs by accounting for prior beliefs about true impacts. The Bayesian decision power framework directly measures the probability of meeting launch criteria while incorporating these same beliefs. Both approaches yield practical tools for more realistic power assessment, enabling experimenters to make better-informed decisions about experiment design.

A Appendix: Proof of Claims

Lemma A.1. *Let Z be a standard normal random variable, with cumulative distribution function $\Phi(z)$ and probability distribution function $\phi(z)$. Then for any $a, b \in \mathbb{R}$:*

$$\int_{-\infty}^{\infty} \Phi(az + b)\phi(z)dz = \Phi\left(\frac{b}{\sqrt{1 + a^2}}\right). \tag{6}$$

Proof. For independent standard normal random variables Z_1 and Z_2:

$$\int_{-\infty}^{\infty} \Phi(az + b)\phi(z)\mathrm{d}z = E(\Phi(aZ_1 + b)) = \Pr(Z_2 \leq aZ_1 + b) = \Pr(Z_2 - aZ_1 \leq b), \tag{7}$$

where $X = Z_2 - aZ_1 \sim N(0, 1 + a^2)$. It follows that:

$$\int_{-\infty}^{\infty} \Phi(az + b)\phi(z)\mathrm{d}z = \Pr\left(\frac{X}{\sqrt{1 + a^2}} \leq \frac{b}{\sqrt{1 + a^2}}\right) = \Phi\left(\frac{b}{\sqrt{1 + a^2}}\right), \tag{8}$$

proving the thesis. □

Claim A.2. *If $G = N(\mu, \sigma^2) = G^N$, then for sampling variance τ^2 and significance level α prior-informed average power is given by:*

$$\bar{\Pi}(G^N, \tau^2, \alpha) = 1 - \Phi\left(\frac{\tau\Phi^{-1}(1-\alpha/2) - \mu}{\sqrt{\tau^2+\sigma^2}}\right) + \Phi\left(\frac{-\tau\Phi^{-1}(1-\alpha/2) - \mu}{\sqrt{\tau^2+\sigma^2}}\right). \tag{9}$$

Proof. Assume $G = G^N$. Then for sampling variance τ^2 and significance level α prior-informed average power is given by:

$$\begin{aligned}
\bar{\Pi}(G^N, \tau^2, \alpha) &= \int_{-\infty}^{\infty} \Pi(\delta, \tau^2, \alpha) \mathrm{d}G^N(\delta) && (10)\\
&= \int_{-\infty}^{\infty} \Pi(\delta, \tau^2, \alpha) \frac{1}{\sigma}\phi\left(\frac{\delta-\mu}{\sigma}\right) \mathrm{d}\delta && (11)\\
&= \int_{-\infty}^{\infty} \left(1 - \Phi\left(\Phi^{-1}(1-\alpha/2) - \frac{\delta}{\tau}\right)\right. \\
&\qquad \left. + \Phi\left(-\Phi^{-1}(1-\alpha/2) - \frac{\delta}{\tau}\right)\right) \frac{1}{\sigma}\phi\left(\frac{\delta-\mu}{\sigma}\right) \mathrm{d}\delta && (12)\\
&= \int_{-\infty}^{\infty} \left(1 - \Phi\left(az+b\right) + \Phi\left(az+c\right)\right) \phi(z) \mathrm{d}z, && (13)
\end{aligned}$$

where the last equality follows from substituting the terms $\delta = z\sigma + \mu$, $a = -\sigma/\tau$, $b = \Phi^{-1}(1 - \alpha/2) - \mu/\tau$, and $c = -\Phi^{-1}(1-\alpha/2) - \mu/\tau$. The claim then follows from lemma A.1. □

Claim A.3. *If $G = N(\mu, \sigma^2) = G^N$, then for sampling variance τ^2 and launch and dial-down thresholds $(\alpha, 1-\beta)$ Bayesian decision power is given by:*

$$\tilde{\Pi}\left(G^N, \tau^2, \alpha, \beta\right) = 1 + \Phi\left(\frac{\mu\sqrt{\tau^2+\sigma^2}}{\sigma^2} - \frac{\tau\Phi^{-1}(\alpha)}{\sigma}\right) - \Phi\left(\frac{\mu\sqrt{\tau^2+\sigma^2}}{\sigma^2} - \frac{\tau\Phi^{-1}(1-\beta)}{\sigma}\right). \tag{14}$$

Proof. Assume $G = G^N$ and define $S = (L + D) \in \{0, 1\}$ to be indicator for whether the launch criteria is met $(L = 1)$ or the dial down criteria is met $(D = 1)$. Then for sampling variance τ^2 and launch and dial-down thresholds $(\alpha, 1-\beta)$ Bayesian decision power is given by:

$$\begin{aligned}
\tilde{\Pi}\left(G^N, \tau^2, \alpha, \beta\right) &= \int_{-\infty}^{\infty} \mathrm{Pr}(S = 1 | \Delta = \delta, \tau^2, G^N, \alpha, \beta) \frac{1}{\sigma}\phi\left(\frac{\delta-\mu}{\sigma}\right) \mathrm{d}\delta && (15)\\
&= \int_{-\infty}^{\infty} \left[\mathrm{Pr}(L = 1 | \Delta = \delta, \tau^2, G^N, \alpha)\right. \\
&\qquad \left. + \mathrm{Pr}(D = 1 | \Delta = \delta, \tau^2, G^N, \beta)\right] \frac{1}{\sigma}\phi\left(\frac{\delta-\mu}{\sigma}\right) \mathrm{d}\delta && (16)
\end{aligned}$$

Next, note that:

$$\begin{aligned}
\mathrm{Pr}\left(L = 1 | \Delta = \delta, \tau^2, G^N, \alpha\right) &= \mathrm{Pr}\left(\Phi\left(\frac{\tilde{\mu}}{\tilde{\sigma}}\right) > \alpha \middle| \Delta = \delta, \tau^2, G^N, \alpha\right) && (17)\\
&= \Phi\left(\frac{\delta + \left(\frac{\tau^2}{\sigma^2}\right)\mu - \tau\sqrt{\frac{\sigma^2+\tau^2}{\sigma^2}}\Phi^{-1}(\alpha)}{\tau}\right). && (18)
\end{aligned}$$

Symmetrically:

$$\Pr\left(D=1|\Delta=\delta,\tau^2,G^N,\beta\right)=\Pr\left(\Phi\left(\frac{\tilde{\mu}}{\tilde{\sigma}}\right)<1-\beta\Big|\Delta=\delta,\tau^2,G^N,\beta\right) \tag{19}$$

$$=1-\Phi\left(\frac{\delta+\left(\frac{\tau^2}{\sigma^2}\right)\mu-\tau\sqrt{\frac{\sigma^2+\tau^2}{\sigma^2}}\Phi^{-1}(1-\beta)}{\tau}\right). \tag{20}$$

Therefore:

$$\Pr(S=1|\Delta=\delta,\tau^2,G^N,\alpha,\beta)=\Phi\left(\frac{\delta+\left(\frac{\tau^2}{\sigma^2}\right)\mu-\tau\sqrt{\frac{\sigma^2+\tau^2}{\sigma^2}}\Phi^{-1}(\alpha)}{\tau}\right) \tag{21}$$

$$+1-\Phi\left(\frac{\delta+\left(\frac{\tau^2}{\sigma^2}\right)\mu-\tau\sqrt{\frac{\sigma^2+\tau^2}{\sigma^2}}\Phi^{-1}(1-\beta)}{\tau}\right). \tag{22}$$

Plugging back into eq. (15) with:

$$\delta=z\sigma+\mu \tag{23}$$

$$a=\sigma/\tau \tag{24}$$

$$b=\frac{1}{\tau}\left(\mu+\mu(\tau^2/\sigma^2)-\tau\sqrt{\frac{\sigma^2+\tau^2}{\sigma^2}}\Phi^{-1}(\alpha)\right) \tag{25}$$

$$c=\frac{1}{\tau}\left(\mu+\mu(\tau^2/\sigma^2)-\tau\sqrt{\frac{\sigma^2+\tau^2}{\sigma^2}}\Phi^{-1}(1-\beta)\right), \tag{26}$$

yields the following:

$$\tilde{\Pi}\left(G^N,\tau^2,\alpha,\beta\right)=\int_{-\infty}^{\infty}\Phi(az+b)\phi(z)\mathrm{d}z+1-\int_{-\infty}^{\infty}\Phi(az+c)\phi(z)\mathrm{d}z. \tag{27}$$

Applying lemma A.1 to both integrals yields:

$$\tilde{\Pi}\left(G^N,\tau^2,\alpha,\beta\right)=\Phi\left(\frac{b}{\sqrt{1+a^2}}\right)+1-\Phi\left(\frac{c}{\sqrt{1+a^2}}\right). \tag{28}$$

Finally, plugging back in for a, b, and c and simplifying yields equation (14). □