Overcoming the Winner's Curse: Leveraging Bayesian Inference to Improve Estimates of the Impact of Features Launched via A/B tests

Ryan Kessler
Amazon.com

Overview: Many data-driven companies measure the impact of product groups and allocate resources across them based on the estimated impacts of features they launch via A/B tests. In this doc, we show that, when based on a standard frequentist estimator of the impact of features, this practice can significantly overstate the impact of product groups and distort the allocation of resources. When this practice is instead based on a Bayesian estimator of the impact of features, there are no such problems when the underlying prior beliefs regarding the distribution of true impacts are correctly specified. To help assess performance of the estimators in practice, we conduct simulations, allowing for different forms of misspecification in prior beliefs regarding the distribution of true impacts. In these simulations, we find that the Bayesian estimator generally outperforms the frequentist estimator, even under certain forms of misspecification. We use both the frequentist and Bayesian estimators to measure cumulative impacts across A/B tests at Amazon, highlighting differences in their overall magnitude and their distribution across product groups.

Setup: We consider a data-driven company that uses A/B tests to determine whether or not to launch new features. The company has a set of candidate features indexed by $w \in \{1, ..., W\}$. For each candidate feature, the company runs an A/B test, exposing the feature to a random subset of traffic to measure the impact Δ_w of the feature on a metric of interest. The A/B test delivers a frequentist estimator $\hat{\Delta}_w$ of Δ_w and an estimator $\hat{\tau}_w^2$ of the sampling variance of $\hat{\Delta}_w$, which we treat as a known constant. The estimator $\hat{\Delta}_w$ could, for example, be the estimated difference in means scaled by the total number of units in the A/B test, with adjustment for pre-determined covariates. We assume that $\hat{\Delta}_w$ is normally distributed, with mean Δ_w and variance $\hat{\tau}_w^2$:

$$\hat{\Delta}_w | \Delta_w, \hat{\tau}_w^2 \overset{\text{ind}}{\sim} N(\Delta_w, \hat{\tau}_w^2) \tag{1}$$

The true impacts Δ_w are distributed according to an unknown distribution G, with hyperparameters $\boldsymbol{\beta}$:

$$\Delta_w | \boldsymbol{\beta} \overset{\text{iid}}{\sim} G(\boldsymbol{\beta}) \tag{2}$$

The company assumes (perhaps naively) that the true impacts Δ_w are normally distributed, with mean μ and variance σ^2:

$$\Delta_w | \mu, \sigma^2 \overset{\text{iid}}{\sim} N(\mu, \sigma^2) \tag{3}$$

Under equations (1) and (3), the true impacts given the A/B test results $(\hat{\Delta}_w, \hat{\tau}_w^2)$ are also normally distributed:

$$\Delta_w | \hat{\Delta}_w, \hat{\tau}_w^2, \mu, \sigma^2 \overset{\text{iid}}{\sim} N(\tilde{\mu}_w(\mu, \sigma^2), \tilde{\sigma}_w^2(\sigma^2)) \tag{4}$$

$$\tilde{\mu}_w(\mu, \sigma^2) = \omega_w(\sigma^2)\hat{\Delta}_w + (1 - \omega_w(\sigma^2))\mu \tag{5}$$

$$\tilde{\sigma}_w^2(\sigma^2) = \omega_w(\sigma^2)\hat{\tau}_w^2 \tag{6}$$

$$\omega_w(\sigma^2) = \sigma^2(\sigma^2 + \hat{\tau}_w^2)^{-1} \tag{7}$$

The company launches a feature if and only if the estimated impact $\hat{\Delta}_w$ is greater than a launch threshold k_w. Let $\phi(\cdot)$ and $\Phi(\cdot)$ be the probability density function and cumulative distribution function of the standard normal distribution, respectively. The launch threshold k_w could, for example, be based on a frequentist decision rule to launch if and only if $\hat{\Delta}_w > 0$ and the p-value is less than α in which case $k_w = \hat{\tau}_w \cdot \Phi^{-1}(1 - \alpha/2)$.

Estimand and estimators: The company's goal is to estimate the cumulative impact of the features it launches:

$$\Delta_{\mathcal{L}} = \sum_{w \in \mathcal{L}} \Delta_w \tag{8}$$

where $\mathcal{L} = \{w \in \{1, ..., W\} : \hat{\Delta}_w > k_w\}$ is the subset of features the company launches given launch thresholds $\boldsymbol{k} = (k_1, ..., k_W)$.

We consider 2, easy-to-compute estimators of $\Delta_{\mathcal{L}}$:

$$\hat{\Delta}_{\mathcal{L}} = \sum_{w \in \mathcal{L}} \hat{\Delta}_w \qquad\qquad \tilde{\mu}_{\mathcal{L}}(\hat{\mu}, \hat{\sigma}^2) = \sum_{w \in \mathcal{L}} \tilde{\mu}_w(\hat{\mu}, \hat{\sigma}^2) \tag{9}$$

The first estimator is the sum of frequentist impacts $\hat{\Delta}_w$ across features the company launches. The second estimator is the sum of Bayesian impacts $\tilde{\mu}_w(\hat{\mu}, \hat{\sigma}^2)$ across features the company launches, given estimates $(\hat{\mu}, \hat{\sigma}^2)$ of (μ, σ^2).

Motivated by equations (1) and (4), we construct (nominal) $(1-\alpha)$ confidence intervals for the estimators via:

$$\hat{\Delta}_{\mathcal{L}} \pm \sqrt{\sum_{w \in \mathcal{L}} \hat{\tau}_w^2} \cdot \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \tag{10}$$

$$\tilde{\mu}_{\mathcal{L}}(\hat{\mu}, \hat{\sigma}^2) \pm \sqrt{\sum_{w \in \mathcal{L}} \tilde{\sigma}_w^2(\hat{\sigma}^2)} \cdot \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \tag{11}$$

exploiting the fact that, given independence across features w, the variance of the sum is equal to the sum of the corresponding variances.

Observations: We explore the bias of $\hat{\Delta}_{\mathcal{L}}$ and $\tilde{\mu}_{\mathcal{L}}(\hat{\mu}, \hat{\sigma}^2)$ and the coverage of their respective confidence intervals. We summarize our findings across 5 main observations, with details relegated to appendix A.

Observation 1: The frequentist estimator $\hat{\Delta}_{\mathcal{L}}$ is biased upwards. If $\hat{\Delta}_w | \Delta_w, \hat{\tau}_w^2 \overset{\text{ind}}{\sim} N(\Delta_w, \hat{\tau}_w^2)$, then:

$$E\left(\hat{\Delta}_{\mathcal{L}} - \Delta_{\mathcal{L}} \middle| \hat{\Delta}_w > k_w \ \text{ for all } \ w \in \mathcal{L}\right) = \sum_{w \in \mathcal{L}} E\left(\hat{\tau}_w \left(\frac{\phi\left(\frac{k_w - \Delta_w}{\hat{\tau}_w}\right)}{1 - \Phi\left(\frac{k_w - \Delta_w}{\hat{\tau}_w}\right)}\right) \middle| \hat{\Delta}_w > k_w\right) > 0 \tag{12}$$

Observation 2: The bias of the frequentist estimator $\hat{\Delta}_{\mathcal{L}}$ is decreasing in statistical power Π_w, with the bias approaching 0 as power approaches $\Pi_w = 1$ for all $w \in \mathcal{L}$.

Observation 3: When statistical power Π_w is low, the $(1-\alpha)$ confidence interval for $\hat{\Delta}_{\mathcal{L}}$ will cover $\Delta_{\mathcal{L}}$ less than $(1-\alpha)$ percent of the time. If $\hat{\Delta}_w | \Delta_w, \hat{\tau}_w^2 \overset{\text{ind}}{\sim} N(\Delta_w, \hat{\tau}_w^2)$ and, for each feature $w \in \mathcal{L}$, $\Pi_w < 0.5$ with probability 1, then:

$$\Pr\left(\left|\hat{\Delta}_{\mathcal{L}} - \Delta_{\mathcal{L}}\right| \leq \sqrt{\sum_{w \in \mathcal{L}} \hat{\tau}_w^2} \cdot \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \middle| \hat{\Delta}_w > k_w \ \text{ for all } \ w \in \mathcal{L}\right) < 1 - \alpha \tag{13}$$

Observation 4: When the company's prior beliefs are correctly specified, the Bayesian estimator $\tilde{\mu}_{\mathcal{L}}(\hat{\mu}, \hat{\sigma}^2)$ is unbiased. If $\hat{\Delta}_w | \Delta_w, \hat{\tau}_w^2 \overset{\text{ind}}{\sim} N(\Delta_w, \hat{\tau}_w^2)$ and $\Delta_w | \mu, \sigma^2 \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, then:

$$E\left(\tilde{\mu}_{\mathcal{L}}(\mu, \sigma^2) - \Delta_{\mathcal{L}} \middle| \hat{\Delta}_w > k_w \ \text{ for all } \ w \in \mathcal{L}\right) = 0 \tag{14}$$

Observation 5: When the company's prior beliefs are correctly specified, the $(1-\alpha)$ confidence interval for $\tilde{\mu}_{\mathcal{L}}(\hat{\mu}, \hat{\sigma}^2)$ will cover $\Delta_{\mathcal{L}}$ $(1-\alpha)$ percent of the time. If $\hat{\Delta}_w | \Delta_w, \hat{\tau}_w^2 \overset{\text{ind}}{\sim} N(\Delta_w, \hat{\tau}_w^2)$ and $\Delta_w | \mu, \sigma^2 \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, then:

$$\Pr\left(\left|\tilde{\mu}_{\mathcal{L}}(\mu, \sigma^2) - \Delta_{\mathcal{L}}\right| \leq \sqrt{\sum_{w \in \mathcal{L}} \tilde{\sigma}_w^2(\sigma^2)} \cdot \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \middle| \hat{\Delta}_w > k_w \ \text{ for all } \ w \in \mathcal{L}\right) = 1 - \alpha \tag{15}$$

Observation 1 formalizes the idea of the so-called "winner's curse" (see, for example, [1], [2], [3]). Intuitively, the company is more likely to launch a feature precisely when frequentist estimates overestimate its impact. This results in an upward bias when using $\hat{\Delta}_{\mathcal{L}}$ to estimate the cumulative impact features that are launched. Observation 2 shows that the magnitude of the winner's curse bias is decreasing in statistical power. Observation 3 shows that, in underpowered A/B tests, the winner's curse also creates challenges for inference, with confidence intervals failing to achieve nominal coverage. Observations 4 and 5 show that the Bayesian estimator $\tilde{\mu}_{\mathcal{L}}(\hat{\mu}, \hat{\sigma}^2)$ eliminates the winner's curse bias and preserves inference on average when the company's prior beliefs are correctly specified. Intuitively, the Bayesian estimator $\tilde{\mu}_{\mathcal{L}}(\hat{\mu}, \hat{\sigma}^2)$ by definition conditions on the results of A/B tests, so further conditioning on meeting the launch thresholds $\boldsymbol{k}$ does not cause bias or distort inference (on average, when the company's prior beliefs are correctly specified).

Simulations: We evaluate the performance of the estimators via Monte Carlo simulations calibrated to mimic A/B tests at Amazon. We obtain the estimated sampling variances $\hat{\tau}_w^2$ for a set of historical Amazon A/B tests $\mathcal{W}^S$. For each A/B test in $\mathcal{W}^S$, we draw from the data generating process (DGP):

$$\Delta_w | M, \nu, s \overset{\text{iid}}{\sim} M + t_\nu \cdot s \tag{16}$$

$$\hat{\Delta}_w | \Delta_w, \hat{\tau}_w^2 \overset{\text{ind}}{\sim} N(\Delta_w, \hat{\tau}_w^2) \tag{17}$$

where M and s are scalars and t_ν is a student-t random variable with ν degrees of freedom. We identify the set of A/B tests $\mathcal{L}$ that meet launch thresholds $\boldsymbol{k}$. Among these A/B tests, we construct the estimators $\hat{\Delta}_{\mathcal{L}}$ and $\tilde{\mu}_{\mathcal{L}}(\hat{\mu}, \hat{\sigma}^2)$, their 95 percent confidence intervals, and the ground truth $\Delta_{\mathcal{L}}$. We repeat this 1K times and estimate bias and coverage via their sample analogues, averaging across the 1K replications and across the set $\mathcal{W}^S$ of A/B tests.

We allow for the company's prior beliefs regarding the distribution of true impacts to be misspecified via different specifications of the DGP parameters (M, s, ν). Note first that when $(M, s, \nu) = (\hat{\mu}, \hat{\sigma}, \infty)$ the company's prior beliefs are correctly specified and the DGP in equations (16) and (17) reduces to the DGP in equations (1) and (3). Relative to this baseline, we consider two forms of misspecification. First, we allow the company to misspecify the degrees of freedom of the distribution of true impacts while correctly specifying the mean and variance. In particular, we consider $(M, s, \nu) = (\hat{\mu}, \hat{\sigma} \cdot \sqrt{(\nu - 2)/\nu}, \nu)$ for different degrees of freedom $\nu > 2$. This could be the case if in practice the distribution of true impacts has fatter-than-normal tails. Second, we allow the company to misspecify the variance of the distribution of true impacts while correctly specifying the mean and degrees of freedom. In particular, we consider $(M, s, \nu) = (\hat{\mu}, \hat{\sigma} \cdot \psi, \infty)$, with a scale multiplier $\psi \in [0.33, 3.00]$. This could be the case if there is heterogeneity in the variance of the distribution of true impacts across different groups of A/B tests.

Figure 1: Simulation results

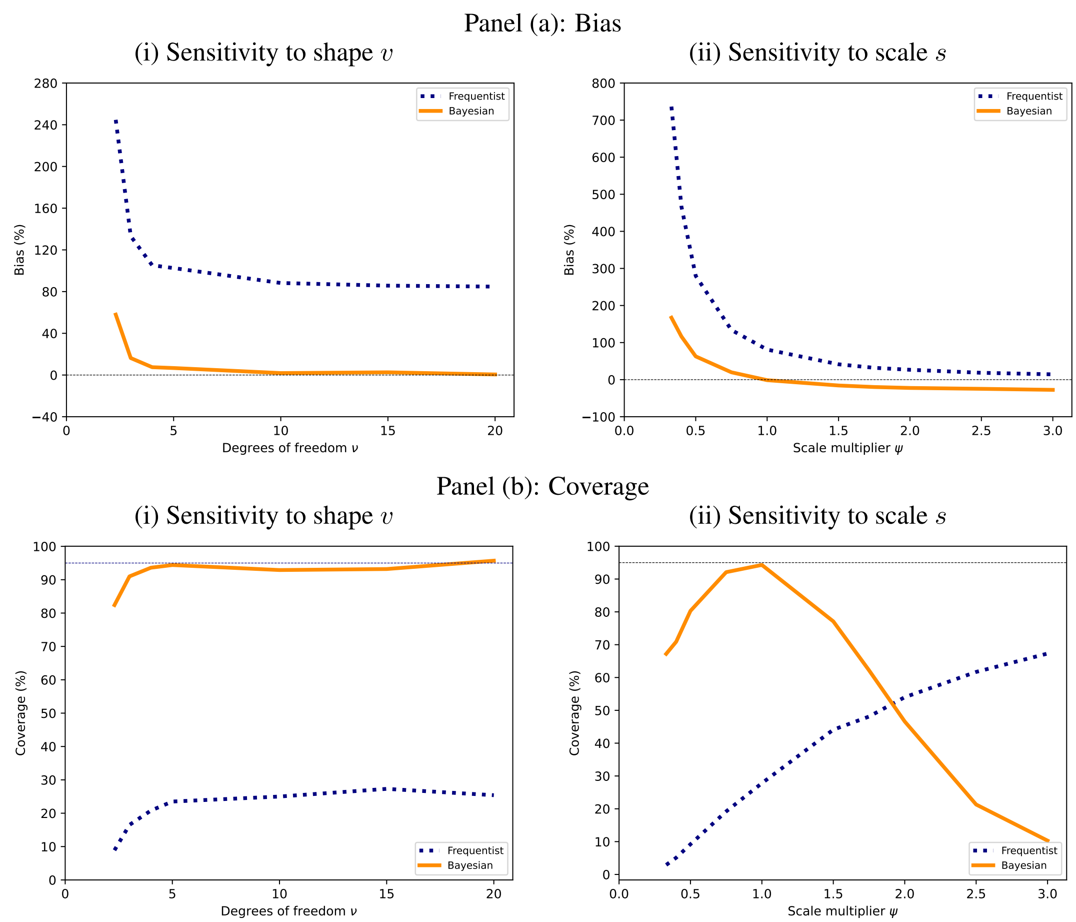

Figure 1 presents our main results. Panel (a) presents results for bias. Consistent with observation 1, figure 1 shows that across the different specifications of (M, s, ν) the frequentist estimator $\hat{\Delta}_{\mathcal{L}}$ is biased upward. The Bayesian estimator $\tilde{\mu}_{\mathcal{L}}(\hat{\mu}, \hat{\sigma}^2)$ is biased upward or downward depending on the specification of (M, s, ν). Relative to the frequentist estimator, the only specification of the DGP parameters (M, s, ν) for which the Bayesian estimator $\tilde{\mu}_{\mathcal{L}}(\hat{\mu}, \hat{\sigma}^2)$ does not reduce the magnitude of the bias is when the scale multiplier $\psi > 2$ in which case the company understates the variance of true impacts by a factor of more than 4. Consistent with observation 4, when the company's prior beliefs are correctly specified, the Bayesian estimator $\tilde{\mu}_{\mathcal{L}}(\hat{\mu}, \hat{\sigma}^2)$ is approximately unbiased. Panel (b) presents results for coverage. Consistent with observation 3, figure 1 shows that across the different specifications of (M, s, ν) the 95 percent confidence interval for the frequentist estimator $\hat{\Delta}_{\mathcal{L}}$ often has coverage significantly below 95 percent. Coverage of the 95 percent confidence interval for the Bayesian estimator $\tilde{\mu}_{\mathcal{L}}(\hat{\mu}, \hat{\sigma}^2)$ is generally higher. As was the case with bias, the only specification of the DGP parameters (M, s, ν) for which the confidence interval for Bayesian estimator $\tilde{\mu}_{\mathcal{L}}(\hat{\mu}, \hat{\sigma}^2)$ does not improve coverage relative to the confidence interval for the frequentist estimator $\hat{\Delta}_{\mathcal{L}}$ is when the scale multiplier $\psi > 2$. Consistent with observation 5, when the company's prior beliefs are correctly specified, the 95 percent confidence interval for the Bayesian estimator $\tilde{\mu}_{\mathcal{L}}(\hat{\mu}, \hat{\sigma}^2)$ has coverage of approximately 95 percent.

Application: We apply the estimators to A/B tests run at Amazon in 2023, assuming for simplicity that they all followed the same, standard launch criteria based on an (obfuscated) engagement metric. We estimate the cumulative impact of launches among the A/B tests meeting the launch criteria, both overall and separately by product group. We present results normalized by the frequentist estimate of the cumulative impact across all product groups. 95 percent confidence intervals are presented in brackets.

Table 1: Estimated cumulative impact of launched features

Product group #	(1) Frequentist $\hat{\Delta}_{\mathcal{L}}$	(2) Bayesian $\tilde{\mu}_{\mathcal{L}}(\hat{\mu}, \hat{\sigma}^2)$
1	0.011	0.004
	[0.009, 0.013]	[0.002, 0.005]
2	0.026	0.011
	[0.021, 0.030]	[0.008, 0.014]
3	0.030	0.006
	[0.026, 0.034]	[0.004, 0.008]
4	0.036	0.010
	[0.032, 0.040]	[0.008, 0.013]
5	0.036	0.019
	[0.031, 0.042]	[0.015, 0.023]
6	0.070	0.020
	[0.060, 0.080]	[0.015, 0.024]
7	0.073	0.042
	[0.066, 0.080]	[0.036, 0.048]
8	0.078	0.027
	[0.071, 0.084]	[0.022, 0.031]
9	0.078	0.036
	[0.069, 0.086]	[0.030, 0.042]
10	0.078	0.035
	[0.070, 0.087]	[0.029, 0.041]
11	0.161	0.073
	[0.146, 0.177]	[0.062, 0.084]
12	0.324	0.130
	[0.308, 0.339]	[0.119, 0.140]
Total	1.000	0.413
	[0.970, 1.030]	[0.393, 0.434]

Table 1 shows significant gaps between the frequentist estimates $\hat{\Delta}_{\mathcal{L}}$ and the Bayesian estimates $\tilde{\mu}_{\mathcal{L}}(\hat{\mu}, \hat{\sigma}^2)$, both in their overall magnitude and their distribution across product groups. Relative to the frequentist estimate, the Bayesian estimate of the cumulative impact of launched features across all product groups is nearly 60 percent smaller in magnitude. Relative to the frequentist estimates, the Bayesian estimates increase the share of the total impact accounted for by product group 7 by around 40 percent while reducing the share of the total impact accounted for by product group 3 around 50 percent.

Limitations: We conclude by emphasizing 3 important limitations of our work. First, our setup assumes the A/B tests are independent of each other in which case the cumulative impact of features that are launched is equal to the sum of the impacts of launching each feature. But this need not be the case in practice. If, for example, 2 features interact and are tested concurrently then the impact of launching them both could be different than the sum of their impacts. Second, we focus on estimators that can easily be computed given standard results from A/B tests. But these estimators may be outperformed by other, more sophisticated estimators (see, for example, [4]). Third, in assessing the relative performance of the frequentist estimator $\hat{\Delta}_{\mathcal{L}}$ and the Bayesian estimator $\tilde{\mu}_{\mathcal{L}}(\hat{\mu}, \hat{\sigma}^2)$ we allow for the company's prior beliefs to be misspecified in ways that we think are most likely in practice. But it's possible that company's prior beliefs could be misspecified in other ways and that under these different types of misspecification we would arrive at different conclusions regarding the relative performance of the estimators.

References

[1] Minyong R. Lee and Milan Shen. Winner's curse: Bias estimation for total effects of features in online controlled experiments. *KDD*, pages 491–499, 2018.

[2] Alex Deng, Yicheng Li, Jiannan Lu, and Vivek Ramamurthy. On post-selection inference in a/b testing. *KDD*, pages 1–10, 2021.

[3] Eric A. Cator and Erik W. van Zwet. The signifiance filter, the winner's curse and the need to shrink. *Statistica Neerlandica*, 75(4):437–452, 2021.

[4] Isaiah Andrews, Toru Kitagawa, and Adam McCloskey. Winner's curse: Bias estimation for total effects of features in online controlled experiments. *Quarterly Journal of Economics*, 139(1):305–358, 2023.

[5] William H. Greene. *Econometric Analysis*. Pearson Prentice Hall, 2008.

[6] George Casella and Roger L. Berger. *Statistical Inference*. Wadsworth Group, 2002.

A Appendix: Proofs

Observation 1: The frequentist estimator $\hat{\Delta}_{\mathcal{L}}$ is biased upwards. If $\hat{\Delta}_w | \Delta_w, \hat{\tau}_w^2 \overset{\text{ind}}{\sim} N(\Delta_w, \hat{\tau}_w^2)$, then:

$$E\left(\hat{\Delta}_{\mathcal{L}} - \Delta_{\mathcal{L}} \middle| \hat{\Delta}_w > k_w \text{ for all } w \in \mathcal{L}\right) = \sum_{w \in \mathcal{L}} E\left(\hat{\tau}_w \left(\frac{\phi\left(\frac{k_w - \Delta_w}{\hat{\tau}_w}\right)}{1 - \Phi\left(\frac{k_w - \Delta_w}{\hat{\tau}_w}\right)}\right) \middle| \hat{\Delta}_w > k_w\right) > 0 \tag{18}$$

Proof: We follow ideas presented in [1]. For each feature $w \in \mathcal{L}$, we have that:

$$E\left(\hat{\Delta}_w - \Delta_w \middle| \hat{\Delta}_w > k_w\right) = E\left(E\left(\hat{\Delta}_w - \Delta_w \middle| \hat{\Delta}_w > k_w, \Delta_w\right) \middle| \hat{\Delta}_w > k_w\right) \tag{19}$$

$$= E\left(E\left(\hat{\Delta}_w \middle| \hat{\Delta}_w > k_w, \Delta_w\right) - \Delta_w \middle| \Delta_w > k_w\right) \tag{20}$$

$$= E\left(\Delta_w + \hat{\tau}_w \left(\frac{\phi\left(\frac{k_w - \Delta_w}{\hat{\tau}_w}\right)}{1 - \Phi\left(\frac{k_w - \Delta_w}{\hat{\tau}_w}\right)}\right) - \Delta_w \middle| \hat{\Delta}_w > k_w\right) \tag{21}$$

$$= E\left(\hat{\tau}_w \left(\frac{\phi\left(\frac{k_w - \Delta_w}{\hat{\tau}_w}\right)}{1 - \Phi\left(\frac{k_w - \Delta_w}{\hat{\tau}_w}\right)}\right) \middle| \hat{\Delta}_w > k_w\right) \tag{22}$$

where equation (19) follows from the law of iterated expectations, equation (20) follows from the linearity of the expectation operator, and equation (21) follows from properties of the truncated normal distribution (see, for example, theorem 24.2 in [5]). Leveraging the independence across features w and the linearity of the expectation operator therefore yields:

$$E\left(\sum_{w \in \mathcal{L}} \hat{\Delta}_w - \sum_{w \in \mathcal{L}} \Delta_w \middle| \hat{\Delta}_w > k_w \text{ for all } w \in \mathcal{L}\right) = \sum_{w \in \mathcal{L}} E\left(\hat{\tau}_w \left(\frac{\phi\left(\frac{k_w - \Delta_w}{\hat{\tau}_w}\right)}{1 - \Phi\left(\frac{k_w - \Delta_w}{\hat{\tau}_w}\right)}\right) \middle| \hat{\Delta}_w > k_w\right) \tag{23}$$

That the bias is positive follows from the fact that, excluding edge cases, $\hat{\tau}_w > 0$, $\phi(\cdot) > 0$, and $\Phi(\cdot) < 1$ □

Observation 2: The bias of the frequentist estimator $\hat{\Delta}_{\mathcal{L}}$ is decreasing in statistical power Π_w, with the bias approaching 0 as power approaches $\Pi_w = 1$ for all $w \in \mathcal{L}$

Proof: For each feature $w \in \mathcal{L}$, we want to show that:

$$\frac{dE\left(\hat{\tau}_w \left(\frac{\phi\left(\frac{k_w - \Delta_w}{\hat{\tau}_w}\right)}{1 - \Phi\left(\frac{k_w - \Delta_w}{\hat{\tau}_w}\right)}\right) \middle| \hat{\Delta}_w > k_w\right)}{d\Pi_w} < 0 \tag{24}$$

To this end, we assume regularity conditions under which the derivative of the expectation is equal to the expectation of the derivative (see section 2.4 of [6]) and define:

$$B_w = \hat{\tau}_w \left(\frac{\phi\left(\frac{k_w - \Delta_w}{\hat{\tau}_w}\right)}{1 - \Phi\left(\frac{k_w - \Delta_w}{\hat{\tau}_w}\right)}\right) \tag{25}$$

By the chain rule, we have that:

$$\frac{dB_w}{d\Delta_w} = \frac{dB_w}{d\Pi_w} \frac{d\Pi_w}{d\Delta_w} \tag{26}$$

We can therefore prove the claim that $dB_w / d\Pi_w < 0$ by showing that (a) B_w is strictly decreasing in Δ_w and that (b) Π_w is strictly increasing in Δ_w.

Toward showing (a), note that:

$$\frac{dB_w}{d\Delta_w} = \frac{-\phi'\left(\frac{k_w-\Delta_w}{\hat{\tau}_w}\right)\left(1-\Phi\left(\frac{k_w-\Delta_w}{\hat{\tau}_w}\right)\right)-\phi\left(\frac{k_w-\Delta_w}{\hat{\tau}_w}\right)^2}{\left(1-\Phi\left(\frac{k_w-\Delta_w}{\hat{\tau}_w}\right)\right)^2} \tag{27}$$

$$= \frac{\left(\frac{k_w-\Delta_w}{\hat{\tau}_w}\right)\phi\left(\frac{k_w-\Delta_w}{\hat{\tau}_w}\right)\left(1-\Phi\left(\frac{k_w-\Delta_w}{\hat{\tau}_w}\right)\right)-\phi\left(\frac{k_w-\Delta_w}{\hat{\tau}_w}\right)^2}{\left(1-\Phi\left(\frac{k_w-\Delta_w}{\hat{\tau}_w}\right)\right)^2} \tag{28}$$

$$= \frac{z\phi\left(z\right)\left(1-\Phi\left(z\right)\right)-\phi\left(z\right)^2}{\left(1-\Phi\left(z\right)\right)^2} \tag{29}$$

where equation (27) follows from the quotient rule, equation (28) follows from the fact that $-\phi'(z) = z\phi(z)$ for any z, and equation (29) follows from the substitution $z = (k_w - \Delta_w)/\hat{\tau}_w$. It follows that $dB_w/d\Delta_w < 0$ if and only if $z\left(1-\Phi\left(z\right)\right) < \phi\left(z\right)$ for all z. This inequality holds trivially for all $z < 0$. For $z > 0$, note that

$$1-\Phi\left(z\right) = \int_z^{\infty}\frac{1}{\sqrt{2\pi}}\exp(-t^2/2)dt \tag{30}$$

$$< \frac{1}{z}\int_z^{\infty}\frac{1}{\sqrt{2\pi}}t\exp(-t^2/2)dt \tag{31}$$

$$= \frac{1}{z}\phi(z) \tag{32}$$

from which the inequality follows. Toward showing (b), note that:

$$\Pi_w = 1-\Phi\left(\Phi^{-1}\left(1-\frac{\alpha}{2}\right)-\frac{\Delta_w}{\hat{\tau}_w}\right)+\Phi\left(-\Phi^{-1}\left(1-\frac{\alpha}{2}\right)-\frac{\Delta_w}{\hat{\tau}_w}\right) \tag{33}$$

$$\frac{d\Pi_w}{d\Delta_w} = \frac{1}{\hat{\tau}_w}\left(\phi\left(\Phi^{-1}\left(1-\frac{\alpha}{2}\right)-\frac{\Delta_w}{\hat{\tau}_w}\right)-\phi\left(-\Phi^{-1}\left(1-\frac{\alpha}{2}\right)-\frac{\Delta_w}{\hat{\tau}_w}\right)\right) \tag{34}$$

For any x and y such that $|x| < |y|$, $\phi(x) > \phi(y)$. Because $\Phi^{-1}\left(1-\frac{\alpha}{2}\right) > 0$ and $\Delta_w > 0$:

$$\left|\Phi^{-1}\left(1-\frac{\alpha}{2}\right)-\frac{\Delta_w}{\hat{\tau}_w}\right| < \left|-\Phi^{-1}\left(1-\frac{\alpha}{2}\right)-\frac{\Delta_w}{\hat{\tau}_w}\right| \tag{35}$$

and therefore $d\Pi_w/d\Delta_w > 0$. Together, the fact that $dB_w/d\Delta_w < 0$ and $d\Pi_w/d\Delta_w > 0$ establish (via equation (26)) that $dB_w/d\Pi_w < 0$, as claimed. We conclude by noting that $\lim_{\Delta_w/\hat{\tau}_w\to\infty} B_w = 0$ □

Observation 3: When statistical power is low, the $(1-\alpha)$ confidence interval for $\hat{\Delta}_{\mathcal{L}}$ will cover $\Delta_{\mathcal{L}}$ less than $(1-\alpha)$ percent of the time. If $\hat{\Delta}_w|\Delta_w, \hat{\tau}_w^2 \overset{\text{ind}}{\sim} N(\Delta_w, \hat{\tau}_w^2)$ and, for each feature $w \in \mathcal{L}$, $\Pi_w < 0.5$ with probability 1, then:

$$\Pr\left(\left|\hat{\Delta}_{\mathcal{L}}-\Delta_{\mathcal{L}}\right| \le \sqrt{\sum_{w\in\mathcal{L}}\hat{\tau}_w^2}\cdot\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\middle|\hat{\Delta}_w > k_w \text{ for all } w\in\mathcal{L}\right) < 1-\alpha \tag{36}$$

Proof: We outline a sketch of a proof under restrictions on k_w and the assumption that the confidence interval for $\hat{\Delta}_{\mathcal{L}}$ will have lower-than-nominal coverage conditional on $\hat{\Delta}_w > k_w$ for all $w \in \mathcal{L}$ if, for each feature $w \in \mathcal{L}$, the confidence interval for $\hat{\Delta}_w$ has lower-than-nominal coverage conditional on $\hat{\Delta}_w > k_w$. We follow ideas in [3]. Toward establishing this latter fact, define $z_\alpha = \Phi^{-1}(1-\alpha/2)$ and note that, for each feature $w \in \mathcal{L}$:

$$\Pr\left(\left|\hat{\Delta}_w-\Delta_w\right| \le \hat{\tau}_w\cdot z_\alpha\middle|\hat{\Delta}_w > k_w\right) = \Pr\left(\Pr\left(\left|\hat{\Delta}_w-\Delta_w\right| \le \hat{\tau}_w\cdot z_\alpha\middle|\hat{\Delta}_w > k_w, \Delta_w\right)\middle|\hat{\Delta}_w > k_w\right) \tag{37}$$

Note also that $\Delta_w < \hat{\tau}_w\cdot z_\alpha$ whenever $\Pi_w < 0.5$. We therefore show that for each feature $w \in \mathcal{L}$:

$$\Pr\left(\left|\hat{\Delta}_w - \Delta_w\right| > \hat{\tau}_w \cdot z_\alpha \middle| \hat{\Delta}_w > k_w, \Delta_w\right) > \alpha \tag{38}$$

for any $0 < \Delta_w < k_w = \hat{\tau}_w \cdot z_\alpha$. By Bayes rule:

$$\Pr\left(\left|\hat{\Delta}_w - \Delta_w\right| > \hat{\tau}_w \cdot z_\alpha \middle| \hat{\Delta}_w > k_w, \Delta_w\right) = \frac{\Pr\left(\hat{\Delta}_w > k_w \middle| \Delta_w, \left|\hat{\Delta}_w - \Delta_w\right| > \hat{\tau}_w \cdot z_\alpha\right) \Pr\left(\left|\hat{\Delta}_w - \Delta_w\right| > \hat{\tau}_w \cdot z_\alpha \middle| \Delta_w\right)}{\Pr\left(\hat{\Delta}_w > k_w \middle| \Delta_w\right)} > \alpha \tag{39}$$

Because $\hat{\Delta}_w | \Delta_w, \hat{\tau}_w^2 \overset{\text{ind}}{\sim} N(\Delta_w, \hat{\tau}_w^2)$:

$$\Pr\left(\left|\hat{\Delta}_w - \Delta_w\right| > \hat{\tau}_w \cdot z_\alpha \middle| \Delta_w\right) = \alpha \tag{40}$$

It therefore suffices to show that:

$$\Pr\left(\hat{\Delta}_w > k_w \middle| \Delta_w, \left|\hat{\Delta}_w - \Delta_w\right| > \hat{\tau}_w \cdot z_\alpha\right) - \Pr\left(\hat{\Delta}_w > k_w \middle| \Delta_w\right) > 0 \tag{41}$$

Note that:

$$\Pr\left(\hat{\Delta}_w > k_w \middle| \Delta_w, \left|\hat{\Delta}_w - \Delta_w\right| > \hat{\tau}_w \cdot z_\alpha\right) - \Pr\left(\hat{\Delta}_w > k_w \middle| \Delta_w\right) = \frac{1}{2} - 1 + \Phi\left(\frac{k_w - \Delta_w}{\hat{\tau}_w}\right) \tag{42}$$

$$= \Phi\left(\frac{k_w - \Delta_w}{\hat{\tau}_w}\right) - \frac{1}{2} \tag{43}$$

which is positive whenever $k_w > \Delta_w$, establishing the claim □

Observation 4: When the company's prior beliefs are correctly specified, the Bayesian estimator $\tilde{\mu}_{\mathcal{L}}(\hat{\mu}, \hat{\sigma}^2)$ is unbiased. If $\hat{\Delta}_w | \Delta_w, \hat{\tau}_w^2 \overset{\text{ind}}{\sim} N(\Delta_w, \hat{\tau}_w^2)$ and $\Delta_w | \mu, \sigma^2 \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, then:

$$E\left(\tilde{\mu}_{\mathcal{L}}(\mu, \sigma^2) - \Delta_{\mathcal{L}} \middle| \hat{\Delta}_w > k_w \text{ for all } w \in \mathcal{L}\right) = 0 \tag{44}$$

Proof: For each feature $w \in \mathcal{L}$, we have that:

$$E\left(\tilde{\mu}_w(\mu, \sigma^2) - \Delta_w \middle| \hat{\Delta}_w > k_w\right) = E\left(E\left(\tilde{\mu}_w(\mu, \sigma^2) - \Delta_w \middle| \hat{\Delta}_w\right) \middle| \hat{\Delta}_w > k_w\right) \tag{45}$$

$$= E\left(\tilde{\mu}_w(\mu, \sigma^2) - E\left(\Delta_w \middle| \hat{\Delta}_w\right) \middle| \hat{\Delta}_w > k_w\right) \tag{46}$$

$$= 0 \tag{47}$$

where equation (45) follows from the law of iterated expectations, equation (46) follows from the linearity of the expectation operator, and equation (47) follows from the fact that if $\hat{\Delta}_w | \Delta_w, \hat{\tau}_w^2 \overset{\text{ind}}{\sim} N(\Delta_w, \hat{\tau}_w^2)$ and $\Delta_w | \mu, \sigma^2 \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ then $\tilde{\mu}_w(\mu, \sigma^2) = E(\Delta_w \big| \hat{\Delta}_w)$. Leveraging the independence across features w and the linearity of the expectation operator therefore yields:

$$E\left(\sum_{w \in \mathcal{L}} \tilde{\mu}_w(\mu, \sigma^2) - \sum_{w \in \mathcal{L}} \Delta_w \middle| \hat{\Delta}_w > k_w \text{ for all } w \in \mathcal{L}\right) = \sum_{w \in \mathcal{L}} E\left(\tilde{\mu}_w(\mu, \sigma^2) - \Delta_w \middle| \hat{\Delta}_w > k_w\right) \tag{48}$$

$$= 0 \tag{49}$$

which establishes the claim □

Observation 5: When the company's prior beliefs are correctly specified, the $(1-\alpha)$ confidence interval for $\tilde{\mu}_{\mathcal{L}}(\hat{\mu}, \hat{\sigma}^2)$ will cover $\Delta_{\mathcal{L}}$ $(1-\alpha)$ percent of the time. If $\hat{\Delta}_w | \Delta_w, \hat{\tau}_w^2 \overset{\text{ind}}{\sim} N(\Delta_w, \hat{\tau}_w^2)$ and $\Delta_w | \mu, \sigma^2 \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, then:

$$\Pr\left(\left|\tilde{\mu}_{\mathcal{L}}(\mu, \sigma^2) - \Delta_{\mathcal{L}}\right| \leq \sqrt{\sum_{w \in \mathcal{L}} \tilde{\sigma}_w^2(\sigma^2)} \cdot \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \middle| \hat{\Delta}_w > k_w \text{ for all } w \in \mathcal{L}\right) = 1 - \alpha \tag{50}$$

Proof: Define $z_\alpha = \Phi^{-1}(1 - \alpha/2)$. By the law of iterated expectations:

$$\Pr\left(\left|\sum_{w \in \mathcal{L}} \tilde{\mu}_w(\mu, \sigma^2) - \sum_{w \in \mathcal{L}} \Delta_w\right| \leq \sqrt{\sum_{w \in \mathcal{L}} \tilde{\sigma}_w^2(\sigma^2)} \cdot z_\alpha \middle| \hat{\Delta}_w > k_w \text{ for all } w \in \mathcal{L}\right) = \Pr\left(\Pr\left(\left|\sum_{w \in \mathcal{L}} \tilde{\mu}_w(\mu, \sigma^2) - \sum_{w \in \mathcal{L}} \Delta_w\right| \leq \sqrt{\sum_{w \in \mathcal{L}} \tilde{\sigma}_w^2(\sigma^2)} \cdot z_\alpha \middle| \{\hat{\Delta}_w\}_{w \in \mathcal{L}}, \mu, \sigma^2\right) \middle| \hat{\Delta}_w > k_w \text{ for all } w \in \mathcal{L}\right) \tag{51}$$

If $\hat{\Delta}_w | \Delta_w, \hat{\tau}_w^2 \overset{\text{ind}}{\sim} N(\Delta_w, \hat{\tau}_w^2)$ and $\Delta_w | \mu, \sigma^2 \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, then $\Delta_w | \hat{\Delta}_w, \mu, \sigma^2 \overset{\text{ind}}{\sim} N(\tilde{\mu}_w(\mu, \sigma^2), \tilde{\sigma}_w^2(\sigma^2))$ and therefore:

$$\sum_{w \in \mathcal{L}} \Delta_w \middle| \{\hat{\Delta}_w\}_{w \in \mathcal{L}}, \mu, \sigma^2 \overset{\text{ind}}{\sim} N\left(\sum_{w \in \mathcal{L}} \tilde{\mu}_w(\mu, \sigma^2), \sum_{w \in \mathcal{L}} \tilde{\sigma}_w^2(\sigma^2)\right) \tag{52}$$

from which it follows that:

$$\Pr\left(\left|\sum_{w \in \mathcal{L}} \tilde{\mu}_w(\mu, \sigma^2) - \sum_{w \in \mathcal{L}} \Delta_w\right| \leq \sqrt{\sum_{w \in \mathcal{L}} \tilde{\sigma}_w^2(\sigma^2)} \cdot z_\alpha \middle| \{\hat{\Delta}_w\}_{w \in \mathcal{L}}, \mu, \sigma^2\right) = 1 - \alpha \tag{53}$$

which establishes the claim □