informs
https://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Articles in Advance, pp. 1–12
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Contextual Areas

Technical Note—Data-Driven Profit Estimation Error in the Newsvendor Model

Andrew F. Siegel,[a] **Michael R. Wagner**[a,*]

[a] Information Systems and Operations Management, Michael G. Foster School of Business, University of Washington, Seattle, Washington 98195

*Corresponding author

Contact: asiegel@uw.edu, https://orcid.org/0000-0002-2144-2591 (AFS); mrwagner@uw.edu, https://orcid.org/0000-0003-2077-3564 (MRW)

Received: February 6, 2023
Revised: May 19, 2023;
June 14, 2023; June 20, 2023
Accepted: July 5, 2023
Published Online in Articles in Advance:
August 9, 2023

Area of Review: Operations and Supply Chains

https://doi.org/10.1287/opre.2023.0070

Abstract. In this note, we identify a statistically significant error in naively estimating the expected profit in a data-driven newsvendor model, and we show how to correct the error. In particular, we analyze a newsvendor model where the continuous demand distribution is not known, and only a sample of demand data is available. In this context, an empirical demand distribution, that is induced by the sample of data, is used in place of the (unknown) true distribution. The quantity at the critical percentile $1 - c/p$ of the empirical distribution is known as the sample average approximation order quantity, where p is the unit revenue and c the unit cost. We prove that, if the empirical distribution is used to estimate the expected profit, this estimate exhibits a positive, statistically significant bias. We derive a closed-form expression for this bias that only depends on p and c and the sample of data. The bias expression can then be used to design an adjusted expected profit estimate, which we prove is asymptotically unbiased. Numerical hypothesis testing experiments confirm that the unadjusted estimation error is statistically significant, whereas the adjusted estimation error is not significantly different from zero. The bias is not negligible in our numerical experiments: For lognormally and normally distributed demand, the unadjusted error is 2.4% and 3.0% of the true expected profit, respectively. A more detailed exploration with exact finite-sample results for exponentially distributed demand demonstrates that the estimation error percentage can be much larger.

Funding: A. F. Siegel gratefully acknowledges the support of the Grant I. Butterbaugh Professorship. M. R. Wagner gratefully acknowledges the support of a Neal and Jan Dempsey Faculty Fellowship. Both are from the Michael G. Foster School of Business, University of Washington.
Supplemental Material: The online appendix is available at https://doi.org/10.1287/opre.2023.0070.

Keywords: newsvendor • estimation error • statistics • data-driven

1. Introduction

In this note, we identify a statistically significant error in naively estimating the expected profit in a data-driven newsvendor model, and we show how to correct the error. In particular, we study a newsvendor model that applies the sample average approximation (SAA) order quantity, an approach that replaces the unknown true demand distribution with a data-driven empirical distribution in the classic newsvendor formulas. We demonstrate that estimating the maximized expected profit using the empirical distribution systematically overestimates the true expected profit, leading a decision maker to believe that the expected profit is larger than it actually is. Practically speaking, the expected profit serves as a forecast for the future realized profit, and we effectively show that this forecast is *biased*. We demonstrate how this bias can be estimated from data, which allows us to adjust the expected profit estimate and obtain an asymptotically unbiased estimate.

We rigorously prove that using the empirical distribution in place of the true distribution will lead to a positive bias in estimating expected profit, on average. We derive the asymptotic bias in closed form, which is a function of the sample size, the selling price and purchasing cost, as well as an estimate of the demand density at the (unknown but estimated) true optimal ordering quantity. Our bias formula can be used directly to adjust the empirical distribution-based SAA expected profit estimate to obtain a new estimate that is provably asymptotically unbiased. We also perform numerical experiments that show that (1) using the empirical distribution to estimate expected profit leads to an expected estimation error that is statistically significant, and (2) our adjusted expected profit estimate has no statistically significant expected estimation error. A more detailed numerical exploration for exponentially distributed demand and exact finite-sample results demonstrates that the bias can be a large percentage of the true expected profit.

1.1. Contributions

The primary contributions of our note are as follows:

- For a generic smooth demand distribution, from which we have only a sample of data, we derive, using basic probability theory along with conditional distribution theory, the asymptotic properties of order statistics, and Taylor series expansions, a closed-form expression for the (positive) expected bias that results from estimating the expected profit using the empirical distribution in place of the true (unknown) distribution. This bias expression can be estimated in a way that only depends on the selling price, the purchase cost, and the sample data. Using the estimated bias formula, we adjust the empirical distribution expected profit estimate, and we prove that the resulting adjusted estimate is asymptotically unbiased.
- Hypothesis testing Monte Carlo simulation experiments confirm our theoretical results. We demonstrate that the unadjusted estimation error is significantly positive, whereas the adjusted estimation error is not significantly different from zero. Notably, the unadjusted estimation errors are 2.4%–3.0% of the true expected profit values in these experiments, although we also demonstrate that the percentage error can be much higher, depending on the economic parameters, sample size, and underlying true demand distribution. We also, for normal and lognormal distributions of demand, find that our analytical adjustments eliminate bias at least as well as a cross-validated estimation of expected profit. In other words, if cross-validation is applied naively, our analytical approach eliminates bias better. In contrast, if cross-validation is properly tuned (e.g., deciding the number of folds), it can match our ability to eliminate bias. However, we argue that our analytical adjustment is preferred to cross-validation because it (1) provides intuition about why there is a bias, (2) is faster than cross-validation, and (3) does not require proper tuning as in the cross-validation case.

1.2. Literature Review

Although there is a vast literature related to the newsvendor model, we focus on the most relevant data-driven newsvendor papers. Kleywegt et al. (2002) analyzed the SAA method that uses a data-driven empirical distribution in place of the unknown true distribution for general stochastic discrete optimization problems and identified a bias in estimating the optimal objective value but did not provide an adjustment term to correct the bias. Levi et al. (2007) showed that, in the newsvendor model, the SAA method provides a solution that is provably near optimal, with high probability, a result that was improved by Levi et al. (2015) by introducing an additive bias into the order quantity. Ban and Rudin (2019) further extended the approach of Levi et al. (2015, 2007) to include explanatory variables that influence the demand distribution using machine learning algorithms. He et al. (2012) similarly studied a features-based newsvendor model for staffing hospital operating rooms, and Ban et al. (2019) studied features of demand in a multiperiod inventory management setting. In these references, the optimal expected profit depends on the true distribution, which is unknown, which implies that the profit is not calculable. Of course, the empirical distribution could be used to estimate the profit in these references; we show that this estimated profit is biased in the newsvendor model, and we also demonstrate how to correct for the bias to obtain an asymptotically unbiased profit estimate.

Other papers have implicitly or explicitly considered bias in a newsvendor context. Liyanage and Shanthikumar (2005) considered the newsvendor model under an exponential demand distribution, where estimation (of the exponential distribution's mean) and optimization (finding the optimal order quantity) are performed simultaneously, which results in the order quantity being intentionally biased to obtain higher expected profit in a data-driven context. Chu et al. (2008) consider parametric demand distributions characterized by location and scale parameters and extend the results in Liyanage and Shanthikumar (2005) using Bayesian analysis. Siegel and Wagner (2021) is the paper most related to ours, in that a bias in estimating profit in the data-driven newsvendor model is identified and corrected. However, the methods and results presented here are considerably different from those of Siegel and Wagner (2021) for several reasons. First, Siegel and Wagner (2021) worked within a finite-dimensional parametric family of distributions, whereas our work here is *nonparametric* in the sense that we are working with only a single unknown smooth distribution with no known parametric form. Therefore, our work here is much more general. Next, Siegel and Wagner (2021) made extensive use of the theory of maximum likelihood estimation to derive their results (along with the many required assumptions), whereas there is no corresponding theory available in the present nonparametric context; instead, we are able to (surprisingly, perhaps) derive our results for a single unknown distribution using primarily smoothness of the unknown density function together with conditional distribution theory and the asymptotic properties of order statistics.

Finally, Ito et al. (2018) study a similar problem of providing unbiased estimates of optimized objective function values, for more general optimization problems, except under stronger parametric and linearity assumptions that we do not need in our note. Furthermore, their results require the solutions of multiple optimization problems and are not closed form, which limits interpretability; in contrast, our results are for an unknown smooth demand distribution (not restricted to any particular parametric family of distributions) in the newsvendor model, data driven and closed form, which facilitates interpretability. Gupta and Rusmevichientong (2021) and Gupta et al. (2023) study similar parameterized linear objective optimization models, focusing on small-data large-scale optimization and out-of-sample estimation, respectively. Our results are again distinct because we do not require a parametric family of distributions or linear assumptions. Indeed, it is unclear how these models would apply to the nonlinear newsvendor problem under an arbitrary continuous demand distribution, the focus of our note.

2. Preliminaries

The newsvendor model determines an order quantity y that maximizes expected profit in the face of random demand X, where the unit sales price is p and the unit procurement cost is c, with $p > c > 0$. We assume that X is a continuous random variable, with continuous density f and cumulative distribution function F, with support on a nondegenerate interval within (or equal to) the nonnegative real numbers. We assume that all relevant moments of X exist and are finite. We also assume that the derivative f' of the density is continuous and bounded. We make no other assumption about the distribution. In particular, we do not assume that we know the value of any summary statistics (mean, variance, etc.), and we do not assume any parametric form for the distribution. The only information about F will be via a data sample from the distribution. A note on notation: We denote random variables as uppercase letters and their realizations (or nonrandom quantities) in lowercase letters.

If the distribution F were known, the classic newsvendor model is $\max_{y\geq 0} E_F(p\min\{X, y\} - cy)$, which has the well-known solution $y^* = F^{-1}(1 - c/p)$, where $1 - c/p$ is known as the *critical ratio*. Unfortunately, in many situations the demand distribution F is not known. Instead, data are typically available. In the next section, we describe how a data-driven newsvendor model can be formulated.

2.1. Data Perspective

In practice, one typically has access to data, which, in the newsvendor context, is demand data $x_1, \ldots, x_n$, assumed here to be obtained from a random sample from F. Because the distribution F of demand X is not known, one may instead approximate F using the empirical distribution $\widehat{F}$, which is a discrete distribution with each of the observations x_i occurring with probability $1/n$. In other words, the newsvendor model is approximated by

$$\max_{y\geq 0} \frac{1}{n}\sum_{i=1}^{n}(p\min\{x_i, y\} - cy). \tag{1}$$

This approach is known as the SAA. Following Levi et al. (2007, p. 825), we define the optimal SAA order quantity y_{saa} to be

$$y_{saa} \triangleq \min_{j=1,\ldots,n}\left\{x_j : \frac{1}{n}\sum_{i=1}^{n} I_{x_i\leq x_j} \geq 1 - c/p\right\}, \tag{2}$$

where I_A is the indicator function, which equals one if the event A is true and zero otherwise. In particular, the expression $\frac{1}{n}\sum_{i=1}^{n} I_{x_i\leq y_{saa}}$ equals $\widehat{F}(y_{saa})$ and approximates $F(y_{saa})$; recall that, in the newsvendor model, $F(y^*) = 1 - c/p$.

Lemma 1. *Definition (2) of the optimal SAA order quantity is equivalent to $y_{saa} = x_{(k)}$ with $k = \lceil n(1 - c/p)\rceil$, where $x_{(k)}$ is the kth order statistic and $\lceil\cdot\rceil$ denotes the ceiling function.*

Strictly speaking, k is a function of n; however, we will use the notation k instead of $k(n)$ for simplicity.

Our note is primarily concerned with estimating expected profit for the y_{saa} order quantity. A natural data-driven estimate for the expected profit is to simply plug this order quantity into the objective of Formulation (1):

$$\frac{1}{n}\sum_{i=1}^{n}(p\min\{x_i, y_{saa}\} - cy_{saa}). \tag{3}$$

Our first main result is to show that this approach results in a biased estimate of the true expected profit; to demonstrate this, we first view the data-driven problem from the perspective of random variables in the next section. Our second main result is to show how to correct this bias to provide an asymptotically unbiased data-driven estimate of expected profit.

2.2. Random Variable Perspective

The SAA order quantity y_{saa} depends on the sample of data that was used to calculate it in Equation (2). We therefore wish to study its dependence on the sampling distribution. In particular, we let $X_1, \ldots, X_n$ denote an independent identically distributed (i.i.d.) sample of size n from the (unknown) demand distribution F, representing a sample of past demand data. We next introduce a probabilistic analogue of Equation (2):

$$Y_{saa} \triangleq \min_{j=1,\ldots,n}\left\{X_j : \frac{1}{n}\sum_{i=1}^{n} I_{X_i\leq X_j} \geq 1 - c/p\right\}. \tag{4}$$

We let g and G denote the density and cumulative distribution function of Y_{saa}, respectively (where the functional forms of g and G will be given in Lemma 7). In Definition (4), Y_{saa} is a random variable that explicitly depends on the random sample $X_1, \ldots, X_n$. Letting $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ denote the order statistics of the sample, we also provide a probabilistic analogue to the alternative characterization of y_{saa} in Lemma 1: $Y_{saa} = X_{(k)}$, where $k = \lceil n(1 - c/p)\rceil$. Let X denote future demand, obtained by drawing a new observation from F, independent of the past demand data $(X_1, \ldots, X_n)$. The true expected profit of the SAA order quantity can be written as

$$\pi_{true} \triangleq E_F(p\min\{X, Y_{saa}\} - cY_{saa}), \tag{5}$$

which is the quantity we are focused on estimating unbiasedly. Examining Equation (3), it turns out that the natural estimate of expected profit associated with the SAA order quantity is a realization of the random variable $\frac{1}{n}\sum_{i=1}^{n}(p\min\{X_i, Y_{saa}\} - cY_{saa})$, whose expectation is

$$\pi_{naive} \triangleq E_F\left(\frac{1}{n}\sum_{i=1}^{n}(p\min\{X_i, Y_{saa}\} - cY_{saa})\right). \tag{6}$$

We argue that this approach leads to erroneous, overly optimistic estimates of the true expected profit. Indeed, Theorem 1 implies that $\pi_{naive} = Bias_{profit} + \pi_{true} + o(1/n)$, where $Bias_{profit} > 0$, which demonstrates that the naive estimate in Equation (3) is biased upward. The expansion of this equation,

$$\begin{aligned} &E_F\left(\frac{1}{n}\sum_{i=1}^{n}(p\min\{X_i, Y_{saa}\} - cY_{saa})\right)\\ &\quad = Bias_{profit} + E_F(p\min\{X, Y_{saa}\} - cY_{saa}) + o\left(\frac{1}{n}\right), \end{aligned} \tag{7}$$

lends to a machine learning interpretation, in terms of in-sample versus out-of-sample performance. In the left expression of Equation (7), there are n i.i.d. realizations of demand $X_1, \ldots, X_n$ and Y_{saa}, which is a function of these n observations, and the entire expression represents the expected *in-sample* maximized profit objective. In contrast, in the right expression of Equation (7), there are $n + 1$ i.i.d. samples, namely $X_1, \ldots, X_n$ (folded into the definition of Y_{saa}), plus X, which represents future demand, and the term $E_F(p\min\{X, Y_{saa}\} - cY_{saa})$ represents the expected *out-of-sample* expected profit exactly as though it had been computed from an infinite virtual holdout sample (of both X_i and X); given the special structure of the newsvendor model, we can derive this result without needing an actual holdout sample.

In addition to identifying an estimation bias, we also show how to correct for it. In particular, in Theorem 2, we show that a data-driven *adjustment* results in $\frac{1}{n}\sum_{i=1}^{n}(p\min\{x_i, y_{saa}\} - cy_{saa}) - adjustment$ being an asymptotically unbiased estimator; in other words, we prove that $\pi_{naive} - E_F(adjustment) = \pi_{true} + o(1/n)$.

Finally, we emphasize that the true expected profit $\pi_{true} = E_F(p\min\{X, Y_{saa}\} - cY_{saa})$ is not computable because F is not known. However, the adjusted estimate $\frac{1}{n}\sum_{i=1}^{n}(p\min\{x_i, y_{saa}\} - cy_{saa}) - adjustment$ is computable, as it is based entirely on the n data observations $x_1, \ldots, x_n$.

3. Naive Estimation of Expected Profit Exhibits Statistical Bias

In this section, we first describe the form of the bias in Section 3.1. We then derive the bias in Section 3.2. In Section 3.3, we use the bias expression to *adjust* the naive expected profit formula, so that it is asymptotically unbiased. Finally, in Section 3.4, we provide an intuitive interpretation of the bias.

3.1. Form of the Bias

The profit bias can be written as

$$\begin{aligned} Bias_{profit} \triangleq \pi_{naive} - \pi_{true} = E_F\left(\frac{1}{n}\sum_{i=1}^{n}(p\min\{X_i, Y_{saa}\} - cY_{saa})\right)\\ - E_F(p\min\{X, Y_{saa}\} - cY_{saa}). \end{aligned}$$

For simplicity, we cancel the $E_F(cY_{saa})$ term and p multiplier, and we instead focus on the sales bias $Bias_{sales} \triangleq E_F(\frac{1}{n}\sum_{i=1}^{n}\min\{X_i, Y_{saa}\}) - E_F(\min\{X, Y_{saa}\})$, where $Bias_{profit} = pBias_{sales}$. We first define the sales from the naive profit expression in Equation (6) as $\tilde{S} \triangleq \frac{1}{n}\sum_{i=1}^{n}\min\{X_i, Y_{saa}\}$, whose expectation we decompose as follows:

$$\begin{aligned} E_F(\tilde{S}) &= E_F\left(\frac{1}{n}\sum_{i=1}^{n}\min\{X_i, Y_{saa}\}\right)\\ &= E_F\left(\frac{1}{n}\sum_{i=1}^{k-1}X_{(i)} + \left(\frac{n+1-k}{n}\right)Y_{saa}\right) \qquad (8)\\ &= \left(\frac{k-1}{n}\right)E_F\left(\frac{1}{k-1}\sum_{i=1}^{k-1}X_{(i)}\right)\\ &\quad + \left(\frac{n+1-k}{n}\right)E_F(Y_{saa}), \qquad (9) \end{aligned}$$

where the second equality is due to $Y_{saa} = X_{(k)}$, and the final expression follows directly; the final expression is convenient for comparing with the sales expression from the exact expected profit. We next define the sales from the exact profit expression in Equation (5) as $S \triangleq \min\{X, Y_{saa}\}$, whose expectation we decompose as follows:

$$\begin{aligned} E_F(S) &= E_F(\min\{X, Y_{saa}\}) = P(X < Y_{saa})E_F(X | X < Y_{saa})\\ &\quad + P(X > Y_{saa})E_F(Y_{saa} | X > Y_{saa})\\ &= \left(\frac{k}{n+1}\right)E_F(X | X < Y_{saa}) + \left(\frac{n+1-k}{n+1}\right)\\ &\quad E_F(Y_{saa} | X > Y_{saa}), \end{aligned} \tag{10}$$

where the second equality is due to the law of total expectation and the final equality is due to properties of order statistics, as shown in Lemma 2 (presented later in Section 3.2.1).

Examining the expressions in Equations (9) and (10), we observe three discrepancies, which may be considered the three potential sources of bias. In particular,

1. The probabilistic multipliers in Equation (9) are $\frac{k-1}{n}$ and $\frac{n+1-k}{n}$, whereas in Equation (10), they are $\frac{k}{n+1}$ and $\frac{n+1-k}{n+1}$, respectively.
2. The first expectation in Equation (9) is the unconditional $E_F\left(\frac{1}{k-1}\sum_{i=1}^{k-1}X_{(i)}\right)$, whereas in Equation (10), it is the conditional $E(X | X < Y_{saa})$.

3. The second expectation in Equation (9) is the unconditional $E_F(Y_{saa})$, whereas in Equation (10), it is the conditional $E(Y_{saa} | X > Y_{saa})$.

The cumulative effect of these discrepancies can be seen by examining the sales bias, which can be written as $Bias_{sales} = E_F(\tilde{S} - S)$. Our first result characterizes this bias asymptotically, with an $o(1/n)$ error term, and we see that it is a positive bias; that is, the naive expected profit estimate is larger than and overestimates the true expected profit, on average. The following theorem corrects for the previous discrepancies and follows immediately from Propositions 1 and 2 in the next section.

Theorem 1. *The sales bias can be written as* $Bias_{sales} = \frac{1}{nf(y^*)}\left(\frac{c}{p}\right)\left(1 - \frac{c}{p}\right) + o\left(\frac{1}{n}\right)$, *where* $y^* = F^{-1}(1 - c/p)$, *and the profit bias can be written as* $Bias_{profit} = \frac{c}{nf(y^*)}\left(1 - \frac{c}{p}\right) + o\left(\frac{1}{n}\right)$.

3.2. Derivation of the Bias

In the following results, and proofs thereof (in the online appendix), we suppress the *saa* subscript for expository clarity, so that $Y = Y_{saa}$. We next decompose the true sales S into the sales that occur under the events of overage and underage, respectively: $S \triangleq \min\{X, Y\} = XI_{X<Y} + YI_{X>Y} = S_{over} + S_{under}$, where $S_{over} \triangleq XI_{X<Y}$ and $S_{under} \triangleq YI_{X>Y}$. In particular, S_{over} is the quantity X of items sold if overage occurs and is zero otherwise, and S_{under} is the quantity Y of items sold if underage occurs and is zero otherwise. Similarly, we may decompose $\tilde{S} \triangleq \frac{1}{n}\sum_{i=1}^{n} \min\{X_i, Y\}$ into $\tilde{S} = \tilde{S}_{over} + \tilde{S}_{under}$, where $\tilde{S}_{over} \triangleq \frac{1}{n}\sum_{i=1}^{k-1} X_{(i)}$ and $\tilde{S}_{under} \triangleq \left(\frac{n+1-k}{n}\right)Y$; see Equation (8). We also introduce the function $h(y) \triangleq \int_0^y xf(x)dx = E(XI_{X<y})$, which features prominently in our analyses. In particular, $E(h(Y)) = E(S_{over})$ will be proven in Lemma 3.

Using these decompositions, we divide the evaluation of the sales bias into two subproblems:

$$\begin{aligned} Bias_{sales} &= E_F(\tilde{S} - S) \\ &= E_F(\tilde{S}_{over} - S_{over}) + E_F(\tilde{S}_{under} - S_{under}), \qquad (11) \end{aligned}$$

where $E_F(\tilde{S}_{over} - S_{over})$ is the sales bias due to overage and $E_F(\tilde{S}_{under} - S_{under})$ is the sales bias due to underage.

3.2.1. Auxiliary Results to Prove Theorem 1.

In this section, we provide a series of auxiliary results that are needed to prove our main propositions, which in turn are used to prove Theorem 1. We make use of the representations $X_i = F^{-1}(U_i)$, $X = F^{-1}(U)$, $X_{(i)} = F^{-1}(U_{(i)})$, $Y = X_{(k)} = F^{-1}(U_{(k)})$, and $F(Y) = U_{(k)}$ where F^{-1} denotes the inverse function of F, $U_1, \ldots, U_n$ is an i.i.d. sample from the uniform distribution on the interval $(0, 1)$, and U is independent of $U_1, \ldots, U_n$ from the same uniform distribution.

Our proof techniques use Taylor series expansions extensively, both for random variables as well as deterministic functions (e.g., expectations). When analyzing deterministic functions, we characterize the finite expansions' errors precisely using asymptotic notation (e.g., $O(1/n)$ or $o(1/n)$). When expanding random variables, we use $\approx$ to represent a Taylor Series expansion up to a specified degree; however, once we take expectations of these random variable expansions, we again characterize their errors analytically.

The diagram in Figure 1 presents a graph theoretic perspective of the dependencies between the various lemmas (labeled L) and propositions (labeled P), which ultimately lead to proving Theorem 1, one of the main contributions of our note. In particular, a directed arrow from, say, node L2 to L4 signifies that Lemma 2 is used to prove Lemma 4. Propositions 1 and 2 characterize the sales bias under the cases of overage and underage, respectively. These two propositions then lead naturally to Theorem 1, which characterizes the overall profit bias.

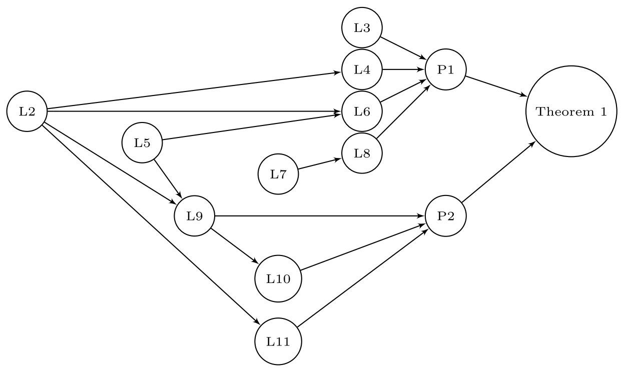
Lemma 2 applies concepts from order statistics to characterize many useful quantities about the SAA order quantity $Y = X_{(k)}$ and, as is evident from Figure 1, influences many of the subsequent lemmas that are needed to prove Theorem 1.

Lemma 2. *Here are some basic facts about the order quantity* $Y = X_{(k)}$ *and the independent future demand observation* X.

- *The probability of underage* $P(X > Y) = (n + 1 - k)/(n + 1)$.
- *The probability of overage* $P(X < Y) = k/(n + 1)$.
- *The expectation of the CDF evaluated at the order quantity is* $E[F(Y)] = k/(n + 1)$.
- *The variance of the CDF evaluated at the order quantity is* $Var[F(Y)] = Var(U_{(k)}) = k(n + 1 - k)/[(n + 1)^2(n + 2)] = O\left(\frac{1}{n}\right)$.
- *The expectation of the order quantity is* $E(Y) = F^{-1}\left(\frac{k}{n+1}\right) - \frac{f'\{F^{-1}[k/(n+1)]\}}{2f^3\{F^{-1}[k/(n+1)]\}}\frac{k(n+1-k)}{(n+1)^2(n+2)} + o\left(\frac{1}{n}\right)$.
- *The variance of the order quantity is* $Var(Y) = \frac{k(n+1-k)}{\{f[E(Y)]\}^2(n+1)^2(n+2)} + o\left(\frac{1}{n}\right) = O\left(\frac{1}{n}\right)$.

Lemma 3 characterizes some expansions of $h(Y) = \int_0^Y xf(x)dx = E(XI_{X<Y})$, where $E(h(Y)) = E(S_{over})$. The first

Figure 1. Dependencies Among Lemmas, Propositions, and Theorem 1



expansion is of a random variable, and we do not characterize the random error term, but the second two expressions (which use the first) are expectations, and we characterize the error terms precisely.

Lemma 3. *Here are some basic facts about* $h(y) \triangleq E(XI_{X<y}) = \int_0^y xf(x)dx$.

- *To first order,* $h(Y) \approx h[E(Y)] + E(Y)\{F(Y) - F[E(Y)]\}$.
- *The covariance of h with F, as functions of the order quantity, is* $Cov[h(Y), F(Y)] = E(Y)Var[F(Y)] + o\left(\frac{1}{n}\right)$.
- *The expected ratio of h to F is* $E\left(\frac{h(Y)}{F(Y)}\right) = \left(1 + \frac{Var[F(Y)]}{\{E[F(Y)]\}^2}\right)\frac{E[h(Y)]}{E[F(Y)]} - \frac{Var[F(Y)]}{\{E[F(Y)]\}^2}E(Y) + o\left(\frac{1}{n}\right)$.
- *The expected value of* $h(Y)$ *is* $E[h(Y)] = E(S_{over})$.

The next three lemmas use basic algebraic manipulations to simplify some expressions that are functions of $k = \lceil n(1 - c/p) \rceil$.

Lemma 4. *This expression will assist in establishing negligibility of some terms in the proof of Proposition 1.* $1 + \frac{Var[F(Y)]}{\{E[F(Y)]\}^2} - \frac{nE[F(Y)]}{k-1} = O\left(\frac{1}{n^2}\right)$.

Lemma 5. *These ratios involving k may be expressed in terms of the critical ratio as follows:* $\frac{k}{n} = 1 - \frac{c}{p} + O\left(\frac{1}{n}\right)$, $\frac{k}{n+1} = 1 - \frac{c}{p} + O\left(\frac{1}{n}\right)$, *and* $\frac{n+1-k}{n} \quad - \frac{n+1-k}{n+1} = \frac{n+1-k}{n(n+1)} = \frac{1}{n}\left(\frac{c}{p}\right) + O\left(\frac{1}{n^2}\right)$.

Lemma 6. *The following is instrumental in establishing Proposition 1:* $\frac{k-1}{n}\frac{Var[F(Y)]}{\{E[F(Y)]\}^2} = \frac{1}{n}\left(\frac{c}{p}\right) + O\left(\frac{1}{n^2}\right)$

Lemma 7 derives the conditional densities of the SAA order quantity Y, given the events of overage and underage, respectively, along with the unconditional density and CDF of Y. These results are used to characterize the conditional expected sales when overage occurs, which is used in Lemma 8, which, in turn, is used in the proof of Proposition 1. These conditional densities of the SAA order quantity may also be of independent interest; hence, we also include the underage case.

Lemma 7. *The conditional densities of Y given overage and given underage are as follows:*

- *The conditional density of Y given overage* $X < Y$ *is* $g_{X<Y}(y) = \frac{F(y)g(y)}{P(X<Y)}$,
- *The conditional density of Y given underage* $X > Y$ *is* $g_{X>Y}(y) = \frac{[1-F(y)]g(y)}{P(X>Y)}$,
- *The unconditional density g of Y is* $g(y) = \frac{n!}{(k-1)!(n-k)!}[F(y)]^{k-1}[1 - F(y)]^{n-k}f(y)$,
- *The unconditional CDF G(y) of Y is given by the incomplete beta function with parameters k and* $n + 1 - k$ *evaluated at F(y), where g denotes the unconditional density of Y.*

Lemma 8 analyzes expressions that are related to the differences between Equations (9) and (10), the naive and true profit estimations discussed in Section 3.1. Per Figure 1, Lemma 8 is the final auxiliary result needed to prove the overage sales bias result in Proposition 1 (presented in the next section).

Lemma 8. *The following are relevant to understanding the case of overage:*

- *This is the conditional expectation of demand, given overage with respect to a constant order quantity.* $E(\frac{1}{k-1}\sum_{i=1}^{k-1} X_{(i)} | Y = y) = E(X|X < y) = \frac{h(y)}{F(y)}$ *is an increasing function of y.*
- *This shows that the expected conditional expectation of demand in overage (given the order quantity Y) is smaller than the conditional expectation of demand given overage.* $E\left(\frac{1}{k-1}\sum_{i=1}^{k-1} X_{(i)}\right) = E\left(\frac{h(Y)}{F(Y)}\right) < E(X|X < Y) = \frac{E[h(Y)]}{E[F(Y)]}$.
- *This same expectation is related to the expected sales in overage as follows:* $\frac{k}{n+1}E\left(\frac{1}{k-1}\sum_{i=1}^{k-1} X_{(i)}\right) < E[XI_{X<Y}] = E[h(Y)] = E(S_{over})$.

Lemma 9 builds on Lemma 2 and provides alternate expressions for certain functions of Y needed for deriving the underage sales bias result in Proposition 2 (presented in the next section).

Lemma 9. *The following are additional facts about F(Y):* $E[F(Y)] - \frac{k-1}{n} = \frac{1}{n}\left(\frac{c}{p}\right) + O\left(\frac{1}{n^2}\right)$, $Var[F(Y)] = \frac{1}{n}\left(\frac{c}{p}\right)\left(1 - \frac{c}{p}\right) + O\left(\frac{1}{n^2}\right)$, *and* $E(\{F(Y) - F[E(Y)]\}^2) = Var[F(Y)] + O\left(\frac{1}{n^2}\right)$.

Lemma 10 is the auxiliary result that identifies the form of the sales bias, which interestingly equals (asymptotically) the covariance between the SAA order quantity Y and its percentile $F(Y)$. Per Figure 1, Lemma 10 is only used to prove the underage sales bias result in Proposition 2, and we learn that the overall profit bias is driven primarily by the underage sales bias; more discussion on this observation is provided in the next section.

Lemma 10. *The covariance of the order quantity with its CDF is* $Cov[Y, F(Y)] = \frac{1}{nf[E(Y)]}\left(\frac{c}{p}\right)\left(1 - \frac{c}{p}\right) + o\left(\frac{1}{n}\right)$.

Lemma 11 is the final auxiliary result needed to prove Proposition 2 and demonstrates that the SAA order quantity Y is an (asymptotically) unbiased and consistent estimator of the true optimal order quantity $y^* = F^{-1}(1 - c/p)$.

Lemma 11. *Considering the true (unknown) optimal order quantity* $y^* = F^{-1}(1 - c/p)$, *we have* $E(Y) \to y^*$, *and that Y is a consistent estimator of* y^*.

3.2.2. Main Propositions.

The following two propositions provide the formulas for the asymptotic biases of the two components of the sales bias.

Proposition 1. *The sales bias in overage is* $E(\tilde{S}_{over}) - E(S_{over}) = -\frac{1}{n}\left(\frac{c}{p}\right)E(Y) + o\left(\frac{1}{n}\right)$.

Proposition 2. *The sales bias in underage is* $E(\tilde{S}_{under}) - E(S_{under}) = \frac{1}{n}\left(\frac{c}{p}\right)E(Y) + \frac{1}{nf(y^*)}\left(\frac{c}{p}\right)\left(1 - \frac{c}{p}\right) + o\left(\frac{1}{n}\right)$.

Together, Propositions 1 and 2 imply Theorem 1, due to Equation (11). Interestingly, the $\frac{1}{n}\left(\frac{c}{p}\right)E(Y)$ term is common to the bias expressions in both the overage and underage cases, which cancel, and the overall sales bias is driven solely by the additional term in the underage sales bias. Fortunately, after this cancellation, all terms in the bias expression are known, except $f(y^*)$, which we address in the next section.

3.3. Correcting for the Bias

The sales bias $Bias_{sales} = \frac{1}{nf(y^*)}\left(\frac{c}{p}\right)\left(1-\frac{c}{p}\right) + o\left(\frac{1}{n}\right)$ from Theorem 1 may be asymptotically unbiasedly estimated using $\widehat{Bias}_{sales} = \frac{1}{n\widehat{f}}\left(\frac{c}{p}\right)\left(1-\frac{c}{p}\right)$, where $\widehat{f}$ is a consistent estimate of the density f at $E(Y)$. To estimate the density term $f(y^*)$, we use the fact that $f = F'$ to allow us to take a numerical derivative of the empirical CDF using two-order statistics with one on each side of $Y = X_{(k)}$. Choosing order statistics symmetrically placed at ranks $k - m$ and $k + m$ with integer $m > 0$, we note that the empirical CDF increases by $2m/n$ as the density's argument increases from $x_{(k-m)}$ to $x_{(k+m)}$. We therefore define our estimate as

$$\widehat{f} \triangleq \frac{2m}{n(x_{(k+m)} - x_{(k-m)})}. \tag{12}$$

To see that $1/\widehat{f}$ is a consistent estimate of $1/f(y^*)$ in the bias equation, we rely on equations (8) and (9) and theorem 1 of Tusnády (1974), where Tusnády's k corresponds to our $2m$. Thus, $\widehat{f}$ is a consistent estimator of $f(y^*)$ because the density is being estimated at $Y = X_{(k)}$ with (from Lemma 11) $E(Y) \to y^*$ and (from Lemma 2) $Var(Y) = O(1/n)$. That a continuous function (in this case, the reciprocal) of a consistent estimator is itself consistent follows, for example, from the theorem on p. 24 of Serfling (1980). Thus, we also have that $1/\widehat{f}$ is a consistent estimator of $1/f(y^*)$ as required. Tusnády's work was extended by Barabás (1987), who shows on p. 122 that, asymptotically, the best choice for m is $m \sim Cn^{2/3}$, which represents a compromise between a larger m with less density estimation variability and a smaller m with less bias; in particular, with this choice for m we have both $m \to \infty$ (so that many data points fall within the interval from $X_{(k-m)}$ to $X_{(k+m)}$) and $m/n \to 0$ (so that the interval tends to a single point). These results show that the bias estimate is asymptotically unbiased, as stated in the following proposition.

Proposition 3. *The estimated sales bias is an asymptotically unbiased estimator of the true sales bias, in the sense that* $E[\widehat{Bias}_{sales}] = Bias_{sales} + o\left(\frac{1}{n}\right)$.

Using these results, we can create an adjusted sales estimate that is asymptotically unbiased. In particular, referring to Equation (3), we define $\widehat{S}_{adjusted} \triangleq \frac{1}{n}\sum_{i=1}^{n}\min\{x_i, y_{saa}\} - \widehat{Bias}_{sales}$, where x_i, $i = 1, \ldots, n$, is the realized sample of demand and y_{saa} is the realized SAA order quantity. Similarly, an adjusted profit estimate is defined as

$$\widehat{\pi}_{adjusted} \triangleq \frac{1}{n}\sum_{i=1}^{n}(p\min\{x_i, y_{saa}\} - cy_{saa}) - p\widehat{Bias}_{sales}. \tag{13}$$

Because we demonstrated in Section 3.1 that the profit bias is driven solely by the sales bias, Proposition 3 implies that the adjusted profit estimate in Equation (13) is also asymptotically unbiased, which we present in the following theorem.

Theorem 2. *The bias-adjusted estimated profit is an asymptotically unbiased estimate of the true profit, in the sense that* $E(\widehat{\pi}_{adjusted}) = \pi_{true} + o(1/n)$.

3.4. Interpretation of the Bias

In this section, we interpret the bias and show that its dominant source is underage. We begin by showing that the canceling terms $\pm\frac{1}{n}\left(\frac{c}{p}\right)E(Y)$ (negative for the bias in overage from Proposition 1, positive in underage from Proposition 2) are artifacts due to an error of estimation of the probabilities of overage and underage as computed within-sample by the naive newsvendor. After correcting for this estimation error, we show that the asymptotic bias from Theorem 1 is due entirely to *underage*. We then understand the bias as a mathematical consequence of the sign of the covariance of the SAA order quantity Y with a monotonic function of itself $F(Y)$.

3.4.1. Bias Is due to Underage Only.

The naive sales estimate in underage is $\tilde{S}_{under} = \left(\frac{n+1-k}{n}\right)Y$ (c.f., the beginning of Section 3.2), which may be interpreted as the within-sample estimated probability of underage, $\frac{n+1-k}{n}$, times the sales Y that would occur with this order quantity when underage occurs. However, the true probability of underage is actually the slightly smaller $\frac{n+1-k}{n+1}$ from Lemma 2. We correct this estimation error by defining the *probability-adjusted naive sales estimate in underage* given Y to be an $O(1/n)$ perturbation of $\tilde{S}_{under}$ formed by replacing the within-sample estimate with the true probability as follows: $\tilde{S}_{under,ProbAdj} \triangleq \left(\frac{n+1-k}{n+1}\right)Y = \tilde{S}_{under} - \left(\frac{n+1-k}{n(n+1)}\right)Y$. The effect of this adjustment on the bias in underage may be computed as follows:

$$\begin{aligned} E(\tilde{S}_{under,ProbAdj}) - E(S_{under}) &= [E(\tilde{S}_{under,ProbAdj}) - E(\tilde{S}_{under})] \\ &\quad + [E(\tilde{S}_{under}) - E(S_{under})] \\ &= \left(\frac{n+1-k}{n+1} - \frac{n+1-k}{n}\right)E(Y) \\ &\quad + [E(\tilde{S}_{under}) - E(S_{under})]. \end{aligned}$$

Using the third assertion of Lemma 5 for the first term and Proposition 2 for the second term, we find the bias of the probability-adjusted sales in underage to be

$$\begin{aligned}
&E(\tilde{S}_{under,\,ProbAdj}) - E(S_{under})\\
&= -\left[\frac{1}{n}\left(\frac{c}{p}\right) + O\left(\frac{1}{n^2}\right)\right]E(Y)\\
&\quad + \left[\frac{1}{n}\left(\frac{c}{p}\right)E(Y) + \frac{1}{nf(y^*)}\left(\frac{c}{p}\right)\left(1-\frac{c}{p}\right) + o\left(\frac{1}{n}\right)\right]\\
&= \frac{1}{nf(y^*)}\left(\frac{c}{p}\right)\left(1-\frac{c}{p}\right) + o\left(\frac{1}{n}\right),
\end{aligned}$$

which we recognize as the full sales bias from Theorem 1 (which combines underage and overage biases). The probability adjustment has eliminated the canceling term $\pm\frac{1}{n}\left(\frac{c}{p}\right)E(Y)$ from the bias in underage.

To preserve the naive estimate in totality, having subtracted $\left(\frac{n+1-k}{n(n+1)}\right)Y$ from $\tilde{S}_{under}$ to obtain $\tilde{S}_{under,\,ProbAdj}$, we must add $\left(\frac{n+1-k}{n(n+1)}\right)Y$ to $\tilde{S}_{over}$ to obtain the *probability-adjusted naive sales estimate in overage* given Y as $\tilde{S}_{over,\,ProbAdj} \triangleq \tilde{S}_{over} + \left(\frac{n+1-k}{n(n+1)}\right)Y = \frac{1}{n}\sum_{i=1}^{k-1} X_{(i)} + \frac{n+1-k}{n(n+1)}X_{(k)}$, which is not (strictly speaking) a pure probability adjustment but, instead, is the required adjustment implied by the probability adjustment made in underage. The resulting term $\tilde{S}_{over,\,ProbAdj}$ may be interpreted as adding a small fraction of $X_{(k)}$ to the overstock estimate $\tilde{S}_{over}$. Adding this small fraction of order $O(1/n)$ greatly simplifies the bias interpretation because we now have asymptotic bias zero, plus $o(1/n)$, in overage. To see this, we use the third assertion of Lemma 5 for the first term and Proposition 1 for the second term and find the bias of the probability-adjusted sales in overage to be

$$\begin{aligned}
&E(\tilde{S}_{over,\,ProbAdj}) - E(S_{over})\\
&= \left(\frac{n+1-k}{n(n+1)}\right)E(Y) + [E(\tilde{S}_{over}) - E(S_{over})]\\
&= \left[\frac{1}{n}\left(\frac{c}{p}\right) + O\left(\frac{1}{n^2}\right)\right]E(Y)\\
&\quad + \left[-\frac{1}{n}\left(\frac{c}{p}\right)E(Y) + o\left(\frac{1}{n}\right)\right] = o\left(\frac{1}{n}\right).
\end{aligned}$$

Thus, the dominant bias is due entirely to underage after the probability adjustment: $E(\tilde{S}) - E(S) = \frac{1}{nf(y^*)}\left(\frac{c}{p}\right)\left(1-\frac{c}{p}\right) + o\left(\frac{1}{n}\right) = E(\tilde{S}_{under,\,ProbAdj}) - E(S_{under})$.

3.4.2. Interpreting the Bias as a Covariance.

Our interpretation of the bias is as a mathematical consequence of the sign of the covariance $Cov[Y,F(Y)] > 0$ of the SAA order quantity Y with $F(Y)$, which is a monotonically increasing function of its argument Y. This covariance emerges from using iterated expectations, first conditioning on Y, and then using the definition of covariance to obtain the true expected sales in underage as follows:

$$\begin{aligned}
E(S_{under}) &= E(YI_{X>Y}) = E[E(YI_{X>Y}|Y)] = E\{Y[1-F(Y)]\}\\
&= E[1-F(Y)]E(Y) + Cov\{Y,[1-F(Y)]\}.
\end{aligned}$$

Next, using Lemma 2 for $E[F(Y)]$ and recognizing that additive constants within a covariance are irrelevant, we find $E(S_{under}) = \left(\frac{n+1-k}{n+1}\right)E(Y) - Cov[Y,F(Y)] = E(\tilde{S}_{under,\,ProbAdj}) - Cov[Y,F(Y)]$, which implies that the bias $E(\tilde{S}_{under,\,ProbAdj}) - E(S_{under}) = Cov[Y,F(Y)]$ is positive because $F(Y)$ is monotonically increasing in Y.

4. Exact Results for an Exponential Distribution of Demand

In this section, we derive exact formulas for the expected true profit, the expected naive profit, and the bias, when the SAA order quantity is used under the assumption that F is an exponential distribution with mean μ. The following proposition provides the closed-form expressions for these quantities.

Proposition 4. *Exact expressions for the expected true profit π_{true}, the expected naive profit π_{naive}, and the bias $Bias_{profit}$, when F is an exponential distribution with mean μ, are as follows: $\pi_{true} = \mu\left[\frac{pk}{n+1} - c\sum_{i=n+1-k}^{n}\left(\frac{1}{i}\right)\right]$, $\pi_{naive} = \mu\left[\frac{pk}{n} - c\sum_{i=n+1-k}^{n}\left(\frac{1}{i}\right)\right]$, $Bias_{profit} = \frac{pk\mu}{n(n+1)}$.*

To prove Proposition 4, we first need the following lemma.

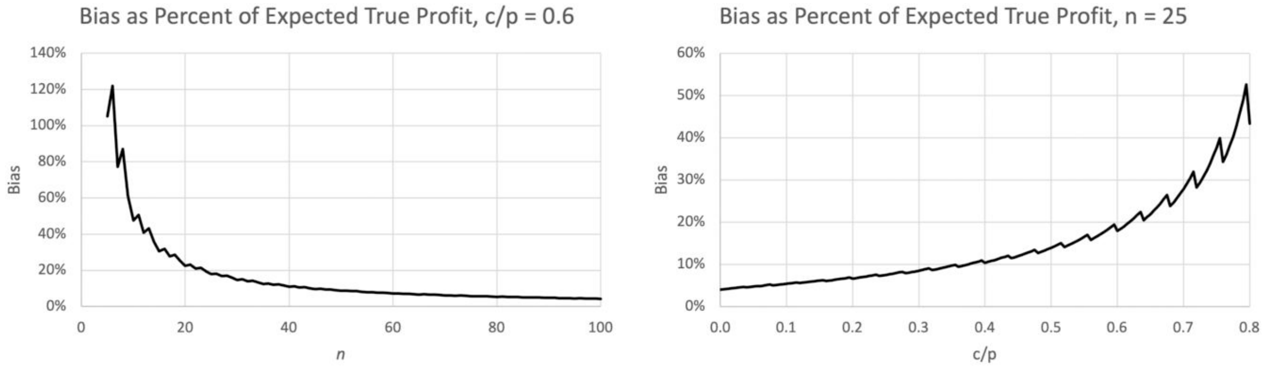
Lemma 12. *When F is an exponential distribution with mean μ, the following expectations hold exactly: $E[\min(X,Y)] = \mu\left(\frac{k}{n+1}\right)$, $E(Y) = \mu\sum_{i=n+1-k}^{n}\left(\frac{1}{i}\right)$, $E\left(\sum_{i=1}^{k-1} X_{(i)}\right) = \mu\left(k-(n+1-k)\sum_{i=n+1-k}^{n}\left(\frac{1}{i}\right)\right)$.*

To compare the exact result $Bias_{profit} = \frac{pk\mu}{n(n+1)}$ for the exponential distribution to our general asymptotic result $Bias_{profit} = \frac{p}{nf(y^*)}\left(\frac{c}{p}\right)\left(1-\frac{c}{p}\right) + o\left(\frac{1}{n}\right)$, we now show that they are equal in the exponential case, although technically we have already proven this in the general case. Nonetheless, it is of interest to see how the reciprocal of the density produces the scale factor μ for the exact bias in the exponential case.

Proposition 5. *The exact profit bias for the exponential distribution is equal to the general asymptotic profit bias formula in the sense that $Bias_{profit} = \frac{pk\mu}{n(n+1)} = \frac{p}{nf(y^*)}\left(\frac{c}{p}\right)\left(1-\frac{c}{p}\right) + o\left(\frac{1}{n}\right)$.*

We conclude this section by evaluating the size of the bias, with respect to the true expected profit. In the left plot of Figure 2, we present $Bias_{profit}$ as a percentage of π_{true}, as a function of the sample size n, for $p = 5$ and $c = 3$; we obtain qualitatively similar results for different values of p and c. We see that, especially for small

Figure 2. Bias as a Percentage of the True Expected Profit for Exponentially Distributed Demand



Notes. (Left) Percentage as a function of the sample size n for $p = 5$ and $c = 3$. (Right) Percentage as a function of c/p for $n = 25$.

values of n, the percentage is large, which then generally decreases as the sample size increases. In the right plot of Figure 2, we present the bias percentage as a function of c/p for $n = 25$; again, qualitatively similar plots are obtained for different values of n. Here, we see that the bias percentage increases with c/p. In both plots, the jagged nature of the lines is due to the value of k changing (Lemma 1).

We conclude this section by pointing out that Siegel and Wagner (2021) also derived an analytical estimate of the bias for exponentially distributed demand, except under the assumption of a parametric demand distribution. Unfortunately, comparing the bias expressions does not lead to an easy analytical conclusion. Consequently, we performed numerical experiments, similar to those herein, that demonstrate that the profit estimation bias is smaller when there is parametric information available (i.e., the fact that demand is exponentially distributed). Intuitively, more information (the parametric form of the distribution) leads to better decision making and hence less bias. However, despite the nonparametric environment (the focus of our note) exhibiting more bias, our data-driven profit adjustment performs as well as that of Siegel and Wagner (2021).

5. Numerical Experiments to Evaluate the Adjusted Profit's Bias Reduction

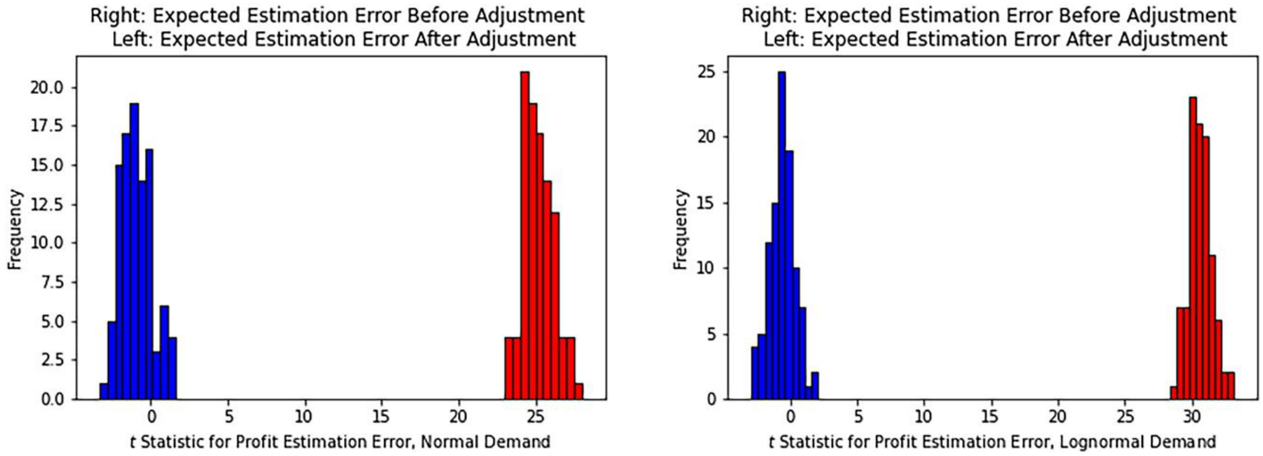
Simulations were performed for the normal and lognormal demand cases, and considerable expected estimation error reduction was observed using our proposed adjustment in Equation (13), which empirically verifies Theorem 2 for these distributions.

Our experimental primitive is as follows: we generate a sample of $n = 25$ observations x_i from the true distribution, from which we obtain the order size $y = x_{(k)}$. We then compute both the naive and the adjusted expected profit, as perceived by the newsvendor for this sample. To improve the efficiency of the Monte Carlo simulations, we decrease the noise involved in computing the true profit as follows: Instead of generating a new observation x from the true distribution and using $p\min(x,y) - cy$, we use analytical expressions for the conditional expectation of the true profit given y, namely $pE_F[\min(X,y)] - cy$, where X is drawn from the true distribution and y is held fixed at its value for this sample. For the normal distribution $X \sim N(\mu,\sigma^2)$, this conditional expectation is $p\left[y - \sigma\varphi\left(\frac{y-\mu}{\sigma}\right) - (y-\mu)\Phi\left(\frac{y-\mu}{\sigma}\right)\right] - cy$, which may be derived using basic probability theory. Similarly, for the lognormal distribution $X = e^{\mu+\sigma Z}$, with standard normal Z, the conditional expectation is $p\left\{e^{\mu+\sigma^2/2}\Phi\left(\frac{\ln y-\mu}{\sigma} - \sigma\right) + y\left[1 - \Phi\left(\frac{\ln y-\mu}{\sigma}\right)\right]\right\} - cy.$

The estimation error before adjustment is formed by subtracting the conditional expectation (of the true profit) from the naive estimate, where both were formed from this sample. Similarly, the estimation error after adjustment is formed by subtracting the conditional expectation (of the true profit) from the adjusted estimate, again, where both were formed from this sample. We compute the t statistic (testing against zero estimation error) for each measure (naive and adjusted) by repeating this procedure for 10,000 independent samples, each of size $n = 25$. The result is a pair of t statistics for the expected estimation error: one before and one after adjustment. We then repeat this procedure 100 times to obtain 100 pairs of t statistics, each based on 10,000 simulations.

In Figure 3, we plot histograms of these 100 paired t statistics for the normal (left) and lognormal (right) distributions. For both distributions, we set the mean at 200 and the standard deviation at 65; in the normal case, we set $\mu = 200$ and $\sigma = 65$, whereas the lognormal case uses $\mu = 5.248112$ and $\sigma = 0.316877$ so that its mean will be $E(X) = e^{\mu+\sigma^2/2} = 200$ and its standard deviation will be $StDev(X) = e^{\mu+\sigma^2/2}\sqrt{e^{\sigma^2} - 1} = 65$. For estimating the density in Equation (12), we use $m = 2$, which is consistent with the $m \sim Cn^{2/3}$ recommendation from Barabás (1987), for $n = 25$ and $C = 1/4$. In both cases, we set $p = 5$ and $c = 3$.

Figure 3. (Color online) Adjustment Eliminates Bias for Normal and Lognormal Demand



Note. Normal distribution (left) and lognormal distribution (right) both have mean 200 and standard deviation 65, with $p = 5$, $c = 3$, and sample size $n = 25$.

We observe strong evidence that our asymptotic adjustment eliminates the statistically significant estimation error very effectively even in these finite samples. As shown in Figure 3, for both the normal and the lognormal distributions, the unadjusted estimation error shows high statistical significance (i.e., the t values in the histograms before adjustment are considerably higher than the standard 1.96 critical value), which is successfully eliminated by the asymptotic correction (i.e., the histogram after adjustment is centered close to zero). For normal demand, the expected profit is 270.4, and the expected estimation error is 8.0, which is a nonnegligible 3.0% of the true expected profit. For lognormal demand, the expected profit is 280.9, and the expected estimation error is 6.8, which is 2.4% of the true expected profit.

We repeated these experiments with smaller values of c, which results in larger newsvendor quantiles $1 - c/p$; these larger quantiles might have caused difficulties in estimating $f(y^*)$, but the experimental results (omitted for brevity) were qualitatively identical to those in Figure 3.

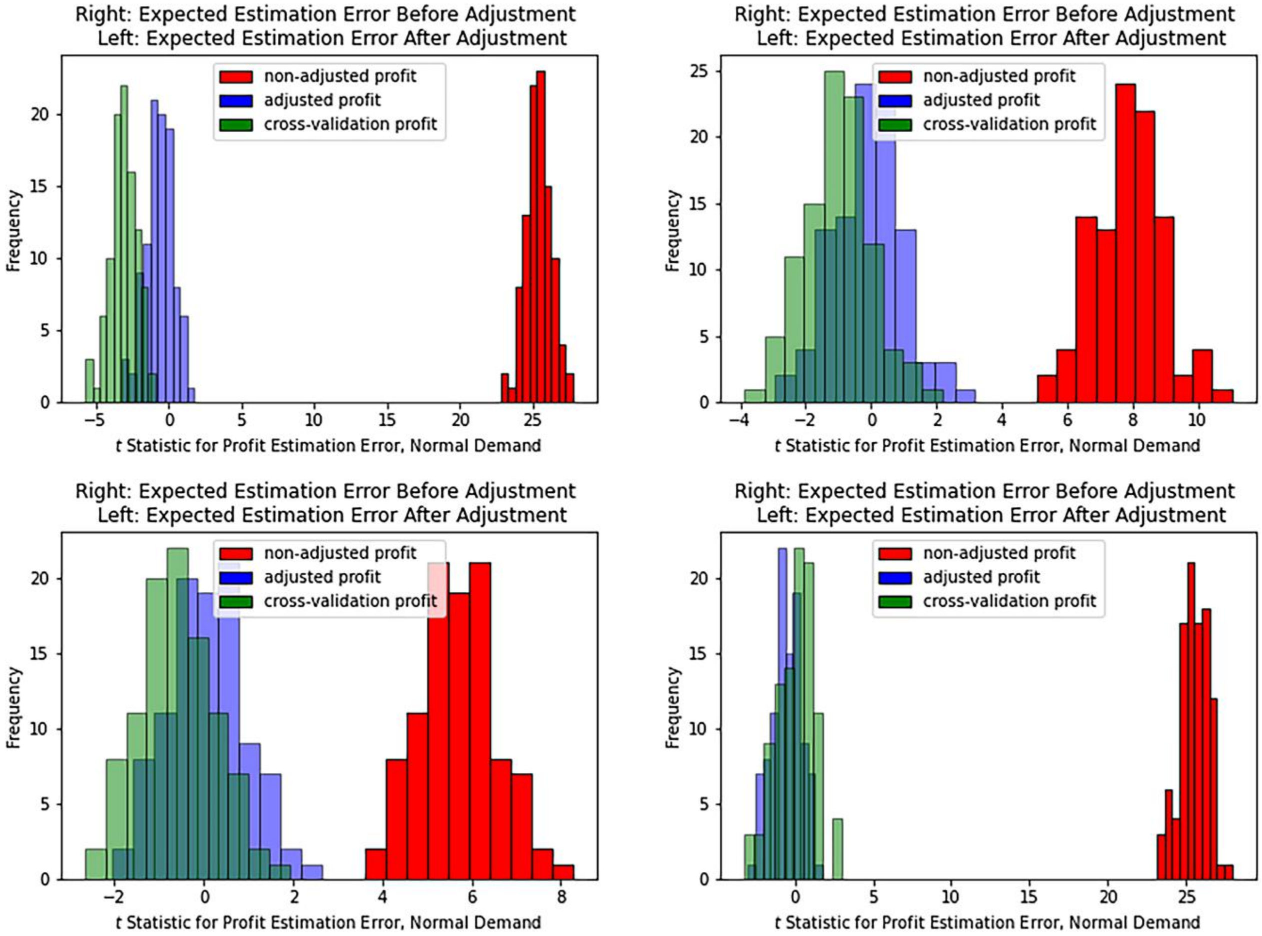
5.1. Comparison with Cross-Validation

In the language of machine learning, we use our sample of data $(X_1, \ldots, X_n)$ as a training data set to calculate Y_{saa}. However, evaluating the performance (i.e., expected profit) of an algorithm on the same data used to train it is a mistake. Typically, in machine learning, data are split into training and testing data sets. The training data set would be used to calculate Y_{saa} and then the *testing* data set would be used to estimate the expected profit. Of course, this expected profit estimation would depend on the random split of the data into training and testing data sets. A better approach is to use cross-validation to estimate the expected profit. In, say, fivefold cross-validation, the data are split into five equally sized sets, or folds, of data. Each of the folds serves as a testing data set, with the remaining four folds serving collectively as the training data. Averaging the five test data set expected profit estimations results in the cross-validated estimate of expected profit. Cross-validation is one of the standard approaches in machine learning to estimate the performance of an algorithm on new data. In this section, we examine how cross-validation performs with respect to our bias adjustment.

We reuse the hypothesis-testing experimental design of this section, where we add the t statistic histograms for fivefold cross-validated estimated expected profits for sample sizes $n \in \{25, 250, 500\}$ and leave-one-out cross-validation for $n = 25$. We report our findings in Figure 4 for normally distributed demand; we observe qualitatively similar results for lognormally distributed demand, which we omit for brevity.

We see that fivefold cross-validation (in green, on left) significantly reduces bias compared with the unadjusted expected profit (in red, on right), as expected, but does not eliminate the bias for smaller sample sizes. For the smaller sample sizes, our analytical bias adjustment reduces bias better than cross-validation: observe that the histograms for our bias-adjusted expected profits (in blue, in middle) are centered more closely around zero, whereas the cross-validated histograms are centered to the left of zero, indicating that cross-validation actually overcompensates and results in a *negative* bias. However, leave-one-out cross-validation performs as well as our analytical correction for small samples sizes, as does fivefold cross-validation on larger sample sizes. Thus, applied properly, cross-validation can match the bias elimination of our analytical correction. However, we argue that our analytical adjustment is preferred to cross-validation because it (1) provides intuition about why there is a bias (see the discussion of underage and

Figure 4. (Color online) Our Analytical Adjustment Compared with Cross-Validation for Normally Distributed Demand with Mean 200, Standard Deviation 65, and $p = 5$, $c = 3$



Note. The types of cross-validation (CV) and sample sizes are as follows: (top left) fivefold CV and $n = 25$, (top right) fivefold CV and $n = 250$, (bottom left) fivefold CV and $n = 500$, (bottom right) leave-one-out CV and $n = 25$.

covariance in Section 3.4), (2) is faster than cross-validation (the calculation of the adjustment is effectively instantaneous), and (3) does not require proper tuning as in the cross-validation case (e.g., selecting the right number of folds versus leave-one-out cross-validation).

6. Conclusion

In this note, we identify a statistically significant error in naively estimating the expected profit in a data-driven newsvendor model, and we show how to correct the error while assuming only a smooth density, without parametric assumptions. In particular, we analyze a newsvendor model, where the SAA order quantity is calculated by substituting the empirical demand distribution, generated by a sample of demand data, for the unknown true distribution. We prove that using the same empirical distribution to estimate the expected profit of the SAA order quantity results in a positive asymptotic bias, which we derive in closed form. Furthermore, we provide an in-depth interpretation of the bias, connecting it mathematically with a covariance term. The bias expression allows us to adjust the expected profit estimate to obtain an asymptotically unbiased expected profit estimate using only information from the sample, whereas the true distribution remains unknown. The bias is nonnegligible: In our numerical experiments, which demonstrate that the estimation error is statistically significant and the adjusted estimation error is generally not significantly different from zero, the bias is approximately 2.4%–3.0% of the true expected profit, when demand is generated by a lognormal and normal distribution, respectively. Exact numerical experiments for exponentially distributed demand demonstrate that the bias can be a much larger percentage of the true expected profit. We also interpret our results intuitively: Although overage contributes negative sales bias, underage contributes a larger magnitude of positive sales bias, which dominates. This implies that the positive profit bias is due to underage.

We conclude by briefly discussing future research directions. First, our results could potentially be extended to a contextual newsvendor model, where a demand observation X is paired with a contextual vector $\mathbf{Z} = (Z_1, \dots, Z_m)$. A machine learning model could be used to predict the distribution of X from $\mathbf{Z}$. If such a prediction model provides an empirical demand distribution, conditional on $\mathbf{Z}$, then the newsvendor could set Y_{saa} to the $1 - c/p$ quantile of this conditional distribution. Ban and Rudin (2019) and Bertsimas and Kallus (2020) study similar models from a variety of perspectives but do not correct for any biases.

References

Ban G-Y, Rudin C (2019) The big data newsvendor: Practical insights from machine learning. *Oper. Res.* 67(1):90–108.

Ban G-Y, Gallien J, Mersereau AJ (2019) Dynamic procurement of new products with covariate information: The residual tree method. *Manufacturing Service Oper. Management* 21(4):798–815.

Barabás B (1987) Estimation of density functions by order statistics. *Periodica Math. Hungarica* 18:115–122.

Bertsimas D, Kallus N (2020) From predictive to prescriptive analytics. *Management Sci.* 66(3):1025–1044.

Chu L, Shanthikumar J, Shen Z (2008) Solving operational statistics via a Bayesian analysis. *Oper. Res. Lett.* 36:110–116.

Gupta V, Rusmevichientong P (2021) Small-data, large-scale linear optimization with uncertain objectives. *Management Sci.* 67(1):220–241.

Gupta V, Huang M, Rusmevichientong P (2023) Debiasing in-sample policy performance for small-data, large-scale optimization. *Oper. Res.* Forthcoming.

He B, Dexter F, Macario A, Zenios S (2012) The timing of staffing decisions in hospital operating rooms: Incorporating workload heterogeneity into the newsvendor problem. *Manufacturing Service Oper. Management* 14(1):99–114.

Ito S, Yabe A, Fujimaki R (2018) Unbiased objective estimation in predictive optimization. Dy J, Krause A, eds. *Proc. 35th Internat. Conf. on Machine Learn.*, vol. 80 (PMLR, New York), 2176–2185.

Kleywegt A, Shapiro A, Homem-De-Mello T (2002) The sample average approximation method for stochastic discrete optimization. *SIAM J. Optim.* 12(2):479–502.

Levi R, Perakis G, Uichanco J (2015) The data-driven newsvendor problem: New bounds and insights. *Oper. Res.* 63(6): 1294–1306.

Levi R, Roundy R, Shmoys D (2007) Provably near-optimal sampling-based policies for stochastic inventory control models. *Math. Oper. Res.* 32(4):821–839.

Liyanage L, Shanthikumar J (2005) A practical inventory control policy using operational statistics. *Oper. Res. Lett.* 33:341–348.

Serfling R (1980) *Approximation Theorems of Mathematical Statistics* (Wiley, New York).

Siegel A, Wagner M (2021) Profit estimation error in the newsvendor model under a parametric demand distribution. *Management Sci.* 67(8):4863–4879.

Tusnády G (1974) On testing density functions. *Periodica Math. Hungarica* 5:161–169.

Andrew F. Siegel is emeritus professor at the University of Washington, Seattle. The Nelson-Siegel model for interest rates is in use at central banks around the world. His research interests include optimization using random data, asset pricing, portfolio geometry, bond markets, genetics, statistics, and probability. His business statistics textbook (now with Michael R. Wagner as co-author) has been translated into Russian and Chinese.

Michael R. Wagner is an associate professor of operations management and a Neal and Jan Dempsey endowed faculty fellow at the Michael G. Foster School of Business, University of Washington, Seattle. His research interests are in crowdsourcing, machine learning, and optimization under uncertainty (stochastic, online, and robust optimization, as well as hybrids thereof). He is co-author, with Andrew F. Siegel, of the eighth edition of a business statistics textbook.