On the Bias of Precision Estimation Under Separate Sampling

Cancer Informatics
Volume 18: 1–9
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1176935119860822

Shuilian Xie and Ulisses M Braga-Neto

Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA.

ABSTRACT: Observational case-control studies for biomarker discovery in cancer studies often collect data that are sampled separately from the case and control populations. We present an analysis of the bias in the estimation of the precision of classifiers designed on separately sampled data. The analysis consists of both theoretical and numerical results, which show that classifier precision estimates can display strong bias under separating sampling, with the bias magnitude depending on the difference between the true case prevalence in the population and the sample prevalence in the data. We show that this bias is systematic in the sense that it cannot be reduced by increasing sample size. If information about the true case prevalence is available from public health records, then a modified precision estimator that uses the known prevalence displays smaller bias, which can in fact be reduced to zero as sample size increases under regularity conditions on the classification algorithm. The accuracy of the theoretical analysis and the performance of the precision estimators under separate sampling are confirmed by numerical experiments using synthetic and real data from published observational case-control studies. The results with real data confirmed that under separately sampled data, the usual estimator produces larger, ie, more optimistic, precision estimates than the estimator using the true prevalence value.

KEYWORDS: Precision, recall, bias, classification, observational study, experimental design

RECEIVED: May 9, 2019. **ACCEPTED:** June 2, 2019.

TYPE: Methodology

FUNDING: The author(s) received no financial support for the research, authorship, and/or publication of this article.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Ulisses M Braga-Neto, Department of Electrical and Computer Engineering, Texas A& M University, College Station, TX 77843, USA.
Email: ulisses@ece.tamu.edu

Introduction

Biomarker discovery is typically attempted by means of observational case-control studies where classification techniques are applied to high-throughput measurement technologies, such as DNA microarrays,[1,2] next-generation RNA sequencing (RNA-seq),[3] or "shotgun" mass spectrometry.[4] The validity and reproducibility of the results depend critically on the availability of accurate and unbiased assessment of classification accuracy.[5,6]

The vast majority of published methods in the statistical learning literature make the assumption, explicitly or implicitly, that the data for training and accuracy assessment are sampled randomly, or unrestrictedly, from the mixture of the populations. However, observational case-control studies in biomedicine typically proceed by collecting data that are sampled with restrictions. The most common restriction, and the one that is studied in this article, is that the data are sampled separately from the case and control populations. That creates an important issue in the application of traditional statistical learning techniques to biomedical data, because there is no meaningful estimator of case prevalences under separate sampling. Therefore, any methodology that directly or indirectly uses estimates of case prevalence could be severely biased.

Precision and *Recall* have become very popular classification accuracy metrics in the statistical learning literature.[7–9] The recall does not depend on the prevalence, while the precision does. Therefore, we investigate in this article the bias of the precision estimator when the typical separate sampling design used in case-control studies is not properly taken into account.

A similar study was conducted previously into the accuracy of cross-validation under separate sampling.[10] It was shown in that study that the usual "unbiasedness" property of k-fold cross-validation does not hold under separate sampling. In fact, the bias can in fact be substantial and systematic, ie, not reducible under increasing sample size. In Braga-Neto et al,[10] modified k-fold cross-validation estimators were proposed for the class-specific error rates. In the case where the true case prevalence is known, those estimators can be combined into an estimator of the overall error rate, which satisfies the usual "unbiasedness" property of cross-validation.

By contrast, the present paper employs analytical and numerical methods to investigate precision estimation under separate sampling. We show that the usual precision estimator is asymptotically unbiased as sample size increases, under the condition that the classification rule has a finite Vapnik-Chervonenkis (VC) dimension. However, under separate sampling, we show that the usual precision estimator will in general display a systematic bias, which cannot be reduced by increasing sample size, if the observed prevalence of cases in the data is different from the true prevalence in the population of interest, and the bias is larger the more different they are. In particular, the bias tends to be large when the true prevalence is small but the training data contain an equal number of examples from both classes, which is a common scenario in practice. If the true case prevalence is known (eg, from public health records), then a modified precision estimator that uses the

known prevalence is shown to be asymptotically unbiased in the separate sampling case, under the condition that the classification rule is sufficiently stable as sample size increases. All of these theoretical results, and the approximations used to derive them, are verified by numerical experiments using both synthetic and real data from published studies.

Materials and Methods

In this section, we define and study the various error rates of interest in this study, including precision and recall.

Population performance metrics

The *feature* vector $\mathbf{X} \in R^d$ summarizes numerical characteristics of a patient (eg, blood concentrations of given proteins). The *label* $Y \in \{0,1\}$ is defined as $Y = 0$ if the patient is from the control population, and $Y = 1$ if the patient is from the case population.

The *prevalence* is defined by

$$\text{prev}=P(Y = 1) \tag{1}$$

ie, the probability that a randomly selected individual is a case subject. The prevalence plays a fundamental role in the sequel.

A *classifier* $\psi : R^d \to \{0,1\}$ assigns $\mathbf{X}$ to the control or case population, according to whether $\psi(\mathbf{X}) = 0$ or $\psi(\mathbf{X}) = 1$, respectively. The classification sensitivity and specificity are defined as follows:

$$\text{sens} = P(\psi(\mathbf{X}) = 1 \mid Y = 1) \tag{2}$$

$$\text{spec} = P(\psi(\mathbf{X}) = 0 \mid Y = 0) \tag{3}$$

The closer both are to 1, the more accurate the classifier is. A noteworthy property of the sensitivity and specificity is that they *do not depend on the prevalence*.

Other common performance metrics for a classifier are the *false-positive* (FP), *false-negative* (FN), *true-positive* (FP), and *true-negative* (FN) rates, given by

$$\text{FP} = P(\psi(\mathbf{X}) = 1, Y = 0) \tag{4}$$

$$= (1-\text{spec})\times(1-\text{prev}) \tag{5}$$

$$\text{FN} = P(\psi(\mathbf{X}) = 0, Y = 1) = (1-\text{sens})\times\text{prev} \tag{6}$$

$$\text{TP} = P(\psi(\mathbf{X}) = 1, Y = 1) = \text{sens}\times\text{prev} \tag{7}$$

$$\text{TN} = P(\psi(\mathbf{X}) = 0, Y = 0) = \text{spec}\times(1-\text{prev}) \tag{8}$$

Unlike sensitivity and specificity, the previous performance metrics *do* depend on the prevalence.

Note that

$$\text{prev}=\text{FN}+\text{TP}, \quad 1-\text{prev}=\text{FP}+\text{TN} \tag{9}$$

$$\text{sens}=\frac{\text{TP}}{\text{TP}+\text{FN}}, \text{spec}=\frac{\text{TN}}{\text{TN}+\text{FP}} \tag{10}$$

Finally, we define the precision and recall accuracy metrics. Precision measures the likelihood that one has a true case given that the classifier outputs a case:

$$\text{prec}=P(Y = 1 \mid \psi(\mathbf{X}) = 1) \tag{11}$$

Applying Bayes'Theorem and using previously derived relationships reveal that

$$\text{prec} = \frac{\text{TP}}{\text{TP}+\text{FP}} = \frac{\text{sens}\times\text{prev}}{\text{sens}\times\text{prev}+(1-\text{spec})\times(1-\text{prev})} \tag{12}$$

On the other hand, recall is simply the sensitivity:

$$\text{rec}=\text{sens}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{13}$$

It follows that precision depends on the prevalence, but recall does not.

Estimated performance metrics

In practice, the performance metrics defined in the previous section need to be estimated from sample data $S_n = \{(\mathbf{X}_1, Y_1), \ldots, \mathbf{X}_n, Y_n)\}$. Let $\hat{P}$ denote the empirical probability measure defined by S_n. The estimator of prevalence is

$$\widehat{\text{prev}}=\hat{P}(Y = 1)=\frac{1}{n}\sum_{i=1}^{n} I_{Y_i=1} \tag{14}$$

where $I_A = 1$ if A is true and $I_A = 0$ if A is false. Similarly,

$$\widehat{\text{FP}} = \hat{P}(\psi(\mathbf{X}) = 1, Y = 0)=\frac{1}{n}\sum_{i=1}^{n} I_{\{\psi(\mathbf{X}_i)=1, Y_i=0\}} \tag{15}$$

$$\widehat{\text{FN}} = \hat{P}(\psi(\mathbf{X}) = 0, Y = 1)=\frac{1}{n}\sum_{i=1}^{n} I_{\{\psi(\mathbf{X}_i)=0, Y_i=1\}} \tag{16}$$

$$\widehat{\text{TP}} = \hat{P}(\psi(\mathbf{X}) = 1, Y = 1)=\frac{1}{n}\sum_{i=1}^{n} I_{\{\psi(\mathbf{X}_i)=1, Y_i=1\}} \tag{17}$$

$$\widehat{\text{TN}} = \hat{P}(\psi(\mathbf{X}) = 0, Y = 0)=\frac{1}{n}\sum_{i=1}^{n} I_{\{\psi(\mathbf{X}_i)=0, Y_i=0\}} \tag{18}$$

The remaining performance metrics estimators are defined analogously, using equations (10), (12), and (13):

$$\widehat{\text{spec}}=\frac{\widehat{\text{TN}}}{\widehat{\text{TN}}+\widehat{\text{FP}}}=\frac{\sum_{i=1}^{n} I_{\{\psi(\mathbf{X}_i)=0,Y_i=0\}}}{\sum_{i=1}^{n} I_{Y_i=0}} \tag{19}$$

$$\widehat{\text{prec}}=\frac{\widehat{\text{TP}}}{\widehat{\text{TP}}+\widehat{\text{FP}}}=\frac{\sum_{i=1}^{n} I_{\{\psi(\mathbf{X}_i)=1,Y_i=1\}}}{\sum_{i=1}^{n} I_{\psi(\mathbf{X}_i)=1}} \tag{20}$$

$$\widehat{\text{rec}}=\widehat{\text{sens}}=\frac{\widehat{\text{TP}}}{\widehat{\text{TP}}+\widehat{\text{FN}}}=\frac{\sum_{i=1}^{n} I_{\{\psi(\mathbf{X}_i)=1,Y_i=1\}}}{\sum_{i=1}^{n} I_{Y_i=1}} \tag{21}$$

Mixture and separate sampling

The usual scenario in Statistical Learning is to assume that $S_n=\{(\mathbf{X}_1,Y_1),\ldots,(\mathbf{X}_n,Y_n)\}$ is an independent and identically distributed (i.i.d.) sample from the true distribution of the pair $(\mathbf{X},Y)$. That makes S_n a sample from the *mixture* of populations, where each label Y_i is distributed as

$$\begin{aligned} P(Y_i=1)&=\text{prev}\\ P(Y_i=0)&=1-\text{prev}\end{aligned} \tag{22}$$

for $i=1,\ldots,n$. Under mixture sampling, $N_0=\sum_{i=0}^{n} I_{Y_i=0}$ and $N_1=\sum_{i=1}^{n} I_{Y_i=1}=n-N_0$ are binomial random variables, with parameters $(n,1-\text{prev})$ and (n,prev), respectively.

By contrast, observational case-control studies in biomedicine typically proceed by collecting data from the populations separately, where the separate sample sizes n_0 and n_1, with $n_0+n_1=n$, are pre-determined and nonrandom, ie, sampling occurs with the restriction $N_0=\sum_{i=0}^{n} I_{Y_i=0}=n_0$ (or, equivalently, $N_1=\sum_{i=1}^{n} I_{Y_i=1}=n_1$). Therefore, all probabilities and expectations over the sample are conditional on $N_0=n_0$. The restriction means that the labels $Y_1,\ldots,Y_n$ are no longer independent, even though the feature vectors $\mathbf{X}_1,\ldots,\mathbf{X}_n$ are still independent given the labels. In fact, under separate sampling, only the order of the labels $Y_1,\ldots,Y_n$ may be random. Thus, $f(Y_1,\ldots,Y_n\,|\,N_0=n_0)$ is a discrete uniform distribution over all $\binom{n}{n_0}$ possible orderings. This can also be obtained by direct computation, as follows:

$$\begin{aligned} f(Y_1,\ldots,Y_n\,|\,N_0=n_0)&=\frac{f(Y_1,\ldots,Y_n,N_0=n_0)}{P(N_0=n_0)}\\ &=\begin{cases}\dfrac{\text{prev}^{n_1}(1-\text{prev})^{n_0}}{\binom{n}{n_0}\text{prev}^{n_1}(1-\text{prev})^{n_0}}=\dfrac{1}{\binom{n}{n_0}}, & \text{if } \sum_{i=1}^{n} I_{Y_i=0}=n_0\\ 0, & \text{otherwise}\end{cases}\end{aligned} \tag{23}$$

It is not difficult to verify that under equation (23), the marginal distribution of each label Y_i is given by

$$\begin{aligned} P(Y_i=1\,|\,N_0=n_0)&=\frac{n_1}{n}\overset{\Delta}{=}r\\ P(Y_i=0\,|\,N_0=n_0)&=\frac{n_0}{n}=1-r\end{aligned} \tag{24}$$

for $i=1,\ldots,n$, where r is the (fixed) sample size ratio under separate sampling. Comparing equations (22) and (24) reveals the main difference between mixture and separate sampling.

Bias of the precision estimator

In this section, we present a theoretical large sample analysis of the bias of the estimators discussed previously, focusing on the precision estimator. Estimation bias is defined as the expectation over the sample data S_n of the difference between the estimated and true quantities.

The situation is clear with the estimator of the prevalence itself, given by equation (14). Under mixture sampling, we have

$$E[\widehat{\text{prev}}]=\frac{1}{n}\sum_{i=1}^{n} E[I_{Y_i=1}]=P(Y_1=1)=\text{prev} \tag{25}$$

so the estimator is unbiased (in addition, as n increases, $\text{Var}(\widehat{\text{prev}})\to 0$ and $\widehat{\text{prev}}\to\text{prev}$ in probability, by the law of large numbers). However, under separate sampling,

$$\begin{aligned} E[\widehat{\text{prev}}\,|\,N_0=n_0]&=\frac{1}{n}\sum_{i=1}^{n} E[I_{Y_i=1}\,|\,N_0=n_0]\\ &=P(Y_1=1\,|\,N_0=n_0)=r\end{aligned} \tag{26}$$

according to equation (24). This also follows directly from the fact that $\widehat{\text{prev}}$ becomes a constant estimator, $\widehat{\text{prev}}\equiv r$, according to equation (14). Thus,

$$\begin{aligned} \text{Bias}_{\text{sep}}(\widehat{\text{prev}})&=E[\widehat{\text{prev}}-\text{prev}\,|\,N_0=n_0]\\ &=r-\text{prev}\end{aligned} \tag{27}$$

Assuming that the sample size ratio $r=n_1/n$ is held constant as n increases (eg, under the common balanced design case, $n_0=n_1=n/2$), then this bias cannot be reduced with increased sample size. Furthermore, the bias is larger the further away prev is from r. In particular, the bias tends to be large when prev is small and $r=1/2$, which is a common scenario in practice.

The situation for $\widehat{\text{FP}}$, $\widehat{\text{FN}}$, $\widehat{\text{FP}}$, and $\widehat{\text{TN}}$ is more complicated. First, we are interested in a classifier ψ_n derived by a classification rule from the sample data $S_n=\{(\mathbf{X}_1,Y_1),\ldots,\mathbf{X}_n,Y_n)\}$. Therefore, all expectations and probabilities in the previous sections are conditional on S_n. Under mixture sampling, the powerful *Vapnik-Chervonenkis Theorem* can be applied to show that all of these estimators are asymptotically unbiased, provided that

classification rule has a finite *VC Dimension*.[11] This includes many useful classification algorithms such as Linear Discriminant Analysis (LDA), linear Support Vector Machines (SVMs), perceptrons, polynomial-kernel classifiers, certain decision trees, and neural networks, but it excludes nearest-neighbor classifiers, for example. Classification rules with finite VC dimension do not cut the feature space in complex ways and are thus generally robust against overfitting.

Assuming mixture sampling and a classification algorithm with finite VC dimension V_C, it can be shown that (the details are omitted; see Braga-Neto and Dougherty[6] for a similar argument)

$$\mathrm{Bias}_{\mathrm{mix}}(\widehat{\mathrm{FP}}) \leq 8\sqrt{\frac{V_C \log(n+1)+4}{2n}} \tag{28}$$

so that the bias vanishes as $n \to \infty$. Similar inequalities apply to $\widehat{\mathrm{FN}}$, $\widehat{\mathrm{FP}}$, and $\widehat{\mathrm{TN}}$. These are distribution-free results; hence, vanishingly small bias is guaranteed if $n \gg V_C$, regardless of the feature-label distribution. For linear classification rules, $V_C = d+1$, where d is the dimensionality of the feature vector. In this case, the $\widehat{\mathrm{FP}}$, $\widehat{\mathrm{FN}}$, $\widehat{\mathrm{FP}}$, and $\widehat{\mathrm{TN}}$ estimators are essentially unbiased if $n \gg d$.

Next we consider the bias of the precision and recall estimators under mixture sampling (the analysis for the sensitivity and specificity estimators is similar; in fact, the former is just the recall estimator). We will make use of the following approximation for the expectation of a ratio of two random variables W and Z (see Appendix 1 for the derivation of this approximation and the conditions under which it is valid):

$$E\left[\frac{W}{Z}\right] \approx \frac{E[W]}{E[Z]} \tag{29}$$

The approximation is quite accurate if W and Z are around $E[W]$ and $E[Z]$, respectively (it is asymptotically exact as $W \to E[W]$ and $Z \to E[Z]$). For the precision estimator,

$$\begin{aligned} E[\widehat{\mathrm{prec}}] &= E\left[\frac{\widehat{\mathrm{TP}}}{\widehat{\mathrm{TP}}+\widehat{\mathrm{FP}}}\right] \approx \frac{E[\widehat{\mathrm{TP}}]}{E[\widehat{\mathrm{TP}}+\widehat{\mathrm{FP}}]} \\ &\approx \frac{E[\mathrm{TP}]}{E[\mathrm{TP}+\mathrm{FP}]} \approx E\left[\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}\right] = E[\mathrm{prec}] \end{aligned} \tag{30}$$

for a sufficiently large sample, where we used the previously established asymptotic unbiasedness of $\widehat{\mathrm{TP}}$, $\widehat{\mathrm{TP}}$, and $\widehat{\mathrm{FN}}$. An entirely similar derivation shows that $E[\widehat{\mathrm{rec}}] = E[\mathrm{rec}]$. Hence, for "well-behaved" classification algorithms (those with finite VC dimension), both the precision and recall estimators are asymptotically unbiased under mixture sampling.

We are not aware of the existence of a VC theory for separate sampling at this time. To obtain approximate results for the separate sampling case, we will assume instead that at large enough sample sizes, the classifier ψ is nearly constant, and invariant to the sample. This assumption is not unrelated to the finite VC dimension assumption made in the case of mixture sampling. Many of the same classification algorithms that have finite VC dimension, such as LDA and linear SVMs, will also become nearly constant as sample size increases. In this case, we have

$$\begin{aligned} E[\widehat{\mathrm{TP}} \mid N_0 = n_0] &= \frac{1}{n}\sum_{i=1}^{n} E[I_{\{\psi(X_i)=1, Y_i=1\}} \mid N_0 = n_0] \\ &= P(\psi(\mathbf{X}_1)=1, Y_1=1 \mid N_0 = n_0) \\ &= P(\psi(\mathbf{X}_1)=1 \mid Y_1=1) P(Y_1=1 \mid N_0 = n_0) \\ &= \mathrm{sens} \times r \end{aligned} \tag{31}$$

where we used the fact that the event $\{\psi(\mathbf{X}_1)=1\}$ is independent of N_0 given Y_1 and equation (24). Note that the equality $P(\psi(\mathbf{X}_1)=1 \mid Y_1=1) = \mathrm{sens}$ depends on the fact that ψ is assumed to be constant, so that $(\mathbf{X}_1, Y_1)$ behaves as an independent test point (also because of a constant ψ, there is no expectation around sens). Hence, $\widehat{\mathrm{TP}}$ is biased under separate sampling, with

$$\mathrm{Bias}_{\mathrm{sep}}(\widehat{\mathrm{TP}}) = \mathrm{sens} \times r - \mathrm{TP} = \mathrm{sens} \times (r - \mathrm{prev}) \tag{32}$$

As in the case with the bias of $\widehat{\mathrm{prev}}$ under separate sampling, the bias of $\widehat{\mathrm{TP}}$ cannot be reduced with increasing sample size. The bias is in fact larger the more sensitive the classifier is. One can derive similar results for $\widehat{\mathrm{FP}}$, $\widehat{\mathrm{FN}}$, and $\widehat{\mathrm{TN}}$.

The recall estimator is approximately unbiased under separate sampling:

$$\begin{aligned} E[\widehat{\mathrm{rec}} \mid N_0 = n_0] &= E\left[\frac{\widehat{\mathrm{TN}}}{\widehat{\mathrm{TN}}+\widehat{\mathrm{FP}}} \mid N_0 = n_0\right] \\ &= E\left[\frac{\widehat{\mathrm{TP}}}{\widehat{\mathrm{prev}}} \mid N_0 = n_0\right] = \frac{E[\widehat{\mathrm{TP}} \mid N_0 = n_0]}{r} \\ &= \frac{\mathrm{sens} \times r}{r} = \mathrm{sens} = \mathrm{rec} \end{aligned} \tag{33}$$

This is a consequence of recall's not being a function of the prevalence. However, for the precision estimator,

$$\begin{aligned} E[\widehat{\mathrm{prec}} \mid N_0 = n_0] &= E\left[\frac{\widehat{\mathrm{TP}}}{\widehat{\mathrm{TP}}+\widehat{\mathrm{FP}}} \mid N_0 = n_0\right] \\ &\approx \frac{E[\widehat{\mathrm{TP}} \mid N_0 = n_0]}{E[\widehat{\mathrm{TP}}+\widehat{\mathrm{FP}} \mid N_0 = n_0]} \\ &= \frac{\mathrm{sens} \times r}{\mathrm{sens} \times r + (1-\mathrm{spec}) \times (1-r)} \\ &\neq \frac{\mathrm{sens} \times \mathrm{prev}}{\mathrm{sens} \times \mathrm{prev} + (1-\mathrm{spec}) \times (1-\mathrm{prev})} = \mathrm{prec} \end{aligned} \tag{34}$$

The precision estimator is thus biased under separate sampling unless the true prevalence matches exactly the sample ratio $r = n_1 / n$; the bias is larger the further away prev is from r.

In case the true prevalence is known, eg, from public health records and government databases, then we show below that the following estimator of the precision,

$$\widehat{\text{prec}}^{\text{prev}} = \frac{\widehat{\text{sens}} \times \text{prev}}{\widehat{\text{sens}} \times \text{prev} + (1 - \widehat{\text{spec}}) \times (1 - \text{prev})} \tag{35}$$

which is based on equation (12), is an asymptotically unbiased estimator of the precision under either mixture or separate sampling. Asymptotic unbiasedness in the mixture sampling case can be shown by repeating the steps in the analysis of the ordinary precision estimator. Under separate sampling, we have

$$\begin{aligned} & E[\widehat{\text{prec}}^{\text{prev}} \mid N_0 = n_0] \\ & \approx \frac{E[\widehat{\text{sens}} \mid N_0] \times \text{prev}}{E[\widehat{\text{sens}} \mid N_0] \times \text{prev} + (1 - E[\widehat{\text{spec}} \mid N_0]) \times (1 - \text{prev})} \\ & = \frac{\text{sens} \times \text{prev}}{\text{sens} \times \text{prev} + (1 - \text{spec}) \times (1 - \text{prev})} = \text{prec} \end{aligned} \tag{36}$$

since $E[\widehat{\text{sens}} \mid N_0 = n_0] = \text{sens}$ and $E[\widehat{\text{spec}} \mid N_0 = n_0] = \text{spec}$, as can be easily shown. Hence, $\widehat{\text{prec}}^{\text{prev}}$ is an asymptotically unbiased estimator of the precision under either mixture or separate sampling. The ordinary precision estimator $\widehat{\text{prec}}$ should not be used under separate sampling, or large and irreducible bias may occur. On the other hand, in the impossibility of obtaining information on the true prevalence value, then no meaningful estimator of the precision is possible.

Results and Discussion

In this section, we employ synthetic and real-world data to investigate the accuracy of the analysis in the previous section and the performance of the precision estimator under separate sampling. Corresponding results for mixture sampling and the recall estimator can be found in the Supplementary Material.

Experiments with synthetic data

We performed a set of experiments employing synthetic data from a homoskedastic Gaussian model, consisting of three-dimensional class-conditional distributions $N(\boldsymbol{\mu}_i, \Sigma)$, for $i = 0,1$, with $\boldsymbol{\mu}_0 = (0,0,0)$, $\boldsymbol{\mu}_1 = (0,0,\theta)$, where $\theta > 0$ is a parameter governing the separation between the classes, and $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \sigma_3^2)$ (ie, a matrix with $\sigma_1^2, \sigma_2^2, \sigma_3^2$ on the diagonal and zeros off the diagonal). We consider two sample sizes, $n = 30$ and $n = 200$, so that we can compare the results for small and large sample sizes. All experiments with separate sampling are performed with sample size ratio $r = n_1 / n \in [0.1, 0.9]$. The synthetic data parameters are summarized in Table 1.

For each value of r and prev, we repeat the following process 1000 times and average the results to estimate expected values:

Table 1. Synthetic data parameters.

PARAMETER	VALUE
Dimensionality/feature size	$D = 3$
Mean difference	$\theta = 2$
Covariance matrix	$\sigma_1^2 = 0.5, \sigma_2^2 = 0.5, \sigma_3^2 = 1$
Sample size	$n = 30,200$
Sample size ratio r	$r = 0.1,0.3,0.5,0.7,0.9$
True prevalence	prev = 0.1,0.3,0.5,0.7,0.9

1. Generate sample data S_n of size n according to r (separate sampling) or prev (mixture sampling);
2. Train a classifier using one of three classification rules:[12] LDA, 3-Nearest Neighbors (3NN), and a nonlinear Radial-Basis Function Support Vector Machine (RBF-SVM).
3. Obtain recall and precision estimates. Compute both the usual precision estimate $\widehat{\text{prec}}$ and the modified precision estimate $\widehat{\text{prec}}^{\text{prev}}$.
4. Obtain accurate estimates of the true precision values using a test set of size 10 000.

Figure 1 displays the results of the experiment. Note that there is only one curve for the traditional precision estimator $\widehat{\text{prec}}$ because it does not employ the actual value of prev. The values of $\widehat{\text{prec}}$ and $\widehat{\text{prec}}^{\text{prev}}$ coincide when prev $= r$, as expected. However, as the values of prev and r become different, their values become quite different, and $\widehat{\text{prec}}^{\text{prev}}$ displays much less bias, ie, it tracks the true precision much more closely, than $\widehat{\text{prec}}$. At the small sample size $n = 30$, both estimators display bias, which is however much larger overall for $\widehat{\text{prec}}$ than for $\widehat{\text{prec}}^{\text{prev}}$. At the large sample size $n = 200$, the bias of $\widehat{\text{prec}}^{\text{prev}}$ nearly disappears for LDA and is reduced for the other classification rules. We note that among these classification rules, LDA is the only one with a finite VC dimension; the fact that the bias in this case shrinks to zero as sample size increases confirms the results of the theoretical analysis in the previous section (convergence is quite fast, and quite evident at $n = 200$, due to the fact that the synthetic data are homoskedastic Gaussian). Note also that the bias of $\widehat{\text{prec}}$ cannot be reduced by increasing sample size, which is also in agreement with the theoretical analysis (and so are the results in the Supplementary Material).

To examine more closely the effect of the difference between prev and r on precision estimation, Figure 2 plots bias estimates for $\widehat{\text{prec}}$ and $\widehat{\text{prec}}^{\text{prev}}$ as a function of the absolute difference between prev and r, using the same data employed in Figure 1. It can be seen that the bias is always positive, indicating optimistic precision estimates. In nearly all cases, $\widehat{\text{prec}}^{\text{prev}}$

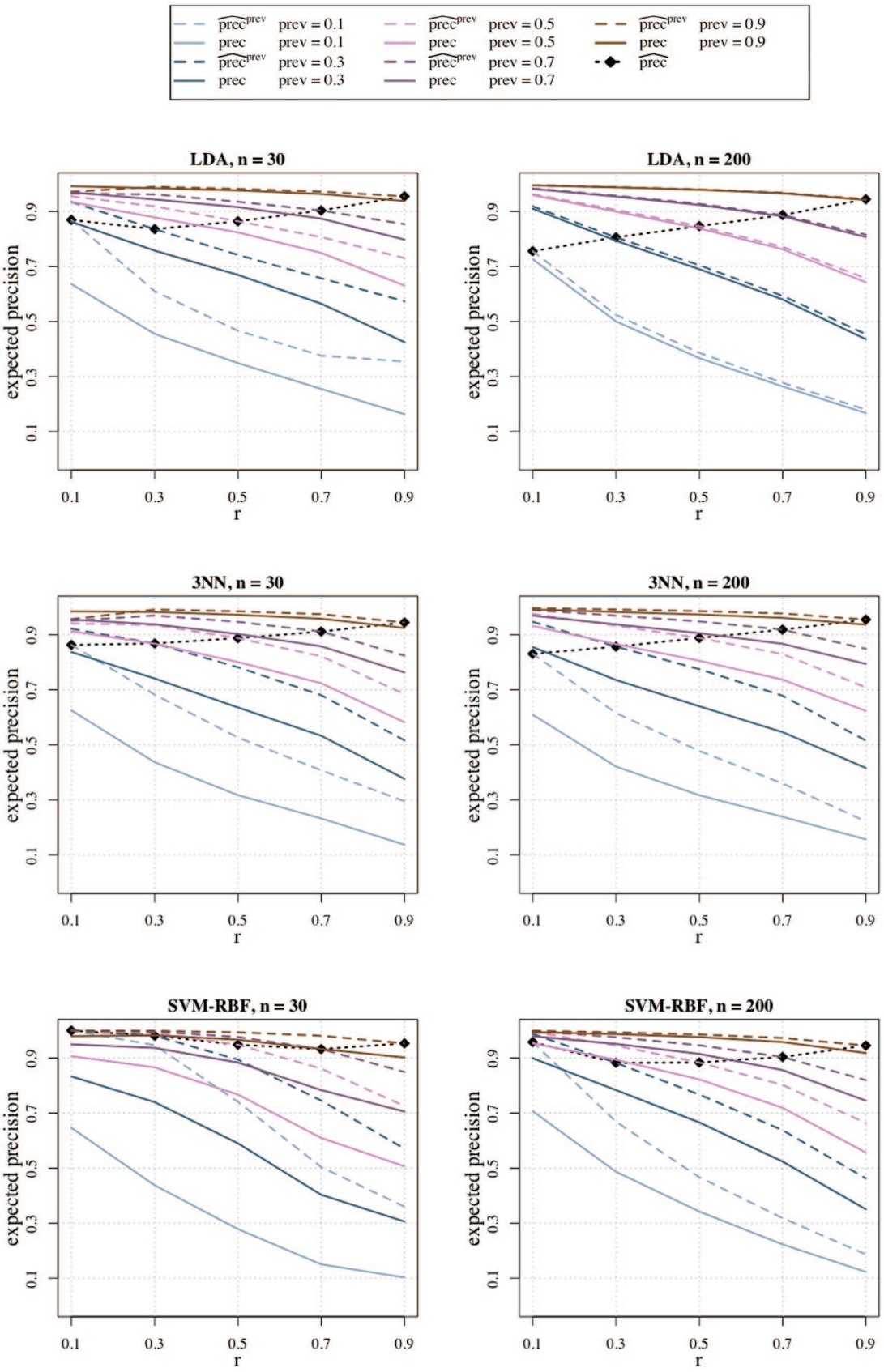

Figure 1. Average true precision (solid curves), average usual precision estimate $\widehat{\text{prec}}$ (dash-diamond curves), and average modified precision estimate $\widehat{\text{prec}}^{\text{prev}}$ (dashed curves), for LDA, 3NN, and RBF-SVM, with sample sizes $n = 30$ and $n = 200$, and different prevalence values, as a function of the sample size ratio. LDA indicates Linear Discriminant Analysis; 3NN, 3-Nearest Neighbors; RBF-SVM, Radial-Basis Function Support Vector Machine.

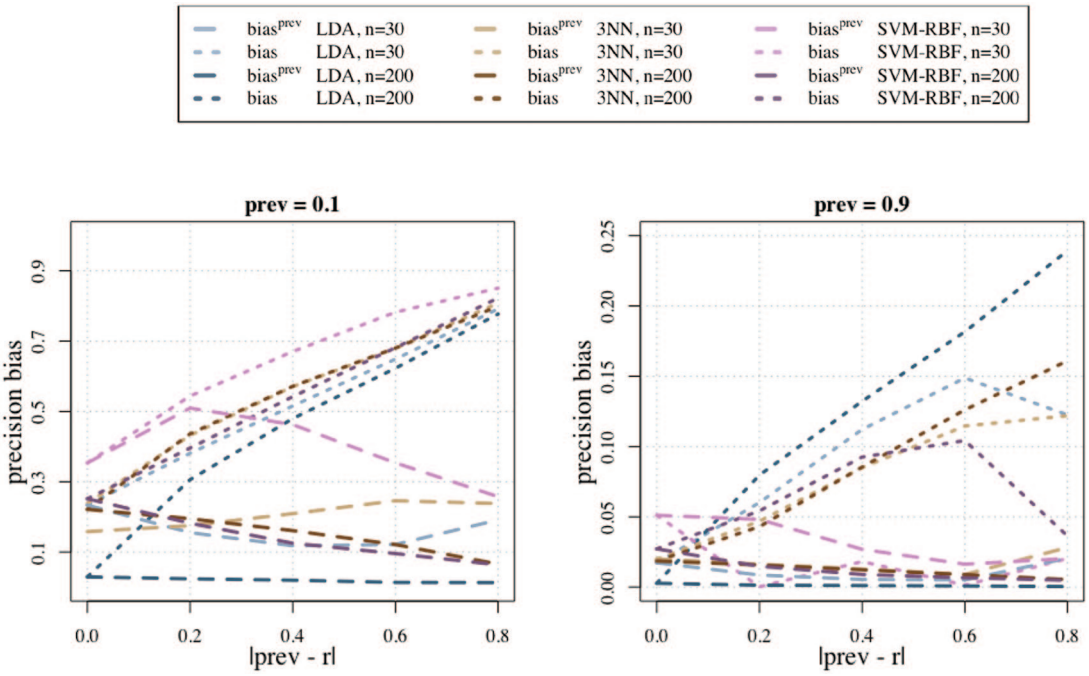

Figure 2. Estimated bias of the usual precision estimator $\widehat{\text{prec}}$ (dotted curves), and the modified precision estimator $\widehat{\text{prec}}^{\text{prev}}$ (dashed curves) for LDA, 3NN, and RBF-SVM, with sample sizes $n = 30$ and $n = 200$, and different prevalence values, as a function of the absolute difference between true prevalence and sample size ratio. LDA indicates Linear Discriminant Analysis; 3NN, 3-Nearest Neighbors; RBF-SVM, Radial-Basis Function Support Vector Machine.

has a smaller bias than $\widehat{\text{prec}}$, and when prev is far from r, the difference in bias becomes quite large.

Case studies with real data

Here we further investigate the bias of precision estimation under separate sampling using real data from three published studies.

Leukemia study. This publication[13] used a tumor microarray data set containing two types of human acute leukemia: acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). Gene expression measurements were taken from 15154 genes from 72 tissue specimens, 47 of which of ALL type (class 0), and 25 of AML type (class 1), so that $r = 0.347$. The estimator $\widehat{\text{prec}}^{\text{prev}}$ was computed using the value $\text{prev} = 0.222$, which is the incidence rate of ALL over AML in the US population.[14]

Breast cancer study. The second publication[15] employed the Wisconsin Breast Cancer (Original) Dataset from the University of California-Irvine (UCI) Machine Learning Repository,[16,17] which has been used by several groups to investigate breast cancer classification methods.[18,19] The data set consists of 699 instances, 458 and 241 of which are from benign and malignant tumors, respectively, and 10 features corresponding to cytological characteristics of breast fine-needle aspirates. According to Wilkins,[20] fewer than 20% of breast lumps are malignant; therefore, we used used $\text{prev} = 0.2$ in the computation of the modified precision estimator $\widehat{\text{prec}}^{\text{prev}}$.

Liver disease study. The final publication[21] employed a liver disease data set, also from the UCI Machine Learning Repository. This data set contains 5 blood test attributes and 345 records, of which 145 belong to individuals with liver disease (class 0) and 200 measurements are taken from healthy individuals (class 1), so that $r = 0.42$. This data set was donated to UCI in 1990, when the prevalence rate for chronic liver disease in the United States was $\text{prev} = 0.1178$,[22] which we use as the prevalence in the computation of the $\widehat{\text{prec}}^{\text{prev}}$ estimator.

All three studies used libraries from the Weka machine learning environment[23] to compute usual precision estimates on separately sampled data, while ignoring true prevalences, for different classification rules: Naive Bayes (NB),[24] C4.5 decision tree,[25] Back-Propagated Neural Networks, 3NN, and Linear SVM.[12] We reproduced the analysis in all three papers using Weka, obtaining almost exactly the same $\widehat{\text{prec}}$ estimates reported in those papers, and added for comparison the $\widehat{\text{prec}}^{\text{prev}}$ using the prevalence values described above. The results, displayed in Figure 3, show that without exception, the usual precision estimates $\widehat{\text{prec}}$ are larger than the more accurate $\widehat{\text{prec}}^{\text{prev}}$ estimates, in agreement with the previously observed fact that $\widehat{\text{prec}}$ displays a larger (optimistic) bias. The bias is particularly large in the case of the liver disease study, reflecting the fact that among the three data sets, this is the one where the value of prev and r differ the most.

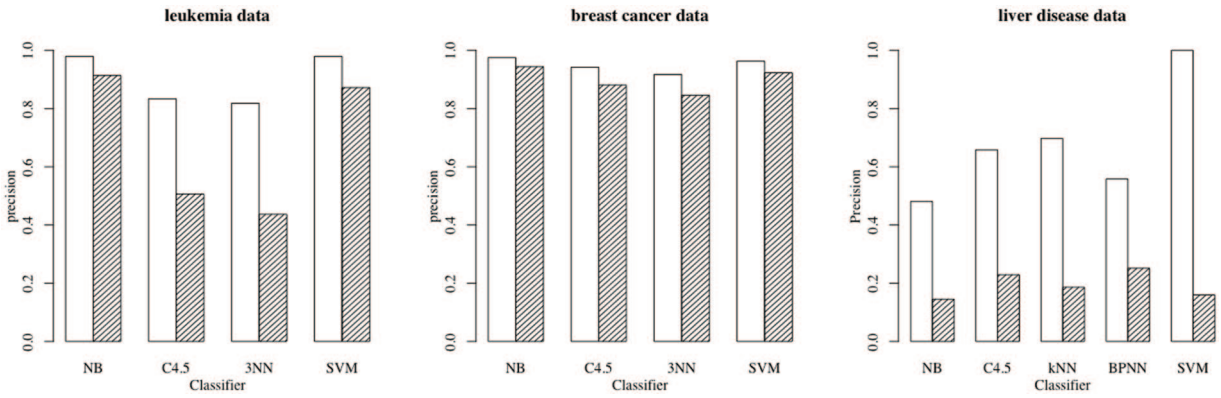

Figure 3. Precision estimates for different classification rules using separately sampled leukemia, breast cancer, and liver disease data. The white bars depict the usual estimated precision estimates, while the shaded bars are for the precision estimates using the true case prevalences. NB indicates Naive Bayes; 3NN, 3-Nearest Neighbors; SVM, Support Vector Machine.

Concluding Remarks

Accuracy and reproducibility in observational studies is critical to the progress of biomedicine, in particular, in the discovery of reliable biomarkers for disease diagnosis and prognosis. In this study, theoretical results confirmed by numerical experiments show that the usual estimator of precision can be severely biased under the typical separate sampling scenario in observational case-control studies. This will be true especially if the true disease prevalence differs significantly from the apparent prevalence in the data. If knowledge of the true disease prevalence is available, or can even be approximately ascertained, then it can be used to define a modified precision estimator, which is nearly unbiased at moderate sample sizes. In all the results using real data sets, we observed that the usual precision estimator produces values that are larger, ie, more optimistic, than the modified one using the true prevalence, which agrees with the results obtained with the synthetic data. Absence of knowledge about the true prevalence means simply that the precision cannot be reliably estimated in observational case-control studies and its use should be discouraged. Finally, we note that in our experiments, we considered the case where the prevalence is between 0.1 and 0.9, not without reason. If the prevalence is significantly under 0.1, as is the case in some rare diseases, then neither the precision, nor in fact the classification error, should be used as a criterion of performance, but rather the sensitivity and specificity need to be considered separately—otherwise, a large precision and small classification error can be achieved by biasing the classification rule to produce FP rates close to zero while ignoring the FN rate.

Author Contributions

UMB-N proposed the original idea of studying precision estimates under separate sampling. SX conducted a detailed bibliographical research on the use of precision in Bioinformatics. SX designed and conducted the numerical experiments using the synthetic and real data sets. Both authors contributed in the discussion of the results. SX prepared the initial draft of the manuscript, and UMB-N contributed in the preparation of the final version.

ORCID iD

Shuilian Xie https://orcid.org/0000-0002-6991-0315

Supplemental Material

Supplemental material for this article is available online.

REFERENCES

1. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*. 1995;270:467–470.
2. Lockhart DJ, Dong H, Byrne MC, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol*. 1996;14:1675.
3. Mortazavi A, Williams B, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat Methods*. 2008;5:621–628.
4. Aebersold R, Mann M. Mass spectrometry-based proteomics. *Nature*. 2003;422:198–207.
5. Braga-Neto U, Dougherty E. Is cross-validation valid for microarray classification? *Bioinformatics*. 2004;20:374–380.
6. Braga-Neto U, Dougherty E. *Error Estimation for Pattern Recognition*. New York, NY: John Wiley & Sons; 2015.
7. Ong MS, Magrabi F, Coiera E. Automated categorisation of clinical incident reports using statistical text classification. *Qual Saf Health Care*. 2010;19:e55.
8. Dang HX, Lawrence CB. Allerdictor: fast allergen prediction using text classification techniques. *Bioinformatics*. 2014;30:1120–1128.
9. Hassanpour S, Langlotz CP, Amrhein TJ, et al. Performance of a machine learning classifier of knee MRI reports in two large academic radiology practices: a tool to estimate diagnostic yield. *Am J Roentgenol*. 2017;208:750–753.
10. Braga-Neto U, Zollanvari A, Dougherty ER. Cross-validation under separate sampling: strong bias and how to correct it. *Bioinformatics*. 2014;30:3349–3355. doi:10.1093/bioinformatics/btu527.
11. Devroye L, Gyorfi L, Lugosi G. *A Probabilistic Theory of Pattern Recognition*. New York, NY: Springer; 1996.
12. Duda RO, Hart PE, Stork DG, et al. *Pattern Classification*. 2nd ed. New York, NY: Springer; 2001:55.
13. Hewett R, Kijsanayothin P. Tumor classification ranking from microarray data. *BMC Genomics*. 2008;9:S21.
14. Howlader N, Noone A, Krapcho M, et al. SEER cancer statistics review 1975-2013. *SEER*. http://seer.cancer.gov/csr/1975_2013/. Updated 2016.
15. Asri H, Mousannif H, Al Moatassime H, et al. Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Proc Comput Sci*. 2016;83:1064–1069.
16. Dua D, Graff C. UCI machine learning repository. *UCI*. http://archive.ics.uci.edu/ml. Updated 2017.
17. Wolberg WH, Mangasarian OL. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proc Natl Acad Sci U S A*. 1990;87:9193–9196. doi:10.1073/pnas.87.23.9193.
18. Shajahaan SS, Shanthi S, ManoChitra V. Application of data mining techniques to model breast cancer data. *Int J Emerg Technol Adv Eng*. 2013;3:362–369.
19. Akay MF. Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Syst Appl*. 2009;36:3240–3247.
20. Wilkins L. *Interpreting Signs and Symptoms* (LWW Medical Book Collection). Philadelphia, PA: Lippincott Williams & Wilkins; 2007.

21. Ramana BV, Prasad MS, Venkateswarlu NB. A critical study of selected classification algorithms for liver disease diagnosis. *Int J Database Manag Syst*. 2011;3: 101–114.
22. Younossi Z, Stepanova M, Afendy M, et al. Changes in the prevalence of the most common causes of chronic liver diseases in the united states from 1988 to 2008. *Clin Gastroenterol Hepatol*. 2011;9:524–530.e1; quiz e60.
23. Holmes G, Donkin A, Witten I. *Weka: A Machine Learning Workbench* (Working paper 94/9). Hamilton, New Zealand: Department of Computer Science, University of Waikato; 1994.
24. Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. *Mach Learn*. 1997;29:131–163.
25. Dietterich T. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Mach Learn*. 2000;40:139–157.

Appendix 1

Here we derive the asymptotic approximation in equation (29). If $f:\mathbb{R}^2 \to \mathbb{R}$ is infinitely differentiable at point (a,b), then it can be expanded by a bivariate Taylor series around (a,b) as

$$f(x,y)=f(a,b)+\frac{\partial f(a,b)}{\partial x}(x-a)+\frac{\partial f(a,b)}{\partial y}(y-b) \\ +\text{second and higher order terms in } x-a \text{ and } y-b \tag{37}$$

Now let X_n and Y_n be sequences of random variables with means μ_X and μ_Y, with $\mu_Y \neq 0$. The ratio x/y is infinitely differentiable at (a,b) if $b \neq 0$; therefore, we can apply the previous result and get

$$\frac{X_n}{Y_n}=\frac{\mu_X}{\mu_Y}+\frac{1}{\mu_Y}(X_n-\mu_X)-\frac{\mu_X}{\mu_Y^2}(Y_n-\mu_Y) \\ +\text{second and higher order terms in } X_n-\mu_X \text{ and } Y_n-\mu_Y \tag{38}$$

Taking expectations on both sides gives

$$E\left[\frac{X_n}{Y_n}\right]=\frac{\mu_X}{\mu_Y}+E\left[\begin{array}{l}\text{second and higher}\\ \text{order terms in } X_n-\mu_X \text{ and } Y_n-\mu_Y\end{array}\right] \tag{39}$$

Except in pathological cases involving heavy-tailed distributions, the remainder in the previous equation becomes negligible as $X_n \to \mu_X$ and $Y_n \to \mu_Y$ in probability. Therefore, we write

$$E\left[\frac{X}{Y}\right] \approx \frac{E[X]}{E[Y]} \tag{40}$$

as long as X and Y are around $E[X]$ and $E[Y]$, respectively (ie, $\mathrm{Var}[X]$ and $\mathrm{Var}[Y]$ are small).