Discussion paper

On the accuracy of estimators based on a binary classifier

Sander Scholtus
Arnout van Delden

February 2020

Content

Summary

Publications in official statistics often involve estimates by domain. The accuracy of these statistics is determined in part by the accuracy with which units are classified in their correct domains. With the increased use of administrative and other non-survey-based data sources in official statistics, as well as the development of automatic classifiers based on machine learning, it is important to account for the effect of classification errors on statistical accuracy.

Although bias and variance formulas for estimated domain totals and growth rates under classification errors have been derived before in the literature, these expressions are relatively complicated and therefore do not provide much insight into the 'typical' effect of classification errors on domain statistics. The aim of the present paper is to provide some guidance, in the form of simple rules-of-thumb, on the behavior of domain statistics that may be expected in practice when classification errors occur. To keep matters as simple as possible, we restrict attention to the common case of counts with a binary classifier. We examine the accuracy of estimated proportions, differences of estimated proportions between two periods, and growth rates of estimated counts between two periods. The results are illustrated using a real text mining application on the prevalence of cybercrime in the Netherlands.

Keywords

classification errors, domain statistics, bias, variance, machine learning classifier

 The authors would like to thank Quinten Meertens for his insightful comments on the first draft of this paper, which have led to several improvements.

1. Introduction

Publications in official statistics often involve estimates by domain. Two well-known examples are: the total number of inhabitants of each municipality and the total quarterly turnover of each economic sector. The accuracy of these statistics is determined in part by the accuracy with which units are classified in their correct domains. For statistics based on traditional sample surveys, it is often assumed that the effect of classification errors is limited and therefore negligible in comparison to the sampling error. For statistics based on administrative data or big data, it is clear that this assumption cannot be made. With the increased use of administrative and other non-survey-based data sources in official statistics, it has therefore become more important to account for the effect of classification errors on statistical accuracy.

Recently, new applications in official statistics have started to emerge that involve the use of machine-learning algorithms for the automatic classification of units into domains of interest. Some recent examples include: the number of innovative businesses by region (Van der Doef et al., 2018), the total turnover of webshops in the European Union generated in the Netherlands (Meertens et al., 2020), and the number of police reports that register cybercrime (Tollenaar et al., 2019). Often, such applications make use of data sets that are very large and unstructured, so that manual classification is not an option other than for small subsamples that may be used to train or validate an automatic classifier. The results in this paper apply in particular, though not exclusively, to classifiers based on machine learning.

Basic expressions for the bias and variance of estimated domain totals under classification errors have been derived by, among others, Kuha and Skinner (1997), Van Delden et al. (2016), and Meertens et al. (2019a). The first reference focuses exclusively on counts, the other references treat the more general case of a domain total of a numerical variable (e.g., total turnover by economic sector). More often than not, official statistics focus also on the development of phenomena over time and therefore involve either a difference or a growth rate estimator between two periods. For estimated domain growth rates (of a numerical variable, with counts as a special case), approximate bias and variance formulas are given by Scholtus et al. (2019). Van Delden et al. (2016) and Scholtus et al. (2019) also proposed bootstrap methods that can be applied to estimate the bias and variance of more complicated statistics under classification errors.

Both the analytical expressions and the bootstrap methods require information on the probabilities with which particular classification errors occur. In applications that use machine learning, reasonable estimates for these probabilities are often obtained as a by-product when validating the algorithm. In other applications, where units are assigned to domains by human annotators or by decision rules, these probabilities may have to be estimated specifically for the purpose of evaluating the accuracy of target statistics (Van Delden et al., 2016).

A drawback of the general bias and variance expressions cited above is that they are relatively complicated — in particular for growth rates — and therefore do not provide much insight into the 'typical' effect of classification errors on domain statistics. The aim of the present paper is to provide some guidance, in the form of simple rules-of-thumb, on the behavior of domain statistics that may be expected in practice when classification errors occur. Quite often, the classification of interest for statistical publications is binary. For instance, in the above machine-learning examples, a business either is innovative or not, a business either has a webshop or not, and a police report describes criminal activities that are either cyber-related or not. To keep matters as simple as possible, in this paper we will therefore focus on the common case of counts with a binary classifier. We will look at the accuracy of estimated proportions, differences of estimated proportions between two periods, and growth rates of estimated counts between two periods.

The remainder of this paper is organized as follows. Section 2 introduces the problem in more detail and reviews previously derived results on the bias and variance under classification errors, applied specifically to a binary classifier. In Section 3, the behavior of these bias and variance expressions is studied under some simplifying assumptions that will hold approximately in many practical applications. To illustrate these results, in Section 4 they are applied to a real example of machine learning on the prevalence of cybercrime. The main results and discussion points are summarized in Section 5.

2. Setup and review of existing results

2.1 Setup and notation

We consider a situation where a classification algorithm (e.g., a machine-learning classifier or a human annotator) assigns cases to two possible classes. As a running example, to be examined in more detail in Section 4, we consider an application by Hooijschuur et al. (2019) on cybercrime-related aspects in police reports of crime victims in 2016. The data consisted of descriptions by victims of the crimes that occurred to them. The authors used a text mining model to derive a binary variable indicating whether a reported crime is cyber-related or not. The aim of this application was to estimate the proportion of cyber-related crimes in the Netherlands in 2016.

In general, we suppose that there is a classification variable with two classes, denoted by 1 and 0. Usually, the class 1 indicates that a case has some property of interest (e.g., a police report that describes criminal activities that are cyber-related) and the class 0 indicates the absence of this property.

We suppose that the classifier is not perfect. In the cybercrime example, the text mining model may classify some reports that in reality are cyber-related as being not cyber-related, and vice versa. We are interested in the effect of these errors on the accuracy of the estimated proportion of cyber-related crimes, in terms of bias (systematic deviation) and variance (noise).

To describe these effects, we need to know how often particular classification errors are expected to occur. In this paper, we will only consider the simplest scenario, where the classifier is assumed to make random errors in the same way for all cases. Let p_{11} denote the probability that a case that in reality belongs to class 1 is assigned to the correct class, and let p_{00} denote the probability that a case that in reality belongs to class 0 is assigned to the correct class. From the point of view of class 1, the expected share of false negatives amongst the true target cases is $1 - p_{11}$ and the expected share of false positives among the true non-target cases is $1 - p_{00}$. These probabilities can be displayed conveniently in a 2×2 matrix, as follows:

$$\mathbf{P} = \begin{pmatrix} p_{11} & 1 - p_{11} \\ 1 - p_{00} & p_{00} \end{pmatrix}. \qquad (1)$$

In this matrix, the rows correspond to the (generally unknown) true classification and the columns correspond to the observed classification. Note that p_{11} is equal to the *recall* (or *sensitivity*) and p_{00} is equal to the *specificity* of the classifier.

In practice, the probabilities in (1) can be estimated if the true classification is known for a random sample of units, such as an annotated test set. For instance,

in the cybercrime application a random sample was available of 168 reports that were known to be cyber-related, of which the text mining algorithm classified 133 as cyber-related and 35 as not cyber-related. From this, the probability p_{11} was estimated to be $133/168 \approx 0.79$. We will discuss the estimated matrix $\mathbf{P}$ for this example in more detail in Section 4 (cf. Table 1).

In general, we suppose that the same classifier is used for two different time periods (e.g., two years), denoted by q and r. Let N^q and N^r denote the total number of cases to be classified in each period; in what follows, these numbers are considered to be known and fixed. Also, for simplicity we assume throughout this paper that the classifier is applied to the entire target population; that is to say, we do not consider sampling issues here.

Denote the true number of cases that belong to class $j \in \{0,1\}$ in both periods as N_j^q and N_j^r. The associated true proportions are denoted as $\alpha_j^q = N_j^q/N^q$ and $\alpha_j^r = N_j^r/N^r$. Based on the observed classification, the numbers of cases in class $j \in \{0,1\}$ are estimated to be $\widehat{N}_j^q$ and $\widehat{N}_j^r$. It is assumed that $\widehat{N}_1^q + \widehat{N}_0^q = N^q$ and $\widehat{N}_1^r + \widehat{N}_0^r = N^r$. It is then natural to estimate the associated proportions by $\hat{\alpha}_j^q = \widehat{N}_j^q/N^q$ and $\hat{\alpha}_j^r = \widehat{N}_j^r/N^r$. As a fictional example, suppose that in 2016 there are $N^{2016} = 1\,000\,000$ police reports on crimes, of which $\widehat{N}_1^{2016} = 100\,000$ are classified as cyber-related by the text mining model. In this example, the estimated proportion of interest is $\hat{\alpha}_1^{2016} = 10\%$; it follows that $\hat{\alpha}_0^{2016} = 90\%$.

We suppose that the classification error probabilities for both periods are described by the same matrix $\mathbf{P}$. Note that this assumption may not hold if the classification model was updated between the two periods, or if the population distribution of features used by the classifier has changed substantially in the interim.

We are interested in the accuracy of the estimated proportion of cases in class 1 in both periods ($\hat{\alpha}_1^q$ and $\hat{\alpha}_1^r$) and also in the development of this proportion between the two periods. This development may be expressed as a difference $D_1^{q,r} = \alpha_1^q - \alpha_1^r$ and estimated by $\widehat{D}_1^{q,r} = \hat{\alpha}_1^q - \hat{\alpha}_1^r$. For instance, suppose that in the above fictional example for 2017 there are again $N^{2017} = 1\,000\,000$ police reports on crimes, but now $\widehat{N}_1^{2017} = 150\,000$ are classified as cyber-related. It follows that $\hat{\alpha}_1^{2017} = 15\%$ and $\widehat{D}_1^{2017,2016} = 5\%$: the estimated prevalence of cybercrime has increased by five percentage points.

Alternatively, the development of a phenomenon over time may be expressed as the growth rate of the number of cases in the population, $G_1^{q,r} = N_1^q/N_1^r$, and estimated by $\hat{G}_1^{q,r} = \widehat{N}_1^q/\widehat{N}_1^r$. In the above fictional example, we find $\hat{G}_1^{2017,2016} = 1.5$. The growth rate $G_1^{q,r}$ might be mainly useful if it holds approximately that $N^q = N^r$. Otherwise, the difference in proportions $D_1^{q,r}$ might be considered more informative. Note that when it holds that $N^q = N^r$ exactly, then $\hat{G}_1^{q,r} = \widehat{N}_1^q/\widehat{N}_1^r = \hat{\alpha}_1^q/\hat{\alpha}_1^r$, which means that the growth rate of the counts of the phenomenon then equals the growth rate of the corresponding proportions. In practice, the development is also sometimes expressed as a *relative* growth rate, which equals $\left(N_1^q - N_1^r\right)/N_1^r = G_1^{q,r} - 1$. In the fictional example, this relative

growth rate would be 50%. Because the expressions for relative growth rates only differ from the growth rates by a constant, they are left out of the remainder of this paper.

2.2 Existing results

In what follows, we will assume that classification errors are the only errors that affect the statistics of interest. We will also assume that classification errors are independent across cases. Finally, we will assume that the errors that affect $\widehat{N}_j^q$ are independent of those that affect $\widehat{N}_j^r$. Under the previous two assumptions, this latter assumption holds in particular if the sets of cases classified by the algorithm are disjoint for periods q and r. Given these three assumptions, the accuracy of the estimators defined in Section 2.1 is determined completely by the probabilities in the matrix $\mathbf{P}$ and the true counts N_j^q and N_j^r. Specific formulae will now be reviewed that are known or easily derived from the existing literature.

2.2.1 Results for estimated counts and proportions

In Burger et al. (2015), the following expressions are derived for the bias and variance of the estimated count $\widehat{N}_1^q$:

$$B\left(\widehat{N}_1^q\right) = (p_{11} - 1)N_1^q + (1 - p_{00})N_0^q;$$
$$V\left(\widehat{N}_1^q\right) = p_{11}(1 - p_{11})N_1^q + p_{00}(1 - p_{00})N_0^q.$$

Since $B\left(\hat{\alpha}_1^q\right) = B\left(\widehat{N}_1^q\right)/N^q$ and $V\left(\hat{\alpha}_1^q\right) = V\left(\widehat{N}_1^q\right)/(N^q)^2$, it follows that the bias and variance of the estimated proportion $\hat{\alpha}_1^q$ are given by:

$$B\left(\hat{\alpha}_1^q\right) = (p_{11} - 1)\alpha_1^q + (1 - p_{00})\alpha_0^q, \tag{2}$$

and

$$V\left(\hat{\alpha}_1^q\right) = \frac{p_{11}(1 - p_{11})\alpha_1^q + p_{00}(1 - p_{00})\alpha_0^q}{N^q}. \tag{3}$$

Similar expressions follow for $B(\hat{\alpha}_1^r)$ and $V(\hat{\alpha}_1^r)$. Formula (2) is equivalent to formula (28.5) in Kuha and Skinner (1997). Formula (3) is also derived in Meertens et al. (2019a).

For future reference, we note that the accuracy of an estimated count, say $\widehat{N}_1^q$, can be summarised in terms of its relative bias $RB\left(\widehat{N}_1^q\right) = B\left(\widehat{N}_1^q\right)/N_1^q$ and its coefficient of variation $CV\left(\widehat{N}_1^q\right) = \sqrt{V\left(\widehat{N}_1^q\right)}/E(\widehat{N}_1^q)$. (Note that the coefficient of variation of an estimator is evaluated with respect to its mean, whereas the relative bias is evaluated with respect to the true value.) For the associated estimated proportion $\hat{\alpha}_1^q$, we find:

$$RB\left(\hat{\alpha}_1^q\right) = \frac{B\left(\hat{\alpha}_1^q\right)}{\alpha_1^q} = \frac{B\left(\widehat{N}_1^q\right)/N^q}{N_1^q/N^q} = \frac{B\left(\widehat{N}_1^q\right)}{N_1^q} = RB\left(\widehat{N}_1^q\right)$$

and

$$CV(\hat{\alpha}_1^q) = \frac{\sqrt{V(\hat{\alpha}_1^q)}}{E(\hat{\alpha}_1^q)} = \frac{\sqrt{V(\hat{N}_1^q)/(N^q)^2}}{E(\hat{N}_1^q)/N^q} = \frac{\sqrt{V(\hat{N}_1^q)}}{E(\hat{N}_1^q)} = CV(\hat{N}_1^q).$$

So it does not matter whether the relative bias and coefficient of variation are computed for estimated counts or for the associated estimated proportions.

2.2.2 Results for estimated differences

For the bias of the estimated difference $\hat{D}_1^{q,r}$, it may be noted first of all that

$$\begin{aligned} B(\hat{D}_1^{q,r}) &= E(\hat{D}_1^{q,r}) - D_1^{q,r} \\ &= E(\hat{\alpha}_1^q) - E(\hat{\alpha}_1^r) - (\alpha_1^q - \alpha_1^r) \\ &= B(\hat{\alpha}_1^q) - B(\hat{\alpha}_1^r). \end{aligned}$$

Hence, it follows directly from (2) that

$$B(\hat{D}_1^{q,r}) = (p_{11} - 1)(\alpha_1^q - \alpha_1^r) + (1 - p_{00})(\alpha_0^q - \alpha_0^r). \tag{4}$$

For the associated variance, the assumption that errors are independent between periods implies that $V(\hat{D}_1^{q,r}) = V(\hat{\alpha}_1^q) + V(\hat{\alpha}_1^r)$, so that

$$\begin{aligned} V(\hat{D}_1^{q,r}) = &\frac{p_{11}(1-p_{11})\alpha_1^q + p_{00}(1-p_{00})\alpha_0^q}{N^q} \\ &+ \frac{p_{11}(1-p_{11})\alpha_1^r + p_{00}(1-p_{00})\alpha_0^r}{N^r}. \end{aligned} \tag{5}$$

2.2.3 Results for estimated growth rates

Finally, again under the assumption of independent classification errors between periods, the following formulae based on Scholtus et al. (2019) may be used for the approximate bias (AB) and approximate variance (AV) of the estimated growth rate $\hat{G}_1^{q,r}$:

$$AB(\hat{G}_1^{q,r}) = (\breve{G}_1^{q,r} - G_1^{q,r}) + \breve{G}_1^{q,r} \frac{p_{11}(1-p_{11})N_1^r + p_{00}(1-p_{00})N_0^r}{[p_{11}N_1^r + (1-p_{00})N_0^r]^2}, \tag{6}$$

and

$$\begin{aligned} AV(\hat{G}_1^{q,r}) = &\frac{p_{11}(1-p_{11})\left[N_1^q + (\breve{G}_1^{q,r})^2 N_1^r\right]}{[p_{11}N_1^r + (1-p_{00})N_0^r]^2} \\ &+ \frac{p_{00}(1-p_{00})\left[N_0^q + (\breve{G}_1^{q,r})^2 N_0^r\right]}{[p_{11}N_1^r + (1-p_{00})N_0^r]^2}. \end{aligned} \tag{7}$$

Here, the symbol $\breve{G}_1^{q,r}$ is used as shorthand for

$$\breve{G}_1^{q,r} = \frac{E\left(\widehat{N}_1^q\right)}{E\left(\widehat{N}_1^r\right)} = \frac{p_{11}N_1^q + (1-p_{00})N_0^q}{p_{11}N_1^r + (1-p_{00})N_0^r}. \quad (8)$$

The approximate nature of these expressions derives from the fact that they are based on Taylor series approximations: formula (6) is based on a second-order Taylor approximation and formula (7) is based on a first-order Taylor approximation to the estimated growth rate $\widehat{G}_1^{q,r}$.

In Scholtus et al. (2019) more general expressions can be found for the approximate bias and variance of an estimated growth rate when the classification errors in both periods are not independent and/or when the growth rate concerns the total of a numerical variable rather than a simple count. Moreover, all papers that have been referred to in this section also provide results for classifications with more than two classes. Here, we will not consider these more complicated situations.

Finally, it is important to note that expressions (2)–(8) contain elements that are unknown in practice: the probabilities p_{11} and p_{00} and true counts (or true proportions) such as N_1^q (or α_1^q). Estimating the latter quantities in practice is not entirely straightforward. We will ignore this issue for now and return to it in the final section (Section 5).

3. Typical properties of estimators under classification errors

In principle, the formulas given in Section 2.2 can be used in practice to evaluate the accuracy (bias and standard error) of estimated proportions, differences of proportions, and growth rates, provided that the matrix $\mathbf{P}$ is known or previously estimated. However, in particular for differences and growth rates, these formulas are relatively complicated.

In the present section, we aim to provide some more insight into the typically expected behavior of these estimators, in terms of their standard error and bias. We will derive simple rules-of-thumb by making some simplifying assumptions that should hold approximately in many practical applications. We will start with estimated proportions $\hat{\alpha}_1^q$ (Section 3.1) and then move on to estimated differences $\hat{D}_1^{q,r}$ (Section 3.2) and estimated growth rates (Section 3.3).

3.1 Accuracy of estimated proportions

3.1.1 Standard error

It follows from expression (3) that the standard error of the estimated proportion $\hat{\alpha}_1^q$ is given by

$$\begin{aligned} SE(\hat{\alpha}_1^q) &= \sqrt{V(\hat{\alpha}_1^q)} \\ &= \frac{1}{\sqrt{N^q}} \sqrt{p_{11}(1-p_{11})\alpha_1^q + p_{00}(1-p_{00})(1-\alpha_1^q)}. \end{aligned} \tag{9}$$

Here, we used that $\alpha_1^q + \alpha_0^q = 1$.

Formula (9) has some interesting features:

- The standard error decreases in proportion to the square root of the total number of classified cases (N^q). Thus, all else being equal, more precise estimators of the true proportion α_1^q are obtained for larger populations. This makes sense, since we have made the assumption that classification errors are independent between cases. Therefore, the uncertainty caused by these errors should tend to 'average out' as the population becomes larger. (Which is not to say that classification errors do not have any effect on estimators for large populations: there could still be bias.)
- For a hypothetical classifier that does not use any background information and simply assigns half of the cases 'completely at random' to the first class, it would hold that $p_{11} = 1/2$ and $p_{00} = 1/2$. Consider classifiers with $p_{11} > 1/2$ and $p_{00} > 1/2$, i.e., classifiers that work better, for both classes, than

classifying the cases 'completely at random'. Since $f(x) = x(1-x)$ is a monotone decreasing function for $1/2 \leq x \leq 1$, it then follows that the standard error of $\hat{\alpha}_1^q$ decreases as p_{11} and p_{00} become larger. Thus, all else being equal, more precise estimators of the true proportion α_1^q are obtained when false negatives and/or false positives become more rare. Again, intuitively this makes sense.[1]

- For $p_{11} = p_{00} = 1$, the standard error is zero, as no classification errors occur. Theoretically, the standard error also vanishes when $p_{11} = p_{00} = 0$, which corresponds to a scenario where all units are misclassified with certainty.
- For a given true proportion α_1^q and population size N^q, the above formula for $SE(\hat{\alpha}_1^q)$ can be solved to find all combinations (p_{11}, p_{00}) for which a desired level of precision $[SE(\hat{\alpha}_1^q) \leq \gamma]$ is achieved:

$$\mathcal{A}(a_1^q, N^q, \gamma) = \{(p_{11}, p_{00}) | p_{11}(1-p_{11})\alpha_1^q + p_{00}(1-p_{00})(1-\alpha_1^q) \leq N^q\gamma^2\}.$$

For given values of α_1^q, N^q and γ, it can be shown that the points (p_{11}, p_{00}) for which the desired accuracy is achieved exactly $[SE(\hat{\alpha}_1^q) = \gamma]$ lie on a curve that in practice will take the form of an ellipse.[2] In the special case that $\alpha_1^q = 1/2$, this ellipse becomes a circle.

To illustrate the last point, Figure 1 shows various ellipses of points (p_{11}, p_{00}) where a certain precision is achieved for $\hat{\alpha}_1^q$ in a particular example. In this example, the population size $N^q = 2000$ and the true proportion of class 1 in the population is $\alpha_1^q = 30\%$. The red curve indicates combinations of p_{11} (horizontal axis) and p_{00} (vertical axis) such that $SE(\hat{\alpha}_1^q)$ is exactly equal to 1.1 percentage points. Any point that lies above this curve will achieve a better precision, so the inequality $SE(\hat{\alpha}_1^q) \leq 1.1\%$ holds for any classifier with (p_{11}, p_{00}) lying above or on the red curve. Similarly, the other curves indicate combinations (p_{11}, p_{00}) such that $SE(\hat{\alpha}_1^q)$ is exactly equal to, respectively, 1.0, 0.9, 0.8, 0.7, and 0.6 percentage points.

[1] From a purely mathematical point of view, for classifiers with $p_{11} < 1/2$ and $p_{00} < 1/2$ it follows analogously that the standard error of $\hat{\alpha}_1^q$ would decrease as p_{11} and p_{00} decrease towards zero. The practical relevance of this result is limited, as the performance of such a classifier (according to any reasonable evaluation criterion) could be improved by switching the two assigned classes.

[2] In two dimensions, a *conic section* consists of points (x, y) such that $ax^2 + 2hxy + by^2 + 2gx + 2fy + c = 0$, for certain constants a, b, c, f, g, and h. An *ellipse* is a special case which occurs when the so-called determinant $\Delta = abc + 2fgh - af^2 - bg^2 - ch^2 \neq 0$ and it holds that $h^2 - ab < 0$; see, e.g., Nelson (2003, pp. 76–78). The points (p_{11}, p_{00}) on the edge of $\mathcal{A}(a_1^q, N^q, \gamma)$ clearly form a conic section with $a = \alpha_1^q$, $b = (1-\alpha_1^q)$, $c = N^q\gamma^2$, $f = -(1-\alpha_1^q)/2$, $g = -\alpha_1^q/2$, and $h = 0$. In this case, the determinant simplifies to $\Delta = \alpha_1^q(1-\alpha_1^q)(N^q\gamma^2 - 1/4)$, which is nonzero unless $\alpha_1^q = 0$, $\alpha_1^q = 1$ or $\gamma^2 = 1/(4N^q)$. In practice, it will nearly always hold that $0 < \alpha_1^q < 1$. Also, it is natural to demand a precision such that $\gamma^2 < 1/(4N^q)$, since the bound $\gamma^2 = 1/(4N^q)$ is achieved by classifying the cases 'completely at random' ($p_{11} = p_{00} = 1/2$). Furthermore, when $0 < \alpha_1^q < 1$ it always holds that $h^2 - ab = -\alpha_1^q(1-\alpha_1^q) < 0$. We conclude that, in practice, the edge of $\mathcal{A}(a_1^q, N^q, \gamma)$ will take the form of an ellipse.

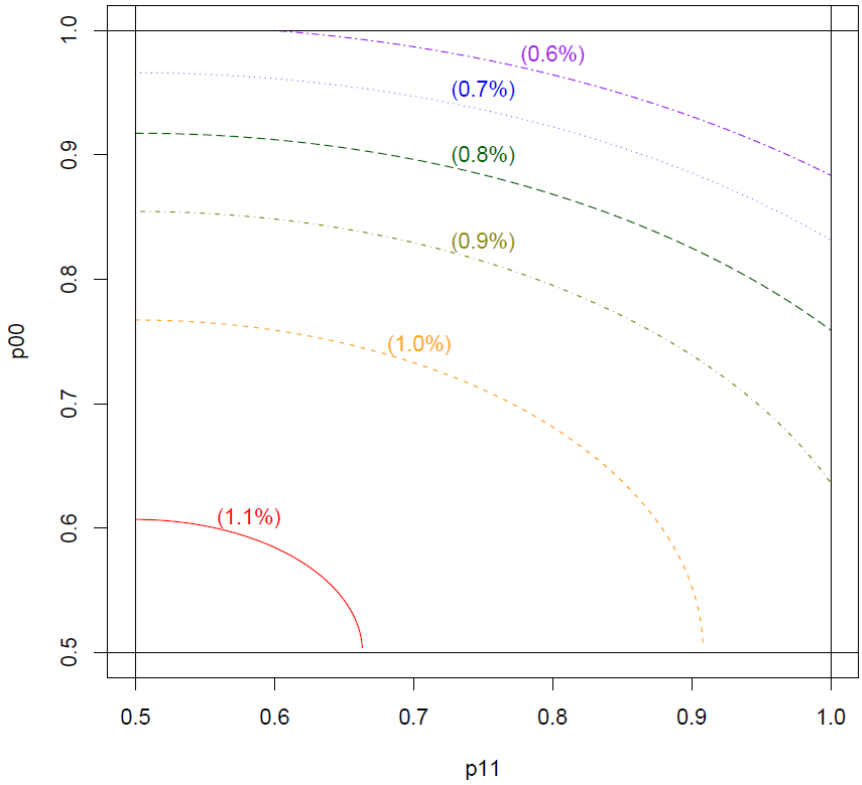

Figure 1. Isocurves (p_{11}, p_{00}) such that $SE(\hat{\alpha}_1^q) = \gamma$ for a classification problem with $N^q = 2000$ and $\alpha_1^q = 30\%$, with different values of γ.

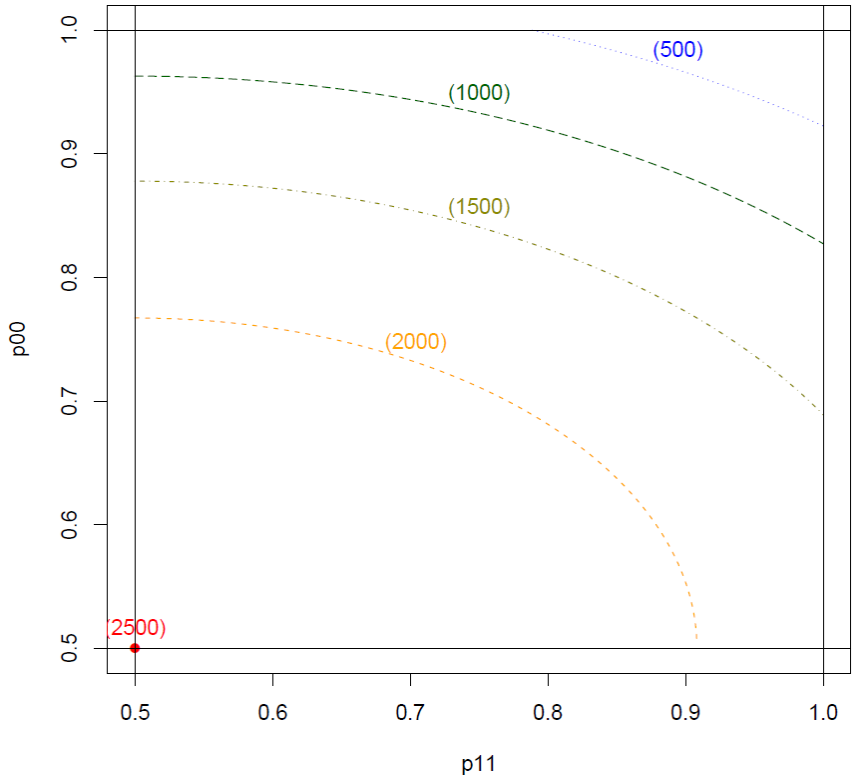

Figure 2. Isocurves (p_{11}, p_{00}) such that $SE(\hat{\alpha}_1^q) = 1.0\%$ for a classification problem with $\alpha_1^q = 30\%$, with different population sizes N^q.

Different classification models with different probabilities (p_{11}, p_{00}) that lie on the same curve will achieve the same standard error for $\hat{\alpha}_1^q$. For example, it is seen using the orange curve in Figure 1 that for this classification problem two different classifiers, one with probabilities $p_{11} = 0.90$ and $p_{00} = 0.55$, and the other with probabilities $p_{11} = 0.70$ and $p_{00} = 0.73$, both achieve a standard error of approximately 1.0 percentage points. It is also seen that, for fixed values of N^q and α_1^q, a relatively large increase in the probabilities p_{11} and/or p_{00} is needed to improve the standard error of the estimator by as little as 0.1 percentage point.

To illustrate the influence of the population size N^q on the precision of $\hat{\alpha}_1^q$, Figure 2 shows ellipses of points (p_{11}, p_{00}) where a standard error of 1.0 percentage points is achieved for the same problem as in Figure 1 but with N^q varying between 500 and 2500. Any point (p_{11}, p_{00}) that lies above a curve corresponds to a classifier with $SE(\hat{\alpha}_1^q) < 1.0\%$ for the population size N^q associated with that curve.

It is seen that the precision improves rapidly as the population size increases. For instance, when $p_{11} = 0.9$, to achieve a standard error of 1.0 percentage points when the population size $N^q = 1000$ (green curve), it is required that $p_{00} \geq 0.88$. But when the population size is $N^q = 2000$ (orange curve), the same precision is already achieved when $p_{00} \geq 0.55$. For $N^q = 2500$, a standard error of 1.0 percentage points can be achieved by classifying the cases 'completely at random', with $p_{11} = p_{00} = 0.50$. Hence, the red curve in Figure 2 is reduced to a single point.

3.1.2 Bias

For the bias of the estimated proportion $\hat{\alpha}_1^q$, it follows from (2) that

$$B(\hat{\alpha}_1^q) = (p_{11} - 1)\alpha_1^q + (1 - p_{00})(1 - \alpha_1^q). \tag{10}$$

Unlike the standard error, this bias does not become smaller for larger populations. Also, the bias does not necessarily decrease when false negatives and/or false positives become more rare. In fact, it is easy to derive from (10) that, for a given true proportion α_1^q, a particular bias $B(\hat{\alpha}_1^q) = \beta$ is achieved exactly when

$$p_{00} = \frac{\alpha_1^q}{1 - \alpha_1^q} p_{11} + \frac{1 - 2\alpha_1^q - \beta}{1 - \alpha_1^q}. \tag{11}$$

Note that the points (p_{11}, p_{00}) where this bias is achieved form a straight line. This is illustrated in Figure 3 for the same example as above, with $\alpha_1^q = 30\%$. In this example, expression (11) can be simplified to $p_{00} = (3/7)p_{11} + (40 - \beta)/70$ (with β expressed in percentage points).

It is seen in (10) that the bias in $\hat{\alpha}_1^q$ is zero if, and only if,

$$(1 - p_{11})\alpha_1^q = (1 - p_{00})(1 - \alpha_1^q). \tag{12}$$

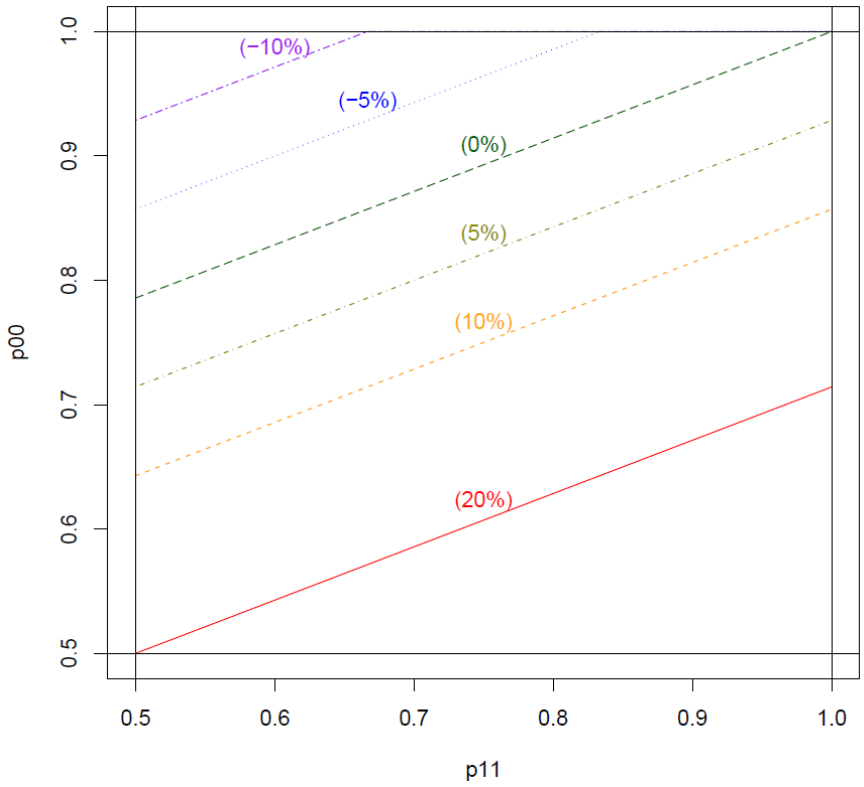

Figure 3. Isolines (p_{11}, p_{00}) such that $B(\hat{\alpha}_1^q) = \beta$ for a classification problem with $\alpha_1^q = 30\%$, with different values of β.

In this situation, the expected number of misclassified cases from class 1 is equal to the expected number of misclassified cases from class 0. As a special case of (11), it follows that the bias is zero if, and only if,

$$p_{00} = \frac{\alpha_1^q}{1-\alpha_1^q} p_{11} + \frac{1-2\alpha_1^q}{1-\alpha_1^q}. \tag{13}$$

Thus, the bias in $\hat{\alpha}_1^q$ is minimized when the probabilities of classification error for the two classes satisfy a certain 'ideal' ratio that depends on the true proportion α_1^q.[3] As can be seen in Figure 3, any changes to the model that result in a deviation from this ratio will increase the absolute bias, even when the quality of the model is improved in terms of false positives and/or false negatives. Van Delden et al. (2016) gave a real-life example where this phenomenon was encountered. On the other hand, it is always possible to simultaneously improve both (p_{11}, p_{00}) as well as the bias as long as $0 < \alpha_1^q < 1$. It follows from Figure 3 that starting from a point (p_{11}, p_{00}) with bias and moving in a straight line towards the point $(1, 1)$ always reduces the bias.

Of course, improving the rate of false positives and/or false negatives could still be considered beneficial in practice, as it improves the precision of the estimator. Conversely, a classifier with very low values of p_{11} and p_{00} that happen to satisfy

[3] Meertens et al. (2019a) observed that condition (12) is equivalent to having a binary classifier with precision exactly equal to recall.

condition (12) would still be considered unacceptable in practice, as the resulting estimator $\hat{\alpha}_1^q$ is too unstable, even when it is theoretically unbiased. The main point here is that some care should be taken when trying to improve the accuracy of a classifier in practice, because in some cases the gain in precision of the estimator $\hat{\alpha}_1^q$ may be offset by an increase in bias.

As was seen in (13), for any value of p_{11} it is possible to choose p_{00} in such a way that $B(\hat{\alpha}_1^q) = 0$. Figure 4 shows the standard errors that are achieved for these combinations of (p_{11}, p_{00}), for the same population sizes as in Figure 2. Again, it is seen that, when the classifier is unreliable, a relatively large increase in the probabilities (p_{11}, p_{00}) is needed to achieve a significant improvement in $SE(\hat{\alpha}_1^q)$.

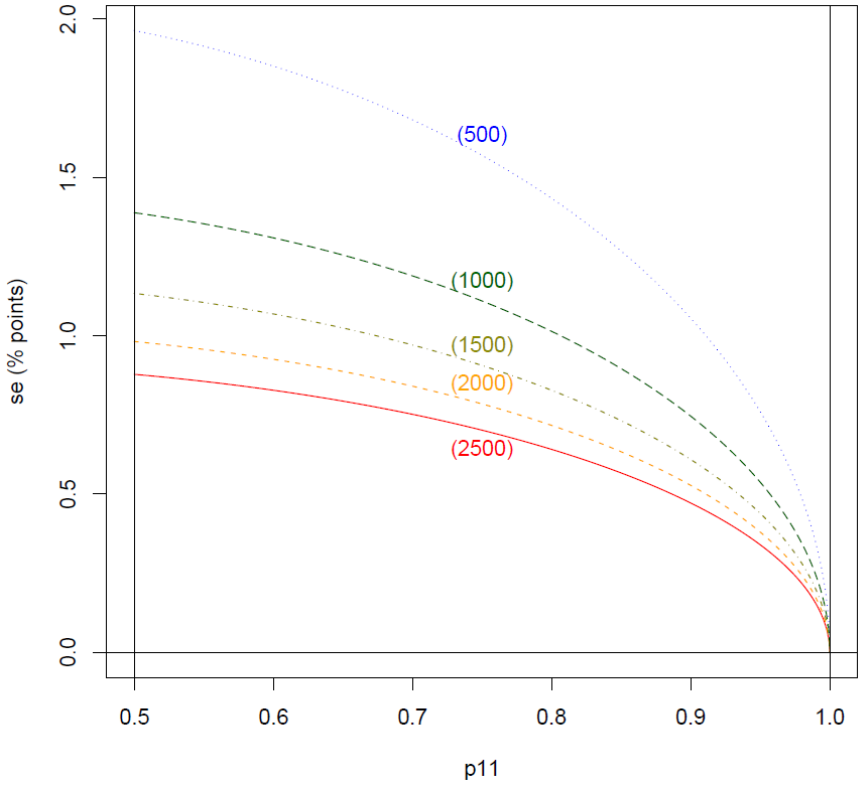

Figure 4. Curves of $SE(\hat{\alpha}_1^q)$ for a classification problem with $\alpha_1^q = 30\%$, with different population sizes N^q, for combinations (p_{11}, p_{00}) such that the bias is zero.

Finally, we observe that it follows from (9) and (10) together that

$$\frac{B(\hat{\alpha}_1^q)}{SE(\hat{\alpha}_1^q)} = \sqrt{N^q}\,\frac{(p_{11}-1)\alpha_1^q + (1-p_{00})(1-\alpha_1^q)}{\sqrt{p_{11}(1-p_{11})\alpha_1^q + p_{00}(1-p_{00})(1-\alpha_1^q)}} = O(\sqrt{N^q}).$$

That is to say, for large population sizes it is expected that the bias dominates the standard error of $\hat{\alpha}_1^q$. Theoretically, an exception would occur if condition (12) holds exactly.

3.1.3 Acceptable ranges of classification error probabilities

As a final result for estimated domain proportions, we remark that graphs of isocurves similar to the ones shown in the above figures may be used in practice to find acceptable ranges of classification error probabilities, given a desired accuracy. This is illustrated in Figure 5. For a particular application, a user may define the maximal absolute bias and the maximal standard error that are acceptable [$\left|B\left(\hat{\alpha}_1^q\right)\right| \leq \beta$ and $SE\left(\hat{\alpha}_1^q\right) \leq \gamma$]. For a given true proportion and population size, isocurves may then be drawn of points (p_{11}, p_{00}) that achieve these bounds exactly: an ellipse for the standard error γ and two lines for the bias $+\beta$ and $-\beta$.

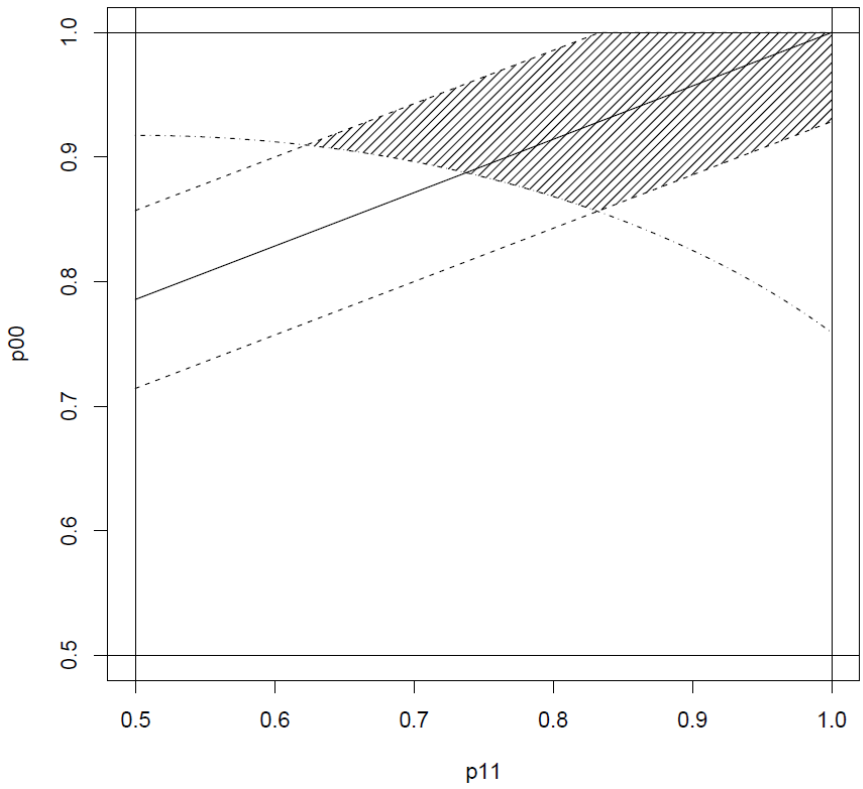

Figure 5. Illustration of the use of isocurves for the bias and standard error of a proportion. The shaded area shows an acceptable region of probabilities (p_{11}, p_{00}) such that $\left|B\left(\hat{\alpha}_1^q\right)\right| \leq \beta$ and $SE\left(\hat{\alpha}_1^q\right) \leq \gamma$. The solid line indicates points where $B\left(\hat{\alpha}_1^q\right) = 0$.

Together, these curves define an acceptable region of points (p_{11}, p_{00}) where both restrictions are satisfied; in Figure 5, this is the shaded region. When training the classification model, the user should then aim to achieve a combination (p_{11}, p_{00}) that lies within this acceptable region. In practice, since the true proportion is not known but estimated, this shaded area itself is known with some uncertainty.

3.2 Accuracy of estimated differences

3.2.1 Standard error

The standard error of the estimated difference of proportions $\widehat{D}_1^{q,r} = \hat{\alpha}_1^q - \hat{\alpha}_1^r$ is given by the square root of expression (5). If the phenomenon under study develops gradually between periods q and r, and the population size in the two periods does not differ by much, then it may be reasonable to assume that $\alpha_1^q \approx \alpha_1^r$ and $N^q \approx N^r$. In this situation, it follows that

$$SE\left(\widehat{D}_1^{q,r}\right) \approx \sqrt{2} \times SE\left(\hat{\alpha}_1^q\right) = \frac{1}{\sqrt{N^q}} \sqrt{2\{p_{11}(1-p_{11})\alpha_1^q + p_{00}(1-p_{00})\left(1-\alpha_1^q\right)\}}. \quad (14)$$

Thus, as a rule-of-thumb it may be expected that by taking the difference of two estimated proportions, the standard error due to classification errors is inflated approximately by a factor $\sqrt{2}$ compared to the estimated proportions themselves. This is due to the assumption made here that classification errors in different periods are independent.

3.2.2 Bias

For the bias of the estimated difference of proportions, it follows from (4) that

$$B\left(\widehat{D}_1^{q,r}\right) = (p_{11}-1)\left(\alpha_1^q - \alpha_1^r\right) + (1-p_{00})\left(\alpha_1^r - \alpha_1^q\right), \quad (15)$$

where we used again that $\alpha_1^q + \alpha_0^q = 1$ and $\alpha_1^r + \alpha_0^r = 1$. It is seen that this bias will vanish when either $\alpha_1^q = \alpha_1^r$ holds or it holds that $\alpha_1^q \neq \alpha_1^r$ and $(p_{11}-1) = (1-p_{00})$. However, this latter condition is only satisfied when $p_{00} = p_{11} = 1$, i.e., when no classification errors occur.

Although the bias of $\widehat{D}_1^{q,r}$ vanishes completely only under these exact conditions, in practice it is reasonable to expect this bias to be relatively small. Namely, it was noted above (4) that

$$B\left(\widehat{D}_1^{q,r}\right) = B\left(\hat{\alpha}_1^q\right) - B(\hat{\alpha}_1^r).$$

Again as a rule-of-thumb, for phenomena that develop gradually over time it may be reasonable to expect that $B\left(\hat{\alpha}_1^q\right) \approx B(\hat{\alpha}_1^r)$, so that $B\left(\widehat{D}_1^{q,r}\right) \approx 0$.

In summary, we conclude that in practice an estimated difference of proportions will typically be less biased than the estimated proportions themselves, but also less precise.

3.3 Accuracy of estimated growth rates

3.3.1 Standard error

Expression (7) for the approximate variance of an estimated growth rate $\hat{G}_1^{q,r}$ is relatively complicated. To get an impression of the typical magnitude of this variance, we now introduce the following additional simplifying assumptions:

1. The relative bias of the estimated number of units in class 1 is equal in both periods: $RB(\hat{N}_1^q) = RB(\hat{N}_1^r)$.
2. It holds that $\alpha_1^q = \alpha_1^r$.
3. The size of the population does not change between the two periods: $N^q = N^r$.

If the phenomenon under study is developing gradually over time, these assumptions may often be reasonable in practice, at least as a first-order approximation.

In situations where the above three assumptions are reasonable, one may use as a practical rule-of-thumb for the standard error of $\hat{G}_1^{q,r}$:

$$SE(\hat{G}_1^{q,r}) \approx \sqrt{2} \times CV(\hat{N}_1^q). \qquad (16)$$

That is to say, the standard error of an estimated growth rate is approximately inflated by a factor $\sqrt{2}$ compared to the coefficient of variation (i.e., the relative standard error) of the underlying estimated population count. For a derivation of expression (16), see the appendix.[4]

As was noted in Section 2.2.1, $CV(\hat{N}_1^q) = CV(\hat{\alpha}_1^q)$, so we could also have used the coefficient of variation of an estimated proportion in formula (16). Moreover, we could also have replaced $CV(\hat{N}_1^q)$ by $CV(\hat{N}_1^r)$ in this formula, since it is not difficult to show that assumptions 2 and 3 together imply that these coefficients of variation are equal as well.

Using the form $SE(\hat{G}_1^{q,r}) \approx \sqrt{2} \times CV(\hat{\alpha}_1^q)$, we obtain as an explicit expression:

$$SE(\hat{G}_1^{q,r}) \approx \frac{1}{\sqrt{N^q}} \frac{\sqrt{2\{p_{11}(1-p_{11})\alpha_1^q + p_{00}(1-p_{00})(1-\alpha_1^q)\}}}{p_{11}\alpha_1^q + (1-p_{00})(1-\alpha_1^q)}. \qquad (17)$$

3.3.2 Bias

For the bias of an estimated growth rate, it may be noted first of all that expression (6) is equivalent to:

[4] It is not difficult to show that assumption 1 is actually implied by assumptions 2 and 3. We have listed assumption 1 separately because it may also hold (approximately) in a situation where assumption 2 or assumption 3 fails. In that case, partially simplified expressions for the standard error and bias of $\hat{G}_1^{q,r}$ can still be obtained from the derivations in the appendix, by using only those parts of the derivations that require assumption 1.

$$AB\left(\hat{G}_1^{q,r}\right) = \left(\breve{G}_1^{q,r} - G_1^{q,r}\right) + \breve{G}_1^{q,r}\frac{V\left(\widehat{N}_1^r\right)}{\left[E\left(\widehat{N}_1^r\right)\right]^2} = \left(\breve{G}_1^{q,r} - G_1^{q,r}\right) + \breve{G}_1^{q,r}\left[CV\left(\widehat{N}_1^r\right)\right]^2. \quad (18)$$

It is shown in the appendix that, again under assumptions 1–3, this expression can be simplified to:

$$AB\left(\hat{G}_1^{q,r}\right) = \left[CV\left(\widehat{N}_1^r\right)\right]^2. \quad (19)$$

Thus, as a rule-of-thumb if these assumptions hold approximately, the bias of an estimated growth rate will be approximately equal to the square of the coefficient of variation of the underlying estimated population count. Note that, again, it holds under the assumptions made here that $CV\left(\widehat{N}_1^q\right) = CV\left(\widehat{N}_1^r\right)$, so it does not matter for which period the coefficient of variation is calculated. Also, we could again replace $CV\left(\widehat{N}_1^r\right)$ in (19) by $CV(\hat{\alpha}_1^r)$, since it was seen in Section 2.2.1 that these coefficients of variation are equal.

Using the form $AB\left(\hat{G}_1^{q,r}\right) = [CV(\hat{\alpha}_1^r)]^2$, we obtain as an explicit bias approximation:

$$B\left(\hat{G}_1^{q,r}\right) \approx \frac{1}{N^q}\frac{p_{11}(1-p_{11})\alpha_1^r + p_{00}(1-p_{00})(1-\alpha_1^r)}{[p_{11}\alpha_1^r + (1-p_{00})(1-\alpha_1^r)]^2}. \quad (20)$$

The above results suggest that, in practice, it will often be reasonable to expect that the bias of $\hat{G}_1^{q,r}$ is negligible. The coefficient of variation of $\widehat{N}_1^r$ (or $\hat{\alpha}_1^r$) will usually be a number between 0 and 1, and ideally much closer to 0 than to 1. The square of this number will then be even smaller, possibly by several orders of magnitude.

More precisely, it follows from (16) and (19) [with $CV\left(\widehat{N}_1^r\right) = CV\left(\widehat{N}_1^q\right)$] that the ratio of the bias and standard error of $\hat{G}_1^{q,r}$ is approximately given by:

$$\frac{B\left(\hat{G}_1^{q,r}\right)}{SE\left(\hat{G}_1^{q,r}\right)} \approx \frac{\left[CV\left(\widehat{N}_1^r\right)\right]^2}{\sqrt{2}\times CV\left(\widehat{N}_1^q\right)} = \frac{CV\left(\widehat{N}_1^r\right)}{\sqrt{2}} = O\left(\frac{1}{\sqrt{N^r}}\right).$$

Thus, the bias of an estimated growth rate should become negligible in comparison to its standard error as the population size increases. Note that in Section 3.1.2 the opposite behavior was found for an estimated proportion $\hat{\alpha}_1^q$.

4. Application

To illustrate the results of Section 3, we revisit the application by Hooijschuur et al. (2019) on cybercrime-related aspects in police registrations of crime victims. As introduced in Section 2, in this application a text mining model was used to estimate the proportion of cyber-related crimes in 2016, which we denote here by α_1^{2016}. This proportion is estimated for the registered crime, which concerns close to one million reports. As a result, the standard error of this estimator is negligible; we therefore limit the presented numbers to the part of the formula that does not depend on the population size.

A random sample of 1718 descriptions was drawn out of a total of 980 thousand reports from 2016. This sample was annotated manually by two persons. For the purpose of the illustration in this section, we consider the manually assigned labels as representing the truth for the cases in the sample. The performance of different text mining models for predicting cyber-related crime was tested on the random sample using a tenfold cross-validation.

Hooijschuur et al. (2019) started with a basic text mining model that was aimed at estimating the overall proportion of cyber-related crimes. This basic model consists of a Support Vector Machine (SVM) model. The SVM model uses a bag of words approach with real words as input. Lemmatisation or stemming was not used, since there are no good tools for that in the Dutch language. A limited set of stop words was removed.

As a next step, the authors wanted to use the model to estimate the proportion of cyber-related crimes at a more detailed level, for nineteen different crime categories. (Examples of crime categories are: 'theft/embezzlement and burglary', 'handling stolen goods', and 'assault'.) The performance of the basic model varied per crime category. Using a log-linear model Hooijschuur et al. (2019) identified seven crime categories with for which the association between the true and the predicted labels differed significantly from the others. Next, the parameters of the model were retrained on each of those seven crime categories separately. All other crime categories were taken as an eighth group on which the model was retrained. Details can be found in Hooijschuur et al. (2019). In this section, we will compare the results of the basic and retrained model.

For the basic model, Table 1 shows how many cases in the random sample were classified correctly and how many misclassifications occurred. Table 2 provides the same information for the retrained model. Note that in both tables the number of true cyber-related crimes is the same (168). Below, we will use the associated proportion in the sample as a proxy for the (unknown) true proportion in the population: $\alpha_1^{2016} = 168/1718 = 9.8\%$.

Table 1. Contingency table of true and predicted labels for the annotated sample (basic model).

True class	Predicted class		Total
	1	0	
1	133	35	168
0	33	1517	1550
Total	166	1552	1718

Table 2. Contingency table of true and predicted labels for the annotated sample (retrained model).

True class	Predicted class		Total
	1	0	
1	139	29	168
0	51	1499	1550
Total	190	1528	1718

Using the results from Section 3.1 and the information in Table 1 and Table 2, we can estimate the bias and standard error of the estimated proportion $\hat{\alpha}_1^{2016}$ for both models. For the basic model, Table 1 yields the following matrix of classification error probabilities $\mathbf{P}$ as defined in (1):

$$\mathbf{P}_{basic} = \begin{pmatrix} 0.7917 & 0.2083 \\ 0.0213 & 0.9787 \end{pmatrix}.$$

Substituting these probabilities into (10) and (9), together with $\alpha_1^{2016} = 9.8\%$, we obtain:

$$B(\hat{\alpha}_1^{2016}) = -0.2083 \times 9.8\% + 0.0213 \times 90.2\% = -0.12\%,$$
$$SE(\hat{\alpha}_1^{2016}) = \frac{\sqrt{0.7917 \times 0.2083 \times 9.8\% + 0.9787 \times 0.0213 \times 90.2\%}}{\sqrt{N}} = \frac{0.1869}{\sqrt{N}}.$$

Note that — since the bias does not depend on the size of the dataset — we can check the outcome for the bias from the numbers in Table 1. Since we used 168/1718 as a proxy for α_1^{2016} we find $B(\hat{\alpha}_1^{2016}) = (166 - 168)/1718 = -0.12\%$. From Table 1 we can also directly verify footnote 3: when recall and precision are equal, the estimated bias is 0.

Similarly, for the retrained model we find from Table 2 that

$$\mathbf{P}_{retrained} = \begin{pmatrix} 0.8274 & 0.1726 \\ 0.0329 & 0.9671 \end{pmatrix}.$$

In comparison to the basic model, the retrained model identified a larger proportion of cyber-related crimes correctly. It also misclassified a slightly larger proportion of non-cyber-related crimes as cyber-related. Detailed results in Hooijschuur et al. (2019) show that the association between the true and predicted labels was more homogenous across different crime categories after

retraining the model; in this respect, the retraining was considered successful because its main goal was achieved.

For the bias and standard error of the estimated overall proportion after retraining, we find:

$$B(\hat{\alpha}_1^{2016}) = -0.1726 \times 9.8\% + 0.0329 \times 90.2\% = 1.3\%,$$
$$SE(\hat{\alpha}_1^{2016}) = \frac{\sqrt{0.8274 \times 0.1726 \times 9.8\% + 0.9671 \times 0.0329 \times 90.2\%}}{\sqrt{N}} = \frac{0.2066}{\sqrt{N}}.$$

Interestingly, the magnitude of the bias was increased by retraining, and its sign was switched. This effect is surprisingly large, given that the misclassification probabilities in $\mathbf{P}_{basic}$ and $\mathbf{P}_{retrained}$ are quite similar. To better understand this result, we have drawn a similar graph to Figure 3 for the classification problem in this application, with $\alpha_1^{2016} = 9.8\%$; see Figure 6.

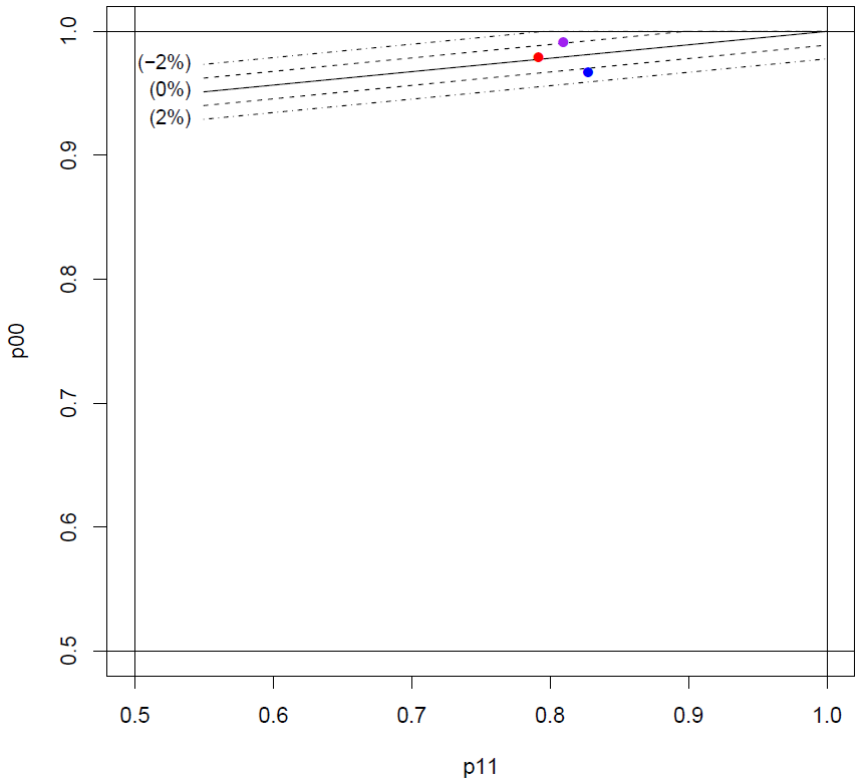

Figure 6. Isolines (p_{11}, p_{00}) such that $B(\hat{\alpha}_1^q) = \beta$ for the cybercrime application with $\alpha_1^q = 9.8\%$ (solid line: $\beta = 0$; dashed lines: $\beta = \pm 1\%$; dot-dashed lines: $\beta = \pm 2\%$). The red, blue and purple points indicate (p_{11}, p_{00}) for the basic, retrained and hypothetical model, respectively.

Given the size of α_1^{2016}, it seems undesirable to have an absolute bias that exceeds 1 or 2 percentage points. Therefore, the lines in Figure 6 show the probability pairs (p_{11}, p_{00}) such that the absolute bias equals 2% (dot-dashed lines) or 1% (dashed lines) or vanishes (solid line). The red dot shows the position of the basic model in this grid with a bias close to zero. By retraining the model, we move to the position

of the blue dot. It is seen in the figure that, although the two points are relatively close together, they lie on different isolines with respect to the bias of $\hat{\alpha}_1^{2016}$.

As a further illustration of the fact that reducing the probabilities of classification errors does not always improve the bias of domain estimators, we consider a hypothetical example. Suppose that we would retrain the model in this application again and somehow end up with the contingency table shown in Table 3.

Table 3. Contingency table of true and predicted labels for the annotated sample (hypothetical model).

True class	Predicted class		Total
	1	**0**	
1	136	32	168
0	14	1536	1550
Total	150	1568	1718

The associated matrix of error probabilities is:

$$\mathbf{P}_{hypothetical} = \begin{pmatrix} 0.8095 & 0.1905 \\ 0.0090 & 0.9910 \end{pmatrix}.$$

In comparison with $\mathbf{P}_{basic}$, both diagonal probabilities in $\mathbf{P}_{hypothetical}$ are closer to 1. Thus, the hypothetical new model is able to predict more accurately whether an *individual* report is cybercrime-related or not, both for reports that truly involve cybercrime and reports that do not. However, on an *aggregated* level we find for the bias and variance of $\hat{\alpha}_1^{2016}$ under this model:

$$B(\hat{\alpha}_1^{2016}) = -0.1905 \times 9.8\% + 0.0090 \times 90.2\% = -1.0\%,$$
$$SE(\hat{\alpha}_1^{2016}) = \frac{\sqrt{0.8095 \times 0.1905 \times 9.8\% + 0.9910 \times 0.0090 \times 90.2\%}}{\sqrt{N}} = \frac{0.1522}{\sqrt{N}}.$$

In comparison with the original model, the absolute bias has increased from 0.12% to 1.0%. In Figure 6, the position of this model on the grid is indicated by the purple dot. It is seen that the pair of probabilities (p_{11}, p_{00}) has been moved in a direction that is not ideal with respect to the bias. A remarkable feature of this example is that the hypothetical new model would be considered an improvement on the original model by many conventional evaluation criteria for machine learning (e.g., recall, precision, accuracy, F1 score).

Note that, in comparison with Figure 3, the isolines in Figure 6 have a smaller slope. In fact, Figure 3 was based on an example with $\alpha_1^q = 30\%$, yielding a slope of $3/7 \approx 0.43$. Here, with $\alpha_1^{2016} = 9.8\%$, the slope of the lines in Figure 6 is $9.8/90.2 \approx 0.11$. It follows from (10) that, for a fixed value of p_{11}, a smaller value of α_1^q implies that the same change in p_{00} will have a larger effect on the bias.

5. Discussion and conclusion

5.1 Summary of results

In this paper, we have presented simplified expressions for the typical bias and standard error of estimated domain proportions, differences and growth rates under random classification errors, for the common situation where the target classification is binary. These simplified expressions provide some insight into the way the bias and standard error are affected by (a) the true proportion of interest, (b) the number of classified cases in the dataset, and (c) the probabilities of correct classification within each class (p_{11}, p_{00}).

In Section 3.1 we used graphs of isocurves to illustrate the influence of these three factors on the bias and standard error of an estimated proportion. Such graphs may be used in practice to find acceptable ranges of classification error probabilities, given a desired accuracy, as was illustrated in Figure 5.

The results for estimated differences and growth rates in this paper hold only under the simplifying assumptions that classification errors at different time points are independent and that the performance of the classification model does not change over time. It is straightforward to relax the latter assumption to allow for different matrices $\mathbf{P}$ at different time points. This is relevant, for instance, when the model has been retrained in the interim. When classification errors for the same unit can be dependent over time, the expressions for the bias and standard error of estimated differences and growth rates become significantly more complicated, even in the binary case; we refer to Scholtus et al. (2019) for details.

In practice, to apply the results in this paper one requires estimates for the probabilities in the matrix $\mathbf{P}$ and for the true proportion α_1^q (and possibly α_1^r). In a supervised machine-learning context, it seems natural to estimate these unknown parameters from the available test set, provided that this is a random sample of sufficient size from the target population. In particular, this approach was used in the application of Section 4. Alternatively, one could construct a dedicated audit sample to estimate $\mathbf{P}$ (Van Delden et al., 2016). Meertens et al. (2019a) developed a Bayesian method for estimating $\mathbf{P}$ which could be useful in particular when only a small audit sample or test set is available. Finally, in the absence of better information, one could estimate α_1^q by the proposed estimator $\hat{\alpha}_1^q$. Estimator $\hat{\alpha}_1^q$ is equivalent to classifying all units in the population by the (trained) classifier and then computing the proportion that is allocated to class 1. Van Delden et al. (2016) showed that in general this yields biased estimates of $B(\hat{\alpha}_1^q)$ and $SE(\hat{\alpha}_1^q)$; see also Meertens et al. (2019b). (Note that in practice the estimation of α_1^q is not the only purpose of using a machine learning model. If it was, one could just estimate α_1^q directly from a sample. For instance, by machine learning one also aims to relate predicted outcomes to background variables.)

5.2 Future work

In an application where the bias of a proposed estimator is found to be too large, a natural follow-up question is how to improve this estimator. For a machine-learning context, we now present four possible approaches to reduce this bias. The relative effectiveness of these approaches, and which of these approaches works best under which conditions, is a point for future research.

First, one could try to improve the classification model, e.g., by changing the type of model or extending the training set. A sufficient improvement is obtained when (p_{11}, p_{00}) lies in its acceptable region as in Figure 5. An interesting idea for future research may be to incorporate the bias directly into the loss function that is minimized when training the model.

Second, one might estimate the proportion $\hat{\alpha}_1^q$ in a different way. The classical machine-learning approach is to classify each case explicitly into class 0 or 1. Alternatively one could first predict for each unit i in the population the *probability* that it belongs to class 1, denoted by P_{1i}^q. Next, one can estimate α_1^q as $\hat{\alpha}_1^q = \sum_i \hat{P}_{1i}^q / N^q$. For many traditional models with estimators based on maximum likelihood or estimating equations, it appears that this approach leads to unbiased estimation of aggregates, provided that all underlying model assumptions are satisfied by the data. For instance, in the case of logistic regression it is well known that the maximum likelihood estimator satisfies $\sum_i \hat{P}_{1i}^q = N_1^q$; see, e.g., Hastie et al. (2009, p. 120). By contrast, explicit classification based on these predicted probabilities generally results in a biased estimator.

Third, one could try to define an alternative estimator that corrects for the bias. In principle, once the bias has been estimated (e.g., using one of the formulas in this paper), a bias-corrected estimator may be obtained immediately by subtracting the estimated bias from the original estimator. In practice, an important drawback of this direct bias correction is that it can significantly inflate the standard error of the estimator. In particular, when the unknown parameters that occur in the bias formula have been estimated from an audit sample or test set, the standard error of the bias-corrected estimator will be determined by the size of this random sample, which may be very small.

Fourth, a more appealing alternative to a bias-corrected estimator may be to correct for the bias at the micro-level. This can be done by imputing a new, classification-error-corrected version of the original predicted classification. This has the advantage that micro-level auxiliary information can be incorporated into the correction procedure. Two approaches that have been proposed recently for correcting classification errors in this way are Multiple Over-imputation (Blackwell et al., 2017) and Multiple Imputation of Latent Classes (Boeschoten, 2019).

References

M. Blackwell, J. Honaker, and G. King (2017), A Unified Approach to Measurement Error and Missing Data: Overview and Applications. *Sociological Methods & Research* **46**, 303–341.

L. Boeschoten (2019), *Consistent Estimates for Categorical Data Based on a Mix of Administrative Data Sources and Surveys*. PhD Thesis, Tilburg University.

J. Burger, A. van Delden, and S. Scholtus (2015), Sensitivity of Mixed-Source Statistics to Classification Errors. *Journal of Official Statistics* **31**, 489–506.

A. van Delden, S. Scholtus, and J. Burger (2016), Accuracy of Mixed-Source Statistics as Affected by Classification Errors. *Journal of Official Statistics* **32**, 619–642.

S. van der Doef, P. Daas, and D. Windmeijer (2018), Identifying Innovative Companies from their Website. Presentation at BigSurv18 conference. Abstract available at http://www.bigsurv18.org/program2018?sess=34#205 (retrieved: June 2019).

T. Hastie, R. Tibshirani, and J. Friedman (2009), *The Elements of Statistical Learning* (Second Edition). Springer, New York.

E. Hooijschuur, A. van Delden, L. Jong, D. Windmeijer, and C. Verkleij (2019), Towards implementing a text mining model to detect cybercrime in police reports. Paper presented to Statistics Netherlands' Advisory Board on Methodology and Quality; available on request.

J. Kuha and C. Skinner (1997), Categorical Data Analysis and Misclassification. In: Lyberg, Biemer, Collins, de Leeuw, Dippo, Schwarz, and Trewin (eds.), *Survey Measurement and Process Quality*. New York: John Wiley & Sons, pp. 633–670.

Q.A. Meertens, C.G.H. Diks, H.J. van den Herik, and F.W. Takes (2019a), A Bayesian Approach for Accurate Classification-Based Aggregates. In: *Proceedings of the 2019 SIAM International Conference on Data Mining*, Calgary, pp. 306–314. Available at https://doi.org/10.1137/1.9781611975673.35 (retrieved: December 2019).

Q.A. Meertens, C.G.H. Diks, H.J. van den Herik, and F.W. Takes (2020), A Data-Driven Supply-Side Approach for Estimating Cross-Border Internet Purchases within the European Union. *Journal of the Royal Statistical Society: Series A* **183**, 61–90.

Q.A. Meertens, A. van Delden, S. Scholtus, and F.W. Takes (2019b), Bias Correction for Predicting Election Outcomes with Social Media Data. Paper presented at the 5th International Conference on Computational Social Science, Amsterdam. Available at

http://www.researchgate.net/publication/333661444_Bias_Correction_for_Predicting_Election_Outcomes_with_Social_Media_Data (retrieved: August 2019).

D. Nelson (ed.) (2003), *The Penguin Dictionary of Mathematics* (3rd Edition). Penguin Books, London.

S. Scholtus, A. van Delden, and J. Burger (2019), Evaluating the Accuracy of Growth Rates in the Presence of Classification Errors. Discussion Paper, CBS, The Hague/Heerlen.

N. Tollenaar, J. Rokven, D. Macro, M. Beerthuizen, and A.M. van der Laan (2019), Predictieve textmining in politieregistraties. Cyber- en gedigitaliseerde criminaliteit. Report (in Dutch), Wetenschappelijk Onderzoek- en Documentatiecentrum. Available at http://www.wodc.nl/binaries/Cahier%202019-2_2849b_Volledige%20tekst_tcm28-375448.pdf (retrieved: December 2019).

Appendix: Additional derivations

Derivation of (16)

Under assumption 1 in Section 3.3.1, it follows that $\check{G}_1^{q,r}$ from expression (8) can be simplified:

$$\check{G}_1^{q,r} = \frac{E(\hat{N}_1^q)}{E(\hat{N}_1^r)} = \frac{N_1^q + B(\hat{N}_1^q)}{N_1^r + B(\hat{N}_1^r)} = \frac{N_1^q}{N_1^r} \times \frac{1 + RB(\hat{N}_1^q)}{1 + RB(\hat{N}_1^r)} = \frac{N_1^q}{N_1^r} = G_1^{q,r}.$$

Substituting this simplification into expression (7) for $AV(\hat{G}_1^{q,r})$, we find:

$$\begin{aligned} AV(\hat{G}_1^{q,r}) &= \frac{p_{11}(1-p_{11})\left[N_1^q + (G_1^{q,r})^2 N_1^r\right] + p_{00}(1-p_{00})\left[N_0^q + (G_1^{q,r})^2 N_0^r\right]}{\left[E(\hat{N}_1^r)\right]^2} \\ &= \frac{p_{11}(1-p_{11})N_1^q(1+G_1^{q,r}) + p_{00}(1-p_{00})N_0^q\left(1 + G_1^{q,r}\frac{N_1^q}{N_1^r}\frac{N_0^r}{N_0^q}\right)}{\left[E(\hat{N}_1^r)\right]^2}. \end{aligned}$$

Here, it was also used that $E(\hat{N}_1^r) = p_{11}N_1^r + (1-p_{00})N_0^r$, as seen in expression (8).

An alternative form for the product at the end of the numerator in the last line may be obtained as follows:

$$\frac{N_1^q}{N_1^r}\frac{N_0^r}{N_0^q} = \frac{N_1^q}{N_1^r}\frac{N^r - N_1^r}{N^q - N_1^q} = \frac{N_1^q}{N_1^r}\frac{N^r}{N^q}\frac{1-\alpha_1^r}{1-\alpha_1^q} = \frac{\alpha_1^q}{\alpha_1^r}\frac{1-\alpha_1^r}{1-\alpha_1^q}.$$

Hence,

$$\begin{aligned} AV(\hat{G}_1^{q,r}) = &\frac{p_{11}(1-p_{11})N_1^q(1+G_1^{q,r}) + p_{00}(1-p_{00})N_0^q(1+G_1^{q,r})}{\left[E(\hat{N}_1^r)\right]^2} \\ &+ \frac{p_{00}(1-p_{00})N_0^q G_1^{q,r}\left(\frac{\alpha_1^q}{\alpha_1^r}\frac{1-\alpha_1^r}{1-\alpha_1^q} - 1\right)}{\left[E(\hat{N}_1^r)\right]^2}. \end{aligned}$$

The first term in this expression is equal to (cf. Section 2.2.1):

$$\begin{aligned} \frac{V(\hat{N}_1^q)}{\left[E(\hat{N}_1^r)\right]^2}(1+G_1^{q,r}) &= \left[CV(\hat{N}_1^q)\right]^2(1+G_1^{q,r})\left[\frac{E(\hat{N}_1^q)}{E(\hat{N}_1^r)}\right]^2 \\ &= \left[CV(\hat{N}_1^q)\right]^2(1+G_1^{q,r})(G_1^{q,r})^2, \end{aligned}$$

where we used again that, under assumption 1, $\check{G}_1^{q,r} = E(\hat{N}_1^q)/E(\hat{N}_1^r) = G_1^{q,r}$.

The second term can be re-written as:

$$\frac{p_{00}(1-p_{00})N_0^q G_1^{q,r}\dfrac{\alpha_1^q-\alpha_1^r}{\alpha_1^r(1-\alpha_1^q)}}{\left[E\left(\widehat{N}_1^r\right)\right]^2}.$$

Under assumption 2, this second term is negligible and we conclude that

$$AV\left(\hat{G}_1^{q,r}\right)=\left[CV\left(\widehat{N}_1^q\right)\right]^2\left(1+G_1^{q,r}\right)\left(G_1^{q,r}\right)^2.$$

Finally, assumptions 2 and 3 together imply that $G_1^{q,r}=1$ and so it follows that

$$AV\left(\hat{G}_1^{q,r}\right)=2\left[CV\left(\widehat{N}_1^q\right)\right]^2.$$

Upon taking the square root, expression (16) follows.

Derivation of (19)

As was seen in the above derivation, assumption 1 implies that $\breve{G}_1^{q,r}=G_1^{q,r}$. Hence, expression (18) for the approximate bias of $\hat{G}_1^{q,r}$ is reduced to:

$$AB\left(\hat{G}_1^{q,r}\right)=G_1^{q,r}\left[CV\left(\widehat{N}_1^r\right)\right]^2.$$

Furthermore, it was noted in the previous derivation that assumptions 2 and 3 together imply that $G_1^{q,r}=1$. Hence, under these assumptions, expression (19) follows.

Explanation of symbols

Empty cell	Figure not applicable
.	Figure is unknown, insufficiently reliable or confidential
*	Provisional figure
**	Revised provisional figure
2017–2018	2017 to 2018 inclusive
2017/2018	Average for 2017 to 2018 inclusive
2017/'18	Crop year, financial year, school year, etc., beginning in 2017 and ending in 2018
2013/'14–2017/'18	Crop year, financial year, etc., 2015/'16 to 2017/'18 inclusive

Due to rounding, some totals may not correspond to the sum of the separate figures.

Colophon

Publisher
Centraal Bureau voor de Statistiek
Henri Faasdreef 312, 2492 JP Den Haag
www.cbs.nl

Prepress
Statistics Netherlands, CCN Creation and visualisation

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contactform: www.cbsl.nl/information