Statistics Norway
Statistical Methods and Standards

Anna-Karin Mevik

Documents

Uncertainty in the Norwegian Business Tendency Survey

Statistics Norway
Statistisk sentralbyrå

1. Introduction **2**

2. Population **2**
2.1. Stratification 2
2.2. The diffusion index 2

3. The sample design **4**

4. Estimation of the diffusion index **4**

5. Design-based analysis of the estimator **5**
5.1. Linearization 7
5.2. Bootstrap 13

6. Model-based analysis of the estimator **13**

7. Illustration **18**

8. An alternative estimator: Based on a corrected Horvitz-Thompson estimator **22**
8.1. Design-based analysis 23
8.2. Model-based analysis 26
8.3. Illustration 28

9. An alternative estimator: Based on best linear unbiased predictor **32**
9.1. Design-based analysis 32
9.2. Model-based analysis 36
9.3. Illustration 38

10. Summary **42**

1. Introduction

In this paper we are going to look at how to measure and estimate the uncertainty in the Norwegian Business Tendency Survey for manufacturing, quarrying and mining. We will disregard the effect of non-response, measurement error and coverage error, and focus on the uncertainty coming from estimating a population total based on a sample. This uncertainty is usually measured by the design-based standard error, but we will also use a model-based measure. In addition to investigating the uncertainty of the estimator in use today (section 4-7), we introduce two alternative estimators for the existing one (section 8 and 9).

The Norwegian Business Tendency Survey is a quarterly survey, and maps out the industrial management leaders judgement of the business situation and the outlook for a fixed set of indicators such as level of production, capacity utilisation and employment. We shall consider the diffusion index, which is a measure of the outlooks for the level of production in the next quarter. (Our analysis and the estimators we derive for this index will also hold for similar indexes in the Norwegian Business Tendency Survey).

In section 2 we describe the population and the diffusion index, and in section 3 the sample design. The estimator of the diffusion index is introduced in section 4, and in section 5 and 6 we give a design-based and a model-based analysis of the estimator. The measurement of uncertainty these analyses give, are used to estimate the uncertainty in the Business Tendency Survey in the period 1999, 2000 and 2002. The result of this estimation is presented in section 7. In section 8 and 9 we consider two alternative estimators of the diffusion index. Finally, a summary is given in section 10.

2. Population

The unit of analysis is the branch unit. The branch unit comprises all establishments within an enterprise belonging to the same 3-digit industry group (SIC94) - in the following referred to as the branch. The population covers all branch units within the industries Mining and quarrying (10,13-14) and Manufacturing (15-37), see the Standard of Industrial Classification 1994 (SIC94). The Business and Enterprise Register defines the population.

2.1. Stratification

Branch and number of employees stratify the population. The employment intervals are 0-99, 100-199, 200-299 and 300-∞. There are about 270 strata in the population.

2.2. The diffusion index

The diffusion index is a measure of how the branch units are judging the level of their productions in the next quarter compared to the current quarter. More precisely the diffusion index for a given quarter is given by

$$d = S + \frac{1}{2}U,$$

where

$$S = \frac{s}{X} \cdot 100 \qquad \text{and} \qquad U = \frac{u}{X} \cdot 100 .$$

Here s is the number of employees in branch units which expect that the production (to the branch unit) will be larger in the next quarter compared to the current quarter, u is the number of employees in branch units which expect that the production (to the branch unit) will be more or less the same in the next quarter compared to the current quarter, and X is the total number of employees for all branch units in the population. That is, d gives the share of employees that are working in a branch unit that expects that its production will increase, plus half the share of employees that are working in a branch unit that expects that its production not will change.

Because the expression $S+1/2U$ is unsuitable for the analyses we are going to do, we define some new variables and use they to obtain a more suitable expression of the diffusion index.

For branch unit i in stratum h, we define the tendency variable

$$y_{hi}=\begin{cases} x_{hi} & \text{,if the branch unit expects its production to increase} \\ \frac{1}{2}x_{hi} & \text{,if the branch unit expects its production to be unchang} \\ 0 & \text{,if the branch unit expects its production to be smaller} \end{cases},$$

where x_{hi} is the employment of the branch unit.

We now define the tendency variable of the stratum by

$$Y_h=\sum_{i\in U_h} y_{hi},$$

where the subscript h refers to the stratum label and U_h denotes the branch units belonging to the stratum. The tendency variable of the stratum is equal to the employment of the branch units (within the stratum) that expect their productions to increase, plus half the employment of the branch units (within the stratum) that expect their productions not to change. As more branch units expect the production to increase the larger Y_h becomes. But the tendency variable could never exceed the employment in the stratum, that is,

$$Y_h \le X_h$$

where $X_h=\sum_{i\in U_h} x_{hi}$ is the employment in the stratum.

Finally, we define the tendency variable of the population by

$$Y=\sum_h Y_h,$$

where the sum $\sum_h$ is over all strata in the population. That is, the tendency variable of the population is equal to the employment of all the branch units that expect their productions to increase, plus half the employment of all the branch units that expect their productions not to change. Like Y_h, the tendency variable Y is larger as more branch units expect the production to increase, but Y will never exceed the total employment in the population.

With these definitions the diffusion index can be written

$$(1) \qquad d = \frac{Y}{X} \cdot 100 = \frac{\sum_h Y_h}{\sum_h X_h} \cdot 100 ,$$

where $X = \sum_h X_h$ is the total employment in the population. In the rest of the paper we will be using this presentation of the diffusion index.

Since $0 \le Y \le X$, the diffusion index will lie between 0 and 100. If $d > 50$ it means that the branch units that expect an increase in the production employ more people than the branch units that expect a reduction. The opposite situation, that is $d < 50$, means that the branch units that expect a reduction in the production employ more people than the branch units that expect an increase. We could say that $d > 50$ indicates expected growth in the production while $d < 50$ indicates an expected reduction.

The diffusion index (1) refers to the entire population. We are also interested in diffusion indexes for domains. A domain could for example be a branch, or a group of branches. The diffusion index for such a domain is also given by (1), but where the sum $\sum_h$ is only over the strata within the domain.

3. The sample design

A new sample is selected each year. If necessary, the sample is updated or rotated during the four quarters. The sample size is about 710. Branch units with more than 300 employees are included as a panel. Branch units with less than 10 employees are never included in the sample, that is, they have probability 0 of being selected.

Proportional allocation is used to decide the size of the stratum samples. The allocation does not ensure that we get a sample form each stratum, so we might have strata without sample.

The selections in the different strata are done independently. Within the stratum the sample is selected with probabilities proportional to size, where the size is the employment to the branch unit (branch units with less than 10 employees have probability 0 of being selected).

4. Estimation of the diffusion index

The estimator of the diffusion index (1) can be written as

$$(2) \qquad \hat{d} = \frac{\sum_h \hat{Y}_h}{\sum_h X_h} \cdot 100 ,$$

where

$$(3) \qquad \hat{Y}_h = \frac{\overline{y}_{s_h}}{\overline{x}_{s_h}} \cdot X_h$$

is a ratio estimator of the tendency variable $Y_h = \sum_{i \in U_h} y_{hi}$. Here, $\overline{x}_{s_h}$ and $\overline{y}_{s_h}$ are given by

$$\overline{x}_{s_h} = \frac{1}{n_h} \sum_{i \in s_h} x_{hi} \qquad \text{and} \qquad \overline{y}_{s_h} = \frac{1}{n_h} \sum_{i \in s_h} y_{hi} ,$$

where s_h is the response sample from stratum h and n_h is the size of s_h.

The estimator (2) requires a sample from each of the strata. As previously mentioned we can have strata with no sample. Such strata are probably removed from the population when $\hat{d}$ is to be calculated, so that the sum $\sum_h$ is only over the strata with sample.

The inequality $0 \le y_{hi} \le x_{hi}$ implies that $0 \le \overline{y}_{s_h} \le \overline{x}_{s_h}$ so that $0 \le \hat{Y}_h \le X_h$. Hence, $\hat{d}$ lies between 0 and 100, which is a necessary property since the diffusion index lies between 0 and 100. In the next two sections we analyse $\hat{d}$ further.

The diffusion index for the domains is also estimated by (2), but again the sum $\sum_h$ is only over the strata within the group.

5. Design-based analysis of the estimator

This section presents a design-based analysis of $\hat{d}$. It means that we are looking at the tendency variables to the branch units as parameters, while the sample is stochastic. Thus in this section the variance, bias etc. will be derived under that situation. We assume a sample from each stratum, and take no account of non-response, measurement error or coverage error.

A desired property of an estimator is unbiasedness. This means that the expectation of $\hat{d}$ should equal d for all possible values of the y_{hi}'s, but this is not the case[1]. To see this we write the expectation as

$$\mathrm{E}\hat{d} = \frac{\sum_h \mathrm{E}\hat{Y}_h}{\sum_h X_h} \cdot 100 ,$$

where $\mathrm{E}\hat{Y}_h$ is the expectation of $\hat{Y}_h$. Later on it is shown that $\hat{Y}_h$ is generally not unbiased: The expectation of $\hat{Y}_h$ can be both greater and smaller than $\hat{Y}_h$, depending on the y_{hi}'s. Hence, the expectation of $\hat{d}$ can be both greater and smaller than d, and $\hat{d}$ is thus not unbiased.

Since we have independence between the strata, the variance of $\hat{d}$ is given by

$$\mathrm{V}\left(\hat{d}\right) = \frac{\sum_h \mathrm{V}\left(\hat{Y}_h\right)}{\left(\sum_h X_h\right)^2} \cdot 100^2 ,$$

[1] Even though $\hat{d}$ is not unbiased, a simulation study shows that $\mathrm{E}\hat{d} \approx d$ for most of the configuration of the y_{hi}'s.

where $\mathrm{V}\left(\hat{Y}_h\right)$ is the variance of $\hat{Y}_h$. The standard error of $\hat{d}$, defined as the square root of the variance, is therefore

$$\text{(4)} \qquad \text{s.e.}\left(\hat{d}\right) = \frac{\sqrt{\sum_h \mathrm{V}\left(\hat{Y}_h\right)}}{\sum_h X_h} \cdot 100 .$$

The uncertainty of an estimator is usually measured by the standard error. But in a situation where the estimator is not unbiased, the mean squared error might be a better measure. The mean squared error is defined as the expectation of the square of the error, and equals the variance plus the square of the bias (the bias of an estimator is defined by the expectation of the error). That is, the mean squared error of $\hat{d}$ is given by

$$\begin{aligned} \mathrm{MSE}\left(\hat{\mathrm{d}}\right) &= \mathrm{E}\left[\left(\hat{d} - d\right)^2\right] \\ &= \mathrm{V}\left(\hat{\mathrm{d}}\right) + \mathrm{Bias}\left(\hat{\mathrm{d}}\right)^2 , \end{aligned}$$

where $\mathrm{Bias}\left(\hat{d}\right) = \mathrm{E}\left[\hat{d} - d\right]$ is the bias of $\hat{d}$.

Despite the fact that $\hat{d}$ is not unbiased, we choose to measure the uncertainty of $\hat{d}$ with the standard error. The reason for this choice is that we are not able to estimate the mean squared error of $\hat{d}$. (We give an upper and a lower bound of the bias of $\hat{d}$ in subsection 5.1).

Since the standard error (4) is unknown it has to be estimated. That is done by estimating the variance of the $\hat{Y}_h$'s. There is no standard estimator of $\mathrm{V}\left(\hat{Y}_h\right)$, and we will use linearization to obtain an approximate variance that can be estimated (subsection 5.1). We have also considered using bootstrap to estimate the variance, but have not found any method suitable for our situation (subsection 5.2).

Before we begin with the linearization in subsection 5.1, we introduce the inclusion probabilities, and use the explicit formula of these probabilities to obtain an upper bound of $\mathrm{V}\left(\hat{Y}_h\right)$ when $n_h = 1$.

The inclusion probability of a branch unit i in stratum h is denoted π_{hi} and is the probability that the branch unit is included in the sample. Since branch units with less than 10 employees are excluded, these units have $\pi_{hi} = 0$. For the other branch units the inclusion probabilities are proportional to the employment (within each stratum). Using the notation

$$U_h^* = \left\{ i \in U_h \,:\, \pi_{hi} > 0 \right\},$$

we have $\pi_{hi} = 0$ when $i \notin U_h^*$, and $\pi_{hi} = n_h \left(x_{hi} / \sum_{i \in U_h^*} x_{hi} \right)$ when $i \in U_h^*$ (provided that $n_h \left(x_{hi} / \sum_{i \in U_h^*} x_{hi} \right) \leq 1$ for all $i \in U_h^*$, which is satisfied for most strata).

Now, let us consider the situation where $n_h = 1$. Then it is impossible to estimate the variance of $\hat{Y}_h$, and we shall use an upper bound as a conservative estimate of the variance. By writing $\hat{Y}_h = \left(y_{hi_s} / x_{hi_s} \right) X_h$ and $\pi_{hi} = x_{hi} / \sum_{i \in U_h^*} x_{hi}$, where i_s denotes the selected branch unit, the expectation can be written

$$
\begin{aligned}
\mathrm{E}\hat{Y}_h &= \sum_{i \in U_h^*} \frac{y_{hi}}{x_{hi}} X_h \pi_{hi} \\
&= \sum_{i \in U_h^*} y_{hi} + \frac{\sum_{i \in U_h^*} y_{hi}}{\sum_{i \in U_h^*} x_{hi}} \sum_{i \notin U_h^*} x_{hi}.
\end{aligned} \tag{5}
$$

This expectation is generally not equal Y_h, unless $U_h^* = U_h$. That is, $\hat{Y}_h$ is generally not unbiased. The variance of $\hat{Y}_h$ can be written as

$$
\mathrm{V}\left(\frac{y_{hi_s}}{x_{hi_s}} X_h \right) = X_h^2 \left(\frac{\sum_{i \in U_h^*} y_{hi}^2 / x_{hi}}{\sum_{i \in U_h^*} x_{hi}} - \left(\frac{\sum_{i \in U_h^*} y_{hi}}{\sum_{i \in U_h^*} x_{hi}} \right)^2 \right). \tag{6}
$$

Using that $y_{hi} \le x_{hi}$, gives

$$
\frac{\sum_{i \in U_h^*} y_{hi}^2 / x_{hi}}{\sum_{i \in U_h^*} x_{hi}} - \left(\frac{\sum_{i \in U_h^*} y_{hi}}{\sum_{i \in U_h^*} x_{hi}} \right)^2 \le \frac{\sum_{i \in U_h^*} y_{hi}}{\sum_{i \in U_h^*} x_{hi}} - \left(\frac{\sum_{i \in U_h^*} y_{hi}}{\sum_{i \in U_h^*} x_{hi}} \right)^2 \le \frac{1}{2} - \left(\frac{1}{2} \right)^2 = \frac{1}{4},
$$

and we get the inequality

$$
\mathrm{V}\left(\frac{y_{hi_s}}{x_{hi_s}} X_h \right) \le \frac{1}{4} X_h^2 .
$$

That is, $\mathrm{V}\left(\hat{Y}_h \right) \le X_h^2 / 4$ when $n_h = 1$, and we use the bound $X_h^2 / 4$ to estimate the variance. (Theoretically we could have found a better bound by maximizing with respect to the possible configuration of the y_{hi}'s. But if N_h is large, the maximisations are impossible).

5.1. Linearization

Generally, assume we are going to estimate the variance $\mathrm{V}\left(\hat{\theta} \right)$, where $\hat{\theta} = g\left(\bar{\boldsymbol{y}} \right)$ is an estimator of $\theta = g\left(\bar{\boldsymbol{Y}} \right)$. Here, $\bar{\boldsymbol{Y}} = \left(\bar{Y}_1, \ldots, \bar{Y}_p \right)$ and $\bar{\boldsymbol{y}} = \left(\bar{y}_1, \ldots, \bar{y}_p \right)$ are the vectors of population means and sample means, respectively. By doing a first order Taylor expansion of g around the expectation $\mathrm{E}\bar{\boldsymbol{y}} = \left(\mathrm{E}\bar{y}_1, \ldots, \mathrm{E}\bar{y}_p \right)$, we obtain the approximation

$$g(\bar{\boldsymbol{y}}) \approx g(\mathrm{E}\bar{\boldsymbol{y}}) + \sum_{j=1}^{p} a_j \left(\bar{y}_j - \mathrm{E}\bar{y}_j\right),$$

where $a_j = \dfrac{\partial g}{\partial y_j}(\mathrm{E}\bar{\boldsymbol{y}})$. This approximation leads to the following approximation of the variance:

$$\mathrm{V}(\hat{\theta}) \approx \sum_{j=1}^{p} a_j^2 \mathrm{V}(\bar{y}_j) + \sum_{j=1}^{p} \sum_{\substack{i=1 \\ i \neq j}}^{p} a_j a_i \mathrm{C}(\bar{y}_j, \bar{y}_i),$$

where $\mathrm{C}(\bar{y}_i, \bar{y}_j)$ is the covariance of $\bar{y}_i$ and $\bar{y}_j$. The variance of $\hat{\theta}$ is now estimated by estimating a_j by $\hat{a}_j = \partial g / \partial y_j (\bar{\boldsymbol{y}})$, and choosing suitable estimators of $\mathrm{V}(\bar{y}_j)$ and $\mathrm{C}(\bar{y}_j, \bar{y}_i)$.

In our situation we shall estimate the variance of $\hat{Y}_h = g(\bar{y}_{s_h}, \bar{x}_{s_h})$, with $g(y_1, y_2) = (y_1 / y_2) X_h$. The approximation of $\hat{Y}_h$ and $\mathrm{V}(\hat{Y}_h)$ becomes, respectively,

(7) $$\hat{Y}_h \approx \frac{\mathrm{E}\bar{y}_{s_h}}{\mathrm{E}\bar{x}_{s_h}} \cdot X_h + \frac{1}{\mathrm{E}\bar{x}_{s_h}} \cdot X_h \cdot \left(\bar{y}_{s_h} - \mathrm{E}\bar{y}_{s_h}\right) - \frac{\mathrm{E}\bar{y}_{s_h}}{\left(\mathrm{E}\bar{x}_{s_h}\right)^2} \cdot X_h \cdot \left(\bar{x}_{s_h} - \mathrm{E}\bar{x}_{s_h}\right)$$

$$\mathrm{V}(\hat{Y}_h) \approx \frac{1}{\left(\mathrm{E}\bar{x}_{s_h}\right)^2} \cdot X_h^2 \cdot \mathrm{V}(\bar{y}_{s_h}) + \frac{\left(\mathrm{E}\bar{y}_{s_h}\right)^2}{\left(\mathrm{E}\bar{x}_{s_h}\right)^4} \cdot X_h^2 \cdot \mathrm{V}(\bar{x}_{s_h}) - 2 \frac{\mathrm{E}\bar{y}_{s_h}}{\left(\mathrm{E}\bar{x}_{s_h}\right)^3} \cdot X_h^2 \cdot \mathrm{C}(\bar{x}_{s_h}, \bar{y}_{s_h}),$$

and the proposed variance estimator of $\hat{Y}_h$ is

$$\hat{\mathrm{V}}(\hat{Y}_h) = \frac{1}{\left(\bar{x}_{s_h}\right)^2} \cdot X_h^2 \cdot \hat{\mathrm{V}}(\bar{y}_{s_h}) + \frac{\left(\bar{y}_{s_h}\right)^2}{\left(\bar{x}_{s_h}\right)^4} \cdot X_h^2 \cdot \hat{\mathrm{V}}(\bar{x}_{s_h}) - 2 \frac{\bar{y}_{s_h}}{\left(\bar{x}_{s_h}\right)^3} \cdot X_h^2 \cdot \hat{\mathrm{C}}(\bar{x}_{s_h}, \bar{y}_{s_h})$$

where $\hat{\mathrm{V}}(\bar{y}_{s_h})$, $\hat{\mathrm{V}}(\bar{x}_{s_h})$ and $\hat{\mathrm{C}}(\bar{x}_{s_h}, \bar{y}_{s_h})$ are estimators of $\mathrm{V}(\bar{y}_{s_h})$, $\mathrm{V}(\bar{x}_{s_h})$ and $\mathrm{C}(\bar{x}_{s_h}, \bar{y}_{s_h})$, respectively.

So we need to find estimators of $\mathrm{V}(\bar{y}_{s_h})$, $\mathrm{V}(\bar{x}_{s_h})$ and $\mathrm{C}(\bar{x}_{s_h}, \bar{y}_{s_h})$. We begin with the variance of $\bar{y}_{s_h}$, which can be written

$$\begin{aligned} \mathrm{V}(\bar{y}_{s_h}) &= \frac{1}{n_h^2} \sum_{i \in U_h^*} \sum_{j \in U_h^*} y_{hi} y_{hj} \left(\pi_{hij} - \pi_{hi}\pi_{hj}\right) \\ &= \frac{1}{n_h^2} \frac{1}{2} \sum_{i \in U_h^*} \sum_{\substack{j \in U_h^* \\ j \neq i}} \left(\pi_{hi}\pi_{hj} - \pi_{hij}\right)\left(y_{hi} - y_{hj}\right)^2. \end{aligned}$$

Here, π_{hij} denotes the probability that branch units i and j will be included in the sample (second-order inclusion probability), and $\pi_{hii} = \pi_{hi}$. Two estimators of this variance are

$$\hat{\mathrm{V}}_{\mathrm{HT}}\left(\bar{y}_{s_h}\right)=\frac{1}{n_h^2}\sum_{i\in s_h}\sum_{j\in s_h}\frac{y_{hi}y_{hj}\left(\pi_{hij}-\pi_{hi}\pi_{hj}\right)}{\pi_{hij}}$$

and

$$\hat{\mathrm{V}}_{\mathrm{SYG}}\left(\bar{y}_{s_h}\right)=\frac{1}{2}\frac{1}{n_h^2}\sum_{i\in s_h}\sum_{\substack{j\in s_h\\ j\neq i}}\frac{\pi_{hi}\pi_{hj}-\pi_{hij}}{\pi_{hij}}\left(y_{hi}-y_{hj}\right)^2 .$$

HT stands for Horvitz-Thompson and SYG stands for Sen-Yates-Grundy. Either estimator is unbiased if $\pi_{hij}>0$ whenever $\pi_{hi}>0$ and $\pi_{hj}>0$[2]. But unfortunately they might take negative values.

To use the estimators $\hat{\mathrm{V}}_{\mathrm{HT}}\left(\bar{y}_{s_h}\right)$ and $\hat{\mathrm{V}}_{\mathrm{SYG}}\left(\bar{y}_{s_h}\right)$ we need to know the π_{hij}'s. When $i=j$ we have $\pi_{hij}=\pi_{hii}=\pi_{hi}$, and when $i\neq j$ along with $\pi_{hi}=1$ or $\pi_{hj}=1$ we have $\pi_{hij}=\pi_{hi}\pi_{hj}$. So we know these π_{hij}'s since the π_{hi}'s are known. But we do not know the π_{hij}'s when $i\neq j$ with $\pi_{hi}\neq 1$ and $\pi_{hj}\neq 1$. Thus, to be able to estimate $\mathrm{V}\left(\bar{y}_{s_h}\right)$ with $\hat{\mathrm{V}}_{\mathrm{HT}}\left(\bar{y}_{s_h}\right)$ or $\hat{\mathrm{V}}_{\mathrm{SYG}}\left(\bar{y}_{s_h}\right)$, we approximate the unknown π_{hij}'s with $\hat{\pi}_{hij}=\dfrac{N_h^*\left(n_h-1\right)}{n_h\left(N_h^*-1\right)}\cdot\pi_{hi}\pi_{hj}$, where N_h^* is the size of U_h^*.[3]

With this approximation of the unknown π_{hij}'s, we get that $\pi_{hi}\pi_{hj}-\pi_{hij}\geq 0$ for all $i\neq j$. This ensures that $\hat{\mathrm{V}}_{\mathrm{SYG}}\left(\bar{y}_{s_h}\right)$ always is positive. It can also be shown that if π_{hij} is proportional to $\pi_{hi}\pi_{hj}$ when $i\neq j$, $i,j\in U_h^*$, then $\hat{\mathrm{V}}_{\mathrm{SYG}}\left(\bar{y}_{s_h}\right)$ becomes larger if we use $\hat{\pi}_{hij}$ instead of π_{hij}, while $\hat{\mathrm{V}}_{\mathrm{HT}}\left(\bar{y}_{s_h}\right)$ becomes smaller. For these reasons we choose to estimate $\hat{\mathrm{V}}\left(\bar{y}_{s_h}\right)$ with $\hat{\mathrm{V}}_{\mathrm{SYG}}\left(\bar{y}_{s_h}\right)$, that is, we estimate $\hat{\mathrm{V}}\left(\bar{y}_{s_h}\right)$ with

(8) $$\hat{\mathrm{V}}_{\mathrm{SYG}}\left(\bar{y}_{s_h}\right)=\frac{1}{2}\frac{1}{n_h^2}\sum_{i\in s_h}\sum_{\substack{j\in s_h\\ j\neq i}}\frac{\pi_{hi}\pi_{hj}-\hat{\pi}_{hij}}{\hat{\pi}_{hij}}\left(y_{hi}-y_{hj}\right)^2$$

where

---

[2] If $\pi_{hij}>0$ when π_{hi} and $\pi_{hj}>0$, we can write $\hat{\mathrm{V}}_{\mathrm{HT}}\left(\bar{y}_{s_h}\right)=\dfrac{1}{n_h^2}\sum_{i\in U_h^*}\sum_{j\in U_h^*}\dfrac{y_{hi}y_{hj}\left(\pi_{hij}-\pi_{hi}\pi_{hj}\right)}{\pi_{hij}}\cdot I_{hi}I_{hj}$. Here, $I_{hi}=1$ if the branch unit i is included in the sample and 0 otherwise. This gives $\mathrm{E}\left[\hat{\mathrm{V}}_{\mathrm{HT}}\left(\bar{y}_{s_h}\right)\right]=$ $\dfrac{1}{n_h^2}\sum_{i\in U_h^*}\sum_{j\in U_h^*}\dfrac{y_{hi}y_{hj}\left(\pi_{hij}-\pi_{hi}\pi_{hj}\right)}{\pi_{hij}}\cdot\mathrm{E}\left[I_{hi}I_{hj}\right]=\dfrac{1}{n_h^2}\sum_{i\in U_h^*}\sum_{j\in U_h^*}\dfrac{y_{hi}y_{hj}\left(\pi_{hij}-\pi_{hi}\pi_{hj}\right)}{\pi_{hij}}\cdot\pi_{hij}=\mathrm{V}\left(\bar{y}_{s_h}\right)$. In a similar way it is shown that $\hat{\mathrm{V}}_{\mathrm{SYG}}\left(\hat{Y}_h\right)$ is unbiased.

[3] This approximation is from a course in model based survey estimation (lectured by Ray Chamber).

$$\hat{\pi}_{hij} = \begin{cases} \pi_{hi} & \text{, when } i = j \\ \dfrac{N_h^*(n_h-1)}{n_h(N_h^*-1)} \cdot \pi_{hi}\pi_{hj} & \text{, when } i \neq j,\ \pi_{hi} \neq 1 \text{ and } \pi_{hj} \neq 1 \\ \pi_{hi}\pi_{hj} & \text{, when } i \neq j \text{ and } \pi_{hi} \text{ or } \pi_{hj} = 1\,. \end{cases}$$

In a similar way as we derived the estimator (8) of $\mathrm{V}\left(\overline{y}_{s_h}\right)$, we have

$$\text{(9)} \qquad \hat{\mathrm{V}}_{\mathrm{SYG}}\left(\overline{x}_{s_h}\right) = \frac{1}{2}\frac{1}{n_h^2}\sum_{i\in s_h}\sum_{\substack{j\in s_h \\ j\neq i}} \frac{\pi_{hi}\pi_{hj} - \hat{\pi}_{hij}}{\hat{\pi}_{hij}}\left(x_{hi} - x_{hj}\right)^2$$

and

$$\text{(10)} \qquad \hat{\mathrm{C}}_{\mathrm{SYG}}\left(\overline{x}_{s_h}, \overline{y}_{s_h}\right) = \frac{1}{2}\frac{1}{n_h^2}\sum_{i\in s_h}\sum_{\substack{j\in s_h \\ j\neq i}} \frac{\pi_{hi}\pi_{hj} - \hat{\pi}_{hij}}{\hat{\pi}_{hij}}\left(x_{hi} - x_{hj}\right)\left(y_{hi} - y_{hj}\right)$$

of $\mathrm{V}\left(\overline{x}_{s_h}\right)$ and $\mathrm{C}\left(\overline{x}_{s_h}, \overline{y}_{s_h}\right)$ respectively. The variance estimator of $\hat{Y}_h$ becomes

$$\text{(11)} \qquad \hat{\mathrm{V}}\left(\hat{Y}_h\right) = \frac{1}{\left(\overline{x}_{s_h}\right)^2}\cdot X_h^2 \cdot \hat{\mathrm{V}}_{\mathrm{SYG}}\left(\overline{y}_{s_h}\right) + \frac{\left(\overline{y}_{s_h}\right)^2}{\left(\overline{x}_{s_h}\right)^4}\cdot X_h^2 \cdot \hat{\mathrm{V}}_{\mathrm{SYG}}\left(\overline{x}_{s_h}\right) - 2\frac{\overline{y}_{s_h}}{\left(\overline{x}_{s_h}\right)^3}\cdot X_h^2 \cdot \hat{\mathrm{C}}_{\mathrm{SYG}}\left(\overline{x}_{s_h}, \overline{y}_{s_h}\right),$$

where $\hat{\mathrm{V}}_{\mathrm{SYG}}\left(\overline{y}_{s_h}\right)$, $\hat{\mathrm{V}}_{\mathrm{SYG}}\left(\overline{x}_{s_h}\right)$ and $\hat{\mathrm{C}}_{\mathrm{SYG}}\left(\overline{x}_{s_h}, \overline{y}_{s_h}\right)$ are given by (8), (9) and (10) respectively. It can be shown that this estimator always is positive.

The estimator (11) applies to strata where $n_h > 1$. When $n_h = 1$ we will estimate the variance with the upper bound $X_h^2/4$ as previously mention. So, since $\mathrm{V}\left(\hat{Y}_h\right) = 0$ when $n_h = N_h^*$, we get the following estimator of the standard error (4):

$$\text{(12)} \qquad \hat{\mathrm{s}}.\mathrm{e}.\left(\hat{d}\right) = \frac{\sqrt{\sum_h \tilde{\mathrm{V}}\left(\hat{Y}_h\right)}}{\sum_h X_h}\cdot 100$$

where

$$\tilde{\mathrm{V}}\left(\hat{Y}_h\right) = \begin{cases} 0 & \text{, when } n_h = N_h^* \\ \dfrac{1}{4}X_h^2 & \text{, when } n_h = 1 < N_h^* \\ \hat{\mathrm{V}}\left(\hat{Y}_h\right) & \text{, when } 1 < n_h < N_h^* \end{cases}$$

and $\hat{\mathrm{V}}\left(\hat{Y}_h\right)$ is given by (11).

We have seen that $\hat{Y}_h$ is, in general, not unbiased when $n_h = 1$, and are now going to show that this is the case when $n_h > 1$ as well. We have no explicit expression of the expectation when $n_h > 1$, but using the approximation (7) gives that

$$\begin{aligned} \mathrm{E}\hat{Y}_h &\approx \frac{\mathrm{E}\overline{y}_{s_h}}{\mathrm{E}\overline{x}_{s_h}} \cdot X_h \\ &= \frac{\sum_{i \in U_h^*} y_{hi}\pi_{hi}}{\sum_{i \in U_h^*} x_{hi}\pi_{hi}} \cdot X_h \ . \end{aligned} \tag{13}$$

This approximation is generally not equal to Y_h. For most values of x_{hi}'s, there exist configurations of the y_{hi}'s that makes the approximation either larger or smaller than Y_h. E.g., assume there is a stratum where $N_h = 5$, $n_h = 2$ and the x_{hi}'s are 100, 120, 130, 130 and 170. If the respective values of the y_{hi}'s are 0, 120/2, 130/2, 130/2 and 170, then $\left(\mathrm{E}\overline{y}_{s_h} / \mathrm{E}\overline{x}_{s_h}\right) \cdot X_h \approx 396$ while $Y_h = 360$. That is, the approximate expectation is larger than Y_h. If, on the other hand, the respective values of the y_{hi}'s are 100, 120/2, 130/2, 130/2 and 0, then $\left(\mathrm{E}\overline{y}_{s_h} / \mathrm{E}\overline{x}_{s_h}\right) \cdot X_h \approx 255$ while $Y_h = 290$. That is, the approximate expectation is now smaller than Y_h. Thus, assuming that the approximation (13) is good enough, $\hat{Y}_h$ is generally not unbiased.

Since $\hat{Y}_h$ is not unbiased, the estimator $\hat{d}$ is not unbiased. This means that the bias of $\hat{d}$, $\mathrm{Bias}\left(\hat{d}\right)$, is not equal to 0 for all possible configurations of the y_{hi}'s. We will now derive upper and lower bounds of the bias.

It can be shown that

$$\mathrm{Bias}\left(\hat{d}\right) = \frac{\sum_h \mathrm{Bias}\left(\hat{Y}_h\right)}{\sum_h X_h} \cdot 100 \, ,$$

where $\mathrm{Bias}\left(\hat{Y}_h\right) = \mathrm{E}\left[\hat{Y}_h - Y_h\right]$ is the bias of $\hat{Y}_h$. For strata where $n_h = 1$ we have from (5) that

$$\mathrm{Bias}\left(\hat{Y}_h\right) = \sum_{i \in U_h} c_{hi} y_{hi} \, ,$$

where $c_{hi} = \sum_{i \notin U_h^*} x_{hi} \Big/ \sum_{i \in U_h^*} x_{hi}$ when $i \in U_h^*$ and -1 when $i \notin U_h^*$. (When $U_h^* = U_h$ we get $\mathrm{Bias}\left(\hat{Y}_h\right) = 0$ as we should). For strata where $n_h > 1$ we make use of the approximation (13) and find

$$\begin{aligned} \mathrm{Bias}\left(\hat{Y}_h\right) &\approx \frac{\sum_{i \in U_h^*} y_{hi}\pi_{hi}}{\sum_{i \in U_h^*} x_{hi}\pi_{hi}} \cdot X_h - Y_h \\ &= \sum_{i \in U_h} b_{hi} y_{hi} \, , \end{aligned}$$

where $b_{hi} = \dfrac{\pi_{hi} X_h}{\sum_{i \in U_h^*} x_{hj} \pi_{hj}} - 1$. Since $\text{Bias}\left(\hat{Y}_h\right) = 0$ when $n_h = N_h$ we have

$$\text{(14)} \qquad \text{Bias}\left(\hat{d}\right) \approx \frac{\sum_{\{h:\, 1<n_h<N_h\}} \left(\sum_{i \in U_h} b_{hi} y_{hi} \right) + \sum_{\{h:\, n_h=1<N_h\}} \left(\sum_{i \in U_h} c_{hi} y_{hi} \right)}{\sum_h X_h} \cdot 100.$$

By using $0 \le b_{hi} y_{hi} \le b_{hi} x_{hi}$ when $b_{hi} \ge 0$, $b_{hi} x_{hi} \le b_{hi} y_{hi} \le 0$ when $b_{hi} \le 0$, and $-\sum_{i \notin U_h^*} x_{hi} \le \sum_{i \in U_h} c_{hi} y_{hi} \le \sum_{i \notin U_h^*} x_{hi}$, we get the inequality

$$\text{L} \le \frac{\sum_{\{h:\, 1<n_h<N_h\}} \left(\sum_{i \in U_h} b_{hi} y_{hi} \right) + \sum_{\{h:\, n_h=1<N_h\}} \left(\sum_{i \in U_h} c_{hi} y_{hi} \right)}{\sum_h X_h} \cdot 100 \le \text{U}$$

where

$$\text{L} = \frac{\sum_{\{h:\, 1<n_h<N_h\}} \left(\sum_{i \in A_h^c} b_{hi} x_{hi} \right) - \sum_{\{h:\, n_h=1<N_h\}} \left(\sum_{i \notin U_h^*} x_{hi} \right)}{\sum_h X_h} \cdot 100 ,$$

$$\text{U} = \frac{\sum_{\{h:\, 1<n_h<N_h\}} \left(\sum_{i \in A_h} b_{hi} x_{hi} \right) + \sum_{\{h:\, n_h=1<N_h\}} \left(\sum_{i \notin U_h^*} x_{hi} \right)}{\sum_h X_h} \cdot 100 ,$$

$A_h = \{i:\ b_{hi} \ge 0\}$ and $A_h^c = \{i:\ b_{hi} < 0\}$. Under the assumption that the approximation (13) is good enough we can, therefore, use L as a lower bound and U as an upper bound of the bias of $\hat{d}$, that is

$$\text{L} \le \text{Bias}\left(\hat{d}\right) \le \text{U} .$$

We note that the bounds L and U depend on the employment of the branch units and the n_h's. Thus the bounds may differ form one quarter to another if the population or the n_h's change. We emphasize that the bias depends on the actual values of the y_{hi}'s. Even though the bounds L and U turn out to be large, the bias may be small.

If the actual values of the y_{hi}'s are so that the bias of $\hat{d}$ is approximately 0, then we can derive a confidence interval of d. To do this we use that

$$\frac{\hat{d}-d}{\text{s.e.}\left(\hat{d}\right)} \approx \frac{\sum_h \left\{\left(\hat{Y}_h - Y_h\right) - \text{E}\left[\hat{Y}_h - Y_h\right]\right\}}{\sqrt{\sum_h \text{V}\left(\hat{Y}_h - Y_h\right)}}$$

when $\text{Bias}\left(\hat{d}\right) \approx 0$. Thus under some conditions on the variance $\text{V}\left(\hat{Y}_h - Y_h\right)$ we get via the Lindeberg theorem that $\left(\hat{d}-d\right)/\text{s.e.}\left(\hat{d}\right)$ is approximately standard normal. An approximate 95% confidence interval of d is therefore given by $\hat{d} \pm 1.96 \cdot \text{s.e.}\left(\hat{d}\right)$. Since $\text{s.e.}\left(\hat{d}\right)$ is unknown it is estimated by $\hat{\text{s}}.\text{e.}\left(\hat{d}\right)$, and the interval

(15) $\quad \hat{d} \pm 1.96 \cdot \hat{\text{s}}.\text{e.}\left(\hat{d}\right)$

is used as a 95% confidence interval for d.

5.2. Bootstrap

The basic idea of bootstrap is to resample a lot of samples form the original sample. The new estimates based on these samples are then used to estimate the variance of the original estimator. The clue is to choose a resembling method so that the bootstrap variance estimator converges to a desired estimator (when the number of resamples approaches infinity).

We are interested in a bootstrap method to estimate the variance of $\hat{Y}_h = \left(\bar{y}_{s_h} / \bar{x}_{s_h}\right) \cdot X_h$. There exist some methods for this type of estimator, but they are all based on a simple random sample. If we use one of these methods, we probably get an estimator that is not suited to estimate the variance of $\hat{Y}_h$. It also exist some methods for unequal probability sampling (Rao and Wu, 1988, and Sitter, 1992). But these methods are not derived for the type of estimator we have.

6. Model-based analysis of the estimator

This section presents a model-based analysis of $\hat{d}$. It means that we consider the tendency variables of the branch units as random variables, and the analysis is done conditional on the observed sample. Since the y_{hi}'s are random, then the Y_h's and d are random as well. This means that we shall predict the value of a random variable, and so $\hat{Y}_h$ and $\hat{d}$ are often referred to as predictors. As in the previous section we assume a sample from each stratum, and take no account of non-response, measurement error or coverage error.

Since the sample usually is selected randomly, it might seem strange to treat the sample not randomly and instead assume a population model. But to assume a population model is not uncommon in surveys. The choice of sample design and estimator is very often based on assumptions on how the variable of interest is distributed in the population. In addition, in the Norwegian Business Tendency Survey the same sample is used for four quarters (or even longer). For each quarter it is observed new values of the y_{hi}'s in the sample. For these reasons it is meaningful to derive an analysis where we treat the sample as given and the y_{hi}'s as random variables. Another argument for the model-based analysis is the Likelihood Principle. The Likelihood Principle points out that two proportional

likelihood functions shall give the same statistical analysis, and leads to that the population must be modelled to do a informative analysis. (For more on the Likelihood Principle, see Bjørnstad, 1995).

The variable y_{hi} is now a random variable that can take the values x_{hi}, $x_{hi}/2$ and 0 with probabilities p_s, p_u and p_m, respectively. That is, p_s is the probability that the branch unit expects its production to increase ($y_{hi} = x_{hi}$), p_u is the probability that the branch unit expects its production to be unchanged ($y_{hi} = x_{hi}/2$), and p_m is the probability that the branch unit expects its production to decrease ($y_{hi} = 0$). With this notation the expectation and variance of y_{hi} can be expressed as

$$\mathrm{E}y_{hi} = \beta x_{hi}$$

and

$$\mathrm{V}(y_{hi}) = \sigma^2 x_{hi}^2,$$

where $\beta = p_s + (1/2)p_u$ and $\sigma^2 = p_s + (1/4)p_u - \left(p_s + (1/2)p_u\right)^2$.

The values of p_s, p_u and p_m may vary with the branch unit. But since branch units within the same stratum are in the same branch, it is not unreasonable to assume that these probabilities are approximately the same for all branch units in the stratum. Hence, β and σ^2 are the same for all branch units in the stratum, say β_h and σ_h^2 for stratum h. The model, denoted by ξ, is therefore given by

$$\mathrm{E}y_{hi} = \beta_h x_{hi} \quad , \forall \ i \in \text{stratum } h$$

$$\mathrm{V}(y_{hi}) = \sigma_h^2 x_{hi}^2 \ , \forall \ i \in \text{stratum } h .$$

In the rest of this section we shall assume this model. We shall also assume that the y_{hi}'s are independent of each other.

Some branch units have employment $x_{hi} = 0$, that is $y_{hi} = x_{hi} = 0$. This means that $p_s = p_u = p_m = 1$, and these probabilities are not equal to the probabilities for the other branch units in the stratum. Thus, the assumptions that led to the model ξ dose not hold for branch units with $x_{hi} = 0$. But since $\mathrm{E}y_{hi} = 0 = \beta_h x_{hi}$ and $\mathrm{V}(y_{hi}) = 0 = \sigma_h^2 x_{hi}^2$ when $x_{hi} = 0$, the model is still valid for these branch units.

The fraction $\bar{y}_{s_h} / \bar{x}_{s_h}$ can be interpreted as an estimator of β_h, and so we use the notation

$$\hat{\beta}_h = \frac{\bar{y}_{s_h}}{\bar{x}_{s_h}},$$

and get

$$\hat{Y}_h = \hat{\beta}_h X_h .$$

By using that the sample s_h is given, we find that

$$\mathrm{E}\hat{\beta}_h = \beta_h$$

so that

$$\mathrm{E}\left[\hat{Y}_h - Y_h\right] = \beta_h X_h - \beta_h X_h = 0$$

and

$$\begin{aligned}\mathrm{E}\left[\hat{d} - d\right] &= \mathrm{E}\left(\frac{\sum_h \left(\hat{Y}_h - Y_h\right)}{\sum_h X_h} \cdot 100\right) \\ &= \frac{\sum_h \mathrm{E}\left[\hat{Y}_h - Y_h\right]}{\sum_h X_h} \cdot 100 = 0\ .\end{aligned}$$

Since this is valid for all s_h, β_h and σ_h, $\hat{Y}_h$ and $\hat{d}$ are unbiased.

Since we are doing a model-based analysis, we are interested in the prediction variance $\mathrm{V}\left(\hat{d} - d\right)$ instead of the variance $\mathrm{V}\left(\hat{d}\right)$ ($\hat{d} - d$ is called the prediction error). By using the independence between the tendency variables we find that

$$\begin{aligned}\mathrm{V}\left(\hat{d} - d\right) &= \mathrm{V}\left(\frac{\sum_h \left(\hat{Y}_h - Y_h\right)}{\sum_h X_h} \cdot 100\right) \\ &= \frac{\sum_h \mathrm{V}\left(\hat{Y}_h - Y_h\right)}{\left(\sum_h X_h\right)^2} \cdot 100^2.\end{aligned}$$

The uncertainty of $\hat{d}$ is now measured by the standard error

$$\begin{aligned}\mathrm{s.e.}\left(\hat{d} - d\right) &= \sqrt{\mathrm{V}\left(\hat{d} - d\right)} \\ &= \frac{\sqrt{\sum_h \mathrm{V}\left(\hat{Y}_h - Y_h\right)}}{\sum_h X_h} \cdot 100\ .\end{aligned} \tag{16}$$

Two arguments for choosing the prediction variance instead of the variance when measuring the uncertainty of a predictor are: 1) Assume that we are able to predict d exactly, that is, we have a predictor $\tilde{d} = d$. Then the measure of uncertainty should be zero since d is predicted exactly. But if we measure the uncertainty with the variance of $\tilde{d}$ ($\mathrm{V}\left(\tilde{d}\right) = \mathrm{V}\left(d\right)$), we have a measure that indicates large uncertainty if the variance of d is large. If we instead measure the uncertainty with the prediction variance ($\mathrm{V}\left(\tilde{d} - d\right) = \mathrm{V}\left(0\right) = 0$), we have a measure that indicates that it is no uncertainty, as we wish. 2) Now, assume we know the expectation of d. If d is predicted with this expectation,

that is $\tilde{d} = \mathrm{E}d$, the measure of uncertainty should indicate large uncertainty if the variance of d is large. This is satisfied with the prediction variance of $\tilde{d}$ ($\mathrm{V}\left(\tilde{d} - d\right) = \mathrm{V}\left(d\right)$), while the variance of $\tilde{d}$ always is zero.

We estimate the standard error (16) by estimating the prediction variance $\mathrm{V}\left(\hat{Y}_h - Y_h\right)$. Since $\hat{Y}_h - Y_h = \hat{\beta}_h \sum_{i \in r_h} x_{hi} - \sum_{i \in r_h} y_{hi}$, we find that

$$(17) \qquad \mathrm{V}\left(\hat{Y}_h - Y_h\right) = \sigma_h^2 \left\{ \left(\frac{\sum_{i \in r_h} x_{hi}}{\sum_{i \in s_h} x_{hi}} \right)^2 \sum_{i \in s_h} x_{hi}^2 + \sum_{i \in r_h} x_{hi}^2 \right\},$$

where r_h is the number of non-sampled branch units in stratum h.

From regression theory we have that

$$\hat{\sigma}_h^2 = \frac{1}{n_h - 1} \sum_{i \in s_h} \frac{1}{x_{hi}^2} \left(y_{hi} - \tilde{\beta}_h x_{hi} \right)^2 ,$$

where

$$\tilde{\beta}_h = \frac{1}{n_h} \sum_{i \in s_h} \frac{y_{hi}}{x_{hi}},$$

is an unbiased estimator of σ_h^2. Hence,

$$\hat{\sigma}_h^2 \left\{ \left(\frac{\sum_{i \in r_h} x_{hi}}{\sum_{i \in s_h} x_{hi}} \right)^2 \sum_{i \in s_h} x_{hi}^2 + \sum_{i \in r_h} x_{hi}^2 \right\}$$

is an unbiased estimator of (17) (provided that $n_h > 1$).

When $n_h = 1$, the prediction variance becomes

$$\mathrm{V}\left(\hat{Y}_h - Y_h\right) = \sigma_h^2 \left\{ \left(\sum_{i \in r_h} x_{hi} \right)^2 + \sum_{i \in r_h} x_{hi}^2 \right\},$$

and is estimated with

$$\tilde{\sigma}_h^2 \left\{ \left(\sum_{i \in r_h} x_{hi} \right)^2 + \sum_{i \in r_h} x_{hi}^2 \right\}$$

where

$$\tilde{\sigma}_h^2 = \frac{1}{n_g - |g|} \sum_{l \in g} \sum_{i \in s_l} \frac{1}{x_{li}^2} \left(y_{li} - \tilde{\beta}_l x_{li} \right)^2 .$$

Here, g denotes the group of strata that have the same employment interval as stratum h, n_g is the number of selected branch units in this group, and $|g|$ is the number of strata in the group. (The estimator $\tilde{\sigma}_h^2$ is an unbiased estimator of σ_h^2 if all σ_l^2, $l \in g$, are equal).

The estimator of the standard error (16) is

$$\text{(18)} \qquad \hat{\text{s}}.\text{e}.\left(\hat{d} - d \right) = \frac{\sqrt{\sum_h \hat{\text{V}} \left(\hat{Y}_h - Y_h \right)}}{\sum_h X_h} \cdot 100$$

where

$$\text{(19)} \qquad \hat{\text{V}} \left(\hat{Y}_h - Y_h \right) = \begin{cases} \hat{\sigma}_h^2 \left\{ \left(\dfrac{\sum_{i \in r_h} x_{hi}}{\sum_{i \in s_h} x_{hi}} \right)^2 \sum_{i \in s_h} x_{hi}^2 + \sum_{i \in r_h} x_{hi}^2 \right\}, & \text{if } n_h > 1 \\ \tilde{\sigma}_h^2 \left\{ \left(\sum_{i \in r_h} x_{hi} \right)^2 + \sum_{i \in r_h} x_{hi}^2 \right\} & , \quad \text{if } n_h = 1 . \end{cases}$$

In addition to estimating d with the estimator $\hat{d}$ we can give a confidence interval for d. We have that

$$\frac{\hat{d} - d}{\text{s.e.}\left(\hat{d} - d \right)} = \frac{\sum_h \left(\hat{Y}_h - Y_h \right)}{\sqrt{\sum_h \text{V} \left(\hat{Y}_h - Y_h \right)}} .$$

Under certain conditions on the variance we get from the Lindeberg central limit theorem that $\left(\hat{d} - d \right) \big/ \text{s.e.}\left(\hat{d} - d \right)$ is approximately standard normal. An approximate 95% confidence interval of d is therefore given by $\hat{d} \pm 1.96 \cdot \text{s.e.}\left(\hat{d} - d \right)$. Since $\text{s.e.}\left(\hat{d} - d \right)$ is unknown it is estimated by $\hat{\text{s}}.\text{e}.\left(\hat{d} - d \right)$, and the interval

$$\text{(20)} \qquad \hat{d} \pm 1.96 \cdot \hat{\text{s}}.\text{e}.\left(\hat{d} - d \right)$$

is used as a 95% confidence interval for d.

7. Illustration

In this section we shall use the uncertainty measures from sections 5 and 6 to estimate the uncertainty of $\hat{d}$ in 1999, 2000 and 2002. The uncertainty measure from section 5 is the design-based standard error $\hat{\text{s}}.\text{e}.\left(\hat{d}\right)$, given by (12), and the uncertainty measure from section 6 is the model-based standard error $\hat{\text{s}}.\text{e}.\left(\hat{d}-d\right)$, given by (18).

In 1999 we have data from 2., 3. and 4. quarter, in 2000 we have data from all four quarters, and in 2002 we have data from 1. and 2. quarter. Since the population data from 1999 and 2000 are on an aggregate level, and we need population data on micro level to calculate the model-based measure $\hat{\text{s}}.\text{e}.\left(\hat{d}-d\right)$, this measure is only calculated for 2002.

In 1999 and 2000 the sample size is about 720. The non-response in this period varies from 125 to 158 branch units. In 2002 the sample size is about 660, and the non-response is 84 branch units in 1. quarter and 97 branch units in 2. quarter. Hence, the non-response is about 19% in 1999 and 2000, and about 14% in 2002.

The population consists of about 30000 branch units, spread over about 270 strata. For about 50 of these strata, the sample equals the total stratum. These are the strata with branch units that have 300 or more employees. For the other strata, a proportional allocation is used to decide the sample size of the strata. As mentioned before, this allocation does not ensure that we get a sample from each stratum, and we have about 70 strata with no data. If we in addition count the strata where all branch units are non-response, we have about 85 strata with no data.

When the diffusion index $\hat{d}$ and the standard errors $\hat{\text{s}}.\text{e}.\left(\hat{d}\right)$ and $\hat{\text{s}}.\text{e}.\left(\hat{d}-d\right)$ are calculated, we have removed from the population strata that have no sample. Properly speaking this means that $\hat{d}$ is an estimator of the diffusion index of the domain defined by the strata with samples. It also means that $\hat{\text{s}}.\text{e}.\left(\hat{d}\right)$ and $\hat{\text{s}}.\text{e}.\left(\hat{d}-d\right)$ are measures of the uncertainty of this estimator, and not of the diffusion index of the entire population.

In addition to estimating the diffusion index of the entire population, we have estimated the diffusion index of three domains. The domains are called E1, E2 and E5. The domain E1 comprises intermediate goods, E2 comprises capital goods and E5 comprises consumer goods.

The values of $\hat{d}$, $\hat{\text{s}}.\text{e}.\left(\hat{d}\right)$ and the design-based interval (15) are given in Table 1.[4] If we compare the estimates of d from one quarter to another, we see that the estimates have become larger from second to third quarter 1999, smaller from third to fourth quarter 1999, and so on throughout 2000. The only exception is domain E2 where the estimate has become larger from third to fourth quarter 2000 while the remaining estimates have become smaller. We also see that all estimates of d for domain E1 are larger than 50. This indicates that the majority of branch units in domain E1 have expected growth in its production in this period.

[4] We do not know if these intervals can be regarded as 95% confidence interval, since we do not know if the values of the y_{hi}'s in the population are so that the design-bias of $\hat{d}$ is approximately 0.

The estimates of the design-based standard error $\text{s.e.}\left(\hat{d}\right)$ (Table 1) are smallest for the entire population and largest for the domain E2. For the entire population, $\hat{\text{s}}.\text{e.}\left(\hat{d}\right)$ varies from 1.43 to 1.56. This indicates small uncertainty when the diffusion index of the population is estimated. For domain E2, $\hat{\text{s}}.\text{e.}\left(\hat{d}\right)$ varies from 3.02 to 3.64 and indicates some uncertainty. For domain E1 and E5, $\hat{\text{s}}.\text{e.}\left(\hat{d}\right)$ varies around 2.40 and 2.25, respectively.

The estimates of the model-based standard error $\text{s.e.}\left(\hat{d}-d\right)$ and the model-based confidence interval (20) are given in Table 2. The figures of $\hat{\text{s}}.\text{e.}\left(\hat{d}-d\right)$ for the population and the domains E1, E2 and E5 are respectively 1.17, 1.91, 2.63 and 1.69 in 1. quarter 2002, and 1.28, 1.81, 2.79 and 2.17 in 2. quarter 2002. Hence, the estimated uncertainty is smallest for the entire population and largest for domain E2.

The design-based and the model-based standard error measure different kinds of uncertainty. The design-based standard error measures the uncertainty we have since a lot of samples that can be selected, while the model-based standard error measures the uncertainty coming from the assumption that the y_{hi}'s are random variables that can take different values. If we compare the estimates, we see that the estimates of $\text{s.e.}\left(\hat{d}-d\right)$ are smaller than the estimates of $\text{s.e.}\left(\hat{d}\right)$. This could mean that the uncertainty measured with $\text{s.e.}\left(\hat{d}-d\right)$ is smaller than the uncertainty measured with $\text{s.e.}\left(\hat{d}\right)$. But since we have used a conservative estimate of $\text{V}\left(\hat{Y}_h\right)$ ($n_h=1$) when $\text{s.e.}\left(\hat{d}\right)$ is estimated, it is also possible that the two measures are of the same size.

As mentioned, $\hat{d}$ is not design-unbiased. To see how large the bias can be, we have calculated the lower and upper bound from subsection 5.1. We got the lower bound -19.23 and the upper bound 19.23 for 1. and 2. quarter 2002.[5] (Since we need population data on micro level to calculate the bounds, they are only calculated for 2002). These bounds are huge, but it does not necessarily mean that the bias is large. It is possible that the y_{hi}'s are so that the bias is small.

To get an idea of the size of the bias with different configurations of the y_{hi}'s, we have simulated 10000 independent random configurations. For each of the configurations we have calculated the bias for 1. quarter 2002. This resulted in 10000 biases in the interval -1.664 to 1.656. Only 106 of the biases were larger than 1 or smaller than -1, while 8030 were in the interval -0.5 to 0.5. This indicates that only a small proportion of all possible configurations of the y_{hi}'s gives a large bias. If the actually y_{hi}'s are not such a configuration the bias is small.

[5] It is no coincidence that the absolute values of the bounds are equal. Using the notation from subsection 5.1 we have that $\sum_{i\in U_h} b_{hi}x_{hi}=0$. This gives $\sum_{i\in A_h^c} b_{hi}x_{hi}=-\sum_{i\in A_h} b_{hi}x_{hi}$, and hence $\text{L}=-\text{U}$.

Table 1: Design-based standard error

2. QUARTER 1999

Aggregated level	Estimated diffusion index	Estimated standard error	Lower bound, 95% confidence interval	Upper bound, 95% confidence interval
Population	47.73	1.59	44.61	50.85
Domain E1	52.40	2.47	47.56	57.24
Domain E2	38.86	3.64	31.73	46.00
Domain E5	49.16	2.40	44.46	53.87

3. QUARTER 1999

Aggregated level	Estimated diffusion index	Estimated standard error	Lower bound, 95% confidence interval	Upper bound, 95% confidence interval
Population	55.03	1.54	52.01	58.05
Domain E1	56.07	2.48	51.22	60.93
Domain E2	39.07	3.43	32.35	45.79
Domain E5	66.04	2.29	61.55	70.53

4. QUARTER 1999

Aggregated level	Estimated diffusion index	Estimated standard error	Lower bound, 95% confidence interval	Upper bound, 95% confidence interval
Population	42.99	1.44	40.18	45.80
Domain E1	51.01	2.33	46.44	55.59
Domain E2	38.30	3.13	32.17	44.44
Domain E5	38.07	2.17	33.82	42.32

1. QUARTER 2000

Aggregated level	Estimated diffusion index	Estimated standard error	Lower bound, 95% confidence interval	Upper bound, 95% confidence interval
Population	63.40	1.52	60.43	66.37
Domain E1	68.71	2.37	64.07	73.35
Domain E2	52.55	3.53	45.63	59.48
Domain E5	66.20	2.18	61.93	70.47

2. QUARTER 2000

Aggregated level	Estimated diffusion index	Estimated standard error	Lower bound, 95% confidence interval	Upper bound, 95% confidence interval
Population	53.50	1.45	50.65	56.34
Domain E1	59.94	2.49	55.06	64.81
Domain E2	50.28	3.02	44.37	56.20
Domain E5	49.35	2.16	45.11	53.60

3. QUARTER 2000

Aggregated level	Estimated diffusion index	Estimated standard error	Lower bound, 95% confidence interval	Upper bound, 95% confidence interval
Population	60.36	1.53	57.36	63.35
Domain E1	62.41	2.47	57.56	67.26
Domain E2	54.71	3.23	48.38	61.05
Domain E5	62.67	2.39	57.98	67.35

4. QUARTER 2000

Aggregated level	Estimated diffusion index	Estimated standard error	Lower bound, 95% confidence interval	Upper bound, 95% confidence interval
Population	53.59	1.43	50.78	56.40
Domain E1	60.26	2.39	55.58	64.95
Domain E2	59.46	3.14	53.31	65.61
Domain E5	42.13	2.07	38.07	46.19

1. QUARTER 2002

Aggregated level	Estimated diffusion index	Estimated standard error	Lower bound, 95% confidence interval	Upper bound, 95% confidence interval
Population	63.43	1.45	60.59	66.27
Domain E1	65.72	2.34	61.12	70.32
Domain E2	56.82	3.27	50.42	63.23
Domain E5	66.09	2.06	62.05	70.12

2. QUARTER 2002

Aggregated level	Estimated diffusion index	Estimated standard error	Lower bound, 95% confidence interval	Upper bound, 95% confidence interval
Population	49.70	1.43	46.89	52.51
Domain E1	54.73	2.17	50.48	58.98
Domain E2	45.31	3.17	39.09	51.52
Domain E5	47.85	2.29	43.36	52.35

Table 2: Model-based standard error and confidence interval

1. QUARTER 2002

Aggregated level	Estimated diffusion index	Estimated standard error	Lower bound, 95% confidence interval	Upper bound, 95% confidence interval
Population	63.43	1.17	61.13	65.73
Domain E1	65.72	1.91	61.98	69.46
Domain E2	56.82	2.63	51.67	61.98
Domain E5	66.09	1.69	62.78	69.40

2. QUARTER 2002

Aggregated level	Estimated diffusion index	Estimated standard error	Lower bound, 95% confidence interval	Upper bound, 95% confidence interval
Population	49.70	1.28	47.18	52.21
Domain E1	54.73	1.81	51.18	58.28
Domain E2	45.31	2.79	39.84	50.77
Domain E5	47.85	2.17	43.61	52.10

8. An alternative estimator: Based on a corrected Horvitz-Thompson estimator

As we have seen, $\hat{d}$ is not design-unbiased. We shall now introduce an estimator that have a smaller design-bias than $\hat{d}$ for most of the configurations of the y_{hi}'s,

$$\tilde{d} = \frac{\sum_h \tilde{Y}_h}{\sum_h X_h} \cdot 100 ,$$

where

$$\tilde{Y}_h = \sum_{i \in s_h} \frac{y_{hi}}{\pi_{hi}} + \tilde{\beta}_h \sum_{i \notin U_h^*} x_{hi} \qquad \text{and} \qquad \tilde{\beta}_h = \frac{1}{n_h} \sum_{i \in s_h} \frac{y_{hi}}{x_{hi}} .$$

When $U_h^* = U_h$ we have $\tilde{Y}_h = \sum_{i \in s_h} y_{hi} / \pi_{hi}$, that is, the Horvitz-Thompson estimator. When $U_h^* \neq U_h$, $\tilde{Y}_h$ can be read as a corrected Horvitz-Thompson estimator, where we correct with $\tilde{\beta}_h \sum_{i \notin U_h^*} x_{hi}$ to reduce the design-bias.

When $n_h = 1$, we have $\tilde{Y}_h = \left(y_{hi_s} / x_{hi_s} \right) X_h$. Hence, we have $\tilde{Y}_h = \hat{Y}_h$ when $n_h = 1$.

We assume a sample from each stratum, and take no account of non-response, measurement error or coverage error when deriving the analyses of $\tilde{d}$.

8.1. Design-based analysis

We start this subsection by explaining why we choose to estimate Y_h with a corrected Horvitz-Thompson estimator instead of the Horvitz-Thompson estimator.

The Horvitz-Thompson estimator $\sum_{i\in s_h} y_{hi}/\pi_{hi}$ is unbiased when $\pi_{hi}>0$ for all $i\in U_h$. That is, the Horvitz-Thompson estimator is unbiased when $U_h^*=U_h$. But when $U_h^*\neq U_h$ the Horvitz-Thompson estimator is biased since $\pi_{hi}=0$ for some $i\in U_h$. The bias is given by

$$\text{Bias}\left(\sum_{i\in s_h}\frac{y_{hi}}{\pi_{hi}}\right)=-\sum_{i\notin U_h^*} y_{hi}\ ,$$

and it is always negative. This means that the Horvitz-Thompson estimator systematically underestimates Y_h when $U_h^*\neq U_h$. To reduce the bias, the sum $\sum_{i\notin U_h^*} y_{hi}$ is estimated with $\tilde{\beta}_h\sum_{i\notin U_h^*} x_{hi}$ and added to the Horvitz-Thompson estimator. This gives the estimator $\tilde{Y}_h$.

Even though this correction reduces the bias of the Horvitz-Thompson estimator for most configurations of the y_{hi}'s, $\tilde{Y}_h$ is still biased unless $U_h^*=U_h$. The expectation and the bias of $\tilde{Y}_h$ are

$$\text{E}\left[\tilde{Y}_h\right]=\sum_{i\in U_h^*} y_{hi}+\frac{1}{n_h}\sum_{i\in U_h^*}\frac{y_{hi}}{x_{hi}}\pi_{hi}\sum_{i\notin U_h^*} x_{hi}$$

and

$$\begin{aligned}\text{Bias}\left(\tilde{Y}_h\right)&=\frac{1}{n_h}\sum_{i\in U_h^*}\frac{y_{hi}}{x_{hi}}\pi_{hi}\sum_{i\notin U_h^*} x_{hi}-\sum_{i\notin U_h^*} y_{hi}\\&=\sum_{i\in U_h} b_{hi}y_{hi}\ ,\end{aligned}$$

where $b_{hi}=\left(\pi_{hi}/n_h x_{hi}\right)\sum_{i\notin U_h^*} x_{hi}$ when $i\in U_h^*$ and -1 otherwise. When $U_h^*=U_h$ we have $\sum_{i\notin U_h^*} x_{hi}=\sum_{i\notin U_h^*} y_{hi}=0$ so that $\text{E}\tilde{Y}_h=Y_h$ and $\text{Bias}\left(\tilde{Y}_h\right)=0$, as it should. But when $U_h^*\neq U_h$ the bias of $\tilde{Y}_h$ is not equal to 0 for all possible values of the y_{hi}'s, that is, $\tilde{Y}_h$ is not unbiased.

Since $\tilde{Y}_h$ is, in general, not unbiased, $\tilde{d}$ is not unbiased. A simulation study shows, however, that the bias of $\tilde{d}$ is approximately 0, and smaller than the bias of $\hat{d}$, for most of the configurations of the y_{hi}'s (see subsection 8.3).

We shall now derive bounds of the bias of $\tilde{d}$. Since $\tilde{Y}_h=Y_h$ when $n_h=N_h$, we have

$$\mathrm{Bias}\left(\tilde{d}\right)=\frac{\sum_{\{h:\, n_h<N_h\}}\mathrm{Bias}\left(\tilde{Y}_h\right)}{\sum_h X_h}\cdot 100$$
$$=\frac{\sum_{\{h:\, n_h<N_h\}}\sum_{i\in U_h} b_{hi}y_{hi}}{\sum_h X_h}\cdot 100\ .$$

By using $-\sum_{i\notin U_h^*} x_{hi}\le \mathrm{Bias}\left(\tilde{Y}_h\right)\le \sum_{i\notin U_h^*} x_{hi}$ we get

$$\mathrm{L}\le \mathrm{Bias}\left(\tilde{d}\right)\le \mathrm{U}\ ,$$

where

$$\mathrm{L}=\frac{-\sum_{\{h:\, n_h<N_h\}}\sum_{i\notin U_h^*} x_{hi}}{\sum_h X_h}\cdot 100$$

and

$$\mathrm{U}=\frac{\sum_{\{h:\, n_h<N_h\}}\sum_{i\notin U_h^*} x_{hi}}{\sum_h X_h}\cdot 100\ .$$

Hence, L is a lower bound and U is an upper bound of $\mathrm{Bias}\left(\tilde{d}\right)$. As will be shown in subsection 8.3, these bounds are significantly smaller than the bounds of $\mathrm{Bias}\left(\hat{d}\right)$, but they are nevertheless large.

The standard error of $\tilde{d}$ is

$$\mathrm{s.e.}\left(\tilde{d}\right)=\frac{\sqrt{\sum_h \mathrm{V}\left(\tilde{Y}_h\right)}}{\sum_h X_h}\cdot 100\ ,$$

where $\mathrm{V}\left(\tilde{Y}_h\right)$ is the variance of $\tilde{Y}_h$. Since $\tilde{Y}_h=\sum_{i\in s_h}\left(1/\pi_{hi}+\left(\sum_{i\notin U_h^*} x_{hi}\right)/n_h x_{hi}\right)y_{hi}$ we find that

$$\mathrm{V}\left(\tilde{Y}_h\right)=\frac{1}{2}\sum_{i\in U_h^*}\sum_{\substack{j\in U_h^*\\ j\neq i}}\left(\pi_{hi}\pi_{hj}-\pi_{hij}\right)\left(v_{hi}-v_{hj}\right)^2\ ,$$

where $v_{hi}=\left(1/\pi_{hi}+\left(\sum_{i\notin U_h^*} x_{hi}\right)/n_h x_{hi}\right)y_{hi}$.

The standard error of $\tilde{d}$ is estimated by estimating the variance of the $\tilde{Y}_h$'s. When $n_h = 1$ it is impossible to estimate the variance. Then we shall use an upper bound as a conservative estimate. Since $\tilde{Y}_h = \hat{Y}_h$ when $n_h = 1$, we have from section 5 that $\mathrm{V}\left(\tilde{Y}_h\right) \leq X_h^2 / 4$, and this is the bound we shall use.

When $n_h > 1$ the variance is estimated with

$$\text{(21)} \qquad \hat{\mathrm{V}}_{\mathrm{SYG}}\left(\tilde{Y}_h\right) = \frac{1}{2} \sum_{i \in s_h} \sum_{\substack{j \in s_h \\ j \neq i}} \frac{\pi_{hi}\pi_{hj} - \hat{\pi}_{hij}}{\hat{\pi}_{hij}} \left(v_{hi} - v_{hj}\right)^2 ,$$

where

$$\hat{\pi}_{hij} = \begin{cases} \pi_{hi} & \text{, when } i = j \\ \dfrac{N_h^*\left(n_h - 1\right)}{n_h\left(N_h^* - 1\right)} \cdot \pi_{hi}\pi_{hj} & \text{, when } i \neq j,\ \pi_{hi} \neq 1 \text{ and } \pi_{hj} \neq 1 \\ \pi_{hi}\pi_{hj} & \text{, when } i \neq j \text{ og } \pi_{hi} \text{ or } \pi_{hj} = 1 . \end{cases}$$

Hence, the standard error of $\tilde{d}$ is estimated by

$$\text{(22)} \qquad \hat{\text{s}}.\text{e}.\left(\tilde{d}\right) = \frac{\sqrt{\sum_h \hat{\mathrm{V}}\left(\tilde{Y}_h\right)}}{\sum_h X_h} \cdot 100 ,$$

where

$$\hat{\mathrm{V}}\left(\tilde{Y}_h\right) = \begin{cases} 0 & \text{, when } n_h = N_h^* \\ \dfrac{1}{4} X_h^2 & \text{, when } n_h = 1 < N_h^* \\ \hat{\mathrm{V}}_{\mathrm{SYG}}\left(\tilde{Y}_h\right) & \text{, when } 1 < n_h < N_h^* \end{cases}$$

and $\hat{\mathrm{V}}_{\mathrm{SYG}}\left(\tilde{Y}_h\right)$ is given by (21).

If the values of the y_{hi}'s are such that the bias of $\tilde{d}$ is approximately 0, then we can derive an approximate 95% confidence interval of d. This is done in a similar way as in subsection 5.1, and gives the interval

$$\text{(23)} \qquad \tilde{d} \pm 1.96 \cdot \hat{\text{s}}.\text{e}.\left(\tilde{d}\right) .$$

Since the bias of $\tilde{d}$ is smaller than the bias of $\hat{d}$ for most of the values of the y_{hi}'s, we will prefer $\tilde{d}$ to $\hat{d}$ if the standard error of $\tilde{d}$ is smaller than the standard error of $\hat{d}$. Unfortunately we are not able to compare the standard errors. Instead we have compared the estimates of the standard errors, and

find that they are quite similar (subsection 8.3). This indicates that $\text{s.e.}\left(\tilde{d}\right)$ and $\text{s.e.}\left(\hat{d}\right)$ is more or less the same. Hence, it does not seem to be any large gain by replacing $\hat{d}$ with $\tilde{d}$.

8.2. Model-based analysis

This subsection presents a model-based analysis of $\tilde{d}$ under model ξ. That is, the sample is treated as given while we assume that

$$\mathrm{E}y_{hi} = \beta_h x_{hi} \quad , \forall \ i \in \text{stratum } h,$$

$$\mathrm{V}(y_{hi}) = \sigma_h^2 x_{hi}^2 \ , \forall \ i \in \text{stratum } h$$

and that the y_{hi}'s are independent of each other.

From these assumptions we have

$$\mathrm{E}\left[\tilde{Y}_h - Y_h\right] = 0$$

and

$$\mathrm{E}\left[\tilde{d} - d\right] = \frac{\sum_h \mathrm{E}\left[\tilde{Y}_h - Y_h\right]}{\sum_h X_h} \cdot 100 = 0 .$$

That is, $\tilde{Y}_h$ is an unbiased estimator of Y_h and $\tilde{d}$ is an unbiased estimator of d.[6]

As argued in section 6, the uncertainty of a predictor is measured by the standard error of the prediction error. Thus, the uncertainty of $\tilde{d}$ is measured with the standard error of $\tilde{d} - d$,

$$\text{s.e.}\left(\tilde{d} - d\right) = \frac{\sqrt{\sum_h \mathrm{V}\left(\tilde{Y}_h - Y_h\right)}}{\sum_h X_h} \cdot 100 .$$

---

[6] To see that $\mathrm{E}\left[\tilde{Y}_h - Y_h\right] = 0$ for strata where $\pi_{hi} = n_h x_{hi} / \sum_{i \in U_h^*} x_{hi} \ \ \forall \ i \in U_h^*$, we write $\tilde{Y}_h = \left(X_h / n_h\right) \sum_{i \in s_h} y_{hi} / x_{hi}$. From this it follows that $\mathrm{E}\left[\tilde{Y}_h - Y_h\right] = 0$. For strata where $n_h x_{hi} / \sum_{i \in U_h^*} x_{hi} > 1$ for some i, say $i \in s1_h$, we have $\pi_{hi} = 1$ for $i \in s1_h$, and $\pi_{hi} = \left(n_h - \left|s1_h\right|\right) x_{hi} / \sum_{i \in U_h^* \setminus s1_h} x_{hi}$ for $i \in U_h^* \setminus s1_h$. Then we can write $\tilde{Y}_h = \sum_{i \in s1_h} y_{hi} + \left(n_h - \left|s1_h\right|\right)^{-1} \left(\sum_{i \in s_h \setminus s1_h} y_{hi} / x_{hi}\right) \sum_{i \in U_h^* \setminus s1_h} x_{hi} +$ $n_h^{-1} \left(\sum_{i \in s_h} y_{hi} / x_{hi}\right) \sum_{i \notin U_h^*} x_{hi}$ (because $s1_h \subset s_h$). From this it follows that $\mathrm{E}\left[\tilde{Y}_h - Y_h\right] = 0$.

We estimate this standard error by estimating the variance $\mathrm{V}\left(\tilde{Y}_h - Y_h\right)$. By using that

$$\tilde{Y}_h - Y_h = \sum_{i \in s_h} \left(\frac{1}{\pi_{hi}} + \frac{\sum_{i \notin U_h^*} x_{hi}}{n_h x_{hi}} - 1 \right) y_{hi} - \sum_{i \in r_h} y_{hi} \text{ , we find that}$$

$$\mathrm{V}\left(\tilde{Y}_h - Y_h\right) = \sigma_h^2 \left\{ \sum_{i \in s_h} \left(\frac{1}{\pi_{hi}} + \frac{\sum_{i \notin U_h^*} x_{hi}}{n_h x_{hi}} - 1 \right)^2 x_{hi}^2 + \sum_{i \in r_h} x_{hi}^2 \right\}.$$

So we need to estimate σ_h^2. Since we use the same model as in section 6, we can use the estimators from that section. That is, we estimate σ_h^2 with

$$\hat{\sigma}_h^2 = \frac{1}{n_h - 1} \sum_{i \in s_h} \frac{1}{x_{hi}^2} \left(y_{hi} - \tilde{\beta}_h x_{hi} \right)^2$$

when $n_h > 1$, and

$$\tilde{\sigma}_h^2 = \frac{1}{n_g - |g|} \sum_{l \in g} \sum_{i \in s_l} \frac{1}{x_{li}^2} \left(y_{li} - \tilde{\beta}_l x_{li} \right)^2$$

when $n_h = 1$. (Here g denotes the group of strata that have the same employment interval as stratum h, n_g is the number of selected branch units in this group, and $|g|$ is the number of strata in the group). As previous mentioned $\hat{\sigma}_h^2$ is an unbiased estimator of σ_h^2, while $\tilde{\sigma}_h^2$ is unbiased if σ_l^2, $l \in g$, are equal.

The estimator of the standard error becomes

$$\text{(24)} \qquad \hat{\text{s}}.\text{e}.\left(\tilde{d} - d\right) = \frac{\sqrt{\sum_h \hat{\mathrm{V}}\left(\tilde{Y}_h - Y_h\right)}}{\sum_h X_h} \cdot 100 \,,$$

where

$$\hat{\mathrm{V}}\left(\tilde{Y}_h - Y_h\right) = \begin{cases} \hat{\sigma}_h^2 \left\{ \sum_{i \in s_h} \left(\frac{1}{\pi_{hi}} + \frac{\sum_{i \notin U_h^*} x_{hi}}{n_h x_{hi}} - 1 \right)^2 x_{hi}^2 + \sum_{i \in r_h} x_{hi}^2 \right\}, & \text{when } n_h > 1 \\ \tilde{\sigma}_h^2 \left\{ \sum_{i \in s_h} \left(\frac{1}{\pi_{hi}} + \frac{\sum_{i \notin U_h^*} x_{hi}}{n_h x_{hi}} - 1 \right)^2 x_{hi}^2 + \sum_{i \in r_h} x_{hi}^2 \right\}, & \text{when } n_h = 1 \,. \end{cases}$$

An approximate 95% confidence interval of d based on $\tilde{d}$ is given by

$$\text{(25)} \qquad \tilde{d} \pm 1.96 \cdot \hat{\text{s}}.\text{e}.\left(\tilde{d} - d\right).$$

Since both the estimators $\tilde{d}$ and $\hat{d}$ are model-unbiased, we will prefer the estimator that have the smallest standard error of the prediction error. If we do a comparison of $\mathrm{V}\left(\tilde{Y}_h - Y_h\right)$ and $\mathrm{V}\left(\hat{Y}_h - Y_h\right)$, we find that $\mathrm{V}\left(\tilde{Y}_h - Y_h\right)$ can be both smaller and larger than $\mathrm{V}\left(\hat{Y}_h - Y_h\right)$ (depending on the sample and the employment in the population). Hence, it is possible that $\mathrm{s.e.}\left(\tilde{d} - d\right)$ can be both smaller and larger than $\mathrm{s.e.}\left(\hat{d} - d\right)$ (this will also depend on σ_h). Based on the estimates we have for 1. and 2. quarter 2002 (see Table 4), it seem like $\mathrm{s.e.}\left(\tilde{d} - d\right)$ is a slightly smaller than $\mathrm{s.e.}\left(\hat{d} - d\right)$ for these two quarters.

8.3. Illustration

To calculate the estimator $\tilde{d}$ and its uncertainty measures $\hat{\mathrm{s}}.\mathrm{e.}\left(\tilde{d}\right)$ and $\hat{\mathrm{s}}.\mathrm{e.}\left(\tilde{d} - d\right)$, given respectively by (22) and (24), we have used the same data as in section 7. Thus, we can only calculate $\hat{\mathrm{s}}.\mathrm{e.}\left(\tilde{d} - d\right)$ for 2002, since we need population data on micro level to calculate this standard error. The values of $\hat{\mathrm{s}}.\mathrm{e.}\left(\tilde{d}\right)$ and the confidence interval (23) are given in Table 3, while the values of $\hat{\mathrm{s}}.\mathrm{e.}\left(\tilde{d} - d\right)$ and the confidence interval (25) are given in Table 4. (To make the comparison of $\tilde{d}$ and $\hat{d}$ easier, we have given the values of $\hat{d}$, $\hat{\mathrm{s}}.\mathrm{e.}\left(\hat{d}\right)$ and $\hat{\mathrm{s}}.\mathrm{e.}\left(\hat{d} - d\right)$ in brackets after the values of $\tilde{d}$, $\hat{\mathrm{s}}.\mathrm{e.}\left(\tilde{d}\right)$ and $\hat{\mathrm{s}}.\mathrm{e.}\left(\tilde{d} - d\right)$, respectively).

If we compare the values of $\tilde{d}$ and $\hat{d}$ we see that they are relatively equal. Some times $\tilde{d}$ is a little larger than $\hat{d}$, other times a little smaller. But usually $\tilde{d}$ is larger than $\hat{d}$ (28 of the 36 estimates with $\tilde{d}$ are larger than the corresponding estimate with $\hat{d}$). We have the largest difference between the estimates in 4. quarter 2000, where $\tilde{d} = 43.95$ and $\hat{d} = 42.13$ for domain E5.

The values of the design-based standard error $\hat{\mathrm{s}}.\mathrm{e.}\left(\tilde{d}\right)$ are smallest for the population and largest for the domain E2. The estimates vary around 1.45, 2.40, 3.20 and 2.20 for the population and the domains E1, E2 and E5, respectively. Based on these figures we may say that the uncertainty of $\tilde{d}$ is small for the population and large for the domain E2.

The estimates of $\mathrm{s.e.}\left(\tilde{d}\right)$ are almost equal to the estimates of $\mathrm{s.e.}\left(\hat{d}\right)$. Some times $\hat{\mathrm{s}}.\mathrm{e.}\left(\tilde{d}\right)$ is smaller than $\hat{\mathrm{s}}.\mathrm{e.}\left(\hat{d}\right)$, other times larger. This suggests that the design-based standard errors $\mathrm{s.e.}\left(\tilde{d}\right)$ and $\mathrm{s.e.}\left(\hat{d}\right)$ are almost equal.

In Table 4 we see that the values of the model-based standard error $\hat{\mathrm{s}}.\mathrm{e.}\left(\tilde{d} - d\right)$ in 1. quarter 2002 equal 1.11, 1.81, 2.49 and 1.61 for respectively the population, domain E1, E2 and E5. The corresponding estimates in 2. quarter 2002 are 1.23, 1.72, 2.66 and 2.10. Again we have the smallest estimates for the population and the largest for the domain E2.

Comparing the estimates of s.e.$\left(\tilde{d}-d\right)$ with the estimates of s.e.$\left(\hat{d}-d\right)$, we find that ŝ.e.$\left(\tilde{d}-d\right)$ is a little smaller than ŝ.e.$\left(\hat{d}-d\right)$. This could indicate that s.e.$\left(\tilde{d}-d\right)$ is a little smaller than s.e.$\left(\hat{d}-d\right)$ in these two quarters.

To see how large the design-based bias of $\tilde{d}$ can be, we have calculated the bounds L and U from subsection 8.1. We got the lower bound -11.00 and the upper bound 11.00 for 1. and 2. quarter 2002. These bounds are much smaller than the bounds ± 19.23 of Bias$\left(\hat{d}\right)$. The bounds are nevertheless large. Fortunately, a simulation study suggests that only a few of all possible configurations of the y_{hi}'s give a large bias. We have simulated 10000 configurations of the y_{hi}'s, and calculated Bias$\left(\tilde{d}\right)$ for 1. quarter 2002. The resulting 10000 values of the bias all fall in the interval -0.479 to 0.428.

We have also compared Bias$\left(\tilde{d}\right)$ and Bias$\left(\hat{d}\right)$ with randomly chosen configurations of the y_{hi}'s. This showed that $\left|\text{Bias}\left(\tilde{d}\right)\right|$ can be both larger and smaller than $\left|\text{Bias}\left(\hat{d}\right)\right|$, but that $\left|\text{Bias}\left(\tilde{d}\right)\right|$ usually is smaller than $\left|\text{Bias}\left(\hat{d}\right)\right|$. (Since we do not know the exact bias of $\hat{d}$, we have used the approximation (14)).

Table 3: Design-based standard error

The figures in brackets are $\hat{d}$ and ŝ.e.$\left(\hat{d}\right)$ from Table 1.

2. QUARTER 1999

Aggregated level	Estimated diffusion index	Estimated standard error	Lower bound, 95% confidence interval	Upper bound, 95% confidence interval
Population	48.56 (47.73)	1.58 (1.59)	45.47	51.66
Domain E1	52.23 (52.40)	2.48 (2.47)	47.37	57.09
Domain E2	40.36 (38.86)	3.68 (3.64)	33.15	47.58
Domain E5	50.54 (49.16)	2.27 (2.40)	46.09	54.99

3. QUARTER 1999

Aggregated level	Estimated diffusion index	Estimated standard error	Lower bound, 95% confidence interval	Upper bound, 95% confidence interval
Population	55.86 (55.03)	1.51 (1.54)	52.89	58.83
Domain E1	57.20 (56.07)	2.51 (2.48)	52.29	62.12
Domain E2	40.76 (39.07)	3.24 (3.43)	34.41	47.12
Domain E5	65.89 (66.04)	2.28 (2.29)	61.43	70.36

4. QUARTER 1999

Aggregated level	Estimated diffusion index	Estimated standard error	Lower bound, 95% confidence interval	Upper bound, 95% confidence interval
Population	43.93 (42.99)	1.45 (1.44)	41.10	46.77
Domain E1	51.68 (51.01)	2.35 (2.33)	47.07	56.29
Domain E2	39.00 (38.30)	3.14 (3.13)	32.84	45.15
Domain E5	39.50 (38.07)	2.20 (2.17)	35.18	43.82

1. QUARTER 2000

Aggregated level	Estimated diffusion index	Estimated standard error	Lower bound, 95% confidence interval	Upper bound, 95% confidence interval
Population	63.44 (63.40)	1.46 (1.52)	60.58	66.29
Domain E1	69.08 (68.71)	2.36 (2.37)	64.45	73.71
Domain E2	52.78 (52.55)	3.32 (3.53)	46.28	59.28
Domain E5	65.77 (66.20)	2.07 (2.18)	61.71	69.83

2. QUARTER 2000

Aggregated level	Estimated diffusion index	Estimated standard error	Lower bound, 95% confidence interval	Upper bound, 95% confidence interval
Population	54.26 (53.50)	1.48 (1.45)	51.35	57.16
Domain E1	59.73 (59.94)	2.48 (2.49)	54.88	64.58
Domain E2	51.21 (50.28)	3.11 (3.02)	45.11	57.31
Domain E5	50.99 (49.35)	2.26 (2.16)	46.56	55.42

3. QUARTER 2000

Aggregated level	Estimated diffusion index	Estimated standard error	Lower bound, 95% confidence interval	Upper bound, 95% confidence interval
Population	60.60 (60.36)	1.52 (1.53)	57.62	63.59
Domain E1	63.12 (62.41)	2.46 (2.47)	58.31	67.93
Domain E2	54.63 (54.71)	3.26 (3.23)	48.23	61.02
Domain E5	62.71 (62.67)	2.37 (2.39)	58.06	67.36

4. QUARTER 2000

Aggregated level	Estimated diffusion index	Estimated standard error	Lower bound, 95% confidence interval	Upper bound, 95% confidence interval
Population	54.52 (53.59)	1.41 (1.43)	51.76	57.29
Domain E1	60.73 (60.26)	2.37 (2.39)	56.09	65.37
Domain E2	59.88 (59.46)	3.05 (3.14)	53.91	65.86
Domain E5	43.95 (42.13)	2.06 (2.07)	39.90	48.00

1. QUARTER 2002

Aggregated level	Estimated diffusion index	Estimated standard error	Lower bound, 95% confidence interval	Upper bound, 95% confidence interval
Population	63.74 (63.43)	1.32 (1.45)	61.15	66.32
Domain E1	66.01 (65.72)	2.27 (2.34)	61.57	70.46
Domain E2	58.23 (56.82)	2.84 (3.27)	52.66	63.79
Domain E5	65.53 (66.09)	1.84 (2.06)	61.92	69.14

2. QUARTER 2002

Aggregated level	Estimated diffusion index	Estimated standard error	Lower bound, 95% confidence interval	Upper bound, 95% confidence interval
Population	49.63 (49.70)	1.41 (1.43)	46.87	52.38
Domain E1	54.04 (54.73)	2.12 (2.17)	49.89	58.18
Domain E2	45.75 (45.31)	3.15 (3.17)	39.57	51.93
Domain E5	48.04 (47.85)	2.22 (2.29)	43.69	52.38

Table 4: Model-based standard error and confidence interval

The figures in brackets are $\hat{d}$ and ŝ.e.$\left(\hat{d}-d\right)$ from Table 2.

1. QUARTER 2002

Aggregated level	Estimated diffusion index	Estimated standard error	Lower bound, 95% confidence interval	Upper bound, 95% confidence interval
Population	63.74 (63.43)	1.11 (1.17)	61.56	65.92
Domain E1	66.01 (65.72)	1.81 (1.91)	62.47	69.56
Domain E2	58.23 (56.82)	2.49 (2.63)	53.35	63.10
Domain E5	65.53 (66.09)	1.61 (1.69)	62.38	68.68

2. QUARTER 2002

Aggregated level	Estimated diffusion index	Estimated standard error	Lower bound, 95% confidence interval	Upper bound, 95% confidence interval
Population	49.63 (49.70)	1.23 (1.28)	47.22	52.03
Domain E1	54.04 (54.73)	1.72 (1.81)	50.66	57.41
Domain E2	45.75 (45.31)	2.66 (2.79)	40.54	50.96
Domain E5	48.04 (47.85)	2.10 (2.17)	43.93	52.14

9. An alternative estimator: Based on best linear unbiased predictor

The estimator $\tilde{d}$ was motivated from a design-based point of view. We shall now present an estimator that is motivated form a model-based point of view. The estimator is

$$\hat{\hat{d}} = \frac{\sum_h \hat{\hat{Y}}_h}{\sum_h X_h} \cdot 100 ,$$

where

$$\hat{\hat{Y}}_h = \sum_{i \in s_h} y_{hi} + \tilde{\beta}_h \sum_{i \in r_h} x_{hi} \qquad \text{and} \qquad \tilde{\beta}_h = \frac{1}{n_h} \sum_{i \in s_h} \frac{y_{hi}}{x_{hi}} .$$

It can be shown that $\hat{\hat{Y}}_h$ is the best linear unbiased predictor of Y_h, under model ξ (Bjørnstad, 1995). That is, among all linear and unbiased predictors of Y_h, $\hat{\hat{Y}}_h$ is the predictor that has smallest prediction variance.

When $n_h = 1$, we have $\hat{\hat{Y}}_h = \left(y_{hi_s} / x_{hi_s} \right) X_h$. Hence, we have $\hat{\hat{Y}}_h = \hat{Y}_h = \tilde{Y}_h$ when $n_h = 1$.

We assume a sample from each stratum, and take no account of non-response, measurement error or coverage error when deriving the analyses of $\hat{\hat{d}}$.

9.1. Design-based analysis

We are not able to find exact expressions of the design-based expectation and variance of $\hat{\hat{Y}}_h$ (except when $n_h = 1$). For this reason we write $\hat{\hat{Y}}_h$ as

$$\hat{\hat{Y}}_h = \sum_{i \in s_h} \left(1 + \frac{X_h}{n_h x_{hi}} \right) \cdot y_{hi} - n_h \overline{\left(\frac{y}{x} \right)}_{s_h} \bar{x}_{s_h} ,$$

where $\overline{(y/x)}_{s_h} = (1/n_h) \sum_{i \in s_h} y_{hi} / x_{hi}$, and do a first order Taylor expansion of the last term. That is, we do the approximation

$$n_h \overline{\left(\frac{y}{x} \right)}_{s_h} \bar{x}_{s_h} \approx n_h \mathrm{E} \overline{\left(\frac{y}{x} \right)}_{s_h} \mathrm{E}\bar{x}_{s_h} + n_h \mathrm{E}\bar{x}_{s_h} \left(\overline{\left(\frac{y}{x} \right)}_{s_h} - \mathrm{E} \overline{\left(\frac{y}{x} \right)}_{s_h} \right) + n_h \mathrm{E} \overline{\left(\frac{y}{x} \right)}_{s_h} \left(\bar{x}_{s_h} - \mathrm{E}\bar{x}_{s_h} \right) ,$$

and obtain

$$\hat{\hat{Y}}_h \approx \sum_{i \in s_h} \left(1 + \frac{X_h}{n_h x_{hi}} \right) \cdot y_{hi} - n_h \mathrm{E} \overline{\left(\frac{y}{x} \right)}_{s_h} \mathrm{E}\bar{x}_{s_h} - n_h \mathrm{E}\bar{x}_{s_h} \left(\overline{\left(\frac{y}{x} \right)}_{s_h} - \mathrm{E} \overline{\left(\frac{y}{x} \right)}_{s_h} \right) - n_h \mathrm{E} \overline{\left(\frac{y}{x} \right)}_{s_h} \left(\bar{x}_{s_h} - \mathrm{E}\bar{x}_{s_h} \right) . \qquad (26)$$

From this we have

$$\begin{aligned}
\mathrm{E}\hat{\hat{Y}}_h &\approx \mathrm{E}\left[\sum_{i\in s_h}\left(1+\frac{X_h}{n_h x_{hi}}\right)\cdot y_{hi}\right] - n_h \mathrm{E}\left(\overline{\frac{y}{x}}\right)_{s_h} \mathrm{E}\bar{x}_{s_h} \\
&= \sum_{i\in U_h^*}\left(1+\frac{X_h}{n_h x_{hi}}\right) y_{hi}\pi_{hi} - \frac{1}{n_h}\sum_{i\in U_h^*}\frac{y_{hi}}{x_{hi}}\pi_{hi}\sum_{i\in U_h^*} x_{hi}\pi_{hi} \\
&= Y_h + \sum_{i\in U_h}\left(y_{hi} - \frac{Y_h}{X_h}x_{hi}\right)\pi_{hi},
\end{aligned}$$

where the last equality refers to strata where $\pi_{hi} = n_h\left(x_{hi}/X_h\right)$ for all $i \in U_h$. The bias of $\hat{\hat{Y}}_h$ is now approximated with

$$\begin{aligned}
\mathrm{Bias}\left(\hat{\hat{Y}}_h\right) &\approx \sum_{i\in U_h^*}\left(\pi_{hi} + \frac{X_h - \sum_{j\in U_h^*} x_{hj}\pi_{hj}}{n_h x_{hi}}\cdot\pi_{hi} - 1\right) y_{hi} - \sum_{i\notin U_h^*} y_{hi} \\
&= \sum_{i\in U_h} b_{hi} y_{hi}\ ,
\end{aligned} \tag{27}$$

where $b_{hi} = \pi_{hi} + \dfrac{X_h - \sum_{j\in U_h^*} x_{hj}\pi_{hj}}{n_h x_{hi}}\cdot\pi_{hi} - 1$ when $i \in U_h^*$ and -1 elsewhere. This approximation is in general not equal to 0.

When $n_h = 1$ we have an exact expression of $\mathrm{Bias}\left(\hat{\hat{Y}}_h\right)$. From section 5 we have that

$$\begin{aligned}
\mathrm{Bias}\left(\hat{\hat{Y}}_h\right) &= \frac{\sum_{i\in U_h^*} y_{hi}}{\sum_{i\in U_h^*} x_{hi}}\sum_{i\notin U_h^*} x_{hi} - \sum_{i\notin U_h^*} y_{hi} \\
&= \sum_{i\in U_h} c_{hi} y_{hi}\ ,
\end{aligned}$$

where $c_{hi} = \sum_{i\notin U_h^*} x_{hi} \Big/ \sum_{i\in U_h^*} x_{hi}$ when $i \in U_h^*$, and -1 elsewhere (because $\hat{\hat{Y}}_h = \hat{Y}_h$ when $n_h = 1$). This bias is in general not equal to 0, unless $U_h^* = U_h$.

The bias of $\hat{\hat{d}}$ is given by

$$\mathrm{Bias}\left(\hat{\hat{d}}\right) = \frac{\sum_h \mathrm{Bias}\left(\hat{\hat{Y}}_h\right)}{\sum_h X_h}\cdot 100\ .$$

Since $\hat{\hat{Y}}_h$ in general is a biased estimator, $\hat{\hat{d}}$ is a biased estimator. To derive upper and lower bounds of the bias, we use the approximation (27) and find that

$$(28)\qquad \mathrm{Bias}\left(\hat{\hat{d}}\right) \approx \frac{\sum_{\{h:\ 1<n_h<N_h\}}\left(\sum_{i\in U_h} b_{hi}y_{hi}\right) + \sum_{\{h:\ n_h=1<N_h\}}\left(\sum_{i\in U_h} c_{hi}y_{hi}\right)}{\sum_h X_h}\cdot 100\,.$$

By using $0 \le b_{hi}y_{hi} \le b_{hi}x_{hi}$ when $b_{hi} \ge 0$, $b_{hi}x_{hi} \le b_{hi}y_{hi} \le 0$ when $b_{hi} \le 0$, and $-\sum_{i\notin U_h^*} x_{hi} \le \sum_{i\in U_h} c_{hi}y_{hi} \le \sum_{i\notin U_h^*} x_{hi}$, we derive the inequality

$$\mathrm{L} \le \frac{\sum_{\{h:\ 1<n_h<N_h\}}\left(\sum_{i\in U_h} b_{hi}y_{hi}\right) + \sum_{\{h:\ n_h=1<N_h\}}\left(\sum_{i\in U_h} c_{hi}y_{hi}\right)}{\sum_h X_h}\cdot 100 \le \mathrm{U}\,,$$

where

$$\mathrm{L} = \frac{\sum_{\{h:\ 1<n_h<N_h\}}\left(\sum_{i\in A_h^c} b_{hi}x_{hi}\right) - \sum_{\{h:\ n_h=1<N_h\}}\left(\sum_{i\notin U_h^*} x_{hi}\right)}{\sum_h X_h}\cdot 100\,,$$

$$\mathrm{U} = \frac{\sum_{h\ \text{s.a.}\ 1<n_h<N_h}\left(\sum_{i\in A_h} b_{hi}x_{hi}\right) + \sum_{h\ \text{s.a.}\ n_h=1<N_h}\left(\sum_{i\notin U_h^*} x_{hi}\right)}{\sum_h X_h}\cdot 100\,,$$

$A_h = \{i:\ b_{hi} \ge 0\}$ and $A_h^c = \{i:\ b_{hi} < 0\}$. Under the assumption that the approximation (28) is good enough we can, therefore, use L as a lower bound and U as an upper bound of the bias of $\hat{\hat{d}}$, that is

$$\mathrm{L} \le \mathrm{Bias}\left(\hat{\hat{d}}\right) \le \mathrm{U}\,.$$

As will be seen in subsection 9.3, these bounds are large. But again, a simulation study shows that most of the configurations of the y_{hi}'s give a relatively small bias.

From (26) we find that the variance of $\hat{\hat{Y}}_h$ can be approximated with

$$
\begin{aligned}
\mathrm{V}\left(\hat{\hat{Y}}_h\right) &\approx \mathrm{V}\left(\sum_{i\in s_h}\left(1+\frac{X_h}{n_h x_{hi}}\right)\cdot y_{hi} - n_h\left(\mathrm{E}\overline{x}_{s_h}\right)\overline{\left(\frac{y}{x}\right)}_{s_h} - n_h\left(\mathrm{E}\overline{\left(\frac{y}{x}\right)}_{s_h}\right)\overline{x}_{s_h}\right) \\
&= \mathrm{V}\left(\sum_{i\in s_h}\left(y_{hi}+\frac{X_h}{n_h x_{hi}}y_{hi} - \frac{y_{hi}}{x_{hi}}\mathrm{E}\overline{x}_{s_h} - x_{hi}\mathrm{E}\overline{\left(\frac{y}{x}\right)}_{s_h}\right)\right) \\
&= \mathrm{V}\left(\sum_{i\in s_h} z_{hi}\right) \\
&= \frac{1}{2}\sum_{i\in U_h^*}\sum_{\substack{j\in U_h^* \\ j\neq i}}\left(\pi_{hi}\pi_{hj}-\pi_{hij}\right)\left(z_{hi}-z_{hj}\right)^2,
\end{aligned}
$$

where $z_{hi}=y_{hi}+\frac{X_h}{n_h x_{hi}}y_{hi}-\frac{y_{hi}}{x_{hi}}\mathrm{E}\overline{x}_{s_h}-x_{hi}\mathrm{E}\overline{\left(\frac{y}{x}\right)}_{s_h}$. Hence, we estimate $\mathrm{V}\left(\hat{\hat{Y}}_h\right)$ with

$$
(29)\qquad \hat{\mathrm{V}}_{\mathrm{SYG}}\left(\hat{\hat{Y}}_h\right)=\frac{1}{2}\sum_{i\in s_h}\sum_{\substack{j\in s_h \\ j\neq i}}\frac{\pi_{hi}\pi_{hj}-\hat{\pi}_{hij}}{\hat{\pi}_{hij}}\left(\hat{z}_{hi}-\hat{z}_{hj}\right)^2,
$$

where

$$
\hat{\pi}_{hij}=\begin{cases}
\pi_{hi} & \text{, when } i=j \\
\dfrac{N_h^*\left(n_h-1\right)}{n_h\left(N_h^*-1\right)}\cdot\pi_{hi}\pi_{hj} & \text{, when } i\neq j,\ \pi_{hi}\neq 1 \text{ and } \pi_{hj}\neq 1 \\
\pi_{hi}\pi_{hj} & \text{, when } i\neq j \text{ og } \pi_{hi} \text{ or } \pi_{hj}=1
\end{cases}
$$

and $\hat{z}_{hi}=y_{hi}+\frac{X_h}{n_h x_{hi}}y_{hi}-\frac{y_{hi}}{x_{hi}}\overline{x}_{s_h}-x_{hi}\overline{\left(\frac{y}{x}\right)}_{s_h}$. When $n_h=1$ we do as for $\hat{Y}_h$ and estimate the variance with the upper bound $X_h^2/4$.

The standard error of $\hat{\hat{d}}$ is given by

$$
\mathrm{s.e.}\left(\hat{\hat{d}}\right)=\frac{\sqrt{\sum_h \mathrm{V}\left(\hat{\hat{Y}}_h\right)}}{\sum_h X_h}\cdot 100
$$

and estimated by

$$
(30)\qquad \hat{\mathrm{s}}.\mathrm{e}.\left(\hat{\hat{d}}\right)=\frac{\sqrt{\sum_h \hat{\mathrm{V}}\left(\hat{\hat{Y}}_h\right)}}{\sum_h X_h}\cdot 100,
$$

where

$$\hat{\mathrm{V}}\left(\hat{\hat{Y}}_h\right)=\begin{cases}0 & \text{, when } n_h = N_h \\ \frac{1}{4}X_h^2 & \text{, when } n_h = 1 < N_h \\ \hat{\mathrm{V}}_{\mathrm{SYG}}\left(\hat{\hat{Y}}_h\right) & \text{, when } 1 < n_h < N_h\end{cases}$$

and $\hat{\mathrm{V}}_{\mathrm{SYG}}\left(\hat{\hat{Y}}_h\right)$ is given by (29).

An approximate 95% confidence interval of d is

(31) $\quad \hat{\hat{d}} \pm 1.96 \cdot \hat{\mathrm{s}}.\mathrm{e}.\left(\hat{\hat{d}}\right),$

provided that the y_{hi}'s are so that the bias of $\hat{\hat{d}}$ is approximately 0.

To determine if $\hat{\hat{d}}$ is a better estimator than $\hat{d}$, we have to compare both the standard errors and the biases. Unfortunately we are not able to compare the standard errors, but based on the estimates they seem to be quite similar. Regarding the biases, a simulation study shows that $\mathrm{Bias}\left(\hat{\hat{d}}\right)$ can be both smaller and larger than $\mathrm{Bias}\left(\hat{d}\right)$. Therefore, we cannot say that one of the estimators always is better than the other one.

9.2. Model-based analysis

In this subsection we shall again treat the sample as given and assume that the y_{hi}'s are distributed according to model ξ. That is, we assume that

$$\mathrm{E}y_{hi} = \beta_h x_{hi} \quad , \forall \ \ i \in \text{stratum } h,$$

$$\mathrm{V}(y_{hi}) = \sigma_h^2 x_{hi}^2 \ , \forall \ \ i \in \text{stratum } h$$

and that the y_{hi}'s are independent of each other.

From these assumptions we have $\mathrm{E}\left[\hat{\hat{Y}}_h - Y_h\right] = 0$ so that

$$\mathrm{E}\left[\hat{\hat{d}} - d\right] = \frac{\sum_h \mathrm{E}\left[\hat{\hat{Y}}_h - Y_h\right]}{\sum_h X_h} \cdot 100 = 0 \ .$$

That is, $\hat{\hat{d}}$ is unbiased.

By using $\hat{\hat{Y}}_h - Y_h = (1/n_h)\sum_{i\in s_h}(y_{hi}/x_{hi})\sum_{i\in r_h} x_{hi} - \sum_{i\in r_h} y_{hi}$ we find that the prediction variance of $\hat{\hat{Y}}_h$ is

$$\mathrm{V}\left(\hat{\hat{Y}}_h - Y_h\right) = \sigma_h^2\left\{\frac{1}{n_h}\left(\sum_{i\in r_h} x_{hi}\right)^2 + \sum_{i\in r_h} x_{hi}^2\right\}.$$

From this expression it is seen that the prediction variance of $\hat{\hat{Y}}_h$ is smallest when the branch units with the largest employment are sampled.

It can be shown that the prediction variance of $\hat{\hat{Y}}_h$ is less than or equal to the prediction variance of all linear and unbiased predictors of Y_h. Since $\hat{Y}_h$ and $\tilde{Y}_h$ are linear and unbiased this means that $\hat{\hat{Y}}_h$ has a smaller (or equal) prediction variance than $\hat{Y}_h$ and $\tilde{Y}_h$. Hence, the standard error

$$\text{(32)} \qquad \text{s.e.}\left(\hat{\hat{d}} - d\right) = \frac{\sqrt{\sum_h \mathrm{V}\left(\hat{\hat{Y}}_h - Y_h\right)}}{\sum_h X_h}\cdot 100$$

is smaller than the corresponding standard errors of $\hat{d}$ and $\tilde{d}$.

We estimate the standard error (32) by estimating the prediction variance of $\hat{\hat{Y}}_h$, and we estimate the prediction variance of $\hat{\hat{Y}}_h$ by estimating σ_h^2. To estimate σ_h^2 we use the same estimator as in section 6 and subsection 8.2, that is

$$\hat{\sigma}_h^2 = \frac{1}{n_h - 1}\sum_{i\in s_h}\frac{1}{x_{hi}^2}\left(y_{hi} - \tilde{\beta}_h x_{hi}\right)^2$$

when $n_h > 1$, and

$$\tilde{\sigma}_h^2 = \frac{1}{n_g - |g|}\sum_{l\in g}\sum_{i\in s_l}\frac{1}{x_{li}^2}\left(y_{li} - \tilde{\beta}_l x_{li}\right)^2$$

when $n_h = 1$. (Here, g denotes the group of strata that have the same employment interval as stratum h, n_g is the number of selected branch units in this group, and $|g|$ is the number of strata in the group). The estimator of the standard error (32) is therefore

$$\text{(33)} \qquad \hat{\text{s}}\text{.e.}\left(\hat{\hat{d}} - d\right) = \frac{\sqrt{\sum_h \hat{\mathrm{V}}\left(\hat{\hat{Y}}_h - Y_h\right)}}{\sum_h X_h}\cdot 100,$$

where

$$\hat{\mathrm{V}}\left(\hat{\hat{Y}}_h - Y_h\right) = \begin{cases} \hat{\sigma}_h^2 \left\{ \dfrac{1}{n_h} \left(\sum_{i \in r_h} x_{hi} \right)^2 + \sum_{i \in r_h} x_{hi}^2 \right\}, & \text{when } n_h > 1 \\ \tilde{\sigma}_h^2 \left\{ \dfrac{1}{n_h} \left(\sum_{i \in r_h} x_{hi} \right)^2 + \sum_{i \in r_h} x_{hi}^2 \right\}, & \text{when } n_h = 1 \end{cases}$$

is a estimator of $\mathrm{V}\left(\hat{\hat{Y}}_h - Y_h\right)$.

We note that $\hat{\mathrm{V}}\left(\hat{\hat{Y}}_h - Y_h\right)$ is smaller than $\hat{\mathrm{V}}\left(\hat{Y}_h - Y_h\right)$ and $\hat{\mathrm{V}}\left(\tilde{Y}_h - Y_h\right)$. This is because the prediction variance of $\hat{\hat{Y}}_h$ is smaller than the prediction variance of $\hat{Y}_h$ and $\tilde{Y}_h$ for all $\sigma_h^2 > 0$, and that σ_h^2 is estimated with the same estimator in all prediction variances. From this it also follows that $\hat{\mathrm{s.e.}}\left(\hat{\hat{d}} - d\right)$ is smaller than $\hat{\mathrm{s.e.}}\left(\hat{d} - d\right)$ and $\hat{\mathrm{s.e.}}\left(\tilde{d} - d\right)$.

An approximate 95% confidence interval of d based on $\hat{\hat{d}}$ is given by

(34) $$\hat{\hat{d}} \pm 1.96 \cdot \hat{\mathrm{s.e.}}\left(\hat{\hat{d}} - d\right).$$

This interval is narrower than the corresponding intervals based on $\hat{d}$ and $\tilde{d}$, since $\hat{\mathrm{s.e.}}\left(\hat{\hat{d}} - d\right)$ is smaller than $\hat{\mathrm{s.e.}}\left(\hat{d} - d\right)$ and $\hat{\mathrm{s.e.}}\left(\tilde{d} - d\right)$.

We have seen that all the estimators $\hat{d}$, $\tilde{d}$ and $\hat{\hat{d}}$ are model-unbiased with model ξ. Since $\mathrm{s.e.}\left(\hat{\hat{d}} - d\right)$ is smaller than $\mathrm{s.e.}\left(\hat{d} - d\right)$ and $\mathrm{s.e.}\left(\tilde{d} - d\right)$ we may say that $\hat{\hat{d}}$ is a better estimator than $\hat{d}$ and $\tilde{d}$. However, whether or not $\hat{d}$ should be replaced by $\hat{\hat{d}}$ depends on how much smaller $\mathrm{s.e.}\left(\hat{\hat{d}} - d\right)$ is compared to $\mathrm{s.e.}\left(\hat{d} - d\right)$. (As will be seen in the next subsection, the estimates of $\mathrm{s.e.}\left(\hat{\hat{d}} - d\right)$ are just a little smaller than the estimates of $\mathrm{s.e.}\left(\hat{d} - d\right)$).

9.3. Illustration

We use the same data to calculate the uncertainty measures of $\hat{\hat{d}}$, as we used to calculate the uncertainty measures of the estimators $\hat{d}$ and $\tilde{d}$. Table 5 presents the values of the design-based standard error $\hat{\mathrm{s.e.}}\left(\hat{\hat{d}}\right)$, given by (30), and the confidence interval (31). Table 6 gives the values of the model-based standard error $\hat{\mathrm{s.e.}}\left(\hat{\hat{d}} - d\right)$, given by (33), and the confidence interval (34).

The values of $\hat{\hat{d}}$ and $\hat{d}$ are quite similar. Usually $\hat{\hat{d}}$ is larger than $\hat{d}$ (26 of the 36 estimates with $\hat{\hat{d}}$ are larger than the corresponding estimate with $\hat{d}$). We have the largest difference in 1. quarter 2002, where $\hat{\hat{d}} = 58.48$ while $\hat{d} = 56.82$ for domain E2. If we compare with the values of $\tilde{d}$ as well, we find that $\hat{\hat{d}}$ falls between $\hat{d}$ and $\tilde{d}$ in 23 cases (and usually we have the relationship $\hat{d} \le \hat{\hat{d}} \le \tilde{d}$).

From Table 5 we have that the estimates of $\text{s.e.}\left(\hat{\hat{d}}\right)$ vary around 1.45 for the population, 2.40 for domain E1, 3.20 for domain E2, and 2.20 for domain E5. We may say that the uncertainty is small for the population, somewhat larger for domain E1 and E5, and largest for domain E2.

The estimates of $\text{s.e.}\left(\hat{\hat{d}}\right)$ are almost equal to the estimates of $\text{s.e.}\left(\hat{d}\right)$. Some times $\hat{\text{s}}\text{.e.}\left(\hat{\hat{d}}\right)$ is smaller than $\hat{\text{s}}\text{.e.}\left(\hat{d}\right)$, other times larger. In one case the estimates are (approximately) the same. This suggests that the design-based standard errors $\text{s.e.}\left(\hat{\hat{d}}\right)$ and $\text{s.e.}\left(\hat{d}\right)$ are almost equal. (If we compare $\hat{\text{s}}\text{.e.}\left(\hat{\hat{d}}\right)$ and $\hat{\text{s}}\text{.e.}\left(\tilde{d}\right)$ we find that they are (approximately) equal in 24 cases).

In Table 6 we see that the model-based standard error $\hat{\text{s}}\text{.e.}\left(\hat{\hat{d}} - d\right)$ in 1. quarter 2002 equals 1.11, 1.80, 2.47 and 1.59 for respectively the population, domain E1, E2 and E5. The corresponding estimates in 2. quarter 2002 are 1.22, 1.72, 2.63 and 2.07.

We know that $\text{s.e.}\left(\hat{\hat{d}} - d\right)$ is smaller than $\text{s.e.}\left(\hat{d} - d\right)$ and $\text{s.e.}\left(\tilde{d} - d\right)$. We also know that $\hat{\text{s}}\text{.e.}\left(\hat{\hat{d}} - d\right)$ is smaller than $\hat{\text{s}}\text{.e.}\left(\hat{d} - d\right)$ and $\hat{\text{s}}\text{.e.}\left(\tilde{d} - d\right)$. From Table 4 and 6 we find that $\hat{\text{s}}\text{.e.}\left(\hat{\hat{d}} - d\right)$ is almost equal to $\hat{\text{s}}\text{.e.}\left(\tilde{d} - d\right)$, and just a little bit smaller than $\hat{\text{s}}\text{.e.}\left(\hat{d} - d\right)$. This could indicate that $\text{s.e.}\left(\hat{\hat{d}} - d\right)$ is almost equal to $\text{s.e.}\left(\tilde{d} - d\right)$, and just a little bit smaller than $\text{s.e.}\left(\hat{d} - d\right)$.

We have calculated the bounds of $\text{Bias}\left(\hat{\hat{d}}\right)$ from subsection 9.1, obtaining the lower bound -12.59 and the upper bound 12.59 (for 1. and 2. quarter 2002). The corresponding bounds of $\hat{d}$ and $\tilde{d}$ are ± 19.23 and ± 11.00, respectively. This means that the bounds of $\text{Bias}\left(\hat{\hat{d}}\right)$ are some smaller than the bounds of $\text{Bias}\left(\hat{d}\right)$, and a bit larger than the bounds of $\text{Bias}\left(\tilde{d}\right)$.

To get a better picture of the design-based bias of $\hat{\hat{d}}$, we have simulated 10000 independent configurations of the y_{hi}'s. For each of these we have calculated the bias of $\hat{\hat{d}}$. This gave 10000 biases in the interval -0.708 to 0.774, and 9903 of these were in the interval -0.5 to 0.5. This indicates that the bias of $\hat{\hat{d}}$ is small for most of the configurations of the y_{hi}'s.

We have also compared $\text{Bias}\left(\hat{d}\right)$ and $\text{Bias}\left(\hat{\hat{d}}\right)$ with randomly chosen configurations of the y_{hi}'s. This showed that $\left|\text{Bias}\left(\hat{\hat{d}}\right)\right|$ can be both larger and smaller than $\left|\text{Bias}\left(\hat{d}\right)\right|$, but that $\left|\text{Bias}\left(\hat{\hat{d}}\right)\right|$ usually is smaller than $\left|\text{Bias}\left(\hat{d}\right)\right|$.

Table 5: Design-based standard error

The figures in brackets are $\hat{d}$ and $\hat{\text{s}}.\text{e}.\left(\hat{d}\right)$ from Table 1.

2. QUARTER 1999

Aggregated level	Estimated diffusion index	Estimated standard error	Lower bound, 95% confidence interval	Upper bound, 95% confidence interval
Population	48.30 (47.73)	1.58 (1.59)	45.20	51.41
Domain E1	52.04 (52.40)	2.48 (2.47)	47.18	56.89
Domain E2	40.31 (38.86)	3.68 (3.64)	33.10	47.51
Domain E5	50.06 (49.16)	2.31 (2.40)	45.53	54.58

3. QUARTER 1999

Aggregated level	Estimated diffusion index	Estimated standard error	Lower bound, 95% confidence interval	Upper bound, 95% confidence interval
Population	55.48 (55.03)	1.51 (1.54)	52.51	58.44
Domain E1	57.13 (56.07)	2.51 (2.48)	52.21	62.04
Domain E2	39.87 (39.07)	3.24 (3.43)	33.52	46.22
Domain E5	65.58 (66.04)	2.27 (2.29)	61.13	70.03

4. QUARTER 1999

Aggregated level	Estimated diffusion index	Estimated standard error	Lower bound, 95% confidence interval	Upper bound, 95% confidence interval
Population	43.72 (42.99)	1.44 (1.44)	40.90	46.53
Domain E1	51.62 (51.01)	2.35 (2.33)	47.01	56.22
Domain E2	38.49 (38.30)	3.09 (3.13)	32.43	44.55
Domain E5	39.34 (38.07)	2.19 (2.17)	35.04	43.64

1. QUARTER 2000

Aggregated level	Estimated diffusion index	Estimated standard error	Lower bound, 95% confidence interval	Upper bound, 95% confidence interval
Population	63.58 (63.40)	1.46 (1.52)	60.72	66.43
Domain E1	69.05 (68.71)	2.36 (2.37)	64.42	73.68
Domain E2	53.01 (52.55)	3.30 (3.53)	46.53	59.49
Domain E5	66.01 (66.20)	2.08 (2.18)	61.94	70.09

2. QUARTER 2000

Aggregated level	Estimated diffusion index	Estimated standard error	Lower bound, 95% confidence interval	Upper bound, 95% confidence interval
Population	54.13 (53.50)	1.48 (1.45)	51.23	57.03
Domain E1	59.72 (59.94)	2.48 (2.49)	54.86	64.57
Domain E2	50.92 (50.28)	3.10 (3.02)	44.84	57.01
Domain E5	50.85 (49.35)	2.26 (2.16)	46.43	55.27

3. QUARTER 2000

Aggregated level	Estimated diffusion index	Estimated standard error	Lower bound, 95% confidence interval	Upper bound, 95% confidence interval
Population	60.42 (60.36)	1.52 (1.53)	57.45	63.40
Domain E1	63.03 (62.41)	2.45 (2.47)	58.22	67.83
Domain E2	54.50 (54.71)	3.24 (3.23)	48.15	60.85
Domain E5	62.40 (62.67)	2.37 (2.39)	57.75	67.04

4. QUARTER 2000

Aggregated level	Estimated diffusion index	Estimated standard error	Lower bound, 95% confidence interval	Upper bound, 95% confidence interval
Population	54.36 (53.59)	1.41 (1.43)	51.60	57.13
Domain E1	60.70 (60.26)	2.37 (2.39)	56.06	65.34
Domain E2	60.04 (59.46)	3.06 (3.14)	54.05	66.02
Domain E5	43.41 (42.13)	2.06 (2.07)	39.37	47.45

1. QUARTER 2002

Aggregated level	Estimated diffusion index	Estimated standard error	Lower bound, 95% confidence interval	Upper bound, 95% confidence interval
Population	63.90 (63.43)	1.32 (1.45)	61.31	66.49
Domain E1	66.04 (65.72)	2.27 (2.34)	61.59	70.49
Domain E2	58.48 (56.82)	2.84 (3.27)	52.90	64.06
Domain E5	65.78 (66.09)	1.85 (2.06)	62.16	69.40

2. QUARTER 2002

Aggregated level	Estimated diffusion index	Estimated standard error	Lower bound, 95% confidence interval	Upper bound, 95% confidence interval
Population	49.69 (49.70)	1.41 (1.43)	46.93	52.45
Domain E1	54.22 (54.73)	2.12 (2.17)	50.06	58.37
Domain E2	46.12 (45.31)	3.15 (3.17)	39.94	52.30
Domain E5	47.72 (47.85)	2.22 (2.29)	43.37	52.07

Table 6: Model-based standard error and confidence interval

The figures in brackets are $\hat{d}$ and ŝ.e.$\left(\hat{d}-d\right)$ from Table 2.

1. QUARTER 2002

Aggregated level	Estimated diffusion index	Estimated standard error	Lower bound, 95% confidence interval	Upper bound, 95% confidence interval
Population	63.90 (63.43)	1.11 (1.17)	61.74	66.07
Domain E1	66.04 (65.72)	1.80 (1.91)	62.51	69.57
Domain E2	58.48 (56.82)	2.47 (2.63)	53.64	63.32
Domain E5	65.78 (66.09)	1.59 (1.69)	62.65	68.90

2. QUARTER 2002

Aggregated level	Estimated diffusion index	Estimated standard error	Lower bound, 95% confidence interval	Upper bound, 95% confidence interval
Population	49.69 (49.70)	1.22 (1.28)	47.30	52.07
Domain E1	54.22 (54.73)	1.72 (1.81)	50.85	57.58
Domain E2	46.12 (45.31)	2.63 (2.79)	40.96	51.28
Domain E5	47.72 (47.85)	2.07 (2.17)	43.67	51.77

10. Summary

In this paper we have seen how the uncertainty of $\hat{d}$ can be measured by design-based standard error and model-based standard error. The design-based standard error measures the uncertainty arising from the fact that a lot of samples can be selected. The model-based standard error measures the uncertainty coming from the assumption that the y_{hi}'s are random variables that can take different values.

If we measure the uncertainty with the design-based standard error, the uncertainty is small when the diffusion index of the population is estimated (the estimates of the standard error vary from 1.43 to 1.59). The uncertainty is a little larger when the diffusion index of domain E1 and E5 is estimated (the estimates of the standard error vary from 2.17 to 2.49 for domain E1 and from 2.06 to 2.40 for domain E5). For domain E2 the estimates of the standard error vary form 3.02 to 3.64 and indicates some uncertainty for this domain.

When we measure the uncertainty with the model-based standard error, it seems to be a relative small uncertainty when the diffusion index of the population and the domain E1 is estimated (the estimates of the standard error for the population equal 1.17 and 1.28). The uncertainty is a little bit larger for the domain E5. We have the largest uncertainty when we estimate the diffusion index of domain E2 (the estimates of the standard error are 2.63 and 2.79).

Which of these measures we should use depends on which uncertainty we want to measure. The design-based measure is often used in survey sampling, but we believe that the model-based measure might be a better measure in the Business Tendency Survey. This is because the same sample is used

for several quarters, and for each quarter it is observed new values of the y_{hi}'s in the sample. Hence it seems reasonable to think of the sample as given and instead treat the y_{hi}'s as random variables.

In addition to measuring the uncertainty of $\hat{d}$ we have analysed two alternative estimators of the diffusion index ($\tilde{d}$ and $\hat{\hat{d}}$). The reason for doing this was to see if the diffusion index can be estimated more accurately with one of these estimators (compared to $\hat{d}$). Whether that is the case depends on whether we have a design-based or a model-based point of view.

If we choose a design-based point of view, we cannot say that one of the estimators always is better than the others. It seems that each of the estimators could be better than the others for some configurations of the y_{hi}'s, but not for all.

On the other hand, if we believe it is more proper to look at the estimators from a model-based point of view, then $\hat{\hat{d}}$ is the best estimator. This is because $\hat{\hat{d}}$ has the smallest model-based standard error. (All of the estimators are unbiased under the assumed model). But the differences between the standard errors are probably quite small, based on the estimates of the standard errors. Hence, it does not seem worthwhile to replace the estimator in use today.

References

Bjørnstad, J.F. (1995): *Utvalgsundersøkelser og prediksjon (in Norwegian. English title: Survey sampling and Prediction)*.

Chambers, R.: Handouts from the course ST640 (Survey Sampling and Estimation II) from MSc in Official Statistics.

Rao, J.N.K. and Wu, C.F.J. (1988): *Resampling Inference With Complex Survey Data.* JASA vol. 83 no. 401, 231-241.

Sitter, R.R. (1992): *A Resampling Procedure for Complex Survey Data.* JASA, vol. 87, no. 419, 755-765.

Recent publications in the series Documents

2002/5 P. Boug, Å. Cappelen and A. Rygh Swensen: Expectations and Regime Robustness in Price Formation: Evidence from VAR Models and Recursive Methods

2002/6 B.J. Eriksson, A.B. Dahle, R. Haugan, L. E. Legernes, J. Myklebust and E. Skauen: Price Indices for Capital Goods. Part 2 - A Status Report

2002/7 R. Kjeldstad and M. Rønsen: Welfare, Rules, Business Cycles and the Employment of Single Parents

2002/8 B.K. Wold, I.T. Olsen and S. Opdahl: Basic Social Policy Data. Basic Data to Monitor Status & Intended Policy Effects with Focus on Social Sectors incorporating Millennium Development Goals and Indicators

2002/9 T.A. Bye: Climate Change and Energy Consequenses.

2002/10 B. Halvorsen: Philosophical Issues Concerning Applied Cost-Benefit Analysis

2002/11 E. Røed Larsen: An Introductory Guide to the Economics of Sustainable Tourism

2002/12 B. Halvorsen and R. Nesbakken: Distributional Effects of Household Electricity Taxation

2002/13 H. Hungnes: Private Investments in Norway and the User Cost of Capital

2002/14 H. Hungnes: Causality of Macroeconomics: Identifying Causal Relationships from Policy Instruments to Target Variables

2002/15 J.L. Hass, K.Ø. Sørensen and K. Erlandsen: Norwegian Economic and Environment Accounts (NOREEA) Project Report -2001

2002/16 E.H. Nymoen: Influence of Migrants on Regional Varations of Cerebrovascular Disease Mortality in Norway. 1991-1994

2002/17 H.V. Sæbø, R. Glørsen and D. Sve: Electronic Data Collection in Statistics Norway

2002/18 T. Lappegård: Education attainment and fertility pattern among Norwegian women.

2003/1 A. Andersen, T.M. Normann og E. Ugreninov: EU - SILC. Pilot Survey. Quality Report from Staistics Norway.

2003/2 O. Ljones: Implementation of a Certificate in Official Statistics - A tool for Human Resource Management in a National Statistical Institute

2003/3 J. Aasness, E. Biørn and T. Skjerpen: Supplement to Distribution of Preferences and Measurement Errors in a Disaggregated Expenditure System

2003/4 H. Brunborg, S. Gåsemyr, G. Rygh and J.K. Tønder: Development of Registers of People, Companies and Properties in Uganda Report from a Norwegian Mission

2003/5 J. Ramm, E.Wedde and H. Bævre: World health survey. Survey report.

2003/6 B. Møller and L. Belsby: Use of HBS-data for estimating Household Final Consuption Final paper from the project. Paper building on the work done in the Eurostat Task Force 2002

2003/7 B.A. Holth, T. Risberg, E. Wedde and H. Degerdal: Continuing Vocational Training Survey (CVTS2). Quality Report for Norway.

2003/8 P.M. Bergh and A.S. Abrahamsen: Energy consumption in the services sector. 2000

2003/9 K-G. Lindquist and T. Skjerpen: Exploring the Change in Skill Structure of Labour Demand in Norwegian Manufacturing

2004/1 S. Longva: Indicators for Democratic Debate - Informing the Public at General Elections.

2004/2 H. Skiri: Selected documents on the modernisation of the civil registration system in Albania.

2004/3 J.H. Wang: Non-response in the Norwegian Business Tendency Survey.

2004/4 E. Gulløy and B.K Wold: Statistics for Development, Policy and Democracy. Successful Experience and Lessons Learned through 10 years of statistical and institutional development assistance and cooperation by Statistics Norway

2004/5 S. Glomsrød and L. Lindholt: The petroleum business environment.

2004/6 H.V. Sæbø: Statistical Metadata on the Internet Revised.

2004/7 M. Bråthen: Collecting data on wages for the Labour Force Survey – a pilot

2004/8 A.L. Brathaug and E. Fløttum: Norwegian Experiences on Treatment of Changes in Methodologies and Classifications when Compiling Long Time Series of National Accounts.

2004/9 L. Røgeberg, T. Skoglund and S. Todsen: Report on the project Quality adjusted input price indices for collective services in the Norwegian national accounts. Report from a project co-financed by Eurostat.