Statistisk sentralbyrå
Statistics Norway

Some inference issues regarding modeling, variance estimation and nonresponse in survey sampling

Jan Bjørnstad, Statistics Norway, email: jab@ssb.no

Workshop on Survey Sampling Theory and Methodology
August 25-29, Kuressaare, Estonia

- Lecture 1: Discussion of design-based versus model-based inference. Likelihood and likelihood principle in sampling
- Lecture 2: Different variance measures and related variance estimation
- Lecture 3: Nonresponse issues and imputation
- Lecture 4: Variance estimation in the presence of nonresponse. Multiple imputation methods for non-Bayesian imputation

Lecture 1:
Theoretical talk- on the foundation of survey sampling

Design-based inference

- Population (Target population): The universe of all units of interest for a certain study: U = {1,2, …, N}
  - All units can be identified and labeled
  - Variable of interest y with population values $\mathbf{y} = (y_1, y_2, ..., y_N)$
  - Typical problem: Estimate total t or population mean t/N
- Sample: A subset s of the population, to be observed
- Sampling design $p(s)$ is known for all possible subsets;
  - The probability distribution of the stochastic sample

Simple random sample (SRS) of size *n*

$$p(s) = 1/\binom{N}{n} \text{ if } |s| = n$$

$$= 0 \text{ if } |s| \neq n$$

Estimation of the population mean, with no auxiliary variables, use the sample mean

$$\bar{y}_s = \sum_{i \in s} y_i / n$$

- Design-unbiased: $E(\bar{y}_s) = \sum_s \bar{y}_s p(s) = t / N = \bar{y}$
- Design-variance:

$$Var(\bar{y}_s) = (1 - f)\frac{S^2}{n},$$

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2 \quad \text{and } f = n / N$$

Statistisk sentralbyrå
Statistics Norway

Problems with design-based inference

- Generally: Design-based inference is with respect to *hypothetical* replications of sampling for a *fixed* population vector **y**
- Variance estimates may fail to reflect information in a *given sample*
- Difficult to combine with models for nonsampling errors like nonresponse
- If we want to measure how a certain estimation method does in quarterly or monthly surveys, then **y** will vary from quarter to quarter or month to month – need to assume that **y** is a realization of a random vector
- **Today's lecture: Likelihood and likelihood principle as guideline on how to deal with these issues**

Statistisk sentralbyrå
Statistics Norway

• Nonexistence of optimal estimators

Theorem

Let $p(s)$ be any sampling design with $p(U) < 1$. Then there exists no uniformly best (minimum variance) estimator for the total t

Proof

1. For any $\hat{t}$ unbiased and population vector $\mathbf{y}_0$ there exists an unbiased estimator $\hat{t}_0$ with variance 0 at $\mathbf{y}_0$

2. If $\hat{t}$ has uniformly minimum variance, it must have variance 0 for all values of $\mathbf{y}$

3. That is impossible

Problem with design-based variance measure Illustration 1

a) $N+1$ possible samples: $\{1\}, \{2\},\ldots,\{N\}, \{1,2,\ldots N\}$

b) Sampling design: $p(\{i\}) = 1/2N$, for $i = 1,..,N$;
$p(\{1,2,\ldots N\}) = 1/2$

c) Use $\bar{y}_s$ as the estimator for the population mean $\bar{y}$

$$\text{Unbiased}: E(\bar{y}_s) = \sum_s p(s)\bar{y}_s = \sum_{i=1}^{N} \frac{1}{2N} y_i + \frac{1}{2}\bar{y} = \bar{y}$$

$$\text{Design - variance}:$$

$$Var(\bar{y}_s) = E(\bar{y}_s - \bar{y})^2 = \sum_{i=1}^{N} (y_i - \bar{y})^2 \cdot \frac{1}{2N} = \frac{1}{2} \cdot \frac{N-1}{N} S^2 = \frac{1}{2} \cdot \tilde{S}^2$$

d) Assume we select the "sample" $\{1,2,\ldots,N\}$. Then we claim that the "precision" of the resulting sample (known to be without error) is $\tilde{S}^2/2$

Problem with design-based variance measure Illustration 2

a) Expert 1 : SRS and estimate $\bar{y}_s$

Precision is measured by $(1 - f)\dfrac{S^2}{n}$

b) Expert 2 : SRS with replacement and estimate $\bar{y}_s$

measures precision by $\tilde{S}^2 / n$

Both experts select the <u>same</u> sample, compute the <u>same</u> estimate, but give <u>different</u> measures of precision…

The likelihood principle, LP general model

$$\text{Model}: X \sim f_\theta(x), \theta \in \Omega; \theta \text{ are the unknown parameters in the model}$$

- The likelihood function, with *data x*: $l_x(\theta) = f_x(\theta)$

  l is quite a different animal than f !!

  Measures the likelihood of different θ values in light of the data x

- LP: The likelihood function contains all information about the unknown parameters
- More precisely: Two proportional likelihood functions for θ, from the same or different experiments, should give identically the same statistical inference

• Maximum likelihood estimation satisfies LP, using the curvature of the likelihood as a measure of precision (Fisher)

• LP is controversial, but hard to argue against because of the fundamental result by Birnbaum, 1962:

• LP follows from sufficiency and conditionality principles that ”no one” disagrees with.

• SP: Statistical inference should be based on sufficient statistics

• CP: If you have 2 possible experiments and choose one at random, the inference should depend only on the chosen experiment

Radical consequences for statistical analysis

- Statistical analysis, given the observed data: The sample space is irrelevant
- The usual criteria like confidence levels and P-values do not necessarily measure reliability for the actual inference given the observed data
- Frequentistic measures evaluate *methods*
  - *not necessarily relevant criteria for the observed data*

Illustration- Bernoulli trials

$X_1, \ldots, X_i, \ldots$

$X_i = 1$ (success) with probability θ

Two experiments to gain information about θ:

$E_1 : n = 12$ observations and observe $Y_1 = \sum_{i=1}^{12} X_i$

E_2 : Continue trials until we get 3 failures (0's) and observe Y_2 = number of successes

Suppose the results are $y_1 = y_2 = 9$

The likelihood functions:

$$l_9^{(1)}(\theta) = \binom{12}{9}\theta^9(1-\theta)^3 \qquad \text{binomial}$$

$$l_9^{(2)}(\theta) = \binom{11}{9}\theta^9(1-\theta)^3 \qquad \text{negative binomial}$$

Proportional likelihoods: $l_9^{(2)}(\theta) = (1/4)\, l_9^{(1)}(\theta)$

LP: Inference about θ should be identical in the two cases

Frequentistic analyses give different results:

F.ex. test $H_0 : \theta = 1/2$ against $H_1 : \theta > 1/2$

$(E_1, 9)$: P - value $= 0.0730$ $\qquad$ $(E_2, 9)$: P - value $= 0.0327$

because different sample spaces: (0,1,..,12) and (0,1,...)

Frequentistic vs. likelihood

- Frequentistic approach: Statistical methods are evaluated pre-experimental, over the sample space
- LP evaluate statistical methods post-experimental, given the data
- History and dicussion after Birnbaum, 1962: An overview in ”*Breakthroughs in Statistics,1890-1989, Springer 1991*”

Likelihood function in design-based inference

- Unknown parameter: $\mathbf{y}=(y_1, y_2 \ldots, y_N)$
- Data: $x = \{(i, y_{obs,i}) : i \in s\}$
- Likelihood function = Probability of the data, considered as a function of the parameters

$$\Omega_x = \{\mathbf{y} : y_i = y_{obs,i} \text{ for } i \in s\}$$

- Sampling design: $p(s)$
- Likelihood function: $l_x(\mathbf{y}) = \begin{cases} p(s) \text{ if } \mathbf{y} \in \Omega_x \\ 0 \ \text{ otherwise} \end{cases}$
   - All possible **y** are equally likely !!

- Likehood principle, LP : The likelihood function contains all information about the unknown parameters
- **According to LP:**
  - The design-model is such that the data contains no information about the unobserved part of $\mathbf{y}$, $\mathbf{y}_{unobs}$
  - One has to assume in advance that there is a relation between the data and $\mathbf{y}_{unobs}$:
    - As a consequence of LP: Necessary to assume a model
  - The sampling design is irrelevant for statistical inference, because two sampling designs leading to the same *s* will have proportional likelihoods

Let p_0 and p_1 be two sampling designs. Assume we get the same sample s in either case. Then the data x are the same and Ω_x are the same for both experiments.

The likelihood function for sampling design p_i , $i = 0{,}1$:

$$l_{i,x}(\mathbf{y}) = \begin{cases} p_i(s) \text{ if } \mathbf{y} \in \Omega_x \\ 0 \text{ otherwise} \end{cases}$$

$$\Rightarrow l_{1,x}(\mathbf{y}) / l_{0,x}(\mathbf{y}) = p_1(s) / p_0(s) \text{ if } \mathbf{y} \in \Omega_x$$

and then for *all* $\mathbf{y}$:

$$l_{1,x}(\mathbf{y}) = \frac{p_1(s)}{p_0(s)} l_{0,x}(\mathbf{y})$$

- Same inference under the two different designs. This is in direct opposition to usual design-based inference, where the only stochastic evaluation is thru the sampling design, for example the Horvitz-Thompson estimator
- Concepts like design unbiasedness and design variance are irrelevant according to LP when it comes to do the actual statistical analysis.
- Note: LP is not concerned about method performance, but the statistical analysis *after* the data have been observed
- This *does not mean* the sampling design is not important. It is important to assure we get a good representative sample. But once the sample is collected the sampling design should not play a role in the inference phase, according to LP

Model-based inference

- Assumes a model for the **y** vector
- Conditions on the actual sample
- Use modeling to combine information
- **Problem:** dependence on model
  - Introduces a subjective element, but no different than usual statistical modeling
  - almost impossible to model all variables in a survey
- Design approach is "objective" in a perfect world of no nonsampling errors

Model-based approach

$y_1, y_2, \ldots, y_N$ are realized values of

random variables $Y_1, Y_2, \ldots Y_N$

Two stochastic elements:

1) sample $s \sim p(\cdot)$ 2) $(Y_1, Y_2, \ldots Y_N) \sim f_\theta$

Treat the sample s as fixed

[Model-assisted approach: use the distribution assumption of Y to construct estimator, and evaluate according to distribution of s, given the realized vector **y**]

We can decompose the total t as follows:

$$t = \sum_{i=1}^{N} y_i = \sum_{i \in s} y_i + \sum_{i \notin s} y_i$$

Since $\sum_{i\in s} y_i$ is known, the problem is to estimate

$z = \sum_{i\notin s} y_i$, the realized value of $Z = \sum_{i\notin s} Y_i$

- The unobserved z is a realized value of the random variable Z, so the problem is actually to *predict* the value z of Z.

Can be done by predicting each unobserved y_i: $\hat{Y}_i, i \notin s$

Estimator : $\hat{T}_{pred} = \sum_{i\in s} y_i + \sum_{i\notin s} \hat{Y}_i = \sum_{i\in s} y_i + \hat{Z}$

$\hat{Z}$ is a predictor for z

- The prediction approach, the prediction based estimator

Determine $\hat{Y}_i$ by modeling,

similar to the model - assisted approach

Predictive likelihood approach

- Prediction problem. May use a likelihood approach
- Data: x, unknown: z. Joint distribution: $f_\theta(x,z)$
- Joint likelihood for the unknown quantities:

$$l_x(z,\theta) = f_\theta(x,z)$$

- Corresponding likelihood principle is implied by principles of prediction suffiency and conditionality
- Aim: To develop a partial likelihood for z, $L(z|x)$, from l_x
- Any such likelihood is called a *predictive likelihood* and gives rise to one particular prediction method

One basic predictive likelihood: Profile PL:

$$L_p(z \mid x) = \max_\theta l_y(z,\theta) = \max_\theta f_\theta(x,z)$$

Any predictive likelihood *L* is assumed normalized as a probability distribution in *Z*

The mean in *L*, $E_{pl}(Z)$, is a predictor for *Z*

3 typical models

I. A model for business surveys, the ratio model:

$$Y_i = \beta x_i + \varepsilon_i \quad \text{with } E(\varepsilon_i) = 0, Var(\varepsilon_i) = \sigma^2 x_i \ \text{ and } Cov(\varepsilon_i, \varepsilon_j) = 0$$

$$\Leftrightarrow E(Y_i) = \beta x_i \ , Var(Y_i) = \sigma^2 x_i \ \text{ and } Cov(Y_i, Y_j) = 0$$

II. A model for social surveys, simple linear regression:

$$Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i \ , \ E(\varepsilon_i) = 0, \ Var(\varepsilon_i) = \sigma^2 \ \text{ and } Cov(\varepsilon_i, \varepsilon_j) = 0$$

- Ex: x_i is a measure of the “size” of unit *i*, and y_i tends to increase with increasing x_i. In business surveys, the regression goes thru the origin in many cases

III. Common mean model:

$$E(Y_i) = \beta \ , \ Var(Y_i) = \sigma^2 \text{ and the } Y_i\text{'s} \text{ are uncorrelated}$$

Remarks:

1. The model-assisted regression estimator has often the form

$$\hat{T}_{reg} = \sum_{i=1}^{N} \hat{Y}_i, \ \hat{Y}_i = \hat{\beta} x_i \ \text{ in case of a ratio model}$$

2. The prediction approach makes it clear: no need to estimate the observed y_i

3. **Any** estimator can be expressed on the "prediction form:

$$\hat{T} = \sum_{i \in s} Y_i + \hat{Z}_{\hat{t}}$$
$$\text{letting } \hat{Z}_{\hat{t}} = \hat{T} - \sum_{i \in s} Y_i$$

4. Can then use this form to see if the estimator makes any sense

Ex 1. $\hat{t} = N\bar{y}_s = \sum_{i\in s} y_i + (N-n)\bar{y}_s = \sum_{i\in s} y_i + \sum_{i\notin s} \bar{y}_s$

Hence, $\hat{z} = \sum_{i\notin s} \bar{y}_s$ and $\hat{y}_i = \bar{y}_s$, for all $i \in s$

Ex.2 $\hat{t}_{HT} = \sum_{i\in s} y_i / \pi_i$ and $\pi_i = nx_i / X$, $X = \sum_{i=1}^{N} x_i$

Reasonable sampling design when *y* and *x* are positively correlated

$$\hat{t}_{HT} = \sum_{i\in s} \frac{X \cdot y_i}{nx_i} = \sum_{i\in s} y_i + \sum_{i\in s} y_i \left(\frac{X}{nx_i} - 1 \right)$$

$$= \sum_{i\in s} y_i + \underbrace{\frac{1}{n} \sum_{i\in s} \frac{y_i}{x_i} \left(\frac{(X - nx_i)}{X - n\bar{x}_s} \right)}_{\hat{\beta}_{HT}} \sum_{i\notin s} x_i = \sum_{i\in s} y_i + \hat{z}_{HT}$$

$$\hat{z}_{HT} = \sum_{i\notin s} \hat{\beta}_{HT} x_i = \sum_{i\notin s} \hat{y}_i$$

$\hat{\beta}_{HT}$ is a rather unusual regression coefficient

Model-based estimators (predictors)

1. Predictor: $\hat{T} = \sum_{i \in s} Y_i + \hat{Z}$
2. Model parameters: θ
3. $\hat{T}$ is model-unbiased if $E_\theta(\hat{T} - T \mid s) = 0 \;\; \forall \theta, \; T = \sum_{i=1}^{N} Y_i$
4. Model variance of model-unbiased predictor is the variance of the *prediction error*, also called the *prediction variance*

$$Var_\theta(\hat{T} - T \mid s) = E_\theta((\hat{T} - T)^2 \mid s)$$

Prediction variance as a variance measure for the actual observed sample

Illustration 1, slide 5

$N+1$ possible samples: $\{1\}, \{2\}, \ldots, \{N\}, \{1,2,\ldots N\}$

Use $\hat{T} = N\overline{Y}_s$ as the estimator for the population total T

Assume we select the "sample" $\{1,2,\ldots,N\}$.

Then $\hat{T} = N\overline{Y} = T$

Prediction variance: $Var(\hat{T} - T) = Var(0) = 0$

Illustration 2, slide 6: Exactly the same prediction variance for the two sampling designs

Linear predictor: $\hat{T} = \sum_{i \in s} a_i(s) Y_i$

5. Optimality:

$\hat{T}_0$ is the best linear unbiased (BLU) predictor for T if

1) $\hat{T}_0$ is model-unbiased

2) $\hat{T}_0$ has uniformly minimum prediction variance among all model-unbiased linear predictors:

For any model-unbiased linear predictor $\hat{T}$

$$Var_\theta(\hat{T}_0 - T) \le Var_\theta(\hat{T} - T) \text{ for all } \theta$$

Lecture 2: Different variance measures and related variance estimation

- We have seen two variance measures:

  Design-based variance

  Model-based (prediction) variance.
- A third variance measure: Anticipated variance (method variance)
- A fourth variance measure:

  Variance in a normalized predictive likelihood

Bootstrap methods for estimating design-based variance

• Bootstrap: Unaided efforts, by one's own bootstrap, self-reliant

• Mention 2 standard methods, for estimating a population quantity θ, function of the population mean

Method 1. Without-replacement bootstrap, BWO

1. Construct a pseudo population U^* from the sample s

   If π_k is the inclusion probability for unit k.

   $U^* : 1/\pi_k$ copies of each $y_k, k \in s$

   Population size : $N^* = \sum_s 1/\pi_k = \hat{N}$

   Population total : $t^* = \sum_s y_k (1/\pi_k) = \hat{t}_{HT}$

Illustration: Simple random sample:

$$1/\pi_k = N/n, N^* = N, t^* = N\overline{y}_s$$

2. From *U** draw B independent ”resamples” with replacement, using the *same sampling design* as for the original sample *s*

3. Estimates: $\hat{\theta}_1^*,...,\hat{\theta}_B^*$. Values of original estimator $\hat{\theta}$.

$$\hat{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*$$

Variance estimate: $\hat{V}(\hat{\theta}) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b^* - \hat{\theta}^*)^2$

Problem: Does not yield reasonable estimates except in the simplest sampling plans

Method 2, developed for stratified samples, BWR

- Draw resamples directly from the sample
- Problem: The original observations are not independent
- Rescale the resampled values:
- For each stratum of simple random sample *s* ; *n,N*

1. Resample: Simple random sample from s, with replacement:

$\{ y_i^*; i = 1,\ldots,m \}$

$$\text{Compute}: \tilde{y}_i = \bar{y}_s + \left[\frac{m}{n-1}(1-f) \right]^{1/2} \left(y_i^* - \bar{y}_s \right), \quad f = n/N$$

2. B independent resamples. Each time compute

$$\tilde{\theta}_b = \hat{\theta} \text{ based on } \{\tilde{y}_i; i = 1,\dots,n\}$$

$$\tilde{\theta} = \frac{1}{B}\sum_{b=1}^{B}\tilde{\theta}_b$$

3. Varians estimate:

$$\hat{V}_{BS} = \frac{1}{B-1}\sum_{b=1}^{B}(\tilde{\theta}_b - \tilde{\theta})^2$$

- Can be used in complex estimation problems.
- Consistent variance estimator as the number of strata goes to infinity
- The expected value over the bootstrap samples reduces to the usual variance estimate in the linear case

Anticipated variance (method variance)

We want a variance measure that tells us about the expected uncertainty in *repeated* surveys

1. Conditional on the sample s, with model-unbiased $\hat{T}$:

$Var(\hat{T} - T)$ measures the uncertainty for *this* particular sample s

2. The expected uncertainty for repeated surveys:

$E_p\{Var(\hat{T} - T)\}$, over the sampling distribution $p(\cdot)$

3. This is called the *anticipated variance*.

4. It can be regarded as a variance measure that describes how the estimation *method* is doing in repeated surveys

Statistisk sentralbyrå
Statistics Norway

If $\hat{T}$ is not model-unbiased, we use

$$E_p\{E(\hat{T}-T)^2\}$$

as a criterion for uncertainty, the anticipated mean square error

Note: If $\hat{T}$ is design-unbiased then

$$E_p\{E(\hat{T}-T)^2\} = E\{E_p(\hat{T}-T)^2 \mid \mathbf{Y})\}$$

and

$$E_p(\hat{T}-T)^2 \mid \mathbf{Y} = \mathbf{y}) = E_p(\hat{t}-t)^2 = Var_p(\hat{t})$$

And the anticipated MSE becomes the expected design-variance, also called the anticipated design variance

$$E_p\{E(\hat{T}-T)^2\} = E\{Var_p(\hat{T})\}$$

Example:
Ratio model and simple random sample

Model :

$$Y_i = \beta x_i + \varepsilon_i \text{ , } E(\varepsilon_i) = 0 \text{ and } Var(\varepsilon_i) = \sigma^2 x_i$$

$$Y_1,\ldots,Y_N \text{ are uncorrelated , } Cov(\varepsilon_i, \varepsilon_j) = 0$$

Auxiliary information **x** known for the whole population

BLU predictor:

$$\hat{T}_{pred} = \sum_{i\in s} Y_i + \sum_{i\notin s} \hat{\beta}_{opt} x_i$$

where $\hat{\beta}_{opt}$ is the best linear unbiased estimator (BLUE) of β

$$\hat{\beta}_{opt} = \frac{\sum_{i\in s} Y_i}{\sum_{i\in s} x_i} = \hat{R}$$

$$\hat{T}_{pred} = \sum_{i \in s} Y_i + \hat{R} \sum_{i \notin s} x_i = X \cdot \hat{R} = \hat{T}_R$$

where $X = \sum_{i=1}^{N} x_i$

The usual ratio estimator : Approximately design unbiased

Let $\bar{x}_r = \sum_{i \notin s} x_i / (N-n)$ and $\bar{x} = X / N$

$$Var(\hat{T}_{pred} - T) = Var(\hat{R} \sum_{i \notin s} x_i - \sum_{i \notin s} Y_i)$$

$$= (N-n)^2 \bar{x}_r^2 \frac{\sigma^2}{n\bar{x}_s} + \sigma^2 (N-n)\bar{x}_r = (N-n)\sigma^2 \bar{x}_r \left[\frac{(N-n)\bar{x}_r + n\bar{x}_s}{n\bar{x}_s} \right]$$

$$= (N-n)\sigma^2 \frac{\bar{x}_r \cdot N\bar{x}}{n\bar{x}_s} = N^2 \frac{1-f}{n} \cdot \frac{\bar{x}_r \bar{x}}{\bar{x}_s} \sigma^2$$

$$E_p\{Var(\hat{T}_{pred} - T)\} = N^2 \frac{1-f}{n} \sigma^2 \bar{x} E_p(\frac{\bar{x}_r}{\bar{x}_s})$$

$$\approx N^2 \frac{1-f}{n} \sigma^2 \bar{x} \cdot \frac{E_p(\bar{x}_r)}{E_p(\bar{x}_s)} = N^2 \frac{1-f}{n} \bar{x} \sigma^2$$

Unbiased estimator of σ^2: Usual least squares estimator:

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i \in s} \frac{1}{x_i} (Y_i - \hat{R} x_i)^2$$

Design-based variance estimate:

$$\hat{V}_{SRS}(\hat{t}_R) = N^2 \cdot \frac{1-f}{n} \cdot \bar{x} \frac{\bar{x}}{\bar{x}_s^2} \cdot \frac{1}{n-1} \sum_s (y_i - \hat{R} x_i)^2$$

Estimation of model-based variance: Robust variance estimation

- The model assumed is really a "working model"
- Especially, the variance assumption may be misspecified and it is not always easy to detect this kind of model failure
  - like constant variance
  - variance proportional to size measure x_i
- Standard least squares variance estimates is sensitive to misspecification of variance assumption
- Concerned with robust variance estimators

The ratio estimator

Working model:

$$Y_i = \beta x_i + \varepsilon_i \ , E(\varepsilon_i) = 0 \ \text{ and } \ Var(\varepsilon_i) = \sigma^2 x_i$$

$$Y_1,\ldots,Y_N \text{ are uncorrelated} \ , Cov(\varepsilon_i, \varepsilon_j) = 0$$

Under this working model, the unbiased estimator of the prediction variance of the ratio estimator is

$$\hat{V}(\hat{T}_R - T) = N^2 \frac{1-f}{n} \cdot \frac{\bar{x}_r \bar{x}}{\bar{x}_s} \hat{\sigma}^2$$

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i \in s} \frac{1}{x_i} (Y_i - \hat{R} \cdot x_i)^2$$

$$\hat{R} = \bar{Y}_s / \bar{x}_s$$

This variance estimator is non-robust to misspecification of the variance model.

Suppose the true model has

$$E(Y_i) = \beta x_i \text{ and } Var(Y_i) = \sigma^2 v(x_i)$$

Ratio estimator is still model-unbiased but prediction variance is now

$$Var(\hat{T}_R - T) = (\sum_{i \notin s} x_i)^2 Var(\hat{R}) + \sigma^2 \sum_{i \notin s} v(x_i)$$

$$= (\sum_{i \notin s} x_i)^2 \frac{\sigma^2 \sum_{i \in s} v(x_i)}{(\sum_{i \in s} x_i)^2} + \sigma^2 \sum_{i \notin s} v(x_i)$$

$$= \sigma^2 \left(\frac{(N-n)^2 \bar{x}_r^2}{n^2 \bar{x}_s^2} \sum_{i \in s} v(x_i) + \sum_{i \notin s} v(x_i) \right)$$

$$Var(\hat{T}_R - T) = \sigma^2 \left(\frac{(N-n)^2 \bar{x}_r^2}{n\bar{x}_s^2} \bar{v}_s + (N-n)\bar{v}_r \right)$$

$$= \sigma^2 N^2 \frac{1-f}{n} \left((1-f)\bar{v}_s (\bar{x}_r / \bar{x}_s)^2 + f \cdot \bar{v}_r \right)$$

$$\bar{v}_s = \sum_{i \in s} v(x_i) / n \quad \text{and} \quad \bar{v}_r = \sum_{i \notin s} v(x_i) / (N-n)$$

Moreover, $\quad E(\hat{\sigma}^2) \neq \sigma^2$:

$$E(\hat{\sigma}^2) = \frac{1}{n-1} \sum_{i \in s} \frac{1}{x_i} E(Y_i - \hat{R} \cdot x_i)^2$$

$$= \sigma^2 \left[(v/x)_s + \frac{1}{n-1} \{ (v/x)_s - \bar{v}_s / \bar{x}_s \} \right], (v/x)_s = \frac{1}{n} \sum_{i \in s} v(x_i) / x_i$$

Statistisk sentralbyrå
Statistics Norway

Robust variance estimator for the ratio estimator

$$Var(\hat{T}_R - T) = \sigma^2 N^2 \frac{1-f}{n}\left((1-f)\bar{v}_s(\bar{x}_r/\bar{x}_s)^2 + f\cdot\bar{v}_r\right)$$

$$= \sigma^2 N^2 \frac{1-f}{n}\left(\bar{v}_s(\bar{x}_r/\bar{x}_s)^2 + f\cdot\{\bar{v}_r - \bar{v}_s(\bar{x}_r/\bar{x}_s)^2\}\right)$$

$$\approx \sigma^2\bar{v}_s \cdot N^2 \frac{1-f}{n}(\bar{x}_r/\bar{x}_s)^2 \ ,$$

the leading term in the prediction variance

and: $\sigma^2\bar{v}_s = \frac{1}{n}\sum_{i\in s}\sigma^2 v(x_i) = \frac{1}{n}\sum_{i\in s} Var(Y_i)$

$$\sigma^2\bar{v}_s = \frac{1}{n}\sum_{i\in s} E(Y_i - \beta x_i)^2 = E\{\frac{1}{n}\sum_{i\in s}(Y_i - \beta x_i)^2\}$$

Statistisk sentralbyrå
Statistics Norway

Suggests we may use:

$$\hat{\sigma}^2_{rob}\bar{v}_s = \frac{1}{n-1}\sum_{i\in s}(Y_i - \hat{R}x_i)^2$$

Leading to the robust variance estimator:

$$\hat{V}_{rob}(\hat{T}_R - T) = (\bar{x}_r / \bar{x}_s)^2 \cdot N^2 \frac{1-f}{n} \cdot \frac{1}{n-1}\sum_{i\in s}(Y_i - \hat{R}x_i)^2$$

Almost the same as the *design* variance estimate in SRS:

$$\hat{V}_{SRS}(\hat{t}_R) = (\bar{x} / \bar{x}_s)^2 \cdot N^2 \frac{1-f}{n} \cdot \frac{1}{n-1}\sum_{i\in s}(y_i - \hat{R}x_i)^2$$

A new interpretation of this variance estimate!!

$$E\left[\hat{V}_{rob}(\hat{T}_R - T)\right] \approx (\bar{x}_r / \bar{x}_s)^2 \cdot N^2 \frac{1-f}{n} \cdot \sigma^2 \bar{v}_s \approx \left[V(\hat{T}_R - T)\right]$$

Approximately model-unbiased

Can we do better?

Require estimator to be exactly unbiased under ratio model, $v(x) = x$:

When $v(x) = x : E\{\frac{1}{n-1}\sum_{i\in s}(Y_i - \hat{R}\cdot x_i)^2\}$

$$= \frac{1}{n-1}\sum_{i\in s} E(Y_i - \hat{R}x_i)^2 = \frac{1}{n-1}\sum_{i\in s}\sigma^2 x_i (1 - \frac{x_i}{n\bar{x}_s})$$

$$= \sigma^2 \bar{x}_s \left(1 - \frac{1}{n}\cdot\frac{s_x^2}{\bar{x}_s^2}\right), \quad s_x^2 = \frac{1}{n-1}\sum_{i\in s}(x_i - \bar{x}_s)^2$$

The prediction variance when $v(x) = x$:

$$V(\hat{T}_R - T) = N^2 \frac{1-f}{n} \cdot \frac{\bar{x}_r \bar{x}}{\bar{x}_s} \sigma^2$$

$$E\{\hat{V}_{rob}(\hat{T}_R - T)\} = N^2 \frac{1-f}{n} (\bar{x}_r^2 / \bar{x}_s) \sigma^2 \left(1 - \frac{1}{n} \cdot \frac{s_x^2}{\bar{x}_s^2}\right)$$

So a robust variance estimator that is exactly unbiased under the working model , $v(x) = x$:

$$\hat{V}_{R,rob}(\hat{T}_R - T)\} = \frac{\bar{x}}{\bar{x}_r} \left(1 - \frac{1}{n} \cdot \frac{s_x^2}{\bar{x}_s^2}\right)^{-1} \hat{V}_{rob}(\hat{T}_R - T)$$

$$= \{1 - n^{-1}(s_x^2 / \bar{x}_s^2)\}^{-1} (\bar{x}_r \bar{x} / \bar{x}_s^2) \cdot N^2 \frac{1-f}{n} \cdot \frac{1}{n-1} \sum_{i \in s} (Y_i - \hat{R} x_i)^2$$

$$= \{1 - n^{-1}(s_x^2 / \bar{x}_s^2)\}^{-1} (\bar{x}_r / \bar{x}) \cdot \hat{V}_{SRS}(\hat{t}_R)$$

General approach to robust variance estimation

1. Find robust estimators of $Var(Y_i)$, that does not depend on model assumptions about the variance

2. $\hat{T} = \sum_{i \in s} w_{is} Y_i$

$$Var(\hat{T} - T) = \sum_{i \in s} (w_{is} - 1)^2 Var(Y_i) + \sum_{i \notin s} Var(Y_i)$$

3. For $i \in s : \hat{V}(Y_i) = (Y_i - \hat{\mu}_i)^2$

$\hat{\mu}_i$ estimate $E(Y_i)$ under true model

4. Estimate only leading term in the prediction variance, typically dominating, or estimate the second term from the more general model

Statistisk sentralbyrå
Statistics Norway

Predictive likelihood variance

Predictive likelihood for Z, *normalized* as a probability distribution: $L(\mathrm{z})$

Predictive likelihood variance, $V_{pl}(Z)$, is the variance in $L(z)$

$V_{pl}(Z)$ is based on the data only, and is therefore automatically a ”variance estimate” or if you like, a data-based measure of uncertainty

Ratio model –estimating the total

$Z = \sum_{i \notin s} Y_i$ to be predicted

The profil predictive likelihood for Z is such that

$$\frac{Z - \hat{R}\sum_{i \notin s} x_i}{\hat{\sigma} N \sqrt{\frac{n-1}{n}} \sqrt{\frac{1-f}{n} \cdot \frac{\bar{x}_r \bar{x}}{\bar{x}_s}}} \sim t_n - \text{distribution}$$

Predictive mean: $E_{pl}(Z) = \hat{R}\sum_{i \notin s} x_i$

$$\Rightarrow E_{pl}(T) = \sum_{i \in s} Y_i + \hat{R}\sum_{i \notin s} x_i = X \cdot \hat{R} = \hat{T}_R$$

Predictive variance:

$$V_{pl}(Z) = \frac{n-1}{n-2}\hat{\sigma}^2 N^2 \frac{1-f}{n} \cdot \frac{\bar{x}_r \bar{x}}{\bar{x}_s}$$

$$= \frac{n-1}{n-2}\hat{V}(\hat{T}_R - T)$$

Some references- after topic

- **Likelihood theory**:
  - Birnbaum (1962). On the foundations of statistical inference (with discussion). *Journal of the American Statistical Association*, 57, 269-306.
  - Bjørnstad (1996). On the generalization of the likelihood function and the likelihood principle. *Journal of the American Statistical Association,* 91, 791-806.
- **Likelihood theory in survey sampling**
  - Godambe (1966). A new approach to sampling from finite populations,I. *Journal of the Royal Statistical Society B,* 28**,** 310 – 331.
  - Basu (1969). Role of the sufficiency and likelihood principles in sample survey theory. *Sankhya,* A 31**,** 441 – 454.
  - Royall (1976). Likelihood functions in finite population sampling survey. *Biometrika*, 63, 605 – 617.

- **Model-based approach**:
  - Valliant, Dorfman and Royall (2000).*Finite Population Sampling and Inference. A Prediction Approach.* Wiley (ch. 5 deals with robust variance estimation)
- **Predictive likelihood**:
  - Bjørnstad (1990).Predictive likelihood: A Review (with discussion). *Statistical Science, 5,* 242-265.
- **Predictive likelihood in survey sampling**:
  - Bolfarine and Zacks (1992). *Prediction theory for Finite Populations.* Springer
  - Bjørnstad and Ytterstad (2008). Two-stage sampling from a prediction point of view when the cluster sizes are unknown. *Biometrika*, 95, 187-204.

Statistisk sentralbyrå
Statistics Norway

- **Boostrap methods for variance estimation**
  – Gross (1980). Median estimation in sample surveys. *Proceedings of the Section on Survey Research,* Amer. Statisti. Assoc., 181-184
  – Rao and Wu (1988). Resampling Inference With Complex Survey Data. *Journal of the American Statistical Association,*83, 231-241.
  – Sitter (1992). Comparing Three Bootstrap Methods for Survey Data. *Canadian Journal of Statisitcs,*135-154.
  – Sitter (1992). A Resampling Procedure for Complex Survey Data. *Journal of the American Statistical Association,* 87, 755-765.