Sampling and Estimation

Day 4: Linearization Methods for Variance Estimation

Stefan Zins[1] and Matthias Sand[2]

September 8, 2015

[1]Stefan.Zins@gesis.org
[2]Matthias.Sand@gesis.org

In practice design based variance estimation will *not* be an option for most surveys.

In practice design based variance estimation will *not* be an option for most surveys. Why?

In practice design based variance estimation will *not* be an option for most surveys. Why? We do not know the π_{kl} and in some cases not even the π_k and even if we would, the response process and frame imperfections remain still unknown.

In practice design based variance estimation will *not* be an option for most surveys. Why? We do not know the π_{kl} and in some cases not even the π_k and even if we would, the response process and frame imperfections remain still unknown. Thus we have to resort to second best solutions:

- Model based estimation (as shown for the estimation of the *deff*)
- Estimation of approximated variances

What exactly has variance estimation to fulfill?

- Deliver adequate quality measures
- Consider practical issues

What exactly has variance estimation to fulfill?

- Deliver adequate quality measures
- Consider practical issues

But

- Can we apply general methods like for point estimation?

In variance estimation we face often following obstacles:

- Non-linear Statistics

In variance estimation we face often following obstacles:

- Non-linear Statistics
- Complex Designs (incl. non-response)

In variance estimation we face often following obstacles:

- Non-linear Statistics
- Complex Designs (incl. non-response)

Two of the commonly used methods to overcome these are

In variance estimation we face often following obstacles:

- Non-linear Statistics
- Complex Designs (incl. non-response)

Two of the commonly used methods to overcome these are

Linearization ; An approximation to the variance of the estimator is sought that is far easier to estimate.

In variance estimation we face often following obstacles:

- Non-linear Statistics
- Complex Designs (incl. non-response)

Two of the commonly used methods to overcome these are

Linearization ; An approximation to the variance of the estimator is sought that is far easier to estimate.

Resampling Methods ; The distribution of the estimator is simulated in order to estimate its variance.

In variance estimation we face often following obstacles:

- Non-linear Statistics
- Complex Designs (incl. non-response)

Two of the commonly used methods to overcome these are

Linearization ; An approximation to the variance of the estimator is sought that is far easier to estimate.

Resampling Methods ; The distribution of the estimator is simulated in order to estimate its variance.

The line between model and design based approaches is not so clear in some cases. There are resampling methods and linearization techniques that assume infinite population, i.e. they rely on models. We try to stay with the finite population *model*.

Recall StrSRS

In StrSRS we have the variance and variance estimator:

$$\mathrm{V}\left(\overline{y}_{\text{str}}\right)_{\text{SRS}} = \sum_{h=1}^{H} \frac{N_h - n_h}{N_h} \gamma_h^2 \frac{V_h^2}{n_h}$$

$$\widehat{\mathrm{V}}\left(\overline{y}_{\text{str}}\right)_{\text{SRS}} = \sum_{h=1}^{H} \frac{N_h - n_h}{N_h} \gamma_h^2 \frac{s_h^2}{n_h}$$

$$s_h^2 = \frac{1}{n_h - 1} \sum_{k \in s_h} \left(y_k - \overline{y}_h\right)^2$$

Recall StrSRS

In StrSRS we have the variance and variance estimator:

$$V(\bar{y}_{\mathrm{str}})_{\mathrm{SRS}} = \sum_{h=1}^{H} \frac{N_h - n_h}{N_h} \gamma_h^2 \frac{V_h^2}{n_h}$$

$$\widehat{V}(\bar{y}_{\mathrm{str}})_{\mathrm{SRS}} = \sum_{h=1}^{H} \frac{N_h - n_h}{N_h} \gamma_h^2 \frac{s_h^2}{n_h}$$

$$s_h^2 = \frac{1}{n_h - 1} \sum_{k \in s_h} (y_k - \bar{y}_h)^2$$

Most surveys in the social science use designs without replacement, thus even for SRS within strata we need to know the N_h, which are often not readily provided (e.g. for reasons of disclosure control).

Recall StrSRS

In StrSRS we have the variance and variance estimator:

$$V\left(\bar{y}_{\text{str}}\right)_{\text{SRS}} = \sum_{h=1}^{H} \frac{N_h - n_h}{N_h} \gamma_h^2 \frac{V_h^2}{n_h}$$

$$\hat{V}\left(\bar{y}_{\text{str}}\right)_{\text{SRS}} = \sum_{h=1}^{H} \frac{N_h - n_h}{N_h} \gamma_h^2 \frac{s_h^2}{n_h}$$

$$s_h^2 = \frac{1}{n_h - 1} \sum_{k \in s_h} (y_k - \bar{y}_h)^2$$

Most surveys in the social science use designs without replacement, thus even for SRS within strata we need to know the N_h, which are often not readily provided (e.g. for reasons of disclosure control). One approach would be to use the variance estimator for sampling with replacement instead:

$$\hat{V}\left(\bar{y}_{\text{str}}\right)_{\text{SRSWR}} = \sum_{h=1}^{H} \gamma_h^2 \frac{s_h^2}{n_h}$$

gesis
Leibniz Institute
for the Social Sciences

Most surveys in the social science use designs without replacement, thus even for SRS within strata we need to know the N_h, which are often not readily provided (e.g. for reasons of disclosure control). One approach would be to use the variance estimator for sampling with replacement instead:

$$\hat{\mathrm{V}}\left(\overline{y}_{\text{str}}\right)_{\text{SRSWR}} = \sum_{h=1}^{H} \gamma_h^2 \frac{s_h^2}{n_h}$$

Since $\mathrm{V}\left(\overline{y}_{\text{str}}\right)_{\text{SRS}} \leqslant \mathrm{V}\left(\overline{y}_{\text{str}}\right)_{\text{SRSWR}}$. we would on average over estimate the variance and thus have over conservative test results. (Note: $\mathrm{E}\left(s_h^2\right)_{SRS} > \mathrm{E}\left(s_h^2\right)_{SRSWR}$, but this effect should be smaller than that of the neglected finite population correction.)

gesis Leibniz Institute for the Social Sciences

$$V\left(\hat{\tau}_{\pi}\right) = \sum_{k\in\mathcal{U}}\sum_{l\in\mathcal{U}}\left(\pi_{kl} - \pi_k\pi_l\right)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} = \left(\check{\mathbf{y}}\right)^{\top}\boldsymbol{\Delta}\check{\mathbf{y}} \text{ and}$$

$$\widehat{V}\left(\hat{\tau}_{\pi}\right)_1 = \sum_{k\in s}\sum_{l\in s}\frac{\left(\pi_{kl} - \pi_k\pi_l\right)}{\pi_{kl}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} = \left(\check{\mathbf{y}}_s\right)^{\top}\check{\boldsymbol{\Delta}}_s\check{\mathbf{y}}_s \ ,$$

where $\boldsymbol{\Delta} = [\pi_{kl} - \pi_k\pi_l]_{k,l\in\mathcal{U}}$, $\check{\mathbf{y}} = [y_k\pi_k^{-1}]_{k,l\in\mathcal{U}}$ and their sample equivalents are $\check{\boldsymbol{\Delta}}_s = [\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}}]_{k,l\in s}$ and $\check{\mathbf{y}}_s = [y_k\pi_k^{-1}]_{k,l\in s}$.

For a fixed size design we may write the variance of $\hat{\tau}_\pi$ as

$$\mathrm{V}\left(\hat{\tau}_\pi\right) = -\frac{1}{2}\sum_{k\in\mathcal{U}}\sum_{l\in\mathcal{U}}\left(\pi_{kl} - \pi_k\pi_l\right)\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 ,$$

which can be estimated by

$$\widehat{\mathrm{V}}\left(\hat{\tau}_\pi\right)_2 = -\frac{1}{2}\sum_{k\in s}\sum_{l\in s}\frac{\left(\pi_{kl} - \pi_k\pi_l\right)}{\pi_{kl}}\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 .$$

Note that $\widehat{\mathrm{V}}\left(\hat{\tau}_\pi\right)_1 = \widehat{\mathrm{V}}\left(\hat{\tau}_\pi\right)_2 + (\breve{\mathbf{y}}_s^2)^\top \breve{\boldsymbol{\Delta}}_s \mathbf{1}_s$, where $\mathbf{1}_s$ is a vector of ones of length n. Hence, $\mathrm{V}\left(\widehat{\mathrm{V}}\left(\hat{\tau}_\pi\right)_1\right) \geqslant \mathrm{V}\left(\widehat{\mathrm{V}}\left(\hat{\tau}_\pi\right)_2\right)$, thus $\widehat{\mathrm{V}}\left(\hat{\tau}_\pi\right)_1$ should never be used for fixed size designs.

Practical Estimators for the Design Variance

Due to the double sum and the required knowledge of the π_{kl}'s, both variance estimators $\widehat{\mathrm{V}}\left(\hat{\tau}_\pi\right)_2$ and $\widehat{\mathrm{V}}\left(\hat{\tau}_\pi\right)_1$ are in practice not really applicable. In order to avoid calculating these double sums and the π_{kl}'s, several approximations have been proposed for a single sum variance estimator:

$$\widehat{\mathrm{V}}\left(\hat{\tau}_\pi\right)_{\text{approx}} = \sum_{k \in s} \hat{b}_k \hat{e}_k^2 \ ,$$

where

$$\hat{e}_k = \frac{y_k}{\pi_k} - \hat{B} \qquad \text{and} \qquad \hat{B} = \frac{\sum_{k \in s} \frac{y_k}{\pi_k} \hat{b}_k}{\sum_{k \in s} \hat{b}_k} \pi_k \ .$$

Practical Estimators for the Design Variance

Numerous choices can be found in the literature for $\hat{b}_k$ [Matei & Tillé, 2005]:

$$
{}_1\hat{b}_k = (1 - \pi_k)\frac{n}{n-1}, \qquad \text{[Hájek, 1981]}
$$

$$
{}_2\hat{b}_k = (1 - \pi_k)\left[1 - \sum_{k \in \mathcal{s}} \left(\frac{1 - \pi_k}{\sum_{l \in \mathcal{s}} (1 - \pi_l)}\right)^2\right]. \qquad \text{[Deville, 1999]}
$$

Practical Estimators for the Design Variance

An approximation that is implement with the `survey` package was published by Brewer (2002)

$$\widehat{\mathrm{V}}\left(\hat{\tau}_\pi\right)_{\text{brewer}} = \sum_{k \in \mathcal{s}} \left(\frac{1}{b_k^*} - \pi_k\right)\left(\frac{y_k}{\pi_k} - \frac{\hat{\tau}_\pi}{n}\right)^2 ,$$

Brewer (2002) presents the following (among others) choices for b_k^*

$$
\begin{aligned}
{}_1\hat{b}_k^* &= \frac{n-1}{n-\pi_k} , \\
{}_2\hat{b}_k^* &= \frac{n-1}{n - n^{-1}\sum_{k\in\mathcal{U}} \pi_k^2} .
\end{aligned}
$$

Not that ${}_2\hat{b}_k^*$ includes a statistic based on the whole population. The version with ${}_1\hat{b}_k^*$ is implement with the `survey` package.

We want to estimate the total number of men in 2004 in Belgium by sampling municipalities proportional to their total population in 2003.

```
library(sampling)
library(survey)
data(belgianmunicipalities)  #load the population
DATA    <- belgianmunicipalities[-2,] #we remove a very large municipality
n       <- 25                #a 4.3% sample
DATA$IP <- inclusionprobabilities(DATA$Tot03,n)
##compute the joint inclusion probabilities (for Sampford sampling)
IPkl    <- UPsampfordpi2(DATA$IP) #this can take some seconds
##Covariance matrix of the sample indicator
DELTA   <- IPkl - DATA$IP%*%t(DATA$IP)
##The tue variance is
V <- with( DATA ,t(cbind(Men04/IP))%*%DELTA%*%(cbind(Men04/IP)) )
sqrt(V)

##            [,1]
## [1,] 20521.55
```

For the estimation we use the `survey` package and define `svydesign` objects with different variance estimators.

```
library(Matrix)  #needed for 'svydesign' objects with 'pps=ppsmat()'
set.seed(5675)
## select a sampling using the Sampford's method
s <- UPsampford(DATA$IP)
DATA.s <- DATA[s == 1, ]
IPkl.s <- IPkl[s == 1, s == 1]
## Variances without replacement: Vest_1
dpps_br <- svydesign(id = ~1, fpc = ~IP, data = DATA.s, pps = "brewer")
dpps_ht <- svydesign(id = ~1, fpc = ~IP, data = DATA.s, pps = ppsmat(IPkl.s))
## Vest_2
dpps_yg <- svydesign(id = ~1, fpc = ~IP, data = DATA.s, pps = ppsmat(IPkl.s),
    variance = "YG")
## The with replacement approximation
dpps_wr <- svydesign(id = ~1, probs = ~IP, data = DATA.s)
## Estimation:
V_1 <- svytotal(~Men04, dpps_ht)
V_2 <- svytotal(~Men04, dpps_yg)
br <- svytotal(~Men04, dpps_br)
wr <- svytotal(~Men04, dpps_wr)
```

These are the results from our estimation:

Table: Estimated Standard Errors and their Relative Error

	Men04	se	rel.er
$\widehat{\mathrm{V}}\left(\hat{\tau}_\pi\right)_1$	4864204.79	167069.45	7.14
$\widehat{\mathrm{V}}\left(\hat{\tau}_\pi\right)_2$	4864204.79	22567.25	0.10
$\widehat{\mathrm{V}}\left(\hat{\tau}_\pi\right)_{\mathrm{brewer1}}$	4864204.79	22965.78	0.12
$\widehat{\mathrm{V}}\left(\hat{\tau}_\pi\right)_{\mathrm{wr}}$	4864204.79	24320.79	0.19

The variance estimate from $\widehat{\mathrm{V}}\left(\hat{\tau}_\pi\right)_1$ is way off, compared to the true value. The approximate method by Brewer $\widehat{\mathrm{V}}\left(\hat{\tau}_\pi\right)_{\mathrm{brewer1}}$ is not much different from $\widehat{\mathrm{V}}\left(\hat{\tau}_\pi\right)_2$, which is good, considering that we didn't need the π_{kl}'s for this. As can be expect the with replacement approximation $\widehat{\mathrm{V}}\left(\hat{\tau}_\pi\right)_{\mathrm{wr}}$ is slightly above $\widehat{\mathrm{V}}\left(\hat{\tau}_\pi\right)_2$ and $\widehat{\mathrm{V}}\left(\hat{\tau}_\pi\right)_{\mathrm{brewer1}}$.

The estimator we use are often non-linear, like $\overline{y}_w$. Which is a problem, in particular for variance estimation. Mind that:

$$\mathrm{E}\left(\frac{\hat{\tau}_{y_1\,w}}{\hat{\tau}_{y_2\,w}}\right) \approx \frac{\mathrm{E}\left(\hat{\tau}_{y_1\,w}\right)}{\mathrm{E}\left(\hat{\tau}_{y_2\,w}\right)}$$

$$\mathrm{V}\left(\frac{\hat{\tau}_{y_1\,w}}{\hat{\tau}_{y_2\,w}}\right) \neq \frac{\mathrm{V}\left(\hat{\tau}_{y_1\,w}\right)}{\mathrm{V}\left(\hat{\tau}_{y_2\,w}\right)}$$

Leibniz Institute for the Social Sciences

In case our statistic of interest θ can be displayed as a function f of Q totals

$$\theta = f(\boldsymbol{\tau}) \ ,$$

with $\boldsymbol{\tau} = (\tau_{x_1}, \ldots, \tau_{x_q}, \ldots, \tau_{x_Q})^\top$ and $\tau_{x_1} = \sum_{k \in \mathcal{U}} x_{kq}$. We estimate θ by

$$\hat{\theta} = f(\hat{\boldsymbol{\tau}}) \ ,$$

with $\hat{\boldsymbol{\tau}} = (\hat{\tau}_{x_1}, \ldots, \hat{\tau}_{x_q}, \ldots, \hat{\tau}_{x_Q})^\top$.

If f is continuously differentiable up to second order between $\boldsymbol{\tau}$ and $\hat{\boldsymbol{\tau}}$ we can use the Taylor series of estimator $\hat{\theta}$ to obtain a linearized version of it

$$\hat{\theta} = \theta + \sum_{q=1}^{Q} \left[\frac{\partial f(t_1, \ldots, t_Q)}{\partial t_q} \right]_{\mathbf{t}=\boldsymbol{\tau}} \left(\hat{\tau}_{x_q} - \tau_{x_q} \right) + R(\hat{\boldsymbol{\tau}}, \boldsymbol{\tau}) \ ,$$

where

$$R(\hat{\boldsymbol{\tau}}, \boldsymbol{\tau}) = \frac{1}{2} \sum_{q=1}^{Q} \sum_{p=1}^{Q} \left[\frac{\partial^2 f(t_1, \ldots, t_Q)}{\partial t_q \partial t_p} \right]_{\mathbf{t}=\ddot{\boldsymbol{\tau}}} (\hat{\tau}_{x_q} - \tau_{x_q})(\hat{\tau}_{x_p} - \tau_{x_p})$$

and $\ddot{\boldsymbol{\tau}}$ is between $\hat{\boldsymbol{\tau}}$ and $\boldsymbol{\tau}$. In most application the remainder term R is ignored for large enough sample sizes.

Thus we can approximate the variance of $\hat{\theta}$ by

$$\begin{aligned} \mathsf{V}\left(\hat{\theta}\right) &\approx \mathsf{V}\left(\sum_{q=1}^{Q}\left[\frac{\partial f(t_1,\ \ldots,\ t_Q)}{\partial t_q}\right]_{\mathbf{t}=\boldsymbol{\tau}}\hat{\tau}_{x_q}\right) \\ &= \sum_{q=1}^{Q} a_q^2 \mathsf{V}\left(\hat{\tau}_{x_q}\right) + 2\sum_{q=1}^{Q}\sum_{\substack{p=1\\p<q}}^{Q} a_q a_p \mathsf{COV}\left(\hat{\tau}_{x_q},\ \hat{\tau}_{x_p}\right)\ , \end{aligned}$$

with $a_q = \left[\frac{\partial f(t_1,\ \ldots,\ t_Q)}{\partial t_q}\right]_{\mathbf{t}=\boldsymbol{\tau}}$.

For $\hat{\boldsymbol{\tau}} = (\hat{\tau}_{x_1 w}, \ldots, \hat{\tau}_{x_q w}, \ldots, \hat{\tau}_{x_Q w})^\top$ Woodruff (1971) proposes the transformation of x_{kq}

$$z_k = \sum_{q=1}^{Q} a_q x_{kq}$$

and use the following expression as an approximate variance of $\hat{\theta}$

$$\mathrm{V}\left(\hat{\theta}\right) \approx \mathrm{V}\left(\sum_{k \in \mathcal{s}} w_k z_k\right) .$$

This approximation is far more convenient to estimate then to estimate all the different variances and covariance in the above formula separately. z_k is also sometimes called the linearized variable.

Suppose we want to estimate $\theta = \frac{\tau_y}{N}$ the population mean of variable $\mathcal{Y}$. To do this we use estimator $\overline{y}_w = \overline{y}_\pi$, i.e. we set $w_k = d_k$ for all $k \in \mathcal{s}$. Now we have

$$\hat{\theta} = f(\boldsymbol{\tau}) = f((\hat{\tau}_{y,\pi}, \hat{\tau}_{x,\pi})) = \frac{\hat{\tau}_{y,\pi}}{\hat{\tau}_{x,\pi}} ,$$

were $\hat{\tau}_{y,\pi} = \sum_{k\in\mathcal{s}} d_k y_k$ and $\hat{\tau}_{x,\pi} = \sum_{k\in\mathcal{s}} d_k$, because $x_k = 1$ for all $k \in \mathcal{U}$. Our estimator is a function of Q=2 totals, and we have

$$a_1 = \frac{1}{\tau_x} \qquad a_2 = -\frac{\tau_y}{\tau_x^2} = -\frac{\theta}{\tau_x}$$

$$z_k = a_1 y_k + a_2 x_k = \frac{1}{\tau_x}(y_k - \theta)$$

To estimate the approximate variance $\mathrm{V}\left(\sum_{k\in\mathcal{s}} d_k z_k\right)$ we need estimates for the z_k's, because they involve unknown statistics τ_x and θ.

$$\hat{z}_k = \frac{1}{\hat{\tau}_{x,\pi}}\left(y_k - \overline{y}_\pi\right)$$

Our variance estimator would be $\widehat{\mathrm{V}}\left(\sum_{k\in\mathcal{s}} d_k \hat{z}_k\right)$ for which $\widehat{\mathrm{V}}\left(.\right)_1$ or $\widehat{\mathrm{V}}\left(.\right)_2$ could be used or an estimator for an approximate design variance.
Note that the variances of the $\hat{z}_k$'s (and the covariances between them) is often thought to be negligible and is therefore usually not considered.

Variance Estimation for Calibration Estimators

The variance of $\hat{\tau}_w$ with GREG weights can be approximated by

$$\mathrm{V}\left(\hat{\tau}_w\right) \approx \mathrm{V}\left(\sum_{k\in\mathcal{s}} d_k e_k\right) \quad ,$$

where the e_k's are the residuals of regression $\hat{y}_k = \mathbf{x}_k^\top \boldsymbol{\beta}$ with

$$\boldsymbol{\beta} = \left(\sum_{k\in\mathcal{U}} c_k \mathbf{x}_k \mathbf{x}_k^\top\right)^{-1} \left(\sum_{k\in\mathcal{U}} c_k \mathbf{x}_k y_k\right) \quad .$$

To estimate $\mathrm{V}\left(\sum_{k\in\mathcal{s}} d_k e_k\right)$ we need to estimates $\hat{e}_k$ for the residuals e_k's. We can obtain them using $\hat{e}_k = y_k - \mathbf{x}_k^\top \hat{\boldsymbol{\beta}}$ where $\hat{\boldsymbol{\beta}}$ is the vector of the estimated regression coefficients used in solving the calibration problem.
Finally we estimate the variance of $\hat{\tau}_w$ by using an estimator $\widehat{\mathrm{V}}\left(\sum_{k\in\mathcal{s}} d_k \hat{e}_k\right)$ that is appropriate for the present sampling design.

Variance Estimation for Calibration Estimators

Deville (1999) showed that for the estimation of non-linear estimator that use calibration weights, like $\overline{y}_w$ with GREG weights, the same technique can be used to obtain a variance estimator. Here the residuals e_k are from the regression of the linearized variable z_k on the auxiliary variables. Then we have

$$\widehat{\mathrm{V}}\left(\sum_{k \in \mathcal{s}} d_k \hat{e}_k^*\right) \ ,$$

with $\hat{e}_k^* = \hat{z}_k - \mathbf{x}_k^\top \hat{\beta}$.

Deville (1999) showed that for the estimation of non-linear estimator that use calibration weights, like $\overline{y}_w$ with GREG weights, the same technique can be used to obtain a variance estimator. Here the residuals e_k are from the regression of the linearized variable z_k on the auxiliary variables. Then we have

$$\widehat{\mathrm{V}}\left(\sum_{k \in s} d_k \hat{e}_k^*\right) ,$$

with $\hat{e}_k^* = \hat{z}_k - \mathbf{x}_k^\top \hat{\beta}$.

With raking weights variance estimation can be done in a similar way, although the used regression is different from the GREG case.

Taylor-linearization with Calibration Weights

Recall the calibration example to estimate total expenditures of hospitals. We know want to estimate the mean using $\overline{y}_w$

```
set.seed(428274453)
sam <- UPsampford(IP)          # now we use Sampford sampling
sam.dat     <- smho.[sam==1, ]
sam.dat$IP <- IP[sam==1]
#1. build a 'design' object
sam.dsgn <-
  svydesign(ids = ~1,          # no clusters
            data = sam.dat, # the sample data
            fpc = ~IP,       # inclusion probabilities
            pps = "brewer") # handeling of 2. order inc.prob.
lmod2 <- lm(EXPTOTAL ~ SEENCNT + EOYCNT + hosp.type:BEDS, data=smho.)
pop.tots <- colSums(model.matrix(lmod2))
#2. use 'calibrate' to compute GREG weights
sam.cal <-
  calibrate(design = sam.dsgn,
            formula = ~ SEENCNT + EOYCNT + hosp.type:BEDS,
            population = pop.tots,
            calfun='linear' )
```

Taylor-linearization with Calibration Weights

Leibniz Institute for the Social Sciences

```
# Estimation with design weights
svymean(~EXPTOTAL, design = sam.dsgn)

##                mean      SE
## EXPTOTAL 14427717 1618361

# and with calibrated weights
svymean(~EXPTOTAL, design = sam.cal)

##                mean     SE
## EXPTOTAL 13480447 920393
```

For `sam.cal` the reported standard errors are estimates of the linearized variance using Brewer's approximation.

Non-linear and Non-differentiable Estimators

There are also those estimators that are non-linear and non-differentiable and thus cannot be linearized using a Taylor series.

Non-linear and Non-differentiable Estimators

There are also those estimators that are non-linear and non-differentiable and thus cannot be linearized using a Taylor series.

A prominent case are estimators for quantiles, like the median.

Non-linear and Non-differentiable Estimators

There are also those estimators that are non-linear and non-differentiable and thus cannot be linearized using a Taylor series.

- A prominent case are estimators for quantiles, like the median.
- But also poverty measures such as the at-risk-of poverty rate and inequality measures as the GINI coefficient and the quintile share ratio.

Non-linear and Non-differentiable Estimators

There are also those estimators that are non-linear and non-differentiable and thus cannot be linearized using a Taylor series.

- A prominent case are estimators for quantiles, like the median.
- But also poverty measures such as the at-risk-of poverty rate and inequality measures as the GINI coefficient and the quintile share ratio.

However it is possible by using the concept of *influence function* or *estimation equations* to obtain a linearized variable for these estimator, too. For example, as the linearized variable of the median we could use

$$z_k = -\frac{1}{NF'_N[\mathrm{MED}(M)]}\left(\mathbb{1}[y_k \leqslant \mathrm{MED}(M)] - 0.5\right) ,$$

where $\mathrm{MED}(M)$ is the median, $F_N(y) = \frac{\sum_{k\in\mathcal{U}} \mathbb{1}[y_k \leqslant y]}{N}$ is the empirical distribution function of variable $\mathcal{Y}$ at point y and F'_N its first derivative. $\mathbb{1}[.]$ is a indicator function assuming the value of 1 if the argument is true and 0 otherwise. The `svyquantile()` function from the `survey` package has such a method implemented.

K. Brewer.
Combined Survey Sampling Inference.
Arnold, 2002.

J.C. Deville.
Variance Estimation for Complex Statistics and Estimators: Linearization and Eesidual Techniques.
Survey Methodology, 1999.

J. Hájek.
Sampling from a Finite Population.
Dekker, 1981.

A. Matei, Y. Tillé.
Evaluation of Variance Approximations and Estimators in Maximum Entropy Sampling with Unequal Probability and Fixed Sample Size.
Journal of Official Statistics, 2005.

R.S. Woodruff.
A Simple Method for Approximating the Variance of a Complicated Estimate.
Journal of the American Statistical Association, 1971.