Surveymetodik
Föreläsning 3

Måns Magnusson

Avd. Statistik, LiU

Section 1

Teori för urvalsundersökningar

- Olika inferensteorier för statistisk inferens för urvalsundersökningar

	Design based	**Model based**	**Bayesian**
Randomness	Home-made	Given by nature	Subjective/Rationality axioms
Main focus	Population	Parameters	Population
Parameters	Population values	Unknown/unobservable	Do not exist, but useful
Inference	Frequency based	Frequency based	Probability-based
Output	Point-estimates/CI	Point-estimates/CI	Posterior distributions
Possible use	Not my problem!	Not my problem!	Interface with decisions

Källa: Thorburn (2009) (med vissa förändringar)

Teori I

- Vår estimator / skattningsmetod för $\bar{y}_{\mathcal{U}}$ är:

$$\hat{\bar{y}}_{\mathcal{U}} = \bar{y}_{\mathcal{S}} = \frac{1}{n} \sum_{i \in \mathcal{S}} y_i$$

- **Designbaserat** förväntat värde:

$$E(\hat{\bar{y}}_{\mathcal{U}}) = \sum_{i=k}^{K} P(\mathcal{S}_k) \hat{\bar{y}}_{\mathcal{U}}$$

- **Designbaserad** varians

$$Var(\hat{\bar{y}}_{\mathcal{U}}) = \sum_{i=k}^{K} P(\mathcal{S}_k) \left[\hat{\bar{y}}_{\mathcal{U}} - E(\hat{\bar{y}}_{\mathcal{U}})\right]^2$$

- Vi är ofta intresserade av att minimera det totala felet:

$$MSE(\hat{\bar{y}}_{\mathcal{U}}) = E\left[(\hat{\bar{y}}_{\mathcal{U}} - \bar{y}_{\mathcal{U}})^2\right] = \sum_{k=1}^{K} P(\mathcal{S}_k) \cdot (\hat{\bar{y}}_{\mathcal{U}} - \bar{y}_{\mathcal{U}})^2$$

- Detta fel kan delas upp i två delar $(*)$
  - Bias: $Bias(\hat{\bar{y}}_{\mathcal{U}})^2 = (E(\hat{\bar{y}}_{\mathcal{U}}) - \bar{y}_{\mathcal{U}})^2$
  - Varians: $Var(\hat{\bar{y}}_{\mathcal{U}})$
  - Härledning finns i Lohr (2009, s. 31)

- Samma exempel som tidigare där $\bar{y}_{\mathcal{U}} = 17.2$ och $Var(y) = 98.886$ då n =6, $N = 15$ och $K = 5005$.

	Obs.1	Obs.2	Obs.3	Obs.4	Obs.5	Obs.6	P_S	y_hat
1	29	21	23	3	22	30	0.0002	21.3
2	29	21	23	3	22	24	0.0002	20.3
3	29	21	23	3	22	6	0.0002	17.3
4	29	21	23	3	22	15	0.0002	18.8
5	29	21	23	3	22	2	0.0002	16.7
6	29	21	23	3	22	4	0.0002	17.0
7	29	21	23	3	22	10	0.0002	18.0

$$E(\hat{\bar{y}}_{\mathcal{U}}) = \sum_{i=k}^{K} P(\mathcal{S}_k)\hat{\bar{y}}_{\mathcal{U}} = 17.2$$

$$Var(\hat{\bar{y}}_{\mathcal{U}}) = \sum_{i=k}^{K} P(\mathcal{S}_k)\left[\hat{\bar{y}}_{\mathcal{U}} - E(\hat{\bar{y}}_{\mathcal{U}})\right]^2 = 9.889(*)$$

- Vi vill att skattningarna ska vara **design-väntevärdesriktiga**
- Vi vill visa att:

$$
\begin{aligned}
E(\bar{y}_{\mathcal{S}}) &= \bar{y}_{\mathcal{U}} \\
Var(\bar{y}_{\mathcal{S}}) &= \left(1 - \frac{n}{N}\right) \frac{S_{\mathcal{U}}^2}{n} \\
E(s^2) &= S_{\mathcal{U}}^2
\end{aligned}
$$

- Vi gör detta genom att införa en egen slumpvariabel (som vi styr själva)

Indikatorvariabeln Z_i

- Vi behöver introducera vår egen "hemmagjorda" slump:

$$\begin{aligned} Z_i &= P(i \in \mathcal{S}) \sim Bernoulli(\pi_i) \\ Z &= \{0, 1\} \end{aligned}$$

Z_i brukar kallas **indikatorvariabel** och har följande egenskaper $(*)$:

$$E(Z_i) = E(Z_i^2) = P(Z_i = 1) = \frac{n}{N}$$

$$Var(Z_i) = \frac{n}{N}\left(1 - \frac{n}{N}\right)$$

$$E(Z_i Z_j) = \frac{n}{N}\frac{n-1}{N-1}$$

$$Cov(Z_i, Z_j) = -\frac{1}{N-1}\left(1 - \frac{n}{N}\right)\frac{n}{N}$$

Indikatorvariabeln Z_i II

- Med indikatorvariabeln kan vi visa att $(*)$

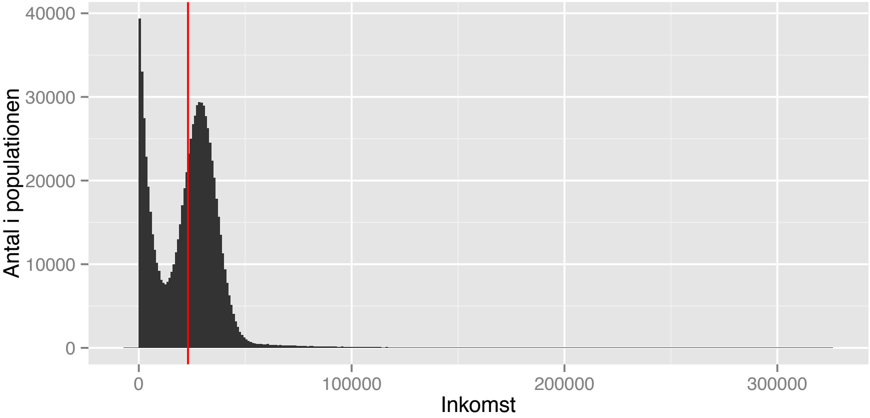
$$\begin{aligned} E(\bar{y}_{\mathcal{S}}) &= \bar{y}_{\mathcal{U}} \\ Var(\bar{y}_{\mathcal{S}}) &= \left(1 - \frac{n}{N}\right) \frac{S_{\mathcal{U}}^2}{n} \\ E(s^2) &= S_{\mathcal{U}}^2 \end{aligned}$$

där

$$S_{\mathcal{U}}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{y}_{\mathcal{U}})^2$$

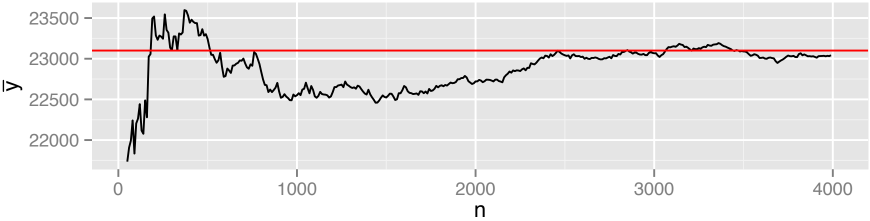
- Härledningar finns också i Lohr (2009, s. 51-54)

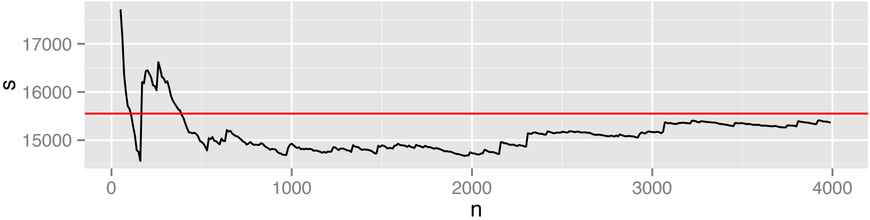
- Vi har en befolkning som består av 815599 personer.

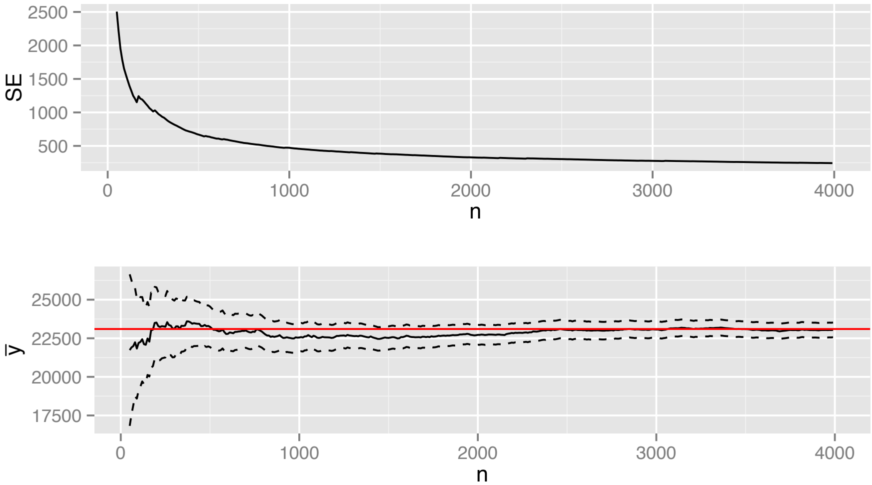


- Den 'sanna' genomsnittsinkomsten ($\bar{y}_{\mathcal{U}}$) är 23100.33 och den 'sanna' standardavvikelsen ($S_{\mathcal{U}}$) är 15552.189.

Exempel: Estimation av $\bar{y}_{\mathcal{U}}$ och $S_{\mathcal{U}}$






2500
2000
1500
1000
500
SE
0
1000
2000
3000
4000
n
25000
22500
20000
17500
ȳ
0
1000
2000
3000
4000
n

Referenser

Lohr, S., 2009. Sampling: design and analysis, 2nd Edition. Thomson.

Thorburn, D., 2009. Bayesian methods in survey sampling, preliminary version, Workshop on Survey Sampling, August 23-27 2009, Kiev.