Stikprøveteori 1. del

Hans Bay

mandag 16. juli 2018

Stikprøveteori definitioner
Univers $Y_1, Y_2, Y_3, ...Y_N$

$$\overline{Y} = \frac{1}{N}\sum_{j=1}^{N} Y_i$$

$$S^2 = \frac{1}{N-1}\sum_{j=1}^{N}(Y_j\text{-}\overline{Y})^2$$

stikprøve $y_1,\ y_2,\ y_3, ...\ y_n$

$$\overline{y}_{si} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$V(\overline{y}_{si}) = \frac{(N-n)}{N}\frac{1}{n}S^2$$

tre dele: endelighed, stikprøve, varians i univers.

Stikprøveteori definitioner
Univers $Y_1, Y_2, Y_3, ...Y_N$

stikprøve $y_1,\ y_2,\ y_3, ...\ y_n$

simpel tilfældig="repræsentativ":

alle mulige stikprøvekombinationer er lige sandsynlige

$\binom{N}{n}$ dette er alle mulige kombinationer

Lille bitte eksempel på auditoriet

Univers: 1,2,3,4 dvs. N=4

.

udregn $\overline{Y} = \frac{1}{N}\sum_{j=1}^{N} y_j$ og $S^2 = \frac{1}{N-1}\sum_{j=1}^{N}(Y_i\text{-}\overline{Y})^2$

.

stikprøve er n=2

Hvor mange mulige stikprøver ?

udregn gennemsnittet og varians for disse stikprøver

udregn middelværdien blandt de mulige udtrukne stikprøver

Stikprøveteori grundlag

$I_j = 1$ hvis nr i er udtaget til stikprøven ellers nul
$P(I_j = 1) = \frac{n}{N}$

$E(I_j) = P(I_j = 1) = \frac{n}{N}$

$V(I_j) = \frac{n}{N}\frac{(N-n)}{N} = \frac{n(N-n)}{N^2}$
$P(I_j = 1) = \frac{\text{gunstige}}{\text{mulige}} = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}$

$E(I_j) = 0 * P(I_j = 0) + 1 * P(I_j = 1) = \frac{n}{N}$

$I_j^2 = I_j \quad V(I_j) = E(I_j^2) - [E(I_j)]^2$

$= \frac{n}{N} - (\frac{n}{N})^2 = \frac{n(N-n)}{N^2}$

- $E(\overline{y}_{si}) = E(\frac{1}{n}\sum_{i=1}^{n} y_i) = E(\frac{1}{n}\sum_{j=1}^{N} I_j y_j) =$
- $\frac{1}{n}\sum_{j=1}^{N} E(I_j y_j) = \frac{1}{n}\sum_{j=1}^{N} y_j E(I_j) =$
- $\frac{1}{n}\sum_{j=1}^{N} y_j \frac{n}{N} = \frac{1}{n}\sum_{j=1}^{N} y_j = \overline{} Y$
- $V(\overline{y}_{si}) = V(\frac{1}{n}\sum_{j=1}^{N} I_j y_j)$

Egenskaber

$$E(\overline{y}_{si}) = \overline{\quad Y}$$

$$V(\overline{y}_{si}) = \frac{(N-n)}{N}\frac{1}{n}S^2$$

$$E(s^2) \ = \ S^2$$

Estimerer variansen i universet ved variansen i stikprøven
$$\widehat{S^2} = s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_i\text{-}\overline{y}_{si})^2$$

eksempel kommuner s. 46

N=275 kommuner i gode gamle dage
Y_j = udgift til vejvæsen i 1.000 kr. pr. indbygger (i 1998) for kommune nr. j

$$\overline{Y} = \frac{1}{275} \sum_{j=1}^{275} Y_i = 0,630746$$

$$S^2 = \frac{1}{275-1} \sum_{j=1}^{N} (Y_i\text{-}\ 0,630746)^2 = 0,023054$$

$$S = \sqrt{S^2} = 0,152$$

Beregn usikkerhed (og tilhørende 95% cf-interval) når man udtager en stikprøve på

n=5, 10,20 og 50

alternativ variation

Alternativ variation, dvs. at $y_j = 0$ eller 1

P$=$ andel af 1'taller $= \overline{\ Y}$

$S^2 = \frac{N}{N-1}P(1-P) \approx P(1-P)$

$\widehat{p} = \overline{y}_{si} = \frac{1}{n}\sum_{j=1}^{n} y_i =$ andel af 1'taller i stikprøven

$\widehat{S^2} = s^2 = \frac{n}{n-1}\widehat{p}(1-\widehat{p})$ (beregnet i stikprøven)

Varians beregninger i alternativt tilfælde

$y_j = 1$ eller 0. dermed $y_j = y_j^2$

$$(N-1)S^2 = \sum_{j=1}^{N}(Y_j\text{-}\overline{Y})^2 = \sum_{j=1}^{N} Y_j^2 + \sum_{j=1}^{N} \overline{Y}^2 - 2\overline{Y}\sum_{j=1}^{N} Y_j =$$

$$\sum_{j=1}^{N} Y_j + N\overline{Y}^2 - 2\overline{Y}N\overline{Y} \quad \text{brug } P = \overline{Y}$$

$$= NP + NP^2 - 2NP^2 = N(P + P^2 - 2P^2) = NP(1-P)$$

$$S^2 = \frac{N}{N-1}P(1-P)$$

Varians beregninger i alternativt tilfælde

$V(\widehat{p}) = V(\overline{y}_{si}) = \frac{(N-n)}{N}\frac{1}{n}S^2 = \frac{(N-n)}{N}\frac{1}{n}\frac{N}{N-1}P(1-P) = \frac{(N-n)}{N-1}\frac{1}{n}P(1-P)$ (hypergeometrisk)

$\widehat{V(\widehat{p})} = \frac{(N-n)}{N}\frac{1}{n}\widehat{S^2} = \frac{(N-n)}{N}\frac{1}{n}s^2 = \frac{(N-n)}{N}\frac{1}{n}\frac{n}{n-1}\widehat{p}(1-\widehat{p}) = \frac{(N-n)}{N}\frac{1}{n-1}\widehat{p}(1-\widehat{p})$

Øvelse se på Greens målinger

Vælg et parti i Greens opinion fra 3. marts 2017
og eftervis konfidensintervallet

Stratifikation

Strat	*Antal*	*Univers*	*stikpr.*	*stikpr.*	*sum*	*gns*	*varians*
		vægte	*antal*	*vægte*			
1	N_1	$W_1 = \frac{N_1}{N}$	n_1	$w_1 = \frac{n_1}{n}$	$Y_{1.}$	$\overline{Y_{1.}}$	S_1^2
2							
K	N_K	$W_K = \frac{N_K}{N}$	n_K	$w_K = \frac{n_K}{n}$	$Y_{K.}$	$\overline{Y_{K.}}$	S_K^2
	N	1	n	1	$Y = Y_{..}$	$-$	$-$

Variansanalyseopspaltningen

$$S^2 = \frac{1}{N-1}\sum_{j=1}^{N}(Y_i\text{-}\overline{Y})^2$$

$$(N-1)S^2 = \sum_{k=1}^{K}[(N_k-1)S_k^2 + N_k(\overline{Y}_{k.}-Y)^2]$$

$$\sum_{k=1}^{K}\sum_{m=1}^{N_k}(y_{km}-Y)^2 = \sum_{k=1}^{K}\sum_{m=1}^{N_k}(y_{km}-\overline{Y}_{k.}+\overline{Y}_{k.}-Y)^2 =$$

$$\sum_{k=1}^{K}\sum_{m=1}^{N_k}(y_{km}-\overline{Y}_{k.})^2 + \sum_{k=1}^{K}\sum_{m=1}^{N_k}(\overline{Y}_{k.}-Y)^2 +$$

$$\sum_{k=1}^{K}\sum_{m=1}^{N_k}2*(y_{km}-\overline{Y}_{k.})*(\overline{Y}_{k.}-Y)$$

$$\sum_{k=1}^{K}\sum_{m=1}^{N_k}(y_{km}-\overline{Y}_{k.})^2 = \sum_{k=1}^{K}(N_k-1)S_k^2$$

$$\sum_{k=1}^{K}\sum_{m=1}^{N_k}(\overline{Y}_{k.}-Y)^2 = \sum_{k=1}^{K}N_k(\overline{Y}_{k.}-Y)^2$$

$$\sum_{k=1}^{K} \sum_{m=1}^{N_k} 2 * (y_{km} - \overline{Y}_{k.}) * (\overline{Y}_{k.} - Y) = 2 \sum_{k=1}^{K} (\overline{Y}_{k.} - Y) \sum_{m=1}^{N_k} (y_{km} - \overline{Y}_{k.})$$

$$\sum_{m=1}^{N_k} (y_{km} - \overline{Y}_{k.}) = 0$$

Stratifikation

$$(N-1)S^2 = \sum_{k=1}^{K} [(N_k - 1)S_k^2 + N_k(\overline{Y}_{k.} - Y)^2]$$

$$(N-K)S_i^2 = \sum_{k=1}^{K} \sum_{m=1}^{N_k} (y_{km} - \overline{Y}_{k.})^2 = \sum_{k=1}^{K} (N_k - 1)S_k^2$$

$$(K-1)S_m^2 = \sum_{k=1}^{K} N_k(\overline{Y}_{k.} - \overline{Y})^2$$

$$(N-1)S^2 = (N-K)S_i^2 + (K-1)S_m^2$$

Stratifikation lille eksempel

Stratum	*Antal*	*sum*	*gns*	*varians*
	N_K		$\overline{Y}_{k.}$	S_k^2
4, 7, 10	3	21	7	18/2
5, 9, 11, 15	4	40	10	52/3
2, 5 , 5 , 4	4	16	4	6/3
sum	11	77	–	–

$$\overline{Y} = \sum_{k=1}^{K} \sum_{m=1}^{N_k} y_{km} = \frac{77}{11} = 7$$

$$\overline{Y} = \sum_{k=1}^{K} N_k \overline{Y}_{k.} = \frac{1}{11}[3 * 7 + 4 * 10 + 4 * 4] = 7$$

Stratifikation lille eksempel

$$S_i^2 = \frac{1}{11-3}\sum_{k=1}^{3}(N_k - 1)S_k^2 = 9,5$$

$$S_m^2 = \frac{1}{3-1}\sum_{k=1}^{3} N_k(\overline{Y}_{k.} - 7)^2 = 36$$

$$S^2 = \frac{1}{11-1}\sum_{j=1}^{11}(Y_j\text{- } 7)^2 = \frac{148}{10}$$

$$(N-1)S^2 = (N-K)S_i^2 + (K-1)S_m^2$$

$$(11-1) * \frac{148}{10} = (11-3) * 9,5 + (3-1) * 36$$

Stratifikation

$$E(\overline{y_{k.}}) = \overline{Y_{k.}}$$

$$V(\overline{y_{k.}}) = \frac{(N_k - n_k)}{N_k} \frac{1}{n_k} S_k^2$$

$$\overline{y}_{strat} = \sum_{k=1}^{K} W_k \overline{y_{k.}}$$

$$E(\overline{y}_{strat}) = \overline{\ \ Y}$$

$$V(\overline{y}_{strat}) = \sum_{k=1}^{K} W_k^2 \frac{N_k - n_k}{N_k} \frac{1}{n_k} S_k^2$$

side 140

Stratum	*Antal*	vægte	*sum*	*gns*	*varians*
	N_K	W_k		$\overline{Y}_{k.}$	S_k^2
1,2,3	3	$\frac{1}{4}$	6	2	1
4,5,6	3	$\frac{1}{4}$	15	5	1
7,8,9	3	$\frac{1}{4}$	24	8	1
10,11,12	3	$\frac{1}{4}$	33	11	1
sum	12	1	78	-	-

$$\overline{Y}=\frac{1}{12}\sum_{j=1}^{12} y_j = \frac{1}{12}\sum_{j=1}^{12} j = \frac{78}{12} = 6,5$$

$$S^2 = \frac{1}{12-1}\sum_{j=1}^{12}(j-6,5)^2 = 13$$

$$V(\overline{y}_{si}) = \frac{(N-n)}{N}\frac{1}{n}S^2 = \frac{12-4}{12}\frac{1}{4}13 = \frac{13}{6}$$

$$V(\overline{y}_{strat}) = \sum_{k=1}^{4} W_k^2 \frac{N_k-n_k}{N_k}\frac{1}{n_k}S_k^2 = \sum_{k=1}^{4}(\frac{1}{4})^2\frac{(3-1)}{3}\frac{1}{1}1 = \frac{1}{6}$$

Stratum	*Antal*	vægte	*sum*	*gns*	*varians*
	N_K	W_k		$\overline{Y}_{k.}$	S_k^2
1,5,9	3	$\frac{1}{4}$	15	5	16
2,6,10	3	$\frac{1}{4}$	18	6	16
3,7,11	3	$\frac{1}{4}$	21	7	16
4,8,12	3	$\frac{1}{4}$	24	8	16
sum	12	1	78	-	-

$$\overline{Y}=\frac{1}{12}\sum_{j=1}^{12} y_j = \frac{1}{12}\sum_{j=1}^{12} j = \frac{78}{12} = 6,5$$

$$S^2 = \frac{1}{12-1}\sum_{j=1}^{12}(j-6,5)^2 = 13$$

$$V(\overline{y}_{si}) = \frac{(N-n)}{N}\frac{1}{n}S^2 = \frac{12-4}{12}\frac{1}{4}13 = \frac{13}{6}$$

$$V(\overline{y}_{strat}) = \sum_{k=1}^{4} W_k^2 \frac{N_k-n_k}{N_k}\frac{1}{n_k}S_k^2 = \sum_{k=1}^{4}(\frac{1}{4})^2\frac{(3-1)}{3}\frac{1}{1}16 = \frac{16}{6}$$

S. 142 kommuner

Stratum	*Antal*		*gns*	*varians*	stik
	N_K	W_k	$\overline{Y}_{k.}$	S_k^2	n_k
HR	50	0,18		0,0174	2
øerne	84	0,31		0,0177	3
Jylland	141	0,51		0,0258	5
sum	275	1,00	–	–	10

$$\overline{Y} = \frac{1}{275}\sum_{j=1}^{275} Y_i = 0,630746$$

$$S^2 = \frac{1}{275-1}\sum_{j=1}^{N}(Y_i\text{- } 0,630746)^2 = 0,023054 = (0,152)^2$$

$$V(\overline{y}_{si}) = \frac{(275-10)}{275}\frac{1}{10}(0,152)^2 = (0,047)^2$$

Udregn varians for stratificeret stikprøve

S. 142 kommuner

Stratum	*Antal*		*gns*	*varians*	stik
	N_K	W_k	$\overline{Y}_{k.}$	S_k^2	n_k
HR	50	0,18		0,0174	2
øerne	84	0,31		0,0177	3
Jylland	141	0,51		0,0258	5
sum	275	1,00	–	–	10

$$\overline{Y} = \frac{1}{275} \sum_{j=1}^{275} Y_i = 0,630746$$

$$S^2 = \frac{1}{275-1} \sum_{j=1}^{N} (Y_i\text{- } 0,630746)^2 = 0,023054 = (0,152)^2$$

$$V(\overline{y}_{si}) = \frac{(275-10)}{275} \frac{1}{10} (0,152)^2 = (0,047)^2$$

$$V(\overline{y}_{strat}) = \sum_{k=1}^{3} W_k^2 \frac{N_k - n_k}{N_k} \frac{1}{n_k} S_k^2 = (0,046)^2$$

Proportional

$\mathrm{n}_k = nW_k$

$\overline{y}_p = \sum_{k=1}^{K} W_k \overline{y_{k.}} = \frac{1}{n} \sum_{i=1}^{n} y_i$ dvs. det "oprindelige" estimat, selvvejende

$V(\overline{y}_p) = \frac{N-n}{N} \frac{1}{n} \sum_{k=1}^{K} W_k S_k^2$
Sammenlignet med simpel tilfældig
$V(\overline{y}_{si}) = \frac{(N-n)}{N} \frac{1}{n} S^2$

optimal

$$\mathrm{n}_k = n\frac{W_k S_k}{\sum\limits_{k=1}^{K} W_k S_k}$$

$$V(\overline{y}_{opt}) = \frac{1}{n}(\sum_{k=1}^{K} W_k S_k)^2 - \frac{1}{N}\sum_{k=1}^{K} W_k S_k^2$$

Generelt gælder at

$$V(\overline{y}_{opt}) <= V(\overline{y}_p) <= V(\overline{y}_{si})$$

S. 142 o s. 166

Strat	*Antal*		*varians*			
	N_K	W_k	S_k^2	S_k	$W_k*\mathrm{S}_k$	$W_k*\mathrm{S}_k^2$
HR	50	0,18	0,0174	0,1319	0,0240	0,0032
øer	84	0,31	0,0177	0,1330	0,0406	0,0054
Jyll.	141	0,51	0,0258	0,1606	0,0824	0,0132
sum	275	1,00	–		0,1470	0,0218

$$\overline{Y} = \frac{1}{275}\sum_{j=1}^{275} Y_i = 0,630746$$

$$S^2 = \frac{1}{275-1}\sum_{j=1}^{N}(Y_i\text{- }0,630746)^2 = 0,023054 = (0,152)^2$$

$V(\overline{y}_{si}) = \frac{(275-25)}{275}\frac{1}{25}(0,152)^2 = (0,0290)^2$
Alloker en proportional og optimal stikprøve n=25
beregn varianser herfor

S. 142 o s. 166

	Antal		*varians*				*P*	*opt*
	N_K	W_k	S_k^2	S_k	$W_k * S_k$	$W_k * S_k^2$		
HR	50	0,18	0,0174	0,1319	0,0240	0,0032	4	4
øer	84	0,31	0,0177	0,1330	0,0406	0,0054	8	7
Jyll.	141	0,51	0,0258	0,1606	0,0824	0,0132	13	14
sum	275	1,00	–		0,1470	0,0218	25	25

$$S^2 = \frac{1}{275-1} \sum_{j=1}^{N} (Y_i\text{- } 0,630746)^2 = 0,023054 = (0,152)^2$$

$$V(\overline{y}_{si}) = \frac{(275-25)}{275} \frac{1}{25} (0,152)^2 = (0,0290)^2$$

$$V(\overline{y}_p) = \frac{N-n}{N} \frac{1}{n} \sum_{k=1}^{K} W_k S_k^2 = \frac{(275-20)}{275} \frac{1}{20} (0,0218) = (0,0282)^2$$

$$V(\overline{y}_{opt}) = \frac{1}{n} (\sum_{k=1}^{K} W_k S_k)^2 - \frac{1}{N} \sum_{k=1}^{K} W_k S_k^2 = \frac{1}{25} (0,1470)^2 - \frac{1}{275} (0,0218) = (0,0280)^2$$

oversigt

	estimator	varians
$\overline{y}_{si}$	$\frac{1}{n}\sum_{i=1}^{n} y_i$	$\frac{(N-n)}{N}\frac{1}{n}S^2$
$\widehat{p}$		$\frac{(N-n)}{N-1}P(1-P) \ =V(\widehat{p})$
$\widehat{p}$		$\frac{(N-n)}{N}\frac{1}{n-1}\widehat{p}(1-\widehat{p}) = \widehat{V(\widehat{p})}$
$\overline{y}_{strat}$	$\sum_{k=1}^{K} W_k\overline{y_{k.}}$	$\sum_{k=1}^{K} W_k^2\frac{N_k-n_k}{N_k}\frac{1}{n_k}S_k^2$
$\overline{y}_p$	$\sum_{k=1}^{K} W_k\overline{y_{k.}} = \frac{1}{n}\sum_{i=1}^{n} y_i$	$\frac{N-n}{N}\frac{1}{n}\sum_{k=1}^{K} W_k S_k^2$
$\overline{y}_{opt}$	$\sum_{k=1}^{K} W_k\overline{y_{k.}}$	$\frac{1}{n}(\sum_{k=1}^{K} W_k S_k)^2 - \frac{1}{N}\sum_{k=1}^{K} W_k S_k^2$

	allokering af n_k	bestemmelse af stikprøve-størrelse
$\overline{y}_{si}$		$\frac{S^2}{(\frac{L_0}{2*1{,}96})^2+\frac{1}{N}S^2}$
$\overline{y}_{strat}$		
$\overline{y}_p$	$n_k = nW_k$	$\frac{\sum_{k=1}^{K} W_k S_k^2}{(\frac{L_0}{2*1{,}96})^2+\frac{1}{N}\sum_{k=1}^{K} W_k S_k^2}$
$\overline{y}_{opt}$	$n_k = n\frac{W_k S_k}{\sum_{k=1}^{K} W_k S_k}$	$\frac{(\sum_{k=1}^{K} W_k S_k)^2}{(\frac{L_0}{2*1{,}96})^2+\frac{1}{N}\sum_{k=1}^{K} W_k S_k^2}$

konfidensintervaller for små stikprøver

En simpel tilfældig stikprøve på n=4.0000 blandt N=4.200.000
Der findes 2 (to) tilfælde af Ziska.
Udregn et 95% konfidensinterval for andelen af Zika tilfælde i Danmark
kommenter dette interval

Brug SAS programmet wright