Stikprøveteori, Brøk

tirsdag 2. uge, juli 2018

Stikprøveteori definitioner

Univers $Y_1, Y_2, Y_3, \ldots\ldots\ldots Y_N$,

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

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

stikprøve $y_1, y_2, y_3, \ldots\ldots y_n$,

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

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

tre dele: endelighed, stikprøve, varians i univers

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

$$\widehat{S^2} = s^2 = \frac{1}{n-1}\sum_{j=1}^{n} (y_i - \overline{y}_{si})^2$$

Brøk (ratio)

Univers $\begin{array}{cccccccc} Y_1 & Y_2 & Y_3 & . & . & . & . & Y_N \\ X_1 & X_2 & X_3 & . & . & . & . & X_N \end{array}$

$$\mathrm{R}=\frac{Y}{X} = \frac{\frac{1}{N}\sum\limits_{j=1}^{N} Y_i}{\frac{1}{N}\sum\limits_{j=1}^{N} X_i} = \frac{\overline{Y}}{\overline{X}}$$

stikprøve $y_1, y_2, y_3, \ldots\ldots y_n,$
stikprøve $x_1, x_2, x_3, \ldots\ldots x_n,$

$\widehat{R} = \frac{\overline{y}}{\overline{x}}$

$\mathrm{E}(\widehat{R}) \approx R$

$V(\widehat{R}) \approx \frac{1}{\overline{X}^2}\frac{N-n}{N}\frac{1}{n}T_z^2$

$T_z^2 = \frac{1}{N-1}\sum\limits_{j=1}^{n}(Y_j\text{-}RX_j)^2$

Brøk

$$\text{Univers} \qquad \begin{array}{cccccccc} Y_1 & Y_2 & Y_3 & . & . & . & . & Y_N \\ X_1 & X_2 & X_3 & . & . & . & . & X_N \\ Z_1 & Z_2 & Z_3 & . & . & . & . & Z_N \end{array}$$

$\mathrm{Z}_j = Y_j - RX_j$

$T_z^2 = \frac{1}{N-1}\sum_{j=1}^{N}(Y_j\text{-}RX_j)^2$

$\frac{1}{N-1}\sum_{j=1}^{n}(Y_j\text{-}RX_j)^2 = [S_y^2 + R^2 S_x^2 - 2R\rho_{xy}S_x S_y]$
estimeres ved
$\widehat{S}_y^2 + \widehat{R}^2\widehat{S}_x^2 - 2\widehat{R}\widehat{\rho}_{xy}\widehat{S}_x\widehat{S}_y$

$$
\text{Univers} \qquad \begin{array}{cccccccc} Y_1 & Y_2 & Y_3 & . & . & . & . & Y_N \\ X_1 & X_2 & X_3 & . & . & . & . & X_N \\ Z_1 & Z_2 & Z_3 & . & . & . & . & Z_N \end{array}
$$

$Z_j = Y_j - RX_j$ hvor R$=\frac{\overline{Y}}{\overline{X}}$

$$\overline{Z} = \frac{1}{N}\sum_{j=1}^{N} Z_j = \frac{1}{N}\sum_{j=1}^{N}(Y_j\text{-}RX_j) = \frac{1}{N}\sum_{j=1}^{N} Y_j - \frac{1}{N}\sum_{j=1}^{N} RX_j =$$

$$=\overline{Y} - R\overline{X} = \overline{Y} - \overline{\frac{\overline{Y}}{\overline{X}}X} = \overline{Y} - \overline{Y} = 0$$

$$var(Z) = \frac{1}{N-1}\sum_{j=1}^{N}(Z_j - \overline{Z})^2 = \frac{1}{N-1}\sum_{j=1}^{N}(Z_j)^2$$

$$T_z^2 = \frac{1}{N-1}\sum_{j=1}^{N}(Y_j\text{-}RX_j)^2$$

Nu haves et univers $Z_{1,}Z_{2,}Z_{3,}$............Z_N
med gennemsnit 0 og varians T_z^2

Brøk

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

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

$$\frac{1}{N-1}\sum_{j=1}^{N}(Y_j\text{-}\overline{Y})^2 = S_y^2$$
$$\frac{1}{N-1}\sum_{j=1}^{N}(\overline{Y} - RX_j)^2 = \frac{1}{N-1}\sum_{j=1}^{N}[R(\overline{X} - X_j)]^2 =$$

$$\frac{1}{N-1}R^2\sum_{j=1}^{N}(-\overline{X} + X_j)^2 = R^2 S_x^2$$

Brøk

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

$$\frac{1}{N-1}\sum_{j=1}^{N} 2R(Y_j\text{-}\overline{Y})(\overline{X} - X_j) =$$
$$\frac{1}{N-1}2R(SAP_{xy}) =$$

$$2R\frac{SAP_{xy}}{\sqrt{SAK_x SAK_y}}\sqrt{\frac{SAK_x}{N-1}\frac{SAK_y}{N-1}} =$$

$2R\rho S_x S_y$
samlet fås:
$$var(Z) = T_z^2 = \frac{1}{N-1}\sum_{j=1}^{N}(Y_j\text{-}RX_j)^2 =$$
$$[S_y^2 + R^2 S_x^2 - 2R\rho_{xy} S_x S_y]$$

BRØK frækt bevis

Fra kap. II fås
$E(\overline{z}_{si}) = \overline{Z} = 0$ og $V(\overline{z}_{si}) = \frac{N-n}{Nn} T_z^2$

$$\widehat{R} - R = \frac{\frac{1}{n}\sum_{i=1}^{n} y_i}{\frac{1}{n}\sum_{i=1}^{n} x_i} - R = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_j - Rx_i)}{\overline{x}}$$

$$E(\widehat{R} - R) = E\left(\frac{\frac{1}{n}\sum_{i=1}^{n}(y_j - Rx_i)}{\overline{x}}\right) \approx \frac{E(\overline{z})}{\overline{X}} = 0$$

BRØK frækt bevis

$$\widehat{R} - R = \frac{\frac{1}{n}\sum_{i=1}^{n} y_i}{\frac{1}{n}\sum_{i=1}^{n} x_i} - R = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_j - Rx_i)}{\overline{x}}$$

$$V(\widehat{R}) = V(\widehat{R} - R) = V\left(\frac{\frac{1}{n}\sum_{i=1}^{n}(y_j - Rx_i)}{\overline{x}}\right) \approx \frac{V(\overline{z})}{\overline{X}^2} =$$

$$= \frac{1}{\overline{X}^2}\frac{N-n}{N}\frac{1}{n}T_z^2$$

taylor

$$f(x) = f(x_0) + f^{'}(x_0)(x - x_0) + \text{ rest}$$

.

$$E[f(x)] = E[f(x_0)] + f^{'}(x_0)E(x - x_0) + \text{ rest} \approx f(x_0)$$

.

$$V[f(x)] = V[f(x_0)] + [f^{'}(x_0)]^2 V(x - x_0) + \text{ rest} \approx [f^{'}(x_0)]^2 V(x)$$

taylor

$$f(x,y) = f(x_0,y_0) + f_x'(x_0,y_0)(x-x_0) + f_y'(x_0,y_0)(y-y_0) + \text{ rest}$$

.

$$E[f(x,y)] \approx E[f(x_0,y_0)] + f_x'(x_0,y_0)E(x-x_0) + f_y'(x_0,y_0)E(y-y_0)$$

.

$$= E[f(x_0,y_0)] = f(x_0,y_0)$$

.

$$V[f(x,y)] \approx V[f(x_0,y_0) + f_x'(x_0,y_0)(x-x_0) + f_y'(x_0,y_0)(y-y_0)] =$$

.

$$V[f_x'(x_0,y_0)(x-x_0)] + V[f_y'(x_0,y_0)(y-y_0)] +$$

.

$$2cov[f_x'(x_0,y_0)(x-x_0), f_y'(x_0,y_0)(y-y_0)]$$

$f(x,y) = \frac{y}{x}$
udvikles i (x_0, y_0) =$(\overline{X}, \overline{Y})$ og $f(\overline{X}, \overline{Y}) = \frac{\overline{Y}}{\overline{X}} = R$

.

$f_x^{'} = (-x^{-2})y$
$f_y^{'} = x^{-1}$
$f(x,y) = \frac{y_0}{x_0} + (-x_0^{-2})y_0 * (x - x_0) + x_0^{-1} * (y - y_0) =$
$f(x,y) = R + (-Rx_0^{-1}) * (x - x_0) + x_0^{-1} * (y - y_0) =$

$$f(x,y) = R + (-Rx_0^{-1}) * (x - x_0) + x_0^{-1} * (y - y_0) =$$

.

$$E[f(x,y)] = E(R) + (-Rx_0^{-1}) * E((x - x_0)) + x_0^{-1} * E((y - y_0)) =$$

.

$$= \; R \;\; + (-Rx_0^{-1}) * 0 + x_0^{-1} * 0 = R$$

$$f(x,y) = R + (-Rx_0^{-1}) * (x - x_0) + x_0^{-1} * (y - y_0)$$

.

$$V[f(x,y)] =$$

.

$$V(R) + (-Rx_0^{-1})^2 * V((x - x_0)) + (x_0^{-1})^2 * V((y - y_0)) +$$

.

$$2(-Rx_0^{-1}) * (x_0^{-1}) * Cov(x,y) =$$

.

$$0 + \frac{R^2}{x_0^2} * V(x) + \frac{1}{x_0^2} * V(y) - 2R(\frac{1}{x_0^2}) * Cov(x,y) =$$

.

$$\frac{1}{x_0^2} * [R^2 V(x) + V(y) - 2R * Cov(x,y)] =$$
$$\frac{1}{x_0^2} * [R^2 V(x) + V(y) - 2R * \rho * S_y * S_x] =$$

$$S_y = \sqrt{V(y)} \quad S_x = \sqrt{V(x)} \quad \rho = \frac{Cov(x,y)}{S_y * S_x}$$
$$x_0 = \overline{X} \quad y_0 = \overline{Y} \quad R = \frac{\overline{Y}}{\overline{X}}$$

$$V(\widehat{R}) \approx (\frac{1}{\overline{X}})^2 * [S_y^2 + R^2 S_x^2 - 2R * \rho * S_y * S_x]$$

Usikkerhed i Lovmodel

Eksempel IV.2.2 side 204
N = 2,3 mio. husstande i Danmark

$\mathrm{X}_j = \#$ personer i husstand nr. j

$$\mathrm{X} = \sum_{j=1}^{N} X_j = 5{,}1 \text{ mio. personer. Det samlede antal personer i Danmark}$$

$\mathrm{Y}_j = \mathrm{X}_j$ hvis husstanden er i Roskilde, ellers er $\mathrm{Y}_j = 0$

$$\mathrm{Y} = \sum_{j=1}^{N} Y_j = \text{det samlede antal personer i Roskilde}$$

$\mathrm{R} = \frac{Y}{X} = \frac{antal\_personer\_i\_Roskilde}{antal\_personer\_i\_Danmark} = 0{,}041950516$
(opslag i Statistisk Årbog, nu bruges statistikbanken)

Stikprøve på n= 76.784 (husstande)

$x_1, x_2, x_3, ...x_n$

.

$$\sum_{i=1}^{76.784} X_i = 171.893 \text{ er det samlede antal personer i de udtrukne husstande}$$

Model 1 (forkert model)
Antag nu at de 171.893 personer er udtrukket simpelt tilfældigt blandt de 5,1 mio. personer i Danmark.
Vi ønsker at estimere andelen af personer der bor i Roskilde.

I stikprøven på de 171.893 personer bor 7.406 personer i Roskilde

$\widehat{R} = \frac{7406}{171896} = 0{,}04308494$

$V(\widehat{R}) = \frac{N-n}{Nn}\frac{N}{N-1}R(1-R)$

$\widehat{V(\widehat{R})} = \frac{N-n}{Nn}\frac{n}{n-1}\widehat{R}(1-\widehat{R}) = \frac{N-n}{N}\frac{1}{n-1}\widehat{R}(1-\widehat{R}) =$

.

husk at udvælgelsen er $1/30$

.

brug $\quad \widehat{R} = \frac{7406}{171896} = 0,04308494$

.

.

$(1-\frac{n}{N})\frac{1}{n-1}\widehat{R}(1-\widehat{R}) = (1-\frac{1}{30})\frac{1}{171893-1}\widehat{R}(1-\widehat{R}) = (0,0004815)^2$

$\mathrm{H}_0 :\ R = 0,041950516 \qquad \mathrm{H}_A : R \neq 0,041950516$

U$= \frac{\widehat{R}-R}{\sqrt{var(\widehat{R})}} = \frac{0,04308494-0,041950516}{0,0004815} = 2,36$ som er udover 1,96 grænsen

model 2 (Brøk estimation)
Stikprøve på n= 76.784 (husstande)
$y_1, y_2, y_3, ...y_n$ (husk at mange af y'erne bliver nul)

$x_1, x_2, x_3, ...x_n$

$$\widehat{R} = \frac{\frac{1}{n}\sum_{i=1}^{n} y_i}{\frac{1}{n}\sum_{i=1}^{n} x_i} = \frac{\overline{y}}{\overline{x}} = \frac{7406}{171896} = 0,04308494$$

vi har at

$$V(\widehat{R}) \approx \frac{1}{\overline{X}^2}\frac{N-n}{N}\frac{1}{n}T_z^2 \text{ som skal estimeres}$$

vi bruger $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{171.893}{76.784} = 2,2387$

$\frac{N-n}{N} = (1-f) = (1-\frac{1}{30})$

$$\widehat{T_z^2} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i\text{-}\widehat{R}x_i)^2 = 0,2905$$

$$V(\widehat{R}) \approx \frac{1}{(2{,}2387)^2}(1-\frac{1}{30})\frac{1}{76.784} * 0,2905 = (0,00085431)^2$$

$H_0: \ R = 0,041950516$ $\qquad$ $H_A: R \neq 0,041950516$

U$= \frac{\widehat{R}-R}{\sqrt{var(\widehat{R})}} = \frac{0{,}04308494-0{,}041950516}{0{,}00085431} = 1,32$ som er under 1,96 grænsen.

ratio estimation

R$=\frac{Y}{X}$ Y=RX

$\widehat{Y} = \widehat{R}X$

$\widehat{Y} = \frac{\overline{y}}{\overline{x}}X \qquad \frac{1}{N}\widehat{Y} = \frac{\overline{y}}{\overline{x}}\frac{1}{N}X$

$\overline{y}_R = \frac{\overline{y}}{\overline{x}}\overline{X} = \overline{y}_{si}\frac{\overline{X}}{\overline{x}} = \overline{y}_{si}\frac{gennemsnit\_i\_univers}{gns\_i\_stikprøve}$

V$(\overline{y}_R) = V(\frac{\overline{y}}{\overline{x}}\overline{X}) = V(\widehat{R})(\overline{X})^2 \approx \frac{N-n}{N}\frac{1}{n}T_z^2$

Regression overspringes

$$\overline{y}_{reg} = \overline{y}_{si} + \beta(\overline{X} - \overline{x})$$

vælg β som hældningskoefficient i universet
dvs $\beta = \frac{\rho_{xy} S_y}{S_x}$

$$\mathrm{V}(\overline{y}_{reg}) = \frac{N-n}{N}\frac{1}{n} S_y^2 (1 - \rho_{xy}^2) = V(\overline{y}_{si})(1 - \rho_{xy}^2)$$

så altid $V(\overline{y}_{reg}) <= V(\overline{y}_R)$

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

når ρ_{xy}^2 stor så $V(\widehat{R}) <= V(\overline{y}_{si})$

eks. IV.3.2 præmieindtægter s. 220

Univers Y_1,Y_2,Y_3, ...Y_{94},indtægt i 1990
$X_{1,}$ $X_{2,}$ $X_{3,}$.......... $X_{94,}$ indtægt i 1988

Univers oplysninger

	1988	1990
gns	$\overline{X} = 58.437$	$\overline{Y} = 67.368$
S	$S_X = 226.535$	$S_Y = 252.730$
R	$R = \frac{67368}{58437} = 1,1528$	
korr.	$\rho_{XY} = 0,99556$	
hældning	$\beta = \frac{0,99556 * 252730}{226535} = 1,1107$	

eks. IV.3.2 præmieindtægter s. 220

Univers $Y_1, Y_2, Y_3, ...Y_{94}$, indtægt i 1990
$X_{1,}$ $X_{2,}$ $X_{3,}.........$ $X_{94,}$ indtægt i 1988

Stikprøve n=5 oplysninger

	1988	1990
Asperup	85	78
Hagl	708	810
Lunde	976	1171
Rødvig	1142	1227
Ulfborg	138	125
gns	$\overline{x} = 610$	$\overline{y} = 682$
S	$s_X = 481$	$s_Y = 554$
R	$\widehat{R}=\frac{682}{610} = 1,1180$	
korr.	$\widehat{\rho}_{xy} = 0,99577$	
hældning	$\widehat{\beta} = \frac{0,99577*554}{481} = 1,1469$	

eks. IV.3.2 præmieindtægter s. 220

Univers $Y_1, Y_2, Y_3, \ldots\ldots\ldots Y_{94}$,indtægt i 1990
$X_{1,}\ X_{2,}\ X_{3,} \ldots\ldots\ldots\ X_{94,}$ indtægt i 1988

	estimator	spredning
$\overline{y}_{si}$	682	75.688
$\overline{y}_R$	$=682*\frac{58437}{610}=65334$	7.676
$\overline{y}_{reg}$	$=682+1{,}1469*(58437-610)=67004$	7.124

Proportional og optimal

$n_k = nW_k$

$\overline{y}_p = \sum_{k=1}^{K} W_k \overline{y_{k.}} = \frac{1}{n} \sum_{i=1}^{n} y_i$ 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

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