Surveymetodik
Föreläsning 8

Måns Magnusson

Avd. Statistik, LiU

Översikt

Section 1

Kvotestimation

Introduktion till kvotestimation

- Vi ksa nu fokusera på estimation - d.v.s. efter att undersökningen är gjord
- Tidigare estimerat populationsparametrar $p_{\mathcal{U}}$, $\bar{y}_{\mathcal{U}}$ och $t_{\mathcal{U}}$ med $p_{\mathcal{S}}, \bar{y}_{\mathcal{S}}$ och $N\bar{y}_{\mathcal{S}}$

- Ett alternativ är **kvotestimation** då vi
  - observerat y
  - observerat **hjälpvariabeln** (auxiliary variable) x

- **Två situationer**:
  - Vi känner till populationstotalen för x (t_x)
  - Vi saknar kunskap om populationstotalen för x (via ex. register)

Introduktion till kvotestimation II

- Exempel på användning om vi **inte känner** till t_x
  - Vi kan vara intresserad av **populationskvoten**

$$\frac{\bar{y}_{\mathcal{U}}}{\bar{x}_{\mathcal{U}}} = \frac{t_y}{t_x} = B$$

  - Vi kan vara intresserade av att skatta i **redovisningsgrupper** (domänestimation). $(*)$
- Exempel på användning om vi **känner** till t_x
  - Vi kan använda t_x för att **förbättra precisionen** i $\hat{\bar{y}}_{\mathcal{U}}$ eller $\hat{t}_y$
  - Vi kan använda t_x för att skapa **totalskattningen** t_y **när** N **är okänd**.
  - Vi kan använda t_x för att **kalibrera** $\hat{\bar{y}}_{\mathcal{U}}$ eller $\hat{t}_y$.
    Används för att hantera bortfallsfel och ramfel.
- Kvotestimator ska användas om **nämnaren** ändras vid ett annat urval.

Populationsparametrar av intresse

$$t_y = t_{\mathcal{U},y} = \sum_{i \in \mathcal{U}} y_i = \sum_{i=1}^{N} y_i = \text{Populationstotalen för variabel } y$$

$$\bar{y}_{\mathcal{U}} = \frac{1}{N} \sum_{i \in \mathcal{U}} y_i = \text{Populationsmedelvärde för variabel } y$$

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

$$B = \frac{\bar{y}_{\mathcal{U}}}{\bar{x}_{\mathcal{U}}} = \frac{t_y}{t_x} = \text{Populationskvoten}$$

$$R = \frac{\sum_{i \in \mathcal{U}} (x_i - \bar{x}_{\mathcal{U}})(y_i - \bar{y}_{\mathcal{U}})}{(N-1) S_y S_x} = \text{Populationskorrelationskoefficienten}$$

- Skattningen av kvoten görs på följande sätt:

$$\hat{B} = \frac{\bar{y}_S}{\bar{x}_S} = \frac{\sum_{i \in S} y_i}{\sum_{i \in S} x_i}$$

- Om vi känner till t_x kan vi använda x som **hjälpvariabel** för att skatta t_y med **bättre precision** på följande sätt:

$$\hat{\bar{y}}_r = \hat{B}\bar{x}_U = \text{kvotskattning av } \bar{y}_U$$

$$\hat{t}_{r,y} = \hat{B}t_x = \text{kvotskattning av } t_y$$

Estimatorer II

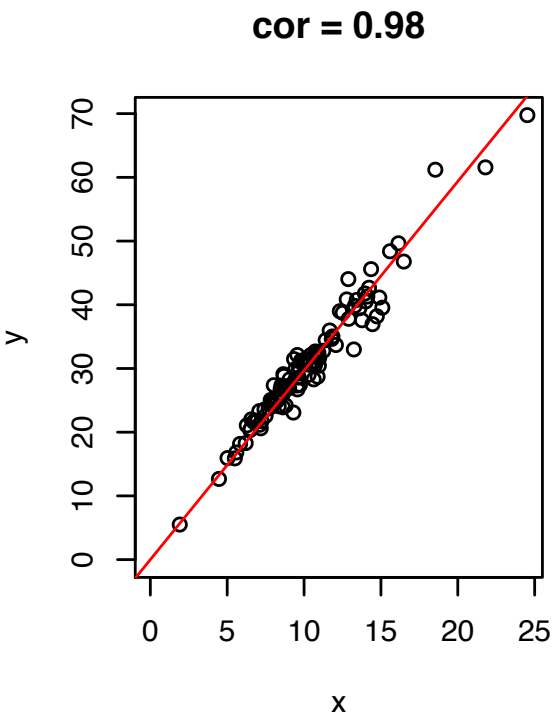
- En kvotskattning bygger på modellen

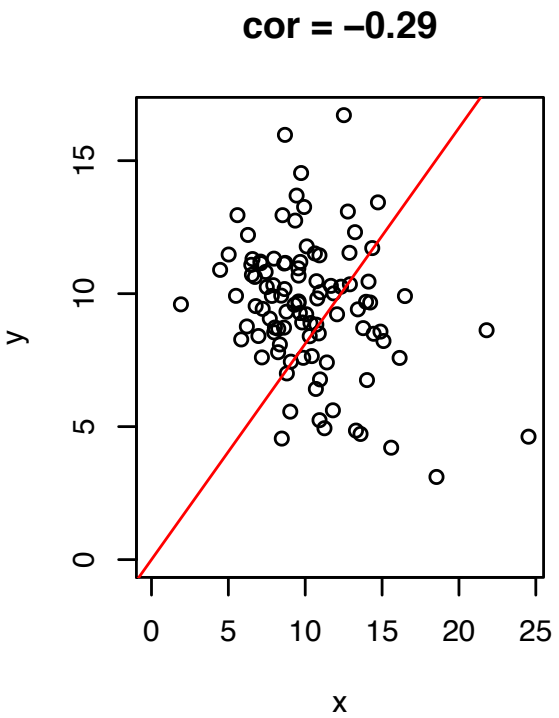
$$y_i = Bx_i$$

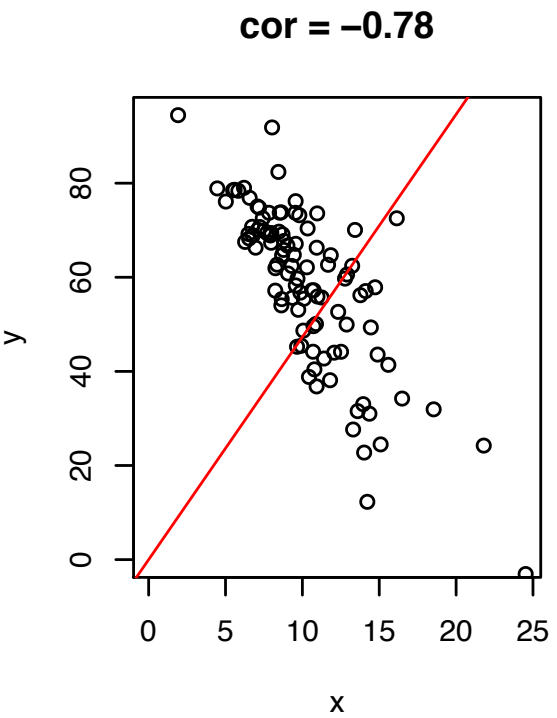
- Detta innebär att kvotestimation ökar precisionen när:
  - (a) Korrelationen i populationen är tillräckligt stor (**och** positiv)
  - (b) När den sanna "regressionslinjen" i populationen, B, går genom 0 (annars använder vi regressionsestimaton)

- Kallas ofta för **modell-assisterad** estimation.

Exempel: Brott i bostadsområden







Bias

- $\hat{\bar{y}}_r$ är dock inte en (design) väntevärdesriktig skattning av $\bar{y}_{\mathcal{U}}$ utan det finns en **bias**
  (se Lohr (2009, s. 125)för detaljer)

$$Bias(\hat{\bar{y}}_r) \approx \left(1 - \frac{n}{N}\right) \frac{\left(BS_x^2 - RS_xS_y\right)}{n\bar{x}_{\mathcal{U}}}$$

- Bias minskas således av
  - n är stort
  - urvalsfraktionen $\frac{n}{N}$ är stor
  - $\bar{x}_{\mathcal{U}}$ är stor och S_x är litet
  - Populationskorrelationen R är hög

Varians

- Variansen för $\hat{B}$ skattas (se Lohr (2009, s. 125 f.) för detaljer)

$$\hat{Var}(\hat{B}) = \left(1 - \frac{n}{N}\right) \frac{s_e^2}{n\bar{x}_S^2}$$

där

$$s_e^2 = \frac{1}{n-1} \sum_{i \in S} e_i^2 \text{ och } e_i = y_i - \hat{B}x_i$$

där e_i är residualerna kring linjen.

- Variansen minskas av
  - Små e_i (bra modell)
  - Liten urvalsfraktion $\frac{n}{N}$
  - Stort n
  - Stort $\bar{x}_S^2$

Varians II

- Om vi **känner till** $\bar{x}_{\mathcal{U}}$ och t_x så kan vi använda detta för att få mindre varians när vi skattar $\hat{\bar{y}}_r$ och $\hat{t}_{r,y}$:

$$\hat{Var}(\hat{\bar{y}}_r) = \hat{Var}(\hat{B}\bar{x}_{\mathcal{U}}) = \bar{x}_{\mathcal{U}}^2 \hat{Var}(\hat{B}) = \left(1 - \frac{n}{N}\right)\left(\frac{\bar{x}_{\mathcal{U}}}{\bar{x}_{\mathcal{S}}}\right)^2 \frac{s_e^2}{n}$$

$$Var(\hat{t}_{r,y}) = Var(\hat{B}t_x) = t_x^2 Var(\hat{B}) = \left(1 - \frac{n}{N}\right)\left(\frac{t_x}{\bar{x}_{\mathcal{S}}}\right)^2 \frac{s_e^2}{n}$$

- Om n är stort blir $\left(\frac{\bar{x}_{\mathcal{U}}}{\bar{x}_{\mathcal{S}}}\right)^2 \approx 1$ och $\left(\frac{t_x}{\bar{x}_{\mathcal{S}}}\right)^2 \approx N^2$
- Vinsten uppstår om modelen är bra $y_i - \hat{B}x_i < y_i - \bar{y}$
  (Återkommer vid regressionsestimation)
- Om urvalen är tillräckligt stora kan vi använda **centrala gränsvärdessatsen** och beräknar konfidensintervall på följande sätt:

$$\hat{B} \pm z_{\alpha/2}\sqrt{\hat{Var}(\hat{B})}, \hat{\bar{y}}_r \pm z_{\alpha/2}\sqrt{\hat{Var}(\hat{\bar{y}}_r)} \text{ och } \hat{t}_{r,y} \pm z_{\alpha/2}\sqrt{\hat{Var}(\hat{t}_{r,y})}$$

Mean squared error (MSE)

- $Bias(\hat{\bar{y}}_r)^2$ minskar snabbare med urvalsstorleken (n) än $Var(\hat{\bar{y}}_r)$ så (lite slarvigt):

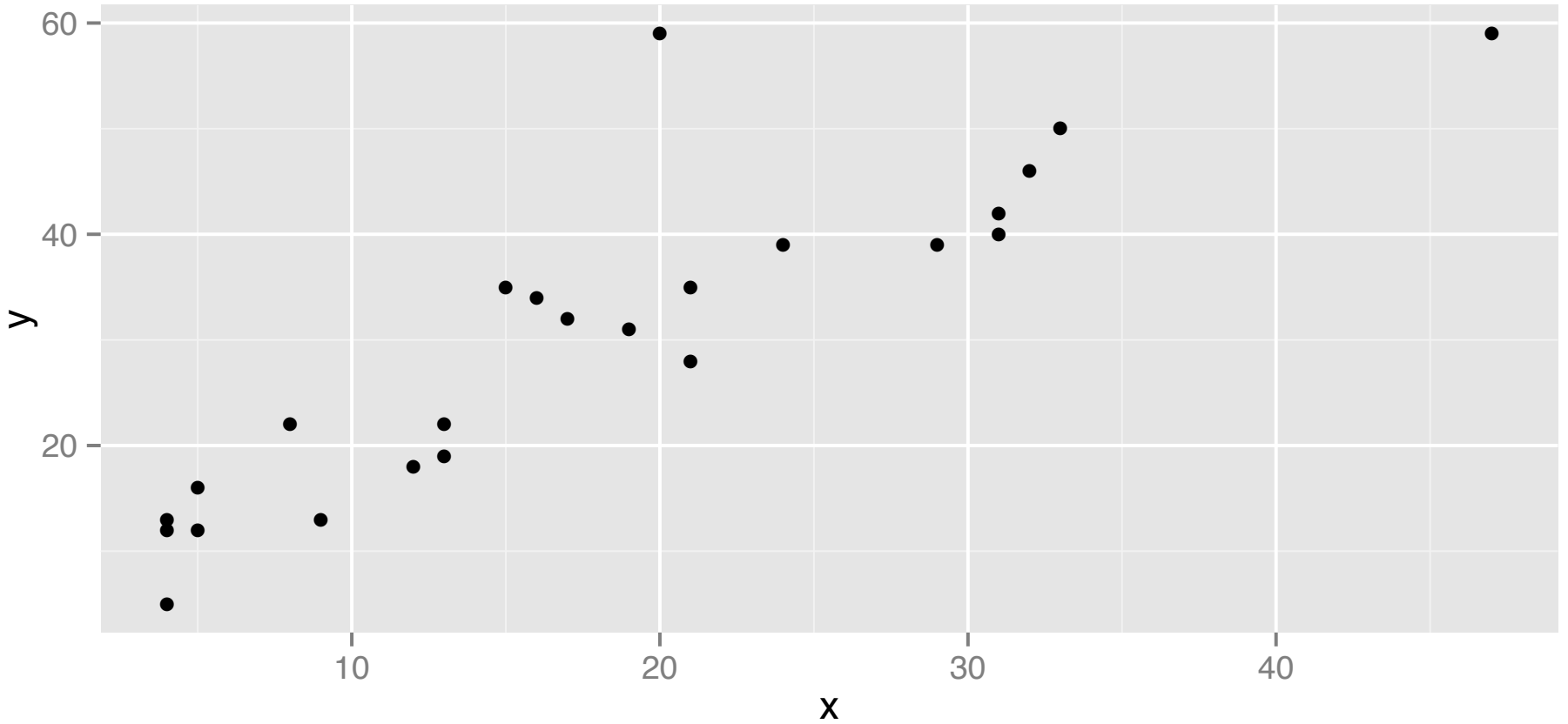
$$n \to \infty \text{ så } MSE(\hat{\bar{y}}_r) \to Var(\hat{\bar{y}}_r)$$

- Om korrelationen mellan x och y är tillräckligt stor($R > \frac{1}{2}$) blir $MSE(\hat{\bar{y}}_r) < MSE(\hat{\bar{y}}_{\mathcal{U}})$ och

- För större urval så är $Var(\hat{\bar{y}}_r) < Var(\hat{\bar{y}}_{\mathcal{U}})$ (se Lohr (2009, s. 133))

Exempel: Brott i bostadsområden

Vi vill uppskatta det totala antalet utsatta för inbrott i 24 bostadsområden (y). Vi vet antalet polisanmälningar (x) för alla områden och drar ett urval på $n = 4$.



Korrelationen i populationen är 0.884.

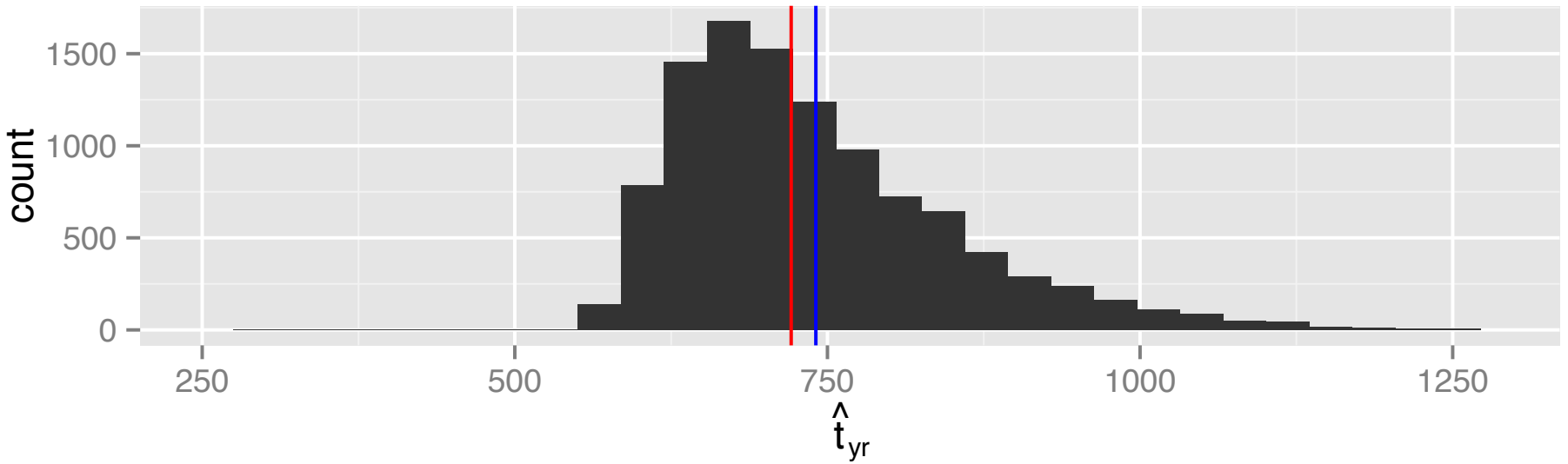
Exempel: Brott i bostadsområden II

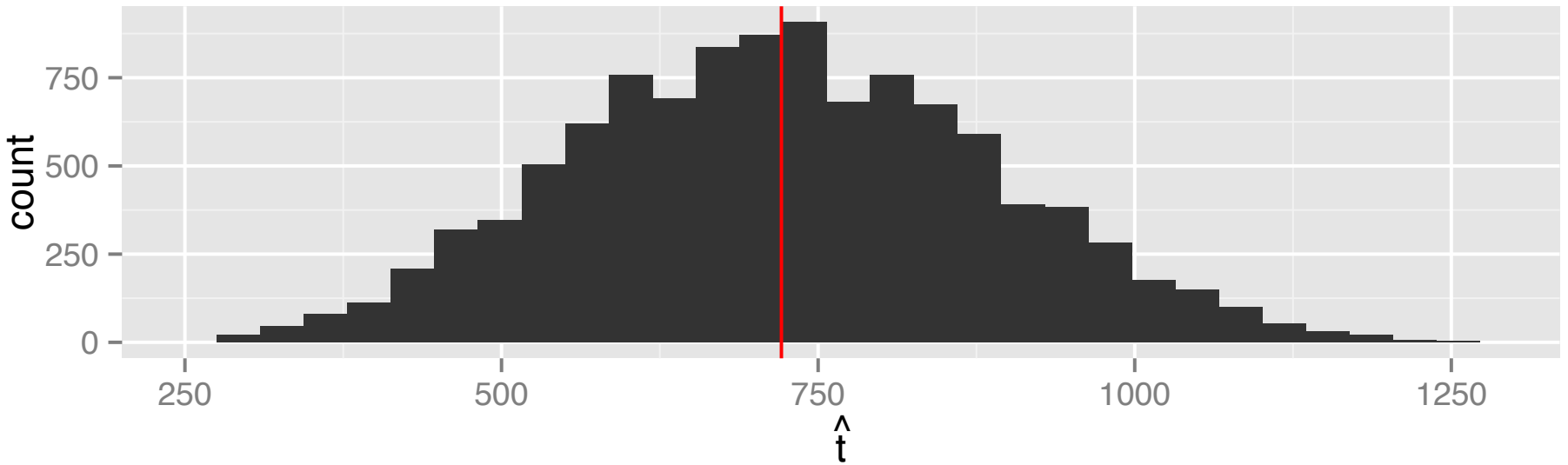
Teoretiska fördelningen med K =10626 stycken teoretiska urval.

	obs.1	obs.2	obs.3	obs.4	t_x	t_hat_x	t_hat_y	B_hat	t_hat_yr
3725	31	13	59	35	433	414	828	2.00	866
590	32	16	35	22	433	306	630	2.06	891
3371	31	16	5	18	433	240	420	1.75	758
5356	16	59	34	39	433	390	888	2.28	986
9469	59	18	35	59	433	600	1026	1.71	740
923	32	13	42	40	433	528	762	1.44	625
4436	31	34	35	40	433	522	840	1.61	697
4078	31	5	12	59	433	450	642	1.43	618
9511	59	35	42	46	433	624	1092	1.75	758
2383	28	13	13	19	433	282	438	1.55	673
4025	31	5	34	12	433	258	492	1.91	826
2585	28	35	18	12	433	312	558	1.79	774
2141	28	16	42	46	433	534	792	1.48	642
7950	35	22	22	59	433	498	828	1.66	720
6590	50	18	35	40	433	582	858	1.47	638
6929	13	5	39	34	433	348	546	1.57	679
7311	13	39	42	19	433	492	678	1.38	597
2977	28	22	18	12	433	306	480	1.57	679
1062	32	35	42	12	433	402	726	1.81	782
9798	22	35	42	12	433	414	666	1.61	697

Exempel: Brott i bostadsområden III

Samlingfördelningen för $\hat{t}_{y,r}$ och $\hat{t}_y$ då t_y =721, n =4 och N =24.





Effekten av kvotestimation

- Skillnaden mellan kvotestimatorn och den "vanliga" estimatorn

$$
\begin{array}{rclrcl}
E(\hat{t}_{yr}) & = & 740.744 & E(\hat{t}_{y}) & = & 721 \\
Var(\hat{t}_{yr}) & = & 12329.028 & Var(\hat{t}_{y}) & = & 27438.043 \\
Bias(\hat{t}_{yr}) & = & 19.744 & Bias(\hat{t}_{y}) & = & 0 \\
MSE(\hat{t}_{yr}) & = & 12718.851 & MSE(\hat{t}_{y}) & = & 27438.043
\end{array}
$$

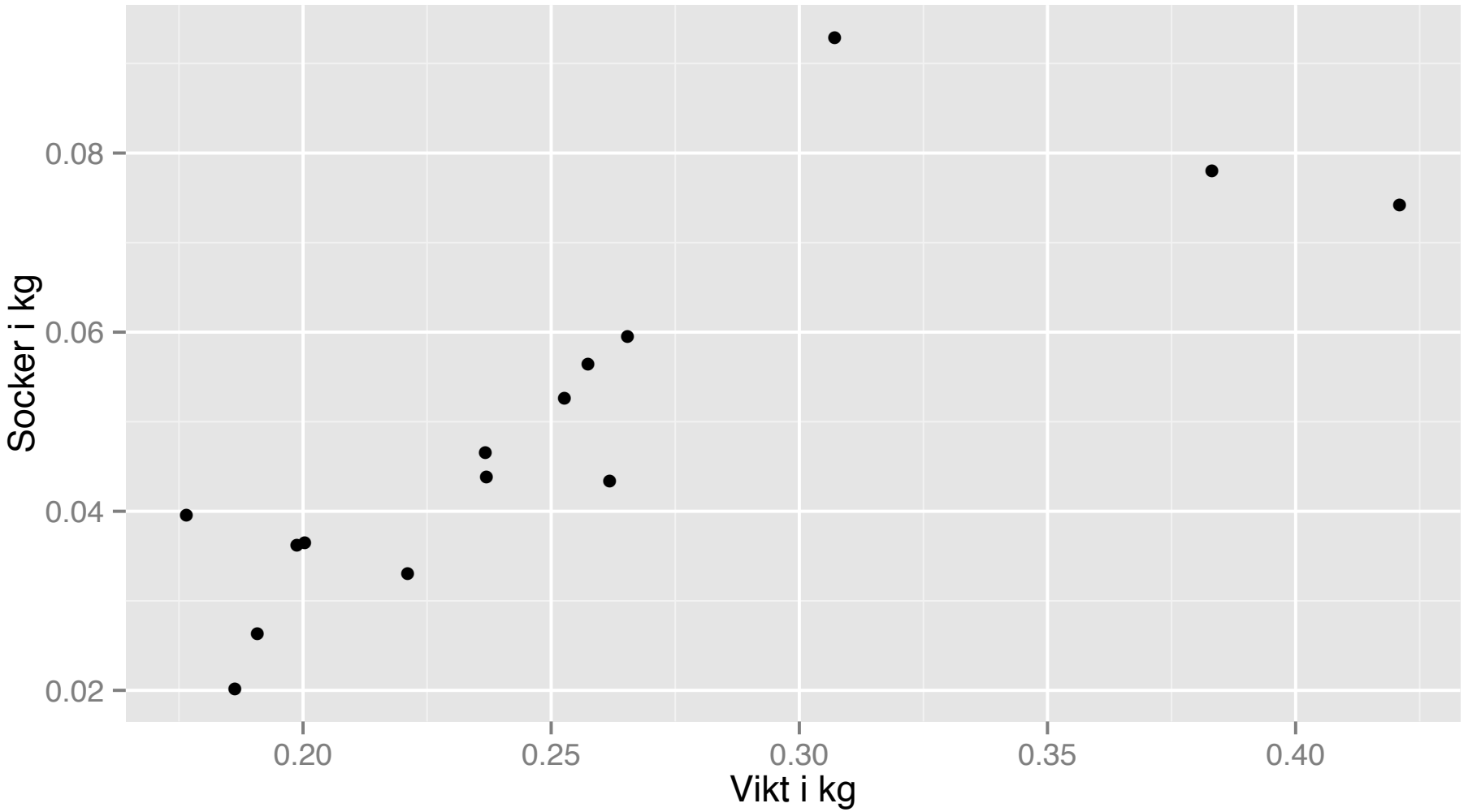
Exempel: Sockermängd i apelsiner

- Intresserade av att uppskatta den **totala** sockermängden i en lastbil apelsiner.
- Vi undersöker ett OSU av 15 apelsiner och för respektive apelsin mäts vikt och sockermängd.
- Den totala vikten för hela lasten är 857 kg.
- Skatta den totala mängden socker i hela lasten med tillhörande konfidensintervall.

Exempel: Sockermängd, data

	socker	vikt	e
1	0.0781	0.383	0.00342
2	0.0365	0.200	-0.00251
3	0.0438	0.237	-0.00234
4	0.0526	0.253	0.00337
5	0.0202	0.186	-0.01612
6	0.0264	0.191	-0.01079
7	0.0434	0.262	-0.00758
8	0.0465	0.237	0.00041
9	0.0564	0.257	0.00627
10	0.0929	0.307	0.03307
11	0.0742	0.421	-0.00776
12	0.0396	0.177	0.00520
13	0.0331	0.221	-0.01000
14	0.0595	0.265	0.00782
15	0.0363	0.199	-0.00247

Exempel: Sockermängd, data



Exempel: Sockermängd i apelsiner V

- Det hade i detta fall inte varit möjligt att räkna ut totala sockermängden utan hjälpinformationen vikt.
- Det är tydligt att i detta fall fungerar kvotestimatorn då y måste vara 0 då $x = 0$.

Subsection 1

Kvotestimatorn som kalibrering

Kvotestimation som kalibrering

- En kvotestimator kan även uttryckas i form av vikter

$$\hat{t}_y = \sum_{i \in \mathcal{S}} w_i y_i \text{ och } \hat{t}_{yr} = \frac{t_x}{\hat{t}_x} \sum_{i \in \mathcal{S}} w_i y_i = \sum_{i \in \mathcal{S}} w_i g_i y_i$$

där

$$g_i = \frac{t_x}{\hat{t}_x}$$

- På detta sätt **kalibreras** vikterna w_i till de **kända populationstotalerna**

$$\sum_{i \in \mathcal{S}} w_i g_i x_i = \frac{t_x}{\hat{t}_x} \sum_{i \in \mathcal{S}} w_i x_i = \frac{t_x}{\hat{t}_x} \hat{t}_x = t_x$$

- Detta är principen när vi bortfallskalibrerar (fast då använder vi ofta regressionsestimatorn istället).

Exempel: Kalibrering

- Vi är intresserade av **antalet** personer som känner sig otrygga under helgkvällar år 2013 och i Norrköpings kommun (N=130 623) varav 64 851 män och 65 772 kvinnor.
- Vi undersöker 1140 personer och av dessa svarar 789 personer.
- Av de svarande är 424 kvinnor och 315 män och av männen svarar 45 att de har känt sig otrygga och bland kvinnorna 192.
- Vid en bortfallsanalys visar det sig att bortfallet är större bland män än bland kvinnor.
- Beräkna g-vikten för män och kvinnor och gör en totalskattning av antalet otrygga.

Subsection 2

Redovisningsgrupper

Redovisningsgrupper

- Ofta finns ett intresse att producera skattningar i mindre delar av populationen, dessa kallas **domäner** eller **redovisningsgrupper**
- Vanliga redovisningsgrupper/domäner är kön, ålder och geografi
- Ibland **väldigt många** domäner

- n i varje domän blir mycket litet (ibland $n = 0$), kallas **Small area estimation**

Redovisningsgrupper II

- Det finns två typer av domänestimation:
  - När urvalet i domänen är **fast** (ex. vid stratifiering)
    Precis som vanligt (fast med mindre urval)
  - När urvalet i domänen är **slumpmässigt**
    Specialfall av kvotskattning där n_d behöver skattas

Redovisningsgrupper - Notation

$$N_d = \text{Antal observationer i domän } d$$

$$\mathcal{U}_d = \{1, 2, 3, ..., N_d\} = \text{Populationsmängden i domän } d$$

$$n_d = \text{Antal observationer i urvalet i domän } d$$

$$\mathcal{S}_d = \{1, 2, ..., n_d\} = \text{Urvalsmängden i domän } d \ (\mathcal{S}_d \subseteq \mathcal{U}_d)$$

$$\bar{y}_{\mathcal{U}_d} = \frac{1}{N_d} \sum_{i \in \mathcal{U}_d} y_i$$

$$\bar{y}_{\mathcal{S}_d} = \frac{1}{n_d} \sum_{i \in \mathcal{S}_d} y_i$$

- Först skapar vi nya variabler för vår domän

$$x_i = \begin{cases} 1 & \text{om } i \in \mathcal{U}_d \\ 0 & \text{om } i \notin \mathcal{U}_d \end{cases}$$

$$u_i = y_i x_i = \begin{cases} y_i & \text{om } i \in \mathcal{U}_d \\ 0 & \text{om } i \notin \mathcal{U}_d \end{cases}$$

- Med hjälp av dessa variabler kan vi skatta $\bar{y}_{\mathcal{U}_d}$ med en kvotestimator

$$t_x = \sum_{i \in \mathcal{U}} x_i = N_d \text{ och } \bar{x}_{\mathcal{U}} = \frac{N_d}{N}$$

$$t_u = \sum_{i \in \mathcal{U}} u_i = \sum_{i \in \mathcal{U}_d} y_i$$

$$\bar{y}_{\mathcal{U}_d} = \frac{t_u}{t_x} = B$$

$$\bar{y}_{\mathcal{S}_d} = \bar{y}_d = \hat{B} = \frac{\bar{u}}{\bar{x}} = \frac{\hat{t}_u}{\hat{t}_x}$$

- För att beräkna medelfelet använder vi medelfelet från kvotestimatorn vilket ger $(*)$
  (se Lohr (2009, s. 135))

$$\hat{Var}(\bar{y}_d) = \left(1 - \frac{n}{N}\right) \frac{n}{n_d^2} \frac{(n_d - 1)s_{yd}^2}{n - 1}$$

där

$$s_{yd}^2 = \frac{\sum_{i \in \mathcal{S}_d} (y_i - \bar{y}_d)^2}{n_d - 1}$$

- Om $E(n_d)$ är stort så är $(n_d - 1)/n_d \approx 1$ och $n/(n-1) \approx 1$ vilket gör att

$$SE(\bar{y}_d) \approx \sqrt{\left(1 - \frac{n}{N}\right) \frac{s_{yd}^2}{n_d}}$$

Exempel: Bygga ett resecentrum

- Opinionsundersökning gällande inställning till att bygga ett resecentrum i Vaxholm
- Vi drar ett OSU och frågar $n = 500$ personer av $N = 11\ 141$ (och har inget bortfall)
- Av 500 svarande anger 277 att de är positiva.
- Vi har även data redovisat i olika åldersgrupper.

Åldersgrupp	Antal i urvalet (n)	Antal positiva
18-30 år	139	103
31-65 år	281	142
66-80 år	80	32

- Vi är intresserade av **andelen** och **totalen** som är positiva i åldersgruppen 18-30 år samt populationen som helhet.

Referenser

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