Surveymetodik
Föreläsning 2

Måns Magnusson

Avd. Statistik, LiU

Översikt

1 Introduktion till urval
- Notation
- Populationsparametrar
- Urvalsteori

2 Obundet slumpmässigt urval
- Inklutionssannolikhet och designvikt
- Estimation vid OSU
- Konfidensintervall vid OSU
- Urvalsdimensionering
- Totaler och proportioner vid OSU

Section 1

Introduktion till urval

Notation I

$$N = \text{Antal observationer i populationen}$$

$$\mathcal{U} = \{1, 2, 3, ..., N\} = \text{Populationsmängden}$$

$$n = \text{Antal observationer i urvalet}$$

$$\mathcal{S} = \{1, 2, ..., n\} = \text{Urvalsmängden}(\mathcal{S} \subseteq \mathcal{U})$$

$$i = \text{Observation } i$$

$$i \in \mathcal{S} = \text{Observation } i \text{ ingår i urvalsmängden } \mathcal{S}\text{, del av urvalet}$$

$$\mathcal{K} = \{\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_K\} \text{=Mängden av alla möjliga urval}$$

Notation II

$y =$ Den egenskap/variabel vi vill undersöka

$y_i =$ egenskapen hos individ i

$t =$ total

$p =$ proportion

$\hat{p} =$ skattning av p

- Populationstotal

$$t_{\mathcal{U}} = \sum_{i=1}^{N} y_i = \sum_{i \in \mathcal{U}} y_i$$

- Populationsmedel

$$\bar{y}_{\mathcal{U}} = \frac{1}{N} \sum_{i=1}^{N} y_i$$

- Populationsvarians

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

- Populationsandel

$$p_{\mathcal{U}} = \frac{\text{antal med egenskap av intresse}}{N}$$

- Urvalsteori är en **annan** teori för **inferens** än den "klassiska" statistiska teorin
  - y_i är ett **fixt** värde, inte utfall en variabel Y
  - **Skapar slumpmässigheten själva** genom att dra ett urval slumpmässigt
  - Syftar **bara** till inferens om ändliga populationer (populationsparametrar, inte μ eller σ i en normalfördelning)
  - Kallas ofta **designbaserad** inferens
- Det finns två sätt att dra (skapa) ett slumpmässigt urval
  - med återläggning
  - utan återläggning
  - vidare kommer endast "utan återläggning" behandlas (om inte annat nämns)

Section 2

Obundet slumpmässigt urval

- Obundet slumpmässigt urval - alla urval har samma sannolikhet
- Mest grundläggande urvalsmetod - grunden för andra mer komplicerade metoder

- Antalet tänkbara urval av storlek n från en population av storlek N (utan återläggning):

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

- **Urvalsfördelningen** för $\mathcal{S}_k$ vid OSU

$$P(\mathcal{S}_k) = \frac{1}{\binom{N}{n}} = \frac{n!(N-n)!}{N!}$$

- Sannolikheten att observation i inkluderas i urvalet: π_i
- **Inklutionssannolikheten** för observation i **vid OSU**

$$\pi_i = \frac{\#\text{ urval som innehåller obs. } i}{\text{Totalt antal urval}} = (*) = \frac{n}{N}$$

- Summan (över hela populationen) av inklutionssannolikheterna vid OSU är:

$$\sum_{i=1}^{N} \pi_i = N \cdot \frac{n}{N} = n$$

Designvikt

- **Urvals-/designvikten** (sampling/design weight) för observation i

$$w_i = \frac{1}{\pi_i} = \frac{N}{n}$$

- Summan av vikterna (över urvalet) vid OSU är

$$\sum_{i=1}^{n} w_i = n \cdot \frac{N}{n} = N$$

- **Tolkning:** Antal personer som representeras av varje respondent i studien

OSU - Exempel 1

- Vad är sannoliheten för att inkludera den vita bollen i vårt urval om $n = 3$ och $\mathcal{U} = \{\text{Grön}, \text{Blå}, \text{Röd}, \text{Svart}, \text{Vit}\}$ $(N = 5)$
- Vi drar ett OSU **utan** återläggning.

- Totalt antal urval vi kan dra är:

$$\binom{N}{n} = \binom{5}{3} = \frac{5!}{(5-3)!3!} = 10$$

- De olika urvalen vi kan dra blir då:

$$\mathcal{S}_1 = \{\mathrm{G}, \mathrm{B}, \mathrm{R}\}, \mathcal{S}_2 = \{\mathrm{G}, \mathrm{B}, \mathrm{S}\}, \mathcal{S}_3 = \{\mathrm{G}, \mathrm{B}, \mathrm{V}\}, \mathcal{S}_4 = \{\mathrm{G}, \mathrm{R}, \mathrm{S}\},$$
$$\mathcal{S}_5 = \{\mathrm{G}, \mathrm{R}, \mathrm{V}\}, \mathcal{S}_6 = \{\mathrm{G}, \mathrm{S}, \mathrm{V}\}, \mathcal{S}_7 = \{\mathrm{B}, \mathrm{R}, \mathrm{S}\}, \mathcal{S}_8 = \{\mathrm{B}, \mathrm{R}, \mathrm{V}\},$$
$$\mathcal{S}_9 = \{\mathrm{B}, \mathrm{S}, \mathrm{V}\}, \mathcal{S}_{10} = \{\mathrm{R}, \mathrm{S}, \mathrm{V}\}$$

- Sannolikheten för respektive urval är $P(\mathcal{S}_k) = \frac{1}{10}$, vilket ger att

$$\pi_{\mathrm{Vit}} = P(\mathcal{S}_3) + P(\mathcal{S}_5) + P(\mathcal{S}_6) + P(\mathcal{S}_8) + P(\mathcal{S}_9) + P(\mathcal{S}_{10}) = \frac{3}{5}$$

- Eller vi kan använda formeln på slide 10:

$$\pi_{\mathrm{Vit}} = \frac{n}{N} = \frac{3}{5}$$

Subsection 2

Estimation vid OSU

Estimation vid OSU

- Vi är intresserade av att skatta eller **estimera** populationsparametrarna $\bar{y}_{\mathcal{U}}$ och $S^2_{\mathcal{U}}$.
- För detta använder vi **estimatorer**.

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

$$\hat{S^2_{\mathcal{U}}} = S^2_{\mathcal{S}} = s^2 = \frac{1}{n-1}\sum_{i\in\mathcal{S}}(y_i - \bar{y})^2$$

- På grund av urvalfördelningen $P(\mathcal{S})$ får estimatorerna en sannolikhetfördelning.

OSU - Exempel 2

- Vi fortsätter på exemplet på slide 12 ovan där $N = 5$, $n = 3$ och $\mathcal{U} = \{\text{Grön}, \text{Blå}, \text{Röd}, \text{Svart}, \text{Vit}\}$ och $P(\mathcal{S}) = \frac{1}{10}$
- De olika bollarna väger olika mycket

$$y = \{5, 10, 8, 7, 12\}$$

OSU - Exempel 2 (forts.)

- Populationsparametrar (de sanna värdena) som vi är intresserade av är:
- Populationsmedelvärdet:

$$\bar{y}_{\mathcal{U}} = \frac{1}{N} \sum_{i \in \mathcal{U}} y_i = \frac{1}{5} \sum_{i \in \{a,b,c,d,e\}} y_i = 8.4$$

- Populationsvarians/standardavvikelse:

$$S^2_{\mathcal{U}} = \frac{1}{N-1} \sum_{i \in \mathcal{U}} (y_i - \bar{y}_{\mathcal{U}})^2 = \frac{1}{4} \sum_{i \in \{a,b,c,d,e\}} (y_i - 8.4)^2 = 7.3$$

$$S_{\mathcal{U}} = \sqrt{7.3} \approx 2.3$$

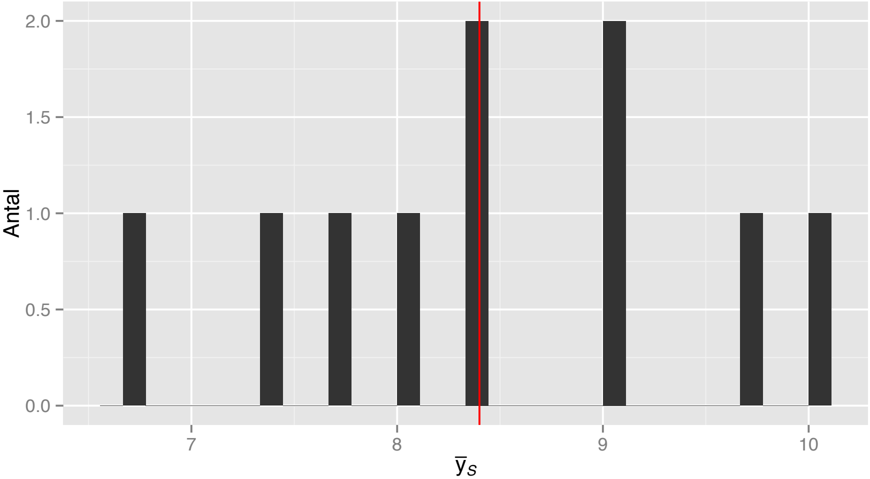
- Vad är samplingfördelningen för estimatorn $\hat{\bar{y}}_{\mathcal{U}}$?

OSU - Exempel 2 (forts. II) I

- Med R kan vi beräkna samplingfördelningen för $\bar{y}_S$ då $n = 3$ och $y =$(5, 10, 8, 7, 12)
- Här används funktionen `samplingDist()` som finns här.

```
   Obs.1 Obs.2 Obs.3 P_S y_hat s_hat
1      5    10     8 0.1  7.67  2.52
2      5    10     7 0.1  7.33  2.52
3      5    10    12 0.1  9.00  3.61
4      5     8     7 0.1  6.67  1.53
5      5     8    12 0.1  8.33  3.51
6      5     7    12 0.1  8.00  3.61
7     10     8     7 0.1  8.33  1.53
8     10     8    12 0.1 10.00  2.00
9     10     7    12 0.1  9.67  2.52
10     8     7    12 0.1  9.00  2.65
```

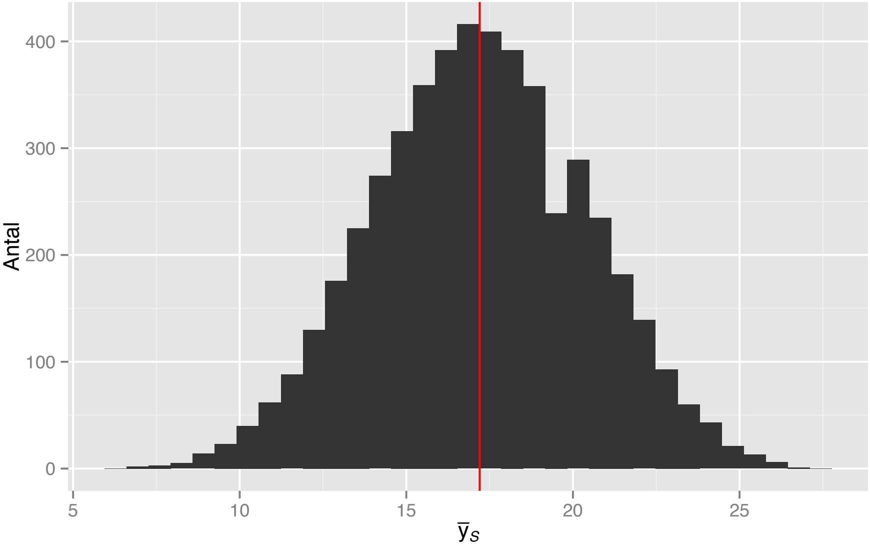
Samplingfördelningen för $\bar{y}_{\mathcal{S}}$



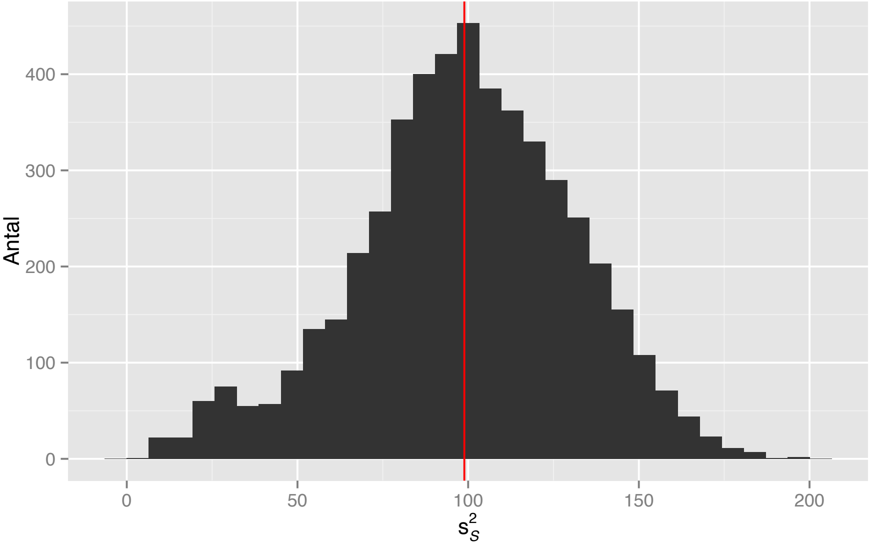
- Samplingfördelningen för $\bar{y}_{\mathcal{S}}$ då $n=6$, $N=15$ och $\binom{15}{6}=5005$.
- $y_{\mathcal{U}}=(29, 21, 23, 3, 22, 30, 24, 6, 15, 2, 4, 10, 16, 27, 26)$.

	Obs.1	Obs.2	Obs.3	Obs.4	Obs.5	Obs.6	P_S	y_hat	s_hat
1	29	21	23	3	22	30	0.0002	21.3	9.73
2	29	21	23	3	22	24	0.0002	20.3	8.94
3	29	21	23	3	22	6	0.0002	17.3	10.37
4	29	21	23	3	22	15	0.0002	18.8	8.95
5	29	21	23	3	22	2	0.0002	16.7	11.33
6	29	21	23	3	22	4	0.0002	17.0	10.83
7	29	21	23	3	22	10	0.0002	18.0	9.59
8	29	21	23	3	22	16	0.0002	19.0	8.88
9	29	21	23	3	22	27	0.0002	20.8	9.26
10	29	21	23	3	22	26	0.0002	20.7	9.14
11	29	21	23	3	30	24	0.0002	21.7	9.79
12	29	21	23	3	30	6	0.0002	18.7	11.54
13	29	21	23	3	30	15	0.0002	20.2	10.05
14	29	21	23	3	30	2	0.0002	18.0	12.49

OSU - Exempel IV



- Samplingfördelningen för $s_{\mathcal{S}}^2$



Estimatorernas osäkerhet

- Oftast är vi intresserade av **osäkerheten** i de punktskattningar vi producerar.
- Osäkerheten kommer från urvalsfördelningen $P(\mathcal{S})$
- Om vi känner till $S^2_{\mathcal{U}}$ (vilket vi aldrig gör) är

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

- Istället använder vi punktskattningen för $S^2_{\mathcal{U}}$, nämligen s^2 och **uppskattar** vår osäkerhet.

$$\hat{Var}(\hat{\bar{y}}_{\mathcal{U}}) = \frac{s^2}{n}\left(1 - \frac{n}{N}\right)$$

  - $\left(1 - \frac{n}{N}\right)$ kallas **ändlighetskorrektion** (finite population corrections)
  - $\frac{n}{N}$ kallas (i dessa sammanhang) **urvalsfraktion** (sampling fraction)

Estimatorernas osäkerhet (forts.)

- Istället för en punktskattnings varians brukar ofta en punktskattnings **medelfel** (standardfel, standard error) användas

$$\hat{SE}(\hat{\bar{y}}_{\mathcal{U}}) = \sqrt{\hat{Var}(\hat{\bar{y}}_{\mathcal{U}})} = \sqrt{\frac{s^2}{n}\left(1-\frac{n}{N}\right)}$$

- **Medelfel** $\hat{SE}(\hat{\bar{y}}_{\mathcal{U}})$ och **urvalsstandardavvikelsen** s^2 är olika saker.

Subsection 3

Konfidensintervall vid OSU

Konfidensintervall

- För att **beskriva** vår osäkerhet bildar vi ofta **konfidensintervall** (KI) för våra estimatorer $\hat{\bar{y}}_{\mathcal{U}}$
- Inom urvalsteorin antar vi **inte** någon fördelning/modell för y - vi har bara samplingfördelningen $P(\mathcal{S})$
- Normalapproximation kan göras med **centrala gränsvärdessatsen (CGS)** om:
  - N och n är "stora" och
  - Fördelningen i urvalet, $y_{\mathcal{S}}$ inte är "alltför" skev
  - Varning för variabler med (vilka ofta kan vara skeva):
    - små proportioner
    - inkomster
  - För tumregler, se Lohr (2009, formel 2.23, s. 44)

Konfidensintervall vid OSU (forts.)

- Om $\bar{y}_{\mathcal{S}}$ är approximativt normalfördelat så

$$\frac{\hat{\bar{y}}_{\mathcal{U}} - \bar{y}_{\mathcal{U}}}{SE(\bar{y}_{\mathcal{S}})} \sim N(0,1)$$

- Så ett $100 \cdot (1-\alpha)\%$ KI för $\bar{y}_{\mathcal{U}}$ är

$$\bar{y}_{\mathcal{S}} \pm z_{\alpha/2} \cdot SE(\bar{y}_{\mathcal{S}}) = \bar{y}_{\mathcal{S}} \pm z_{\alpha/2} \cdot \sqrt{\frac{S^2_{\mathcal{U}}}{n}\left(1-\frac{n}{N}\right)}$$

  där $z_{\alpha/2}$ är den $\alpha/2$ percentilen i normalfördelningen
- $z_{\alpha/2} \cdot \hat{SE}(\bar{y}_{\mathcal{S}})$ brukar kallas för **felmarginal**.

- t-fördelningen används ofta i statistikprogram som R och SAS.
  - Detta för att ta ta hänsyn till osäkerheten i s^2

- Konfidensintervall för $\hat{y}_{\mathcal{U}}$ då $n = 6$, $N = 15$ från det föregående exemplet.
- Det sanna värdet i populationen: $\bar{y}_{\mathcal{U}} = 17.2$.
- Nedan följer de 8 första möjliga urvalen (kombinationer) av totalt $K = 5005$ teoretiskt möjliga urval.

	Obs.1	Obs.6	P_S	y_hat	s_hat	SE_hat	t	KI.low	KI.up	in.KI
1	29	30	0.0002	21.3	9.73	3.08	2.57	13.42	29.2	1
2	29	24	0.0002	20.3	8.94	2.83	2.57	13.07	27.6	1
3	29	6	0.0002	17.3	10.37	3.28	2.57	8.91	25.8	1
4	29	15	0.0002	18.8	8.95	2.83	2.57	11.56	26.1	1
5	29	2	0.0002	16.7	11.33	3.58	2.57	7.46	25.9	1
6	29	4	0.0002	17.0	10.83	3.42	2.57	8.20	25.8	1
7	29	10	0.0002	18.0	9.59	3.03	2.57	10.20	25.8	1
8	29	16	0.0002	19.0	8.88	2.81	2.57	11.78	26.2	1

- Täckningsgraden (baserat på t-fördelningen) för konfidensintervallen är 0.946.

Subsection 4

Urvalsdimensionering

- Hur stort urval ska vi dra? Vanlig (kostnads-) fråga!
- Beror på hur säkra vi vill vara. Bredd på KI.

- Vad kan vi påverka i formeln för KI?

$$\bar{y}_{\mathcal{S}} \pm z_{\alpha/2} \cdot \hat{SE}(\bar{y}_{\mathcal{S}}) = \bar{y}_{\mathcal{S}} \pm z_{\alpha/2} \cdot \sqrt{\frac{s^2}{n}\left(1 - \frac{n}{N}\right)}$$

- Minska konfidensgraden (α)
  - ett KI med $\alpha = 0.2$ är kortare än ett med $\alpha = 0.01$

- Urvalsstorleken (n)
  - ett större urval ger ett smalare KI

- Hur stort urval ska vi dra för att få en viss bredd på KI?
  - Vi behöver veta $S^2_{\mathcal{U}}$
    (tar vi från annan undersökning eller gissar)
- Låt e vara den felmarginal vi vill ha

$$e = z_{\alpha/2} \cdot \sqrt{\frac{S^2_{\mathcal{U}}}{n}\left(1 - \frac{n}{N}\right)}$$

- Lös ut n så fås $(*)$

$$n = \frac{n_0}{1 + \frac{n_0}{N}}, \text{ där } n_0 = \frac{z^2_{\alpha/2} S^2_{\mathcal{U}}}{e^2}$$

- Då $N \to \infty$ så går $n \to n_0$

- Vi vet vill dimensionera en studie för att uppskatta kostnaderna för magsjuka i Katrineholm med $N = 4328$.
- Baserat på tidigare studier antar vi att $S_{\mathcal{U}} = 938$.
- Vi är intresserade av ett konfidensintervall (95 %) på maximalt ± 100 kr.

Vi har att $N = 4328, z_{2.5\%} = 1.96$, $e = 100$ och $S_{\mathcal{U}} = 938$.
Vi använder oss av:

$$n_0 = \frac{z^2_{\alpha/2} S^2_{\mathcal{U}}}{e^2} = \frac{1.96^2 \cdot 938^2}{100^2} = 338.001$$

Sedan använder vi

$$n = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{338.001}{1 + \frac{338.001}{4328}} = 313.516 \approx 314$$

Således behöver vi i detta fall ett urval på 314.

Subsection 5

Totaler och proportioner vid OSU

Totaler och proportioner vid OSU

- Vanliga skattningar av intresse är **totaler** och **proportioner**
- Båda är specialfall av $\bar{y}$
- Total (t)
  - Populationsparameter:

$$t_{\mathcal{U}} = \sum_{i \in \mathcal{U}} y_i = \sum_{i=1}^{N} y_i = N\bar{y}_{\mathcal{U}}$$

- Proportion (p)
  - Populationsparameter av intresse:

$$p_{\mathcal{U}} = \frac{\#\text{ med egenskap av intresse}}{N}$$

  - Om y antar värden $\{0, 1\}$ så

$$p_{\mathcal{U}} = \frac{1}{N}\sum_{i=1}^{N} y_i = \bar{y}_{\mathcal{U}}$$

Estimation av totaler

- På följande sätt estimeras totalen (**totalestimator**) vid OSU

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

- Med medelfelet $(*)$

$$\hat{SE}(\hat{t}_{\mathcal{U}}) = N \cdot \sqrt{\frac{s_y^2}{n}\left(1-\frac{n}{N}\right)}$$

- Detta ger följande KI för $\hat{t}_{\mathcal{U}}$

$$N\bar{y}_{\mathcal{S}} \pm z_{\alpha/2} \cdot N \cdot \sqrt{\frac{s_y^2}{n}\left(1-\frac{n}{N}\right)}$$

Obs! s_y^2 är den uppskattade populationsvariansen för y.

- Vi vill uppskatta de totala sjukvårdsomkostnaderna för magsjuka i befolkningen (15 - 64 år).
- Vi har gjort en undersökning med

$$\begin{aligned} n &= 2731 \\ N &= 4942059 \\ \bar{y}_{\mathcal{S}} &= 76.5 \\ s &= 191 \end{aligned}$$

- Beräkna en skattning av de totala sjukvårdskostnaderna för magsjuka med tillhörande 95%-igt konfidensintervall.

- Detta ger följande totalskattning:

$$\hat{t}_{\mathcal{U}} = N\bar{y}_{\mathcal{S}} = 4942059 \cdot 76.5 = 378067513.5$$

med konfidensintervallet

$$\begin{aligned}
& \hat{t}_{\mathcal{U}} \pm z_{\alpha/2} \cdot N \cdot \sqrt{\frac{s_y^2}{n}\left(1-\frac{n}{N}\right)} \\
= \quad & 378067513.5 \pm 1.96 \cdot 4942059 \cdot \sqrt{\frac{191^2}{2731}\left(1-\frac{2731}{4942059}\right)} \\
= \quad & 378067513.5 \pm 1.96 \cdot 4942059 \cdot \sqrt{\frac{191^2}{2731}(0.999)} \\
= \quad & 378067513.5 \pm 1.96 \cdot 18057616.004 \\
\Rightarrow \quad & [413460440.868 : 342674586.132]
\end{aligned}$$

Estimation av proportioner

- På följande sätt estimeras proportionen (då $y = \{0, 1\}$)

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

- Standardfelet $(*)$

$$\hat{SE}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}\left(1-\frac{n}{N}\right)}$$

  är ett specialfall då $y = \{0, 1\}$.

- Vilket ger följande KI för$\hat{p}_{\mathcal{U}}$

$$\hat{p}_{\mathcal{U}} \pm z_{\alpha/2,n-1} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}\left(1-\frac{n}{N}\right)}$$

Estimering av proportioner - Exempel

- Vi vill estimera andelen personer i befolkningen 16-79 år som utsatts för misshandel under 2011. (se Irlander and Hvitfeldt (2012))

  Slog, sparkade eller utsatte någon dig med avsikt för något annat fysiskt våld, så att du skadades eller så att det gjorde ont, under förra året (2011)?

- Följande uppgifter har vi

$$\begin{aligned} N &= 7297354 \\ n &= 13386 \\ \sum_{i \in \mathcal{S}} y_i &= 363 \end{aligned}$$

- **Obs!** I NTU används stratifierat urval och kalibrerade vikter, så "på riktigt" får vi en annan siffra.

Urvalsdimensionering - Exempel, lösning

- Detta ger följande skattning av populationsandelen:

$$\hat{p}_{\mathcal{U}} = \frac{1}{n}\sum_{i\in\mathcal{S}} y_i = \frac{363}{13386} = 0.027$$

med konfidensintervallet

$$\begin{aligned}
& \hat{p}_{\mathcal{U}} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}\left(1-\frac{n}{N}\right)} \\
= \quad & 0.027 \pm 1.96 \cdot \sqrt{\frac{0.027(1-0.027)}{13386-1}\left(1-\frac{13386}{7297354}\right)} \\
= \quad & 0.027 \pm 1.96 \cdot \sqrt{\frac{0.026}{13385}(0.998)} \\
= \quad & 0.027 \pm 1.96 \cdot 0.001 \\
\Rightarrow \quad & [0.03 : 0.024]
\end{aligned}$$

- Brås skattning är 0.025.

Referenser

Irlander, Å., Hvitfeldt, T., 2012. NTU 2011 : om utsatthet, trygghet och förtroende. Brottsförebyggande rådet, Stockholm.

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