F12

Mera om olika urvalsmetoder: Stratifierat urval, systematiskt urval, gruppurval

Förra gången (F11)

- Om $X \in N(\mu_X, \sqrt{V(X)}) \Rightarrow \overline{X} \in N\left(\mu_X, \sqrt{\frac{V(X)}{n}}\right)$
- Om vi drar alla möjliga urval av en viss storlek (n), beräknar X-bar för alla urval och bildar ett intervall som är $\pm 1{,}96$ standardavvikelser (för X-bar) runt X-bar så vet vi att 95% av dessa intervall kommer täcka det sanna medelvärdet.
  - Antagande: Fördelningen för X-bar är normalfördelad
- Konfidensintervall för μ:

$$\hat{\mu} \pm 1{,}96 \times \sqrt{V(\hat{\mu})}$$

Förra gången (F11)

- Konfidensintervall för μ vid OSU-urval utan återläggning:

$$\bar{x} \pm 1{,}96 \times \sqrt{\left(1 - \frac{n}{N}\right)\frac{s^2}{n}}$$

- Tolkning: Med 95% konfidens ligger det sanna medelvärdet inom gränserna…

Urvalsfel

- Skillnaden mellan erhållet och sant värde
- Kan uttryckas som
  - Variansen/standardavvikelsen för skattningen
  - Felmarginalen
  - Längden på konfidensintervallet

Längden på intervallet

$$längd = 2 \times \left(1{,}96 \times \sqrt{V(\hat{\mu})}\right)$$

- Vad påverkar längden?
  - Ex vid OSU utan återläggning:

$$längd = 2 \times \left(1{,}96 \times \sqrt{\left(1 - \frac{n}{N}\right)\frac{\sigma^2}{n}}\right)$$

  - Stickprovsstorlek, n
  - Varians
  - Konfidensgrad
- Vad kan vi påverka?

Längden på intervallet

- Vad kan vi påverka för att minska längden på intervallet?
  - Ta ett större stickprov
  - Använda en lägre konfidensgrad – men då är vi inte lika ”säkra”
  - Kan vi göra något med variansen?

Exempel OSU och stratifierat urval (ex i KD)

- OSU av element
  - Population: N=1000 personer anställda i ett verkstadsföretag
  - Sökt: 95% k.i. för μ_x=medellönen= $\frac{\sum x_i}{1000}$
  - Välj n=200 anställda genom OSU utan återläggning
  - Skatta μ_x med stickprovsmedelvärdet $\bar{x} = \frac{\sum_{i=1}^{200} x_i}{200}$
  - Skatta populationsvariansen σ^2 med stickprovsvariansen $s^2 = \frac{\sum_{i=1}^{200} (x_i - \bar{x})^2}{n-1}$

Ex forts.

- Numeriska värden:
  $\bar{x} = 109{,}34 kr/tim \quad s^2 = 3{,}814 \quad s = 1{,}953$
- Här gäller att n>30 så att X-bar är approximativt normalfördelad enligt CGV.
- Ett 95%igt k.i. för μ fås som
  $\bar{x} \pm 1{,}96\sqrt{\left(1-\frac{n}{N}\right)\frac{\sigma^2}{n}}$
- Här är σ^2 okänt i praktiken, skattas med s^2:
  $109{,}34 \pm 1{,}96\sqrt{\left(1-\frac{200}{1000}\right)\frac{3{,}814}{n}}$
- Med 95% konfidens ligger medellönen för samtliga anställda i intervallet 109,10 till 109,58 kr/tim.

Kan vi utnyttja resurserna bättre?

- Vi vet att kvinnor och män har olika lön
  - Det finns ett samband mellan kön och lön
- Dela in populationen i två grupper, kvinnor och män, samt dra ett OSU av personer från varje grupp (=från varje **stratum**)
  - Kön är här en **stratifieringsvariabel**

Exempel forts.

Stratum	N_i	n_i
Kvinnor	250	?
Män	750	?
Totalt	N=1000	n=200

- Hur ska vi fördela n=200 mellan grupperna, dvs hur många kvinnor ska väljas och hur många män? Hur ska urvalet *allokeras*?
- Förslag: gör ett *proportionellt stratifierat urval* (PSU)

Proportionellt stratifierat urval (PSU)

- Fördela urvalet i samma proportioner som strata förhåller sig till hela populationen

$$n_1 = 200 \times \frac{250}{1000} = 200 \times 0{,}25 = 50$$

$$n_2 = 200 \times \frac{750}{1000} = 200 \times 0{,}75 = 150$$

- Generellt

$$n_i = n \times \frac{N_i}{N}$$

Exempel forts.

Stratum	N_i	n_i	$\bar{x}_i$	s_i
Kvinnor	250	50	106,87	1,052
Män	750	150	110,45	0,789

- Väg ihop skattningarna från respektive stratum

$$\bar{x}_{str} = \frac{250 \times 106{,}87 + 750 \times 110{,}45}{1000} = 109{,}56$$

- Formelmässigt har vi använt

$$\bar{X}_{str} = \frac{N_1 \times \bar{X}_1 + N_2 \times \bar{X}_2 +}{N} = \frac{N_1}{N} \times \bar{X}_1 + \frac{N_2}{N} \times \bar{X}_2 = W_1 \times \bar{X}_1 + W_2 \times \bar{X}_2$$

- Där W_1 och W_2 är stratum vikterna, i exemplet 0,25 resp 0,75

Proportionellt stratifierat urval (PSU)

- Nu ska osäkerheten mätas i denna skattning $V(\overline{X}_{str}) = V(W_1 \times \overline{X}_1 + W_2 \times \overline{X}_2)$
- Eftersom urvalen från varje stratum görs *oberoende* av varandra, blir stickprovsmedelvärdena från varje stratum oberoende slumpvariabler
- Vikterna är konstanter
- Då gäller:

$$V(\overline{X}_{str}) = W_1^2 \times V(\overline{X}_1) + W_2^2 \times V(\overline{X}_2)$$

PSU

- Eftersom urvalen inom strata görs med OSU utan återläggning, kan variansen för varje stratumstickprovsmedelvärde bestämmas enligt formlerna för OSU
- Då vi också måste skatta stratumvarianserna med motsvarande stickprovsvarianser fås

$$\hat{V}(\overline{X}_{str}) = W_1^2\left(1-\frac{n_1}{N_1}\right)\frac{s_1^2}{n_1} + W_2^2\left(1-\frac{n_2}{N_2}\right)\frac{s_2^2}{n_2}$$

- och med numeriska värden:

$$\hat{V}(\overline{X}_{str}) = 0{,}25^2 \times \left(1-\frac{50}{250}\right) \times \frac{1{,}052^2}{50} + 0{,}75^2 \times \left(1-\frac{150}{750}\right) \times \frac{0{,}789^2}{150} =$$

$$= 0{,}001106704 + 0{,}001867563 = 0{,}002974267$$

PSU: exempel forts.

- Ett 95%igt k.i. för μ fås som

$$109{,}56 \pm 1{,}96\sqrt{0{,}002974267} = 109{,}56 \pm 0{,}11$$

skattning $\pm$ felmarginal

- Med 95% konfidens ligger medellönen i intervallet (109,45; 109,67) kr/timme
- Antagande: stora stickprov inom varje stratum, linjärkombination av oberoende NF-variabler ger en NF variabel
- Jämför med resultatet från OSU med samma urvalsstorlek!

Stratifierat urval

- Vi kan få högre precision (lägre varians) i skattningen vid stratifierat urval jämfört med OSU-urval
- När får vi det?
  - När vi delar in populationen i strata (grupper) som är homogena med avseende på undersökningsvariabeln
  - Om grupperna är lika (homogena) inom sig, så får vi lägre varians inom varje grupp (stratum)
  - Målet när vi delar in populationen i strata är att de ska bli så lika inom strata som möjligt och så olika mellan strata (heterogena) som möjligt med avseende på *undersökningsvariabeln*

Stratifierat urval

- Dela in populationen i strata
- Dra urval från varje stratum

Stratifierat urval

- Att välja stratumgränser
  - Hur ska vi välja antal strata samt stratumgränser så strata blir så homogena som möjligt?
  - Ju fler strata desto mindre precisionsvinst
  - Rekommendation för SDAII: 3-4 strata
  - Stratumgränser?
    - Om undersökningsvariabeln är lön och vi ”vet” att det finns relativt tydliga ”grupperingar” med höga, mellan och låga löner kan vi försöka bilda strata utifrån dessa grupper. Problem?

Problem?

- Vi vill ha homogena strata med avseende på undersökningsvariabeln – ex lön – och vill dela in populationen så att vi får ett strata med låg lön och ett med hög lön alternativt ett med låg, ett med mellan och ett med hög lön.
  - Vi har inte tillgång till värden på undersökningsvariabeln lön!
  - Vi måste ta hjälp av någon *stratifieringsvariabel* (hjälpvariabel) som vi tror har ett SAMBAND med undersökningsvariabeln
  - Vi måste ha en *ram* sorterad efter stratifieringsvariabeln
  - Om vi tror att kön har ett samband med lön kan vi dela in populationen i två strata: män och kvinnor.
  - Om vi tror att ålder har ett samband med lön kan vi försöka hitta homogena strata med avseende på ålder (stratifieringsvariabeln) och om det finns ett samband mellan ålder och lön så hoppas vi att strata även blir homogena med avseende på lön (undersökningsvariabeln)

Planering av ett stratifierat urval

- Vilken stratifieringsvariabel ska väljas?
- Hur många strata?
- Var ska stratumgränserna dras?
- Hur ska allokeringen göras?
  - Hur ska stickprovet fördelas?

Allokering

- Vi drar ett OSU ur varje stratum – hur ska vi välja hur många element som ska dras från varje stratum?
- Proportionell allokering
  - Proportionellt stratifierat urval (PSU)
  - Välj proportionellt sett lika många element ur varje stratum som stratumet utgör av populationen
    - Ex om stratum 1 utgör 30% av populationen – välj 30% av stickprovsstorleken från stratum 1
  - Urvalet blir ”självvägt”
  - Se till att alltid dra minst 2 element från varje stratum!
  - Om ett stratum är ”litet” kan man istället för ett PSU välja att göra en totalundersökning i detta stratum

Allokering

- Vi kan istället för PSU välja att dra lika många element från varje stratum
  - Exempelvis då vi är intresserade av att göra gruppjämförelser såsom att jämföra mäns och kvinnors medellöner
- Optimal allokering (Neyman allokering)
  - Dra fler element från stratum med större varians och färre från stratum med mindre varians
- Inklusionssannolikhet
  - Sannolikheten är inte samma för varje element att väljas (som för OSU)
  - n_i/N_i

Systematiskt urval

- Antag att vi vill undersöka medellönen i ett företag på N=1000 anställda och vill dra ett urval på n=100.
- Antag att vi har en ram sorterad efter personnummer (dvs ålder)
- Vi skulle kunna dra var 10:e personnummer från ramen och undersöka de valdas löner.
- Vi drar en slumpmässig start mellan 1 och 10, säg 4, och urvalet består då av element 4, 14, 24, … 994 från ramen sorterad efter personnummer

Systematiskt urval

- Vi har en ram som består av N element av vilka n ska väljas
- Bilda kvoten r=N/n och avrunda nedåt till närmsta heltal
- Välj med lika sannolikhet ett tal mellan 1 och r
- Dra sedan var r:te element tills *hela ramen* är genomgången

Möjliga urval

- Exempel: vi har en population av N=12 element varav n=3 ska väljas (se exempel i KD)
- Vi drar var 4:de (12/3=4) element
- De möjliga urvalen är:

1	2	3	4
5	6	7	8
9	10	11	12

- Inklusionssannolikheten för varje element är 1/r dvs ¼ i detta fall.
- Vi har lika inklusionssannolikheter för varje element (som vid OSU) men alla tänkbara urval har *inte* samma sannolikhet

Systematiskt urval

- Obs! Viktigt att genomlöpa hela ramen
  - Annars risk för systematiska fel
  - Vad gör vi om vi får ”för många” element?
    - I praktiken går det ofta bra att behålla dem
    - Kan annars slumpa bort överflödiga
- Alternativ: se ramen som en cirkulär förteckning

Systematiskt urval

- Om ramen är sorterad efter någon variabel som är korrelerad med undersökningsvariabeln (det finns ett samband mellan denna och undersökningsvariabeln – jmfr stratifierat urval) får vi ett approximativt PSU urval
  - Högre precision än OSU
  - Medelvärdet blir självvägt
- Se upp för periodicitet!
  - Använd inte systematiskt urval om det finns periodicitet i ramen.
  - Ex om vi väljer var 7:de dag för att studera annonser i tidningen
  - Risk för systematiskt fel samt lägre precision (högre varians) för skattningarna

Gruppurval (klusterurval)

- Vi är intresserade av att undersöka element
- Vi väljer slumpmässigt (ex med OSU) ut *grupper* av element
- Vi undersöker samtliga element i de valda grupperna
  - Enstegs gruppurval
- Vi drar urval av element ur grupperna
  - Tvåstegs (flerstegs) gruppurval

Gruppurval

- Grupperna
  - I motsats till stratifierat urval vill vi ha så *heterogena* grupper som möjligt
- Varför gruppurval?
  - Ramproblem
  - Geografisk spridning
  - Kostnader

Vilken metod ska vi välja?

- OSU
  - Teoretiskt enkelt
  - Kräver ram
- Stratifierat urval
  - Kan ge bättre precision än OSU
  - Bra vid sneda fördelningar
  - Bra vid gruppjämförelser
  - Mer komplicerat att genomföra än OSU
- Systematiskt urval
  - Enkelt att genomföra
  - Kan ge bättre precision än OSU
  - Risk för periodicitet

Vilken metod ska vi välja?

- Gruppurval
  - Billigt per element
  - Ingen element-ram krävs
  - Kan vara geografiskt spridda element
  - Ofta sämre precision än OSU