Det Nationale Forskningscenter for Arbejdsmiljø

Repræsentativitet af panelundersøgelser - forløb og tværsnit

Peter Linde, forsknings- og analysechef, pli@nfa.dk

Dagorden

- Hvad er repræsentativitet?
- Eks 1: Hvordan vægter man, hvis stikprøven er stratificeret så den fortsat er repræsentativ?
- Eks 2: Et stratum – hvordan opdateres stikprøven?
- Eks 3: To virksomhedsstrata. Hvordan opdatering næste gang?
- Opsamling på de tre eksempler
- SAMU generelt
- SAMU og udskiftning i paneler
- SAMU - koordineret og ikke koordineret udvælgelse
- Hvordan hvis ikke brugte SAMU første gang?
- CPR numre og SAMU
- Spørgsmål

Repræsentativitet

- Den danske ordbog:
- ”Typisk eller dækkende for nogen eller noget”

- Tre egenskaber til et repræsentativt udsnit:
- 1) Kan ligge lidt under eller lidt over det rigtige – gerne tæt
- 2) Det skal gælde for **alle** forhold – universelt!
- 3) Hvor flere der udvælges, hvor tættere på det rigtige

- Disse egenskab opfylder kun simpel tilfældig udvælgelse fra hele populationen.
  Det er **universel** repræsentativitet.

Repræsentativitet (2)

Ca. 60.000 udvalgte i omnibus undersøgelser. Har ca. en stikprøvefejl på op til 0,4%

Køn	Pop.	Udvalgt
• Mænd	50,2%	50,2%
• Kvinder	49,8%	49,8%
Alder		
• 16-19 år	6,6%	6,5%
• 20-29 år	17,6%	17,4%
• 30-39 år	15,9%	16,0%
• 40-49 år	18,7%	18,8%
• 50-59 år	18,1%	18,4%
• 60-74 år	23,1%	22,9%

Repræsentativitet (3)

Region	Pop.	Udvalgt
• Nordjylland	10,3%	10,5%
• Midtjylland	22,6%	22,7%
• Syd Danmark	21,0%	21,0%
• Hovedstaden	31,7%	31,5%
• Sjælland	14,4%	14,3%

Strata - vægtning så stikprøven fortsat er repræsentativ

Eksempel 1

- Man skal vægte stratum for stratum
- I hvert stratum er vægten populationen divideret med stikprøven: N/n
- **Man vægter op i det stratum hvor man er valgt, og hvor man har ens udvalgssandsynlighed**

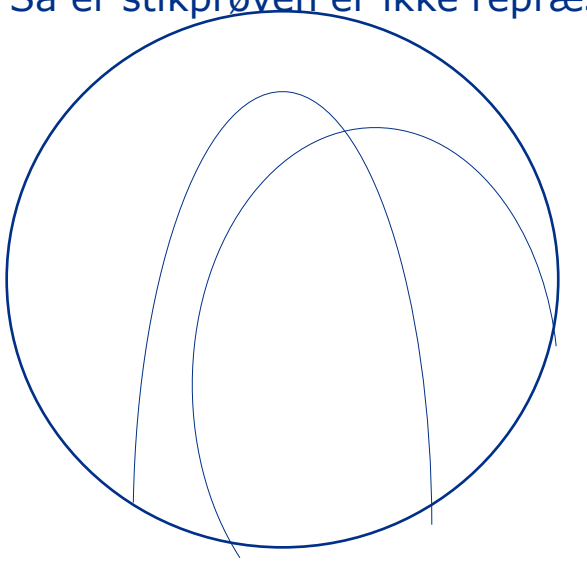
- Eksempel 1: Har en population fra 2018 og har trukket en simpel tilfældig stikprøve.
- Og en population fra 2017, hvor man også har en simpeltilfældig stikprøve.
- Og en population fra 2016, hvor man også har en simpel tilfældig stikprøve
- I 2018 kontaktes de, der er udvalgt i 2018 populationen og de fra 2017 stikprøven, der i er live og en del af 2018 populationen, og de fra 2016 stikprøven, der er i live og en del af 2018 populationen.

Strata - vægtning så stikprøven fortsat er repræsentativ (2)

Vi skal tilbage til begyndelsen – **Noas ark princippet**.

Vi har reelt fire strata i 2018:
Fællesmængden. De der både var en del af 2018, 2017 og 2016 populationen. (tre muligheder for valg)
Fælles 2016 og 2018 populationen, men ikke 2017 populationen (to muligheder for valg)
Fælles 2017 og 2018 populationen, og ikke en del af 2016 populationen (to muligheder for valg).
Udenfor: De der kun er en del af 2018 populationen (en mulighed for valg)

Strata er forskellige, fx ældre. Og forskellige udvalgssandsynligheder: I hver af de fire strata: N/n
Hvad hvis man ikke kender N? – Så er stikprøven er ikke repræsentativ. Fx web paneler.



Et stratum – hvordan opdateres stikprøven
Eksempel 2

- Case: En population, dvs. et stratum. Og simpel tilfældig udvælgelse. Alt lige til og OK.
- Populationen kan være de 18-70-årige, beskæftigede eller en anden gruppe.
- Populationen er på 1.000.000 og vi ønsker at vælge 0,1%, dvs. 1.000 personer.
- Der går fint det første år. Hvis alle svarer har vi vægten 1.000 bag hvert svar.
- Næste gang/år er populationen lidt ændret. 1.050.000 tilhører nu populationen.
- Vi har fået lidt ekstra midler og vil gerne vælge 1.100 denne gang.
- Af de 1.000.000 fra året før er 50.000 ikke mere i populationen, 950.000 fortsat i populationen og 100.000 kommet til. Så 1.050.000 i alt. Og af de 1.000 udvalgte sidst er 950 i den nye population.
- Løsning 1: Vi vælger de ekstra 150 (1.100-950) i den hele nye population på 1.050.000 – dog ikke blandt de 950 allerede udvalgte.
- Løsning 2: Vi vælger de ekstra 150 blandt de 100.000, der er kommet til population.

Et stratum – hvordan opdateres stikprøven (2)

Eksempel 2

Løsning 1: Vi vælger de ekstra 150 i den nye population på 1.050.000 – dog ikke blandt de 950 allerede udvalgte.

Løsning 2: Vi vælger de ekstra 150 blandt de 100.000, der er kommet til population.

Og løsning 3. Vi vælger først 100 ekstra blandet de 100.000, der er kommet til populationen. Så er der balance. Og derefter 50 i hele den nye population på 1.050.000. Så er der stadig balance.

Kommentar 1: Ses ret ofte. De 150 vil ca. fordele sig med 14 fra de nye 100.000 og 136 fra de overlevende 950.000. Dvs. 136 + 950 = 1.086 fra de 950.000 overlevende i populationen og 14 fra de 100.000 nye. Det giver to vægte (Noas ark princippet): 950.000/1.086=875 og 100.000/14=7.142. Dvs. to opregningsstrata med to meget forskellige vægte. Og ellers tab af repræsentativitet.

Kommentar 2: Ses også. Giver også to opregningsstrata hvis repræsentativiteten skal sikres, men ikke helt så forskellige vægte. Nemlig 950.000/950=1.000. Og 100.000/150=667. Derfor faktisk bedre end løsning 1 - mindre stikprøvevarians.

Kommentar 3: **Den selvvejede løsning**. Kræver ikke to opregningsstrata. Alle har samme vægt: 1.050.000/1.100=955.

To strata – hvordan opdateres stikprøven?
- eksempel 3

- Vi har to strata – den kan være virksomheder med 10-100 og 100+ ansatte:
- I stratum 1 vælges 1.000 af 100.000, dvs. 1/100. Vægten for de 1.000 svar er 100.
- I stratum 2 vælges 1.000 af 10.000, dvs. 1/10. Vægten for de 1.000 svar er 10.

Tjek: Populationen er 110.000. Vægtene summer til: 1.000*100+1.000*10=110.000

Året efter er nogle af virksomhederne lukket og andre kommet til. Og nogle har skriftet stratum.

Forskeren vil gerne have så mange som muligt genudvælges og derudover suppleret op så der samlet er valgt 1/100 af de mindst og 1/10 af de store på mindst 100+ ansatte.

Hvordan gøres det?

To strata (2)

Eksempel 3

Først lidt info:

I gl. stratum 1 er 5.000 lukket, 1.000 kommet under 10 ansatte og 2.000 flyttet til stratum 2.

I gl. stratum 2 er 500 lukket, 500 kommet under 10 ansatte og 500 flyttet til til stratum 1.

	Sidst	Lukket	< 10	Stratum 1	Stratum 2	
Gl. stratum 1	100.000	5.000	1.000	92.000 (920)	2.000 (20)	
Gl. stratum 2	10.000	500	500	500 (50)	8.500 (850)	
I alt	110.000	1.000	1.500	92.500 (970)	10.500 (870)	
Nye				7.500	1.500	
Samlet				**100.000**	**12.000**	I alt **112.000**
Kvote (andel som sidst i nye strata)				1.000	1.200	

Løsning 1: 30 (1.000-970) vælges i stratum 1 og 330 (1.200-870) i stratum 2

Løsning 2: 30 vælges blandt de 7.500 nye i stratum 1 og 330 i blandt de 1.500 nye i stratum 2.

To strata
Eksempel 2

Løsning 1: 30 vælges i stratum 1 og 330 i stratum 2

Løsning 2: 30 vælges blandt de 7.500 nye i stratum 1 og 330 i blandt de 1.500 nye i stratum 2.

Og løsning 3: Vi søger en selvvejet løsning inden for hvert stratum.

Kommentar 1: Hvis vi giver alle sammen vægt i stratum 1 (100.000/1.000=100) og stratum 2 (12.000/1.200=10) har vi ikke en repræsentativ stikprøve. Virksomheder i fremgang får for lille vægt og virksomheder i tilbagegang for stor vægt.

Kommentar 2: Løsning 2 har det samme problem som i kommentar 2. Hvis man vil genskabe repræsentativiteten skal man bruge vægte svarende til de vægte de er født med, eller som de fik som nye i populationen da de blev udvalgt i de to strata. Dvs. for de oprindelige vægten 10 eller 100. Og for de nye den konkrete nye vægt. Fx i løsning 2: 7.500/30=250 for nye i stratum 1 og 1.500/330=4,45 for nye i stratum 2.

Kommentar 3: Man kan også forøge hhv. frigøre i de enkelte undergrupper i to strata så der alle steder igen vælges 1/100 i stratum 1 og 1/10 i stratum 2. Konkret frigiver vi fx 45 af de 50 i stratum 1, der kommer fra stratum 2. Og vælger 1/100 del af de 7.500 nye svarende til 75. Så er der i alle delgrupper valgt 1/10 i stratum 1, og alle i stratum 1 har derfor den samme vægt som sidst.

Opsamling

- Søg den selvvejede løsning første gang – vægtene bliver ellers bare værre og værre for hver gang. Det gælder om af skabe balance mellem undergrupper inden for strata (**forbundne badekar princippet**)
- Hvis ikke selvvejet skal der vægte indenfor strata – ellers ryger repræsentativitet og biasen bliver let større end stikprøve fejlen. Og så taber panelet ligesom sin værdi.
- Søg tilbage til fødslen af vægtene (**Noas ark princippey**) , dvs. del op efter designets delgrupper. Og håndter delgruppe for delgruppe. Bedst med et selvvejet design og ellers med vægte.
- Det sidste kræver let en del opdelinger og kan give fejl. Er der ikke en let løsning. Jo ☺

SAMU – en løsning der altid virker

- SAMU står for **Sam**ordnet **U**dvælgelse og stammer fra Sveriges Statistik. Kilde kan skaffes.
- Alle enheder i populationen tildeles et tilfældigt tal mellem 0 og 1. Når man får en ny population beholder de der ”overlever” sit samu-nummer. Og de nye får et tilfældigt tal mellem 0 og 1. På den måde har alle et samu-nummer.
- Inden for hvert stratum – hvis der kun ét bare generelt – bruges samu-nummeret til at bestemme hvem der er med. Hvis der fx skal vælges 1.200 af 12.000 tager man de med det laveste samu-nummer indtil man når 1.200. Dvs. ca. op til 0,1.
- Det giver præcis løsning 3 i eksempel 2 og 3. Og er helt generelt måden let at fjerne og tilføje selvvejet på.

SAMU og udskiftning paneler

- I et panel, hvor man er med år efter år, vil der opstår paneltræthed.
- Der skal derfor udskiftes løbende. Start med det samme, ellers bliver det bare værre og værre.
- Man kunne fx ønske at man var med i 5 år og så slap fri.
- Det løser SAMU let.

1) Det tilfældige tal afrundes til præcist 10 decimaler.
2) De tilfældige tal ordnes efter rækkefølge.
3) Der sættes på plads nummer 11 og 12 to ekstra cifre ind.
4) Startende i 00, 01, 02,…, 99, og så igen 00, 01, 02 osv.
5) De sidste to cifre bruges nu til at ændre samu-nummeret med.
6) Hvis man fx vil frigøre efter 5 år, lægger man et tal til alle samu-numre, der ender på 00 til 19. Det tal sættes så højst at man ikke bliver valgt igen.
   Hvis man fx højst vælge 1/10 lægger vi 0,1 til alle samu-numre, der ender på 00 til 19. Hvis samu-nummeret kommer over 1, trækker vi 1 fra. Så det igen kommer mellem 0 og 1.
   Næste år lægger 0,1 til de samu-numre der ender på 20-39. Osv.

SAMU koordineret og ikke koordineret udvælgelse

- Man kan også bruge SAMU til at styre flere undersøgelser
- Hvis man ønsker de skal have så stort overlap som muligt lader man udvælgelsen starte fra samme samu-værdi. Fx 0,0.
- Hvis man har flere man ikke ønsker der skal udvælgelse til samtidigt lader man dem starte fra tilpas forskellige værdier.

Hvordan hvis ikke SAMU første gang?

- Hvis man ikke fik givet SAMU numre første gang – hvad så?
- Det er altid muligt at give samu-numre og starte på en frisk. Man laver en ”**harmonika**”.
- Hvis man sidste gang fx har 100.000 i population og har udvalgt 1.000. dvs. 1/100:

  De 1.000 udvalgte får et tilfældigt nummer mellem 0,00 og 0,01. Det var jo dem man valgte.

  Og de andre 99.000 et tilfældigt nummer mellem 0,01 og 1,00.

  Og sådan fra stratum til stratum.

CPR numre og SAMU

- Hvis man har adgang til CPR numre (det har alle til statistiske undersøgelser) kan man ikke altid give **hele** population et samu-nummer fordi man bestiller den hos fx cpr kontoret. Så laves et skygge nummer.
- Det løses ved at tildele alle lovlige cpr-numre et samu-nummer.
- Hvert enkelt ciffer af personnummerets ni første cifre ganges med hvert af cifrene i 432765432, deles med 11 - og derpå trækkes 11 fra menten. Nørdet, men muligt. Giver 35.000.000 numre.
- Man bestiller nu relevant mange tilfældige cpr numre baseret på samu, så man har nok til sin stikprøve. I praksis ca. 6 gange flere end kvoten. Konkret fås dermed navn og adresse for de cpr numre, der virker. Og man er i gang med sit repræsentative forskningsprojekt.

Spørgsmål

Tak for ordet