



Sveriges lantbruksuniversitet
Swedish University of Agricultural Sciences

Institutionen för skoglig resurshushållning

Egenskaper hos bra urvalsdesigner

Anton Grafström

Professor i skoglig inventering och sampling

Sveriges lantbruksuniversitet

Institutionen för skoglig resurshushållning

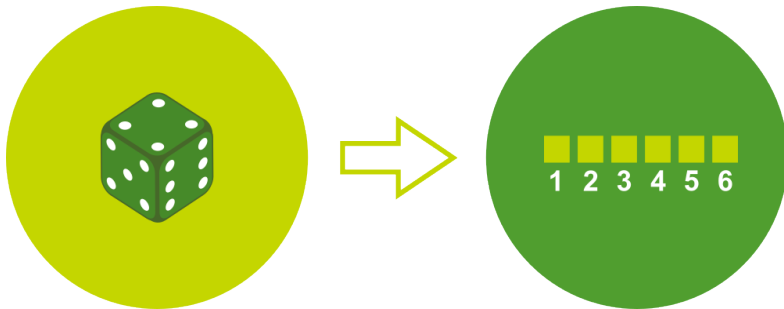


Problem och mål

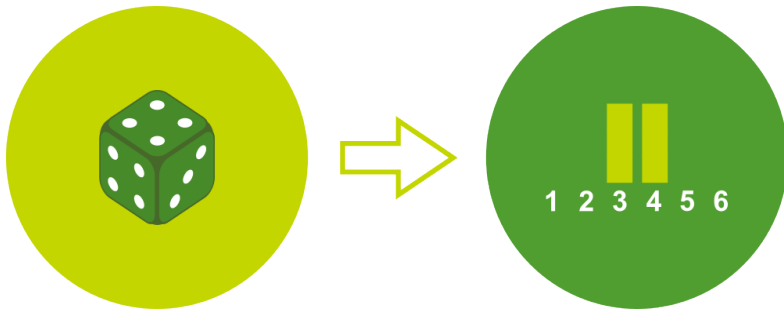
Problem: Ta reda på något om en population genom att välja och observera ett slumpmässigt urval från populationen.

Mål: Skapa eller välja en design som ger mest värde givet de resurser vi har.

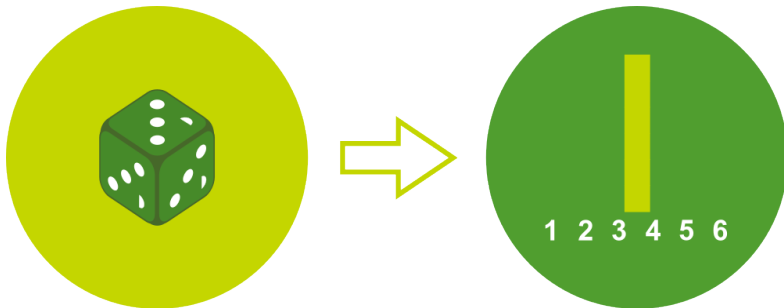
Slumpmässigt urval och kast med tärning



Vad händer om vi fördelar om prickarna?



Tråkig tärning, men underbar design!





Notation

$U - \{1, 2, \dots, N\}$ ändlig population

y - målvariabel

x - hjälpvariabel

S - slumpmässigt urval

i, j - enheter i populationen

π_i - inklusionssannolikhet för enhet i

π_{ij} - andra ordningens inklusionssannolikhet

I_i - inklusionsindikator

$I(A)$ - indikator för händelsen A



En bra design måste använda hjälpinformation, men hur?

Några alternativ

- pps (π ps) - Normalt en hjälpvariabel (helst proportionell mot målvariabel, med ökande spridning för stora värden)



En bra design måste använda hjälpinformation, men hur?

Några alternativ

- pps (π ps) - Normalt en hjälpvariabel (helst proportionell mot målvariabel, med ökande spridning för stora värden)
- Stratifiering - En eller flera variabler



En bra design måste använda hjälpinformation, men hur?

Några alternativ

- pps (π ps) - Normalt en hjälpvariabel (helst proportionell mot målvariabel, med ökande spridning för stora värden)
- Stratifiering - En eller flera variabler
- Balanserade urval (kubmetoden) - Numeriska variabler, stratifiering specialfall



En bra design måste använda hjälpinformation, men hur?

Några alternativ

- pps (π ps) - Normalt en hjälpvariabel (helst proportionell mot målvariabel, med ökande spridning för stora värden)
- Stratifiering - En eller flera variabler
- Balanserade urval (kubmetoden) - Numeriska variabler, stratifiering specialfall
- Rumsligt balanserade, väl spridda urval (avståndsmått, fungerar för alla typer av variabler)

Definitioner, balans och rumslig balans

- Balanserad sampling är att välja slumpmässigt urval med inklusionssannolikheter π_i så att

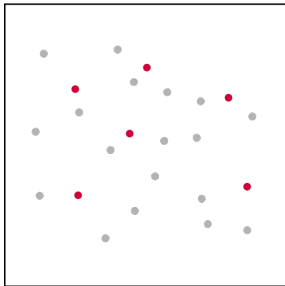
$$\sum_{i \in S} \frac{\mathbf{x}_i}{\pi_i} \approx \sum_{i \in U} \mathbf{x}_i.$$

HT-skattningen av totalerna för hjälpvariablerna är lika eller nästan lika populationstotalerna.

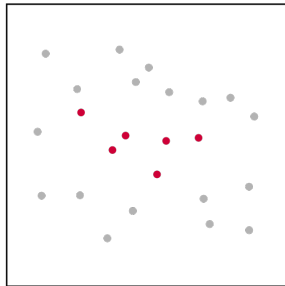
- Rumsligt balanserad sampling är att välja slumpmässigt urval som är så väl spritt som möjligt i hjälpvariablerna samtidigt som inklusionssannolikheterna respekteras.

De två typerna av balans

a)

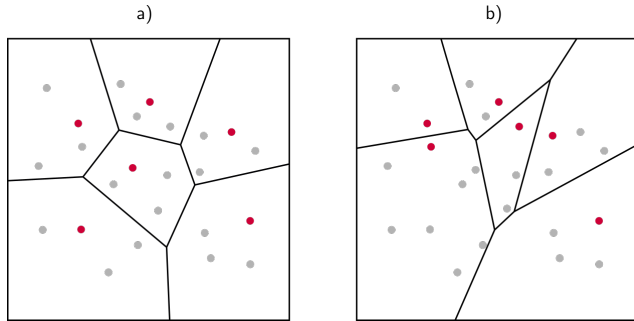


b)



Båda urvalen är balanserade, men endast a) är rumsligt balanserat.
Rumslig balans medför balans.

Voronoipolytoper kan användas för att mäta hur välsprikt ett urval är. Polytopen p_i för enhet $i \in s$ innehåller alla populationsenheter närmare i än någon annan enhet i urvalet.



Låt

$$v_i = \sum_{j \in p_i} \pi_j,$$

så att v_i är summan av inklusionssannolikheterna i polytop i . Ett urval sägs vara välsprikt (eller rumsligt balanserat) om varje v_i är nära 1.

$$B(s) = \frac{1}{n} \sum_{i \in s} (v_i - 1)^2$$

kan användas som ett mått på rumslig balans för ett urval s , och $E(B(S))$ som ett mått på hur välspridda urval en design ger i genomsnitt.



En bra design

1. Använder tillgänglig hjälpinformation
2. Ger (ungefär) samma multivariata fördelning för hjälpvariabler i alla stickprov
3. Ger (ungefär) samma multivariata fördelning för alla målvariabler som (till någon del) förklaras av hjälpvariablerna
4. Fördelar om prickarna på tärningen så att vi får ungefär lika många varje gång (minskar variansen)



Fördelningsfunktioner

Population cdf för x

$$F_U(x) = \frac{1}{N} \sum_{i \in U} I(x_i \leq x)$$

Stickprovets ecdf för x med antagande om fixt n

$$F_S(x) = \frac{1}{n} \sum_{i \in U} I_i I(x_i \leq x)$$



Förväntad fördelningsfunktion

Förväntad värde för stickprovets ecdf för x

$$F_{\pi}(x) = \frac{1}{n} \sum_{i \in U} \pi_i I(x_i \leq x)$$

Om $\pi_i = n/N$, då $F_{\pi}(x) = F_U(x)$.

F_S är väntevärdesriktig för F_π , och för $\pi_j = n/N$ även väntevärdesriktig för F_U . Om vi betingar på x , då får vi följande varians för $F_S(x)$

$$V(F_S(x)) = V \left(\frac{1}{n} \sum_{i \in U} I_i I(x_i \leq x) \right) =$$

$$- \frac{1}{2} \sum_{i \in U} \sum_{j \in U} \frac{\pi_{ij} - \pi_i \pi_j}{n^2} (I(x_i \leq x) - I(x_j \leq x))^2$$

på grund av fix n .



Stickprovets ecdf

Betingat på x , minimeras variansen $V(F_S(x))$ om vi har oberoende mellan i och j då $I(x_i \leq x) \neq I(x_j \leq x)$. Alltså om designen är stratifierad vid x .

Betingat på x , minimeras variansen $V(F_S(x))$ om vi har oberoende mellan i och j då $I(x_i \leq x) \neq I(x_j \leq x)$. Alltså om designen är stratifierad vid x .

För att få F_S nära F_π måste vi stratifiera på alla x -värden på samma gång. Ej möjligt. Det närmaste vi kan komma är att sprida urvalet så bra som möjligt i x , dvs, välja ett urval som är rumsligt balanserat på x . Då kommer varje utfall av F_S vara nära F_π .

Urvalet måste spridas så mycket som möjligt i hjälpvariablerna. Det finns flera metoder som åstadkommer detta!

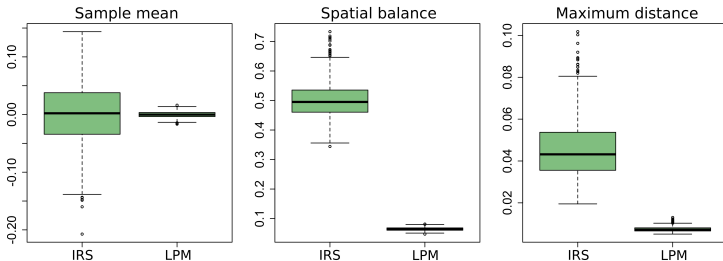
- Local Pivotal Method (LPM)
- Spatially correlated Poisson sampling
- Locally correlated Poisson sampling
- Doubly balanced sampling
- ...



Exempel $N(0,1)$

Låt hjälpvariabelns fördelning vara $X \sim N(0, 1)$. Vi gör en simulering med 1000 urval av storlek $n = 350$, både med oberoende observationer och med LPM.

Exempel $N(0,1)$



Där “maximum distance” är största avståndet mellan stickprovets ecdf och $N(0,1)$. Alla urval med LPM hamnar närmare än det bästa urvalet med oberoende observationer.



Exempel $N(0,1)$

Förväntad rumslig balans är en form av varians för F_S , och mäter hur nära F_S är i genomsnitt till F_π . I detta exempel svarar LPM urval med 350 observationer mot 35 000 oberoende observationer. Det syns tydligt att rumslig balans medför approximativ balans för hjälpvariabeln.



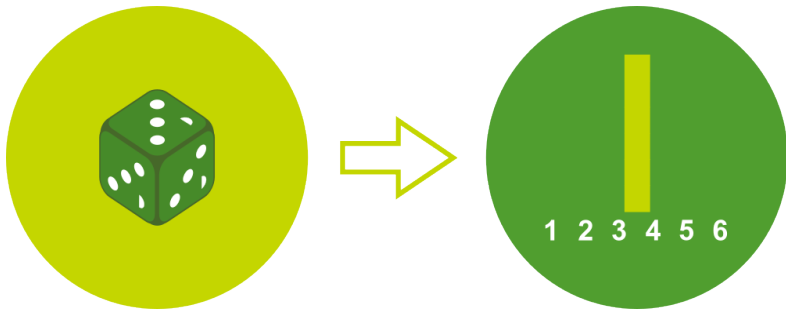
Slutsats

Det är otroligt osannolikt att hamna nära den förväntade fördelningen vid normalstora urval från enkla designner. Vi måste aktivt se till att fånga fördelningen med designen.

Det är otroligt osannolikt att hamna nära den förväntade fördelningen vid normalstora urval från enkla designer. Vi måste aktivt se till att fånga fördelningen med designen.

Det beror på att det finns extremt många urval av storlek n , medan mängden bra urval av storlek n är extremt liten i jämförelse. Sannolikheten att få ett bra urval om vi inte använder hjälpinformation är i princip noll.

Välj den design som bäst fångar fördelningen för hjälpvariablerna. Då har du störst chans att också fånga fördelningen för målvariablerna. Gör simuleringar och använd ett mått för hur mycket den empiriska fördelningen för hjälpvariablerna avviker från sin förväntade fördelning. Alltså, mät rumslig balans.



- Prentius, W., & Grafström, A. (2024). How to find the best sampling design: A new measure of spatial balance. *Environmetrics*, 35(7).
- Grafström, A., Prentius, W., & Lisic, J. (2024). *BalancedSampling: Balanced and spatially balanced sampling*. R-package, version 2.1.1.
- Grafström, A., & Schelin, L. (2014). How to select representative samples. *Scandinavian Journal of Statistics*, 41(2), 277-290.
- Grafström, A., & Lundström, N. L. (2013). Why well spread probability samples are balanced. *Open Journal of Statistics*, 3(1), 36-41.
- Grafström, A., Lundström, N. L., & Schelin, L. (2012). Spatially balanced sampling through the pivotal method. *Biometrics*, 68(2), 514-520.