

Stockholm, 13 mars, 2008

Statistikens Framtid

Lennart Bondesson¹

Sammanfattning

En aning om det förflutna, mer om nuet, diffust om framtiden.
Eller: Lite om mycket.

¹ *Inst f Matematik & Matematisk Statistik, Umeå Universitet*

Lennart B:

63———72-74——-77——-80 ——-83———-94 ——-99———-08

Lund M-L SCB Lund Umeå SLU, Um Uppsala Umeå

Inomvetenskapliga innovationer:

60 ———— X ——— X ——— X ——— X ———— X ——— X ——— X ———— 08 —

T.ex. spatial statistik, ARIMA, överlevnadsanalys, EM, bootstrapping, GREG, MCMC, m.m.

Nya statistiska verksamhetsområden:

60 ———— X ———— X ———— X ———— X ——— X ——— 08 —

T.ex. Kvalitetsteknik, biostatistik, kemometri, bild- och fjärranalys, bioinformatik, finansiell matematik m.m.

Dator & programvaruutveckling, grafik, Internet, data- & textmining, m.m.

Märkta Poissonprocesser?

Lévy-processer med stora hopp och många små hopp?

Tidsserieanalys för prediktion av Statistikens Framtid?

Eller bara spåna?

*The ability to guess the unseen from the seen constitutes our
experience. – Henry James*

Statistiska forskningsnätet i Umeå på Matematik inst:

Spatial statistik (Sara de Luna, LB)

Stokastiska processer (Oleg Seleznev, LB)

Datamining (Oleg Seleznev)

Sampling (Lennart B)

Bioinformatik (Patrik Rydén)

Finansiell matematik (Kaj Nyström)

Biologisk modellering (Åke Brännström)

+ **Doktorander**

Lennart B: Samplingteori

Miljö (nu):

G Kulldorff (grå eminens), I Traat (Tartu), G Ståhl (SLU)

LB

K Meister, A Lundqvist, A Grafström, J Bygren

D Thorburn, Y Tillé, Baltiska gäster, m fl.

π ps-sampling

Population $\{1, 2, \dots, N\}$.

Hur dra n enheter ur den så att man får givna inklusionssannolikheter $\pi_1, \pi_2, \dots, \pi_N$? ($\sum \pi_i = n$)

Systematisk π ps sampling är en möjlighet.

Poissonsampling en annan men ger ej fix stickprovsstorlek.

Historik (med kända namn):

Rao (1963), Brewer (1963) Hajek (1964), Durbin (1967)

Sampford (1967)

Brewer (1975, 1983)

Hájek (1981)

Sunter (1986),

Ohlsson (1990), Rosén (1997a,b), Aires (1999),

Brewer (2002), Deville & Tillé (1998, 2004), Tillé (2006)

Traditionellt:

Ett sampel utan återläggning s är en delmängd av populationen $\mathcal{U} = \{1, 2, \dots, N\}$ med slhsfunktion $p(s)$ så att $\sum_s p(s) = 1$.

Imbi Traat, Tartu (med multivariat bakgrund):

Ett sampel är en binär N -vektor av typ:

$$\mathbf{x} = (x_1, x_2, \dots, x_N) = (0, 1, 0, 1, 1, 1, \dots, 1).$$

Slhsfunktion: $p(\mathbf{x}) = P(\mathbf{I} = \mathbf{x})$.

Denna borde vara central. En bra idé, i all sin enkelhet.

LB indragen pga MCMC intresse.

Conditional Poisson sampling (Rao, Hajek):

$$p^{CP}(\mathbf{x}) = C \prod_{i=1}^N p_i^{x_i} (1 - p_i)^{1-x_i}, \quad |\mathbf{x}| = n$$

Man måste anpassa $p_i : na$ så att $E(I_i) = \pi_i$.

Dragningstekniken bygger på Poissonsampling och acceptans-rejektion.

Sampford sampling (Sampford, 1967):

$$p^S(\mathbf{x}) = C \prod_{i=1}^N \pi_i^{x_i} (1 - \pi_i)^{1-x_i} \times \sum_{k=1}^N (1 - \pi_k)^{x_k}, \quad |\mathbf{x}| = n.$$

Sampfords dragningssteknik är inte snabb.

Men då slhsfunktionen är känd kan man även nyttja MCMC!

Pareto sampling (B Rosén, 1997):

Låt U_1, U_2, \dots, U_N vara oberoende slumpstal från $U(0, 1)$. Välj som sampel de n enheter med minst värden på rankningsvariablerna

$$Q_i = \frac{U_i/(1 - U_i)}{\pi_i/(1 - \pi_i)}. \quad i = 1, 2, \dots, N.$$

Snabb metod. Approximativt korrekta inklusionsslh.

BTL (2006):

$$p(\mathbf{x}) = \prod_{i=1}^N \pi_i^{x_i} (1 - \pi_i)^{1-x_i} \times \sum_{k=1}^N c_k x_k, \quad |\mathbf{x}| = n, \quad c_k \stackrel{approx}{\propto} 1 - \pi_k.$$

Slutsats: Pareto sampling ligger nära Sampfordsampling.

Konsekvens: Vi kan få Sampfordsampel via preliminära Paretosampel och A-R teknik.

Bättre:

$$c_k \stackrel{approx}{\propto} (1 - \pi_k) \left(1 + \frac{\pi_k(\pi_k - \frac{1}{2})}{d}\right) \quad \text{där} \quad d = \sum_{k=1}^N \pi_k(1 - \pi_k)$$

Konsekvens: Rankningsvariablerna

$$\tilde{Q}_i = Q_i \exp\left(\frac{\pi_k(1 - \pi_k)(\pi_k - \frac{1}{2})}{d^2}\right)$$

ger korrektare inklusionssannolikheter än vanlig Paretosampling.

Splitting metoden (Tillé, 2006):

Ett sampel u.å är av formen $\mathbf{x} = (0, 1, 1, \dots, 0, 1)$, men även ett hörn i den N-dimensionella kuben

$$\mathcal{C} = \{\mathbf{x}; 0 \leq x_i \leq 1\}.$$

För att få ett π ps-sampel kan man genomföra en allmän slumpvandring **utan drift** inom och på sidorna och subsidorna av kuben. **START** i $\pi \in \mathcal{C}$ och **STOPP** när hörn har nåtts.

Martingal: $\pi(t) = \pi(t-1) + \mathbf{e}(t)$, där $E(\mathbf{e}(t)|\mathcal{F}_{t-1}) = 0$.

Splitting metoden ger ett rikt urval av π ps-metoder både med fix och variabel sampel storlek. Den ger också möjlighet till olika typer av balanserade urval, **kubmetoden**.

Lite Sen Litteratur:

- Bondesson, L., Traat, I. & Lunqvist, A. (2006) Pareto Sampling versus Sampford and conditional Poisson sampling. *Scand. J. Statist.* **33**, 699-720.
- " ——— " & Thorburn D. (2008). A list sequential sampling method suitable for real-time sampling. To appear in *Scand. J. Statist.*
- " ————" (2008a). Unequal probability sampling designs with high entropy. To appear in *Encyclopedia of Statistical Sciences*, Wiley.
- " ————" (2008b). The splitting method for unequal probability sampling. To appear in *Encyclopedia of Statistical Sciences*, Wiley.
- Grafström, A. (2007). On generalization of Poisson sampling. Submitted.
- Meister, K (2005). *On methods for real time sampling and distributions in sampling*. PhD-thesis, Dept of Math Statistics, Umeå university.
- Tillé (2006). *Sampling Algorithms*. Springer, New York.

Två gamla slutcitat:

Om jag skulle behöva använda statistik för att analysera ett av mina försök skulle jag hellre göra ett nytt försök.

– Ernest Rutherford

Statistiskt tänkande kommer en dag att vara lika nödvändigt för effektivt medborgarskap som förmågan att läsa och skriva.

– H.G Wells

TACK FÖR MIG