STATISTISKA CENTRALBYRÅN

1978-11-01

PROMEMORIOR FRÅN P/STM

NR 1

BAYESIANSKA IDÉER VID PLANERINGEN AV SAMPLE SURVEYS

AV

LARS LYBERG

INLEDNING

TILL

Promemorior från P/STM / Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån, 1978-1986. – Nr 1-24.

Efterföljare:

Promemorior från U/STM / Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån, 1986. – Nr 25-28.

R & D report : research, methods, development, U/STM / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1987. – Nr 29-41.

R & D report : research, methods, development / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1988-2004. – Nr. 1988:1-2004:2.

Research and development : methodology reports from Statistics Sweden. – Stockholm : Statistiska centralbyrån. – 2006-. – Nr 2006:1-.

Promemorior från P/STM 1978:1. Bayesianska idéer vid planeringen av sample surveys / Lars Lyberg.
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-PM-PSTM-1978-1

STATISTISKA CENTRALBYRÅN

1978-11-01

PROMEMORIOR FRÅN P/STM

NR 1

BAYESIANSKA IDÉER VID PLANERINGEN AV SAMPLE SURVEYS

AV

LARS LYBERG

INLEDNING

1 Allmänt om intresset för Bayesiansk statistisk teori

Thomas Bayes (1702-1761) skrev en uppsats kallad "Essay towards solving a problem in the doctrine of chances", som publicerades två år efter hans död. Uppsatsen finns återgiven i Biometrika, vol 45, 1958. Bayes' idéer i form av främst Bayes' teorem kom att inkorporeras i dåtidens statistiska teoriutveckling. Tankarna kom sedermera att överges av modernare teoribildare som R A Fisher och J Neyman-E Pearson. Fisher kunde exempelvis inte acceptera "inverse probabilities" som grundval för inferens och Neyman ansåg att Bayes' teorem hade ett alltför snävt användningsområde. Både Fisher och Neyman accepterade dock Bayes' teorem per se.

Sedan början av 1950-talet har vi kunnat bevittna ett pånyttfött intresse av att utveckla statistisk teori karaktäriserad av användningen av Bayes' teorem i förening med ett subjektivt sannolikhetsbegrepp. Bland ledarna för denna utveckling kan nämnas L J Savage, D Lindley, H Raiffa, R Schlaifer, B de Finetti och W A Ericson. Orsaken till det nyväckta intresset kan väl delvis spåras till de essäer om subjektiva sannolikheter, som i början av detta århundrade författades av F Ramsey och B de Finetti. (Essäerna finns återgivna i Ramsey (1964) och de Finetti (1964).) Det är tänkbart att denna bayesianska teori visar sig så slagkraftig, att en mängd hittills olösbara eller dåligt lösta problem blir lösbara. Den nya falangen statistiker, som kallas bayesianer eller neobayesianer, ser Bayes' idéer och då främst Bayes' teorem som ett hjälpmedel att genomföra subjektivistisk inferens och beslutsfattande. Bayesianerna brukar dock själva framhålla att det inte råder likhet mellan bayesianism och subjektivism. Det karaktäristiska för bayesianismen är att man utnyttjar metoder som reviderar sannolikheter i skenet av ny information.

2 Artikelns syfte

Denna artikel bygger till stor del på Lyberg (1973), som är en uppsats utarbetad inom Tore Dalenius' forskningsprojekt FEL I UNDERSÖKNINGAR, Statistiska institutionen, Stockholms Universitet. Syftet är att redovisa vissa möjligheter att tillämpa bayesianska idéer inom området för urvalsundersökningar (sample surveys).

Inledningsvis diskuteras bayesiansk statistisk teori relativt ingående. Därefter diskuteras tillämpningsmöjligheterna på samplingområdet. Därvid behandlas både generella tillämpningsmöjligheter vid sampling från finita populationer såsom de beskrivs i främst Ericson (1969 a) och mer speciella tillämpningar beskrivna i exempelvis Ericson (1965), Brown (1969) och Zacks (1970). Artikeln avslutas med ett avsnitt om kritiken av bayesianska tankegångar och med ett försök till syntes.

Jag vill i detta sammanhang tacka Claes Cassel och Jan Hagberg för synnerligen värdefulla kommentarer till ett tidigare utkast.

AVD A BAYESIANSK TEORI - EN ÖVERSIKT

3 Allmänt om statistiska teorier

Gemensamt för de flesta statistiska teorier är att de är orienterade mot behandlingen av osäkerhetsproblem. Teorierna skiljer sig åt bl a med hänsyn till arten av osäkerhet, som är föremål för teorin och sättet att angripa osäkerheten. Den teori som förknippas med t ex Neyman-Pearson behandlar det slags osäkerhet, som föreligger hos fenomen kallade stokastiska experiment. Sannolikhetsteorin är en modell för detta slags fenomen. Den blir därför sådan, att den tolkas i frekvenstermer. Anhängare av dessa tankegångar kallas bl a frekventister, objektivister eller "klassiker".

Men det finns också förespråkare för andra synsätt. I Savage (1962) gör M S Bartlett följande distinktion beträffande dessa övriga.

a) Vissa är fortfarande frekventister

b) Vissa förespråkar utnyttjande av andra sannolikheter, t ex à priorisannolikheter och kanske mått på nyttan. Dessa i sin tur kan indelas i

i) sådana som betraktar dessa sannolikheter som personliga eller subjektiva (bayesianer)

ii) sådana som betraktar dessa sannolikheter som opersonliga eller objektiva.

Gemensamt för neo-bayesianska teorier är att osäkerheten är en beskrivning av vår bristande kunskap om något (en parameter eller dylikt). Osäkerheten uttryckes i termer av subjektiva sannolikheter. Detta innebär givetvis att den ibland även kan utgöra inslag i en modell, där fenomen observeras i termer av frekvenser. Bayesiansk teori har alltså (sett ur bayesiansk synvinkel) ett vidare tillämpningsobjekt än stokastiska experiment.

Det bör dock här upprepas att bayesianerna själva inte sätter likhetstecken mellan subjektivism och bayesianism.

4 Bayesiansk statistisk teori

4.1 Bayes' teorem

Grunden i den bayesianska doktrinen står att finna i Bayes' teorem. Det kan ses som ett instrument att välja mellan hypoteser på grundval av data och lyder:

$$P(A|B) = \frac{P(B|A)\, P(A)}{P(B)}$$

Här uttrycks den betingade sannolikheten för en händelse A givet en annan händelse B, såsom en funktion av den betingade sannolikheten för B givet A och av den obetingade sannolikheten för A. $1/P(B)$ betraktas ofta som en proportionalitetsfaktor på så sätt att teoremet skrivs som

$$P(A|B) = \alpha\, P(B|A)\, P(A).$$

Teoremet är inte begränsat till två hypoteser. I sin allmänna form ser det ut på följande sätt.

$$P(A_i|B) = \frac{P(B|A_i)\, P(A_i)}{\sum_{j=1}^{k} P(A_j)\, P(B|A_j)}$$

där k är antalet hypoteser. Ibland skrivs teoremet med A utbytt mot "naturtillståndet" θ, där θ är en diskret variabel, och B mot utfallet t. $P(\theta|t)$ kallas à posteriorisannolikheten för θ och $P(\theta)$ kallas à priorisannolikheten för θ.

Om man med en sannolikhet menar en numerisk representation av ett informationstillstånd, så måste denna representation revideras då ny information tillkommer. Bayes' teorem utgör en möjlighet att formalisera en sådan revidering.

À priorisannolikheten $P(A)$ revideras till $P(A|B)$ då B observeras. $P(B|A)$ är den s k likelihooden för sampleinformationen B givet att A är sann. Denna à posteriorisannolikhet kan sedan användas som à priorisannolikhet om ytterligare information insamlas.

För att vi ska se hur teoremet fungerar betraktar vi ett enkelt exempel. Det är hämtat från Phillips (1973).

Det existerar två urnor som för en betraktare verkar identiska. Vi vet dock att den ena innehåller 70 blå kulor och 30 röda medan den andra innehåller 40 blå och 60 röda. Vi kallar urnorna X och Y.

Med hjälp av exempelvis slantsingling med ett unbiased mynt väljer vi nu en av urnorna. Eftersom vi inte känner innehållet vet vi inte vilken urna som valts. Det är det vi vill göra ett uttalande om.

Låt oss ansätta à priorisannolikheterna 0,5 för de båda tänkbara alternativen. Vi drar nu slumpmässigt en kula ur den valda urnan, och noterar dess färg varefter kulan återlägges. Förfarandet upprepas tills, säg, 12 kulor dragits varvid vi konstaterar att 8 är blå och 4 är röda. Det verkar ju onekligen som om urna X dragits med tanke på att vi vet att dess innehåll till största delen består av blå kulor, men hur sannolikt är det? Vilka är à posteriorisannolikheterna för de båda urnorna?

Våra hypoteser är:

A_1: Urna X har valts
A_2: Urna Y har valts

De data, B, vi har är att samplet innehöll 8 blå och 4 röda kulor.

À priorisannolikheterna är $P(A_1) = P(A_2) = 0.5$

Likelihood ges av:

$P(B|A_1) = \binom{12}{8} (0,7)^8 (0,3)^4 = 495 \cdot 0{,}000467$

$P(B|A_2) = \binom{12}{8} (0,4)^8 (0,6)^4 = 495 \cdot 0{,}0000849$

och täljarna i Bayes' teorem blir

$P(B|A_1) \cdot P(A_1) = 495 \cdot 0{,}000467 \cdot 0{,}5 = 495 \cdot 0{,}0002335$

$P(B|A_2) \cdot P(A_2) = 495 \cdot 0{,}0000849 \cdot 0{,}5 = 495 \cdot 0{,}00004245$

Nämnaren blir

$$\sum_{i=1}^{2} P(B|A_i) \cdot P(A_i) = 495 \cdot 0{,}00027595$$

À posteriorisannolikheterna blir

$$P(A_1|B) = \frac{495 \cdot 0{,}0002335}{495 \cdot 0{,}00027595} = 0{,}85$$

$$P(A_2|B) = \frac{495 \cdot 0{,}00004245}{495 \cdot 0{,}00027595} = 0{,}15$$

Med bayesianskt uttryckssätt kan vi nu konstatera att vi är 85 %-igt säkra på att urna X var den som valdes. Innebörden i ett uttalande av detta slag återkommer vi till.

Bayes' teorem kan naturligtvis också tecknas för kontinuerliga slumpvariabler. I detta fall får det utseendet

$$f(\theta|y) = \frac{h(\theta)\, g(y|\theta)}{\int_{-\infty}^{\infty} h(\theta)\, g(y|\theta)\, d\,\theta}$$

där $\int_{-\infty}^{\infty} h(\theta)\, g(y|\theta)\, d(\theta) > 0$

och där $h(\theta)$ är à priorifördelningen för en stokastisk variabel H, där $f(\theta|y)$ är à posteriorifördelningen för H och där $g(y|\theta)$ är likelihoodfunktionen för ett utfall y givet $H = \theta$.

Bayes' teorem är matematiskt korrekt och oantastligt. Det är således inte på denna punkt som någon skiljelinje uppstår mellan frekventister och bayesianer. Det är möjligheten att ansätta en subjektiv sannolikhetsfördelning $h(\theta)$ som inte kan accepteras av de förra. I de fall sannolikheterna är inslag i en modell, där fenomen observeras i termer av frekvenser, kan satsen användas av även icke-bayesianer. Bayes' teorem behandlas utförligt i Cornfield (1967), Edwards et al (1963) och Placket (1966).

Låt oss nu betrakta ytterligare en illustration av användningen av Bayes' teorem.

Exemplet är hämtat från Roberts (1963). Antag att vi skall inferera till en Bernoulliprocess med parametern p. Om inga direkta sampleutfall från processen finns att tillgå, så kan man tänka sig att ansätta en subjektiv sannolikhetsfördelning för $\hat{p}$ (parametern p betraktar vi nu som en stokastisk variabel, vilket symboliseras av $\hat{p}$). Denna fördelning utgör alltså nu en à priorifördelning. På grundval av exempelvis erfarenheter från liknande processer eller andra kunskaper ansätts i detta exempel en i intervallet $[0.1]$ likformig fördelning för $\hat{p}$. Låt Bernoulliprocessen ifråga spegla upprepade oberoende kast med ett mynt och att parametern p är sannolikheten för krona i ett visst kast. Antag nu att vi vid tre kast med myntet observerar krona, krona, klave. Sannolikheten att observera detta sample givet p är $p^2(1-p)$. Detta är alltså samplets likelihood. Via Bayes' teorem skall vi nu erhålla en à posteriorifördelning för $\hat{p}$. I fördelningstermer kan vi skriva

$$f''(p|r,n) = \frac{f'(p)\ p^r(1-p)^{n-r}}{\int_0^1 f'(p)\ p^r(1-p)^{n-r}dp}$$

där $f'(p)$ är à priorifördelningen för $\hat{p}$, där $p^r(1-p)^{n-r}$ är likelihoodfunktionen om r krona noteras i samplet och där $f''(p|r,n)$ utgör à posteriorifördelningen för $\hat{p}$ givet sampleutfallet. I exemplet är $f'(p) = 1$, $(0 \leq p \leq 1)$, $r = 2$, $n = 3$ och

$$\int_0^1 f'(p)\, p^r(1-p)^{n-r}dp = \int_0^1 p^2(1-p)dp = 1/12$$

$$\therefore\ f''(p|r = 2, n = 3) = \begin{cases} 12\, p^2(1-p), & 0 \leq p \leq 1 \\ 0 & \text{eljest} \end{cases}$$

Vi erhåller således som slutprodukt en sannolikhetsfördelning för $\hat{p}$. Denna fördelning speglar våra bedömningar av $\hat{p}$ i skenet av sampleinformationen. Med hjälp av fördelningen kan vi nu enkelt räkna fram sannolikheten för olika $\hat{p}$-värden. Vi kan ange x %-iga tilltrosintervall och vi kan ange à posteriorifördelningens medelvärde om det är intressant.

4.2 Subjektiva sannolikheter

Som nämnts ovan existerar flera sannolikhetsbegrepp i den statistiska teorin. Vid sidan om den traditionella frekvenstolkningen existerar olika varianter med ett inslag av subjektivism i sannolikhetsbegreppet. Några olika definitioner har givits i Ramsey (1964), de Finetti (1964), Savage (1962 b) och Anscombe och Aumann (1963). I den subjektiva tolkningen av sannolikhetsbegreppet är sannolikheten ett mått på graden av tilltro hos en individ (degree of belief) visavi en händelse A. En möjlighet att formellt bestämma storleken av denna sannolikhet ligger i att definiera den i termer av odds. Vi kan tala om en vadslagningssituation. Detta innebär att om en person subjektivt tillordnar en sannolikhet som mått på en viss grad av sin tilltro till något, så skall detta tillordnade speglas av en motsvarande villighet att anta ett vad. Om oddsen till förmån för en händelse är a mot b, så är sannolikheten för denna händelse a/(a+b). Andra definitioner innehåller ett mått av nytta.

Om en person tillordnar subjektiva sannolikheter för ett fenomen, gör han detta på ett inkonsistent eller inkoherent sätt om han, i en vadslagningssituation, bjuder emot sig själv. Nödvändiga och tillräckliga koherensvillkor är

$0 \leq P(A) \leq P(S) = 1$

$P(A \cup B) = P(A) + P(B)$

där S är den universella händelsen, och där A och B är två godtyckliga varandra uteslutande händelser.

Subjektiva sannolikheter kan sägas utgöra subjektiva attityder gentemot sinnevärlden. Dessa attityder varierar naturligtvis mellan individer.

I Raiffa (1962) rättfärdigas ansättandet av en subjektiv sannolikhetsfördelning för en parameter och användandet av denna fördelning på samma sätt som vore den frekvensbaserad med följande motiveringar.

i) Icke-bayesianernas konfidensintervall, signifikansnivåer etc är endast försök att "designa" bort användningen av subjektiva sannolikheter.

ii) Det existerar ett koherensvillkor d v s man förutsätter ett konsistent uppträdande av den subjektive.

iii) Det är möjligt att ge en operationell mening åt beslutsfattarens subjektiva sannolikhetsfördelningar i termer av dennes grundläggande val i beslutssituationer.

iv) Om en liberal tolkning av sannolikhetsbegreppet tillåtes, så ökar tillämpningsområdet för den matematiska sannolikhetsteorin avsevärt. Det gäller exempelvis beslutsproblem inom medicin, ekonomi, juridik etc.

Bayesianerna erkänner dock att allmän vaghet och obeslutsamhet gör att inte alla fenomen har en väldefinierad sannolikhet.

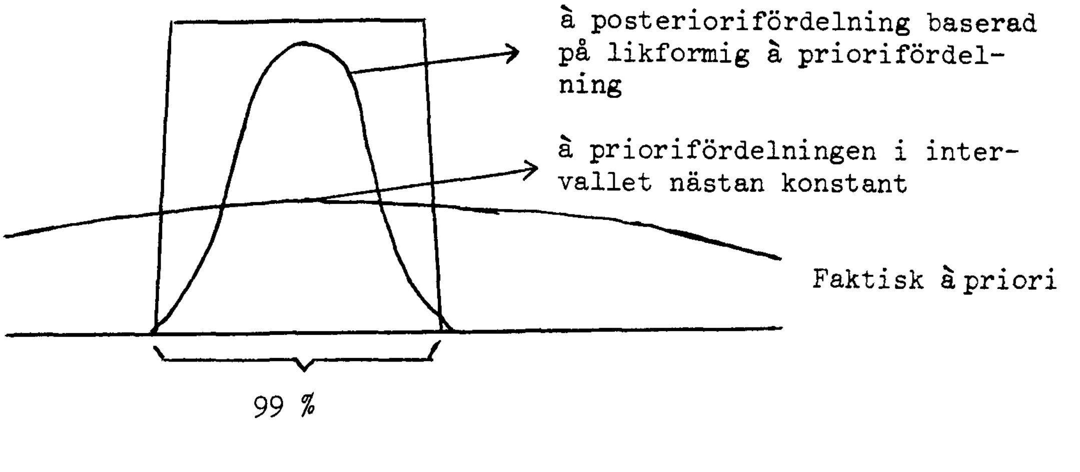
Den subjektiva sannolikhetsfilosofin behandlas utförligt i bl a Edwards et al (1963), Clutton-Brock (1965), Cornfield (1969), Smith (1965), Savage (1961), Good (1950), de Finetti (1972), de Finetti (1974 a) och de Finetti (1974 b).

4.3 Några attraktiva egenskaper

I bl a Savage (1962 b), Edwards et al (1963) och Roberts (1963) talas om vissa speciella och attraktiva egenskaper hos den bayesianska statistiska teorin. Bland dessa egenskaper kan nämnas de två som kallas stabil estimation och likelihood-principen.

Många gånger har à priorifördelningen inte så stor betydelse för à posteriorifördelningens utseende. Den senare bestäms istället i stor utsträckning av sampleinformationen. Om det finns mycket sådan information, så kan olika tillordnare av à priorifördelningar komma att närma sig varandra à posteriori-fördelningsmässigt efter det att hänsyn tagits till sample-informationen. Detta är innebörden av <u>stabil estimation</u>. Man kan också säga att innebörden är att man ibland kan anta en likformig à priorifördelning fastän den faktiska à priori-fördelningen inte är likformig, nämligen då den faktiska à priorifördelningen är nästan konstant i det område som täcks av ett, säg, 99 %-igt tilltrosintervall (se avsnitt 4.5) baserat på à posteriorifördelningen. Se figuren nedan.

Fig



Låt mig göra några försök till klarlägganden. Med faktisk à priorifördelning menar jag en som tillordnats formellt. Den ska spegla tillordnarens tilltro beträffande variabelns fördelning. Tilltron kan vara baserad på subjektiva bedömningar eller kända frekvenser. Den behöver alltså inte vara likformig, bara den uppfyller kravet om att vara nästan konstant enligt ovan.

En annan fråga är följande om nu à prioriinformationen helt överskuggas av sampleinformationen, så kan man undra var vinsten ligger. Ja, föremålet för den statistiska analysen är ju fortfarande à posteriorifördelningen.

Betydelsen av sampleinformationen speglas i likelihoodfunktionen. Likelihoodprincipen innebär följande. Om x_1 och x_2 är två "datamängder" med samma likelihoodfunktion frånsett en konstant faktor k (dvs $p(x_1|\theta) = k\, p(x_2|\theta)$ för alla $\theta \in \Theta$, där k är oberoende av θ), så skall beslut och inferens vara identiska för x_1 och x_2. Denna princip medför vissa konsekvenser som inte står i samklang med den traditionella synen. Exempelvis, så görs ingen skillnad mellan sekvensanalys och analys vid fix urvalsstorlek när data väl finns. Den s k stoppregeln är irrelevant. Invers binomial sampling är således ekvivalent med den obundet slumpmässiga motsvarigheten med fix urvalsstorlek n. I detta fall är likelihoodfunktionen i båda metoderna $\theta^r(1-\theta)^{n-r}$, men ändå skattar frekventisterna θ med r/n respektive (r-1)/(n-1).

4.4 À priorifördelningen

4.4.1 Allmänt

Valet av à priorifördelningar är ett viktigt steg i bayesanalysen, trots det som ovan sagts om stabil estimation. À priorifördelningen skall spegla den uppfattning man har om en parameter θ. Det finns situationer, där à priorifördelningens utseende inte kan negligeras. Några sådana situationer är då

i) både à priorifördelningen och sample informationen är vad man kallar "vag" eller "diffus"

ii) observationskostnaderna är så höga att det blir alltför dyrt att sampla för en stabil estimation

iii) man befinner sig i en beslutssituation som förutsätter detaljuppfattningar om à priorifördelningen.

Men i andra situationer behöver inte vagheten ha någon större betydelse, eftersom à posteriorifördelningen kan vara relativt okänslig för à priorifördelningens utseende. Man kan ha en känsla av att à priorifördelningen är så gott som likformig. En sådan vaghet kan uttryckas med en diffus, icke-informativ eller lokalt likformig à priorifördelning. En likformig à priorifördelning är användbar för statistiska rapporter eftersom den leder till à posteriorifördelningar från vilka likelihoodfunktionen är lätt härledd. Användandet av vissa typer av diffusa à priorifördelningar ger resultat som ofta är numeriskt ekvivalenta med de klassiska inferensmetodernas. I Bos och Tiao (1965) redovisas en sådan studie.

Motexempel saknas dock inte. I Stone (1976) visas hur likformiga à priorifördelningar kan leda till s k strong inconsistencies. Det kanske ska tilläggas att motexempel till motexemplen inte heller saknas.

Vad gör man då i en situation där man inte behöver förutsätta detaljuppfattningar om à priorifördelningen? Ja, då ska man välja à priorifördelning från en fördelningsfamilj F som, med Savages' språkbruk, passar en elefant.

Dessutom skall F vara analytiskt attraktiv på så sätt att

i) det med lätthet skall gå att räkna fram à posteriorifördelningen för en given à priorifördelning och ett givet urval

ii) om à priorifördelningen är medlem av F så skall à posteriorifördelningen vara medlem av F, det s k konjugatvillkoret. Konjugatvillkoret är ett sätt att klara svåra beräkningar vid tillämpning av Bayes' teorem. För närvarande spelar konjugatfördelningarna i bayesiansk forskning endast rollen som "beräkningsförenklare".

4.4.2 Konstruktion av à priorisannolikheter

Bayesananalysens första steg är konstruktionen av à priorisannolikheter. Det finns en hel del skrivet om sådana konstruktioner. Några referenser är Winkler (1967 a), (1967 b), (1969), Savage (1971) och Staël v Holstein (1970). Mer formella ansatser för att skatta en à priorifördelning med hjälp av observationerna ges i Clutton-Brock (1965) och Hendricks (1964).

En "god" sannolikhetstillordnare bör uppfylla två villkor. Dels skall koherensvillkoret vara uppfyllt, dels skall tillordnaren besitta nödvändiga kunskaper om den parameter eller händelse som betraktas. Sällan är båda dessa villkor uppfyllda samtidigt. Det går emellertid att träna upp tillordnare och med en effektiv feedback kan bedömningen hos personen förbättras. Tillordnarens förmåga kan laboratoriemässigt evalveras löpande med hjälp av s k straffunktioner, poängregler o d. I praktiken uppstår dock problem av olika slag.

En vanlig ansats vid tillordnandet av subjektiva sannolikheter är att härleda en kumulativ fördelningsfunktion. Detta innebär att man bestämmer fraktiler genom s k subdivisioner eller direkta frågor snarare än att be tillordnaren om en täthetsfunktion. I Morrison (1967) ges en metod att testa tillordnarens förmåga att konsistent bestämma sådana fraktiler. Ett exempel på denna subdivisionsteknik är följande:

a) Man ber tillordnaren bestämma en punkt $x(0.5)$ som är sådan att det för denne är lika troligt att θ är mindre än $x(0.5)$ som att θ är större än $x(0.5)$.

b) Man meddelar nu tillordnaren att θ i själva verket är mindre än $x(0.5)$ och ber denna att <u>nu</u> bestämma en punkt $x(0.25)$ sådan att det lika troligt att $\theta<x(0.25)$ som att $x(0.25)<\theta<x(0.5)$.

c) Man fortsätter att fråga tillordnaren tills man har tillräckligt många fraktiler för att kunna forma en fördelning.

Detta är ett exempel på en indirekt metod att ansätta à priorifördelningar. Den direkta metoden består i att tillordnaren direkt anger hela fördelningen.

Ansättningen kan enligt Schlaiffer (1969) i detta senare fall grunda sig på

i) rent subjektiv bedömning
ii) "historiska" data eller
iii) prognoser och estimat.

Ytterligare en dimension i detta sammanhang är att det slutliga tillordnandet kan ske av en enda person eller alternativt en grupp av personer. En genomgång av litteraturen på ansättningsområdet ges i Hampton et al (1973) och i Hogarth (1975). I den senare redovisas en kritisk granskning av litteraturen på detta område och där man speciellt diskuterar det meningsfulla i att ansätta subjektiva sannolikheter.

4.4.3 Consensusproblemet

I den mån man har flera personer som tillordnar subjektiva sannolikhetsfördelningar så uppstår det s k consensusproblemet. Hur skall man jämka ihop de subjektiva fördelningarna $f_1(\theta)$, $f_2(\theta), \ldots, f_n(\theta)$ till en enda fördelning $f(\theta)$ som kan användas som à priorifördelning för θ? Den metod som ligger närmast till hands är att göra en linjärkombination. Varje tillordnare erhåller en vikt w_i. Därefter kan en aggregerad täthetsfunktion $f(\theta)$ erhållas såsom $\Sigma w_i \, f_i \, (\theta)$. Svårigheten med detta förfarande är givetvis att bestämma vikterna. I Staël v Holstein (1970) och Winkler (1968) föreslås flera viktningsregler.

En annan lösning på consensusproblemet är att jämka ihop de n fördelningarna i ett feedbacksystem. Man kan då hoppas att uppfattningarna om θ i skenet av den totala mängden uppfattningar skall kunna konvergera mot en enda, som alla n tillordnare kan enas om.

I Gough (1973) redovisas några alternativa aggregeringsmetoder. Där hävdas att viktbestämningen påverkar den slutliga à priorifördelningen mindre än valet av aggregeringsmetod.

I De Groot (1974) presenteras en annan consensusmodell där vikterna kan beräknas och där det kan avgöras om en consensus alls är möjlig. Modellen postulerar att var och en av tillordnarna i gruppen själv viktar sin egen och de övrigas fördelningar. Successiva revideringar ger sedan om möjligt en slutlig à priorifördelning.

4.4.4 Några användbara fördelningar

I syfte att förenkla analysen har stor uppmärksamhet riktats mot de naturligt konjugata à priorifördelningarna, som nämnts ovan. Givetvis måste en naturligt konjugat à priorifördelning åtminstone approximativt spegla den existerande à prioriinformationen för att fördelningen skall kunna utnyttjas.

I Grizzle och Novick (1965) ges exempel på hur en binomialfördelad à priorifördelning ger en betafördelad à posteriorifördelning.

En vanlig situation är den där à prioriinformationen är diffus relativt sampleinformationen. Den vanliga ansatsen är att då använda sig av en likformig fördelning för θ i hela intervallet. Kritik kan dock riktas mot ett sådant val. Om en likformig fördelning speglar θ bra, så borde den också kunna spegla θ^2, θ^3 och andra monotona funktioner. Det är inte säkert att den gör det.

4.5 Några ytterligare tekniska aspekter

De båda återstående komponenterna i Bayes' teorem, likelihoodfunktionen och à posteriorifördelningen, har redan berörts ovan. Likelihoodfunktionen som speglar sampleinformationen omformar à priorifördelningen till en à posteriorifördelning. Denna senare sannolikhetsfördelning utgör den bayesianska analysens slutprodukt. Detta att presentera en hel fördelning är en skiljelinje jämfört med den klassiska analysen, där man oftast presenterar konfidensintervall. Det är vanligt att bayesianer tillsammans med à posteriorifördelningen även presenterar någon form av intervall. Dessa kan kallas tilltrosintervall (credible intervall). Ett Y %-igt tilltrosintervall är ett där man mellan intervallets ändpunkter kan återfinna Y% av den yta som täcks av à posteriorifördelningen. Ibland betraktar man det kortaste av alla dessa möjliga intervall som ett "bästa" intervall för en parameter. I Peers (1968) redovisas teorin för tilltrosintervall närmare.

I Reilly (1976) beskrivs en metod för numerisk beräkning av à posteriorifördelningen med hjälp av datorteknik. Vid datorlagringen betraktar man parameterfördelningarna som vektorer vilket leder till flexibla ansatser inom exempelvis området för icke-linjär regression.

AVD B TILLÄMPNINGEN I SAMPLE SURVEYS

5 Allmänt

Inom området för sample surveys används ofta à prioriinformation om kostnads- och varianssituationen liksom allmänna erfarenhetsdata om specifika felkällor i form av exempelvis bortfall för att erhålla bättre tillförlitlighet i resultaten. Sådan information är ofta subjektiv även om den är baserad på historiska data[1]. Det vi här kommer att betrakta är den situation där à priorisannolikheter revideras i skenet av data; dvs vi studerar hur subjektiv à prioriinformation formellt kan inkorporeras i ett surveydesign med hjälp av Bayes' teorem. Framställningen blir relativt "smal" vilket är en spegling av forskningens nuvarande nivå på detta område. De bayesianska tankegångarna är emellertid på stark frammarsch vilket visat sig i en allt större andel bayes-artiklar och läroböcker med inslag av bayesiansk filosofi.

Innan vi går vidare ges en kortfattad beskrivning av problemområdet. Framställningen bygger på Dalenius (1971).

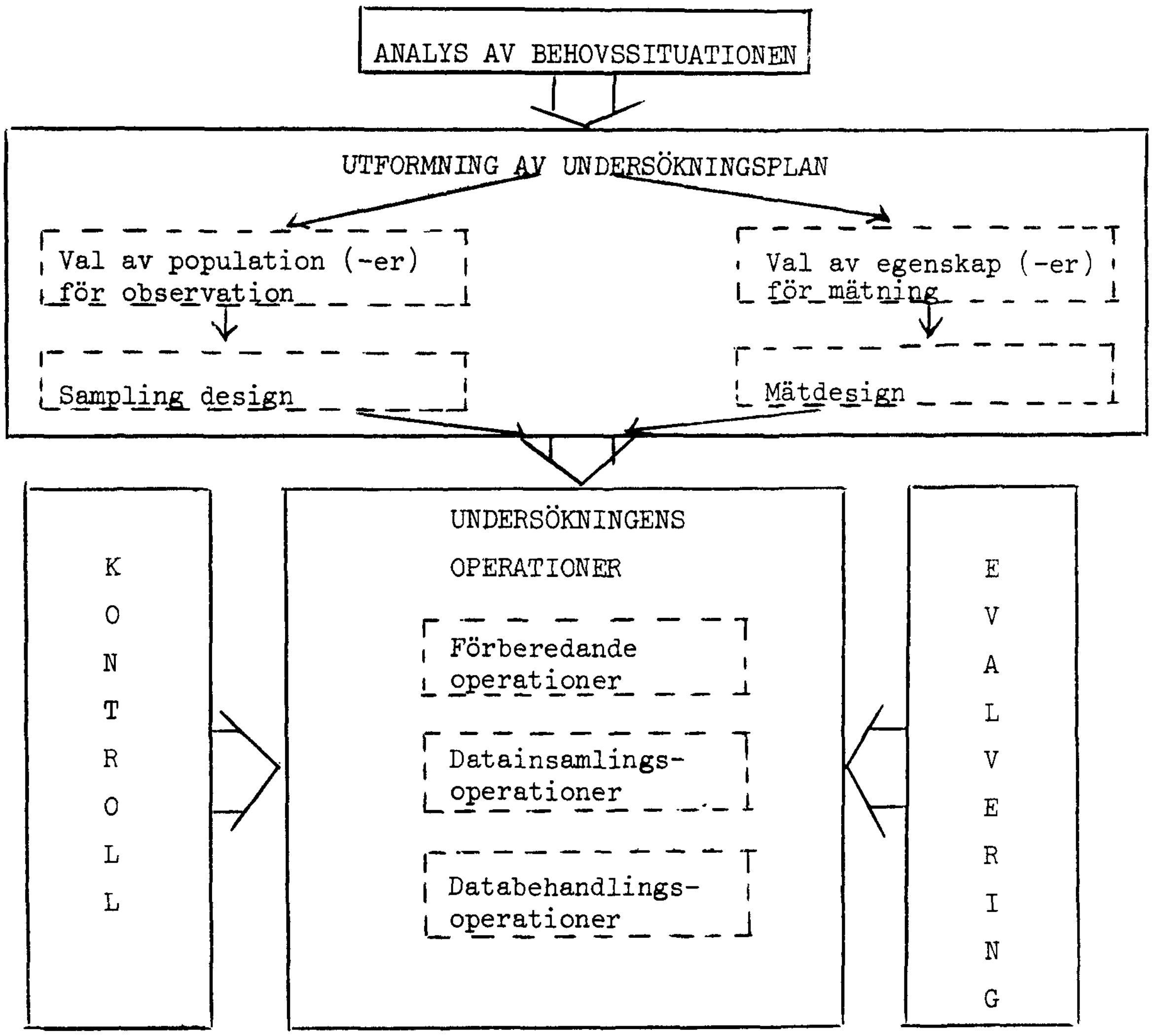
6 Något om sample survey-området

6.1 Begreppet sample survey

Det finns en mängd olika definitioner av begreppen sample survey, sampling survey och survey. Det kan dock konstateras att begreppen kan ges en vid innebörd. Det kan därför vara mer konstruktivt att redovisa de huvudfaser, som ingår i en dylik undersökning. Betrakta därför nedanstående figur, som är ett exempel på en sådan redovisning.

---

[1] Några sådana ansatser finns nämnda i Dalenius (1974).



6.2 Allmänt om planeringen av en sample survey

6.2.1 Kriteriefrågan

Ett nödvändigt element i planeringsarbetet är ett kriterium för avvägning mellan kostnad och egenskaper hos slutresultatet. Ett vanligt kriterium är att välja den plan, som ger den efterfrågade noggrannheten i undersökningsresultatet till lägsta möjliga kostnad. Den efterfrågade noggrannheten brukar specificeras av ett mått på medelkvadratavvikelsen. På senare tid har andra typer av kriterier kommit att diskuteras. I dessa laboreras vanligen med olika typer av admissibilitet men även olika slag av minimaxkriterier har förts fram som alternativ. I Isaki (1970) ges en redovisning av dessa nyare tankegångar.

6.2.2 Kartläggning och evalvering av undersökningssituationen

Det är nödvändigt i planeringsarbetet att noggrant studera och pröva undersökningssituationen. Därvid bör man betrakta de noggrannhets- och kostnadsvillkor, som uppställs av konsumenten. Vidare måste de mät- och samplingmetodologiska resurserna inventeras. Dessutom bör tillgängligt informationskapital i form av kunskaper om varians- och kostnadssituationen utnyttjas. Efter det att dessa faktorer kartlagts och prövats återstår en evalvering av undersökningssituationen. Denna evalvering skall resultera i ett beslut, som antingen innebär att undersökningen skall genomföras eller att den inhiberas.

6.2.3 Den tekniska planeringen

Givet att evalveringen resulterar i att undersökningen skall genomföras så återstår den tekniska planeringen. Denna bör göras på så sätt att ett antal alternativa planer konstrueras varefter en plan väljes i enlighet med planeringskriteriet. Utformningen av de alternativa planerna kan göras i tre steg. Dessa är

i) indelning av populationen i samplingenheter
ii) sampling av dessa enheter och
iii) estimation av parametrar.

6.3 Några grundläggande urvalsscheman

Här uppräknas några av de urvalsschema, som förekommer i nästa avsnitt.

Obundet slumpmässigt urval (OSU) innebär slumpval av n element från en population om N element. Karaktäristiskt är att varje kombination av n element har samma sannolikhet att väljas.

Clusterurval innebär urval av m clusters från en population om M clusters. Varje cluster består av ett eller flera element. Alla element i de uttagna m clustren utgör undersökningens sample av element.

Stratifierat urval innebär att populationen av element eller cluster indelas i delpopulationer. Dessa delpopulationer kallas strata och urval görs från varje stratum.

Variationer och hybrider av dessa schema kan göras. Några exempel är flerstegsurval, dubbelsampling och systematiskt urval.

7 Bayesiansk samplingteori

7.1 Referenser

Under det senaste decenniet har en speciell bayesiansk samplingteori utvecklats. Denna teori kännetecknas av att den är Bayesiansk "från början till slut". Detta medför vissa konsekvenser, som berörs nedan. Denna nyare samplingteori finns utvecklad i bl a Zacks (1969), (1970) och Ericson (1969a), (1969b).

7.2 Bayes samplingschema enligt W A Ericson

I detta avsnitt redovisas relativt kort innebörden i det Bayesianska samplingschema, som utvecklats av W A Ericson. Framställningen är hämtad ur Ericson (1969b).

Vi definierar en ändlig population om N element genom indexmängden $\underset{\sim}{N} = \{1, 2, \ldots, n\}$ och mängden $\underset{\sim}{X} = \{X_1, X_2, \ldots, X_N\}$, där X_i är värdet på någon karakteristika hos det i:te elementet i populationen. Inferens kan göras till någon enkel funktion $g(\underset{\sim}{X})$ av $\underset{\sim}{X}$ såsom

$$\sum_1^N X_i \text{ eller } \mu = \frac{\sum_1^N X_i}{N} .$$

Ett sample, s^*, om storleken m element definieras som en ordnad sekvens $i_1^*, i_2^*, \ldots, i_m^*$ av populationens element tillsammans med en sekvens av motsvarande observerade värden på elementens karakteristikor $\underset{\sim}{x}^* = \{x_{i_1^*}, x_{i_2^*}, \ldots\ldots\ldots\ldots, x_{i_m^*}\}$. Ett sample design definieras nu som en uppräknelig mängd S^* av ordnade sekvenser, s^*, tillsammans med ett sannolikhetsmått, som tilldelas genom val av en funktion $p(s^*) \geq 0$, $\sum_{s^* \in S^*} p(s^*) = 1$, där $p(s^*)$ är sannolikheten att välja samplet s^*.

För varje sådant sample (s^*, x^*) definieras statistikan $(s; x_i, i \in s)$ som mängden av distinkta populationselement $s = \{i_1, \ldots\ldots, i_n\} \subseteq \underset{\sim}{N}$ ingående i den observerade sekvensen s^* tillsammans med de observerade värdena x_j på X_j, $j \in s$. Givet ett godtyckligt sample s^* som innehåller de n distinkta enheterna $s = \{i_1, \ldots\ldots, i_n\}$ definieras matrisoperatorn S sådan att $S(\underset{\sim}{X}) = (X_{i_1}, \ldots\ldots, X_{i_n})$, där $i_1 < i_2 < \ldots\ldots < i_n$, komplementet $\bar{S}$ sådant att $\bar{S}(\underset{\sim}{X}) = (X_{j_1}, X_{j_2}, \ldots\ldots, X_{j_{N-n}})$ för alla $j_i \in \underset{\sim}{N} - s$ $(j_1 < j_2 < \ldots\ldots < j_{N-n})$ samt vektorn $\underset{\sim}{x} = (x_{i_1}, \ldots\ldots, x_{i_n})$ för observerade värden på $S(\underset{\sim}{X})$. Om vi à priori antar en simultanfördelning $p'(X_1, X_2, \ldots X_N)$ för $\underset{\sim}{X}$, så är à posteriorifördelningen för $\underset{\sim}{X}$ givet samplet $(s^*, \underset{\sim}{x}^*)$ precis densamma som den givet statistikan $(s, \underset{\sim}{x})$. Med bayesianska ögon blir då (s, x) en s k "sufficient statistic".

Givet $(s^*, \underset{\sim}{x}^*)$ ges likelihood-funktionen för $\underset{\sim}{X}$ av

$$\ell(\underset{\sim}{X}; (s^*, \underset{\sim}{x}^*)) = \ell(\underset{\sim}{X}; (s, x)) =$$

$$= \begin{cases} kp(s^*) \text{ för } \underset{\sim}{X} | S(\underset{\sim}{X}) = \underset{\sim}{x} \\ 0 \text{ eljest} \end{cases}$$

där $k > 0$ är en godtycklig konstant.

Vidare gäller att givet en godtycklig simultan N-dimensionell à priorifördelning för $\underset{\sim}{X}$, $p'(X_1, \ldots\ldots, X_N)$, så ges à posteriorifördelningen för $\underset{\sim}{X}$ givet ett sample av ovan beskrivna typ av

$$p(\underset{\sim}{X} | (s, \underset{\sim}{x})) = \begin{cases} p'(\underset{\sim}{X})/p'_{S(\underset{\sim}{X})}(\underset{\sim}{x}) \text{ för } \underset{\sim}{X} | S(\underset{\sim}{X}) = \underset{\sim}{x} \\ 0 \text{ eljest} \end{cases}$$

där $p_{S(\underset{\sim}{X})}(\underset{\sim}{x}) \neq 0$ är den marginella à prioritätheten för $S(\underset{\sim}{X})$.

En fördel med detta schema är att à posteriorifördelningen inte beror på sampledesignet S^* och funktionen $p(s^*)$, varför dessa är irrelevanta vid inferensen till $\underset{\sim}{X}$. Även om designet inte spelar någon roll vid inferensen, så har det sin avgjorda betydelse à priori. Det gäller ju att sträva efter någon form av optimal sampling. Man kan exempelvis välja det design, som ger minsta förväntade à posteriorivarians för given kostnad. Ericson betraktar speciellt klassen av utbytbara à priorifördelningar. Sådana à priorifördelningar är symmetriska med hänsyn till möjliga permutationer. Detta innebär att variablerna $X_1, X_2, \ldots, X_N$ sägs vara utbytbara om var och en av de $N!$ permutationerna har samma simultana sannolikhetsfördelning. Om alla X_i är utbytbara i enlighet med ens à prioikunskaper, så innebär tillordnandet av en sådan à priorifördelning

$$p'(\tilde{X}) = \int_\theta \prod_{i=1}^{N} P(X_i|\theta)\, d\, F(\theta)$$

ett

i) val av en parametrisk fördelningsfamilj $p(X_i|\theta)$

och

ii) val av en à priorifördelning $F(\theta)$ för den okända parametern θ.

Dessa specifikationer är de som krävs för en Bayesiansk ansats i de flesta parametriska inferensproblem.

Genereringen av en simultan à priorifördelning på detta sätt är ekvivalent med att betrakta den finita populationen som ett sample från en oändlig superpopulation med den okända parametern θ. Parametern θ åsätts själv en subjektiv sannolikhetsfördelning.

7.3 En speciell konsekvens

Det har i ovan visats att randomiseringsmomentet kan sakna betydelse i ett bayesianskt analyssammanhang i den meningen att à posteriorifördelningen betingat av den sufficienta sample-statistikan (s,x) är oberoende av hur urvalet har dragits. Om man förutsätter utbytbara à priorifördelningar, så är alla urval om n likvärdiga. Man kan då i själva verket vara relativt indifferent gentemot randomisering och eventuella randomiseringskostnader kan vara en avgörande faktor, som talar för icke-randomisering. Om observationskostnaderna är olika för olika element, så kvarstår dock frågan om ekonomisk urvalssammansättning.

Randomisering garanterar givetvis en viss "opartiskhet" hos den som planerar och genomför en undersökning. Det är också möjligt att då i förväg specificera analysprocedurer som har kända egenskaper såsom "unbiasedness".

Vissa bayesianer är dock dogmatiska på denna punkt. I Solomon och Zacks (1970) definieras exempelvis ett ofullständigt bayesdesign såsom ett innehållande ett moment av randomisering. Randomiseringsfrågan diskuteras även i Zacks (1969) och Royall (1968).

8 Bayesansatser i sample surveys

8.1 Allmänt

I detta avsnitt redovisas något om hur bayesianska tankegångar skulle kunna tillämpas vid planeringen och genomförandet av sample surveys. Därvid tas ett antal surveyoperationer fram och bayesdoktrinens potentialitet avseende dessa operationer studeras. Den icke-bayesianske surveyplaneraren använder à prioriinformation om kostnads- och varianssituationen som ett informellt planeringsinstrument. Det tycks alltså inte vara något märkligt i sig att utnyttja mer eller mindre diffus à prioriinformation om en parameter. Se exempelvis Dalenius (1974). Skillnaden ligger i om man kan acceptera tanken att parametern kan ha en sannolikhetsfördelning. Det bör poängteras att de bayesansatser som kan bli aktuella mycket väl kan utgöra begränsade bayesansatser. Den ovan redovisade bayesianska samplingteorin är Bayes "helt igenom" och medför vissa konsekvenser som man kanske inte är beredd att acceptera. I en begränsad bayesansats kan tyngdpunkten ligga i att finna lämpliga representationer för ens à priorikunskaper.

8.2 Val av stickprovsstorlek

Val av stickprovsstorlek i en undersökning bestäms vanligen genom att ett precisionskrav av typ $r = \sqrt{1-n/N}\ \mathrm{Var}/\sqrt{n}$ föreligger eller att det finns en budgetrestriktion. I det senare fallet är det en vanlig situation att man samplar så långt budgeten medger medan man i det första fallet kan beräkna erforderlig stickprovsstorlek n om variansen är känd. Variansen är naturligtvis aldrig känd i en surveysituation men den kan ofta uppskattas à priori. Uppskattningen görs med hjälp av tidigare erfarenheter från liknande undersökningar eller med hjälp av speciella pilotundersökningar. I Stein (1945) redovisas en procedur för bestämning av n där tvåfasurval används. Med hjälp av ett första urval om n_1 skattas variansen till s_1. Erforderligt n blir s_1^2/r^2. Om $n \leq n_1$ används n_1. Om $n > n_1$ tas ett nytt urval om $n_2 = n-n_1$ med samma procedur.

I Palit och Guttman (1973) redovisas en bayesiansk metod att bestämma erforderlig stickprovsstorlek i samma situation. Tvåfasurvalet används här för att undvika problemet att estimera variansen. Man vill här skatta medelvärdet $\bar{X}$ i en ändlig population där $\bar{X}$ kan betraktas som en summa av tre komponenter

$$\bar{X} = \frac{1}{N}\left[n_1\,\bar{x}_1 + n_2\,\bar{x}_2 + (N-n)\,\bar{y}\right]$$

där $\bar{x}_1$ och $\bar{x}_2$ är medelvärdena i de båda faserna och där n_1 och n_2 är motsvarande samplestorlekar. $\bar{y}$ är medelvärdet för återstoden av populationselementen, N är antalet element i populationen och $n = n_1 + n_2$.

Som estimator väljer författarna

$$E(\bar{X}|\underline{x}_1,\ \underline{x}_2) = \frac{1}{N}\left[n_1\,\bar{x}_1 + n_2\,\bar{x}_2 + (N-n)\,E(\bar{y}|\underline{x}_1,\ \underline{x}_2)\right]$$

där $\underline{x}$ symboliserar urvalen. Estimatorn är det förväntade à posteriorivärdet för $\bar{X}$, där

$$E(\bar{y}|\underline{x}_1,\ x_2) = \frac{1}{N-n}\sum_{i=1}^{N-n} E(y_i|x_1,\ x_2) \quad \text{med}$$

$$E(y_i|\underline{x}_1,\ \underline{x}_2) = \int\ldots\int y_i\,p(\underline{y}|\underline{x}_1,\ x_2)\,dy_1\ldots dy_{N-n}$$

där $\underline{y}$ är vektorn för de okända värdena, dvs $\underline{y} = (y_1,\ldots y_{N-n})$.

Variansen är, säg,

$$V(\bar{X}|\underline{x}_1, \underline{x}_2) = (\frac{N-n}{N})^2 \quad V(\bar{y}|\underline{x}_1, \underline{x}_2) = \xi.$$

Problemet är nu att välja n_2 så att ett totalt precisionskrav uppfylls.

Två tänkbara ansatser är

(1) Vi väljer n_2 så att $E(\xi|\underline{x}_1, n_2)$ ger den önskade precisionen och förbehåller oss rätten att välja ett tredje fasens urval ifall det visar sig nödvändigt.

(2) Använd fördelningen för $\underline{y} = (y_1,\ldots\ldots, y_{N-n})$ givet $\underline{x}_1$ och välj n_2 så att sannolikheten att nå eller överträffa önskad precision ligger på en specifik nivå.

Om vi antar superpopulationstanken att den ändliga populationen $X_1,\ldots\ldots, X_N$ är ett slumpmässigt urval från en normal population med okänt medelvärde och varians och att vi ansätter icke-informativa à priorifördelningar för μ och σ^2 så blir à posteriorifördelningen för μ och σ^2

$$p(\mu,\sigma^2|\underline{x}_1, \underline{x}_2) \propto (\sigma^2)^{-(\frac{n}{2}+1)} \quad \exp\left\{-\frac{1}{2\sigma^2}\left[Q + n(\mu-m_x)^2\right]\right\}$$

där

$$m_x = \frac{n_1\,\bar{x}_1 + n_2\,\bar{x}_2}{n_1 + n_2}$$

och

$$Q = S_1^2\,(\underline{x}_1) + S_2^2\,(\underline{x}_2) + \frac{n_1 n_2}{n_1+n_2}\,(\bar{x}_2 - \bar{x}_1)^2$$

med

$$S_i^2\,(x_i) = \sum_{j=1}^{n_i} (x_{ji} - \bar{x}_i)^2 \quad , \quad i = 1,2$$

Det går att visa att

$$E(\bar{y}|\underline{x}_1, \underline{x}_2) = m_x \quad \text{och}$$

$$Var(\bar{y}|\underline{x}_1, \underline{x}_2) = \left(\frac{N}{N-n_1-n_2}\right) \frac{Q/(n_1+n_2-3)}{n_1 + n_2}$$

och att

$$E(\bar{X}|\underline{x}_1, \underline{x}_2) = m_x \quad \text{och}$$

$$Var(\bar{X}|\underline{x}_1, \underline{x}_2) = \left(\frac{N-n}{N}\right) \frac{Q/(n-3)}{n} = \xi$$

Om vi ska kunna använda ansatserna (1) och (2) ovan så kräver detta att vi känner fördelningen för ξ, såsom den just definierats.

I artikeln visas att

$$E[\xi|\underline{x}_1] = \left(\frac{N-n}{N}\right) \frac{S_1^2(\underline{x}_1)/n_1-3}{n_1 + n_2}$$

Efter många och långa beräkningar får vi att

$$E[\xi|\underline{x}_1] = \left(\frac{N-n}{N}\right) \frac{S_1^2(\underline{x}_1)/(n_1-3)}{n_1 + n_2}$$

Om alternativ (1) ovan används så blir

$$n_2 = \frac{N\, S_1^2(\underline{x}_1)/(n_1-3)}{N\, V^2 + S_1^2(\underline{x}_1)/(n_1-3)} - n_1$$

där V^2 är precisionskravet i form av en varians.

Detta resultat är numeriskt ekvivalent med ett "klassiskt" resultat i Cox (1952), som bygger på Steins' idé.

Om alternativ (2) används så väljs n_2 så att

$$p(\xi < V^2|\underline{x}_1, n_2) = 1 - \alpha$$

där α är en acceptabel risk att V^2 överskrids. Efter en hel del algebra kan man visa att n_2 måste satisfiera

$$\frac{NV^2\ (n_1+n_2)(n_1+n_2-3)}{(N-n_1-n_2)\ s_1^2\ (x_1)} - 1 = \frac{n_2}{n_1-1}\ F_\alpha$$

där F_α är punkten som överskrids med sannolikheten α då vi använder F-fördelningen med n_2, n_1-1 frihetsgrader.

I artikeln ges tabeller och diagram över erforderliga n_2-värden.

I Rao och Ghangurde (1972) används en fördelningsfri ansats som leder till en förväntad à posteriorivarians för medelvärdet efter första fasens sampling som överensstämmer med formeln ovan

$$E(\xi|\underline{x}_1) = \frac{N-n}{N}\ \frac{s_1^2\ (\underline{x}_1)/(n_1-3)}{n_1 + n_2}$$

med det undantaget att n_1-3 ersätts av n_1+1.

Detta resultat, oavsett om det är n_1-3 eller n_1+1 som gäller, leder till en erforderlig stickprovsstorlek n_2 som approximativt överensstämmer med det icke-bayesianska resultat som gavs av Cox givet normalfördelningsantagandet. Rao och Ghangurde visar att detta inte är fallet då vi har att skatta en proportion P. Om vi antar en diffus à priorifördelning h(P) enligt Lindley där $h(P) \propto \left[P(1-P)\right]^{-1}$ så blir

$$n_2 = \frac{q_1}{ap_1}\ \left(\frac{n_1}{n_1+1}\right) - n_1$$

där $p_1 = r_1/n_1$ är sampleproportionen i första samplet och $\sqrt{a}$ den önskade variationskoefficienten. Om man antar Jeffries invarianta à priorifördelning $h(P) \propto \left[P(1-P)\right]^{-1/2}$ så **erhålls**

$$n_2 = \frac{\tilde{q}_1}{a\tilde{p}_1}\left[\frac{n_1+1}{n_1+2}\right] - n_1 - 1$$

där $\tilde{p}_1 = (r_1 + \frac{1}{2})/(n_1 + 1)$

Urvalsstorleken enligt Cox' icke-bayesianska ansats blir

$$n_2 = \frac{q_1}{ap_1} + \frac{3}{p_1q_1} + \frac{1}{an_1p_1} - n_1$$

såvida inte P är alltför liten.

Ett enkelt numeriskt exempel där n_1=30, p_1=0.6 och $\sqrt{a}$ =0.1 ger

$n_{2(\text{Lindley})} = 35$

$n_{2(\text{Jeffrey})} = 35$

$n_{2(\text{Cox})} = 55$

vilket alltså visar att de angivna bayesianska ansatserna i detta exempel kräver mindre urvalsstorlekar än den icke-bayesianska föreslagen av Cox vid estimation av proportioner.

8.3 Val av estimator

Valet av estimator kan göras "situationsberoende" i många surveysituationer. Innebörden av detta uttryck är att exempelvis enkla medeltalsestimatorer ersätts med ett alternativ som speglar den gällande estimationssituationen. Ett exempel på en icke-bayesiansk sådan ansats ges i Christianson (1974). Där ges exempel på ett problem som uppstår vid användande av s k multiple choice-frågor vid kunskapsmätning. Problemet är att respondenter kan gissa rätt svarsalternativ vilket ger en bias i den vanliga medeltalsestimatorn vid obundet slumpmässigt urval, p = y/n,

där y är antalet rätta svar och n stickprovsstorleken.

Istället föreslås den "situationsberoende" estimatorn

$p' = (Rp-1)/(R-1)$

där R är antalet svarsalternativ.

I Press och Yang (1974) ges en bayesiansk ansats för lösandet av ett annat problem där en konventionell estimator skulle ge en bias. Det gäller här det ytterst vanliga surveyproblemet att estimera kategoritillhörighet då ett antal respondenter svarar "vet ej", "ingen åsikt", "har ej bestämt mig" etc i syfte att inte röja sin kategoritillhörighet.

Ibland fördelar man dessa respondenter i de olika klasserna i enlighet med något enkelt viktningsschema. Hur man än väljer vikterna torde det leda till ett biased förfarande. Om man istället öppet redovisar respondenterna i en "vet ej"-kategori kan man fråga sig om man mäter det man vill mäta dvs om estimatorn för kategoritillhörighet är relevant.

I Press och Yang antas att det existerar en huvudsaklig fråga i undersökningen men att det dessutom existerar ett antal med den huvudsakliga frågan relaterade hjälpfrågor. Problemet är att estimera kategoritillhörighet för den variabel som den huvudsakliga frågan avser. Detta kan göras på bayesianskt sätt genom att utnyttja följande tre informationskällor.

(1) Subjektiv à prioriinformation

(2) Samplefrekvenserna för de respondenter som svarat "korrekt" på den huvudsakliga frågan.

(3) Den information som finns i svaren på hjälpfrågorna.

Antag att n_i respondenter säger sig tillhöra kategori i beträffande den huvudsakliga frågan där i=1,....., M. Antag vidare att de återstående "obestämda" respondenterna uppgår till ett antal av m. Låt q_i vara sannolikheten att en slumpmässigt vald respondent hamnar i kategori i. En bayesestimator för denna sannolikhet är medelvärdet för en à posteriorifördelning

$$E(q_i | n_i, \ldots, n_{M-1}; z^{(1)}, \ldots, z^{(m)})$$

Denna skattning kan skrivas som

$$\hat{q}_i = \frac{(n_i + \alpha_i) + \sum_{j=1}^{m} P\{\Pi_i | z^{(j)}\}}{m + \sum_{j=1}^{M} (n_j + \alpha_j)}$$

$i = 1, \ldots, M$

där α_i:na är parametrarna i à priorifördelningen för q_i:na ($\alpha_i = 1$ om man använder en vag à priorifördelning),

$P\{\Pi_i|z^{(j)}\}$ är den marginella prediktionssannolikheten att klassificera den j:te "obestämde" respondenten till kategori Π_i för huvudfrågan, givet dennes svar på hjälpfrågorna $z^{(j)}$.

Man kan visa att $\hat{q}_i$ kan uttryckas på följande sätt.

$$\hat{q}_i = \frac{n_i + \alpha_i}{m + \Sigma_{j=1}^{M} (n_j + \alpha_j)} \left[\Sigma_{j=1}^{m} \left[\frac{h(z^{(j)}|\Pi_i)}{\Sigma_{t=1}^{M} (n_t + \alpha_t)\, h(z^{(j)}|\Pi_t)} \right] \right]$$

där α_i är parametrarna i den à priorifördelningen (Dirichlet) som antas gälla alla q_j. $h(z^{(j)}|\Pi_i)$ betecknar den predikterade marginaltätheten för svaret på hjälpfrågorna $z^{(j)}$ för den j:te "obestämde" respondenten, om han hade svarat att han tillhörde kategori i beträffande den huvudsakliga frågan; dvs under antagandet att han tillhörde kategorin Π_i.

Hjälpfrågorna väljs efter grad av diskriminans och antalet efter hur liten den prediktiva felklassificeringssannolikheten bör vara. Antag att hjälpfrågorna alla avser diskreta variabler och att det finns S celler i en korstabell över svaren till mängden av hjälpfrågor. Då är fördelningen för sannolikheten att den j:te "obestämde" respondenten hamnar i cell k i korstabellen samtidigt som han tillhör kategori i beträffande den huvudsakliga frågan.

$$h(z^{(j)}=u_K|\Pi_i) = \frac{x_{k|i} + \Delta_{k|i}}{n_i + \Sigma_{t=1}^{S} \Delta_{t|i}}$$

där k=1, ..., S, i=1, ..., M, där u_K betecknar en Sx1 vektor med en "etta" i cell k och nollor för övrigt och där $x_{k|i}$ betecknar antalet respondenter som svarade "korrekt" att de tillhörde kategori i beträffande den huvudsakliga frågan och som befann sig i cell k beträffande hjälpfrågorna. Vidare betecknar $\Delta_{k|i}$ parametrarna i à priorifördelningen för cellsannolikheterna för svaren på hjälpfrågorna.

Det är också nödvändigt att definiera den betingade sannolikheten $p_{k|i}$ att en respondents svar på de aktuella hjälpfrågorna kommer att hamna i kategori k givet att respondenten tillhör kategori i beträffande huvudfrågan. Man kan exempelvis ansätta en Dirichlet à priorifördelning för $p_i = (p_{1|i},\ldots\ldots, p_{S-1|i})$.

I artikeln ges ett numeriskt exempel på denna estimation. Antag att 100 respondenter svarar på den huvudsakliga frågan där det existerar tre alternativ förutom det obestämda. Undersökningen ger vid handen att 20 tillhör kategori 1, 20 kategori 2, 20 kategori 3 medan 40 är obestämda. Antag att vi bara har en hjälpfråga med fyra svarsalternativ.

Tabellen visar hur de 60 "bestämda" respondenterna svarat på hjälpfrågorna.

Huvud-fråga	Hjälpfråga Alt 1	Alt 2	Alt 3	Alt 4	Totalt
Alt 1	17	1	1	1	20
Alt 2	5	5	5	5	20
Alt 3	1	1	1	17	20

De 40 "obestämda" fördelar sig beträffande hjälpfrågan enligt

Alt 1	Alt 2	Alt 3	Alt 4
25	5	5	5

Om man antar vaga à priorifördelningar för q_i och p_i så blir $\hat{q}_1 = 0.40$, $\hat{q}_2 = 0.34$ och $\hat{q}_3 = 0.26$. Om man inte brytt sig om de "obestämda" (vilket alltså ofta är fallet i undersökningar hade proportionerna skattats till $\hat{q}_1 = \hat{q}_2 = \hat{q}_3 = 20/60 = 1/3$.

Låt mig avslutningsvis också nämna följande. Man väljer ibland kvot- eller regressionsestimatorer som ett alternativ till unbiased estimatorer. Man gör det i situationer där hjälpvariabelns medelvärde och sambandet mellan den skattade variabeln och hjälpvariabeln är kända och utnyttjar den kunskapen vid skattningen. I Tomlin (1977) redovisas en semiklassisk, semi-bayesiansk ansats vid tillämpningen av dessa estimatorer. Där utnyttjas även information om populationens utseende eftersom detta påverkar estimatorernas egenskaper. God sådan information kan leda till stora precisionsvinster jämfört med vanliga kvot- och regressionsestimatorer.

8.4 Stratifiering

En Bayesansats vid stratifierat urval aktualiserar åtminstone två problem. Det gäller att välja eller konstruera en à priorifördelning och det gäller att allokera urvalet optimalt. Om à priorikunskaperna om en population kan speglas med utbytbara à priorifördelningar, så kan man tänka sig att dela populationen i delpopulationer sådana att man inom varje delpopulation betraktar variabelvärdena X_i som utbytbara. Delningen kan då betraktas som en stratifiering. Antag att X_{ji} betecknar värdet på enheten i i stratum j. Antag vidare att det finns N_j enheter i det j:te stratumet och att $\sum_1^k N_j = N$ där N är antalet element i populationen.

Låt $\underset{\sim}{X}_j = (X_{j1},\ldots\ldots, X_{jN_j})$ $j = 1,\ldots, k$ beteckna variabelvärdena för objekten i stratum j och $\underset{\sim}{X} = (\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_k)$ beteckna variabelvärdena för enheterna i hela populationen. En klass av à priorifördelningar för X ges av den allmänna täthetsfunktionen

$$p'(X) = \prod_{j=1}^{k} p'_j\,(\underset{\sim}{X}_j)$$

där $p'_j\,(\underset{\sim}{X}_j)$ speglar utbytbarhet beträffande alla X_{ji} där $i = 1,\ldots, N_j$. Alla $\underset{\sim}{X}_j$ kan nu betraktas som à priori oberoende. Om stratumstorlekarna N_j är stora kan fördelningarna för X_{ji}:na många gånger illustreras med approximativa normalfördelningar.

I Ericson (1965) och (1969b) redovisas lösningen på allokeringsproblemet under olika förutsättningar.

En annan situation kan vara den, där det förutsätts att det existerar sådan à prioriinformation om de enskilda stratummedelvärdena μ_j, att den kan approximeras med normalfördelningar. Allokeringsproblemet blir att finna den vektor $n = (n_1, \ldots, n_k)$ som definierar det optimala stratifierade urvalet i den meningen att à posteriorivariansen v'' för μ minimeras för en given budget. Låt π_j vara den kända proportionen enheter, som finns i stratum j. Låt σ_j^2 beteckna kända inomstratumvarianser. Vidare gäller att samplemedelvärdesvektorn $x = (\bar{x}_1, \ldots, \bar{x}_k)$ har en k-variat normalfördelning definierad av en medelvärdesvektor och en varians-kovariansmatris, vars diagonalelement är $v_{jj} = \sigma_j^2/n_j$.

Dessutom förutsättes en kostnadsvektor $c = (c_1, \ldots, c_k)$ där c_j är observationskostnaden per element i stratum j. Vi söker nu mängden $\{n_j\}$ som minimerar

$$v'' = \sum_{j=1}^{k} \frac{\pi_j^2}{(n'_{jj} + n_j/\sigma^2_j)}$$

med hänsyn till budgetvillkoret

$$\sum_{j=1}^{k} c_j n_j = C$$

där $n'_{jj} = 1/v'_{jj}$ och där v'_{jj} är elementen i à priorifördelningens varians-kovariansmatris. Lösningen på allokeringsproblemet ligger i att sampla i ett urval av alla strata. Låt strata numreras så att $B_0 > B_1 \geq B_2 > \ldots \geq B_k$, där $B_j = \sigma_j \sqrt{c_j} | \pi_j \, v'_{jj}$ och $B_0 = \infty$.

Intervallet $C > 0$ kan nu uppdelas på följande sätt. För $r = 0, 1, \ldots, k-1$ låt

$$I_r = \{C | C_r < C \leq C_{r-1}\}$$

där $C_{-1} = \infty$ och

där $C_r = B_{r+1} \sum_{j=r+1}^{k} \pi_j \, \sigma_j \sqrt{c_j} - \sum_{j=r+1}^{k} c_j \, \sigma_j^2 \, n'_{jj}$

Den optimala urvalsstorleken, n_j^o, givet $C \epsilon I_r$, ges av

$$\begin{cases} n_j^0 = 0 \quad \text{för } j \leq r \\ n_j^0 = \dfrac{\pi_j \sigma_j}{\sqrt{c_j}} \left[\dfrac{C + \sum\limits_{j=r+1}^{k} \dfrac{c_j \sigma_j^2}{v'_{jj}}}{\sum\limits_{j=r+1}^{k} \pi_j \sigma_j \sqrt{c_j}} \right] - \dfrac{\sigma_j^2}{v'_{jj}} \quad \text{för } j > r \end{cases}$$

Strata har alltså ordnats och gränsen r är en funktion av informationsmängd och samplingkostnader för olika strata.

Vid en budgetökning börjar man sampla i stratum (r-1) därefter i (r-2), etc, tills man för tillräckligt stor budget är kapabel att sampla från alla strata. Detta förfaringssätt strider mot klassiskt stratifierat urval. Vi har dock en parallell som används inom icke-bayesiansk sampling, nämligen s k cut-off-sampling. Principen är där att populationen kan indelas i två strata varefter endast ett stratum undersöks eftersom man där kan finna nästan all information om den sökta parametern. Förfarandet är dock inte där formaliserat på samma sätt som i den bayesianska ansatsen.

Om à prioriinformationen ovan är diffus i den meningen att à priorivarianserna v'_{jj} alla är mycket stora, så erhålls den välkända "klassiska" allokeringsformeln

$$n_j^0 = \frac{\pi_j \sigma_j}{\sqrt{c_j}} \frac{C}{\sum\limits_{j=1}^{k} \pi_j \sqrt{c_j}\, \sigma_j}, \quad j=1, \ldots, k$$

eftersom $v'_{jj} \to \infty$.

I Palit och Guttman (1973) betraktas en superpopulationsmodell för stratifierat urval. De N_i stratumelementen i stratum i betraktas här som ett urval om N_i oberoende observationer från en oändlig superpopulation där fördelningen för den intressanta variabeln X_{ji} är $f_j(x|\underline{\theta}_j)$.

Om alla θ_j à priori betraktas som oberoende blir à posteriorimedelvärdet för ett populationsmedelvärde $\bar{X}$

$$E(\bar{X}|\underline{x}_1,\ldots\ldots, \underline{x}_k) = \frac{1}{N} \sum_{j=1}^{k} N_j\ E(\bar{X}_j|\underline{x}_j)$$

där $\underline{x}_j$ är samplet från stratum j och där $E(\bar{X}_j|\underline{x}_j)$ är bayes-estimatorn för stratum j:s medelvärde.

Variansen blir

$$Var(\bar{X}|\underline{x}_1,\ldots, \underline{x}_k) = \sum_{j=1}^{N} (\frac{N_j}{N})^2\ Var(\bar{X}_j|\underline{x}_j)$$

I Zacks (1970) redovisas Bayesestimatorer för en- och tvåfas-stratifierat urval. I fallet med tvåfasurvalet är situationen följande.

Ett problem är att skatta en proportion θ i en population uppdelad i k strata. Populationsstorleken är $N = \{N_1,\ldots\ldots, N_k\}$. Ett annat problem är att allokera budget och urvalsstorlekar mellan faserna och bland strata. Låt $m = \{m_1,\ldots\ldots, m_k\}$ beteckna förstafasallokeringen och $n = \{n_1,\ldots\ldots, n_k\}$ andrafasallokeringen. Låt $X = (X_1,\ldots\ldots\ldots, X_k)$ vara observationsvektorn för första fasens oberoende variabler. Variablerna kan vara antalet defekta, dvs i första fasen observeras X_k defekta enheter i stratum k. Låt $Y = (Y_1, \ldots, Y_k)$ vara motsvarande vektor i den andra fasen. Antag vidare att $X_1, \ldots, X_k$ och $Y_1, \ldots, Y_k$ är respektive hypergeometriskt fördelade. Givet allokeringen (m, n) och observationerna (X, Y) så är bayesestimatorn för proportionen $\theta = \sum^{k} \lambda_j P_j$ lika med

$$\hat{\theta}(X,Y,m,n) = \sum_{j=1}^{k} \lambda_j \left[\frac{X_j+Y_j+1}{n_j+m_j+2} \quad (1+\frac{2}{N_j}) - \frac{1}{N_j} \right]$$

där λ_j är en specifik konstant och P_j är proportionen defekta i stratum j.

Man kan efter första fasen beräkna preposterioririsken $R(X,m,n) = E_{Y|X;\ m,n} \{R(X,Y;\ m,n)\}$ där $R(X,Y,;\ m,n)$ är à posterioririsken. (Det skulle föra för långt att här ytterligare diskutera de i designet involverade riskfunktionerna.) Den optimala allokeringen i den andra fasen är den som minimerar denna preposterioririsk. Det ställer sig besvärligare att allokera optimalt i den första fasen. Man kan dock tänka sig en pseudo-optimal allokering i enlighet med ren enfas-allokering.

Låt P_j ovan definieras som M_j/N_j (där M_j är faktiska antalet defekta i stratum j). I Solomon och Zacks (1970) visas att om alla M_j är à priori oberoende binomialfördelade så är det optimala designet sådant att det utgörs av ett enfasurval. Om däremot alla M_j är à priori oberoende diskret likformigt fördelade, så är det bättre med tvåfasurval. Man kan visa att tvåfasurval mestadels knappast är värt mödan då M_j är beta- eller binomialfördelade. I Aggarwal (1959) visas att vid OSU från en finit population är samplemedelvärdet en minimaxestimator för populationsmedelvärdet. Det är också möjligt att ansätta en minimaxallokeringsvektor för stratifierat urval. Denna allokering skall minimera risken

$$R(n) = \sum_{j=1}^{k} \left[\left(\frac{N_j}{N}\right)^2 \left(\frac{1}{n_j} - \frac{1}{N_j} \right) \sigma_j^2 + c_j n_j \right]$$

där σ_j^2 är variansen i stratum j och c_j samplingskostnaden per element i stratum j.

Minimaxallokeringen ges av

$$n_j^* = \text{heltalet närmast} \left[\frac{N_j^2 \sigma_j^2}{N^2 c_j} + \frac{1}{4} \right]^{1/2} \quad \text{och} \leq N_j$$

för $j = 1, \ldots, k$.

I Draper och Guttman (1968a) och (1968b) visas hur man allokerar sina observationer i den andra fasen i ett tvåfasurval och därvid utnyttjar informationen i den första. Före den första allokeringen antas à prioriinformationen representerad av oberoende, lokalt likformiga à priorifördelningar för μ_j (medelvärdet i stratum j) och σ_j^2. Den simultana à posteriorifördelningen för alla μ_j och σ_j^2 kan sedan användas som à prioriinformation före det andra steget. Vidare ges allokeringsmetoder vid tvåfasurvalets andra steg vid sampling av k variabler.

Draper och Guttmans ansats gäller skattning av medelvärdet i en oändlig population. Skattning med hjälp av tvåfasurval kan också tillämpas vid ändliga populationer vilket visas i Palit och Guttman (1972). Detta att man utnyttjar den information som erhålls i den första fasen leder till s k databaserade à priori-fördelningar.

I Ericson (1968) tas hänsyn till att såväl fasta som rörliga **kostnader är förknippade med sampling** från ett visst stratum. Den optimala allokeringen ges för ett antal olika fördelningsantaganden. En tolkning ges också till två- och flerstegsurval då det finns à prioriinformation rörande förstastegsenheterna. De resultat som Ericson härleder kan användas för att erhålla optimal allokering i en tvåstegsurvalsmodell.

I Aggarwal (1966) ges en clustersamplingtillämpning. Principiellt kan man se clusterurval som ett urval av h strata från k i en population. Därefter allokeras ett urval bland de h valda. Aggarwal använder sig av en förlustfunktion

$$L(\hat{\mu},\mu) = (\hat{\mu}-\mu)^2 + c_b h + \sum_{j=1}^{h} c_j n_j$$

där c_b är kostnaden att observera en primär urvalsenhet (PUE), och där $\hat{\mu}$ är estimatorn för populationsvärdet μ. Genom att betrakta klassen av alla à priorifördelningar för vilka den betingade variansen för stratummedelvärdena μ_j givet μ begränsas av σ_b^2 (som är en given konstant) och för vilka den betingade inomstratum à priorivariansen begränsas av σ_j^2, så erhålles minimaxestimatorn

$$\mu^* = \sum_{j=1}^{h} W_j \bar{X}_j \Big/ \sum_{j=1}^{h} W_j$$

där $W_j = n_j\,(\sigma_j^2 + n_j\,\sigma_b^2)^{-1}$ för $j = 1,\ldots, h$ och där $\bar{X}_j$ är observerat stratummedelvärde.

Motsvarande minimaxrisk är

$$R^*(h,n) = \left[\sum_{j=1}^{h} W_j\right]^{-1} + h\sigma_b^2 + \sum_{j=1}^{h} c_j n_j .$$

där n är allokeringsvektorn för urvalet bland de h valda strata.

Problemet med bayesiansk allokering vid stratifierat urval tas också upp i Khan (1976).

Vi har sett att förutsättningarna i de olika ansatserna varierar. Ericson, exempelvis, antar

i) normalfördelade inomstratumpopulationer
ii) kända stratumvarianser
iii) normalfördelade à priorifördelningar för μ_j

och en algoritm som konstruerar och minimerar à posteriori-variansen för medelvärdet i populationen under en budget-restriktion och att $n_j \geq 0$.

Draper och Guttman utvidgar tillämpningsområdet till att gälla okända stratumvarianser genom användande av en databaserad à priorifördelning som en kombination av diffus à priorifördelning och likelihooden för förstafasurvalet.

Rao och Ghangurde föreslår allokeringslösningar som är fördelningsfria och fria från restriktioner som "oändlig population och kända varianser".

I Scott och Smith (1971) redovisas en speciell problemsituation vid stratifierat urval. Ofta vill man estimera parametrar för delgrupper i en population, s k domains of study, men man känner inte i förväg vilka som tillhör delgruppen. Det går därför inte att bygga in dessa i designet som tillhörande ett eget stratum utan vid t ex stratifierat urval kommer delgruppsmedlemmarna att fördela sig på flera strata. I regel är antalet medlemmar, M_j, i stratum j tillhörande delgrupper okänt. Antag att det existerar viss à prioriinformation om delgruppsstorlekarna $M_1, M_2, \ldots\ldots, M_j, \ldots\ldots, M_L$ där L är antalet strata och där N_j är antalet element i stratum j. Denna à prioriinformation kan eventuellt representeras som en produkt av oberoende s k polyavariabler med täthetsfunktionerna

$$P(M_j|N_j,\gamma_j,p_j) = \binom{N_j}{M_j} \frac{\prod_{i=1}^{M_j-1} (p_j + i\gamma_j) \prod_{i=1}^{N_j-M_j-1} (1-p_j+i\gamma_j)}{\prod_{i=1}^{N_j-1} (1 + i\,\gamma_j)}$$

där $0 < p_j < 1$ och antingen $(N_j-1)\,\gamma_j \geq -\min(p_j, 1-p_j)$ eller $\gamma_j = -T_j^{-1}$ där T_j och $T_j p_j$ är heltal med $T_j \geq N_j$. Denna fördelning har medelvärde $N_j p_j$ och varians

$$N_j p_j\,(1-p_j)(1+N_j\gamma_j)/(1+\gamma_j).$$

Om ett slumpmässigt urval genererar delgruppsstorlekar på m_1, $m_2, \ldots, m_j, \ldots, m_L$ i urvalet så är dessa m_j oberoende hypergeometriskt fördelade

$$p(m_j|M_j) = \binom{M_j}{m_j}\binom{N_j-M_j}{n_j-m_j} \Big/ \binom{N_j}{n_j}$$

Det kan visas att à posteriorifördelningen för M_j-m_j också är av polyaform

$$P(M_j-m_j|N_j-n_j,\ \gamma_j^*,\ p_j^*)$$

där $\gamma_j^{*-1} = \gamma_j^{-1} + n_j$

och $p_j^*/\gamma_j^* = p_j/\gamma_j + m_j$

À posteriorimedelvärdet för M_j blir

$$E(M_j|m_j) = m_j + (N_j-n_j)\,p_j^*$$

och à posteriorivariansen

$$\mathrm{Var}(M_j|m_j) = (N_j-n_j)\,p_j^*\,(1-p_j^*)\,\frac{(1+N_j\gamma_j)}{(1+(n_j+1)\,\gamma_j)}$$

Om vi nu vill estimera t ex totalen Y för någon parameter i delgruppen så kombineras à posteriorivärdena för M_j med data om sampleutfallet av typ

$E(Y|M,\underline{y})$ där $\underline{y}$ är sampletotalen eventuellt vägd med à prioriinformation om y och där à posterioriinformationen om M sätts in.

I icke-bayesianska lösningar på detta problem förutsätts att inget är känt om delgruppselementens värden och att $\{M_j\}$ antingen är helt känd eller också inte alls. Detta motsvaras av γ_j-värden på $-N_j^{-1}$ eller $\gamma_j \to \infty$.

För γ_j-värden däremellan blir bayesestimatet för totalen ett vägt medelvärde av de icke-bayesianska lösningar som motsvarar de båda extremvärdena på γ_j vilket väl kan tyckas intuitivt tilltalande.

8.5 Bortfall

Bortfall innebär att ett eller flera samplade element av olika anledningar inte har kunnat mätas. Konventionell sampling hänvisar i sådana situationer till att exempelvis subsampla bland de element som ingår i bortfallet. En sådan ansats ges i Hansen och Hurwitz (1946). Där är syftet att skatta parametern i bortfallsdelen. Vanligt är också att man använder en rad preventiva och andra metoder vars syfte är att reducera effekterna och omfattningen av bortfallet. Bortfallet är i många undersökningar en mycket besvärande specifik felkälla. À priorikunskaper om bortfallets storlek och struktur kan dock många gånger hjälpa statistikern i planeringsarbetet. En bayesiansk ansats i en bortfallssituation redovisas i Ericson (1967). I denna inkorporeras formellt à priorikunskaper av olika slag.

Antag att problemet är att estimera ett populationsmedelvärde μ. Man kan därvid se populationen som uppdelad i två strata. Det ena består av element, som man kan mäta (medverkandestratum), och det andra av de element som utgör bortfallet (bortfallsstratum). I dessa strata antas respektive medelvärde vara μ_1 och μ_2. Stratumstorlekarna har proportionerna Π och $(1-\Pi)$ respektive. Med hjälp av tidigare undersökningar, provundersökningar etc, så ansätts en simultan à priorifördelning för (μ_1, μ_2, Π). Man väljer nu ett urval om n element. Av dessa kan man få mätvärden från r. Ett sådant resultat ger direkt information om både μ_1 och Π via mätvärdena $X_1, \ldots, X_r$ och kanske indirekt om μ_2, ifall man känner några samband mellan μ_1 och μ_2. Med hjälp av Bayes' teorem erhålles en simultan à posteriorifördelning för (μ_1, μ_2, Π). Denna fördelning kan användas som à priorifördelning, då det gäller att bestämma en i någon mening optimal urvalsstorlek om m vid urval bland de (n-r) element, som utgör bortfallet. Om detta urval kan genomföras utan nya bortfall, erhålles mätresultaten V_1, ..., V_m. Detta andra urval används för att erhålla en slutlig à posteriorifördelning för (μ_1, μ_2, Π) som i sin tur ger fördelningen för $\mu = \Pi\mu_1 + (1-\Pi)\mu_2$.

Låt nu à priorifördelningen för Π representeras av en betafördelning och à priorifördelningen för (μ_1, μ_2) som en av denna betafördelning oberoende bivariat normalfördelning. Den simultana à priorifördelningen för (μ_1, μ_2, Π) är då en "beta-normal" fördelning. Man kan vidare anta att antalet medverkande r är en binomialfördelad variabel med parametrarna n och Π. Beträffande dessa förutsättningar kan man säga att oberoendet mellan individer kan vara naturligt men att principvägrare inte passar in i modellen.

Under förutsättning att element nr i mätes, så kan det uppmätta värdet X_i ses som en observation från en normalfördelning med medelvärdet μ_1 och känd varians σ_1^2. På samma sätt kan V-värdena i andra steget ses som observationer på en normalfördelning med medelvärde μ_2 och känd varians σ_2^2. På så sätt kan en för de båda urvalen gemensam likelihoodfunktion bestämmas. Det bayesianska optimala estimatet av μ, säg $\hat{\mu}$, erhållas genom att minimera en förlustfunktion av typen

$$E_\mu^* \; L(n,r,\bar{X}_r,m,\bar{V}_m,\hat{\mu},\mu) = K\, E_\mu^* \, (\mu-\hat{\mu}) + c_1 n + c_2 r + c_3 m$$

där E_μ^* betecknar förväntan för μ avseende à posteriorifördelningen givet de båda urvalen och där c_i betecknar elementkostnaden i respektive urval och där K representerar en slags "trade-off" mellan estimationsfelet och kostnaderna. Det värde på $\hat{\mu}$ som minimerar denna förväntade förlust är $\hat{\mu}_0$, det slutliga à posteriorimedelvärdet för μ. Värdet kan skrivas

$$\hat{\mu}_0 = \frac{r'+r}{n'+n}\,\mu_1^* + \frac{n'+n-r'-r}{n'+n}\,\mu_2^*$$

där μ_1^* och μ_2^* ingår som komponenter i medelvärdesvektorn i den slutliga à posteriorifördelningen och där n' och r' är parametrarna i à posteriorifördelningen för Π.

Om ens à prioriuppfattning om (μ_1, μ_2, Π) är vag, dvs då à priorifördelningen för Π är likformig och à priorifördelningen för (μ_1, μ_2) har stor varians, så är den optimala estimatorn

$$\hat{\mu}_0 = \frac{r}{n} \bar{X}_r + \frac{n-r}{n} \bar{V}_m$$

där

$$\bar{X}_r = \frac{1}{r} \sum_{i=1}^{r} \bar{x}_i \quad \text{och} \quad \bar{V}_m = \frac{1}{m} \sum_{i=1}^{m} v_i$$

Den optimala urvalsstorleken, m_0, i det andra urvalet är den som minimerar den förväntade förlusten $L(n,r,\bar{X}_r,m)$. Den optimala urvalsstorleken, n_0, i det första urvalet är den som minimerar den förväntade förlusten $L(n,r,\bar{X}_r)$ med avseende på den simultana marginalfördelningen för r och $\bar{X}_r$. Resonemanget ovan förutsätter att inget bortfall sker bland de m_0 valda i det andra urvalet. I en strävan att nå ett 100%-igt medverkande (exempelvis genom att hämta mätvärden från externa informationskällor, öka trycket mot respondenten, tillämpa ekonomisk ersättning till de medverkande etc) kan man få en bias. Denna kan ta sig uttryck på så sätt att den betingade förväntan för $\bar{V}_m|\mu_2$ inte behöver vara μ_2 utan kanske μ_2 + B, där B är en biasterm. En sådan biasterm kan inkorporeras i planeringen genom att ansätta en fördelning för B baserade på tidigare erfarenheter. Om en biasterm inkorporeras reduceras den optimala urvalsstorleken från m_0 till m_0'. Detta speglar det faktum att biased observationer är mindre värda än unbiased.

I Rao och Ghangurde (1972) ges en liknande lösning.

I Kaufman och King (1973) ges en lösning i det binomiala fallet. Antag att vi vill estimera proportionen p_f med en viss egenskap i en population. Det existerar ett bortfall så att följande parametermängd uppstår. (Beteckningarna är de som använts i referensen).

	Svarsstratum	Bortfallsstratum	Totalt
Egenskap	p_{fr}	p_{fs}	p_f
Icke-egenskap	p_{ur}	p_{us}	p_u
Totalt	p_r	p_s	1

En urvalsstorlek på n fördelar sig analogt

n_{fr}	n_{fs}	n_f
n_{ur}	n_{us}	n_u
n_r	n_s	n

Uppenbarligen kan man, förutom det givna n, endast observera n_{fr}, n_{ur} och n_s.

För att lösa problemet att estimera p_f införs en alternativ parametrisering, nämligen

$p_{r \cdot f}$ = den betingade sannolikheten för "icke-bortfall" givet "egenskap"

$p_{r \cdot u}$ = den betingade sannolikheten för "icke-bortfall" givet "icke-egenskap"

$p_{s \cdot f} = 1 - p_{r \cdot f}$

$p_{s \cdot u} = 1 - p_{r \cdot u}$

$p_{f \cdot r}$ = den betingade sannolikheten för "egenskap" givet "icke-bortfall"

$p_{f \cdot s}$ = den betingade sannolikheten för "egenskap" givet "bortfall"

$$p_{u\cdot r} = 1 - p_{f\cdot r}$$

$$p_{u\cdot s} = 1 - p_{f\cdot s}$$

En bekväm à priorifördelning för $\tilde{p}_r$, $\tilde{p}_{f\cdot r}$ och $\tilde{p}_{f\cdot s}$ är en som består av produkten av oberoende betafördelningar

$$p_r^{n'_r-1} \quad p_s^{n'_s-1} \quad p_{f\cdot r}^{n'_{fr}-1} \quad p_{u\cdot r}^{n'_{ur}-1} \quad p_{f\cdot s}^{n'_{fs}-1} \quad p_{u\cdot s}^{n'_{us}-1}$$

där alla n' symboliserar parametrarna i denna simultana à priorifördelning och där

$$n'_f = n'_{fr} + n'_{fs} \qquad \text{och} \qquad n'_u = n'_{ur} + n'_{us}$$

Antag att data genereras i enlighet med en s k dubbel dikotomiprocess. Sannolikheten att observera ett sample om n bestående av n_{fr} respondenter med egenskapen, n_{ur} som ej har egenskapen och n_s bortfalls individer är då proportionell mot

$$(p_{fs} + p_{us})^{n_s} \quad p_{fr}^{n_{fr}} \quad p_{ur}^{n_{ur}}$$

I termer av de betingade sannolikheterna p_r, $p_{f\cdot r}$ och $p_{f\cdot s}$ blir likelihooden proportionell mot

$$p_s^{n_s} \quad p_r^{n_r} \quad p_{f.r}^{n_{fr}} \quad p_{u\cdot r}^{n_{ur}}$$

Om vi använder den ovan angivna beta à priorifördelningen för p_f, $p_{r\cdot f}$ och $p_{r\cdot u}$ och multiplicerar med likelihooden enligt Bayes' teorem erhålls

$$p_f^{n''_f-1} \quad p_u^{n''_u-1} \quad p_{r\cdot f}^{n''_{fr}-1} \quad p_{s\cdot f}^{n'_{fs}-1} \quad p_{r\cdot u}^{n''_{ur}-1} \quad p_{s\cdot u}^{n'_{us}-1} \quad \left[p_f p_{s\cdot f} + p_u p_{s\cdot u}\right]^{n_s}$$

där

$$n''_f = n'_f + n_{fr} \qquad \text{och} \qquad n''_u = n'_u + n_{ur}$$

À posteriorifördelningen för p_f blir en beta med medelvärde

$$\bar{p}''_f = (n''_f + \bar{n}''_{fs})/n''$$

där

$$\bar{n}''_{fs} = n_s \, n'_{fs}/n'_s$$

$$n' = n'_f + n'_u = n'_r + n'_s$$

$$n = n_f + n_u = n_r + n_s$$

$$n'' = n' + n$$

Medelvärdet $\bar{p}''_f$ kan tolkas på följande sätt: Om antalet bortfallsindivider med egenskapen vore känt, så skulle à posteriorimedelvärdet för p_f vara $(n''_f + n_{fs})/n''$. Eftersom vi inte känner n_{fs} ersätts denna med dess förväntade värde $\bar{n}''_{fs}$.

Om man däremot ansätter icke-informativa à priorifördelningar för p_r, $p_{f\cdot r}$ och $p_{f\cdot s}$ à la Jeffreys (se avsnitt 8.2) och därefter observerar data leder detta till en databaserad à priorifördelning som kan användas vid fortsatt analys.

I ansatserna ovan antar man att bortfallet uppföljs med hjälp av urval i bortfallsstratum. I Rubin (1977) utnyttjas tillgänglig hjälpinformation. Ett subjektivt intervall konstrueras med hjälp av

a) data och
b) subjektiva parametrar som specificerar à priorifördelningen för bortfallets parametrar givet svarsstratums parametrar.

Antag att proportionen bortfall är p i ett urval om n och att man är intresserad av en parameter $\bar{Y}$. En ideal skattning vore

$$\bar{y} = (1-p)\,\bar{y}_R + p\,\bar{y}_{NR}$$

där $\bar{y}_R$ är det observerade medelvärdet i svarsstratum och y_{NR} är motsvarande okända storhet i bortfallsstratum.

Antag att det finns bakgrundsinformation om q variabler $\underline{X} = (X_1, X_2, \ldots, X_q)$ för alla n objekten i urvalet. Om $\bar{Y}$ har ett linjärt samband med $\underline{X}$ i både svars- och bortfallsstratum så gäller

$$E(\bar{y}_R) = \alpha_R + \underline{\beta}_R \, \underline{\bar{X}}'_R$$

$$E(\bar{y}_{NR}) = \alpha_{NR} + \underline{\beta}_{NR} \, \underline{\bar{X}}'_{NR}$$

där $\underline{\bar{X}}'_R$ och $\underline{\bar{X}}'_{NR}$ är de observerade medelvärdena för $\underline{X}$ i urvalet och resten är okända parametrar.

Tyvärr kan man inte skatta α_{NR} och $\underline{\beta}_{NR}$. Nu definieras två subjektiva parametrar, θ_1 och θ_2. Parametern θ_1 formaliserar subjektiva antaganden om hur lika lutningen Y på X kan vara i de båda strata. Parametern θ_2 formaliserar subjektiva antaganden om hur lika E(Y) är i de båda strata för sådana individer som har $\bar{\underline{X}}$ ungefär lika med $\underline{\bar{X}}_R$. Om θ_1 är subjektiva variationskoefficienten för bortfallsindividernas regressionskoefficienter och $\underline{\beta}_R^{(i)}$ och $\underline{\beta}_{NR}^{(i)}$ är dom i:te komponenterna i $\underline{\beta}_R$ och $\underline{\beta}_{NR}$ så blir ett 95 %-igt tilltrosintervall för $\beta_{NR}^{(i)}$

$$\beta_R^{(i)}(1 \pm 2\ \theta_1)$$

Man har då antagit att intervallet för $\underline{\beta}_{NR}$ är centrerat kring $\underline{\beta}_R$.

Parametern θ_2 låter vi vara subjektiva variationskoefficienten för $\bar{Y}$ i en grupp bortfallsindivider vars $\underline{X}$-fördelning är ungefär samma som de svarandes. Om $\alpha_R + \underline{\beta}_R \, \underline{\bar{X}}'_R$ är E(Y) för en grupp svarande med $\underline{X}$ lika med $\underline{\bar{X}}_R$, så blir ett 95 %-igt tilltrosintervall för E(Y) bland bortfallsindivider vars $\underline{X}$ är ungefär lika med $\underline{\bar{X}}_R$

$$(\alpha_R + \underline{\beta}_R \, \underline{\bar{X}}'_R)(1 \pm 2\ \theta_2)$$

Om man använder parametriseringen (θ_1, θ_2) och antar att regressionerna Y på X har identiska och oberoende fördelade normala residualer, så kan man visa att ett subjektivt intervall för $\bar{y}$ är

$$\bar{y}_R \left[1 + h_0 \pm z\sqrt{\theta_1^2 h_1^2 + \theta_2^2 h_2^2 + h_3^2}\right]$$

där h_1, h_2 och h_3 är funktioner av observerade data och h_0 är en relativ bias, som speglar olika $\underline{X}$ för medverkande och bortfall. z är ≈ 2 om den önskade nivån är 95 %. I ett exempel redovisas i detalj användning av h_i. Alla h_i är relativa varianser.

Enklast varierar man θ_1 och θ_2 för att kontrollera effekten på $\bar{y}$. Intervallet är mer känsligt för variationer i θ_2 än i θ_1.

8.6 Biasproblemet

Den klassiska statistiska teori som Gauss utvecklade var avsedd för naturvetenskapliga tillämpningar. Ett exempel kan vara vägning av ett föremål. En n gånger upprepad vägning ger observationsserien:

$$y_1, y_2, \ldots\ldots, y_n$$

Dessa värden kan betraktas som observationer på en stokastisk variabel

$$Y = \mu + d$$

där μ är den konstanta sanna vikten medan d är resultatet av mätfel och där $E(d) = 0$. Antag att vågen är felinställd. Vi erhåller då observationsserien:

$$y_1', y_2', \ldots\ldots, y_n'$$

Dessa värden kan betraktas som observationer på en stokastisk variabel

$$Y' = \mu + B + d$$

där B är biasen hos vågen och lika med $E(Y') - E(Y)$.

Detta betraktelsesätt kan lätt överföras till surveyområdet. En tänkt serie om m observationer av en individ J kan betraktas som observationer på en stokastisk variabel

$$Y_J = \mu_J + B_J + d_J$$

där $B_J = E(Y_J) - \mu_J$.

Biaskomponenten är en svår term att uppskatta i de flesta surveysituationer. Det krävs uppenbarligen kunskap om μ_J i en eller annan form för att problemet ska bli lösbart.

I teorin går biasestimation till på det sättet att man för ett urval av element skaffar fram sanna värden. Det vanliga är att ett sådant anskaffande inte låter sig göras p g a att sanna värden inte existerar eller för att kostnaderna att ta fram dessa är prohibitiva. Ett praktiskt alternativ är då att skaffa information om en approximation μ'_J av μ_J. Tillvägagångssättet är att "bättre" information samlas in om ett urval av element och att man samtidigt hoppas att denna information är behäftad med ingen eller ringa bias.

I Schlaiffer (1959) ges ett exempel på hur man kan ansätta en à priorifördelning för biastermen B baserad på historiska data. I detta fall betraktas förstås alla komponenterna i relationen

$$Y = \mu + B + d$$

som stokastiska variabler.

Om à priorifördelningen för Y är normal, om à priorifördelningen för μ är normal och om $\sigma^2(Y)$ är känd så är à posteriorifördelningen för μ normal där förväntan är

$$\frac{\frac{1}{\sigma^2(\mu)} E(\mu) + \frac{1}{\sigma^2(Y)} \left[\bar{y} - E(B)\right]}{\frac{1}{\sigma^2(\mu)} + \frac{1}{\sigma^2(Y)}}$$

där $\bar{y}$ är sampleinformation och resten av komponenterna à priori fastställda.

8.7 Kvalitetskontroll

Industriell statistisk kvalitetskontroll förekommer inte sällan vid kontroll av enskilda operationer i surveys. Exempel på sådana operationer är kodning och stansning. Sådan kvalitetskontroll tillgår i regel på så sätt att en urvalsplan utformas vars syfte är att upprätthålla en viss given kvalitet (t ex proportion defekta element) på operationens output. Ett exempel är att en viss utgående kvalitet garanteras med en viss AOQL (= Average Outgoing Quality Limit). Vanligen indelas hela produktionen i väldefinierade bitar, s k lots eller batches, där urval av producerade element sker inom varje batch. På grundval av urvalsresultatet bestäms huruvida batchen ska förkastas eller accepteras. I det förra fallet kontrolleras hela batchen medan ingen ytterligare åtgärd vidtas i det senare fallet. Förfarandet kan kallas acceptanssampling. En plan för sådant urval definieras av urvalsstorleken, n, och acceptanstalet, t. För att erhålla något slags optimal urvalsplan krävs kunskap om proportionen defekta i produktionsprocessen. I konventionell sampling används oftast sådan kunskap endast om den kan antas vara mer eller mindre exakt. Man är inte speciellt angelägen att betrakta proportionen defekta som en stokastisk variabel utan väljer hellre en plan med alltför stort n.

I Barnett (1973) diskuteras hur à prioriinformation om proportionen defekta kan användas i en bayesansats vid kvalitetskontroll. Antag att proportionen defekta i batch (k+1) är θ_{k+1}. En enkel punktestimation av θ_{k+1} kan göras genom att sampla element ur batch k+1 och skatta proportionen med $\hat{\theta}_{k+1} = r/n$ där r är antalet defekta och n samplestorleken. I många situationer kan vi ha en mer eller mindre klar bild av produktionsprocessens tidigare led, nämligen framställning av batcharna 1, 2,...., k. Man kan exempelvis tänka sig att de k tidigare batcharna totalkontrollerats medan man för batch k+1 och ett tag framåt övergår till urvalskontroll. Det är alltså tänkbart att vi har en uppsättning värden $\theta_1, \theta_2, \ldots, \theta_k$ som är kända. Den fördelning som dessa θ-värden uppvisat kan nu betraktas som en à priorifördelning där θ_{k+1} är en observation från denna fördelning. Uppsättningen $\theta_1, \theta_2, \ldots\ldots, \theta_k$ kan representeras av exempelvis ett histogram för vilket man kan anpassa en kontinuerlig funktion. Denna kan sedan tjäna som à priorifördelning för batchkvaliteten.

I artikeln ges ett numeriskt exempel (k=498) där den anpassade fördelningen representeras mycket väl av en $\beta(20.80)$

dvs $$\Pi(\theta) = \frac{\Gamma 100}{\Gamma 20\ \Gamma 80}\ \theta^{19}(1-\theta)^{79}$$

Dess typvärde är $\tilde{\theta} \approx 0.194$.

I den batch (k+1) som vi samplar ur är punktestimatet av proportionen defekta $\hat{\theta} = 0.3$. Om vi nu på bayesianskt vis kombinerar à prioriinformationen med sampleinformationen istället för att enbart utnyttja den senare så erhålls en beta à posteriorifördelning som är $\beta(80.220)$(r=60, n=200). Typvärdet i denna fördelning är $\hat{\theta} = 79/298$ eller 0.265 och vi ser hur à posteriorivärdet är en kompromiss mellan sample- och à prioriinformation.

I en acceptanssamplingssituation där önskemålet är att utforma en plan med minimering av kostnaderna är jämviktskvaliteten $\frac{b}{b+c}$, där b är vinsten av att godkänna en icke-defekt enhet och c förlusten av att godkänna en defekt enhet och där urvalskostnaden är 1. Vid jämviktskvalitet är det lika lönsamt att förkasta som att acceptera en batch. Om à priorifördelningen för inkommande kvalitet θ är $\beta(1,m)$ och urvalsstorleken är n så accepteras enligt Wetherhill (1969) batchen om $r \leq t$ och förkastas om $r > t$ där

$$t = (1+m+n)\left(\frac{b}{b+c}\right) - 1$$

Om b/b+c är 1/4 i exemplet ovan med $\Pi(\theta) = \beta(20.80)$ och n = 200, så blir t=55. Men r=60 varför batchen förkastas.

å priorifördelningar i form av olika betafördelningar är ett vanligt sätt att representera information om proportionen defekta i en batch. I Hendricks (1964) ges ytterligare ett exempel på beta som å priorifördelning i en acceptanssampling-situation. Men man kan givetvis tänka sig andra representationer.

Bayesanalys i acceptanssamplingsituationer diskuteras ytterligare i bl a Hald (1968), Weiler (1965), Ladany (1976).

Ett annat exempel på hur å prioriinformation skulle kunna utnyttjas i kvalitetskontrollsyfte är följande.

I en survey förekommer databearbetningsoperationer som kodning och stansning. Dessa operationer är i relativt stor utsträckning felgenererande och man vill därför i många fall ha en uppfattning om den utgående kvaliteten. Antag exempelvis att man önskar skatta antalet felstansade kort i en hålkortsmassa. Ett obundet slumpmässigt urval om n kort uttages. Vid kontroll av dessa n kort visar det sig, att m kort innehåller ett eller flera stansfel. Enligt gängse metodik skattas nu proportionen felaktiga kort med $\hat{p} = m/n$. Men man kan också tänka sig att utnyttja resultat från tidigare kontrollstansningar. Låt oss anta att man från tidigare stansningar av samma slag har erhållit kännedom om felprocenten givet en viss storlek på undersökt hålkortsmassa. Man känner m a o tabellen

p_j (felprocent)	$P_0(p_j)$ (relativ frekvens)
p_1	$P_0(p_1)$
p_2	$P_0(p_2)$
.	.
.	.
.	.
.	.
p_ℓ	$P_0(p_\ell)$
	$\sum_{j=1}^{\ell} P_0(p_j) = 1$

Vi kan för praktiska ändamål tänka oss att reducera värdeförrådet för p_j, så att det består av exempelvis 0 och multipler av tiondels procent. Vid konstruktionen av tabellen kan man då justera observerade p_j-värden till närmaste tiondel i värdeförrådet för p_j i tabellen. Tabellen fungerar nu som en diskret à priorifördelning.

För alla p_j-värden beräknas nu $P(x=m|p_j)$, där x är antalet felaktigt stansade kort i urvalet. Antagandet om oberoende medför att

$$P(x=m|p_j) = \binom{n}{m} p_j^{\,m} (1-p_j)^{n-m}$$

Enligt Bayes' teorem erhålles nu

$$P'(p_j) = P(p_j|x=m) = \frac{P(x=m|p_j)\, P_0(p_j)}{\sum\limits_{j=1}^{\ell} P(x=m|p_j) P_0(p_j)}$$

En skattning av stansfelet blir nu

$$\hat{p}' = \max_j \{P'(p_j)\}$$

8.8 Evalvering av en survey

Evalvering innebär att ett mått framräknas för det totala felet i en survey. Detta tillgår vanligtvis så att surveyresultatet jämförs med ett estimat av ett "sant" värde. I Brown (1963), (1967) och (1969) redovisas en subjektivistisk ansats för evalvering. Ansatsen är inte bayesiansk eftersom Bayes' teorem inte används men jag har ändå valt att redovisa den p g a det subjektiva inslaget. Ansatsen kan kallas tilltrosanalys och den innebär att subjektiv information används för att korrigera framräknade estimat grundade på urval.

Den brownska filosofin är att den tilltro, som ligger i ett surveyresultat, är den logiska konsekvensen av tilltron beträffande var och en av surveyens olika led. Om dessa "tilltroskomponenter" kan kvantifieras, så finns möjligheter att räkna fram surveyens totala tilltro med hjälp av konventionella statistiska mått (varianser och kovarianser).

Antag, att vi beräknar en mängd delestimat. "Tvivel" beträffande ett slutgiltigt estimat beror då inte endast på varje delestimat utan också på den subjektive bedömarens värderingar beträffande kovariansen mellan estimaten. För bedömaren gäller inte endast att bestämma tilltron för varje delestimat utan också att formalisera denna tilltro. Anta, att man vill estimera en parameter T med hjälp av en urvalsundersökning. Undersökningen ger ett estimat E av parametern. Det är nu möjligt att betrakta kvoten mellan T och E. Denna kvot kan betraktas som en produkt av två eller flera andra kvoter. Kvoterna korrensponderar mot olika felkällor. Ett exempel på detta är

$$\frac{T}{E} = \frac{T}{M} \cdot \frac{M}{E}$$

där M är det värde, som erhålls genom att mäta hela populationen på samma sätt som urvalet. Om T/M är skilt från 1, så föreligger mätfel och om M/E är skilt från 1, så är urvalet inte representativt. För var och en av kvoterna kan en tilltrosfördelning ansättas. Evalveringen av mät- och urvalsfelen ger då värden på varianserna i dessa tilltrosfördelningar. Med hjälp av delresultaten kan så det slutliga "tvivlet" (= variansen) beträffande undersökningsresultatet T/E räknas fram. Denna variant av tilltrosanalys benämns "error ratio analysis", "felkvotsanalys". En förutsättning för att felkvotsanalys skall vara meningsfull är givetvis att delkvoterna kan evalveras lättare än det totala undersökningsresultatet.

I Frankel (1971) framförs kraftig kritik mot den av Brown föreslagna ansatsen. Kritiken riktar sig främst mot att felkvotsanalysen helt grundar sig på subjektiva uppfattningar om kvoterna ifråga. Inslaget av empiriska data kan i klassiska och neo-bayesianska design reducera effekten av felaktiga à prioriföreställningar. Någon sådan gardering finns inte i felkvotsanalysen.

8.9 Några andra operationer

I en survey ingår operationer som ibland innehåller moment som ej är traditionella surveymoment. Sådana moment kan vara variansanalys och experimentsplanering. Här ges några referenser till bayesansatser vid sådana moment.

Ansatser inom experimentplanering diskuteras i Grizzle och Novick (1965), Placket (1966) och Smith (1965).

Bayesanalys i en variansanalyssituation redovisas i Hoadley (1969).

8.10 Några speciella à priorifördelningar i surveys

Ibland kan ett slags allmänna à priorifördelningar uppstå i surveysituationer. I Ericson (1965) ges några exempel. Om det exempelvis gäller att kombinera tidsseriedata med sample surveydata och à priorifördelningen bestäms på basis av regressionsmodeller med användande av tidigare estimat på medelvärden μ_i, så kan en simultan k-variat normal à priorifördelning, där kovariansen mellan olika medelvärden inte nödvändigtvis är noll, vara naturlig. Om ett stratifierat urval dras så kommer den resulterande à posteriorifördelningen att vara en naturlig kombination av tidsserier eller regression och sample surveyinformation.

En annan allmän väg sådana à priorifördelningar kan uppstå på är i repetetiva surveysituationer. Antag att μ_h är en kxl observationsvektor för k stratummedelvärden i en i:te tidigare period. Dessa μ_h (h=1,....., n) kan betraktas som ett urval av n från en multivariat normalfördelning. À priorifördelningen för den innevarande (n+1):te perioden kan då ses som en k-variat normalfördelning med medelvärde och kovariansmatris estimerad från de n tidigare μ_h:na.

Superpopulationstanken kan också konkretiseras. Exempelvis kan okända stratummedelvärden μ_j (j=1,...., L) liknas vid en realisering av L oberoende urval av en superpopulation med medelvärde m och varians v.

8.11 Allmänt om bayesanalys i surveys

I Winkler (1973) görs några allmänna reflexioner kring utnyttjandet av bayesansatser i statistiska undersökningar. Man kan fastslå att bayesansatser ännu så länge sällan används i praktiken. Orsakerna kan bl a vara att

i) traditionen talar för klassiska ansatser

ii) träningen i utnyttjandet är liten

iii) teorin kan vara svårtillgänglig

iv) beräkningssvårigheterna kan vara stora om inte konjugatfördelningar och datorer används

v) rapporteringssvårigheter föreligger

vi) ett visst motstånd föreligger från tidskriftsredaktörernas sida

vii) filosofiskillnad existerar.

I praktiken är kanske den senaste orsaken den som har den minsta betydelsen. Traditionen betyder troligen mer. Vad sägs om följande ritual?

Vid icke-bayesiansk hypotesprövning används ofta stora urval vid standardtestning av s k skarpa nollhypoteser; dvs om $\theta=0$. Egentligen är man oftast intresserad av om θ ligger i närheten av 0. Stora urval ger ofta vid handen att nollhypotesen bör förkastas. Men man förkastar därmed en hypotes som man egentligen inte är intresserad av. Testet är alltför precist för att avgöra om θ ligger nära 0. I bayesiansk hypotesprövning ges istället relativ tilltro till alternativa hypoteser.

Bayesianska metoder har inte gjorts tillgängliga för undersökningsanalytiker. Metoderna finns inte i "kokbokstappning", Vidare existerar endast ett fåtal dataprogram och tabellverk för bayesanalys varav några finns i Novick och Jackson (1974).

Tillämpningen av bayesansatser kräver dessutom mer tid, arbete och grad av sofisticering än traditionella ansatser.

En fördel med en bayesansats är att resultaten kan presenteras i en lättfattlig form. Ett intervallestimat blir av typen P(θ ligger i intervallet) istället för det traditionella P(intervallet innesluter θ). (Många gånger tolkar icke-bayesianerna de facto sina resultat på det förra sättet.) Som nämndes ovan ges en möjlighet att besvara frågan: "Vilken hypotes är troligast?"

Trots det som sagts om möjligheterna att presentera bayeslösningar i en lättfattlig form får bayesianen ofta en extra börda i och med att alla procedurer som ingår i analysen måste förklaras. Ett traditionellt t-test är så välbekant att det kan rapporteras i form av ett t-värde. Därför uppstår ännu så länge kommuniceringssvårigheter med resultatanvändaren.

En bayesrapport måste innehålla resultat där man gör en avvägning mellan fullständig och koncis framställning. Fullständighet kräver att alla rådata publiceras. Om likelihoodfunktionen publiceras kan olika datautnyttjare använda sina egna à priorifördelningar för erhållande av à posteriorifördelningar. För att minska presentationsbördan kan man ange flera olika à posteriorifördelningar korrensponderande mot alternativa à priorifördelningar. En möjlighet är att utnyttja principen för stabil estimation och ange en à posteriorifördelning korresponderande mot en diffus à priorifördelning. En annan möjlighet är att ange funktioner som relaterar parametrarna i konjugata à priori- och à posteriorifördelningar. Om funktionerna presenteras grafiskt är det lätt att se hur känslig à posteriorifördelningen är för förändringar i à priorifördelningen.

I de fall à posteriorifördelningen är komplex, t ex i multivariata sammanhang är det önskvärt att sammanfatta fördelningen i termer av parametrar, sannolikheten för intervall etc.

En ytterligare variant är att redovisa hur känslig à posteriorifördelningen är för ändringar i sampleinformationen.

AVD C AVSLUTNING

9 Kritik mot den neo-bayesianska ståndpunkten

Mycket kritik har riktats mot den ståndpunkt som företräds av neo-bayesianerna. Kritik i den motsatta riktningen är dock relativt sett vanligare och detta är naturligt. Det gäller ju här att propagera för något nytt. Snart sagt varje bayesartikel innehåller jämförelser med den klassiska skolan och jämförelserna är givetvis oftast till förmån för den egna skolan.

Beträffande den kritik som riktats mot bayesianerna så har den undergått förändringar under årens lopp. Ett citat från Fisher kan vara belysande.

"The theory of inverse probability is founded upon an error and must be wholly rejected. Inferences respecting populations from which known samples have been drawn, cannot by this method be expressed in terms of probability".

Sedan detta yttrande fälldes har både kretsen av bayesianer utvidgats och kritiken nyanserats. Många icke-bayesianer anser att kontroversen inte längre är av filosofisk natur, utan snarare att bayesanalys många gånger är komplex och svårgenomförbar. Det praktiska värdet av bayesanalys ifrågasätts också ofta. Många anser att à priorikunskaper mestadels är av så vag natur att de är av ringa eller intet värde.

En annan invändning då det gäller surveytillämpningar är att varje användare av en survey skulle tvingas att ansätta egna à priorifördelningar. Detta är orimligt eftersom användarna oftast är andra individer än planerarna av undersökningen. I ett sådant läge med många användare tycks den enda vettiga lösningen vara att ansätta diffusa à priorifördelningar. Detta ger resultat som ligger mycket nära dem som erhålles via ett konventionellt genomförande av surveys. Detta betyder från bayesiansk horisont att om à prioriinformationen är vag är ingen skada skedd eftersom resultatet ligger nära det traditionella. Däremot kvarstår styrkan i Bayes' teorem när à prioriinformationen är specifik.

Det anses av vissa att teorin för utbytbara à priorifördelningar kan bli en stark länk mellan bayesiansk och klassisk teori. Tvåstegsansatsen att generera en à priorifördelning genom att definiera en parametrisk superpopulation och sedan välja en à priorifördelning för parametrarna i denna utgör också en naturlig korrespondens mellan bayesiansk samplingteori och annan statistisk teori.

I bl a Ericson (1969a), Smith (1965) och Berkson (1977) redovisas kritik mot bayesansatser.

10 Syntes

Som nämnts ovan är framställningen om inkorporering av bayesianska idéer i surveyplaneringen relativt "smalt" inriktad. Sålunda ligger tyngdpunkten på tillämpningar vid stratifierat urval. Men även där är tillämpningen "smal". Allmänt förutsätts exempelvis kända varianskomponenter.

Kan man då använda bayesianska metoder i surveys? Frågan kan besvaras med ett ja. I själva verket finns en bayesiansk survey gällande ett författarskapsproblem redovisad i Mosteller och Wallace (1963) och (1964) där det gällde att bestämma à posteriorisannolikheter för författarskap där sampleinformationen utgjordes av ordfrekvenser. I denna studie framgick med önskvärd tydlighet att avsaknaden av praktiska försök var stor. Så värst många ytterligare försök har heller inte gjorts sedan Mostellers och Wallaces studie genomfördes.

Några exempel finns dock i Zellner och Williams (1973) och Murphy och Winkler (1973).

Däremot har som ovan visats en mängd teoretiska artiklar arbetats fram. Detta beror säkerligen till stor del på att metoderna är svårtillämpade snarare än kontroversiella. Det behövs nämligen en större arsenal av à priorifördelningar med hänsyn till genomförande och faktisk representation av företeelser. Så snart man överger de enklaste bayestillämpningarna så kommer man i matematiska svårigheter. Därför krävs en systematisk utveckling gentemot enklare förfaranden där exakta lösningsmetoder kan negligeras.

Själva idén med formellt hänsynstagande till à prioriinformationen är givetvis tilltalande. Konsekvenserna av de sekvensiella bayesdesign som bl a redovisas i Solomon och Zacks (1970) leder i en surveysituation till att man kan tänka sig att sampla kunder i en affär tills stängningsdags, tills man blir trött, tills man tror man har tillräckligt evidence för något man vill påvisa etc.

Den springande punkten i en tillämpningssituation är dock à priorifördelningens utseende. Enligt bayesianskt synsätt kan denna konstrueras helt subjektivt. Man kan dock tänka sig naturliga hjälpmedel härvidlag såsom provundersökningar. Som nämnts ovan gäller det dock att konstruera "standards" för olika à prioriinformationssituationer. Man måste finna à priorifördelningar som passar komplexa urvalsdesign. Det behövs à priorifördelningar som speglar snedhet osv.

Man behöver dock inte använda bayeslösningar från början till slut i en survey. Man kan mycket väl tänka sig att enbart begränsa bayesutnyttjandet till enskilda operationer såsom att försöka lösa bortfallsproblemet i en survey eller att utnyttja à prioriinformation för batchkvalitet vid kvalitetskontroll av exempelvis stansning som ju förekommer i de flesta större undersökningar.

11 Referenser

Aggarwal, O P (1959): Bayes and minimax procedures in sampling from finite and infinite populations. The Annals of Mathematical Statistics, sid 206-218.

Aggarwal, O P (1966): Bayes and minimax procedures for estimating the arithmetic mean of a population with two-stage sampling. The Annals of Mathematical Statistics, sid 1186-1195.

Anscombe, F J och Auman, R J (1963): A definition of subjective probability. The Annals of Mathematical Statistics, sid 199-205.

Barnett, V (1973): Bayesian, and decision thecretic, methods applied to industrial problems. The Statistician, sid 199-226.

Bayes, T (1763): An essay towards solving a problem in the doctrine of chances. Återtryckt i Biometrika, 1958, sid 293-315.

Berkson, J (1977): My encounter with neo-Bayesianism. International Statistical Review, sid 1-11.

Box, G E P och Tiao, G C (1965): Multiparameter problems from a Bayesian point of view. The Annals of Mathematical Statistics, sid 1468-1482.

Brown, R V (1963): Credence analysis. A guide to choosing research strategies. Metra, sid 361-383.

Brown, R V (1967): Evaluation of total survey error by error ratio analysis. Metra, sid 593-613.

Brown, R V (1969): Research and the credibility of estimates. Harvard University.

Christianson, A (1974): Uppskattning av kunskap med multiple choicefrågor. Statistisk Tidskrift, nr 4.

Clutton-Brock, M (1965): Using the observations to estimate the prior distribution. Journal of the Royal Statistical Society, serie B, sid 17-27.

Cornfield, J (1967): Bayes' theorem. Review of the International Statistical Institute, sid 34-49.

Cornfield, J (1969): The Bayesian outlook and its applications. Biometrics, sid 617-642.

Cox, D R (1952): Estimation by double sampling. Biometrika, sid 217-227.

Dalenius, T (1971): Principer och metoder för planering av samplingundersökningar. Intern handbok nr 4, Statistiska centralbyrån, Stockholm.

Dalenius, T (1974): Utnyttjandet av diffus à prioriinformation vid sample surveys. Artikel presenterad vid SCBs forskarkonferens i statistisk metodik 1974.

de Finetti, B (1964): Foresight: its logical laws, its subjective sources. I Kyburg och Smokler: Studies in subjective probability, Wiley.

de Finetti, B (1972): Subjective or objective probability: is the dispute undecidable? Istituto Nazionale di alta matematica, Symposia Mathematica, Vol IX, s 21-36.

de Finetti, B (1974a): Theory of probability, vol 1. Wiley.

de Finetti, B (1974b): Bayesianism: Its unifying role for both the foundations and applications of statistics. Review of the International Statistical Institute.

De Groot, M (1974): Reaching a consensus. Journal of the American Statistical Association, sid 118-121.

Draper, N R och Guttman, I (1968a): Some Bayesian stratified two-phase sampling results. Biometrika, sid 131-139.

Draper, N R och Guttman, I (1968b): Bayesian stratified two-phase sampling results: k characteristics. Biometrika, sid 587-588.

Edwards, W, Lindman, H och Savage, L J (1963): Bayesian statistical inference for psychological research, Psychological Review, sid 193-242.

Ericson, W A (1965): Optimum stratified sampling using prior information. Journal of the American Statistical Association, sid 750-771.

Ericson, W A (1967): Optimal sample design with nonresponse. Journal of the American Statistical Association, sid 63-78.

Ericson, W A (1968): Optimal allocation in stratified and multistage samples using prior information. Journal of the American Statistical Association, sid 964-983.

Ericson, W A (1969a): Subjective Bayesian models in sampling finite populations. Journal of the Royal Statistical Society, serie B. sid 195-232.

Ericson, W A (1969b): Subjective Bayesian models in sampling finite populations: stratification. I N L Johnson och H Smith (ed): New developments in survey sampling, Wiley, sid 326-357.

Frankel, L (1971): Recension av "Brown, R V: Research and credibility of estimates". Journal of the American Statistical Association, sid 656-657.

Good, I J (1950): Probability and the weighing of evidence. Griffin.

Gough, R G (1973): The effect of group format on aggregate subjective probability distributions. Artikel presenterad vid 1973 års konferens om subjektiva sannolikheter, Rom.

Grizzle, J F och Novick, M R (1965): A Bayesian approach to the analysis of data from clinical trials. Journal of the American Statistical Association, sid 81-96.

Hald, A (1968): Bayesian single sampling attribute plans for continous prior distributions. Technometrics, sid 667-683.

Hansen, M H och Hurwitz, W N (1946): The problem of non-response in sample surveys. Journal of the American Statistical Association, sid 517-529.

Hampton, J M, Moore, P G och Thomas, H (1973): Subjective probability and its measurement. Journal of the Royal Statistical Society, serie A, sid 21-42.

Hendricks, W A (1964): Estimation of the probability that an observation will fall into a specified class. Journal of the American Statistical Association, sid 225-232.

Hoadley, B (1969): The compound multinomial distribution and Bayesian analysis of categorical data from finite populations. Journal of the American Statistical Association, sid 216-229.

Hogarth, R M (1975): Cognitive processes and the assessment of subjective probability distributions. Journal of the American Statistical Association, sid 271-289.

Isaki, C T (1970): Survey design utilizing prior information. University Microfilms, Ann Arbor, Michigan.

Kaufman, G M och King, B (1973): A Bayesian analysis of nonresponse in dichotomous processes. Journal of the American Statistical Association, sid 670-675.

Khan, M Z (1976): Optimum allocation in Bayesian stratified two-phase sampling when there are m-attributes. Metrika, sid 211-219.

Ladany, S P (1976): A practical meaning of the Bayesian decision making approach to acceptance sampling. Journal of Quality Technology, sid 127-132.

Lyberg, L (1973): Användandet av neo-bayesianska idéer vid sample surveys. Forskningsprojektet FEL I UNDERSÖKNINGAR, rapport nr 66. Statistiska institutionen, Stockholms universitet.

Morrison, G (1967): Critique of "Ranking procedures and subjective probability distributions". Management Science, sid B253-254.

Mosteller, F och Wallace, D L (1963): Inference in an authorship problem. Journal of the American Statistical Association, sid 275-309.

Mosteller, F och Wallace, D L (1964): Inference and disputed authorship: The Federalist, Addison-Wesley.

Murphy, A H och Winkler, R L (1973): Subjective probability forecasting in the real world: some experimental results. Artikel presenterad vid 1973 års konferens om subjektiva sannolikheter, Rom.

Novick, M R och Jackson, P H (1974): Statistical methods for educational and psychological research. McGraw-Hill.

Palit, C D och Guttman, I (1972): Bayesian estimation procedures. Technical Report 131, Centre de Recherches Mathématiquea, Université de Montréal, Québec, Canada.

Palit, C D och Guttman, I (1973): Bayesian estimation procedure for finite populations, single stage designs, and normal populations. Communications in Statistics, sid 93-111.

Peers, H W (1968): Confidence properties of Bayesian interval estimation. Journal of the Royal Statistical Society, serie B, sid 535-544.

Phillips, L D (1973): Bayesian statistics for social scientists. Nelson.

Placket, R L (1966): Current trends in statistical inference. Journal of the Royal Statistical Society, serie A, sid 249-267.

Press, S J och Yang, C-F (1974): A Bayesian approach to second guessing "undecided" respondents. Journal of the American Statistical Association, sid 58-67.

Raiffa, H (1962): Bayesian decision theory. I Robert Machol och Paul Gray (ed): Recent developments in information and decision processes, sid 92-101.

Ramsey, F P (1964): Truth and probability. I Kyburg och Smokler (ed): Studies in subjective probability, Wiley.

Rao, J N K och Ghangurde, P D (1972): Bayesian optimization in sampling finite populations. Journal of the American Statistical Association, sid 439-443.

Reilly, P M (1976): The numerical computation of posterior distributions in bayesian statistical inference. Applied Statistics, sid 201-209.

Roberts, H V (1963): Bayesian inference. Proceedings of the Social Statistics Section, The American Statistical Association, sid 76-80.

Royall, R (1968): An old approach to finite population sampling theory. Journal of the American Statistical Association, sid 1269-1279.

Rubin, D B (1977): Formalizing subjective notions about the effect of nonrespondents in sample surveys. Journal of the American Statistical Association, sid 538-543.

Savage, L J (1961): The subjective basis of statistical practice. Stencil.

Savage, L J (1962a): The foundation of statistical inference. Methuen.

Savage, L J (1962b): Bayesian statistics. I Robert Machol och Paul Gray (ed): Recent developments in information and decision processes, sid 161-194.

Savage, L J (1971): Elicitation of personal probabilities and expectations. Journal of the American Statistical Association, sid 783-801.

Schlaiffer, R (1959): Probability and statistics for business decisions. Mc Graw-Hill.

Schlaiffer, R (1969): Analysis of decisions under uncertainty. McGraw-Hill.

Scott, A och Smith T M F (1971): Bayes estimates for subclasses in stratified sampling. Journal of the American Statistical Association, sid 834-836.

Smith, C A B (1965): Personal probability and statistical analysis. Journal of the Royal Statistical Society, serie A, sid 469-489.

Solomon, H och Zacks, S (1970): Optimal design of sampling from finite populations. Journal of the American Statistical Association, sid 653-677.

Staël v. Holstein, C-A S (1970): Assessment and evaluation of subjective probability distributions. The Economic Research Institute, Stockholm, Sweden.

Stein, C (1945): A two sample test for a linear hypothesis whose power is independent of the variance. Annals of Mathematical Statistics, sid 243-258.

Stone, M (1976): Strong inconsistency from uniform priors. Journal of the American Statistical Association, sid 114-116.

Tomlin, P H (1977): Ratio- and regression-type estimators employing à priori knowledge. American Statistical Association. Proceedings of Social Statistics Section.

Weiler, H (1965): The use of incomplete beta functions for prior distributions in binomial sampling. Technometrics, sid 335-347.

Wetherill, G B (1969): Sampling inspection and quality control. Methuen.

Winkler, R L (1967a): The assessment of prior distributions in Bayesian analysis. Journal of the American Statistical Association, sid 776-800.

Winkler, R L (1967b): The quantification of judgment: some methodological suggestions. Journal of the American Statistical Association, sid 1105-1120.

Winkler, L (1968): The consensus of subjective probability distributions. Management Science, sid B61-B75.

Winkler, R L (1969): Scoring rules and the evaluation of probability assessors. Journal of the American Statistical Association, sid 1073-1078.

Winkler, R L (1973): Statistical analysis: Theory versus practice. Preliminary report.

Zacks, S (1969): Bayes sequential designs of fixed size samples from finite populations. Journal of the American Statistical Association, sid 1342-1349,

Zacks. S (1970): Bayesian design of single and double stratified sampling for estimation proportions in finite populations. Technometrics, sid 119-130.

Zellner, A och Williams, A D (1973): Bayesian analysis of the federal reserve - MIT - penn model's Almon lag consumption function. Journal of Econometrics, sid 267-299.