Är bayesianska metoder användbara?
–
Tankar kring representativitet, bortfall och andra kvalitetsproblem

Daniel Thorburn
Stockholms universitet

Frimis, Linköping
28 nov 2018

Uppläggning

- Introduktion till Bayes 3
- Bortfall/representativitet 7
  - Hjälpvariabler 8
  - Svarsordning 12
  - Ingen information 19
- Andra kvalitetsaspekter 24

Bayesianska metoder

- Med Bayesianska metoder kan mer kunskap eller sidoinformation modelleras i sannolikhetstermer. Mer kan alltså ingå i analysen. Bra när man har kunskap från andra källor eller gissningar, som man vill kunna använda t ex
  - Kunskaper om parametern från andra källor
  - Erfarenhet av andra felkällor som svarsfel, bortfallsorsaker, över- eller undertäckning …
  - Betydelsen av precision i relation till kostnaden

- I genomgången skall jag undvika alla formler och försöka förenkla härledningarna så att idéerna framgår (på bekostnad av stringensen)

- Börjar med ett enkelt exempel på Bayes formel
  - Vad vi vet innan + vad undersökningen säger = vad vi vet efteråt

Enkelt exempel: Tänk er en opinionsundersökning

- Bakgrund:
  - Förra månaden hade ett visst parti 17 % och osäkerheten var 1 % (savv).
  - Erfarenhetsmässigt vet man att andelen förändras med 1 % (savv) mellan månaderna, lika gärna upp som ner.
- Á priori denna månad:
  - Utan mer information är andelen alltså ca 17 % med ett fel av 1,4 % (variansen är1+1=2).
- Mätning
  - Denna månad; 19 % med ett slumpfel av 0,7 %.
- Á posteriori denna månad:
  - Dessa siffror kan vägas samman optimalt med hänsyn till precisionen Den bästa skattningen blir 18,6 % med ett slumpfel av 0,6 %.

Opinionsundersökningar (forts)

- Denna process går att upprepa varje månad. Den senaste månadens skattning blir nästa månads á priori men med lite extra osäkerhet tillagd.

- Modellen går att använda för att beräkna olika sannolikheter för olika händelser t ex att SD blir största parti vid nästa ordinarie val/om ett halvår eller för att KD åker ur riksdagen. Men det är inte syftet nu.

Opinionsundersökningar (forts)

- Beskrivningen är förenklad – man kan bygga in mer t ex
  - En bakomliggande trend,
  - Att förändringar sker snabbare under valkampanjer,
  - Olika instituts skattningar vägs in till en helhet. Man kan behöva lägga in att de har olika systematiska fel

  Allt detta kan skattas historiskt

- Man kan också lägga in mer subjektiva värderingar t ex
  - partier utan regeringsansvar brukar gynnas,
  - speciellt mindre partier i koalitioner missgynnas,
  - vid utrikes kriser gynnas sittande regering

  Rekommenderas inte att statistikproducenten gör. Det får läsaren göra

Hur kan man hantera bortfallsfelet bayesianskt?

- Första frågan: Vad har man för information om bortfallet som kan användas?

- Tre tänkbara svar (minst)
  - Bias för andra variabler med känt svar (t ex register eller annan statistik). Om andra variabler har stora fel gäller det förmodligen även den vi tittar på.
  - Hur skiljer sig de som svarar snabbt eller långsamt. Stora skillnader tyder på att insamlingsförfarandet har stor effekt dvs stort bortfallsfel.
  - Ingen kunskap – välj ”likformig” á priori – dvs använd informationen att vi inget vet
  - Information från psykokometriska studier (tar jag inte upp)

- Hur skall denna information modelleras för att kunna ingå i analysen?

Första svaret: Andra variabler

- Tänk er att man har fem variabler där man vet svaret, kön, ålder, inkomst, utbildning, sysselsättningsgrad. För dem är relativa biasen är 15, –4, – 11, 7, 5 %. (Relativ bias är biasen genom standardavvikelsen per enhet)

- Om man inte har någon orsak att tro att den variabel som studeras skulle skilja sig från dessa har vi har sex dragningar från samma aprioriförde lning där den sjätte är den studerade. Då kan man tex säga att sannolikheten är 1/6 att den studerade variabeln har störst relativ bias, dvs är större än 15.

- Om man vågar anta normalfördelning kan man skatta spridningen s = 9 %. Den relativa biasen understiger alltså 18 % med 95 % sannolikhet.

- Detta kan sedan kombineras med det vanliga slumpfelet. Med 400 obs är det relativa slumpfelet 5 % och den totala standardavvikelsen 10,5 %. Med 95 % sannolikhet understiger den 21 %

- Detta förutsätter vanlig enkel datainsamling Om man använt hjälpvariablerna vid insamlingen t ex stratifierad insamling eller adaptiv sampling måste metoden modifieras.

- Samma metod fungerar hjälpligt också för att skatta representativitetsfelet vid t ex Web-paneler eller situationer med ofullständiga ramar (med ovanstående reservation).

Bortfallskorrigerade skattningar

- Om man har hjälpinformation försöker man ofta använda den för att förbättra skattningen med t ex vägning, kalibrering eller poststratifiering.

- Då tror man att skattningen blir bättre – men hur skall man uppskatta storleken på bortfallsfelet efter bortfallsvägning?

Gör samma sak (nästan)!

- Skatta de kända variablerna, en i taget, så bra som möjligt med samma typ av kalibrering (alla övriga hjälpvariabler). Beräkna relativ bortfallsbias för dem.

- De kvarvarande bortfallsfelen för kön, ålder, inkomst, sysselsättningsstatus blir kanske 8, 1, – 9, – 4 och 2 (istället för 15, – 4, – 11, 7, 5).

- Då kan man tro att sannolikheten är 1/6 att kvarvarande bias är större än 9.

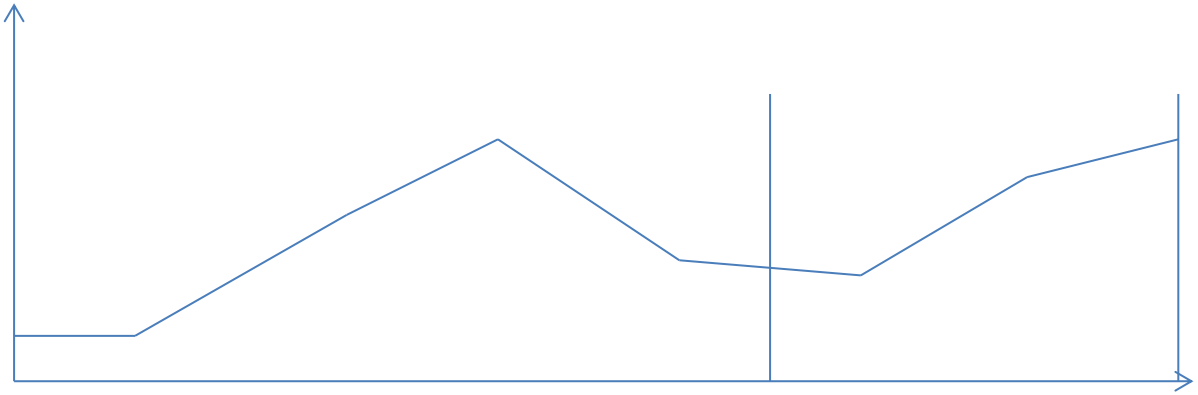
- Med normalfördelning kan bortfallsbiasens relativa storlek nu skattas till s = 6 och felet blir allts högst 12 med 95 % säkerhet.

- Om det vanliga slumpfelet efter kalibrering skattades till 3 så blir skattningen av totalfelet 7 (och högst 14 med 95 % säkerhet.

Andra svaret: info om svarsordningen

De som svarar snabbt skiljer sig från dem som behöver påminnelser, flera uppringningar eller tar mer tid på sig.

Utvecklingen kan se ut t ex så här

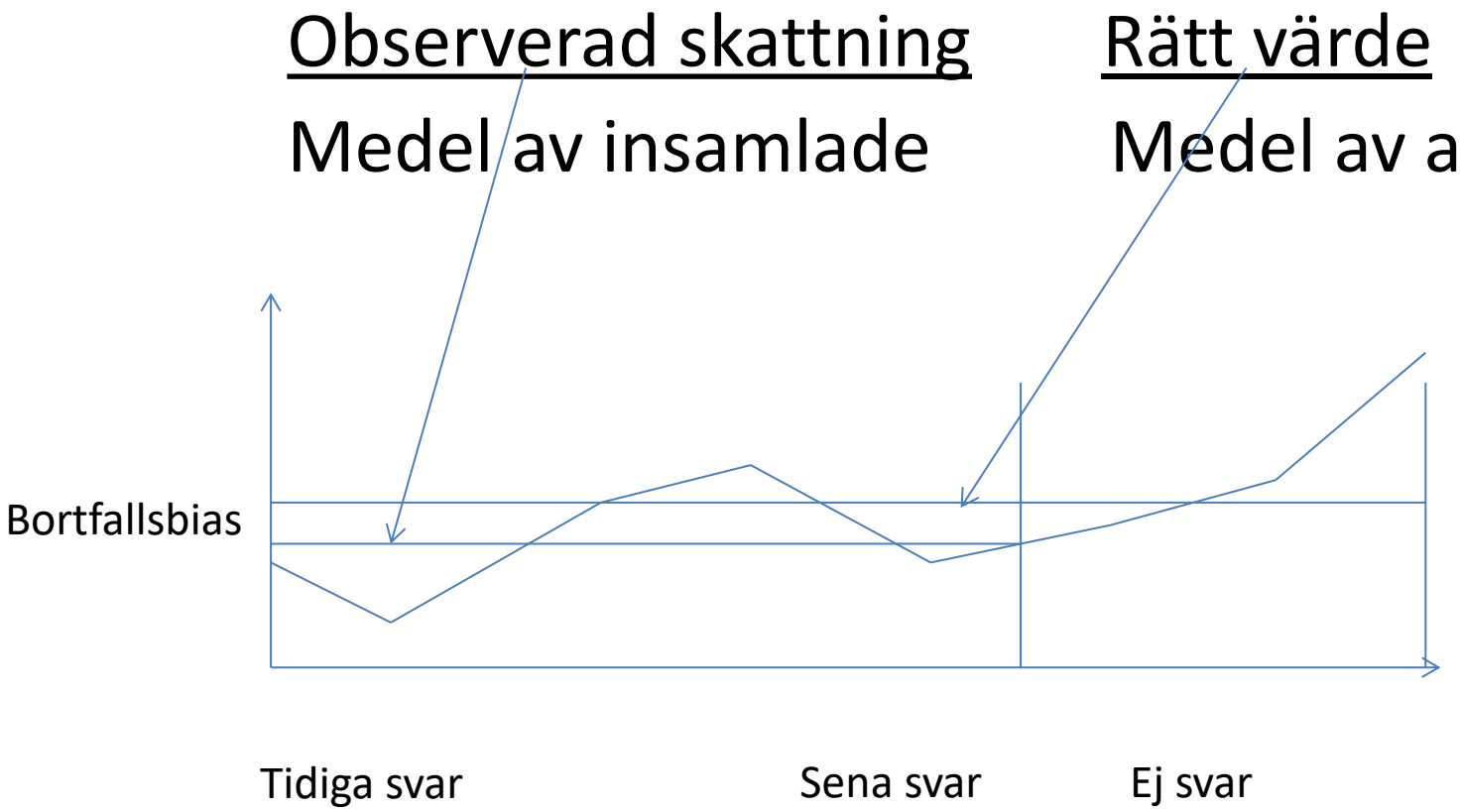


Tidiga svar

Sena svar

Behöver tex fler påminnelser
Dessa ser vi ej.
Måste modelleras

Dessa ser vi efter datainsamling


Observerad skattning
Medel av insamlade
Rätt värde
Medel av alla
Bortfallsbias
Tidiga svar
Sena svar
Ej svar

- Om man antar att kurvan är kontinuerlig men att man inte vet om den kommer att gå upp eller ner men är ungefär lika hackig så är vanlig slumpvandring en rimlig modell (à priori).

- Då kan man visa att standardavvikelsen för bortfallsbiasens är proportionell mot bortfallet (andel).

- Proportionalitetskonstanten (Hur hackig kurvan är) kan skattas om man noterar när svaren kommer (i dagar eller timmar eller antal uppringningar eller… .

- Man ser tex att bortfallsbiasen blir dubbelt så stor för en välgjord webb-panel-enkät som för en lika välgjord vanlig enkät med 50 % bortfall. (Slumpfelet går dock inte att skatta)

Praktiskt exempel, AKU

- Jag har fått data från SCB och kunnat skatta standardavvikelser för bortfallsbias för tre variabler
  - Månadsinkomst 4000 kr
  - Kön 0,5 %
  - Arbetslöshet 0,2 %
- Men data är inte särskilt bra och osäkerheterna i dessa skattningar är rätt stor

Undersökningsplanering

- Nu har vi en modell med både slumpfel och bortfallsfel i samma termer. Det betyder att de kan läggas samman, så att vi kan minimera totalfelets standardavvikelse.

- Om man kan beskriva kostnaden för att uppnå ett bestämd bortfallsnivå och urvalsstorlek så kan man lätt beräkna vilken kombination som ger det minsta totalfelet till given kostnad.

- I praktiken har man sällan hela svarsordningen men ofta har man t ex svarsdag, antal påminnelser eller antal uppringningsförsök. Detta brukar gå att använda även om man kan behöva modifiera skattningsmetoden,

Undersökningsplanering

- Speciellt vid relativt små ekonomiska resurser kan det visa sig att 100 % bortfall blir optimalt – vilket svarar mot en välgjord webb-panelstudie.

- Eller mot en vanlig erfarenhet – om man besöker ett land kan man fråga några i omgivningen om politisk inställning. Man lär sig mycket av de första gångerna men knappast så mycket efter de tio första.

- (Vid dåligt gjorda webb-paneler kan inte standardavvikelsen beräknas vare sig i förväg eller efteråt och de kan alltså inte hanteras så här)

Med hjälpvariabler

- Går att utvidga analysen till fallet med hjälpvariabler
- Men komplicerade formler – flera varianser/ kovarianser inblandade. De som har med bortfallet att göra och de som inte har det
- Slutsatsen blir att de vanliga uppskattningarna av precisionen blir nästan alltid för små (osäkerhetsintervallen blir för korta)
- En vägning med t ex regressionsskattning (eller propensity score) tar i stort sett bort det fel som beror på hjälpvariablerna men tillför själv ett fel som kan vara stort genom sk ”spurious correlation”.
- De exempel, som jag räknat på, med denna modell tyder på att man inte bör använda metoden om korrelationerna understiger 0,25.

Tredje svaret: Ingen information

- Ibland vet man ingenting (tror man).
- Men man kan visa att bortfallsfelet är proportionellt mot korrelationen mellan svarsbenägenheten och den studerade variabeln (och lite annat)
- Korrelationen ligger någonstans mellan -1 och 1. Vi vet inte var så vi ansätter en likformig fördelning.
- Det ger inte mycket. Men med en hjälpvariabel så blir det intressant.
- Om korrelationsmatrisen för svarsbenägenhet, hjälpvariabel och studerad variabel är likformigt fördelad kan man beräkna sannolikheten att den vanliga bortfallskorrektionen ger en förbättring

- Man kan alltid skatta korrelationen mellan hjälpvariabel och bortfallssannolikhet.

- Korrelationen mellan studerad variabel och hjälpvariabel kan hjälpligt skattas med hjälp av de svarande

- Givet dessa två värden kan man se hur bra/dålig bortfallskorrektionen blir (givet att man inte vet något om korrelationen mellan studerad variabel och bortfallssannolikhet.

Sannolikheten att biasen blir mindre efter omvägning med regressionsskattning

som funktion av hjälpvariabelns korrelationer med bortfallssannolikheten resp. med studerad variabel.
Likformig fördelning apriori för korrelationsmatrisen

		0,03	0,27	0,51	0,75	0,99
0,03		0,5	0,5	0,5	0,51	0,55
0.27		0,5	0,52	0,54	0,58	0,99
0,51		0,5	0,54	0,59	0,67	1
0,75		0,51	0,58	0,67	0,82	1
0,99		0,55	0,99	1	1	1

Ganska starka korrelationer krävs för att korrektionen avsevärt påverkar bortfallsbias. (Men slumpfelet kan minska mycket)

Förväntad relativ minskning av absoluta biasen jämfört med redovisad förändring av skattningen

som funktion av hjälpvariabelns korrelation med bortfallssannolikheten resp. med studerad variabel.
Likformig fördelning apriori för korrelationsmatrisen

		0,03	0,27	0,51	0,75	0,99
0,03		0	0	0,01	0,02	0,11
0.27		0	0,04	0,08	0,16	0,75
0,51		0,01	0,08	0,18	0,34	0,88
0,75		0,02	0,16	0,34	0,61	0,88
0,99		0,11	0,75	0,88	0,94	0,99

Ganska starka korrelationer krävs för att bortfallsbias skall ha minskat med mer än hälften av observerad ändring.

Slutsats

- Med bayesianska metoder och en del ganska rimliga antaganden kan man säga en hel del om bortfallets effekter och hur bra bortfallskorrektionen är.
- Man kan jämföra medelfelet vid olika bortfallsstorlekar och välja optimalt. Bra webbpaneler kan ses som nära 100% bortfall.
- Men även annat kan modelleras:

Andra felorsaker som också kan behandlas bayesianskt och inkluderas i analysen

- Variabelval, svarsfel
  - Krångliga och svåra frågor om rätt sak eller enklare och lättbesvarade frågor om något besläktat
  - Tidsfel (t ex val av intermittens eller längd på undersökningsperiod)
  - Värde och omfattning av frågetest
  - Avvägning mellan förändrings- och nivåskattning

- Ramfel
  - T ex begränsad population med hög svarsfrekvens kontra hela populationen med stort bortfall och större mätfel (t ex antal anställda, bara svensktalande). De bortdefinierade tas om hand på annat sätt.

Bayes-metoder för TQM – TSE

Total Quality Management – Total Survey Errors

- Till exempel kan man väga relevans mot svarsfrekvens. Är det bättre att fråga om månadsinkomst, som är lätt att besvara eller om årlig inkomst från alla källor som kanske är mer relevant

- Man får göra lite subjektiva (bayesianska) uppskattningar t ex
  - månadsinkomst får 50 % lägre varians (per år) än totalinkomst
  - men leder till en okänd underskattning på 10 +/– 5 % (per år)
  - och att den partiella bortfallssannolikheten ökar från 0 till 5 %.

- Om man räknar på detta får man att om urvalet är mindre än 783 skall man fråga om månadsinkomst och vid större urval om alla inkomstslag

Bayes-metoder för TQM – TSE

Total Quality Management – Total Survey Errors

- Så kan man räkna på nästan alla designfrågor t ex
  - Questionnaire design
  - Intervjulängd
  - Mode
  - Urvalsstorlek
  - Påminnelser
  - Intervjuarutbildning
  - Val av urvalsram (RDD eller RTB eller…)
  - Presentation

- Det gäller bara att, som vi gjorde ovan, tänka igenom vilka konsekvenserna blir, hur säker man är på dem och försöka värdera dem med samma mått och slutligen räkna igenom.

- En bra undersökare gör detta omedvetet, men det kan vara bra att veta att det går att räkna på det.

Tack för att ni lyssnade!