

## 2 Bortfallsbegrepp och bortfallsmått

### 2.1 Vad är bortfall?

Med bortfall avses generellt avsaknad av värden på en eller flera variabler som ska inhämtas i en undersökning. Ofta tänker man på bortfall bara som svarsbortfall i en undersökning (med individer eller företag som uppgiftslämnare), men det kan t.ex. även vara bortfall av värden avsedda att inhämtas från ett administrativt register. En utförligare beskrivning av bortfall ges i SCB (2001a, s. 8):

*Svarsbortfall*, vanligen förkortat till *bortfall*, uppstår när värden på en eller flera observationsvariabler inte kan inhämtas. Termen skall uppfattas i den vida mening som indikeras av engelskans "missing value". Den omfattar t.ex. att uppgifter saknas därför att de av en eller annan anledning inte kunde överföras från ett administrativt register. Om man för ett objekt inte erhåller några användbara uppgifter föreligger *objektbortfall*. Om användbara uppgifter inhämtats för vissa, men inte alla, observationsvariabler föreligger *partiellt bortfall* (för variabler utan värden).

### 2.2 Vad menas med mål- och rampopulation?

När man redovisar bortfallet i en statistisk undersökning bör man vara införstådd med ett par olika populationsbegrepp och några termer som avsluter till dessa. I undersökningens fokus ligger *målpopulationen*: den grupp av *element* eller *objekt* (exempelvis företag, individer eller händelser) som man är intresserad av. För att möjliggöra informationsinsamling fordras en förteckning eller *ram* som ger tillgång till elementen i målpopulationen. Gruppen av alla ramelement kallas rampopulation. Ramelementen kan, men måste inte, vara direkt kopplade till elementen i målpopulationen. (Låt oss säga att målpopulationen består av alla personer som bor i ett visst bostadsområde. För att välja ut personer drar man slumpmässigt ett antal hus från en förteckning över alla hus i området, och undersöker de personer som bor i de utvalda husen. Här är målpopulationens element personer medan ramelementen är hus.) I denna framställning antar vi för enkelhetens skull att ramelementen har en en-till-en relation (Dalenius, 1986, p. IV-3) med målpopulationens element. Urval från en sådan ram, som alltså direkt identifierar elementen i målpopulationen, kallar vi *direkt elementurval* (Särndal et al. 1992, p. 10).

I en ram som används för direkt elementurval ska rampopulationen och målpopulationen idealt sammanfalla. Det typiska är dock att grupperna i någon mån skiljer sig åt, varvid ramen säges lida av täckningsfel. Speciellt definieras *övertäckningsobjekt* som element som ingår i rampopulationen trots att de inte tillhör målpopulationen, och *undertäckningsobjekt* som element i målpopulationen som saknas i ramen. Se vidare SCB (2001a).

I en *urvals-* eller *stickprovsundersökning* (begreppen betraktas här som synonyma) används insamlade data för att uppskatta olika storheter (såsom medelvärden) för populationen som helhet. Om täckningsfelen är stora kan olika typer av korrigerande åtgärder behöva vidtas för att undersökningsresultaten ska vara giltiga för målpopulationen.

## 2.3 Hur kategoriserar vi stickprovet?

Utgångsläget här är att man slumpmässigt drar ett direkt elementurval från rampopulationen och försöker samla in data för utvalda element. Det kan då hända att man får in svar som avslöjar sina upphovsmän som ej tillhörande undersökningens målpopulation. Hur dessa element, så kallade *svarande övertäckningsobjekt*, ska hanteras är inte självklart. I SCB (2001b, kap. 11) beskrivs metoder för att hantera bortfall och ramproblem när man uppskattar populationsstorheter utifrån stickprovsdata. Där används uppgiften om samtliga svarande (inklusive de svarande i övertäckningen) i variansskattningen. Här fokuserar vi på själva bortfallet och försöker uppskatta hur stort detta är. I detta sammanhang är det lämpligt att exkludera svarande övertäckningsobjekt från stickprovet. I Sektionen för surveystatistik (2005, kap. 5) beskrivs hur kvarvarande element kan delas in i fyra huvudkategorier. Dels en kategori av svarande målpopulationselement,

*S*: Element tillhörande målpopulationen för vilka fullständiga eller partiella svar erhålls,

dels tre kategorier av element för vilka svar ej erhålls:

*B*: Element tillhörande målpopulationen som utgör (objekt)bortfall,

*O*: Element vars målpopulationsstatus är okänd, och

$\emptyset$ : Element som inte tillhör målpopulationen.

Element i bortfallet om vilka det råder osäkerhet om de ingår i målpopulationen eller ej faller i kategori *O*. Någon kategori för undertäckningsobjekt finns inte, eftersom de inte har någon chans att inkluderas i stickprovet.

Antalet element i *S*, *B* och *O* betecknar vi  $n_s$ ,  $n_b$  och  $n_o$ . Med hjälp av dessa byggstenar kan en rad intressanta bortfallsmått beräknas, se avsnitt 2.4.

Den del av bortfallet som tillhör kategori *O* är dock prekär. För att kunna beräkna olika bortfallsmått behöver andelen  $u$  av *O*-elementen som tillhör *B* uppskattas (resterande element antas sedan tillhöra  $\emptyset$ ). I AAPOR (2004, s. 29) finner vi följande allmänna rekommendation:

In estimating  $[u]$ , one must be guided by the best available scientific information on what share eligible cases make up among the unknown cases and one must not select a proportion in order to boost the response rate.

För en genomgång av olika föreslagna metoder för att skatta  $u$ , och referenser, se Smith (2003). Smith konstaterar att alla metoder har sina tillkortakommanden, och att ingen av dem för närvarande kan betraktas som "gold standard". I avsaknad av stark kunskap om  $O$ -elementens fördelning över  $B$  och  $\emptyset$  rekommenderar vi användning av  $u=1$ , som ger en försiktig uppskattning av undersökningens svarsandel.

## 2.4 Vilka bortfallsmått kan användas?

Bortfallsmått används huvudsakligen i två syften: dels för att utvärdera kvaliteten i fältarbetet, dels som indikatorer på kvaliteten i den statistik som redovisas. Måtten kan vara vägda eller ovägda med hänsyn till urvalsdesignen: i den mån elementen i stickprovet har varierande urvalssannolikheter tar vägda mått hänsyn till detta, ovägda gör det inte. I allmänhet används ovägda mått för att utvärdera kvaliteten i fältarbetet. Bortfallsandelen, som vi snart ska återkomma till, används huvudsakligen i detta syfte. I AAPOR (2004, s. 30-32) finner vi andra fältarbetsrelaterade mått såsom andel genomförda intervjuer av totalt antal kontakter ("cooperation rate") och andel vägrare av av totalt antal potentiella uppgiftslämnare ("refusal rate"). När bortfallsredovisningen syftar till att utvärdera kvaliteten i den framtagna statistiken bör i första hand vägda mått användas.

I det följande koncentrerar vi oss på mått som speglar objektbortfalllets storlek, men noterar samtidigt att dessa mått, efter smärre justeringar, kan användas även för att redovisa partiellt bortfall. Måtten är hämtade från kap. 5 i Sektionen för surveystatistik (2005), men de flesta återfinns även i exempelvis AAPOR (2004).

Vi börjar med den enklaste redovisningen: den ovägda svarsandelen

$$(1) SA(\text{ovägt}) = \frac{n_s}{n_s + n_B + u \cdot n_O}$$

och tillhörande bortfallsandel

$$(2) BA(\text{ovägt}) = 1 - SA(\text{ovägt})$$

Vid antagandet att  $u=1$  förenklas formel (1) till

$$SA(\text{ovägt}) = \frac{n_s}{n_s + n_B + n_O}$$

Formlerna (1)-(2) kan användas både för urvals- och totalundersökningar. De kan också lätt modifieras för att speciellt betona i vilken utsträckning man fått in *fullständiga* svar. Låt  $F$  beteckna delmängden av element i  $S$  för vilka fullständiga svar erhållits, och låt antalet element i  $F$  betecknas  $n_F$ . Nu kan vi beräkna modifierade ovägda mått som

$$(3) SA^F(\text{ovägt}) = \frac{n_F}{n_S + n_B + u \cdot n_O}$$

och

$$(4) BA^F(\text{ovägt}) = 1 - SA^F(\text{ovägt}).$$

Vi övergår nu till mått som är *vägda* med hänsyn till urvalsdesignen. Antag att elementen i stickprovet är numrerade  $1, 2, \dots, k, \dots, n$ , och låt designvikten (inversen av urvalssannolikheten) för element  $k$  betecknas  $d_k$ . Summan av designvikterna för de element som tillhör kategori  $S$  betecknas  $\sum_S d_k$ , summan av designvikterna för kategorierna  $B$ ,  $O$  och  $F$  på motsvarande sätt. De vägda motsvarigheterna till formel (1)-(2) ges nu som:

$$(5) SA(\text{vägt}) = \frac{\sum_S d_k}{\sum_S d_k + \sum_B d_k + u \cdot \sum_O d_k}$$

$$(6) BA(\text{vägt}) = 1 - SA(\text{vägt})$$

Även formlerna (5)-(6) kan modifieras för att betona i vilken utsträckning man fått in *fullständiga* svar. Vi får då:

$$(7) SA^F(\text{vägt}) = \frac{\sum_F d_k}{\sum_S d_k + \sum_B d_k + u \cdot \sum_O d_k}$$

$$(8) BA^F(\text{vägt}) = 1 - SA^F(\text{vägt})$$

Utöver vägning med hänsyn till urvalsdesignen kan det ibland bli aktuellt att dessutom *storleksväga* med någon hjälpvariabel  $x$  som har samband med undersökningsvariabeln. I SCB (1980, s. 5:4) exemplifieras detta med en undersökning av utbetalda lönesummor i industrin. Ett relativt stort bortfall av företag kan vara betydelselöst i sammanhanget om det hänförs till övervägande små företag: storleksvägning med antal anställda ökar bortfallsmåttets relevans. Vid storleksvägning modifieras formlerna (5)-(6) enligt följande:

$$(9) SA(\text{vägt}) = \frac{\sum_S d_k x_k}{\sum_S d_k x_k + \sum_B d_k x_k + u \cdot \sum_O d_k x_k}$$

$$(10) BA(\text{vägt}) = 1 - SA(\text{vägt})$$

där  $x_k$  är hjälpvariabelvärdet för element  $k$ .

## 2.5 Hur redovisar vi bortfall?

Både användare och producenter av statistik behöver ha tillgång till information om bortfallet för att kunna tolka resultat, jämföra med annan statistik och vidarebearbeta data. En bra bortfallsredovisning ska således tillgodose både användarnas och statistikproducentens behov av information om bortfallets omfattning och effekter.

SCB:s mall för kvalitetsdeklARATION av statistik (SCB 2001a, s. 14-17) ger anvisningar för hur svarsbortfallet i en undersökning ska redovisas:

Ange kvantitativa och kvalitativa aspekter på bortfallet; bortfallsandel (vägd och/eller ovägd), uppdelning i ”avböjd medverkan” respektive ”ej anträffad”, kända eller förmodade bortfallsmönster, åtgärder för att minska bortfallet m.m. Redovisa bortfallets bedömda konsekvenser för statistikens tillförlitlighet. Beskriv metod(er) för bortfallskompensation i skattningsförfarandet [...].

En bortfallsredovisning bör således omfatta:

- objektbortfallets storlek för olika typer av objekt relaterat till urvalet (ovägt) och populationen (vägt).
- omfattning av partiellt bortfall för viktiga variabler och redovisningsgrupper (se kapitel 3)
- orsaker till bortfall (se kapitel 4)
- åtgärder som vidtagits för att reducera bortfallet
- en bedömning av bortfallets effekter på undersökningsresultaten
- metoder för bortfallskompensation i skattningsförfarandet.

En bortfallsredovisning kan även omfatta bortfallets utveckling över tiden.

Statistikproducenten behöver dessutom få information om bortfallet för olika delprocesser för att kunna utvärdera och förbättra statistikproduktionen (se kapitel 21).

## Referenser

AAPOR (2004). *Standard Definitions. Final Dispositions of Case Codes and Outcome Rates for Surveys*. On-line Edition 3.1, revised February, 2005. The American Association for Public Opinion Research. Ann Arbor, Michigan. Hemsidaadress: [www.aapor.org/pdfs/standarddefs\\_3.1.pdf](http://www.aapor.org/pdfs/standarddefs_3.1.pdf).

Dalenius, T. (1986). *Elements of Survey Sampling*. Stockholm: Sarec, SCB.

SCB (1980). *Räkna med bortfall – En handbok om statistiska metoder i samband med bortfall*.

SCB (2001a). *Kvalitetsbegrepp och riktlinjer för kvalitetsdeklaration av officiell statistik*. MIS 2001:1. Kan beställas från SCB, Publikationstjänsten, 701 89 Örebro, e-post: [publ@scb.se](mailto:publ@scb.se).

SCB (2001b). *Estimation in the presence of Nonresponse and Frame Imperfections*. Kan beställas från SCB, Publikationstjänsten, 701 89 Örebro, e-post: [publ@scb.se](mailto:publ@scb.se).

Särndal, C.-E., Swensson, B. och Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer.

Sektionen för surveystatistik (2005). *Standard för bortfallsberäkning*. Hemsidaadress: [www.bortfallssnuran.se](http://www.bortfallssnuran.se).

Smith, Tom W. (2003). *A Review of Methods to Estimate the Status of Cases with Unknown Eligibility*. Report prepared for the AAPOR Standard Definitions Committee. Presented to AAPOR, Phoenix, May, 2004. Version 1.1. Hemsidaadress: [www.aapor.org/pdfs/erate.pdf](http://www.aapor.org/pdfs/erate.pdf).