

Sluppmässiga stickprov och livslängdsdata

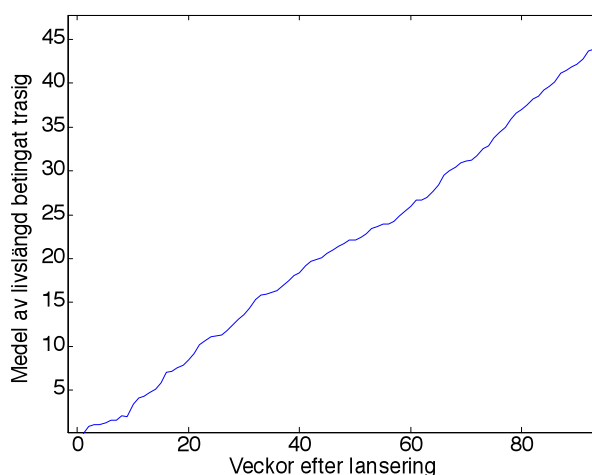
Det där med sluppmässiga stickprov är inte alltid lätt att förstå sig på, vilket följande lilla berättelse belyser.

Företaget Elektro-Toast AB tillverkar brödrostar. En dag kan de stolt lansera en ny version av sin brödrost. Den säljer bra. Första veckan säljs 3000 exemplar, liksom andra veckan. Och så fortsätter det, vecka efter vecka.

Fem veckor efter lanseringen börjar ingenjörerna bli bekymrade. En och annan kund har uppsökt återförsäljaren för att klaga på att brödrosten slutat fungera. Det visar sig vid en analys att man fått brott på fjädermekanismen. Det är vid den tidpunkten totalt sju brödrostar som fått detta problem och de uppvisar livslängderna

$$(t_1, t_2, t_3, t_4, t_5, t_6, t_7) = (1.1, 1.4, 0.6, 2.9, 1.8, 0.2, 0.9) \text{ [veckor]}$$

vilket ger medelvärdet 1.3. Man bokar genast tid hos företagsledningen och redovisar där den katastrofala nyheten att fjädrarna i den nya brödrosten har en livslängd på 1.3 veckor. Det får till följd att den ena problemlösninggruppen efter den andra tillsätts. Man vidtar vissa åtgärder och redovisar återigen för företagsledningen efter 10 veckor på marknaden. Det visar sig då att medellivslängden stigit till 3.4 veckor. Efter 20 veckor är det dags igen. Det enda kvarvarande problemlösningsteamet slår sig för bröstet för nu kan de med en medellivslängd på 8.5 veckor påvisa "continuous improvement". Ledningen följer sedan arbetet vecka för vecka tills förbättringen blivit så omfattande att förbättringsteamet belönas med varsin biobiljett och utmärkelsen "Chairman's Quality Award of the Year".



Medelvärdet av livslängden på de brödrostar som gått sönder där man bortsett från brödrostar som ännu är hela.

Det kanske är lätt för oss som kan statistik att inse att man i just detta fall inte tagit ett stickprov från den avsedda populationen. Men det finns svårare situationer. Jag har själv gjort misstag med just sluppmässiga stickprov och livslängdsdata, om än inte så flagrant som i detta exempel, utan att inse det förrän jag under redovisningen av mina resultat hörde någon i skaran av lyssnare säga till den som satt bredvid att det var något av det mest imbecilla han sett.

I andra sammanhang kan detta med slumpmässigt stickprov *från den avsedda populationen* vara ännu besvärligare. Hur många har inte vid något tillfälle varit intresserade av långtidsvariationen men ändå av misstag tagit ett stickprov från en enda batch?

Datamaterialet ovan är simulerade data från en exponentialfördelning med väntevärde 8000 veckor, det vill säga cirka 150 år. Det är en medellivslängd som skulle vara helt OK för min del som köpare av en brödrost. I figuren har jag ritat ”medelvärdet”, beräknat som i exemplet ovan, som funktion av tiden från simulerade data.

Hur har man då gjort sina beräkningar på Elektro-Toast och hur borde man gjort? Att man räknat ut medelvärdet som

$$\bar{t} = \frac{1}{7} \sum_{i=1}^7 t_i = \frac{1.1 + 1.4 + 0.6 + 2.9 + 1.8 + 0.2 + 0.9}{7} = 1.4 \text{ [veckor]}$$

kanske är tydligt utifrån exemplet. Men man borde dessutom ha tagit hänsyn till de

$$5 \times 3000 - 7 = 14993 \text{ brödrostar}$$

som ännu inte gått sönder. Dessa observationer sägs vara censurerade. Antag exempelvis att en brödrost fortfarande inte gått sönder efter 4 veckor. Det enda vi vet om dess livslängd är då att den ligger i intervallet $[4 ; \infty]$ veckor. Det finns visserligen inte så mycket information i detta, men i alla fall något, och för att få en rättvisande bild måste man ta hänsyn därtill. Det gör man enklast genom att använda maximum-likelihoodmetoden för att skatta parametrar. Vi kommer inte att redogöra för hur, bara konstatera att de flesta statistiska programpaket (exempelvis Matlab statistiktoolbox och Minitab) kan hantera censurerade livslängdsdata med maximum-likelihoodmetoden.

Magnus Arnér