Stickprovsparadoxen – partistorlekens ringa inverkan på resultatet

När man diskuterar ett stickprovsförfarande på ett företag blir det alltid tal om partistorlekar och att '...självklart måste man ta hänsyn till dessa då man tar sitt prov...'. Visst verkar det naturligt att öka stickprovets storlek då man har större partier.

Det visar sig dock, kanske paradoxalt, att partistorleken spelar ingen, eller mycket ringa, roll vid praktiskt stickprovstagning. Detta faktum kallas ibland på engelska för 'The Sampling Paradox' och nedan reder vi ut en del fakta med hjälp av statistikteori.

1. Den hypergeometriska fördelningen

Antag att vi har ett parti om N stycken detaljer och att det finns M stycken felaktiga detaljer i partiet. Antag dessutom att vi tar ett slumpmässigt stickprov om n detaljer och avsynar dessa. Vi är nu intresserade av sannolikheten att 0, 1, 2, 3, ... , n felaktiga detaljer i stickprovet.

Den sannolikhetsfördelning som vi då har går under det kryptiska namnet *hypergeometrisk fördelning*. (Söker man på namnet i litteraturen hittar man vaga förklaringar som man inte blir klokare av.) Nedan sammanfattar vi några av fördelningens egenskaper men vi visar inte uttrycken för de enskilda sannolikheterna (finns lätt tillängligt i litteraturen):

$$\mu = \frac{n \cdot M}{N}$$

$$\sigma = \sqrt{\frac{n \cdot M}{N} \cdot \left(1 - \frac{M}{N}\right) \cdot \frac{N - n}{N - 1}}$$

- μ är det förväntade antal felaktiga i ett stickprov om n detaljer.
- σ är standardavvikelsen hos antal felaktiga i stickproven, dvs variationen mellan stickproven.

2. Inspektion av uttrycket för standardavvikelse

Eftersom det finns M felaktiga bland totalt N produkter i partiet kan vi kalla kvoten för en ’felkvot’ och beteckna den p. Vi skriver om formlerna med hjälp av detta:

$$p = \frac{M}{N}$$

$$\mu = n \cdot p$$

$$\sigma = \sqrt{n \cdot p \cdot (1 - p) \cdot \underbrace{\frac{N - n}{N - 1}}_{a}}$$

Med hjälp av p skrivs formlerna om så att det blir tydligare att resonera om dess egenskaper.

Faktorn a i uttrycket för standardavvikelsen är speciellt intressant. När stickprovets storlek n blir mindre i förhållande till N blir faktor a allt närmare 1. När man dessutom drar kvadratroten ur uttrycket för sigma blir betydelsen av faktor a ännu mindre.

Vi kan också vända på resonemanget. Om stickprovets storlek n ökas och till sist blir lika med N då blir faktor a noll och hela uttrycket noll, dvs ingen variation. Inte så konstigt, vi gör ju då en totalkontroll och då finns det inget utrymme för statistik osäkerhet.

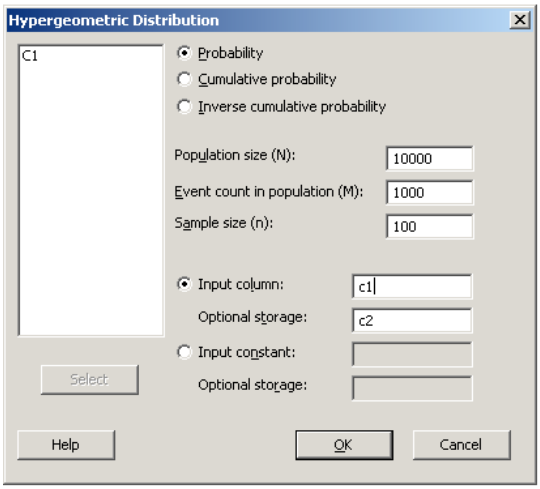
3. Beräkning av sannolikheter

Här gör vi en jämförelse mellan två olika partier, A och B, som har exakt samma felkvot, 0.1 (10%):

	Parti A	**Parti B**
N (partistorlek)	10 000	50 000
M (antal felaktiga)	1 000	5 000
n (stickprovstorlek)	100	100

Bägge partierna, A och B, har samma felkvot, 10%, men parti B är fem gånger större än parti A. Med hjälp av Minitab och menyn [Calc]>[Probability Distributions]>[Hypergeometric...] beräknar vi sannolikheten att få 0, 1, 2, 3… osv felaktiga i stickprovet. Först skapar vi en kolumn med talen 0, 1, 2, 3, ... Sedan beräknar vi sannolikheterna.

```
set c1              # Lagrar 0 1 2...25 (25 tycks vara en lagom övre gräns).
0:25
end
```

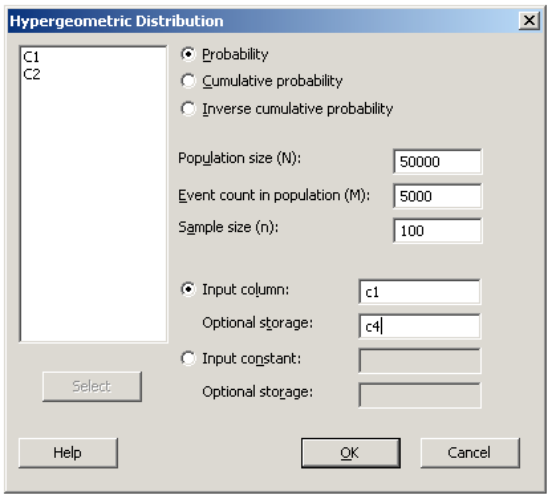


Parti A

Beräknar sannolikheten att få 0, 1, 2, 3, … i ett stickprov om 100 enheter dragna från parti A (N = 10000) med felkvoten 10%, partiet innehåller 1 000 felaktiga detaljer.

Resultatet lagras i kolumn c2 så att vi senare kan visa resultatet i diagramform.

Vi använder nu samma meny och gör beräkningarna för parti B:



Parti B

Beräknar sannolikheten att få 0, 1, 2, 3, … i ett stickprov om 100 enheter dragna från parti B (N = 50 000) med felkvoten 10%, dvs partiet innehåller 5 000 felaktiga detaljer.

Resultatet lagras i kolumn c4 så att vi senare kan visa resultatet i diagramform.

De två resultaten listas ned och de tycks vara väldigt lika (längre ned presenteras resultaten i diagamform):

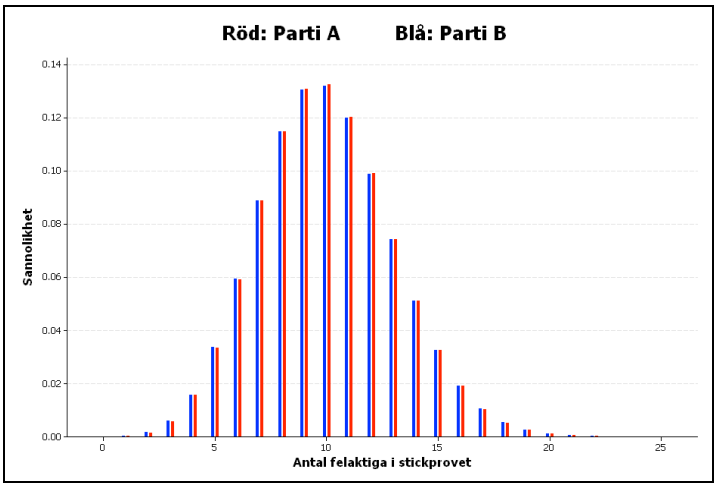
Sann parti A

```
0.000025 0.000282 0.001568 0.005743 0.015594 0.033488 0.059238 0.088774 0.115040 0.130941
0.132530 0.120468 0.099152 0.074400 0.051193 0.032462 0.019052 0.010388 0.005280 0.002509
0.001118 0.000468 0.000184 0.000068 0.000024 0.000008
```

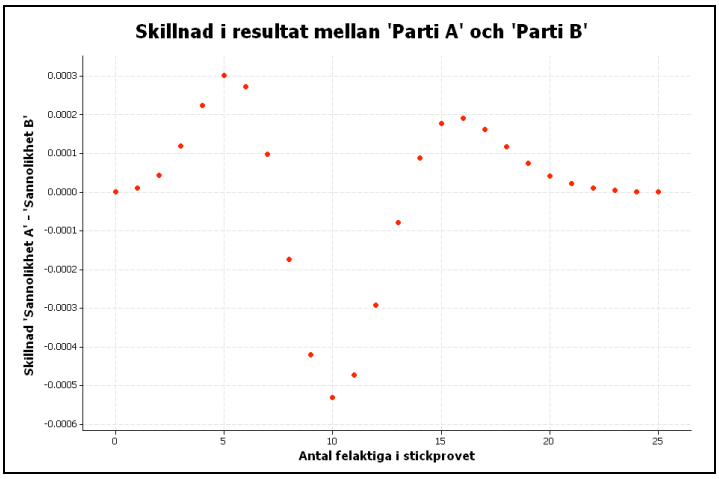
Sann parti B

```
0.000026 0.000293 0.001612 0.005862 0.015819 0.033790 0.059511 0.088871 0.114866 0.130521
0.131997 0.119995 0.098861 0.074322 0.051282 0.032639 0.019244 0.010551 0.005397 0.002583
0.001160 0.000490 0.000195 0.000073 0.000026 0.000009
```

4. Resultat i diagramform



Om man ritar de två sannolikhets-fördelningarna i samma diagram ser man en stor likhet.



Ett vanligt knep är ju att beräkna skillnaden och rita denna i ett diagram.

Man ser att skillnaden är för-hållandevis liten, från -0.0005 till 0.0003, dvs på fjärde decimalen.

5. Sammanfattning

Detta dokument visar ett partistorleken *N* har ringa eller inte någon betydelse för den statistiska osäkerheten eller resultatet i praktisk stickprovtagning. ■