

### *Tankar om urvalsfel och bortfallsfel vid webbenkäter – idéer till en totalmodell*

Daniel Thorburn, Stockholms universitet  
Måns Magnusson, Linköpings universitet

Surveyföreningens konferens  
**Nya möjligheter med webbpaneler**  
Stockholm, 18 februari 2016

Projektet är bara i sin linda ännu. En utgångspunkt är:  
**Måns Magnusson**  
Webb-paneler, kvalitet och sannolikhetsurval  
Quintensen 2-3, 2015, sid 23-24

Men vi har även några andra idéer som kommit fram när jag och Måns diskuterat. Vi började strax innan jul, så vi vet ännu inte riktigt var vi hamnar.

Några citat från Måns artikel:

- "Osäkerhetsintervall är det självklara sättet vi alltid bör presentera statistisk osäkerhet på"
- "Istället för att lägga vår kraft på om de få som deltar i Webb-paneler är insamlade med sannolikhetsurval eller inte bör vi lägga fokus på de bortfallskompensatoriska modellerna och den teori som redan finns och skapa metoder för att presentera osäkerhet från självselektade urval"
- "Särskilt den Bayesianska metodiken har här fördelar som kan utnyttjas då vi kan utnyttja metodologisk kunskap och resultat som stöd i våra skattningar"

### En totalmodell

- Tre olika felorsaker som alla skall modelleras och kopplas samman
  - Studerade variabelers (Y) variation (dvs urvalsfel)
  - Selektion/bortfallsfel (dvs samband Y & bortfall)
  - Y och bortfallets samband med hjälpvariabler – förklaringsgrader!
- Jämför epidemiologi!
- Bayes-modeller är mycket användbara – i detta fall eftersom de går att koppla samman
- Många delar i en modell; några exempel:

### Exempel: Bortfallsfel och urvalsfel

- Vid urvalsfel är man van vid att beskriva felet med medelfel eller varians.
  - Det handlar om det förväntade felet.
  - Det verkliga felet kan av en slump vara litet även om medelfelet är stort och vice versa
- På samma sätt skall vi behandla selektionsfelet
  - Dvs som en okänd variabel med osäkerhetsgränser
  - Om man sedan gör en uppföljningsundersökning kan det visa sig vara litet även när vi varit mycket osäkra eller stort när vi varit relativt säkra.

### Exempel: Förklaringsgrad

- Klassisk regression
  - Slumpfel efter regressionen =  $\sigma_v^2(1-R)$ , där R är förklaringsgrad
- Selektionsbias
  - (Här blir förklaringen ett lite krångligare uttryck)
  - Före korrektion: bortfallsbias =  $\text{Cov}(y, r)/(1-b)$  där r står för svarssannolikhet och b för totalbortfall
  - Efter korrektion: bortfallsbias =  $(\text{Cov}(y, r) - \text{cov}(y, x)\text{cov}(x, r)/\text{var}(x)) / (1-b)$  (exakt uttryck vid kalibrering med regressionskattning)

### Några ytterligare exempel på modelldelar finns nedan

Men jag lovade fatta mig kort, så

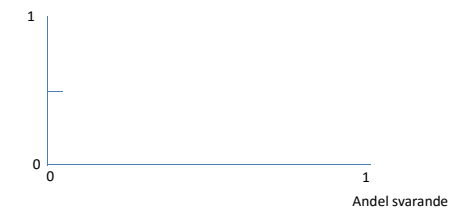
**Tack för intresset**

Jag hoppas Måns eller jag får återkomma vid senare tillfälle

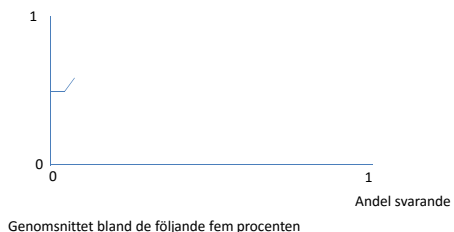
### Exempel: samband olika bortfallsnivåer och web-paneler

- En av våra modeller för att koppla samman olika bortfallsnivåer och web-paneler

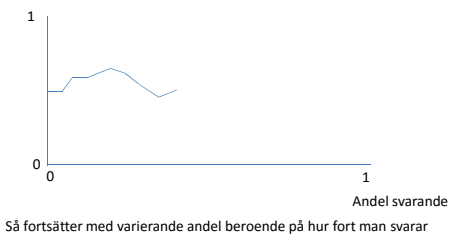
Ex: Tänkbar kurva över andelen rödgröna efter hur fort man svarar



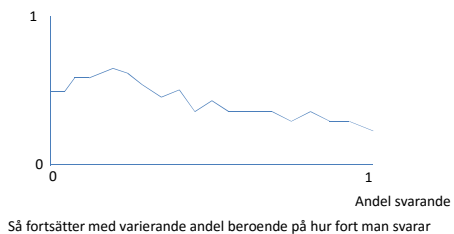
Ex: Tänkbar kurva över andelen rödgröna efter hur fort man svarar



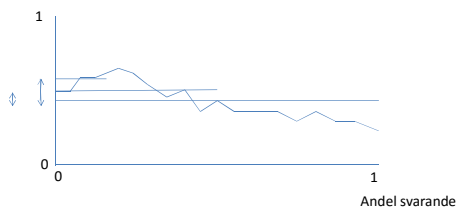
Ex: Tänkbar kurva över andelen rödgröna efter hur fort man svarar



Ex: Tänkbar kurva över andelen rödgröna efter hur fort man svarar

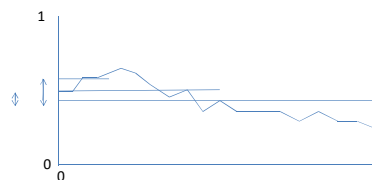


### Ex: Tänkbar kurva över andelen rödgröna efter hur fort man svarar



Genomsnitt om svarsfrekvensen är 20 %, 50 % och 100 % dvs inget bortfall.  
Skillnaden beskriver bortfallsfelet

### Ex: Tänkbar kurva över andelen rödgröna efter hur fort man svarar



Om man inte vågar tro mycket om kurvan mer än att den är kontinuerlig kan man ansätta en Wienermodell. Då blir variansen för proportionell mot bortfallet (här 80 resp 50)

Proportionalitetskonstanten beror på frågan, hur omsorgsfull man är vid insamlingen och hur bra förklaringsvariablerna förklarar skevheten  
Web-paneler svarar mot en kumulativ svarsfrekvens som är nästan 0.