Statistical Paradises and Paradoxes in Big Data

Tankar om Xiao-Li Mengs artikel

Dan Hedlin

Statistiska institutionen,

Stockholms universitet

Ur Qvintensen 2010/2

Kärleksmodulen – hur kan man hitta lämpliga par?

Om att reformera statistikutbildningen

"Happy course" med chokladprovning, champagne och studenternas egna filminspelningar om diverse statistiska fenomen. Xiao-Li Meng berättade om nya grepp inom statistikutbildningen på Harvard.

i kursen, med champagne i. Men det behövdes också kärlek och pengar. I kärleksmodulen funderar studenterna över hur man kan hitta lämpliga par, som datingsajter försöker göra, och hur man kan formulera frågor om personlighet och önskemål om partner.

"Happy course" har funnits i fem år. Numera består den av fem moduler som illustrerar statistiska problem och som tilltalar studenterna; ekonomi, kärlek, medicin, juridik/valundersökningar och choklad. Kursen ska fylla gapet mellan introduktionskurser och kurser på högre nivåer, tanken är att ge djup och känsla

Att undervisa

Nödvändigheten att snabbt få in nya lärare på kurser har tvingat fram nytänkande. Sedan tidigare fanns en central resurs för lärarnas kompetensutveckling. Derek Bok Center for Teaching and Learning. Där finns bland annat skådespelare som lär ut kroppsspråk och röstteknik, grupper som lär ut tentarättning med liten variation mellan lärare och mycket annat. På institutionen kompletterar man det med små diskussionsgrupper där erfarna lärare ställer den typ av frågor som studenter ställer. Vet du vilken fråga från studenter som är svårast ge en bra

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. The Annals of Applied Statistics.

”För att kunna dra statistiskt säkra slutsatser till en population från ett stickprov, krävs att man har dragit ett slumpmässigt urval utifrån en aktuell ram över de som ingår i mål-populationen, och som är möjliga att nå.” (Dahmström 2011, p. 88)

Vad gäller för urval och bortfall?

- **Ignorerbar** urvals- och svarsmekanism: Inget samband mellan urvals- och svarssannolikheter och det man undersöker

  (Little 1982, Smith 1983)
- Gäller vid slumpmässigt urval och slumpmässigt bortfall men även vid vissa icke-slumpmässiga urval

- Men även om det finns “samband”, kan man justera för det så är det ok
- **Poststratifiering**, “vägning”
- SCB använder **kalibrering** (Särndal och Lundström 2005)

Felet = Differens mellan skattat medelvärde och sant medelvärde

Felet är en produkt av

1. **Datakvalitet**
2. **Datakvantitet**
3. **Problemets svårighetsgrad**

(Meng 2018)

Register / Big data

- **Felet, produkt (multiplikation) av:**

1. Korrelation mellan selektion och undersökningsvariabel
2. Roten ur ”drop-out odds”: andel ej undersökta genom andel undersökta
3. Undersökningsvariabelns spridning (standardavvikelse)

Bias orsakad av bortfall i under-sökningar med slumpmässigt urval

- **Relativ bias = (Väntevärde av skattning – sant värde) dividerat med sant värde**

Produkt:

1. Korrelation mellan selektion och undersökningsvariabel (ρ)
2. Svarssannolikheternas spridning (cv)
3. Undersökningsvariabelns spridning (cv)

(Bethlehem 1988)

Vad är bäst?

1. Slumpmässigt urval av 1% av populationen men 40% bortfall

2. Register som täcker 80% av populationen

Exempel från Meng (2018)

Beror på tre faktorer

1. Kvoten av två *drop-out odds:*
   - Urvalsundersökning $(1 - f_s)/f_s$, $f_s = \frac{n}{N}$
   - Register f_r
   - Oddskvoten: urvalsundersökningens drop-out odds dividerad med registrets drop-out odds, *OK*
2. Registrets korrelation, ρ_r
3. Urvalsundersökningens korrelation, ρ_s

Villkor för att register ska vara bäst

- $|\rho_r| \leq |\rho_s| \cdot \sqrt{OK}$ (Meng 2018)

Mengs exempel:

1. Slumpmässigt urval av 1% av populationen men 40% bortfall
2. Register som täcker 80% av populationen

- Roten ur *OK* är ungefär 26
- Ganska säkert att registret är bäst

Ett till exempel, population 7.5 miljoner

1. Slumpmässigt urval av 0.1% av populationen men 60% bortfall, 3000 svarande
2. Panel med 100 000 personer, alla svarar

- Roten ur *OK* är ungefär 6
- Svårt att säga vad som är bäst

Designeffekt =

den varians man får med vald metod

Dividerad med

den varians man hade fått med OSU

"Lack-of-design effect"

- MSE (big data) / MSE (OSU utan bortfall) =
  Populationsstorleken gånger 'data defect index'

- *Data defect index* är väntevärdet av kvadraten av korrelationen för big data
- **Felet i skattningen blir större ju större population!**
- **'Return of the long-forgotten monster *N'***

(Meng 2018, s. 698)

"Law of large populations"

- Felet / roten ur variansen under OSU
  =
  Roten ur populationsstorleken gånger korrelationen för big data

- **"The bigger the data, the surer we fool ourselves"**

(Meng 2018, s. 702)

Antag:

1. Big data som täcker halva populationen
2. Korrelationen för big data är 0.05

- ***Kan inte*** bli bättre än ett slumpmässigt urval på 400 personer, utan bortfall

(Meng 2018)

Men man kan ju väga data…

- Den gamla metoden poststratifiering fungerar generellt bra
- Men man kommer aldrig att reducera bortfallsfelet eller big data-felet till noll
- För det krävs att man har all information som förklarar selektionsmekanismen, och det har man nu inte.

- "… what is big about Big Data is the number of intellectually and technologically challenging problems that keep many of us sleepless either because we are too excited or too frustrated."

(Meng 2018, s. 722)

References

- Bethlehem, J. (1988). Reduction of nonresponse bias through regression estimation. Journal of Official Statistics, 4(3), 51-60.
- Dahmström, K. (2011). Från datainsamling till rapport: att göra en statistisk undersökning. 5th ed. Lund: Studentlitteratur.
- Little, R. J.A. (1982). Models for Nonresponse in Sample Surveys. Journal of the American Statistical Association, 77, 237-250.
- Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. The Annals of Applied Statistics.
- Smith, T.M.F. (1983). On the validity of inferences from non-random sample. Journal of the Royal Statistical Society, Series A, 146, 394-403.
- Särndal, C.-E. och Lundström, S. (2012). Estimation in Surveys with Nonresponse. New York: Wiley.