Fakta om metoden multipel imputering

Der er tre grunde til, at man bruger imputering i PISA – og i stikprøveundersøgelser generelt:

1. Det giver mulighed for at stille mange testspørgsmål og dermed dække flere områder af elevernes færdigheder og dermed øge målingssikkerheden.
2. Det reducerer den statistiske stikprøveusikkerhed og øger dermed den statistiske inferens, hvilket vil sige, at man fx kan afgøre, om de opnåede resultater for forskellige grupper af elever er forskellige.
3. Det øger brugernes muligheder for at lave analyser og tilgå data.

Internationalt har imputering været brugt de sidste 50 år i moderne statistik og surveyforskning. Vi skal helt tilbage til 1977, hvor professor Rubin fra Harvard Universitet kom med centrale bidrag til dette forskningsområde med EM algoritmen. Senere har professor Rubin sammen med en række medforfattere udgivet adskillige artikler og bøger om emnet "non-response" multipel imputering. Multipel imputering er altså en "gammel" og anerkendt metode, som i dag benyttes i langt de fleste stikprøveundersøgelser over hele verden, hvor de førende institutioner ligger i USA, Canada og Australien.

I Danmark foregår forskningen i surveymetoder, dvs. sampling, stikprøveteori, imputering, vægtning, non-response bl.a. på CBS, hvor man arbejder med imputering, missing data og EM algoritmen. Der er også forskningsmiljøer inden for dette område på bl.a. Økonomisk institut, København Universitet og på Danmarks Statistik. Multipel imputering er i dag indført som standardværktøjer i de anerkendte og benyttede statistikprogrammer, fx SPSS, SAS, R eller STATA, som benyttes verden over i kvantitative analyser.

Hvad er ideen i multipel imputering?

Et eksempel på, hvornår mutipel imputering er nyttig, er, når en respondent i en spørgeskemaundersøgelse har sprunget besvarelsen af et spørgsmål, fx indkomst, over. Man kan, ud fra de andre svar denne person har givet om fx køn, uddannelse, arbejdstid, branche, anciennitet og stilling, give et godt skøn for, om personen har en indkomst over eller under gennemsnitsindkomsten. Man har i praksis tre muligheder:

1. Man kan smide besvarelsen væk, men det svarer til, man lader de andre besvarelsers gennemsnit bestemme gennemsnittet for den gruppe, som personen tilhører, og dermed mister man information.
2. Man kan beholde besvarelsen men udelade respondenten, når der laves analyser om indkomst. Men det har med hensyn til indkomst samme betydning som den første mulighed – man benytter ikke al information.
3. Man kan ud fra andre respondenters (med det samme køn, uddannelse, arbejdstid, branche, anciennitet og stilling) besvarelser give et skøn (ud fra imputering) over indkomsten. Er der fx 20 besvarelser fra andre respondenter med samme køn, uddannelse, arbejdstid, branche, anciennitet og stilling, kan man tage gennemsnitsindkomsten for disse 20 og lægge dette gennemsnit ind som svaret for den person, der ikke har svaret. Denne metode er fin til at beregne summer og gennemsnit med, men kan ikke bruges til analyser af, hvor stor en andel, der fx har under 200.000 kr. i årsindtægt. Man kan også vælge et af de 20 andre respondenters svar tilfældigt og sætte det i stedet for det udeladte svar. Det vil i det store perspektiv, når man analyserer hele datamaterialet, i gennemsnit give det rigtige svar. Men ved denne løsning har man ikke brugt muligheden for imputering fuldt ud, og man kan ikke lave statistisk inferens,

fordi man ikke kan medregne usikkerhed fra imputeringen. Her kommer de gentagende imputeringer, multipel imputering, som bl.a. professor Rubin har forsket i, ind som en mulighed. Hvis man fx fem gange valgte en af de 20 andre besvarelser tilfældigt, analyserede dem hver for sig, og bagefter lagde resultaterne sammen, så kan man reducere usikkerheden. Dette øger altså sikkerheden af det nye erstattede svar (imputeringen) og giver samtidig mulighed for at lave statistiske analyser, fordi sikkerheden af imputeringen kan beregnes og dermed medregnes.

Principperne i eksemplet ovenfor kan udbygges til at gælde generelt for, hvor der er manglende svar. I PISA, hvor eleverne har svaret på forskellige delopgaver, men ikke dem alle, kan man altså ligeledes imputere svarene fra de, der har svaret. Man imputerer blandt de elever, der minder mest om dem, der mangler at svare, og overtager altså deres svar. Det ændrer ikke på, at statistikken er repræsentativ. Man deler bare svarene på flere. Det kan selvfølgelig ikke bruges til at bestemme præcist, hvad den enkelte konkrete elev ville have svaret. Man kan aldrig sige noget om den enkelte elevs præstationer. Men det kan bruges til at lave statistiske analyser – altså til at sige noget om gennemsnit for forskellige elevgrupper og til at bestemme, om der er forskelle i præstationerne mellem disse grupper.

Ved at benytte multipel imputering udnyttes al information fra datasættet på optimal vis, således at elevernes færdigheder inden for flere fagområder – og flere dimensioner inden for disse fagområder – dækkes. Desuden bliver den statistiske stikprøve-usikkerhed betydeligt reduceret, når multipel imputering benyttes, fordi man med imputering udnytter graden af forklaret variation med hensyn til de variabler, der imputeres efter, svarende til den reduktion, man opnår i en tilsvarende flerdimensionel statistisk analyse. Og endelig, så kan forskeren bruge hele datasættet uden at skulle korrigere for spørgsmål, der er sprunget over.

Yderligere information:
Peter Linde, kontorchef i Danmarks Statistik
E-mail: pli@dst.dk