Kvalitetsmåling i statistik

Peter Linde. Survey and Methods. Statistics Denmark. pli@dst.dk

Opfyldelse af brugernes behov

Over- eller undervurderer statistikken den faktiske udvikling, og hvor sikre er tallene? Det er det, som brugerne vil vide. Brugerne ønsker en kort og præcis vurdering og ikke lange tekniske beskrivelser som eksperter måske kan få noget ud af. Hvis vi ikke som statistikere tør svare på brugernes spørgsmål, hvem skal så? Og hvis vi ikke kan svare, kan brugerne så overhovedet anvende statistikken til noget?

Statistikere forsøger at måle befolkningens og virksomhedernes adfærd, holdninger og økonomisk aktivitet. Det er knap så almindeligt at måle, hvor retvisende den statistik vi udgiver er. Diskussionen om kvalitetsmåling af statistik ender nogle gange med den konklusion, at det er umuligt at måle kvaliteten, og andre gange med meget simple indikatorer, der kun viser en lille del af sandheden. Dokumentation bliver som en konsekvens heraf, enten en meget tekniske beskrivelse af den statistiske usikkerhed og processerne, eller meget generel. Denne artikel kommer med et forslag til en mere pragmatisk måde at måle kvaliteten på, der direkte tager stilling til fortolkningen af outputtet.

Ambitionen er at komme hele paletten rundt og sætte fingeren på de steder, vi ved det gør ondt. Og ikke kun begrænse os til de simple indikatorer, der kan måles med et tal som fx non-response, og som ofte kun viser toppen af isbjerget. Ingen statistik er bedre end det svageste led, og de fleste os, der har lavet rigtigt mange fejl og arbejdet i "maskinrummet", ved også, hvor vi skal kigge: Fx anvendeligheden af spørgeskemaet, spørgsmål der er svære at svare retvisende på (der har høj byrde), afgræsningen af populationen, bortfald og uoplyste svar, samt skæv udvælgelse, ikke systematisk fejlsøgning eller imputering. Erfaringen viser, at de systematiske fejl normalt er langt større end den eventuelle tilfældige stikprøvefejl. Da det kun er den sidste fejltype, man kan beregne konkret, er der brug for nye måder at kvantificere kvaliteten på.

De fleste kvalitetsmål, der foreslås i denne artikel, kan bruges både til registre- og stikprøveundersøgelser, da der i begge tilfælde er tale om en dataindsamling med de samme målingsproblemer (på nær stikprøveusikkerheden). Kun en mindre del af kvalitetsmålene retter sig til den specifikke indsamlingsmetode.

Bløde og hårde kvalitetsmål

Kvalitetsmål skal afspejle, hvor sikkert en statistik beskriver virkeligheden. Både kvantitative og kvalitative metoder kan være nyttige hertil. Det kræver så, at man bagefter kan sammenligne og generalisere ud fra forskellige metoder – dvs. har en fælles skala eller referenceramme. Men først vil jeg se på de tre metoder, der i denne artikel bruges til at beskrive kvalitet med:

1. **Numeriske indikatorer**, der med et tal beskriver kvaliteten af outputtet, fx stikprøveusikkerheden, bortfaldet eller andelen af oplyste.
2. **Standarder** for produktionen (som fx ISO9001), der beskriver krav som skal opfyldes i produktionsprocessen og hvordan man kontrollerer og tjekker dette.
3. **Analyser** - både kvalitative og kvantitative, fx analyser af bortfaldet, resultatet af test af spørgeskemaet eller beskrivelsen af en metode, hvordan man indsamler priser til en 'basiskurv' eller produktionsværdien i Nationalregnskabet.

Alle kvalitetsmål har fire niveauer

For at kunne sammenligne kvalitetsmålene foreslås de opdelt i fire niveauer på baggrund af en på forhånd fastlagt faglig standard med fokus på deres betydning for outputtet. Denne standardisering er udtryk for, hvad vi som statistikproducent mener betyder noget for outputtet, men også hvad der er muligt at opnå, og hvad vi som statistikproducenter selv sætter som mål for vores statistik. Dette bliver selvfølgelig en delvist subjektiv faglig vurdering. Pointen er, at vi tager stilling på forhånd ud fra vores bedste erfaring og bagefter måler mod denne standard. Fx at mindre end 1% uoplyste værdier svarer til *meget høj kvalitet* og mere end 3% uoplyste værdier svarer til *usikker kvalitet*. Eller at et spørgeskema, der testes hvert år med kvalitative metoder og brugervurderinger, har en *høj kvalitet*, og hvis det aldrig er blevet testet er der tale om en usikker *kvalitet* (fordi man jo ikke kan sige noget når man ikke har undersøgt det).

Konkret forslås en opdeling i fire grader:

A - Meget høj kvalitet

B - Høj kvalitet

C - Rimelig kvalitet

D - Usikker kvalitet

Senere i denne artikel beskrives, hvordan analyser af fx betydningen af andelen af uoplyste værdier kan ændre ovenstående kvalitetsvurdering. Det er den tællingsansvarlige, der skal fastlægge, hvilken af de fire grader af kvalitet, der er tale om – ud fra de skriftlige retningslinier. De to højeste kvalitetsniveauer A og B kræver altid uddybet dokumentation, der er tilgængelig skriftligt for andre internt og eksternt. Man kan fx ikke bare selv vurdere spørgeskemaet til at være af *høj kvalitet*, hvis man ikke kan dokumentere det overfor andre. Så hvis vi vender tilbage til eksemplet med kvalitative test af spørgeskemaet, kræver det at der findes en dokumentation for den årlige kvalitativ test, der beskriver hvor mange test man har lavet og hvad resultatet var. Hvis man ikke kan fremlægge en sådan dokumentation, kan man ikke gøre krav på at spørgeskemaet har en *høj kvalitet*.

Man kan mod denne tilgang sige, at det er fair nok, at man skal kunne dokumentere sine kvalitative test af spørgeskemaet overfor andre for at testen har værdi, men at det omvendt ikke er en garanti for, at skemaet så er bedre end et skema, der ikke er testet. Det er selvfølgeligt muligt. Ligesom en statistik aldrig er 100% sikker, kan et mål for kvaliteten heller ikke være 100% sikker. Men hvis flere tællinger har dokumentation for deres kvalitative test af spørgeskemaet, der er tilgængelig for brugerne, vil det danne baggrund for en fælles faglig standard, og dermed kunne højne kvaliteten af alle spørgeskemaer.

Stærke og svage numeriske indikatorer

Eurostat bruger en række simple numeriske indikatorer som standard i de nationale statistikbureauers kvalitetsrapporter. Med ordet indikatorer antydes, at de ikke er udtryk for et sikkert kvalitetsmål. Og med flertalsformen, at der er mere end én måde at måle kvalitet på. Derfor er indikatorer alligevel vigtig, men en indikator kan omvendt ikke stå alene, da den netop kun er en indikator. *Ikke alt der kan tælles tæller og ikke alt der tæller kan tælles,* som Einstein sagde engang. Det har også delvis gyldighed, når man vil beskrive en statistiks kvalitet – noget af det der bruges mest tid på i statistikproduktionen, fx fejlsøgningen, er også det, der er sværest at beskrive kvaliteten af med en numerisk indikator. I denne artikel arbejdes med tre niveauer af numeriske kvalitetsindikatorer:

1. **Universelle indikatorer**, fx spredningen eller varianskoefficienten i en stikprøve. Målet kan sammenlignes over tid, mellem statistikker og lande.

2. **Stærke indikatorer**, fx bortfaldet i en undersøgelse eller andelen af uoplyste eller manglende fletning af populationer i en registerstatistik. Disse mål giver det mening at sammenligne over tid for en konkret tælling og mellem tællinger af samme type, fx personstikprøver i Danmarks Statistik. Men man kan ikke nødvendigvist sige, at Arbejdskraftundersøgelsen (AKU) med 25% bortfald i Sverige er bedre end den samme undersøgelse med 45% bortfald fra Danmark. Det afhænger af betydningen af bortfaldet – om det er skævt eller tilfældigt. Men et større bortfald er en stærk indikator for en kvalitetsforskel, men skal følges op af en kvalitativ eller kvantitativ analyse.
3. **Svage indikatorer**, fx andel af fejl i en fejlsøgning. Denne indikator kan kun sammenlignes over tid inden for en tælling, da den afhænger af, hvor mange ressourcer man har brugt til fejlsøgningen og hvor gode ens metoder er til at finde fejl, og ikke måler hvor mange fejl der er tilbage.

Ikke mindst svage indikatorer kan med fordel kombineres med mål for opfyldelsen af standarder, kvalitative mål og analyser.

Grundlægge kvalitet og endelig kvalitet

Analyser af problemer, der kan øge forståelsen af statistikkens styrker og svagheder, øger også statistikkens anvendelighed for brugeren. Lad os fx vender tilbage til eksemplet med bortfald for AKU i Danmark og Sverige. Pga. af det højere bortfald i Danmark kunne det være rimeligt at tildele AKU i Danmark den **grundlæggende** kvalitetsvurdering D: *Usikker kvalitet (*pga. 45% bortfald) og i Sverige C: *Rimelig kvalitet* (pga. 25% bortfald)*.* Det er muligt at reparere for bortfaldet ved at kalibrere skævheden i bortfaldet op mod kendte registeroplysninger, fx om indkomst, historisk beskæftigelse, alder eller uddannelse. Hvis man har gennemført en sådan bortfaldsanalyse og efterfølgende kalibrering (efterstratifikation) af bortfaldet, er det udtryk for en forbedret kvalitet og anvendelighed af undersøgelsen for brugeren. Det skal derfor også afspejles i den **endelige** kvalitetsvurdering. Meget af kvalitetsarbejdet i statistikbureauerne er netop denne form for reparation. Hvis dette arbejde ikke giver ny værdi og kvalitet, er der jo ingen grund til at lave det – og tilsvarende skal det så også indgå i den endelige kvalitetsvurdering. Fx vil efterstratifikation efter uddannelse, indkomst og nationalitet betyde, at estimationen af hvor mange, der har adgang til Internettet hjemmefra ikke bliver overestimeret. Det vil omvendt være tilfældet, hvis man ikke efterstratificere, da bortfaldet mht. uddannelse, indkomst og nationalitet er højst i de undergrupper, der har mindst adgang til Internettet hjemmefra.

En efterstratifikation, der reducerer betydningen af bortfaldet, vil normalt medføre en forøgelse af den tilfældige stikprøvefejl (CV) pga. vægtenes variation. Det er derfor afgørende, at en efterstratifikation også tæller for det positive den betyder for den samlede usikkerhed. For hvis det eneste en efterstratifikation tæller for er en større stikprøvefejl, fremtræder efterstratifikation jo som en tilsyneladende kvalitetsforringelse. Det kunne få nogle til at foreslå at undlade at korrigere for det skæve bortfald, da statistikken så ville fremstå som mere rigtigt. Så konsekvensen af en mekanisk anvendelse af numeriske indikatorer uden også at inddrage analyser vil være katastrofal. Indikatorer kan ikke alene danne udgangspunkt for, hvordan man laver statistik med høj kvalitet.

Ideen med at inddrage den foreslåede bortfaldsanalyse er, at det skal være muligt at hæve kvalitetsvurderingen med ét niveau, hvis der findes en dokumentation for den nævnte analyse og efterstratifikation. For at den tællesansvarlige skal kunne løfte den grundlæggende kvalitetsvurdering af bortfaldet én grad fra C til B eller fra E til C, kræver det en dokumentation, der er tilgængelig for brugeren. For ellers har det jo ingen værdi for brugerne.

Eksemplet med bortfaldet AKU-en kunne lægge op til en endnu bedre kvalitetssikring, nemlig en vurdering af effekten af opregningen. Dette ville være muligt i AKU-en, da vi efter en

periode har ledighedskoden fra registrene for de udvalgte personer. Dette registersvar kan vi bruge i den vægtede stikprøve og sammenligne med den tilsvarende kendte samlede registerledighed. En sådan analyse kunne danne baggrund for at hæve den grundlæggende kvalitetsvurdering med yderligere ét niveau. Fx vil en sådan analyse kunne vise, at efterstratifikationen er så effektiv, at registerledigheden estimeret ved stikprøvevægtene kun har en bias på 0,1%.

Begrundelsen for at analyser af et problem skal kunne påvirke kvaliteten positivt, er først og fremmest at det øger brugernes korrekte anvendelse af statistikken, men også at det er her, den tællingsansvarlige direkte kan påvirke kvaliteten, hvilket vi ønsker at fremme.

Generelle kvalitetsdimensioner og specifikke

I Danmarks Statistik arbejder vi hen imod cirka 15 generelle kvalitetsdimensioner, der skal dække alle typer statistik, fx både registerbaseret og stikprøvebaseret. Derudover kan en konkret tælling udarbejde sine egne specifikke mål med fire niveauer. Det sidste kan fx være hvor mange observationer eller andel af markedet et konkret prisindeks skal være baseret på. Et andet eksempel på et specifikt kvalitetsmål i en statistik kunne være en række standarder for fejlsøgningen, hvor meget den fejlsøges og principper herfor. Statistikkontoret kan så fastlægge hvad der skal til for at have en af de fire grader. Pointen er, at kriterierne for disse tællingsspecifikke kvalitetsdimensioner er offentlige og beskrevet præcist på forhånd, så man fx ved hvad der forstås med at fejlsøgningen har en *høj kvalitet*. Hvis flere af de foreslåede 15 dimensioner ikke passer til den konkrete statistik må vi som statistikere kunne formulere mål for hvad vi i stedet forstår med kvaliteten af vores arbejde for den pågældende statistik. Det må være et rimeligt krav til et nationalt statistikbureau.

Anvendelsen af kvalitetsmålene

I og med kvalitetsvurdering A eller B altid kræver udbygget dokumentation kan man indvende, at vurderingen er C - *Rimelig kvalitet*, er mere gratis at bruge, og den så bliver den normale standard. Samt at graden D - *Usikker kvalitet* ikke bliver brugt. Generelt gælder der for begge indvendinger, at man jo stadig skal svare rigtigt på de forskellige kvalitetsmål, selv om kravet til dokumentation måske er mindre. Fx foreslås det at et bortfald på mere end 40% svare til vurderingen D: *Usikker kvalitet*.

Mht. til det sidste af de to indvendinger skal det derudover nævnes, at det vil være i statistikkontorets interesse at skrive, at en konkret kvalitetsdimension er usikker (når den er det), for ellers vil man kunne risikere, at brugerne overfortolker statistikkens kvalitet og problemet vender tilbage som en boomerang. Der er også eksempler på, at der er offentliggjort statistikker, hvor der er taget markante forbehold for sikkerheden af dele af data.

Samlet vurdering af kvaliteten

Det ultimative mål er at lave en samlet vurdering af kvaliteten. En statistik kan ikke være meget bedre end sit svageste led, så det er dette, der skal danne udgangspunkt for den samlede vurdering. Hvis bortfaldet fx er 60% og det har givet den laveste vurdering, hjælper det jo ikke meget at hæve stikprøvestørrelsen markant, for at opnå en mindre stikrøveusikkerhed. Princippet om at det svageste led bestemmer afhænger selvfølgelig af, hvor vigtigt det svageste led er. Hvis det svageste led fx er sæsonkorrektionen pga. af en kort tidsserie og sæsonkorrektionen er et supplement til statistikken er problemet ikke så stort. Men som udgangspunkt er en statistiks samlede kvalitet den laveste af samtlige vurderinger (eller fx tre laveste), med mindre der er argumenter for, at denne kvalitetsdimension ikke er vigtig for den pågældende statistik.

Målet er derudover, at den tællingsansvarlige tør komme med en samlet ærlig vurdering af omfanget af usikkerheden af hensyn til brugerne. Fx formuleringer som *ændringer på op til 1% kan skyldes usikkerheden i statistikken*. Meget få har forsøgt sig med sådanne formuleringer. På den anden side kender den tællingsansvarlige typisk godt statistikkens smertegrænse for fortolkninger, når hun bliver kontaktet af fx journalister. Bag den foreslåede formulering af *ændringer på op til 1% kan skyldes usikkerheder i statistikken* kunne ligge, at stikprøveusikkerheden var 0,5% og de andre usikkerhedskilder blev vurderet til at have den samme størrelsesorden. Det er klart dette er en faglig vurdering. Men at påstå at den eneste usikkerhed er stikprøveusikkerheden ville være en fejlinformation til brugerne.

Så sammenfattende: Selv om det er svært at lave en samlet vurdering af usikkerheden i en statistik, er det omvendt det, som brugerne ønsker for at kunne anvende vores statistikker. Hvis vi ikke laver en sådan vurdering, har statistikken så overhovedet værdi for brugerne?

Bilag Konkrete kvalitetsmål og emner

Nedenstående er der den korte version af nogle af eksemplerne på de ca. 15 konkrete kvalitetsmål og overskrifter for andre. De uddybes i det mundtlige indlæg.

	Usikker	Rimelig	Høj	Meget høj
Sammenlignelighed over tid uden databrud	Under 5 år	Over 5 år	Over 10 år	Over 20 år
Uoplyste værdier for hovedvariabler	Over 3%	Under 3%	Uden 2%	Under 1%
Sæsonkorrektion. Serier der opfylder EU's krav	Mindre 75%	Mindst 75%	Mindst 85%	Mindst 95%
Fejlsøgning. **S**elektiv hhv **M**acro	Ingen fejlsøgning	Ikke selektiv micro	Enten S/M	Både S/M
Cut-off (i erhvervs-statistik)	Over 20% af oms.	Under 20% af oms.	Under 10% af oms.	Under 5% af omsætning.
Stikprøvefejl	Over 3%	Under 3%	Under 2%	Under 1%
Bortfald	Over 40%	Under 40%	Under 20%	Under 5%
Test af skema **P**ilottest hhv. **K**valitativ test **B**rugervurdering	Ingen	Enten P/K/B	Mindst to	Både P/K/B

Alle målene kan opjusteres 1 eller 2 grader, hvis der foreligger en uddybende analyse af problemet eller effekter. Hvad der kan medføre en opgradering afhænger af emnet og er fastlagt på forhånd. Fx en analyse af omfanget og effekten af databruddet.

Andre kvalitetsmål: Dækningsgrad af population, beskrivelse af populationsudvikling over tid, imputerede værdier, revision samt foreløbige og endelig tal, registre, stikprøves repræsentativitet, alder på stikprøve, opregning og indeks.