Berlingske Tidende
Kronik:

Den statistiske løgn

”Statistik er blevet en lettilgængelig og billig knage, som vi kan hænge vores tynde argumenter op på.”

Af Peter Svarre, digital strateg og forfatter, cand.scient.pol.
20. april 2013

Peter Svarre: Moderne teknologi gør i stigende grad indsamling og beabejdning af data til en svir for såvel privatpersoner som medier, virksomheder og det offentlige. Problemet er bare, at vores evner til at behandle og omgås disse data ikke er fulgt med den eksponentielt stigende mængde af dem. Tværtimod.

Vi lever i en verden, hvor omkostningerne ved at indsamle og lagre data tordner ned ad bakke. Enhver idiot kan lave sin egen opinionsundersøgelse med en gratis spørgeskemaservice, som man kan finde på nettet. Enhver virksomhed kan købe sig til millioner af kundedata, og det offentlige har mere end nogensinde adgang til hobevis af økonomiske, sociologiske og personlige data. Problemet er bare, at vores evner til at behandle og omgås alle disse data ikke er fulgt med den eksponentielt stigende mængde. Tværtimod virker det som om, at den stigende mængde af data er blevet akkompagneret af en skødesløshed og ligegyldighed, hvor data sjældent betragtes som et middel til at blive klogere, men oftere bliver et mål i sig selv.

Meningsmålinger indgår i den offentlige debat som orakelagtige sandheder på trods af deres indbyggede usikkerheder. Konklusioner fra nye statistiske sundhedsundersøgelser plastres på forsiden af medierne og forårsager en opblomstren af en hel industri af paleo-low-carb-raw-food-kure, som lever lige præcis frem til den næste nye sundhedsstatistiske undersøgelse. Og i Deadline diskuterer seriøse politikere og journalister økonomiske forudsigelser ned til det sidste decimaltal, selv om det året efter viser sig, at forudsigelsen ikke ramte ved siden af med decimaler, men med hundredvis af procenter.

Tal og statistik kan være fantastiske redskaber til at blive klogere på vores verden, men det er desværre ikke sådan, vi bruger statistik i dag. Statistik er blevet en lettilgængelig og billig knage, som vi kan hænge vores tynde argumenter op på.

Misbrugen af statistik og data kan inddeles i fire typer, som man stort set møder dagligt i nye, gamle og sociale medier.

Den første og mest udbredte type af statistisk misbrug handler om de statistiske usikkerheder. 26. februar kunne man læse følgende overskrift i MetroXpress: »Chokmåling: DF er nu større end S«. Overskriften er baseret på en undersøgelse fra YouGov, som viser, at Dansk Folkeparti står til at få 17,4 procent af stemmerne, mens Socialdemokraterne står til 17,2 procent. Lige ved siden af artiklen kan man dog læse en note, hvor det fremgår, at den statistiske usikkerhed på tallene i undersøgelsen er på +/- 2,5 procentpoint, hvilket med andre ord betyder, at artiklens overskrift er ren og skær fantasi. Når man indregner den statistiske usikkerhed, kan det nemlig lige så vel være tilfældet, at Socialdemokraterne er noget større end Dansk Folkeparti.

Flere og flere medier er blevet bedre til at rapportere usikkerhederne, men det pudsige er, at journalisterne alligevel vælger at ignorere usikkerhederne og skrive bombastiske overskrifter, som sjældent har noget belæg i det statistiske materiale. Journalister, eksperter og meningsdannere kan med andre ord skrive artikler, kronikker, kommentarer og

ledere på en konklusion, som i princippet lige så godt kunne være den stik modsatte – hvis man altså tog usikkerhedsintervallerne seriøst.

Den anden type misbrug handler om, at man får svar, som man spørger. 5. november 2012 publicerede Kulturministeriet en undersøgelse af danskernes kulturvaner, som meget opløftende viste, at danskerne gik langt mere i teatret, end vi alle gik rundt og troede. Problemet var bare, at undersøgelsen var baseret på en spørgeskemaundersøgelse og ikke på tal for teatrenes billetsalg. Efter at historien havde taget en hurra-runde i det selvrefererende mediecirkus, kunne lektor i teatervidenskab Stig Jarl nemlig afsløre, at tallene fra spørgeskemaundersøgelsen var cirka fire gange større end de tal, som man kunne måle ude i den virkelige virkelighed.

Hvis man spørger danskerne, om vi har været i teatret inden for det seneste år, har vi alle en tendens til at overdrive vores kulturforbrug. Ikke fordi vi bevidst ønsker at lyve, men fordi vi har et billede af os selv som kloge og veluddannede mennesker, som går i teatret. Så hvis man ikke lige har været i teatret i 2012, så var man vel i 2011, og det tæller vel også – gør det ikke? Internetbaserede meningsmålingsfirmaer har gjort det muligt at lave hurtige og billige online surveys, hvor man kan spørge folk om dit og dat, men problemet er bare, at man ikke kan spørge folk om alt – i hvert fald ikke hvis man forventer at få et sandfærdigt svar. Lige så snart man begynder at spørge mennesker om emner, der har betydning for deres identitet og selvforståelse, vil man opleve at folk lyver, overdriver og har selektiv hukommelse.

Den tredje type af misbrug handler om, hvad der egentlig forårsager hvad. I en artikel i Berlingske 2. januar 2013 kunne vi læse, at overvægtige mennesker havde seks procent mindre risiko for at dø i en given periode end normalvægtige. Den naturlige konklusion hos læseren er naturligvis, at man hellere må se at blive overvægtig, hvis man vil leve længere.

Men hvad nu, hvis det ikke er overvægten, der forårsager livsforlængelsen? Hvad nu, hvis overvægtige mennesker typisk lever et mere afslappet og mindre stressende liv end normalvægtige (som jo hele tiden er på paleo-low-carb-raw-food-kure)? Så er det jo i virkeligheden ikke normalvægten, men stressen, som tager livet af de normalvægtige. Der kan være et utal af bagvedliggende og mellemliggende variable, som er de virkelige årsager til en hændelse. Og når ret skal være ret, kontrollerer forskere ofte for mange af disse variable, men livet er en kompleks størrelse, og det kan sjældent lade sig gøre at kontrollere for alle bagved- og mellemliggende variable.

Det største problem i vores lemfældige omgang med årsagssammenhænge er dog, at de sjældent behandles i mediernes overfladiske referater af komplekse undersøgelser. Som læser sidder man derfor næsten altid tilbage med flere spørgsmål end svar, når man læser om nye revolutionerende og banebrydende undersøgelser. Hvorfor er der ikke flere journalister, der stiller spørgsmål til forskere og udfordrer de påståede årsagssammenhænge, som de udleder af de statistiske sammenhænge? Skal der virkelig tusindvis af døde forsøgsmus, omfattende svindel med forskningsresultater og en rød Toyota cabriolet til, før journalister gider stille kloge spørgsmål til forskningsresultater?

Den fjerde type misbrug hænger sammen med vores nyligt erhvervede adgang til enorme datamængder. Tænk på eksemplet med de overvægtiges længere liv. Undersøgelsen legitimeres af, at den er baseret på et aggregat af 100 statistiske sundhedsundersøgelser fra hele verden, og at de tilsammen indeholder data om 2,8 millioner mennesker. Det lyder fantastisk troværdigt, men det er det ikke nødvendigvis. Det er blevet ufattelig billigt og ufattelig nemt at indsamle, sammenkøre og krydse data og dermed udlede sammenhænge af disse data. Problemet er blot, at statistiske sammenhænge ikke er ensbetydende med virkelige sammenhænge, og hvis man har adgang til millioner af datapunkter og krydser dem systematisk, vil man uundgåeligt ende med at finde statistiske sammenhænge.

Enhver person, der har snuset til et kursus i statistik, vil vide, at de fire typer af misbrug på ingen måde er nye. Folk, der har arbejdet seriøst med statistik og data, har altid kendt til disse helt grundlæggende begrænsninger i statistikkens muligheder. Det nye er, at statistik og data er blevet allemandseje, og at statistik indgår i alle vores politiske, økonomiske og personlige beslutninger. Vi planlægger vores kostvaner, vores politiske indgreb og vores virksomhedsstrategier på grundlag af statistik og data, som oven i købet ofte filtreres igennem mediernes overskriftshungrende optik. Så længe forskere, journalister og borgere ikke forholder sig seriøst til statistikkens begrænsninger, vil vi bygge vores samfund, vores virksomheder og vores liv på en statistisk velunderbygget løgn.