SSFr
Svenska statistikfrämjandet

MINNESORD/
JAN LANKE

SIDAN 14

EN STATISTISK
PARADOX

SIDORNA 15-17

Qvintensen

SVENSKA STATISTIKFRÄMJANDET NR 2 2016

BIG DATA/
– big accountability – ett modeord eller?

SIDORNA 8-13

»Utan statistik är vi ingenting!»

– JOURNALIST FREDRIK LAURIN PÅ ÅRETS STÖRSTA STATISTIKMÖTE

SIDORNA 2-4

Statistik för kvalitetsteknik

SIDORNA 24-26

Anna Rosling-Rönnlund lägger idag det mesta av sin tid på satsningen Dollar street och vill med hjälp av fotografier och statistik visa hur världen verkligen ser ut.

» Hon förnekar inte att det finns stora utmaningar, men allt för många tror att utvecklingen går åt fel håll. Enligt Anna är det tvärtom.»

Nordiska statistikermötet har arrangerats sedan 1927. I år var det Sveriges tur.

Årets största statistikermöte

När SCB den 22-24 augusti stod värd för det 27:e Nordiska statistikermötet, NSM, samlades över 350 statistiker i Stockholm. NSM arrangerades första gången 1927 och äger rum vart tredje år. De nordiska ländernas statistikbyråer turas om att arrangera och det var alltså 15 år sedan NSM ägde rum i Sverige.

Stefan Lundgren, generaldirektör på SCB, pekade ut två viktiga trender.

Big data på agendan

Temat för NSM 2016 var Statistics in a changing world – towards 2020 and beyond. SCB:s generaldirektör Stefan Lundgren tog i sitt invigningstal på Stockholm Waterfront Congress Centre upp två viktiga trender som statistiksamfundet måste vara uppmärksam på. Den ena är att efterfrågan på statistik från samhällets olika aktörer blir allt större och att vi idag förväntar oss att beslut är evidensbaserade. Den andra trenden är big data och de allt större mängderna information som nu finns tillgängliga. Det senare är en utveckling som inte saknar risker. I framtiden kan nya aktörer ta över uppgifter som tidigare varit vikta för statistikbyråer.

Big data togs upp vid flera tillfällen under NSM. Exempelvis på Can Tongurs, metodstatistiker på SCB, seminarium med rubriken Statistics in the digital era. Under seminariet försökte Can bland annat besvara frågan om och i så fall hur man kan använda sig av big data för att producera statistik. Frågeställningen gav upphov till diskussioner på det fullsatta seminariet.

Nya perspektiv

Ett viktigt syfte med NSM är att nätverka och utbyta erfarenheter. Därför är de presentationer som statistikbyråernas medarbetare håller en central del av NSM. I år hölls ett 70-tal korta presentationer som handlade om statistikens roll i samhället, kommunikation med användare samt hur datainsamling och statistikproduktion kan bli effektivare. Samtliga papers som presenterades finns på www.scb.se/nsm2016 under punkten speakers.

Nya perspektiv och idéer utifrån är ett annat viktigt inslag på NSM. En av de inbjudna talarna var Xavier Badosa från Statistical Institute of Catalonia. Ämnet för hans master class var Consuming Nordic Statbank with JSON-stat. Hans tre timmar långa seminarium lockade ett 40-tal deltagare som fick en grundlig genomgång av hur man använder JSON-stat som utvecklats för att bland annat göra det enklare att visualisera data och att arbeta med öppna data. PX-web, som SCB och övriga nordiska statistikbyråer använder sig av, stöder JSON-stat. Därför kunde Badosa guida seminariedeltagarna genom JSON-stat med hjälp av bland annat svensk arbetslöshetsstatistik och norska befolkningssiffror.

En annan uppskattad talare var Jussi Karlgren från KTH och Gavagai, där han är verksam inom området datorer och lingvistik. Även Jussi intresserar sig för big data, främst då stora textmängder, som han försöker hitta mönster i. Han betonade vikten av kunskap och analys för att undvika risken att dra helt felaktiga slutsatser av de stora datamängder man granskar.

Utan statistik är vi ingenting

Journalisten Fredrik Laurin, känd från TV:s Uppdrag granskning, var inbjuden för att tala om hur journalister använder sig av data och statistik i reportage och artiklar. Fredrik inledde sitt seminarium med att utbrista "Utan statistik »»»

HÄNT OCH HÖRT

Disputationer

13 nov 2015 **ROBERT SZULKIN**, inst. för medicinsk epidemiologi och biostatistik, KI. Opponent: Paul Frenks.

2 juni 2016 **MIKAEL PETERSSON**, matematisk statistik, Stockholms universitet. Opponent: Nikolaos Limnios.

15 juni 2016 **JENS MALMROS**, matematisk statistik, Stockholms universitet. Opponent: Erik Volz.

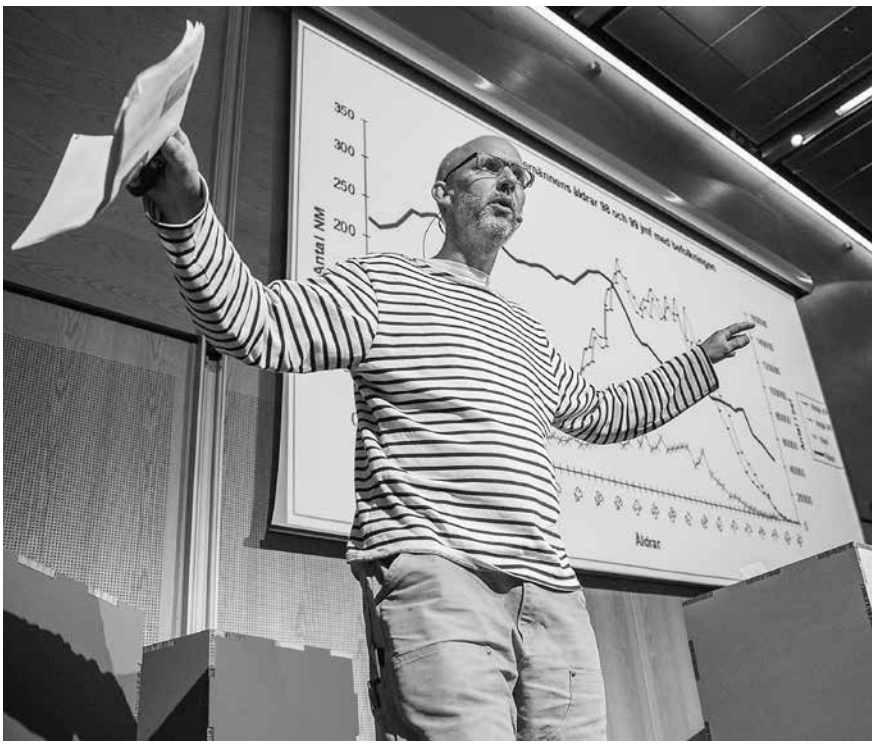
Théâtre Royal i Marrakech.

Kommande konferenser

New Techniques and Technologies for Statistics
(Organized by Eurostat on new techniques and methods for official statistics, and the impact of new technologies on statistical collection, production and dissemination systems. Papers presented are eligible for submission to a special NTTS issue of the Journal of Official Statistics.)
14 – 16 Mars 2017, Bryssel.

International Conference on Official Statistics
30-31 Mars 2017, Sarajevo.

61st ISI World Statistics Congress
16 – 21 Juli 2017, Marrakech, Marocko



Fredrik Laurin.

»»» ÅRETS STÖRSTA STATISTIKERMÖTE

är vi ingenting!". Med stor inlevelse berättade han sedan om ett antal exempel där fakta varit helt avgörande för att bevisa missförhållanden i samhället. I Uppdrag granskning barndom var det inte alltid så noga med tolkningen av statistiken vilket straffade sig. De senaste tio åren har redaktionen därför alltid anlitat experter före publicering.

Visualisering för en begripligare värld

Visualisering av data är ett effektivt verktyg för att nå fler och för att levandegöra statistiken. Om detta pratade både Mikael Ahlström, entreprenör och grundare av utbildningsföretaget Hyper island, och Anna Rosling-Rönnlund som tillsammans med Ola och Hans Rosling grundat Gapminder.

Mikael Ahlströms budskap var "Visualize data and people will start to care!" och det var något som Anna Rosling-Rönnlund kunde skriva under på. Hon lägger idag det mesta av sin tid på satsningen Dollar street och vill med hjälp av fotografier och statistik visa hur världen verkligen ser ut. Syftet är att ge en motbild till den negativa bild av världen som medierna ofta förmedlar. Hon förnekar inte att det finns stora utmaningar, men allt för många tror att utvecklingen går åt fel håll. Enligt Anna är det tvärtom.

Nordiska värderingar i Baltikum

Om tre år är det Finlands tur att bjuda in till NSM. Och då kanske de baltiska länderna också är representerade. Idén föddes under Bi Puranens presentation av World Value Survey. Hon hade nyligen varit i Baltikum och upptäckt att de baltiska staternas medborgare värderingsmässigt står nära Norden. Innan NSM 2016 avslutades fick deltagarna därför med hjälp av mentometer svara på frågan om de tyckte att det var en god idé att bjuda in grannarna på andra sidan Östersjön. En majoritet svarade ja på frågan.

LARS PYK
STATISTISKA CENTRALBYRÅN

Transaktionsdata och officiell statistik

Detta är en sammanfattning av min presentation vid Nordiska statistikermötet i Stockholm 2016, under temat transaktionsdata i officiell statistik. Jag tog upp vissa studier av alternativa datakällor och deras kvalitetsaspekter samt redogjorde för praktisk användning av transaktionsdata i Konsumentprisindex (KPI). Det verkar finnas en del nya datakällor med potential att användas i officiell statistikproduktion, om än med vissa utmaningar.

Stora datamängder som genereras exempelvis genom Internet, olika applikationer så som sensorer, eller transaktioner mellan olika handelsled, skulle kunna ge en bild av hur människor och företag agerar. De senaste åren har det gjorts en del studier där alternativa datakällor används för att exempelvis replikera eller förekomma officiell statistik. Ofta är det ostrukturerade data fångade via Internet, så kallade Big Data. Ofta är den officiella statistiken något av ett riktmärke eller målparameter för dessa studier, och då vanligtvis på aggregerad nivå – mer sällan avseende de delar som bildar aggregaten. Inte sällan föreligger en del automatiserad datatvätt i processen.

»Stora datamängder som genereras exempelvis genom Internet, olika applikationer så som sensorer, eller transaktioner mellan olika handelsled, skulle kunna ge en bild av hur människor och företag agerar.»

Det förekommer även kvalificerade vetenskapliga studier där alternativa datakällor, exempelvis mobiltelefonidata, har kunnat användas för att studera vissa typer av medicinska och sociala problem, särskilt i utvecklingsländer, och då kan det finnas stora vinster att hämta i termer av folkhälsa och kapacitetsbyggande för exempelvis epidemier.

På SCB har vi, som nämnts, börjat använda transaktionsdata inom prismätningar av dagligvaror i KPI. Detta tack vare att de stora kedjorna inom dagligvaruhandeln förser SCB med aggregerade data över transaktioner per produkt (men inte per kund!). Sverige är inte ensamt om att använda denna typ av data i den officiella prisstatistiken men hör, jämte Nederländerna, Norge och Schweiz, till de ledande länderna i världen vad gäller analyser av datakvalitet och kvalificerade metodmässiga överväganden.

Ett känt exempel på ett annat förfarande för prisinsamling är forskningsprojektet The Billion Prices Project vid MIT, där man har kunnat efterlikna USAs officiella konsumentprisindex (någorlunda) genom att använda s.k. web scraping, vilket innebär att Internet skannas av maskinellt med så kallade robotar för att fånga priser för olika produkter, med daglig frekvens. Metoden förefaller vettig, men den förutsätter, på samma sätt som KPI, att man väljer produkter enligt etablerade indelningar och väger samman priser till produktgrupper i en ”varukorg” enligt faktiska konsumtionsmönster. Det senare är, till min vetskap, inte uppfyllt, och då kan man fråga sig om beräkningarna fyller något syfte? I slutändan handlar det hela om lite mer än att enbart fånga data.

För etablerad officiell statistik som ska utgöra beslutsunderlag för olika samhällsfunktioner kan utrymmet att använda alltför aparta datakällor vara litet, med tanke på precisionskraven för estimaten, särskilt om det handlar om inflation (jämförelser av lika med lika, över tid) eller arbetslöshet, totalt och inom demografiska grupper. Därmed inte sagt att officiell statistik är perfekt, men det ligger ändock ett omfattande vetenskapligt arbete bakom den redan etablerade datafångsten, med syfte att tjäna specifika frågeställningar.

Det verkar finnas lovande datamängder som inte är ”designdata”/ surveydata, nämligen administrativa data hos exempelvis företag, liksom även transaktionsdata liknande de försäljningsuppgifter vi tar emot på SCB. När det finns en struktur och tydlighet i data, när informationsinnehållet är otvetydigt och stort kontra bruset, då föds möjligheten att utreda vad datakällan kan tillföra officiell statistik, även om det kräver metodförändringar. Detta är ett välkommet inslag i tider med bortfall och krav på minskad uppgiftslämnarbörda. Vi är ännu bara i början av dataåldern, trots allt.

CAN TONGUR,
STATISTISKA CENTRALBYRÅN

Några tankar om folkomröstningar

I förra numret av Qvintensen, liksom även i detta nummer, så är det en dominans av bidrag med samhällsvetenskaplig inriktning. Det är naturligtvis en återspegling av statistikens viktiga roll i samhället.

Men det finns också mycket annat, som kunde passa i Qvintensen, till exempel:

- Artiklar med inriktning mot andra områden, till exempel medicinsk statistik eller industriell statistik.
- Diskussioner och debattinlägg.
- Yrkeserfarenheter. Hur är det att jobba som statistiker inom olika specialområden?
- Synpunkter på undervisning, läromedel och kursinnehåll i dagens statistikundervisning.
- Statistik i massmedia.
- Historiska och filosofiska synpunkter på statistisk verksamhet och statistiska resonemang.
- En och annan överraskning kanske?
- Interna frågor rörande Statistikfrämjandet.
- Informationen om genomförda disputationer och lic.-seminarier är för närvarande extremt ofullständig. Jag skulle önska att varje institution utser någon som rutinmässigt rapporterar till Qvintensen.

Tidigare har Qvintensen publicerat ett kalendarium över kommande konferenser. Jag är osäker på om vi ska fortsätta med detta. Det är tidskrävande, och det kan vara svårt att veta var man ska dra gränsen för vad som är av intresse för Qvintensens läsare. De som arbetar på institutioner eller organisationer får nog snabbare information på sina arbetsplatser. Dessutom är det lätt att leta på nätet för den som är intresserad. Men det hindrar inte att vi i notisform kan informera om sådana kommande konferenser, som vi vill rikta speciell uppmärksamhet på.

Slutligen några privata funderingar kring folkomröstningar, ett aktuellt ämne. Det finns likheter, men också olikheter, mellan folkomröstningar och opinionsundersökningar. I båda fallen vill man veta vad folk tycker. Men en olikhet är att en folkomröstning alltid är en totalundersökning, medan en opinionsundersökning i regel är baserad på ett stickprov från den aktuella populationen. I en opinionsundersökning finns vanligen svarsalternativ, som uttrycker grader av instämmande i ett påstående, till exempel ”instämmer helt”, ”instämmer delvis”, ”ingen bestämd uppfattning”, och liknande. I en folkomröstning däremot finns inga svarsalternativ som uttrycker osäkerhet, utan här är det bara ”ja” eller ”nej”.

Vid opinionsundersökningar anses i regel stort bortfall vara en källa till bekymmer. Man använder ofta den ena eller andra tekniken för att försöka i någon mån korrigera för bortfallets eventuella snedvridande effekt. Men vid folkomröstningar talas det inte mycket om bortfall, trots att det ofta kan vara en stor andel av befolkningen som inte röstar. Här några aktuella exempel. (För enkelhets skull har jag i alla exemplen avrundat de procenttal jag hittat i media)

Exempel 1. Resultatet från den engelska folkomröstningen (juni 2016) om Brexit brukar i media redovisas så här. :

Lämna EU	52%
Stanna i EU	48%
Totalt	100%

Man har dragit slutsatsen att en knapp majoritet av befolkningen vill lämna EU. Men intrycket blir ett annat, om man även redovisar de icke-röstande:

Lämna EU	37%
Stanna i EU	35%
Ej röstat	28%
Totalt	100%

Nu verkar inte slutsatsen lika självklar längre.

Exempel 2. Vid folkomröstningen i Colombia (oktober 2016) frågades: Stöder du fredsavtalet mellan regeringen och Farc-gerillan? Resultatet brukar redovisas så här:

Ja	49,8%
Nej	50,2%
Totalt	100%

Alltså en majoritet (om än mycket knapp) för nej-sidan säger man. Men om man tar hänsyn till de icke-röstande blir resultatet (efter avrundning):

Ja	18%
Nej	19%
Ej röstat	63%
Totalt	100%

Ett osäkert underlag att dra några slutsatser från.

Exempel 3. Vid folkomröstningen i Ungern (oktober 2016) var frågan: Vill du att EU ska kunna tvinga Ungern att ta emot icke-ungerska immigranter utan parlamentets medgivande? Resultatet brukar redovisas som:

Ja	1,7%
Nej	98,3%
Totalt	100%

En förkrossande seger för nej-sidan ser det ut som. Om hänsyn tas till de icke-röstande blir resultatet (efter avrundning):

Ja	1%
Nej	43%
Ej röstat	56%
Totalt	100%

Här tillkommer komplikationen att det krävdes minst 50 procents valdeltagande för att folkomröstningen skulle betraktas som giltig. Man vet också att ja-sidan i förväg bestämt sig för att bojkotta folkomröstningen, varför de icke-röstande kan antas till viss del bestå av ja-anhängare. Hur många är förstås okänt, men nej-anhängarnas seger ser nog i alla fall större ut på papperet än vad som gäller i verkligheten.

Kanske finns det några av Qvintensens läsare som vill kommentera eller korrigera? I så fall är ni välkomna!

JAN WRETMAN

Qvintensen

Ansvarig utgivare
John Öhrvik

Redaktör
Jan Wretman, 070-781 78 35

Redaktion
Marie Linder, FMS
Jonas Ortman, Surveyföreningen
Hans Ekholm
Alf Fyhrlund
Anna Torrång

E-post wretman.jan@gmail.com

Produktion
Form och redigering: Mezzo Media AB
Tryckeri: Trydells Tryckeri AB

Annonser
Annonser i Qvintensen bokas med redaktören.
Annonsutskick per e-post bokas med Svenska statistikfrämjandets sekreterare.
Priset är 6 000 kronor för institutionsmedlem eller motsvarande, och
8 500 kronor för övriga annonsörer. På alla priser tillkommer 25 % moms.

Svenska statistikfrämjandet

SVENSKA STATISTIKFRÄMJANDETS STYRELSE
Ordförande
John Öhrvik, 070-666 08 08,
John.Ohrvik@ki.se

Vice ordförande Patrik Rydén

Skattmästare Maria Juhlin

Sekreterare Ingrid Lönnstedt

Hemsidesansvarig Alf Fyhrlund

Klubbmästare Vakant

Surveyföreningen
Åke Wissing

FMS
Anna Lindgren

Industriell statistik
Hans Alberg

Cramérsällskapet
Rebecka Jörnsten

Övriga ledamöter
Inger Eklund, Karin Lindgren

Adress
Svenska statistikfrämjandet
Sekreterare: Ingrid Lönnstedt

E-post sekrfram@gmail.com

Webbplats www.statistikframjandet.se

Nästan 70 procent av de tillfrågade svarade att de är oroliga över hur deras personuppgifter hanteras.»

DANIEL TORNBERG OCH ALEXANDER JUTE
SIDORNA 8-10

Qvintensen
SVENSKA STATISTIKFRÄMJANDET
NR 2 2016

Innehåll

ÅRETS STATISTIKMÖTE 2-4

TRANSAKTIONSDATA 5

REDAKTÖRENS RUTA 6

BIG DATA/
Big data, big accountability 8-10
"Biga data" och "Predictive Analytics" – modeord eller...? 10-13

MINNESORD
Vi minns Jan Lanke 14

EN STATISTISK PARADOX 15-17

Våra föreningar

SURVEYFÖRENINGEN
Ordföranden har ordet 18
Uppsatstävling 19
Workshop om surveymetodik 19

FÖRENINGEN FÖR MEDICINSK STATISTIK
Ordföranden har ordet 20
FMS-stipendium 21
Konferensrapporter 21-22

FÖRENINGEN FÖR INDUSTRIELL STATISTIK
Ordföranden har ordet 22
Om terminologi för riskanalys 23
Statistik för kvalitetsteknik 24-26

ISSN 2000-1819
Författare och andra som utan ersättning bidrar med material för publicering behåller upphovsrätten till sina verk. Svenska statistikfrämjandet äger rätten att publicera dessa verk i valfri form och upplaga samt att behålla alla eventuella intäkter från sådan publicering, om inget annat avtalas. Författare och andra som utan ersättning bidrar med material för publicering garanteras rätten att själva få sprida sitt verk under förutsättning att Qvintensen anges som originalkälla med angivande av nummer och år. Svenska statistikfrämjandet äger upphovsrätten till publicerat material då det är framställt av redaktionsmedlem, styrelsemedlem, någon till Främjandet knuten förenings styrelse eller person som erhållit arvode för sitt bidrag. © Svenska statistikfrämjandet 2016.

Advokaterna Daniel Tornberg och Alexander Jute, båda vid advokatfirman MarLaw, har mångårig erfarenhet av rådgivning kring företag och organisationers personuppgiftshantering och är flitigt anlitade föreläsare på området. I följande artikel delar de med sig av den senaste utvecklingen på området och ger dessutom några praktiska tips.

BIG DATA GACCO

Inledning

Big Data är ett brett begrepp och kan ha en mängd olika betydelser beroende på sammanhang. Ur ett legalt perspektiv tar termen främst sikte på stora mängder data från olika typer av källor som kan innehålla personuppgifter. Utifrån dessa uppgifter kan man bilda sig en uppfattning om personen bakom uppgifterna. Många företag samlar in stora mängder data för att erhålla information om sina kunder, t.ex. för att kunna effektivisera marknadsföringskampanjer. Den tekniska utvecklingen har på senare tid möjliggjort detaljerade kartläggningar av personers beteenden, inte minst i onlinemiljö, och fler och fler uppfattar detta som problematiskt. Från legalt håll ställer denna typ av åtgärder vissa krav på insamlingen och hanteringen av personuppgifter. Detta har även fått högsta aktualitet med anledning av att EU nu beslutat om en ny dataskyddsförordning som kommer bli direkt tillämplig i samtliga medlemsstater inom två år.

Att frågorna om hantering av personuppgifter är aktuella visar inte minst en undersökning utförd av TNS Opinion & Social på uppdrag av EU-kommissionens generaldirektorat för rättsliga frågor och konsumentfrågor. Personer representerande hela EU fick bland annat frågan: *Myndigheter och privata företag som innehar information om dig kan ibland, utan att informera dig, använda den i annat syfte än det den insamlades för (t.ex. för direktmarknadsföring, riktad onlinereklam, profilering). Hur orolig är du över att din information används på detta sätt?*

»Nästan 70 procent av de tillfrågade svarade att de är oroliga över hur deras personuppgifter hanteras.»

Nästan 70 procent av de tillfrågade svarade att de är oroliga över hur deras personuppgifter hanteras. Se Special Eurobarometer 431, publicerad juni 2015.

Ett prioriterat område inom EU

Frågor om hantering av personuppgifter har fått förnyad aktualitet, men faktum är att de länge varit ett prioriterat område. Detta framgår inte minst av att skyddet för personuppgifter är reglerat i EU:s stadga om de grundläggande rättigheterna, artikel 8, vilken anger att var och en har rätt till skydd av de personuppgifter som rör honom eller henne samt att uppgifterna ska behandlas lagenligt för bestämda ändamål och på grundval av den berörda personens samtycke eller någon annan legitim grund. Dessutom stadgas att en oberoende myndighet ska kontrollera att dessa regler efterlevs.

Dessa övergripande bestämmelser är sedan 1995 kompletterade av ett direktiv, Europaparlamentet och rådets direktiv 95/46/EG av den 24 oktober 1995 om skydd för enskilda personer med avseende på behandling av personuppgifter och om det fria flödet av sådana uppgifter (dataskyddsdirektivet). I direktivet fastställs strikta begränsningar för insamling och användning av personuppgifter. En av utgångspunkterna i direktivet, i likhet med stadgan, är att personuppgifter endast får behandlas om den registrerade otvetydigt har lämnat sitt samtycke.

Direktivet har implementerats i svensk rätt genom personuppgiftslagen (1998:204) (PuL). I den inledande paragrafen anges att syftet med lagen är att skydda människor mot att deras personliga integritet kränks genom behandling av personuppgifter. 3 § PuL innehåller definitioner av några begrepp som är väsentliga att känna till för den kommande framställningen. Med *personuppgifter* avses all slags information som direkt eller indirekt kan hänföras till en fysisk person som är i livet. Det kan exempelvis vara allt från ett namn, foto, e-postadress, bankuppgifter, inlägg på sociala medier, medicinsk information, IP-adresser, m.m. Med *behandling* av personuppgift avses varje åtgärd eller serie av åtgärder som vidtas i fråga om personuppgifter, vare sig det sker på automatisk väg eller inte. Exempelvis kan det röra sig om insamling, registrering, organisering, lagring bearbetning, användning, utlämnande genom översändande, sammanställning eller samkörning, utplåning eller förstöring. Med *samtycke* avses varje slag av frivillig, särskild och otvetydig viljeyttring genom vilken den registrerade, efter att ha fått information, godtar behandling av personuppgifter som rör honom eller henne.

Även om det finns reglering kring hanteringen av personuppgifter har EU identifierat ett behov av en översyn av personuppgiftsskyddet inom EU, inte minst på grund av att nuvarande ordning, med direktiv som implementerats på olika sätt i olika medlemsstater, har gjort det svårt för multinationella företag att behandla uppgifterna på ett enhetligt sätt även inom Europa. Dessutom har den tekniska utvecklingen ställt nya krav på en modern reglering. Direktivet är från 1995 då internet fortfarande var i sin linda. Mot bakgrund av detta har man nu tagit fram och även beslutat om en ny dataskyddsförordning (beslutat 14 april 2016) som träder i kraft om två år.

Överföring till tredje land

En särskild, om än inte ovanlig, form av hantering av personuppgifter är överföring till tredje

UNTABILITY

land, dvs. överföring av personuppgifter från en medlemsstat till ett land utanför EU, t.ex. USA. Exempelvis innefattar många molntjänster och sociala medier sådan överföring då tjänsterna ofta är baserade utanför EU. För att överföra personuppgifter till ett tredje land ställs krav på att det tredje landet har en adekvat skyddsnivå. EU- kommissionen har fattat ett flertal beslut om att specifika länder måste ha en adekvat skyddsnivå, de s.k. Safe harbour-reglerna. Ett av länderna som kommissionen ansett ha ett adekvat skydd är USA och sedan 2000 har överföring av personuppgifter till USA kunnat göras med stöd av Safe harbour-reglerna. Dock har det genom olika avslöjanden kommit fram uppgifter om att USAs myndigheter har långtgående möjligheter att få fram uppgifter från amerikanska bolag samt att myndigheter i USA massövervakar medborgare även i EU. Med anledning av dessa avslöjanden anmälde en österrikisk medborgare Facebook Ireland Ltd till den irländska dataskyddsmyndigheten och hävdade att Facebook överförde Facebookanvändares personuppgifter till USA, trots att USA inte garanterade en adekvat skyddsnivå för personuppgifterna. Ärendet slutade i EU-domstolen, som i oktober förra året fann att USA saknade en adekvat skyddsnivå och att överföring av personuppgifter till USA baserad på Safe harbour-reglerna inte var tillåten.

»Ärendet slutade i EU-domstolen, som i oktober förra året fann att USA saknade en adekvat skyddsnivå ... »

Detta innebär dock inte att EU och USA ska sluta bedriva handel mellan varandra, men det har medfört att ett sedan tidigare pågående projekt mellan EU och USA, ”Privacy Shield”, intensifierats för att ersätta Safe harbour-reglerna. Projektet innebär bland annat strängare skyldigheter för företag i USA för att kunna erbjuda ett effektivt skydd av EU-medborgarnas rättigheter. Projektet är i princip i mål, dock återstår ett slutligt beslut mellan parterna. Redan nu har projektet fått kritik från tillsynsmyndigheterna i Europa som bland annat ifrågasätter vilken rättskraft de europeiska dataskyddsmyndigheternas beslut kommer att få i USA, hur federala myndigheter kommer att kunna garantera att reglerna följs, och om avtalet verkligen är anpassat till den nya dataskyddsförordningen, eller om man måste förhandla om avtalet om två år när förordningen träder i kraft. Detta innebär att försiktighet ska tillämpas vid överföringar av personuppgifter till USA.

Hantering av personuppgifter i undersökningar

När det gäller undersökningar så gäller givetvis de vanliga reglerna som genomgåtts ovan. Därutöver har även ESOMAR tagit fram en ”Data protection checklist” som förklarar hur personuppgiftsreglering bör tillämpas vid undersökningar, och vilken ansvar undersökningsföretag bör åta sig. I korthet innehåller dokumentet förklaringar till hur man begränsar användning av uppgifter, hur samtycke bör inhämtas, hur data får överföras och hur man bör använda en privacy policy. Dessutom ges rekommendationer för vissa speciella fall som rör insamling av data från barn, B2B, fotografier, ljud- och bildinspelningar, molntjänster m.m.

Vilka nyheter innebär dataskyddsförordningen?

Dataskyddsförordningen, som nu beslutats av EU, syftar slutligt till att möjliggöra en enhetlig ”digital single market”, dvs. att det ska bli enklare för multinationella företag inom EU att hantera personuppgifter, och att EU-medborgare ska få utökad insyn i vilka uppgifter som behandlas om dem. Några av de viktigaste bestämmelserna för enskilda rör rätten att begära att personuppgifter raderas, rätten att få åtkomst till sina personuppgifter för att föra över dem till en annan leverantör av elektroniska tjänster för kommunikation, med mera. Dessutom innebär bestämmelserna tydligare regler om ansvar hos dem som behandlar personuppgifter, främst personuppgiftsansvariga men också personuppgiftsbiträden. Vidare finns det krav på inbyggda dataskyddsgarantier (inbyggd integritet/privacy by design) samt skyldighet att anmäla eventuella incidenter till tillsynsmyndigheten. Förordningen innehåller också regler om förstärkt samarbete mellan de olika EU-medlemsstaternas dataskyddsmyndigheter och en s.k. one-stop-shop-mekanism som ska underlätta för sådana personuppgiftsansvariga som är verksamma i flera medlemsstater (till exempel internationella företagskoncerner) – tanken är att de endast ska behöva vara i kontakt med en av de olika nationella tillsynsmyndigheterna. Dessutom införs utökade sanktionsmöjligheter, som bl.a. innebär administrativa avgifter på upp till 20 miljoner euro eller 4 % av den globala årsomsättningen för företag som inte följer regleringen.

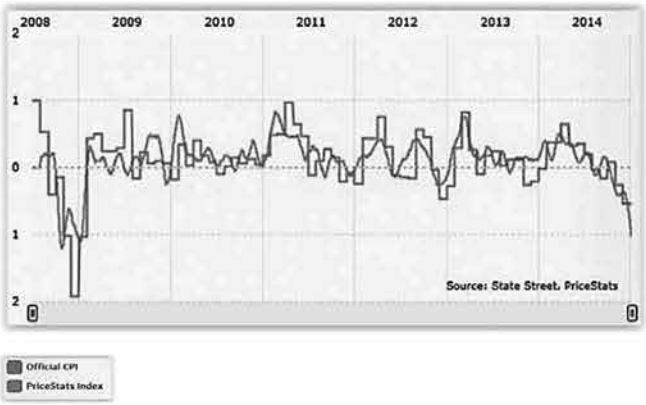
Företagens utökade insyn ställer bland annat långtgående informationskrav vid insamling av personuppgifter, bland annat ska företagen informera om följande när uppgifter samlas in: syftet med insamlingen, om den personuppgiftsansvarige avser att överföra personuppgifter till ett tredje land eller en internationell organisation och huruvida ett beslut av kommissionen om adekvat skyddsnivå föreligger eller saknas, den »»»

period under vilken personuppgifterna kommer att lagras, att det föreligger en rätt att av den personuppgiftsansvarige begära tillgång till och rättelse eller radering av personuppgifter eller begränsning av behandling som rör den registrerade eller att invända mot behandling samt rätten till dataportabilitet, m.m.

Slutsatser/Att tänka på

Hanteringen av personuppgifter är ständigt aktuell varför det är viktigt att noga följa lagstiftningsutvecklingen. I och med att dataskyddsförordningen är beslutad och innebär att företag och organisationer inom två år ska ha vidtagit nödvändiga åtgärder för att kunna efterleva det nya regelverket, rekommenderas att man redan nu påbörjar det förändringsarbete som den nya förordningen kräver. Vi ser ofta ett behov av att utarbeta, eller göra översyn av befintlig personuppgiftspolicy, samt att en generell översyn görs av vilka behov av data ett företag eller organisation verkligen har. Använder/behöver man verkligen alla data som det idag är möjligt att samla in? Tänk på att stora mängder data också ställer stora krav på hanteringen och ansvaret för dessa. Utgå därför från behovet och ta därefter fram en policy för hur företaget eller organisationen ska hantera data. En policy möjliggör ett enkelt sätt att såväl internt som externt kommunicera vilka värderingar ett företag eller en organisation har kring sin personuppgiftshantering.

DANIEL TORNBERG OCH
ALEXANDER JUTE



Figur 1. Inflationen I USA: PriceStats Index vs. Official CPI. Från AAPOR (2015). Figuren finns också i färg på sidan 27.

Tre tillämpningar

Tre framgångsrika tillämpningar av ”Big Data” och ”Predictive Analytics” nämns ofta i litteraturen: Google Translate, Konsumentprisindex (KPI) och Google Flu Trends.

Google Translate är ett genombrott inom automatisk översättning. I ett halvsekel har forskare försökt skapa datorprogram för översättning mellan naturliga språk – utan att lyckas. De misslyckade försöken har i huvudsak byggt på teorier från språkvetenskapen. Google Translate har en annan ansats, baserad på statistiska analyser av dokument, som översatts av professionella översättare. Översättningarna från Google Translate är inte perfekta men ofta fullt begripliga och användbara, när man inte har tid eller råd att anlita professionella översättare. För detaljer om metoder och algoritmer i Google Translate, gör en Google-sökning på ”google translate methodology” eller läs Wikipedias artikel: https://en.wikipedia.org/wiki/Google_Translate

Konsumentprisindex (KPI). PriceStats-projektet, se AAPOR (2015), använder prisdata från Internet för att framställa konsumentprisindex som alternativ till den traditionella metoden, som innebär att man låter fältarbetare hämta in prisuppgifter från fysiska butiker. Den nya metoden är billigare och snabbare och ger resultat som i huvudsak överensstämmer med traditionellt framtagna KPI; se Figur 1, som finns här i grått och i färg på sidan 27.

Google Flu Trends använder statistiska analyser av Google-sökningar avseende influensasymptom för att förutse influensaspridning. Google Flu Trends förutsåg influensans spridning ca en månad snabbare än traditionella epidemiologiska metoder. Efter en tids användning började algoritmerna dock ge sämre resultat än tidigare. För detaljer om metoden, algoritmerna och kvalitetsproblemen, se Wikipedias artikel i ämnet: https://en.wikipedia.org/wiki/Google_Flu_Trends.

Definitioner och begrepp

”Big Data” avser själva datamängderna och deras hantering. ”Predictive Analytics” är en klass av statistiska analysmetoder som ofta används i kombination med ”Big Data”.

”Big Data” karakteriseras ofta i litteraturen med tre V: ”Volume”, ”Velocity” och ”Variety”:

Volume: Stora datamängder hämtas från Internet och från operativa system som affärssystem, offentliga administrativa system och övervakningssystem, t.ex. system för trafikövervakning, med hjälp av nya datainsamlingsmetoder som sociala media och sensorer (”Internet of Things”). Datamängderna kan lagras på traditionellt sätt i databaser och datalager men kan även analyseras ”i flykten” för att ge feed-back till styrningen av de processer, som de genereras från.

Velocity: Data genereras ofta fortlöpande som s.k. strömmande data (”streaming data”) av händelser som inträffar med hög frekvens och som måste tas om hand med hög hastighet.

Variety: Många olika slags data förekom-

De flesta statistiker har nog åtminstone hört talas om ”Big Data”, och kanske något färre också har hört det kombineras med ”Predictive Analytics”. Men vad är egentligen innebörden av dessa termer, och representerar de något i grunden nytt, eller är de bara nya varianter av begrepp som redan är kända från områden som beslutsstödssystem, ”data mining” och ”business intelligence”?

“Big Data” och “Predictive Analytics” – modeord eller …?

mer och måste ofta hanteras i kombination med varandra: mer eller mindre strukturerade data, fritextdata, bilder, ljud, videos m.m. Hanteringen kräver nya typer av databaser utöver traditionella SQL-databaser.

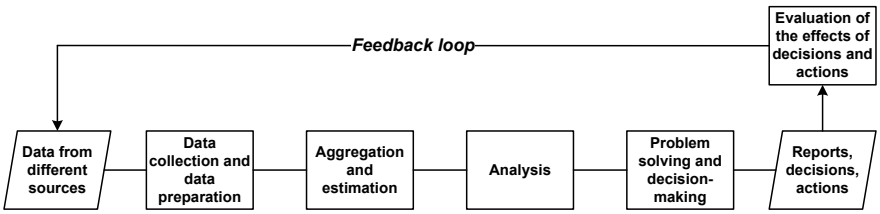
Man kan skilja mellan upphittade data (”found data”) och tillverkade data (”made data”). Vid statistiska undersökningar och experiment utformas datainsamlingen och efterföljande processer av användare, forskare och statistiker som beställer och genomför undersökningarna och experimenten. Data ”tillverkas” alltså på ett medvetet och kontrollerat sätt. När man använder sig av data från existerande källor, t.ex. administrativa register, affärssystem, övervakningssystem eller sociala medier, så har undersökaren ingen kontroll över hur data genereras eller kvalitetssäkras, utan får acceptera ”upphittade data” som de är och eventuellt komplettera med egna kontroller och bearbetningar för att i görligaste mån säkra kvaliteten.

Beteckningen ”Predictive Analytics” används för en klass av statistiska metoder som ofta används för att upptäcka och analysera statistiska samband i ”Big Data”. Metoderna kan delas in i två huvudkategorier:

- Metoder som bygger på korrelation och regression
- Metoder som bygger på s.k. ”machine learning”, t.ex. neurala nätverk och mönsterigenkänning

Trots ordet ”predictive” i ”Predictive Analytics” används metoderna även i andra sammanhang än för att göra förutsägelser och prognoser. Några vanliga tillämpningsområden är:

- Riskbedömningar i samband med kreditgivning och försäkringsavtal
- Bedömningar av olika kundkategoriers potential i samband med marknadsföring
- Bedömningar av säkerhetsrisker och risker för bedrägerier



Figur 2. Huvudprocesser vid analytisk användning av information. Från Sundgren (2015).

Vad är nytt?

”Big Data” och ”Predictive Analytics” kan ses som det senaste steget i en utvecklingskedja som startade för flera årtionden sedan, och som inkluderar:

- Beslutsstödssystem, ”Decision Support Systems” (DSS)
- Expertsystem
- ”Data Mining”
- ”Business Intelligence” (BI)

Alla dessa metoder kan beskrivas med hjälp av en grundläggande processmodell, se figur 2.

Det som är nytt med ”Big Data” är bl.a. det som karakteriseras med de tre V:na: datavolymerna, den hastighet, varmed data genereras och måste tas omhand, samt mångfalden av datatyper, ofta i kombination med varandra. Internet, strömmande data och datagenererande sensorer (”Internet of Things”) har härvidlag skapat radikalt nya förutsättningar. Nya verktyg för datahanteringen krävs.

En annan viktig nyhet är frikopplingen av analyserna från domänspecifika teorier och modeller. En fundamental tes inom traditionell statistik är att statistiska samband som sådana inte kan bevisa några orsakssamband såvida de inte kombineras med domänspecifika teorier och modeller. De framgångsrika applikationerna av ”Big Data” och ”Predictive Analytics” visar emellertid att man åtminstone ibland kan få fram mycket användbara kunskapsresultat och tillämpningar utan att nödvändigtvis behöva bygga på orsakssamband.

Referenser

AAPOR (2015). AAPOR Report on Big Data.

https://www.aapor.org/getattachment/EducationResources/Reports/BigDataTaskForceReport_FINAL_2_12_15_b.pdf.aspx

Anderson, C. (2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Wired Magazine. http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory

Evelson, B. (2015). BI and Big Data: Same or Different - An Approach To Converge The Worlds of Big Data And BI. Information Management Blogs, http://www.information-management.com/blogs/Business-Intelligence-Big-Data-Same-Different-10026730-1.html

Gold, K. (2011). Norvig vs. Chomsky and the Fight for the Future of AI. http://www.tor.com/2011/06/21/norvig-vs-chomsky-and-the-fight-for-the-future-of-ai/

Kuhn, T. S. (1962). The structure of scientific revolutions. University of Chicago Press. http:// projektintegracija.pravo.hr/_download/repository/Kuhn_Structure_of_Scientific_Revolutions.pdf

Norvig, P. (2012). Colorless green ideas learn furiously: Chomsky and the two cultures of statistical learning. Significance, August 2012, p 30-33.

http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2012.00590.x/epdf

Ogden, C.K. et. al. (1923, 1956) The Meaning of Meaning: A study of the Influence of Language upon Thought and of the Science of Symbols. Harcourt Brace, New York.

http://editura.mttlc.ro/carti/55_Charles_Ogden_The_Meaning_of_Meaning_volume_one.pdf

Russell, S., & Norvig, P. (1995). Artificial intelligence: a modern approach. Prentice-Hall, Englewood Cliffs. http://aima.cs.berkeley.edu/

Significance (2012). Special Issue: Big Data. American Statistical Association and Royal Statistical Society. Wiley Online Library. http://onlinelibrary.wiley.com/doi/10.1111/sign.2012.9.issue-4/issuetoc

Sundgren, B. (2015). The concept of information. Pro Libera Scio.

Sundgren, B. (2016). Big Data and predictive analytics - a scientific paradigm shift? Pro Libera Scio.

Turing, A. (1950). Computing machinery and intelligence. Mind 49: 433-460.

http://cogprints.org/499/1/turing.HTML

Referenser till webbplatser:

Bo Sundgren. https://sites.google.com/site/bosundgren/. Bibliografiska uppgifter och fritt nedladdningsbara publikationer.

Pro Libera Scio. https://sites.google.com/site/proliberascio/. Artiklar med anknytning till e-samhället. Open access. ISSN 2001-8932.

Wikipedia. https://en.wikipedia.org/wiki/Google_Flu_Trends. Artikel om Google Flu Trends.

Wikipedia. https://en.wikipedia.org/wiki/Google_Translate. Artikel om Google Translate.

»»» "BIG DATA" OCH "PREDICTIVE ...

Paradigmskifte?

Thomas Kuhn (1962) införde begreppen ”paradigmskifte” och ”vetenskaplig revolution” för att beskriva omvälvande förändringar i vetenskapliga teorier, förändringar som också påverkar vår världsbild och verklighetsuppfattning, t.ex. när världsbilden med jorden i centrum ersattes av en världsbild med solen i centrum, eller när Einstein ersatte Newtons mekanik med nya teorier, som tills vidare anses vara sanna. (Alla teorier visar sig förr eller senare vara falska eller ofullständiga.)

Man kan fråga sig om ”Big Data” och ”Predictive Analytics” verkligen utgör ett paradigmskifte. Vissa forskare anser detta, både sådana som bejakar användningen av statistiska analysmetoder utan koppling till domänspecifika teorier och sådana som fördömer sådana användningar, inkluderande både framstående statistiker som Gary King vid Harvard University och domänteoretiker som språkforskaren Noam Chomsky.

Några av dem som driver och bejakar utvecklingen har drabbats av ett visst övermod, t.ex. AI-giganten Peter Norvig – se Russel & Norvig (1995, 2010), Norvig (2012) – och tidskriftsredaktören och debattören Chris Anderson – se Anderson (2008). Den senare utropar:

• “All models are wrong, but some are useful.” So proclaimed statistician George Box 30 years ago, and he was right. But what choice did we have? Only models … seemed to be able to consistently, if imperfectly, explain the world around us. Until now. …

• This is a world where massive amounts of data and applied mathematics replace every other tool ... Out with every theory of human behaviour, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with

BIG DATA

Hur ska man skriva ”big data”?

■ I amerikanska rapporter verkar praxis vara att man skriver ”Big Data”, med stora begynnelsebokstäver (versaler) i båda orden. I svenska texter har man i stor utsträckning övertagit detta skrivsätt. Men vid kontakt med Språkrådet har jag fått rekommendationen att i svensk text skriva ”big data”, alltså med små begynnelsebokstäver (gemena). Så skriver man också i Nationalencyklopedins nätupplaga. I framtiden får vi väl försöka vänja oss vid att skriva ”big data”. Men det är för sent att göra några ändringar i det här numret av Qvintensen.

JW

unprecedented fidelity. With enough data, the numbers speak for themselves. …

• Scientists are trained to recognize that correlation is not causation, that no conclusions should be drawn simply on the basis of correlation between X and Y (it could just be a coincidence). Instead, you must understand the underlying mechanisms that connect the two. Once you have a model, you can connect the data sets with confidence. Data without a model is just noise. …

• There is now a better way. We can stop looking for models. We can analyse the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.

• The new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world. …

• There's no reason to cling to our old ways. It's time to ask: What can science learn from Google?

Många exempel på kritiska synpunkter, inte minst vad gäller kvalitetsfrågor, redovisas i AAPOR (2015) och Sundgren (2015, 2016), där det också finns referenser till många artiklar, som behandlar dessa frågor.

Språkvetaren Noam Chomsky invänder att ingen verklig kunskap om ett ämne kan uppnås utan välunderbyggda, ämnesspecifika teorier och modeller. ”Statistisk datamuddring” kan på sin höjd vara praktiskt användbar i vissa fall men bidrar inte till verklig förståelse. Statistikern Gary King tillägger att man inte kan vara säker på någonting, om man inte känner till vilka variabler som är viktiga för ett problem och kan mäta dessa på ett tillförlitligt sätt. Alla förutsättningar är dessutom föränderliga över tiden.

Genomgripande förändringar av metoder och teknik – ”disruptive innovations”

Det är omstritt om ”Big Data” och ”Predictive Analytics” innebär ett paradigmskifte, och om det i så fall är ett önskvärt sådant. Både entusiaster och kritiker förefaller vara överens om att de nya metoderna kan vara användbara, även om man inte riktigt förstår dem, och även om de inte återspeglar hur människor tänker. Kanske bör man därför snarare se ”Big Data” och ”Predictive Analytics” som en s.k. ”disruptive innovation”, en genomgripande förändring av metod och teknik. Vi har sett många sådana under det senaste århundradet, t.ex. kylskåpet som ersatte isutkörare, digital teknik som ersatte mekanik, digitala massmedia och Internet som ersatte pappersmedia, grammofonskivor, videofilmer och traditionell distribution av dessa media.

Syntes

De vetenskapliga striderna kring “Big Data” och ”Predictive Analytics” kan ses som ytterligare en fas i kampen mellan empirister och rationalister, mellan dem som tror att mänsklig kunskap främst baseras på empiriska observationer och dem som betonar förnuftets och teoriernas betydelse. De senaste århundradenas exempellösa framgångar inom naturvetenskapen baseras på en syntes av dessa synsätt: teorier byggs utifrån tillgängliga observationer och testas därefter genom kontrollerade experiment.

Måhända får vi en ny syntes mellan ”dumma” statistiska analyser baserade på enorma mängder av empiriska data och kommande teorier som kan få oss att förstå dessa statistiska metoder bättre, varför de kan fungera, och vilka fallgropar de inrymmer. Utvecklingen kan säkert också bidra till framväxten av mera realistiska och empiriskt välunderbyggda teorier inom olika vetenskapliga domäner.

Det finns anledning att påminna om Alan Turings kloka ord i Turing (1950), där han föreslår att frågan ”Can machines think?” ersätts med frågan ”Can machines do what we (as thinking entities) can do?” Kanske kan datorstödda system laddade med Big Data hjälpa oss uppnå nya resultat genom att använda metoder som är fundamentalt annorlunda än människors sätt att tänka, modellera, analysera, resonera och härleda ny kunskap.

BO SUNDGREN, PROFESSOR
HÖGSKOLAN DALARNA

MINNESORD ÖVER JAN LANKE

Jan Lanke.

Jan Lanke, professor i biostatistik i Lund, avled den 15 mars 2016 i en ålder av 81 år.

Jan Lanke föddes i Lund, och var staden trogen hela livet, med avstickare till Sheffield, UK, och Waterloo, Canada, i samband avhandlingsskrivandet. Med en licentiatexamen i matematik blev han nära medarbetare till docenten och professorn Gunnar Blom när ämnet matematisk statistik återuppstod i Lund 1959. Jan satte sin karakteristiska prägel på utbildningen, först vid naturvetenskapliga fakulteten och från 1962 vid den nya tekniska högskolan där han blev universitetslektor 1964. Det noggranna upplägget av undervisningen och det briljanta framförandet vid svarta tavlan slogs av ingen. Han blev en förebild som föreläsare både till stil och till innehåll.

I början på 1960-talet utbildade Jan Lanke sig själv till statistiker genom att undervisa! Med sin matematiska bakgrund saknade han förutfattade meningar om vad ämnet matematisk statistik borde innehålla och han lät i stället sin nyfikenhet bestämma. Hans föreläsningar på andra terminen (”tvåbetygskursen”) gav en bred och mycket precis översikt över (tillämpad) sannolikhetsteori och processer, försäkringsmatematik, sampling, varians- och regressionsanalys, försöksplanering. Allt mycket verklighetsanknutet. Hans förmåga att kombinera matematisk precision med en översättning mellan det matematisk statistiska språket och verkligheten inspirerade många studenter att fortsätta med ämnet, däribland flera av undertecknarna.

I själva verket inspirerade han sig själv till att bli doktorand i matematisk statistik och han disputerade 1975 på en avhandling om sampling, ett då mycket försummat forskningsområde inom matematisk statistik. Valet av avhandlingsämne var ingen slump – legendaren Tore Dalenius insåg vad som var Jans styrka som forskare och övertygade honom om att survey sampling var ett lämpligt och givande forskningsområde.

Avhandlingen ledde så småningom till docentur i statistik vid samhällsvetenskapliga fakulteten, vilken 1989 omvandlades till en professur i biostatistik med finansiering från fyra håll: tre fakulteter, samhällsvetenskaplig, medicinsk och odontologisk, och en institution, matematisk statistik. Från 1980-talet arbetade Jan huvudsakligen med medicinska tillämpningar. Värdefulla insatser gjordes bl a vid studiet av psykiska sjukdomar i Lundbykohorten och i stora och inflytelserika studier av blodtryckmediciner, studier som publicerades i The Lancet. Aktiv in i det sista, arbetade Jan med flera statistiska projekt, som utveckling av metoder för analys av överlevnad efter stamcellstransplantation.

Jan Lanke var ingen vän av uppmärksamhet kring sin person men hade inget emot att påverka samhället. En förändring i den nya regeringsformen som trädde i kraft 2015 väntas leda till en bättre proportionalitet vid fördelning av riksdagsmandaten. Ändringen är underbyggd med Jans noggranna studier av det svenska valsystemet.

Jan Lanke betydde mycket som samarbetspartner och som bollplank genom sitt genuina intresse för statistiska problem och sitt prestigelösa och öppna sätt. Hans analytiska läggning och skolning var en stor styrka i karriären. Vi är tacksamma att han delade med sig till oss.

HARALD ANDERSON, LARS HOLST, ULLA HOLST,
GEORG LINDGREN, KRZYSZTOF NOWICKI

En statistisk paradox

Inledning

Den här artikeln handlar om något som brukar kallas för ”Benfords lag”. När man har data i form av en stor samling tal visar det sig ofta (men inte alltid) att första siffran i dessa tal mycket ofta är något av talen 1, 2 eller 3, medan talen 7, 8 och 9 förekommer betydligt mindre ofta som första siffra. Fenomenet blev mer allmänt känt när Benford (1938) skrev om detta; därav benämningen ”Benfords lag”. Ibland menas med ”Benfords lag” en speciell sannolikhetsfördelning för talets första siffra, när man har en stor samling tal. Benford var inte den förste som benhandlade det här. Redan 57 år tidigare hade Newcomb (1881) gjort samma iakttagelse som Benford.

Att inom vetenskapen namnge en lag, ett teorem eller överhuvudtaget ett begrepp efter en specifik person är alltid förenat med stora risker. Stigler (1980) introducerade “*Stigler's Law of Eponymy*”, som kort kan formuleras “*No scientific discovery is named after its original discoverer*”. Stigler själv var övertygad att även hans egen lag följer denna regel. Enligt Atle Selberg ”*sa André Weil engang at hvis noe har fått navnet etter noen person, så hadde vedkommende som regel veldig litt med med det och gøre*” (Baas, N. A. & Skau, C. F.; 2008). Weils uttalande gjordes självfallet långt före Stigler.

Några resultat rörande Benfords lag presenteras här utan matematiska bevis. En mera mångsidig artikel över Benfords lag med en fullständigare litteraturförteckning utgör Fellman (2014). För svenska läsare kan det vara intressant att veta att Lennart Råde (1964) skrev en artikel om Benfords lag med anknytning till Poincarés roulettproblem.

Empirisk beskrivning av Benfords lag.

Vi har redan nämnt att Benford och före honom Newcomb konstaterade att den inledande siffran i tal ur stora datamaterial med stor variation kunde uppvisa en ovanlig fördelning med övervikt för de låga siffrorna 1 - 3. Denna upptäckt gjorde de utifrån iakttagelsen att deras logaritmtabeller var ojämnt slitna. (Det här var ju före datorernas tid.) Början av tabellen, där talen började med 1, 2 och 3, var mera sliten än slutet, där talen började med 7, 8 och 9. Båda konstaterade att av allt att döma hade de data de arbetade med oftare något av de låga talen som första siffra och mera sällan de höga.

Forskare som studerat Benfords lag empiriskt har försökt identifiera specifika datamaterial som uppvisat god överensstämmelse med denna lag. Man har också undersökt olika talföljder för att identifiera sådana för vilka lagen gällde (benfordföljder) och sådana för vilka lagen inte gäller. Dylika problem har ingående analyserats av Berger & Hill (2015).

Benfords teoretiska fördelning

I denna artikel använder vi genomgående beteckningen ”log” för de briggska logaritmerna med basen tio. Benfords teoretiska fördelning definieras nu såsom

$$P(D = d) = \log(d + 1) - \log(d);\ d = 1, 2, \ldots, 9$$

där D är den observerade första siffran hos ett tal. Det följer härav att

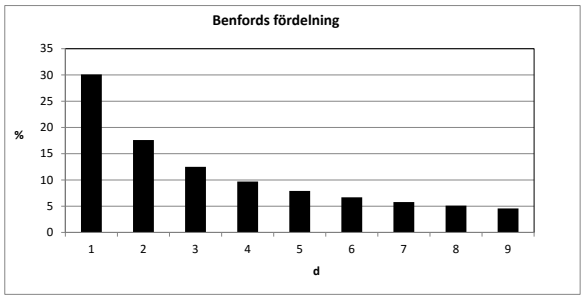
$$P(D \leq d) = \log(d+1);\ d = 1, 2, \ldots, 9$$

Fördelningen presenteras i Tabell 1. Notera i tabellen att över 30 procent av talen börjar med en etta och endast knappa 5 procent med en nia. Andelen tal som börjar med ett, två eller tre utgör sammanlagt drygt 60 procent av alla.

Tabell 1. Den teoretiska fördelningen för den första siffran (*D*) hos tal som är fördelade enligt Benfords lag. Notera att log(1)=0 och att log(10)=1

d	$P(D = d)$	*%*	$P(D \leq d)$	*%*
1	$\log(2) - \log(1)$	*30,10*	$\log(2)$	*30,10*
2	$\log(3) - \log(2)$	*17,61*	$\log(3)$	*47,71*
3	$\log(4) - \log(3)$	*12,49*	$\log(4)$	*60,21*
4	$\log(5) - \log(4)$	*9,69*	$\log(5)$	*69,90*
5	$\log(6) - \log(5)$	*7,92*	$\log(6)$	*77,82*
6	$\log(7) - \log(6)$	*6,69*	$\log(7)$	*84,51*
7	$\log(8) - \log(7)$	*5,80*	$\log(8)$	*90,31*
8	$\log(9) - \log(8)$	*5,12*	$\log(9)$	*95,42*
9	$\log(10) - \log(9)$	*4,58*	$\log(10)$	*100,00*
Totalt	*1*	*100*		

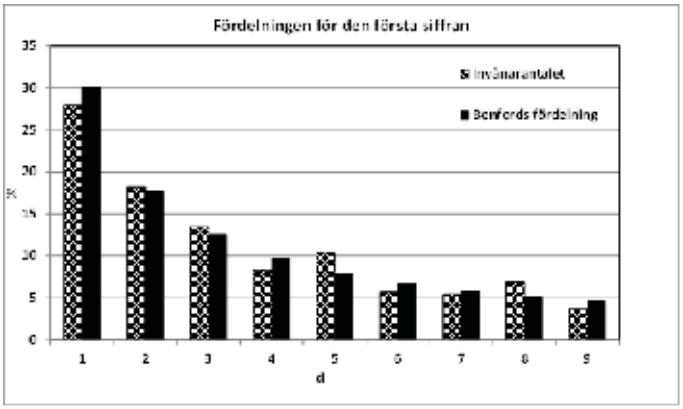
Figur 1 ger en grafisk bild av fördelningen.



Figur 1. Benfords teoretiska fördelning för den första siffran.

Befolkningstalen i Finlands kommuner år 1982

Nu introducerar jag ett litet numeriskt datamaterial, som jag använt i olika sammanhang och som är tillräckligt stort för att egenskaper hos Benfords lag kan spåras och belysas (Fellman, 2014). År 1982 var Finland indelat i 461 kommuner, och befolkningstalen varierade mellan 147 i Sottunga på Åland och 484 260 i Helsingfors. Datamaterialets storlek och stora variation är tillräckliga för att dessa data kan utgöra ett acceptabelt exempel på Benfords lag. I Figur 2 jämförs den empiriska fördelningen över invånarantalet med Benfords teoretiska fördelning. Överstämmelsen är uppenbar. Då vi statistiskt testar diskrepansen erhåller vi $\chi^2 = 10{,}42$, med 8 frihetsgrader vilket innebär ingen statistiskt signifikant avvikelse från Benfords lag.



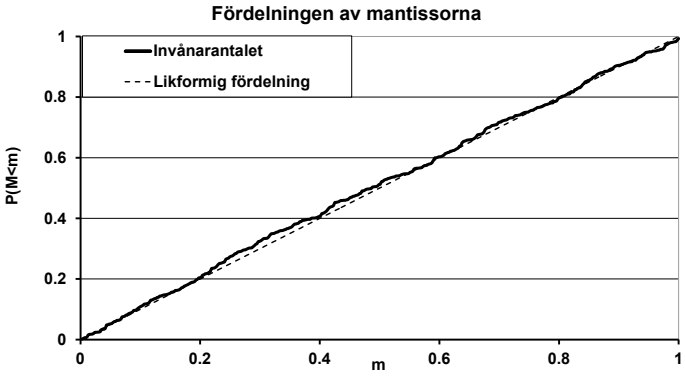
Figur 2. Jämförelse mellan kommunernas befolkningstal fördelade enligt första siffran och Benfords fördelning. Man kan konstatera en god överensstämmelse som kommenteras i texten.

Fördelningen för mantissan i den briggska logaritmen

En briggsk logaritm består av två delar: (1) heltalsdelen, *karakteristikan*, som anger storleksordningen av talet och utgör antalet gällande siffror före decimalkommat minskat med ett; och (2) decimaldelen, *mantissan*, som är entydigt bestämd av sifferkombinationen i talet.

Man kan visa att om det ursprungliga datamaterialet följer Benfords lag, så är mantissorna hos det logaritmerade datamaterialet likformigt fördelade över intervallet (0, 1). Ovannämnda resonemang kan också genomföras i motsatt riktning. Om mantissan är likformigt fördelad är alltså utgångsdata benfordfördelade. Utan formellt bevis gavs detta resultat redan av Newcomb.

I Figur 3 presenteras fördelningen för mantissorna i de logaritmerade invånarantalen i de finska kommunerna, och det visar sig att överensstämmelsen med en likformig fördelning är mycket god.



Figur 3. Fördelningen för mantissorna för invånarantalet. Den goda anpassningen till den räta linjen stöder antagandet om en likformig fördelning för mantissan och Benfords lag för fördelningen av invånarantalen.

Villkor för Benfords lag

Följer alla datamaterial Benfords lag? Svaret är ett entydigt nej. Tänk till exempel på kroppslängden i centimeter för en årskull rekryter. Materialet är omfattande, men absoluta majoriteten av måtten börjar med ett och en liten minoritet med två. Andra möjligheter finns inte. Benford noterade att olika datamaterial kunde ha olika god anpassning till lagen, men att materialen sammantagna till ett enda datamaterial gav en bättre anpassning. Ser man till de matematiska bevis och analyser av empiriska resultat som presenterats i litteraturen, framgår att datamaterial följer Benfods lag om följande villkor gäller:

1. materialet skall vara tillräckligt stort, och
2. de enskilda talen skall uppvisa en så stor variation att de har olika storleksordning.

Smith (2007) framhöll att variationen på den aritmetiska skalan skall vara av högre storleksordning än på den logaritmiska skalan.

Fördelningen för de efterföljande siffrorna

Går man till fördelningen för de efterföljande siffrorna i talen, finner man att det sker en snabb utjämning i riktning mot den likformiga fördelningen, samtidigt som modellens teoretiska formler blir mera komplicerade. De matematiska detaljerna ges inte här. Redan Newcomb gav fördelningen för den andra siffran. I Tabell 2 ges en jämförelse mellan fördelningen för den första, den andra, den tredje och den fjärde siffran. Talen i de olika positionerna är inte statistiskt oberoende. Notera den snabba konvergensen mot en likformig fördelning.

Tabell 2. Den procentuella fördelningen för *D* för olika ordningsnummer *s*=1, 2, 3 och 4.

s	0	1	2	3	4	5	6	7	8	9	Totalt
1		30,10	17,61	12,49	9,69	7,92	6,69	5,80	5,12	4,58	100
2	11,97	11,39	10,88	10,43	10,03	9,67	9,34	9,04	8,76	8,50	100
3	10,18	10,14	10,10	10,06	10,02	9,98	9,94	9,90	9,86	9,83	100
4	10,02	10,01	10,01	10,01	10,00	10,00	9,99	9,99	9,99	9,98	100

Tillämpningar av Benfords lag

Nigrini och Woods (1995) har använt Benfords lag för att studera fördelningen av befolkningen enligt 1990 års folkräkning i USA. De konstaterar god överensstämmelse och finner det plausibelt att man med Benfords lag skulle kunna prognostisera framtida befolkningstal.

Sandron (2002) studerar befolkningstalen i 198 olika länder och konstaterar en god överstämmelse med Benfords lag. Likaså analyserar han sambandet mellan Benfords lag och befolkningstillväxten.

Ley (1996) undersöker hur väl avkastningar på aktiemarknader följer Benfords lag. Han studerar ettdygnsavkastningarna av *Dow-Jones Industrial Average Index* och *Standard&Poor´s Index* och konstaterar att de rätt väl följer Benfords lag.

Nigrini och Mittermaier (1997) undersöker om lagen kan användas för att spåra bokföringsfusk. Grundidén är att bokföringsdata borde följa Benfords lag i och med att data är omfattande och de enskilda talen visar stor variation. Däremot skulle fusk kunna skapa diskrepanser från Benfords lag. De presenterar statistiska test med vilka man kan se hur väl bokföringsdata överensstämmer med Benfords lag.

el Sehity et al. (2005) diskuterar prissättningen på varor. De konstaterar att prissättningen av försäljningspsykologiska skäl inte kan förväntas vara benfordfördelad. De studerar hur övergången till euro rubbar prissättningen och på detta sätt eliminerar till sin omfattning och variation inflytandet av den psykologiska prissättningen. Detta skulle kunna leda till bättre anpassning till Benfords lag.

Gonzalez-Garcia och Pastor (2010) prövar möjligheten att med Benfords lag testa kvaliteten hos makroekonomiska data. De noterar att största delen av dylika data satisfierar Benfords lag. Författarna konstaterar dock att konformiteten med Benfords lag bör jämföras med andra kvalitetstest. De utvecklar studierna hos Nigrini och Mittermaier (1997) genom att utvidga mängden makroekonomiska data så att de täcker de centrala sektorerna av ekonomin. Ytterligare utvidgas studien till att omfatta 80 länder indelade i sex sektorer. Deras resultat är att de makroekonomiska data satisfierar Benfords lag. Författarna konstaterar dock att ett förkastande av Benfords lag inte behöver vara en indikator för låg kvalitet hos data. Övertygande konstaterar de att en god överensstämmelse mellan testresultat och datakvalitet inte kan etableras. Deras analyser visar att avvikelse från Benfords lag kan, oberoende av datakvaliteten, betingas av markanta ekonomiska fluktuationer samt strukturförändringar hos data.

Diskussion

Benfords lag är inte endast ett kuriosum, utan forskare har visat den stort intresse både vid teoretiska analyser och vid försök att tillämpa den på olika områden. På olika sätt och under olika antaganden har man kommit fram till denna fördelning. Genomgående kan man konstatera att oavsett om studierna är teoretiska eller empiriska, så är det centrala intresset att se hur generell denna fördelning är och vilka villkor ett datamaterial skall uppfylla för att Benfords lag skall gälla.

JOHAN FELLMAN
PROFESSOR EMERITUS I STATISTIK
SVENSKA HANDELSHÖGSKOLAN I HELSINGFORS

Litteratur

Baas, N. A. & Skau, C. F. (2008). Selbergintervjuet del 1- Matematisk oppvekst. *Nordisk matematisk tidskrift (Normat), 56 (1)*:5-23.

Benford, F. (1938). The law of anomalous numbers. *Proceedings of American Philosophical Society, 78*:551-572.

Berger, A. & Hill, T. P. (2015). An introduction to Benford´s law. Princeton University Press. viii+248 pp.

el Sehity, T., Hoelzl, E. & Kirchler, E. (2005) Price developments after a nominal shock: Bedford´s law and psychological pricing after the euro introduction. *Intern. J. of Research in Marketing. 22*:471-480.

Fellman, J. (2014). The Benford paradox. Journal of Statistical and Econometric Methods,3(4):1-20.

Gonzalez-Garcia, J. & Pastor, G. (2010). *Benford´s law and macroeconomic data quality*. IMF Working Paper, 20 pp.

Ley, E. (1996). On the peculiar distribution of the U.S. stock indices digits. *The American Statistician, 50*:311-313.

Newcomb, S. (1881). Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics, 4*:39-40.

Nigrini, M. J. & Mittermaier, L. J. (1997). The use of Benford´s law as an aid in analytical procedures. *Auditing: A journal of Practice & Theory, 16*:52-67.

Nigrini, M. J. & Woods, W. (1995). *Assessing the Integrity of Tabulated Demographic Data*. Preprint. Univ. Cincinnati and St. Mary's Univ. Citerad av Hill (1995b).

Råde, L. (1964). Den första siffran och Poincarés rouletteproblem. Nordisk matematisk tidskrift, 12 (1-2): 45-52.

Sandron, F. (2002). Do populations conform to the law of anomalous numbers?. *Population, 57*:755-762.

Smith, S. W. (2007). Explaining Benford´s law. Chapter 34 (701-722) ingår i *The Scientist and Engineer´s Guide to Digital Signal Processing*.

Stigler, S. M. (1980). Stigler's law of eponymy. *Transactions of the New York Academy of Sciences, Series II Vol. 39*:147-158.

ORDFÖRANDEN HAR ORDET

»Texten var nog inte så kontroversiell, men jag tror mig veta att det finns starka åsikter ... »

Det händer väldigt sällan att jag får någon specifik respons på den här spalten, men mina reflektioner senast angående design- kontra modellbaserade metoder inom urvalsteorin blev uppskattade av åtminstone en läsare. Texten var nog inte så kontroversiell, men jag tror mig veta att det finns starka åsikter i frågan om vilken ansats som är lämpligast, särskilt i dessa tider med höga bortfallsandelar i urvalen.

Resultatet av det brittiska parlamentsvalet 2015 blev överraskande jämfört med opinionsundersökningarna före valet. I en stor majoritet av dessa skattades skillnaden i väljarstöd mellan partierna Conservatives och Labour till mindre än två procentenheter. I många fall väntade man sig rent av ”dött lopp”: 34% för båda partierna. Men i valet fick Conservatives 37,8% av rösterna mot 31,2% för Labour.

Nu föreligger en rapport [1], skriven av en kommitté ledd av Patrick Sturgis, universitetet i Southampton. Den 120-sidiga rapporten försöker klargöra varför de flesta opinionsundersökningarna inför detta val i ett historiskt perspektiv var ovanligt dåliga på att förutsäga valresultatet. Kommittén fick tillgång till ”rådata” från flera opinionsinstitut och kunde därigenom kontrollera gjorda beräkningar. Dessa visade sig alla vara korrekt gjorda. Det gick heller inte att påvisa någon systematisk skillnad i resultaten mellan olika datainsamlingssätt (”modes”).

Variationerna för skattningarna av väljarstödet för Conservatives blev under den sista valkampanjveckan allt mindre ju närmre valdagen undersökningarna gjordes. En möjlig orsak till ett sådant fenomen kan vara så kallad flockmentalitet (”herding”) bland instituten. I dess mest extrema form innebär detta att resultat som kraftigt avviker från tidigare mätningars (egna och/eller andras) inte ens publiceras. Rapportens slutsats efter redovisade analyser är att tendenser till flockmentalitet är påvisbara givet stickprovsstorlekarna.

En uppenbar möjlig orsak till att valresultatet kraftigt skilde sig från vad opinionsundersökningarna visade är att många väljare i sista stund ändrade sig. I rapporten argumenteras det dock för att en sådan förändring av opinion (”late swing”), om den alls fanns, endast var marginell.

Huvudorsaken till diskrepansen mellan opinionsundersökningarnas resultat och valresultatet var i stället, enligt rapporten, att urvalen inte var representativa för den verkliga opinionen. Labour-anhängarna var systematiskt överrepresenterade i urvalen och justeringar (omviktningar) mildrade inte den effekten i tillräckligt hög grad.

Opinionsundersökningarna byggde på kvoturval, och i rapporten finns också en diskussion om sådana urvalsmetoder kontra rena sannolikhetsurval. Kort efter valet gjordes för övrigt två stora undersökningar där individer valdes med slumpmekanism från ett adressregister. Trots höga bortfallsandelar på nära 50 %, erhölls skattade partisympatiandelar som låg ganska nära valresultatet och mycket närmre än de föregående opinionsundersökningarnas.

Rapporten är ett läsvärt dokument både för praktiker och metodintresserade, och läsaren ges möjligheter att begrunda argument för eller emot olika tillvägagångssätt för urval och skattningar.

PER GÖSTA ANDERSSON

[1] Sturgis, P. Baker, N. Callegaro, M. Fisher, S. Green, J. Jennings, W. Kuha, J. Lauderdale, B. and Smith, P. (2016) Report of the Inquiry into the 2015 British General Election Opinion Polls, London: Market Research Society and British Polling Council.

Tävling om bästa uppsats

Kom ihåg att anmäla goda uppsatser inom surveyområdet till tävlingen om bästa uppsats!

Uppsatsen ska ha lagts fram mellan den 25 januari 2016 och den 24 januari 2017.

Nominerade uppsatser bedöms utifrån graden av originalitet och nyskapande, relevans i metodansats med avseende på vald problemställning, genomförande och klarhet i framställningen.

Vinnande uppsats utses av en jury bestående av styrelsen för Surveyföreningen. Prissumman är på 7 500 kronor och överlämnas i samband med Surveyföreningens årskonferens den 17 mars 2017 i Lund. Syftet med tävlingen är att stimulera intresset för samhällsstatistisk undersökningsmetodik bland studenter och handledare i grundutbildningen.

Skicka din nominering senast den 1 februari 2017
till Per Gösta Andersson på per.gosta.andersson@stat.su.se

Workshop om surveymetodik

Den 26-27 maj 2016 anordnade Surveyföreningen och Statistiska institutionen vid Stockholms universitet gemensamt en workshop med titeln ”Learn about new methods needed in a time of hardening survey climate”. Workshopen ägde rum på Stockholms universitet med cirka fyrtio deltagare, huvudsakligen från Sverige. Presentationer hölls av föredragshållare från Sverige, Estland, Finland, och Norge. Det övergripande temat var urval, datainsamling och estimation i ett läge när man har stora bortfallsproblem. Tonvikten låg på tillämpningar inom samhällsvetenskap och officiell statistikproduktion. Bidragen var huvudsakligen tekniska till sin natur.

I ett inledningsföredrag uppehöll sig Carl-Erik Särndal vid så kallat adaptivt urval som ett sätt att handskas med bortfallsproblemet. Länge har ”klassiskt” sannolikhetsurval setts som eftersträvansvärt. Det är fortfarande en förebild, men har blivit allt svårare att genomföra i praktiken, eftersom det förutsätter obefintligt eller mycket litet bortfall, vilket numera ofta är omöjligt att uppnå, främst av ekonomiska skäl. Här kommer adaptivt urval in i bilden som en ny teknik att handskas med bortfallsproblemet. Adaptivt urval innebär att man under pågående urval successivt modifierar urvalsdesignen i syfte att erhålla ett slutligt urval som, trots bortfall, uppfyller krav på ”representativitet” med avseende på vissa kända bakgrundsvariabler. Detta skulle innebära en besparing såtillvida att man inte behövde lägga stora resurser på ofta dyrbara ansträngningar att minska bortfallets storlek.

Workshopens övriga presentatörer och deras ämnen listas här kortfattat.

Anton Johansson, SCB: Datainsamlingsstrategier för att nå icke-svarande i SCB:s arbetskraftsundersökningar.

Gelaye Hailemichael, Stockholms universitet: En jämförande studie rörande sannolikheturval gentemot icke-sannolikhetsurval, bland annat vid webbpanelundersökningar.

Thomas Laitila, Örebro universitet och SCB: En svarsmodell vid QR ansats.

Risto Lehtonen, Helsingfors universitet: Kalibreringsmetoder för domänestimation.

Melike Alper, Statistisk sentralbyrå, Norge: En empirisk likelihood-ansats för komplexa parametrar och komplexa surveydata.

Edgar Bueno, Stockholms universitet: Jämförelse mellan stratifierat urval och pps-uval.

Kaur Lumiste, Tartu universitet: Studie av om skattningar förbättras vid försök att reducera obalans i urvalet.

Peter Lundquist, Stockholms universitet och SCB: Användning av indikatorvariabler vid adaptivt urval.

JAN WRETMAN

FMS

ORDFÖRANDEN HAR ORDET

»one of the main messages was the importance of statistical models and methods for harvesting value from big data»

On September 1-2 2016, Karolinska Institutet hosted a symposium recognising Juni Palmgren's work and vision during more than three decades of engagement in statistical science in biomedical research. The aims of the symposium were twofold, firstly to recognise Juni's broad contributions to biostatistics and, secondly, to articulate the challenges that lie ahead in the big data era.

Juni Palmgren started her career at the University of Helsinki and at the National Public Health Institute in Finland. She was recruited as Professor of Biostatistics to the Department of Mathematical Statistics at Stockholm University 1997. She held a Guest Professorship at Karolinska Institutet from 1999 and was awarded an Honorary Doctorate in Medicine MDHc 2007 for her efforts to promote Biostatistics at Karolinska Institutet. She was Secretary General for Research Infrastructure at the Swedish Research Council 2010-2015 and Finland Distinguished Professor, FiDiPro at the Institute for Molecular Medicine Finland, FIMM 2011-2015. In 2013 she was appointed full Professor of Biostatistics at Karolinska Institutet. Juni was president of FMS from 1998 to 2000.

The symposium assembled experiences from Juni's former students, postdocs and protégés across her 30 years of engagement in promoting medical biostatistics in Finland and Sweden. Ten talks were presented on a variety of topics, including genetic analyses, statistical models, causal inference, clinical drug design, nutritional epidemiology, register studies, and health policy. The span was broad, with an international perspective and a sense of opportunity and confidence. Details of the program, and the speakers biographies, are available on the symposium webpage: http://ki.se/en/meb/palmgren-symposium-2016.

The second aim was focused on Juni's vision for the future of not just statistical science in biomedical research, but biomedical research in general, as big data technologies reshape the field. The symposium addressed the scientific, political, educational and organisational challenges that lie ahead in a changing world where the role of data science is becoming increasingly important. Three international biostatisticians/data scientists gave special invited talks: Louise Ryan from Sydney, Els Goethgebeur from Ghent, and Arnoldo Frigessi from Oslo.

Especially thoughtful and inspiring was the panel discussion on 'Big Data, e-Science and e-Infrastructure', moderated by Hans-Olov Adami. The panel included besides Ryan, Gothgebeur and Frigessi also Henrik Grönberg from KI, Olli Kallioniemi from SciLifeLab and Erik Lindahl from SeRC.

Given the background of the speakers, it was not surprising that one of the main messages from the symposium was the importance of statistical models and methods for harvesting value from big data. However, the speakers also emphasised the importance of combining optimal statistical methods with appropriate computational methods. A common theme was the need for scientists with training in both statistics and computer science and the need for cross-campus collaboration.

PAUL DICKMAN

FOTO: ULF SIRBORN/BILDMAKARNA

FMS-stipendier. FMS utlyser varje år ett eller flera stipendier på ett sammanlagt belopp av högst 20 000 kr. Stipendiet finansieras primärt av intäkter från företag och institutioner som är partners till föreningen och kan sökas av medlemmar som vill främja den egna professionella utvecklingen som statistiker genom deltagande i konferenser, kurser eller dylikt. Nedan följer konferensrapporter från Hannah Bower och Sarah Burkill, båda doktorander vid Karolinska institutet, vilka tilldelades stipendium 2016.

Report on the ICPE conference 2016

The 2016 International Conference for Pharmacoepidemiology in Dublin from 24th to 28th August provided an invaluable experience for me to improve my analytical abilities, particularly with regard to health based and drug utilisation research. The methodological aspects of pharmacoepidemiological research were covered by both the course and the conference, including a diverse range of topics such as the appropriate use of propensity scores, and study design. Various options for case-control studies as well as cohort studies and the pros and cons of each design were discussed, providing valuable knowledge which can be applied to my own studies. The pre-conference courses offered an opportunity to hear new ideas and introduce concepts which will be helpful in my development as a statistician.

During my time at the conference I gained valuable experience in scientific communication and dissemination of research, as I had the opportunity to present findings on research projects I have been working on through both a poster and an oral presentation, as well as to meet with potential collaborators to discuss the possibility of future projects. Presenting the work also gave the opportunity to receive feedback and to understand how the projects can be improved and possible next stages. The oral presentation was entitled 'Pharmacological treatments preceding the diagnosis of progressive multifocal leukencephalopathy' and was presented in a session on the use of immunomodulators, and the title of the poster was 'Pharmacological treatments preceding the diagnosis of idiopathic intracranial hypertension'.

I am a PhD student based at Karolinska Institutet within the Centre for Pharmacoepidemiology. My thesis will be mostly on outcomes for multiple sclerosis patients in Sweden, but will also include a collaborative study between Sweden, Finland and Norway. The topics are quite varied, and will include a study on a potential genetic link between psychosis and multiple sclerosis. The methods for this study have not been finalized, and some of the presented topics discussed at ICPE provided insightful ideas as to how this could be done taking into account issues including the clustered nature of the data. Aside from my doctoral research, I also work on other projects including the two projects I presented at ICPE. Both these presentations were on rare diseases, and used a different approach to study design and data collection to the data I am using in my thesis. Working with a variety of subjects and using data of different designs means being aware of the strengths and weaknesses of different study designs, and the issues which should be taken into consideration when analyzing each type of data. Attendance at the ICPE conference was very useful for gaining a better understanding and greater awareness of such issues.

Sarah Burkill.

The stipend contributed substantially towards the costs of the conference and pre-conference courses, as well as the accommodation and travel costs. The stipend also helped with payments for poster printing.

SARAH BURKILL

»»» FMS-STIPENDIER

Report on the ISCB conference 2016

Thanks to the stipend I was awarded by FMS, I was able to attend and present at the 37th ISCB (International Society for Clinical Biostatisticians) annual conference in Birmingham, United Kingdom between 21st and 25th August 2016. The aim of the conference was to "provide a scientific forum for international exchange of theory, methods and applications for clinical biostatistics among biostatisticians, epidemiologists and other medical researchers".

Information about the conference can be found on their website http://www.birmingham.ac.uk/facilities/mds-cpd/conferences/iscb-2016/index.aspx .

Hanna Bower.

The conference was attended by over 600 people coming from over 40 different countries; this meant that there were many experts in a variety of different statistical areas. The conference began with a day of courses, then three days of presentations, followed by a student's day and mini-symposium. The three day conference program contained 21 different sessions with 7 sessions running parallel at a time; each session contained 5 presentations with time for questions and discussion. There were also two plenary talks by Diego Kuonen and David Spiegelhalter which were highly interesting. The session topics included Advances in statistical issues in epidemiology, Missing data, Adaptive trial design and analysis, Analysis of observational studies, Meta-analysis, and medical decision making; among many others.

I am a PhD student at the Department of Medical Epidemiology and Biostatistics at Karolinska Institutet. My research focuses on the development and application of statistical methods for population-based cancer survival studies. I have worked largely with a measure called the loss in expectation of life, which quantifies the reduction in life expectancy due to a diagnosis of cancer. Attending ISCB gave me the opportunity to present one of my current doctoral projects titled "Stratifying population life tables by socioeconomic status and comorbidity by using a control population". This work should enable the estimation of the loss in expectation of life by socioeconomic status using data available to our research group. I was pleased to receive some feedback and questions from the audience.

Since I have not attended a large, biostatistical conference before, the FMS stipend gave me an opportunity learn about other types of biostatistical research is taking place on an international level which gave me ideas about what sort of work I would like to do in the future. The whole conference was extremely enjoyable, increased my statistical knowledge in both my field and others, gave me the opportunity to present my work and get valuable feedback, and increased my network so that statistical collaboration is a higher probability for the future. I wish to thank FMS for giving me this opportunity.

HANNAH BOWER

ORDFÖRANDEN HAR ORDET

Årsmöte i föreningen

Årsmöte för Föreningen för industriell statistik ägde rum den 15 mars 2016. Vi var ca 25 deltagare, vilket kan anses bra. Jag vill tacka Love Florgård för ordförandeskap på mötet och Åsa Wikenstål för protokollgranskning, samt de medlemmar som kommit till årsmötet.

Mötet inleddes med en serie seminarier om beslutskvalitet. Kjell Borking, Preference, talade om riskhantering i samband med försäljningsprocessen hos Ericsson AB. Tomas Torstensson, Skandia, presenterade därefter utmaningar rörande datakvalitet inom livförsäkring. Lars Håkansson, professor vid Blekinge tekniska högskola, visade på nyttan med Fouriertransformationen vid ljud- och vibrationsanalyser. Seminarierna avslutades av Bengt Johannesson, Volvo Global Trucks Technology, med ett föredrag om designstrategier för att säkra uthållighetskraven för Volvo truckbromsar.

Efterföljande diskussioner och frågor visade tydligt på statistikens användbarhet som metodik, verktyg och kunskap. Jag vill tacka alla våra föredragshållare och organisatörer för ett mycket givande, intressant och minnesvärt möte.

Årsmötesförhandlingarna resulterade i att den sittande styrelsen fick ansvarsfrihet och valdes om för ytterligare ett år. I samband med omvalet utökades styrelsen med en person, Mats Franzén, som vi hälsar välkommen. Mats har under en tidigare period varit föreningens kassör. Mer information om mötet och styrelsen finns på vår hemsida Startsida Industriell Statistik.

FOTO: LINA LANDE

GÖRAN LANDE

Om terminologi för riskanalys

Under året har vi noterar ökande intresse för analys av risker där flera organisationer skrivit standarder och guider. Terminologin i dessa dokument hade kunnat bli bättre om den hämtats från statistikläran. Detta visar på att information om statistisk terminologi behövs och bör spridas inom t.ex. riskhantering och standardiseringsarbete.

Definitionen av risk varierar mellan olika branscher, och även själva analysmetoderna varierar. Detta skapar en komplexitet, där analysarbete kompromissas till en enklare översyn av risker istället för att göra en statistisk riskanalys. I enklare fall används ofta påhittade skalor för t.ex. sannolikhet som sedan multipliceras med ett tal för t.ex. konsekvens. Att multiplicera ihop några tal och eventuellt aggregera resultatet torde vara enkelt, men det kan vara nog så svårt att bestämma numeriska värden på de ingående faktorerna i produkten.

Ett kärnkraftverk kan utgöra ett exempel på sådana svårigheter. Där finns ofta ett antal säkerhetsbarriärer som skall fallera innan en händelse inträffar. Man kan tänka sig att dessa barriärer är kopplade på något sätt som gör det svårt att rätt härleda sannolikheten för denna händelse.

Även formerna och verktygen för riskhantering skiljer sig mellan olika organisationer och verksamheter. Antagligen därför har vi flera standarder och guider, eftersom det är svårt att i ett enda dokument beskriva riskhanteringen i detalj.

Forsmarks tre kärnreaktorer.

Sedan 2009 har ett antal standarder tillkommit för att skapa samsyn beträffande riskhanteringens begrepp, processer, mm. Exempel på sådana standarder är:

- ISO 31000:2009, Risk management – Principles and guidelines.
- ISO/IEC Guide 73:2009, Definitions of generic terms related to risk management.

ISO/IEC Guide 73 definierar risk som kombinationen av sannolikheten för en viss händelse och dess konsekvenser.

ISO 31000 föreslår att den tidigare definitionen av risk ändras från "slump eller sannolikheten för förlust", till "effekten av osäkerhet för att uppnå mål". Den nya riskdefinitionen som en "effekt" får nog betraktas som ett olycksfall i standardiseringsarbetet.

ISO/IEC Guide 73 definierar riskvärdering som en övergripande process med riskanalys och riskbedömning där olika risker hanteras var och en för sig. Det kan leda till att man missar effekten av t.ex. risksamverkan. Här borde i stället förordas användning av etablerade metoder för analys av samverkan mellan flera riskfaktorer.

En bredare användning av statistisk terminologi för riskanalys kan bidra till tydligare definitioner och säkrare riskhantering.

GÖRAN LANDE

Statistik för kvalitetsteknik

Inledning

Kvalitetsteknik är en samlingsbeteckning för de verktyg, metoder och teorier som handlar om hur varu- och tjänsteproducerande organisationer skall kunna nå en hög kvalitet på sina processer, tjänster och produkter. Tekniken utvecklades redan i början av 20-talet inom industrin, men används idag inom andra verksamheter, av myndigheter, kommuner, landsting, mm. Syftet med tekniken är att minimera förluster, eliminera fel och därigenom få nöjda kunder och förbättra anseendet. Förluster kan orsakas av t.ex. onödiga väntetider, fel, missförstånd, garantikostnader, defekter, bristande kommunikation eller organisation. Kvalitetsteknik används idag inom många områden som t.ex. verksamhetsutveckling och för att minska variationen i olika processer. Inom industrin är det känt att variation kan bäst identifieras och minskas med hjälp av statistiska metoder som används för t.ex. processtyrning, förkortat SPS (statistisk processtyrning).

Utvecklingen av kvalitetsteknik

Kvalitetstekniken började utvecklas när industrin ville öka sin produktivitet. Teorin för kvalitetsstyrning (Quality Control) av produktegenskaper utvecklades av bland annat Walter Shewhart, Edwards Deming och Genichi Taguchi. Dessa utvecklade framförallt de statistiska metoder som ända sedan dess anses utgöra kvalitetsteknikens vetenskapliga kärna. Bland annat utvecklades metoder för statistisk acceptanskontroll av produkter, statistisk processtyrning och försöksplanering – en metodik för att identifiera processens styrande faktorer med hjälp av regressionsanalys. Genom dessa metoder lärde sig många industriföretag tidigt hur man kunde bygga in kvalitet i produkter redan från början genom att noga kontrollera och övervaka sina tillverkningsprocesser. Dessutom lärde man sig hur man kunde reducera sina försöksplaner och ändå erhålla nödvändig information. På detta sätt kunde man minska kostnaderna för provtagning och kassationer.

Kvalitetsteknik fick sitt egentliga genombrott efter andra världskriget, när den japanska industrins kvalitet behövde förbättras. Japanska produkter ansågs vid denna tid ofta hålla en undermålig kvalitet. Utbildning av ingenjörer i kvalitetsteknik samt effektiv opinionsbildning bland företagsledare blev en viktig del i det så kallade japanska undret.

Översyn av dagens kvalitetsteknik visar på en vidareutveckling till mer statistisk kvalitetsteknik. Kunskaper i statistik tycks vara den avgörande skillnaden mellan en dåligt fungerande verksamhetsutveckling och ett effektivt förbättringsarbete.

Verksamhetsutvecklare börjar inse att ISO-standarder för kvalitetssystem enbart anger minimikraven för en kvalitetssäkring och att kunskaper i statistik är nödvändiga för att säkra bättre kvalitet.

Därför behövs det kompetens i både kvalitetsteknik och statistik för att kunna hantera och förbättra processer, analysera stora datamängder, processdata, optimera och öka produktivitet och kvalitet. Där bör statistiken ses som en strategisk kompetens för att kunna skilja signaler från bruset i den stora informationsmängd som moderna IT-system gör tillgänglig.

Här nedan ges några exempel på metoder som kan användas för att förbättra arbetet med t.ex. målstyrning, marknadsanalyser och produktionsstyrning.

Målstyrning och uppföljning

Begreppet målstyrning omfattar oftast målframtagning, målnedbrytning och uppföljning. Även om arbetet med målframtagning och uppföljning tar relativt kort tid så har det mycket stor betydelse under resten av året, eftersom det styr företagets samtliga aktiviteter. Kanske den viktigaste uppgiften för en statistiker är just att hjälpa till med en bra utformning och definition av målen så att de är väldefinierade, mätbara och verifierbara.

Oftast används en fallande rak linje för att exemplifiera en minskning av t.ex. felutfall med tiden. Tyvärr fungerar det inte så bra i praktiken. Bättre är att räkna fram en teoretisk modell för en ständig förbättring, (Se diagram i figur nr 1). Där utgår man från en uppmätt felkvot (den första tidpunkten i diagrammet) och målvärdet på felkvoten efter 12 månader (sista tidpunkten i diagrammet).

Exempel i figur nr 1 illustrerar en modell för minskning av felutfall, där modellen, (heldragen linje), visar hur felutfallet bör kunna minska under perioden av 12 månader. Cirklarna är mätningar av felkvoter för olika provtagningar under året. Dessa värden används för att undersöka samstämmigheten mellan mätningarna och modellen. (Här används simulerade värden som beräknades med hjälp av logistisk regression i datorprogrammet Minitab och ett makro skapat av Ingemar Sjöström).

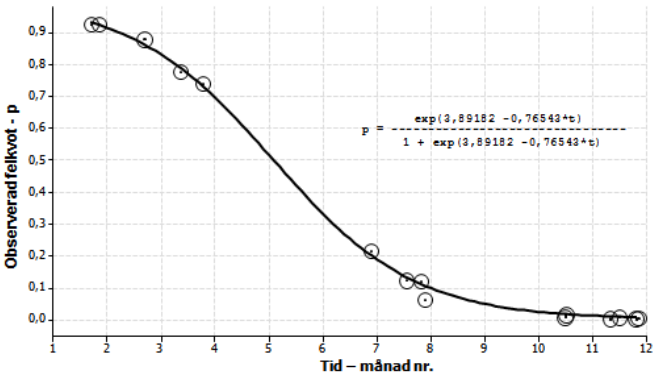


Fig. 1. Exempel på en modell för minskning av felutfall. Modellen, (heldragen linje), visar hur felen ska kunna minskas med tiden. Cirklarna visar mätresultat.

Marknadsanalys

Statistisk analys av t.ex. kundnöjdhetsdata kan förse företaget med viktig information om t.ex. kundernas nöjdhet med en viss produkt, visa marknadstrender och bidra till mer effektiva marknadssatsningar. En kunddataanalys görs vanligtvis genom intervjuer av kundgrupper med frågor till kunder om t.ex. hur nöjda de är med en viss produkt eller varumärke. Antalet kunder som intervjuas kan variera mellan olika intervjutillfällen och den andel kunder som svarar att de t.ex. är ”mycket nöjda” redovisas som en proportion. Spännvidden mellan styrgränserna varierar beroende på antalet kunder som ingår i undersökningen. Spännvidden mellan styrgränserna blir mindre ju flera kunder som intervjuas och säkerheten för ett skattat värde blir högre. Exempel på en kunddataanalys visas här i figur nr 2, där mätvärdena, (proportion), presenteras med hjälp av ett styrdiagram, som vanligtvis används för styrning av processer. Vitsen med att presentera data just på detta sätt är att det illustrerar och verifierar om en observerad förändring, eller förbättring är verklig och inte bara resultat av slumpen.

Data i form av cirklar innanför styrgränserna kan vara resultat av slumpen. De cirklar som ligger utanför styrgränserna beror med stor sannolikhet inte på slumpen och visar på att en förbättring med tiden faktiskt kan konstateras.

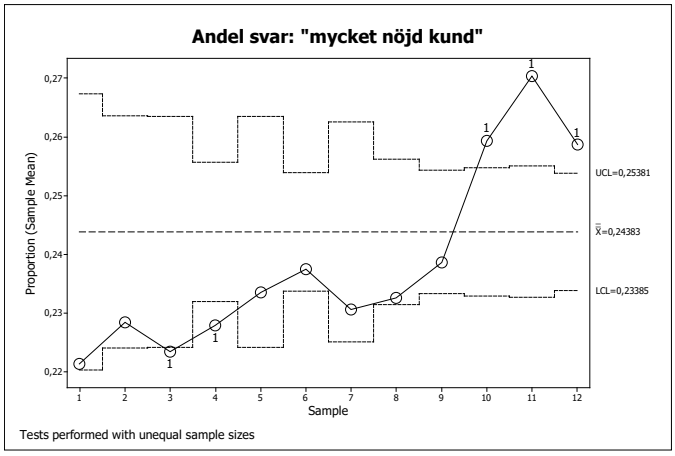


Fig. 2. Exempel på kunddataanalys presenterad i form av ett styrdiagram. Den horisontella axeln visar månadens nummer och den vertikala visar proportion, dvs. andel svar från ”mycket nöjda kunder”. Streckad linje, (0,24), visar mätningarnas medelvärde. De prickade linjerna visar styrgränserna. Cirklarna märkta med siffran 1 markerar data som visar en relevant förändring. Cirklar utan siffra, (innanför styrgränserna), visar mätdata som kan bero på slumpen.

Analysen visar att andelen av ”mycket nöjda kunder” ökade under året och det verifierar en faktisk förbättring.

För jämförelse presenteras här samma data i ett enklare linjediagram, se figur nr. 3. Ett sådant diagram ger inte samma möjlighet att avgöra om en förändring verkligen har inträffat.

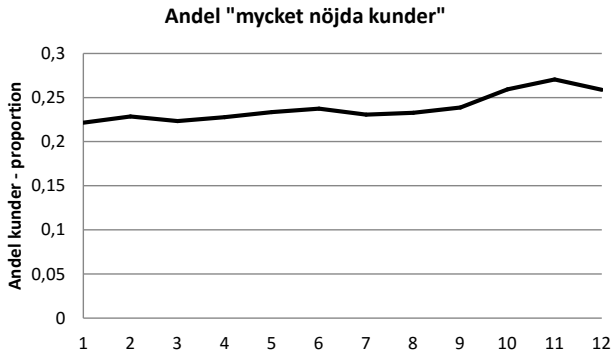


Fig. 3. Exempel på en enklare kunddataanalys presenterad i form av ett linjediagram. Den horisontella axeln visar månadernas nummer och den vertikala visar proportion, dvs. andel kunder som svarade att de är ”mycket nöjda kunder”. I det här diagrammet är det svårare att se om en förändring med tiden är verklig eller kan bero på slumpen.

Kapabilitetsanalys - analys av processduglighet

I praktiken är det vanligt att de flesta processer uppvisar en naturlig variation, ”shift” och ”drift” av medelvärde över tiden. Under kortare tidsperioder kan processer betraktas som stabila om det är enbart slumpen som påverkar processen. Under sådana omständigheter är processdata oftast normalfördelade och processen kan beskrivas med hjälp av målvärde, medelvärde, variation och dess spridningsmått – standardavvikelse i förhållande till givna specifikationsgränser. I de fall där processens spridning befinner sig inom dessa gränser anser vi processen vara kapabel eller duglig. Kapabiliteten mäts med hjälp av olika kapabilitetsindex. Kapabilitetsanalys och kapabilitetsindex har presenterats tidigare i Qvintensen 2011:2 av Magnus Arnér och Kerstin Vännman.

»»» STATISTIK FÖR KVALITETSTEKNIK

Processkapabilitet eller duglighet har att göra med förmågan hos en process att uppfylla de krav vi har på den i form av givna specifikationsgränser, som ofta benämns toleransgränser. Vid kapabilitetsstudier undersöker man hur bra resultatet från t.ex. en maskin eller tillverkningsprocess är, jämfört med toleransgränserna. Förhållandet mellan toleransgränserna och processpridningen anges i form av kapabilitetsindex. Ju högre värde på kapabilitetsindex desto större är tillförlitligheten hos processen. Kapabilitetsanalys används även för att identifiera möjligheter till förbättringar i produktion av produkter och tjänster.

Exempel på nyttan med kapabilitetsanalys illustreras här som jämförelse mellan en statistisk analys av processens kapabilitet (figur nr 4), och en enkel presentation av processdata i form av ett punktdiagram (figur nr 5). I figurerna jämförs samma processdata från 29 mätningar gjorda på kretskort. Data består av spänningsmätningar med krav på max. spridning inom intervallet 4,95 – 5,05V.

I figur nr 4 presenteras de erhållna mätvärdenas fördelning i form av ett histogram med lika klassbredd. Man ser här hur de erhållna värdena på spänningsmätningarna fördelar sig i förhållande till specifikationsgränserna och målvärdet. En normalfördelning har sedan anpassats till givna data (den tjockare streckade linjen i diagrammet), och denna ses som en förklaringsmodell. (Processdata är i regel normalfördelade men även icke-normalfördelade data kan ofta efter transformering behandlas som normalfördelade.)

I nästa steg jämförs läget av modellens medelvärde med det specificerade målvärdet (”target”), och modellens spridning med specifikationsgränserna LSL (lägre specifikationsgräns) och USL (övre specifikationsgräns). Vi beräknar dels avvikelsen från målvärdet, dels den summerade ytan av fördelningens svansar som råkar ligga utanför specifikationsgränserna). Det senare ser vi som en skattning av andelen felaktiga enheter som vi riskerar att producera.

Processen anses vara duglig (med kapabilitet = 1) om så gott som hela ytan under normalfördelningen (svarande mot medelvärdet ± tre standardavvikelser) ligger mellan de båda specifikationsgränserna. I figur nr 4 ser vi att 7,44 procent av ytan ligger utanför dessa gränser (se den svarta pilen i diagrammets vänstra del; 7,44 = 0,08+7,36). Vi tolkar detta som att processen är mindre duglig, och att vi troligen kommer att producera 7,44 procent enheter ned felaktig spänning, som ska behöva kasseras.

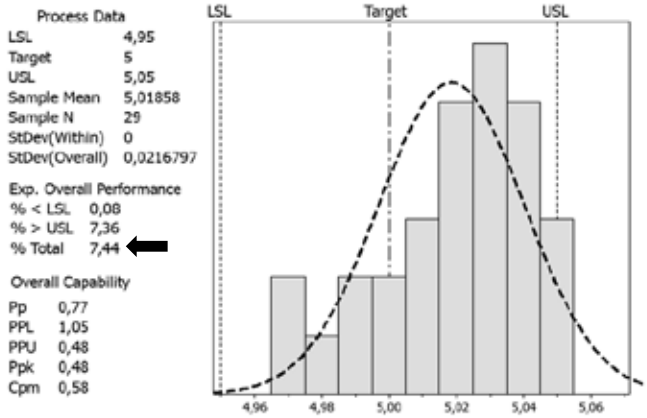


Fig. 4. Exempel på en standard kapabilitetsanalys gjord med datorprogrammet Minitab. Data består av mätningar av spänning på 29 st. kretskort, med krav på max. spridning inom intervallet 4,95 – 5,05V. Diagrammet visar att processdugligheten är låg och att vi kan förvänta oss fel på ca 7,44 % av producerade enheter. (märkt med svart pil). Det betyder att produktionen inte klarar marknadskraven och det kan leda till kassation.

Som jämförelse kan vi konstatera att den enkla presentationen i figur nr 5 nedan inte ger någon information om trolig procentuell andel felaktiga enheter i kommande produktion.

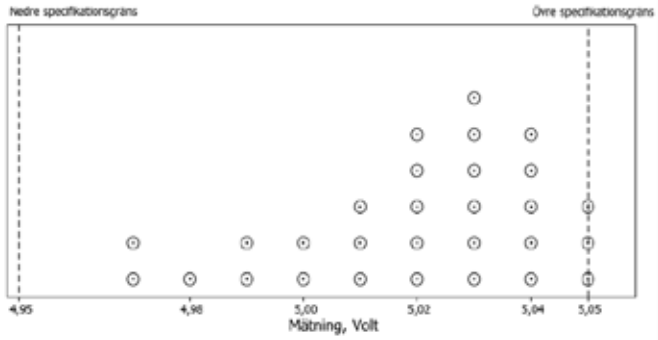


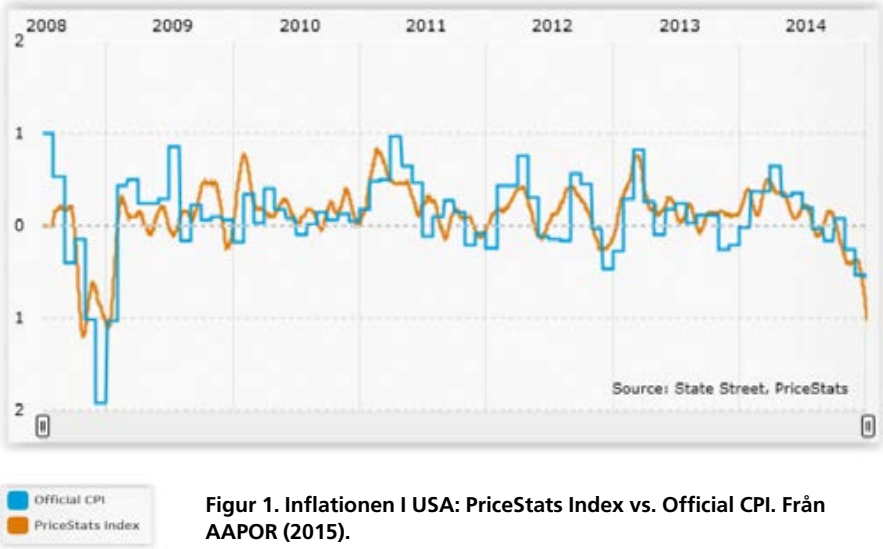
Fig. 5. Exempel på punktdiagram för presentation av uppmätta processdata i förhållande till specifikationsgränserna. (Här används samma data som för beräkning av kapabilitet i figur nr 4). Diagrammet visar att alla 29 enheter klarar marknadskraven.

Avsikten med de presenterade exemplen är att uppmuntra till en bredare användning av statistisk analys inom kvalitetsteknik och verksamhetsutveckling. Rätt analys kan avgöra skillnaden mellan ett bra och ett dåligt beslutsunderlag.

GÖRAN LANDE

SIDAN 10-13/

“Big Data” och “Predictive Analytics” – modeord eller …?



Figur 1. Inflationen I USA: PriceStats Index vs. Official CPI. Från AAPOR (2015).

Har du flyttat?

■ Meddela din nya adress genom att skicka ett e-brev till Ingrid Lönnstedt på sekrfram@gmail.com.

FOTO: GULLERS, K W/ NORDISKA MUSEET

Nils Ferlin – tillfällig gäst i Qvintensen

■ Som motto för SCB:s statistikproduktion skulle man kunna ha:

Allt är rejält och fakta.
Kom och se och beakta.
Intet rosas och risas.
Allt bara förevisas.

(Från Nils Ferlins dikt *”Vaxkabinett”* i diktsamlingen Goggles, 1938.)

■ En annan Ferlindikt återger, efter lätt uppdatering, en upplevelse som vi nog alla känner igen:

Du har glömt ditt hemliga lösenord,
du barfotabarn i livet.
Så sitter du åter vid datorns bord
och gråter så övergivet.
Vad var det för ord – var det långt eller kort,
var det väl eller illa skrivet?
Tänk efter nu – förrn vi kopplar dig bort,
du barfotabarn i livet.

(Baserat på Ferlins dikt *”Du har tappat ditt ord”* i diktsamlingen Barfotabarn,1933.)

JW

Qvintensen

POSTTIDNING B

SVERIGE
PORTO BETALT
PORT PAYÉ

Avsändare/Sender: Svenska statistikfrämjandet
c/o John Öhrvik, Västmanlands sjukhus Västerås, Centrum
för klinisk forskning, 721 89 Västerås, Sweden

Bli medlem i Svenska statistikfrämjandet

Svenska statistikfrämjandets syfte är bland annat att främja sund användning av statistik som beslutsunderlag och att väcka och sprida intresse för statistik i samhället.

Om du önskar bli medlem i Svenska statistikfrämjandet skicka ett e-brev till Ingrid Lönnstedt på sekrfram@gmail.com. Du får Qvintensen i brevlådan och platsannonser via e-post.

Det ställs inga krav för att bli medlem; alla som är intresserade av statistik och vill stödja statistikens roll i samhället är välkomna.

SSFr
Svenska statistikfrämjandet