

Qvintensen

NR 2 2019

STATISTIK-
UTBILDNING
OCH
SIMULERING

SIDORNA 30-31

TEMA/

P-värde
– ett redskap
i den statistiska
verktygslådan

SIDORNA 11-25

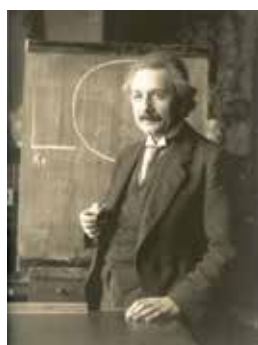




Carl von Linné.



Zlatan Ibrahimovic.



Albert Einstein.

Namn	Levnadsår	Antal Wikipedia-upplagor (Juni 2019)	Antal sidvisningar 2018 alla språk
Georg Stiernhielm	1598-1672	16	11 642
Olof Rudbeck d.ä,	1630-1702	25	33 333
Urban Hjärne	1641-1724	8	8 216
Christoffer Polhem	1661-1751	25	50 551
Anders Celsius	1701-1744	83	463 771
Carl von Linné	1707-1778	156	2 383 926
Pehr Wargentin	1717-1783	9	4 364
Torbern Bergman	1735-1784	27	19 951
Carl Wilhelm Scheele	1742-1786	63	218 838
Simon Stevin	1548-1620	38	139 550
Isaac Newton	1642-1727	210	10 149 008
Gottfried Achenwall	1719-1772	15	41 723
Adolphe Quetelet	1796-1874	37	89 542
Jacobus Kapteyn	1851-1922	20	15 994
Anders Lindstedt	1854-1934	7	2 194
Albert Einstein	1879-1955	204	19 418 199
Hans Rosling	1948-2017	31	443 941
Zlatan Ibrahimovic	1981-	84	13 728 298

Pehr Wilhelm Wargentin.

Ett sätt att mäta en persons berömmelse idag är att kontrollera i hur många Wikipedia-upplagor vederbörande förekommer med egen artikel. Man kan också titta på hur många gånger en viss Wikipedia-artikel blivit läst. I tabellen nedan jämförs de "svenska snillena" i Anderssons bok med några andra vetenskapsmän, samt med fotbollsstjärnan Zlatan Ibrahimovic.

Ett bortglömt jubileum

En upptäckt som gjordes något för sent

I Sankt Jacobs kyrka vid Kungsträdgården i Stockholm kan man beskåda ett epitafium med följande text:

Här vilar Pehr Wilhelm Wargentin, Kungl. Vetenskapsakademiens ständige sekreterare och astronom. Född i Sunne, Jämtl. 1717. Död i Stockholm 1783. Kungliga Vetenskapsakademien uppsatte denna minnestavla 1937.

Kanske uppmärksammade Kungliga Vetenskapsakademien (KVA) 220-årsminnet av Wargentins födelse.

I samband med 200-årsminnet av Wargentins död gav SCB 1983 ut en minnesskrift. Eftersom Wargentin föddes den 11 september 1717 (den 22 september enligt modern tideräkning), så borde 200-årsminnet av hans födelse ha uppmärksamats 2017, vilket mig veterligen inte gjordes av vare sig KVA, SCB eller Svenska Statistikfrämjandet.

Själv konfirmerades jag i Sankt Jacobs kyrka 1967, några månader före Wargentins 250-års jubileum, men lade då inte då märke till epitafiet. Mitt intresse för vetenskapshistoria hade inte vaknat ännu. Jag har sedan besökt samma kyrka vid ett antal tillfällen men började intressera mig för texten först 2018, således ett år för sent för att uppmärksamma jubileet.

Vem var Pehr Wargentin?

I boken *Svenska snillen* från 1964 så påpekar civilingenjör P. Gunnar Andersson att

Trots sina lysande insatser för de svenska vetenskaperna under deras blomstringstid på 1700-talet är dock Wargentin för nutiden ett relativt okänt namn utom för ett fåtal specialister.

Andersson framhåller Wargentins arbeten rörande Jupitermånarna, samt pionjärinsatserna med befolkningsstatistiken.

Som tidigare nämnts så föddes Wargentin 1717 i Jämtland. Hans far var kyrkoherde, men hade ambitioner inom fysiken och var påtänkt till en professur i Dorpat (Tartu). Skolrektorn på Frösön var så pass imponerad av den unge Pehr, att han hade tänkt skicka honom som fjorton-åring till Uppsala. Fadern tyckte dock att man skulle skynda långsamt och sände honom istället till gymnasiet i Härnösand.

År 1735 skrivs Pehr Wargentin in som student i Uppsala och börjar studera teologi, men går snart över till astronomin, inspirerad av Anders Celsius. Han har även matematikern Samuel Klingenskierna som lärare. P.g.a. ekonomiska problem tvingas Wargentin göra uppehåll i studierna, men promoveras till magister 1743.

År 1748 är ett framgångens år: Wargentin blir utnämnd till adjunkt i matematik, och dessutom medlem av Kungliga Vetenskapsakademien samt korresponderande ledamot av franska vetenskapsakademien. Ett år senare blir han också ständigt sekreterare i Kungliga Vetenskapsakademien.

Wargentin hade, vid sidan av sitt värv som ständigt sekreterare i Kungliga Vetenskapsakademien, en mängd andra uppdrag, bl.a. var han den mest drivande vad gällde att bygga upp den så kallade Tabellkommissionen, en myndighet som inrättades 1756. Den var en föregångare till SCB, och kan ses som världens äldsta statistiska myndighet. Tabellkommissionen hade som uppgift att ansvara för det så kallade Tabellverket. Observera att "Tabellverket" inte var namnet på någon myndighet, utan i stället namnet på en verksamhet där data insamlades från kyrkobokföringen i alla församlingar i landet för att sammanställas centralt. Tabellverket hade påbörjats redan 1749 på initiativ av Pehr

Elvius (1710-1749), som var Wargentins föregångare som Vetenskapsakademiens ständige sekreterare, Wargentin avled 1783, vilket enligt Kungliga Vetenskapsakademiens hemsida, "markerade det definitiva slutet på Akademiens första storhetstid".

Wargentins internationella inflytande

Sverige var på sjuttonhundratalet en tämligen fattig nation, i utkanten av Europa. Dock hade landet ett gott rykte som vetenskapsnation. Wargentin bidrog verksamt till detta rykte genom att raskt och noggrant besvara brev från utländska vetenskapsmän. I N.V.E. Nordenmarks monografi över Wargentin kan man läsa att "Wargentin var väl i utlandet näst Linné den mest kände representanten för svensk vetenskap under denna tid."

Speciellt bör nämnas kopplingen till Gottfried Achenwall (1719-1772), professor i natur- och folkrätt vid universitetet i Göttingen. Han anses vara den förste som använde ordet "statistik", och han ivrade för att statistik skulle bli ett universitetsämne. Ordet "statistik" hade dock på den tiden en något annan innebörd än nu, och skulle närmast motsvara en (både verbal och numerisk) beskrivning av stater och deras förhållanden. Achenwall lär ha följt Wargentins arbete med befolkningsstatistik med stort intresse. (Achenwall föddes den 20 oktober, vilket råkar vara den internationella statistikdagen. Det har inte gått att utröna om detta är en tillfällighet.)

England hade på den här tiden ett väl utvecklat försäkringsväsende, men saknade befolkningsstatistik. Man kunde dock utnyttja Wargentins siffror

Kombinationen statistik och astronomi

Som ovan nämnts så presenteras Wargentin som astronom på epitafiet i Sankt Jacobs kyrka. Idag »»»

»»» ETT BORTGLÖMT JUBILEUM

så tänker vi inte i första hand på astronomi när vi talar om tillämpningar av statistik, men historiskt sett så har kombinationen astronomi och statistik inte varit helt ovanlig. Som exempel kan nämnas:

- Belgaren Adolphe Quetelet (1796-1874) var intresserad av att på ett statistiskt sätt beskriva egenskaper hos kollektiv av människor. Är idag kanske mest känd som upphovsman till det begrepp som numera kallas Body Mass Index (BMI). Han var även chef för det under hans ledning upprättade observatoriet i Bryssel.
- Anders Lindstedt (1854-1939) verkade i början av sin karriär som astronom i Lund, Hamburg och Dorpat (Tartu). Blev sedan professor i matematik vid KTH (Lindstedtsvägen på KTH-området är uppkallad efter honom). Kom att intressera sig för försäkringslagtiftning och socialförsäkringar, och är känd som "den svenska socialförsäkrings fader".
- Carl Charlier (1862-1934) var en svensk astronomiprofessor som intresserade sig för matematisk statistik, vilket han sedan tillämpade inom olika grenar av astronomin.

- Den kände nederländske astronomen Jacobus Kapteyn (1851-1922) använde statistiska metoder för att studera egenskaper hos vårt stjärnsystem. Han publicerade också 1903 en artikel om statistiska metoder relevanta för biologer.

Är Wargentin bortglömd idag?

Stämmer det som P. Gunnar Andersson skrev 1964 fortfarande? Kanske stämmer det ännu mer nu än då: 200-årsminnet av Wargentins födelse verkar inte ha uppmärksammat av vare sig KVA, SCB eller Svenska Statistikfrämjandet. Tidigare har dock hans insatser uppmärksammat, vilket framgår av litteraturlistan.

Ett sätt att mäta en persons berömmelse idag är att kontrollera i hur många Wikipedia-upplagor vederbörande förekommer med egen artikel. Man kan också titta på hur många gånger en viss Wikipediaartikel blivit läst. I tabellen på sidan 2 jämförs de "svenska snillena" i Anderssons bok med några andra vetenskapsmän, samt med fotbollsstjärnan Zlatan Ibrahimovic.

Dagens Wikipediastatistik tycks således bekräfta Anderssons utsaga.

HANS ALBERG

Referenser

Andersson, P. Gunnar (1964). *Svenska snillen: från Stiernhielm till Scheele*. Stockholm: Natur och kultur.

Jorner, Ulf (2008). *Summa summa-rum – SCB:s första 150 år*. Stockholm: SCB. (Kan även läsas på SCB:s hemsida; se nedan.)

Nordenmark, N.V.E. (1939). *Pehr Wilhelm Wargentin – Kungl. Vetenskapsakademiens sekreterare och astronom*. Uppsala: Kungl. Vetenskapsakademien.

Statistiska centralbyrån (1983). *Pehr Wargentin – den svenska statistikens fader*. Stockholm: SCB.

På SCB:s hemsida kan man bl.a. läsa:

<https://www.scb.se/om-scb/scbs-verksamhet/scbs-historia/> [= SCB:s historia i sammandrag]

http://share.scb.se/OV9993/Data/Publikationer/statistik/_publikationer/OV9999_2008A01_BR_X05BR0801.pdf [= Jorner (2008); se ovan]

Kommentar med anledning av Johan Brings och Lars Rönnegårds artikel i Qvintensen 2019:1

Reflektioner gällande skelettmognad och åldersbestämning

■ I två nu historiska arbeten (Edgren et al., 1974 och 1976) kunde vi konstatera att skelettmognaden kan ske i olika takt inom olika populationer. Till exempel noterade vi att hos samerna i Finland vilkas exakta ålder var känd ur kyrkböckerna skedde mognaden ca ett år senare än inom den övriga befolkningen. Om dylika populationsskillnader är mera generella innebär det enligt min mening att mätningar genomförda för en population inte behöver vara representativa för individer hörande till en annan. Detta medför uppenbar osäkerhet när skelettmognaden används vid en exakt gränsdragning när individen anses uppnå fullvuxen ålder.

JOHAN FELLMAN
FELLMAN@HANKEN.FI

Referenser

Edgren, J., Fellman, J. & Lewin, T. (1974). Skeletal maturation of the hand and wrist, a longitudinal study covering two years among children of Finnish lapps. *Acta Morph. Neerl.-Scand.* 1974:12, sid.1-7.

Edgren, J., Bryngelsson, C., Lewin, T., Fellman, J. & Skrobak-Kaczynski, J. (1976). Skeletal maturation of the hand and wrist in Finnish lapps. *Proc. 3rd Int. Symp., "Circumpolar Health"*, Yellowknife, Canada, ed. R.J. Shephard & S. Itoh, Univ. of Toronto Press, Toronto and Buffalo, 1976, sid. 248-254.



Ur Qvintensen nr 1/2019

Qvintensen

Ansvarig utgivare

Joakim Malmdin

Redaktör

Jan Wretman, 070-781 78 35

Redaktion

Hans Alberg

Jens Malmros

Rolf Larsson

Marie Linder

Marika Wenemark

E-post wretman.jan@gmail.com

Produktion

Form och redigering: Mezzo Media AB

Tryckeri: Trydells Tryckeri AB

Annonser

Annonser i Qvintensen bokas med redaktören.

Annonssutskick per e-post bokas med 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 STATISTIKFRÄMJANDETS STYRELSE

Ordförande

Joakim Malmdin 010-479 49 45,
ordforande.framjandet@gmail.com

Vice ordförande Maria Karlsson

Skattmästare Anders Haraldsson

Sekreterare Mattias Strandberg

Hemsidesansvarig Jens Malmros

Klubbmästare Moa Thyni

Representant Surveyföreningen

Åke Wissing

Representant FMS

Fredrik Norström

Representant Industriell statistik

Hans Alberg

Representant Cramérskapskapet

Fredrik Olsson

Övrig ledamot

Linnea Johansson Kreuger

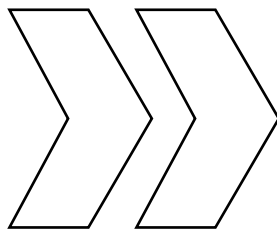
Adress

Svenska statistikfrämjandet

Sekreterare: Mattias Strandberg

E-post sekrfram@gmail.com

Webbplats www.statistikframjandet.se



Marknadskrafter och näringsliv bör utnyttja de statistikexperter som finns tillgängliga och be dem om hjälp.»

GÖRAN LANDE, SIDORNA 30-31

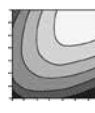
Qvintensen

STATISTIKFRÄMJANDET

NR 2 2019

Innehåll

ETT BORTGLÖMT JUBILEUM	2-4
REFLEKTIONER GÄLLANDE SKELETT-MOGNAD OCH ÅLDERSBESTÄMNING	4
REDAKTÖRENS RUTA/ Jan Wretman	6
ATT VARA STATISTIKER INOM MEDICINSK FORSKNING	7-8
MINNESORD/ Bengt Swensson	8-9
RAPPORT FRÅN KONFERENS	10
TEMA: P-VÄRDEN/ Något om P-värden	11-14
Defending the P-value	15-20
A discussion on the use of the p-value	21-22
Kan Statistikfrämjandet bidra till bättre rapportering?	23-25
STATISTIKUTBILDNING OCH SIMULERING	30-31



Våra föreningar

INDUSTRIELL STATISTIK

Ordförande har ordet 26

SURVEYFÖRENINGEN

Nya i styrelsen 27

Ordförande har ordet 26

CRAMÉRSÄLLSKAPET

Höstmötet 2019..... 28

Nominera till Cramérpriset!..... 28

Nya i styrelsen 29

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 2019.

Om p-värden och lite annat

Under min tid som lärare i statistik kom det ofta på min lott att ge slentrianmässiga nybörjarkurser i statistik för studenter av vilka de flesta var totalt ointresserade av ämnet. De tog den här kursen därför att den var ett obligatoriskt inslag i den examen de hade tänkt ta. Min egen pedagogiska förmåga var inte tillräcklig för att skapa något större intresse. Som kurslitteratur användes alltid någon jättetjock, jättepratig och jätteointressant amerikansk lärobok, antagligen utformad för att passa amerikanska collegestudenter. Ett av kursens mål verkade vara att så många studenter som möjligt skulle bli godkända, så att institutionen inte skulle få kritik. För att uppnå detta gällde det att trivialisera kursinnehållet.

Ett viktigt inslag i kursen var så kallad hypotesprövning. Man skulle, utan närmare eftertanke, formulera en nollhypotes och en mothypotes (= den ur forskningssynpunkt intressanta hypotesen). Man skulle bestämma en signifikansnivå, alltid fem procent. Man skulle, utifrån sina data, beräkna värdet på en teststatistika. Om värdet hamnade utanför vissa gränser, blev slutsatsen "nollhypotesen förkastas". Om värdet inte hamnade utanför dessa gränser blev slutsatsen "nollhypotesen kan inte förkastas". Lärobokens exempel var ofta hämtade från amerikansk collegemiljö. Jag försökte varna studenterna för att gå ut och okritiskt göra så som det stod i läroboken, men jag tror inte att jag lyckades.

I det här numret av Qvintensen handlar ett antal artiklar om problematiken i samband med så kallade p-värden. Ovanstående minnen dök upp, när jag läste igenom dessa artiklar. "Klassisk signifikanstestning" (som den beskrivits i föregående stycke) och "p-värdesresonemang" är ju beslätade förfaringssätt att redovisa slutsatser från experiment. I båda fallen finns risk att av okunnighet formulera slutsatser som inte är berättigade.

Kritiska röster angående p-värden har hörts då och då, och under de senaste åren har debatten tagit ny fart. Uppaktan verkar nu ha varit

en artikel i The American Statistician, som varnade för ogenomtänkt användning av p-värden, men där man egentligen inte hade några alternativ att komma med. Både kritik och försvar har därefter framförts. Om detta kan vi läsa i flera artiklar i det här numret av Qvintensen. Observera att kritiken i regel inte avser begreppet p-värde i sig. Det är ju ett klart definierat begrepp. Kritiken avser snarare tvivelaktig användning och tolkning av p-värden. Här kortfattat några sådana exempel (ibland lite slarvigt formulerade):

- Slentrianmässigt krävs ibland att ett p-värde måste vara mindre än 0,05 för att ett vetenskapligt försök ska vara värt att rapportera.
- Forskare i karriären kan ibland frestas att inrikta sin forskning mot just sådana frågor där man kan vänta sig att få låga p-värden.
- Ibland tolkas $p < 0,05$ som att en viss forskningshypotes är praktiskt taget bekräftad.
- Ibland föreställer man sig att p-värdet ger exakt kvantitativ information om precis hur mycket (eller hur lite) man ska våga tro på en viss forskningshypotes.
- Ibland kan vetenskapliga tidskrifter ha som princip att bara acceptera manuskript, där forskare redovisar p-värden mindre än 0,05. Då får läsaren i det långa loppet en skev information om forskningsläget. Läsaren får bara kännedom om de fall där en viss forskningshypotes har fått stöd, men inte om de fall där samma hypotes inte har fått stöd.
- Om man har tillräckligt stort antal observationer, så kommer nollhypotesen (till exempel att en viss behandling inte har effekt) nästan alltid att förkastas. Även mycket små skillnader mellan försöksgrupp och kontrollgrupp kommer



att ge utslag. I praktiken bör man nog skilja mellan formell signifikans och praktiskt betydelsefull signifikans.

Mer information om dessa saker får läsaren i artiklar i detta nummer. Observera att det inte är fråga om någon egentlig debatt mellan olika författare. Varje författare har lämnat sitt bidrag utan att ha kännedom om vad de övriga har skrivit. Därför har en viss överlappning av innehåll inte kunnat undvikas. Läsaren får stå ut med detta. Som vanligt

gäller att någon referee-granskning inte förekommer i Qvintensen. Varje författare får själv stå för vad han eller hon har skrivit. De aktuella artiklarna ger inte uttryck för någon åsikt som sanktionerats av Statistikfrämjandet. Beträffande själva uttrycket "p-värde" har varje författare fått behålla sitt eget skrivsätt. Det verkar inte finnas någon vedertagen praxis. (I litteraturen ser man ibland stort P, ibland litet p, ibland kursiverat, ibland icke-kursiverat.)

Själv har jag mest varit verksam i surveybranschen. Där har jag aldrig stött på några p-värden. Men jag kan inte låta bli att tänka på ett problem även där, som också har att göra med redovisning och tolkning av resultat. Vid mätning av politiska opinioner (till exempel vilket parti man gillar mest) redovisas ibland förändringar från en tidpunkt till en annan. Sådana förändringar uppges ibland vara "statistiskt säkerställda" (då blir de intressanta), och ibland "inte statistiskt säkerställda" (då struntar man i dem). Det låter ju vetenskapligt, kvantitativt och fint. Men det finns två problem: (1) man redovisar nästan aldrig hur beräkningarna har gjorts, och (2) det finns ofta anledning att misstänka att beräkningarna har gjorts enligt enkla standardformler, där förutsättningarna för dessa beräkningar inte är uppfyllda.



Ida Hed Myrberg jobbar på Karolinska Institutet.

Att vara statistiker inom medicinsk forskning

"Vad är ett protokoll?" Den frågan ställde jag inte, istället nickade jag och instämde hummande med allt som professorn i klinisk barnonkologi – som snart skulle bli min chef – berättade. "Vad är ett pek?" Det frågade jag inte heller, utan bet ihop och låtsades förstå tills mötet med doktoranden och hans handledare var över så att jag kunde gå tillbaka till mitt arbetsrum och googla. För er som är lika obevandrade i den medicinska forskarvärlden som jag en gång var kan jag berätta att ett "pek" är slang för en vetenskaplig artikel, och att ett protokoll är en detaljerad plan för studieupplägget i en klinisk studie.

Bland det första jag fick lära mig på mitt nya jobb som statistiker på en medicinsk forskningsenhet inom barncancerforskning var att våga ställa dumma frågor. Jag, som inte ens gått naturinriktning på gymnasiet, insåg snabbt att mina biologikunskaper från högstadiet var minst sagt rostiga. Det hjälpte inte heller att forskarna envisades med att använda en uppsjö av förkortningar och latinska begrepp – ibland båda två i kombination. Jag blev till exempel inte klokare när jag fick berättat för mig att HLH stod för Hemofagocyterande Lymfohistiocytos, ord som jag ännu i skrivande stund har svårt att uttala och komma ihåg. Något som underlättade var dock forskarnas villighet att förklara svåra begrepp på ett enkelt sätt – bara de fick frågan.

Jag i min tur fick vänja mig vid att bli bombarderad med statistikfrågor. Vad är skillnaden mellan medelvärde och median och när ska jag

använda vilken? Varför logaritmerar du variabeln? Hur ska en oddskvot tolkas? Hur många studiedeltagare behöver jag i min studie? Att då komma dragande med någon krånglig matematisk formel för att förklara skapar bara ytterligare förvirring. Jag fick snabbt lära mig att översätta de statistiska begreppen till forskarnas eget språk, rita upp enkla figurer och hitta nya sätt att förklara, när forskarnas ansikten såg ut som frågetecken.

Något annat som var nytt för mig var hur det gick till med alla dessa "pek" som folk pratade om. För att få pengar till sin forskning måste forskaren ha bra publikationer i bagaget, gärna i väl ansedda medicinska tidskrifter med en hög "impact factor". (Med en tidskrifts impact factor för t.ex. år 2018 menas kvoten mellan (1) antal citeringar 2018 av artiklar publicerade 2016 och 2017, och (2) antal citerbara artiklar publicerade 2016 och 2017.) En tidskrift som citeras flitigt har alltså en hög impact factor, och tidskrifter som Nature och Science låg 2018 på 42 respektive 41. Att stå som medförfattare till en artikel i en högt ansedd tidskrift är alltså eftersträfvansvärt, och bäst är om ditt namn står först eller sist i listan av författare. Hur många författare kan en publikation rimligen ha, tänker kanske den oinvidiga läsaren. Jag själv är medförfattare till en publikation

med hela 34 stycken. Det kan tyckas vara väldigt många, men om en person uppfyller Vancouverrekommendationerna för medförfattarskap bör personen stå med som författare.

En av de största utmaningarna med jobbet var dock inte den akademiska jargongen eller mina bristande biologikunskaper, utan krocken med verkliga data. Som nyutexaminerad från en utbildning i matematisk statistik kom jag från en värld av tydliga ramar, tillrättalagda exempel och metodval som var självklara bara genom kursnamnet. Nu möttes jag istället av stökiga datafiler, ibland oklara frågeställningar,

små stickprovsstorlekar och data som saknades. Dessutom var det oftast inte alls givet vilken statistisk metod jag skulle välja. Jag fick snabbt inse att jag inte kan ha 7,5 högskolepoäng i alla metoder jag någon gång behöver använda. Till exempel överlevnadsanalys och mixed models var nytt för mig, och jag fick lära mig att tillämpa och tolka metoderna på egen hand

genom att läsa böcker och artiklar, och titta på inspelade föreläsningar. Men metodval visade sig inte bara vara en fråga för statistikern. Inom enskilda forskningsfält kan forskarna ha vant sig vid att använda ett fåtal specifika statistiska metoder. Då kan det vara svårt att motivera valet av en alternativ metod, lämpligare för den aktuella frågeställningen och datamaterialet. Det »»»»

»En av de största utmaningarna med jobbet var [---] krocken med verkliga data.»

»»» ATT VARA STATISTIKER ...

är inte heller bara forskarna i den aktuella studien som ska övertalas. När publikationen väl är inskickad till en vetenskaplig tidskrift ska den granskas av andra forskare och experter inom samma fält, och dessa kan vara skeptiska till statistiska metoder de inte känner igen.

När det kommer till att presentera resultaten från statistiska analyser för en forskare gäller det att hålla tungan rätt i mun. Risken är att forskaren stirrar sig blind på p-värdet och hurrar om detta är lägre än 5%, men glömmer bort den kliniska signifikansen. Med tillräckligt stora datamaterial går det att hitta statistisk signifikans för pyttesmå skillnader/effekter som inte har någon som helst klinisk betydelse. Och tvärtom – en studie kan visa på betydelsefulla skillnader/effekter, men på grund av en för liten stickprovsstorlek uppnås inte statistisk signifikans, och detta tolkas felaktigt som att effekt saknas.

Små stickprovsstorlekar visade sig på mitt nya jobb mer vara regel än undantag. Forskarna var ofta väl medvetna om begränsningarna hos sina data, men kunde ändå be om olika former av statistiska tester. Att i ett sådant läge avfärda studien och be forskarna samla in mer data kan i många fall vara orealistiskt. Till exempel djurstudier bör av etiska skäl vara så små som möjligt. Kliniska data kan vara svåra att samla in, till exempel på grund av att diagnosen som studeras är mycket ovanlig, eller på grund av kostsamma, tidskrävande och ibland smärtsamma undersökningar. Det är viktigt att som statistiker ha respekt för sådana omständigheter och att försöka sätta sig in i vad det innebär att samla in data. Jag själv, som när jag anställdes hade mycket liten inblick i klinisk och preklinisk forskning, samt i sjukvårdens metoder, passade på att följa med forskarna till labbet när jag fick möjlighet. Dessutom deltog jag själv som frisk kontroll i en klinisk studie där mitt huvud blev undersökt med magnetkamera och PET-kamera.

Trots – eller kanske tack vare – alla ovan nämnda utmaningar har jag blivit kvar på samma forskningsenhet i snart fyra år, parallellt med att på deltid ha läst en masterutbildning i matematisk statistik. Det fantastiska med statistik är, bland mycket annat, att det går att tillämpa på så många olika områden. Det är en lyx att kunna bli anställd på en medicinsk forskningsenhet utan att ha någon medicinsk bakgrund, med enbart sin statistiska kompetens i bagaget, samt en vilja att utveckla och tillämpa denna på bästa sätt. Att i mitt arbete möta engagerade, nyfikna och kloka forskare, och tillsammans med dessa komma fram till resultat som i slutändan kan rädda och förbättra livet för människor – det är något jag vill hålla på med länge.

IDA HED MYRBERG
BARNCANCERFORSKNINGSENHETEN,
INSTITUTIONEN FÖR KVINNORS
OCH BARNS HÄLSA
KAROLINSKA INSTITUTET

Källor och referenser

Källa: Journal Citation Reports

Referens: <http://www.icmje.org/icmje-recommendations.pdf>

Från tidigare doktorander

Professor emeritus Bengt Swensson, Örebro, har avlidit vid 82 års ålder. Närmast anhöriga är hustrun Lena samt barnen John, Li, Towe och Hannah med familjer.

"Den gamle" har gått bort. Vi som var hans doktorander kallade Bengt så ibland när vi pratade om honom. Benämningen syftade inte i första hand på åldersskillnaden, utan var snarare ett uttryck för respekt. Inte bara för hans kunande inom området survey sampling, utan även för hans person. Bengt var vår handledare, om inte alltid formellt så i praktiken. Han engagerade sig, ett engagemang som innebar höga krav på arbetet. Även om man inte förväntades göra mer än överenskommet, skulle det som gjordes vara bra. Engagemanget omfattade inte bara arbetet. Utan att han nödvändigtvis var medveten om det, påverkade han även utbildningsinriktning och livshållning. Bengt var en i ordets bästa mening moralisk person med stor integritet, och utgjorde i detta avseende en sann förebild.

Bengt var som statistiker verksam både vid Statistiska Centralbyrån, SCB, och högskolan/universitetet i Örebro. Under sin tid vid SCB på 70- och 80-talet bidrog han starkt till att surveyteorin tillämpades och utvecklades inom verket. Vi var då en handfull anställda som genom hans stora pedagogiska förmåga bibringades klarhet och förståelse om vad denna teori faktiskt innebar och hur den kunde tillämpas. Bengt besvarade sällan frågor direkt, man fick vara glad om han tipsade om några relevanta referenser.

Under sin tid på SCB inledde Bengt, tillsammans med kollegorna Carl-Erik Särndal och Jan Wretman, författandet av läroboken *Model Assisted Survey Sampling*. Boken utkom 1992 och betraktas i dag över hela världen som ett standardverk. När boken väl utkom hade Bengt lämnat SCB för Högskolan i Örebro, sedermera Örebro universitet, där han förblev under resten av sin yrkesverksamma karriär.

Kontakten med Bengt var en smula oregelbunden efter avhandlingsarbetet, men den fanns där. Ett återkommande inlägg var matcherna mellan ÖSK och MFF, där Bengt med glädje återföll i skånsk lokalpatriotism om omständigheterna så tillät. På senare tid träffades vi ibland för att äta lunch med honom och Lena. Det var muntra och uppskattade tillfällen där hans deltagande kommer att saknas.

CLAES ANDERSSON, FD
MARTIN AXELSON, FD
ANDERS HOLMBERG, FD

MINNESORD ÖVER BENGT SWENSSON

Från mångåriga medarbetare

Bengt Swensson, Örebro, avled den 4 april, i en ålder av 82 år. Han efterlämnar hustru Lena och barn med familjer.

Bengt var professor emeritus i statistik vid Örebro universitet. Hans specialområde var matematisk/statistisk metodik för urvalsundersökningar (sample surveys), framför allt så som den används i framställning av officiell statistik. Ett viktigt exempel är statistiken om sysselsättning och arbetslöshet, utförd i många länder med datainsamling från ett urval av personer ur populationen – i regel ett sannolikhetsurval – följd av avancerad estimationsteknik. I Sverige gör Statistiska Centralbyrån (SCB) detta i Arbetskraftsundersökningarna (AKU).

Bengt doktorerade 1977 i statistik vid Stockholms Universitet med avhandlingen *Survey measurement of sensitive attributes: Some contributions*. Området är känt som "Randomized response", en teknik som kanadensaren Stanley Warner introducerade 1965 för surveys med känsliga frågor, där intervjupersonernas tendens till att ge undvikande eller medvetet felaktiga svar skulle kunna orsaka skeva skattningar, men där datainsamlingen från uttagna personer kan modifieras så att skattningar ändå blir väntevärdesriktiga.

Efter att ha doktorerat kom Bengt att arbeta en period som metodstatistiker vid SCB i Örebro samtidigt som han fortsatte att meritera sig vetenskapligt. Han återvände till universitetsvärlden och blev så småningom den förste professorn i statistik vid Örebro Universitet, där han bidrog starkt till uppbyggandet av den statistiska institutionen.

Handledare för Bengts doktorsarbete var professor Tore Dalenius. Tore var en föregångsman när vetenskaplig surveymetodik – med det som nu kallas sannolikhetsurval – slog igenom på bred front i Sverige, på SCB och andra ställen, från 1950-talet och framåt. Bengt brukade säga att vad som tidigt attraherade honom till den klassiska surveyteorin med dess sannolikhetsurval – så som J. Neyman hade framställt den i

viktiga teoretiska arbeten på 1930-talet – var att *objektiv kunskap* om populationen blev möjlig.

Med "objektiv" menade Bengt att sannolikhetsurvalet medger inferens utan några som helst antaganden eller modeller om populationens form och andra okända egenskaper. Väntevärdesriktiga skattningar och deras precision kan härvid beräknas utifrån kunskap om själva urvalsdragningen. Det blir möjligt genom att denna, alltifrån ett enkelt OSU till en komplex design i flera steg, har tilldelat populationens alla objekt kända urvalssannolikheter. "Man-made randomization" har det kallats; som en följd kan *designbaserad* inferens – som det kom att heta – ske rörande den ändliga populationens parametrar, såsom medeltal, median, regressions- och korrelationskoefficienter o.s.v.

Bengt blev en auktoritet på statistisk metodik för urvalsundersökningar, välkänd som medförfattare – med undertecknade två kollegor – till boken *Model Assisted Survey Sampling* (Springer, 1992). Den behandlar surveymetoder – urval och estimation – som är generellt tillämpbara och den blev en internationell standardreferens; dess tänkesätt kom att bli använt i statistikproduktion vid nationella statistikbyråer i världen och har lett till tusentals citat i den vetenskapliga och tillämpade litteraturen om sample surveys.

Tänkesättet kallade vi för *modellassisterad* – men fortfarande *designbaserad* – inferens, det vill säga en form av inferens som utnyttjar hjälpvariabler, vilkas samband med undersökningsvariabeln kan uttryckas i en antagen modell. Trots ovissheten om modellens "riktighet" bibehålls den designbaserade formen av inferens. Dess konservativa opartiskhet – inferens utan den risk som ligger i att förlita sig "rakt ut" på en osäker modell – låg djupt rotad i de nationella statistik-byråernas medvetande, alltsedan den designbaserade inferensens genombrottsperiod 1950 – 1970.

Det ska påpekas att diskussionen om olika former för inferens – rent modellbaserad, eller bayesiansk, eller annan form – sköt fart kring

1970; diskussionens vågor gick höga vid internationella konferenser på 1970- och 1980-talet. Givetvis utan att någon klar vinnare kunde koras. Men kombinationen *modellassisterad-designbaserad* inferens kom så småningom att övertyga många.

I linje med bokens filosofi och tänkesätt handledde Bengt ett antal doktorander och licentiander fram till fullbordad avhandling. Fem av hans doktorer i statistik – med specialisering på sample surveys – är eller var metodexperter vid SCB.

Bengts statistiska expertis togs i anspråk i olika sammanhang; han medverkade till exempel i en statlig utvärdering av undervisning och forskning i ämnet statistik vid de svenska universiteten och han engagerades som statistisk expert vid rättsfall i Marknadsdomstolen.

Bengt hade bestämda uppfattningar och kunde vara sträng i sina omdömen. Han var mycket mån om tilliten till den officiella statistiken; ställde höga krav på tillförlitlighet och exakthet, utifrån perspektivet "undvikande av grovt felaktiga beslut på hög nivå". Han bekymrade sig på senare år för de negativa konsekvenserna för statistikens tillförlitlighet, till följd av tilltagande svarsobenägenhet och andra typer av bortfall i sample surveys.

Bengt kunde vara fordrande när han såg brister i viljan att med krafttag ta i tu med kvalitetsproblemen i statistiken, i viljan att göra investeringar för att komma tillrätta med dessa. Han framförde sin kritik med auktoritet.

Vi som under lång tid arbetade nära tillsammans med Bengt minns honom som en inspirerande arbetskamrat och god vän med mycket mänsklig värme, något som vi också fick uppleva från Bengts familj i Örebro. Han hade vid sidan av statistiken breda intressen inom humaniora och samhällsfrågor. Han kommenterade ofta det han läst på kultur- och debattsidor i de stora dagstidningarna eller inhämtat på annat sätt. Att diskutera med honom var alltid givande. Vi saknar honom.

CARL-ERIK SÄRNDAL
JAN WRETMAN

Rapport från konferens för statistikstuderande på mastersnivå

Statistikfrämjandets styrelse anordnade den 20-21 maj detta år en konferens, som vände sig till studerande vid mastersprogram med statistikinriktning vid svenska universitet. Konferensen ägde rum på Garnisonen i Stockholm. Huvudansvarig för arrangemanget var Fredrik Norström (styrelsemedlem) med bistånd av Moa Thuni (klubbmästare), Mattias Strandberg (sekreterare) och Joakim Malmelin (ordförande). Vid arrangemanget deltog 15 studenter, 9 av dessa från Umeå universitet, 3 från Lunds universitet, 2 från Örebro universitet och 1 från Uppsala universitet. Programmet, som sträckte sig från lunch till lunch under de två dagarna, bestod av föredrag av 7 inbjudna föreläsare samt ett pass där de studerande gavs möjlighet att presentera sitt studiearbete. Av de inbjudna föreläsarna var det 2 stycken som presenterade aktuell statistisk forskning, samt 5 stycken som hade ett fokus mot hur det är att jobba som statistiker inom olika områden. De fem sistnämnda föreläsarna presenterade innehåll med anknytning till de olika medlemsföreningarna inom Statistikfrämjandet. Samtliga föreläsningar filmades och kommer efter lite redigering att läggas upp på Statistikfrämjandets hemsida, så att de blir tillgängliga även för dem som inte var med på konferensen. Föreläsarna var följande:

- **ARVID SJÖLANDER**, docent vid Karolinska institutet, gav en föreläsning med fokus på

kausal inferens. Arvids föreläsning byggde på en egen tvillingstudie och behandlade hur man ska ta hänsyn till confounders, mediatorer och colliders vid analys för att på bästa sätt mäta kausal effekt.

- **MATTIAS VILLANI**, professor vid både Linköpings universitet och Stockholms universitet, gav en föreläsning med fokus på användandet av statistik inom artificiell intelligens och machine learning. Mattias beskrev hur man med olika metoder som t.ex. dynamic multi-layered networks och diagonal orthant latent dirichlet allocation (DOLDA) kan lösa problem.
- **OLA ANDERSSON**, anställd hos Ivbar, gav exempel på arbetsuppgifter i samband kvalitetssäkring inom hälso- och sjukvård, vilket ligger inom Ivbars arbetsområde. Anknytningen var här till Föreningen för industriell statistik.
- **NURGUL BATYRBETAKOVA**, Scandinavian Development Services, gav exempel på arbetsuppgifter som biostatistisk konsult när det gäller statistiska metoder för medicinska tillämpningar Dessutom redovisades erfarenheter som doktorand. Anknytningen var här till Föreningen för medicinsk statistik.
- **KARIN NELSON**, Inizio, gav exempel på arbetsuppgifter i samband med surveyverksamhet. Inizio är ett ledande institut för opi-

nionsmätningar i Sverige. Anknytningen var här till Surveyföreningen.

- **STEFAN EKBERG**, Folksam, berättade om hur det är att jobba som aktuarie inom försäkringsbranschen, där Folksam är ett av de största försäkringsbolagen i Sverige. Anknytningen var här till Cramérsällskapet.
- **IDA HED MYRBERG**, Karolinska institutet, berättade om hur det är att som nyutexaminerad statistiker komma till medicinsk forskning som statistisk konsult. Anknytningen var även här till Föreningen för medicinsk statistik.

Vid arrangemanget deltog även styrelsemedlemmar från Statistikfrämjandet och Surveyföreningen för att i pauser informera om sin verksamhet. Statistiska centralbyrån sponsrade arrangemanget och informerade även om sin verksamhet. Statistikfrämjandet gav ett resebidrag till deltagande studenter utanför Stockholms närområde.

Den generella uppfattningen om arrangemanget var att det var mycket uppskattat av studenterna såväl som av föreläsarna. Målsättningen är därför att göra konferensen till en regelbunden aktivitet som arrangeras av Statistikfrämjandet.

FREDRIK NORSTRÖM



Har du flyttat?

■ Du kan själv ändra dina uppgifter genom att gå till <https://www.membit.net/m4-member/login> där du loggar in med dina personliga inloggningsuppgifter (finns på din senaste inloggningsavi). Vid frågor kontakta Mattias Strandberg på sekrfram@gmail.com.

Något om P-värden

Inledning

I denna text tänkte jag ge en liten ingång till P-värden och debatten kring dessa, som fått fart igen på senare tid. Redan från början ska jag säga att jag inte har lusläst allt av det senaste som har skrivits. Jag utgår ifrån något av det, samt även från en del annat som jag tidigare har hört och läst. Min ambition är att till stor del hålla mig på en ganska elementär nivå, så att läsare med olika stora förkunskaper i ämnet alla kan få ut något av texten.

Vad är ett P-värde?

Först av allt, vad är egentligen ett P-värde? (Den läsare som redan är väl förtrogen med begreppet kan med fördel hoppa över detta avsnitt.)

Låt oss tänka oss att en ny medicinsk behandling, eller drog, har tagits fram, och att vi vill testa om den är bättre än den tidigare behandlingen. I mitt tänkta

exempel gäller det en blodtryckshöjande medicin. Vi har en slumpmässigt utvald patientgrupp, där var och en först får ta den gamla medicinen. Sedan mäter man blodtrycket på varje patient.

Efter en kort behandlingspaus ger man sedan den nya medicinen till samma grupp av patienter. Blod-

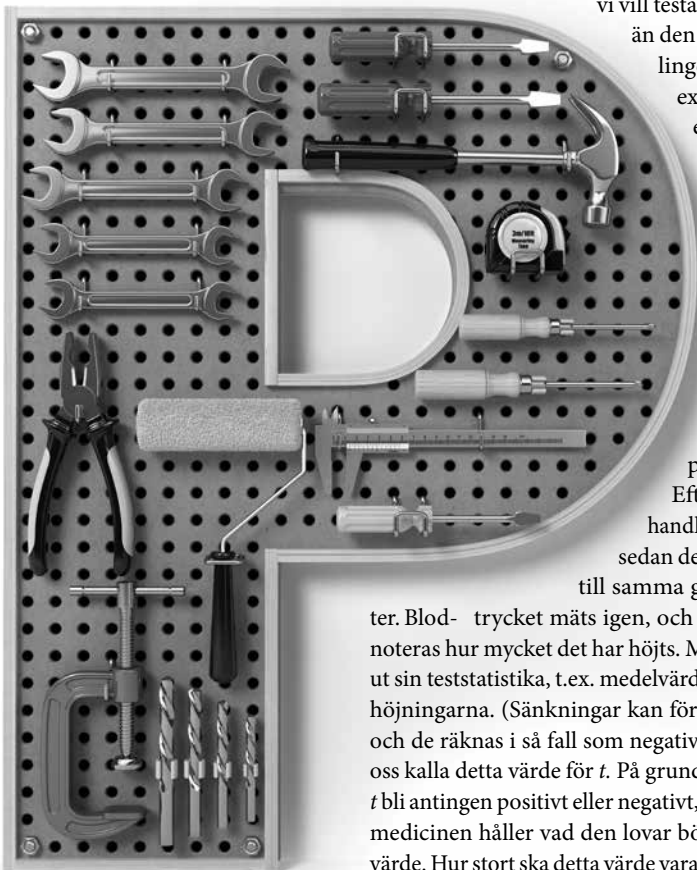
trycket mäts igen, och för varje patient noteras hur mycket det har höjts. Man räknar sedan ut sin teststatistika, t.ex. medelvärden av blodtryckshöjningarna. (Sänkningar kan förstås förekomma, och de räknas i så fall som negativa höjningar.) Låt oss kalla detta värde för t . På grund av slumpen kan t bli antingen positivt eller negativt, men om den nya medicinen håller vad den lovar bör t få ett positivt värde. Hur stort ska detta värde vara för att vi ska våga

tro på att den positiva avvikelser från noll (inte bättre än den gamla medicinen) inte bara beror på slumpen?

Om t är tillräckligt stort säger vi att den nya medicinen har en "signifikant större effekt" än den gamla, och det är här som P-värdet kommer in. Vi säger att vi testar *nollhypotesen* att medicinerna har lika effekt mot *alternativhypotesen* (eller *mothypotesen*) att den nya har större effekt. För att göra detta räknar vi ut vårt (observerade) P-värde såsom: sannolikheten att få ett t -värde lika stort som, eller större än, det vi faktiskt har observerat, under förutsättning att nollhypotesen om ingen effektskillnad är sann. Om denna sannolikhet är tillräckligt liten, säg mindre än fem procent, anser vi att det inte är troligt att vi skulle kunna få ett sådant utfall *ifall nollhypotesen var sann*. Vi *förkastar* då vår nollhypotes till förmån för alternativet, alltså att den nya medicinen har en större positiv effekt än den gamla. (Vi har från början utestutit att den har mindre effekt.)

Ett annat sätt att göra detta är att först fastlägga en *signifikansnivå*, säg på fem procent, sedan teoretiskt räkna ut hur stort t behöver vara för att vi ska förkasta nollhypotesen, och till sist undersöka om vårt observerade t är så stort eller inte. Rent logiskt blir denna procedur likvärdig med den som beskrevs ovan. Det faktum att P-värdet är mindre än fem procent (säg) motsvaras nämligen precis av att man förkastar nollhypotesen, vilket ju i sin tur innebär att vårt observerade t är "tillräckligt stort" (vi har en signifikant effektskillnad).

Ett annat exempel är politiska opinionsundersökningar, där man t.ex. vill studera om andelen sympatisörer för ett visst parti har förändrats jämfört med en tidigare undersökning, eller om den observerade förändringen ligger "inom den statistiska felmarginalen". Då är nollhypotesen att andelen inte har ändrats, och mothypotesen är att den har det, åt endera hållet. I detta scenario är alltså mothypotesen tvåsidig, och vi förkastar nollhypotesen vid stora värden av förändringens absolutvärde, dvs. hur långt från noll den är, oavsett tecken. (Vi har lärt oss i ett tidigare nummer av Qvintensen att det långt ifrån alltid är så att urvalen är slumpmässiga i opinionsundersökningar, och i så fall kan man inte beräkna någon statistisk felmarginal.) »»»



NÅGOT OM P-VÄRDEN

Mer allmänt har vi två likvärdiga definitioner av p-värdet:

1. P-värdet är sannolikheten att få det observerade utfallet *eller något extremare i riktning mot alternativhypotesen*, beräknad under förutsättningen att nollhypotesen är sann.
2. P-värdet är *den minsta signifikansnivån* för vilken vi kan förkasta nollhypotesen.

Här är den första definitionen mer intuitiv, medan den andra kanske kan kännas mer matematiskt stringent.

För den läsare som vill veta hur man konkret kan räkna ut P-värdet i olika exempel (det kommer också ett sådant nedan), samt få en fördjupad förståelse av P-värdet och relaterade begrepp, hänvisar jag till Alm och Britton (2008), som är den bok vi använder då vi undervisar i inferensteori på elementär nivå för studenter på kandidatprogrammet i matematik i Uppsala. En annan bra referens, med inriktning mot biostatistik, är Holm (2008).

Två diskussioner i en

Som jag ser det verkar det finnas två utgångspunkter för diskussionen om P-värden. Den som kanske mest har lyfts fram i senare artiklar handlar om den magiska gränsen 5%. Men det finns också en vidare aspekt, nämligen om man det överhuvudtaget är lämpligt att använda sig av begreppet P-värde vid statistisk slutledning. Det är främst detta jag kommer att diskutera här. Men jag inleder med något om den första frågan.

Varför fem procent?

Kanske tycker läsaren att det känns lite väl riskabelt med en signifikansnivån så stor som fem procent i mitt blodtrycksexempel ovan? Detta innebär ju i praktiken att om man testar tjugo mediciner på raken som inte är bättre än den gamla, så kommer i genomsnitt en av dem att (falskeligen) uppvisa en signifikant förbättring! Om signifikansnivån hade valts till något lägre, t.ex. en procent, skulle detta bara ske en gång på hundra. Å

andra sidan skulle en alltför rigid nivå, t.ex. 0,001 procent, göra att kravet på att förkasta nollhypotesen blev så hårt att detta i praktiken aldrig skulle kunna inträffa även om det vore berättigat! Vad ska man egentligen välja?

Varför har just signifikansnivån fem procent blivit ett så populärt val? Den historiska artikeln av Kennedy-Shaffer (2019) ger några svar. Den berömda statistikern och genetikern R.A. Fisher var en av de första som föreslog nivån fem procent. Alla som någon gång gått på en elementär kurs i statistik har fått lära sig att man (vid fem procents signifikansnivå och normalfördelningsapproximation) förkastar en enkel nollhypotes mot en tvåsidig mothypotes, om den standardiserade teststatistikan antar ett absolutvärde större än 1,96. På motsvarande sätt konstrueras ett 95-procentigt konfidensintervall (vid approximation med normalfördelning) som en skattning plus/minus 1,96 gånger standardfelet för skattningen. Och 1,96 är ju ungefär lika med 2, så grovt kan man säga att om det observerade värdet för skattningen avviker från det värde som anges av nollhypotesen med mer än två standardfel åt något håll, så har vi en signifikant stor avvikelse. En enkel och naturlig regel, eller hur?

»Som jag ser det verkar det finnas två utgångspunkter för diskussionen om P-värden.»

Detta är nog bekant för de flesta som läser denna text. Men i det ovan nämnda tidskriftsnumret ges en annan intressant vinkling av

Greenland (2019), som i alla fall var ny för mig när jag läste om den. Tänk dig att du har ett mynt, och att du med upprepade kast vill ta reda på om chansen att krona kommer upp är 0,5 (alltså att myntet är välbalanserat). Du misstänker att denna chans kan vara större. Säg nu att du bestämmer dig för att göra fyra kast. Hur många gånger ska du få krona på dessa kast för att du inte längre ska tro att myntet är välbalanserat, till förmån för att det ger krona oftare än klave? Svår fråga, men kanske borde du inte nöja dig med mindre än att du får krona i alla de fyra kasten? Och vad är sannolikheten

för detta, ifall myntet är välbalanserat (alltså P-värdet)? Jo, det är ju en halv upphöjt till fyra, alltså $1/16 = 0,0625$. Man kan också uttrycka detta i termer av den s.k. Shannon-entropin, S , som är minus två-logaritmen av sannolikheten (det tal man ska upphöja en halv till för att få denna sannolikhet). I detta fall är $S = 4$. För sannolikheten 0,05 blir Shannonentropin cirka 4,3. Så även här kommer 5% in som ett ganska rimligt P-värde, om man så vill. (I detta exempel ser vi att det minsta P-värde man kan få, 0,0625, är större än fem procent, så här är försöket för litet för att man ska ha någon möjlighet att förkasta nollhypotesen på signifikansnivån 5 procent.)

Varför P-värde?

Jag går nu över till den mer allmänna frågan om huruvida P-värdet som sådant kan vara ett lämpligt verktyg vid statistisk slutledning.

I nyss nämnda artikel (Greenland, 2019) listas och bemöts ett antal vanliga argument som talar emot användning av P-värdet. Jag rekommenderar här läsaren att bilda sig sin egen uppfattning, men mitt intryck är att listan mest består i argument mot traditionell frekventistisk hypotesprövning, snarare än mot P-värdet i sig. Till exempel diskuteras det välkända problemet att, givet noll- och mothypotes, P-värdet tenderar att minska då stickprovsstorleken ökar. (Alltså: sannolikheten för små P-värden ökar.) På så vis kan det lilla P-värdet som fås vid stora stickprov tyckas ge ett "bevis" för att nollhypotesen är falsk, trots att avvikelsen kanske är så liten att den är praktiskt ointressant. I mitt inledande exempel med medicinerna skulle detta motsvaras av att patientgruppen är så stor att även en väldigt liten höjning av blodtrycksmedelvärdet leder till ett lågt P-värde, och därmed till förkastande av nollhypotesen om ingen skillnad.

Omvänt ska ett stort P-värde inte tolkas som ett bevis för att nollhypotesen är sann. En rimligare tolkning är att vi inte har tillräcklig information för att förkasta nollhypotesen på den önskade signifikansnivån. (Till exempel kunde en förespråkare av det gamla läkemedlet med flit använda en så liten patientgrupp

att endast en mycket stor höjning av blodtrycksmedelvärdet skulle betraktas som signifikant.)

Men detta är ju, som sagt, inte en kritik av P-värdet i sig, utan snarare av den allmänna metodik där P-värdet ingår som ett viktigt verktyg.

En allvarigare aspekt på P-värdet tas upp i Schervish (1996). Närmare bestämt handlar det om logiska problem som uppstår om man tolkar P-värdet som ett mått på stödet för nollhypotesen.

Säg att vi på signifikansnivån 0,05 vill testa om en okänd parameter μ har värdet m . Till detta använder vi teststatistikan T , som antas (approximativt) ha en standardiserad normalfördelning, $N(0, 1)$, ifall nollhypotesen är sann. Om alternativhypotesen är att μ är skild ifrån (större eller mindre än) m ska vi förkasta nollhypotesen om det observerade värdet på T är större än 1,96 eller mindre än -1,96. Vi kan också ha alternativet att μ är större än m , och därmed låta alla μ som är mindre än eller lika med m ingå i nollhypotesen. Då ska vi förkasta om T är större än 1,645.

Så långt allt väl. Säg nu att vi observerar värdet $T=1,7$. Eftersom $1,7 < 1,96$ förkastar det tvåsidiga testet inte nollhypotesen på nivån 5%. P-värdet är alltså större än 5% (i själva verket cirka 9%). Men det ensidiga testet förkastar nollhypotesen: vi har ju $1,7 > 1,645$. Så för det ensidiga testet är P-värdet mindre än 5% (cirka 4,5%). Om man nu tolkar ett större P-värde som ett större stöd för nollhypotesen så finner man att nollhypotesen med endast $\mu = m$ har ett större stöd (P-värde 9%) än vad nollhypotesen med det större området $\mu \leq m$ har (P-värde 4,5%). Med andra ord har vi ett kraftigare "bevis" för att $\mu = m$ än för att $\mu \leq m$, trots att det förra logiskt ingår i det senare!

Kommer man ifrån detta logiska problem om man avstår från att jämföra nollhypoteser av olika typ? Nej, Schervish visar att om man endast betraktar nollhypoteser i form av slutna intervall (att parametern ligger mellan två på förhand bestämda värden eller är lika med något av dem), så kan P-värdet för ett intervall

vara större än för ett annat intervall som det ingår som en del i.

Så detta visar att det inte bara är olämpligt att tolka P-värdet som ett mått på hur väl en nollhypotes stämmer med data, det är även logiskt fel!

Vad säger Bayesianerna?

Kritik av frekventistiska procedurer, alltså till exempel sådana som använder P-värden, mynnar ofta ut i att det är mer tilltalande att använda Bayesianska metoder. Som många läsare säkert vet går sådana metoder ut på att väga samman vad man på förhand (a priori) "tror" om parametern/parametrarna och vad man observerar till vad man i efterhand (a posteriori) ska "tro" (vad sannolikheterna blir för olika möjliga parametervärden). Det Bayesianska alternativ till P-värdet som brukar föras fram är den så kallade Bayesfaktorn: kvoten mellan sannolikheten för att observera det värde man fick på teststatistikan uträknad under nollhypotesen dividerad med motsvarande kvantitet uträknad under mothypotesen. Ju större värde som erhålls på denna kvot, desto större "stöd" har nollhypotesen.

Detta låter ju bra, men tyvärr visar Lavine och Schervish (1999) att Bayesfaktorn lider av samma problem som P-värdet då den tolkas som ett mått på stödet för nollhypotesen. Om man däremot beräknar det så kallade a posteriori oddset för nollhypotesen jämfört med mothypotesen, vilket är något annat, så råkar man inte ut för problemet.

Diskussionen om frekventism kontra Bayes är intressant, men alltför komplicerad för att rymmas inom denna text. Låt mig bara konstatera att fördelar med den Bayesianska analysen är den enkla tolkningen och dess logiska koherens, men den stora nackdelen (om man strävar efter så objektiva slutsatser som möjligt) är att den kan tyckas vara mer subjektiv eftersom den kräver att en apriorifördelning för parametrarna specificeras. Dock tror jag att de flesta statistiker numera har en ganska pragmatisk syn på det hela, och väljer den metodik som passar bäst beroende på situation.

Konfidens

Ett alternativ till att presentera P-värdet för ett test är att ge ett konfidensinter-

vall för någon rimlig konfidensnivå. Om denna konfidensgrad till exempel är 95 procent så betyder ju detta att alla parametervärden som ingår i konfidensintervallet är sådana att de inte skulle förkastas på signifikansnivån fem procent, om de bildade en enkel nollhypotes. På detta sätt kommer man ifrån lite av problemet med att för-

kasta för ointressant små avvikelser från nollhypotesen som jag beskrev ovan. På konfidensintervallet ser man ju precis hur stora (eller små) avvikelser som är "rimliga".

Å andra sidan är ju en fördel med P-värdet att det ger information om på vilka signifikansnivåer man kan förkasta nollhypotesen. Denna information mister man om man bara presenterar ett enda konfidensintervall med fastlagd konfidensnivå. Lösningen på dilemmaet är att i en figur åskådliggöra de konfidensintervall man kan få på alla möjliga nivåer. Detta kallas också för ett konfidensnät.

Hur kan detta åstadkommas? Ett sätt är att tänka sig alla ensidiga test med alternativet att parametern är större än det under nollhypotesen fixerade värdet. Låt det observerade värdet på teststatistikan vara givet. Sedan ritar man upp en funktion (kurva) som visar P-värdet för varje sådant test som funktion av parametervärdet under nollhypotesen, då detta får variera över hela x-axeln. Vad får denna funktion för egenskaper? Ja, om man låter parametern gå mot minus oändligheten så kommer P-värdet att gå mot noll, eftersom alternativet täcker fler och fler möjliga värden (det blir mer och mer säkert att vi förkastar). Går den istället mot plus oändligheten går P-värdet mot ett. Ju större parametervärdet under nollhypotesen är, desto svårare blir det ju att förkasta mot alternativet att det har ett ännu större värde. På detta sätt ser man också att funktionen måste vara monotont växande i parametern. Alltså uppfyller den samma villkor som en fördelningsfunktion. Vi säger att vi har

»Dock tror jag att de flesta statistiker numera har en ganska pragmatisk syn på det hela, och väljer den metodik som passar bäst beroende på situation.«

»»» NÅGOT OM P-VÄRDEN

ritat upp konfidensfördelningen. Deriveras funktionen erhålls konfidenstätheten. Den kan ses som ett alternativ till den Bayesianska a posterioritätheten för parametern. Observera finessen att ingen a priorifördelning behövs!

Hur får vi ur detta en graf med alla konfidensintervall? Jo det är ”bara” att inverta konfidensfördelningen, med hänsyn tagen till att ett tvåsidigt konfidensintervall kan fås genom att kombinera två ensidiga.

För mera läsning om konfidens hänvisas till översiktsartikeln Xie och Sing (2013) och till Schweder (2003), som ger en relativt enkel förklaring av metoden och tillämpar den på räkning av valar. Tilläggas kan att redan R.A. Fisher var inne på dessa tankebanor. Han talade om fiduciella sannolikheter för parametern. Det visade sig att Fishers idé inte riktigt höll måttet då man studerade andra situationer än de allra enklaste, och han har fått mycket kritik för den genom åren. Nu är det så att konfidensfördelningen bara sam-

manfaller med den fiduciella i just de enkla fallen där Fishers idé förefaller att fungera väl. Så det finns hopp om att detta område kan visa sig bli ett intressant alternativ både till Bayesianska metoder och till andra frekventistiska.

Avslutande kommentarer

Kanske kan min text tolkas som att jag tycker att det är ganska bra att använda den fixerade signifikansnivån fem procent men att P-värdet i sig är problematiskt. I så fall är det för att jag skrivit som ”djävulens advokat”, för jag tycker nog egentligen tvärtom. Jag tycker att man kan se P-värdet som en ”mjuk övergång” mellan en ”hård” signifikansnivå och en rent subjektiv analys. Det är skadligt att stirra sig blind på om P-värdet är större eller mindre än fem procent. Men P-värdet i sig är, använt på rätt sätt, ett fint verktyg i den statistiska verktygslådan!

ROLF LARSSON,
PROFESSOR I
TILLÄMPAD MATEMATIK
OCH STATISTIK,
UPPSALA UNIVERSITET

Referenser

- Alm, S.-E. och Britton, T. (2008). *Stokastik – Sannolikheteori och statistikteori med tillämpningar*. Liber.
- Greenland, S. (2019). Valid P-values behave exactly as they should: Some misleading criticisms of P-values and their resolution with S-values. *The American Statistician*, 73:sup1, 106-114. [DOI: 10.1080/00031305.2018.1529625]
- Holm, S. (2008). *Biostatistisk analys*. Studentlitteratur.
- Kennedy-Shaffer, L. (2019). Before $p < 0.05$ to beyond $p < 0.05$: Using history to contextualize p-values and significance testing. *The American Statistician*, 73:sup1, 82-90. [DOI: 10.1080/00031305.2018.1537891]
- Lavine, M. och Schervish, M.J. (1999). Bayes factors: What they are and what they are not. *The American Statistician*, 53, 119-122.
- Schervish, M.J. (1996). P values: What they are and what they are not. *The American Statistician*, 50, 203-206.
- Schweder, T. (2003). Abundance estimation from multiple photo surveys: Confidence distributions and reduced likelihoods for bowhead whales of Alaska. *Biometrics*, 59, 974-983.
- Xie, M. and Singh, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: a review. *International Statistical Review*, 81, 3-39.

Defending the P-value

Introduction

If you are an applied statistician, especially in medical areas, you are very likely using significance tests and producing P-values regularly. So it may not escape your attention that there are numerous attacks on the P-value. These attacks are of course not new, but the recent ones appear to come from mainstream sources and their target range has widened. The level varies from benign, via medium, to severe. On the benign end: Benjamin et al (2018) in *Nature Human Behaviour* simply suggested that we stop using 0.05 cutoff and start using 0.005 instead. Amrhein et al (2019) in *Nature* stepped up the attack and reported more than 800 supporting votes, including many in academia, in their call to “retire” statistical significance. They are not banning the P-value per se, but calling for the dismantling of significance testing. According to them, we should not say “statistically significant” and “not statistically significant”; in general we should “quit categorizing”. They do not mind confidence intervals, but are of course against using them as the basis for claiming statistically significant results. “[A]uthors will emphasize their estimates and the uncertainty in them – for example, by explicitly discussing the lower and upper limits of their intervals.”

Going further on the severity spectrum: the journal *Epidemiology* bans the P-value altogether, but not confidence intervals; actually this has been its practice since its founding in 1990. On the most extreme end of the spectrum, the journal *Basic and Applied Social Psychology* since 2015 had banned the P-value, or more generally the Null Hypothesis Significance Testing Procedure. They claim that this procedure is invalid, because it suffers from the ‘inverse inference problem.’ The confidence interval also suffers from the same problem, so it is also banned. For the Bayesian procedures, the Editors “reserve the right to make case-by-case judgements”, so they are neither required nor banned.

The non-technical meta-literature on the P-value is enormous. Just the special issue of *The American Statistician* – an official journal of the American Statistical Association – in March 2019 alone contains 43 papers (401 total pages). It is practically

impossible for a person to absorb the diversity of opinion among the experts, let alone the non-experts. What I find most interesting is perhaps the explicit opinion of the Editors, as to some extent they represent mainstream academic statisticians. Echoing Amrhein et al (2019), the Recommendation 2 in their editorial clearly states “Don’t Say ‘Statistically Significant’, and they wrote: “... it is time to stop the term ‘statistically significant’ entirely. Nor should variants such as ‘significantly different’, ‘ $p < 0.05$ ’, and ‘nonsignificant’ survive.”

This debate is not likely to die out soon. To appear in the December 2019 issue of the *Journal of American Medical Association (JAMA)* Psychiatry is an article “Is It Time to Ban the P-value?” by Helena Kraemer from the Department of Psychiatry and Behavioral Sciences, Stanford University, not from some out-of-the-way places. Book-length treatment is not as common. The title *The Cult of Statistical Significance* and the subtitle *How the Standard Error Costs Us Jobs, Justice, and Lives* of the book by Stephen Ziliak and Deirdre McCloskey (2008) clearly convey the views of the authors.

How are we to react to these attacks on the P-value and significance test? As you continue to read this article it should become obvious which side I am on, so I might as well say it explicitly now: Banning P-value or significance test is a bad idea; “quit categorizing” defeats the purpose of scientific investigation; reducing the threshold from 0.05 or 0.005 is sensible for new discoveries, but not a good idea as a blanket rule.

When the P-value is unavoidable

There are a number of common statistical problems and applications where banning the P-value is awkward and not sensible. Here are some of them:

- Displaying and comparing Kaplan-Meier survival curves. These are among the most common plots in medical applications. What if we are not allowed to report the standard log-rank statistic and

»How are we to react to these attacks on the P-value and significance test?«

»»» DEFENDING THE P-VALUE

its P-value? We could report the hazard ratio and its confidence intervals, but this is model-dependent, for example it assumes proportional hazards. We could provide confidence bands, but then what do we state: are we allowed to focus on the areas where the confidence bands do not overlap? On the technical side, there is of course an issue of simultaneous inference across time axis, but this effort would also violate the idea of not using confidence intervals for significance testing.

- Wilcoxon and non-parametric rank tests in general. Technically there is an underlying parameter associated with a rank test, but who in practice knows what specific parameter is being tested by the Wilcoxon test? It is not at all clear whether an estimate and confidence interval for such a parameter, e.g. using the Hodges-Lehmann estimator, are informative for the readers.
- Trend tests across an ordinal axis; e.g., is there a trend in the risk of death across the low-medium-high exposure group? The test may depend on an arbitrary scaling of the x-axis, so the estimated slope and its confidence interval are not meaningful numbers.
- Any test with more than one degree of freedom: ANOVA tests of multiple groups or interactions, or χ^2 test of independence or goodness-of-fit or model-checking. When we test the association between, say, 4 eye colors and 3 hair colors, the χ^2 -statistic has $(4-1) \times (3-1) = 6$ degrees of freedom. This means there are 6 independent underlying parameters that are being tested. What are they? It is possible to write them down, but they are simply not interesting. Providing the estimate and confidence interval for each would be absurd.
- Because of its importance in molecular medical research, in the genome-wide association studies, the plot of a very large collection of P-values – also called the Manhattan plot – is a primary output of the analysis. It is simply not sensible to show one million confidence intervals.
- In the ubiquitous multiplicity problems, the P-value is often the raw ingredient for further adjustment to account for multiplicity. It is very simple

to adjust the P-value, for example, using the Bonferroni correction; if estimates and confidence intervals are given, with some calculations, though not in your head, you can recompute the confidence intervals to adjust for multiplicity, but this is rarely done.

- From a theoretical point view, the construction of confidence interval requires a continuous parametric model, or at least semi-parametric, whereas the P-value only requires one model under the null hypothesis. For example, in spatial point-process applications we first want to test whether a homogeneous Poisson process fits the data. If yes, there is not much more we can investigate. It does not make sense to build a more complex nonhomogeneous Poisson model, for example, a competition process or a cluster point process, and to provide an estimate and confidence interval for the parameters of that process.
- A great contribution of statistics is in the theory of – or at least in the explicit considerations for -- design of experiments. Above all, thinking about design issues prior to starting the study forces a good discipline on the experimenters. Theoretically the concepts of testing, test statistic, significance level, power, and sample size are closely connected. Banning the P-value must also mean banning the test statistic, since there is one-to-one map between them. Without the test statistic it does not make sense to talk about power and sample size. Overall, much of the good statistical ideas from the theory of design of experiments would have to be abandoned.

Reasons to ban the P-value

Let us assess some commonly stated reasons to ban the P-value. Number one, because people often misinterpret it as the probability of the null hypothesis. Is that really a mortal sin? P-value *is not* the probability of observed data given the null hypothesis (that would be the likelihood of the null). P-value is more complex: the probability of the observed

data or *something more extreme than the observed data* assuming the null hypothesis is true. So P-value is not a transversal of the Bayesian posterior:

$$\text{P-value} = P(\text{observed data or more extreme} | H_0)$$

$$\text{Bayesian posterior} = P(H_0^* | \text{observed data}).$$

(The notation H_0^* is used for the Bayesian posterior to indicate it may not be exactly the same set as H_0 for the P-value. E.g. $H_0: \mu = 0$, but $H_0^*: \mu \leq 0$ for a typical testing of the normal mean parameter.)

In regular problems, the kind that lead to complaints about the inferential meaning of P-value, the P-value is indeed not a probability of the null hypothesis, but *it is* a valid measure of “confidence” (Pawitan 2001, Chapter 5). This non-Bayesian concept of confidence is the same “confidence” that we use in ‘confidence interval,’ so it is a main-stream concept. Unfortunately it is rarely used in its full capacity as a measure of uncertainty. The recent book by Schweder and Hjort (2016) gives a comprehensive exposition. Thus, we may report, say, the 95% confidence interval $2.1 < \mu < 3.9$, meaning that specific interval has 95% confidence. The (one-sided right-side) P-value for the null hypothesis $H_0: \mu = 0$ is the *confidence associated with $\mu \leq 0$* . Furthermore, theoretically, the confidence distribution is a Bayesian posterior for a certain implied prior distribution (implied by the

choice of P-value) (Fraser, 2011; Efron and Tibshirani, 1993, Chapter 24). In summary, mathematically we may interpret the P-value as a confidence measure and as a Bayesian posterior probability. So the reason to ban the P-value because it is not a probability of null hypothesis is not fully justified. It is a confidence measure in a relevant null hypothesis (such as $\mu \leq 0$ for one sided P-value) and it has the same logical status and epistemic value as the Bayesian posterior probability for a certain choice prior distribution. Now we can see also that to ban the P-value, but “might consider” the Bayesian alternative, shows a lack of theoretical understanding.

»Banning the P-value must also mean banning the test statistic, since there is one-to-one map between them.»

The second common reason: people use the P-value carelessly, thus produce many misleading results, primarily false positives or hyped claims. We know this because many results indeed fail to replicate. Statisticians should of course care about not producing misleading results. Most of our theories in statistics are based on optimizing error rates. But, as we all know, there is a type-I error and a type-II error. Whatever we say about a hypothesis, there is a chance we make an error. If we declare significance we may make type-I error, which is a sin of commission; if non-significance, we might commit type-II error, a sin of omission. A good procedure should balance between the two errors, and a good statistician or scientist should be aware of both at all times. But how is that achieved by banning the P-value?

Benjamin et al (2018)'s modest suggestion of reducing the P-value threshold of significance to 0.005 goes a long way towards reducing false positive rates. But even this sensible suggestion cannot be applied as a blanket rule to all studies, because it will increase the type-II error rate. For example, for a validation study, a 0.05 threshold could be sufficient, especially if the validation study has a small sample size, so it is not well powered. Together with the original study, the cumulative evidence for real effect is much stronger. But, in practical scientific activities, this is where a grey area appears: what constitutes a 'validation' study? Does it have to be by the same research group as the original discovery? If not, does it count as validation if the idea of the study comes after reading a journal article? But scientists will argue that most of their studies are based on a chain of previous studies and results, either by themselves or by their collaborators or by complete strangers; the studies are never completely new, based on an idea that appears out-of-blue. Also, to be a 'validation' study, does the study have to be as similar as the original? The original study could be done in mice, does a human version constitute a validation? A phase-III clinical trial to get the US Federal Drug Agency's approval for a drug typically use 0.05 threshold for significance. At that point the drug has already undergone thorough and numerous investigations to establish its

biological basis, so the human trial is a form of validation. Yet, as a practicing statistician who has produced literally millions of P-values – in genetics/genomics applications -- I know that a P-value around 0.05 is highly unstable; it can easily move to 0.10, for example, after a re-definition of inclusion or exclusion criteria, or a change in the handling of missing values. So in practice it is rarely warranted to get too excited about novel discoveries with P-value less than but near 0.05.

Perhaps the most legitimate concern about the P-value is that it does not capture the increased uncertainty due to various steps and decisions – some documented and some undocumented -- we make during data analysis: choice of wording in a questionnaire, definition of new variables for analysis, choice of transformation, choice of split when analyzing categories, choice of model class, decisions regarding inclusion and exclusion criteria, choice of variable(s) for sub-group analysis, handling of missing data, etc. Explicit model selection, for example, the best-subset regression, is known to produce optimistic P-values from the final selected model. But how do we account for all various steps, some of which are often taken with the thought of increasing the sensitivity of the experiment, hence potentially giving an optimistic P-value?

Overcoming the weakness of the P-value as a discovery tool, science has a safety measure by evolving a skeptical culture. It is rare for any scientific result to be accepted by the relevant research community – the group most knowledgeable with the details of the area, hence not easily fooled like the general public. The simple requirement is that a result gets established by external validation and reproducibility. The harder the validation the stronger the result would be. What constitute a "hard validation"? As I have indicated above, a scientific validation does not mean a result based on a study as similar as possible to the original study. In fact, it is better to validate in distinct studies. For example, a biological finding from human study is validated in zebra-fish, then in mice, then back in human. All relates to the same biology, but perhaps with different measurements. Good scientific results

should lead to a deeper theory, which in turn should produce novel predictions for future experiments. These future experiments, even with completely distinct measurements, will constitute a validation of the original study or theory. In summary, 'reproducibility' is a must in science, though it must be seen as a general concept.

Most opponents of the P-value and its 0.05 threshold point their finger to R.A. Fisher, one of the greatest statisticians of the 20th century. With his highly influential *Statistical Methods for Research Workers*, including the practical Tables for various common statistics, Fisher certainly popularized the 0.05 limit. In a paper in 1926 he wrote (with my emphasis):

... Personally, the writer prefers to set a low standard of significance at the **5 per cent** point, and ignore entirely all results which fail to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance.

As a practicing scientist, Fisher started with a "low standard" 0.05 limit, but note the key words "*rarely fail*" as it clearly implies that he required successful validation studies before accepting a scientific fact as established. He was also careful to state that this was *his personal preference*. The proper balance between type-I and type-II error is highly context-dependent. In research areas where discoveries are rare, or, if a potential discovery – if real – is fundamental, or, if false positives are easy to dismiss, it is not sensible to have a very stringent P-value cut-off, as it lowers sensitivity. In such a situation the researchers would be happy to pursue any reasonable lead. On the other hand, when potential discoveries are abundant and establishing a scientific fact is expensive – for example, in genome-wide association studies (GWAS) or in high-energy physics – one should be very stringent with P-value cutoff. The standard cutoff is 5×10^{-8} in GWAS, and 3×10^{-7} in high-energy physics.

How to quit categorizing?

I believe the proposal to ban the term "statistically significant" or "significantly different", and to "quit categorizing" is

Table 2. Key Secondary End Points.*				
End Point	TAVR (N = 496)	Surgery (N = 454)	TAVR vs. Surgery (95% CI)†‡	P Value§
New-onset atrial fibrillation at 30 days — no./total no. (%)¶	21/417 (5.0)	145/369 (39.5)	0.10 (0.06 to 0.16)	<0.001
Length of index hospitalization — median no. of days (inter-quartile range)	3.0 (2.0 to 3.0)	7.0 (6.0 to 8.0)	−4.0 (−4.0 to −3.0)	<0.001
Death from any cause, stroke, or rehospitalization at 1 year — no. (%)§	42 (8.5)	68 (15.1)	0.54 (0.37 to 0.79)	0.001
Death, KCCQ score of <45, or decrease from baseline in KCCQ score of ≥10 points at 30 days — no./total no. (%)	19/492 (3.9)	133/435 (30.6)	−26.7 (−31.4 to −22.1)	<0.001
Death or stroke at 30 days — no. (%)§	5 (1.0)	15 (3.3)	0.30 (0.11 to 0.83)	0.01
Stroke at 30 days — no. (%)§	3 (0.6)	11 (2.4)	0.25 (0.07 to 0.88)	0.02

Table 2.

»»» DEFENDING THE P-VALUE

downright unhelpful. Unfortunately the proposers do not give good examples how to report statistical results. So let us try this ourselves. Mack et al (2019) reported the result of the study on transcatheter aortic-valve replacement with a balloon-expandable valve in the *New England Journal of Medicine*. Here is how the result and conclusion paragraphs appear in the Abstract. Some offending parts are highlighted with bold or underline, and both if we quit categorizing and ban the P-values; they do cover much substance of the results:

RESULTS At 71 centers, 1000 patients underwent randomization. The mean age of the patients was 73 years, and the mean Society of Thoracic Surgeons risk score was 1.9% (with scores ranging from 0 to 100% and higher scores indicating a greater risk of death within 30 days after the procedure). The Kaplan–Meier estimate of the rate of the primary composite end point at 1 year was **significantly lower** in the TAVR group than in the surgery group (8.5% vs. 15.1%; absolute difference, −6.6 percentage points; 95% confidence interval [CI], −10.8 to −2.5; **P<0.001 for noninferiority**; hazard ratio, 0.54; 95% CI, 0.37 to 0.79; P=0.001 for superiority). At 30 days, TAVR resulted in a **lower rate of stroke than surgery (P=0.02)** and **in lower rates of death or stroke (P=0.01)** and **new-onset atrial fibrillation (P<0.001)**. TAVR also resulted in a **shorter index hospitalization than surgery (P<0.001)** and **in a lower risk of a poor treatment outcome (death or a low Kansas City Cardiomyopathy Questionnaire**

score) at 30 days (P<0.001). There **were no significant between-group differences** in major vascular complications, new permanent pacemaker insertions, or moderate or severe paravalvular regurgitation.”

CONCLUSIONS Among patients with severe aortic stenosis who were at low surgical risk, the rate of the composite of death, stroke, or rehospitalization at 1 year was **significantly lower** with TAVR than with surgery.

Here is what happens if we clean the Abstract from the offending parts, still allowing confidence intervals, but not P-values:

At 71 centers, 1000 patients underwent randomization. The mean age of the patients was 73 years, and the mean Society of Thoracic Surgeons risk score was 1.9% (with scores ranging from 0 to 100% and higher scores indicating a greater risk of death within 30 days after the procedure). The Kaplan–Meier estimate of the rate of the primary composite end point at 1 year was 8.5% in the TAVR group and 15.1% in the surgery group (absolute difference, −6.6 percentage points; 95% confidence interval [CI], −10.8 to −2.5; hazard ratio, 0.54; 95% CI, 0.37 to 0.79).

As a scientific statement, the last sentence looks acceptable, but, without categorical direction, the statement seems dry and harder to absorb than the original. The second half of the original abstract must go because all statements are categorical. Finally, as a plea to all proposers of ban-

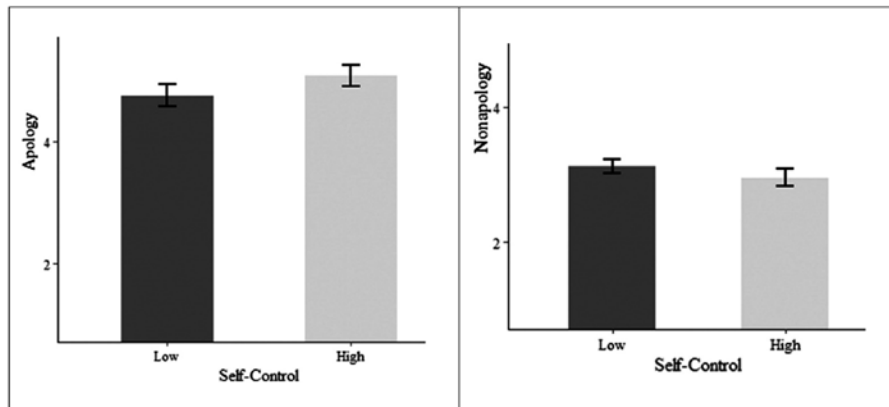
ning the significance test and the term ‘significantly different’, please tell us how to formulate the conclusion statement. If ‘significantly lower’ is not allowed, how about just ‘lower’? But it is still categorical, and it will be less informative to the readers. The second half of the Results paragraph looks like ‘sizeless science’ as complained by Ziliak and McCloskey (2008), but the results could of course be reported in terms of statistics and confidence intervals. They would just entail an undigested list of numbers as given in Table 2 of the paper (above). Thus effect sizes of the secondary end-points are indeed available within the paper, but the authors have to make choices due to limited space in the Abstract. Even stylistic considerations alone would stop most authors from simply giving a tedious list of numbers in their Abstract.

Life with any statistical inference

How is life without the P-value or any statistical inference in the journal *Basic and Applied Social Psychology* (BASP)? Here is one example from a recent article: “Sorry is the Hardest Word to Say: The Role of Self-Control in Apologizing” by Guilfoyle et al (2019). The Abstract is as follows; the key statistical results are highlighted.

“Apologizing is an effective strategy for reconciling relationships after transgressions. However, transgressors often resist or refuse to apologize. The current research investigated the role of self-control in apologizing. **In Study 1, self-control was associated with participants’ proclivity to apo-**

Figures 1 and 2 from Guilfoyle et al (2019: Effect of self-control on apology and nonapology. The error bars represent standard errors.



logize and apologetic and nonapologetic behavior. In Studies 2 and 3, self-control was manipulated to test the causal relationship. **Both studies found participants with high self-control were more apologetic and less nonapologetic and were more likely to use apologetic statements in e-mails to their victims.** Overall, these studies suggest that transgressors with high self-control are more apologetic than those with low self-control.”

Now, in the body of the paper, we can find the supporting evidence for the statements in the Abstract:

..., those in the high self-control group reported greater apology ($M = 5.08$, $SD = 1.45$) than those in the low self-control group ($M = 4.76$, $SD = 1.50$), $d = 0.22$ (Figure 1). We also found that those in the low self-control condition reported greater nonapology ($M = 3.12$, $SD = 0.79$) than those in the high self-control group ($M = 2.95$, $SD = 1.00$), $d = 0.19$ (Figure 2).

Hence, as promised, no statistical tests need to be reported, though it is not clear if the authors had actually performed them prior to submission to BASP. Interestingly, the text does not state the standard errors, but the figures do. Using the standard deviations (SDs) in the figures would give alarmingly large error bars; but it is arguable whether you should be allowed to show standard errors, since they are inferential quantities, not summary statistics of the data. Is this kind of reporting meant to be better than reporting with P-values? Remarkably, the

impact factor of BASP increased from 1.3 in 2015 to 3.4 in 2017, no doubt giving the editors a great feeling of vindication, but it then went down to 1.0 in 2018.

Like me, you are probably curious as to what kind of people want to ban the P-value and statistical inference. Among the 43 papers in the 2019 Special Issue of *The American Statistician*, there is a curious paper by Valentin Amrhein, David Trafimow and Sander Greenland. Now S. Greenland is well recognized to many of us, particularly for his text book *Modern Epidemiology*. V. Amrhein is the one who wanted to ban statistical significance, and D. Trafimow is the Editor of BASP. The article is titled “Inferential statistics as descriptive statistics: there is no replication crisis if we don't expect replication.” Here is the Abstract, which makes for strange reading for a scientist, and I highlight one sentence that perhaps captures the attitude of the authors:

“Nonetheless, considerable non-replication is to be expected even without selective reporting, and generalizations from single studies are rarely if ever warranted. Honestly reported results *must* vary from replication to replication because of varying assumption violations and random variation; excessive agreement itself would suggest deeper problems. ... Since a small P-value could be large in a replication study, and a large P-value could be small, **there is simply no need to selectively report studies based on statistical results.** Rather than focusing our study reports on uncertain conclusions, we should thus

focus on describing accurately how the study was conducted, what problems occurred, what data were obtained, what analysis methods were used and why, and what output those methods produced.”

In the text, we can read

“Reaching for statistical tests to force out “inferences” (whether traditional “ $p \leq \alpha$ ” testing or substitutes like tests using Bayes-factor criteria) is, like drinking alcohol, a culturally ingrained habit. Statistical testing (like alcohol) often gives the wrong impression that complex decisions can be oversimplified without negative consequences.... And many researchers are addicted to such oversimplification. These addictions are worth breaking.”

Conclusions

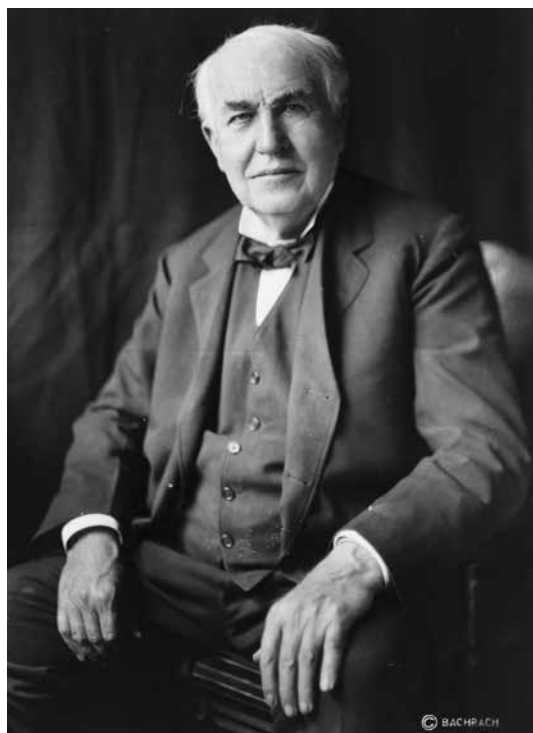
I guess the people who write proposals to ban the P-value or significance tests, or the editors of journals that ban them, are well-meaning with no hidden agenda. I cannot see any profit to be made by holding such a point of view. However, we know that it is possible to be well-meaning and unhelpful. There is a tendency in their writing to consider only the common problem of comparing two groups with simple parametric one-degree of freedom tests; I mentioned several common problems where it is awkward not to use the P-value. The unfortunate result is that good suggestions – e.g. to pay attention to the estimates of effect sizes – are mixed with bad proposals and arguments. For example, Ziliak and McCloskey (2008) invented the term ‘sizeless

»»» DEFENDING THE P-VALUE

science' as the straw man in their polemical book. In practice, when results are written down for publications, other factors come into play. As we show above, in the Abstract section of a paper, limited space considerations might lead to few details about effect sizes. But it would be highly unusual for the Results section not to report effect sizes.

In the process of developing the light-bulb, between 1878 and 1880 the legendary inventor Thomas Alva Edison (and his associates) famously worked on at least 3,000 different theories, trying out no fewer than 6,000 materials. Let us imagine his state of mind during the period: would he be arbitrary, not paying attention to the signal-to-noise ratio? How stringent would he be in the beginning investigation of a new material or design? How stringent would he be when a filament is showing promise? Finally, how stringent would he be on the design that he would be happy to patent? I imagine, like Fisher, most scientists today would work in a similar way: in the beginning quite happy looking for many possible leads, not afraid of false positives, but not going in completely arbitrary direction either; and, once there is a promising direction, putting a lot of efforts to validation studies.

The biggest difference is of course that Edison did not have to publish anything. Modern scientists have to publish even not-so-final findings; once published, there is no control who would consume the information and how the information is further disseminated. The skeptical scientists have no problem with the tentative nature of scientific results, even when presented categorically, as they would expect that validation studies will eventually screen out false positives. But the



general public might get confused with this potentially chaotic process; this issue will deserve its own discussion session. The point I would like to make here is that banning the P-value and significance tests seems to be motivated by the wish to reduce/avoid public confusion of scientific results, which are tentative by their nature, but this wish is contrary to the daily activities and needs of practical scientists.

Finally, as is well-known in statistics we have a long-standing unresolved philosophical debate between the frequentist and the Bayesian schools. Most Bayesians do not accept the P-value; theoretically the P-value violates the likelihood principle (Pawitan, 2001, Chapter 7). And, they also claim that the P-value misrepresents the level of evidence against the null hypothesis. They would instead suggest the Bayes Factor as an alternative to the P-value. So, an attack on the P-value from the Bayesians is nothing new. It is possible that some of the current attackers are partly motivated by the old Bayesian dislike of the P-value, but the proposal to ban significance tests altogether should also apply to use of the Bayes Factor, so it appears distinct from the Bayesian-frequentist debate.

YUDI PAWITAN
DEPARTMENT OF
MEDICAL EPIDEMIOLOGY
AND BIostatISTICS
KAROLINSKA INSTITUTET

**Hur tänkte
Thomas
Alva Edison
när han
utvecklade
glödlampan?**

References

- Amrhein, V., Greenland, S. and McShane, B. (2019). Scientists rise up against statistical significance. *Nature* 567(7748): 305-307.
- Amrhein, V., Trafimow, D. and Greenland, S. (2019). Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *The American Statistician* 73: 262-270.
- Benjamin, D. et al. (2018). Redefine statistical significance. *Nature Human Behaviour* 2: 6-10.
- Efron, B. and Tibshirani, R. (1993). *An introduction to the bootstrap*. Boca Raton: Chapman and Hall.
- Fisher, R. A. (1926). Arrangement of field experiments. *Journal of Ministry of Agriculture XXXIII*: 503-13.
- Fraser, D.A.S. (2011). Is Bayes Posterior a quick and dirty confidence. *Statistical Science* 26: 299-316.
- Guilfoyle, J.R., Struthers, W., van Monsjou, E. and Shoikhetbrod, A. (2019). Sorry is the hardest word to say: The role of self-control in apologizing. *Basic and Applied Social Psychology* 41: 72-90.
- Kraemer, H.C. (2019). Is it time to ban the P-value? *JAMA Psychiatry*. Published online August 7, 2019.
- Mack, M.J. et al. 2019. Transcatheter aortic-valve replacement with a balloon-expandable valve in low-risk patients. *New England Journal of Medicine* 380: 1695-1705.
- Pawitan, Y. (2001). *In All Likelihood*. Oxford: Oxford University Press.
- Schweder, T. and Hjort, N.L. (2016). *Confidence, Likelihood and Probability*. Cambridge: Cambridge University Press.
- Ziliak, S. and McCloskey, D. (2008). *The Cult of Statistical Significance*. Ann Arbor: University of Michigan Press.

A discussion on the use of the p-value

Introduction to the p-value

The p-value is used in statistical significance testing to determine whether the null hypothesis of a research question should be accepted or rejected. From the ASA statement on statistical significance and p-values, “a p-value is the probability under a specified statistical model that a statistical summary of the data would be equal to or more extreme than its observed value”. The consensus used in most research fields is that a p-value of 0.05 or below can be considered ‘statistically significant’, which gives cause for the null hypothesis to be rejected. This cut off was initially proposed by Fisher in 1925. The appropriate use and interpretation of the p-value is often poorly considered in research, with particular issues arising when p-values are interpreted as the probability that a hypothesis is true given that a certain result has been observed. This is not the same as the probability that a result has occurred given that a certain hypothesis is true.

Advantages of the p-value

Although p-values have been used inappropriately in some instances, there are positives to such an approach. Hypothesis based estimates taken from p-values allows for generalizability to populations outside of the sample, which other approaches such as log likelihood testing are less accommodating of. The p-value considers the variability within a sample, since the p-value is derived using the standard error, and therefore provides an estimate of the degree to which the sample outcomes deviate from the ‘real’ outcome for wider populations. If the assumptions on distribution, independence of observations, and randomness of the data hold, it is possible to provide at least

some indication of the range of possible values for the ‘true’ outcome, and can therefore be a useful tool.

Disadvantages of the p-value

The p-value is prone to overly simplistic and often misleading interpretation. In particular, the fifth principle from the ASA statement is often misinterpreted or ignored. This principle states that “A p-value, or statistical significance, does not measure the size of an effect or the importance of a result” which is closely related to the Altman and Bland statistics notes in BMJ 1995 “Absence of evidence is not evidence of absence”¹. The standard significance level of 0.05 is a statistical convention used to be able to draw a decision line for inferring results coming from subpopulations to more general populations. Since this value is arbitrary, and has simply become convention, the significance level should be cautiously used to establish presence or absence of effects. Non-significant results can still hide important signals which would need more evidence to be generalizable. Certain signals should be taken as a starting point for new investigations of the same research question, for example using data coming from other settings. To build robust evidence significant results from a single study is not the absolute key, and support from validating studies is needed. In both randomized and observational studies, non-significant results are given a negative interpretation and are considered to show the absence of the investigated effect. Such results, however, should be interpreted as the absence of evidence for the effect

in the specific study, or alternatively in the specific data sample. In a follow-up editorial from Alderson “Absence of evidence is not evidence

»The p-value is prone to overly simplistic and often misleading interpretation.»

of absence - We need to report uncertain results and do it clearly”, in BMJ 2004, the author states that “We need to create a culture that is comfortable with estimating and discussing uncertainty”, which is the real key when trying to understand the truth.

Problems with the use of the p-value in the research environment

Whether or not scientific journals decide to accept or reject an academic paper may have less to do with the quality of the paper, and whether or not the method is robust, and more to do with whether the results are considered to be ‘statistically significant’ according to the conventional 0.05 cut off. Such an effect is known as publication bias, and this process can have a huge impact on available published evidence. In particular, the effect on meta-analyses, often used to build an overarching picture of the existing literature, can be drastic, and is particularly concerning given that meta-analyses are often considered by for example media sources to be the ‘gold-standard’ when answering an overarching research question². The inherent bias by academic journals toward results shown to be statistically significant leads on to the detrimental practice of p-mining, and to the ways that p-values can be potentially misused for reasons which do not comply with good scientific practice.

»»»

»»» A DISCUSSION ON THE USE OF THE P-VALUE

p-mining and conflict of interest

p-mining and reporting only of significant results can represent a large conflict of interest. p-values have been known to make-or-break careers with significant findings being preferentially published. Lack of a “positive” or otherwise known as significant result can drastically reduce research funding opportunities, and impact on career opportunities, as CVs are evaluated and often ranked based on how many high-impact journals appear in the publication list. But this affects more than individual researchers and can have far-reaching consequences: a significant p-value is in many cases the green-light for clinical and medical guidelines³. The future of a pharmaceutical medication can depend on the p-values of the clinical trial, and when a small number, potentially only one published study, is considered to contain the truth, this can have large and long lasting ramifications. The ramifications are widespread not only in health care, but in other fields as well. Over 20 years ago, a significant p-value was published linking the measles, mumps, rubella vaccination and autism: a claim that was made from researchers with extreme conflicts of interest which were not declared at the point of submission. Subsequent investigations unearthed that the study was funded by lawyers suing vaccine manufacturers⁴. This serves to highlight the dominance of the p-value and the importance that is placed on obtaining a significant finding, but it should not be the only measure of the final scientific conclusion or implication. It should be taken together with the magnitude of the association, underlying mechanisms, replication of results, and previous research findings. A small p-value does not necessarily mean “relevant” given that an intrinsic property of a smaller p-value is that it can be obtained in nearly all cases with a large sample size⁵. Larger sample size does not necessarily replace quality of a study and could be instead more reflective of the size of the funding supporting the research.

Alternatives to the p-value

An increasingly popular alternative to formal hypothesis and significance testing using p-values is the concept of inference based on probabilities of outcomes occurring, or of belonging to a certain group. In particular, Bayesian approaches which use pre-existing information within data to determine the likelihood of future possible outcomes are sometimes used in place of statistical significance testing. Rather than testing a research question and accepting or rejecting

the null hypothesis, the fit of the model can be assessed instead, and information provided on whether the addition of more groups or variables to the model is improving the extent to which the model explains the data. Values which can be used in these kinds of probabilistic methods include Akaike’s information criteria (AIC)⁶, formally presented by Hirotugu Akaike in 1974, and the Bayesian Information Criteria (BIC) developed by Gideon E. Schwarz in 1978⁷. These values are data driven and penalize complexity and number of variables, and therefore aid in model construction as well as interpretation. The field of for example trajectory analysis uses a probabilistic approach to estimate expected outcomes over time for different groups of individuals, which are defined by the model rather than by a hypothesis. An advantage to such methods is the ability to assess uncertainty as well as predicting outcomes, and so takes what could be described as a more holistic approach to data analysis relative to the use of p-values. A disadvantage is the reliance on the data itself, which can reduce generalizability, and interpretation can be more complicated when using these approaches.

Conclusions

Often, p-values are used as an easily interpretable indication of what results mean for the benefit of individuals who may have minimal or no statistical training. Use of the p-value as a simplification of whether an association exists is problematic for all the above mentioned reasons, in particular the intrinsic quality of the p-value reducing as the sample size increases is often overlooked. The allocation of research funding towards hot topic studies in which a new association is believed to have been found (for example the Wakefield study) is not only detrimental to scientific understanding, but it can be dangerous to human health. To address this, journals need to move away from the temptation to publish primarily ‘significant’ results which can be neatly packaged as answering a certain question conclusively. When p-value results are interpreted with care, and considering all the assumptions involved, they can provide a useful tool when addressing scientific questions. Discussion of the uncertainty involved in a study, and a move away from headline inducing studies with a focus on novelty and statistical significance is an important first step.

SARAH BURKILL, LAURA PAZZAGLI,
AND KELSI A. SMITH
CENTRE FOR PHARMACOEPIDEMIOLOGY
KAROLINSKA INSTITUTET

References:

1. Altman, D.G. and Bland, J.M. (1995). Statistics notes: Absence of evidence is not evidence of absence. *British Medical Journal* 1995, 311-485.
2. Mlinarić, A., Horvat, M., and Šupak Smolčić, V. (2017). Dealing with the positive publication bias: Why you should really publish your negative results. *Biochemia Medica*, 27(3), 030201 (DOI:10.11613/BM.2017.030201)
3. Head, M.L., Holman, L., Lanfear, R. et al. (2015). The extent and consequences of P-hacking in science. *PLoS Biology*, 2015, 13(3):e1002106 (DOI: 10.1371/journal.pbio.1002106)
4. Canadian Medical Association Journal. (2010). Lancet retracts 12-year old article linking autism to MMR vaccines. *Canadian Medical Association Journal News*, 2010, 182(4):E199-E200. (DOI:10.1503/cmaj.109-3179)
5. Bang, H. and Kim, J. (2016). Three common misuses of P values. *Dental Hypotheses* 7(3): 73-80. (DOI: 10.4103/2155-8213.190481)
6. Akaike, H. A. (1998). New Look at the Statistical Model Identification. In: *A New Look at the Statistical Model Identification*, Parzen, E., Tanabe, K., and Kitagawa, G. (eds). New York: Springer.
7. Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics* 1978(6): 461-464.

Kan Statistik-främjandet bidra till bättre rapportering av forskningsresultat?

Efter ett nyligen genomfört gympass hörde jag ledaren diskutera kostråd med några av deltagarna. Hon konstaterade ”ena året är det ena nyttigt, nästa år något annat”.

Denna inställning är nog inte helt ovanlig, och inte heller helt obegriplig. Ibland kan kritiskt tänkande vara nyttigt, men vetenskapsskepticism kan leda till bl.a. felaktiga politiska beslut eller försämrad folkhälsa (exempelvis genom obefogat vaccinationsmotstånd).

Nyhetsmedias rapportering av vetenskapliga upptäckter baseras ofta på pressreleaser från olika lärosäten. Dessa baseras i sin tur på studier där forskare försökt ta reda på hur det förhåller sig i verkligheten. I vart och ett av dessa steg kan det naturligtvis uppkomma missförstånd.

Vad gäller det första steget (d.v.s. forskarens beskrivning av en upplevd verklighet) så kan man till exempel hänvisa till Ioannidis artikel med den alarmistiska titeln ”Why most published research findings are false” (Ioannidis

2005), vilken i skrivande stund lästs närmare tre miljoner gånger. En viktig aspekt är här felaktigt användande av det s.k. p-värdet. Detta kan tolkas som sannolikheten att uppnå ett minst lika extremt utfall som det erhållna, givet att en s.k. nollhypotes är uppfylld.

Det har pågått en hel del diskussioner om nyttjandet av p-värdet. En fullständig redogörelse för dessa diskussioner ligger långt ovanför ambitionsnivån för detta inlägg, men några aspekter kan vara värda att nämna:

- Vid tillräckligt stort antal observationer kommer nollhypotesen nästan alltid att förkastas. Man måste skilja mellan formell signifikans och praktiskt betydelsefull signifikans.
- Om en nollhypotes inte förkastas, så betyder det inte att den kan betraktas som någorlunda sann. Det finns många andra tänkbara nollhypoteser, som inte heller skulle ha förkastats.
- Användning av 5% signifikansnivå sker ofta slentrianmässigt.
- Om forskare bara redovisar experiment där



Ett pressmeddelande om hur massiv
nyodling av skog kan rädda klimatet
startade omfattande diskussioner.

»»» KAN STATISKFRÄMJANDET BIDRA TILL BÄTTRE RAPPORTERING AV FORSKNINGRESULTAT?

$p < 0,05$ (eller om tidskrifter bara publicerar sådana manuskript), så ger det en felaktig bild av forskningsläget.

Debatten ledde 2016 till ett uttalande från ASA (American Statistical Association) som ansågs vara unikt i föreningens historia (Wasserstein 2016). Dokumentet har lästs och citerats flitigt och mig veterligen har det inte direkt ifrågasatts. Däremot har synpunkten framförts att dokumentet innehåller en del varningar, men inte tillräckligt med anvisningar om hur man borde göra istället. Därför anordnade ASA ett symposium som ledde till ett antal artiklar vilka publicerades 2019 (Wasserstein 2019). Enligt Wasserstein (personlig kommunikation) så pågår en del ytterligare uppföljningsaktiviteter; bl.a. är diskussioner med redaktionsrådet för *The Journal of the American Medical Association* planerade. Vidare hävdar han att det är mindre troligt att man kan hitta en universell metod att omvandla data till insikter; med andra ord så är vetenskapligt arbete komplicerat.

Aktiviteter pågår också på andra håll, bl.a. under ledning av professor Claudia Becker vid universitetet i Halle, vilka är inriktade mot ekonometri (Hirschauer et al. 2018).

Även om forskaren hanterar data på ett statistiskt korrekt sätt och drar korrekta slutsatser så finns det en risk att den som skriver pressreleasen "bättrar på slutsatserna". I en brittisk studie, publicerad 2014 (Summer et al. 2014), framhölls bl.a. att 40% av pressreleaserna innehöll mer direkta och explicita råd än de publicerade artiklarna. En liknande nederländsk studie, (Schat et al. 2018) kom fram till att 40 av 142 studerade pressreleaser överdrev slutsatserna. Även

om dessa siffror bygger på ett begränsat material kan man lugnt konstatera att denna felkälla inte är försumbar.

Det kan tänkas att nyhetsmedia "bättrar på" slutsatserna ytterligare, men det tycks mindre vanligt, i varje fall enligt den ovan nämnda nederländska studien. I vissa fall tycks journalisterna till och med gräva djupare än bara läsa pressreleasen.

Ett aktuellt exempel är en pressrelease från ETH (Tekniska Universitetet i Zürich, ett av världens högst rankade lärosäten) rörande en studie om möjligheterna att rädda klimatet genom massiv nyodling av skog. ETH:s pressrelease överdrev tydligen slutsatserna från studien vilket ledde till omfattande diskussioner; se bl.a. inlägg från professor Stefan Rahmstorf (se Twiterkonto). Att denna pressrelease utsattes för en så pass omfattande granskning berodde säkerligen på att den avhandlade ett mycket aktuellt ämne.

Min slutsats är att man bör vara mer tydlig med osäkerheter, även om det strider mot den akademiska traditionen. Då statistik mycket handlar om att hantera osäkerheter torde statistiker kunna spela en större roll vad gäller att tolka resultat. Det förutsätter närmare samarbete mellan statistiker och övriga discipliner. Svenska statistikfrämjandet skulle därmed kunna göra en viktig insats vad gäller att genom propagerandet för sund användning av statistik begränsa spridandet av vetenskapliga missförstånd. Ett intressant initiativ av detta slag är det tyska "Unstatistik des Monats" (se länk nedan) bakom vilken ligger en psykolog, en ekonom och en statistiker. Där redovisas regelbundet felaktigt användning av statistik.

HANS ALBERG

Referenser

Hirschauer, N., Grüner, S., Mußhoff, O., Becker, C. (2018). Pitfalls of significance testing and p-value variability: An econometrics perspective. *Statistics Surveys*, 12(2018): 136-172.

Ioannidis, J.P.A. (2005). Why Most Published Research Findings Are False. *PLoS Med* 2(8): e124. <https://doi.org/10.1371/journal.pmed.0020124>

Ioannidis J.P.A. (2019). What Have We (Not) Learnt from Millions of Scientific Papers with P Values?, *The American Statistician*, 73:sup1, 20-25, DOI: 10.1080/00031305.2018.1447512

Schat, J. et al. (2018). Overdreven gezondheidsnieuws. Relatie tussen overdrijving in academische persberichten en in nieuwsmedia [Exaggerated health news. The relation between exaggeration in academic press releases and in news media], *Nederlands Tijdschrift voor Geneeskunde*, 162 (1), 13-17.

Sumner, P. et al (2014). The association between exaggeration in health related science news and academic press releases: retrospective observational study. *British Medical Journal*. 2014 Dec 9; 349: g7015.

Wasserstein, R.L., Nicole L.A. (2016). The ASA Statement on p-Values: Context, Process, and Purpose, *The American Statistician*, 70:2, 129-133, DOI: 10.1080/00031305.2016.1154108

Wasserstein, R.L., Schirm A.L., Lazar L.A. (2019). Moving to a World Beyond "p < 0.05", *The American Statistician*, 73:sup1, 1-19, DOI: 10.1080/00031305.2019.1583913

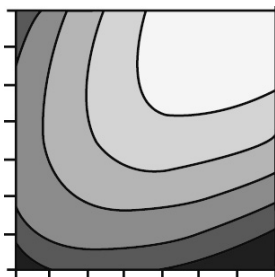
Informella referenser

ETH pressrelease "Wie Bäume das Klima retten konnten. (04.07.2019)".

Rahmstorf, S, Twitter, <https://twitter.com/rahmstorf/status/1151174593043935232>

Unstatistik des Monats, <http://www.rwi-essen.de/unstatistik/>

Wasserstein, R. personlig kommunikation.



ORDFÖRANDEN HAR ORDET

Modellen behöver anpassas till den digitala världen

Statistisk metodik är basen för allt förbättringsarbete, allt kvalitetsarbete och all verksamhetsutveckling inom industri, företagande och förvaltning. Grunden för detta tänkande lades av Bergman och Klevsjö, som skapade den svenska kvalitetsutvecklingsmodellen för nästan 12 år sedan. Men modellen har – med all respekt – åldrats med tiden.

Vid modellens introduktion hade man fokus på statistisk metodik, men efter ett tag flyttades fokus successivt till administrativa ISO-standarder för revision av verksamheten. Lukrativa kurser i revision ersatte kurserna i grundläggande statistik. Därför kan frågan ställas om den nuvarande modellen fortfarande är aktuell eller om den behöver förnyas.

Anledningen till frågan är nya utmaningar som t.ex.

- Digitalisering och automatisering
- IoT (Internet of Things), konceptet Industri 4.0
- Maskininlärning och utveckling kring AI (artificiell intelligens)

Bergman-Klevsjö-modellen för offensiv kvalitetsutveckling har varit bra, men vi måste inse att dess ”bäst före datum” antagligen har passerats på grund av dessa nya utmaningar.

En av utmaningarna är att följa med utvecklingen inom t.ex. IoT och dess konsekvenser. Det är troligt att IoT och AI med maskininlärning kommer att dominera den industriella utvecklingen framöver. Basen för den utvecklingen är tillämpad matematisk statistik, vilket ställer krav på kunskaper i statistik överhuvudtaget. Inom många verksamheter är dessa kunskaper en bristvara.

Vi behöver inse att Bergman-Klevsjö-modellen behöver uppdateras och anpassas till verkligheten i den ”nya” digitala världen. Det är t.ex. bra med ledningens engagemang, men det är inte tillräckligt för att lyckas med modern, autonom produktion. Det är en utmaning för ledningen att säkra den egna kompetensen inom resp. ansvarsområde. Inom ledarskapet kommer det att krävas mer kunskaper och kompetens inom flera discipliner än tidigare. T.ex. IoT kräver kunskaper i att kontinuerlig bevaka och följa statistisk analys av hela systemets status. Maskininlärning kräver tillämpning och analys av avancerade statistiska modeller. AI kräver kunskaper i matematisk statistik av bl.a. Bayesiansk typ, samt simulering.

Följaktligen blir nuvarande kvalitetsfunktioner och kontroller obehövlige när produkternas kvalitet kommer att avgöras automatiskt inom autonoma produktionssystem, styrda av AI. Insikt om begränsningen av föråldrade kvalitetssystem och modeller kan vara det första steget till anpassning av kvalitetsutveckling till verkligheten. Är man inom industrin, läroanstalter och myndigheter beredd att möta nämnda utmaningar?



FOTO: LINA LANDE

GÖRAN LANDE



Nya i styrelsen

■ **MARCUS BERG** (ordf). Jag är senior data scientist på Annalect, en del av Omnicom Media Group. Jag har bott i Tokyo i ett par år, där jag arbetade som data scientist. Innan dess föreläste jag om undersökningsmetodik och regressionsanalys på A-nivå på Stockholms universitet. Jag har också suttit som Surveyföreningens webbsidesansvarig under perioden 2014-16, det är kul att vara tillbaka!



■ **RIKARD GARD**. Jag är metodstatistiker på SCB och har varit det i tre år. Tidigare arbetade jag med befolkningsprognoser och befolkningsstatistik på SCB. Mina arbetsuppgifter idag är väldigt varierande. Det är allt ifrån design och urval av undersökningar och nya prognosmetoder till att hitta metoder att kvantifiera felen i våra register. Väldigt roligt jobb!



»... i kölvattnet av politiker som skriker fake news ...»

Så har det hänt igen: Surveyföreningen har fått en ny ordförande. Och den här gången så är det en inte så blyg 32-åring vars största merit är att han arbetat som data scientist i Tokyo under ett par år. Data scientist, ni vet dom analytiker som förlitar sig på machine learning mer än på statistiska metoder, och som snackar om AI till höger och vänster. Hur tänkte egentligen valberedningen när man nominerade mig till ordförande?

Jag tror att de tänkte tillbaka på tiden när jag satt med i styrelsen som webbsidesansvarig mellan 2014 och 2016, innan jag flyttade till Tokyo och bodde och arbetade där i tre år. Jag tror att de tänkte tillbaka på det halvdagsseminarium jag var med och anordnade om Big Data hösten 2015, och på rapporten om Big Data som jag var med och skrev åt AAPOR 2015. Jag har en klassisk utbildning i statistik, men har alltid låtit mig dras åt det nya och buzzword-iga: "Data är den nya oljan", "big data", "machine learning", "AI", allt det där, ni vet...
...och jag tror att jag, och min resa, kan fungera som en allegori för hela branschen och för den resa vi har varit med om sedan 2013, the International Year of Statistics: Tanken var att 2013 skulle vara

startskottet för att statistisk kunskap och vetenskap skulle sprida sig i samhället, men istället så fastnade vi statistiker i en duell med black box-algoritmer vars enda krav är mängden data och processorkraften på datorerna i molnet. Vi fann oss i kölvattnet av politiker som skriker *fake news* om alla statistiskt bevisade samband som inte passar deras agenda. Och jag lät mig slitas med: Om jag står inför en kund och har att välja mellan två modeller, en statistisk modell och en machine learning-modell, där den senare är en black box, så att kunden ifråga inte kan syna den allt för hårt i sömmarna, vad spelar det då för roll att jag inte kan visa några osäkerhetsmått för skattningarna? *Den träffar ju rätt på både träningsdata och testdata.*

Men sen så kommer whiplashen, eller baksmällan, eller vad vi nu ska kalla den för: Dagen då du inser att hela din verktygslåda av machine learning-algoritmer bara kan svara på en liten del av alla frågeställningar som man stöter på ute i den stora världen. Machine learning kan hjälpa oss att predicera och klassificera, men vad gör vi om vi vill ta reda på någons åsikt eller ta fram någon form av populationsskattning? Ja, då måste vi söka oss tillbaka till undersökningsmetodiken.

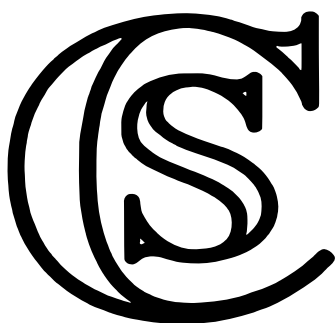
Och det är därför jag är här nu, om du frågar mig. Jag ser ett större behov i samhället än någonsin av statistisk bildning i allmänhet och surveymetodik i synnerhet, och jag tror att Surveyföreningen har en stor roll att spela.

Men det är inte bara jag som är ny i styrelsen, även Rikard Gard från SCB är ny. Här är hela styrelsen som valdes på årsmötet 2019:

- Marcus Berg, Annalect (nyval, ordförande)
- Ann-Marie Flygare, Örebro Universitet (omval, kassör)
- Gösta Forsman, Statisticon (omval, vice ordförande)
- Rikard Gard, SCB (nyval)
- Johan Martinsson, SOM-institutet Göteborgs Universitet (omval)
- Mats Nyfjäll, Statisticon (omval)
- Karin Schneller, TNS sifo (omval)
- Yingfu Xie, SCB (omval)
- Åke Wissing, Åke Wissing & Co (omval)

Ella Adem och Karin Nelsson avgick, och vi tackar dem varmt för deras tid och engagemang! Vi i nya styrelsen tackar för förtroendet att driva föreningens verksamhet 2019/20, det ska bli en spännande resa tillsammans!

MARCUS BERG



Cramérsällskapets höstmöte 2019

Cramérsällskapet arrangerade, efter en tids vilande, ett höstmöte om utbildningsfrågor 30/9 – 1/10 2019. Mötet hade temat ”Ny statistik – ny undervisning?” och ägde rum från lunch till lunch i SLU:s vackra miljö utanför Uppsala i ljuvlig höstskrud. Mötet samlade 38 deltagare med representanter från 12 svenska lärosäten och 14 institutioner. Endast Lund saknades.

Under mötet redovisade alla institutioner vilka förändringar som skett och sker inom ut-

bildningen med tanke på den nya beräknings-intensiva statistiken (som ibland på svengelska kallas ”data science” eller ”machine learning”). Vi fick även höra lite längre presentationer från två nordiska grannar: **OLE WINTHER** från DTU berättade om hur machine learning förts in i merparten civilingenjörsutbildningar i Danmark, och **GEIR STORVIK** redogjorde för det nya kursutbudet inom både kandidat- och masterutbildningen i matematik och statistik vid Oslo universitet. Mötet avslutades med en

paneldebatt i vilken det bl.a. framkom synpunkter om att viss traditionell statistik kan minskas ned, men att små datamängder kommer att fortsätta finnas, varför traditionella statistiska metoder kommer att ha fortsatt användning. Många deltagare påtalade också det positiva i att utbyta erfarenheter med varandra, och såg fram mot liknande aktiviteter i Cramérsällskapets regi framgent.

TOM BRITTON

NOMINERA TILL 2019 ÅRS CRAMÉRPRIS!

Cramérpriset utdelas årligen för bästa doktorsavhandling till en person som disputerat i statistik/matematisk statistik under kalenderåret. Den huvudsakliga bedömningsgrunden är avhandlingens kvalitet, men pedagogisk skicklighet (t ex manifesterad i avhandlingens kapa) och andra personliga egenskaper kan också väga in. Tack vare en donation har pris-summan höjts till tiotusen kronor.

Vi uppmanar till nomineringar till detta pris. Motiveringstexten bör innehålla information om avhandlingens innehåll, graden av självständighet, kvalitet och kreativitet. Andra intressanta omständigheter kan t.ex. vara vilka publikationer avhandlingen lett till, om kandidaten blivit klar på eller före utsatt tid.

En länk till avhandlingen, eller pdf-version, skall även skickas in.

Cramérsällskapets styrelse beslutar om pristagare. Pristagaren förväntas presentera sin avhandling vid Cramérsällskapets årsmöte den 20 mars 2020.

Sista datum för nominering är den **15 januari 2020**. Skicka nominering till Cramérsällskapets sekreterare, Jesper Rydén: jesper.ryden@slu.se.

Cramérsällskapets nya styrelse



Tom Britton.



David Bolin.



Yuli Liang.

■ **TOM BRITTON** (ordf), professor i matematisk statistik vid Matematiska institutionen, Stockholms universitet. Tom forskar inom stokastisk och statistisk modellering av problem inom biologi och medicin, speciellt smittsamma sjukdomar, nätverk och fylogeni. Undervisar i kurser på grundläggande, avancerad och doktorandnivå och har bl.a. skrivit två läroböcker. Han var under 7,5 år prefekt vid institutionen och är nu sektionsdekanus i Matematik-Fysik.

■ **DAVID BOLIN**, docent i matematisk statistik vid Matematiska Vetenskaper, Göteborgs Universitet. Davids forskning handlar mestadels om spatial statistik och statistiska modeller baserade på stokastiska partiella differentialekvationer. Undervisar i statistik på alla nivåer vid Göteborgs Universitet och Chalmers och är för närvarande associate editor för *Scandinavian Journal of Statistics*.

■ **YULI LIANG**, biträdande lektor i statistik vid Handelshögskolan, Örebro universitet. Yulis forskning handlar om att utveckla metoder för att kunna dra rättvisande slutsatser från olika mönstrade kovariansmatriser till exempel Toeplitz, cirkulär Toeplitz och blockkovariansstrukturer. Dessutom forskar Yuli inom officiell statistik, speciellt utveckling av nya metoder för att minska bortfällskeheter i olika undersökningar. Undervisar grundläggande statistik. Yuli var tidigare metodstatistiker på avdelningen för process- och metodutveckling vid SCB.

■ **ANDERS LUNDQUIST** (kassör), universitetslektor i statistik vid Handelshögskolan, Umeå Universitet. Anders forskar om metoder för analys av longitudinella data med icke-slumpmässigt bortfall. Han har en del av sin tjänst på Umeå Universitets centrum för funktionell hjärnabbildning (UFBI). Undervisar på grundläggande och avancerad nivå, med såväl kurser som uppsatshandledning.

■ **FREDRIK OLSSON**, universitetslektor och ETP (Excellent Teaching Practitioner) i matematisk statistik vid Matematikcentrum, Lunds universitet. Fredrik forskar inom operationsanalys, speciellt lagerstyrning och produktionsstyrning. Har undervisat i kurser på både avancerad nivå och grundnivå, exempelvis kurser inom: operationsanalys, köteori, försöksplanering, simulering, matematisk analys, linjär algebra och grundkurser i matematisk statistik.

■ **JESPER RYDÉN** (sekr), universitetslektor i biometri vid Sveriges lantbruksuniversitet, SLU (docent i matematisk statistik). Forskningsintressen berör statistisk modellering i geovetenskaperna, t.ex. skattningsmetodik för återkomstvärden i hydrologi och meteorologi. Har undervisat i kurser på samtliga nivåer och skrivit två läroböcker. En viktig arbetsuppgift i tjänsten är statistisk konsultation för forskare vid SLU.



Anders Lundquist.



Fredrik Olsson.



Jesper Rydén.

Statistikutbildning och simulering

Bakgrund

Det är fortfarande så att inom industrin – och troligen även inom andra näringar – är bristen på tid det största problemet, och att det därför är svårt att attrahera ledningsgrupper och chefer till deltagandet i några längre statistikkurser. Tidigare ansågs det viktigt att utbilda företagsledare och chefer i grundläggande statistik. Många tyckte då att det var en positiv satsning, men alla klagar nu på den ständiga tidsbristen för kompetensutveckling, och satsningarna har blivit mindre frekventa.

Nu – med facit i hand – kan vi konstatera att behovet och intresset hos företagsledarna för statistikutbildning inte har ökat nämnvärt, fastän verkligheten talar för att det borde vara tvärtom. Särskilt nu när digitaliseringsfrågor diskuteras i TV och press. Det som kanske märks mest är den ökande efterfrågan på kompetenta individer som kan analysera data och gräva fram det viktiga som behövs för ett bra och riskfritt beslutsfattande.

Dagens ledningsfunktioner behöver snabba besked, on line-analyser och framtagning av trender för att kunna följa med och fatta snabba beslut. Det är lönsamheten och snabbheten som promoveras i affärsvärlden och därför efterfrågas snabba beslut med stöd av expertkunskaper. Det är viktigare att ha omedelbar tillgång till kompetens inom statistisk analys av data än att satsa på kursdagar avseende grundkunskaper.

Jag vill här argumentera för att det bör finnas tid för visualisering och pedagogisk användning av det som vanligtvis kallas för statistisk simulering.

Behovet av simulering.

Simulering av en process kräver att man har en matematisk-sannolikhetsteoretisk modell, som kan antas ge en någorlunda hygglig beskrivning av processen. Själva simuleringen innebär att man låter datorn generera ett antal utfall av processen. Genom att generera en stor mängd utfall av samma process kan man få en bild av variabiliteten hos utfallen. Speciellt kan man få en uppfattning om risken att få vissa icke-önskade utfall. Genom att göra flera simuleringssstudier av samma process, med olika värden på modellens

parametrar i de olika studierna, kan man få en bild av hur processutfallet påverkas av parametrarnas värden. Om modellen är någorlunda realistisk, kan man då bilda sig en uppfattning om hur processen i verkligheten påverkas av ändrade rutiner.

Fördelen med en simulering är snabbheten i att visualisera resultatet utan att besvära mottagaren med tidskrävande information om teoretiska härledningar och sammanhang. Man kan direkt se hur utfallen blir. Detta under förutsättning att simuleringen görs av en kompetent och kunig yrkesperson. Kanske en s.k. certifierad statistiker eller modernare statistikingenjör.

Ett exempel på simuleringens användbarhet är ett scenario där företaget behöver veta om det går att klara ett av projektmålen, som t.ex. leveransdatum. Det handlar då om att styra projektresurserna så att eventuellt uppkommande problem kan lösas och projektet avslutas vid utlovad tid.

Viktigt är att inse att det finns ett behov av kompetenta, yrkesverksamma statistiker och dataanalytiker, som kan anlitas som rådgivare. I en situation där bristen på tid är markant, är behovet av statistikkompetens mycket stort. Marknadskrafter och näringsliv bör utnyttja de statistikexperter som finns tillgängliga och be dem om hjälp med att simulera olika processer med hjälp av statistiska metoder som till exempel att göra Monte Carlo-simuleringar, göra riskanalyser, och uppskatta sannolikheter för att lyckas med olika åtgärder. Idag insamlas data i stora mängder. Allt finns tillgängligt i olika databaser. Det som behövs är tillförlitliga analyser av dessa stora datamängder, och framför allt visualisering av nyttan med sådana analyser.

Problem med simulering

Förutom problemet med att utbilda tillräckligt många industristatistiker i moderna metoder för simulering, så finns det också ett par problem av mer icke-teknisk natur.

Man måste skilja mellan modell och verklighet. För att en simuleringssstudie ska vara till

nytta krävs att den teoretiska modellen verkligen ger en hyfsat bra beskrivning av den verkliga situation som man är intresserad av. Att ta fram realistiska men ändå hanterbara modeller för att

beskriva olika industriella processer är en viktig uppgift. Det är inte säkert att gamla vanliga standardmodeller alltid passar bra.

Det finns också ett mer pedagogiskt problem. Den som ska presentera resultatet av en simuleringssstudie för höga beslutsfattare måste kunna göra

detta på ett lättbegripligt men ändå korrekt sätt. Det gäller att få beslutsfattaren att förstå i princip (men utan detaljer) hur studien är gjord och vad studien kommit fram till. Det gäller också att få beslutsfattaren att inse vilka osäkerheter studien är behäftad med. Studien bygger ju på vissa teoretiska modellantaganden, som kan ge en mer eller mindre bra bild av hur verkligheten är beskaffad.

Vad kan vi göra?

Simuleringsteknik och simuleringmetoder utvecklas snabbt och när allt bredare kretsar av användare. Med tanke på dagens datorkraft bör vi förorda visualisering av olika processer genom simulering, när processdata finns tillgängliga i någon form. Statistiker behöver lära sig nya simuleringmetoder och utbildas i takt med tillkomsten av nya tekniker. Sådana skapas idag och utvecklas av andra än traditionella statistiker, som t.ex. datavetare och analytiker inom områden som biokemi, data science, AI, radiologi, geovetenskap, mm.

En viktig uppgift, inte bara för industristatistiker, är att använda simulering för att visualisera den variation som finns i olika processer. Företag, myndigheter och kommuner bör sträva efter att alla ledningsfunktioner ska känna till variationens betydelse och beakta det vid beslutsfattande. Besluten blir då betydligt säkrare och kanske mindre kostsamma med tanke på den variation som ständigt påverkar alla våra processer.

GÖRAN LANDE



Alt finns tillgängligt i olika databaser. Det som behövs är tillförlitliga analyser av dessa stora datamängder, och framför allt visualisering av nyttan med sådana analyser.

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.

För att bli medlem, gå till <http://www.statistikframjandet.se> och läs mer i högerspalten under "Vill du bli medlem?". Har du frågor kontakta Mattias Strandberg 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.

