Adam Taube, professor, statistiker, institutionen för informationsvetenskap, Uppsala universitet *(adam.taube@dis.uu.se)*
Jörgen Malmquist, docent i internmedicin, Höllviken *(jorgen.malmquist@mailbox.swipnet.se)*

Räkna med vad du tror

Bayes – inte P-värdet – mäter tilltron

Det är tyvärr alldeles för vanligt att medicinska forskare inte är medvetna om att varje statistisk analys grundas på en teoretisk modell, och att de slutsatser som kan dras från en statistisk analys av data gäller under just denna modell. Valet av analysmodell styr därför alltid resultatet – anpassar man t ex en rät linje till tvådimensionella data blir resultatet just en rät linje och inget annat, oavsett om detta är en realistisk beskrivning eller ej. När det gäller situationer där signifikansanalys och konfidensintervall brukar användas kan valet av en bayesiansk modell ibland leda till fundamentalt annorlunda tolkningar än den traditionella s k frekventistiska analys som för närvarande helt dominerar den medicinska litteraturen.

Vid en skattning av t ex risken för ventrombos hos kvinnor som använder p-piller förutsätter den traditionella, frekventistiska modellen att en »sann« risk (med ett exakt numeriskt värde, traditionellt betecknat med en grekisk bokstav) existerar och att de data man insamlar i en viss undersökning ger ett estimat av denna bakomliggande fixa okända parameter. Den frekventistiska ansatsen, som förklaras i följande stycke, har under decennier välsignat medicinska artiklar med ett ymnigt regn av signifikansstjärnor och P-värden som inte sällan feltolkats och ibland till och med saknat relevans [1].

Tolkningen av ett konfidensintervall

Anta t ex att man beräknat ett 95 procents konfidensintervall för patienternas genomsnittliga blodtryckssänkning vid en klinisk prövning av ett nytt blodtryckssänkande läkemedel, och att detta intervall blev 6,2 ± 4,7 mm Hg. Den korrekta, traditionella tolkningen är då: »detta intervall hade (märk imperfektum) sannolikheten 95 procent att falla så att det täckte det sanna (men okända) värdet för blodtryckssänkningen«. Den modell som hela kalkylen baseras på förutsätter alltså att det existerar ett enda sant, fixt värde som man söker estimera så bra som möjligt. Resonemanget utgår från vad som kan förväntas i det långa loppet: om man skulle upprepa undersökningen många gånger och varje gång beräkna ett 95 procents konfidensintervall från de erhållna data skulle i genomsnitt 19 av 20 sådana intervall falla så att de innefattar »det sanna värdet« (därav termen »frekventistisk«). Man har därmed 95 procents konfidens (förtroende) för detta intervall.

Inte sällan förekommer i medicinska sammanhang en annan, felaktig formulering som i det här fallet skulle lyda: »med 95 procents säkerhet är det sant att den verkliga genomsnittliga blodtryckssänkningen ligger i intervallet 6,2 ± 4,7«. Eftersom det aktuella intervallet begränsas av fixa numeriska värden förutsätter denna formulering att det sanna värdet på något sätt skulle vara variabelt. Detta leder emellertid till en helt annan teoretisk modell och avspeglar väl närmast hur mången medicinare skulle önska att ett dylikt intervall kunde

SAMMANFATTAT

Bayesiansk analys av medicinska data beskrivs och exemplifieras i denna artikel, som är den andra av två.

Grundtanken i traditionell (frekventistisk) statistisk analys förklaras med utgångspunkt i begreppet konfidensintervall.

Bayesiansk analys i diagnostik exemplifieras: i valet mellan olika diagnoser sammanvägs tillståndens relativa förekomst i patientpopulationen med sensitiviteterna för ett diagnostiskt test.

Skillnaderna mellan traditionell och bayesiansk analys av resultaten av en klinisk prövning beskrivs, och fördelarna hos den senare förklaras.

En kort litteratursammanfattning visar den utbredda tillämpbarheten av bayesiansk analys inom medicinsk statistik.

Evidensbaserad medicin

En tidigare artikel om Bayes-analys har publicerats i Läkartidningen 24/2001.

Thomas Bayes (1702–1761).

Tabell I. Bayesiansk kalkyl vid diagnostik med hjälp av cancermarkören CA 125.

1 Sjukdomskategori	2 Andel bland ifrågavarande patienter (före-sannolikhet)	3 Sensitivitet hos CA 125-analys (likelihood)	4 Produkten av 2 och 3	5 Efter-sannolikhet
Ovarialcancer	0,40	0,85	0,34	0,73
Annan cancer	0,10	0,50	0,05	0,11
Benign tumör	0,50	0,15	0,075	0,16
Summa	1,00		0,465	1,00

tolkas. Denna tolkning stämmer, som skall beröras ytterligare, snarare överens med den bayesianska modellen än med den frekventistiska.

Sammanvägning av tillgänglig information

Vår föregående artikel [2] presenterade Bayes' sats i en situation med dikotomier – »sjuk/ej sjuk« och »positivt/negativt testutfall«. Här skall strukturen belysas med ett fiktivt exempel där flera sjukdomar är aktuella. Vi utgår från kvinnor som vid gynekologisk undersökning befunnits ha en tumörmisstänkt förändring intill uterus. Undersökningen kompletteras med analys av tumörantigenet CA 125 i serum. Förhöjd nivå av denna indikator antas ha följande sensitivitetsvärden: vid ovarialcancer omkring 85 procent och vid andra tänkbara cancerformer ca 50 procent. Vidare antas 15 procent av patienter med benign tumör ha förhöjd CA 125-nivå. Anta vidare att man vid kliniken har erfarenheten att bland patienter med palpationsfyndet ifråga 40 procent har ovarialcancer, 10 procent annan cancer och 50 procent benign tumör. Detta innebär bl a att man i avsaknad av ytterligare diagnostisk information anser sannolikheten vara 40 procent för att en patient med detta palpationsfynd har ovarialcancer.

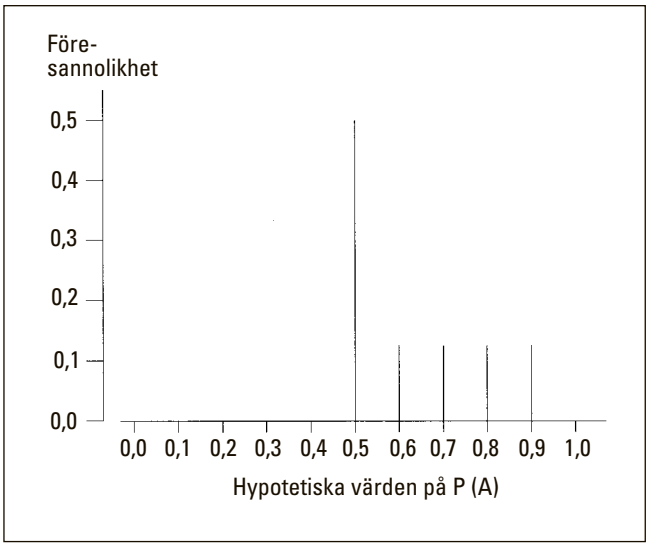
Bland dessa patienter blir andelen kvinnor med förhöjt CA 125-värde sammanlagt $0{,}85 \times 0{,}4 + 0{,}5 \times 0{,}1 + 0{,}15 \times 0{,}5 = 0{,}465$ och andelen som har ovarialcancer samt förhöjt CA 125-värde $0{,}85 \times 0{,}4 = 0{,}34$, se Tabell I. Sannolikheten för att en viss kvinna med förhöjt CA 125 har ovarialcancer blir givetvis andelen ovarialcancerfall bland alla dem som har förhöjt CA 125, dvs 0,34/0,465 = 0,73.

Utan att beakta annan diagnostisk information skulle man alltså på grundval av palpationsfynd och förhöjt CA 125 kunna hävda att sannolikheten är 73 procent att denna kvinna har ovarialcancer. Denna kalkyl grundas på uppfattningen om proportionerna (0,4 respektive 0,1 och 0,5) mellan de tre diagnoskategorierna. Dessa proportioner har vägts samman med respektive sensitiviteter. Varje sensitivitetstal anger troligheten (likelihood) att en person inom respektive sjukdomskategori skall komma med bland dem som har ett positivt CA 125-värde.

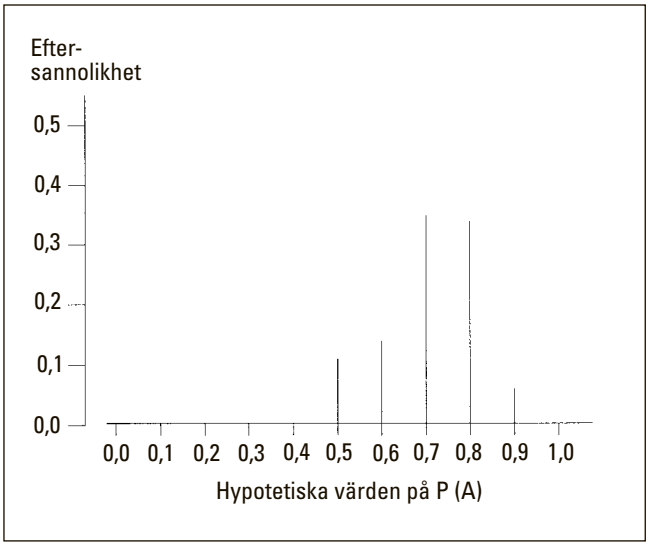
I detta exempel ändrades alltså före-sannolikheten (prior probability) 40 procent genom det förhöjda CA 125-värdet till efter-sannolikheten (posterior probability) 73 procent. Det är uppenbart att denna kalkyl är starkt beroende av de nämnda erfarenhetsmässiga proportionerna mellan olika patientkategorier; skulle man ändra dessa före-sannolikheter skulle slutresultatet bli ett annat. Det kan också vara fallet att sådana värden inte alls är kända och att berörda läkare endast har på känn ungefär hur stora de kan vara. Men även detta är kunskap.

Traditionell »frekventistisk« analys

Anta att man vill jämföra två migränpreparat genom en överkorsningsstudie där 20 migränpatienter får ange vilket preparat de föredrar utifrån bästa symtomlindrande effekt. Av de båda medlen A och B innehåller B ett standardpreparat, medan A innehåller samma preparat plus en tillsats i form av en ny substans. Den senares eventuella förmåga att förstärka effekten skall studeras. Det anses finnas goda skäl att anta att tillsatsen inte orsakar försämrad effekt, men det är ovisst om den har någon positiv verkan. Nollhypotesen enligt gängse analysmodell är att medlen är lika effektiva, vilket innebär att det kan förväntas att antalet patienter som föredrar A är detsamma som antalet som föredrar B. Detta innebär att sanno-



Figur 1a. *A priori-fördelning av sannolikheterna för att olika andelar av patienterna föredrar läkemedel A i en prövning av A mot B.*



Figur 1b. *A posteriori-fördelning av sannolikheterna för att olika andelar av patienterna föredrar läkemedel A. Dessa efter-sannolikheter är beräknade med ledning av prövningsresultatet, där 15 av 20 patienter föredrog A. Se Tabell II.*

Tabell II. Bayesiansk resultatkalkyl vid överkorsningsprövning av läkemedlen A och B, varvid 15 av 20 patienter föredrog A.

1 Hypotes (andel av patienter som föredrar A)	2 Före-sannolikhet	3 Likelihood (× 100 000)	4 Produkten av 2 och 3 (× 100 000)	5 Efter-sannolikhet
0,5	0,500	0,0954	0,0477	0,114
0,6	0,125	0,4815	0,0602	0,144
0,7	0,125	1,1537	0,1442	0,345
0,8	0,125	1,1258	0,1407	0,336
0,9	0,125	0,2059	0,0257	0,061
Summa	1,000		0,4185	1,000

likheten för att föredra A är P(A) = 0,5. Om A har större effekt än B skall man förvänta sig att P(A) > 0,5. Man beslutar att resultatet skall signifikansprövas på 5-procentsnivån med enkelsidigt test.

Resultatet blev att 15 patienter föredrog A och fem föredrog B. Detta ger ett s k P-värde som är 0,02 (enligt exakt binomialfördelning med n=20 och π=0,5). Det man nu vet är alltså: »sannolikheten för att få det resultat som verkligen erhölls (eller något ännu extremare) var så liten som 0,02 – om nollhypotesen verkligen varit sann«. Observera att detta inte innebär att sannolikheten för att nollhypotesen är sann är 0,02.

Konfidensintervallet (95 procent) för proportionen patienter som föredrar A blir 0,51 till 0,91. Ett 95 procents konfidensintervall, som ju innehåller alla tänkbara hypotetiska värden på P(A) som inte kan förkastas på grundval av aktuella data, täcker alltså här nästan hela skalan av värden större än 0,5. Detta innebär att man har 95 procents konfidens för det aktuella intervallet men ingen bestämd sannolikhet för att enskilda värden är sannna, även om den funna proportionen 15/20 = 0,75 är en mycket rimlig gissning som den mest troliga (maximum likelihood-estimat).

Analys enligt Bayes

Vi utgår från samma prövning som i föregående avsnitt. Anta att prövarna först tillfrågade en rad kolleger om vad de trodde om möjligheten att den nya tillsatsen skulle ha en effektförbättrande verkan. Kollegerna fick helt enkelt, på grundval av sin kunskap, erfarenhet och intuition, föreslå (gissa) ett numeriskt värde på P(A), för enkelhets skull angivet endast som hela tiondelar, dvs 0,1, 0,2 etc. Det visade sig att ungefär hälften inte trodde att den nya tillsatssubstansen skulle ha någon märkbar effekt, dvs de angav som troligt värde P(A) = 0,5 (nollhypotesen). De som inte trodde på nollhypotesen var alla inne på att tillsatsen kunde förväntas ha en positiv effekt, dvs att värdet på P(A) borde vara >0,5. Vidare precisering hade man svårt att enas om, men ingen trodde att tillsatsen skulle vara så effektiv att alla patienter skulle föredra medlet A. Man enades om att som alternativ till P(A) = 0,5 studera värdena 0,6, 0,7, 0,8 och 0,9 vilka av de berörda ansågs vara ungefär lika troliga. Den tilltro som en samlad expertis alltså tillmätte de olika alternativa värdena på P(A) kunde därmed sammanfattas i den »a priori-fördelning« som illustrerats i Figur 1a.

Hur förändras nu tilltron till tillsatssubstansens effekt genom resultatet från prövningen? Betrakta t ex alternativet P(A) = 0,6. Då är likelihood-värdet för att man skall få just den kombination av resultat som prövningen gav

$$(0,6)^{15} \times (0,4)^{5} = 0,4815 \times 10^{-5}$$

Likelihood-värdet anger alltså sannolikheten för att få det resultat som faktiskt erhållits, om just denna hypotes vore riktig (jfr sensitivitetens roll i exemplet ovan med diagnostisk användning av CA 125!). För varje värde på P(A) i a priori-fördelningen beräknas på detta sätt en likelihood, se Tabell II. Sedan sammanvägs de olika a priori-sannolikheterna med respektive likelihood. Efter justering så att totalsumman blir 1 erhålles den a posteriori-fördelning som anges i sista kolumnen i Tabell II och som illustreras i Figur 1b. Denna fördelning skildrar alltså efter-sannolikheterna, dvs den tilltro till var och en av de uppställda hypoteserna som man får fram genom en kombination av vad man trodde före prövningen (före-sannolikheterna) och det man fick fram vid själva prövningen.

A posteriori-fördelningen visar dels att det blir störst tilltro till värdena P(A) = 0,7 och P(A) = 0,8 med över 30 procent för vardera, dels att tilltron till nollhypotesen har ett värde på ungefär 11 procent. Detta skiljer sig påtagligt från det P-värde som erhölls vid den traditionella frekventistiska analysen och som mätte något annat, nämligen risken att felaktigt förkasta nollhypotesen. Skillnaden beror givetvis på att man i den bayesianska analysen har vägt in det faktum att hälften av experterna från början inte trodde att tillsatssubstansen skulle ha någon märkbar effekt.

Bayes-analys i litteraturen

Här följer några korta referat av publikationer om användning av bayesiansk analys inom medicinen.

Analys av resultat från kliniska prövningar. Ett exempel på bayesiansk värdering av prövningsdata är en ofta citerad artikel av Brophy och Joseph [3]. Den granskade rapporten från GUSTO-studien som jämförde alteplas (t-PA) med streptokinas vid hjärtinfarkt. Rapporten hade konkluderat att alteplas var överlägset och att nollhypotesen (ingen skillnad) kunde förkastas med P=0,006. Brophy och Joseph visar i siffror och grafik de efter-fördelningar som blev resultatet då GUSTO-data analyserades med användning av resultaten av två tidigare jämförande prövningar som före-fördelningar, vilka gavs tilltro med 0, 10, 50 eller 100 procent. De fann att det rådde betydande osäkerhet om existensen av en kliniskt relevant effektskillnad mellan preparaten.

Planering av kliniska prövningar. Vid planering av behandlingsprövningar i fas I–II är en bayesiansk ansats av värde. Man kan integrera den mycket preliminära kunskapen om ett nytt medels effekter med olika antaganden om storleken av önskvärda effekter och biverkningar, och därigenom få hållpunkter för sannolikt bästa utformning av studien, regler för monitorering och hur många personer som behöver studeras [4].

Epidemiologiska data. En artikel av Lilford och Braunholtz [5] tar sin utgångspunkt i den omdiskuterade eventuella skillnaden i risk för trombos mellan andra och tredje »generationen« av p-piller. Författarna beskriver hur Bayes-analys ger en

bättre grund än konventionell statistisk analys när man, med data från epidemiologiska undersökningar, skall bedöma sannolikheten för att en väsentlig riskdifferens finns. Artikeln ger också en introduktion till bayesianska grundbegrepp. Författarna betecknar en ökad användning av bayesianska metoder som en nödvändighet för välgrundade beslut inom »public policy«.

Medicinsk teknologivärdering. En monografi från en brittisk utvärderingsmyndighet [6] ger en grundlig beskrivning och exemplifiering av bayesiansk analys, i jämförelse med klassisk statistisk analys, inom detta område. Publikationen innehåller en beskrivande förteckning över analysprogram som kan hämtas från Internet.

Litteratur med metodologiska kommentarer. Bayes-analys berörs kort i flera böcker om medicinsk statistik, t ex det välkända verket av Armitage och Berry [7], som också hänvisar till litteratur som är helt inriktad på Bayes-metodik.

En redaktionell artikel [8] berör skillnaden mellan klassisk (frekventistisk) och bayesiansk analys, och framför skäl för ökad användning av Bayes-metodik vid bedömning av resultaten av kliniska prövningar.

Ett antal skribenter kombinerar en beskrivning av Bayesanalysens fördelar med en redogörelse för nackdelarna med fixering vid P-värden och åtföljande rigida uppdelning av forskningsresultat i »signifikanta« och »icke-signifikanta«. Det framhålls att detta binära tänkande bör ersättas av att alla data så långt möjligt tolkas inom ramen för övrigt tillgängligt vetande. Sterne och Davey Smith [9] har nyligen publicerat en artikel där schablonmässig signifikanstestning kontrasteras mot bayesiansk bedömning.

En mycket grundlig och intressant genomgång av tankegångarna bakom signifikanstestning, hypotesprövning och bayesiansk analys har gjorts av Goodman [10, 11]. Han ger argument för en markant kritisk inställning till icke-bayesiansk analys. Han konstaterar bland annat att ordinära metaanalyser visar tendens till bayesiansk anda genom att göra en integrerad bedömning av alla tillgängliga data, men förklarar varför de inte kan anses ge en tillräckligt djupgående analys. Goodmans artiklar sammanfattades och kommenterades positivt av redaktören för Annals of Internal Medicine [12].

Argument kring Bayes-analys

Bayesiansk analys ifrågasätts ibland på den grunden att det är diskutabelt att nya data modifieras av a priori-kunskaper, eftersom de senare kan vara mycket osäkra eller begränsade. Motargumentet är att det är irrationellt att bortse från befintlig kunskap eller uppfattning när man värderar nya resultat. Graden av osäkerhet hos a priori-kunskaperna kan man ta hänsyn till i den bayesianska kalkylen. Integrationen av nya data med före-sannolikheter kan utföras ett antal gånger med varierade förutsättningar. Man kan införa korrektionsfaktorer för påvisad eller förmodad bias i de tidigare eller de nya data, och man kan studera den effekt som uppkommer om vissa data modifieras (känslighetsanalys). Vidare ger Bayes-analys den värdefulla möjligheten att beräkna sannolikheten för att skillnaden mellan två behandlingsmetoders effekter överstiger ett visst värde. Man kan också beräkna s k prediktionsintervall (på engelska även kallade credible intervals). Dessa är besläktade med konfidensintervallen i klassisk analys. Ett prediktionsintervall har emellertid fördelen att ha den något annorlunda och mer lättfattliga innebörden att det sanna populationsvärdet med den beräknade sannolikheten befinner sig inom intervallet.

Kalkylarbetet kan vara mödosamt. De exempel vi använt är enkla: det rör sig om ett begränsat antal observationer och diskreta utfallsmöjligheter (modeller). Uträkningen av tillhörande likelihoodvärden är därför enkel. Bland annat när det gäller kontinuerliga variabler i stället för diskreta kan uträkningsarbetet bli mycket omfattande. Detta är en av orsakerna till att användningen av Bayes-analys hittills varit begränsad. De vanliga kommersiellt tillgängliga mjukvarorna för statistiska kalkyler inkluderar för närvarande inte bayesianska uträkningar, men detta kommer sannolikt att förändras inom kort. Vidare finns bayesianska kalkylprogram tillgängliga på Internet, se ovan. Ett annan hindrande faktor har utgjorts av att uträkningarna kan kräva stor datorkraft, men dagens datorer har i regel tillräckliga prestanda.

Referenser

1. Taube A. Det borde (inte) vara stjärnor... Läkartidningen 1985;82: 2422-4.
2. Taube A, Malmquist J. Räkna med vad du tror. Bayes' sats i diagnostiken. Läkartidningen 2001;98:2910-3.
3. Brophy JM, Joseph L. Placing trials in context using Bayesian analysis. GUSTO revisited by Reverend Bayes. JAMA 1995;273:871-5.
4. Thall PF, Simon RM, Estey EH. New statistical strategy for monitoring safety and efficacy in single-arm clinical trials. J Clin Oncol 1996;14:296-303.
5. Lilford RJ, Braunholtz D. The statistical basis of public policy: a paradigm shift is overdue. BMJ 1996;313:603-7.
6. Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR. Bayesian methods in health technology assessment: a review. Health Technol Assess 2000; 4: No 38. Även tillgänglig via Internet: http://www.-ncchta.org
7. Armitage P, Berry G. Statistical methods in medical research. 3rd ed. Oxford: Blackwell Science, 1994.
8. Freedman L. Bayesian statistical methods. A natural way to assess clinical evidence. BMJ 1996;313:569-70.
9. Sterne JAC, Davey Smith G. Sifting the evidence – what's wrong with significance tests? BMJ 2001;322:226-31.
10. Goodman SN. Toward evidence-based medical statistics. 1 The P value fallacy. Ann Int Med 1999;130:995-1004.
11. Goodman SN. Toward evidence-based medical statistics. 2 The Bayes factor. Ann Int Med 1999;130:1005-13.
12. Davidoff F. Standing statistics right side up. Ann Int Med 1999;130: 1019-21.

SUMMARY

Count on your beliefs
Bayes – not the p value – measures credence

Adam Taube, Jörgen Malmquist
Läkartidningen 2001; 98: 3208-11

This article (the second of two) describes traditional (frequentist) statistical analysis in the context of the confidence interval.
Bayesian analysis is described in two settings. In the choice between diagnostic alternatives, the bayesian approach offers useful integration of new information with previous knowledge.
With regard to the evaluation of clinical trial data, this article exemplifies bayesian analysis as contrasted with traditional analysis, and advantages of the former are cited.
A brief literature review exposes the wide applicability of bayesian analysis in medical statistics.

Correspondence: Adam Taube, Dept of Information Science, Uppsala University, Box 513, SE-751 20 Uppsala, Sweden. (adam.taube@dis.uu.se)