Jonas E Ludvigsson, med dr, barnkliniken, Universitetssjukhuset i Örebro, samt enheten för klinisk epidemiologi, Karolinska Universitetssjukhuset Solna *(jonasludvigsson@yahoo.com)*

Kort om stickprovsstorlek

Verkligheten begränsar ofta möjligheterna att bedriva forskning. De där 80 patienterna man hade tänkt inkludera visar sig vara omöjliga att rekrytera, kostnadsberäkningen för analyserna är alltför optimistisk och man tvingas begränsa antalet inkluderade patienter. Tyvärr kan detta leda till att patientantalet blir alltför litet och att studien saknar förutsättningar för att besvara de frågor vi ställt upp. Vi har för låg styrka (power), dvs ingen rimlig chans att påvisa skillnad mellan jämförda grupper. En sådan studie riskerar att bli bortkastad, vi lyckades ju inte besvara frågan!

Stickprovsstorlek (sample size) är det antal patienter/människor/försöksdjur/observationer vi behöver inkludera i en studie för att kunna förkasta nollhypotesen. Nollhypotesen är alltid att det inte finns någon skillnad mellan jämförda grupper: t ex ingen skillnad i blodtryck mellan diabetiker och patienter med celiaki. Vanligtvis talar man dock om sin alternativa hypotes, t ex: Diabetiker har högre blodtryck än celiakipatienter. Stickprovsstorleken är beroende av flera faktorer, och fyra begrepp är av intresse för den som vill räkna ut stickprovsstorlek.

Sammanfattat

Verkligheten kan ofta begränsa möjligheterna att bedriva forskning. En studie kan t ex stupa på att antalet inkluderade personer är för litet.

Att känna till principen för stickprovsstorlek (sample size) kan då vara till god hjälp.

Klinisk forskning

Läs mer på *www.lakartidningen.se*

Styrka (power) är chansen att hitta en sann skillnad, och vanligtvis väljer man värdet 80 procent. Av detta följer att man vid en styrka på 80 procent har 20 procents risk (1–0,80) att missa en sann skillnad mellan två grupper. Dessa 20 procent är också lika med risker för typ 2-fel (risk att missa en sann skillnad).

Signifikansnivå (ofta angivet som P) är risken att den skillnad

Att räkna ut stickprovsstorlek

Jag minns min första kontakt med begreppet stickprovsstorlek. Vi skulle beräkna antalet patienter som behövdes för att jämföra en behandling av två jämnstora grupper av IBD-patienter [1]. Efter diverse letande i bokhyllor hittade vi ett nomogram som gav en ganska klar uppfattning om behövt antal patienter i var grupp. Men beräkningar av stickprovsstorlek kan göras enklare, och med ett enkelt förfarande skulle fler räkna på styrka före sina studier – det finns goda skäl för det.

Syftet med beräkning av stickprovsstorlek är att ta reda på det behövda antalet patienter i en studie för att utifrån givna förutsättningar kunna påvisa en statistiskt signifikant skillnad i fråga om effektmått.

Alltför få patienter/observationer leder till att man riskerar att missa en faktisk skillnad på populationsnivå, och alltför många patienter/observationer gör studien onödigt dyr och i vissa fall rent av oetisk. Man kan även använda beräkningar av stickprovsstorlek efter genomförd studie för att undersöka vilken styrka man har för en viss skillnad mellan två eller flera grupper.

Förutsättningen för beräkning av stickprovsstorlek är kännedom om ett antal variabler: Typ 1-felet (α) motsvarar risken att hitta en skillnad som inte motsvaras av en faktisk skillnad i populationen (falskt positiv skillnad); α brukar sättas till 0,05. Typ 2-felet (β) motsvarar risken att inte hitta en faktisk skillnad; detta brukar sättas till 0,20. Ofta talar vi om styrka (power), vilket motsvarar »1–β«, som då blir 80 procent.

Vidare måste man bestämma hur stor skillnad man vill kunna (eller man via en pilotstudie förväntar sig) upptäcka. Ju större skillnad, desto större chans att kunna påvisa en statistiskt signifikant skillnad. De ingående värdenas spridning (standardavvikelsen i normalfördelade grupper) måste också anges. I fallet med två grupper, vars värden ligger tätt samlade runt medelvärdet (låg standardeviation), är det förstås lättare att påvisa en skillnad.

Sample Power från SPSS är ett utmärkt program. Dess enda tre nackdelar är att det inte finns för Macintosh, det kostar en hel del och möjliggör inte stickprovsstorleksberäkningar baserade på icke-parametriska test. Programmet är lätt att installera och mycket överskådligt.

Ovannämnda variabler (se föregående) fylls i i de tomma fälten, varefter programmet ger styrkan. Genom att använda funktionen »spin control« (klicka på ikonen som visar en kikare) får man upp antalet patienter som behövs för att precis nå upp till 80 procents styrka. Det går även att räkna ut stickprovsstyrka med olikstora grupper, vilket undertecknad haft stor glädje av i samband med epidemiologiska studier, där man kanske vill studera t ex graviditetsutfall hos en mindre grupp patienter (t ex N = 50 patienter) i en kohort som i övrigt består av »friska« (t ex N = 15 000) [2].

Olika scenarier kan sparas i programmet. En oerhört elegant funktion är dessutom rapportskrivandet. Genom att klicka på en

man hittar mellan sina grupper inte motsvaras av en sann skillnad mellan grupperna om vi tittat på alla individer med de undersökta egenskaperna (typ 1-fel). För att förkasta nollhypotesen vill vi ha ett P-värde under 0,05, men även lägre P-värden brukar redovisas i vetenskapliga arbeten.

Variabilitet är spridningen eller variationen i uppmätta värden. Denna brukar ofta uttryckas som standardavvikelse (SD). Standardavvikelsen kan skilja sig mellan friska (ett slumpmässigt uppmätt blodsocker hos 100 friska personer varierar kanske mellan 3,5 och 6,5) och sjuka patienter (varierar kanske mellan 2,9 och 18,3; större spridning än bland friska). Ibland måste man gissa sig till ett ungefärligt värde på spridningen; man kanske är den första som genomför en viss typ av undersökning. I andra fall, t ex vid mätning av HbA_{1c}, vet man kanske vilka medelvärden som är vanliga hos välinställda diabetiker och vilken spridning dessa värden brukar ha – spridningen mäts vanligtvis som standardavvikelse hos variabler som har en jämn spridning (normalfördelning).

Minsta intressanta skillnad eller effekt uttrycks normalt som en skillnad mellan de undersökta grupperna och brukar motsvara en skillnad som man betraktar som »kliniskt viktig«.

Ett fingerat exempel: Vi vill beräkna stickprovsstorleken (antal patienter och kontroller som behövs) för att visa på en skillnad i blodtryck mellan två behandlingsregimer. Vi utgår från standardvärden för styrka (80 procent) och signifikansnivå (5 procent). För olika värden av styrka och signifikansnivå har man förbestämda koefficienter som används för att beräkna stickprovet. Koefficienten kallas för power index (PI). För en styrka på 80 procent och en signifikansnivå på 5 procent är PI = 2,8. Vi antar att en förändring av blodtrycket på 5 mm Hg är av intresse och att standardavvikelsen för blodtrycket är ca 12 mm Hg. N är antal patienter i vardera gruppen (multipliceras med en faktor 2 för att få fram det antal patienter vi behöver inkludera). Den formel vi använder för ett s k oparat t-test (två oberoende grupper) skrivs:

$$N = 2 \times (PI \times SD/\text{minsta intressanta skillnad})^2.$$

I den här formeln finns styrka och signifikans (PI), spridning (SD) och minsta intressanta skillnad. I vårt fall behöver vi:

$$N = 2 \times (2{,}8 \times 12/5)^2 = 91 \text{ (avrundat uppåt)}$$

Vi behöver alltså inkludera 91 personer i varje grupp (sammanlagt 182). Om spridningen ökar (12 blir 15) ökar också det antal personer som behövs. Om minsta intressanta skillnad minskar (5 blir 4) ökar också det antal personer som behövs. Givetvis finns det formler även för parat t-test och för s k χ^2-test när man jämför proportioner. För dessa hänvisar jag till någon bok i statistik [1-3].

Det viktiga är inte att kunna stickprovsstorleksformeln utantill utan att känna till principerna för den:

- Om vi vill påvisa små skillnader behöver vi ett stort stickprov, och om vi vill påvisa en stor skillnad behöver vi färre personer.
- Om spridningen på våra mätvärden är liten behöver vi ett litet stickprov (mindre risk att gruppernas observerade värden ska överlappa).
- Om vi vill vara säkra på att kunna påvisa en skillnad mellan två grupper kanske vi väljer styrkan 90, då ökar PI och stickprovsstorlek.
- Om vi vill minska risken för att göra ett typ 1-fel sätter vi kanske signifikansnivån till 0,01, även då ökar PI och stickprovsstorlek.

*

Potentiella bindningar eller jävsförhållanden: Författaren har utan kostnad erhållit testversioner av SPSS Sample Power och JMP.

Referenser

1. Altman DG. Practical statistics for medical research. London: Chapman and Hall; 1991.
2. Armitage P, Berry G, Matthews J. Statistical methods in medical research. Oxford: Blackwell Science Ltd; 2002.
3. Hassard TH. Understanding biostatistics. St Louis: Mosby Year Book; 1991.

= artikeln är referentgranskad

knapp ger Sample Power en engelsk formulering med de nyss inknappade värdena i sitt sammanhang. Beskrivningen kan sedan klistras in i en artikel eller annan rapport.

SPSS Sample Power kan även uttrycka förhållandet mellan patientantal och styrka grafiskt. Sample Power är helt fristående från SPSS övriga statistikprogram och kan köpas oberoende av vilket statistikprogram man använder.

JMP är ett statistikprogram som tillverkas av SAS (SAS är också ett statistikprogram). JMP har ett tilltalande utseende och är liksom Sample Power lätt att installera. JMP finns för Macintosh och är ett komplett statistikprogram utöver att det kan räkna ut stickprovsstorlek. Dess riktigt stora fördel ligger dock i priset. Institutioner och studenter kan köpa en version som fungerar i blott fyra år men å andra sidan bara kostar 800 kr exklusive moms (studentlegitimation eller fakultetslegitimation krävs).

Både JMPs sticksprovsstorleksmodul och Sample Power är intuitiva. Med små grundkunskaper i statistik kan man hoppa över manualen och gå direkt på programmet (i JMP hittar man stickprovsstorleksmodulen under rubriken DOE).

I JMP kan man inte räkna med olikstora grupper, vilket är en nackdel. Programmet är heller inte i övrigt lika omfattande som Sample Power, i synnerhet inte vad gäller mer avancerade beräkningar som logistisk regression och överlevnadsanalyser. För många användare torde dock JMP räcka till. Så kallade animation scripts (grafer) för att åskådliggöra skillnader mellan två grupper med kontinuerliga effektmått är dessutom mer tilltalande i JMP än i SPSS.

Sammanfattningsvis finns det flera bra program för stickprovsstorleksberäkning. Såväl JMP som Sample Power har sina för- och nackdelar. De flesta kliniker med forskningsverksamhet är dock betjänta av program för att beräkna stickprovsstorlek.

Tillägg: Det finns även ett antal böcker som innehåller mer [3] eller mindre [4] om stickprovsstorleksberäkningar och även program för beräkning av stickprovsstorlek att ladda ned gratis från Internet, men tillförlitligheten i dem kan undertecknad inte garantera. Recensioner av programvaror publiceras regelbundet i tidskriften Biotech Software and Internet Report.

Kostnad: SPSS Sample Power ca 8 900 kr (http://www.spss.com/se/). JMP professional (corporate version) ca 10 000 kr. JMPin (slutar fungera på din dator efter fyra år – till för studenter och institutioner) 800 kr (http://www.corpus-datamining.com/).

Referenser

1. Ludvigsson J, Krantz M, Bodin L, Stenhammar L, Lindquist B. Elemental versus polymeric enteral nutrition in paediatric Crohn's disease: a multi-centre randomised controlled trial. Acta Paediatr 2004;93:327-35.
2. Ludvigsson JF, Ludvigsson J. Coeliac disease in the father affects the newborn. Gut 2001;49:169-75.
3. Machin D, Campbell M, Fayers P, Pinol A. Sample size tables for clinical studies (With 3 5 Inch Diskette for Windows 3 1 Or 95). Oxford: Blackwell Science Inc; 1997.
4. Altman DG. Practical statistics for medical research. London: Chapman and Hall; 1991.