Metodstöd vid experimentell utvärdering

Innehåll

1. Inledning

Detta dokument, som utgör ett komplement till ”Experimentell utvärdering – en handbok”, vänder sig till metodstatistiker. Det bygger på ett utkast, som har ett bredare innehåll och som togs fram samtidigt som handboken, Björnram m.fl. (2004). Mycket av det som beskrivs nedan gäller även för pilotundersökningar, men fokus ligger på experiment som är inbäddade i pågående undersökningar.

De typer av test som har använts vid SCB för inbäddade experiment är vanliga inom biostatistiken. I det enklaste fallet tänker man sig att det finns två behandlingar: en ny metod som ska jämföras med den som används för närvarande, den ”gamla” metoden. Nedan benämns, med terminologi från biostatistiken, den grupp som får den nya behandlingen experimentgrupp, medan den grupp som får den gamla behandlingen kallas kontrollgrupp.

Det är undersökningens urvalsdesign och den frågeställning som ska besvaras som ligger till grund för hur experimentet ska läggas upp. En viktig förutsättning är att de objekt som ska ingå har valts till experiment- och kontrollgrupp på ett sätt som är slumpmässigt med kända urvalssannolikheter.

Ett genomfört inbäddat experiment utvärderas vanligen med statistisk hypotesprövning. Målet med hypotesprövningen är att ge stöd vid ett beslut, t.ex. om en undersökning ska gå över till en ny insamlingsmetod, en ny kontaktstrategi eller ett förändrat mätinstrument.

En parameter θ används för att beskriva skillnaden mellan den nya och den gamla behandlingen, och θ_0 betecknar det värde på θ som ska testas. Ofta är $\theta_0 = 0$. Nollhypotesen kan då vara $\theta \leq 0$ och alternativhypotesen $\theta > 0$. En annan möjlighet är nollhypotesen $\theta = 0$ och alternativhypotesen $\theta \neq 0$.

Experiment används även för att studera mer nyanserade frågeställningar om olika behandlingars effekter (vilket påverkar θ_0):

- om den nya behandlingen är *betydligt bättre än* den behandling som används för närvarande och inte sämre än eller ungefär likvärdig
- om den nya behandlingen är *bättre än eller ungefär likvärdig* med den behandling som används för närvarande och inte betydligt sämre
- om behandlingarna är *ekvivalenta* i någon mening; att effekterna är ungefär likvärdiga och inte klart olika.

De fortsatta avsnitten behandlar val av effektvariabel, formulering av statistiska test, risker och styrka hos test, urvalsstorlekar och allokering, experimentdesign samt några praktiska aspekter.

2. Val av effektvariabel och hypotestester

2.1. Effektvariabel

Man vill, eller överväger att, införa en ny metod (behandling) i en undersökning. För att kunna jämföra den nya metoden med den befintliga (”gamla”) metoden behöver en effektvariabel definieras. Effektvariabeln ska vara central för undersökningen, och den ska förväntas fånga upp eventuella skillnader mellan den nya och den gamla metoden. I en del experiment är det svårt att definiera behandlingseffekten, det vill säga den resulterande skillnaden mellan metoderna. Hur mäter man t.ex. en högre kvalitet på data eller minskad uppgiftslämnarbörda?

Man kan behöva skapa en särskild variabel som fångar upp syftet – effekten – med den nya metoden. Det kan vara en kombination av befintliga variabler, processdata eller en ny variabel/fråga som skapas just för experimentet. Det kan finnas en frestelse att välja en lättfångad effektvariabel (t.ex. andelen svar i undersökningen). Om denna variabel inte mäter det som man vill få svar på (t.ex. tillförlitlighet) bör man välja en annan variabel.

Ett vanligt fall är att jämföra två metoder, eller *behandlingar* med den biostatistiska termen. Det är en *ny* och en *gammal* behandling, och man vill kunna skatta skillnaden mellan de båda behandlingarna med avseende på en effektvariabel, som kallas y. När skillnader mellan behandlingar diskuteras nedan, är det utan att varje gång understryka att det är med avseende på effektvariabeln.

Den statistiska inferensen kan vara till den ändliga populationen. Det är det fall som beskrivs i fortsättningen. Det är fullt möjligt att välja något annat mål för inferensen om det är mer relevant, t.ex. en delpopulation eller en tänkt superpopulation.

Låt y_{ib} vara det y-värde för det i:te objektet i den ändliga populationen U som fås vid behandling b (som är *ny* eller *gammal*). De parametrar som ska skattas är populationsmedelvärdena för effektvariabeln y om alla populationens N objekt utsätts för behandling:

$$\mu_{ny} = \frac{1}{N}\sum_{i\in U} y_{i.ny} \text{ och } \mu_{gammal} = \frac{1}{N}\sum_{i\in U} y_{i.gammal}$$

I de experiment som förekommer vid SCB är det vanligast att det bara går att genomföra en mätning på varje objekt, och det är den situationen som betraktas här.

För att kunna skatta μ_{ny} och μ_{gammal} dras ett urval, s, bestående av n objekt (i två delmängder, se vidare nedan) ur U. Genom att bilda skattningar av medelvärdena för de båda behandlingarna kan man uttala sig om effekter på populationsnivå (eller för en delpopulation eller vad man har valt för den statistiska inferensen).

2.2. Några hypotestester vid experiment

Beroende på experimentets syfte kan man välja något av följande hypotestester, vilka har kopplingar till punkterna i avsnitt 1 ovan.

1. En undersökning vill införa en ny, mer resurskrävande, metod. Man tycker att den nya metoden ska vara klart bättre för att det ska vara värt att byta. Då passar *superiority-test*.
2. Man har för en undersökning tagit fram en ny metod, och beslutet är att införa denna om inte ett experiment visar att den nya metoden är klart sämre med avseende på den eller de variabler som studeras i experimentet. Då passar *non-inferiority-test*.
3. I en undersökning funderar man på att genomföra vissa tekniska förändringar i kontakten med uppgiftslämnarna. Dessa förväntas inte påverka uppgiftslämnarnas svar. För att få detta prövat i ett experiment passar *ekvivalenstest*.
4. Man har redan genomfört en förändring, och man vill se om det blev någon mätbar skillnad. Det är då mer informativt att skatta förändringen med ett konfidensintervall än att utföra ett hypotestest.

I fall 1–3 ovan är den intressanta parametern vanligen skillnaden mellan experiment- och kontrollgruppernas parametrar:

$$\theta = \mu_{ny} - \mu_{gammal} \tag{1}$$

Parametern μ_{ny} kan t.ex. vara den svarsandel som det nya utskicket (teoretiskt) skulle ge med en totalundersökning. Parametern θ_0 betecknar som förut det värde på θ som ska testas; i regel är $\theta_0 = 0$.

2.3. Ett traditionellt ensidigt test

En hypotesprövning med en noll- och en alternativhypotes kan t.ex. ha nedanstående utseende, om testet är ensidigt.

$$\begin{aligned} H_0&: \theta \leq \theta_0 \\ H_1&: \theta > \theta_0 \end{aligned} \tag{2}$$

Om $\theta_0 = 0$ är nollhypotesen att den nya metoden ger högst samma värde som den gamla. Alternativhypotesen är att den nya metoden ger ett högre värde.

Låt $\hat{\theta}$ beteckna en estimator av parametern θ i (1). Om $\hat{\theta}$ kan antas vara normalfördelad med känd varians kan följande teststatistika användas:

$$Z = \frac{\hat{\theta} - \theta_0}{\sqrt{Var(\hat{\theta})}}. \tag{3}$$

Om

$$Z > c \tag{4}$$

där c är ett förutbestämt värde, så förkastas nollhypotesen. Värdet c har bestämts så att önskad signifikansnivå på testet uppnås. Se vidare nedan om variansskattningar och signifikansnivåer.

2.4. Toleransgräns

En vanlig situation i experimentell utvärdering är att numeriskt små skillnader mellan den nya och den gamla behandlingen är oviktiga. Genom att bestämma en *toleransgräns*, ett tal $\delta > 0$, kan hypotesen i (2) omformuleras så att skillnader som saknar praktisk betydelse, eller är oviktiga, inte betraktas som signifikanta.

För att synliggöra toleransgränsen ersätts θ_0 med $\theta_0 + \delta$ eller $\theta_0 - \delta$, där δ är ett litet positivt tal. Ofta är $\theta_0 = 0$ och då handlar toleransgränsen om triviala avvikelser från 0 i $\theta = \mu_{ny} - \mu_{gammal}$; avvikelser som är ”praktiskt taget 0”. Se t.ex. avsnitt 2.6.

Toleransgränsen δ kan sättas ur en ”ämnessynvinkel”: en gräns sätts mellan intressant skillnad och föga intressant skillnad. Det är även möjligt att diskutera δ i relation till andra parametrar, se avsnitt 3.1.

Även det omvända kan gälla, att man t.ex. av kostnadsskäl vill byta metod enbart om den nya är klart bättre än den gamla, se avsnitt 2.5.

2.5. Superiority-test

I situationer där man vill visa att en metod är klart bättre än den tidigare passar ett superiority-test. Man sätter då upp noll- och alternativhypoteserna nedan för att se om ”θ är minst δ större än θ_0”.

$$\begin{aligned} &H_0: \theta \leq \theta_0 + \delta \\ &H_1: \theta > \theta_0 + \delta \end{aligned} \tag{5}$$

Nollhypotesen förkastas om

$$Z = \frac{\hat{\theta} - (\theta_0 + \delta)}{\sqrt{Var(\hat{\theta})}} > c. \tag{6}$$

2.6. Non-inferiority-test

En ofta förekommande situation är att man vill undersöka om den nya metoden är minst lika bra som den gamla. Om man väljer en toleransgräns $\delta > 0$ kan hypotesen i (2) omformuleras till ett non-inferiority-test som testar

$$\begin{aligned} &H_0: \theta \leq \theta_0 - \delta \\ &H_1: \theta > \theta_0 - \delta \end{aligned} \tag{7}$$

Nollhypotesen innebär att ”θ understiger θ_0 med minst δ”. Den förkastas om

$$Z = \frac{\hat{\theta} - (\theta_0 - \delta)}{\sqrt{Var(\hat{\theta})}} > c. \tag{8}$$

2.7. Ekvivalenstest

Ett test som är tänkt att påvisa likvärdighet mellan metoder benämns ekvivalenstest.

Testet är användbart vid utvärdering t.ex. då teknisk utrustning byts men frågeformuläret är oförändrat. Man vill se om svaren på frågorna är väsentligen desamma för gammalt och nytt upplägg. Det resultat som skulle vara avvikande (att byte av utrustning påverkar svaren med mer än toleransgränsen δ) är därför satt som nollhypotes.

En toleransgräns δ används, men noll- och alternativhypoteserna formuleras lite annorlunda än tidigare:

$$\begin{aligned} H_0&: |\theta - \theta_0| \geq \delta \\ H_1&: |\theta - \theta_0| < \delta. \end{aligned} \tag{9}$$

Noll- och alternativhypoteserna har bytt plats i jämförelse med (5) och (7). Även teststatistikan blir lite annorlunda. Med $\theta_0 = 0$ (en vanlig situation) förkastas nollhypotesen om

$$\frac{\hat{\theta} - \delta}{\sqrt{Var(\hat{\theta})}} < -c \ \textbf{och} \ \frac{\hat{\theta} + \delta}{\sqrt{Var(\hat{\theta})}} > c. \tag{10}$$

Ekvivalenstestet är två simultana enkelsidiga test. För en utförligare beskrivning se bilagan, där intervall och val av kvantil (percentil) behandlas.

3. Avvägningar och val

3.1. Avvägning mellan risker

Den verklighet som ett hypotestest ska pröva kan i princip förhålla sig på två sätt: antingen noll- eller alternativhypotesen är sann. Vartdera förhållandet är förknippat med en risk: att testet felaktigt kommer fram till att det motsatta förhållandet skulle vara sant, se beslutsmatrisen nedan. Det är inte ett symmetriskt förhållande.

Tillstånd / Beslut	H_0 är sann	H_0 är falsk
Förkasta H_0	Typ I-fel. P(typ I-fel) = α	
Förkasta ej H_0		Typ II-fel. P(typ II-fel) = β

Parametern α motsvarar risken för att få ett fel av typ I, det vill säga att förkasta nollhypotesen i det fall då nollhypotesen är sann. Denna risk kontrolleras genom att i testet välja c i (6), (8) respektive (10) så att önskad risk erhålls. När nollhypotesen är ett intervall är α det högsta värdet på risken.

Även om valet $\alpha = 0{,}05$ är konventionellt kan man i många sammanhang tänka sig att utforma experimentet så att α är större, t.ex. 0,10. Detta kan t.ex. gälla i ett non-inferiority-test om konsekvenserna av ett felbeslut inte är alltför stora.

Parametern β motsvarar risken för ett fel av typ II, att underlåta att förkasta nollhypotesen om den är falsk. Sannolikheten för motsatsen, det vill säga att testet korrekt förkastar nollhypotesen (för ett visst värde på parametern θ; alternativhypotesen är ofta ett intervall), kallas testets styrka, $1 - \beta$. Det råder ett motsatsförhållande mellan α och β. Det man tjänar på ett större α är en förbättring av testets styrka: sannolikheten att korrekt förkasta nollhypotesen i det fall då den är falsk. Experimentets test kan läggas upp med målsättningen att denna sannolikhet, testets styrka, är minst ett givet värde på $1 - \beta$ (för något valt värde på θ).

Valet av α och β bör avspegla de risker man är beredd att ta. Det är vanligt att välja $\alpha = 0{,}05$ och $\beta = 0{,}10$ eller $\beta = 0{,}20$. Dessa val avspeglar det vanliga förhållandet att risken bör vara asymmetriskt fördelad. Det är en avvägning mellan risken α att ovetandes gå över till en ny metod som är i någon mening sämre än den befintliga meto-

den och risken β att gå miste om den nya metoden trots att den hade varit bättre.

En annan aspekt är vad som är möjligt och realistiskt. Det finns fyra parametrar att bestämma i experimentplaneringen: utöver α och β är det toleransgränsen δ – se (5), (7) respektive (9) – samt urvalsstorleken n. Om tre av parametrarna är bestämda, kan den fjärde parametern härledas.

I ett experiment som är helt inbäddat i en ordinarie undersökning är n bestämt. Det kan då bli så att de önskade α och β skulle leda till ett oönskat stort δ. Detta upplägg kan bara undvikas genom att göra ändringar i de tre parametrarna (α, β och δ) utöver n eller att lägga upp experimentet på annat sätt.

3.2. Urvalsstorlekar och styrka

I designen av en urvalsundersökning tar man ofta hänsyn till att man vill ha god precision (liten varians) i skattningarna av olika redovisningsgrupper för någon viktig undersökningsvariabel. I kliniska prövningar beräknas behandlingsgruppernas storlek med hjälp av den primära effektvariabeln, vilket är en parallell till val av design i en urvalsundersökning. För att kunna upptäcka meningsfulla skillnader krävs ett tillräckligt stort urval i såväl kontroll- som experimentgrupp.

Möjligheten att bestämma urvalsstorleken beror givetvis av tillgängliga resurser, men även av vilken typ av undersökning som experimentet är inbäddat i. Om det är en pågående undersökning är urvalet bestämt på förhand och möjligheten att välja gruppstorlekar är begränsat. Det kan hända att urvalet i sig är tillräckligt stort för att man med hög styrka, exempelvis $1 - \beta \geq 0{,}80$, ska kunna upptäcka en meningsfull skillnad vid ett bestämt värde på risken α. Det kan också vara så att det krävs ett extra stort urval under den produktionsomgång då experimentet görs, alternativt att experimentet pågår under flera produktionsomgångar.

Om det är en stor skillnad i resultaten mellan experiment- och kontrollgrupperna kan kanske endast data från den ena gruppen användas i redovisningen av statistiken. En säkerhetsåtgärd kan därför vara att begränsa experimentgruppens storlek. Det är vanligt att ansvarig(a) är beredd att riskera en viss försämring i precisionen för den redovisade statistiken och föreslår gruppstorlekarna med hänsyn till detta. Det som då kan göras är att beräkna toleransgränsen för de valda gruppstorlekarna.

I en pilotundersökning är det möjligt att beräkna gruppstorlekarna i förväg för att bestämma experimentets storlek. Om huvudskälet till pilotundersökningen är en experimentell utvärdering bör gruppernas

storlek väljas efter en förbestämd toleransgräns och styrka. Om det är andra skäl som väger tyngre begränsas möjligheterna att välja gruppstorlekar.

3.3. Allokering med hänsyn till grupper

Hur bestäms antalet objekt i varje behandlingsgrupp i den typ av experiment som har beskrivits?

I de flesta fall är gruppstorlekar inte baserade på formella styrkeberäkningar utan på vad man ”har råd med”. Då görs approximativa styrkeberäkningar för att visa hur stor osäkerhet som kan förväntas vid ett statistiskt test. Detta ger de ansvariga en möjlighet att förändra förutsättningarna. I de fall man har valt att gå vidare med ett experiment trots låg styrka, vet man vad man kan förvänta sig.

Experimentets syfte påverkar gruppstorlekarna (eller toleransgränsen). Nedan beskrivs hur allokering görs för att uppnå önskad risknivå och styrka när man vill skatta skillnaden mellan två olika behandlingar med superiority-, non-inferiority- respektive ekvivalenstest.

Som förut används beteckningen $\theta = \mu_{ny} - \mu_{gammal}$ för att beskriva skillnaden mellan den nya och den gamla behandlingen, och θ_0 betecknar det värde på θ som ska testas (i regel är $\theta_0 = 0$). Testerna bygger på att urvalen är (tillräckligt) stora och att teststatistikan betraktas som normalfördelad. Om urvalsdesignen är ett OSU utan återläggning, ändlighetskorrektioner kan försummas och kovarianser är approximativt noll är följande en approximation av variansen:

$$Var\left(\hat{\theta}\right) = Var\left(\bar{y}_{ny} - \bar{y}_{gammal}\right) \approx \tag{11}$$

$$Var\left(\bar{y}_{ny}\right) + Var\left(\bar{y}_{gammal}\right) = \frac{S_{ny}^2}{n_{ny}} + \frac{S_{gammal}^2}{n_{gammal}}$$

n_{ny} och n_{gammal} är antal objekt allokerade till de två behandlingarna och S_{ny}^2 och S_{gammal}^2 är populationsvarianserna, som ofta antas vara lika. I formlerna nedan används det gemensamma S^2 vilket approximativt motsvarar variansen σ^2 under normalfördelningsantagandet. Notera att andra approximationer kan vara att föredra vid andra urvalsdesigner. Om σ^2 är ”känd”, t.ex. från tidigare genomföranden av undersökningen, kan det vara en möjlighet att använda det värdet.

Exempel 1 Superiority- och non-inferiority-test

För superiority-test formuleras följande hypotes, givet att $\delta > 0$:

$H_0: \theta \leq \theta_0 + \delta$
$H_1: \theta > \theta_0 + \delta$

Om i stället $\delta < 0$ blir det ett non-inferiority-test. Då kan man, som förut, i stället skriva $\theta_0 - \delta$ och ha $\delta > 0$.)

Under alternativhypotesen, H_1, och för givna värden på z_α, z_β och δ, där z_α och z_β är α- respektive β-kvantilen i standardnormalfördelningen, kan beräkningar göras. Även θ och θ_0 måste bestämmas, vanligen sätts $\theta - \theta_0 = 0$. Urvalsstorlekar kan då beräknas enligt nedan.

Eftersom experimentet bara har två grupper måste $n_{ny} + n_{gammal} = n$ och därmed kan n_{gammal} skrivas som $n_{gammal} = \omega n_{ny}$ och

$$n_{ny} = \frac{(z_\alpha + z_\beta)^2 \, S^2 (1 + \frac{1}{\omega})}{(\theta - \theta_0 - \delta)^2} \tag{12}$$

Omskrivning av formeln ger följande uttryck

$$(\theta - \theta_0 - \delta)^2 = \left(z_\alpha + z_\beta\right)^2 S^2 (\frac{1}{n_{ny}} + \frac{1}{n_{gammal}})$$

genom vilket toleransnivån kan bestämmas. Under alternativhypotesen, H_1, gäller $\theta - \theta_o - \delta > 0$, vilket i situationen $\theta - \theta_0 = 0$ ger följande toleransgräns.

$$\delta = -(z_\alpha + z_\beta)\sqrt{S^2 \, (\tfrac{1}{n_{ny}} + \tfrac{1}{n_{gammal}})}.$$

Formeln gäller även non-inferiority-test om man har $\delta > 0$.

Exempel 2 Ekvivalenstest

För ekvivalenstest formuleras följande hypotes (vilket betyder att $\delta \geq 0$):

$H_0: |\theta - \theta_0| \geq \delta$
$H_1: |\theta - \theta_0| < \delta.$

Under alternativhypotesen, H_1, och för givna värden på z_α, z_β och δ, där z_α och z_β är α- respektive β-kvantilen i standardnormalfördelningen, kan beräkningar göras. Även θ och θ_0 måste bestämmas, vanligen sätts $\theta - \theta_0 = 0$. Urvalsstorlekar kan då beräknas enligt nedan.

Eftersom experimentet bara har två grupper måste $n_{ny} + n_{gammal} = n$, och därmed kan n_{gammal} skrivas som $n_{gammal} = \omega n_{ny}$ och

$$n_{ny} = \frac{\left(z_\alpha + z_{\beta/2}\right)^2 \, S^2 \left(1 + \frac{1}{\omega}\right)}{(\delta - |\theta - \theta_0|)^2}$$

Även i detta fall skrivs formeln om för att beräkna toleransnivån:

$$(\delta - |\theta - \theta_0|)^2 = \left(z_\alpha + z_{\beta/2}\right)^2 S^2 (\frac{1}{n_{ny}} + \frac{1}{n_{gammal}})$$

Under alternativhypotesen, H_1, gäller $|\theta - \theta_0| < \delta$ vilket i situationen $\theta - \theta_0 = 0$ ger följande toleransgräns:

$$\delta = -\left(z_\alpha + z_{\beta/2}\right)\sqrt{S^2\left(\frac{1}{n_{ny}} + \frac{1}{n_{gammal}}\right)}$$

3.4. Flera tester i samma experiment

Ofta vill man med ett experiment mäta flera olika saker. Det blir då ännu viktigare än annars att formulera vad experimentet ska besvara för frågor. Ett problem är att man med stor sannolikhet kommer att få statistiskt signifikanta utfall på vissa av testerna, utan att kunna avgöra vilka av dem som korrekt påvisar en verklig skillnad och vilka som beror på slumpen. Om tio test görs, alla individuellt med $\alpha = 0{,}05$, blir risken för minst en falsk signifikans (det vill säga fel av typ I) ungefär 40 %. Den sammanlagda risken för fel av typ I kallas ”experimentfel”; i det här fallet är den risken 0,40.

För att minimera experimentfelet är det viktigt att experimentets syfte är klarlagt och att relevanta effektvariabler väljs. Den första frågan att ställa sig är om de som planerar experimentet verkligen är intresserade av att få svar på alla frågor som man funderat på, även om det innebär ökade kostnader? Om man ser att en stor mängd frågor som man vill testa formellt alla är av ungefär samma typ, är det lämpligt att dela in testerna i grupper efter behandling. Nedan diskuteras två typer av experiment där man vill genomföra flera tester.

Om man i ett experiment vill få bekräftat att ett byte av teknik inte påverkar skattningarna för någon av 50 variabler i 20 redovisningsgrupper, måste $50 * 20 = 1000$ test genomföras. Om alla test genomförs, kan antalet signifikanta utfall med $\alpha = 0{,}05$ användas för att testa om andelen signifikanta utfall är väsentligt större än 5 %. Om inte, finns ingen anledning att tro att teknikbytet har påverkat skattningarna. Här söker man alltså svaret på frågan om det finns *någon* skillnad mellan skattningarna för experiment- och kontrollgrupp, inte vilka dessa skillnader i så fall är.

Om experimentet å andra sidan gäller t.ex. byte av frågeformulär med radikalt nya formuleringar för ett antal frågor kan man tänka sig att svaren på en del av dem skiljer sig väsentligt från dem för de gamla frågorna. Här vill man försöka identifiera de ”verkliga” skillnaderna, i motsats till de som uppstår av en slump.

Experimentfelet kan bli stort om alla frågor testas var för sig med $\alpha = 0{,}05$. Det finns tre sätt att hantera detta:

- acceptera stort experimentfel och den därmed sammanhängande oklarheten i tolkningen av resultaten
- höja kraven på vad som ska betraktas som statistiskt signifikant
- begränsa antalet test som görs eller fokusera mer på vissa test än andra.

Ett vanligt sätt att höja kraven på vad som ska betraktas som statistiskt signifikant är att dividera α med antalet test, vilket är en så kallad Bonferroni-justering. Planerar man tre test, kan det vara lämpligt att sätta $\alpha = 0{,}05/3$. Dock är det kanske inte rimligt att ha $\alpha = 0{,}00005$ för 1 000 test, eftersom även mycket stora skillnader mellan experiment- och kontrollgrupp då kommer att bedömas vara icke-signifikanta. I Holms test rangordnas testerna, säg m stycken, efter deras teststatistikor med den största först. Den största testas med $\alpha = 0{,}05/m$, den näst största med $\alpha = 0{,}05/(m-1)$, etc. Se t.ex. Holm (1979).

Det finns mycket som talar för alternativet att begränsa antalet test. Den primära effektvariabeln specificeras, och den testas med ett av de beskrivna testen. Sedan kan man därutöver specificera en eller ett par sekundära effektvariabler, testa dem individuellt med t.ex. $\alpha = 0{,}05$ och acceptera ett visst experimentfel. De övriga frågor som man gärna vill ha svar på därutöver lämnas därhän eller så testas de. Då har problemet åtminstone flyttats bort från de viktigaste frågorna.

4. Experimentdesign

4.1. Inledning

Normalt sett är det bästa att testa en förändring i taget för att veta vad som verkligen ger resultat. Ibland är man kanske övertygad om att det krävs mer än en åtgärd för att förbättra undersökningen. I sådana fall är ett ”åtgärdspaket” som omfattar alla de föreslagna åtgärderna det mest kostnadseffektiva. I ett sådant kontrollerat experiment jämför man åtgärdspaketet med den gamla metoden eller behandlingen som är kontrollgruppen.

Den strategi som hittills har beskrivits är att man så långt det är möjligt väljer ut ett primärt syfte med experimentet. Om det inte går bör man försöka att rangordna syftena. Det blir då lättare att kartlägga förutsättningarna och börja sökandet efter lämpliga effektvariabler. Rangordningen av syftena medför att även effektvariablerna blir rangordnade.

Om det är bestämt att man ska mäta behandlingseffekten på undersökningsvariabler så måste man på något sätt försöka välja ut någon eller några variabler. Det finns ibland ett beroende mellan variablerna, vilket gör att om en variabel påverkas av ett experiment så påverkas också ett antal andra. Det räcker då att välja ut en av dessa variabler. Här bör ansvariga fundera på vilka variabler som kan komma ifråga; vilka är statistikanvändarna mest intresserade av?

Vidare bör man fundera över vilka ytterligare faktorer som kan påverka valda effektvariabler. Det varierar mellan undersökningar, och man bör gå igenom undersökningens val av

- urvalsobjekt (individer, företag etc.)
- intervjuare (om intervjuarledd undersökning)
- urvalsdesign (stratifieringsvariabler, hjälpinformation etc.).

Ska experimentet genomföras på hela urvalet eller är det särskilt riktat mot en delpopulation? Vill man t.ex. studera företag med utländska ägare eller invandrare med akademisk utbildning? Det kan tänkas att det av praktiska skäl blir enklare att genomföra experimentet i ett delurval, men att resultaten ändå går att generalisera till hela populationen. Om man t.ex. på goda skäl kan anta att geografisk region inte påverkar effektvariablerna kan man i en intervjuundersökning av ekonomiska skäl välja att genomföra experimentet i en geografisk intervjuarregion. Om det är ett delurval i undersökningen som experimentet genomförs på, måste man se till att tillräckligt många objekt ingår för att en statistisk utvärdering ska kunna göras.

Många av SCB:s undersökningar använder stratifierat urval. Detta måste experimentdesignen ta hänsyn till. Eftersom urvalsobjekt från

samma stratum är mer lika varandra än urvalsobjekt från olika strata, bör stratifieringsvariablerna ingå som faktorer i experimentdesignen. Det bästa är om randomiseringen på behandling sker stratumvis. Om det finns många strata kan det vara svårt att genomföra en randomisering stratumvis. Man kan då behöva slå ihop strata för att få tillräckligt många objekt i de olika behandlingarna.

Vid valet av faktorer som ska ingå i experimentdesignen får man göra en avvägning och en anpassning till undersökningen som experimentet är inbäddat i. Det kan t.ex. vara svårt att få med alla strata och alla intervjuare som faktorer i experimentdesignen. Det viktiga är att man inte utelämnar någon central faktor så att förutsättningarna att utvärdera behandlingarna förstörs. Två vanliga experimentdesigner som tar hänsyn till olika faktorer som kan påverka effektvariabeln beskrivs i avsnitt 4.2 och 4.4.

Behandlingsgrupperna kommer att innehålla färre observationer än det antal som vanligen används i samband med estimationen. En fördelning av de olika behandlingarnas objekt på hjälpvariablerna gör det möjligt att avgöra om en reducering av hjälpinformationen i estimationen är nödvändig. Relevanta hjälpvariabler bestäms av metodstatistiker och ansvariga för att kunna skapa nya vikter och en ny estimator.

4.2. Fullständigt randomiserad design (CRD)

En fullständigt randomiserad design (completely randomized design, CRD) bygger på att det ursprungliga urvalet (oavsett urvalsdesign) delas upp slumpmässigt på de olika behandlingarna, som ses som underurval. Varje sådant underurval kan då betraktas som ett slumpmässigt urval från populationen. Om den ursprungliga designen är ett obundet slumpmässigt urval så kommer underurvalen också att vara det. I de flesta fall, när mer sofistikerade urvalsdesigner används, överförs inte den ursprungliga urvalsdesignen på underurvalen.

4.3. Skattningar vid CRD

Ett urval s dras från den ändliga populationen U bestående av N objekt. Låt, för den ursprungliga urvalsdesignen, $\pi_i, i \in s$ beteckna första ordningens inklusionssannolikhet. Ur detta urval s dras nu med OSU underurval s_b av storleken n_b för varje behandling, b =1, ..., B. Inklusionssannolikheten för behandling b blir då $\pi_i^* = \left(\frac{n_b}{n}\right)\pi_i$ för $i \in s$, och Horvitz-Thompsons estimator för medelvärdet vid behandling b kan skrivas

$$\hat{\bar{Y}}_b = \frac{1}{N}\sum_{i \in s_b} \frac{y_i}{\pi_i^*}$$

Betrakta, för att beskriva variansskattningen, två behandlingar: *ny* och *gammal* (metoden kan generaliseras). Man vill nu hitta variansestimatorn för följande differens:

$$\hat{\theta} = \hat{\bar{Y}}_{ny} - \hat{\bar{Y}}_{gammal}$$

Eftersom underurvalen s_{ny} och s_{gammal} är dragna från s utan återläggning är $\hat{\bar{Y}}_{ny}$ och $\hat{\bar{Y}}_{gammal}$ beroende, vilket innebär att man har en kovarians i urvalsmening mellan underurvalen. En estimator av variansen kräver dessutom att båda behandlingarna är gjorda på samtliga objekt i det ursprungliga urvalet s, vilket inte motsvarar situationen här, eftersom varje objekt bara väljs ut till en behandling. Detta innebär att man inte kan få en väntevärdesriktig variansestimator för $\hat{\theta}$ (under den ursprungliga urvalsdesignen och underurvalen). I litteraturen ges emellertid en approximativt väntevärdesriktig variansestimator för $\hat{\theta}$ som kan skrivas

$$\widehat{Var}(\hat{\theta}) = \hat{d}_{ny} + \hat{d}_{gammal} \tag{13}$$

där

$$\hat{d}_b = \frac{1}{n_b}\frac{1}{n_b - 1}\sum_{i \in s_b}\left[\frac{ny_i}{N\pi_i} - \frac{n}{n_b}\sum_{j \in s_b}\frac{ny_j}{N\pi_j}\right]^2 \tag{14}$$

för b=*ny* och b=*gammal*.

Beroende på hur många behandlingar man studerar kan lämpliga teststatistikor formuleras (för Wald- eller t-test). Se van den Brakel (2001, 2013).

4.4. Randomiserad blockdesign (RBD)

I många fall går det inte att använda en CRD av praktiska skäl. Exempelvis måste i en besöksundersökning intervjuarnas geografiska arbetsområden begränsas av ekonomiska skäl. Det kan då vara praktiskt att t.ex. gruppera ihop angränsande arbetsområden med hänsyn till antalet behandlingar och randomisera behandling på intervjuare. Varje sådan gruppering av intervjuare kan då betraktas som ett block. Inom varje block genomförs sedan en slumpmässig uppdelning på behandling. Då får man ett experiment med randomiserad blockdesign (RBD).

4.5. Skattningar vid RBD

På samma sätt som för CRD dras ett urval s från den ändliga populationen U bestående av N objekt där $\pi_i, i \in s$ betecknar första ordningens inklusionssannolikhet. Objekten i detta urval s delas nu på ett deterministiskt sätt in i K block där s_k motsvarar delmängden

för block *k* av storleken n_k. Varje objekt i s_k randomiseras till en behandling, *b* =1, ..., *B*, där s_{kb} i sin tur motsvarar delmängden för behandling *b* i block *k* bestående av n_{kb} objekt. Inklusionssannolikheten för behandling *b* i block *k* blir då $\pi_i^* = \left(\frac{n_{kb}}{n_k}\right)\pi_i$ för $i \in s_{kb}$ och Horvitz-Thompson-estimatorn för behandling *b* kan skrivas

$$\hat{\bar{Y}}_b = \frac{1}{N}\sum_{i \in s_b}\frac{y_i}{\pi_i^*}$$

En variansestimator behövs för följande differens:

$$\hat{\theta} = \hat{\bar{Y}}_{ny} - \hat{\bar{Y}}_{gammal}$$

Motiveringen för RBD-estimatorn är densamma som för CRD-estimatorn, och en approximativt väntevärdesriktig variansestimator för $\hat{\theta}$ kan skrivas

$$\widehat{Var}(\hat{\theta}) = \hat{d}_{ny} + \hat{d}_{gammal}$$

där

$$\hat{d}_b = \sum_{k=1}^{K}\frac{1}{n_{kb}}\frac{1}{n_{kb}-1}\sum_{i \in s_{kb}}\left[\frac{ny_i}{N\pi_i} - \frac{n}{n_b}\sum_{j \in s_b}\frac{ny_j}{N\pi_j}\right]^2 \qquad (15)$$

för *b* = *ny* och *b* = *gammal.*

Beroende på hur många behandlingar man studerar kan lämpliga teststatistikor formuleras.

5. Praktiska aspekter

5.1. Implementering av randomiseringsdesignen

I samband med att experimentplanen fastställs måste man också planera hur man i praktiken ska gå tillväga när man skapar gruppindelningar, det vill säga hur man drar underurvalen, hur gruppindelningarna kommer att användas i experimentet och hur man kommer att säkerställa att den gjorda gruppindelningen följs. Ansvarig(a) ska utses.

Man bör också ha en plan för hur man ska agera om det visar sig att gruppindelningen av någon anledning inte har följts.

Man måste bestämma hur och var randomiseringsinformationen ska lagras. Om viss personal som deltar i arbetet med undersökningen eller experimentet inte ska känna till gruppindelningen (t.ex. på grund av ”blindning”, se nedan) måste man även besluta hur man säkerställer att informationen inte sprids till dessa ”obehöriga”.

5.2. Blindning (främst intervjuundersökningar)

I biostatistiska experiment talar man ofta om blindning, det vill säga att försökspersonerna inte vet vilken behandling de får. Om inte heller ”behandlarna” vet vilken försöksperson som får vilken behandling talar man om ett dubbelblint experiment. I ett experiment måste man ta ställning till om urvalspersonerna (urvalsobjekten) ska vara ”blindade”, och vidare, om det är en intervjuundersökning, om intervjuarna ska få veta att de ingår i ett experiment och vilken behandlingsgrupp de tillhör. Ansvaret för att besluta om blindning ligger främst på metodstatistiker, produktansvarig och undersökningsledare (motsvarande).

Det är egentligen vetskapen om behandlingsgruppstillhörighet som avgör om experimentet är ”blint” i traditionell mening. Det är enbart om personerna känner till att de deltar i ett experiment som det även är aktuellt att ta ställning till om de ska få veta vilken behandlingsgrupp de tillhör.

När det gäller urvalspersonerna (urvalsobjekten), väljer man vanligtvis att inte tala om att de deltar i ett experiment eftersom det skulle kunna medföra att de blev mer negativa till att delta. Intervjuarna däremot känner nästan alltid till att de deltar i ett experiment, och ofta är det också nödvändigt (eller oundvikligt) att de även får vetskap om vilken behandlingsgrupp de tillhör. Däremot får intervjuarna inte alltid information om vilken behandlingsgrupp urvalspersonerna tillhör.

5.3. Hantering av avvikelser och oväntade resultat

För att kunna hantera avvikelser i undersökningen eller experimentet behövs en plan för hur man ska göra i de situationer som man, i detta skede, föreställer sig kan komma att inträffa. Det handlar både om att identifiera omständigheter i själva undersökningen som kan få negativa konsekvenser och att upptäcka möjliga problem i experimentet. Vissa avvikelser och onormala resultat kanske är acceptabla, medan andra kan antas komma att ”förstöra” undersökningen eller experimentet om inget görs i tid.

Om man måste avbryta experimentet, är det bra om man även har en förutbestämd avbrottsstrategi. En form av strukturerad avbrottsstrategi, som ger möjlighet till en statistisk analys av den primära effektvariabeln, är *sekventiell experimentdesign*. Designen inbegriper bl.a. en eller flera interimsanalyser, det vill säga analyser efter en viss tid av experimentet, för att se om resultaten tyder på att det är värt att fullfölja experimentet. Det kan t.ex. handla om att kontrollera om resultaten pekar åt ett visst håll eller om det finns störande faktorer som gör att man inte kan analysera den primära effektvariabeln på det sätt som man hade tänkt.

Om något skulle gå snett i experimentet, exempelvis om det är drastiskt stigande bortfall i en av behandlingsgrupperna, bör man ha en färdig plan för hur man ska hantera detta, så att undersökningen som experimentet är inbäddat i inte ”förstörs”, och så att experimentet, om möjligt, går att analysera. Man bör även ha avbrottsstrategier för situationer som inte täcks in av en sekventiell experimentdesign, t.ex. avvikelser från experimentplanen eller extrema resultat för andra variabler än den primära effektvariabeln.

Bilaga: Hur ekvivalenstest fungerar

I många sammanhang är det lämpligt att testa om något är ”tillräckligt lika” (bl.a. avsnitt 2.7 ovan). Det kan göras med följande ekvivalenshypotes:

$$H_0: |\theta - \theta_0| \geq \delta \qquad (16)$$
$$H_1: |\theta - \theta_0| < \delta.$$

där θ är den intressanta parametern, θ_0 det värde på θ som man testar – för enkelhets skull sätts $\theta_0 = 0$ nedan – och $\delta > 0$ är en toleransgräns. Ett vanligt sätt att utföra testet i (16) är Westlakes 2α-procedur, även kallad ”Two One-Sided Tests”, som förkortas TOST. Det vanligaste sättet att utföra ett TOST är att beräkna ett konfidensintervall.

$$\left(\hat{\theta} - z_{1-\alpha}\sqrt{Var(\hat{\theta})}\,, \hat{\theta} + z_{1-\alpha}\sqrt{Var(\hat{\theta})}\right) \qquad (17)$$

Om $\alpha = 0{,}05$ så är $z_{1-\alpha}$ den 95:e percentilen för normalfördelningen (och inte 97,5 %). Om urvalet är litet och normalfördelning kan antas, så kan *t*-fördelningen användas i stället. Om intervallet i (17) ligger helt inom $[-\delta, \delta]$ förkastas H_0.

Ett test motsvaras vanligen av en ekvivalent procedur som innebär att man beräknar ett konfidensintervall och ser om det ligger helt inom det kritiska området. Om det gör det, så förkastas nollhypotesen. Om konfidensintervallet sticker ut med någon bit bortom det kritiska området så kan nollhypotesen inte förkastas.

För ett lämpligt $100(1-\alpha)$ % – konfidensintervall (D^-, D^+) kan man tycka att TOST borde vara ekvivalent med proceduren att om $(D^-, D^{+)}$ konstateras vara helt inkluderat i $[-\delta, \delta]$ så förkastas H_0.

Det konfidensintervall som används i ett TOST är emellertid – vilket kan vara förvirrande – ett $100(1-2\alpha)$ % − intervall, och inte motsvarande $100(1-\alpha)$ % − intervall. För detaljer, se t.ex. Berger och Hsu (1996) samt Walker och Nowacki (2011); även den tidiga diskussionen i Kirkwood och Westlake (1981).

Det finns flera sätt att utföra ett ekvivalenstest än Westlakes 2α–procedur som har beskrivits ovan. I de situationer som är vanliga vid SCB är denna procedur enligt erfarenheten hittills att föredra.

Referenser och vidare läsning

Berger, R.L., och Hsu, J.C. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. Statistical Science, **11**, no. 4, 283–319; inkl. kommentarer och diskussion http://projecteuclid.org/download/pdf_1/euclid.ss/1032280304

Björnram, A., Boynton, I.-M., Hedlin, D. och Lundquist P. (2004). Experimentell utvärdering, Del 2. Utkast juni 2004. SCB.

van den Brakel, J. (2001). Design and analysis of experiments embedded in complex sample surveys. Ph.D. dissertation, Erasmus Universiteit, Rotterdam.

van den Brakel, J. (2013). Design-based analysis of factorial designs embedded in probability samples. Survey Methodology, **39**, 323–349.

Casella, G. och Berger, R. L. (1990). Statistical Inference. Belmont: Duxbury Press.

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. Scand. J. Statist., **6**, 65–70.

Kirkwood, Th. B. L. och Westlake, W. J. (1981). Bioequivalence Testing – A Need to Rethink. Biometrics, **37**, 589–594.

Montgomery, D. (2009). Design and Analysis of Experiments. Wiley. (Finns flera upplagor.)

SCB (2005). Experimentell utvärdering – en handbok. SCB. Handbok 2005:1

Walker, E., och Nowacki, A. S. (2011). Understanding Equivalence and Noninferiority Testing. *J. Gen. Intern. Med.*, *26*(2), 192–196. Online 2010: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3019319/pdf/11606_2010_Article_1513.pdf