P-värdet får en annan tolkning i situationer där man väljer ut hypoteser och antalet möjliga statistiska test är stort.

Varför är subgruppsanalyser Och vad är ett

Varför är en subgruppsanalys mindre trovärdig? Jag tror jag har många statistiker med mig när jag säger att detta är en väldigt vanlig fråga. Ett resultat är ju exakt vad det är och inte mindre verkligt bara för att det inträffat i en oväntad subgrupp. Detsamma gäller p-värden. Varför är inte ett p-värde ett p-värde, oavsett var man beräknat det? Och vad menas med ”nominellt” p-värde. Enligt SAOL så betyder nominell: ”till namnet men inte i realiteten”, hur kopplar vi det till p-värden?

Veckans vinlotteri

Låt oss börja med ett enkelt exempel. Vad är sannolikheten att just DU vinner tre gånger i rad i veckans vinlotteri på jobbet (vinlotteri var nog ett typiskt 1990-talsfenomen, men kanske har du själv deltagit). Tänk att ni är 20 anställda och varje person köper en lott. Sannolikheten att just du skall vinna 3 veckor i rad är då 0,05 x 0,05 x 0,05 = 0,000125, det vill säga ungefär 1/10.000. Med andra ord detta kommer inte att hända – du får basera dig på egna vininköp också den närmaste tiden. En annan, men besläktad fråga är: Hur stor är möjligheten att NÅGON vinner 3 gånger i rad under det närmaste året? För att beräkna detta så använder vi sannolikheten ovan, men multiplicerar den med 20 (antal personer på arbetsplatsen och sedan med 49 (inom ett år betyder ju att sekvensen med tre vinster måste påbörjas inom 49 veckor). Då får vi sannolikheten 0,1225 eller 12,3 procent. Något som var extremt osannolikt (att du vinner de närmaste tre veckorna) förvandlas till något som kan hända med lite tur genom att vi släpper på förutsättningarna för beräkningarna.

Frågorna ovan är besläktade på samma sätt som följande frågor:

1. Vad är sannolikheten att primary endpoint blir statistiskt signifikant av en slump (dvs. p-värdet <0,05 även om effekten är likvärdig i behandlingsarmarna)?

2. Vad är sannolikheten att någon subgruppsanalys blir statistiskt signifikant av en slump (predefinierade och post hoc-analyser)?

Vad betyder p-värdet?

Ett p-värde uppfattas oftast som ett mått på om vi ska tro på studiens resultat och p-värdet anger ”risken att vi tar fel” – (läs gärna den tidigare artikeln om p-värden i Pharma Industry nr 4, 2015). För att ett p-värde ska betyda det vi vill, måste vi enkelt uttryckt begränsa oss till en primary endpoint, det vill säga ta en enda lott i vinlotteriet. Börjar vi testa multipla hypoteser blir situationen som ovan, det blir ganska sannolikt att någon av hypoteserna och testerna får ett p-värde som ligger under 0,05. Samtidigt betyder inte p-värdet längre det vi vill att det skall betyda: Att risken att vi observerar en så stor skillnad som vi gör av en ren slump är max 5 procent. Slumpen blir ”hjälpt av” att vi gör många test och sannolikheten (eller risken) för en slumpmässig signifikans är inte längre 5 procent. Ett ord som ofta används i diskussioner runt detta är ”multiplicitet” som förstås refererar till att det rör sig om många test utan att vi tar hänsyn till detta.

mindre trovärdiga? nominellt p-värde?

Nominellt p-värde

P-värden är väldigt etablerade i medicinsk forskning och presenteras en subgruppsanalys utan p-värde blir ofta första frågan: Vad är p-värdet? Rent tekniskt är det inte något problem att beräkna ett p-värde för vilken analys som helst, problemet är att ett beräknat p-värde inte betyder det vi vill att ett p-värde ska betyda på grund av multiplicitetsproblematiken. Någon väljer att komma runt detta genom att kalla det för nominellt p-värde, det vill säga beräknat på samma sätt men inte korrigerat för antalet möjliga test. Andra förhåller sig mer kallsinniga och anger inget p-värde, utan presenterar resultaten deskriptivt med skattningar och konfidensintervall. Det man som statistiker kan anmärka mot användandet av nominella p-värden är att många som refererar till dessa faktiskt inte vet skillnaden mellan ett riktigt p-värde och ett nominellt. För den oinvigde blir det lätt att fokusera på att det är ett p-värde utan att ta in betydelsen av att det är ett nominellt p-värde.

Är subgruppsanalyser suspekta?

Självklart är det inte så. Tvärt om, det är viktigt att ha möjligheten att undersöka vilka patientgrupper som profiterar på en ny behandling, om någon grupp reagerar oväntat, men också att göra oplanerade post-hoc analyser för att generera hypoteser inför nya kliniska studier. Forest-plottar som ofta presenteras i publikationer har just detta syfte; att se om det finns subgrupper som avviker. Problemet uppstår när subgruppsresultat, speciella i oväntade, oplanerade analyser, upphöjs till ”sanning”. Förplanerade subgruppsanalyser som har en tydlig medicinsk rational och där andra supporterade analyser visar på samma effekter kan givetvis vara trovärdiga, både medicinskt och statistiskt. Även oplanerade subgruppsanalyser kan vara trovärdiga om man har resultat från flera oberoende studier som pekar åt samma håll. Då bekräftar dessa analyser varandra indirekt, något som kan bli väldigt trovärdigt, om det inte också finns resultat som pekar i motsatt riktning.

Hur hanterar man multiplicitetsproblematik?

Det enklaste är förstås att förhålla sig till primary endpoint, men då går man ju eventuellt miste om värdefulla resultat. Ett alternativ är att korrigera p-värden genom att multiplicera dem med en faktor som kompenserar för antalet test. För en oplanerad post hoc-analys finns det inget specificerat antal, det presenterade resultatet är en av många möjliga analyser som kanske gjorts, eller kanske inte. För planerade endpoints, primära och sekundära, kan man använda hierarkisk testning, det vill säga man ordnar de planerade testen i en hierarki och fortsätter beräkna p-värden neråt i hierarkin tills ett test blir icke-signifikant, då upphör formell testning. Det finns en enkel metod som tar höjd för alla möjliga situationer och det är kritiskt tänkande.

Leukemi vid kraftledningar

Problematiken visar sig exempelvis i diskussionen om ökad risk för leukemi vid kraftledningar och oväntade fynd i epidemiologiska studier. Det är svårt att veta om man verkligen observerar en överrisk för cancer, när slutsatsen baseras på observationer där sannolikheten att det vi ser är av en ren slump. Hur många kraftledningar finns det där vi skulle ha kunnat observera flera cancerfall? Detsamma gäller epidemiologiska studier där man inte har prespecificerade analyser på samma sätt som i kliniska studier. Det intressanta fyndet som publiceras är en möjlig analys av ett stort antal analyser och vi kan inte bedöma trovärdigheten endast baserat på ett p-värde. Speciellt inte som p-värdet inte har den strikta betydelsen vi vill att ett p-värde ska ha. Det finns otaliga intressanta fynd i subgrupper och epidemiologiska studier, som senare inte har kunnat replikeras och därför avskrivits som Typ-I-fel, det vill säga att analysen visade på en statistisk signifikant skillnad av en slump.

ANNA TÖRNER
statistiker, verkställande direktör,
Scandinavian Development Services