HYPOTESPRÖVNING. (TESTTEORI.)

Oscar har varit på en spelhåla och förlorat pengar. Med sig från spelhålan har han en tärning, som han misstänker har för stor sannolikhet att visa sexa. För att undersöka saken bestämmer han sig för att kasta tärningen 180 gånger och se hur många sexor det blir. Låt oss beteckna antalet sexor han då får med X. Blir X "för stort" tänker Oscar gå tillbaka till spelhålan och anklaga den för falskspel.

Vad är då "för stort"? Om tärningen är symmetrisk så bör ungefär 1/6 av kasten ge en sexa, dvs man bör få ungefär 30 sexor. Man inser kanske direkt att Oscar knappast tänker bråka om han får 31 sexor. Inte heller 32 bör föranleda Oscar att bråka. Fast hur är det om han får t ex 39 sexor? Bör han bråka då?

För att kunna utreda dethär måste vi nalkas det hela formellt med hjälp av teorin för hypotesprövning. Låt därför p beteckna tärningens sannolikhet att visa sexa. (p är ett okänt tal som vi misstänker är större än 1/6.)

Oscar ställer nu upp hypoteserna H_0:p=1/6 och H_1:p>1/6. Sedan behöver han en s k testvariabel. Han väljer X: antalet kast (av de 180) som ger en sexa. Om X >C tänker Oscar förkasta H_0: p=1/6 och gå tillbaka och bråka.

Hur ska nu C bestämmas? Ja, det beror på de risker som är förenade med att bråka med spelhålan i de fall då det inte är något fel på tärningen. Säg att Oscar är villig att ta (högst) 1% risk för detta. I så fall ska vi bestämma C så att P(X>C)=1% när H_0 är sann. Att H_0 är sann betyder att p=1/6 och om kasten sker oberoende av varandra har vi standardsituationen för binomialfördelning dvs X är Bin (180,1/6). Då vi inte har någon tabell över binomialfördelningen för dessa parametervärden måste vi approximera fördelningen. Enligt FS 6 gäller att Bin(180,1/6) $\approx$ N(30,5).

Om man nu normerar testvariabeln får vi P(Z>(C-30)/5)=1% där Z=(X-30)/5 är $\approx$ N(0,1). Enligt tabell 2 i FS får man (jämför med den övre figuren bredvid tabellen):
(C-30)/5=2.3263. Detta ger C$\approx$ 41.6 dvs (eftersom C måste vara ett heltal) C=42.

Får Oscar 42 eller fler sexor ska han alltså gå tillbaka och bråka. Den risk (=testets FELRISK!) han då tar att felaktigt anklaga spelhålan för att använda falska tärningar är (högst) 1%. (Andra namn på felrisk är "signifikansnivå"eller "risknivå".)

(Observera att ju större värde på C vi väljer, ju mindre risk tar vi för att komma med en felaktig anklagelse. Men väljer vi C stort minskar å andra sidan vår sannolikhet för att upptäcka om vi blivit lurade!)

Den s k styrkefunktionen har lite karaktär av "överkurs", men är samtidigt ett bra sätt att förstå hur ett test verkligen fungerar. (Särskilt svårt är det inte.) Styrkefunktionen S(p) är sannolikheten att förkasta H_0 som funktion av det verkliga värdet på p. Med hjälp av N-approximationen ovan bestämmer vi lätt S(p).

För att kunna bestämma S(p) måste man veta testets felrisk. Låt oss anta att Oscar (som ovan) valt felrisken till 1 %. Vi får då S(p)= P(förkasta H_0)=
P(X>41.6)= P((X-180p)/$\sqrt{\text{180p(1-p)}}$ > (41.6-180p)/$\sqrt{\text{180p(1-p)}}$) $\approx 1-\phi(\frac{41.6-180p}{\sqrt{180p(1-p)}})$

Ritar du upp denna funktion (med hjälp av tabell 1) ser du att den har sitt minimum (=felrisken!) för p=1/6 och är växande för p>1/6. Sannolikheten att förkasta H_0 är alltså minst då H_0 är sann och allt större "ju falskare tärningen är". Dvs testet fungerar precis som man vill att det ska. (Att S(p) är mindre än 1% för p<1/6 behöver man inte bry sig om eftersom man redan vid uppställandet av H_0 och H_1 gjorde klart att man ansåg det uteslutet att p<1/6 gällde.)

Testvariabeln ska väljas så att den får en känd fördelning när H_0 är sann. I exemplet ovan var detta inget problem, åtminstone inte om man kände till binomialfördelningen. Nu följer ett mer komplicerat exempel där valet av testvariabel inte är riktigt lika självklart.

Två legeringar 1. och 2. skulle jämföras med avseende på smältpunkt. Låt oss kalla de verkliga värdena på smältpunkterna för μ_1 och μ_2. (μ_1 och μ_2 är okända konstanter.) Man har mätt smältpunkterna för de bägge legeringarna 5 resp. 10 gånger. Mätfelen antas vara oberoende och N(0,σ)-fördelade och eftersom de uppmätta värdena kan ses som summan av det verkliga värdet på smältpunkten och mätfelet, får vi följande modell: $X_1, X_2 \cdots X_5$ är N(μ_1, σ) och $Y_1, Y_2 \cdots Y_{10}$ är N(μ_2, σ). (Samtliga 15 variabler oberoende.) Data: $\sum_{i=1}^{5} x_i =$ 5260.1 $\sum_{i=1}^{5} x_i^2 =$ 5 533 744.2 $\sum_{i=1}^{10} y_i = 10509.2$ $\sum_{i=1}^{10} y_i^2 =$ 11 044 365.1.

Man misstänker att μ_1 är större än μ_2 dvs att $\mu_1 - \mu_2 > 0$ gäller och vill testa detta. Vi väljer därför $\mu_1 - \mu_2 > 0$ som H_1. Som H_0 skulle man nu kunna välja $\mu_1 - \mu_2 \leq 0$, men eftersom man inte alls misstänker att $\mu_1 - \mu_2 < 0$ gäller nöjer man sig med $H_0 : \mu_1 - \mu_2 = 0$. Eftersom $\bar{X} - \bar{Y}$ skattar $\mu_1 - \mu_2$ bör testet grunda sig på denna skattning. För att få en känd fördelning väljer vi T=$\frac{\bar{X}-\bar{Y}}{S\sqrt{(1/5+1/10)}}$ som testvariabel.(Enligt FS 11.2d har T ju en t(13) fördelning om H_0 är sann.) Om H_1 gäller dvs om $\mu_1 - \mu_2 > 0$ så kommer $\bar{X} - \bar{Y}$ att tendera att bli >0 dvs testets utseende blir: T>C $\Rightarrow$ H_0 förkastas. Återstår att bestämma C. Med felrisken (t ex) = 5% får vi 5 % =P(förkasta H_0 när H_0 är sann) dvs 5 % = P(T>C) där T är t(13) fördelat. Tabell 3 i FS ger C=1.77. Slutligen utförs testet. Från våra data ovan beräknar vi $T_{obs} = 1.02$. Eftersom detta är mindre än C förkastas H_0 ej. (Detta betyder att data inte motsäger att $\mu_1 = \mu_2$ är vad som gäller.)

Låt oss nu behålla allting i testet ovan oförändrat (inklusive $H_0 : \mu_1 - \mu_2 = 0$), men ändra H_1. Lägg noga märke till hur testets utseende påverkas av denna ändring. Har vi t ex $H_1 : \mu_1 - \mu_2 < 0$ istället så blir testet: T< C $\Rightarrow$ H_0 förkastas. Och skulle vi ha $H_1 : \mu_1 - \mu_2 \neq 0$ så får vi testet: $\mid T \mid >$ C $\Rightarrow$ H_0 förkastas.

I det sista fallet hade vi också kunnat använda den s k konfidensmetoden för att testa. Metoden innebär här att man bildar ett 95 % igt konfidensintervall för $\mu_1 - \mu_2$ och förkastar H_0 om 0 ej tillhör intervallet. Som alltid kan man bara uttala sig om H_0 och det enda man kan säga är "förkasta H_0" eller "ej förkasta H_0".

Man väljer vidare H_0 och H_1 så att det man uppfattar som det farligaste felet får felrisken som sannolikhet. (Felrisken bestämmer man ju själv.) Ett bra exempel på detta:

Om ölflaskor som är märkta med "max 3.5 %" innehåller mer än 3.5 % kan bryggeriet råka ut för rättsliga åtgärder. Man vet att alkoholhalten i en slumpvis uttagen flaska är N(μ, 0.10). (Observera att vi räknar med % som sort.) Vi har uppmätt alkoholhalten i 50 flaskor oberoende av varandra, dvs vår modell blir $X_1, X_2 \cdots X_{50}$ oberoende N(μ, 0.10).

Vill man testa alkoholhalten och väljer $H_0 : \mu >$3.5 och $H_1 : \mu \leq$3.5 avspeglar detta att man ser allvarligast på risken att råka ut för rättsliga åtgärder. (Att välja H_0 och H_1 tvärtom betyder att man är mera mån om att kunden inte ska få för svagt öl!) Genom en lite allmännare definition av felrisken (som vi inte behöver gå in på) kan man visa att vi kan ändra $H_0 : \mu > 3.5$ till $H_0 : \mu = 3.5$ och utgå från det istället utan att testet förändras. Som testvariabel väljer vi T= $\frac{\bar{X}-3.5}{0.10/\sqrt{50}}$. När H_0 är sann (dvs när μ=3.5 gäller) har T en N(0,1) fördelning. Av utseendet på H_1 ser man att testet blir $T \leq C \Rightarrow H_0$ förkastas. Välj nu felrisk = 1 promille och visa att $\bar{x} = 3.46$ då medför att H_0 ej förkastas. (Jfr tal 13.20).

När man utför ett test med hjälp av ett dataprogram används ofta den s k P-metoden. Antag t ex att Oscar fick 39 sexor. Datorn beräknar då (exakt!) att sannolikheten för ett så stort (eller större!) antal sexor är 4.807 % när H_0 är sann. Detta betyder att om felrisken väljs till 5 % så förkastas H_0, men om felrisken istället väljs till 1% så förkastas H_0 inte.