Statistiska modeller och inferens

"Artificial intelligence is no substitute for the real thing"
Robert Gentleman

Innehåll

1 Inledning

En modell är ett generellt begrepp som betyder olika saker i olika sammanhang. I allmänhet beskriver en modell något fenomen, t. ex. ett samband mellan två kvantiteter, så som förekomsten av hudcancer i en population och den mängd UV-strålning individerna i populationen tenderar att utsätta sig för under en viss tidsperiod. Men modeller kan även beskriva informationsflödet i en organisation eller beräkningsstegen i en datoralgoritm.

Vi har redan konstaterat att statistik är vetenskapen om hantering av osäkerhet i diverse former. Därmed är det föga förvånande att vi främst är intresserade av stokastiska modeller inom statistiken (vi kallar sådana beskrivningar statistiska modeller fortsättningsvis). Vad är då stokastiska (eller statistiska) modeller för något? Man kan exempelvis konstatera att de representerar motsatsen till deterministiska modeller som läsaren förmodligen redan tidigare har stött på i skolfysiken. Deterministiska modeller beskriver generellt exakta samband mellan olika storheter (t. ex. massan och accelerationen). Statistiska modeller beskriver i sin tur nivåer av osäkerhet som förknippas med de storheter vi är intresserade av, givet den mängd information som finns tillgänglig för oss.

Det är viktigt att förstå statistiska modellers ursprung - **vi** bygger dem för diverse behov! Modeller ploppar helt enkelt ej ur intet, utan vi formar dem enligt den förståelse vi har för den osäkerhet som finns inblandad i en viss situation. Därmed finns det allt från bra till dåliga modeller, helt beroende på vår förmåga att beskriva osäkerheten på ett ändamålsenligt sätt.

Varför behövs då statistiska modeller? I många sammanhang måste vi uppskatta mängden av något (t. ex. bly i svampar) någonstans (t. ex. skogarna i Söderkulla i Sibbo) utan att ha tillgång till exakt information. Det kan vara för dyrt, tidskrävande, eller rentav omöjligt att skaffa sig exakt information om det vi är intresserade av. Enkätundersökningar gällande attityder, åsikter, köpesvanor mm. hos olika grupper av människor, representerar typiska exempel på sådana situationer. Men statistiska modeller duger till mycket mer än enbart till att uppskatta det som redan finns. En viktig aspekt av dylika modeller är nämnligen förmågan att skapa prognoser och kvantifiera osäkerhet hos något vi ej kan observera. Detta kan motsvara en prognos om framtiden, exempelvis inflationsnivån under nästa budgetår eller den genomsnittliga sommartemperaturen i Medelhavsregionen under nästa decennium, men även mängden solenergi en solpanel kan absorbera om ytbeläggningen modifieras på ett visst sätt.

Vad är då inferens för något? Statistisk inferens betyder allmänt de formella matematiska procedurerna vi använder för att dra slutsatser på basen av ett empiriskt datamaterial. Slutsatserna handlar i detta sammanhang alltsom oftast om diverse egenskaper hos de statistiska modeller vi plockat fram för tillämpningen i fråga. Ofta kallas sådana egenskaper *parametrar* och de bestämmer hur en modell beskriver osäkerheten. Somliga kallar själva slutsatserna om parametrarna inferens. Det väsentliga är dock att vi strävar efter ett logiskt eller rationellt beteende hos den metod vi använder för att dra slutsatserna. Det är viktigt att förstå att statistiska procedurer ej alltid beter sig rationellt, speciellt då vi tillämpar dem på något som de ursprungligen ej är tänkta för!

I denna text behandlas särskilt problematiken kring hypotesprövning och jämförelse av olika alternativa modeller. Modellbaserade prognoser och skattning av parametrar, samt konfidensintervall för dem betraktas senare, i ett separat material.

2 Hypotesprövning

2.1 Allmänt on hypotesprövning

Bakom det statistiska konceptet **hypotesprövning**, ligger en stark vetenskapsfilosofisk tradition gällande falsifieringsprincipen och den logiska positivismen, där stora namn som Karl Popper, Rudol Carnap m. fl. har varit aktiva. Utifrån statistikens synvinkel, har hypotesprövningen intimt kopplats till den frekventistiska inferenstraditionen, ävenom man visst kan pröva hypoteser med andra angreppssätt. Vi försöker ändå beskriva hypotesprövningen i den frekventistiska kontexten och betrakta dess goda och aviga sidor.

Generellt i litteraturen brukar man använda H_0 för att beteckna en sk nollhypotes och H_1 för att betrakta en sk mothypotes. Förenklat motsvarar en viss hypotes i statistikens värld en viss statistisk modell för empiriska observationer, dvs. hypotesen hävdar att observationerna är fördelade enligt någon sannolihetsfördelning. Då vi arbetar utifrån falsifieringsprincipen, strävar vi efter att utesluta teorier (dvs. nollhypoteser H_0) genom att falsifiera dem med hjälp av empiriska observationer.

2.2 Den generella proceduren för hypotesprövning

Själva hypotesprövningen, som ibland kallas ett statistiskt test, utförs i allmänhet enligt följande formella procedur (där ordningen på de fyra första stegen visserligen kan variera):

1. Bestäm nollhypotesen H_0 och mothypotesen H_1.
2. Observera ett datamaterial $\mathbf{x}$.
3. Bestäm en lämplig signifikansnivå α ($0 \leq \alpha \leq 1$).
4. Bestäm en lämplig testkvantitet (en sk teststatistika) T, för vilken ett observerat värde T_{obs} erhålls från $\mathbf{x}$.
5. Beräkna sannolikheten $P(T \geq T_{obs}|H_0)$, vilken motsvarar händelsen att ett lika eller mer extremt värde på testkvantiteten skulle observeras, givet att ett lika stort datamaterial skulle genereras under H_0.
6. Förkasta H_0, ifall $P(T \geq T_{obs}|H_0) < \alpha$.

Steget 5 ovan ger oss något som allmänt kallas ett **p-värde**, som motsvarar den lägsta signifikansnivån på vilken H_0 kan förkastas för det observerade datamaterialet $\mathbf{x}$. Det är i en viss mening ett mått på hur överraskande vårt datamaterial är, ifall nollhypotesen är sann. Det är alltså ingalunda sannolikheten för att nollhypotesen är sann, vilken är den vanligaste missuppfattningen om p-värdet! Ett litet p-värde tyder alltså på att våra data är mindre rimliga under nollhypotesen. Vanligen förekommande signifikansnivåer i vetenskapliga arbeten är 5%, 1% och 0.1%.

Observera att nollhypotesen har en väldigt speciell roll i den frekventistiska hypotesprövningen. Man är alltså ute efter att förkasta den, eftersom proceduren **aldrig** kan ge ett direkt mått på **stöd** för en viss hypotes. Vi diskuterar senare diverse problem förknippade med hypotesprövning, nu är det dags att titta på kanske den oftast förekommande jämförelsen av två modeller. En rätt bra beskrivning av p-värden hittar man på `http://www.wikipedia.org/P_value`. Sidan innehåller även länkar till ytterligare material som skapats för pedagogiska syften. Den artikel i British Medical Journal (Sifting the evidence - What's wrong with significance tests") som citeras av Wikipedia-sidan, är mycket läsvärd.

Example 1 ***Jämförelse av medelvärden för två populationer****. Anta att vi är intresserade av hur två olika typers choklad påverkar mängden kolesterol i blodådrorna. Den ena typen är vanlig mjölkchoklad (innehåller ca 30% kakao) och den andra typen är mörk choklad med hög kakaohalt (86%). Vissa ämnen i kakao spekuleras påverka cellernas ämnesomsättning, vilket eventuellt kunde leda till en förändring gällande kolesterolhalterna (och kanske även kolesterolens relativa sammansättning). Ett antal försökspersoner fördelas slumpmässigt i två grupper, varav den ena (gruppen A) inmundigar dagligen 50 gram vanlig mjölkchoklad och den andra (gruppen B) samma mängd mörk choklad. Efter den tre månader långa uppföljningsperioden mäter man kolesterolhalt (X) hos alla individerna i bägge grupperna.*

Example 2 ***Jämförelse av medelvärden för två populationer (fortsättning)****. Låt oss anta att kolesterolhalterna är fördelade enligt en normalfördelning i den population som representeras av försökspersonerna, så att* $X \sim N(\mu_A, \sigma^2)$ *för dem som inmundigar mjölkchoklad och* $X \sim N(\mu_B, \sigma^2)$ *för dem som inmundigar mörk choklad. Här antar vi alltså att spridningen (som representeras av variansen* σ^2*) är densamma i båda grupperna, men att de genomsnittliga kolesterolhalterna kan vara olika i A och B. Nollhypotesen och mothypotesen formuleras då oftast enligt följande:* $H_0 : \mu_A = \mu_B, H_1 : \mu_A \neq \mu_B$*. Logiken bakom detta resonemang är att datamaterialet skall innehålla motstridig information med avseende på påståendet* $\mu_A = \mu_B$*, innan vi kan hävda att den mörka chokladens effekt gällande kolesterolhalten är annorlunda än mjölkchokladens.*

I exemplet ovan är nollhypotesen H_0 en sk **enkel** (eller **skarp**) hypotes och H_1 motsvarar något som kallas en **sammansatt** hypotes. Det specifika hos en enkel hypotes är att den bestämmer fullständigt värdet på den parameter (eller de parametrar) vi främst är intresserade av. Om vi tittar riktigt noga på exemplet, så ser vi att nollhypotesen i detta fall egentligen bestämmer entydigt värdet på **skillnaden** mellan μ_A och μ_B, dvs. att den skillnaden är noll.

En sammansatt hypotes däremot omfattar ett område av värden som anses samstämmiga med hypotesen. I chokladexemplet är vi i första hand intresserade av medelvärdenas eventuella skillnad, men inte av variansen, ävenom vi måste ta hänsyn till den i modellen. Den är ju trots allt okänd för oss, och måste därmed uppskattas från det observerade datamaterialet!

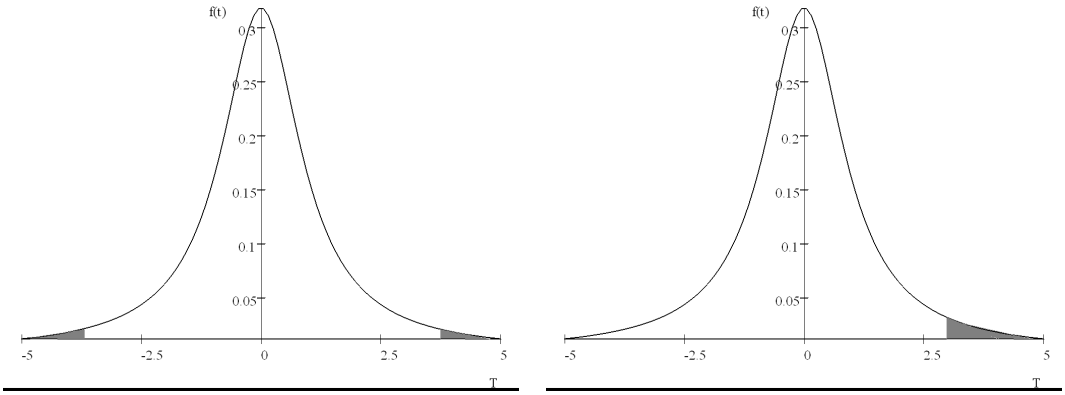
Variansen kallas ibland en **störande** parameter, då vi inte är intresserade av att jämföra skillnader i spridning mellan två eller fler grupper. Generellt

betraktas sådana parametrar vi inte är direkt intresserade av, men som ändå är nödvändiga att sättas in i modellen, som störande parametrar. Dylika parametrar måste tas i beaktande på ett lämpligt sätt då vi utför statistisk inferens, men vi skall ej här gå närmare in på dem.

Hypoteserna givna ovan i samband med chokladstudien leder till ett **dubbelsidigt** test, eftersom vi inte har bestämt i förväg om skillnaden mellan μ_A och μ_B bör vara negativ eller positiv under mothypotesen. I en annan situation kunde den underliggande biologin leda oss t. ex. till en hypotes om att både medelvärdet och variationen gällande kolesterolhalten sjunker, då man inmundigar mörk choklad regelbundet. Denna mer specifika biologiska kunskap kunde representeras av hypoteserna: $H_0 : \mu_A = \mu_B, \sigma_A^2 = \sigma_B^2$, $H_1 : \mu_A > \mu_B, \sigma_A^2 > \sigma_B^2$. De test som motvarar dylika hypoteser kallas **enkelsidiga**, eftersom avvikelser endast i en viss riktning från nollhypotesen kan leda till förkastning.

Signifikans hos dubbelsidiga och enkelsidiga test bestäms på olika sätt, vilket illustreras i bilderna nedan. I bilden till vänster betraktar man ett dubbelsidigt test, där ett observerat värde på testkvantiteten, som faller inom det gråa området, leder till ett signifikant resultat på nivån α (t. ex. 5%). Kurvan i bilden representerar hur sannolikhetsmassan fördelar sig för testkvantiteten under nollhypotesen och bägge gråa områden har en massa som motsvarar sannolikheten $\alpha/2$ (t. ex. 2.5%). Detta sätt att beräkna signifikansen beror på att avvikelser i båda riktiningar från nollhypotesen betraktas relevanta i sammanhanget. Logiken bakom hypotesprövningen säger att då en testkvantitets värde är inom det gråa området, gäller antingen: 1) nollhypotesen är sann och en osannolik händelse har inträffat, eller 2) nollhypotesen är ej sann. Vanligtvis väljer man då det senare alternativet och förkastar därmed nollhypotesen.

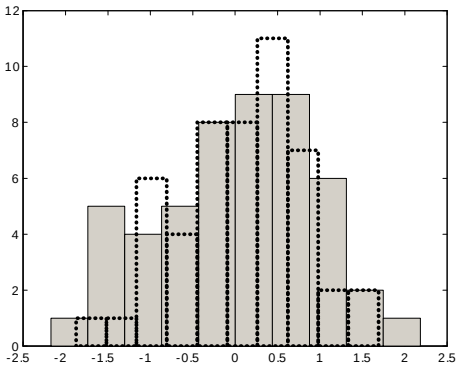
I bilden till höger visar det gråa området signifikanta värden på ett enkelsidigt test, där endast avvikelser i en viss riktning betraktas relevanta. Notera att ett mindre värde på testkvantiteten T i absolutbelopp leder nu till ett signifikant resultat, jämfört med det dubbelsidiga testet. Detta beror på att hypoteserna är nu formulerade så att skillnader i en viss riktning anses irrelevanta, och därmed leder ej till signifikanta testresultat (oavsett hur stor skillnaden är).



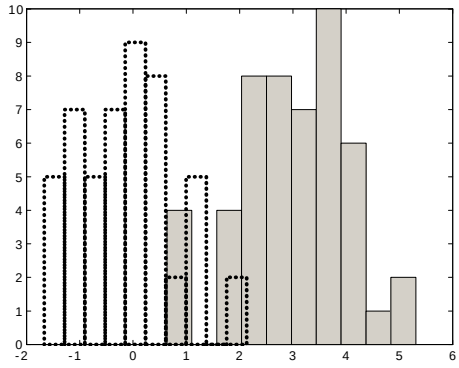
Example 3 ***Jämförelse av medelvärden för två populationer (fortsättning)****. Låt x_{ij} representera mätvärdet på kolesterolhalten för individen i hos gruppen j (alltså $j = A$ eller $j = B$). Vi betecknar antalet försökspersoner i de två grupperna med n_A respektive n_B. Den matematiska proceduren för hypotesprövningen i chokladexemplet består av följande element. 1. Beräkna medelvärden $\bar{x}_A, \bar{x}_B$ och varianserna s_A^2, s_B^2 för grupperna och bestäm den sk poolade (dvs. sammansatta) variansen $s_{A,B}^2 = \frac{(n_A-1)s_A^2+(n_A-1)s_A^2}{n_A+n_B-2}$. 2. Beräkna värdet på testkvantiteten $T_{obs} = \frac{|\bar{x}_A-\bar{x}_B|}{\sqrt{s_{A,B}^2(\frac{1}{n_A}+\frac{1}{n_B})}}$. Obs! $|x|$ betecknar här ett absolutbelopp av x, dvs. talets storlek i förhållande till 0. 3. Förkasta H_0 om $P(T \geq T_{obs}) < \alpha$. Detta är ett dubbelsidigt test, eftersom både positiva ($\bar{x}_A > \bar{x}_B$) och negativa ($\bar{x}_A < \bar{x}_B$) skillnader mellan medelvärden betraktas, då man avgör signifikans.*

Example 4 ***Jämförelse av medelvärden för två populationer (fortsättning)****. Testproceduren som presenterades ovan är ett sk t-test för medelvärdena hos två populationer, då vi erhållit oberoende stickprov från dem och variansen antas lika i bägge populationerna. Namnet t-test syftar till testkvantitetens fördelning, som heter t-fördelning. För denna testsituation kan p-värdet bestämmas enkelt och testkvantitetens form skvallrar om att det beror på avståndet mellan gruppernas medelvärden i förhållande till nivån av spridning. Testkvantiteten är lika med noll endast om medelvärdena är exakt lika hos A och B. Bilderna nedan representerar två olika situationer med 50 observationer i bägge grupperna. I den vänstra bilden är histogrammen ritade för ett datamaterial som härstammar från en och samma fördelning (N(0,1), A:*

gråa stolpar, B: stolpar utan färg). I den högra bilden motsvarar histogrammet utan färg datamaterial från fördelningen N(0,1) och det andra histogrammet datamaterial från fördelningen N(3,1). Ett t-test på nivån 5% skulle ej förkasta H_0 för datat i fallet 1, men nog för datat i fallet 2. Gränsvärdet gällande signifikans skulle i detta fall vara ca 1.98 för testkvantiteten.



$1 : H_0$ sann ($\mu_A = \mu_B$)



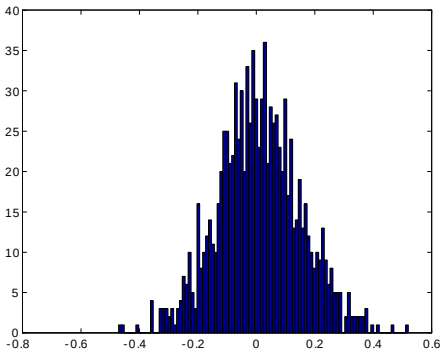
$2 : H_1$ sann ($|\mu_A - \mu_B| = 3$)

2.3 Vissa egenskaper hos testproceduren

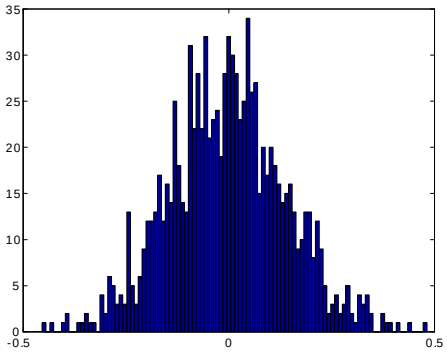
Vad kan då gå åt pipan med den testprocedur vi skissade precis? Uppenbarligen, det kan hända att nollhypotesen trots allt gäller, ävenom vi förkastat den på signifikansnivån α (t. ex. 5%). Proceduren utmynnas i det att vi reglerar α, vilket avgör hur troligt det är att begå ett fel av **första slaget** (Type I error), dvs. felaktigt förkasta nollhypotesen. Denna sannolikhet är förknippad med tanken om replikering. Den frekventistiska grundtanken bakom hypotesprövning är att kontrollera sannolikheter för felaktiga slutsatser, då vi upprepade gånger plockar fram datamaterial som antas bete sig enligt sannolikhetsfördelningen given av den ena eller den andra hypotesen.

Testets beteende kan studeras med hjälp av en datorsimulering. Vi slumpar först fram två oberoende datamaterial med varsitt set på 50 observationer från en N(0,1) fördelning, beräknar sedan värdet på testkvantiteten, och avgör till sist om H_0 skall förkastas på nivån 5% ($T_{obs} > 1.98$) eller ej ($T_{obs} \leq 1.98$). Proceduren upprepas 1000 gånger för att vi skall se hur ofta vi gör en felaktig slutsats. Bilderna nedan visar hur medelvärden i de två

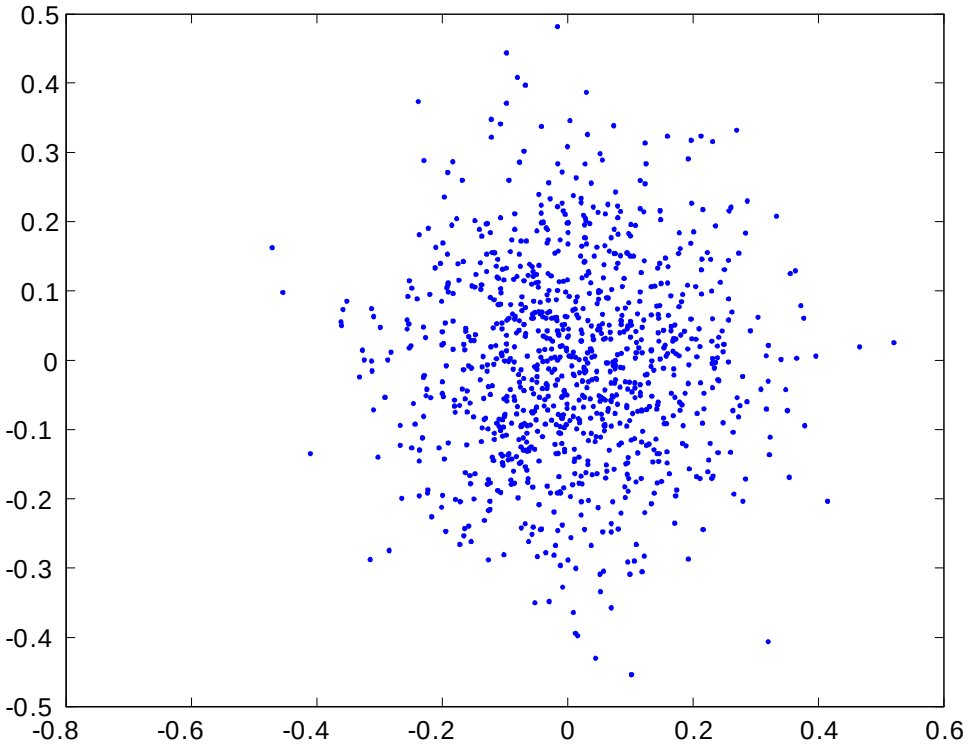
datamaterialen varierar över de 1000 upprepningarna, samt hur deras skillnad beter sig. I denna simulering förkastas H_0 49 gånger av 1000 försök, dvs den empiriska frekvensen för ett fel av första slaget är 4.9%. Testet verkar alltså hålla vad det lovar. Genom att minska på α, skulle den förväntade empririska frekvensen av dylika fel sjunka vidare. Nackdelen blir då att alltmer data krävs för att riktiga skillnader skall kunna påvisas.



Fördelningen för $\bar{x}_A$



Fördelningen för $\bar{x}_B$



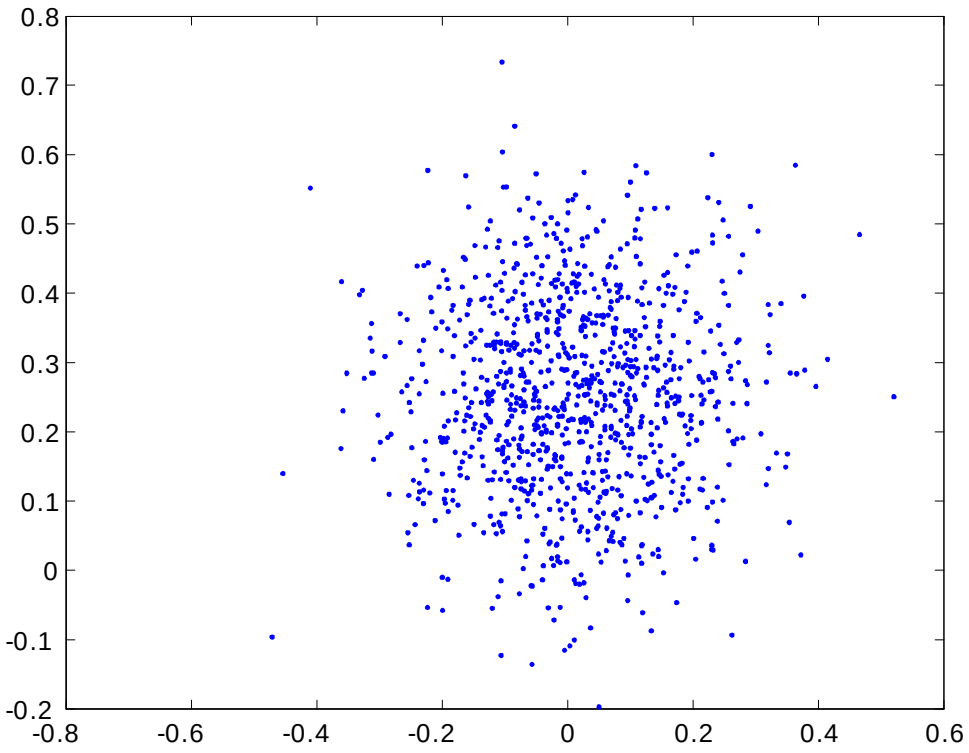
Spridningsdiagrammet för värdena $\bar{x}_A$ mot $\bar{x}_B$ då H_0 är sann ($\mu_A = \mu_B$).

Vad annat kan gå åt pipan med den testprocedur vi skissade ovan? Ja, det kan hända att vi accepterar nollhypotesen, men i verkligheten finns det en skillnad mellan medelvärdena μ_A och μ_B. Detta kallas för fel av andra slaget (Type II error) och sannolikheten för att vi blir utsatta för det beror på testet **styrka**, dvs. förmågan att upptäcka om H_0 ej är sann. **Styrkan beror på: 1) hur stor skillnaden mellan grupperna är, 2) spridningsnivån i populationen och 3) på datamaterialets storlek, dvs. antalet observationer**.

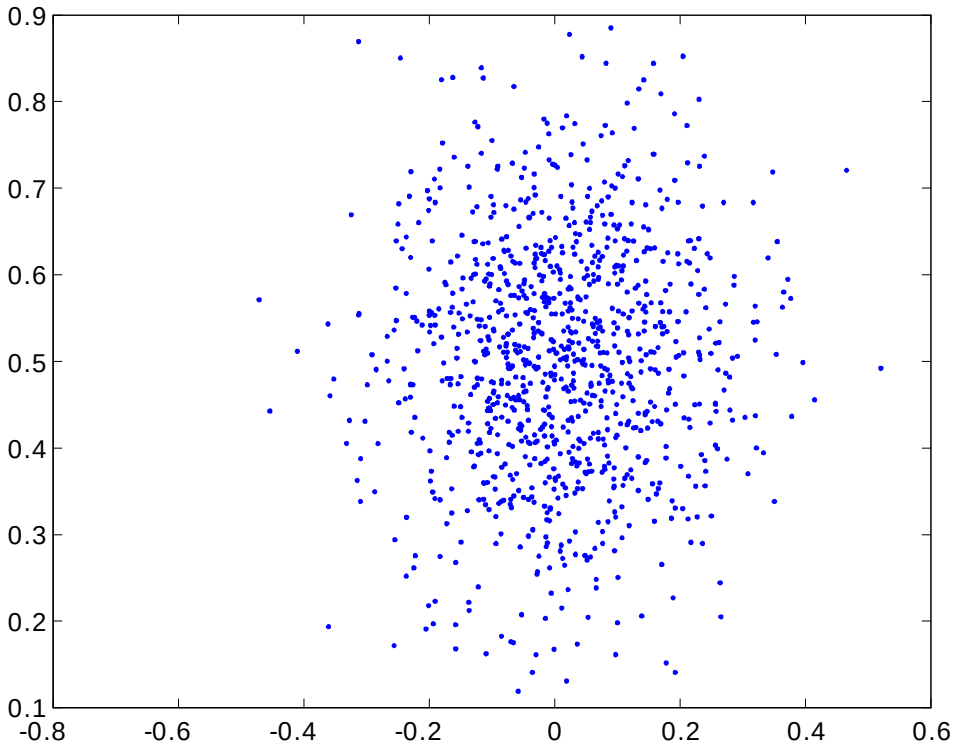
Låt oss studera styrkan med hjälp av en dylik datorsimulering som användes tidigare för fel av första slaget. För en uppsättning av värden på μ_A och μ_B, slumpar vi 1000 gånger fram 50 observationer från motsvarande fördelningar, och kontrollerar hur ofta testet upptäcker den underliggande skillnaden mellan populationerna. Bilderna nedan visar spridningsdiagrammen

för medelvärden $\bar{x}_A, \bar{x}_B$ över tre olika situationer: $\mu_A = 0, \mu_B = 0.25; \mu_A = 0, \mu_B = 0.50; \mu_A = 0, \mu_B = 3.0$. I samtliga fall är variansen lika med 1 hos båda populationerna. Ett t-test på nivån 5% förkastar H_0 enligt följande tabell över de 1000 replikationerna:

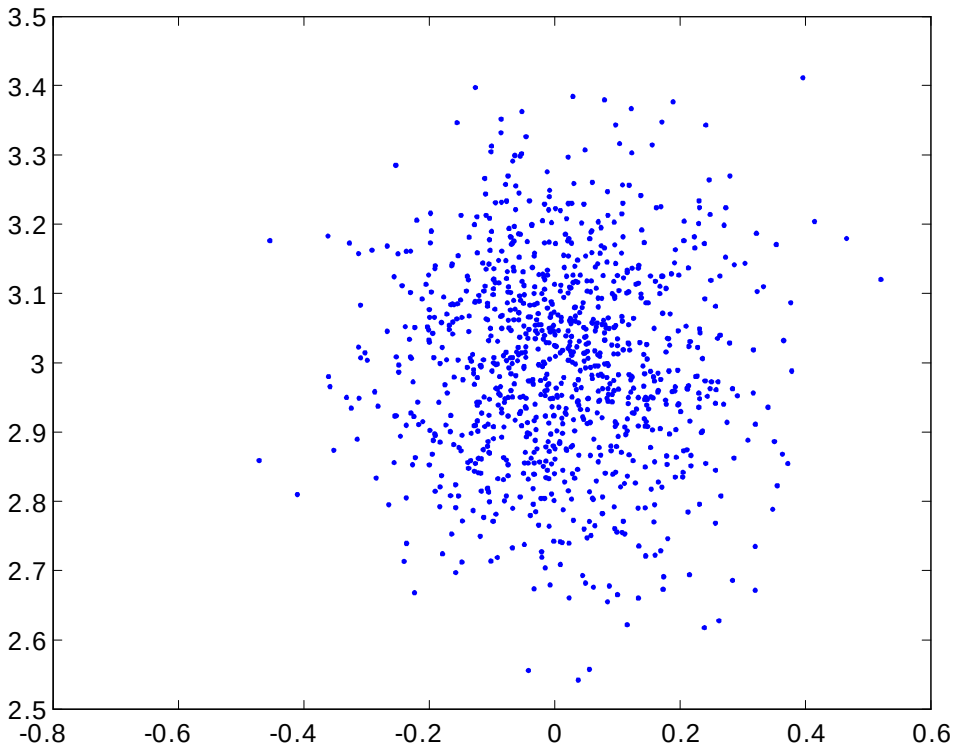
	Antalet accepterade H_0	Antalet förkastade H_0
$\mu_A = 0, \mu_B = 0.25$:	759	241
$\mu_A = 0, \mu_B = 0.50$:	312	688
$\mu_A = 0, \mu_B = 3.0$	0	1000



Spridningsdiagram för värdena $\bar{x}_A$ mot $\bar{x}_B$, då $\mu_A = 0, \mu_B = 0.25$



Spridningsdiagram för värdena $\bar{x}_A$ mot $\bar{x}_B$, då $\mu_A = 0, \mu_B = 0.50$



Spridningsdiagram för värdena $\bar{x}_A$ mot $\bar{x}_B$, då $\mu_A = 0, \mu_B = 3.0$

En sammanfattning av olika beslut i samband med en hypotesprövning och deras konsekvenser finns i tabellen nedan.

Verklighet\Beslut	Förkastar H_0	Accepterar H_0
H_0 är sann	Fel av första slaget (Typ I)	:)
H_1 är sann	:)	Fel av andra slaget (Typ II)

En testprocedur är allmänt utformad enligt följande logik. Fel av Typ I kontrolleras först med hjälp av signifikansnivån α. Genom att välja ett litet tal ser man till att dylika felaktiga beslut fattas mer sällan **under den antagna modellen**. Sedan väljer man en testkvantitet som maximerar styrkan, dvs. minimerar sannolikheten för beslut som leder till fel av Typ II.

2.4 Kritisk användning av hypotesprövning

I den statistiska litteraturen beskrivs ett enormt antal olika statistiska test och de vanligaste testen finns alltid inprogrammerade i statistiska programpaket. Man kan t. ex. testa signifikans för korrelations- eller regressionskoefficienter, eller för samband mellan kategoriska variabler. Det viktiga i detta sammanhang är att veta antaganden som ligger bakom ett test och för vilka typer av situationer det lämpar sig för. En lista över ca 10 vanliga test med kort beskrivning av de underliggande antaganden finns på `http://www.wikipedia.org/Hypothesis_test`. Exempelvis kunde antagandet om normalfördelning vara i sin helhet en grov missuppfattning, eller så kunde datamaterialet innehålla några avvikande observationer vars läge och/eller spridning är avsevärt annorlunda jämfört med de övriga observationerna. **Det är ytterst viktigt att försöka kontrollera antaganden vid användning av statistiska test, eftersom resultaten annars kan vara missvisande åt det ena eller det andra hållet!** Grafiska presentationer är mycket användbara för detta syfte.

Hypotesprövning, enligt den procedur vi ägnat uppmärksamhet åt hittills, kan kriticeras utifrån ett vetenskapsfilosofiskt perspektiv. Det mest fundamentala problemet är att man ej kan presentera grader av evidens för en hypotes mot en annan hypotes (eller mot en uppsättning alternativa hypoteser) med hjälp av hypotesprövning. Falsifieringsprincipen haltar i den bemärkelsen att den ignorerar de sannoliksmässiga **prognoser** som statistiska modeller oundvikligen levererar oss **för och från data**. Vi skall dock nu fokusera på mer praktiska problem som är förknippade med vissa sätt att tillämpa hypotesprövning.

1. Genom ett enögt fokus på signifikans, kan man välja modeller där en eller flera egenskaper i modellen saknar relevans i själva tillämpningen. Exempelvis kan skillnaden mellan A och B grupperna i chokladexemplet vara signifikant på 5%-nivån, men ändå vara så liten att den saknar medicinsk relevans. **Lösning**: Lita ej på p-värden som sådana, utan studera alltid även storleken på själva skillnaden, korrelationskoefficienten, regressionskoefficienten, osv.

2. Med stora datamaterial tenderar nollhypoteserna i samband med de flesta vanliga modellerna, så som regressionsmodellerna, alltid bli förkastade. Detta beror på att modeller sällan kan felfritt beskriva information i data, dvs. det brukar alltid finnas en viss diskrepans mellan

modellen och observationerna. Hur liten den skillnaden än är, kommer hypotesprövningen att upptäcka den och förkasta nollhypotesen, bara man skaffar sig tillräckligt med data. Hypotesprövningen tar ej generellt till hänsyn att datamaterialet kan vara ännu mer orimligt under en mothypotes jämfört med nollhypotesen, ävenom testet förkastar H_0 på en relativt liten signifikansnivå. **Lösning**: Delvis densamma som vid punkt 1, men även andra lösningar finns, t. ex. sk. informationskriterier (så som BIC, AIC) kan användas för att utesluta orimliga modeller.

3. Det är ibland nödvändigt att utföra ett stort antal test. Tänk Dig en situation, där man betraktar sambandet mellan förekomsten av migrän och ett stort antal variabler som beskriver individers personliga egenskaper och livsstil. Vi upprepar att man kontrollerar frekvensen av felaktiga beslut av Typ I (H_0 är sann, men förkastas ändå) genom ett litet värde på signifikansnivån α. Om M oberoende statistiska test utförs för att studera sambandet mellan migrän och de M variablerna, och nollhypotesen är sann (inget samband) i samtliga fall, förväntar vi oss αM felaktigt förkastade nollhypoteser. Exempelvis, med M =100 och $\alpha = 0.05$, förväntar vi oss alltså upptäcka 5 falska samband mellan migrän och diverse faktorer. **Lösning**: Man kan försöka hantera problemet med multipla hypotesprövningar genom olika korrektioner på de signifikansnivåer som används vid enskilda test. Traditionella metoder i detta sammanhang, så som sk Bonferroni-korrektioner, fungerar dock dåligt, eftersom de sänker testens styrka mot noll. De modernare metoderna, så som False Discovery Rate kontrollen, är att föredra. Även informationskriterier (så som BIC, AIC) kan återigen användas.

4. I många situationer vill man rangordna modeller enligt deras anpassningsnivå i förhållandet till data. Om modellerna är av olika typer, eller innehåller olika antal parametrar, är p-värden ej väl kalibrerade för rangordningssyftet. Ett konkret exempel på en dylik situation är regressionsanalys med multipla potentiella förklarande variabler, säg M stycken. Man vill alltså förklara variationen hos en (eller fler) beroende variabel Y, med hjälp av en möjligast bra delmängd av samtliga förklarande variabler $X_1, X_2, ..., X_M$. Då man testar den allmänna anpassningsgraden för en regressionsmodell, där en viss uppsättning av vari-

abler bland de M möjliga används, brukar man använda ett sk. likelihood-kvot test mot den modell där samtliga M används (mer om detta generella test senare). Ett litet p-värde indikerar då att anpassningsgraden för modellen inte är särskilt bra. Problemet är dock att p-värdet ökar automatiskt, då man lägger till fler förklarande variabler. Alltså, ju mer komplicerad modell (fler parametrar), desto bättre anpassningsgrad till de observerade datat enligt testet. Därför är det inte meningsfullt att försöka rangordna sådana modeller enligt p-värden. Utöver detta problem, måste man vanligtvis även tampas med problematiken gällande ett stort antal test (se punkt 3), om man använder den testbaserade ansatsen för att välja en lämplig uppsättning förklarande variabler. **Lösning**: Använd informationskriterier (så som BIC, AIC) för att välja rimliga modeller.

För samtliga problempunkter i listan ovan, kan en lösning även erhållas genom användning av sk Bayesiansk statistik. I många sammanhang kräver detta dock mer expertis för att kunna tillämpas på ett pålitligt sätt, eftersom den Bayesianska analysen kräver oftast att användaren ger sannolikhetsmässiga specifikationer gällande den aktuella osäkerheten för en modell och dess parametrar (sk *a priori* uppfattning). Vi noterar ändå att användandet av en del informationskriterier motsvarar approximativt en Bayesiansk analys i en modellvalssituation, se t. ex. `http://www.wikipedia.org/Bayesian_information_criterion`. Tyvärr krävs det en hel del matematiska förkunskaper för att studera närmare dylika kriterier och deras egenskaper.

Statistiska modeller ses oftast som något slags data-genererande maskiner, precis som i det experiment som utfördes för att beräkna den empiriska frekvensen av felaktigt förkastade nollhypoteser (se avsnitt 2.3). Detta är en användbar, men ibland en allför snäv tolkning av dem, som kan ställa till det rejält vid vissa situationer. Ett representativt exempel på detta problem är statistisk inferens av evolutionsmässigt släktskap mellan olika levande organismer (detta kallas ofta fylogenetik, se `http://www.wikipedia.org/Phylogenetics`). Inom det frekventistiska synsättet måste vi då inbilla oss multipla replikat av evolutionen, som pågått på dylika planeter som jorden, för att kunna få en tolkning av signifikansnivåer och p-värden. Detta kan minsann vara svårt att föreställa sig! En mycket enklare och intuitiv syn på osäkerheten om evolutionens gång får man genom att betrakta dess betingade sannolikhetsfördelning, givet det molekylära och morfologiska data vi har till förfogande. Denna syn motsvarar återigen den Bayesianska analysen och

kräver därmed vissa insikter i sannolikhetskalkyl för en kritisk betraktelse.

3 Direkt jämförelse av modellernas användbarhet

3.1 Likelihood-kvot

Då hypotestprövningen (eller det statistiska testet) introducerades, konstaterade vi att den inte fungerar som ett verktyg för en direkt jämförelse av modellernas användbarhet. Hur kunde man uppnå en sådan jämförelse? Vi närmar oss problemet genom ett konkret exempel.

Example 5 ***Graden av släktsskap****. Vi skall bedöma graden av släktskap hos två personer med hjälp sk DNA-markörer. Enligt vår nollhypotes är personerna i fråga bröder, medan de är halvsyskon enligt mothypotesen. De sju valda DNA-markörerna (kallas oftast loci och de motsvarar bestämda platser i arvsmassan) har ett givet antal olika värden (alleler) som kan observeras hos människor. Vi kunde ge dessa värden etiketter A,B,... för en viss markör och studera värdena för de två personerna, men för att förenkla beteckningarna, betraktar vi enbart indikatorvariabeln X_i som antar värdet 0, ifall de två personerna bär olika alleler vid lokuset i, och värdet 1 ifall individerna har samma allel vid lokuset i. Symbolen i fungerar här som ett index och antar värden mellan 1 och 7, eftersom vi har sju loci totalt. Detta brukar man beteckna som* $i = 1, ..., 7$ *i matematiska sammanhang. Observera att informationen ovan motsvarar en förenkling av verkligheten, eftersom människan är en diploid organism och har därmed två alleler vid varje lokus i genomet. Enligt den grundläggande genetiken skulle två bröder ha identiska alleler vid ca 50% av loci, medan två halvsyskon endast vid 25% av loci. Därenligt förväntar vi oss att antalet 'ettor' vi ser efter DNA-testet, jämfört med antalet 'nollor', tenderar vara olika beroende på vilken av hypoteserna är sann (observera att bägge hypoteserna kan vara felaktiga!). Både* H_0 *och* H_1 *är* ***enkla*** *hypoteser i situationen ovan, eftersom de entydigt bestämmer observationernas sannolikhetsfördelning.*

Example 6 ***Graden av släktsskap (fortsättning)****. För att komma åt en sannolikhetsbaserad beskrivning av datamaterialet vi erhåller i exemplet, är*

det nyttigt att betrakta en matematiskt analogisk situation. Tänk Dig en 8-sidig tärning (sådana förekommer t. ex. i vissa rollspel) med sidorna numrerade med 1,2,...,8. Vi antar att tärningen är balanserad, så att vid varje enskilt kast har samtliga sidor lika stor tendens att hamna uppåt. Låt oss namnge följande händelser vid ett enskilt kast: "vi slår något av siffrorna 1,2,3,4", vilket motsvarar att $X = 1$*; vi slår något av siffrorna 5,6,7,8", vilket motsvarar att* $X = 0$*. Om vi nu gör sju kast och registrerar alla händelserna (t. ex.* $X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 1, X_5 = 1, X_6 = 0, X_7 = 0$*), får vi ett datamaterial som är genererat enligt en sannolikhetsmodell som motsvarar nollhypotesen* H_0 *(två bröder). Värdena på* X_i *(*$i = 1, ..., 7$*) är alltså tänkta som* ***betingat oberoende*** *utfall, givet att hypotesen (dvs. modellen) håller. Det betingade oberoendet låter oss beräkna den sammansatta (eller simultana) sannolikheten för en viss händelse gällande samtliga variabler, som en produkt av sannolikheter för de enstaka händelserna, dvs.*

$$\begin{aligned} P(X_1 &= 1, X_2 = 0, X_3 = 1, X_4 = 1, X_5 = 1, X_6 = 0, X_7 = 0) = \\ &= P(X_1 = 1)P(X_2 = 0)P(X_3 = 1)P(X_4 = 1)P(X_5 = 1)P(X_6 = 0)P(X_7 = 0) \end{aligned}$$

Example 7 ***Graden av släktsskap (fortsättning)****. Precis som för fallet "*H_0*, två bröder", kan vi använda analogin för att åstadkomma en genererande modell för fallet "*H_1*, halvsyskon". Vi tittar fortfarande på samma tärning som förut, men namnger isället följande händelser vid ett enskilt kast: "vi slår antingen 1 eller 2", vilket motsvarar att* $X = 1$*; vi slår något av siffrorna 3,4,5,6,7,8", vilket motsvarar att* $X = 0$*. Om vi nu gör sju kast och registrerar alla händelserna (t. ex.* $X_1 = 1, X_2 = 0, X_3 = 0, X_4 = 0, X_5 = 1, X_6 = 0, X_7 = 0$*), får vi ett datamaterial som är genererat enligt en sannolikhetsmodell som motsvarar mothypotesen* H_1 *(två halvsyskon).*

Example 8 ***Graden av släktsskap (fortsättning)****. I detta exempel är 'ettorna' och 'nollorna' vi observerar, fördelade enligt en Bernoulli-fördelning(p), där parametern p är bestämd av hypotesen, dvs.* $p = 0.5$ *om vi tror på* H_0*, och* $p = 0.25$ *om vi tror på* H_1*. Låt* $\mathbf{x}$ *beteckna kompakt vår observationsserie* $x_1, ..., x_7$ *(de små bokstäverna betecknar att varje* X_i *har tilldelats ett visst värde, antingen 1 eller 0). Låt vidare y vara antalet 'ettor' vi har, därmed blir antalet 'nollor' lika med* $7 - y$*. Ett intuitivt redskap för jämförelse av två*

enkla hypoteser är en sk likelihood-kvot, som blir nu

$$\begin{aligned} L &= \frac{p(\mathbf{x}|H_0)}{p(\mathbf{x}|H_1)} = \\ &= \frac{p(x_1|H_0)\cdot p(x_2|H_0)\cdot p(x_3|H_0)\cdot p(x_4|H_0)\cdot p(x_5|H_0)\cdot p(x_6|H_0)\cdot p(x_7|H_0)}{p(x_1|H_1)\cdot p(x_2|H_1)\cdot p(x_3|H_1)\cdot p(x_4|H_1)\cdot p(x_5|H_1)\cdot p(x_6|H_1)\cdot p(x_7|H_1)} \\ &= \frac{(0.5)^y(0.5)^{7-y}}{(0.25)^y(0.75)^{7-y}} \end{aligned}$$

Likelihood-kvoten säger hur många gånger bättre sannolikhetsfördelningen enligt hypotesen i täljaren beskriver det observerade datamaterialet $\mathbf{x}$*, jämfört med fördelningen enligt hypotesen i nämnaren. Alltså, ett* $L = 3$ *säger att modellen enligt* H_0 *är tre gånger bättre än modellen enligt* H_1 *i detta hänseende. Om vi antar att observationerna består av serien* $X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 1, X_5 = 1, X_6 = 0, X_7 = 0,$ *blir* $y = 4$ *och* $L = 4.74,$ *dvs. nollhypotesen får ca fem gånger mer stöd än mothypotesen i detta fall.*

Example 9 ***Graden av släktsskap (fortsättning)****. Man bör komma ihåg att statistiska jämförelser av modeller kan vara ytterst vilseledande, t. ex. om båda hypoteserna råkar vara felaktiga. Detta blir mer förståeligt genom ett exempel. Anta att vi observerar följande händelser i släktsskapsstudien:* $X_1 = 0, X_2 = 0, X_3 = 0, X_4 = 0, X_5 = 0, X_6 = 0, X_7 = 0,$ *dvs.* $y = 1$*. Hypoteserna är fortfarande desamma som ovan, vilket leder till*

$$\begin{aligned} L &= \frac{p(\mathbf{x}|H_0)}{p(\mathbf{x}|H_1)} = \\ &= \frac{(0.5)^y(0.5)^{7-y}}{(0.25)^y(0.75)^{7-y}} = 0.0585, \end{aligned}$$

dvs. H_1 *är ca 17.08 gånger bättre än* H_0*. Alltså, vi får nu en stark indikation på att individerna ifråga är halvsyskon. Men, om vi betraktar närmare själva observationerna, tyder de på att individerna i själva verket inte alls är närbesläktade, eftersom ingen utav sju möjliga alleler är matchande för dem. Summan av kardemumman är att observationerna är rätt osannolika både under nollhypotesen och under mothypotesen, vilket leder till det till synes starka stödet för* H_1*, eftersom den hypotesen är mindre dålig. Då man lägger fram hypoteser, bör man fundera på om de överhuvudtaget kan vara rimliga representationer av data!*

Tyvärr kan likelihood-kvoten användas till en direkt jämförelse av två hypoteser endast när de är **enkla båda två**. För många typer av modelljämförelser som omfattar sammansatta hypoteser, använder man dock likelihood-kvoten som en testkvantitet, eftersom den har vissa goda matematiska egenskaper. Själva beräkningen blir mer komplicerad, då sammansatta hypoteser betraktas. Vi skall studera en dylik situation genom att modifiera ovanstående exemplet med graden av släktsskap.

Example 10 ***Graden av släktsskap (fortsättning)****. Vi fortsätter med att formulera nollhypotesen i enlighet med: "$H_0 : p = 0.5$, två bröder". Mothypotesen ändras däremot, så att enligt den kan de två individerna ha en godtycklig grad av släktskap, vilket översätts till en modell där $H_1 : p \in [0, 1]$. Sannolikheten för en 'etta' är nu ej given av hypotesen, utan den måste* ***skattas*** *med hjälp av observationerna. Skattningen görs i detta sammanhang enligt en generell princip som kallas likelihood-principen, och enligt den maximerar vi $p(\mathbf{x}|H_1)$ som en funktion av p för att åstadkomma $\hat{p}(\mathbf{x}|H_1)$. I vårt exempel ges maximum likelihood -lösningen $\hat{p}$ av den relativa frekvensen 'ettor', dvs. $\hat{p} = y/7$. Jämförelsen av de två hypoteserna leder då till*

$$
\begin{aligned}
L &= \frac{p(\mathbf{x}|H_0)}{\hat{p}(\mathbf{x}|H_1)} = \\
&= \frac{p(x_1|H_0) \cdot p(x_2|H_0) \cdot p(x_3|H_0) \cdot p(x_4|H_0) \cdot p(x_5|H_0) \cdot p(x_6|H_0) \cdot p(x_7|H_0)}{\hat{p}(x_1|H_1) \cdot \hat{p}(x_2|H_1) \cdot \hat{p}(x_3|H_1) \cdot \hat{p}(x_4|H_1) \cdot \hat{p}(x_5|H_1) \cdot \hat{p}(x_6|H_1) \cdot \hat{p}(x_7|H_1)} \\
&= \frac{(0.5)^y (0.5)^{7-y}}{\hat{p}^y (1-\hat{p})^{7-y}}.
\end{aligned}
$$

Om vi fortfarande antar att observationerna består av serien $X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 1, X_5 = 1, X_6 = 0, X_7 = 0$, blir $y = 4$ och $\hat{p} = 4/7$, vilket leder till $L = 0.93$. Det är ytterst viktigt att notera skillnaden till den tidigare jämförelsen - för sammansatta hypoteser H_1, där H_0 kan ses som ett specialfall, är $\hat{p}(\mathbf{x}|H_1)$ **aldrig** mindre än $p(\mathbf{x}|H_0)$. Därmed kan likelihood-kvoten ej direkt användas till att jämföra graden av stödet för hypoteserna, om de ej är enkla båda två. Eftersom fördelningen för $\log L$ är matematiskt hanterbar för generella modeller under nollhypotesen, används ett sk likelihood-kvot test väldigt allmänt för att avgöra om avvikelser från en viss hypotes är signifikanta.

3.2 Bayes faktor

I föregående avsnitt presenterades likelihood-kvoten som ett alternativ, då man ville direkt jämföra två hypotesers förmåga att förutspå ett observerat datamaterial. Ävenom likelihood-kvoten har en intuitiv tolkning för fallet med två enkla hypoteser, saknar den ett visst element som är relevant i många sammanhang. Detta element är *a priori* uppfattningen om hypotesernas giltighet gentemot varandra, dvs. uppfattningen innan vi observerade datamaterialet $\mathbf{x}$, samt *a priori* uppfattningen om osäkerheten gällande de parametrar som modellerna innehåller. Vi noterar att vid enkla hypoteser har man ingen osäkerhet om parametrar kvar, **efter** att man har fixerat modellens struktur.

Den sk **Bayes faktorn** representerar ett sätt att jämföra direkt två hypotesers förmåga att förutspå datamaterialet, och den definieras som

$$BF = \frac{p(\mathbf{x}|H_0)}{p(\mathbf{x}|H_1)},$$

dock så att $p(\mathbf{x}|H_0)$ och $p(\mathbf{x}|H_1)$ är exakt lika med den tidigare presenterade likelihood-kvoten **endast om** hypoteserna är **enkla**. Om Bayes faktorn är större än 1, ger modellen i täljaren (här H_0) en bättre prognos för datat än modellen i nämnaren, givet vår *a priori* uppfattning och datat. Om däremot Bayes faktorn får ett värde mindre än 1, är modellen i nämnaren ett bättre alternativ. Stora värden (t. ex $BF > 10$) indikerar ett starkt stöd för modellen i täljaren och små värden (t. ex. $BF < 0.10$) ett starkt stöd för modellen i nämnaren.

Anta att vår *a priori* uppfattning om osäkerheten säger att hypotesen H_0 torde gälla med sannolikheten $P(H_0)$. På motsvarande sätt har vi en *a priori* sannolikhet $P(H_1) = 1 - P(H_0)$ för mothypotesen. Bayes faktorn kan nu ses som den faktor som behövs för att komma från *a priori* till *a posteriori* odds för H_0 mot H_1, dvs.

$$\begin{array}{lll} \textit{a posteriori odds} & = \textit{Bayesfaktor} & \times \ \textit{a priori odds} \\ \dfrac{p(H_0|\mathbf{x})}{p(H_1|\mathbf{x})} & = & \dfrac{p(\mathbf{x}|H_0)}{p(\mathbf{x}|H_1)} \times \dfrac{P(H_0)}{P(H_1)} \end{array}$$

Datat kan alltså ses som en mängd information som antingen ökar eller minskar odds för H_0 mot H_1 från den ursprungliga uppfattningen. Notera att odds betyder här exakt samma sak som i vadslagning.

Bayes faktorn kan fortfarande användas och tolkas på samma sätt när den ena hypotesen (eller bägge två) är sammansatta, men det blir då mer krångligt att räkna ut den. Sannolikheten för datat, dvs. $p(\mathbf{x}|H_0)$ eller $p(\mathbf{x}|H_1)$, är i sådana fall en sk marginell likelihood (eller marginell fördelning för datat), som erhålls genom integration med avseende på en *a priori* fördelning för de parametrar som finns i modellen. Vi skall ej gå in på detaljerna för denna operation, utan det väsentliga är att Bayes faktorn lämpar sig för en direkt jämförelse av två statistiska modeller, oavsett om de motsvarande hypoteserna är enkla, sammansatta, eller en blandning av dessa. Wikipedia innehåller en kortfattad beskrivning av Bayes faktorn, se `http://www.wikipedia.org/Bayes_factor`. En allmän artikel om Bayes faktorn med diverse tillämpningar är Kass, R. and Raftery, A. (1995). Bayes factors. *J. Amer. Stat. Assoc.* **90**, 773-795.