10. STATISTISK INFERENS

10.1 Inledning

När datainsamlingen är avklarad blir nästa steg i den empiriska undersökningen att analysera datat. Det behöver knappast påpekas att det finns en stor mängd olika metoder för att analysera data, var och en av metoderna bygger på sina förutsättningar. Vi begränsar diskussionen här till att endast omfatta sådana metoder som förutsätter att vi har en sannolikhetsmodell som beskriver det studerade fenomenet. Sådana metoder kallas gemensamt för statistiska inferensmetoder.

Att analysera ett datamaterial med hjälp av statistisk inferens innebär att göra jämförelser mellan observationer och teoretiskt härledda förutsägelser. Analysen kan t ex innebära att mönster jämförs eller att samband mellan olika variabler undersöks. I analysen skall datamaterialets uppgifter omvandlas till meningsfylld information. Detta kräver god kännedom om problemområdet och att en väl underbyggd teori om problemområdet föreligger. Analysarbetet får inte urarta till att bli ett mekaniskt räknande som producerar nya meningslösa uppgifter ur gamla.

Under arbetet med problemformuleringen strukturerar man ofta frågeställningen i ett antal delfrågor. Analysarbetet kan då struktureras påsamma sätt, så att man i tur och ordning undersöker ett antal välavgränsade analysproblem. Valet av specifik analysmetod beror bl a på den modell vi har ställt upp och av den frågeställning vi vill analysera.

Ett viktigt hjälpmedel för statistisk inferens är s k empiriska fördelningar. Empiriska fördelningar är det empiriska observationsmaterialets motsvarighet till teorins sannolikhetsfördelningar. Vi diskuterar empiriska fördelningar närmare i avsnitt 10.2.

De två viktigaste problemtyperna som förekommer inom statistisk inferens är uppskattningsproblemet och inferensproblemet. I uppskattningsprob-

lemet söker vi en uppskattning av värdena på okända parametrar i den teoretiska modellen. I prövningsproblemet önskar vi pröva hypoteser om okända parametrar i den teoretiska modellen. I både uppskattningsproblemet och prövningsproblemet låter vi resultaten, uppskattningarna respektive utfallen från hypotesprövningar bero på vad som har observerats. Uppskattningsproblemet och prövningsproblemet diskuteras något ytterligare i avsnitt 10.3 respektive 10.4 och är föremål för mer omfattande diskussioner i kapitlen 16 -18.

10.2 Empiriska fördelningar

Ett av stegen i kunskapsbyggnadsprocessen är att samla in observerationer på verkligheten. Då modellen innehåller stokastiska variabler, innebär detta att vi gör observationer på stokastiska variabler. Vi vill nu använda dessa observationer tillsammans med modellen för att dra slutsatser om verkligheten. Detta är möjligt genom att den information som observationerna är bärare av kan uttolkas med hjälp av modellen. Det finns flera olika sätt att representera den information som ett datamaterial är bärare av. Ett sätt är att helt enkelt räkna upp alla observerade värden. Om vi t ex gör observationer på en stokastisk variabel X, kan vi låta x_1 representera värdet på den första observationen, x_2 värdet på den andra observationen, o s v. Om vi sammanlagt gör n observationer är x_1, x_2, ..., x_n en uppräkning av alla gjorda observationer, vilket är ett sätt att representera den information som data är bärare av.

Om n är stort, dvs om vi gör många observationer, blir ett datamaterial lätt oöverskådligt om man bara anger de observerade värdena. Alternativt kan man bestämma den så kallade empiriska fördelningsfunktionen $F_n(x)$.

Definition 1 *: Den empiriska fördelningsfunktionen definieras genom*

$$F_n(x) = \frac{\text{Antal observationer som är midre än eller lika med } x}{n}$$

Vi ser här att $F_n(x)$ är en så kallad trappstegsfunktion. Dess värde är noll då argumentet x är mindre än det minsta observerade värdet. För varje observerat värde blir det sedan ett trappsteg hos $F_n(x)$, där höjden av trappsteget är antalet observationer med värdet x dividerat med totala antalet

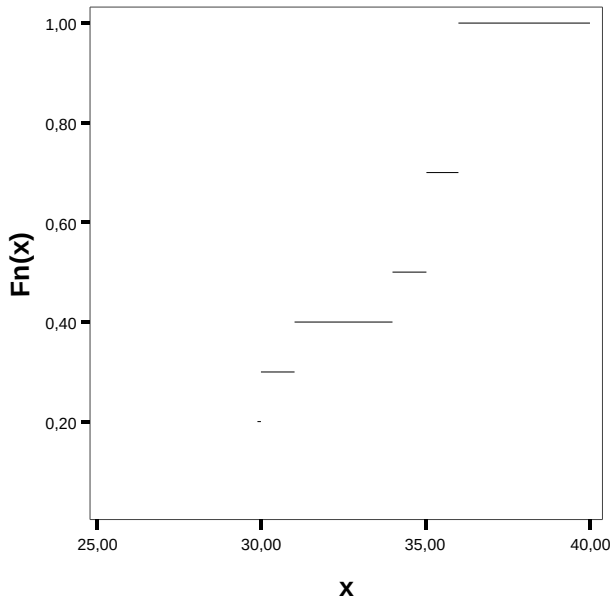
observationer, n. Väljer vi ett tillräckligt stort värde på argumentet x så att alla observationerna är mindre än eller lika med x, ser vi att $F_n(x) = 1$. Observera här likheten i egenskaper mellan den empiriska fördelningsfunktionen $F_n(x)$ och fördelningsfunktionen $F(x)$ för den stokastiska variabeln X. För varje par av tal a och b, $a \leq b$, kan vi t ex beräkna

$$F_n(b) - F_n(a) = \text{proportionen observationer i intervallet } a < x \leq b.$$

Observera också den begreppsmässiga skillnaden mellan den empiriska fördelningsfunktionen och fördelningsfunktionen för en stokastisk variabel. Medan den empiriska fördelningsfunktionen representerar vad vi faktiskt har observerat, så representerar fördelningsfunktionen sannolikheter enligt vår postulerade modell för vad som kan observeras. På så vis kan man säga att den empiriska fördelningsfunktionen ger en observerad bild av sannolikhetsfördelningen.

Exempel 10.1

Följande 10 observationer avser priset på en viss vara i 10 olika butiker: x_1 =30:-, $x_2 = 34$:-, $x_3 = 31$:-, $x_4 = 36$:-, $x_5 = 30$:-, $x_6 = 40$:-, $x_7 = 36$:-, $x_8 = 40$:-, $x_9 = 40$:- och $x_{10} = 35$:- (sortenhet kr). I figur 10.1 visas den empiriska fördelningensfunktionen. Det minsta observerade värdet är 30:- kr. Två av observationerna har detta värde (x_1 och x_5). Därför gör $F_n(x)$ ett språng med höjden 2/10 i punkten $x = 30$. En observation har värdet 31:- kr (x_3) varför $F_n(x)$ gör ett språng med höjden 1/10 i punkten $x = 31$. Nästa språng hos $F_n(x)$ är inte förän i punkten $x = 34$ eftersom ingen observation har ett värde mellan 31:- kr och 34:- kr. På så vis fortsätter $F_n(x)$ att språngvis ta sig upp mot värdet 1.00, vilket den får då x är större än eller lika med det största observerade värdet, 40 kr. ■



Figur 10.1 Empiriska fördelningsfunktionen för prismaterialet.

Om ett observationsmaterial innehåller väldigt många olika värden kan det vara praktiskt att klassindela materialet. Man får på så sätt en approximation till den empiriska fördelningsfunktionen. Klassindelning behandlas utförligare i kapitel 11. För diskreta variabler kan man som ett alternativ till den empiriska fördelningsfunktionen bestämma den empiriska frekvensfunktionen. Grafiskt illustreras den med ett stolpdiagram. För observationer på kontinuerliga variabler blir situationen lite mer komplicerad. För att efterlikna täthetsfunktionens egenskaper, vill vi att den empiriska täthetsfunktionen konstrueras så att arean under den empiriska motsvarigheten, $f_n(x)$, och mellan två punkter, a och b, är lika med relativa andelen observerade värden i intervallet $a < x \leq b$. Tyvärr är det inte möjligt att konstruera en funktion som har den egenskapen. Däremot kan vi med hjälp av en klassindelning av observationsmaterialet konstruera en empirisk täthetsfunktion vars egenskaper så långt som möjligt har denna egenskap. Grafiskt illustreras den empiriska täthetsfunktionen med ett histogram.

Exempel 10.1 (fortsättning).

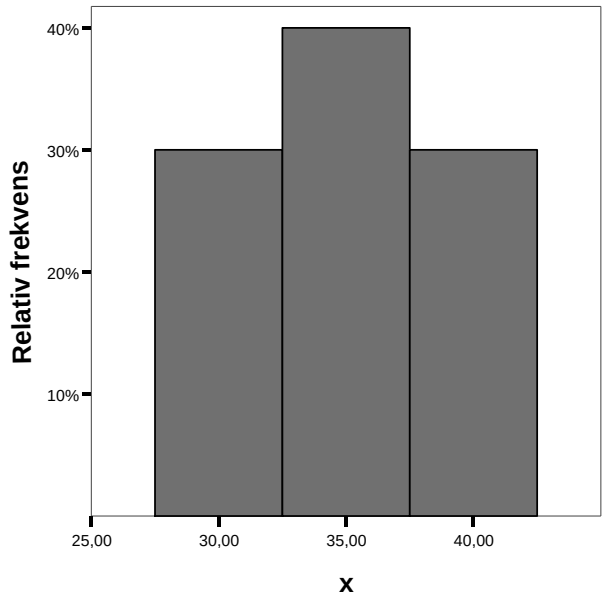
En (av flera möjliga) klassindelningar av prisuppgifterna redovisas i tabell 10.1. Motsvarande empiriska täthetsfunktion illustreras i figur 10.2.

Tabell 10.1 En klassindelning av prisuppgifterna

Prisklass (kr)	frekvens (antal)	relativ frekvens
27:50-32:50	3	0.3
32:50-37:50	4	0.4
37:50-42:50	3	0.3
Totalt	10	1.0

Antag nu att den stokastiska variabeln X vi gör våra observationer på följer en sannolikhetsfördelning och att θ är en parameter i den fördelningen. Vi vet då att sannolikhetsfördelningen för X karaktäriseras av värdet på θ och att olika sannolikhetsfördelningar karaktäriseras av olika värden på θ. På motsvarande sätt kan vi definiera storheter, så kallade urvalskaraktäristikor, som karaktäriserar den empiriska fördelningen. Olika empiriska fördelningar karaktäriseras då av olika värden på urvalskaraktäristikorna. Problemet att bestämma värdet på urvalskaraktäristikor brukar kallas uppskattningsproblemet och diskuteras ytterligare i avsnitt 10.3.

Även om en sannolikhetsfördelning fullständigt specifiseras av dess parametrar brukar man ofta ange fördelningens väntevärde $E[X]$ och varians $V[X]$ som allmänna karakteristika på fördelningen. Dessa värden anger ett lägesmått respektive ett variationsmått för sannolikhetsfördelningen. Motsvarigheterna för empiriska fördelningar är medelvärdet $\bar{x}$ och (urvals-) variansen s^2.



Figur 10.2 Histogram över den empiriska täthetsfunktion som svarar mot klassindelningen i tabell 10.1

Definition 2 *: Medelvärdet och urvalsvariansen definieras genom*

$$\bar{x} = \sum_{i=1}^{n} x_i$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Som spridningsmått använder vi även (urvals-) standardavvikelsen, s, som är kvadratroten ur variansen.

Empiriska fördelningar och deras karaktäristika behandlas utförligare i kapitlen 11 - 13.

Lägg noga märke till att de begrepp vi har diskuterat i det här avsnittet är observationer av verkligheten och att dessa begrepp har sina motsvarigheter i modellvärlden. T ex svarar den empiriska fördelningen mot modellvärldens sannolikhetsfördelning, den empiriska fördelningens medelvärde svarar mot sannolikhetsfördelningens väntevärde o s v.

Några av verklighetens och modellvärldens begrepp och hur de kan kopplas samman visas i figur 10.3

Modell	**Verklighet**
Stokastisk variabel X	Observation x
Fördelningsfunktion $F(x)$	Empirisk fördelningsfunktion $F_n(x)$
Sannolikhetsfördelning	Empirisk fördelning
Väntevärde $E[X]$	Medelvärde $\overline{x}$
Varians $V[X]$	Urvalsvarians s^2

Figur 10.3 Några motsvarigheter i modellvärlden och i verkligheten

10.3 Uppskattningsproblemet

Betrakta det slumpmässiga försöket "ett barn föds och vi noterar dess kön". Vi har modellen

$\Omega = \{\text{pojke, flicka}\}$

$P(\text{pojke}) = \pi$

$P(\text{flicka}) = 1 - \pi$

Modellen innehåller en okänd parameter, π, som anger sannolikheten för utfallet "pojke". Den frekventistiska tolkningen av denna parameter i modellvärlden är "proportionen pojkar i alla födslar". I många tillämpningar (t ex när vi gör befolkningsprognoser) är vi intresserade av att veta värdet på π, eller åtminstone ett närmevärde till π. Antag att vi har observerat 100 födslar och noterat "53 barn var pojkar". Det är nu lockande att induktivt dra slutsatsen "53% av alla barn som föds är pojkar", eller formulerat i modelltermer "$\pi = 0.53$". Observera att vi gör då ett uttalande om alla födslar, alltså även födslar som är helt okända för oss. Vår slutledning kan sålunda vara felaktig. Däremot kan vi använda talet 0.53 som en uppskattning av

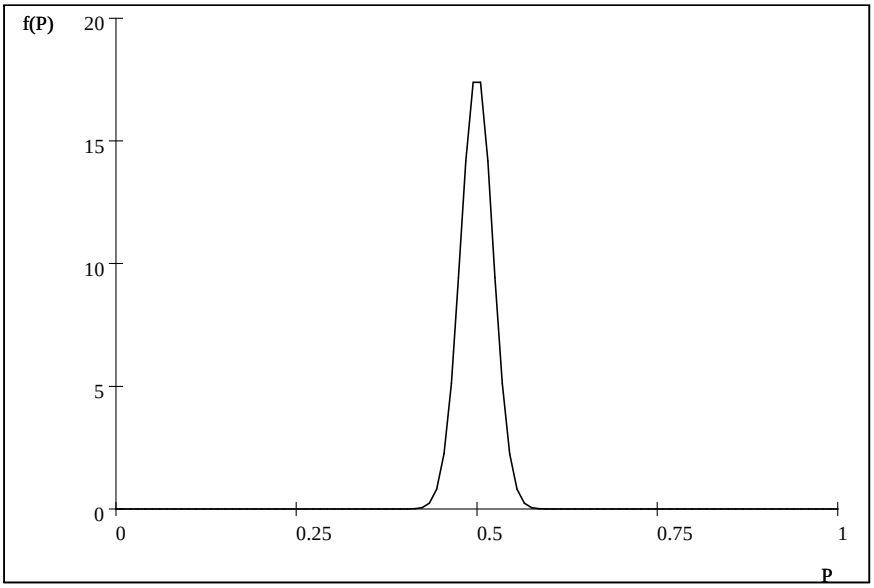
π. Ibland används även det engelska ordet estimat för uppskattning. Om vi betecknar en uppskattning av π med $\widehat{\pi}$, så är den slutsats vi drar "$\widehat{\pi} = 0.53$".

En uppskattad parameter är en induktivt dragen slutsats och är därför behäftad med en viss osäkerhet. Om vi gör fler observationer bör vi dock bli säkrare i våra slutsatser. Observerar vi t ex 1000 födslar är vår osäkerhet mindre än om vi observerar endast 100 födslar o s v. Förutom att få en uppskattning av okända modellparametrar är det också önskvärt att kunna tala om hur osäkra uppskattningarna är. Vi skall därför inte bara ägna oss åt att uppskatta okända parametrar, utan också försöka ge ett mått på osäkerheten i uppskattningarna.

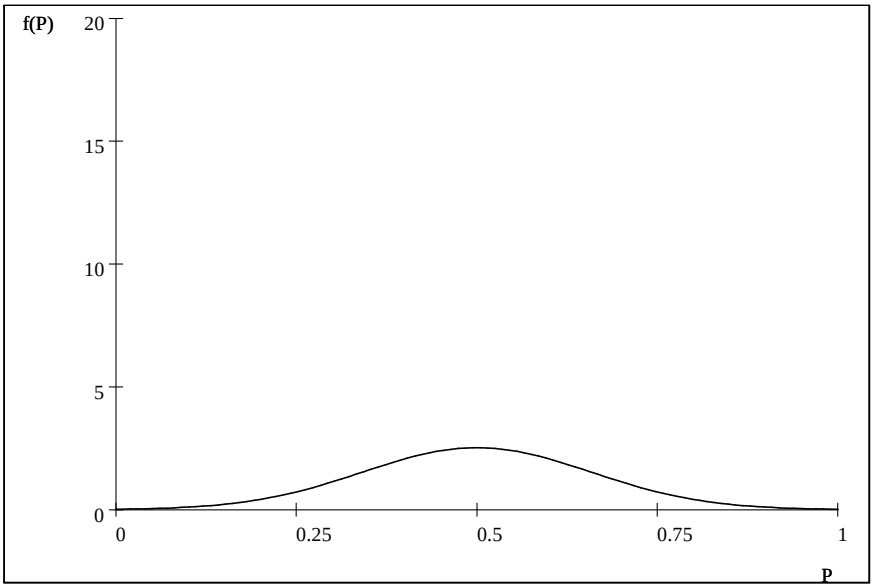
Lägg märke till att vår uppskattning baserar sig på observationer på slumpmässiga försök. Värdet på uppskattningen är beroende på resultaten i de enskilda födslarna. Om t ex 52 av barnen hade varit pojkar hade vi uppskattat π snarare med $\widehat{\pi} = 0.52$ än med $\widehat{\pi} = 0.53$. Tydligen är uppskattningen resultatet av ett antal slumpmässiga försök. Vi kan också betrakta alla födslar tillsammans som ett enda stort försök. Om födslarna är oberoende av varandra blir antalet födslar en binomialfördelad stokastisk variabel, X. Vi bildar sedan en ny stokastisk variabel, P, genom en transformation av X,

$$P = X/n,$$

där n är antalet observerade födslar. Vår uppskattning $\widehat{\pi}$ är sålunda en observation på den stokastiska variabeln P. Om sannolikhetsfördelningen för P är koncentrerad i närheten av den för oss okända populationsproportionen π, så är det stor sannolikhet att observationen, dvs $\widehat{\pi}$ är nära π. Se figur 10.4. Om däremot sannolikhetsfördelningen för P är utbredd och har stor sannolikhetsmassa långt ifrån π, är det inte särskillt troligt att $\widehat{\pi}$ kommer att ligga i närheten av π. Se figur 10.5. Genom att studera sannolikhetsfördelningen för P kan vi alltså bilda oss en uppfattning om uppskattningens precision.



Figur 10.4 Sannolikhetsfördelningen för P är koncentrerad kring π. En observaion på P kommer därför att med stor sannolikhet ligga nära π.



Figur 10.5 Sannolikhetsfördelningen för P är utbredd. En observaion på P kan därför vara långt från π.

Antag nu mer allmänt att vi gör observationerna x_1, x_2, ..., x_n på en stokastisk variabel X och att X har en sannolikhetsfördelning som innehåller en okänd parameter θ. Problemet att finna en uppskattning av värdet på θ med hjälp av de observerade värdena x_1, x_2, ..., x_n kallas uppskattningsproblemet. Den uppskattning vi får baseras på observationerna och är därför själv en observation på en stokastisk variabel. Den sannolikhetsfördelning som svarar mot en uppskattning kallas samplingfördelning.

Eftersom samplingfördelningen ger information om uppskattningens egenskaper, är det viktigt att känna till samplingfördelningar och att kunna arbeta med dem. Vi ägnar därför kapitel 15 åt en relativt noggrann genomgång av samplingfördelningar.

Vi skall avslutningsvis poängtera att uppskattningsproblemet behandlas under förutsättning att den modell vi har satt upp för att beskriva observationerna är korrekt. Skulle modellen vara fel förlorar de uppskattningar vi härleder sin betydelse. Att t ex försöka uppskatta en proportion i en fördelning som saknar en sådan parameter är givetvis en meningslös sysselsättning.

10.4 Prövningsproblemet

Antag nu att vi har en fullständigt specificerad modell och att vi vill pröva den. En sådan modell över det slumpmässiga försöket "ett barn föds och vi noterar dess kön" är t ex

$\Omega = \{\text{pojke, flicka}\}$

$P(\text{pojke}) = 0,5$

$P(\text{flicka}) = 0,5$

I denna modell finns det dels ett påstående om utfallsrummet, dels ett påstående om sannolikheterna för utfallen. Påståendet om utfallsrummet är tämligen okontroversiellt och behöver knappast bli föremål för en prövning. Däremot måste påståendet om sannolikheterna prövas. I detta avsnitt redogör vi för de grundläggande begreppen då sådana påståenden prövas.

Ett påstående om verkligheten, t ex $P(\text{pojke}) = 0,5$, kallas en hypotes. Vi betecknar allmänt hypoteser med H. Den hypotes vi vill pröva kallas nollhypotes och betecknas med H_0.

Exempel 10.2

Antag att X är en normalfördelad stokastisk variabel. Om vi vill pröva hypotesen att väntevärdet μ är lika med 5, skall vi alltså pröva

$$H_0 : \mu = 5.$$

Vill vi pröva att variansen σ^2 är lika med 2 har vi

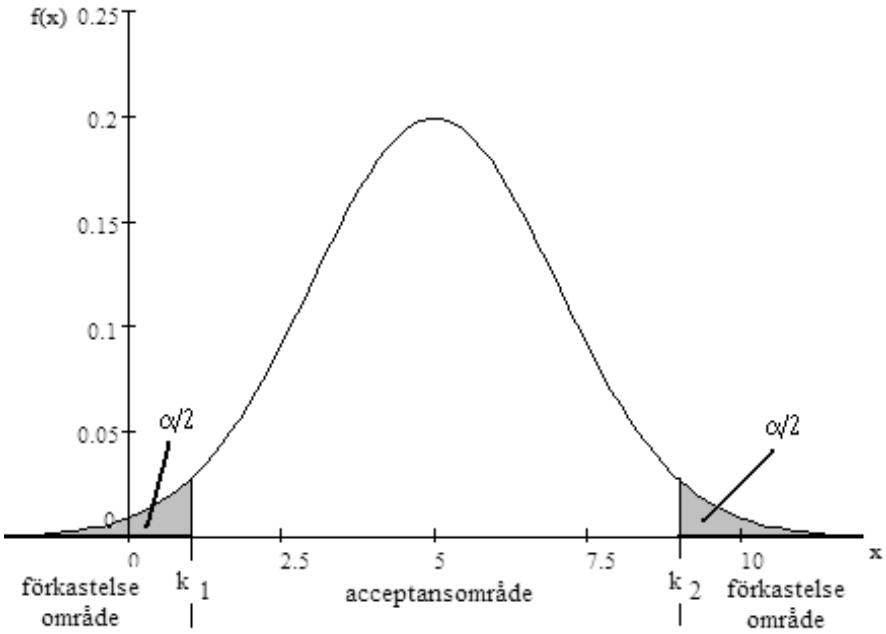
$$H_0 : \sigma^2 = 2. \blacksquare$$

För att utföra en prövning av en hypotes måste vi göra observationer på vår stokastiska variabel och på basis av observationerna induktivt avgöra om påståendet är rimligt eller ej. Om observationerna inte överensstämmer med påståendet, tror vi att påståendet är falskt. Vi säger då att vi förkastar hypotesen. Om däremot påståendet verkar överensstämma med observationerna säger vi att vi inte förkastar hypotesen. Det finns många sätt att avgöra om vi skall förkasta eller inte förkasta en hypotes. Vi skall här endast diskutera en av dessa metoder.

Exempel 10.3

Antag att vi gör observationer på diametrarna på n st rör och beräknar medelvärdet. Vi betecknar detta medelvärde (medeldiametern) med $\overline{X}$. Eftersom värdet på $\overline{X}$ beror på ett slumpmässigt försök (de enskilda observationerna $X_1, X_2, ..., X_n$) ser vi att $\overline{X}$ är en stokastisk variabel. Sannolikhetsfördelningen för $\overline{X}$ är följaktligen en samplingfördelning. Om de observerade diametrarna hos rören är normalfördelade med väntevärdet μ_0 och standardavvikelsen σ, kan man visa att (se kapitel 15) samplingfördelningen för $\overline{X}$ är normalfördelad med väntevärde μ_0 och standardavvikelsen $\sigma/\sqrt{n}$. Förmodligen kommer det observerade medelvärdet $\bar{x}$ inte att vara lika med

μ_0. Frågan är nu om skillnaden mellan μ_0 och det observerade medelvärdet är ett uttryck för slumpmässig variation, eller om skillnaden är så stor att vi kan anta att en stystematisk avvikelse från μ_0 föreligger. Vi har alltså att avgöra vad som är rimliga värden på $\bar{x}$ om hypotesen är sann. ∎



Figur 10.6 Uppdelning av utfallsrummet för $\overline{X}$ i ett acceptansområde och ett förkastelseområde

För att pröva en hypotes om väntevärdet i en sannolikhetsfördelning delas utfallsrummet för $\overline{X}$ upp i två delar. Den ena delen, den som innehåller rimliga värden på $\overline{X}$ om hypotesen är sann, kallas acceptansområde. Den andra delen kallas förkastelseområde. Se figur 10.6. Om det observerade värdet på $\overline{X}$ hamnar i acceptansområdet, dvs $\bar{x}$ är rimligt om hypotesen är sann, kan vi inte förkasta hypotesen. Å andra sidan, om det observerade värdet på $\overline{X}$ hamnar i förkastelseområdet, dvs $\bar{x}$ är orimligt om hypotesen är sann, förkastar vi hypotesen.

I figur 10.6 begränsas acceptansområdet av två tal, k_1 och k_2. Dessa tal kallas kritiska värden. Vår uppgift är att bestämma de kritiska värdena. Tydligen

gäller det att

om $\bar{x} < k_1$ eller $\bar{x} > k_2$ så förkastar vi hypotesen

om $k_1 \leq \bar{x} \leq k_2$ så förkastar vi inte hypotesen.

Även om hypotesen är sann är det möjligt att vi får ett observerat värde på $\overline{X}$ som ligger i förkastelseområdet. Det är alltså möjligt att förkasta en sann hypotes. Vi säger då att vi gör ett typ I-fel (eller $\alpha-$fel). Vi betecknar sannolikheten att göra ett typ I-fel med α, dvs

$$P(\text{vi förkastar } H_0 | H_0 \text{ är sann}) = \alpha$$

Ibland kallas också α för testets signifikansnivå. Sannolikheten α representeras av den streckade ytan i figur 10.6. Den streckade ytan till vänster om k_1 svarar alltså mot sannolikheten $\alpha/2$, dvs

$$P(\bar{X} \leq k_1 | H_0 \text{ är sann}) = \alpha/2$$

Om samplingfördelningen för $\overline{X}$ är känd då H_0 är sann, är det nu lätt att beräkna det kritiska värdet k_1.

Exempel 10.4

Antag att $\overline{X}$ är normalfördelad med väntevärdet 5 och standardavvikelsen 2 då H_0 är sann. Om vi är beredda att göra ett typ I-fel med sannolikheten 0,05 gäller det att

$$P(\overline{X} \leq k_1 | H_0 \text{ är sann}) = \frac{0,05}{2}$$

Eftersom $(\overline{X}-5)/2$ är en normalfördelad stokastisk variabel med väntevärdet 0 och variansen 1, får vi

$$P(\frac{\overline{X}-5}{2} \leq \frac{k_1-5}{2}) = 0,025$$

Men

$$\Phi(-1,96) \approx 0,025$$

så att

$$\frac{k_1 - 5}{2} \approx -1,96$$

eller

$$k_1 \approx -1,96 \cdot 2 + 5 = 1,08 \blacksquare$$

På samma sätt svarar den streckade ytan till höger om k_2 i figur 10.6 mot sannolikheten $\alpha/2$. Därför gäller det att

$$P(\overline{X} \geq k_2 | H_0 \text{ är sann}) = \alpha/2.$$

Lägg märke till hur den logiska strukturen i prövningsproblemet följer den hypotetisk-deduktiva metoden. För prövningsproblemet identifierar vi först en logisk implikation:

"Om H_0 är sann, så är det rimligt att det observerade värdet på $\overline{X}$ hamnar i intervallet $[k_1, k_2]$".

Med logiska symboler skriver vi det som

$$H_0 \xrightarrow{p} I,$$

där H_0 är den hypotes vi önskar pröva och I är den observationssats ($\overline{X}$ hamnar i intervallet $[k_1, k_2]$) som med stor sannolikhet inträffar om H_0 är sann. Symbolen $\xrightarrow{p}$ betecknar just att implikationen inträffar med stor sannolikhet. I själva verket är den sannolikheten $P\left(k_1 \leq \overline{X} \leq k_2 \mid H_0 \text{ är sann}\right) = 1 - \alpha$, där vi i den här testproceduren på förhand kan välja värdet på α. Vi

identifierar sedan en av två möjliga observationssatser i prövningsproblemet, nämligen

$$I : \text{“Det observerade värdet på } \overline{X} \text{ hamnade i intervallet } [k_1, k_2]\text{”}$$

eller

$$\overline{I} : \text{“Det observerade värdet på } \overline{X} \text{ hamnade ej i intervallet } [k_1, k_2]\text{”}.$$

Den slutsats vi sedan drar beror på observationssatsen. Om vi observerar $\overline{I}$ förkastas hypotesen H_0 medan om vi observerar I inte förkastar H_0.

Övning 10.1

Diskutera hur antalet observationer påverkar säkerheten i Arbuthnot's slutsats i avsnitt 4.4.