15. SAMPLINGFÖRDELNINGAR OCH CENTRALA GRÄNSVÄRDESSATSEN

15.1 Samplingfördelningar

I tidigare kapitel har vi diskuterat stokastiska variabler och deras sannolikhetsfördelningar då variablerna har representerat observationer på individer. I statistiska undersökningar brukar man emellertid göra observationer på många individer, ett urval, och beräkna någon karakteristika, t ex urvalsmedelvärdet. Eftersom olika slumpmässiga urval kan bestå av olika individer kommer urvalskarakteristikor att få olika värden beroende på vilka individer som råkar ingå i urvalet. Därmed kan man också se urvalskarakteristikorna som stokastiska variabler. Sannolikhetsfördelningen för en urvalskarakeristika kallas samplingfördelning. I detta avsnitt skall vi studera samolingfördelningar för urvalsmedelvärden, urvalsvarianser och urvalsproportioner. En speciell egenhet som urvalsmedelvärden har är att under vissa villkor blir deras fördelning mer och mer lik en normalfördelning. Det resultatet kallas Centrala gränsvärdessatsen. Tack vare Centrala gränsvärdessatsen kan sannolikhetsfördelningen för urvalsmedelvärden ofta approximeras med en normalfördelning, något som förenklar beräkningarna högst väsentligt.

Exempel 15.1.

Antag vi har en populationen bestående av en årskull skolelever vars kroppsvikt vi observerar. Antag vidare att medelvikten i hela populationen är μ och att variansen av vikterna är σ^2. Vi utför nu det slumpmässiga försöket "att slumpmässigt välja ut en elev ur populationen, notera elevens vikt och sedan återföra eleven till populationen". Låter vi X_i vara den i-te observerade elevens vikt (den i-te upprepningen av försöket) gäller det att X_i har väntevärdet μ och variansen σ^2. ■

Exempel 15.2.

Betrakta populationen bestående av mängden av alla tänkbara kast med ett mynt. Till varje kast tillordnas en stokastisk variabel X_i såden att

$$X_i = \left\{ \begin{array}{l} 1 \text{ om “klave” inträffar} \\ 0 \text{ om “krona” inträffar} \end{array} \right\}$$

För ett symmetriskt mynt har således varje stokastisk variabel X_i

$$E(X_i) = 0,5$$

$$V(X_i) = 0,25. \blacksquare$$

Antag att vi har n stycken stokastiska variabler X_1, $X_2, ..., X_n$, alla med väntevärde μ och varians σ^2. För varje urval om n observationer, x_1, $x_2, ...x_n$, kan vi bilda medelvärdet

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

och stickprovsvariansen

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

Flera andra karakteristikor kan beräknas, men här betraktar vi endast $\bar{x}$ och s^2. För ett visst urval kommer $\overline{x}$ att få ett visst värde och s^2 ett visst värde. Drar vi ett nytt urval kommer förmodligen $\overline{x}$ och s^2 att få andra värden, dvs de värden vi får på $\overline{x}$ och s^2 är resultat av ett slumpmässigt försök, nämligen det slumpmässiga försöket att dra ett slumpmässigt urval om n stycken individer, observara dem och beräkna observationernas medelvärde. Detta innebär att vi kan betrakta $\bar{x}$ och s^2 som observationer på de stokastiska variablerna $\overline{X}$ och S^2 definierade genom

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

respektive

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Sannolikhetsfördelningarna för $\overline{X}$ och S^2 kallas samplingfördelningen för urvalsmedelvärdet respektive samplingfördelningen för urvalsvariansen.

Låt $\mu_{\overline{X}}$ och $\sigma^2_{\overline{X}}$ beteckna väntevärdet respektive variansen för $\overline{X}$. Med hjälp av räknereglerna för väntevärden får vi

$$\mu_{\overline{X}} = E\left(\overline{X}\right) = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}E(\sum_{i=1}^{n} X_i)$$

$$= \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\ \mu = \mu,$$

dvs om vi drar många urval ur denna population, alla med storleken n, kommer medelvärdet $\bar{x}$ att i medeltal bli lika med populationens medelvärde μ.

Antag nu att de stokastiska variablerna X_1, $X_2, ..., X_n$ är oberoende. Enligt räknereglerna för varianser får vi

$$\sigma^2_{\overline{X}} = V(\overline{X}) = V(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}V(\sum_{i=1}^{n} X_i)$$

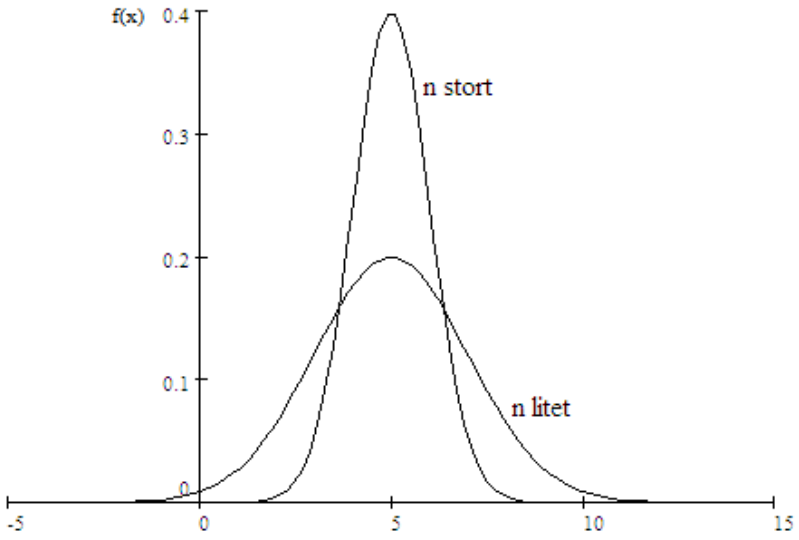
$$= \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{1}{n^2} n\ \sigma^2 = \frac{\sigma^2}{n}.$$

Om individerna dras slumpmässigt från en oändlig population kommer observationerna att vara oberoende. Om däremot populationen är ändlig kan observationerna bli beroende. Detta avgörs om urvalet sker med eller utan återläggning. Vid dragning med återläggning väljs först en individ ur populationen, mätningen utförs, X_1 noteras och individen "läggs tillbaka" i (återförs till) populationen, varefter en ny dragning genomförs och värdet X_2 noteras osv tills vi har gjort n stycken observationer. Att "lägga tillbaka" en individ innebär att en individ kan bli dragen flera gånger. För att oberoende skall råda får ju inte utfallet från ett slumpmässigt försök (händelsen att en viss individ har blivit dragen) påverka sannolikheterna vid en upprepning av försöket (sannolikheten att en viss individ skall bli dragen). Vid dragning utan återläggning däremot, kan en individ som tidigare har blivit dragen inte bli dragen en gång till. Därför kommer observationerna att bli beroende vid dragning utan återläggning. Speciellt innebär detta att variansformeln för urvalsmedelvärdet ovan inte gäller. Man kan visa att för dragning utan återläggning gäller det att

$$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right),$$

där N är antal individer i hela populationen. Vi ser att det som skiljer de båda variansformlerna är en korrektionsfaktor $(N-n)/(N-1)$ i fallet utan återläggning. För stora populationer då N är stort i förhållande till urvalsstorleken n, blir korrektionsfaktorn nära 1, varför den enklare formeln för fallet med återläggning kan användas som en god approximation.

Från variansformlerna framgår att $\sigma^2_{\overline{X}}$ beror av urvalsstorleken n på sådant sätt att ju större urval vi tar, desto mindre blir $\sigma^2_{\overline{X}}$. Detta kan tolkas som att ju fler observationer vi gör, desto bättre information får vi om populationens medelvärde, vilket är ett högst rimligt resultat! I figur 15.1 visas hur samplingfördelningen för $\overline{X}$ förändras då urvalsstorleken ökar, i det fall då fördelningen är kontinuerlig och symmetrisk kring populationsmedelvärdet μ.



Figur 15.1 Ett exempel på hur samplingfördelninen för $\overline{X}$ kan se ut då urvalet är litet respektive stort.

Exempel 15.3.

Vi illustrerar samplingfördelningen med utgångspunkt från en mycket liten population. Antag att vår population består av tre personer ($N = 3$) med kroppsvikterna 62, 65 respektive 68 kg. Tydligen är populationsmedelvärdet

$$\mu = \frac{1}{3}(62 + 65 + 68) = 65$$

och populationsvariansen

$$\sigma^2 = \frac{1}{3}\left((62 - 65)^2 + (65 - 65)^2 + (68 - 65)^2\right) = 6$$

Vi bildar nu alla tänkbara urval om två individer ($n = 2$) som kan dras med återläggning ur populationen. För varje sådant urval beräknar vi urvasmedelvärdet $\overline{X}$.

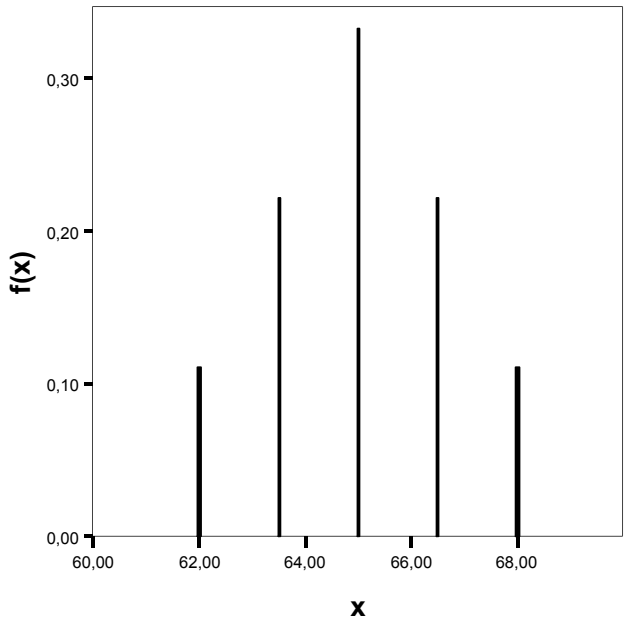
Tabell 15.1 De olika tänkbara urvalen (9 stycken) och hur samplingfördelningen för $\overline{X}$ genereras.

1:a dragningen	2:a dragningen	sannolikhet	medelvärde
62	62	1/9	62,0
62	65	1/9	63,5
62	68	1/9	65,0
65	62	1/9	63,5
65	65	1/9	65,0
65	68	1/9	66,5
68	62	1/9	65,0
68	65	1/9	66,5
68	68	1/9	68,0

Den högra kolumnen i tabell 15.1 visar de olika urvalsmedelvärden vi kan få. Tydligen varierar urvalsmedelvärdet $\overline{X}$ från 62,0 kg till 68,0 kg. Vissa värden på $\overline{X}$ förekommer i två eller flera urval. Exempelvis förekommer $\overline{X} = 63,5$ kg två gånger. Detta innebär att sannolikheten att få urvalsmedelvärdet 63,5 vid slumpmässig dragning med återläggning av två element ur populationen är 2/9, eftersom varje urval har sannolikheten 1/9. Samtliga värden på $\overline{X}$ med tillhörande sannolikheter utgör samplingfördelningen för $\overline{X}$ då urvalsstorleken är 2. Denna fördelning visas i tabell 15.2 och i figur 15.2.

Tabell 15.2 Samplingfördelningen för $\overline{X}$.

$\bar{x}$	$f(\bar{x})$	$\overline{x}f\left(\overline{x}\right)$	$\overline{x}^2 f\left(\overline{x}\right)$
62,0	1/9	62/9	3844,0/9
63,5	2/9	127/9	8064,5/5
65,0	3/9	195/9	12675,0/9
66,5	2/9	133/9	8844,5/9
68,0	1/9	68/9	4624,0/9
Summa	1,0	585/9	38052,0/9



Figur 15.2 Samplingfördelingen för $\overline{X}$ då urvalsstorleken n är 2

För samplingfördelningen för $\overline{X}$ gäller tydligen att

$$E\left(\overline{X}\right) = \mu_{\overline{X}} = \frac{585}{9} = 65 = \mu$$

dvs väntevärdet i samplingfördelningen är lika med populationsmedelvärdet.

Variansen i samplingfördelningen skiljer sig däremot från populationsvariansen. Detta inser man också genom att studera figuren över samplingfördelningen. En beräkning av variansen för $\overline{X}$ ger

$$V\left(\overline{X}\right) = \sigma^2_{\overline{X}} = \frac{38052}{9} - 65^2 = \frac{27}{9} = 3 = \frac{\sigma^2}{n},$$

dvs variansen i samplingfördelningen är lika med populationsvariansen dividerat med 2, urvalsstorleken. ■

Exempel 15.4.

Antag att vi gör 60 kast med ett mynt där

$$P(\text{klave}) = 0,41$$

$$P(\text{krona}) = 1 - 0,41 = 0,59$$

.

Om klave kommer upp låter vi en stokastisk variabel X få värdet ett. Kommer krona upp får X värdet noll. Den stokastiska variabeln X har alltså väntevärdet $E(X) = 0,41$ och variansen $V(X) = 0,41 \cdot 0,59 = 0,2419$. Medelvärdet $\overline{X}$ definieras nu som

$$\overline{X} = \frac{\sum X_i}{n} = \frac{\text{antal klave}}{60}$$

och samplingfördelningen för $\overline{X}$ har karakteristikorna

$$\mu_{\overline{X}} = 0,41$$

$$\sigma^2_{\overline{X}} = \frac{0{,}2419}{60} \approx 0{,}004032$$

Om vi låter T vara den stokastiska variabeln "antalet klave bland de 60 kasten" får vi

$$T = \sum_{i=1}^{60} X_i = 60\overline{X}$$

och T är binomialfördelad med parametrarna $n = 60$ och $p = 0,41$. ■

15.2 Samplingfördelningen för urvalsmedelvärdet vid observationer från en normalfördelad population då variansen är känd

Antag att vi har n stycken oberoende stokastiska variabler, X_1, X_2, ..., X_n, och att alla är normalfördelade med väntevärdet μ och variansen σ^2. Man kan visa att då är fördelningen för urvalsmedelvärdet, $\overline{X}$, också normalfördelad med väntevärdet μ, men med variansen σ^2/n. Följaktligen blir standardavvikelsen för $\overline{X}$ lika med $\sigma/\sqrt{n}$. Detta innebär att i normalfördelningsfallet så har urvalsmedelvärdet en mindre varians än de enskilda observationerna (precis som vi väntar oss), dvs samplingfördelningen för urvalsmedelvärdet blir med koncentrerad än populationens fördelning. Däremot förändras inte samplingfördelningens övriga egenskaper. Om variansen σ^2 är känd innebär detta att en rad sannolikhetsteoretiska och inferensteoretiska problem kan lösas på ett enkelt sätt.

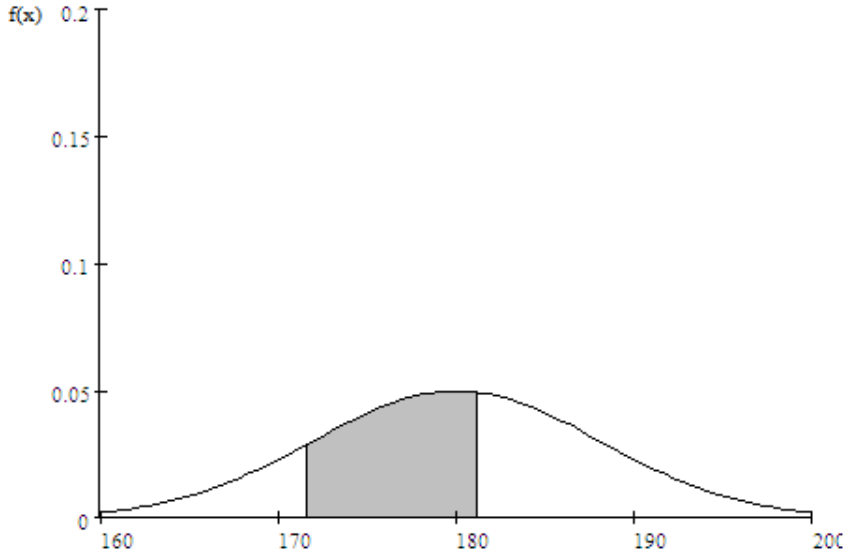
Exempel 10.5.

Antag att populationen utgörs av svenska män över 18 år, vilka observerats med avseende på kroppslängden. Beteckna variabeln kroppslängd med X och antag att den är normalfördelad med $\mu = 180$ cm och $\sigma = 8$. Vi söker sannolikheten att längden av en slumpmässigt vald person i populationen har en kroppslängd mellan 172 cm och 182 cm.

Den sökta sannolikheten representeras av den andel av fördelningsytan som ligger mellan de vertikala linjerna vid $x = 172$ och $x = 182$ i figur 15.3. För att kunna avläsa denna ytandel i tabellen över den standardiserade normalfördelningen transformerar vi X till Z och finner att det z-värde som motsvarar $x = 172$ är

$$z = \frac{x - \mu}{\sigma} = \frac{172 - 180}{8} = -1$$

och det som motsvarar $x = 182$ är på motsvarande sätt $z = 0,25$. Eftersom transformationen endast innebär att man "byter skala" under fördelningen, så är andelen mellan $x = 172$ och $x = 182$ i fördelningen för X densamma som andelen mellan $z = -1$ och $z = 0,25$ i fördelningen för Z.



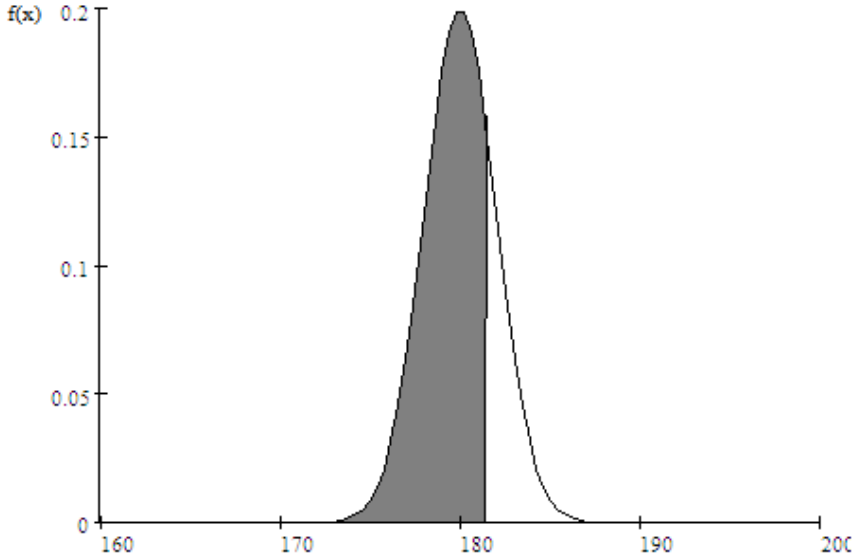
Figur 15.3 Fördelningen av kroppslängd hos svenska män över 18 år.

Från tabellen över den standardiserade normalfördelningen får vi att andelen fördelningsyta till vänster om $z = 0,25$ är 59,9% och den till vänster om $z = -1$ är 15,9%. Kort uttryckt är alltså sannolikheten att $172 \leq X \leq 182$ lika med sannolikheten att $-1 \leq Z \leq 0,25$, dvs

$$P(172 \leq X \leq 182) = P(\frac{172-180}{8} \leq \frac{X-\mu}{\sigma} \leq \frac{182-180}{8})$$

$$= P(-1 \leq Z \leq 0,25) = \Phi(0,25) - \Phi(-1) \approx 0,599 - 0,159 = 0,440.$$

Vi söker nu sannolikheten att medelvärdet $\overline{X}$ för ett slumpmässigt urval om $n = 16$ individer ligger mellan 172 cm och 182 cm. Variabeln $\overline{X}$ är normalfördelad med väntevärdet $\mu_{\overline{X}} = 180$ och standardavvikelsen $\sigma_{\overline{X}} = \sigma/\sqrt{n} = 8/\sqrt{16} = 2$. Se figur 15.4.



Figur 15.4 Samplingfördelningen för $\overline{X}$ då $n = 16$.

Transformationen till den standardiserade normalfördelningen ger nu

$$P(172 \leq \overline{X} \leq 182) = P(\frac{172 - 180}{2} \leq \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} \leq \frac{182 - 180}{2})$$

$$= P(-4 \leq Z \leq 1) = \Phi(1) - \Phi\,(-4) \approx 0.841 - 0.000 = 0.841$$

Ytan mellan $\bar{x} = 172$ och $\bar{x} = 182$ är alltså densamma som ytan mellan $z = -4$ och $z = 1$ i figur 15.4. Observera att sannolikheten att medelvärdet är mellan 172 cm och 182 cm är större än sannolikheten att en enskild observation ligger i samma intervall. Detta beror på att samplingfördelningen för $\overline{X}$ är mer koncentrerad kring sitt medelvärde än fördelningen för X. En jämförelse av figurerna 15.3 och 15.4 ger en grafisk illustration av detta fenomen. ■

Notera att för att kunna utföra z-transformationen måste urvalsmedelvärdets varians, $\sigma^2_{\overline{X}}$, vara känd. I många tillämpningar är så inte fallet. För att klara

av även situationer då $\sigma^2_{\overline{X}}$ är okänd måste vi införa två nya fördelningar, de så kallade χ^2- och t-fördelningarna. Dessa två fördelningar behandlas i följande två avsnitt.

15.3 χ^2-fördelningen

Låt X_1, X_2, ...X_n vara n stycken oberoende normalfördelade stokastiska variabler, alla med väntevärdet μ och variansen σ^2. Vi vet då att variablerna

$$Z_i = \frac{X_i - \mu}{\sigma}, \text{ för } i = 1, 2, ..., n$$

är oberoende och normalfördelade med väntevärdet noll och variansen ett, dvs Z_1, Z_2, ..., Z_n är oberoende standardiserade normalfördelade variabler. Bilda nu en ny stokastisk variabel χ^2 genom att kvadrera de stokastiska variablerna Z_i, dvs beräkna Z_i^2 för $i = 1, 2, ..., n$, och summera kvadraterna

$$\chi^2 = \sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2}$$

.

χ^2 säges då vara en χ^2-fördelad stokastisk variabel med n frihetsgrader. Vi betecknar denna fördelning med symbolen $\chi^2(n)$. Antalet termer i summan, n, kallas fördelningens frihetsgradtal.

χ^2-fördelningen förekommer i flera tillämpningar vi kommer att studera och den spelar en viktig roll i inferensteorin. I figur 15.5 visas χ^2-fördelningen för några olika frihetsgradtal. Observera att den stokastiska variabeln χ^2 inte kan anta negativa värden. Om frihetsgradtalet är större än två, så har täthetsfunktionen ett maximum för värdet $n - 2$. Man kan visa att

$$E(\chi^2) = n,$$

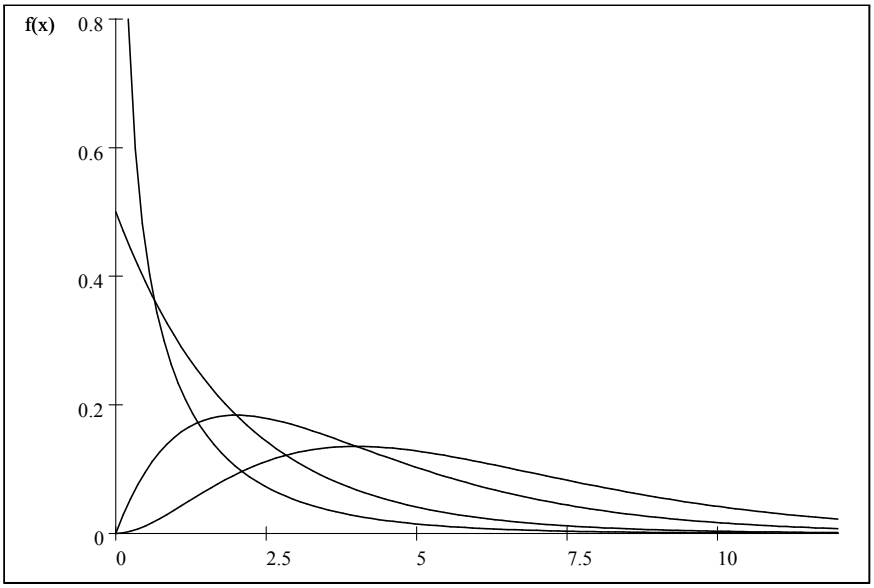
$$V(\chi^2) = 2n.$$

Då frihetsgradtalet ökar, blir χ^2-fördelningen mer symmetrisk. För stora värden på frihetsgradtalet liknar χ^2-fördelningen en normalfördelning. Vi

kan då approximera χ^2-fördelningen med normalfördelningen. Approximationen blir bättre då frihetsgradtalet ökar. Percentiler i χ^2-fördelningen kan fås från tabeller över χ^2-fördelningen eller genom datorprogram som beräknar motsvarande värden. I bilagan till detta kapitel finns en tabell över χ^2-fördelningen.

Exempel 15.6.

Låt χ^2 vara en χ^2-fördelad stokastisk variabel med 10 frihetsgrader. Vi vill bestämma den 95-te percentilen, dvs det tal K som är sådant att $P(\chi^2 \leq K) = 0,95$. Från tabellen över χ^2-fördelningen ser vi att $P(\chi^2 \leq 18,3) \approx 0,95$, dvs det sökta talet är $K \approx 18,3$. ■



Figur 15.5 χ^2-fördelningens täthetsfunktion för frihetsgradtalen n=1, n=2, n=4 och n=6.

Vi visar nu två viktiga tillämpningar av χ^2-fördelningen. Om vi först byter ut väntevärdet μ mot urvalsmedelvärdet $\overline{X}$ i definitionen av χ^2-fördelningen får vi

$$C^2 = \sum_{i=1}^{n} \frac{X_i - \overline{X}}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

där C^2 är en stokastisk variabel. Det verkar rimligt att anta att C^2 är χ^2-fördelad med några frihetsgrader. Så är också fallet. Man kan nämligen visa att C^2 är χ^2-fördelad med $n-1$ frihetsgrader.

Betrakta nu formeln för urvalsvariansen

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

Vi vet att S^2 är en stokastisk variabel, men vilken fördelning har S^2? Låt oss göra ett trick! Multiplicera S^2 med $n-1$, dividera med σ^2 och se vad vi får:

$$\frac{(n-1)\,S^2}{\sigma^2} = \frac{n-1}{\sigma^2} \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

$$= \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \overline{X})^2 = C^2.$$

Till vår stora överraskning och glädje finner vi att den stokastiska variabeln $(n-1)S^2/\sigma^2$ är χ^2-fördelad med $n-1$ frihetsgrader. Därur kan vi också härleda samplingfördelningen för urvalsvariansen då vi har oberoende normalfördelade observationer.

Den andra tillämpningen av χ^2-fördelningen har vi redan stött på i avsnitt 14.3 i samband med analys av modeller för multinomialfördelade stokastiska variabler. Genom möjligheten att approximera binomialfördelningen med normalfördelningen får vi, med samma beteckningar som i avsnitt 14.3, att den stokastiska variabeln

$$Z = \frac{y_1 - \widehat{y}_1}{\sqrt{n\pi_1 (1 - \pi_1)}}$$

är approximativt $N(0,1)$-fördelad, då n är stort. Ur definitionen på χ^2-fördelningen följer det därför att

$$Q_1 = Z^2 = \frac{(y_1 - \widehat{y}_1)^2}{n\pi_1 (1 - \pi_1)} = \frac{(y_1 - \widehat{y}_1)^2}{\widehat{y}_1} + \frac{(y_2 - \widehat{y}_2)^2}{\widehat{y}_2}$$

är approximativt χ^2-fördelad med en frihetsgrad. Det är ingen svårighet att generalisera detta till fallet med ν grupper. Vi får då att

$$Q_{\nu-1} = \sum_{i=1}^{\nu} \frac{(y_i - \widehat{y}_i)^2}{\widehat{y}_i}$$

är approximativt χ^2-fördelad med $\nu - 1$ frihetsgrader, då n är stort. Som tumregel på hur stort n behöver vara för att approximationen skall vara god brukar man kräva att $\hat{y}_i = n\pi_i > 5$ och $n - \widehat{y}_i = n(1 - \pi) > 5$, för $i = 1$, 2,..., ν.

Exempel 15.7.

Låt X vara en stokastisk variabel med en godtycklig, men känd fördelning. Antag att vi gör n stycken oberoende observationer på X och klassindelar observationerna i ν stycken klasser. Bilda ν stycken stokastiska variabler Y_1, $Y_2, ..., Y_\nu$, där Y_i anger antalet observationer som hamnar i den i-te klassen. Eftersom fördelningen för X är känd kan vi beräkna hur många av de n observationerna vi förväntar oss hamna i varje klass, dvs vi kan beräkna $E(Y_1)$, $E(Y_2), ..., E(Y_\nu)$. Vi kan då bilda

$$Q_{\nu-1} = \sum_{i=1}^{\nu} \frac{(Y_i - E(Y_i))^2}{E(Y_i)}$$

och om n är stort, är tydligen $Q_{\nu-1}$ en approximativt χ^2-fördelad stokastisk variabel med $\nu - 1$ frihetsgrader. ■

15.4 t-fördelningen

Låt Z vara en standardiserad normalfördelad stokastisk variabel (så att den har väntevärde noll och varians ett) och låt χ^2 vara en χ^2-fördelad stokastisk variabel med ν frihetsgrader. Antag nu att Z och χ^2 är stokastiskt oberoende av varandra och bilda en ny stokastisk variabel T genom

$$T = \frac{Z}{\sqrt{\chi^2/\nu}}$$

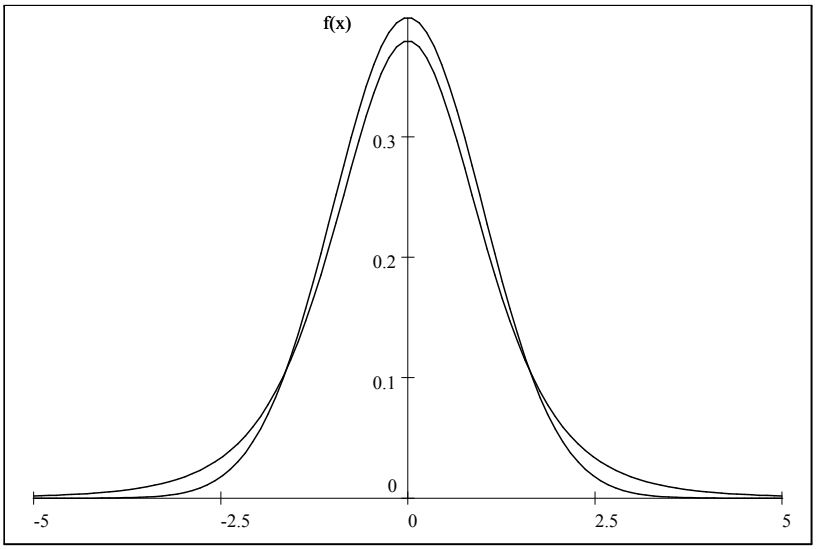
Då säges T vara en t-fördelad stokastisk variabel med ν frihetsgrader. Vi betecknar denna fördelning med $t(\nu)$. I figur 15.6 visas täthetsfunktionen för t-fördelningen med $\nu = 5$ frihetsgrader. I figuren visas även täthetsfunktionen för den standardiserade normalfördelningen. De två fördelningarna ser tämligen lika ut. t-fördelningen har dock "tjockare svansar" än normalfördelningen. Man kan visa att då antalet frihetsgrader ökar blir t-fördelningen mer och mer lik den standardiserade normalfördelningen. t-fördelningen kan då approximeras med normalfördelningen. Normalt brukar man kräva att frihetsgradtalet skall överstiga 30 för att en sådan approximation skall vara god. Percentiler för t-fördelningen kan fås från tabeller eller från datorprogram som beräknar motsvarande värden. I bilagan till detta kapitel finns en tabell över t-fördelningen.

Exempel 15.8.

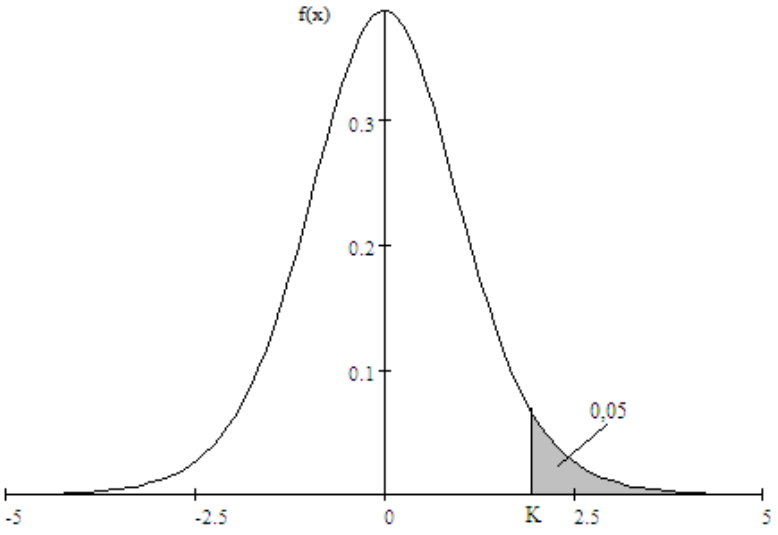
Låt T vara en t-fördelad stokastisk variabel med 10 frihetsgrader. Vi vill bestämma den 95-te percentilen, dvs det tal K som är sådant att $P(T \leq K) = 0,95$.

Ur tabellen över t-fördelningen får vi att $P(T \leq 1,812) \approx 0,95$, dvs vårt sökta tal är $K \approx 1,812$.

Om vi i stället studerar en standardiserad normalfördelad stokastisk variabel Z, så ger $P(Z \leq K) = \Phi(K) = 0,95$ ett annat värde på K, nämligen $K \approx 1,65$. Det större värdet för t-fördelningen beror tydligen på t-fördelningens "tjockare svansar". ■



Figur 15.6 Täthetsfunktionerna för t-fördelningen med $\nu = 5$ frihetsgrader och för den standardiserade normalfördelningen



Figur 15.7 t-fördelningen med 10 frihetsgrader. K är den 95-te percentilen, dvs det tal som är sådant att $P(T \leq K) = 0,95$.

15.5 Samplingfördelningen för urvalsmedelvärdet i en normalfördelad population då variansen är okänd

Enligt diskussionen i avsnitt 15.2 är $Z-$transformationen mycket användbar då vi skall göra inferens om populationsmedelvärdet och populationsvariansen är känd. Om speciellt X_1, X_2, ..., X_n är n stycken oberoende normalfördelade stokastiska variabler alla med väntevärdet μ och variansen σ^2 definieras $Z-$transformationen av

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

där X är, som tidigare, urvalsmedelvärdet. Denna transformation är möjlig att använda endast om populationsvariansen σ^2 är känd. Om den inte är känd kan $Z-$transformationen inte användas. Ett tänkbart alternativ är då att ersätta det okända σ^2 med uppskattningen S^2. Vi får då en s k $t-$transformation, definierad enligt

$$t = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

Att $t-$transformationen har en teoretisk motivering, och inte bara heuristisk, har vi redan sett i avsnitt 14.4. Där visas hur en tillämpning av bl a likelihoodkriterier för test av hypotesen $H_0 : \mu = \mu_0$ mot $H_1 : \mu \neq \mu_0$ leder till ett studium av den stokastiska variabeln

$$F = \left(\frac{\overline{X} - \mu}{S/\sqrt{n}}\right)^2$$

eller, med andra ord, studium av t-transformationen i kvadrat,

$$F = t^2$$

Inferens om μ då σ^2 är okänd kräver sålunda att vi känner sannolikhetsfördelningen för t eller, alternativt, för F. Låt oss utveckla uttrycket för t något,

$$t = \frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{\frac{1}{\sigma}\left(\overline{X} - \mu\right)}{\frac{1}{\sigma}\frac{S}{\sqrt{n}}} = \frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\frac{1}{\sigma}S} = \frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{S^2/\sigma^2}} =$$

$$\frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2(n-1)}}} = \frac{Z}{\sqrt{C^2/\left(n-1\right)}}$$

där $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ är en normalfördelad stokastisk variabel med väntevärdet 0 och variansen 1 och $C^2 = \frac{(n-1)S^2}{\sigma^2}$ är, som diskuterades i avsnitt 15.3, en χ^2-fördelad stokastisk variabel med $n-1$ frihetsgrader. Uttryckt på detta sätt ser vi att $t-$transformationen inte är någonting annat än en standardiserad normalfördelad stokastisk variabel dividerad med kvadratroten ur en χ^2-fördelad stokastisk variabel som i sin tur är dividerad med sitt frihetsgradtal. Vidare är det möjligt att visa att Z och C^2 är stokastiskt oberoende av varandra. Därmed följer det att t är $t-$fördelad med $n-1$ frihetsgrader.

Det är här inressant att jämföra uttrycket för t med uttrycket för Z-transformationen vi utför i det fall då σ^2 är känd. De båda uttrycken är identiska så när som på att det okända σ^2 har ersatts med dess uppskattning s^2. t-fördelningens tjockare svansar är det pris vi får betala för att inte känna σ^2, dvs vi får en samplingfördelning som är något mer utbredd (har större varians). Vi vet också att t-fördelningen blir mer och mer lik normalfördelningen då frihetsgradtalet ökar. Detta innebär att då antalet observationer ökar blir fördelningen för t mer och mer lik den standardiserade normalfördelningen och skillnaden mot det fall då σ^2 är känd blir mindre och mindre. Heuristiskt kan vi tolka detta som att ju fler observationer vi gör, ju bättre kan S^2 uppskatta σ^2 och t-transformationen blir mer och mer lik en Z-transformation.

Exempel 15.9.

Antag att vi har en modell som säger att längden av värnpliktiga är normalfördelad med det okända väntevärdet μ och den okända variansen σ^2. Antag vidare att vi gör $n = 4$ oberoende observationer. Om vi får ett urval

med $s^2 = 11,56$, vad är då sannolikheten att $\overline{X} \leq \mu + 4$ cm? Eftersom $\mu_{\overline{X}} = \mu$ gäller det att

$$P(\overline{X} \leq \mu + 4) = P(\overline{X} - \mu \leq 4) = P(\frac{\overline{X} - \mu_{\overline{X}}}{S/\sqrt{n}} \leq \frac{4}{\sqrt{11,56/4}})$$

$$\approx P(t \leq 2,35) \approx 0,95.$$

där t är $t-$fördelad med $n - 1 = 3$ frihetsgrader. ■

15.6 Centrala gränsvärdessatsen

I de föregående avsnitten redogör vi för samplingfördelningen för urvalsmedelvärdet då observationerna dras ur en normalfördelad population. I detta avsnitt redogör vi för samplingfördelningen för urvalsmedelvärdet då observationerna inte nödvändigtvis är normalfördelade. Vi stöder oss då på ett resultat som brukar kallas Centrala gränsvärdessatsen. Den säger oss att

Formen på samplingfördelningen för en Z-transformation av urvalsmedelvärdet närmar sig formen för en standardiserad normalfördelning då urvalsstorleken ökar.

Centrala gränsvärdessatsen gäller allmänt oberoende av formen på den fördelning urvalet hämtas ur. Ett viktigt krav är dock att variansen måste vara ändligt stor (dvs $\sigma^2 < \infty$).

Exempel 15.10.

Antag att vi har en mycket stor population bestående av familjer där den intressanta egenskapen är antalet barn per familj. Vi definierar en variabel X som antal barn per familj. Fördelningen för X framgår av figur 15.8 respektive tabell 15.2.

För denna population gäller tydligen att populationsmedelvärdet är

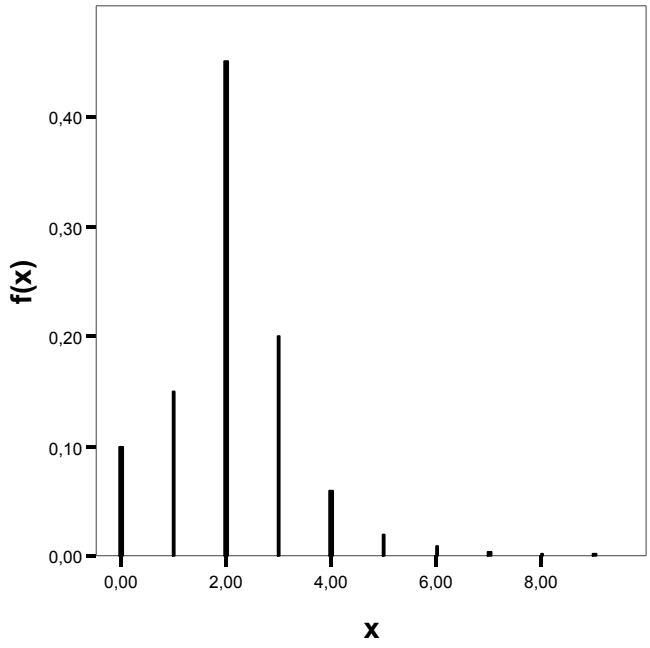
$$\mu = \sum x \ f(x) = 2,1275$$

och populationsvariansen är

$$\sigma^2 = 6,0775 - 2,1275^2 \approx 1,5512$$

med populationsstandardavvikelsen

$$\sigma \approx \sqrt{1,5512} \approx 1,2455.$$



Figur 15.8 Fördelningen för antal barn per familj.

Tabell 15.2 Fördelningen för antal barn per familj.

x	$f(x)$	$xf(x)$	$x^2f(x)$
0	0,10	0,00	0,00
1	0,15	0,15	0,15
2	0,45	0,90	1,80
3	0,20	0,60	1,80
4	0,06	0,24	0,96
5	0,02	0,10	0,50
6	0,01	0,06	0,36
7	0,005	0,035	0,245
8	0,0025	0,020	0,160
9	0,0025	0,0225	0,2025
Summa	1,0000	2,1275	6,0775

Låt oss nu tänka oss att vi drar alla tänkbara urval med återläggning av storlek $n = 50$ ur populationen och för varje urval beräknar urvalsmedelvärdet. De värden på urvalsmedelvärdet $\overline{X}$ vi då får utgör samplingfördelningen för urvalsmedelvärdet. Vi vet att väntevärdet för $\overline{X}$ är lika med populationens väntevärde

$$\mu_{\overline{X}} = \mu = 2,1275$$

och dess standardavvikelse är lika med populationsstandardavvikelsen dividerat med kvadratroten ur urvalsstorleken

$$\sigma_{\overline{X}} = \sigma/\sqrt{n} \approx 1,2455\sqrt{50} \approx 0,1761. \blacksquare$$

Enligt Centrala gränsvärdessatsen är samplingfördelningen för $Z = \left(\overline{X} - \mu_{\overline{X}}\right)/\sigma_{\overline{X}}$ approximativt normalfördelad med väntevärdet 0 och variansen 1. Detta innebär att vi kan använda den standardiserade normalfördelningen för att approximativt beräkna sannolikheter om urvalsmedelvärdet $\overline{X}$.

Centrala gränsvärdessatsen förutsätter eller kräver inte någonting om populationens form. I exemplet ovan är ju fördelningen vad man kallar sned. Allmänt gäller att:

- Om den population man drar urvalet ur är normalfördelad, kommer samplingfördelningen för urvalsmedelvärdet $\overline{X}$ att vara en normalfördelning, oberoende av urvalsstorleken.

- Ju större skillnader populationens fördelning företer gentemot en normalfördelning, desto större urval krävs för att transformationen $Z = \left(\overline{X} - \mu_{\overline{X}}\right)/\sigma_{\overline{X}}$ skall kunna betraktas som en approximativt normalfördelad stokastisk variabel.

- I regel krävs att urvalsstorleken är större än 30 för att Centrala gränsvärdessatsen skall kunna tillämpas.

Exempel 15.10 (*fortsättning*).

Vi söker nu sannolikheten att $\overline{X} \leq 2,25$. Antag först att populationsvariansen σ^2 är känd, dvs vi vet att $\sigma = 1,2455$ så att $\sigma_{\overline{X}} = 0,1761$. Då är

$$P(\overline{X} \leq 2,25) = P(\frac{\overline{X} - \mu}{\sigma_{\overline{X}}} \leq \frac{2,25 - 2,1275}{0,1761}) \approx P(Z \leq 0,70)$$

Enligt centrala gränsvärdessatsen kan vi approximera fördelningen för Z med den standardiserade normalfördelningen. Vi får därför

$$P(\overline{X} \leq 2,25) \approx \Phi(0,70) \approx 0,7580$$

Om å andra sidan σ^2 vore okänd, men uppskattades till $s = 1,2455$, skulle detta inte påverka våra slutsatser. Vi skulle uppskatta standardavvikelsen för $\overline{X}$ med $s_{\overline{X}} = 0,1761$ och kvantiteten $Z = (\bar{X} - \mu_{\overline{X}})/s_{\overline{X}}$ skulle, eftersom urvalet är så pass stort ($n = 50$), enligt centrala gränsvärdessatsen vara approximativt normalfördelad. Den sökta sannolikheten approximeras således även i detta fall med $\Phi(0,70) \approx 0,7580$. ■

Som en följd av centrala gränsvärdessatsen kan man approximera samplingfördelningen för $\overline{X}$ med en normalfördelning med väntevärdet $\mu_{\overline{X}}$ och variansen $\sigma^2_{\overline{X}}$ då urvalsstorleken är stor. På motsvarande sätt är det möjligt att approximera samplingfördelningen för summan av observationerna med en normalfördelning. Detta följer av att medelvärdet är lika med summan av

observationerna, dividerat med antalet observationer, så att summan T fås som antalet observationer multiplicerat med medelvärdet

$$T = n\overline{X} = \sum_{i=1}^{n} X_i$$

Samplingfördelningen för T har därför karakteristikorna

$$\mu_T = E(T) = E(n\overline{X}) = nE(\overline{X}) = n\mu$$

$$\sigma_T^2 = V(T) = V(n\overline{X}) = n^2 V(\overline{X}) = n^2\sigma^2/n = n\sigma^2$$

I vissa fall kan Centrala gränsvärdessatsen tillämpas även på T, så att T kan approximeras med en normalfördelad stokastisk variabel.

15.7 Samplingfördelningen för proportioner

I många tillämpningar är man intresserad av proportioner, t ex proportionen pojkar bland nyfödda barn och proportionen röstberättigade som sympatiserar med ett visst politiskt parti. I det här avsnittet diskuterar vi samplingfördelningen för observerade proportioner i urval.

Antag att vi söker proportionen av en egenskap A i en population. Utför det slumpmässiga försöket "välj slumpmässigt ut en individ ur populationen och notera om individen har egenskapen A eller ej". Till varje upprepning av detta försök definierar vi en stokastisk variabel X_i genom

$$X_i = \left\{ \begin{array}{l} 1 \text{ om individ } \ i \text{ har egenskap } A \\ 0 \text{ om individ } \ i \text{ inte har egenskap } A \end{array} \right\}$$

Om populationen är oändlig eller om den är ändlig och vi gör dragningarna med återläggning blir urvalet X_1, X_2, ..., X_n oberoende stokastiska variabler med

$$E(X_i) = \pi$$

$$V(X_i) = \pi(1-\pi)$$

där π är proportionen A i populationen.

Bilda nu urvalsproportionen P genom

$$P = \frac{1}{n}\sum_{i=1}^{n} X_i$$

.

Eftersom P är ett urvalsmedelvärde följer ur Centrala gränsvärdessatsen att P är approximativt normalfördelad då n är stort. För att approximationen skall vara god brukar man i allmänhet kräva att $n\pi \geq 5$ och $n(1-\pi) \geq 5$. Slutligen följer det att

$$E(P) = \pi$$

$$V(P) = \pi(1-\pi)/n.$$

Exempel 15.10 (*fortsättning*).

Låt A vara egenskapen "barnfamilj" och låt P vara den stokastiska variabeln "proportionen barnfamiljer i ett slumpmässigt urval om $n = 50$ familjer". Vi ser att proportionen barnfamiljer i den studerade populationen är $\pi = 0.90$. Med $n = 50$ får vi att $n\pi = 45$ och $n(1-\pi) = 5$, dvs enligt Centrala gränsvärdessatsen kan vi förvänta oss att normalfördelningen är en god approximation av samplingfördelningen för P. Väntevärdet och variansen i samplingfördelningen för P är

$$E\left(P\right) = \pi = 0,90$$

$$V\left(P\right) = \frac{\pi\left(1-\pi\right)}{n} = \frac{0,90\left(1-0,90\right)}{50} = 0,0018.$$

Sannolikheten att P ligger i intervallet [0,85, 0,95] är

$$P(0,85 \leq P \leq 0,95)$$

$$= P(\frac{0,85 - 0,90}{\sqrt{0,0018}} \leq \frac{P - \pi}{\sqrt{\pi\,(1-\pi)\,/n}} \leq \frac{0,95 - 0,90}{\sqrt{0,0018}})$$

$$\approx P(-1,18 \leq Z \leq 1,18).$$

Enligt centrala gränsvärdessatsen är Z approximativt normalfördelad med väntevärdet noll och variansen ett, så att

$$P(0,85 \leq P \leq 0,95) \approx \Phi(1,18) - \Phi(-1,18)$$

$$\approx 0,8810 - 0,1190 \approx 0,7620.$$

Övningar

15.1 Man gör en serie på n stycken kast med en symmetrisk tärning. Hur stor är sannolikheten att relativa frekvensen sexor skall ligga i intervallet $\frac{1}{6} \pm \frac{1}{18}$ (inklusive intervallgränserna) ifall

a) $n = 20$

b) $n = 40$

c) $n = 80$

15.2 Ett företag tillverkar och säljer pappersark i balar om 5000 ark. Vikten av ett 1 m^2 stort pappersark kan anses normalfördelad med väntevärdet 80 gram och standardavvikelsen 4 gram. Om balens vikt understiger 399,5 kg kommer en köpare att reklamera balen.

a) Vad är sannolikheten att företaget får en reklamation.

b) Företaget har nyligen sålt 1000 balar. Hur många balar kan man förvänta sig kommer att reklameras?

c) Ett ark kostar företaget 30 öre att tillverka. Arket säljs för 40 öre. Om en bal reklameras säljs den till det reducerade priset 1200 kr. Beräkna förväntad vinst per bal för företaget.

d) Lönar det sig för företaget att sänka genomsnittsvikten på ett pappersark till 79,8 gram (standardavvikelsen oförändrad)? Kostnaden för företaget att producera ett ark skulle då sjunka till 28 öre.

15.3 En skidlift är dimensionerad för max 100 personer och klarar av en totalvikt av 8200 kg (max vikt för de 100 personerna). Om medelvikt och standardavvikelse är 79 resp 14 kg, vad är sannolikheten att den maximalt tillåtna vikten överskrids då 100 personer åker samtidigt i liften?

15.4 En stokastisk variabel X har okänt väntevärde, μ, och variansen $\sigma^2 = 4$. Hur stora stickprov skall tas för att sannolikheten åtminstone skall vara 0,95 för händelserna A: "Urvalsmedelvärdet avviker från μ med högst 0,1" och B : "Urvalsmedelvärdet avviker från μ med högst 10% av standardavvikelsen för X" skall inträffa.

15.5 En metod att bedöma kvaliteten hos papper är att fastställa en undre gräns för papperets dragstyrka. Om dragstyrkan understiger ett på förhand bestämt värde, C, måste papperet säljas som "lågkvalitetspapper" och därmed till ett lägre pris. Följande beslutsregel gäller: För en pappersrulle tas 10 prover. Om minst tre av dessa har en dragstyrka understigande C enheter, klassificeras pappersrullen som "lågkvalitet". Om produktionen är under kontroll kan dragstyrkan betraktas som en normalfördelad stokastisk variabel med väntevärde 750 och varians 100. Den undre gränsen, C, för dragstyrkan har fastställts till 737,2. Beräkna sannolikheten att en slumpmässigt vald pappersrulle klassificeras som "lågkvalitet" även om produktionen är under kontroll.

15.6 Ett mindre trafikflygplan kan ta fyra passagerare utom piloten. Säkerhetstesten säger att flygplanet får starta om de fyra passagerarnas samman-

lagda vikt ej överstiger 325 kg. Antag att passagerarna kan anses slumpmässigt valda från en normalfördelad population med väntevärde 75 kg och standardavvikelse 16 kg. Vad är sannolikheten att planet får startförbud på grund av övervikt vid ett tillfälle då fyra passagerare medföljer planet?

15.7 Ett läkemedelsföretag har ett mycket stort parti tabletter. Vikten (i gram) av en slumpmässigt vald tablett kan med god noggrannhet anses vara en observation på en normalfördelad stokastisk variabel med väntevärde μ och standardavvikelse 0,02 g. För kontroll av vikten tar man ut ett antal tabletter och väger dem. Antag att $\mu = 0,65$ g.

a) Beräkna sannolikheten att vikten av en slumpmässigt vald tablett ligger i intervallet (0,64, 0,66).

b) Beräkna sannolikheten att medelvärdet av vikten av 30 slumpmässigt valda tabletter ligger utanför intervallet (0,64, 0,66).

c) Hur många tabletter bör man väga om man vill att sannolikheten skall vara högst 0,05 att medelvärdet kommer att ligga utanför intervallet (0,64, 0,66).