
11. DESKRIPTION — EN VARIABEL

11.1 Inledning

I detta och nästa två kapitel introduceras en enkel typ av dataanalys kallad deskription. Deskription innebär att mer informellt presentera en observerad empirisk fördelning eller att jämföra den med en teoretisk sannolikhetsfördelning. Det finns en stor spännvidd mellan olika ansatser till deskriptioner av datamaterial. Speciellt kan kopplingen till teorier om problemområdet vara mer eller mindre uttalat i olika ansatser. Självklart kan en deskription inte vara helt ”teorilös”. Redan valet av variabler, valet av jämförelser och valet av presentationssätt innebär ett teoribasrat ställningstagande.

Vi har tidigare konstaterat att empiriska fördelningar bildlikt motsvarar modellvärldens sannolikhetsfördelningar. De tabelleringstekniker och grafiska tekniker som kommer till användning vid deskription är därför nära släkt med motsvarande tekniker som används för att presentera och jämföra sannolikhetsfördelningar.

För att begränsa diskussionen berör vi endast några av de enklaste fallen. Detta kapitel behandlar det fall då vi gör oberoende observationer på en stokastisk variabel medan nästa kapitel behandlar det fall då vi gör oberoende observationer på flera stokastiska variabler. Antagandet om oberoende observationer är särskilt vanligt då materialet består av $s \times k$ tvärsnittsdata, dvs uppgifter som avser flera olika individer vid samma tidpunkt eller samma tidsperiod. Däremot kan antagandet om oberoende observationer ofta ifrågasättas om materialet består av $s \times k$ tidsseriedata, d v s uppgifter som avser en följd av tidpunkter eller perioder. Kapitel 13 behandlar några metoder för att analysera en tidsmässig utveckling av en variabel. Detta fall kan ofta tolkas som observationer på en följd av beroende stokastiska variabler. Efter de tre deskriptionskapitlen påbörjas en diskussion om mer formell analys av data.

11.2 Tabeller

Det första steget i en statistisk analys av data består ofta i att sammanfatta datamaterialet i form av den empiriska fördelningsfunktionen eller i en empirisk frekvensfunktion. Dessa funktioner kan i sin tur redovisas och framställas på många olika sätt. Vanligen redovisas de empiriska funktionerna i form av tabeller och diagram. Det här avsnittet är ägnat åt att redovisa några enkla tabelleringsprinciper, medan avsnitt 11.3 ägnas åt den grafiska framställningen av empiriska fördelnings- och frekvensfunktioner.

Om variabeln är *kvalitativ* är det normalt inga problem att konstruera tabeller. Man anger helt enkelt hur många observationer som finns i varje klass (kategori). På så vis får vi en tabell som beskriver den empiriska frekvensfunktionen. En sådan tabell kallas en (empirisk) frekvenstabell. Om det t ex finns 35 arbetare, 10 tjänstemän och 5 försäljare på ett företag med 50 anställda får vi den empiriska frekvensfunktionen för variabeln personalkategori redovisad i den *envägsindelade frekvenstabellen* 11.1.

Tabell 11.1 Anställda vid X-företaget den 1 jan 2008, fördelade på personalkategorier

Personalkategori	Antal
Arbetare	35
Tjänstemän	10
Försäljare	5
Summa	50

Källa: X-företagets personalförteckning

Tabell 11.1 innehåller *absoluta frekvenser*, dvs antal anställda. Variabelns fördelning kan också beskrivas med hjälp av *relativa frekvenser*, dvs proportioner anställda. Relativa frekvenser beräknas genom att dividera frekvenserna i respektive kategori med totala antalet observationer. Detta visas i tabell 11.2. Procenttal är ofta lättare att tolka än antal, t ex får vi en bättre uppfattning av fördelningen av anställda på personalkategorier om vi får veta att det finns 70% arbetare i ett företag, än om vi får veta att det finns 35

arbetare i företaget. Observera här hur de relativa frekvenserna har egenskapen att de summerar till ett (100%) liksom fallet är för sannolikheter i sannolikhetsfördelningar.

Tabell 11.2 Anställda vid X-företaget den 1 jan 2008, fördelade på personalkategorier

Personalkategori	Procent
Arbetare	70
Tjänstemän	20
Försäljare	10
Summa	100

Antal anställda: 50

Källa: X-företagets personalförteckning

Om antalet observationer som utgör basen för procentberäkningarna är litet, kommer varje observation att representeras av många procentenheter. En antalsmässigt liten skillnad mellan olika kategorier kan i sådana fall ge upphov till stora procentuella skillnader. När man redovisar en tabell med relativa frekvenser är det därför viktigt att även ange det totala antalet observationer.

Det är viktigt att en tabell tydligt anger vad den visar. Tabellens rubrik skall ge en allmän beskrivning av vad den presenterar. Om den variabel som presenteras kan tänkas förändras över tiden är det också viktigt att ange vilken tidpunkt observationerna avser. Tabellens kolumner skall identifiera variablerna och tydligt ange i vilken måtenhet de är mätta.

Kategorierna hos en kvalitativ variabel har ingen naturlig ordning. De kan därför i princip anges i vilken ordning som helst i en frekvenstabell. Det är dock ofta lämpligt att ange kategorierna i fallande frekvensordning, dvs kategorin med den högsta frekvensen anges först, därefter kategorin med den näst högsta frekvensen osv.

Grundprincipen vid tabellering av kvantitativa variabler är liksom för kvalitativa variabler, att man redovisar den empiriska frekvensfunktionen i form av en envägsindeldad frekvenstabell. Kvantitativa variabler har dock en naturlig ordning. Observationerna ordnas därför lämpligen i *storleksordning* så att

frekvensen för det minsta värdet anges först, därefter frekvensen för det näst minsta värdet, o s v.

Det enklaste fallet av frekvenstabell får vi då variabern är diskret och endast antar ett fåtal värden. Ett exempel på detta är observationer på matematikbetyget i ett urval om 40 studenter. De observerade värdena är

4 3 5 4 1 2 4 4 3 3 2 3 3 3 3 2 4 2 3 5 5 1 3 3 5 3 3 3 4 2 2 5 2 4 3 5 3 4 2 3

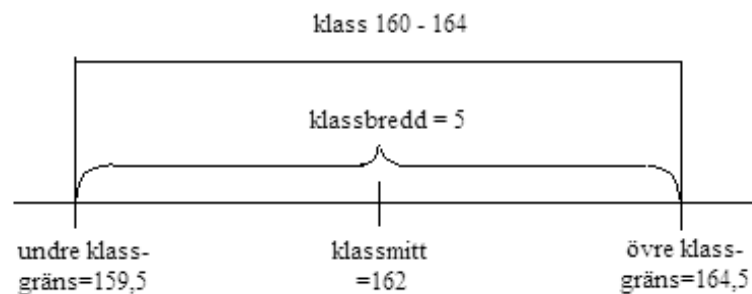
vars empiriska frekvensfunktion redovisas i tabell 11.3. I tabell 11.3 visas både de absoluta och de relativa frekvenserna, dvs tabellen är sammansatt av två envägsindelade frekvenstabeller.

Tabell 11.3 Fördelning av matematikbetyg hos 40 slumpvis utvalda studenter.

Betyg	Antal	Procent
1	2	5
2	8	20
3	16	40
4	8	20
5	6	15
Summa	40	100

Om variabeln är diskret och antar många värden eller om den är kontinuerlig delar man upp variationsområdet i ett lämpligt antal intervall eller *klasser*. Om t ex variabeln är kroppslängd och har mätts upp i hela centimeter kan variationsområdet delas in i klasserna ..., 155-159 cm, 160-164 cm, 165-169 cm,... I modellvärlden är variabeln kroppslängd en kontinuerlig variabel så att vi kan tänka oss personer med en längd som faller mellan klasserna. T ex kan det finnas personer med en längd mellan 159 cm och 160 cm eller mellan 164 cm och 165 cm osv. Om avrundningarna vid mätningarna är gjorda så att värden mellan 159,5 cm och 160,0 cm höjs till 160 cm kommer klassen 160-164 cm att innehålla mätvärden som överstiger 159,5 cm men inte att innehålla några längder som understiger 159,5 cm. Man säger då att klassens *undre klassgräns* är 159,5 cm. Om på samma sätt värden mellan 164,0 cm och 164,5 cm sänks till 160 cm kommer klassen 160-164 cm att innehålla mätvärden som understiger 164,5 cm men inte att innehålla längder som överstiger 164,5

cm. Klassens *övre klassgräns* är då 164,5 cm. Klasser som har en fixerad övre och en undre klassgräns kallas *slutna klasser*. Skillnaden mellan den övre och den undre klassgränsen kallas *klassbredden*. I vårt exempel med kroppslängder är klassbredden 5 cm för de klasser vi har diskuterat hittills. När man gör en klassindelning är det viktigt att specificera klasserna så att varje mätvärde hamnar i exakt en klass. Detta kan uttryckas så att klasserna skall vara ömsesidigt uteslutande, dvs ett mätvärde kan inte tillhöra två klasser, och klasserna skall vara uttömmande, dvs varje mätvärde skall hamna i någon klass. Om den fortsatta bearbetningen skall basera sig på det klassindelade materialet brukar man specificera ett värde för varje klass som representant för klassen. Man väljer då ofta *klassmitten* i varje klass. Att arbeta enbart med klassmitten i stället för de enskilda observationerna innebär givetvis en smärre approximation som får vägas mot den förenkling man åstadkommer. Om observationerna inom respektive klass är någorlunda jämnt fördelade inom klassen är approximationsfelet i de flesta tillämpningar förhållandevis litet. I vårt exempel med kroppslängder ligger klassmitten för klassen 160-164 sålunda mitt emellan 159,5 cm och 164,5 cm, dvs 162 cm. Dessa begrepp illustreras i figur 11.1. En fullständig envägsindeldad frekvenstabell över kroppslängdsmaterialet sedan det klassindelats visas i tabell 11.4.



Figur 11.1 Principskiss över klassindelning.

Tabell 11.4 Kroppslängd hos 1000 värnpliktiga

Kroppslängd	Klassgränser	Klassmitt	Antal
-154	-154.5		3
155 - 159	154.5-159.5	157	32
160 - 164	159.5-164.5	162	83
165 - 169	164.5-169.5	167	113
170 - 174	169.5-174.5	172	233
175 - 179	174.5-179.5	177	294
180 - 184	179.5-184.5	182	167
185 - 189	184.5-189.5	187	52
190 - 194	189.5-194.5	192	21
195 -	194.5-		2

Vi lägger här märke till att tabellen innehåller två *öppna klasser*, dvs endast en klassgräns är fixerad. Öppna klasser används då det inte finns någon naturlig övre eller undre klassgräns. Givetvis finns det ingen klassmitt i öppna klasser och det är inte heller trivialt att hitta någon annan representant för en öppen klass. Både avsaknaden av klassrepresentant och en klassgräns i öppna klasser kan medföra tolkningsproblem.

Tabell 11.5 Kroppslängd hos 1000 värnpliktiga

Kroppslängd	Klassgränser	Klassmitt	Antal
- 159	-159.5		35
160-169	159.5-169.5	164,5	196
170-174	169.5-174.5	172	233
175-189	174.5-189.5	182	513
190 -	189.5-		23

I tabell 11.5 redovisas en frekvenstabell över kroppslängdsmaterialet, men nu basedad på en annan klassindelning. Vi lägger här märke till att tabellerna 11.4 och 11.5 ger två olika bilder av samma fördelning och att denna skillnad beror på att variabeln har klassindelats på två olika sätt. Valet av antal klasser och klassgränser kräver eftertanke och gott omdöme. Då man väljer antal klasser måste man bland annat kompromissa mellan de motstridiga

kraven att mycket information kräver många klasser vilket ger svårläst tabell och en lättläst tabell vilket kräver få klasser och ger lite information

I vissa situationer kan man välja klassgränser så att man får lika breda klasser medan det i andra situationer är lämpligare med olika breda klasser. Då man t ex studerar åldersfördelningar kan det ofta vara lämpligt att använda åldersklasserna 0-6 år, 7-15 år, 16-64 år och över 65 år. Det viktigaste kravet, som vi aldrig får rucka på, är att utforma tabellerna så att de ger den information som är relevant för undersökningsproblemet.

Alla tabeller i detta avsnitt har varit *frekvenstabeller*. Gemensamt för sådana tabeller är att de består av en *summa* (här frekvenser eller relativa frekvenser) som har delats upp i ett antal delposter.

Övning 11.1

Vid antagning till en viss utbildning fick 32 sökande göra ett kunskapsprov. Testdeltagarna fick följande poäng på provet: 34, 38, 21, 39, 32, 24, 52, 41, 31, 44, 46, 39, 29, 49, 40, 38, 29, 48, 42, 18, 41, 49, 31, 39, 41, 28, 51, 35, 45, 43, 42, 44.

- a) Gör en klassindelning av materialet och redovisa provresultaten i en frekvenstabell.
- b) Flytta klassgränserna tre steg neråt, dvs minska klassgränserna med tre enheter. Vad händer? Hur påverkas fördelningens form?

Övning 11.2

Samla in uppgifter på kön, längd och antal syskon hos dina kamrater i gruppen. Gör lämpliga frekvenstabeller för de tre variablerna.

Övning 11.3

För att testa ut prestanda för en ny bilmodell har ett urval om 49 bilar provkörts och bensinförbrukningen har uppmätts. Följande mätvärden (liter per mil) erhöles:

0,76 0,74 0,78 0,74 0,73 0,71 0,75 0,76 0,74 0,73 0,76 0,77 0,72 0,78 0,75 0,73
0,72 0,74 0,77 0,77 0,74 0,75 0,76 0,72 0,74 0,75 0,73 0,76 0,79 0,76 0,73 0,72
0,75 0,77 0,74 0,75 0,73 0,75 0,74 0,77 0,72 0,75 0,74 0,74 0,72 0,75 0,74 0,73

Gör en lämplig tabell över bilarnas bensinförbrukning.

Övning 11.4

En tillverkare av sopsäckar prövar säckarnas hållfasthet genom att lasta 40 slumpmässigt valda säckar tills hantagen går sönder eller säckarna spricker. Vikten innan säcken går sönder noteras varvid följande observationer (kg) erhöles:

48,5 52,3 53,5 50,5 50,3 49,6 51,0 48,3 50,6 50,2 52,5 47,5 50,9 49,8 50,0 50,8
53,0 50,9 49,9 50,1 50,7 48,2 51,5 49,0 51,7 53,2 51,1 52,6 51,2 49,5 49,4 51,9
52,0 48,8 46,8 51,3 49,3 54,0 49,2 51,4

Gör en lämplig tabell över sopsäckarnas hållfasthet.

11.3 Grafisk framställning

I detta avsnitt presenterar vi några enkla principer för att grafiskt redovisa empiriska frekvens- och fördelningsfunktioner. Den grafiska framställningen är i form av diagram vars syfte i första hand är att förmedla ett helhetsintryck av den empiriska fördelningen. Liksom vid tabellkonstruktionen är det viktigt att skilja mellan olika typer av variabler även vid diagramkonstruktion.

Kvalitativa variabler illustreras som regel med *ytdiagram* bestående av geometriska figurer såsom rektanglar, kvadrater eller cirklar. I ytdiagrammet låter man ytan av de geometriska figurerna svara mot frekvenserna. De två vanligaste typerna av ytdiagram är *stapeldiagram* och *cirkeldiagram*.

Staplarna i ett stapeldiagram kan delas upp eller grupperas. Då man använder uppdelade staplar svarar stapelns hela area mot hela materialet. Stapeln delas sedan in så att de olika delarnas areor svarar mot de olika kategoriernas frekvenser. Då man använder grupperade staplar får varje kategori sin egen stapel vars area svarar mot kategoriernas frekvens. Eftersom staplarna är lika breda blir arean proportionell mot staplarnas höjd, dvs stapelns höjd får

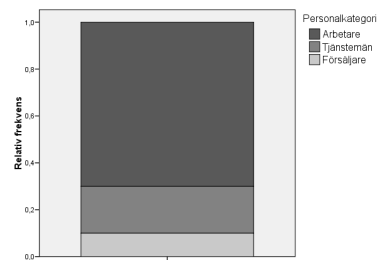
representera frekvensen. I figur 11.2 visas exempel på en uppdelad stapel och på grupperade staplar för ett datamaterial i föregående avsnitt.

Principen för *cirkeldiagrammet* är att en cirkel delas upp i olika cirkelsektorer och arean av varje sektor svarar mot frekvensen för en kategori. Om exempelvis en kategori omfattar 70% av materialet (som för kategorin “Arbetare” i exemplet ovan), blir medelpunktsvinkeln $0,70 \cdot 360 = 252$ grader för motsvarande sektor. I figur 11.2 c visas ett exempel på ett cirkeldiagram.

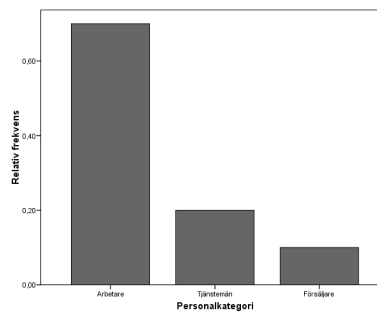
De grundläggande reglerna för diagramritning kan utvidgas och varieras på en mängd olika sätt. Tex kan diagrammen bli mer tilltalande och få en ökad slagkraft om de ritas i tre dimensioner i stället för två. Om man använder en tredje dimension måste man dock komma ihåg att frekvenserna då svarar mot volymer av olika geometriska figurer och att det kan vara svårt att uppfatta storleksförhållanden mellan olika figurer. Även om denna diagramtyp kan sägas ha ett tilltalande utseende bör man alltså göra klart för sig att läsaren relativt sett har svårare för att göra direktavläsningar i ett sådant diagram.

För kvantitativa variabler är det ofta lämpligt att använda andra typer av diagram. Om variabeln är diskret och endast antar ett fåtal variabelvärden, använder man ett *stolpdiagram* för att illustrera den empiriska frekvensfunktionen. För varje variabelvärde finns en stolpe vars höjd är proportionell mot variabelvärdets frekvens. I figur 11.3 visas ett stolpdiagram över fördelningen av matematikbetyg hos 40 slumpvist valda studenter i tabell 11.3.

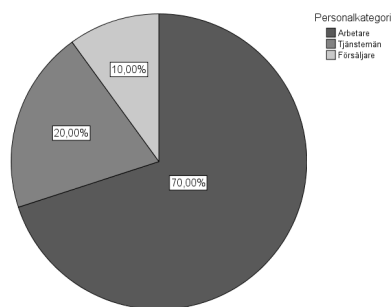
Om undersökningsvariabeln är klassindeldad åskådliggörs den empiriska frekvensfunktionen med ett *histogram*. I ett histogram representeras varje klass av en stapel vars area är proportionell mot frekvensen i klassen och vars bas motsvaras av klassbredden. I figur 11.4 visas ett histogram för den empiriska fördelningen för variabeln kroppslängd enligt tabell 11.4. Förutom de två öppna klasserna har alla klasserna samma klassbredd. Det medför att stolparna får lika bredd och därmed blir staplarnas höjder proportionella mot klassfrekvenserna. De öppna klassernas frekvenser markeras med streckade linjer. Lägg också märke till att variabelaxeln är stympad. Stympningen görs av utrymmesskäl och våglinjen markerar att skalan inte utgår från noll.



a) Uppdelad stapel

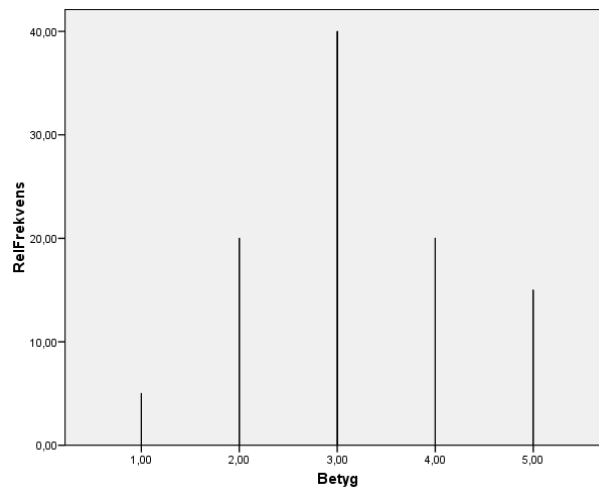


b) Grupperade staplar

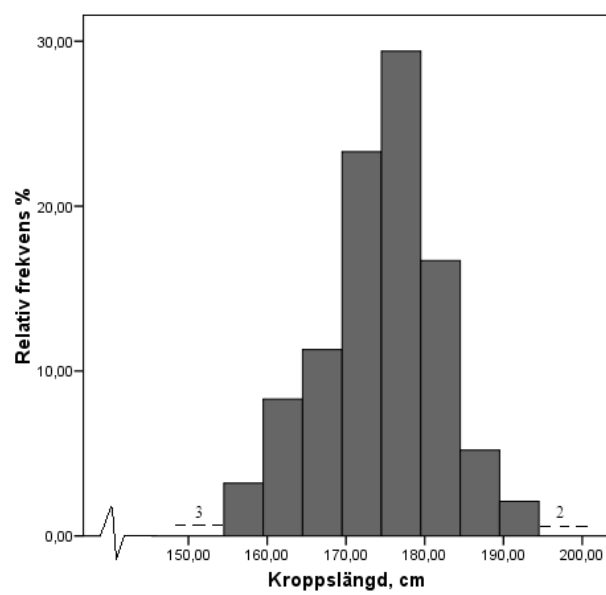


c) Cirkeldiagram

Figur 11.2 Anställda fördelade på personalkategorier enligt tabell 11.1

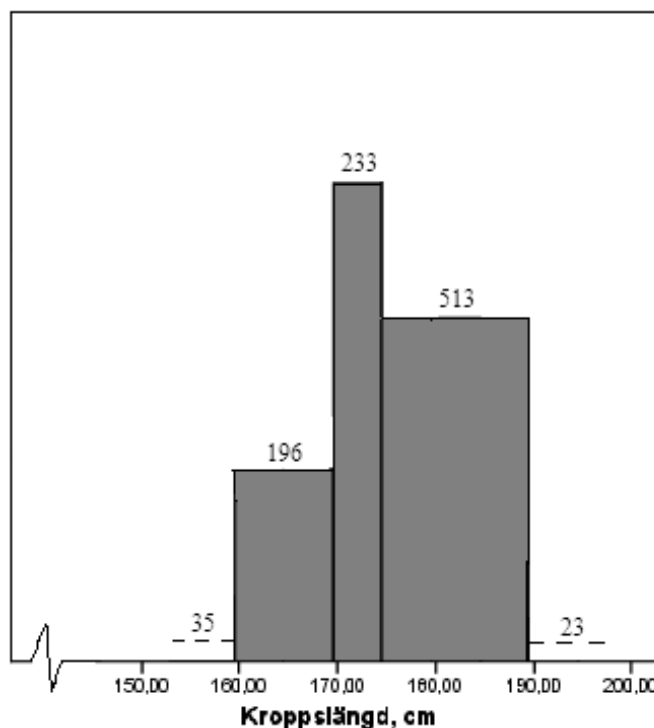


Figur 11.3 Stolpdiagram över matematikbetyget hos 40 slumpvist valda studenter enligt tabell 11.3.



Figur 11.4 Histogram över fördelningen av kroppslängd hos 1000 värnpliktiga enligt tabell 11.4.

När materialet är indelat i klasser med olika bredder, blir det mer komplicerat att konstruera ett histogram. För att efterlikna sannolikhetsfördelningens täthetsfunktion konstrueras histogrammets staplar så att arean för en klass är proportionell mot frekvensen i klassen och stapelns bas är proportionell mot klassbredden. Om en klass är bredare, måste stapelhöjden därför minskas. Oavsett vilken klass en observation hamnar i, skall den representeras av en viss area i histogrammet. Den area som representerar en observation skall dessutom vara lika stor för alla observationer.



Figur 11.5 Histogram över fördelningen av kroppslängd hos 1000 värnpliktiga enligt tabell 11.5.

I figur 11.5 visas ett exempel på ett histogram med olika klassbredder. Histogrammet är baserat på kroppslängdsmaterialet med klassindelningen i tabell 11.5. Eftersom arean av en stapel skall vara proportionell mot frekvensen

och arean är lika med stapelns höjd multiplicerad med stapelns bredd, följer det att stapelns höjd skall vara proportionell mot frekvensen dividerad med stapelns bredd. Vidare vet vi att stapelns bredd skall vara proportionell mot klassbredden. För t ex klassen 160-169 cm är frekvensen 196 och klassbredden är 10. Höjden på motsvarande stapel skall således vara proportionell mot $196/10=19,6$.

I ett histogram med olika klassbredder kan frekvenserna inte avläsas på den lodräta axeln. I stället kan man ange klassfrekvenserna inom eller över respektive stapel. Även om klassfrekvenserna skrivs ut i diagrammet skall man hålla i minnet att ett histogram med olika klassbredder är svårtolkat och bör därför användas mycket sparsamt.

Stam och lövdiagram är ett alternativ till histogram. Ett stam och löv diagram kan liknas med ett histogram där även observationernas värden redovisas. På så vis innehåller ett stam och lövdiagram mer information om materialet än ett histogram. När ett stam och löv diagram konstrueras delas först varje observation i två delar, stammen och lövet. T ex kan man dela varje observation vid decimalkommat, vid tiotalen etc. Stammen är den del som finns till vänster om delningspunkten och lövet är den del som finns till höger om delningspunkten. Observationer på kroppslängd kan vara lämpliga att dela vid tiotalen. Observationen 168 cm delas då så att 16 är stammen och 8 är lövet.

I stam och löv diagrammet representeras varje förekommande värde på stammen av en rad. På respektive rad skrivs sedan värdena på motsvarande löv. På så vis bildar stammarna en klassindelning och antalet löv bildar frekvenser. Genom att lövens värde skrivs ut ser man dessutom de olika observationernas värden. Ett stam och löv diagram presenterar därför materialet delvis i tabellform och delvis i diagramform.

I figur 11.6 redovisas ett stam och löv diagram för följande delmängd av kroppslängdsmaterialet:

172 178 168 185 178 192 176 178 190 171 182 188

För att underlätta avläsningen av storleken på observationerna bör man till diagrammet lägga till en uppgift om lövstorleken och hur stam och löv skall

kombineras för att återfå observationsvärdena.

Frekvens	Stam	Löv
1	16	8
6	17	126888
3	18	258
2	19	02

Figur 11.6 Stam och lövdiagram över kroppslängd hos 12 värnpliktiga.

Syftet med att grafiskt presentera ett datamaterial är att försöka bidra till förståelsen och tolkningen av vad som har observerats. Det är därför viktigt att alltid fråga sig vad man ser i en figur, vare sig man har gjort den själv eller om någon annan har gjort den. Först bör man se om det finns något generellt mönster i figuren, att försöka beskriva med ord figurens centrum, utbredning och form. Man bör även se efter om det finns några avvikelser från det generella mönstret. Man försöker då att ta reda på om det finns observationer som beter sig annorlunda än huvuddelen av materialet. Finns det sådana observationer är det viktigt att identifiera dem. Avvikande observationer kan vara felregistreringar eller komma från fel population, men de kan också innehålla ny och överraskande information om det problem vi studerar.

Övning 11.5

Rita lämpliga figurer för att presentera datamaterialet i övning 11.1

Övning 11.6

Rita lämpliga figurer för att presentera datamaterialen i övning 11.2

Övning 11.7

Rita lämpliga figurer för att presentera datamaterialen i övning 11.3

Övning 11.8

Rita lämpliga figurer för att presentera datamaterialen i övning 11.4

Övning 11.9

Vid en studie av väntetider vid medicinska undersökningar gjordes bl a ett antal mätningar av hur lång tid det tar att få ett prov analyserat i lab. Följande observationer i timmar erhöles:

6,1 2,1 2,1 3,5 1,3 5,7 1,3 15,7 3,9 8,4 24,0 1,7 17,8 5,3 17,5 1,2 5,0 4,7 3,4
4,8 9,1 8,7 3,9 7,1 1,2 9,3 6,5 1,3 4,9 13,9 5,2 24,5 4,4 2,9 8,3 1,1 1,1 2,6 5,1
8,1 1,7 5,6 1,1 2,2 4,3 3,2 4,2 4,4 3,0 2,0 1,8 11,9 24,8 2,5 4,0 2,8 3,0 4,5 12,7
2,6 2,4 1,9 13,0 4,0 5,0 8,8 1,3 7,3 16,2 2,7 5,2 2,2 3,0 24,0 16,2 6,7 5,2 8,3
4,4 5,7 1,6 16,7 12,1 6,4

- Konstruera ett histogram över väntetiderna.
- Gör en bedömning genom att titta på histogrammet om väntetiden vid de flesta analyserna är kortare än några timmar.
- Har några analyser mycket lång väntetid? Vilka analyser bör studeras speciellt noga för att ta reda på anledningen till långa väntetider?

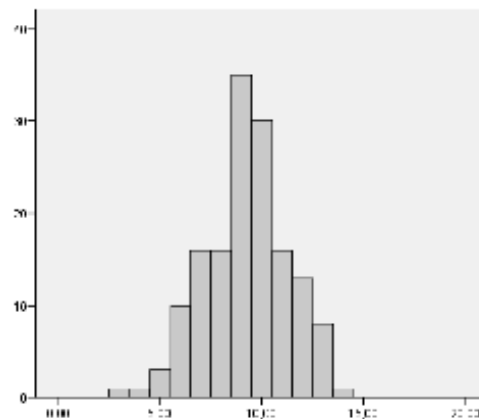
Övning 11.10

Leta rätt på ett cirkeldiagram i en tidning eller en tidskrift. Skriv ner en analys av den information som presenteras i figuren.

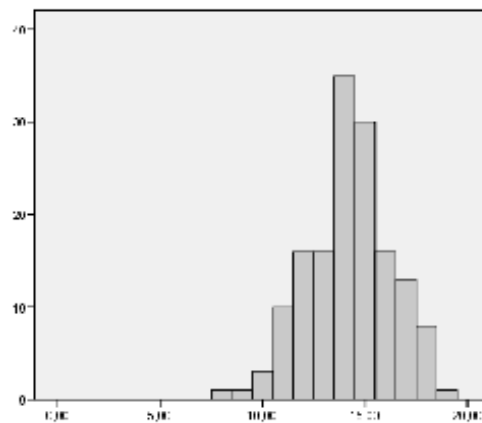
11.4 Lägesmått och kvartiler

På samma sätt som sannolikhetsfördelningar kan karaktäriseras och beskrivas av olika mått kan vi karaktärisera och beskriva empiriska fördelningar med olika mått. I det här avsnittet presenteras några lägesmått och kvartiler för empiriska fördelningar medan nästa avsnitt ägnas åt några spridningsmått för empiriska fördelningar. Syftet med ett lägesmått är att ange storleksordningen eller något slags "genomsnittsmått" på observationerna. I figur 11.7 visas histogram över två datamaterial. Det som skiljer materialen åt är att alla observationerna i det ena materialet har mätvärden som är 500 enheter större än i det andra materialet. Histogrammen har lika breda och lika många klasser samt lika stora frekvenser i respektive klass. Skillnaderna i mätvärden tar sig i uttryck att histogrammet över materialet med de större mätvärdena ligger längre till höger på talaxeln. Ett lägesmått har

till uppgift att beskriva var på talaxeln ett histogram befinner sig eller har sin tyngdpunkt. För materialet med de mindre mätvärdena har vi sålunda ett mindre värde på lägesmåttet och vice versa. De lägesmått vi beskriver här är medelvärdet och medianen.



a) Mindre mätvärden; mindre lägesmått



b) Större mätvärden; större lägesmått

Figur 11.7 Histogram över två fördelningar med olika lägesmått

*Medelvärde*t är det vanligaste lägesmåttet. För ett material bestående av n stycken mätvärden $x_1, x_2, \dots, x_n$, beräknas medelvärdet $\bar{x}$ genom

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

I exemplet med matematikbetygen i avsnitt 11.1 är medelbetyget för de 40 studenterna

$$\bar{x} = \frac{4 + 3 + \dots + 3}{40} = \frac{128}{40} = 3.2$$

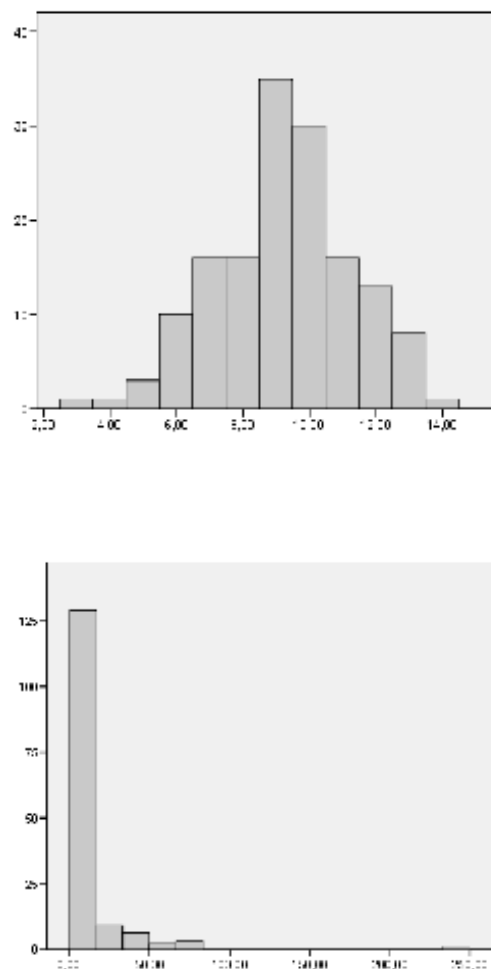
Medianen delar in ett observationsmaterial i två lika stora delar så att hälften av mätvärdena är mindre än eller lika med medianen och hälften av mätvärdena är större än medianen.

Medianen lämpar sig väl som ett lägesmått. Om antalet observationer är udda är medianen det mittersta mätvärdet av de rangordnade mätvärdena. Om antalet är jämnt brukar man låta medianen vara medelvärdet av de två mittersta mätvärdena. I exemplet med matematikbetygen för de 40 studenterna är medianbetyget 3.

Om materialet är symmetriskt kring medelvärdet sammanfaller medelvärdet och medianen. Om däremot materialet är skevt påverkas medelvärdet mer av extremvärdena i svansen av fördelningen varför medelvärde och median blir olika stora. Se figur 11.8 för en illustration.

De tre kvartilerna delar in ett observationsmaterial i fyra (ungefär) lika stora delar. 25% av mätvärdena skall vara mindre än första (nedre) kvartilen. I betygsmaterialet är antalet observationer 40 så att $0.25 \cdot 40 = 10$ mätvärden skall vara mindre än eller lika med den första kvartilen. Samtidigt skall $0.75 \cdot 40 = 30$ mätvärden vara större än den första kvartilen. I betygsexemplet är det 10 observationer som är mindre än eller lika med mätvärdet 2 och det är 30 mätvärden som är 3 eller större. Den första kvartilen är alltså inte entydigt bestämd. I sådana fall brukar man ta det värde som ligger mitt emellan de två mätvärden som "stänger in" kvarilen. I betygsexemplet blir då den första kvartilen 2,5.

Den andra kvartilen delar in observationsmaterialet i två lika stora delar och sammanfaller därför med medianen. 75% av mätvärdena skall vara mindre än eller lika med den tredje (övre) kvartilen. För att bestämma tredje kvartilen i betygsmaterialet ser vi först att $0.75 \cdot 40 = 30$, dvs tredje kvartilen är det 30:e mätvärdet i det rangordnade materialet. I vårt exempel med betyg är det



Figur 11.8 Två histogram som illustrerar effekten av sneda fördelningar på medelvärde och median. I a) är medelvärdet 9,25 och medianen 9,00, medan i b) är medelvärdet 8,79 och medianen 1,00

30:e rangordnade mätvärdet lika med 4. Det skall även vara $0.25 \cdot 40 = 10$ mätvärden som är större än den tredje kvartilen. I betygsmaterialet är det 14 observationer som har mätvärdet 4 eller högre, så det blir ingen problem med entydigheten i detta fall, den tredje kvartilen är 4.

På motsvarande sätt kan man beräkna deciler och percentiler som delar ett material i 10 respektive 100 (ungefär) lika stora delar.

Avslutningsvis poängterar vi att olika lägesmått ger olika karaktäriseringar av en fördelning. Detta är åter ett uttryck för att olika sätt att bearbeta ett material kan ge olika bild av materialet. På samma sätt som att en bearbetningsmetod inte kan sägas vara riktig eller felaktig, men däremot mer eller mindre lämplig med tanke på undersökningsproblemet, kan man inte säga att ett visst lägesmått är riktigt eller felaktigt, men väl mer eller mindre lämpligt för ett visst syfte.

11.5 Spridningsmått

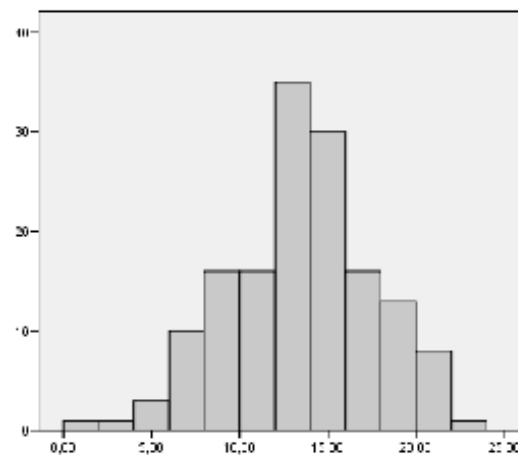
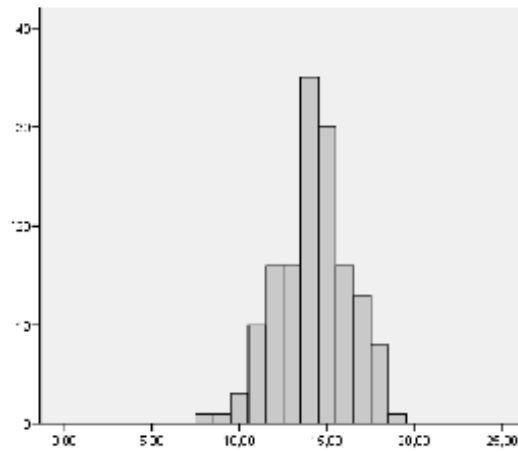
Syftet med ett spridningsmått är att ange hur mycket mätvärdena avviker från varandra eller från ett lägesmått. Ett spridningsmått ger således information om hur stort område mätvärdena är spridda över eller, alternativt, över hur stort område den empiriska fördelningen breder ut sig. I figur 11.9 visas två histogram för empiriska fördelningar med samma lägesmått men med olika spridningsmått.

De spridningsmått som behandlas här är variationsvidden, kvartilavvikelsen och standardavvikelsen.

Variationsvidden definieras, såsom namnet anger, som skillnaden mellan det största och det minsta mätvärdet. I exemplet med matamatikbetygen är det största mätvärdet 5 och det minsta 1, så att variationsvidden är $5 - 1 = 4$.

Kvartilavvikelsen beskriver hur mycket kvartilerna i genomsnitt avviker från medianen bortsett från tecken.

$$\text{Kvartilavvikelsen} = \frac{\text{tredje kvartilen} - \text{första kvartilen}}{2}$$



Figur 11.9 Histogram för två fördelningar, en med liten och en med stor spridning

Denna definition kan motiveras på följande sätt. Vi kan använda första kvartilen som ett lägesmått för den hälft av mätvärden som är mindre än medianen. Följaktligen representerar skillnaden mellan första kvartilen och medianen dessa mätvärdens avvikelse från medianen. För den hälft av mätvärden som är större än medianen kan vi på samma sätt använda skillnaden mellan

medianen och tredje kvartilen som ett mått på dessa observationers avvikelse från medianen. Bildar vi nu medelvärde av dessa två avvikelser får vi

$$\frac{(\text{medianen} - \text{första kvartilen}) + (\text{tredje kvartilen} - \text{medianen})}{2}$$

$$= \frac{\text{tredje kvartilen} - \text{första kvartilen}}{2} = \text{kvartilavvikelsen}$$

För de 40 matematikbetygen är första kvartilen = 2,5 och tredje kvartilen = 4. Kvartilavvikelsen är därför $(4 - 2,5)/2 = 0,75$.

Det vanligaste spridningsmåttet är *standardavvikelsen*. Den anger hur mycket mätvärdena i genomsnitt avviker från medelvärdet. De avvikelser vi bildar kvadreras så att avvikelsernas tecken inte spelar någon roll. Ett mätvärde som är ett visst antal enheter större än medelvärdet värderas lika som ett mätvärde som är lika många enheter mindre än medelvärdet. Ofta betecknas standardavvikelsen med s .

Antag att vi har n stycken mätvärden $x_1, x_2, \dots, x_n$ med medelvärdet $\bar{x}$. Då definieras standardavvikelsen av

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

För de 40 observationerna på matematikbetyg får vi att

$$s = \sqrt{\frac{1}{40-1} ((4-3.2)^2 + (3-3.2)^2 + \dots + (3-3.2)^2)}$$

$$= \sqrt{\frac{46.4}{39}} \approx \sqrt{1.1897} \approx 1.09.$$

Kvadraten på standardavvikelsen, s^2 , kallas (urvals-) variansen.

Ibland förekommer formler för s och s^2 där man dividerar med n i stället för $n - 1$. Här använder vi huvudsakligen den definition där man dividerar med $n - 1$.

Vi introducerar nu en användbar olikhet kallad Tchebysheff's olikhet. Även om olikheten är enkel att bevisa utesluter vi beviset här.

Tchebysheff's olikhet: Låt k vara ett tal större än eller lika med ett. Då är proportionen observationer i intervallet från $\bar{x} - ks$ till $\bar{x} + ks$ minst $1 - 1/k^2$.

Som en illustration av tillämpningen av Tchebysheff's olikhet kan vi ta $k = 1.5$. Vi får då att minst $1 - 1/1.5^2 \approx 0.56$ % av observationerna i ett datamaterial ligger i intervallet från $\bar{x} - 1.5s$ till $\bar{x} + 1.5s$. I exemplet med matematikbetyg innebär det att minst 56% av observationerna ligger mellan $3.2 - 1.5 \cdot 1.09 = 1.56$ och $3.12 + 1.5 \cdot 1.09 = 4.84$. En inspektion av materialet ger vid handen att 32 observationer, dvs $32/40=80\%$ ligger i det givna intervallet.

Låter vi $k = 2$ får vi på samma sätt att minst $1 - 1/2^2 = 3/4 = 75\%$ av observationerna ligger mellan $\bar{x} - 2s$ och $\bar{x} + 2s$. Vidare får vi att minst $1 - 1/3^2 = 8/9 \approx 88.9\%$ av observationerna ligger mellan $\bar{x} - 3s$ och $\bar{x} + 3s$. För materialet med matematikbetyg ligger samtliga observationer i båda dessa intervall. Gränserna i Tchebysheff's olikhet understiger således de faktiska proportionerna med bred marginal i detta exempel. Så är fallet också i allmänhet. Man säger att gränserna i Tchebysheff's olikhet är konservativa. En anledning till att Tchebysheff's olikhet är "försiktig" i sina uttalanden är att olikheten gäller för alla empiriska fördelningar, hur de än ser ut.

Om datamaterialets fördelning liknar normalfördelningen brukar man ibland använda normalfördelningens egenskaper för att approximera proportionen observationer i ett intervall. Detta leder till den s k Empiriska regeln:

Empiriska regeln: Om datamaterialets fördelning liknar normalfördelningen så innehåller intervallet

från $\bar{x} - s$ till $\bar{x} + s$ ungefär 68% av observationerna

från $\bar{x} - 2s$ till $\bar{x} + 2s$ ungefär 95% av observationerna

från $\bar{x} - 3s$ till $\bar{x} + 3s$ nästan alla observationer.

Observera att Empiriska regeln endast är en approximation och att approximationen kan förväntas vara god endast under förutsättning att den empiriska fördelningen verkligen liknar normalfördelningen.

Övning 11.11

Förklara hur medelvärdet och medianen förhåller sig till varandra då fördelningen är

- a) symmetrisk
- b) skev med en svans åt vänster
- c) skev med en svans åt höger
- d) ungefär normalfördelad

Övning 11.12

Förklara med egna ord begreppet variationsvidd.

Övning 11.13

Diskutera hur variansen och standardavvikelsen mäter variation.

Övning 11.14

Empiriska regeln och Tchebysheff's olikhet gör liknande uttalanden om ett material. Beskriv med egna ord vilken typ av uttalanden de gör och vad som skiljer dem.

Övning 11.15

Bestäm medelvärde, median, kvartiler, variationsvidd, kvartilavvikelse och standardavvikelse för materialet i övning 11.1. Kontrollera Tchebysheff's olikhet och empiriska regeln för några värden på k .

Övning 11.16

Bestäm medelvärde, median, kvartiler, variationsvidd, kvartilavvikelse och standardavvikelse för materialen i övning 11.2. Kontrollera Tchebysheff's olikhet och empiriska regeln för några värden på k .

Övning 11.17

Medelvärde och standardavvikelse för hållfastheten hos de 40 sopsäckarna i övning 11.4 är $\bar{x} = 50,575$ respektive $s = 1,6438$.

- Vilken slutsats kan man dra från histogrammet i övning 11.8 om empiriska regeln kan användas för att beskriva hållfasthetsmaterialet?
- Använd empiriska regeln för att bestämma intervall som innehåller c:a 68%, 95% och 99% av hållfasthetsmaterialet.
- Ger intervallet som innehåller 99,73% av alla mätvärden något stöd för påståendet att nästan alla sopsäckar har en hållfasthet som överstiger 45 kg? Motivera ditt svar.
- Hur stora andelar av observationerna på sopsäckarnas hållfasthet hamnar faktiskt i intervallen $[\bar{x} \pm s]$, $[\bar{x} \pm 2s]$ och $[\bar{x} \pm 3s]$? Jämför dessa andelar med de som anges av empiriska regeln. Ger dessa jämförelser stöd för de påståenden du gjorde i uppgift b och c?

11.6 Box plotdiagrammet

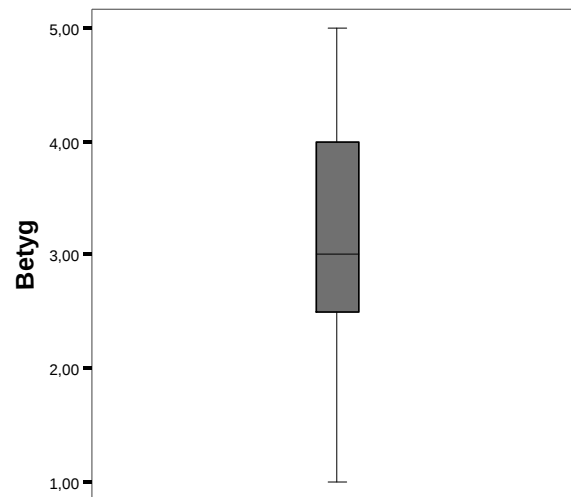
Box plotdiagrammet är en diagramtyp som syftar till att illustrera läges och spridningsmått för en empirisk fördelning samt i detalj beskriva fördelningens "svansar", d v s hur de minsta och de största observationsvärdena förhåller sig till de övriga observationsvärdena. Figur 11.9 visar en box plot över matematikbetygen för de 40 slumpvis utvalda studenterna som redovsas i tabell 11.3. I centrum av figuren finns en "låda" (box) som begränsas av första och tredje kvartilerna. Lådans längd blir därmed lika med kvartilavståndet, d v s två gånger kvartilavvikelsen. Därigenom representerar lådans längd ett spridningsmått. I lådan markeras även medianen som ett lägesmått.

Lådan i en box plot markerar således värdena hos 50% av observationerna. Vidare vet vi att 25% av observationerna har värden som ligger över lådan och

25% av observationerna har värden som ligger under lådan. Dessa observationers värden markeras särskilt i box plotdiagrammet. Vi identifierar först observationer som avviker mer än 1.5 gånger kvartilavvikelsen från någon av kvartilerna. Det största respektive det minsta observerade värdet som avviker mindre än 1.5 gånger kvartilavståndet från undre respektive övre kvartilen markeras med ett streck och en linje in till lådan. Observationer som avviker mellan 1.5 och 3 gånger kvartilavvikelsen från undre respektive övre kvartilen markeras med "o" och observationer som avviker mer än 3 gånger kvartilavståndet från övre respektive undre kvartilen markeras med "*". Dessa typer av observationer kallas ibland något oegentligt för outliers respektive extremvärden. Dessa observationer kan ha ett stort inflytande på urvalsmedelvärdet och urvalsvariansen samt på resultaten från mer sofistikerade analyser av datamaterialet. Det är därför viktigt att identifiera vilka observationer som hamnar i dessa grupper samt att försöka förklara varför de avviker så kraftigt från medianen. Ofta är det helt naturligt att en del observationer avviker så pass kraftigt som t ex tre gånger kvartilavståndet från en kvartil, eller mer. Många gånger är det dock ett mätfel eller ett skrivfel som ligger bakom ett sådant mätvärde. Om så är fallet måste det felaktiga värdet korrigeras eller, om det inte är möjligt, i värsta fall uteslutas. En annan förekommande anledning till avvikande mätvärden är att observationerna har gjorts under speciella betingelser. Antingen tillhör då individerna som observationerna är gjorda på inte den studerade populationen eller så tillhör de en speciell delmängd av populationen. Det senare fallet är ofta mycket intressant eftersom det antyder ett samband mellan mätvärdenas storlek och de speciella betingelser som definierar den delmängd av populationen som individen tillhör.

Övning 11.18

Gör ett box plotdiagram över observationerna på hållfasthet av sopsäckar i övning 11.4. Jämför diagrammet med histogrammet i övning 11.4 och de slutsatser du drog i övning 11.17.



Figur 11.9 Box plot över matematikbetyget för 40 slumpvis utvalda studenter