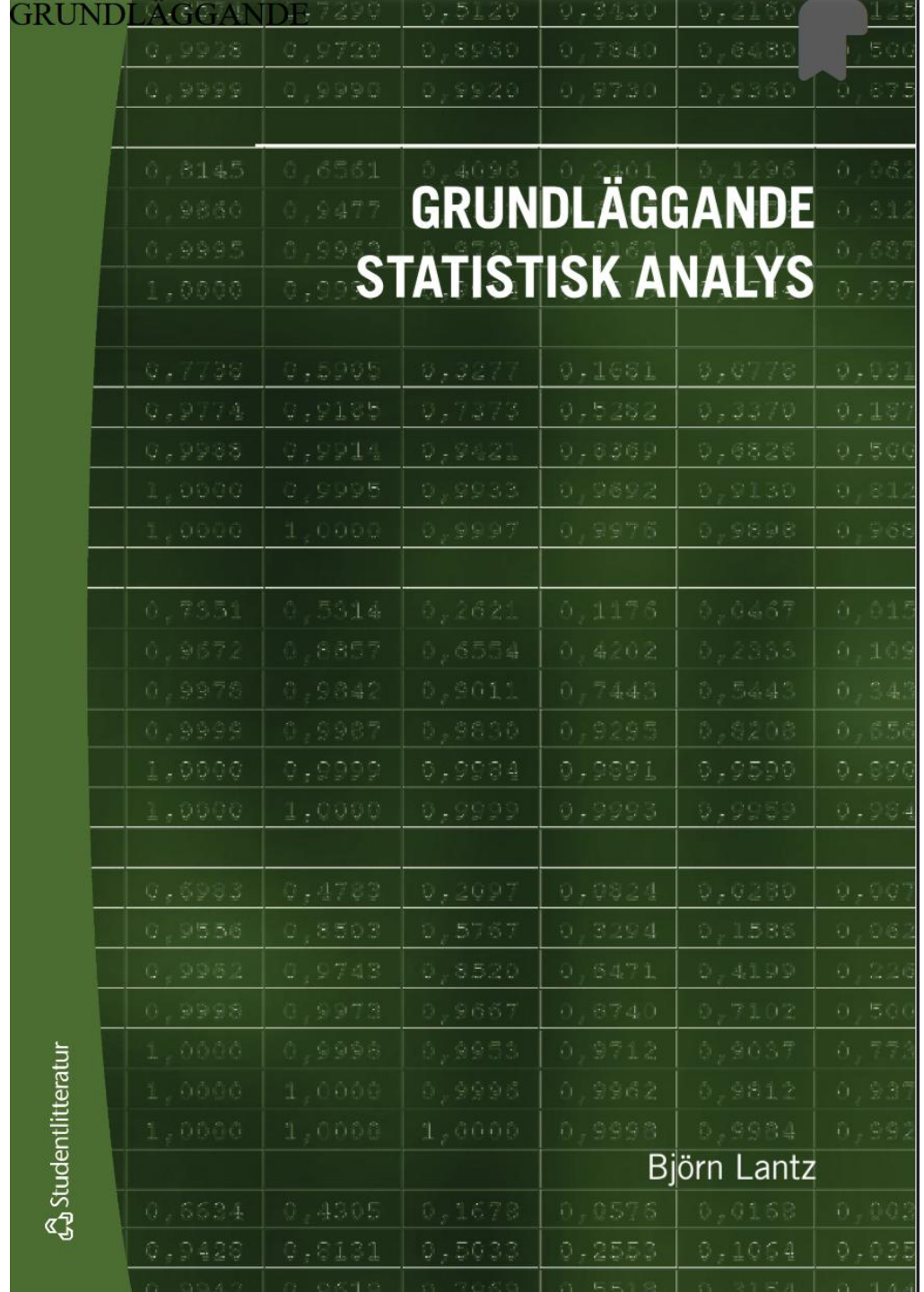
GRUNDLÄGGANDE



GRUNDLÄGGANDE STATISTISK ANALYS

Björn Lantz

Studentlitteratur

1 Introduktion till statistisk analys

1.1 Inledning

Statistik handlar om att samla in, sammanställa, analysera och tolka data. Ekonomer och ingenjörer (liksom många andra yrkesgrupper) har större nytta av att behärska grundläggande statistiska metoder och modeller än vad man kanske tror. Det finns nämligen många olika användningsområden för statistik, och det är ofta så att man kan skapa en grund för bättre beslut om man använder statistiska metoder för att analysera olika situationer. I den här boken kommer många exempel på sådana situationer att ges, t.ex.:

- när man ska göra trovärdiga prognoser om efterfrågan,
- när man ska kartlägga olika produktionsprocessers kvalitet,
- när man ska undersöka om en viss typ av reklam ger bättre utfall än en annan,
- när man ska analysera kostnadsstrukturen för en viss produkt, eller
- när man ska fastställa vilka faktorer som är viktigast för kunders beslut att köpa en viss produkt.

Statistikämnet kan (grovt) delas in i *beskrivande* statistik och *analytisk* statistik. Den beskrivande statistiken syftar till att på ett överskådligt och strukturerat sätt skildra de väsentliga dragen i en mängd data som samlats in. Ofta utnyttjas olika typer av grafiska figurer för att ge läsaren en så klargörande beskrivning som möjligt över materialet. Den analytiska statistiken handlar främst om att dra slutsatser om hur verkligheten ser ut i olika avseenden med hjälp av stickprov (*statistisk inferens*), att dra slutsatser om sambandet mellan olika företeelser ser ut (*sambandsanalys*) och att göra fördjupade analyser av hur säkra dessa slutsatser är.

Det är lätt att göra fel med hjälp av statistik om man inte förstår anledningen till varför man gör som man gör i olika situationer. Anta t.ex. att vi vill göra en kundundersökning i en affär. Vi skickar därför en person att intervjua de första fem kunderna som kommer in genom dörren om deras syn på affären. Intervjuarens uppfattning är att tre av de fem kunderna i stora drag tyckte att affären var bra. Ska vi nu dra slutsatsen att 60 % av affärens kunder tycker att affären är bra?

Nej, det ska vi inte göra! Om vi hade dragit denna slutsats så hade vi nämligen gjort ett antal grova fel. För det första frågade vi väldigt få kunder. Det kan mycket väl vara så att de kunder vi råkade träffa på av en ren slump råkade vara ovanligt positiva – eller ovanligt negativa – jämfört med affärens totala kundkrets. Ju färre kunder vi bestämmer oss för att fråga, desto större är risken att slumpen råkar spela oss ett sådant spratt. Detta visar på en fundamental princip i den statistiska vetenskapen: ju större stickprov man tar, desto säkrare kan man räkna med att resultatet blir. Att eventuella slumpfel kan förväntas ”bli mer utspädda” ju större stickprov man använder sig av är en intuitivt tilltalande logik (som dessutom stämmer).

För det andra kan det vara så att den tidpunkt vi valde för våra intervjuer leder till att bara en viss typ av kunder blir intervjuade. Om affären t.ex. öppnar kl. 07.00 och alla fem intervjuerna äger rum under första halvtimmen av öppethållandet så kan det innebära att kunderna är väldigt stressade för att de har bråttom till jobbet, att kunderna är positiva bara för att de ska köpa nybakt bröd till frukost, eller att morgontrötta kunder inte får möjlighet att lämna sin åsikt. Alla affärens kunder har inte haft samma chans att bli utvalda för en intervju, och risken med det är att resultaten kan snedvridas. Systematiken i urvalet kan ha varit ”för grov”.

För det tredje var det intervjuarens personliga uppfattning att tre av de fem kunderna tyckte affären var bra. I verkligheten kanske två av kunderna uppehöll sig vid saker som de tyckte kunde förbättras, även om de i stora drag tyckte att affären var bra. Eller tvärtom. Vi kan i alla fall inte vara säkra på att intervjuaren faktiskt inte har missförstått vad kunderna egentligen tyckte, eftersom detta omdöme tydligen var en subjektiv helhetsbedömning.

För att undvika den här typen av enkla fällor är det viktigt att inte bara veta *hur* man gör när man arbetar med statistik, utan också *varför*. Syftet med denna bok är att inte bara visa när och hur man använder grundläggande statistiska metoder och modeller, utan också att ge läsaren en bra bild av den teoretiska bakgrunden till varför man gör på det sätt man gör. Det är författarens förhoppning att läsaren därmed ska kunna få en bättre förståelse för statistikämnet.

1.2 Mått på centraltendens

Med statistiska analysmetoder kan man ofta med viss säkerhet förklara *hur* något förhåller sig, men man kan inte alltid (långt därifrån) förklara *varför*. Statistiken använder numerisk information (d.v.s. data) av olika slag, och fenomen som inte går att uttrycka numeriskt kan vi därför inte heller analysera med hjälp av statistik. Några exempel på data:

- 92 kunder var kunder i en viss affär på en viss dag.
- Tidsåtgången för att producera en viss vara är 18 minuter.
- Andelen högutbildade på en viss arbetsplats är 72 %.
- Antalet försök som krävs för att klara en viss uppgift är 3.
- Diametern på ett visst rör är 56 mm.

Data härrör från någon typ av observation eller mätning av ett *element*. Med element avses ett objekt, fenomen eller en person för vilken någon *egenskap* kan observeras eller mätas. Mätningen eller observationen av egenskapen resulterar i ett värde på en *variabel*.

För att beskriva en datamängd bestående av ett antal uppmätta eller observerade variabelvärden används ofta något mått på datamängdens *centraltendens*. De viktigaste måtten på centraltendens är *typvärde*, medelvärde och *median*.

Typvärdet för en datamängd är det värde i datamängden som är vanligast förekommande.

Medelvärdet för en datamängd är dess genomsnitt, d.v.s. summan av alla datamängdens variabelvärden delat med antalet variabelvärden.

Medianen för en datamängd är det värde som är det mittersta om datamängden rangordnas. Observera att om datamängden består av ett jämnt antal värden finns två värden i mitten, och då beräknas medianen som medelvärdet av dessa två värden.

Om vi har n observationer, betecknade med x_1, x_2, …, x_n, så beräknas medelvärdet av dessa observationer, betecknat med $\bar{x}$, som

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}.$$

Exempel 1-1

Antalet sålda bilar per dag på Stures Bil AB de senaste 7 dagarna är 3, 5, 4, 3, 3, 6, 11 stycken. Vad är typvärdet, medianen och medelvärdet för antalet sålda bilar under perioden?

Lösning

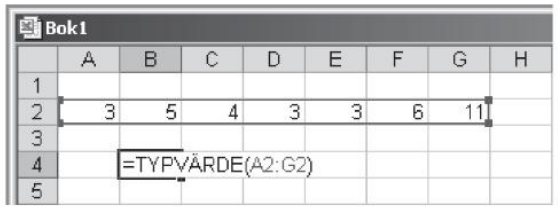
Typvärdet för en datamängd är det värde som är vanligast förekommande, och det enda värde som förekommer mer än en gång är 3. Typvärdet för antalet sålda bilar är alltså 3.

Medianvärdet får vi enklast genom att rangordna datamängden: 3, 3, 3, 4, 5, 6, 11. Vi ser då att värdet 4 hamnar i mitten. Medianen för antalet sålda bilar är således 4.

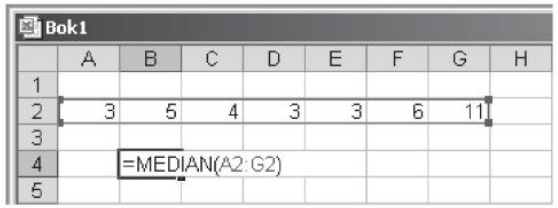
Medelvärdet för antalet sålda bilar blir

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{3+5+4+3+3+6+11}{7} = \frac{35}{7} = 5.$$

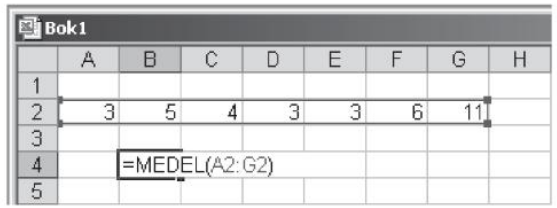
I Excel kan man utnyttja funktionerna TYPVÄRDE, MEDIAN och MEDELVÄRDE för att beräkna de aktuella måtten på centraltendens för en datamängd. I samtliga fall anges som argument det område i kalkylarket där datamängden finns. Hur det kan se ut i fallet i exempel 1-1 innan man tryckt på ”enter” framgår av figurerna 1.1, 1.2 och 1.3 nedan.



Figur 1.1: Typvärde i Excel – inmatning



Figur 1.2: Median i Excel – inmatning



Figur 1.3: Medelvärde i Excel – inmatning

1.3 Variabeltyper

Vilket av de tre grundläggande måtten ska man då använda sig av om man vill beskriva centraltendensen för en datamängd? Svaret är att man gärna får använda alla tre – om man kan! Beroende på vilken slags data man arbetar med är det inte säkert att det går att använda alla tre måtten. Den variabel som observationerna eller mätningarna har gett värden på kan nämligen vara av en typ som inte tillåter att man till exempel beräknar ett medelvärde.

En variabel – oavsett typ – får alltså ett värde till följd av en mätning eller en observation. Om ett variabelvärde exempelvis representerar en *kvantitet* (d.v.s. variabelvärdet är svaret på ”hur många”) så är variabeln *kvantitativ*. Motsatsen är om ett variabelvärde representerar en kategori, t.ex. en persons kön, färgen på en bil eller rangordningen (”placeringen”) för en viss tävlingsdeltagare. I detta fall sägs variabeln vara *kvalitativ*.

Kvantitativa variabeldata sägs ofta vara ”starkare” än kvalitativ, vilket syftar på att fler slags operationer är tillåtna med hjälp av datan. Den allra svagaste formen av data är kvalitativa data som endast kategoriserar datan utan att ge de olika kategorierna någon inbördes ordning, t.ex. civilstånd, färg eller postnummer. Sådan data kallas *nominaldata*. För att beskriva centraltendensen för en datamängd bestående av nominaldata kan endast typvärde användas. Eftersom man inte kan rangordna datan kan median inte användas, och eftersom variabelvärdena inte motsvarar kvantiteter kan medelvärde inte beräknas.

Om kvalitativa data beskrivs av kategorier mellan vilka det finns en inbördes (rang-) ordning så kallas den ordinaldata. Några exempel på ordinaldata är utbildningsnivå, betyg och preferensordning. Just möjligheten att rangordna gör att det finns operationer som kan utföras på ordinaldata som inte är möjliga på nominaldata. Den uppenbara skillnaden jämfört med nominaldata är att centraltendensen för en datamängd förutom typvärdet även kan beskrivas med dess medianvärde. Medelvärde kan dock inte beräknas eftersom mätpunkterna är kategorier som saknar matematisk betydelse.[1]

[1] Observera att ett betyg eller en attityd på t.ex. skalan 1–5 alltså *inte* är kvantitativa data. Även om kategorierna namnges med hjälp av siffror så representerar dessa siffror inte kvantiteter. Med andra ord: de svarar inte på frågan

Kvantitativa data kan däremot beskrivas med såväl medelvärde och median som typvärde, vilket framgick av exempel 1-1 ovan. När variabelvärdena motsvarar kvantiteter så vet vi att det är samma differens mellan 7 och 5 som mellan 4 och 2. Eftersom variabelvärdena representerar kvantiteter kan de alltså alltid adderas (och subtraheras), vilket är vad som krävs för att man ska kunna beräkna deras medelvärde.

Ett kvantitativt variabelvärde kan däremot inte alltid relateras inbördes till ett annat kvantitativt variabelvärde. Om utomhustemperaturen var 2 grader Celsius i går och 4 i förrgår så är det inte rimligt att säga att det var hälften som varmt i går som i förrgår med motiveringen att 2 / 4 = 0,5. För hur skulle man i så fall uttrycka temperaturrelationen mellan två dagar med exempelvis –1 respektive 3 grader Celsius?

En familj med 2 barn har däremot hälften så många barn som en familj med 4 barn, just därför att 2 / 4 = 0,5. Skillnaden mellan dessa båda fall är att vi i det senare fallet har en absolut nollpunkt på mätskalan. Man kan inte ha färre än 0 barn. Temperaturskalan för grader Celsius har däremot ingen absolut nollpunkt. För att vi ska kunna dividera två variabelvärden med varandra måste de alltså komma från en kvantitativ skala med en absolut nollpunkt. Data av denna typ kallas därför för *kvotdata*. Kvantitativa data som saknar absolut nollpunkt kallas för *intervalldata.*[2]

”hur många”. Eftersom ett sådant variabelvärde inte har någon reell matematisk betydelse så är datan inte heller av kvantitativ karaktär. Betyget 4 är bättre än betyget 2 vilket innebär att betygen är rangordningsbara, men det är helt ologiskt att säga att betyget 4 är ”dubbelt så bra” som betyget 2, vilket innebär att det inte rör sig om kvantitativa data. Eftersom sådana mätvärden inte kan adderas kan deras medelvärde heller inte beräknas.

[2] I praktiken uppstår sällan problemet med intervalldata. Det är svårt att komma på något realistiskt exempel med intervalldata utom just temperaturskalor. Grader Celsius (eller grader Fahrenheit) kan dessutom räknas om till grader Kelvin, där det finns en absolut nollpunkt.

Nominaldata är kvalitativa data som kan kategoriseras men inte rangordnas.

Ordinaldata är kvalitativa data som både kan kategoriseras och rangordnas.

Intervalldata är kvantitativa data utan absolut nollpunkt.

Kvotdata är kvantitativa data med absolut nollpunkt.

Exempel 1-2

På Stures Bil AB har man haft problem med klimatanläggningen på en viss biltyp och man vill nu försöka kartlägga detta bättre. Man skickar därför ut en enkät med frågor om

a) Kön på den som oftast kör bilen.

b) Antal mil bilen körs per år.

c) Normalt inställd temperatur på klimatanläggningen.

d) Hur nöjd kunden var med bemötandet hos Stures Bil AB.

Vilken typ av data erhålls som svar på respektive fråga? Vilken typ av mått på centraltendens kan man använda för att beskriva den insamlade datan?

Lösning

a) Nominaldata. Svar på frågan är möjliga att kategorisera men inte möjliga att rangordna. Endast typvärde kan användas för att beskriva centraltendens.

b) Kvotdata. Exempelvis är 1 000 mil dubbelt så långt som 500 mil. Medelvärde, median och typvärde kan användas för att beskriva centraltendens.

c) Intervalldata. 20 grader är 4 grader mer än 16 grader, men inte 4 / 16 = 25 % mer. Medelvärde, median och typvärde kan användas för att beskriva centraltendens.

d) Ordinaldata. En mycket nöjd kund är mer nöjd än en ganska nöjd kund, som i sin tur är mer nöjd än en missnöjd kund. Men man kan inte säga hur stor skillnaden är. Således kan median och typvärde – men inte medelvärde – användas för att beskriva centraltendens.

1.4 Populationer och stickprov

Som vi sa tidigare så syftar den analytiska statistiken i hög grad till att dra slutsatser om hur verkligheten ser ut med hjälp av stickprov. Den verklighet vi vill observera består ofta av många element – för många för att vi ska kunna undersöka alla. Ofta kan då analys av stickprov vara en framkomlig väg.

En avgränsningsbar mängd element kallas för en *population*. Några exempel på populationer:

- Alla människor i Sverige som är över 18 år.
- Alla registrerade studenter på Göteborgs universitet under höstterminen 2004.
- Alla fakturor som mottogs på ett visst företag under ett visst år.

De ovan nämnda exemplen på populationer har det gemensamt att de är *ändliga* – det är alltså möjligt (åtminstone teoretiskt) att specificera det antal element som ingår i dessa populationer. Men en population kan även bestå av ett oändligt antal presumtiva element. När element är resultat från en process eller ett experiment som kan upprepas hur många gånger som helst, motsvarar dessa element en *oändlig* population. Några exempel:

- Bilar som produceras på ett löpande band.
- Kunder i en viss affär.
- Studenter som har haft, har eller kommer att ha denna bok som kurslitteratur.

Att undersöka alla element i en population – att göra en *totalundersökning* – låter sig sällan göras, beroende på t.ex. att populationen är oändlig, för stor eller att undersökningen i sig är destruktiv för de undersökta elementen. Statistiken erbjuder dock ofta metoder för att analy-

sera mindre stickprov från en population på ett sätt som möjliggör slutsatser med god säkerhet om t.ex. vilken centraltendens som gäller för någon viss parameter i populationen. Det är just sådana metoder som den här boken huvudsakligen handlar om.

För att informationen i ett stickprov ska kunna återspegla populationen som man tar stickprov från gäller som huvudregel att *slumpen ska styra vilka element från populationen som hamnar i stickprovet*. I det enklaste fallet låter man slumpen styra fullt ut genom att låta alla element i populationen få exakt samma chans att komma med i stickprovet. Detta kallas *obundet slumpmässigt urval* (OSU) och är den vanligaste och viktigaste principen för statistisk stickprovsanalys. I och med att inget annat än slumpen styr, kan man *i genomsnitt* förvänta sig att stickprovet kommer att vara en miniatyrkopia av populationen från vilken elementen kommer. Observera dock att detta bara gäller i genomsnitt. Just därför att slumpen styr, kommer det ibland att bli så att det blir mer eller mindre extrema element som hamnar i stickprovsurvalet. Vi måste därför analysera stickprovet statistiskt för att se hur pass trovärdiga slutsatser vi kan dra om den bakomliggande populationen med hjälp av det.

Ibland används någon form av systematik för att åstadkomma ett stickprovsurval från en population. Om man t.ex. ska göra en kundundersökning i en affär så kanske man ställer frågor till var 20:e kund som kommer in genom dörren i stället för att renodlat slumpa fram vilka kunder man ska fråga. Ett sådant *systematiskt urval* kan användas som en god approximation av ett OSU om, men bara om, det inte finns något i systematiken som riskerar att snedvrida urvalet. Det åligger den som ska göra undersökningen att noggrant överväga sådana risker. Om man frågar var 20:e kund som kommer in i en affär så är denna risk normalt mycket liten, men om man t.ex. frågar besökarna på en viss webbsida rörande deras köpvanor på internet så måste man inse att resultatet inte kommer att vara giltigt för människor i allmänhet. Risken att folk som surfar på internet tenderar att i högre grad handla på internet än människor i allmänhet måste anses vara mycket stor. I detta fall bör populationen kanske snarare definieras som folk som surfar på internet eller folk som surfar in på den aktuella sidan. När ett urval utgörs av personer som själva aktivt har valt att vara med i en undersökning föreligger dessutom en stor risk för att dessa personer i genomsnitt har

mer extrema åsikter i det ämne som undersökningen handlar om än människor i allmänhet.

Om den population man vill uttala sig om består av klart avgränsade grupper, kan det finnas anledning att hjälpa slumpen en smula. Ett typiskt exempel är om de klart avgränsade grupperna består av element som är mycket lika varandra men mycket olika de element som finns i andra grupper. Ett enkelt exempel kan vara en bransch där det finns många små och få stora företag. Om vi tar ett OSU från denna population så är risken stor att bara små företag kommer med, varvid stickprovets egenskaper inte kommer att vara en spegling av den population från vilken stickprovet kommer. I detta läge kan man dela in populationen i två underpopulationer (s.k. *strata*) och ta ett OSU från var och en av dessa populationer, för att sedan väga ihop resultatet så att det sammanlagda stickprovet kan förväntas återspegla hela branschen. Detta kallas *stratifierat urval*.

En liknande situation uppstår om de klart avgränsade gruppernas sammansättning är mycket lika varandra men kännetecknas av att det råder stor variation inom varje grupp. Detta kan ofta vara fallet när en population är stor och är spridd över ett stort geografiskt område med många tydliga regionala grupper. I detta läge kan man få större precision och/eller lägre kostnad i sin undersökning genom att först göra ett OSU av grupper, för att sedan totalundersöka eller göra ett OSU från var och en av de utvalda grupperna. Detta kallas *klusterurval*.

Ett obundet slumpmässigt urval (OSU) innebär att alla element i en population har samma sannolikhet att komma med i urvalet.

Ett systematiskt urval innebär att urvalet baseras på att elementen i populationen finns i någon slags ordning.

Ett stratifierat urval innebär att urvalet baseras på att populationen kan delas in i grupper bestående av liknande element där grupperna sinsemellan är olika.

Ett klusterurval innebär att urvalet baseras på att populationen kan delas in i grupper med stor variation inom varje grupp men med liknande sammansättning som de andra grupperna.

1.5 Mått på spridning

När man beskriver en datamängd med dess centraltendens vill man ofta även visa hur stor spridning kring centraltendensen som datamängden uppvisar. Beroende på vilken typ av data man arbetar med finns olika möjligheter att använda mått på sådan spridning. När det exempelvis rör sig om nominaldata kan man uppenbarligen inte tala om ”spridning kring centraltendens”, eftersom det inte finns någon inbördes ordning mellan kategorierna. Illustration av sättet på vilket datan fördelar sig på de olika kategorierna får då ske genom att man just visar vilket antal observationer (*frekvenser*) som finns för respektive kategori.

När man arbetar med data som kan rangordnas kan man däremot tala om ”spridning kring centraltendens”. Jämför till exempel datamängden 3, 4, 4, 5, 5 med datamängden 1, 3, 4, 5, 8. Båda datamängderna har medianen 4 men den senare av de båda datamängderna har uppenbarligen en större spridning. Ett sätt att beskriva spridningen för datan är att ange dess *variationsvidd*, vilken motsvaras av differensen mellan högsta och lägsta värde i datamängden. Den första datamängden har variationsvidden 5 – 3 = 2 medan den andra har variationsvidden 8 – 1 = 7.

När vi har rangordnat en datamängd för att ta fram dess median och dess variationsvidd kan vi beskriva den ännu tydligare. Medianen är som sagt det mittersta värdet i datamängden – det värde som delar datamängden i två lika stora delar. De tre värden som delar en datamängd i fyra lika stora delar kallas *kvartiler*. Under den första kvartilen finns 25 % av datamängden, den andra kvartilen är samma sak som medianen och under den tredje kvartilen finns 75 % av datamängden. Genom att ange kvartilerna för en datamängd så får man också en bild av spridningen i datamängden. En uppenbar nackdel med variationsvidd som spridningsmått är att ett enstaka extremt värde påverkar spridningsmåttet starkt. Genom att man i stället använder kvartilerna så elimineras detta potentiella problem.

Om en datamängd består av totalt n rangordnade värden så är första kvartilen det värde som finns på position $(n+1)/4$, och tredje kvartilen det värde som finns på position $3(n+1)/4$. (Andra kvartilen är helt enkelt medianen, som med samma logik finns på position $2(n+1)/4$.) Om inte $n+1$ är jämt delbart med 4 krävs därför interpolering mellan två värden för att beräkna första och tredje kvartilen. Anta t.ex. att en rangordnad datamängd består av de sex värdena 1, 3, 4, 6, 9 och 10. Medianen är medelvärdet av 4 och 6, d.v.s. 5. Eftersom antalet rangordnade värden är $n = 6$ blir den första kvartilen är det värde som finns på den fiktiva positionen $(6+1)/4 = 1{,}75$, d.v.s. efter exakt 75 % av vägen från värdet på position 1 till värdet på position 2. Interpolering mellan dessa båda värden ger då $0{,}75 \cdot 3 + 0{,}25 \cdot 1 = 2{,}5$. Med samma logik blir tredje kvartilen det värde som finns på den fiktiva positionen $3(6+1)/4 = 5{,}25$, d.v.s. efter exakt 25 % av vägen från värdet på position 5 till värdet på position 6. Interpolering mellan dessa båda värden ger då $0{,}25 \cdot 10 + 0{,}75 \cdot 9 = 9{,}25$.

De vanligaste, och även viktigaste, spridningsmåtten är *varians* och *standardavvikelse*. Variansen för en datamängd om N observationer som utgör en population betecknas med σ^2 och definieras som den genomsnittliga kvadrerade avvikelsen från det sanna medelvärdet i populationen, som i sin tur betecknas med μ. Variansen för en datamängd är mer informativ än enklare mått som t.ex. variationsvidden eftersom alla observationer i datamängden ingår i måttet. Varians är alltså ett mått på hur stora observationernas avvikelser från medelvärdet är i en datamängd som helhet.

Variansen för en population med N element:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

Standardavvikelsen för en datamängd som utgör en population betecknas med σ och är kvadratroten ur populationens varians.

Standardavvikelsen för en population med N element:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

Om datamängden däremot utgörs av ett *stickprov* om n observationer från en större population så dividerar man inte med n när man beräknar varians och standardavvikelse för stickprovet, vilket man spontant skulle kunna tro. Detta beror på att vi inte har tillgång till populationens sanna medelvärde μ när vi bara har tagit ett stickprov. I stället måste vi använda stickprovets medelvärde $\bar{x}$ som en *skattning* av μ när vi räknar varians och standardavvikelse för stickprovet. Risken är att stickprovets medelvärde inte överensstämmer med sanna medelvärdet, och för att justera för denna risk dividerar vi med n–1 i stället för med n. Oavsett om man överskattar eller underskattar populationens sanna medelvärde i denna process så blir effekten att täljaren i variansuttrycket blir mindre (pröva själv!). För att kompensera för detta så använder man alltså n–1, som ju är mindre än n, i nämnaren.[3] I övrigt räknar man på samma sätt som när datamängden utgörs av en population.

Variansen för ett stickprov betecknas med s^2 och standardavvikelsen för ett stickprov betecknas med s. Standardavvikelsen är som tidigare roten ur variansen.

Variansen för ett stickprov om n observationer:

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

Standardavvikelsen för ett stickprov om n observationer:

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

Det kommer att framgå längre fram i den här boken varför varians och standardavvikelse är de viktigaste spridningsmåtten för en datamängd.

[3] Ett mer formellt sätt att förklara n-1 i nämnaren på är att man förlorar en *frihetsgrad* när man skattar en av parametrarna i beräkningen – i detta fall medelvärdet – med hjälp av stickprovsdata. Begreppet frihetsgrader kommer att diskuteras mer i detalj längre fram i den här boken.

En nackdel med dem är emellertid det faktum att kvadreringen leder till att enstaka extrema avvikelser får stort genomslag i måttet. Ett sätt att undvika detta är att i stället använda den genomsnittliga absoluta avvikelsen (*Mean Absolute Deviation*, MAD) från medelvärdet som spridningsmått. Varje avvikelse kommer då att väga proportionellt sett lika tungt, medan större avvikelser proportionellt sett väger tyngre i en varians och i en standardavvikelse.

Genomsnittlig absolutavvikelse för en datamängd om n observationer:

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

Exempel 1-3

Antalet sålda bilar per dag på Stures Bil AB de senaste 10 dagarna är 5, 5, 8, 4, 5, 4, 3, 3, 6, 11 stycken. Vad är kvartilerna och variationsvidden i denna datamängd? Se datamängden som ett stickprov och beräkna därvid även den genomsnittliga absolutavvikelsen, variansen och standardavvikelsen.

Lösning

Datamängden består av 10 observationer, och rangordning ger: 3, 3, 4, 4, 5, 5, 5, 6, 8, 11.

Första kvartilen är värdet på den fiktiva positionen (10+1)/4 = 2,75, vilket motsvaras av 0,75·4+0,25·3 = 3,75. Medianen, d.v.s. andra kvartilen, är 5. Tredje kvartilen är värdet på den fiktiva positionen 3(10+1)/4 = 8,25, vilket motsvaras av 0,75·6+0,25·8 = 6,5.

För att kunna beräkna genomsnittlig absolutavvikelse, varians och standardavvikelse måste vi först beräkna medelvärdet:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{3+3+4+4+5+5+5+6+8+11}{10} = \frac{54}{10} = 5{,}4.$$

Genomsnittliga absolutavvikelsen blir

$$MAD = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} = \frac{|3-5{,}4|+|3-5{,}4|+...+|11-5{,}4|}{10} = 1{,}76\,.$$

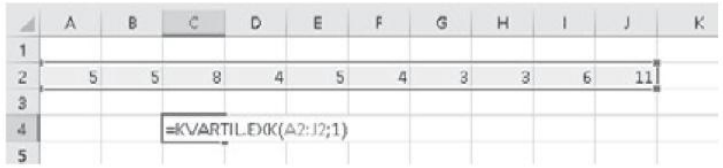
Variansen beräknas till

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{(3-5{,}4)^2+(3-5{,}4)^2+...+(11-5{,}4)^2}{10-1} = 6{,}0444$$

och standardavvikelsen

$s = \sqrt{6{,}0444} = 2{,}4585.$

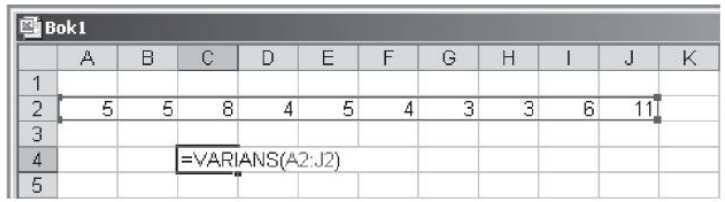
I Excel kan funktionen KVARTIL.EXK med syntaxen KVARTIL.EXK(dataområde;kvartil) användas för att beräkna kvartilerna för en datamängd. Av figur 1.4 framgår hur man matar in exemplets data för att beräkna första kvartilen.



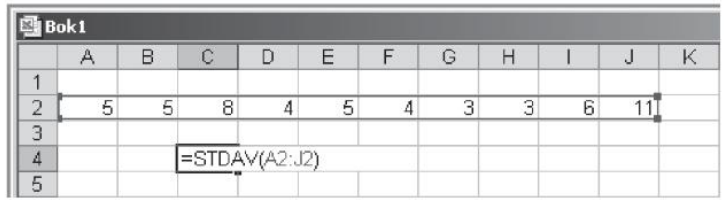
Figur 1.4: Kvartiler i Excel – inmatning

Vidare kan funktionerna VARIANS och STDAV användas för att beräkna respektive spridningsmått för stickprovsdata. I båda fallen anges som argument det område i kalkylarket där datamängden finns. Av figurerna 1.5 och 1.6 framgår hur inmatningen av exemplets data kan se ut innan man har tryckt på ”enter”. Om datamängden som man arbetar med är en population i stället för ett stickprov används i stället funktionerna

VARIANSP och STDAVP. Syntaxen är likadan, och enda beräkningsmässiga skillnaden är alltså att n då används i nämnaren i stället för n–1.



Figur 1.5: Stickprovsvarians i Excel – inmatning



Figur 1.6: Stickprovsstandardavvikelse i Excel – inmatning

Spontant kanske man undrar vad skillnaden mellan att använda standardavvikelse och varians såsom spridningsmått för en datamängd är. Standardavvikelsen är ju bara roten ur variansen, så vad är skillnaden i information hos de båda måtten? Jo, skillnaden beror just på det faktum att variansen består av kvadrerade data. Den ”enhet” i vilken variansen uttrycks är därför kvadraten av enheten i vilken den ursprungliga datamängden uttrycks. För datamängden i exempel 1-3 är medelvärdet 5,4 bilar men variansen måste alltså tolkas som 6,0444 ”kvadratbilar”. Genom att dra roten ur variansen får vi tillbaks den ursprungliga enheten, vilket innebär att standardavvikelsen i exemplet är 2,4585 bilar. Innebörden hos en standardavvikelse är alltså lättare att tolka rent praktiskt.

Vi kan också notera att formeln för s^2 (och därmed givetvis även för s eftersom s alltid är roten ur s^2) går att förenkla, vilket underlättar när man t.ex. ska räkna för hand. Den enklare versionen, vilken naturligtvis är matematiskt ekvivalent, framgår nedan.

Förenklad formel för variansen för ett stickprov om n observationer:

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

I exempel 1-3 ovan skulle variansen med denna formel ha beräknats som

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1} = \frac{3^2 + 3^2 + \ldots + 11^2 - \frac{(3+3+\ldots+11)^2}{10}}{10-1} = \frac{346 - \frac{54^2}{10}}{9}$$
$$= 6{,}0444.$$

Fördelen med den förenklade formeln ovan är att medelvärdet $\bar{x}$ inte alls behöver räknas fram för att man ska kunna beräkna variansen. Om man ändå redan har beräknat medelvärdet så finns det ytterligare en ekvivalent men förenklad formel för variansen som framgår nedan.

En annan förenklad formel för variansen för ett stickprov om n observationer:

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

I exempel 1-3 ovan skulle variansen med denna formel ha beräknats som

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2}{n-1} = \frac{(3^2 + 3^2 + \ldots + 11^2) - 10 \cdot 5{,}4^2}{10-1} = 6{,}0444\,.$$

1.6 Klasser och stapeldiagram

En datamängd kan ofta med fördel delas in i grupper. Om man ska göra en arbetsmarknadsundersökning kan det kanske vara lämpligt att dela in människor åldersmässigt i t.ex. grupperna ”under 18 år”, ”18–65 år” samt ”över 65 år”. En sådan grupp kallas för en *klass*. Sättet på vilket man delar in en datamängd i klasser kan vara godtyckligt, men faller sig ibland tämligen naturligt som i det nyss nämnda exemplet. En fördel med att dela in ett datamaterial i klasser på detta sätt är att man ofta kan presentera det på ett mycket överskådligare sätt. En nackdel är å andra sidan att man förlorar information då effekten blir att man gör om kvotdata till ordinaldata eller att man gör om ordinaldata till mindre exakt ordinaldata.

Ett mycket vanligt sätt att presentera klassificerade data är att använda sig av *stapeldiagram*. I ett stapeldiagram representeras varje klass av en stapel där stapelns höjd återspeglar det antal observationer som finns i den aktuella klassen. Sättet på vilket de olika klassernas frekvenser är fördelade på de olika klasserna kallas *frekvensfördelning*. Frekvenserna kan vara absoluta eller relativa; relationen mellan staplarnas höjd i diagrammet påverkas inte. Om respektive klass utgörs av endast ett variabelvärde som är diskret och kvantitativt så kallas diagrammet istället för stolpdiagram.

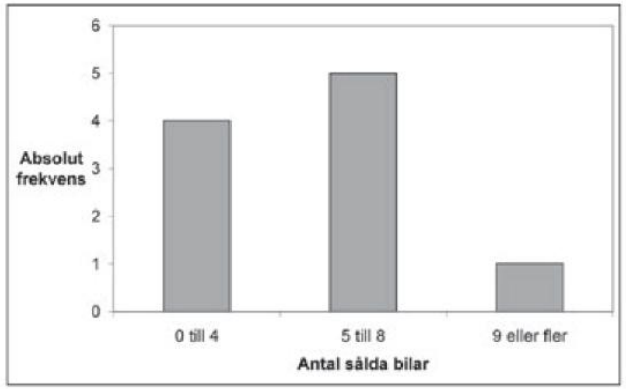
Exempel 1-4

Antalet sålda bilar per dag på Stures Bil AB de senaste 10 dagarna är 5, 5, 8, 4, 5, 4, 3, 3, 6, 11 stycken. Klassificera datan på lämpligt sätt och illustrera den med stapeldiagram.

Lösning

Eftersom det inte finns några självklara klassgränser är det godtyckligt hur dessa väljs. Exempelvis kan vi använda oss av de 3 klasserna ”0 till sålda 4 bilar”, ”5 till 8 sålda bilar” och ”9 eller fler sålda bilar”. I dessa klasser hamnar då 4, 5 respektive 1 av de 10 dagarna som datamängden omfattar. Ett stapeldiagram som illustrerar denna frekvensfördelning

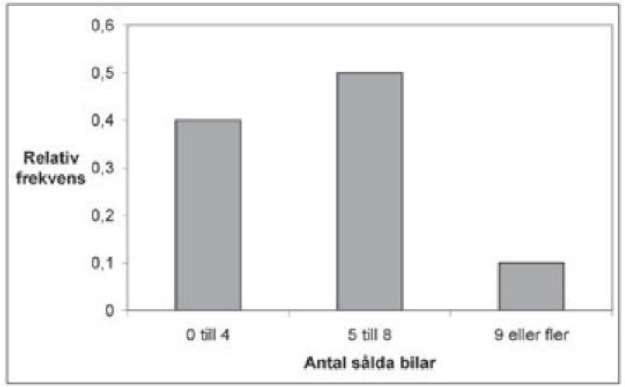
finns i figur 1.7 nedan. Summan av staplarnas höjder blir alltid lika med antalet observationer.



Figur 1.7: Stapeldiagram med absoluta frekvenser

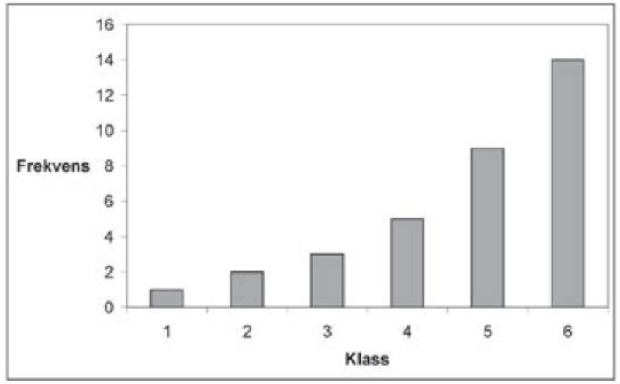
Man kan naturligtvis även illustrera datamängden med ett stapeldiagram utan att klassificera datan. Varje möjligt variabelvärde får då ”sin egen” stapel i diagrammet. Observera att när variabelvärdena bara kan vara heltal, som i exemplet ovan, så ritas staplarna med mellanrum. Om variabeln vars fördelning diagrammet illustrerar kan anta vilka värden som helst, alltså inte bara heltalsvärden, så ritas staplarna i diagrammet utan mellanrum för att illustrera kontinuiteten i variabeln. Diagrammet kallas då för histogram.

Man kan också låta staplarna belysa den relativa frekvensfördelningen. I detta fall kommer höjden på en stapel att återspegla den *andel* av den totala datamängden som stapelns klass omfattar. Summan av staplarnas höjder blir då alltid 1. Ett stapeldiagram som illustrerar detta finns i figur 1.8.

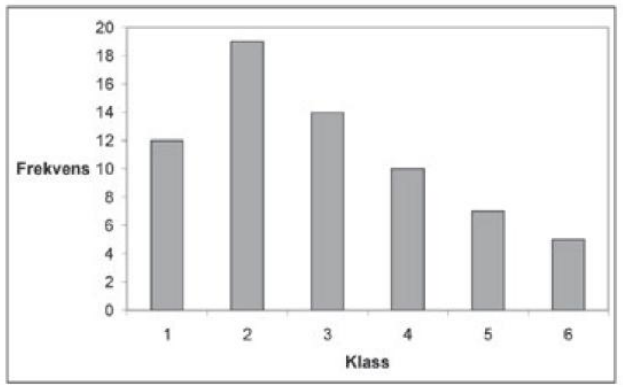


Figur 1.8: Stapeldiagram med relativa frekvenser

En frekvensfördelning kan alltså illustreras på ett tydligt sätt med ett stapeldiagram, och den kan beskrivas med mått på centraltendens och spridning. En annan egenskap hos en frekvensfördelning som kan vara intressant är dess *snedhet*. Graden av snedhet hos en frekvensfördelning är synonymt med graden av asymmetri. Ju snedare en frekvensfördelning är, desto mindre symmetrisk är den. Om den är utsträckt till vänster så sägs den vara sned åt vänster och tvärtom (se figur 1.9 och 1.10). En fördelning som är sned åt vänster kännetecknas allmänt av att medelvärdet är lägre än medianen, medan det motsatta förhållandet gäller när snedhet till höger råder.



Figur 1.9: En frekvensfördelning som är sned åt vänster



Figur 1.10: En frekvensfördelning som är sned åt höger

Snedheten för en datamängd beräknas som ett mått (formeln finns här nedanför) och resultatet blir ett tal som är större än 0 då snedhet åt höger råder respektive mindre än 0 om snedhet åt vänster råder. Om snedheten beräknas till 0 så innebär det att fördelningen är symmetrisk.

Snedheten för en datamängd med n element beräknas som

$$\sum_{i=1}^{n}\left[\frac{x_i-\bar{x}}{s}\right]^3\cdot\frac{n}{(n-1)(n-2)}$$

Exempel 1-5

En datamängd består av värdena 3, 7, 2, 4, 2, 2, 1. Beräkna dess medelvärde och standardavvikelse samt analysera dess snedhet.

Lösning

Medelvärde: $\bar{x}=\dfrac{2+7+3+4+2+2+1}{7}=3$

Standardavvikelse:

$$s=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}=\sqrt{\frac{(2-3)^2+(7-3)^2+\ldots+(1-3)^2}{7-1}}=2$$

Snedhet:

$$\sum_{i=1}^{n}\left[\frac{x_i-\bar{x}}{s}\right]^3\cdot\frac{n}{(n-1)(n-2)}=\left(\left[\frac{2-3}{2}\right]^3+\left[\frac{7-3}{2}\right]^3+\ldots+\left[\frac{1-3}{2}\right]^3\right)\cdot\frac{7}{6\cdot5}$$

$= 1{,}575$

Snedhetsmåttet är större än 0, vilket innebär att snedhet åt höger råder. Medianen är också lägre än medelvärdet (2 är lägre än 3). I Excel kan funktionen SNEDHET användas för att beräkna snedheten, varvid det område i kalkylarket där datamängden finns används som argument. Av figur 1.11 framgår hur man matar in exemplets data för att beräkna datamängdens snedhet.