
12. DESKRIPTION — FLERA VARIABLER

12.1 Jämförelser i frekvenstabeller

I detta avsnitt diskuterar vi hur man studerar beroenden och samband mellan olika stokastiska variabler då vi gör oberoende observationer på variablerna.

Liksom i fallet med oberoende observationer på en stokastisk variabel sammanfattar vi ett datamaterial bestående av observationer på flera stokastiska variabler i form av empiriska fördelningar. Det enklaste sättet att redovisa en empirisk fördelning för två eller flera variabler är i form av en korstabell. Korstabeller är den empiriska motsvarigheten till teorins simultana sannolikhetsfördelning. Variablernas värden definierar sålunda tabellens marginaler och observerade frekvenser skrivs in i tabellens celler. På motsvarande sätt som i sannolikhetsteorin kan vi även här summera frekvenserna ut till marginalerna och där avläsa variablernas empiriska marginalfördelningar.

Exempel 12.1

Vid en undersökning av hushållens konsumtionsvanor gjordes observationer på bl a variablerna “konsumtion av livsmedel” och “hushållens inkomst” hos 100 slumpmässigt utvalda hushåll. För att förenkla analysen delar vi in materialet i tre grupper efter hushållens inkomst. “Låginkomsthushåll” är sådana som har en inkomst som är lägre än 50% av urvalets medelinkomst och “Höginkomsthushåll” är sådana hushåll som har en inkomst som är större än 150% av urvalets medelinkomst. Gruppen däremellan utgör “Medelinkomsthushåll”.

I tabell 12.1 redovisas de observerade mätvärdena i form av en korstabell. Lägg märke till att korstabellen är en tvåvägsindeldad frekvenstabell. Lägg också märke till att kolumnen med beteckningen “samtlige” anger den empiriska marginalfördelningen för variabeln “konsumtion” dvs den envägsindeldade frekvenstabell vi skulle få om vi bara tabellerade variabeln “konsumtion”. På samma sätt anger raden med beteckningen “samtlige” den empiriska

marginalfördelningen för variabeln “inkomst”. Cellerna i tabellen anger de simultana frekvenserna, dvs det finns fem hushåll i materialet som är “låginkomsthushåll” och har en konsumtion som är högst 3900 kr.

Tabell 12.1 Hushållens konsumtion av livsmedel i tusentals kronor fördelade på inkomstnivåer

Konsumtion (kr)	Inkomstnivå			Samtliga
	låg	mellan	hög	
-3900	5	5	1	11
4000-5900	9	20	2	31
6000-7900	12	22	6	40
8000-9900	1	7	3	11
10000-11900	0	2	3	5
12000-	0	0	2	2
Samtliga	27	56	17	100

I korstabellen kan även relativa frekvenser anges. Detta åstadkomms genom att tabellens alla tal divideras med antalet observationer. Eftersom vi har 100 observationer i detta exempel får vi de relativa frekvenserna genom att tolka talen i tabell 12.1 som procenttal.

Vi påminner oss nu sannolikhetsteorins multiplikationssats som implicerar att två stokastiska variabler är oberoende om och endast om den simultana sannolikheten är lika med produkten av marginalsannolikheterna. På motsvarande sätt får vi en indikation om oberoende föreligger i verkligheten om den simultana empiriska relativa frekvensen är lika med produkten av de empiriska relativa marginalfrekvenserna. Vi ser t ex att för “låginkomsthushåll” med en konsumtion som är högst 3900 kr är den simultana empiriska relativa frekvensen 0,05. De empiriska relativa marginalfrekvenserna är 0,27 och 0,11 med produkten 0,0297. Att denna produkt avviker från 0,05 kan bero på att ett beroende mellan variablerna “inkomstnivå” och “konsumtion” föreligger. Skillnaden kan emellertid också vara ett uttryck för en slumpmässig avvikelse eftersom vi studerar endast ett urval av hushåll och inte hela populationen. I kapitel 18 diskuterar vi en formell metod, χ^2 -metoden, för att analysera en korstabell med avseende på beroende mellan variablerna. Denna metod

bygger just på att jämföra produkten av observerade marginalfrekvenser med observerade simultana frekvenser. ■

När vi talar om samband mellan variabler är det viktigt att ha sambandens *riktningar* klara för sig, dvs det är viktigt att hålla isär vilken eller vilka variabler som påverkar vilken annan variabel. Det är givetvis den teori vi stöder oss på som avgör sambandens riktningar. En variabel som påverkar andra variabler kallas ibland för förklaringsvariabel och en variabel som påverkas kallas ibland för effektvariabel. Vi låter här X beteckna en förklaringsvariabel och Y beteckna en effektvariabel, dvs vi följer beteckningskonventionen att X påverkar Y .

För att analysera samband kan man med fördel arbeta med betingade empiriska fördelningar. Betingade empiriska fördelningar är den empiriska motsvarigheten till teorins betingade sannolikhetsfördelningar. För ett visst givet värde på en förklaringsvariabel, säg $X = x_1$, bestämmer vi den empiriska fördelningen för en effektvariabel, säg Y . Denna bestämning sker helt enligt de principer vi diskuterade i kapitel 10 och på motsvarande sätt som betingade sannolikhetsfördelningar bestäms. I detta fall får vi den betingade empiriska frekvensen för Y givet $X = x_1$ som

$$f_n(y_j \mid x_1)$$

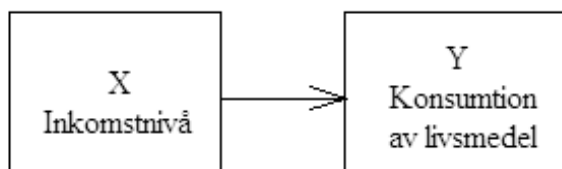
$$= \frac{\text{antal observationer där } X = x_1 \text{ och } Y = y_j}{\text{antal observationer där } X = x_1}$$

Vi har då bestämt den empiriska fördelningen för Y givet $X = x_1$. Därefter förfar vi på samma sätt för att bestämma den betingade empiriska fördelningen för Y givet ett eller flera andra värden på X , säg $X = x_2$, $X = x_3$ osv. Genom att jämföra de betingade fördelningarna vi får för Y kan vi få ett empiriskt stöd för eller emot att variabeln X påverkar variabeln Y . Om den betingade empiriska fördelningen för Y är densamma för de olika värdena på X , har X ingen empiriskt påvisbar effekt på Y . Om, å andra sidan, den betingade empiriska fördelningen för Y är olika för olika värden på X , har vi fått ett mönster som stöder en hypotes om att X påverkar Y .

Exempel 12.1 (fortsättning)

Enligt ekonomisk teori finns ett samband mellan inkomst och konsumtion. För livsmedel gäller att en ökad inkomst leder till en något ökad konsumtion. Vi skall därför studera om det finns skillnader i konsumtionen för de olika inkomstnivåerna som kan ge ett empiriskt stöd åt den ekonomiska teorin.

Variabeln “inkomstnivå”, med värdena “låg”, “mellan” och “hög” är således förklaringsvariabel, X , och variabeln “livsmedelskonsumtion” är effektvariabel, Y . Figur 12.1 visar ett exempel på hur detta samband kan åskådliggöras i ett pilschema.



Figur 12.1 Sambandet mellan “inkomstnivå” och “konsumtion”

De betingade empiriska fördelningarna för konsumtion av livsmedel givet inkomstnivå i vårt exempel redovisas i tabell 12.2. Tabellen är konstruerad så att förklaringsvariabeln sätts upp i tabellhuvudet och dess värden definierar olika kolumner i tabellen. Effektvariabeln anges i marginalen och dess värden får definiera olika rader i tabellen. De relativa frekvenserna beräknas så att varje kolumn summerar till 100% (alternativt till 1.00). De betingade fördelningarna jämförs genom att rader i tabellen jämförs. Det kan då inträffa att inga skillnader förekommer på någon rad. Detta tolkas som att det inte finns några skillnader mellan de olika betingade fördelningarna, dvs vi ser inget som tyder på att förklaringsvariabeln påverkar effektvariabeln. Om små skillnader förekommer mellan kolumnerna tolkas detta som att det finns små skillnader mellan de betingade fördelningarna, dvs X har en liten effekt på Y och, slutligen, om det förekommer stora skillnader mellan kolumnerna, tolkas detta som att X har en stor effekt på Y .

Tabell 12.2 Hushållens konsumtion av livsmedel fördelade på inkomstnivåer. %

Konsumtion (kr)	Inkomstnivå			Samtliga
	låg	mellan	hög	
-3900	18.5	8.9	5.9	11.0
4000-5900	33.3	35.7	11.8	31.0
6000-7900	44.5	39.7	35.3	40.0
8000-9900	3.7	12.5	17.6	11.0
10000-11900	0.0	3.6	17.6	5.0
12000-	0.0	0.0	11.8	2.0
Summa	100.0	100.0	100.0	100.0

Tabell 12.2 antyder stora skillnader i konsumtion av livsmedel för olika inkomstnivåer. En större andel av låginkomsthushållen har en låg konsumtion medan en större andel av höginkomsthushållen har en hög konsumtion. I tabell 12.3 visas ett exempel på en tabell utan samband. Där har konsumtionen samma betingade empiriska fördelning för alla inkomstnivåer. ■

Tabell 12.3 Fiktiv tabell över hushållens konsumtion av livsmedel fördelade på inkomstnivåer. %

Konsumtion (kr)	Inkomstnivå			Samtliga
	låg	mellan	hög	
-3900	8.9	8.9	8.9	8.9
4000-5900	35.7	35.7	35.7	35.7
6000-7900	39.7	39.7	39.7	39.7
8000-9900	12.5	12.5	12.5	12.5
10000-11900	3.6	3.6	3.6	3.6
12000-	0.0	0.0	0.0	0.0
Summa	100.0	100.0	100.0	100.0

Genom analys av betingade empiriska fördelningar (eller med hjälp av någon annan analysteknik) kan man få en bild av sambandet mellan två variabler. Den bild man får av sambandet i en tabell beror delvis på hur tabellen har utformats. Vi vet sedan tidigare att ett ändrat klassantal och

ändrade klassgränser i en frekvenstabell kan ge olika bilder av en empirisk fördelning och därmed även ge olika bilder av sambandet mellan variabler. Vidare påverkas bilden av sambandet av en mängd andra faktorer. Några av de vanligast förekommande faktorerna som stör bilden av samband är förekomsten av systematiska och/eller slumpmässiga fel, olikheter i definitioner, mätmetoder etc och förekomsten av andra viktiga förklaringsvariabler.

Övning 12.1

Tabellen nedan ger rökvanor hos studenter och deras föräldrar. Hur många studenter ingår i undersökningen? Hur stor andel av studenterna röker? Beräkna marginalfördelningen för föräldrarnas rökvanor.

	Studenten röker	Studenten röker inte
Båda föräldrarna röker	200	780
En förälder röker	208	1035
Ingen förälder röker	94	868

Övning 12.2

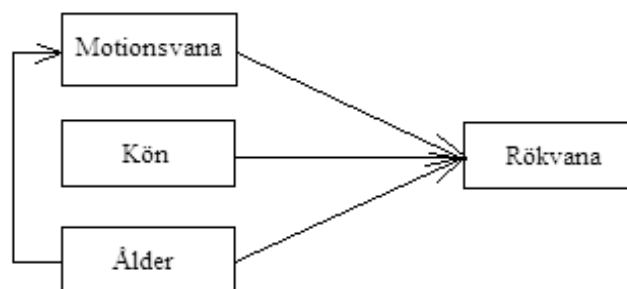
En fabrik tillverkar tre olika produkter, A, B och C. Produktionen av dessa sker vid två olika avdelningar, Avdelning 1 och Avdelning 2. De färdiga produkterna kontrolleras innan de levereras varvid defekta enheter sorteras bort. Man vill nu göra en jämförelse mellan produktionsavdelningarna avseende den totala kassationsfrekvensen. Man vet att kassationsfrekvensen inte är lika för de tre typerna av produkter och måste därför ta hänsyn till detta vid jämförelsen av avdelningarna. I nedanstående tabell redovisas antalet producerade enheter och antalet utsorterade enheter under 2007 fördelade efter produkt och avdelning.

Typ	Antal producerade enheter			Antal utsorterade enheter		
	Avd 1	Avd 2	Summa	Avd 1	Avd 2	Summa
A	3000	2000	5000	210	180	390
B	2000	4000	6000	40	160	200
C	1000	4000	5000	80	240	320
Summa	6000	10000	16000	330	580	910

Genomför jämförelsen av produktionsavdelningarna genom att ta hänsyn till skillnader i produktionsfördelning. Vad drar du för slutsats?

Övning 12.3

För att undersöka eventuella skillnader i rökvanor mellan män och kvinnor samlade man in vissa uppgifter från 60 slumpmässigt utvalda personer. Man utgick ifrån följande modell:



Analysera sambandet mellan kön och rökvana, Sätt upp en lämplig tabell och tolka resultatet i ord. Data till övningen finns i Excel-filen Rökdata.

12.2 Jämförelser i kvottabeller

I föregående avsnitt studerade vi samband mellan variabler genom att jämföra betingade empiriska fördelningar. Fördelningar kan också jämföras genom att jämföra karakteristika såsom läges och spridningsmått. I det här avsnittet behandlar vi hur samband mellan variabler kan studeras genom att jämföra betingade medelvärden, dvs medelvärden i de betingade fördelningarna. Dessa medelvärden sammanställs i så kallade envägsindelade kvottabeller. Även dessa tabeller ställer vi upp så att man gör radvisa jämförelser. Då antyder alltså inga skillnader mellan kolumner i tabellen att inga skillnader mellan de betingade fördelningarna och inget samband mellan variablerna föreligger. Små kolumnskillnader tolkas som att det finns små skillnader mellan de betingade fördelningarna, dvs förklaringsvariabeln har en inverkan effekt på effektvariabeln. Om, slutligen, det finns åtminstone någon stor

skillnad mellan kolumner tolkas detta som att förklaringsvariabeln kan ha en stor inverkan på effektvariabeln.

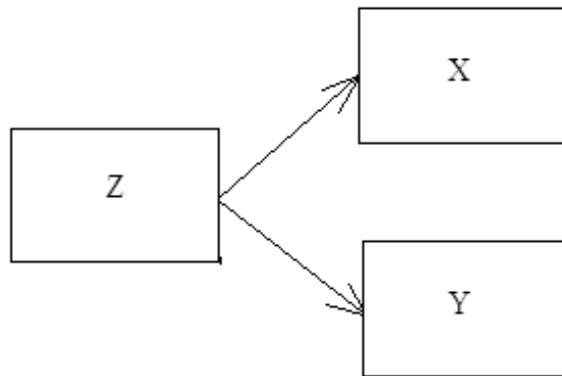
Exempel 12.1 (fortsättning)

Betingade medelvärden redovisas i kvottabellen i tabell 12.4. Tabellen antyder stora skillnader i konsumtion mellan olika inkomstnivåer. Hushåll med högre inkomst tenderar att ha en högre konsumtion. ■

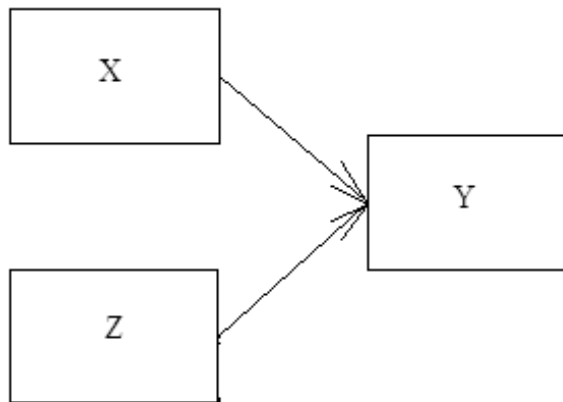
Tabell 12.4 Hushållens medelkonsumtion av livsmedel (kr).

Inkomstnivå			Samtliga
låg	mellan	hög	
5358	6226	8530	6384

Det är här på sin plats att notera en viktig skillnad mellan frekvenstabeller och kvottabeller, nämligen att i frekvenstabeller kan värden summeras till marginalernas värden vilket inte är möjligt i kvottabeller.



Figur 12.2 En variabel Z påverkar både X och Y .



Figur 12.3 Två förklaringsvariabler, X och Z , påverkar effektvariabeln Y .

Observerade skillnader i en tabell kan påverkas mycket om man har uteslutit någon viktig förklaringsvariabel i analysen. Figur 12.2 visar ett pilschema där en tredje variabel, Z , kan påverka både förklaringsvariabeln X och effektvariabeln Y . Här kan ett observerat samband mellan X och Y uppstå genom att Z påverkar X och Z påverkar Y , utan att det därför finns någon direkt koppling mellan X och Y . Ett berömt exempel är förekomsten av storkar, X , och nativiteten, Y , som båda beror på urbaniseringsgrad (tätort/landsbygd), Z . Ett sådant förhållande kan ge att där det finns många storkar, föds det också många barn och där det finns få storkar föds det få barn, utan att förekomsten av storkar har något direkt att göra med barnafödande.

Figur 12.3 visar en annan vanlig situation där Z är en andra förklaringsvariabel. Familjens storlek, mätt t ex med antal barn, är ett exempel på en sådan variabel som kan vara aktuell i exempel 12.1. Följande tre exempel visar hur en extra förklaringsvariabel kan påverka jämförelser mellan betingade medelvärden.

Exempel 12.2.

Antag att vi har ett material som består av konsumtion av livsmedel för sex hushåll:

Konsumtion	Inkomstnivå	Antal barn
2000	Låg	0
3200	”	0
5600	”	4
2600	Hög	0
6000	”	4
5200	”	4

Om vi bortser från förklaringsvariabeln “antal barn” får vi de betingade medelvärden som redovisas i den envägsindelade kvottabellen 12.5.

Tabell 12.5 Betingad medelkonsumtion av livsmedel för hushåll med låg respektive hög inkomstnivå.

Inkomstnivå	
Låg	Hög
3600	4600

I tabell 12.5 finns stora skillnader mellan medelkonsumtionen av livsmedel för hushåll i de två inkomstnivåerna: hushåll med hög inkomst har en hög medelkonsumtion. Om vi nu tar hänsyn till variabeln “antal barn” och betingar även på den, får vi de betingade medelvärden som redovisas i den tvåvägsindelade kvottabellen 12.6.

Tabell 12.6 Betingad medelkonsumtion av livsmedel för hushåll med låg respektive hög inkomstnivå.

Antal barn	Inkomstnivå	
	Låg	Hög
0	2600	2600
4	5600	5600

Även i denna tabell gör vi radvisa jämförelser. Varje sådan radvis jämförelse innebär att hushåll med låg respektive hög inkomst men med samma antal barn jämförs. En eventuell skillnad kan då inte bero på variabeln “antal barn”, utan tolkas som en skillnad mellan enbart inkomstnivåer. Den bild

av sambandet som tabell 12.5 ger är alltså i det avseendet missvisande. Det är inte inkomstnivån som påverkar konsumtionen av livsmedel, utan det är olikheter i hushållens storlek som ger de observerade skillnaderna i konsumtion. ■

Exemplet visar hur en tvåvägsindeldad kvottabell kan användas för att studera sambandet mellan en förklaringsvariabel X och en effektvariabel Y så att analysen inte störs av en annan förklaringsvariabel Z . Tabellen konstrueras så att den förklaringsvariabel X vars effekt på effektvariabeln Y vi önskar studera sätts upptill i tabellhuvudet och den andra förklaringsvariabeln, Z , anges i marginalen. I tabellens olika celler anges betingade medelvärden av Y givet både X och Z . Genom att göra radvisa jämförelser i denna tabell undersöker man sambandet mellan X och Y . När man gör en sådan radvis jämförelse varierar X medan den andra förklaringsvariabeln Z är konstant. Man ser då effekten på Y av att endast X varierar, dvs X -variabelns effekt blandas inte ihop med den andra förklaringsvariabelns effekt på Y . Följande två exempel visar hur bilden av ett sambandsmönster mellan två variabler X och Y kan förändras då vi tar hänsyn till en annan förklaringsvariabel Z .

Exempel 12.3.

Antag att vi har följande material:

Konsumtion	Inkomstnivå	Antal barn
3200	Låg	0
5200	”	4
6000	”	4
3200	Hög	0
4000	”	0
6000	”	4

Tabell 12.7 Betingad medelkonsumtion av livsmedel för hushåll med låg respektive hög inkomstnivå.

Inkomstnivå	
Låg	Hög
4800	4400

I tabell 12.7 redovisas olika medelkonsumtion för olika inkomstnivåer: hushåll med låg inkomst har högre medelkonsumtion. Om vi nu också tar hänsyn till variabeln “antal barn” får vi de betingade medelkonsumtionerna som redovisas i tabell 12.8. Där framgår att för hushåll med samma antal barn har de med högre inkomstnivå också en högre medelkonsumtion. Sambandet i tabell 12.7 har alltså en helt annan karaktär än sambandet i tabell 12.8. ■

Tabell 12.8 Betingad medelkonsumtion av livsmedel för hushåll med låg respektive hög inkomstnivå.

Antal barn	Inkomstnivå	
	låg	hög
0	3200	3600
4	5600	6000

Exempel 12.4.

Antag att vi har följande material:

Konsumtion	Inkomstnivå	Antal barn
1400	Låg	0
4200	”	4
4600	”	4
2000	Hög	0
3200	”	0
5000	”	4

Tabell 12.9 Betingad medelkonsumtion av livsmedel för hushåll med låg respektive hög inkomstnivå

Inkomstnivå	
Låg	Hög
3400	3400

Tabell 12.10 Betingad medelkonsumtion av livsmedel för hushåll med låg respektive hög inkomstnivå.

Antal barn	Inkomstnivå	
	låg	hög
0	1400	2600
4	4400	5000

I tabell 12.9 antyds inga skillnader i konsumtion av livsmedel mellan hushåll med olika inkomstnivåer. Tar vi däremot även hänsyn till variabeln “antal barn” ser vi i tabell 12.10 att skillnader i konsumtion föreligger: hushåll med högre inkomst har en högre medelkonsumtion av livsmedel. ■

En viktig slutsats från exemplen 12.2 - 12.4 är att analys av sambandet mellan en förklaringsvariabel X och en effektvariabel Y påverkas om det finns samband mellan en förklaringsvariabel Z och Y och det finns ett samband mellan X och Z .

När vi studerar samband mellan variabler är det vanligt att analysera detta i termer av förändringar i betingade medelvärden på det sätt vi har diskuterat här. Det är emellertid möjligt att även tänka sig att ett förändrat värde på förklaringsvariabeln skall resultera i förändringar i andra karakteristika för den betingade fördelningen än det betingade medelvärdet. Vi skall dock inte diskutera sådana varianter här. Däremot är det viktigt att komma ihåg att ett medelvärde kan ge mer eller mindre träffande beskrivning av läget hos ett material beroende på spridningen i materialet. Det kan därför vara viktigt att även studera spridningarna i de betingade empiriska fördelningarna, eftersom deras storlek har en viss betydelse för hur sambandet skall tolkas. Följande två exempel illustrerar detta.

Exempel 12.5.

Betrakta först följande sex observationer på konsumtion av livsmedel hos hushåll (med samma antal barn)

Inkomstnivå	
låg	hög
4800	6000
4400	5600
4900	5200

De betingade medelvärdena, redovisade i tabell 12.11, antyder en viss skillnad i konsumtion mellan hushåll med låg respektive hög inkomstnivå. Eftersom spridningarna i de båda betingade fördelningarna är små, ger de betingade medelvärdena en god beskrivning av fördelningarnas läge. Vi säger att variabeln ”inkomstnivå” förklarar mycket av observerade skillnader i variabeln ”konsumtion”, eller, alternativt, att vi har ett starkt samband i statistisk bemärkelse.

Tabell 12.11 Betingad medelkonsumtion av livsmedel för hushåll med låg respektive hög inkomstnivå.

Inkomstnivå	
låg	hög
4400	5600

Betrakta nu följande sex observationer på konsumtion av livsmedel hos hushåll

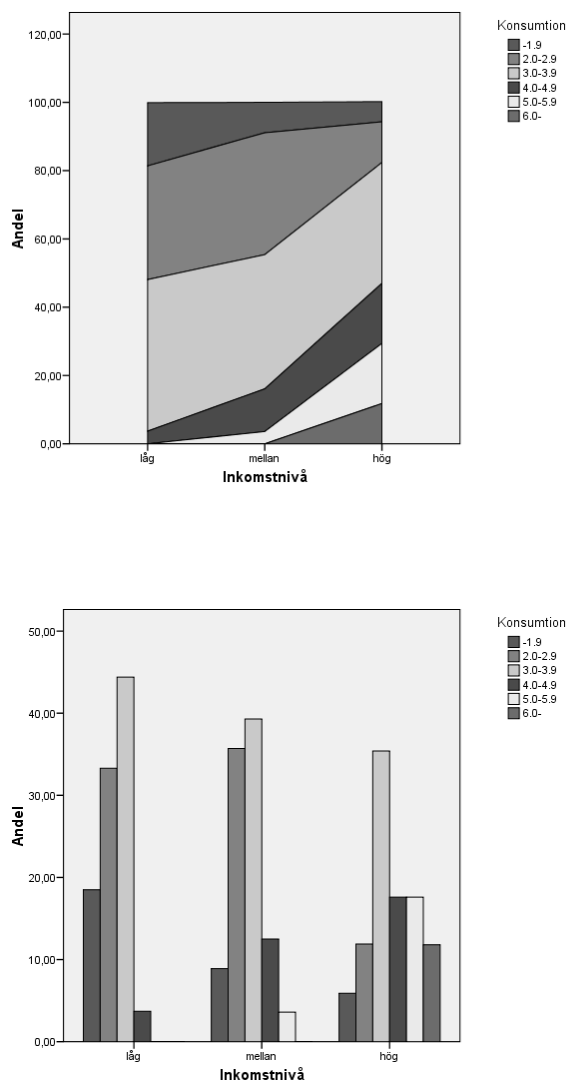
Inkomstnivå	
låg	hög
7800	9200
4400	5600
1000	2000

Detta material har samma betingade medelvärden som det förra materialet, redovisade i tabell 12.11. Däremot har de betingade fördelningarna för detta material stora spridningar. De betingade medelvärdena ger därför en dålig beskrivning av fördelningarnas läge. Vi säger i detta fall att variabeln ”inkomstnivå” förmår att förklara endast en liten del av observerade skillnader i variabeln ”konsumtion”. ■

Slutligen skall påpekas något som alltför lätt glöms bort: den statistiska analysen kan inte ”bevisa” att några samband mellan variabler existerar, den kan endast ge ett empiriskt stöd åt teoretiskt motiverade samband.

12.3 Grafisk framställning

Den grafiska framställningen av datamaterial som omfattar flera variabler följer samma grundprinciper som den grafiska framställningen av observationer på en variabel. En viktig skillnad är dock att man i flervariabelfallet



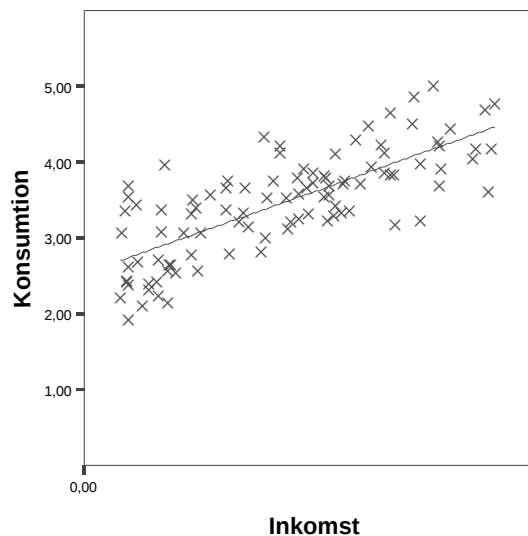
Figur 12.4 Hushållens konsumtion av livsmedel fördelade på inkomstnivåer.

ofta vill göra jämförelser mellan olika kategorier eller delgrupper. Detta åstadkomms genom att diagram över betingade empiriska fördelningar framställs. I figur 12.4 visas några exempel på diagram baserade på tabell 12.2.

Antag nu att vi har en förklaringsvariabel, X , och en effektvariabel, Y , och att båda är kontinuerliga. Om vi betraktar observationerna på X och Y som talpar, (x_1, y_1) , (x_2, y_2) , ..., (x_n, y_n) , kan dessa läggas in som punkter i ett koordinatsystem. Det diagram vi då får kallas punktdiagram.

Exempel 12.1 (fortsättning)

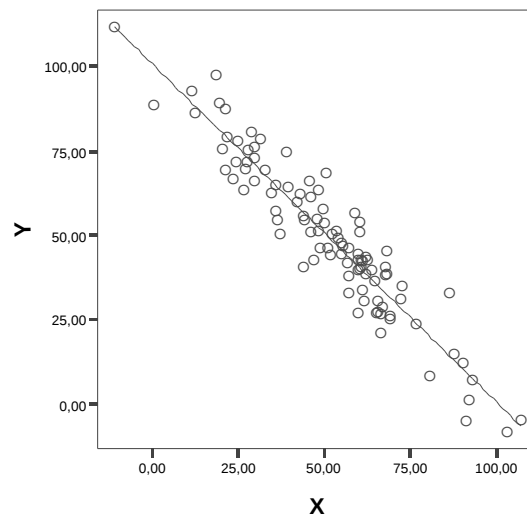
Både "hushållens inkomst" och "konsumtion av livsmedel" är kontinuerliga variabler. I figur 12.5 visas ett punktdiagram över materialet.



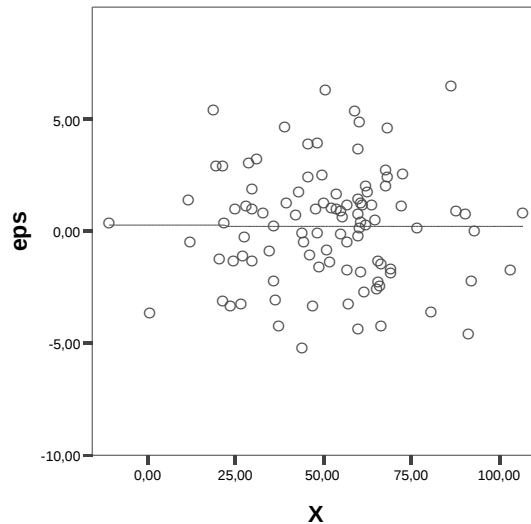
Figur 12.5 Observationer på hushållens inkomst och konsumtion av livsmedel redovisade i ett punktdiagram.

Av diagrammet framgår att hög inkomst är förknippad med hög konsumtion och att låg inkomst är förknippad med låg konsumtion. Vi säger då att det finns ett positivt samband mellan variablerna.

I diagrammet har vi även ritat in en rät linje, en så kallad regressionslinje. Den beskriver det betingade medelvärdet för konsumtion givet inkomsten som en linjär funktion av inkomsten. I den linjära regressionsanalysen studeras hur man bestämmer ekvationen för en sådan linje, något som emellertid faller utanför ramen för den här framställningen. ■



Figur 12.6 Negativt samband mellan X och Y .

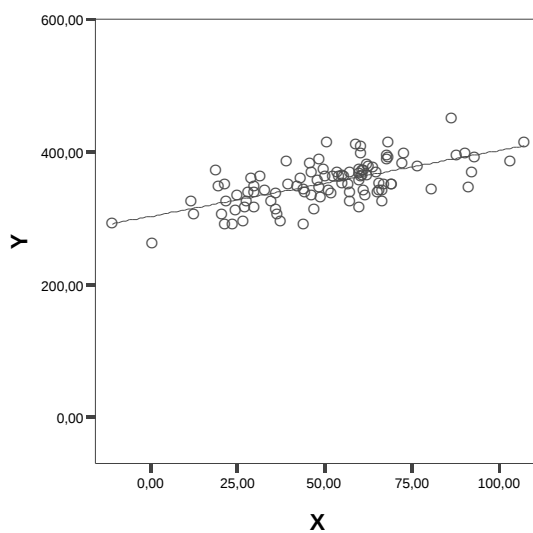
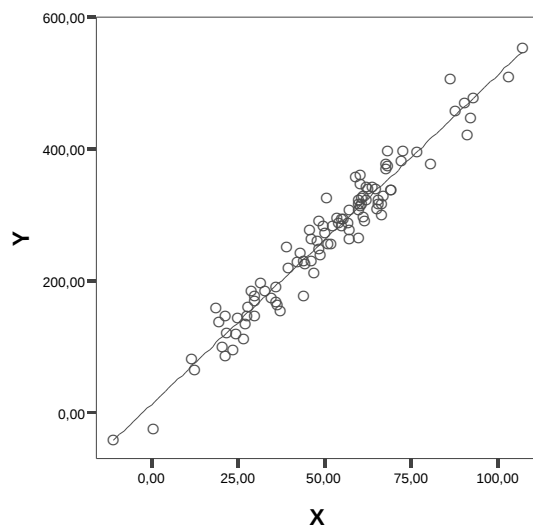


Figur 12.7 Inget samband mellan X och Y

Vi ger nu några exempel på punktdiagram som illustrerar olika datamaterial där sambandet mellan förklaringsvariabeln X och effektvariabeln Y har olika karaktär. I figur 12.6 visas ett exempel på ett negativt samband mellan X och Y . Då X ökar, avtar Y och tvärt om. Om en rät linje anpassas till observationerna i koordinatsystemet så har den en negativ lutning.

Ett exempel där inget samband mellan X och Y råder visas i figur 12.7. Då X ökar varken ökar eller minskar Y . Om en rät linje anpassas till punkterna i punktdiagrammet så har den lutningen noll.

Om det betingade medelvärdet förändras mycket då förklaringsvariabeln förändras får regressionslinjen en brant lutning. Förklaringsvariabeln har då en kraftig inverkan på effektvariabeln. Om, å andra sidan, det betingade medelvärdet endast har en marginell förändring då förklaringsvariabeln förändras får regressionslinjen en flack lutning och förklaringsvariabeln har en svag inverkan på effektvariabeln. Detta illustreras i de två punktdiagrammen i figur 12.8



Figur 12.8 Starkt respektive svagt samband mellan förklaringsvariabeln och effektvariabeln.

I föregående avsnitt diskuterade vi betydelsen av spridningen kring ett betingat medelvärde. Spridningen kring en regressionslinje har motsvarande betydelse. I figur 12.9 visas två exempel med olika stor spridning kring regressionslinjen. I exemplet med liten spridning är sambandsmönstret tydligt. Vi säger då att vi har en stor förklaringsgrad eller ett starkt samband i statistisk bemärkelse. I exemplet med stor spridning kring regressionslinjen är sambandsmönstret oklart. Vi säger då att vi har en liten förklaringsgrad eller ett svagt samband i statistisk bemärkelse. Notera här att i båda exemplen är regressionslinjens lutning densamma.

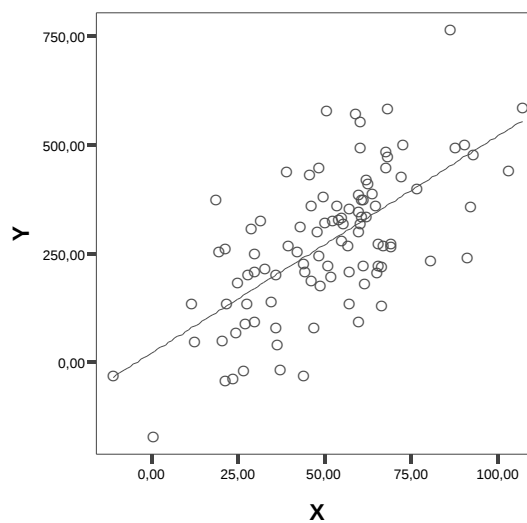
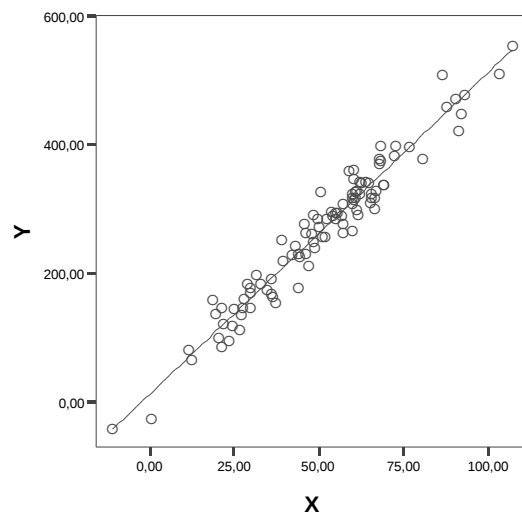
Det betingade medelvärdet behöver naturligtvis inte vara en linjär funktion av förklaringsvariabeln. I figur 12.10 visas ett exempel på ett fall då sambandet (och därmed regressionslinjen) inte är linjära.

Övning 12.4

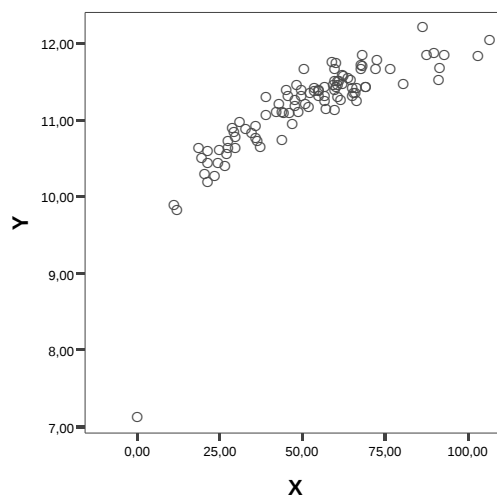
Hur påverkas bränsleförbrukningen av hastigheten hos bilar? I tabellen nedan redovisas observationer på hastighet (km/h) och bensinförbrukning (l/mil) för en Ford Escort.

Hastighet (km/h)	Förbrukning (l/mil)	Hastighet (km/h)	Förbrukning l/(mil)
10	2,10	90	0,76
20	1,30	100	0,83
30	1,00	110	0,90
40	0,80	120	0,99
50	0,70	130	1,08
60	0,59	140	1,18
70	0,63	150	1,28
80	0,70		

a) Gör ett punktdiagram (vilken är förklaringsvariabeln?). Beskriv formn på sambandet. Ge en förklaring på formen på sambandet. Förklara varför sambandet varken är positivt eller negativt. Avgör om sambandet är starkt.



Figur 12.9 Starkt resprktive svagt samband i statistisk bemärkelse mellan förklaringsvariabeln och effektvariabeln.



Figur 12.10 Icke linjärt samband

12.4 Kovarians och korrelation

I kapitel 6.7 definierade vi kovarians mellan två stokastiska variabler, och diskuterade hur kovariansen kan användas för att beskriva ett linjärt samband mellan variablerna. På motsvarande sätt kan vi definiera en observerad kovarians mellan variabler då vi har ett urval med observationer på variablerna. Låt därför $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ vara n par av observationer på två variabler. Urvalskovariansen definieras då genom

$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

Då man skall beräkna en kovarians är det ofta enklare att använda uttrycket

$$s_{XY} = \frac{1}{n-1} \left(\sum_{i=1}^n x_i y_i - \frac{\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n} \right).$$

Exempel 12.6

En studie hade som syfte att studera eventuellt samband mellan kroppsvikt och energiomsättning. Tabellen nedan visar observationer på kroppsvikt utan fett i kg, x , och energiförbränning under 24 timmar i kcal, y , hos 12 kvinnor som ingick i studien. Tabellen visar också beräkningar för att bestämma kovariansen mellan variablerna kroppsvikt utan fett och energiförbränning.

i	x	y	xy	x^2	y^2
1	42,0	1422	59724,0	1764,00	2022084
2	48,5	1398	67803,0	2352,25	1954404
3	54,2	1432	77614,4	2937,64	2050624
4	36,5	995	36317,5	1332,25	990025
5	51,2	1525	78080,0	2621,44	2325625
6	42,0	1256	52752,0	1764,00	1577536
7	40,3	1179	47513,7	1624,09	1390041
8	33,3	951	31668,3	1108,89	904401
9	41,2	1208	49769,6	1697,44	1459264
10	51,1	1345	68729,5	2611,21	1809025
11	34,2	1052	35978,4	1169,64	1106704
12	42,0	1118	46956,0	1764,00	1249924
Summa	516,5	14881	652906,4	22746,85	18839657

Med hjälp av summorna i tabellen kan nu kovariansen beräknas enligt

$$\begin{aligned}s_{XY} &= \frac{1}{12-1} \left(652906,4 - \frac{516,5 \cdot 14881}{12} \right) = \frac{1}{11} (652906,4 - 640503,0417) \\ &= 12403,3583/11 = 1127,5780 \blacksquare\end{aligned}$$

Tolkningen av s_{XY} svarar mot tolkningen av $Cov(X, Y)$. Observationer (x_i, y_i) där både x_i och y_i är större än sina respektive medelvärden kommer att bidra positivt till summan i definitionen av s_{XY} . På samma sätt bidrar observationer där både x_i och y_i är mindre än sina respektive medelvärden positivt till summan. Om, å andra sidan, en av x_i och y_i är större än sitt medelvärde och den andra är mindre än sitt medelvärde får vi en negativ

term i summan. Positiva samband kommer därför att ge upphov till positiva kovarianser och negativa samband ger upphov till negativa kovarianser.

Kovarianser beräknade på observationsmaterial är beroende av skalorna hos de ingående variablerna. Man får olika kovarianser om man mäter variablerna i meter eller om man mäter dem i centimeter, t ex. Detta är en egenskap som delas med kovarianser för stokastiska variabler. Det är därför ofta av intresse att beräkna urvalskorrelationen, definierad geom

$$r_{XY} = s_{XY}/s_x s_Y,$$

där s_X och s_Y är standardavvikelserna för observationerna x_i respektive y_i .

Man kan visa att $-1 \leq r_{XY} \leq 1$ och att $|r_{XY}|$ är nära 1 då det råder ett nästan exakt linjärt samband mellan variablerna.

Exempel 12.6 (fortsättning)

Ur tabellen över datamaterialet finns också uträknat de summor vi behöver för att räkna ut varianserna och standardavvikelserna.

$$\begin{aligned} s_X^2 &= \frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - \frac{\sum_{i=1}^n x_i}{n} \right) = \frac{1}{12-1} \left(22746,85 - \frac{516,5^2}{12} \right) \\ &= \frac{1}{11} (22746,85 - 22231,0208) = 46,8936 \end{aligned}$$

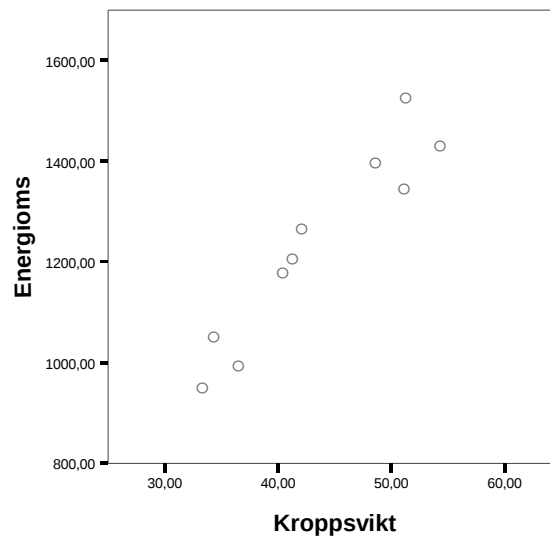
och på samma sätt

$$\begin{aligned} s_Y^2 &= \frac{1}{12-1} \left(18839657 - \frac{14881^2}{12} \right) \\ &= \frac{1}{11} (18839657 - 18453680,0833) = 35088,8106. \end{aligned}$$

Korrelationen blir därför

$$r_{XY} = \frac{1127,5780}{\sqrt{46,8936}\sqrt{35088,8106}} = 0,8790$$

Vi ser att korrelationen är, liksom kovariansen, positiv, dvs det råder ett positivt samband mellan variablerna kroppsvikt och energiomsättning. Vi ser vidare att sambandet är starkt eftersom korrelationen är relativt stor. Ett punktdiagram över observationerna visas i figur 12.11. ■



Figur 12.11 Punktdiagram över observationer på energiomsättning och kroppsvikt

12.5 Mera om betingade medelvärden - samspelseffekt

I detta avsnitt fördjupar vi diskussionen kring problemet hur man studerar ett samband mellan en förklaringsvariabel X och en effektvariabel Y då sambandet påverkas av en annan förklaringsvariabel Z . Diskussionen sker inom ramen för följande exempel.

Exempel 12.1 (fortsättning)

Vi vill studera sambandet mellan “hushållens inkomstnivå”, X , och “konsumtion av livsmedel”, Y , då vi kompenserar för eventuella effekter orsakade av familjens storlek uttryckt i “antal barn”, Z . En lämplig tvåvägsindeld kvottabell redovisas i tabell 12.12.

Tabell 12.12 Betingade medelvärden för konsumtion av livsmedel.

Antal barn	Inkomstnivå		
	låg	mellan	hög
0	1292	4660	5760
1	3254	4720	5772
2	4828	6272	7848
3	6660	7558	12668
4	7900	9500	-
Samtliga	5358	6232	8529

På raden som betecknas “Samtliga” redovisas de betingade medelvärdena givet endast inkomstnivå (dvs samma som redovisas i tabell 12.4). Om man jämför de betingade medelvärdena givet inkomstnivå finns det risk för att man får en missvisande bild av sambandet — inkomstnivåns effekt på konsumtionen kan blandas samman med effekter av andra förklaringsvariabler. I tabellens celler redovisas betingad medelkonsumtion givet både inkomstnivå och antal barn, dvs medelkonsumtionen i de olika delgrupperna av hushåll som definieras av värdena på de båda förklaringsvariablerna “inkomstnivå” och “antal barn”. När man jämför betingade medelvärden av konsumtionen på samma rad så blandas inte effekten av “antal barn” samman med effekten av “inkomstnivå”. Om vi t ex jämför medelvärdena för inkomstnivåerna “låg” och “mellan” ser vi att skillnaderna i medelvärdena är

för 0 barn: $4660 - 1292 = 3368$

för 1 barn: $4720 - 3254 = 1466$

för 2 barn: $6272 - 4828 = 1444$

för 3 barn: $7558 - 6660 = 8898$

för 4 barn: $9500 - 7900 = 1\,600$

för samtliga: $6226 - 5358 = 868$

Vi ser här att de betingade medelvärdena givet endast “inkomstnivå” underskattar skillnaden i konsumtion mellan dessa två inkomstnivåer. Eftersom hushållsstorlekens fördelning är olika i dessa inkomstnivåer beror skillnaden mellan totalmedelvärdena både på skillnaden mellan inkomstnivå och på skillnader i hushållsstorleken mätt som antal barn.

Då vi jämför de betingade medelvärdena i tabell 12.12 ser vi att hushåll med låg inkomstnivå har låg medelkonsumtion och hushåll med hög inkomstnivå har hög medelkonsumtion. Detta är ytterligare ett exempel på vad vi i föregående avsnitt kallar positivt samband. Om istället låg inkomstnivå haft hög konsumtion och hög inkomstnivå låg konsumtion skulle vi haft ett negativt samband.

Vi ser vidare att skillnaderna mellan medelkonsumtionerna är olika för olika storlekar på hushållet. Detta är ett exempel på vad som kallas *samspelseffekt* (interaktion). Man säger att det finns samspel mellan variablerna “inkomstnivå” och “antal barn”, dvs mellan de två förklaringsvariablerna. En tabell utan samspel skulle kunna se ut som tabell 12.13

Tabell 12.13 Fiktiva betingade medelvärden för konsumtion av livsmedel.

Antal barn	Inkomstnivå		
	låg	mellan	hög
0	2660	4660	6260
1	2720	4720	6320
2	4272	6272	7872
3	5558	7558	9158
4	7500	9500	11100
Samtliga	4712	6232	7688

Även i denna tabell finns ett samband mellan “inkomstnivå” och “konsumtion”. Hushåll med högre inkomstnivå har en högre medelkonsumtion än hushåll med lägre inkomstnivå. Däremot finns inget samspel mellan “inkomstnivå” och “antal barn” eftersom skillnaden i konsumtion mellan hushåll i de olika inkomstnivåerna är densamma för alla familjestorlekar. ■

12.6 Standardvägning — standardpopulationsmetoden

I de föregående avsnitten har vi diskuterat en metod baserad på jämförelser av betingade medelvärden för att analysera samband mellan variabler. I det här avsnittet beskriver vi en alternativ metod att studera samband. Vi börjar med att studera hur totalmedelvärden och betingade medelvärden är relaterade till varandra. Man kan då bättre förstå varför effekten av olika förklaringsvariabler kan blandas samman om man jämför totalmedelvärden. Vi genomför beräkningarna för exemplet med hushållens konsumtionsvanor, som torde vara ett välbekant exempel vid det här laget.

Tabell 12.14 Antal barn och konsumtion av livsmedel hos låginkomsthushåll

Antal barn	Antal hushåll	Procent	Konsumtion
0	1	3.7	1292
1	5	18.5	3254
2	9	33.3	4828
3	9	33.3	6660
4	3	11.1	7900
Samtliga	27	100.0	5358

Exempel 12.1 (fortsättning)

Betingade medelvärden och antal barn för låginkomsthushållen redovisas i tabell 12.14. Den genomsnittliga konsumtionen av livsmedel för låginkomsthushåll är 5358 kr. Detta kan beräknas genom att dividera låginkomsthushållens sammanlagda konsumtion med antalet låginkomsthushåll. Den sammanlagda konsumtionen är, i sin tur, summan av konsumtionerna för varje hushållsstorlek. Eftersom det finns ett hushåll med 0 barn och medelkonsumtionen är 1292 kr är totala konsumtionen för den hushållsstorleken $1 \cdot 1292 = 1292$ kr. Det finns fem hushåll med ett barn. Medelkonsumtionen för den hushållsstorleken är 3254 kr varför gruppens totala konsumtion är $5 \cdot 3254 = 16\,270$ kr, o s v. Den sammanlagda konsumtionen blir därmed $1 \cdot 1292 + 5 \cdot 3254 + \dots + 3 \cdot 7900 = 1292 + 16270 + \dots + 23700 = 144654$ kr och medelkonsumtionen blir $144654/27 = 5358$ kr. Detta medelvärde kan också beräknas med hjälp av de betingade medelvärdena på följande sätt:

$$\begin{aligned}
5358 &= \frac{1 \cdot 1292 + 5 \cdot 3254 + 9 \cdot 4828 + 9 \cdot 6660 + 3 \cdot 7900}{27} \\
&= \frac{1}{27} \cdot 1292 + \frac{5}{27} \cdot 3254 + \frac{9}{27} \cdot 4828 + \frac{9}{27} \cdot 6660 + \frac{3}{27} \cdot 7990 \\
&\approx 0,037 \cdot 1292 + 0,185 \cdot 3254 + 0,333 \cdot 4828 + 0,333 \cdot 6660 + 0,111 \cdot 7990
\end{aligned}$$

Det betingade medelvärdet givet inkomstnivå fås alltså genom att väga samman de betingade medelvärdena givet "inkomstnivå" och "antal barn" med de relativa frekvenserna för "antal barn" som vikter. För beräkning av låginkomsthushållens betingade medelvärde används således viktsystemet:

$$0,037 \ 0,185 \ 0,333 \ 0,333 \ 0,111$$

De betingade medelvärdena givet inkomstnivå för medel- respektive höginkomsthushåll får vi genom att väga samman deras respektive betingade medelvärden givet inkomstnivå och antal barn (se tabell 12.12) med relativa frekvenser för antal barn i respektive inkomstnivå som vikter. För medelinkomsthushållen får vi alltså

$$\begin{aligned}
&0,036 \cdot 4660 + 0,304 \cdot 4720 + 0,339 \cdot 6272 + 0,286 \cdot 7558 + 0,036 \cdot 9500 \\
&\approx 6232
\end{aligned}$$

Vi har alltså använt viktsystemet

$$0,036 \ 0,304 \ 0,339 \ 0,286 \ 0,036.$$

På motsvarande sätt genomför vi beräkningarna för höginkomsthushållen

$$\begin{aligned}
&0,235 \cdot 5760 + 0,118 \cdot 5772 + 0,353 \cdot 7848 + 0,294 \cdot 12668 + 0 \\
&= 8529
\end{aligned}$$

Här är viktsystemet således

$$0,235 \ 0,118 \ 0,353 \ 0,294 \ 0,000$$

Skillnaden mellan de betingade medelvärdena givet endast “inkomstnivå” beror alltså dels på att hushåll med olika inkomstnivå har olika medelkonsumtion av livsmedel och dels på att man använder olika viktsystem vid beräkningen (på grund av att variablerna “inkomstnivå” och “antal barn” samvarierar). ■

I fall då både betingade medelvärden och viktsystem är olika, kommer effekterna av olika förklaringsvariabler att sammanblandas. Observera här att det viktsystem vi använder är faktiskt den betingade empiriska fördelningen för den variabel vi vill kompensera för, säg Z , givet den förklaringsvariabel vi vill studera effekten av, säg X . Olika viktsystem innebär alltså att den betingade empiriska fördelningen för Z givet X är olika för olika värden på X , eller med andra ord, förklaringsvariablerna X och Z samvarierar. Om å andra sidan X och Z inte samvarierar så är den betingade fördelningen för Z givet X densamma oavsett värdet på X , dvs vi får samma viktsystem och det föreligger ingen risk för sammanblandning av förklaringsvariablernas effekt på effektvariabeln. Lägg också märke till att de beräkningar vi har genomfört för att bestämma de betingade medelvärdena inte är någonting annat än en empirisk motsvarighet till satsen om total sannolikhet som vi känner igen från sannolikhetsteorin.

Om man vill jämföra betingade medelvärden för effektvariabeln Y givet värden på förklaringsvariabeln X utan att jämförelsen störs av förklaringsvariabeln Z , måste man använda samma viktsystem i alla medelvärdesberäkningar. Vi säger då att vi beräknar standardvägda medelvärden. Det viktsystem vi använder brukar kallas standardpopulation. För att beräkna standardvägda medelvärden måste vi alltså välja en lämplig standardpopulation

Exempel 12.1 (fortsättning)

Antag att vi väljer fördelningen för antal barn bland höginkomsthushållen som standardpopulation. Det standardvägda medelvärdet för låginkomsthushållen blir då

$$0,235 \cdot 1292 + 0,118 \cdot 3254 + 0,353 \cdot 4828 + 0,294 \cdot 6660 + 0,000 \cdot 7900 \\ \approx 4350$$

och för medelinkomsthushållen får vi

$$0,235 \cdot 4660 + 0,118 \cdot 4720 + 0,353 \cdot 6272 + 0,294 \cdot 7558 + 0,000 \cdot 9500 \\ \approx 6088$$

Detta innebär att låginkomsthushållens medelkonsumtion av livsmedel hade varit 4350 kr om de hade haft samma fördelning av barn som höginkomsthushållen och medelinkomsthushållens konsumtion varit 6088 kr om deras fördelning av antal barn hade varit samma som för höginkomsthushållen. Jämför man de standardvägda medelvärdena, kommer jämförelserna inte att störas av variabeln "antal barn", eftersom vi har använt samma viktsystem i alla beräkningarna. Resultaten sammanfattas i tabell 12.15.

Tabell 12.15 Betingade medelvärden och standardvägda medelvärden för konsumtion av livsmedel

	Inkomstnivå		
	låg	mellan	hög
Betingade medelvärden	5358	6232	8529
Standardvägda medelvärden	4350	6088	8529

Låt oss nu jämföra skillnaderna mellan betingade och standardvägda medelvärden för hög och låginkomsthushållen. Vi får då att

$$8529 - 5358 = (8529 - 4350) - 1008$$

$$3171 = 4179 - 1008$$

där komponenten 4179 kr beror på samvariationen mellan inkomstnivå och konsumtion. Höginkomsthushåll har i genomsnitt 4179 kr högre konsumtion än låginkomsthushåll med samma antal barn. Komponentens -1008

kr beror på att fördelningen av barn är olika i de olika inkomstnivåerna. Låginkomsthushåll har i genomsnitt en konsumtionsökning på 1008 kr jämfört med höginkomsthushåll på grund av att låginkomsthushållen i allmänhet har fler barn. ■

Olika standardpopulationer kan ge olika mått på skillnaden i medelvärden. Detta inträffar om det finns samspel mellan förklaringsvariablerna. Om inget samspel föreligger blir skillnaden mellan de standardvägda medelvärdena densamma oavsett vilket viktsystem vi använder. Om samspel mellan förklaringsvariablerna föreligger måste man därför vara mycket noga med valet av standardpopulation. Om olika standardpopulationer, som alla kan anses vara adekvata, ger helt olika resultat, är det olämpligt att överhuvudtaget använda standardvägning som metod att analysera samband.

Övning 12.5

Lönesättningen hos två företag skall jämföras. För detta ändamål har följande löner fördelade på utbildning observerats:

	Företag 1		Företag 2	
	Ant anställda	Medellön	Ant anställda	Medellön
Akademiker	144	32.300:-	26	31.300:-
Gymnasieutb	72	26.400:-	217	27.900:-
Lägre utb	144	22.600:-	52	22.100:-

- Beräkna medellönen totalt för de båda företagen.
- Gör en standardvägning av medellönerna som eliminerar skillnaderna i utbildningsstrukturen hos de båda företagen. Kommentera resultaten. Diskutera även värdet eller faran med metoden i detta fall.

Övning 12.6

För en studie av sjukfrånvaron i en viss industribransch och dess eventuella samband med kön drog man två stickprov, ett bland männen och ett bland kvinnorna anställda i branschen. En del av resultaten sammanfattas i nedanstående tabell.

Ålder	Män		Kvinnor	
	Antal i urvalet	Genomsnittlig sjukfrånvaro	Antal i urvalet	Genomsnittlig sjukfrånvaro
- 19	4	3	2	3
20 - 24	90	2	41	3
25 - 34	80	4	67	6
35 - 44	49	4	71	5
45 - 54	38	6	73	5
55 - 59	38	9	53	5
60 -	2	16	5	9
Summa	301		312	

- a) Åldersfördelningen verkar vara olika för män och kvinnor. Inled därför analysen med att illustrera åldersfördelningarna för män och kvinnor i histogram, beräkna medianåldrarna och kvartilavvikelsen för åldrarna. Kommentera eventuella skillnader mellan könen.
- b) Vid en tidigare undersökning har medelåldern för kvinnor anställda i branschen beräknats till 41,1 år. Kommentera vilka egenskaper hos datamaterialet som har lett till överensstämmelse eller avvikelse med medianåldern i detta datamaterial.
- c) Beräkna den genomsnittliga sjukfrånvaron för män och kvinnor i de båda stickproven.
- d) För att eliminera eventuella effekter av olika åldersfördelningar på genomsnittlig sjukfrånvaro skall du göra en standardvägning. Använd kvinnorna som standardpopulation. Kommentera resultatet.