

LMA201/LMA522: Faktorförsök

Föreläsning 1

Anders Hildeman

- Försöksplanering
- Faktorförsök med två nivåer
- Skattning av effekterna.
- Diagram för huvudeffekter
- Diagram för samspelseffekter
- Paretodiagram

- Den här veckan kommer tillägnas faktorförsök.
- Bok: “Försöksplanering - faktorförsök” av Ulla Blomqvist (den lilla svarta boken). Om ni köpt en äldre bok så är boken lila och Ulla heter Dahlbom i efternamn istället. Den äldre boken innehåller i alla viktiga aspekter samma material och fungerar bra den också.

Vi vill veta

Hur påverkar en eller flera faktorer, x (oberoende variabler), någon given storhet, y (beroende variabel).

För att kunna svara på detta så behöver man data på y vid olika värden på x . Man kan få tillgång till sådan data på två olika sätt:

- Planerat försök: Man sätter själv nivåerna på x och mäter y .
- Observationsstudie: Man observerar y och antecknar vilka nivåer x hade under motsvarande observation.

Data från ett planerat försök är oftast att föredra då:

- Man kan balansera antalet observationer av y för de olika värdena av x .
- Man kan lättare utesluta störfaktorer, faktorer som inte är av intresse men som påverkar y .

Nackdelen är dock att:

- Ett planerat försök kräver möjlighet att kontrollera nivåerna.
- Ett planerat försök kräver oftast större arbetsinsats/kostnad.

- Om x är det enda som påverkar y och både x och y är exakt mätbara så är det inte någon slump inblandad. Problemet är då enbart av matematisk natur och statistik behövs inte.
- I verkligheten är dock slump ofta inblandat, t.ex. mätningar är inexakta, det finns ytterligare faktorer som man inte känner till eller kan mäta. Med sådan osäkerhet så betar sig den uppmätta storheten till viss del slumpmässigt \Rightarrow behov av statistik.

Eftersom slumpen är inblandad så modelleras den beroende variabeln som en slumpvariabel (Y). Man kan då observera slumpvariabel betingat på olika värden hos de kända faktorerna, ($Y|\mathbf{x}$).

Om man bara har diskreta värden hos faktorerna (nivåer) så kan man dela upp observerad data i grupper. En grupp för varje unik uppsättning av nivåer hos faktorerna.

Exempel

- $Y \in (0, \infty)$: Tiden det tar för 11:ans spårvagn att åka sin rutt.
- x_1 : Tiden på dagen ($\{ 'morgon', 'dag', 'kväll' \}$)
- x_2 : Väderlek ($\{ 'sol', 'regn' \}$)

Antal grupper: $3 \cdot 2 = 6$.

I exemplet med spårvagnen så kan man fråga sig om sannolikhetsfördelningen för dessa 6 olika grupper är olika och i så fall på vilket sätt?

Ofta nöjer man sig med att ta reda på om väntevärdena är olika och vilken grupp som ger det bästa eller värsta värdet (den kortaste eller längsta tiden i exemplet med spårvagnen).

Exempel

Är $\mathbb{E}[Y|x_1 = \text{morgon}, x_2 = \text{sol}] < \mathbb{E}[Y|x_1 = \text{dag}, x_2 = \text{regn}]$?

Den här veckan kommer vi lära oss planera försök med två nivåer. D.v.s. hur vi skall utföra ett planerat försök då faktorerna bara kan anta två möjliga nivåer.

Då vi endast tänker oss två nivåer per faktor så kan man koda dessa som 'låg nivå' (-1) och 'hög nivå' ($+1$). Nivåerna behöver således inte vara kvantitativa värden.

Man tänker sig att väntevärdet, $\mu|\mathbf{x}$, för $Y|\mathbf{x}$ kan modelleras som en funktion av faktornivåerna med parametrar $\beta = [\beta_a, \beta_b, \beta_{ab}, \beta_c, \dots]$.

Exempel: Två faktorer

$$\mathbb{E}[Y|\mathbf{x}] =: \mu|\mathbf{x} = \beta_0 + \beta_a x_a + \beta_b x_b + \beta_{ab} x_a x_b$$

Termen β_0 motsvarar det medelvärdesbildade värdet av $\mu|\mathbf{x}$ över alla grupper. Om alla grupper är lika sannolika skulle detta motsvara det obetingade μ .

Termen $2\beta_a$ motsvarar hur väntevärdet förändras om man låter faktor A gå från 'låg' till 'hög' och medelvärdesbildar över motsvarande grupper för faktor B. Om alla grupper är lika sannolika skulle det motsvara hur det betingade väntevärdet enbart med avseende på x_a förändras mellan 'låg' och 'hög'.

På samma sätt så är β_b påverkan av faktor B om man medelvärdesbildar bort faktor A.

Termen β_{ab} är en samspelseffekt. β_{ab} förklarar hur de två faktorerna tillsammans förstärker eller motverkar den påverkan som redan givits av β_a - och β_b -effekterna.

Det kan kanske vara lättare att förstå vad β -termerna egentligen symboliserar ifall man beskriver dem som funktioner av de betingade väntevärdena istället för tvärtom. Låt

$$\mu_{(k,l)} = \mu|(x_a = k, x_b = l).$$

$$\beta_0 = \frac{\mu_{(-1,-1)} + \mu_{(+1,-1)} + \mu_{(-1,+1)} + \mu_{(+1,+1)}}{4}$$

$$\beta_a = \frac{\mu_{(+1,-1)} + \mu_{(+1,+1)}}{4} - \frac{\mu_{(-1,-1)} + \mu_{(-1,+1)}}{4}$$

$$\beta_b = \frac{\mu_{(-1,+1)} + \mu_{(+1,+1)}}{4} - \frac{\mu_{(-1,-1)} + \mu_{(+1,-1)}}{4}$$

$$\beta_{ab} = \frac{\mu_{(+1,+1)} + \mu_{(-1,-1)}}{4} - \frac{\mu_{(-1,+1)} + \mu_{(+1,-1)}}{4}$$

Målet med ett planerat försök är ofta att upptäcka bästa eller värsta kombinationen av faktorer för ett önskat resultat. Det kan t.ex. vara hur man skall ställa in produktionsprocessen i en fabrik, hur man skall äta och träna för att göra ett bra resultat i en tävling eller hur man får sina medarbetare nöjda.

Om $\beta_a > 0$ så innebär det t.ex. att om man låter faktor A vara hög så kommer väntevärdet av Y vara större än om faktor A är låg (om man medelvärdesbildar över de andra faktorerna). På samma sätt för faktor B. Det kan dock vara så att samspelseffekten är såpass starkt negativ att t.ex. $\beta_a + \beta_b + \beta_{ab} < \beta_a - \beta_b - \beta_{ab}$ även om $\beta_a > 0$ och $\beta_b > 0$.

Mer än två faktorer?

Fungerar på exakt samma sätt. Man behöver nu en parameter för varje faktor (K st om vi antar K faktorer) samt en parameter för varje samspel. Samspel kan vara allt från två-faktorsspel till K -faktorsspel. Alltså finns det $\sum_{k=2}^K \binom{K}{k}$ möjliga samspel.

För tre faktorer:

$$\begin{aligned}\mu|\mathbf{x} = & \beta_0 + \beta_a x_a + \beta_b x_b + \beta_c x_c + \dots \\ & + \beta_{ab} x_a x_b + \beta_{ac} x_a x_c + \beta_{bc} x_b x_c + \dots \\ & + \beta_{abc} x_a x_b x_c\end{aligned}$$

Skattade effekter

- Eftersom vi har slump i vår data så kan vi aldrig veta våra väntevärden exakt. Därför kan vi inte heller ta reda på våra β -parametrar exakt.
- Precis som vi tidigare gjort med väntevärden så kan vi göra punktskattningar på våra effekter. Givet vår data så tar vi alltså fram vår bästa gissning av effekterna.
- Låt oss kalla den skattade effekten av $2\beta_a$ för I_a , alltså $\mathbb{E}[I_a] = 2\beta_a$, på samma sätt $I_b \Leftrightarrow 2\beta_b$ osv.
- Det skattade “obetingade väntevärdet” kallar vi för M och är en skattning av β_0 (märkväl inte en skattning av $2\beta_0$).

Givet att vi gör minst en mätning för varje grupp kan vi då göra punktskattningar av β -parametrarna. T.ex.

$$\beta_a = \frac{\mathbb{E}[Y|(+1, +1)] + \mathbb{E}[Y|(+1, -1)]}{4} - \frac{\mathbb{E}[Y|(-1, +1)] + \mathbb{E}[Y|(-1, -1)]}{4}$$

så l_a blir

$$l_a = l_{a^+} - l_{a^-} = \frac{\bar{y}_{(+1,+1)} + \bar{y}_{(+1,-1)}}{2} - \frac{\bar{y}_{(-1,+1)} + \bar{y}_{(-1,-1)}}{2},$$

där $\bar{y}_{(+1,+1)} = \frac{1}{n_{(+1,+1)}} \sum_{i=1}^{n_{(+1,+1)}} y_{i,(+1,+1)}$, $n_{(+1,+1)}$ är antalet mätningar i grupp $(+1, +1)$ och $y_{i,(+1,+1)}$ är den i :te mätningen i grupp $(+1, +1)$.

På samma sätt skattar vi l_b , l_c , osv.

Vi kan förenkla livet med en tabell.

Försök nr	M	A	B	AB	Resultat
1	+	-	-	+	$\bar{y}_1 = 4$
2	+	+	-	-	$\bar{y}_2 = 2$
3	+	-	+	-	$\bar{y}_3 = 3$
4	+	+	+	+	$\bar{y}_4 = 4$

Se på tabellen vilka y-värden som skall vara positiva och negativa för att räkna ut respektive effekt.

Vi kan förenkla livet med en tabell.

Försök nr	M	A	B	AB	Resultat
1	+	-	-	+	$\bar{y}_1 = 4$
2	+	+	-	-	$\bar{y}_2 = 2$
3	+	-	+	-	$\bar{y}_3 = 3$
4	+	+	+	+	$\bar{y}_4 = 4$

Se på tabellen vilka y-värden som skall vara positiva och negativa för att räkna ut respektive effekt.

$$M = \frac{4 + 2 + 3 + 4}{4} = 3.25$$

$$I_a = \frac{-4 + 2 - 3 + 4}{2} = -0.5$$

$$I_b = \frac{-4 - 2 + 3 + 4}{2} = +0.5$$

$$I_{ab} = \frac{4 - 2 - 3 + 4}{2} = +1.5$$

- Som ni ser så kommer varje skattning av en effekt att beräknas med hjälp av alla mätvärdena \bar{y}_i .
- Varför inte bara slippa allt krångel och titta på för vilket i vi fick bästa värdet på \bar{y}_i ?

Det finns två anledningar att "krångla" till det.

- 1 β -parametrarna säger mer än bara vilken grupp som ger bäst resultat. De säger på vilket sätt varje faktor påverkar resultatet och hur de samspelar. Det kan vara viktigt att veta vilka faktorer som förbättrar, försämrar, inte påverkar, samt hur de interagerar med varandra.
- 2 För tre faktorer eller fler så blir skattningarna av l -parametrarna stabilare (mindre variation) jämfört med \bar{y}_i .

- Som ni ser så kommer varje skattning av en effekt att beräknas med hjälp av alla mätvärdena \bar{y}_i .
- Varför inte bara slippa allt krångel och titta på för vilket i vi fick bästa värdet på \bar{y}_i ?
Det finns två anledningar att "krångla" till det.
 - 1 β -parametrarna säger mer än bara vilken grupp som ger bäst resultat. De säger på vilket sätt varje faktor påverkar resultatet och hur de samspelar. Det kan vara viktigt att veta vilka faktorer som förbättrar, försämrar, inte påverkar, samt hur de interagerar med varandra.
 - 2 För tre faktorer eller fler så blir skattningarna av l -parametrarna stabilare (mindre variation) jämfört med \bar{y}_i .

Om man antar samma varians, σ^2 , på mätningar oavsett grupp så blir variansen av de skattade effekterna:

$$\text{Var}[l_a] = \frac{4\sigma^2}{N},$$

där N är antal grupper. Alltså, $N = 2^K$ där K är antalet faktorer.

Försök nr	A	B	AB	Resultat
1	-	-	+	$\bar{y}_1 = 4$
2	+	-	-	$\bar{y}_2 = 2$
3	-	+	-	$\bar{y}_3 = 3$
4	+	+	+	$\bar{y}_4 = 4$

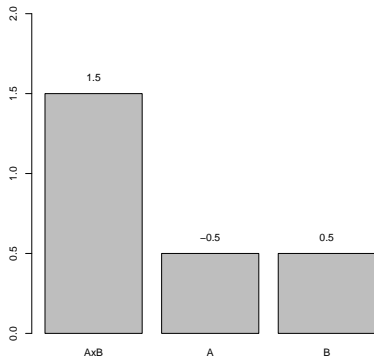
\bar{y}_1 är här stickprovsmedelvärdet av alla mätningar gjorda med uppsättningen av faktornivåer (-1, -1).

Om det är mer än en mätning för varje uppsättning faktornivåer, kom ihåg att välja ut när mätningarna sker i experimentet slumpmässigt. Alltså, gör inte alla mätningar för (-1,-1) efter varandra utan slumpa ut ordningen.

Detta minskar risken att något systematiskt fel av misstag inkluderas i analysen. (tänk t.ex. att utövaren blir bättre på mätningar med tiden eller att temperaturen kanske stiger i rummet med tiden)

Paretdiagram

För att tydligt kunna se storleken på de uppskattade effekterna så kan man rita upp ett Paretdiagram.

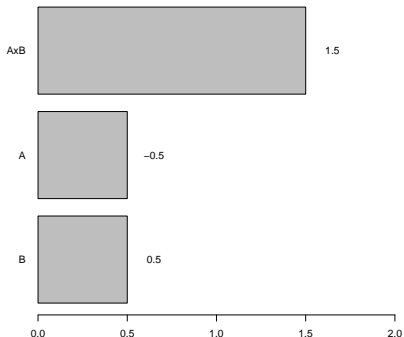


- Rangordna effekterna i storleksordning utan hänsyn till tecken.
- Rita in staplarna (störst till vänster och sedan i avtagande ordning).

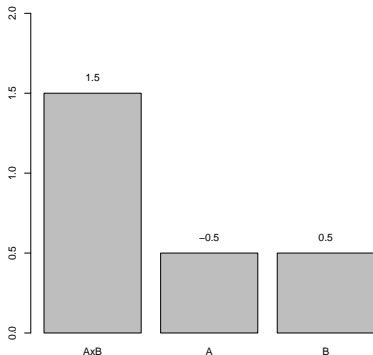
Figur: Paretdiagram över de uppskattade effekterna från tidigare exempel.

Paretdiagram

I boken ritas de dem liggande.

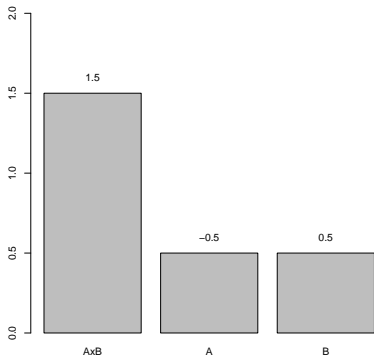


Figur: Paretdiagram över de uppskattade effekterna från tidigare exempel.



Exempel

- Hur bör vi välja våra faktorer för att få så högt värde som möjligt?
- Hur bör vi välja våra faktorer för att få så lågt värde som möjligt?



Exempel

- Hur bör vi välja våra faktorer för att få så högt värde som möjligt?
- Hur bör vi välja våra faktorer för att få så lågt värde som möjligt?

$$\hat{\mu}_{(+1,+1)} = 3.25 - 0.5 + 0.5 + 1.5 = 4.75$$

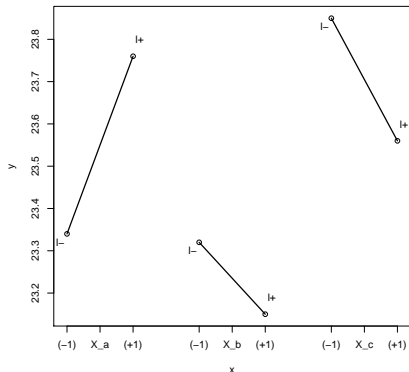
$$\hat{\mu}_{(-1,+1)} = 3.25 + 0.5 + 0.5 - 1.5 = 2.75$$

$$\hat{\mu}_{(+1,-1)} = 3.25 - 0.5 - 0.5 - 1.5 = 0.75$$

$$\hat{\mu}_{(-1,-1)} = 3.25 + 0.5 - 0.5 + 1.5 = 4.75$$

Diagram för huvudeffekter

Man kan rita ut huvudeffekterna i ett diagram för att lättare få känsla för vad effekterna är när faktorn är låg resp. hög. För faktor A ritas man då ut I_{a-} och I_{a+} -värdena på y-axeln och motsvarande -1 och $+1$ på x-axeln. Man fortsätter därefter med faktor B, C, etc på samma sätt.

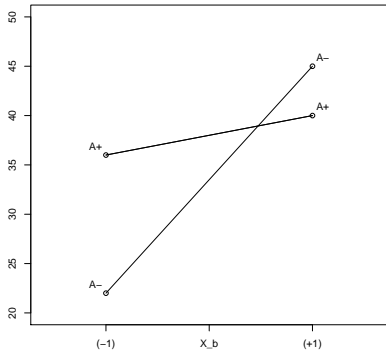


I den här figuren ser man att faktor A:s huvudeffekt har störst påverkan eftersom det är störst skillnad mellan I_+ och I_- för faktor A. Dessutom ser vi att det är en

positiv påverkan då $I_+ > I_-$. (B och C:s huvudeffekter påverkar negativt.)

Diagram för samspelseffekter

Man kan även rita ut diagram för samspelseffekterna. Är man t.ex. intresserad av samspelseffekten β_{ab} så kan man rita ett streck mellan medelvärdet inom grupperna där faktor B är låg respektive hög då faktor A är låg respektive hög.



I den här figuren ser man att det existerar ett samspel eftersom linjerna inte är parallella mot varandra.

$$\mu|x = \beta_0 + \beta_a x_a + \beta_b x_b + \beta_{ab} x_a x_b$$

Samspelet är negativt eftersom $A_+ B_+ - A_- B_+ < A_+ B_- - A_- B_-$.

- Vi vill utföra planerade faktorförsök för att ta reda på hur en viss storhet påverkas av förändringar i bakomliggande faktorer.
- Vi bryr oss nu bara om hur väntevärdet förändras då faktorerna skiftas mellan “höga” och “låga” nivåer.
- Om vi har K olika faktorer så kommer vi behöva göra minst 2^K olika mätningar för att uppskatta alla effekter i vår modell.
- Vid mer än en mätning för varje nivåuppsättning, välj ut ordningen på mätföljden slumpmässigt. Då undviker man att få med eventuella systematiska fel.
- Visualisera de skattade effekterna med diagram.

Exempel

Försök nr	A	B	C	AB	BC	AC	ABC	Uppmätt
1	-	-	-	+	+	+	-	$y_1 = [3.7, 2.8]$
2	+	-	-	-	+	-	+	$y_2 = [4.8, 4.8]$
3	-	+	-	-	-	+	+	$y_3 = [18.7, 17.1]$
4	+	+	-	+	-	-	-	$y_4 = [13.5, 14.1]$
5	-	-	+	+	-	-	+	$y_5 = [10.1, 11.7]$
6	+	-	+	-	-	+	-	$y_6 = [8.8, 9.3]$
7	-	+	+	-	+	-	-	$y_7 = [17.7, 16.9]$
8	+	+	+	+	+	+	+	$y_8 = [0.4, -0.2]$

Försök nr	A	B	C	AB	BC	AC	ABC	Resultat
1	-	-	-	+	+	+	-	$\bar{y}_1 = 3.25$
2	+	-	-	-	+	-	+	$\bar{y}_2 = 4.80$
3	-	+	-	-	-	+	+	$\bar{y}_3 = 17.9$
4	+	+	-	+	-	-	-	$\bar{y}_4 = 13.8$
5	-	-	+	+	-	-	+	$\bar{y}_5 = 10.9$
6	+	-	+	-	-	+	-	$\bar{y}_6 = 9.05$
7	-	+	+	-	+	-	-	$\bar{y}_7 = 17.30$
8	+	+	+	+	+	+	+	$\bar{y}_8 = 0.10$

$$M = \frac{3.25 + 4.80 + 17.9 + 13.8 + 10.9 + 9.05 + 17.30 + 0.10}{8} = 9.64$$

$$I_a = \frac{4.80 + 13.8 + 9.05 + 0.10}{4} - \frac{3.25 + 17.9 + 10.9 + 17.30}{4} = -5.4$$

Försök nr	A	B	C	AB	BC	AC	ABC	Resultat
1	-	-	-	+	+	+	-	$\bar{y}_1 = 3.25$
2	+	-	-	-	+	-	+	$\bar{y}_2 = 4.80$
3	-	+	-	-	-	+	+	$\bar{y}_3 = 17.9$
4	+	+	-	+	-	-	-	$\bar{y}_4 = 13.8$
5	-	-	+	+	-	-	+	$\bar{y}_5 = 10.9$
6	+	-	+	-	-	+	-	$\bar{y}_6 = 9.05$
7	-	+	+	-	+	-	-	$\bar{y}_7 = 17.30$
8	+	+	+	+	+	+	+	$\bar{y}_8 = 0.10$

$$I_b = \frac{17.9 + 13.8 + 17.3 + 0.1}{4} - \frac{3.25 + 4.8 + 10.9 + 9.05}{4} = 5.28$$

$$I_c = \frac{10.9 + 9.05 + 17.3 + 0.1}{4} - \frac{3.25 + 4.8 + 17.9 + 13.8}{4} = -0.60$$

Försök nr	A	B	C	AB	BC	AC	ABC	Resultat
1	-	-	-	+	+	+	-	$\bar{y}_1 = 3.25$
2	+	-	-	-	+	-	+	$\bar{y}_2 = 4.80$
3	-	+	-	-	-	+	+	$\bar{y}_3 = 17.9$
4	+	+	-	+	-	-	-	$\bar{y}_4 = 13.8$
5	-	-	+	+	-	-	+	$\bar{y}_5 = 10.9$
6	+	-	+	-	-	+	-	$\bar{y}_6 = 9.05$
7	-	+	+	-	+	-	-	$\bar{y}_7 = 17.30$
8	+	+	+	+	+	+	+	$\bar{y}_8 = 0.10$

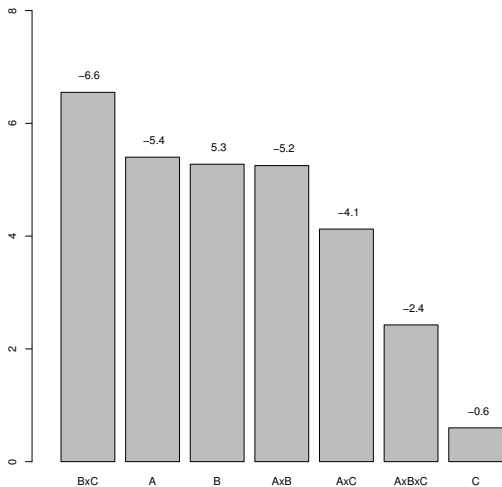
$$I_{ab} = \frac{3.25 + 13.8 + 10.9 + 0.1}{4} - \frac{4.8 + 17.9 + 9.05 + 17.3}{4} = -5.25$$

$$I_{bc} = \frac{3.25 + 4.8 + 17.3 + 0.1}{4} - \frac{17.9 + 13.8 + 10.9 + 9.05}{4} = -6.55$$

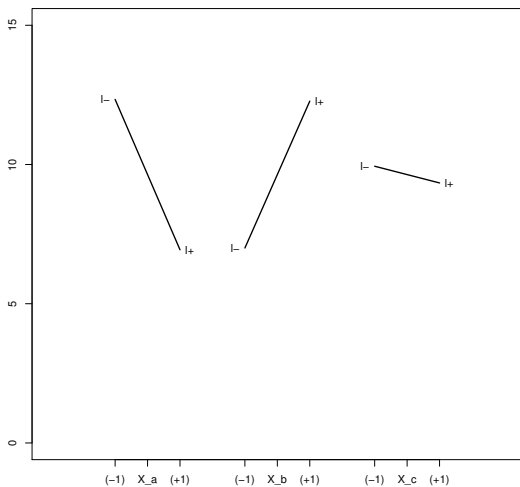
Försök nr	A	B	C	AB	BC	AC	ABC	Resultat
1	-	-	-	+	+	+	-	$\bar{y}_1 = 3.25$
2	+	-	-	-	+	-	+	$\bar{y}_2 = 4.80$
3	-	+	-	-	-	+	+	$\bar{y}_3 = 17.9$
4	+	+	-	+	-	-	-	$\bar{y}_4 = 13.8$
5	-	-	+	+	-	-	+	$\bar{y}_5 = 10.9$
6	+	-	+	-	-	+	-	$\bar{y}_6 = 9.05$
7	-	+	+	-	+	-	-	$\bar{y}_7 = 17.30$
8	+	+	+	+	+	+	+	$\bar{y}_8 = 0.10$

$$I_{ac} = \frac{3.25 + 17.9 + 9.05 + 0.1}{4} - \frac{4.8 + 13.8 + 10.9 + 17.3}{4} = -4.13$$

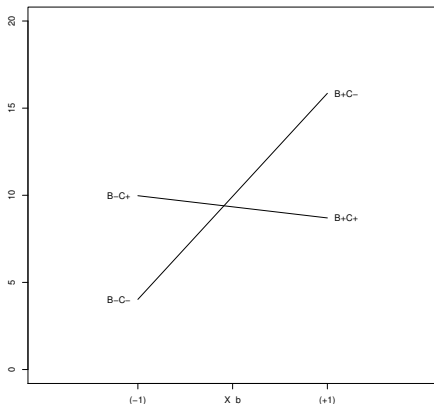
$$I_{abc} = \frac{4.8 + 17.9 + 10.9 + 0.1}{4} - \frac{3.25 + 13.8 + 9.05 + 17.3}{4} = -2.43$$



Figur: Paretodigram.



Figur: Diagram för huvudeffekter.



(a) Samspelsgraf för BxC.

- B_+C_- : medelvärdet av alla \bar{y}_i där faktor B är hög och faktor C är låg.
- Samspelseffekten är tydlig då linjerna inte är parallella. Att vinkeln är stor mellan linjerna visar att samspelseffekten är stor jämfört med de båda huvudeffekterna.
- Negativt samspel då $B_+C_+ - B_+C_- < B_-C_+ - B_-C_-$.