

Statistik och dataanalys I

F6: Samband mellan numeriska variabler och korrelation

Valentin Zulj

Statistiska institutionen

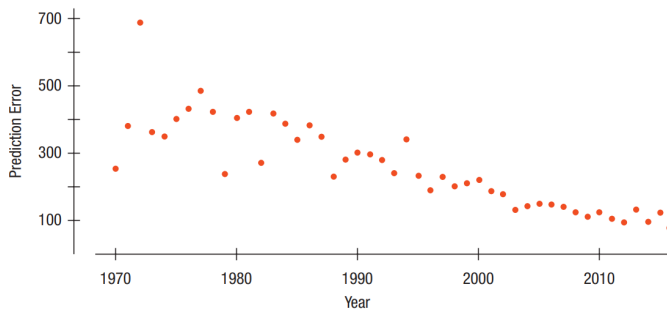


Stockholm
University

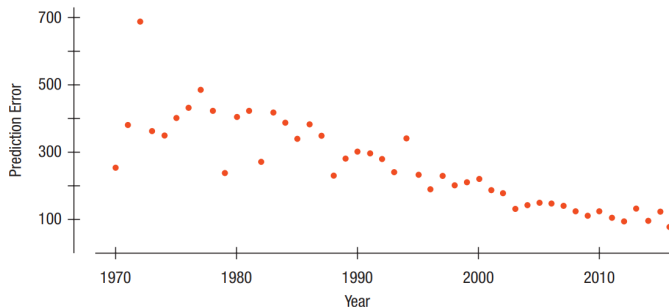
- Som statistiker är vi ofta intresserade av att studera samband mellan variabler
- Hittills har vi tittat på
 - Samband mellan två eller flera kategoriska variabler, med hjälp av korstabeller och stapeldiagram
 - Samband mellan en numerisk och en kategorisk variabel, med hjälp av låddiagram och histogram
- Ett naturligt steg framåt är att nu börja titta på **samband mellan två numeriska variabler**

- Vi ska som sagt prata om att undersöka samband mellan numeriska variabler
- Denna typ av sambandsstudier är väldigt vanliga, och förekommer ofta i studentuppsatser, publicerade forskningsartiklar, och övriga tillämpningar
- Mer specifikt så kommer vi att titta på
 - Spridningsdiagram
 - Linjära samband och korrelation
 - Korrelation och kausalitet
- Detta leder oss fram till ämnet för nästa föreläsning, nämligen **linjär regression**

Spridningsdiagram



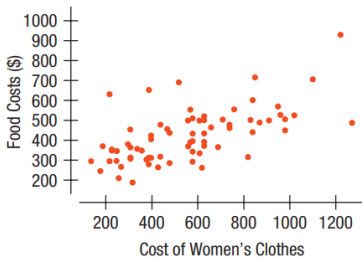
- Vi har redan tittat på diagram över tidsserier, där vi studerar hur en numerisk variabel utvecklas över en given tidsperiod
- Detta är exempel exempel på samband mellan två numeriska variabler, då även tiden är en numerisk variabel
- Figuren visar en tidsserie av årsvisa genomsnittliga prediktionsfel i nautiska mil i lokalisering av orkaner i Atlanten



- Samband mellan två numeriska variabler kan illustreras med hjälp av ett **spridningsdiagram** (*scatter plot*)
- Diagrammet ovan illustrerar en tidsserie, och är alltså ett exempel på ett spridningsdiagram

- Vi har ett datamaterial med information om olika prisnivåer i 73 städer
- I nästa slide finns fyra olika spridningsdiagram, där varje röd punkt representerar en stad
- När vi läser av diagrammen kan vi fundera på
 - Sambandets riktning – Är det positivt eller negativt?
 - Sambandets styrka – Är det starkt (tydligt) eller svagt (otydligt)?
 - Huruvida sambandet är linjärt – Följer mönstret en rät linje?
- Vi börjar med att titta närmare på Diagram A

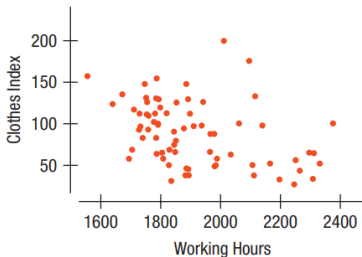
Att läsa av ett spridningsdiagram – Exempel



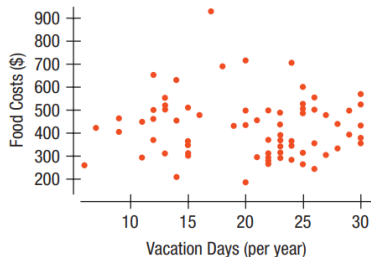
A



B



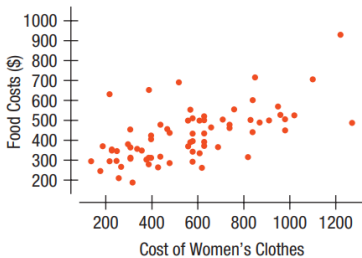
C



D

- I **Diagram A** har vi matpriser på y -axeln, och klädpriser på x -axeln
- Är sambandet positivt eller negativt?
- Vi ser att punktsvärmen “lutar uppåt” – generellt är matpriser höga när klädpriser är höga, och vice versa
- Vi säger att sambandet är **positivt**
- Är sambandet linjärt eller icke-linjärt?
- Punktsvärmen ser ut att vara ganska samlad runt en rät linje
- Vi säger att sambandet är på ett ungefär **linjärt**
- Vi fortsätter med att titta på Diagram B

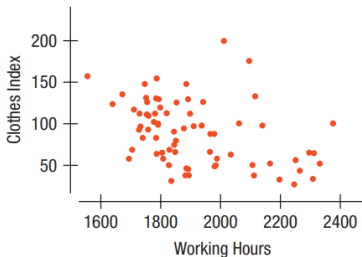
Att läsa av ett spridningsdiagram – Exempel



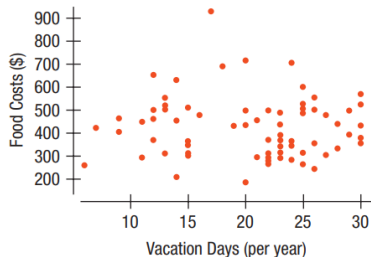
A



B



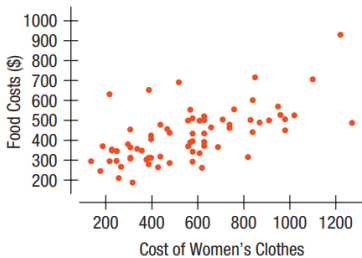
C



D

- I **Diagram B** har vi antalet arbetstimmar som krävs för att tjäna ihop till en iPhone på y -axeln, och medellön per timme på x -axeln
- Är sambandet positivt eller negativt?
- Vi ser att punktsvärmen “lutar neråt” – generellt paras små x -värden med stora y -värden, och stora x -värden med små y -värden
- Vi säger att sambandet är **negativt**
- Är sambandet linjärt eller icke-linjärt?
- Punktsvärmen ser **inte** ut att vara samlad runt en rät linje
- Vi säger att sambandet är **icke-linjärt**
- Vi fortsätter med att titta på Diagram C

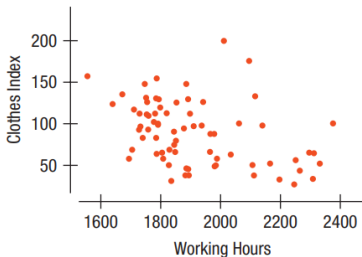
Att läsa av ett spridningsdiagram – Exempel



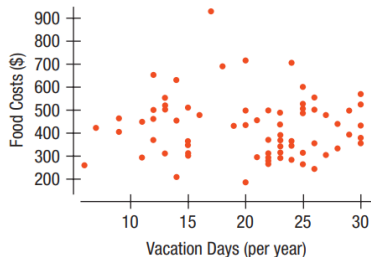
A



B



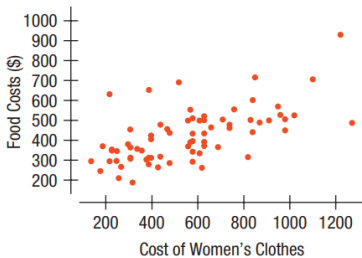
C



D

- I **Diagram C** har vi klädprisindex på y -axeln, och arbetstid per år på x -axeln
- Vi ser – enligt samma principer som tidigare – att sambandet är negativt
- Vi ser dock också att sambandet är ganska **svagt** – “nedåtlutningen” är inte lika tydlig som t.ex. “uppåtlutningen” i Diagram A
- Vi ser även att sambandet är någorlunda linjärt
- Vi fortsätter med att titta på Diagram D

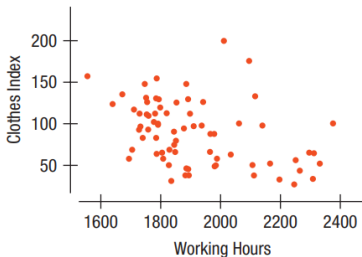
Att läsa av ett spridningsdiagram – Exempel



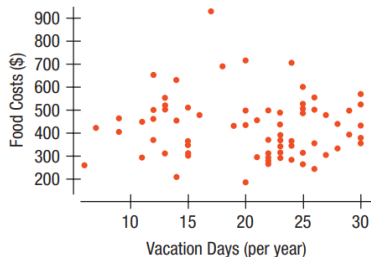
A



B



C

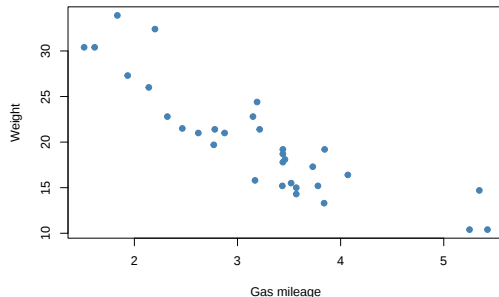
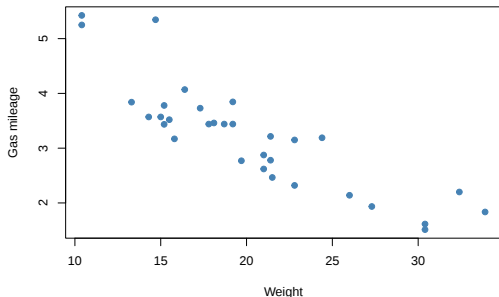


D

- I **Diagram D** har vi matpriser på y -axeln, och semesterdagar per år på x -axeln
- Kommentarer?
 - Är sambandet positivt eller negativt?
 - Är sambandet linjärt eller icke-linjärt?
- Här finns inte något samband alls, vare sig positivt eller negativt
- **Viktig slutsats:** Det finns inga negativa bieffekter av längre semester!

- Hur bestämmer vi vilken variabel som är y och vilken som är x ?
- Ofta finns det inget korrekt svar, men vi kan använda tumregler för vad som är rimligast
- Enklast att tänka i termer av prediktion
 - Den variabel som vi har intresse av att prediktera väljs som y
 - Den variabel som hjälper oss att prediktera/förklara väljs som x
- Är det rimligt att fråga sig om variabeln på y -axeln påverkas av variabeln på x -axeln?
Blir frågan mer rimlig om variablerna byter plats?

- Bilden visar räckvidden per gallon bensin mot vikten hos en bil
- Vilken av följande frågor låter rimligast?
 - Brukar bensinförbrukningen vara högre för bilar som väger mer?
 - Brukar bilar vara tyngre om de förbrukar mer bensin?
- Om den första frågan låter mer rimlig väljer vi grafen till vänster

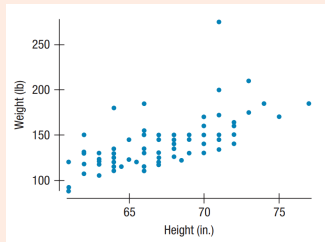
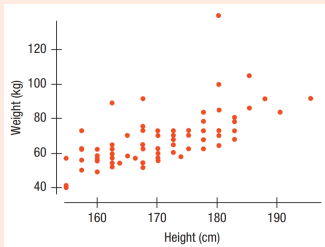


- Det finns flera olika namn som används för x - respektive y -variabeln
- Variabeln y kallas i vissa sammanhang **responsvariabel** (*response variable*)
- Variabeln x kallas ibland för **förklaringsvariabel** (*explanatory variable*)
- Ett annat vanligt namn för y är **beroende variabel** (*dependent variable*), och x blir då en **oberoende variabel** (*independent variable*)
- Andra vanligt förekommande namn för x -variabeln är **prediktor** (*predictor*), **kovariat** (*covariate*), och **förklarande variabel** (*explanatory variable*)
- Inom maskininlärning kallas förklarande variabler ofta för **features**

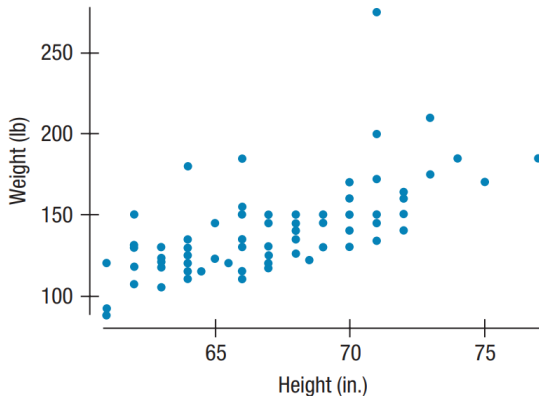
Linjära samband och korrelation

- Vi har pratat vagt om styrkan i ett samband, men kan vi på något sätt mäta hur **starkt** ett **linjärt samband** mellan två numeriska variabler är?
- Vi lägger till att vi gärna vill ha ett mått som är oberoende av enheter

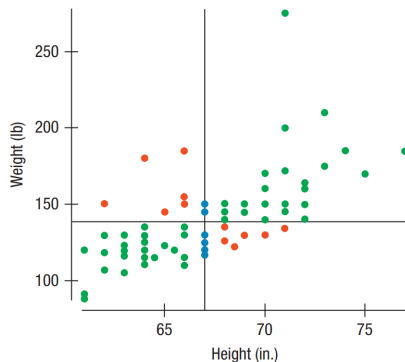
- Figur 6.2 och 6.3 i De Veaux et al. (2021) visar ett antal personers vikt och längd, och vi ser liknande samband oavsett om vikt ges i kg eller lbs



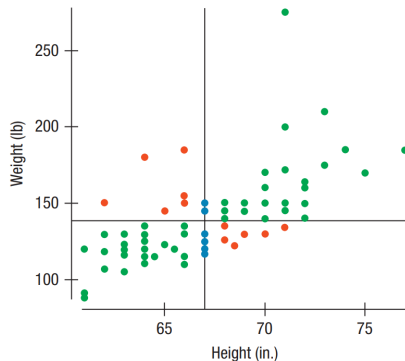
- Den enda skillnaden är olika skalor på y -axelarna i de två graferna – och vi vill därför ha ett mått som ger samma resultat i båda fallen (och alltså inte påverkas av skala)



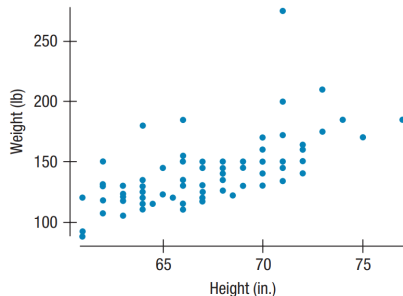
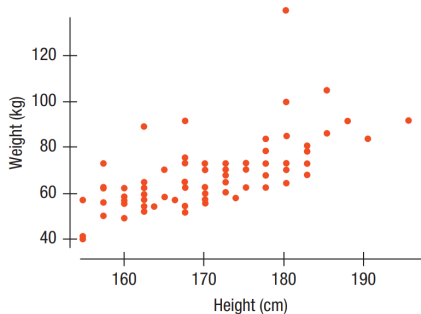
- Vi ser på bilden att sambandet är positivt – de datapunkter som har ett stort värde på x -axeln tenderar att också ha ett stort värde på y -axeln
- För att göra sambandet ännu tydligare kan vi rita ut variablernas medelvärden i grafen
- Vi markerar $\bar{x}$ med ett vertikalt streck och $\bar{y}$ med ett horisontellt streck (se nästa sida)



- Linjerna som representerar variablernas medelvärden delar in grafen i fyra delar, och färgen på punkterna skiljer sig åt mellan delarna
- Punkterna är **gröna** om *både x och y är större* än respektive medelvärde, eller om *både x och y är mindre* än respektive medelvärde
- Punkterna är **röda** om de är mindre än medelvärdet på den ena axeln och större än medelvärdet på den andra axeln

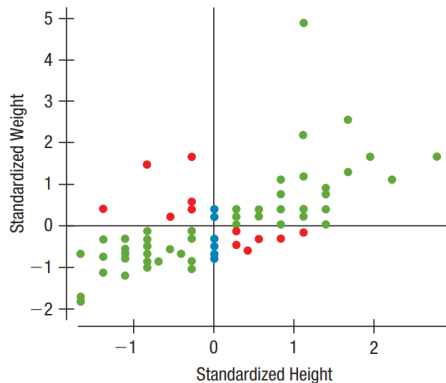
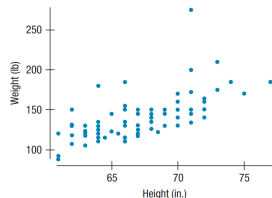
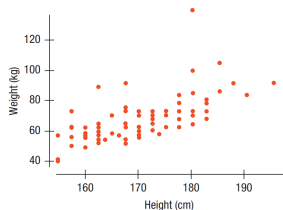


- De **gröna** punkterna påverkar sambandet i *positiv* riktning
- De **röda** punkterna påverkar sambandet i *negativ* riktning
- De gröna punkterna är fler, vilket indikerar att sambandet är positivt
- Fler röda punkter hade indikerat ett *negativt* samband



- Att räkna punkter i olika delar av bilden är ett enkelt och intuitivt sätt att bedöma riktningen på sambandet
- **Men:** Det är inte ovanligt att ha tusentals datapunkter (jobbigt att räkna)
- **Och:** Vi vill ofta kunna säga något om *hur starkt* det linjära sambandet är
- Vi vill så ha ett mer konkret mått, som inte beror på vilka enheter vi använder

- Vi har tidigare räknat med **z -värden** för numeriska variabler – dessa har fördelen att de är lika stora oavsett enhet som en variabel mäts i
- Vi kallar z för en *standardiserad* variabel, då den har en “standardenheter”



- Vi förtydligar detta med ett exempel i R, med variablerna `weight_pounds` och `weight_kg` (vikt på olika bilmodeller)
- `weight_pounds` och `weight_kg` anger vikt på **samma bilar i olika enheter**, och vi ska visa att de får samma värden efter standardisering

```
data(mtcars)

weight_pounds <- mtcars$wt # vikt i pounds
head(weight_pounds)
## [1] 2.620 2.875 2.320 3.215 3.440 3.460

weight_kg <- weight_pounds * 0.454 # vikt i kg
head(weight_kg)
## [1] 1.18948 1.30525 1.05328 1.45961 1.56176 1.57084
```

- Vi standardiserar var och en av variablerna, och får två uppsättningar z -värden enligt

$$z_i = \frac{x_i - \bar{x}}{s}, \quad \bar{x} = \frac{\sum_i x_i}{n}, \quad s = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n - 1}}.$$

- Genom standardisering byter vi enhet från *kg* eller *pounds* till *standardavvikelser*
- Eftersom variablerna mäter samma sak förväntar vi oss att de ska få samma standardiserade värden

```
## Beräkningar i pounds
m_pounds <- mean(weight_pounds) # medelvärde
s_pounds <- sd(weight_pounds)    # standardavvikelse

z_pounds <- (weight_pounds - m_pounds)/s_pounds # z

## Beräkningar i kg
m_kg <- mean(weight_kg)
s_kg <- sd(weight_kg)

z_kg <- (weight_kg - m_kg)/s_kg

## z för vikt i pounds
head(z_pounds) |> round(3)
## [1] -0.610 -0.350 -0.917 -0.002  0.228  0.248

## z för vikt i kg
head(z_kg) |> round(3)
## [1] -0.610 -0.350 -0.917 -0.002  0.228  0.248
```

- För att mäta styrkan på det linjära sambandet använder vi **korrelationskoefficienten** (*correlation coefficient*), som beräknas enligt

$$r = \frac{\sum z_x z_y}{n - 1}$$

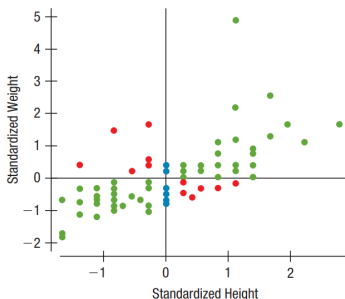
- Eftersom beräkningen använder de standardiserade variablerna z_x och z_y blir korrelationskoefficienten oberoende av enheterna för x och y
- Korrelationen är alltid ett tal mellan -1 och 1, dvs $-1 < r < 1$
- Om r ligger nära 1 eller -1 visar det att korrelationen är stark, men om r ligger nära 0 visar det att korrelationen är svag
- Vi kan ibland använda beteckningen r_{xy} för att betona att det är korrelationen mellan variablerna x och y

- Om korrelationskoefficienten är positiv har vi ett positivt linjärt samband mellan variablerna
- Om korrelationskoefficienten är negativ har vi ett negativt linjärt samband
- Korrelationsmåttet är *symmetriskt*, vilket betyder att korrelationen mellan x och y är densamma som korrelationen mellan y och x

- Vi kommer ihåg att korrelationskoefficienten beräknas som

$$r = \frac{\sum z_x z_y}{n - 1}$$

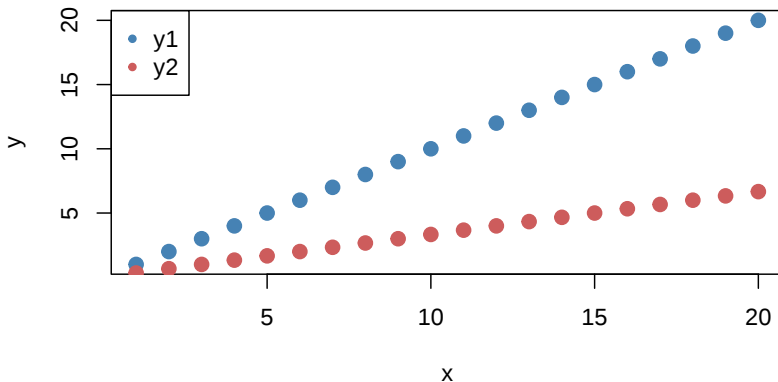
- Vi ser i formeln att korrelationskoefficienten blir positiv om uttrycket $\sum z_x z_y$ är positivt, och att korrelationen är starkare när uttrycket är stort
- När uttrycket $\sum z_x z_y$ är negativt blir korrelationen negativ, och om uttrycket är ett stort negativt tal är den negativa korrelationen starkare
- Vi tittar närmare på intuitionen bakom detta i nästa bild!



- Vi kopplar formeln för r till bilden från tidigare – kom ihåg att $z_y = (y - \bar{y})/s_y$
- Notera att produkten $z_x z_y$ är **positiv** när
 - både z_x och z_y är positiva
 - både z_x och z_y är negativa
- På motsvarande sätt är produkten $z_x z_y$ är **negativ** när
 - z_x är positiv och z_y är negativ, eller vice versa

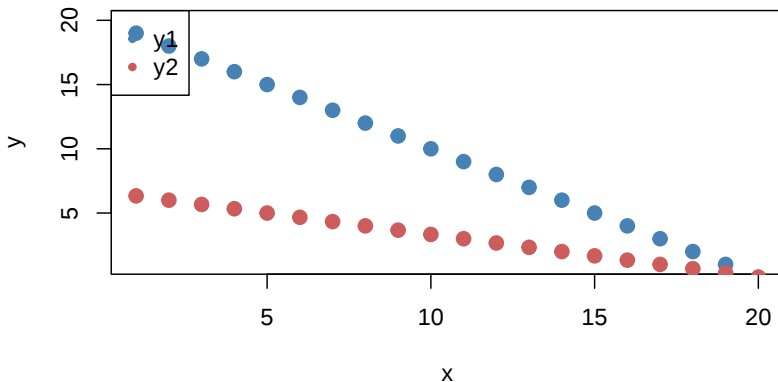
- Figuren nedan visar två exempel på hur ett **perfekt positivt** linjärt samband kan se ut
- Både y_1 och y_2 är perfekt korrelerade med x , dvs $r = 1$ för båda sambanden, trots att linjerna har olika lutning

y1 vs x and y2 vs x, $r = 1$

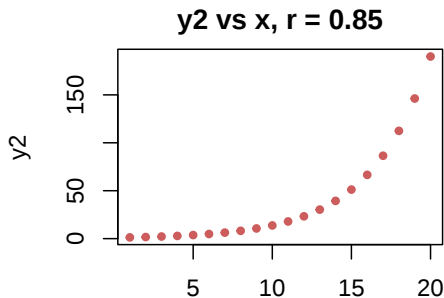
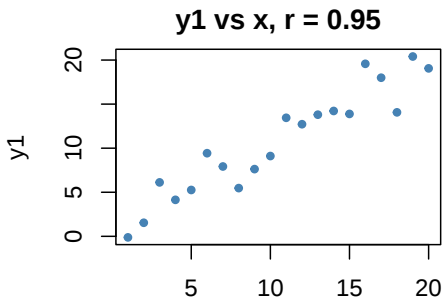


- Figuren nedan visar två exempel på hur ett **perfekt negativt** linjärt samband kan se ut
- Både y_1 och y_2 är perfekt korrelerade med x , dvs $r = -1$ för båda sambanden, trots att linjerna har olika lutning

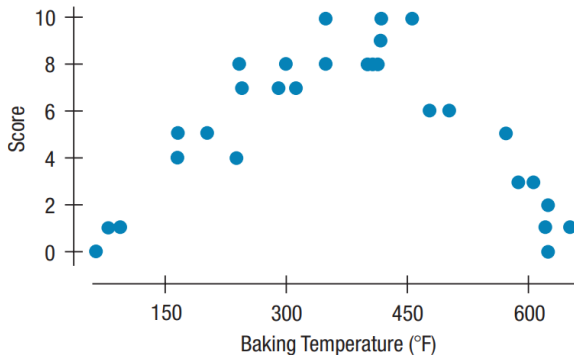
y1 vs x and y2 vs x, $r = -1$



- Figurerna visar exempel på två olika sorters samband
- De blå punkterna visar ett en **starkare** korrelation än de röda punkterna, trots att de blå punkterna är utspridda medan de röda punkter följer en tydlig linje
- Även om de blå punkterna är mer spridda följer de en **rak** linje, medan de röda punkterna följer en böjd linje (kurva)
- Kom ihåg att korrelation **bara** mäter styrkan i det **linjära** sambandet – linjärt syftar här på en **rät** linje

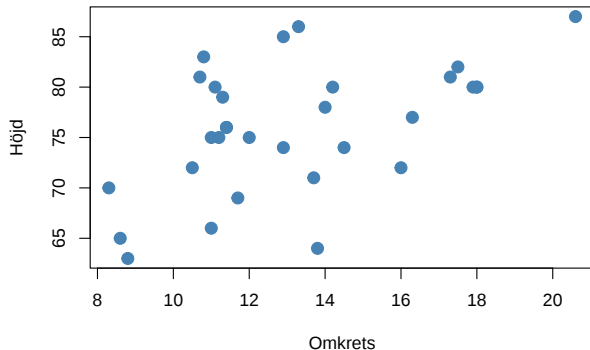


- Bara för att det finns *något* samband betyder det **inte** att det finns korrelation, utan korrelation mäter **bara** styrkan på **linjära** samband
- I figuren nedan finns ett uppenbart samband mellan god smak (y) och baktemperatur (x) för kladdkakor – ändå är korrelationen nära noll!



- I R kan vi enkelt rita ett spridningsdiagram för 2 numeriska variabler
- Datamaterialet `trees` anger omkrets (`girth`), höjd (`height`) och volym (`volume`) för 31 träd

```
data(trees)
plot(x=trees$Girth, y=trees$Height, pch=19, cex=1.5,
     col="steelblue", xlab="Omkrets", ylab="Höjd")
```



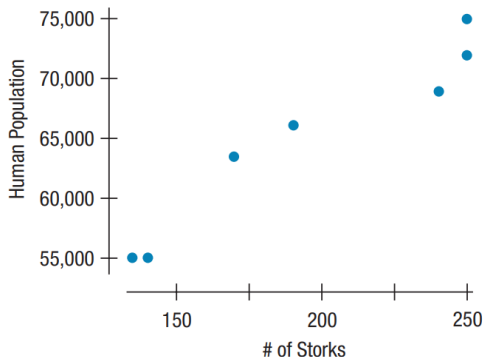
- Vi kan räkna ut korrelationskoefficienten för två variabler med funktionen `cor()`

```
cor(trees$Girth, trees$Height) # Korrelation mellan omkrets och höjd  
## [1] 0.5192801
```

- Detta ser ut att stämma relativt väl överens med bilden på föregående slide

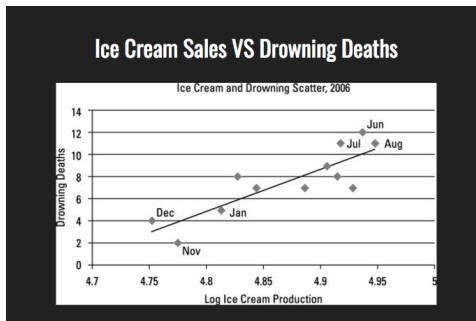
Korrleation och kausalitet

- Figur 6.10 i De Veaux et al. (2021) visar ett tydligt linjärt samband mellan antalet storkar och befolkningens mängden i en stad i Tyskland (data för sju år)
- Det sägs ibland att storkar kommer med barn, och det skulle kunna förklara varför antalet människor blir fler när det kommer många storkar
- Korrelationskoefficienten är $r = 0.97$, så det linjära sambandet är starkt



- Slutsatsen om att fler storkar leder till fler människor är en **kausal** tolkning av korrelationen – den gör en utsaga om ett **orsakssamband**
- Är slutsatsen rimlig? Skulle vi kunna få barn även om storkarna försvann?
- Kanske är sambandet det omvända? Storkar bygger bon på skorstenar, och fler människor betyder fler hus med skorstenar där storkarna kan bygga bon
- Allmänt bör vi vara försiktiga med att dra snabba slutsatser av det här slaget
- Det kan t.ex. finnas en **dold variabel** som förklarar både x och y , och som gör det svårt för oss att identifiera kausaliteten helt
- Detta betyder dock inte att kausalitet alltid saknas – ibland finns det kausalitet och ibland finns det inte
- Huvudsaken är att det krävs mer utförliga och specifikt designade studier för att etablera kausala relationer

- Harvard Onlines FB-sida: drunkningsolyckor (y) och glassproduktion (x) under ett år

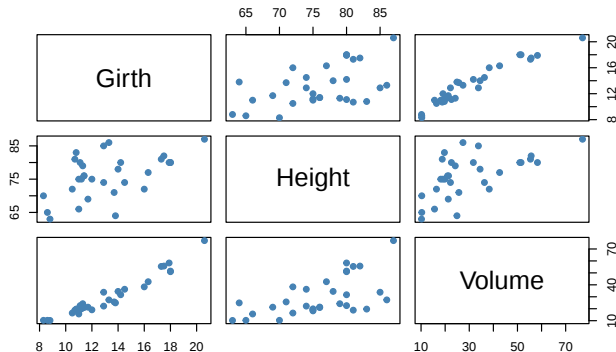


- **Kausal tolkning:** När det produceras mer glass måste folk äta upp glassen, och för mycket glass gör att människor får håll och drunknar i badet
- **Omvänd kausal tolkning:** Folk blir ledsna och tröstäter glass när de läser om drunkningsolyckor, så mer glass produceras i tider då många drunknar
- Den senare verkar raljant, men i kontexten är ingen mer rätt/fel än den andra

Samband mellan fler än två numeriska variabler

- Om vi har tre eller fler numeriska variabler kan vi göra parvisa spridningsdiagram med funktionen `pairs()` i R

```
# trees är en data frame  
pairs(trees, col="steelblue", pch=19, cex=1)
```



- Om vi har flera variabler kan vi skapa en tabell med parvisa korrelationer
- Tabell 6.1 i De Veaux et al. (2021) innehåller t.ex. en korrelationstabell för olika finansiella mått från Forbes

	Assets	Sales	Market Value	Profits	Cash Flow	Employees
Assets	1.000					
Sales	0.746	1.000				
Market Value	0.682	0.879	1.000			
Profits	0.602	0.814	0.968	1.000		
Cash Flow	0.641	0.855	0.970	0.989	1.000	
Employees	0.594	0.924	0.818	0.762	0.787	1.000

- Varför är alla värden längs diagonalen 1?
- Vad kan vi säga om den tomma delen av tabellen ovanför diagonalen?

- I R kan funktionen `cor()` även skapa en korrelationstabell, men då måste vi ge en hel data frame som input
- **Viktigt!** Detta fungerar bara om alla variabler i vår data frame är **numeriska**

```
cor(trees)
##           Girth    Height    Volume
## Girth  1.0000000  0.5192801  0.9671194
## Height 0.5192801  1.0000000  0.5982497
## Volume 0.9671194  0.5982497  1.0000000
```

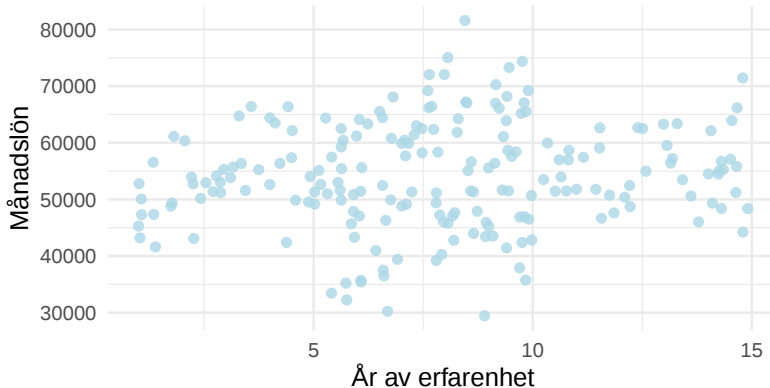
- Jämför matrisens nedre vänstra del med dess övre högra del!

- Vi ska nu titta på ett exempel där en dold variabel (lurking variable) helt förändrar vår tolkning av sambandet
- **Scenario:** Vi undersöker sambandet mellan lön och yrkeserfarenhet hos ett företag
- När vi plottar alla data tillsammans ser vi en typ av samband
- När vi betingar på en nyintroducerad variabel, yrkeskategori, så blir bilden en annan
- Notera att detta är påhittade data som jag har simulerat i R

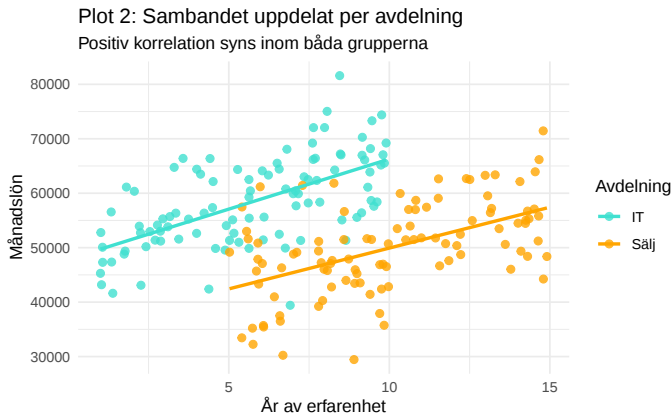
- Vi kan inte se något (eller möjligtvis ett väldigt svagt) samband mellan erfarenhet och lön

Samband mellan erfarenhet och lön (Totalt urval)

Övergripande korrelation $r = 0.13$



- Nu kan vi tydligt se ett samband mellan erfarenhet och lön, när vi har betingat på yrkeskategori
- Detta är ett annat exempel på Simpson's paradox (se föreläsning 3)



Tack för idag!!