Stockholm University

Matias Quiroz[1]

[1]Statistiska institutionen, Stockholms universitet

VT 2023

Innehåll

- Punktdiagram för att beskriva samband mellan två numeriska variabler.
- Korrelation som mått för linjära samband.
- Korrelation gentemot kausalitet.
- Parvisa samband mellan flera numeriska variabler.
- Transformationer för att uppnå linjära samband.

Statistik studerar samband

- En viktig del i statistik är att **studera samband mellan variabler**.
- Samband hjälper oss att förstå hur två eller fler variabler förhåller sig till varandra.
- Samband mellan variabler är också användbara vid **prediktion**. **Regressionsanalys** (Föreläsning 8 och 9).
- Om två variabler har ett samband kan vi använda en variabels värde för att prediktera (förutsäga) värdet på den andra.
- Vi har studerat följande samband:
  - Samband mellan två kategoriska variabler. Korstabeller.
  - Samband mellan tre kategoriska variabler. Korstabeller.
  - Samband mellan en numerisk variabel och en kategorisk. Låddiagram eller histogram för varje grupp i den kategoriska variabeln.
- Nästa naturliga steg är att studera **sambandet mellan två numeriska variabler**.

Samband mellan två numeriska variabler

- Vi har tidigare stött på tidsserier.
- Tidsserieanalys studerar hur en variabel (eller flera) varierar över tid. Samband mellan två numeriska variabler.
- Punktdiagram (**scatter plot** på engelska) visualiserar samband mellan två numeriska variabler.
- Tidsserie av prediktionsfel i lokalisering av orkaner i Atlanten:

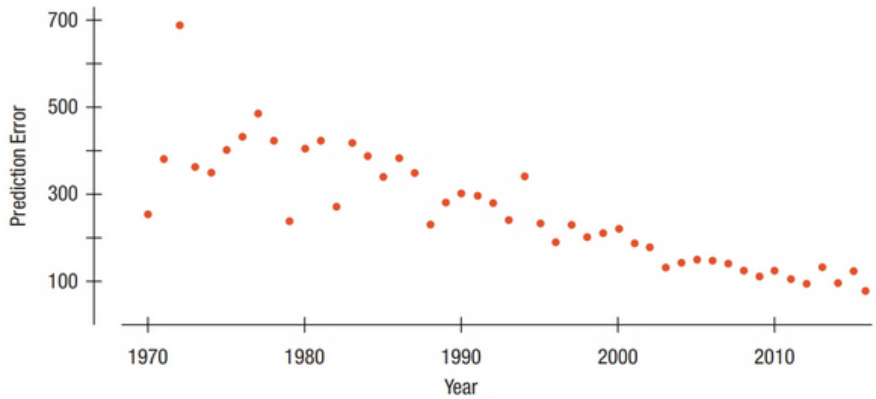

Figure 1: Figur 6.1 i De Veaux et al. (2021).

- Negativt samband: När x axeln ökar så minskar y variabeln. Sambandet ter sig (approximativt) linjärt. Lokalisering av orkaner blir säkrare med tid.

Samband mellan två numeriska variabler, forts.

- Data från 73 olika länder. Några samband:

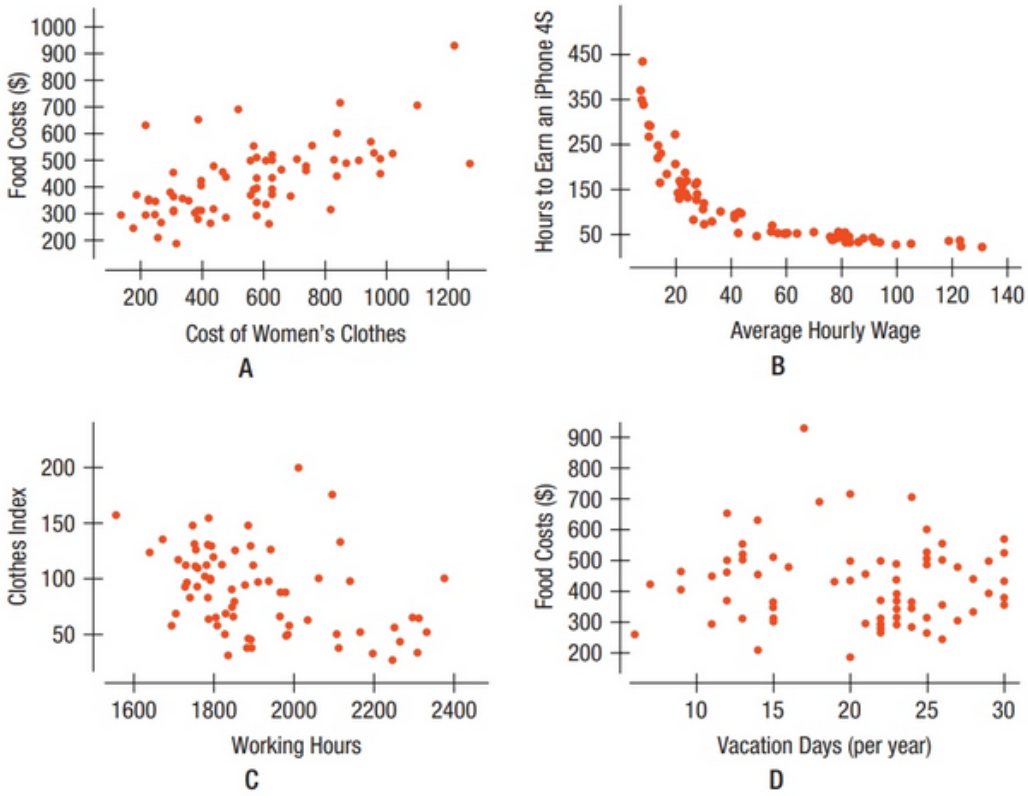

Figure 2: Figur från s.195 i De Veaux et al. (2021).

- Kommentarer:
  - A: Linjärt (approximativt) positivt samband. När $x \uparrow$, $y \uparrow$. När $x \downarrow$, $y \downarrow$. Länder med dyrare kvinnokläder tenderar att ha högre matpriser.

Samband mellan två numeriska variabler, forts.

- Data från 73 olika länder. Några samband:

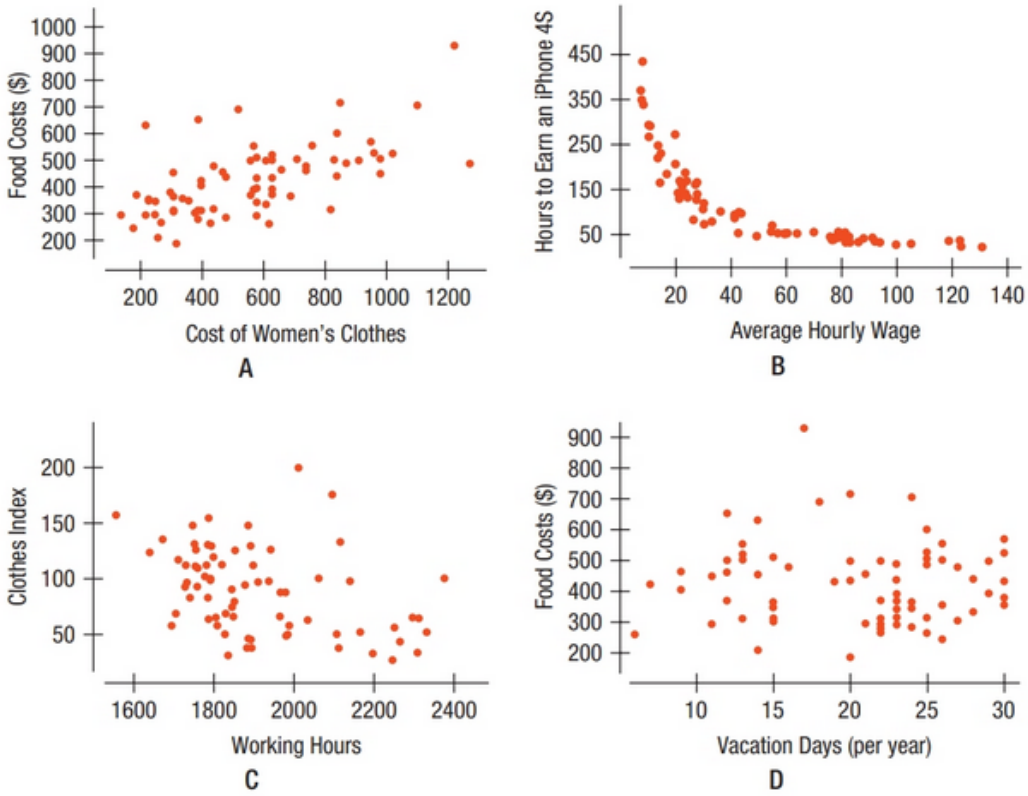

Figure 3: Figur från s.195 i De Veaux et al. (2021).

- Kommentarer (forts.):
  - B: Icke-linjärt negativt samband. När $x \uparrow$, $y \downarrow$. När $x \downarrow$, $y \uparrow$. I länder med högre timlön tenderar antal arbetstimmar för att köpa en iPhone vara lägre.

Samband mellan två numeriska variabler, forts.

- Data från 73 olika länder. Några samband:

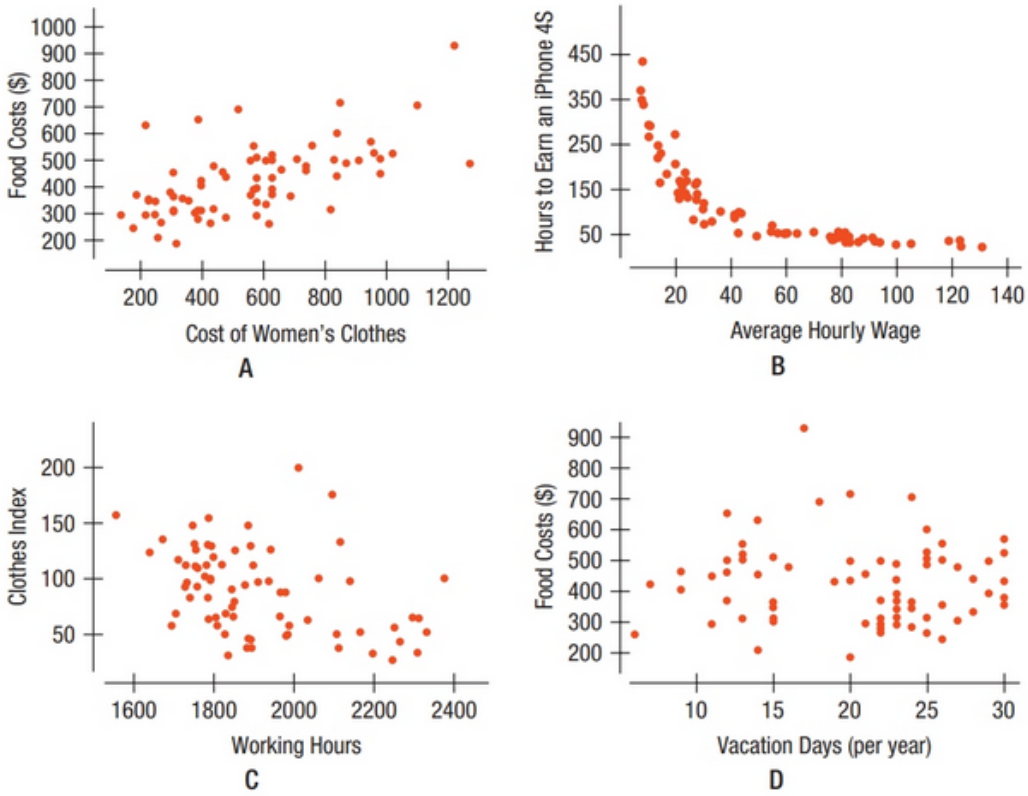

Figure 4: Figur från s.195 i De Veaux et al. (2021).

- Kommentarer (forts.):
  - C: Linjärt (approximativt) negativt samband. När $x \uparrow$, $y \downarrow$. När $x \downarrow$, $y \uparrow$. I länder där man arbetar mer är clothes index lägre. Svagare samband än i A.

Samband mellan två numeriska variabler, forts.

- Data från 73 olika länder. Några samband:

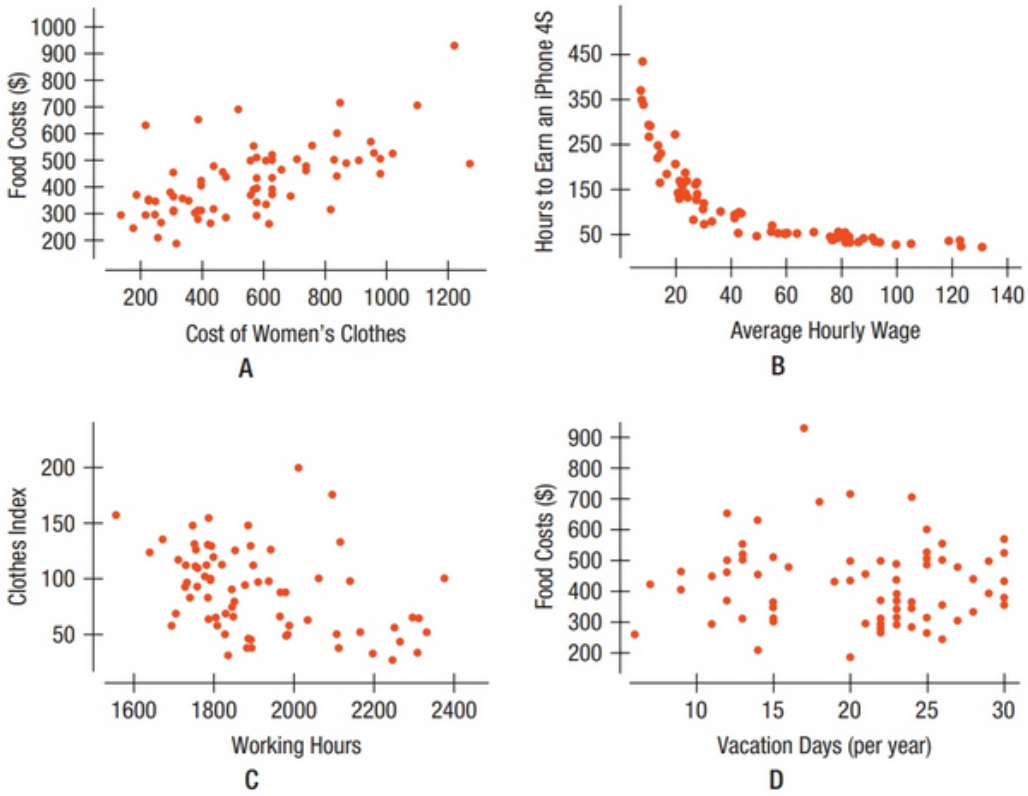

Figure 5: Figur från s.195 i De Veaux et al. (2021).

- Kommentarer (forts.):
  - D: Inget samband. När $x \uparrow$, $y \downarrow\uparrow$. När $x \downarrow$, $y \uparrow\downarrow$. Inget samband mellan antal semesterdagar och matpriser.

Samband mellan två numeriska variabler, forts.

- Hur bestämmer vi vilken variabel som är y och vilken som är x?
- Enklast att ha prediktion i åtanke för att avgöra frågan.
- Variabeln **av intresse att prediktera** väljs som y.
- Variabeln som **hjälper oss att prediktera** väljs som x.
- Variabeln y kallas också för responsvariabeln (**response variabel** på engelska).
- Variabeln x kallas också för den förklarande variabeln (**explanatory variabel** på engelska).
- Ett annat vanligt namn för y är beroende variabel (**dependent variable** på engelska), och oberoende variabel (**independent variable**) för x .
- Andra vanligt förekommande namn för x variabeln: **Prediktor** och **kovariat**.
- Inom maskininlärning kallas förklarande variabler **features**.

Samband mellan två numeriska variabler, forts.

- Betrakta återigen datan från 73 olika länder:

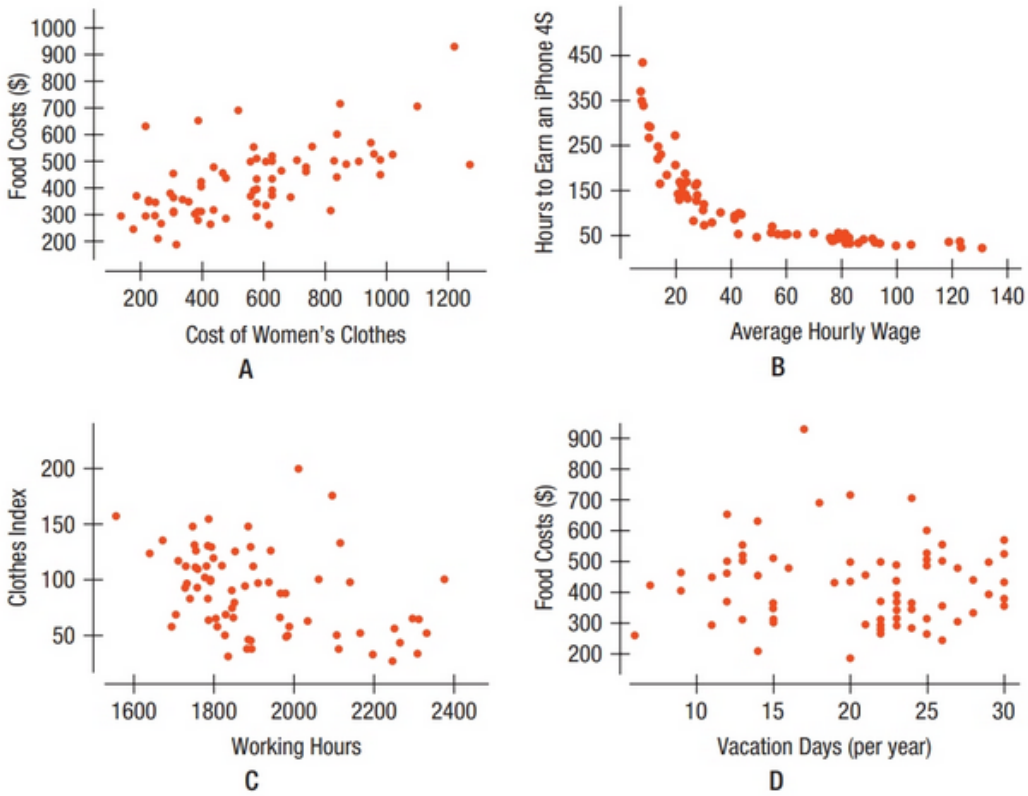

Figure 6: Figur från s.195 i De Veaux et al. (2021).

- Olika styrkor på sambanden.
- Bland de linjära sambanden tycks A vara starkare (mer uttalat) än C.

Samband mellan två numeriska variabler, forts.

- Vårt mål är att kunna mäta linjära samband med ett tal mellan -1 och 1.
- -1 beskrivet ett perfekt negativt samband.
- 1 beskrivet ett perfekt positivt samband.
- Styrkan på sambandet bör vara oberoende av vilken enhet mätningarna är i.
- Exempel: Studenters vikt och längd. Samma positiva samband oavsett enheter.

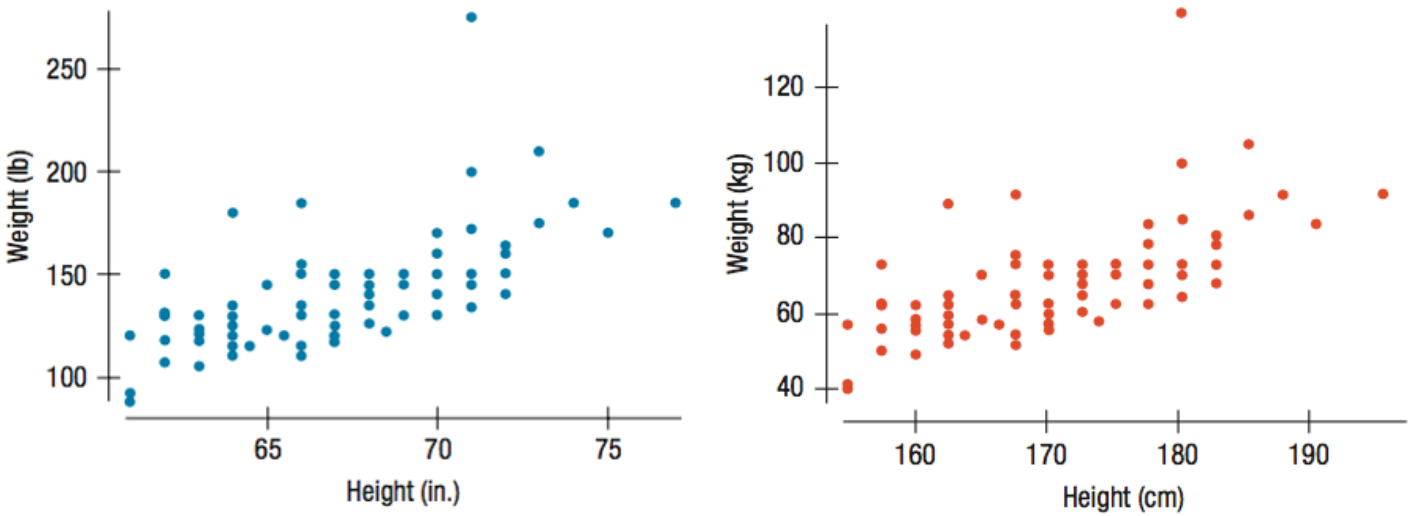

Figure 7: Figur 6.2 och 6.3 i De Veaux et al. (2021).

Samband mellan två numeriska variabler, forts.

- Betrakta studenters vikt och längd igen:

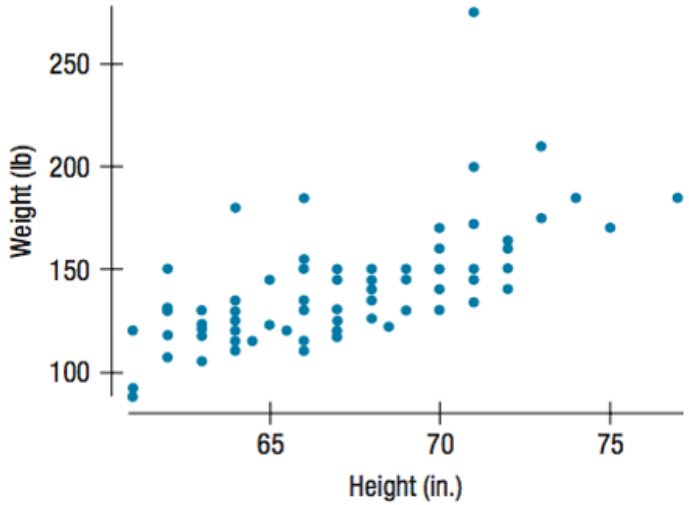

Figure 8: Figur 6.2 och 6.3 De Veaux et al. (2021).

- Vilka datapunkter bidrar till ett **positivt** linjärt samband?
- Vilka datapunkter bidrar till ett **negativt** linjärt samband?
- Låt oss markera medelvärdet för x, dvs $\overline{x}$, som ett vertikalt streck och medelvärdet för y, dvs $\overline{y}$, som ett horisontellt streck i figuren.

Samband mellan två numeriska variabler, forts.

- Studenters vikt och längd med respektive medelvärden utmarkerade:

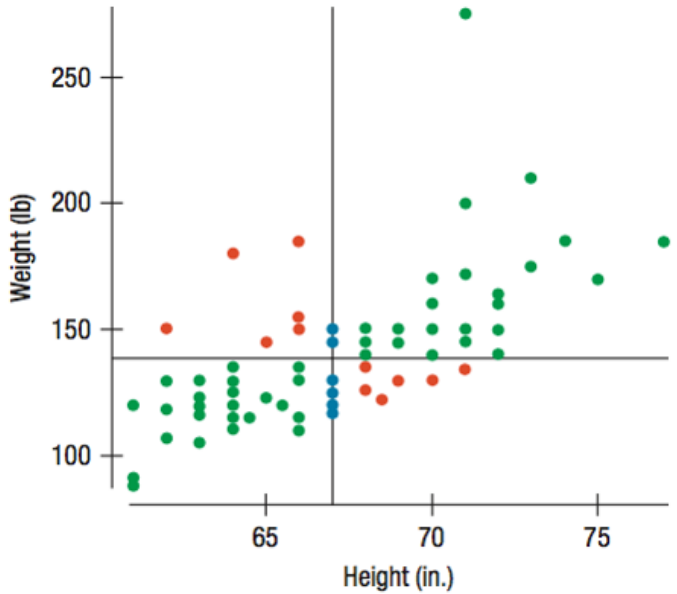

Figure 9: Figur 6.4 i De Veaux et al. (2021).

- Gröna punkter bidrar till ett linjärt **positivt samband**.
  - Första kvadranten: $x - \overline{x} > 0$ och $y - \overline{y} > 0$.
  - Tredje kvadranten: $x - \overline{x} < 0$ och $y - \overline{y} < 0$.

Samband mellan två numeriska variabler, forts.

- Studenters vikt och längd med respektive medelvärden utmarkerade:

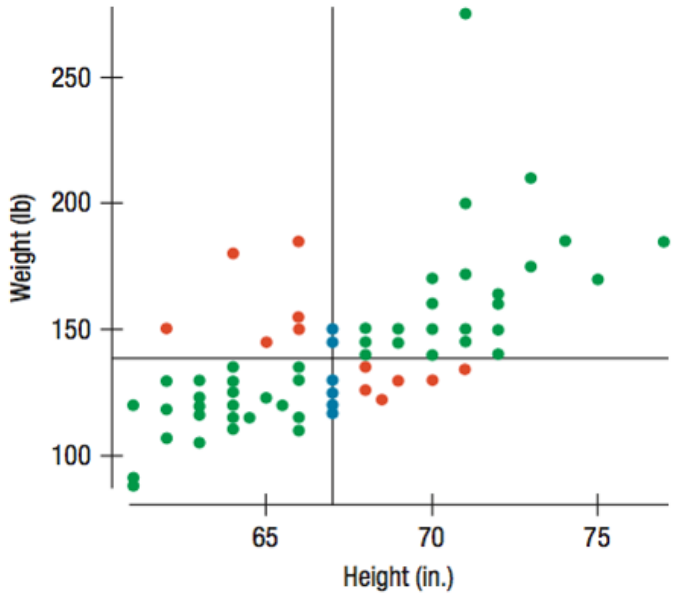

Figure 10: Figur 6.4 i De Veaux et al. (2021).

- Röda punkter bidrar till ett linjärt **negativt** samband.
  - Andra kvadranten: $x - \overline{x} < 0$ och $y - \overline{y} > 0$.
  - Fjärde kvadranten: $x - \overline{x} > 0$ och $y - \overline{y} < 0$.

Samband mellan två numeriska variabler, forts.

- Studenters vikt och längd med respektive medelvärden utmarkerade:

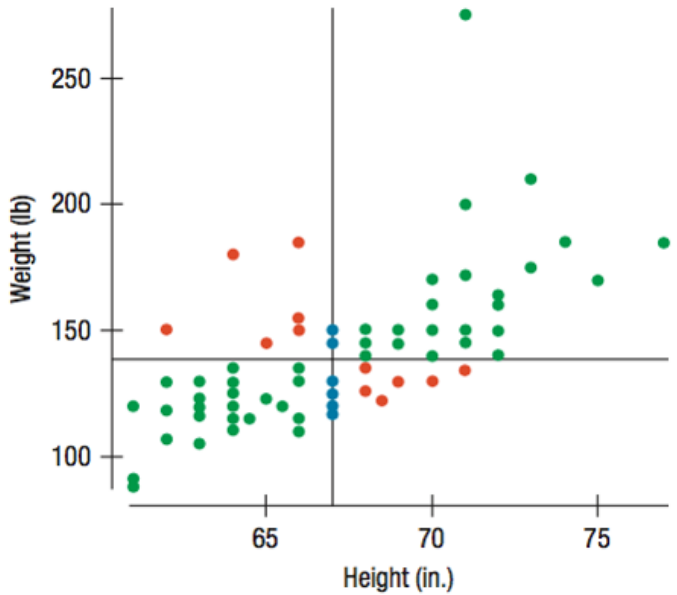

Figure 11: Figur 6.4 i De Veaux et al. (2021).

- Blå punkter **bidrar inte** till ett linjärt samband.
  - $x - \overline{x} = 0$ och/eller $y - \overline{y} = 0$.
- I det här datasetet är $x - \overline{x} = 0$ för de blåa punkterna.

Samband mellan två numeriska variabler, forts.

- Studenters vikt och längd med respektive medelvärden utmarkerade:

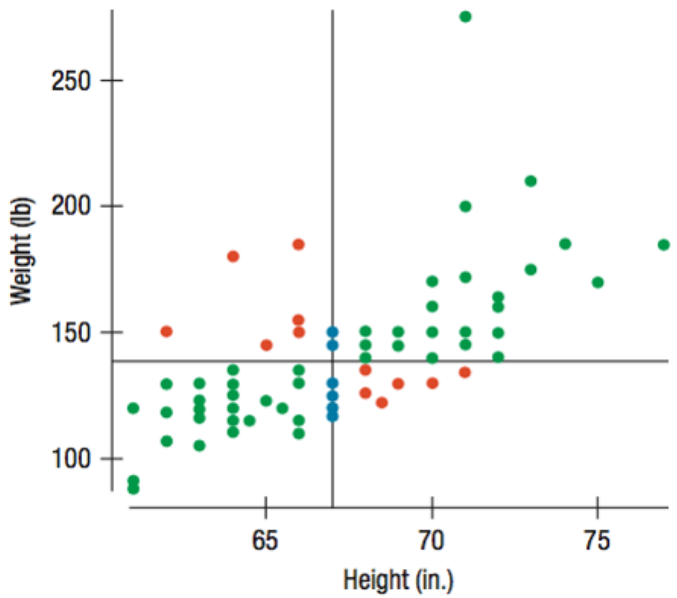

Figure 12: Figur 6.4 i De Veaux et al. (2021).

- Eftersom vi har fler gröna punkter, med längre sammanlagt avstånd till de nya axlarna än de röda, är det sammanvägda linjära sambandet positivt.
- Hur kan vi sammanfatta det linjära samtalet med ett enda tal?
- Ett av våra önskemål var att **måttet inte ska bero på enheten**. Har vi stött på en variabel vars värden är enhetslösa?

Samband mellan två numeriska variabler, forts.

- Standarisering! Standarisera båda variablerna separat.
- Beräkna z-värdet för x variabeln,

$$z_x = \frac{x - \overline{x}}{s_x}, \ \overline{x} = \frac{\sum x}{n}, \ s_x = \sqrt{\frac{\sum (x - \overline{x})^2}{n-1}}.$$

  och z-värdet för y variabeln,

$$z_y = \frac{y - \overline{y}}{s_y}, \ \overline{y} = \frac{\sum y}{n}, \ s_y = \sqrt{\frac{\sum (y - \overline{y})^2}{n-1}}.$$

- Både z_x och z_y har medelvärden 0 och standardavvikelser 1.

Samband mellan två numeriska variabler, forts.

- Punktdiagram för (z_x, z_y) (istället för (x, y))

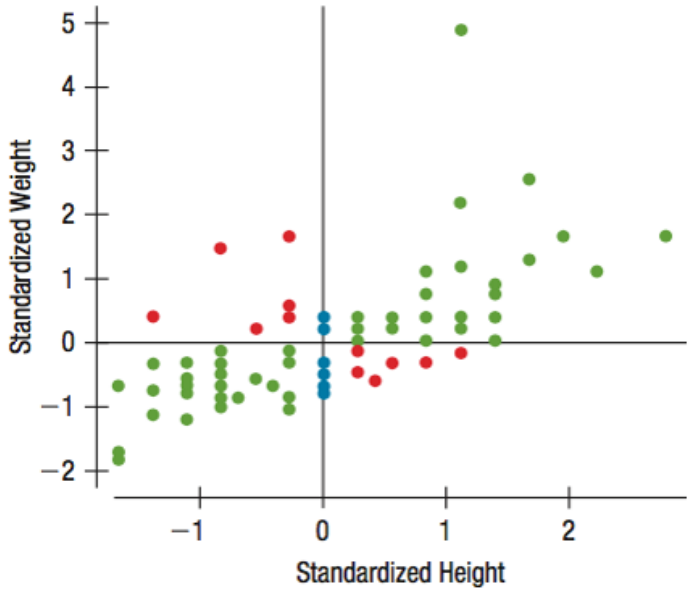

Figure 13: Figur 6.5 i De Veaux et al. (2021).

- Gröna punkter bidrar till ett linjärt positivt samband.
  - Första kvadranten: $z_x > 0$ och $z_y > 0$.
  - Tredje kvadranten: $z_x < 0$ och $z_y < 0$.
- I bägge fall är **produkten positiv**, dvs $z_x z_y > 0$.

Samband mellan två numeriska variabler, forts.

- Punktdiagram för (z_x, z_y) (istället för (x, y))

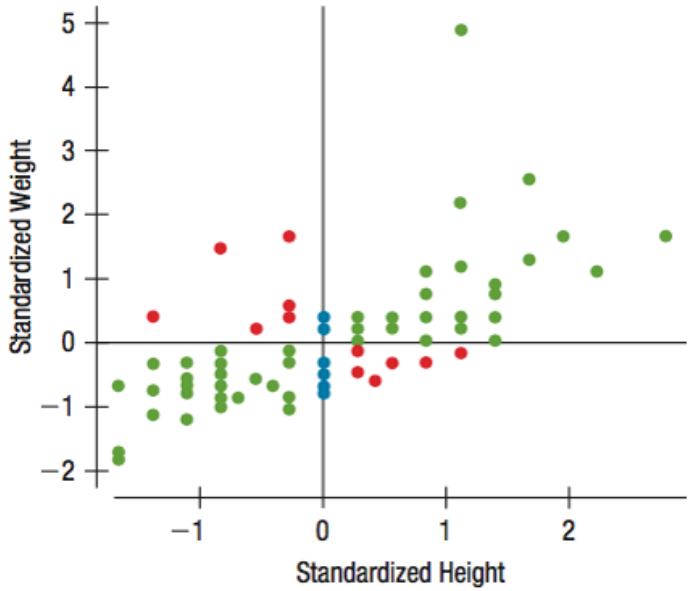

Figure 14: Figur 6.5 i De Veaux et al. (2021).

- Gröna punkter bidrar till ett linjärt positivt samband.
  - Andra kvadranten: $z_x < 0$ och $z_y > 0$.
  - Fjärde kvadranten: $z_x > 0$ och $z_y < 0$.
- I bägge fall är **produkten negativ**, dvs $z_x z_y < 0$.

Samband mellan två numeriska variabler, forts.

- Punktdiagram för (z_x, z_y) (istället för (x, y))

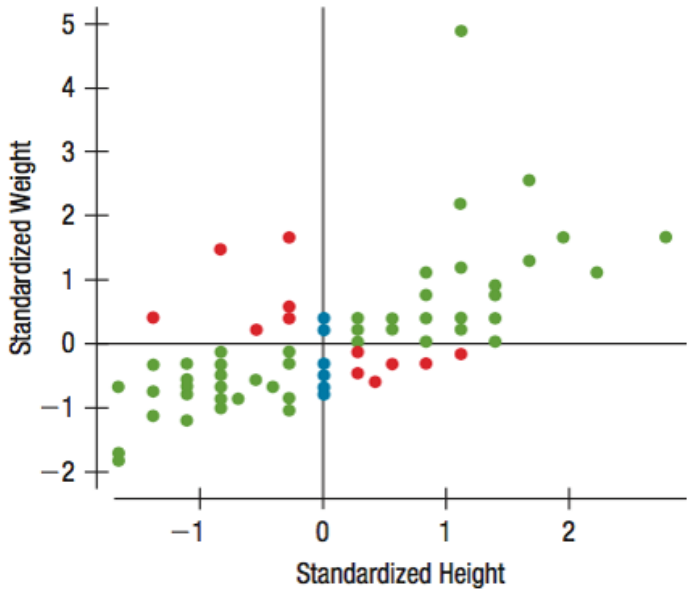

Figure 15: Figur 6.5 i De Veaux et al. (2021).

- Blå punkter bidrar inte till ett linjärt positivt samband.
  - $z_x = 0$ och/eller $z_y = 0$.
- I bägge fall är **produkten noll**, dvs $z_x z_y = 0$.
- I det här datasetet är $z_x = 0$ för de blåa punkterna.

Samband mellan två numeriska variabler, forts.

- Ett möjligt mått på det linjära sambandet är $\sum z_x z_y$.
- Observationer med större (oavsett tecken) $z_x z_y$ bidrar mer till $\sum z_x z_y$.
- $\sum z_x z_y$ växer med n, antal observationer.
- Betrakta (nästan) medelvärdet av $\sum z_x z_y$ istället,

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

- Kvantiteten r kallas för korrelationskoefficienten för stickprovet (**sample correlationcoefficient** på engelska). Betecknas r_{xy} ibland.
- r mäter **korrelationen mellan två numeriska variabler**.
- Varför delar vi med $n-1$ i (1) och inte med n? Vi behöver lära oss mer statistik innan vi kan förklara det.
- **Korrelationsmåttet är symmetrisk**: Korrelationen mellan x och y är densamma som mellan y och x.

Samband mellan två numeriska variabler, forts.

- Man kan visa att $-1 \leq r \leq 1$. Vidare:
  - $r = 1$ ger perfekt positiv korrelation. Punkterna ligger på en exakt linje med positiv lutning.
  - $r = -1$ ger perfekt negativ korrelation. Punkterna ligger på en exakt linje med negativ lutning.
  - $r > 0$ för ett linjärt positivt samband. Ju närmare 1, desto starkare det positiva linjära sambandet.
  - $r < 0$ för ett linjärt negativt samband. Ju närmare 1, desto starkare det negativa linjära sambandet.
  - Tecknet, samt storleken, på r kan vara känsligt för outliers.
- **Korrelation är inte samma sak som samband**.
- Ett samband är en association mellan två variabler som kan vara väldigt komplex, inte nödvändigtvis linjärt.
- Vi kan tala om samband för såväl numeriska som kategoriska variabler.
- Korrelation är ett **linjärt samband mellan två numeriska variabler**.

Samband mellan två numeriska variabler, forts.

- Smakbetyg mot bakningstemperatur för kladdkakor:

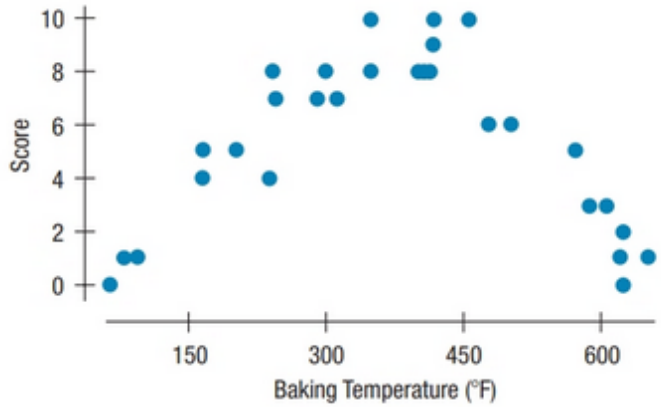

Figure 16: Figur från s.212 i De Veaux et al. (2021).

- Korrelationskoefficienten är $r \approx 0.05$, väldigt liten. Felaktig att säga att det inte finns ett samband mellan smakbetyget och bankningstemperatur.
- Använd rätt språk: Korrelation är ett linjärt samband.
- Frånvaro av korrelation betyder inte nödvändigtvis att ett samband inte finns. Bra att plotta!

Samband mellan två numeriska variabler, forts.

- Korrelation innebär inte kausalitet. På samma sätt som ett samband inte innebär ett kausalt samband.
- Låt oss bevisa storkmyten med statistik genom att vantolka korrelation.
- Storkmyten är en berättelse om hur barn blir till. Det är storkfåglar som levererar barn.
- Befolkningsmängd mot antalet storkar under sju år i en stad i Tyskland.

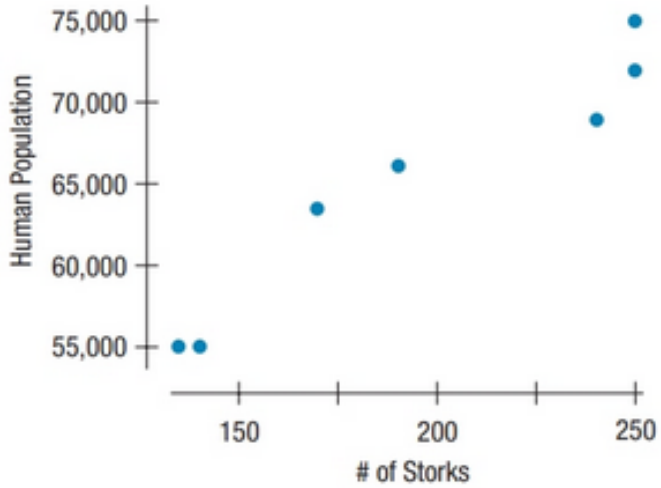

Figure 17: Figur 6.10 i De Veaux et al. (2021).

- Linjärt samband föreligger, variablerna är positivt korrelerade $r = 0.97$.

Samband mellan två numeriska variabler, forts.

- Kausal tolkning: Fler storkar kan leverera fler bebisar och ger därmed en större befolkningsmängd. Data stödjer storkmyten.
- I det här fallet är kausaliteten snarare omvänd enligt följande resonemang.
- Storkar bygger ofta sina bon på skorstenar. Ökad befolkningsmängd innebär fler hus på vilka storkar kan bygga sina bon.
- Alltså är det ökning av befolkningsmängden som orsakar fler storkar, och inte tvärtom.
- Ibland finns det inget kausalt samband åt något håll, utan en tredje variabel som istället förklarar både x och y. **Lurking variable** på engelska.

Samband mellan två numeriska variabler, forts.

- Drunkningsolyckor mot glassproduktion:

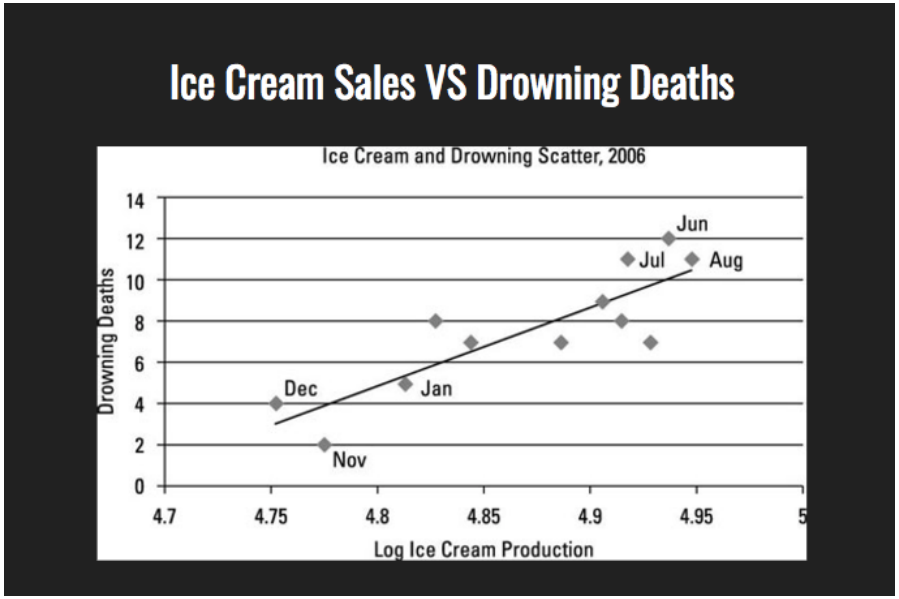

Figure 18: Figur från Harvard Onlines `Facebooksida`.

- Kausal tolkning: Mer glassproduktion innebär större glasskonsumption. Överkonsumption av glass gör folk till sämre simmare, vilket medför fler drunkningsolyckor.

Samband mellan två numeriska variabler, forts.

- Drunkningsolyckor mot glassproduktion:

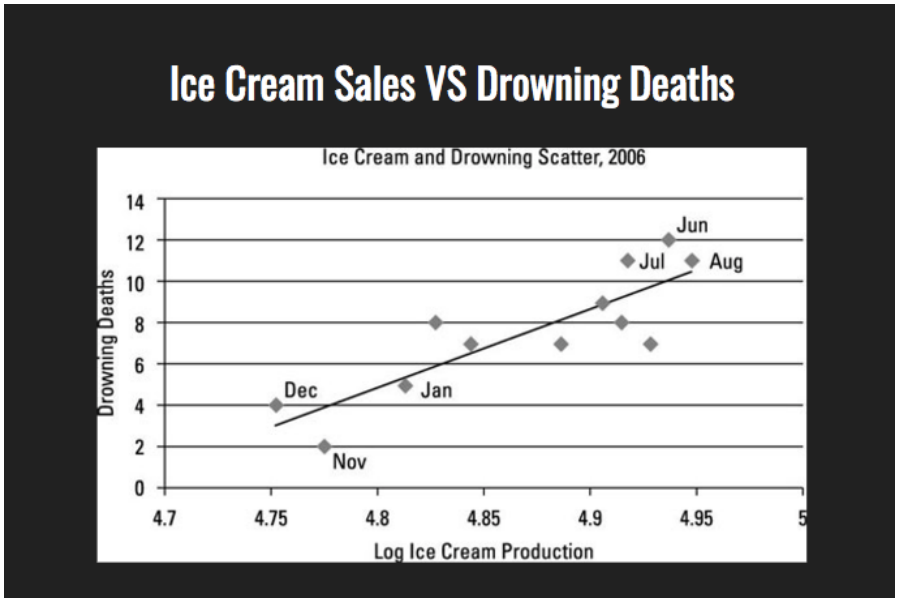

Figure 19: Figur från Harvard Onlines `Facebooksida`.

- Omvänd kausal tolkning: Folk blir deprimerade när de läser om drunkningsolyckor vilket gör dom glassugna. Efterfrågan på mer glass innebär att glassproduktionen ökar.

- Både drunkningsolyckor och glassproduktion beror på årstid. Årstiden (här i månader) är en lurking variabel som förklarar både x och y.

Parvisa samband mellan flera numeriska variabler, forts.

- När vi har tre eller fler numeriska variabler kan vi göra parvisa punktdiagram för att visuellt studera samband. `pairs` funktionen i R.
- Parvisa punktdiagram för omkrets, längd och volym för 31 träd:

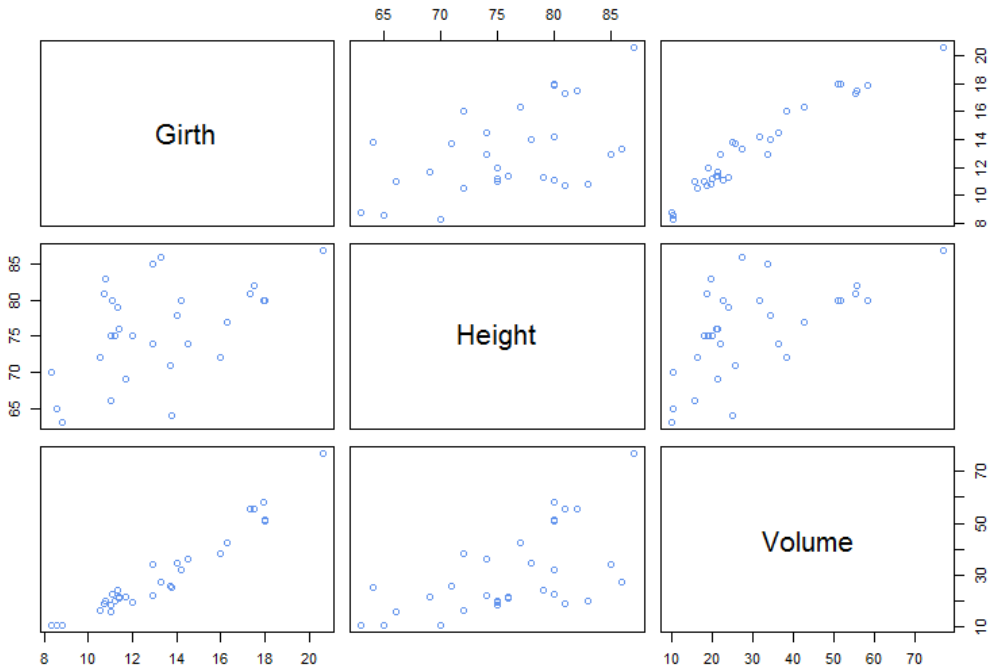

Figure 20: Skapad med `pairs(trees, col="cornflowerblue")` i R.

Parvisa samband mellan flera numeriska variabler, forts.

- För att studera parvisa linjära samband, dvs korrelationer, kan vi göra en korrelationstabell (**correlation table** på engelska).
- En korrelationstabell innehåller alla parvisa korrelationer.
- 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

Figure 21: Tabell 6.1 i De Veaux et al. (2021).

- Varför är diagonalen 1?
- Vad kan vi säga om korrelationerna ovanför diagonalen?

Vad händer om vi får nya data?

- Som vanligt: Nya data ger ny deskriptiv statistik.
- Om vi tar ett nytt stickprov ändras korrelationskoefficienten.
- Hur varierar korrelationskoefficienten från stickprov till stickprov?
- Ett stickprov på 50 observationer från populationen föräldrar i USA.

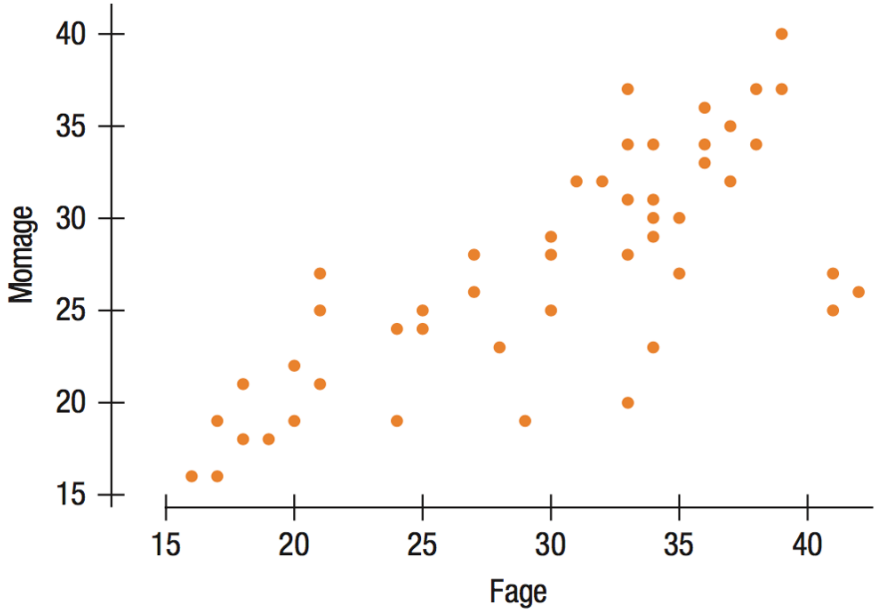

Figure 22: Figur 6.6 i De Veaux et al. (2021).

- Tag 10 000 nya stickprov, varje stickprov består av 50 observationer.

Vad händer om vi får nya data?, forts.

- Histogram för korrelationer från 10 000 olika stickprov:

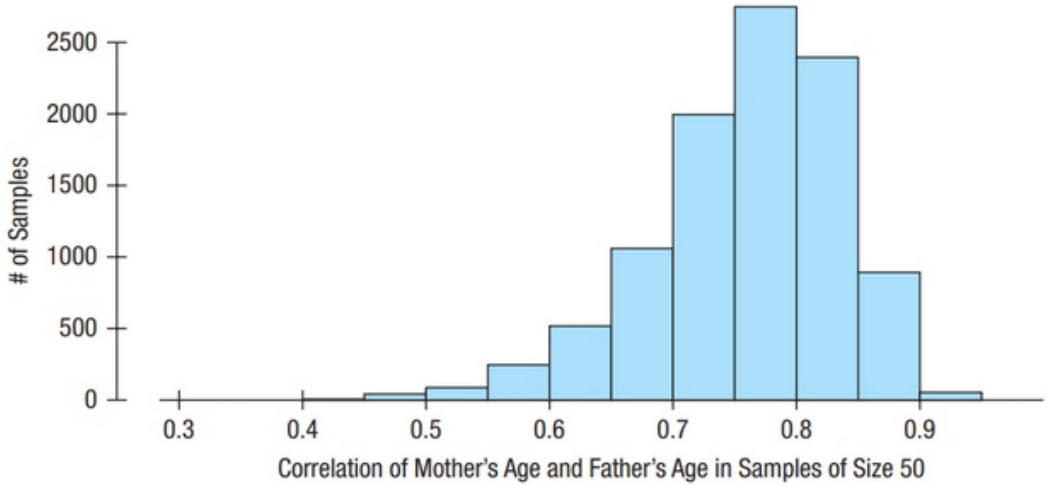

Figure 23: Figur 6.7 i De Veaux et al. (2021).

- Detta är **samplingfördelningen** för korrelationskoefficienten.
- I det här exemplet har vi tillgång till populationen, och kan räkna dess korrelationscoefficient till 0.757.
- Notera att samplingfördelningen för korrelationskoefficienten är skev.

Transformationer

- Vi har tidigare använt transformationer för att bland annat få en mer symmetrisk fördelning. Föreläsning 5.
- Transformationer kan också användas för att "räta ut" ett punktdiagram, dvs få ett mer linjärt förhållande mellan y och x.
- Vitsen med detta är att kunna använda korrelation för att beskriva linjära samband.
- För att få en meningsfull tolkning av korrelation måste det transformerade datat visa ett linjärt samband.

Transformationer, forts.

- Stege av potenstransformationer (inklusive ingen transformation y):

$$\text{(toppen)} \quad y^2, y, y^{1/2}, \log(y), -y^{-1/2}, -y^{-1} \text{ (botten)},$$

  kallas **ladder of powers** på engelska.

- Gå igenom stegen tills sambandet ter sig någorlunda linjärt.
- Notera att $y^{1/2} = \sqrt{y}$, $-y^{-1/2} = -1/\sqrt{y}$ och $-y^{-1} = -1/y$.
- Utan det negativa tecknet på de två sista transformationer så bevaras inte ordningen. Exempel om sista transformationen inte har ett minustecken:

$$3 < 6, \text{ och för de transformerade värden } 1/3 > 1/6.$$

- Med det **negativa tecknet bevaras ordningen**, eftersom

$$3 < 6, \text{ och för de transformerade värden } -1/3 < -1/6.$$

- Vi kommer lära oss tumregler för transformationer när vi går igenom regression på nästa föreläsning.

References I

De Veaux, R. D., Velleman, P., and Bock, D. (2021). *Stats: Data and Models*. Pearson, Harlow, United Kingdom, fifth edition.