

Note: Statistik og kombinatorik

Population og stikprøve

En population af er en stor mængde af individer. Det kan være mennesker, dyr, planter, ting eller mere abstrakte genstande. Populationen er dog almindeligvis for stor til, at vi kan undersøge alle individer i den. Ofte interesserer vi os for den *andel* af populationens individer, som har en bestemt egenskab. Vi kalder andelen for *populationsandelen* p .

For at bestemme populationsandelen udtager vi n individer på en tilfældig måde. De udtagne individer kalder vi for en *stikprøve*. Andelen af individer som har egenskaben vi undersøger for, kaldes for *stikprøveandelen* $\hat{p}$.

Konfidensinterval

Statistisk set kan vi ikke forvente at $p = \hat{p}$. Omvendt kan vi godt forvente, at afvigelsen mellem p og $\hat{p}$ ikke er alt for stor. Afvigelsen kaldes den statistiske usikkerhed s .

Hvis stikprøvens størrelse er n og den målte stikprøveandel $\hat{p}$, så vil usikkerheden være givet ved $s = k \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}$. Her er k et tal der angiver *konfidensniveauet*, det vil sige hvor sikre vi er på at den faktiske værdi af p ligger højest s fra $\hat{p}$. Hvis sandsynligheden skal være ca. 95%, vælges $k = 2$.

På et 95% konfidensniveau bliver den statistiske usikkerhed altså til:

$$s = 2 \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}$$

For $k = 1$ fås et konfidensniveau på ca. 68% og for $k = 3$ fås et niveau på 99,7%.

Vi kan således opstille et 95% *konfideninterval*, som er et interval af andele som den rigtige populationsandel p med 95% sikkerhed ligger i:

$$[\hat{p} - s; \hat{p} + s] = \left[\hat{p} - 2 \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}; \hat{p} + 2 \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} \right]$$

Kombinatorik

Sætning:

Antallet af kombinationer vi kan lave af r elementer valgt blandt n elementer er:

$$K(n, r) = \frac{n!}{r! \cdot (n - r)!}$$

Bevis:

Hvis vi har n elementer, kan disse sættes i rækkefølge på $n!$ måder.

Dette følger af at vi på første plads har n muligheder, på næste $n - 1$, på næste $n - 2$, og så videre, indtil vi på næstsidste plads har 2 muligheder og på sidste plads kun 1. Med både-og-princippet fås:

$$n \cdot (n - 1) \cdot (n - 2) \cdot \dots \cdot 2 \cdot 1 = n!$$

Hvis vi kun skal lave rækkefølger af r elementer valgt blandt de n , vil vi have n muligheder på første plads, $n - 1$ på næste plads, $n - 2$ på tredje plads, osv. ned til plads $n - (r + 1)$. Med både-og-princippet bliver dette til:

$$\begin{aligned} & n \cdot (n - 1) \cdot (n - 2) \cdot \dots \cdot (n - (r + 1)) = \\ & \frac{n \cdot (n - 1) \cdot (n - 2) \cdot \dots \cdot (n - (r + 1)) \cdot (n - r) \cdot (n - r - 1) \cdot \dots \cdot 2 \cdot 1}{(n - r) \cdot (n - r - 1) \cdot \dots \cdot 2 \cdot 1} = \\ & \frac{n!}{(n - r)!} \end{aligned}$$

Antal rækkefølger (permutationer) af r elementer valgt blandt n elementer kan altså bestemmes som:

$$P(n, r) = \frac{n!}{(n - r)!}$$

For r specifikke elementer valgt blandt n gælder, at de kan sættes i $r!$ rækkefølger. Antallet af mulige kombinationer af r elementer valgt blandt n elementer kan således bestemmes som:

$$\frac{\text{Antal rækkefølger af } r \text{ valgt blandt } n}{\text{Antal rækkefølger af } r \text{ elementer}} = \frac{P(n, r)}{r!} = P(n, r) \cdot \frac{1}{r!} = \frac{n!}{(n - r)!} \cdot \frac{1}{r!} = \frac{n!}{(n - r)! \cdot r!}$$

Dermed er det vist, at antallet af kombinationer af r elementer valgt blandt n er:

$$K(n, r) = \frac{n!}{(n - r)! \cdot r!}$$

Lineær regression

Antag at vi har en serie observerede sammenhørende data: $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$.

På tabelform kan det skrives:

x	x_1	x_2	$\dots$	x_n
y	y_1	y_2	$\dots$	y_n

Vi kalder disse data for ”observerede værdier”. Særligt interesseret er vi i de observerede y -værdier.

Derudover har vi en lineær model der søger at beskrive hvordan y afhænger af x :

$$f(x) = a \cdot x + b$$

For hver observeret x -værdi fremkommer da en modelværdi for y (også kaldet *forventet værdi*):

$$f(x_1), f(x_2), \dots, f(x_n)$$

Vi interesserer os nu for forskellen mellem den observerede y -værdi og den forventede modelværdi for y . Forskellen kalder vi for *residualet* og beregner som:

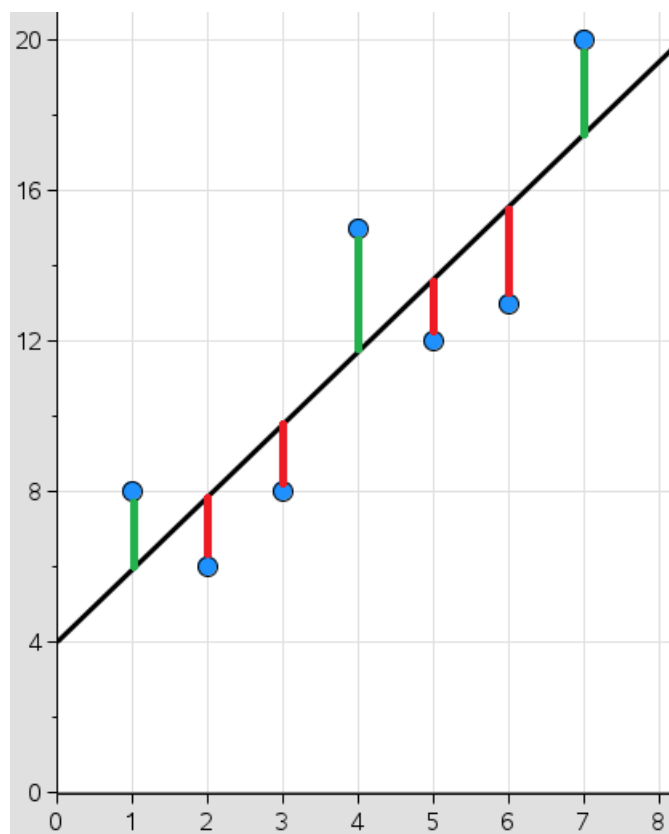
$$r_i = y_i - f(x_i)$$

Grafisk svarer residualet til den lodrette afstand fra det observerede punkt til den linje der er graf for modellen.

Der regnes dog med fortegn.

Hvis punktet ligger over linjen, er residualet positivt.

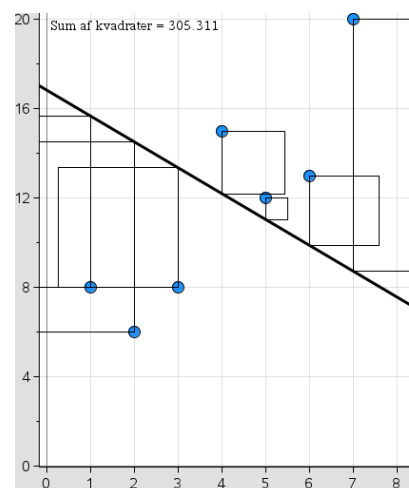
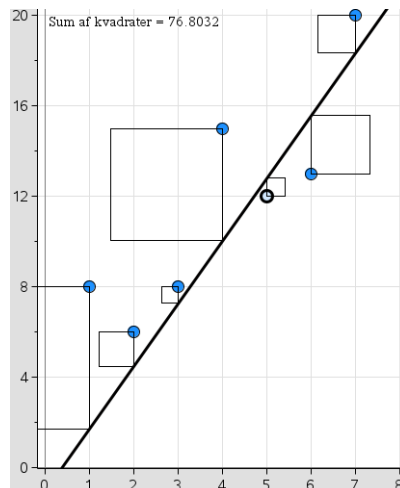
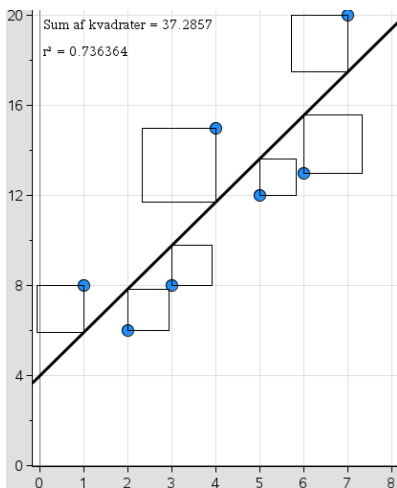
Hvis punktet ligger under linjen, er residualet negativt.



Når man skal vurdere hvor en ret linje er til at beskrive punkterne, ser man på summen af de kvadrerede residualer:

$$S = r_1^2 + r_2^2 + \dots + r_n^2$$

Grafisk set svarer denne sum til det samlede areal af de kvadrater man kan lave ud fra afstandene mellem punkt og linje. Her er tre eksempler, hvor det første er den ”bedste rette linje” - det vil sige den linje som har den mindst mulige værdi af S .



For at vurdere hvor god en bedste rette linje er, regnes gennemsnittet af de observerede y -værdier ud:

$$\bar{y} = \frac{y_1 + y_2 + \dots + y_n}{n}$$

Derpå findes residualkvadratsummen T for linjen: $y = \bar{y}$

Symbolet $\bar{y}$ udtales ” y -streg”. Dette giver:

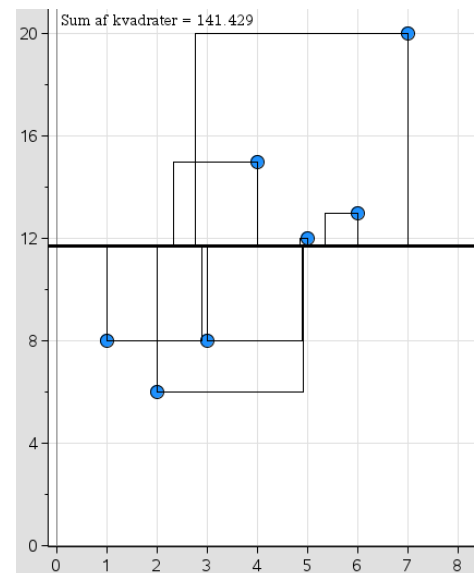
$$T = (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \dots + (y_n - \bar{y})^2$$

Denne linje repræsenterer den antagelse, at y -værdierne er helt uafhængige af x -værdierne.

Vi vurderer derfor de to kvadratsummer i forhold til hinanden:

$$r^2 = 1 - \frac{S}{T}$$

Eksemplet her giver: $r^2 = 1 - \frac{37,2857}{141,429} = 0,7364$



Perfekt sammenhæng

Hvis alle punkter ligger på den bedste rette linje, fås $y_i = f(x_i)$ for alle punkter.

Dermed fås residualerne til $r_i = y_i - f(x_i) = 0$ for alle punkter.

Dermed fås $S = r_1^2 + r_2^2 + \dots + r_n^2 = 0^2 + 0^2 + \dots + 0^2 = 0$.

Dermed fås: $r^2 = 1 - \frac{S}{T} = 1 - \frac{0}{T} = 1 - 0 = 1$

Hvis alle punkter ligger på ”bedste rette linje”, vil man altså få $r^2 = 1$.

Vi siger at der er en perfekt sammenhæng mellem x og y .

Ingen sammenhæng

Hvis y -værdierne er helt uafhængige af x -værdierne, vil variationen i y -værdier være helt uafhængig af x -værdierne. Det betyder at y -værdierne vil være tilfældigt fordelt rundt om linjen $y = \bar{y}$, som dermed bliver bedste rette linje.

Da S er residual-kvadrat-summen for bedste rette linje og T er residual-kvadrat-summen for gennemsnitslinjen $y = \bar{y}$, vil der gælde $S = T$ hvis $y = \bar{y}$ er den bedste rette linje. Dermed fås:

$$r^2 = 1 - \frac{S}{T} = 1 - \frac{T}{T} = 1 - 1 = 0$$

For den fuldstændige mangel på sammenhæng gælder altså $r^2 = 0$.