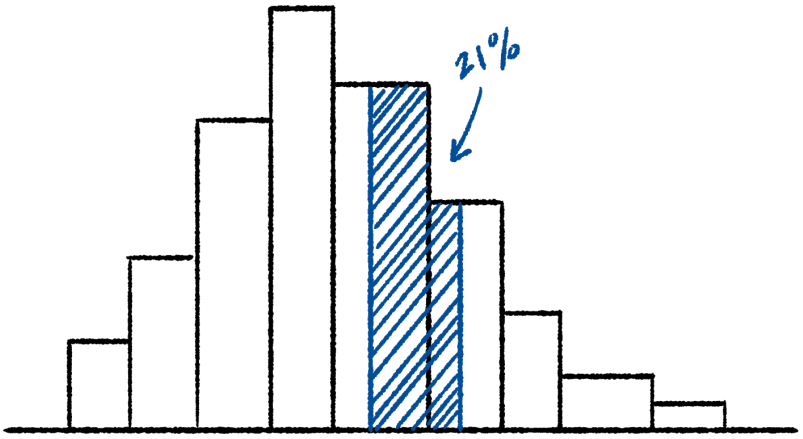
Statistik

Version 0.10
2. oktober 2020



Mike Vandal Auerbach
www.mathematicus.dk

Statistik
Version 0.10, 2020

Disse noter dækker kernestoffet i statistik på stx A- og B-niveau efter gymnasiereformen 2017.

Sandsynlighedsregning er ikke medtaget i disse noter, men noterne kan næsten læses uden viden om sandsynlighedsteoretiske begreber; den eneste undtagelse er de afsluttende bemærkninger om residualer og normalfordelingen, samt afsnittet om konfidensintervaller.

Noterne er skrevet vha. tekstformateringsprogrammet LaTeX, se www.tug.org og www.miktex.org. Figurer og diagrammer er fremstillet i *pgf/TikZ*, se www.ctan.org/pkg/pgf.

Disse og andre noter kan downloades fra www.mathematicus.dk.

Indhold

1 Hvad er statistik? 5
- 1.1 Repræsentativitet og systematiske fejl . . . 6
- 1.2 Skjulte variable . . . 7

2 Ugrupperet statistik 9
- 2.1 Variationsbredde, typetal og middelværdi . . . 10
- 2.2 Kvartiler . . . 11
- 2.3 Outliers . . . 13
- 2.4 Skævhed . . . 14
- 2.5 Spredning . . . 15
- 2.6 Grafiske afbildninger . . . 16

3 Grupperet statistik 19
- 3.1 Middelværdi og spredning . . . 19
- 3.2 Grafiske afbildninger . . . 20

4 Lineær regression 25
- 4.1 Forklaringsgrad . . . 28
- 4.2 Residualplot og residualspredning . . . 30
- 4.3 Konfidensintervaller . . . 31
- 4.4 Andre typer regression . . . 33

Bibliografi 35

Hvad er statistik?

1

Statistik er et område af matematikken, hvor man undersøger datasæt for at give en beskrivelse eller finde sammenhænge mellem de observationer, der er gjort. Det datasæt, man undersøger, kalder man *populationen*. Populationen i statistik er altså hele mængden af personer, genstande eller abstrakte objekter, man vil vide noget om.

Den størrelse, man måler, kalder man så en (statistisk) variabel. En variabel i statistik er ikke nødvendigvis et tal. Hvis populationen består af en bestemt gruppe mennesker (f.eks. borgere i Danmark), kan man måle deres højde eller vægt, og denne statistiske variabel kan beskrives med et tal – men man kan også registrere deres hårfarve, og den kan ikke beskrives med et tal. En variabel der beskrives med et tal, kaldes en *kvantitativ* variabel, mens en variabel som ikke er et tal (men f.eks. en egenskab), kaldes en *kvalitativ* variabel.

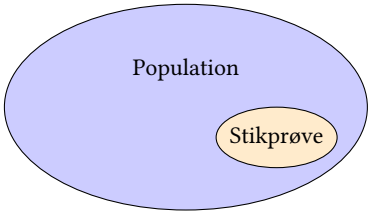
Statistik kan bruges til udelukkende at beskrive forskellige måledata, sådan at man kan skabe et overblik. Dette kalder man *deskriptiv statistik*.[1] Man forsøger altså at skabe et overblik over et datamateriale, der måske umiddelbart virker uoverskueligt.

[1] Jf. engelsk »describe«, *beskrive*.

Dette overblik kan skabes på flere måder, f.eks. kan man

- lave en tabel over datamaterialet og evt. gruppere nogle af dataene,
- udregne nogle deskriptorer, dvs. enkelte tal der beskriver datamaterialet, eller
- tegne diagrammer over dataene.

I mange tilfælde vil ens måledata dog nærmere bestå af en *stikprøve*. Hvis man vil lave en valgundersøgelse, er det umuligt at ringe til alle stemmeberettigede borgere for at høre, hvordan de vil stemme ved et kommende valg. Derfor laver man i stedet en stikprøve, hvor man måske spørger 1000 mennesker om deres politiske standpunkt. Her bliver man så nødt til at sørge for, at stikprøven er *repræsentativ*, dvs. at de resultater man kommer frem til ud fra stikprøven, nogenlunde repræsenterer hele befolkningen. Sammenhængen mellem population og stikprøve fremgår af de næste to eksempler (se også figur 1.1).



Figur 1.1: Stikprøven er den delmængde af populationen, som man undersøger.

Eksempel 1.1 Man vil undersøge vælgertilslutningen til et bestemt politisk parti. Populationen er i dette tilfælde hele befolkningen. Stikprøven vil så være et udsnit af befolkningen.

Eksempel 1.2 Et firma vil undersøge, om der er lige mange af hver farve vingummi i deres slikposer. Populationen er her alle de slikposer, firmaet producerer. En stikprøve kunne være et antal tilfældigt valgte poser fra lageret.

En sidste ting statistik bruges til, er at finde frem til matematiske modeller ud fra givne data. Måler man på to forskellige statistiske variable, kan man f.eks. udføre en regressionsanalyse for at se, om der er en matematisk sammenhæng. Når man analyserer data på denne måde, skal man være opmærksom på at en tilsyneladende sammenhæng også kan skyldes en tredje såkaldt »skjult« variabel (se afsnit 1.2 nedenfor).

1.1 Repræsentativitet og systematiske fejl

Når man vælger en stikprøve, er det vigtigt at stikprøven er *repræsentativ*. Dvs. at stikprøvens sammensætning skal være sådan, at stikprøvens karakteristika svarer til populationens som helhed. Hvis stikprøven ikke er repræsentativ, taler man om *systematisk fejl*.

Eksempel 1.3 En avis vil undersøge befolkningens holdning til digitalisering i det offentlige. De opretter derfor et spørgeskema på avisens hjemmeside.

Her er problemet, at man må forvente, at en del af de personer, der er kritiske over for digitalisering, ikke er på internettet i nævneværdig grad. Deres holdning vil derfor mangle i undersøgelsen. Stikprøven er derfor ikke repræsentativ.

Systematiske fejl forekommer altså, hvis nogle bestemte holdninger eller egenskaber er over- eller underrepræsenterede i stikprøven i forhold til populationen. Et ofte brugt eksempel på systematiske fejl i udvægelsen af en stikprøve er Literary Digests forudsigelse af, hvem der ville vinde det amerikanske præsidentvalg i 1936:

Eksempel 1.4 I 1936 forudsagde det amerikanske tidsskrift Literary Digest, at Alfred Landon ville vinde det amerikanske præsidentvalg med 57% af stemmerne. I stedet vandt den siddende præsident Franklin D. Roosevelt valget med 62% af stemmerne. Det fejlagtige resultat fik tidsskriftet på trods af, at de havde sendt spørgeskemaer ud til 10 mio. amerikanere, hvoraf 2,4 mio. svarede.[2]

Det der gik galt i undersøgelsen, var to ting. For det første havde tidsskriftet fået adresserne på de 10 mio. amerikanere der fik tilsendt et spørgeskema, fra bilklubber, telefonbøger og kartoteket over deres egne abonnenter. I 1936 var depressionen på sit højeste, dvs. de amerikanere der ejede en bil eller en telefon eller abonnerede på et tidsskrift, tilhørte med al sandsynlighed den rigeste del af befolkningen.

Det var dog sandsynligvis den anden fejl i designet der gav den falske forudsigelse.[5] Undersøgelsen baserede sig nemlig på de svar, som tidsskriftet modtog – dvs. det er muligt at der var en overvægt af folk med en bestemt holdning blandt dem som gav sig tid til at svare.

Som det fremgår af ovenstående eksempler, skal man altså overveje nøje hvordan man udvælger en stikprøve. I valgundersøgelser og lignende hvor man undersøger holdningen i en befolkning, vælger man typisk en stikprøve på omkring 1000 mennesker. Det er som regel tilstrækkeligt til at stikprøvens sammensætning kan afspejle befolkningen – men man skal være påpasselig med udvælgelsen, og man skal sørge for at få de mennesker repræsenteret der ikke gider svare på sådan en undersøgelse.

1.2 Skjulte variable

Målet for statistiske undersøgelser kan være at påvise en sammenhæng mellem forskellige størrelser. Her skal man passe på, at den sammenhæng man ser, ikke i virkeligheden skyldes noget helt tredje. I sådanne tilfælde taler man om *skjulte variable*.

Eksempel 1.5 Hvis man sammenholder salget af is med antallet af drukneulykker, finder man ud af at i perioder med højere issalg er der også flere drukneulykker. Man kunne derfor få den tanke at isspisning øger risikoen for at drukne.

Dette er selvfølgelig noget vås. Undersøger man i stedet begge variable (issalg og antal ulykker) i forhold til vejret, finder man hurtigt ud af at på varme dage sælges der flere is, og der er også flere folk der bader – hvilket fører til at antallet af drukneulykker øges.

Det er altså varmen der i dette tilfælde er den skjulte variabel som de andre afhænger af.

Ugrupperet statistik

2

Den ugrupperede statistik handler om at beskrive adskilte data. Man kan f.eks. forestille sig, at man har spurgt en gymnasieklasse på 25 elever, hvor mange gange de har været i biografen i løbet af det sidste år. Svarene kunne fordele sig som i tabel 2.1.

Dette giver ikke så meget overblik over tallene. Det første man kan gøre er derfor at sortere dem. Det er gjort i tabel 2.2.

Som man kan se i tabel 2.2, er der nogle af tallene der optræder flere gange. Det kan derfor være en god ide at lave en tabel over de forskellige tal og deres *hyppigheder* (dvs. hvor mange gange de optræder). Tabellen kunne se således ud:

Tabel 2.1: Antal biografbesøg (usorteret).

1	0	3	2	4
2	3	4	6	5
4	2	3	4	0
4	4	5	3	3
1	0	0	4	5

Tabel 2.2: Antal biografbesøg (sorteret).

0	0	0	0	1
1	2	2	2	3
3	3	3	3	4
4	4	4	4	4
4	5	5	5	6

Observation, x	Hyppighed, $h(x)$	Frekvens, $f(x)$	Kum. fr., $F(x)$
0	4	16%	16%
1	2	8%	24%
2	3	12%	36%
3	5	20%	56%
4	7	28%	84%
5	3	12%	96%
6	1	4%	100%
I alt	25	100%	

De første to kolonner i tabellen viser hhv. observationen, dvs. antallet af biografbesøg og hyppigheden. Derefter udregnes frekvensen, som er den relative hyppighed, dvs. hvor stor en brøkdel udgør denne observation af det samlede observationssæt.

Den sidste kolonne viser den *kumulerede frekvens* som angiver hvor stor en brøkdel af observationssættet der udgøres af observationerne *til og med* den observation man kigger på. Den kumulerede frekvens for 3 biografbesøg er f.eks. 56% fordi 56% har været i biografen højst 3 gange – eller med andre ord: hvis man tæller til og med 3 biografbesøg, har man talt 56% af eleverne.

I det følgende defineres de tre størrelser der knyttes til hver observation:

Definition 2.1

For et datasæt med n observationer $x_1, x_2, \ldots, x_n$ defineres følgende størrelser:

1. *Hyppigheden* $h(x)$ er antallet af gange observationen x optræder i datasættet.
2. *Frekvensen* $f(x)$ er hyppigheden i forhold til det samlede antal, dvs. $f(x) = \frac{h(x)}{n}$.
3. Den *kumulerede frekvens* $F(x)$ er summen af frekvenserne *til og med* frekvensen for den observation, man betragter, dvs.[1]

$$F(x) = \sum_{t \le x} f(t) \, .$$

[1] Sumtegnet $\sum_{t \le x}$ betyder i denne sammenhæng at man summerer over alle de værdier der er mindre end eller lig med x.

Frekvensen i tabellen ovenfor angiver altså, hvor stor en del af de adspurgte elever, der har været i biografen 0 gange, 1 gang, osv. Dette kan være nyttigt, hvis man skal sammenligne to klasser, der ikke har lige mange elever. Tallet angives ofte i procent, men det er ikke nødvendigt.

Den kumulerede frekvens angiver hvor mange elever der har været i biografen x gange *eller færre*. Den kumulerede frekvens for observationen $x = 2$ er 36%. Det betyder at 36% af eleverne har været i biografen 2 eller færre gange. Tallet kan findes ved at lægge frekvenserne for observationerne $x = 0$, $x = 1$ og $x = 2$ sammen:

$$F(2) = f(0) + f(1) + f(2) = 16\% + 8\% + 12\% = 36\% \, .$$

2.1 Variationsbredde, typetal og middelværdi

Selv om en tabel som den ovenstående kan lette overblikket over et datasæt, så er det nogle gange nemmere at sammenligne forskellige datasæt hvis man kan beskrive dem med nogle få tal, såkaldte *deskriptorer*.

F.eks. kan man se i tabellen at den mindste værdi er 0, og den største er 6. Herudfra kan man beregne den såkaldte *variationsbredde* som er forskellen på den største og den mindste værdi. I dette tilfælde er variationsbredden altså

$$x_{\max} - x_{\min} = 6 - 0 = 6 \, .$$

Typetallet er den observation, der optræder flest gange. I dette tilfælde er typetallet 4 – dvs. der er flest elever, der været i biografen 4 gange.

En deskriptor der kræver lidt mere udregning, er middelværdien som fortæller hvad den gennemsnitlige observation er. Middelværdien finder man ved at lægge alle observationerne sammen og dele med det samlede antal. For tallene i tabel 2.2 er middelværdien

$$\overline{x} = \frac{0 + 0 + 0 + 0 + 1 + \cdots + 5 + 5 + 5 + 6}{25} = 2{,}88 \, .$$

Når nu hyppighederne for de enkelte observationer allerede er opregnet (i tabellen ovenfor er f.eks. angivet, at observationen »2« optræder 5 gange)

kan man også anvende hyppighederne fra tabellen, så regnestykket bliver

$$\overline{x} = \frac{0 \cdot 4 + 1 \cdot 2 + 2 \cdot 3 + 3 \cdot 5 + 4 \cdot 7 + 5 \cdot 3 + 6 \cdot 1}{25} = 2{,}88 \ .$$

Resultatet ændrer sig selvfølgelig ikke.

Idet frekvenserne fås ved at dele hyppighederne med det samlede antal kunne man også dividere alle hyppighederne med 25 til at begynde med, hvorved man får[2]

$$\overline{x} = 0 \cdot 0{,}16 + 1 \cdot 0{,}08 + 2 \cdot 0{,}12 + 3 \cdot 0{,}20 + 4 \cdot 0{,}28 + 5 \cdot 0{,}12 + 6 \cdot 0{,}04 = 2{,}88 \ .$$

[2] Bemærk at frekvenserne her angives som decimaltal. Frekvensen for den første observation er f.eks. ikke 16, men 16%, hvilket jo er det samme som 0,16.

Definition 2.2

For et datasæt med n observationer $x_1, x_2, \ldots, x_n \in X$, defineres middelværdien $\overline{x}$ som[3]

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{x \in X} x \cdot h(x)}{n} = \sum_{x \in X} x \cdot f(x) \ .$$

[3] Sumtegnet $\sum_{i=1}^{n}$ angiver her at man summerer over de enkelte observationer fra nr. 1 til nr. n, mens $\sum_{x \in X}$ angiver at der summeres over alle *forskellige* observationer.

X er i denne sammenhæng mængden af mulige observationer.

Middelværdien angiver den gennemsnitlige observation. Når $\overline{x} = 2{,}88$ for datasættet ovenfor betyder det altså, at de 25 gymnasieelever gennemsnitligt har været i biografen 2,88 gange.

Man bruger somme tider symbolet μ for middelværdien af den bagvedliggende population. De 25 adspurgte gymnasieelever kan f.eks. ses som en stikprøve af den samlede population af alle danske gymnasieelever (eller danske unge mellem 15 og 20). I dette tilfælde bliver $\overline{x}$ et *estimat* for populationens middelværdi μ der er middelværdien for *alle* danske gymnasieelevers biografbesøg (og denne værdi er i sig selv ukendt).

2.2 Kvartiler

Middelværdien af et datasæt kan påvirkes stærkt hvis der optræder ekstreme værdier. Hvis der f.eks. havde været en enkelt elev der havde været i biografen 40 gange, ville middelværdien have været meget højere. Derfor giver det nogle gange mening at beskrive et datasæt vha. *medianen* som er den midterste observation.

Stiller man samtlige 25 tal fra tabel 2.2 op på en lang række, er medianen altså det midterste tal, dvs. tal nr. 13:[4]

median

0, 0, 0, 0, 1, 1, 2, 2, 2, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 6

[4] Hvis der er et lige antal observationer er medianen gennemsnittet af de to midterste observationer.

Medianen for elevernes biografbesøg er altså 3. Dvs. halvdelen af eleverne har været i biografen 3 gange eller færre. Den anden halvdel har været i biografen 3 gange eller mere. Det er vigtigt at bemærke, at dette intet har at gøre med middelværdien, og som man kan se er de to tal også forskellige.

Nogle gange kan man ønske lidt mere information end medianen alene giver. Medianen findes ved at dele observationssættet over i 2 halvdele. Mere information finder man, hvis man deler talsættet op i fire fjerdedele. Herved finder man de såkaldte *kvartiler*:

nedre kvartil median øvre kvartil

0, 0, 0, 0, 1, 1, | 2, 2, 2, 3, 3, 3, | 3, | 3, 4, 4, 4, 4, 4, | 4, 4, 5, 5, 5, 6

Nedre kvartil er medianen af den nederste halvdel. Da den nederste halvdel indeholder et lige antal observationer (12) bliver nedre kvartil gennemsnittet af de to midterste værdier (nr. 6 og 7). Nedre kvartil er derfor

$$Q_1 = \frac{1+2}{2} = 1{,}5 \ .$$

Medianen er det samme som før. Dvs. medianen er

$$m = 3 \ .$$

Øvre kvartil er medianen af den øverste halvdel. Her skal der igen tages gennemsnittet af to værdier, dvs.

$$Q_3 = \frac{4+4}{2} = 4 \ .$$

De tre tal Q_1, m og Q_3 kaldes tilsammen kvartil*sættet*.

Kvartilsættet for elevernes biografbesøg er $(1{,}5; 3; 4)$.

I stedet for betegnelserne »nedre kvartil, median og øvre kvartil« bruger man også nogle gange »første, andet og tredje kvartil«.

Definition 2.3

For et ugrupperet datasæt definerer man

- *Medianen* (eller *andet kvartil*) m, som er den midterste observation. Er der et lige antal observationer, er medianen gennemsnittet af de to midterste.
- *Nedre* (eller *første*) *kvartil* Q_1, som er medianen af den første halvdel af observationerne.
- *Øvre* (eller *tredje*)*kvartil* Q_3, som er medianen af den sidste halvdel af observationerne.

Talsættet $(Q_1; m; Q_3)$ bestående af alle kvartilerne, kaldes *kvartilsættet*. Talsættet $(x_{\min}; Q_1; m; Q_3; x_{\max})$ bestående af mindste observation, kvartilsættet og største observation kaldes *det udvidede kvartilsæt*.

At nedre kvartil for biografbesøgene er 1,5 viser, at en fjerdedel (25%) af eleverne var i biografen 1,5 gange eller færre, mens tre fjerdedele (75%) var i biografen 1,5 gange eller flere.

Øvre kvartil, $Q_3 = 2$ viser på samme måde, at tre fjerdedele af eleverne var i biografen 4 gange eller færre, mens en fjerdedel var i biografen 4 gange eller flere.

Hvis man vil have et fuldstændigt overblik over fordelingen, angiver man også sommetider det *udvidede kvartilsæt*, der som beskrevet består af kvartilsættet samt den mindst og den største observation. I dette tilfælde er det udvidede kvartilsæt

$$(0; 1{,}5; 3; 4; 6) ,$$

dvs. mindste observation er 0, nedre kvartil er 1,5, medianen er 3, øvre kvartil er 4 og den største observation er 6.

En yderligere størrelse, man taler om i denne forbindelse er *kvartilbredden*, der er afstanden mellem Q_1 og Q_3, dvs. i tilfældet med biografbesøgene er kvartilbredden

$$Q_3 - Q_1 = 4 - 1{,}5 = 2{,}5 .$$

Stikprøven over biografbesøgene kan samlet set altså beskrives med de følgende deskriptorer

Deskriptor		Værdi	
Minimum	$x_{\min}$	0	Udvidet kvartilsæt
Nedre kvartil	Q_1	1,5	
Median	m	3	
Øvre kvartil	Q_3	4	
Maksimum	$x_{\max}$	6	
Middelværdi	$\overline{x}$	2,88	
Typetal		4	
Kvartilbredde	$Q_3 - Q_1$	2,5	
Variationsbredde	$x_{\max} - x_{\min}$	6	

2.3 Outliers

Man kan forestille sig at man har spurgt 10 andre gymnasieelever om hvor mange gange de har været i biografen og fået svarene i tabel 2.3. Dette observationssæt har middelværdien

$$\overline{x} = 2{,}8 ,$$

og det udvidede kvartilsæt er

$$(0; 2; 2{,}5; 4; 8) .$$

Tabel 2.3: Ny undersøgelse af biografbesøg

4	3	2	0	2
3	2	4	8	0

Kigger man på datasættet, ser man, at en enkelt observation (den ene elev der har været i biografen 8 gange) er ret stor i forhold til de andre. I dette tilfælde kan der være tale om en såkaldt *outlier*, dvs. en observation der ligger langt fra den typiske observation. Den følgende definition viser hvor stor eller lille en værdi skal være for at man kalder den en outlier.

Definition 2.4

En observation x i et observationssæt kaldes en *outlier* hvis den ligger mere en halvanden kvartilbredde under nedre kvartil eller mere end halvanden kvartilbredde over øvre kvartil.

Dvs. x er en outlier hvis

$$x < Q_1 - 1{,}5 \cdot (Q_3 - Q_1) \quad \text{eller} \quad x > Q_3 + 1{,}5 \cdot (Q_3 - Q_1) \, .$$

I tilfældet ovenfor er kvartilbredden

$$Q_3 - Q_1 = 4 - 2 = 2 \, .$$

Dvs. halvanden kvartilbredde under første kvartil hhv. over trejde kvartiel svarer til

$$\begin{aligned} Q_1 - 1{,}5 \cdot (Q_3 - Q_1) &= 2 - 1{,}5 \cdot 2 = -1 \\ Q_3 + 1{,}5 \cdot (Q_3 - Q_1) &= 4 + 1{,}5 \cdot 2 = 7 \end{aligned}$$

Idet observationen 8 er større end 7 er der altså tale om en outlier. På samme måde vil alle tal under -1 i princippet være outliers (men i dette tilfælde er det umuligt at få negative tal som observation).

2.4 Skævhed

For de første 25 adspurgte elever fandt man en middelværdi på 2,88 og en median på 3. En sådan fordeling hvor middelværdien er mindre end medianen, kaldes en *venstreskæv* fordeling.

I det foregående afsnit blev der kigget på et datasæt (10 elever) hvor middelværdien var 2,8, og medianen var 2,5. Her er middelværdien altså større end medianen. Denne fordeling er derfor *højreskæv*.

Definition 2.5

Et datasæt har en

- *venstreskæv fordeling* hvis middeltallet er mindre end medianen, $\overline{x} < m$,
- *ikke-skæv fordeling* hvis middeltallet er lig med medianen $\overline{x} = m$, eller en
- *højreskæv fordeling* hvis middeltallet er større end medianen, $\overline{x} > m$.

Hvis fordelingen er ventre- eller højreskæv vil det kunne ses på grafiske afbildninger af fordelingen, jf. afsnit 2.6.

2.5 Spredning

Spredningen er et mål for, hvor langt observationerne i gennemsnit ligger fra middelværdien. Spredningen af en population er defineret som

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}} = \sqrt{\frac{\sum_{x \in X}(x - \mu)^2 \cdot h(x)}{n}} \,. \tag{2.1}$$

Problemet er her, at denne størrelse ikke kan beregnes ud fra en stikprøve. Man kan nemlig ikke bestemme en populations sande middelværdi μ ud fra en stikprøve, man kan kun estimere den ved stikprøvens middelværdi $\overline{x}$. Bruger man blot $\overline{x}$ i stedet for μ vil man dog konsekvent vurdere spredningen til at være for lille.

Eksempel 2.6 På et tilfældigt gymnasium er middelværdien for højden af drengene 173 cm. Nu tager man en stikprøve for at estimere denne middelværdi. Man måler højderne af 3 drenge og får hhv. 168, 176 og 181 cm. Middelværdien af disse højder er så

$$\overline{x} = \frac{168 + 176 + 181}{3} = 175 \,.$$

Denne størrelse er et estimat for den sande middelværdi på 173 cm.

Bruger man den sande middelværdi til at beregne spredningen, får man

$$\sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}} = \sqrt{\frac{(168 - 173)^2 + (176 - 173)^2 + (181 - 173)^2}{3}} = 5{,}72 \,.$$

Kender man ikke den sande middelværdi, bliver man nødt til at estimere den og i stedet bruge $\overline{x}$, men så får man

$$\sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}} = \sqrt{\frac{(168 - 175)^2 + (176 - 175)^2 + (181 - 175)^2}{3}} = 5{,}35 \,.$$

Man får altså et for lille estimat for σ, hvis man bruger $\overline{x}$ som estimat for μ.

Hvis man i stedet for at dividere med 3 i formlen ovenfor dividerer med 1 mindre (altså 2) får man:

$$\sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} = \sqrt{\frac{(168 - 175)^2 + (176 - 175)^2 + (181 - 175)^2}{2}} = 6{,}56 \,.$$

Dette tal er et for højt estimat for den virkelige spredning; men det er altid at foretrække frem for et for lavt estimat.

Fordi man får et for lille estimat, hvis man anvender $\overline{x}$ i stedet for μ i formlen (2.1), dividerer man altså med $n - 1$ i stedet for med n fordi man så får et bedre estimat for spredningen:

Definition 2.7

For en stikprøve bestående af elementerne $x_1, x_2, \ldots, x_n$, defineres *stikprøvespredningen* som tallet

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} = \sqrt{\frac{\sum_{x \in X}(x - \overline{x})^2 \cdot h(x)}{n-1}} .$$

Kigger man igen på undersøgelsen af de 25 elevers biografbesøg, finder man her en stikprøvespredning på

$$\begin{aligned} s &= \sqrt{\frac{\sum_{x \in X}(x - \overline{x})^2 \cdot h(x)}{n-1}} \\ &= \sqrt{\frac{(0 - 2{,}88)^2 \cdot 4 + (1 - 2{,}88)^2 \cdot 2 + \cdots + (6 - 2{,}88)^2 \cdot 6}{25 - 1}} \\ &= 1{,}76 . \end{aligned}$$

For de 10 elever der blev spurgt efterfølgende bliver stikprøvespredningen $s = 2{,}30$. Spredningen bliver større her fordi dette datasæt indeholder en outlier.

2.6 Grafiske afbildninger

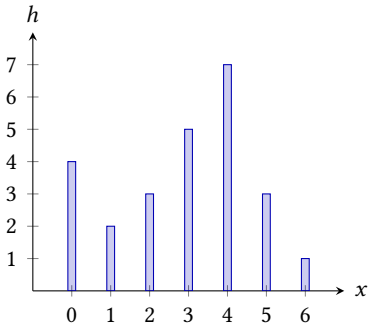
I dette afsnit gennemgås 3 forskellige afbildninger til at få overblik over et ugrupperet datasæt:

- Et stolpediagram, som er godt til at få overblik over et enkelt datasæt.
- Et trappediagram.
- Et boksplot, som er godt til sammenligning af forskellige datasæt.

Stolpediagram

Undersøgelsen om de 25 gymnasieelevers biografbesøg gav følgende tabel:

Observation, x	Hyppighed, $h(x)$	Frekvens, $f(x)$	Kum. fr., $F(x)$
0	4	16%	16%
1	2	8%	24%
2	3	12%	36%
3	5	20%	56%
4	7	28%	84%
5	3	12%	96%
6	1	4%	100%
I alt	25	100%	



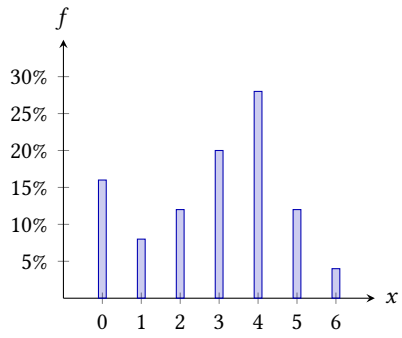
Figur 2.4: Elevernes biografbesøg som stolpediagram med hyppigheden op ad andenaksen.

Herudfra kan man tegne et stolpediagram. Det består i, at man afsætter de enkelte observationer ud ad en førsteakse, og derefter tegner stolper, hvis højde er lig hyppigheden eller frekvensen.

På figur 2.4 ses et stolpediagram, hvor højden på søjlerne er givet ved hyppigheden. Figur 2.5 er det samme stolpediagram, men her er højderne givet ved frekvensen. Som man kan se, er de to stolpediagrammer ens, bortset fra inddelingen på andenaksen.

Hvis man blot vil have et overblik over datamaterialet er det udmærket at anvende hyppighederne. Vil man imidlertid sammenligne to datasæt, er det en god ide at anvende frekvenserne. Det gør det nemmere at sammenligne, især hvis de to datasæt man sammenligner ikke indeholder lige mange observationer. Det kunne f.eks. være hvis man sammenlignede to klasser med et forskelligt antal elever.

Fordelingen af observationerne i dette datasæt er i øvrigt, som tidligere omtalt, venstreskæv. Kigger man på stolpediagrammet, kan man se at »vægten« i diagrammet ser ud til at være forskudt mod venstre. Havde fordelingen i stedet været højreskæv, ville stolperne have set ud til at skubbe sig mere mod højre.

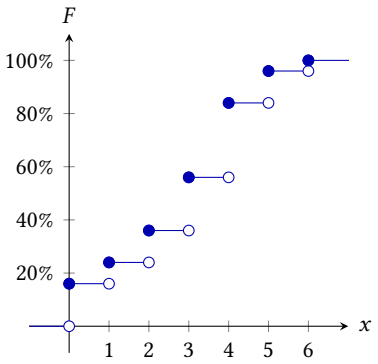


Figur 2.5: Elevernes biografbesøg som stolpediagram med frekvensen op ad andenaksen.

Trappediagram

Et trappediagram er et diagram over de kumulerede frekvenser. Man afsætter den kumulerede frekvens ud for observationen og bevæger sig derefter vandret, ind til man kommer til den næste observation, hvor man springer op til den næste kumulerede frekvens. Herved fremkommer et diagram, der ligner en trappe, se figur 2.6.

Det er muligt at anvende et trappediagram til at sammenligne forskellige observationssæt, men boksplottet, som introduceres nedenfor, er et mere overskueligt redskab til sammenligninger.



Figur 2.6: Trappediagram over elevernes biografbesøg.

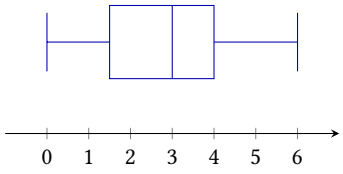
Boksplot

Et boksplot er et diagram, som laves udelukkende ud fra det udvidede kvartilsæt. Herved kasserer man en masse information, men til gengæld får man et diagram, der på meget overskuelig vis kan vise, hvordan tallene fordeler sig.

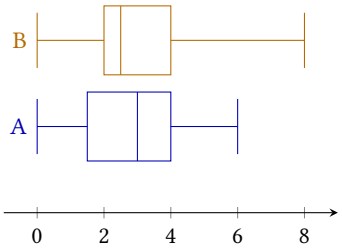
Et boksplot over elevernes biografbesøg kan ses på figur 2.7. Der tegnes lodrette streger ud for den mindste observation (0), nedre kvartil (1,5), medianen (3), øvre kvartil (4) og den største observation (6). Dernæst forbindes de lodrette streger som på figuren.

Selve boksen kommer derved til at indeholde den midterste halvdel af observationerne, mens de vandrette linjestykker i enderne viser biografbesøgene for den nederste hhv. den øverste fjerdedel af klassen.

Når man tegner boksplot over fordelinger, bliver det nemt at sammenligne. Hvis man har det udvidede kvartilsæt for begge de to undersøgelser (de første 25 elever (A) og de efterfølgende 10 (B)) kunne man for eksempel have følgende tabel:



Figur 2.7: Boksplot over elevernes biografbesøg.



Figur 2.8: Sammenligning af biografbesøg vha. boksplots.

Datasæt	Mindste værdi	Q_1	m	Q_3	Største værdi
A	0	1,5	3	4	6
B	0	2	2,5	4	8

Kigger man blot på tallene, kan det være svært at se forskel på de to klasser. Tegner man derimod et boksplot over begge fordelinger i samme diagram (se figur 2.8) bliver det pludselig nemmere at sammenligne.

F.eks. kan man se, at selv om B har den elev, der har det største antal biografbesøg, så har de nederste 50% af eleverne i B været lidt sjældnere i biografen end de nederste 50% i A. Den midterste halvdel i B ligger også tættere end i A, dvs. kvartilbredden er mindre her (B har dog stadig større stikprøvespredning end A, jf. foregående afsnit).

Grupperet statistik

3

Man taler om grupperet statistik når de data man ser på, er arrangeret i intervaller. Dette kan f.eks. ske hvis der er tale om et meget stort datasæt, eller hvis man måler data med mange decimaler. Her vil man typisk have et meget stort antal forskellige observationer, og det giver derfor mening at gruppere dem i intervaller. Typisk vil intervallerne grænse op til hinanden; det er dog ikke strengt nødvendigt. Men intervallerne skal altid være adskilte – dvs. den samme værdi må ikke høre til flere forskellige intervaller.

I tabel 3.1 ses en stikprøve fra et firma der producerer poser med sukker. Tabellen viser en kontrolvejning af en stikprøve på 500 poser. Der er her tale om så mange tal, at det ikke kan betale sig at liste alle de vejninger der ligger til grund for tabellen. Derfor har man i stedet opdelt de forskellige vægte i intervaller.

Tabel 3.1: Stikprøve af vægten af sukkerposer.

Interval (i gram)	Antal
800-850	11
850-900	17
900-950	53
950-1000	208
1000-1050	125
1050-1100	86

Når man ser på tabellen, kan man ikke umiddelbart se om en vægt på 850 g skal regnes med i det første eller det andet interval. Der kan derfor være en god ide i at bruge intervalnotation for at beskrive om de vægte der ligger på grænserne skal regnes med det ene eller det andet sted.

Man kan også se at hyppighederne i dette tilfælde er nogle temmeligt store tal. De tilhørende frekvenser ser således ud:

Interval	Hyppighed, h	Frekvens, f	Kumuleret frekv., F
$[800; 850[$	11	2,2%	2,2%
$[850; 900[$	17	3,4%	5,6%
$[900; 950[$	53	10,6%	16,2%
$[950; 1000[$	208	41,6%	57,8%
$[1000; 1050[$	125	25,0%	82,8%
$[1050; 1100[$	86	17,2%	100,0%
I alt	500	100,0%	

3.1 Middelværdi og spredning

Middelværdien og stikprøvespredningen kan ikke beregnes på samme måde som det blev gjort for den ugrupperede statistik. Det kan den ikke fordi man ikke ved hvordan vægtene fordeler sig i de enkelte intervaller. De rå data der ligger til grund for tabellen, har man ikke.

Det man så gør i stedet, er at antage at vægtene ligger jævnt fordelt i intervallerne. Så kan man nemlig bruge intervallernes midtpunkter som enkelte observationer.

Definition 3.1

For et datasæt, der er grupperet i n intervaller, $[a_1; b_1[$, $[a_2; b_2[$, …, $[a_n; b_n[$, som i alt indeholder N observationer, er middelværdien

$$\overline{x} = \frac{\sum_{i=1}^{n} m_i \cdot h_i}{N} = \sum_{i=1}^{n} m_i \cdot f_i \ ,$$

og stikprøvespredningen er

$$s = \sqrt{\frac{\sum_{i=1}^{n} (m_i - \overline{x})^2 \cdot h_i}{N - 1}} \ .$$

h_i er intervallets hyppighed, f_i er frekvensen, og m_i er intervallets midtpunkt, $m_i = \frac{a_i + b_i}{2}$.

For at beregne middelværdien for datasættet ovenfor tilføjer man en kolonne med intervalmidtpunktet:

Interval	Intervalmidtpunkt, m	Frekvens, f
[800; 850[	825	2,2%
[850; 900[	875	3,4%
[900; 950[	925	10,6%
[950; 1000[	975	41,6%
[1000; 1050[	1025	25,0%
[1050; 1100[	1075	17,2%

Middelværdien er så

$$\overline{x} = 825 \cdot 0{,}022 + 875 \cdot 0{,}034 + \cdots + 1075 \cdot 0{,}172 = 992{,}7 \ .$$

Gennemsnitsvægten i tabellen er altså på 992,7 g.

3.2 Grafiske afbildninger

I dette afsnit gennemgås tre afbildninger for grupperede data:

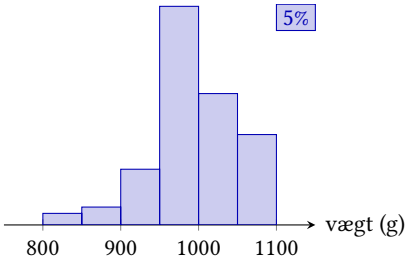
- Histogrammer, som svarer til de stolpediagrammer, man kan tegne for ugrupperede data.
- Sumkurver, som svarer til trappediagrammer og bl.a. kan anvendes til at bestemme kvartilsættet.
- Boksplots, som er fuldstændigt magen til den tilsvarende afbildning for ugrupperede data.

Histogrammer

I et histogram afbilder man intervallernes frekvens som søjler. I modsætning til et stolpediagram kan man dog ikke lade søjlernes højde være afgjort af frekvensen. Hvis man gjorde det, ville brede intervaller få mere vægt i afbildningen end smalle intervaller, og det duer ikke.

I stedet lader man søjlernes *areal* svare til frekvensen, se figur 3.2.

Når man bruger arealet til at angive frekvensen er det nødvendigt at angive, hvor stort et areal, der svarer til en bestemt procentdel. Dette er illustreret på figuren, hvor rektanglet i øverste højre hjørne viser, hvor stort et areal, der svarer til 5%.

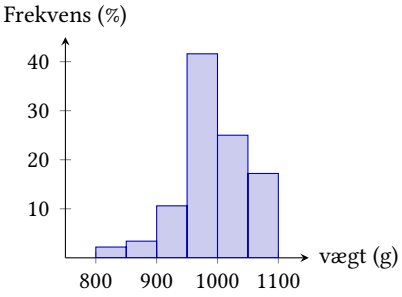


Figur 3.2: Histogram over vægtfordelingen. Arealet angiver frekvensen.

Da det er arealet, der angiver frekvensen, er der ikke brug for en andenakse, og den udelades derfor som regel.

Hvis det forholder sig sådan, at alle intervallerne er lige brede, kan man dog lade højden svare til frekvensen. Mange CAS-værktøjer arbejder på denne måde. Men det er vigtigt at huske, at så *skal* intervallerne være lige brede.

I dette tilfælde er intervallerne faktisk lige brede, dvs. det er tilladt at tegne histogrammet som på figur 3.3.



Figur 3.3: Histogram over vægtfordelingen. Højden angiver frekvensen.

For et histogram er det altså vigtigt at huske, at

> intervalfrekvensen er *arealet* af den pågældende søjle – med mindre alle intervallerne er *lige brede*.

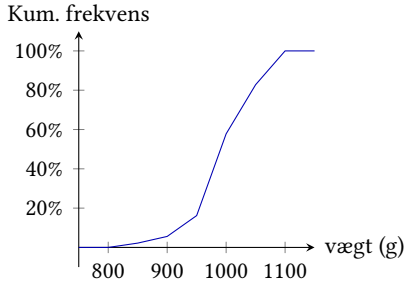
Sumkurver

En sumkurve tegnes på baggrund af de kumulerede frekvenser. Kurven illustrerer, hvor mange procent af datasættet der ligger under en bestemt værdi. Da kurven skal vise hvor mange procent der ligger *under* en given værdi, er det de *højre* intervalendepunkter der afsættes ud ad førsteaksen.

Man kan derfor tilføje en kolonne med intervalendepunkter til tabellen ovenfor:

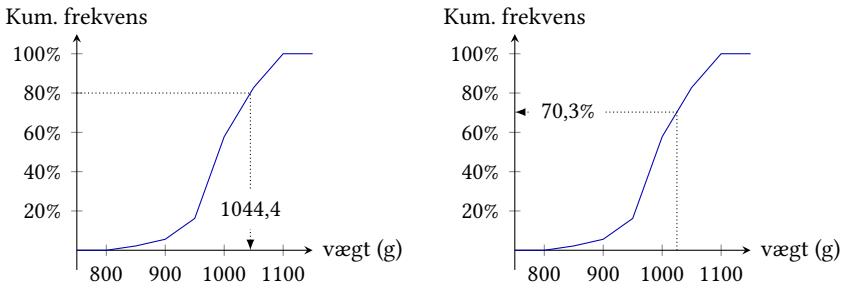
Interval	Højre intervalendepunkt	Kumuleret frekvens
[800; 850[	850	2,2%
[850; 900[	900	5,6%
[900; 950[	950	16,2%
[950; 1000[	1000	57,8%
[1000; 1050[	1050	82,8%
[1050; 1100[	1100	100,0%

Sumkurven tegnes derefter ved at afsætte de højre intervalendepunkter ud af førsteaksen og de kumulerede frekvenser ud ad andenaksen. De afsatte punkter forbindes derefter med linjestykker.



Figur 3.4: Sumkurve over vægtfordelingen.

Da sumkurven angiver, hvor mange procent der ligger under en given værdi, kan den bruges til at undersøge, f.eks. hvor mange procent af poserne der



(a) Aflæsning af 0,8-fraktilen

(b) Hvor mange poser vejer under 1025 g?

Figur 3.5: På figuren til venstre aflæses 0,8-fraktilen. Tallet viser at 80% af poserne vejer under 1044,4 g.
Til højre aflæses ud for 1025 på førsteaksen. Tallet her viser, at 70,3% af poserne vejer 1025 g eller mindre mindre.

vejer under 1025 g, eller hvad den maksimale vægt er for de 80% af poserne der vejer mindst. Det sidste kalder man *0,8-fraktilen* eller *80%-fraktilen*. Der gælder følgende definition:

Definition 3.2

For et statistisk talmateriale betegner p-fraktilen den værdi i observationssættet, der har en kumuleret frekvens på p.

På figur 3.5 kan man se en aflæsning af 0,8-fraktilen. Man går ud fra 80% på andenaksen og aflæser den tilsvarende værdi på førsteaksen. Det aflæste tal 1044,4 viser at 80% af poserne i stikprøven vejer 1044,4 g eller mindre. Tilsvarende vejer 20% af poserne derfor 1044,4 g eller mere.

Figuren viser også en aflæsning af hvilken fraktil der svarer til en vægt på 1025 g. Her går man ud fra 1025 på førsteaksen og aflæser det tilsvarende tal på andenaksen. Det aflæste tal er 70,3%. Det betyder at 70,3% vejer under 1044,4 g, altså vil 29,7% af poserne veje mere end 1025 g.

Ud fra sumkurven definerer man også kvartilsættet.

Definition 3.3

På en sumkurve for et givet statistisk materiale kan man aflæse kvartilsættet $(Q_1; Q_2; Q_3)$.

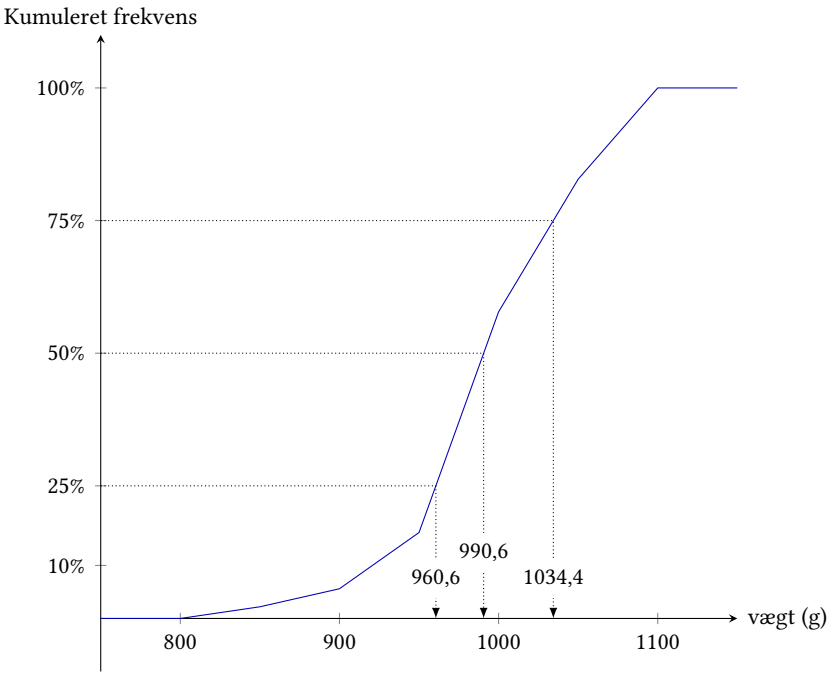
1. *Nedre kvartil*, Q_1 er 25%-fraktilen.
2. *Medianen*, Q_2 er 50%-fraktilen.
3. *Øvre kvartil*, Q_3 er 75%-fraktilen.

På figur 3.6 ses en aflæsning af kvartilsættet. Man går ud fra hhv. 25%, 50% og 75% på andenaksen og aflæser de tilsvarende værdier på førsteaksen. Her ser man, at kvartilsættet er

$$(960{,}6; 990{,}6; 1034{,}4)\ .$$

Disse tal viser, at

- 25% af poserne vejer 960,6 g eller mindre,
- 50% af poserne vejer 990,6 g eller mindre, og
- 75% af poserne vejer 1034,4 g eller mindre.



Figur 3.6: Aflæsning af kvartilsættet på sumkurven for vægtfordelingen.

Boksplot

At tegne et boksplot for et grupperet datasæt er ikke anderledes end at tegne et boksplot for et ugrupperet datasæt. Der hvor de to ting adskiller sig er i, hvordan man skal finde kvartilsættet. Når først det er fundet så er fremgangsmåden fuldstændigt den samme.

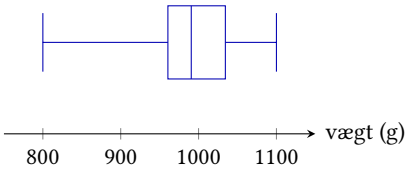
For vægtfordelingen der er set på ovenfor, var kvartilsættet

$$(960{,}6;\ 990{,}6;\ 1034{,}4)\ .$$

Den mindste værdi var 800 og den største 1100. Det udvidede kvartilsæt er altså

$$(800;\ 960{,}6;\ 990{,}6;\ 1034{,}4;\ 1100)\ ,$$

og et boksplot over denne fordeling vil derfor se ud som på figur 3.7.



Figur 3.7: Boksplot over vægtfordelingen.

Lineær regression

4

Hvis man har målt en række sammenhørende værdier af to variable hvor den ene variabel afhænger af den anden, kan man i nogle tilfælde opstille en model over sammenhængen mellem de to variable.

Når man har et datasæt $(x_1; y_1), (x_2; y_2), \ldots, (x_n; y_n)$ kan man prøve at modellere sammenhængen med en funktion f, sådan at grafen for f kommer til at ligge så tæt på punkterne som muligt. Da der altid er måleusikkerhed forbundet med målinger af virkelige data, vil en sådan graf aldrig gå gennem alle punkterne. Afvigelsen mellem modellens y-værdi $\hat{y}_i = f(x_i)$ (også kaldet den *estimerede* værdi) og den målte y-værdi y_i kalder man *residualet*. For punktet $(x_i; y_i)$ er residualet

$$r_i = y_i - \hat{y}_i \ .$$

En måde at bestemme funktionen på består i at finde den funktion f hvor residualerne samlet set bliver så små som muligt. Et mål for den samlede afvigelse er givet ved kvadratsummen af residualerne[1]

$$SSE = r_1^2 + r_2^2 + \cdots + r_n^2 \ .$$

[1] *SSE* er en forkortelse for »sum of squares of errors of prediction«, altså kvadratsummen af fejlene/afvigelserne, dvs. residualerne.

Denne størrelse ønsker man så at minimere. Fordi man ser på kvadraterne af residualerne, kaldes metoden også for *mindste kvadraters metode*.

Hvis funktionen f nævnt ovenfor er en lineær funktion, $f(x) = ax + b$, finder man ved hjælp af metoden den rette linje der passer bedst på de n punkter $(x_1; y_1), (x_2; y_2), \ldots, (x_n; y_n)$. Residualerne har så formen

$$r_i = y_i - (ax_i + b) \ .$$

Kvadratsummen *SSE* af residualerne bliver så

$$SSE = r_1^2 + r_2^2 + \cdots + r_n^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2 \ . \tag{4.1}$$

Den rette linje $y = ax + b$ vi søger, er så den linje hvor kvadratsummen *SSE* er mindst mulig.

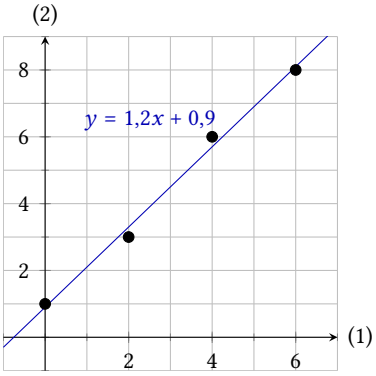
Det viser sig at tallene a og b for denne linje kan beregnes på følgende måde:

Tabel 4.1: Sammenhørende målte værdier af x og y.

x	y
0	1
2	3
4	6
6	8

Tabel 4.2: x, y, $x \cdot y$ og x^2. Den nederste række viser gennemsnittene.

x	y	$x \cdot y$	x^2
0	1	0	0
2	3	6	4
4	6	24	16
6	8	48	36
$\overline{x}$	$\overline{y}$	$\overline{x \cdot y}$	$\overline{x^2}$
3	4,5	19,5	14



Figur 4.3: Den bedste rette linje gennem de 4 punkter.

Sætning 4.1

For datapunkterne $(x_1; y_1), (x_2; y_2), \ldots, (x_n; y_n)$ findes den bedste rette linje $y = ax + b$, hvor

$$a = \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2},$$

og

$$b = \overline{y} - a \cdot \overline{x}.$$

I denne sætning er $\overline{x}$ gennemsnittet af x-værdierne, $\overline{y}$ er gennemsnittet af y-værdierne, $\overline{x \cdot y}$ er gennemsnittet af $x \cdot y$, osv.

Eksempel 4.2 Tabel 4.1 viser sammenhørende værdier af den uafhængige variabel x og den uafhængige variabel y. For at kunne bruge formlerne skal man bruge en række gennemsnit. Disse er beregnet i tabel 4.2.

Man kan nu beregne

$$a = \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{19{,}5 - 3 \cdot 4{,}5}{14 - 3^2} = 1{,}2.$$

og

$$b = \overline{y} - a \cdot \overline{x} = 4{,}5 - 1{,}2 \cdot 3 = 0{,}9.$$

Den bedste rette linje har derfor ligningen

$$y = 1{,}2x + 0{,}9.$$

Punkterne og linjen kan ses på figur 4.3.

Som det måske fremgår af eksemplet, kan det hurtigt blive besværligt at beregne den bedste rette linje vha. formlerne i sætning 4.1. Heldigvis er metoden indbygget i de fleste CAS-værktøjer, så man kan nøjes med at indtaste punkterne og få værktøjet til at beregne ligningen.

Bevis for formlerne

For at kunne bevise formlerne i sætning 4.1 er det nødvendigt at vide hvornår en kvadratsum antager sit minimum:

Sætning 4.3

Kvadratsummen $q(c) = \sum_{i=1}^{n}(z_i - c)^2$ antager sit minimum, når $c = \overline{z}$.

Bevis

Man har kvadratsummen

$$q(c) = \sum_{i=1}^{n}(z_i - c)^2.$$

$q(c)$ er altså en funktion af c, som er givet ved

$$q(c) = (z_1 - c)^2 + (z_2 - c)^2 + \cdots + (z_n - c)^2.$$

Man kan nu omskrive udtrykket for $q(c)$ på følgende måde,[2]

$$\begin{aligned} q(c) &= \sum_{i=1}^{n}(z_i - c)^2 \\ &= \sum_{i=1}^{n}\left(z_i^2 + c^2 - 2z_i c\right) \\ &= nc^2 - \left(2\sum_{i=1}^{n} z_i\right)c + \sum_{i=1}^{n} z_i^2 \\ &= nc^2 - (2n\overline{z})c + \sum_{i=1}^{n} z_i^2 \ . \end{aligned}$$

[2] Undervejs benyttes, at $\sum_{i=1}^{n} c^2 = nc^2$, og at $\sum_{i=1}^{n} z_i = n\overline{z}$.

$q(c)$ er altså et andengradspolynomium i c. Et andengradspolynomium $y = Ac^2 + Bc + C$ hvor $A > 0$, antager sit minimum i toppunktet, og her er $c = -\frac{B}{2A}$.

I $q(c) = nc^2 - (2n\overline{z})c + \sum_{i=1}^{n} z_i^2$ er koefficienterne

$$A = n \ , \quad B = -2n\overline{z} \quad \text{og} \quad C = \sum_{i=1}^{n} z_i^2 \ .$$

$q(c)$ antager sit derfor sit minimum, når

$$c = -\frac{-2n\overline{z}}{2n} = \overline{z} \ . \qquad \blacksquare$$

Af sætning 4.3 får man umiddelbart at S er mindst når[3]

$$b = \overline{y - ax} = \overline{y} - a \cdot \overline{x} \ . \tag{4.2}$$

[3] Man sætter $z_i = y_i - ax_i$ og $c = b$ i udtrykket i sætningen.

Man har altså nu et udtryk for den rette linjes skæring med andenaksen. For at finde et udtryk for hældningskoefficienten a indsættes udtrykket (4.2) i udtrykket for kvadratsummen SSE fra (4.1):

$$\begin{aligned} SSE &= \sum_{i=1}^{n}(y_i - ax_i - b)^2 \\ &= \sum_{i=1}^{n}(y_i - ax_i - \overline{y} + a\overline{x})^2 \\ &= \sum_{i=1}^{n}((y_i - \overline{y}) - a(x_i - \overline{x}))^2 \\ &= \left(\sum_{i=1}^{n}(x_i - \overline{x})^2\right)a^2 - \left(2\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})\right)a + \sum_{i=1}^{n}(y_i - \overline{y})^2 \ . \end{aligned}$$

Dette er et andengradspolynomium i a, som antager sit minimum, hvor

$$a = -\frac{-2\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{2\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \ .$$

Man kan vha. en del udregning vise, at denne brøk kan skrives som

$$a = \frac{\sum_{i=1}^{n} x_i y_i - n \cdot \overline{x} \cdot \overline{y}}{\sum_{i=1}^{n} x_i^2 - n \cdot \overline{x}^2} \ ,$$

hvilket igen kan reduceres så man får formlerne i sætning 4.1.

4.1 Forklaringsgrad

Man kan altid beregne den bedste rette linje vha. formlerne i sætning 4.1, men det er selvfølgelig ingen garanti for at punkterne rent faktisk ligger på nogenlunde på en linje. Man har derfor fundet på et mål for hvor godt den fundne linje passer på datapunkterne, den såkaldte *forklaringsgrad*.

Hvis der slet ingen sammenhæng er mellem x og y i datapunkterne, så kan man forvente at alle y-værdierne varierer omkring gennemsnittet $\overline{y}$ uafhængigt af x-værdien. Dette giver en kvadratsum på forskellen mellem den målte y-værdi og den forventede y-værdi (som her er $\overline{y}$) på

$$S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2 \ .$$

S_{yy} er altså kvadratsummen på afvigelserne når man antager at der slet ingen sammenhæng er mellem x og y.

Men i virkeligheden forventer man en sammenhæng mellem x og y, og i dette tilfælde kan afvigelserne beskrives ved kvadratsummen SSE som er kvadratsummen på afvigelserne når man modellerer de givne data med en lineær funktion. Denne vil være mindre end S_{yy}. Man beregner så forklaringsgraden som

$$R^2 = \frac{S_{yy} - SSE}{S_{yy}} \ .$$

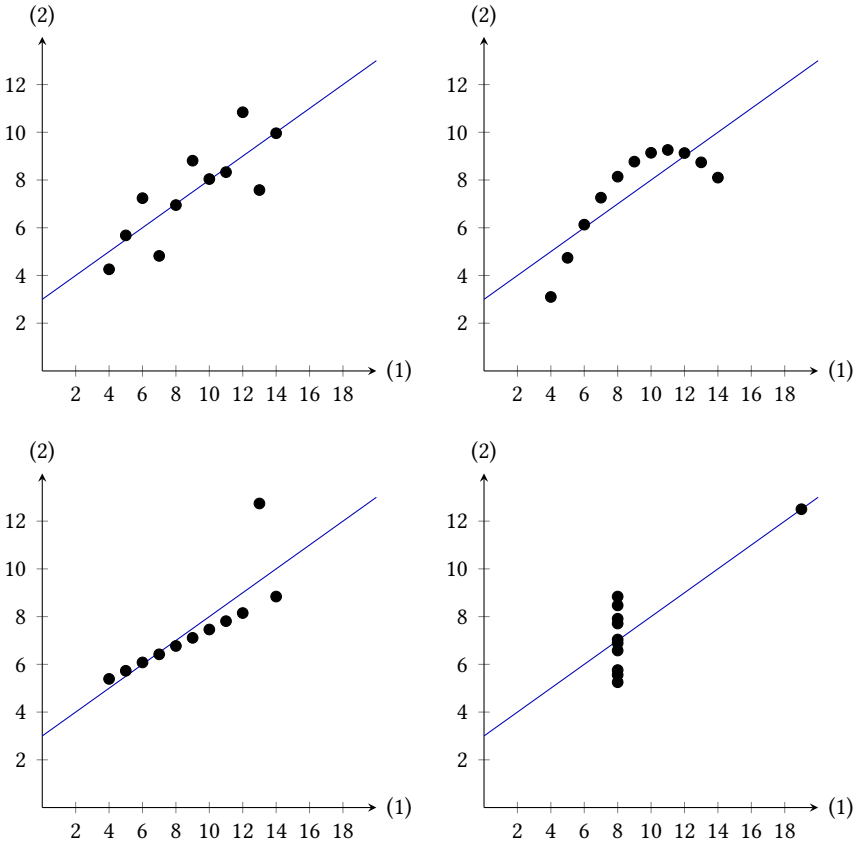
Tallet R^2 viser altså hvor mange procent SSE er mindre end S_{yy}. Hvis SSE er meget lille i forhold til S_{yy} vil dette tal ligge tæt på 1, mens det vil ligge tæt på 0 når SSE er næsten lige så stor som S_{yy}, dvs. når den lineære funktion ikke ligger meget tættere på punkterne end hvis man blot så på et rent gennemsnit af y-værdierne.

Forklaringsgraden er et udmærket mål for hvor godt linjen passer på de givne punkter, men den kan ikke stå alene. Når man foretager lineær regression for at finde den bedste rette linje, kan man gøre det uden at tegne grafen. Man kan sådan set nøjes med at få et CAS-værktøj til at beregne linjens ligning og forklaringsgraden R^2. Forklaringsgraden kan så bruges til at afgøre, om det er fornuftigt at modellere sammenhængen med en ret linje.

I praksis er det dog altid fornuftigt at tegne grafen, for det viser sig, at man kan få den samme rette linje og forklaringsgrad ud fra vidt forskellige data.

Statistikeren Francis Anscombe beskrev i en artikel i 1973 fire forskellige datasæt, som havde den samme regressionsligning og forklaringsgrad, men så vidt forskellige ud.[1] De fire datasæt kan ses i tabel 4.4.

Indsætter man de fire datasæt i hver deres koordinatsystem får man billedet på figur 4.5. Det ses tydeligt, at de fire datasæt er udtryk for vidt forskellige fordelinger. Det første datasæt ser ud til nogenlunde at kunne modelleres med en ret linje. Punkterne ligger i hvert fald nogenlunde tilfældigt omkring linjen. Det næste datasæt (øverst til højre) viser en tydelig sammenhæng – men den er bestemt ikke lineær. De sidste to datasæt har et enkelt punkt, der ligger helt anderledes end alle de andre (en outlier).



Figur 4.5: Anscombes fire datasæt afbildet i hver deres koordinatsystem.
Her ses det tydeligt, at det drejer sig om vidt forskellige sammenhænge.

På trods af deres forskellighed har de alle den samme regressionslinje og forklaringsgrad, nemlig

$$y = 0{,}50 \cdot x + 3{,}00\ , \quad R^2 = 0{,}67\ .$$

Forklaringsgraden alene er altså ikke tilstrækkelig til at afgøre om en lineær model passer »godt« på de givne data. Det er derfor en god ide at tegne grafen så man kan få et overblik over hvordan punkterne fordeler sig, før man evt. foretager lineær regression.

I det tilfælde, hvor et enkelt punkt ligger helt anderledes end de andre,

x	y
4	4,26
5	5,68
6	7,24
7	4,82
8	6,95
9	8,81
10	8,04
11	8,33
12	10,84
13	7,58
14	9,96

x	y
4	3,1
5	4,74
6	6,13
7	7,26
8	8,14
9	8,77
10	9,14
11	9,26
12	9,13
13	8,74
14	8,1

x	y
4	5,39
5	5,73
6	6,08
7	6,42
8	6,77
9	7,11
10	7,46
11	7,81
12	8,15
13	12,74
14	8,84

x	y
8	6,58
8	5,76
8	7,71
8	8,84
8	8,47
8	7,04
8	5,25
8	5,56
8	7,91
8	6,89
19	12,5

Tabel 4.4: Anscombes fire datasæt. Fra [1].

giver det god mening at undersøge dette punkt grundigere. Kunne det evt. være resultatet af en fejlmåling? Og hvis grafen krummer på en karakteristisk måde, kunne det måske være at man skulle anvende en helt anden regressionstype.

4.2 Residualplot og residualspredning

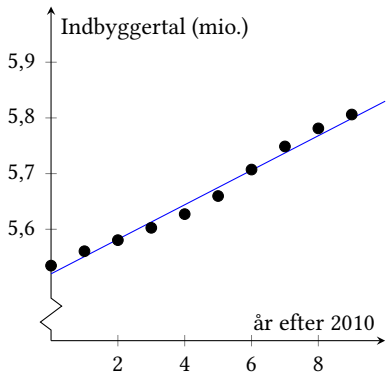
Når nu forklaringsgraden ikke kan bruges som objektivt mål, giver det altså god mening at undersøge grafen. Men det kan være svært at se med det blotte øje, om punkterne ligger tæt på grafen, eller om der er en karakteristisk afvigelse mellem punkterne og linjen.

Man bør derfor altid kigge på residualplottet, dvs. et plot af residualerne som funktion af de tilhørende x-værdier. Disse skal helst være små i forhold til de målte y-værdier, og de må ikke udvise noget karakteristisk mønster.

Men man kan faktisk sige noget mere end det. Hvis der er en lineær sammenhæng i dataene, så vil afvigelserne (dvs. residualerne) være udtryk for måleusikkerhed. Man kan derfor lave statistik på residualerne for at undersøge dette.

Tabel 4.6: Danmarks indbyggertal 2010–2019.[4]

Årstal	Indbyggertal
2010	5 534 738
2011	5 560 628
2012	5 580 516
2013	5 602 628
2014	5 627 235
2015	5 659 715
2016	5 707 251
2017	5 748 769
2018	5 781 190
2019	5 806 081



Figur 4.7: Regression over Danmarks indbyggertal 2010–2019.

Når man foretager lineær regression, bliver residualerne

$$r_i = y_i - ax_i - b \ ,$$

dvs. gennemsnittet af residualerne er

$$\overline{r} = \overline{y} - a\overline{x} - b \ .$$

Men da $b = \overline{y} - a\overline{x}$, bliver $\overline{r} = 0$. Gennemsnittet af residualerne er altså 0.

Spredningen af residualerne kan estimeres vha. den såkalde residualspredning. Den beregnes på følgende måde:

Definition 4.4

Hvis en række datapunkter $(x_1; y_1), \ldots, (x_n; y_n)$ modelleres med den rette linje $y = a \cdot x + b$, er *residualspredningen* givet ved

$$s = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{r_1^2 + r_2^2 + \cdots + r_n^2}{n-2}} \ ,$$

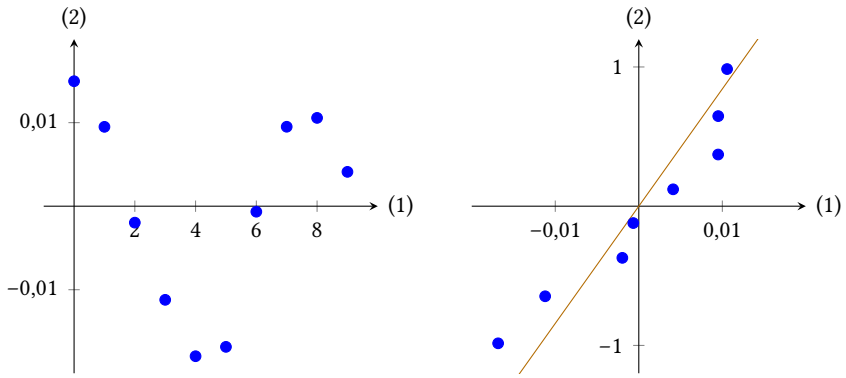
hvor $r_1, r_2, \ldots, r_n$ er residualerne.

Det viser sig at måleusikkerheder typisk er normalfordelte. Hvis data kan modelleres fornuftigt med en lineær model, skal residualerne være normalfordelte med middelværdi 0 og spredning s.[3]

Eksempel 4.5 Tabel 4.6 viser Danmarks indbyggertal i årene 2010–2019. Hvis man foretager lineær regression på disse data, får man grafen på figur 4.7. Regressionsligningen er

$$y = 0{,}031x + 5{,}520 \ ,$$

hvor x er antal år efter 2010, og y er indbyggertallet i mio.



(a) Residualplot. (b) Normalfordelingsplot.

Figur 4.8: Residualplot og normalfordelingsplot over residualerne ved lineær regression over Danmarks indbyggertal 2010–2019.

Forklaringsgraden og residualspredningen kan beregnes til

$$R^2 = 0{,}985 \quad \text{og} \quad s = 0{,}013\ .$$

Forklaringsgraden viser at punkterne ligger ganske tæt på den rette linje, og residualspredningen er lille set i forhold til y-værdierne der alle ligger mellem 5 og 6.

Figur 4.8 viser residualplottet og et normalfordelingsplot over residualerne. Ud fra residualplottet kan man argumentere for at der muligvis er et mønster i residualerne, men uden flere data kan man lige så vel argumentere for at residualerne svinger tilfældigt.

Normalfordelingsplottet viser til gengæld at residualerne er nogenlunde normalfordelte da punkterne ligger tæt på den rette linje. Samlet set kan man ud fra de to parametre, grafen, samt de to plots på figur 4.8 argumentere for at Danmarks indbyggertal i den givne periode udmærket kan beskrives ved en lineær model.

4.3 Konfidensintervaller

Hvis det er rimeligt at beskrive data med en lineær model, er den beregnede hældningskoefficient og skæring med andenaksen et estimat for de virkelige værdier i den bagvedliggende model. I denne sammenhæng skelner man mellem de virkelige værdier for de to parametre a og b og de estimerede tal $\hat{a}$ og $\hat{b}$ som er beregnet ud fra stikprøven.

Man er typisk interesseret i at finde ud af hvor godt estimatet egentlig er. Man kan derfor beregne et såkaldt *konfidensinterval* for hældningskoefficienten a som den med en hvis procentdels sandsynlighed vil ligge i. Ofte beregner man det såkaldte 95% konfidensinterval: Hvis dataene er udtryk for en stikprøve fra en bagvedliggende lineær model, vil 95% af de hældningskoefficienter man beregner på baggrund af en stikprøve ligge i dette interval. Man kan derfor med nogenlunde rimelighed påstå at 95% konfidensintervallet angiver hvor den »rigtige« hældningskoefficient med 95% sikkerhed befinder sig.

Idet måleusikkerheder forventes at være normalfordelte, kan man udlede at de hældningskoefficienter man kan beregne ud fra stikprøver, er

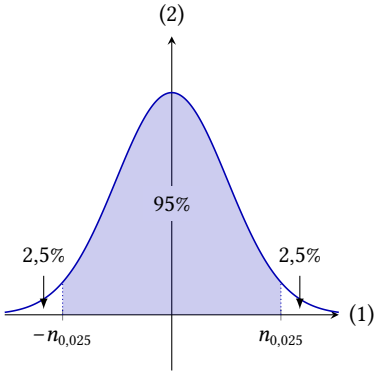
normalfordelte med middelværdi a og spredning

$$\sigma_a = \frac{\sigma}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}} = \frac{\sigma}{\sqrt{S_{xx}}},$$

hvor σ er den teoretiske spredning af residualerne. Man kan finde et såkaldt 95% konfidensinterval ved at beregne intervalgrænserne

$$a \pm n_{0,025} \cdot \frac{\sigma}{\sqrt{S_{xx}}}.$$

Tallet $n_{0,025}$ er den værdi som 2,5% af standardnormalfordelingen (dvs. med $\mu = 0$, $\sigma = 1$) ligger over. Dvs. 95% af normalfordelingen ligger mellem $\pm n_{0,025}$ (se figur 4.9). Da normalfordelingen skalerer pænt, ligger 95% af en hvilken som helst normalfordeling altså mellem de to værdier $\mu \pm n_{0,025} \cdot \sigma$.



Figur 4.9: For standardnormalfordelingen ligger 95% af fordelingen mellem $\pm n_{0,025}$.

95% af de mulige stikprøver vil altså resultere i en estimeret værdi $\hat{a}$ der ligger mellem grænserne

$$a \pm 1{,}96 \cdot \frac{\sigma}{\sqrt{S_{xx}}}.$$

Problemet er nu at man hverken kender den teoretiske residualspredning σ eller den sande værdi af a, men kun den estimerede residualspredning s og den estimerede hældning $\hat{a}$. Og når man bruger denne estimerede spredning, er de estimerede $\hat{a}$-værdier ikke længere normalfordelte – de fordeler sig derimod efter den såkaldte t-fordeling.[4] Den estimerede spredning for a bliver[7]

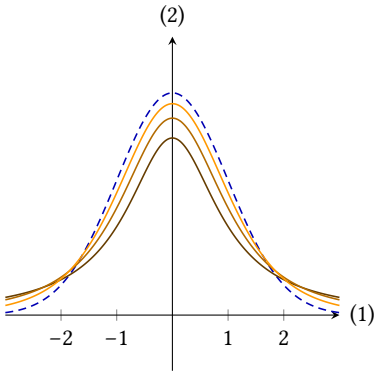
$$s_a = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}} = \frac{s}{\sqrt{S_{xx}}},$$

hvor s er den estimerede residualspredning.

[4] t-fordelingen er en fordeling man finder når man undersøger normalfordelte data på baggrund af en middelværdi og en spredning som er estimeret ud fra en stikprøve.[6]

Størrelsen $\hat{a}$ er som nævnt t-fordelt. t-fordelingen er den fordeling man finder hvis man kigger på fordelingen af en normalfordelt parameter der er estimeret ud fra en stikprøve. Idet man kun kender en stikprøve og ikke hele det underliggende datasæt, får man en frekvensfordeling der ligner normalfordelingen, men med »tykkere haler« fordi en større procentdel af de målte værdier vil ligge længere væk fra middelværdien når middelværdi og spredning kun er estimater.

Ydermere vil t-fordelingen afhænge af hvor stor stikprøven er – det såkaldte antal *frihedsgrader*. Som man kan se på figur 4.10 vil t-fordelingen komme tættere og tættere på normalfordelingen jo flere frihedsgrader der er, dvs. jo større stikprøven er. Det skyldes at jo flere data man har, jo tættere vil den estimerede middelværdi og spredning ligge på de rigtige værdier – og den estimerede fordeling vil så ligge tættere på den teoretiske normalfordeling.



Figur 4.10: Standardnormalfordelingen (stiplet) og t-fordelingen med hhv. 1, 2 og 5 frihedsgrader.

Har man n datapunkter vil man så kunne beregne et 95% konfidensinterval for parameteren a som er

$$\hat{a} \pm t_{0,025} \cdot \frac{s}{\sqrt{S_{xx}}}.$$

hvor $t_{0,025}$ svarer til tallet $n_{0,025}$, men for t-fordelingen med $n-2$ frihedsgrader.

På samme måde kan man i øvrigt bestemme et konfidensinterval for parameteren b (linjens skæring med andenaksen). Her finder man konfidensintervallet[7]

$$\hat{b} \pm t_{0,025} \cdot \frac{s}{\sqrt{n}} ,$$

hvor s er den estimerede residualspredning, og n er antallet af datapunkter.

Alle argumenterne ovenfor kan samles i denne sætning:

Sætning 4.6

Hvis man foretager lineær regression på et datasæt, kan man bestemme $(1-\alpha)$-konfidensintervaller for parametrene a og b i den lineære model $y = ax + b$ som

$$\hat{a} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{S_{xx}}} ,$$

og

$$\hat{b} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} .$$

Tallet α i sætningen ovenfor er 5% (dvs. 0,05) for et 95% konfidensinterval, 1% (dvs. 0,01) for et 99% konfidensinterval, osv.

4.4 Andre typer regression

Sammenhængen mellem x og y i et datasæt $(x_1; y_1), \ldots, (x_n; y_n)$ er ikke nødvendigvis lineær. Der kunne for eksempel også være tale om en potens-, en eksponentiel eller en polynomiel sammenhæng.

Potens- og eksponentiel regression foretages af de fleste programmer ved at transformere datasættet og derefter lave lineær regression på de transformerede data. Hvis sammenhængen mellem x og y er eksponentiel, har man nemlig

$$\begin{aligned} y &= b \cdot a^x \\ \log(y) &= \log(b \cdot a^x) \\ \log(y) &= \log(a) \cdot x + \log(b) , \end{aligned}$$

dvs. hvis sammenhængen mellem x og y er eksponentiel, så vil sammenhængen mellem x og $\log(y)$ være lineær. Man kan derfor lave lineær regression på datasættet $(x_1; \log(y_1)), \ldots, (x_n; \log(y_n))$. Herved finder man $\log(a)$ og $\log(b)$ hvorefter man kan transformere tilbage til a og b.

Man kan tilsvarende transformere en potenssammenhæng til en lineær sammenhæng ved at tage en logaritme til både x- og y-værdierne i datasættet.

Pointen er her at hvis man foretager regression på baggrund af transformerede data, vil R^2 også være beregnet på baggrund af transformerede data. Man skal altså her være ekstra forsigtig med fortolkningen, og det er derfor vigtigt at man også kigger på både grafen og residualplottet.

Bibliografi

[1] F. J. Anscombe. »Graphs in Statistical Analysis«. I: *The American Statistician* 27.1 (feb. 1973), s. 17–21.

[2] Dan Bobkoff. *A magazine once polled millions on the presidential election – and got the results dead wrong*. Business Insider. 23. aug. 2016. URL: http://nordic.businessinsider.com/magazines-presidential-poll-was-dead-wrong-2016-8 (bes. 07.06.2018).

[3] Per Bruun Brockhoff, Claus Thorn Ekstrøm og Ernst Hansen. »Lineær regression – lidt mere tekniske betragtninger om R^2 og et godt alternativ«. I: *LMFK-bladet* nr. 2 (2017).

[4] Danmarks Statistik. *FOLK2: Folketal 1. januar efter køn, alder, herkomst, oprindelsesland og statsborgerskab*. URL: https://statistiksbanken.dk.

[5] Dominic Lusinchi. »'President' Landon and the 1936 Literary Digest Poll: Were Automobile and Telephone Owners to Blame?« I: *Social Science History* 36.1 (2012), s. 23–54.

[6] Christian Walck. *Hand-book on statistical distributions for experimentalists*. University of Stockholm, 10. sep. 2007.

[7] Thomas H. Wonnacott og Ronald J. Wonnacott. *Introductory Statistics for Business and Economics*. 2. udg. John Wiley & Sons, Inc., 1977.