Projekt 8.7 Konfidensintervaller bestemt ved bootstrapping

(*Dette projekt er en udbygning af afsnit 4 i HEM3, kapitel 8, der handler om estimering af parametrene i en lineær regression, specielt om at bestemme konfidensintervaller. Det kan gøres med værktøjsprogrammer, men det kan også gøres med en særlig simuleringsteknik, der kaldes bootstrapping. Det er hovedsigtet i dette projekt at dykke ned i denne betydningsfulde teknik i moderne statistik. Vi giver også i projektet en formel, der giver mulighed for umiddelbart at bregne konfidensintervallet ud fra data – ligesom man kan beregne den lineære regression ud fra de givne data. De første 2 sider af projektet er identisk med første del af afsnit 4.*)

1. Lineær regression som statistisk metode

I *Hvad er matematik*? 2, kapitel 8 har vi analyseret den lineære regressionsmodel, og udledt formlen for, hvordan vi bedst estimerer parametrene $\hat{a}$ og $\hat{b}$ i den lineære model ud fra det givne datasæt. Vi forestiller os i hele den analyse, at der findes en "sand lineær model", $y = a_0 \cdot x + b_0$, som ligger bag både de givne data, og data fra andre lignende eksperimenter, man kunne foretage. Vi kan aldrig nå frem til med sikkerhed at sige: "Dette er den sande model". Men vi forestiller os, at

- De *empiriske* dataværdier y_i er tæt ved den *teoretiske* værdi fra sande model, $a_0 \cdot x_i + b_0$
- Forskellen er et mindre *residual*, r_i, således at vi kan skrive $y_i = a_0 \cdot x_i + b_0 + r_i$
- Residualer kan beskrives ved en normalfordelt stokastisk variabel R.

Dvs normalfordelingsteorien kan bidrage til at vi udbygger vores lineære regressionsmodel

Øvelse 1: Estimat for middelværdi og spredning i Galton datasættet

Hent Galtons datasæt på ***her***. Anvend i det følgende metoder fra eksempel 1 i afsnit 3.4.

a) Udfør lineær regression på datasættet med brug af et værktøjsprogram.
b) Bestem residualerne.
c) Bestem et estimat for middelværdien af residualerne i Galton datasættet.
d) Bestem et estimat for spredningen af residualerne i Galton datasættet.
e) Tegn et normaliseret histogram for fordelingen af residualerne, dvs et histogram med areal 1
f) Tegn grafen for tæthedsfunktionen for den overlejrende normalfordeling sammen med det normaliserede histogram.

2. Usikkerhed på estimatet for hældning og konstantled

Estimaterne for hældningen $\hat{a}$ og konstantleddet $\hat{b}$ er bestemt ud fra data i en stikprøve, og vi kan forestille os, at vi med en anden stikprøve vil få andre estimater. Vi vil i det følgende prøve at undersøge denne statistiske usikkerhed på hældning og konstantled i den lineære regressionsmodel. Vi kunne i praksis forsøge at indsamle nye datasæt, men det er ikke den metode, som vi vil forfølge i dette afsnit.

Vi vil i stedet forsøge at frembringe en variation i estimaterne ud fra de eksisterende residualer vha. den statistiske metode, der kaldes *bootstrapping*. Metoden er vidt udbredt i anvendt statistik fx i hele det farmaceutiske område, når nye præparater skal testes. Vi vil her tage de første skridt ind i brugen af denne metode.

Ideen er, at nye datapunkter kan fremkomme ved, at vi bytter rundt på residualerne! Til denne ombytning vælger vi *med* tilbagelægning fra stikprøven af residualer.

Vi ser på en stikprøve på 10 datapunkter:

nummer	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Farens højde	186,9	184,6	185	182,1	179,4	178,4	179,7	179,7	176,5	173,4
Sønnens højde	183,4	172	179	165,4	155,4	160,3	165	168,7	160,3	157,5

Øvelse 2: **Hældning og konstantled ud fra 10 datapunkter**

Udfør regression og bestem estimater for hældning og konstantled.
Du skal få: til $\hat{a}=1{,}94$ og $\hat{b}=-183{,}58$.

Øvelse 3: residualer.
Bestem residualerne og opret en ny linje hertil i tabellen: Du skal få:

Residualer	4,42	-2,52	3,71	-4,27	-9,03	-2,19	-0,01	3,69	1,5

Øvelse 4: Frembring nye datasæt ved bootstrapping.

a) Vælg fra listen med residualer, du lige har bestemt, 10 residualer vha. et matematisk værktøjsprogram. Du skal vælge *med* tilbagelægning. Du kan fx få:

s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}
-4,27	1,50	-4,27	-0,01	4,71	3,71	-2,52	-4,27	-2,52	4,71

b) Frembring et nyt datasæt med de 10 udvalgte residualer ud fra $z_i=\hat{a}\cdot x_i+\hat{b}+s_i$
Med ovenstående residualer ville man som det første element få: $1{,}94\cdot 186{,}9-183{,}58-4{,}27=174{,}74$
Bemærk: Du vil få andre værdier, da du givetvis har et andet udtræk af residualer end ovenstående.

Øvelse 5: **Hældning og konstantled for det første datasæt i din bootstrapping**
Bestem nye estimater for hældning og konstantled ud fra det bootstrappede datasæt.
Estimaterne ud fra det angivne bootstrappede datasæt er $\hat{a}=1{,}48$ og $\hat{b}=-100{,}65$.

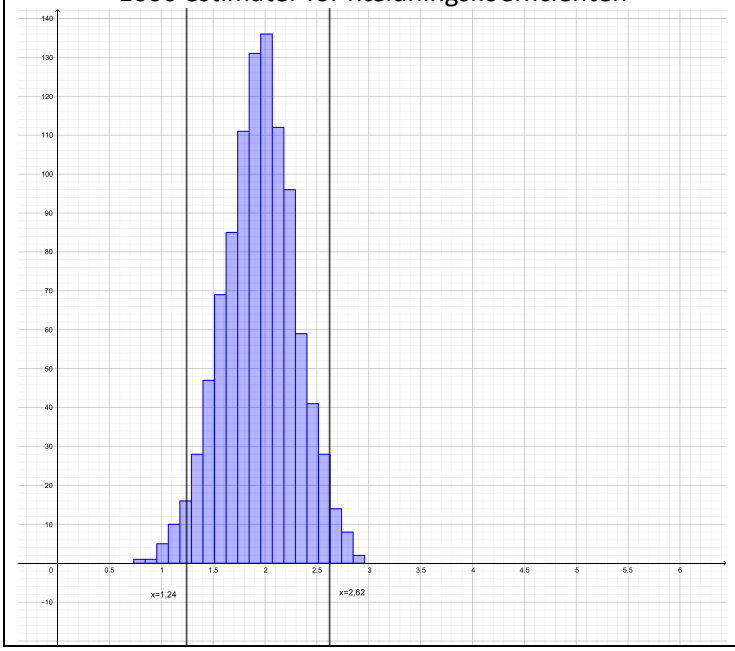
Øvelse 6: **Gentag processen**

a) Gentag processen i øvelse 5 1000 gange. I kan evt. i klassen / på holdet tilrettelægge arbejdet, så hver af jer indsamler et vist antal, som I så samler i ét regneark.

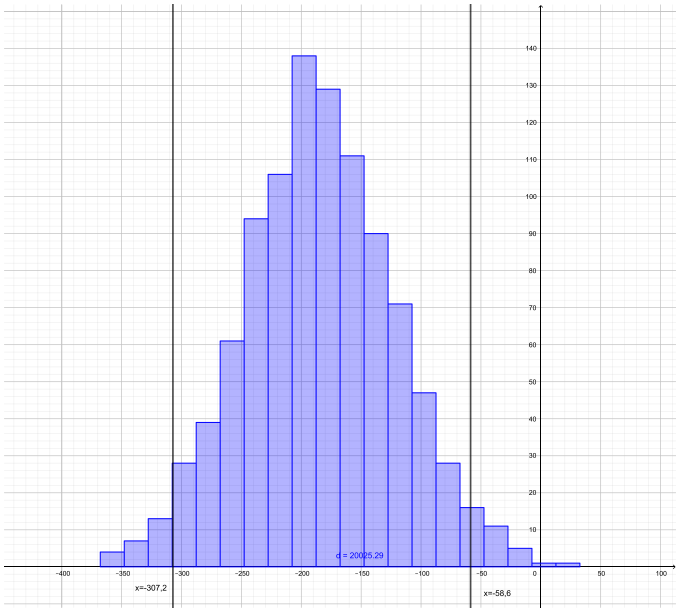
b) Tegn et histogram over de 1000 estimater for hældning *a* og konstantled *b*.

Da hvert bootstrappet datasæt er genereret af dit matematiske værktøjsprogram, så vil du ikke få de samme datasæt og endeligt histogram, men det vil ligne:

1000 estimater for hældningskoefficienten



1000 estimater for konstantleddet



Vi husker, at bootstrapping denne gang er foregået på grundlag af stikprøven på 10 dataværdier. Resultaterne kan naturligvis ikke blive meget anderledes end stikprøvens, og stikprøven er taget som de første 10 datasæt i Galtons sæt med 952 punkter. Du vil se nedenfor, at dette datasæt langtfra er repræsentativ! Men metoden er den samme.

Øvelse 7. Bestem konfidensintervaller ud fra histogrammerne

Aflæs nu på dine egne histogrammer de relevante grænser til at bestemme konfidensintervaller. For de histogrammer du ser ovenfor får vi:

a) Grænsen for 2,5% mindste estimater for hældningen bestemmes til 1,24.

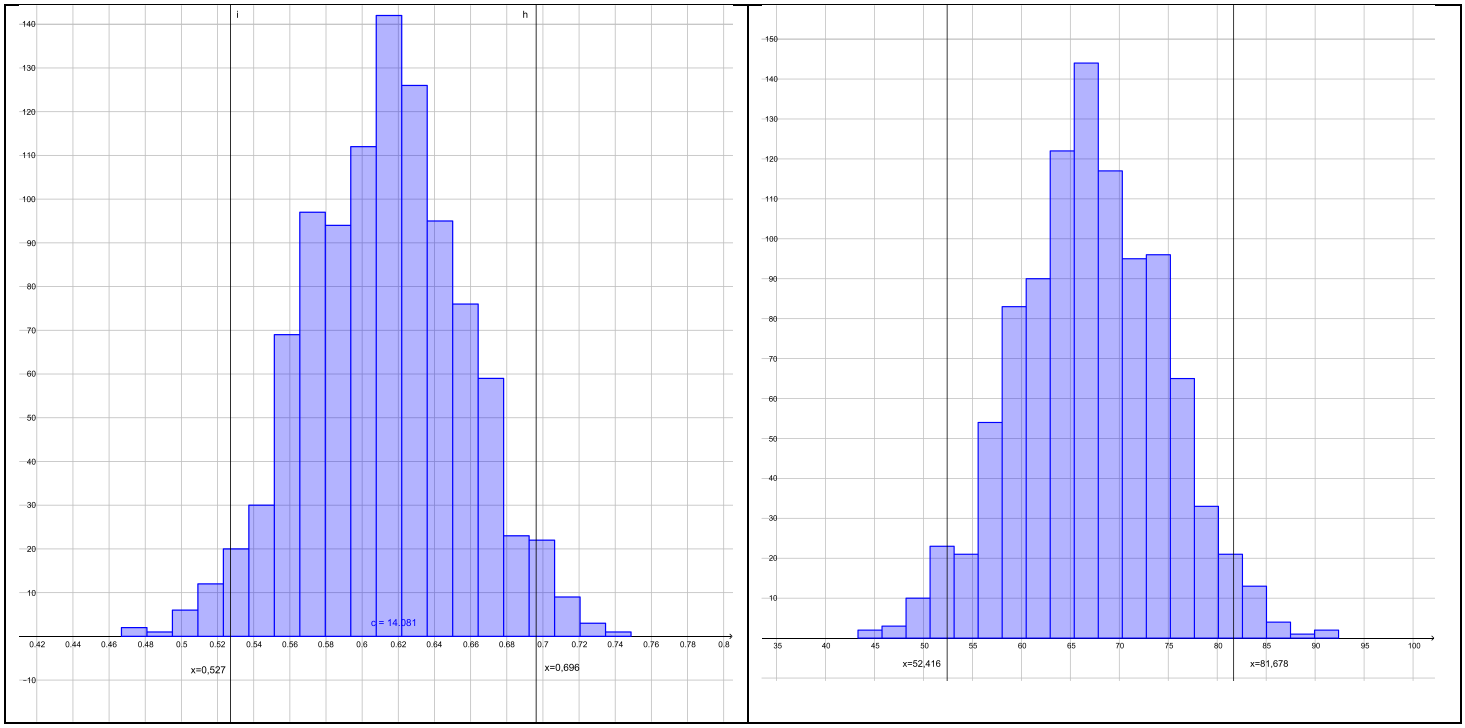
b) Grænsen for de 2,5% største estimater for hældningen bestemmes til 2,62.

c) Et 95% konfidensinterval for hældningen *a* bestemmes til $[1{,}24;2{,}62]$.

d) Et 95% konfidensinterval for konstantleddet *b* bestemmes til $[-307{,}2;-58{,}6]$.

Øvelse 8. Gentag processen på hele Galtons datasæt

1. Gentag processen ovenfor på Galtons datasæt på 952 datapunkter. Du vil få histogrammer, der ser nogenlunde således ud:.



2. Aflæs på baggrund af de 1000 bootstrappede datasæt ud fra det store datasæt med 952 datapunkter. Vi aflæser af ovenstående:

2.1. 95% konfidensinterval for hældningen *a* er $[0{,}527;0{,}696]$. Da 0 ikke ligger i intervallet, så har vi belæg for at sige, at der er en lineær sammenhæng mellem fædres højde og sønners højde.

2.2. 95% konfidensinterval for hældningen *b* er $[52{,}416;81{,}678]$.

3. Formelbaseret bestemmelse af konfidensintervaller for estimater

Som omtalt i indledningen findes en formel, der ud fra datasættet direkte giver mulighed for at beregne konfidensintervaller. Det er en ret avanceret formel, der inddrager *t*-fordelingen, og vi angiver den alene for at vise, at i den teoretiske statistik arbejdes både med simuleringer af empiriske data, og med teorien bag.

Praksisboks: Formel for 95% konfidensinterval for hældningen *a*

Et 95% konfidensinterval for hældningen a i en lineær regressionsmodel er givet ved

$$\left[\hat{a}-t_{0,975}\cdot\sqrt{\frac{\frac{1}{n-2}\cdot\sum_{i=1}^{n}r_i^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}};\hat{a}+t_{0,975}\cdot\sqrt{\frac{\frac{1}{n-2}\cdot\sum_{i=1}^{n}r_i^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}}\right]$$

I formlen for 95% konfidensintervallet genkender vi residualspredningen som tælleren i brøken under kvadratroden. Nævneren har vi ligeledes set tidligere i formlen for estimatet for hældningen $\hat{a}$ i kapitel 8 i bog 2. Faktoren $t_{0,975}$ er ny og kan bestemmes ved hjælp af et matematisk værktøjsprogram.

Eksempel: Udregning af 95% konfidensinterval for Galton datasættet ud fra formel

Vi har tidligere bestemt $\hat{a}=0{,}613$, $\frac{1}{n-2}\cdot\sum_{i=1}^{n}r_i^2=\frac{1}{952-1}\cdot 34040{,}85=35{,}7948$, $\sum_{i=1}^{n}(x_i-\overline{x})^2=19596{,}5062$ og $n=952$.

$t_{0,975}=\text{inverseTDistribution}(951;0{,}975)=1{,}962.$

Ud fra formlen får vi $\left[0{,}613-1{,}962\cdot\sqrt{\frac{35{,}7948}{19596{,}5062}};0{,}613+1{,}962\cdot\sqrt{\frac{35{,}7948}{19596{,}5062}}\right]=\left[0{,}5291;0{,}6969\right]$.

Hvilket er i overensstemmelse med 95% konfidensintervallet for *a* bestemt ved bootstrapping metoden.

4. Bestemmelse af konfidensintervaller for estimater ved hjælp af værktøjsprogram

Vi kan beregne konfidensintervallerne i alle de gængse værktøjsprogrammer. Det kan se sådan ud:

Eksempel: Udregning af 95% konfidensinterval for Galton datasættet ud fra et matematisk værktøjsprogram

testLin(GaltonData)

	a	b
Koefficient	0.612762	66.981821
Standardfejl	0.042761	7.417231
t-stat	14.329898	9.030570
p-værdi	0.000000	0.000000
Nedre 95.00%	0.528845	52.425772
Øvre 95.00%	0.696680	81.537871
Frihedsgrader	950	

Hvilket igen er i overensstemmelse med 95% konfidensintervallet for *a* bestemt ved bootstrapping metoden.

På bogens website (Hvad er matematik? 3 – LRU.praxis.dk) finde en vejledning til brug af de gængse værktøjsprogrammer til bestemmelse af 95% konfidensinterval for hældningen og konstantled. Filen hedder *Konfidensintervaller på parametrene i linreg*.