Kapitel 7

UAFHÆNGIGE IDENTISK FORDELTE NORMALE OBSERVATIONER

Da normalfordelingen er en kontinuert fordeling, optræder der følgende forskel fra de diskrete modeller vi hidtil har studeret. Ved analyse af statistiske modeller for kontinuert fordelte observationer definerer man likelihoodfunktionen som værdien af *tætheden* i det observerede punkt, som funktion af de ukendte parametre. Dette er i virkeligheden ikke en ændring i forhold til definitionen i det diskrete tilfælde. Man må tænke på, at data i praksis altid vil være givet på afrundet form, så det man observerer er en hændelse af formen "$X \in A_x$", hvor A_x er den lille mængde af kontinuerte observationer, som ved afrunding giver netop den observation x man har fået. Hermed er man tilbage i et diskret problem, hvor likelihoodfunktionen kan defineres på sædvanlig måde ved

$$L(\vartheta) = P_\vartheta(X \in A_x).$$

Men da sandsynligheden for at den kontinuerte observation havner i A_x netop — ifølge tæthedens fortolkning — er approksimativt proportional med tæthedens værdi i et punkt af A_x, bliver denne "rigtige likelihood" (på nær den approksimation, som under alle omstændigheder ligger i afrundingen) proportional med værdien af tætheden i det observerede punkt x.

7.1. Uafhængige identisk fordelte observationer, kendt varians.

Vi går nu over til at kalde observationerne for $y, y_i, \ldots$, fordi bogstavet x skal bruges til noget andet i forbindelse med regressionsanalysen.

Lad $y_1, \ldots, y_n$ være observerede værdier af n uafhængige, normalfordelte stokastiske variable $Y_1, \ldots, Y_n$ med middelværdi μ og varians σ^2. Vi begynder med at opfatte variansen σ^2 som kendt, og diskuterer estimation af middelværdien μ og test for simpel hypotese af formen $\mu = \mu_0$.

Likelihoodfunktionen bliver

$$\begin{aligned} L(\mu) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mu)^2\right) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right), \end{aligned}$$

og log–likelihooden (idet vi kan se bort fra normeringsfaktoren, da den kun afhænger af σ^2)

$$l(\mu) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2 .$$

Maksimum–likelihood estimatet for middelværdien er således den værdi af μ der minimerer kvadratsummen

$$\sum_{i=1}^{n}(y_i-\mu)^2 .$$

Som vi skal se er det et generelt træk ved normalfordelingsmodellerne, at estimaterne for middelværdiparametrene fås ved minimering af kvadratsummen af observationernes afvigelser fra deres middelværdier. Derfor bruger man ofte betegnelsen *mindste kvadraters metode* for denne estimationsmetode. Blandt økonomer bruges også forkortelsen OLS (af engelsk **O**rdinary **L**east **S**quares).

Minimering af ovenstående kvadratsum foretages lettest ved følgende omskrivning (som i øvrigt er velkendt fra sandsynlighedsregningen, idet det stort set er identiteten $\mathrm{E}\left((Y-\mu)^2\right) = \mathrm{E}\left((Y-\mathrm{E}Y)^2\right) + (\mathrm{E}Y-\mu)^2$ for en stokastisk variabel Y, som følger den empiriske fordeling bestemt ved $y_1,\ldots,y_n$). Med betegnelsen $\bar{y} = \frac{1}{n}\sum y_i$ for observationernes gennemsnit har vi

$$\sum(y_i-\mu)^2 = \sum\left((y_i-\bar{y})+(\bar{y}-\mu)\right)^2$$

$$= \sum(y_i-\bar{y})^2 + \sum(\bar{y}-\mu)^2 + 2\sum(y_i-\bar{y})(\bar{y}-\mu)$$

$$= \sum(y_i-\bar{y})^2 + n(\bar{y}-\mu)^2.$$

Af omskrivningen følger umiddelbart, at likelihooden antager sit maksimum for

$$\hat{\mu} = \bar{y}.$$

Eksempel. Ved gentagne målinger (såsom vejninger, kemiske koncentrationsbestemmelser og lignende) anbefales det ofte at angive målingernes gennemsnit som "facit". Dette kan åbenbart begrundes teoretisk, som en anvendelse af maksimum likelihood estimation i en model hvor målingerne fortolkes som uafhængige, identisk normalfordelte, med den "sande værdi" af det der skal måles som middelværdi. Den centrale grænseværdisætning giver i et vist omfang belæg for den opfattelse, at målefejl følger en normal fordeling.

Med henblik på at angive en usikkerhed på estimatet $\hat{\mu} = \bar{y}$ bemærker vi, at den stokastiske variable $\bar{Y}$ ifølge den normale fordelings foldningsegenskab er normalfordelt med middelværdi μ og varians $\frac{\sigma^2}{n}$. Med sandsynlighed 0.95 gælder derfor

$$-1.96 \leq \frac{\mu - \bar{Y}}{\sqrt{\sigma^2/n}} \leq 1.96$$

eller

$$\bar{Y} - 1.96\sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \bar{Y} + 1.96\sqrt{\frac{\sigma^2}{n}},$$

hvilket gør det naturligt at angive estimationsresultatet på formen

$$\mu = \bar{y} \pm 1.96\sqrt{\frac{\sigma^2}{n}}.$$

Intervallet

$$\left[\bar{y} - 1.96\sqrt{\frac{\sigma^2}{n}}, \bar{y} + 1.96\sqrt{\frac{\sigma^2}{n}}\right]$$

kaldes i denne forbindelse et 95 % *konfidensinterval* eller *sikkerhedsinterval* for μ, og dets endepunkter kaldes for *konfidensgrænser* eller *sikkerhedsgrænser*. Generelt defineres et 95 % konfidensinterval for en parameter i en statistisk model som et interval, udregnet på basis af observationerne, der med sandsynlighed 0.95 indeholder den sande værdi af parameteren.

7.2. Test for simpel hypotese når variansen er kendt.

Antag, at vi i modellen for n identisk normalfordelte variable med ukendt middelværdi μ og kendt varians σ^2 ønsker at teste en hypotese af formen $\mu = \mu_0$. Kvotientteststørrelsen bliver

$$2(l(\bar{y}) - l(\mu_0)) = 2\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \bar{y})^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_0)^2\right)$$

$$= \frac{n}{\sigma^2}(\bar{y} - \mu_0)^2 = \left(\frac{\bar{y} - \mu_0}{\sqrt{\sigma^2/n}}\right)^2.$$

Under hypotesen $\mu = \mu_0$ er

$$U = \frac{\bar{Y} - \mu_0}{\sqrt{\sigma^2/n}}$$

normeret normalfordelt. Heraf følger at $-2\log Q = U^2$ er χ^2–fordelt med 1 frihedsgrad (eksakt, til en afveksling). Testets P–værdi kan altså udtrykkes som

$$P\left(U^2 \geq \frac{n}{\sigma^2}(\bar{y}-\mu_0)^2\right) = P\left(|U| \geq \left|\frac{\bar{y}-\mu_0}{\sqrt{\sigma^2/n}}\right|\right) = 2P\left(U \geq \left|\frac{\bar{y}-\mu_0}{\sqrt{\sigma^2/n}}\right|\right)$$

hvor U er en normeret normalfordelt stokastisk variabel (og U^2 dermed en χ^2–fordelt variabel med 1 frihedsgrad). Det sidste udtryk for P–værdien antyder hvordan testet kan foretages som et *tosidet U–test*, dvs. et test hvor størrelsen $u = (\bar{y}-\mu_0)/\sqrt{\sigma^2/n}$ vurderes ved opslag i en normalfordelingstabel. Fordelen ved at gøre det på den måde er, at man så samtidig holder styr på hvilken vej afvigelsen går. Bemærk, at P–værdien bliver ≥ 0.05 hvis og kun hvis værdien μ_0 ligger i 95 % konfidensintervallet for μ.

Normalfordelingsmodeller med kendt varians er ikke så relevante i praksis. I det følgende vil vi udelukkende beskæftige os med modeller, hvor variansen er ukendt. Her skal blot bemærkes, at de lineære normalfordelingsmodeller med kendt varians generelt har den pæne egenskab, at kvotientteststørrelserne er eksakt χ^2–fordelte. De generelle resultater vedrørende approksimativ χ^2–fordeling af kvotientteststørrelser bygger på approksimation af statistiske modeller med normalfordelingsmodeller af denne simple slags, baseret på den centrale grænseværdisætning og Taylorudvikling af log–likelihooden.

7.3. Uafhængige identisk fordelte observationer, ukendt varians.

Lad igen $y_1,\dots,y_n$ være observerede værdier af n uafhængige normalfordelte stokastiske variable $Y_1,\dots,Y_n$ med middelværdi μ og varians σ^2. Nu opfatter vi imidlertid både μ og σ^2 som ukendte parametre. Log–likelihooden bliver så (idet logaritmen til normeringsfaktoren nu skal medtages, bortset fra at konstanten $(1/\sqrt{2\pi})^n$ naturligvis kan ignoreres)

$$l(\mu,\sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2 .$$

Maksimering m.h.t. μ for fast σ^2 giver samme resultat som i modellen med kendt varians, så vi har stadig $\hat{\mu} = \bar{y}$. Hvis vi indsætter dette får vi den *delvis maksimerede log–likelihood*

$$l(\bar{y},\sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\bar{y})^2 = -\frac{1}{2}\left(n\log\sigma^2 + \frac{\mathrm{SSD}_y}{\sigma^2}\right)$$

hvor vi her og i det følgende anvender betegnelsen SSD_y (**S**um of **S**quares of **D**eviations) for kvadratsummen af observationernes afvigelser fra deres gennemsnit. Hvis vi kan maksimere dette udtryk som funktion af σ^2

har vi åbenbart fundet maksimaliseringsestimatorerne. Differentiation m.h.t. σ^2 giver

$$\frac{d}{d\sigma^2} l(\bar{y}, \sigma^2) = -\frac{1}{2}\left(\frac{n}{\sigma^2} - \frac{\mathrm{SSD}_y}{(\sigma^2)^2}\right) = -\frac{1}{2\sigma^2}\left(n - \frac{\mathrm{SSD}_y}{\sigma^2}\right).$$

Da denne funktion af σ^2 er nul hvis og kun hvis $\sigma^2 = \frac{\mathrm{SSD}_y}{n}$, positiv til venstre for denne værdi og negativ til højre for, følger det at likelihoodfunktionen har entydigt maksimum i punktet

$$\sigma^2_{ML} = \frac{\mathrm{SSD}_y}{n}.$$

Bemærk til senere brug at log–likelihoodens værdi i maksimumspunktet er

$$l(\hat{\mu}, \sigma^2_{ML}) = -\frac{1}{2}\left(n \log \frac{\mathrm{SSD}_y}{n} + n\right).$$

Korrektion af variansestimatet. Grunden til, at vi har brugt den underlige betegnelse σ^2_{ML}, er at man i praksis plejer at bruge en anden estimator, nemlig den man får ved at dividere kvadratafvigelsessummen med $n-1$ i stedet for med n. For denne *korrigerede* estimator vil vi anvende den mere gængse betegnelse

$$\hat{\sigma}^2 = \frac{\mathrm{SSD}_y}{n-1}.$$

En begrundelse for at foretage denne korrektion er, at vi herved får en estimator der netop har middelværdi σ^2, jvf. Ssr. opgave 4.4.2. En estimator med den egenskab, at dens middelværdi netop er den parameter den skal estimere, kalder man *central* eller *middelværdiret* (engelsk: *unbiased*). Dette er helt klart en ønskværdig (omend ikke absolut nødvendig) egenskab ved en estimator.

Et yderligere argument (eller, om man vil, det samme argument i et ekstremt tilfælde) for korrektionen er følgende. For $n = 1$ er den egentlige maksimaliseringsestimator veldefineret og lig med 0. Den korrigerede estimator er udefineret, fordi den involverer division med 0. Når man tænker på, hvor meningsløst det er at udtale sig om variansen på grundlag af en enkelt observation, virker den sidste egenskab måske mest naturlig.

Bemærk at vi ved kvotienttestning benytter likelihoodfunktionens rigtige maksimum $l(\hat{\mu}, \sigma^2_{ML})$ som udregnet ovenfor — altså *ikke* den værdi, man får ved at indsætte det korrigerede variansestimat.

Af Ssr. sætning 7.2.2 følger at SSD_Y er χ^2–fordelt med $n-1$ frihedsgrader og skalaparameter σ^2. Heraf følger (igen) at estimatoren $\hat{\sigma}^2$ er central, og dette fordelingsresultat kan man i øvrigt udnytte til at angive eksakte konfidensgrænser for σ^2.

7.4. Eksakte konfidensgrænser for middelværdien.

Til vurdering af usikkerheden på middelværdiestimatet $\hat{\mu} = \bar{y}$ er det nærliggende at benytte angivelsen

$$\mu = \bar{y} \pm 1.96\sqrt{\frac{\hat{\sigma}^2}{n}}$$

ud fra den primitive betragtning, at når vi ikke kender variansen må vi indsætte estimatet for den i stedet for. Det er overvejelser af denne type, der ligger bag de angivelser af estimationsusikkerhed, vi har benyttet i de diskrete modeller.

Men det kan vi gøre meget bedre her. Af Ssr. sætning 7.2.2 følger jo, at størrelsen

$$T = \frac{\sqrt{n}(\bar{Y} - \mu)}{\sqrt{\frac{1}{n-1}\mathrm{SSD}_Y}} = \frac{\bar{Y} - \mu}{\sqrt{\hat{\sigma}^2/n}}$$

er T–fordelt med $n-1$ frihedsgrader. Hvis vi med $t_{n-1}(97.5)$ betegner 97.5%–fraktilen i denne T–fordeling, har vi derfor med sandsynlighed 0.95

$$-t_{n-1}(97.5) \leq \frac{\mu - \bar{Y}}{\sqrt{\hat{\sigma}^2/n}} \leq t_{n-1}(97.5)$$

eller

$$\bar{Y} - t_{n-1}(97.5)\sqrt{\frac{\hat{\sigma}^2}{n}} \leq \mu \leq \bar{Y} + t_{n-1}(97.5)\sqrt{\frac{\hat{\sigma}^2}{n}}.$$

Sagt i ord: *Eksakte* 95%–konfidensgrænser får man ved at modificere konfidensgrænserne fra tilfældet hvor variansen er kendt, idet variansen σ^2 erstattes med med sit estimatet $\hat{\sigma}^2$, og 1.96 (som er den normerede normalfordelings 97.5%–fraktil) erstattes med 97.5%–fraktilen i T–fordelingen med $n-1$ frihedsgrader.

7.5. Test for simpel middelværdihypotese når variansen er ukendt.

Betragt en simpel hypotese af formen $\mu = \mu_0$. Log–likelihooden som funktion af σ^2 under hypotesen er

$$l(\mu_0, \sigma^2) = -\frac{1}{2}\left(n \log \sigma^2 + \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu_0)^2\right).$$

Udtrykket ligner den delvis maksimerede likelihood (side 62) fra den større model. Den eneste forskel er at SSD_y er blevet erstattet med hypotesens kvadratafvigelsessum

$$\sum_{i=1}^{n}(y_i - \mu_0)^2 = \mathrm{SSD}_y + n(\bar{y} - \mu_0)^2.$$

Heraf følger umiddelbart, at maksimum antages for

$$\hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu_0)^2 = \frac{1}{n}\left(\mathrm{SSD}_y + n(\bar{y}-\mu_0)^2\right).$$

Indsætning af dette estimat i log–likelihooden giver

$$l(\mu_0, \hat{\sigma}_0^2) = -\frac{1}{2}\left(n \log \frac{\mathrm{SSD}_y + n(\bar{y}-\mu_0)^2}{n} + n\right).$$

Kvotientteststørrelsen for hypotesen $\mu = \mu_0$ bliver så

$$-2\log q = 2\left(l(\hat{\mu}, \sigma_{ML}^2) - l(\mu_0, \hat{\sigma}_0^2)\right)$$

$$= 2\left(-\frac{1}{2}\left(n\log\frac{\mathrm{SSD}_y}{n} + n\right) + \frac{1}{2}\left(n\log\frac{\mathrm{SSD}_y + n(\bar{y}-\mu_0)^2}{n} + n\right)\right)$$

$$= n\log\frac{\mathrm{SSD}_y + n(\bar{y}-\mu_0)^2}{\mathrm{SSD}_y} = n\log\left(1 + \frac{n(\bar{y}-\mu_0)^2}{\mathrm{SSD}_y}\right).$$

Det sidste udtryk viser, at kvotientteststørrelsen er en monotont voksende funktion af størrelsen $\frac{n(\bar{y}-\mu_0)^2}{\mathrm{SSD}_y}$; eller, ækvivalent hermed, af størrelsen

$$f = \frac{n(\bar{y}-\mu_0)^2}{\mathrm{SSD}_y/(n-1)}.$$

Men fordelingen af den tilsvarende stokastiske variable

$$F = \frac{n(\bar{Y}-\mu_0)^2}{\mathrm{SSD}_Y/(n-1)} = \left(\frac{\sqrt{n}(\bar{Y}-\mu_0)}{\sqrt{\frac{1}{n-1}\mathrm{SSD}_Y}}\right)^2$$

under hypotesen $\mu = \mu_0$ kender vi fra Ssr. sætning 7.2.2. Ifølge denne sætning er størrelsen

$$T = \frac{\sqrt{n}(\bar{Y}-\mu_0)}{\sqrt{\frac{1}{n-1}\mathrm{SSD}_Y}} = \frac{\bar{Y}-\mu_0}{\sqrt{\hat{\sigma}^2/n}}$$

T–fordelt med $n-1$ frihedsgrader, og kvadratet på den er (jvf. kommentarer til sætning 7.1.4 i Ssr.) F–fordelt med $(1, n-1)$ frihedsgrader. Testet kan derfor udføres enten ved vurdering af, om teststørrelsen f ovenfor er ekstremt stor i en F–fordeling med $(1, n-1)$ frihedsgrader, eller ved tosidet vurdering af størrelsen

$$t = \frac{\bar{y}-\mu_0}{\sqrt{\hat{\sigma}^2/n}}$$

i en T–fordeling med $n-1$ frihedsgrader.

Bemærk, at T–størrelsen ligner den der blev benyttet ved test for simpel hypotese i modellen med kendt varians. Blot er den kendte varians σ^2 blevet erstattet med variansestimatet $\hat{\sigma}^2$, og den normerede normalfordeling er udskiftet med T–fordelingen. T–fordelingen har tykkere haler end normalfordelingen, hvilket kompenserer for den usikkerhed der ligger i estimationen af variansen. For store værdier af frihedsgradsantallet $n-1$, hvor variansestimatet er ret nøjagtigt, ligner T–fordelingen den normerede normalfordeling.

Bemærk også at testet giver godkendelse på niveau $\alpha = 0.05$ hvis og kun hvis μ_0 ligger i det tidligere udledte 95% konfidensinterval.

7.6. Modelkontrol og udregninger i praksis.

Undersøgelse af, om et givet observationssæt $y_1, \ldots, y_n$ kan beskrives ved en normal fordeling, foretages primært ved tegning af et *histogram*. Et passende interval, som indeholder alle observationerne, inddeles i lige store delintervaller. Over hvert interval tegnes en kasse, hvis højde er lig med (eller proportional med) antallet af observationer i intervallet. Herved får man — hvis man har valgt de små intervallers bredde fornuftigt — en tegning, der gerne skulle ligne den normale fordelings tæthed, på nær tilfældig variation.

En mere raffineret metode går ud på at tegne et såkaldt *fraktildiagram* eller *probitdiagram*. Dette går groft sagt ud på at tegne den empiriske fordelingsfunktion for observationerne, men med den lodrette akse skaleret på en sådan måde at normalitet svarer til at man får en ret linie. Mere præcist: Hvis

$$y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}$$

betegner de ordnede observationer, så vil man forvente at der approksimativt (på nær tilfældig variation) gælder

$$r_{(i)} = \Phi\left(\frac{y_{(i)} - \mu}{\sigma}\right) \approx \frac{i}{n+1}$$

fordi punkterne $r_{(i)}$ kan fortolkes som de ordnede, observerede værdier af n uafhængige normeret rektangulært fordelte stokastiske variable, som ifølge de store tals lov vil fordele sig jævnt ud over enhedsintervallet. Ved at anvende Φ^{-1} på begge sider af denne approksimative relation får vi

$$\frac{y_{(i)} - \mu}{\sigma} \approx \Phi^{-1}\left(\frac{i}{n+1}\right)$$

som viser at punkterne

$$\left(y_{(i)}, \Phi^{-1}\left(\frac{i}{n+1}\right)\right) \ , \quad i = 1, \ldots, n$$

— som er dem man afsætter i fraktildiagrammet — må forventes at ligge på en ret linie. Væsentlige afvigelser fra normalitet vil give sig udslag i krumning eller S–form.

Der findes specielt papir til tegning af fraktildiagrammer, såkaldt "normalfordelingspapir" eller (med en totalt forvirrende betegnelse, som formodentlig skyldes en oversættelsesfejl) "sandsynlighedspapir", hvor den lodrette akse på forhånd er inddelt svarende til transformation med Φ^{-1}.

Både for histogrammet og (i særdeleshed) for fraktildiagrammet gælder, at det kræver ret stor erfaring at vurdere, om der er en signifikant afvigelse fra normalitet. Hvis man er i tvivl kan man sætte en computer til at simulere et antal tilsvarende tegninger for "ægte" (altså simulerede, se Ssr. side 114–115) normalfordelte observationer, for at få en fornemmelse for hvor store de tilfældige udsving kan være. Det man skal kikke efter er især skævhed af fordelingen, som ytrer sig ved at den ene hale er væsentligt tykkere end den anden. Skævheder kan man i nogle tilfælde rette op ved at transformere observationerne. Et særdeles almindeligt fænomen er, at et sæt af ikke–negative observationer (f. eks. kemiske koncentrationer og mange typer af økonomiske data) udviser en skævhed, som svarer til en ophobning af værdier nær 0 og en forholdsvis tyk højre hale. En sådan skævhed kan man i heldige tilfælde fjerne ved at erstatte alle tallene med deres logaritmer.

Angående udregningerne skal siges, at det betaler sig at bruge formlerne

$$\bar{y} = \mathrm{S}_y/n \text{ og } \mathrm{SSD}_y = \mathrm{SS}_y - \mathrm{S}_y^2/n,$$

hvor

$$\mathrm{S}_y = \sum y_i \ , \ \ \mathrm{SS}_y = \sum y_i^2 \ .$$

Den besværlige del af udregningerne reducerer hermed til beregning af sum og kvadratsum af alle observationerne, og det kan man på de fleste lommeregnere klare uden at skulle indtaste tallene mere end én gang. Men man er selvfølgelig nødt til at gentage hele forestillingen, hvis man vil sikre sig mod indtastningsfejl.

Hvis alle tallene ligger i et interval af formen $[S, S+\Delta]$, hvor S er positiv og stor i forhold til Δ, kan der opstå numeriske problemer ved udregning af SSD, fordi højre side af formlen er en differens mellem to næsten lige store tal. Det kan man løse ved på forhånd at trække en passende konstant fra alle tallene. Ligeledes kan man, for at undgå besværet med indtastning af decimalpunktum, gange alle tallene med en passende titalspotens under indtastningen, og bagefter korrigere resultaterne på indlysende måde.

Eksempel 7.1. (Kilde A. Hald: Statistiske Metoder, Akademisk Forlag 1968.) Ved fabrikation af hørgarn udtages regelmæssigt prøver til bestemmelse af garnets styrke. Nedenfor gengives 50 sådanne prøveresultater, ordnet efter størrelse.

Brudstyrker af hørgarn i kg

1.40	1.52	1.63	1.69	1.73	1.73	1.78	1.89	1.92	1.95
1.98	1.99	2.02	2.03	2.07	2.12	2.12	2.13	2.15	2.16
2.20	2.23	2.26	2.30	2.31	2.32	2.35	2.36	2.37	2.39
2.40	2.40	2.44	2.47	2.50	2.52	2.55	2.60	2.63	2.64
2.65	2.71	2.74	2.77	2.79	2.86	2.92	2.94	3.02	3.30

På side 69 ses histogram (ved opdeling i intervaller af længde 0.2) og fraktildiagram for disse data. Der er bestemt ikke noget her, som giver anledning til tvivl om normalfordelingsantagelsen. På side 70 er de to tegninger gengivet i miniatureformat sammen med 15 tegninger af samme slags for computersimulerede datasæt af samme størrelse. Det ser ud som om fordelingen af hørgarns brudstyrke overgår selveste den normale fordeling i normalitet. Men det giver selvfølgelig ikke mening. Forklaringen er snarere at dette datasæt ikke er helt tilfældigt valgt.

Vi udregner

$$\mathrm{S} = 1.40 + 1.52 + \cdots + 3.30 = 114.95$$

og

$$\mathrm{SS} = 1.40^2 + 1.52^2 + \cdots + 3.30^2 = 272.5443$$

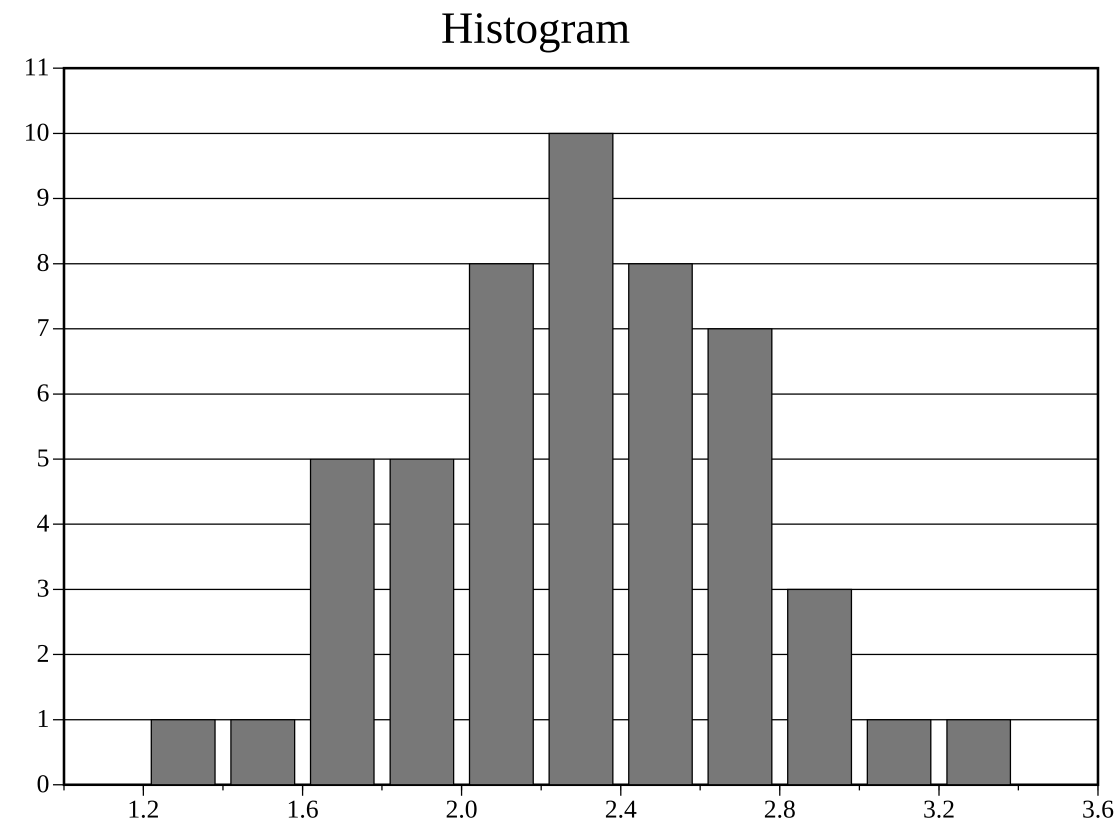
og får derefter

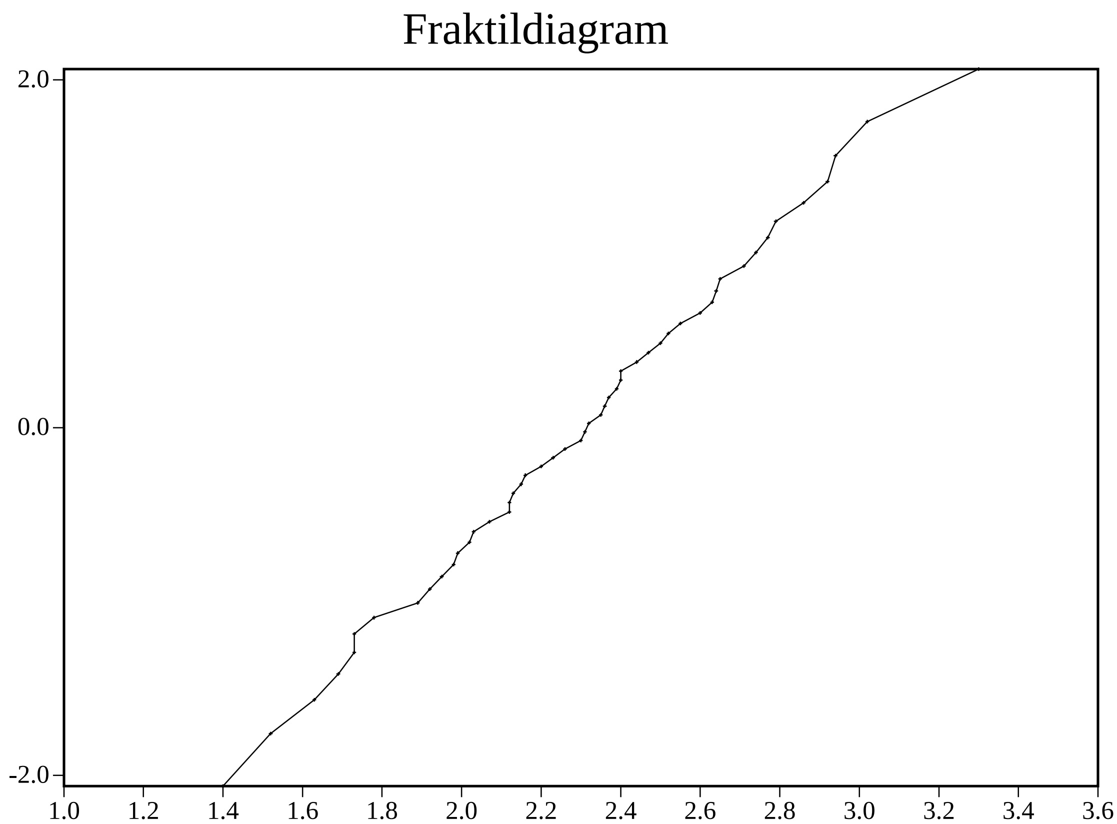
$$\begin{aligned} \hat{\mu} = \bar{y} &= 114.95/50 = 2.299, \\ \mathrm{SSD} &= 272.5443 - 114.95^2/50 = 8.2743, \\ \hat{\sigma}^2 &= 8.2743/49 = 0.1689, \\ \hat{\sigma} &= \sqrt{0.1689} = 0.4109. \end{aligned}$$

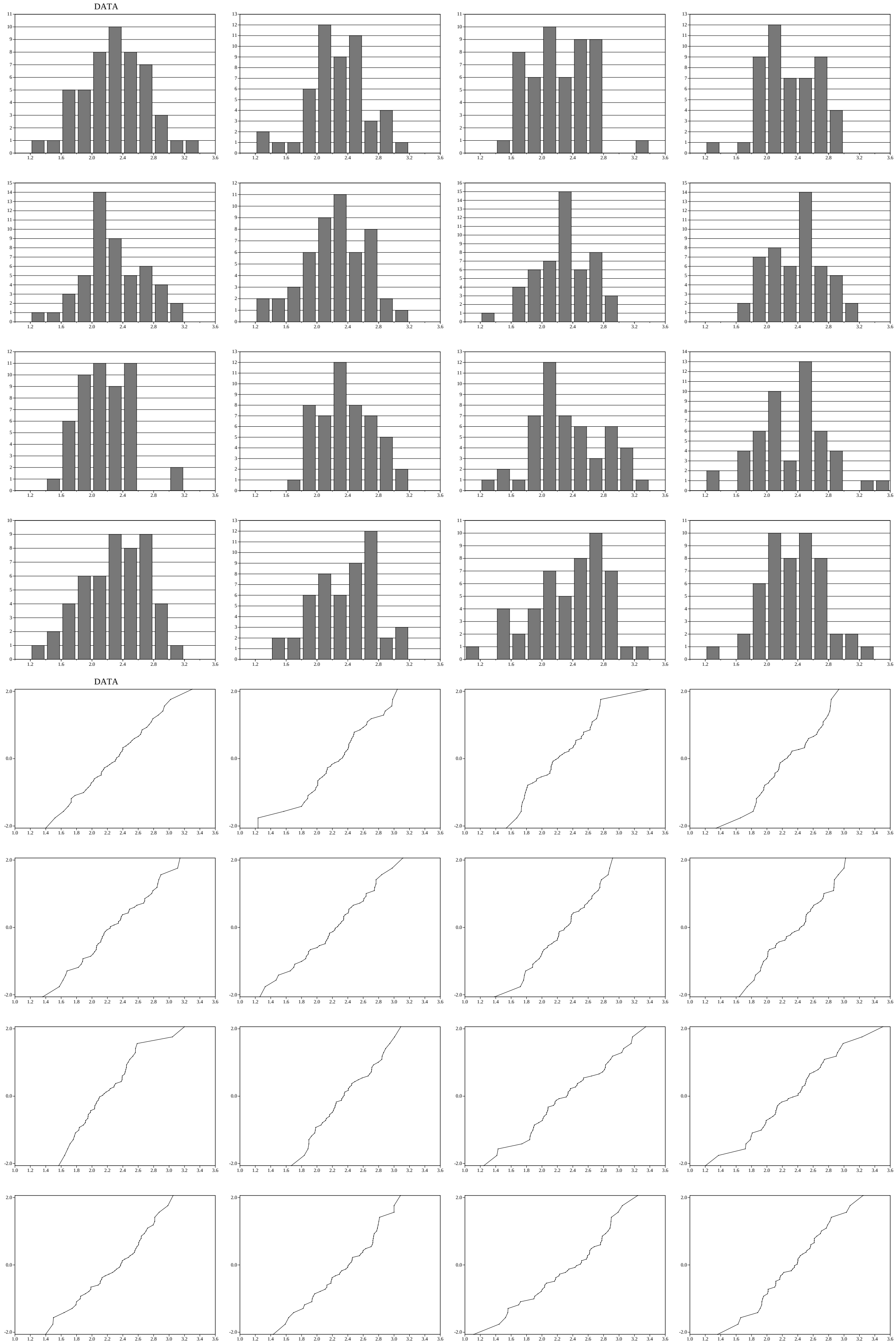
95% konfidensgrænserne for middelværdiestimatet bliver (idet 97.5%–fraktilen i T–fordelingen med 49 frihedsgrader er 2.010)

$$2.299 \pm 2.010\sqrt{\frac{0.1689}{50}} = 2.299 \pm 0.117,$$

Vi kan altså med “95% sikkerhed” hævde at fordelingens sande middelværdi ligger i intervallet [2.182,2.416].


Histogram
11
10
9
8
7
6
5
4
3
2
1
0
1.2
1.6
2.0
2.4
2.8
3.2
3.6


Fraktildiagram
2.0
0.0
-2.0
1.0
1.2
1.4
1.6
1.8
2.0
2.2
2.4
2.6
2.8
3.0
3.2
3.4
3.6


DATA
DATA