Deskriptiv statistik ud fra berømte måleserier

Newcombes måling af lysets hastighed[1]

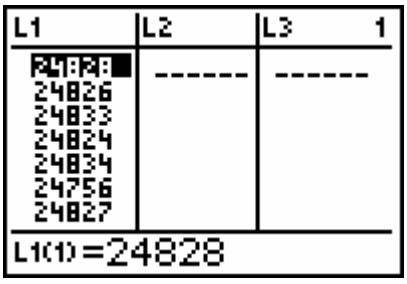
Newcombe arbejdede sammen med Michelson i slutningen af forrige århundrede og indførte nye teknikker til målingen af lysets hastighed. Det resulterede bl.a. i en serie på 66 præcisionsmålinger af lysets hastighed, som Newcombe foretog i perioden juli-september 1882 ved at måle returtiden for en lysstråle, der blev sendt ned af Potomac-floden og tilbage igen. Derved tilbagelagde lyset i alt en distance på knap 7½ kilometer, hvilket derfor tog i størrelsesordenen

$$\frac{7.5\,\text{km}}{300000\,\text{km/s}} = 2.5\cdot 10^{-5}\text{s} = 25000\,\text{ns} \quad .$$

De 66 målinger fremgår af det følgende skema med de første 11 målinger i den første kolonne, de næste 11 målinger i den næste kolonne osv.:

24828	24822	24836	24826	24828	24828
24826	24824	24832	24830	24827	24824
24833	24821	24836	24832	24831	24825
24824	24825	24828	24836	24827	24832
24834	24830	24825	24826	24826	24825
24756	24823	24821	24830	24833	24829
24827	24829	24828	24822	24826	24827
24816	24831	24829	24836	24832	24828
24840	24819	24837	24823	24832	24829
24798	24824	24825	24827	24824	24816
24829	24820	24828	24827	24839	24823

Det kan jo godt virke lidt uoverskueligt med disse mange data, så for at få et overblik over dem taster i dem ind i en liste på vores lommeregner[2]:



Spørgsmålet er nu, hvordan vi skal forholde os til disse data? Hvordan kan vi fx trække en rimelig værdi ud for lysets returtid og derigennem finde en rimelig værdi for lysets hastighed? Som altid kan det da betale sig først at kigge lidt på dataene før vi kaster os ud i vilde beregninger.

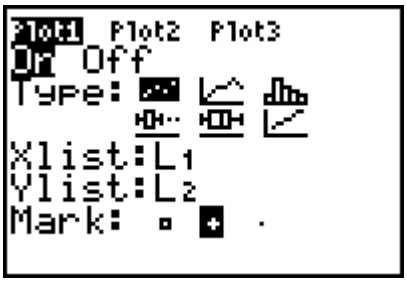
---

[1] Efter More and McCabe: Introduction to the Practice of Statistics, 2nd ed, Freeman, side 3.

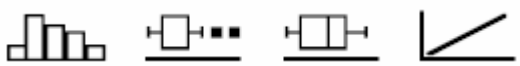
[2] Du kan også indsætte disse data via den ovenstående tabel eller fra den regnearksfil, der er lagt ud på EMU sammen med dette dokument.

Grafisk inspektion af data

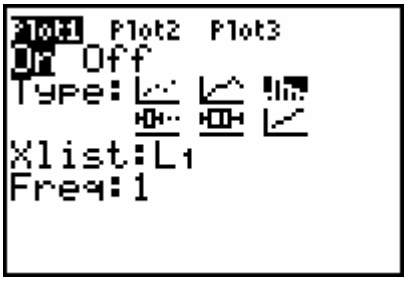
Vi vil nu foretage forskellige grafiske afbildninger af Newcombes data for at forsøge at forstå, hvordan vi kan trække en rimelig præcis værdi ud for lysets returtid. Sådanne statistiske plots indskrives i 2nd Y= , dvs. i STAT PLOT-menuen:



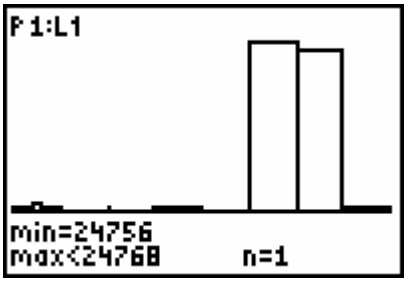
De to første plottyper: *Punkt-grafer* og *linje-grafer* benyttes til 2-dimensionale datasæt, så vi skal have fat i de fire sidste plottyper, dvs. *histogrammet*, det *udvidede boksplot*, det *almindelige boksplot* og *normalfordelingsplottet*:



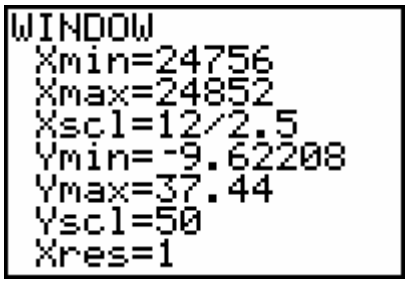
Vi starter med **histogrammet**:

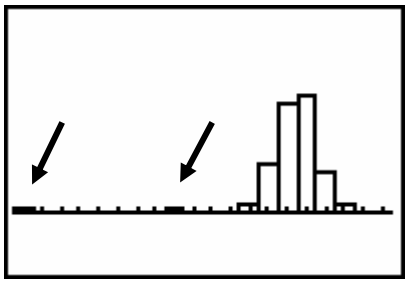


Ligeså snart vi har valgt histogrammet ændres de nederste linjer til at vi nu kun kan vælge en x-liste og en hyppighedsliste (idet *frequency* på engelsk betyder *hyppighed* på dansk). Den sidste er valgfri, og hvis vi ikke oplyser en hyppighedsliste, sættes alle hyppighederne bare til 1, som vist! Da vores datasæt netop ligger i listen L_1, er vi altså færdige med at indstille plottypen. For at få den tegnet vælger vi nu som sædvanlig ZoomStat (dvs. Zoom 9) og overlader det til maskinen at sætte plottet rimeligt op:



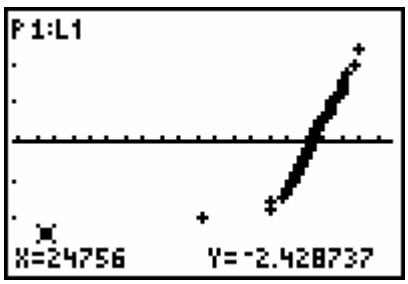
Som det ses er der afsat rimelig plads til oplysningerne fra en trace. Vi ser også at næsten alle målingerne ligger i to stort set lige store blokke, og at der er afsat plads til 8 blokke i alt. Det er standard for et histogram tegnet med ZoomStat. Men vi kan selvfølgelig nemt gå ind i WINDOW-menuen og ændre på dette. Læg mærke til at bredden af de enkelte blokke styres af Xscl. Det er lidt ukonventionelt, men meget praktisk. For at sætte antallet af blokke op, skal vi altså blot sætte Xscl tilsvarende ned. Hvis vi fx deler Xscl med 2.5 får vi altså i stedet 20 blokke:

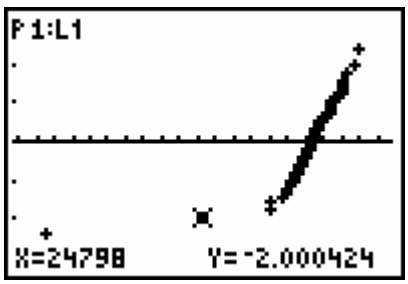




Vi får da en mere klokkeformet fordeling men stadig med to tydelige undtagelser til venstre for klokken. Noget tyder altså på at de to laveste målinger er atypiske. Den klokkeformede fordeling tyder også på at fordelingen godt med tilnærmelse kunne være normalfordelt. Det kan vi checke med et **normalfordelingsplot**. Vi ændrer derfor plottypen:

Igen ændrer det drastisk på de følgende indtastningslinjer. Først skal vi oplyse en dataliste (og her er L_1 jo igen netop det rigtige valg). Men så derefter kan vi vælge mellem to forskellige typer normalfordelingsplot, idet vi kan afsætte dataene vandret ud af x-aksen eller lodret op af y-aksen. Da vi har tradition for det første godkende vi maskinens forslag, og beder igen maskinen om selv at tilrettelægge plottet med ZoomStat:





Igen skinner de to laveste målinger i øjnene, idet de tydeligt bryder med det retlinjede mønster, der ellers præger dataene. Det ser altså virkeligt ud, som om de to laveste målinger er problematiske!

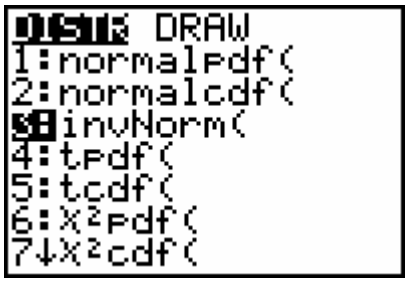
Inden vi forlader normalfordelingsplottet vil vi lige komme med en enkelt kommentar til, hvordan det bliver til: Den laveste måling, her 24756, har den kumulerede frekvens 1/66 og afsættes derfor med y-koordinaten

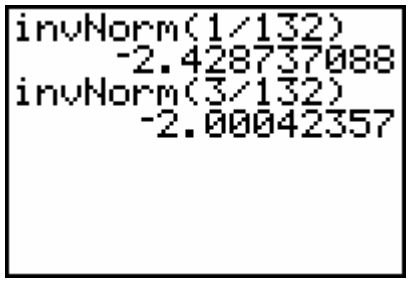
invNorm(1/132) = -2.4287...,

idet 1/132 er midtpunktet i det første kumulerede frekvensinterval [0;1/66]. Den næstlaveste måling, her 24798, har den kumulerede frekvens 2/66 og afsættes derfor med y-koordinaten

invNorm(3/132) = -2.0004...,

idet 3/132 er midtpunktet i det næste kumulerede frekvensinterval [1/66;2/66] osv.





Atypiske målinger: Hvordan kan vi nu indføre et passende kriterium for hvornår målinger er så atypiske, at de bør give anledning til særlige overvejelser? Her kan det være nyttigt med et **udvidet boksplot**. Det bygger på en undersøgelse af kvartilerne for det pågældende datasæt. Først må vi derfor forstå kvartilbegrebet, der bygger på en amerikansk konvention, og derved for eksempel adskiller sig fra en engelsk konvention, der igen adskiller sig fra en dansk konvention. Kvartiler er altså ikke noget særligt veldefineret begreb indenfor international statistik! Først definerer vi *medianen* Med som den midterste observation, hvis antallet af observationer er ulige. Ellers definerer vi det som *gennemsnittet af de to midterste observationer*, hvis antallet af observationer er lige. Medianen behøver altså ikke selv være en observation! Så definerer vi første halvdel af datasættet, som mængden af de observationer, der går forud for medianen. Tilsvarende definerer vi sidste halvdel af datasættet som mængden af de observationer, der følger efter medianen. Første kvartil Q1 er så medianen for første halvdel af datasættet og sidste kvartil Q3 er medianen for sidste halvdel af datasættet. Her er et par eksempler:

Eksempel 1: Datasættet {2, 5, 7, 8, 10, 12} har medianen Med = (7+8)/2 = 7.5 (som er gennemsnittet af de to midterste observationer). Første halvdel af datasættet består så af observationerne {2, 5, 7}, hvorfor den første kvartil er Q1 = 5, mens sidste halvdel af datasættet består af observationerne {8, 10, 12}, hvorfor den sidste kvartil er Q3 = 10.

Eksempel 2: Datasættet {1, 3, 3, 5, 7, 7, 10, 12, 12} har medianen Med = 7 (som er den midterste observation). Første halvdel af datasættet består så af observationerne {1, 3, 3, 5}, hvorfor den første kvartil er Q1 = (3+3)/2 = 3, mens sidste halvdel af datasættet består af observationerne {7, 10, 12, 12}, hvorfor den sidste kvartil er givet ved Q3 = (10+12)/2 = 11.

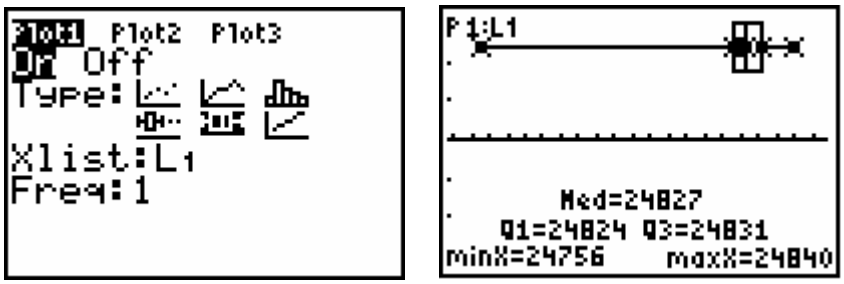
I et **boksplot** afsætter man nu alle observationerne i forhold til de **fem statistiske nøgletal**:

mindste observation, første kvartil, medianen, sidste kvartil, største observation

minX - Q1 - Med - Q3 - maxX

Midterområdet fra første til sidste kvartil, dvs. *kvartilboksen*, indeholder derfor (stort set) halvdelen af observationerne. Forskellen mellem første og sidste kvartil, dvs. tallet Q3 - Q1 kaldes *kvartilbredden*. Det angiver altså kvartilboksens længde.

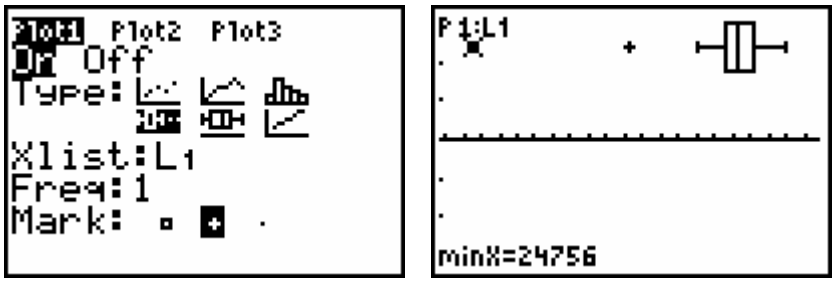
Vælger vi det almindelige boksplot, fås derfor (her vist som en montage af fem plots):



Den engelske statistiker Tukey foreslog nu i 70'erne at man skulle anvende den følgende regel for *atypiske observationer*:

Tukeys regel
En observation er **atypisk**, hvis dens afstand fra den nærmeste kvartil overstiger 1½ kvartilbredde

Der er forskellige måder at begrunde denne regel på, men her noterer vi os blot at den minder om definitionen på et exceptionelt udfald, som er et udfald, hvis afstand til middelværdien er større end 3 spredninger. I det **udvidede boksplot** afsættes atypiske observationer nu som enkeltpunkter:



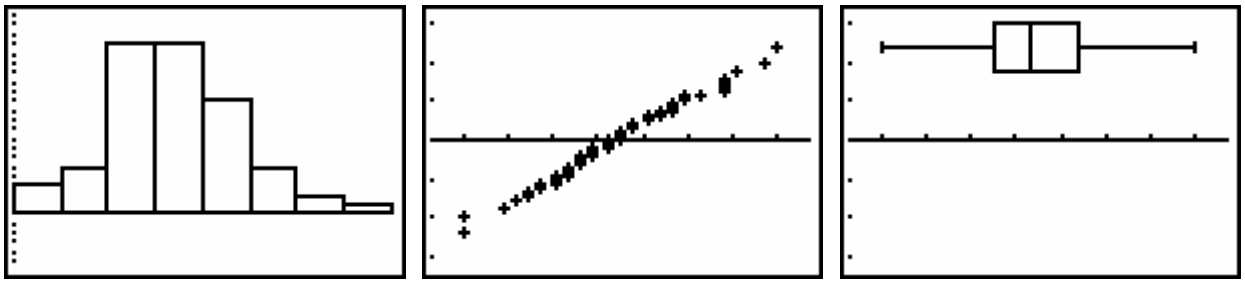
Begge de to laveste observationer er altså klart atypiske! Spørgsmålet er så blot, hvad vi skal gøre ved dem? Hvis de skyldes fejl i målingerne skal vi selvfølgelig smide dem ud, men de kunne også rumme interessante oplysninger! Den klassiske fejl er at smide dem ud pr automatik: Fx vil mange edb-overvågningsprogrammer smide atypiske målinger ud, fordi man går ud fra at de repræsenterer fejlmålinger, der kan forstyrre den efterfølgende databehandling. Da man benyttede satellitter til at holde øje med ozon-koncentrationen i atmosfæren smed man således i en årrække sommerens ozon-tal for sydpolen ud, da de var atypisk små. Først senere gik det

op for forskerne, at der ikke var tale om fejlmålinger, men om et reelt og problematisk hul i ozonlaget.

Newcombe selv besluttede at smide den laveste væk, men beholde den næstlaveste. Vi vælger at smide dem begge ud, og fortsætter herefter med dataanalysen!

Den hurtigste metode til at finde de to mindste observationer i datalisten er nok at sortere listen efter observationernes størrelse. Det gøres ved hjælp af kommandoen SortA(fra [2nd] [LIST]- menuen (hvor A står for Ascending):

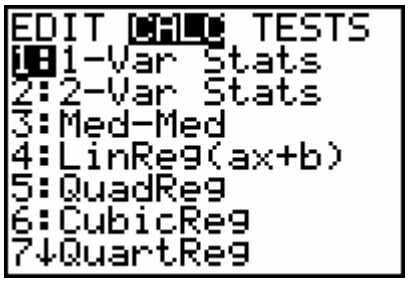
Efter at de er smidt ud, får vi nu pæne histogrammer, normalfordelingsplot og udvidede boksplot. De bekræfter vores forventning om, at de resterende 64 målinger opfører sig pænt, dvs. som målinger med tilfældige fejl, der klumper sig sammen om den 'sande værdi':

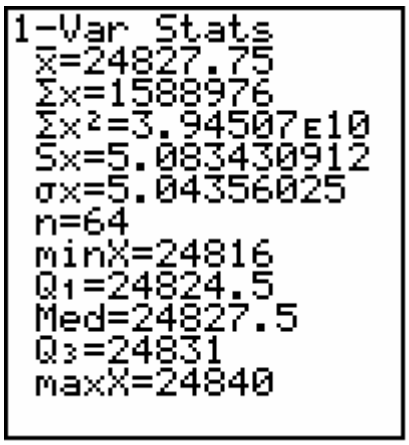


Således opmuntrede er vi nu grundigt forberedte til at diskutere den **numeriske analyse** af dataene!

Enkeltvariabel statistik

Den numeriske analyse af datasæt styres via kommandoen 1-var-stats på CALC-menuen under STAT:





Vi får da først *gennemsnittet* og *spredningen* oplyst for dataene:

$$\bar{x} = 24827.75 \qquad \sigma x = 5.04356025$$

NB: Der er to spredningsmål, både standardafvigelsen Sx og spredningen σx . Vi vil ikke her gå i dybden med forskellen på de to spredningsmål.

Tilsvarende får vi også *medianen* og *kvartilerne* oplyst:

$$Q1 = 24824.5 \qquad Med = 24827.5 \qquad Q3 = 24831$$

Det minder meget om de foregående oplysninger: medianen har næsten samme værdi som middelværdien og kvartilbredden på 6.5 er af samme størrelsesorden som spredningen. De to typer *deskriptorer* har hver deres fordele og ulemper: Middelværdien er *demokratisk* og inddrager alle observationerne på lige fod i udregningen af gennemsnittet. Medianen derimod er *robust* og påvirkes ikke synderligt af en enkelt eller to fejlmålinger. Men da vi har smidt de atypiske målinger ud, synes det rimeligt at bruge middelværdien som det bedste bud på den 'sande returtid'.

$$\text{målt returtid} \approx 24827{,}75$$

Men hvor præcis er den? Da vi har mange observationer til rådighed synes det rimeligt at tro, at gennemsnittet er behæftet med langt mindre usikkerhed end de enkelte observationer. De enkelte observationer har usikkerheden

$$\sigma_x = 5.04356025$$

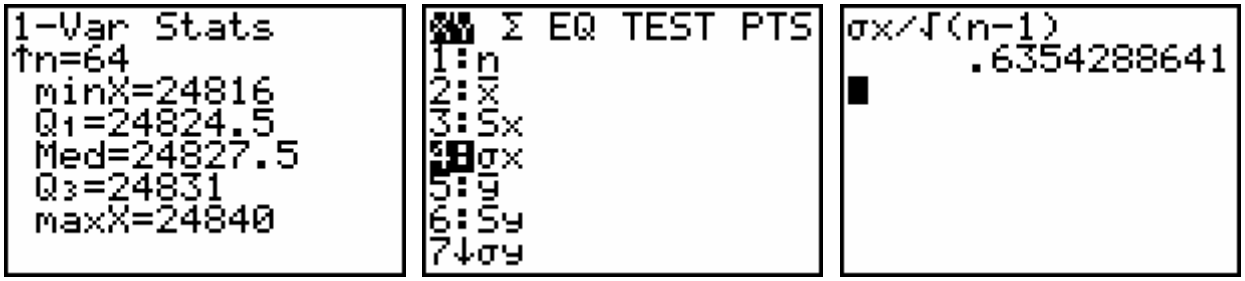
Standardreglen i statistik om virkningen af at tage gennemsnittet af n uafhængige målinger er nu at nedsætte denne usikkerhed med faktoren $\sqrt{n-1}$ svarende til kvadratroden af *antallet af frihedsgrader*:

$$\text{usikkerheden på gennemsnittet} = \frac{\text{usikkerheden på den enkelte måling}}{\sqrt{\text{antal frihedsgrader}}}$$

Det kan godt virke lidt ejendommeligt, at skulle tage hensyn til antallet af frihedsgrader, men forestil dig at du skulle beregne usikkerheden på gennemsnittet ud fra blot en enkelt observation. Det giver tydeligvis ingen mening. Der skal mindst 2 observationer til, før vi kan vurdere usikkerheden. Man siger derfor, at *det er observationerne* 2,..,*n*, *der bærer informationen om usikkerheden*, og derfor er antallet af frihedsgrader netop *n*–1. I dette tilfælde er usikkerheden på gennemsnittet derfor givet ved

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{63}} = 0.6354288641$$

Det udregnes på lommeregneren ved at hente den statistiske variabel på VARS-Statistics skærmen:



På den baggrund kan vi altså slutte, at vi nok kun skal opgive returtiden med højst 1 decimal:

$$\text{målt returtid} = 24827{,}8 \pm 0{,}6$$

Men er det så virkelig et 'sandt resultat'? Det kan man ikke afgøre statistisk! Sagen er at Newcombes målinger var de første af sin art og byggede på delikate kalibreringer af hans apparatur. Kun efterfølgende og mere sofistikerede målinger kan vise, hvor nøjagtige Newcombes kalibreringer egentlig var. Sådanne senere målinger blev foretaget af bl.a. Michelson i 1927. I dag kender vi lysets hastighed langt mere præcist end Newcombe – faktisk så præcis, at vi bruger dens værdi som en af de grundlæggende eksakte værdier i SI-systemet:

$$c = 299792458 \text{ m/s}$$

Med den viden vi har i dag kan vi regne os frem til, at Newcombe burde have fået returtiden:

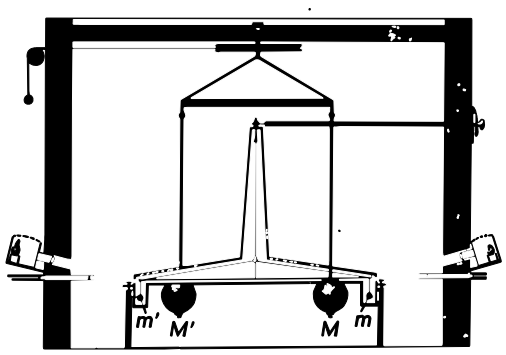
$$\text{Sand(!) returtid} = 24833.02 \text{ ns}$$

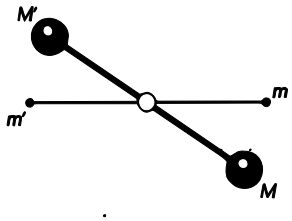
Den ligger klart uden for Newcombes usikkerhedsinterval. Faktisk er forskellen overraskende stor:

$$\frac{24833.02 - 24827.75}{0.6354288641} = 8.29...$$

dvs. den sande værdi ligger mere end 8 standardafvigelser fra Newcombes resultat. Det er altså højst exceptionelt, så i dag ved vi derfor, at Newcombe havde problemer med kalibreringen af sit udstyr. Men set med samtidens øjne var hans målinger et afgørende fremskridt!

Projekt: **Cavendish' målinger af Jordens densitet**[3]





I 1798 fandt den engelske fysiker Cavendish Jordens (gennemsnitlige) densitet ved omhyggelige målinger af gravitationskonstanten G udført med en præcis torsionsbalance. Cavendish udtrykte jordens densitet i forhold til vands densitet og opnåede i 29 successive målinger de følgende resultater:

5,50	5,47	5,29	5,55	5,75	5,27
5,57	4,88	5,34	5,34	5,29	5,85
5,42	5,62	5,26	5,30	5,10	5,65
5,61	5,63	5,44	5,36	5,68	5,39
5,53	5,07	5,46	5,79	5,58	***

Undersøg disse data grafisk og numerisk med henblik på bl.a. at diskutere de følgende spørgsmål:

Er der atypiske målinger?

Er målingerne med tilnærmelse normalfordelte?

Hvor præcis en værdi kan man udtrække for Jordens densitet fra disse data?

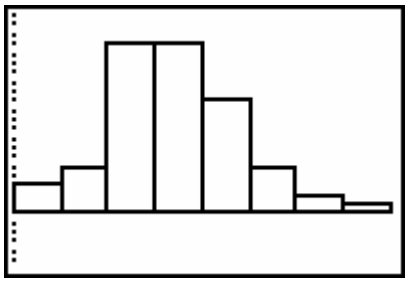
Hvad er den moderne værdi for Jordens densitet?

Stemmer den overens med Cavendish' målinger?

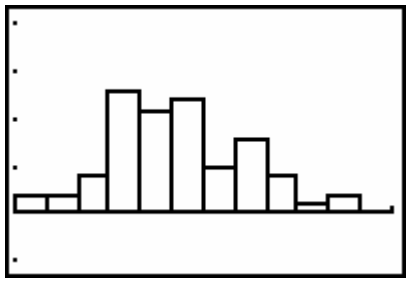
[3] Efter More and McCabe: Introduction to the Practice of Statistics, 2nd ed, Freeman, side 28.

Avancerede emner

1) Normalfordelingsapproksimationen



Zoom 9



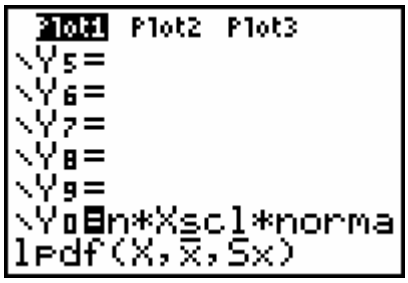
med 12 bokse

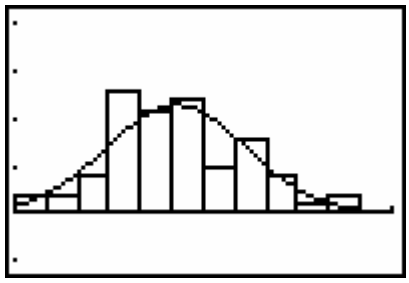
Som vi har set, kan vi tegne histogrammer over måledata nemt og smertefrit, men hvordan får vi indtegnet normalfordelingsapproksimationen til histogrammet? Der skal vi udnytte, at de enkelte kasser i histogrammet har bredden Xscl. Det samlede areal under histogrammet er derfor givet ved n*Xscl, hvor n er antallet af observationer. Når vi skal indtegne sandsynlighedstætheden for normalfordelingen, der jo har det samlede areal 1, skal den derfor ganges med n*Xscl. Så skal vi have fat i nogle rimelige bud på middelværdien og spredningen for normalfordelingen. Men dem kan vi jo få fra en enkeltvariabelstatistik. Det fineste er da at bruge standardafvigelsen Sx som bud på spredningen, men det gør ikke megen forskel om vi skulle komme til at bruge σx i stedet.

Før vi tegner normalfordelingsapproksimationen skal vi altså huske at udføre en enkeltvariabel statistik, dvs. 1-var Statistics på måledataene! Derefter indskriver vi den følgende graffunktion

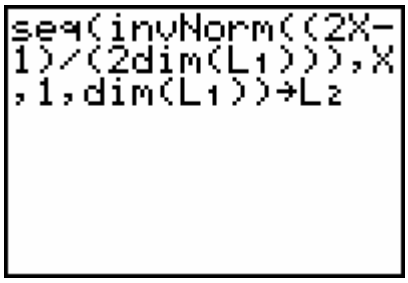
$$n*Xscl*Normalpdf(X, \bar{x}, Sx)$$

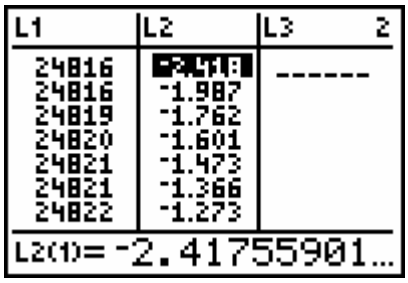
i graflisten (hvor pdf står for point distribution function). Her henter vi de enkelte statistiske deskriptorer under VARS -Window- og VARS -Statistics-menuen:



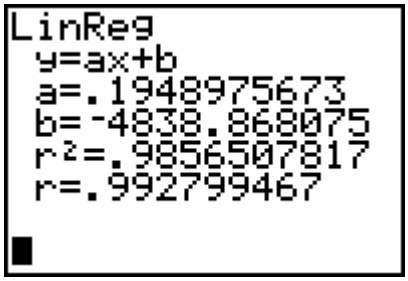


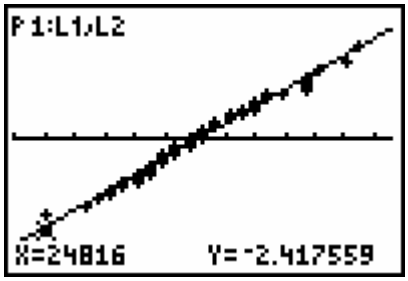
Spørgsmålet er så selvfølgelig om denne approksimation er overbevisende! Men det er jo det samme problem vi har med normalfordelingsplottet: Hvornår ligner det i tilstrækkelig grad en ret linje. Det ville altså være rart om vi kunne checke lidt mere håndfast, hvor god normalfordelingsapproksimationen egentlig er. Men det kunne vi jo fx gøre ved at udføre en lineær regression på normalfordelingsplottet. Her er X-dataene ikke noget problem. De skal bare sorteres med en SORTA(kommando som forklaret tidligere. Problemet er blot at vi ikke har direkte adgang til Y-dataene! Vi må derfor beregne dem selv, og da sandsynlighedsfordelingerne *ikke* er liste-robuste, dvs. de virker ikke på lister, må vi selv beregne dem én for én med en sekvens-kommando (fra 2nd List-Ops-menuen):





Dermed er vejen banet for at få udført en almindelig lineær regression på L_1 og L_2 og se, hvor godt det passer med en ret linje:

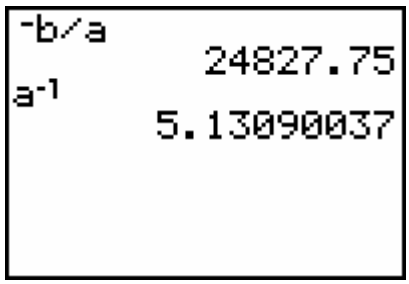




Det ser jo rigtigt pænt ud med en forklaringsgrad på 98,56%. Vi kan så også trække de relevante parametre ud for normalfordelingsapproksimationen. Den rette linje har ligningen

$$Y = \frac{X-\mu}{\sigma} = \frac{1}{\sigma}\cdot X - \frac{\mu}{\sigma}$$

Spredningen er altså givet ved den reciprokke hældning a^{-1}, mens middelværdien er givet ved $-b/a$:



Middelværdien er ikke overraskende givet ved den gamle kending 24827.75, mens spredningen på 5.13090037 ligger en anelse højere end de tidligere skøn baseret på enkelt variabel statistik: Sx = 5.0834.. og σx = 5.0435..! Men det ændrer selvfølgelig ikke væsentligt ved de tidligere konklusioner.

2) Tukeys regel

Ifølge Tukeys regel er en observation atypisk, hvis dens afstand til nærmeste kvartilværdi overstiger 1,5 kvartilbredder. Hvordan kan man nu begrunde en sådan regel? Den nemmeste måde at forstå den på er at se på et normalfordelt observationsmateriale. Hvis datasættet er standardnormalfordelt med middelværdi 0 og spredning 1, så ligger første kvartil efter de første 25% af observationerne, dvs.

$$Q1 = \text{invNorm}(0.25) = -0.6744897495$$

eller ca. –2/3. Tilsvarende ligger sidste kvartil efter de første 75% af observationerne, dvs.

$$Q3 = \text{invNorm}(0.75) = 0.6744897495$$

eller ca. 2/3.

I runde tal er kvartilbredden for en standardnormalfordeling derfor 4/3. Halvanden kvartilbredde er derfor i runde tal givet ved $3/2 \cdot 4/3 = 2$. Vi ser derfor at for et normalfordelt observationsmateriale svarer 1½ kvartilbredde netop til 2 standardafvigelser.

Der er tradition for at regne udfaldene *indenfor to standardafvigelser* for *normale*, og tilsvarende regne observationerne *udenfor tre standardafvigelser* for *exceptionelle*. Her imellem ligger en *gråzone*, men jo tættere vi kommer på de exceptionelle udfald, jo mere atypiske er observationen selvfølgelig. Da vi lægger 1½ kvartilbredde til kvartilen vil en atypisk observation ifølge Tukeys regel for *normalfordelte* observationer derfor ligge $2\frac{2}{3}$ standardafvigelser fra middelværdien. De er altså ikke helt exceptionelle, men alligevel tilstrækkeligt tæt på til at retfærdiggøre reglen!

Hvis omvendt observationerne er ligefordelte over intervallet [0; 1], så optræder alle observationerne jo på lige fod, og der bør derfor *ikke* kunne forekomme atypiske observationer i et sådant datasæt! Men for en sådan ligefordeling er kvartilbredden jo netop ½ og den sidste kvartil er ¾. Vi ser derfor at eventuelle store atypiske udfald skal være større end

$$\frac{3}{4} + \frac{3}{2} \cdot \frac{1}{2} = \frac{3}{2}$$

dvs. et godt stykke uden for intervallet [0; 1]. Reglens sikrer altså at der ikke forekommer atypiske observationer i ligefordelte observationssæt.

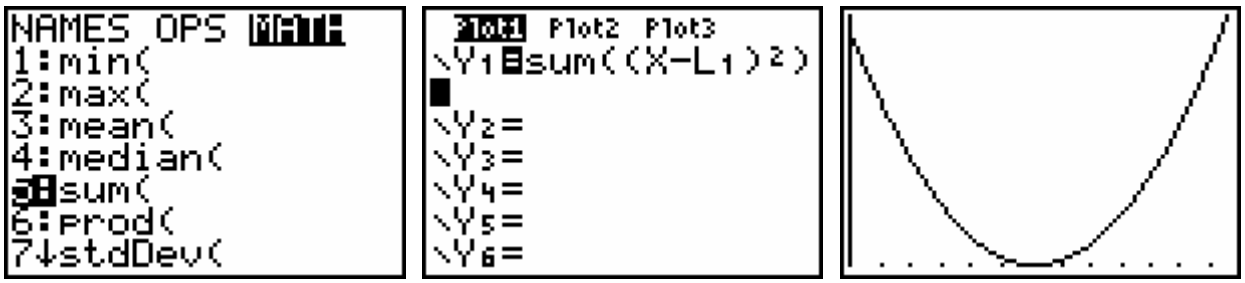
3) Middelværdien som punktet med det mindste *samlede afstandskvadrat.*
Vi kan give en avanceret karakterisering af *middelværdien* på den følgende måde: Vi ser på vores datasæt $\{x_1, x_2, x_3, ..., x_n\}$ og stiller os det følgende spørgsmål:

> Hvordan skal vi vælge tallet x, så det samlede afstandskvadrat til observationerne $\{x_1, x_2, x_3, ..., x_n\}$, dvs. summen
> $$y = (x - x_1)^2 + (x - x_2)^2 + (x - x_3)^2 + ... + (x - x_n)^2$$
> er mindst mulig?

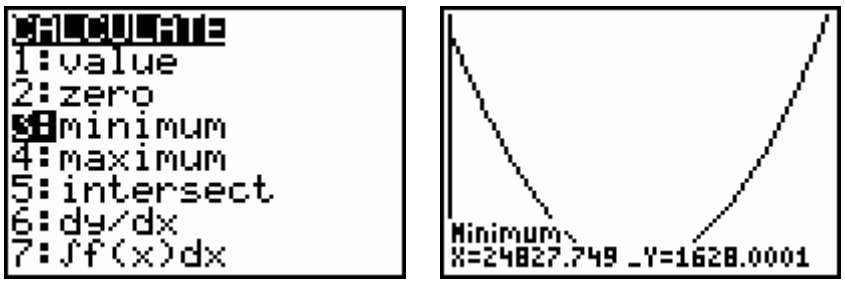
Vi kan få en fornemmelse for problemstillingen ved at undersøge den på grafregneren. Her ligger datasættet i listen L_1 og den ovenstående sum af afstandskvadrater defineres derfor på følgende måde:

$$Y_1 = \text{sum}((X - L_1)^2)$$

Men så kan vi jo tegne *grafen* for summen og derved bestemme dens minimum. Når vi skal indtaste det ovenstående udtryk skal vi huske på, at alle liste-kommandoerne ligger på [2nd] LIST-menuen:

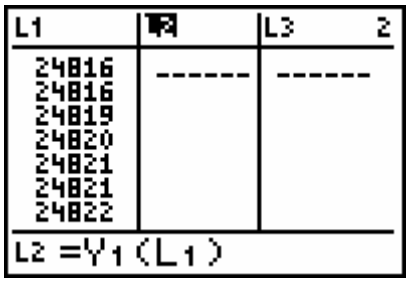


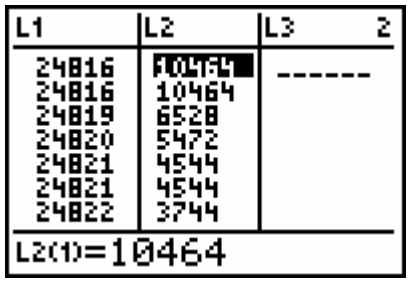
Grafen er tegnet med ZoomFit over et X-interval, som rummer såvel den mindste som den største observation. Den tager lidt tid om at blive tegnet, men det skyldes jo at udregningen af hver eneste Y-værdi er en lidt langsommelig affære, da det jo er en sum af 64 led! Vi kan så bestemme minimumspunktet ved at gå ind på [2nd]-CALC-menuen:

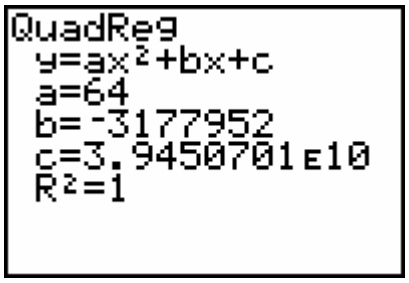


Det kunne jo godt se ud, som om vi har fanget middelværdien $\bar{x} = 24827.75$ og at den tilhørende sum af afstandskvadrater er givet ved sådan ca. 1628. De grafiske rutiner i CALC-menuen arbejder jo kun med en endelig præcision.

Men kig lige på grafen en gang til: Det kunne godt ligne en parabel! Faktisk er det nemt at checke numerisk, at det ikke bare ligner en parabel: Det er en parabel! Hertil skal vi blot have konstrueret en tabel over de tilhørende Y-værdier:



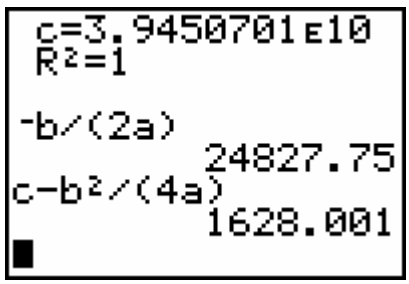




Der er altså netop tale om en perfekt parabel, og oven i købet med simple koefficienter, idet

$$a = 64,\ b = -3177952 \text{ og } c = 3.9450701\text{û}10\ .$$

Men så kan vi jo bruge toppunktsformlen til at beregne den præcise værdi af minimumspunktet, idet koefficienterne til andengradspolynomiet jo som sædvanlig findes på VARS-Statistics-menuen:



Hvordan kan vi nu forstå dette teoretisk? Jo summen af alle afstandskvadraterne består af en masse led, der hver for sig kan regnes ud til et andengradspolynomium. Men så kan summen jo også udregnes som et andengradspolynomium:

$$\begin{aligned} y &= (x-x_1)^2 + (x-x_2)^2 + (x-x_3)^2 + \ldots + (x-x_n)^2 \\ &= (x^2 - 2x_1 \cdot x + x_1^{\,2}) + (x^2 - 2x_2 \cdot x + x_2^{\,2}) + (x^2 - 2x_3 \cdot x + x_3^{\,2}) + \ldots + (x^2 - 2x_n \cdot x + x_n^{\,2}) \\ &= (x^2 + x^2 + x^2 + \ldots + x^2) - (2x_1 \cdot x + 2x_2 \cdot x + 2x_3 \cdot x + \ldots + 2x_n \cdot x) + (x_1^{\,2} + x_2^{\,2} + x_3^{\,2} + \ldots + x_n^{\,2}) \\ &= n \cdot x^2 - 2 \cdot (x_1 + x_2 + x_3 + \ldots + x_n) \cdot x + (x_1^{\,2} + x_2^{\,2} + x_3^{\,2} + \ldots + x_n^{\,2}) \end{aligned}$$

Så det viser netop at summen af afstandskvadraterne er et andengradspolynomium med koefficienterne

$$A = n, \quad B = -2 \cdot (x_1 + x_2 + x_3 + \ldots + x_n) \quad \text{og} \quad C = (x_1^{\,2} + x_2^{\,2} + x_3^{\,2} + \ldots + x_n^{\,2})$$

Af toppunktsformlen får vi så den følgende værdi for x-koordinaten i minimumspunktet:

$$x_T = -\frac{B}{2A} = -\frac{-2 \cdot (x_1 + x_2 + x_3 + \ldots + x_n)}{2n} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \bar{x}$$

Når man skal finde det tal x, der har det mindste samlede afstandskvadrat til et datasæt, skal man altså vælge gennemsnittet af tallene!

4) Medianen som punktet med den mindste *samlede afstand.*

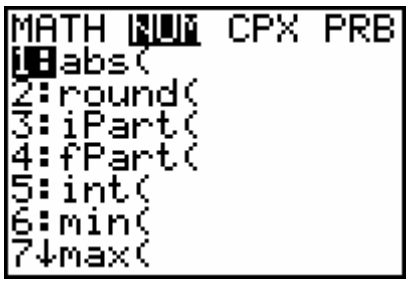
Vi kan give en tilsvarende avanceret karakterisering af *medianen* på den følgende måde: Vi ser på vores datasæt $\{x_1, x_2, x_3, ..., x_n\}$ og stiller os det følgende spørgsmål:

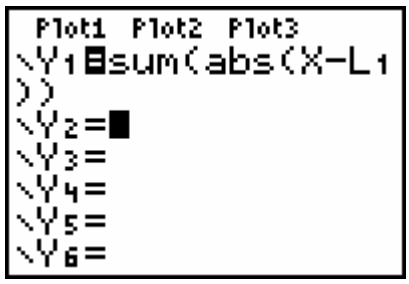
> Hvordan skal vi vælge tallet x, så den samlede afstand til observationerne $\{x_1, x_2, x_3, ..., x_n\}$, dvs. summen
>
> $$y = |x - x_1| + |x - x_2| + |x - x_3| + ... + |x - x_n|$$
>
> er mindst mulig?

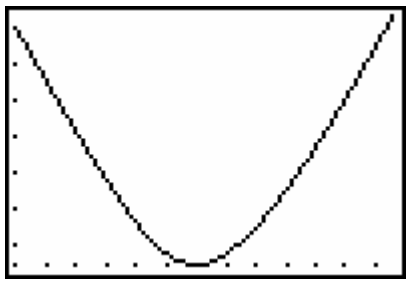
Vi kan igen få en fornemmelse for problemstillingen ved at undersøge den på grafregneren. Her ligger datasættet i listen L_1 og den ovenstående sum af afstandene defineres derfor på følgende måde:

$$Y_1 = \text{sum}(\text{abs}(X\text{-}L_1))$$

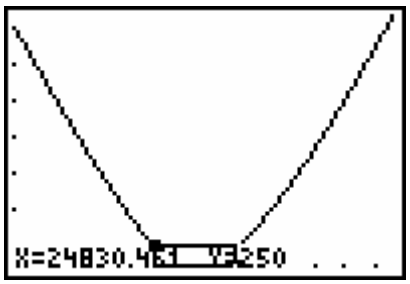
Men så kan vi jo tegne *grafen* for summen og derved bestemme dens minimum:

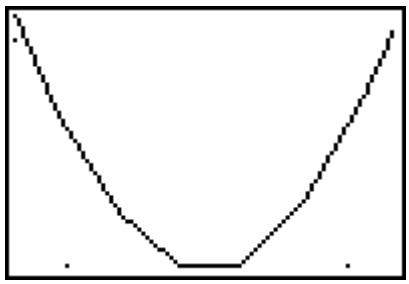


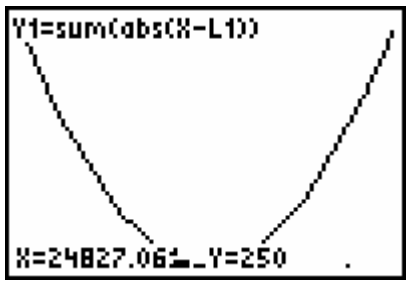




Grafen er tegnet med ZoomFit over et X-interval, som rummer såvel den mindste som den største observation. Denne gang skulle vi gerne kunne se, at den ikke krummer lige så pænt som en parabel. Faktisk er det en sum af stykvis lineære funktioner, idet grafen for en absolutværdifunktion jo er stykvis lineær. Zoomer vi ind på bunden af grafen med Zbox fremtræder denne stykvis lineære karakter tydeligere (med hældningerne 0, ±1, ±2, …):



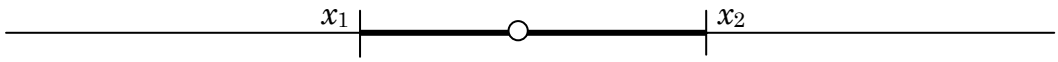




Denne gang kan vi derfor ikke bruge minimumsrutinen fra 2nd-CALC-menuen til at finde minimumspunktet. Til gengæld kan vi trace os frem til bundpunktet, eller som i dette tilfælde, *bundstykket*, idet grafen munder ud i en vandret linje på det laveste stykke, hvor X-værdien går fra 24827 til 24828 med en konstant Y-værdi på 250. Men det er jo netop de *to midterste observationer*, så de løser problemet sammen med alle de mellemliggende værdier. I praksis benytter man derfor – af symmetrigrunde – deres gennemsnit, dvs. tallet 24827.5, altså netop *medianen* af observationssættet!

Hvordan kan vi nu forstå dette teoretisk? Denne gang er det lidt mere tricket, fordi vi ikke har en færdig formel til at finde toppunkter for stykvis lineære funktioner. I stedet bruger vi bare den generelle strategi, at absolutværdifunktioner har deres minimum i et knækpunkt, hvorfor y-værdien må være minimal i et af knækpunkterne, dvs. én af observationerne. Det gør det nemt at finde minimumsværdien i praksis, da der kun er et endeligt antal punkter vi skal prøve igennem, men vi skulle jo også gerne kunne argumentere teoretisk for løsningen.

Det er da praktisk at starte med at se på det simplest mulige datasæt bestående af to forskellige punkter x_1 og x_2:



Der er det nu klart, at hvis tallet x ligger *indenfor* de to punkter x_1 og x_2 er den samlede afstand til de to punkter simpelthen $|x_2 - x_1|$, dvs. specielt er den konstant på dette stykke. Hvis der imod tallet x ligger *udenfor* de to punkter er den samlede afstand summen af afstanden til det nærmeste datapunkt og de to datapunkters indbyrdes afstand, dvs. den er større!

Vi ser nu på et vilkårligt datasæt $\{x_1, x_2, x_3, ..., x_n\}$, som vi har stillet op i *stigende rækkefølge*, dvs.:

$$x_1 \le x_2 \le x_3 \le ... \le x_n$$

Her kan vi nu parre den første observation x_1 med den sidste x_n, den anden observation x_2 med den næstsidste x_{n-1}, osv. Hvis der er et lige antal observationer går parringen op, ellers bliver der den sidste midterste observation tilbage, som vi i givet fald lader danne par med sig selv. *Bredden* af disse par kaldes d_1, d_2, d_3, …, dvs.

$$d_1 = |x_1 - x_n| \quad \text{osv.}$$

Den samlede bredde af alle parrene kaldes D, dvs. $D = d_1 + d_2 + ...$. Så længe vi befinder os udenfor et par kan vi nu gøre den samlede afstand mindre ved at rykke tallet x indenfor parret. Den mindste samlede afstand fås derfor indenfor det inderste par! Hvis der er et lige antal observationer er det altså mellem de to midterste observationer. Hvis der er et ulige antal observationer lander vi derimod præcis i den midterste observation. I begge tilfælde kan vi derfor bruge *medianen* som det tal x, der har den samlede mindste samlede afstand til datasættet, og den samlede afstand er netop summen af bredderne, dvs. tallet D.

Når man skal finde det tal x, der har den mindste samlede afstand til et datasæt, skal man altså vælge medianen af tallene!

Bemærkning: Men læg mærke til, at vi *altid* kan løse problemet ved at vælge tallet x som en *central observation*. Med et ulige antal observationer er løsningen entydig. Men ved et lige antal observationer er der to centrale observationer, der begge kan bruges til at løse problemet. Den mindste samlede afstand antages altså som tidligere bemærket i et eller flere observationspunkter!