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 vi dem ind i en tabel i DataMeter[2]:

Newcombes data

	Returtid	<ny>
1	24828	
2	24826	
3	24833	
4	24824	
5	24834	
6	24756	
7	24827	
8	24816	

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.

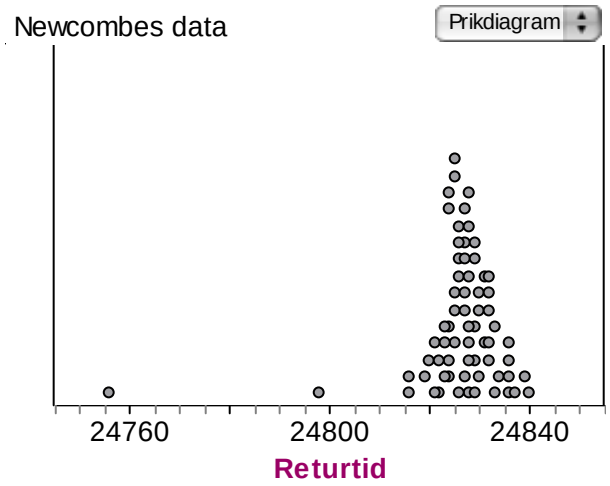
---

[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 DataMeterfil, der er lagt ud i programmet som en eksempelfil under **Filer ➔Åbn eksempel ➔**Statistik ➔ Deskriptiv statistik.

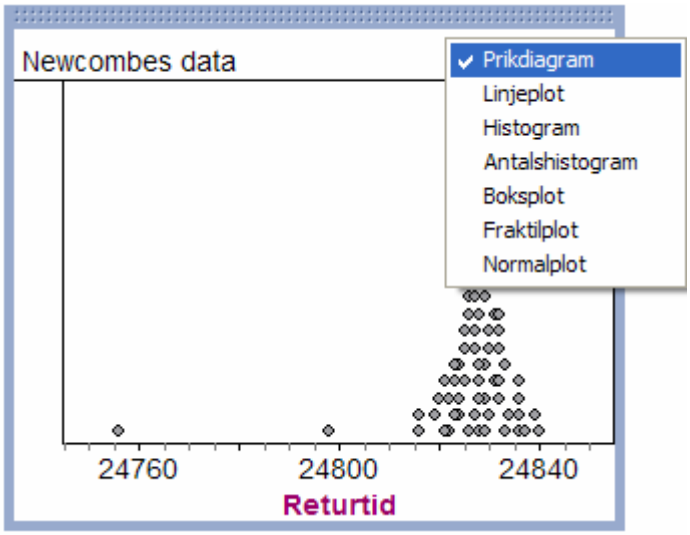
Grafisk inspektion af data

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. Vi trækker derfor et grafvindue ned i dokumentet og trækker variablen **returtid** ind på førsteaksen:

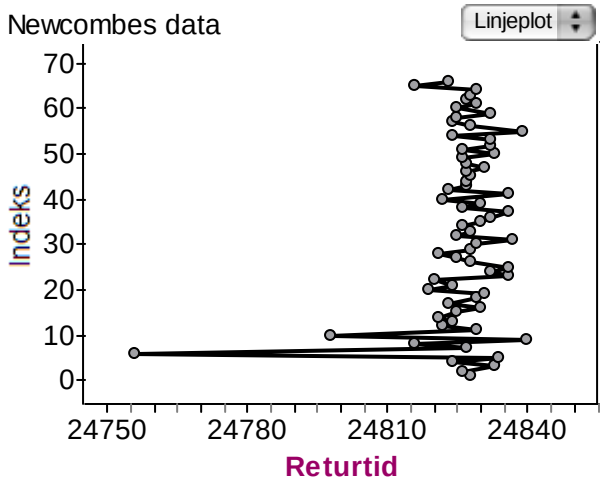


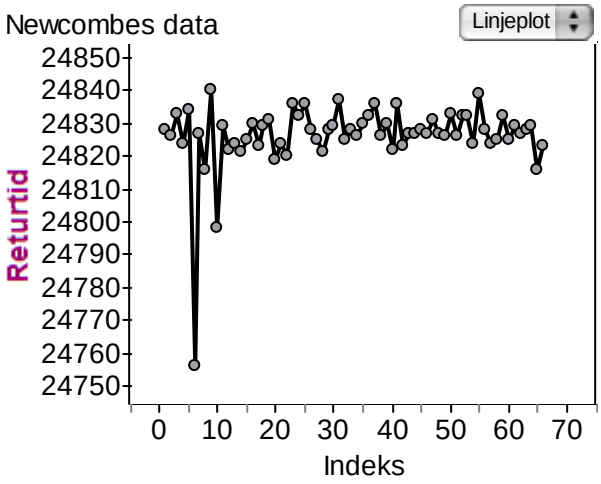
Vi får da umiddelbart et prikdiagram, der giver en første fornemmelse for data. De fleste data ligger klumpet sammen oppe omkring 24830, men der er også to observationer, der skiller sig tydeligt ud fra de andre!

DataMeter tillader selvfølgelig også mange andre plottyper:



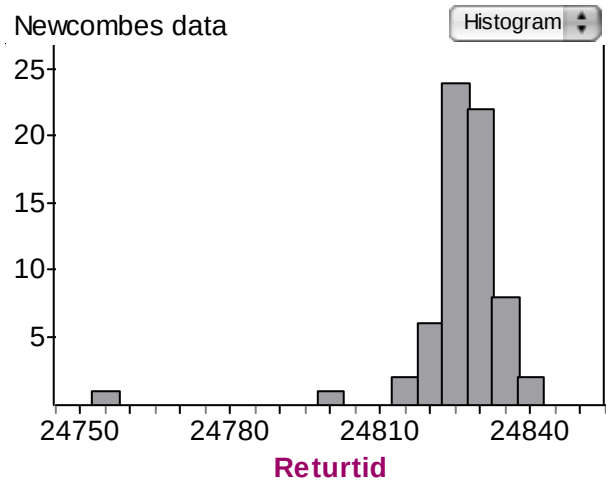
Den næste plottype er **Linjeplottet**, der viser observationerne afsat i rækkefølge. Umiddelbart får vi data afsat lodret, men ved at trække returtiden op på andenaksen fås et tydeligere linjeplot, hvor rækkefølgen (indekset) er afsat ud af førsteaksen. Hvis man kan regne med at rækkefølgen svarer til den tidslige rækkefølge af observationerne (hvilket er korrekt i dette tilfælde) svarer det altså til en tidsserie for datasættet:

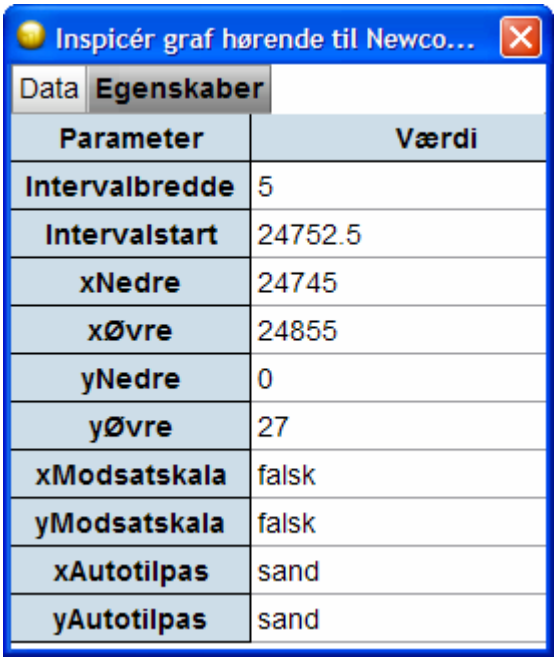




Det er bemærkelsesværdigt at de to 'skæve målinger' forekommer lige i starten. Man kan altså forestille sig at der har været problemer med at indstille udstyret, men at målingerne forholdsvis hurtigt stabiliseres. Der synes i øvrigt ikke at være andre generelle tendenser i tidsserien (der er for eksempel intet der tyder på at målingerne bliver mere og mere præcise som tiden går, idet grafen ikke indsnævres).

Vi kigger derefter på **histogrammet**:

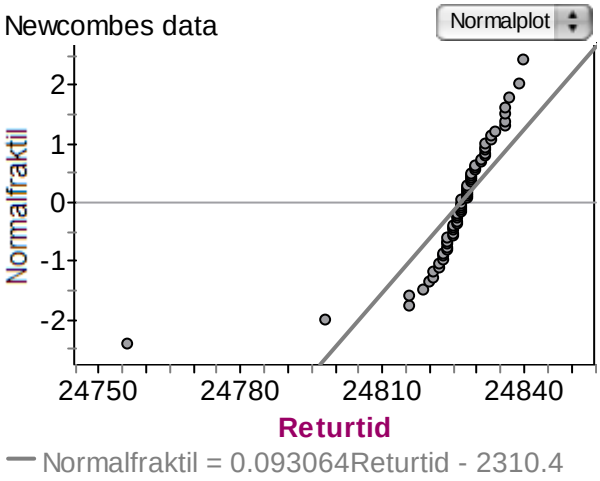


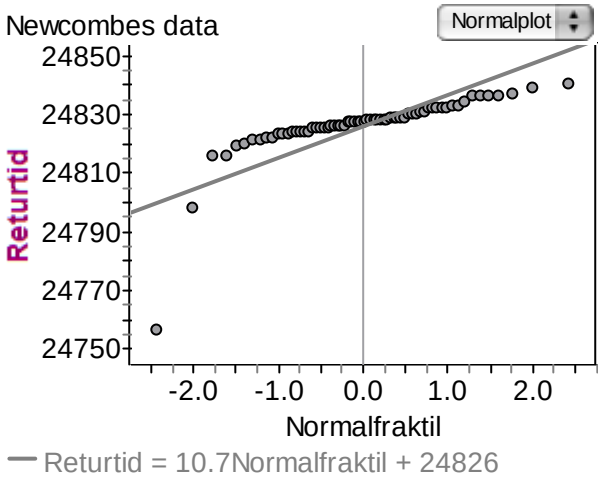


Parameter	Værdi
Intervalbredde	5
Intervalstart	24752.5
xNedre	24745
xØvre	24855
yNedre	0
yØvre	27
xModsatskala	falsk
yModsatskala	falsk
xAutotilpas	sand
yAutotilpas	sand

Vi får altså en klokkeformet fordeling omkring 24830, men stadig med to tydelige undtagelser til venstre for klokken. Noget tyder altså på at de to laveste målinger er perifere. Vi kan selvfølgelig regulere intervalbredderne for histogrammet ved enten at trække i skillelinjerne med musen, eller ved at dobbeltklikke i grafvinduet og gå ind og ændre dem i grafinspektøren. I dette tilfælde får vi dog fra starten en rimelig god oversigt over dataene og beholder derfor blot de foreslåede indstillinger.

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 til et **normalplot**:





Igen skinner de to laveste målinger i øjnene, idet de tydeligt bryder med det retlinjede mønster, der ellers præger dataene. Som det ses følger dataene overhovedet ikke den tilhørende normalfordeling med samme middelværdi og spredning. Læg også mærke til at hvis vi som vist vender om på akserne, dvs. flytter returtiden op på andenaksen kan vi direkte aflæse middelværdien som skæringen med andenaksen og spredningen som hældningen! 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

$$\text{invNorm}(1/132) = -2.4287...,$$

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

$$\text{invNorm}(3/132) = -2.0004...,$$

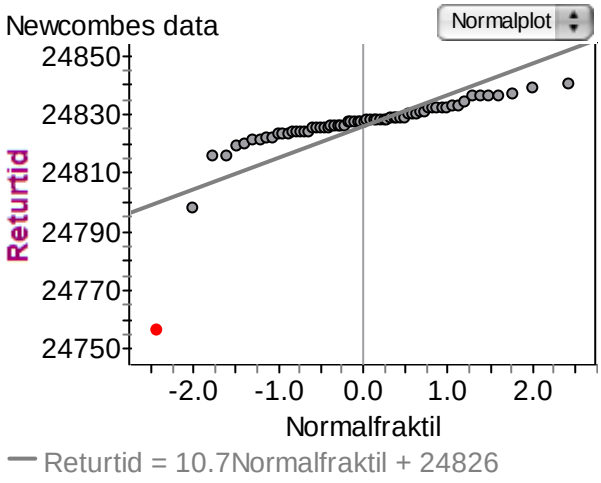
idet 3/132 er midtpunktet i det næste kumulerede frekvensinterval [1/66;2/66] osv.

Ingen data

Slip en variabel her	
	-2.4287371
	-2.0004236

R1 = normInv $\left(\frac{1}{132}\right)$

R2 = normInv $\left(\frac{3}{132}\right)$



(24756 , -2.4287)

Perifere målinger: Hvordan kan vi nu indføre et passende kriterium for hvornår målinger er så perifere, at de kan give anledning til særlige overvejelser? Her kan det være nyttigt med et **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 osv. 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. Generelt er der to særlige krav[3] man ønsker at overholde for medianer, kvartiler osv.:

1. Symmetri
2. Sammenfald med 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. De giver ingen problemer med et lige antal observationer. Men ved et ulige antal observationer er det lidt mere tricket, for her er medianen selv en observation. Vi kan derfor gøre forskellige ting. Vi kan lade medianen høre med til begge halvdele, vi kan udelukke den fra begge halvdele eller vi kan sommetider gøre det ene og somme tider det andet. DataMeter vælger den sidste strategi, idet den udnytter medianen til så vidt muligt at sikre at også kvartilerne falder sammen med observationer, idet vi ved et ulige antal observationer splitter i to halvdele, der igen indeholder et ulige antal observationer. Det kan vi netop opnå fordi vi selv kan bestemme om medianen skal regnes med eller ej.

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, 5,7}, hvor vi har tilføjet medianen for at få et ulige antal observationer. Altså er den første kvartil givet ved Q1 = 3. Tilsvarende består den sidste halvdel af datasættet af observationerne {7, 7, 10, 12, 12}, hvor vi igen har tilføjet medianen for at sikre os at der er et ulige antal observationer, hvorfor den sidste kvartil er givet ved Q3 = 10.

[3] I Danmark har vi tidligere vægtet det sidste krav højest og dermed brudt symmetrien ved at insistere på at medianen altid skal falde sammen med en observation. Med et lige antal observationer har man derfor tidligere ret så arbitrært fastlagt medianen til at være den største af de to midterste observationer.

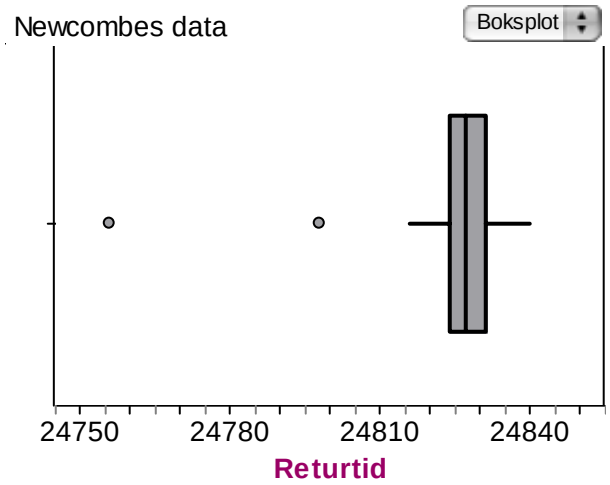
I et **boksplot** afsætter man nu alle observationerne i forhold til de **fem statistiske nøgletal**:

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

min - Q1 - Med - Q3 - maks

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

Vælger vi boksplottet, fås derfor:



Igen skinner de to perifere målinger tydeligt igennem! Men hvordan afgør man i praksis om en observation er perifer? Den amerikanske statistiker Tukey foreslog i 70'erne at man skulle anvende den følgende regel for *perifere observationer*:

Tukeys regel
En observation er **perifer**, 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 **boksplottet** afsættes perifere observationer altså som enkeltpunkter. Begge de to laveste observationer er altså klart perifere! 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 måling 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 at højreklikke på variablen og vælge kommandoen **Sortér stigende**, hvorefter vi kan slette de to første variable:

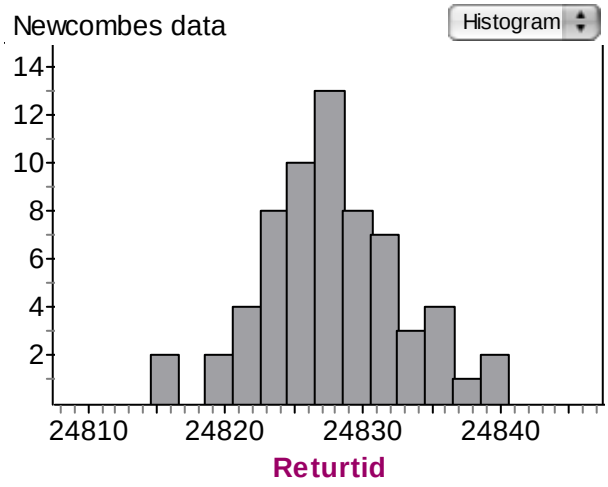
Newcombes data

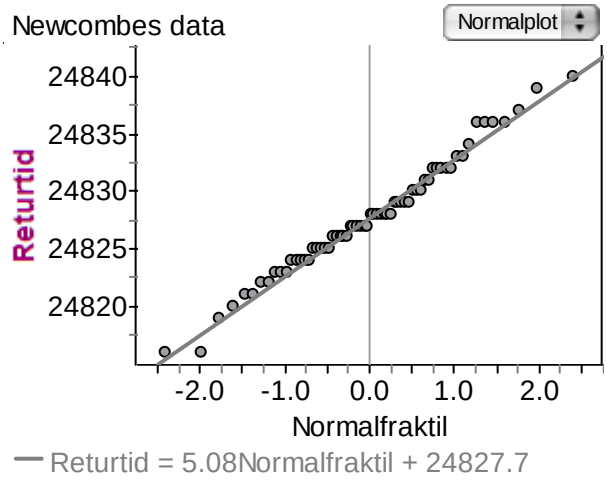
	Returtid	<ny>
1	24756	
2	24798	
3	24816	
4	24816	
5	24819	
6	24820	
7	24821	
8	24821	

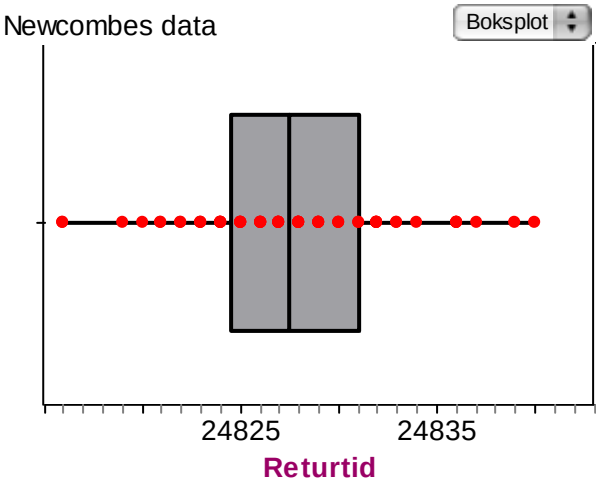
Newcombes data

	Returtid	<ny>
1	24816	
2	24816	
3	24819	
4	24820	
5	24821	
6	24821	
7	24822	
8	24822	

Efter at de er smidt ud, får vi nu pæne histogrammer, normalfordelingsplot og boksplot (efter at vi har tilpasset akserne til grafen og sat intervalbredden til 2).







De bekræfter vores formodning om, at de resterende 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 beregningsværktøjet, som vi trækker ned i dokumentet, hvorefter vi trækker variablen **Returtid** ind i oveersigtstabellen:

Newcombes data

Returtid	24827.75

R1 = middel ( )

I første omgang får vi da kun oplyst middeltallet. Men herefter kan vi ved at højreklikke dels tilføje **den grundlæggende statistik**, dels **fempunktsoversigten**:

Newcombes data

Returtid	24827.75 64 5.0834309 0.63542886 0 24816 24824.5 24827.5 24831 24840

R1 = middel ( )
R2 = tæl ( )
R3 = stdAfv ( )
R4 = stdFejl ( )
R5 = tæl (mangler ( ))
R6 = min ( )
R7 = Q1 ( )
R8 = median ( )
R9 = Q3 ( )
R10 = maks ( )

De første fem oplysninger er middeltallet, antallet af observationer, standardafvigelsen (spredningen), standardfejlen, og hvor mange uoplyste observationer, der foreligger.

De sidste fem observationer er netop de fem nøgletal. Som det ses er der ikke stor forskel på median (24827.5) og middeltal (24827.75), hvilket tyder på en høj grad af symmetri. Tilsvarende får vi også *kvartilerne* oplyst:

Q1 = 24824.5 Q3 = 24831

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 perifere målinger ud, synes det rimeligt at bruge middelværdien som det bedste bud på den 'sande returtid'.

målt returtid ≈ 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

$$\sigma_x = 5.0834\ldots$$

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

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

I dette tilfælde er usikkerheden på gennemsnittet derfor givet ved

$$\frac{\sigma_x}{\sqrt{64}} = 0.63542886$$

Det er altså netop standardfejlen!

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:

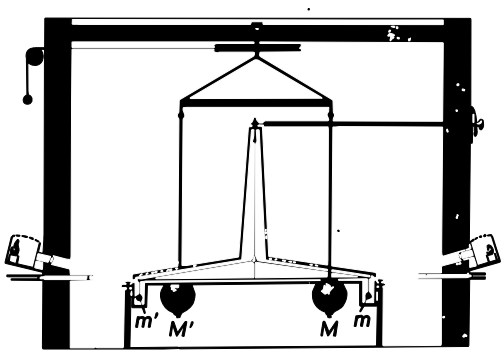
$$\text{Sand(!) returtid} = 24833.02 \text{ ns}$$

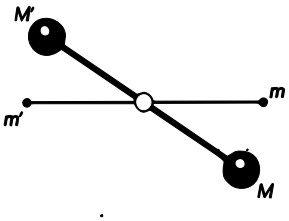
Den ligger klart uden for Newcombes usikkerhedsinterval. Faktisk er forskellen overraskende stor:

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

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**[4]





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 perifere 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?

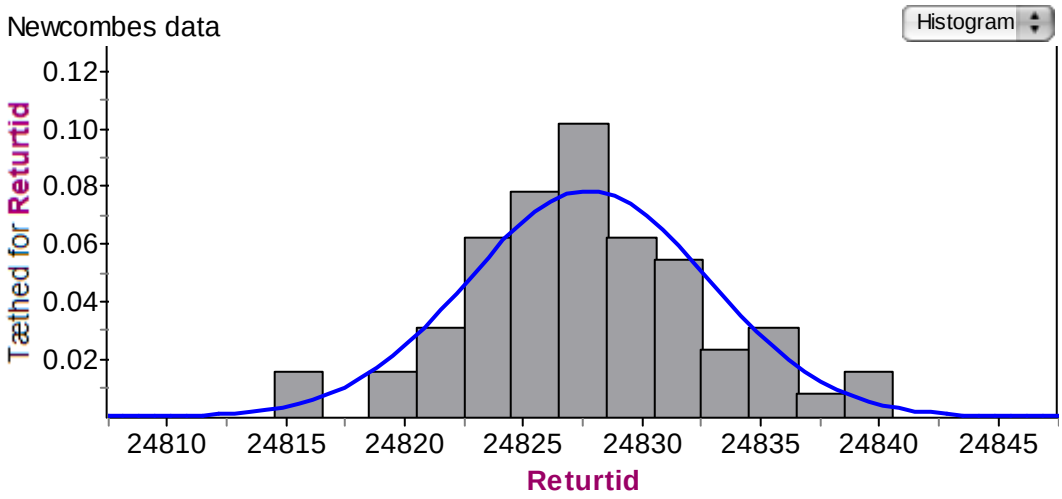
Stemmer den overens med Cavendish' målinger?

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

Avancerede emner

1) Normalfordelingsapproksimationen

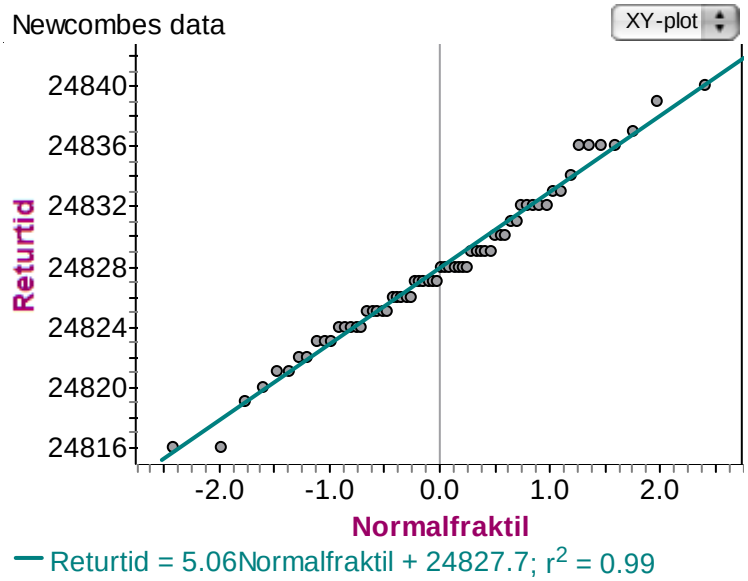
Som vi har set, kan vi tegne histogrammer over måledata nemt og smertefrit, men hvordan får vi indtegnet normalfordelingsapproksimationen til histogrammet? Der skal vi først højreklikke i histogrammet og vælge histogramskalaen **tæthed**. Derved normeres højderne, så det samlede areal under histogrammet netop er 1. Dernæst plotter vi den tæthedsfunktion for normalfordelingen, der har den samme middelværdi og spredning som datasættet:



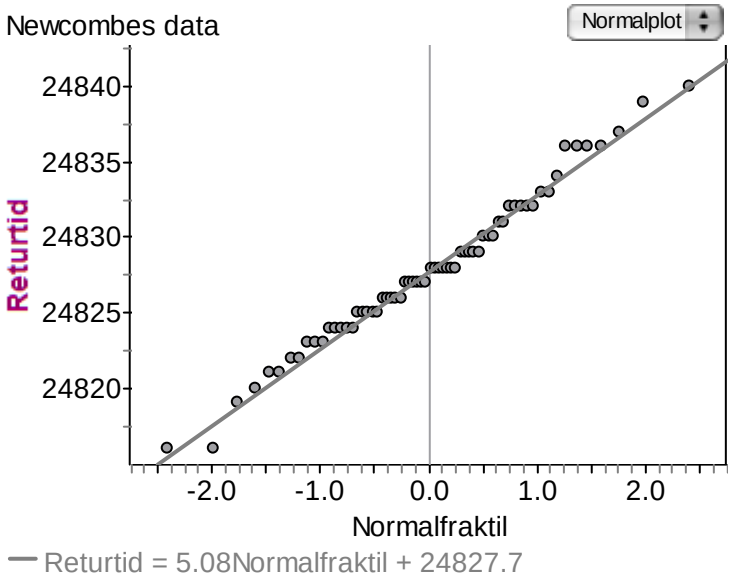
Spørgsmålet er så selvfølgelig om denne approksimation er overbevisende! Men det er jo det samme problem vi har med normalfordelingsplottet: Hvornår ligner det i tilstrækkelig grad en ret linje. Det ville altså være rart om vi kunne checke lidt mere håndfast, hvor god normalfordelingsapproksimationen 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 Sortér stigende kommando som forklaret tidligere. Problemet er blot at vi ikke har direkte adgang til Y-dataene! Vi må derfor beregne dem selv:

Newcombes data

	Returtid	Normalfraktil	<ny>
=		normInv ((2 Indeks − 1) / (2 tæl (Indeks)))	
1	24816	-2.41756	
2	24816	-1.98743	
3	24819	-1.76167	
4	24820	-1.60101	
5	24821	-1.47347	
6	24821	-1.3662	
7	24822	-1.2727	
8	24822	-1.18916	
9	24823	-1.11319	



Dermed er vejen banet for at få udført en almindelig lineær regression på datasættet, og se hvor godt det passer med en ret linje. Som det ses er forklaringsgraden ret så høj 99%), ligesom koefficienterne til mindste kvadraters linje passer fint sammen med den observerede middelværdi og spredning:



En mere præcis udregning af koefficienterne fås fra en beregningsboks (men et større antal decimaler giver selvfølgelig ikke mindre usikkerhed i tallene!)

Newcombes data

	0.98565078
	5.057276
	24827.75

R1 = forklaringsgrad (Normalfraktil; Returtid)
R2 = hældning (Normalfraktil; Returtid)
R3 = skæring (Normalfraktil; Returtid)

2) Tukeys regel

Ifølge Tukeys regel er en observation perifer, 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.

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

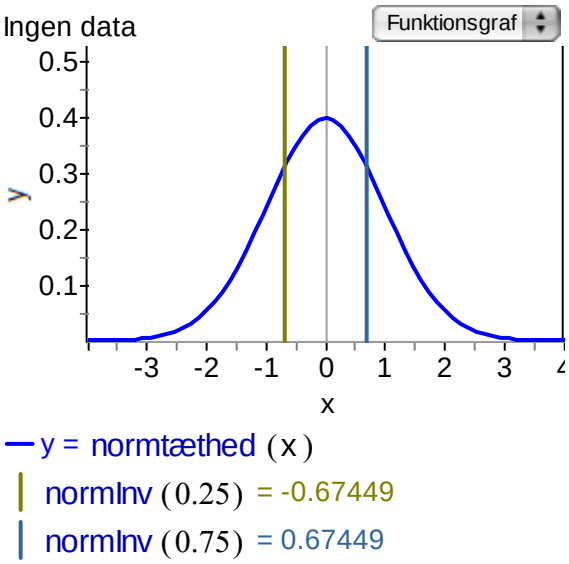
eller ca. 2/3.

Ingen data

Slip en variabel her	
	-0.67448975 0.67448975

$R1 = \text{normInv}(\frac{1}{4})$

$R2 = \text{normInv}(\frac{3}{4})$



I runde tal er kvartilbredden for en standardnormalfordeling derfor 4/3. Halvanden kvartilbredde er derfor i runde tal givet ved 3/2·4/3 = 2. Vi ser derfor at for et normalfordelt observationsmateriale svarer 1½ kvartilbredde ret godt 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 perifere 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 også at der ikke kan forekomme atypiske observationer i ligefordelte observationssæt.

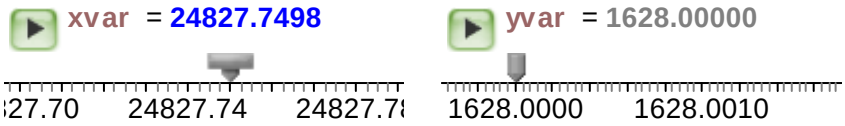
3) Middelværdien som punktet med det mindste *samlede afstandskvadrat.*
Vi kan give en avanceret karakterisering af *middelværdien* på den følgende måde: Vi ser på vores datasæt $\{x_1, x_2, x_3, ..., x_n\}$ og stiller os det følgende spørgsmål:

> 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 grafisk for et konkret datasæt, her Newcombes data. Desværre er det en smule kompliceret at konstruere den pågældende graf, men vi vil forsøge alligevel! Nøglen er dynamiske parametre, hvor den første repræsenterer de uafhængige variabel *xvar* (som vi lader løbe fra den mindste til den største værdi, dvs. fra 24816 til 24840), mens den anden repræsenterer den afhængige variabel *yvar* givet ved formlen:

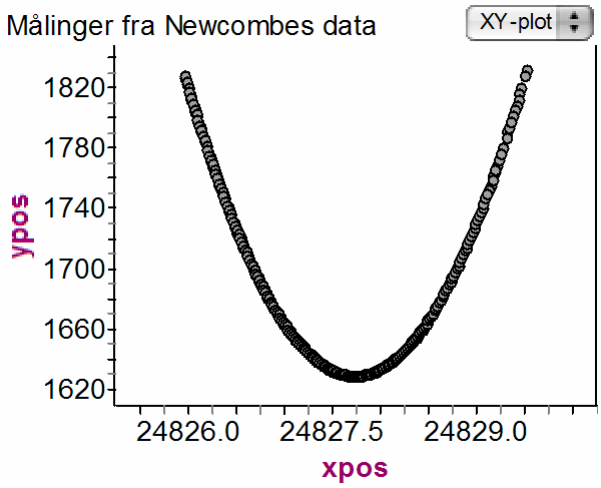
yvar =sum((xvar - Returtid)^2)

Her skal vi huske at trække datasættet ind i den dynamiske skyder for at få adgang til at regne på returtiden! Med disse to skydere til rådighed kan vi allerede få en første fornemmelse for hvor den minimale værdi af *yvar* ligger. Efter en del indzoomning finder vi således:



Det ligger retså tæt på middelværdien for returtiden!

Men vi kan også tegne *grafen* for summen og derved bestemme dens minimum ved at benytte et lidt beskidt trick. Vi opretter to målinger **xpos** og **ypos**, der netop registrerer værdien af **xvar** og **yvar**. Dernæst udfører vi gentagne målinger og trækker de to målinger ind i et grafrum. Endelig sørger vi for at nye målinger tilføjes kontinuerligt, når der sker ændringer i datasættet. Ved at animerer skyderen for xvar får vi nu aftegnet grafen for sammenhængen mellem **xvar** og **yvar**:



Men kig lige på grafen en gang til: Det kunne godt ligne en parabel! Faktisk er det overkommeligt at checke numerisk, at det ikke bare ligner en parabel: Det er en parabel!

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 $\{x_1,x_2,x_3,\ldots,x_n\}$ og stiller os det følgende spørgsmål:

Hvordan skal vi vælge tallet x, så den samlede afstand til observationerne $\{x_1,x_2,x_3,\ldots,x_n\}$, dvs. summen

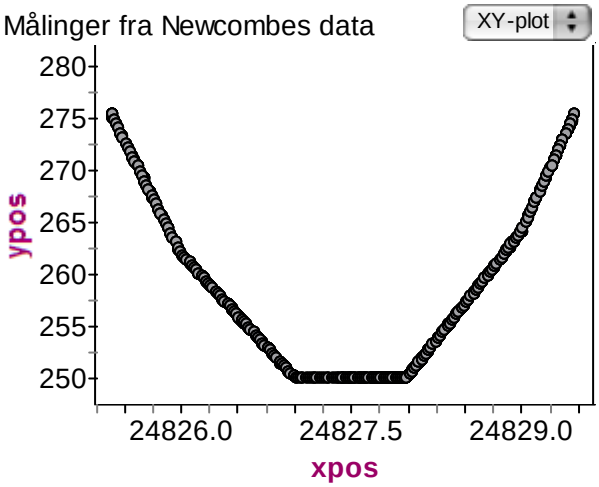
$$y=|x-x_1|+|x-x_2|+|x-x_3|+\ldots+|x-x_n|$$

er mindst mulig?

Vi kan igen få en fornemmelse for problemstillingen ved at undersøge den i DataMeter ved hjælp af den samme komplicerede metode som i det foregående eksempel. Den ovenstående sum af afstandene defineres denne gang på følgende måde:

```
yvar = sum(abs(xvar - returtid))
```

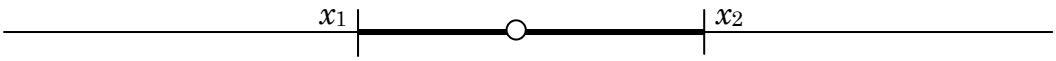
Men så kan vi jo tegne *grafen* for summen og derved bestemme dens minimum:



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. Til gengæld kan vi spore 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!