Note: Deskriptiv statistik med Nspire

KBJ, april 2023

Statistikkens grundlag

Udgangspunktet for statistik er *observationssæt*. Et observationssæt er en liste over observeret data. I et observationssæt med N observationer kan vi repræsentere listen over observationerne ved:

$$o_1, o_2, \dots, o_N$$

Hvert o_i i listen er en observeret værdi. Det meste statistik forudsætter at værdierne er tal, men det kan også være kategorier (f.eks. farver eller partier). Tallet N kaldes *observationssættets størrelse*.

Ugrupperede observationer

I nogle observationssæt optræder den samme observerede værdi mange gange. I en liste over alderen på medlemmerne i en forening vil hver enkelt alder, f.eks. 18 år, optræde mange gange. De værdier der optræder kalder vi *observationsværdierne*. Hvis der er n sådanne kalder vi dem:

$$x_1, x_2, \dots, x_n$$

Listen af observationsværdier er ordnet sådan at: $x_1 < x_2 < x_3 < \cdots < x_n$.

Antallet af gange observationsværdien x optræder i observationssættet kaldes *hyppigheden* af x:

$h(x)$: Hyppigheden af x – antal gange x optræder i observationssættet.

Specielt gælder: $h(x_1) + h(x_2) + \cdots + h(x_n) = N$

Antallet af observationer der er mindre end eller lig med x_i kaldes den *kumulerede frekvens* af x_i:

$$H(x_i) = h(x_1) + h(x_2) + \cdots + h(x_i) = H(x_{i-1}) + h(x_i)$$

Andelen af gange observationsværdien optræder i observationssættet kaldes *frekvensen* af x:

$f(x)$: Frekvensen af x – andel gange x optræder i observationssættet.

Det gælder at: $f(x) = \frac{h(x)}{N}$. Og deraf følger at der gælder: $f(x_1) + f(x_2) + \cdots + f(x_n) = 1$.

Andelen af observationer som er mindre end eller lig med x_i kaldes den *kumulerede frekvens* af x_i:

$$F(x_i) = f(x_1) + f(x_2) + \cdots + f(x_i) = F(x_{i-1}) + f(x_i) = \frac{H(x_i)}{N}$$

Observationsværdien med størst hyppighed/frekvens kaldes for observationssættets *typetal*.

Observationssættets største tal kaldes *maksimum* (max) og det mindste tal kaldes *minimum* (min).

Observationssættets *variationsbredde* kan bestemmes som: $max - min$

Observationssættets *middeltal* eller *gennemsnit* er det tal, som hvis alle observationer erstattes af det, så vil summen af observationerne være det samme. Middeltallet μ bestemmes ved:

$$\mu = \frac{o_1+o_2+\cdots+0_N}{N} = \frac{x_1 \cdot h(x_1)+x_2 \cdot h(x_2)+\cdots+x_n \cdot h(x_n)}{N} = x_1 \cdot f(x_1) + x_2 \cdot f(x_2) + \cdots + x_n$$

Observationssættets *spredning* er et tal som beskriver hvor langt fra middeltallet observationerne i gennemsnit ligger. Spredningen σ bestemmes ved:

$$\sigma = \sqrt{\frac{(o_1-\mu)^2+(o_2-\mu)^2+\cdots+(o_N-\mu)^2}{N}} = \sqrt{\frac{(x_1-\mu)^2 \cdot h(x_1)+(x_2-\mu)^2 \cdot h(x_2)+\cdots+(x_n-\mu)^2 \cdot h(x_n)}{N}}$$

$$\sigma = \sqrt{(x_1-\mu)^2 \cdot f(x_1) + (x_2-\mu)^2 \cdot f(x_2) + \cdots + (x_n-\mu)^2 \cdot f(x_n)}$$

Hvis alle observationerne lægges i voksende rækkefølge fås et *ordnet observationssæt*.

For ulige N vil *medianen* m da være værdien af den midterste af de ordnede observationer, mens at det for lige N vil være gennemsnittet af de to midterste observationer.

For ulige N vil *nedre kvartil* Q_1 være medianen af de observationer der ligger til venstre for den midterste observation, mens det for lige N vil være medianen af den venstre halvdel.

Tilsvarende gælder for ulige N at *øvre kvartil* Q_3 er medianen af de observationer der ligger til venstre for den midterste observation, mens det for lige N vil være medianen af den venstre halvdel.

Generelt siges om et observationssæt at mindst 50% af observationerne er mindre end eller lig m, mens mindst 25% er mindre end eller lig Q_1 og mindst 75% er mindre end eller lig Q_3.

Observationssættets kvartilbredde bestemmes da som: $q = Q_3 - Q_1$.

En observation der ligger mere end 1,5 kvartilbredde under nedre kvartil eller over øvre kvartil kaldes en *outlier*. En observation o er altså en outlier hvis:

$$o < Q_1 - 1{,}5 \cdot q \quad \text{eller} \quad o > Q_3 + 1{,}5 \cdot q$$

Et observationssæt siges at være *venstreskævt* hvis $\mu < m$ og højre skævt hvis $\mu > m$. Om et observationssæt hvor $\mu \approx m$ vil vi sige at det er *ikke-skævt*.

Grupperede observationer

Hvis antallet af observationsværdier er meget stor og/eller hvis hver enkelt værdi har en hyppighed på 1 – vælger man almindeligvis at optælle dem i intervaller. Vi laver således en serie af observationsintervaller: $I_1, I_2, \dots, I_n$ vor $I_1 =]a_0; a_1]$, $I_2 =]a_1; a_2], \dots, I_n =]a_{n-1}; a_n]$.
Ofte vælges intervalbredden til at være den samme for alle I, dvs. $a_i - a_{i-1}$ er ens for alle i.
Intervalmidtpunkterne er da givet ved $m_1, m_2, \dots, m_n$, hvor $m_i = \frac{a_{i-1}+a_i}{2}$

Antal observationer i intervallet I kaldes *intervalhyppigheden* $h(I)$. Antallet af observationer som er mindre end a_i kaldes *kumuleret intervalhyppighed* af I_i: $H(I_i) = h(I_1) + h(I_2) + \cdots + h(I_i)$.

Andel observationer i intervallet I kaldes *intervalfrekvensen* $f(I)$. Andelen af observationer som er mindre end a_i kaldes *kumuleret intervalfrekvens* af I_i: $F(I_i) = f(I_1) + f(I_2) + \cdots + f(I_i)$.

Hvis man kender hele observationssættet kan mange af de statistiske deskriptorer beregnes med samme formler som for ugrupperede observationer. Men ofte kender man kun data i grupperet form. I så fald kan

Middeltallet μ af et grupperet observationssæt kan bestemmes ved formlen:

$$\mu = m_1 \cdot f(I_1) + m_2 \cdot f(I_2) + \cdots + m_n \cdot f(I_n)$$

I denne beregning antages at observationerne inden for et givent interval er jævnt fordelt.

Spredningen kan på tilsvarende vis beregnes som:

$$\sigma = \sqrt{(m_1 - \mu)^2 \cdot f(I_1) + (m_2 - \mu)^2 \cdot f(I_2) + \cdots + (m_n - \mu)^2 \cdot f(I_n)}$$

Intervallet med størst hyppighed/frekvens kaldes for *typeintervallet*.

Ud fra intervallerne kan variationsbredden beregnes som: $a_n - a_0$.

Kvartilsættet er vanskeligt at beregne, men vil almindeligvis skulle aflæses fra en sumkurve.

Stikprøver

Hvis vores observationssæt er en stikprøve og formålet er at estimere fordelingen i en popukation, så estimeres *spredningen* med en lettere modficeret formel:

$$\sigma = \sqrt{\frac{(o_1-\mu)^2+(o_2-\mu)^2+\cdots+(o_N-\mu)^2}{N-1}} = \sqrt{\frac{(x_1-\mu)^2 \cdot h(x_1)+(x_2-\mu)^2 \cdot h(x_2)+\cdots+(x_n-\mu)^2 \cdot h(x_n)}{N-1}}$$

Oprettelse af *lister* i Nspire

	A data
=	
1	3.
2	4.
3	3.
4	5.
5	2.
6	3.
7	5.

Et observationssæt kan bearbejdes i Nspire som en *liste*. Definitionen af observationssættet som listen "data" i et beregninger-vindue kan ske med kommandoen:

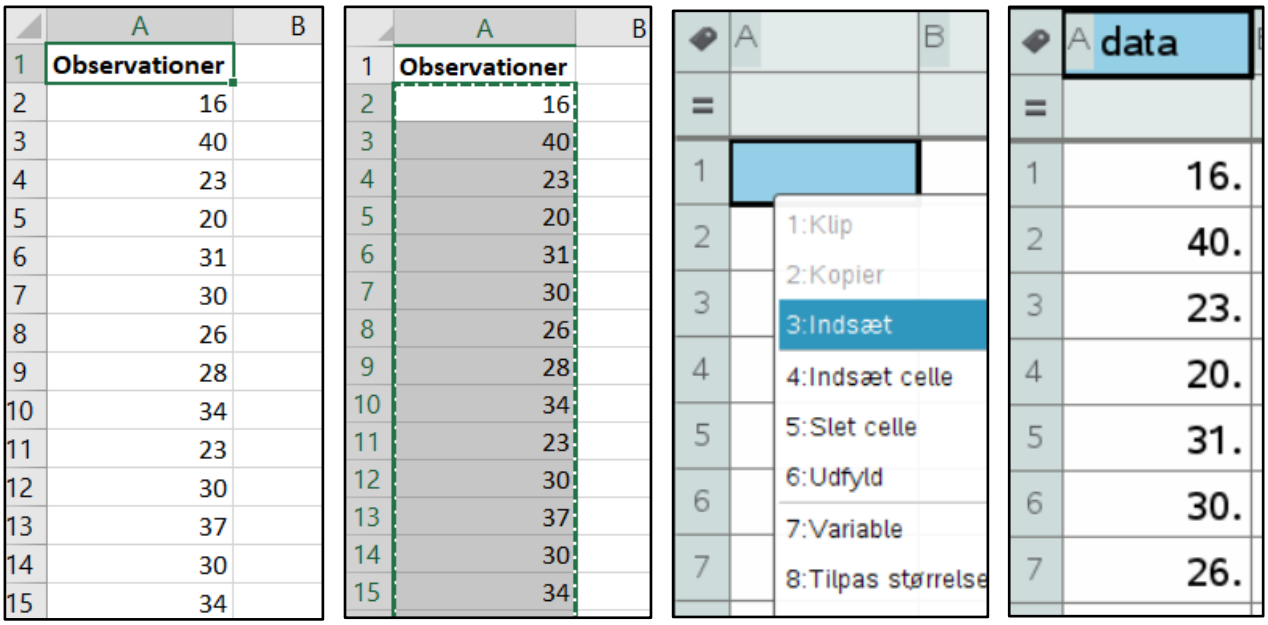
`data:=`$\{o_1, o_2, \ldots, o_N\}$

$data:=\{3,4,3,5,2,3,5,6,7,7,3,1,3\}$

$data:=\{3,4,3,5,2,3,5,6,7,7,3,1,3\}$
$\{3.,4.,3.,5.,2.,3.,5.,6.,7.,7.,3.,1.,3.\}$

En liste kan også oprettes i et "lister og regneark"-vindue ved at sætte listens navn i øverste felt og indtaste værdierne nedenunder, som vist ovenfor til højre.

Hvis en meget stor liste kommer i en Excel-fil som vist herunder, markeres og kopieres alle data i listen (lad overskriften stå). I øverste celle af en liste i et "lister og regenark"-vindue i Excel sættes tallene ind. I øverste felt skrives et ord, som bliver navnet på listen – f.eks. "data".



Proceduren ovenfor virker kun, hvis tallene i Excel er hele tal. Er der komma-tal i mellem, og er der brugt "dansk komma", skal disse konverteres til punktum, for at Nspire forstår dem korrekt.

Det nemmeste i det tilfælde er at kopiere søjlen fra Excel ind i et tomt Word-dokument og vælge "Søg og Erstat" og derefter erstatte "," med ".". Hvis erstat-funktionen i Excel anvendes, virker det ikke altid optimalt. På nogle Mac-computere kommer der et mellemrum mellem hver observation når det indklippes i Nspire. Disse kan normalt ignoreres.

Statistik på lister

Antal observationer i listen "data" bestemmes med kommandoen: `count(data)`

Kommandoen "countif" kan bruges til at tælle bestemte observationer i en liste, som lever op til et bestemt kriterium. I kriteriet anvendes tegnet "?" som repræsentant for "en vilkårlig observation":

Hyppigheden af observationen T i listen "data" bestemmes med: `countif(data,?=T)`

Den *mindste* observation i listen "data" bestemmes med: `min(data)`

Den *største* observation i listen "data" bestemmes med: `max(data)`

Variationsbredden i listen "data" bestemmes derfor med: `max(data)-min(data)`

Summen af observationerne i listen "data" bestemmes med: `sum(data)`

Kvadratsummen af observationerne i listen "data" bestemmes med: `sum(data^2)`

Gennemsnittet af observationer i listen "data" bestemmes med: `mean(data)`

Variansen af observationer i listen "data" bestemmes med: `varpop(data)`

Spredningen af observationer i listen "data" bestemmes med: `stdevpop(data)`

Medianen af observationer i listen "data" bestemmes med: `median(data)`

Observation nr. n i listen "data" kan hentes med kommandoen: `data[n]`

Listen "data" kan gøres til en ordnet liste (voksende) med: `sorta(data)`

Listen "data" kan gøres til en ordnet liste (aftagende) med: `sortd(data)`

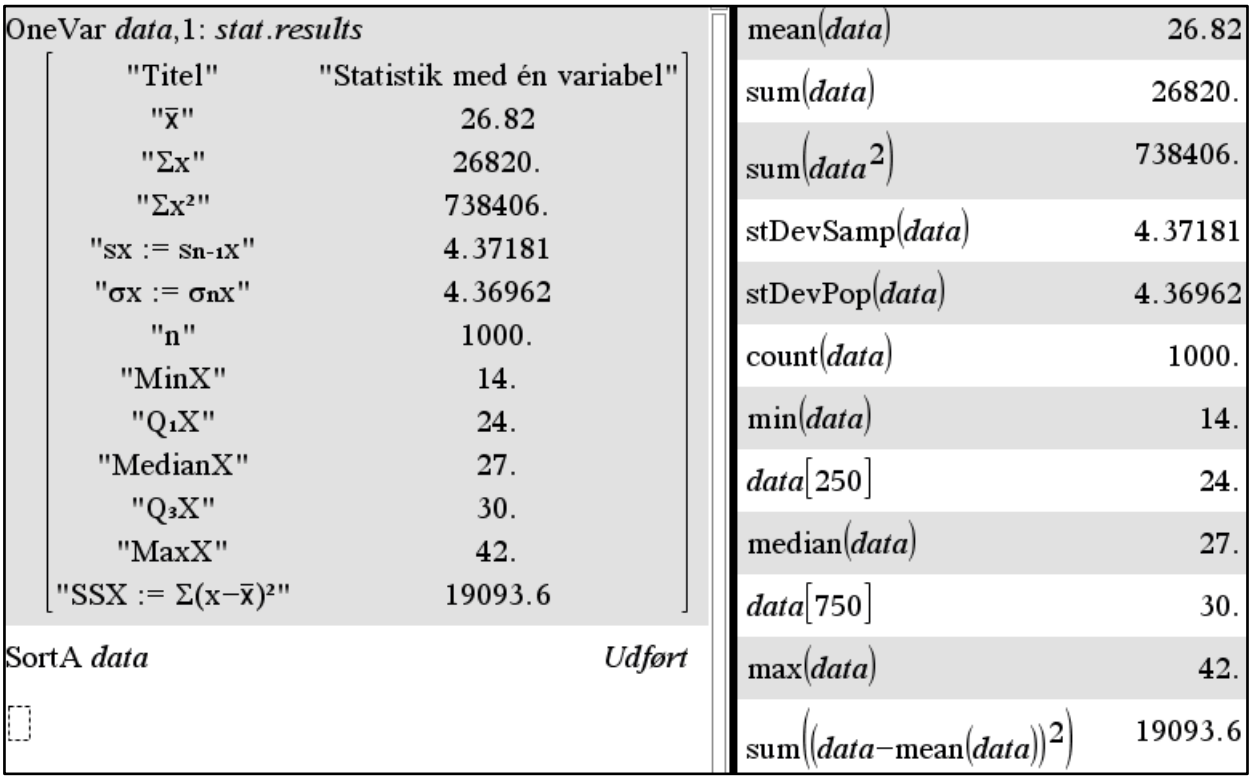
Sorteringskommandoerne virker ved at ændre listen – den definerer ikke en ny liste.
Nedre- og *øvre kvartil* kan således bestemmes ved at sortere listen i voksende rækkefølge og derefter kalde relevant nummer i rækken. Hvis N er delelig med 4, vil kvartilsættet for store observationssæt med stor præcision kunne bestemmes med kommandoerne:

Q_1: `data[N/4]` Median: `data[N/2]` Q_3: `data[3*N/4]`

Hvis et kvartil er lokaliseret som observation nr. n benyttes `data[n]`. Hvis et kvartil er lokaliseret som værdien midt mellem observation n og $n+1$ benyttes: (`data[n]+data[n+1])/2`

I et "Beregninger"-vindue kan man få det meste udregnet på én gang hvis man i værktøjer vælger: "6. Statistik" → "1. Statistike beregninger" → "1. En-variabel statistik…"

Sæt "Antal lister" til "1" og sæt "X1-liste" som "data". Resultatet ses herunder til venstre.
Til højre ses de tilsvarende kommandoer anvendt direkte.



Det ses at der skelnes mellem to typer af *spredning*. Der skelnes mellem om den beregnede spredning opfattes som spredningen i en hel population eller om det er et estimat for *spredningen* i en population ud fra en stikprøve.

Spredning (standardafvigelse) for hel population: `stdevpop(data)`

Stikprøvebaseret estimat for *spredning* i population: `stdevsamp(data)`

Varians (standardafvigelse) for hel population: `varpop(data)`

Stikprøvebaseret estimat for *varians* i population: `varsamp(data)`

En-variabel-statistikken kan kaldes direkte med: `onevar data,1:stat.results`

Tabel over hyppighed, frekvens og kumuleret frekvens for UGRUPPERET data.

Ofte har man brug for en tabel hvor hyppighed, frekvens og kumuleret frekvens af de forskellige observationsværdier optræder. Nogle opgaver starter ved tabellen. Da kan følgende også bruges, i det man så typisk må indtaste observationsværdier og hyppigheder eller frekvenser manuelt.

Der oprettes et "lister og regneark"-vindue. I første søjle vil vi gerne have *observationsværdierne*. Vi kalder derfor søjlen for "obs". Observationsværdierne kan indtastes manuelt. Hvis det er en serie af tal fra *mindste* til *største* værdi kan i formelfeltet bruges kommandoen:

`seq(x,x,min(data),max(data))`

I stedet for "min(data)" og "max(data)" kan tastes andre start- og slutværdier for rækken af tal. Efter slutværdien kan tilføjes et step mellem tallene. Endvidere kan indføres at de gennemløbne tal skal sendes gennem en funktion $f(x)$: `seq(f(x),x,start,slut,step)`.

I anden søjle vil vi gerne have *hyppighederne*, så vi kalder listen "hyp". I øverste celle indtastes:

`=countif(data,?=a1)`

Denne kommando skal kopieres til søjlens øvrige celler – referencen til cellen "a1" tilpasses automatisk ved kopiering. Alternativt kan i formelfeltet bruges: `frequency(data,obs)`

For at beregne *frekvenserne* laves søjlen "frk" med kommandoen: `hyp/count(data)`
Har man ikke en liste med observationer, men kun med hyppigheder, bruges: `hyp/sum(hyp)`

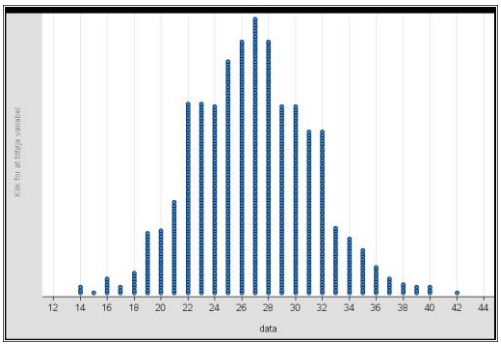
For at beregne *kumuleret frekvens* laves søjlen "kumfrk" med: `cumulativesum(frk)`
Kumuleret hyppighed kan laves som søjlen "kumhyp" med: `cumulativesum(hyp)`

	A obs	B hyp	C frk	D kumfrk
=	=seq(x,x,min(data),max(data))		=hyp/(count(data))	=cumulativesum(frk)
1	14.	=countif(data,?=a1)	0.003	0.003
2	15.	1.	0.001	0.004
3	16.	6.	0.006	0.01

I en tabel over et stort datasæt med obsevationsværdier og kumuleret frekvens kan aflæses:
Nedre kvartil: Den mindste observationsværdi med en kumuleret frekvens på mindst 25%.
Median: Den mindste observationsværdi med en kumuleret frekvens på mindst 50%.
Øvre kvartil: Den mindste observationsværdi med en kumuleret frekvens på mindst 75%.
Endvidere kan gennemsnit nu beregnes som: `sum(obs*frk)`
Og spredning vil kunne udregnes som: `sum((obs-sum(obs*frk))^2*frk)`

Grafiske fremstillinger af UGRUPPERET observationssæt

I et "Diagrammer og statistik"-vindue kan de gængse grafiske repræsentationer af et ugrupperet datasæt laves.



På førsteaksen kan man placere listen med observationssættet ("data"). Som standard tegner Nspire da et *prikdiagram*, som vist her til højre.

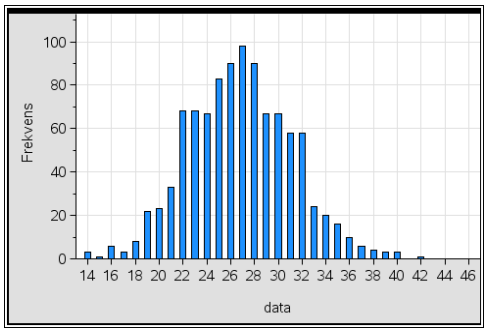
Søjlediagram

Ønskes i stedet et *søjlediagram* kan man i menuen "1. Diagramtyper" vælge "Histogram" (alternativt højreklik på diagram og vælg "Histogram".

Det viste *histogram* er ikke et korrekt søjlediagram. For at lave det højreklikkes på histogrammet og "Søjle-indstillinger" vælges og derpå "Lige store intervaller".

Sæt "Bredde" til "0.5" og Søjlestart til "-0.25".

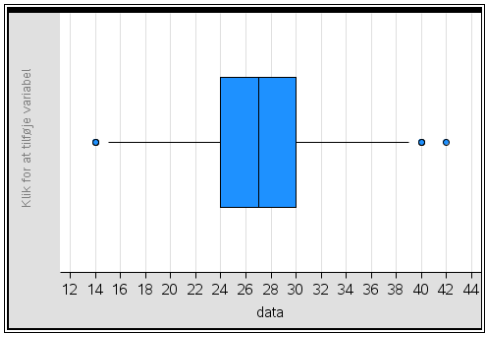
Ved at vælge "Skala" kan vælges om diagrammet skal vise *hyppigheder* på 2.aksen ("tæl") eller *frekvenser* ("procent"). Bemærk at i Nspire hedder hyppighed "frekvens".



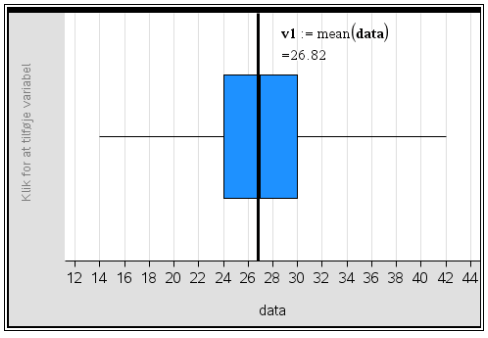
Ved at vælge "Zoom" og "Indstillinger for vindue" kan det justeres hvilken del af 1.aksen der skal vises og hvor høj 2.aksen skal være. Man bør tilpasse dette til diagrammet.

Boksplot

Den sidste grafiske repræsentation er et "Boksplot". Ved at højreklikke i diagram-vinduet og vælge "Boxplot" fås figuren her til højre. Bemærk at der uden for antennerne er "prikker" som repræsenterer datasættets "outlier" – det vil sige observationer der ligger mere end 1,5 kvartilbredde under nedre/over øvre kvartil.



Ved at højreklikke på boksplottet og vælge "Udvid boxplotgrænser" ses boksplottet uden angivelse af outliers. Middeltal kan markeres lodret ved under "Undersøg data" at vælge "plot funktion" og sætte $v1 := mean(data)$.



Tabel over hyppighed, frekvens og kumuleret frekvens for GRUPPERET data.

Hvis et datasæt egnet til gruppering er givet i sin fulde størrelse, kan Nspire lave samme statistik på det, som på ét egnet til at være ugrupperet. Der er således ingen grund til at gentage dette.

Hvis der derimod skal laves en tabelopstilling ser det noget anderledes ud. I stedet for en liste med observationsværdier har vi nu brug for to lister med hhv. start og slutpunkt for hvert observationsinterval. Vi kalder disse to lister for ”start” og ”slut”. Intervallerne skal vælges så de dækker fra mindste til største observation, ofte med samme bredde – hvilket dog ikke er et krav. Først bør altid være et interval der slutter hvor første observationsinterval starter – det kan ofte vælges til at starte i 0. Intervallet skal have intervalhyppighed og kumuleret intervalfrekvens på 0.

Ofte kan med fordel beregnes en liste ”midt” med intervalmidtpunkterne: `(start+slut)/2`

Intervalhyppighederne beregnes i en liste med navnet ”hyp”. I første celle skrives kommandoen:

`countif(data,a1<?<=b1)`

Denne kommando kopieres til de øvrige celler i søjlen. Referencen til a1 og b1 rettes automatisk.

Intervalfrekvenserne beregnes i søjlen ”frk” med kommandoen: `hyp/count(data)`
Hvis ikke der er en liste med hele observationssættet bruges: `hyp/sum(hyp)`

De kumulerede intervalfrekvenser beregnes i søjlen ”kumfrk” med: `cumulativesum(frk)`

	A start	B slut	C midt	D hyp	E frk	F kumfrk	G
=			=(start+slut)/2		=hyp/(sum(hyp))	=cumulativesum(frk)	
1	0.	40.	20.	=countif(**data**,*a1*<?<=*b1*)	0.	0.	
2	40.	50.	45.	3.	0.003	0.003	
3	50.	60.	55.	11.	0.011	0.014	

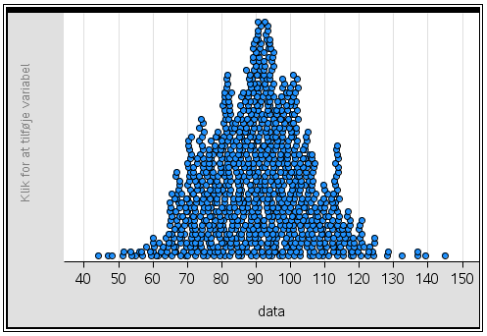
I opgaver hvor kun observationsintervaller og deres intervalhyppighed eller –frekvens kendes, oprettes tabellen på samme måde, idet de kendte oplysninger indtastes direkte.

Middelværdien kan beregnes som: `sum(midt*frk)`

Spredningen kan beregnes som: `sqrt(sum((midt-sum(midt*frk))^2))`

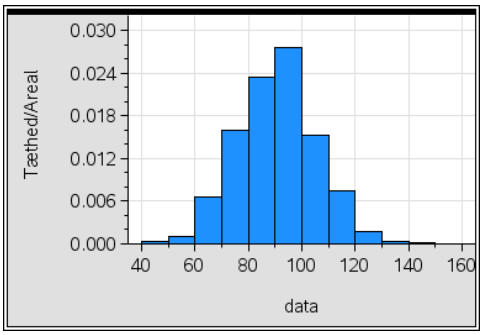
Ud fra de kumulerede frekvenser kan man se hvilket interval nedre og øvre kvartil samt median ligger i, men de kan ikke aflæses/bestemmes. Dette kan kun ske ved aflæsning på en sumkurve.

Grafiske fremstillinger af GRUPPERET observationssæt



I et "Diagrammer og statistik"-vindue sættes listen med data på førsteaksen og der vises et prikdiagram (som er noget mere uorganiseret end for ugrupperet data). Et sådan prikdiagram regnes almindeligvis ikke for særlig interessant.

Histogram

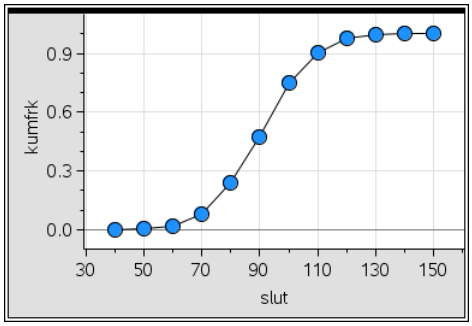


Grupperet data præsenteres med et ægte histogram. Her er det vigtigt at søjlerne fylder hele bredden af det interval hvis frekvens søjlen repræsenterer. Indstil via "Søjleindstillinger" og "Lige store intervaller" til passende intervalbredde (f.eks. den samme som i en evt. tabel) og sæt søjlestart til samme start-værdi som det første observationsinterval.

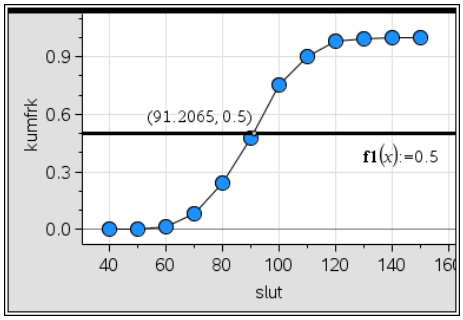
Et rigtigt histogram afspejler intervalfrekvensen af et interval ved arealet af den søjle der står "på det". Derfor bør man i "Skala" vælge "Tæthed/Areal". Vælges i stedet "Tæl" vil højden af søjlen angive intervalhyppigheden og vælges "procent" vil højden angive "intervalfrekvensen". Så længe alle intervaller er lige bredde, er det ikke decideret forkert at vælge "Tæl" eller "Procent" som skala.

I fald man ikke har en liste med alle data, men kun en tabel, tegnes histogrammet ved at sætte listen "start" (med interval-startpunkterne) på førsteaksen og højreklikke på 2.-aksen og vælge "Tilføj Y-værdiliste" og her vælge "frk" (altså listen med frekvenser).

Sumkurve



For at tegne en sumkurve må der først laves en tabel som vist på forrige side. Derpå sættes listen "slut" på førsteaksen og listen "kumfrk" på andenaksen. Der højreklikkes på et punkt i diagrammet og vælges "Forbind datapunkter"



Med "Plot funktion" i menuen "4. Undersøg data" kan man indtegne grafen for $f1(x) = q$ for at aflæse q-fraktilen som førstekoordinatet til skæring mellem linje og sumkurve. Specielt kan *medianen* aflæses som 0,5-fraktilen, *nedre kvartil* som 0,25-fraktilen og *øvre kvartil* som 0,75-fraktilen. Et punkt kan afsættes på den vandrette linje ved at højreklikke på den og vælge "Grafsporing". Det ses at medianen er 91,2.

OVERSIGT OVER NSPIRE-KOMMANDOER

Beskrivelse	**Kommando**
Definition af liste med navnet ”data”	`data:={x,y,...,z}`
Antal værdier i listen ”data”	`count(data)`
Antal værdier i listen ”data” med værdien v	`countif(data,?=v)`
Antal værdier i listen mellem a og b.	`countif(data,a<?<b)`
Mindsteværdi i listen ”data”	`min(data)`
Maksimumværdi i listen ”data”	`max(data)`
Variationsbredde i listen ”data”	`max(data)-min(data)`
Gennemsnit af værdier i listen ”data”	`mean(data)`
Varians af værdier i listen ”data” som population	`varpop(data)`
Varians i population med ”data” som stikprøve	`varsampl(data)`
Spredning af værdier i listen ”data” som population”	`stdevpop(data)`
Spredning i population med ”data” som stikprøve	`stdevpop(data)`
Median af værdier i listen ”data”	`median(data)`
Sortering (stigende) af listen ”data”	`sorta(data)`
Sortering (faldende) af listen ”data”	`sortd(data)`
Hent observation nr. n i listen ”data”	`data[n]`
Sum af værdier af i listen ”data”	`sum(data)`
Sum af kvadrerede værdier i listen ”data”	`sum(data^2)`
Samlet en-variabel-statistik på listen ”data”	`onevar data,1:stat.results`
Liste med værdier fra m til n	`seq(x,x,m,n)`
Listen ”frk” med frekvenser ud fra listen ”hyp”	`frk:=hyp/sum(hyp)`
Kumuleret sum af værdier i rækkefølge i listen ”frk”	`cumulativesum(frk)`